1. For this problem, use the M and M's data set from the University of Puget Sound Data Hoard (link on course calendar). The goal of this problem is to write a simulation to see whether a global $F$ test is suitable for a model of M and M mass modeled on type (a 1-way ANOVA).

(a) Read the background of the data set on the website, so that you know what you're working with.

(b) Inspect simultaneous density plots of mass grouped by type. You should see some issues with the equal variance sampling variability assumption already, since the residuals of the model you are about to fit are just shifted versions of the mass.

(c) Fit a linear model of mass modeled on type.

(d) Inspect simultaneous density plots of the residuals of the fitted model grouped by type. Now you can really see some issues with the equal variance sampling variability assumption.

(e) In the next several steps you'll write a simulation to assess the effects of these violations of the assumption. To start this, use the aggregate() function to compute (and store) the standard deviation of mass for each of the three types of M and M. For our purposes here, we'll call these $s_1$, $s_2$, and $s_3$.

(f) Use the table() function (or some other function) to find (and store) the counts of the three types of M and M's in the data set. For our purposes here, we'll call these $n_1$, $n_2$, and $n_3$.

(g) Now figure out how to do one simulation. For each of the three types, use the rnorm() function to draw (and store) $n_1$ values from a normal distribution with mean 0 and standard deviation $s_1$. Do the same for the second and third types of M and M's too, again using a mean of 0 and the appropriate standard deviation. Call the three vectors that you make this way simulatedPeanut, simulatedPB, and simulatedPlain.

(h) Combine the three simulated vectors into a single vector with

```
simulatedMasses <- c(simulatedPeanut, simulatedPB, simulatedPlain)
```

(i) Make a “grouping vector” that tells the types of the values in the simulatedMasses vector with:

```
simulatedTypes <- c(rep("peanut", n_1), rep("peanut butter", n_2), rep("plain", n_3))
```

You will need to fill in the names above for whatever you called $n_1$, $n_2$, and $n_3$ in R.

(j) Fit a linear model of simulatedMasses modeled on simulatedTypes, and call the fitted model simulatedModel.

(k) The $p$-value for a global $F$-test can be extracted from this fitted model with:

```
anova(simulatedModel)[1, "Pr(>F)"]
```

You have now successfully simulated a data set for which the null hypothesis of the global $F$-test is known to be true: all the simulated masses came from normal distributions with the same means. The simulated sample sizes in each type are the same as the sample sizes in our data set. Also, the standard deviations of the simulated masses match our estimates for the standard deviations of the populations of M and M's of the three types. These standard deviations, however, are not equal, so we can use this simulation to assess whether or not a significance level of 0.05 retains its usual meaning (which is that on average about 5% of the time the test will flag the results as being statistically significant evidence against the null hypothesis).

Now that you have simulated such a data set once, do it $N = 1000$ times and count the number of times that the resulting $p$-value shows statistically significant evidence against the null hypothesis at the usual 0.05 significance level. Use `set.seed(100)` immediately before the loop that runs the simulation.

Interpret what you find from your simulation. In particular, state what effect (if any) this violation of an assumption appears to have on the global $F$-test in this situation. Also, state whether you think that you should proceed with such an $F$-test in this particular case or not.

You do not need to include any R code or output for the above enumerated steps (where you learned to do a single simulation). However, do cut and paste your R code and output for your full simulation at the end of your write-up for this problem.
2. For this problem, use the data set given at:

http://www.statsci.org/data/general/poison.html

Assuming the results of this experiment can be extended to a broader population of animals, which treatment should be used to increase survival time in cases where the type of poison ingested (from among these three) is not known?

You should conduct a full analysis of this problem for your own edification, but in your write-up include only a Tukey comparison of all possible treatment levels (with the associated adjusted \( p \)-values, confidence intervals, and compact letter display grouping) and your interpretation of that. Also cut and paste the R code and R output for this part of your analysis at the end of your write-up of this problem. (The other aspects of the analysis are important too, but because this is only meant to be a short homework assignment rather than a full project, I will not ask you to write those up here.)

A few things are noteworthy on this problem:

(a) You will need to convert both \texttt{Poison} and \texttt{Treatment} to factors immediately after you read the data set into R, since they are labeled by number.

(b) As per the suggestion on the website, you will need to transform \texttt{Time} to \( 1/\text{Time} \) and use \texttt{that} as the response variable in your model. Otherwise the sampling variability assumptions will be way off and the analysis won’t be appropriate.

(c) Although there are ways to get at this using the \texttt{TukeyHSD()} command, I don’t want you to approach this problem in that manner. For this problem, you are instead required to use the \texttt{multcomp} R package for the multiple testing.