I. Differences Between Two Groups

For this problem set you will collect some data and test some hypotheses. The general question you will be asking is "is the mean quantity of A equal to the mean quantity of B?" You will gather data on A and data on B and make a comparison. Make sure you have at least 25 observations on quantity A and at least 25 observations on quantity B. You may choose any of the following questions, or propose one of your own after discussing it with me. Please come see me if you have any questions. You are welcome to work on this with another student in the class.

a. Do men study more than women?
b. Do women exercise more than men?
c. Are female seniors taller than first year female students?
d. Are male seniors heavier than male first year students?
e. Do expensive cookies taste better than similar looking non-expensive cookies?
f. Is cereal at Vons more expensive than cereal at Stater Bros?
g. Do barber shops and beauty salons charge more for women's haircuts than men's haircuts?
h. Does Coke taste better than Pepsi?
i. Is the daily high temperature greater in November than in February?
j. Is the unemployment rate usually higher during the summer than during the fall?
k. Do stock prices go up more on Mondays than on Fridays?
l. Do poor countries grow more quickly than rich countries?
m. Do seniors have a higher GPA than first year students?
n. Do Pitzer students study more than Scripps students?
o. Are first born children more conservative than later born children?
p. Do textbooks for humanities courses cost more than textbooks for social science courses?
q. Are Pitzer students more liberal than Pomona students?

Answer the following questions. Use Excel for 1-3. The data you have collected should be presented clearly in an appendix.

1. What is the sample mean of A and the sample mean of B?
2. What is the sample variance of A and the sample variance of B?
3. What is the sample standard deviation of A and the sample standard deviation of B?
4. Test the hypothesis that the mean of A is equal to some value of interest to you.
5. Test the hypothesis that the variance of B is equal to some value of interest to you.
6. Test the hypothesis that the variance of A is equal to the variance of B at the 10% level.
7. Test the hypothesis that the mean of A is equal to the mean of B at the 5% level. State your assumptions explicitly.

Please hand in this section by Wednesday 18 April
II. Simple Regression
Select one of the following topics below. You will need to collect the data and type it into an Excel spreadsheet. You will run a simple linear regression. You should report the mean and standard deviation of your two variables, their correlation coefficient, the regression intercept, regression slope, and the regression $R^2$. You should also submit a scatterplot of your data and the residuals. You should test the hypothesis that the explanatory variable helps to explain the dependent variable. If you do not like any of these projects, you are welcome to propose your own project after discussing it with me. You are welcome to work on this with another student in the class.

1. Collect data for at least 25 years on the cost of attending Pitzer or any college. Taking into account the increase in the overall price level during these years, use the simple regression model to see whether there has been a trend in the real cost of attending Pitzer.

2. Find data for at least 25 countries on the rate of population growth and the level of per capita gross domestic product (GDP). Make sure that your GDP data are measured in a common currency, such as dollars. Use the simple regression model to see if there is a relationship between population growth and GDP.

3. Go to Zillow.com and look at homes for sale in Claremont (or any other community). Use a random sample of at least 25 houses to estimate a simple regression model with asking price as the dependent variable and square footage as the explanatory variable.

4. Pick a date and approximate time of day (for example, 10am on April 1) for scheduling nonstop flights from LAX to at least 25 large U.S. cities. By searching on the web, determine the cost of an economy seat on each of these flights and the distance covered by each flight. Use your data to estimate a simple regression model with ticket cost as the dependent variable and the distance as the explanatory variable. Are there any outliers?

5. Obtain data on the number of homicides each month for the last 3 years in Los Angeles County. Also gather data on the average high temperature for each of these months. Is there a relationship between homicides and temperature?

6. Go to Huntley Bookstore (or Amazon.com) and randomly select 30 new hardcover textbooks. For each book, record the number of pages and the price. Run a regression to see if there is a relationship between the number of pages and the price of the textbook.

7. For the last two major league baseball seasons, identify those players with at least 400 official times at bat each season. Use the simple regression model to see how well each player’s batting average in 2017 is predicted by his batting average in 2016.

8. An old Wall Street saying is, “As January goes, so goes the year.” Use the simple regression model to see whether the February-through-December percentage change in the Dow Jones Industrial average of stock prices is well predicted by the percentage change in January. Consider the past 30 years.
9. Use the U.S. Department of Energy’s monthly publication *Monthly Energy Review* to obtain U.S. data back to 1950 on total energy consumption (in quadrillion BU). The *Statistical Abstract of the United States*, or the Census homepage at http://www.census.gov, shows the U.S. population over this same period. See if there is a trend in per capita energy consumption. Estimate the equation $y = \alpha + \beta x + \varepsilon$, where $y$ is per capita energy consumption and $x$ is the year 1950, and so on.

10. Go to Edmunds.com and obtain data for at least 25 randomly selected passenger cars on the car’s weight and overall gasoline mileage. Is there a relationship between the two variables?

11. For each of the 50 states, calculate Barack Obama’s percentage of the 2008 election votes cast for the Democratic and Republican presidential candidates; do not include the votes cast for other candidates. Do the same for the 2012 election. Is there a relationship between these two sets of data? Are there any apparent outliers or anomalies?

12. Use the simple regression model to see whether the annual percentage change in the Dow Jones industrial average of stock prices is well predicted by the percentage change the preceding year.

13. Randomly select 30 mutual funds to investigate how well current mutual fund performance is predicted by past performance. Is average performance over the past five years associated with performance in the current year?

14. Randomly select 30 male (female) students and see if there is a relationship between their height and their biological father’s (mother’s) height.

*Due Wednesday 25 April*