Estimation and Analytic Issues for Rare Events in NHANES

Lester R. Curtin, Deanna Kruszon-Moran, Margaret Carroll, and Xianfen Li
Lester R. Curtin PhD
National Center for Health Statistics, Centers for Disease Control
3311 Toledo Road, Room 4413, Hyattsville, Maryland, 20782

Keywords: Binomial distribution, Wald test, complex survey, rare events, Design effects

1. Introduction

Most national surveys, including the National Health and Nutrition Examination Survey (NHANES), produce descriptive statistics such as means and proportions. Major Federal policy decisions and funding allocations are made based on these types of statistics. In fact, NHANES often provides the first national data or the only national data for many relevant health issues. As such, there are major implications when assessing the statistical significance of NHANES measures. Assessment of statistical significance, even when based on well known simple statistical methods, can be problematic in certain situations. If the validity of a statistical method is an issue dealing with simple random sample data, one can infer that there will also be an issue when that method is modified for complex survey data. Certain aspects of the design for the NHANES survey make such situations even more problematic.

With easily assessable pubic use data files and with the complexity of the NHANES data sets, it has been long recognized that there is a need for written analytic guidelines and for the wide variety of users of the publicly released micro-data. NHANES, as with many national surveys, is designed with analytic specifications in terms of reliability of estimates, yet data users often go beyond those design limits. One particularly vexing problem is the analysis and interpretation of statistical measures for rare events. For NHANES, with many measures collected and many possible demographic groups, the number of situations involving rare events makes the "rare event" a fairly common occurrence.

There are a number of analytic issues related to rare events in NHANES. For complex surveys, approximations (usually based on linearization or replication) are most often used to estimate sampling errors. These methods are based on asymptotic results and may not generalize to rare events in a relatively small survey such as NHANES. Even when the sample variance is properly estimated, there are several alternative methods for computing confidence intervals. Another complication comes from the design of NHANES where sample weights can be quite heterogeneous, thus creating some controversy on when to use weighted versus unweighted data. The issues of alternative methods and weights are then confounded by the population heterogeneity of many health measures. This presentation addresses these issues and their impact on constructing confidence intervals for proportions based on survey data with emphasis on rare events.

The following presentation will address the general analytic issues for rare events then the focus will shift to specific issues in the construction of confidence intervals for "rare" proportions. Section 2 will review alternative statistical methods for proportions when the data are based on simple random samples. Section 3 will review extensions to complex survey case. For the complex survey case, evaluation of methods is often done through simulations, so the results may only generalize to surveys that correspond to the design used in the simulations. Section 4 discusses the NHANES design features that impact on the problem of confidence intervals for proportions and summarizes a set of empirical comparisons of alternative methods when applied to NHANES data. This section also includes a brief note on the application of the results to the issue of constructing confidence intervals for percentile estimates using the Woodruff method. Section 5 discusses commercial software considerations. Section 6 provides a preliminary recommendation that will be included in the next set of NHANES analytic guidelines.

2. The Simple Random Sample Case

Classical statistical methods are applied when data are assumed to be collected by a simple random sample. The data observations are assumed to be independent and identically distributed. A count of the number of events with a specific characteristic is most often assumed to follow a binomial distribution.

The binomial distribution arises naturally as the sum of n independent Bernoulli random variables. It is typical to define $X_i = 1$ if the i-th sample person has a specific characteristic (success or failure, a positive test result,

classified as overweight, etc), with probability p for i = 1, 2,…, n or $X_i = 0$ if the person does not have the characteristic with probability 1-p for i = 1, 2,…, n Then $X = \sum X_i$ is distributed as a Binomial random variable with parameters (n, p). A key consideration is the assumption of homogeneity, that p (the Bernoulli probability of "success") is constant for all i observations.

A binomial random variable has mean np and variance np(1-p). The maximum likelihood estimate for p is given by $\hat{p}$ = x/n and the variance of p is estimated by n $\hat{p}$ (1 - $\hat{p}$). The underlying assumption for the construction of confidence intervals for proportions in the classical setting is, for a sufficiently large number of trials, n. The estimated proportion of successes, $\hat{p}$, is approximately normally distributed with mean p and variance $Var(p) = \frac{p(1-p)}{N}$ which is estimated by $\text{var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$. Consequently,

$$\frac{\hat{p} - p}{\sqrt{\text{var}(\hat{p})}} \sim N(0, 1)$$

This is referred to as the pivotal statistic and is approximately normally distributed with zero mean and unit variance. In a survey setting this translates into simple random sampling where n is the sample size; depending upon the application there may or may not be a larger population where N is the population size.

The corresponding two sided (1- α) % confidence interval is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\text{var}(\hat{p})}$$

where $z_{\alpha/2}$ is the standard normal deviate. This is often referred to as the Wald confidence interval. Note that the standard normal critical value is used and this is usually based on the assumption that the variance is known, not estimated, and that the sample size is large enough to apply the Central Limit Theorem. For small sample size, one could replace the standard normal critical value with a critical value based on the student t with degree of freedom d = n – 1. The use of t_d instead of z doesn't seem to be considered in the textbook/simple random sample case (although it will be in the complex survey case).

In most textbooks, it is generally stated to use the Wald interval if the sample size is greater than 30 and the proportion is greater than 0.10 or 10 percent But Brown et al (2001) and many others have shown that exact coverage of the Wald interval behaves erratically even for p close to 0.5 or 50 percent.

One basic problem is due to the discrete outcome space for the distribution of a binomial random variable. Consider the following table based on n = 10 with a true p = 0.50. The two-sided 95% Wald confidence interval is based on the estimate $\hat{p}$ and not the true p. The true p (or p under the null hypothesis) is used in the calculation of Pr(X=x). The coverage of the Wald confidence interval is calculated by summing Pr(X=x) for all x where the confidence interval includes the true value – in this case for x = 3, 4, 5, 6, 7 and the coverage is 0.89 (less than the nominal coverage of 0.95)

X	Pr(X=x)	$\hat{p}$	Lower CL	Upper CL
0	.000977	0.0	0.0	0.3
1	.009765	0.1	-0.08974	0.289737
2	.043945	0.2	-0.05298	0.452982
3	.117188	0.3	0.010172	0.589828
4	.205078	0.4	0.090161	0.709839
5	.246094	0.5	0.183772	0.816228
6	.205078	0.6	0.290161	0.909839
7	.117188	0.7	0.410172	0.989828
8	.043945	0.8	0.547018	1.052982
9	.009765	0.9	0.710263	1.089737
10	.000977	1.0	0.7	1.0

As the true p changes from .5 to .46, each probability Pr (X=x) changes slightly, but the intervals containing the true p are the same. Only when p =.45 does the interval corresponding to x = 2 now become included in the two-sided coverage. The following illustrates how the coverage varies with true p

True proportion	Coverage
0.50	0.089
0.49	0.088
0.48	0.086
0.47	0.085
0.46	0.084
0.45	0.945

In the above table, it is seen that only values of $\hat{p}$ corresponding to observations x = 0, 1, 2, … can be observed. The coverage depends on the true value of p and the number of intervals containing the true value. This illustrates that it is the discrete nature of the outcome space that creates the erratic coverage for true p in the neighbourhood of 0.5

Because observations from simple random samples have equal (or no) weights, a short digression is

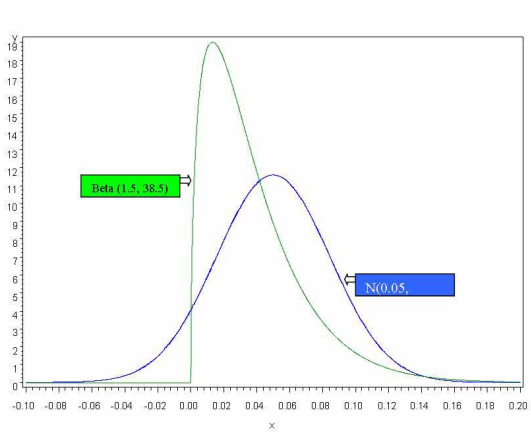
presented here for the complex survey case with differential weights, where the outcome space for possible estimates $\hat{p}$ can be increased. This is illustrated in the following example. Suppose n = 4 and the observations have weights 11, 23, 36, and 49. Now instead of 5 discrete outcomes for $\hat{p}$ = (0, 0.25, 0.5, 0.76 and 1.0), there are 17 possible outcomes. If the true proportion is 0.05, then we get the following table for the usual Wald two- sided 95% interval:

X_{wt}	X	P(X=x)	$\hat{p}$	Lower	Upper
0	0	0.6561	0.000	0.0	0.75
11	1	0.0729	0.092	-0.197	0.382
23	1	0.0729	0.193	-0.202	0.588
34	2	0.0081	0.286	-0.166	0.737
36	1	0.0729	0.303	-0.157	0.761
47	2	0.0081	0.395	-0.094	0.884
49	1	0.0729	0.412	-0.080	0.904
59	2	0.0081	0.496	-0.004	0.996
60	2	0.0081	0.504	0.004	1.004
60	3	0.0009	0.504	0.004	1.004
72	2	0.0081	0.605	0.116	1.093
83	3	0.0009	0.697	0.238	1.156
85	2	0.0081	0.714	0.262	1.166
96	3	0.0009	0.807	0.412	1.201
108	3	0.0009	0.908	0.618	1.197
119	4	0.0001	1.000	0.25	1.000

The above table indicates that, with the estimation space being increased when data are weighted, perhaps the erratic behaviour for the coverage of the Wald may not be as great in the complex survey case.

The above tables illustrate that a confidence interval for p can be calculated even when the observed number of "successes" is zero, that is x = 0 and $\hat{p}$ = 0. Hanley (1995) gives a 95 percent confidence interval for p in this case as (0, 3/n). Alternatively, when x =n and $\hat{p}$ = 1.0, the 95 percent confidence interval is given by (1-3/n, 1). This is supported by Louis (1976). The (0, 3/n) interval is used in the above examples; but alternatives have been proposed for the x = 0 case, see (Olivier and May, 2006)

When p is close to 0 or 1, the binomial assumption has an additional problem. The underlying distribution of $\hat{p}$ is not symmetric, it is skewed. This is illustrated in the following figure for n = 40, p = 0.05 where the normal with the binomial mean (np = 2) and variance (np(1-p) = 1.9) is compared to the comparable Beta(a,b) distribution with parameters a=20 and b= 20.



For the Wald confidence interval for $\hat{p}$ it is possible to obtain a negative lower limit. This is recognized in standard texts; there are a number of ways of avoiding this problem.

One way to avoid the possibility of obtaining a negative lower limit is to apply a transformation that results in a standard normally distributed random variable on the transformed scale, construct a Wald confidence interval on the transformed scale, then back transform using the inverse transformation to get the confidence interval on the original scale. Two particular transformations are the most widely used, the arcsine (square root) $y = \sin^{-1}(\sqrt{\hat{p}})$ and the logit, defined for values of p greater than zero but less than unity, $y = \log(\hat{p}/(1-\hat{p}))$.

The confidence interval for the arc-sine square root is given by

$$\sin^{-1}(\sqrt{\hat{p}}) \pm z_{\alpha/2}\sqrt{\frac{1}{4n}}$$

The confidence interval for the logit is given by

$$\log it(\hat{p}) \pm \frac{z_{\alpha/2}}{\sqrt{n\hat{p}(1-\hat{p})}}$$

Another variation on the Wald concept, similar to a continuity correction for discreteness, was considered by Agresti and Coull (1998). The Agresti-Coull Interval is given by

$$\tilde{p} \pm \kappa\sqrt{\tilde{p}\tilde{q}/\tilde{n}}$$

Where κ = normal deviate; x* = x+c; n* = n + c ; $\tilde{p}$ = x*/n* and $\tilde{q}$ = 1 - $\tilde{p}$. Agresti and Coull use c = 2 but Olivier and May (2006) consider the more general case.

The above approaches are all variations of the Wald confidence interval. A slightly different approach is to use the Wilson Score confidence interval (Wilson, 1927). This confidence interval uses the pivotal statistic

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$$

which is approximately normally distributed and is based on the quadratic form

$$\hat{p} \pm \frac{(1-2\hat{p})\frac{z_{\frac{\alpha}{2}}^2}{2n} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+(\frac{z_{\frac{\alpha}{2}}^2}{4n^2})}}{1+\frac{z_{\alpha/2}^2}{n}}$$

This is obtained by solving the equations

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \pm z_{\alpha/2}$$

for p.

Note that the denominator of the pivotal statistic is given in terms of the population parameter, p, where the denominator of the Wald pivotal statistic is given in terms of $\hat{p}$ rather than p.

In order to deal with both discreteness and skewness, it can be recognized that the discrete binomial distribution corresponds to either a beta or F-distribution (Johnson and Kotz, 1978). This is analogous to the discrete Poisson distribution corresponding to a Chi-square distribution. By using the formulation as a continuous distribution, "exact" binomial confidence intervals can be constructed as originally done by Clopper and Pearson (1934). These intervals always achieve at least the nominal coverage, but can be conservative (that is too wide). In terms of the Beta distribution, the lower limit of the Clopper-Pearson Interval for x=n $\hat{p}$ is given by

$$L_{CP}(x) = \beta(\alpha/2; x, n-x+1)$$

And the upper limit for x is given by

$$U_{CP}(x) = \beta(1-\alpha/2; x+1, n-x)$$

Alternatively, the Clopper-Pearson can be given in terms of the F-distribution as

$$L_{CP}(x) = \frac{\nu_1 F_{\nu_1,\nu_2}(\alpha/2)}{\nu_2 + \nu_1 F_{\nu_1,\nu_2}(\alpha/2)}$$

$$U_{CP}(x) = \frac{\nu_3 F_{\nu_3,\nu_2}(1-\alpha/2)}{\nu_4 + \nu_3 F_{\nu_3,\nu_4}(1-\alpha/2)}$$

Where $\nu_1 = 2x$, $\nu_2 = 2(n-x+1)$,

and $\nu_3 = 2(x+1)$, $\nu_4 = 2(n-x)$

The Clopper-Pearson interval corresponds closely to a Bayesian confidence interval (see Brown et al 2001). In this a Bayesian approach, using a Jeffreys prior, the lower and upper confidence limits are given as:

$$L_J(x) = \beta(\alpha/2; x+1/2. n-x+1/2)$$
$$U_J(x) = \beta(1-\alpha/2; x+1/2. n-x+1/2)$$

For simple random sample data, the Wald, transformed Wald, Agresti-Coull, Clopper-Pearson and Bayesian approaches have been compared by a number of authors. Newcombe (2001) found that the coverage of Logit intervals vary slightly when the intervals are narrow, but much more marked when they are wide (i.e. n small or p close to 0 or 1). Chen (1990) concluded that, among transformations, the arcsine is "almost the optimal transformation." However, Brown, et al (2001) found that while the arcsine preformed well for many values of p, it did not perform well for p close to 0 or 1. Newcombe (1998) found that the Clopper-Pearson method "unnecessarily conservative." And that the Wilson method produced reasonable intervals for all values of p, including extreme ones. However, Brown, et al (2001) found the "Wilson performs quite well in terms of coverage for p away from 0 or 1 but the interval was unnecessarily long and exceeded that of the Clopper-Pearson interval when p was close to 0 or 1. Despite some inconsistencies, the general conclusions in the simple random sample case, as summarized by Brown et al (2001) are:

1. Do not use text-book Wald interval for any p.
2. Confidence intervals for transformations are generally too wide.
3. The Clopper-Pearson is too conservative (others affirm that the Clopper-Pearson is the preferred interval).
4. Likelihood ratio interval (not examined here) was too hard to compute for the average user.
5. For not so rare events - use the Wilson Score interval or Agresti-Coull (easier to compute).
6. For rare events– use the Wilson Score interval

3. Applications to Complex Samples

Adjustments to statistical methods are generally needed when data come from a Complex survey rather than a simple random sample. Observations from a complex survey may not be independent. Many complex survey designs involve clustering of sample persons (as do many clustered-randomization trials). Not only are the observations not independent, but the complex variance for clustered designs are typically larger than the hypothetical simple random sample variance for the same sample size. Many surveys involve differential sampling fractions for specified demographic sub-domains (age, race, and sex); that and ratio adjustments (for non-response and post-stratification) yield differential sampling weights. Ignoring sample weights can lead to biased estimates for proportions and the differential weighting also increases the complex variance. For many variables, especially in health data, the assumption of homogeneity for p across sample persons (or between geographic areas) is also violated.

Before examining the alternative methods for confidence intervals for proportions based on survey data, it is useful to summarize a few issues and concepts in variance estimation for complex surveys. These issues are (1) alternative methods for variance estimation for complex survey data, (2) the relationship of complex variance to simple random sampling variance (the design effect) and (3) issues in the stability of complex survey variance estimates (the degrees of freedom).

For simple random samples, variance estimators can be given as closed form expressions. For complex survey estimators, the design and the ratio adjustments create a situation where there are no closed form expressions for variance estimators. Asymptotic approximations for complex variance estimators have been developed. The most widely used of these estimators fall into two distinct classes: linearization and replication. Replication methods are typically some form of balanced repeated replication or a jackknife approach. For a more complete discussion see Rust (1998) or Wolter (1985)

In many instances, the linearization and replication variance estimates produced for many statistics are very similar, but there can be differences depending upon the statistic and the survey design. For small proportions and an NHANES type survey design, further investigation is required to determine if the asymptotic results of previous studies really hold for a design with few Primary sampling units (PSU's) or for the estimation of a rare proportion. However, for the rest of this presentation we shall simply assume that a reasonable estimator exists for our situation.

Kish (1965, 1995) has popularized the term design effect, often denoted Deff, as an indicator of the increase (usually) in variance for an estimate when a complex design is used. He defined the Deff as

$$\text{Deff} = \text{Var}_{\text{complex}}(p)/\text{Var}_{\text{srs}}(p)$$

It should be noted that the $\text{Var}_{\text{srs}}(p)$ is a hypothetical value that assumes the same estimate of p as if it were based on the same sample size. Kish assumes simple random sampling with replacement. For a complex design involving clustering and differential weighting, the Deff can be modeled (Kish, 1992) as

$$\text{Deff} = \left\{1 + \rho(\overline{m} - 1)\right\}\left\{1 + CV_{Wts}^{2}\right\}$$

Where $\overline{m}$ is the average cluster size, ρ is the intra-class correlation and CV^2_{wts} is the coefficient of variation of the sampling weights. This formulation has been further justified by Park and Lee (2004) and by Gabler, Haeder and Lahiri (2001). Usually the deff is greater than 1 as the complex variance is greater than the hypothetical simple random sampling variance. Some efficient designs can yield a true Deff less than 1. In practice, because the complex sample variance is estimated and subject to it's own sampling error, an estimated Deff can be less than 1.

One application of the Deff is to compute what has become known as the effective sample size, that is if n is the sample size for a complex design then the effective sample size is given by $n_e = n/\text{Deff}$. The effective sample size can be interpreted as the hypothetical simple random sample size required yielding the same variance. For example, if a complex design of sample size 300 yields a complex variance with a deff of 1.5, then the effective sample size is 200 and a hypothetical simple random sample variance based on that sample size of 200 is the same as the complex variance based on the true sample size of 300. The effective sample size can be used to compare alternative survey designs or to adjust simple random sample test statistics for complex designs.

There are some issues in computing the Deff for NHANES. The definition of Deff, as used in commercial software, can vary. For example, SUDAAN has options for 4 different definitions (NHANES typically uses the second definition, namely simple random sampling with replacement). WESVAR uses simple random sampling without replacement in

defining the Deff. In NHANES, Deff vary by demographic sub domain which creates another problem of heterogeneity in the data. When sample size is small, or an event is rare, demographic groups are often combined to yield a statistically reliable estimate. The Deff for the combined domain, say all ages combined, is typically greater than the individual Deff for the sub domain, in this case, age-specific Deff. This is mostly due to the differential sampling weights by domain. Even for small demographic domains, some Deff are very large (see Lacher, Curtin, and Carroll, 2001); it is not clear if the adjustments for complex survey data, to be discussed below, will generalize to the case of very large Deff.

One further background issue must be addressed. As previously stated, an estimated variance is also subject to (sampling) variation. Under appropriate regularity conditions, a variance estimator can be considered to follow a Chi-square distribution with d degrees of freedom. For complex variance estimators, and a classic two PSU per stratum design, the nominal degree of freedom is considered to be the number PSUs minus the number of strata (see Cochran, 1977 or Korn and Graubard, 2002). Again, this is an asymptotic result and may not hold for estimates from the NHANES survey. Alternative characterization for the degrees of freedom has been proposed by Satterthwaite (1934) and, specifically for NHANES, by Jain and Eltinge (1998).

The degrees of freedom has an impact on the critical value to use when hypothesis testing or calculating confidence intervals. For estimates based on survey data, simple hypothesis tests for comparing sub-domain differences often use the student t statistic t_d instead of the standard normal score z. The NHANES survey also used the nominal degrees of freedom as publication criteria in the NHANES III analytic guidelines (reference). Specifically, a statistical reliability criteria is considered where a minimum of 12 degrees of freedom should be used as the criteria for displaying a sample error estimate (24 PSUs in 12 strata). Note the nominal degrees of freedom can differ in a survey for specific race/ethnic groups as not every PSU has members of that race/ethnic group in the underlying finite population of inference.

Because Deff are themselves estimates, and subject to sampling error, average deff effect models are sometimes used (for NHANES, see Johnson and Kovar, 1982). When analysis uses an average deff, it is not clear if the degrees of freedom should also be modified. This is an area that requires additional research. This presentation is restricted to the case where average deff are not used.

Given these background issues, several references in the literature have examined alternative methods for computing confidence intervals for proportions based on complex survey data. Approximations for complex survey variance estimation for proportions are discussed in Rust and Rao (1998). For confidence intervals for proportions, some references are Gross and Frankel (1971), Korn and Graubard (1998), Kott et al (2003). A re-sampling approach is given by Grey et al (2004) but this is beyond the scope of this presentation. Software manuals for survey data (in particular STATA and WESVAR) also contain explanatory information.

Korn and Graubard (1998) examined the Wald, the Logit transformation, a Poisson approximation (Breeze, 1990) and the Clopper-Pearson, or exact binomial. In addition, the text by Korn and Graubard (1999) examines the Wilson score method (although they use the terminology "quadratic form" and do not specifically refer to this as a Wilson score interval). Kott et al (2004) examined Wald, a modification due to Andersson-Neurman (2001), the Wilson interval and a proposed Modified Wilson approach.

The general approach of Koran and Graubard is to consider the alternative methods as derived for the simple random sample case, consider $\hat{p}$ as the weighted estimate based on the survey, calculate the complex variance of $\hat{p}$ and it's estimated Deff, and calculate the nominal degrees of freedom as the number of PSU's minus the number of strata. There are two alternatives for adjusting the sample size. One is to use the classic effective sample size $n_e=n/Deff$ and the second is to also adjust the sample size for the degrees of freedom by considering $n_e^* = (t_d/z)n_e$. Korn and Graubard suggest that the adjusted sample size should not be used if the effective sample size becomes less than the original sample size. Given the revised sample size, a "scaled" number of observations (the x in the usual binomial notation) can be calculated as $x_e = \hat{p}\, n_e$ or $x_e^* = \hat{p}\, n_e^*$. To modify the alternative methods for confidence intervals, simply substitute n_e or n_e^* for n and x_e or x_e^* for x. If using n_e instead of n_e^*, also substitute t_d for z. Thus, in the complex sample setting, the alternative methods are modified as follows:

Wald, or normal, interval

$\hat{p} \pm \kappa\sqrt{Var(\hat{p})}$ with $\kappa = t_d$ Using the definition of Design effect, this is equivalent to

$\hat{p} \pm \kappa\sqrt{DEFF * \hat{p}\hat{q}/n}$ and in terms of effective sample size this is also equivalent to

$$\hat{p} \pm \kappa\sqrt{\hat{p}\hat{q}/n_e}$$

The transformation such as arcsine and logit follow directly as:

$$\sin^{-1}(\sqrt{\hat{p}}) \pm t_{d,\alpha/2}\sqrt{\frac{1}{4n_e}}$$

And

$$\text{logit}(\hat{p}) \pm \frac{t_{d,\alpha/2}}{\sqrt{n_e\hat{p}(1-\hat{p})}}$$

There has been no attempt to modify the generalized Agresti-Coull interval in complex survey setting; although the Korn and Graubard type substitution could be used it is not clear if the additive constant "c" should also depend on the survey design.. For the generalized Agresti-Coull, a further simulation is planned to see if the constant "c" can be optimized for survey data.

The modification to the Wilson Score Interval is

$$\frac{1}{1+\kappa^2 n_e^{-1}}\left\{\left(\hat{p} + \frac{\kappa^2}{2}n_e^{-1}\right) \pm \kappa n_e^{-1/2}\sqrt{\hat{p}\hat{q} + \frac{\kappa^2}{4}n_e^{-1}}\right\}$$

Note: If n_e is used, κ becomes t_d if n_e* is substituted for n_e is used, κ becomes z.

For the Jeffreys and Clopper–Pearson intervals, the deviate κ (t_d versus z) is not directly used, so n_e* and x_e* should be substituted for the sample observed values of n and x.

The Jeffreys interval for x (not p) becomes

$$L_J\left(x^*_e\right) = \beta\left(\alpha/2; x^*_e + 1/2, n^*_e - x^*_e + 1/2\right)$$

$$U_J\left(x^*_e\right) = \beta\left(1-\alpha/2; x^*_e + 1/2, n^*_e - x^*_e + 1/2\right)$$

Clopper-Pearson Interval for x (not p), given as a Beta is

$$L_{CP}\left(x^*_e\right) = \beta\left(\alpha/2; x^*_e, n^*_e - x^*_e + 1\right)$$

$$U_{CP}\left(x^*_e\right) = \beta\left(1-\alpha/2; x^*_e + 1, n^*_e - x^*_e\right)$$

And the Clopper Pearson given as an F distribution (as k is not directly used) becomes

$$L_{CP}(x_e^*) = \frac{\nu_1 F_{\nu_1,\nu_2}(\alpha/2)}{\nu_2 + \nu_1 F_{\nu_1,t_2}(\alpha/2)}$$

$$U_{CP}(x_e^*) = \frac{\nu_3 F_{\nu_3,\nu_2}(1-\alpha/2)}{\nu_4 + \nu_3 F_{\nu_3,t_4}(1-\alpha/2)}$$

Where $\nu_1 = 2x_e^*$, $\nu_2 = 2(n_e^* - x_e^* + 1)$,

and $\nu_3 = 2(x_e^* + 1)$, $\nu_4 = 2(n_e^* - x_e^*)$

Again, the Beta and F-distribution approaches do not explicitly use t_d or z so the Korn and Graubard approach is to use the degrees of freedom adjusted effective sample size n_e*. Once the confidence interval for the number of events is calculated, the confidence intervals for the estimated proportion follow directly

Based on their simulations, Korn and Graubard concluded that the Wald interval was not good for rare events and the Logit transform was too wide; they recommended the Modified Clopper-Pearson interval using the effective sample size adjusted for the survey's nominal degrees of freedom. Kott et al (2004) did not examine the exact binomial but concluded to use modified Wilson score method.

4. Issues specific to NHANES design

There are several survey design characteristics that affect the application of these methods to NHANES survey data. The current NHANES survey is actually a continuous series of annual surveys that can be aggregated into 2, 4, 6 or more years for estimation. There are 15 clusters (PSUs) per year, with a sample size of approximately N = 5,000 per year. Because NHANES can combine years of data as 2, 4 or 6 years, the degrees of freedom for the estimated sampling error changes accordingly as 15, 30, and 45. There are specified sampling fractions for each of 72 age-race-ethnic-sex domains, with "oversampling" so that 50 percent of the sample is below age 20 years; Mexican Americans and Black Americans are also oversampled. This creates sampling weights that are very heterogeneous. Most survey measures vary by age race and sex; many characteristics have a large between PSU variance component and/or a large intra-class correlation within clusters. The design specifications require that an estimate for a 10 percent statistic have less than or equal to 30 percent RSE. With a desired Deff = 1.5 this implies n = 150 per analytic subdomain (or n_e = 100).

Given the design characteristics of NHANES, it can not be assumed that previous results, based on particular simulations, necessarily apply to the NHANES situation. In terms of simulations, Korn and Graubard (1998) let p vary by PSU, used simplified weights (either 1 or 10 and either informative or non-informative with respect to the probability of success), 32 Strata, a sample size of n = 100 per strata, and conducted 100,000 draws when comparing the coverage of alternative methods. Kott et al (2004) used p as constant between PSU, they allowed more heterogeneity in the sample weights by assuming the weights followed a chi-square distribution, and their results were based on 50,000 draws. Neither simulated design was quite like NHANES.

For this presentation, the NHANES 1999-2002 data was combined as a single, 4 year survey to empirically examine the alternative confidence interval methods for proportions. A number of variables/outcomes were examined for age, sex and race/ethnic groups: Elevated blood lead (Prevalence varies by age), Rubella seropositivity (prevalence close to 99%), Rubella seronegativity (prevalence 1%), Stahlococcus aureus (SA) (prevalence 32.4%), Methicillin Resistant Staphlococcus aureus (MRSA) (prevalence 0.4%) and HIV Status (prevalence 0.5%). Estimated proportions, standard errors, design effects, and alternative confidence intervals were constructed for these measures over age, sex, and race/ethnicity sub-domains. Different sub-domains also provided a range in the nominal degrees of freedom.

The empirical results for two-sided 95 percent confidence intervals were generally consistent across the range of estimates produced. The lower Wald confidence interval may be negative, the width of the arcsin is comparable to the Wald but shifted upward (thus avoiding negative lower limits), the width of the Wilson is greater than the arcsin and the logit gives the widest confidence interval. The Wilson interval is generally, but not always, slightly smaller than the Clopper-Pearson interval. Across all measures and subdomains, there was little variability between methods when analyzing all confidence intervals or when looking at sub-domains where estimates are stable (sufficient degrees of freedom). Methods varied more among estimates with high relative standard errors, smaller numbers of positives, lowest seroprevalence, and fewer degrees of freedom.

The following table (from Carroll and Curtin, 2001), for percent Rubella seropositivity, is typical. In this table, the upper confidence limit is multiplied by the population size to get an "estimated number" with the positive test result. The logit confidence interval is important because it is the most widely used of the alternative methods for rare events.

Method	Lower CL	Upper CL	Upper Est
Logit	0.8	9.3	334,000
Wilson	0.8	8.5	305,000
Arcsin	0.4	7.2	258,000
Wald	-0.7	6.2	223,000

The results are comparable to the next table for HIV status (also from Carroll and Curtin). Both of these tables illustrate a problem with the logit transformation. Although in a probability sense, the logit works well because the coverage is greater than the nominal coverage, there is a practical consideration. Using the upper confidence limit for the proportion to get an upper confidence limit for the estimated number, it can be seen that the alternative methods give a very different upper limit; this upper limit could be quite important as planning on a possible 50,000 cases per year (based on arcsin) is quite different from planning on nearly twice as many, 98,900 for the logit interval.

Method	Lower CL	Upper CL	Upper Est
Logit	0.1	3.7	98,900
Wilson	0.1	2.8	74,900
Arcsin	0.0	1.9	50,800
Wald	-0.5	1.4	37,450

For a very rare event, Rubella seropositivity and seronegativity, results are presented below, showing the symmetry in the methods for p and 1-p

	Pos(+)	Pos(+)	Neg(-)	Neg(-)
Method	Lower	Upper	Lower	Upper
Logit	97.0	99.9	0.1	3.0
Wilson	97.3	99.8	0.2	2.7
Exact	97.6	99.9	0.1	2.4
Arcsin	98.0	100	0.0	2.0
Wald	98.3	100.3	-0.3	1.7

As opposed to previously reported results for the Wald interval, for the complex survey case the Wald was quite similar to other methods for p or (1-p) close to 0.5 This is illustrated in the following table for Stahlococcus aureus.

Methods	Lower CL	Upper CL
Logit	34.3	39.3
Wilson	34.6	39.2
Exact	34.8	39.1
Arcsin	34.6	39.2
Wald	34.6	39.2

The logit almost always gave the widest confidence intervals and the largest upper confidence limit, again leading to an interpretation that the logit is too conservative. Over the range of estimates, the Wilson and Clopper-Pearson were almost always very similar. Usually the Wilson provided wider intervals than the Clopper-Pearson but not always. Of course, there will always be extreme examples, as illustrated for MRSA below:

Method	Lower CL	Upper CL
Logit	0.1	100
Wilson	3.2	99.0
Exact	37.0	89.0
Arcsin	48.0	20.1
Wald	-95.4	228.9

Here both the Wald and the arcsin give nonsensical results, but the prevalence is so small and the observed number so small that even the Wilson and logit basically indicate the prevalence is between 0 and 100 percent. Considering the exact method (Clopper-Pearson) meets the nominal coverage, its confidence interval performs reasonable well even in this extreme example

A brief note on the application of these results to the calculation of confidence intervals for percentiles. NHANES data are often used to estimate percentile distribution for environmental containments such as blood lead, mercury, persistent pesticides and other chemicals (see NCEH, 2002). One procedure for estimating the Confidence intervals for percentiles is the Woodruff procedure (Woodruff, 1952). This method requires the calculation of the corresponding estimates for proportions (that is the 99 percentile requires the confidence interval for a proportion corresponding to 99 percent. Therefore, any recommendation for the confidence interval for a proportion should carry over to implementing the Woodruff procedure. The NCEH report used the Korn and Graubard modification to the Clopper-Pearson in implementing the Woodruff method.

5 Software considerations

As previously stated, analytic guidelines for NHANES need to take into consideration methods as currently available in commercial software packages. Software such as SUDAAN and SAS currently include only the Wald and the logit transformations for survey data. In STATA, one has to be careful in that different procedures implement different approaches. The procedure sysmean can be used to estimate a proportion (as the mean of a 0,1 variable) but gives a Wald confidence interval. The STATA manual states the logit is used for procedure sysprop, but the example of output shows a "adjusted wald" that appears to be the modified Clopper-Pearson interval (although the documentation in the manual and associated technical report is not specific). WESVAR appears to be the only package that currently (as of September, 2006) uses the Wilson score confidence interval for survey data. The options available for these survey packages are summarized below

	SUDAAN	STATA	WESVAR	SAS
Var	Lin, Rep	Lin	Rep	Lin
DEFF	4	1	2	1
DF	Nom	Nom,Sa	Nom	Nom
CL	Yes	Yes	Yes	Yes
Trans	Logit	Logit	Logit	Logit
Wilson	No	No	Yes	No
Exact	No	Yes	No	No

6 Preliminary Conclusions for NHANES Analytic Guidelines

Based upon previous published research and empirical assessment using NHANES 1999-2004 data, a set of recommendations is being prepared for publication on the CDC/NHANES website as a new set of analytic guidelines, replacing the NHANES III analytic guidelines. The basic recommendation will include:
(1) Always use weights to compute the estimated proportion and it's standard error,
(2) No specific recommendation is planned on preference for linearization versus replication methods.
(3) The Wald interval may be used for $.25 < p < .75$
(4) The Wald interval should be avoided for small p, say $p < 0.25$ The choice of an alternative method is dependent on software availability. If the data analyst is limited to SAS or SUDAAN and must use the logit, it should be recognized that the interval is conservative and the data analyst should consider using a 90% confidence interval and not a 95% interval.
(5) If using WESVAR, for small p ($p < .25$) the recommended procedure is the Wilson Score interval. For STATA users, the "adjusted Wald" in the sysprop procedure (i.e. the modified Clopper-Pearson interval) should be used.

For SUDAAN and SAS users, CDC/NHANES plans on incorporating sample programs in Analytic guidelines to compute both the Wilson interval and Clopper-Pearson interval as modified by Korn and Graubard. Once these programs are available, recommendation (4) will be revised to recommend the use of either the Wilson or the Clopper-Pearson for NHANES data when p or 1-p is small.

These should be considered preliminary recommendations as additional research is required for the nuances of the NHANES design. For rare events and small degrees of freedom additional simulations are needed (1) to compare replication versus linearization for Var(p), (2) to examine the use of the estimated DEFF or an average DEFF, and (3) to determine the most appropriate estimate for the degrees of freedom. Such a research project is currently under way.

References

Anscombe, F.J. (1956). On Estimating Binomial Response Relations, Biometrika. 43:461-464.

Agresti, A and Coull, BA.(1998). Approximate is Better than "Exact" for Interval Estimation of Binominal Proportions, The American Statistician, 52:119-126

Agresti, A. (2003) Dealing with discreteness: making 'exact' confidence intervals for proportions, and odds ratios more exact. Statistical Methods in Medical Research 12:3-21.

Andersson, P G and Nerman, O (2000) Balanced Adjusted Confidence Interval, Procedure Applied to Finite Population Sampling. Presented at the International Conference of Establishment Surveys, Buffalo, New York.

Barker, L (2002). A comparison of nine confidence intervals for a Poisson parameter when the expected number of events is less than 5. American Statistician 56:85-89

Berger RL and Coutant BW (2001) Comment on small sample interval estimation of Bernoulli and Poisson parameters by D Wardell. The American Statistician 55:85

Berger YG and Skinner CJ (2003). Variance estimation for a low income proportion. Applied Statistics 82: 457-468

Bohning, D and Viwatwongkasem, C (2005). Revisiting proportion estimators. Statistical Methods in Medical Research 14:147-169

Breeze E (1990). General Household survey report on sampling error. London Her Majesty's Stationary office (Office of Population and Surveys)

Brown, L.D., Cai, T.T. and Das Gupta, A. (2001). Interval Estimation for a Binomial Proportion, Statistical Science. 16:101-133.

Blyth CR and Still HA (1983). Binomial confidence intervals. JASA 78:108-116

Carroll MD and Curtin LR (2004). On the construction of Confidence intervals for rare events based on data from a complex survey. Proceedings of ASA

Chen, H (1990) The accuracy of approximate intervals for a Binomial parameter. JASA 85:514-518

Clopper, C.J. and Pearson, E.S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, 26 404-413.

Cochran WG (1977) Sampling Techniques Wiley, New York

Cox, D.R. and Snell, E.J. (1989). Analysis of Binary Data, 2nd ed. Chapman and Hall: London.

Gabler S, Haeder S, Lahiri (1999). A model based justification of Kish's formula fro design effects for weighting and clustering. Survey Methodology 25:105-106

Gray, A., Haslett, S. and Kuzmicich, G. (2004). Confidence Intervals for Proportions from Complex Survey Designs. Journal of Official Statistics, 20, 705-723.

Gross and Frankel (1991). Confidence limits for small proportions in complex samples. Communications in statistics theory and methods 20:951-975

Hanley JA Lippman-Hand A (1995) If nothing goes wrong, is everything all right? JASA 249:1743-1745.

Jain DS and Eltinge JL (1995) Empirical assessment of the stability of variance estimators based on a two cluster per stratum design. Contract report to NCHS

Jennings DE (1987) How do we judge confidence interval adequacy. American statistician 41:335-337

Johnson and Kotz (1978) Continuous Distributions, Vol 1. Wiley Press

Johnson and Kovar (1983) The average design effect in NHANES Survey data . Proceedings of ASA

Kish, L (1995) Methods for Design Effects, Journal of Official Statistics. 11(1): 55-77.

Kish (1965) Survey methods, Wiley, New york

Kott PS (1994) A Hypothesis test of Linear Regression Coefficients of survey data. Survey Methodology 20, 159-164

Kott PS and Carr DA (1997) Developing an estimation strategy for a pesticide data program Journal of Official Statistics 13: 367-383

Kott, Phillip S. Andersson, Per Gosta, and Nerman, Olle (2004) Two sided coverage of intervals for small proportions based on survey data. Federal Committee on statistical methodology.

Korn, L.K. and Graubard, B.I. (1998). "Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data," Survey Methodology, 24, 193-201.

Korn and Graubard (1999). The analysis of complex survey data. Wiley InterScience, New York

Lacher, Curtin and Carroll (2001) Extreme design effects and what to do about them. Proceedings of ASA

Louis TA (1981). Confidence Intervals for a binomial parameter after observing no success. The American Statistician 35:154

Newcombe, RG (1998). Two sided confidence

intervals for the single proportion: comparisons of seven methods Statistics in Medicine 17:857-872.

Newcombe (2001). Logit confidence intervals and the Inverse Sinh Transformation. The American Statistican. 55:200-202

Centers for Disease Control (2003) Second national report on human exposure to environmental chemicals DHHS report (NCEH Pub no 02-0716)

Olivier, J and May, WL (2006). Weighted confidence interval construction for binomial parameters. Statistical methods in medical research 15:37-46

Park I and Lee H (1994) Design effects for weighted means and total estimators under complex survey sampling survey methodology 30:183-193

Rao JNK and Wu CFJ (1985) Inferences from stratified samples secondary analysis of three methods for nonlinear statistics. JASA 80 620-630

Rust KF (1995) Variance estimation in complex surveys Journal of Official Statistics

Rust KF and Rao JNK (1996) variance estimation for complex surveys using replication techniques. Statistical methods in medical research 5:283-310

Satterthwaite FE (1946). An approximate distribution of estimates of variance components Biometrics 2:110-114

Sitter, R.R. and Wu, C. (2001). A Note on Woodruff Confidence Intervals for Quantiles. Statistics and Probability Letters, 52, 353-358.

StataCorp. Survey Data Manual (2005) Stata statistical software release 9 College Station TX

SUDAAN User's Manual Release 8. (2002) Research Triangle Park, North Carolina.

Vollset SE (1993) Confidence intervals for a binomial proportion. Statistics in Medicine 12: 809-824

Vos PW and Hudson S (2005) Evaluation criteria for discrete confidence intervals beyond coverage and length. American Statistician 59:137-142

Wardell DGH (1997). Sample interval estimation of Bernoulli and Poisson parameters. American Statistician 51: 321-325

Wendall JP and Cox SP (2005). Coverage properties of Optimized confidence intervals for proportions. Journal of modern applied statistical methods 4:43-52

Westat (2000). A User's Guide to WESVAR, Version 4. Rockville, Md.: Westat

Wilson, E.B. (1927). "Probable Inference, the Law of Succession, and Statistical Inference," Journal of the American Statistical Association, 22, 209-212.

Woodruff RS (1952) Confidence intervals for medians and other position measures JASA 47:635-646

Wolter, KM, Introduction to Variance Estimation, Springer-Verlag, New York, Berlin, Heidelberg and Tokyo, 1985. (Chapter 6)