Movie ratings

The file `moviesall.txt` contains various information on movies released in the year 2003. Altough it is not really a random sample of movies, we will treat it as a random sample of movies for the purpose of this lab. You may think of it as representative of movies in general in the past decade.

For each movie, we have it's rating (G, PG, PG-13, R), it's genre, it's box-office gross (in millions of dollars), it's run time (in minutes), and it's score on `rottentomatoes.com` (higher scores mean better movies). For this lab, you will investigate differences in run time between the movie ratings. Summary information about runtimes is given below:

	n	sample mean	sample sd
G	4	82.75	12.58
PG	21	95.33	10.78
PG-13	65	110.69	24.09
R	50	110.60	19.06

Analysis of Variance (ANOVA)

The parametric method for comparing a quantitative variable between K groups is Analysis of Variance, or ANOVA. ANOVA, like the t-test, is a test about population means. Our hypotheses are as follows:

H_0 : All populations have the same mean. $(\mu_1 = \mu_2 = \cdots = \mu_K)$

H_1 : At least one population has a different mean. ($\mu_i \neq \mu_j$ for at least one pair (i, j))

Notice that ANOVA does an overall test for a difference between the group means. It does not consider which mean(s) are different or the direction ($<$, $>$) of the difference.

The test-statistic for ANOVA is the F-statistic, which is usually calculated through construction of an ANOVA table. Given a sample of size n_i from group i, with sample mean $\overline{X}_i$ and standard deviation S_i, we can construct the following table:

	SS	df	MS	F
Treatment	$SSTr = \sum_{i=1}^{K} n_i(\overline{X}_i - \overline{X}_{overall})^2$	$K-1$	$MSTr = \frac{SSTr}{K-1}$	$F = \frac{MSTr}{MSE}$
Error	$SSE = \sum_{i=1}^{K}(n_i - 1)S_i^2$	$N-K$	$MSE = \frac{SSE}{N-K}$	
Total	$SSTotal = SSTr + SSE$	$N-1$		

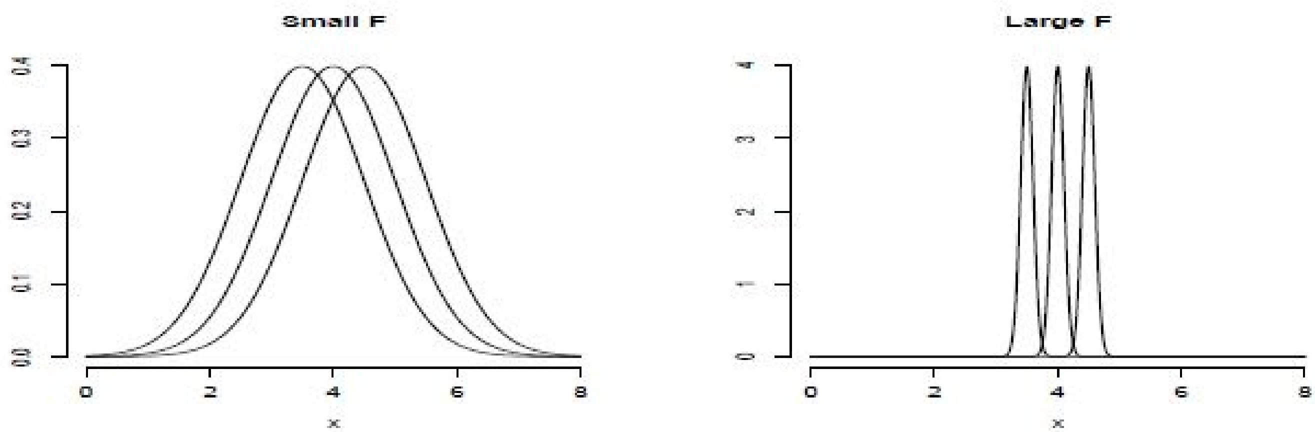
Here the SS column stands for 'sum of squares', the *df* column is degrees of freedom, and the MS column stands for 'mean square'. So $SSTr$ in the 'sum of squares treatment' and MSE is the 'mean square error'. Also, $N = n_1 + n_2 + \cdots + n_K$ is the total number of observations from all of the groups.

1. Construct an ANOVA table for determining whether movies with different ratings have the same mean run time. That is, construct a blank table like the one above and fill in all of the values of $SSTr$, SSE, etc. You might find the following code helpful as a starting point:

```
n = c(4, 21, 65, 50); N = sum(n); K = length(n)
xbar = c(82.75, 95.33, 110.69, 110.60)
s = c(12.58, 10.78, 24.09, 19.06)
xbar.overall = sum(n*xbar)/N

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

You can think of the F-statistic as a measure of the difference in means between the groups ($SSTr$), relative to the variability within the groups (SSE). If F is large, then we have evidence the means are different. If F is small, then we do not have evidence the means are different.



Provided certain assumptions are met, this F-statistic has an F-distribution with $K-1$ and $N-K$ degrees of freedom. Since large values of F give us evidence against H_0, we can calculate our p-value as `1 - pf(F, K-1, N-K)`.

2. Calculate a p-value for testing whether different movie ratings have the same mean run time. State a conclusion based on your p-value.

ANOVA in R

We can also let R produce the ANOVA table and calculate the p-value for us, using the following code:

```
runtime.anova = lm(runtime ~ rating)
anova(runtime.anova)
```

3. Use R to run an ANOVA of runtime on movie rating. Confirm that the test statistic and p-value that R gives you is the same as the one you calculated by hand.

ANOVA Assumptions

The fact that F has an F-distribution depends on the following three assumptions:

- The K samples are randomly selected from the K populations of interest.
- Each of the K populations have a normal distribution.
- All K populations have the same variance.

Now you will check these assumptions with respect to the movie data.

4. Check the normality assumption by producing histograms of run time for each group. Comment on whether you feel this assumption is violated.

5. Check the assumption of equal variance by producing side-by-side boxplots of run time for the different groups. Comment on whether you feel this assumption is violated.

   ```
   boxplot(runtime ~ rating)
   ```

The permutation F-test

The F-statistic is a good measure of the difference in means between the groups, but we cannot rely on p-values calculated from the F-distribution if the above assumptions are violated. Instead, we can perform a permutation version of the F-test.

The hypotheses for a permutation F-test are as follows:

$H_0 : F_1(x) = F_2(x) = \cdots = F_K(x)$

$H_1 : F_i(x) \leq F_j(x)$ or $F_i(x) \geq F_j(x)$ for at least one pair (i, j), with strict inequality for at least one x.

Now you will perform this permutation F-test:

6. How many possible assignments do you need to consider to perform an exact permutation test in this case?

With this large of a data set, we will use random sampling of the permutations to calculate an approximate p-value. Consider the following code. This code is different than the code we have previously used, so be sure you understand what each line is doing.

```
### calculate the observed F-statistic
runtime.anova = lm(runtime ~ rating)
teststat.obs = summary(runtime.anova)$fstatistic[1]

teststat = rep(NA, 1000)
for(i in 1:1000) {

### randomly "shuffle" the rating labels for the movies
ratingSHUFFLE = sample(rating)

### compute the F-statistic for the shuffled data
SHUFFLE.anova = lm(runtime ~ ratingSHUFFLE)
teststat[i] = summary(SHUFFLE.anova)$fstatistic[1]

}
### calculate the approximate p-value
sum(teststat >= teststat.obs)/1000
```

7. Convince yourself that the first section of this code does indeed calculate the observed test statistic by verifying that it gives you the F-statistic you calculated in #1 and #3.

   ```
   ### calculate the observed F-statistic
   runtime.anova = lm(runtime ~ rating)
   teststat.obs = summary(runtime.anova)$fstatistic[1]
   ```

8. Explain why we calculate the p-value by `sum(teststat >= teststat.obs)/1000`. Will this always be the case, or are there circumstances where this might change?

9. Run the entire chunk of code to calculate a p-value. Report your p-value and state a conclusion.