Statistics 302 Spring 2014 Larget Assignment #11 posted April 24, 2014

Due Wednesday, April 30

Textbook Exercises

8.47, 8.48, 9.18, 9.21, 9.42, 9.59, 9.60

Computer Exercises

For each R problem, turn in answers to questions with the written portion of the homework. Send the R code for the problem to Katherine Goode. The answers to questions in the written part should be well written, clear, and organized. The R code should be commented and well formatted.

R problem 1 In this problem, you will do a simulation exercise to see the distribution of p-values for the t-test for the difference in population means when applied to the samples with the largest and smallest means for a setting where all sample sizes are 8, there are 7 populations (and so 21 possible pairwise comparisons), and the null hypothesis is true with all population means equal to 100 and all population standard deviations equal to 20.

The following function will do the following:

1. Generate random samples.
2. Use `lm()` to fit a linear model.
3. Determine which sample means are largest and smallest.
4. Find the two-sided p-value for the test of the difference of these two population means using

$$\text{SE} = \sqrt{\text{MSE}} \times \sqrt{\frac{1}{8} + \frac{1}{8}}$$

for the standard error and a t distribution with $n - k = 56 - 7 = 49$ degrees of freedom to find the p-value.

```
sim = function(npop = 7, ni = 8, mu = 100, sigma = 20) {
    # Create data
    group = factor(rep(LETTERS[1:npop],each=ni))
    y = rnorm(n=npop*ni,mean=mu,sd=sigma)
    # Find all sample means, and largest and smallest
    all.means = as.vector( by(y,group,mean) )
    max.mean = max(all.means)
    min.mean = min(all.means)
    # Fit the linear model (ANOVA)
    fit = lm(y ~ group)
    # Find the df and sqrt of MSE and the SE
    mse = anova(fit)$Mean[2]
    df = fit$df.residual
    se = sqrt(mse) * sqrt(2/ni)
    # Compute and return p-value
    t.stat = (max.mean - min.mean) / se
    p.value = 2*pt(-abs(t.stat),df)
    return( p.value )
}
```

1. In R, create variables `npop`, `ni` `mu`, and `sigma` with the specified values. Then, enter each line of the function `sim()` into R separately and explain what it does. Display the result of each object created.
2. Write a `for()` loop to run `sim()` 10,000 times, saving the p-values into a vector. (This will take up to a minute to actually run.)
3. Display the distribution of p-values with `ggplot2` and `geom_density()`. Describe the shape of p-values. Are they approximately uniform from 0 to 1, or is the center of the distribution shifted left or right?
4. What fraction of the p-values are less than 0.05?
5. Explain what this simulation result means with respect to interpreting p-values from pairwise comparisons following an ANOVA analysis with regard to the issue of multiple testing.

R Problem 2 This problem will teach you to do many steps in a simple linear regression analysis.

1. Load the data from **InkjetPrinters** into R (`library(Lock5Data); data(InkjetPrinters)`).
2. Plot a scatterplot of the data using *CostColor* as the explanatory variable and *Price* as the response variable. Use `ggplot()`.
3. Fit the simple linear regression model. Print a simple summary. Pull data from the summary to write an expression for the regression line and an estimate of σ. Your expression should be like this, but with numbers instead of a and b.

$$(Price) = a + b(CostColor)$$

```
require(Lock5Data)
data(InkjetPrinters)
fit = lm(Price ~ CostColor, data = InkjetPrinters)
summary(fit)
```

4. Make a plot of residuals versus *CostColor*. Are there any patterns to suggest nonlinearity or nonconstant variance as x changes?

```
resid = residuals(fit)
d = data.frame(CostColor = InkjetPrinters$CostColor, Residuals = resid)
```

5. Use `predict()` to find a 95% confidence interval for the mean price of all inkjet printers where the cost per page of color printing is 10 cents. Verify that the numerical results match those from the equations on page 553.

```
predict(fit,data.frame(CostColor=10),interval="confidence")
```

6. Use `predict()` to find a 95% prediction interval for the price of a single inkjet printer where the cost per page of color printing is 10 cents. Verify that the numerical results match those from the equations on page 553.

```
predict(fit,data.frame(CostColor=10),interval="prediction")
```

7. Briefly explain why the prediction interval is wider than the confidence interval.