Anmol More (11915043)

Please use table of contents (PDF Bookmarks) for easy navigation

```
house_prices <- read.csv('houseprices.csv')
summary(house_prices)
```

```
##      Price         Living.Area     Bathrooms        Bedrooms   
##  Min.   : 16858   Min.   : 672   Min.   :1.000   Min.   :1.000  
##  1st Qu.:112014   1st Qu.:1336   1st Qu.:1.500   1st Qu.:3.000  
##  Median :151917   Median :1672   Median :2.000   Median :3.000  
##  Mean   :163862   Mean   :1807   Mean   :1.918   Mean   :3.183  
##  3rd Qu.:205235   3rd Qu.:2206   3rd Qu.:2.500   3rd Qu.:4.000  
##  Max.   :446436   Max.   :4534   Max.   :4.500   Max.   :6.000  
##     Lot.Size           Age           Fireplace     
##  Min.   :0.0000   Min.   :  0.00   Min.   :0.0000  
##  1st Qu.:0.2100   1st Qu.:  6.00   1st Qu.:0.0000  
##  Median :0.3900   Median : 18.00   Median :1.0000  
##  Mean   :0.5696   Mean   : 28.06   Mean   :0.5931  
##  3rd Qu.:0.6000   3rd Qu.: 34.00   3rd Qu.:1.0000  
##  Max.   :9.0000   Max.   :247.00   Max.   :1.0000  
```

Part (A)

1. a) Your friend claims that the average house price in this area is above $150K. Do you agree ? Briefly explain what the p-values in these cases mean ?

Let, μ be average house prices in area

null hypothesis, $H_0 : \mu \leq \$150K$

alternate hypothesis, $H_a : \mu > \$150K$

Population mean is known, Sample standard deviation, s is known and sample size is known. Since sample size is large enough, greater than 30 we use Test Statistic, t here.

p-value is probability of getting a stronger evidence to reject null hypothesis. Here, it can be given by probability of seeing sample mean of more than $163K, if population mean, μ is $150K

```
n <- length(house_prices$Price)
s = sd(house_prices$Price)
x_bar = mean(house_prices$Price)
mu = 150000

print(paste("Sample Mean :", x_bar))
```

```
## [1] "Sample Mean : 163862.125119389"
```

```
t <- (x_bar -  mu)/(s/sqrt(n))
p_value <- 1 - pt(t, n-1)
print(paste("p value :", p_value))
```

```
## [1] "p value : 2.68138844461419e-11"
```

Since, p-value < 0.05 we reject the null. We agree with the claim that house prices are aove \$150K at 5% significance level

1. b) He also claims that the average living area is more than 1800 Sq. Ft. Do you agree with this? (Use a 5% significance level for both.). Briefly explain what the p-values in these cases mean?

Let, μ be average living area of all houses

null hypothesis, $H_0 : \mu \leq 1800$

alternate hypothesis, $H_a : \mu > 1800$

Population mean is known, Sample standard deviation, s is known and sample size is known. Since sample size is large enough, greater than 30 we use Test Statistic, t here.

p-value is probability of getting a stronger evidence to reject null hypothesis. Here, it can be given by probability of seeing sample mean of more than 1807, if population mean, μ is 1800

```r
n <- length(house_prices$Living.Area)
s = sd(house_prices$Living.Area)
x_bar = mean(house_prices$Living.Area)
mu = 1800

print(paste("Sample Mean :", x_bar))
```

```
## [1] "Sample Mean : 1807.30276981853"
```

```r
t <- (x_bar -  mu)/(s/sqrt(n))
p_value <- 1 - pt(t, n-1)
print(paste("p value :", p_value))
```

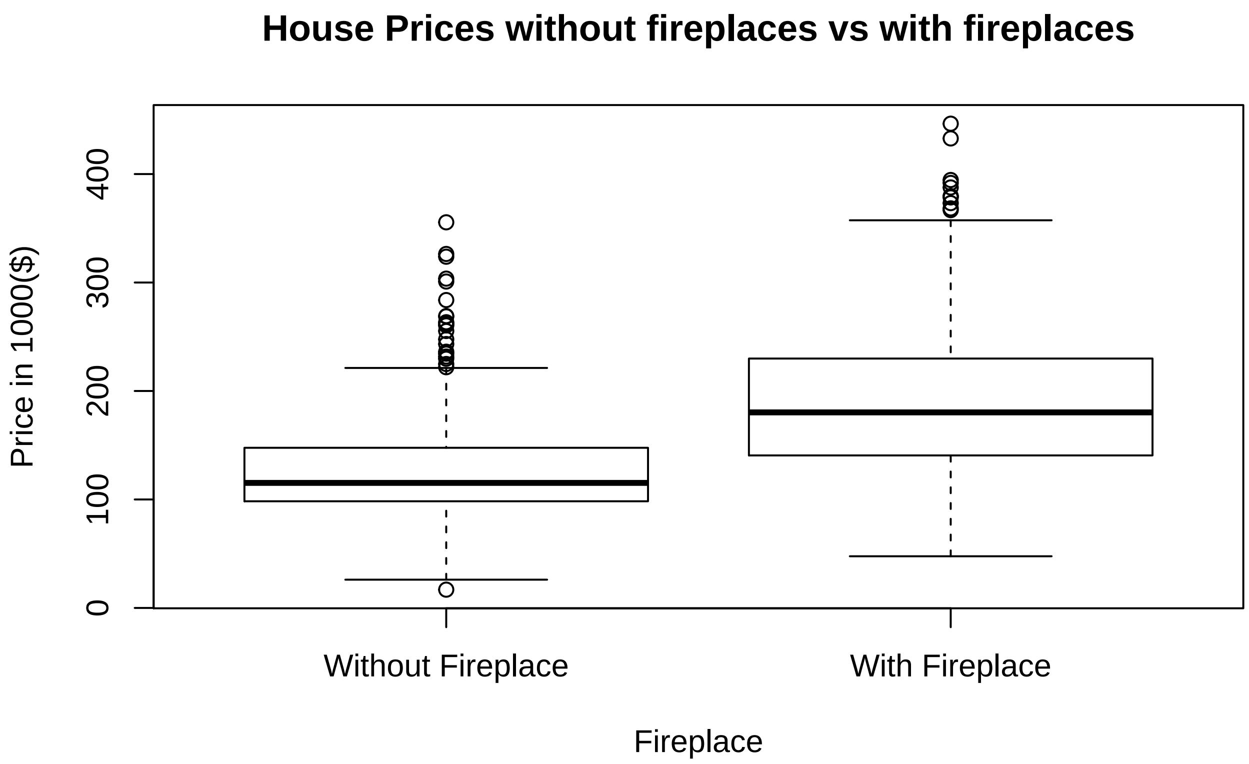
```
## [1] "p value : 0.356333895962788"
```

Since, p-value > 0.05 (alpha) we do not reject the null, We do not agree with the claim that average living area is more than 1800 Sq. Ft. at 5% significance level

2. Are the home prices higher for houses with fireplaces as compared to those without?

a) Create side-by-side box plots of the house prices of the two groups and comment them.

Looking at boxplots below, we can say that houses with fireplaces have higher prices than without fireplace



b) Formulate an appropriate hypothesis and test it in order to check the above claim. Assume that the population standard deviations of house prices in the two groups are equal.

Let, μ_f be average prices of house with fireplace and μ_{wf} be average prices of house without fireplace

null hypothesis, H_0 : μ_f - μ_{wf} ≤ 0

alternate hypothesis, H_a : μ_f - μ_{wf} > 0

p-value is probability of getting a stronger evidence to reject null hypothesis. Here, it can be given by probability of seeing another sample with mean of house prices with fire - mean of house prices without fire being less than 0

Population standard deviation is unknown, but given that $\sigma_f = \sigma_{wf}$, so we can use pooled sample standard deviation

```
houses_with_fireplace <- subset(house_prices, house_prices$Fireplace == 1)
houses_without_fireplace <- subset(house_prices, house_prices$Fireplace == 0)

#sample sizes for houses with and without fireplace
n1 <- length(houses_with_fireplace$Price)
n2 <- length(houses_without_fireplace$Price)

#standard deviation of samples houses with and without fireplace
s1 <- sd(houses_with_fireplace$Price)
s2 <- sd(houses_without_fireplace$Price)

#sample mean of houses with and without fireplace
x_bar_f <- mean(houses_with_fireplace$Price)
x_bar_wf <- mean(houses_without_fireplace$Price)
```

```r
#standard deviation of pooled sample
s_pooled <- sqrt(((n1-1)*s1^2 + (n2-1)*s2^2) / (n1+n2-2))
df = n1+n2-2
D_0 <- 0

#Going by formula, with sigma1 = sigma2
t <- ((x_bar_f - x_bar_wf) - D_0) / (s_pooled*(sqrt(1/n1+1/n2)))
p_value <- 1 - pt(t, df)
print(p_value)
```

```
## [1] 0
```

Since, p-value is 0, so there is 0% chance to get a sample evidence with mean house prices with fireplaces - mean house prices without fireplaces < 0. So we reject null, and can say that house prices with fireplaces having higher prices than without fireplaces is TRUE

3. Any house aged more than 30 years is considered an "old" house. Your friend claims that old houses have larger lot sizes than new houses. Do you agree? Explain. Use a significance level of 5% for your test. Historical data suggests that old houses include some very large and some very small lot sizes but new houses are more homogeneous in their lot sizes.

Let, μ_{old} be lot size of old house and μ_{new} be lot size of new house

null hypothesis, H_0 : μ_{old} - μ_{new} ≤ 0

alternate hypothesis, H_a : μ_{old} - μ_{new} > 0

p-value is probability of getting a stronger evidence to reject null hypothesis. Here, it can be given by probability of seeing another sample with mean of lot sizes in old houses - mean of lot sizes in new houses less than 0

Population standard deviation is unknown, also given that $\sigma_{old} \neq \sigma_{new}$

```r
old_houses <- subset(house_prices, house_prices$Age > 30)
new_houses <- subset(house_prices, house_prices$Age <= 30)

#sample sizes for old and new houses
n1 <- length(old_houses$Lot.Size)
n2 <- length(new_houses$Lot.Size)

#standard deviation of lot size of old and new houses
s1 <- sd(old_houses$Lot.Size)
s2 <- sd(new_houses$Lot.Size)

#sample mean of old and new houses
x_bar_old <- mean(old_houses$Lot.Size)
x_bar_new <- mean(new_houses$Lot.Size)

#degree of freedom
df <- ((s1^2/n1 + s2^2/n2)^2) / (((s1^2/n1)^2 / (n1-1)) + ((s2^2/n2)^2 / (n2-1)))
D_0 <- 0

#Going by formula, with sigma1 = sigma2
t <- ((x_bar_old - x_bar_new) - D_0) / (sqrt((s1^2)/n1+(s2^2)/n2))
```

```r
p_value <- 1 - pt(t, df)
print(paste0("p-value :",p_value))
```

```
## [1] "p-value :0.722382522539582"
```

Since, p-value is 0.722 $>$ 0.05, we do not reject the null hypothesis, and can say that chance to getting a sample evidence with lot size of older houses being - lot size of newer houses $<$ 0 is high !

4. Based on the evidence available here, would you be willing to claim that fireplaces have become more fashionable? For simplicity, it is OK to compare only "new" houses and "old" houses. Use a significance level of 5% for your test. Use a significance level of 5% for your test.

Let, p_{nf} be proportion of new houses with fireplace and p_{of} be proportion of old houses with fireplace

null hypothesis, H_0 : p_{nf} - p_{of} ≤ 0

alternate hypothesis, H_a : p_{nf} - p_{of} > 0

p-value is probability of getting a stronger evidence to reject null hypothesis. Here, it can be given by probability of seeing another sample with proportion of new houses with fireplace - proportion of old houses withour fireplace less than 0

Population standard deviation is unknown, also given that $\sigma_{old} \neq \sigma_{new}$

```r
#set of old and new houses
old_houses <- subset(house_prices, house_prices$Age > 30)
new_houses <- subset(house_prices, house_prices$Age <= 30)

#set of old and new houses with fireplace
new_houses_with_fireplace <- subset(house_prices, house_prices$Age <= 30 & house_prices$Fireplace == 1)
old_houses_with_fireplace <- subset(house_prices, house_prices$Age > 30 & house_prices$Fireplace == 1)

n_nf <- nrow(new_houses_with_fireplace)
n_of <- nrow(old_houses_with_fireplace)

p_nf <- nrow(new_houses_with_fireplace)/nrow(new_houses)
p_of <- nrow(old_houses_with_fireplace)/nrow(old_houses)

p_bar <- (p_nf*n_nf + p_of*n_of) / (n_nf + n_of)

z <- (p_nf - p_of) / sqrt(p_bar*(1-p_bar)*(1/n_nf + 1/n_of))
p_value <- 1 - pnorm(z)
print(paste0("p value :",p_value))
```

```
## [1] "p value :7.71237021235383e-06"
```

Since, p-value is close to 0 $<$ 0.01, we reject null hypothesis and we can claim that fireplaces have become more fashionable and proportion of new houses with fireplace will be always high

5. Suppose that houses with 1-2 bedrooms are considered to be "Small Houses", those with 3-4 are "Medium Houses" and 5-6 as "Big Houses". Can we conclude that the prices of Small, Medium and Big houses are not the same, at 1% level of significance?

Let, μ_i be price of house in each group

null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3$ alternate hypothesis, H_a : Not all μ_i are equal

Using statistical test for equality of means

```
small_houses <- subset(house_prices, house_prices$Bedrooms < 3)
medium_houses <- subset(house_prices, house_prices$Bedrooms == 3 | house_prices$Bedrooms == 4)
big_houses <- subset(house_prices, house_prices$Bedrooms > 4)
small_houses$Group <- "Small Houses"
medium_houses$Group <- "Medium Houses"
big_houses$Group <- "Big Houses"
print(length(small_houses$Price))
```

```
## [1] 179
```

```
print(length(medium_houses$Price))
```

```
## [1] 843
```

```
print(length(big_houses$Price))
```

```
## [1] 25
```

```
houses_by_bedrooms <- rbind(rbind(small_houses,medium_houses),big_houses)

anova <- aov(houses_by_bedrooms$Price~houses_by_bedrooms$Group)
summary(anova)
```

```
##                            Df    Sum Sq   Mean Sq F value Pr(>F)    
## houses_by_bedrooms$Group    2 4.840e+11 2.420e+11   58.71 <2e-16 ***
## Residuals                1044 4.303e+12 4.122e+09                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since, 2e-16 is close to 0, we reject null hypothesis, even at significance level of 1%.

we can say that it is highly unlikely that all three groups of small/medium/big houses will have same mean price