Biostatistical Methods I (BIOS 5710)
Breheny

Assignment 10
Due: Wednesday, December 3

1. It has been hypothesized that allergies result from a lack of early childhood exposure to antigens. If this hypothesis were true, then we would expect allergies to be more common in very hygienic households with low levels of bacteria and other infectious agents. To test this theory, researchers at the University of Colorado sampled the houses of 61 children 9-24 months old and recorded two variables: (1) whether the child tested positive for allergies and (2) the concentration of bacterial endotoxin in the house dust. Their results are available on the course website.

   (a) Plot a histogram of bacterial endotoxin levels. Do the data appear to be skewed to the left, skewed to the right, or not skewed?

   (b) Test whether the observed difference in endotoxin levels between the "sensitive" and "normal" groups could be due to chance using a t-test.

   (c) Carry out a permutation test of the null hypothesis that endotoxin levels are independent of allergen sensitivity, using the absolute difference of means as your test statistic $T(x)$.

   (d) Carry out a permutation test of the null hypothesis that endotoxin levels are independent of allergen sensitivity, using the absolute difference of medians as your test statistic $T(x)$.

   (e) Test for a difference in endotoxin levels between the two groups using a Mann-Whitney/Wilcoxon rank sum test.

   (f) Plot a histogram of the log-transformed bacterial endotoxin levels. Do the data appear to be skewed to the left, skewed to the right, or not skewed?

   (g) Test for a difference in endotoxin levels between the two groups using a t-test applied to the log-transformed endotoxin levels.

   (h) Why don't the tests in (b), (c), (d), (e), and (g) agree with each other? Which test do you feel is best in this situation? Justify your opinion (briefly).

   (i) Construct a 95% confidence interval for the difference in mean endotoxin levels between the two groups.

   (j) Construct a 95% confidence interval for the difference in mean log-endotoxin levels between the two groups.

   (k) Let r denote the ratio of geometric means, comparing endotoxin levels in normal children to endotoxin levels in allergen-sensitive children. Construct a 95% confidence interval for the ratio of how much higher endotoxin levels are in normal children than allergen-sensitive children.

   (l) Use the bootstrap to obtain a 95% confidence interval for the difference in median endotoxin levels between the two groups.

   (m) Do the results of this study lend support to the hypothesis that allergies result from a lack of early childhood exposure to antigens, or do they contradict this hypothesis? Or is the study inconclusive?

(n) Decide on a single parameter that, in your opinion, is the most relevant to the scientific goals of this study and write a short, simple sentence describing what this study found out about that parameter. By "simple sentence", I mean one that uses no technical jargon – any literate English-speaking person should be able to understand your sentence.

2. In class I said that permutation tests have low power for small sample sizes. Let's explore further. Suppose we are measuring a continuous outcome, and obtain values $\{1, 2, 3\}$ in group 1 and values $\{101, 102, 103\}$ in group 2.

   (a) What is the p-value for a t-test comparing the means of those two groups?

   (b) What is the p-value of a permutation test comparing the means of those two groups? Explicitly calculate an exact analytic answer, and show your work (i.e., do not use a computer).

   (c) Is it possible to obtain a significant p-value from a permutation test for a two-group comparison when $n = 3$ in both groups? If not, how large does n have to be before it is possible to obtain $p < 0.05$?

3. In class I said that the Wilcoxon rank sum test was "about 95% as powerful" as the t-test even when the data were normally distributed. Let's investigate the accuracy of this statement with a simulation.

   (a) Suppose we are sampling independently from two populations following normal distributions such that $\mu_1 - \mu_2 = 1$ and $\sigma = 3$. What sample size (per group) do we need to obtain 80% power? Do not round your answer.

   (b) Take your answer from part (a) and divide by 0.95. We will use this as the sample size for the Wilcoxon rank sum test (i.e., the t-test sample size will be 95% of the rank sum test sample size). Carry out a simulation according to the description in (a), but use the Wilcoxon rank sum approach to test for the difference between groups. What is the power of the Wilcoxon rank sum test?

4. (a) Suppose the governor of Iowa proposes to cut the salary of all state employees by \$100 per month. What is the correlation between the pre- and post-cut salaries?

   (b) Suppose instead that he proposes to cut the salary of all state employees by 5%. What is the correlation between the pre- and post-cut salaries?

5. Based on data from the March 1995 Current Population Survey, the correlation between the percentage of a state's residents that are foreign-born and the average income for that state is .52. This suggests that foreign-born individuals tend to have higher incomes than native-born individuals.

   (a) At the individual level, will the association be stronger, weaker, or the same?

   (b) Do you think this is convincing evidence that foreign-born individuals make more money than native-born individuals?

6. Generate a sample such that for $i = 1, 2, \ldots 49$, $X_i \sim \mathrm{N}(0, 1)$ and $Y_i \sim \mathrm{N}(0, 1)$, with X and Y independent. However, $X_{50} = Y_{50} = 20$.

   (a) What is the Pearson correlation between X and Y?

   (b) What is the Spearman (i.e., rank) correlation between X and Y?

   (c) Use $B = 1000$ replicates to calculate a bootstrap confidence interval for the Pearson correlation.

   (d) Use $B = 1000$ replicates to calculate a bootstrap confidence interval for the Spearman correlation.