Confidence Intervals for Quantile Estimation from Complex Survey Data

Babubhai V. Shah[1] and Akhil K. Vaish[2]
[1] Safal Institute, Durham, NC 27713 (babu@SAFALinc.com)
[2] RTI International, P.O. Box 12194, Durham, NC 27709 (avaish@rti.org)

Abstract

The basic concept for deriving the confidence limits for quantiles is well known. Woodruff (1952) presented a method for estimating confidence interval for quantiles. The approach consists of three steps: a point estimate of the cumulative distribution function, a confidence interval for the point estimate, and converting it into confidence interval for the quantile. A point estimate for the distribution function for complex survey data is not available in the literature. There are several approaches for deriving binomial confidence intervals. Francisco-Fuller (1991) has suggested another method, different than that by Woodruff, for inverting binomial confidence limits into limits for quantiles. Thus there are too many variations and no consensus about the best practice.

We derive an estimate for complex survey data that is consistent with the one for simple random sample. This paper evaluates several alternatives through extensive simulations to determine the exact specification for the best approach.

Key Words: Quantile estimation, Confidence intervals, Complex survey data.

1. Introduction

The approach for deriving confidence intervals of quantile estimate involves three steps:

1. Compute an estimate of the distribution function $\hat{F}(x)$. The estimate for quantile Q_0 is x_0, where $\hat{F}(x_0) = Q_0$.
2. Obtain confidence interval for the function $\hat{F}$. Since $\hat{F}(x_0)$ is the probability that $x < x_0$, the variable $y = (x < x_0)$ is a binomial variable.
3. Convert confidence interval for $\hat{F}$ into the confidence interval for the estimated quantile.

In section 2, we derive a point estimate of the cumulative distribution function from complex survey data using sample weights. In section 3, we list the several approximations for binomial confidence intervals. Section 4 describes two approaches for inverting confidence interval for the distribution function to derive those of quantiles. Options presented in sections 3 and 4 provide many alternatives. An evaluation strategy to determine optimal method is presented in section 5. The details of the simulations that were performed are in Section 6. Section 7 summarises the results from the simulations. Concluding section 8 recommends an optimal method in practice.

2. A point estimate of the distribution function

Let there be n observations with values $x_1, x_2, \cdots, x_n$ with weights $W_1, W_2, \cdots, W_n$ respectively. Define corresponding normalized weights as $w_i = W_i / \bar{W}$ where $\bar{W} = \frac{1}{n}\sum_{j=1}^{n} W_j$ .We represent the ordered x values and corresponding normalized weights as $x_{[1]}, x_{[2]}, \cdots, x_{[n]}$ and $w_{[1]}, w_{[2]}, \cdots, w_{[n]}$, respectively. Most commonly used estimate of the distribution function $\hat{F}(x)$ is defined as $\hat{F}(x_{[i]}) = S_i = \frac{1}{n}\sum_{k=1}^{i} w_{[k]}$. This estimate $\hat{F}$ implies that the value at $x_{[i]}$ jumps from S_{i-1} to S_i. Consequently, inverse of $\hat{F}$ is not unique at this point. This problem is not addressed in the literature; see Woodruff (1952) and Francisco-Fuller (1991). In general, $F(x)$ is monotonic and the monotonic property is necessary to uniquely determine quantile x_0, such that $\hat{F}(x_0) = Q_0$. In case of a simple random sample, all weights are equal; and the estimator that is monotonic and avoids jumps is given by

$$\hat{F}(x_{[i]}) = \frac{i}{n+1}.$$

To derive an equivalent function for the weighted data, we assume that $\hat{F}$ is of the form: $\hat{F}(x_{[i]}) = a + bw_{[i]} + cS_i$. When the weights are equal, the resulting function must be identical to the one for simple random sample; the resulting conditions are:

$$a + bw_i + cS_i = i/(n+1) \quad \forall i$$

.

We need an additional equation to solve for constants *a* and *b*. The resulting function is invariant under ascending and descending order. Let us assume $G(y)$ is the distribution function for $y = -x$, then

$$F(x_0) = 1 - G(y_0), \text{ if } y_0 = -x_0$$

For the rank ordered values ordering of x and y are reversed, the result is $y_{[i]} = x_{[n+1-i]}$

$$a + bw_{[i]} + cS_i = 1 - (a + bw_{[i]} + cS_i^*)$$

where

$$S^* = \frac{1}{n}\sum_{k=1}^{i} w_{[k]}^* = \frac{1}{n}\sum_{k=n+1-i}^{n} w_{[k]}$$

$S_i^* + S_i = n + w_i$. On solving for *a, b,* and c, we obtain

$$\hat{F}(x_{[i]}) = \frac{1}{(n+1)}\left(S_i + \frac{1}{2} - \frac{w_{[i]}}{2}\right).$$

We propose the above function since it has the desirable properties. This is a new estimator for the weighted data that is equivalent to the estimator $i/(n+1)$, in case of a simple random sample.

3. Binomial confidence intervals

The distribution function at a given point x is the proportion of the observations that have values less than x. For the weighted data, the statistic $\hat{F}(x)$ is a ratio of linear functions of observed variables and its approximate variance based on a survey design can be easily computed. Assuming that the computed estimate of the variance is $\hat{V}[\hat{F}(x)]$, the approximate α% confidence interval for $\hat{F}(x)$, based on normal approximation is given by $\hat{F}(x) \pm \psi_\alpha \sqrt{\hat{V}[\hat{F}(x)]}$ where the α% confidence bound for the standard normal distribution is ψ_α. This is the large sample approximation that works well in most cases but may fail in case of proportion close to zero or one.

One may consider Logistic transformation $(\hat{L} = Logit(\hat{F}) = \ln[\hat{F}/(1-\hat{F})])$ and derive confidence interval assuming that $\hat{L}$ is normally distributed with the large sample approximate variance, $(\hat{V}(\hat{L}) = \hat{V}(\hat{F})/[\hat{F}/(1-\hat{F})]^2)$. Better approximations have been proposed by Korn and Graubard (1998). Let us consider the case of simple random sample of size n. The distribution of $n\hat{F}$ is known to be Binomial, and we can derive exact bounds of $\hat{F}$ without using normal approximation, as follows: let $p_0 = F(X \le x_0) = F(x_0)$ and k be the number of observations in the sample less than or equal to x_0 then $\hat{F}(x_0) = k/n$ and α% upper confidence bound $(\hat{U})$ for p_0 is given by the solution of the following equation: $\sum_{x=k}^{n} \binom{n}{x} p_0^x (1-p_0)^{(n-x)} = \alpha$. The left hand side of the above equation is equal to the incomplete Beta function and can be written as $I_{p_0}(k, n-k+1) = \alpha$. To adjust for continuity, k is replaced by k+p, k+0.5, k+1-p, or k+1. As a result, user is faced with a difficult choice. In simulation results, we shall label these methods as Beta0, BetaP, Beta0.5, Beta1-P, and Beta1, respectively.

For the complex surveys, exact probabilities cannot be computed. For evaluating $\hat{U}$, the Francisco-Fuller method recommends estimating the variance under survey design and then using normal approximation. Korn and Graubard (1998) suggest computing effective sample size n_d under the sample design as $n_d = \frac{\hat{F}(x_0)(1-\hat{F}(x_0))}{\hat{V}_d(\hat{F}(x_0))}$. then deriving appropriate confidence bounds by solving the resulting equation $I_{p_0}(k, n_d - k + 1) = \alpha$. In section 5, we present an evaluation plan that compares all seven methods: normal, logit, and incomplete beta function with five possible values for k.

4. Confidence intervals for quantiles

There are two approaches to convert the confidence interval for the estimated distribution function into the confidence levels for the quantiles. Francisco-Fuller method requires estimation of the three functions $\hat{F}$, $\hat{L}$, and $\hat{U}$. Further more the evaluation functions $\hat{L}$ and $\hat{U}$ needs computation of $\hat{V}[\hat{F}(x)]$ at many points. The point estimate of the quantile is given by the equation $\hat{F}(x_0) = p_0$. The upper confidence level for the quantile x_0 is

given by $\hat{F}(x_L) = p_0$ and the lower confidence level for the quantile x_0 is given by $\hat{F}(x_U) = p_0$. Woodruff (1952) suggested an approach that requires the evaluation of the functions only at one point x_0 where $\hat{F}(x_0) = U_0$. This method was used by Woodruff (1952) for confidence intervals of medians. The approach requires computing confidence interval for $\hat{F}(x_0)$. Assuming these values are L_0 and U_0, the confidence interval for x_0, namely (x_L, x_U) is implicitly defined by the equations: $\hat{F}(x_L) = L_0$ and $\hat{F}(x_U) = U_0$.

5. Evaluation plan

There are two approaches to convert the confidence interval for the estimated distribution function into the confidence levels for the quantiles. We have seven alternatives to drive confidence intervals. As a result, there are fourteen possible methods. Our objective is to find the one that is best for deriving confidence intervals for quantiles. Our strategy has evolved over time and its understanding is necessary to comprehend the simulation results. Each step o evolution was to achieve greater discrimination among the methods. In general two sided intervals are constructed using two one sided intervals. Hence, it is sufficient to evaluate one sided confidence interval $[Q_0, \infty]$ that includes true quantile Q_T at a given confidence level α. In this paper, we evaluate only onesided confidene intervals.

The common approach for evaluating confidence intervals involves selecting few levels for α, such as (5, 10, ... , 95)%, generating a large number of samples, and counting the percentage P_α of samples for which the confidence interval at a given α level includes the true quantile Q_T. This approach fails to provide conclusive evidence to identify the best method.

For a sample, we compute the minimum value for α_T, at which the confidence interval includes true quantile Q_T. For a sample with $\alpha_T \le \alpha_0$, the confidence interval at α_0 will include the true quantile Q_T. For an ideal method, the expected proportion of such samples should be α_0, the distribution of α_T should be uniform [0, 1]. The goodness of fit of the simulated distribution of α_T to the uniform distribution can be checked by Kolmogorov-Smirnov's test or by Cramer-Von Mises test. The method with the smallest value for these statistics is the best. We computed these tests based on 10,000 samples. We repeated groups of samples several times, and took average of Kolmogorov or Cramer statistics, and realized that we needed a different approach. The situation here is similar to the problem of determining the population with the highest mean, among several populations, using only sample data. It is not possible to arrive at the correct decision based on sample means. We may rank populations based on sample means, and repeat process several times. We expect that the population with the highest mean will be ranked first or second most often. We rank methods by Kolmogorov or Cramer statistic, repeat it in various situations and count the number of times a method is ranked first or second. To manage complexity of the task, we evaluate methods using Woodruff

approach and those using Francisco-Fuller approach separately, then compare best of each one against the other.

6. Simulations

All simulations were carried out with the finite population created below:

- Four variables (Stratum, PSU, systolic, and diastolic blood pressure) were extracted from NHANES3 data for the years 1999-2009. After deleting records with missing data, there were 33,672 records.
- Since there were only two PSUs per stratum, each PSU was split up in two or three PSUs, for simulating sampling of PSU.
- Two hypothetical variables Cauchy and Pareto were generated, using parameters that were randomly assigned to each PSU.
- For probability proportion to size (PPS), we generated a random normal variable with mean equal to the size of the PSU, and standard deviation equal to 10% of the mean.
- The population values were computed for each of the nine deciles for each of the four variables: Cauchy, Pareto, systolic, and diastolic blood pressure. These are the "true" deciles for evaluating various methods.

All the samples were generated from the population described above. For each sample we computed confidence level α_T , which is the minimum confidence interval that includes true quantile Q_T . The value of α_T was produced for each of the four variables, for each of the nine deciles, by each of the fourteen methods, resulting in 504 values for α_T , from each sample.

For evaluating the goodness of fit of the sample α_T values to the uniform distribution by Kolmogorov-Smirnov's test, 10,000 samples were generated from the empirical distribution of each of 504 α_T . Kolmogorov statistic was computed for the entire distribution and for each of the two tails [0.0, 0.1] and [0.9, 1.0], resulting in 1512 (504 x 3) Kolmogorov statistics. The entire process was repeated 10 times and averages of 10 results were produced for each of the 1512 values. The goodness of fit statistics as measured by average Kolmogorov test was produced for 7 methods with woodruff's approach and 7 methods with Francisco-Fuller approach in 96 instances formed by 4 variables, 8 deciles and 3 tests. Methods were ranked with in both groups separately from 1 to 7, to determine the best method using woodruff's approach and the best one by Francisco-Fuller's approach. The entire simulations were carried out for three different sample designs: Simple Random Sample (SRS), Equal Probability Sample (EPS), and Probability proportional to Size (PPS). We also considered three different sample sizes: 300, 600, and 1200. The simulation results for all nine sample designs are presented in the next section. Three methods using incomplete beta function were identical for medians and hence median as not included. The summary tables include only eight deciles.

7. Summary of results

We computed an average of 10 Kolmogorov tests, with each test based on 10,000 α_T values. Table 1 presents the results for the first decile of the variable "Cauchy" for the 9 sample designs by the 14 methods. It can be seen from Table 1 that the values are very similar and discrimination among methods is not easy.

Table 1. Average of 10 Kolmogorov tests for the first decile of the variable "Cauchy" by sample design and size by the 14 methods

	Probability Proportional to Size			Equal Probability Sample			Simple Random Sample		
	300	600	1200	300	600	1200	300	600	1200
Method	Using Woodruff's Approach								
Beta0	0.0379	0.0355	0.0356	0.0403	0.0327	0.0305	0.0368	0.0358	0.035
BetaP	0.0272	0.0302	0.032	0.0278	0.0284	0.028	0.0296	0.0297	0.033
Beta0.5	0.0272	0.0302	0.032	0.0278	0.0284	0.028	0.0296	0.0297	0.033
Beta1+P	0.0272	0.0302	0.032	0.0278	0.0284	0.028	0.0296	0.0297	0.033
Beta1	0.0372	0.0369	0.0365	0.0382	0.0357	0.032	0.0408	0.0364	0.0397
Logit	0.0272	0.0303	0.0321	0.0278	0.0285	0.0281	0.0297	0.0298	0.0331
Normal	0.0272	0.0303	0.032	0.0278	0.0285	0.0281	0.0297	0.0297	0.0331
	Using Francisco-Fuller's Approach								
Beta0	0.0372	0.0367	0.036	0.0382	0.0357	0.032	0.0409	0.036	0.0389
BetaP	0.0271	0.0299	0.0319	0.0278	0.0284	0.028	0.0293	0.0296	0.033
Beta0.5	0.0273	0.0301	0.032	0.0278	0.0285	0.0281	0.0295	0.0298	0.0332
Beta1+P	0.0275	0.0303	0.0322	0.0279	0.0286	0.0281	0.0297	0.03	0.0334
Beta1	0.038	0.0356	0.0365	0.0403	0.0327	0.0305	0.0371	0.0365	0.036
Logit	0.0272	0.03	0.032	0.0278	0.0285	0.028	0.0294	0.0297	0.0331
Normal	0.0273	0.0301	0.032	0.0278	0.0285	0.0281	0.0295	0.0298	0.0332

For each combination of variable (4), decile (8), and tail (3), we ranked the methods form 1to 7 based on Kolmogorov test. The rankings were done separately for Woodruff's and Francisco-fuller approaches. The number of times a method was ranked first or second is presented in Table 2.

Table 2. The number of instances the method was ranked first or second by sample design and size

		Probability Proportional to size			Equal Probability Sample			Simple Random Sample		
	Total	300	600	1200	300	600	1200	300	600	1200
Method	Using Woodruff's Approach									
Beta0	144	4	13	27	3	18	33	9	10	27
BetaP	355	42	48	28	48	40	23	50	44	32
Beta0.5	590	90	71	48	83	61	37	83	68	49
Beta1+P	264	25	28	36	22	35	43	21	22	32
Beta1	140	5	9	30	2	13	32	4	17	28
Logit	190	14	18	26	17	18	32	17	19	29
Normal	261	36	29	21	41	31	16	32	36	19
	Using Francisco-Fuller's Approach									
Beta0	210	20	20	29	21	24	33	16	19	28
BetaP	656	90	80	62	90	69	39	87	79	60
Beta0.5	215	23	25	22	27	29	15	30	24	20
Beta1+P	114	7	9	20	6	13	21	6	12	20
Beta1	238	19	27	30	20	26	37	22	25	32
Logit	357	46	40	37	37	40	48	35	36	38
Normal	154	11	15	16	15	15	23	20	21	18

The preponderance of evidence from these counts, suggests that the best method for Woodruff's approach is Beta0.5 and that for Fuller-Francisco approach is BetaP, The replication of the entire simulation confirmed this conclusion (see Table 3).

Table 3. The number of instances the method was ranked first or second by sample design and size (Second replication)

		Probability Proportional to Size			Equal Probability Sample			Simple Random Sample		
	Total	300	600	1200	300	600	1200	300	600	1200
Method	Using Woodruff's Approach									
Beta0	104	3	9	20	3	4	30	5	9	21
BetaP	363	43	45	31	46	44	21	50	50	33
Beta0.5	626	92	76	50	93	69	38	87	73	48
Beta1+P	204	19	18	30	11	28	48	14	12	24
Beta1	93	3	7	23	0	1	28	0	8	23
Logit	186	11	16	25	10	23	27	15	26	33
Normal	368	45	45	37	53	47	24	45	38	34
	Using Francisco-Fuller's Approach									
Beta0	166	16	16	23	16	19	27	14	14	21
BetaP	678	91	85	58	99	71	44	93	77	60
Beta0.5	199	17	25	21	24	23	17	21	30	21
Beta1+P	82	4	6	15	4	8	16	5	9	15
Beta1	202	17	21	30	16	21	31	19	18	29
Logit	449	60	49	45	48	52	57	50	44	44
Normal	168	11	14	24	9	22	24	14	24	26

The comparison of the best methods for the two approaches is presented in Table 4. The theoretical specification for Francisco-Fuller approach assumes that binomial confidence interval is computed at every point. In general this is not practical and one evaluates

binomial confidence interval at the few points only and linear interpolation is applied for all other points. To evaluate this practical aspect, we computed binomial confidence interval at 99 points, corresponding to $\hat{F}(x) = (0.01, 0.02, ..., 0.99)$. The ideal application of Francisco-Fuller approach is superior to Woodruff's approach. However, the practical implementation results in a performance worse than that of Woodruff's approach.

Table 4. Comparing Woodruff's approach with both ideal and interpolated Francisco-Fuller's approach

			FF ideal		FF interpolated	
Design	Sample Size	Total Instances Compared	Instances FF (ideal) better	Average Ratio WF / FF	Instances FF (inter) better	Average Ratio WF / FF
EPS	300	108	85	1.21	30	0.95
EPS	600	108	84	1.12	1	0.60
EPS	1200	108	83	1.08	0	0.42
PPS	300	108	78	1.16	25	0.92
PPS	600	108	83	1.12	2	0.61
PPS	1200	108	76	1.07	0	0.39
SRS	300	108	86	1.17	5	0.78
SRS	600	108	76	1.09	0	0.50
SRS	1200	108	72	1.04	0	0.35
Results from the second replication						
EPS	300	108	89	1.21	34	0.94
EPS	600	108	87	1.13	1	0.60
EPS	1200	108	79	1.08	0	0.42
PPS	300	108	80	1.16	28	0.93
PPS	600	108	84	1.13	3	0.62
PPS	1200	108	80	1.07	0	0.40
SRS	300	108	88	1.17	7	0.78
SRS	600	108	72	1.08	0	0.49
SRS	1200	108	68	1.03	0	0.35

8. Conclusion

The best practical approach to derive the confidence interval for the quantiles is as follows:

- Compute point estimate of the distribution function as described in section 2.
- Obtain binomial confidence interval for $\hat{F}$ using incomplete beta function.
- Convert interval for $\hat{F}$ into interval for quantiles using Woodruff's approach.

References

Carol A. Francisco, Wayne A. Fuller (1991), Quantile Estimation with a Complex Survey Design, Annals of Statistics, 19, 454-469

Korn, E.L., and B.I. Graubard. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. Survey Methodology 24:193-201.

Philip J. McCarthy (1965). Stratified Sampling and Distribution-Free Confidence Intervals for a Median. Journal of the American Statistical Association, 60, 772-783

Research Triangle Institute (2004) SUDAAN User's Manual. Research Triangle Institute, Research Triangle Park, NC.

Woodruff R. S. (1952). Confidence intervals for medians and other positional measures. Journal of the American Statistical Association, 47, 635-646.