STAT 500 HOMEWORK #2
FALL 2021

1. Install the `R` package `fishdata` and refer to the data set `juvenile_metrics`. This data set contains measurements on a random sample of *Galaxis maculatus* fish caught from a population of fish in a New Zealand lake.

   a) Plot a histogram of the variable `growth_rate` giving the average daily growth of each fish caught (in mm). Comment on whether the data appears to follow a normal distribution.

   b) Use the function `ks.test` in `R` to test whether the null hypothesis the data follows a normal distribution should be rejected.

   c) Log-transform the data and carry out a one-sample $t-$test of the hypotheses $H_0 : \tilde{X} \geq 1.7$ versus $H_a : \tilde{X} < 1.7$ at $\alpha = 0.05$ where $\tilde{X}$ denotes the true median growth rate. What do you conclude?

2. Refer to the data set `Birthweight_reduced_kg_R.csv` on physical attributes of a random sample of babies born at a certain hospital.

   a) Plot a comparative boxplot of birthweights versus mother's smoking habits (smoker or not).

   b) Carry out a two-sample t-test at $\alpha = 0.10$ of the hypothesis of no difference in mean birthweight between babies born to mothers who smoke versus non-smoking mothers.

   c) Find a 90% confidence interval for the difference in average birthweight.

   d) Draw a conclusion based on your test. Be careful to take into account random sampling and randomization in your answer.

3. Refer to the data set `Cholesterol_R.csv`. This data frame presents data on 18 randomly-sampled individuals diagnosed with high cholesterol who replaced butter in their diets with a brand (A or B) of margarine. The brand of margarine was randomized. Their doctors recorded their blood cholesterol levels at the beginning of the experiment, after four weeks of their diets, and again after eight weeks.

   a) Identify the experimental units, population, treatments, and response.

   b) Plot a comparative boxplot of 8-week cholesterol reduction versus margarine brand.

   c) Test the hypothesis of no difference in mean cholesterol reduction between the two brands of margarine after 8 weeks of use. Use a two-sample t-test at $\alpha = 5\%$.

   d) Compute a 95% confidence interval for the difference in mean cholesterol reduction.

   e) Draw a conclusion based on your test. Be careful to take into account random sampling and randomization in your answer.

4. Refer to the data set `attendance.csv` that lists fan attendance at home (`group = 1`) and away (`group = 0`) games for a local semi-professional baseball team.

   a) Plot a comparative boxplot of attendance versus home/away location.

   b) Plot a qq-plot of standardized attendance for home games against a standard normal distribution. Do the same for away games. What does it indicate?

   c) Plot the autocorrelation function of attendance for home games and also for away games using `acf` in `R`. What does it indicate?

d) Suppose you work for the team and are asked to analyze attendance differences between home and away games. Does a two-sample t-test adequately account for the variation in attendance? Why or why not?

5. Refer to the data set `grades.csv` that displays (fictional) final grades of students in a particular statistics course over two semesters (`class = 0` versus `group = 1`).

   a) Plot a comparative boxplot of grades versus the semester the course was given.

   b) Plot normal qq-plots of grades for each semester. And, also plot histograms of grades for each semester. What do the plots indicate?

   c) Try out a log transformation of grades. Does it improve normality of the grades?

   d) Perform a two-sided rank-sum test for a difference in mean grade over the two semesters at $\alpha = 0.10$.

   e) What hypotheses does the above test evaluate? Explain in the context of the problem. And, provide your conclusion based on the test's approximate p-value.