STAT 140: Lab 2

Exploring the data set

```
## Print the first few rows of Mroz
head(Mroz)
```

```
##   lfp k5 k618 age  wc hc       lwg    inc
## 1 yes  1    0  32  no no 1.2101647 10.910
## 2 yes  0    2  30  no no 0.3285041 19.500
## 3 yes  1    3  35  no no 1.5141279 12.040
## 4 yes  0    3  34  no no 0.0921151  6.800
## 5 yes  1    2  31 yes no 1.5242802 20.100
## 6 yes  0    0  54  no no 1.5564855  9.859
```

```
## Uncomment the line below to use the help functionality to learn about the
## variables Note, this will only work for datasets built in to packages
## ?Mroz

## Apply str() to the data frame to examine it
str(Mroz)
```

```
## 'data.frame':    753 obs. of  8 variables:
##  $ lfp : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ k5  : int  1 0 1 0 1 0 0 0 0 0 ...
##  $ k618: int  0 2 3 3 2 0 2 0 2 2 ...
##  $ age : int  32 30 35 34 31 54 37 54 48 39 ...
##  $ wc  : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 2 1 1 1 ...
##  $ hc  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lwg : num  1.2102 0.3285 1.5141 0.0921 1.5243 ...
##  $ inc : num  10.9 19.5 12 6.8 20.1 ...
```

Summary statistics

In this section, you will learn functions for finding the sample statistics that we talked about in lecture.

In the chunk below, apply the summary function (from Lab 1) to the Mroz data set. Comment on the output for each variable. Are there any variables that are being summarized as numerical variables, but should really be treated as categorical (factors, in R language)?

Answer:

```
## apply summary function
```

Measures of center: sample statistics

Perhaps the most familiar measure of center is the sample mean, or average. The formula for this is $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}$. This is the more appropriate measure of center in cases where the distribution of the variable is symmetric.

A faster way to do this is to use the function mean(), which takes a numerical variable as an argument and returns the average. Remember that you will have to extract a single column from the dataframe as we did in Lab 1. Use this function to find the sample mean income (inc) in the Mroz dataset. Store the value for the mean as an object called `mean_inc`. On the next line, type `mean_inc` to print the mean.

```
## apply mean function to find the mean inc
```

The median is another measure of center. It is an example of a *robust statistic*, meaning that it is not influenced by unusually large or unusually small observations of a variable. It is a more appropriate measure of center when the distribution of the variable is skewed (either left-skewed or right-skewed).

We can calculate the median of a variable using the function median(), which takes a numerical variable as an argument and returns the average. Find the sample median income (inc) in the Mroz dataset. Store the value for the median as an object called `median_inc`. On the next line, type `median_inc` to print the median.

```
## apply median function to find the median inc
```

Is the median larger or smaller than the mean? What does this mean about the shape of the distribution?

Answer:

Measures of spread: sample statistics

Variance is a common measure of spread (how spread out the observations are from the mean). The sample variance is calculated by taking the sum of the squared deviances and dividing by $n-1$ (rather than n) for reasons that go beyond the scope of this class: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

In R, we can use the var() function to find the variance of a numerical variable. Find the sample variance of income. Store the value for the variance as an object called `var_inc`. On the next line, type `var_inc` to print the variance.

```
## apply var function to find the variance of inc
```

The standard deviation, s, is the positive square root of the variance. This is a better measure of spread to report than the variance because it is on the same scale (with the same units) as the original data (so in this case, dollars). We can find the standard deviation of income in two ways. Like the mean, the standard deviation (and variance) are more suitable measures of spread when the distribution of the variable is symmetric.

First, apply the sqrt() function (this takes the square root) to var_inc. Store this value as an object called `sd_inc_1`. On the next line, type `sd_inc_1` to print the value. Second, apply the sd() function to the income variable, and store this value as `sd_inc_2`. On the next line, type `sd_inc_2` to print the value. Verify that these two values are the same.

```
## apply the sqrt function to var_inc to find the standard deviation of inc

## apply sd function to find the standard deviation of inc
```

The interquartile range (IQR) is another measure of spread, which is obtained by taking the difference between the third quartile (Q3, also known as the 75th percentile) and the first quartile (Q1, also known as the 25th percentile): $Q_3 - Q_1$. Like the median, IQR is an example of a *robust statistic*. It is a more appropriate measure of spread when the distribution of the variable of interest (here income) is skewed (not symmetric).

Apply the iqr() function to the income variable to find the IQR. Store this value as an object called `iqr_inc`. On the next line, type `iqr_inc` to print the value.

```
## apply iqr function
```

Based on your comparison of the mean and median for income, is it better to report the IQR or the standard deviation as a measure of spread for income?

Answer:

Plots

The basic anatomy of a plot made in ggplot is as follows:

```
## Skeleton code - should not run anything
(ggplot(data=<name of data frame>,
        aes(x=<variable for x axis>, y=<variable for y axis>,
            color = <variable for color lines>,
            fill= <variable for color area>))
        + geom_<geometry type>()
        + <optional other things like axis labels, ...>)
```

Scatterplots

We saw example code to make a scatterplot using the ggplot() function in Lab 1. Using that example, plot inc (the y variable) versus age (the x variable).

```
## Scatterplot of inc vs age.
```

Histograms

Make a frequency histogram of income. You do not need a y-variable for this, and the geometry type is histogram.

```
## Histogram of income
```

Boxplots

Make a boxplot for income. You do not need an x-variable for this, and the geometry type is boxplot.

```
## Boxplot of income
```

Bar charts

Bar charts differ from histograms in that they are used to summarize categorical variables (or a discrete numerical variable with a small range of values), while histograms are used for numerical variables. The order of the bars in a bar chart does not matter (but it does in a histogram). Also, the bars do not touch in a bar chart, but they do in a histogram.

Make a bar chart for income and wc. The geometry type is bar, and you need both an x and y variable.

```
## Bar chart of wc
```

Challenge: side-by-side boxplots

If you have completed the rest of the lab, try to to make a side-by-side boxplot, which allows us to compare the medians, etc., of a numerical variable between groups. Use the numerical variable, inc., and the factor variable wc. You can look at the ggplot2 example here: http://homepages.gac.edu/~anienow2/MCS_142/R/R-boxplot2.html and try to modify it.

```
## Side-by-side boxplot
```