**Statistics using R**

**Olatunbosun Akinade**

**D00252416**

The data set use for this project are information collected on a small sample of customer that purchase 6 types of products from a business. The data set contain the customer information on 1995 observations which was original 2250 but 4 row was omitted as the income for these rows were x6 the largest income from the data set used which was 98777 and 251 where duplicate data sets.

The information in the data set contained Year of birth, Education, Marital status, Income, Number of children, the product types they purchase from the business (Wine, Fruit, Meat, Fish, Sweet), the method that was used for purchase by the customer (Catalog, Store and Web) and the number of deals claimed by each customer.

The variable in the data set are Year\_Birth, Education, Marital\_Status, Income, Children, Wines\_Products, Fruits\_Products, Meat\_Products, Fish\_Products, Sweet\_Products, PurchasesFromCatalog, PurchasesFromStore, PurchasesFromWeb, Deals\_Claimed

|  |  |
| --- | --- |
| **Variable Type** | |
| **Categorical** | Year\_Birth, Education |
| **Numerical Discrete** | Children, Deals\_Claimed |
| **Numerical Continuous** | Marital\_Status, Income, Wines\_Products, Fruits\_Products, Meat\_Products, Fish\_Products, Sweet\_Products, PurchasesFromCatalog, PurchasesFromStore, PurchasesFromWeb |

The research questions that will be Annalise in the projected are:

What variable have the most impact on the number of products bought.

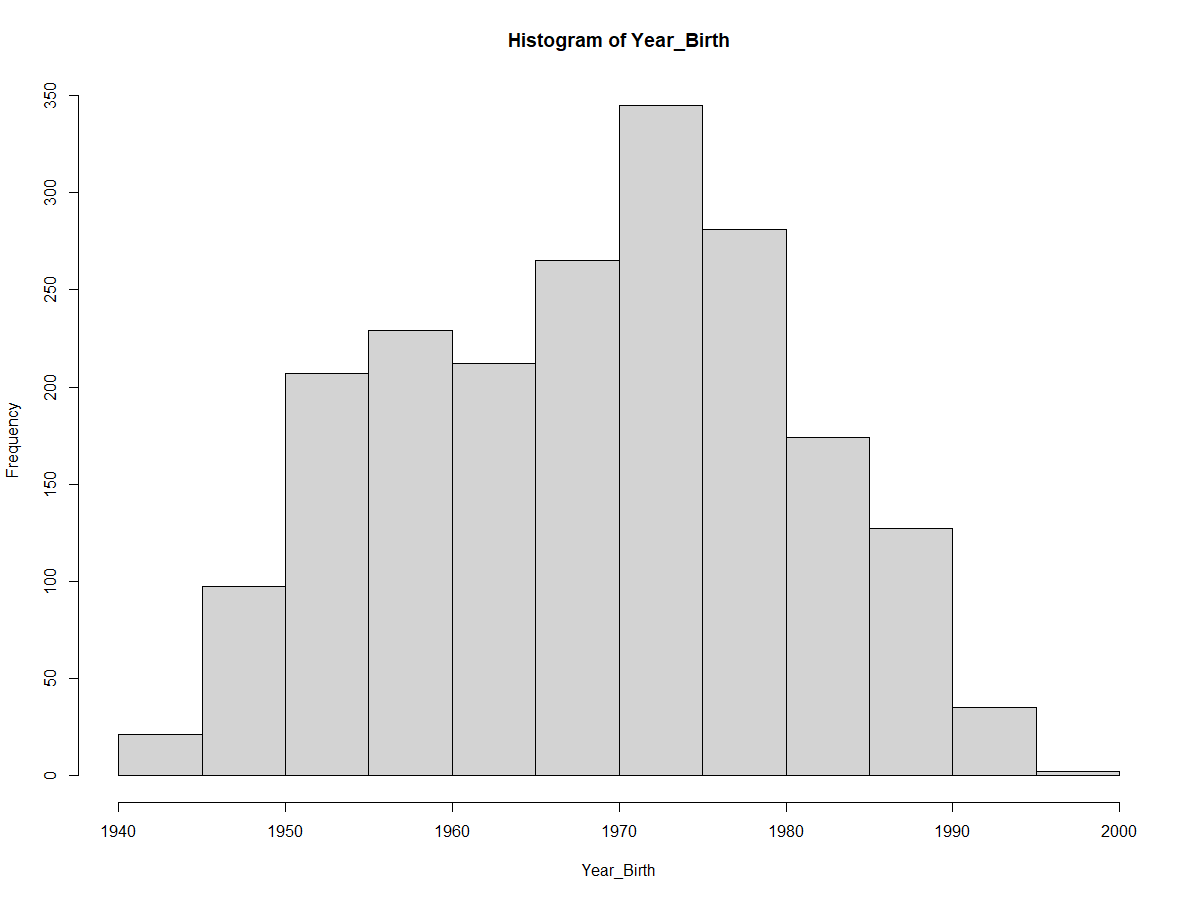
What method of purchase is the most used by the customers?

Does the number of children influence the amount of product being purchased?

Does the number of incomes influence the amount of product being purchased?

Does the deal influence the amount of product bought?

* Year of Birth



mean (Year\_Birth) : 1968.9

median (Year\_Birth): 1970

The year of birth of the sample customers have a normal distribution with the mean of 1968.9

* Education

Chart, bar chart

Description automatically generated

Proportion

Basic: 0.02456140, Graduate: 0.50426065 Master: 0.16541353 PhD: 0.21503759 Secondary: 0.09072682

The customer education is most made up off graduate, which takes up 0.5 % of the population.

Basic level of education was the least represented in the sample population of 0.02% of the population

* Marital Status

Chart, bar chart

Description automatically generated

Proportion

Divorced: 0.10526316 Married: 0.38997494 Single: 0.21804511 Together: 0.25213033 Widow:0.03458647

The marital status that is most represented by the sample population are married at 0.4% of the population, customer follow by single and together marital service with widow being the least represented.

* Income

Chart, box and whisker chart

Description automatically generated

The income boxplot shows a normalize spread with the median of 51373, upper quarter of 98777 and lower quarter of 1730.

* Children

Chart, histogram

Description automatically generated

Proportion

0: 0.27919799 1: 0.51027569 2: 0.18796992 3: 0.02255639

Most of the customer have 1 or no children, with customer with 1 child being more represented

* Product

Chart

Description automatically generated

The product being buy the most are the wine products with sweet product being bought the least,

The boxplot all shows skewed population with outliners present in all product groups.

* Deal Claimed

Chart, box and whisker chart

Description automatically generated

Deals most of the customer only claim deal the total of 2 times with outliers going up to 15

Purchase Methods

Chart, box and whisker chart

Description automatically generated

Most of the product are sold in store (46%) follow by website Purchases (33%) then catalog purchases (21%)

Marital\_Status VS Incomes

Chart, box and whisker chart

Description automatically generated

All income shows a normal distribution spread by marital status group with overlapping distributions but with widow having a higher median.

Marital\_Status VS Children

Chart, bar chart

Description automatically generated

Married couple have the most children among he group.

Income vs Products

Qr code

Description automatically generated

The Pair graph shows a correlation between income and the amount of product bought.

Children VS Product

Diagram, engineering drawing, qr code

Description automatically generated

The Pair graph shows no correlation between children and the amount of product bought.

Deals claimed Vs Product bought

Diagram, engineering drawing, qr code

Description automatically generated

The amount of product bought decreases with the number of deals claimed.

Education Vs Income

Chart, box and whisker chart

Description automatically generated

The plot shows a normal distribution between Education Vs Income,

Education Vs Income shows a influence of education on the income as the customer with higher level of education earn more the secondary and extremely more than the basic.

**#1 sample hypothesis tests to for means####**

#Ho : The income mean of the Data sample is the income mean of population that buys from the business

#H1: The income mean of the Data sample is not the income mean of population that buys from the business

#Assumptions of the one-sample

#Independent random sample = True

#Large sample size = True

#Normally distributed population:

hist(Income)

#=true

#Standard deviation of the population known= False

#this mean A T-Test will be done

#N = 2196

#mean = 51484.11

Income\_mean=mean(Income)

#sample\_SD = 20584.98

Income\_sample\_SD = sd(Income)

t\_income = t.test(Income, mu=51484.11)

t\_income$p.value

t\_income$conf.int

#p.value > 0.05 we fail to reject the H0

#> t\_income$p.value

#[1] 0.9999929

#> t\_income$conf.int

#[1] 50622.67 52345.54

#attr(,"conf.level")

#[1] 0.95

t.test(Income, mu=51484.11, alternative="two.sided"

**#1 sample hypothesis tests to for proportions####**

#Ho= product are sold in store makes up 46% of the total population purchases

#H1= product are sold in store does not makes up 46% of the total population purchases

#Ho: p = 0.46

#hi: p not = 0.46

#p is the true portion of product brought from the store

sum\_wp=sum(NumWebPurchases)

sum\_wp

sum\_cp=sum(PurchasesFromCatalog)

sum\_cp

sum\_sp=sum(NumStorePurchases)

sum\_sp

purchase\_no=sum(sum\_wp,sum\_cp,sum\_np)

purchase\_no

p = 0.46

n = purchase\_no

n

phat = sum\_sp /n

#np>=10, n(1-p)>=10

n\*p

n\*(1-p)

#Assumptions:

#Observations are independent = True

#Sample size is large enough = True

SE=sqrt(p\*(1-p)/n)

SE

z= (phat-p)/SE

z

#since 2-sided, it can be more extreme in both directions

#P(z<-1.331762) or P(z>1.331762)

p.value<-2\*pnorm(z)

p.value: 1.658765

#p.value>0.05, fail to reject H0.

**#2 sample hypothesis tests to for means independent sample####**

# Research Question:

#Does customer with Married and Together marital statues have the same income mean?

#Null hypothesis: the population mean income mean is the same in both groups

#H0: µ1 = µ2

#or equivalently

#H0: µ1 - µ2 = 0

#Alternative hypothesis: the population mean income mean is not the same in both groups

#H1: µ1 ¹ µ2

#or equivalently

#H0: µ1 - µ2 ¹ 0

# µ1, µ2 = the population mean with marital statues Married and those with marital statues Together, respectively

income\_married= Income[marital\_status=="Married"]

income\_married

hist(income\_married)

income\_Together= Income[marital\_status=="Together"]

income\_Together

hist(income\_Together)

boxplot(income\_Together,income\_married)

#The sample

sd(income\_married)

sd(income\_Together)

var(income\_married)

var(income\_Together)

#Assumption = Meet

#Two samples must be independent and random = True

#The underlying populations must not be skewed (normally distributed)=True

#The two population must be equal spread=True

t.test(income\_married, income\_Together)

#p-value=0.98>0.05, therefore fail to reject, and based on data, no evidence to suggest a difference.

#95% confidence interval, 95% the true population mean difference for Income between Married and Together marital statues lies between -2177.326 and 2141.297

#Confidence interval also leads to the same conclusion as it does cover 0

**#2 sample hypothesis tests to for means/proportions independent sample####**

#H0: no.of children and martial status are independent of each other

#H1: no.of children and martial status are dependent of each other

marital\_childrens\_table= table(marital\_status,Childrens)

marital\_childrens\_chisq<-chisq.test(marital\_childrens\_table)

str(marital\_childrens\_chisq)

marital\_childrens\_chisq$expected

marital\_childrens\_chisq$p.value

###looking expected values I have enough enough data to

#meet my assumptions as expected values >=5

marital\_childrens\_chisq$p.value

#p-value=0.038 < 0.05, no.of children and martial status are independent of each other are disease independent.