

## Exam (Long Answer)

Question 1: Consider the RealEstate.csv data set.

- Plot the distribution of the price per unit area variable.
- Describe the shape of the distribution.
- Compute an estimate of the mean price per unit area.
- Bootstrap 10000 sample means of the price per unit area. Save the bootstrapped means to a vector called `boot_means`.
- Plot the sampling distribution to see if it appears to be normally distributed.
- Use the `qnorm` function (consider the distribution of the sample mean) to compute a 95% confidence interval for the mean.
- Now, imagine you are not sure the distribution is normal. Use the quantile function to compute a 95% confidence interval for the mean price per unit area.

# (a) Plot the distribution of the price per unit area variable.

# Answer (a) below:

```
RE = read.csv(file =
"/Users/itagakikouki/stat123/lab11/RealEstate.csv", header = TRUE)
dim(RE)
```

```
## [1] 414 8
```

```
head(RE)
```

```
##      No X1.transaction.date X2.house.age
```

X3.distance.to.the.nearest.MRT.station

```
## 1 1          2012.917          32.0
```

84.87882

```
## 2 2          2012.917          19.5
```

306.59470

```
## 3 3      2013.583      13.3
```

561.98450

##	4	4	2013.500	13.3
----	---	---	----------	------

561.98450

```
## 5 5      2012.833      5.0
```

390.56840

##	6	6	2012.667	7.1
----	---	---	----------	-----

2175.03000

```
##      X4.number.of.convenience.stores X5.latitude X6.longitude
```

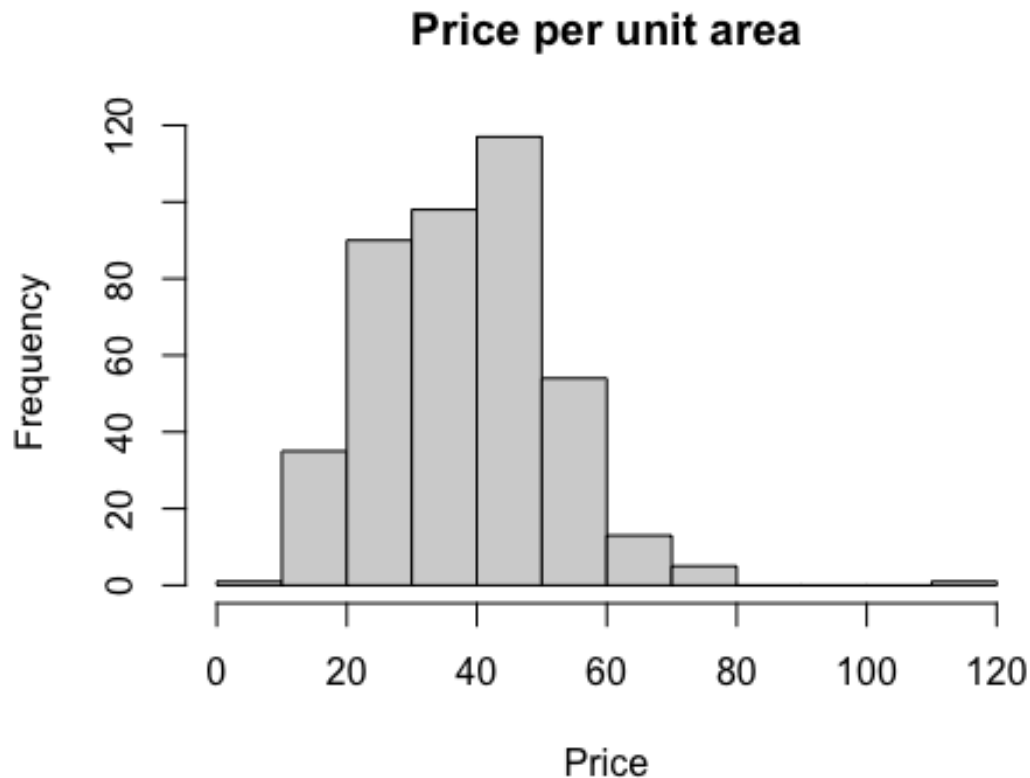
##	X1.V.Name	Conv	ConvLen	ConvCS	X2.V.Name	Conv	ConvLen	ConvCS
## 1		10	24.98298				121.5402	

## 2	9	24.98034	121.5395
------	---	----------	----------

```
## 3          5    24.98746    121.5439
## 4          5    24.98746    121.5439
## 5          5    24.97937    121.5425
## 6          3    24.96305    121.5125
## Y.house.price.of.unit.area
## 1          37.9
## 2          42.2
## 3          47.3
## 4          54.8
## 5          43.1
## 6          32.1
```

*#Since numerical*

```
hist(RE$Y.house.price.of.unit.area, main = "Price per unit area", xlab =
"Price")
```



*# (b) Describe the shape of the distribution.*

*# Answer (b) below:*

*#This is right-skewed*

```

# (c) Compute an estimate of the mean price per unit area.

# Answer (c) below:
xbar = mean(RE$Y.house.price.of.unit.area)
xbar

## [1] 37.98019

# (d) Bootstrap 10000 sample means of the price per unit area. Save the
# bootstrapped means to a vector called boot_means.

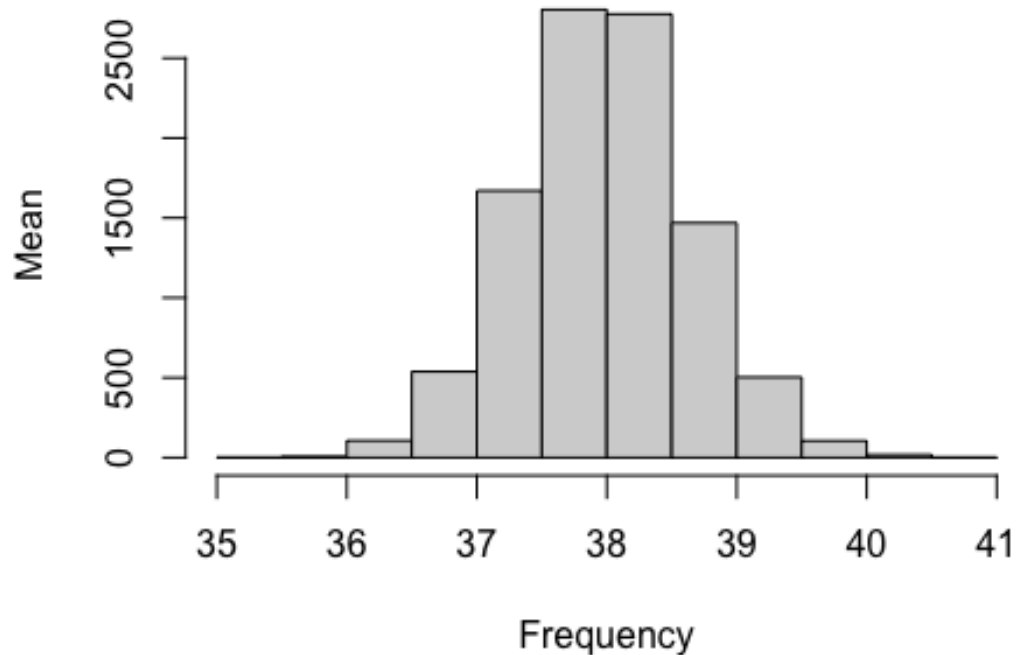
# Answer (d) below:
n = length(RE$Y.house.price.of.unit.area)
# Declare boot_mean as numeric vector with length 10000
boot_mean = numeric(length = 10000)
for(i in 1:10000){
  # Since replace is true, we can sample the same value twice.
  temp_sample = sample(RE$Y.house.price.of.unit.area ,n ,replace = TRUE)
  boot_mean[i] = mean(temp_sample)
}

# (e) Plot the sampling distribution.

# Answer (e) below:
hist(boot_mean,main = "The histogram of 10000 sample means of the price per
unit area"
      ,xlab = "Frequency", ylab = "Mean")

```

e histogram of 10000 sample means of the price per unit



```
# (f) Use the quantile function to compute a 95% confidence interval for the
mean
# price per unit area.
```

```
# Answer (f) below:
quantile(boot_mean)
```

```
##      0%      25%      50%      75%     100%
## 35.19469 37.54173 37.98116 38.42929 40.78889
```

```
sd<-sd(boot_mean)
ese = sd/sqrt(n)
critical_val = qnorm(boot_mean)
```

```
## Warning in qnorm(boot_mean): NaNs produced
```

```
lower_ci<-xbar - critical_val*ese
upper_ci<-xbar + critical_val*ese
```

```
cat(paste("The confidence interval is [",lower_ci, ",",upper_ci))
```

```
## The confidence interval is [ NaN , NaN The confidence interval is [ NaN ,
NaN The confidence interval is [ NaN , NaN The confidence interval is [ NaN ,
```

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



```
# (g) Use a different method (which considers the distribution of the sample mean)
# to compute a 95% confidence interval for the mean.
```

```
quantile(boot_mean,c(0.025,0.975))
```

```
##      2.5%      97.5%
## 36.69153 39.28628
```

Question 2: Load the media\_spend.csv dataset into R and save it to df. Dataset contains information about a fictitious company that's trying to determine how much money to spend on various types of advertising for the coming year. They have historical data showing sales (in millions of dollars) and the amount they spent on TV, Radio, and Newspaper advertising that year (in thousands of dollars). The goal is to determine which of the types of advertising effects sales the most. We are trying to see how the advertising effects sales, therefore, sales is our response variable and the other columns are our regressors.

- Plot the response variable (as the y-axis) against each of the regressor variables (one plot for each regressor). Use the `par(mfrow = c())` function so that all the plots are displayed at once.
- Looking only at the plots, which type of advertising do you think will have the largest effect on sales?
- Perform a linear regression for each form of advertising vs the response variable, sales.
- Print out the summary for each of these regressions and take note of the p-value for the t-test on the significance of the coefficient for each. Which of the regressors is the most significant?
- Create a vector of these p-values and name each element with the corresponding type of advertising

```
# (a) Plot the response variable (as the y-axis) against each of the regressor variables
# (one plot for each regressor). Use the par(mfrow = c()) function so that all the plots
# are displayed at once
```

```
# Answer (a) below:
```

```
df<- read.csv(file =
"/Users/itagakikouki/stat123/lab11/media_spend.csv",header = TRUE)
head(df)
```

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2   22.1
## 2  44.5  39.3      45.1   10.4
```

```
## 3  17.2  45.9      69.3  12.0
## 4 151.5  41.3      58.5  16.5
## 5 180.8  10.8      58.4  17.9
## 6   8.7  48.9      75.0   7.2

par(mfrow = c(2,2))
par(mar = c(4.8,4.5,5,2.1))
colo<-c("blue","yellow", "green")
cname<-c("TV","Radio","Newspaper")
for(i in 1:3){
  title<-paste("Sales vs", cname[i])
  plot(df[,i],df[,4],col = colo[i], xlab = cname[i], ylab = "Sales", main =
        title)
}
mtext("SAles vs Media type",outer = TRUE,line = -1.5)
```

*# (b) Looking only at the plots, which type of advertising do you think will have the  
# largest effect on sales?*

*# Answer (b) below:  
#According to the graph, TV have the*

*# (c) Perform a linear regression for each form of advertising vs the  
response variable,  
# sales.*

*# Answer (c) below:*

```
fit_news<-lm(Sales~Newspaper +TV + Radio, data = df)
```

*# (d) Print out the summary for each of these regressions and take note of  
the p-value for # the t-test on the significance of the coefficient for each.  
Which of the regressors is  
# the most significant?.*

*# Answer (d) below:  
summary(fit\_news)*

```
##
## Call:
## lm(formula = Sales ~ Newspaper + TV + Radio, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.3034 -0.8244 -0.0008  0.8976  3.7473
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.6251241   0.3075012   15.041  <2e-16 ***
## Newspaper    0.0003357   0.0057881    0.058   0.954
## TV           0.0544458   0.0013752   39.592  <2e-16 ***
## Radio        0.1070012   0.0084896   12.604  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.662 on 196 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9011
## F-statistic: 605.4 on 3 and 196 DF,  p-value: < 2.2e-16

# (e) Create a vector of these p-values and name each element with the
# corresponding type
# of advertising

# Answer (e) below:

res<-numeric(3)
names(res)<- c("Newspaper","TV","Radio")
res[1]<- summary(fit_news)$coefficients[2,4]
res[2]<- summary(fit_news)$coefficients[3,4]
res[3]<- summary(fit_news)$coefficients[4,4]

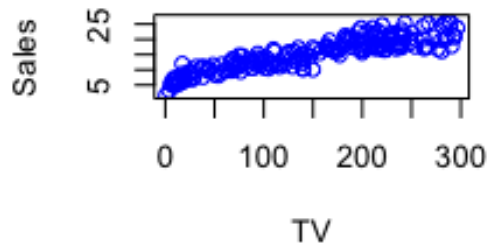
#res[2]<-summary(lm(Sales ~TV,data = df))$coefficients[2,4]
#res[3]<-summary(lm(Sales ~Radio,data = df))$coefficients[2,4]

print(res)

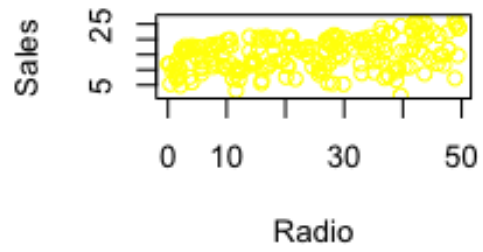
##      Newspaper          TV          Radio
## 9.538145e-01 1.892945e-95 4.602097e-27
```

## Sales vs Media type

### Sales vs TV



### Sales vs Radio



### Sales vs Newspaper

