DESCRIPTIVE STATISTICS

Experiment-1

JUNE 18, 2021

BIMAL PARAJULI
20BDS0405

Descriptive Statistics

AIM:

Using R software, compute all descriptive statistics and interpret the result.

R-Syntax:

R-Code	Description
Mean(X)	To compute the mean of X
Median(X)	To obtain the median of X
Quartile(X)	Find all the quartiles of X.
Range(X)	To find the range of X
Var(X)	To find the variance of X
Table(X)	To create the frequency table of X
X[n]	To obtain the data in nth column of data vector X.
Var(X, Y)	Calculate the covariance of X and Y
IQR(X)	Find the interquartile range of X.
Length(X)	Find the length of vector X.

Tools Used:

R-Studio (IDE)
R (programming language)

Problem 1:

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)          #given age data
> x
 [1] 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
> md=median(x)                                                               #median
> md
[1] 20.5
> y=x[x<25]                                                                 #data of ages under 25
> y
 [1] 18 19 19 19 19 20 20 20 20 20 21 21 21 21 22 23 24
> median(y)
[1] 20
> rx=table(x)                                                               #frequency table of given data
> rx
x
18 19 20 21 22 23 24 27 30 36
 1  4  5  4  1  1  1  1  1  1
> mode=which(xr==max(xr))                                                   #mode
> mode
20
 3
```

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 the mean and median of all students.

b) Find median age of all students under 25 years.

c). Find the modal age of all students.

R code :-

```
> x = c(
> mean(x)   # mean
 [1] 22
> median(x)   # median
 [1] 20.5
> y = x[x<25]   # median of under 25.
> md = median(y)
> md
 [1] 20

> xr = table(x)
> mode = which(xr == max(xr))   # mode.
> mode
 20
  3
```

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 institutes 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)                                  #given X data
> x
[1] 0 1 2 3
> f=c(8,11,5,1)                                    # Corresponding frequency data
> f
[1]  8 11  5  1
> y=rep(x,f)                                       # Individual data
> y
 [1] 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 3
> mean=(sum(y)/sum(f))                             #mean
> mean
[1] 0.96
> median(y)                                        #median
[1] 1
>
```

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 institutes have served on the faculty? Following is the frequency distribution of their responses.

X	0	1	2	3
f	8	11	5	1

Find the mean and median of the distribution:

Rcode:-

```
>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
```

(sum(y))

Problem 3:

Compute mean, median 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

R- Code:

```
Console  Terminal  Jobs
~/
> 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)
> f
[1]  4  6 28 58 64 30  5  5
> mean=sum(x*f)/sum(f)           #mean
> mean
[1] 165.175
>
>
```

```
Console  Terminal  Jobs
~/
>
> #For median:
> cumulative= cumsum(f)             #cumulative frequency og given distribution
> cumulative
[1]   4  10  38  96 160 190 195 200
>
> N=sum(f)                         #total number of observations
> N
[1] 200
> position_Q2=min(which(cumulative>n/2))  #serial number of median class
Error in which(cumulative > n/2) : object 'n' not found
> position_Q2=min(which(cumulative>N/2))  #serial number of median class
> position_Q2
[1] 5
>
> h=5
> h
[1] 5
>
> f_q2=f[position_Q2]             #frequency corresponding to median
> f_q2
[1] 64
>
> cumulative_q2=cumulative[position_Q2-1]
> cumulative_q2
[1] 96
>
> lower=x[position_Q2]-h/2
> lower
[1] 165
>
> median=lower+(((n/2)-cumulative_q2)/f_q2)*h
Error: object 'n' not found
> median=lower+(((N/2)-cumulative_q2)/f_q2)*h
> median
[1] 165.3125
> #Hence the median is calculated.
>
```

```
>
>
> #For Mode:
> position_mode=which(f==max(f))
> position_mode
[1] 5
>
> f0=f[position_mode-1]
> f0
[1] 58
>
> f1=f[position_mode]
> f1
[1] 64
>
> f2=f[position_mode+1]
> f2
[1] 30
>
> lower_mo=x[position_mode]-h/2
> lower_mo
[1] 165
>
> mode=lower_mo+ ((f1-f0)/(2*f1-f0-f2))*h
> mode
[1] 165.75
>
> #Hence the mode of data is calculated
>
```

Problem 3: Compute mean, median and mode 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

Rcode:-

```
>mid = seq(147.5, 182.5, 5)
>mid
 [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) :
>f
 [1] 4 6 28 58 64 30 5 5
>mean=sum(x*f)/sum(f)          # mean
>mean
 [1] 165.175
>#For Median:
```

```
>cl=cumsum(frequency)
>cl
 [1]  4  10  38  96  160  190  195  200
>n=sum(frequency)
>n
 [1] 200
>ml=min(which(cl>=n/2))        # The serial number of the median class.
>ml
 [1] 5
>h=5
>h
 [1] 5
>f=frequency[ml]               # frequency of the median class.
>f
 [1] 64
>C=cl[ml-1]                    # Cumulative frequency of median class.
>C
 [1] 96
>l=mid[ml]-h/2
>l
 [1] 165
>median=l+h*((n/2)-C)/f)       # median.
>median
 [1] 165.3125
```

Mode:-

```
>m=which(frequency==max(frequency))   # serial number of modal class.
>m
 [1] 5
>fm=frequency[m]               # frequency of modal class.
>fm
 [1] 64
>f1=frequency[m-1]             # frequency of pre modal class.
>f2=frequency[m+1]             # frequency of post modal class.
>f1
 [1] 58
>f2
 [1] 30
```

```
>l=midx[m]-h/2
>l
 [1] 165
>mode=l+(fm-f1)/(2*fm-f1-f2))*h
>mode
 [1] 165.75
```

Problem 4:

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, and 1.3.

Compute all the measures of dispersion.

```
Console ~/R/
> 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)                                #range of given data
> res
[1] 1.0 1.7
>
> diff(res)
[1] 0.7
>
> var(x)                                      #Variance of X
[1] 0.049
>
> sd(x)                                       #standard Deviation of X
[1] 0.2213594
>
> quartile(x)                                 #quartiles of x
Error in quartile(x) : could not find function "quartile"
> quantile(x)                                 #quartiles of x
   0%   25%   50%   75%  100%
1.000 1.200 1.300 1.475 1.700
>
> IQR(x)                                      #inter-quartile_range of X
[1] 0.275
>
>
> #For mean deviation about mean, median, mode:
> y=abs(x-mean(x))                            #absolute deviations from the mean.
> y
 [1] 0.13 0.07 0.03 0.27 0.33 0.17 0.37 0.23 0.13 0.03
>
> md_mean=sum(y)/length(y)                    # Mean Deviation from mean
> md_mean
[1] 0.176
>
> md_median=abs(x-median(x))                  #absolute deviations from median.
> md_median
 [1] 0.1 0.1 0.0 0.3 0.3 0.2 0.4 0.2 0.1 0.0
>
>
> #Since, this problem is a bimodal problem, mode calculation is not possible as of now.
>
```

An entomologist studying morphological variation in species of mosquito recorded the following data on body length:

1.2, 1.4, 1.3, 1.6, 10, 1.5, 1.7, 1.1, 1.2, 1.3

Compute all the measures of dispersion.

Rcode:-

```
>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)
 [1]. 0.049
>sd(x)
  [1]. 0.02213594
>quantile(x)
  0%  25%  50%  75%  100%.
1.000 1.200 1.300 1.475 1.700.
>IQR(x)                          # Interquartile range of X.
 [1] 0.275
>y=abs(x-mean(x))                # absolute deviations from mean.
>md_mean=sum(y)/length(y)        # Mean deviation from mean.
>md_median=sum(abs(x-median(x)))/length(x)  # Mean deviation from Q2.
># since,this is bimodal, mode is not possible to calculate.
```

CORRELATION AND REGRESSION

Experiment-2

JUNE 18, 2021

BIMAL PARAJULI
20BDS0405

Correlation Definition:-

Correlation refers to the relationship between two or more variables. Simple correlation studies the relationship between two variables. Correlation analysis attempts to determine the degree of relationship between variables.

Measures of Correlation:

- Scatter Diagram
- Karl Pearson's Coefficient of Correlation
  It is defined as the ratio of covariance between x and y say Cov (X, Y) to the product of the standard deviations of X and Y, say σ (X) and σ (Y)

$$i.e \qquad r_{XY} = \frac{Cov(XY)}{\sigma_X \sigma_Y}$$

- SPEARMAN'S RANK CORRELATION COEFFICIENT
  Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient ρ is given by

$$\rho = 1 - \left[\frac{6\sum d^2}{n(n^2-1)}\right]$$

- KENDALL'S COEFFICIENT OF CONCURRENT DEVIATIONS
  The Kendall's coefficient of concurrent deviations is denoted by rc and defined as

$$r_c = \pm\sqrt{\pm\left[\frac{2C-n}{n}\right]}$$

  Where, C = Number of concurrent deviations or position signs of (DX, DY); n = Number of pairs of deviations

Regression:

DEFINITION

Regression analysis is a statistical method of determining the mathematical functional relationship connecting independent variable(s) and a dependent variable.

Its types are:

- Simple linear Regression

  In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete and nature of relationship is linear. This relationship can be expressed using a straight line equation (linear regression) that best approximates all the individual data points.
  The general form of the simple linear regression equation is Y = a + bX + e, where 'X' is independent variable, 'Y' is dependent variable, a' is intercept, 'b' is slope of the line and 'e' is error term.

- Multiple linear Regression

  Multiple linear regression uses two or more independent variables to estimate the value(s) of the response variable (Y). The general form of the multiple linear regression equation is Y = a + b1X1 + b2X2 + b3X3 + ... + btXt + e

- Non Linear Regression

Problem 1:

Calculate the Coefficient of correlation of x and y from the given data:

X	23	27	28	28	29	30	31	33	35	36
y	18	20	22	27	21	29	27	29	28	29

R- Code:

```
Console ~/R/
> x=c(23,27,28,28,29,30,31,33,35,36)                    # Given X data
> x
 [1] 23 27 28 28 29 30 31 33 35 36
> y=c(18,20,22,27,21,29,27,29,28,29)                    # Given Y data
> y
 [1] 18 20 22 27 21 29 27 29 28 29
>
> var(x)                                                #variance of X
[1] 15.33333
> var(y)                                                #variance of Y
[1] 18.22222
> var(x,y)                                              #Co-variance of X and Y
[1] 13.66667
>
> var(x,y)/sqrt(var(x)*var(y))                          #Coefficient of Correlation
[1] 0.8176052
> cor.test(x,y,method="pearson")

        Pearson's product-moment correlation

data:  x and y
t = 4.0164, df = 8, p-value = 0.003861
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3874142 0.9554034
sample estimates:
      cor
0.8176052

> |
```

Calculate the coefficient of correlation of x and y from given data.

x	23	27	28	28	29	30	31	33	35	36
y	18	20	22	27	21	29	27	29	28	29

R-code:-

```
>x=c(23,27,28,28,29,30,31,33,35,36).
>y=c(18,20,22,27,21,29,27,29,28,29).
>var(x)
 [1] 15.33333
>var(y)
 [1] 18.22222
>var(x,y)
 [1] 13.6667
>r=var(x,y)./sqrt(var(x)*var(y))
>r
[1] 0.8176052
>cor.test(x,y, method="pearson").
```

output:-

```
        Pearson's product-moment correlation.
data: x and y.
t=4.0164, df=8, p-value= 0.003861
alternative hypothesis: true correlation is not equal to 0.
95 percent confidence interval:
  0.387142    0.9554034
Sample estimates:
   cor
0.8176052
```

Problem 2:

Twelve recruits were subjected to selection test to ascertain their sustainability for a certain course of training. At the end of training, they were given a proficiency test. The marks scored by the recruits are recorded below:

Recruit	1	2	3	4	5	6	7	8	9	10	11	12
Selection Test Score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency test Score	48	55	45	60	43	80	58	50	77	46	47	65

R- Code:

```
Console ~/R/
> selection=c(44, 49, 52, 54, 47, 76, 65, 60, 63, 58, 50, 67)        #Selection Test Score
> selection
 [1] 44 49 52 54 47 76 65 60 63 58 50 67
>
>
> proficiency=c(48, 55, 45, 60, 43, 80, 58, 50, 77, 46, 47, 65)     # Proficiency Test Score
> proficiency
 [1] 48 55 45 60 43 80 58 50 77 46 47 65
>
>
> cor.test(selection, proficiency, method = "spearman")

        Spearman's rank correlation rho

data:  selection and proficiency
S = 80, p-value = 0.01102
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7202797

>
```

Problem :- Twelve recruits were subjected to selection test to ascertain their sustainability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below :-

Recruit	1	2	3	4	5	6	7	8	9	10	11	12
Selection Test score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency Test score	48	55	45	60	43	80	58	50	77	46	47	65

Calculate the rank correlation coefficient and comment on result

Solution :-

```
>selection = c(44, 49, 52, 54, 47, 76, 65, 60, 63, 58, 50, 67).
>proficiency = c(48, 55, 45, 60, 43, 80, 58, 50, 77, 46, 47, 65).
>cor.test (selection, proficiency, method = "spearman").
```

output :

```
Spearman's rank correlation rho
data: selection and proficiency.
S = 80, p-value = 0.01102
alternative hypothesis: true rho is not equal to 0.
sample estimates:
rho
0.7202797.
```

Problem 3:

The body weight and BMI of 12 school going children are given in the following table. Fit a simple regression model of BMI on weight and examine the results.

Weight	15	26	27	25	25.5	27	32	18	22	20	26	24
BMI	13.35	16.12	16.74	16.00	13.59	15.73	15.65	13.85	16.07	12.8	13.65	14.42

R- Code:

```
Console ~/R/
> weight=c(15, 26, 27, 25, 25.5, 27, 32, 18, 22, 20, 26, 24)              #Given weight data
> weight
 [1] 15.0 26.0 27.0 25.0 25.5 27.0 32.0 18.0 22.0 20.0 26.0 24.0
>
>
>
> bmi=c(13.35, 16.12, 16.74, 16.00, 13.59, 15.73, 15.65, 13.85, 16.07, 12.8, 13.65, 14.42)   #Given BMI data
> bmi
 [1] 13.35 16.12 16.74 16.00 13.59 15.73 15.65 13.85 16.07 12.80 13.65 14.42
>
>
> cor(weight, bmi)                #Correlation between weight and BMI
[1] 0.5790235
>
> model<- lm(bmi~weight)
> summary.lm(model)

Call:
lm(formula = bmi ~ weight)

Residuals:
     Min       1Q   Median       3Q      Max
-1.52988 -0.75527  0.04426  0.95286  1.57397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.73487    1.85405   5.790 0.000175 ***
weight       0.17096    0.07612   2.246 0.048524 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.155 on 10 degrees of freedom
Multiple R-squared:  0.3353,    Adjusted R-squared:  0.2688
F-statistic: 5.044 on 1 and 10 DF,  p-value: 0.04852

>
```

Problem:-

The body weight and BMI of 12 school going children are given in the following table:- Fit a simple regression model BMI on weight and examine the results.

Weight	15	26	27	25	25	27	32	18	22	20	26	24
BMI	13.35	16.12	16.74	16.00	13.59	15.73	15.65	13.85	16.07	12.8	13.65	14

R-code:

```
>weight = c(15,26,27,25,255, 27,32,18,22,20,26,24).
> bmi = c(13.35,16.12, 16.74,16.00,13.59,15.73,15.65,13.85,16.07,12.8,
                                                    13.65, 14.42).
> cor (weight, bmi).
> model <- lm(bmi~weight).
> summary.lm(model).
```

Output:

```
Call:
lm(formula = bmi ~ weight).

Residuals:
     Min       1Q   Median       3Q      Max
-1.52988 -0.78527  0.04426  0.95858  1.57397

Coefficients:-
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.73487    1.85405   5.790 0.000175***
weight       0.17096    0.07612   2.246 0.048524*
---
Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1

Residual standard Error: 1.155 on 10 degrees of freedom.
Multiple R-squared: 0.3353, Adjusted R-squared: 0.2688
f-Statistic: 5.044 on 1 and 10 DF, p-value: 0.04852.
```

BINOMIAL, POISSON AND NORMAL DISTRIBUTION

Experiment 3

JUNE 30, 2021

BIMAL PARAJULI

20BDS0405

Exp-3 Binomial, Poissons and Normal Distributions.

Using R software, obtain the binomial probabilities, poisson probability and normal probability and also express it diagramatically.

⇒ 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 the vector of numbers.
p is the vector of probabilities.
n is the number of observations.
size is the number of trials.
prob is the probability of success of each trial.

#pbinom()

Example1:- The probability of getting 26 or less heads from 51 tosses of a coin.

R-code:-

```
> x<-pbinom(26,51,0.5)
> print(x)
```

[1] 0.610116.

dbinom()

This function gives the probability distribution at each point.

```
> x <- seq(0,50,by=1)        # Creates a sample of 50 numbers increments by 1.
> y<- dbinom(x,50,0.5)       # Creates binomial distribution.
> png(file="dbinom.png")     # Gives the chart a filename.
> plot(x,y)                  # plots the graph of sample.
> dev.off()                  # Saves the file.
```

* qbinom()

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

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

R-code:-

```
X ← qbinom(0.25, 51, 1/2)
print(X).
[1] 23
```

* rbinom()

This function ~~takes the probability value and gives a~~ generates required number of random values of a given probability from the given sample.

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

R-code:-

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

Poisson's Distribution in R:

R-codes:-

- dpois(x, lambda) # the probability of x success in a period when 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 specified cumulative propability (percentile) p.
- rpois(n, lambda) # returns n random numbers from the Poisson distribution :-

Eg:-

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

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

2) What is P(X>=2) with lambda 3?

```
> 1- ppois(2,3)
[1] 0.5768099
```

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

It's standard notation 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, it's transformation $Z = \frac{X-\mu}{\sigma}$ follows standard normal distribution (mean 0 and unit variance).

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

The distribution function of standard normal distribution is:-

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

R has four built in functions to generate normal distributions. They are:-

- dnorm(x, mean, sd) # Calculates the height of probability distribution at each point for given mean and standard deviation.
- pnorm(x, mean, sd) # Gives the probability of a randomly distributed random number to be less than the value of given number. It is also called "Cumulative Distribution Function".
- qnorm(p, mean, sd) # It takes the probability value and gives a number whose cumulative value matches the probability value.
- rnorm(n, mean, sd) # It generates random numbers whose distribution is normal. It takes sample size as input and generates that many random numbers.

Problems:-

1). If a committee has 7 members, find the probability of having more female members than male members given that the probability of having a male or female member is equal.

Sol".

The probability of having female member = 0.5

The probability of having male member = 0.5.

To have more female members, the number of females should be greater than or equal to 4.

R-code:-

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

```
> #probability of having a female member is : 0.5
> #probability of having a male member is : 0.5
>
> #probability of having more female than male is same as having 4 or more females.
>
> 1-pbinom(3,7,0.5)
[1] 0.5
>
> #Hence the probability of having more women than
  men is 0.5
```

2). The weekly wages of 1000 workman are normally distributed around a mean of Rs 70 with SD 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.

R-codes:-

```
>#(i) Between Rs 69 and Rs 72.
>(pnorm(72, mean=70, sd=5) - pnorm(69, mean=70, sd=5))*1000
 [1] 234.6815
>#(ii) Hence, the number of workers whose wage lie between 69 and 72 is 234
>#(ii) Less than Rs 69.
>pnorm(69, mean=70, sd=5))*1000.
 [1] 420.7403
>#Hence, the number of workers whose wave is less than Rs 69 is 421.
>#(iii) More than Rs 72.
>(1-pnorm(72, mean=70, sd=5))*1000
 [1] 1344.5783.
```

```
> #(i)between Rs 69 and Rs 72
> (pnorm(72, mean=70,sd=5) - pnorm(69, mean=70, sd=5))*1000
[1] 234.6815
> #(ii)Less than Rs69.
> (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 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
>
```

SAMPLING TECHNIQUES (LARGE SAMPLING)

LAB Experiment 4

JUNE 28, 2021

BIMAL PARAJULI
20BDS0405

Exp-4

Sampling techniques

Large sample Test (n>30)

Z-test (One Sample)

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

Problem-1

Suppose a 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 lasted only 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 $\mu \geq 10000$.

R-code:-

```
>xbar=9900                          #Sample mean
> mu0 = 10000                       #hypothesised 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 0.05 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 liftime of the bulb is above 10,000 hours.

```
> xbar = 9900                                   #Sample mean
> mu0  = 10000                                  #Hypothesized Value
> sigma = 120                                   #population Standard Deviation
> n = 30                                        #Sample Size
> z = (xbar - mu0 )/(sigma/sqrt(n))             #Test statistic
> z
[1] -4.564355
>
>
> 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 -1.669. Hence, at .05 significance level, we
 reject tha claim that mean lifetime of the bulbnis above 10,000 hours.
>
```

Upper Tail Test of Population Mean with Known Variance.

The null hypothesis of upper tail test of population mean can be expressed as $\mu \leq \mu_0$
Let's define test statistic Z in terms of sample size, mean, standard deviation, $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Then, Null hypothesis is to be rejected if $z > z_0$.

Problem-2:-

Suppose the food label on a cookie bag states that there is at most 2grams 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?

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

R-code:-

```
>xbar=2.1                    # sample mean
>muD=2                       # hypothesized value.
>sigma=025                   # population standard deviation.
>n=35                        # Sample Size
>z=(xbar-muD)/(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 1.649. Here, at .05 significance level, we reject that claim that there is at most 2 grams of saturated fat in a cookie.

```
> 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))              # test statistic
> z
[1] 2.366432
>
> alpha =.05
> z.alpha = qnorm(1- alpha)
> z.alpha                                       # Critical Value
[1] 1.644854
>
>
> #Intepretation...
> #The test statistic 2.3664 is greater than the critical value of 1.6449. Hence, at .05 significance level,
 we reject that claim that there is at most 2 grams of saturated fat in a cookie.
>
```

SMALL SAMPLE TEST (T-TEST)

LAB Experiment 5

JULY 6, 2021

BIMAL PARAJULI
20BDS0405

Small Sample Test

- t-test for Single Mean.
  and
  t-test for difference of Mean.

Problem-1

An outage 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 cream. 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 of Salmonella in ice cream greater than 0.3 MPN/g?

R-codes:-

```
>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).
```

Output:-

```
        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
```

Inference:-

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

```
> x=c(0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418)
> x
[1] 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

>
```

Problem-2

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

```
>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).
```

Outputs:-

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

Inferances:-

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

```
> Unit_A =c(14.1, 10.1, 14.7, 13.7, 14.0)
> Unit_A
[1] 14.1 10.1 14.7 13.7 14.0
>
> Unit_B = c(14.0, 14.5, 13.7, 12.7, 14.1)
> Unit_B
[1] 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

>
```

CHI SQUARE TEST
GOODNESS OF FIT AND INDEPENDENCE OF ATTRIBUTES

LAB Experiment 6

JULY 6, 2021

BIMAL PARAJULI
20BDS0405

Lab 6

Chi-Square Test.

Goodness of fit and Independence of Attributes.

1). The below table gives the distribution of students according to family type and 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), nrol = 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.

```
>
> # Chi Square Test
> data <- matrix(c(35, 43, 61, 48, 51, 68), ncol = 3, byrow = T)
> data
     [,1] [,2] [,3]
[1,]   35   43   61
[2,]   48   51   68
> chisq.test(data)

        Pearson's Chi-squared test

data:  data
X-squared = 0.53926, df = 2, p-value = 0.7637

> #Here, P value (0.7637) > 0.05. Hence, there is no evidence to reject the null hypothesis . So, we consider
 the anxiety level and family type as independent
>
```