HW\_4\_Template

Lina Cao , Thursdays 1:15PM

SDGB 7844; Prof. Nagaraja; Fall 2017

## Question 1

Because they can use the small sample size to test the effect of mail advertising, including whether the advertising method gains, how much the company can earn from the mail advertising and even predict the revenue if entire name list is been used, etc. And by not buying the whole list at first, they can avoid further loss on asking price ($3 per name) if the advertising does not work.

## Question 2

The holder does earn from the small amount of purchase. In addition, a small sample can be more easily to get a satisfying result, therefore his potential purchaser can understand the whole mail advertising process and see the effect of it, which is also an approach to promote his "product". If the potential purchaser comes back to gain the list, the holder will earn more, however, if not, the holder at least have the revenue from the sample list price.

## Question 3

The null hypothesis is that the gross dollar sale (gross profit) generated by direct mailing is less or equal to 3 dollar per name.(However, in H0, I simply use equal 3) The alternative hypothesis is that the gross profit generated by mailing is greater than 3 dollar per name. i.e. H0: mu=3 H1: mu>3 mu here means the true gross dollar of sale earned by direct mail advertising with the whole list.

## Question 4

Population | sales generated from every name (and address) on the whole list

Parameter | mu, the true average sales in population

Sample | sales generated from randomly sampled 225 name (and address) on the list

Statistic | sample mean of sales, but once we use t distribution to do the test, we will calculate the test | statistic, which is t=(sample mean-3)/sd of sample

## Question 5

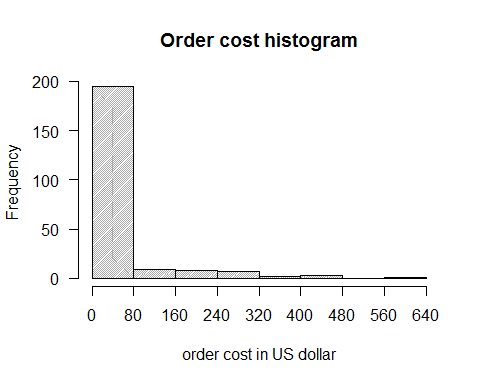
Type I error in this case means that the sales is truely 3 dollar, or below, but we reject the null hypothesis. The probability of it is represented by alpha (0.05). Here in this case, the probability of making Type I error is 0.05.

## Question 6

### read the excel file  
# install.packages("openxlsx")  
library(openxlsx)

## Warning: package 'openxlsx' was built under R version 3.4.2

setwd("F:/0000Fordham New Life/4 SM&C.statistics methods and computation I/HW4")  
aaa <- read.xlsx("direct\_mail.xlsx")  
### define order cost data  
ccc <- as.numeric(aaa$order\_cost)  
### (a).construct the histogram  
hist(ccc, breaks = seq(from = 0, to = 640, by = 80),  
 main = "Order cost histogram", xlab = "order cost in US dollar",  
 axes = FALSE,  
 col = "darkgray", angle = 45, density = 65, border = "black")  
axis(side=1,at = seq(from=0,to=640,by=80))  
axis(side=2,at = seq(from=0,to=200,by=50),las = TRUE)



### (b).summary statistics  
# minimum  
min(ccc)

## [1] 0

# median  
median(ccc)

## [1] 0

# mean  
mean(ccc)

## [1] 33.66436

# max  
max(ccc)

## [1] 638.92

# standard deviation  
sd(ccc)

## [1] 97.01957

# fill them in a table  
sss <- data.frame(min(ccc),median(ccc),mean(ccc),max(ccc),sd(ccc),stringsAsFactors = FALSE)  
library(knitr)  
kable(sss)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| min.ccc. | median.ccc. | mean.ccc. | max.ccc. | sd.ccc. |
| 0 | 0 | 33.66436 | 638.92 | 97.01957 |

### (c).compute fraction  
# number of people bought nothing  
zero <- length(ccc[ccc==0])  
# fraction  
zero/length(ccc)

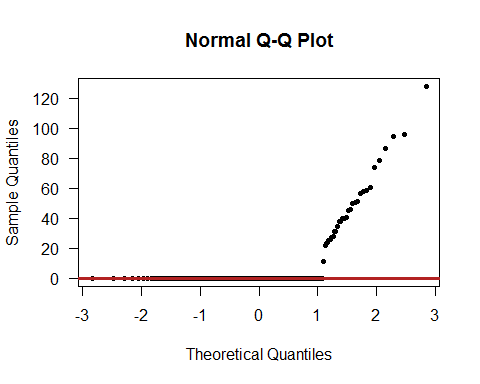
## [1] 0.8622222

rm(zero)

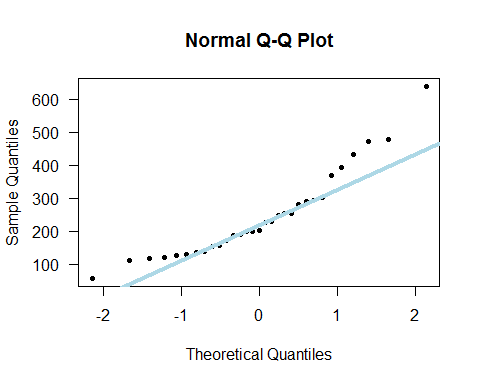
1. From the plot, order cost (sales for the company) data is extremely right skewed. Sales within 0~100 dollar take up most of the dataset.
2. The minimun and median of the cost is 0 dollar. The mean of the cost is 33.66 dollar. The maximum of the cost is 638.92 dollar. The standard deviation is 97.02 dollar.
3. 86.2% people among the sample bought nothing from Performance Tires.

## Question 7

# the second assumption  
# check the q-q plot to see if it is normal distribution  
# the data should be multiplied by 0.2, since the result is the profit for the company  
qqnorm(ccc\*0.2, pch=20, las=TRUE)  
# also add line to check  
qqline(ccc\*0.2,col="firebrick",lwd=3)



# if the zeroes are omitted  
qqnorm(ccc[ccc!=0], pch=20, las=TRUE)  
qqline(ccc[ccc!=0], col="lightblue",lwd=4)



The assumptions of the one-sample t-test are: (source: <https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/One-Sample_T-Test.pdf>)

1. The data are continuous (not discrete).

2. The data follow the normal probability distribution.

3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

The profits that can be earned are continuous number, the first assumption is satisfied. The 225 names are randomly selected, which also meets the third one, "equal probability to be chosen" assumption.

As for the second one, although from the first plot the sample data is not normally distributed, we can still believe the data satisfy normal distribution. Because firstly, the whole name list must have more than 225 names, which is a large size to be regard as normal distribution. Additionally, the situation is caused by too many zeroes in the sample data. The non-zero data is more linear correlated. But the slightly "s-shape" indicates a heavy tailed distribution if only consider the 31 non-zero data. Therefore, when we consider the amount in a large data size, I think this data satisfied for assumption 2.

## Question 8

# data vector : ccc\*0.2, because the 20% of each sales is their profit  
# alternative : from question 6, H1 is mu>3, set "greater"  
# mu : from H0, mu=3  
# confidence level : 0.05, therefore conf.level=1-0.05=0.95  
t.test(ccc\*0.2,alternative = "greater",mu=3,conf.level=0.95)

##   
## One Sample t-test  
##   
## data: ccc \* 0.2  
## t = 2.8857, df = 224, p-value = 0.002144  
## alternative hypothesis: true mean is greater than 3  
## 95 percent confidence interval:  
## 4.596261 Inf  
## sample estimates:  
## mean of x   
## 6.732871

# check 0.95 quantile of the t distribution, with its degree of freedom 224  
qt(p=0.95, df=224)

## [1] 1.651685

The test statistic is the standardized sample mean profit, which is calculated by the formula: t=(mean(cost\**0.2)-3)/sd(cost\**0.2).

The sample size is 225, therefore the degree of freedom is 225-1, i.e 224. From the result, the t statistic is 2.8857, which is much larger than the 95% quantile of t(224) distribution; also, the p-value of the one sample t-test, 0.0021, is much smaller than 0.05, therefore we should reject the null hypothesis mu=3 and choose the alternative hypothesis. Thus, from the test we conclude that the average sales from the direct mailing is greater than 3 dollar per name.

The result turns out to be positive to use the whole name list, I think the Performance Tires should do further advertising using the whole name and address list.