Problem Set 5

Statistics 104

Due April 02, 2020 at 11:59 pm

Problem set policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, be sure to include the code in your solution.*

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Problem 1.

Suppose that a friend concentrating in Environmental Science and Public Policy comes to you for some statistical consulting. He is working on an analysis using data collected at 200 locations across Europe and the continental United States in January of 2018 and January of 2008. Your friend proposes the following approach: for each location, conduct a two-sided hypothesis test at the $\alpha = 0.05$ significance level to compare the mean temperature in January 2018 to the mean temperature in January 2008. He plans to conclude there is evidence of temperature warming for the locations at which mean temperature in January 2018 is significantly higher than mean temperature in January 2008, and specifically present only those significant results to his adviser.

Based on your knowledge of statistics, critique your friend's analysis plan and provide specific advice for addressing any problems you identify. Limit your response to at most ten sentences.

Problem 2.

A medical research group is recruiting people to complete short surveys about their medical history. For example, one survey asks about family history of cancer. So far, people who sign up complete an average of 4 surveys, with standard deviation 2.2. The research group wants to try a new interface that they think may encourage new enrollees to complete more surveys. They plan to randomize each enrollee to either the old or new interface.

a) How many new enrollees do they need for each group (old or new interface) to detect an effect size of 0.5 surveys per enrollee, if the desired power level is 80%? Let $\alpha = 0.05$.

b) Explain the effect of increasing α on the power of the test. What is one disadvantage to increasing α, from a decision-making standpoint?

Problem 3.

Recent research suggests that the use of financial incentives can be effective in promoting weight loss and healthy behaviors. In one study, 104 employees of the Children's Hospital of Philadelphia with BMIs in the 30 - 40 kg/m^2 range who volunteered to participate were randomized to one of three experimental groups:

- Control group: participants were provided information about strategies for losing weight.
- Individual incentive group: participants were provided the same information as those in the control group, but also informed that each time they met or exceeded their monthly target weight loss, they would receive \$100.
- Group incentive group: participants were provided the same information and financial incentive as those in the individual incentive group, but also informed that they had been sorted into anonymous groups of five participants. At the end of each month, participants who met or exceeded their monthly target weight loss would equally split an additional \$500 among their assigned group. For example, a participant who met or exceeded their target loss could gain \$500 at the end of a month if they were the only person who did so in their assigned group, or an additional \$100 if all individuals in the group did so.

The study ran for 6 months. The total change in weight (in pounds) from the beginning of the study to the end of the study was recorded in `weight.csv`. The variable `Change` was calculated as $weight_{end} - weight_{beginning}$.

a) Create a plot that shows the association between total change in weight and experimental group. Describe what you see.

b) Conduct an analysis to determine whether the mean total weight loss varies among the experimental groups.

   i. Assess whether the assumptions for the analysis method are reasonably satisfied.

   ii. Summarize the conclusions and comment on the generalizability of the study results.

c) Suppose the participants in the group incentive group had instead been told that at the end of each month, the participant who lost the most weight (relative to starting weight) within their assigned group of five individuals would receive the \$500.

   i. Speculate as to how the study results might have differed (or not differed) from the obtained data.

   ii. Speculate as to whether it would be advisable to offer the described incentive to study participants.

Problem 4.

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure and is planning a clinical trial to test the drug's effectiveness. Participants are randomized to one of two treatments, either a currently accepted medication or the new drug. At the end of the study, a hypothesis test will be conducted to assess whether there is evidence that the new drug performs better than the standard medication.

In this problem, you will perform simulations under the assumption that the alternative hypothesis is true, and there is a difference in population mean blood pressure between the two groups. For clinical trial data, it is standard practice to test the two-sided alternative.[1]

Suppose the alternative hypothesis $H_0 : \mu_{treatment} \neq \mu_{control}$ is true. Let the mean systolic blood pressure be 140 mm Hg in the control group and 138 mm Hg in the treatment group. Assume that standard deviation in both groups is 10 mm Hg and that blood pressures are normally distributed.

a) Simulate blood pressure values for 25 individuals in the control group and 25 individuals in the treatment group.

    i. Conduct a two-sided test of the null hypothesis from the simulated data. What is the conclusion of the test?

    ii. Does the conclusion from part i. represent an instance of Type II error? Explain your answer.

b) *Power and Sample Size*. Write a simulation that repeats the experiment in part a) 1,000 times.

    i. With 25 individuals in each group, what is the estimated power?

    ii. How does power change with increasing sample size? Estimate the power of the test as n changes to 50, 100, and 200 (leaving all other parameters the same).

c) *Power and Standard Deviation*. For simplicity, these simulations assume that the standard deviation of the treatment and control groups are equal.

    i. Would you expect the probability of rejecting H_0 when H_A is true to increase or decrease if there is more variation in the observations? Explain your answer.

    ii. Using a simulation similar to the one in part b), explore how power changes with increased standard deviation. Estimate the power of the test as standard deviation within each group changes to 5, 10, and 15. Summarize your findings.

d) *Power and Effect Size*. Effect size refers to the difference between the population means, $\mu_{treatment} - \mu_{control}$. In the simulations so far, $\mu_{treatment} - \mu_{control} = 138 - 140 = -2$ mm Hg.

    In a realistic setting, the effect size is chosen to be the incremental value of the intervention that would justify changing current clinical recommendations from an existing intervention to a new one. The simulations so far mimic a setting in which researchers decide they are interested in detecting an effect on blood pressure that is 2 mm Hg or greater, when comparing the new drug to the old drug.

[1] Note that for this setting, there is no actual population of individuals taking the new drug (since the drug is not yet available on the market). Regardless, the observations on the participants assigned to take the new drug are treated as if they are a random sample from a hypothetical population.

i. If the true difference in the group means is relatively large (e.g., 5 mm Hg), as opposed to relatively small (e.g., 1 mm Hg), would you expect the probability of rejecting H_0 when H_A is true to be relatively large or relatively small?

ii. How does power change with effect size? Using a simulation similar to the one in part b), estimate the power of the test as effect size increases; change `treatment.mean` from 138 to 137 and then to 136 (leaving all other parameters the same).

Problem 5.

For this problem, do not use the functions `cor()` or `lm()` (except for checking your calculations); however, you are welcome to use R to make plots and compute values. This is the largest dataset for which you will be asked to estimate a regression line 'by hand'.

Consider the five ordered (x, y) pairs (1, 5), (2, 4), (2.5, 3.5), (3, 0.5), and (4, 0).

a) Plot the data.

b) Calculate the correlation between x and y. Show your work, either in the form of algebraic work or from using R as a calculator.

c) Estimate the slope and y-intercept of the least-squares regression line predicting y from x for these data.

d) Plot the data with the regression line calculated in part c).

Problem 6.

The `utility` dataset contains the average utility bills (in USD) for homes of a particular size and the average monthly outdoor temperature in Fahrenheit.

a) Plot the data. From a visual inspection, does there seem to be a linear relationship between the variables? Why or why not?

b) Fit a linear model to the data. From the linear model, estimate the average utility bill if the average monthly temperature is 120 degrees. Explain whether the estimate is reasonable.

Problem 7.

Suppose that a class of 360 students has just taken an exam. The exam consisted of 40 true-false questions, each of which was worth one point. A diligent teaching fellow has recorded the number of correct answers (Y) and the number of incorrect answers (X) for each student. Suppose that the teaching fellow then fits a linear model to estimate Y from X. *Note: no data are necessary to solve this problem.*

a) What will be the values of b_0, b_1, and R^2?

b) Is this a useful model to fit? Explain your answer.

Problem 8.

The international bank UBS regularly produces a report on prices and earnings in major cities throughout the world. Three of the measures they include are the prices of basic commodities: 1 kg of rice, 1 kg of bread, and the price of a Big Mac hamburger at McDonald's.

An interesting feature of the prices they report is that prices are measured in the minutes of labor required for a "typical" worker in that location to earn enough money to purchase the commodity. Using minutes of labor corrects at least in part for currency fluctuations, prevailing wage rates, and local prices.

The data file `ubsprices.csv` includes measurements for rice, bread, and Big Mac prices from the 2003 and 2009 reports. The year 2003 was before the major recession hit much of the world around 2006, and the year 2009 may reflect changes in price due to the recession.

In this problem, you will model rice prices in 2009 as the dependent variable and rice prices in 2003 as the independent variable.

a) Plot the data and the $y = x$ line.

b) Briefly explain the key difference between points above versus points below the $y = x$ line.

c) Fit a linear model to the data and interpret the slope coefficient. For these data, what is the key difference between a slope larger than 1 versus smaller than 1?

d) From a visual inspection of the plot, identify two potentially influential points and explain your reasoning.

e) Fit a new linear model to the data excluding the two points identified in part d). Based on the results, does it seem these points are influential? Explain your answer.

Problem 9.

A study was conducted on children in cities from the Flanders region in Belgium to assess whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children with access to the supply.

The file `water.Rdata` contains some observations from the study. The fluoride content of the public water supply in each city, measured in parts per million (ppm), is saved as the variable `fluoride`; the number of dental caries per 100 children examined is saved as the variable `caries`. The number of dental caries is calculated by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth at the time of the study.

a) Create a plot that shows the relationship between fluoride content and caries experience. Add the least squares regression line to the scatterplot.

b) Based on the plot from part a), comment on whether the model assumptions of linearity and constant variability seem reasonable for these data.

c) Use a residual plot to assess the model assumptions of linearity and constant variability. Comment on whether the residual plot reveals any information that was not evident from the plot from part b).

The file `water_new.Rdata` contains data from a more recent study conducted across 175 cities in Belgium. Repeat the analyses from parts a) - c) with the new data.

d) Create a plot that shows the relationship between fluoride content and caries experience in the new data. Add the least squares regression line to the scatterplot.

e) Based on the plot from part d), comment on whether the model assumptions of linearity and constant variability seem reasonable for these data.

f) Use a residual plot to assess the model assumptions of linearity and constant variability. Comment on whether the residual plot reveals any information that was not evident from the plot from part e).