OUTER CONFIDENCE INTERVALS FOR QUANTILE INTERVALS OF FINITE POPULATIONS

John S. Meyer, Albion College

1. INTRODUCTION

Assume we have a finite population of N elements, with distinct X values associated with the elements. The ordered population is then denoted as

$$X_{(1)} < X_{(2)} < \dots < X_{(N)}.$$

Let t and u be fixed integers in the range $1 \leq t < u \leq N$. Then $X_{(t)}$ and $X_{(u)}$ are the (t/N)th and (u/N)th quantiles of the population.

A simple random sample of size n is selected without replacement from the population. Denote the order statistics from the sample as

$$x_{(1)} < x_{(2)} < \dots < x_{(n)}.$$

We wish to form a confidence interval for the quantile interval $[X_{(t)}, X_{(u)}]$ of the form $[x_{(r)}, x_{(s)}]$ where $1 \leq r < s \leq n$, and $r \leq t$.

For example, if we have a population of size N = 399, we may wish to find a 95 % or greater confidence interval for the interquartile interval $[X_{(100)}, X_{(300)}]$, that is, the interval in which the middle half of the population values falls. We can show that, based on a simple random sample of size n = 20, that the 2-nd and 19-th order statistics of the sample yields a confidence interval with confidence coefficient of 95.7 %.

This confidence interval is similar to the <u>outer confidence interval</u> for quantile intervals as defined by Wilks (1962, p. 332). However, Wilks assumes a population having a continuous distribution function F. The population quantile ξ_p of order p ($0 < p < 1$) is defined by $\xi_p = \min\{x : F(x) \geq p\}$. For $0 < p_1 < p_2 < 1$, the outer confidence interval for the quantile interval (ξ_{p_1}, ξ_{p_2}) based on the order statistics $x_{(r)}$ and $x_{(s)}$ ($1 \leq r < s \leq n$) is the random interval $[x_{(r)}, x_{(s)}]$.

Krewski (1976) gives a closed form expression for the confidence coefficient of this interval, upper and lower bounds for the coefficent, and numerical examples of each. Reiss and Ruschendorf (1976) give exact formulas and a recurrence relation for the confidence coefficient and also improved bounds for it. They also discuss asymptotic approximations, and numerical examples are given. Sathe and Lingras (1981) give sharper bounds than Krewski and Reiss and Ruschendorf, using properties of convex functions.

Our problem differs from those above, in that we are assuming a finite population, and form a random interval for the (t/N)th and (u/N)th quantile interval. In Section 2 we state the probability of coverage. Section 3 tabulates some examples, and in Table 2, compares the results with those of Krewski. Section 4 states extensions and special cases. The Appendix gives the derivation of the main result.

2. THE CONFIDENCE COEFFICIENT

We now state the formula for the exact confidence coefficient for outer confidence intervals for quantile intervals from finite populations.

<u>Theorem 2.1.</u> A population consists of N elements with distinct X values, ordered as $X_{(1)} < \dots < X_{(N)}$. Let $x_{(1)} < \dots < x_{(n)}$ be a simple random sample of size n drawn from the population without replacement. Assume $1 \leq t < u < N$, $1 \leq r < s \leq n$ and $r \leq t$. Then

$$P[x_{(r)} \leq X_{(t)} < X_{(u)} \leq x_{(s)}]$$

$$= \sum_{r^*=r}^{s} \sum_{s^*=0}^{s-1} \binom{t}{r^*}\binom{u-1-t}{s^*-r^*}\binom{N-u+1}{n-s^*} \Big/ \binom{N}{n}.$$

(We use the convention that $\binom{n}{r} = 0$ if r is a negative integer or $r > n$.)

The proof of this theorem is in the Appendix.

3. NUMERICAL EXAMPLES

A Pascal quadruple precision program was written to compute the confidence coefficient $P[x_{(r)} \leq X_{(t)} < X_{(u)} \leq x_{(s)}]$ for various values of r, s, t, u, N (the population size), and n (the sample size).

Table 1 gives selected values for the confidence coefficient using population sizes of N = 47, 99, 199, 399 and 799, and forming the quantile interval $[X_{((N+1)/4)}, X_{(3(N+1)/4)}]$ (essentially the 25-th and 75-th percentile interval). The tabluations are for a sample size of n = 10, and r and s such that s = n - r + 1 and s = n-r. Due to symmetry, the confidence interval for (r,s) = (r, n - r + 2) has the same confidence coefficient as (r-1, n - r + 1), r = 2, ... , [(n-1)/2].

Table 2 is similar to Table 1, except that the sample size is 20. The additional column labeled K gives the exact confidence coefficient for the continuous case, with n = 20, for $P[x_{(r)} \leq \xi_{.25} < \xi_{.75} \leq x_{(s)}]$ as given in Krewski (1976).

In Table 3 we list the confidence coefficients for samples of size n = 40, from populations of sizes 47, 99, 199, and 399.

Although the tables use (essentially) symmetric intervals for the approximate 25-th and 75-th quantiles, the computer program is written with enough generality to be used for any values of r, s, t, u, n, and N.

Table 1.
$P[x_{(r)} \leq X_{((N+1)/4)} < X_{(3(N+1)/4)} \leq x_{(s)}]$
n=10

r	s	N = 47	N = 99	N = 199	N = 399	N = 799
1	10	.92931	.90830	.89837	.89337	.89086
1	9	.76758	.73424	.71964	.71252	.70901
2	9	.61948	.57827	.56098	.55271	.54866
2	8	.36323	.33809	.32773	.32280	.32039
3	8	.19191	.17869	.17330	.17076	.16952
3	7	.06245	.06176	.06132	.06109	.06096
4	7	.01575	.01678	.01715	.01731	.01739
4	6	.00195	.00244	.00265	.00275	.00280
5	6	.00012	.00018	.00021	.00023	.00023

Table 2.
$P[x_{(r)} \le X_{((N+1)/4)} < X_{(3(N+1)/4)} \le x_{(s)}]_{n=20}$

r	s	N = 47	N = 99	N = 199	N = 399	N = 799	K
1	20	.99933	.99725	.99567	.99472	.99420	.99366
1	19	.99434	.98477	.97900	.97584	.97421	.97253
2	19	.98935	.97230	.96238	.95704	.95429	.95148
2	18	.95869	.92158	.90339	.89420	.88960	.88499
3	18	.92813	.87156	.84558	.83281	.82649	.82023
3	17	.82667	.75149	.72073	.70621	.69916	.69224
4	17	.72926	.63975	.60593	.59045	.58303	.57581
4	16	.53922	.46637	.44032	.42859	.42301	.41760
5	16	.38310	.32743	.30846	.30006	.29608	.29225
5	15	.20500	.18543	.17816	.17483	.17323	.17167
6	15	.10040	.09753	.09601	.09524	.09485	.09445
6	14	.03365	.03931	.04094	.04159	.04188	.04251
7	14	.00979	.01408	.01563	.01632	.01664	.01695
7	13	.00178	.00367	.00452	.00492	.00512	.00531

4. EXTENSIONS AND SPECIAL CASES

4.1 Non-distinct Population Values

It can be shown that the confidence coefficient we stated in Section 2 is a lower bound to the confidence coefficient for the outer confidence interval for quantile intervals if the population values are not distinct.

4.2 Systematic Sampling

If instead of simple random sampling without replacement we do systematic sampling (eg., choose every tenth item in the population), and assume that the population is in random order, Theorem 2.1 still holds.

4.3 Symmetric Confidence Intervals

The formulas for symmetric outer confidence intervals for symmetric quantile intervals are simpler, but depend on whether the population sizes are even or odd.

REFERENCES

KREWSKI, DANIEL (1976), "Distribution-Free Confidence Intervals for Quantile Intervals", Journal of the American Statistical Association, 71, 420-422.

REISS, ROLF D. and RUSCHENDORF, LUDGER (1976), "On Wilks' Distribution-Free Confidence Intervals for Quantile Intervals", Journal of the American Statistical Association, 71, 940-944.

SATHE, Y.S. and LINGRAS, S. R. (1981), "Bounds for the Confidence Coefficient of Outer and Inner Confidence Intervals for Quantile Intervals", Journal of the American Statistical Association, 76, 473-475.

SEDRANSK, J. and MEYER, J. (1978), "Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling", Journal of the Royal Statistical Society B, 40, 239-252.

WILKS, S. S. (1962), Mathematical Statistics, New York: John Wiley.

Table 3.
$P[x_{(r)} \le X_{((N+1)/4)} < X_{(3(N+1)/4)} \le x_{(s)}]_{n=40}$

r	s	N = 47	N = 99	N = 199	N = 399
1	40	1.00000	.99999	.99999	.99999
1	39	1.00000	.99999	.99996	.99992
2	39	1.00000	.99999	.99992	.99984
2	38	1.00000	.99993	.99962	.99930
3	38	1.00000	.99987	.99931	.99875
3	37	1.00000	.99938	.99763	.99617
4	37	1.00000	.99889	.99596	.99359
4	36	1.00000	.99628	.98951	.98479
5	36	1.00000	.99368	.98307	.97600
5	35	.99999	.98357	.96437	.95299
6	35	.99998	.97347	.94572	.93009
6	34	.99946	.94402	.90322	.88232
7	34	.99895	.91472	.86128	.83538
7	33	.99145	.84900	.78452	.75545
8	33	.98396	.78493	.71091	.67936
8	32	.93245	.67335	.60150	.57233
9	32	.88093	.57074	.50275	.47635
9	31	.69950	.43269	.38336	.36448
10	31	.53578	.31978	.28586	.27304
10	30	.26629	.20400	.19045	.18493
11	30	.11947	.12486	.12262	.12136
11	29	.02316	.06401	.06932	.07100
12	29	.00390	.03099	.03741	.03980
12	28	.00000	.01227	.01747	.01965
13	28	.00000	.00452	.00769	.00919
13	27	.00000	.00133	.00289	.00374

APPENDIX

Proof of Theorem 2.1:

$P[x_{(r)} \le X_{(t)} < X_{(u)} \le x_{(s)}]$

$$= P[x_{(r)} \le X_{(t)}) - P(x_{(r)} \le X_{(t)} \cap x_{(s)} \le X_{(u-1)}]. \quad (2.1)$$

Since

$P[x_{(r)} \le X_{(t)} \cap x_{(s)} \le X_{(u-1)}]$

$$= P[x_{(s)} \le X_{(u-1)} | x_{(r)} \le X_{(t)}) \cdot P(x_{(r)} \le X_{(t)}] \quad (2.2)$$

we have, using (2.1) and (2.2),

$P[x_{(r)} \le X_{(t)} < X_{(u)} \le x_{(s)}]$

$$= P[x_{(r)} \le X_{(t)}] \cdot \{1 - P[x_{(s)} \le X_{(u-1)} | x_{(r)} \le X_{(t)}]\}. \quad (2.3)$$

Now, $P[x_{(r)} \le X_{(t)}]$

$$= \sum_{r^*=r}^{\min(t,n)} P[\text{exactly } r^* \text{ are } \le X_{(t)}]$$

$$= \sum_{r^*=r}^{\min(t,n)} \binom{t}{r^*}\binom{N-t}{n-r^*} \Big/ \binom{N}{n} \qquad (2.4)$$

as given in Sedransk and Meyer (1978).

Also,

$$P[x_{(s)} \le X_{(u-1)} | x_{(r)} \le X_{(t)}]$$

$$= \sum_{r^*=r}^{\min(t,s)} P[x_{(s)} \le X_{(u-1)} | \text{exactly } r^* \text{ are } \le X_{(t)}]$$

$$= \sum_{r^*=r}^{\min(t,s)} \sum_{s^*=s}^{\min(n,u-1)} P[\text{exactly } s^* \text{ are } \le X_{(u-1)} | \text{exactly } r^* \text{ are } \le X_{(t)}]$$

$$= \sum_{r^*=r} \sum_{s^*=s}^{u-1} \binom{u-1-t}{s^*-r^*}\binom{N-u+1}{n-s^*} \Big/ \binom{N-t}{n-r^*} \qquad (2.5)$$

Evaluating (2.3) using formulas (2.4) and (2.5), we have

$$P[x_{(r)} \le X_{(t)} < X_{(u)} \le x_{(s)}]$$

$$= \sum_{r^*=r}^{\min(t,s)} \binom{t}{r^*}\binom{N-t}{n-r^*} \Big/ \binom{N}{n}$$

$$\times \left[1 - \sum_{s^*=\max(r^*,s)}^{\min(n,u-1)} \binom{u-1-t}{s^*-r^*}\binom{N-u+1}{n-s^*} \Big/ \binom{N-t}{n-r^*} \right]$$

$$= \sum_{r^*} \binom{t}{r^*}\left[\binom{N-t}{n-r^*} - \sum_{s^*=s}^{n} \binom{u-1-t}{s^*-r^*}\binom{N-u+1}{n-s^*}\right] \Big/ \binom{N}{n}$$

$$= \sum \binom{t}{r^*} \sum_{s^*=0}^{s-1} \binom{u-1-t}{s^*-r^*}\binom{N-u+1}{n-s^*} \Big/ \binom{N}{n}$$

$$= \sum_{r^*=r}^{s} \sum_{s^*=0}^{s-1} \binom{t}{r^*}\binom{u-1-t}{s^*-r^*}\binom{N-u+1}{n-s^*} \Big/ \binom{N}{n} \qquad (2.6)$$