LIR 832: Computer Problem set #1

How to do computer problem sets:

1. Problem sets can be done and turned in as a group of two.

2. If you are not able to make progress on a step in a problem set in 15 minutes and are becoming frustrated, take a break. You aren’t going to solve it by hitting your head against the problem time and time again. Take a break, let your mind stew over the problem, talk with someone else about it.

3. The content of the homework is important and should be readily accessible to the grader. Show your work (typically this will be the log file from MINITAB), answer the questions and compose any graphs carefully. For example, the proper labeling and structuring of graphs is important, graphs should convey information in a compact and readily accessible form.

3. Although neatness is encouraged, this class is about substance more than presentation. It is important to show your work so that I can review any errors and make suggestions on how to correct them. However, neatly laying out the graph on a page and making any text flow around it, a la corporate reports, is not important to this class.
Computer Problem Set #1:

Learning to Use Minitab

Use the 20 percent subsample for managers and professionals for this assignment. This can be downloaded from the course datasets and links web page.

We will be working with three variables from the managers and professionals data set: uhour1 (usual hours on the main job), gender and race. You will want to download and print out the Unicon Codebook. This codebook has detailed information on all of the variables in the data set. You will need this to understand and answer the questions.

1. Before we can begin working with data, we need to learn about the variables and understand how we can use them. Review the codebook for uhour1, gender and race and answer the following questions.

   a. What values can gender take on? Is gender a qualitative or quantitative variable? What values does race take on? Is it qualitative or quantitative? Explain your answer.

   b. Although both of these variables are numeric, it may be more useful to turn them into text (name) variables. The recode function is under Data.

   \[ \text{Data} \rightarrow \text{Code} \rightarrow \text{numeric to text} \]

   (Note that you can either give the “new” column a name or specify a column location. If you do the first, MINITAB will toss the new column somewhere out to the right. If you do the second, you can place the new column closer to your data, but you then need to go into the worksheet and put a name in)

Recode each variable to a text variable with the following categories:

   Gender: (call this Sex)
   - Man
   - Woman

   Race (call this Racetxt)
   - White
   - African-American
   - Native American
   - Asian

   You may be creative about the actual text assigned to each outcome. Which numeric values are associated with each outcome?
c. Now consider uhour1. What values can uhour take on? Is it qualitative or quantitative? If we took an average of this variable as it is currently constructed, would the average be meaningful? Explain.

Use the graphing capabilities of MINITAB to make a histogram of uhour1 and indicate where the difficulty is in this data. (Graph » histogram)

To obtain a meaningful variable, we need to recode the data so that the variable hour designation is not averaged into the data. Let's create a new variable, call this UHOUR1rc (for UHOUR1 recode)

\[ Data \rightarrow Code \rightarrow numeric to numeric \]

Make the new variable name uhour1rc

Code the -4 value as *
Leave everything else blank

Taking a look at the data file, what happens to the -4 value in the new variable? What happens to all of the other values? Why is this useful?

2. Calculate descriptive statistics and a summary graph for Uhour1rc.

\[ Stat \rightarrow Basic Statistics \rightarrow Display Descriptive Statistics \rightarrow Uhour1rc \]
\[ Stat \rightarrow Basic Statistics \rightarrow Graphical Summary \rightarrow Uhour1rc \]

What have you learned about usual hours of work from the descriptive statistics and the graphical summary? Now create a graphical summary of uhour1. What are the differences between uhour1rc and hour1? What are the strengths and weaknesses of the recoded uhour1 variable?

3. Can you calculate descriptive statistics for Sex or for Racetxt (the text versions of the Gender and Hispanic variables)? What about for Gender and Race? Are the results of the Race and Gender variables useful? Explain.
4. There are two approaches to obtaining useful information about qualitative variables. First, although we cannot calculate descriptive statistics on qualitative data, we can get frequency distributions on the outcomes. We do this with the Table command in MINITAB.

   a. To use the table command to obtain frequency distributions on Sex

      Statistics ▶ Tables ▶ Tally Individual Variables (put Sex into the variable box). Try both the counts and total percent commands. Then repeat this with the recoded race variable. What is the distribution of managers and professionals by race.

   b. Another approach is to working with qualitative data is to recode the variables into indicator variables and compute descriptive statistics. Indicator variables show whether a particular condition exists or doesn’t exist. To do this with the Gender variable, use the recode command.

      Data ▶ Code ▶ numeric to numeric (remember you need to designate a new variable)

      Recode the value 1 to 0 (man)
      Recode the value 2 to 1 (woman)

      Call this new variable female. Now run descriptive statistics on this new variable. What is the interpretation of the mean? How does the mean compare to the statistics you generated with the Table command?

5. Now calculate descriptive statistics for uhour1rc and split it by gender. To do this, after you designate uhour1rc as your variable, make the by variable Sex. Do these results suggest a difference in typical hours of work between men and women? Explain. Split the sample by race using the same approach. Is there evidence of a difference in typical hours of work by race?
6. Let's move ahead and test several hypotheses about usual hours of work.

   a. Using the full sample, test the hypothesis that, the workforce works, on average, 40 hours per week? To do this,

   Statistics  ► basic statistics  ► one sample t

   Set the test mean to 40.

   You also need to set whether this is a one or two tailed test. You can set this under “option”. You set the greater than, not equal to, less than so that it is consistent with the alternative hypotheses. In this case, our null is that, on average, managers and professionals work 40 hour per week; the alternative is that they do not work 40 hours. MINITAB has you designate the alternative. As we are doing a two tailed test, the correct setting for the alternative is not equal.

   Can we reject the null of 40 hours at the 10% level? At the 5% level? At the 1% level? What is your conclusion?

   b. Repeat (a) but now test the alternative that managers and professionals are working less than 40 hours per week. Again, you will have to set the type of test under the “options” button and again, the sign you set in MINITAB is consistent with the alternative hypotheses. Can we reject the null at the 10% level? At the 5% level? At the 1% level? What is your conclusion?

   c. MINITAB will also test hypotheses for differences in the means of two samples. Let's test two hypotheses about men and women’s working hours.

      First, let's test whether, on average, men and women managers and professionals work the same number of hours? What we really believe is that they don't work the same hours. So our alternative would be:

      \[ H_{\text{alternative}}: \mu_{\text{men}} \neq \mu_{\text{women}} \]

      while

      \[ H_{\text{null}}: \mu_{\text{men}} = \mu_{\text{women}} \]
To set up the two sample test,

Statistics \rightarrow \text{basic statistics} \rightarrow \text{two sample t}

\text{under subscript, use Sex}
\text{Under option, make the alternative “not equal to”}

What is the P score for the two sample t? Can reject the null of equality in a 10%, 5% or 1% test?

d. Now test the null whether men’s hours of work are greater than women’s. Now our alternative and null are:

\begin{align*}
H_{\text{alternative}}: \mu_{\text{men}} &> \mu_{\text{women}} \\
\text{while} \\
H_{\text{null}}: \mu_{\text{men}} &\leq \mu_{\text{women}}
\end{align*}

or, using our preferred approach

\begin{align*}
H_{\text{alternative}}: \mu_{\text{men}} - \mu_{\text{women}} &> 0 \\
H_{\text{null}}: \mu_{\text{men}} - \mu_{\text{women}} &\leq 0
\end{align*}

Note: I like the preferred approach because we have set it up in terms of the relationship of the difference in the means to zero. This gives us the direction of the test to use in MINITAB.

Again, go into the menu:

Statistics \rightarrow \text{basic statistics} \rightarrow \text{two sample t}

\text{under subscript, use the Sex variable}
\text{Under option, make the alternative greater than}

Run the test. What is the p-value for the null? Can we reject this null in a 10, 5 or 1% test? What do you conclude about men’s and women managers hours of work?

A Warning: MINITAB loves alphabetical ordering. So, if you set up the gender labels as male and female, rather than man and woman, it will automatically reverse your hypothesis test. You need to be aware of this as an issue in using MINITAB.
7. A bit of graphing

a. Returning to descriptive statistics and using the graphical summary, create a histogram of Uhour1rc with a normal plot overlay. What are the strengths and weaknesses of this graph?

b. Now go to histograms under graphs and create a histogram of uhour1rc (it’s the same as the last graph).

   graph ➔ histogram ➔ with fit

   Its not much better so let create a superior graph. What is nice about MINITAB14 is that you can edit the graph directly.

1. Print out the graph you have produced (copy it into a report or Word with cut and paste)

2. The scale at the bottom isn’t very intuitive. Let’s alter the ticks at the bottom to be more useful. You need to ‘click’ on the X (horizontal) axis until you get a box of commands. You will have a command ‘edit X scale’. Click on this. Under the tab marked ‘Scale’ click the custom scale box (it defaults to automatic) and put ticks in at 0 10 20 30 40 50 60 70 80 90 100. Click ‘OK’ and watch the graph change. Why are these ticks more useful than the automatic ticks? (By the way, add this graph to your report)

3. This is better but it remains difficult to work with. It is just too hard to have as many bars as we do. We need to consolidate the data. We can do this by resetting the width of the bars, the “binning”. Again click on the X axis and bring up the command box. Go to binning tab and (1) reset to cutpoint intervals. This means that the vertical lines indicate the border between categories. Now go to interval definition and set midpoint/cutpoint positions. This will be most interesting if we set them as

   graph ➔ histogram ➔ options ➔ midpoint/cutpoint positions

   Set these at 0 10 20 30 35 40 45 50 60 70 80 90 100

   Now create the histogram. Why is this histogram more useful than the previous graph?

4. You probably want to reset the tick points to correspond to the revised bars on the histogram. Do this now and add this graph to your report.

   The X-axis is getting a little crowded, so let’s play with fonts and alignment. Click on the X–axis, go into edit the X-axis and to the font tab. Reset the font to 9 point. Now go to the alignment tab. Reset the alignment from 0 to 20. What does this do to the graph? (Add the graph to your report).

5. Turning to the Y-axis, frequencies are useful but percentages are more useful. Why?
Now click on the Y-axis, go to edit Y axis, go to the type tab and shift to a percent. You will get a nasty little error message but pay no attention. Now you have percent of the sample on your vertical scale.

6. Now it is time for an evocative title on this. Left click almost anywhere on the graph and go to select item. Then go to title. Let’s call this graph “Distribution of Hours for Managers and Professionals.” Next, the X-Axis title is nothing to write home about. Again, left click, go to select item and to X Axis Label. Re-label this “Usual Hours of Work.”

7. It is good form to let your reader know where you got your data from, so lets add a footnote. To do this, do a left click, to to “Add” go to “footnote” and add the text “Data from 2000 Outgoing Rotation files of the CPS”. Create the graph. The default font is a bit large, so double click on the footnote and reset the font to 8 point type.

8. As a final exercise with this data, lets create separate graphs for men and women. We are going to lose much of our nice formatting, but it don’t worry about that. Instead, go into limit our graph to women. To do this, go to

   Graph ▶ histogram ▶ simple graph (set uhour1rc) ▶ multiple graphs

   ▶ in separate panels of the same graph (force Y and X axis to be identical)
   ▶ by variable: by group in separate panels

   Run the histogram and save it. How do men and women’s hours compare?

c. That was tough, lets go to box plots. Under graph go to box plot and then to box plot by groups. Make the graph variable Uhour1rc and the group variable sex. What does this tell you about men’s and women’ usual hours? Why are there only two rather than three lines for Q1, the Median and Q3?

Repeat this exercise (just the box plot part of the assignment) with your race variable rather than Sex. What can you tell about the hours of different groups of managers by race?