Homework 1
Due August 31st at 11:59 pm

Directions: Write out the solution to the three problems in the assignment. Additional practice problems have been provided, but you do not need to turn them in. ***Save your work as a pdf file before submitting it with the Assignment.***

Answer all questions at the 95% confidence level. For the first two problems, you should run the appropriate hypothesis test to answer each question. Your work should include the following steps:

- Explain which test you need to use.
- State your hypotheses.
- Calculate the appropriate test statistic.
- Calculate the p-value that corresponds to the test statistic.
- Interpret the p-value and draw a conclusion from your results. Your conclusion should include not just a statement of statistical significance, but also an indication of the effect size.

1. The file mortgage_payments.xlsx shows two random samples, one of mortgage payments from this year, the other of mortgage payments from five years ago.
   (a) Assuming that both samples were collected from homeowners living in the same house as they were five years ago, what type of test would you run to see if there is a difference between mortgage payments now and five years ago? Run the appropriate test and draw a conclusion from your results.
   (b) What if, instead of being from homeowners in the same house, each sample is a random sample of local homeowners, but there is no connection between homeowners included in this years sample versus the sample from five years ago. How would that change both the test you run to determine if the mortgage payments are different? Does your conclusion change as a result of running a different test?
   (c) What is the estimate of the difference between mortgage payments for both parts (a) and (b)? Why is the margin of error smaller in part (a)?

2. The Society for Human Resource Management (SHRM) collaborated with Globoform on a series of organizational surveys with the goal of identifying challenges that HR leaders face and what strategies help them conquer those challenges. A 2016 survey indicates that employees retention/turnover (46%) and employee engagement (36%) were cited as the most important organizational challenges currently faced by HR professionals. One strategy that may have impact on employee retention, turnover, and engagement is a successful employee recognition program. Surveying small organizations, (with 500 to 2499 employees) and large organizations (with more than 10,000 employees), SHRM and Globoforce showed that 326 of the 423 small organizations and 167 of the 192 large organizations have employee recognition programs.
   (a) Use a hypothesis test to determine if there is a difference between the proportion of small and large organizations that have employee recognition programs.
   (b) What is the best estimate for the difference between the proportion of small and large organizations that have employee recognition programs? Is your estimate consistent with the results of your hypothesis test in part (a)? Explain your answer.

3. A used car dealership wants to build a model for the price of used cars based on the miles on the odometer. The dealership pulls the information on used Toyota Corollas and Honda Civics that they have sold. Use the file used_cars.xlsx to create a simple linear regression model for Corollas and a separate model for Civics and answer the following questions for each model.

   (a) Interpret the variable coefficient in terms of how the independent variable effects the dependent variable.

   (b) How much of the variation in the dependent variable is explained by the model?

   (c) What price would you estimate for a used car that has 15,000 miles on it?

   (d) Would you want to use your regression equation to estimate the price of a used car that has 30,000 miles? Explain your answer.

Additional Problems:

1. A hotel booking company is interested in determining whether two hotel chains offer comparable prices. They take a random sample of the price for similar rooms in 160 different cities. Use the file hotel_costs.xlsx to answer the following questions.

   (a) If the company runs the analysis treating the samples as independent, can they conclude that there is a difference in price between the average cost for a hotel room for the two chains?

   (b) Does your conclusion change if the samples for both hotel rooms in a given city are treated as dependent?

   (c) Is the difference in the price of comparable hotel rooms both statistically and practically significant? Explain your answer.

2. Researchers are interested in investigating whether the proportion of people who use eReaders. is equal to 10%. Randomly selected participants in a study were divided into two age groups. In the 16- to 29-year-old group, 44 of the 628 surveyed use eReaders, while 254 of the 2309 participants 30 years old and older use eReaders.

   (a) Is the percent of either age group that use eReaders consistent with a value of 10%?

   (b) Is there a difference between the proportion of people who use eReaders based on age?

3. A box office analyst seeks to predict opening weekend box office gross for movies. The analyst collects a random sample of the number of YouTube trailer views from the release of the trailer through the Saturday before a movie opens and the opening weekend box office grow (in millions of USD) for 200 movies. The data is in the file movie_gross.xlsx.

   (a) Create a model for the analyst to use that includes all 200 data points. Describe the model, including your choice of dependent and independent variables, and interpret its results. Your discussion should include the following:

      - Evidence that your model is statistically significant,
      - How well your model explains the variation in the dependent variable,
      - Estimates and interpretations for your coefficients, and
      - The regression equation you would use to predict the expected opening weekend box office gross based on YouTube trailer views.

   (b) What is your biggest concern looking at the scatter plot?

   (c) Create a new model that addresses your concerns. How does your new model change your interpretation of the results? Do you feel your concerns were justified?

   (d) What would you expect for the amount of money brought in for its opening weekend by a movie with 10 million YouTube trailer views? Report your answer as both an interval estimate (value ± margin of error) and as a confidence interval (LCL, UCL).

   (e) Should you use your model to predict the opening weekend gross earnings for a movie with only half a million YouTube trailer views? Explain your answer.

4. A book publishing company is interested in keeping track of whether their customers are more inclined to read physical or electronic copies of books. They survey 150 customers, asking them how many physical books and ebooks they have read in the past twelve months. Use the dataset book_type.xlsx to answer the following questions:

(a) What is the estimate for the mean value of the number of physical books and the number of ebooks their customer's read in the past year? Give your answer both as an interval estimate (value ± margin of error) and as a confidence interval (LCL, UCL).

(b) Run a hypothesis test to determine if there is a difference between the number of physical books and the number of ebooks. Your work should include the following steps:

- Explain which test you need to use.
- State your hypotheses.
- Calculate the appropriate test statistic.
- Calculate the p-value that corresponds to the test statistic.
- Interpret the p-value and draw a conclusion from your results.

(c) Do more than two-thirds of the company's customers read twenty or more books a year? What is your best estimate for the minimum value of the proportion of customers who have read twenty or more books in the last twelve months?

5. To help determine the impact of using a cell phone on driving safely, researchers have measured the reaction times of drivers. Half of the sample was tested while talking on their cell phones and the other half was not on a phone. The data is located in the cell_phone2.xlsx file.

   (a) Determine whether there is a difference between the reaction time of drivers based on whether or not they are on their cell phones. Calculate the p-value for your hypothesis test and determine whether to reject the null hypothesis.

   (b) What is your estimate for the difference between the reaction times of drivers talking on their cell phones or not on their phones? Draw a sketch representing your confidence interval.

   (c) Are your answers in parts (a) and (b) consistent? Explain.

6. Three Chinese restaurants, Happy Panda, Jade Dragon, and Lotus Leaf, each claim to be the fastest at delivering meals to customers. You collect a random sample of times it takes for a meal to be delivered from the time the order is place. The data is located in the file chinese_food.xlsx.

   (a) Is there a difference between the delivery times for the restaurants? Use the results of your ANOVA test to explain your conclusion?

   (b) Calculate the Tukey comparisons and determine which pairs of restaurants, if any, have a significant difference in their delivery times.