HW1
ABG 327: Data Analysis
Spring 2018

For this assignment, you may work in pairs. The submission will be **one** Excel file to PolyLearn. If you work with partners both names must be on the first two lines of the workbook. Only one person needs to submit the assignment for each pair. Formatting counts. If it is hard to find the correct answer(s) you will be marked wrong. Make sure all your responses are well-labelled and to use text boxes for longer answers. **Your hypothesis tests should include all 6 steps and written conclusion. You should highlight your calculated answers.**
Never accept population variances as known unless told otherwise. Always treat variances as sample variances.

Every component of this course incorporates hypothesis testing and it is an essential element of most quantitative analyses. The following problems provide practice on solving one population mean and proportion hypothesis tests. Make sure you understand how these distinct tests differ!

Stand-Alone Problems

1) The SAT exam is scaled to have a standard deviation of 110. So this is one of the few examples where we know (mostly) the population standard deviation. The average SAT score in 2017 was 1083. Suppose we want to test whether the average Cal Poly CAFES SAT score is higher than 1083. Based on a sample of 2,268 new freshmen we have a sample mean of 1315. Run this hypothesis test using an alpha of .05. **Make sure to go through all 6 steps and highlight your calculated answers**

2) Now run the hypothesis test that the average CAFES SAT score is equal to 1300, using an alpha of .03. We still have a sample mean of 1315 and population standard deviation of 110.

For the following questions you will rely on the Excel file posted on PolyLearn "HW1 Data."
We will be using this dataset throughout the rest of the quarter. For every county, it includes agricultural and demographic data, as well as the number of farmers' markets.

3)
   a. Test whether the average county size (in acres) is less than 600,000 acres, at a 10% level.
   b. Now run the same test without outliers.
   c. Did your conclusion change? If so, how? What does this suggest about the importance/impact of outliers in this context? Why?

4)
   a. Test whether the average number of farmed acres in a county is greater than 300,000 acres, at a 3% level.
   b. Now run the same test without outliers.
   c. Did your conclusion change? If so, how? What does this suggest about the importance/impact of outliers in this context?

5)

a. A trend in U.S. land use has been a transition of land from agricultural to non-agricultural purposes. What percentage of U.S. counties are farmland (on average)?
b. Test whether the average percentage of farmland is less than 40%, using a level of significance of .05.
c. Test whether the proportion of counties that are more than 50% farmland is greater than .5. Use an alpha of .05.

6) The average number of farmers' markets in our sample is 2.6.

a. What proportion of our counties have at least 3 farmers' markets?
b. Why is the proportion of counties that have an above average number of farmers' markets not equal to 50%? Give an explanation other than rounding up from 2.6 to 3.
c. Test whether the proportion of counties that have at least 3 farmers' markets is less than .5. Use an alpha of .03.

HW2
ABG 327: Data Analysis
Spring 2018

For this assignment, you may work in pairs. The submission will be **one** Excel file to PolyLearn, saved as "LastName_FirstName_HW2. Include your name on the first line of the workbook; if you work with partners both names must be on the first two lines of the workbook. Only one person needs to submit the assignment for each pair. Formatting counts. If it is hard to find the correct answer(s) you will be marked wrong. Make sure all your responses are well-labelled and to use text boxes for longer answers. **Your hypothesis tests should include all 6 steps and written conclusion. You should highlight your calculated answers.**
Never accept population variances as known unless told otherwise. Always treat variances as sample variances.

Many hypothesis tests concern understanding how values differ between groups of people. The following problems will give you experience in performing two-population hypothesis tests—make sure you understand how to choose the correct derivation for a given question and dataset.

Stand-Alone Problems

1) Suppose you work in a medical laboratory. You would like to know the impact of sugar on weight gain. You get 100 rats and break them into two equal groups: those that get soda during each feeding and those that just get water. At the end of the study these are your results:

	Rats w/ Soda	Rats w/o Soda
Avg. Weight Gain (lbs)	5.2	2.5
Sample Std. Dev	0.8	1

Though you do not know the population standard deviation, you can assume the population variances are equal. Test that weight gain for soda rats was greater than the weight gain of non-soda rats, using an alpha of .05. Make population 1 the first population in the claim (soda rats).

2) Career change! After joining PETA you've decided to concentrate on agricultural experiments. You are testing the impact of a new fertilizer on tomato crop yields. Half of your 70 tomato plants gets the new fertilizer and the other half get the current fertilizer. At the end of your study period these are your yields:

	Tom w/ new fertilizer	Tom w/ old fertilizers
Avg. Production (lbs)	24	16
Sample Std. Dev	2.6	3.1

Though you do not know the population standard deviation, you can assume the population variances are equal. Test whether the type of fertilizer impacted yield using a significance level of .03. Make population 1 the tomatoes that received the new fertilizer.

3) You are now writing a report on how the lime market has changed in the last half of the 1990's. For 10 states, you know how the number of pounds per capita changed between 1995 and 2000.

	State 1	State 2	State 3	State 4	State 5	State 6	State 7	State 8	State 9	State 10
Lbs pc in 1995	2	2.4	3.8	1.6	2.6	3.1	2.9	3.2	2.3	1.9
Lbs pc in 2000	4.3	3.6	4.8	3.4	2.4	3.7	4.6	5.4	4.5	4.2

At a 10% significance level, test whether there has been an increase in the number of pounds per capita of limes produced.

4) In a time you probably don't remember, the average consumer was easily able to recycle their cans and plastic bottles because every grocery store had a recycling machine. I was able to put all my bottles and cans in, they were crunched up, and I got cash in return! You would like to assess what impact the removal of those machines might have had on recycling rates. Due to data constraints you are again only able to collect information for 10 states; for each state you know how many pounds of consumer cans and bottles were recycled.

	State 1	State 2	State 3	State 4	State 5	State 6	State 7	State 8	State 9	State 10
Lbs in 1995	580	369	798	345	842	1080	678	546	984	327
Lbs in 2010	475	380	567	124	628	997	590	324	768	221

a. At a 5% significance level, test whether the amount recycled changed over time.
b. What is the problem with using this data to make conclusions about the impact of machine removal on amount recycled?

For the following questions you will rely on the Excel file posted on PolyLearn "HW2 Data." We will be using this dataset throughout the rest of the quarter. For every county, it includes agricultural and demographic data, as well as the number of farmers' markets.

5) Let's begin by looking at county size.
    a. What are the top 10 counties by population size? Create a table for those 10 with the rank, population value, County Fips and state name. *Hint: Formulas like index and match could be helpful.* What states appear at least twice?
    b. What are the top 10 counties by area (ie # of acres)? Create a table for those 10 with the rank, population value, County Fips and state name. What states appear at least twice?
    c. Are the same counties in both tables? Why might we see this result?

6) We will now look at the relationship between the number of farmers' markets and county size.
    a. What is the mean population? What population is the 75^{th} percentile? Explain why they are different and what is surprising about the result.
    b. Define a "large county" as any county that has at least 66,700 people. Perform the hypothesis test, at a 3% level, that large and small counties have the same number of farmers' markets.
    c. Now perform the hypothesis test, at a 3% level, that large counties have more farmers' markets than small counties.
    d. Now define a "large county" as any county that has at least 588,198 acres. Perform the hypothesis test, at a 3% level, that large counties have more farmers' markets than small counties.

7) For this question, we will further explore county size. For our purposes, we define a county as being "primarily dedicated to Ag" if at least 51% of the county is Ag Land.
    a. At a 5% level, test whether the proportion of counties that are dedicated to Ag is higher in low population counties than high population counties.
    b. At a 5% level, test whether the proportion of counties that are dedicated to Ag is higher in high area counties than lower area counties.

HW3
ABG 327: Data Analysis
Spring 2018

For this assignment, you may work in pairs. The submission will be **one** Excel file to PolyLearn, saved as "LastName_FirstName_HW3. Include your name on the first line of the workbook; if you work with partners both names must be on the first two lines of the workbook. Only one person needs to submit the assignment for each pair. Formatting counts. If it is hard to find the correct answer(s) you will be marked wrong. Make sure all your responses are well-labelled and to use text boxes for longer answers. **Your hypothesis tests should include all 6 steps and written conclusion. You should highlight your calculated answers.**
Never accept population variances as known unless told otherwise. Always treat variances as sample variances.

For the following questions you will rely on the Excel file posted on PolyLearn "HW3 Data." We will be using this dataset throughout the rest of the quarter. For every county, it includes agricultural and demographic data, as well as the number of farmers' markets.

1) When breaking data down into categories it is important to justify your decisions. Based on U.S. Census Bureau reports it would appear that one metric they study is number of counties with at least one million people. So we are going to break our counties into 3 populations: those with 0-499,999 residents, those with 500,000-999,999 and those with 1,000,000+. For these three populations,
    a. Test whether median household income differs by county size, at a 5% level.
    b. Test whether mean household income differs by county size, at a 5% level.
    c. How do these test results differ?
    d. Test whether the number of farmers' markets differs by county size, at a 5% level.
    e. Why might we see the results in part d?

2) Now we will look at the relationship between income and the number of farmers' markets. Based on U.S. census categories, split the counties into three separate populations by median income: No more than $44,999, $45,000-$69,999 and at least $70,000.
    a. Test whether the number of farmers' markets differs by county income, at a 5% level.
    b. Explain why we might see these results.
    c. Now test whether the number of farmers' markets differs by county income, using state as a blocking factor.
    d. Test if the blocking factor in part b was effective.
    e. Explain why we used a block in part b. In general, what is the intended effect of blocks in an ANOVA setting?

3) Thus far we have analyzed the relationship between county size, income and farmers markets/food access. We will now look at state and regional level information. The "Variables" tab states what states are in four regions: Northeast, Midwest, South, West.
    a. Test whether the number of farmers' markets differ by region, at a 1% level.
    b. Using the "Poverty Rates" data, test whether the U.S. poverty rate differs by region, at a 1% level.
4) Give a brief synopsis of what you have learned about the distribution of farmers' markets and poverty in the U.S.

HW4
ABG 327: Data Analysis
Spring 2018

For this assignment, you may work in pairs. The submission will be **one** Excel file to PolyLearn, saved as "LastName_FirstName_HW4. Include your name on the first line of the workbook; if you work with partners both names must be on the first two lines of the workbook. Only one person needs to submit the assignment for each pair. Formatting counts. If it is hard to find the correct answer(s) you will be marked wrong. Make sure all your responses are well-labelled and to use text boxes for longer answers. **Your hypothesis tests should include all 6 steps and written conclusion. You should highlight your calculated answers.**
Never accept population variances as known unless told otherwise. Always treat variances as sample variances.

1. You have finally realized your dream of opening an ice cream store, Udderly Delicious. You want to determine if the proportion of households that go to ice cream shops depends on the number of children. Suppose the following data were collected from 515 randomly selected households:

		Visited ice cream shop in past six months	
Number of children		*Yes*	*No*
	0	20	112
	1	45	86
	2	44	82
	3+	48	78

   a. Using α=0.01, perform a chi-sq test to determine if the proportion of households that visit ice cream shops differ by number of children.
   b. Explain where we see the largest differences.

2. Let's suppose you now want to determine whether consumers of different ages have different preferences for ice cream flavors. Over a given week you keep track of ice cream purchases and make the following table:

	Age Range		
Ice Cream Flavor	*Under 18*	*18-54*	*55+*
Vanilla	10	7	14
Mint Chocolate Chip	8	18	8
Chocolate	24	13	8

   a. Using α=0.05, perform a chi-sq test to determine if preferred ice cream flavor differs by age.
   b. How would you use these results when it comes to marketing?

For these problems, use the data on PolyLearn.

3. Create a well-labelled table of summary statistics (including mean, median, std. dev, Coefficient of Variation, min, and max) for # farmers' markets, county size, # of farms, average farm size, population and median income. Note: You will have to create some of these variables.
    a. Using these numbers, explain the benefit of using Coefficient of Variation rather than Std. Dev. to compare variability.
    b. Additionally, break down the mean number of farmers' markets by region. Analyze these results
4. Develop a correlation matrix for these continuous variables. Explain all the farmers' market correlations (both what the values mean and why we may see those relationships).
5. The Northeast and West had the highest average number of farmers' markets. Using α=0.05, test whether there is a difference in variability of income between Northeastern and Western counties. Explain and interpret these results.
6. For this next question you will have to create a table with market frequency categories across the top and population frequency categories down the side. Split the counties into those that have 0 farmers' markets, 1-3 markets and greater than 3 markets. This is your first categorical variable. Also split the counties into those that have a population of 0-99,999, those that have a population of 100,000-499,999 and those that have a population of 500,000 or greater.

	# of Markets		
Pop	*0 Markets*	*1-3 Markets*	*> 3 Markets*
0-99,999			
100,000-499,999			
500,000+			

   At a 3% level, test whether farmers' markets and population are independent. Interpret this result. Where do you see the largest deviations?

7. Now, test whether the proportion of counties that have more than 1 farmers' market differs by region. Interpret these results. How do they compare to your summary statistics?

For questions 8-12 you will look at the impact of median income on farmers markets.

8. Which is the independent variable and which is the dependent variable? Why?
9. Create a well-labelled scatter plot. Now create a second scatter plot omitting the two counties with more than 60 farmers' markets (hint: you only need to change the axes). Add a trendline and trendline equation. Explain what the trendline means.
10. Calculate the Total Sum of Squares, Regression Sum of Squares and Error Sum of Squares. Calculate R^2 and interpret that value.
11. Test for the significance of the coefficient of determination, using α=0.05.
12. Test for the significance of the slope of the regression equation, using α=0.05.

For questions 13-16 you need to create two new variables: population in 1000's and % of county that is farming/Ag Land (0-100). You will use the Data Analysis regression method.

13. What is the impact of % farming on number of farmers' markets? Additionally explain the output of "R Square", "F", "Significance F", "t-stat" for % farming and "P-value" for % farming.
14. What is the impact of total population (not yet in 1000s) on number of farmers' markets?
15. Which model is a better predictor of farmers' markets? Why?
16. Now run the regression of total population (in 1000s) on number of farmers' markets. What changes do you see in the coefficients from question 14? Why do you think this is? Which is a better model to report in a paper? Why?

HW5
ABG 327: Data Analysis
Spring 2018

For this assignment, you may work in pairs. The submission will be **one** Excel file to PolyLearn, saved as "LastName_FirstName_HW5. Include your name on the first line of the workbook; if you work with partners both names must be on the first two lines of the workbook. Only one person needs to submit the assignment for each pair. Formatting counts. If it is hard to find the correct answer(s) you will be marked wrong. Make sure all your responses are well-labelled and to use text boxes for longer answers. **Your hypothesis tests should include all 6 steps and written conclusion. You should highlight your calculated answers.**

For this homework we are trying to learn what impacts the number of farmers' markets in a county. These are the sorts of questions asked by policymakers that are trying to increase food access or improve the local food environment. *Note: All percentages should be from 0 to 100 so that a one unit increase is a one percentage point increase.*

1. We begin with a model that looks at how general county demographic variables relate to the number of farmers' markets: # Farmers' Markets=f(# of Farming Operations, Total Population, Median Age, # of Households, Median Household Income).
    a. Create a correlation matrix of these variables and report in a new tab labeled Answers. Look at the correlations with number of farmers' markets. Do they all have signs that make sense to you, why or why not? Explain using one sentence for each variable.
    b. Based on your correlations, which variable do you expect to have the largest impact on the # of Farmers' Markets? Why (use both statistics and logic)?
    c. Run the regression model. In the answers tab, create a table with the Adjusted R^2, F-test p-value, coefficients, critical values, test statistics, p-values and your verdict of significant or insignificant (at the .05 level). Use proper formatting.
    d. Explain the significant coefficients.

2. We now want to get a more specific, and further refine our demographic explanatory variables. Our model is now # Farmers' Markets=f(# of Farming Operations, Total Population, Median Age, # of Households, % of residents that are children (under 18), % of residents that are elderly (older than 65), % of residents that are female, % of residents that are white).
    a. Run the regression model. In the answers tab, create a table with the Adjusted R^2, F-test p-value, n, coefficients, critical values, test statistics, p-values and your verdict of significant or insignificant (at the .05 level). Use proper formatting.
    b. Explain the coefficients on children and the elderly. Why might we see these results?
    c. Now run the same model, with the addition of "median household income". Did any variables change your conclusion of significant (at a 5% level)? Explain why that might be, using both statistical reasoning as well as theory (logic).
    d. Why does Excel report the p-value of the F-test as 0? How should you report it in a paper? Why?

3. For our final model we will look at the effect of poverty on food access. We will be looking at the model # Farmers' Markets=f(% of land that is dedicated to farming, % of residents that are children (under 18), % of residents that are elderly (older than 65), % of residents that are female, % of the population that has at least a Bachelor's degree, % of population that does not have access to grocery stores, median household income).
    a. Why do we use percentage variables instead of the counts?
    b. Run the regression model. In the answers tab, create a table with the Adjusted R^2, F-test p-value, n, coefficients, critical values, test statistics, p-values and your verdict of significant or insignificant (at the .05 level).
    c. Explain the coefficients on education and low food access.
    d. Which is a better measure of wealth in a county, when it comes to the decision to open a farmers' market: median or mean income? Why? Use both statistics and theory (logic) to explain your response.

4. We now need to make a decision on what we are going to report!
    a. Which was a better model: 1c, 2c or 3b? Justify your response. Why do you think this is?
    b. Summarize what you have learned about the variation in the number of farmers' markets in each county.