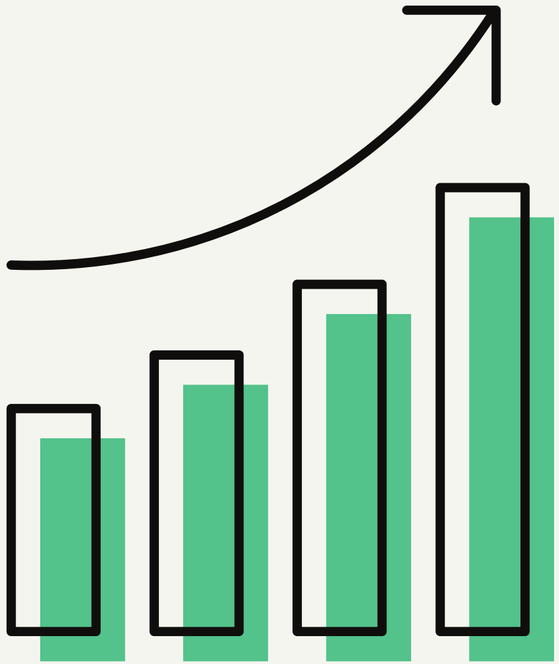
APPLICATION OF CHI-SQUARE DISTRIBUTION



Prepared by -

RISHI R

SUHASI GOHIL

MOHITA AHUJA

Guided by -

DR.SANTOSHA CD

Table of Contents

Table of Contents

Introduction

Part 01

Chi-square distribution

Karl Pearson in 1990, developed a test known as $\chi 2$-test to test if the deviation between observation (experiment) and theory may be attributed to chance (fluctuations of sampling) or if it is really due to the inadequacy of the theory to fit the observed data.

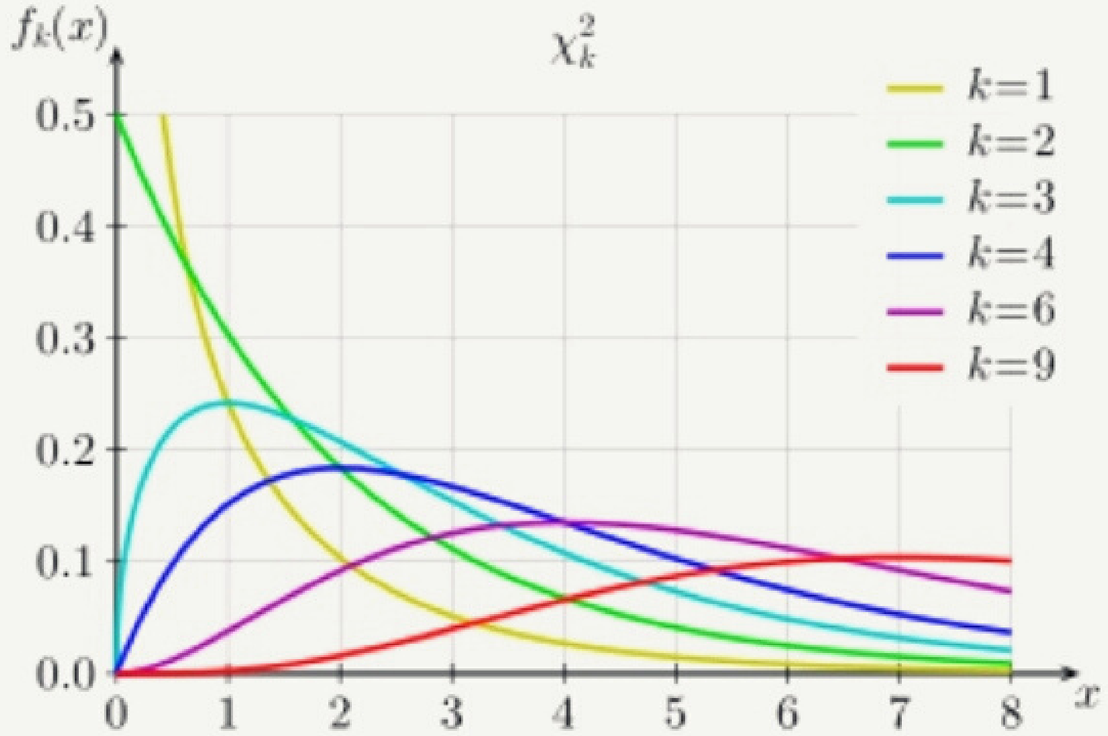
The probability density function (pdf) of the chi-square distribution with k degrees of freedom is,

$$f(x) = \begin{cases} \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{\frac{-x}{2}} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where k is a positive integer denoting degrees of freedom

Γ = Gamma function

The probability density curve can be given as,



Characteristics of chi-square distribution with k degrees of freedom

- Cumulative distribution function: $\frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})}$
- Moment generating function: $(1- 2t)^{\frac{-k}{2}}$, $t < 1/2$
- Mean: k
- Median: $k\left(1 - \frac{2}{9k}\right)^3$
- Mode: max (k – 2, 0)
- Variance: 2k
- Skewness: $\frac{\text{Mean-Mode}}{\text{S.D}} = \frac{k-(k-2)}{\sqrt{2k}} = \sqrt{\frac{2}{k}}$
- Kurtosis: $\frac{12}{k}$

POINTS TO REMEMBER

Applications of chi-square distribution

- Chi-square test for goodness of fit

Chi-Square goodness of fit test is a non-parametric test that is used to find out how the observed value of a given phenomena is significantly different from the expected value. It establishes the discrepancy between the observed values and the expected values from a normal distribution case.

- Chi-square test for independence

The Chi-square test of independence determines whether there is a statistically significant association between two categorical variables. This test utilizes a contingency table to analyze the data. A contingency table is an arrangement in which data is classified according to two categorical variables. Each cell reflects the total count of cases for a specific pair of categories.

Assumptions for applying chi-square test

- The data are randomly drawn from a population
- The values in the cells are considered adequate when expected counts are not < 5 and there are no cells with zero count
- The sample size is sufficiently large. The application of the chi-square test to a smaller sample could lead to type II error (i.e. accepting the null hypothesis when it is actually false).
- The variables under consideration must be mutually exclusive (independent). It means that each variable must only be counted once in a particular category and should not be allowed to appear in other category.

Test procedure for chi-square test for goodness of fit

- **Null and Alternative hypothesis**

H_o : Theoretical frequency distribution is a good fit to the observed frequency distribution

Vs

H_1 : Theoretical frequency distribution is not a good fit to the observed frequency distribution

- **Test statistic**

The test statistic is given by, $\chi^2 = \sum\left[\frac{(f_0-f_e)^2}{f_e}\right] \sim \chi^2$ distribution with (n - 1) degrees of freedom

in which fo = experimented frequency of the observed facts

fe = expected frequency of occurrence on null hypothesis.

- **Critical region and Conclusion**

At α level of significance, the critical value is $\chi^2_{\alpha}(n-1)$

Reject the null hypothesis if, $\chi^2_{cal} > \chi^2_{\alpha}(n-1)$

Or reject the null hypothesis if, $P\left(\chi^2 > \chi_{cal}^{\ 2}\right) \leq \alpha$

Test procedure for chi-square test for independence

- **Null and Alternative hypothesis**

H_o : The two attributes are independent of each other

Vs

H_1 : The two attributes are dependent on each other

- **Test statistic**

The test statistic is given by, $\chi^2 = \sum\left[\frac{(f_0-f_e)^2}{f_e}\right] \sim \chi^2$ distribution with (r-1)(c-1) degrees of freedom

where r = the number of rows and c = number of columns

also, fo = experimented frequency of the observed facts

fe = expected frequency of occurrence on null hypothesis.

- **Critical Region and Conclusion**

At α level of significance, the critical value is $\chi^2_{\alpha}(r-1)(c-1)$

Reject the null hypothesis if, $\chi^2_{cal} > \chi^2_{\alpha}(r-1)(c-1)$

Or reject the null hypothesis if, $P\left(\chi^2 > \chi^2_{cal}\right) \leq \alpha$

To illustrate the above mentioned applications,
we undertook a team project.

PROJECT TITLE

TECH IN LOCK-DOWN

Project details

Part 02

Overview of our project

- Objective
  - " To capture your mode of entertainment via technology in the wake of covid'19 outbreak "
- Domain

Overview of our project

- Methodology

Data collection method - Primary data collected through an online questionnaire as Network sampling technique

Test used - Chi-square test for goodness of fit
Chi-square test for independence

Sample description - Target group, respondents' age ranging from 15 - 42
300+ responses from various parts of India

Project Timeline

01

Phase 01
Deciding on the broader question

02

Phase 02
Working out the confined domains and sub-domains

03

Phase 03
Having clarity on the objective of the study

04

Phase 04
Constructing the questionnaire and collecting data

05

Phase 05
Presentation preparation and Analysis of data.

Analysis

Part 03

Chi-square test for independence

- Variables taken -

Type of OTT platform (i.e. Netflix, Amazon prime etc) and Occupation of the respondent

- Null and Alternative hypothesis

H_o : The choice of OTT platform and occupation are independent of each other

Vs

H_1 : The choice of OTT platform and occupation are dependent on each other

- R programme -

```
Tech_OTT<-read.csv("Tech_OTT.csv")
Tech_OTT
head(Tech_OTT)
str(Tech_OTT)
library(gmodels)
Tech_Table1<-CrossTable(Tech_OTT$OTT.Platforms,Tech_OTT$Occupation,chisq
= TRUE,expected =TRUE,prop.r = FALSE,prop.c = FALSE,prop.t = F,prop.chisq = F )
Tech_Count<-Tech_Table1$t
Tech_Count
barplot(Tech_Count,beside
=T,col=c("Turquoise","Cadetblue","Cornflowerblue","Darkblue"),ylab="Frequency",xla
b="Occupation",legend=T)
```

- R Output obtained –

Tech_OTT$OTT.Platforms	Tech_OTT$Occupation Employed	Self Employed	Student	Row Total
Amazon Prime Video	14 8.667	8 6.167	24 31.167	46
Hotstar	6 4.899	4 3.486	16 17.616	26
Netflix	18 27.130	14 19.304	112 97.565	144
Youtube	14 11.304	11 8.043	35 40.652	60
Column Total	52	37	187	276

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 15.12815 d.f. = 6 p = 0.01928306

- Contingency table –

OTT Platforms	Frequency	Occupation			Total	$\chi^2_{(6)}$	p-value
		Employed	Self Employed	Student			
Netflix	Observed Frequency	18 (34.62%)	14 (37.84%)	112 (59.89%)	144	15.13	0.019283
	Expected Frequency	27.13	19.304	97.565			
Amazon Prime Video	Observed Frequency	14 (26.92%)	08 (21.62%)	24 (12.83%)	46		
	Expected Frequency	8.67	6.167	31.167			
YouTube	Observed Frequency	14 (26.92%)	11 (29.73%)	35 (18.72%)	60		
	Expected Frequency	11.304	8.043	40.652			
Hotstar	Observed Frequency	06 (11.54%)	04 (10.81%)	16 (8.56%)	26		
	Expected Frequency	4.899	3.486	17.616			
	Total	52	37	187	276		

Table 1.1: Association between OTT platform and occupation of the respondent

- Critical value and conclusion –

Under H_o , the obtained χ^2 value = 15.13 at 5% level of significance

Here the critical value is, $\chi^2_{\alpha}(r-1)(c-1) = \chi^2_{0.05(6)} = 12.592$

Since, χ^2 value > $\chi^2_{\alpha}(r-1)(c-1)$ we reject the null hypothesis

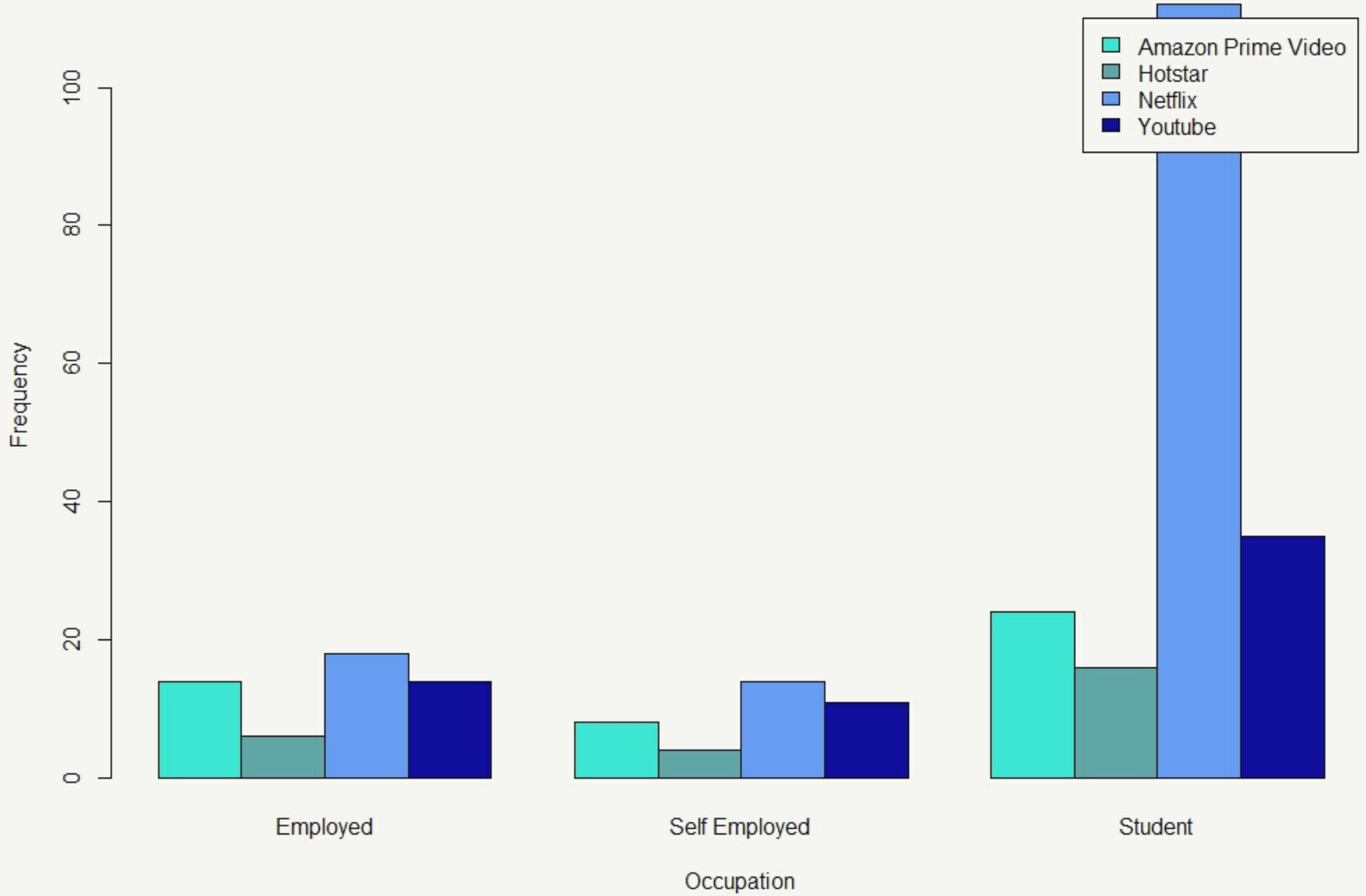
Also,

The obtained p value = 0.019283 at 5% level of significance

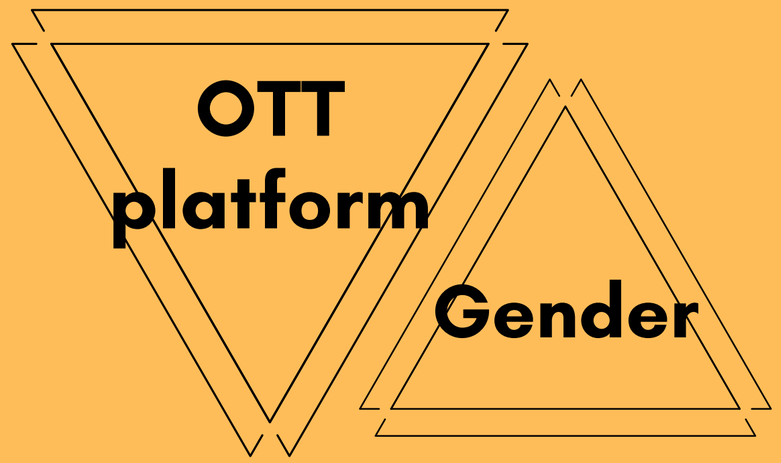
Since, p value < 0.05, we reject the null hypothesis

i.e. the choice of OTT platform and occupation of the respondent are dependent on each other

- Clustered bar chart –



- Test with other socio-demographical variables -

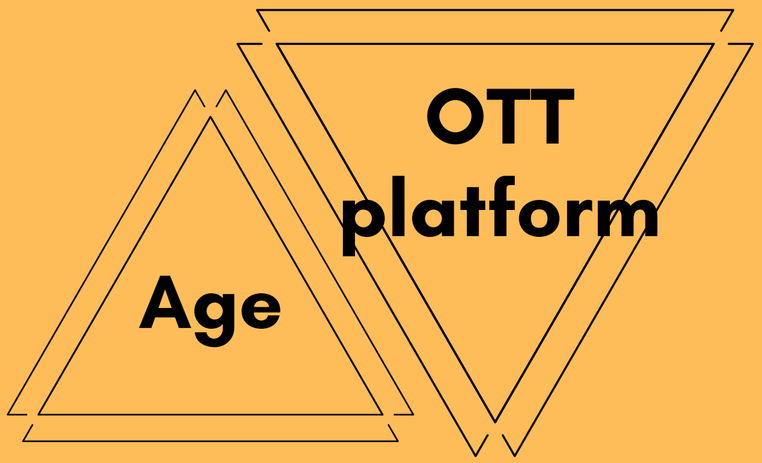


Associated

P VALUE = 0.039

Associated

P VALUE = 0.003



Chi-square test for independence

- Variables taken -

PUBG players and age of the respondent

- Null and Alternative hypothesis

H_o : The choice to play PUBG and age are independent of each other

Vs

H_1 : The choice to play PUBG and age are dependent on each other

- R programme -

```
Tech_Gaming<-read.csv("Tech_Gaming.csv")
Tech_Gaming
table(Tech_Gaming$Call.Of.Duty.Players)
head(Tech_Gaming)
str(Tech_Gaming)
library(gmodels)
Tech_Table3<-CrossTable(Tech_Gaming$PUBG.Players,Tech_Gaming$Age,chisq
= TRUE,expected =TRUE,prop.r = FALSE,prop.c = FALSE,prop.t = F,prop.chisq = F )
Tech_Table3
Tech_Count<-Tech_Table3$t
Tech_Count
barplot(Tech_Count,beside=T,col=c("Maroon","Yellow"),ylab="Frequency",xlab="Age",legen
d=T,width = 15, xlim =c(0,120))
```

- R Output obtained –

```
Total Observations in Table:  122


                             | Tech_Gaming$Age
Tech_Gaming$PUBG.Players     | 25 or more |    Under 25 |  Row Total |
-----------------------------|------------|-------------|------------|
                  Non-player |         26 |          52 |         78 |
                             |     19.180 |      58.820 |            |
-----------------------------|------------|-------------|------------|
                     Players |          4 |          40 |         44 |
                             |     10.820 |      33.180 |            |
-----------------------------|------------|-------------|------------|
               Column Total |         30 |          92 |        122 |
-----------------------------|------------|-------------|------------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  8.915591     d.f. =  1     p =  0.002827463
```

- Contingency table -

PUBG	Frequency	Age Group		Total	$\chi^2_{(1)}$	p-value
		Under 25	25 or more			
Players	Observed Frequency	40 (43.48%)	4 (13.33%)	44	8.92	0.002827
	Expected Frequency	33.18	10.82			
Non-Players	Observed Frequency	52 (56.52%)	26 (86.67%)	78		
	Expected Frequency	58.82	19.18			
	Total	92	30	122		

Table 1.2: Association between PUBG players and age of the respondent

- Critical value and conclusion –

Under H_o , the obtained χ^2 value = 8.92 at 5% level of significance

Here the critical value is, $\chi^2_{\alpha}(r-1)(c-1) = \chi^2_{0.05(1)}$ = 3.841

Since, χ^2 value > $\chi^2_{\alpha}(r-1)(c-1)$ we reject the null hypothesis

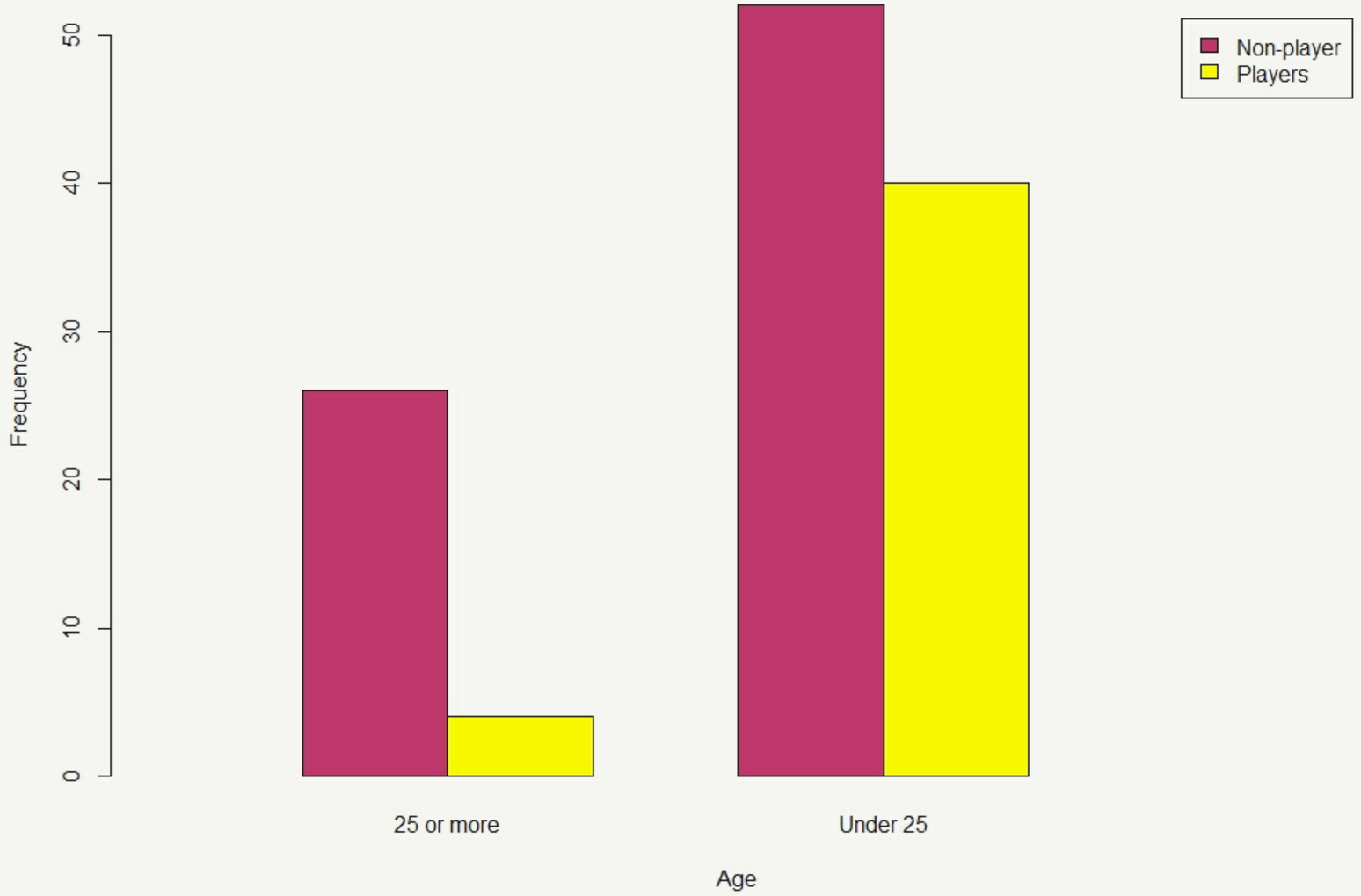
Also,

The obtained p value = 0.002827 at 5% level of significance

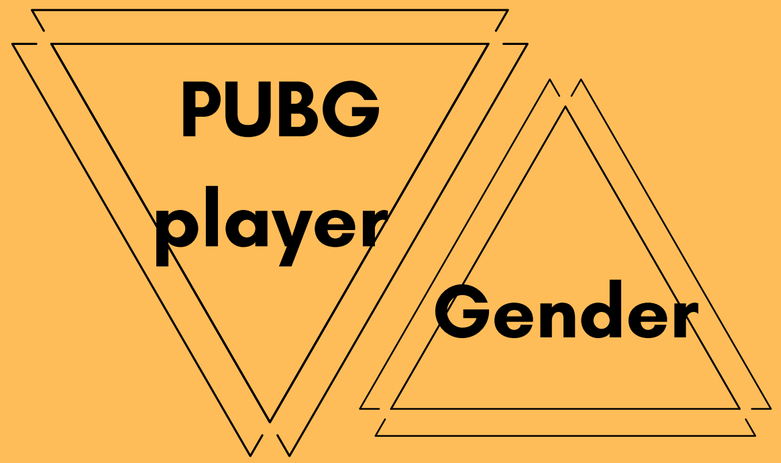
Since, p value < 0.05, we reject the null hypothesis

i.e. the choice to play PUBG and age of the respondent are dependent on each other

- Clustered bar chart –



- Test with other socio-demographical variables -

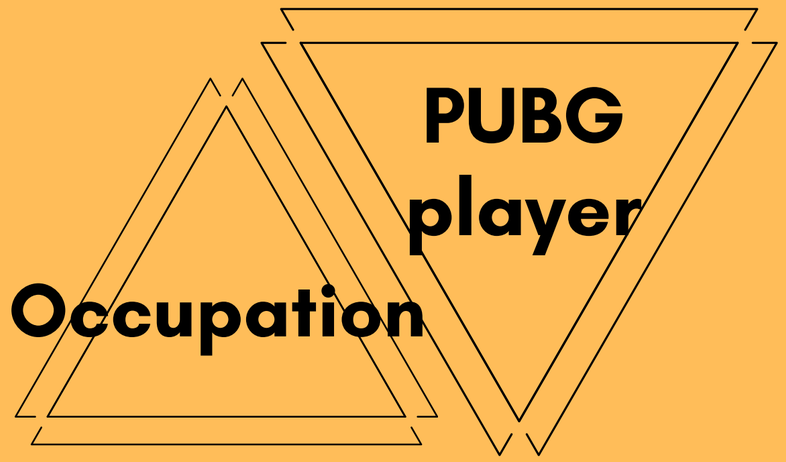


Not associated

P VALUE = 0.979

Associated

P VALUE = 0.020



Chi-square test for independence

- Variables taken -

  Whatsapp users and occupation of the respondent

- Null and Alternative hypothesis

H_o : The usage of whatsapp and occupation are independent of each other

Vs

H_1 : The usage of whatsapp and occupation are dependent on each other

- R programme -

```
Tech_SocialMedia<-read.csv("Tech_SocialMedia.csv")
Tech_SocialMedia
head(Tech_SocialMedia)
str(Tech_SocialMedia)
library(gmodels)
Tech_Table2<-
CrossTable(Tech_SocialMedia$Whatsapp.User,Tech_SocialMedia$Occupation,chisq
= TRUE,expected =TRUE,prop.r = FALSE,prop.c = FALSE,prop.t = F,prop.chisq = F )
Tech_Table2
```

- R Output obtained –

```
|-------------------------|
|                       N |
|              Expected N |
|-------------------------|


Total Observations in Table:  280


                                | Tech_SocialMedia$Occupation
Tech_SocialMedia$Whatsapp.User |      Employed | Self Employed |       Student |     Row Total |
-------------------------------|---------------|---------------|---------------|---------------|
                     Non-user |             4 |            10 |            75 |            89 |
                              |        15.893 |        13.668 |        59.439 |               |
-------------------------------|---------------|---------------|---------------|---------------|
                         User |            46 |            33 |           112 |           191 |
                              |        34.107 |        29.332 |       127.561 |               |
-------------------------------|---------------|---------------|---------------|---------------|
                 Column Total |            50 |            43 |           187 |           280 |
-------------------------------|---------------|---------------|---------------|---------------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  20.46134     d.f. =  2     p =  3.604757e-05
```

- Contingency table -

Whatsapp	Frequency	Occupation			Total	$\chi^2_{(2)}$	p-value
		Self Employed	Employed	Student			
Users	Observed Frequency	33 (76.74%)	46 (92%)	112 (59.89%)	191	20.46	0.001
	Expected Frequency	29.33	34.11	127.56			
Non-users	Observed Frequency	10 (23.26%)	04 (8%)	75 (40.11%)	89		
	Expected Frequency	13.67	15.89	59.44			
	Total	43	50	187	280		

Table 1.3: Association between Whatsapp users and occupation of the respondent

- Critical value and conclusion –

Under H_o , the obtained χ^2 value = 20.46 at 5% level of significance

Here the critical value is, $\chi^2_{\alpha}(r-1)(c-1) = \chi^2_{0.05(2)} = 5.991$

Since, χ^2 value > $\chi^2_{\alpha}(r-1)(c-1)$ we reject the null hypothesis

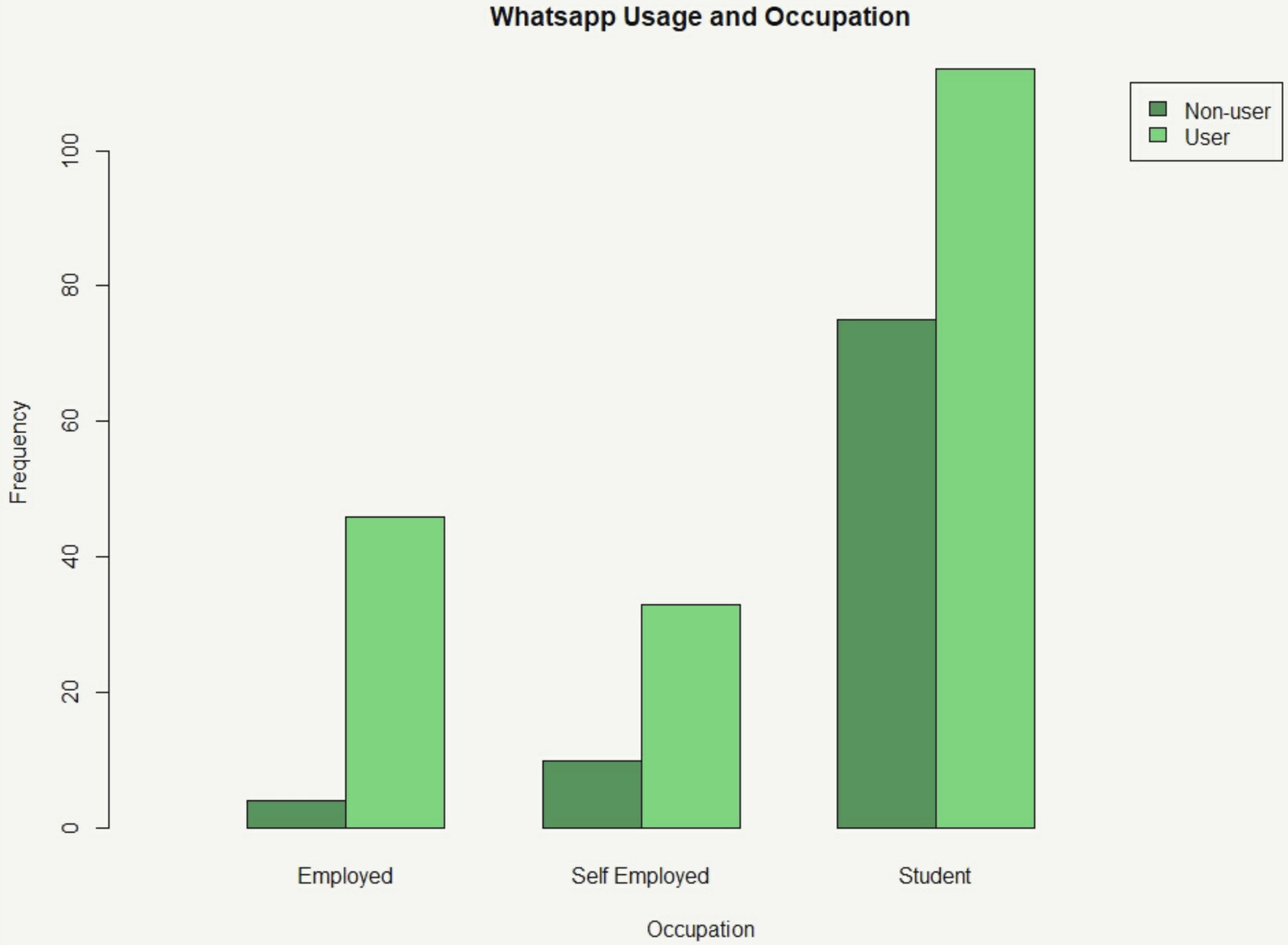
Also,

The obtained p value = 0.001 at 5% level of significance

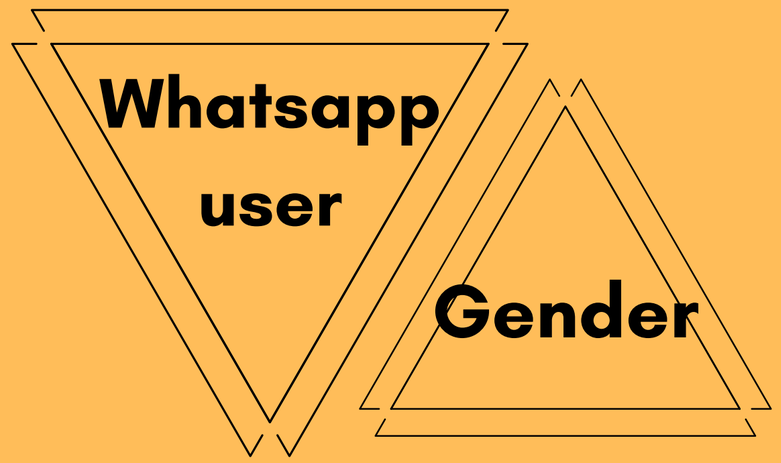
Since, p value < 0.05, we reject the null hypothesis

i.e. The choice to use whatsapp and occupation of the respondent are dependent on each other

- Clustered bar chart –



- Test with other socio-demographical variables -

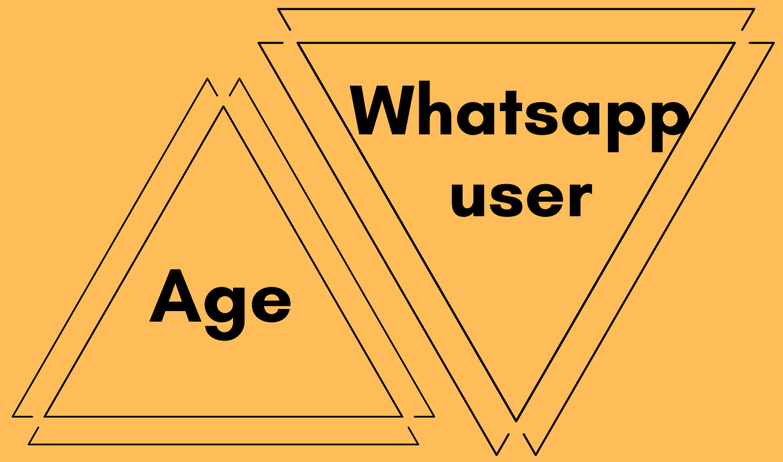


Not associated

P VALUE = 0.889

Associated

P VALUE = 0.001



Chi-square test for goodness of fit

- Variables taken –

  Various online applications used for video calling

- Null and Alternative hypothesis

H_0 : The usage of video call application is equally distributed in the data

Vs

H_1 : The usage of video call application is not equally distributed in the data

- R programme -

```
Vcall<-c(68,69,71,72)
names(Vcall)<-c("Google Meet","Google Duo","Zoom","Social Media Apps")
Vcall
Vcall/sum(Vcall)
Probability<-c(0.25,0.25,0.25,0.25)
#Ho:Proportion of usage of applications for video calls is 0.25
#Ha:Proportion of usage of applications for video calls is not same
chisq.test(Vcall,p=Probability)
```

- R output obtained

```
> #Test of Goodness of Fit
> Vcall<-c(68,69,71,72)
> names(Vcall)<-c("Google Meet","Google Duo","Zoom","Social Media Apps")
> Vcall
      Google Meet        Google Duo              Zoom Social Media Apps
               68                69                71                72
> Vcall/sum(Vcall)
      Google Meet        Google Duo              Zoom Social Media Apps
        0.2428571         0.2464286         0.2535714         0.2571429
> Probability<-c(0.25,0.25,0.25,0.25)
> #Ho:Proportion of usage of applications for video calls is 0.25
> #Ha:Proportion of usage of applications for video calls is not same
> chisq.test(Vcall,p=Probability)

        Chi-squared test for given probabilities

data:  Vcall
X-squared = 0.14286, df = 3, p-value = 0.9862
```

- Frequency table -

Video call Apps	Observed Frequency	Expected Frequency	$\chi^2_{(3)}$	p-value
Google Duo	69(24.29%)	70	0.14	0.9862
Google Meet	68 (24.64%)	70		
Zoom	71 (25.36%)	70		
Social Media Apps	72 (25.71%)	70		
Total	280			

Table 2.0: Frequency distribution of usage of video call applications

- Critical value and conclusion –

Under H_o , the obtained χ^2 value = 0.14 at 5% level of significance

Here the critical value is, $\chi^2_\alpha(n-1) = \chi^2_{0.05(3)}$ = 7.815

Since, χ^2 value < $\chi^2_\alpha(n-1)$ we accept the null hypothesis

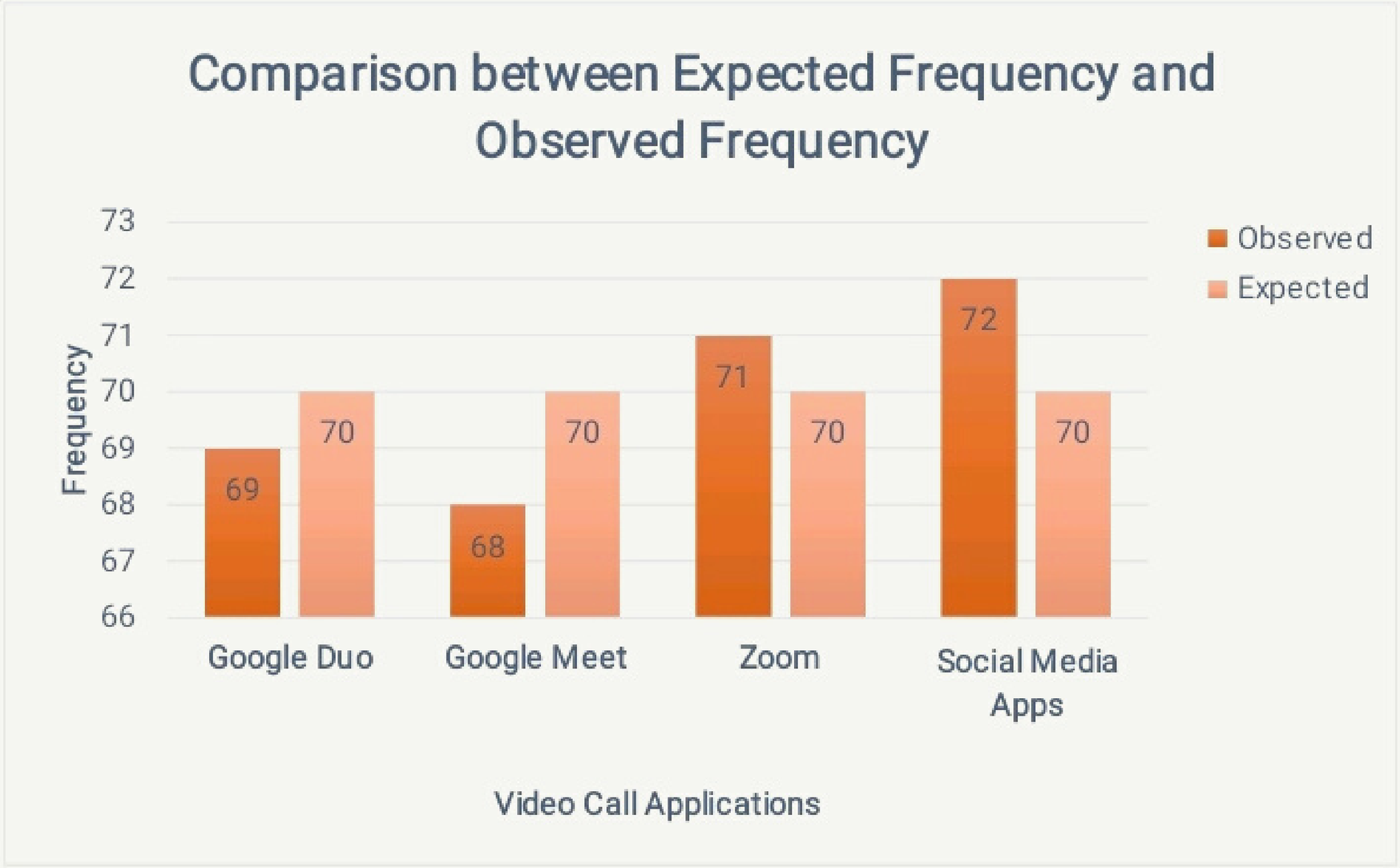
Also,

The obtained p value = 0.9862 at 5% level of significance

Since, p value > 0.05, we accept the null hypothesis

i.e The usage of video call application is equally distributed in the data

- Bar chart –


Comparison between Expected Frequency and Observed Frequency
Frequency
73
72
71
70
69
68
67
66
69
70
68
70
71
70
72
70
Observed
Expected
Google Duo
Google Meet
Zoom
Social Media Apps
Video Call Applications

Limitations and challenges

- The target for the number of responses being 400, was not achieved
- Analysis was entirely based on the number and kind of responses we receive
- As the project was time bound, we faced time constraints
- Questionnaire was accessible to only those who had internet connection

Scope of further study

On account of the pandemic and also it's consequences, technology in it's various forms has influenced the lives of human beings of all age groups.

It is thus evident that technology has played a vital role not only in entertainment but also in the health regime and to enhance our abilities in various fields of interest as well.

Acknowledgement

We would like to express our special thanks of gratitude to
Prof. Kavya S and Prof. Preeti Ravikiran for their able guidance and support

We would also like to extend our gratitude to our guide
Dr. Santosha C D and Dr. Namrata P for their constant advice and encouragement

- The following books were referred for the completion of this project

  - Fundamentals of Statistics by S.C Gupta

THANK YOU FOR WATCHING!