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

Homework #5
Due: Monday, November 11, 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.)*

In all problems requesting "statistical analyses" (either descriptive or inferential), you should present both

- ***Methods:** A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE R CODE.*
- ***Inference**: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details on Canvas in the "Supplementary Material" Folder.*

This homework uses the same dataset on a sample of generally healthy elderly subjects from four U.S. communities from the previous four homework assignments. In this homework, we are interested in assessing associations between serum creatinine level and 5-year all-cause mortality by comparing risk and odds across groups. 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".

Questions:

1. For this question, we are interested in assessing if there is any association between serum creatinine level and 5 year all-cause mortality by comparing the odds of dying within 5 years across groups defined by whether the subjects have "high" or "low" creatinine levels, where serum creatinine levels greater than 1.2 are considered to be "high", (i.e., "high" corresponds to creatinine > 1.2 mg/dl and "low" corresponds to creatinine ≤ 1.2 mg/dl).
    a. What is the estimated probability of dying within 5 years for subjects with high creatinine levels? What is the estimated odds of dying within 5 years for subjects with high creatinine levels?

b. What is the estimated probability of dying within 5 years for subjects with low creatinine levels from the sample? What is the estimated odds of dying within 5 years for subjects with low creatinine levels from the sample?

c. Give full inference for an association between 5-year all-cause mortality and serum creatinine levels based on the ratio of the odds from questions 1a and 1b, i.e., the **odds ratio (OR)** of dying within 5 years for subjects with high creatinine levels and subjects' low creatinine levels

d. How do the odds in part 1a and 1b change if the response variable of interest is changed from dying within five years to **surviving at least 5 years?** Explain and provide evidence to support your answer.

e. How does the odds ratio in 1c change when the response variable of interest is changed from dying within five years to **surviving at least 5 years?** Does the statistical evidence for an association between 5-year all-cause mortality and serum creatinine level change when using an **odds ratio for surviving at least 5 years** versus using an **odds ratio for dying within 5 years?** Explain and provide evidence to support your conclusion.

2. For question 1, a prospective association analysis is conducted where differences in the distribution of death within 5 years were compared across the two groups defined by high and low serum creatinine levels. In this question, you will now conduct a **retrospective association analysis** where we will compare the distribution of serum creatinine levels collected at baseline (i.e., at the start of the study) across groups defined by vital status at year 5 of the study. For this retrospective analysis, the response variable of interest is an **indicator variable for having high serum creatinine level**, and the predictor of interest is an **indicator variable for death within 5 years**. (Only provide a formal report of inference when asked to.)

   a. What is the estimated probability of having high serum creatinine for subjects who die within 5 years? What is the estimated odds of having high creatinine level for subjects who die within 5 years?

   b. What is the estimated probability of having high serum creatinine for subjects who survive at least 5 years? What is the estimated odds of having high creatinine level for subjects who survive at least 5 years?

   c. Give full inference for an association between 5-year all-cause mortality and serum creatinine levels based on the ratio of the odds from questions 2a and 2b, i.e., the **odds ratio (OR)** of having high serum creatinine for subjects who die within 5 years and subjects who survive at least 5 years.

   d. Compare the retrospective odds ratio and association results obtained in part 2c to the prospective odds ratio and association results obtained in part 1c. Briefly describe any similarities or differences.

3. Now suppose that we are interested in evaluating an association between 5-year all-cause mortality and serum creatinine level by comparing the **risk of death** (or probability of death) within 5 years across the two groups defined by high and low serum creatinine levels.

    a. Give full inference for an association between 5-year all-cause mortality and serum creatinine level using the **risk difference (RD) of death within 5 years** for subjects with high serum creatinine and subjects with low serum creatinine.

    b. Give full inference for an association between 5-year all-cause mortality and serum creatinine level using the **risk ratio (RR)** of death within 5 years for subjects with high serum creatinine and subjects with low serum creatinine.