Confidence Intervals for Quantile Estimation from Complex Survey Data

Babubhai V. Shah[1] and Akhil K. Vaish[2]
Safal Institute, Durham, NC 27713, babushah@earthlink.net[1]
RTI International, P.O. Box 12194, Durham, NC 27709, avaish@rti.org[2]

Abstract

A method for estimating quantiles and their confidence intervals based on the paper by Francisco-Fuller (1991) has been implemented in SUDAAN® software. There have been problems encountered in practical application of the Francisco-Fuller method: It requires evaluation of bounds at many points, the resulting limits are, at times, not monotonic or internally consistent, and the accuracy is not very good. The objective of this paper is to develop an improved method that overcomes these problems. An adjustment to the empirical distribution is proposed to achieve internal consistency and reduce bias.

Keywords: Quantile estimation, Confidence intervals, Complex survey data.

1. Introduction

A method for estimating quantiles and their variances based on the paper by Francisco-Fuller [1991] has been implemented in SUDAAN® [2004] software. There have been problems encountered in practical application of the Francisco-Fuller method. The specific problems reported are:

a. Too often, estimates of the quantile or its variance or confidence limits are missing.
b. Estimated variance is smaller than expected.
c. Francisco-Fuller (1991) method requires evaluation of bounds for $\hat{F}(x)$ at many points.

The objective of this paper is to present a simpler method that avoids the above problems and performs as well as the Francisco-Fuller method. A point estimate $\hat{X}_p$ of p is defined implicitly, by the equation: $\hat{F}(\hat{X}_p) = p.$ Consequently, it is not possible to explicitly derive confidence levels for $\hat{X}_p$. The common approach is to:

1. Compute an estimate of the distribution function $\hat{F}$ to obtain a point estimate of the quantile.
2. Obtain confidence intervals for the function $\hat{F}$. Since $\hat{F}$ at a given point x_0 is the proportion of the binomial variable $X < x_0$. This step is equivalent to deriving confidence interval for binomial parameter from a complex survey data.
3. Convert confidence intervals for $\hat{F}$ into the confidence interval for the estimated quantile for $\hat{X}_p$.

An approach using the above three steps was first used by Woodruff (1952) for obtaining confidence interval for medians. Francisco-Fuller (1991) presented a similar three-step approach for confidence interval for quantiles but used a different method for Step 3. If any of the three steps of a method is changed, different set of confidence intervals is obtained. In this paper, we combine "best" method for each step to produce a new method for estimating confidence intervals for a quantile.

First we propose a new estimator for the weighted data that is equivalent to a well-known estimator for a simple random sample. Second we suggest either Logistic or incomplete Beta function for confidence intervals for the function $\hat{F}$. Korn and Graubard (1998) present methods for confidence intervals for the proportions. Lastly, for the step 3, we apply Woodruff's approach for converting confidence intervals for $\hat{F}$ into the confidence interval for the estimated quantile for $\hat{X}_p$.

2. Estimate of the Distribution Function

Let there be n observation 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 $\hat{F}(x_{[i]}) = S_i = \frac{1}{n}\sum_{k=1}^{i} w_{[k]}$.

Woodruff (1952) and Francisco-Fuller (1991) both used the above estimator. The above estimator implies that the distribution has no observable values greater than the maximum x_i , which is not true. Secondly, if the data were sorted in a descending order the resulting estimate will be inconsistent with the estimate obtained by using data sorted in ascending order.

In case of a simple random sample, all the weights are equal and the equivalent empirical distribution reduces to $\hat{F}(x_{[i]}) = \frac{i}{n}$. The most common adjustment to avoid the two anomalies is to use the estimator:

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

To derive an equivalent function for the weighted data, we assume that $\hat{F}$ is of the form: $\hat{F}(x_{[i]}) = a + bw_{[i]} + cS_i$. To solve for the unknown constants, we impose two conditions: when the weights are equal, the resulting function will be identical to the one for simple random sample; and that the resulting function is invariant under ascending and descending order. 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. Confidence Intervals for the Distribution Function

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 $\alpha\%$ 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 ψ_α is the $\alpha\%$ confidence bound for the standard normal distribution. This is the large sample approximation that works well in most cases but fails incase of proportions close to zero or one.

The problem with the normal approximation is that the upper bound may be greater than 1 or the lower bound may be less than 0. An improvement that avoids this problem is to use logistic transformation $\hat{L} = \ln[\hat{F}/(1-\hat{F})]$. Large sample approximate variance of $\hat{L}$ is $\hat{V}(\hat{L}) = \hat{V}(\hat{F})/[\hat{F}/(1-\hat{F})]^2$. An upper bound for $\hat{L}$ is $\hat{L}_U = \hat{L} + \psi_\alpha \sqrt{\hat{V}(\hat{L})}$ then the upper bound for $\hat{F}$ is $\hat{F}_U = \exp(\hat{L}_U)/[1+\exp(\hat{L}_U)]$. Similarly, a lower bound on $\hat{F}$ could be also be obtained.

Korn and Graubard (1998) have suggested a better approximation for complex data that is analogous to the exact method in case of a simple random sample. Let us assume that in a simple random sample of size n , there are k values which are less than a given value x_0 . Then $\hat{F}(x_0) = k/n$ and the $\alpha\%$ upper bound is obtained by solving for p_U in the equation: $\sum_{i=k}^{n} \binom{n}{i} p_U^i (1-p_U)^{(n-i)} = \alpha.$ The left hand side of the above equation is equal to the incomplete Beta function and can be written as $I_{p_U}(k, n-k+1) = \alpha.$

For applying the incomplete Beta approximation for the upper bound of $\hat{F}$, the effective sample size under the design is:

$$n_d = \frac{\hat{F}(x_0)(1-\hat{F}(x_0))}{\hat{V}_d(\hat{F}(x_0))}.$$

The upper bound for $\hat{F}(x_0)$ is p_U , where p_U is the solution of the equation $I_{p_U}(k_d, n_d - k_d + 1) = \alpha$ where $k_d = n_d \hat{F}(x_0)$.

4. Confidence Intervals for Quantile

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 $\hat{F}(x_U) = p_0$. We suggest 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 $\hat{F}(x_U) = U_0$.

In the next section, we evaluate the proposed approach that uses:

1. The new estimator of $\hat{F}$ derived in Section 2.
2. The confidence interval for $\hat{F}$ using the incomplete Beta function or Logistic approximation.
3. The conversion of the confidence interval for $\hat{F}$ into the confidence interval for the quantile using the method similar to the one applied by Woodruff (1952).

5. Simulation Results

The simulation was carried out using a stratified clustered sample with unequal probabilities from a finite population. For creating the large population, we selected the data from the 2004 Behavioral Risk Factor Surveillance System (BRFSS). We selected two variables State and Body Mass Index (BMI) from the data set. The deletion of observation with missing values yielded 289444 records from 52 States. We used States as strata and arbitrarily created Primary Sampling Units (PSUs or clusters) within each State. We also generated a new dummy variable with high intra class correlation of 0.1. The deciles for both the variables for the population were the "True values" for our simulation.

We generated 10,000 samples, with unequal probabilities of selection for PSU's and equal probability of selection for each record within a PSU. For each of the sample, all the nine deciles were computed. For each of the deciles, we obtained one sided confidence interval at 10, 20, 30, 40, 50, 60, 70, 80, and 90 percent confidence levels. We then counted the number of samples for which the corresponding "true" or population decile fell below the confidence limit for that quantile. We would expect 1000 samples for 10% confidence limit, 2000 samples for 20% confidence limit, and so on. The results of the simulation study are presented in Table I for the first decile and in Table II for the median. The results for other deciles are similar, and not presented here to conserve space. They are available from the authors on request.

We also obtained 95% two tail confidence limits by Francisco-Fuller method using SUDAAN® software. We also computed 2.5% lower and 97.5% upper limits by the proposed method with Beta approximation. For each of the deciles, we counted the number samples in which the "true" decile was below the lower tail, as well as those samples in which the "true" decile was above the upper tail. These sample counts corresponding to the confidence limits for Body Mass Index are shown in Table III, and those for the Dummy Variable are in shown in Table IV.

Table I. Number of Samples below the Confidence Level of the 1st Decile

Alpha	Body Mass Index		Dummy Variable	
	Logit	Beta	Logit	Beta
10%	830	947	837	952
20%	1769	1928	1864	2017
30%	2778	2936	2901	3082
40%	3810	3935	3925	4074
50%	4768	4970	4951	5105
60%	5798	5935	5963	6123
70%	6760	6892	6950	7091
80%	7683	7813	7870	8009
90%	8710	8815	8805	8897

Table II. Number of Samples Below the Confidence Level of the Median

Alpha	Body Mass Index		Dummy Variable	
	Logit	Beta	Logit	Beta
10%	973	973	1113	1114
20%	1914	1915	2131	2131
30%	2859	2859	3163	3163
40%	3817	3817	4148	4148
50%	4780	4780	5163	5163
60%	5794	5794	6212	6211
70%	6753	6752	7211	7211
80%	7795	7795	8213	8212
90%	8827	8826	9101	9101

6. Conclusion

Based on the limited simulation study the proposed method performs well under a wide variety of circumstances. For a variable such a BMI, where the data are highly clumped at several points, the Francisco-Fuller method's performance is erratic. The Francisco-Fuller method requires estimation of complete functions for upper and lower confidence levels, and that results in evaluation of variance at many points for $\hat{F}$ and also more interpolation for other points. Consequently, the Francisco-Fuller method's performance is erratic for BMI. Both methods perform well for the dummy variable with a smoother distribution function generated through pseudo-random numbers.

Table III. Beta vs. Francisco-Fuller (FF) Method (95% Confidence levels for BMI): Number of Samples with True Value in Tails

Decile	Beta (Proposed)		FF (SUDAAN)	
	Lower	Upper	Lower	Upper
0.1	211	235	218	198
0.2	244	210	509	117
0.3	243	240	607	155
0.4	260	297	268	445
0.5	282	339	2278	72
0.6	326	282	214	788
0.7	259	294	151	307
0.8	203	284	236	236
0.9	234	283	384	129

Table IV. Beta vs. Francisco-Fuller (FF) Method (95% Confidence Levels for Dummy Variable) Number of Samples with True Value in Tails

Decile	Beta (Proposed)		FF (SUDAAN)	
	Lower	Upper	Lower	Upper
0.1	398	364	249	135
0.2	401	384	364	128
0.3	407	316	293	170
0.4	374	271	276	239
0.5	349	235	204	324
0.6	306	289	226	270
0.7	341	272	216	295
0.8	446	301	206	274
0.9	445	290	186	255

The proposed method is much simpler to implement, because it requires estimation of confidence limits for $\hat{F}$ at a single point. Lastly, it uses the new estimator for the distribution function F which has the desirable properties for the weighted data.

7. References

Francisco, Carol A., 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.

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.