Statistical Inference: Case Study

Session Pedagogy

- Data Understanding
- ***Business Objective given***
- Class Discussion to
  - Conceptualize the Output
  - Approach the Analysis in R
- ***Participants Create the R code independently***
- Final code and output shown

DATA SCIENCE INSTITUTE

Background

- Large FMCG company
- Pan country presence
- 3 business lines- Ice Cream, Chocolates and Cakes-Biscuits
- Large amount of data is available

Supply Chain

‘Customer’ in our case study

Data Snapshots

Customer Profile Data

CUSTID	REGION
10000	North
10001	South
10002	West
10003	South
10004	East
10005	West
10006	West
10007	South
10008	East

NPS Data

CUSTID	NPS
22929	6
12089	5
19120	6
10155	8
13085	9
14784	4
17714	5
19735	6
23779	2
10477	9
22958	3
21552	9
10594	3

Market survey data is available for 107 customers(NPS Data). The survey recorded 'Net promoter Score' .

Net Promoter Scores are based on response to single question (0-10 scale)
How likely is it that you would recommend [brand] to a friend or colleague?

Get Started

- Import data sets: CUST_PROFILE and NPSDATA.
- CUST_PROFILE data has custid and region.
- NPSDATA has custid and Net Promoter Score measured on 0-10 scale.
- Import and check dimensions and number of unique customers in each data set.
- Merge two data sets.

Get Started- R codes and output

```
custprofile<-read.csv(file.choose(),header=T)
npsdata<-read.csv(file.choose(),header=T)

dim(custprofile)
[1] 15001     2

dim(npsdata)
[1] 107   2

length(unique(custprofile$CUSTID))
[1] 15001

 length(unique(npsdata$CUSTID))
[1] 107
```

DATA SCIENCE INSTITUTE

Get Started- R codes and output
Merging Two Data Sets

```
npsregion<-merge(npsdata,custprofile,by="CUSTID",all.x=T)
head(npsregion)
```

```
    CUSTID  NPS  REGION
1  10155     8      South
2  10211     7      West
3  10271     7      North
4  10477     9      South
5  10535     8      West
6  10564     7      South
```

DATA SCIENCE INSTITUTE

1. Assessing Net Promoter Score

- What is the Net Promoter Score on an average?

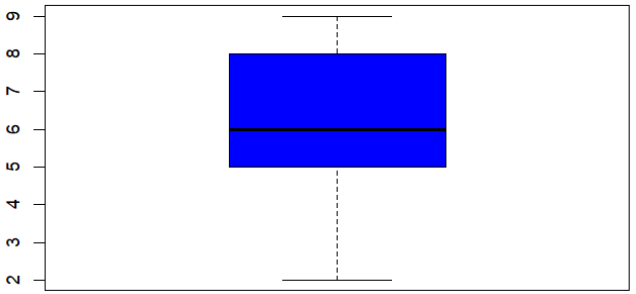
  Suggestion: Use median as well as mean since the measurement scale is ordinal.(median is better measure)

- Describe NPS graphically.

- Suggestion: Use Box-Whisker plot or bar chart to plot median

Assessing Net Promoter Score
R Code and Output

```
> round(length(npsregion$NPS),0)
[1] 107

> round(mean(npsregion$NPS),2)
[1] 6.2

> round(median(npsregion$NPS),2)
[1] 6

boxplot(npsregion$NPS,col="blue")
```



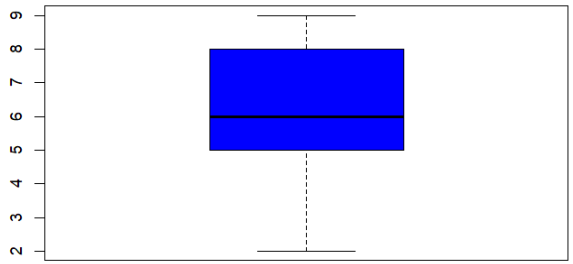
DATA SCIENCE INSTITUTE

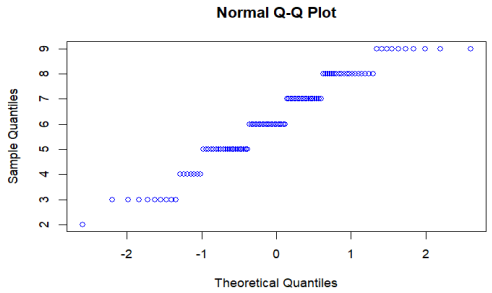
2. Is NPS significantly more than '6'?

- Which test to be used?
- Can we assume 'Normality' of the distribution?
- Is NPS significantly more than '6'?

Can we assume 'Normality' of the distribution?
R Code and Output

```
boxplot(npsregion$NPS,col="blue")
qqnorm(npsregion$NPS,col="blue")

shapiro.test(npsregion$NPS)

Shapiro-Wilk normality test
data:  NPS
W = 0.94709, p-value = 0.000326

library(nortest)
lillie.test(npsregion$NPS)
 Lilliefors (Kolmogorov-Smirnov) normality test
data:  NPS
D = 0.12501, p-value = 0.0002978
```





DATA SCIENCE
INSTITUTE

Is NPS significantly more than '6'?
R Code and Output

```
wilcox.test(npsregion$NPS,mu=6,alternative="greater")

Wilcoxon signed rank test with continuity correction

data:  NPS
V = 2166, p-value = 0.0989
alternative hypothesis: true location is greater than 6
```

Conclusion: Do not reject H0. Average NPS is not significantly more than '6'.

DATA SCIENCE INSTITUTE

3. Compare NPS Region-wise

- Which region has on an average highest NPS?

  Suggestion: Use mean as well as median.

- Is region wise difference in NPS significantly different?

- Present region wise NPS graphically.

Which region has on an average highest NPS?
R Code and Output

```
aggregate(NPS~REGION,data=npsregion,FUN=mean)
REGION      NPS
1   East      6.250000
2  North      6.208333
3  South      6.057143
4. West       6.321429

aggregate(NPS~REGION,data=npsregion,FUN=median)
 REGION NPS
1   East     6
2  North     6
3  South     6
4   West     7
```

DATA SCIENCE INSTITUTE

Is region wise difference in NPS significantly different?

R Code and Output

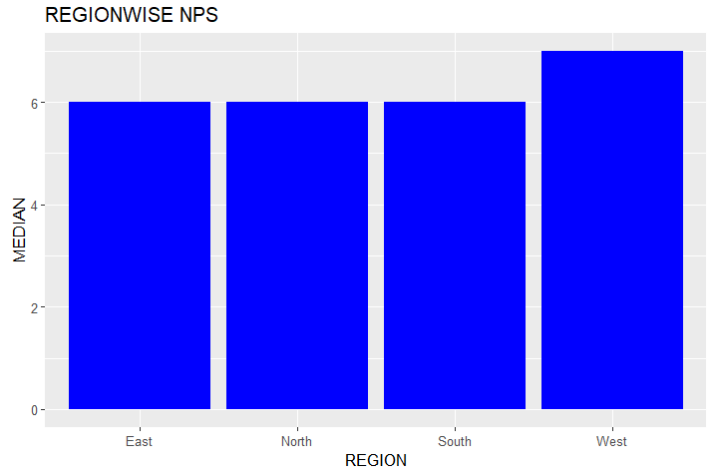
```
kruskal.test(NPS~REGION,data=npsregion)

 Kruskal-Wallis rank sum test

data:  NPS by REGION
Kruskal-Wallis chi-squared = 0.52336, df = 3, p-value = 0.9137
```

Region wise difference in NPS is not significant.

DATA SCIENCE INSTITUTE

Median NPS -Bar Diagram

```
med<-aggregate(NPS~REGION,data=npsregion,FUN=median)
library(ggplot2)
ggplot(med,aes(x=REGION,y=NPS))+
geom_bar(stat="identity",fill="blue")+
labs(x="REGION",y="MEDIAN",title="REGIONWISE NPS")
```



DATA SCIENCE INSTITUTE

4. Detractors vs. Non-detractors

- What percentage of customers are 'detractors'?

  Detractor: NPS score of less than or equal to 6

  Suggestion: Derive a new variable 'detractor' having values 'YES' or 'NO'.

- Is percentage of detractors significantly greater than 40%?

What percentage of customers are 'detractors'?
R Code and Output

```
npsregion$detractor[npsregion$NPS<=6]<-"YES"
npsregion$detractor[npsregion$NPS>6]<-"NO"
head(npsregion)
t<-table(npsregion$detractor)
t
```

```
detractor
 NO YES
 48  59
```

```
prop.table(t)
detractor
      NO       YES
0.4485981 0.5514019
```

55% are detractors

DATA SCIENCE INSTITUTE

Is percentage of detractors significantly greater than 40%?
R Code and Output

```
prop.test(t["YES"],sum(t),0.4,alternative='greater')

1-sample proportions test with continuity correction

data:  t["YES"] out of sum(t), null probability 0.4
X-squared = 9.5985, df = 1, p-value = 0.0009737
alternative hypothesis: true p is greater than 0.4
95 percent confidence interval:
 0.4673912 1.0000000
sample estimates:
      p
0.5514019
```

% of detractors significantly more than 40%.

DATA SCIENCE INSTITUTE

5. Association between region and detractor(Y/N)

- Summarize 'detractor' by region.

- Test for association between region and detractor(Y/N)

  Suggestion: Use 'gmodels' package.

Testing Association
R Code and Output

```
library(gmodels)
CrossTable(npsregion$REGION, npsregion$detractor,prop.c=FALSE,prop.t=FALSE,
           prop.chisq=FALSE,chisq=TRUE)


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  1.711052     d.f. =  3     p =  0.6344795
```

No association between Region and 'detractor'

DATA SCIENCE INSTITUTE