**COURSE\_NUM – Bivariate statistics - Workshop Activity**

**Task 1:** Using the GSS 2018 SURF aggregate data (“Task 1 + 2” tab), create a bar chart that displays differences between men and women in the proportion of who report feeling lonely some, most, or all of the time, by age group.

*Note. We are working with small cell sizes here and these statistics may not be trustworthy or be large enough to pick up statistical differences. We’ll talk about this throughout the course.*

1. *Create percents from raw population numbers in new separate columns*

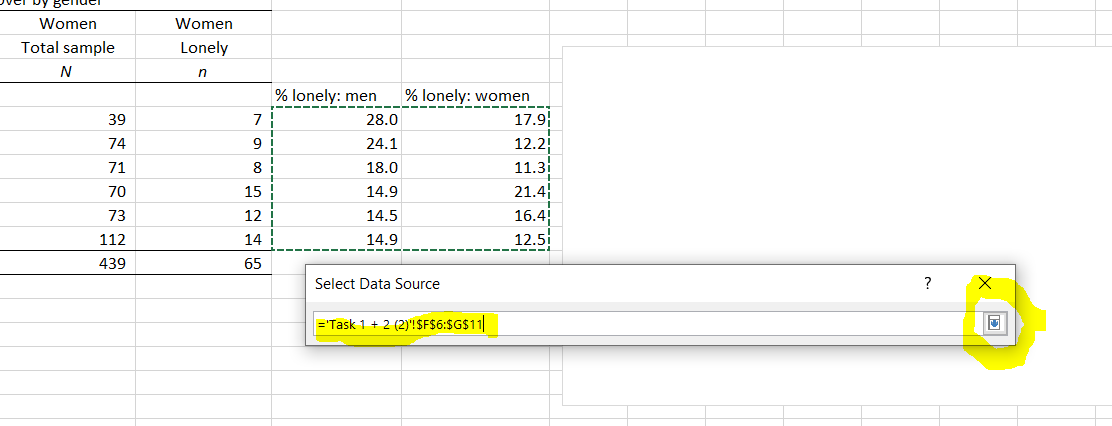
* Formula for percents is the population of lonely people (numerator) divided by the total sample (denominator)
  + For men: “=C6/B6” *or* “=(C6/B6)\*100”
  + For women: “=E6/D6” *or* “=(E6/D6)\*100”

1. *Select chart*

* “Insert” tab, “charts” ribbon tab
* Click on “2-D column”

1. *Assign data*

* Right-click on chart, click on “Select data”
* Click on little box with arrow next to “Chart data range”
* Select cells with percent values you want included on the chart



1. *Assign value labels to axis and series*

* Horizontal axis:
  + Click “Edit” button under “Horizontal (Category) Axis Labels”
  + Select cells with label values (i.e., the age groups in column A)
* Series labels:
  + Highlight the series you want to change name of
  + Click “Edit” button under “Legend Entries (Series)”
  + Select cells with series label values
    - For the men series, select any cell that says “men”
    - For the women series, select any cell that says “women”

1. *Add axis and chart titles*

* “Design” tab, click on “Add chart element”
  + Click on “Axis titles” to add horizontal and vertical axes
  + Click on “Chart title” to add chart title to top of chart
  + Click on “Legend” to add a legend to the chart to denote which bar represents women and which bar represents men

1. *Make it pretty*

* Right-click on chart, click on “Format Chart Area”
* Click on elements of chart you want to change

**Task 2:** Using GSS 2018 SURF aggregate data (“Task 1 + 2” tab), demonstrate how the age composition of loneliness differs for men versus women.

1. *Two options for organising data:*
2. Create within-year column percents;
   * + Copy and paste age group labels in rows under the data
     + Estimate what proportion of lonely men are aged between 15-24 years, 25-34 years, etc., by dividing the number of lonely men in each age group by the total number of lonely men.
     + Example: for men in age group: =14 (number of 15-24 year old lonely men) / 100 (total lonely men) = 14.0% of lonely men are aged between 15-24 years
       - The Excel equation is: “= (C6/C$12)\*100”
       - Drag down to copy and paste the equation
       - Repeat for women (=E6/E$12)\*100
     + Use column percents in cumulative column graphs
3. Use raw count data in cumulative column graphs.
4. *Follow the steps above except chart type will be “100% stacked column.”*

**Task 3:** Use the unit/person-level data to produce descriptive statistics. This data set includes a sample of 997 people.

*Variables:*

id: Unique identifying number

Sex: 0 = man; 1 = woman

Age: 1 = 15-24 years; 2 = 25-34 years; 3 = 35-44 years; 4 = 45-54 years; 5 = 55-64 years; 6 = 65 years or older.

lonely: 1 = lonely none of the time; 2 = lonely a little of the time; 3 = lonely some of the time; 4 = lonely most of the time; 5 = lonely all of the time.

lonely\_yes: 0 = reports feeling lonely none or a little of the time; 1 = reports feeling lonely some, most, or all of the time.

***Fill out the following table***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Question | Calculate: | Population | Formula | Answer |
| A | Mean | Proportion of women feeling lonely some, most, or all of the time | =AVERAGEIFS(E2:E998,C2:C998,"=1")  *Note.* This says create an average of the values in the E column, but only include the value if the values in the C column that say “1”, which indicates the respondent is a woman. |  |
| B | Mean | Proportion of men feeling lonely some, most, or all of the time | =AVERAGEIFS(E2:E998,C2:C998,"=0")  *Note.* This says create an average of the values in the E column, but only include the value if the values in the C column that say “0”, which indicates the respondent is a man. |  |
| C | Mean | Proportion of women feeling lonely aged 15-***34*** years | =AVERAGEIFS(E2:E998,C2:C998,"=1",B2:B998,">0",B2:B998,"<3") |  |
| D | Mean | Proportion of men feeling lonely aged 15-***34*** years | =AVERAGEIFS(E2:E998,C2:C998,"=1",B2:B998,">0",B2:B998,"<3") |  |
| E | Mean | Average level of loneliness among men | =AVERAGEIFS(D2:D998,C2:C998,"=0") |  |
| F | Standard deviation | Standard deviation of loneliness among men | =STDEV.S(IF(C2:C998=K12,D2:D998))  *Note.* You need to create three new cells: One that says “sex” (K11),one that has a “0” (K12), and one that has “1” (K13).  Here you ask Excel to match criteria, where values on the C column match K11 (0 = man). |  |
| G | Mean | Average level of loneliness among women | =AVERAGEIFS(D2:D998,C2:C998,"=1") |  |
| H | Standard deviation | Standard deviation of loneliness among women | =STDEV.S(IF(C2:C998=K13,D2:D998)) |  |

**Task 4:** Conduct a Chi2 test using unit/person-level data.

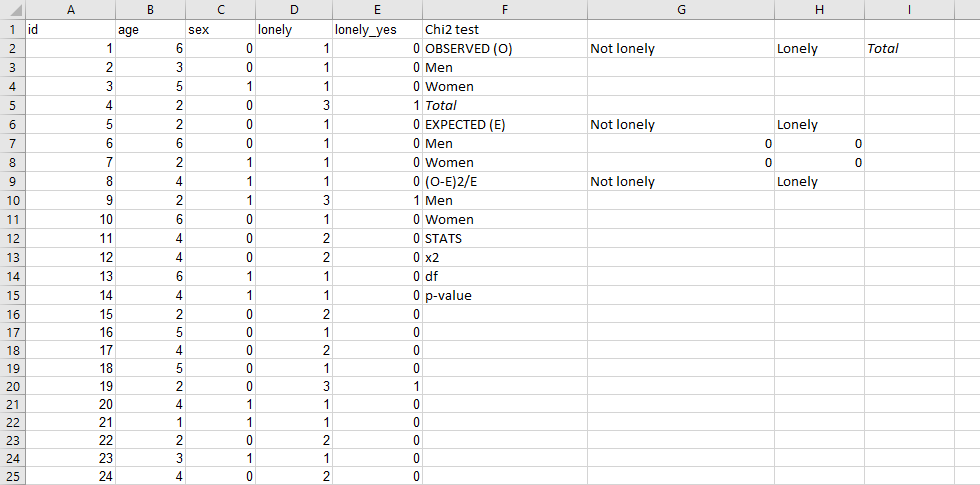
1. *Write out the null and alternate hypotheses testing whether the proportion of people feeling lonely some, most, or all of the time that are* ***aged 15-34 years*** *differs by gender.*

*H0:*

*HA:*

1. *Conduct a Chi2 test to determine whether the percent of people* ***aged 15-34 years*** *who are lonely differs by gender (men vs. women).*
   1. *Prepare a small table where you will record your observed, expected, and calculated values, like this:*

*Note: The rest of the instruction will only make sense if you set up your table exactly like this, in the same cells.*



* 1. *Create the observed values*
  + In cell G3, count the number of non-lonely people who are men aged between 15-34 years old, by entering: =COUNTIFS($E$2:$E$998,"=0",$C$2:$C$998,"=0",$B$2:$B$998,">0",$B$2:$B$998,"<3")
  + Copy and paste this equation into remaining three cells (i.e., lonely men, non-lonely women, lonely women), altering the gender and whether counting lonely people or non-lonely people.
  + Sum the number of non-lonely people and lonely people by gender: for example, in cell I3, type =sum(G3:H3)
  + In cell I5, create a sum of either the row or the column
  1. *Create the expected values: The expected value is calculated by multiplying the row total and column total and dividing by the total sample size*
  + Cell G7 – Non-lonely people among men: =(I3\*G5)/I5
  + Cell G8 – Non-lonely people among women: =(I4\*G5)/I5
  + Cell H7 – Lonely people among men: =(I3\*H5)/I5
  + Cell H8 – Lonely people among women: =(I4\*H5)/I5
  1. *Calculate the [observed – expected]2 divided by the expected value*
  + Cell G10: =(G3-G7)^2/G7
  + Copy and paste this equation into the remaining three cells. The equation will adjust to select the correct cells as you paste.
  1. *Run the Chi2 test*
  + Calculate the Chi2 critical value in cell G13 by summing the values produced in step d above: =SUM(G10:H11)
  + Calculate the degrees of freedom in G14, which it the number of rows in your matrix minus 1, multiplied by the number of columns in your matrix minus 1: =(2-1)\*(2-1)
  + Run the Chi2 test in G15: =CHISQ.DIST.RT(G13,G14)

**Task 5:** Conduct a two sample t-test using unit/person-level data

*Note: Your version of Excel may not have the data package that is needed to run this test loaded. To load the package:*

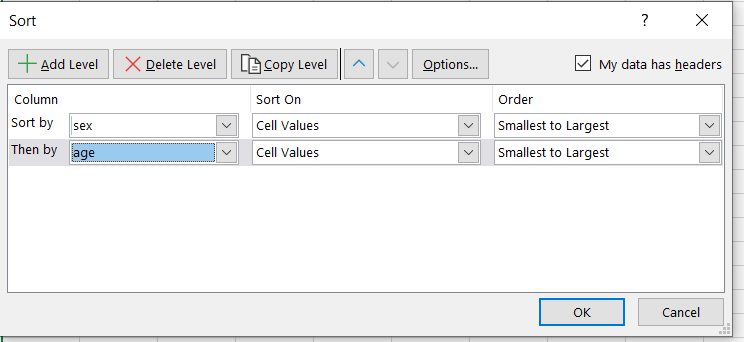
* *Click “Data” in the ribbon*
* *Click “Analysis tools.” A pop up box will appear*
* *Tick the “Analysis ToolPak” box, if it is not ticked, and click “OK.”*

1. *Write out the null and alternate hypotheses testing whether the average levels of loneliness (i.e., the “lonely” variable) among those* ***aged 15-34 years*** *differs between men and women.*

*H0:*

*HA:*

1. Conduct the t-test
   1. *The first step is to rearrange your data in preparation for the t-test. You will need to create two columns for the two groups you are comparing that hold the outcome variable data. Use the sheet “Task 5.” Because we are examining the loneliness levels of women and men aged between 15-34 years, we need to create two columns: 1) loneliness among young men; 2) loneliness among young women.*
   * Highlight columns A through E by clicking on the “A” at the top of column, holding down the mouse and dragging until columns A through E are selected.
   * Click “Data” in the ribbon, then click “Sort”.
   * Sort by “sex” and the option “smallest to largest”, then click the “+” symbol in the lower lefthand corner and sort by “age”



* + Rename column G “Loneliness among young men” and column H “Loneliness among young women”
  + Copy and paste into column G the values in the “lonely” column (column D) all the way until the end of where sex=0 and age=1 or 2
  + Copy and paste into column H the values in the “lonely” column (column D) where sex=1 and age=1 or 2
  1. Click “Data in the ribbon”
  2. Click the “Data Analysis” button on the right. A pop-up box will appear.
  3. Click on “t-Test: Two-Sample Assuming Equal Variances” option
  4. Click on the box with the arrow in the same line as “Variable 1 range”. Click and drag down to highlight all the values in column G, including the variable name at the top of the column (i.e., loneliness of men aged 15-34 years old, plus the name of the variable). Click the box with the arrow again.
  5. Repeat this step: Click on the box with the arrow in the same line as “Variable 2 range”. Click and drag down to highlight all the values in column H (i.e., loneliness of women aged 15-34 years old), including the variable name at the top of the column. Click the box with the arrow again.
  6. Enter “0” as the hypothesized mean value (i.e., the null hypothesis: there is no difference in the mean loneliness between men and women aged 15-34 years old).
  7. Click the “Labels” box.
  8. Leave alpha at “0.05”
  9. In output range, click the arrow button, and click “J2” or any cell that does not have data in it nor data to the right of it.

The “P(T<=)two-tail” provides you with the p-value that determines whether the mean differences in loneliness between men and women aged 15-24 years old are different from each other.

**Task 6:** Using the information you produced in tasks 1-5, briefly describe your findings. In particular, identify how overall rates of loneliness differ by gender and by age.