ON ESTIMATION OF POPULATION PROPORTIONS BY DIRECT AND INVERSE SAMPLING

R. P. Chakrabarty and P. J. Tsai
Jackson State University and University of Georgia

Summary

In this paper, we compare the relative merits of four sampling methods, (1) Direct sampling without replacement, (2) Direct sampling with replacement, (3) Inverse sampling without replacement and (4) Inverse sampling with replacement that may be used to estimate population proportions. The stability of the variance estimator in addition to the efficiency of the estimator is taken into consideration in judging the performance of an estimator. Conditions under which one method is preferable to another are obtained. Sample size needed for the use of asymptotic results are indicated by simulation.

1. INTRODUCTION

In socio-economic and ecological surveys, one is frequently interested in estimating either the number NP or proportion P which possesses a specified characteristic in a finite population of size N. The unbiased estimator of the proportions, its variance and the data-base variance estimator for direct sampling with and without replacement are well known (see Cochran 1963). Haldane (1945) first obtained an unbiased estimator of P and its asymptotic variance for the case of inverse sampling with replacement. Recently Best (1974) obtained a closed form for the variance of Haldane's unbiased estimator of the P and compared its efficiency with that of the maximum likelihood estimator given by Feller (1968). Best's numerical evaluation of efficiency showed that the unbiased estimator of P was generally more efficient than the maximum likelihood estimator. Scheaffer (1974) compared the unbiased estimator in direct sampling with replacement and the biased estimator in inverse sampling with replacement, and obtained the asymptotic stabilities of the variance estimators. He has shown that under certain conditions inverse sampling with replacement provides a more stable variance estimator in large samples.

It may be noted that the results of Haldane, Feller, Best and Scheaffer were obtained assuming infinite populations. Some new results for the important case of sampling without replacement from a finite population are obtained in this paper. Only unbiased estimators are considered in view of Bests' results. Special emphasis has been given to obtain results that are exact for any sample size whenever it is possible. We note that the estimate of the variance of an estimator is often used in drawing statistical inferences (e.g., confidence limits for the parameters). It is, therefore, desirable that the variance estimator should be as stable as possible. The importance of investigation of the stability of the variance estimator was pointed out by Chakrabarty and Rao (1967) and Chakrabarty (1973) in estimation of ratios. We, therefore, consider the stability of the variance estimator associated with an estimator in addition to the efficiency of the estimator itself in judging the performance of an estimator. Without loss of generality, we shall discuss the problem of estimation of the proportion P only in a finite population of N units.

Some notations are introduced here that are used in subsequent sections:

- N : number of units in the population.
- N_1 : number of units in the population that possesses a specified attribute or characteristic or falls into specified class.
- n : total sample size.
- n_1 : sample size from the specified class.
- P : $P = N_1 / N$ population proportion to be estimated.
- Q : $Q = 1 - P$.
- p_1 : the unbiased estimator of P by direct sampling with replacement; DWR.
- p_2 : the unbiased estimator of P by direct sampling without replacement; DWTR.
- p_3 : the unbiased estimator of P by inverse sampling with replacement; IWR.
- p_4 : the unbiased estimator of P by inverse sampling without replacement; IWTR.
- q_i : $q_i = 1 - p_i$, $i = 1, 2, 3, 4$.
- f : sampling fraction, $f = \frac{n}{N}$.
- c : finite population correction factor.
- $\bar{c}$: $1 - f = 1 - n/N$.
- $V(p_i)$: the variance of p_i, $i = 1, 2, 3, 4$.
- $v(p_i)$: the estimator of $V(p_i)$, $i = 1, 2, 3, 4$.
- $V_e(v(p_i))$: the exact variance of $v(p_i)$, $i = 1, 2$.
- $V_a(v(p_i))$: asymptotic variance of $v(p_i)$, $i = 1, 2, 3, 4$.

2. DIRECT SAMPLING

2.1 Direct Sampling with Replacement; DWR. The probability density function (p.d.f.) of n_1 is

$$f(n_1) = \binom{n}{n_1} P^{n_1} Q^{n-n_1} \quad ; \quad n_1 = 0, 1, 2, \ldots n$$

$$P + Q = 1 \quad , \quad P, Q \geq 0$$

It is well known that the unbiased estimator of P, its variance and unbiased variance estimator are given by

$$p_1 = n_1/n$$

$$V(p_1) = PQ/n$$

and

$$v(p_1) = p_1 (1-p_1) / (n-1) \qquad (2.1)$$

respectively (see Cochran).

It can be shown that the exact variance of $v(p_1)$ is given by

$$V_e(v(p_1)) = (A_1 + A_2 + A_3 + A_4) / (n^4 (n-1)^2)$$

where

$$A_1 = m_1^{(4)} - 2(n-3)m_1^{(3)}$$

$$A_2 = (n^2 - 6n + 7)m_1^{(2)}$$

$$A_3 = (n-1)^2 m_1^{(1)}$$

$$A_4 = \left\{(n-1)m_1^{(1)} - m_1^{(2)}\right\}^2 .$$

$$m_1^{(k)} = E(n_1^{(k)}) = E\left[n_1(n_1-1)\ldots(n_1-k+1)\right];$$

$$k = 1, 2, 3, 4 \qquad (2.2)$$

The asymptotic variance of $v(p_1)$ by using the well known formula (Kendall and Sturat, 1968)

$$V(g(x)) \doteq g'(E(x))\ V(x) \quad ,$$

is obtained as

$$V_a(v(p_1)) = \left\{v'(p)\right\}^2 V(p_1) \qquad (2.3)$$

$$= (1-2P)^2 PQ/(n(n-1)^2)$$

2.2 Direct Sampling without Replacement; DWTR

The p.d.f. of n_1 is

$$f(n_1) = \frac{\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}{\binom{N}{n}},$$

$$n_1 = 0, 1, \ldots\ldots, \min(n, N_1) \qquad (2.4)$$

The unbiased estimator of P, its variance and unbiased variance estimator are given by:

$$p_2 = n_1/n$$

$$V(p_2) = PQ\,(1-(n-1)/N-1))/n$$

and $v(p_2) = p_2 q_2 (1-n/N)/(n-1)$ (2.5)

respectively. Using the relations between moments about origin and descending factorial moments of n_1 for the p.d.f. (2.4) it can be shown that the exact variance of $v(p_2)$ is given by:

$$V_e(v(p_2)) = C^2 (B_1+B_2+B_3+B_4)/(n^4 (n-1)^2)$$

where

$$B_1 = m_2^{(4)} - 2(n-3)m_2^{(3)}$$

$$B_2 = (n^2 - 6n + 7)\, m_2^{(2)}$$

$$B_3 = (n-1)^2 m_2^{(1)}$$

$$B_4 = -((n-1)m_2^{(1)} - m_2^{(2)}).$$

$$m_2^{(k)} = E(n_1^{(k)}) = n^{(k)} N_1^{(k)} / N^{(k)};$$

$$k = 1, 2, 3, 4 \qquad (2.6)$$

The asymptotic variance of $v(p_2)$ is given by

$$V_a(v(p_2)) = \left(\frac{c}{(n-1)}\right)(1-2P))^2 V(p_2)$$

$$= (1-2P)^2 PQ\,(1-n/N)^2\,(1-(n-1)/$$

$$(N-1))\ /(n(n-1)^2) \qquad (2.7)$$

3. INVERSE SAMPLING

Inverse Sampling is a procedure where sampling is continued until n_1 units possessing the specified characteristic have been obtained. The number n_1 is pre-determined and the Sample Size $n(\geq n_1)$ is thus a random variable.

3.1. Inverse Sampling with Replacement: IWR. The p.d.f of n is

$$f(n) = \binom{n-1}{n_1-1} P^{n_1} Q^{n-n_1} \quad ; \quad n = n_1, n_1+1, \ldots \qquad (3.1)$$

Haldane (1945) first gave the unbiased estimator of P and its variance as

$$P_3 = (n_1-1)/(n-1)$$

and $V(p_3) = \frac{P^2 Q}{n_1}\left(1 + \frac{2!Q}{n_1+1} + \frac{3!Q^2}{(n_1+1)(n_1+2)} + \ldots\right)$ (3.2)

respectively. Note that

$E(n) = n_1 / P$ and $V(n) = n_1 Q / P^2$ (3.3)

consequeltly, the asymptotic variance of p_3 is

$$V(p_3) \doteq V(n)\left[(n_1-1)^2 / (E(n)-1)^4\right]$$
$$= n_1 (n_1-1)^2 P^2 Q / (n_1 - P)^4 \quad (3.4)$$

Now, $E\left[\frac{(n_1-1)(n_1-2)}{(n-1)(n-2)}\right] = \sum_{n=n_1}^{\infty} \frac{(n_1-1)(n_1-2)}{(n-1)(n-2)} \times$

$$\binom{n-1}{n_1-1} P^{n_1} Q^{n-n_1} = P^2$$

$$E\left[\left(\frac{n_1-1}{n-1}\right)^2 - \frac{(n_1-1)(n_1-2)}{(n-1)(n-2)}\right]$$
$$= E(p_3^2) - E^2(p_3) = V(p_3)$$

Thus $v(p_3) = \left(\frac{n_1-1}{n-1}\right)^2 - \frac{(n_1-1)(n_1-2)}{(n-1)(n-2)}$

$$= \frac{(n_1-1)(n-n_1)}{(n-1)^2(n-2)} = \frac{p_3(1-p_3)}{(n-2)} \quad (3.5)$$

is an unbiased estimator of $V(p_3)$. The asymptotic variance of $v(p_3)$ is obtained by using (2.3) as

$$Va(v(p_3)) = \left[n_1 (n_1-1)^2 \left\{(3P-2)n_1^2 + (2-5P)n_1 P + 2P^2\right\}^2 P^4 Q\right] / \left[(n_1-P)^6 (n_1-2P)^4\right] \quad (3.6)$$

3.2 Inverse Sampling without Replacement; IWTR

The p.d.f. of sample size n is

$$f(n) = \frac{\binom{N_1}{n_1-1}\binom{N-N_1}{n-n_1}}{\binom{N}{n-1}} \cdot \frac{N_1-n_1+1}{N-n+1};$$

$$n = n_1, n_1+1, \ldots\ldots N \quad (3.7)$$

The kth ascending factorial moment of n is

$m_4^{[k]} = E(n^{[k]}) = n_1^{[k]} (N_1+1)^{[k]} / (N_1+1)^{[k]}$

Where $n^{[k]} = n(n+1)(n+2)\ldots\ldots(n+k-1)$ (3.8)

consequently, $E(n) = n_1 (N+1) / (N_1+1)$ (3.9)

and $V(n) = \frac{n_1(N+1)(N-N_1)(N_1-n_1+1)}{(N_1+1)^2(N_1+2)}$

or

$$V(n) = \frac{n_1\left(1+\frac{1}{N}\right)\left(P - \frac{n_1-1}{N}\right)Q}{\left(P+\frac{1}{N}\right)^2\left(P+\frac{2}{N}\right)} \quad (3.10)$$

Since

$$E\left(\frac{n_1-1}{n-1}\right) = \sum_{n=1}^{N} \frac{(n_1-1)\binom{N_1}{n-1}\binom{N-N_1}{n-n_1}}{(n-1)\binom{N}{n-1}}.$$

$$\frac{N_1-n_1+1}{N-n+1} = \frac{N_1}{N} = P$$

$$p_4 = \frac{n_1-1}{n-1} \quad (3.11)$$

is an unbiased estimator of P,

The asymptotic variance of p_4 is

$$V(p_4) \doteq \frac{(n_1-1)^2}{\left\{E(n)-1\right\}^4} V(n) \quad (3.12)$$

where E(n) and V(n) are given in (3.9) and (3.10) respectively.

We show that an unbiased estimator of the exact variance of p_4 can be obtained even though an explicit expression for $V(p_4)$ could not be obtained. By definition, we have

$$V(p_4) = E(p_4^2) - P^2.$$

Since

$$\left(\frac{n_1-1}{n-1}\right)^2 - \frac{(n_1-1)(n_1-2)}{(n-1)(n-2)} = \frac{p_4 q_4}{n-2}$$

and $E((n_1-1)(n_1-2)/(n-1)(n-2)) = P(N_1-1)/(N-1)$

We get $E(p_4q_4/(n-2)) = V(p_4) + PQ/(N-1)$

Further, $E(p_4q_4) = PQ - V(p_4)$

Consequently, $v(p_4) = (p_4q_4)(1 - (n-1)/N)/(n-2)$

(3.13)

is an unbiased estimator of $V(p_4)$.

Now, the derivative of $v(p_4)$ with respect to n is

$$\frac{dv(p_4)}{dn} = (n_1-1)\left(\frac{-2n^2 + (2+3n_1)n + 2 - 5n_1}{(n-1)^3(n-2)^2}\cdot\left(1 - \frac{n-1}{N}\right) - \frac{n-n_1}{(n-1)^2(n-2)}\cdot\frac{1}{N}\right)$$

$$= g(n) \quad \text{(say)} \qquad (3.14)$$

Accordingly, the asymptotic variance of $v(p_4)$ is

$$V_a(v(p_4)) \doteq (g(E(n)))^2 V(n)$$

where E(n) and V(n) are given in (3.9) and (3.10) respectively. The above expression is too complicated for analytical comparison but can be easily evaluated numerically.

4. COMPARISON OF DIRECT AND INVERSE SAMPLING

Any comparison of direct sampling with a fixed sample size n to inverse sampling with variable sample size n would be made on the basis of expected sample size, E(n). Note that in

IWR: $E(n) = n_1/P$ or $n_1 = P.E(n)$

IWTR: $E(n) = n_1(1+1/N)/(P+1/N)$

$= n_1/P$ for large N (4.1)

Therefore, n_1 is replaced by P.E(n) in variance formulas in inverse sampling for comparison with variance formulas in direct sampling. The results are given in the sequal.

4.1. Efficiencies of Unbiased Estimators pi

It can be shown that for large n or E(n)

$$V(p_1) = PQ/n \quad , \quad V(p_2) \doteq PQ(1-n/N)/n$$

$$V(p_3) \doteq PQ/E(n) \quad , \quad V(p_4) \doteq PQ(1-E(n)/N)/E(n) \qquad (4.2)$$

Thus p_2 in DWTR is more efficient than p_1 in DWR and p_4 in IWTR is more efficienct than p_3 in IWR. But p_1 and p_3 as well as p_2 and p_4 are equally efficient in large samples.

Sampling without replacement is definitely preferable but one cannot choose between DWTR and IWTR on the basis of the efficiency of unbiased estimators of P.

4.2. Stabilities of Variance Estimators v(pi)

Following Rao and Chakrabarty (1967) and Chakrabarty (1973) the stability of the variance estimator $v(p_i)$ relative to $v(p_j)$ is given by

$(CV(v(p_i)))^2/(CV(v(p_j)))^2$ (4.3)

where $(CV(v(p_i)))^2 = V(v(p_i))/(V(p_i))^2$

$i = 1,2,3,4$ is the relative variance of $v(p_i)$ since $E(v(pi)) = V(p_i)$. The variance estimator $v(p_i)$ is more stable than $v(p_j)$ if

$(CV(v(p_i)))^2 < (CV(v(p_j)))^2$

Now from (2.2), (2.4), (2.6), and (2.7) it can be shown that $Ve(v(p_1)) > Ve(v(p_2))$ and $Va(v(p_1)) > Va(v(p_2))$ for all P.

Substituting n_1 by P.E(n) in variance formulas in inverse sampling and equating E(n) in inverse sampling to n in direct sampling we obtain after some algebraic manipulations the conditions under which one variance estimator is more stable than another. The details which are given in a technical report by the authors (1976) are omitted here and only the results are summarized below:

(1) $v(p_2)$ in DWTR is more stable than $v(p_1)$ in DWR for all P.

(2) $v(p_3)$ in IWR is more stable than $v(p_1)$ in DWR iff $P > 0.6$.

(3) $v(p_4)$ in IWTR is more stable than $v(p_1)$ in DWR iff

$$P > \frac{2-d-(1-d)^{\frac{1}{2}}}{3-2d+2(1-d)^{\frac{1}{2}}}$$

(4) $v(p_3)$ in IWR is more stable than $v(p_2)$ in in DWTR iff

$$\frac{2-(1-d)^{\frac{1}{2}}}{3+2(1-d)^{\frac{1}{2}}} < P < \frac{2+(1-d)^{\frac{1}{2}}}{3-2(1-d)^{\frac{1}{2}}}$$

(5) $v(p_4)$ in IWTR is more stable than $v(p_2)$ in DWTR iff

$$P > \frac{3-2d}{5-4d}$$

(6) $v(p_4)$ in IWTR is more stable than $v(p_3)$ in IWR iff

$$P > \frac{2-d+2(1-d)^{\frac{1}{2}}}{3-2d+3(1-d)^{\frac{1}{2}}}.$$

where $d=n/N$

These results are shown graphically indicating the regions of stable variance estimator. The implication of these results are that if we have prior knowledge about the range of the proportion parameter P, then it is possible to choose the optimum sampling techniques from the four techniques we have discussed, particularly between DWTR and IWTR.

We may mention that we have also evaluated numerically the stabilities of the variance estimators by generating 1000 samples, each of size 100 from a population with N=1000 and P=.8 on cyber 74. DWTR and IWTR was equally efficient but IWTR provided the most stable variance estimator.

We have also investigated the sample size needed for the use of asymptotic resluts by simulation for different values of P. The asymptotic variances were found to be almost equal to exact and/or simulated variances for moderate sample sizes. These results are given in a technical report by the authors (1976).

REFERENCES

Best, D. J., (1974). "The Variance of the Inverse Binomial Estimator," Biometrika, 385-386.

Chakrabarty, R. P. and Rao, J. N. K., (1967). "The Bias and Stability of Jack-Knife Variance Estimator in Ratio Estimation," Proc. Amer. Statist. Assoc., (Social Statistics section), 326-331.

Chakrabarty, R. P., (1973). "A Note on the Small Sample Theory of the Ratio Estimator in Certain Specified Populations," Jour. Indian Soci. Agri. Statist., Vol. 25, No. 2, 49-57.

Cochran, W. G., (1963). Sampling Techniques, Wiley, New York.

Feller, W., (1968). An Introduction to Probability Theory and Its Applications, Vol 1, 3rd Ed., Wiley, New York.

Haldane, J. B. S., (1945). "On a Method of Estimating Frequencies," Biometrika, 33:222-225.

Kendall, M. G. and Stuart, A., (1968). The Advanced Theory of Statistics, Vol. 1, London, Griffin.

Scheaffer, R. L., (1974). "On Direct Versus Inverse Sampling form Mixed Populations", Biometrics, 30:187-198.

Tsai, P. J. and Chakrabarty, R. P., (1976). "Comparison of Direct and Inverse Sampling for Estimation of Population Proportions," Technical Report, Dept. of Statistics and Computer Science, Univ. of Georgia.

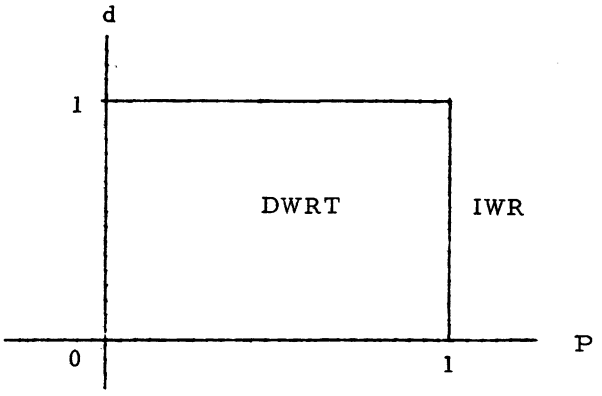

Figure 1: The Region of the Stable Variance Estimators in DWR and DWTR

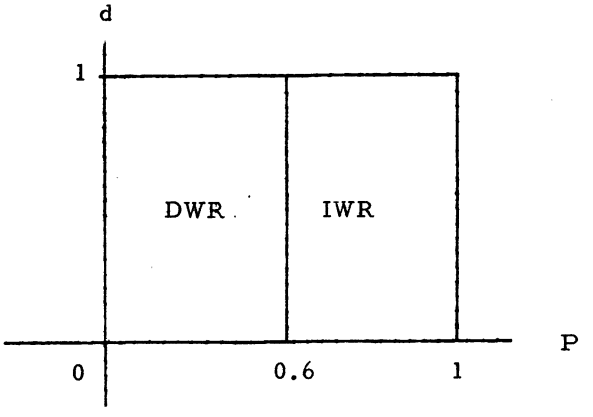

Figure 2: The Region of the Stable Variance Estimators in DWR and IWR

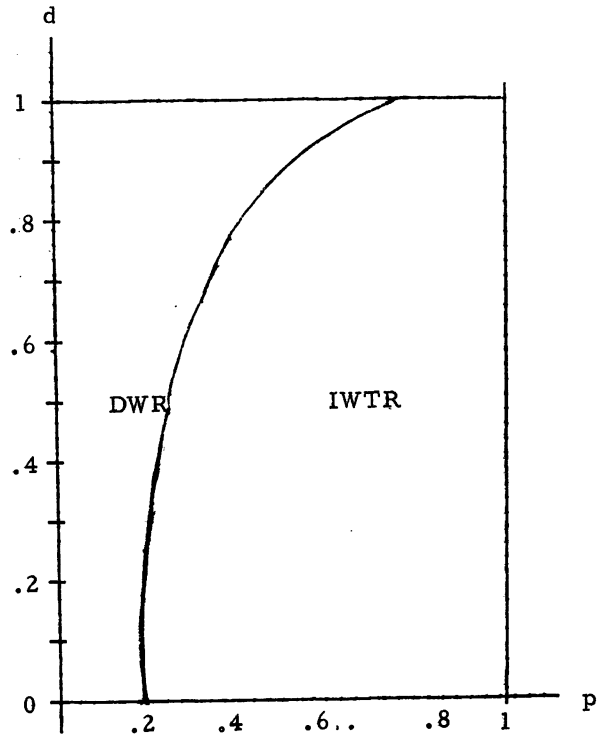

Figure 3: The Region of the Stable Variance Estimator in DWR and IWTR.

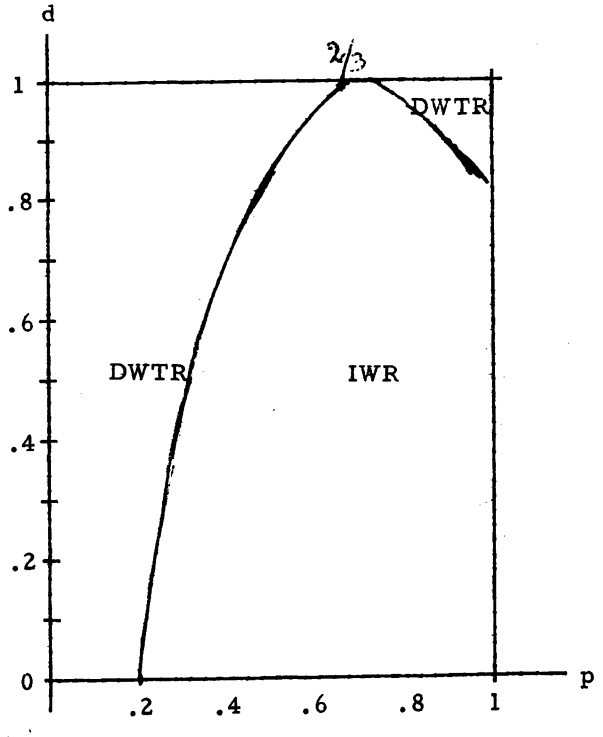

Figure 4: The Region of the Stable Variance Estimator in IWR and DWTR

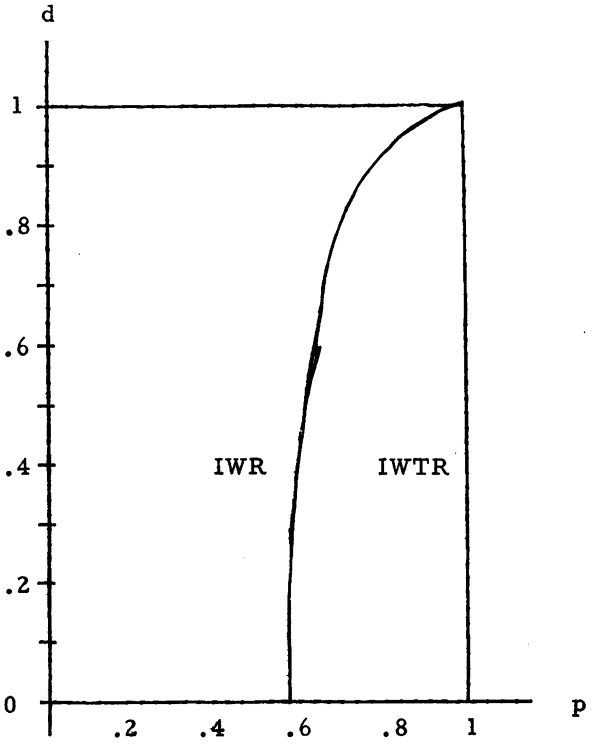

Figure 6: The Region of the Stable Variance Estimator in IWR and IWTR

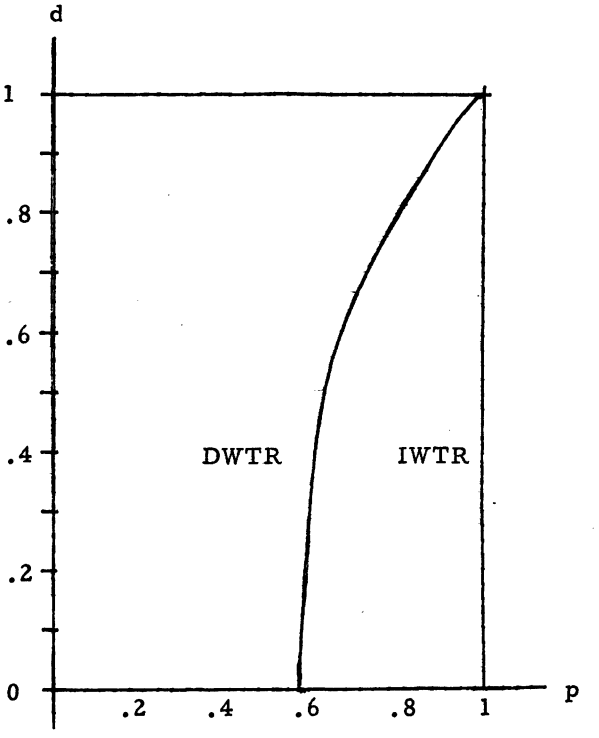

Figure 5: The Region of the Stable Variance Estimator in DWTR and IWTR