FORMULAS FOR CORRELATION COEFFICIENTS UNDER SPECIAL RELATIONSHIPS BETWEEN ESTIMATED CHARACTERISTICS

Richard Griffin and Gary M. Shapiro
U. S. Bureau of the Census

I. INTRODUCTION

The following are some examples of situations in which the results of this article are applicable.

A. Suppose you have conducted a survey in which double sampling was used and you are interested in the ratio of Spanish persons with less than 12 years of education to total Spanish persons, where one quantity comes from the large sample and one from the small sample. This article gives a simple algebraic formula for the correlation coefficient between them so that the variance of the ratio can be estimated.

B. Suppose you want to compare the unemployment rate for Blacks to that of the total population. This article gives a simple algebraic formula for the correlation coefficient between the two estimated rates and also for the variance of the difference between the two rates.

The major impetus for this article came from a desire to estimate covariances among estimates from different subsamples of the 1970 Decennial Census. This was part of an effort to determine standard errors for the ratio of the number of certain minority groups who are U.S. citizens 18 years of age and older to the total number of U.S. citizens 18 years of age and older in individual counties. [2] This was of interest to us in connection with determining which jurisdictions should be included in coverage under the 1975 amendments to the Voting Rights Act. Consider, for example, the minority group Spanish.

Data on whether a person is of Spanish heritage or not is available from the 15 percent subsample of the decennial census. Citizenship data, however, is available from the 5 percent subsample. Thus we had to deal with ratios involving data from different subsamples of the census. We used Taylor series approximations on the ratios which left us with sums involving variances and covariances. Estimates of the variances were available. Many of the formulas in this article were derived for use in estimating the covariances.

Section II of this article gives the basic notation and assumptions. Section III contains all the formulas for correlation coefficients and, in four cases, the related variance estimates. Finally, the appendix covers proofs of all results given in section III.

It should be noted that the results in this paper have been proven only for simple random sampling. However, the authors have applied these results to systematic cluster samples and believe they yield reasonable approximations.

II. NOTATION

Suppose we have a population of N units. A simple random sample without replacement (S.R.S.W.O.R.) of size n_α is taken from this population. From the units in this sample, a S.R.S.W.O.R. of size n_β is taken. Among the n_α units in the first sample, there are n_γ units not selected in the second sample ($n_\alpha = n_\gamma + n_\beta$). The second sample is also a S.R.S.W.O.R. of size n_β from the entire population of N units. It is also true that the n_γ remaining units constitute a S.R.S.W.O.R. of size n_γ from the entire population.

Let x_A, x_B, and x_C be sample estimates of the number of units in the population that have characteristics A, B, and C respectively. Estimate x_A is calculated from the first sample, estimate x_B from the second sample, and estimate x_C from the n_γ remaining units. X_A, X_B, and X_C are the respective expected values of x_A, x_B, x_C.

Let M_A, M_B, and M_C represent the sets of all units in the population that have characteristics A, B, and C respectively. Also let $f_\alpha = \frac{n_\alpha}{N}$, $f_\beta = \frac{n_\beta}{N}$, and $f_\gamma = \frac{n_\gamma}{N}$. V is used to denote the coefficient of variation.

Let $x_{A'}$, $x_{B'}$, $M_{A'}$, $M_{B'}$, $X_{A'}$, $X_{B'}$ be defined in the same manner as x_A, x_B, M_A, M_B, X_A, and X_B respectively, except that A' and B' are different characteristics than A and B.

III. RESULTS

A. Let $M_A \subset M_B$, then $\rho_{x_A, x_B} =$

$$\frac{(1-f_\alpha)f_\beta}{f_\alpha(1-f_\beta)} \frac{V_{x_B}}{V_{x_A}}$$

This formula applies when the set of all units in the population that have the characteristic estimated from the large sample is a subset of the set of units in the population that have the characteristic estimated from the subsample. For example, consider unemployment and civilian labor force where civilian labor force is estimated from a subsample of the sample from which unemployment is estimated.

1. Special Case #1:

$$n_\alpha = n_\beta \Rightarrow \rho_{x_A, x_B} = \frac{V_{x_B}}{V_{x_A}}$$

This formula applies in the case where the set of units that have one of the characteristics is a subset of the set of units that have the other characteristic and both characteristics are estimated from the same sample.

2. Special Case #2:

$$A=B \Rightarrow \rho_{x_A, x_B} = \sqrt{\frac{\frac{n_\beta}{n_\alpha} - f_\beta}{1-f_\beta}}$$

(This result due to Bershad [1].)

This case applies when there is one characteristic estimated twice, the second estimate being formulated from a subsample of the sample used to produce the first estimate.

B. Let $M_B \subset M_A$, then $\rho_{x_A, x_B} = \frac{V_{x_A}}{V_{x_B}}$

This formula applies when the set of units in the population that have the characteristic estimated from the subsample is a subset of the set of units in the population that have the characteristic estimated from the large sample.

C. Let $M_C \subset M_B$, then $\rho_{x_B, x_C} = \frac{-f_\beta}{1-f_\beta} \frac{V_{x_B}}{V_{x_C}}$

This formula applies when a sample is taken and one characteristic is estimated from a subsample of the sample and the other characteristic is estimated from the units of the first sample that are not in the subsample. Furthermore, the set of units that have one of the characteristics must be a subset of the set of units that have the other characteristic. For example, consider lawyers and white collar workers where the number of lawyers is estimated from the 1970 Census 5 percent sample and the number of white collar workers is estimated from the 1970 Census 15 percent sample.

1. Special Case:

$$B=C \Rightarrow \rho_{x_B, x_C} = -f_\alpha \frac{\sqrt{\frac{n_\beta}{n_\alpha} \frac{n_\gamma}{n_\alpha}}}{\sqrt{(1-f_\beta)(1-f_\gamma)}}$$

(This result due to Bershad [1].)

This case applies when there is one characteristic estimated twice, first from a subsample of an original sample and secondly from the units of the original sample that are not in the subsample.

D. Let $M_A \cap M_B = \phi$,

$$\text{then } \rho_{x_A, x_B} = -\frac{1-f_\alpha}{f_\alpha (N-1)} \frac{1}{V_{x_A} V_{x_B}}$$

$$VAR(x_A - x_B) = \sigma^2_{x_A} + \sigma^2_{x_B} + \frac{2 X_A X_B (1-f_\alpha)}{f_\alpha (N-1)}$$

This is applicable when the set of units that have the characteristic estimated from the large sample is disjoint from the set of units that have the characteristic estimated from the subsample.

E. Let $M_B \cap M_C = \phi$,

$$\text{then } \rho_{x_B x_C} = \frac{1}{(N-1) V_{x_B} V_{x_C}}$$

$$VAR(x_B - x_C) = \sigma^2_{x_B} + \sigma^2_{x_C} - \frac{2 X_B X_C}{(N-1)}$$

This applies when a sample is taken and one characteristic is estimated from a subsample of the sample and the other characteristic is estimated from the units in the first sample that are not in the subsample. Furthermore, the set of units that have one of the characteristics is disjoint from the set of units that have the other characteristic. For example, consider college graduates and persons with only a high school education where the number of college graduates is estimated from the 1970 Census 15 percent sample and the number of persons with only a high school education is estimated from the 1970 Census 5 percent sample.

F. Let $M_B \subset M_A$, $M_{B'} \subset M_B$, $M_{A'} \subset M_A$, $M_{B'} \subset M_{A'}$,

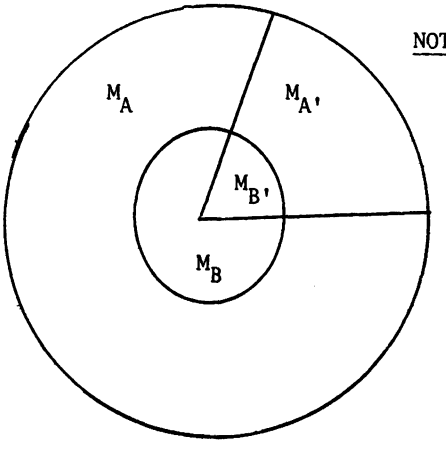
$$n_\alpha = n_\beta, \; p_{A'} = \frac{x_{A'}}{x_A}, \text{ and } p_{B'} = \frac{x_{B'}}{x_B},$$

$$\text{then } \rho_{p_{A'}, p_{B'}} \cong \frac{X_B \sigma_{p_{B'}}}{X_A \sigma_{p_{A'}}} \quad \text{and}$$

$$VAR(p_{A'} - p_{B'}) \cong \sigma^2_{p_{A'}} + \left(1 - 2\frac{X_B}{X_A}\right) \sigma^2_{p_{B'}}$$

(This result due to Tomlin [3].)

Illustration:



NOTE: M_A is the whole circle and M_B is the whole inner circle.

As an example of a situation in which these formulas apply, consider the unemployment rate for blacks and the overall unemployment rate, where both are estimated from the same sample.

G. Let A=B, $M_{A'} \subset M_A$, $M_{B'} \subset M_B$, $M_{A'} \cap M_{B'} = \phi$,

$n_\alpha = n_\beta = n$, $p_{A'} = \frac{x_{A'}}{x_A}$, $p_{B'} = \frac{x_{B'}}{x_B}$, $P_{A'} = \frac{X_{A'}}{X_A}$,

and $P_{B'} = \frac{X_{B'}}{X_{B'}}$

then

$$\rho_{p_{A'}, p_{B'}} \cong -\sqrt{\frac{P_{A'} P_{B'}}{(1-P_{A'})(1-P_{B'})}} \quad \text{and}$$

$$\mathrm{VAR}(p_{A'} - p_{B'}) \cong \sigma^2_{p_{A'}} + \left(1 + \frac{2P_{A'}}{1-P_{B'}}\right) \sigma^2_{p_{B'}}$$

(This result due to Tomlin [3].)

As an example of a situation in which these formulas apply, consider the proportion of employed persons who are chemists and the proportion of employed persons who are engineers, where both are estimated from the same sample.

IV. ACKNOWLEDGEMENTS

We would like to thank Paul H. Tomlin and Kirk Wolter, both of the Bureau of the Census, for helpful comments on this paper.

APPENDIX - PROOFS OF RESULTS

I. Proof of A

$$\mathrm{cov}(x_A, x_B) = \mathrm{cov}(E(x_A/n_\alpha), E(x_B/n_\alpha)) + E[\mathrm{cov}(x_A, x_B/n_\alpha)]$$

$$= \mathrm{cov}(x_A, y_B) + 0$$

where $y_B = \frac{N}{n_\alpha} \sum_{i=1}^{n_\alpha} s_i$, where $s_i = \begin{cases} 1 \text{ if unit i has characteristic B} \\ 0 \text{ otherwise} \end{cases}$

$$\left(\text{since } E\left(\frac{N}{n_\beta} \sum_{i=1}^{n_\beta} s_i/n_\alpha\right) = y_B\right)$$

$$\text{now } \mathrm{cov}(x_A, y_B) = N \frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} \left(\sum_{i=1}^{N} r_i s_i - \frac{\sum_{i=1}^{N} r_i \sum_{i=1}^{N} s_i}{N} \right) \quad (r_i \text{ is the indicator function for characteristic A})$$

$$= N \frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} X_A \left(1 - \frac{X_B}{N}\right)$$

$$\text{thus } \rho_{x_A, x_B} = \frac{N \frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} X_A \left(1 - \frac{X_B}{N}\right)}{\sigma_{x_A} \sigma_{x_B}}$$

$$= \frac{(N-n_\alpha) n_\beta}{n_\alpha (N-n_\beta)} \frac{V_{x_B}}{V_{x_A}} \left(\text{since } \sigma^2_{x_B} = N \frac{N-n_\beta}{n_\beta} \frac{1}{N-1} X_B \left(1 - \frac{X_B}{N}\right) = \frac{(1-f_\alpha) f_\beta}{f_\alpha (1-f_\beta)} \frac{V_{x_B}}{V_{x_A}}\right.$$

II. Proof of B

Using the same argument as in A, we have:

$$\rho_{x_A, x_B} = \frac{N \frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} X_B \left(1 - \frac{X_A}{N}\right)}{\sigma_{x_A} \sigma_{x_B}}$$

$$= \frac{V_{x_A}}{V_{x_B}}$$

III. Proof of C

$$\text{cov}(x_B, x_C) = \text{cov}(E(x_B/n_\alpha), E(x_C/n_\alpha)) + E(\text{cov}(x_B, x_C/n_\alpha))$$

Consider

$$E(x_B x_C/n_\alpha) = E\left(N \frac{\sum_{i=1}^{n_\alpha} v_i s_i}{n_\beta} \; N \frac{\sum_{i=1}^{n_\alpha} w_i t_i}{n_\gamma}\right)$$

(s and t are indicator functions for characteristics B and C respectively and v and w are the indicator functions for the samples of size n_β and n_γ, respectively.)

$$= \frac{N^2}{n_\beta n_\gamma}\left(\sum_{i=1}^{n_\alpha} s_i t_i E(v_i w_i) + \sum_{i \neq j}^{n_\alpha} s_i t_j E(v_i w_j)\right) = \frac{N^2}{n_\beta n_\gamma} \frac{n_\beta}{n_\alpha} \frac{n_\gamma}{(n_\alpha - 1)} \sum_{i=1}^{n_\alpha} t_i \left(\sum_{i=1}^{n_\alpha} s_i - 1\right)$$

so

$$E(\text{cov}(x_B x_C/n_\alpha)) = E\left[\frac{N^2}{n_\alpha(n_\alpha-1)} \sum_{i=1}^{n_\alpha} t_i \left(\sum_{i=1}^{n_\alpha} s_i - 1\right) - \frac{N^2}{n_\alpha^2} \sum_{i=1}^{n_\alpha} s_i \sum_{i=1}^{n_\alpha} t_i\right]$$

$$= \frac{N^2}{n_\alpha^2(n_\alpha-1)}\left(\sum_{i=1}^{N} t_i s_i E(u_i^2) + \sum_{i \neq j}^{N} t_i s_j E(u_i u_j)\right) - \frac{N^2}{n_\alpha(n_\alpha-1)} \sum_{i=1}^{N} t_i E(u_i)$$

(u is the indicator function for the sample of size n_α.)

$$= \frac{N^2}{n_\alpha^2(n_\alpha-1)}\left(X_C \frac{n_\alpha}{N} + X_C (X_B - 1) \frac{n_\alpha(n_\alpha-1)}{N(N-1)}\right) - \frac{N^2}{n_\alpha(n_\alpha-1)} X_C \frac{n_\alpha}{N}$$

$$= \frac{N}{n_\alpha(n_\alpha-1)} X_C + \frac{N}{n_\alpha(N-1)} X_C (X_B - 1) - \frac{N}{(n_\alpha-1)} X_C$$

$$= \frac{N - Nn_\alpha}{n_\alpha(n_\alpha-1)} X_C + \frac{N}{n_\alpha(N-1)} X_C (X_B - 1)$$

next

$$\text{cov}(E(x_B/n_\alpha), E(y_C/n_\alpha)) = N \frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} X_C \left(1 - \frac{X_B}{N}\right)$$

and then

$$\rho_{x_B, x_C} = \frac{N \frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} X_C \left(1 - \frac{X_B}{N}\right) + \frac{N - Nn_\alpha}{n_\alpha(n_\alpha-1)} X_C + \frac{N}{n_\alpha(N-1)} X_C (X_B - 1)}{\sigma_{x_B} \sigma_{x_C}} = \frac{-f_B}{1-f_B} \frac{V_{x_B}}{V_{x_C}}$$

(This equality results after the use of a lot of algebra which is not presented here.)

IV. Proof of D

Using arguments similar to those used in A and B, we have:

$$\rho_{x_A, x_B} = \frac{-N \frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} \frac{X_A X_B}{N}}{\sigma_{x_A} \sigma_{x_B}} = -\frac{1-f_\alpha}{f_\alpha(N-1)} \frac{1}{V_{x_A} V_{x_B}}$$

V. Proof of E

$$\text{cov}(x_B, x_C) = E\left[\text{cov}(x_B, x_C/n_\alpha)\right] + \text{cov}\left[E(x_B/n_\alpha), E(x_C/n_\alpha)\right]$$

$$\text{now } \text{cov}\left(E(x_B/n_\alpha), E(x_C/n_\alpha)\right) = \text{cov}(y_B, y_C) = -\frac{N-n_\alpha}{n_\alpha} \frac{1}{N-1} X_B X_C$$

(y_B is defined as in the proof of A and y_C is defined in the same manner.)

Next, consider

$$E\left[\text{cov}(x_B, x_C/n_\alpha)\right] = E\left[E(x_B x_C/n_\alpha) - E(x_B/n_\alpha) E(x_C/n_\alpha)\right]$$

$$E(x_B x_C/n_\alpha) = E\left(N \frac{\sum_{i=1}^{n_\alpha} v_i s_i}{n_\beta} \frac{N \sum_{i=1}^{n_\alpha} w_i t_i}{n_\gamma}\right)$$

(s and t are the indicator functions for characteristics B and C respectively and v and w are the indicator functions for the sample of size n_β and the sample of size of n_γ respectively.)

$$= \frac{N^2}{n_\beta n_\gamma} \left(\sum_{i=1}^{n_\alpha} s_i t_i E(v_i w_i) + \sum_{i \neq j}^{n_\alpha} s_i t_j E(v_i w_j) \right)$$

$$= \frac{N^2}{n_\beta n_\gamma} \sum_{i \neq j}^{n_\alpha} s_i t_j E(v_i w_j) = \frac{N^2}{n_\beta n_\gamma} \frac{n_\beta}{n_\alpha} \frac{n_\gamma}{n_\alpha - 1} \sum_{i \neq j}^{n_\alpha} s_i t_j$$

and

$$E(x_B/n_\alpha) \; E(x_C/n_\alpha) = \frac{N^2}{n_\alpha^2} \sum_{i=1}^{n_\alpha} s_i \sum_{i=1}^{n_\alpha} t_i = \frac{N^2}{n_\alpha^2} \sum_{i \neq j}^{n_\alpha} s_i t_j$$

so

$$E\left[\text{cov}(x_B, x_C/n_\alpha)\right] = E\left(\frac{N^2}{n_\alpha(n_\alpha-1)} \sum_{i \neq j}^{n_\alpha} s_i t_j - \frac{N^2}{n_\alpha^2} \sum_{i \neq j}^{n_\alpha} s_i t_j \right)$$

$$= \frac{N^2}{n_\alpha^2(n_\alpha-1)} E\left(\sum_{i \neq j}^{n_\alpha} s_i t_j \right) = \frac{N^2}{n_\alpha^2(n_\alpha-1)} \sum_{i \neq j}^{N} s_i t_j E(u_i u_j)$$

(u is the indicator function for the sample of size n_α.)

$$= \frac{N^2}{n_\alpha^2(n_\alpha-1)} X_B X_C \frac{n_\alpha(n_\alpha-1)}{N(N-1)} = \frac{N}{n_\alpha(N-1)} X_B X_C$$

Thus

$$\rho_{x_B, x_C} = \frac{\frac{N}{n_\alpha(N-1)} X_B X_C - \frac{N-n_\alpha}{n_\alpha} \frac{1}{(N-1)} X_B X_C}{\sigma_{x_B} \sigma_{x_C}} = \frac{\frac{1}{N-1} X_B X_C}{\sigma_{x_B} \sigma_{x_C}}$$

VI. Proof of F

Let $M_D = M_A \cap M_B^C$ and $P_{D'}$ be the estimated proportion of units in M_D that have characteristic A'. (M_B^C is the complement of M_B). $\text{cov}(p_{B'}, p_{D'}) = E\left[\text{cov}(p_{B'}, p_{D'})/(x_B, x_D)\right]$

$+ \operatorname{cov}\left[E\left(p_{B'}/(x_B, x_D)\right), E\left(p_{D'}/(x_B, x_D)\right)\right] = 0$, where x_D is the sample estimate of the number of units in the population that have characteristic D. Note that for cluster sampling rather than simple random sampling, this covariance may not always be close to zero.

Now, $p_{A'} \cong \frac{X_B}{X_A} p_{B'} + \frac{X_D}{X_A} p_{D'}$ Thus, $\operatorname{cov}(p_{A'}, p_{B'}) \cong \frac{X_B}{X_A} \operatorname{VAR}(p_{B'}) + \frac{X_D}{X_A} \operatorname{cov}(p_{B'}, p_{D'}) = \frac{X_B}{X_A} \operatorname{VAR}(p_{B'})$

$$\Rightarrow \rho_{p_{A'}, p_{B'}} \cong \frac{X_B \sigma_{p_{B'}}}{X_A \sigma_{p_{A'}}}$$

So that $\operatorname{VAR}(p_{A'} - p_{B'}) \cong \sigma^2_{p_{A'}} + \sigma^2_{p_{B'}} - 2 \frac{X_B}{X_A} \sigma^2_{p_{B'}} = \sigma^2_{p_{A'}} + \left(1 - 2 \frac{X_B}{X_A}\right) \sigma^2_{p_{B'}}$

VIII. Proof of G

Let r, r', and s' be the indicator functions for characteristics A, A', and B' respectively. Let u be the indicator function for the sample.

Then let $k = \sum_{i=1}^{N} u_i r_i$, $\ell = \sum_{i=1}^{N} u_i r'_i$, and $m = \sum_{i=1}^{N} u_i s'_i$

So $\operatorname{cov}(p_{A'}, p_{B'}) = \operatorname{cov}\left(\frac{\ell}{k}, \frac{m}{k}\right) = \operatorname{cov}\left(E\left(\frac{\ell}{k}/k\right), E\left(\frac{m}{k}/k\right)\right) + E\left(\operatorname{cov}\left(\frac{\ell}{k}, \frac{m}{k}/k\right)\right) = 0 + E\left(\frac{1}{k^2} \operatorname{cov}(\ell, m/k)\right)$

$E(\ell/k) = k P_{A'}$ and $E(m/k) = k P_{B'}$ $\quad E(\ell m/k) = E\left(\sum_{i=1}^{X_A} z_i r'_i \sum_{i=1}^{X_A} z_i s'_i\right)$

where z is the indicator function for a S.R.S.W.O.R. of k units that have characteristic A from the population of units that have characteristic A.

$$= E\left(\sum_{i=1}^{X_A} z_i^2 r'_i s'_i + \sum_{i \neq j}^{X_A} z_i z_j r'_i s'_j\right) = \sum_{i \neq j}^{X_A} r'_i s'_j E(z_i z_j) = \frac{k(k-1)}{X_A(X_A - 1)} X_{A'} X_{B'}$$

Thus $\operatorname{cov}(\ell, m/k) = \frac{k(k-1)}{X_A(X_A-1)} X_{A'} X_{B'} - \frac{k^2 X_{A'}}{X_A} \frac{X_{B'}}{X_B} = -k P_{A'} P_{B'} \left(k - \frac{X_A}{X_A - 1}(k-1)\right)$ (since A=B) $\cong -k P_{A'} P_{B'}$

(A good approximation for large X_A)

So, $\operatorname{cov}(p_{A'}, p_{B'}) \cong E\left(-\frac{1}{k} P_{A'} P_{B'}\right) \cong -\frac{1}{k} P_{A'} P_{B'}$

From this we have, $\operatorname{VAR}(p_{A'} - p_{B'}) \cong \frac{P_{A'}(1-P_{A'})}{k} + \frac{P_{B'}(1-P_{B'})}{k} + 2\frac{P_{A'} P_{B'}}{k} \cong \sigma^2_{p_{A'}} + \sigma^2_{p_{B'}} \left(1 + \frac{2P_{A'}}{1-P_{B'}}\right.$

and,

$$\rho_{p_{A'}, p_{B'}} \cong -\sqrt{\frac{P_{A'} P_{B'}}{(1-P_{A'})(1-P_{B'})}}$$

REFERENCES

1. Bershad, Max A., "Correlations Between Half Samples," Internal Census Bureau memorandum to William N. Hurwitz, July 1, 1968.

2. Miskura, Susan, "Documentation of the Derivation of Estimates and Variance Estimates for the Voting Rights Amendment Tabulations," Internal Census Bureau memorandum to Gary M. Shapiro, November 10, 1975.

3. Tomlin, Paul H., "A Possible Method of Presentation of Comparisons of Proportions in Source and Reliability Statements," Internal Census Bureau memorandum to Gary M. Shapiro, February 19, 1976.