Statistical Methods II

General Comments on the Final Examination

Problem 1: A Flighty Problem

A quick re-reading of this problem shows that I did not tell you how many planes took off and landed before or after deregulation. I also did not specify the total number of miles flown (before or after). Thus, just being given the number of crashes leaves out important information. The reality is that the number of miles flown after deregulation increased. As such, we would expect the number of crashes to increase, as well, *ceteris paribus*.

Problem 2: Where's Old MacDonald?

The same issue arises in this problem, as well. We spent much time in that little room upstairs discussing why we need different experimental designs. Just being told that Farmer Jim gets X bushels per acre and Farmer Bob gets Y bushels per acre does not tell us anything about the fertilizer. Farmer Bob may have some very productive cropland. Farmer Jim may be trying to grow his wheat in a swamp. As such, there is no way we can state (with any level of certainty) that Farmer Jim should change fertilizers.

Problem 3: Airport Fodder

Here, we are faced with a poll stating one thing, and the commentator concluding something not involved in the poll. Just because a person thinks the country is headed in the wrong direction does not mean he or she connects that with the President (or with Congress). Do the Westboro Baptists think the country is headed in the wrong direction?

Furthermore, even if you do connect the "heading in the wrong direction" to the government, the first poll was before the Republicans took over the House, while the second was after they did. It is quite reasonable that Democrats would be more likely to think the country is headed in the wrong direction due to the increase of power of the Republicans.

Finally, Even if a person does think the country is headed in the wrong direction and does blame President Obama, that does not mean the individual will vote for someone else in the next election; that person may think Obama is too right-wing.

Problem 4: A Power Defect

I guess using the word "power" was a good hint. While few of you actually stated the issue was the power of the test Kip performed, all of you described what the problem was. The sample size was too small to detect any appreciable change in production.

Those discussing causality, I let it slide since you got the power issue. However, let me ask you this: How do you define causality?

Problem 5: The Full Monte

Everyone got the process correct: Draw a sample of size 35, perform the t-test, store the p-value, repeat until your computer starts to smoke. Then plot a histogram of the p-values and hope they are uniformly distributed (not normally distributed). The problems were conceptually minor. The distribution to draw from is the uniform distribution. There are two ways of doing this (the way chosen depends on how you were raised, I guess):

```
runif(n)*100
runif(n, max=100)
```

Both methods will give you what you want. The results are that a sample of size n=35 is large enough. In fact, before the advent of cheap computers, we would generate Normal variables from the average of 5 uniform variables (as random-number tables were quite prevalent).

Problem 6: Hit the Upper V

It was pretty clear that the grade distributions in the three sports groups was not Normally distributed. However, there are still assumptions of the Kruskal-Wallis test that need to be tested. Remember that the Kruskal-Wallis test assumes the distributions are identical except for the mean (or median). This assumption is easily tested and needed to be. The conclusion is that there is no statistically significant difference in grades among the three sports groups.

Problem 7: It's in the Gbagbo

This analysis is very similar to those we have done in the past. The one major addition is that you needed to determine which candidate won each region --- not *localite*. There are scripting ways of doing this, but the easiest way is to download the data and do it by hand.

Scripting it does create a better sense of keeping the data pristine. Here is a script to do it

```
lev <- levels(cdi$REGION)
reg.oua <- numeric()
reg.gba <- numeric()

for(i in lev) {
  reg.oua[i] <- sum(cdi$OUATTARA[cdi$REGION==i])
  reg.gba[i] <- sum(cdi$GBAGBO[cdi$REGION==i])
}
```

```
won.ou  <- reg.oua > reg.gba
won.oua <- numeric()

for(i in 1:length(cdi$REGION)) {
 won.oua[i] <- won.ou[cdi$REGION[i]]
}
```

Also, some of you decided to get creative with your model formula. There is no excuse for having your dependent variable appear on the independent variable side. Models have meaning. You should not write out a model without understanding what it means; that is, do not use it in the model if you cannot interpret the results.

Problem 8: Friends, Romans, Countrymen!

(… lend me your ears.) In this final question, you are charged with predicting peer-reviewed funding based on the value of earmarks received and the classification of the university. In other words, the dependent variable is peer-reviewed funding --- not earmark funding. So, prediction functions that fail to provide values for earmarks but provide values for peer-reviewed funding are doomed to failure. You must pay attention to what you write.

It is obvious that an identity link will not work, as many of the values are near a real boundary of zero. It was interesting to see how everyone tried to fit this. While the identity link was right out, the reciprocal link was an option, as was the logarithm link.

As for families, there should not be much difference between the Gaussian family and the Poisson family (fit with quasi-likelihood methods). I'm not sure anything else would work in this situation.

The plots were just as interesting. Some produced a scatterplot with two blobs of color, as the plotting character was too large for the scale chosen. The prediction lines tended to be essentially horizontal, which is what I got.

Remember, just because you think there is a relationship, there may not be one.

Final Thoughts

The script is posted to the website (and accompanies this email). I had fun with this class. I hope you enjoyed it, too.

Ciao!