Journal of Modern Applied Statistical Methods

Volume 10 | Issue 1 Article 2

5-1-2011

A Test That Combines Frequency and Quantitative Information

Norman Cliff
University of Southern California, nrcliff5@q.com

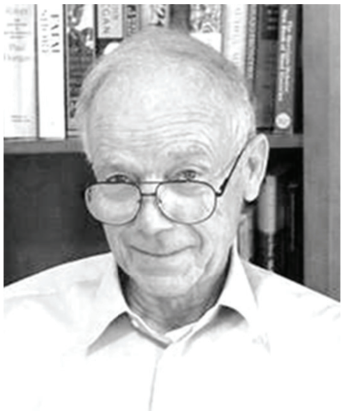
Recommended Citation
Cliff, Norman (2011) "A Test That Combines Frequency and Quantitative Information," *Journal of Modern Applied Statistical Methods*: Vol. 10 : Iss. 1 , Article 2.
DOI: 10.22237/jmasm/1304222460

Journal of Modern Applied Statistical Methods
May 2011, Vol. 10, No. 1, 2-7

1538 – 9472/11/$95.00

Invited Articles

A Test That Combines Frequency and Quantitative Information



Norman Cliff
University of Southern California,
Los Angeles, CA USA

In many simple designs, observed frequencies in subclasses defined by a qualitative variable are compared to the frequencies expected on the basis of population proportions, design parameters or models. Often there is a quantitative variable which may be affected in the same way as the frequencies. Its differences among the groups may also be analyzed. A simple test is described that combines the effects on the frequencies and on the quantitative variable based on comparing the sums of the values for the quantitative value within each group to the random expectation. The sampling variance of the difference is derived and is shown to combine the qualitative and quantitative aspects in a logical way. A version of the test based on enumeration of the possible values of the sum is described, an example is analyzed, factors affecting the test's power are discussed and extensions are suggested.

Key words: Combined effects, frequencies, means.

Introduction

A method is proposed that combines two simple analytical paradigms dating from the early 20th century. In one, the frequencies with which observations fall into each of a set of categories are compared to the frequencies that are expected under the hypothesis of an a priori model. The discrepancies are analyzed according to the elementary Chi-square test. In the other paradigm, the groups are compared on a quantitative variable through a *t*-test or analysis of variance. The Chi-square and the *F*- or *t*-test are among the most elementary of inferential methods and hundreds of sources describe them. Alternatively, a more recent variant, such as a loglinear model (e.g., Agresti, 1983), can be used instead of the Chi-square, and location comparisons can be made by more

Norman Cliff is Professor Emeritus in the Department of Psychology. Norman Cliff has degrees in psychology from Wayne State and Princeton Universities. He was the longtime head of the Quantitative Psychology Program at the University of Southern California. He has over 100 research publications and three books. He is a Fellow of the American Association for the Advancement of Science. E-mail him at: nrcliff5@q.com.

modern methods such as those described by Wilcox (1990, 2012) or an ordinal alternative (Cliff, 1993, 1996). The method described herein combines the frequency and quantitative information into a single test.

A common simple research context is one in which a population can be divided into g subpopulations, each with known relative frequency or theoretical probability π_i. Some members of the subpopulations then fall into a certain class, and the research question is whether members of the class come disproportionately from different sub-populations. A simple example is the hoary beads-of-different-colors-in-a-bag from which a random sample of beads is drawn and various questions about the contents of the bag can be investigated. Empirical examples might include investigating whether individuals with a certain disease come disproportionately from different city precincts, ethnicities, age groups, etc., each of whose population sizes are known; whether students taking high school advanced placement exams tend to come differentially from different schools, genders, etc.; whether psychotics tend to come from certain neighborhoods; whether the number of germinating plants tend to come from certain seed stocks, or netted fish tend to come from certain stocked batches or subspecies; whether defaulting mortgages tend to come from certain banks. Alternatively, there could be a model that determines the π.

Experimental contexts can also occur. Suppose three different inoculation regimens are employed, each on a large group. At a later time, incidences of the disease are recorded and the number that comes from each treatment is compared to its expected frequency based on the original group sizes. In all these cases it is assumed that there is an a priori probability that a random member of the observed class will come from subpopulation i. The observed number f_i that comes from i can be compared to $n\pi_i$, where n is the total number observed to fall in the class and elementary significance tests are applied to the results.

This article elaborates on such methods to cases where there is also an expected effect on an associated quantitative variable, specifically, for data such as: numbers written on the beads-in-a-bag; a measure of severity of disease; scores of students on an exam; the measured sizes of plants or fish; amounts or dates of loan defaults. The groups could compared quantitatively using some form of location comparison, such as, analysis of variance, *t*-test, modern more robust methods (Wilcox, 1990, 2012) or ordinal comparisons (Cliff, 1993, 1996).

Can the quantitative and qualitative information in testing a random model be combined? The traditional way this might be accomplished is to divide the quantitative variable into categories to form a cross-classification and then calculate expected cell frequencies or fit a loglinear model, etc. The qualitative variable could also be coded in some rational way and treated in parallel with the quantitative one via the general linear model. Here, combining quantitative and qualitative data more directly is suggested.

New Test Description

There are two beads-in-a-bag models to which the method can be applied. In the first, there is a large sack containing red and white beads. The supplier indicates that some beads, an equal number of red and white, have numbers written on them, and that the means of the red-bead numbers and white-bead numbers are the same. A sample of beads is taken, discarding those that do not have numbers, resulting in n numbered beads, some red and some white. The goal is to test the supplier's assurance of equal frequencies and equal means. A priori probabilities, π_r and π_w of 0.50, state that a numbered bead is white or red and the further hypothesis is that the means are the same for both red and white. In the general case, the a priori probabilities could be different, and/or there could be more than two colors of beads.

The second bead model uses two bags of beads, one red and one white. By hypothesis, equal proportions from red and white are numbered and the means of the numbers from red and white are equal. In this model, the plan is to sample s_w from the white bag and s_r from the red bag, once again discarding any unnumbered beads, and to determine how many of each are numbered and what the numbers are: if red and white beads are equally likely to be numbered, the probability that a numbered bead is white is $s_w/(s_w + s_r)$. The objective is to test the combined hypothesis that the probabilities are as assumed and that the means are equal and the method generalizes to more than two colors.

The natural way to test either of the models is to calculate $\Sigma_j x_{ij}$, the sum of the scores x_{ij} by the j^{th} member of subpopulation i who are in the class, that is, are a numbered bead, and compare it to a random expectation. Here, the obvious candidate is $n\pi_i m$, where m is the overall mean of X, n is the number falling in the class (numbered beads in the examples), and π_i is the a priori probability that the member came from subpopulation i. The difference, $\Sigma_j x_{ij} - n\pi_i m$, is a random variable that can be expected to be approximately normal under a wide variety of circumstances.

In order to assess whether the deviation could be consistent with a random model, it is essential to know the standard error of this difference. Its sampling variance, d_i^2, is $E[(\Sigma_j(x_{ij}) - \pi_i m)]^2$. To determine its form, first consider the expectation at a fixed cell frequency f_i and make use of $\Sigma_j x_{ij} = f_i m_i$ where m_i is the mean of the x_{ij} in i. The expected value of d_i^2 at a given f_i is

$$E(d_i^2) = E[f_i^2 m_i^2 - 2f_i m_i n \pi_i + (n\pi_i m)^2],$$

and, because m_i is the mean of f_i cases,

$$E(f_i^2 m_i^2] = f_i^2 \mu^2 + f_i \sigma_x^2.$$

Next, take the expectation across the possible sample values of f_i; where μ^2 and σ_x^2 are constants, and f_i is a binomial in π_i and n. Thus, because the expected value of a squared random variable is again the sum of its squared mean and its variance,

$$E[f_i^2] = (n\pi_i)^2 + n(\pi_i - \pi_i^2).$$

Putting this back into d_i^2 and collecting terms yields:

$$E(d_i^2) = n\mu^2(\pi_i - \pi_i^2) + n\pi_i \sigma_x^2.$$

This is exactly what one would expect: that the expected squared deviation under the null hypothesis is the sum of a term reflecting the expected deviation of the frequency from expectation and one reflecting the expected deviation of the subgroup mean from the overall mean. Under the null hypothesis, the two deviations are independent; their terms are therefore additive.

Under broad conditions, that is, when $n\pi_i$ is not too close to either n or zero and X is not far from normal with homogeneous variances across groups, the deviations $\Sigma x_{ij} - n\pi_i\mu$ are approximately normal with the given variance, in this case the obvious test is to compute the ratio of the observed difference to d_i. In the application that this method was developed to solve, X was the first n integers so that μ and σ^2 were known parameters – in which case the ratio can be taken as a standard normal deviate.

However, in most applications m and s^2 are estimates from the sample, the latter being a within-cells estimate. As was noted, d^2 has two components, one identical to the denominator of the Chi-square test and one derived from the variance. When the latter is a sample estimate, the ratio is no longer a normal deviate, but tends to resemble a t-ratio to some degree. (Note that the unbiased estimate of μ^2 is $m^2 - s^2/n$.) Consequently, a slightly conservative approach is to interpret the ratio as a t with $n - k$ df , k being the number of groups, although the expectation is that, in most contexts, the null sampling distribution may be very close to normal due to the influence of the first term in d^2. The method can be adapted to situations where $n\pi_i$ is close to the extremes, offering some special advantages over simply comparing frequencies under those circumstances.

Example

Table 1 contains artificial data that is used to illustrate the procedure. The data are analogous to what might be found if two groups of animals are given different cancer treatments. After a time the occurrence and size of lesions are determined, so x_{ij} is the size of the lesion in animal j from group i; originally, there were s_1 = 15 animals in treatment 1 and 10 in treatment 2, so the a priori probabilities that a given lesioned animal is in a given group are $\pi_1 = 0.6$ and $\pi_2 = 0.4$, analogously to the second bead example. The expectation is that lesions will be more common and larger in Group 1. Thirteen animals are found to have lesions in Group 1 and three in Group 2, so n = 16. The sizes of the lesions in each group are given in the upper part of the table along with the statistics for each group.

The lower part of the Table shows the components of d_i^2 and the t's for each group, which are found to be significant at the $\alpha = 0.05$ level, one-tailed. A SAS macro was written to perform the analysis (by Professor Du Feng), but it is easily carried out in small samples with the aid of a pocket calculator. An analysis based on the rank-order version of the data gave highly similar results.

Power Considerations

It would seem natural to expect that including quantitative information would increase power over the simple frequency analysis, but one may wonder about the circumstances under which this might actually be true. Note that d_i^2 has the appearance of combining expected frequency deviations and subgroup mean deviations, by adding these two components $f_i m_i - n\pi_i m$ can be made into a form:

$$f_i m_i - n\pi_i = [f_i m_i - f_i m] + [f_i m - n\pi_i m].$$

After squaring the second bracketed term, call it a^2, and comparing it to the frequency part of d_i^2, their ratio would give exactly the same result as would be obtained in computing the $\underline{i}^{th}$ component of the Chi-square for testing observed frequencies; thus, the frequency component of this test is similar to the traditional test.

The mean difference component resembles a component of the F-test on mean differences, but is not identical. Dividing $f_i^2(m_i - m)^2$ by $f_i s^2$ would give a component of F, but the corresponding term in d_i^2, $n\pi_i$, is the expected frequency, not f_i, of the observed group size itself, thus, these terms are similar, but are not the same.

However, the general circumstances under which using the combined test would be more powerful than simply using the frequencies can still be investigated. If b is defined as $f_i[m_i - \mu]$ and e^2 as the variance part of $E(d_i^2)$, then the ratio from the combined test is $(a + b)^2/(c^2 + e^2)$. The new ratio will be greater than the frequency ratio when

$$(a + b)^2/(c^2 + e^2) > a^2/c^2,$$

and, collecting some terms, this will be true when

$$(2ab + b^2)/e^2 > a^2/c^2.$$

This relation indicates that the new procedure is more likely to detect effects than simply testing the frequencies when both the mean and frequency effects are in the same direction as well as when the mean effect is relatively large.

Table 1: Artificial Animal Data to Illustrate the Combined Frequency and Quantitative Test

Group	Data	Statistics
1	13.1, 2.5,9.2, 6.2, 15.0, 12.1, 10.4, 17.4, 15.1, 6.0, 16.0, 6.1, 11.2	$\bar{x}$ =11.55
2	3.1, 9.3, 8.6	$\bar{x}$ = 7.00 m = 10.69 s^2 = 16.86

Group	Analysis			
	Σx_{ij}	$n\pi_i m$	d_i^2	t-ratio
1	150.1	102.66	578.06	1.973
2	21.0	68.44	531.75	2.057

Another consequence of examining the ratio in this way is seeing that its two aspects are implicitly weighted by the relative magnitudes of variance and squared mean. The other factors, $\pi_i - \pi_i^2$ and π_i, are similar in magnitude, their ratio being between 0.5 and 1.0. When the data consist of the first n positive integers, the ratio of squared mean to variance approaches 3.0 as n increases, indicating that frequency effects will always be emphasized relative to mean effects in such data.

The difference in influence can be even greater with some psychological variables whose mean and variance are set by convention. Many scholastic aptitude tests are scaled to have a mean about 500 and variance about 10,000, giving a ratio of about 25.0; the IQ scale is even more extreme, giving a ratio of squared mean to variance of more than 40.0. In such circumstances, the mean part of the proposed ratio has little effect because the proposed ratio approaches the traditional one for frequencies as μ^2/σ^2 increases.

In some research contexts, X has a well-established and empirically meaningful zero point. However, in others, such as the SAT and IQ scales, it merely represents a convenient reference. Where the origin of the scale is arbitrary, the user may feel that it is justifiable to give more nearly equal a priori weights to μ^2 and σ^2. However, it seems desirable that the lowest possible Σx_{ij} value should be zero, occurring when $f_i = 0$. Thus, subtracting a constant to make the lowest observed score slightly positive seems to be the most that can be done to equate influences. However, if X is quasi-normal with lowest standardized value of around −3.5 or −3.0, the ratio is still 9.0 to 12.0. Thus, making the analysis ordinal by converting the observed variable to the first n integers may be the most that can be done in equating influences of mean and variance.

Exact Version

When $n\pi_i$ is smaller than about five, the normality of the distribution of differences is likely to break down, making the assumed boundaries for an acceptance region unrealistic. In that circumstance, the researcher can construct cutoff values for the sum that correspond nearly exactly to a given rejection probability. These probabilities are now defined under a randomization hypothesis rather than on the basis of parameter estimates.

A given set of n x_{ij} values, that is, from all groups in the sample, defines 2^n possible values for $\Sigma_j x_{ij}$; of these, a certain fraction, corresponding to the desired rejection level, give the smallest (largest) values for the sum. These can be enumerated; if the obtained sum falls within this set, the null hypothesis is rejected. This enumeration process may improve power in such cases by defining a finer-grained rejection region than the corresponding test that is based only on the frequencies or only on the means.

The method is suggested by the beads-in-a-bag models. Consider an obtained sum for Group i and ask: What is the probability of obtaining a sum this small (large) or smaller (larger) when drawing n times with probability π_i? To illustrate with the example, the sum for Group 2 is 21.0, n is 16 and π_2 is 0.40.

There are $2^{16} = 65{,}336$ possible outcomes of randomly drawing a sum. Which are less than 21.0 and what are their respective probabilities? Of these outcomes, one has a sum of 0.0, that with $f_2 = 0$. This will happen with binomial probability 2.82×10^{-4}. There are 16 draws with $f_2 = 1$, each with probability 1.88×10^{-4}, and all have sums less than 21.0. There are 120 with $f_2 = 2$, all with probability 1.25×10^{-4}, but only 55 of them have sums less than 21.0. When $f_2 = 3$, there are only 23 that are less than 21.0, each having probability 8.35×10^{-5}. No combination of four has a sum below that limit.

Summing the probabilities of the instances that have sums less than 21.0 it is found that, under randomization, $0.000282 + 16 \times 0.000188 + 55\times 0.000125 + 23\times 0.0000835 = 0.012192$ is the probability of obtaining a sum of 21.0 or less for Group 2, which is just short of the 0.01 significance level. By contrast, if only the frequencies are considered, the corresponding binomial probability of $f_2 = 3$ or fewer is 0.0652. Also, the t-test in Table 1 yielded a significance level of about 0.04, less extreme than the probability obtained by enumeration.

Applications

Applied contexts having the characteristics that are appropriate to the method seem likely to be fairly common. Consider a state infectious disease-monitoring agency that observes an outbreak of a disease such as meningitis, and tabulates the locations, by district, of the disease. It might hope to identify the origin of the outbreak by tabulating frequency by district and comparing them to expectations based on district sizes. Here, n is the total number of meningitis cases and the π_i are defined by the relative sizes of the populations of the different districts. If the agency records the days since diagnosis of each case and uses it as the quantitative variable in the present method, an easier identification of the outbreak's focus may be possible.

Consider also a bank-regulating agency such as the Federal Deposit Insurance Corporation that is observing a group of banks to assess the riskiness of their policies. It knows the number of mortgages issued by the banks and records the defaults that occur for each, n being the total number of mortgages that are in default and the p_i are defined by the number of mortgages issued by each bank. Using either the days since default or the amount of the default as well as the frequency of default might well give a more sensitive measure of the banks' statuses than frequency alone.

In psychology, suppose individuals are given training in problem-solving. After training, they and a control group are given a problem to solve under a time-limit. Some individuals are successful and some not, n being successful, and the time taken to success is recorded. If there are s_t individuals in the trained group and s_c in the control, $\pi_t = s_t/(s_t + s_c)$, and similarly for π_c, represent the a priori probabilities that a success comes from the respective groups. Here, in order for the time variable to operate in the appropriate direction, it is best recorded as time remaining before the cut-off signal in order that small means and small frequencies are expected to go together.

In a study of differences in criminal recidivism, released convicts who have been under different prison regimens or treatments or who belong to different natural groups can be followed for a period. The frequency of re-incarceration can be combined with the length of sentence and analyzed in the proposed way. The method could also be applied to studies of the effects of educational treatments.

Many other potential applications exist; the key to the relevance of the method is the expectation that frequency and some quantitative variable will act in the same direction. It has been noted that treating the quantitative variable as a rank order may have some advantages.

It has been assumed that the qualitative variable consists of a single dimension of classification, but it seems in principle that this limitation is not necessary. The classification could have two or more ways as in a factorial or nested design and the relevant quantities could be computed for various effects. Another possible complication is dealing with more than one quantitative variable. Could the variables be combined by forming an optimally weighted composite of the observed variables? That optimization might be complicated by the necessity of keeping the composite positive. Investigation of such a possibility is beyond the scope of the present article.

Acknowledgements

This article benefited from discussions with and suggestions by Texas Tech University Professors Clyde Martin, Department of Mathematics and Statistics, e-mail: clyde.f.martin@ttu.edu; Du Feng, Human Development and Family Studies, e-mail: du.feng@ttu.edu; and Rand Wilcox, Department of Psychology, University of Southern California, e-mail: rwilcox@wilcox.usc.edu.

References

Agresti, A. (1984). *Analysis of ordinal categorical data.* New York: Wiley.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis.* Mahwah, NJ: Erlbaum.

Wilcox, R.W. (1990). Comparing the means of independent groups. *Biometrical Journal, 32,* 771-780.

Wilcox, R. W. (2012). *Statistics for the social and behavioral sciences*. New York: CRC Press.