CONJOINT

A Mu Sigma University Publication, May 2015

Mu Sigma

In This Issue...

Abstract

Permutation methods are computer-intensive and the number of permutations can be prohibitively large even for modest sample sizes. In such a situation, a satisfactory solution is to take a random sample of permutations and generate a reference distribution of a statistic. The method of randomization will depend on the nature of the data collection design and the hypothesis to be tested. This article describes a few examples of how to go about selecting these random samples of permutations. Read the post on "Permutation Tests" in the March 2015 issue of **Conjoint**.

Randomization Tests

- By T. Krishnan

Random Sample of Permutations

Permutation methods are useful alternatives to parametric and nonparametric methods since they do not rely on any assumptions and use only the data at hand. However, they consume a great deal of computing resources and are well nigh impossible to implement because the number of permutations is very large even for moderate sample sizes. For instance, when you have 20 paired (x, y) samples, the number of permutations is $2^{20} = 1048576$ and when you have a sample on two groups of 10 each, the number of permutations is $\binom{20}{10} = 184756$.

Under these circumstances, a random sample of the permutations of a manageable size will be adequate to generate the required sampling distribution of the statistic of interest. The sampling distribution so generated is often called the **reference distribution**, to distinguish it from the null distribution of the test statistic usually generated from models or theoretical considerations. The generation of a random sample of permutations depends on how the data have been generated and what hypothesis is to be tested.

If all permutations are examined, it is often called a permutation test, or an exact test. If only a sample of permutations is examined, it is called a **randomization test**.

Once you have selected the random sample of n permutations, then you do what you do for permutation tests, namely

1. Choose a statistic relevant for data design and the hypothesis to be tested.
2. For each sample, compute the value of the chosen statistic.
   - Suppose the data consists of paired observations (x, y) for which in the parametric case you might do a paired t-test; here you would compute for each data point $z = x - y$ and carry out a one-sample t-test based on the z values. Suppose the data consists of independent sets of x values for two groups; here you would compute the 2-sample t test.
   - Thus depending on the nature of the data, for each random sample of permutations you might compute the corresponding type of t statistic; or you might decide to simply compute the mean $\bar{z}$ for the paired sample and the difference of the means of the two groups in the 2-sample case.
3. Then the distribution of the set of n statistics z or t is the reference distribution for the hypothesis test.
4. The p-value is simply the proportion of the statistics more extreme than the observed value in the actual sample. This is the same as what we do in a hypothesis test with the sampling distribution of the test statistic.

CONJOINT

A Mu Sigma University Publication, May 2015

Mu Sigma

In This Issue...

Sampling Algorithms

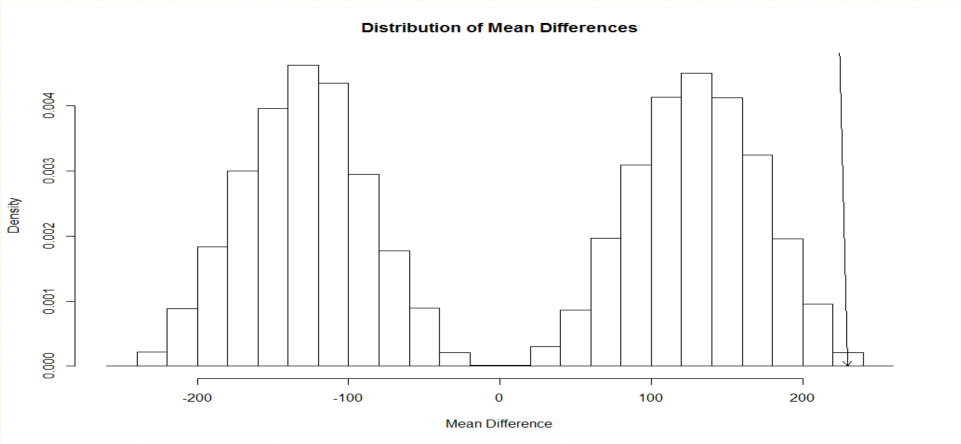
We discuss some examples of how random samples of permutations are selected. Generally, sampling would mean that we have a list often called **the sampling frame** of elements with their IDs; a sample is selected of IDs from this sampling frame and **observations are made on these IDs**. But the re-sampling we are discussing here is of a different kind. Here **observations have already been made**. What we need is a collection of random IDs from the IDs of the possible configurations (2^{20} in our example of paired t-test). This selection will depend on the nature of the configurations. Let us discuss a few examples.

Selection of Re-samples: Paired Data Case

Let us take the example of 20 paired observations where there are 2^{20} configurations from which we need say, $n = 10000$ random samples. Let the 20 observations be $(x_i, y_i), i = 1, 2, \dots, 20$. Let $z_i = x_i - y_i$, and $s_i = |x_i - y_i|$. The 2^{20} configurations yield the 2^{20} possible configurations of values of s_i with a $+$ or a $-$ sign attached to each s_i. Thus a random re-sample is equivalent to selecting a set of n 20-vectors of $\pm$ from the collection of 2^{20} possible 20-vectors of $\pm$. Once a random re-sample is selected, then we need to compute the mean of s_i with $\pm$ signs. If a t statistic is needed, it can be computed from the selected $\pm s_i$ values.

A random sample from the permutations can be selected by selecting 20 $\pm$ signs independently each with a probability of 0.5 for each of $+$ and $-$. This will produce a random permutation. This is repeated n times. This procedure will result in a random sample of n permutations **with replacement**, a randomized set of permutations. The n values of the statistic—mean or t—yields a reference distribution, which is used in the usual manner to compute the p value.

For the example on husband-wife driving hours we considered in the article "Permutation Tests" in the March 2015 issue, we found the p-values for the paired t-test, the permutation test, and the randomization test (based on 10000 re-samples) to be respectively 0.0973, 0.0034, and 0.0018. The permutation test p-value (exact p-value) of 0.0034 is the best choice. The histogram of the 10000 chisquare values for the randomization samples is given below.



CONJOINT

A Mu Sigma University Publication, May 2015

Mu Sigma

In This Issue...

Selection of Resamples: Two-Sample Case

Let us take the example of 20 observations on x, 10 from each of two groups A and B, for which in the parametric hypothesis test, we would use a t statistic with 18 degrees of freedom to test the equality of the means of A and B. Suppose we need a random sample of say $n = 10000$ from the $\binom{20}{10} = 184756$ complete set of permutations. A sample permutation is selected as follows: From the 20 x values select a subset of 10 values (evidently **without** replacement) and assign them to group A and the remaining 10 to group B. Repeat this exercise n times. This gives you a **with replacement** sample of n permutations from the entire collection of permutations. A compilation of the difference between the means of groups A and B or a two-sample t computed from each re-sample yields the reference distribution.

Re-sampling a Frequency Table of Nominal Variables

The idea of a permutation test or exact test originated with Sir Ronald Fisher in the analysis of a $r \times c$ table of frequencies (r rows and c columns) of combinations of two categorical variables with small frequencies where the well-known chisquare test of independence or homogeneity is not applicable. Let us consider the case of a 2 × 2 table of frequencies where the rows are “campaign” (A and B) and columns are “result” (success or failure) (Y and N). Here consider 100 (fixed) customers who were subjected to campaign A and 200 (fixed) to campaign B with results as follows:

Observed Data (Frequencies)

Campaign⇓	Result		
	Y	N	Total
A	46	54	100
B	108	92	200
Total	154	146	300

Possible Configurations

Campaign⇓	Result		
	Y	N	Total
A	m_A	$100 - m_A$	100
B	m_B	$200 - m_B$	200
Total	$m_A + m_B$	$300 - m_A - m_B$	300

- The hypothesis of no association between campaign and result is that the proportion of success is the same for campaigns A and B.
- The chisquare statistic with 1 degree of freedom turns out to be 1.7079 with a p-value = 0.1913.
- The set of permutations for this data are various possible frequency tables with the row totals fixed and each of the 300 cases falling independently in the Y column or the N column with the same probabilities p and $1 - p$ respectively, consistent with the hypothesis to be tested.
- In the table, the marginal probabilities of Y and N are respectively 0.5133 and 0.4867.
- We then draw 100 Bernoulli(0.5133) random variables for the first row and assign cases with value 1 to the Y column and 0 to the N column.
- Similarly 200 cases in the second row are allotted to the Y or N column.
- Then we get a frequency table like the data table.
- A chi-square statistic is then calculated for the table.
- This process is repeated a large number of times (say, $n = 10,000$).

CONJOINT

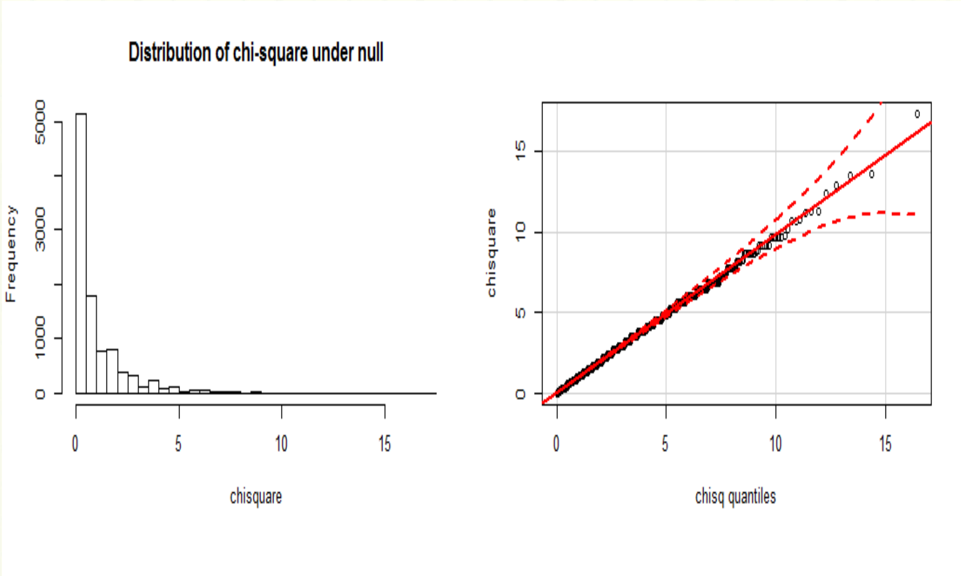
A Mu Sigma University Publication, May 2015

Mu Sigma

In This Issue...

- The p-value for this randomized test is computed as the proportion of the $n = 10,000$ tables whose chisquare values exceed that of the data table.
- Although this p-value is not exact in the sense of being derived from a full set of permutations, with 10,000 samples the probability will be very close to exact.
- And since the cell frequencies in our example are large, it should be close to the p-value given by Pearson's chisquare.

The results of such a randomized test based on 10,000 samples are given in the histogram of the 10,000 chisquare values, the QQ Plot of the chisquare (demonstrating a reasonable chisquare distribution for them) and the p-value from the randomized test as 0.1969. This is very close to the data table p-value of 0.1913.



An R code for this computation can be found in
https://www.uvm.edu/ dhowell/StatPages/ResamplingWithR/RandomMatchedSample/RandomMatchedSampleR.html

References

- T.M.Carsey & J.J.Harden (2013): *Monte Carlo Simulation and Resampling Methods for Social Science*. Thousand Oaks, CA: Sage Publications.
- P.Good (2005): *Introduction to Statistics Through Resampling Methods and R/S-PLUS*. New York: Wiley.