3659-2019

Using SAS® MACRO to Demonstrate Data Presentation Standards for Proportions

Mary Ann Bush, Nazik Elgaddal, National Center for Health Statistics

ABSTRACT

This paper describes a method to apply the National Center for Health Statistics' (NCHS) data presentation standards for reporting proportions in NCHS reports and data products. This method was developed by the Data Suppression Workgroup at NCHS. Standards are particularly important for large, cross-cutting reports where estimates cannot be individually evaluated and indicators of precision cannot be included alongside the estimates. A SAS® MACRO, with code for direct computation based on mathematical formulae, is used to apply the multistep data presentation standards based on minimum denominator sample size and on the absolute and relative widths of a Confidence Interval (CI) calculated using the Clopper-Pearson method, or the Korn and Graubard method for complex surveys. The MACRO provides a clear, efficient, and transferrable method to implement the SAS custom code. This MACRO is useful to SAS programmers if they want to replicate/match/benchmark against published NCHS estimates or if they want to use with their own data. Data from the National Health Interview Survey (NHIS) is used to demonstrate the application of the data presentation standards for proportions.

INTRODUCTION

NCHS is one of the 13 U.S. federal statistical agencies, and is responsible for collecting, analyzing, and disseminating official health statistics on a broad range of health topics through diverse publications, databases, and tables. NCHS uses a variety of data collection mechanisms to fulfill its mission to collect and report reliable statistics.

The National Vital Statistics System (NVSS) provides the nation's official vital statistics data. NCHS obtains data from all birth and death records filed in the 50 states and the District of Columbia. NCHS also conducts national surveys based on samples designed to represent the American population such as the National Health Interview Survey (NHIS). The data collected in the NHIS are obtained through a complex, multistage sample design that involves stratification, clustering, and oversampling of specific population subgroups.

Guidelines for data presentation have differed across data systems and have changed over time. The Relative Standard Error (RSE=SE/estimate) had been the most frequently used guideline in NCHS reports, where estimates with an RSE > 30% (or some other threshold) were identified as less reliable or even suppressed. When dividing the SE by very small proportions, the RSE can be too conservative, or conversely, when dividing the SE by very large proportions, the RSE can be too liberal. The NCHS Data Suppression Workgroup was formed to develop updated presentation criteria to be readily and transparently applied to proportions (usually multiplied by 100 and expressed as percentages)(1). Proportions are the most commonly published estimates in NCHS reports. These presentation criteria are being applied to all NCHS publications.

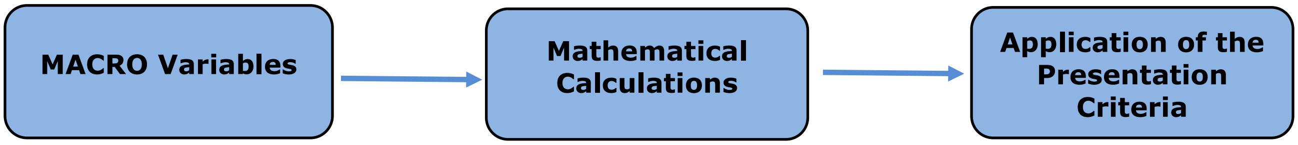
The NCHS Data Presentation Standards for Proportions rely on both minimum denominator sample size and the absolute and relative confidence interval (CI) widths. Estimates for absolute and relative CI widths are based on the 95% two-sided Clopper-Pearson CI or the Korn-Graubard method for complex surveys (2, 3).

A SAS MACRO with code for direct computation based on mathematical formulae is used to apply the multistep data presentation standards. For complex sample surveys, default calculations from survey software may not always be appropriate or even feasible for all situations including:

- Age-adjusted estimates
- Estimates based on multiple imputation
- Estimates for subgroups represented in only a subset of primary sampling units (PSUs) (e.g., some racial and ethnic groups and region-specific estimates)
- Estimates calculating annual or survey cycle estimates using a multiyear or multicycle data file.

The SAS MACRO provides programmers and analysts enhanced flexibility to apply the guidelines in an automated and efficient manner across their analytic programming.

THE SAS MACRO DESCRIPTION


MACRO Variables
Mathematical Calculations
Application of the Presentation Criteria

MACRO Variables

The SAS macro requires the input of keyword parameters with both the name and value. See Table **1** below for macro variable contents.

Macro Variable	Description
percent = or p =	percent of interest or proportion of interest
sepercent = or sep =	standard error of percentage or standard error of the proportion
nsum =	unweighted denominator or nominal sample size of the proportion
df =	degrees of freedom (degrees of freedom are not applicable to vital statistics data)

Table 1 Macro Variable List and Description

Percent (&percent) and Standard Error of Percent (&sepercent) or Proportion (&p) and Standard Error of the Proportion (&sep)

The variables (&percent and &sepercent) or (&p and &sep) may be passed into the MACRO. If the MACRO variables (&percent and &sepercent) are used, the MACRO will convert these to proportions. For NVSS data, the proportion and its standard error can be obtained through direct calculation using formulas. The number of births or deaths reported for an area represent complete counts of such events. As such, they are not subject to sampling error but these vital events are subject to random variation (4,5). Proportions and standard errors obtained through survey data require analytic procedures to handle the clustering protocols that are used in the multistage selection sample. Several software packages are available for analyzing complex samples such as the SAS Survey Software.

Sample Size (&nsum)

The &nsum parameter for sample size (denominator) is an important indicator of an estimate's precision. For vital statistics, the sample size for a proportion is the number of births (or deaths) in the denominator. For an estimate from a complex survey, the sample size is the unweighted denominator of the proportion.

Degrees of Freedom (&df)

For complex sample surveys, the precision of the estimated variance is approximately inversely related to the square root of the degrees of freedom. The use of SE's, derived from variance calculations with low precision, to assess estimated proportions may lead to poor measures of effective sample size and CI widths. Estimates based on fewer than 8 degrees of freedom are flagged for statistical review by a clearance official or subject matter expert. This review may result in either the presentation or suppression of the proportion.

The degrees of freedom can be calculated as the number of primary sampling units (PSU's) minus the number of strata. This calculation is used in most NCHS surveys. This method is also used by survey software packages, but specific calculations can vary across packages. Default calculations of degrees of freedom from survey software may not be appropriate for subgroups represented in only a subset of PSU's (e.g., some racial and ethnic groups and region-specific estimates) and when calculating annual or survey cycle estimates using a multiyear or multicycle data file. In these instances, the relevant information should be extracted and the degrees of freedom directly calculated to assess estimate precision. The calculation of degrees of freedom as a measure of precision for the SE may not be applicable for all surveys (see survey-specific documentation).

SAS MACRO code for the parameters is presented below:

```
 %macro Presentation_Code(percent=, sepercent=, p=, sep=, nsum=, df=);
 *p and its standard error should be decimal numbers between 0 and 1
 *check if user has entered a percent or a proportion;
    %if &percent^= %then %do;  *convert percent to proportion;
          p=&percent/100;
          sep=&sepercent/100;
          q=1-p;
    %end;
    %else %if &p^= %then %do; *assume that user has entered a proportion;
          p=&p;
           sep=&sep;
          q=1-p;
    %end;

    nsum=&nsum;
    df_flag=0;

* check if user entered a df value
* if not, then set df to sample(nsum) - 1;
    %if &df= %then %do;
          df=nsum-1;
    %end;
    %else %do;
          df = &df;
          if df<8 then df_flag=1;
    %end;
```

Mathematical Calculations

The MACRO computes various statistics necessary to calculate a 95% two-sided confidence interval using the Clopper-Pearson method, or the Korn and Graubard method for complex surveys. Table **2** lists these calculations with a summary of their computational purpose.

Macro Calculation	Purpose
Effective sample size ($\boldsymbol{n}_e$)	For complex surveys: the effective sample size is the sample size, n, divided by the design effect. For NVSS data, the effective sample is the number of births (or deaths) in the denominator
Korn and Graubard Adjustment to the effective sample size for complex survey data	Applies the adjustment to sample size suggested by Korn and Graubard for complex survey data that incorporates information from the survey design, including the effective sample size and, when appropriate, the degrees of freedom.
Confidence Interval	Calculate the 95% two-sided CI's using the Clopper-Pearson method or Korn and Graubard method for complex surveys.
Absolute and relative widths for proportions and for the complement of the proportion	For absolute widths: subtract the value of the lower confidence limit from the value of the upper confidence limit. For relative widths: Divides the absolute CI width by the proportion and multiplied by 100%.

Table 2 Macro Calculations and Purpose

Calculation of Effective Sample Size

The sample size (denominator) is an important indicator of an estimate's precision. The variance of a proportion is directly related to the sample size. For purposes of the presentation criteria, the effective sample size for estimated proportions should be based on a minimum denominator sample size. The effective sample size for a complex survey such the NHIS, is defined as the sample size, *n*, divided by the design effect, which is the ratio of the calculated variance to the variance expected with a simple random sample. The equation below presents a method to calculate (n_e) for estimated proportions from complex surveys:

$$n_e = \frac{\hat{p}(1-\hat{p})}{\widehat{var}(\hat{p})}$$

where $\hat{p}$ is the estimated proportion and $\widehat{var}(\hat{p})$ is the estimated variance of the proportion under the complex survey design. If the number of events is 0, the estimated proportion will be 0, the estimated variance of the proportion will be 0 and the effective sample size will be undefined. In this case, the effective sample size will be the sample size, n.

For complex sample surveys, due to sampling design and variability, there may be cases where the effective sample size is greater than the sample size. When the effective sample size is greater than the sample size, the sample size should be used to determine whether the minimum sample size criterion is met, and it should be used for CIs and other computations that include the effective sample size.

SAS MACRO Code for Calculations of Effective Sample:

```
*Effective sample size
*compute n effective
*note: for proportions from vital data files where SE=(p*q)/N, n_eff
 will equal to N;
   if (0<p<1) then n_eff=(p*(1-p))/(sep**2);
   else n_eff=nsum;

   if (n_eff=. or n_eff>nsum) then n_eff=nsum;
```

Calculation of the Korn and Graubard Adjustment to the Effective Sample Size for Complex Survey Data

The Korn and Graubard adjustment to sample size incorporates information from the survey design, including the effective sample size and, when appropriate, the degrees of freedom. The degrees of freedom adjusted effective sample size (n_e^*) is computed by applying the degrees of freedom adjustment to the effective sample size, n_e, where df is the degrees of freedom and $t_{df}(1-\alpha/2)$ is the 100(1 – α/2)*th* percentile of the *t* distribution with df degrees of freedom.

$$n_e^* = n_e \left(\frac{t_{(n-1)}(1-\alpha/2)}{t_{df}(1-\alpha/2)} \right)^2$$

SAS MACRO Code for Calculations the Korn and Graubard adjustment to the effect sample size:

```
*Ratio of ts: adjustment to sample size suggested by Korn and Graubard
 for complex survey data;
*A two-sided α (0.05/2 or 0.025) is used in the equation below: 1-0.025 =
0.975 ;
   if df > 0 then rat_squ=(tinv(.975,nsum-1)/tinv(.975,df))**2;
   else rat_squ=0;          *limit case: set to zero;

*df-adjusted effective sample size (can be no greater than the sample
size);
   if p > 0 then n_eff_df=min(nsum,rat_squ*n_eff);
   else n_eff_df=nsum;      *limit case: set to sample size;
```

Calculation of Confidence Intervals

Figure **1** contains the Korn and Graubard modified equation for the Clopper-Pearson confidence intervals for a proportion (6).

$$P_L \;=\; \left(1+\frac{n_{\mathrm{e}}^{*}-\hat{p}n_{\mathrm{e}}^{*}+1}{\hat{p}n_{\mathrm{e}}^{*}\,F\left(\alpha/2\,,\;2\hat{p}n_{\mathrm{e}}^{*},\;2\left(n_{\mathrm{e}}^{*}-\hat{p}n_{\mathrm{e}}^{*}+1\right)\right)}\right)^{-1}$$

$$P_U \;=\; \left(1+\frac{n_{\mathrm{e}}^{*}-\hat{p}n_{\mathrm{e}}^{*}}{\left(\hat{p}n_{\mathrm{e}}^{*}+1\right)F\left(1-\alpha/2\,,\;2\left(\hat{p}n_{\mathrm{e}}^{*}+1\right),\;2\left(n_{\mathrm{e}}^{*}-\hat{p}n_{\mathrm{e}}^{*}\right)\right)}\right)^{-1}$$

where $F(\alpha/2,\ b, c\,)$ is the $\alpha/2$ percentile of the F distribution with b and c degrees of freedom, n_{e}^{*}, is the adjusted effective sample size, and $\hat{p}$ is the proportion estimate.

Figure 1 Modified Clopper-Pearson Confidence Intervals For A Proportion.

A confidence interval provides a way to assess an estimate's precision. The coverage of a 95% Clopper-Pearson CI is generally conservative, meaning that the CI includes the true proportion more than 95% of the time. For calculation of Clopper-Pearson CI, with Korn and Graubard adjustment for complex surveys, the degrees of freedom adjusted effective sample size (n_{e}^{*}) is substituted for the sample size and $(n_{\mathrm{e}}^{*}\,\hat{p}\,)$ represents the number of positive outcomes. The Clopper-Pearson interval can also be computed using the Beta distribution, an accepted alternative (7, 8). The MACRO code shown below calculates parameters for beta confidence limits. This method was chosen for numerical efficiency. The lower bound is set to 0 when *p* = 0, and the upper bound is set to 1 when *p* =1. The MACRO code for calculation of the beta confidence intervals makes use of the SAS BETAINV Function.

SAS MACRO Code for Calculation of Clopper-Pearson CI (with Korn and Graubard adjustment):

```
*Parameters for beta confidence limits;
   x=n_eff_df*p;
   v1=x;
   v2=n_eff_df-x+1;
   v3=x+1;
   v4=n_eff_df-x;

*lower and upper confidence limits for Korn and Graubard interval
*Note: Using inverse beta instead of ratio of Fs for numerical efficiency
*if (0<p<1), otherwise set lower limit to 0 when p=0 and upper limit to 1
 when p=1
*A two-sided α (0.05/2 or 0.025) is used in the equations below: 0.025 and
 0.975;
   if (v1=0) then kg_l=0;
   else kg_l=betainv(.025,v1,v2);
   if (v4=0) then kg_u=1;
   else kg_u=betainv(.975,v3,v4);
```

Calculation of Absolute and Relative Widths for Proportions and for the Complement of the Proportion

From a calculated CI, the absolute CI widths for the estimate and the complement of the estimate are obtained by subtracting the value of the lower confidence limit from the value of the upper confidence limit. The relative CI widths for the estimates and the complement of the estimate are calculated by dividing the absolute CI widths by the proportion or the complement of the proportion and multiplied by 100%.

SAS MACRO Code for Absolute and Relative Confidence Interval Widths:

```
*Korn and Graubard CI absolute width;
         kg_wdth=kg_u - kg_l;

*Korn and Graubard CI relative width for p;
   if (p>0) then kg_relw_p=100*(kg_wdth/p);
   else kg_relw_p=.;

*Korn and Graubard CI relative width for q;
   if (q>0) then kg_relw_q=100*(kg_wdth/q);
   else kg_relw_q=.;
```

Application of the Presentation Criteria

Table **3** Summarizes the NCHS Data Presentation Standards for Proportions. The code below shows the process used to implement the Standards. When applying the NCHS Data Presentation Standards for Proportions, estimates identified as unreliable will be suppressed and replaced with an asterisk. Other estimates will be flagged for statistical review by a clearance official or a subject matter expert. Results from the MACRO provide reasons that an estimate was suppressed or flagged for statistical review. These reasons may then be published along with the estimate table.

Statistic	Criteria
Sample Size	Estimates with either a minimum denominator sample size or effective denominator sample size (when applicable) less than 30 should be suppressed. If all other criteria are met for presentation, an estimate based on 0 events should be flagged for statistical review by the clearance official. The review could result in either the presentation or the suppression of the proportion.
Confidence interval	If the effective sample size criterion is met, calculate a 95% two-sided confidence interval using the method of Clopper and Pearson, or the method of Korn and Graubard for complex surveys, and obtain its width.
Small absolute confidence interval width	If the absolute confidence interval width is greater than 0 and less than or equal to 0.05 then the proportion can be presented, or
Large absolute confidence interval width	if the absolute confidence interval width is greater than or equal to 0.30 then the proportion should be suppressed, or

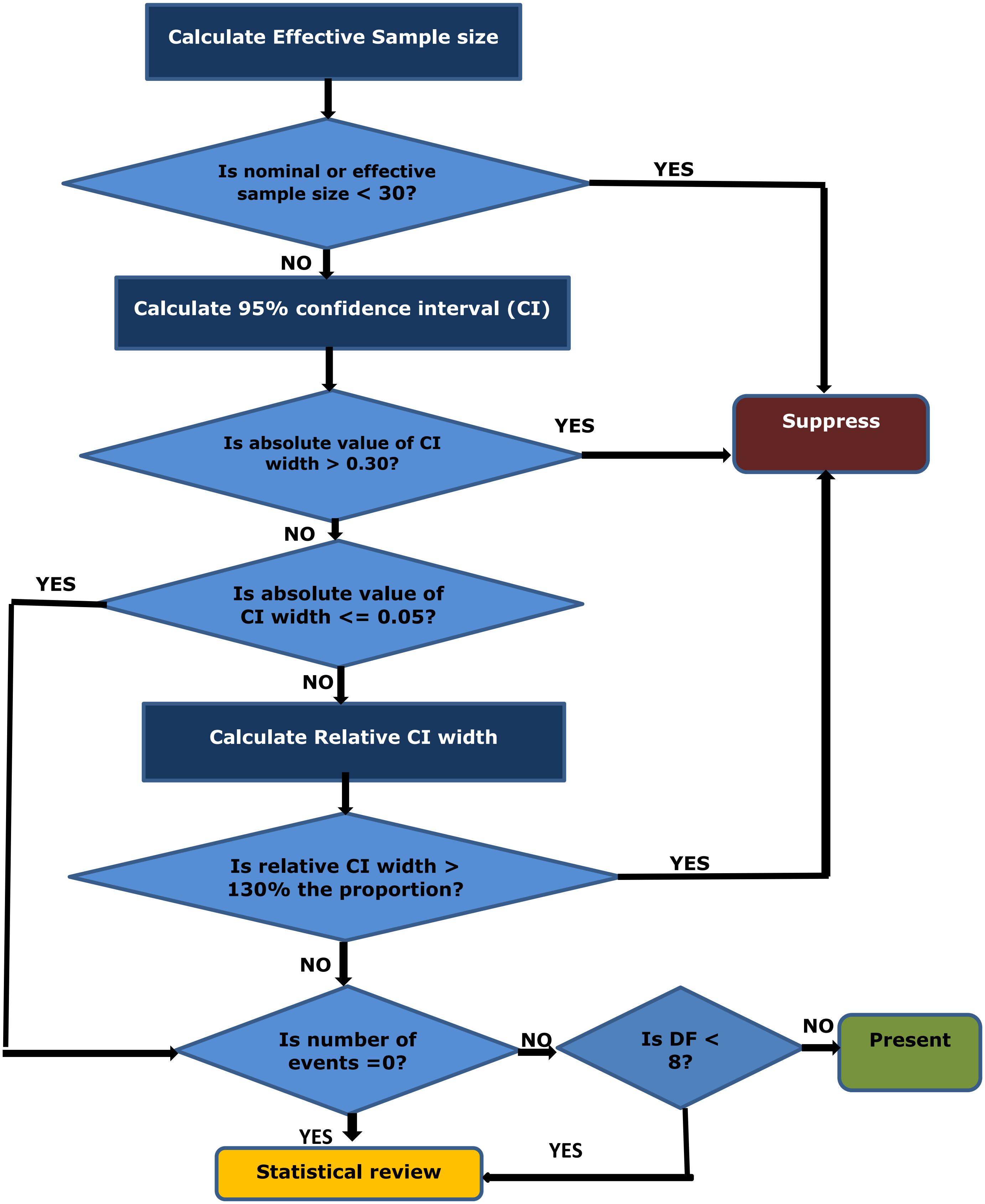
Relative confidence interval width	if the absolute confidence interval width is between 0.05 and 0.30 and the relative confidence interval width is more than 130% times the proportion then the proportion should be suppressed, or
Relative confidence interval width	if the absolute confidence interval width is between 0.05 and 0.30 and the relative confidence interval width is less than or equal to 130% times the proportion then the proportion can be presented.
Degrees of freedom	If for complex surveys, the sample size and confidence interval criteria are met for presentation and the degrees of freedom are fewer than 8, then the proportion should be flagged for statistical review by the clearance official. This review may result in presentation or suppression of the proportion.
Complementary proportions	If all criteria are met for presenting the proportion but not for its complement then the proportion should be shown. A footnote indicating that the complement of the proportion may be unreliable should be provided.

Table 3 Summary of Presentation Criteria.

MACRO Code for application of the presentation criteria:

```
*Proportions with CI width <= 0.05 are reliable, unless;
   p_reliable=1;
*Effective sample size is less than 30;
   if n_eff < 30 then p_reliable=0;

*Absolute CI width is greater than or equal 0.30;
   else if kg_wdth ge 0.30 then p_reliable=0;

*Relative CI width is greater than 130%;
   else if (kg_relw_p > 130 and kg_wdth > 0.05) then p_reliable=0;

*Determine if estimate should be flagged as having an unreliable
complement;
   if (p_reliable=1) then do;

*Complementary proportions are reliable, unless;
       q_reliable=1;

*Relative CI width is greater than 130% ;
       if (kg_relw_q > 130 and kg_wdth > 0.05) then q_reliable=0;
   end;

   p_staistical=0;
         if p_reliable=1 then do;

*Estimates with df < 8 or percents = 0 or 100 or unreliable complement are
 flagged for clerical or ADS review;
   if df_flag=1 or p=0 or p=1 or q_reliable=0 then p_staistical =1;
   end;

%mend Presentation_Code;
```

Implementation of NCHS Data Presentation Standards for Proportions


Calculate Effective Sample size
Is nominal or effective sample size < 30?
YES
NO
Calculate 95% confidence interval (CI)
Is absolute value of CI width > 0.30?
YES
Suppress
NO
YES
Is absolute value of CI width <= 0.05?
NO
Calculate Relative CI width
Is relative CI width > 130% the proportion?
YES
NO
Is number of events =0?
NO
Is DF < 8?
NO
Present
YES
Statistical review
YES

We will be using data from the National Health Interview Survey (NHIS) to illustrate the MACRO. The NHIS provides information on the health of the U.S. civilian noninstitutionalized population through confidential interviews conducted in households. NHIS is one of the nation's largest in-person household health surveys. It provides data for analyzing health trends and tracking progress toward achieving national health objectives. NHIS public use microdata is downloadable for free. The NHIS 2016 data can be accessed through the following link: https://www.cdc.gov/nchs/nhis/nhis_2016_data_release.htm

Percent of Children Under 18 Years Of Age With Any Emergency Room Visits in the Past 12 Months – NHIS 2016

Age and Race	Nominal Sample Size	Effective Sample Size	Percent Estimate	Standard Error of Percent	Relative Standard Error, (RSE)	Lower Bound	Upper Bound	Absolute CI Width	Relative CI Width	Degrees of Freedom
All children, White	2644	1619.22	21.4	1	4.77	19.4	23.4	.04	18.97	2643
All children, Black	446	302.28	25.4	2.5	9.86	20.6	30.7	.10	39.82	445
Children 0-5 years of age, American Indian/Alaskan Native ONLY	67	29.47	*	*	22.77	*	*	0.37	93.55	66
Children 0-5 years of age, ASIAN ONLY	206	206	1.1	.6	52.61	0.2	3.7	.04	312.36	205

NOTES

Estimates shown in black would be presented under the new Standards and it would have been presented based on the RSE criterion.

Estimates shown in red would be suppressed/flagged under the new Standards because effective sample size is less than 30, but it would have been presented based on the RSE criterion.

Estimates shown in green would be presented under the new Standards, but they would have been suppressed/flagged based on the RSE criterion.

CONCLUSION

This paper provided the ability to use a MACRO to apply the NCHS presentation criteria broadly and efficiently across existing production programs. The MACRO is transferable and designed for any SAS user interested in producing estimates for proportions with code for direct computation based on mathematical formulae.

REFERENCES

1. Parker JD, Talih M, Malec DJ, et al. National Center for Health Statistics Data Presentation Standards for Proportions. National Center for Health Statistics. Vital Health Stat 2(175). 2017. https://www.cdc.gov/nchs/data/series/sr_02/sr02_175.pdf
2. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26(4):404–13. 1934.
3. Korn EL, Graubard BI. Confidence intervals for proportions with small expected number of positive counts estimated from survey data. Surv Methodol 24(2):193–201. 1998.
4. National Center for Health Statistics. Vital statistics of the United States: Mortality, 1999. Technical appendix. Hyattsville, MD. 2004. Available from: HTTPS://WWW.CDC.GOV/NCHS/DATA/STATAB/TECHAP99.PDF.
5. National Center for Health Statistics. User Guide to the 2010 Natality Public Use File. Hyattsville, Maryland: National Center for Health Statistics. Annual product 2012. Available for downloading at: ftp://ftp.cdc.gov/pub/health_statistics/nchs/dataset_documentation/dvs/natality/userguide2010.pdf.
6. SAS Institute Inc. 2010. SAS/STAT® 9.22 User's Guide. Cary, NC: SAS Institute Inc.
7. Abraham, J. Computation of CIs for Binomial proportions in SAS and its practical difficulties. Kreara Solutions Pvt. Ltd., Thiruvananthapuram, India. PhUSE 2013 – Paper SP05. 2013.
8. Disney, S.M. "Revisiting activity sampling: a fresh look at binomial proportion confidence intervals", European Journal of Industrial Engineering, Vol. 10, No. 6, pp724-759. ISSN 1751-5254. DOI: 10.1504/EJIE.2016.081021. 2016.

ACKNOWLEDGMENTS

SAS® and all other SAS® Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

We are indebted to Cathy Duran whose coding expertise provided the foundation for the presentation criteria MACRO illustrated in this paper. We also want to thank Makram Talih who patiently reviewed our coding and testing efforts.

We want to thank Cynthia Reuben for providing the chart.

We are indebted to Renee Gindi, Julie Weeks, Barnali Das and Jennifer Parker for reviewing this paper and for their valuable suggestions and comments.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Mary Ann Bush
National Center for Health Statistics
Address: 3311 Toledo Road
City, State ZIP: Hyattsville, MD 20782
Work Phone: (301) 458-4130
Fax: (301) 458-4038
Email: MBUSH@CDC.GOV

Nazik Elgaddal
National Center for Health Statistics
Address: 3311 Toledo Road
City, State ZIP: Hyattsville, MD 20782
Work Phone: (301) 458-4538
Fax: (301) 458-4038
Email: NELGADDAL@CDC.GOV