ANOVAs

William H. Knapp III

March 26, 2016

1 When to Use ANOVA

If your design compares two or more groups of people and the dependent variable is approximately continuous, or if you have more than one discrete independent varaibles, ANOVAs are probably what you need to analyze your data. There are a number of different flavors of ANOVAs. In this document, we'll consider one-way ANOVAs for within- and between-subjects manipulations. We'll also take a look at two-way ANOVAs for within-, between-, and mixed- subjects manipulations.

2 ANOVAs

We're going to need the ggplot2, gplots, and dplyr packages for this assignment, so let's load them up first.

```
> library(ggplot2)
> library(gplots)
> library(dplyr)
```

We also need to set our working directory. Once that's set, we can read in the data set that we'll use.

```
> dat<-read.csv("example6.csv")
```

This dataframe consists of all the data we'll need to run various analyses. We won't use all of the data for some of our analyses. Unfortunately, this

complicates things a bit for the analyses we'll run here, in example6.Rmd, and homework6.Rmd. But your data for your capstone project will be better organized, which will simplify your analyses. Specifically, the data set contains the percent participants remember for items in various positions in a list when they can take the test immediately or have to perform a distractor task inbetween learning the lists and recalling the information from the lists.

If you look at the structure of the data you'll see that position, subject, and subject two are all integer variables. We don't want that. Instead we want these variables to be treated as factor variables. We can do accomplish this using the following code.

```
> dat$position<-as.factor(dat$position)
> dat$subject<-as.factor(dat$subject)
> dat$subject2<-as.factor(dat$subject2)
```

2.1 One-Way Between-Subjects ANOVAs

You use a one-way ANOVA when you have one independent variable with 2 or more levels. Here we'll examine the serial position effects when there's no distractor.

The previous statment ends with a comma inside the brackets, what this tells R is that we want all of the columns of dat that are associated with no distractor. Checking the structure, you can see that you now have half the data you had previously.

To conduct our ANOVAs we'll use two functions: aov(), which conducts the analysis of variance, and summary(), which presents the results of the ANOVA in an easier to interpret structure. We'll also need to create models of our data. Model creation is fairly simple. You specify the dependent variable and then indicate the variables the DV is supposed to depend on. We do this in R using the following template, which we'll expand as our models become more complex.

```
DV~IV
```

Translated into English the tilda (i.e.) means as a function of, or as predicted by. Thus this model indicates that the DV is a function of the IV.

Although the data were created so we could use within-subjects ANOVAs, by ignoring the subject variables we can perform the between-subjects analysis. Specifically, we'll look to see if there are any differences in how well

participants remember items in different positions of the list. In the between-subjects design, this would mean that different subjects recalled items from the different positions. This wouldn't happen in a real experiment, as all the participants would have data for each position, but this is an example to show you how to run a simple one-way between-subjects ANOVA.

```
> summary(aov(percent~position, data=temp))

             Df Sum Sq Mean Sq F value   Pr(>F)
position      9   8945   993.9   14.45 3.43e-15 ***
Residuals   110   7564    68.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of our analysis, we can see that position does affect the percent of items recalled from the various positions, $F(9,110) = 14.5$, $p < .05$.

2.2 One-Way Within-Subjects ANOVAs

If the same participants participate in all of the conditions in a one-way ANOVA, we need to change our model to account for the subject variable. Specifically, the subject variable will change the error terms for our ANOVAs so we'll modify the basic model as follows.

```
DV~IV+Error(subjectvariable/IV)
```

We'll use the first subject variable (i.e. "subject") to serve as our subject variable for this analysis. You can see in the temp data frame that we only have subjects 1-12, so using subject is the appropriate choice for this analysis.

```
> summary(aov(percent~position+Error(subject/position)))

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 11  10319   938.1

Error: subject:position
          Df Sum Sq Mean Sq F value Pr(>F)
```

```
position   9  14416  1601.8   71.34 <2e-16 ***
Residuals 99   2223    22.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 120  10310   85.92
```

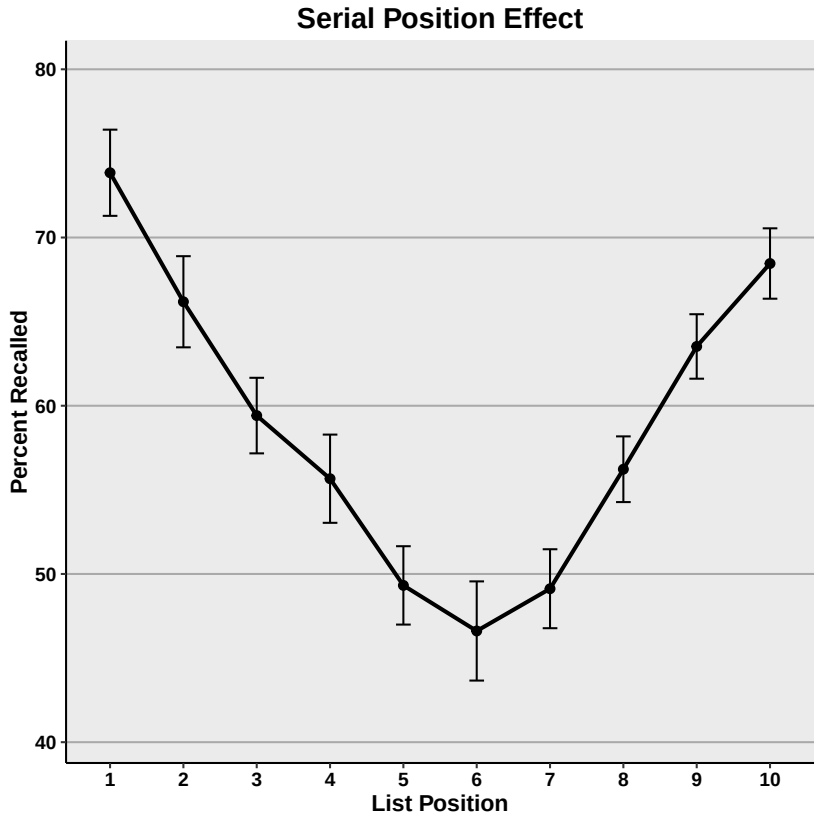
Notice how the within-subjects analysis is much more powerful. From this analysis we can conclude that position affects the percent recalled from words in various positions in lists of words, $F(9,99) = 71.3$, $p < .05$.

Because we have a significant result in both cases, it might be nice to create a figure to go along with our analysis. However, because position is ratio scale variable, we'll use a line graph to display our results. Try to follow along with the following commands and compare it to the commands we used last time to create the bar graphs.

First we'll need to summarize the data to get the means and standard errors of the means for the various positions in the list. Then we can create our figure.

```
> temp<-temp%>%group_by(position)%>%
+   summarize(means=mean(percent),
+             sems=sd(percent)/sqrt(length(percent)))
> f<-ggplot(temp, aes(x=as.factor(position),
+                     y=means,
+                     group=1))+
+   geom_line(size=1)+
+   geom_point(size=2)+
+   geom_errorbar(aes(ymax=means+sems,
+                     ymin=means-sems),
+                 width=.2)+
+   ggtitle("Serial Position Effect")+
+   labs(x="List Position",y="Percent Recalled")+
+   theme(plot.title=element_text(size=15,face="bold",vjust=.5))+
+   theme(axis.title.x=element_text(size=12,face="bold",vjust=-.25))+
+   theme(axis.title.y=element_text(size=12,face="bold",vjust=1))+
+   theme(axis.text.x=element_text(size=10,face="bold",color="black"))+
```

```
+   theme(axis.text.y=element_text(size=10,face="bold",color="black"))+
+   coord_cartesian(ylim=c(min(temp$means)-2*max(temp$sems),
+                          max(temp$means)+2*max(temp$sems)))+
+   theme(panel.border=element_blank(),axis.line=element_line())+
+   theme(panel.grid.major.x=element_blank())+
+   theme(panel.grid.major.y=element_line(color="darkgrey"))+
+   theme(panel.grid.minor.y=element_blank())
> f
```



2.3 2-Way Between-Subjects ANOVA

When we have multiple independent variables, we can observe effects of each of the independent variables and interactions between them. To tell R that we want to include the interactions in our analyses, we'll use asterisks that separate the different variables. Thus, our basic model template takes the following form:

```
DV~IV1*IV2
```

In plain English, the previous says, DV depends on IV1, IV2, and the interaction between IV1 and IV2. As before position is one of our independent variables. The other one is whether or not participants engaged in a distractor task between learning the list and taking a memory test. Thus, we'll need to use all the data again. As this is a between subjects analysis, we can again ignore the subjects variables.

```
> summary(aov(percent~position*distractor, data=dat))

                     Df Sum Sq Mean Sq F value   Pr(>F)
position              9  14416  1601.8   22.70  < 2e-16 ***
distractor            1   2206  2205.6   31.26 6.65e-08 ***
position:distractor   9   5124   569.4    8.07 2.55e-10 ***
Residuals           220  15522    70.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

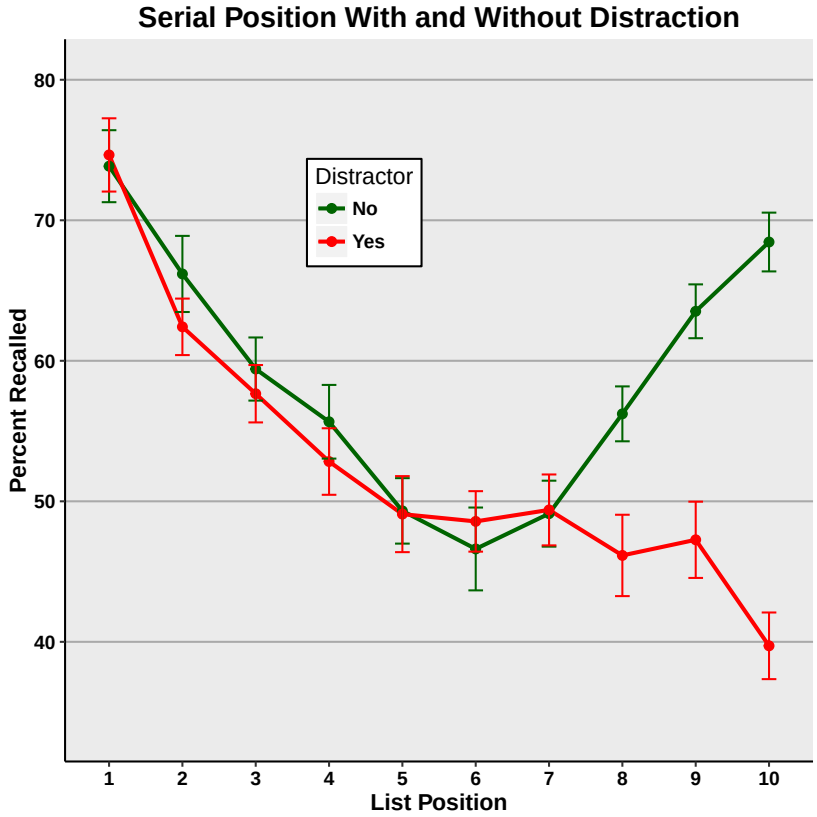
Our analyses indicate that there's a main effect of position, $F(9,220) = 22.7$, $p < .05$. There's also a main effect of distractor, $F(1,220) = 31.3$, $p < .05$. Finally, there's an interaction between position and distractor, $F(9,220) = 8.1$, $p < .05$.

In order to understand the effects and interactions, we should plot the data. Again, we'll use a line graph to display the data. I'll also show you how to create a grouped bar graph incase the IV you're plotting on the x-axis isn't ratio, interval, or an ordinal scale.

Again, we'll need to use dplyr to summarize our data before we can create our figure.

```
> temp<-dat%>%group_by(position, distractor)%>%
+   summarize(means=mean(percent),
+             sems=sd(percent)/sqrt(length(percent)))
> f<-ggplot(temp, aes(x=position,
+                     y=means,
+                     group=distractor,
+                     color=distractor))+
+   geom_line(size=1)+
+   geom_point(size=2)+
```

```
+   scale_color_manual(values=c("darkgreen","red"),
+                      name="Distractor",
+                      breaks=c("no","yes"),
+                      labels=c("No", "Yes"))+
+   geom_errorbar(aes(ymax=means+sems, ymin=means-sems),width=.2)+
+   ggtitle("Serial Position With and Without Distraction")+
+   labs(x="List Position",y="Percent Recalled")+
+   theme(plot.title=element_text(size=15,face="bold",vjust=.5))+
+   theme(axis.title.x=element_text(size=12,face="bold",vjust=-.25))+
+   theme(axis.title.y=element_text(size=12,face="bold",vjust=1))+
+   theme(axis.text.x=element_text(size=10,face="bold",color="black"))+
+   theme(axis.text.y=element_text(size=10,face="bold",color="black"))+
+   coord_cartesian(ylim=c(min(temp$means)-2*max(temp$sems),
+                          max(temp$means)+2*max(temp$sems)))+
+   theme(panel.border=element_blank(),axis.line=element_line())+
+   theme(panel.grid.major.x=element_blank())+
+   theme(panel.grid.major.y=element_line(color="darkgrey"))+
+   theme(panel.grid.minor.y=element_blank())+
+   theme(legend.position=c(.4,.76))+
+   theme(legend.background=element_blank())+
+   theme(legend.background=element_rect(color="black"))+
+   theme(legend.title=element_blank())+
+   theme(legend.title=element_text(size=12))+
+   theme(legend.title.align=.5)+
+   theme(legend.text=element_text(size=10,face="bold"))
> f
```



From examining the figure along with our earlier statistical statements, we can see that when there was no intervening task, participants performed best at the beginning and end list positions. This shows the standard serial position effect. However, when participants have are temporarily distracted, they perform best on the earliest list positions and worst on the most recent positions.

As I mentioned previously, I also want to show you how to create a grouped bar graph that you can use if the variable on your x-axis is a categorical variable.

```
> f<-ggplot(temp, aes(x=position,
+                     y=means,
+                     fill=distractor))+
+   geom_bar(stat="identity",position=position_dodge())+
+   scale_fill_manual(values=c("darkgreen","red"),
+                     name="Distractor",
```

```
+                      breaks=c("no","yes"),
+                      labels=c("No", "Yes"))+
+   geom_errorbar(aes(ymax=means+sems,
+                     ymin=means-sems),
+                 width=.2,
+                 position=position_dodge(.9))+
+   ggtitle("Serial Position With and Without Distraction")+
+   labs(x="List Position",y="Percent Recalled")+
+   theme(plot.title=element_text(size=15,face="bold",vjust=.5))+
+   theme(axis.title.x=element_text(size=12,face="bold",vjust=-.25))+
+   theme(axis.title.y=element_text(size=12,face="bold",vjust=1))+
+   theme(axis.text.x=element_text(size=10,face="bold",color="black"))+
+   theme(axis.text.y=element_text(size=10,face="bold",color="black"))+
+   coord_cartesian(ylim=c(min(temp$means)-2*max(temp$sems),
+                          max(temp$means)+2*max(temp$sems)))+
+   theme(panel.border=element_blank(),axis.line=element_line())+
+   theme(panel.grid.major.x=element_blank())+
+   theme(panel.grid.major.y=element_line(color="darkgrey"))+
+   theme(panel.grid.minor.y=element_blank())+
+   theme(legend.position=c(.4,.76))+
+   theme(legend.background=element_blank())+
+   theme(legend.background=element_rect(color="black"))+
+   theme(legend.title=element_blank())+
+   theme(legend.title=element_text(size=12))+
+   theme(legend.title.align=.5)+
+   theme(legend.text=element_text(size=10,face="bold"))
> f
```

2.4 2-Way Within-Subjects ANOVA

As with the within-subjects one-way ANOVA, we need to indicate which of our variables is associated with subjects and adjust our Error term appropriately. Our model now becomes a bit more complex.

```
DV~All*IVs+Error(subject/(within*subjects*IVs))
```

So let's run an ANOVA using this template examining for effects of both variables and their interaction.

```
> summary(aov(percent~position*distractor+
+               Error(subject/(position*distractor))))

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
```

```
Residuals 11  10319   938.1

Error: subject:position
          Df Sum Sq Mean Sq F value Pr(>F)
position   9  14416  1601.8   71.34 <2e-16 ***
Residuals 99   2223    22.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subject:distractor
           Df Sum Sq Mean Sq F value   Pr(>F)
distractor  1   2206  2205.6   128.4 2.09e-07 ***
Residuals  11    189    17.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subject:position:distractor
                    Df Sum Sq Mean Sq F value Pr(>F)
position:distractor  9   5124   569.4   20.19 <2e-16 ***
Residuals           99   2792    28.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This output is quite a bit more complicated because it breaks up the error variance in a number of pieces. Read my interpretation here and compare what I say with the output from the ANOVA so you can understand where I'm getting the numbers from. From our analysis, we can see that overall position has an effect, $F(9,99) = 71.34$, $p < .05$. In general participants perform better at the beginning and ending list positions. Distractor also has an effect, specifically, participants perform better when there's no intervening distractor task, $F(1,11) = 128.4$, $p < .05$. Additionally, there's an interaction between list position and distractor, F(9,99). As I mentioned before, the interaction shows that people in the undistracted conditions perform better at the beginning and end of the list, while the distracted individuals perform best at the beginning positions only.

2.5 2-Way Mixed ANOVA

When you have one variable that varies within subjects and another variable that varies between subjects your model will be the same as I specified earlier. Consider the situation in which distractor varies between participants, but list position varies within subjects. The subject2 variable was set up to perform this mixed ANOVA. You'll see that participants 1-12 participated with distraction and participants 13-24 participated without distraction. So let's perform our Mixed 2-way ANOVA.

```
> summary(aov(percent~position*distractor+
+               Error(subject2/position)))

Error: subject2
           Df Sum Sq Mean Sq F value Pr(>F)
distractor  1   2206  2205.6   4.618 0.0429 *
Residuals  22  10508   477.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subject2:position
                     Df Sum Sq Mean Sq F value Pr(>F)
position              9  14416  1601.8   63.25 <2e-16 ***
position:distractor   9   5124   569.4   22.48 <2e-16 ***
Residuals           198   5014    25.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this analysis and our earlier graphs we can see that distracted participants performed worse overall than the non-distracted participants, $F(1,22) = 4.6$, $p = .0429$. We can also see that words from earlier positions are generally remembered better than those from later positions, $F(9,198) = 63.3$, $p < .05$. Finally, those who didn't perform an intervening distractor task performed best at the beginning and end positions, while those who were distracted did best at the beginning of the list, $F(9,198) = 22.5$, $p < .05$.

3 Homework

You're homework is to complete homework6.Rmd. You can use code from this document or from my example6.Rmd and this document to complete the assignment. Once you've completed the assignment and have made sure that the document knits properly. Sync the changes with your Github account using the add, commit, and push functions in Git.

4 Summary

After working through these documents, you should have everything you'll need to analyze your data. It should just be a matter of identifying the appropriate analysis and then using the code that I've provided in these documents to conduct appropriate analyses and create appropriate graphs. Good luck and feel free to Skype with me or email me if you're encountering difficulties.

5 Tips for the Future

Although I analyze the same data set in multiple ways to demonstrate how to conduct various analyses, for your projects you'll only perform one primary analysis (e.g. an ANOVA) and potentially follow up analyses (e.g. t-tests) to identify which data points are different from one another.

Similarly, I've created two graphs to show you the same data. For your projects, you'll only need to create a single graph to display the data.