Lab #5, Part #1 – ANOVA and Speed Dating Participants

The data for this lab come from a survey of participants in a speed dating event, which was administered before participating in actual speed dating. It includes information on the gender, race, and age of the participant, as well as the self-reported importance of religion and race when seeking a date. There are also items in the survey measured on a scale of 1 to 10 (with 10 being the highest) asking "How interested are you in the following activities?" and "Overall, how happy do you expect to be with the people you meet during the speed-dating event?" There is a key at the end of this lab to inform you of all variable names and descriptions.

The data come from the following link, and I've reduced it and cleaned it up a good bit for this lab: https://www.kaggle.com/annavictoria/speed-dating-experiment. You can download the clean file on Cavnas called `speedDating.csv`. We'll use this data set exclusively for this lab on ANOVA.

Please be sure to include all code and all visuals and tables in your submission on Canvas. Note that all responses to questions should be in complete sentences and provide details and explanations when asked.

Activity #1: Open R-Studio and Load in Data.
Go ahead and download the files (`speedDating.csv`) from the Canvas Lab #5 page and save them to a new lab folder in your class folder directory (Lab 4 would be a good choice for your new folder) as .csv files. Then open up R-Studio, set your working directory with `setwd()`, and load the data. Name the data `spdate`.

```
#set working directory and load up the data.
setwd("c:/Users/…")
spdate <- read.csv(file = "speedDating.csv", h = T)
head(spdate)
```

Be sure that you've told R-Studio that the data has variable names (headers) with `h = T`, and confirm that they show up correctly.

Activity #2. Explore Your Data.
Using the `summary()` function, summarize all variables in your data in a neat table. Report the mean, standard deviation, minimum, and maximum for all numeric variables (we'll treat the scale from 1 to 10 as numeric for our purposes, rather than ordinal). For `gender` and `race`, report the proportion of participants in each category using `table()`.

2.1 Make sure to put these in a neatly organized table rather than just copying and pasting from R-Studio. Include this in your lab report and provide some descriptions of the data in complete sentences.

2.2 Now that you have some summaries, let's make some visuals. Using the `hist()` function, take a look at the histogram for `watchsport`, `gaming`, `dining`, and `reading`. Include these in your lab report and be sure to label axes appropriately, and give each a title. Feel free to use color if you'd like. In complete sentences, describe what the distribution of each looks like. Given what you see, do you think that the data come from a normally distributed population for each of these variables? Do you think this is a problem? Why or why not?

2.3 Repeat 2.2, but this time make separate histograms for Male and Female participants. Do this just for `watchsport`. You can do this any way you'd like, but I'd suggest using the `par(mforw = c(1,2))` function to do side-by-side Male/Female histograms (and you can use color to discern each group, too, if you'd like). Make sure that the x-axis and y-axis across histograms matches. Describe what you see.

2.4 Repeat 23, but this time make separate histograms by `race`. You can do this any way you'd like, but I'd suggest using the `par(mfrow = c(2,3))` function to make a matrix of 6 histograms. Make sure that the x-axis and y-axis across histograms matches. Describe what you see.

Activity #3. Do a t-test as review.

Let's start by doing a t-test similar to Lab #4-2. For this, we'll just do an independent samples t-test comparing differences in interest in watching sport (`watchsport`) by `gender`.

3.1 Start by making a table of the mean and standard deviations of `watchsport`, separated by gender. Remember that you can do this with the `aggregate()` function. Note that this should reflect what you see in 2.3 above. Include this in your lab report.

For now, we'll skip the null and alternative hypotheses setup from last week, but note that we want to test if there are differences between males and females with respect to their interest in watching sports. The null is "no difference" and the alternative is "there is a difference."

3.2 Use the `t.test()` function to do a two-tailed test of differences in `watchsport` across males and females. Treat this test as one where variance is unequal across groups (Welch's test). Report the test in your lab report, and use a couple sentences to describe your conclusion in the context of the data and the table you made in 3.1. We'll return to this in Activity #6.

Activity #4. Make New Groups.

Although we have continuous variables for Age and the Importance of Religion in our data, because this is an ANOVA lab, we're going to consolidate these variables and make new groups. Usually, we would not want to do this, since it means we lose information about more granular values, something we'll talk about with regression. But for our purposes, this will work fine.

4.1 Make a new variable that groups together age. We'll group into the following categories: Under 18-25, 26-30, and 31+. Use the code below, then report the number of observations in each group neatly in a table in your lab report.

```
#make a new variable called ageGroup
spdate$ageGroup <- ifelse(spdate$age < 26, "Young", NA)
spdate$ageGroup <- ifelse(spdate$age > 25 & spdate$age < 31, "Mid",
     spdate$ageGroup)
spdate$ageGroup <- ifelse(spdate$age > 30, "Old", spdate$ageGroup)
```

4.2 Make a new variable that groups together Importance of Religion. Group together as low (1-3), mid (4-6), and high (7-10). Report the number of observations in each group neatly in a table in your lab report.

Activity #5. One-Way ANOVA Practice.

For this activity, I'll walk through one example of ANOVA with a new function, `aov()`, and then ask you to do two on your own. Let's start with comparing differences in `watchsport` across the `race` variable. First, set up our null and alternative hypotheses:

H_0: $\mu_{Asian/Pacific} = \mu_{Black} = \mu_{White} = \mu_{Hispanic} = \mu_{Other}$. Interest levels in watching sport across speed dating participants is equal.

H_A: Interest levels in watching sport across speed dating participants is not the same.

Now we know we're directly testing the average response about watching sports across race groups. We can make use of a new function as follows:

```
#test interest in watching sports across race with an ANOVA
test1 <- aov(watchsport ~ race, data = spdate)
```

Notice that this function using the ~, which we used in the `aggregate() function`. Essentially, this is telling R-Studio that we want to know the differences in means of the variable on the left hand side of the ~, as a function of the group the participants are in on the right hand side of the ~.

We give the result a name, because we actually want to use another function on the result. Aov() essentially fits a model to your data, but we'll need to use summary() to tell R-Studio to give us the test statistic and p-value. You can use the code below for this:

```
#report F-statistic and p-value
summary(test1)
```

You should notice that this table looks like the one we saw in the lecture slides.

5.1 Report the results of the ANOVA in your lab report. For this, report the F-statistic (with the degrees of freedom) and p-value. In a couple sentences, note whether you reject or do not reject the null hypothesis, and state your conclusions in the context of the data.

5.2 Now use the `reading` variable and test across `ageGroup`. In your lab report, be sure to include the null and alternative hypotheses in words. Report your F-statistic, degrees of freedom, and p-value, and come to a reject/do not reject decision. Explain what this means in words and in the context of the data.

5.3 For this problem, use the `gaming` variable and test across `ageGroup`. In your lab report, be sure to include the null and alternative hypotheses in words. Report your F-statistic, degrees of freedom, and p-value, and come to a reject/do not reject decision. Explain what this means in words and in the context of the data.

5.4 For this problem, use the `dining` variable and test across `religGroup`. In your lab report, be sure to include the null and alternative hypotheses in words. Report your F-statistic, degrees of freedom, and p-value, and come to a reject/do not reject decision. Explain what this means in words and in the context of the data.

Activity #6. Post-Hoc Tests.

For this activity, we will introduce a few new things. We want to be able to run the correct post-hoc test with the right standard error used for the paired comparisons. We'll use a variant of post-hoc test called Tukey's HSD, which is slightly different from what we did in the lectures. The general idea is the same, and it's easy to implement Tukey in R-Studio.

For this, we'll introduce installing new packages (free extensions to R-Studio built by programmers), loading them into your R-Studio instance, and using a new function for multiple comparisons after an ANOVA.

Let's start with installing a new package. We'll be using a package called `multcompView` and a new function called `install.packages()`. Try the following code:

```
#load new package
install.packages("multcompView")
```

When you do this, you'll most likely have a window pop up asking you to choose a CRAN Mirror. All this means is that there are repositories for these packages in various places online for R. Most of these will work fine. I learned R when I lived in Michigan, so I always choose the Michigan one. Go ahead and choose Michigan and press OK.

From there, it should start installing. Once it's finished up, you're not done yet. You actually need to load it directly into your current R-Studio instance with the `library()` function. This function

just calls the package from a folder R-Studio created on your computer that holds various packages you use. From there, you can use the function inside the `multcompView` library.

```
#load the library for use in R
library(multcompView)
```

Notice that to install the package, you needed to include quotes around its name. But, when you load the library after it's installed, just the name is required inside the parentheses.

Now you're ready to do your pairwise comparisons. We'll just do this for one of our examples from Activity #5: comparing `gaming` by `ageGroup`. Note that I have named my ANOVA `test3`, which I'll use in my code below. Be sure you use the name you gave this in your own R-Studio instance.

The function we're interested in from this package is called `TukeyHSD()`. This will calculate the difference in each group pair as well as provide an adjusted p-value. All we need to do is use it

```
#do post-hoc comparisons on the test3 ANOVA
TukeyHSD(test3)
```

Notice that it also spits out some confidence intervals. While we didn't talk about confidence intervals much in the context of t-tests and post-hoc tests, they follow similar logic as when we calculated them for other scenarios. In a more advanced class, we would use this information to make plots of the confidence intervals with this package. But we'll leave that for another day.

6.1 Report the pairwise group differences and p-values in a neat table in your lab report. Using an alpha level of 0.05 – and without any additional necessary corrections to this alpha level – describe the results of the post-hoc pairwise analysis. What does this tell you about the results you found in 5.3? Does this confirm what you found there from a qualitative standpoint? State in words and in the context of the data in complete sentences.

Activity #7. Repeat Activity #3 with `aov()`.

Note that in Activity #3, we only had 2 groups, which necessitates the t-test. But, as it turns out, the t-test is just a special case of an ANOVA! In fact, it turns out that the F-statistic is just our t-statistic squared. We won't derive that here, but you'll confirm it below.

7.1 We can show this in R by using the `aov()` function with a two-group comparison. Go a head and use the function on the `watchsport` variable, using `gender` as the grouping variable. Report your F-statistic, degrees of freedom, and p-value, and come to a reject/do not reject decision.

7.2 If you square your t-statistic from Activity #3, is it approximately equal to what R-Studio reports as the F-statistic? Relate this back to our discussion in the lectures relating to the F-statistic always being a positive value.

Data Variable Key

`gender`: self-reported binary gender (Male or Female)
`age`: age in years
`race`: 6-category self-reported race

`imprace`: importance of person you date be of same race/ethnicity (1-10)
`imprelig`: importance of person you date be of the same religious background(1-10)

How interested are you in the following activities, on a scale of 1-10?
`playsport`: playing sports/athletics
`watchsport`: watching sports
`exercise`: body building/exercising
`dining`: dining out
`museums`: museums/galleries
`art`: making or viewing art
`hiking`: hiking/camping
`gaming`: gaming
`clubbing`: dancing/clubbing
`reading`: reading
`television`: watching tv
`theater`: going to the theater (live)
`movies`: going to movies
`concerts`: going to concerts
`music`: listening to music
`shopping`: shopping
`yoga`: doing yoga/meditation

`exphappy`: expectation about being happy with speed dating (1-10)