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

Homework #3
Due: Monday, October 21, 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 homework #2, and uses the same data 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.

In this homework, all questions relate to associations between death from any cause, smoking behavior, as measured in pack years, creatinine level, and sex.

In this homework, we will use the following variables: smoking history in pack years (*packyrs*); creatinine level (*crt*); sex (*male*); length of follow-up (*obstime*); and vital status at the end of follow-up (*death*).

Questions:

1. For this problem, we are interested in assessing if there is an association between smoking history in pack years and 5-year all-cause mortality.
    a. Provide a histogram displaying the distribution of pack years of smoking. Make sure that the histogram has an appropriate main title and an appropriate label for the x-axis. Is the distribution symmetric, right skewed, or left skewed?
    b. Among subjects who died within 5 years, for pack years, 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. Among subjects who survived at least 5 years, for pack years, 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.
    d. Provide boxplots for pack years of smoking for groups defined by vital status at 5 years. The two boxplots should appear in the same figure. Make sure that the graph has a title and there are appropriable labels on the axes for each of the boxplots. Compare the two

boxplots. Briefly describe similarities and/or differences in the distribution of pack years of smoking for the two groups defined by vital status at 5 years.

e. What are the point estimate and the 95% confidence interval for mean pack years of smoking for a population of similar subjects that survives at least 5 years?

f. What are the point estimate and the 95% confidence interval for mean pack years of smoking for a population of similar subjects that dies within 5 years?

g. What is the point estimate for **difference in mean pack years of smoking** between a population of similar subjects that survives at least 5 years and a population of similar subjects that dies within 5 years? Construct 95% confidence interval the difference in mean pack years for the two populations defined by vital status at 5 years, allowing for unequal variances between the two populations. What conclusions can you reach, if any, about differences in mean pack years of smoking for the two populations defined by vital status a 5 years.

h. Conduct a **t-test that allows for heteroscedasticity** (i.e., the possibility of unequal variances across groups) to assess if the two populations defined by vital status at 5 years have the same mean pack years. Clearly state the null hypothesis and the alternative hypothesis. What is the P value for testing the hypothesis that the two populations have the same mean pack years? What conclusions do you reach about a statistically significant association between mean pack years and 5-year all-cause mortality?

i. Briefly compare the results and inference obtained from the statistical analyses conducted in questions 1g and 1h for there being an association with mean pack years of smoking and 5-year all-cause mortality?

j. Now conduct a **t-test that presumes homoscedasticity** (i.e., equal variances across groups) to assess if the two populations defined by vital status a 5 years have the same mean pack years. What is the P value for testing the null hypothesis? What conclusions do you reach about a statistically significant association between mean pack years and 5-year all-cause mortality?

k. Compare the results for 1h and 1j, where a statistical analyses are conducted using a **t-test that allows for heteroscedasticity and t-test that presumes homoscedasticity,** respectively. What analysis would you have preferred a priori in order to answer the question about an association between smoking history and 5-year all-cause mortality? Briefly describe why.

2. For this problem we are interested in assessing if there is a difference in smoking history by pack years across groups defined by sex.

a. What is the point estimate for the **difference in mean pack years** between a population of similar subjects of elderly females and a population of similar subjects of elderly males? Construct 95% confidence interval for the difference in mean pack years of smoking for elderly females and elderly males, allowing for unequal variances between the two populations. What conclusions can you reach, if any, about mean pack years being the same for elderly females and elderly males.

b. Perform a statistical analysis evaluating an association between smoking history in pack years and sex by comparing mean pack years of smoking for females and males using a **t-test that allows for heteroscedasticity** (i.e., the possibility of unequal variances across groups). Clearly state the null hypothesis and the alternative hypothesis. What is the P value for testing the hypothesis that the population of elderly females and population of

elderly males have the same mean pack years? What conclusions do you reach about a statistically significant association between pack years of smoking and sex?

c. Compare the inference obtained in 2a and 2b for an association between smoking history by pack years and sex.

3. For this problem, we are interested in assessing if there is any association between serum creatinine level and 5-year all-cause mortality.

   a. What is the point estimate for the **difference** in mean creatinine level between a population of similar subjects that survives at least 5 years and a population of similar subjects that dies within 5 years? Construct 95% confidence interval the difference in mean creatinine level for the two populations defined by vital status at 5 years, allowing for unequal variances between the two populations. What conclusions can you reach, if any, about mean creatinine level being the same for the two populations defined by vital status at 5 years.

   b. Perform a statistical analysis evaluating an association between serum creatinine level and 5-year all-cause mortality by comparing mean creatinine levels across groups defined by vital status at 5 years using a **t-test that allows for heteroscedasticity**. Clearly state the null hypothesis and the alternative hypothesis. What is the P value for testing the hypothesis that the two populations have the same mean creatinine level? What conclusions do you reach about a statistically significant association between serum creatinine and 5-year all-cause mortality?

   c. Compare the inference on an association between serum creatinine level and 5-year all-cause mortality obtained in 3a and 3b.