HW 3

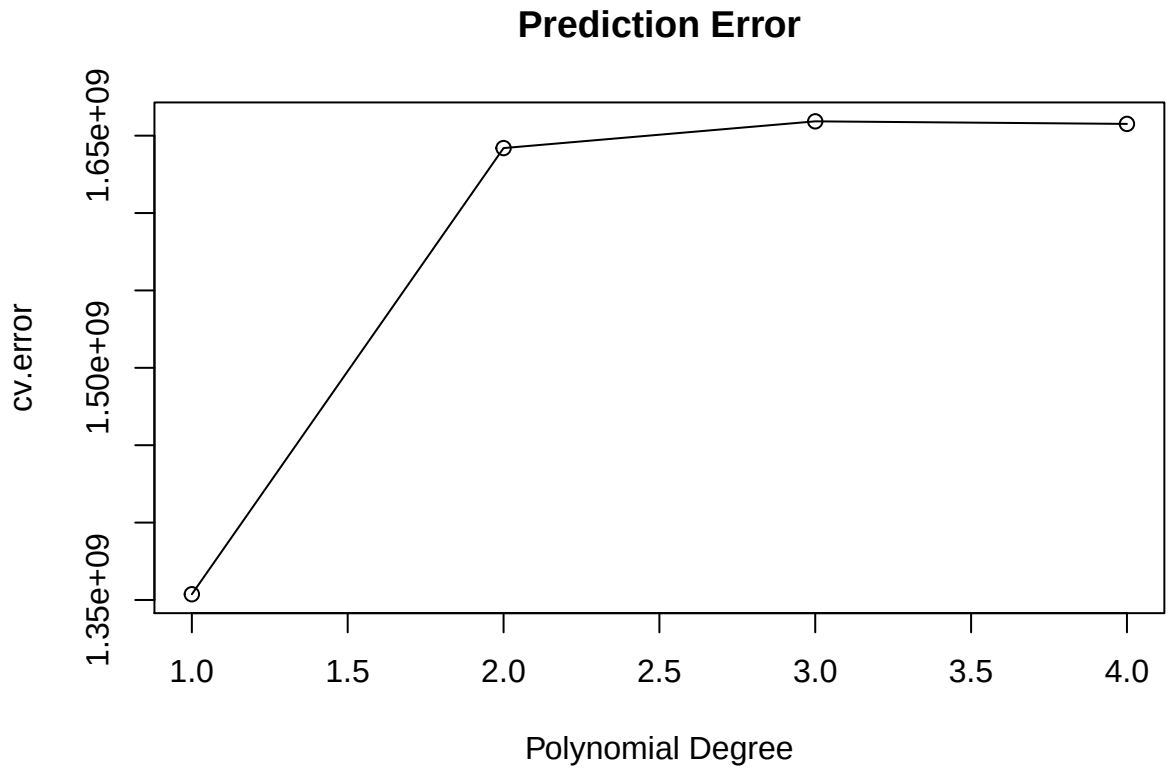
Ben Buzzee

9/20/2019

1. Cross Validation

We will use cross validation to find out which degree polynomial has the lowest prediction error when predicting incomes.

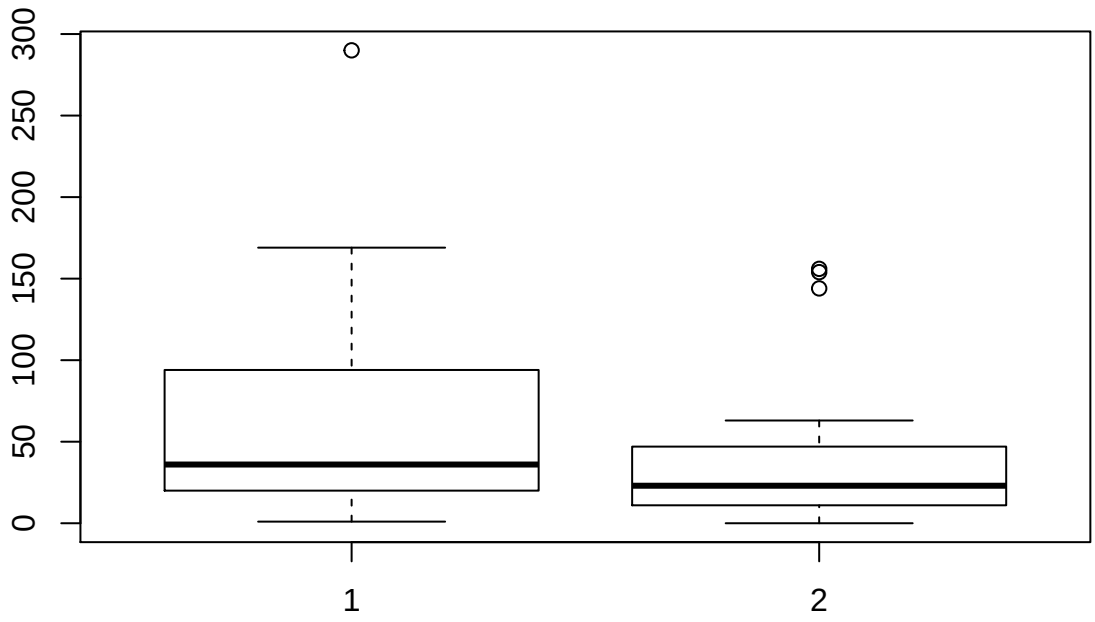
```
income <- read.csv("income.txt", sep = "", header = T)

# linear
glm.fit1 = glm(y~x, data=income)
cv.error1=cv.glm(income, glm.fit1, K=5)

# quadratic model
glm.fit2 = glm(y~poly(x,2), data=income)
cv.error2=cv.glm(income, glm.fit2, K=5)

# cubic model
glm.fit3 = glm(y~poly(x,3), data=income)
cv.error3=cv.glm(income, glm.fit3, K=5)

# quartic model
glm.fit4 = glm(y~poly(x,4), data=income)
cv.error4=cv.glm(income, glm.fit4, K=5)


cv.error <- c(cv.error1$delta[1], cv.error2$delta[1], cv.error3$delta[1], cv.error4$delta[1])

d=c(1,2,3,4)
plot(d, cv.error, main = "Prediction Error", xlab = "Polynomial Degree")
lines(d, cv.error)
```



There seems to be a lot of variation in our prediction error. We could repeat the CV process several times and compute an average error, but given that a degree of one is the simplest most interpretable model, I'd stick with a degree of one even if it means a slight increase in prediction error.

2. Mann-Whitney Test

We want to test the hypothesis that children exposed to violent TV were slower to react to real-life violence than children who were not exposed to violent TV. The response times were measured in seconds for both groups. First let's examine the data visually:



(a)

If we let F be the CDF of violent tv watcher's reaction times, and G be the cdf of reaction times of children not exposed the violent TV, our hypothesis statements would be $H_0 : F(t) = G(t)$ vs $H_A : F(t) = G(t - \Delta)$ for some $\Delta < 0$

```
wilcox.test(x,y, alternative = "greater")
```

```
##
##  Wilcoxon rank sum test
##
## data:  x and y
## W = 274, p-value = 0.09222
## alternative hypothesis: true location shift is greater than 0
```

With a U test statistic of 274 and a p-value of .09, would fail to reject the null hypothesis. However, I think that would be a misleading conclusion based on an arbitrarily chosen "significance" level and that a p-value of .09 suggest there is evidence that reaction times were longer for children exposed to violent TV.

To convert our U statistic to the W statistic, we can subtract $\frac{n(n+1)}{2}$ which in this case is 231, so our W statistic would be 43.

To conduct a hypothesis test based on a large sample approximation, we can set the exact argument to false in the wilcox.test function.

```
wilcox.test(x,y, alternative = "greater", exact = FALSE)
```

```
##
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 274, p-value = 0.09122
## alternative hypothesis: true location shift is greater than 0
```

And we come to the same conclusion as before with our p-value roughly the same.

3. Hodges and Lehman Estimator

(a) Estimate $\hat{\Delta}$:

```
# data
X <-  c(84.5, 81.0, 82.6, 80.5, 82.1, 83.4, 79.7)
Y <- c(82.4, 83.9, 86.3, 86.6, 87.8, 84.1)

# initialize empty matrix
diff_mat <- matrix(nrow = length(X), ncol = length(Y))

# for every combo of x and y compute the difference
for (i in 1:length(X)){
  for (j in 1:length(Y)){
  diff_mat[i,j] <- X[i] - Y[j]
  }
}

median(diff_mat)
```

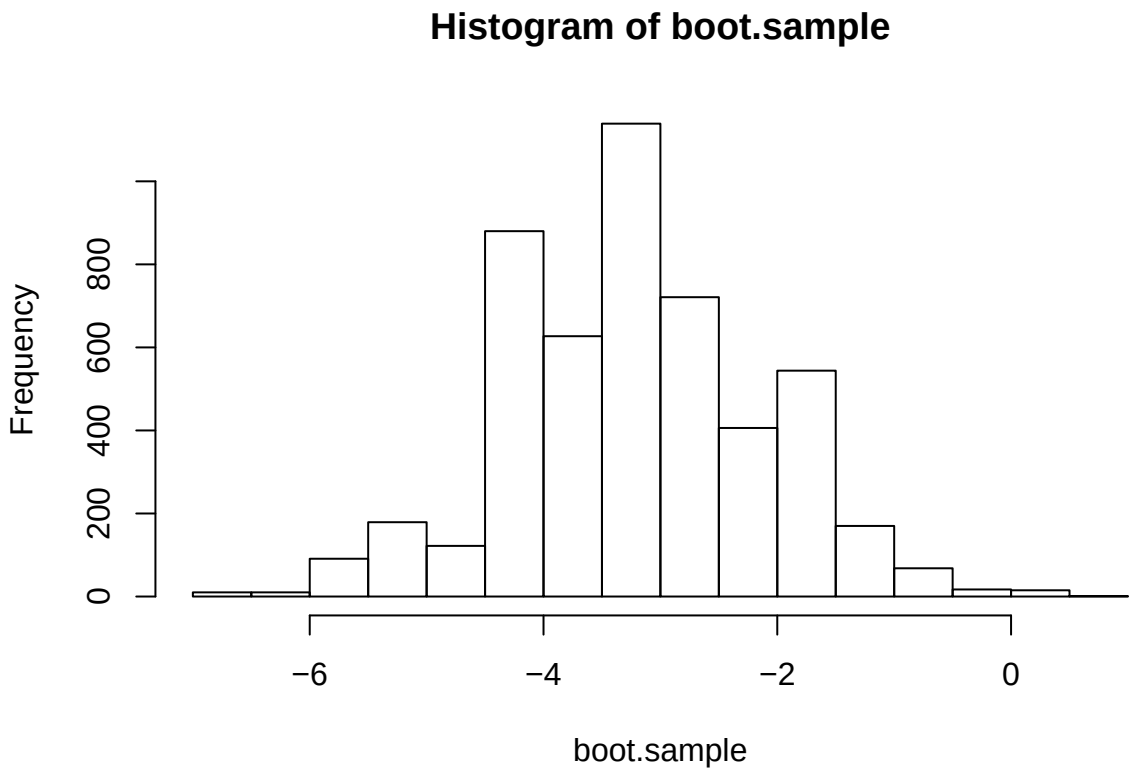
```
## [1] -3.25
```

We arrive at an estimate for $\hat{\Delta}$ of -3.25.

(b) Bootstrap Distribution of $\hat{\Delta}$

```
nreps <- 5000
boot.sample <- rep(0, nreps)
diff_mat <- matrix(nrow = length(X), ncol = length(Y))

for (z in 1:nreps){
  # resample
  newX <- sample(X, replace = T)
  newY <- sample(Y, replace = T)

  # find delta for new sample
  for (i in 1:length(newX)){
    for (j in 1:length(newY)){
      diff_mat[i,j] <- newX[i] - newY[j]
      delta <- median(diff_mat)
    }
  }
  # add new delta to sample
  boot.sample[z] <- delta
```

```
}

hist(boot.sample)
```




```
## [1] "Mean of Bootstrap Distribution:"
## [1] -3.19985
## [1] "SD of Bootstrap Distribution:"
## [1] 1.093925
## [1] "95% Confidence Interval:"
##     2.5%    97.5%
## -5.45000 -1.29875
```

Since our confidence interval is entirely negative, we have evidence that the wing lengths for migratory (X) juncos are smaller than the wing lengths of non-migratory juncos (Y).

4. Computing W by hand

(a)

```
x=sort(c(2.1, 1.9, 2.6, 3.3))
y=sort(c(1.9, 2.6, 3.7, NA))
rank_y <- c(1.5,4.5,7)
```

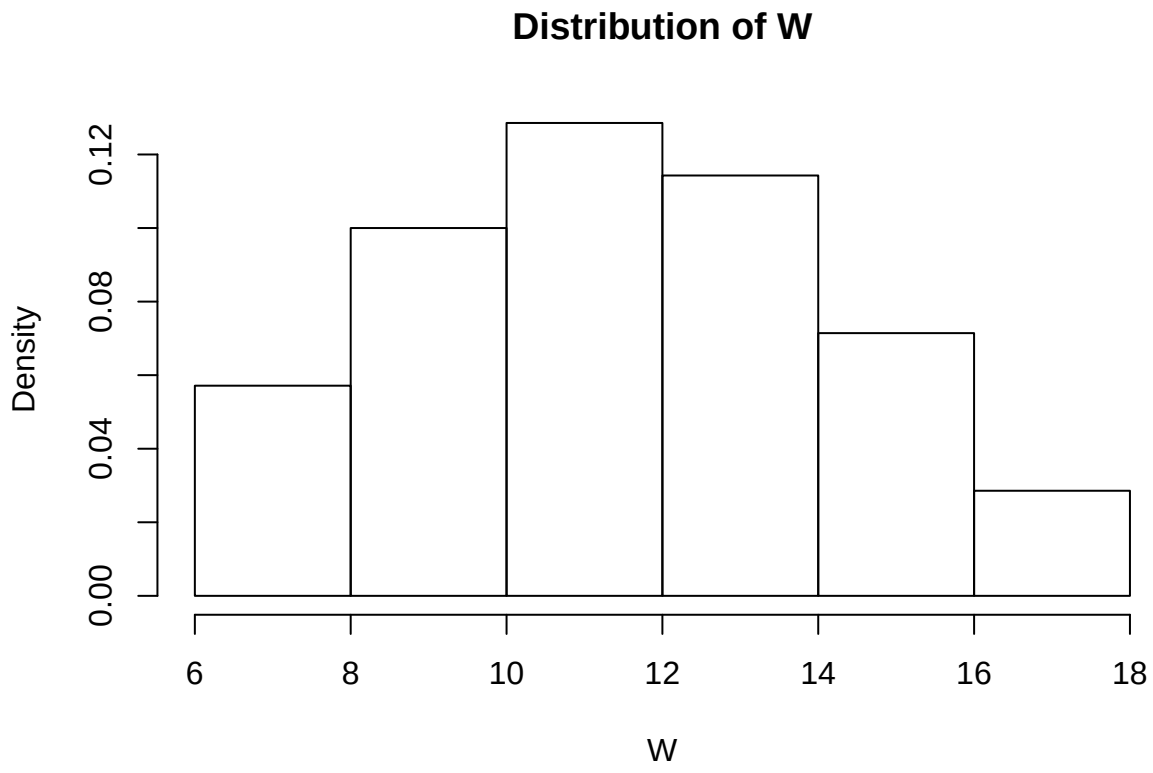
```r
df  <- data.frame(x = x, y = c(y, NA), rank_y = c(rank_y, NA))

df
```

```
##     x   y rank_y
## 1 1.9 1.9    1.5
## 2 2.1 2.6    4.5
## 3 2.6 3.7    7.0
## 4 3.3  NA     NA
```

We can find the W statistic by summing the ranks of Y. This gives us a W statistic of 13.

(b)

```r
# number of ways we can get 3 ranks given seven options, all equally likely under null
ranksums <- colSums(combn(1:7, 3))


hist(ranksums, probability = T, main = "Distribution of W", xlab = "W")
```



(c)

To find how extreme W is, we can look at a tabled version of the above distribution:

```
table(ranksums)
```

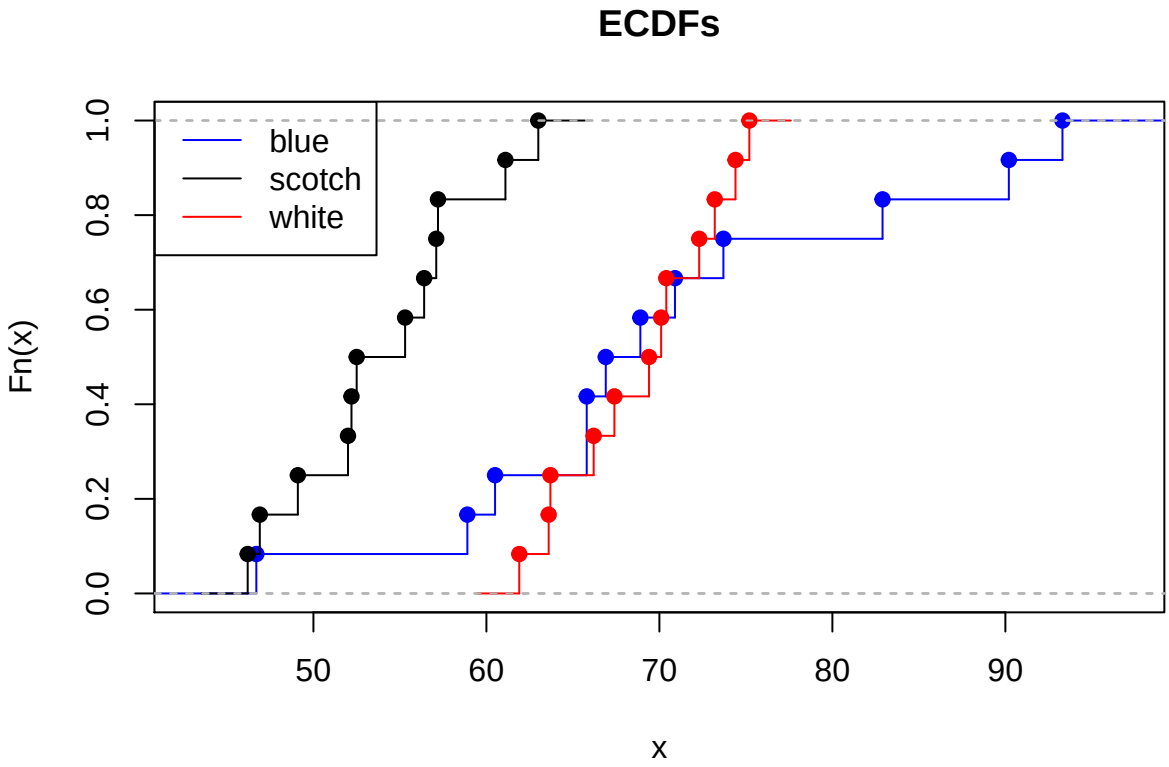
```
## ranksums
##  6  7  8  9 10 11 12 13 14 15 16 17 18
##  1  1  2  3  4  4  5  4  4  3  2  1  1
```

Our p-value would be the probability we get a test statistic as large or larger than 13 under the null hypothesis. This would be 15/35 or .43. So we would fail to reject the null hypothesis and have no evidence of a difference in the location of the distributions for X and Y.

5. Kruskal-Wallace

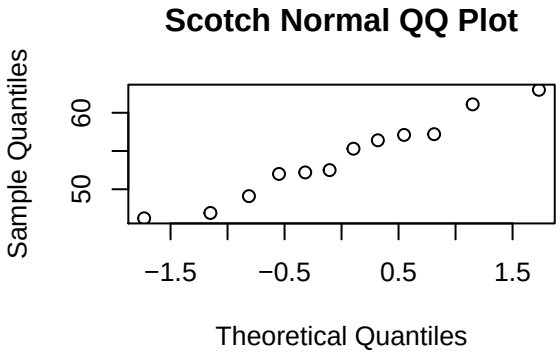
(a)

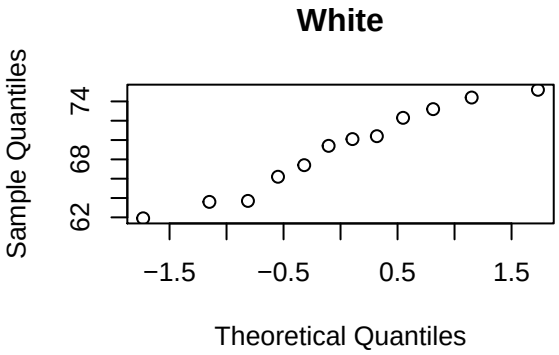
The assumptions of independence and iid observations from a continuous cdf must be assumed from the context of the problem. We can plot the empirical distributions of each group to check the third assumption that our three variables come from the same CDF with a possible shift parameter.

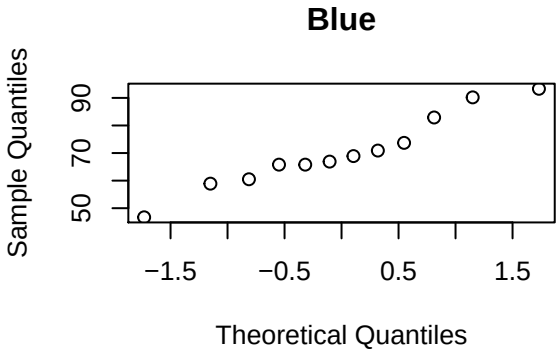


If the CDFs of our variables only differed by a location parameter we would expect the "steepness" of our empirical distributions to be similar. From the above plot it appears that white pines differ substantially from blue and scotch pines by more than just a location parameter.

To check the assumptions of an ANOVA modal we can check that each variable is normally distributed with QQ







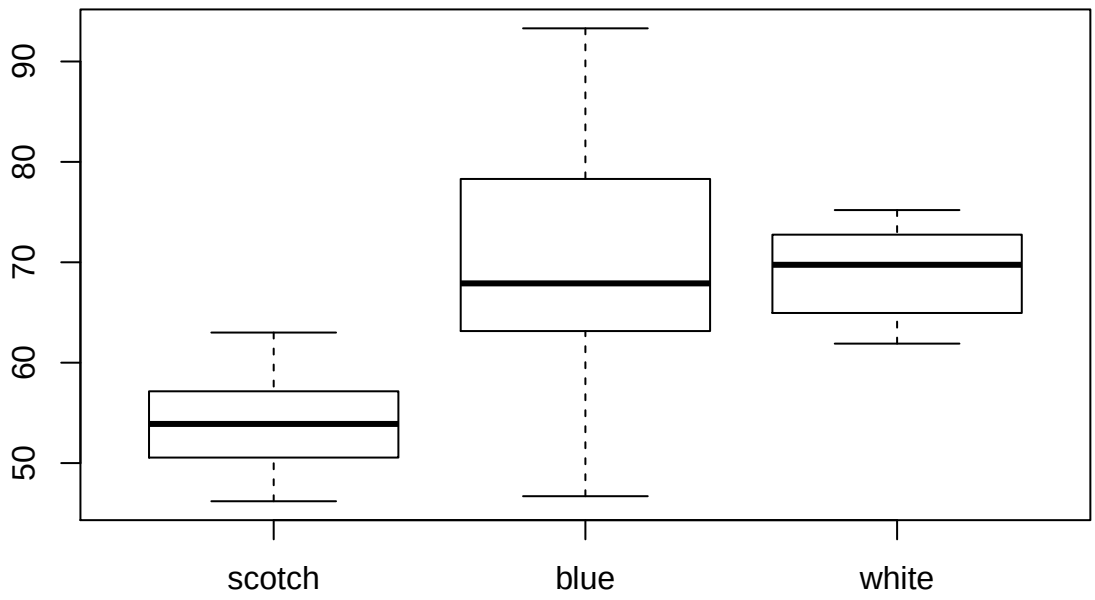
plots:

Since we only have twelve observations for each variable it would be difficult to test and assume normality. All the variables seem to deviate at the tails from what we would expect from a normal distribution, so an ANOVA may be inappropriate.

(b)

First we'll look at a boxplot, to see if we might expect a difference in location parameters:

```
boxplot(pines)
```



To formalize our results, we will conduct a KW test:

```
scotch <- pines$scotch
white <- pines$white
blue <- pines$blue

pKW(x=list(scotch, white, blue))
```

```
## Ties are present, so p-values are based on conditional null distribution.
## Group sizes: 12 12 12
## Kruskal-Wallis H Statistic: 18.6006
## Monte Carlo (Using 10000 Iterations) upper-tail probability: 1e-04
```

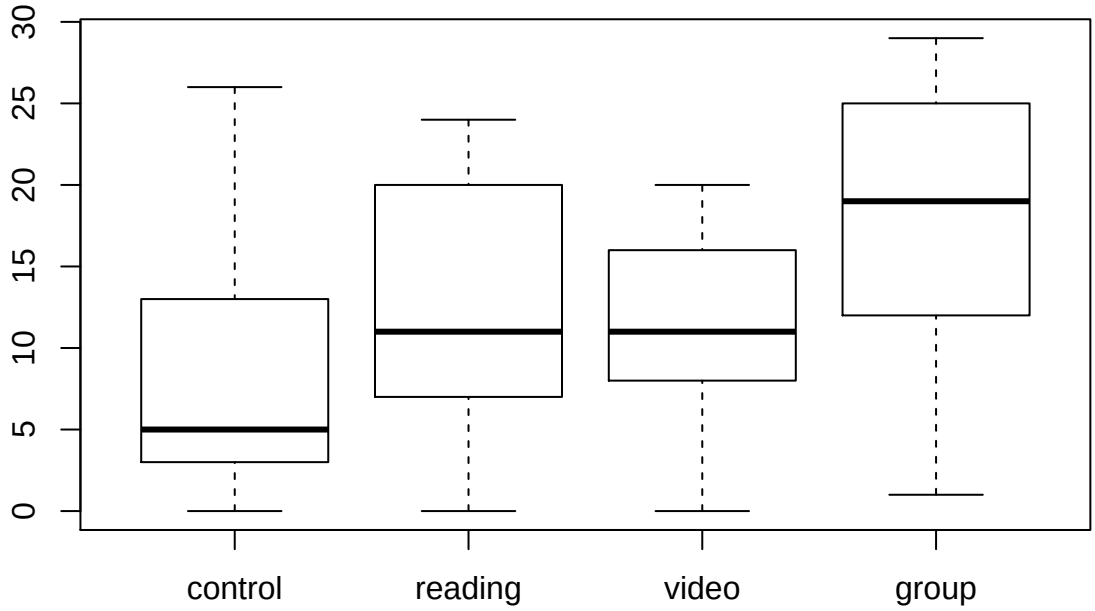
Hypothesis statements: $H_0 : \tau_1 = \tau_2 = \tau_3$ vs $H_A : \tau_i \neq \tau_j$ for some $i \neq j$. With a test statistic of 18.6 and a p-value of approximately 0, we can conclude at least two of the location parameters are different.

6. KW Large Sample Approximation

Again we'll start by examining the data visually and then perform a KW test.

```
therapy <- read.csv("therapy.txt", sep = " ")

control <- therapy$control
reading <- therapy$reading
video <- therapy$video
group <- therapy$group
```

```r
boxplot(therapy)
```




```r
kruskal.test(x = list(control, reading, video, group))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(control, reading, video, group)
## Kruskal-Wallis chi-squared = 4.2646, df = 3, p-value = 0.2343
```

From the large sample approxiation, we arrive at a test statistic of H = 4.26. Next we will try to correct for ties and compute H' by hand:

```r
# create a vector of all values
vec <- c(control, reading, video, group)

# retrieve the values that have ties and how many other values they are tied with
# top row is the value, bottom row is how many ties it has
table(vec)[table(vec) != 1]
```

```
## vec
##  0  1  3  5  9 11 13 17 20
##  3  2  2  3  2  2  4  2  3
```

So in the formula for H', the total number of tied values, g, is the sum of the bottom row above. The number of groups in each tie, t_j, would be the individual values on the bottom row.

```r
# total number of ties
g <- sum(table(vec)[table(vec) != 1])

# total observations
N <- 36

# first we nee to expand the above table so we can loop over it
# vec1 will be our values of t_j, or the number of groups that share the tied value
# for example, the first three values will be 3, 3, 3 indicating that three groups
# share the first three ties--this is probably a misinterpretation of the notes :)

vec1 <- NULL
for (i in 1:9){
  vec1 <- c(vec1, rep(table(vec)[table(vec) != 1][[i]], times = table(vec)[table(vec) != 1][[i]]))
}


# now we can compute the sum in the denominator of the formula for H'
correction <- sum(vec1^3 - vec1)/(N^3 - N)


Hprime <- 4.26/(1-correction)


Hprime
```

```
## [1] 4.307678
```

So we arrive at a corrected test statistic of 4.30, which is not much different than the original.

To finish the hypothesis test, we can compare our test statistic to the null distribution.

```r
pchisq(4.30, df = 3, lower.tail = F)
```

```
## [1] 0.2308387
```

So with a test statistic of 4.30 and a p-value of .23, we would fail to reject the null and have no evidence different pretherapy treatments have different effects (or median effects).