## Appendix D: Confidence Interval Estimation for Percentiles

A large body of literature describes various methods to estimate percentiles and to derive the variance and confidence intervals for complex survey data. Highlighted in the literature are the following methods: Woodruff method (Woodruff 1952), "test inversion" method (Francisco and Fuller 1991), the Normal transformation method (Korn and Graubard 1999), and Replication methods (Kovar 1988; Rogers 2003).

Confidence intervals for percentiles in this report were calculated with the Woodruff method. This method uses the standard error of the empirical distribution function at the selected percentile and constructs a 95% confidence interval, followed by back transformation using the inverse of the empirical distribution. The previous National Report on Biochemical Indicators of Diet and Nutrition in the U.S. Population, 1999–2002 used a variation of the Woodruff method by combining it with the method of Clopper and Pearson proposed by Korn and Graubard (1999) for complex surveys. This approach was used previously because large-sample normal approximations used to calculate confidence intervals for proportions close to zero or 1 can lead to confidence intervals with poor coverage properties. However, a paper by Sitter and Wu (2001) concluded that despite the fact that confidence intervals around the empirical distribution function at tail regions perform poorly, the Woodruff confidence intervals obtained by inverting these poorly behaved intervals perform very well for percentiles. Therefore, the confidence intervals presented in this report are based on the Woodruff approach with no further modifications, as described in the steps below.

### Background

Define an arbitrary percentile $\mathbf{X_p}$, such that $\mathbf{F(X) = P(X \leq X_p) = p}$. This is pictured in Figure 1, where the y-axis displays the empirical distribution function (cdf) over a set of hypothetical values. In this example, p = 0.5 and so $\mathbf{X_{0.5}}$ is the median. Both SAS (version 9.2) and SUDAAN (version 10.0) find $\mathbf{X_p}$ through linear interpolation. Let $\hat{F}(x_j)$ be an estimate of the empirical distribution function at x and assume data $x_1, x_2, \ldots x_n$ are a rank ordered listing of the sampled values, such that $x_1$ is the minimum value and $x_n$ is the maximum value; then the estimated percentile by use of linear interpolation is calculated as:

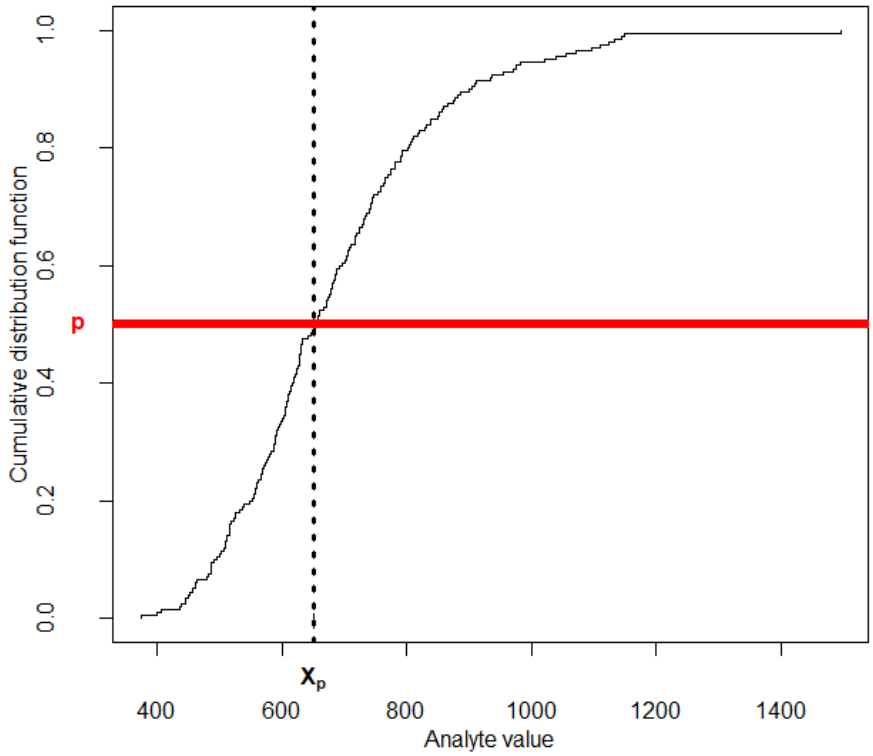
$$\hat{X}_p = x_j + \frac{p - \hat{F}(x_j)}{\hat{F}(x_{j+1}) - \hat{F}(x_j)}\left(x_{j+1} - x_j\right) \qquad \hat{F}(x_j) \leq p < \hat{F}(x_{j+1}).$$

*Description: Equation for estimating a percentile from an empirical distribution using linear interpolation. The estimated pth percentile equals the lower ranked value (x sub j) plus a weighted fraction of the distance between consecutive ranked values (x sub j and x sub j plus 1). The weight is*

*determined by where the target percentile p falls between the empirical cumulative distribution values F hat(x sub j) and F hat(x sub j plus 1). The formula applies when F hat(x sub j) ≤ p < F hat(x sub j plus 1).*

To find the percentiles and confidence intervals in this report, we used results derived from SUDAAN's PROC DESCRIPT (DESIGN=WR) PERCENTILE statement and results from the Histogram output group.

**Figure 1:** Definition of a percentile




*Description: Empirical cumulative distribution function (ECDF) of analyte values. The x-axis represents analyte value and the y-axis represents the cumulative distribution function, ranging from 0 to 1. The black stepwise curve shows the cumulative proportion of observations at or below each analyte value. A red horizontal line marks the cumulative probability level p (approximately 0.5). The point where this line intersects the ECDF identifies the percentile estimate, X sub p. A black vertical dashed line extends through X sub p to indicate the analyte value at which p percent of observations fall at or below that value. The figure illustrates the graphical determination of percentile X sub p from an empirical cumulative distribution.*

**Step 1**

Use SUDAAN (DESIGN=WR) to estimate the percentiles, the empirical distribution, and the standard error of the empirical distribution function at each point. SUDAAN uses a Horowitz-Thompson estimator of the empirical distribution function at each value. The estimated empirical distribution function can be outputted into a dataset using the HISTPCT statement in conjunction with an OUTPUT statement in PROC DESCRIPT. By default, SUDAAN estimates the empirical distribution function by using a maximum of 100 equally spaced percentages to divide the population into bins. We used the option /NPCT in the HISTPCT statement to change this default to allow a jump in the empirical distribution function at every unique data value, up to 2950 unique values. An output file is generated by SUDAAN to contain: the upper endpoint of the current bin in the histogram, the cumulative percent less than or equal to the upper endpoint of the current bin and the respective estimated standard error of the cumulative percent. This file is used to obtain 95% confidence intervals of the empirical distribution function at the selected percentile. In some rare cases, using the values of the upper endpoint available in this file may differ slightly from a rank order list of the weighted sampled values if there are more than 2950 distinct values. This difference may lead to very small differences when comparing the confidence limits to other software which uses Woodruff confidence limits based on the weighted sample values.

Sample SUDAAN code for serum folate (FOL) is as follows:

```
PROC DESCRIPT DATA=NHANES03_06 FILETYPE=SAS DESIGN=WR;
NEST SDMVSTRA SDMVPSU/MISSUNIT;
WEIGHT WTMEC4YR;
   VAR LBXFOL;
PERCENTILES 5 10 90 95 /MEDIAN ;
HISTPCT /NPCT=2950;
OUTPUT QTILE /FILENAME=PCTILES
        FILETYPE=SAS REPLACE ;
OUTPUT UPPEREND CUMPCT SECUMPCT   /FILENAME=HIST FILETYPE=SAS
        REPLACE;
```

If you change the first OUTPUT statement in the above program to

OUTPUT QTILE LOWQTILE UPQTILE /FILENAME=PCTILES FILETYPE=SAS REPLACE ;

SUDAAN will provide the upper and lower confidence limits based on the "test inversion" method of Francisco and Fuller. However, as mentioned earlier, this report does not use SUDAAN's default method to generate confidence intervals for percentiles.

**Step 2**

Using SAS DATA steps to manipulate the output files from Step 1, find the value of the estimated cumulative distribution function that is less than or equal to **p** using the values of the cumulative percent produced by SUDAAN in the output file HIST:

$$\hat{F}(x_j) \leq p < \hat{F}(x_{j+1})$$

Save $\hat{F}(x_j)$ and the corresponding standard error (SE) estimate at $\hat{F}(x_j)$ and proceed to step 3.

**Step 3**

Using **p (the desired percentile)** and the standard error of the estimate of $\hat{F}(x_j)$, compute the 95% confidence interval for p: ($p_L$, $p_U$) as $p \pm t_{0.025,DF} SE$, where the degrees of freedom (DF) are the number of primary sampling units minus the number of strata and SE is the standard error from step 2. To get the appropriate degrees of freedom for each subgroup use the ATLEVEL1 and ATLEVEL2 options in SUDAAN's PROC DESCRIPT to count up the number of strata and PSUs with valid data. This must be done in a separate call to PROC DESCRIPT than the one that calculates the percentiles because the HISTPCT statement is not available with ATLEVEL. Note: SAS (version 9.2) uses the empirical point estimate at the desired percentile, $\hat{p} = \hat{F}(x_j)$, in order to calculate the 95% confidence interval as $\hat{p} \pm t_{0.025,DF} SE$.

**Step 4**

Let $\hat{L}_p$ be the lower confidence limit of the estimated percentile and $\hat{U}_p$ be the upper confidence limit of the estimated percentile. $x_1, x_2, \ldots x_n$ are a rank ordered listing of the values as produced by SUDAAN using HISTPCT, such that $x_1$ is the minimum value and $x_n$ is the maximum value; then these can be found from the following expressions:

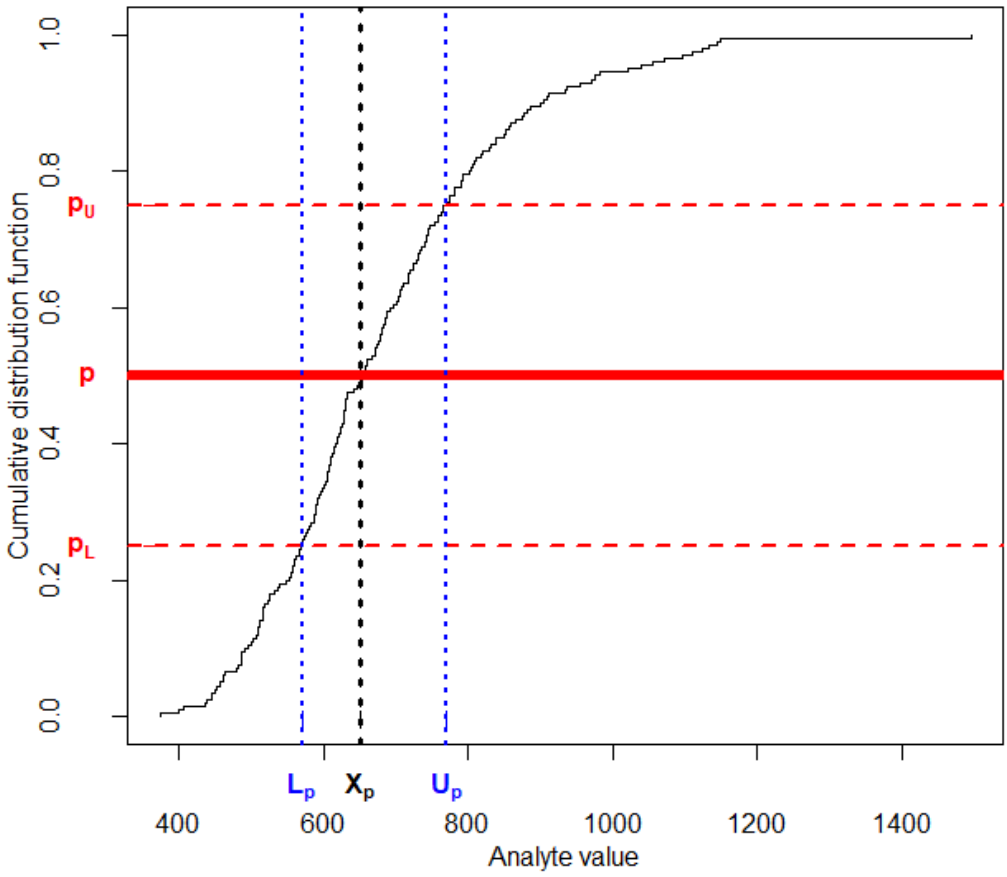
$$\hat{L}_p = \begin{cases} x_1 & p_L < \hat{F}(x_1) \\ x_j + \dfrac{p_L - \hat{F}(x_j)}{\hat{F}(x_{j+1}) - \hat{F}(x_j)}\left(x_{j+1} - x_j\right) & \hat{F}(x_j) \le p_L < \hat{F}(x_{j+1}) \\ x_n & p_L = 1 \end{cases}$$

*Description: The formula estimates the lower confidence bound of a percentile from an empirical distribution using linear interpolation. The estimated lower confidence bound, L sub p, equals the lower ranked value (x sub j) plus a weighted fraction of the distance between consecutive ranked values (x sub j and x sub j plus 1). The weight is determined by where the lower confidence probability, p sub L, falls between the empirical cumulative distribution values F hat (x sub j) and F hat( x sub j plus 1). The formula applies when F hat( x sub j) is less than or equal to p sub L, and p sub L is less than F hat( x sub j plus 1). If p sub L is less than F hat(x sub 1), then the estimated lower confidence bound, L sub p, equals the minimum value, x sub 1. If p sub L equals 1, then the estimated lower confidence bound, L sub p, equals the maximum value, x sub n.*

$$\hat{U}_p = \begin{cases} x_1 & p_U < \hat{F}(x_1) \\ x_j + \dfrac{p_U - \hat{F}(x_j)}{\hat{F}(x_{j+1}) - \hat{F}(x_j)}\left(x_{j+1} - x_j\right) & \hat{F}(x_j) \le p_U < \hat{F}(x_{j+1}) \\ x_n & p_U = 1 \end{cases}$$

*Description: The formula estimates the upper confidence bound of a percentile from an empirical distribution using linear interpolation. The estimated upper confidence bound, U sub p, equals the lower ranked value (x sub j) plus a weighted fraction of the distance between consecutive ranked values (x sub j and x sub j plus 1). The weight is determined by where the upper confidence probability, p sub U, falls between the empirical cumulative distribution values F hat( x sub j) and F hat(x sub j plus 1). The formula applies when F hat( x sub j) is less than or equal to p sub U, and p sub U is less than F hat ( x sub j plus 1). If p sub U is less than F hat( x sub 1), then the estimated upper confidence bound, U sub p, equals the minimum value, x sub 1. If p sub U equals 1, then the estimated upper confidence bound, U sub p, equals the maximum value, x sub n.*

**Figure 2: Empirical Cumulative Distribution Function Illustrating Percentile $X_p$ and confidence interval**




*Description: Empirical cumulative distribution function (ECDF) of analyte values. The x-axis shows analyte value and the y-axis shows cumulative distribution function, ranging from 0 to 1. A black stepwise curve increases from left to right, representing the cumulative proportion of observations. A solid red horizontal line marks percentile level p (approximately 0.5). The intersection of the ECDF with this line identifies the percentile estimate, X sub p, shown by a black vertical dashed line. Red dashed horizontal lines mark the lower and upper confidence levels, p sub L and p sub U. Their intersections with the ECDF correspond to the lower and upper confidence bounds for the percentile, labeled L sub p and U sub p, shown as blue vertical dashed lines. The figure illustrates that X sub p is the analyte value at or below which p percent of observations fall, with L sub p and U sub p representing the 95% confidence interval for that percentile estimate.*

**Commercial Software**

PROC DESCRIPT in SUDAAN (version 8.0 and higher) calculates confidence limits for the percentiles using the "test-inversion" method by Francisco and Fuller, as noted in Step 1.

PROC SURVEYMEANS (SAS version 9.1 and higher) can be used to obtain Woodruff like confidence intervals for percentiles. However, the SURVEYMEANS method differs slightly from the traditional Woodruff method as noted in Step 3.

## References


Francisco CA, Fuller WA. Quantile estimation with a complex survey design. Annals Statist. 1991;19:454–469.

Korn EL, Graubard BI. Analysis of Health Surveys. Wiley, New York; 1999.

Kovar JG, Rao JNK, Wu CFL. Bootstrap and other methods to measure errors in survey estimated. Can J Statist. 1988;16S:25–45.

Research Triangle Institute (2008). ***SUDAAN Language Manual, Release 10.0*** Research Triangle Park, NC: Research Triangle Institute.

Rogers JW. Estimating the variance of percentiles using replicate weights. Proceedings of the Section on Survey Research Methods. 2003.

Sitter RR, Wu C. A note on Woodruff confidence intervals for quantiles. Statist Probabil Letters. 2001;52:353–358.

Woodruff RS. Confidence intervals for medians and other position measures. J Am Statist Assoc. 1952;57:622–627.

U.S. Centers for Disease Control and Prevention. NHANES Analytic guidelines, the Third National Health and Nutrition Examination Survey, NHANES III (1988–94). Hyattsville (MD): National Center for Health Statistics; October 1996 [cited January 5, 2026]. Available from: https://wwwn.cdc.gov/nchs/data/nhanes/analyticguidelines/88-94-analytic-reporting-guidelines.pdf.