Correlation Exercise

Research Methods, Fall 2015

[these instructions were made using Microsoft Excel for Windows 2013]

**If you get to a step that says “you should see...” and you see something different, flag me down and I’ll help!  
Remember, you can always “undo” a step if something gets messed up (keyboard shortcut control z on Windows, command z on mac).**

1. Open the Excel file: CorrelationExercise\_Fall2015.xlsx  
   Make sure you’re in the Sheet1 worksheet tab.  
   These are some data from 10 imaginary students. In most research you don’t have peoples’ real names in a data file, but instead use some kind of unique code for each participant. We have two variables: number of absences, and grade on an exam. These are both ratio level variables. I’ve labeled them X and Y. We have a value of each variable for every participant (student).  
   **Note the way the data are organized: each participant has one row, and each variable has one column.**  
   The cells you will be filling in are colored bright yellow.
2. First, let’s figure out what the mean value is for each of the two variables. We’re going to do that using a formula in Excel, which will do the basic arithmetic for us. We’ll start with the Number of Absences (X variable).
   * Double-click in cell C15 so you can start typing into it.
   * Note that the first thing you type in order to use an Excel formula in a cell will always the equal symbol.
   * So in cell C15, type the following: =AVERAGE(C4:C13)  
     (it doesn’t actually matter if you use all caps or not)
   * Press the return or enter key, and Excel will calculate the result of the formula.
   * **You should now see** the value 3.4 displayed in cell C15. That is the mean number of absences among these 10 students.
   * Note that 3.4 is the displayed result of the formula that is contained in cell C15. If you double-click on C15 again, you’ll see that you can edit the formula. Notice also that a cell’s formula is visible in the formula bar at the top of the spreadsheet window.
3. Now we’ll do the mean for Exam Grade in a slightly different way.
   * Now double-click in cell D15, and type just the following: =AVERAGE(
   * Use the mouse to click and drag to select cells D4 through D13. You’ve selected all the values in Exam Grade column.
   * Now you can just press the return or enter key again to complete the formula. (Excel will automatically add the closing parenthesis).
   * **You should see** the value 76 displayed in cell D15.
4. Now let’s figure out the standard deviation (*SD*) for each of the two variables. The standard deviation is a number that represents how spread out a distribution of scores is (how far the scores tend to be from their mean).
   * Double-click in cell C16, and type: =STDEVP(C4:C13)  
     then press the return or enter key
     + note you want to use the formula STDEVP(), **not** just STEDV(). Make sure you have the P! I have an explanation for this, but it’s technical.
   * **You should now see** the value 2.3748684 displayed in cell C16. That’s good that we have such a precise number, but we don’t need to be taxing our limited attentional resources looking at all those decimal places. Right-click on cell C16 and choose “Format cells” and then in the Number tab, click on Number instead of General. Click OK. [or find the format dropdown menu and change from General to Number] Now you should just see 2.37. Much nicer to look at! [There are multiple ways to do this in Excel’s interface.] Note that Excel still has the full number stored, it’s just limiting the decimal places it displays to us.
5. Now for the SD of Exam Grade, you’re going to copy and paste your formula from cell C16 to cell D16
   * Right-click once on cell C16 and choose COPY from the contextual menu that pops up.
   * Left click once on cell D16 to highlight it. Note that cell C16 is surrounded by an animated dashed line, indicating that we’ve just copied the formula from that cell.
   * Now paste into D16, by right-clicking and choosing PASTE from the contextual popup menu, or by going to the Edit menu and choosing PASTE, or by using a keyboard shortcut (control V on Windows, command V on Mac).
   * **You should now see** the value 15.04 displayed in D16. Look at the formula bar. The formula should show as =STDEVP(D4:D13) . Excel automatically changed the references from the C column to the D column, because they are relative references. (To make a reference to another cell an absolute reference that won’t be automatically changed when you copy/paste, you have to a put a dollar sign in front of the reference for both row and column, so for example $C$4:$C$13. You don’t want to do that just yet though.) Note also that the cell format, which is to only show you the first 2 decimal places, also got copied over.
6. A formula for the Pearson correlation coefficient (*r*) is:   
   * So we need to calculate the z-score (standardized score) for each participant for both variables.
   * We’ll start by doing the standardized # Absences for Student01. Standardization of a single score is always done with respect to some distribution of scores; in this case the distribution is our sample of size ten. For Student01 we need to calculate the difference between this participant’s score (4) and the mean of this variable in our sample (3.4), and then divide that difference by the standard deviation of this variable in our sample (2.37).
   * Double-click in cell F4 and type: =(C4-C15)/C16 t  
     hen press return/enter
     + Make sure you get the parentheses right!
   * **You should see** the value 0.252645576 in cell F4.
   * Let’s try to save ourselves some work by copying the formula from that cell to the rest of the cells in this column. Right-click once and copy cell F4. Then left-click and drag to select cells F5 through F13. Now paste.
   * **You should see...** OMG nooooooo, what happened?! #DIV/0! all over the place instead of the values we want. Double-click into cell F5 (Student02’s row) to see what’s going on.
   * Excel has color-coded highlighted the reference cells to show you which cells the formula in F5 is referring to. What we want to be happening in this cell is (4-3.4)/2.37. But when we copied and pasted the formula from F4 (Student01) to F5 (Student02), Excel automatically changed the rows for all the references in the formula. Now the formula in F5 is pointing to an empty cell (C17) instead of the standard deviation like we want. That’s because we were using relative references. Let’s fix that.
   * Go back to cell F4 and change the formula to this: =(C4-$C$15)/$C$16  
     We’re now using absolute references for the mean and SD, which locks the references in place so they won’t change when we copy/paste the formula to another cell. Note that we’re leaving C4 as a relative reference, because we DO want that one to change automatically so it points to the # Absences for the participant’s row.
   * Now you can try again to copy/paste the formula from F4 to the other cells in this column (F5 through F13).
   * **You should see** -0.168430384 in cell F13. Double-click on F13 to see how it’s pointing to the correct cells.
7. Now let’s do the standardized scores for Exam Grades.
   * In cell G4, type the following: =(D4-$D$15)/$D$16   
     then press return/enter
   * Then copy that cell and paste to the rest of the cells in this column (G5 through G13) just like we did with the # Absences.
   * **You should see** 0.132979194 in cell G13
8. We’re getting close! Let’s take a look at that formula for the correlation coefficient again:  
   * So what we need to do is, for each participant we multiply their two z-scores together. We’ll do that in column I. Then we’ll add up those products across all participants and divide by *N*, which is the number of participants (10).
   * In cell I4, type: =F4\*G4  
     or instead of typing you can just type the equal sign, then click once on F4, then type the asterisk, then click once on G4. See, you can make formulas through a mixture of typing and point-and-click.
   * **You should see** 0.1008 in cell I4
   * Copy/paste that cell to the rest of the cells in that column (I5 through I13)
   * **You should see** -0.022 in cell I13
   * Now we need to add up those products (this is what the Σ symbol tells us to do). In cell I15 type: =sum(I4:I13)
   * **You should see** -8.511 in cell I15
   * The last component we need is *N,* the number of participants in our sample. We know that *N* = 10 in this case, so you could just type 10 into cell K4. But sometimes you’ll have a much bigger data set and maybe you’ve had to remove some participants’ rows (e.g., because they were texting while doing your task), or maybe you’re missing data from some participants (empty cells!), so you want to have Excel double-check exactly how many participants’ rows of data you’ve got here now. There’s a COUNT() formula you can use. In K4, type: =COUNT(I4:I13)
   * Now, FINALLY, you can make a formula in the final cell, for the correlation coefficient *r*. In cell K8, type: =I15/K4
   * **You should see** -0.851 in cell K8. That’s a really strong negative correlation!
9. Okay, so now I’ll show you a much quicker way to calculate the correlation coefficient in Excel. Don’t be mad at me; it was a worthwhile learning experience for you to have done everything you did so far!
   * Click over to Sheet2
   * Now there’s one new yellow cell for you to fill in: M4.
   * So, it turns out there’s a function in Excel called CORREL() . Between those parentheses you have to enter the range of cells for the first variable, then a comma, then the range of cells for the second variable. So, in M4, type: =CORREL(C4:C13,D4:D13)  
     or you could alternatively type just =CORREL( and then click-drag to select C4:C13, then type a comma, then click-drag to select D4:D13
   * **You should see** -0.851 in cell M4
   * Note how it’s the same value as we got in K8! Using the CORREL() function saves you all the middle steps of making the standardized scores and multiplying them. You wouldn’t even need to make columns F, G, and I.
   * So this correlation coefficient seems pretty strong. But how do we tell if it’s statistically significantly different from zero? That is, how do we tell if we can reject the null hypothesis? A *t*-test! I’ve included the formulas that do a t-test in the 3 cells to the right of M4. *df* is degrees of freedom. That’s *N*-2 for a correlation. *t* is the *t*-test statistic, and if you double-click to see the formula you’ll see *t* can be calculated from *r* and *df*. (note the SQRT() function just does a square root). Finally, *p* is the probability that we would have gotten a *t* value as extreme as we did just due to chance. *p* is less than .05 (5%), which by convention is the cutoff for rejecting the null hypothesis. So yes, this correlation is statistically significant. More absences is associated with lower exam grades.
   * **You should see** 0.0037 in cell P4
10. Now let’s make a scatterplot to visualize the relationship between the two variables.
    * Click-drag to highlight cells C3 through D13. (note this includes the two cells that label the variables, # Absences and Exam Grade)
    * Click on the “Insert” tab in the “ribbon” toward the top of the window.
    * In the Charts section of the ribbon, click the button that looks like a little scatterplot and choose Scatter (top-left option).
    * Ta-da, a scatterplot! Too bad Excel is kinda stupid and probably put the scatterplot right on top of the data. You can click and drag to move the whole chart further down onto the worksheet so it’s not covering up the data. The rest of the steps here are to make the scatterplot a lot less stupid looking.
    * Click the “Home” tab in the “ribbon” toward the top of the window.
      + Set the font to Arial (instead of Calibri) and size 14 (instead of 10). Also set the Font Color to black (not the Fill Color!).
    * Click the “Chart Tools: Design” tab in the ribbon toward the top of the window.
      + (If you don’t see the “Chart Tools” tabs, you may have clicked off the scatterplot into a cell. Click once on the scatterplot and the “Chart Tools” tabs should appear in the ribbon.)
    * Click Add Chart Element, then Chart Title 🡪 None.
    * Click Add Chart Element, then Gridlines 🡪 Primary Major Horizontal. This should **remove** the horizontal gridlines that were included by default.
    * Click Add Chart Element, then Gridlines 🡪 Primary Major Vertical. This should **remove** the vertical gridlines that were included by default.
    * Click Add Chart Element, then Axis Titles 🡪 Primary Horizontal. Now type: “Number of Absences” without quotes then press enter/return.
    * Click Add Chart Element, then Axis Titles 🡪 Primary Vertical. Now type: “Exam Grade” without quotes then press enter/return.
    * Now click over to the tab “Chart Tools: Format” in the ribbon.
    * There’s a dropdown menu in the top-left of the ribbon now. It probably right now says “Vertical (Value) Axis Title”. Click on this dropdown menu and choose “Vertical (Value) Axis” (not the title, just the axis). Then click the Format Selection button just below the dropdown menu.
      + A formatting panel should now be opened for the vertical axis. In the “Axis Options” section of this panel (indicated by a little icon that looks like a bar graph), you want to set the Maximum bound to 100 instead of 120. (Why? Because 100 is the maximum score participants could get, so it doesn’t make sense to show larger values on the scale.) So type 100 into the Maximum box and press enter/return.
      + Also, under Tick Marks, set Major Type to Outside.
      + In the formatting panel, you now need to get to the “Fill & Line” subsection by clicking on the paint bucket icon. Change the Line color (not Fill!) from gray to black.
    * Now in that dropdown menu in the top-left of the “Chart Tools: Format” ribbon, you want to choose “Horizontal (Value) Axis”.
      + In the formatting panel, get to the Axis Options subsection by clicking the icon that looks like a bar graph. You’ll see that the Maximum bound is already set to 10. But you nevertheless need to type 10 into the Maximum box and press enter/return. This will prevent the maximum from being changed automatically later.
      + Also, under Tick Marks, set Major Type to Outside.
      + In the formatting panel, you now need to get to the “Fill & Line” subsection by clicking on the paint bucket icon. Change the Line color (not Fill!) from gray to black.
    * The dots in the scatterplot are probably blue by default and we need to make them black. The color doesn’t convey information, and anyway most journals still have a standard of figures being in grayscale. In that dropdown menu in the top-left of the “Chart Tools: Format” ribbon, you want to choose “Series ‘Exam Grade’”.
      + In the formatting panel, you need to get to the “Fill & Line” subsection by clicking on the paint bucket icon. Then click “Marker”.
      + Under Fill, change the color to black.
      + Under Border, change the color to black.
      + Under Marker Options, you can make the dots bigger by clicking Built-in then increasing the size value. You can also change the dot shapes if you want.
    * In that dropdown menu in the top-left of the “Chart Tools: Format” ribbon, choose “Chart Area”.
      + In the formatting panel, in the “Fill & Line” subsection (paint bucket icon), change the Border to No line.
    * In that dropdown menu in the top-left of the “Chart Tools: Format” ribbon, choose “Plot Area”.
      + In the formatting panel, in the “Fill & Line” subsection (paint bucket icon), change the Border to Solid line.
    * Just look at that totally legit scatterplot! You can resize it a bit if you want, by clicking and dragging one of the corners of the whole thing.
    * Click over to Sheet3 to see a scatterplot that I made. Yours should look something kind of like that. Aww yeah.
11. So we did a correlation analysis, and made a scatterplot. Want to do a regression analysis too? The answer is yes, yes you do.
    * Go back and take a look at the regression slide in the Chapter 6 Powerpoint to remind yourself what the heck regression is about.
    * Click over to Sheet4.
    * Note you’ve now got a Regression section down at cell L17.
    * Excel has some handy functions for simple linear regression. In cell N19, type just the following: =INTERCEPT(  
      There should be a little hovering window that shows you what the function wants inside the parentheses: known\_y’s, known\_x’s
      + First, click-drag to select the Y values, that is: cells D4 thru D13  
        or just type D4:D13
      + Then type a comma
      + Then, click-drag to select the X values, that is: cells C4 thru C13
      + Then press enter/return to complete the formula (Excel automatically adds the closing parenthesis)
      + **You should see** 94.326 in cell N19.
      + This is the intercept of the regression line that would best fit these data. This value means that based on the data we have, we’d predict that a student with zero absences would get a 94.326 exam grade, on average.
    * Now for the slope.
      + In cell N20, type just the following: =SLOPE(
      + First, click-drag to select the Y values, that is: cells D4 thru D13  
        or just type D4:D13
      + Then type a comma
      + Then, click-drag to select the X values, that is: cells C4 thru C13
      + Then press enter/return to complete the formula (Excel automatically adds the closing parenthesis)
      + **You should see** -5.39 in cell N20.
      + This is the slope of the regression line that would best fit these data. This value means that based on the data we have, for each additional absence a student has we predict an exam grade that is lower by 5.39 points, on average.
    * Finally, let’s add a regression line to the scatterplot to visualize it.
      + Click once on the scatterplot to activate it.
      + Click the “Chart Tools: Design” tab in the ribbon toward the top of the window.
      + Click Add Chart Element, then Trendline 🡪 Linear Forecast. w00t, check that out!!
      + Now go to the “Chart Tools: Format” tab in the ribbon, then in that dropdown menu in the top-left of the ribbon, choose “Series ‘Exam Grade’ Trendline1”.
        - In the formatting panel, if the regression line is dotted blue, which it might be by default, go to the Fill & Line subsection (paint bucket icon), under Line, set the color to black and Dash Type to Solid.
        - In the formatting panel, in the Trendline Options subsection (icon that looks like a bar graph), check the box for Display equation on chart”. It’s formatted kinda dumb, with too many decimal places and the intercept after the slope, but now you can see the equation that corresponds to the regression line.
      + **You should see** Y = -5.3901x + 94.326 on the scatterplot.
      + Click over to Sheet5 to see what your Sheet4 should look like, more or less.
12. You’re done! Just need to turn in your work.
    * Save your Excel file as “YourName\_CorrelationExercise.xlsx” (except replace “YourName” with your name)
    * Upload it to Blackboard by going to the class Blackboard page, then Course Content, then In-Class Activities, then Correlation Exercise, then “Correlation Exercise turn in here”