Week 2 Assignment

ALY 6015 Intermediate Analytics

Submitted to:

Joseph Manseau

Date :03/03/2020

Submitted by:

Deepak Natarajan (001088182)

Ashlesha Kshirsagar(001082234)

Introduction:

Hypothesis testing or significance testing is a method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample. In this method, we test some hypothesis by determining the likelihood that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true. (Sagepub). Inferential statistics is very useful in measuring the behavior of samples to learn more about the population behavior because finding the behavior of people is not that easy since the data is too large. To identify the behavior of sample we perform different tests like one sample t-test, two-sample t-test, paired t-test, proportion test and F-test

PART A

1.1 One sample t-test

Dataset

A numeric vector of 24 determinations of copper in whole meal flour, in parts per million.

Hypothesis

Null Hypothesis: The floor production company is producing whole meal flour with less than 1 part per million copper in it. Ho: μ<=1

Alternative Hypothesis: The floor production company is producing whole meal flour with greater than 1 part per million copper in it. Ha: μ>1

Assumption

- Data is independent.
- Data is collected randomly.
- The data is approximately normally distributed.

Code:

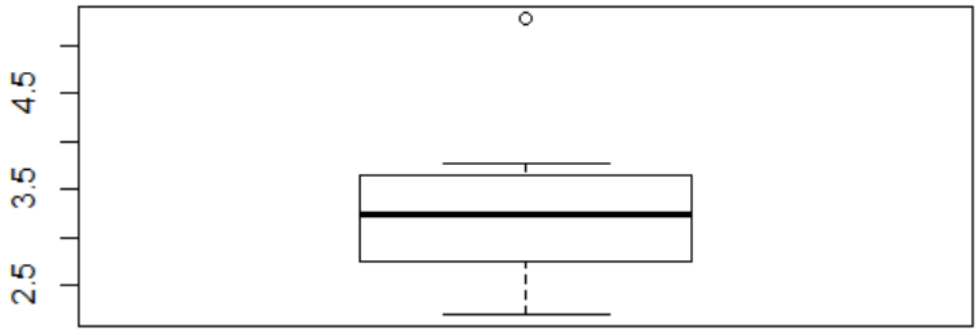
```
chem
CI(chem,ci=0.95)
chem= c(2.9, 3.1, 3.4, 3.4, 3.7 ,3.7 ,2.8 ,2.5, 2.4, 2.4, 2.7, 2.2 ,5.28
,3.37, 3.03, 3.03, 2.95, 3.77, 3.4 ,2.2, 3.5, 3.6, 3.7, 3.7)
t.test(chem, alternative = "greater", mu = 1)
```

Output:

```
        One Sample t-test

data:  chem
t = 3.0337, df = 23, p-value = 0.002952
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 2.427162      Inf
sample estimates:
mean of x
 4.280417
```

Graph



Observation:

This is right tail hypothesis t test. The significance level is 0.05.the test statistics value from R output is 3. 0337.The P value obtained 0.002952 which less than significance level 0.05.So we can reject the null hypothesis.

Result:

At 95% confidence level there is statistically significant evidence that the floor production company is producing whole meal flour with greater than 1 part per million copper in it

1.2 Two sample T test

Dataset

The heart and body weights of samples of male and female cats used for digitalis experiments.

The cats were all adult, over 2 kg body weight. We have performed two sample unpaired t test .

An unpaired t-test is used to compare two population means.

Hypothesis

Null Hypothesis: Male and Female cats have the same body weight. Ho: $\mu 1 - \mu 2 = 0$

Alternative Hypothesis: Male and Female cats have different body weight. Ha: $\mu 1 - \mu 2 \neq 0$

Code:

```
# c] - Two Sample Test
cats
summary(cats)
female<- subset(cats,subset=(cats$Sex=="F"))
male<- subset(cats,subset=(cats$Sex=="M"))
t.test(male["Bwt"] ,female["Bwt"] )
```

Output

```
        Welch Two Sample t-test

data:  male["Bwt"] and female["Bwt"]
t = 8.7095, df = 136.84, p-value = 8.831e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4177242 0.6631268
sample estimates:
mean of x mean of y
 2.900000  2.359574
```

Observation

It is observed that test Statistics is 8.709. The P value obtained from R is very less than significance level 0.05. Also, the value doesn't lie in the Confidence interval So we reject the null hypothesis.

Result

95% confidence level we have statistically significant evidence that male and female cats have different body weights.

1.3 Paired T test

Dataset

A list of two vectors, giving the wear of shoes of materials A and B for one foot each of ten boys. We are using this dataset to do a paired t-test. A **paired t-test** is used when we are interested in the difference between two variables for the same subject.

Hypothesis

Null Hypothesis: Wear of material A (µ1) was less than wear of Material B (µ2). Ha: µ1- µ2<=0

Alternative Hypothesis: Wear of material A (µ1) was greater than wear of Material B (µ2).

Ha: µ1- µ2>0

Code

```
#D] - Paired T test
shoes
t.test(shoes[["A"]] ,shoes[["B"]] ,alternative = "greater",paired=TRUE)
```

Output

```
        Paired t-test

data:  shoes[["A"]] and shoes[["B"]]
t = -3.3489, df = 9, p-value = 0.9957
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.6344264        Inf
sample estimates:
mean of the differences
                 -0.41
```

Observation

The p-value is 0.9957 which is greater than 0.05 and the confidence interval is -0.6355 to infinity. 0 lies between the confidence interval thus we fail to reject the null hypothesis.

Result

we do not have enough evidence to say that the wear of material A was greater than wear of material B

1.4Test of equal or given proportion

Dataset

Tests of the presence of the bacteria H. influenzae in children with otitis media in the Northern Territory of Australia. Using this dataset, we will be testing if the drug treatment resulted in a significant effect of presence of bacteria compared with the placebo.

Hypothesis

Null Hypothesis: Proportion of bacteria presence (P1) is equal to proportion of placebo (P2)

Ha: P1- P2 =0

Alternative Hypothesis: Proportion of bacteria presence (P1) is not equal to proportion of placebo (P2) Ho: P1- P2 ≠ 0

Code

```
# E] - Test if equal or given proportions
bacteria
View(bacteria)
my_table<- table(bacteria$y,bacteria$ap)
View(my_table)
prop.test(x=c(177,96),n=c(220,220),alternative = "two.sided",conf.level = 0.95)
```

Output

```
2-sample test for equality of proportions with continuity correction

data:  c(177, 96) out of c(220, 220)
X-squared = 61.767, df = 1, p-value = 3.867e-15
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2797293 0.4566343
sample estimates:
   prop 1    prop 2
0.8045455 0.4363636
```

Observation

Presence of Bacteria	Presence of active or placebo	Frequency
n	a	31
y	a	93
n	p	12
y	p	84

From Table we can say that the first value in x vector is 84+93 = 177 and second value is 84+12 = 96. The total of all the frequencies is 'n' which is 220. P value is less than 0.05 and 0 does not fall in the 95% confidence interval. Thus, we reject the null hypothesis

Result

At 95% confidence there is a statistically significant evidence that the proportion of presence of bacteria is not equal to the proportion of placebo

1.5 F test

Dataset

We will be using the cats dataset to test the following hypothesis using F-test:

Hypothesis

Null Hypothesis: There is no difference in the variance of female cat weight and male cat weight

Alternative Hypothesis: There is a difference in the variance of female cat weight and male cat weight

Code

```
# F]- F-test
male1 <-subset(cats,subset=(cats$Sex=="M"))
female1 <-subset(cats,subset=(cats$Sex=="F"))
var.test(male1$Bwt,female1$Bwt)
```

Output

```
        F test to compare two variances

data:  male1$Bwt and female1$Bwt
F = 2.9112, num df = 96, denom df = 46, p-value = 0.0001157
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.723106 4.703057
sample estimates:
ratio of variances
          2.911196
```

Observation

The p value 0.0001157 is less than 0.05. Also 1 does not fall in the 95% confidence interval thus we reject the null hypothesis also The F value here also indicates that there is a difference in the variances.

Result

There is enough evidence at 95% confidence level that there is a difference in variance of male cats and female cats.

PART 2

Dataset

We will be using medical charges data by health insurance to perform hypothesis testing using various tests like One sample t-test, two sample t-test and F test.

```
insurance_data <- read.csv('insurance.csv')
head(insurance_data)
```

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622

1.1 One sample t-test

One sample t test is used to compare the mean of sample to a hypothesized value.

Assumptions

- Data is independent.
- Data is collected randomly.
- The data is approximately normally distributed.

Hypothesis

Null Hypothesis: Body mass index(BMI) of male customers is less than or equal to 29.

Ho: $\mu \leq 29$.

Alternative Hypothesis: Body mass index(BMI) of male customers is greater than 29.

Ha: $\mu > 29$.

Code and Output

```
male_customers <- subset(insurance_data, subset=(insurance_data$sex == 'male'))
male_sample_bmi <- sample(male_customers$bmi, size = 40)# Random sampling
t.test(male_sample_bmi, alternative = "greater", mu = 29)#one sample t test
```

```
        One Sample t-test

data:  male_sample_bmi
t = 1.953, df = 39, p-value = 0.02901
alternative hypothesis: true mean is greater than 29
95 percent confidence interval:
 29.25818      Inf
sample estimates:
mean of x
 30.88063
```

Graph

```
options(repr.plot.width=5, repr.plot.height=4)
boxplot(male_sample_bmi,xlab = 'Body Mass Index')
```

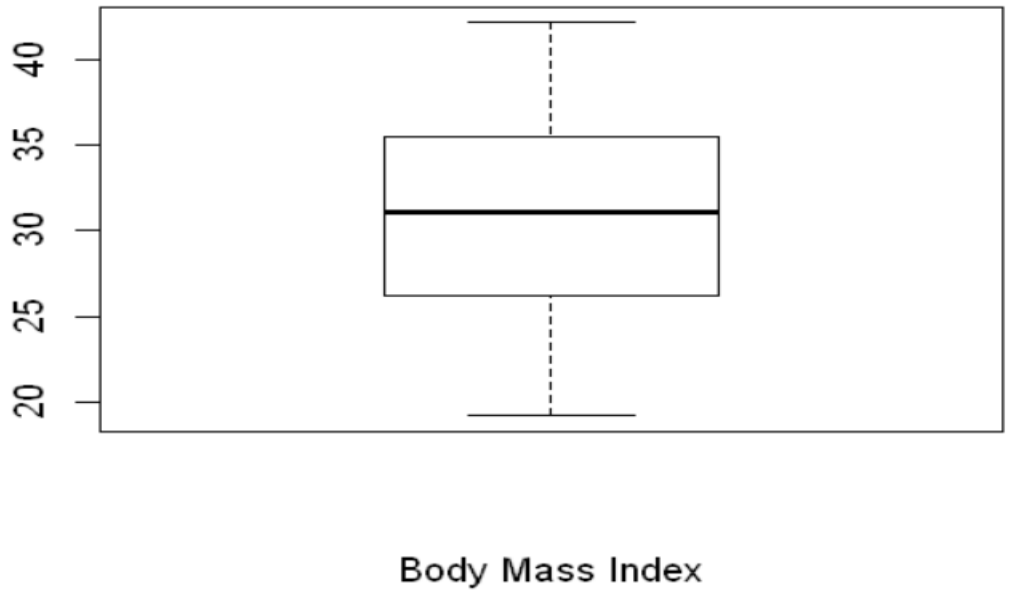


Fig 1.1 Box plot of BMI values of male customers

The box plot shows that the body mass index(BMI) of male customers is normally distributed.

Observation

This is Right tail hypothesis test. We performed random sampling to select sample data from the population to perform hypothesis testing. Since p-value (0.0290) is less than significance level(0.05). There is enough evidence that the BMI of male customers is greater than 29. So, we can reject Null Hypothesis. The Confidence interval range for BMI of male customers is from 29.26 to infinity and mean value of male customers BMI is 30.88. These findings show that the average Body Mass Index of male customers is greater than 29.

1.2 Two sample t-test

A two-sample t-test is used when you want to compare the means of two independent groups.

Assumptions

- Data is independent.
- Data is collected randomly.
- The data is approximately normally distributed.

Hypothesis

Null Hypothesis: Male smokers and Male non-smokers has same body mass index(BMI). Ho:

$\mu 1 - \mu 2 = 0$

Alternative Hypothesis: Male smokers and Male non-smokers have different body mass

index(BMI). Ho: $\mu 1 - \mu 2 \neq 0$

Code and Output

```
male_smokers <- subset(male_customers, subset=(male_customers$smoker == 'yes'))
male_non_smokers <- subset(male_customers, subset=(male_customers$smoker == 'no'))

# sampling
male_smokers_sample = sample(male_smokers$bmi, size = 40)
male_non_smokers_sample = sample(male_non_smokers$bmi, size = 40)

# two sample t test
t.test(male_smokers_sample ,male_non_smokers_sample, conf.level = 0.95)
```

```
        Welch Two Sample t-test

data:  male_smokers_sample and male_non_smokers_sample
t = 0.59914, df = 77.096, p-value = 0.5508
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.001096  3.723596
sample estimates:
mean of x mean of y
 31.25400  30.39275
```

Graph

```
options(repr.plot.width=5, repr.plot.height=5)
boxplot(male_smokers_sample,male_non_smokers_sample, xlab = 'Male smokers and non Smokers')
```

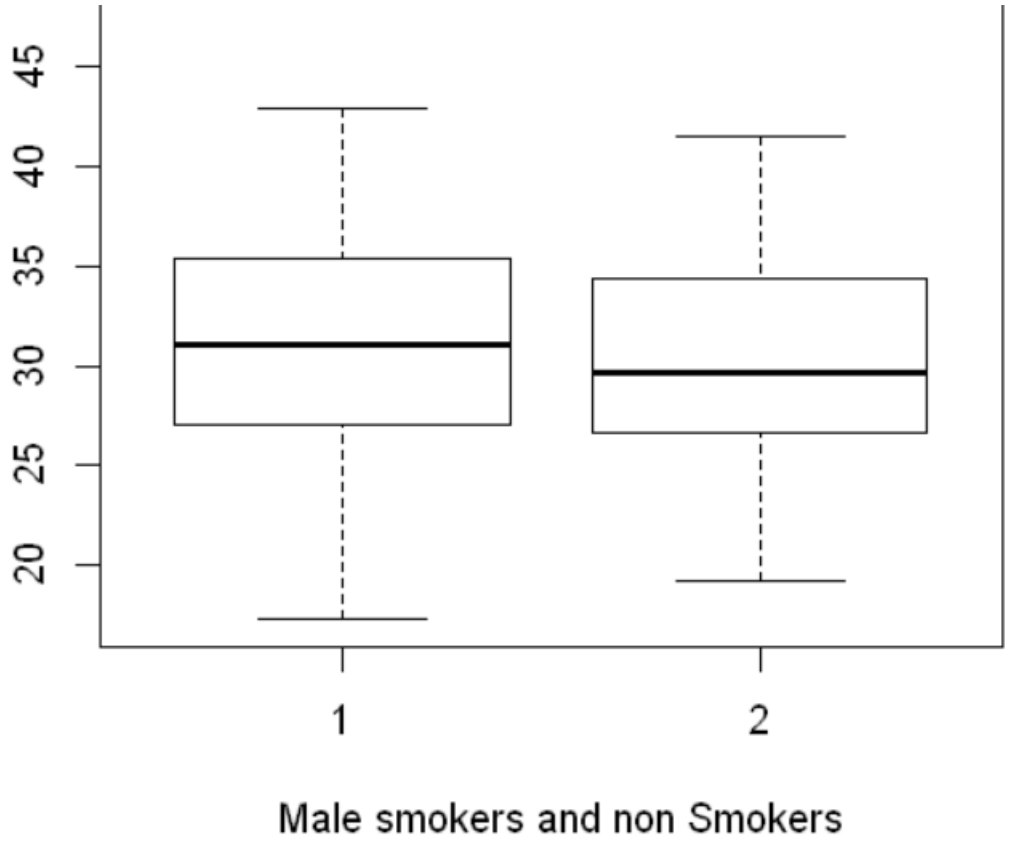


Fig 1.2 Box plot of BMI values of male smokers and male non-smokers

The box plot shows that the body mass index(BMI) of male smokers and male non-smokers are normally distributed.

Observation

This is two-tailed hypothesis test. We performed random sampling to select sample data from male smokers and male non-smokers to perform hypothesis testing. Since p-value (0.5508) is greater than significance level(0.05). we don't have enough evidence to prove that male smokers and nonsmokers have different Body mass index. So, we fail to reject Null Hypothesis. The

Confidence interval range for mean difference between two groups of male customers is from -2.001 to 3.724 and mean value of male smokers and nonsmokers are 31.25 and 30.39 respectively. These findings show that the irrespective of whether the male customers smoking or not body mass index is remains same.

1.3 F-test

A two-sample t-test is used when you want to compare the population variances of two groups.

Assumptions

- Data is independent and collected randomly.
- The data is approximately normally distributed.

Hypothesis

Null Hypothesis: Male customers and Female customers has same body mass index(BMI). Ho: $\sigma 1^2 = \sigma 2^2$

Alternative Hypothesis: Male customers and Female customers have different body mass index (BMI). Ho: $\sigma 1^2 \neq \sigma 2^2$

Code and Output

```
#F test

male_customers <- subset(insurance_data, subset=(insurance_data$sex == 'male'))
female_customers <- subset(insurance_data, subset=(insurance_data$sex == 'female'))

# sampling
male_customers_sample = sample(male_customers$bmi, size = 40)
female_customers_sample = sample(female_customers$bmi, size = 40)

var.test(male_customers_sample, female_customers_sample)
```

```
        F test to compare two variances

data:  male_customers_sample and female_customers_sample
F = 0.71973, num df = 39, denom df = 39, p-value = 0.3086
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3806628 1.3608002
sample estimates:
ratio of variances
         0.7197263
```

Graph

```
options(repr.plot.width=5, repr.plot.height=5)
boxplot(male_customers_sample,female_customers_sample, xlab = 'Male and female customers')
```

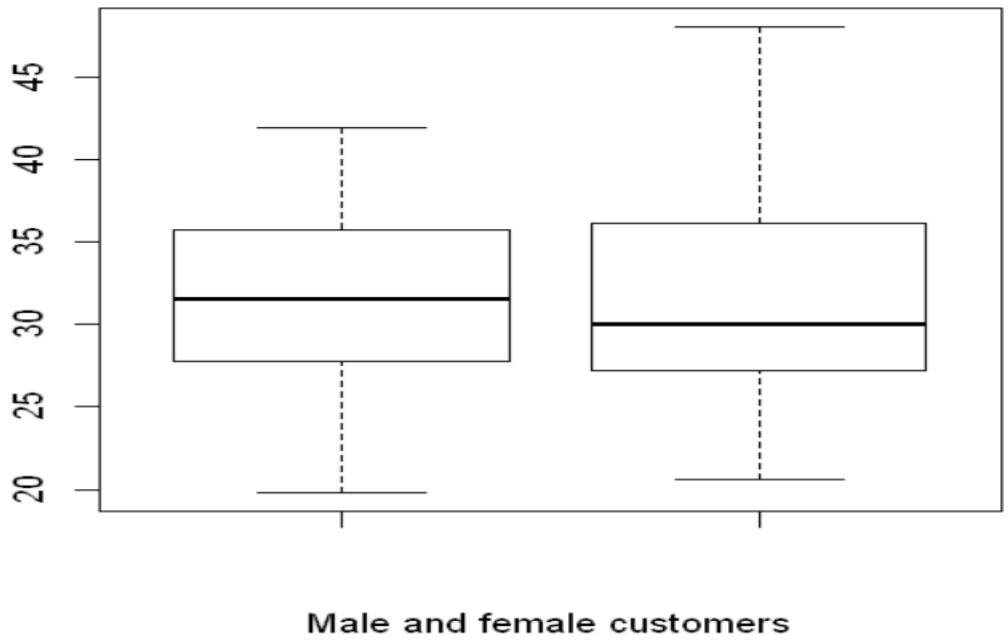


Fig 1.3 Box plot of BMI values of male and female customers

The box plot shows that the body mass index(BMI) of male and female customers are normally distributed.

Observation

This is two-tailed hypothesis test. We performed random sampling to select sample data from male and female customers to perform hypothesis testing. Since p-value (0.6892) is greater than significance level(0.05). we don't have enough evidence to prove that male and female customers have different Body mass index variances. So, we fail to reject Null Hypothesis. The Confidence interval range for ratio of variances between two groups is from 0.886 to 1.20 and ratio of variances of male and female customers is 1.0315. These findings show that the irrespective of whether the customers are male or female variances of body mass index is remains same.

Conclusion

In sum, we studied and learnt the importance of hypothesis testing which is utilized to take care of numerous business-related issues and to make better strategic decision making for business in various areas of financial market, medical field, and in politics and elections to make prediction of the winner. We analyzed the data and performed hypothesis testing on different dataset. We used inbuilt functions to perform the testing. Based on p value and significance level, we reject or fail to reject null hypothesis.

Reference:

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Scholkopf, and A. Smola, A Kernel Two-Sample Test, Journal of Machine Learning Research, 13 (2012), 723-773

Bluman Allan. (2017). Elemenatry Statistics A Step by Step Approach. New York, NY: McGraw-Hill