MET3220C
Computational Statistics

Nonparametric Hypothesis Tests: Resampling

(Chapter 5.3 of Wilks' book)

Key Points:
1) Introduction to Resampling
2) Permutation Tests
3) Bootstrap

http://campus.fsu.edu/
bourassa@met.fsu.edu

Introduction to Resampling

- Resampling techniques attempt to improve a description of a data set, by repeatedly subsampling the available data to artificially increase the number of samples.
  - The resampling (repeated subsampling) is done in a manner consistent with the null hypothesis.
  - Other names for this approach are randomization tests, rerandonization tests, and Monte-Carlo techniques.
  - The concept is highly adaptable. These approaches can easily be designed for to meet specific needs.
- The concept is to build a collection of artificial data sets of the same size as the actual data.
  - Then compute the chosen test statistic for each of the artificial data sets.
  - This results in a set of test statistics, with a size equal to the number of artificial data sets.
  - This set of tests statistics is used to make a null distribution.

Introduction to Resampling II

- What do we do with the estimated null distribution?
- The test statistics from the original data can be evaluated through use of the null distribution.
- Note that resampling is ideally suited to computer programming.
  - Given a uniform random number generator, which is built into most languages, it is very easy to program resampling.
- Advantages of resampling:
  - Does not require any assumption of an underlying parametric distribution.
  - Any computable test can be applied. However, some are better than others.

Desirable Characteristics Of Test Statistics

- There are three desirable characteristics for a test statistic (Good 2000).
- **Sufficiency**: All the information about the distribution (or physical phenomena) contained in the data is also reflected in the test statistic. In other words, gathering more data will not add more useful information, OR there is sufficient information.
  - Would this condition invalidate any tests that we have discussed.
- **Invariance**: The test statistic should be invariant to arbitrary transformations.
  - Example: changes in units, e.g., °F to °C.
  - Note this condition assumes that the comparison data (if any) is also transformed in the same manner!
- **Loss**: The consequence for differences from expectations, as used in the test statistic, should be consistent with the problem.
  - Example: Squared differences are more sensitive to large differences. That is not always consistent with the question.

Permutation Tests

- The Wilcoxon-Mann-Whitney test is a special case of a permutation test.
  - These tests are applied to two (or more) sample problems.
  - Think of permutation tests as similar to a Wilcoxon-Mann-Whitney test, but applied to a chosen test statistic, and with 2+ sets of samples (rather than exactly 2).
  - These types of tests rely on the principle of exchangeability.
  - The null hypothesis is that the statistics does not change if data are exchanged.
- Recall that for a two sample case the number of permutations is equal to $n! / [(n_1!)(n_2!)]$. This number can easily be enormous.
- There is usually no need to calculate more than 10,000 values of the test statistic, unless a good description of extremes of the null hypothesis are critical.
- Note that individual resampled data sets are created without replacement, similar to rearranging data in the Wilcoxon-Mann-Whitney test.

Example: Does Cloud Seeding Influence Lightning Strikes?

TABLE 5.6 Illustration of the procedure of the rank-sum test using the cloud-to-ground lightning data in Table 5.5. In the left portion of this table, the $n_1 + n_2 = 23$ counts of lightning strikes are pooled and ranked. In the right portion of the table, the observations are segregated according to their labels of seeded (S) or not seeded (N) and the sums of the ranks for the two categories (R_1 and R_2) are computed.

Pooled Data			Segregated Data			
Strikes	Seeded?	Rank				
0	N	1			N	1
2	S	2	S	2		
4	S	3	S	3		
9	S	4	S	4		
10	N	5.5			N	5.5
10	S	5.5	S	5.5		
12	S	7	S	7		
16	S	8	S	8		
18	S	9	S	9		
20	N	10			N	10
22	S	11	S	11		
26	S	12	S	12		
29	S	13	S	13		
30	N	14			N	14
33	N	15			N	15
34	S	16	S	16		
45	N	17			N	17
49	S	18	S	18		
61	N	19			N	19
62	N	20			N	20
63	N	21			N	21
82	N	22			N	22
358	N	23			N	23
		Sums of Ranks:	$R_1 = 108.5$		$R_2 = 167.5$	

- The Wilcoxon-Mann-Whitney test evenly weights all observations.
- If we think that extremes are more important, we can weight the extreme more heavily.
- Example test statistics:
  - Test = $\sum_i (Rank(x_i) - n/2)^2$
- Example that focuses on values near meal
  - Test = $\sum_i Min(Rank(x_i), n - Rank(x_i))^2$
- Tests do not have to use ranks.

Bootstrap Methods

- Bootstrap methods are used when
  - (1) you have only one sample, and/or
  - (2) when the principle of exchangeability cannot be applied to the null hypothesis.
- Why are these cases more complicated?
- These are the two conditions for which a permutation will not work.
- Bootstrap techniques are relatively newly developed (Efron 1979).
- Bootstrap techniques assume that the distribution from which a sample data set has been drawn is well represented by that data set.
  - The assumptions in a permutation test are always more robust, provided that a permutation can be applied.
  - ONLY USE A BOOTSTRAP METHOD WHEN YOU CAN'T USE A PERMUTATION TEST!

Bootstrap – The Mechanics

- Key difference from permutation methodology is data are sampled with replacement.
  - If you are randomly selecting n values from the data set, you 'return' each number right after it is drawn.
  - This means some numbers will probably be drawn more than once, and some will not be drawn at all.
- The construction of the new sample set (above) is repeats a large number of times n_B.
  - n_B might be 10,000 (example of a typical value).
  - Results in n_B samples of size n.
- The test statistics is computed for each of the n_B samples.
- The distribution of the resulting test statistics is used to approximated the null distribution.