# Project 1

Chloe Chen & Kangrui Liu

9/8/2023

## 1. Use `hprice` in `faraway` package.

```
#prepare data set
library("faraway")
data("hprice")
dim(hprice)
```

```
## [1] 324   8
```

```
hprice$homeprice<-exp(hprice$narsp)*1000
```

### 1) What are the mean and the variance of homeprice? What do they mean?

```
mean(hprice$homeprice)
```

```
## [1] 94411.42
```

```
var(hprice$homeprice)
```

```
## [1] 1583110349
```

- The mean is 94411.42 and the variance is 1583110349.
- The terms "mean" and "variance" refer to the average of a data set and the deviation of a data point from the mean, respectively.

## 2) Construct a 95% confidence interval of the average homeprice. What does the confidence interval imply?

```
n<-dim(hprice)[[1]] # find the number of homeprice
samvar<-var(hprice$homeprice)/(n-1)  # sampling var
samvar
```

```
## [1] 4901270
```

```
t.score<-qt(p=.05/2, df=n-1, lower.tail=F)
t.score
```

```
## [1] 1.967336
```

```
lowCI <- mean(hprice$homeprice)-t.score*sqrt(samvar)
upCI <- mean(hprice$homeprice)+t.score*sqrt(samvar)
print(c(lowCI,upCI))
```

```
## [1] 90055.97 98766.87
```

```
t.test(hprice$homeprice, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  hprice$homeprice
## t = 42.711, df = 323, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  90062.70 98760.14
## sample estimates:
## mean of x
##  94411.42
```

- The 95% confidence interval of $\mu_{homeprice}$ is $[90055.97, 98766.87]$

## 3) Estimate the average homeprice by whether the MSA was adjacent to a coastline, noted in ajwtr, and the standard errors.

```r
subset1 <- subset(hprice, hprice$ajwtr == 1)# yes
subset2 <- subset(hprice, hprice$ajwtr == 0)# no
mean1 <- mean(subset1$homeprice)
mean2 <- mean(subset2$homeprice)
se1 <- sd(subset1$homeprice)/sqrt(nrow(subset1))
se2 <- sd(subset2$homeprice)/sqrt(nrow(subset2))
```

- We estimate the average homeprice by the MSA was adjacent to a coastline is 90055.97 and the standard error of it is 4655.88.
- We estimate the average homeprice by the MSA was not adjacent to a coastline is 82388.89 and the standard error of it is 1228.66.

## 4) Test the difference in homeprice between coastline MSAs and non-coastline MSAs. Clearly state the formula for the hypothesis, the test method and your rationale for selecting the method. What do you conclude about the hypothesis?

- Step1: Check means and variances

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
hprice%>%
  group_by(ajwtr)%>%
  summarize(m=mean(homeprice),
            v=var(homeprice))
```

```
## # A tibble: 2 x 3
##   ajwtr       m           v
##   <fct>   <dbl>       <dbl>
## 1 0      82389.  285315565.
## 2 1     111243. 2926428411.
```

- $\hat{\mu}_{ajwtr=0} = \hat{\bar{y}}_{ajwtr=0} = 82388.89$ and $\sigma^2_{ajwtr=0} = s^2_{ajwtr=0} = 285315565$
- $\hat{\mu}_{ajwtr=1} = \hat{\bar{y}}_{ajwtr=1} = 111242.96$ and $\sigma^2_{ajwtr=1} = s^2_{ajwtr=1} = 2926428411$
- Step2: Test equal variance ( with F test)
  Corresponding hypothesis: $H_0 : \sigma^2_{ajwtr=0} = \sigma^2_{ajwtr=1}$ vs. $H_A : \sigma^2_{ajwtr=0} \neq \sigma^2_{ajwtr=1}$.

```
var.test(homeprice ~ ajwtr, hprice, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  homeprice by ajwtr
## F = 0.097496, num df = 188, denom df = 134, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.07088604 0.13297389
## sample estimates:
## ratio of variances
##          0.09749617
```

- The range of 95 percent confidence interval is $[0.07088604, 0.13297389]$, and the ratio of variances is 0.09749617. Therefore, We conclude that $H_0 : \sigma^2_{ajwtr=0} = \sigma^2_{ajwtr=1}$ does not hold.
- Step 3: Conduct proper testing(t-test)

```
t.test(homeprice ~ ajwtr, hprice, var.equal=FALSE, conf.int = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  homeprice by ajwtr
## t = -5.9922, df = 152.79, p-value = 1.43e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not e
## 95 percent confidence interval:
```

```
##  -38367.19 -19340.96
## sample estimates:
## mean in group 0 mean in group 1
##        82388.89       111242.96
```

- We reject $H_0 : \mu_{=1} = \mu_{=0}$, meaning that the homeprice is not the same for those living on different coastlines.

## 5) Estimate the Pearson correlation coefficient between homeprice and per capita income of the MSA of a given year, noted in ypc.

```
cor(hprice$homeprice, hprice$ypc, method="pearson")
```

```
## [1] 0.7437474
```

- The Pearson correlation coefficient between homeprice and ypc is 0.7437474.

## 6) Test whether the correlation coefficient between homeprice and ypc is 0 or not. Clearly state the hypothesis including the formula. What do you conclude?

```
cor.test(hprice$homeprice, hprice$ypc, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  hprice$homeprice and hprice$ypc
## t = 19.965, df = 322, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6907661 0.7887854
## sample estimates:
##       cor
## 0.7437474
```

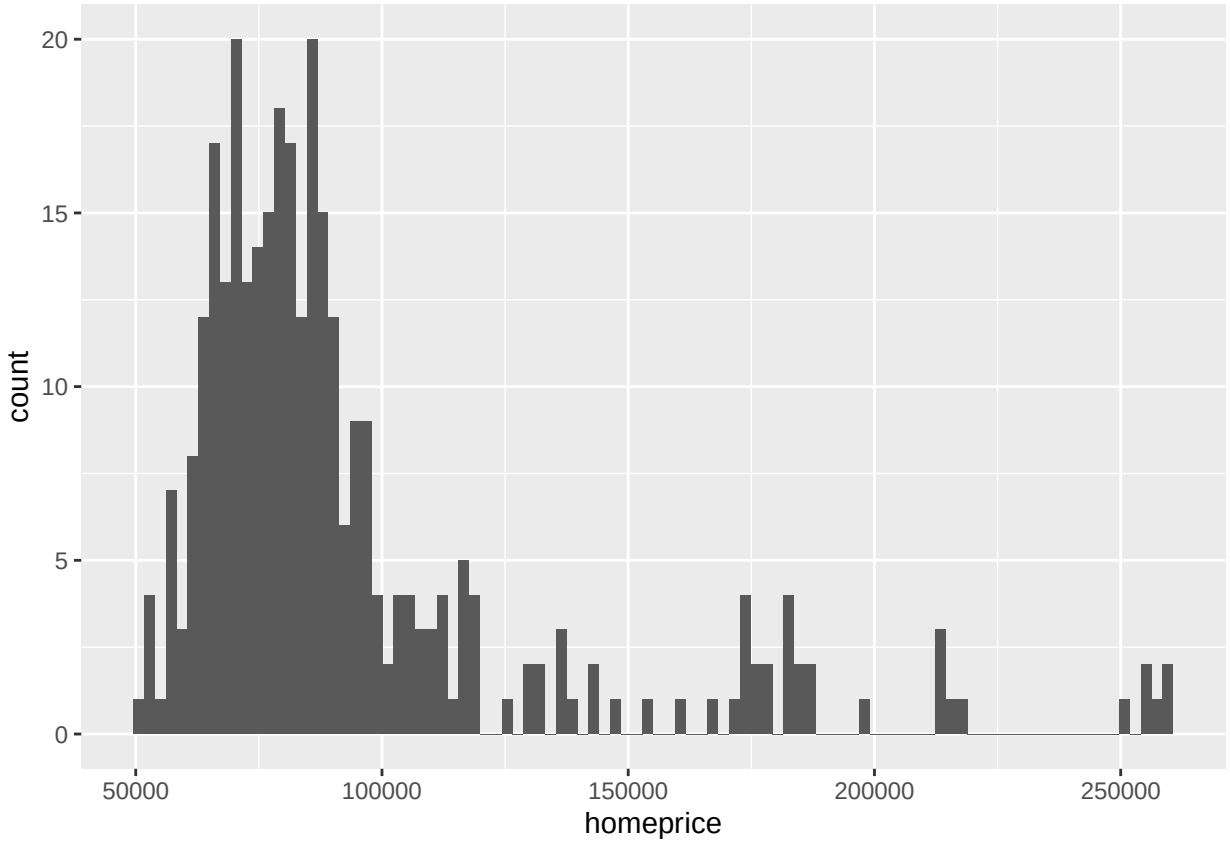
- Corresponding hypothesis: $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$.
- $\bar{\rho} = 0.7437474$, and 95% CI is $[0.6907661, 0.7887854]$, So reject $H_0$.

## 7) Can you say that per capita income has an effect on the home sales price using the results from #6)? Why or why not?

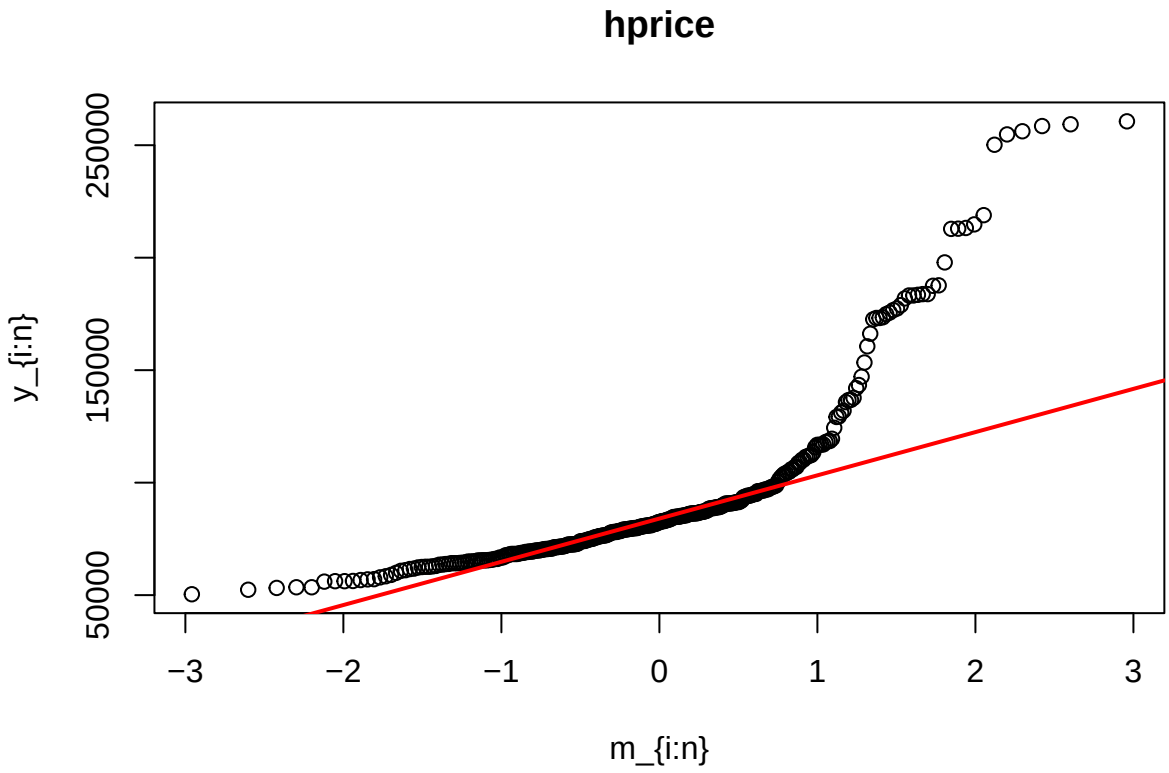
- The fact that two variables A and B have a significant correlation coefficient does not imply that A has a causal effect on B. Correlation measures the strength and direction of a linear relationship between two variables, but it does not establish causation. Causation would require further research or causal inference methods to determine whether one variable directly influences the other. Therefore, significant correlation alone does not imply a causal relationship.

## 8) Test the normality of homeprice. Would this test result change your responses to #1) to 7)? Why or why not?

```
library(ggplot2)
ggplot(hprice, aes(x=homeprice)) + geom_histogram(binwidth=2200)
```




```
qqnorm(hprice$homeprice, main="hprice", ylab="y_{i:n}", xlab="m_{i:n}")
qqline(hprice$homeprice, col="red",lwd=2)
```




```
shapiro.test(hprice$homeprice)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hprice$homeprice
## W = 0.7327, p-value < 2.2e-16
```

- In Shapiro-Wilk normality test, p-value $< 0.05$, and from the histogram, as well as the qq-plot, `homeprice` does not appear to follow a normal distribution.But this wouldn't change the responses to the previous questions, because this still make sense with a sufficiently large sample(N=324), according to CLT.