A TEST OF HOMOGENEITY FOR A STRATIFIED SAMPLE

Tomas Garza-Hernandez, C-E-I-R de Mexico, and
Philip J. McCarthy, Cornell University

A common problem in the analysis of data in the social sciences is that of testing homogeneity on distributions of qualitative variables. This problem arises when a population has been divided into several so called "domains", and at the same time a qualitative variable is defined on each domain. In this situation it is often required to compare the distribution of the qualitative variable across these groups, that is, to test whether the proportion of elements belonging to each category of the qualitative variable is the same in each of the domains.

In general, the distribution of the population elements among the several domains is not known beforehand. This fact introduces a complication in the analysis, since it is not possible then to design a sampling procedure which will yield a specified number of elements from a given domain. Moreover, it sometimes happens in practice that a prior stratification, unrelated to the domains of study, has been used in selecting the sample, thus introducing more difficulties in the solution of the problem, which of course remains unchanged, since the stratification is not relevant to the purposes of the investigation.

An example of the situation described above is that of a survey conducted in 1952 in a Canadian Maritime Province, whose specific aims were to assertain the incidence of psychiatric disorders within different subgroups of the population under study. This survey was conducted in the following manner (1):

1. A county was divided into three strata, corresponding distinct geographical and social areas. From each stratum according to some preassigned sampling rates, a sample of households was selected, and either the male or female in the household was interviewed.

2. The set of domains of study corresponded to categories of an index called "Occupational Disadvantage", roughly a measure of average well-being in various occupational levels. For example, one domain corresponds to owner, salaried and professional occupations, and another to self-employed workers in agriculture and fishing.

3. The qualitative variable is an overall judgment made by the project psychiatrists: a person belongs to category 1 if he is a well person, to category 2 if he would almost qualify for psychiatric attention or therapy, and to category 3 if the diagnosis is doubtful.

In this example, the hypothesis to be tested is that the proportion of persons within each category of "psychiatric status" is the same for both categories of "occupational disadvantage", the strata to play no role in the analysis.

A first approach to the solution might be through the application of a χ^2-test to the separate strata, and then combining the results for all the strata taken together, as described, e.g., by Kendall (2). But perhaps an overall test is required, relating to the entire population, regardless of the stratification, among other reasons because a χ^2-test would be meaningless whenever no observations are obtained from a given domain, and this may easily happen in an actual situation.

Let N_i, $i=1,2,\ldots,L$, denote the number of elements in the i-th stratum, and within this particular stratum, let $N_{ia}, N_{ib}, \ldots, N_{id}$ denote the number of elements in domains a, b, ..., d, respectively.

Within each stratum, say the i-th, and within domain a, say, in this stratum, let $P_{ia(1)}, P_{ia(2)}, \ldots, P_{ia(k)}$ denote the proportions of elements pertaining respectively, to categories 1, 2, .., k.

Then, in the whole population, we are interested in comparing the proportions $P_{a(j)}$, $j=1,2,\ldots,k$, where

$$P_{a(j)} = \frac{\sum N_i \pi_{ia} P_{ia(j)}}{\sum N_i \pi_{ia}}$$

and where we denote N_{ia}/N_i by π_{ia}, for domain a, and similarly for other domains.

We now want to consider the following problem: To test the hypothesis H: $P_{a(j)}=P_{b(j)}=\ldots=P_{d(j)}$, for each j=1, 2, ..., k, on the basis of a sample drawn at random from the separate strata.

In tabular form, we have the following situation:

$P_{a(1)}$	$P_{a(2)}$	----	$P_{a(k)}$
$P_{b(1)}$	$P_{b(2)}$	----	$P_{b(k)}$
$P_{d(1)}$	$P_{d)2)}$	----	$P_{d(k)}$

and it is desired to test whether the quantities appearing in a given column are the same no matter what the actual value is, that is, if the distribution of elements among the categories is the same for all domains.

Noting that the sum across columns equals 1 for any one row, we may leave the last column out of the analysis, and we then state our hypothesis as follows:

$$H: \quad P_{a(j)}=P_{b(j)}= \ldots = P_{d(j)}$$

simultaneously for $j=1,2, \ldots, k-1$.

a. Estimates of the parameters: their variances and covariances.

In accordance with the usual approach of finite population methodology, the estimates that we propose for the above parameters are simply given by

$$p_{a(j)}=\hat{P}_{a(j)}= \frac{\sum N_i \frac{n_{ia}}{n_i} p_{ia(j)}}{\sum N_i \frac{n_{ia}}{n_i}},$$

$j = 1,2,\ldots,k-1.$

where n_i is the size of the sample drawn at random from the i-th stratum, n_{ia} is the number of elements of the n_i which fall in domain a, and $p_{ia(j)}$ is the proportion of elements in domain a, stratum i, that fall in the j-th category.

The variance of this estimate follows, after some manipulations:

$$\mathrm{Var}(p_{a(j)}) \doteq \frac{1}{N_a^2} \sum \frac{N_i^2 (N_i-n_i)}{n_i (N_i-1)} \pi_{ia} P_{ia(j)}(1-P_{ia(j)}) (1-\pi_{ia}) \pi_{ia} (P_{ia(j)}-P_{a(j)})^2 \quad \text{where } N_a=\sum N_{ia}$$

This formula agrees with Hartley's (3), p. 15, formula (33), and although Professor Hartley remarks that it "appears to be restricted to a proportional allocation of the sample to strata", we have derived it with no assumptions on the sample allocation.

There are two kinds of covariances between our estimates. The first one arises since the sample of size n_i is composed of $n_{ia}, n_{ib}, \ldots, n_{id}$, all these adding up to n_i, thus giving correlations across domains. The second one comes from the fact that in a given stratum we have $\sum P_{ia(j)}=1$, and hence there are correlations between estimates across categories.

It can be shown, then, that

$$\mathrm{Cov}(P_{a(j)}, p_{b(j)}) \doteq -\frac{1}{N_a N_b} \sum \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i - 1} \pi_{ia} \pi_{ib} (P_{ia(j)}-P_{a(j)})(P_{ib(j)}-P_{b(j)}),$$

and

$$\mathrm{Cov}(P_{a(j)}, P_{a(\ell)}) \doteq \frac{1}{N_a^2} \left\{ -\sum N_i^2 \frac{N_i - n_i}{N_i - 1} \frac{\pi_{ia} P_{ia(j)} P_{ia(\ell)}}{n_i} + \sum N_i^2 \frac{N_i-n_i}{N_i-1} \frac{\pi_{ia}(1-\pi_{ia})}{n_i} (P_{ia(j)}-P_{a(j)})(P_{ia(\ell)} - P_{a(\ell)}) \right.$$

b. Derivation of the test procedure.

A test procedure for the stated hypothesis should be based on the behavior of the estimates in the sampling process. This behavior is described by specifying the joint probability distribution of the estimates. No attempt will be made to derive the exact form of this distribution; instead we shall make an assumption concerning the joint distribution of the d(k-1) random variables $p_{a(j)}, \ldots, p_{d(j)}$; $j=1,2,\ldots,k-1$.

For the sake of simplicity, let us rewrite the parameters $P_{a(1)}, P_{b(1)}, \ldots, P_{d(k-1)}$ as $P_1, P_2, \ldots, P_{d(k-1)}$, the same change is notation holding for the estimates of the parameters. In the new notation our hypothesis is now

$H: P_j=P_{j+1}=\ldots=P_{j+d-1}$, for each $j=1, d+1, 2d+1, \ldots, (k-2)d+1$.

Making a straightforward generalization of the often used approximation to the distribution of a proportion, or of a difference of proportions, we propose the following assumption: Let P denote a column vector whose components are $P_1, P_2, \ldots, P_{(k-1)d}$, and p denote a column vector whose components are $p_1, p_2, \ldots, p_{(k-1)d}$. Let A denote the $(k-1)d \times (k-1)d$ matrix whose element in the (i, j)-th place is the covariance between p_i and p_j. Then, generalizing the result for the univariate case, let the joint density of the random variables $p_1, p_2, \ldots, p_{(k-1)d}$ be given by

$$\varphi(p,P) = c \cdot \exp\left\{-\frac{1}{2}(p-P)' A^{-1} (p-P)\right\} \quad \ldots\ldots(1)$$

a multivariate normal density, where c is a constant, and A^{-1} is the inverse matrix of A.

A likelihood-ratio criterion will now be

derived to test H. The maximum of $\varphi(p,P)$ over the whole parameter space is clearly seen to be attained for p=P, and let it be denoted by φ_Ω

In order to find the maximum of $\varphi(p,P)$ subject to the condition that H: $P_j = P_{j+1} = \ldots = P_{j+d-1}$, for j=1, d+1, ..., (k-2) d+1, we proceed as follows: Since the problem is equivalent to finding the maximum of the exponent in (1), we shall use the method of Lagrange multipliers, trying to express the restrictions under which the minimum is to be attained in vector form.

Let $S_1, S_2, \ldots, S_{(d-1)(k-1)}$ be a set of vectors, where S_1 has 1 as its first component, -1 as its second component and zeros all the way down. S_r is constructed from $S_{(r-1)}$ by shifting all its elements one place downwards, and replacing the first one by zero. Then, the product $S_1'P=0$ expresses the fact that $P_1=P_2$. Also, $S_2'P=0$ means $P_2=P_3$, and so on until we express $P_{d-1}=P_d$

The minimization procedure is then carried out by the usual method:

Let
$D=(p-P)'A^{-1}(p-P)+m_1 S_1'P+m_2 S_2'P+\ldots+ m_{(k-1)(d-1)} S'_{(k-1)(d-1)}P$, where the m's are constants.

Taking the partial derivative of D with respect to P, and setting it equal to zero, we have the set of (k-1)(d-1) +1 simultaneous equations:

$$\frac{\partial D}{\partial P} = 0$$

$$m_1 S_1'P = 0$$

$$m_{(k-1)(d-1)} S'_{(k-1)(d-1)} P=0$$

The solution of this set of equations gives the value of P which maximizes (1) subject to H. Let this maximum be φ_ω.

Then, the test criterion is as follows: If $\lambda = \varphi_\omega / \varphi_\Omega$ is greater than or equal to λ_0, reject H, where λ_0 is chosen according to the level of significance desired, and using the fact that $-2 \log\lambda$ follows a χ^2-distribution with d(k-2) degrees of freedom.

REFERENCES

1. Hughes, C.C., et al. (1960). People of Cove and Woodlot. New York: Basic Books, Inc.

2. Kendall, M. G. (1945-46). The Advanced Theory of Statistics. London: C. Griffin.

3. Hartley, H. O. (1959). Analytic Studies of Survey Data. Reprint Series #63. Statistical Laboratory, Iowa State University, Ames, Iowa.