# CMDA-2006 Stat Exam 1

Due March 20th (11:59am Noon) as a .pdf upload

You may discuss the problems with your classmates. However, your work must be your own! **Copying anyone else's work is a violation of the honor code.**

To discourage any temptation of copying, every student has a **unique** dataset for each data problem. Therefore, every student will have different answers and potentially different conclusions!

**Therefore copying will be caught very easily, so just don't do it!**

## Primary Instructions:

**Download the .zip file (that contains your data) corresponding to your PID from Canvas under: Files > Exams - Stat > Exam 1 > Data Files, then complete the exam to the best of your ability.**

---

Be efficient and Don't Panic! This exam is meant to be done 100% by R and in very few lines! You've seen how to do all of these types of questions in class notes or solutions by now. Be sure to justify your methods and write appropriate conclusions.

You must show your work in your R Markdown write-up, meaning you need to show the code that you have used to solve the problems. There should be zero physical pencil-and-paper writing for this exam, everything should be typed. You are required to comment your code to explain what's happening.

The use of R Markdown is required and you must compile your document directly to a .pdf (no conversions from HTML or Word). Don't forget to properly **answer the questions!**

---

Finally, submit your solutions in a formal writeup `.pdf` file name of the form (`Lastname_Firstname_Stat_Ex1.pdf`).

Additionally, you must submit your functions for problems 3 and 4 in a separate .R file of the form: (`Lastname_Firstname_Stat_Ex1.R`) (These will be scanned and compared to your classmates for honor code violations.)

Every problem identifies how much it is worth in total.

## Problem 1 (10 pts)

Methamphetamine comes in several forms and can be smoked, inhaled (snorted), injected, or orally ingested. The time it takes for a euphoric response after a release of dopamine (rush) in the brain, regardless of method of delivery, is exponentially distributed.

Suppose we wish to compare the dopamine release time when methamphetamine is smoked versus injected in rats. It is hypothesized that smoking might lead to a quicker dopamine release than injection.

In order to investigate each method of delivery was run 12 times, and the time it takes for a dopamine rush is measured. The results, in milliseconds, are found in `meth.csv`.

(a) Based upon the data, what method should you use to analyze the data? Use plots and statistical tests to justify your answer.

(b) State the null and alternative hypothesis in words and/or symbols.

(c) Using the method that you selected and justified, carry out that method. Test at the 5% level. Show the appropriate R output.

(d) State the conclusion of the test and provide an interpretation of the conclusion in the context of this particular problem.

## Problem 2 (20 pts)

In August and September 2005, Hurricanes Katrina and Rita caused extraordinary flooding in New Orleans, Louisiana. Many homes were severely damaged or destroyed, of those that survived, many required extensive cleaning. It was thought that cleaning flood-damaged homes might present a health hazard due to the large amounts of mold present in many of the homes. The article "Health Effects of Exposure to Water-Damaged New Orleans Homes Six Months After Hurricanes Katrina and Rita" (K. Cummings, J. Cox-Ganser, et al., American Journal of Public Health, 2008:869-875) looked at a sample of residents who had participated in the cleaning of one or more homes and a sample of residents who had not participated in cleaning.

Some members of each group experienced symptoms of wheezing.

The data can be found in `hurricane.csv`.

The focus of this study is to compare the proportion of people who develop wheezing symptoms in the two population groups (those who participated in cleanup and those who did not).

(a) Make a table of the results.

(b) Generate a 95% confidence interval (using the Agresti Method) for the difference in the proportions for those with wheezing symptoms in the two groups. (This cannot be done directly with any of the R functions that I have taught you, but can be done in a couple of lines that you write yourself easily enough.) Be sure to interpret this confidence interval in the context of the study.

(c) Suppose someone makes the claim that the frequency of wheezing symptoms is greater among those residents who participated in the cleaning of flood-damaged homes? State the null and alternative hypotheses using symbols for this situation.

(d) Conduct the hypothesis test in part (c) using a $z$-test at the 5% significance level. Be sure to state the conclusion of the test and provide an interpretation of the conclusion in the context of this particular problem.

(e) Instead of conducting a $z$-test to answer the question in part (c), conduct a chi-square test instead.

## Problem 3 (40 pts)

After reading the article "Determination of Carboxyhemoglobin Levels and Health Effects on Officers Working at the Istanbul Bosphorus Bridge" (G. Kocasoy and H. Yalin, Journal of Environmental Science and Health, 2004:1129-1139), you felt compelled to conduct your own study. This study much like what was presented in the above paper is concerned with assessing health outcomes of people working in an environment with high levels of carbon monoxide (CO).

To obtain data, you contacted a regional factory that routinely has workers who work in one of three shifts (Morning, Evening, Night). This factory also keeps a strict record of employees who report various medical ailments that may or may not be job related. These symptoms include Influenza, Headache, Weakness, and Shortness of Breath.

The data, found in `workerhealth.csv`, contains observations noting the shift of the worker and the symptom being reported.

Can you conclude that the proportions of workers with the various symptoms differ among the shifts? Test at the 1% level.

To answer this question, do the following:

(a) Write a function in R called `my.chi.test()` that can carry out both chi-square tests for a single categorical variable and chi-square tests for contingency tables and has exactly the following format (I must be able to call your function):

```
# Usage:
#  x is a vector of categorical observations
#  y is a vector of categorical observations
# If only x is given, then this is a chi-squared goodness of fit test
#  the default probability vector is equal probabilities (shown below),
#  you'll have to specify a different p = pvec for different probabilities
# If x and y are given, your function should automatically make the table of observations.

my.chisq.test <- function(x, y = NULL, p = rep(1/length(x), length(x)) ){
    your code goes here...
    return( list("Xsquared" = xs, "df" = df.xs, "Pvalue" = pvalue) )
}
```

Do not write your function in a separate .R file, keep it in your main write-up. Of course your function should not call the `chisq.test()` function or anything similar in R.

(b) Make a table of the results for all of the data. Additionally add the marginal totals and grand total as well. The `table()` function will give you a partial answer.

(c) Plot the data using a stacked relative frequency barplot with `shift` on the $x$-axis.

(d) State the null hypothesis (using symbols only) and alternative hypothesis (using words only).

(e) **You must use your R function** to compute the test statistic, the degrees of freedom, and the P-value. Of course, you can check your answer with `chisq.test()` to make sure your function is working. Show the R output when using your function on the data.

(f) State the conclusion of the test and provide an interpretation of the conclusion in the context of this particular problem.

(g) Regardless of the time of shift, it has been hypothesized that the proportion of workers who report influenza is 25%, headaches is 40%, weakness is 20%, and shortness of breath is 15%. Is the data consistent with this hypothesis?. Test this claim, by first stating the null hypothesis (in symbols) and alternative hypothesis (in words). Show the appropriate data that is used as evidence. Then use your function (show the output) to conduct the test at 5% significance level. Comment on your results.

## Problem 4 (20 pts)

Write a function for carrying out the $F$ Test of equal variances. Only the two-sided version is needed. The general format for your function is shown below:

```
# Usage:
# x and y are both numeric vectors

my.var.test <- function(x, y){
  your code goes here...
  return( list("Fstat" = Fs, "df1" = df1, "df2" = df2, "Pvalue" = pval) )
}
```

Of course your function must not call `var.test()` or any similar function in R.

Use your function to answer the following questions.

A procurement specialist has purchased 25 resistors from vendor 1 and 35 resistors from vendor 2. The data can be found in `procurement.csv`.

(a) What distributional assumption is needed to test the claim that the variance of resistance of product from vendor 1 is not significantly different from the variance of resistance of product from vendor 2? Perform a graphical procedure to check this assumption.

(b) Perform an appropriate statistical hypothesis-testing procedure to determine whether the procurement specialist can claim that the variance of resistance of product from vendor 1 is significantly different from the variance of resistance of product from vendor 2. **You must use your function to answer this question.** Test at the 5% significance level.

## Problem 5 (10 points)

Using the data found in `reactions.csv` correspond to reaction times (in seconds) for two different reagents in a laboratory setting. Answer the following questions:

(a) Determine and interpret a 95% confidence interval for the **median** reaction time for reagent 1.

(b) Determine if reagent 1 has a different **median** reaction time than reagent 2. Test at the 5% significance level. (Hint: This is not a test of the means and there is no standardization involved here. You are determining how likely your original differences of the medians are amongst all repeated re-samples.)