To: Med students
From: Jessica Ancker
Date: March 2018
Re: Instructions for lab 1

OVERVIEW:

This lab is designed to give everyone the opportunity to learn how to compute descriptive and inferential statistics in Excel, and provide practice in interpreting statistics. The lab consolidates information from sessions 1 through 4 in the stats module.

- Please bring a laptop if possible – if not possible, you can share with people from your lab session;
- Please report to the small classroom assigned for this session, where you will be joined by 10-11 other students and one instructor;
- Please download the Excel spreadsheet **Data_for_lab_2018.xlsx** from Canvas, preferably before the lab;
- Some of the questions on Homework 1 use this data set and are very similar to questions in the lab session -- feel free to work on these questions during the lab or afterwards.

NOTE: The data are a subset of the National Health and Nutrition Examination Survey (NHANES), a large data set collected by the CDC annually. http://wwwn.cdc.gov/Nchs/Nhanes/Search/nhanes13_14.aspx . I've included 3000+ patients in this data set for your use. Explanations of the variables we will focus on during the lab are provided below in the Appendix. I have included a variety of other variables – in general, 1 = yes and 2 = no, but for more information on any of the variables in the spreadsheet, you can search: http://wwwn.cdc.gov/Nchs/Nhanes/Search/default.aspx

STEP-BY-STEP INSTRUCTIONS:

1. **Review the variable names and check the meaning of the codes in the Appendix below.**
    a. Identify 5 continuous variables
    b. Identify 5 categorical variables

2. **Compute descriptive statistics for variables *age* and *total cholesterol (mg/dL).* Interpret these statistics in English**

Code in Excel:

a. Place your cursor in an empty cell
b. To compute the average of all values in column C, type "= average(Data_for_lab!C:C)"
c. To compute the standard deviation of all values in column C, type "=stdev(Data_for_lab!C:C)"
d. To identify the smallest of all values in column C, type "=min(Data_for_lab!C:C)"
e. To identify the 25% percentile of all values in column C, type "=percentile.exc(Data_for_lab!C:C, 0.25)"
f. To identify the median of all values in column C, type "=median(Data_for_lab!C:C)"
g. To identify the 75% percentile of all values in column C, type "=percentile.exc (Data_for_lab!C:C, 0.75)"
h. To identify the maximum of all values in column C, type "=max(Data_for_lab!C:C)"

3. **Compute descriptive statistics (frequencies) for the variables *hepatitis A antibody* and *marital status.***

Code in Excel using PivotTable:

a. Use your cursor to select all columns with data in them. Click the Insert tab at the top of the page, then select PivotTable. (Alternatively, you may click the Insert Tab – PivotTable first, and then type in "Data_for_lab!Data_for_lab" in Table/Range field.)
b. Confirm that the Table/Range field contains the appropriate data and the New Worksheet option is checked, then select OK. This will bring you to a new worksheet with a blank PivotTable, with all of the available variables listed on the right-hand side of your spreadsheet.
c. On the right-hand side of your spreadsheet, drag the variable *hepatitis A antibody* down into the panel labeled Rows
d. Drag PAT_ID into the panel labeled Values. By default, it may show "Sum of PAT_ID" which may not have any meaning. Click "Sum of PAT_ID", select "Value Field Settings", select "Count", and click "OK." This will provide a count of individual patients in each of the hepatitis answer categories. *(cf. In "Data_for_lab" you have records from row #2 to row #3916, so you have a total of 3915 records. Therefore, your "Grand Total" in you PivotTable should be 3915.)*
e. You can now manually compute the proportions. Or else, dragging PAT_ID into the panel labeled Values again. Click "Sum of PAT_ID", select "Value Field Settings", select "Count". This time, click the tab "Show Values As." On the dropbox menu, select % OF COLUMN TOTAL. *(cf. You should get 46.26% and 53.74%. Make sure you did not get the percentages of "Sum of PAT_ID," which may be meaningless.)*

4. **Compute the relative risk (RR), odds ratio (OR), and absolute risk difference (ARD) of hepatitis A among women versus among men**
   a. Looking at the percentages, does it appear to you that there's any association between gender and hepatitis A?
   b. Compute and interpret the relative risk, the odds ratio, and the absolute risk difference – do these statistics suggest any association between gender and hepatitis A?

Code in Excel using PivotTable:

a. Either create a new Pivot Table or clear your existing one by unselecting all the variables that you previously used.
b. Drag the variable GENDER into the panel labeled Rows, *hepatitis A antibody* into Columns, and PAT_ID into VALUES.
c. Refer to the instructions above to compute the percentages.(Percentages: Note that you have choice of showing "% of Grand Total", "% of Column Total", and "% of Row Total". Try all to see how they might help you interpret your findings.)
d. To calculate a relative risk (RR), place your cursor in an empty cell and type the code "=[cell 1]/[cell 2]" where cell 1 and cell 2 are the 2 Excel spreadsheet cells with the relevant data. Identify the relevant cells to compute the relative risk for women versus men.
e. Follow these instructions to compute the odds of hepatitis A among women, the odds of hepatitis A among men, and the odds ratio for women versus men.
f. To subtract 2 values, use the equation "= [cell 1] - [cell 2]". Use this approach to compute the absolute risk difference for women versus men.

5. **Perform a chi-squared test to determine whether the hepatitis A rates are significantly different among women and men**

Code in Excel:

a. Create a new 2 x 2 table in Excel, with women/men as rows, and Hep A/no Hep A as columns. Put the right values into the interior cells and the margins. This shows the <u>observed values</u> of the data.

	HepA	No HepA	Total
Women	Observed number here (a)	Observed number here (b)	Total across (a+b)
Men	Observed number here (c)	Observed number here (d)	Total across (c+d)
Total	Total down (a+c)	Total down (b+d)	Grand total (a+b+c+d)

b. Create a new 2 x 2 table in Excel to show the expected values.

	HepA	No HepA	Total
Women	(a+b)*(a+c)/ (a+b+c+d)	(a+b)*(b+d)/ (a+b+c+d)	Total from the observed table (a+b)
Men	(c+d)*(a+c)/ (a+b+c+d)	(c+d)*(b+d)/ (a+b+c+d)	Total from the observed table (c+d)
Total	Total from the observed table (a+c)	Total from the observed table (b+d)	Grand total from the observed table (a+b+c+d)

c. Conduct a hypothesis test using the chi-squared test.
   a. Compute the (observed-expected)^2/(expected) for all four cells of the table, and add them up using =SUM(a, b, c, d). This is the chi-squared value (test statistic).
   b. The code to produce the P value is "=chisq.dist.rt([test statistic], [degrees of freedom])". The degrees of freedom for chi-squared statistic is (number of rows - 1) * (number of columns - 1).
   c. Make a decision – reject the null, or fail to reject?
   d. Interpret your decision in English!

6. **Perform a chi-squared analysis to determine whether there is any association between GENDER and SERVED ACTIVE DUTY IN US ARMED FORCES.**
   Use the code from the previous question.

7. **Given the mean total cholesterol (mmol/L) in the data set, estimate the population mean cholesterol (mmol/L) and compute a 95% confidence interval for the mean**
   Recall that a 95% confidence interval for a population mean is $\overline{x} \pm (1.96) * (standard\ error\ of\ the\ mean)$. Furthermore, $standard\ error\ of\ the\ mean = (SD)/\sqrt{sample\ size}$.
   Code in Excel:
   Here are the functions you will need in Excel - you can do the steps one at a time, or compile them into a single cell if you want:
   a. =average([cell range])
   b. =stdev([cell range])
   c. =count([cell range]) --- NOTE: This gives you a count of all of the nonmissing values, i.e., the sample size
   d. =sqrt([cell range])

8. **Compute the 95% confidence interval for the mean *HDL (mmol/L)* among men only and also among women only. Judging by the confidence intervals, are men and women significantly different in HDL levels?**

   Code in Excel:
   - You can accomplish this in PivotTable. Drag the "Gender" variable into the ROWS, and "Count of PATID" into VALUES. Now put the HDL variable into the VALUES column and use Value Field Settings to show the average of HDL. You can put the HDL variable into the VALUES column a second time and use Value Field Settings to show the standard deviations as well. Now you can compute the two confidence intervals using the formulas above.

9. **Do a hypothesis test for independent samples of continuous data: Compare *Direct HDL (mmol/L)* levels among men and women by a) computing descriptive statistics within each group, and 2) computing the test statistic, and c) a p value.**
   Code in Excel
   a. Compute descriptive stats on each group. How big is the difference between groups?
   b. Compute the student's t statistic (test statistic) using the formula t = $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\text{SE}_{\text{Welch}(\bar{X}_1-\bar{X}_2)}}$ under the null hypothesis (μ_{diff}=0).
      i. Note that we have to calculate SE from two independent group. Calculation of SE requires some additional assumption. When we cannot assume an equal variance among the comparison groups, a good choice is to use Welch's t-test (a.k.a. unequal variances t-test, unpooled variances t-test). Therefore, let us use Welch's approximation using the sample standard deviations (s_1, s_2) as follows:

$$\text{SE}_{\text{Welch}(\bar{X}_1-\bar{X}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

      ii. There are two assumptions required for Welch's t-test: (1) the comparison groups are independent simple random samples (or a randomized experiment), (2) normally distributed populations (or large enough sample size to meet the Central Limit Theorem). However, the advantage of Welch's t-test is that it does not require an assumption of equal variance among the comparison groups.
   c. Compute the p value using " = t.dist.rt(test statistic, degrees of freedom) * 2.
      i. Note that the degrees of freedom are also different for Welch's t-test. The degrees of freedom is calculated from Welch-Satterthwaite's equation as follows:

$$df_{\text{Welch}-\text{Satterthwaite}} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

        ii.

d. Make a decision (reject the null or fail to reject the null)

e. Interpret your finding in English!

f. Why is this an independent-sample test?

10. Do a hypothesis test for paired continuous data: Compare *Direct HDL (mmol/L)* with *Second HDL* among all people in the sample

Code in Excel:

a. Compute the individual differences between the two HDL measurements on each person using the formula =(cell1) – (cell2). You can put the formula in one cell and then copy and paste to put it into the entire column.

b. Compute the mean and standard deviation of the differences. How big is the average difference?

c. Compute the student's t statistic using the formula t = $\frac{\bar{X}_{diff} - \mu_{diff}}{SE_{diff}}$ under the null hypothesis (μ_{diff}=0).

d. Compute the p value (see above)

e. Make a decision (reject the null or fail to reject the null)

f. Why is this a paired t-test?

APPENDIX: Variables in Dataset_for_lab.xlsx

PAT_ID = patient ID

Gender:

Code or Value	Value Description
1	Male
2	Female
	Missing

Race/Hispanic origin w/ NH Asian

Code or Value	Value Description
1	Mexican American
2	Other Hispanic
3	Non-Hispanic White
4	Non-Hispanic Black
6	Non-Hispanic Asian
7	Other Race - Including Multi-Racial
	Missing

Served active duty in US Armed Forces

Code or Value	Value Description
1	Yes
2	No
7	Refused
9	Don't Know
	Missing

Served in a foreign country : (question administered to those who answered yes to history of military service)

Code or Value	Value Description
1	Yes
2	No
7	Refused

9	Don't Know
	Missing

Country of birth

Code or Value	Value Description
1	Born in 50 US states or Washington, DC
2	Others
77	Refused
99	Don't Know
	Missing

Education level - Adults 20+

Code or Value	Value Description
1	Less than 9th grade
2	9-11th grade (Includes 12th grade with no diploma)
3	High school graduate/GED or equivalent
4	Some college or AA degree
5	College graduate or above
7	Refused
9	Don't Know
	Missing

Marital status

Code or Value	Value Description
1	Married
2	Widowed
3	Divorced
4	Separated
5	Never married
6	Living with partner
77	Refused
99	Don't Know
	Missing

Pregnancy status at exam
(administered to women 20 through 44)

Code or Value	Value Description
1	Yes, positive lab pregnancy test or self-reported pregnant at exam
2	The participant was not pregnant at exam
3	Cannot ascertain if the participant is pregnant at exam
	Missing

household_income

Code or Value	Value Description
1	$ 0 to $ 4,999
2	$ 5,000 to $ 9,999
3	$10,000 to $14,999
4	$15,000 to $19,999
5	$20,000 to $24,999
6	$25,000 to $34,999
7	$35,000 to $44,999
8	$45,000 to $54,999
9	$55,000 to $64,999
10	$65,000 to $74,999
12	$20,000 and Over
13	Under $20,000
14	$75,000 to $99,999
15	$100,000 and Over
77	Refused
99	Don't know
	Missing

hep_A

Code or Value	Value Description
1	Positive
2	Negative
3	Indeterminate
	Missing