# Summarizing and Plotting Psych 315 Survey Data in R

January 25 2019

This tutorial shows you how to load the survey data and how to calculate summary statistics and visualize different kinds of variables.

A script containing these R commands can be found at:

http://www.courses.washington.edu/psy315/R/SurveyAnalysis.R

First we'll clear the workspace and load in the survey data:

```
rm(list = ls())
survey <-read.csv("http://www.courses.washington.edu/psy315/datasets/Psych315W21survey.csv")
```

Our new variable 'survey' has a bunch of fields associated with it that correspond to each of the questions. To see the actual questions associated with each field and the list of options for responses you can open the file:

http://www.courses.washington.edu/psy315/excel_files/SurveyQuestions.xls

Survey data comes in two levels of measurement, nominal and ratio. In this tutorial we'll go through how to summarize and visualize any of the nominal scale questions, any of the ratio scale questions, and how to compare any two (nominal vs. nominal, ratio vs. ratio, and nominal vs. ratio).

If you're using R Studio' a good way to see the list of fields is to go to the 'Data' window, find the 'survey' variable and click on the blue triangle. You'll see things like:

gender : Factor w /2 levels "Female", "Male": 1 2 2 ...

This means that there is a field 'gender' which you can access with the dollar sign (survey$gender)

'Factor' means that this field is nominal data, and you can see that the 2 levels are 'Female' and 'Male'.

Other fields are either 'int' (integers) or 'num' (decimals), which are both ratio scale data for our survey.

## Visualizing the distribution of ratio scale data

Ratio scale data is best visualized with a histogram with class intervals that you can set.

For example, I asked you about your heights (in inches). This can be found in the field 'height' and can be accesed with the dollar sign:

```
survey$height
```

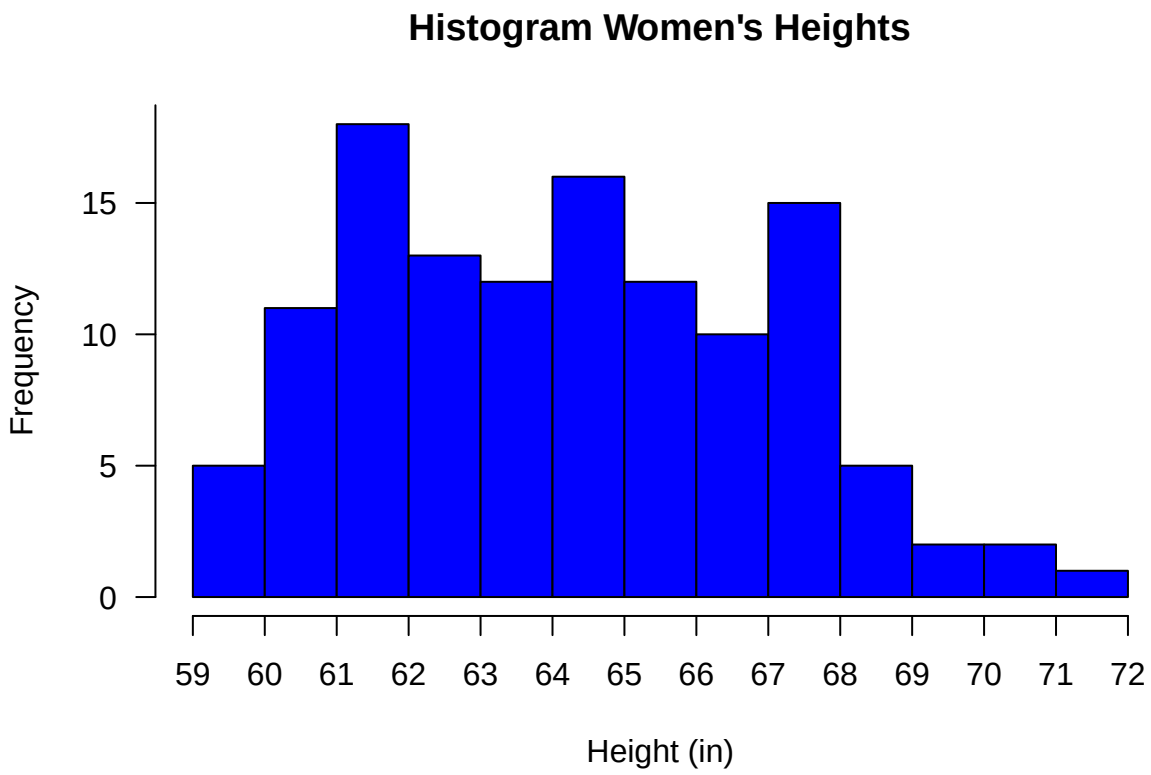
```
##   [1] 68 64 63 60 66 70 62 66 61 65 62 72 61 64 67 69 72 71 62 66 62 65 67 60 62
##  [26] 65 62 72 61 65 70 68 67 68 64 66 61 66 67 63 62 64 61 63 68 62 69 72 68 61
##  [51] 67 65 69 62 63 72 68 63 63 68 74 72 61 68 70 70 62 68 61 62 59 63 62 64 71
##  [76] 62 65 62 70 63 67 66 72 66 66 60 65 62 68 75 65 70 65 68 62 62 72 66 74 71
## [101] 68 68 67 64 67 63 65 66 66 65 66 60 67 64 64 62 61 68 66 64 64 67 67 62 65
## [126] 68 63 64 65 74 71 61 68 68 68 63 72 65 64 61 63 66 65 66 69 69 66 63 73 61
## [151] 65 74
```

We can look at the heights for female students like this:

```
ratio.data <- survey$height[survey$gender == "Female"]
head(ratio.data)
```

```
## [1] 64 63 60 66 62 66
```

```
# define class intervals based in the min and max:
class.interval <- seq(min(ratio.data),
                      max(ratio.data),
                      1)
hist(ratio.data,
     main="Histogram Women's Heights",
     xlab="Height (in)",
     col="blue",
     xaxt='n',
     yaxt = 'n',
     breaks =class.interval
)

# and then adding your own axes with the 'axis' function
# Axis 1 is 'x' and 2 is 'y':
axis(1, at=class.interval)
axis(2, at=seq(0,100,5),las = 1)
```




We can summarize a ratio-scale value with means and standard deviations.

```
mean(ratio.data)
```

```
## [1] 64.70492
```

```
sd(ratio.data)
```

```
## [1] 2.824592
```

Wikipedia says that the average US woman is 64 inches tall. Later on in the quarter we'll use a 't-test' to determine the probability of drawing our mean from the survey by chance, if the true mean is 64 inches.

## Visualizing the distribution of nominal scale data

nominal scale data can be visualized with a histogram too, but the x-axis categories are names, not numbers. R has a function 'table' that counts up the frequencies for nominal data. For example, for Gender

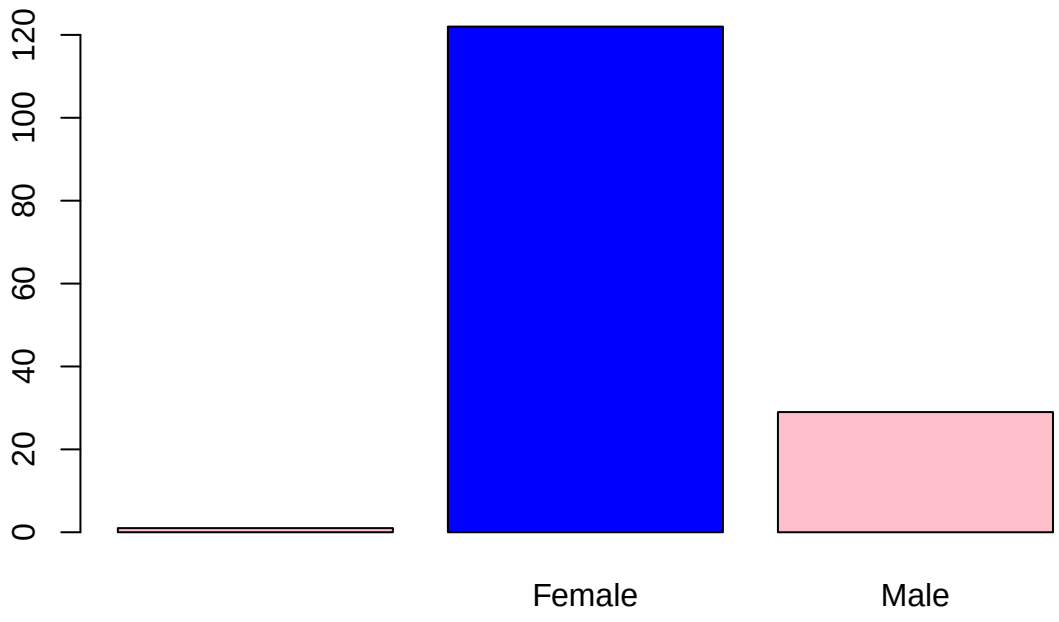
```
nominal.data <- survey$gender

freqs <- table(nominal.data)
freqs
```

```
## nominal.data
##        Female   Male
##      1    122     29
```

gender.freqs is a special list of numbers where the columns have names. In our case, the names the genders.

You can visualize the frequencies for nominal data with 'barplot'. Here we'll color the bars using the 'col' option by their gender stereotypical colors:

```
barplot(freqs,
        col = c("pink","blue"))
```




Is there an equal ratio of men to women in this class? Later on we'll run a 'Chi-squared' test for frequency to determine the probability of getting a distribution like this by chance.

Now that we've seen how to visualize the frequency distribution for ratio scale and nominal scale data, we'll

move on to visualizing how to compare different variables. With these two types of data, there are three kinds of comparisons we can make: ratio to ratio, ratio to nominal, and nominal to nominal.

## Comparing nominal to nominal scale data

Comparing nominal data to nominal data is typically asking if the distribution of frequencies for one variable depends on the level of another. For example, does where you like to sit in the class depend on gender? Both 'gender' and 'sit' are nominal data variables.

R's 'table' function conveniently tabulates frequencies for more than one nominal variable:

```
myTable <- table(survey$gender,survey$sit)
```

The result is a table with both rows and columns, with labels:

```
myTable
```

```
##
##          In the middle Near the front Toward the back
##                      1              0               0
##   Female            56             44              22
##   Male              16              3              10
```

The labels can be pulled out using 'row.names' and 'colnames' (note the inconsistency using '.' in the function names)

```
row.names(myTable)
```

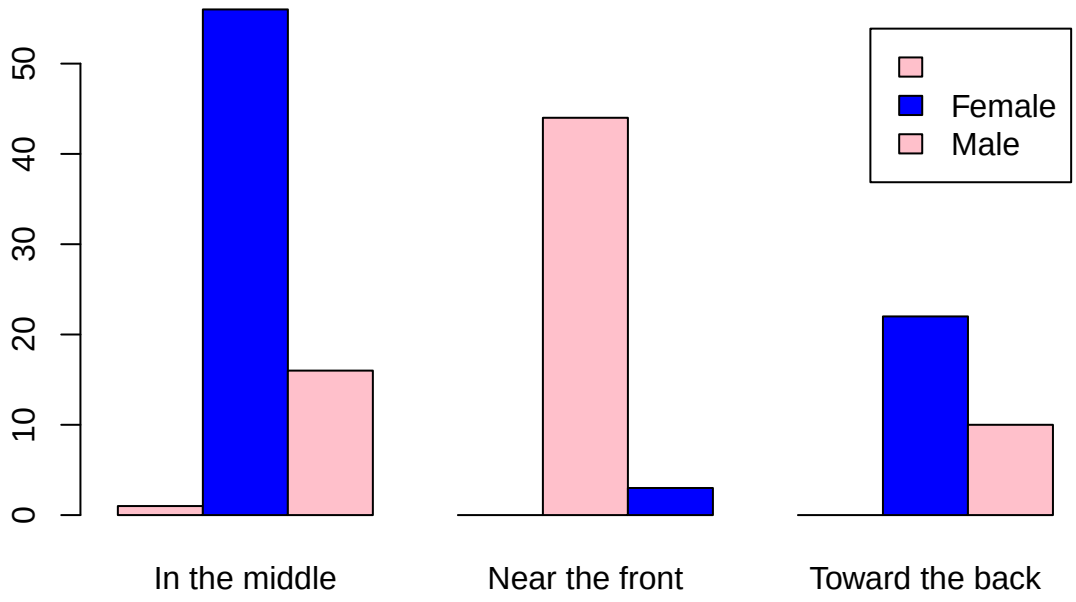
```
## [1] ""       "Female" "Male"
```

```
colnames(myTable)
```

```
## [1] "In the middle"   "Near the front"  "Toward the back"
```

You may or may not see a dependency on where you sit with gender. To visualize these frequencies, use 'barplot' again.

```
barplot(myTable,
        beside=TRUE,
        legend = row.names(myTable),
        col = c("Pink","Blue"))
```




I prefer 'beside=TRUE' over the default which stacks the bars on top of each other (try it)

Can you see a difference in the frequency of where you sit for the two genders? Later on in the quarter we'll run a 'Chi-squared test for independence' which will determine the probability of getting results like this by chance.

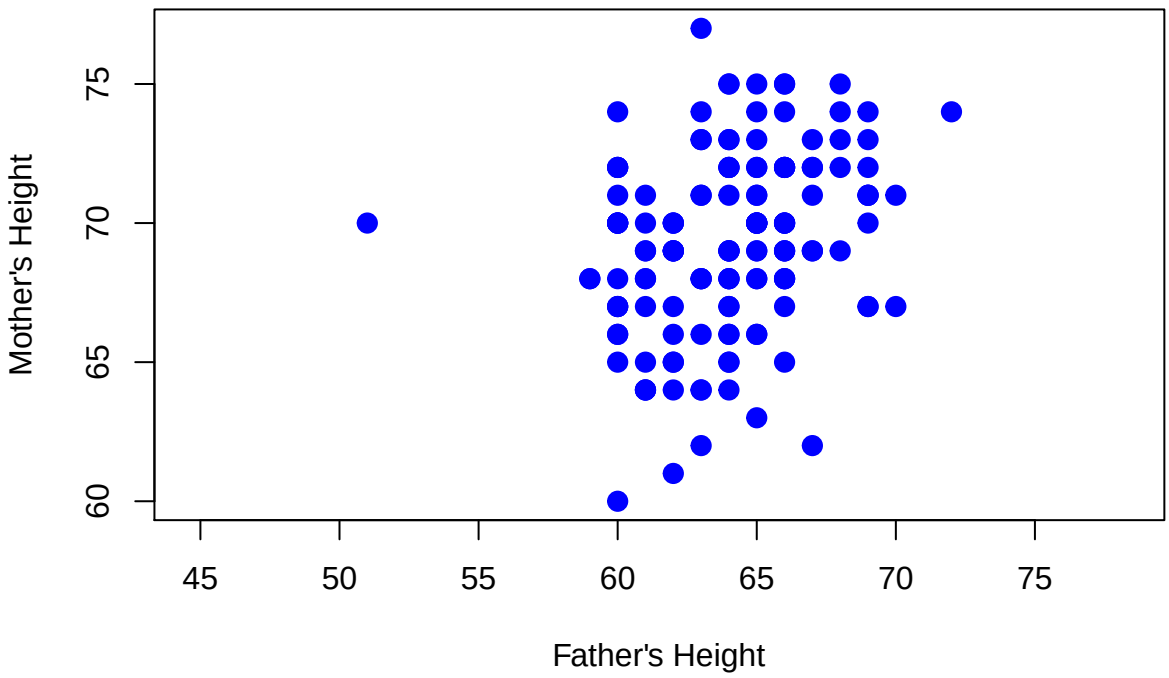
### Comparing ratio to ratio scale data

Comparing ratio scale data to ratio scale data is best done with a scatterplot. For example, to compare your father's heights to your mother's heights, use:

```
ratio.data.x <- survey$mheight
ratio.data.y <- survey$pheight
```

Then run the rest of the code:

```
plot(ratio.data.x,ratio.data.y,
     xlab = "Father's Height",
     ylab = "Mother's Height",
     pch = 20,
     col = "blue",
     as = 1,
     cex = 2)
```




We quantify this relation with the Pearson Correlation

```
cor(ratio.data.x,ratio.data.y, use = "complete.obs")
```

```
## [1] 0.283387
```

Later on we'll find the probability of obtaining a correlation this large by chance using a 'correlation test for r=0'.

## Comparing ratio to nominal scale data

This last comparison involves calculating the mean value for a ratio scale variable for each level of an nominal scale variable.

For example, does the average grade you expect in exam 1 depend on where you sit in the class? This requires pulling out the values in the field 'Exam1' for each of the levels in 'sit'. To find the means for each of the three levels of 'sit' we could do something like this:

```
mean(survey$Exam1[survey$sit=="Toward the back"])
```

```
## [1] 85.3125
```

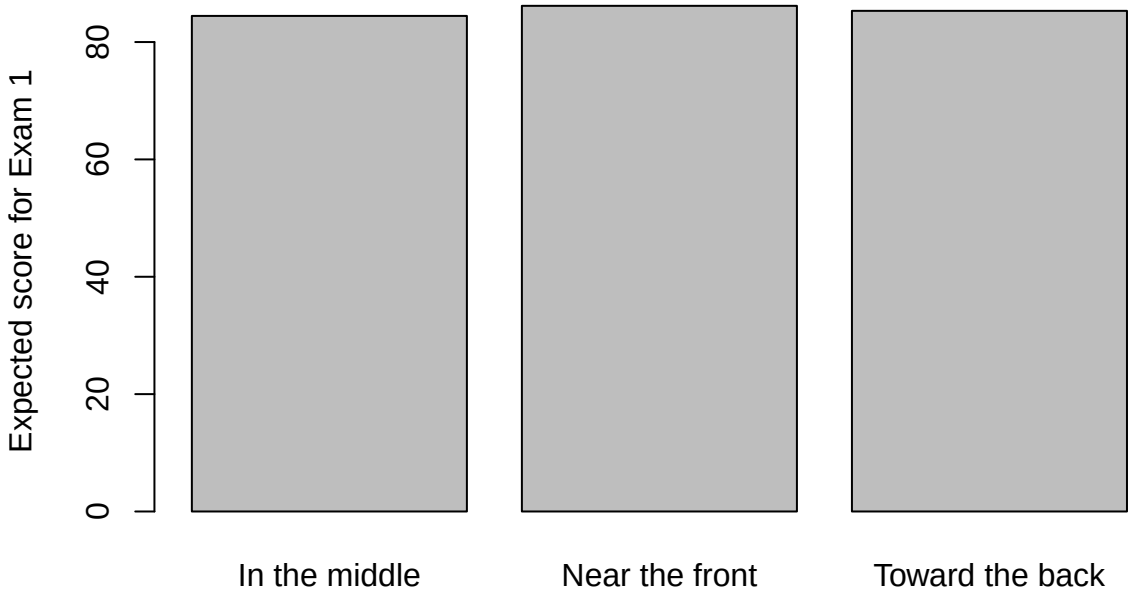
And do it two more times for the other two levels of 'sit'. Fortunatlely, R has a function 'tapply' that calculates a summary statistic for each level of a nominal variable automatically:

```
means <- tapply(survey$Exam1,survey$sit,mean)
means
```

```
##   In the middle  Near the front Toward the back
##        84.45205        86.17021        85.31250
```

We can use 'barplot' to plot these means:

```
barplot(means, ylab = "Expected score for Exam 1")
```




Does it look like the exected Exam 1 score varies across where you like to sit?

Later on in the quarter we'll run an 'ANOVA', or analysis of variance, to determine the probability of getting means this far apart from each other by chance.

## Summary

This tutorial has shown yo 5 ways of summarizing and plotting survey data:

1) histograms for ratio scale data
2) bar graphs for nominal scale data
3) bar graphs of nominal vs. nominal scale data
4) scatterplots of ratio vs. ratio scale data
5) bar graphs of means of ratio scale vs. nominal scale data

Each of these 5 plots has a corresponding statistical test that we'll be covering later in the quarter.