Statistics 5303
Spring 2016
Lab #4

Q: Hey mister, how do I get to Carnegie Hall? A: Practice, boy, practice!

Today we want to get practice looking at non-constant variance plots and normal probability plots. The best way to learn what to expect from these plots is to look at a lot of them. So we will. We'll also experiment a little looking at how nonconstant variance affects p-values.

Task 1

From within R give the command:

```
source("http://www.stat.umn.edu/~gary/classes/5303/lab4.R")
```

This file defines the functions we need, and "source" is like typing the file contents into R. You should now have four functions to practice diagnostics with.

Task 2

We start by looking at nonconstant variance plots. You need to get a feel for how those look when variance is constant and when variance is nonconstant. It can be trickier than you'd think. You just read in the `plotncvar` function. A typical use for this function is as follows:

```
plotncvar(nis=c(2,2,4,4,6),vars=c(1,1,2,2,3),reps=10,which=1)
```

The `nis` argument contains the treatment group sample sizes (the n_i values), the vars argument contains the error variance for each treatment group (σ_i^2), the reps argument tells how many random data sets to plot, and which tells which kind of plot to make (types 1 and 3 are useful for detecting nonconstant variance). The function generates data with these sample sizes and counts, fits the linear model, and makes your requested plot of the residuals. This is done reps times. It will pause before every plot, and you'll need to type a "return" or "enter" to move to the next plot. (The function sets the treatment means to 1 through g.)

> Try the `plotncvar` function with different combinations of sample sizes and variances. Get some practice. Try with 3, 5, and 7 groups. Try with all n_is equal, and with them unequal. Try with all variances equal (do you see nonconstant variance when it isn't there?), and try with modest nonconstant variance (factors of about 2, perhaps), and larger factors as well. How easily can you detect non-constant variance?

Task 3

The function `trynpp()` generates data sets and makes a normal probability plot for each data set. It will pause before each plot; press return to see the plot. By default, it will make plots for 20 data sets of normally distributed data with sample size 50 each. You can change the number of plots and size of the data sets. It can also generate other example distributions; distributions are numbered 1 (normal), 2 (long tailed), 3 (short tailed), 4 (skewed to the right), 5 (skewed to the left). A typical use is

```
trynpp(n=20,reps=10,dist=3)
```

which makes 10 plots of short tailed data with 20 points in each plot.

> Using `trynpp()`, look at data sets of size 50 from each of the five distributions to get a feel for what the plots look like, and how much variation there can be in plots from the same distribution. Try again with normal data and samples of size 20 to see how much variation there is in those plots. What about size 10?

Task 4

The function `guessdist()` chooses a random distribution, generates data from that distribution, and then plots the normal probability plot. It will pause after the plot; press return and the plot will be redrawn with a title giving the distribution. Press return again to move on to the next plot. By default, it will make plots for 20 data sets with sample size 50 each. You can change the number and size of the data sets, for example, `guessdist(n=20,reps=10)` makes 10 plots of data with 20 points in each plot.

> Try `guessdist()` with the default size (50), and also with a smaller size (say 20). Can you reliably distinguish between the distributions?

Task 5

There is a function called `nonconvar()` that illustrates one of the problems that arises when we have nonconstant variance. The usage is illustrated in the following:

```
nonconvar(nis=c(2,2,4,4,6),vars=c(1,1,2,2,3),reps=1000)
```

`nonconvar()` will generate `reps` random data sets (here 1000) from the null hypothesis (all means equal); it will do the Anova for each data set; and it will compute the fraction of data sets for which the nominal p-values are .05 or less and .01 or less. These fractions ought to be .05 and .01, and the extent to which they differ from .05 and .01 tells us how much nonconstant variance is affecting our procedure. As in the previous function, the nis argument gives the sample sizes and the vars argument gives the error variances for each group.

> Try this function using a variety of sample size/variance scenarios, including those you used in Task 2 above. Just how sensitive is Anova to nonconstant variance? Can you detect situations where nonconstant variance is, in fact, affecting your inferences?