PSYC 480 -- Dr. King

Single-Factor Repeated Measures ANOVA

Today we will be working with some data from a Psyc 497 project. You will need to download the repeated.RData file from the website to your Desktop if you haven't already. Then open R, clear the workspace, set the working directory to the Desktop, and attach repeated.RData.

```r
> my.matching = matching
> summary(my.matching)

Gender order          NMTx     CMTx    IRMTx
F:27  CNR:9  Min.  :141.0  Min.  :156.0  Min.  :151.0
M:25  CRN:9  1st Qu.:175.2  1st Qu.:176.8  1st Qu.:167.0
      NCR:9  Median :187.0  Median :192.0  Median :180.0
      NRC:9  Mean   :188.6  Mean   :189.3  Mean   :182.0
      RCN:8  3rd Qu.:202.5  3rd Qu.:201.0  3rd Qu.:192.5
      RNC:8  Max.   :242.0  Max.   :232.0  Max.   :232.0
```

These data are from Wes Rowe (Psyc 497 Spring 2009). Subjects performed a card matching task (a similar task is illustrated below) under each of three different background music conditions. Time required to complete all matches was recorded in seconds. All subjects were CCU students living in student housing and volunteered to participate. Variables are:

- Gender: F=female M=male
- order: the order in which the tasks were presented to the subjects
- Solving times in seconds...
  - NMTx: no music time
  - CMTx: classical music time
  - IRMTx: instrumental rap music time

Note: As there is a column for each level of the IV in the data frame, and no column for the IV, and no subjects identifier, this tells us the data frame is in wide format.

**The Task.** Cards are turned over in pairs. Matching pairs of cards are removed. If the turned cards do not match, they are turned face down again and the subject continues until all the cards have been matched.

Anyone care to state an experimental hypothesis at this time? Frankly, I found the results rather surprising.

What are the IV and DV of interest? What is the importance of the "order" information in the data frame? Which of the variables in the data frame are between groups?

**Graphing.** There are two columns in the data frame that we would not be interested in including in a graph of the data, so we have to take a little care with our graphics commands.
Assumptions. The effect doesn't look impressive in the boxplots. However, the ability of the repeated measures analysis to partial out variability due to differences among subjects makes it a very powerful method for finding even small effects. The price we pay for this power is that the technique is not at all robust to violations of its assumptions. Even small violations can lead to a wrong result.

There are three assumptions we need to be concerned with:
- sampling from normally distributed populations - hard to test
- homogeneity of variance - we will use the usual test
- compound symmetry - explained below

The ANOVA. We're going to put off checking these assumptions until we run the ANOVA. The built-in R functions, at least in the base packages, will not handle data in wide format data frames. Hence, the rmsaov.test function has been created. It is in the repeated.RData file, which must be attached to make it work.

The rmsaov.test function requires the data to be handed to it in the form of a matrix, with the levels of the IV in the columns. We've got a bit of a problem here in that our data frame contains columns that are not levels of the IV. So, we shall first create a matrix consisting of only those columns that contain DV values.

> my.matrix = as.matrix(my.matching[,3:5])
> head(my.matrix)
     NMTx  CMTx  IRMTx
1   218  203  161
2   181  196  167
3   158  164  180
4   169  178  151
5   190  182  166
6   141  169  178
That looks as if it might have worked! So on with the ANOVA.

> rmsaov.test(my.matrix)

Oneway ANOVA with repeated measures

data:  my.matrix
F = 11.5808, df treat = 2, df error = 102, p-value = 2.935e-05

MS.error  MS.treat  MS.subjs
72.00289 833.85256 922.62544

Treatment Means
   NMTx   CMTx   IRMTx
188.5577 189.3269 182.0385

Compound Symmetry Check (Variance-Covariance Matrix)
   NMTx   CMTx   IRMTx
NMTx  441.7025 313.6180 278.8213
CMTx  313.6180 294.2244 258.1833
IRMTx 278.8213 258.1833 330.7044

Estimated Greenhouse-Geisser Epsilon and Adjusted p-value
epsilon    p-adjust
0.805520067 0.000127389

The results of the repeated measures ANOVA are in the first part of the output. We find the null hypothesis to be thoroughly rejected, $F(2, 102) = 11.58$, $p < .001$. Notice that the treatment degrees of freedom follows the usual rule, number of levels of the IV minus one, but the error degrees of freedom does not follow the old rule in repeated measures analyses. The error df should be the treatment (or IV) df times the subjects df (the number of subjects minus 1). There are 52 subjects (each tested 3 times, but still only 52 independent subjects, and 2 treatment degrees of freedom, so $2 \times (52 - 1) = 102$.

No ANOVA table is given in this output, but if we wanted one, it could be reconstructed from the mean squares given in the second part of the output. (Can you do it?) One of those mean squares will be particularly important to us, MS.error (the error term), so make sure you know where to find it. However, another MS that you might want to take notice of is the MS.subjs, variability due to subject differences. If this were a between groups design, all of that variability would be in the error term, and our treatment effect would be utterly swamped. Therein lies the power of repeated measures!

The treatment means are given in the third part of the output. We've already seen them in the summary.

Assumptions Revisited. Next, we find something called a variance-covariance matrix. This allows us to check two of our three assumptions: homogeneity of variance and compound symmetry. The group (condition) variances are on the main diagonal of that matrix, i.e., they begin in the upper left position and end in the lower right position. In this case, those variances are 441.7, 294.2, and 330.7. The ratio of the largest to the smallest is...

> 441.7 / 294.2
[1] 1.50136

So it looks like we are okay on the homogeneity front. Repeated measures analysis adds a new level of complexity, however. Since the three columns of data are from the same subjects measured three
times, those columns should be positively correlated. And as it is assumed that the treatment affects each of the subjects the same way—as otherwise there would be little point in large N research—the three columns should be correlated to the same degree. Covariance is a measure of correlation, and the pairwise covariances appear off the main diagonal in the variance-covariance matrix. These values should also be the same, although they do not have to have the same value as the variances do. The three covariances we have are: N to C 313.6, N to I 278.8, and C to I 258.2. Those don't appear to be too terribly different, but how much difference is too much?

Well, in principle, any! So several correction factors have been invented to try to adjust the p-value for violations. ("Correction factor" is another way of saying "fudge." We should be cautious about putting our complete faith in the ability of these correction factors to completely correct the problem.) One such correction factor is called the Greenhouse-Geisser epsilon. You can see it in the last part of the output.

If there is no violation of compound symmetry, the GGE will be 1, a perfect result and the highest possible value of GGE that it is possible to obtain. The worst possible value of GGE is 1/(k-1), where k is the number of levels of the IV. In this case, the worst possible value of GGE would be 1/2. The value of GGE can be used to "correct" or "adjust" the p-value from the ANOVA test. The adjusted value of p is given along with the GGE value at the bottom of the output. Notice here that we paid a substantial penalty for violation of this assumption, in that our p-value increased by

\[ \frac{0.00127389}{2.935 \times 10^{-5}} \]

4.340341 times. (In fact, this is not such a severe correction. It is fairly typical. Severe violations can result in corrections of tens, hundreds, or even thousands of times.)

(Some technical notes. Use of the GGE correction factor is complicated. I've presented a somewhat simplified version here. Also, the actual assumption we are trying to meet is not compound symmetry but sphericity, a subset of compound symmetry. Sphericity is hard to check, however, so we usually check a more conservative assumption called compound symmetry. Finally, there is a significance test for sphericity, called Mauchly's test for sphericity. Some stat packages--SPSS for example--give the result of this test by default. It's probably not a good idea to pay too much attention to it, as Mauchly's test is not powerful enough to detect violations when sample size is small and tends to overdetect violations, i.e., is too conservative, when sample size is large.)

**Post hoc Tests.** The same post hoc tests are available: Tukey HSD, Fisher LSD, and Bonferroni. They all have to be calculated by hand. (If we had an aov.out object, the TukeyHSD test could be calculated as usual, as was mentioned in the lab.) I propose a Fisher LSD test, as we have only three comparisons and Fisher LSD, as long as it's protected, does not inflate the alpha level with three comparisons. It can be calculated by the method I demonstrated in class when we were talking about between groups designs. Or you can use the `t.contrast` function.

This function requires five pieces of information: the group means, the contrast matrix defining the comparison we want, the number of subjects per group, the pooled SD (square root of MS.error), and the error degrees of freedom. A file containing the function must be attached to use it.

Question: If your graduate adviser gets snippy about there being no p-value adjustment, how would you...
quickly apply the Bonferroni correction here?

```r
> t.contrast(means=c(188.5577, 189.3269, 182.0385), codes=c(1,-1,0),
+     n=52, spool=sqrt(72.00289), dfe=102)
g       SE        t       df  p.value        d
-0.7692   1.6641  -0.4622 102.0000   0.6449   0.0906
```

```r
> t.contrast(means=c(188.5577, 189.3269, 182.0385), codes=c(1,0,-1),
+     n=52, spool=sqrt(72.00289), dfe=102)
g       SE        t       df  p.value        d
6.5192   1.6641   3.9175 102.0000   0.0002   0.7683
```

```r
> t.contrast(means=c(188.5577, 189.3269, 182.0385), codes=c(0,1,-1),
+     n=52, spool=sqrt(72.00289), dfe=102)
g       SE        t       df  p.value        d
7.2884   1.6641   4.3797 102.0000   0.0000   0.8589
```

**Practice, Practice, Practice!** The following practice data sets are available to you in the repeated.RData file.

"baserunning" - These data come from the help(friedman.test) file. The description of them from that source follows (with a few corrections). Hollander & Wolfe (1973), p. 140ff. Comparison of three methods ("round out", "narrow angle", and "wide angle") for rounding first base. For each of 22 players and the three methods, the average time of two runs from a point on the first base line 35ft from home plate to a point 15ft short of second base is recorded. (Times are obviously in seconds.) The obvious question is, is one method better than the others, and if so, by how much?

"chimps" - These data are from an article by Roger Fouts (Acquisition and testing of gestural signs in four young chimpanzees, Science, 1973, 978-980) in which he describes teaching American Sign Language signs to four chimp, two males and two females. Each chimp was taught the same 10 signs (apparently in the same order), and minutes of training required for each chimp to reach a criterion indicating the sign had been learned was recorded for each sign. The questions concern whether some signs were learn consistently more easily than others, and whether there were differences among the chimp in how quickly they learned signs. Data are in wide format that exactly mimics the form in which the data were presented in the original article on page 978. Chimp names and sign names used as column and row names respectively. Be careful! Is this the format you want for testing the effect of signs?

"earthquake" - from the book, Section 18.1

"groceries" - Grocery prices in four local stores in Conway, SC, circa 1999. Two of these stores (B & C) no longer exist, having gone out of business shortly after the appearance of store A. Store D is near a minority neighborhood. The data are in wide format with grocery items in the row names. Are food prices different from store to store?

"headaches" - from the book, Exercise 18.1

"matching" - see above

"rtcolor" - Data are from Josh Weiner Psyc 497 Fall 2009. A reaction time experiment in which he tested each subject for 5 trials on each of four different colored light stimuli. Recorded reaction time is
the mean of those 5 trials for each color by subject. Column names are the color of the stimulus light (and Gender of the subject). Reaction times are in milliseconds. Do subjects respond with different speeds depending upon the color of the stimulus light? What is the potential confound here? Do you know your S & P?

"rttask" - Reaction Time and Decision Complexity. These data were collected in an undergraduate experimental psychology class. Each student in the class did three kinds of reaction time tasks: simple RT (in which the S was required to react as quickly as possible to a stimulus light), disjunctive RT (in which the S was presented with a red or a green light but was instructed to react only to one of them), and choice RT (in which the S was instructed to make one response to the green light and another response to the red light). The responses were to press a key very much like a telegraph key. Until the stimulus light came on, the S had to hold down another key or else the trial was aborted. The numbers in this data set are means of 5 trials in each condition. "RT" is reaction time in milliseconds. Students from two sections of the "lab" (here indicated as A and B) took part in the experiment. (Thus, "lab" is a between-groups variable.) Other variables: "subject" is a subject identifier, "task" gives the type of task, and "color" gives the color of the light the S was responding to. These types of tasks were originally used by Donders about 150 years ago to measure the time it takes for "thoughts" to occur, once believed to be instantaneous. Similar experiments were continued in Wundt's lab after that was established in Leipzig.

"schizo.long" - used in Lab 11

"schizophrenia" - used in Lab 11

"sleepdrugs" - This object is already a matrix. The data are from A.R. Cushny & A.R. Peebles (1905), The action of optical isomers. II. Hyoscines, Journal of Physiology, v.32, pgs.501-510. The data come from a table in that article. The researchers were testing the ability of three hypnotic drugs to produce sleep in people with varying degrees of insomnia. The DV is sleep time in hours. Each patient was measured several times with each drug, and the sleep times are averages of those trials. The object, of course, is to find if the drugs differ in effectiveness. (Note for future psychopharmacology students: L-Hydroscyamine is also called L-atropine. It is the chemical precursor of Hyoscine, which is also called scopolamine, at one time an ingredient in certain cold medicines, but since removed as it is considered too dangerous. Both are muscarinic acetylcholine antagonists.) Historical note: these data were used by Wm. Gosset, aka A.Student, in his development of the t-test.