Hypothesis Testing

Problem statement 1

A F&B manager wants to determine whether there is any significant difference in the diameter of the cutlet between two units. A randomly selected sample of cutlets was collected from both units and measured? Analyze the data and draw inferences at 5% significance level. Please state the assumptions and tests that you carried out to check validity of the assumptions.

Answer:

Rcode:

```
# Hypothesis testing assignment. 2- Sample T test


cutlets <- read.csv(file.choose())

View(cutlets)

attach(cutlets)


# normality test

shapiro.test(Unit.A)

shapiro.test(Unit.B)


# Variance test

var.test(Unit.A, Unit.B)


# Two sample T-test

t.test(Unit.A, Unit.B, alternative = "two.sided", conf.level = 0.95, correct= TRUE)

t.test(Unit.A, Unit.B, alternative = "greater", conf.level = 0.95, correct= TRUE)

t.test(Unit.A, Unit.B, alternative = "less", conf.level = 0.95, correct= TRUE)
```

Console:

```
> cutlets <- read.csv(file.choose())
> View(cutlets)
```

hypothesis_1.R × | cutlets × | hy

Filter

	Unit.A	Unit.B
1	6.8090	6.7703
2	6.4376	7.5093
3	6.9157	6.7300
4	7.3012	6.7878
5	7.4488	7.1522
6	7.3871	6.8110
7	6.8755	7.2212
8	7.0621	6.6606
9	6.6840	7.2402
10	6.8236	7.0503

Showing 1 to 10 of 35 entries, 2 total columns

```
> attach(cutlets)

> # normality test
> shapiro.test(Unit.A)

	Shapiro-Wilk normality test

data:  Unit.A
W = 0.96495, p-value = 0.32

> shapiro.test(Unit.B)

	Shapiro-Wilk normality test

data:  Unit.B
W = 0.97273, p-value = 0.5225

> # Variance test
> var.test(Unit.A, Unit.B)

	F test to compare two variances

data:  Unit.A and Unit.B
F = 0.70536, num df = 34, denom df = 34, p-value = 0.3136
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3560436 1.3974120
sample estimates:
ratio of variances
         0.7053649

> # Two sample T-test
> t.test(Unit.A, Unit.B, alternative = "two.sided", conf.level = 0.95,
correct= TRUE)
```

```
        Welch Two Sample t-test

data:  Unit.A and Unit.B
t = 0.72287, df = 66.029, p-value = 0.4723
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.09654633  0.20613490
sample estimates:
mean of x mean of y 
 7.019091  6.964297 

> 
> t.test(Unit.A, Unit.B, alternative = "greater", conf.level = 0.95, c
orrect= TRUE)

        Welch Two Sample t-test

data:  Unit.A and Unit.B
t = 0.72287, df = 66.029, p-value = 0.2362
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.07166173         Inf
sample estimates:
mean of x mean of y 
 7.019091  6.964297 

> 
> t.test(Unit.A, Unit.B, alternative = "less", conf.level = 0.95, corr
ect= TRUE)

        Welch Two Sample t-test

data:  Unit.A and Unit.B
t = 0.72287, df = 66.029, p-value = 0.7638
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.1812503
sample estimates:
mean of x mean of y 
 7.019091  6.964297 
```

Explanation –

In data set y variable i.e dependent variable is in continuous form and x variable i.e independent variable is in discrete form

Here we are comparing two units with each other so we use 2-t sample test

Error α = 0.05 and 95% confidence interval are given

Business problem – Is there any significant difference in diameter of the cutlet between two units?

1) Normality Test
   H_o = Data is normally distributed.
   H_a = Data is not normally distributed.

For unit A, P value= 0.32 > 0.05.(p high null fly)accepting null hypothesis. hence the data is normally distributed
For unit B, the P value= 0.5225 > 0.05. (p high null fly)accepting null hypothesis. hence the data is normally distributed

Conclusion- both have probability value greater than 0.05 so, we fail to reject Null

2) Variance Test
H_o = Variance of unit A is equal to the Variance of unit B.
H_a = Variance of unit A is not equal to the Variance of unit B.

The P value= 0.3136> 0.05, (p high null fly)accepting null hypothesis. hence the variance is equal.

Conclusion –we fail to reject Null

3) 2- sample T-test

a)

H_o = Average dimension produced by unit A is equal to average dimension produced by unit B.
H_a = Average dimension produced by unit A is not equal to average dimension produced by unit B.

The P value = 0.4723>0.05, (p high null fly) accepting null hypothesis. Hence the Two saple t- test is equal

Conclusion - We fail to reject null. The P value > 0.05

b)

H_o = Average dimension produced by unit A is greater than or equal to average dimension produced by unit B.
H_a = Average dimension produced by unit A is less than average dimension produced by unit B.

The P value = 0.2362>0.05, (p high null fly)accepting null hypothesis. Hence the Two sample t- test is equal

Conclusion - We fail to reject null. The P value > 0.05

c)

H_o = Average dimension produced by unit A is less than or equal to average dimension produced by unit B.
H_a = Average dimension produced by unit A is greater than average dimension produced by unit B.

The P value = 0.7638>0.05, (p high null fly)accepting null hypothesis. Hence the Two sample t- test is equal

Conclusion - We fail to reject null. The P value > 0.05

The Inference from the business problem is that there is no significant difference in the diameter of the cutlets produced by unit A and B.

Problem statement 2

A hospital wants to determine whether there is any difference in the average Turn Around Time (TAT) of reports of the laboratories on their preferred list. They collected a random sample and recorded TAT for reports of 4 laboratories. TAT is defined as sample collected to report dispatch.

Analyze the data and determine whether there is any difference in average TAT among the different laboratories at 5% significance level.

Answer:

Rcode:

```
lab_tat <- read.csv(file.choose())

View(lab_tat)

attach(lab_tat)


# Normality test

shapiro.test(Laboratory.1)

shapiro.test(Laboratory.2)

shapiro.test(Laboratory.3)

shapiro.test(Laboratory.4)


# Variance test

var.test(Laboratory.1, Laboratory.2)

var.test(Laboratory.2, Laboratory.3)

var.test(Laboratory.3, Laboratory.4)


# Anova test

stacked_data <- stack(lab_tat)
```

View(stacked_data)

anova_result <- aov(values ~ ind, data = stacked_data)

summary(anova_result)

Console:

```
> # one way annova test
> lab_tat <- read.csv(file.choose())
> View(lab_tat)
```

hypothesis_2.R × | lab_tat × | hypothesis_3.R × | hypothesisi_4.R ×

Filter

	Laboratory.1	Laboratory.2	Laboratory.3	Laboratory.4
1	185.35	165.53	176.70	166.13
2	170.49	185.91	198.45	160.79
3	192.77	194.92	201.23	185.18
4	177.33	183.00	199.61	176.42
5	193.41	169.57	204.63	152.60
6	179.45	197.00	181.51	161.12
7	191.37	166.36	214.21	154.02
8	166.81	169.60	183.43	163.25
9	158.81	175.36	191.60	152.79
10	165.88	198.68	208.43	161.98

Showing 1 to 10 of 120 entries, 4 total columns

```
> attach(lab_tat)
> # Normality test
> shapiro.test(Laboratory.1)

        Shapiro-Wilk normality test

data:  Laboratory.1
W = 0.99018, p-value = 0.5508
> shapiro.test(Laboratory.2)

        Shapiro-Wilk normality test

data:  Laboratory.2
W = 0.99363, p-value = 0.8637
> shapiro.test(Laboratory.3)

        Shapiro-Wilk normality test

data:  Laboratory.3
W = 0.98863, p-value = 0.4205

> shapiro.test(Laboratory.4)
```

```
       Shapiro-Wilk normality test

data:  Laboratory.4
W = 0.99138, p-value = 0.6619

> # Variance test
> var.test(Laboratory.1, Laboratory.2)

       F test to compare two variances

data:  Laboratory.1 and Laboratory.2
F = 0.77573, num df = 119, denom df = 119, p-value = 0.1675
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5406345 1.1130690
sample estimates:
ratio of variances
         0.7757342

> var.test(Laboratory.2, Laboratory.3)

       F test to compare two variances

data:  Laboratory.2 and Laboratory.3
F = 0.81785, num df = 119, denom df = 119, p-value = 0.2742
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5699887 1.1735038
sample estimates:
ratio of variances
         0.8178532

> var.test(Laboratory.3, Laboratory.4)

       F test to compare two variances

data:  Laboratory.3 and Laboratory.4
F = 1.2021, num df = 119, denom df = 119, p-value = 0.3168
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8377527 1.7247817
sample estimates:
ratio of variances
          1.202057

> # Anova test
> stacked_data <- stack(lab_tat)
> View(stacked_data)
```

hypothesis_2.R | stacked_data

Filter

	values	ind
1	185.35	Laboratory.1
2	170.49	Laboratory.1
3	192.77	Laboratory.1
4	177.33	Laboratory.1
5	193.41	Laboratory.1
6	179.45	Laboratory.1
7	191.37	Laboratory.1
8	166.81	Laboratory.1
9	158.81	Laboratory.1

Showing 1 to 10 of 480 entries, 2 total columns

```
> anova_result <- aov(values ~ ind, data = stacked_data)
> summary(anova_result)
             Df Sum Sq Mean Sq F value Pr(>F)    
ind           3  79979   26660   118.7 <2e-16 ***
Residuals   476 106905     225                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Explanation –

In data set y variable i.e dependent variable is in continuous form and x variable i.e independent variable is in discrete form

Here we are comparing more than two sample with each other so we use ANOVA sample test

Error α = 0.05 and 95% confidence interval are given

Business problem – Is there any difference in average TAT of reports of the laboratories on their preferred list ?

1) Normality Test
   H_o = Data is normally distributed.

H_a = Data is not normally distributed.

For laboratory 1, P value= 0.5508 > 0.05.(p high null fly)accepting null hypothesis. hence the data is normally distributed

For laboratory 2, P value= 0.8637 > 0.05.(p high null fly)accepting null hypothesis. hence the data is normally distributed

For laboratory 3, P value= 0.4205 > 0.05.(p high null fly)accepting null hypothesis. hence the data is normally distributed

For laboratory 4, P value= 0.6619 > 0.05.(p high null fly)accepting null hypothesis. hence the data is normally distributed

Conclusion- probability values are greater than 0.05 so, we fail to reject Null

2) Variance Test
   H_o = Variance of TAT of all the 4 laboratories are equal.
   H_a = Variance of TAT of all the 4 laboratories are not equal.

   For laboratory 1 and 2 ,The P value= 0.1675> 0.05, (p high null fly)accepting null hypothesis. hence the variance is equal.

   For laboratory 2 and 3 ,The P value= 0.2742> 0.05, (p high null fly)accepting null hypothesis. hence the variance is equal.

   For laboratory 3 and 4 ,The P value= 0.3168> 0.05, (p high null fly)accepting null hypothesis. hence the variance is equal.

Conclusion –we fail to reject Null

3) ANOVA test

   a)

   H_o = Average TAT of all the 4 laboratories are equal.
   H_a = Average TAT of all the 4 laboratories are not equal.

   The P value = 0.4723>0.05, (p high null fly) accepting null hypothesis. So average TAT of all the 4 laboratories are equal

   Conclusion - We fail to reject null. The P value > 0.05

b)

H_o = Average dimension produced by unit A is greater than or equal to average dimension produced by unit B.
H_a = Average dimension produced by unit A is less than average dimension produced by unit B.

The P value = 0.2362>0.05, (p high null fly)accepting null hypothesis. Hence average dimension produced by unit A is greater than or equal to average dimension produced by unit B

Conclusion - We fail to reject null. The P value > 0.05

c)

H_o = Average dimension produced by unit A is less than or equal to average dimension produced by unit B.
H_a = Average dimension produced by unit A is greater than average dimension produced by unit B.

The P value = 2e-16<0.05, (p low null go) accepting alternative hypothesis. So Average dimension produced by unit A is greater than average dimension produced by unit B

Conclusion - We accept alternative hypothesis. The P value < 0.05

The Inference from the business problem is that there is difference in average TAT of reports of the laboratories on their preferred list.

Problem statement 3

TeleCall uses 4 centers around the globe to process customer order forms. They audit a certain % of the customer order forms. Any error in order form renders it defective and has to be reworked before processing. The manager wants to check whether the defective % varies by centre. Please analyze the data at *5%* significance level and help the manager draw appropriate inferences

Answer:

Rcode:

buyer.ratio <- read.csv(file.choose())

View(buyer.ratio)

stack the data.

stacked_data <- stack(buyer.ratio)

View(stacked_data)

attach(stacked_data)

table(ind, values)

chi-square test

chisq.test(table(ind, values))

Console:

```
> buyer.ratio <- read.csv(file.choose())
> View(buyer.ratio)
```

hypothesis_3.R | buyer.ratio | stacked_data | hypothesisi_

Filter

	Observed.Values	East	West	North	South
1	Males	50	142	131	70
2	Females	435	1523	1356	750

```
> ## stack the data.
> stacked_data <- stack(buyer.ratio)
> View(stacked_data)
```

hypothesis_3.R × stacked_data ×

Filter

	values	ind
1	50	East
2	435	East
3	142	West
4	1523	West
5	131	North
6	1356	North
7	70	South
8	750	South

```
> attach(stacked_data)
> table(ind, values)
       values
ind     50 70 131 142 435 750 1356 1523
  East   1  0   0   0   1   0    0    0
  West   0  0   0   1   0   0    0    1
  North  0  0   1   0   0   0    1    0
  South  0  1   0   0   0   1    0    0
> ## chi-square test
> chisq.test(table(ind, values))

        Pearson's Chi-squared test

data:  table(ind, values)
X-squared = 24, df = 21, p-value = 0.2931
```

Explanation –

In data set y variable i.e dependent variable is in discrete form and x variable i.e independent variable is in discrete form

Here we are comparing more than two sample with each other so we use chi-square test

Error $\alpha = 0.05$ and 95% confidence interval are given

Business problem – whether the defective % varies by centre?

H_o =proportion of male-female ratio across 4 regions are same.

H_a = proportion of male-female ratio across 4 regions are different.

Here, p value = 0.2931 > 0.05, hence we fail to reject null. The proportions are equal.

The Inference from the business problem is that the proportions are equal. We fail to reject null

Problem statement 4

Fantaloons Sales managers commented that *%* of males versus females walking in to the store differ based on day of the week. Analyze the data and determine whether there is evidence at *5 %* significance level to support this hypothesis.

Answer:

Rcode:

```
fantaloons <- read.csv(file.choose())

View(fantaloons)


attach(fantaloons)

tablef <- table(Weekdays, Weekend)

tablef


### As the data is character we cannot carry normality test and variance test.

## 2 proportion test

prop.test(x= c(66,47), n=c(233, 167), conf.level = 0.95, alternative = "greater")

prop.test(x= c(66,47), n=c(233, 167), conf.level = 0.95, alternative = "two.sided")

prop.test(x= c(66,47), n=c(233, 167), conf.level = 0.95, alternative = "less")
```

Console:

```
> ## 2 proportion test
> fantaloons <- read.csv(file.choose())
> View(fantaloons)
```

hypothesisi_4.R | fantaloons

Filter

	Weekdays	Weekend
1	Male	Female
2	Female	Male
3	Female	Male
4	Male	Female
5	Female	Female
6	Female	Male
7	Female	Female
8	Female	Male
9	Female	Female
10	Female	Male

Showing 1 to 10 of 400 entries, 2 total columns

```
> attach(fantaloons)
> tablef <- table(Weekdays, Weekend)
> tablef
          Weekend
Weekdays Female Male
  Female    167  120
  Male       66   47
> ### As the data is character we cannot carry normality test and vari
ance test.
> ## 2 proportion test
> prop.test(x= c(66,47), n=c(233, 167), conf.level = 0.95, alternative
= "greater")

	2-sample test for equality of proportions with continuity
	correction

data:  c(66, 47) out of c(233, 167)
X-squared = 2.4909e-30, df = 1, p-value = 0.5
alternative hypothesis: greater
95 percent confidence interval:
 -0.07505851  1.00000000
sample estimates:
   prop 1    prop 2
0.2832618 0.2814371

> prop.test(x= c(66,47), n=c(233, 167), conf.level = 0.95, alternative
= "two.sided")

	2-sample test for equality of proportions with continuity
	correction

data:  c(66, 47) out of c(233, 167)
X-squared = 2.4909e-30, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.08943773  0.09308708
```

```
sample estimates:
   prop 1    prop 2 
0.2832618 0.2814371 
> prop.test(x= c(66,47), n=c(233, 167), conf.level = 0.95, alternative
= "less")

	2-sample test for equality of proportions with continuity
	correction

data:  c(66, 47) out of c(233, 167)
X-squared = 2.4909e-30, df = 1, p-value = 0.5
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.07870786
sample estimates:
   prop 1    prop 2 
0.2832618 0.2814371 
```

Explanation:

In data set y variable i.e dependent variable is in discrete form and x variable i.e independent variable is in discrete form

Here we are comparing two sample with each other so we use 2-proportion test

Error $\alpha = 0.05$ and 95% confidence interval are given

Business problem – whether proportion of male and female walking into a store is differ based on day of week or not?

H_o = Proportions of Male and Female are equal.

H_a = Proportions of Male and Female are not equal.

Here, p value = 1 > 0.05, hence we fail to reject null. The proportions are equal.

Now we will try to find out whose proportion is higher. We create another hypothe

H_o = Proportions of Male is less than or equal to Female

H_a = Proportions of Male is greater than Female

P value = 0.5 > 0.05 Hence proportion of Male is less than or equal to Female.

The Inference from the business problem is that the proportions of males walking in the store are less than or equal to the proportion of the female walking in the store based on the day of week . We fail to reject null