

Statistics 406 Problem Set 10

Due in lab, Tuesday December 11th

1. The survey strata are defined by the variable `SDMVSTRA`. Calculate the sample Pearson correlation coefficient between BMI (`BMXBMI`) and systolic blood pressure (`BPXSY1`) separately for the individuals within each stratum. Then calculate 95% confidence intervals for each correlation coefficient. Briefly describe your findings.
2. Suppose we have a dataset that is partitioned into strata. The intraclass correlation coefficient (ICC) for a quantitative variable Z , where Z_{ij} is the j^{th} subject in stratum i , is

$$1 - \frac{\text{var}(Z_{ij} - \bar{Z}_i)}{\text{var}Z_{ij}},$$

where \bar{Z}_i is the mean of the data in the i^{th} stratum.

- (a) Calculate intra-class correlation coefficients for each of columns 6 through 40 for the NHANES data set, based on the strata defined by `SDMVSTRA`.
 - (b) Use the bootstrap to produce a 95% confidence interval for the intraclass correlation coefficient of `RIDAGEYR`.
 - (c) Use permutation analysis to obtain an empirical null distribution for the intraclass correlation of `ridageyr`. Specifically, randomly permute the stratum labels (but not `ridageyr`) 1000 times, and calculate an ICC for each permuted dataset. Report the mean ICC value, and a permutation p-value for the observed ICC.
3. Consider the set of people with asthma (`MCQ010 = 1`). Generate a matched comparison sample as follows. For each patient with asthma, identify the subset of all people with the same gender (`RIAGENDR`), age (`RIDAGEYR`), and stratum (`SDMVSTRA`), but who do not have asthma. Then generate a match for the subject with asthma by sampling a subject at random from this list.

Next, perform a 2-sample paired z-test between the asthma and non-asthma pairs, for each column of the data set from 9 through 28. Do this with both the raw data, and with log-transformed data.

Report the number of variables (from 9 through 28) that have the opposite direction of change in the raw scale and log-transformed data.

Report the number of variables that have a significant difference in either the log-scale or raw-scale data, but not both.