Biost 517: Applied Biostatistics I
Biost 514: Biostatistics I
Autumn 2019

Homework #2
Due: Monday, October 14, 2019 by 9:00 AM

Written problems: To be submitted as a pdf or MS-Word compatible file via the canvas course website.

> *On this (as all homeworks) R code and unedited R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the R output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

This homework builds on the analyses performed in homework #1 and uses the same data from homework 1 on a subset of information that was collected to examine magnetic resonance imaging (MRI) changes in the brain in a sample of generally healthy elderly subjects in four U.S. communities and the relationship with aging, cardiovascular disease, cerebrovascular disease, and mortality.

The data can be found on the Canvas web page by clicking on the "Files" link and then accessing the "Datasets" folder. The file "mri.txt" contains the data and the documentation is in the file "mri.pdf".

In this homework, we will use the following variables: creatinine level (*crt*), which is a waste product of muscles that is excreted by the kidneys; length of follow-up (*obstime*); and vital status at the end of follow-up (*death*).

Questions:

1. In this question, you will perform statistical analyses for inference on the distribution of creatinine levels across groups defined by vital status at 5 years.
    a. Among subjects who survived at least 5 years, for the creatine level variable, provide the number of valid observations, the number of missing observations, the mean, the standard deviation, the minimum, 25th percentile, median, 50th percentile, 75th percentile, and the maximum.
    b. Among subjects who died within 5 years, for the creatine level variable, provide the number of valid observations, the number of missing observations, the mean, the standard deviation, the minimum, 25th percentile, median, 50th percentile, 75th percentile, and the maximum.
    c. Briefly describe any similarities or differences in the distribution of creatine level across groups defined by vital status at 5 years.

d. What are the point estimate, the estimated standard error of the point estimate, and the 95% confidence interval for mean creatinine level in a population of similar subjects who would survive at least 5 years?

e. What are the point estimate, the estimated standard error of the point estimate, and the 95% confidence interval for the mean creatinine level in a population of similar subjects who would die within 5 years?

f. Compare the confidence intervals obtained in questions 1d and 1e. What conclusions can be drawn about differences in mean creatinine levels between the two populations defined by vital status at 5 years? Briefly explain.

g. In question 1f, why is it appropriate to compare the two confidence intervals based on groups defined by a dichotomized time to death variable with censoring according to death within 5 years of study enrollment or death after 5years. Provide descriptive statistics that support your answer.

h. In question 1e, you were asked to provide a 95% confidence interval for mean creatinine level for a population of subjects who would die within 5 years. Now provide a 90% confidence interval and a 99% confidence interval for the same population parameter. Compare the 90%, 95%, and 99% confidence intervals, and briefly explain how the confidence intervals change as the confidence levels increase/decrease.

2. In this question, you will perform statistical analyses for inference on the geometric mean of creatinine levels across groups defined by vital status at 5 years.

   a. What are the point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for the **geometric mean** creatinine level in a population of similar subjects who would survive at least 5 years?

   b. What are the point estimate, the estimated standard error of that point estimate, and the 95% confidence interval for the **geometric mean** creatinine level in a population of similar subjects who would die within 5 years?

3. As indicated in the documentation for this study, “normal” creatinine levels are approximately 0.5 to 1.2 mg/dl. As such, for this question we will define “high” creatinine levels to be greater than 1.2 mg/dl.

   a. Provide evidence for or against a “high” mean creatinine level in a population of similar subjects who would survive at least 5 years. Clearly state the basis for your conclusion

   b. Provide evidence for or against a “high” mean creatinine level in a population of similar subjects who would die within 5 years. Clearly state the basis for your conclusion