Statistic for Data Analytics: CA1

November 2022

Submitted by:
Student Name: Vaibhav Tiwari
Student ID: 10621051
Lecturer Name: Dr Shahram Azizi

Group Members:
Aniket Bachewar : 10621078
Prathamesh Jadhav : 10621081
Vaibhav Tiwari : 10621051

Table of Content:

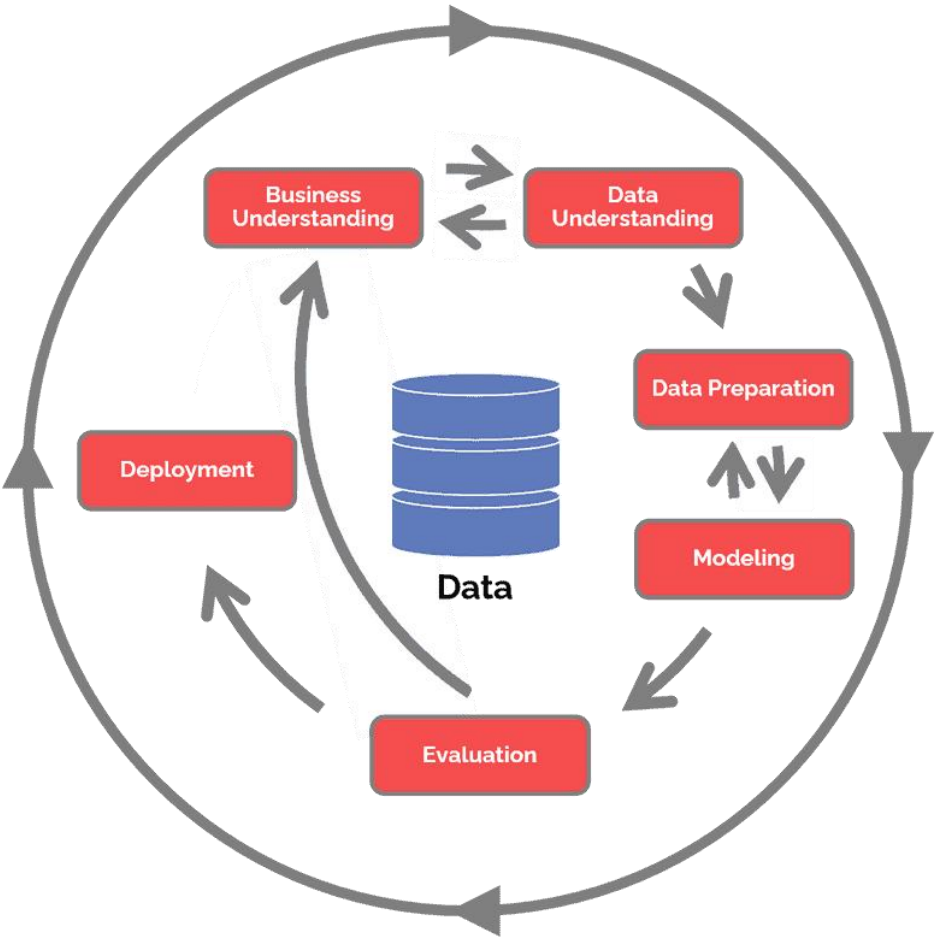
Code colab Link :
https://colab.research.google.com/drive/1T4MTQAGXKN_hnXNF1D5UYeBdvBKpwYCY#scrollTo=59flCJC5rK1V

Dataset :
https://www.kaggle.com/ahmedmohamedmahrous/churn-modelling1

1. Introduction

Statistics in mathematics involves collecting, organizing, analysing, interpreting, and presenting data. Descriptive analysis is one way to collect, summarize, and make sense of the historical data.
Statistics provides us with the most useful techniques to analyse and interpret the data. We have chosen the Bankers dataset as this dataset has sufficient data - continuous and discrete numerical variables and binary variables that we can work on.
The dataset itself has 14 variables. For our implementation, we have primarily selected “CreditScore”, “Age” and “HasCrCard as our variables of choice for showcasing the different models.

We have articulated and developed this assignment on Google Colab and have followed the CRISP-DM methodology to sub divide and understand the different stages that were required.



Descriptive Analytics being simplest form of Data Analytics involves summarizing data set’s main features or the past data into a readable format. This helps in aggregating our findings from historic for the purpose of identifying patterns or tends.

Some real-world examples to mention are revenue per customer and the average time customers take to pay bills.

Descriptive Analytics techniques:

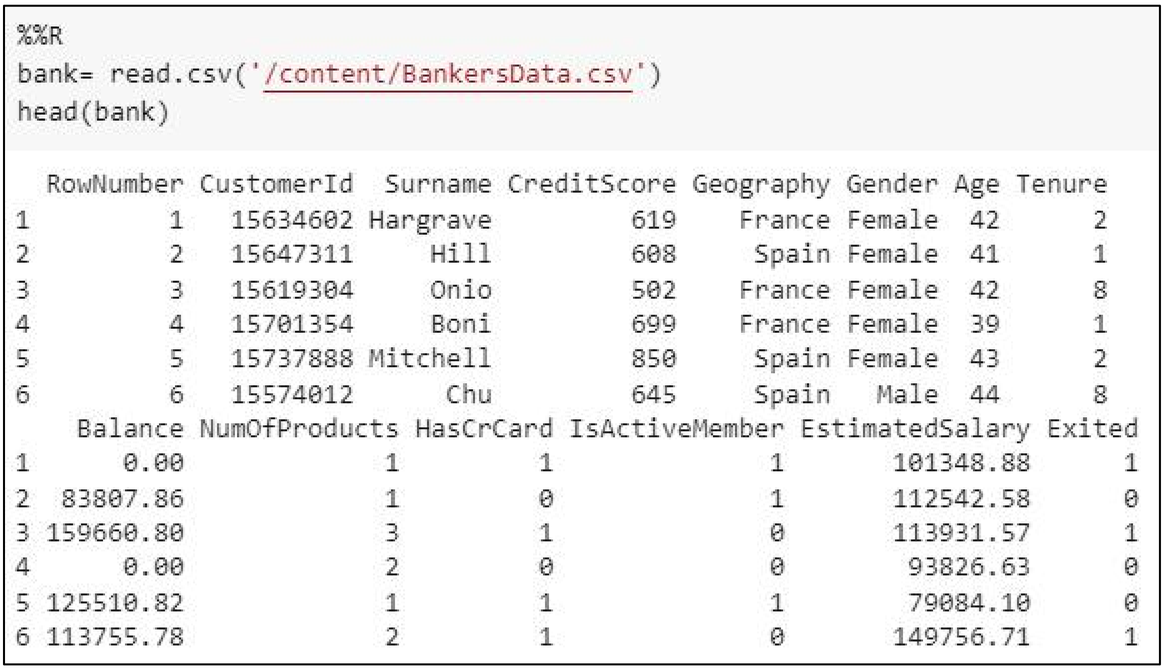
i. Data Aggregation: This is a process of compiling information from databases for data processing.

ii. Data mining: This is a process of extracting and discovering patterns in large data sets

2. Dataset

The dataset describes the banking information of the customers. The dataset contains the details of the existing and old customers of the bank. After reading through all the attributes of the dataset, we get various information of the customers such as age, credit score, location, gender, tenure, bank balance, has credit card, is active customer and is currently a customer with the bank or not.

Below is the snapshot of the dataset wit head function returning first 10 rows of the dataset. The dataset contains information of 10000 customers.



```
%%R
bank= read.csv('/content/BankersData.csv')
head(bank)
```

```
  RowNumber CustomerId  Surname CreditScore Geography Gender Age Tenure
1         1   15634602 Hargrave         619    France Female  42      2
2         2   15647311     Hill         608     Spain Female  41      1
3         3   15619304     Onio         502    France Female  42      8
4         4   15701354     Boni         699    France Female  39      1
5         5   15737888 Mitchell         850     Spain Female  43      2
6         6   15574012      Chu         645     Spain   Male  44      8
    Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
1      0.00             1         1              1       101348.88      1
2  83807.86             1         0              1       112542.58      0
3 159660.80             3         1              0       113931.57      1
4      0.00             2         0              0        93826.63      0
5 125510.82             1         1              1        79084.10      0
6 113755.78             2         1              0       149756.71      1
```

3. Data Cleaning and Data Preparation

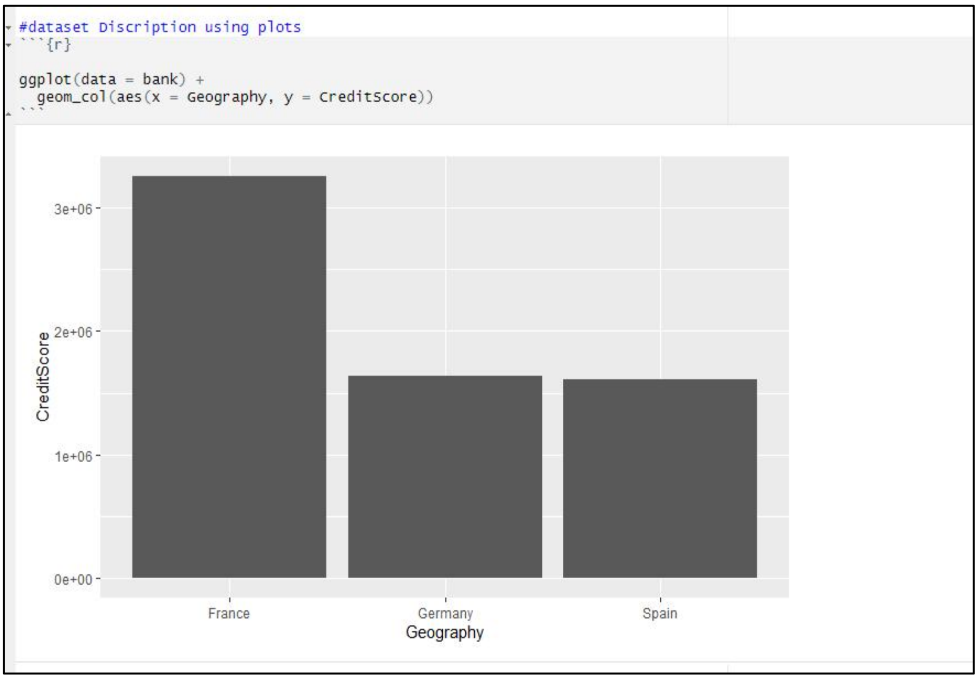
As per above described dataset, the dataset is already cleaned with appropriate values in all the columns. So there is no need for explicit data cleaning and data preparation.

4. Graph Plotting

Data visualization and representation is very simple using the graphs.

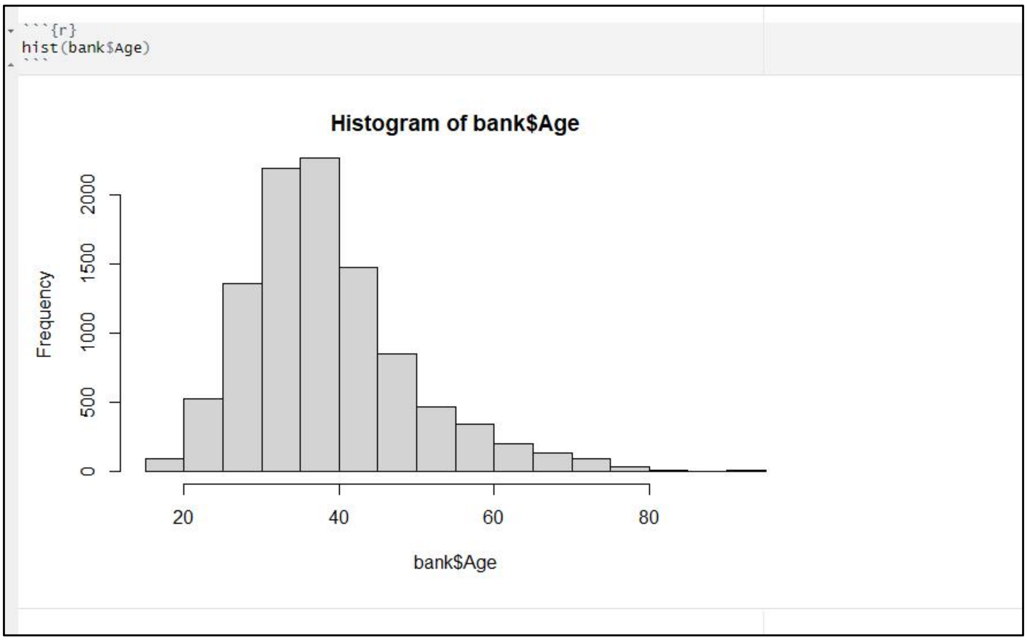
Considering the banker's dataset, we can represent/visualize the credit score of the customers as per the location. As per the below graph, it is clearly understood that customers from France have a better credit score as compared to others.

The graphs are plotted with the help of ggplot.



~~~
#dataset Discription using plots
```{r}
ggplot(data = bank) +
  geom_col(aes(x = Geography, y = CreditScore))
```
~~~

Below is the Histogram representation of Age attribute.



~~~
```{r}
hist(bank$Age)
```
~~~

5. Numerical Measures

Numerical measures are of 2 types:
a) Measures of Central Tendency
b) Measure of Variability

i. Measures of Central Tendency

Measures of Central Tendency is used to derive a single/center value that describes or represents the set of data using the central measures.
Most widely used Central Measures are:

i. Mean
ii. Median
iii. Mode

a. Mean:

- Mean value of a column is a average value of the targeted column.
- Average value is derived by using all values of a targeted column and dividing it by the number of rows.

Formula:

$$\bar{x} = \frac{\sum x}{N}$$

$$\sum x = \text{ the sum of } x$$

$$N = \text{ number of data}$$

Example:
Calculating mean of first five number using the above formula:
Mean = (1+2+3+4+5)/5
Mean of first 5 numbers is 3.

b. Median

- Median value of a column is the middle value of the targeted column.
- The middle value is derived by sorting the column values in ascending order and then extracting the middle value.

Formula:

When n is even

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}$$

When n is odd

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

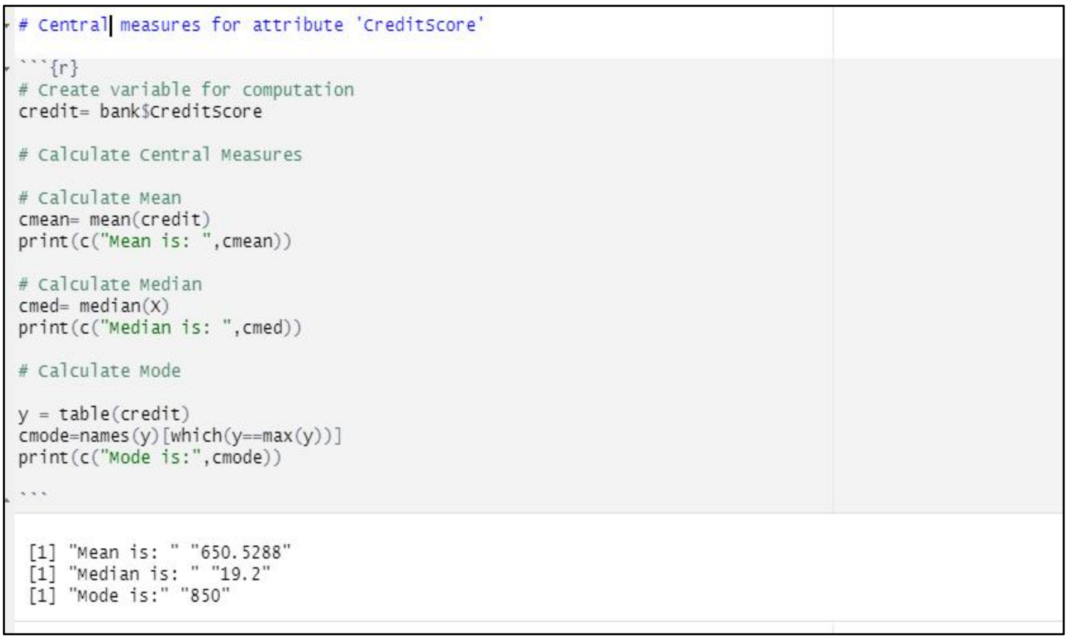
Example:
Calculating median of numbers (12,4,6,2,14) using the above formula:
Median = 2,4,6,12,14 is 6.

c. Mode

- Most occurring number from the targeted column is considered as mode value.
- The mode is the value with the highest frequency.

Example:
Deriving mode from list of values [1,2,3,2,4,2,5,2].
Mode of [1,2,3,2,4,2,5,2] is 2, as 2 has the highest frequency of 4 in the list.

Code: Compute Central Tendency measures for the Credit Score attribute from the dataset.



~~~~
# Central measures for attribute 'CreditScore'

```{r}
# Create variable for computation
credit= bank$CreditScore

# Calculate Central Measures

# Calculate Mean
cmean= mean(credit)
print(c("Mean is: ",cmean))

# Calculate Median
cmed= median(X)
print(c("Median is: ",cmed))

# Calculate Mode

y = table(credit)
cmode=names(y)[which(y==max(y))]
print(c("Mode is:",cmode))

```

 [1] "Mean is: " "650.5288"
 [1] "Median is: " "19.2"
 [1] "Mode is:" "850"
~~~~

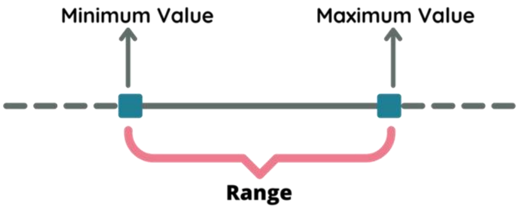
ii) Measures of Variability

Measures of Variability describes to what level the distribution is stretched within a dataset.
The measures of variability are as given below:

a) Range
b) Interquartile Range
c) Variance
d) Standard Deviation

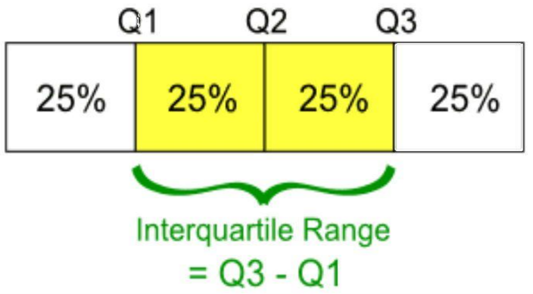
a. Range

The Range is the difference between the lowest and highest values of the selected variable within a dataset.
For example: In the given dataset,
A = {5,7,8,10,5,4} the lowest value is 4 and highest is 10
So, **maximum** data point – **minimum** data point = Range
Here the range, is 10 - 4 = 6



b. Interquartile Range (IQR)

To calculate the interquartile range, we first need to sort the data in ascending order. Once sorted, the IQR describes the 50% of the middle values. The lower and upper half after calculating median is the Q1 and Q3. The difference between Q3 and Q1 is the IQR.



c. Variance:

It is the dispersion of data around the mean value of dataset. It is calculated by taking squared deviations from mean.

$$s^2 = \frac{\Sigma\,(X - \bar{x})^2}{n - 1}$$

d. Standard deviation:

Standard deviation helps us to understand the spread between the data in relation to the mean.

Below is the formula to calculate Standard deviation, where:

x is the value from the dataset
μ is the mean of the dataset
N is the total number of values from the given dataset

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

6. Probability:

The probability model is a mathematical explanation for random events. Before implementing a few distribution models, we have to understand that the data can be of the following two types :
Discrete Data, which can take only specified values.
For example, when you roll a dice, the possible outcomes are 1, 2, 3, 4, 5 or 6 and not 1.5 or 2.45. These specific whole number outcomes are called Discrete values.
Continuous Data can take any value within a given range. The range may be finite or infinite. For example, A person's weight or height, the length of the road. The weight of a person
can take any value like 60 kgs, or 60.5 kgs, or 60.1234 kgs.
We have demonstrated a few of these models in the subsequent sections.

We have implemented below distribution model using the Banker’s dataset.

a) Bernoulli Distribution:

- Bernoulli Distribution comes from a random event or trial which gives rise to only two outcomes which are ubiquitous in nature.
- These outcomes are usually labelled as “success” or “failures”.
- In order to denote the same, if p denotes a possibility of success, then p-1 denotes a possibility of failure, so the Bernoulli probability of the function becomes:

P (0) = (1 - p) and
P (1) = p

Formula:

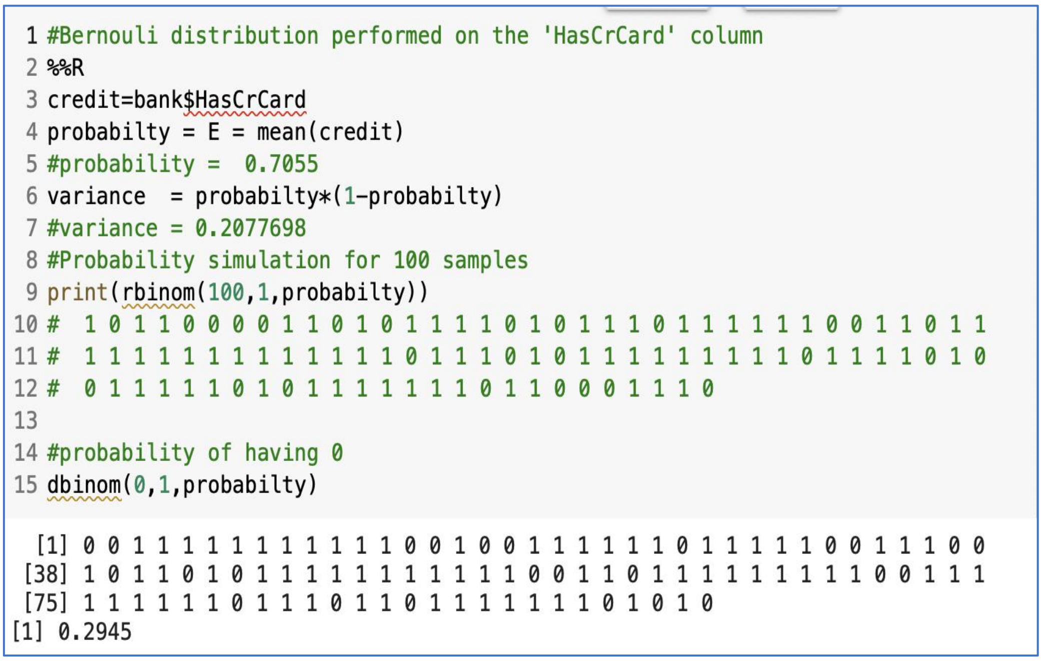
$$P(n) = \begin{cases} 1-p & \text{for } n = 0 \\ p & \text{for } n = 1 \end{cases}$$

labelled by,
n = 0 and n = 1 in which
n = 1 ("success") occurs with probability p and
n = 0 ("failure")

Which can be also written as,

$$P(n) = p^n (1-p)^{1-n}.$$

In the code below, we have performed Bernoulli distribution performed on the **'HasCrCard'** column:



```
#Bernouli distribution performed on the 'HasCrCard' column
%%R
credit=bank$HasCrCard
probabilty = E = mean(credit)
#probability =  0.7055
variance  = probabilty*(1-probabilty)
#variance = 0.2077698
#Probability simulation for 100 samples
print(rbinom(100,1,probabilty))
#  1 0 1 1 0 0 0 0 1 1 0 1 0 1 1 1 1 0 1 0 1 1 1 0 1 1 1 1 1 1 0 0 1 1 0 1 1
#  1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 0
#  0 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 0 1 1 0 0 0 1 1 1 0

#probability of having 0
dbinom(0,1,probabilty)

  [1] 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 0 1 1 1 0 0
 [38] 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1
 [75] 1 1 1 1 1 1 0 1 1 1 0 1 1 0 1 1 1 1 1 1 1 0 1 0 1 0
[1] 0.2945
```

Strength of Bernoulli Distribution:

- Two Bernoulli events are independent of each other as the number of trials and the probability of success and failure always will remain fixed.
- The Bernoulli Distribution inherits several other probability distribution:
- The binomial distribution
- The geometric distribution
- The negative binomial distribution

Weaknesses of Bernoulli Distribution:

- If in case during a Bernoulli Distribution event, there is a single trial, then the outcome simulated value would either be 0 or 1.

Real World Example:

- A possible where there would be exactly 2 outcomes like, a team will win a championship or not
- a student will pass or fail an exam
- a rolled dice will either show a 6 or any other number.

b) Poisson Distribution:

In statistics **Poisson distribution** is a discrete probability which determines the likeliness an event will occur in each period of time.
The Poisson distribution is denoted as 'Poi' and is expressed as follows:

Poisson distribution is denoted by **Poi** Y ~ Po (λ)

$$\Pr(x = j|\lambda) = \frac{e^{-\lambda}\lambda^{j}}{j!}, \ j = 0,1,2,\ldots, \ \lambda > 0$$

$$\mathrm{E}(x|\lambda) = \lambda$$

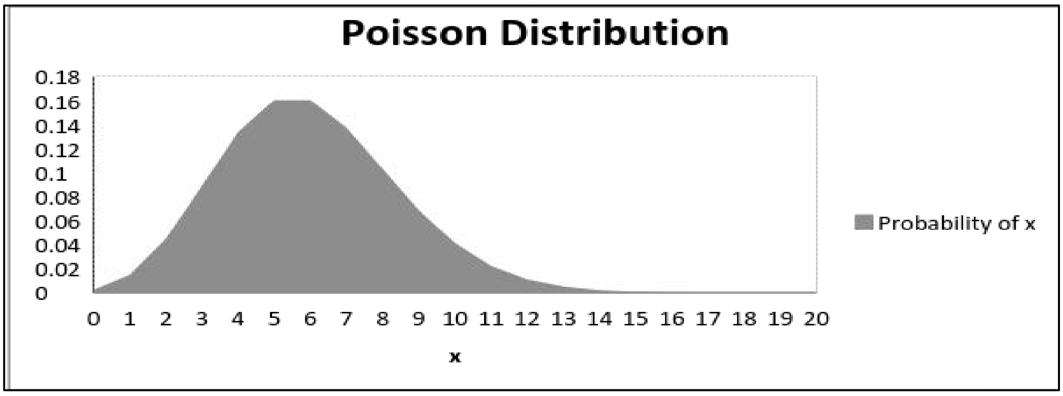
$$\mathrm{Var}(x|\lambda) = \lambda$$

Where, parameter λ corresponds to the expected number of occurrences in that time slot.

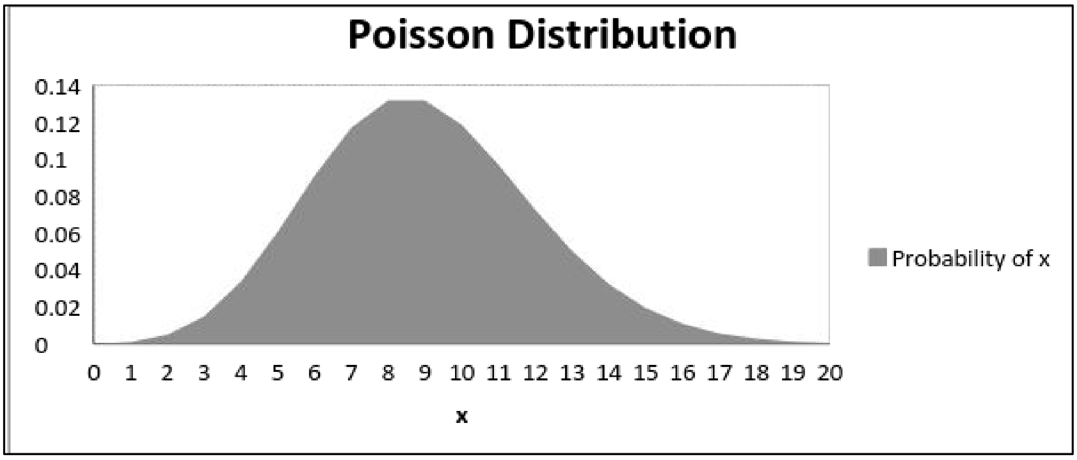
A distribution is called **Poisson distribution** when the following assumptions are valid:

- Any successful event should not influence the outcome of another successful event.
- The probability of success over a short interval must equal the probability of success over a longer interval.
- The probability of success in an interval approaches zero as the interval becomes smaller.

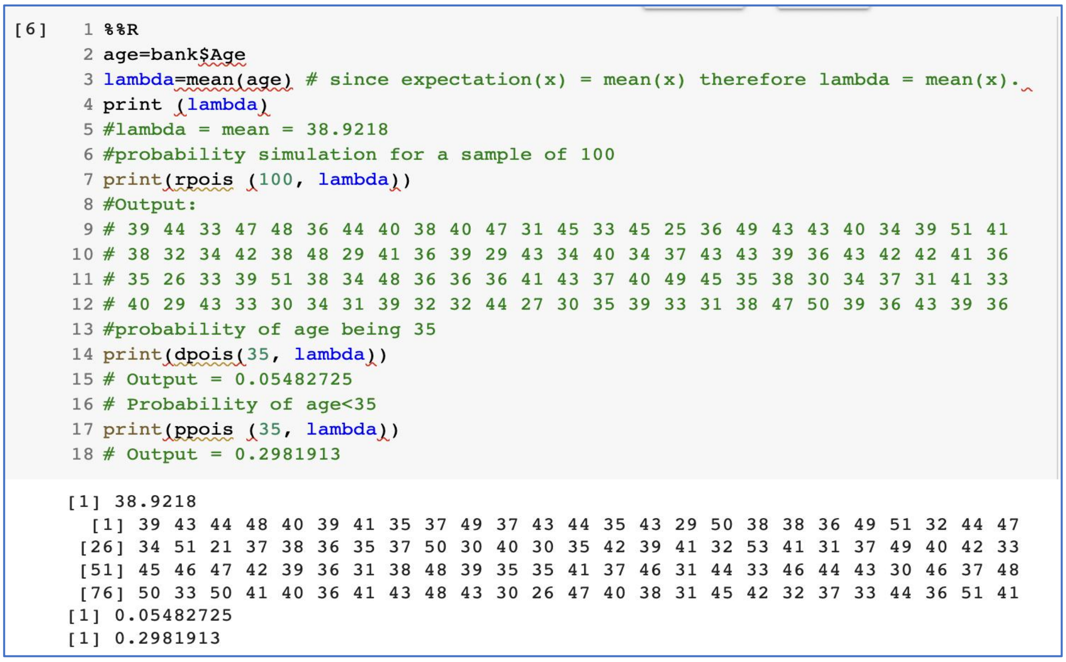
The mean μ is the parameter of this distribution. μ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown below:



The graph shown below illustrates the shift in the curve due to increase in mean. (It is perceived that as the mean increases, the curve shifts to the right.)



Code:
We will be performing Poisson Distribution on the 'Age' attibute from the Bankers dataset.



```
[6]
%%R
age=bank$Age
lambda=mean(age) # since expectation(x) = mean(x) therefore lambda = mean(x).
print (lambda)
#lambda = mean = 38.9218
#probability simulation for a sample of 100
print(rpois (100, lambda))
#Output:
# 39 44 33 47 48 36 44 40 38 40 47 31 45 33 45 25 36 49 43 43 40 34 39 51 41
# 38 32 34 42 38 48 29 41 36 39 29 43 34 40 34 37 43 43 39 36 43 42 42 41 36
# 35 26 33 39 51 38 34 48 36 36 36 41 43 37 40 49 45 35 38 30 34 37 31 41 33
# 40 29 43 33 30 34 31 39 32 32 44 27 30 35 39 33 31 38 47 50 39 36 43 39 36
#probability of age being 35
print(dpois(35, lambda))
# Output = 0.05482725
# Probability of age<35
print(ppois (35, lambda))
# Output = 0.2981913
```

```
[1] 38.9218
  [1] 39 43 44 48 40 39 41 35 37 49 37 43 44 35 43 29 50 38 38 36 49 51 32 44 47
 [26] 34 51 21 37 38 36 35 37 50 30 40 30 35 42 39 41 32 53 41 31 37 49 40 42 33
 [51] 45 46 47 42 39 36 31 38 48 39 35 35 41 37 46 31 44 33 46 44 43 30 46 37 48
 [76] 50 33 50 41 40 36 41 43 48 43 30 26 47 40 38 31 45 42 32 37 33 44 36 51 41
[1] 0.05482725
[1] 0.2981913
```

Strengths of Poisson Distribution:

- The outcome or an event occurrence can be counted and be a discreate number.
- Events are independent of one another which means the probability of a success event at given time interval is independent of previous occurrence.

Weaknesses of Poisson Distribution:

- Can be used when the average of occurrence does not change during the given period of time.

Typical real-world examples of Poisson Distribution are:

- The total number of car accidents occurring on a busy freeway on a given day.
- The number of technical issues in a gadget from a given brand such as a laptop or a mobile phone.

c) Normal Distribution:

- As the name suggest, normal distribution denotes a symmetrical pattern to which most of the natural events adhere to.
- Normal distribution also known as Gaussian distribution, is denoted by,
  Y ~ (μ, σ2)

68% of all its all values should fall in the interval, i.e. (μ – σ , μ+ σ)
E(Y) = μ
Var(Y) = σ2

Normal distribution can be standardized using Z Table.

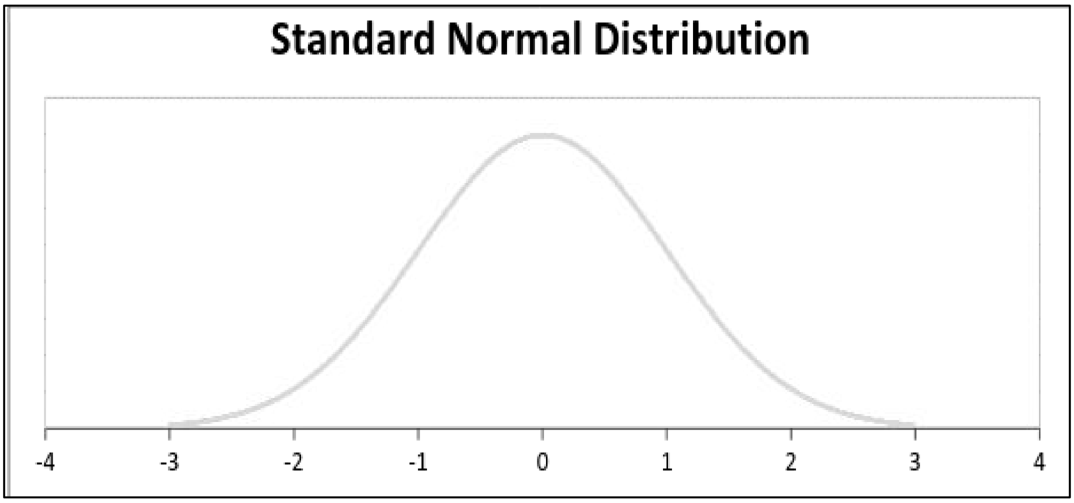
$$z = \frac{y - \mu}{\sigma}$$

Any distribution is known as Normal distribution if it has the following characteristics:

1. The mean, median and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center and the other half to the right.

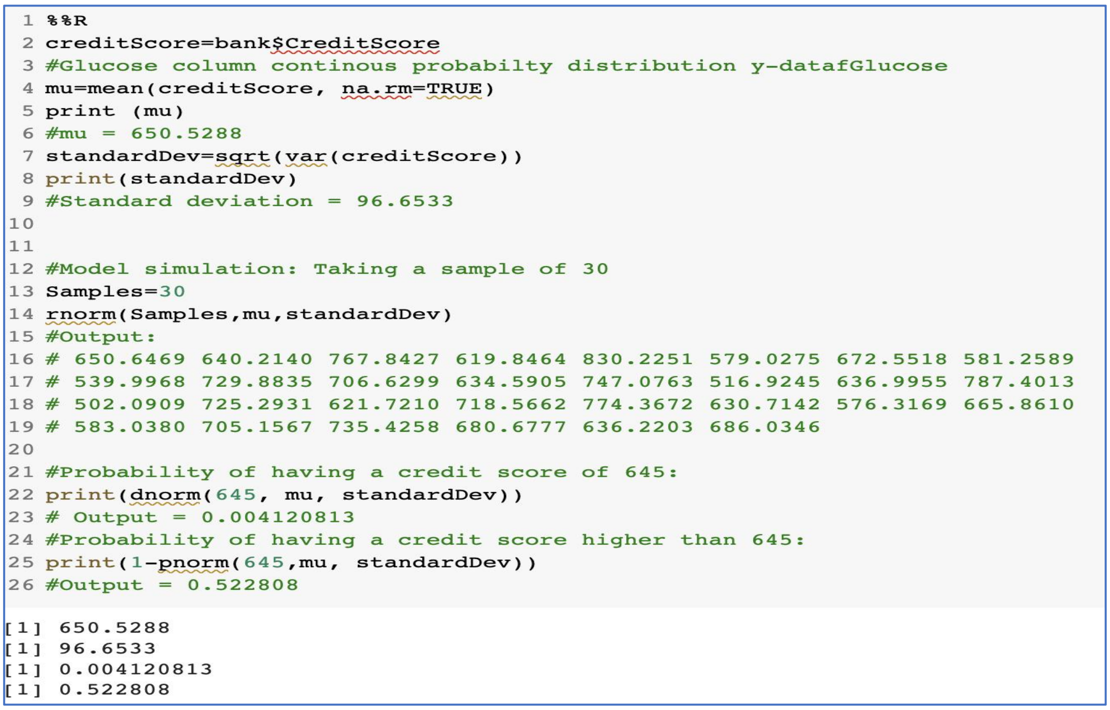
A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1. For such a case, the PDF becomes:



Code:

We perform Normal Distribution on the "EstimatedSalary" attribute from 'Bankers' dataset. We choose "EstimatedSalary" column because approximately 68% value fall between mean.



```
%%R
creditScore=bank$CreditScore
#Glucose column continous probabilty distribution y-datafGlucose
mu=mean(creditScore, na.rm=TRUE)
print (mu)
#mu = 650.5288
standardDev=sqrt(var(creditScore))
print(standardDev)
#Standard deviation = 96.6533


#Model simulation: Taking a sample of 30
Samples=30
rnorm(Samples,mu,standardDev)
#Output:
# 650.6469 640.2140 767.8427 619.8464 830.2251 579.0275 672.5518 581.2589
# 539.9968 729.8835 706.6299 634.5905 747.0763 516.9245 636.9955 787.4013
# 502.0909 725.2931 621.7210 718.5662 774.3672 630.7142 576.3169 665.8610
# 583.0380 705.1567 735.4258 680.6777 636.2203 686.0346

#Probability of having a credit score of 645:
print(dnorm(645, mu, standardDev))
# Output = 0.004120813
#Probability of having a credit score higher than 645:
print(1-pnorm(645,mu, standardDev))
#Output = 0.522808
```

```
[1] 650.5288
[1] 96.6533
[1] 0.004120813
[1] 0.522808
```

Strength of Normal Distribution:

- Normal Distribution is an important distribution as many continuous data in real world show a bell-shaped curved when it is compiled and graphed.
- The mean and the median are equals for normal distribution and 68% data falls within single data deviation.

Weaknesses of Normal Distribution:

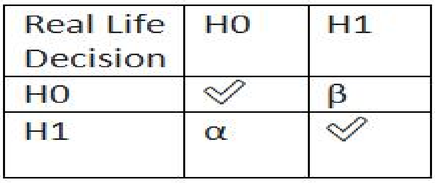
- Incomplete data for missing data for Normal Distribution can make it look completely scattered.
- In certain scenarios, Normal Distribution could make lot of assumption about the underlying data.

Real World Example:

- The distribution height of a person based on gender will reveal a histogram in bell shape.
- The measure of an individual reaction time in situation like driving, flying a fighter jet or an intense sport game. No two person are likely to have a similar reaction time.
- Health vitals like blood pressure within older men is different and their mean can be displayed in normal distribution with a bell shape.

7. Hypothesis Testing

Hypotheses Testing is a well-structured technique in Statistics and it applies binary decision making to test a claim for mean, variance or proportion of a population where the sample data and the level of significance provided for the testing. We need three important information to perform hypotheses testing: the claim, the significance level and the details of the sample data (Azizi lecture notes). For the Hypotheses Testing, we determine two hypotheses: H0 as null hypothesis and H1 as the alternative hypothesis. If we apply the decision rules into H0 and H1 hypotheses model, we will get the table below:



Real Life Decision	H0	H1
H0	✓	β
H1	α	✓

Alpha in the Table 1 shows the significance level (Type-1 Error) of our testing where Beta shows the Type-2 Error. There are two approaches in hypotheses testing: p-value with significance value (Alpha) approach, and test statistics (t-value) with critical value (c-value) approach (Shahram's lecture notes). We will use the test statistics and critical values approach for our hypotheses testing examples in our assignment. For each testing we perform, we share the decision rules with the necessary explanations.

a) Test of Mean

The Test of mean hypothesis compares the mean of a sample to a pre-specified value and tests for a deviation from that value. A one-sample test of means compares a sample's mean to a predetermined value and looks for deviations from it.

$$Test\ Value(Z) = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$Critical\ Value\ =\ |Z_\alpha|$$

Test Of Mean(x) --- normal population

i) two tailed tests
1: H0: x=x0; H1: x! =x0
2: alpha = 0.05
3: test_value = (X-p0)/(s/sqrt(n) X: sample mean; s: Standard deviation
4: c_value = qnorm(alpha/2)
5: if abs(test_value) > c_value ----> Reject H0

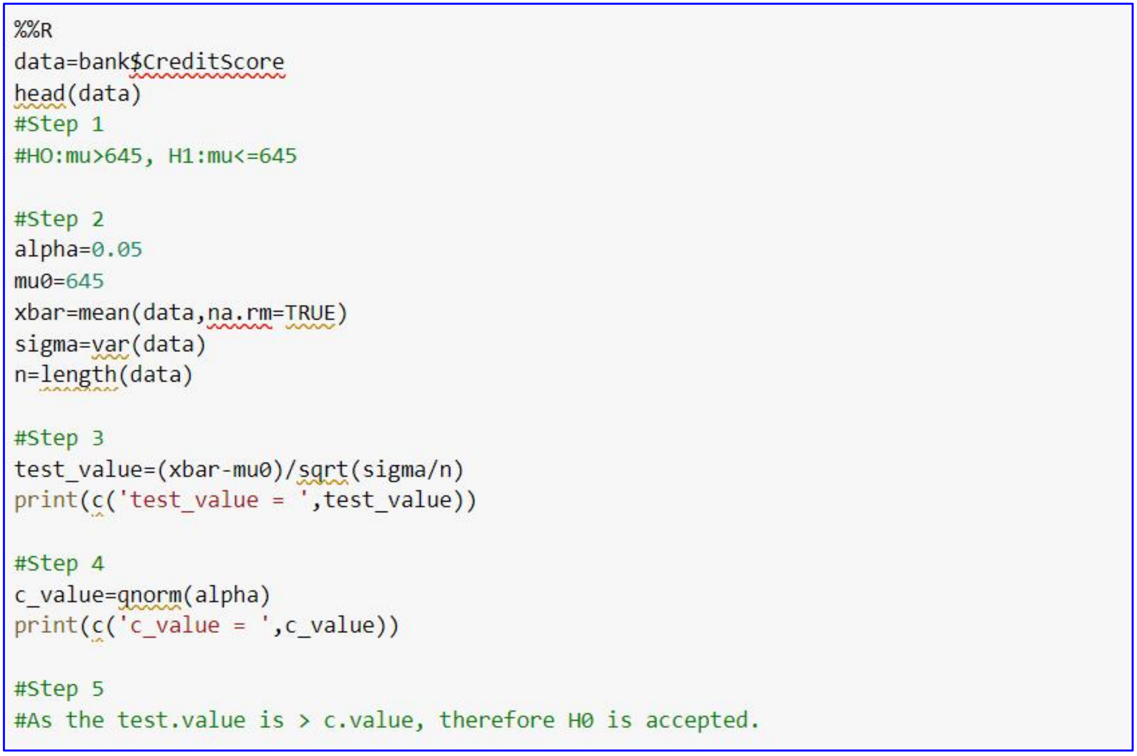
ii) lower tailed test
1: H0: p >= p0; H1: p < p0
2: alpha = 0.05
3: test_value = (X-p0)/(s/sqrt(n) X: sample mean; s: Standard deviation
4: c_value = qnorm(alpha)
5: if test_value <= c_value ----> Reject H0

iii) upper one tailed test

1: H0: p <= p0; H1: p > p0
2: alpha = 0.05
3: test_value = (X-p0)/(s/sqrt(n) X: sample mean; s: Standard deviation
4: c_value = qnorm(alpha)
5: if test_value => c_value ----> Reject H0

Example: - We have applied "Lower One-Sided Hypotheses Testing of Mean" on "CreditScore" data and analysed whether the mean of Credit scores level is greater than 645 at the 5% level of significance.

Step 1: State the hypotheses
H0: μ > μ0
H1: μ ≤ μ0
Step 2: Set the significance level
α = 0.05
Step 3: Compute the test.value
Step 4: Find the c.value
c.value = qnorm(α)
Step 5: Specify the decision rule
if test.value ≤ c.value therefore H0 is rejected
Step 6: Make a decision and conclusion
If the condition in step 5 is satisfied, then the population mean is at most equal to μ0.

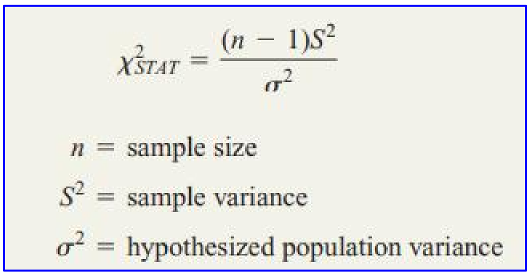


```
%%R
data=bank$CreditScore
head(data)
#Step 1
#HO:mu>645, H1:mu<=645

#Step 2
alpha=0.05
mu0=645
xbar=mean(data,na.rm=TRUE)
sigma=var(data)
n=length(data)

#Step 3
test_value=(xbar-mu0)/sqrt(sigma/n)
print(c('test_value = ',test_value))

#Step 4
c_value=qnorm(alpha)
print(c('c_value = ',c_value))

#Step 5
#As the test.value is > c.value, therefore H0 is accepted.
```

Result: -
After getting the results of the test, we interpret that the average of "CreditScore" is greater than 645 for our case at 5% significance level.

b) Test of Variance:

This module calculates the sample size and performs power analysis for hypothesis tests concerning a single variance.



$$\chi^2_{STAT} = \frac{(n-1)S^2}{\sigma^2}$$

n = sample size

S^2 = sample variance

σ^2 = hypothesized population variance

(sigma2 as the parameter of interest) ----- one population

i) two tailed test
   1: H0: sigma2=sigma2_0; H1: sigma2! = sigma2_0
   2: alpha =0.05
   3: test_value= (n-1) *s2/sigma2_0 ; s2: sample variance
   4: c_value1 = qchisq(alpha/2); c_value2 = qchisq(1-alpha/2)
   5: if test_value in (c_value1, c_value2) ------> accept H0 otherwise reject it

ii) upper one tailed test
   1: H0: sigma2 =< sigma2_0; H1: sigma2 > sigma2_0
   2: alpha =0.05
   3: test_value= (n-1) *s2/sigma2_0; s2: sample variance
   4: c_value2 = qchisq(1-alpha)
   5: if test_value < c_value2 ------> accept H0 otherwise reject it

iii) lower one tailed test
   1: H0: sigma2 >= sigma2_0; H1: sigma2 < sigma2_0
   2: alpha =0.05
   3: test_value= (n-1) *s2/sigma2_0; s2: sample variance
   4: c_value1 = qchisq(alpha)
   5: if test_value > c_value1 ------> accept H0 otherwise reject it

Example: We have applied "Two-Sided Hypotheses Testing of Variance for Two Populations on "CreditScore" and "EstimatedSalary" columns. We aim to test whether there is a difference between the variance at the 5% significance level

Let p, p̂, n be respectively population proportion, sample proportion and sample size.

Step 1: State the hypotheses

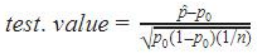
H_0: $p > p_0$

H_1: $p \le p_0$

Step 2: Set the significance level

α = 0.02

Step 3: Compute the test.value



$$test.\ value = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)(1/n)}}$$

Step 4: Find the c.value

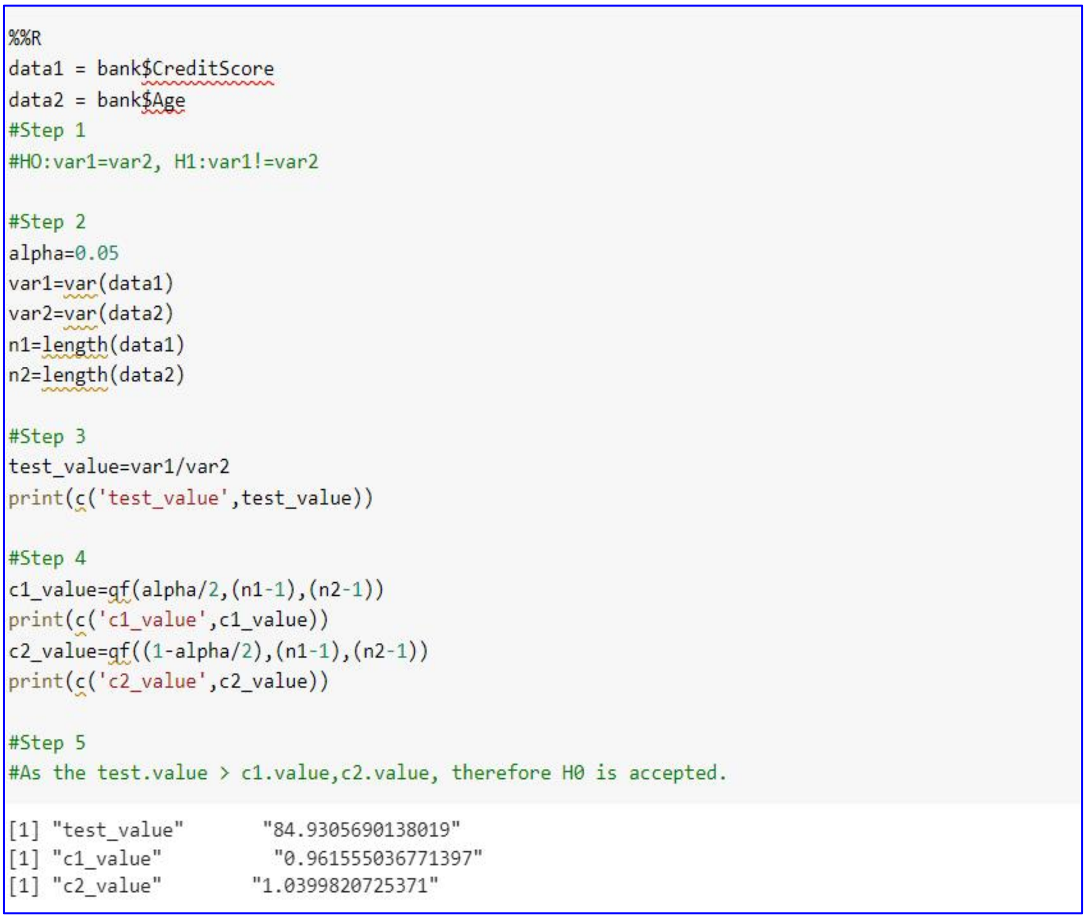
c.value = qnorm(α)

Step 5: Specify the decision rule

if *test.value* ≤ *c.value* therefore H_0 is rejected

Step 6: Make a decision and conclusion

If the condition in step 5 is satisfied, then the population proportion is at most p_0 at the significance level α.



```
%%R
data1 = bank$CreditScore
data2 = bank$Age
#Step 1
#H0:var1=var2, H1:var1!=var2

#Step 2
alpha=0.05
var1=var(data1)
var2=var(data2)
n1=length(data1)
n2=length(data2)

#Step 3
test_value=var1/var2
print(c('test_value',test_value))

#Step 4
c1_value=qf(alpha/2,(n1-1),(n2-1))
print(c('c1_value',c1_value))
c2_value=qf((1-alpha/2),(n1-1),(n2-1))
print(c('c2_value',c2_value))

#Step 5
#As the test.value > c1.value,c2.value, therefore H0 is accepted.
```

```
[1] "test_value"       "84.9305690138019"
[1] "c1_value"          "0.961555036771397"
[1] "c2_value"        "1.0399820725371"
```

Result: -

After getting the results of the test, we can clearly say that the variances of “CreditScore” and “EstimatedSalary” data are not equal at 5% significance level.

c) Test of Proportional:

A test of proportion will assess whether or not a sample from a population represents the true proportion from the entire population.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

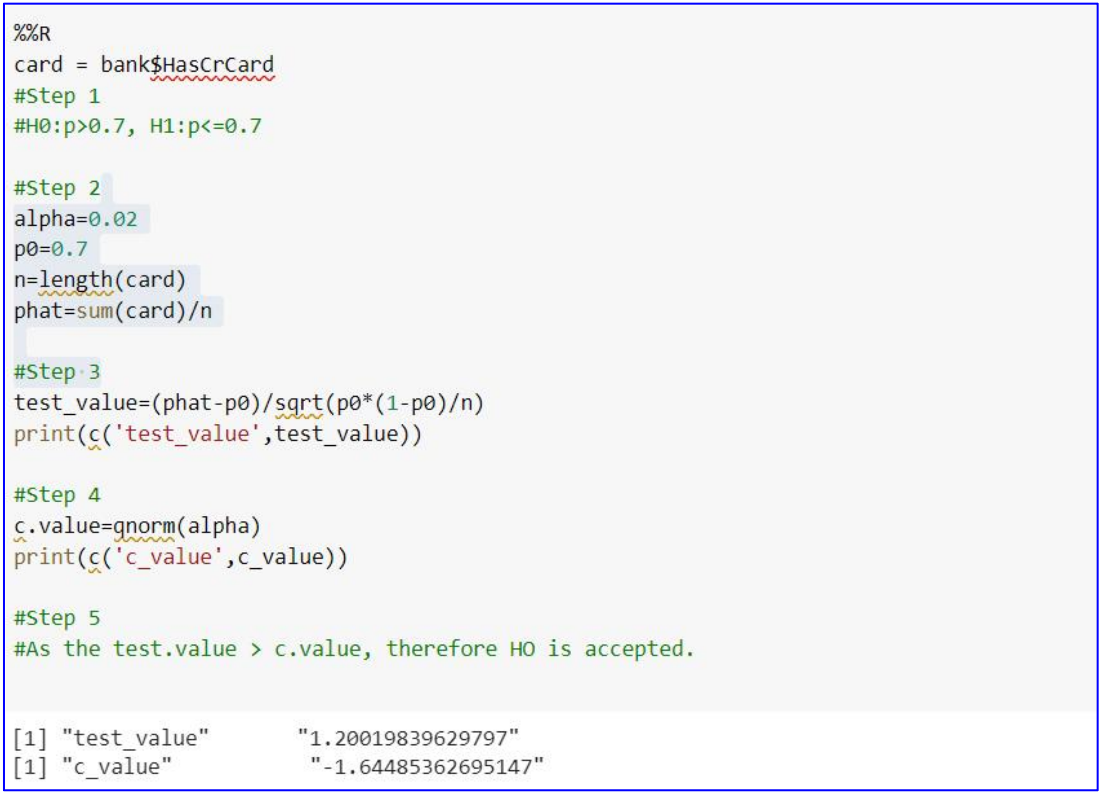
```
Test Of proportion(p) --- one population
    i) two tailed test
        1: H0: p=p0;   H1: p! = p0
        2: alpha = 0.05
        3: test_value = (phat-p0)/sqrt(p0*(1-p0)/n); phat: the ratio of attribute in sample
        4: c_value = qnorm(1-alpha/2)
        5: if abs(test_value) > c_value ----> Reject H0

    ii) lower tailed test
        1: H0: p >= p0; H1: p < p0
        2: alpha = 0.05
        3: test_value = (phat-p0)/sqrt(p0*(1-p0)/n); phat: ratio of sample
        4: c_value = qnorm(alpha)
        5: if test_value < c_value ----> Reject H0

    iii) upper one tailed test
        1: H0: p <= p0; H1: p > p0
        2: alpha = 0.05
        3: test_value = (phat-p0)/sqrt(p0*(1-p0)/n); phat: ratio of sample
        4: c_value = qnorm(1-alpha)
        5: if test_value > c_value ----> Reject H0
```

Example: We have applied “Lower One-Sided Hypotheses Testing of Proportion on our binary variable column: "HasCrCard". We want to test whether the ratio of ones in the "HasCrCard" data is greater than 50% at the level of 2% significance level.

Step 1: State the hypotheses
H0: p > p0
H1: p ≤ p0
Step 2: Set the significance level
α = 0.02
Step 3: Compute the test.value

Step 4: Find the c.value
c.value = qnorm(α)
Step 5: Specify the decision rule

if test.value ≤ c.value therefore H0 is rejected
Step 6: Make a decision and conclusion
If the condition in step 5 is satisfied, then the population proportion is at most p0 at the significance level α.



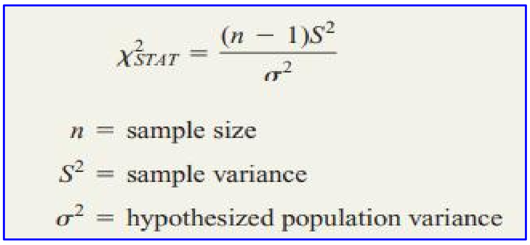
```
%%R
card = bank$HasCrCard
#Step 1
#H0:p>0.7, H1:p<=0.7

#Step 2
alpha=0.02
p0=0.7
n=length(card)
phat=sum(card)/n

#Step 3
test_value=(phat-p0)/sqrt(p0*(1-p0)/n)
print(c('test_value',test_value))

#Step 4
c.value=qnorm(alpha)
print(c('c_value',c_value))

#Step 5
#As the test.value > c.value, therefore HO is accepted.


[1] "test_value"       "1.20019839629797"
[1] "c_value"           "-1.64485362695147"
```

Result: -

Based on the findings of the hypothesis testing we performed above; we can say that the proportion of ones in "HasCrCard" data is greater than 70%.

4.3. Test of Variance:

This module calculates the sample size and performs power analysis for hypothesis tests concerning a single variance.



$$\chi^2_{STAT} = \frac{(n-1)S^2}{\sigma^2}$$

n = sample size

S^2 = sample variance

σ^2 = hypothesized population variance

(sigma2 as the parameter of interest) ----- one population

iv) two tailed test

1: H0: sigma2=sigma2_0; H1: sigma2! = sigma2_0
2: alpha =0.05
3: test_value= (n-1) *s2/sigma2_0 ; s2: sample variance
4: c_value1 = qchisq(alpha/2); c_value2 = qchisq(1-alpha/2)

5: if test_value in (c_value1, c_value2) ------> accept H0 otherwise reject it

v) upper one tailed test
1: H0: sigma2 =< sigma2_0; H1: sigma2 > sigma2_0
2: alpha =0.05
3: test_value= (n-1) *s2/sigma2_0; s2: sample variance
4: c_value2 = qchisq(1-alpha)
5: if test_value < c_value2 ------> accept H0 otherwise reject it

vi) lower one tailed test
1: H0: sigma2 >= sigma2_0; H1: sigma2 < sigma2_0
2: alpha =0.05
3: test_value= (n-1) *s2/sigma2_0; s2: sample variance
4: c_value1 = qchisq(alpha)
5: if test_value > c_value1 ------> accept H0 otherwise reject it

Example: We have applied “Two-Sided Hypotheses Testing of Variance for Two Populations on “CreditScore” and “EstimatedSalary” columns. We aim to test whether there is a difference between the variance at the 5% significance level

Let p, p̂, n be respectively population proportion, sample proportion and sample size.

Step 1: State the hypotheses

H_0: $p > p_0$
H_1: $p \leq p_0$

Step 2: Set the significance level

α = 0.02

Step 3: Compute the test.value

$$test.\,value = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)(1/n)}}$$

Step 4: Find the c.value

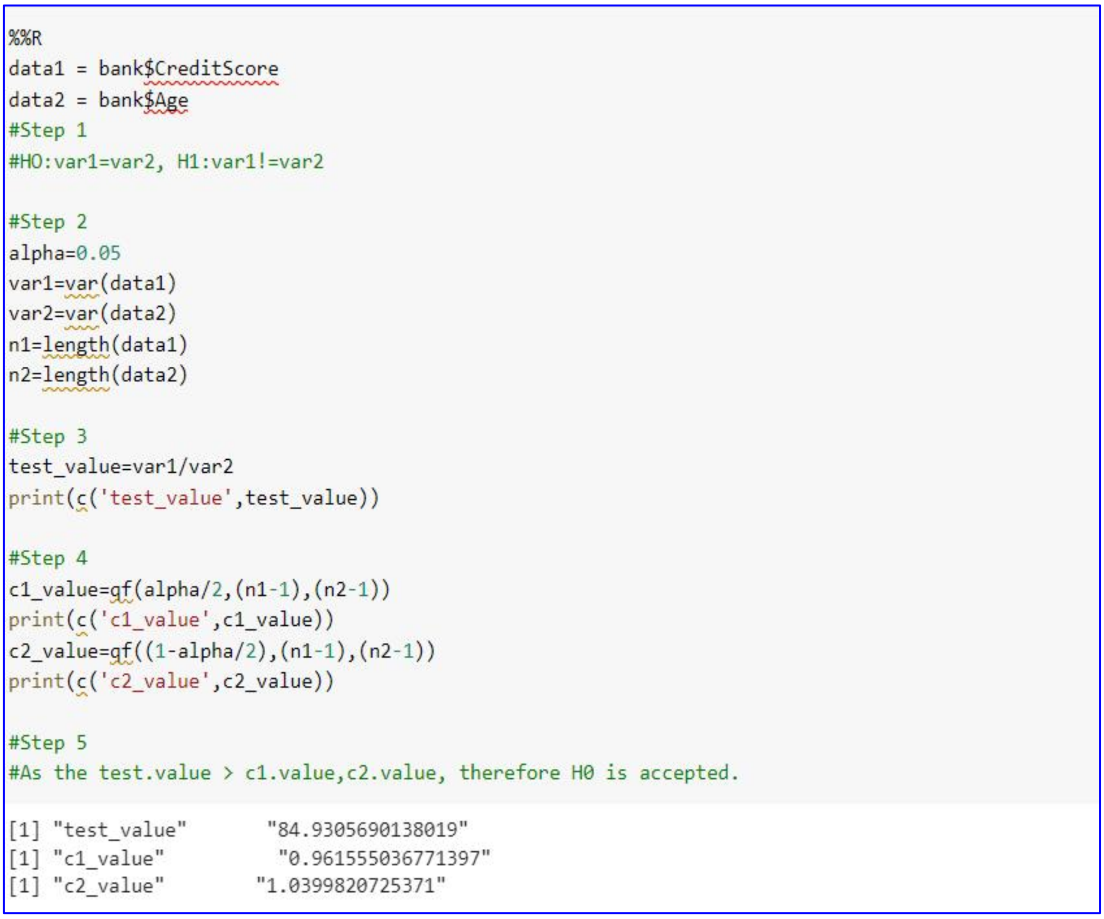
c.value = qnorm(α)

Step 5: Specify the decision rule

if *test.value* ≤ *c.value* therefore H_0 is rejected

Step 6: Make a decision and conclusion

If the condition in step 5 is satisfied, then the population proportion is at most p_0 at the significance level α.



```
%%R
data1 = bank$CreditScore
data2 = bank$Age
#Step 1
#H0:var1=var2, H1:var1!=var2

#Step 2
alpha=0.05
var1=var(data1)
var2=var(data2)
n1=length(data1)
n2=length(data2)

#Step 3
test_value=var1/var2
print(c('test_value',test_value))

#Step 4
c1_value=qf(alpha/2,(n1-1),(n2-1))
print(c('c1_value',c1_value))
c2_value=qf((1-alpha/2),(n1-1),(n2-1))
print(c('c2_value',c2_value))

#Step 5
#As the test.value > c1.value,c2.value, therefore H0 is accepted.
```

```
[1] "test_value"       "84.9305690138019"
[1] "c1_value"          "0.961555036771397"
[1] "c2_value"        "1.0399820725371"
```

Result: -
After getting the results of the test, we can clearly say that the variances of “CreditScore” and “EstimatedSalary” data are not equal at 5% significance level.

d) HT of Independence for Two Categorical Variables (p.value approach)

Since applying p-value approach takes less time and needs less steps to be followed by, it is one of the most efficient ways to perform hypotheses testing.
Example: - Test whether the probability of each class of the 'Tenure' variable in the 'Bankers' dataset is same at level alpha = 0.05
The Chi-square test of independence is a statistical hypothesis test used to determine whether two categorical or nominal variables are likely to be related or not.
Example : we will use p.value approach to test whether “CreditScore” and “Age” are independent at 0.02 significance level. The steps used for computing the respected values are given below -

Let X_1 be the first categorical variable, and X_2 be the second categorical variable.
Step 1: State the hypotheses
H_0: X_1 and X_2 are independent
H_1: X_1 and X_2 are dependent

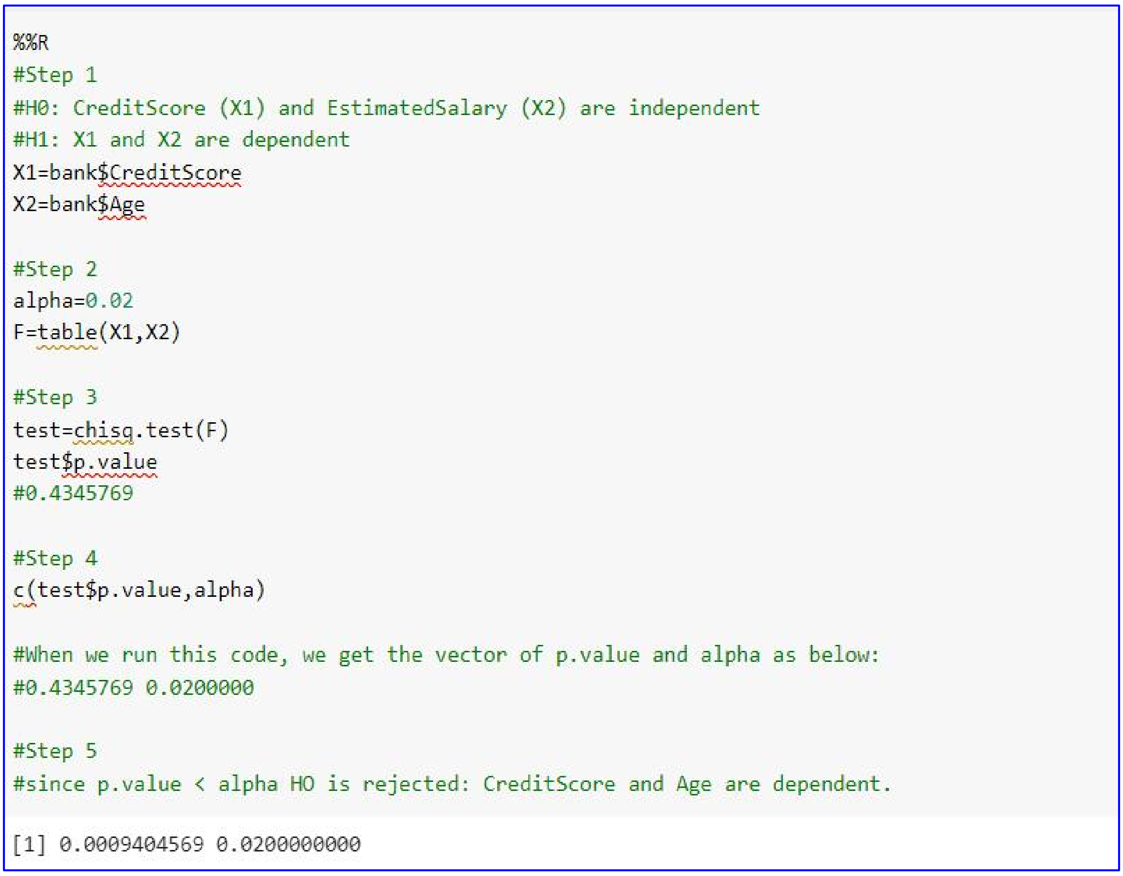
Step 2: Set the significance level
α = 0.02

Step 3: Find the p.value
Let *F* be the frequency table of classes for the variables.
Then the p.value is computed in Python as:
test=chisq.test(F)
test$p.value

Step 4: Compare the p.value and alpha
(p.value, alpha)

Step 5: Specify the decision rule
if *p.value* < *alpha* therefore H_0 is rejected

Step 6: Make a decision and conclusion
If the condition in step 5 is satisfied, then two categorical variables are dependent at the significance level α.



```
%%R
#Step 1
#H0: CreditScore (X1) and EstimatedSalary (X2) are independent
#H1: X1 and X2 are dependent
X1=bank$CreditScore
X2=bank$Age

#Step 2
alpha=0.02
F=table(X1,X2)

#Step 3
test=chisq.test(F)
test$p.value
#0.4345769

#Step 4
c(test$p.value,alpha)

#When we run this code, we get the vector of p.value and alpha as below:
#0.4345769 0.0200000

#Step 5
#since p.value < alpha HO is rejected: CreditScore and Age are dependent.
```

```
[1] 0.0009404569 0.0200000000
```

Result: -
At 0.02 significance level our hypotheses test shows CreditScore and Age variables are correlated, in other words dependent.

8. Appendix:

➢ Appendix A : Individual contributions

Initially, we had setup weekly recurring meetings to discuss about the approach.

- Determined the dataset which could help us understand the details about statistical methods and their concepts.
- As per the basic concepts of Statistics, Numerical Measures are the core part of summarizing the data. Each group member had initially spent time to understand and implement the concepts.
- The data was visualized with the help of graph plotting and this helped the group members to get a detailed understanding of visual representation of dataset and it’s variables.
- Probablity distribution was divided between the group members and knowledge sharing sessions were conducted to explain the distribution techniques.
- The group members decided on which probablity models we should implement and also which data variables should be considered for each distribution technique
- Aniket, Prathamesh and Vaibhav analyzed and performed Bernoulli , Poisson and Normal Distribution respectively and then implemented the code for the same. After the knowledge sharing sessions, each distribution technique was cross validated by group members.
- All members also had to explain why they used this specific model for the selected data variables.
- Hypothesis Testing was broken down into multiple steps, to determine if the Test of Mean, Test of Variance, Test of Proportion and Test of Independence provided the expected results as the outcome.
- Each module had a specific timelines/deadlines for understanding and implementing the learnings.

➢ Appendix B : Dataset

- Researching and selecting the appropriate dataset for the assessment.
- Banker’s Dataset [online] Kaggle.com.
  Available at : https://www.kaggle.com/ahmedmohamedmahrous/churn-modelling1

9. References

Bankers Data [online] Kaggle.com. Available at:
https://www.kaggle.com/ahmedmohamedmahrous/churn-modelling1

https://study.dbs.ie/2122/msc-data/B9DA101/u2/index.html#/

ScienceDirect., 2017 Poisson Distribution <https://www.sciencedirect.com/topics/mathematics/poisson-distribution>>[Accessed 29 June 2021].

Rao, T., 2019. Different Types Of Probability Distribution (Characteristics & Examples) - Databasetown. [online] DatabaseTown. Available at: <https://databasetown.com/types-probability-distribution-characteristics-examples/> [Accessed 29 June 2021].

Advanced Analytical Models, 2015. Understanding and Choosing the Right Probability Distributions. pp.899-917. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119197096.app03>[Accessed 29 June 2021].

Lisa, S., 2017. Hypothesis Testing For Means & Proportions. [online] Sphweb.bumc.bu.edu. Available at: <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTest-Means-Proportions/BS704_HypothesisTest-Means-Proportions_print.html> [Accessed 29 June 2021].

Statistical Software, Chptr 650.Tests for One Variance https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Tests_for_One_Variance.pdf [Accessed 29 June 2021].

https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/

https://spmaddmaths.blog.onlinetuition.com.my/2013/10/7-1a-mean.html

https://mathworld.wolfram.com/BernoulliDistribution.html
https://www.simplilearn.com/tutorials/data-science-tutorial/bernoulli-distribution
https://www.statisticshowto.com/bernoulli-trials/