STOR 455 - Class 2 - R Notebook

# loads packages needed  
# install a package before first using it for the first time  
  
library(readr)  
library(mosaic)  
  
# loads the DistanceHome dataframe into the environment from github  
  
DistanceHome <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/DistanceHome.csv")  
  
# Alternative way to load dataframe (remove # to use)  
# DistanceHome.csv must be saved in the same folder as this notebook!  
  
#DistanceHome <- read\_csv("DistanceHome.csv")  
  
# Shows the variables and first 6 cases (by default)  
head(DistanceHome)

## # A tibble: 6 x 3  
## Distance Hours Introvert   
## <dbl> <dbl> <chr>   
## 1 7606 21 Introversion  
## 2 7606 21 Introversion  
## 3 3800 20 Introversion  
## 4 7102 20 Introversion  
## 5 6000 20 Introversion  
## 6 7756 18 Introversion

**Example: Distance to Home** - *Question:* How can we predict the distance from campus to home for Carolina students? - *Data:* Estimated distance to home (in miles) for students taking STOR 455 in a previous semester. - *Predictor variables:* Start with none.

**Example: Constant Model** Y = c + Error

Where c = an unknown constant

**Terminology** - “The constant c” = **parameter** of the model - “Sample estimate” - use data to provide a sample estimate of c

*How should we estimate 𝑐 from data?*

*Below:* Summarize the Distance variable - Numerical: mean and median

# dataframe$variable\_name  
  
mean(DistanceHome$Distance)

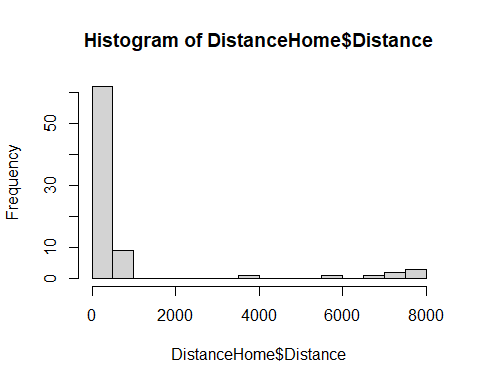
## [1] 844.6234

median(DistanceHome$Distance)

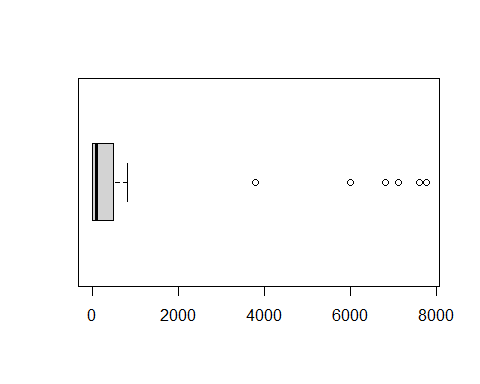
## [1] 113

*Below:* Summarize the Distance variable - Graphical: histogram, boxplot

hist(DistanceHome$Distance, breaks=20)



boxplot(DistanceHome$Distance, horizontal = TRUE)



**Predicted Value for Response** - Get an estimate for Y using the predictors and the model with estimated parameters.

*Notation:* The predicted y is denoted yhat

For the constant Model: yhat = chat

Examples: yhat = chat = ybar (\*sample mean) yhat = chat = m (sample median)

**Can we use a predictor to improve the model?**  X = Hours to travel home? X = Introvert? - Two sample t test for a difference in means

**Model with a Binary Predictor** Y = f(x) + Error where X = introversion, mu1 = mean distance for Extroverts mu2 = mean distance for Introverts

mean(Distance~Introvert, data=DistanceHome)

## Extroversion Introversion   
## 365.6026 1288.5939

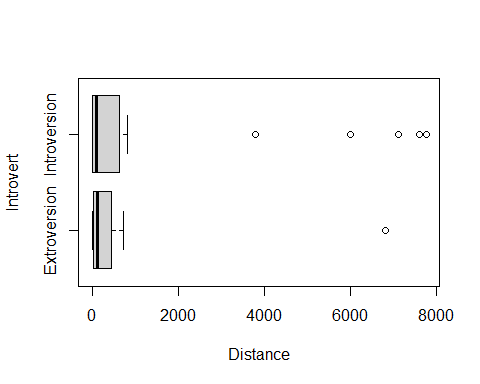
tally(~Introvert, data=DistanceHome)

## Introvert  
## Extroversion Introversion   
## 38 41

sd(Distance~Introvert, data=DistanceHome)

## Extroversion Introversion   
## 1094.443 2559.412

boxplot(Distance~Introvert, data=DistanceHome, horizontal=TRUE)



**Two-sample T-Test Difference in Means** *Hypothesis* Ho: Mu1 = Mu2 Ha: Mu1 != Mu2

*Compare to a t-dist*

**P-value** - The p-value is the proportion of samples, when the H0 is true, that would be as (or more) extreme as the observed sample.

*Below, Conclusion:* Decision: Reject H0 only when the p-value is small.

t.test(Distance~Introvert, data=DistanceHome)

##   
## Welch Two Sample t-test  
##   
## data: Distance by Introvert  
## t = -2.1103, df = 55.025, p-value = 0.03939  
## alternative hypothesis: true difference in means between group Extroversion and group Introversion is not equal to 0  
## 95 percent confidence interval:  
## -1799.48957 -46.49298  
## sample estimates:  
## mean in group Extroversion mean in group Introversion   
## 365.6026 1288.5939

**Normality?** - The two-sample t-test assumes both samples are from normal populations

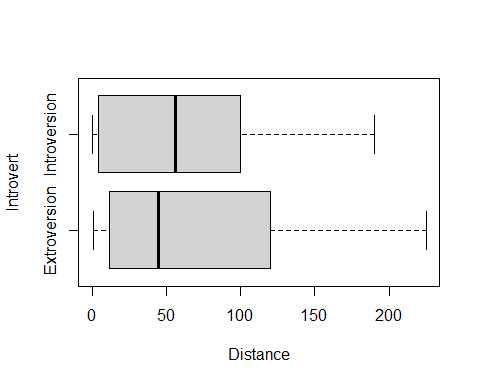
**Domestic Distance** - Suppose that we look only at distances that are really “drivable”? *See below*

Domestic = subset(DistanceHome, Distance<250)  
head(Domestic)

## # A tibble: 6 x 3  
## Distance Hours Introvert   
## <dbl> <dbl> <chr>   
## 1 225 3.5 Extroversion  
## 2 120 3 Extroversion  
## 3 190 3 Introversion  
## 4 172 3 Introversion  
## 5 167. 3 Extroversion  
## 6 190 3 Introversion

* For distance home (only including students less than 250 miles from home) the Introvert variable does not improve the model significantly.

boxplot(Distance~Introvert, data=Domestic, horizontal=TRUE)



t.test(Distance~Introvert, data=Domestic)

##   
## Welch Two Sample t-test  
##   
## data: Distance by Introvert  
## t = 0.09629, df = 51.968, p-value = 0.9237  
## alternative hypothesis: true difference in means between group Extroversion and group Introversion is not equal to 0  
## 95 percent confidence interval:  
## -33.17687 36.52132  
## sample estimates:  
## mean in group Extroversion mean in group Introversion   
## 67.16667 65.49444

**Inference Review: Hypothesis Testing** - Suppose that we look only at distances that are really “drivable”? *Test* Ho: mu1 = Mu2 Ha: mu1 != Mu2

There is a 93.3% chance that we would receive a samples with a difference as extreme as we did if the null hypothesis is true. p-value = 0.933

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. There is not evidence to suggest that there is a difference in the number of miles from home Carolina students are (of those students 250 miles or less) based on if they are introverts or extroverts.

There is a 93.3% chance that we would receive a samples with a difference as extreme as we did if the null hypothesis is true. p-value = 0.933

Domestic=subset(DistanceHome,Distance<250)

**Domestic Distance** For distance home (only including students less than 250 miles from home) the Introvert variable does not improve the model significantly.

t.test(Distance~Introvert, data=Domestic)

##   
## Welch Two Sample t-test  
##   
## data: Distance by Introvert  
## t = 0.09629, df = 51.968, p-value = 0.9237  
## alternative hypothesis: true difference in means between group Extroversion and group Introversion is not equal to 0  
## 95 percent confidence interval:  
## -33.17687 36.52132  
## sample estimates:  
## mean in group Extroversion mean in group Introversion   
## 67.16667 65.49444