Test 2

STAT218

2024-03-16

### Instructions

You have 48 hours from the release of this assignment to complete and submit your work. You may refer to all class materials, notes, and textbooks, but **must complete this assignment on your own**. By submitting your work, you are affirming that your work is your own and you have not consulted with anyone else in preparing your answers. Please use the word document provided (download from the class website) and write in your answers below each prompt. Please ensure that question numbering is preserved in your document. Submit your work via Gradescope.

The test comprises two parts: concepts, containing short-answer and multiple-choice questions; and applications, which requires some data analysis.

## Part I: Concepts

1. [L10] Suppose you use a simple linear regression model to predict summer high temperatures on city blocks from percent tree cover. For a block with 6% tree cover, you obtain two interval estimates for the predicted temperature. Which one is the prediction interval, and which one is the confidence interval, and why?
   1. (89.7, 92.2)
   2. (78.5, 100.6)
2. [L10] Continuing the example above, suppose your estimate for the slope parameter is -0.7. Select the correct interpretation of this parameter.
   1. Every 1% increase in tree cover is associated with an estimated 0.7-degree decrease in mean summer high temperature.
   2. Every doubling of percent tree cover is associated with an estimated 0.7-degree decrease in mean summer high temperature.
   3. Every 1% increase in tree cover is associated with an estimated 0.7-degree increase in mean summer high temperature.
   4. Every 1% increase in tree cover is associated with a 0.7-degree decrease in mean summer high temperature.
3. [L8] Suppose you are interested in studying whether dairy consumption is linked to a rare disease, so you sample 40 individuals with the disease and 50 individuals without the disease, and determine whether each individual has a high or low level of dairy consumption. Which measure(s) of association are appropriate for analyzing the association between dairy consumption and disease risk?
   1. Difference in probability of disease
   2. Relative risk of disease
   3. Odds ratio
   4. All of the above
4. [L6] A study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black () was calculated to be (0.02, 0.06). Select the proper interpretation of this interval.
   1. With 95% confidence, it is estimated that between 2% and 6% more men prefer the color black compared with women.
   2. With 95% confidence, the percentage of men whose favorite color is black is estimated to be between 2 and 6 percentage points higher than the corresponding percentage of women.
   3. With 95% confidence, the probability that a man’s favorite color is black is estimated to be between 2% and 6% higher than the probability that a woman’s favorite color is black.
   4. With 95% confidence, men are between 2% and 6% more likely to prefer black than women.
5. [L6] It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher conducts a study to estimate the proportion of acetaminophen users who have liver damage, and estimates that the proportion is between 0.05 and 0.15 with 95% confidence. What was the sample size?
   1. 54
   2. 144
   3. 20
   4. 87
6. [L7] Three treatments were compared to test their relative efficacy (effectiveness) in treating Type 2 Diabetes in patients aged 10-17 who were being treated with metformin. The primary outcome was lack of glycemic control (or not); lacking glycemic control means the patient still needed insulin, which is not the preferred outcome for a patient. The outcome was considered a ‘success’ if glycemic control was achieved. The chi-square test of association produced a -value of 0.01687. Interpret this result in context.
7. [L7] Continuing the previous problem, the residuals from the chi-square test are shown below. Identify which treatment(s) performed *better* than would be expected had there been no effect of treatment.

Chi-square test residuals

|  | failure | success |
| --- | --- | --- |
| **lifestyle** | 0.2139 | -0.196 |
| **met** | 1.373 | -1.258 |
| **rosi** | -1.584 | 1.451 |

## Part II: Applications

1. [L10] The dataset Sleuth3::ex1220 contains species counts for 30 islands in the Galapagos, along with information about the island geography. Estimate the relationship between island area and the number of observed species. You should include a summary of your fitted model, diagnostic plots, and a visualization of the model. Provide a 95% confidence interval for the model parameter that captures the relationship and interpret the interval in context.
2. [L10] The dataset Sleuth3::ex0817 contains measurements from a study on the effectiveness of biological control of the weed tansy ragwort using flea beetles. Each observation is made on a plot that was planted with high density of ragwort; flea beetles were introduced at different levels on each plot and the average dry mass after harvest in the following season was measured. The objective is to determine whether dry mass decreases for plots with a higher flea beetle load.
   1. Construct a scatterplot of the data. Is the relationship linear?
   2. Construct a scatterplot of the log of the dry mass against flea beetle load. Is this relationship linear?
   3. Fit an appropriate SLR model to estimate the relationship between beetle load and dry mass of ragwort weed among the experimental plots.
   4. Show and comment on the diagnostic plots for the model you fit.
   5. Interpret the significance test on the slope parameter.
   6. Compute and interpret a 95% confidence interval for the slope parameter.
3. [L6] Suppose that among a random sample of 144 U.S. adults 65 and older, 16 have diagnosed Alzheimer’s disease. Calculate and interpret a 95% confidence interval for the prevalence of Alzheimer’s based on the sample.
4. [L7, L8] The dataset openintro::migraine contains observations from an experiment evaluating the effectiveness of acupuncture, relative to sham acupuncture, in treating migraines.
   1. Test for a treatment effect using an appropriate method. (*Hint*: if there’s no effect, we’d observe no association between treatment groups and pain outcomes.)
   2. Estimate the effecct of the acupuncture treatment relative to the control group. Choose an appropriate measure of association and provide a point estimate and 95% confidence interval.
5. [L8] An observational study of 227,571 Medicare beneficiaries who initiated treatment with one of two diabetes medications recorded whether each patient reported cardiovascular problems. Determine which medication is linked to a higher probability of cardiovascular problems, and provide a point and interval estimate for an appropriate measure of comparison.
6. [L7] The dataset openintro::census contains a random subsample of 500 respondents in the 2000 U.S. Census. Construct a contingency table of marital status by sex, and test for an association. Interpret the results of your test.
7. [L6] A random sample of 1629 U.S. adults obtained from NHANES data was used to investigate whether asthma is more common in women than in men. Estimate the difference in asthma prevalence between women and men; provide both a point estimate and an interval estimate.
8. [L10] The dataset openintro::babies\_crawl contains observations of average crawling ages for babies born in each month of the year, along with the average outdoor temperature at six months of age. Fit a simple linear regression model to the data and respond to the following prompts.
   1. How much variation in average crawling age is explained by average outdoor temperature at 6 months old?
   2. Estimate the association between average temperature at 6 months old and average crawling age.
   3. Predict the average crawling age for babies when the average outdoor temperature at 6 months is 70 degrees.
9. [L8] Suppose a retrospective study is done in which data are collected on men ages 50-54 who died over a 1-month period. Of 185 men who died from CVD, 15 had a diet with high salt intake before they died, while of the 196 men who died from other causes, 6 had a diet with high salt intake. Test for an association between salt intake and CVD mortality, and use an appropriate measure of association to estimate the relationship; provide both point and interval estimates.