Introduction to Biostatistics (171:161)
Breheny

Assignment 12
Due: Tuesday, April 29

1. You are conducting a study to investigate whether an intervention designed to promote healthier eating habits really works; your primary outcome of interest is systolic blood pressure.

   (a) Is it a good idea or a bad idea to remove from the data set an individual whose systolic blood pressure is recorded as being 0?

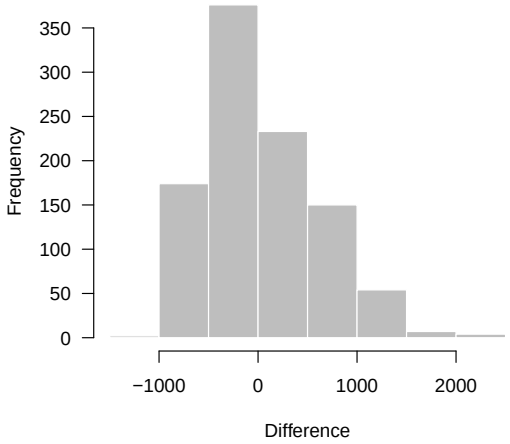
   (b) What about an individual whose systolic blood pressure is 4 standard deviations above the mean?

2. ⌨ 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) I randomly assigned children to the "Sensitive" and "Normal" groups 1,000 times and calculated the difference between the mean Endotoxin levels of the two groups. Below is a histogram of my results.



   What is the p-value associated with this test?

   (d) What is the name for the type of test conducted in part (c)?

(e) Test for a difference in endotoxin levels between the two groups using a Mann-Whitney-Wilcoxon 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 Student's t-test applied to the log-transformed endotoxin levels.

(h) Why don't the tests in (b), (c), (e), and (g) agree with each other?

(i) Which test do you feel is most appropriate in this situation? Why?

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

(k) Construct a 95% confidence interval for the ratio of how much higher endotoxin levels are in normal children than allergen-sensitive children.

(l) Suppose I wanted to construct a nonparametric confidence interval for the difference in median endotoxin levels between the two groups. What statistical technique might I consider using?

(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?

(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. Note: a p-value is **not** a parameter.