R Final Project

Luis Arriaga

March 19, 2017

Background

Behavioral experiments are often performed by measuring an individual’s response multiple times, varying the treatment each time. This often requires the use of a repeated-measures ANOVA. This is done to control for variation among individuals that could affect the results, or to decrease the number of individuals needed. For example (this example will be used throughout this tutorial), let’s imagine that you want to test whether male fish prefer larger, conspecific females over smaller females. One approach would be to get 30 males to “rate” their preference for a small female by spending time associating with her. Another 30 males would be tested with a medium female, and another 30 with a large female. This requires 90 different males, and it is impossible to control for all possible individual variation among the males. At most, you might be able to control for things like size, weight, or age. With a repeated measures approach, you can control nearly all individual variation by measuring the preferences of a single set of 30 males sequentially (and randomly) presented with small, medium, and large females.

Because of this, you decide to spend two years of your life gathering this data for an extremely endangered fish that lives in underwater tree cavities. It’s a miracle you were able to find 30 males. Unfortunately, as you find out when searching for ways to analyze your data, all the intuitive programs you could use to analyze this data are obscenely expensive and produce terrible graphs anyway. Because R is free and produces very nice graphs, you decide to try to analyze your data with R.

The problem with R is that it is very unintuitive, the help pages are worse than worthless since they end up confusing you even more, and every tutorial or forum assumes that the only thing you don’t know how to do in R is the very specific thing you asked about. Because of this, all the tutorials and the people answering your question immediately start talking about things you had never heard about or don’t know how to use. Trying to find accessible information about these things therefore feels like walking through an endless, dark labyrinth of ineffable anguish and desperation. It almost makes you want to quit science altogether and join your Aunt’s business making pizzas for dogs.

This tutorial will try to reduce all this pain and trauma by setting up the foundation of how to perform a simple, repeated-measures ANOVA in R. This procedure can then be added to or modified in order to add covariates, between-subjects effects, and a number of other things, but this tutorial does not cover them.

Tutorial

Each item corresponds to each line of code:

1. The first thing that must be done to analyze a data set is to load it into R `read.csv(“”)` and assign it `<-` to an object, which we’ll call “fish”.
2. To make sure that your data is in the right format, it helps to look at it’s structure with `str( )`. In this case we see that the factor “Size” lists three levels, going from Large down.
3. R plots axes with categorical variables in the order that they’re listed, which means that a plot made with our data will go from size Large down to Small, which looks odd. To fix this, we have to tell R to use a different order

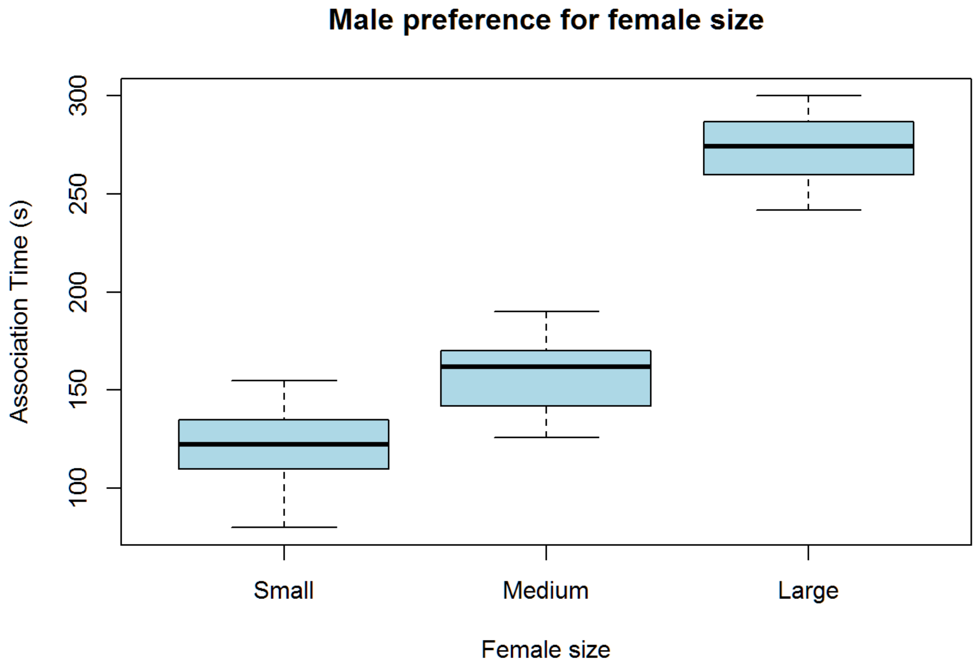
```
fish <- read.csv ("fish.csv")
str(fish)
```

```
## 'data.frame':    90 obs. of  3 variables:
##  $ ID  : int  1 1 1 2 2 2 3 3 3 4 ...
##  $ Size: Factor w/ 3 levels "Large","Medium",..: 3 2 1 3 2 1 3 2 1 3 ...
##  $ Time: int  110 153 285 120 173 290 120 172 252 129 ...
```

```
fish$Size  <- factor(fish$Size, levels=c("Small", "Medium", "Large")) #This basically just tells R that "Small", "Medium", and "Large" are the three levels of the column "Size", and to use them in that order. Without specifying this, R plots the x-axis from Large to Small in the plot below.
```

1. We can now take a look at a boxplot of the data (using `plot`. `boxplot` can also be used, but R will default to a boxplot with data like this).

```
plot(Time ~ Size, data=fish, xlab="Female size", ylab="Association Time (s)", col="light blue", main= "Male preference for female size") #The format is: plot(yaxis ~ xaxis, data= object, xlab="label for the x axis", ylab="label for the y axis", col="fill color of the boxes", main="title of the plot")
```



One of the assumptions of ANOVAs is that the the data is distributed normally. To statistically test this, we can use a Shapiro-Wilk test with `shapiro.test()`. Each test below specifies one size category to be tested.

1. This test requires subsetting the data, as you want to know whether the times under each of the size categories are normally distributed. The `fish$Size == “Small”` part tells R to only use the level “Small” (“Small” is a true condition) in the “Size” column of the object “fish”.

   The `fish$Time[]` part then tells R to look at the Time column in the fish data. Together, it means you are only performing the Shapiro-Wilk test on the times for the small females.
2. The test is then performed on the Medium females
3. And on the Large females

```
shapiro.test(fish$Time[fish$Size == "Small"]) #Shapiro-Wilk test.
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fish$Time[fish$Size == "Small"]
## W = 0.97448, p-value = 0.6673
```

```
shapiro.test(fish$Time[fish$Size == "Medium"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fish$Time[fish$Size == "Medium"]
## W = 0.97271, p-value = 0.6156
```

```
shapiro.test(fish$Time[fish$Size == "Large"])
```

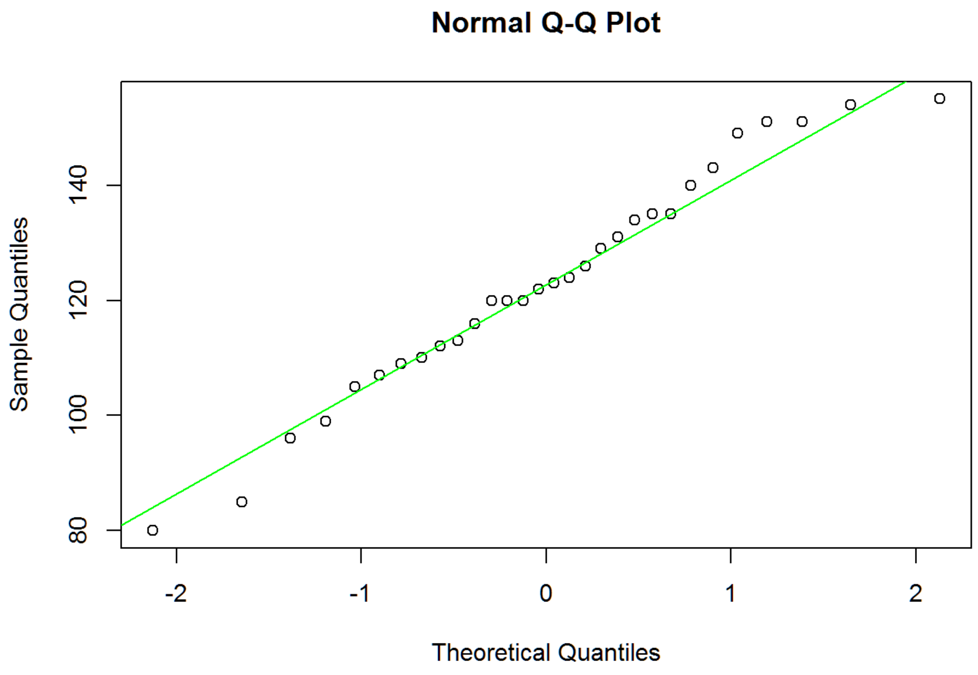
```
##
##  Shapiro-Wilk normality test
##
## data:  fish$Time[fish$Size == "Large"]
## W = 0.95554, p-value = 0.2373
```

Because all of the p-values are greater than .05, the null hypothesis that the data is normally distrubuted is not rejected. Thus, the normality assumption of the ANOVA will not be violated, increasing our confidence that the final results of the ANOVA we will perform will be valid. However, it is sometimes useful to get an intuitive feeling for how the data is distributed in a visual way.

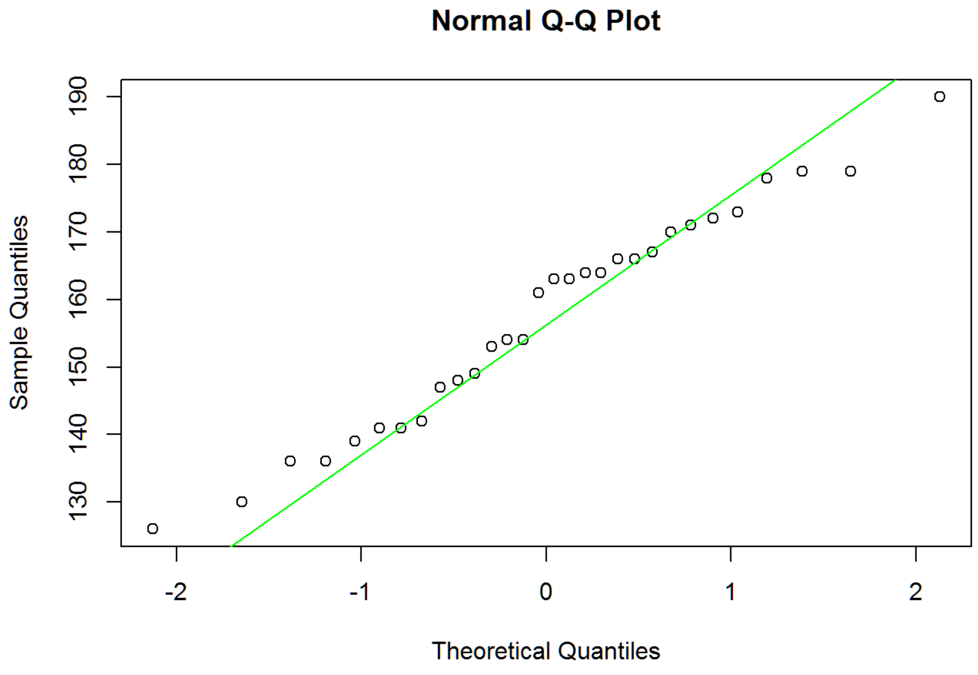
1. One can use `qqnorm` to create a quantile-quantile plot, using the same subsetted data as in the previous test.
2. `qqline` then adds a line that passes through the points which would be expected in a normal distribution

```
qqnorm(fish$Time[fish$Size == "Small"]) #creates quantile-quantile plot.
qqline(fish$Time[fish$Size == "Small"], col = "green") #adds a line through  theoretical normally-distributed
```

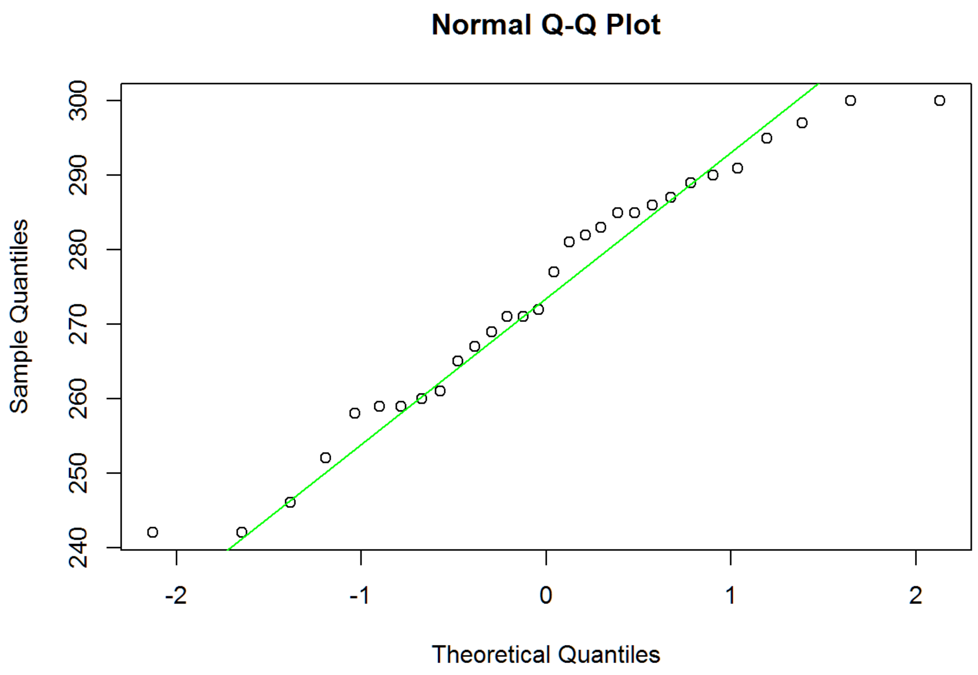
```
data. col allows you to set the color of the line.
```



```
qqnorm(fish$Time[fish$Size == "Medium"])
qqline(fish$Time[fish$Size == "Medium"], col = "green")
```



```
qqnorm(fish$Time[fish$Size == "Large"])
qqline(fish$Time[fish$Size == "Large"], col = "green")
```



Before running the ANOVA, however, it is necessary to set the contrasts defaults. The vast majority of ANOVAs should use the sum-to-zero convention for effect weights that is the generally accepted method and which nearly every statistical package uses. For some reason, however, R by default does not use this convention

```
options(contrasts=c("contr.sum","contr.poly"))
```

One of the assumptions of a repeated-measures ANOVA has to do with Mauchly's sphericity. Most methods of performing repeated-measures do not allow Mauchly's sphericity test to be performed, which is irritating because this means we have to use `lm`, which requires reorganizing the data

1. First, we will reorganize the data, and assign it `<-` to an object called fish2. It is important to create a new object in case we make a mistake or we need the data in the original form later on (and we will). `cbind` creates columns based on the sets of data that you give it, and then puts them into an object. Here, it is taking all of the times of the "Small" females in the "Size" column, and putting them on one column. Then it takes the times of the "Medium" females in the "Size" column and makes another column, and then does the same with the times for the Large females.
2. Because no names were specified, the new vector has no names. We can use `colnames()` to give them names. `c()` creates a sequence of elements, in this case of names. We'll need the names in a variable next, so we also assign that sequence to an object called "cols".

```
fish2 <- cbind(fish$Time[fish$Size=="Small"], fish$Time[fish$Size=="Medium"], fish$Time[fish$Size=="Large"])
colnames(fish2) <- c("Small","Medium","Large") -> cols
```

1. Now you need to create a model using `lm`, which is a function that fits linear models. This line just gives you the coefficients of the three size classes
2. The next steps require the car package, which must be installed.
3. R requires you to then load the package, which you can do with `library()`
4. R needs to be told "cols" is factor with `factor`
5. Now we finally perform the anova with `Anova` and assign it to an object we'll call "model.aov". In this case we're using a type III ANOVA. `idata` just defines the intra-subject levels (the repeated variable), and `idesign` is a formula that uses what was defined in `idata` that specifies the design
6. to get information about the anova, we must tell R to tell us what it just did with `summary`

```
model <- lm(fish2 ~ 1)
#install.packages("car")  #must install car package if not already installed.
library(car)  #even if installed, the package needs to be loaded into R.
```

```
## Warning: package 'car' was built under R version 3.3.3
```

```
factor(cols) #R needs to be told "cols" is a factor so that it can be used in the lm
```

```
## [1] Small  Medium Large
## Levels: Large Medium Small
```

```
model.aov <- Anova(model, idata = data.frame(cols), idesign = ~cols, type="III")
```

```
summary(model.aov, multivariate=FALSE)
```

```
## Warning in summary.Anova.mlm(model.aov, multivariate = FALSE): HF eps > 1
## treated as 1
```

```
## 
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
## 
##                  SS num Df Error SS den Df       F    Pr(>F)    
## (Intercept) 3075812      1    12105     29 7368.69 < 2.2e-16 ***
## cols         375662      2    15286     58  712.72 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## 
## Mauchly Tests for Sphericity
## 
##      Test statistic p-value
## cols        0.96707 0.62576
## 
## 
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
## 
##       GG eps Pr(>F[GG])    
## cols 0.96812  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##        HF eps   Pr(>F[HF])
## cols 1.036204 1.488171e-41
```

The Pr(>F) column tells you the P value. The three asterisks in this case tell you that the p-value is very significant. The Mauchly tests for sphericity have a p-value greater than .05, which means the null hypothesis cannot be rejected. This tells you that the data meets the sphericity assumption of repeated-measures ANOVAs. Since our data meets this assumption, we can ignore the “Greenhouse-Geisser and Huynh-Feldt Corrections for Departure from Sphericity”. We must now perform post-hoc tests to see which of the groups are different from each other. Because a lot of the time the sphericity assumption isn’t completely met, and because one cannot directly do a Tukey test at this point, it is often best to just perform a t-test.

1. This can be done with `t.test`

```
t.test(fish2[,"Medium"]-fish2[,"Small"])
```

```
## 
##  One Sample t-test
## 
## data:  fish2[, "Medium"] - fish2[, "Small"]
## t = 7.6065, df = 29, p-value = 2.19e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  25.05312 43.48022
## sample estimates:
## mean of x 
##  34.26667
```

```
t.test(fish2[,"Large"]-fish2[,"Medium"])
```

```
## 
##  One Sample t-test
## 
## data:  fish2[, "Large"] - fish2[, "Medium"]
## t = 30.354, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  108.8056 124.5277
## sample estimates:
## mean of x 
##  116.6667
```

Because both of the t-tests have a very small p-value (far smaller than our alpha level of 0.05), we can conclude that our male fish significantly prefer medium females over small females, and that they significantly prefer large females over medium females.