Aplikace matematiky

Jaroslav Hájek

On the theory of ratio estimates

Aplikace matematiky, Vol. 3 (1958), No. 5, 384–398

Persistent URL: http://dml.cz/dmlcz/102632

ON THE THEORY OF RATIO ESTIMATES

JAROSLAV HÁJEK

(Received 29 November 1957.) DT:519.27.001

Estimated variances, yielded by large sample approach, are adjusted by a proportional regression approach; subsequently, under the assumption of normality, exact statements on confidence intervals are arrived at. The paper deals too, with complex types of ratio estimates, as well as with modifications needed, when stratification, multiple stages, or some special methods of first-stage sampling are present.

1. Introduction

Ratio estimates belong to the most efficient techniques of modern sample survey practice. Some of them are very ingenious, (see, for example [2], vol. I, p. 413), and make use of very elaborate supplementary information. Nevertheless, the present theory of ratio estimates has not gone beyond approximating their variances. At the same time, the validity of estimated variances is inferred from the discrepancy between the confidence interval, yielded by them, and Fieller's confidence interval. We shall show, however, that, under conditions typical for sample surveys, the Fieller's confidence interval is less advantageous than the usual one, because it is longer for any sample outcome, and, despite this, covers the true value with a smaller probability. In addition, Fieller's device cannot be used for ratio estimates of more complex type. Thus the Fieller's method, though very useful outside the sample surveys domain, (because of the generality of conditions under which it works), cannot be considered as preferable in sample survey conditions.

Let us select n elements from a population of N elements by simple random sampling without replacement, observe values y_i, x_i, z_i of some variables, and wish to estimate the total

$$Y = \sum_{i=1}^{N} y_i \,, \tag{1}$$

under the supposition, that we know totals[1]) and some subtotals of related

[1]) If we only estimate the ratio of totals, $\frac{Y}{X}$, we need not know the total X.

values x_i and z_i, $i = 1, \ldots, N$. Now, let each value y_i, x_i, z_i split into a sum of values y_{aci}, x_{aci}, z_{aci}, respectively, so that

$$y_i = \sum_a \sum_c y_{aci}, \; x_i = \sum_a \sum_c x_{aci}, \; z_i = \sum_a \sum_c z_{aci}, \; a \in A, \; c \in C \,. \tag{2}$$

As a rule, y_{aci}, x_{aci}, z_{aci} will refer to a two-way classification of conditions within the i-th element; for example, y_{aci} may be number of workers in the a-th age-sex group, within the c-th rural-urban part of the i-th county, y_{aci} and z_{aci} may refer to the present and last census number of all persons, respectively, again in the a-th age-sex group within the c-th rural-urban part of the i-th county.

Symbols Y, Y_a, Y_{ac}, X_a, X_{ac}, Z_c and y, y_a, y_c, y_{ac}, x_a, x_{ac}, z_c will denote population and sample totals, respectively, extended over all subscript letters that are not indicated. For example,

$$X_a = \sum_c \sum_{i=1}^{N} x_{aci} \,, \quad y_{ac} = \sum_i y_{aci} \,, \quad \text{etc.,}$$

where $\sum_i$ denotes the sum extended over subscripts of sampled elements, $i = i_1, \ldots, i_n$.

We shall consider the following three types of ratio estimates:

$$X \frac{y}{x} \,, \tag{3-1}$$

$$\sum_a X_a \frac{y_a}{x_a} \,, \tag{3-2}$$

$$\sum_a X_a \frac{\sum_c Z_c \frac{y_{ac}}{z_c}}{\sum_c Z_c \frac{x_{ac}}{z_c}} \,. \tag{3-3}$$

The estimates $(3 - 1)$ and $(3 - 2)$, obviously, are special cases of the estimate $(3 - 3)$.

2. Large — sample approach

The large — sample approach rests upon approximating estimates (3) by a linear function of the sample totals y, x, y_a, x_a, x_c, z_c, y_{ac} involved; the latter function can be obtained from the usual Taylor expansion about corresponding expectations. Carrying out this operation we get the following approximate expressions for estimates (3):

$$Y + \frac{N}{n}\left(y - \frac{Y}{X} x\right), \tag{4-1}$$

$$Y + \frac{N}{n} \sum_a \left(y_a - \frac{Y_a}{X_a} x_a\right), \tag{4-2}$$

$$Y + \frac{N}{n} \sum_a \sum_c \left[y_{ac} - \frac{Y_{ac}}{Z_c} z_c - \frac{Y_a}{X_a}\left(x_{ac} - \frac{X_{ac}}{Z_c} z_c\right)\right]. \tag{4-3}$$

Estimated variances of "concomitant" linear estimates (4) can be calculated in the usual way with the only modification, that

$$y_i - \frac{Y}{X} x_i, \tag{5-1}$$

$$\sum_a \left(y_{ai} - \frac{Y_a}{X_a} x_{ai}\right), \tag{5-2}$$

$$\sum_a \sum_c \left[y_{aci} - \frac{Y_{ac}}{Z_c} z_{ci} - \frac{Y_a}{X_a}\left(x_{aci} - \frac{X_{ac}}{Z_c} z_{ci}\right)\right], \tag{5-3}$$

are treated as single observations, and that sample ratios $\frac{y_a}{x_a}, \frac{y_{ac}}{z_c}, \frac{x_{ac}}{z_c}$ are substituted for population ratios $\frac{Y_a}{X_a}, \frac{Y_{ac}}{Z_c}, \frac{X_{ac}}{Z_c}$, respectively. In this manner we get estimated variances of estimates (4) in the general form

$$\frac{(N - n)n}{N(n - 1)} \sum_i \Delta_i^2 \tag{6}$$

where, for individual estimates (3), Δ_i equal

$$\Delta_i = \begin{cases} \frac{N}{n}\left(y_i - \frac{y}{x} x_i\right) & (7\text{-}1) \\ \sum_a \frac{N}{n}\left(y_{ai} - \frac{y_a}{x_a} x_{ai}\right) & (7\text{-}2) \\ \sum_a \sum_c \frac{N}{n}\left[y_{aci} - \frac{y_{ac}}{z_c} z_{ci} - \frac{y_a}{x_a}\left(x_{aci} - \frac{x_{ac}}{z_c} z_{ci}\right)\right]. & (7\text{-}3) \end{cases}$$

respectively. See, too, equations (11).

We shall not enlarge upon well-known asymptotical properties of estimates (3) and their estimated variances (6); reference is made to the paper [6]. The main point of this section was to show how to get estimated variances in a simple form even for complex ratio estimates.

Remark 2.1. Inserting (7 — 3) into (6), we obtain the estimated variance which is much simpler but otherwise equivalent to that recommended in [2], vol. II., p. 226.

3. Proportional regression approach

The expressions (5) suggest that the corresponding ratio estimates (3) are fitted to the following regression models:

$$M(y_i|x_i) = \frac{Y}{X} x_i \,, \tag{8-1}$$

$$M(y_{ai}|x_{a'i}, a' \in A) = \frac{Y_a}{X_a} x_{ai} \,, \quad a \in A \,, \tag{8-2}$$

$$M(y_{aci}|x_{a'c'i}, z_{c'i}, a' \in A, c' \in C) = \frac{Y_{ac}}{Z_c} z_{ci} + \frac{Y_a}{X_a}\left(x_{aci} - \frac{X_{ac}}{Z_c} z_{ci}\right), a \in A \,, c \in C \,. \tag{8-3}$$

Here, and in what follows, $M(y|x_\alpha, \alpha \in H)$ and $D(y|x_\alpha, \alpha \in H)$ denote the conditional expectation and variance of y with respect to a finite family of random variables $\{x_\alpha, \alpha \in H\}$. Linear regression, which passes through the origin, we shall call, for brevity, proportional regression[2]). All regressions, given by the equations (8) are proportional regressions. Relation (8-2) implies that y_{ai}, given x_{ai}, is linearly independent of $x_{a'i}$, $a' \neq a$. Relation (8-3) means that y_{ai} and x_{ai} both are proportional to z_{ci}, and moreover, that the residual $y_{aci} - \frac{Y_{ac}}{Z_c} z_{ci}$ is proportional to the residual $x_{aci} - \frac{X_{ac}}{Z_c} z_{ci}$.

When sampling is from finite populations, relations (8) must be, generally, reinterpreted, (see Remark 3.1). Let us consider, therefore, the parallel problem when the sample consist of n independent observations from a multidimensional distribution governed by one of the relations (8). Under this condition, the conditional expectations of ratio estimates (3), for any fixed x_{aci} and z_{aci}, will equal Y. It suffices to show it for the most general estimate (3 — 3):

$$M\left(\sum_a X_a \frac{\sum_c Z_c \frac{y_{ac}}{z_c}}{\sum_c Z_c \frac{x_{ac}}{z_c}} \,\middle|\, x_{a'c'1}, \ldots, x_{a'c'n}, z_{c'1}, \ldots, z_{c'n}, a' \in A, c' \in C\right) =$$

$$= \sum_a \sum_c \frac{X_a \frac{Z_c}{z_c}}{\sum_c Z_c \frac{x_{ac}}{z_c}} \sum_{i=1}^{n} M(y_{aci}|x_{a'c'i}, z_{c'i}, a' \in A, c' \in C) =$$

[2]) The proportionality relation is irreversible: If $M(y|x) = \varphi x$ and $M(x) \neq 0$, then $M(x|y) \neq \psi y$, unless $y = \varphi x$ with probability 1.

$$= \sum_a \sum_c \frac{X_a \frac{Z_c}{z_c}}{\sum_c Z_c \frac{x_{ac}}{z_c}} \sum_{i=1}^{n} \left[\frac{Y_{ac}}{Z_c} z_{ci} + \frac{Y_a}{X_a}\left(x_{aci} - \frac{X_{ac}}{Z_c} z_{ci}\right)\right] =$$

$$= \sum_a \sum_c \frac{Y_a Z_c \frac{x_{ac}}{z_c}}{\sum_c Z_c \frac{x_{ac}}{z_c}} + \sum_a \frac{\sum_c X_a \frac{Z_c}{z_c}\left(Y_{ac} \frac{z_c}{Z_c} - \frac{Y_a}{X_a} X_{ac} \frac{z_c}{Z_c}\right)}{\sum_c Z_c \frac{x_{ac}}{z_c}} =$$

$$= \sum_a Y_a + \sum_a \frac{X_a Y_a - Y_a X_a}{\sum_c Z_c \frac{x_{ac}}{z_c}} = Y .$$

For this reason, the adequate confidence interval should be based on estimated conditional variance $\frac{n}{n-1} \sum_i \Delta_i^2$, where, for individual estimates (3), Δ_i equals

$$\Delta_i = \begin{cases} \frac{X}{x}\left(y_i - \frac{y}{x} x_i\right) & (11\text{-}1) \\ \sum_a \frac{X_a}{x_a}\left(y_{ai} - \frac{y_a}{x_a} x_{ai}\right) & (11\text{-}2) \\ \sum_a \sum_c \frac{X_a}{\sum_c Z_c \frac{x_{ac}}{z_c}} \frac{Z_c}{z_c}\left[y_{aci} - \frac{y_{ac}}{z_c} z_{ci} + \frac{y_a}{x_a}\left(x_{aci} - \frac{x_{ac}}{z_c} z_{ci}\right)\right] & (11\text{-}3) \end{cases}$$

Returning, now, to sampling without replacement from finite populations, we may, by analogy, hope, that the Δ_i's given by (11), when inserted into (6), will give better estimated variances than the Δ_i's given by (7). Asymptotically both considered methods of estimating the variance are equivalent, because for large n and (N-n)

$$\frac{X}{x} \approx \frac{N}{n}, \quad \frac{X_a}{x_a} \approx \frac{N}{n}, \quad \frac{X_a}{\sum_c Z_c \frac{x_{ac}}{z_c}} \approx 1, \quad \frac{Z_c}{z_c} \approx \frac{N}{n}.$$

For small samples, however, the improvement may be essential, generally in the sense that (11) will yield considerably greater estimated variance, on the average, than (7). Furthermore, without using (11), we cannot detect the possible negative effect of splitting x_i into too many components x_{ai}, (see [2], vol. II, p. 139).

Remark 3.1. In finite populations, relations (8) are reflected in such a manner that the regression of y_{aci} upon x_{aci} and z_{ci} passes through the origin. Exceptionally, they may be fulfilled precisely, for example if x_i equals 0 or 1, and if $x_i = 0$ implies $y_i = 0$.

Remark 3.2. Inserting (11 — 2) into (6), we obtain the estimated variance, which for the special case $n = 2$, and stratified sampling, has been derived in [5].

4. Normal theory approach

Assuming normality and (8), we shall prove that confidence intervals

$$X \frac{y}{x} \pm \frac{X}{x} t_\alpha \sqrt{\frac{n}{n-1} \sum_i \left(y_i - \frac{y}{x} x_i\right)^2}\,, \tag{12-1}$$

$$\sum X_a \frac{y_a}{x_a} \pm t_\alpha \sqrt{\frac{n}{n-1} \sum_i \left[\sum_a \frac{X_a}{x_a}\left(y_{ai} - \frac{y_a}{x_a} x_{ai}\right)\right]^2}\,, \tag{12-2}$$

$$\sum_a X_a \frac{\sum\limits_c Z_c \frac{y_{ac}}{z_c}}{\sum\limits_c Z_c \frac{x_{ac}}{z_c}} \pm$$

$$\pm t_\alpha \sqrt{\frac{n}{n-1} \sum_i \left\{\sum_a \sum_c \frac{X_a Z_c}{z_c \sum\limits_{c'} Z_{c'} \frac{x_{ac'}}{z_{c'}}} \left[y_{aci} - \frac{y_{ac}}{z_c} z_{ci} - \frac{y_a}{x_a}\left(x_{aci} - \frac{x_{ac}}{z_c} z_{ci}\right)\right]\right\}^2}\,, \tag{12-3}$$

where t_α denotes the critical value of Student's distribution with $n - 1$ degrees of freedom for significance level α, will cover Y with probability greater than $1 - \alpha$. In other words, confidence intervals (12) are "conservative".

The statistic $\sum\limits_{i=1}^{n} \left(y_i - \frac{y}{x} x_i\right)^2$ does not possess chi-square distribution and is not independent of (y, x). Nevertheless, the following general inequality holds:

Theorem. *If* $\{y_{aci}, x_{aci}, z_{ci}, a \in A, c \in C\}$, $i = 1, \ldots, n$, *is a sample from a multidimensional normal distribution, then*

$$\sum_{i=1}^{n} \left\{\sum_a \sum_c K_{ac} \left[y_{aci} - \frac{y_{ac}}{z_c} z_{ci} - \frac{y_a}{x_a}\left(x_{aci} - \frac{x_{ac}}{z_c} z_{ci}\right)\right]\right\}^2 \geqq \chi^2_{n-1} \sigma^2\,, \tag{13}$$

where

$$y_{ac} = \sum_{i=1}^{n} y_{aci}\,, \quad x_{ac} = \sum_{i=1}^{n} x_{aci}\,, \quad z_c = \sum_{i=1}^{n} z_{ci}\,, \quad a \in A\,, \quad c \in C\,,$$

K_{ac} *are arbitrary constants, which may depend on* y_{ac}, x_{ac}, z_c, $a \in A$, $c \in C$,

χ^2_{n-1} *is a random variable which is independent of* $\{y_{ac}, x_{ac}, z_c, a \in A, c \in C\}$, *and possesses the chi-square distribution with* $n - 1$ *degrees of freedom, and*

σ^2 *is the conditional variance of* $\sum_a \sum_c K_{ac} y_{aci}$ *w. r. t.* $\{x_{aci'}, z_{ci'}, a \in A, c \in A, i' = 1, \ldots, n\}$.

Proof. Let us choose an orthogonal transformation, given by the matrix $\{b_{ij}\}$ such that $b_{1j} = \frac{1}{\sqrt{n}}$, $j = 1, \ldots, n$, and denote

$$u_{aci} = \sum_{j=1}^{n} b_{ij} y_{acj}, \; v_{aci} = \sum_{j=1}^{n} b_{ij} x_{acj}, \; w_{ci} = \sum_{i=1}^{n} z_{cj}, \; a \in A, \; c \in C\,.$$

Obviously $u_{ac1} = \frac{y_{ac}}{\sqrt{n}}$, $v_{ac1} = \frac{x_{ac}}{\sqrt{n}}$, $w_{c1} = \frac{z_c}{\sqrt{n}}$. Then, first owing to

$$\sum_a \sum_c K_{ac} \left[u_{aci} - \frac{y_{ac}}{z_c} w_{ci} - \frac{y_a}{x_a} \left(v_{aci} - \frac{x_{ac}}{z_c} w_{ci} \right) \right] =$$

$$= \sum_{j=1}^{n} b_{ij} \left\{ \sum_a \sum_c K_{ac} \left[y_{acj} - \frac{y_{ac}}{z_c} z_{cj} - \frac{y_a}{x_a} \left(x_{acj} - \frac{x_{ac}}{z_c} z_{cj} \right) \right] \right\}, \quad i = 1, \ldots, n\,,$$

and to

$$\sum_{i=1}^{n} \sum_a \sum_c K_{ac} \left[y_{aci} - \frac{y_{ac}}{z_c} z_{ci} - \frac{y_a}{x_a} \left(x_{aci} - \frac{x_{ac}}{z_c} z_{ci} \right) \right] = 0\,,$$

the following well-known algebraic identity (see [7], p. 116)

$$\sum_{i=1}^{n} \left\{ \sum_a \sum_c K_{ac} \left[y_{aci} - \frac{y_{ac}}{z_c} z_{ci} - \frac{y_a}{x_a} \left(x_{aci} - \frac{x_{ac}}{z_c} z_{ci} \right) \right] \right\}^2 =$$

$$= \sum_{i=2}^{n} \left\{ \sum_a \sum_c K_{ac} \left[u_{aci} - \frac{y_{ac}}{z_c} w_{ci} - \frac{y_a}{x_a} \left(v_{aci} - \frac{x_{ac}}{z_c} w_{ci} \right) \right] \right\}^2 \tag{14}$$

will hold, and, secondly, random vectors $\{u_{aci}, v_{aci}, w_{ci}, a \in A, c \in C\}$, $i = 2, \ldots, n$,

1° will be independent one of each other,

2° will be independent of $\{y_{ac}, x_{ac}, z_c, a \in A, c \in C\}$,

3° will possess mean values

$$M(u_{aci}) = M(v_{aci}) = M(w_{ci}) = 0\,, \quad a \in A\,, \quad c \in C\,, \quad i = 2, \ldots, n\,,$$

4° will be governed by the same variance covariance matrix as random vectors $\{y_{aci}, x_{aci}, z_{ci}, a \in A, c \in C\}$, $i = 1, \ldots, n$.

From 1°, 2° and 3° it follows, that

$$\sum_{i=2}^{n}\left\{\sum_{a}\sum_{c}K_{ac}\left[u_{aci}-\frac{y_{ac}}{z_c}w_{ci}-\frac{y_a}{x_a}\left(v_{aci}-\frac{x_{ac}}{z_c}w_{ci}\right)\right]\right\}^2=$$
$$=\chi^2_{n-1}M\left\{\left(\sum_{a}\sum_{c}K_{ac}\left[u_{ac2}-\frac{y_{ac}}{z_c}w_{c2}-\right.\right.\right.$$
$$\left.\left.\left.-\frac{y_a}{x_a}\left(v_{ac2}-\frac{x_{ac}}{z_c}w_{c2}\right)\right]\right)^2\right|y_{ac},x_{ac},z_c,a\in A,c\in C\right\}. \tag{15}$$

It remains to be proved that the latter conditional mean value is always greater or equal to σ^2. As, according to 2°, random vectors $\{u_{ac2}, v_{ac2}, w_{c2}, a \in A, c \in C\}$ and $\{y_{ac}, x_{ac}, z_c, a \in A, c \in C\}$ are independent, we may write

$$M\left\{\left(\sum_{a}\sum_{c}K_{ac}\left[u_{ac2}-\frac{y_{ac}}{z_c}w_{c2}-\right.\right.\right.$$
$$\left.\left.\left.-\frac{y_a}{x_a}\left(v_{ac2}-\frac{x_{ac}}{z_c}w_{c2}\right)\right]\right)^2\right|y_{ac},x_{ac},z_c,a\in A,c\in C\right\}=$$
$$=M^*\left\{\left(\sum_{a}\sum_{c}K_{ac}\left[u_{ac2}-\frac{y_{ac}}{z_c}w_{c2}-\frac{y_a}{x_a}\left(v_{ac2}-\frac{x_{ac}}{z_c}w_{c2}\right)\right]\right)^2\right\}, \tag{16}$$

where the asterisk denotes that $y_{ac}, x_{ac}, z_c, a \in A, c \in C$, are treated as constants. The mean square of any normal random variable cannot be smaller than its conditional variance, i. e.

$$M^*\left\{\left(\sum_{a}\sum_{c}K_{ac}\left[u_{ac2}-\frac{y_{ac}}{z_c}w_{c2}-\frac{y_a}{x_a}\left(v_{ac2}-\frac{x_{ac}}{z_c}w_{c2}\right)\right]\right)^2\right\}\geqq$$
$$\geqq D^*\left\{\sum_{a}\sum_{c}K_{ac}\left[u_{ac2}-\frac{y_{ac}}{z_c}w_{c2}-\frac{y_a}{x_a}\left(v_{ac2}-\frac{x_{ac}}{z_c}w_{c2}\right)\right]\right\}\geqq$$
$$\geqq D^*\left\{\sum_{a}\sum_{c}K_{ac}u_{ac2}\mid v_{ac2},w_{c2},a\in A,c\in C\right\}. \tag{17}$$

Now, according to 4°,

$$D^*\{\sum_{a}\sum_{c}K_{ac}u_{ac}|v_{ac2},w_{c2},a\in A,c\in C\}=$$
$$=D^*\{\sum_{a}\sum_{c}K_{ac}y_{aci}|x_{aci},z_{ci},a\in A,c\in C\}. \tag{18}$$

Finally, as vectors $\{y_{aci}, x_{aci}, z_{ci}, a \in A, c \in C\}, i = 1, \ldots, n$, are independent, we may write

$$D^*\{\sum_{a}\sum_{c}K_{ac}y_{aci}|x_{aci},z_{ci},a\in A,c\in C\}=$$
$$=D\{\sum_{a}\sum_{c}K_{ac}y_{aci}|x_{aci'},z_{ci'},a\in A,c\in C,i'=1,\ldots,n\}=\sigma^2. \tag{19}$$

By the chain of relations from (15) up to (19) the proof is completed.

In the special case, where the conditions (8 — 3) are fulfilled and

$$K_{ac} = \frac{X_a Z_c}{z_c \sum_{c'} Z_{c'} \frac{x_{ac'}}{z_{c'}}},$$

σ^2 becomes the conditional variance of

$$\sum_a X_a \frac{\sum_c Z_c \frac{y_{aci}}{z_c}}{\sum_c Z_c \frac{x_{aci}}{z_c}} \tag{20}$$

for fixed $\{x_{aci}, z_{ci}, a \in A, c \in C, i = 1, \ldots, n\}$. The estimate given in (3 — 3) and (12 — 3) is the sum of the observations (20). Observations (20) are independent (for fixed $\{x_{aci}, z_{ci}, a \in A, c \in C, i = 1, \ldots, n\}$) and, as we have seen in § 3, their sum is an unbiassed estimator of Y. Inequality (13) shows that the estimated variance used in (12 — 3) is greater or equal to an estimator of variance that is (a) unbiassed, (b) chi-square distributed, and (c) independent of the sum of observations (20). Consequently, the above statement concerning the confidence interval (12 — 3) is really true. Intervals (12 — 1) and (12 — 2) need no specific consideration, as they are special cases of the interval (12 — 3).

Remark 4.1. Our results could be generalized for the case where the same intraclass correlation is present, as in sampling without replacement. The only modification needed is to replace the coefficients t_α in (12) by $t_\alpha \sqrt{1 - \frac{n}{N}}$. Thus we obtain confidence intervals corresponding to simple random sampling without replacement.

5. Comparison with Fieller's method

The confidence interval (12 — 1), corresponding to the simple ratio estimate, can be compared with that provided by Fieller's method. We shall show that Fieller's interval is always longer.

Let us have a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ from a two-dimensional normal distribution, and put $\xi = Mx_i$, $\eta = My_i$. Fieller's solution rests upon the obvious fact, that the random variables $y_i - \frac{\eta}{\xi} x_i$ have mean values 0, so that

$$t = \frac{y - \frac{\eta}{\xi} x}{\sqrt{\frac{n}{n-1} \sum_{i=1}^{n} \left[y_i - \bar{y} - \frac{\eta}{\xi} (x_i - \bar{x}) \right]^2}} \tag{21}$$

is governed by Student's distribution with $n - 1$ degrees of freedom. Thus for any $\frac{\eta}{\xi}$, with probability $1 - \alpha$, it holds that $|t| \leq t_\alpha$, where t_α is the critical value of Student's distribution with $n - 1$ degrees of freedom for significance level α. By the inequality $|t| \leq t_\alpha$, where t is given by (21), a certain confidence region for $\frac{\eta}{\xi}$ is defined, which covers the true value of $\frac{\eta}{\xi}$ with probability $1 - \alpha$. Fieller has shown in [1], that, under the condition

$$1 - t_\alpha^2 c_1^2 > 0 \tag{22}$$

the confidence region is equal to the interval with endpoints

$$\frac{y}{x} \frac{1 - t_\alpha^2 c_{12} \pm \sqrt{(1 - t_\alpha^2 c_{12})^2 - (1 - t_\alpha^2 c_1^2)(1 - t_\alpha^2 c_2^2)}}{1 - t_\alpha^2 \, c_1^2} \tag{23}$$

where

$$c_1^2 = \left(\frac{1}{x}\right)^2 \frac{n}{n - 1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \,,$$

$$c_2^2 = \left(\frac{1}{y}\right) \frac{n}{n - 1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \,,$$

$$c_{12} = \frac{1}{xy} \frac{n}{n - 1} \sum_{i=1}^{n} (x_i - x)(y_i - \bar{y}) \,.$$

If (22) does not hold, it may be proved, that the Fieller's confidence region equals the whole line, or half-line, or complement of a finite interval. The square of the length of the interval with end-points (23), say d_F, equals

$$d_F = 4 \left(\frac{y}{x}\right)^2 \frac{(1 - t_\alpha^2 c_{12})^2 - (1 - t_\alpha^2 c_1^2)(1 - t_\alpha^2 c_2^2)}{(1 - t_\alpha^2 c_1^2)^2} =$$

$$= 4 \left(\frac{y}{x}\right)^2 t_\alpha^2 \, \frac{(1 - t_\alpha^2 c_1^2) \left(c_2^2 - \frac{c_{12}^2}{c_1^2}\right) + \left(c_1 - \frac{c_{12}}{c_1}\right)^2}{(1 - t_\alpha^2 c_1^2)^2} \,. \tag{24}$$

Let us compare this length with the length of the interval obtained from (12 — 1) after dividing by X:

$$\frac{y}{x} \pm \frac{1}{x} t_\alpha \sqrt{\frac{n}{n - 1} \sum_{i=1}^{n} \left(y_i - \frac{y}{x} x_i\right)^2} \,. \tag{25}$$

The square of the length of the interval (25), say d_A, equal

$$d_A = 4\left(\frac{t_\alpha}{x}\right)^2 \frac{n}{n-1} \sum_{i=1}^{n} \left(y_i - \frac{y}{x} x_i\right)^2 = 4\left(\frac{y}{x}\right)^2 t_\alpha^2 (c_2^2 - 2c_{12} + c_2^2) =$$

$$= 4\left(\frac{y}{x}\right)^2 t_\alpha^2 \left[c_2^2 - \frac{c_{12}^2}{c_1^2} + \left(c_1 - \frac{c_{12}}{c_1}\right)^2\right] \leqq$$

$$\leqq 4\left(\frac{y}{x}\right)^2 t_\alpha^2 \frac{(1 - t_\alpha^2 c_1^2)\left(c_2^2 - \frac{c_{12}^2}{c_1^2}\right) + \left(c_1 - \frac{c_{12}}{c_1}\right)^2}{(1 - t_\alpha^2 c_1^2)^2} = d_F\,, \qquad (26)$$

i. e. the Fieller's interval (23) is always longer than the usual interval (25).

6. Some other methods of sampling-estimating

Ratio estimates are generally biassed. In the paper [9], however, it is shown, that the bias of the simple ratio estimate $X \frac{y}{x}$ can be completely removed by the following method of sampling. In the first phase, we select one element with probabilities $\frac{x_i}{X}$, $i = 1, \ldots, N$, in the second phase, we select $n - 1$ element from remaining $N - 1$ elements by simple random sampling. The estimated variance, of course, at least for small n, must be changed accordingly. Following Yeates and Grundy, (see [4]), we get the estimated variance in the form

$$\left(\frac{X}{x}\frac{n}{N}\right)^2 \sum_i \sum_j \left(y_i - \frac{y}{x} x_i - y_j + \frac{y}{x} x_j\right)^2 \left(\frac{\pi_i \pi_j}{\pi_{ij}} - 1\right), \qquad (27)$$

where π_i denotes the probability that the i-th element will be included in sample, and π_{ij} denotes the probability that the i-th and j-th element both will be included in sample simultaneously. For the "two-phase" sampling described above, it may be easily shown that

$$\pi_i = \frac{x_i}{X}\frac{N-n}{N-1} + \frac{n-1}{N-1}\,, \quad i = 1, \ldots, N\,, \qquad (28)$$

$$\pi_{ij} = \frac{x_i + x_j}{X}\frac{n-1}{N-2}\frac{N-n}{N-1} + \frac{n-1}{N-1}\frac{n-2}{N-2}\,, \quad i, j = 1, \ldots, N\,. \qquad (29)$$

Let us mention briefly two other methods of sampling-estimating. First method: We choose, n-times independently, one element with probabilities $\frac{x_i}{X}$, $i = 1, \ldots, N$, and use the following point and interval estimate for Y:

$$\frac{X}{n}\left[\sum_{i}^{n_0}\frac{y_i}{x_i}+(n-n_0)\frac{\sum\limits_{i}^{n_0} y_i}{\sum\limits_{i}^{n_0} x_i}\right] \pm t_\alpha \sqrt{\frac{X^2}{n(n-1)}\sum_{i}^{n}\left[\frac{y_i}{x_i}-\frac{1}{n}\sum_{i}^{n}\frac{y_i}{x_i}\right]^2}\,, \tag{30}$$

where $\sum\limits_i^n$ and $\sum\limits_i^{n_0}$ denote the sums extended over all sampled elements and over distinct sampled elements, respectively, and n_0 denotes the number of distinct elements in the sample; (some of elements may appear twice or more times in the sample).

Second method: We carry out N independent experiments, the i-th of which decides with probability $c\,\frac{x_i}{X}$ and $1-c\,\frac{x_i}{X}$ if or if not the i-th element will be included in the samplo, so that the number n of elements included in sample will be a random variable with $M(n)=c$. Then we use the following point and interval estimates for Y:

$$\frac{X}{n}\sum_i\frac{y_i}{x_i}\pm t_\alpha\sqrt{\frac{X^2}{n(n-1)}\sum_i\left(\frac{y_i}{x_i}-\frac{1}{n}\sum_i\frac{y_i}{x_i}\right)^2\left(1-\frac{cx_i}{X}\right)}\,. \tag{31}$$

7. Stratification and subsampling

We shall only touch this topic very briefly. Let us label the strata, and statistics referred to strata, by the subscript $h=1,\ldots,L$. If the stratified sampling is proportionate, i. e. $\frac{N_h}{n_h}=\frac{N}{n}$, $h=1,\ldots,L$, then formulae (3) need no change, and the estimated variance (6) should be replaced by

$$\sum_{h=1}^{L}\frac{(N_h-n_h)n_h}{N_h(n_h-1)}\left[\sum_i\Delta_{hi}^2-\frac{1}{n_h}\left(\sum_i\Delta_{hi}\right)^2\right], \tag{32}$$

where

$$\Delta_{hi}=\frac{N_h}{n_h}\left(y_{hi}-\frac{y}{x}x_{hi}\right),\ \text{etc.},$$

in accordance with (7).

If the stratified sampling fails to be proportionate, then the sample totals y, x, y_a etc. in formulas (3) should be replaced by

$$\sum_{h=1}^{L}\frac{N_h}{n_h}\sum_i y_{hi}\,,\quad \sum_{h=1}^{L}\frac{N_h}{n_h}\sum_i x_{hi}\,,\quad \sum_{h=1}^{L}\frac{N_h}{n_h}\sum_i y_{hai}\,,\quad \text{etc.}, \tag{33}$$

respectively. Formula (32) needs no change. The theorem of § 4 could be generalized for stratified sampling in an obvious manner. In corresponding statements concerning confidence, of course, will denote the critical value based on the generalized Student's distribution, (see [8]).

On some occasions, other ratio estimates may be useful in the case of stratified sampling:

$$\sum_{h=1}^{L} X_h \frac{y_h}{x_h}, \tag{34-1}$$

$$\sum_{h=1}^{L}\sum_{a} X_{ha} \frac{y_{ha}}{x_{ha}}, \tag{34-2}$$

$$\sum_{a} X_a \frac{\sum\limits_{h=1}^{L}\sum\limits_{c} Z_{hc} \dfrac{y_{hac}}{z_{hc}}}{\sum\limits_{h=1}^{L}\sum\limits_{c} Z_{hc} \dfrac{x_{hac}}{z_{hc}}}. \tag{34-3}$$

The corresponding estimated variances equal

$$\sum_{h=1}^{L} \frac{(N_h - n_h)n_h}{N_h(n_h - 1)} \sum_{i} \Delta_{hi}^2, \tag{35}$$

where

$$\Delta_{hi} = \frac{X_h}{x_h}\left(y_{hi} - \frac{y_h}{x_h} x_{hi}\right), \quad \text{etc.,}$$

in accordance with (11). The estimates (34) are advantageous in such cases, where the ratios $\frac{Y_h}{X_h}$, etc., vary substantially from stratum to stratum. On the other hand, they may be heavily biassed, unless "two-phase" sampling described in § 6, is used in each stratum. Moreover, if $\frac{Y_h}{X_h}$ does not vary very much, and if the number of strata is large and n_h's are small, then the variance of estimates (34) may even be greater than the variance of estimates (3).

Finally, when more stages are involved in our sampling plan, then the only modification needed is to replace the values y_i, y_{ai}, y_{aci}, and, possibly, values x_i, x_{ai}, x_{aci}, z_{ci} by their estimates obtained from subsampling. After this adaptation formulae (12) are valid, in a conservative manner, for multistage sampling, too. Subsampling rates are often chosen so that the estimate turns out to be an unweighted sum of the ultimate observations. For example, let the hi-th element contains M_{hi} subsampling elements, m_{hi} of which we select by subsampling. If we put

$$m_{hi} = \frac{1}{k}\frac{M_{hi}X_h}{x_h}, \quad h = 1, \ldots, L; \; i = 1, \ldots, N_h, \tag{36}$$

then the estimate (34 — 1) turns out to equal ky', where y' denotes the sum of ultimate observations.

8. Some concluding remarks

The above discussion, based on a combination of the large—sample, regression and normal theory approaches, will also apply to other possible types of ratio estimates. Some important problems, however, remain to be open. Let us mention two of them: (a) For what set of normal distributions is it true that the confidence intervals (12) cover Y with a probability greater or equal $1 - \alpha$? (b) How can we simplify the computations of estimated variances considerably? (The estimated variances, derived in this paper, despite their relative simplicity, are to complicated for daily practice, particularly, when hundreds of items are tabulated simultaneously.)

REFERENCES

[1] *Fieller, E. C.,* A fundamental formula in the statistics of biological assay, and some applications, Quart. Journ. Pharm. 17 (1944), 117—123.

[2] *Hansen, M. H., Hurwitz, W. N., Madow, W. G.,* Sample Survey Methods and Theory, vol. I, II, New York, London, 1953.

[3] *Hansen, M. H., Hurwitz, W. N.,* On the theory of sampling from finite populations, Ann. Math. Stat. 14 (1943), p. 393.

[4] *Yeates, F., Grundy, P. M.,* Selection without replacement from within strata with probability proportional to size, Journ. Royal. Stat. Soc., B, XV, 1953, 253—261.

[5] *Keyfitz, W.,* Estimates of sampling variance where two units are selected from each stratum, Journ. Am. Stat. Ass. 52 (1957), 503—510.

[6] *Madow, W. G.,* On the limiting distributions of estimates based on samples from finite universes, Ann. Math. Stat. 19 (1948), 535—545.

[7] *Cramér, H.,* Mathematical methods of statistics, Princeton 1946.

[8] *Hájek, J.,* Nerovnosti pro zobecněné Studentovo rozdělení a jejich použití, Časopis pro pěstování matematiky 82 (1957), 182—194.

[9] *Hájek, J.,* Representativní výběr skupin metodou dvou fází, Statistický obzor, XXIX, 1949, 384—394.

Souhrn

O TEORII POMĚROVÝCH ODHADŮ

JAROSLAV HÁJEK

(Došlo dne 29. listopadu 1957.)

V této práci je teorie poměrových odhadů budována kombinací metody velkých výběrů (asymptotické metody), metody lineární regrese, a metody založené na předpokladu normálního rozdělení. Například, odhady rozptylů, jak vyplývají z asymptotické teorie, jsou pro konečný rozsah výběru přizpůsobovány pomocí vhodných předpokladů o regresi. Obdržené výsledky jsou pak ověřovány za předpokladu normálního rozdělení, a ukazuje se, že v poměrech příznačných pro výběrové šetření interval spolehlivosti, sestrojený na základě Studentova rozdělení, pokrývá správnou hodnotu s větší pravděpodobností, než je ta, které odpovídá použití koeficient t_α. Je ukázáno, že zmíněný interval spolehlivosti je vždy kratší než interval spolehlivosti Fiellerův.

Práce dále pojednává o tom, jak získat odhad rozptylu poměrného odhadu, když (a) je tento poměrový odhad velmi složitý, nebo je-li (b) výběr je stratifikovaný a vícestupňový, či je-li (c) výběr je prováděn metodou dvou fází, které činí poměrový odhad nestranným.

Резюме

О ТЕОРИИ ,,ПРОПОРЦИЙНЫХ“[1]) ОЦЕНОК

ЯРОСЛАВ ГАЕК (Jaroslav Hájek)

(Поступило в редакцию 29/XI 1957 г.)

В настоящей работе развивается теория пропорционых оценок комбинированием метода больших выборок (асимптотический метод), метода линейной регрессии и метода, основанного на предположении нормального распределения. Например, оценки дисперсий, полученные по асимптотическому методу, приспособляются для конечного объема выборки при помощи подходящих предположений о регрессии. Полученные результаты затем проверяются при условии нормального распределения, и оказывается, что при обстоятельствах, характерных для выборочного исследования, доверительный интервал, построенный на основании распределения Стьюдента, покрывает точное значение с большей вероятностью, чем вероятность, которой соответствует примененный коэффициент t_α. В статье показано, что указанный доверительный интервал короче доверительного интервала Филлера.

Далее в работе рассматривается вопрос, как получить оценку дисперсии пропорцийной оценки в случае, когда (a) эта относительная оценка является очень сложной, или когда (b) выборка является выборкой по группам (типической) и многоступенчатой, или же когда (c) выборка проводится методом двух фаз, который делает относительную оценку несмещенной.

[1]) По англ. ,,ratio estimate“.