Two-Way Independent Samples ANOVA with R

Download ANOVA2.txt. The data are those that appear in Table 17-3 of Howell’s *Fundamental statistics for the behavioral sciences* (8th ed.) and in Table 13.2 of Howell’s *Statistical methods for psychology* (8th ed.). Dr. David Howell created these data so that the groups (cells) would have the same means and standard deviations obtained by Eysenck. You should read the original research article: Eysenck. M. W. (1974). Age differences in incidental learning. *Developmental Psychology, 10*, 936 – 941.

The independent/grouping/classification variables are Age of participant (young or old) and Depth of cognitive processing (manipulated by the instructions given to participants prior to presentation of a list of words). The dependent/outcome/comparison variable is number of words correctly recalled later. Startup RStudio and import the data.

```r
> anova2 <- read.table("C:/Users/Vatia/Desktop/anova2.txt", header=TRUE, quote="\"")
> View(anova2)
```

**Simple Descriptive Statistics**

Now to get the sample sizes, means, and standard deviations.

```
> table(anova2$Age, anova2$Depth)

          Adjective Counting Imagery Intentional Rhyming
    Old       10      10      10          10      10
    Young     10      10      10          10      10
```

As you can see, we have a “balanced” design, one where the two factors are absolutely independent of each other. Were you to do a Chi-Square contingency table analysis on these sample sizes you would get \( \chi^2 = 0 \). If the sample sizes are not balanced, the code given here will not do a correct analysis. In that case, you would need to do a least-squares ANOVA.

```r
> aggregate(anova2$Words, by=list(anova2$Age, anova2$Depth), FUN=mean)

          Group.1 Group.2  x
 1   Old  Adjective  11.0
 2   Young Adjective 14.8
 3   Old  Counting  7.0
 4   Young Counting  6.5
 5   Old  Imagery 13.4
 6   Young Imagery 17.6
 7   Old Intentional 12.0
 8   Young Intentional 19.3
 9   Old  Rhyming  6.9
10  Young  Rhyming  7.6
```

© Copyright 2015, Karl L. Wuensch - All rights reserved.
aggregate(anova2$Words, by=list(anova2$Age, anova2$Depth), FUN=sd)

<table>
<thead>
<tr>
<th>Group.1</th>
<th>Group.2</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Old</td>
<td>Adjective</td>
</tr>
<tr>
<td>2</td>
<td>Young</td>
<td>Adjective</td>
</tr>
<tr>
<td>3</td>
<td>Old</td>
<td>Counting</td>
</tr>
<tr>
<td>4</td>
<td>Young</td>
<td>Counting</td>
</tr>
<tr>
<td>5</td>
<td>Old</td>
<td>Imagery</td>
</tr>
<tr>
<td>6</td>
<td>Young</td>
<td>Imagery</td>
</tr>
<tr>
<td>7</td>
<td>Old</td>
<td>Intentional</td>
</tr>
<tr>
<td>8</td>
<td>Young</td>
<td>Intentional</td>
</tr>
<tr>
<td>9</td>
<td>Old</td>
<td>Rhyming</td>
</tr>
<tr>
<td>10</td>
<td>Young</td>
<td>Rhyming</td>
</tr>
</tbody>
</table>

Look at the within-cell standard deviations. In the text book, Howell says "it is important to note that the data themselves are approximately normally distributed with acceptably equal variances." I beg to differ. $F_{\text{max}}$ is $4.5^2 / 1.4^2 > 10$ -- but I am going to ignore that here.

aggregate(anova2$Words, by=list(anova2$Age), FUN=mean)

<table>
<thead>
<tr>
<th>Group.1</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Old</td>
</tr>
<tr>
<td>2</td>
<td>Young</td>
</tr>
</tbody>
</table>

aggregate(anova2$Words, by=list(anova2$Depth), FUN=mean)

<table>
<thead>
<tr>
<th>Group.1</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Adjective</td>
</tr>
<tr>
<td>2</td>
<td>Counting</td>
</tr>
<tr>
<td>3</td>
<td>Imagery</td>
</tr>
<tr>
<td>4</td>
<td>Intentional</td>
</tr>
<tr>
<td>5</td>
<td>Rhyming</td>
</tr>
</tbody>
</table>

The Omnibus Factorial ANOVA

eysenck <- aov(anova2$Words ~ anova2$Age*anova2$Depth)  
summary(eysenck)

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>anova2$Age</td>
<td>1</td>
<td>240.2</td>
<td>240.2</td>
<td>29.936</td>
<td>3.98e-07 ***</td>
</tr>
<tr>
<td>anova2$Depth</td>
<td>4</td>
<td>1514.9</td>
<td>378.7</td>
<td>47.191</td>
<td>&lt; 2e-16 ***</td>
</tr>
<tr>
<td>anova2$Age:anova2$Depth</td>
<td>4</td>
<td>190.3</td>
<td>47.6</td>
<td>5.928</td>
<td>0.000279 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>90</td>
<td>722.3</td>
<td>8.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Simple Main Effects

Since the interaction is significant, we should do simple main effects analysis. First, let's look at the simple main effects of depth of cognitive processing. We start by creating two subsets of the data, one for the old folk, the other for the young folk.

age.old = subset(anova2, anova2$Age=="Old")
age.young = subset(anova2, anova2$Age=="Young")

Now one ANOVA on each subset.
depth.old = aov(Words~Depth, data=age.old)
\texttt{summary(depth.old)}

\begin{verbatim}
 Df  Sum Sq  Mean Sq  F value  Pr(>F)
Depth       4  351.5   87.88   9.085  1.82e-05 ***
Residuals   45  435.3    9.67
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
\end{verbatim}

$\eta^2 = \frac{351.5}{(351.5 + 435.3)} = .447$

\texttt{depth.young = aov(Words~Depth, data=age.young)}

\texttt{summary(depth.young)}

\begin{verbatim}
 Df  Sum Sq  Mean Sq  F value  Pr(>F)
Depth       4   1354   338.4   53.06 <2e-16 ***
Residuals   45    287     6.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
\end{verbatim}

$\eta^2 = \frac{1354}{(1354 + 287)} = .825$. The effect of depth of cognitive processing is greater among the young than among the old.

We could do pairwise comparisons within each age group, but I don’t think that would be very revealing.

From the other perspective, the simple main effects of age at each level of depth of cognitive processing.

\texttt{depth.1 = subset(anova2, anova2$Depth=="Counting")}
\texttt{depth.2 = subset(anova2, anova2$Depth=="Rhyming")}
\texttt{depth.3 = subset(anova2, anova2$Depth=="Adjective")}
\texttt{depth.4 = subset(anova2, anova2$Depth=="Imagery")}
\texttt{depth.5 = subset(anova2, anova2$Depth=="Intentional")}

\texttt{age.dep1 = aov(Words~Age, data=depth.1)}

\texttt{summary(age.dep1)}

\begin{verbatim}
 Df  Sum Sq  Mean Sq  F value  Pr(>F)
Age       1    1.25    1.250    0.464  0.504
Residuals 18   48.50   2.694
\end{verbatim}

\texttt{age.dep2 = aov(Words~Age, data=depth.2)}

\texttt{summary(age.dep2)}

\begin{verbatim}
 Df  Sum Sq  Mean Sq  F value  Pr(>F)
Age       1    2.45    2.450    0.586  0.454
Residuals 18   75.30   4.183
\end{verbatim}

\texttt{age.dep3 = aov(Words~Age, data=depth.3)}

\texttt{summary(age.dep3)}

\begin{verbatim}
 Df  Sum Sq  Mean Sq  F value  Pr(>F)
Age       1   72.2    72.2    7.848  0.0118 *
\end{verbatim}
Residuals  18  165.6     9.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

age.dep4 = aov(Words~Age, data=depth.4)
summary(age.dep4)

Df  Sum Sq Mean Sq  F value  Pr(>F)
---
Age      1    88.2   88.20    6.539 0.0198 *
Residuals 18  242.8   13.49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

age.dep5 = aov(Words~Age, data=depth.5)
summary(age.dep5)

Df  Sum Sq  Mean Sq  F value  Pr(>F)
---
Age      1  266.4  266.45    25.23 8.84e-05 ***
Residuals 18  190.1   10.56
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Notice that the effect of age is significant at the three higher levels of cognitive processing but not at the two lower levels.

**Interaction Plot**

interaction.plot(anova2$Depth, anova2$Age, anova2$Words, type="b", col=c(“red”, “blue”), pch=c(16, 18), main = "Interaction between Age and Depth of Cognitive Processing")

The plot appears in the lower right pane of RStudio, all scrunched up. Click on “Export” and select “Copy to Clipboard.” We need to make this plot wider, so adjust the width and height and then “Update Preview.”
Notice that the lines are not labelled. There is probably a way in R to get them labelled, but I just copied the image (Copy Plot) into Photoshop and cleaned it up there.
Now I am going to create an interaction plot with confidence intervals. First I need to install and activate the “sciplot” package.

```r
> install.packages("sciplot")
Installing package into ‘C:/Users/Vati/Documents/R/win-library/3.2’
(as ‘lib’ is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/sciplot_1.1-0.zip'
Content type 'application/zip' length 25934 bytes (25 KB)
downloaded 25 KB

library (sciplot)

lineplot.CI(x.factor=anova2$Depth, response= anova2$Words, group= anova2$Age, 
trace.label="Age", xlab="Depth of Processing", ylab="Mean Words Recalled", 
main="Age x Depth of Cognitive Processing")
```

Error: invalid graphics state

I don't know what to make of this error message, but I got the plot anyhow.

I had to use Photoshop to label the lines here too. The plot looks pretty good, but I do not like the fact that the depth of processing groups are arranged in alphabetical order. I am pretty sure there is a way to fix that, but sometimes it is just easier to switch to a different tool, and that I was I did, using Microsoft Office to create the plot at the bottom of this document.
**Interpretation of Main Effects**

Some folks argue that you should never interpret a main effect if it participates in a significant interaction. IMHO, if the magnitude of the interaction is small compared to the magnitude of the main effect, then it may still make sense to interpret the main effect. The magnitude of the interaction here ($\eta^2 = .07$) is small compared to that of depth of cognitive processing ($\eta^2 = .57$), and I'll show you, for pedagogical reasons, how to interpret the main effect of age.

The interpretation of the effect of age is straightforward -- the youngsters recalled significantly more items than did the oldsters, 3.1 items on average. The pooled within-age standard deviation is computed by taking the square root of the mean of the two groups' variances --

$$s_{pooled} = \sqrt{\frac{4.007^2 + 5.787^2}{2}} = 4.977.$$  The standardized difference, $d$, is then $3.1/4.977 = .62$. Using Cohen's guidelines, that is a medium to large sized effect. In terms of percentage of variance explained, $\eta^2 = \frac{SS_{age}}{SS_{Corrected\_Total}} = \frac{240.25}{2667.79} = .09$.

The interpretation of the recall condition means is also pretty simple. With greater depth of processing, recall is better, but the difference between the intentional condition and the imagery condition is too small to be significant, as is the difference between the rhyming condition and the counting condition. The pooled standard deviation within the intentional recall and the counting conditions is $s_{pooled} = \sqrt{\frac{4.902^2 + 1.618^2}{2}} = 3.65$. Standardized effect size, $d$, is then $\frac{15.65 - 6.75}{3.65} = 2.44$, an enormous effect. In terms of percentage of variance explained by recall condition, $\eta^2 = \frac{1514.94}{2667.79} = .57$.

**Interpretation of Interaction**

Look at the plot. The plot makes it pretty clear that there is an interaction here. The difference between the oldsters and the youngsters is quite small when the experimental condition is one with little depth of cognitive processing (counting or rhyming), but much greater with higher levels of depth of cognitive processing. With the youngsters, recall performance increases with each increase in depth of processing. With the oldsters, there is an interesting dip in performance in the intentional condition. Perhaps that is a matter of motivation, with oldsters just refusing to follow instructions that ask them to memorize a silly list of words.

The results show that the youngsters recalled significantly more items than did the oldsters at the higher levels of processing (adjective, imagery, and intentional), but not at the lower levels (counting and rhyming). The tests we have obtained here employ individual error terms -- that is, each test is based on error variance from only the two groups being compared. Given that there is a problem with heterogeneity of variance among our cells, that is actually a good procedure. If we did not have that problem, we might want to get a little more power by using a pooled error term. What we would have to do is take the treatment $MS$ for each of these tests, divide it by the error $MS$ from the overall factorial analysis, and evaluate each resulting $F$ with the same error $df$ used in the overall ANOVA. Our error $df$ would then be 90 instead of 18, which would give us a little more power.
Confidence Intervals for $\eta^2$

If you use the $F$ statistics here to compute confidence intervals for the main effects and interactions you will obtain confidence intervals for **partial** eta-square, not regular eta-squared. I am wary of the partial eta-squared for several reasons, one being that it is possible to get values of partial eta-squared that sum to more than 100%. In fact, with these data, the values of partial eta-squared are .2496 (age), .6771 (condition), and .2085 (interaction). These sum to 1.1352 – wow, we have explained 114% of the variance in the number of words correctly recalled.

The question answered by partial eta-squared is this: Of the variance that is not explained by other effects in the model, what proportion is explained by this effect. The question answered by regular eta-squared is this: Of all of the variance in the dependent variable, what proportion is explained by this effect.

To get the confidence intervals for regular eta-squared, we need adjust the values of $F$, adding back into the error term the sums of squares and degrees of freedom for the other effects in the analysis. For example, for the interaction:

$$MSE = \frac{SS_{Total} - SS_{Interaction}}{df_{Total} - df_{Interaction}} = \frac{2667.79 - 190.3}{99 - 4} = 26.0788$$

$$F = \frac{47.575}{26.0788} = 1.8243\text{ on } 4, 95 df$$

library(MBESS)

```r
library(MBESS)

ci.pvaf(F.value=1.8243, df.1=4, df.2=95, N=100, conf.level=.90)
```

[1] "The 0.9 confidence limits (and the actual confidence interval coverage) for the proportion of variance of the dependent variable accounted for by knowing group status are given as:
[1] 0
[1] 0.1324388"

Oh crap, the confidence interval for the interaction includes zero, even though the $F$ test for the interaction was significant. Well, this can happen with eta-squared, but not with partial eta squared. For partial eta-squared for the interaction

```r
library(MBESS)

ci.pvaf(F.value=5.928, df.1=4, df.2=90, N=100, conf.level=.90)
```

[1] "The 0.9 confidence limits (and the actual confidence interval coverage) for the proportion of variance of the dependent variable accounted for by knowing group status are given as:
[1] 0.06713947
[1] 0.2867359"

For the main effects, you can get the adjusted $F$ values simply by running a one-way ANOVA for each effect.
```r
depth1way <- aov(anova2$Words ~ anova2$Depth)
summary(depth1way)

Df  Sum Sq Mean Sq  F value  Pr(>F)
anova2$Depth  4   1515  378.7    31.21  <2e-16 ***
Residuals    95   1153   12.1

ci.pvaf(F.value=31.21, df.1=4, df.2=95, N=100, conf.level=.90)
[1] 0.4407301
[1] 0.6334927

age1way <- aov(anova2$Words ~ anova2$Age)
summary(age1way)

Df  Sum Sq Mean Sq  F value Pr(>F)
anova2$Age   1   240.2  240.25    9.699 0.00242 **
Residuals   98 2427.5   24.77

ci.pvaf(F.value=9.699, df.1=1, df.2=98, N=100, conf.level=.90)
[1] 0.01980898
[1] 0.186721

Now for confidence intervals for eta-squared for each simple effect.

Depth, Oldsters
> ci.pvaf(F.value=9.085, df.1=4, df.2=45, N=50, conf.level=.90)
[1] 0.2166165
[1] 0.5448327

Depth, Youngsters
> ci.pvaf(F.value=53.06, df.1=4, df.2=45, N=50, conf.level=.90)
[1] 0.7216638
[1] 0.8588249

Age, Counting
> ci.pvaf(F.value=.464, df.1=1, df.2=18, N=20, conf.level=.90)
[1] 0
[1] 0.2109827

Age, Rhyming
> ci.pvaf(F.value=.586, df.1=1, df.2=18, N=20, conf.level=.90)
[1] 0
```
Age, Adjective
> ci.pvaf(F.value=7.848, df.1=1, df.2=18, N=20, conf.level=.90)
[1] 0.04318947
[1] 0.5118951

Age, Imagery
> ci.pvaf(F.value=6.539, df.1=1, df.2=18, N=20, conf.level=.90)
[1] "The 0.9 confidence limits (and the actual confidence interval coverage) for the proportion of variance of the dependent variable accounted for by knowing group status are given as:"
[1] 0.02560128
[1] 0.4816712

Age, Intentional
> ci.pvaf(F.value=25.23, df.1=1, df.2=18, N=20, conf.level=.90)
[1] 0.2861817
[1] 0.7164766

Pairwise Comparisons Among Marginal Means

Note that the effect of depth of cognitive processing is significant for both age groups, but is larger in magnitude for the youngsters ($\eta^2 = .83$) than for the oldsters ($\eta^2 = .45$). Given that the magnitude of both of these simple main effects is much greater than that of the interaction, I think it of some value to conduct pairwise comparisons on the marginal means for depth of processing.

TukeyHSD(eysenck, "anova2$Depth")
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = anova2$Words ~ anova2$Age * anova2$Depth)

<table>
<thead>
<tr>
<th></th>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting-Adjective</td>
<td>-6.15</td>
<td>-8.6439282</td>
<td>-3.656072</td>
<td>0.00000000</td>
</tr>
<tr>
<td>Imagery-Adjective</td>
<td>2.60</td>
<td>0.1060718</td>
<td>5.093928</td>
<td>0.0366538</td>
</tr>
<tr>
<td>Intentional-Adjective</td>
<td>2.75</td>
<td>0.2560718</td>
<td>5.243928</td>
<td>0.0231283</td>
</tr>
<tr>
<td>Rhyming-Adjective</td>
<td>-5.65</td>
<td>-8.1439282</td>
<td>-3.156072</td>
<td>0.00000001</td>
</tr>
<tr>
<td>Imagery-Counting</td>
<td>8.75</td>
<td>6.2560718</td>
<td>11.243928</td>
<td>0.0000000</td>
</tr>
<tr>
<td>Intentional-Counting</td>
<td>8.90</td>
<td>6.4060718</td>
<td>11.393928</td>
<td>0.0000000</td>
</tr>
<tr>
<td>Rhyming-Counting</td>
<td>0.50</td>
<td>-1.9939282</td>
<td>2.993928</td>
<td>0.9806524</td>
</tr>
<tr>
<td>Intentional-Imagery</td>
<td>0.15</td>
<td>-2.3439282</td>
<td>2.643928</td>
<td>0.9998204</td>
</tr>
<tr>
<td>Rhyming-Imagery</td>
<td>-8.25</td>
<td>-10.7439282</td>
<td>-5.756072</td>
<td>0.0000000</td>
</tr>
<tr>
<td>Rhyming-Intentional</td>
<td>-8.40</td>
<td>-10.8939282</td>
<td>-5.906072</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>
Writing up the Results – Here is an Example

A 2 x 5 factorial ANOVA was employed to estimate the effects of age group and depth of cognitive processing (DCP) on participants’ recall of the items. A .05 criterion of statistical significance was employed for all tests. The main effects of age, $F(1, 90) = 29.94, p < .001, \eta^2 = .07$, 90% CI [.02, .19], and DCP, $F(4, 90) = 47.19, p < .001, \eta^2 = .57, 90\%$ CI [.44, .63], were statistically significant, as was their interaction, $F(4, 90) = 5.93, p < .001, \eta^2 = .21, 90\%$ CI [.00, .13]; $MSE = 8.03$ for each effect. Overall, younger participants recalled more items ($M = 13.16$) than did older participants ($M = 10.06$). The Tukey HSD procedure was employed to conduct pairwise comparisons on the marginal means for DOP. As shown in the table below, recall was better for the conditions which involved greater depth of processing than for the conditions that involved less cognitive processing.

<table>
<thead>
<tr>
<th>Recall Condition</th>
<th>Counting</th>
<th>Rhyming</th>
<th>Adjective</th>
<th>Imagery</th>
<th>Intentional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>6.75$^A$</td>
<td>7.25$^A$</td>
<td>12.90$^B$</td>
<td>15.50$^C$</td>
<td>15.65$^C$</td>
</tr>
</tbody>
</table>

Note. Means sharing a letter in their superscript are not significantly different from one another according to Tukey HSD tests.

The interaction is displayed in the following figure. DCP had a significant simple main effect in both the younger participants, $F(4, 45) = 53.06, MSE = 6.38, p < .001, \eta^2 = .83, 90\%$ CI [.72, .86] and the older participants, $F(4, 45) = 9.08, MSE = 9.68, p < .001, \eta^2 = .45, 90\%$ CI [.22, .54], but the effect was clearly stronger in the younger participants than in the older participants. The younger participants recalled significantly more items than did the older participants in the adjective condition, $F(1, 18) = 7.85, MSE = 9.2, p = .012, \eta^2 = .30, 90\%$ CI [.04, .51], the imagery condition, $F(1, 18) = 6.54, MSE = 13.49, p = .020, \eta^2 = .27, 90\%$ CI [.03, .48], and the intentional condition, $F(1, 18) = 25.23, MSE = 10.56, p < .001, \eta^2 = .58, 90\%$ CI [.29, .72], but the effect of age fell well short of significance in the counting condition, $F(1, 18) = 0.46, MSE = 2.69, p = .50, \eta^2 = .03, 90\%$ CI [.00, .21], and in the rhyming condition, $F(1, 18) = 0.59, MSE = 4.18, p = .45, \eta^2 = .03, 95\%$ CI [.00, .22].
Recall in Young and Old Participants

<table>
<thead>
<tr>
<th>Recall Condition</th>
<th>Mean Items Recalled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting</td>
<td>Young: 7, Old: 6</td>
</tr>
<tr>
<td>Rhyming</td>
<td>Young: 9, Old: 8</td>
</tr>
<tr>
<td>Adjective</td>
<td>Young: 12, Old: 10</td>
</tr>
<tr>
<td>Imagery</td>
<td>Young: 15, Old: 13</td>
</tr>
<tr>
<td>Intentional</td>
<td>Young: 20, Old: 17</td>
</tr>
</tbody>
</table>

Return to Wuensch’s R Lessons

Copyright 2015, Karl L. Wuensch - All rights reserved.