MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #3 (Permutation Tests)

Last modified: February 10, 2014

**Reading from Chihara and Hesterberg**

- Chapter 3, sections 3.1 through 3.4
- Appendix A.3 and Appendix B.1
- The F distribution is covered in section B.13 on pages 391-393. Feel free to ignore it if you wish.
- The hypergeometric distribution is covered in Appendix B.5. Don't worry about Theorem B.4.

**Optional Reading from Haigh**

- Look up the Bernoulli, binomial, and hypergeometric distributions in the index if you want to see more examples.

**Proof of the Week**

- Prove that the sum of $n$ independent Bernoulli random variables, each with parameter $p$, is a binomial random variable $Y \sim \text{Binom}(n, p)$, and that

$$E[Y] = np, \text{ Var } Y = np(1-p).$$

**R scripts**

- Script 3A-Permutation Test. This script explains how to carry out a permutation test for hypothesis testing, both in the case where is is possible to look at all possible permutations and in the more common case where it is necessary to sample random permutations.
  - Topic 1 - a tiny permutation test
  - Topic 2 - a permutation test with a larger data set
  - Topic 3 - data from an actual experiment
  - Topic 4 - generating the subsets by random sampling
- Script Script 3B-PermSkewed.R Section 3.3, pages 44-51. This script deals with a data set where classical methods fail because the data are skewed and the subgroups being compared are wildly unequal in size.
  - Topic 1 - dealing with skewed data and unequal sample size
  - Topic 2 - Using other statistics
- Script 3C-PermTestVariance.R Example 3.5 on page 51. This script shows how a permutation test can be used to test for unequal variance instead of for unequal mean. The result, in some cases, can be replicated by using Fisher's F distribution. We will not cover this in any detail, but you might find it useful to know how it arises.
  - Topic 1 - A permutation test for equal variance
  - Topic 2 - Testing real-world data for unequal variance
  - Topic 3 - Testing for equal variance in skewed data
- Script 3D-GSSContingency.R Based on pp. 54-57 of the textbook.
  This script explains how to use permutation tests to explore whether two factors (columns in a data frame) might not be independent. It demonstrates the use of the chi-square statistic, for which the theory will appear much later in the course.
  - Topic 1 - the chi-square statistic
  - Topic 2 - A permutation test for independence
- Script 3E-HyperPerm.R Based on Appendix B.5 This script replicates the results of an exact permutation test by using the hypergeometric distribution and by using the Fisher exact test that is based on it.
- Script 3P-Proof 3.R This script illustrates sampling that is done using the built-in R functions for the Bernoulli and binomial distributions.

**Mathematical notes**

1. Proof of the week
   Prove that the sum of $n$ independent Bernoulli random variables, each with parameter $p$, is a binomial random variable $Y \sim \text{Binom}(n, p)$, and that

$$E[Y] = np, \ \text{Var } Y = np(1-p).$$

   (a) A Bernoulli random variable $X$ has the value 1 with probability $p$, 0 with probability $1 - p$. Calculate its expectation and variance.

   (b) (The easy way) A binomial random variable $Y$ is the sum of $n$ independent Bernoulli random variables: $Y = X_1 + X_2 + \cdots + X_n$. Calculate its expectation and variance from this property alone.

   (c) If $Y$ has the value $r$, then $r$ of the $X_i$ have the value 1, $n - r$ have the value 0. Calculate the probability of this event, which can happen in many ways, and so determine the mass (density) function $P(Y = r)$.

   (d) (The hard way – not required) Calculate $E[Y]$ and $\text{Var}[Y]$ directly from the mass function for the binomial distribution.

2. The hypergeometric distribution and permutation testing

Although the hypergeometric distribution is not mentioned in Chapter 3, it is well explained in Appendix B.5, and it is supported in R by the usual set of four functions. It describes exactly what happens when you do a permutation test with two factors. Using the example of Appendix B.5, suppose that you have a data frame with two columns. Column Gender has $M$ "'Woman" and $N$ "Man. Column Commitee has $n$ "On" and $M+N-n$ "Off." The expected number of women on the committee is of course $\frac{M}{M+N}$.

The adminstration chooses a committee with $W$ women, and you suspect that women are underrepresented. So you permute the second column and make a histogram of the number of women $x$ who end up on the committee to see how frequently the event $x \leq W$ occurs.

There are $\binom{M+N}{n}$ ways to select a committee with $n$ members. There are $\binom{M}{x}$ ways to select $x$ of the $M$ women. There are $\binom{M}{n-x}$ ways to select $n-x$ of the $N$ men.

So if the committee is formed by permuting the second column and choosing the rows where the second column has "On," the probability that there are $x$ women on the committee is

$$P(X = x) = \frac{\binom{M}{x}\binom{M}{n-x}}{\binom{M+N}{n}}.$$

This probability is given by the R function `dhyper(x, M, N, n)`

The contingency table for the two factors looks like this:

| | Women | Men |
|---|---|---|
| On | $x$ | $M+N-x$ |
| Off | $M-x$ | $N-n+x$ |

For a $2 \times 2$ test, the R function `phyper(x, M, N, n)` replicates the result of a perfect permutation test (using all permutations instead of a random sample). The R function `fisher.test()` automates this test. The chi-square test described in the textbook gives a good approximation when none of the expected counts is too small.

3. A contingency table with one row - why does chi-square work?

   Suppose we make $n$ independent Bernoulli trials, each with probability $p$.

   A table of the results will look like this:

| Success | Failure |
|---|---|
| $x$ | $n - x$ |

   A table of the expected results will look like this:

| Success | Failure |
|---|---|
| $np$ | $n - np$ |

   The chi-square distribution with one degree of freedom has an expectation of 1. Show that the random variable that results from summing over both entries in the table also has an expectation of 1.

**Section problems** I have no idea what the answers to the following questions will turn out to be. Perhaps they will give you ideas for projects.

The Red Sox data came from `http://www.baseball-reference.com`.

The Olympic results came from `www.sochi2014.com`.

1. The file Slopestyle.csv (on the Class 3 page) contains scores for the first and second runs of the recent slopestyle snowboarding competition in Sochi. Carry out permutation tests to answer the following questions. Different members of the class can work on different parts of the problem.

   (a) Is there a significant difference in the mean score (including both runs) for men and for women?

   (b) Is there a significant difference in the median best score for men and for women?

   (c) Is there a significant difference in the absolute value of the difference between the scores on the two runs for men and for women?

2. The file RedSox2013.csv (on the Class 3 page) includes results from every Red Sox game in 2013. Carry out permutation tests to answer the following questions. Different members of the class can work on different parts of the problem.

   (a) Conduct a permutation test for independence of WonLost and Away. Compare with the result of the built-in chi-square test.

   (b) Is there a statistically significant difference between the average duration of a day game and of a night game?

   (c) Is there a statistically significant difference between the variance in attendance at home games and away games?

   (d) Was the number of games that the Red Sox won in the postseason (after game 162) unexpectedly high? Confirm the result of the permutation test by using the hypergeometric distriibution.

**Homework assignment** This assignment should be submitted as a single R script. Include enough comments so that it is clear what you are doing and where each problem begins. You can upload it to the dropbox on the Class 3 page of the Web site.

It is OK to paste and edit lines from the scripts on the course Web site. It is not OK to paste lines from your classmates' solutions!

1. Problem 9 on page 70 (permutation test for a numeric variable)

2. Problem 12 on page 71 (permutation and chi square test for a 2 x 2 contingency table)

3. Problem 7 on page 69 (flight delays, including comparison of variances)

4. In the 2013 regular season, the San Diego Chargers played 16 games, winning 9 and losing 7, in the sequence LWLWLWWLLLWLWWWW.

   It is generally felt that they "finished strong" in winning four of their last six games.

   (a) Carry out an exact permutation test to determine the probability that is the wins and losses were scrambled, the last six games would include four or more wins.

   (b) Replicate your answer by using the multinomial distribution.

   (c) Do a binomial approximation to find the probability that if the Chargers play six games with a probability $p = 9/16$ of winning each, they will win four or more of the six games.

   (d) Rerun the permutation test by using 10,000 randomly chosen samples of 6 games.

5. Problem 10a on page 70. First carry out the built-in chi-square test in R (The hard part may be getting the data into the right format), then repeat the test by creatting a data frame with two columns and 286 rows and carrying out a permutation test using the chi square statistic. You should get good but not perfect agreement.