610 - R1A "Make friends" with your data
Psychology 610, University of Wisconsin-Madison **Prof Colleen F. Moore**

Note: The metaphor of 'making friends' with your data was used by Tukey in some of his writings.

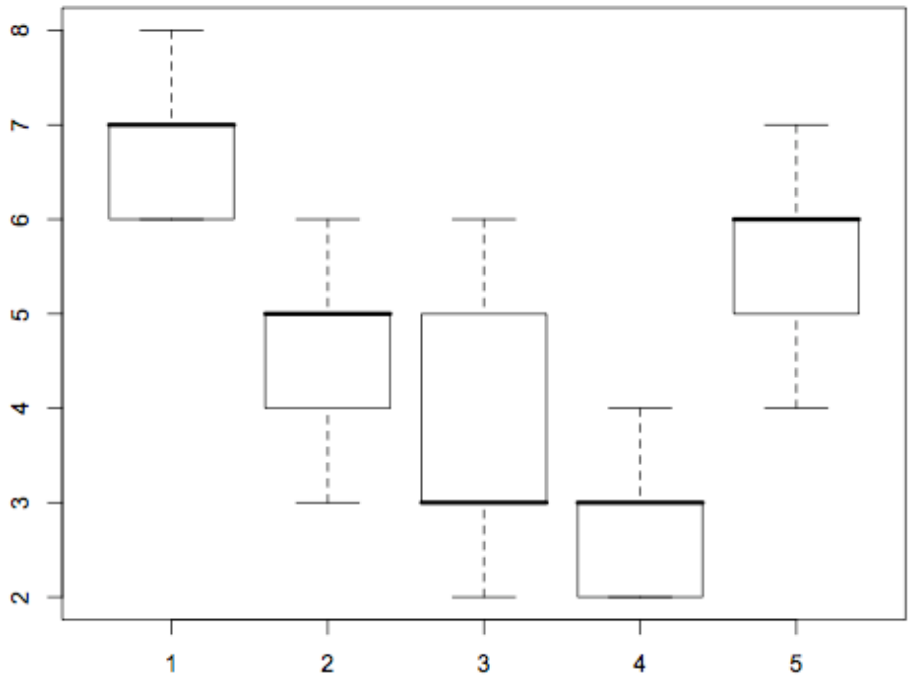
1. Bring the data into R.
We're using the data from the course HO#3 for this example:

```
> datafilename=file.choose()
> your.data=read.table(datafilename,header=T)
> your.data
   group y
1      1 8
2      1 6
3      1 7
4      1 7
5      1 6
6      2 4
7      2 5
8      2 5
9      2 6
10     2 3
11     3 5
12     3 3
13     3 3
14     3 6
15     3 2
16     4 3
17     4 4
18     4 2
19     4 2
20     4 3
21     5 6
22     5 7
23     5 5
24     5 4
25     5 6
```

```
> grp=factor(group)  # because I entered my group codes as numbers, I have to tell R to make it a factor
```

2. Make some graphs of the raw data

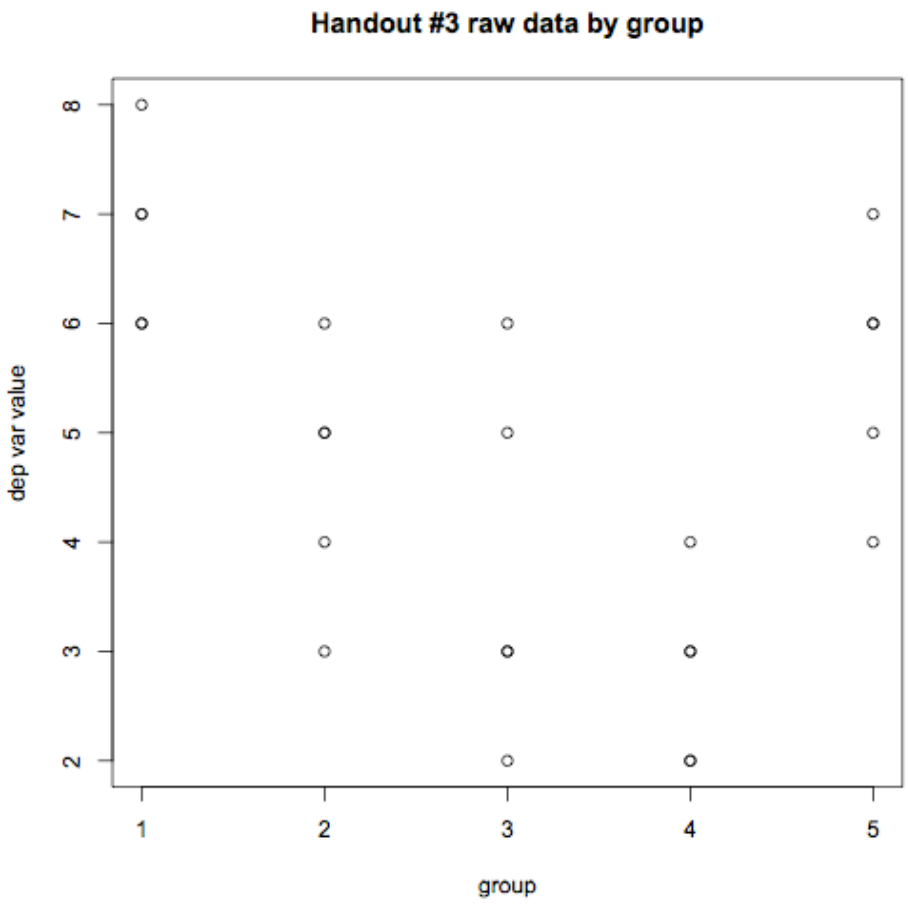
```
> boxplot(y~grp,data=your.data)
```
this makes the boxplot. The bars show the max and min scores, the heavy line shows the median, the box shows the 1st and 3rd quartiles. The means are not shown.



With a small data set, you might want to just plot all the data. You can do this for the groups by making a scatter plot, IF your grouping variable is numerical.

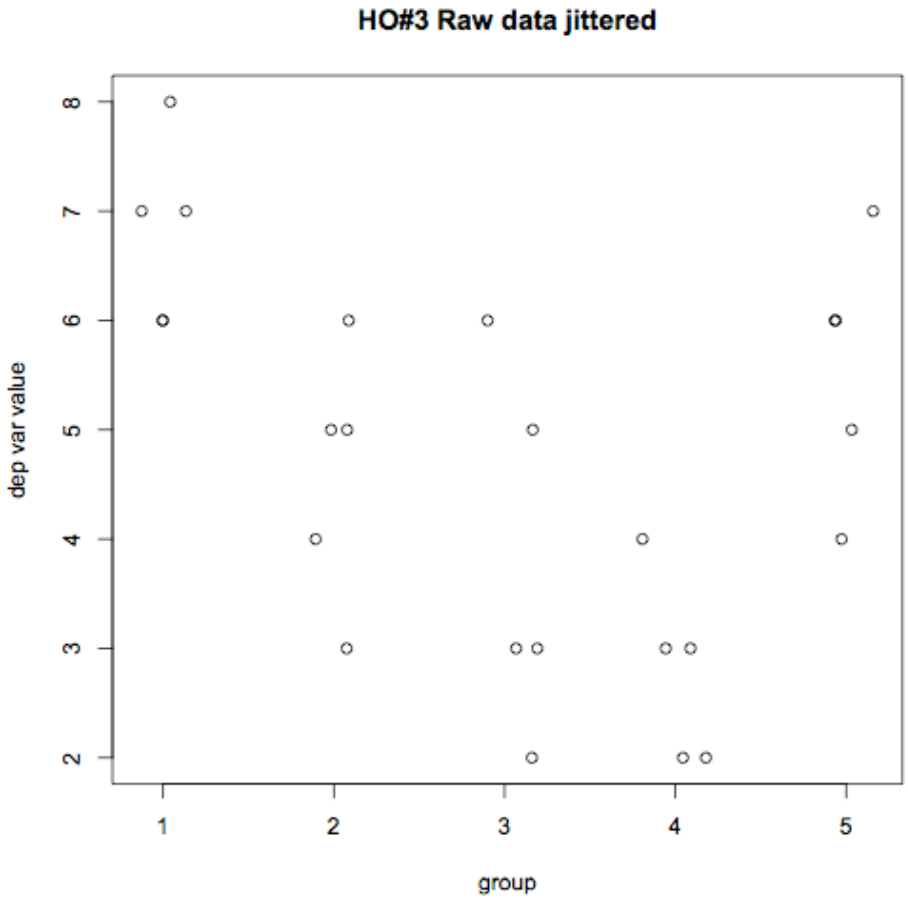
> plot(group, y, main="Handout #3 raw data by group", xlab="group", ylab="dep var value") # note that I use my original grouping variable, "group", not the factor I created called "grp".

Because some scores are the same, not all the data show up. So next I am going to "jitter" the plot so all the individual scores show.



> plot(jitter(group),y,xlab="group",ylab="dep var value",main="HO#3 Raw data jittered")

Eyeball the results below. How does the vertical spread of scores compare across treatments? (the horizontal spread is just due to the 'jitter' so that all the scores show). Group 3 has the largest variance, and groups 1 and 4 have the smallest. With a very small data set it doesn't make much sense



3. Fit the anova model and then plot the residuals. The residuals are supposed to be normally distributed.

> aov.ho3 = aov(y~grp,data=your.data) # This says to do the anova of the y scores with the variable 'grp' as a factor. Store the data in the object called 'aov.ho3'.

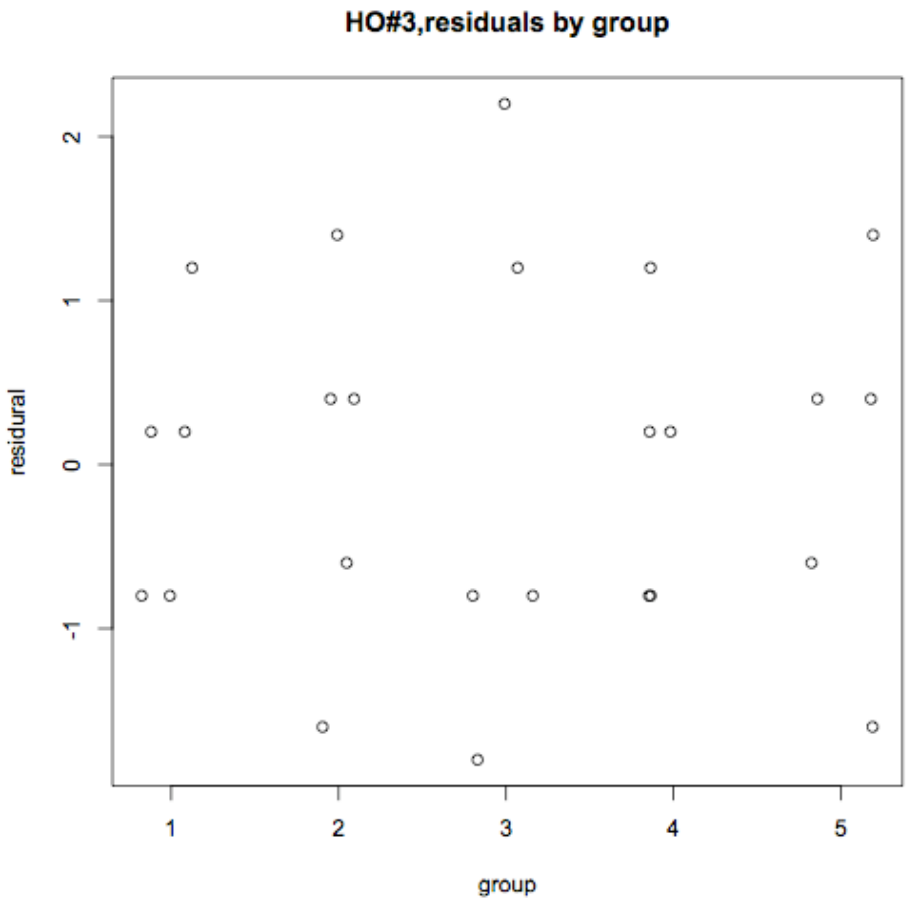
> summary(aov.ho3,intercept=T) # this tells it to make an anova summary table for you:

```
            Df Sum Sq Mean Sq F value    Pr(>F)
(Intercept)  1 556.96  556.96  415.64 7.486e-15 ***
grp          4  48.24   12.06    9.00 0.0002508 ***
Residuals   20  26.80    1.34
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> res=residuals(aov.ho3) # this says to make a variable called 'res' that contains the residuals from the anova that we did called 'aov.ho3'

Now plot the residuals by group. The mean of the residuals will be zero for each group. The spread should be the same as in the original raw data. Now it is easier to eyeball the variances because the group means were removed by fitting the anova model. What's left is the error.

```
> plot(jitter(group),res,main="HO#3,residuals by group", xlab="group", ylab="residual")
```



3. Test some assumptions. See the textbook for comments on the tests and violations of assumptions.

3a. Test homogeneity of variance using Levene's test. This is what spss does for you as an option. You can do this test on the absolute values of the residuals or on the squared residuals from the anova model. I don't know which one SPSS does. One problem with SPSS is the lack of documentation. The 'grp' effect is nonsignificant, so I think homogeneity of variance looks ok.

```
> leveneabs=aov(abs(res)~grp)  # this uses the residuals, named 'res' created above
> summary(leveneabs,intercept=T)
            Df  Sum Sq Mean Sq F value    Pr(>F)
(Intercept)  1 19.3600 19.3600 67.7871 7.455e-08 ***
grp          4  1.7280  0.4320  1.5126    0.2364
Residuals   20  5.7120  0.2856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> levenesq=aov((res^2)~grp)
> summary(levenesq,intercept=T)
            Df  Sum Sq Mean Sq F value    Pr(>F)
(Intercept)  1 28.7296 28.7296 21.6991 0.0001514 ***
grp          4  8.5504  2.1376  1.6145 0.2096017
Residuals   20 26.4800  1.3240
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Or, we can do Bartlett's test for homogeneity of variance:

```
> bartlett.test(y~grp)  # can also use bartlett's test

        Bartlett test of homogeneity of variances

data:  y by grp
Bartlett's K-squared = 2.3953, df = 4, p-value = 0.6635
```

3b. Test normal distribution using Shapiro-Wilk test:

```
> shapiro.test(res)  # I used the residuals from the anova model. I created those earlier.

        Shapiro-Wilk normality test

data:  res
W = 0.9475, p-value = 0.2201
```