*The Annals of Mathematical Statistics*
1972, Vol. 43, No. 2, 656-658

# SENSITIVITY OF RAKED CONTINGENCY TABLE TOTALS TO CHANGES IN PROBLEM CONDITIONS

BY B. D. CAUSEY

*Bureau of the Census*

This paper describes a way of investigating how the entries in a "raked" contingency table—one based on a given contingency table but having row and column totals equal to specified quantities in such a way as to minimize "discrimination information" between the two tables—vary as functions of the entries of the original table and the quantities to which the row and column totals of the "raked" table are constrained to be equal.

**1. Introduction.** Suppose that we have a $p \times q$ contingency table with given entries $n(i, j)$, $i = 1, \cdots, p$, $j = 1, \cdots, q$, and that we find a corresponding set of (noninteger) entries $x(i, j)$ satisfying $\sum_j x(i, j) = r(i)$ and $\sum_i x(i, j) = c(j)$ for given $r(i)$ and $c(j)$. In practice we do not determine $x(i, j)$ exactly but use the fact [2] that $x(i, j) = \lim_{k\to\infty} z(i, j, k)$—where, letting $z(i, j, 0) = n(i, j)$, $v(i, k) = \sum_j z(i, j, k)$, and $w(j, k) = \sum_i z(i, j, k)$, we have for $k$ odd $z(i, j, k) = z(i, j, k - 1)r(i)/v(i, k - 1)$ and for $k$ even $z(i, j, k) = z(i, j, k - 1) \times c(j)/w(j, k - 1)$, $k = 1, 2, \cdots$. We refer to the procedure of iteratively computing the quantities $z(i, j, k)$ from the quantities $z(i, j, k - 1)$ as "raking". We may approximate the quantities $x(i, j)$ as closely as we like by doing a sufficiently large number of iterations. The chief topic of this paper is investigation of the behavior of the quantities $x(i, j)$ (available only by approximation) as functions of changing quantities $r(i)$, $c(j)$, and $n(i, j)$. The investigation described in this paper began as an attempt to estimate the variance of the quantities $x(i, j)$ as induced by variance in the quantities $n(i, j)$; we deal with this particular question in Section 3.

Letting $u$ and $v$ denote the grand totals $\sum\sum n(i, j)$ and $\sum r(i)$ respectively, the quantities $x(i, j)$ have the property [2] that for $y(i, j) = x(i, j)$ the discrimination information function $\sum\sum y(i, j) \log (uy(i, j)/vn(i, j))$ is minimized, subject to $\sum_j y(i, j) = r(i)$ and $\sum_i y(i, j) = c(j)$. It may be shown that $x(i, j) = a(i)b(j)n(i, j)$ for some quantities $a(i)$ and $b(j)$, so that our basic problem boils down to investigating changes in the quantities $a(i)$ and $b(j)$ as functions of the quantities $r(i)$, $c(j)$, and $n(i, j)$.

**2. Basic problem and solution.** We now allow the quantities $r(i)$, $c(j)$, and $n(i, j)$ to be given as known (noninteger) differentiable functions $r(i, t)$, $c(j, t)$, and $n(i, j, t)$ of a scalar $t$ in a neighborhood of $t = 0$; we must always, of course, have $\sum_i r(i, t) = \sum_j c(j, t)$. We abbreviate $r(i, 0)$, $c(j, 0)$, and $n(i, j, 0)$

Received February 19, 1971.

to just $r(i)$, $c(j)$, and $n(i, j)$, and denote the (known) derivatives $dr(i)/dt$, $dc(j)/dt$, and $dn(i, j)/dt$ evaluated at $t = 0$ by $R(i)$, $C(j)$, and $N(i, j)$.

We know that the products $a(i)b(j)$ are unique. In order that the quantities $a(i)$ and $b(j)$ may all be uniquely determined, we introduce the constraint $(1/p) \sum a(i) = (1/q) \sum b(j)$. Suppose that $a(i) = a(i, 0)$ and $b(j) = b(j, 0)$ have been explicitly found corresponding to $t = 0$; we let $A(i)$ and $B(j)$ denote the (unknown) derivatives of $a(i, t)$ and $b(j, t)$ evaluated at $t = 0$. We thus may obtain

$$0 = (1/p) \sum A(i) - (1/q) \sum B(j) . \tag{2.1}$$

Also,

$$\begin{aligned} R(i) = {} & A(i)(\textstyle\sum_j b(j)n(i, j)) + a(i)(\sum_j B(j)n(i, j)) \\ & + a(i)(\textstyle\sum_j b(j)N(i, j)) , \qquad i = 1, \cdots, p . \end{aligned} \tag{2.2}$$

Rearranging (2.2), multiplying it on both sides by $a(i)r(i)$ for $i = 1, \cdots, p$, adding (2.1), and in general letting $d(h, k) = 1$ for $h = k$ and 0 for $h \neq k$, we obtain

$$\begin{aligned} & (a(i)/r(i))(R(i) - a(i)(\textstyle\sum_j b(j)N(i, j))) \\ & \quad = \textstyle\sum_I (d(i, I) + (1/p))A(I) + \sum_j ((a^2(i)n(i, j)/r(i)) \\ & \qquad - (1/q))B(j) , \qquad i = 1, \cdots, p . \end{aligned} \tag{2.3}$$

Likewise we may obtain (this time subtracting (2.1))

$$\begin{aligned} & (b(j)/c(j))(C(j) - b(j)(\textstyle\sum_i a(i)N(i, j))) \\ & \quad = \textstyle\sum_i ((b^2(j)n(i, j)/c(j)) - (1/p))A(i) \\ & \qquad + \textstyle\sum_J (d(j, J) + (1/q))B(J) , \qquad j = 1, \cdots, q . \end{aligned} \tag{2.4}$$

Thus we have a linear system of $p + q$ equations in the $p + q$ unknowns $A(i)$, $i = 1, \cdots, p$, and $B(j)$, $j = 1, \cdots, q$. Once having found these, we easily have that $X(i, j)$—i.e. $dx(i, j)/dt$ at $t = 0$—equals $A(i)b(j)n(i, j) + a(i)B(j)n(i, j) + a(i)b(j)N(i, j)$.

The coefficient matrix $M$ does not depend on the problem derivatives, so that one may find its inverse $M^{-1}$, or, better yet, transform it to a new matrix $L$ for use in conjunction with the Crout procedure [1] for solving simultaneous linear equations, and then use either $M^{-1}$ or $L$ in conjunction with as many sets of $R(i)$, $C(j)$, and $N(i, j)$ as we like, to find the corresponding quantities $A(i)$ and $B(j)$.

We may express $M$ in block form as $\left(\begin{smallmatrix} M(1,1) & M(1,2) \\ M(2,1) & M(2,2) \end{smallmatrix}\right)$ where (1) $M(1, 1)$ is a $p \times p$ matrix with entries $1 + 1/p$ along the diagonal and $1/p$ off, (2) $M(2, 2)$ is a $q \times q$ matrix with entries $1 + 1/q$ along the diagonal and $1/q$ off, and (3) in the particular situation where the quantities $a(i)$ $(b(i))$ are all the same the average of the entries in $M(1, 2)$ $(M(2, 1))$ is exactly zero, while when these quantities are not all the same these entries should tend to be centered around zero.

**3. Corollary results.** For most cases of interest, the functions $r(i, t)$, $c(j, t)$, and $n(i, j, t)$ should be simply linear, with at most one or two of them varying in terms of $t$ in a simple fashion such as $r(i, t) = r(i) + t$. We may make a first-order approximation that for a change $d$ in $t$ from 0, $x(i, j)$ will change by an amount $X(i, j)d$. If $t$ can be construed as a random variable with mean 0 and variance $v$, and $f$ is a differentiable function of the quantities $x(i, j)$, we may estimate Var $f$ by

$$v\left(\frac{df}{dt}\right)^2 = v\left(\sum\sum \frac{df}{dx(i, j)} X(i, j)\right)^2 ;$$

more generally, if $t(k)$, $k = 1, \cdots, m(t)$, is a series of random scalars defined like $t$, and $f(h)$, $h = 1, \cdots, m(f)$, is a series of functions defined like $f$, we may estimate Cov $(f(h), f(H))$ by

$$\sum_k \sum_K \frac{df(h)}{dt(k)} \frac{df(H)}{dt(K)} \text{Cov}\,(t(k), t(K)) , \quad \text{with} \quad \frac{df(h)}{dt(k)} = \sum_i \sum_j \frac{df(h)}{dx(i, j)} \frac{dx(i, j)}{dt(k)} .$$

One can easily investigate special cases of interest such as the quantities $t(k)$ corresponding singly to the quantities $n(i, j)$, the quantities $f(h)$ corresponding singly to the quantities $x(i, j)$, and/or the covariances Cov $(n(i, j), n(I, J))$ (corresponding to Cov $(t(k), t(K))$) based on simple random sampling.

**4. Generalizations.** The ideas of this paper extend in principle from two-way contingency tables to tables of higher order, where marginal sum constraints (corresponding to $r(i)$ and $c(j)$) may be at any level.

Also, in Section 2 we may consider a series of scalars $t(1), \cdots, t(m(t))$, and solve for second derivatives $d^2a(i)/dt(k)dt(K)$ and $d^2b(j)/dt(k)dt(K)$ after finding the first derivatives. Using this information, the first-order estimates of Section 3 may be replaced by second-order estimates. Such derivatives may be found in principle to as high an order as we like, although beyond even the first order it might not be worthwhile to do all the necessary computations.

## REFERENCES

[1] HILDEBRAND, F. B. (1961). *Methods of Applied Mathematics*. Prentice-Hall, Englewood Cliffs, New Jersey.

[2] IRELAND, C. T. and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188.