Experiment1-Descriptive Statistics

Descriptive statistics is a set of math used to summarize data. Descriptive statistics can be distribution, central tendency, and dispersion of data.The distribution can be a normal distribution or binomial distribution. The central tendency can be mean, median, and mode. The dispersion or spreadness can be the range, interquartile range, variance, and standard deviation.In this session, you will import a CSV file, Excel file and you will perform basic data processing. I will explain descriptive statistics, central tendency measurements, dispersion measurements. You will look into how R programming can be used to calculate all these values.

What Is Descriptive Statistics?

Descriptive statistics summarizes the data and usually focuses on the distribution, the central tendency, and dispersion of the data. The distributions can be normal distribution, binomial distribution, and other distributions like Bernoulli distribution. Binomial distribution and normal distribution are the more popular and important distributions, especially normal distribution. When exploring data and many statistical tests, you will usually look for the normality of the data, which is how normal the data is or how likely it is that the data is normally distributed. The Central Limit Theorem states that the mean of a sample or subset of a distribution will be equal to the normal distribution mean when the sample size increases, regardless whether the sample is from a normal distribution. The central tendency, not the central limit theorem, is used to describe the data with respect to the center of the data. Central tendency can be the mean, median, and mode of the data. The dispersion describes the spread of the data, and dispersion can be the variance, standard deviation, and interquantile range. Descriptive statistics summarizes the data set, lets us have a feel and understanding of the data and variables, and allows us to decide or determine whether we should use inferential statistics to identify the relationship between data sets or use regression analysis to identify the relationships between variables.

Reading Data Files

R programming allow you to import a data set, which can be comma-separated values (CSV) file, Excel file, tab-separated file, JSON file, or others. Reading data into the R console or R is important, since you must have some data before you can do statistical computing and understand the data. Before you look into importing data into the R console, you must determine your workplace or work directory first. You should always set the current workspace directory to tell R the location of your current project folder. This allows for easier references to data files and scripts.

To print the current work directory, you use the getwd() function:

```
# get the current workspace location
print(getwd());
> print(getwd());
[1] "C:/Users/gohmi/Documents"

#set the current workspace location

setwd("D:/R"); #input your own file directory, for here
we use "D:/R"
> setwd("D:/R");
```

To get the new work directory location, you can use the getwd() function:

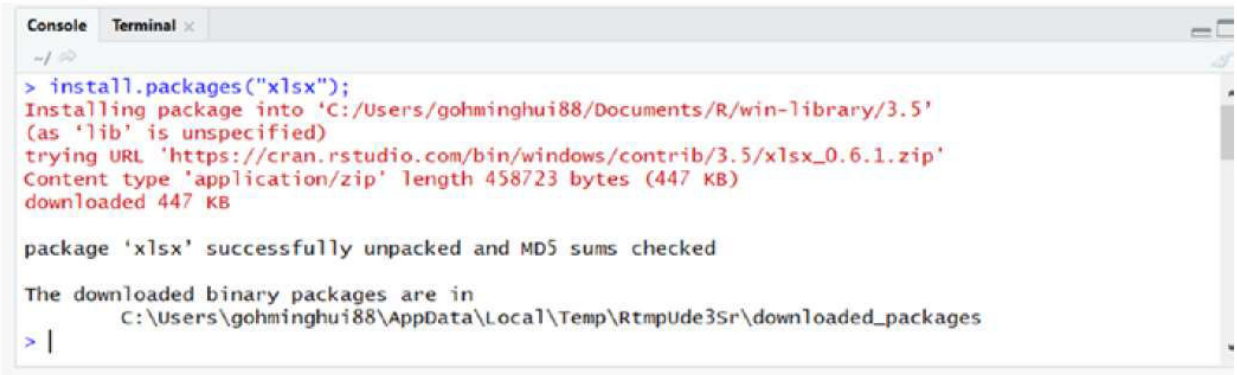
```
#get the new workspace
print(getwd());
> print(getwd());
[1] "D:/R"
```

You can put the data.csv data set into D:/R folder.

Reading an Excel File

The data set can also be in the Excel format or .xlsx format. To read an Excel file, you need to use the xlsx package. The xlsx package requires a Java runtime, so you must install it on your computer. To install the
xlsx package, go to the R console and type the following, also shown in Figure

```
> install.packages("xlsx");
```



To use the xlsx package, use the require() function:

```
> require("xlsx");
Loading required package: xlsx
```

To read the Excel file, you can use the read.xlsx() function:

> data <- read.xlsx(file="data.xlsx", 1);

file is the location of the Excel file. 1 refers to sheet number 1.To view the data variable, you can use the View() function or click the data variable in the Environment portion of RStudio, as shown in Figure.

To look for the documentation of read.xlsx(), you can use the following code.

> help(read.xlsx);

The data variable is of the data frame data type:
> class(data);
[1] "data.frame"

Writing an Excel File

To write a Excel file, you can use the write.xlsx() function:

> write.xlsx(data, file="data2.xlsx", sheetName="sheet1", col.names=TRUE, row.names=FALSE);
data is the variable of data frame type to export to Excel file, file is the file location or path, sheetName is the sheet name, and col.names and row.names are logical values to state whether to export with column names or row names. To view the documentation of the write.xlsx() function or any R function, you can use the help() function.

Basic Data Processing

After importing the data, you may need to do some simple data processing like selecting data, sorting data, filtering data, getting unique values, and removing missing values.

data=read.csv("C:/Users/dkalp/OneDrive/Desktop/spreadsheet.csv")

Mode, Median, Mean

Mean, median, and mode are the most common measures for central tendency. Central tendency is a measure that best summarizes the data and is a measure that is related to the center of the data set.

Mode

Mode is a value in data that has the highest frequency and is useful when the differences are non-numeric and seldom occur.

To get the mode in R, you start with data:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8); #To get mode in a vector, you create a frequency table:
> y <- table(A);
> y;
A
1 2 3 4 5 6 7 8
1 1 1 1 3 1 1 1
```

You want to get the highest frequency, so you use the following to get the mode:

```
> names(y)[which(y==max(y))];
[1] "5"
```

Median

The median is the middle or midpoint of the data and is also the 50 percentile of the data. The median is affected by the outliers and skewness of the data. The median can be a better measurement for centrality than the mean if the data is skewed. The mean is the average, which is liable to be influenced by outliers, so median is a better measure when the data is skewed.

In R, to get the median, you use the median() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> median(A);
[1] 5
```

Mean

The mean is the average of the data. It is the sum of all data divided by the number of data points. The mean works best if the data is distributed in a normal distribution or distributed evenly. The mean represents the expected value if the distribution is random.

In R, to get the mean, you can use the mean() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> mean(A);
[1] 4.6
```

Handle NA Values with mean Function

A typical problem occurs when the data contains NAs. Let's modify our example vector to simulate such a situation:

```
>B=c(A,NA)
>B
[1]  1  2  3  4  5  5  5  6  7  8 NA
```

Our new example vector looks exactly the same as the first example vector, but this time with an NA value at the end. Let's see what happens when we apply the mean function as before:

```
>mean(B)
> [1] NA
```

The RStudio console returns NA – not as we wanted. Fortunately, the mean function comes with the na.rm (i.e. NA remove) option, which can be used to ignore NA values. Let's do this in practice:

```
>mean(B,na.rm=TRUE)
>[1] 4.6
```

As you can see, we get the same mean output as before.
Note:The na.rm option can also be used to ignore NaN or NULL values.

Problem1:Twenty students , graduates and undergraduates, were enrolled in a statistics course. Their ages were
18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36.
a) Find Mean and Median of all students
b) Find median age of all students under 25 years.
c) Find modal age of all student

R code:- >
x=c(18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36)
> mean(x) #mean
[1] 22
> median(x) #median
[1] 20.5
> y=x[x<25]
>median(y)
[1] 20
> xr=table(x) #mode
> mode=which(xr==max(xr))
> mode
20
3

Measures of central tendency for frequency table:-

Problem 2 : A survey of 25 faculty members is taken in a college to study their vocational mobility.They were asked the question “In addition to your present position ,at how many educational instistutes have served on the faculty?.Following is the frequency distribution of their responses .

X	0	1	2	3
f	8	11	5	1

Find mean and median of the distribution
R code:
> x=c(0,1,2,3)
> f=c(8,11,5,1)
> y=rep(x,f)
> mean=(sum(y))/(length(y)) #mean
> mean

[1] 0.96

> median(y) #median

[1] 1

Problem 3 : Compute mean ,median , 1st Quartile, 3rd Quartile and mode of for the following frequency Distribution:

Height in Cm	*145-150*	*150-155*	*155-160*	*160-165*	*165-170*	*170-175*	*175-180*	*180-185*
No. of Adult men	*4*	*6*	*28*	*58*	*64*	*30*	*5*	*5*

```
> x=seq(147.5,182.5,5)
> x
[1] 147.5 152.5 157.5 162.5 167.5 172.5 177.5 182.5
> f=c(4,6,28,58,64,30,5,5)
> mean=sum(x*f)/sum(f)
> mean
[1] 165.175

For Median:

> c=cumsum(f)
> cl=cumsum(f)
> cl
[1]   4  10  38  96 160 190 195 200
> N=sum(f)
> N
[1] 200
> ml=min(which(cl>N/2))
> ml
[1] 5
> h=5
> h
[1] 5
> fm=f[ml]
> fm
[1] 64
> cf=cl[ml-1]
> cf
[1] 96
> l=x[ml]-h/2
> l
[1] 165
> median=l+(((N/2)-cf)/fm)*h #median
> median
[1] 165.3125
```

To find Quartile 1:

```
> Q1=min(which(cl>N/4))
> Q1
[1] 4
> fq1=f[Q1]
> fq1
[1] 58
> cf1=cl[Q1-1]
> cf1
[1] 38
> l=x[Q1]-h/2
```

```
> l
[1] 160
> quartile1=l+(((N/4)-cf1)/fq1)*h
> quartile1
[1] 161.0345
```

To find Quartile 3:

```
> Q3=min(which(cl>3*N/4))
> Q3
[1] 5
> fq3=f[Q3]
> fq3
[1] 64
> cf2=cl[Q3-1]
> cf2
[1] 96
> l=x[Q3]-h/2
> l
[1] 165
> quartile3=l+(((3*N/4)-cf2)/fq3)*h
> quartile3
[1] 169.2188
```

Mode:

```
> m=which(f==max(f))
> m
[1] 5
> f0=f[m]
> f0
[1] 64
> f1=f[m-1]
> f1
[1] 58
> f2=f[m+1]
> f2
[1] 30
> l=x[m]-h/2
> l
[1] 165
> mode=l+((f0-f1)/(2*f0-f1-f2))*h
> mode
[1] 165.75
```

Range,Interquartile Range, Variance, Standard Deviation

Measures of variability are the measures of the spread of the data. Measures of variability can be range, interquartile range, variance, standard deviation, and more.

Range

The range is the difference between the largest and smallest points in the data.
To find the range in R, you use the *range()* function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> range(A);
[1] 1 8
```

To get the difference between the max and the min, you can use

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> res <- range(A);
> diff(res);
[1] 7
```

You can use the min() and max() functions to find the range also:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> min(A);
[1] 1
> max(A);
[1] 8
> max(A) - min(A);
[1] 7
```

To get the range for a data set:

```
> diff(res);
[1] 10.65222
```

Interquartile Range

The interquartile range is the measure of the difference between the 75 percentile or third quartile and the 25 percentile or first quartile.
To get the interquartile range, you can use the IQR() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> IQR(A);
[1] 2.5
```

You can get the quartiles by using the quantile() function:

```
> quantile(A);
0% 25% 50% 75% 100%
1.00 3.25 5.00 5.75 8.00
```

You can get the 25 and 75 percentiles:

```
> quantile(A, 0.25);
25%
3.25
```

```
> quantile(A, 0.75);
75%
5.75
```

The IQR() and quantile() functions can have NA values removed using na.rm = TRUE.

Range measures the maximum and minimum data value , and the interquartile range measures where the majority value is.

Example:
An entomologist studying morphological variation in species of mosquito recorded the following data on body length: 1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3.Compute all the measures of disersion.

```
> x=c(1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3)
> x
 [1] 1.2 1.4 1.3 1.6 1.0 1.5 1.7 1.1 1.2 1.3
> res=range(x)
> res
[1] 1.0 1.7
> diff(res)
[1] 0.7
> var(x)   # Variance
[1] 0.049
> sd(x)    # standard deviation
[1] 0.2213594
> quantile(x)
  0%   25%   50%   75%  100%
1.000 1.200 1.300 1.475 1.700

First Quartile  is 1.2
Second Quartile is 1.3
Third quartile is 1.475

> IQR(x)   # Inter quartile range
[1] 0.275
```

Mean deviation about Mean, Median and Mode:

```
> y=abs(x-mean(x))
> M1=sum(y)/length(y)  # mean deviation about Mean
> M1
[1] 0.176
> z=abs(x-median(x))
> M2=sum(z)/length(z)  # Mean deviation about median
> M2
[1] 0.17
Mean deviation about Mode # in this Problem ,it is a bi-model series (Mode is not
possible)
```

References

1. Biological data analysis, Tartu 2006/2007 (Tech.). (n.d.). Retrieved September 1, 2018, from www-1.ms.ut.ee/BDA/BDA4.pdf.

2. Calculate Standard Deviation. (n.d.). Retrieved from https://explorable.com/calculate-standard-deviation.

3. Descriptive Statistics. (n.d.). Retrieved from http://webspace.ship.edu/cgboer/descstats.html.

4.Descriptive statistics. (2018, August 22). Retrieved from https://en.wikipedia.org/wiki/Descriptive_statistics.

5.Donges, N. (2018, February 14). Intro to Descriptive Statistics – Towards Data Science. Retrieved from https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9.

6. How to Make a Histogram with Basic R. (2017, May 04). Retrieved from www.r-bloggers.com/how-to-make-a-histogram-with-basic-r/.

Exp:2- Correlation and Regression

Correlation Coefficient

The **correlation coefficient** of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related.

Karl Pearson's Correlation

The most commonly used type of correlation is Pearson correlation, named after Karl Pearson, introduced this statistic around the turn of the 20th century.Pearson's r(x ,y) measures the linear relationship between two variables, say X and Y. A correlation of 1 indicates the data points perfectly lie on a line for which Y increases as X increases. A value of -1 also implies the data points lie on a line; however,Y decreases asX increases. The formula for r(x,y) is

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

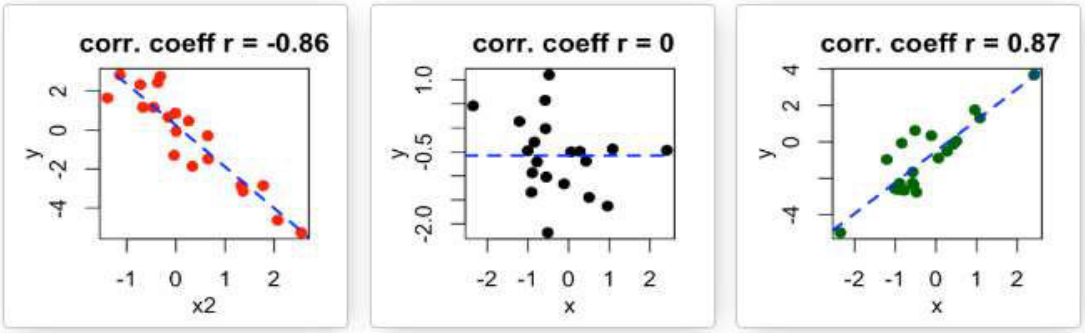
Also,if the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

The Pearson correlation has two assumptions:

1.The two variables are normally distributed. We can test this assumption using
   a.A statistical test (Shapiro-Wilk)
   b.A histogram
   c.A QQ plot

2.The relationship between the two variables is linear. If this relationship is found to be curved, etc. we need to use another correlation test. We can test this assumption by examining the scatterplot between the two variables.

Correlation coefficient is comprised between -1 and 1:



R-code:

To calculate Pearson correlation, we can use the cor() function. The default method for cor( ) is the Pearson correlation. Getting a correlation is generally only half the story, and you may want to know if the relationship is statistically significantly different from 0.

- H_0: There is no correlation between the two variables: $\rho = 0$
- H_a: There is a nonzero correlation between the two variables: $\rho \neq 0$

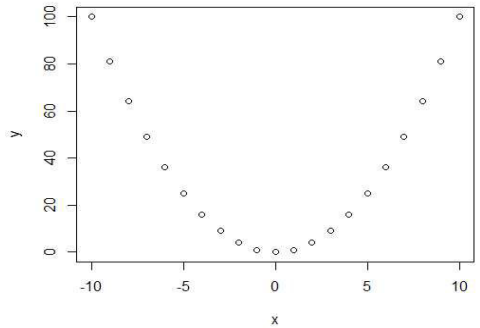
To assess statistical significance, you can use cor.test() function.

Example-1

How the Pearson measure is dependent on the data distribution assumptions (in particular linearity), observe the following deterministic relationship:

$$y = x^2$$

```
> x=seq(-10,10,1)
> y=x^2
> plot(x,y)
> cor(x,y)
[1] 0
```



The R code below computes the correlation between mpg and wt variables in mtcars data set:

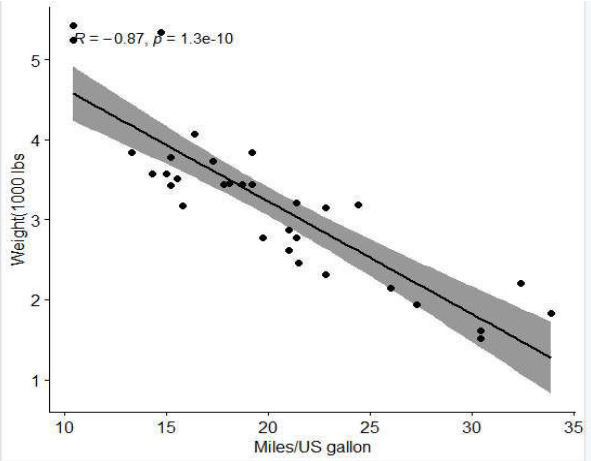
```
> my_data=mtcars
```

```
> head(my_data,6)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Visualize your data using scatter plots

To use R base graphs, click this link: scatter plot - R base graphs. Here, we'll use the ggpubr R package.

```
> library("ggpubr")
>ggscatter(my_data,x="mpg",y="wt",add="reg.line",conf.int=TRUE,cor.coef=TRUE
,cor.method="pearson",xlab="Miles/US gallon",ylab="Weight(1000 lbs")
```



Preleminary test to check the test assumptions

1. Is the covariation linear? Yes, from the plot above, the relationship is linear. In the situation where the scatter plots show curved patterns, we are dealing with nonlinear association between the two variables.
2. Are the data from each of the 2 variables (x, y) follow a normal distribution?
   - Use Shapiro-Wilk normality test –> R function: shapiro.test()
   - and look at the normality plot —> R function: ggpubr::ggqqplot()

Shapiro-Wilk test can be performed as follow:

- Null hypothesis: the data are normally distributed
- Alternative hypothesis: the data are not normally distributed

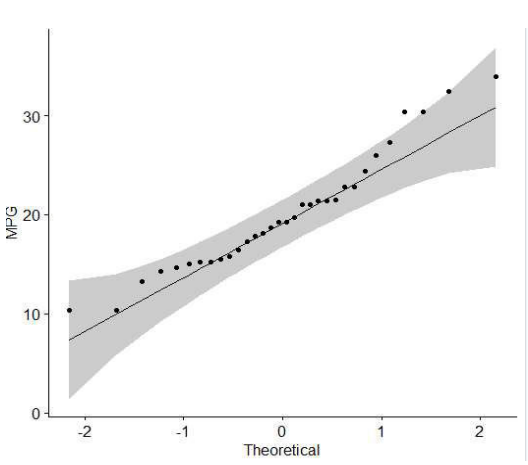
```
# Shapiro-Wilk normality test for mpg
>shapiro.test(my_data$mpg)    # => p = 0.1229
# Shapiro-Wilk normality test for wt
```

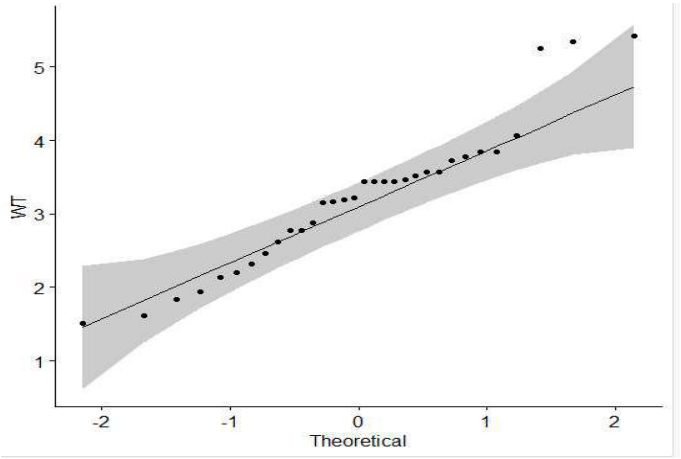
shapiro.test(my_data$wt) # => p = 0.09

From the output, the two p-values are greater than the significance level 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

Visual inspection of the data normality using **Q-Q plots** (quantile-quantile plots). Q-Q plot draws the correlation between a given sample and the normal distribution.

```
>ggqqplot(my_data$mpg,ylab ="MPG")
>ggqqplot(my_data$wt, ylab = "WT")
```





From the normality plots, we conclude that both populations may come from normal distributions.

Note that, if the data are not normally distributed, it's recommended to use the non-parametric correlation, including Spearman and Kendall rank-based correlation tests.

Pearson correlation test

Correlation test between mpg and wt variables:

```
> cor.test(my_data$wt,my_data$mpg,method="pearson")

        Pearson's product-moment correlation

data:  my_data$wt and my_data$mpg
t = -9.559, df = 30, p-value = 1.294e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9338264 -0.7440872
```

```
sample estimates:
       cor
-0.8676594
```

In the result above :

- •t is the t-test statistic value (t = -9.559),
- •df is the degrees of freedom (df= 30),
- •p-value is the significance level of the t-test (p-value = 1.29410^{-10}).
- •conf.int is the confidence interval of the correlation coefficient at 95% (conf.int = [-0.9338, -0.7441]);
- •sample estimates is the correlation coefficient (Cor.coeff = -0.87).

Interpretation of the result

The **p-value** of the test is 1.29410^{-10}, which is less than the significance level alpha = 0.05. We can conclude that wt and mpg are significantly correlated with a correlation coefficient of -0.87 and p-value of 1.29410^{-10} .

Spearman's Correlation Coefficient in R

When the data contains no ties,ρ can be found by taking the difference of the ranked values using the following equation:

$$\rho = 1 - \left[\frac{6\sum d^2}{n(n^2-1)}\right] \qquad \text{[Read the symbol (as 'Rho'.]}$$

Where, $\sum d^2$ = Sum of squares of differences of ranks between paired items in two series n = Number of paired items'

When the data contain ties, the following equation can be used:

$$\rho = 1 - \frac{6(\sum d^2 + C.F_i)}{n(n^2-1)}$$

Spearman's rank correlation coefficient can easily be calculated by again using the cor.test() function in R. The only difference from the Pearson example is the method argument.

```
> cor.test(x=cars$speed,y=cars$dist,method='spearman')

        Spearman's rank correlation rho

data:  cars$speed and cars$dist
S = 3532.8, p-value = 8.825e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8303568
```

Warning in cor.test.default(x = cars$speed, y = cars$dist, method = "spearman"): Cannot compute exact p-value with ties.

Spearman's ρ is reported as .83, slightly higher than Pearson's r of .80, indicating a stronger correlation than previously estimated with Pearson's correlation. Since Pearson's r measures the linear association between two variables, it is likely the outlier in the top right corner of the first graph was causing under-reported values as linear relationships can be extremely affected by outliers.

Obtain the rank correlation coefficient for the following data :

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

```
> x=c(68,64,75,50,64,80,75,40,55,64)

> y=c(62,58,68,45,81,60,68,48,50,70)
> cor.test(x,y,method="spearman")

        Spearman's rank correlation rho

data:  x and y
S = 73.326, p-value = 0.09542
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5555979
Warning message:

In cor.test.default(x, y, method = "spearman") :
  Cannot compute exact p-value with ties

> cor.test(x,y,method="spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  x and y
S = 73.326, p-value = 0.09542
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5555979
```

Linear Regressions

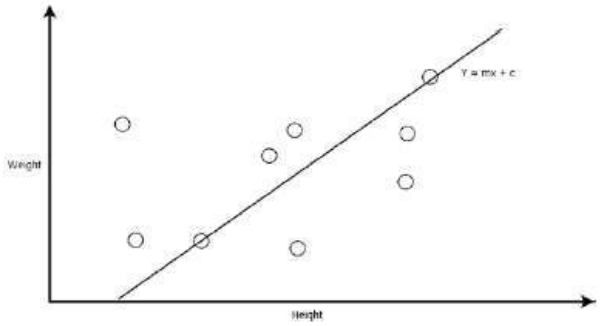
Regression analysis is a form of predictive modelling techniques that identify the relationships between dependent and independent variables(s). The technique is used to find causal effect relationships between variables.

The benefit of using regression analysis is that it identifies the significant relationships between dependent and independent variables and the strength of the impact of multiple independent variables on independent variables.

Linear regression finds the relationship between one dependent variable and one independent variable using a regression line.

The linear regression equation is $y = b_0 + b_1 x$

Where y is the dependent variable and x is the independent variable.



To calculate the slope, you can use

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

To calculate the intercept, you can use

$$b_0 = \bar{y} - b_1 \bar{x}$$

If $b_1 > 0$, x and y have a positive relationship.
If $b_1 < 0$, x and y have a negative relationship.

Steps to Establish a Regression

A simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person.

The steps to create the relationship is −

- Carry out the experiment of gathering a sample of observed values of height and corresponding weight.
- Create a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equation using these
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the weight of new persons, use the **predict()** function in R.

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for **lm()** function in linear regression is −

lm(formula,data)

Following is the description of the parameters used −

- **formula** is a symbol presenting the relation between x and y.
- **Data** is the vector on which the formula will be applied.

Create Relationship Model & get the Coefficients

```
> x <- c(151,174,138,186,128,136,179,163,152,131)

> y <- c(63,81,56,91,47,57,76,72,62,48)
> relation=lm(x~y)
> print(relation)

Call:
lm(formula = x ~ y)

Coefficients:
(Intercept)            y
     61.380        1.415

> print(summary(relation))
```

```
Call:
lm(formula = x ~ y)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0529 -2.4833 -0.0912  1.3774 10.0562

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.3803     7.2653   8.448 2.94e-05 ***
y             1.4153     0.1089  12.997 1.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.712 on 8 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06
```

```
> cor(x,y)
[1] 0.9771296
```

Therefore the regression line is y=61.3803+1.4153x where slope is 1.4153 and intercept is 61.3803.Also there is a positive correlation between x and y.

Interpretation of R-Squared

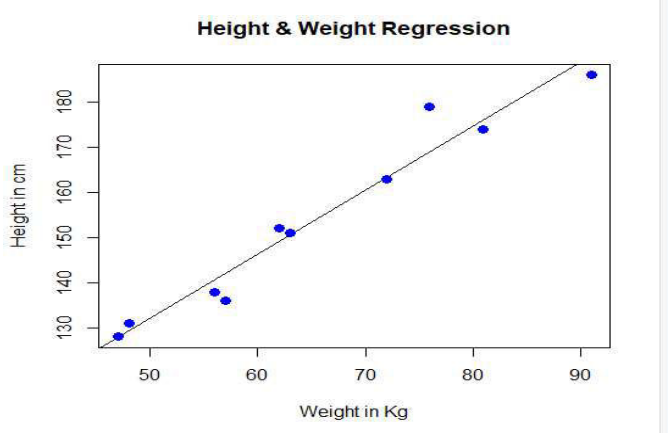
R-squared can take any values between 0 to 1. The R-squared value from the summary is 0.9548, suggesting (correctly here) that X is a good predictor of Y. That is an r-squared of 95% reveals that 95% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

However, it is not always the case that a high r-squared is good for the regression model. The quality of the statistical measure depends on many factors, such as the nature of the variables employed in the model, the units of measure of the variables, and the applieddata transformation. Thus, sometimes, a high r-squared can indicate the problems with the regression model.

A low r-squared figure is generally a bad sign for predictive models. However, in some cases, a good model may show a small value.

Visualize the Regression Graphically

```
> png(file="linear regression.png")
> plot(y,x,col="blue",main ="Height & Weight Regression",abline(lm(x~y)),cex =
1.3,pch = 16,xlab ="Weight in Kg",ylab ="Height in cm")
```



R - Multiple Regression

Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression is −

y = a + b1x1 + b2x2 +...bnxn

Following is the description of the parameters used −

- **y** is the response variable.
- **a, b1, b2...bn** are the coefficients.
- **x1, x2, ...xn** are the predictor variables.

We create the regression model using the **lm()** function in R. The model determines the value of the coefficients using the input data. Next we can predict the value of the response variable for a given set of predictor variables using these coefficients.

lm() Function: This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for **lm()** function in multiple regression is −

```
lm(y ~ x1+x2+x3...,data)
```

Following is the description of the parameters used −

- **formula** is a symbol presenting the relation between the response variable and predictor variables.
- **Data** is the vector on which the formula will be applied.

Example:

```
input <- mtcars[,c("mpg","disp","hp","wt")]
```

```
> model=lm(mpg~disp+hp+wt,data=input)
>
> model

Call:
lm(formula = mpg ~ disp + hp + wt, data = input)

Coefficients:
(Intercept)         disp           hp           wt
  37.105505    -0.000937    -0.031157    -3.800891
```

Based on the above intercept and coefficient values, we create the mathematical equation.

```
Y = 37.15+(-0.000937)*x1+(-0.0311)*x2+(-3.8008)*x3
```

We can use the regression equation created above to predict the mileage when a new set of values for displacement, horse power and weight is provided.

For a car with disp = 221, hp = 102 and wt = 2.91 the predicted mileage is −

```
Y = 37.15+(-0.000937)*221+(-0.0311)*102+(-3.8008)*2.91 = 22.7104
```

References:

1. Correlation. (n.d.). Retrieved from www.mathsisfun.com/data/correlation.html.
2. Covariance. (n.d.). Retrieved from http://mathworld.wolfram.com/Covariance.html.
3. Kabacoff, R. (n.d.). Multiple (Linear) Regression. Retrieved from www.statmethods.net/stats/regression.html.

Exp 3a-Binomial and Poisson Distributions

The Binomial distribution

Consider the following circumstances (binomial scenario):

1. There are n trials.

2. The trials are independent.

3. On each trial, only two things can happen. We refer to these two events as success and failure.

4. The probability of success is the same on each trial. This probability is usually called p.

5. We count the total number of successes. This is a discrete random variable, which we denote by X, and which can take any value between 0 and n (inclusive).

• The random variable X is said to have a binomial distribution with parameters n and p; abbreviated

$$X \sim \text{Bin}(n, p)$$

- It is easy to show that if $X \sim \text{Bin}(n, p)$ then

$$\text{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$.

- $\binom{n}{k}$ is the *binomial coefficient* and is the number of sequences of length n containing k successes.

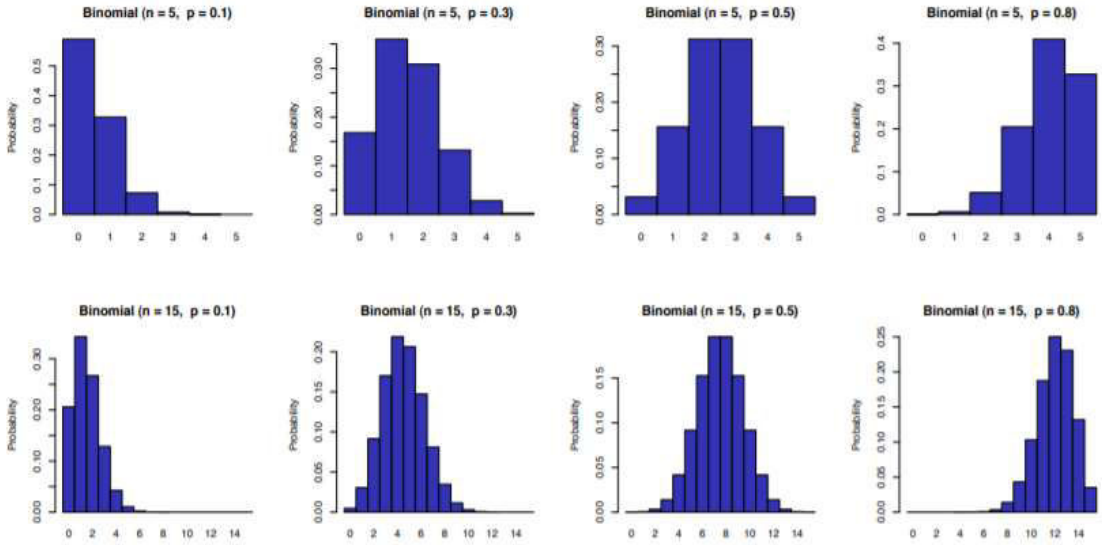
$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

- The expectation and variance of X are given by

$$\text{E}[X] = np$$
$$\text{Var}[X] = np(1-p)$$

The Binomial Distribution: Example

The shape of the distribution depends on n and p.



R has four in-built functions to generate binomial distribution. They are described below.

```
dbinom(x, size, prob)
pbinom(x, size, prob)
qbinom(p, size, prob)
rbinom(n, size, prob)
```

Following is the description of the parameters used −

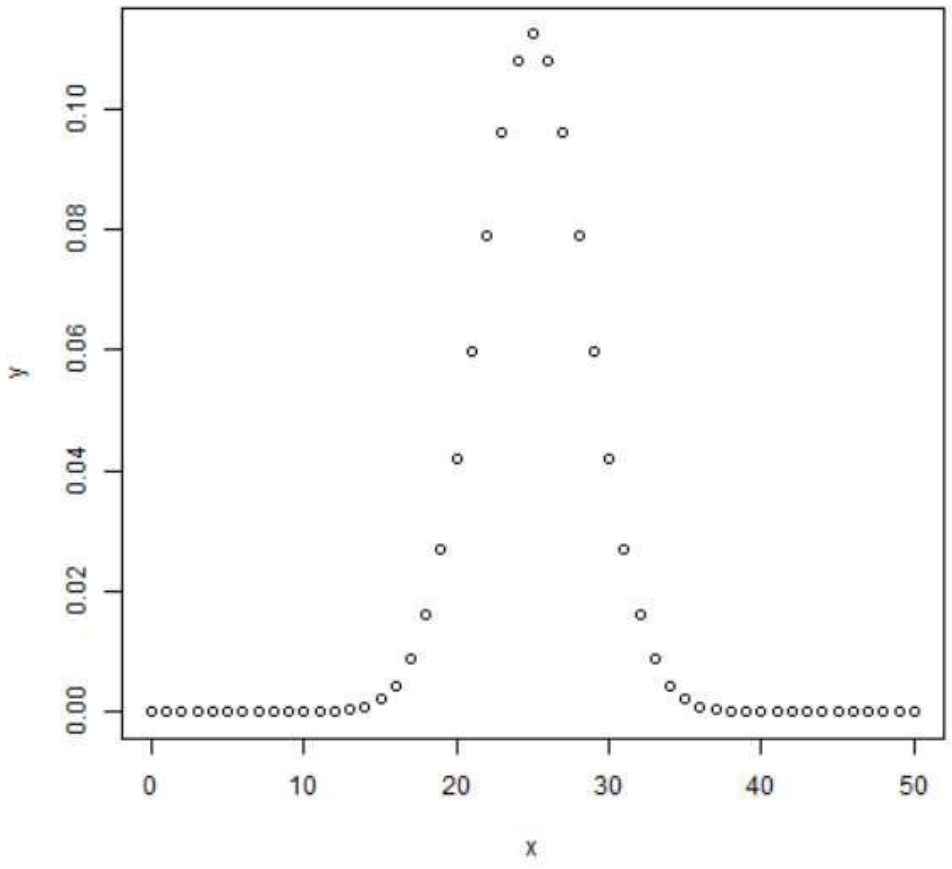
- **x** is a vector of numbers.
- **p** is a vector of probabilities.
- **n** is number of observations.
- **size** is the number of trials.
- **prob** is the probability of success of each trial.

dbinom()

This function gives the probability density distribution at each point.

```
# Create a sample of 50 numbers which are incremented by 1.
x <- seq(0,50,by = 1)

# Create the binomial distribution.
y <- dbinom(x,50,0.5)
```

```
# Give the chart file a name.
png(file = "dbinom.png")

# Plot the graph for this sample.
plot(x,y)

# Save the file.
dev.off()
```



pbinom()

This function gives the cumulative probability of an event. It is a single value representing the probability.

```
# Probability of getting 26 or less heads from 51 tosses of a coin.
x <- pbinom(26,51,0.5)

print(x)
```

[1] 0.610116

qbinom()

This function takes the probability value and gives a number whose cumulative value matches the probability value.

How many heads will have a probability of 0.25 will come out when a coin is tossed 51 times.

```
x <- qbinom(0.25,51,1/2)
print(x)
```
[1] 23

rbinom()

This function generates required number of random values of given probability from a given sample.

Find 8 random values from a sample of 150 with probability of 0.4.

```
x <- rbinom(8,150,.4)
print(x)
```
[1] 58 61 59 66 55 60 61 67

Example:

1. Let X~Bin(5,0.9). Find (a) $P(X \leq 4)$ and $P(X = 4)$

(a)> sum(dbinom(0:4,5,0.9))
[1] 0.40951

(b) > dbinom(4,5,0.9)
[1] 0.32805

2. The proportion of students wearing spectacles is 40%. Let X be the number of students wearing spectacles in a random sample of 10 students. Find

(a) $P(X \leq 2)$; (b) $P(2 \leq X < 5)$; (c) $P(X > 2)$;

(a) > sum(dbinom(0:2,10,0.4))
[1] 0.1672898
Or
>pbinom(2,10,0.4)

[1] 0.1672898

(b)> sum(dbinom(2:4,10,0.4))
[1] 0.5867459

(c) $P(X > 2) = 1 - P(X \leq 2)$
> 1-pbin
om(2,10,0.4)
[1] 0.8327102

3. If a committee has 7 members, find the probability of having more female me mbers than male members given that the probability of having a male or a femal e member is equal.

Sol: The probability of having a female member = 0.5
The probability of having a male member = 0.5
To have more female members, the number of females should be greater than or equal to 4.

```
> 1-pbinom(3,7,0.5)
[1] 0.5
```

4. In a box of switches it is known 10% of the switches are faulty. A technician is wiring 30 circuits, each of which needs one switch. What is the probability that (a) all 30 work, (b) at most 2 of the circuits do not work?

(a) Probability that all 30 work is $P(X = 30) = {}^{30}C_{30}(0.9)^{30}(0.1)^0 = 0.04239$

(b) The statement that "at most 2 circuits do not work" implies that 28, 29 or 30 work. That is $X \geq 28$

$$
\begin{aligned}
P(X \geq 28) &= P(X = 28) + P(X = 29) + P(X = 30) \\
P(X = 30) &= {}^{30}C_{30}(0.9)^{30}(0.1)^0 = 0.04239 \\
P(X = 29) &= {}^{30}C_{29}(0.9)^{29}(0.1)^1 = 0.14130 \\
P(X = 28) &= {}^{30}C_{28}(0.9)^{28}(0.1)^2 = 0.22766
\end{aligned}
$$

Hence $P(X \geq 28) = 0.41135$

```
> dbinom(30,30,0.9)
[1] 0.04239116
```

```
> 1-pbinom(27,30,0.9)
[1] 0.4113512
```

5. If 10% of the Screws produced by an automatic machine are defective, find the probability that out of 20 screws selected at random, there are

(i) Exactly 2 defective (ii) At least 2 defectives

(iii) Between 1 and 3 defectives (inclusive)

(i) # Exactly 2 defective

```
dbinom(2,20,0.10)
[1] 0.2851798
```

(ii) At least 2 defectives

```
1-pbinom(2,20,0.10)
[1] 0.3230732
```

(iii) Between 1 and 3 defectives (inclusive)

```
sum(dbinom(1:3,20,0.10))
[1] 0.74547
```

Poisson Distribution in R

We call it the distribution of rare events., a Poisson process is where DISCRETE events occur in a continuous, but finite interval of time or space in R

The following conditions must apply:

- For a small interval, the probability of the event occurring is proportional to the size of the interval.
- The probability of more than one occurrence in the small interval is negligible.
- Each occurrence must be independent of others and must be at random.
- The events are often defects, accidents or unusual natural happenings, such as an earthquake.
- The parameter for the Poisson distribution is a lambda. It is average or mean of occurrences over a given interval.
- The probability function is: for $x = 0,1.2,3$ ....

Difference between Binomial and Poisson Distribution in R

Binomial Distribution:

- Fixed no. of Trials (n) [10 pie throws], although, only two possible outcomes are possible.
- A probability of success is constant(p).
- Each trial is independent.
- Also, it predicts no.s of successes within a set no. of trials.
- We use it to test for independence.

Poisson Distribution

- Infinite no. of trials.
- Also, it has unlimited no. of outcomes possible.
- The mean of the distribution is the same for all intervals.
- No. of occurrence in any given interval independent of others.
- Also, it predicts no. of occurrences per unit, time, space.
- We use it to test for independence.

R-Code

- dpois(x, lambda) # **the probability of x successes in a period when the expected number of events is lambda**
- ppois(q, lambda) # **the cumulative probability of less than or equal to q successes**
- qpois(p, lambda) # **returns the value (quantile) at the specified cumulative probability (percentile) p**
- rpois(n, lambda) # **returns n random numbers from the Poisson distribution**

Practice problems:

1. What is P(X = 4) with lambda 2.6?

```
> dpois(4, lambda = 2.6)
[1] 0.1414218
```

2. What is $P(X \geq 2)$ with lambda 3?

> **1-ppois(2,3)**

[1] 0.5768099

2. Consider a computer system with Poisson job-arrival stream at an ` average of 2 per minute. Determine the probability that in any one-minute interval there will be

(i) 0 jobs
(ii) Exactly 3 jobs
(iii) at most 3 arrivals

Solution:

Job arrivals lambda =2

(i) No job arrivals

```
> dpois(0,2)
[1] 0.1353353
```

(ii) Exactly 3 jobs

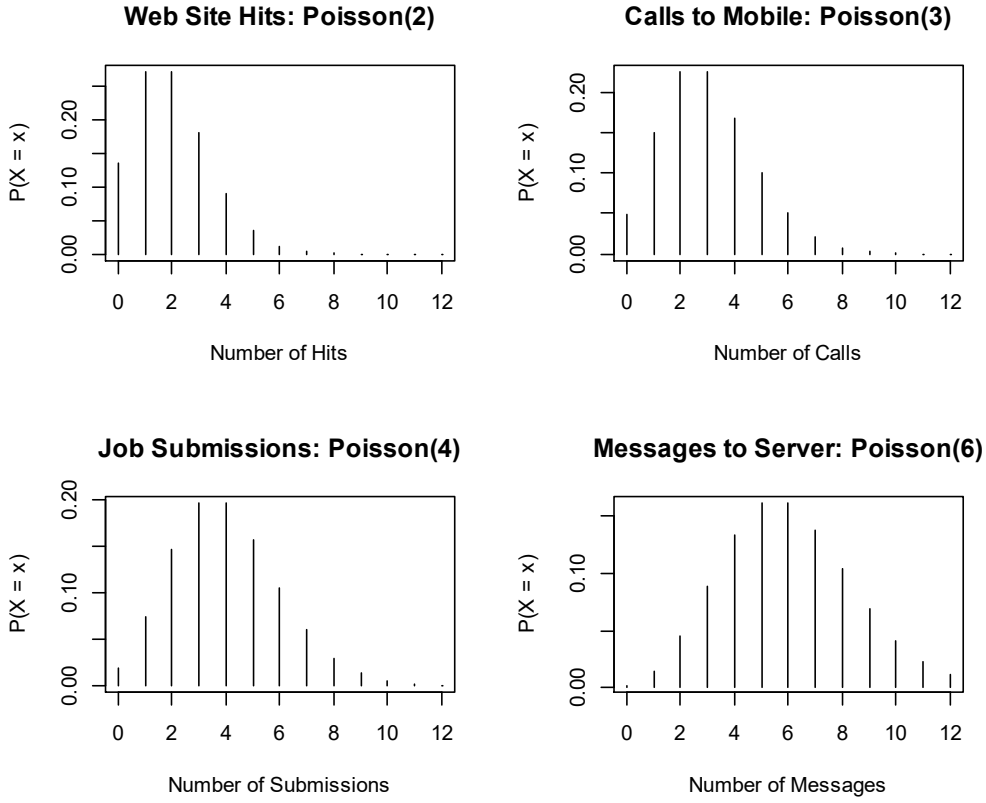
```
> dpois(3,2)
[1] 0.180447
```

(iii) Atmost 3 job arrivals

```
> ppois(3,2)
[1] 0.8571235
```

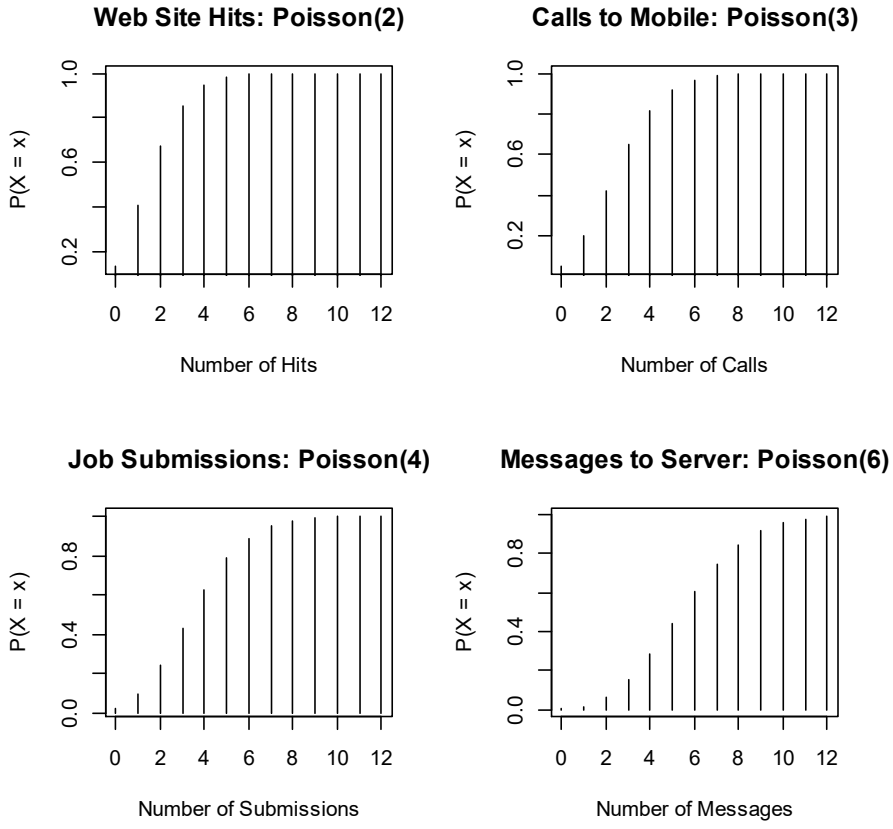
Poisson Probability Density Functions

```
par(mfrow = c(2,2))
# multiframe
x<-0:12 #look at the first 12 probabilities
plot (x, dpois(x, 2), xlab = "Number of Hits", ylab = "P(X = x)", type = "h
", main= "Web Site Hits: Poisson(2)")
plot (x, dpois(x, 3), xlab = "Number of Calls", ylab = "P(X = x)", type = "
h", main= "Calls to Mobile: Poisson(3)")
plot (x, dpois(x, 4), xlab = "Number of Submissions", ylab = "P(X = x)", t
ype = "h", main= "Job Submissions: Poisson(4)")
plot (x, dpois(x, 6), xlab = "Number of Messages", ylab = "P(X = x)", typ
e = "h", main= "Messages to Server: Poisson(6)")
```



Poisson Cumulative Distribution Functions

```
par(mfrow = c(2,2))

# multiframe

x<-0:12   #look at the first 12 probabilities

plot (x, ppois(x, 2), xlab = "Number of Hits", ylab = "P(X = x)", type =
"h", main= "Web Site Hits: Poisson(2)")

plot (x, ppois(x, 3), xlab = "Number of Calls", ylab = "P(X = x)", type =
"h", main= "Calls to Mobile: Poisson(3)")

plot (x, ppois(x, 4), xlab = "Number of Submissions", ylab = "P(X = x)",
type = "h", main= "Job Submissions: Poisson(4)")

plot (x, ppois(x, 6), xlab = "Number of Messages", ylab = "P(X = x)",
type = "h", main= "Messages to Server: Poisson(6)")
```



Practice problems:

1. A recent national study showed that approximately 55.8% of college students have used Google as a source in at least one of their term papers. Let X equal the number of students in have used Google as a source:

a) Find the probability that X is equal to 17
b) Find the probability that X is at most 13.
c) Find the probability that X is bigger than 11.
d) Find the probability that X is at least 15.
e) Find the probability that X is between 16 and 19,
f) Give the mean of X
g) Give the variance of X.
h) Find E(4X + 51.324)

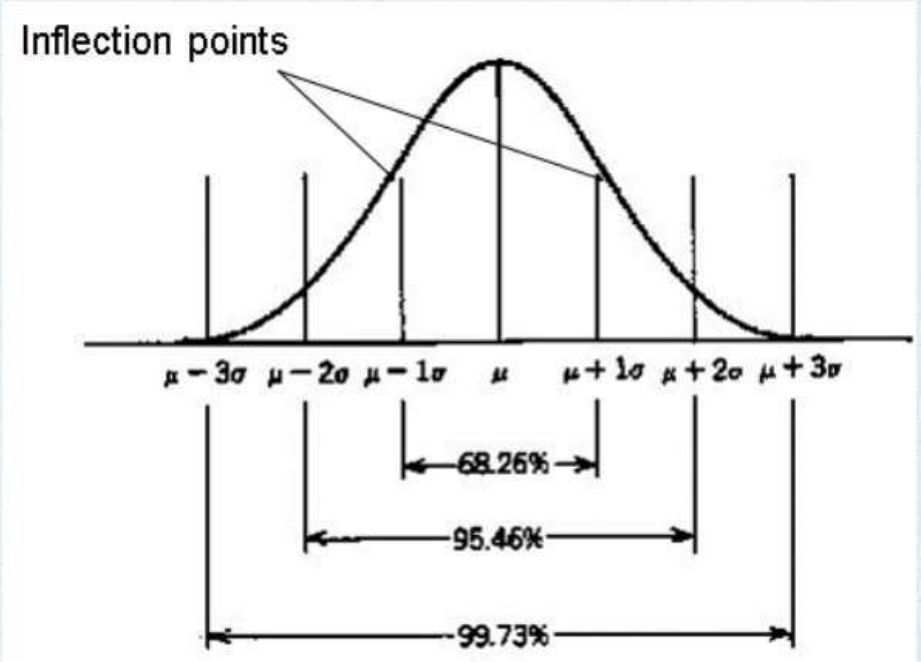
Exp 3b- Normal Distribution

In a random collection of data from independent sources, it is generally observed that the distribution of data is normal. Which means, on plotting a graph with the value of the variable in the horizontal axis and the count of the values in the vertical axis we get a bell shape curve. The centre of the curve represents the mean of the data set. In the graph, fifty percent of values lie to the left of the mean and the other fifty percent lie to the right of the graph. This is referred as normal distribution in statistics.

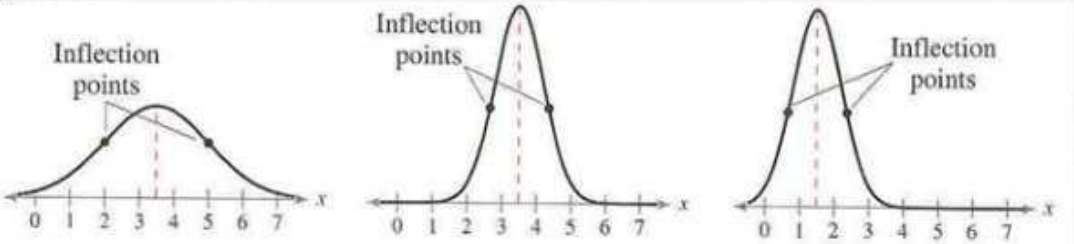
Properties:

A normal distribution is a continuous probability distribution for a random variable, x. The graph of a normal distribution is called the normal curve. A normal distribution has the following properties.

1. The mean, median and mode are equal.
2. The normal curve is bell-shaped and is symmetric about the mean.
3. The total area under the normal curve is equal to 1.
4. The normal curve approaches, but never touches the x-axis as it extends farther and farther away from the mean.
5. Between $\mu - \sigma$ and $\mu + \sigma$ (in the center of the curve) the graph curves downward. The graph curves upward to the left of $\mu - \sigma$ and to the right of $\mu + \sigma$. The points at which the curve changes from curving upward to curving downward are called **inflection points.**



6. A normal distribution can have any mean and any positive standard deviation. These two parameters, μ and σ completely determine the shape of a normal curve. The mean gives the location of the line of symmetry and the standard deviation describes how much the data are spread out.



Mean: $\mu = 3.5$
Standard deviation:
$\sigma = 1.5$

Mean: $\mu = 3.5$
Standard deviation:
$\sigma = 0.7$

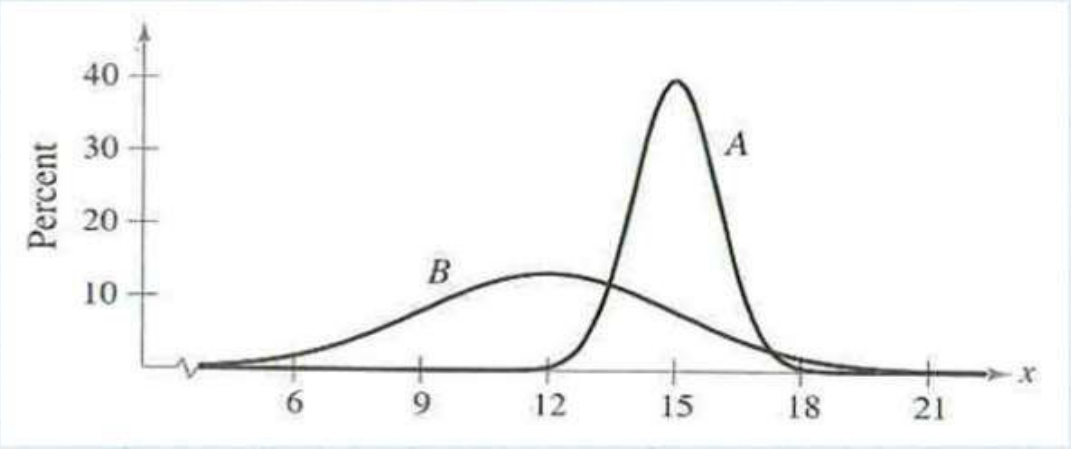
Mean: $\mu = 1.5$
Standard deviation:
$\sigma = 0.7$

See the line of symmetry for each? That's the mean. However, if it is fatter, then the standard deviation is greater. That's the difference.

Understanding Mean & Standard Deviation

Which normal curve has a greater mean?

Which normal curve has a greater standard deviation



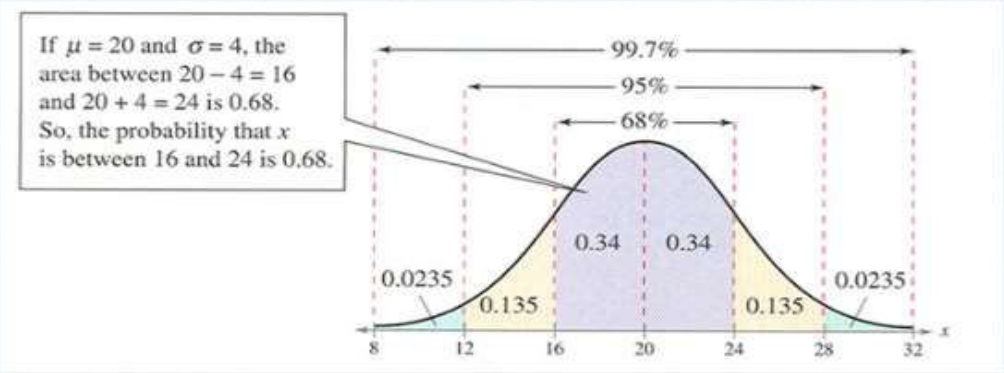
The line of symmetry of curve A occurs at x = 15. The line of symmetry of curve B occurs at x = 12. So, curve A has a greater mean.

Curve B is more spread out than curve A, so curve B has a greater standard deviation.

The Empirical Rule

In a normal distribution with mean μ and standard deviation σ, you can approximate areas under the normal curve as follows:

1. About 68% of the area lies between μ - σ and μ + σ
2. About 95% of the area lies between μ - 2σ and μ + 2σ
3. About 99.7% of the area lies between μ - 3σ and μ + 3σ



Normal Distribution

A random variable X is said to possess normal distribution with mean μ and variance σ^2, if its probability density function can be expressed of the form,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \qquad -\infty < x < +\infty$$

The standard notation used to denote a random variable to follow normal distribution with appropriate mean and variance is, $X \sim N(\mu , \sigma^2)$

STANDARD NORMAL DISTRIBUTION

If a random variable X follows normal distribution with mean μ and variance σ^2 , its transformation $Z = \frac{X-\mu}{\sigma}$ follows standard normal distribution (mean 0 and unit variance)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad , \quad -\infty < z < +\infty$$

The distribution function of the standard normal distribution

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx$$

R has four in built functions to generate normal distribution. They are described below.

dnorm(x, mean, sd)
pnorm(x, mean, sd)
qnorm(p, mean, sd)
rnorm(n, mean, sd)

Following is the description of the parameters used in above functions −

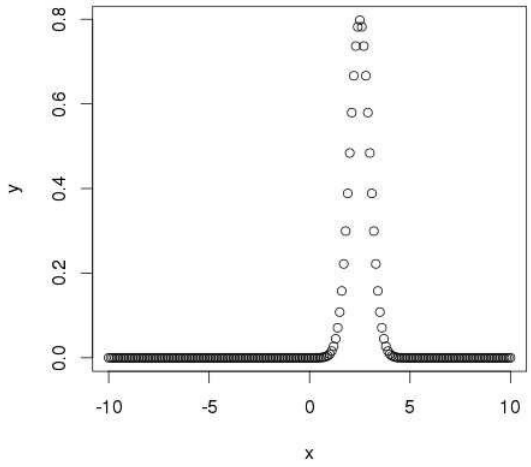
- **x** is a vector of numbers.
- **p** is a vector of probabilities.
- **n** is number of observations (sample size).
- **mean** is the mean value of the sample data. Its default value is zero.
- **sd** is the standard deviation. Its default value is 1.

dnorm()

This function gives height of the probability distribution at each point for a given mean and standard deviation.

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.1.
x <- seq(-10, 10, by = .1)

# Choose the mean as 2.5 and standard deviation as 0.5.
y <- dnorm(x, mean = 2.5, sd = 0.5)

# Give the chart file a name.
png(file = "dnorm.png")
plot(x,y)

# Save the file.
dev.off()
```

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

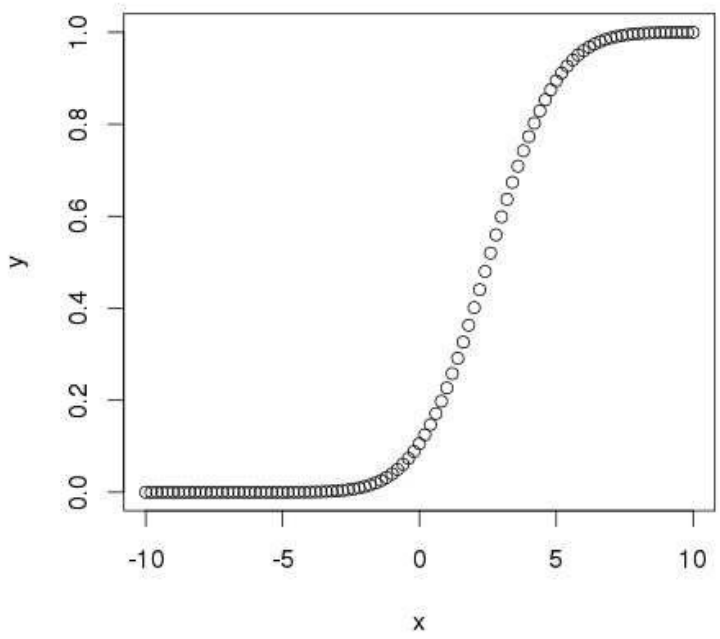


pnorm()

This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.2.
x <- seq(-10,10,by = .2)

# Choose the mean as 2.5 and standard deviation as 2.
y <- pnorm(x, mean = 2.5, sd = 2)

# Give the chart file a name.
png(file = "pnorm.png")

# Plot the graph.
plot(x,y)

# Save the file.
dev.off()
```

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

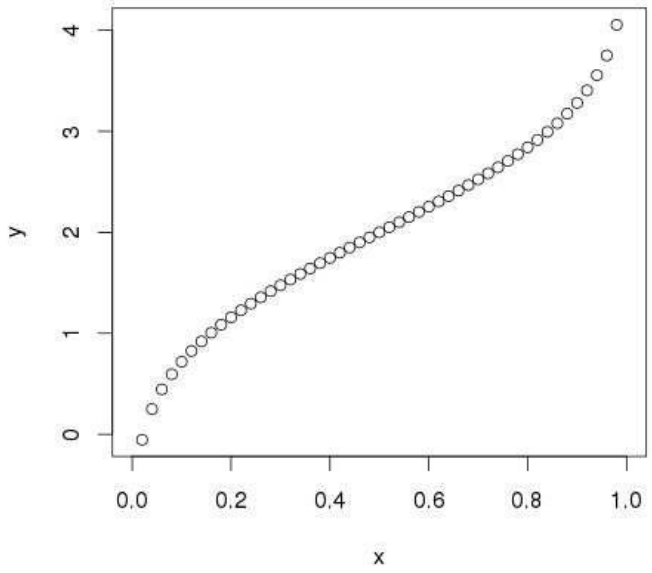


qnorm()

This function takes the probability value and gives a number whose cumulative value matches the probability value.

```
# Create a sequence of probability values incrementing by 0.02.
x <- seq(0, 1, by = 0.02)

# Choose the mean as 2 and standard deviation as 3.
y <- qnorm(x, mean = 2, sd = 1)

# Give the chart file a name.
png(file = "qnorm.png")

# Plot the graph.
plot(x,y)

# Save the file.
dev.off()
```

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

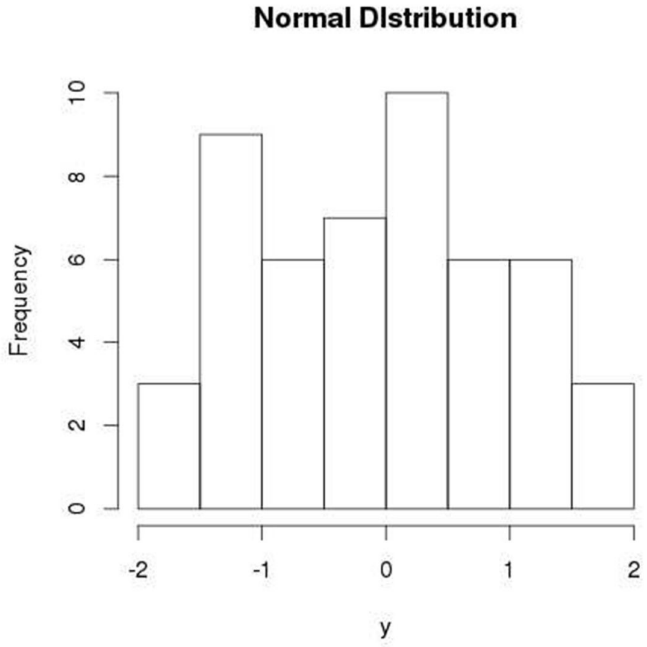


rnorm()

This function is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.

```
# Create a sample of 50 numbers which are normally distributed.
y <- rnorm(50)

# Give the chart file a name.
png(file = "rnorm.png")

# Plot the histogram for this sample.
hist(y, main = "Normal DIstribution")

# Save the file.
dev.off()
```

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

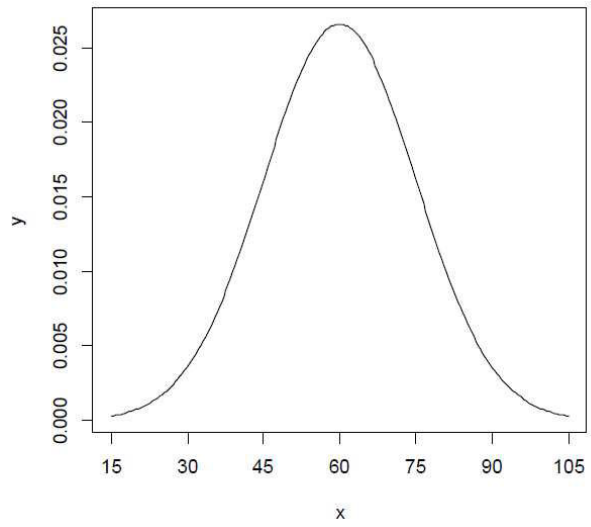


1. The weekly wages of 1000 workmen are normally distributed around a mean of Rs. 70 with S.D of Rs 5.Estimate the number of workers whose weekly wages will be
(i) Between Rs 69 and Rs 72
(ii) Less than Rs 69
(iii) More than Rs 72

```
> #(i)Between Rs 69 and Rs 72
> ( pnorm(72, mean=70, sd=5) - pnorm(69, mean=70, sd=5))*1000
[1] 234.6815
> #The number of workers whose wages lies between Rs.69 and Rs.72 is 234
> #(ii) Less than Rs 69
> (pnorm(69, mean=70, sd=5))*1000
[1] 420.7403
> #The number of workers whose wages is less than Rs.69  is 421
> #(iii) More than Rs 72
> (1 - pnorm(72, mean=70, sd=5))*1000
[1] 344.5783
> #The number of workers whose wages is More than Rs.72  is 345
```

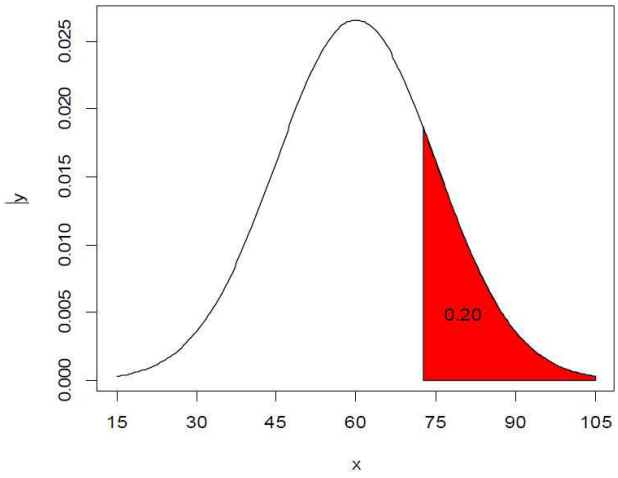
2. Draw a normal distribution with a mean=60 and a standard deviation=15.
>x=seq(15,105,length=200)
>y=dnorm(x,mean=60,sd=15)

```
>plot(x,y,type="l",xaxt="n")
>axis(1,at=c(15,30,45,60,75,90,105))
```



3.Shade the top 20% of the area under the normal density curve

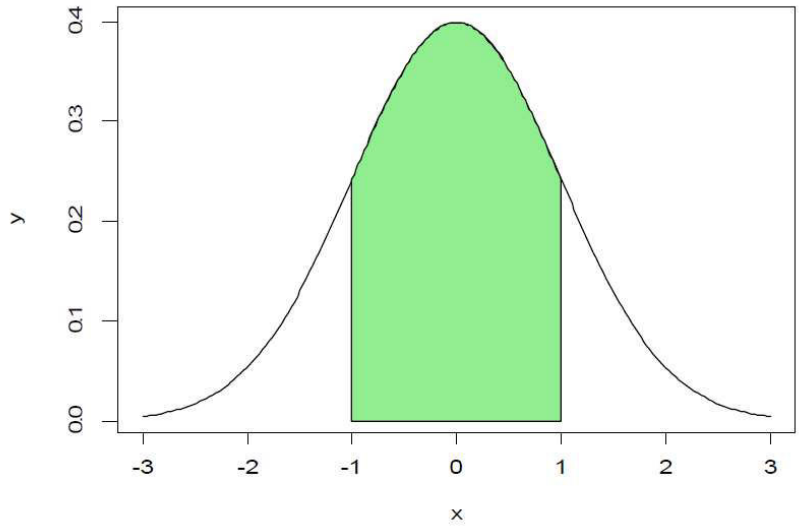
```
>x=seq(15,105,length=200)
>y=dnorm(x,mean=60,sd=15)
>plot(x,y,type="l",xaxt="n")
>axis(1,at=c(15,30,45,60,75,90,105))
>x=seq(72.62,105,length=100)
>y=dnorm(x,mean=60,sd=15)
>polygon(c(72.62,x,105),c(0,y,0),col="red")
>text(80,0.005,"0.20")
```



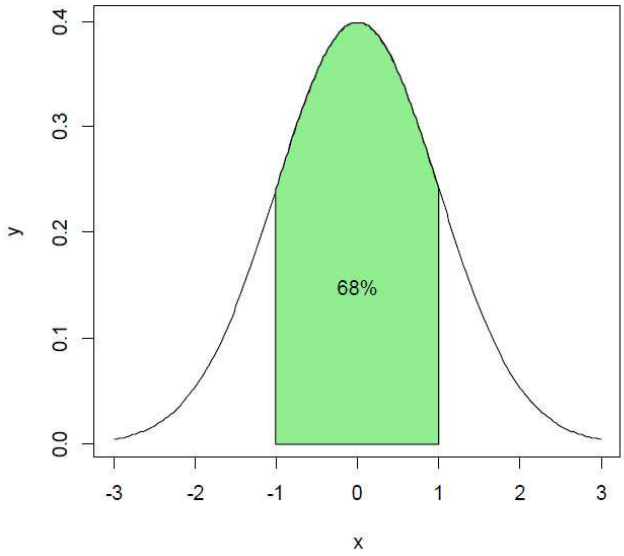
3. Simulate a standard normal density curve (mean=0 and standard deviation=1)

```
>x=seq(-3,3,length=200)
>y=dnorm(x)
```

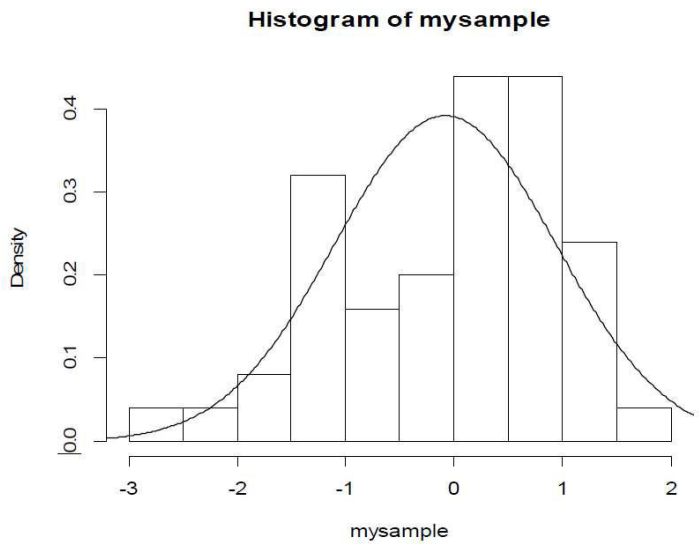
>plot(x,y,type="l")
>x=seq(-1,1,length=100)
>y=dnorm(x)
>polygon(c(-1,x,1),c(0,y,0),col="lightgreen")



>text(0,0.15,"68%")

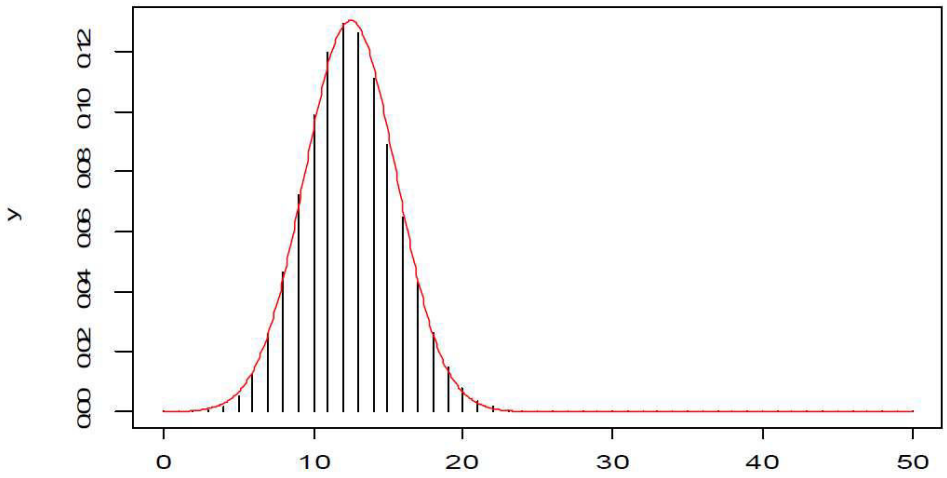


4. Generate 50 (standard) normally distributed random numbers and to display them as a histogram.
>mysample <- rnorm(50)
>hist(mysample, prob = TRUE)
>mu <- mean(mysample)
>sigma <- sd(mysample)
>x <- seq(-4, 4, length = 500)
>y <- dnorm(x, mu, sigma)
>lines(x,y)



5. Approximation of the binomial distribution with the normal distribution

```
> x <- 0:50
> y <- dbinom(x, 50, 0.25)
> plot(x, y, type="h")
> x2 <- seq(0, 50, length = 500)
> y2 <- dnorm(x2, 50*0.25, sqrt(50*0.25*(1-0.25)))
> lines(x2, y2, col = "red")
```



Practice:-

1. Suppose X is normal with mean 527 and
standard deviation 105. Compute P(X <=310)

```
>pnorm(310,527,105)
[1] 0.01938279
```

2. Find P(80 pts< x<95 pts.)

```
>pnorm(95, mean=100, sd=15) - pnorm(80,mean=100, sd=15)
[1] 0.2782301
```

3. In a test on 2000 Electric bulbs ,it was found that the life of particular make, was normally distributed with an average life of 2040 hours and S.D of 60 hours. Estimate the number of bulbs likely to burn for:
(i) More than 2150 hours
(ii) Less than 1950 hours
(iii) More than 1920 hours but less than 2160 hours
(iv) More than 2150 hours

```
> (1 - pnorm(2150, mean=2040, sd=60))*2000
[1] 66.75302
> (pnorm(1950, mean=2040, sd=60))*2000
[1] 133.6144
> ( pnorm(2160, mean=2040, sd=60)-pnorm(1920,mean=2040,sd=60))*2000
[1] 1908.999
```

(i) The number of bulbs expected to burn for more than 2150 hours is 67 (approximately)
(ii) The number of bulbs expected to burn for less than 1950 hours is 134 (approximately)
(iii) The number of bulbs expected to burn more than 1920 hours but less than 2160 is 1909 (approximately

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, volume 1, chapter 13. Wiley, New York.

Sampling Techniques

Large and Small Sample Test

Large Sample Test ($n \geq 30$)
Z- test (One Sample)

Test for single Mean - Lower Tail Test of Population Mean with Known Variance

The null hypothesis of the lower tail test of the population mean can be expressed as follows:

$$\mu \geq \mu_0$$

Where μ_0 is a hypothesized lower bound of the true population mean μ

Z- test cont…

Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ :

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Then the null hypothesis of the lower tail test is to be rejected if $z < -z_{\alpha}$, where z_{α}, is the $100(1-\alpha)$ percentile of the standard normal distribution.

Problems on Single Mean Test

- ***Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. In a sample of 30 light bulbs, it was found that they only last 9,900 hours on average. Assume the population standard deviation is 120 hours. At .05 significance level, can we reject the claim by the manufacturer?***

The null hypothesis is that $\mu \geq 10000$

```
> xbar = 9900                  # sample mean
> mu0 = 10000                  # hypothesized value
> sigma = 120                  # population standard deviation
> n = 30                       # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                            # test statistic
[1] -4.564355
```

Critical Value

We then compute the critical value at .05 significance level.

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> -z.alpha                    # critical value
[1] -1.644854
```

Interpretation

The test statistic -4.5644 is less than the critical value of -1.6449. Hence, at .05 significance level, we reject the claim that mean lifetime of a light bulb is above 10,000 hours.

Alterative Comparison P- Value

- The p-value is the level of marginal significance within a statistical hypothesis test representing the probability of the occurrence of a given event. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.

P- Value Cont….

- The p-value approach to hypothesis testing uses the calculated probability to determine whether there is evidence to reject the null hypothesis. The null hypothesis, also known as the conjecture, is the initial claim about a population of statistics. The alternative hypothesis states whether the population parameter differs from the value of the population parameter stated in the conjecture. In practice, the p-value, or critical value, is stated in advance to determine how the required value to reject the null hypothesis.

P –Value Comparison

- ***Instead of using the critical value, we apply the pnorm function to compute the lower tail p-value of the test statistic. As it turns out to be less than the .05 significance level, we reject the null hypothesis that μ ≥ 10000.***

```
> pval = pnorm(z)
> pval                    # lower tail p-value
[1] 2.505166e-06
```

Upper Tail Test of Population Mean with Known Variance:

- ***The null hypothesis of the upper tail test of the population mean can be expressed as follows***

$$\mu \leq \mu_0$$

where μ_0 is a hypothesized upper bound of the true population mean μ.

- ***Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ :***

Upper Tailed Test cont…

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Then the null hypothesis of the upper tail test is to be rejected if $z \geq z_\alpha$ ***where*** z_α ***is the 100(1− α) percentile of the standard normal distribution***

Problem

- ***Suppose the food label on a cookie bag states that there is at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that the mean amount of saturated fat per cookie is 2.1 grams. Assume that the population standard deviation is 0.25 grams. At .05 significance level, can we reject the claim on food label?***

R - Code

- ***The null hypothesis is that $\mu \leq 2$. We begin with computing the test statistic.***

```
> xbar = 2.1                    # sample mean
> mu0 = 2                       # hypothesized value
> sigma = 0.25                  # population standard deviation
> n = 35                        # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                             # test statistic
[1] 2.366432
```

Critical Value Comparison

```
> alpha = .05
> z.alpha = qnorm(1-alpha)
> z.alpha                    # critical value
[1] 1.644854
```

Interpretation

- ***The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at .05 significance level, we reject the claim that there is at most 2 grams of saturated fat in a cookie.***

Two-Tailed Test of Population Mean with Known Variance

- ***The null hypothesis of the two-tailed test of the population mean can be expressed as follows:***

$$\mu = \mu_0$$

where μ_0 is a hypothesized value of the true population mean . Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ :

$$z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

- ***Then the null hypothesis of the two-tailed test is to be rejected if*** $z \leq -z_{\alpha/2}$ ***or*** $z \geq z_{\alpha/2}$ ***, where*** $z_{\alpha/2}$ ***is the*** $100(1 - \alpha/2)$ ***percentile of the standard normal distribution.***

Problem

- ***Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation is 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?***

R – Code

- **The null hypothesis is that $\mu = 15.4$.**

```
> xbar = 14.6                    # sample mean
> mu0 = 15.4                     # hypothesized value
> sigma = 2.5                    # population standard deviation
> n = 35                         # sample size
> z = (xbar-mu0)/(sigma/sqrt(n))
> z                              # test statistic
[1] -1.893146
```

Comparison using p- value

```
> pval = 2 * pnorm(z)     # lower tail
> pval                    # two-tailed p-value
[1] 0.05833852
```

The p – value turns out to be greater than the .05 significance level, we do not reject the null hypothesis that μ = 15.4.

Lower Tail Test of Population Proportion

- ***The null hypothesis of the lower tail test about population proportion can be expressed as follows:***

$$P \geq P_0$$

where p_0 is a hypothesized lower bound of the true population proportion p. Let us define the test statistic z in terms of the sample proportion and the sample size:

$$z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$$

- ***Then the null hypothesis of the lower tail test is to be rejected if $z < -z_{\alpha}$, where z_{α} is the 100(1− α) percentile of the standard normal distribution.***

Problem

Suppose 60% of citizens voted in last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?

R – Code

The null hypothesis is that $p \geq 0.6$

```
> pbar = 85/148                  # sample proportion
> p0 = .6                        # hypothesized value
> n = 148                        # sample size
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z                              # test statistic
[1] -0.6375983
```

P-value comparison

```
> pval = pnorm(z)
> pval
[1] 0.2618676
```

As p-value turns out to be greater than the .05 significance level, we do not reject the null hypothesis that p ≥ 0.6.

Upper Tail Test of Population Proportion

- The null hypothesis of the upper tail test about population proportion can be expressed as follows:

$$P \leq P_0$$

where p_0 is a hypothesized upper bound of the true population proportion p. Let us define the test statistic z in terms of the sample proportion and the sample size:

$$z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$$

Problem

- ***Suppose that 12% of apples harvested in an orchard last year was rotten. 30 out of 214 apples in a harvest sample this year turns out to be rotten. At .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?***

R - Code

- **The null hypothesis is that $p \leq 0.12$**

```
> pbar = 30/214                  # sample proportion
> p0 = .12                       # hypothesized value
> n = 214                        # sample size
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z                              # test statistic
[1] 0.908751
```

P- Value comparison

- ***The p- value turns out to be greater than the .05 significance level, we do not reject the null hypothesis that $p \leq 0.12$.***

```
> pval = pnorm(z, lower.tail=FALSE)
> pval                          # upper tail p?value
[1] 0.1817408
```

Two-Tailed Test of Population Proportion

- The null hypothesis of the two-tailed test about population proportion can be expressed as follows:

$$P = P_0$$

- where p_0 is a hypothesized value of the true population proportion p. Let us define the test statistic z in terms of the sample proportion and the sample size:

$$z = \frac{p - p_0}{\sqrt{p_0 q_0 / n}} \sim N(0,1)$$

- Then the null hypothesis of the two-tailed test is to be *rejected* if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$, where $z_{\alpha/2}$ is the 100(1 − α) percentile of the standard normal distribution.

Problem

- ***Suppose a coin toss turns up 12 heads out of 20 trials. At .05 significance level, can one reject the null hypothesis that the coin toss is fair?***

The null hypothesis is that p = 0.5

```
> pbar = 12/20                    # sample proportion
> p0 = .5                         # hypothesized value
> n = 20                          # sample size
> z = (pbar-p0)/sqrt(p0*(1-p0)/n)
> z                               # test statistic
[1] 0.8944272
```

P- value comparison

```
> pval = 2 * pnorm(z, lower.tail=FALSE)  # upper tail
> pval
[1] 0.3710934
```

P- value turns out to be greater than the .05 significance level, we do not reject the null hypothesis that p = 0.5.

Alternate Calculation for One Sample Proportion Test

- ***To apply the prop.test function to compute the p-value directly***
- ***Syntax:***
- prop.test(x,n,p, alt =c(“greater”, “lesser”, “two.sided”),conf.level=.95, correct = FALSE)

Problem 1

Suppose that 12% of apples harvested in an orchard last year was rotten. 30 out of 214 apples in a harvest sample this year turns out to be rotten. At .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?

```
> prop.test(30, 214, p=.12, alt="greater", correct=FALSE)

        1-sample proportions test without continuity correction

data:  30 out of 214, null probability 0.12
X-squared = 0.82583, df = 1, p-value = 0.1817
alternative hypothesis: true p is greater than 0.12
95 percent confidence interval:
 0.1056274 1.0000000
sample estimates:
        p
0.1401869
```

Problem 2

Suppose a coin toss turns up 12 heads out of 20 trials. At .05 significance level, can one reject the null hypothesis that the coin toss is fair?

```
> prop.test(12, 20, p=0.5, correct=FALSE)

        1-sample proportions test without continuity correction

data:  12 out of 20, null probability 0.5
X-squared = 0.8, df = 1, p-value = 0.3711
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3865815 0.7811935
sample estimates:
  p
0.6
```

Large Sample Test

Z - Test for Two Samples

Two Proportion Test

R-Code:-

Tests about a proportion using x and n using the prop.test function:

Usage: prop.test(c(x1,x2), c(n1,n2), correct=, alternate =).

1. x1 and x2 are the number of successes in sample 1 and 2 respectively.
2. n1 and n2 are the sample sizes or number of trials.
3. correct = TRUE (use a continuity correction factor) or FALSE (do not).
4. alternate = "two.sided" (default), "less", or "greater".

Problem

A popular cold-remedy was tested for it's efficacy. In a sample of 150 people who took the remedy upon getting a cold, 117 (78%) had no symptoms one week later. In a sample of 125 people who took the placebo upon getting a cold, 90 (75%) had no symptoms one week later. The table summarizes this information. Test the claim that the proportion of all remedy users who are symptom-free after one week is greater than the proportion for placebo users. Test this claim at the 0.05 significance level.

Group	#who are symptom Free after one week(x)	Total # in group (n)	Proportion $\hat{p} = x / n$
Remedy	117	150	0.78
Placebo	90	120	0.75

R Code:-

```
> x<-c(117,90)
> n<-c(150,120)
> prop.test(x,n,alternative="greater",correct=FALSE)

        2-sample test for equality of proportions without continu
        correction

data:  x out of n
X-squared = 0.3354, df = 1, p-value = 0.2812
alternative hypothesis: greater
95 percent confidence interval:
 -0.05557192  1.00000000
sample estimates:
prop 1 prop 2
  0.78   0.75
```

We fail reject the null hypothesis because the P-value (.2812) is greater than the significance level. Therefore, we can't support the claim.

Problem 2

- *The Trial Urban District Assessment (TUDA) is a study sponsored by the government of student achievement in large urban school district. In 2009, 1311 of a random sample of 1900 eighth-graders from Houston performed at or above the basic level in mathematics . In 2011, 1440 of a random sample of 2000 eighth-graders from Houston performed at or above the basic level . (The study reports the proportions).*

(A)Is there an increase in the proportion of eighth-graders who performed at or above the basic level in mathematics from 2009 to 2011 at the 5% significance level?

Compute the 95% confidence interval for the difference in proportion of eighth-graders who performed at or above the basic level in mathematics from 2009 to 2011.

Let p1 and p2 be the proportions of eighth-graders that performed at or above he basic level in mathematics in 2011 and 2009,respectively .

H_0 : *p1=p2 against* H_1 = *p1> p2*

```
> prop.test(c(1440,1311),c(2000,1900),alternative="greater",correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(1440, 1311) out of c(2000, 1900)
X-squared = 4.2197, df = 1, p-value = 0.01998
alternative hypothesis: greater
95 percent confidence interval:
 0.005972807 1.000000000
sample estimates:
prop 1 prop 2
  0.72   0.69
```

- The p-value=0.02 < 0.05 so we reject H0. Thus, there is evidence that there is an increase from 2009 to 2011 in the proportion of eighth-graders who performed at or above the basic level at the 5% significance level.

Solution to part (b)

```
> prop.test(c(1440,1311),c(2000,1900),correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(1440, 1311) out of c(2000, 1900)
X-squared = 4.2197, df = 1, p-value = 0.03996
alternative hypothesis: two.sided
95 percent confidence interval:
 0.001369833 0.058630167
sample estimates:
prop 1 prop 2
  0.72   0.69
```

Thus, we are 95% confident that the percent of eighth-graders who performed at or above the basic level in mathematics in 2011 is between 0:14% and 5:86% higher than in 2009.

Problem 3

- *The use of helmet among recreational alpine skiers and snowboarders are generally low. A study from Norway wanted to examine if helmet use reduces the risk of head injury. In the study, they compared the helmet use among skiers and snowboarders that was injured with a control group. The control group consisted of skiers and snowboarders that was uninjured. 96 of 578 people with head injuries used a helmet and 656 of 2992 people in the uninjured group used a helmet. Is helmet use lower among skiers and snowboarders who had head injuries?*
-

Let p1 be the proportion of helmet use among injured skiers and snowboarders.
Let p2 be the proportion of helmet use among uninjured skiers and snowboarders
H0 : p1 = p2 against H1 : p1 < p2

```
> prop.test(c(96,656),c(578,2992),alternative=c("less"),correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(96, 656) out of c(578, 2992)
X-squared = 8.2336, df = 1, p-value = 0.002056
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000 -0.02482216
sample estimates:
   prop 1    prop 2
0.1660900 0.2192513
```

The p-value= 0.0021 < 0.01 so we have strong evidence that helmet use is lower among skiers and snowboarders who had head injuries compared to uninjured skiers and snowboarders.

Problem 4

- *A survey is taken two times over the course of two weeks. The pollsters wish to see if there is a difference in the results as there has been a new advertising campaign run. Here is the data*

	Week1	Week2
Favorable	45	56
Unfavorable	35	47

H0: P1 = P2

H1: P1 =P2 (two- sided)

R - Code

```
> prop.test(c(45,56),c(45+35,56+47))

        2-sample test for equality of proportions with continuity correction

data:  c(45, 56) out of c(45 + 35, 56 + 47)
X-squared = 0.010813, df = 1, p-value = 0.9172
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1374478  0.1750692
sample estimates:
   prop 1    prop 2
0.5625000 0.5436893
```

we observe that the p-value is 0.9172 so we accept the null hypothesis that P1 =P2.

Two mean Test

The following data shows the heights of individuals of two different countries with the population variance of 5 and 8.5 respectively. Is there any significant difference between the average heights of two groups.

A: 175	168	168	190	156	181	182	175	174	179
B: 185	169	173	173	188	186	175	174	179	180

R – Code

```
> a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)
> b = c(185, 169, 173, 173, 188, 186, 175, 174, 179, 180)
> n1=length(a)
> n2=length(b)
> zeta = abs((mean(a) - mean(b)) / (sqrt(var(a)/n1 + var(b)/n2)))
> zeta
[1] 0.947373
```

P- value comparison

```
> pvalue=2*pnorm(zeta)
> pvalue
[1] 1.656551
```

Since it turns out to be greater than the .05 significance level, we do not reject the null hypothesis

Practice Problems on Large Samples

- In the sample of 1000 people in Maharashtra,540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance

- A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the average lifespan was 46,500 miles with a standard deviation of 9800 miles. Do the data support the claim at the 5% level?

Practice problems cont….

- In the large city A,20 per cent of Random sample of 900 School children had defective eye –sight. In the large city B,15 percent of random sample of 1600 school children had the same defective. Is this Difference between the two Proportions Significant? Obtain 95% confidence limits of the difference in the population proportions.
- A cigarette manufacturing firm claims its brand A of the cigarettes outsells its brand B by 8%.if its found that 42 out sample of 200 smoker prefer brand A and 18 out of another random sample of 100 smokers prefers brand B, test whether the 8% difference is a valid cliam.

Practice problems cont….

- **The average number of sick days an employee takes per year is believed to be about 10. Members of a personnel department do not believe this figure. They randomly survey 8 employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let X = the number of sick days they took for the past year. Should the personnel team believe that the average number is about 10?**
- **In 1955, *Life Magazine* reported that the 25 year-old mother of three worked [on average] an 80 hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the average work week has increased. 81 women were surveyed with the following results. The sample average was 83; the sample standard deviation was 10. Does it appear that the average work week has increased for women at the 5% level?**

Practice problems cont….

- **A sample of 100 tyres is taken from a lot. The mean life of tyres is found to be 39, 350 kilo meters with a standard deviation of 3, 260. Could the sample come from a population with mean life of 40, 000 kilometers?**
- **The mean life time of a sample of 400 fluorescent light bulbs produced by a company is found to be 1, 570 hours with a standard deviation of 150 hours. Test the hypothesis that the mean life time of bulbs is 1600 hours against the alternative hypothesis that it is greater than 1, 600 hours at 1% and 5% level of significance**

Small Sample Test

size less than thirty

Small Sample Test

t-test for single mean and t-test for difference of means

- The t.test( ) function produces a variety of t-tests. Unlike most statistical packages, the default assumes unequal variance.

Syntax: one sample – single mean

- t.test(y, mu=3, alt = "greater“/ “lesser”, var.equal = TRUE/FALSE)

Ho: mu=3

Problem

- An outbreak of salmonella-related illness was attributed to ice produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches ice crean.The levels(in MPN/g) were:

0.593	0.142	0.329	0.691	0.231	0.793	0.519	0.392	0.418

Is there evidence that the mean level pf Salmonella in ice cream greater than 0.3 MPN/g?

R- Code & Interpretation

```
> x=c(0.593,0.142,0.329,0.691,0.231,0.793,0.519,0.392,0.418)
> t.test(x,alternative="greater",mu=0.3)

        One Sample t-test

data:  x
t = 2.2051, df = 8, p-value = 0.02927
alternative hypothesis: true mean is greater than 0.3
95 percent confidence interval:
 0.3245133       Inf
sample estimates:
mean of x
0.4564444
```

From the output we see that the p-value = 0.029. Hence, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3MPN/g.

Problem

Suppose that 10 volunteers have taken an intelligence test; here are the results obtained. The average score of the entire population is 75 in the same test. Is there any significant difference (with a significance level of 95%) between the sample and population means, assuming that the variance of the population is not known

.

Scores: 65, 78, 88, 55, 48, 95, 66, 57, 79, 81

R- Code & Interpretation

```
> a = c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)
> t.test (a, mu=75)

        One Sample t-test

data:  a
t = -0.78303, df = 9, p-value = 0.4537
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 60.22187 82.17813
sample estimates:
mean of x
     71.2
```

the p-value with a significance level of 95%. If ***p-value is lesser than 0.05*** *hence we* ***reject the null hypothesis***

T- test for two samples (independent)

- Two-Tailed Test: t.test(x, y, mu = ,)
- Right-Tailed Test: t.test(x, y, mu= , alternative="greater")
- Left-Tailed Test: t.test(x, y, mu= ,alternative="less")

Problem

Comparing two independent sample means, taken from two populations with unknown variances. The following data shows the heights of individuals of two different countries with unknown population variances. Is there any significant difference between the average heights of two groups.

A:	175	168	168	190	156	181	182	175	174	179
B:	185	169	173	173	188	186	175	174	179	180

R- Code & Interpretation

```
> a = c(175, 168, 168, 190, 156, 181, 182, 175, 174, 179)
> b = c(120, 180, 125, 188, 130, 190, 110, 185, 112, 188)
> t.test(a,b, var.equal=FALSE, paired=FALSE)

        Welch Two Sample t-test

data:  a and b
t = 1.8827, df = 10.224, p-value = 0.08848
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.95955 47.95955
sample estimates:
mean of x mean of y
    174.8     152.8
```

The p-value > 0.05, we conclude that the means of the two groups are significantly similar

Problem

- Suppose the recovery time for patients taking a new drug is measured (in days). A placebo group is also used to avoid the placebo effect. The data are as follows

with drug	: 15 10 13 7 9 8 21 9 14 8
placebo	: 15 14 12 8 14 7 16 10 15 2

Is there any significant difference between the average effect of these two drugs?

R- Code & Interpretation

```
> x = c(15, 10, 13, 7, 9, 8, 21, 9, 14, 8)
> y = c(15, 14, 12, 8, 14, 7, 16, 10, 15, 12)
> t.test(x,y,alt="less",var.equal=TRUE)

        Two Sample t-test

data:  x and y
t = -0.53311, df = 18, p-value = 0.3002
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.027436
sample estimates:
mean of x mean of y
     11.4      12.3
```

P value (0.3002) > 0.05 then there is no evidence to reject our Null hypothesis

Problem

- Six subjects were given a drug (trearment group) and an additional 6 subjects a placebo(control group).Their reaction time to stimulus was measured(in ms).We want to perform a two sample t-test for comparing the means of the treatment and control groups

Control	91	87	99	77	88	91
Treatment	101	110	103	93	99	104

R- Code and Inference

```
> 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-value = 0.003136
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -6.082744
sample estimates:
mean of x mean of y
 88.83333 101.66667
```

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

        Welch Two Sample t-test

data:  control and Treat
t = -3.4456, df = 9.4797, p-value = 0.003391
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -6.044949
sample estimates:
mean of x mean of y
 88.83333 101.66667
```

From both the output we see that the p-value = 0.003136(equal) and 0.003391(Unequal). Therefore, it infers that there is different between treatment and control group.

Paired t-test (Dependent Sample)

```
# paired t-test

> t.test(y1,y2,paired=TRUE)        # where y1 & y2 are numeric
```

Problem

- *A school athletics has taken a new instructor, and want to test the effectiveness of the new type of training proposed by the new instructor comparing the average times of 10 runners in the 100 meters. The results are given below(time in seconds)*

Before training	*12.9*	*13.5*	*12.8*	*15.6*	*17.2*	*19.2*	*12.6*	*15.3*	*14.4*	*11.3*
After training	*12.7*	*13.6*	*12.0*	*15.2*	*16.8*	*20.0*	*12.0*	*15.9*	*16.0*	*11.1*

- *Solu:*
- *In this case we have two sets of paired samples, since the measurements were made on the same athletes before and after the workout. To see if there was an improvement, deterioration, or if the means of times have remained substantially the same (hypothesis H0), we need to make a Student's t-test for paired samples, proceeding in this way*

R- Code & Inference

```
> before = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
> after = c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)
> t.test(before,after, paired=TRUE)

        Paired t-test

data:  before and after
t = -0.21331, df = 9, p-value = 0.8358
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5802549  0.4802549
sample estimates:
mean of the differences
                 -0.05
```

Interpretation :-

The p-value is greater than 0.05, then we do not reject the hypothesis H_0 of equality of the averages and conclude that the new training has not made any significant improvement to the team of athletes.

Problem

Suppose now that the manager of the team (given the results obtained) fired the coach who has not made any improvement, and take another, more promising. We report the times of athletes after the second training:

Before training:	*12.9*	*13.5*	*12.8*	*15.6*	*17.2*	*19.2*	*12.6*	*15.3*	*14.4*	*11.3*
After the second training:	*12.0*	*12.2*	*11.2*	*13.0*	*15.0*	*15.8*	*12.2*	*13.4*	*12.9*	*11.0*

Solu:

Now we check if there was actually an improvement, ie perform a t-test for paired data, specifying in R to test the alternative hypothesis H1 of improvement in times. To do this simply add the syntax alt = "less" when you call the t-test

R- Code & Inference

```
> before=c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
> after = c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)
> t.test(before,after, paired=TRUE, alt="less")

        Paired t-test

data:  before and after
t = 5.2671, df = 9, p-value = 0.9997
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.170325
sample estimates:
mean of the differences
                  1.61
```

In response, we obtained a p-value well above 0.05, which leads us to conclude that we can reject the null hypothesis H_0 in favour of the alternative hypothesis H1: the new training has made substantial improvements to the team

Problem

- *Consider the paired data below that represents cholesterol levels on 10 men before and after a certain medication Test the claim that, on average, the drug lowers cholesterol in all men. i.e., test the claim that* $\mu_d > 0$. *Test this at the 0.05 significance level*

Before(x)	*237*	*289*	*257*	*228*	*303*	*275*	*262*	*304*	*244*	*233*
After(y)	*194*	*240*	*230*	*186*	*265*	*222*	*242*	*281*	*240*	*212*

R- Code and Interpretation

```
> before=c(237,289,257,228,303,275,262,304,244,233)
> after=c(194,240,230,186,265,222,242,281,240,212)
> t.test(before,after,paired=TRUE,alternative="greater",mu=0)

        Paired t-test

data:  before and after
t = 6.5594, df = 9, p-value = 5.202e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 23.05711      Inf
sample estimates:
mean of the differences
                    32
```

We can reject the null hypothesis and support the claim because the P-value ($5.2\ X10^{5}$) is less than the significance level

F- Test (Variance Ration Test)

- Syntax:

```
var.test(x, y)
```

Problem

- *Five Measurements of the output of two units have given the following results (in kilograms of material per one hour of operation) .Assume that both samples have been obtained from normal populations, test at 10% significance level if two populations have the same variance*

Unit A	*14.1*	*10.1*	*14.7*	*13.7*	*14.0*
Unit B	*14.0*	*14.5*	*13.7*	*12.7*	*14.1*

$H_0: S_1^2 = S_2^2$

$H_1: S_1^2 \neq S_2^2$

R- Code and Inference

```
> Unit_A=c(14.1,10.1,14.7,13.7,14.0)
> Unit_B=c(14.0,14.5,13.7,12.7,14.1)
> var.test(Unit_A,Unit_B)

        F test to compare two variances

data:  Unit_A and Unit_B
F = 7.3304, num df = 4, denom df = 4, p-value = 0.07954
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.7632268 70.4053799
sample estimates:
ratio of variances
          7.330435
```

Here p valuve >0.05 ,then there is no evidence to reject the null hypothesis

Practice Problems

- A certain stimulus administered to each of the 13 patients resulted in the following increase of blood pressure: 5, 2, 8,-1, 3, 0, -2, 1, 5, 0, 4, 6, 8. Can it be concluded that the stimulus, in general, be accompanied by an increase in the blood pressure.
- The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with a standard deviation of 5 months. Random samples of 6 such bulbs have the following values: Life of bulbs in months: 24, 20, 30, 20, 20, and 18. Can you regard the producer's claim to valid at 1% level of significance

Practice Problems cont…

- The life time of electric bulbs for a random sample of 10 from a large consignment gave the following data: 4.2, 4.6, 3.9, 4.1, 5.2, 3.8, 3.9, 4.3, 4.4, 5.6 (in '000 hours). Can we accept the hypothesis that the average life time of bulbs is 4, 000 hours
- Data on weight (grams) of two treatments of NMU (nistroso- methyl urea) are recorded. Find out whether these two treatments have identical effects by using t test for sample means at 5% level of significance.

Sample	1	2	3	4	5	6	7	8	9	10	11	12
Treatments 0.2 %	2.0	2.7	2.9	1.9	2.1	2.6	2.7	2.9	3.0	2.6	2.6	2.7
0.4%	3.2	3.6	3.7	3.5	2.9	2.6	2.5	2.7				

Small Sample Test

size less than thirty

Small Sample Test

t-test for single mean and t-test for difference of means

- The t.test( ) function produces a variety of t-tests. Unlike most statistical packages, the default assumes unequal variance.

Syntax: one sample – single mean

- t.test(y, mu=3, alt = "greater“/ “lesser”)

Problem

- An outbreak of salmonella-related illness was attributed to ice produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches ice crean.The levels(in MPN/g) were:

0.593	0.142	0.329	0.691	0.231	0.793	0.519	0.392	0.418

Is there evidence that the mean level pf Salmonella in ice cream greater than 0.3 MPN/g?

R- Code & Interpretation

```
> x=c(0.593,0.142,0.329,0.691,0.231,0.793,0.519,0.392,0.418)
> t.test(x,alternative="greater",mu=0.3)

        One Sample t-test

data:  x
t = 2.2051, df = 8, p-value = 0.02927
alternative hypothesis: true mean is greater than 0.3
95 percent confidence interval:
 0.3245133       Inf
sample estimates:
mean of x
0.4564444
```

From the output we see that the p-value = 0.029. Hence, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3MPN/g.

Problem

Suppose that 10 volunteers have taken an intelligence test; here are the results obtained. The average score of the entire population is 75 in the same test. Is there any significant difference (with a significance level of 95%) between the sample and population means, assuming that the variance of the population is not known

.

Scores: 65, 78, 88, 55, 48, 95, 66, 57, 79, 81

R- Code & Interpretation

```
> a = c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)
> t.test (a, mu=75)

        One Sample t-test

data:  a
t = -0.78303, df = 9, p-value = 0.4537
alternative hypothesis: true mean is not equal to 75
95 percent confidence interval:
 60.22187 82.17813
sample estimates:
mean of x
     71.2
```

the p-value with a significance level of 95%. If ***p-value is lesser than 0.05*** *hence we* ***reject the null hypothesis***

T- test for two samples (independent)

- Two-Tailed Test: t.test(x, y, mu = ,)
- Right-Tailed Test: t.test(x, y, mu= , alternative="greater")
- Left-Tailed Test: t.test(x, y, mu= ,alternative="less")

Problem

Comparing two independent sample means, taken from two populations with unknown variances. The following data shows the heights of individuals of two different countries with unknown population variances. Is there any significant difference between the average heights of two groups.

A:	175	168	168	190	156	181	182	175	174	179
B:	185	169	173	173	188	186	175	174	179	180

R- Code & Interpretation

```
> x=c(175,168,168,190,156,181,182,175,174,179)
> y=c(120,180,125,188,130,190,110,185,112,188)
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = 1.8827, df = 10.224, p-value = 0.08848
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.95955 47.95955
sample estimates:
mean of x mean of y
    174.8     152.8
```

The p-value > 0.05, we conclude that the means of the two groups are significantly similar

Problem

- Suppose the recovery time for patients taking a new drug is measured (in days). A placebo group is also used to avoid the placebo effect. The data are as follows

with drug	: 15 10 13 7 9 8 21 9 14 8
placebo	: 15 14 12 8 14 7 16 10 15 2

Is there any significant difference between the average effect of these two drugs?

R- Code & Interpretation

```
> x=c(15,10,13,7,9,8,21,9,14,8)
> y=c(15,14,12,8,14,7,16,10,15,12)
> t.test(x,y,alt="less")

        Welch Two Sample t-test

data:  x and y
t = -0.53311, df = 16.245, p-value = 0.3006
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.044664
sample estimates:
mean of x mean of y
     11.4      12.3
```

P value (0.3002) > 0.05 then there is no evidence to reject our Null hypothesis

Paired t-test (Dependent Sample)

```
# paired t-test

> t.test(y1,y2,paired=TRUE)        # where y1 & y2 are numeric
```

Problem

- *A school athletics has taken a new instructor, and want to test the effectiveness of the new type of training proposed by the new instructor comparing the average times of 10 runners in the 100 meters. The results are given below(time in seconds)*

Before training	*12.9*	*13.5*	*12.8*	*15.6*	*17.2*	*19.2*	*12.6*	*15.3*	*14.4*	*11.3*
After training	*12.7*	*13.6*	*12.0*	*15.2*	*16.8*	*20.0*	*12.0*	*15.9*	*16.0*	*11.1*

- *Solu:*
- *In this case we have two sets of paired samples, since the measurements were made on the same athletes before and after the workout. To see if there was an improvement, deterioration, or if the means of times have remained substantially the same (hypothesis H0), we need to make a Student's t-test for paired samples, proceeding in this way*

R- Code & Inference

```
> before = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
> after = c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)
> t.test(before,after, paired=TRUE)

        Paired t-test

data:  before and after
t = -0.21331, df = 9, p-value = 0.8358
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5802549  0.4802549
sample estimates:
mean of the differences
                 -0.05
```

Interpretation :-

The p-value is greater than 0.05, then we do not reject the hypothesis H_0 of equality of the averages and conclude that the new training has not made any significant improvement to the team of athletes.

Problem

Suppose now that the manager of the team (given the results obtained) fired the coach who has not made any improvement, and take another, more promising. We report the times of athletes after the second training:

Before training:	*12.9*	*13.5*	*12.8*	*15.6*	*17.2*	*19.2*	*12.6*	*15.3*	*14.4*	*11.3*
After the second training:	*12.0*	*12.2*	*11.2*	*13.0*	*15.0*	*15.8*	*12.2*	*13.4*	*12.9*	*11.0*

Solu:

Now we check if there was actually an improvement, ie perform a t-test for paired data, specifying in R to test the alternative hypothesis H1 of improvement in times. To do this simply add the syntax alt = "less" when you call the t-test

R- Code & Inference

```
> before=c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
> after = c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)
> t.test(before,after, paired=TRUE, alt="less")

        Paired t-test

data:  before and after
t = 5.2671, df = 9, p-value = 0.9997
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.170325
sample estimates:
mean of the differences
                  1.61
```

In response, we obtained a p-value well above 0.05, which leads us to conclude that we can reject the null hypothesis H_0 in favour of the alternative hypothesis H1: the new training has made substantial improvements to the team

Problem

- *Consider the paired data below that represents cholesterol levels on 10 men before and after a certain medication Test the claim that, on average, the drug lowers cholesterol in all men. i.e., test the claim that* $\mu_d > 0$. *Test this at the 0.05 significance level*

Before(x)	*237*	*289*	*257*	*228*	*303*	*275*	*262*	*304*	*244*	*233*
After(y)	*194*	*240*	*230*	*186*	*265*	*222*	*242*	*281*	*240*	*212*

R- Code and Interpretation

```
> before=c(237,289,257,228,303,275,262,304,244,233)
> after=c(194,240,230,186,265,222,242,281,240,212)
> t.test(before,after,paired=TRUE,alternative="greater",mu=0)

        Paired t-test

data:  before and after
t = 6.5594, df = 9, p-value = 5.202e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 23.05711      Inf
sample estimates:
mean of the differences
                    32
```

We can reject the null hypothesis and support the claim because the P-value ($5.2\ X10^{5}$) is less than the significance level

F- Test (Variance Ration Test)

- Syntax:

```
var.test(x, y)
```

Problem

- *Five Measurements of the output of two units have given the following results (in kilograms of material per one hour of operation) .Assume that both samples have been obtained from normal populations, test at 10% significance level if two populations have the same variance*

Unit A	*14.1*	*10.1*	*14.7*	*13.7*	*14.0*
Unit B	*14.0*	*14.5*	*13.7*	*12.7*	*14.1*

$H_0: S_1^2 = S_2^2$

$H_1: S_1^2 \neq S_2^2$

R- Code and Inference

```
> Unit_A=c(14.1,10.1,14.7,13.7,14.0)
> Unit_B=c(14.0,14.5,13.7,12.7,14.1)
> var.test(Unit_A,Unit_B)

        F test to compare two variances

data:  Unit_A and Unit_B
F = 7.3304, num df = 4, denom df = 4, p-value = 0.07954
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.7632268 70.4053799
sample estimates:
ratio of variances
          7.330435
```

Here p valuve >0.05 ,then there is no evidence to reject the null hypothesis

Practice Problems

- A certain stimulus administered to each of the 13 patients resulted in the following increase of blood pressure: 5, 2, 8,-1, 3, 0, -2, 1, 5, 0, 4, 6, 8. Can it be concluded that the stimulus, in general, be accompanied by an increase in the blood pressure.
- The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with a standard deviation of 5 months. Random samples of 6 such bulbs have the following values: Life of bulbs in months: 24, 20, 30, 20, 20, and 18. Can you regard the producer's claim to valid at 1% level of significance

Practice Problems cont…

- The life time of electric bulbs for a random sample of 10 from a large consignment gave the following data: 4.2, 4.6, 3.9, 4.1, 5.2, 3.8, 3.9, 4.3, 4.4, 5.6 (in '000 hours). Can we accept the hypothesis that the average life time of bulbs is 4, 000 hours
- Data on weight (grams) of two treatments of NMU (nistroso- methyl urea) are recorded. Find out whether these two treatments have identical effects by using t test for sample means at 5% level of significance.

Sample	1	2	3	4	5	6	7	8	9	10	11	12
Treatments 0.2 %	2.0	2.7	2.9	1.9	2.1	2.6	2.7	2.9	3.0	2.6	2.6	2.7
0.4%	3.2	3.6	3.7	3.5	2.9	2.6	2.5	2.7				

Chi-square Test

Goodness of Fit and Independence of Attributes

Chi-square test for independence of attributes

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another. Assume Oij is the observed frequency count of events belonging to both i-th category of x and j-th category of y. Also assume Eij to be the corresponding expected count if x and y are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α.

$$\chi^2 = \sum \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

Problem 1 :The below table gives the distribution of students according to the family type and the anxiety level

Family type	Anxiety level		
	Low	Normal	High
Joint family	35	42	61
Nuclear family	48	51	68

R- Code and Interpretation

```
> data<-matrix(c(35,42,61,48,51,68),ncol=3,byrow=T)
> data
     [,1] [,2] [,3]
[1,]   35   42   61
[2,]   48   51   68
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 0.53441, df = 2, p-value = 0.7655
```

Here P value (0.7655) > 0.05. Hence there is no evidence to reject the Null hypothesis. So we consider the anxiety level and family type as independent.

Problem

In the built-in data set survey, the Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".We can tally the students smoking habit against the exercise level with the table function in R. The result is called the contingency table of the two variables.

```
> library(MASS)
> tbl = table(survey$Smoke, survey$Exer)
> tbl
```

	Freq	None	Some
Heavy	7	1	3
Never	87	18	84
Occas	12	3	4
Regul	9	1	7

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

R- Code and Interpretation

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
```

As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

Enhanced Solution

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl.

```
> ctbl = cbind(tbl[,"Freq"], tbl[,"None"] + tbl[,"Some"])
> ct

> ctbl
      [,1] [,2]
Heavy    7    4
Never   87  102
Occas   12    7
Regul    9    8
```

R- Code

```
> chisq.test(ctbl)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571
```

Goodness of Fit

A biologist is conducting a plant breeding experiment in which plants can have one of four phenotypes. If these phenotypes are caused by a simple Mendelian model, the phenotypes should occur in a 9:3:3:1 ratio. She raises 41 plants with the following phenotypes.

Phenotype	*1*	*2*	*3*	*4*
count	*20*	*10*	*7*	*4*

Should she worry that the simple genetic model doesn't work for her phenotypes?

R- Code & Inference

```
> plants <- c(20, 10, 7, 4)
> chisq.test(plants, p = c(9/16, 3/16, 3/16, 1/16))

        Chi-squared test for given probabilities

data:  plants
X-squared = 1.9702, df = 3, p-value = 0.5786

Warning message:
In chisq.test(plants, p = c(9/16, 3/16, 3/16, 1/16)) :
  Chi-squared approximation may be incorrect
```

The Chi-squared distribution is only an approximation to the sampling distribution of our test statistic, and the approximation is not very good when the expected cell counts are too small. This is the reason for the warning.

Here the probability value p is greater than alpha level (0.05), so we do not reject the null hypothesis.

Fitting of Binomial Distribution with Goodness of Fit

A survey of 320 families with 5 children each revealed the following distribution:

Number of Boys	*5*	*4*	*3*	*2*	*1*	*0*
No of Girls	*0*	*1*	*2*	*3*	*4*	*5*
No of families	*14*	*56*	*110*	*88*	*40*	*12*

Is this result consistent with the hypothesis that male and female births are equally possible?

Solution :

Let us setup the null hypothesis that the data are consistent with the hypothesis of equal probability for male and female births.

R- CODE & INFERENCE

```
> x=c(5,4,3,2,1,0)
> n=5                          #Probability of 'r' male births in a family
> N=320                        #Total Number of families
> P<-0.5                       #Probability of Male Birth
> Obf<-c(14,56,110,88,40,12)   #Obsevred frequencies
> exf<-dbinom(x,n,P)*320       #Expected frequencies
> # check the Condintion Sum of Observed and Expected are Equal
> sum(Obf)
[1] 320
> sum(exf)
[1] 320
> chisq<-sum((Obf-exf)^2/exf)
> chisq
[1] 7.16
> qchisq(0.95,5)
[1] 11.0705
```

Calculated value of chi-square is less than the tabulated value ,it is not significant at 5 % level of significance and hence the null hypothesis of equal probability for male and female births.

Fitting of Poisson Distribution with Goodness of Fit

Fit a Poisson distribution to the following data and test the goodness of fit

X	0	1	2	3	4	5	6
f	275	72	30	7	5	2	1

```
> x<-0:6
> f<-c(275,72,30,7,5,2,1)
> lambda<-(sum(f*x)/sum(f))     #mean
> expf <-dpois(x,lambda)*sum(f)    #expcted frequencies
> f1=round(expf)
> # check obserevd and Expected frequencies Total
> sum(f)
[1] 392
> sum(f1)
[1] 393
> # here subtrat '1' from expected frequencies
> #The last 3 frequencies are less than 5 so combine these frequencies in Observation and Expected
> obf<-c(275,72,30,15)
> exf<-c(242,117,28,6)
> chisq<-sum(((obf-exf)^2)/exf)
> chisq
[1] 35.45055
> qchisq(0.95,2)
[1] 5.991465
```

Inference

Since calculated value of χ^2 =35.45055 is much greater than 5.99,it is highly significant .Hence we conclude that poisson distribution is not good fit to the given data

Fitting the Normal Distribution with Goodness of Fit

Problem : The following table displays a frequency distribution of heights of trees in a certain locality. Fit a normal distribution to the data and test the goodness of fit.

Class Interval	Frequency
13.20 – 20.90	2
20.90 – 28.60	10
28.60 – 36.30	16
36.30 – 44.00	37
44.00 – 51.70	43
51.70 – 59.40	39
59.40 – 67.10	29
67.10 – 74.80	13
74.80 – 82.50	06
82.50 – 90.20	05

Heights of Trees (in inches)

```
> midy<-seq(17.05,86.5,length=10)
> f<-c(2,10,16,37,43,39,29,13,6,5)
> mean<-sum(f*midy)/sum(f)
> sd<-sqrt(sum(f*(midy-mean)^2)/sum(f))
> l<-seq(13.2,82.5,length=10)
> l<-c(l,90.2)
> cdf<-pnorm(l,mean,sd)
> cdf<-c(0,cdf,1)
> pcf<-diff(cdf)
> f<-c(0,f,0)
> ex<-round(pcf*sum(f),4)
> fr<-data.frame(f,ex)
> obf<-c(12,16,37,43,39,29,13,11)
> exf<-c(sum(ex[c(1,2,3)]),ex[c(4:9)],sum(ex[c(10,11,12)]))
> sum(obf)
[1] 200
> sum(exf)
[1] 200
> chisq<-sum((obf-exf)^2/exf)
> chisq
[1] 2.153974
> qchisq(0.95,5)
[1] 11.0705
```

Inference

Here chi-square cal value is less than chi-square tab value then there is no evidence to reject our null hypothesis.ie the fit of normal distribution is good

Practice Problems

1. The following data come from a hypothetical survey of 920 people (Men, Women) that ask for their preference of one of the three ice cream flavors (Chocolate, Vanilla, Strawberry). Is there any association between gender and preference for ice cream flavor?

Gender\flavor	Chocolate	Vanilla	Strawberry
Men	100	120	60
Women	350	320	150

2. As a part of quality improvement project focused on a delivery of mail at a department office within a large company, data were gathered on the number of different addresses that had to be changed so that the mail could be redirected to thee correct mail stop. Table shows the frequency distribution. Fit binomial distribution and test goodness of fit

x	0	1	2	3	4
fx	5	20	45	20	10

The number of Addresses Needing Change