SCB STATISTISKA CENTRALBYRÅN

Memo, February 1985

PROMEMORIOR FRÅN P/STM

NR 16

VARIANCE ESTIMATORS OF THE GINI COEFFICIENT

- SIMPLE RANDOM SAMPLING

AV ARNE SANDSTRÖM, JAN WRETMAN OCH BERTIL WALDÉN

INLEDNING

TILL

Promemorior från P/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1978-1986. – Nr 1-24.

Efterföljare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

Promemorior från P/STM 1985:16. Variance estimators of the Gini coefficient – simple random sampling / Arne Sandström m.fl.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-PM-PSTM-1985-16

SCB STATISTISKA CENTRALBYRÅN

Memo, February 1985

PROMEMORIOR FRÅN P/STM

NR 16

VARIANCE ESTIMATORS OF THE GINI COEFFICIENT

- SIMPLE RANDOM SAMPLING

AV ARNE SANDSTRÖM, JAN WRETMAN OCH BERTIL WALDÉN

Memo, February 1985

VARIANCE ESTIMATORS OF THE GINI COEFFICIENT - SIMPLE RANDOM SAMPLING

Arne Sandström, Jan Wretman and Bertil Waldén[1)]

ABSTRACT: Computations of income inequality measures are usually based on data from sample surveys. The most well-known measure is the Gini coefficient, which is a ratio statistic. We study both the exact sampling distribution obtained by simple random sampling without replacement from two small populations and approximated sampling distributions based on simulations. Four variance estimators are compared.

KEY-WORDS: Gini coefficient, Variance estimators, Simulations.

[1)]Sandström and Wretman are Directors, Statistical Research Unit, Statistics Sweden, S-115 81 Stockholm, Sweden and Waldén is B.A., University of Linköping, S-581 83 Linköping, Sweden. The authors are greatful to Fredrik Nygård, Department of Statistics, Swedish University of Turku, Finland, for valuable comments on an earlier draft.

1. Introduction

During the last decades the interest in measuring income inequality has substantially increased. One reason for this is the fact that many economic-political steps are taken to promote equality between individuals and/or households. The most well-known measure of income inequality is the Gini coefficient.

Analyses of income distributions including computations of inequality measures are usually based on sample surveys. However, a discussion of the sampling properties of the Gini coefficient (and other measures of income inequality) is usually ignored, a fact which is probably due to its 'intractability'.

In this paper we will study both the exact sampling distributions obtained by a simple random sampling (srs) design without replacement from two small parent populations of size N = 11 and approximated sampling distributions based on simulations. In the latter case we use parent populations of size N = 10,000 obtained from eleven continuous distributions. The design is srs with replacement.

The main objective of this study is to compare four variance estimators of the estimated Gini coefficient but also to indicate the behavior of the estimator of the Gini coefficient since it is a ratio statistic. The numerator of the estimated Gini coefficient can be viewed as an L-statistic with scores depending on the sample. The first variance estimator ignores this fact and may be treated as a rough estimator. We call it a ratio estimator ($\hat{V}_R$). In the second estimator we take account of the fact that the scores depend on the sample and call it a Taylor estimator ($\hat{V}_T$). The third variance estimator is based on an asymptotic variance ($\hat{V}_A$) and the fourth on a jackknife procedure ($\hat{V}_J$).

In Section 2 the Gini coefficient is defined and the four variance estimators presented. A main conclusion is drawn from the discussion of the exact sampling distributions in Section 3: Let the values $y_1,\ldots,y_N$, associated with the units of a finite population, take on both negative and positive values. If we translate the finite population distribution function F_N so that $\bar{y}_N = N^{-1} \sum_{k=1}^{N} y_k$ tends towards zero than the finite population Gini coefficient will tend towards infinity and the bias in both the estimator and the variance estimators will be very large. In such cases, when $\bar{y}_N$ is close to zero, we recommend that the Gini coefficient is not used.

In the simulation study, discussed in Section 4, we took 500 sample replicates of sizes $n = 5, 10, 20$, and 100. The study showed that $\hat{V}_A$ and $\hat{V}_J$ are the best variance estimators (even for small sample sizes). $\hat{V}_R$ was poor and $\hat{V}_T$ performed quite well for $n \geq 20$. When $n \to \infty$, and $N \to \infty$, $f_n = n/N \to f$, $0 < f < 1$, then $\hat{V}_T$ and $\hat{V}_A$ are identical in the srs design.

2. The estimation problem

We first give a general definition of Gini's mean difference and the Gini coefficient associated with a distribution function.

Let F be a distribution function (df, for short), by which we mean a real-valued function defined on $(-\infty,\infty)$ that is nondecreasing, right continuous and satisfies $F(-\infty) = 0$ and $F(+\infty) = 1$. In terms of the Lebesgue-Stieltjes intergral, Gini's mean difference, G, associated with F is defined as

$$G = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x-y| \, dF(x)dF(y) \; . \tag{2.1}$$

The Gini coefficient, R, associated with F is defined as

$$R = \frac{G}{2\mu} \tag{2.2}$$

where

$$\mu = \int_{-\infty}^{\infty} y \, dF(y) \neq 0 \tag{2.3}$$

Formally, G and R may be considered as functionals: To any given df F, (2.1) and (2.2) assign values G and R. (For the definition of R to be meaningful, it is required that $\mu \neq 0$.)

We note that G is a location-free quantity, in the sense that it is unaffected by a shift of F. (If F_1 and F_2 are two df's such that $F_1(y) = F_2(y+c)$, then F_1 and F_2 will have the same G.) For a sequence of df's such that G is constant but μ tends to zero, R will tend to infinity. We also note that $G \geq 0$, and hence that R will have the same sign as μ, provided $G > 0$.

Gini coefficients are frequently used in the study of income distributions, where R is a measure of the income inequality among units in a population.

Let $y_1, \ldots y_N$ be values associated with the units of a finite population of size N, and let $y_{N:1} \leq y_{N:2} \leq \ldots \leq y_{N:N}$ be the same values arranged in nondecreasing order of magnitude. Let F_N be the finite population df, that is, $F_N(y)$ is the proportion of units such that $y_k \leq y$. Inserting F_N for F in (2.1) - (2.3) we obtain

$$G_N = \frac{1}{N^2} \sum_{k=1}^{N} \sum_{\ell=1}^{N} |y_k - y_\ell|$$

$$= \frac{2}{N^2} \sum_{i=1}^{N} (2i-N-1) y_{N:i} \tag{2.4}$$

$$\mu_N = \bar{y} = \frac{1}{N} \sum_{k=1}^{N} y_k$$

$$R_N = \frac{G_N}{2\mu_N} = \frac{1}{2N^2\bar{y}} \sum_{k=1}^{N} \sum_{\ell=1}^{N} |y_k - y_\ell|$$

$$= \frac{2}{N^2 \bar{y}} \sum_{i=1}^{N} i y_{N:i} - 1 - \frac{1}{N}$$

$$= \frac{\sum_{k=1}^{N} w_k y_k}{\sum_{k=1}^{N} y_k} - 1 - \frac{1}{N} . \tag{2.5}$$

The quantity R_N defined by (2.5) is a finite population parameter. It is essentially a ratio between two finite population totals, $\Sigma_1^N w_k y_k$ and $\Sigma_1^N y_k$. Estimating R_N from a probability sample of units is not quite straightforward, however, because the w_k-values of the sampled units will remain unknown and have to be estimated in some way. An estimator R_N under a general probability sampling design was suggested by Nygård and Sandström (1985).

In this study we consider simple random sampling (both with and without replacement). Let s be the set of sampled units, i.e., a subset of $\{1,2,\ldots,N\}$, with sample mean $\bar{y}_s = n^{-1}\Sigma_{k\in s} y_k$, and let $y_{s:1} \leq y_{s:2} \leq \ldots < y_{s:n}$ be the values of the sampled units arranged in nondecreasing order of size. We will consider the following estimator of R_N, which is simply the sample analog of R_N,

$$\hat{R}_N = \frac{1}{2\, n^2\, \bar{y}_s} \sum_{k\in s} \sum_{\ell\in s} |y_k - y_\ell|$$

$$= \frac{2}{n^2\, \bar{y}_s} \sum_{i=1}^{n} i\, y_{s:i} - 1 - \frac{1}{n}$$

$$= \frac{\Sigma_{k\in s} w_k^* y_k}{\Sigma_{k\in s}\, y_k} - 1 - \frac{1}{n} \tag{2.6}$$

Our problem is to estimate the sampling variance of $\hat{R}_N$, denoted $V(\hat{R}_N)$, under simple random sampling, using data from one single sample. Since $\hat{R}_N$ is a "complex" statistic, it is not possible to estimate its variance by traditional methods of unbiased variance estimation. We have to rely on some approximate variance estimation technique. Four methods of this kind will be considered here.

i) The first method uses a variance estimation formula obtained by analogy with the well-known formula for estimating the approximate variance of a ratio estimator, based on first-order Taylor approximation (see, e.g., Cochran (1977)). Observing that

$$\hat{R}_N + 1 + \frac{1}{n} = \frac{\Sigma_{k\in s}\ w_k^* y_k}{\Sigma_{k\in s}\ y_k} \tag{2.7}$$

we thus have the variance estimation formula

$$\begin{aligned}\hat{V}_R &= \frac{1}{\bar{y}_s^2}\,\frac{1-f}{n}\,\frac{1}{n-1}\ \Sigma_{k\in s}\{w_k^* y_k - (\hat{R}_N + 1 + \frac{1}{n})y_k\}^2 \\ &= \frac{1}{\bar{y}_s^2}\,\frac{1-f}{n}\,\frac{1}{n-1}\sum_{i=1}^{n}(\frac{2i}{n} - \hat{R}_N - 1 - \frac{1}{n})^2\, y_{s:i}^2\end{aligned} \tag{2.8}$$

where $f = n/N$. Note that the analogy with the ratio estimator is not perfect, however, since the w_k^*'s in (2.7) are sample-dependent, in the sense that the value of w_k^* is not determined once we know k, but is determined also by the other units in s and their y-values. This random variation in the w_k^*-values is not taken care of by formula (2.8).

ii) The second method is based on the same first-order Taylor approximation as used in i) above, which suggests a variance estimation formula of the type

$$\begin{aligned}\hat{V}(\hat{R}_N) = \frac{1}{n^2\,\bar{y}_s^2}\ \{&\hat{V}(\sum_{k\in s} w_k^* y_k) + (\hat{R}_N + 1 + \frac{1}{n})^2\ \hat{V}(\sum_{k\in s} y_k) - \\ &- 2(\hat{R}_N + 1 + \frac{1}{n})\ \hat{C}(\sum_{k\in s} w_k^* y_k,\ \sum_{k\in s} y_k)\ \}\end{aligned} \tag{2.9}$$

Using $\hat{V}(\Sigma_s\ w_k^* y_k)$ and $\hat{C}(\Sigma_s\ w_k^* y_k,\ \Sigma_s y_k)$ as suggested by Nygård and Sandström (1985), which account for the randomness involved in the w_k^*-values, we obtain a somewhat complicated variance estimation formula denoted $\hat{V}_T$, defined in Table 2.1.

iii) The third method is motivated by a model-based reasoning, assuming the actual finite population to be generated by some underlying probability model. The resulting variance estimation formula, denoted $\hat{V}_A$, is a consistent estimator of the asymptotic expression for the "model-expected squared error" $\mathcal{E}(\hat{R}_N - R_N)^2$, where $\mathcal{E}$ denotes expectation with respect to the assumed model, for a fixed sample s (For details, see Sandström (1983), and Nygård and Sandström (1985)). The formula is given in Table 2.1.

Table 2.1 Variance estimators $\hat{V}_R$, $\hat{V}_T$ and $\hat{V}_A$ for the Gini coefficient.

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_n)^2 , \qquad \bar{y}_n = n^{-1} \sum_{i \in s} y_i$$

$$\hat{R}_y = \frac{2}{n^2 \bar{y}_n} \sum_{i \in s} i y_{(i)} - 1 , \qquad \hat{R}_z = \frac{2}{n^2 \bar{z}_n} \sum_{i \in s} i z_{(i)} - 1, \qquad z_i = y_i^2$$

$$\hat{B}_y = \frac{2}{n^2 \bar{y}_n} \sum_{i \in s} i y_{(i)} - 1 - \frac{1}{n} , \qquad \bar{z}_n = n^{-1} \sum_{i \in s} y_i^2$$

$$T_i = \sum_{j=1}^{i} j y_{(j)} , \qquad T_0 \equiv 0$$

$$f_n = n/N$$

$$A_n = \frac{1}{n^3} \sum_{i \in s} i^2 y_{(i)}^2 \qquad B_n = \frac{1}{n^3} \sum_{i \in s} T_{i-1} y_{(i)}$$

1. "Ratio estimator"

$$\hat{V}_R(\hat{R}_N) = \frac{(1-f_n)}{n \bar{y}_n^2} \{4(\tfrac{n}{n-1}) A_n - (s_y^2 + \bar{y}_n^2(\tfrac{n}{n-1}))[2(\hat{R}_y+1)(\hat{R}_z+1)-(\hat{R}_y+1)^2]\}$$

2. "Taylor estimator"

$$c_1 = \frac{(N-2)n(n-1)}{(N-1)(n-2)(n-3)} \qquad c_2 = \frac{4Nn^2-6(N+n)n+6n}{(N-1)(n-2)(n-3)}$$

$$c_3 = 2 - 3c_1 + \frac{c_2}{n} + \frac{2}{n-2} + \frac{n}{N-1} \qquad c_4 = \frac{5}{2} - 4c_1 + \frac{c_2}{n} + \frac{3}{n-2} + \frac{n}{N-1}$$

$$\hat{V}_T(\hat{R}_N) = \frac{(1-f_n)}{n \bar{y}_n^2} \{4c_1 A_n + 16 c_1 B_n +$$

$$+ s_y^2 [(\hat{B}_y+1)^2 - 2(\tfrac{n-1}{n-2})(\hat{B}_y+1)(\hat{R}_z+1) + 2\frac{(2n-1)}{n(n-2)}(\hat{B}_y+1) + \frac{1}{n^2} + 2\frac{(n-1)}{n^2} c_3(\hat{R}_z+1) - 4 \frac{(n-1)}{n^3} c_4] +$$

$$+ \bar{y}_n^2 [4(\tfrac{n}{n-2})(\hat{B}_y+1)^2 - c_2(\hat{R}_y+1)^2 - 2(\tfrac{n}{n-2})(\hat{B}_y+1)(\hat{R}_z+1) - 2(\tfrac{n+1}{n-2})(\hat{B}_y+1) - 4 \frac{c_2}{n^2} +$$

$$+ 4 \frac{c_2}{n}(\hat{R}_y+1) - \frac{4}{n-2} (\hat{R}_y+1) + \frac{2}{n} (\tfrac{n+4}{n-2}) - 8 \frac{c_1}{n} + 2\frac{1}{n} c_3(\hat{R}_z+1) - 4 \frac{1}{n^2} c_4]\}$$

3. "Asymptotic estimator"

$$\hat{V}_A(\hat{R}_N) = \frac{(1-f_n)}{n \bar{y}_n^2} \{4A_n + 16 B_n +$$

$$+ s_y^2 [(\tfrac{n-1}{n})(\hat{R}_y+1)^2 - 2(\tfrac{n-1}{n})(\hat{R}_y+1)(\hat{R}_z+1) + 2(\tfrac{n-1}{n^2})(\hat{R}_y+1) + 2(\tfrac{n-1}{n^2})(\hat{R}_z+1)] -$$

$$- \bar{y}_n^2 [2(\hat{R}_y+1)(\hat{R}_z+1) + 2(\tfrac{n-3}{n})(\hat{R}_y+1) + 4 \frac{(n+1)}{n^2} - 2\frac{1}{n} (\hat{R}_z+1)]\}$$

iv) The fourth method is based on a jackknife technique. One observation at a time is deleted from the sample. Each time we calculate $\hat{R}_N^{(j)}$, analogous to $\hat{R}_N$, but based only on the remaining n-1 observations (deleting the j^{th} observation; j=1, 2,..., n). The variance estimation formula is

$$\hat{V}_J = (n-1)n^{-1} \sum_{j=1}^{n} (\hat{R}_N^{(j)} - \hat{R}_N^{(\cdot)})^2$$

where $\hat{R}_N^{(\cdot)} = n^{-1} \sum_{j=1}^{n} \hat{R}_N^{(j)}$.

As $N \to \infty$, $n \to \infty$, and $n/N \to f$ $(0 < f < 1)$, the two variance estimators $\hat{V}_T$ and $\hat{V}_A$ become identical, as seen from Table 2.2.

Table 2.2 Approximated variance estimators $\hat{V}_R$, $\hat{V}_T$ and $\hat{V}_A$ for the Gini coefficient when N and n are large.

Notation: See Table 2.1.

$n \to \infty$, $N \to \infty$, $\hat{B}_y = \hat{R}_y$

$$C_n = 4A_n - (s_y^2 + \bar{y}_n^2)\ [2(\hat{R}_y+1)(\hat{R}_z+1) - (\hat{R}_y+1)^2]$$

$$D_n = 16B_n - \bar{y}_n^2\ [(\hat{R}_y+1)^2 + 2(\hat{R}_y+1)]$$

1. <u>"Ratio estimator"</u>

$$\hat{V}_R(\hat{R}_N) = \frac{(1-f_n)}{n\ \bar{y}_n^2}\ C_n$$

2. <u>"Taylor estimator"</u>

$$\hat{V}_T(\hat{R}_N) = \frac{(1-f_n)}{n\ \bar{y}_n^2}\ \{C_n + D_n\}$$

3. <u>"Asymptotic estimator"</u>

$$\hat{V}_A(\hat{R}_N) = \frac{(1-f_n)}{n\ \bar{y}_n^2}\ \{C_n + D_n\}$$

In Section 3, exact sampling distributions of $\hat{R}_N$ and the variance estimators $\hat{V}_R$, $\hat{V}_T$, $\hat{V}_A$ and $\hat{V}_J$ will be obtained under simple random sampling of n = 5 units from various finite populations of size N = 11. The finite populations are constructed to illustrate how the sampling performance of $\hat{R}_N$ and of the variance estimators are becoming more and more erratic, as the population mean $\bar{y}$ approaches zero.

In Section 4, the sampling behavior of $\hat{R}_N$ and of the variance estimators for larger sample sizes will be studied in a simulation experiment involving simple random sampling with replacement from finite populations of various shapes, such as Pareto, Normal, Lognormal and Weibull.

3. Small Populations: Studies of the Exact Sampling Distributions

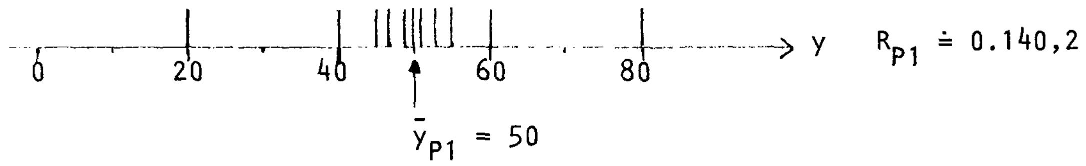
To illustrate the behavior of the point estimator $\hat{R}_N$ of the finite population Gini coefficient and the four variance estimators when the location parameter $\bar{y}_N = N^{-1} \sum_{i=1}^{N} y_i$ is moved towards zero we will make use of two small populations, both of the size N=11. The first population (P1) is symmetric and the second (P2) is positive skew. P1 and P2 are depicted in Figure 3.1.

The arithmetic means in the two populations are $\bar{y}_{P1} = 50$ and $\bar{y}_{P2} = 395/11 \approx 35.91$, respectively, and their Gini coefficients are $R_{P1} = 424/3025 \approx 0.1402$ and $R_{P2} = 232/869 \approx 0.2670$, respectively.

The sample design to be used is srs without replacement with a sample size of n = 5, i.e. the total number of possible samples is 462. New populations have been obtained by letting the location parameters tend towards zero. Totally we have four populations based on P1 and three based on P2. They are all summarized in Table 3.1.

Figure 3.1 The two small parent populations (N=11) used in this study.

P1: Parent Population



P2: Parent Population

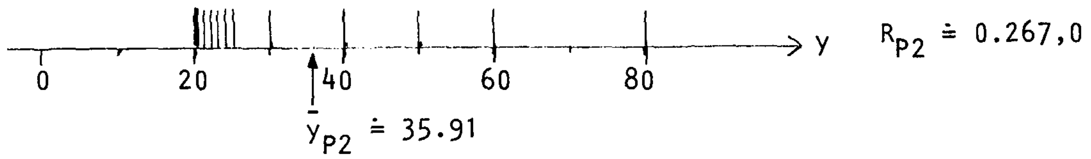


Table 3.1 A summary of the four symmetric populations (P1) and the three skew populations (P2) from which the samples are taken and some characteristics of the sample distributions of $\hat{R}_N$

	Parent Population						
	$P1_1$	$P1_2$	$P1_3$	$P1_4$	$P2_1$	$P2_2$	$P2_3$
Translations	Y_{P1}	Y_{P1}-34.9	Y_{P1}-39.9	Y_{P1}-49.9	Y_{P2}	Y_{P2}-24.7	Y_{P2}-35.7
$\bar{y}_N$	50	15.1	10.1	0.1	35.91	11.21	0.21
R_N	0.1402	0.4641	0.6939	70.083	0.2670	0.8708	45.85
$E(\hat{R}_N)$	0.1245	0.4606	0.9571	0.0915	0.2270	0.8211	-0.2265
Bias	-0.0157	-0.0035	+0.2632	-69.991	-0.0400	-0.0497	-46.08
$V(\hat{R}_N)$	0.002931	0.0635	1.5808	250.6	0.003649	1.418	287.3
Min $\hat{R}_N$	0.0224	0.0718	0.1026	-112	0.03478	-14.4	-101.6
Max $\hat{R}_N$	0.2285	0.9811	17.3333	+112	0.3133	11.2	100.8

Since $R_N = \frac{G_N}{2\bar{y}_N}$, and G_N is location-free, all changes are due to changes in the location parameter $\bar{y}_N$. Note that R_N is not bounded to $[0,1-1/N]$ when $Y \in]-\infty,\infty[$. The sampling distributions of $\hat{R}_N$ are given in Figure 3.2.

In Table 3.1 we have also summarized the results of the sampling distributions of $\hat{R}_N$. It is notable that the bias is of the same order of magnitude (with change of sign) as the true value of R_N when $\bar{y}_N$ is close to zero. This indicates that one should be very careful in estimating Gini coefficients from populations with low arithmetic means ($Y \in]-\infty,\infty[$).

REMARK 3.1 It is possible to have distributions with both negative and positive incomes if the income definition is based on e.g. the rules of the taxation system. At least one of the definitions of the entrepreneurial income in the Swedish income distribution surveys has this property.

In Table 3.2 we have summarized the results of the sampling distributions of the four variance estimators. In the first three symmetric populations ($P1_1$ - P_3) the asymptotic variance estimator $\hat{V}_A$ seems to be the best as measured by the Relative Mean (=Relative Bias + 1). In $P1_4$ (where $\bar{y}_N$=0.1) the jackknife variance estimator is best according to the Relative Mean. In the skew populations the jackknife estimator performs best according to the same criteria. The bad performance of the ratio estimator $\hat{V}_R$ in $P1_1$ and $P2_1$ is confirmed in the simulation study in Section 4. The possibility of negative Taylor variance estimators, $\hat{V}_T$, is also confirmed in Section 4 (at least for small samples).

The sample distributions of the four variance estimators from $P1_1$ and $P2_1$ are illustrated in Figure 3.3.

The results of this study shows that whenever the location parameter $\bar{y}_N$ is close to zero, with y taking on both negative and positive values,

Figure 3.2 The sampling distributions of $\hat{R}_N$ when sampling (srs) without replacement from P1 and P2. Note that the scales on $\hat{R}_N$ are different.

Symmetric parent population P1 | Skew parent population P2

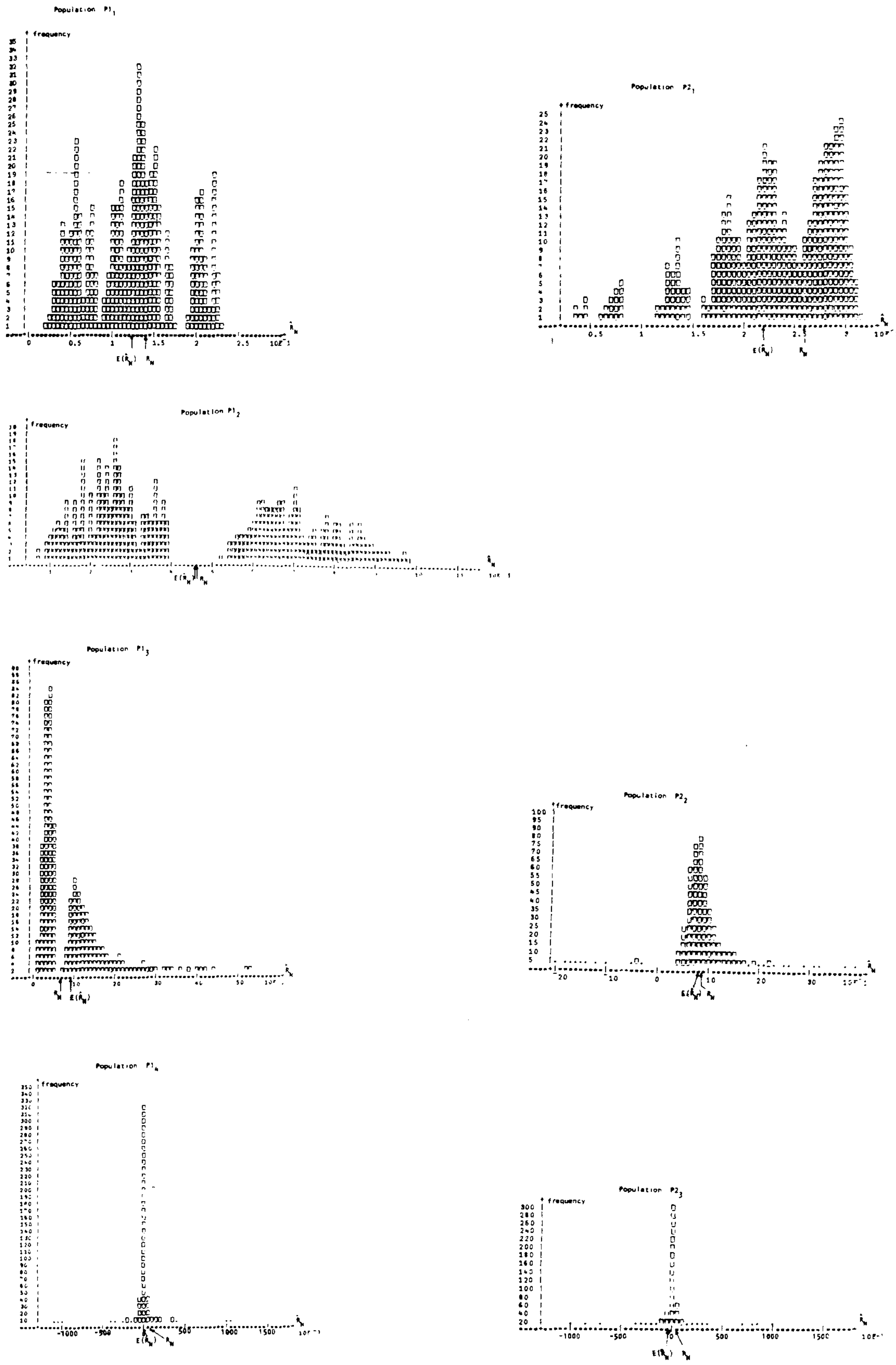


Table 3.2 A summary of the sampling distributions of the four variance estimators $\hat{V}_R$, $\hat{V}_T$, $\hat{V}_A$ and $\hat{V}_J$.

	Parent Population						
	$P1_1$	$P1_2$	$P1_3$	$P1_4$	$P2_1$	$P2_2$	$P2_3$
$V(\hat{R}_N)$	0.0029	0.0635	1.5808	250.6	0.0036	1.418	287.3
$E(\hat{V}(\hat{R}_N))$							
$\hat{V}_R$	0.0416	0.1188	145.04	990760	0.0452	76.9	1286000
$\hat{V}_T$	0.0033	0.1678	148.62	990850	0.0026	77.1	1286000
$\hat{V}_A$	0.0026	0.1517	119.93	792630	0.0012	61.71	1029000
$\hat{V}_J$	0.0049	0.1809	163.49	7486	0.0087	12.45	1245
REL.MEAN = $E(\hat{V})/V$							
$\hat{V}_R$	14.34	1.87	91.75	3953.55	12.56	54.23	4476.16
$\hat{V}_T$	1.14	2.64	94.02	3953.91	0.72	54.37	4476.16
$\hat{V}_A$	0.90	2.39	75.87	3162.93	0.33	43.52	3581.62
$\hat{V}_J$	1.69	2.85	103.42	29.87	2.42	8.78	4.33
Min $\hat{V}$ / Max $\hat{V}$							
$\hat{V}_R$	0.0335	0.0319	0.0214	0.0028	0.0341	0.0040	0.0267
	0.0515	0.8817	53068	68461000	0.0594	23030	71600000
$\hat{V}_T$	0.000008	0.000111	0.000231	0.0030	-0.002716	0.0041	0.000026
	0.0080	1.348	54021	68468000	0.0159	23450	71850000
$\hat{V}_A$	0.000009	0.000106	0.000246	0.0028	0.000046	0.0016	0.000085
	0.0068	1.1469	43467	54934000	0.0055	17960	57060000
$\hat{V}_J$	0.000044	0.000374	0.000713	0.0094	0.000083	0.0078	0.000238
	0.0114	1.2584	4585.5	14940	0.0514	385.7	32580
Coverage rate for the 462 confidence intervals of type $\hat{R}_N \pm 1.96\ \hat{V}^{1/2}$							
$\hat{V}_R$	100 %	100 %	79.9 %	15.2 %	100 %	85.9 %	18.2 %
$\hat{V}_T$	72.7 %	60.2 %	53.0 %	15.2 %	59.1 %	86.6 %	18.2 %
$\hat{V}_A$	65.2 %	57.6 %	55.8 %	14.1 %	64.7 %	94.8 %	18.2 %
$\hat{V}_J$	72.7 %	69.0 %	62.8 %	9.1 %	87.9 %	94.4 %	9.5 %

Figure 3.3 Sampling distributions of variance estimators for the estimated Gini coefficient.

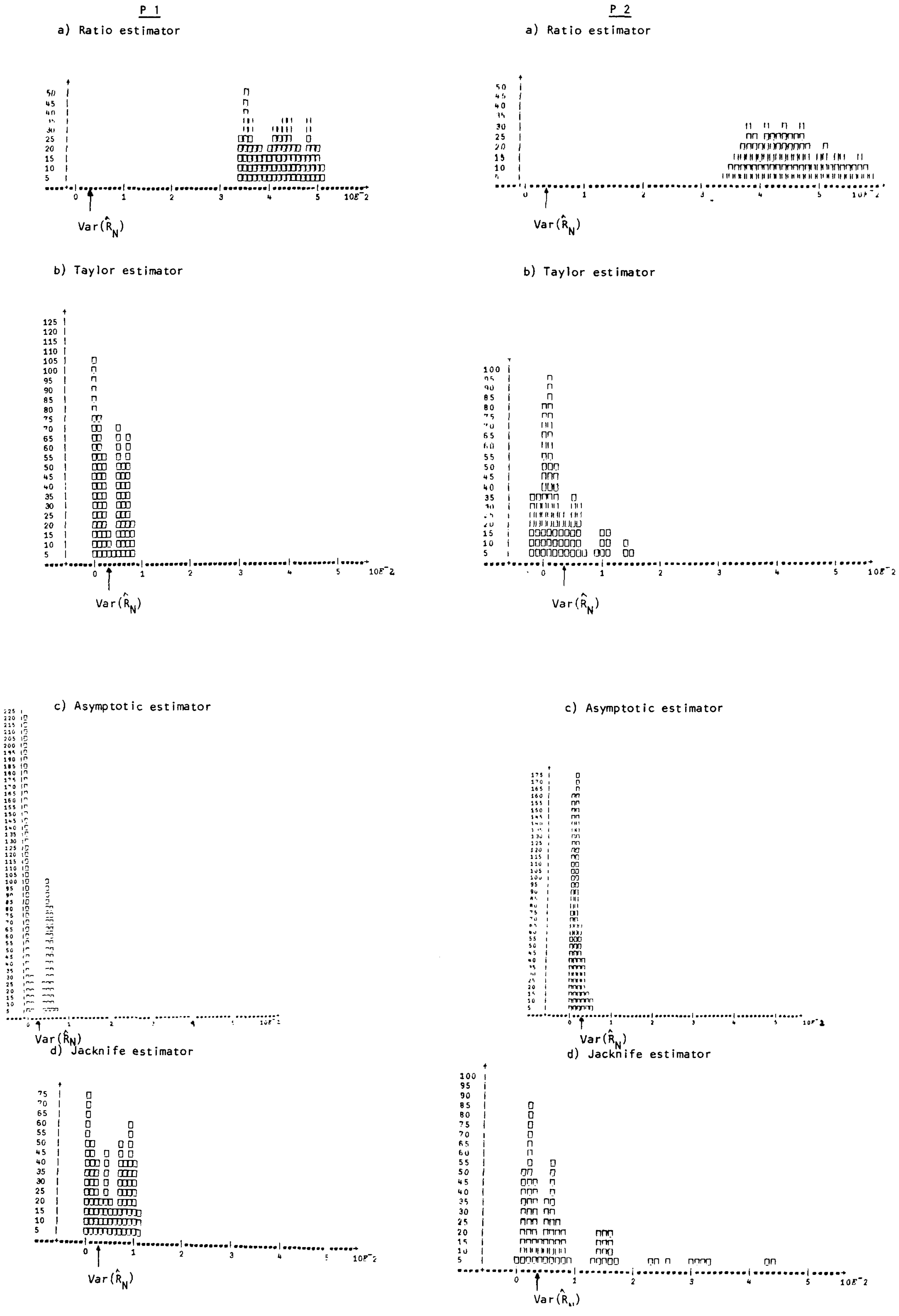


Table 3.3 The correlation coefficients between the variance estimators and between $\hat{R}_N$ and the variance estimators

Correlation coefficient of	Parent Population					
	$P1_1$	$P1_2$	$P1_4$	$P2_1$	$P2_2$	$P2_3$
$\hat{V}_R$ vs $\hat{V}_T$	-0.70	0.99	1.00	0.83	1.00	1.00
$\hat{V}_R$ vs $\hat{V}_A$	-0.85	0.98	1.00	-0.75	1.00	1.00
$\hat{V}_R$ vs $\hat{V}_J$	-0.61	0.97	0.47	0.76	0.27	-0.04
$\hat{V}_T$ vs $\hat{V}_A$	0.96	1.00	1.00	-0.33	1.00	1.00
$\hat{V}_T$ vs $\hat{V}_J$	0.99	0.97	0.47	0.96	0.28	-0.04
$\hat{V}_A$ vs $\hat{V}_J$	0.92	0.99	0.48	-0.29	0.27	-0.04
$\hat{R}_N$ vs $\hat{V}_R$	-0.44	0.75	-0.02	-0.18	-0.38	-0.03
$\hat{R}_N$ vs $\hat{V}_T$	0.82	0.83	-0.02	0.14	-0.39	-0.03
$\hat{R}_N$ vs $\hat{V}_A$	0.77	0.86	-0.01	0.40	-0.37	-0.02
$\hat{R}_N$ vs $\hat{V}_J$	0.86	0.86	0.13	0.32	-0.45	-0.02

one should be careful in trying to estimate the Gini coefficient $R_N = \frac{G_N}{2\bar{y}_N}$ and the variance of $\hat{R}_N$. It may be observed that the correlation between $\hat{G}_N$ and $\hat{\bar{Y}}_N$ is 0 and 0.9041 when the samples are taken from P1 and P2, respectively. The zero correlation follows theoretically from a Theorem by Hogg (1960). The correlations between the variance estimators are given in Table 3.3 together with the correlations between $\hat{R}_N$ and the variance estimators.

4. Large Populations: Monte Carlo Studies of the Sampling Distributions

To illustrate the performance of the four variance estimators $\hat{V}_R$, $\hat{V}_T$, $\hat{V}_A$ and $\hat{V}_J$ a Monte Carlo study was designed. Eleven continuous parent distributions were used, viz. Logistic, Uniform, Normal, Lognormal, Pareto and standard Weibull (6 values on the parameter α). To each continuous df we constructed a finite population from which the samples were taken. The population size was N = 10 000. The finite normal parent population was constructed by use of the Box-Muller method and from this parent population the finite lognormal parent was obtained. For the other distributions the values of the finite population were obtained through the inverse df F_*^{-1}, where F_* is the rounded F $[0,1]$ such that $F_* = 10^{-5}(10k + 5)$, k = 0(1)9999.

In Table 4.1 we have summarized the df's under consideration together with the formulas for the Gini coefficient R and its values according to the selected parameters of the df's. The values of the Gini coefficient in the finite population approximations are also given in Table 4.1. The deviation between the latter values and those obtained from the continuous df's indicate how good our finite population approximations are.

The sampling design used was simple random sampling (srs) with replacement. The sample sizes were n = 5, 10, 20, and 100 and all simulations are based on 500 replicates. All computations were made in APL on the IBM 370 at Statistics Sweden.

Table 4.1 The df's used in the simulation study together with its specific parameter values and the Gini coefficient of F and its finite population approximation F_N, N = 10,000

Distribution function F(x)	Parameters	Theoretical Gini coefficient	Gini coefficients in the specific df. R	the finite population R_N
1. LOGISTIC				
$\{1+e^{-(x-\alpha)/\beta}\}^{-1}$	$\alpha = 5$	β/α	$\frac{1}{5} = 0.2$	0.1998
$-\infty<x<\infty, -\infty<\alpha<\infty$ $\beta>0$	$\beta = 1$			
2. UNIFORM				
$\frac{x-\alpha}{\beta-\alpha}$	$\alpha = 4$	$\frac{1}{3}\left(\frac{\beta-\alpha}{\beta+\alpha}\right)$	$\frac{1}{15} \doteq 0.0666$	0.0666
$\alpha<x<\beta$	$\beta = 6$			
3. NORMAL				
$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2}\,dt$	$\mu = 5$	$\frac{\sigma}{\mu\sqrt{\pi}}$	$\frac{1}{5\sqrt{\pi}} \doteq 0.1128$	0.1127
$-\infty<x<\infty, -\infty<\mu<\infty$ $\sigma^2>0$	$\sigma^2 = 1$			
4. LOGNORMAL				
$\frac{1}{\sigma\sqrt{2\pi}}\int_{0}^{x}\frac{1}{y}e^{-(\log y-\mu)/2\sigma^2}\,dy$	$\mu = 5$ $\sigma^2 = 1$	$2N(\frac{\sigma}{\sqrt{2}}; 0,1)-1$	$\doteq 0.5204$	0,5185
$X=e^U, U\sim N(\mu,\sigma^2)$				
$x>0, -\infty<\mu<\infty$				
$\sigma^2>0$				
5. Pareto				
$1 - a^\alpha x^{-\alpha}$	$a = 1$	$\frac{1}{2\alpha-1}$	$\frac{1}{2} = 0.5$	0.4886
$x\geq a, \alpha>0$	$\alpha = 1.5$			
6. Weibull				
$1-e^{-(\frac{1}{\beta}x)^\alpha}$	$\beta = 1$	$1 - \frac{1}{2^{1/\alpha}}$		
$x>0, \beta>0, \alpha>0$	$\alpha =$			
	0.8		0.5796	0.5795
(Exponential)	1.0		1/2=0.5	0.5000
(Rayleigh)	2.0		0.2929	0.2929
	3.6		0.1751	0.1751
	4.8		0.1345	0.1345
	10.0		0.06697	0.06697

Table 4.2 Estimates of $E(\hat{R}_N)$ and $V(\hat{R}_N)$ based on 500 replicates for eleven parent distributions. The estimates are denoted $Ea(\hat{R}_N)$ and $Va(\hat{R}_N)$.

	Parent Population		Sample size, n = 5	10	20	100	R_N
$Ea(\hat{R}_N)$	Logistic		.1645	.1798	.1896	.1994	.1998
	Uniform		.0545	.0595	.0637	.0656	.0666
	Normal		.0893	.1019	.1078	.1115	.1127
	Lognormal		.3886	.4456	.4817	.5053	.5185
	Pareto		.2909	.3458	.3770	.4515	.4886
	Weibull	α= 0.8	.4615	.5201	.5477	.5730	.5795
	Weibull	α= 1.0	.3976	.4514	.4759	.4940	.5000
	Weibull	α= 2.0	.2442	.2648	.2827	.2907	.2929
	Weibull	α= 3.6	.1482	.1607	.1688	.1734	.1751
	Weibull	α= 4.8	.1093	.1216	.1302	.1324	.1345
	Weibull	α=10.0	.0546	.0612	.0639	.0666	.0670
$Va(\hat{R}_N)$ $10^{-6}\times$	Logistic		72920	34130	16360	360	
	Uniform		227	100	49	10	
	Normal		1232	617	334	65	
	Lognormal		15990	9146	6251	1390	
	Pareto		25820	21690	15390	9305	
	Weibull	α= 0.8	14850	8955	4686	971	
	Weibull	α= 1.0	13830	7992	3876	895	
	Weibull	α= 2.0	7130	3504	1830	404	
	Weibull	α= 3.6	2635	1632	820	160	
	Weibull	α= 4.8	2013	1017	534	118	
	Weibull	α=10.0	540	251	147	34	

4.1 Point Estimates

Estimates of $E(\hat{R}_N)$, based on 500 replicates, are given in Table 4.2 for the samples (n = 5, 10, 20, and 100) taken from the eleven parent populations. In the same table we have also given the variances among the 500 replicated estimates $\hat{R}_N$.

To illustrate the effect of the sample size on the bias of the estimated $E(\hat{R}_N)$, Figure 4.1 gives the relative mean (Rel Mean) together with the relative maximum and minimum. The two latter ratios are computed as Max $\hat{R}_N/R_N$ and Min $\hat{R}_N/R_N$, respectively, where the maximum and minimum values are taken among the 500 estimates $\hat{R}_N$. As is seen in the figure we are, on the average, underestimating R_N irrespective of the parent population and the Rel Mean increases most slowly when data is taken from parent populations with positive skewness. For samples of the size n = 100 and greater the bias is negligible. For small samples the bias could be reduced considerably if we defined $R_N^* = \frac{N}{N-1} R_N$ and $\hat{R}_N^* = \frac{n}{n-1} \hat{R}_N$.

The appearance of the approximate sampling distributions for $\hat{R}_N$ taken from the eleven populations is illustrated in Figure 4.2 when the sample size equals n = 20. The coefficients of skewness of these distributions are given in Table 4.3 for n = 5, 10, 20, and 100. It is remarkable how symmetric these distributions are when the parent populations are skew.

4.2 Variance Estimates

Of the four variance estimators, $\hat{V}_R$, $\hat{V}_T$, $\hat{V}_A$, and $\hat{V}_J$, only the 'Ratio estimator' $\hat{V}_R$ considers the weights to the incomes in the Gini coefficient as constants. As is seen in Table 4.4 and Figure 4.3 this fact highly affects the variance estimation. The Rel Mean of the estimated $E(\hat{V}_R)$, relative to V, is a factor 2-300 times higher than the Rel Mean of the other three variance estimators! $V = V(\hat{R}_N)$ is the variance taken over the 500 replicated estimates $\hat{R}_N$. It is notable that the Rel Mean of the Ratio estimator $\hat{V}_R$ is lowest when the data is taken from positi-

Figure 4.1 The estimated $E(\hat{R}_N)$ relative to R_N and the relative maximum and minimum, based on 500 replicates of sample sizes $n = 5, 10, 20$, and 100.

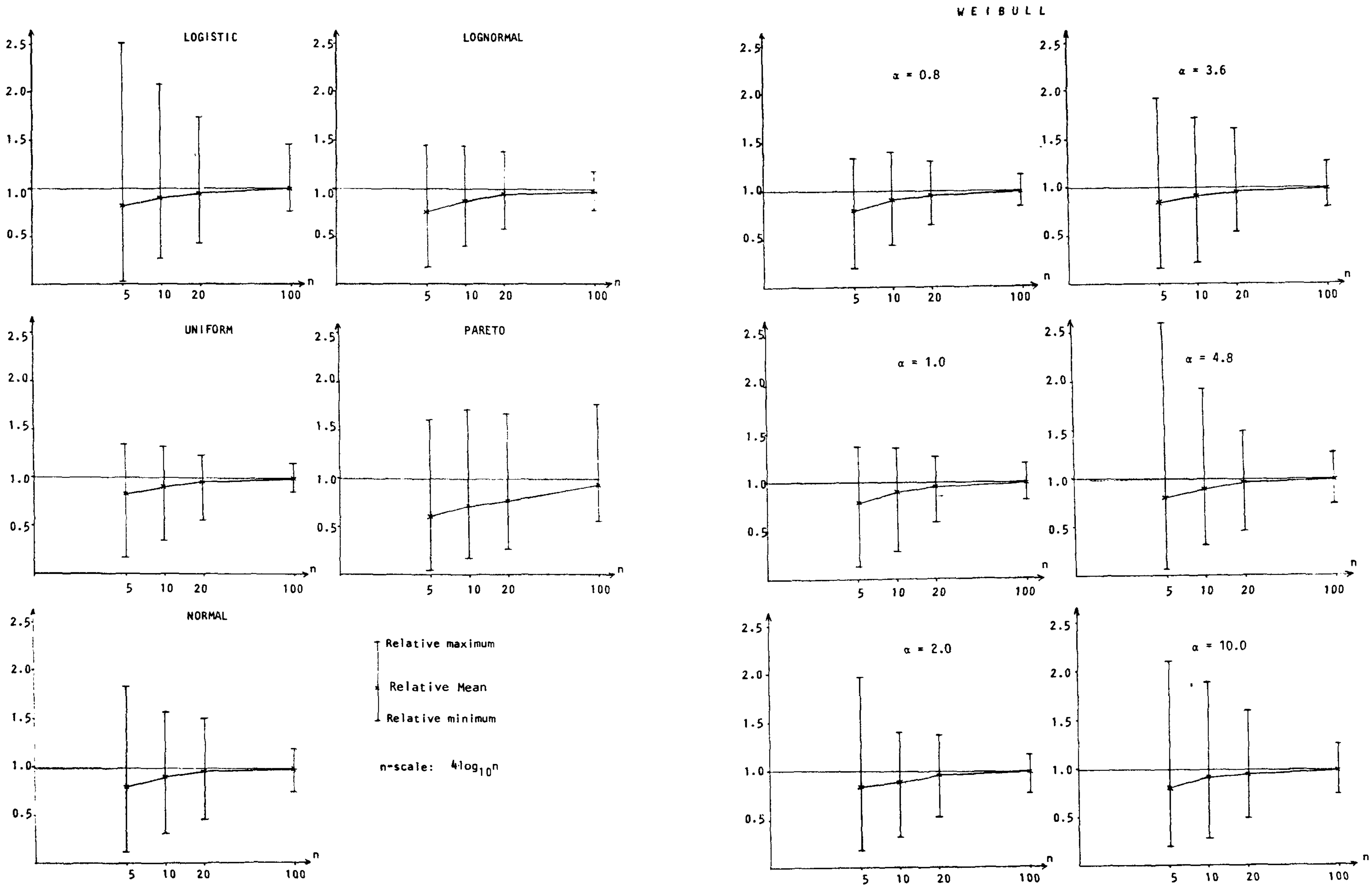


Figure 4.2 The simulated sampling distributions of $\hat{R}_N$ based on 500 replicates of sample size n = 20.

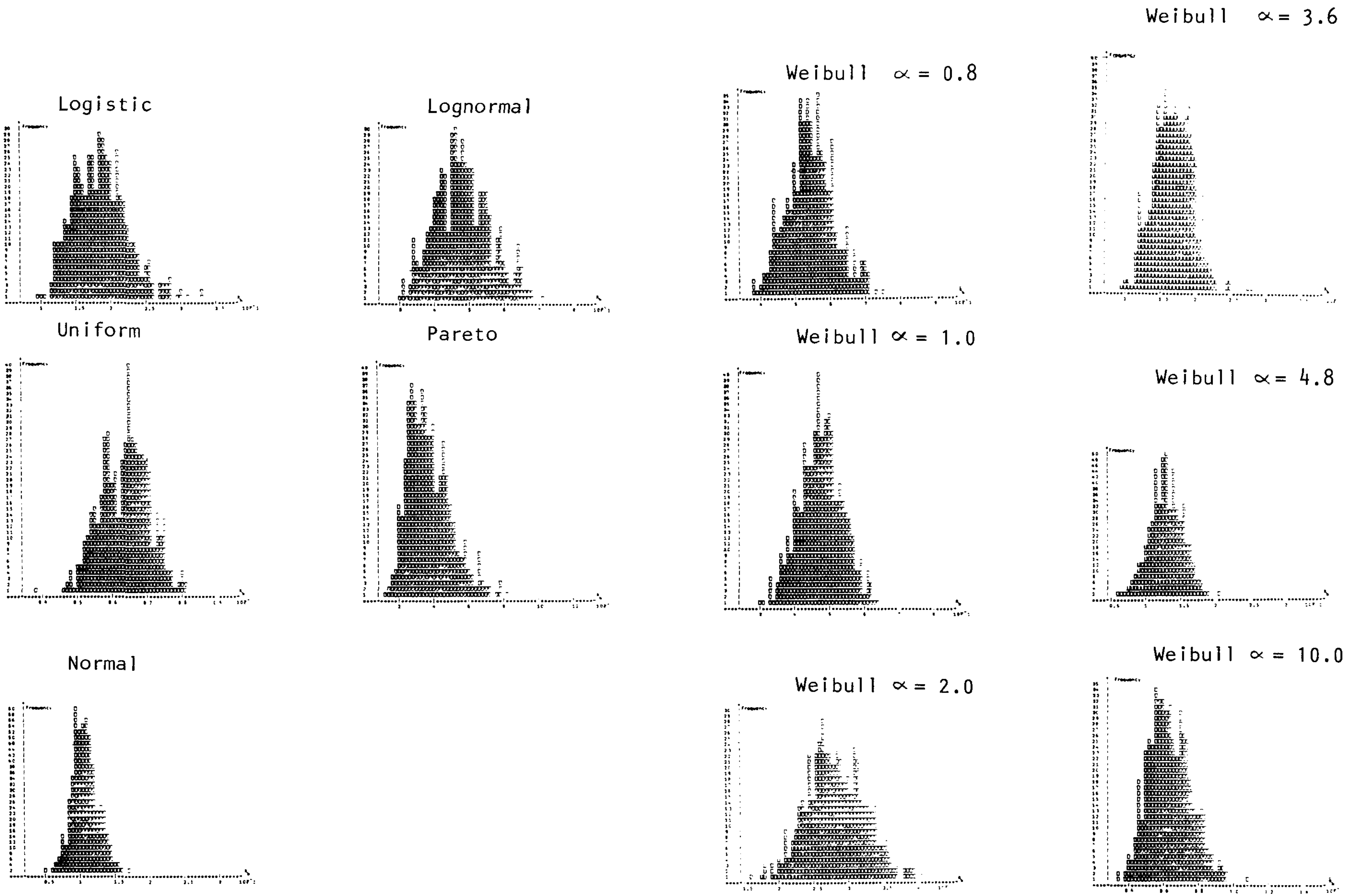


Table 4.3 The coefficient of skewness of the simulated sampling distirbutions of R_N when the sample size is $n = 5, 10, 20$ and 100.

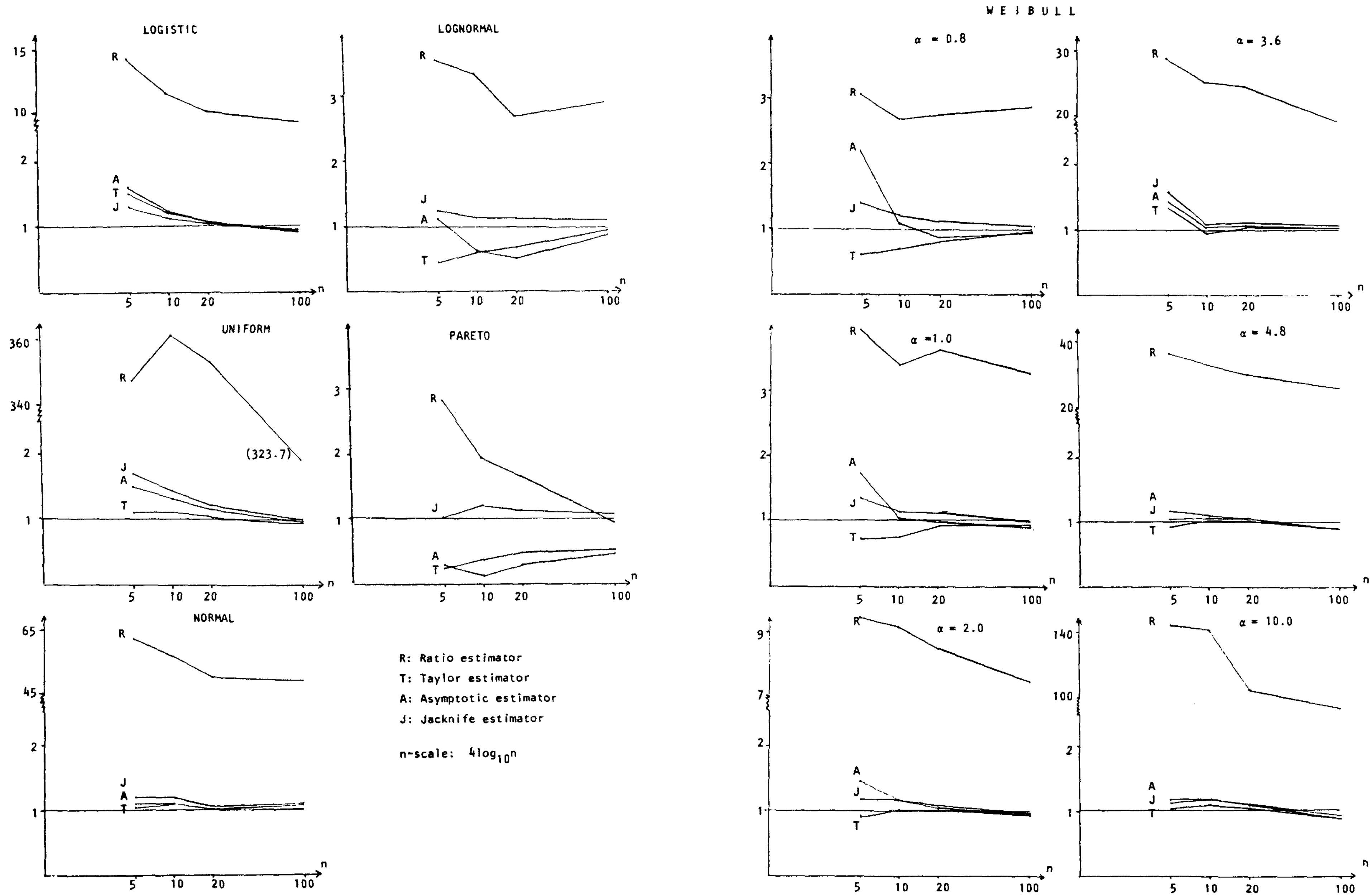
Parent Popularion		Sample size, n = 5	10	20	100
Logistic		1.061	0.932	0.571	0.561
Uniform		-0.213	-0.294	-0.130	-0.106
Normal		0.446	0.318	0.237	-0.087
Lognormal		0.142	0.075	0.236	0.159
Pareto		0.901	0.728	0.754	1.226
Weibull	α= 0.8	-0.171	-0.190	0.129	0.016
Weibull	α= 1.0	0.029	-0.151	-0.026	0.108
Weibull	α= 2.0	0.319	0.037	0.124	0.024
Weibull	α= 3.6	0.416	0.103	0.248	0.386
Weibull	α= 4.8	0.903	0.412	-0.069	0.194
Weibull	α=10.0	0.757	0.490	0.304	0.058

Table 4.4 Estimates of $V(\hat{R}_N)$ based on 500 replicates of sample sizes n = 5, 10, 20 and 100. V is the variance among the 500 replicates of $\hat{R}_N$.

Parent Population	Variance	Sample size, n = 5	10	20	100
Logistic	V	5129	2915	1611	363
10^{-6} ×	$\hat{V}_R$	72920	34130	16360	3152
	$\hat{V}_T$	7727	3548	1715	334
	$\hat{V}_A$	8221	3576	1708	332
	$\hat{V}_J$	6649	3331	1679	336
Uniform	V	227	100	49	10
10^{-6} ×	$\hat{V}_R$	78960	36160	17230	3282
	$\hat{V}_T$	251	110	50	10
	$\hat{V}_A$	342	132	56	10
	$\hat{V}_J$	389	144	59	10
Normal	V	1232	617	334	65
10^{-6} ×	$\hat{V}_R$	77540	35510	16910	3233
	$\hat{V}_J$	1266	672	336	70
	$\hat{V}_A$	1325	671	336	69
	$\hat{V}_J$	1436	728	351	71
Lognormal	V	15990	9146	6251	1073
10^{-6} ×	$\hat{V}_R$	57390	30520	16960	4073
	$\hat{V}_T$	7302	5545	4204	1332
	$\hat{V}_A$	18100	5723	3258	1235
	$\hat{V}_J$	19990	10600	7102	1547
Pareto	V	25820	21690	15390	9305
10^{-6} ×	$\hat{V}_R$	73670	42410	25380	8844
	$\hat{V}_T$	6519	8123	7204	4913
	$\hat{V}_A$	7354	2740	4090	4403
	$\hat{V}_J$	26030	25990	17430	10050

Parent Population	Variance	Sample size, n = 5	10	20	100
Weibull	V	14850	8955	4686	971
α= 0.8	$\hat{V}_R$	45980	24260	13020	2820
10^{-6} ×	$\hat{V}_T$	9272	6356	3827	944
	$\hat{V}_A$	33140	9994	4172	930
	$\hat{V}_J$	21060	11110	5327	1029
Weibull	V	13830	7902	3876	895
α= 1.0	$\hat{V}_R$	54830	27450	14080	2935
10^{-6} ×	$\hat{V}_T$	10140	5984	3511	814
	$\hat{V}_A$	24180	8197	3768	807
	$\hat{V}_J$	18810	9083	4327	862
Weibull	V	7130	3504	1830	404
α= 2.0	$\hat{V}_R$	67720	32310	15630	3022
10^{-6} ×	$\hat{V}_T$	6679	3586	1847	381
	$\hat{V}_A$	10390	4104	1937	383
	$\hat{V}_J$	8502	4085	1975	390
Weibull	V	2635	1632	820	160
α= 3.5	$\hat{V}_R$	73840	34070	16060	3097
10^{-6} ×	$\hat{V}_T$	3560	1599	850	166
	$\hat{V}_A$	4199	1753	890	167
	$\hat{V}_J$	3801	1714	876	169
Weibull	V	2013	1017	534	118
α= 4.8	$\hat{V}_R$	75300	34360	16410	3131
10^{-6} ×	$\hat{V}_T$	1872	1033	539	104
	$\hat{V}_A$	2338	1122	577	104
	$\hat{V}_J$	2094	1085	554	105
Weibull	V	540	251	147	33
α=10.0	$\hat{V}_R$	77940	35520	16850	3212
10^{-6} ×	$\hat{V}_T$	558	277	154	30
	$\hat{V}_A$	627	293	157	31
	$\hat{V}_J$	597	293	158	30

Figure 4.3 The Relative **Mean** of the estimated $E(\hat{V}(\hat{R}_N))$ relative to $V(\hat{R}_N)$, where the latter variance is taken over 500 replicates of $\hat{R}_N$. Sample sizes: 5, 10, 20, and 100.



vely skewed populations - the Rel Mean is here of order 0.95-4.0 - but when data is taken from symmetrical or negatively skewed populations the Rel Mean is of order 10-360. The highest values are obtained when data comes from the uniform parent population.

In Table 4.4 the estimated values of $E(\hat{V})$ are given for the four estimators. The coverage rates for the 500 confidence intervals of type $\hat{R}_N \pm 1.96\ \hat{V}^{\frac{1}{2}}$ are given in Table 4.5. If we reduce the bias using the factor $n/(n-1)$ times the endpoints of the confidence interval $\hat{R}_N \pm 1.96\ \hat{V}^{\frac{1}{2}}$ then the coverage rates will increase. In the Normal parent population case the increase is 10-15% when the sample size is n=5 but only 1% when n=100 and in the Lognormal case the increase is 9-17% if n=5 but 1-2% when n=100. This is valid for $\hat{V}_T$, $\hat{V}_A$ and $\hat{V}_J$. In the $\hat{V}_R$ case small decreases was also obtained.

From these facts we may conclude that the Ratio estimator, $\hat{V}_R$, is a bad estimator of V at least in small samples ($n \leq 20$). The Taylor estimator ($\hat{V}_T$), on the other hand, shows to have a defect for sample sizes n<20, viz. that it takes on negative values!

To illustrate this fact and also showing the discrepancies between the four estimators their approximated sampling distributions for sample sizes n=5, 10, and 20 are given in Figure 4.4 for the Lognormal parent population. The figures show the general tendency which is attained in all the approximated sampling distributions from the eleven parent populations: $\hat{V}_T$ takes on negative values when n<20, but not for n>20. $\hat{V}_R$ and $\hat{V}_T$ are more flatted when n=5 than $\hat{V}_A$, and $\hat{V}_J$, and $\hat{V}_R$ has larger variance than $\hat{V}_T$ when n>20.

Table 4.5 Coverage rates (%) for the 500 confidence intervals of type $\hat{R}_N \pm 1.96\ \hat{V}^{1/2}$ of sample sizes n = 5, 10, 20 and 100.

Parent Population	Variance	Sample size, n = 5	10	20	100
Logistic	$\hat{V}_R$	100	100	100	100
	$\hat{V}_T$	64.8	78.8	85.6	93.8
	$\hat{V}_A$	70.6	79.0	85.6	93.4
	$\hat{V}_J$	73.6	81.0	86.2	94.2
Uniform	$\hat{V}_R$	100	100	100	100
	$\hat{V}_T$	53.0	78.4	90.8	93.0
	$\hat{V}_A$	77.4	87.2	93.2	93.2
	V_J	79.0	89.6	94.4	93.6
Normal	$\hat{V}_R$	100	100	100	100
	$\hat{V}_T$	56.8	79.6	89.6	94.2
	$\hat{V}_A$	64.2	81.8	89.2	94.2
	$\hat{V}_J$	69.8	84.4	90.6	94.4
Lognormal	$\hat{V}_R$	99.2	99.8	100	99.8
	$\hat{V}_T$	50.2	65.0	77.4	86.6
	$\hat{V}_A$	70.2	73.2	76.0	86.2
	$\hat{V}_J$	69.2	76.6	80.4	87.6
Pareto	$\hat{V}_R$	98	97	93.6	89.4
	$\hat{V}_T$	29.8	43.2	48.2	68.4
	$\hat{V}_A$	26.8	29.2	40.6	66.8
	$\hat{V}_J$	42.8	53.6	54.4	72.0
Weibull $\alpha = 0.8$	$\hat{V}_R$	99.6	99.8	100	100
	$\hat{V}_T$	59.4	79.6	86.2	93
	$\hat{V}_A$	86.6	89.8	87.2	92.8
	$\hat{V}_J$	82.8	87.8	88.2	94
Weibull $\alpha = 1.0$	$\hat{V}_R$	100	99.8	100	100
	$\hat{V}_T$	62.8	77.8	87.6	92.2
	$\hat{V}_A$	80.6	88.2	90.4	92.2
	$\hat{V}_J$	81.6	86.0	89.4	92.6
Weibull $\alpha = 2.0$	$\hat{V}_R$	100	100	100	100
	$\hat{V}_T$	63.0	83.4	91.2	94.4
	$\hat{V}_A$	80.4	86.2	91.8	94.2
	$\hat{V}_J$	80.0	87.4	91.6	94.4
Weibull $\alpha = 3.6$	$\hat{V}_R$	100	100	100	100
	$\hat{V}_T$	63.6	79.8	89.2	94
	$\hat{V}_A$	76.8	81.4	89.8	94
	$\hat{V}_J$	78.6	83.6	90.4	94.2
Weibull $\alpha = 4.8$	$\hat{V}_R$	100	100	100	100
	$\hat{V}_T$	57.2	79.2	89.4	90.2
	$\hat{V}_A$	71.6	82.4	89.2	90
	$\hat{V}_J$	73.0	82.2	89.8	90.2
Weibull $\alpha = 10.0$	$\hat{V}_R$	100	100	100	100
	$\hat{V}_T$	54.6	80.6	86.0	91.8
	$\hat{V}_A$	67.6	82.2	87.0	91.8
	$\hat{V}_J$	70.0	83.2	86.8	91.8

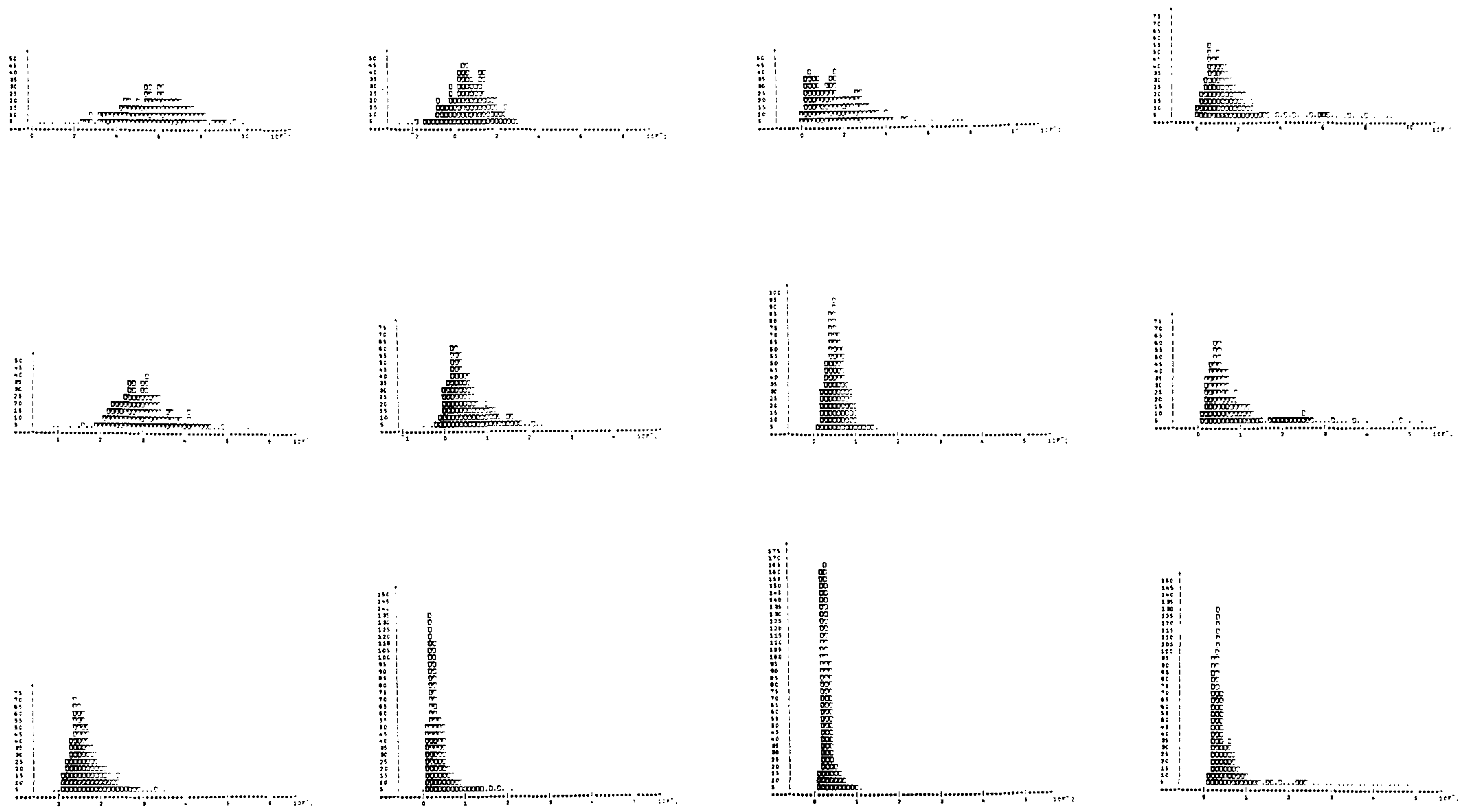
Figure 4.4 Approximated sampling distribution of the four variance estimators based on 500 replicates from the Lognormal parent population. Sample sizes, n = 5, 10 and 20.

Ratio estimator, $\hat{V}_R$ | Taylor estimator, $\hat{V}_T$ | Asymptotic estimator, $\hat{V}_A$ | Jacknife estimator, $\hat{V}_J$

n = 5

n = 10

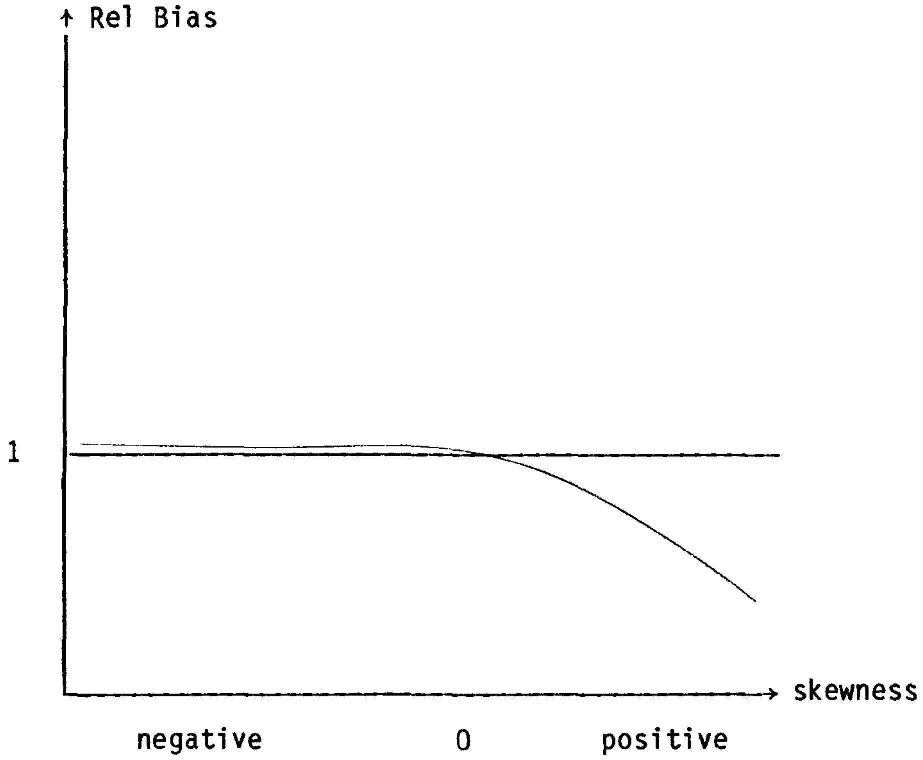
n = 20



From the simulations made on the variance estimators the following schedule of conclusions may be done:

Parent Population	
Symmetric	Asymmetric
. the Rel Mean tends towards 1, as the sample size increases: quite fast	. Rel Mean tends towards 1, as the sample size increases: slowly
. $\hat{V}$ overestimates V	. when the parent population is positively skewed $\hat{V}_T$ underestimates V, $\hat{V}_A$ underestimates V in some cases. $\hat{V}_R$ and $\hat{V}_J$ overestimates V.
. the largest Rel Mean is obtained by $\hat{V}_R$ (a factor 10-360)	. the largest Rel Mean is obtained by $\hat{V}_R$ (a factor 1-10 for positively skewed parent populations and a factor 20-140 for negatively skewed populations)
. the sampling distribution of $\hat{V}$: n=5 positively skew n=10 less variance n=20 symmetrical, with tail n=100 "-	. the sampling distribution of $\hat{V}$: n=5 flat, $\hat{V}_T$ often negative n=10 less flat, $\hat{V}_T$ can be negative n=20 peaked, $\hat{V}_T$ not negative n=100 "-
. According to the results, the best variance estimates are n=5 $\hat{V}_A$, $\hat{V}_J$ n=10 $\hat{V}_A$, $\hat{V}_J$, $\hat{V}_T$ n=20 $\hat{V}_A$, $\hat{V}_J$, $\hat{V}_T$ n=100 $\hat{V}_A$, $\hat{V}_J$, $\hat{V}_T$	. According to the results, the best variance estimates are n=5 $\hat{V}_A$, $\hat{V}_J$ n=10 $\hat{V}_A$, $\hat{V}_J$ n=20 $\hat{V}_A$, $\hat{V}_J$, $\hat{V}_T$ n=100 $\hat{V}_A$, $\hat{V}_J$, $\hat{V}_T$ (Pareto:$\hat{V}_R$!) for negatively skewed parent populations the tableau is similar to that for symmetric populations.

The following figure shows the tendency in the Rel Mean (except for $\hat{V}_R$) with respect to the skewness in the parent population when n is small



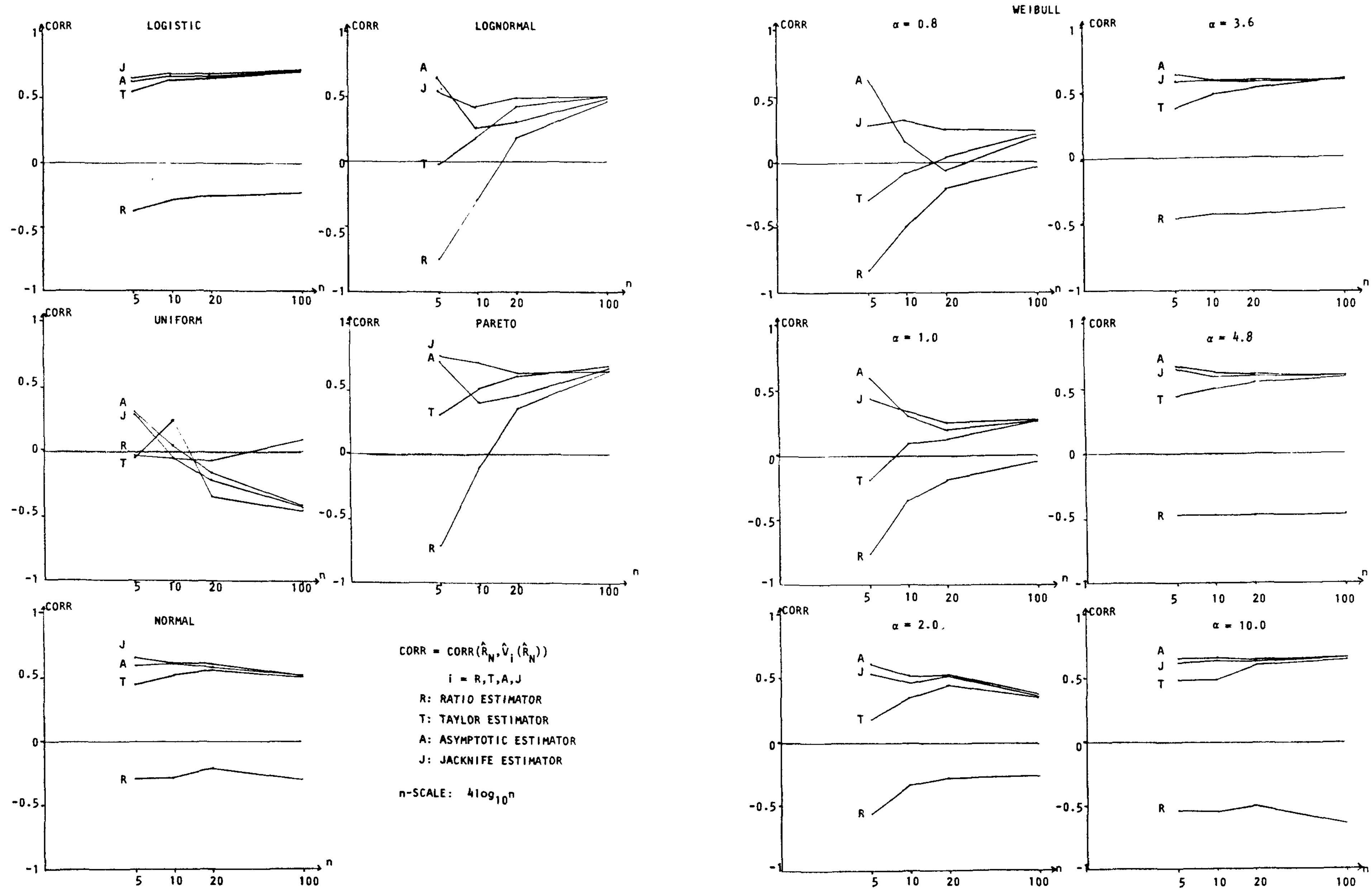
As for the exact sampling distributions in Section 3 we have also studied the correlation ($\hat{\rho}$) between $\hat{G}_N$ and $\hat{Y}_N$ based on the 500 replicates taken from the eleven parent populations. The following was observed: the correlation increases slowly with n and also with the skewness of the parent population. $\hat{\rho}$ is approximately - 0.35 when sampling from Weibull, α = 10.0, zero when the parent populations are symmetrical and 0.40-0.45 when sampling from Weibull α = 2.0. The most positively skewed populations gave rise to the following correlations: Weibull, α = 0.8: 0.90, and Lognormal: 0.95. The Pareto population gave a correlation around 1.

When the parent populations are symmetric or negatively skewed there is a clear tendency in the correlations between variance estimators. In nearly all cases the correlation between $\hat{V}_R$ and $\hat{V}_i$, i=T,A,J, is negative (two positive correlations are observed when the parent population is Logistic). The correlations between the remaining pairs are of size 0.95-1.00 when n = 10 and of size 0.79-0.99 when n=5. The negative correlation observed when $\hat{V}_R$ is involved also shows the defectiveness of the estimator.

The tendency is not as clear as above when the parent population is positively skewed. But when n = 20 and we are sampling from the Lognormal, Pareto and Weibull, α = 0.8, all correlations are of size 0.50-0.97, except $\hat{\rho}_{\hat{V}_R,\hat{V}_A}$ = 0.20 when sampling from Weibull, α = 0.8.

The correlations between $\hat{R}_N$ and $\hat{V}_i$, i=R,T,A,J, respectively, are illustrated in Figure 4.5. When the parent population is symmetric or negatively skewed, with the exception of the Uniform, $\hat{\rho}_{\hat{R}_N,\hat{V}_R}$ is arround -0.25 to -0.50 and the other three coefficients of correlation are gathered just above 0.50. An interesting pattern is found when the parent population is positively skewed. The reader may look at the figures for Lognormal, Pareto and Weibull (α = 0.8, 1.0) parent populations in Figure 4.5. The irregularity found for the Uniform population may perhaps be explained by the fact that the distribution is equally thick.

Figure 4.5 The correlation between $\hat{R}_N$ and the four variance estimators $\hat{V}$, respectively, based on 500 replicates and with sample sizes n = 5, 10, 20, and 100.



References

Cochran, W.G. (1977): Sampling Techniques, 3rd ed., John Wiley & Sons, New York

Hogg, R.V. (1960): Certain Uncorrelated Statistics, JASA, Vol. 62, pp. 265-267

Nygård, F., and Sandström, A. (1985): Estimating Gini and Entropy Inequality Parameters, Memo No. 13, Statistical Research Unit, Statistics Sweden, 1985-01-09

Sandström, A (1983): Estimating Income Inequality, Large Sample Inference in Finite Populations, Dept. of Statist., Univ. of Stockholm, Research Report 1983:5

Tidigare nummer av Promemorior från P/STM:

NR

1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)

2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)

3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)

4 Handledning i AID-analys. Anders Norberg (1980-10-22)

5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)

6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)

7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)

8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)

9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)

10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)

11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)

12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)

13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)

14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)

15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från Elseliv Lindfors, P/STM, SCB, 115 81 Stockholm, eller per telefon 08 7834178

☆Garnisonstryckeriet Stockholm 1985☆