WEIGHTING INDIVIDUAL DATUM FOR NONPARAMETRIC ANALYSIS[1]

J. A. Tucker,[2] and M. Ortiz

Abstract: Many mining permits require a comparison of a stratified reference area to the reclaimed area for bond release. While this is easily accomplished by performing a test that uses stratified sample estimates of means and variances, such as the parametric t-test, nonparametric tests are often required by regulators if the data are not normally distributed or have small sample sizes. We propose three methods of weighting data that weight individual data points such that they can be used for nonparametric comparisons. These methods allow weights to be generated based on the pre-mining distribution of reference areas, the number of samples taken in the reference areas, and a combination of both of the aforementioned weights. These weights maintain comparability of the reference data to the reclaim data by confining the mean of the weights to one. This method provides a solution for nonparametric analysis where permits require comparison of a stratified reference to the reclaimed area.

Additional Key Words: statistics, stratified samples, bond release, reference areas, statistical tests

---

[1] Paper was presented at the 2007 National Meeting of the American Society of Mining and Reclamation, Gillette, WY, 30 Years of SMCRA and Beyond June 2-7, 2007. R.I. Barnhisel (Ed.) Published by ASMR, 3134 Montavesta Rd., Lexington, KY 40502.

[2] Justin Tucker is the project development manager for Buchanan Consultants, Ltd., Farmington NM, 87401; Mel Ortiz is a Professor of Biostatistics at the University of Texas, School of Public Health, El Paso, TX, 79968.

Proceedings America Society of Mining and Reclamation, 2007 pp 831-836
DOI: 10.21000/JASMR07010831
http://dx.doi.org/10.21000/JASMR07010831

Introduction

The use of reference areas as a reclamation success standard is prevalent in surface mining reclamation. Most often, the reference area is not one contiguous homogeneous vegetation type. Instead, reference areas comprise a variety of different vegetation types or strata. The reclamation standard thus becomes a combination of all of these different reference areas, or in other words, a stratified population. The reference area is stratified and weighted in order to adjust for the pre-mining distribution of the vegetation type on the disturbed lands. The use of stratified reference areas presents a problem for nonparametric statistical tests, which at present have no well accepted method to account for a stratified sample. This paper's objective is to provide methods for weighting data such that they can be used in nonparametric analysis where a stratified reference area serves as the reclamation standard.

Parametric statistical tests (i.e. Student's t-test) can easily be performed using a mean and variance from a stratified population, but confusion exists with regard to the appropriateness of a parametric test that uses the functional form of a normal distribution if the data gathered are not normally distributed. The Central Limit Theorem implies that parametric tests can be used when the sample size is large, without regard to the distribution of the sample data (Greenberg and Webster 1983, Greene 1993, Snedecor and Cochran, 1980). Parametric tests use means and variances to determine the likelihood of a sample having been drawn from a theoretical population (assumed to have a normally distributed mean), or other reference group (also assumed to have a normally distributed mean). The sample distributions (determined using these means and variances) are then compared through the test. The Student's t-test is easily weighted because the weights can be multiplied by the means and summed to an overall weighted mean.

The Central Limit Theorem suggests that the distribution of the sample mean is distributed approximately normal as a sample becomes larger (Greenberg and Webster 1983, Greene, 1993). Most often this is interpreted as meaning that parametric tests are appropriate where a dataset contains at least 30 samples. In other words, even if the data gathered are not normally distributed, Student t-tests can be used as long as there are at least 30 observations. Some confusion exists among regulators and the public about the meaning or implications of the Central Limit Theorem with regard to its application on non-normally distributed data and the use of parametric tests for bond release purposes. Specifically, some regulators require that tests for normality are conducted as a prerequisite for the use of a parametric statistical test. If the normality tests do not support that the data are distributed approximately normal, regulators often require the use of a nonparametric test (e.g. NMMMD, 2000).

Nonparametric tests have less restrictive assumptions about the underlying data and are generally considered less powerful than their parametric counterparts (NMMMD, 2000). The argument for using the nonparametric tests is that the nonparametric tests do not use parameters that are highly affected by extreme values in the data, such as the mean. Nonparametric statistical tests use the individual data points to calculate a t-statistic instead of using the mean and variance, as a parametric test would. Unfortunately, nonparametric tests lack the ability to adjust or compensate for a weighted comparison group, such as a stratified reference area. This poses a problem for determining reclamation success when a mine permit requires a stratified reference area to be used and a parametric test is not acceptable to regulators or the public.

In this paper we provide a brief description of nonparametric tests and propose 3 methods for weighting sample data from a stratified population for use in nonparametric statistical analyses. This is followed by a brief discussion of their use and limitations.

Nonparametric Weights

Theory

Nonparametric counterparts of the t-test compare the probability distributions of the sampled populations rather than specific parameters of these populations (e.g. means or variances). If it can be inferred that the distribution of one sampled population lies to the right or left of the other sampled population, the implication is that the two populations are different. Nonparametric methods often use the relative ranks of the sample observations rather than their actual numerical values to compare these probability distributions. The numerical values are only used to rank or order the observations. Statistics based on ranks of observations are called rank statistics.

For such an instance, a mean and variance of the entire sample has no bearing on the outcome of the statistical test. Not even the relative distance from the standard is considered, only the observations. Individually weighting the value of each observation is the only manner to correct for samples being drawn from differently weighted strata, simply because the individual data are the sole consideration for nonparametric tests.

Weights based on Area

Suppose it is of interest to compare a mean response such as total cover in a reclaimed area to the mean of the response in a reference area. Further, suppose the reference area is composed of T strata. Now let p_i be the proportion of the reference area that stratum *i* represents. Note that

(1)

$$\frac{\sum_{i=1}^{T} T * p_i}{T} = \frac{T * \sum_{i=1}^{T} p_i}{T} = \frac{T * 1}{T} = 1$$

That is, the mean of the proportions each multiplied by T equals 1. If weights based on area (A_i) are assigned to observations within each stratum as

(2)

$$A_i = T * p_i ,$$

then, because the mean of the $T * p_i$ equals one, observations that were taken in the bigger strata are adjusted upwards and those observations from the smaller strata are adjusted downwards. Moreover, the mean of the overall reference area remains the same.

Weights based on Sample Size

It is possible that the strata are not each sampled with the same number of observations, that is, stratum i has n_i observations. Let

(3)

$$\sum_{i=1}^{T} n_i = N$$

be the total number of observations in the reference area.

It is desirable to give individual observations within a heavily sampled stratum less weight than individual observations in a lightly sample stratum. For example, consider two strata with one stratum having twice the number of observations as a second stratum. Each individual observation in the less sampled stratum should weigh twice each observation in the more sampled stratum. Consider the inverse of the proportions of the overall sample that each stratum has. These are:

(4)

$$\frac{N}{n_1}, \frac{N}{n_2}, \frac{N}{n_3} \ldots \frac{N}{n_T}$$

and are candidates for weights of each observation. But the mean of these weights does not equal to one. By confining the mean of the weights to one, you can ensure comparability of the stratified reference area data to the data gathered from the reclaimed area. One way to force the mean to be equal to one is to divide each of the above weights by their mean. This mean is obtained as:

(5)

$$\frac{\sum_{i=1}^{T} \frac{N}{n_i}}{T} = \frac{N * \sum_{i=1}^{T} \frac{1}{n_i}}{T} = \frac{N * M}{T}$$

After letting M = $\sum_{i=1}^{T} \frac{1}{n_i}$.

Now dividing each $\frac{N}{n_i}$ by the mean results in the stratum weights (S_i).

(6)

$$S_i = \frac{\frac{N}{n_i}}{\frac{N * M}{T}} = \frac{T}{n_i * M}$$

These values of weights for each stratum now have a mean of one. The A_i weights based on areas and/or the S_i weights based on sample size can be used to adjust each individual observation. These adjusted values have the same mean as the original observations, but now reflect the areas of the strata from which they were obtained and the sample size allocated to each stratum.

Weights based on Area and Sample Size

In most cases, an adjustment will need to be made for both sample size and proportion of the area that the stratum represents. Multiplying the A_i and S_i weights yields a weight (W_i) that adjusts for both the sample size and area that the stratum represents.

(7)

$$W_i = A_i * S_i$$

Because both the A_i and S_i weight have been corrected such that their means individually equal one, the Wi weight needs no further correction.

Using the Weights

Unlike the parametric t-test with weighted data which uses means and variances from the strata to produce a mean and variance, the nonparametric weights are multiplied with each individual observation in the dataset prior to conducting the statistical analysis.

(8)

$$X_W = X_i * A_i$$
$$Or$$
$$X_W = X_i * S_i$$
$$Or$$
$$X_W = X_i * W_i$$

This yields a dataset with weighted data for each observation. It is important to note that weights for nonparametric analysis can only be used for continuous variables such as total or perennial cover estimates. After each observation has been weighted, the nonparametric analysis can be conducted.

Conclusion

We have proposed three methods for weighting observations (data) so that they can be used to conduct nonparametric statistical tests. Due to the conceptual underpinnings of nonparametric statistics, traditional methods of weighting data from a stratified reference area are inadequate for use in nonparametric analysis. Although means and variances are easily obtained from stratified populations, and the Central Limit Theorem argues that parametric tests such as the t-test can be used when a sample is sufficiently large, nonparametric tests are often requested by regulatory agencies. Where nonparametric tests have succeeded in relaxing the assumptions of normality, they have fallen short with their inability to be modified to test using a stratified or weighted reference area. Weighting of individual datum, or data points, is one method to overcome this shortcoming of nonparametric tests. Caution should be exerted when weighting individual level data. Specifically, this method of data weighting assumes that the observations are from a continuous type variable. Weighting ordinal or categorical variables is not appropriate using this method.

While parametric tests are easier and often more statistically powerful, these weighting methods provide one solution for reclamationists who must pacify regulators and the public by providing nonparametric tests using a stratified reference area as the success standard.

Literature Cited

Greenberg E. and C. Webster. 1983. Advanced Econometrics: A Bridge to the Literature. New York: Wiley.

Greene, William H. 1993. Econometric Analysis. New York: Macmillan Publishing

New Mexico Mining and Minerals Division. 2000. Coal Mine Reclamation Program Vegetation Standards. New Mexico Energy, Minerals, and Natural Resources Dept. Mining and Minerals Division. Santa Fe, New Mexico.

Snedecor G.W. and W.G. Cochran. 1980 (7^{th} ed.) Statistical Methods. Ames, Iowa:Iowa State University Press