Repeated Measures ANOVA
Introduction to Statistics Using R (Psychology 9041B)

Paul Gribble

Winter, 2016

1 One-way Repeated-Measures ANOVA

First let's load in the sample data. It's "stacked" in the usual way, where each row is a single observation, and columns code variables — in this case `treatment`, which has four levels, and `subject`, which has 10 levels.

```
> datafilename <- "http://www.gribblelab.org/stats/data/oneWayRepdata.csv"
> mydata <- read.table(datafilename, sep=",", header=TRUE)
> mydata$treatment <- factor(mydata$treatment)
> mydata$subject <- factor(mydata$subject)
> mydata
```

```
   dv treatment subject
1   8         1       1
2  10         2       1
3   7         3       1
4   5         4       1
5   9         1       2
6   9         2       2
7   8         3       2
8   6         4       2
9   7         1       3
10  5         2       3
11  8         3       3
12  4         4       3
13  9         1       4
14  6         2       4
15  5         3       4
16  7         4       4
17  8         1       5
18  7         2       5
19  7         3       5
20  6         4       5
```

```
21  5         1       6
22  4         2       6
23  4         3       6
24  3         4       6
25  7         1       7
26  6         2       7
27  5         3       7
28  4         4       7
29  8         1       8
30  8         2       8
31  6         3       8
32  6         4       8
33  9         1       9
34  8         2       9
35  6         3       9
36  5         4       9
37  7         1      10
38  7         2      10
39  4         3      10
40  5         4      10
```

1.1 demo: ignore subjects

As a demonstration, let's run this *as if* it were simply a between-subjects ANOVA, in other words, ignoring the fact that observations from the four levels of `treatment` come from the same subjects:

```
> am1 <- aov(dv ~ treatment, data=mydata)
> summary(am1)

            Df Sum Sq Mean Sq F value  Pr(>F)
treatment    3   38.9  12.967   6.062 0.00189 **
Residuals   36   77.0   2.139
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case the F-test of the `treatment` effect is significant at $p = 0.001890$.

1.2 demo: include subjects as if it were a two-factor between-subjects design

Now as another demonstration, let's run this ANOVA again *as if* it were a two-factor design, in which we include `subject` as a factor. In this case, note that we cannot include both the main effects *and* the interaction effect — since there is only one observation from each subject in each condition, there are not enough degrees of freedom. So we will simply specify a model that has the two main effects and leaves out the interaction effect:

```
> am2 <- aov(dv ~ treatment + subject, data=mydata)
> summary(am2)

            Df Sum Sq Mean Sq F value   Pr(>F)
treatment    3   38.9  12.967  12.241 3.06e-05 ***
subject      9   48.4   5.378   5.077 0.000471 ***
Residuals   27   28.6   1.059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note how the SS, df, and MS for the `treatment` effect are identical to the previous case where the `subject` factor was not included in the model. What's changed, is that the *error term* (the term denoted `Residuals`) is now smaller. In other words, in the first model (`am1`), the SS was 77.00 — this represents the "leftover" variability in the dependent variable that is unaccounted for by the `treatment` factor. In the `am2` model, the `subject` factor has a SS of 48.4, which is now sliced out of the error term, leaving only 28.6. In other words, the `subject` factor has accounted for a chunk of the variance in the dependent variable, leaving us with a smaller error term in the ANOVA. This is good, because it means that the F-ratio for the `treatment` effect is larger (and hence the p-value is smaller) — since it is equal to the MS for `treatment` divided by the MS for the error term.

Note that this decrease in MS error is not "free", in other words there is a cost: we give up degrees of freedom (was 36 in `am1` but is now reduced to 27 in `am2`). Lower degrees of freedom in the error term means that (all other things being equal) the MS error term will be larger, which (all other things being equal) means a smaller F-ratio for our test of the `treatment` effect. Now all other things are *not* equal, because the inclusion of `subjects` has lowered the error term SS as well ... so whether the inclusion of additional factors helps us, in the end, or not, ultimately depends on the usual question: is the reduction in error gained by the inclusion of the additional term in the model *worth it*, given that we have to "pay" a certain number of degrees of freedom?

1.3 Correct method (univariate approach)

Now the correct way of running a repeated-measures ANOVA is not to pretend that this is a 2-factor between-subjects design, minus the interaction term — (even though in the end, the computations of the ANOVA table are the same). Here is the correct way of doing it:

```
> am3 <- aov(dv ~ treatment + Error(subject/treatment), data=mydata)
> summary(am3)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  9   48.4   5.378

Error: subject:treatment
          Df Sum Sq Mean Sq F value   Pr(>F)
treatment  3   38.9  12.967   12.24 3.06e-05 ***
```

```
Residuals 27   28.6   1.059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the specification of the model has a new element we have not seen before, namely the addition of an error term: `+Error(subject/treatment)`. This notation tells R to slice out of the model an additional error term corresponding to the variance accounted for by subjects. The `subjects/treatment` notation tells R that the `treatment` factor is the repeated-measures factor over which subjects applies.

As you can see the output is essentially the same, just reorganized a bit. The variance accounted for by `subjects` has not only been removed from the model's overall error term, but has even been physically separated in the output. So now, the effect of `treatment` is significant at $p = 3.06e - 05$.

This is the correct method of running a single factor repeated-measures ANOVA, using what is called a *univariate approach*. That is to say, there is a single dependent variable (called `dv` in our data frame `mydata`). We have been doing univariate procedures all along in the course so far.

1.4 Multivariate approach

There is another general method of performing ANOVA when there are one or more repeated-measures factors, called the multivariate approach. Maxwell & Delaney devote a lot of material to going over the rationale. In short, although it is slightly more complex conceptually, one benefit of using a multivariate approach is that it provides ways of testing the *sphericity assumption* — the assumption of homogeneity of variance of differences between groups — as well as providing corrected versions of the F-test assuming the assumption has been violated. As you know from your reading, repeated-measures ANOVA is actually quite sensitive to the sphericity assumption, and what's more, the sphericity assumption is very often violated.

Here is how to perform a repeated-measures ANOVA in R using a multivariate approach. First, we must reorganize the data into a format in which each row represents a single subject, and columns represent levels of the `treatment` factor. This may be familiar to those of you who have used SPSS — this is the way repeated-measures factors are organized in SPSS data tables as well.

```
> response <- with(mydata,cbind(dv[treatment==1], dv[treatment==2],
+          dv[treatment==3], dv[treatment==4]))
```

Our data are now essentially in a matrix format:

```
> response

     [,1] [,2] [,3] [,4]
[1,]    8   10    7    5
[2,]    9    9    8    6
[3,]    7    5    8    4
```

```
 [4,]    9    6    5    7
 [5,]    8    7    7    6
 [6,]    5    4    4    3
 [7,]    7    6    5    4
 [8,]    8    8    6    6
 [9,]    9    8    6    5
[10,]    7    7    4    5
```

We now form the multivariate model using the `lm()` function in `R`:

```
> mlm1 <- lm(response ~ 1)
> mlm1

Call:
lm(formula = response ~ 1)

Coefficients:
             [,1]  [,2]  [,3]  [,4]
(Intercept)  7.7   7.0   6.0   5.1
```

The `1` notation simply tells R that there are no between-subjects factors here ... in other words, only fit the model using intercepts. If you type `mlm1` to look at the model object you will see that four intercepts were fit — one representing the mean of each of the four levels of the dependent variable:

Now we must set up a variable that defines the *design* of our study, which is simple in this case, it's a single factor with four levels:

```
> rfactor <- factor(c("r1", "r2", "r3", "r4"))
```

Now we must load the `car` library into R (we only have to do this once in a given session), because we are going to make use of the `Anova()` function (note the capital `A`, this is different than the `anova()` function). The `Anova()` function calculates ANOVA tables for a number of different kinds of model objects including multivariate model objects.

```
> library(car)
```

We now define a new anova model object `mlm1.aov` by a function call to `Anova()`:

```
> mlm1.aov <- Anova(mlm1, idata=data.frame(rfactor), idesign = ~rfactor, type="III")
```

The first argument, `mlm1`, is our multivariate model defined above. The second argument, `idata=data.frame(rfactor)` passes information about the within-subjects variable, in other words how many levels there are. The third argument, `idesign= rfactor`, passes information about the within-subjects design, in other words that the variable that `rfactor` describes is the repeated-measures variable. The fourth argument, `type="III"`, instructs `Anova()` to calculate the "Type-III" sums of squares when forming the ANOVA table. This

only matters for so-called "unbalanced designs" in which there are different numbers of observations in different groups. Maxwell and Delaney do a good job of explaining how this is related to different ways of computing sums of squares. In the `summary()` command we specify `multivariate=FALSE` because we want to suppress the portion of the output of the `Anova()` function that is related to multivariate statistical tests — these are really only relevant when the experimental design truly is a multivariate design (in other words when there are multiple dependent variables).

Here is the output we get:

```
> summary(mlm1.aov, multivariate=FALSE)

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                SS num Df Error SS den Df       F    Pr(>F)
(Intercept) 1664.1      1     48.4      9 309.440 2.808e-08 ***
rfactor       38.9      3     28.6     27  12.241 3.060e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Mauchly Tests for Sphericity

        Test statistic p-value
rfactor        0.34613 0.14884


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

        GG eps Pr(>F[GG])
rfactor 0.7426  0.0002388 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           HF eps   Pr(>F[HF])
rfactor 0.9981017 3.106252e-05
```

First we see that the test of the repeated measures factor (called `rfactor` here) is significant at $p = 3.06e-05$. Then we see that a test of sphericity was performed (Mauchly's test), which is not significant ($p = 0.14884$). Nevertheless, we get two corrected versions of the F-test: one called Greenhouse-Geisser, and another, slightly less conservative correction, called Huynh-Feldt. In this case, both are also significant (the p-values for GG and HF are marked `Pr(>F[GG])` and `Pr(>F[HF])`, respectively). See your Maxwell and Delaney readings for details about how these corrections are computed.

1.5 Comparisons between individual means

There are two legitimate ways of testing differences between individual means. The first is to use the same F-ratio that we have seen from the between-subjects ANOVA, for testing contrasts ψ:

$$F_{comp} = \frac{\psi^2 / \sum_{j=1}^{a}(c_j^2/n_j)}{MS_{error}} \tag{1}$$

where ψ is the contrast of interest:

$$\psi = \sum_{j=1}^{a}(c_j \bar{Y}_j) \tag{2}$$

So for example if we want to test the mean of group 1 minus the mean of group 2, the weights on the contrast would be $c_1 = (+1)$ and $c_2 = (-1)$. We would simply compute the MS_{error} term based on the ANOVA output. Remember, $MS_{error} = SS_{error}/df_{error}$:

```
> df2 <- 27 # df for error term
> sserr <- 28.6 # ss error term
> mserr <- sserr/df2 # compute mserr
> n <- 10 # num subjects per group
> a <- 4 # num groups
> mresp1 <- mean(response[,1]) # mean of resp1
> mresp2 <- mean(response[,2]) # mean of resp2
> sscomp <- (mresp1-mresp2)^2 # ss comparison
> dfcomp <- 1 # df comparison
> Fcomp <- (n*sscomp/2)/(mserr) # Fobs for comparison
> Fcomp
```

```
[1] 2.312937
```

We can then compute a probability using the `pf()` function:

```
> pcomp <- 1-pf(Fcomp, dfcomp, df2) # pobs for comparison
> pcomp
```

```
[1] 0.1399272
```

In this case, the mean of group 1 is not significantly different than the mean of group 2, $p = 0.1399$. Note that this is a pairwise comparison that is uncorrected for Type-I error. We can, for example, perform this as a Tukey test, by transforming $Fcomp$ into a value of q (studentized range statistic), and then computing a probability (see Maxwell & Delaney, pg. 550, equation 41)

```
> qobs <- sqrt(2*Fcomp) # compute q value
> pt <- 1 - ptukey(qobs, 4, (4-1)*(n-1)) # df for q are (a,(a-1)(n-1))
> pt
```

```
[1] 0.4394847
```

Again, mean 1 is not significantly different than mean 2, $p = 0.5605$.

1.5.1 If you're concerned about sphericity

If you're concerned about violating the sphericity assumption (heterogeneity of variances of differences between groups), then you might notice that the method above is based on using the MS_{error} term from the ANOVA for the denominator of the F-test. Essentially this is a sort of pooled estimate of variance based on all groups. If sphericity is violated, then it may be a better approach to perform a pairwise test in a way that only uses variances of the groups of interest to perform the test. One way to do this is by forming a contrast, based on a set of weights (as above), but that is applied not to the means of each group, but to individual subject scores, to create composite scores ψ_i:

$$\psi_i = \sum_{j=1}^{a} c_j Y_{ij} \tag{3}$$

Then the values of the composite scores ψ_i are evaluated using a t-test. The extent to which they are non-zero reflects the extent to which the contrast is significant. In other words the t-test tests the hypothesis that the composite scores were drawn from a zero-mean population. Here is an example, again comparing group 1 vs group 2:

```
> # fancy matrix multiplication way of computing composite scores:
> mycontrast <- c(+1, -1, 0, 0)
> dscores <- response %*% mycontrast
> # or you can simply do this:
> dscores <- response[,1]-response[,2]
> t.test(dscores)

        One Sample t-test

data:  dscores
t = 1.655, df = 9, p-value = 0.1323
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.2567852  1.6567852
sample estimates:
mean of x
      0.7
```

So the difference between groups 1 and 2 is not significant ($p = 0.1323$). Note that this test is uncorrected for Type-I error. You could always transform a value of the t statistic into an F (when $df_{num} = 1$, $F = t^2$), then into a q statistic (as above) to perform it as a Tukey test.