Midterm Project

Due March 1, 2011

The midterm project consist of writing an R function that performs the nonparametric Gehan test, applying the function to two datasets, and creating a useful graphical description of each dataset. The method for the Gehan test is given below along with an example. Followed by guidelines for writing the function and performing the analysis.

Gehan Test

The Gehan test is a nonparametric procedure for comparing the medians of two independent samples that may contain data that is left truncated at different values. We must assume that the truncation mechanism is the same for both populations. However we do not have to make any assumptions about the variance of the population distributions.

This procedure is frequently used with environmental data to determine if levels of a chemical at a site are different then levels that naturally occur in nearby areas. For example, arsenic, a known carcinogen, naturally occurs in soil and is also a byproduct of mining activity. The Gehan test can be used to compare soil samples from a mining site to nearby areas that are not affected by mining.

Furthermore, environmental data is often left truncated. Laboratory machines have a detection limit. If the concentration of a chemical is lower then this limit, the concentration cannot be detected. In which case the data is left truncated at a known detection limit.

Procedure

The following procedure is taken from the 2002 Naval Facilities Engineering Command *Guidance for Environmental Background Analysis, Volume 1: Soil*, available at the Argonne National Laboratory website, http://www.ead.anl.gov/.

Suppose m background samples and n site samples are collected. Site refers to a potentially hazardous site that is being investigate and background refers to nearby areas that reflect naturally occurring levels. If an observation is a non-detect, that is laboratory machines are unable to detect the chemical concentration, then the detection limit of the machine is given along with a less-than sign to denote the truncated observation.

The following procedure is used to test the hypothesis,

$$\begin{aligned} H_0&: \text{ Median of Site} = \text{Median of Background} \\ H_a&: \text{ Median of Site} \geq \text{Median of Background} \end{aligned}$$

1. List the combined m background and n site measurements, including non-detect values, from smallest to largest. The total number of combined samples is $N = m + n$. Use the given detection limit for non-detect data.

2. Determine the N ranks, $R_1, R_2, \ldots, R_N$, for the N ordered data values using the method described in the example below.

3. Compute the N scores, $a(R_1), a(R_2), \ldots, a(R_N)$, where $a(R_i) = 2R_i - N - 1$, for $i = 1, 2, \ldots, N$.

4. Compute the Gehan statistic, G,

$$G = \frac{\sum_{i=1}^{N} h_i a(R_i)}{\left[\frac{mn \sum_{i=1}^{N} [a(R_i)]^2}{N(N-1)}\right]^{1/2}}$$

   where h_i is an indicator, $h_i = 1$ if the ith observation is from the site population and $h_i = 0$ if the ith observation is from the background population.

5. Calculate the p-value. When $m \geq 10$ and $n \geq 10$ calculate the p-value using a large-sample approximation. Otherwise for small m and n calculate the p-value using a permutation test.

   - For large samples, the distribution of G is approximately standard normal. Therefore, reject the null hypothesis if $G \geq Z_{1-\alpha}$, where $Z_{1-\alpha}$ is the $100(1-\alpha)$th percentile of the standard normal distribution.
   - To perform a permutation test for small samples,
     (a) Take a random sample of size n from the pooled data *without replacement*. These n values represents site data and the other m observations are the background data.
     (b) Calculate G for this resample.
     (c) Repeat steps (a) and (b) several thousand times.
     (d) The distribution of the test statistics calculated in step (c) approximates the sampling distribution under the null hypothesis. The permutation p-value is the proportion of resamples that give a result at least as great as the observed G.

Example Below are 10 samples from site and background areas. The < denotes a non-detect observation, data that is left truncated at the detection limit.

Background:	1	<4	5	7	<12	15	18	<21	<25	27
Site:	2	<4	8	17	20	25	34	<35	40	43

The following steps are used to create this table which is then used to calculate G.

Data	h_i	δ_i	d_i	e_i	R_i	$a(R_i)$	Data	h_i	δ_i	d_i	e_i	R_i	$a(R_i)$
1	0	1	1	0	4	-13	18	0	1	8	3	12.5	4
2	1	1	2	0	5	-11	20	1	1	9	3	13.5	6
<4	0	0	2	1	4.5	-12	<21	0	0	9	4	8	-5
<4	1	0	2	2	4.5	-12	<25	0	0	9	5	8	-5
5	0	1	3	2	7	-7	25	1	1	10	5	15.5	10
7	0	1	4	2	8	-5	27	0	1	11	5	16.5	12
8	1	1	5	2	9	-3	34	1	1	12	5	17.5	14
<12	0	0	5	3	6	-9	<35	1	0	12	6	9.5	-2
15	0	1	6	3	10.5	0	40	1	1	13	6	19	17
17	1	1	7	3	11.5	2	43	1	1	14	6	20	19

1. List the combined m background and n site measurements in column 1 of the Table from smallest to largest. Use the given detection limit for non-detect data.

2. Place a 0 or 1 in the second column of the Table, h_i, using the following rule:
   $h_i = 1$ If the ith measurement is from the site
   $h_i = 0$ If the ith measurement is from background

3. Place a 0 or 1 in the third column of the Table 1, δ_i, using the following rule:
   $\delta_i = 1$ If the ith measurement is a detection
   $\delta_i = 0$ If the ith measurement is a non-detect

4. Determine the values of d_i and e_i using these rules:
   - If the first value is a detect, that is, if $\delta_1 = 1$, then set $d_1 = 1$ and $e_1 = 0$.
   - If the first value is a non-detect, that is, if $\delta_1 = 0$, then set $d_1 = 0$ and $e_1 = 1$.
   - For each successive row increase d_i by 1 when $\delta_i = 1,\ i = 2, \ldots, 20$.
   - For each successive row increase e_i by 1 when $\delta_i = 0,\ i = 2, \ldots, 20$.

5. Let T denote the total number of non-detect values in the pooled background and site datasets. For this dataset there are $T = 6$ non-detects. Compute the rank of the ith observation by,
   - $R_i = d_i + (T + e_i)/2$ if $\delta_i = 1$.
   - $R_i = (T + 1 + d_i)/2$ if $\delta_i = 0$.

6. Compute the $N = 20$ scores, $a(R_1), a(R_2), \ldots, a(R_{20})$, where $a(R_i) = 2R_i - N - 1$.

Using the columns of the Table, the Gehan statistic is $G = 1.77$, since $G > 1.645 = Z_{1-.05}$, we reject the null hypothesis at the 0.05 level.

Project Guidelines

1. The goal of the project is to develop a function(s) that will be useful to other statisticians who are interested in using the Gehan test for large and small samples. You will need to decide on the structure and organization of your function(s). For example, you could write one function that performs the large-sample approximation and the permutation test. This function could include an argument that indicates which method to use with a default approach based on the sample size. Or you could write two separate functions one for each approach. You will also need to carefully consider the format and type of arguments that would be most appropriate for general use of the Gehan test. Do whatever you think is best, but keep in mind that you are writing a general program for others to use.

   Requirements
   - Write an `R` function(s) that performs the Gehan test using the large-sample approximation and the permutation test.
   - Include code that checks if the arguments are valid and returns an error message if there is an invalid argument.
   - Return a warning message if the large-sample test is used when $m < 10$ and $n < 10$.
   - Function(s) should return an object that contains the test statistic, the p-value, and the method.
   - Create an S3 class for the object your function returns and write an S3 print method. The print method should output: the name of the test (Gehan), the method used, the test statistic, and the p-value.
   - Comment your code. Including a description of the function(s), the type and format of the arguments, and a description of the values returned. Your comments should act like a manual for how to use the function. Also include comments in the body of the function that describe what is going on.

2. Apply your function(s) to the following two datasets. Use the large-sample approximation for the first dataset and the permutation test for the second dataset.

   Dataset 1

Background:	4	<18	13	27	39	11	<23	<6	<3	9	29	<19	36
Site:	49	10	<17	<28	50	30	32	20	<26	34	37	48	45

   Dataset 2

Background:	18	<10	27	22	<3
Site:	30	44	23	<16	13

   A < denotes non-detect data, in which case the detection-limit is given.

3. Create at least one graph for each dataset that will be useful for comparing site and background data. For non-detect data use the detection-limit.

Please turn-in a hard copy of your `R` code along with a technical report of your results (do not submit raw output). Also submit via e-mail (njc23@pitt.edu) a copy of your `R` code.