Stat 204 **Exam 1** **Advent 2020**

1. To determine how the calcium level of water affects respiration rates for fish, 90 fish were randomly divided into three tanks with different levels of calcium (*low*, *medium*, and *high*) and their respiration rates were recorded. The results are in the book's **FishGills3** dataset.

   (a) Use StatKey to create a single histogram of *all* of the respiration rates, i.e., for all 3 groups combined. Describe the distribution, determine its summary statistics, and explain the relationship between its mean and median.

   (b) Using the histogram you just made, estimate the percentage of *all* of the respiration rates that are within 1 standard deviation of the overall average. Within 2 standard deviations? Within 3 standard deviations?

   (c) Use StatKey to create boxplots of the respiration rates for the 3 experimental groups.

   (d) Determine the average, standard deviation, and five-number summary for each group's respiration rates.

   (e) Based on this data, summarize the effect of calcium levels of water on fish respiration rates.

2. The drug finasteride is marketed as Propecia to help protect against male pattern baldness; it may also protect against prostate cancer. A large sample of healthy men over age 55 were randomly assigned to receive either a daily finasteride pill or a placebo. The study lasted 7 years; the men received annual checkups and a biopsy at the end of the study. Prostate cancer was found in 804 of the 4,368 men taking finasteride and in 1,145 of the 4,692 men taking a placebo.

   (a) This study was double-blind. What does that mean?

   (b) Was this an experiment or an observational study?

   (c) Summarize the results in a two-way table and include the row and column totals.

   (d) What percentage of the men in the study received finasteride? What percentage of all the men in the study had prostate cancer? Use correct notation for both proportions.

   (e) Compare the prostate cancer proportions for the two groups using appropriate notation.

   (f) Is there evidence that finasteride protects against prostate cancer? State the relevant null and alternative hypotheses, use StatKey to create a randomization distribution based on this data and the null hypothesis, obtain the p-value, and state your conclusion clearly.

3. The **USStates** dataset includes the variables *HouseholdIncome*, the mean household income for a state (in thousands of dollars), and *College*, the percentage of a state's population age 25 or older who graduated from college. You are going to use this data to investigate the relationship between *College* and *HouseholdIncome*; read through all of the parts of this question to get your bearings before you do anything!

   (a) Use StatKey to create a scatterplot of *College* versus *HouseholdIncome*. Describe the trend shown and determine the correlation between these two variables.

   (b) What is the regression equation for predicting *College* from *HouseholdIncome*?

   (c) What does the slope of this equation tell you? Interpret it in context.

   (d) What does the intercept of this equation tell you? Interpret it in context.

(e) If the average household incomes for two states differ by \$10,000, how do their college graduation percentages differ, on average?

(f) For a state with a mean household income of \$50,000, what is the predicted percentage of adults over 25 who have graduated from college?

4. The book's **Atlanta Commute (distance)** dataset (available directly on StatKey) provides the commute distances in miles for 500 randomly chosen people who work in the Atlanta metropolitan area.

(a) Use StatKey to obtain the mean and standard deviation of the commute distances in this sample.

(b) Use StatKey to create 5000 bootstrap samples and a bootstrap distribution of sample means. What are the mean and standard error of your bootstrap distribution?

(c) Use your bootstrap distribution to determine 95% and 99% confidence intervals for the average commute distance of workers in the Atlanta metropolitan area. Determine the margin of error for each of these confidence intervals. What happens to the margin of error as the confidence level increases?

(d) Based on your bootstrap distribution and your interval estimates, which of the following is *most likely* and which is *least likely*?

   i. the average commute distance for people who work in the Atlanta metropolitan area is less than 20 miles

   ii. the average commute distance for people who work in the Atlanta metropolitan area is around 20 miles

   iii. the average commute distance for people who work in the Atlanta metropolitan area is greater than 20 miles

5. To analyze how well lie detectors perform when subjects are stressed, 48 randomly chosen subjects were connected to a lie detector and asked to read true statements out loud while receiving an electric shock. The lie detector incorrectly reported that 27 of the 48 participants were lying.

Note: For this problem, you will have to enter the counts directly into StatKey. Do not use any of the book's available datasets to answer the following questions!

(a) Use StatKey to determine the best estimate, based on this data, of the proportion of times the lie detector yields false positives, i.e., inaccurately reports deception.

(b) Use StatKey to create 5000 bootstrap samples and a bootstrap distribution of sample proportions. What are the center and standard error of your bootstrap distribution?

(c) Use StatKey and your bootstrap distribution to find a 95% confidence interval for the overall percentage of false positives reported by the lie detector. What is the margin of error?

(d) Does this sample provide evidence that lie detectors give inaccurate results more than half the time when subjects are stressed? State the relevant null and alternative hypotheses, use StatKey to create a randomization distribution based on this sample and the null hypothesis, obtain the p-value, and state your conclusion clearly.