Lab 5

Frank Woodling

March 22, 2016

1.

```r
movies <- read.table('moviesall.txt', header = T)
attach(movies)

g.ss <- subset(movies, rating == 'G')
g <- mean(g.ss$runtime)

pg.ss <- subset(movies, rating == 'PG')
pg <- mean(pg.ss$runtime)

pg13.ss <- subset(movies, rating == 'PG-13')
pg13 <- mean(pg13.ss$runtime)

r.ss <- subset(movies, rating == 'R')
r <- mean(r.ss$runtime)

n = c(dim(g.ss)[1], dim(pg.ss)[1], dim(pg13.ss)[1], dim(r.ss)[1])

N = sum(n)
K = length(n)

xbar = c(g, pg, pg13, r)

s = c(sd(g.ss$runtime), sd(pg.ss$runtime), sd(pg13.ss$runtime), sd(r.ss$runtime))

xbar.overall = sum(n*xbar)/N

SSTr = sum(n*(xbar - xbar.overall)^2)
SSE = sum((n-1)*s^2)
SSTotal = SSTr + SSE

MSTr = SSTr/(K-1)
MSE = SSE/(N-K)
F = MSTr/MSE

# SSTr
# SSE
# SSTotal
#
# MSTr
# MSE
# F

r1 <- c(SSTr, K-1, MSTr, F)
```

```
r2 <- c(SSE, N-K, MSE, '')
r3 <- c(SSTotal, N-1, '','')
tbl <- rbind(r1, r2, r3)
rownames(tbl) <- c('Treatment', 'Error', 'Total')
colnames(tbl) <- c('SS', 'df', 'MS', 'F')
tbl
```

```
##           SS                 df    MS                 F
## Treatment "6701.28003663004" "3"   "2233.76001221001" "5.25961286252206"
## Error     "57759.2628205128" "136" "424.700461915535" ""
## Total     "64460.5428571429" "139" ""                 ""
```

2.

```
1 - pf(F, K-1, N-K)
```

```
## [1] 0.001831341
```

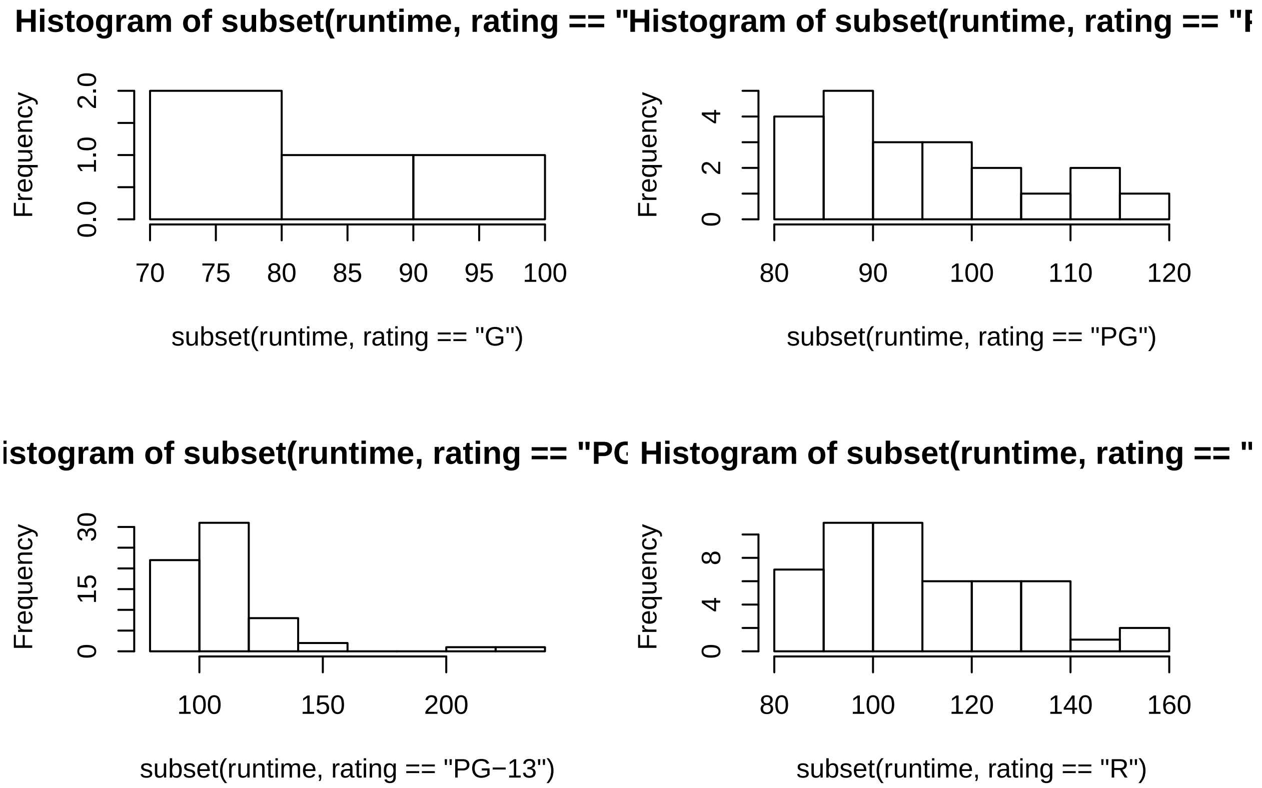
3.

```
runtime.anova <- lm(runtime~rating)
anova(runtime.anova)
```

```
## Analysis of Variance Table
## 
## Response: runtime
##            Df Sum Sq Mean Sq F value   Pr(>F)   
## rating      3   6701  2233.8  5.2596 0.001831 **
## Residuals 136  57759   424.7                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.

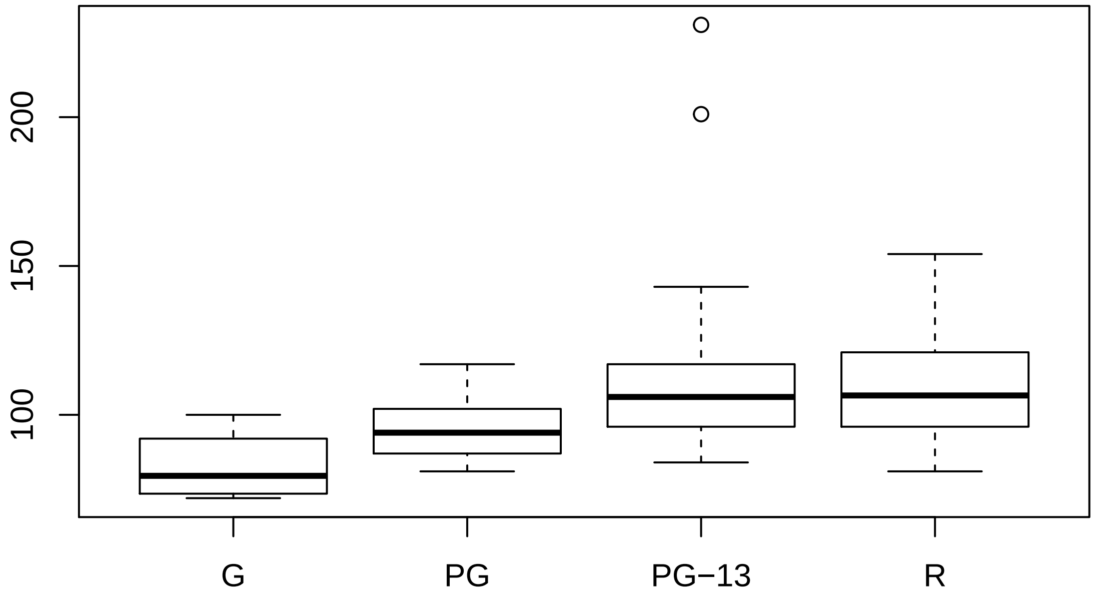
```
plot.new()
par(mfrow=c(2,2))
hist(subset(runtime, rating == 'G'))
hist(subset(runtime, rating == 'PG'))
hist(subset(runtime, rating == 'PG-13'))
hist(subset(runtime, rating == 'R'))
```



Each distribution is right-skewed. None of the histograms are close to normal. The mean and median are not the same so we cannot assume this is a normal distribution.

5.

```
plot.new()
par(mfrow=c(1,1))
boxplot(runtime~rating)
```



The variances do not seem equal when comparing with the boxplot. If they were equal the interquartile ranges would be near equal, but G and PG-13 are much smaller than the other two.

6.

```r
runtime.anova = lm(runtime ~ rating)
teststat.obs = summary(runtime.anova)$fstatistic[1]

teststat = rep(NA, 1000)
for(i in 1:1000){
  ratingSHUFFLE = sample(rating)
  SHUFFLE.anova = lm(runtime ~ ratingSHUFFLE)
  teststat[i] = summary(SHUFFLE.anova)$fstatistic[1]
}

sum(teststat >= teststat.obs)/1000
```

```
## [1] 0.011
```

```r
# Find number of combinations
(factorial(dim(movies)[1]))/
  factorial(dim(g.ss)[1])*factorial(dim(pg.ss)[1])*factorial(dim(pg13.ss)[1])*factorial(dim(r.ss)[1])
```

```
## [1] Inf
```

In order to do an exact permutation test we would need to consider how many observations are in each category and how many we choose. The resulting answer is too large for R to calcuate.

$\frac{N!}{n_1!n_2!...n_k!}$

$\frac{140!}{4!65!50!140!}$

$\frac{1.346201e+241!}{24*5.109094e+19*8.247651e+90*3.041409e+64}$

7.

```
teststat.obs
```

```
##    value
## 5.259613
```

```
F
```

```
## [1] 5.259613
```

Both F-stastistics are 5.2596.

8.

Since a bigger F-statistic means that there is evidence that the means are different we use that for the observed statistic. We can then compare which F-stastistics are greater than our observed statistic.

why is it upper tail test? think about the hypothesis we are testing SSTr captures bigger ssr means bigger F

9.

```
sum(teststat >= teststat.obs)/1000
```

```
## [1] 0.011
```

Since it is random the p-value returns values ranging from 0.002 to 0.008 or so. All of the values reject the null hypothesis. We can conclude that at least one of the treatments has a different mean runtime than the others.