CHAPTER 9

Ratios and Ratio Estimators

9.1 Introduction

For the math (y_i) and verbal (x_i) scores examined in Chapter 2, the ratio of the math to verbal score, $r_i = y_i/x_i$, or their ratios to the total score, $r_i = y_i/(x_i + y_i)$ and $r_i = x_i/(x_i + y_i)$, are frequently of importance. Similarly, the ratios of the employment and establishment figures examined in Chapters 7 and 8 are of practical interest. Per capita figures of agricultural productions, Gross National Products (GNPs), energy consumptions, and similar characteristics for each country are obtained by dividing their totals (y_i) by the population sizes (x_i). The average of these types of ratios $\bar{R} = \Sigma r_i/N$ can be estimated from the sample mean $\bar{r} = \Sigma r_i/n$ of the ratios. Just as the sample mean $\bar{y}$ is unbiased for the population mean $\bar{Y}$, this average of the sample ratios is unbiased for $\bar{R}$. Its variance and estimator of variance are easily obtained from (2.12) and (2.14) by replacing y_i with r_i.

In contrast to the above average of the ratios, the ratio of the means or totals of two characteristics is given by $R = \bar{Y}/\bar{X} = Y/X$. Ratio of the average of the math scores to the average of the total (math + verbal) scores is an example. Another illustration is provided by the ratio of the average of the employment figures to the average of the establishments of the type examined in Chapter 7. From the observations (x_i, y_i), $i = 1, 2, \ldots, n$ of a random sample selected without replacement from the N population units, $\hat{R} = \bar{y}/\bar{x}$ is an estimator for R. In single-stage as well as multistage sampling, several estimators take the form of the ratio of two means. Chapters 7 and 8 have shown that in the case of clusters of unequal size, the population total or mean can be estimated with high precision by this type of ratio, with x_i representing the cluster size.

In the above type of illustrations, x_i can also provide concomitant or supplementary information on y_i. If x_i and y_i are positively correlated, $\bar{Y}/\bar{X}$ can be expected to be close to the sample ratio $\bar{y}/\bar{x}$ and

$\bar{Y}$ can be estimated from $(\bar{y}/\bar{x})\bar{X}$ with high precision. Further gains in precision can be obtained by combining this procedure with stratification. P.S.R.S. Rao (1998a) presents a summary of the ratio estimation procedures.

9.2 Bias and variance of the sample ratio

The sample ratio $\hat{R}$ is biased for R even when x_i and y_i are uncorrelated. From the **linearization** method in Appendix A9, this bias becomes negligible for large n. In such a case, the MSE of $\hat{R}$, which becomes the same as its variance, is approximately given by

$$V(\hat{R}) = (1-f)S_d^2/n\bar{X}^2 \tag{9.1}$$

where

$$\begin{aligned} S_d^2 &= \sum_1^N (y_i - Rx_i)^2/(N-1) \\ &= S_y^2 + R^2S_x^2 - 2RS_{xy} = S_y^2 + R^2S_x^2 - 2R\rho S_xS_y. \end{aligned}$$

An estimator of S_d^2 is obtained from

$$s_d^2 = \sum_1^n (y_i - \hat{R}x_i)^2/(n-1) = s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}.$$

Hence, one may estimate (9.1) from

$$v(\hat{R}) = (1-f)s_d^2/n\bar{x}^2. \tag{9.2}$$

Example 9.1. Heights and weights: For the sake of illustration, Table 9.1 presents heights, weights, and systolic and diastolic blood pressures of $N = 15$ candidates along with their means, variances, standard deviations, and correlations. Consider height (x_i) and weight (y_i). The population ratio of the averages of the weights and heights is $R = 155.53/67.73 = 2.3$. Further, $S_d^2 = 187.12 + (2.3)^2(8.49) - 2(2.3)(0.62)(2.91)(13.68) = 118.5$. If one considers a random sample of $n = 5$, from (9.1), $V(\hat{R}) = 0.0034$ and hence S.E.$(\hat{R}) = 0.0581$. In a random sample of size five, units (4, 6, 9, 12, 15) were selected. From the observations of these units, the means and variances of heights and weights are $(\bar{x}, \bar{y}) =$

(66, 158.6) and $(s_x^2, s_y^2) = (24, 314.8)$. The estimate of R is $\hat{R} = 158.6/66 = 2.4$. Further, $s_{xy} = 84.5$ and the sample correlation coefficient is 0.97. With these figures, $s_d^2 = [314.8 + 5.76(24) - 2(2.4)(84.5)] = 47.44$. Now, from (9.2), $v(\hat{R}) = 0.0015$ and S.E.$(\hat{R}) = 0.0381$.

9.3 Confidence limits for the ratio of means

Approximate confidence limits for R can be obtained from $\hat{R} \pm Z\sqrt{v(\hat{R})}$. If (x_i, y_i), $i = 1, 2, \ldots, n$, are considered to be observations from a bivariate normal distribution, Z is replaced by the $(1 - \alpha)$ percentile of the t-distribution with $(n - 1)$ d.f.

Alternative limits for R may be obtained by adopting Fieller's (1932) approach. If (x, y) follow a bivariate normal distribution, $\bar{d} = (\bar{y} - R\bar{x})$ follows a normal distribution with mean zero and variance $V(\bar{d}) = V(\bar{y}) + R^2V(\bar{x}) - 2R\text{Cov}(\bar{y}, \bar{x})$. For sampling from a finite population, an unbiased estimator of this variance for a *given* R is $v(\bar{d}) = [(1 - f)/n](s_y^2 + R^2s_x^2 - 2Rs_{xy})$. Now, confidence limits for R are obtained by expressing $t_{n-1}^2 = (\bar{y} - R\bar{x})^2/v(\bar{d})$ as a quadratic equation in R and finding its two roots.

Example 9.2. Confidence limits for the ratio of means: This example examines the limit for the ratio of the heights and weights of Example 9.1. For the probability of 0.95, $t_4 = 2.7764$. Since t_4 S.E.$(\hat{R}) = 1.06$, confidence limits for R are given by 2.4 ± 1.06, that is, (1.34, 3.46). With the normal approximation, Z S.E.$(\hat{R}) = 0.07$, and the limits are given by 2.4 ± 0.07, that is, (2.33, 2.47). The limits from the t-distribution are wider; that is, they are *conservative*.

For Fieller's method, $(2.7764)^2(2/15)[314.8 + 24R^2 - 2R(84.5)] = (158.6 - 66R)^2$. Simplifying this equation, $R^2 - 4.87R + 5.73 = 0$. The roots of this equation are (1.99, 2.88). In this example, these limits are also wider compared with the limits obtained above from the normal approximation.

9.4 Ratio estimators for the mean and total

If x_i and y_i are positively correlated and $\bar{X}$ is known, to estimate $\bar{Y}$, one may consider the ratio estimator:

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{R}\bar{X}. \tag{9.3}$$

Table 9.1. Heights, weights, and blood pressures.

Unit	Height	Weight	Systolic Pressure	Diastolic Pressure
1	65	140	120	90
2	63	145	125	80
3	65	150	140	95
4	68	148	140	84
5	70	150	135	86
6	72	160	150	96
7	70	175	150	90
8	70	155	140	82
9	72	180	160	100
10	68	175	160	95
11	64	155	140	84
12	66	135	130	80
13	68	145	130	90
14	65	150	140	85
15	70	170	160	100
Mean	67.73	155.53	141.33	89.13
Variance	8.49	187.12	158.81	46.55
S.D.	2.91	13.68	12.60	6.82
Correlations				
	1			
	0.62	1		
	0.64	0.92	1	
	0.54	0.69	0.70	1

The bias of $\hat{\bar{Y}}_R$, $B(\hat{\bar{Y}}_R) = E(\hat{\bar{Y}}_R) - \bar{Y} = \bar{X}[E(\hat{R}) - R]$ becomes negligible for large n. The large sample MSE or variance of $\hat{\bar{Y}}_R$ is given by

$$V(\hat{\bar{Y}}_R) = E(\hat{\bar{Y}}_R - \bar{Y})^2 = \bar{X}^2 E(\hat{R} - R)^2$$
$$= \frac{(1-f)}{n} S_d^2 = \frac{(1-f)}{n}(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y). \quad (9.4)$$

From (2.12) and (9.4),

$$V(\bar{y}) - V(\hat{\bar{Y}}_R) = \frac{(1-f)}{n} R^2 S_x^2 \left(2\rho \frac{C_y}{C_x} - 1\right), \quad (9.5)$$

where $C_x = S_x/\bar{X}$ and $C_y = S_y/\bar{Y}$ are the coefficients of variation of the two characteristics. Hence, the ratio estimator has smaller variance

than the sample mean if $\rho > (C_x/2C_y)$. Large positive correlation is helpful for the ratio estimator.

An estimator for (9.4) is given by

$$v(\hat{\bar{Y}}_R) = \frac{(1-f)}{n}s_d^2 = \frac{(1-f)}{n}(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}). \qquad (9.6)$$

An alternative estimator is $v^1(\hat{\bar{Y}}_R) = \bar{X}^2 v(\hat{R}) = (\bar{X}/\bar{x})^2(1-f)s_d^2/n$.

The ratio estimator for the total Y is $\hat{Y}_R = N\hat{\bar{Y}}_R$ and its large sample variance is given by $V(\hat{Y}_R) = N^2V(\hat{\bar{Y}}_R)$, which can be estimated from $v(\hat{Y}_R) = N^2v(\hat{\bar{Y}}_R)$ or $V(\hat{Y}_R) = N^2v^1/(\hat{\bar{Y}}_R)$.

Example 9.3. Gain in precision for the ratio estimator: For the illustration in Example 9.1 with $n = 5$, $V(\hat{\bar{Y}}_R) = (2/3)(118.5)/5 = 15.8$ and hence S.E.$(\hat{\bar{Y}}_R) = 3.97$. In this case, $V(\bar{y}) = (2/3)(187.12)/5 = 24.95$. Thus, the gain in precision for the ratio estimator relative to the sample mean is $(24.95 - 15.8)/15.8 = 0.58$ or 58%.

9.5 Confidence limits for the mean and total

With the five observed units in the sample, $\hat{\bar{Y}}_R = 2.4(67.73) = 162.55$, $v(\hat{\bar{Y}}_R) = [(15-5)/(15\times 5)](47.44) = 6.33$ and S.E.$(\hat{\bar{Y}}_R) = 2.52$. Now, the 95% confidence limits for $\bar{Y}$ are given by $162.55 - 1.96(2.52) = 157.61$ and $162.55 + 1.96(2.52) = 167.49$.

From the above sample of five units, $\hat{Y}_R = 15(162.55) = 2438.25$ and S.E.$(\hat{Y}_R) = 15(2.52) = 37.8$. Hence, the 95% confidence limits for Y are given by $2438.25 - 1.96(37.8) = 2364.16$ and $2438.25 + 1.96(37.8) = 2512.34$. These limits can also be obtained by multiplying the above limits for the mean by the population size 15.

9.6 Differences between ratios, means, or totals

A common supplementary variable

Consider estimating the difference in the blood pressure–weight ratios for the systolic and diastolic measurements. With the subscripts 1 and 2 denoting the two types of measurements, these ratios are $R_1 = \bar{Y}_1/\bar{X}$ and $R_2 = \bar{Y}_2/\bar{X}$. From a sample of n units, $\hat{R}_1 = \bar{y}_1/\bar{x}$ and $\hat{R}_2 = \bar{y}_2/\bar{x}$ provide the estimators for these ratios. For large samples, the difference $\hat{R}_1 - \hat{R}_2$ is approximately unbiased for $R_1 - R_2$. With $d_{1i} = y_{1i} - \hat{R}_1 x_i$,

$d_{2i} = y_{2i} - \hat{R}_2 x_i$, and $s_d^2 = \Sigma_i (d_{1i} - d_{2i})^2/(n-1)$, an estimate of its variance is given by

$$v(\hat{R}_1 - \hat{R}_2) = \frac{(1-f)}{n\bar{x}^2} s_d^2. \tag{9.7}$$

The ratio estimators for $\bar{Y}_1$ and $\bar{Y}_2$ are given by $\hat{\bar{Y}}_{R1} = \hat{R}_1 \bar{X}$ and $\hat{\bar{Y}}_{R2} = \hat{R}_2 \bar{X}$, which are approximately unbiased. The difference $(\bar{Y}_1 - \bar{Y}_2)$ can now be estimated from $(\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2})$, and an estimator for its large sample variance is provided by $(1-f)s_d^2/n$ with the above expression for s_d^2.

Example 9.4. Systolic and diastolic pressures: For the five sample units (4, 6, 9, 12, 15) considered in Example 9.1, the mean and variance of the systolic and diastolic pressures, respectively, are $(\bar{y}_1, s_1^2) = (148, 170)$ and $(\bar{y}_2, s_2^2) = (92, 88)$. The sample covariance of these two types of measurements is $s_{12} = 120$. The mean and variance of the weights are $(\bar{x}, s_x^2) = (158.6, 314.8)$.

Ignoring the information on the weights, an estimator for the difference of the means of the systolic and diastolic pressures is $(\bar{y}_1 - \bar{y}_2) = 56$. Following Section 3.5, $v(\bar{y}_1 - \bar{y}_2) = [(15-5)/(15 \times 5)](170 + 88 - 240) = 2.4$, and hence S.E.$(\bar{y}_1 - \bar{y}_2) = 1.55$.

From the sample, ratios of the means of the systolic and diastolic pressures to the mean of the weights are $\hat{R}_1 = 148/158.6 = 0.9332$ and $\hat{R}_2 = 92/158.6 = 0.5801$. Since $\bar{X} = 155.53$, the ratio estimators for the population means of the two types of pressures are $\hat{\bar{Y}}_{R1} = (0.9332)155.53 = 145.14$ and $\hat{\bar{Y}}_{R2} = (0.5801)155.53 = 90.22$. Hence, $\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2} = 54.92$. From the sample observations, $s_d^2 = 9.5797$. Now, from (9.7), $v(\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2}) = 1.28$ and S.E.$(\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2}) = 1.13$.

The actual difference of the population means is $141.33 - 89.13 = 52.2$, and the estimates from the sample means as well as the ratio estimation are not too far from this figure.

From the sample means, an estimate for the difference of the population totals is $15(56) = 840$, and it has an S.E. of $15(1.55) = 23.25$. From the ratio method, an estimate for this difference is $15(54.92) = 819.3$ which has an S.E. of $15(1.13) = 16.95$.

Different supplementary variables

Denote the initial observations on the supplementary and main variables of a population of N units by (x_{1i}, y_{1i}) and their observations on another occasion by (x_{2i}, y_{2i}). From a random sample of size n from the

N units, the population ratios on the two occasions, $R_1 = \bar{Y}_1/\bar{X}_1$ and $R_2 = \bar{Y}_2/\bar{X}_2$, can be estimated from $\hat{R}_1 = \bar{y}_1/\bar{x}_1$ and $\hat{R}_2 = \bar{y}_2/\bar{x}_2$. With $d_{1i} = y_{1i} - \hat{R}_1 x_{1i}$ and $d_{2i} = y_{2i} - \hat{R}_2 x_{2i}$, the variance of $(\hat{R}_1 - \hat{R}_2)$ can be estimated from

$$v(\hat{R}_1 - \hat{R}_2) = \frac{(1-f)}{n} \sum_i (d_{1i}/\bar{x}_1 - d_{2i}/\bar{x}_2)^2/(n-1). \qquad (9.8)$$

The ratio estimators for $\bar{Y}_1$ and $\bar{Y}_2$ are now given by $\hat{\bar{Y}}_{R1} = \hat{R}_1\bar{X}_1$ and $\hat{\bar{Y}}_{R2} = \hat{R}_2\bar{X}_2$. The variance of $(\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2})$ can be estimated from

$$v(\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2}) = \frac{(1-f)}{n} \sum_i (d_{1i} - d_{2i})^2/(n-1). \qquad (9.9)$$

Example 9.5. Reduction in the systolic pressure after an exercise program: Represent the initial measurements of the 15 sample units in Table 9.1 by the subscript 1. The initial means and variances of the weights and systolic pressures for the five sample units (4, 6, 9, 12, 15), respectively, are (158.6, 314.8) and (148, 170).

One can represent the measurements after a physical fitness program by the subscript 2. With this notation, the population means of the weights and systolic pressures in this case can be denoted by $(\bar{X}_2, \bar{Y}_2)$. For the five sample units, as an illustration, consider (x_{2i}, y_{2i}) = (140, 135), (152, 145), (170, 155), (130, 128), and (165, 158). From these observations, the sample means and variances of the weights and systolic pressures now are (151.4, 279.8) and (144.2, 163.7). The sample covariance of the systolic pressures on the two occasions is $s_{y12} = 165.5$.

An unbiased estimator of $(\bar{Y}_1 - \bar{Y}_2)$ is $(\bar{y}_1 - \bar{y}_2) = 148 - 144.2 = 3.8$, that is, the systolic pressure has decreased by 3.8 after the fitness program. The sample variance of this estimator is $v(\bar{y}_1 - \bar{y}_2) = (2/15)[170 + 163.7 - 2(165.5)] = 0.36$, and hence S.E.$(\bar{y}_1 - \bar{y}_2) = 0.6$.

As before, $\hat{R}_1 = 148/158.6 = 0.9332$ and $\hat{\bar{Y}}_{R1} = 0.9332(155.53) = 145.14$. For the observations after the program, $\hat{R}_2 = 144.2/151.4 = 0.9524$. If $\bar{X}_2 = 150$, $\hat{\bar{Y}}_{R2} = 0.9524(150) = 142.86$. Now, the estimate of $(\bar{Y}_1 - \bar{Y}_2)$ is $145.14 - 142.86 = 2.28$. From (9.9), $v(\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2}) = (2/15)(0.4183) = 0.0558$ and hence S.E.$(\hat{\bar{Y}}_{R1} - \hat{\bar{Y}}_{R2}) = 0.24$.

9.7 Regression through the origin and the BLUEs

This section examines the classical regression through the origin for **infinite** populations, and recognizes its role in providing motivation for ratio-type estimators.

Least squares

Regressions of agricultural yields on acreage, industrial productions on employee sizes, household expenses on family sizes, and the like can be represented by the model

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1,\ldots,n. \tag{9.10}$$

The expectation of the residual or error ε_i at a given x_i, $E(\varepsilon_i | x_i)$, is assumed to be zero. For this **regression through the origin**, the mean of y_i at x_i, $E(y_i | x_i)$ is given by βx_i; it becomes zero at $x_i = 0$. The plot of $E(y_i | x_i)$ against x_i is a straight line going through the origin $(x_i, y_i) = (0, 0)$. The coefficient β is the **slope** of this line, which is the rate of change of $E(y_i | x_i)$ with respect to x_i.

For some applications, the variance of y_i at a given x_i, $V(y_i | x_i)$, which is the same as the variance of ε_i at a given x_i, $V(\varepsilon_i | x_i)$, is found to be proportional to x_i; that is, it is of the form $\sigma^2 x_i$. The variance of the transformed variable $y_i/x_i^{1/2}$ becomes σ^2. The residuals ε_i and ε_j at x_i and x_j are assumed to be uncorrelated. Now, minimization of $\Sigma_1^n[(y_i - \beta x_i)^2/x_i]$ results in the least squares (LS) estimator $\hat{\beta} = \Sigma y_i / \Sigma x_i = \bar{y}/\bar{x}$ for the slope β. This estimator is unbiased and its variance is $V(\hat{\beta}) = \sigma^2/n\bar{x}$. Estimates of $E(y_i | x_i)$ and predictions of individual y_i are obtained from $\hat{y}_i = \hat{\beta} x_i = (\bar{y}/\bar{x}) x_i$.

An unbiased estimator of σ^2 is given by the residual or error mean square, $\hat{\sigma}^2 = \Sigma_1^n[(y_i - \hat{\beta} x_i)^2/x_i]/(n-1)$, which has $(n-1)$ d.f. Now, the variance of $\hat{\beta}$ can be estimated from $v(\hat{\beta}) = \hat{\sigma}^2/n\bar{x}$, and the S.E.$(\hat{\beta})$ is obtained from the square root of this expression. The null hypothesis that $\beta = 0$ against the alternative hypotheses that $\beta > 0$ or $\beta < 0$ can be tested from the statistic $t = \hat{\beta}/\text{S.E.}(\hat{\beta})$, which follows Student's t-distribution with $(n-1)$ d.f. If the null hypothesis is rejected, the regression of y_i on x_i cannot be used for the estimation of $E(y_i | x_i)$ or the prediction of y_i.

For the sake of illustration, consider the observations on height (x_i) and weight (y_i) of the 15 candidates of Table 9.1. If the regression of y_i on x_i is of the above type and the 15 candidates are considered to constitute a random sample from a large population, the LS estimate of the slope is $\hat{\beta} = \bar{y}/\bar{x} = 155.53/67.73 = 2.3$. The estimates $\hat{y}_i = \hat{\beta} x_i$ at $x_i = (63, 64, 65, 66, 68, 70, 72)$ are (144.9, 147.2, 149.5, 151.8, 156.4, 161, 165.6). The observations on the 15 units and the estimate of the regression line are presented in Figure 9.1.

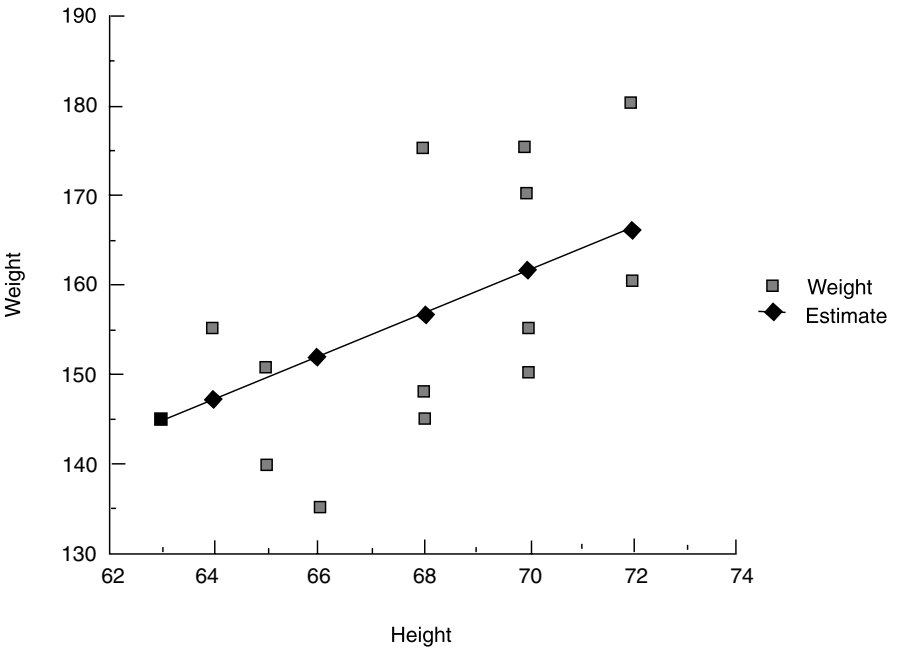

Figure 9.1. Regression through the origin of weight on height.

The estimate for σ^2 is 1.73. Thus, the variance of $\hat{\beta}$ is 1.73/67.73 = 0.0255, and hence its S.E. is 0.1598. The value of $t = 2.3/0.1598 = 14.4$ with 14 d.f. is highly significant at the significance level of 0.001 or much smaller.

In general, the variance of ε_i can be of the form $\sigma^2 x_i^h = \sigma^2/W_i$, where $W_i = 1/x_i^h$. In several applications, the coefficient h has been found to be between zero and two. The approximate value of h can be found by plotting y_i against x_i and examining the variance of y_i at selected values of x_i. Now, for the generalized least squares (**GLS**) or weighted least squares (**WLS**) procedure, the slope is estimated by minimizing $\Sigma W_i (y_i - \beta x_i)^2$. As a result, $\hat{\beta} = \Sigma W_i x_i y_i / \Sigma W_i x_i^2$, which is unbiased with variance $V(\hat{\beta}) = \sigma^2/\Sigma W_i x_i^2$.

The LS estimator $\hat{\beta} x_i$ is unbiased for $E(y_i | x_i)$ and its variance is given by $x_i^2 V(\hat{\beta})$. This estimator is linear in the observations y_i, unbiased, and its variance is smaller than that of any other linear unbiased estimator; hence, it is known as the best linear unbiased estimator (BLUE).

An unbiased estimator of σ^2 is given by the residual or error mean square (EMS), $\hat{\sigma}^2 = \Sigma W_i(y_i - \hat{\beta}x_i)^2/(n-1)$. The error SS (ESS) can be expressed as $\Sigma W_i(y_i - \hat{\beta}x_i)^2 = \Sigma W_i y_i^2 + \hat{\beta}^2 \Sigma W_i x_i^2 - 2\hat{\beta}\Sigma W_i x_i y_i = \Sigma W_i y_i^2 - \hat{\beta}^2 \Sigma W_i x_i^2$ = total SS (TSS) − regression SS (RSS). For this regression through the origin, the d.f. for the total SS, regression SS, and error SS are n, 1, and $(n-1)$, respectively.

It has been noted that for some practical situations, $V(y_i \mid x_i)$ usually takes the form of σ^2, $\sigma^2 x_i$ or $\sigma^2 x_i^2$. For the first case, $V(y_i \mid x_i)$ is a constant; for the second, it is proportional to x_i; and for the third, it is proportional to x_i^2. For these three cases, the WLS estimator for β is given by $\Sigma x_i y_i / \Sigma x_i^2$, $\Sigma y_i / x_i = \bar{y}/\bar{x}$ and $\Sigma(y_i/x_i)/n$, respectively. From the above general expression, the ESS for these cases can be expressed as $\Sigma y_i^2 - \hat{\beta}^2 \Sigma x_i^2$, $\Sigma y_i^2/x_i - \hat{\beta}^2 \Sigma x_i$, and $\Sigma(y_i/x_i)^2 - n\hat{\beta}^2$, respectively. The first term in each of these expressions is the total SS, and the second term is the regression SS.

For the data on heights and weights in Table 9.1, $\hat{\beta}$ is given by 10557.93/4595.73 = 2.2973, 155.53/67.73 = 2.2963, and 34.428/15 = 2.2952 for the above three cases, respectively. Thus, there is little difference in the slopes for these cases. For these cases, the regression and residual SS are (1652.55, 363826), (5357.18, 24.16), and (79.02, 0.35), respectively. The ratio of the regression MS to the residual MS, which follows the F-distribution with 1 and $(n-1) = 15$ d.f. is highly significant for all three cases. From this result, the slope is inferred to be significantly greater than zero for all three cases.

The above regression approach provides motivation for estimating $\bar{Y}$ from $\bar{X}\Sigma W_i x_i y_i / \Sigma W_i x_i^2$. In particular, this estimator becomes the same as $\hat{\bar{Y}}_R = (\bar{y}/\bar{x})\bar{X}$ in (9.3) when $V(y_i \mid x_i) = \sigma^2 x_i$. Note that, for the regression approach, the means and variances of y_i conditional on x_i are considered. The bias, variance, and MSE of the ratio estimators derived in the previous sections depend on the means and variances of both x and y, as well as their correlation.

Superpopulation model for evaluations and estimation

Assuming that the finite population of N units is a sample from an infinite **superpopulation**, following the model $y_i = \alpha + \beta x_i + \varepsilon_i$ with $V(y_i \mid x_i) = \sigma^2 x_i^h$, P.S.R.S. Rao (1968) compared the expected values of the MSE of $\hat{\bar{Y}}_R$ and the variances of $\hat{\bar{Y}}_{SS}$ and $\hat{\bar{Y}}_{RHC}$ described in Section 7.13; the cluster size is represented by x_i. This investigation showed that for $\alpha = 0$, that is, for the model in (9.10), $\hat{\bar{Y}}_R$ and $\hat{\bar{Y}}_{SS}$ are preferable when $V(y_i \mid x_i)$ is proportional to x_i and $\hat{\bar{Y}}_{RHC}$ when this variance is proportional to x_i^2.

When the finite population is assumed to be a sample from a superpopulation following the model in (9.10) with $V(y_i \mid x_i) = \sigma^2 x_i$, Royall (1970) considers $\hat{Y} = \Sigma_1^n c_i y_i$ to estimate Y. He shows that when the variance of $\hat{Y}$, $V(\hat{Y}) = E(\hat{Y} - Y)^2$, is minimized with the **model-unbiasedness** condition, $E(\hat{Y}) = E(Y)$, it becomes the same as the ratio estimator $\hat{Y}_R = \hat{R}X = (\bar{y}/\bar{x})X$. The variance of $\hat{Y}_R$ now becomes $V(\hat{Y}_R) = \sigma^2(N\bar{X} - n\bar{x})\,N\bar{X}/n\bar{x}$. The estimator for σ^2 is given by $\hat{\sigma}^2 = \Sigma_1^n [(y_i - \hat{R}\,x_i)^2/x_i]/(n-1)$, which is the same as the least squares estimator presented earlier in this section. Cochran (1977, p. 159) also presents the derivations for $\hat{Y}_R$, $V(\hat{Y}_R)$, and $\hat{\sigma}^2$.

For the above approach, the estimator for $\bar{Y}$ is the same as $(\bar{y}/\bar{x})\bar{X}$ and its variance is estimated from $\hat{\sigma}^2(N\bar{X} - n\bar{x})\,\bar{X}/Nn\bar{x}$. Through the model in (9.10), P.S.R.S. Rao (1981b) compares this variance estimator with $v(\hat{\bar{Y}}_R)$ in (9.6).

9.8 Ratio estimation vs. stratification

If the available supplementary variable is of the qualitative or categorical type, low–medium–high, for example, it can be used for stratifying the population, but not for the ratio estimation. However, if it is of the continuous type, ratio estimation is preferred provided the assumptions described in Section 9.7 are satisfied at least approximately.

To examine the two procedures, consider wheat production in 1997 (y) for the first $N = 20$ countries in Table T8 in the Appendix; the five countries with the largest production are not included in this group. Summary figures for these data are presented in Table 9.2.

If a sample of $n = 8$ countries is considered for the estimation of the average production, $V(\bar{y}) = [(20 - 8)/160](55.09) = 4.13$ and hence S.E.$(\bar{y}) = 2.03$. If one considers the supplementary data of 1995 (x), $R = 9.95/9.18 = 1.0839$ and $S_d^2 = 2.84$. Now, from (9.4), $V(\hat{\bar{Y}}_R) = [(20 - 8)/160](2.84) = 0.213$ and S.E.$(\hat{\bar{Y}}_R) = 0.46$. Thus, the S.E. of the ratio estimator is only about 20% of that of the sample mean.

To examine the effect of stratification, consider $G = 2$ strata, with the first $N_1 = 11$ and the next $N_2 = 9$ countries of Table T8 in the Appendix along with the summary figures in Table 9.2. For both proportional and Neyman allocations of the sample of $n = 8$ units for the two strata, $n_1 = 4$ and $n_2 = 4$ approximately. For these sample sizes, the variances of the sample means in the two strata are $V(\hat{\bar{Y}}_1) = [(11 - 4)/44](8.14) = 1.3$ and $V(\hat{\bar{Y}}_2) = [(9 - 4)/36](15.94) = 2.21$. Now, from (5.13), $V(\hat{\bar{Y}}_{st}) = (11/20)^2(1.3) + (9/20)^2(2.21) = 0.84$ and hence

Table 9.2. Wheat production.

	1990	1995	1997
	20 Countries		
Total	212.1	183.5	198.9
Mean	10.61	9.18	9.95
Variance	82.68	50.88	55.09
S.D.	9.09	7.13	7.42
	11 Countries		
Total	53.1	41.2	45
Mean	4.83	3.75	4.09
Variance	21.29	8.08	8.14
S.D.	4.61	2.84	2.85
	9 Countries		
Total	159	142.3	153.9
Mean	17.67	15.81	17.1
Variance	67.74	20.66	15.94
S.D.	8.23	4.55	3.99
	Correlations		
	(1990, 1995)	(1990, 1997)	(1995, 1997)
20 countries	0.89	0.90	0.98
11 countries	0.78	0.90	0.96
9 countries	0.77	0.81	0.88

Note: Productions for these 20 countries are presented in Table T8 in the Appendix.

S.E.($\hat{Y}_{st}$) = 0.92. Although this S.E. for stratification is only about 45% that of the sample mean, it is twice that of the ratio estimator.

In this illustration, the ratio estimator has much smaller S.E. than the sample mean since the correlation of the 1995 and 1997 production is 0.98, which is very high. The S.E. of the stratified estimator will be reduced if the 20 countries are divided into two strata with smaller within variances.

9.9 Ratio estimation with stratification

Since stratification as well as ratio estimation produce gains in precision, one may combine both procedures for further gains. There are two methods of achieving this objective, but first consider the benefits of ratio estimation in a single stratum.

Single stratum

When the mean of the supplementary variable $\bar{X}_g$ is known, the ratio estimators for the total Y_g and mean $\bar{Y}_g$ are

$$\hat{Y}_{Rg} = \frac{y_g}{x_g} X_g = \frac{\bar{y}_g}{\bar{x}_g} X_g \tag{9.11}$$

and

$$\hat{\bar{Y}}_{Rg} = \frac{\bar{y}_g}{\bar{x}_g} \bar{X}_g. \tag{9.12}$$

If n_g is large, the bias of (9.12) becomes negligible and following (9.4) its variance is approximately given by

$$\begin{aligned} V(\hat{\bar{Y}}_{Rg}) &= (1 - f_g) S_{dg}^2 / n_g \\ &= \frac{(1 - f_g)}{n_g} [S_{yg}^2 + R_g^2 S_{xg}^2 - 2\rho_g S_{yg} S_{xg}], \end{aligned} \tag{9.13}$$

where $R_g = \bar{Y}_g / \bar{X}_g$. Note that $S_{dg}^2 = \Sigma_1^{N_g} (y_{gi} - R_g x_{gi})^2 / (N_g - 1)$ and ρ_g is the correlation between x and y in the gth stratum. An estimator for this variance is

$$v(\hat{\bar{Y}}_{Rg}) = \frac{(1 - f_g)}{n_g} s_{dg}^2 = \frac{(1 - f_g)}{n_g} [s_{yg}^2 + \hat{R}_g^2 s_{xg}^2 - 2\hat{R}_g s_{xyg}], \tag{9.14}$$

where $\hat{R}_g = \bar{y}_g / \bar{x}_g$ and $s_{dg}^2 = \Sigma_1^{n_g} (y_{gi} - \hat{R}_g x_{gi})^2 / (n_g - 1)$.

The separate ratio estimator

If the regression of y on x approximately passes through the origin in each of the strata, one can consider the estimator in (9.11) separately for each of the strata totals. Now, adding these estimators, the ratio estimator for the population total $Y = \Sigma Y_g$ is given by

$$\hat{Y}_{RS} = \sum_1^G \hat{Y}_{Rg} = \sum_1^G \frac{y_g}{x_g} X_g = \sum_1^G N_g \frac{\bar{y}_g}{\bar{x}_g} \bar{X}_g \tag{9.15}$$

The subscripts R and S refer to ratio estimation and stratification.

Dividing (9.15) by N, the estimator for $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{\text{RS}} = \sum_{1}^{G} W_g \hat{\bar{Y}}_{Rg} = \sum_{1}^{G} W_g \frac{\bar{y}_g}{\bar{x}_g} \bar{X}_g \tag{9.16}$$

The variance and estimator of variance of (9.16) are

$$V(\hat{\bar{Y}}_{\text{RS}}) = \sum W_g^2 V(\hat{\bar{Y}}_{Rg}) \tag{9.17}$$

and

$$v(\hat{\bar{Y}}_{\text{RS}}) = \sum W_g^2 v(\hat{\bar{Y}}_{Rg}). \tag{9.18}$$

With proportional allocation of the sample, the variance in (9.17) becomes

$$V_{\text{prop}}(\hat{\bar{Y}}_{\text{RS}}) = \frac{(1-f)}{n} \sum_{1}^{G} W_g S_{dg}^2. \tag{9.19}$$

The variance in (9.17) or (9.19) clearly becomes small if y_{gi} is highly positively correlated with x_{gi} within each stratum. If these correlations are high and R_g vary, (9.19) can be expected to be much smaller than the variance of $\hat{\bar{Y}}_R$ in (9.4).

For Neyman allocation, the variance in (9.17) is minimized for a given $n = \Sigma n_g$. As a result, n_g is chosen proportional to $N_g S_{dg}$. If the cost function in (5.33) is suitable, optimum allocation results in choosing n_g proportional to $N_g S_{dg}/\sqrt{e_g}$. These procedures require additional information.

Example 9.6. Wheat production: Consider estimation of the average wheat production for the 20 countries for 1997 (y) with the information from 1995 (x) through stratification and the ratio method. From the summary figures in Table 9.2, $S_{d1}^2 = 0.85$ and $S_{d2}^2 = 5.68$. For samples of size four from each of the strata, $V(\hat{\bar{Y}}_{R1}) = (7/44)(0.85) = 0.1352$ and $V(\hat{\bar{Y}}_{R2}) = (5/36)(5.68) = 0.7889$. Now, from (9.17), $V(\hat{\bar{Y}}_{\text{RS}}) = (11/20)^2(0.1352) + (5/36)^2(0.7889) = 0.20$ and S.E.$(\hat{\bar{Y}}_{\text{RS}}) = 0.45$. This variance is only one fourth of the variance of 0.84 found in Section 9.8 for the stratified mean and only slightly smaller than the variance of 0.213 for the ratio estimator without stratification.

The combined ratio estimator

If the ratios R_g for the strata do not differ much, as suggested by Hansen et al. (1946), it is possible first to obtain the common ratio as $\hat{R}_C = \hat{\bar{Y}}_{RC}/\hat{\bar{X}}_{RC} = \Sigma W_g \bar{y}_g / \Sigma W_g \bar{x}_g$ and for $\bar{Y}$ consider the **combined ratio estimator**:

$$\hat{\bar{Y}}_{RC} = \hat{R}_C \bar{X}. \tag{9.20}$$

For large samples, the bias of this estimator vanishes and its variance approximately becomes

$$\begin{aligned} V(\hat{\bar{Y}}_{RC}) &= E(\hat{\bar{Y}}_{RC} - R\hat{\bar{X}}_{RC})^2 \\ &= \Sigma W_g^2 \frac{(1-f_g)}{n_g}(S_{yg}^2 + R^2 S_{xg}^2 - 2R\rho_g S_{yg} S_{xg}). \end{aligned} \tag{9.21}$$

This variance can be estimated from

$$v(\hat{\bar{Y}}_{RC}) = \Sigma W_g^2 \frac{(1-f_g)}{n_g}(s_{yg}^2 + \hat{R}_c^2 s_{xg}^2 - 2\hat{R}_c s_{xyg}). \tag{9.22}$$

Derivations of (9.21) and (9.22) are outlined in Appendix A9.

From the figures in Table 9.2, $R = 9.95/9.18 = 1.084$. Now from (9.21), $V(\hat{\bar{Y}}_{RC}) = 0.1936$ and S.E.$(\hat{\bar{Y}}_{RC}) = 0.44$. This S.E. is almost the same as that of the separate estimator. One reason for this result is that the ratios of the means for the first and second strata are $R_1 = 4.09/3.75 = 1.09$ and $R_2 = 17.1/15.81 = 1.08$, which are almost the same.

Statistical test for the equality of the slopes

From samples of sizes n_g selected from large-size strata, the null hypothesis that their slopes are all the same against the alternative that they are unequal can be tested as follows.

1. Fit the G regressions through the origin, obtain the pooled residual sum of squares, SS_{sep}, with $(n_1 - 1) + (n_2 - 1) + \cdots + (n_G - 1) = (n - G)$ d.f. and obtain the estimate of σ^2 from $\hat{\sigma}^2 = SS_{sep}/(n - G)$.
2. Fit the regression through the origin from combining all the n observations and find the residual sum of squares,

SS_{comb}, with $(n - 1)$ d.f. The difference, $SS_{diff} = (SS_{comb} - SS_{sep})$ has $(n - 1) - (n - G) = (G - 1)$ d.f.

3. Let $MS_{diff} = SS_{diff}/(G - 1)$. The ratio $F = MS_{diff}/\hat{\sigma}^2$ follows the F-distribution with $(G - 1)$ and $(n - G)$ d.f.

Percentiles of the F-distribution are tabulated and are also available through computer software packages. For a specified significance level, the null hypothesis is rejected if the computed value of F exceeds the actual percentage point. Rejection of the null hypothesis suggests the separate estimator; the combined estimator otherwise.

9.10 Bias reduction

From Appendix A9, ignoring terms of order $(1/n^2)$ and smaller, an estimate of the bias of $\hat{R}$ is given by

$$\hat{B}(\hat{R}) = (1 - f)\hat{R}(c_{xx} - c_{xy})/n, \tag{9.23}$$

where $c_{xx} = s_x^2/\bar{x}^2$ and $c_{xy} = s_{xy}/\bar{x}\,\bar{y}$. Subtracting this expression from $\hat{R}$, an estimator for R is obtained from

$$\hat{R}_T = \hat{R}[1 - (1 - f)(c_{xx} - c_{xy})/n]. \tag{9.24}$$

The bias of this estimator is of order $1/n^2$. The corresponding estimator for $\bar{Y}$ is given by $\hat{\bar{Y}}_T = \bar{X}\hat{R}_T$, which was suggested by Tin (1965).

The approach in Appendix A9 for expressing a nonlinear estimator such as $\hat{R}$ in a series and ignoring higher-order terms is known as the **linearization** procedure. As seen above, the bias of an estimator can be reduced by this approach. Further, as shown in Appendix A9, an approximation to the MSE of $\hat{R}$ and its estimator can also be obtained by this procedure. Chapter 12 compares this approach with the **jackknife** and **bootstrap procedures**.

Hartley and Ross (1954) derive an unbiased estimator for $\bar{Y}$ by removing the bias from $\bar{X}\bar{r}$. The investigation by P.S.R.S. Rao (1969) through the model in (9.10) showed that $\hat{\bar{Y}}_R$ and the Hartley–Ross estimators have relatively smaller MSEs when $V(y_i|x_i)$ is proportional to x_i and x_i^2, respectively.

9.11 Two-phase or double sampling ratio estimators

If $\bar{X}$ is not known, it can be estimated from the mean $\bar{x}_1$ of a large sample of size n_1 selected from the N population units. Now, $(\bar{x}, \bar{y})$ are obtained from a subsample of size n selected from the n_1 units. As an illustration, the first sample may provide an estimate for the average family size for the households in a region. The subsample at the second phase provides the averages for the family size and a major characteristic of interest, for example, savings. This type of two-phase sampling is feasible if the cost of sampling the first sample is cheaper than the second.

$\bar{Y}$ can now be estimated from

$$\hat{\bar{Y}}_{\mathrm{Rd}} = (\bar{y}/\bar{x})\bar{x}_1. \tag{9.25}$$

Following Appendix A3, for large n, the bias of this estimator becomes small and its variance becomes

$$V(\hat{\bar{Y}}_{\mathrm{Rd}}) = (N - n_1)S_y^2/Nn_1 + (n_1 - n)S_d^2/n_1 n. \tag{9.26}$$

Note that S_d^2 has the same expression as in Section 9.2.

The bias of $\hat{\bar{Y}}_{\mathrm{Rd}}$ can be reduced or eliminated through modifying and extending the procedures in Section 9.10. P.S.R.S. Rao (1981a) evaluates the efficiencies of nine such procedures through the model in (9.10). A summary of the double-sampling ratio estimators is presented in P.S.R.S. Rao (1998b).

In some situations, the means $\bar{x}_1$ and $(\bar{x}, \bar{y})$ are obtained from samples of sizes n_1 and n selected independently from the N population units. In such a case, one can consider the combined estimator $\bar{x}_a = \mathbf{a}\bar{x}_1 + (1 - \mathbf{a})\bar{x}$ for $\bar{X}$. This estimator is unbiased and the coefficient $\mathbf{a}$ is obtained by minimizing its variance. As a result, the optimum value of $\mathbf{a}$ is given by $(N - n)n_1/[(N - n)n_1 + (N - n_1)n]$. Alternatively, one may remove the duplications from the two samples and consider the mean $\bar{x}_v$ of the v distinct units to estimate $\bar{X}$. The variances of both these alternative estimators are smaller than the variance of $\bar{x}$. Now, one can consider $(\bar{y}/\bar{x})\bar{x}_a$ or $(\bar{y}/\bar{x})\bar{x}_v$ to estimate $\bar{Y}$. P.S.R.S. Rao (1975b) evaluates the merits of these two estimators relative to $\hat{\bar{Y}}_{\mathrm{Rd}}$. Similar procedures for the Hartley–Ross type of estimation are examined in P.S.R.S. Rao (1975a).

9.12 Ratio estimator with unequal probability selection

As we have seen in Section 7.13 for the PPSS procedure, $\bar{X}(\bar{y}/\bar{x})$ is unbiased for $\bar{Y}$ if the sample (x_i, y_i), $i = 1, 2, \ldots, n$, is selected with probability proportional to Σx_i.

If the units are selected with probabilities ϕ_i and without replacement, as described in Section 7.12 for the Horvitz–Thompson estimator, $\Sigma_1^n(y_i/\phi_i)$ and $\Sigma_1^n(x_i/\phi_i)$ are unbiased for $Y = \Sigma_1^N y_i$ and $X = \Sigma_1^N x_i$, but their ratio is not unbiased for R. For $\bar{Y}$, one can still consider the ratio-type estimator $[\Sigma_1^n(y_i/\phi_i)/\Sigma_1^n(x_i/\phi_i)]\bar{X}$. P.S.R.S. Rao (1991) empirically compared this procedure with the PPSS and pps procedures in Section 7.13 and the ratio estimator $\bar{X}(\bar{y}/\bar{x})$ with equal probability selection, with the sample size $n = 2$. The investigation showed that the MSE for this procedure can be smaller than that of the ratio estimator with equal probability selection and the variances of the estimators of the PPSS and pps procedures provided ϕ_i is proportional to y_i.

9.13 Multivariate ratio estimator

If the population means $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p$ of $p(>1)$ supplementary characteristics are available, the ratio estimators $\hat{\bar{Y}}_{R1} = \bar{X}_1(\bar{y}/\bar{x}_1)$, $\hat{\bar{Y}}_{R2} = \bar{X}_2(\bar{y}/\bar{x}_2), \ldots, \hat{\bar{Y}}_{R2} = \bar{X}_2(\bar{y}/\bar{x}_2)$ can be considered for $\bar{Y}$. The means $(\bar{y}; \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$ are obtained from a sample of size n. The estimator corresponding to the supplementary variable that has the highest correlation with the principal characteristic (y) can be expected to have the smallest MSE. For $\bar{Y}$, Olkin (1958) considers the **multivariate** estimator $W_1\hat{\bar{Y}}_{R1} + W_2\hat{\bar{Y}}_{R2} + \cdots + W_p\hat{\bar{Y}}_{Rp}$, $\Sigma_1^p W_i = 1$. The weights $(W_1, W_2, \ldots, W_p)$ are obtained by minimizing the MSE of this estimator.

Exercises

9.1. *Project*. As in Exercise 2.10, consider the 20 possible samples of size three from the six candidates in Table 2.1. (a) Find the expectation of $\hat{R} = \bar{y}/\bar{x}$ and find its bias for estimating $R = \bar{Y}/\bar{X} = Y/X$. (b) Compare the exact variance and MSE of $\hat{R}$ with the approximate variance in (9.1). (c) Find the bias of (9.2) for estimating the approximate variance in (9.1) and the exact MSE of $\hat{R}$.

9.2. *Project*. With the results in Exercise 9.1, (a) find the bias of $\hat{\bar{Y}}_R$ for estimating $\bar{Y}$, (b) compare its approximate variance in (9.4) with its exact MSE, and (c) find the biases of $v(\hat{\bar{Y}}_R)$ and $v'(\hat{\bar{Y}}_R)$ for estimation of $V(\hat{\bar{Y}}_R)$.

9.3. *Project*. From the 20 samples in Exercise 9.1, (a) find the expectation of $\hat{R} = \bar{y}/(\bar{x} + \bar{y})$ and its bias in estimating $R = \bar{Y}/(\bar{X} + \bar{Y}) = Y/(X + Y)$. (b) Find the MSE of $\hat{R}$ and compare it with its approximate variance.

9.4. (a) From Table 9.1, find the population average $\bar{R}$ and variance $V(\bar{r})$ for the ratios of the systolic pressure (y) to weight (x). (b) From the sample units (4, 6, 9, 12, 15) selected in Example 9.1, estimate $\bar{R}$, find the sample variance of the estimate and the 95% confidence limits for $\bar{R}$.

9.5. From the five sample units (4, 6, 9, 12, 15) selected in Example 9.1, estimate $R = \bar{Y}/(\bar{X} + \bar{Y})$, find the S.E. of the estimate and the 95% confidence limits for this ratio.

9.6. With the sample units (3, 6, 8, 12, 13) selected from the 15 units of Table 9.1, (a) estimate the ratio $R = \bar{Y}/\bar{X}$ of the systolic pressure to weight and find the S.E. of the estimate, and (b) find the ratio estimator for $\bar{Y}$ and estimate its S.E.

9.7. With the sample units (3, 6, 8, 12, 13) selected from the 15 units of Table 9.1, (a) estimate the difference in the ratio estimates for the averages of the systolic and diastolic pressures with weight as the supplementary variable, (b) estimate the S.E., and (c) find the 95% confidence limits for the difference.

9.8. For the five sample units (4, 6, 9, 12, 15) selected in Example 9.1, consider the diastolic pressures (80, 92, 95, 80, 93) after the fitness program. Find the ratio estimate for the difference in the averages for the diastolic pressures before and after the fitness program and find the S.E. of the estimate.

9.9. From the data in Table 7.1, for a sample n clusters, (a) find the variances of the ratio estimators for the average employment per cluster and per county with the number of establishments as the supplementary variable. (b) Find the precisions of these estimators relative to the sample means.

9.10. To estimate the average employment of the unequal size clusters in Table 7.3, as seen in Example 7.5, the ratio

estimator $(\bar{y}/\bar{m})\overline{M}$ has considerably higher precision than $\bar{y}$. An alternative ratio estimator with the number of establishments as the supplementary variable is $(\bar{y}/\bar{x})\overline{X}$. For another procedure, let $u_i = x_i/M_i$ and $v_i = y_i/M_i$, denote their sample means by $\bar{u}$ and $\bar{v}$, and consider the ratio estimator $(\bar{v}/\bar{u})\overline{X}$. For a sample of n clusters, find the precision of the three ratio estimators relative to the sample mean.

9.11. For the population of the ten states with the highest enrollments presented in Table T4 in the Appendix, denote the public and private enrollments for 1990 by (x_1, y_1) and for 1995 by (x_2, y_2). (a) Find the means, standard deviations, covariances, and correlations of these four characteristics. (b) For both the years, with the total enrollment as the supplementary variable, find the variance of the ratio estimator for the mean of the public enrollments for a sample of four states. (c) Compare the variances in (b) with those of the sample means.

9.12. (a) Find the variance of the difference of the ratio estimators in Exercise 9.11(b) for 1990 and 1995 and (b) compare it with the variance of the difference of the sample means.

9.13. For the population of the ten states with the smallest enrollments presented in Table T5 in the Appendix, denote the public and private enrollments for 1990 by (x_1, y_1) and for 1995 by (x_2, y_2). (a) Find the means, standard deviations, covariances, and correlations of these four characteristics. (b) For both years, for a sample of four units, find the variance of the ratio of the totals of the public and private enrollments. (c) Find the variance of the difference of the ratios in (b).

9.14. Consider the first six and the remaining nine units of Table 9.1 as two strata. With weight as the supplementary variable, compare the variances of the separate and combined ratio estimators for the mean of the systolic pressures for a total sample of size five. Consider both proportional and Neyman allocations.

9.15. *Project.* For each of the 120 samples of size three that can be selected from the ten states in Table T7 in the Appendix, find the difference of the ratios of the means of the two types of expenditures to the total expenditure as in Section 9.6. (a) From the average and variance of

the 120 estimates, find the bias and MSE of the above estimator. (b) Compute the average of the approximate estimates in (9.7) for each of the 120 samples, and compare their average with the exact MSE in (a).

9.16. Consider the observations in Table 9.1 to be obtained from a sample of 15 units of a large population, and plot the observations on systolic pressure (y_i) against weight (x_i). As described in Section 9.7, consider regression through the origin of y_i on x_i when $V(y_i|x_i)$ equals σ^2, $\sigma^2 x_i$, and $\sigma^2 x_i^2$. For each of these three cases, estimate (a) the slope, (b) σ^2, and (c) the S.E. of the slope. (d) Test the hypothesis that $\beta = 0$ against the alternative hypothesis that $\beta > 0$. (e) From the above results, which of the three procedures do you recommend for estimating the mean of the systolic pressure; give reasons.

9.17. Divide the 25 countries in Table T9 in the Appendix into two strata consisting of the eight largest petroleum-producing countries and the remaining 17 countries with smaller or no production. To estimate the average per capita energy consumption for the 25 countries with petroleum production as the supplementary variable, find the S.E. of the separate ratio estimator for samples of sizes four and eight from the two strata.

Appendix A9

Bias of the ratio of two sample means

The bias of $\hat{R} = \bar{y}/\bar{x}$ is given by

$$B(\hat{R}) = E(\hat{R}) - R = E[(\bar{y} - R\bar{x})/\bar{x}] = E[(\bar{y} - R\bar{x})/\bar{X}(1 + \delta)],$$

where $\delta = (\bar{x} - \bar{X})/\bar{X}$. Expressing $(1 + \delta)^{-1}$ as $1 - \delta + \delta^2 - \cdots$ and retaining only the first two terms,

$$\begin{aligned} B(\hat{R}) &= E[(\bar{y} - R\bar{x})(1 - \delta)/\bar{X}] \\ &= E(\bar{y} - R\bar{x})/\bar{X} - E[(\bar{y} - R\bar{x})(\bar{x} - \bar{X})]/\bar{X}^2. \end{aligned}$$

The first term vanishes, and the bias now becomes

$$B(\hat{R}) = -[\mathrm{Cov}(\bar{x}, \bar{y}) - RV(\bar{x})]/\bar{X}^2 = (1-f)R(C_{xx} - C_{xy})/n,$$

where $C_{xx} = S_x^2/\bar{X}^2$ and $C_{xy} = S_{xy}\bar{X}\bar{Y}$. This bias becomes negligible for large n.

The above procedure of expressing a nonlinear estimator such as the ratio in a series and ignoring higher-order terms is known as the **linearization method**.

MSE of the ratio of two sample means

From the definition,

$$\mathrm{MSE}(\hat{R}) = E(\hat{R} - R)^2 = E[(\bar{y} - R\bar{x})/\bar{x}]^2.$$

Following the linearization approach, for large n, the bias of $\hat{R}$ becomes negligible and this expression approximately becomes the variance of $\hat{R}$, given by

$$V(\hat{R}) = E(\bar{y} - R\bar{x})^2/\bar{X}^2.$$

To obtain an explicit expression for this equation, let $d_i = (y_i - Rx_i)$. The population mean of d_i is $\bar{D} = (\Sigma_1^N d_i)/N = \bar{Y} - R\bar{X} = 0$. Its population variance is

$$S_d^2 = \sum_1^N (y_i - Rx_i)^2/(N-1) = \sum_1^N [(y_i - \bar{Y}) - R(x_i - \bar{X})]^2/(N-1)$$
$$= S_y^2 + R^2 S_x^2 - 2RS_{xy} = S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y.$$

The sample mean of the d_i is $\bar{d} = (\Sigma_1^n d_i)/n = (\bar{y} - R\bar{x})$. The expectation of $\bar{d}$ is zero and its variance is $E(\bar{d} - \bar{D})^2 = E(\bar{d}^2) = (1-f)S_d^2/n$. Thus, from the above two expressions,

$$V(\hat{R}) = V(\bar{d})/\bar{X}^2 = (1-f)S_d^2/n\bar{X}^2.$$

This expression can also be obtained by writing $E(\bar{y} - R\bar{x})^2$ as $E[(\bar{y} - \bar{Y}) - R(\bar{x} - \bar{X})]^2 = V(\bar{y}) + R^2V(\bar{x}) - 2R\ \mathrm{Cov}(\bar{x}, \bar{y})$.

From the sample,

$$s_d^2 = \sum_1^n (y_i - \hat{R}x_i)^2/(n-1) = \sum_1^n [(y_i - \bar{y}) - \hat{R}(x_i - \bar{x})]^2/(n-1)$$
$$= s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}$$

is an estimator for s_d^2. Hence, $V(\hat{R})$ may be approximately estimated from $v(\hat{R}) = (1-f)s_d^2/n\bar{x}^2$.

Variance of the combined ratio estimator

The mean of $d_{gi} = y_{gi} - Rx_{gi}$ for the n_g units of the sample is $\bar{d}_g = \bar{y}_g - R\bar{x}_g$, and its mean for the N_g units of the gth stratum is $\bar{D}_g = \bar{Y}_g - R\bar{X}_g$. The population mean of d_{gi} is $\bar{D} = 0$. The variance of d_{gi} for the gth stratum is $S'^2_{dg} = \Sigma(d_{gi} - \bar{D}_g)^2/(N_g - 1) = S^2_{yg} + R^2S^2_{xg} - 2RS_{xyg}$. Since $\bar{d}_{\mathrm{st}} = \Sigma W_g\bar{d}_g = \hat{\bar{Y}}_{\mathrm{RC}} - R\hat{\bar{X}}_{\mathrm{RC}}$, (9.21) can be expressed as

$$V(\hat{\bar{Y}}_{\mathrm{RC}}) = \sum W_g^2 V(\bar{d}_g) = \sum W_g^2 \frac{(1-f_g)}{n_g} S'^2_{dg}.$$

With the sample variance $s'^2_{dg} = \Sigma(d_{gi} - \bar{d}_g)^2/(n_g - 1) = s^2_{yg} + \hat{R}_c^2 s^2_{xg} - 2\hat{R}_c s_{xyg}$, an estimator for the variance in (9.21) is given by

$$v(\hat{\bar{Y}}_{\mathrm{RC}}) = \sum W_g^2 \frac{(1-f_g)}{n_g} s'^2_{dg}.$$