Test 4

Categorical data analysis [L6, L7, L8]; regression [L10]

STAT218

2024-06-07

## 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 or generated your answers or analyses using AI. Failure to adhere to this expectation will be considered an act of academic dishonesty and result in loss of credit.

You will find a project with a mostly empty script in the class Posit cloud workspace; use this to complete your analyses where required. Note that not all parts require you to perform any calculations; some questions are purely qualitative. Use the prompts as your guide, not the script.

Once you have completed your analyses for the portions requiring use of statistical software, submit your work by filling out the test 2 form posted on the course website. The form will automatically save your work, so you can return to it over the course of the 48-hour test window.

The form will stop accepting responses at the deadline, so **make sure you submit by 5pm on Friday 6/7**. Lastly, keep in mind that you will be given the opportunity to revise problems that you miss the first time around to earn back credit.

## Problems

1. [L3, L6, L7, L8] In a study examining the association between green tea consumption and esophageal carcinoma, researchers recruited 300 patients with carcinoma and 571 without carcinoma and administered a questionnaire about tea drinking habits. Out of the 47 individuals who reported that they regularly drink green tea, 17 had carcinoma. Out of the 824 individuals who reported they never drink green tea, 283 had carcinoma. The greentea dataset contains the participant-level observations.
   1. [L3] Construct a contingency table of tea consumption by carcinoma status.
   2. [L6] Estimate the proportion of patients with carcinoma that regularly consume green tea; provide a point estimate and 95% confidence interval and interpret the estimates in context.
   3. [L6] Estimate the proportion of patients without carcinoma that regularly consume green tea; provide a point estimate and 95% confidence interval and interpret the estimates in context.
   4. [L7] Check the assumptions for a test of association. If they hold, perform the test at the 5% significance level.
   5. [L8] Compute a point estimate for the appropriate measure of association. What does the estimate suggest about the direction of association?
2. [L3, L6, L7] A 2010 Pew Research poll asked 1,306 Americans, “From what you’ve read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?” The warming dataset contains the observations from this poll: each participant’s party or ideology and whether they answered affirmatively or negatively.
   1. [L6] Estimate the proportion of Americans in 2010 who believe there is solid evidence of climate change. Provide both a point estimate and a 95% confidence interval and interpret the estimates in context following the narrative style from class.
   2. [L6] What is the margin of error for this poll?
   3. [L3] Compute the frequency (proportion) of each answer by party/ideology and construct a stacked barplot showing the frequencies.
   4. [L7] Test for an association between climate change opinion and party/ideology. Verify assumptions and carry out the test at the 1% significance level. Interpret the result in context following the narrative style from class.
   5. [L7] If the test is significant, conduct a residual analysis to determine which responses were more or less frequent among each party/ideology than would be expected if opinion and ideology were unrelated.
   6. [LX] Determine the expected *proportions* of responses by ideology under the assumption that ideology and opinion are unrelated.
   7. [LX] Using a Bonferroni correction, compute simultaneous 95% confidence intervals for the proportions of Americans in each ideological group in 2010 who believe there is solid evidence of climate change.
3. [L8] The sulphin dataset contains observations from an experiment studying the efficacy of Sulphinpyrazone for treating patients who have had a heart attack.
   1. [L8] At the 10% significance level, test for an effect of treatment and compute an interval estimate for the relative risk of death in the treatment group compered with the control group at the appropriate confidence level. Be sure to check assumptions and use the appropriate test.
   2. [LX] (Extra credit) Provide a point estimate and confidence interval for the efficacy of Sulphinpyrazone with respect to reducing the risk of death.
4. [L3, L10] The galapagos dataset contains observations of (log-transformed) island area () and (log-transformed) total number of observed species for 30 of the Galapagos islands recorded in 1973.
   1. [L3] Construct a scatterplot of the log total number of species against the log area, and compute the correlation between the two. Based on your results, comment on the apparent linearity, direction, and strength of the relationship.
   2. [L10] Estimate the relationship between log-species and log-area using a simple linear regression model; write the fitted model equation and add the fitted line to your plot in (a).
   3. [L10] Following the narrative style from class, report the proprtion of variance explained and significance of the relationship at the 1% level.
   4. [L10] Provide point and interval estimates at the appropriate confidence level for the model parameter of interest.
   5. [LX] (Optional extra credit) Write the model equation on the original (*i.e.*, not log-transformed) scale; re-interpret the interval estimate from part (d) in terms of the power law relationship.

## Extra credit

1. [L10] The Hubble constant is a fundamental cosmological constant that relates a galaxy’s relative distance and velocity as . The value of the constant can be used to estimate the age of the universe in years is obtained via the conversion . The hubble dataset comprises observations of relative velocities (km/sec) and distances (Mpc) for 24 galaxies.
   1. Use the data provided to estimate the reciprocal of the Hubble constant by fitting a regression model with distance as the response and velocity as the explanatory variable. Fit the model *without an intercept* using the model specification formula = <RESPONSE> ~ <EXPLANATORY> - 1.
   2. Obtain a 90% confidence interval for the reciprocal of the Hubble constant .
   3. Multiply the endpoints of your interval by the conversion factor to obtain an interval estimate for the age of the universe in years.
   4. Report the interval in billions of years and interpret your interval in context.