Quantitative Methods II
Assignment 4
September 18, 2011

This is the fourth homework assignment for the course. Its purpose is to continue increasing your proficiency in using a statistical program, producing presentation-worthy graphics, and writing to get your point across.

Remember that all statistics are an attempt to gather information about a process (or population) from a sample of data generated by the process. As such, we will never be able to answer a statistical question with absolute certainty, only with statements of confidence ranges and of expected Type I Error rates.

This assignment covers Monte Carlo methods. Remember: If you need assistance using `R`, *do not* hesitate to ask for it. To get the most out of such assistance, you will need to explicitly explain your issue, attach the **code** you have already written, and start asking earlier than Sunday.

⋆ ⋆ ⋆

When you hand in this assignment, you will email to me two separate files, your typed solutions to the questions asked in the homework and a separate script file. The script file allows me to check that you did the correct analysis. The solution file allows me to see that you can answer the questions in complete and coherent sentences, weaving in graphics and statistics appropriately.

The email must include, as its subject line:

`POLS6123: Assignment 4`

Note. *Make sure you include neither code nor raw results in the write-up. The code needs to be attached to the email in the separate script file.*

Problem: The t-test: Is 30 enough? [[10]]

During Chapter 2, we discussed the inappropriateness of the z-test. We also mentioned that many Statistics textbooks state that when the sample size is greater than 30, then you can use the z-test. Here, let us put that to the test. (You may wish to review Chapter 2 before you start.)

As we recall, given a sample from a Normal distribution, plus its population variance, we can properly use the z-test. In Chapter 2, we examined what happened when we did not know the population variance and used the sample variance in its stead (Section 2.2.1). Now, let us know the population variance, but let us not have our sample Normally distributed. In fact, let us assume our sample is Exponentially distributed. Is $n = 30$ still enough? Is $n = 50$ enough?

So, the only two changes to the script listing on Page 38 are in Lines 7 and 9. Line 7: Instead of x being Normally distributed, set

```
x <- rexp(n,1)
```

Line 9: Instead of setting `sigmax` to the sample standard deviation, set it equal to 1 (and the expected value, `mu=1`). (PS: You can also remove Line 8.)

Thus, the core of the Monte Carlo routine will be

```
x <- rexp(n,1)
p[i] <- z.test(x, mu=1, sigmax=1)$p.value
```

Now, create a histogram for $n = 30$ and for $n = 50$. Include the horizontal line and the Kolmogorov-Smirnov test results. Interpret the historgrams and the K-S test results in light of the typical suggestion that "30 is enough."

Problem: Firefighting deaths are a roll of the die [[15]]

As mentioned in Section A.1.5, the Binomial distribution most commonly reflects aspects of reality that interest us. Unfortunately there is no test that allows us to compare the success probabilities of two Binomially-distributed random variables. As such, we need to resort to an approximation using t-tests, a transformation that usually will not work, or non-parametric statistics that have low power.

The final alternative is to use Monte Carlo methods to create an empirical distribution of your test statistic. In this problem, you will create a test statistic related to the null hypothesis, create an empirical distribution for that test statistic, and make an appropriate conclusion with respect to the null hypothesis.

We are interested in fatalities in first responders between two fire stations. These two stations are alike in every way except that Station One has professional firefighters, while Station Two uses volunteers. Thus, our null hypothesis is

$$\mu_1 = \mu_2$$

Again, note that the null hypothesis contains the 'no-effect' position.

We have data going back 20 years for both stations. In those 20 years, Station One had a death in four years. Station Two had a death in eight years. Thus, the distribution of 'deaths in a year' for Station One is approximately

$$X_1 \sim Bin(20, 0.2);$$

for Station Two

$$X_2 \sim Bin(20, 0.4).$$

From whence did these numbers come?

As we wish to determine if $\mu_1 = \mu_2$, what shall be your test statistic? There is no reason to get fancy, just create a test statistic that is close to zero when the null hypothesis is correct.

Next, alter the script in Section A.1.5 as needed. Besides the comments, the only needed adjustment to the script in Section A.1.5 is the following: The value for `pp` in Line 7 needs to be changed. (But, to what?) Also, depending on the test statistic you created, you may have to change Line 14.

Now, test the null hypothesis, making an appropriate three-part conclusion. Finally, include a histogram of your test statistic.