Master of Science in Data Analytics

Dataset of World University Ranking

Submitted by

Mohanraj Jayakumar - 10577457

Team Members

Killian Tol – 10581177

Vinay Ramasare Kurmi – 10576078

Supervised by Dr.Shahram Azizi

ACKNOWLEDGEMENT

I would like to take this opportunity to thank my professor Dr.Shahram Azizi for sharing his knowledge and ideas on the subject - B9DA101 Statistics for Data Analytics. From attending his lectures through zoom and taking part in class tutorials I have learned a vast range of statistical concepts that have helped achieve my goal in completing this project. His teaching skills and the way he portrays the module have helped me greatly in my understanding of how statistics plays a vital role in the modern world of data analytics.

Table of Contents

Introduction

The dataset we have chosen describes the rankings of the world's best universities from around the world including USA, Canada, United Kingdom, Japan, Switzerland, Israel, France, South Korea, Russia and many more from the years 2012-2015. We have simplified the data to the top 50 rankings of these countries since this satisfies our requirements. There can be many complications when judging the ranks for universities. It is a difficult, political and controversial practice. Many countries disagree with the terms and conditions for choosing the best universities. Things like research power are chosen over quality of tuition in some rankings and some institutes are chosen over others based on the fact that they teach in English rather than the quality of teaching itself. Our dataset is specifically based on only one statistical data report from the website "Centre for World University Ranking".

For our project we will be using the Crisp DM methodology (Wikipedia, 2020) for various phases within processing and reporting of this dataset. This method segregates the implementation into 6 phases which pragmatically explains the process of implementation of a data project.

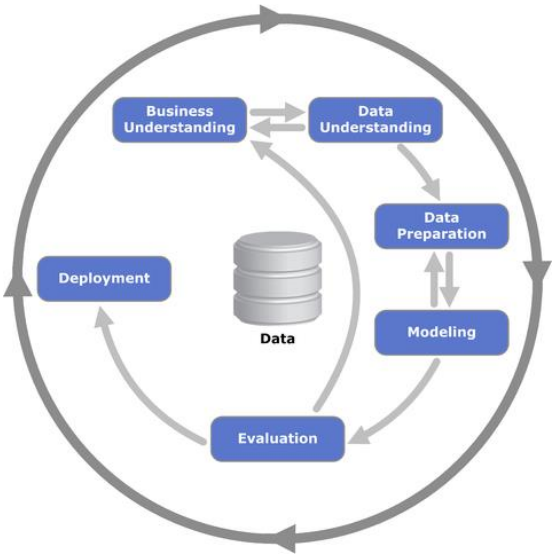


Fig 1: Different phases of CRISP DM

Business Understanding

The dataset we are using can be useful in various aspects of education, to both universities and students seeking admission to higher education. Some of the benefits of our analysis have been listed below:

- It provides an overview on which countries have the most efficient universities while describing the output of the education provided at the institute by various measures such as employed alumni's, publications and patents.
- It tracks the progress of institutes through its ranking when compared to others in the world or even within the country from 2012 to 2015.
- It provides detailed information on how alumni's employment (with help of campus recruitment) directly affects the ranking of the institute.
- It displays how influence over students and industry might affect the rankings of the institute.

Data Understanding

For our project we simplified the columns and rows making it slightly different to the original dataset which was taken from the website "Kaggle". Kaggle is an open source website where datasets are readily available and easy to work with on many platforms. For our project we used a csv file. This made it easier to simplify/transform the data in accordance with our needs and it could be implemented in any type of programming environment. After cleaning and simplifying the dataset we ended up with 200 rows and 13 columns of useful data. Our dataset contains variables of type categorical, double and integer for example. Columns/variables used for our analysis are:

Table 1:

world_rank	Defines the ranking for the institute among all other institutes in the world.
institution	Defines the institute name.
country	Defines the country in which the institute belongs.
national_rank	Defines the ranking of the institute among other institutes within the country.
quality_of_education	Defines the quality of education provided within the institute.
alumni_employment	Defines the number of alumni's employed with help of campus placement.
quality_of_faculty	Defines the quality of the faculty based on their experience and educational background.
publications	Defines the number of reports published by the institute.
influence	Defines the influence of the institute on industry and students.
citations	Defines the number of citations by the institute.
patents	Defines the number of patents filed by the institute.
score	Defines the score of the institute based on evaluation by the international panel.
year	Defines the year of grading of the institute.

Data Preparation

The international ranking system considers all of the above benchmarks when deciding the best institutes and universities around the world. Before considering a university for ranking it should have all the characteristics mentioned in our dataset since these are important fundamentals for being included in the world university rankings. After ensuring all universities held these characteristics we began with the first step of analysis of the dataset. To process the dataset, we used R Studio and the R programming language which is an industry standard tool for machine learning and data science projects.

Importing the required libraries

The first and most critical step was to identify and import the packages necessary for our analysis. We used two libraries for analysis of this project: tidyverse for cleaning the dataset and ggplot2 for graphical representation of data. Below is an illustration on how we imported our libraries:

```
library(tidyverse)
library(ggplot2)
```

Reading the csv data files

We then imported the csv data file into R Studio to be processed using various methodologies. To import the data we used the read.csv method as seen below:

```
wrk_dir = "I:/DBS/Statistics for Data Analytics/CA1/world_university_ranking/data"
setwd(wrk_dir)
getwd()
data <- read.csv("cwurData.csv")
head(data)
```

Fig 2:

	world_rank <int>	institution <chr>	country <chr>	national_rank <int>	quality_of_education <int>	alumni_employment <int>	quality_of_faculty <int>	publications <int>	influence
1	1	Harvard University	USA	1	7	9	1	1	1
2	2	Massachusetts Institute of Technology	USA	2	9	17	3	12	4
3	3	Stanford University	USA	3	17	11	5	4	2
4	4	University of Cambridge	United Kingdom	1	10	24	4	16	16
5	5	California Institute of Technology	USA	4	2	29	7	37	22
6	6	Princeton University	USA	5	8	14	2	53	33

6 rows | 1-10 of 14 columns

We then had a quick glimpse into our unfiltered data which we acquired from Kaggle.

```
nrow(data)
glimpse(data)
colSums(is.na(data))
```

Fig 3:

```
[1] 2200
Rows: 2,200
Columns: 14
$ world_rank           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
$ institution          <chr> "Harvard University", "Massachusetts Institute of T...
$ country              <chr> "USA", "USA", "USA", "United Kingdom", "USA", "USA"...
$ national_rank        <int> 1, 2, 3, 1, 4, 5, 2, 6, 7, 8, 9, 10, 11, 1, 12, 1, ...
$ quality_of_education <int> 7, 9, 17, 10, 2, 8, 13, 14, 23, 16, 15, 21, 31, 32,...
$ alumni_employment    <int> 9, 17, 11, 24, 29, 14, 28, 31, 21, 52, 26, 42, 16, ...
$ quality_of_faculty   <int> 1, 3, 5, 4, 7, 2, 9, 12, 10, 6, 8, 14, 24, 31, 20, ...
$ publications         <int> 1, 12, 4, 16, 37, 53, 15, 14, 13, 6, 34, 22, 9, 8, ...
$ influence            <int> 1, 4, 2, 16, 22, 33, 13, 6, 12, 5, 20, 21, 10, 19, ...
$ citations            <int> 1, 4, 2, 11, 22, 26, 19, 15, 14, 3, 28, 16, 8, 23, ...
$ broad_impact         <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ patents              <int> 5, 1, 15, 50, 18, 101, 26, 66, 5, 16, 101, 10, 9, 3...
$ score                <dbl> 100.00, 91.67, 89.50, 86.17, 85.21, 82.50, 82.34, 7...
$ year                 <int> 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 201...
          world_rank           institution               country         national_rank 
                   0                     0                     0                     0 
quality_of_education    alumni_employment   quality_of_faculty          publications 
                   0                     0                     0                     0 
           influence             citations         broad_impact               patents 
                   0                     0                   200                     0 
               score                  year 
                   0                     0 
```

In the above output we see a brief description of our original dataset. We see the number of rows i.e. 2200 for our dataset using the nrow() method. The type of data columns has also been described and some initial values of those corresponding rows can be seen using the glimpse function from dplyr package. At last, we check for columns with null values using the is.na method.

Filtering the Data

Since our original dataset included a lot of irrelevant data, (which did not help us fulfil our business requirements) we had to filter the data according to the requirements.

```
#Filter the data
fil_data <- filter(data,data$world_rank <= 50)

#Structuring the data columns
fil_data$country <- as.factor(fil_data$country)
fil_data$institution <- as.factor(fil_data$institution)
fil_data$year <- as.factor(fil_data$year)

#Overview of filtered data
colSums(is.na(fil_data))
glimpse(fil_data)
head(fil_data)
summary(fil_data)
nrow(fil_data)
```

Fig 4:

```
> nrow(fil_data)
[1] 200
> summary(fil_data)
   world_rank                                institution           country    national_rank   quality_of_education
 Min.   : 1.0   California Institute of Technology:  4   USA           :135   Min.   : 1.00   Min.   :  1.00
 1st Qu.:13.0   Columbia University               :  4   United Kingdom: 17   1st Qu.: 2.00   1st Qu.: 15.00
 Median :25.5   Cornell University                :  4   Japan         : 16   Median : 9.00   Median : 34.00
 Mean   :25.5   Duke University                   :  4   Canada        :  8   Mean   :12.44   Mean   : 61.55
 3rd Qu.:38.0   Harvard University               :  4   France        :  8   3rd Qu.:22.00   3rd Qu.: 91.00
 Max.   :50.0   Hebrew University of Jerusalem   :  4   Israel        :  8   Max.   :37.00   Max.   :367.00
                (Other)                          :176   (Other)       :  8
 alumni_employment quality_of_faculty  publications      influence        citations         patents
 Min.   :  1.00    Min.   :  1.00     Min.   :  1.00   Min.   :  1.00   Min.   :  1.00   Min.   :  1.00
 1st Qu.: 17.75    1st Qu.: 13.00     1st Qu.: 13.00   1st Qu.: 13.00   1st Qu.: 13.00   1st Qu.: 15.75
 Median : 44.50    Median : 26.00     Median : 27.00   Median : 26.00   Median : 26.00   Median : 40.00
 Mean   : 85.07    Mean   : 43.95     Mean   : 54.12   Mean   : 45.46   Mean   : 54.94   Mean   : 65.24
 3rd Qu.:101.00    3rd Qu.: 52.50     3rd Qu.: 52.00   3rd Qu.: 53.00   3rd Qu.: 52.50   3rd Qu.: 80.25
 Max.   :567.00    Max.   :218.00     Max.   :406.00   Max.   :389.00   Max.   :645.00   Max.   :871.00

     score          year
 Min.   : 49.84   2012:50
 1st Qu.: 56.13   2013:50
 Median : 61.09   2014:50
 Mean   : 66.92   2015:50
 3rd Qu.: 77.33
 Max.   :100.00

> nrow(fil_data)
[1] 200
```

After filtering the original dataset according to our needs, we were left with 200 rows and 13 columns.

Exploratory Analysis / Graphical Descriptions

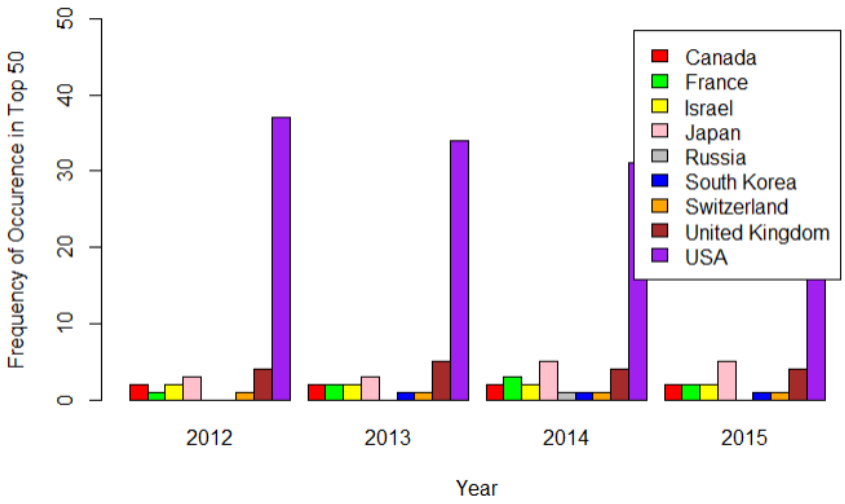
In this section we plotted some graphs and tried to understand the data we acquired.

Bar Plot

Code:

```
# bar graph for year vs frequency of institutions from each country
qualitytab <-table(fil_data$country,fil_data$year)
qualitytab
barplot(qualitytab,col=c("red","green","yellow","pink","grey","blue","orange","brown","purple"),
        ylim=c(0,50),
        xlab = "Year",
        ylab = "Frequency of Occurence in Top 50",
        legend = rownames(qualitytab),
        beside = TRUE)
```

Plot:



The figure above represents the statistical reports of frequency of institute from all countries that occurred in top 50 ranks. The X axis on the graph represents the years from 2012-2015 and the y axis represents the frequency of occurrence in the top 50 colleges. This bar plot helps us understand the amount of institutes from a country and overall efforts of a nation to build its educational system to represent at global level. This bar graph shows that institutes from the USA hold the prominent frequency of occurrence in all the years between 2012 and 2015. USA is in a class of its own compared to other countries shown in the chart. Japan holds the next place in the bar graph and Russia holds the lowest frequency. Other countries like United Kingdom, Switzerland, South Korea, Israel, France and Canada fall well below the frequency of USA but lay above Russia in the ranking as seen above.

Quantiles

Code:

```
# Quantile of number of patents
Q=quantile(fil_data$patents)
Q
```

Quantile Ranges:

```
   0%    25%    50%    75%   100%
 1.00  15.75  40.00  80.25 871.00
```

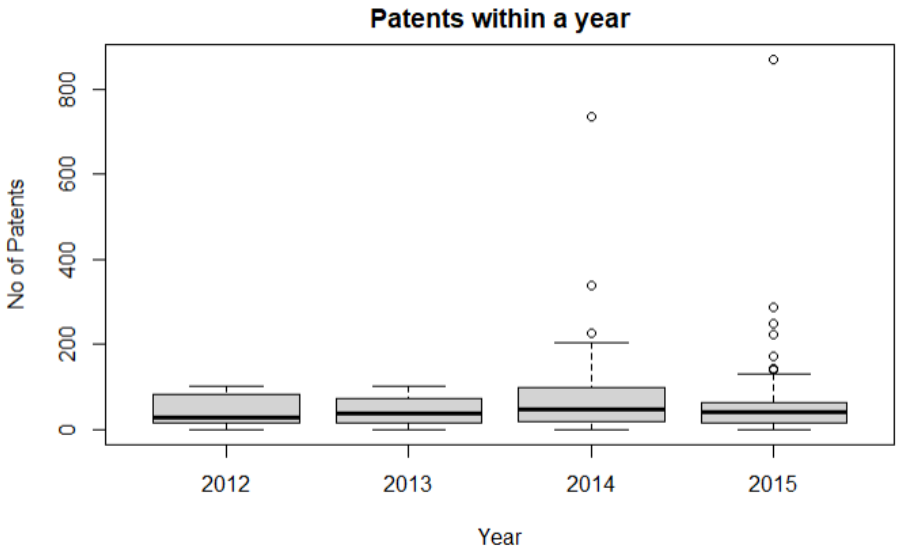
This returns the quantile distribution for number of patents registered by various institutes.

Box Plot

Code:

```
# Boxplot for patents vs year for outlier detection
boxplot(patents~year,data=fil_data,  main="Patents within a year",xlab="Year", ylab="No of Patents")
```

Plot:



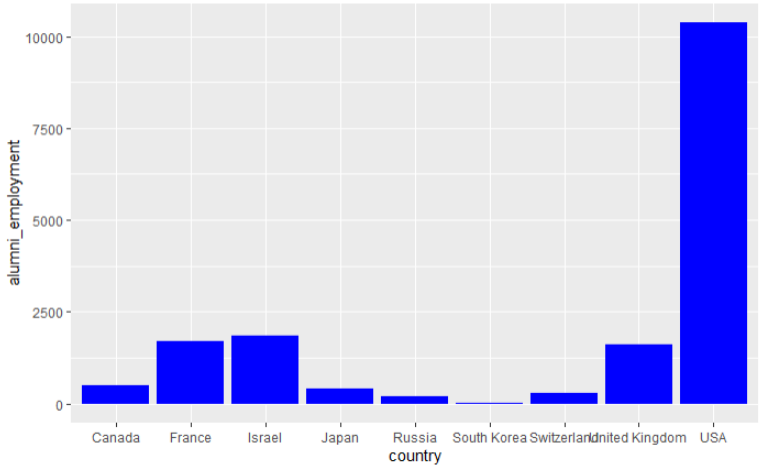
We used the boxplot technique for outlier detection for number of patents published each year. For years 2012 and 2013, most of the universities were producing number of patents that were within the upper and lower whisker of the data plot. While in years 2014 and 2015, the number of patents produced by institutes was overwhelmed which resulted in outliers beyond the upper and lower whisker of the plot.

Pareto Chart

Code:

```
#Pareto chart for country vs alumni employment
ggplot(fil_data, aes(x=country)) + geom_bar(aes(y=alumni_employment), fill="blue", stat="identity")
```

Plot:



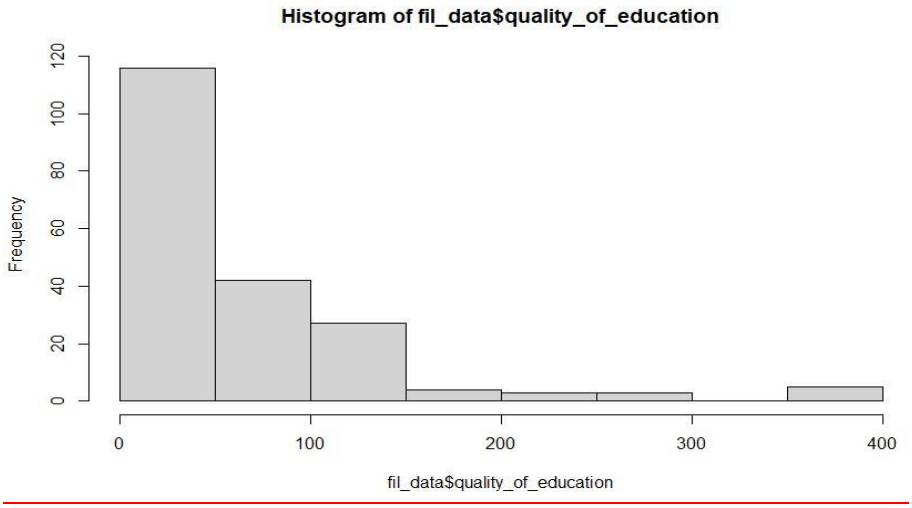
In this chart there is a comparison between the countries and alumni's who got employed in the years 2012-2015. The x axis depicts the countries and the y axis shows the employment rate. From this plot we understand that USA leads the alumni employment criteria followed by Israel and France, whereas students from South Korean Universities lack in acquiring employment immediately after their education.

Histogram

Code:

```
#Histogram of quality of education
hist(fil_data$quality_of_education)
```

Plot:



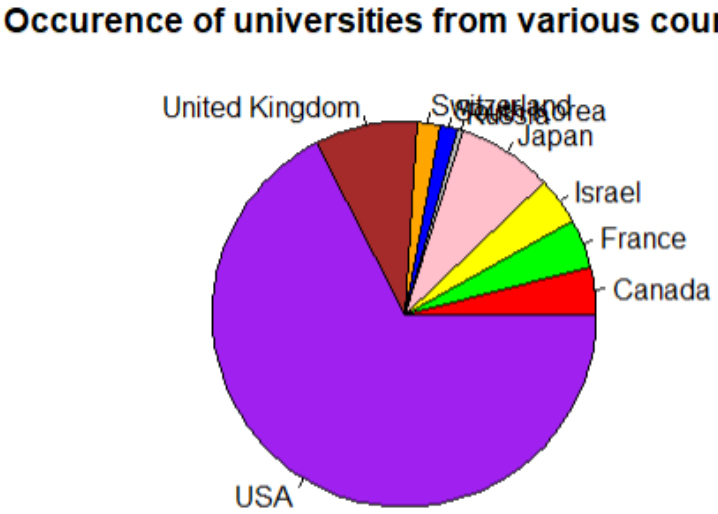
The quality of education is measured using the histogram bar. This plot depicts the aggregation of quality of education from the acquired data.

Pie Chart

Code:

```
pie(table(fil_data$country),
    col = c("red","green","yellow","pink","grey","blue","orange","brown","purple"),
    main="Occurence of universities from various countries")
```

Chart:



The pie chart above depicts the proportion of universities that has been ranked globally. Each country has been given diverse colours for identification purposes. From the diagram, it is clear that USA holds the major share of universities among the top 50 followed by United Kingdom and Japan.

Central Tendency & Measure of variance

Using mean we determined the mean number of patents registered among whole data.

Code:

```
# mean of patents within the institutions
patent_mean <- mean(fil_data$patents)
patent_mean
```

```
[1] 65.24
```

Using the sd() function, we saw how far the publications variable deviates from the mean.

Code:

```
# standard deviation of publications
publication_sd <- sd(fil_data$publications)
publication_sd
```

```
[1] 81.15522
```

Using the var function we calculated the variance of the variable alumni_employment.

Code:

```
# variane for alumni employment
alumni_employment_var <- var(fil_data$alumni_employment)
alumni_employment_var
```

```
[1] 14341.71
```

As seen below, we defined a mode function which was used to determine the number of citations that occurred the most in the dataset.

```
# mode of citations
def_mode <- function(var)
{
  unique_var <-unique(var)
  unique_var[which.max(tabulate(match(var,unique_var)))]
}

citation_mode <- def_mode(fil_data$citations)
citation_mode
```

```
[1] 101
```

Probability Distributions

Probability distributions can be defined as mathematical calculations using statistics which provide us with the possible outcomes of a data point in different situations. For our gathered data we used the below listed models for probability calculations:

Multinomial Model:

This model was used to determine the possibility of occurrence of each institute as next data point.

Code:

```
#Probability of next occurrence in data for each institute
x <- fil_data$institution
phat = (table(x)/length(x))
phat
```

Output:

```
Ã\u2030cole normale supÃ©rieure - Paris                Ã\u2030cole Polytechnique
                                0.010                                0.010
          California Institute of Technology                Carnegie Mellon University
                                0.020                                0.005
                  Columbia University                       Cornell University
                                0.020                                0.020
                   Dartmouth College                          Duke University
                                0.010                                0.020
                   Harvard University                Hebrew University of Jerusalem
                                0.020                                0.020
                Imperial College London                   Johns Hopkins University
                                0.020                                0.020
                      Keio University                          Kyoto University
                                0.010                                0.020
      Lomonosov Moscow State University             Massachusetts Institute of Technology
                                0.005                                0.020
                    McGill University                       New York University
                                0.020                                0.020
             Northwestern University                 Ohio State University, Columbus
                                0.020                                0.010
                     Osaka University       Pennsylvania State University, University Park
                                0.020                                0.015
      Pierre-and-Marie-Curie University                     Princeton University
                                0.005                                0.020
      Purdue University, West Lafayette                     Rockefeller University
                                0.015                                0.020
       Rutgers University-New Brunswick                     Seoul National University
                                0.015                                0.015
                Stanford University Swiss Federal Institute of Technology in Zurich
                                0.020                                0.020
            University College London                       University of Arizona
                                0.020                                0.005
      University of California, Berkeley                University of California, Davis
                                0.020                                0.005
       University of California, Irvine             University of California, Los Angeles
                                0.005                                0.020
     University of California, San Diego           University of California, San Francisco
```

0.020	0.020
University of California, Santa Barbara	University of Cambridge
0.010	0.020
University of Chicago	University of Colorado Boulder
0.020	0.010
University of Edinburgh	University of Illinois at Urbanaâ€“Champaign
0.005	0.020
University of Michigan, Ann Arbor	University of Minnesota, Twin Cities
0.020	0.020
University of North Carolina at Chapel Hill	University of Oxford
0.020	0.020
University of Paris-Sud	University of Pennsylvania
0.015	0.020
University of Pittsburgh - Pittsburgh Campus	University of Southern California
0.005	0.010
University of Texas at Austin	University of Texas Southwestern Medical Center
0.020	0.010
University of Tokyo	University of Toronto
0.020	0.020
University of Utah	University of Virginia
0.010	0.010
University of Washington - Seattle	University of Wisconsinâ€“Madison
0.020	0.020
Waseda University	Washington University in St. Louis
0.010	0.005
Weizmann Institute of Science	Yale University
0.020	0.020

Now consider one specific institute and check the possibility of its occurrence in a specific situation.

#QUESTION: if we select the next 7 institutes, what is the probability that there will be 3 occurrences of Washington University in St. Louis, since probability of Washington University in St. Louis ~ 0.005.

Code:

```
set.seed(456)
1 - pbinom(0.005,7,phat[62])
```

Output:

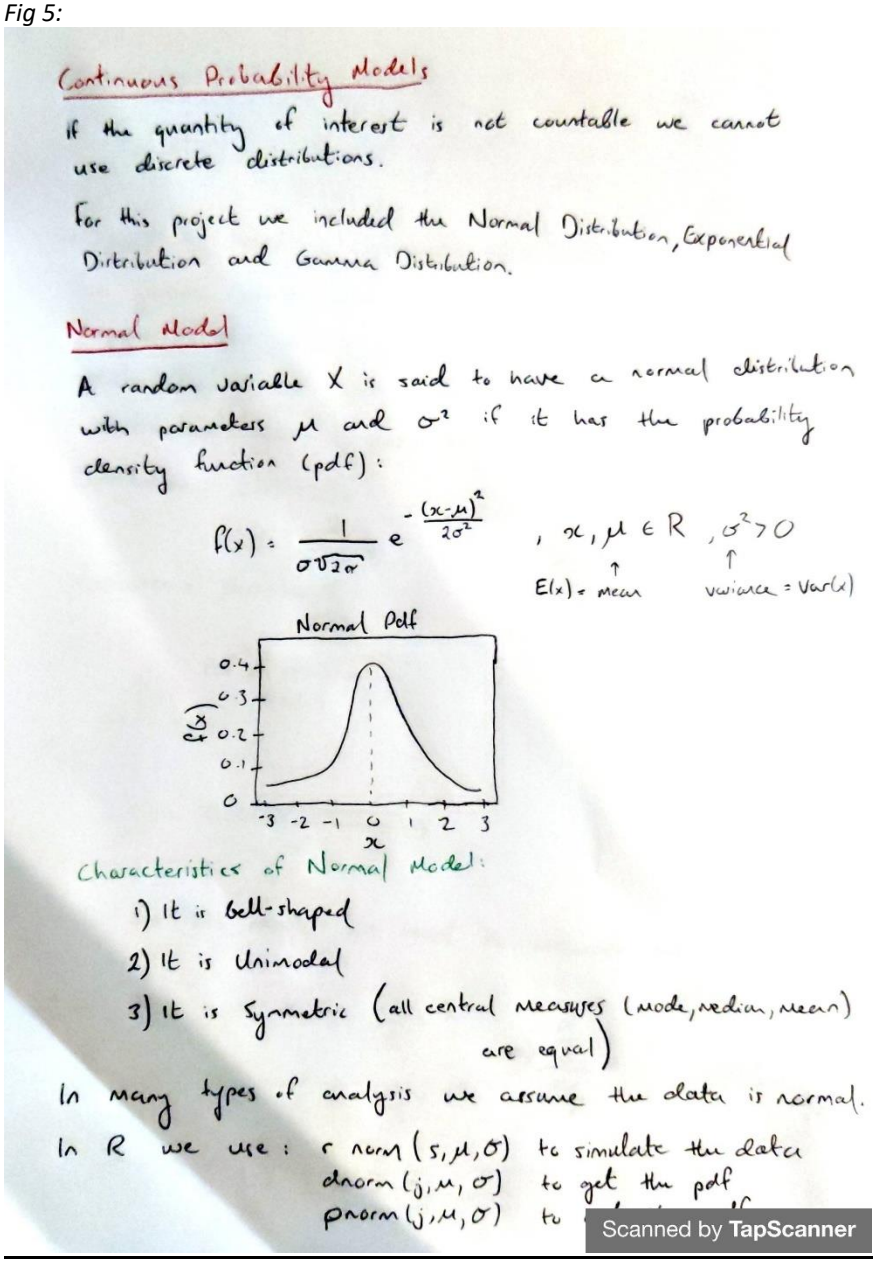
```
[1] 0.03447935
```

Result: This model shows that there is a possibility of 3.44% that there will be 3 occurrences of Washington University in St. Louis in next 7 instances.

Normal Model:

Below we have included the mathematical requirements and rules associated with the Normal Distribution used for continuous variables.

Fig 5:



We modelled the "quality_of_education" variable in the world universities dataset using the normal distribution. We estimated the parameters and used these to make a prediction. The parameters used for the normal distribution are mu and sigma. To estimate mu we took the mean of the variable "quality of education" and to estimate sigma we took the standard deviation of the variable "quality of education".

Code:

```
x <- fil_data$quality_of_education
mu=mean(x)
mu                  #[1] 275.1005
sigma=sd(x)
sigma               #[1] 121.9351
```

We then had to make a prediction. For continuous variables, to make a prediction, we need to simulate an amount of samples (1000) from the corresponding distribution and then compute the mean of these simulated values. This will be the prediction. To have the same result we used set.seed() and entered the same seed.

Code:

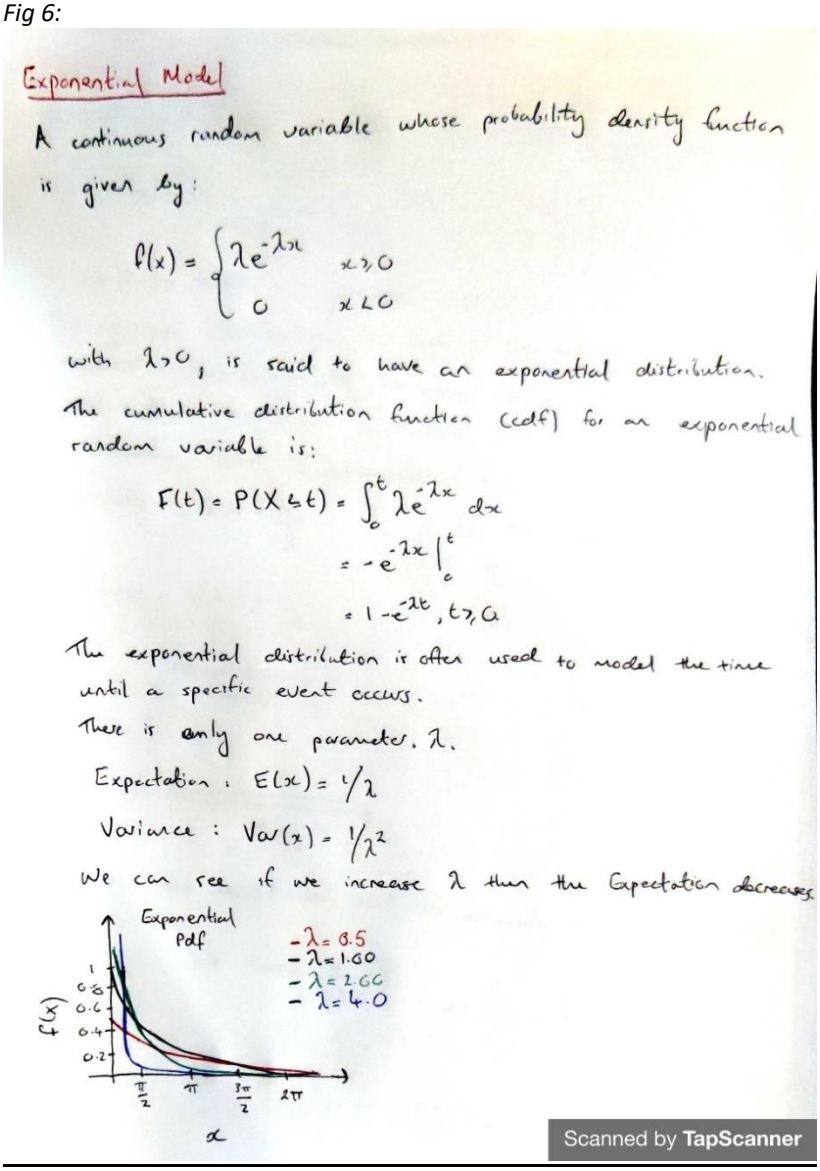
```
set.seed(4567)
sim=rnorm(1000,mu,sigma)
pred=mean(sim)
pred                    #[1] 272.5365
```

So, our predicted value was 272.5365. This was close to the mean we got earlier (275.1005) but was a more accurate result. We applied randomization to get a stochastic prediction based on the normal distribution where the parameters were from the data.

Exponential Model:

Below we have included the mathematical requirements and rules associated with the Exponential Distribution used for continuous variables.

Fig 6:



In the exponential model the quantity of interest is positive. We can see that the “citations” variable consisted of positive real values so the exponential model was appropriate. For the exponential model there is only one parameter, lambda. We know the expectation of x, E(x), is 1/lambda (i.e. the reciprocal to lambda).
This means to get lambda we needed 1/Exp(x).

Code:

```
x <- fil_data$citations
mu=mean(x)
mu                      #[1] 413.4173
lambda=1/mu
lambda                  #[1] 0.002418864
```

We then had to make a prediction. As previously stated, for continuous variables, to make a prediction, we need to simulate 1000 samples from the corresponding distribution and then compute the mean of these simulated values. This will be the prediction. To have the same result we had to define a seed. We used set.seed(). We then generated the data from the exponential distribution. It is common practice in the continuous case to make predictions using the parameters, then simulate the data from the model and the parameters, and then take the mean of the simulated data to make our prediction.

Code:

```
set.seed(45798)
sim=rexp(1000,lambda)     #We used rexp for data simulation in the exponential model.
pred=mean(sim)
pred                       #[1] 428.252
```

Our predicted value was 428.252. This was close to the mean we got earlier (413.4173) but was a more accurate result. However, was it the most accurate result?

We could also model the citations variable with the normal distribution and then compare results to see which was more accurate. Once again for the normal distribution we had the parameters mu and sigma.
To estimate mu we took the mean of the variable "citations" and to estimate sigma we took the standard deviation of "citations".

Code:

```
x <- fil_data$citations
mu=mean(x)
mu                      #[1] 413.4173
sigma=sd(x)
sigma                   #[1] 264.3665
```

We then made a prediction. Once again, for continuous variables, to make a prediction, we needed to simulate 1000 samples from the normal distribution and then compute the mean of these simulated values. This was our prediction.
To have the same result we used set.seed() with the same seed.

Code:

```
set.seed(45798)
sim=rnorm(1000,mu,sigma)
pred=mean(sim)
pred                    #[1] 404.1792
```

So, we got a predicted value for the "citations" variable from both the exponential distribution and the normal distribution, but which was better? To find out which distribution was more accurate for this variable we needed to compute the estimation of model variance and the lowest value of variance would be the best model. Therefore, the predicted value for the corresponding model would be the best prediction. So for both models we got:

Code:

```
Var_e=(mean(x))^2          #This gets the variance for the Exponential model
Var_e                      #[1] 170913.8
Var_n=var(x)               #This gets the variance for the Normal model
Var_n                      #[1] 69889.67
```

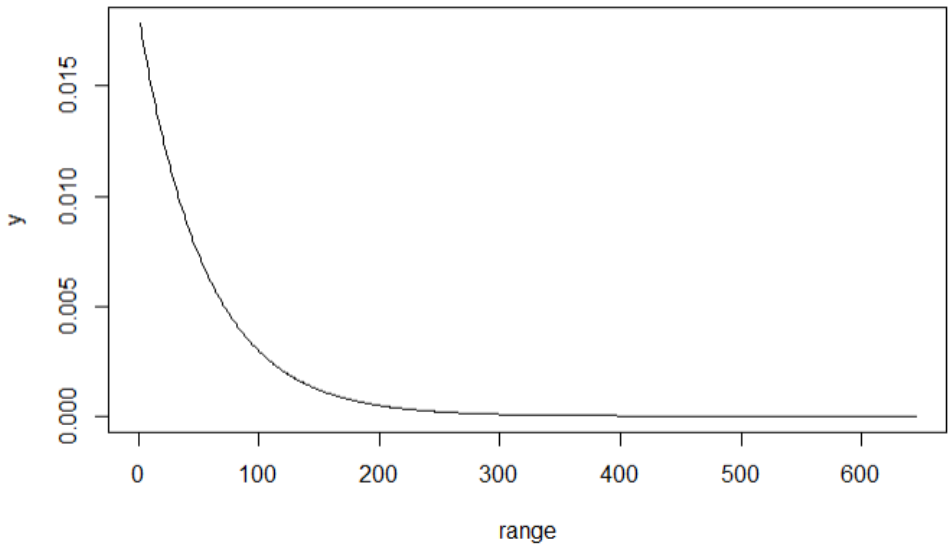
Since the estimation of variance using the normal distribution is lower, the more accurate prediction is 404.1792, (The prediction from the normal distribution).

We also posed another question. We found the probability that the next value of number of citations would be greater than or less than 50.

Code:

```
x <- fil_data$citations
mu = mean(x)
lambda = 1/mu
range = seq(min(x),max(x),1)
y = dexp(range,lambda)
plot(range,y,type = 'l')
```

Plot:



Exponential distribution of citations variable.

```
#Probability that the next data point produced less than 50 citations:
pexp(50,lambda)
```

Output:

[1] 0.5975095

Conclusion: There is a 59.7% possibility that the next data point will have less than 50 citations.

```
#Probability that the next data point produced more than 50 citations
1 - pexp(50,lambda)
```

Output:

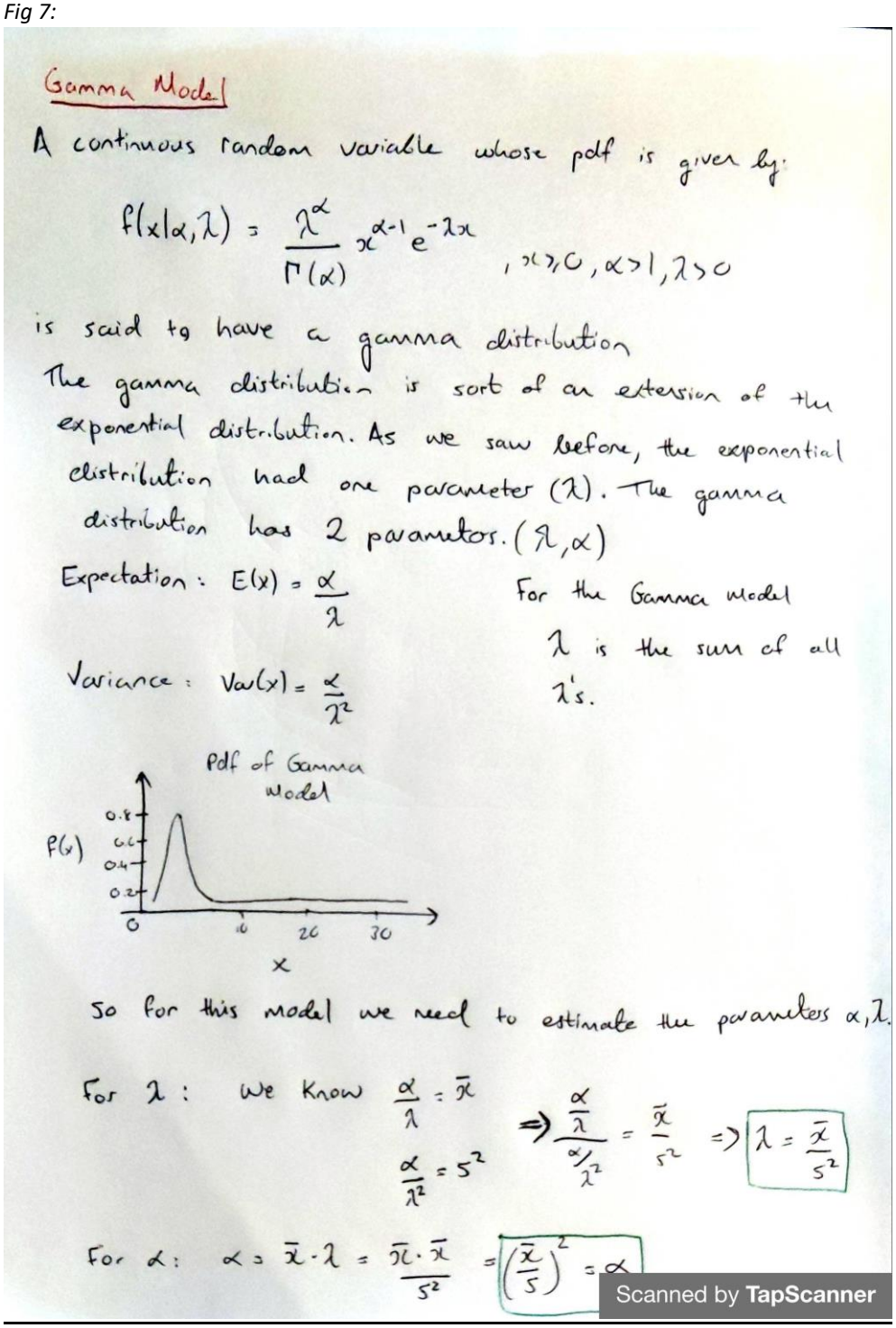
[1] 0.4024905

Result: There is a 40.2% possibility that the next data point will have more than 50 citations.

Gamma Model:

Below we have included the mathematical requirements and rules associated with the Gamma Distribution used for continuous variables.

Fig 7:



Next we used the Gamma distribution to model the "quality_of_faculty" variable.
For this model there are two parameters, alpha and lambda.

Code:

```
#Gamma Model
y <- fil_data$quality_of_faculty
xbar = mean(y)
s = sd(y)

#Parameter estimation
alpha = (xbar/s)^2
lam = xbar/(s^2)

#Prediction
set.seed(4567)
sim = rgamma(1000,alpha,lam)
pred = mean(sim)
pred
```

Output:

```
[1] 177.0844
```

Result:

After parameter estimation using the available data and simulation of 1000 data points, we used simulation to predict the next value for our dataset. The next possible value for quality of faculty would be 177.0844.

Hypothesis Testing

Hypothesis testing can be defined as the test to make a deduction whether the value of an argument associated with a particular whole value i.e. greater than/ less than or equal to (James McClave, 2018). These tests generally include 5 steps which are described below:

Step 1: State your claim/null hypothesis, denoted by **H0**.

State alternative claim/hypothesis, denoted by **H1**.

Step 2: Set alpha or significance level, denoted by **α .** It is usually given but if not take it as 0.05.

Step 3: Compute the test value using the available data and **H0.**

Step 4: Find c value using **α** and **H1.**

Step 5: Using decision rule we make a decision using step 3 and step 4 to check if our claim is being accepted or rejected.

For our data we will now use various types of hypothesis testing to determine if our null hypothesis or claim is accepted or rejected.

Lower One Tail Test

Code:

```
# Claim: h_0 ~ We claim that the average score for institutions in the dataset is greater than 90% at the level of alpha = 0.05.

x <- fil_data$score
# h_0 = mu > 90 ~ mu0
# h_1 = mu <= 90
alpha = 0.05
mu0 = 90
n = nrow(fil_data)
xbar = mean(x)
sigma_2= var(x)
test_value = (xbar - mu0)/(sigma_2/sqrt(n))
c_value = qnorm(alpha)
test_value
c_value

if (test_value <= c_value){
  print('H_0 Rejected')
}else{
  print("Accept H_0")
}
```

Output:

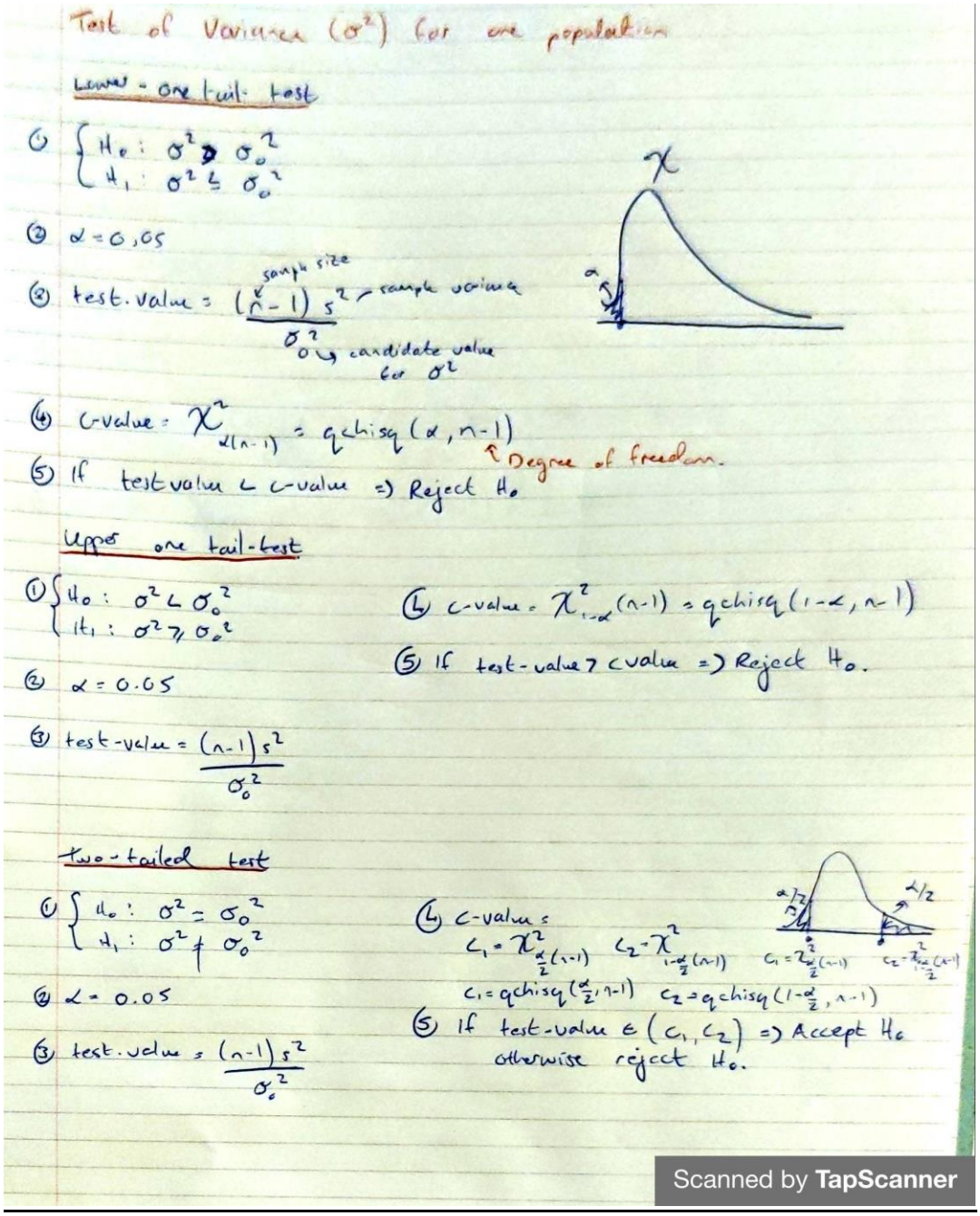
```
[1] -1.557621    #test value
[1] -1.644854     #c value
[1] "Accept H_0"
```

Result:

Since the test value here was greater than the c value, we accepted the null hypothesis.

Test for Variance

Fig 8:



Code:

```
# Claim that average score of quality of education within the institutions is 40. Design a hypothesis test whether claim is valid or not.
# we examine all products and record their average quality of education and standard deviation

# Test whether variance of the data is greater than 70 at alpha=0.05.
x <- fil_data$quality_of_education
# H0 : sigma^2 > nrow(fil_data)
# H1 : sigma^2 <= nrow(fil_data)
alpha = 0.05
n = nrow(fil_data)
```

```
s = sd(x)
sigma_2 = nrow(fil_data)
test.value = (n-1)*(s^2)/sigma_2
test.value
c.value = qchisq(alpha,n-1)    #n-1 Is degree of freedom
c.value

if (test.value < c.value){
  print("Reject H0")}else{
    print('Accept H0')
    }
```

Output:

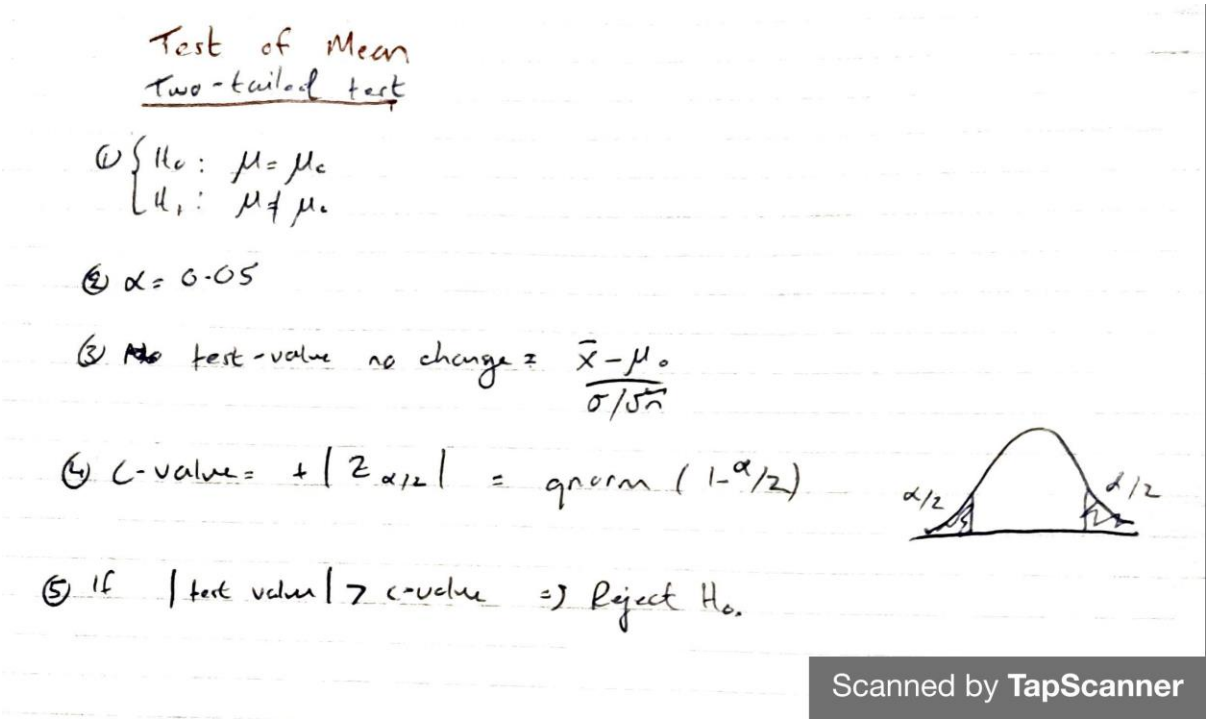
```
[1] 5059.158
[1] 167.361
[1] "Accept H0"
```

Result:

Since our test value here is greater than c value, hence we can say that our null hypothesis can be accepted.

Test of mean (Two-tailed test)

Fig 9:



Continuing with our previous claim.

Code:

```
#Apply test of mean for average score of quality for all institutions at alpha = 0.05
x <- fil_data$quality_of_education
# H0 : mu = 40
# H1 : mu =! 40
mu0 = 40
n = nrow(fil_data)
```

```
s = sd(x)
xbar = mean(x)

test.value = (xbar-mu0)/(s/sqrt(n))
test.value
c.value = qnorm(1-alpha/2)
c.value

if(abs(test.value)>c.value){
  print('Reject H_0')
}else{
  print("Accept H_0")
}
```

Output:

```
[1] 4.274002
[1] 1.959964
[1] "Reject H_0"
```

Result:

After test of mean, since absolute test value is greater than c value, we rejected the null hypothesis.

Bibliography

James McClave, T. S., 2018. Statistics by McClave Thirteenth Edition. In: s.l.:Pearson, p. 397.

Wikipedia, 2020. *Wikipedia.* [Online]
Available at: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining
[Accessed 15 March 2021].

Appendix A

Group Report

For this assignment we had to find people from our class to work as a group to complete a project. It was not possible to have any physical meetings due to lockdown so we decided to create a Whatsapp group chat where we could discuss our ideas and how to commence the project. We then organised a time that suited all of us for a zoom meeting. During this meeting, we introduced ourselves and informed each other of our past educational and employment background to see which parts of the project we would be stronger at individually. From this meeting we came to know about each other's skills and knowledge which we have listed below:

- Killian completed a Bachelor of Science degree in Applied Mathematics in 2020 and has also tutored maths to leaving certificate students for the past 5 years which proved to be accommodating to other team members in situations where we were unable to understand the mathematics backing our probability models and hypothesis testing.
- Mohan holds a Bachelors of Computer Applications completed in 2019 after which he also worked as a Learning Management System Administrator at GP Strategies Corporation for 1 year 6 months. This gave him skills and knowledge in fields such as understanding the commercial aspect of data and concisely reporting the end product of our project.
- Vinay holds a Bachelors of Engineering in Information Technology completed in 2016 after which he worked as a Data Processing Executive for 3 years 11 months at Yougov, which made him extremely equipped in R and Python programming skills and with process and various methodologies of data manipulations.

After this meeting, we discovered that all of us had some skills and knowledge to contribute to the project. We also decided to search individually for a relevant dataset which could be used for the assignment. All of us explored the web for a relevant dataset for a week and proposed each of our datasets in the next zoom meeting. Out of the three datasets we decided to go with this educational dataset, since the education system and its relevance to students and their families has drastically changed since the pandemic began.

We then had to collaborate and successfully execute the project. To do this we used R Studio along with Github to collaboratively complete the project and maintain a transparency throughout the project lifecycle. Since we had finalised a huge dataset of multiple universities from around the globe for over 4 years, we decided to clean and filter the data for us to have a meaningful dataset.

We then assigned each one of us a task to recognize the inconsistencies within the data to be eliminated. Each of us dedicated a week to understand the data and point out any inconsistencies within the data.

Mohan recognized that it was impossible to represent a huge chunk of missing data for the variable "broad_impact" and also was loosely correlated to other columns due to which we had no other option other than to remove that variable from consideration for our models.

Killian recognized that it was impossible to precisely describe the original data using graphs, hence we had to reduce our data to the top 50 universities for 4 years and only use this data for representation and our data models.

Vinay recognized that it was possible to transform 3 chr variables to categorical because all of the 3 variables had repetitive responses in rows. After reading the data on R studio each of us did the cleaning and preparation of data and pushed their code for merging on Github.

For descriptive analysis we decided to take one variable each and described our data using graphs and measures of central tendency. Vinay used bar plots to describe the bar graph for year vs number of institutes from each country within the dataset. Killian used quantile ranges and boxplot to identify outliers in patents for each year

of the data and Mohan used pareto chart to describe the number of alumni's employed from universities collaboratively from each country. We also proposed and implemented some central tendency measures for various variables within our dataset.

At this stage of the project each of us were assigned a task to identify one variable each for the probability models. Killian wrote out the mathematics describing each model to help establish a better understanding of how the models operate. We had to identify the pro and cons of the model we chose, the mathematics behind the model and why do we use it on the variable. We took a week to understand various probability models and on which variables those models can be used. Each one of us then reported our findings back to the team and proposed why we should use the specific model on a variable.

Killian implemented the exponential model on the citations variable and also decided to do a comparison with the Normal model to see which model was more accurate. Mohan implemented the Gamma model on the "quality_of_faculty" variable and Vinay implemented the multinomial model on the discrete variable institution. We all decided to implement the normal distribution on the "quality_of_education" variable to further enhance our knowledge and understanding of probability models. All of us took around 2 weeks to implement and justify our probability models and at the end we merged our codes on Github.

To finish off our project we decided to take one hypothesis each and work along with it while describing it and concluding results for each of them.

As a team and working together we contributed our findings in various phases of this project and worked hard towards its successful execution and report. All together it took us around 4-5 weeks to complete the assignment and report it. We are happy with how it turned out and have a much better understanding of the statistical concepts involved in modern day data analytics.