Influential Observations in the National Health and Nutrition Examination Survey, 1999-2000

Margaret D. Carroll, MSPH and Lester R. Curtin, PhD
Margaret D. Carroll,
National Center for Health Statistics Centers for Disease Control,
6525 Belcrest Road, Hyattsville, Maryland, 20782

Keywords: Outlier, influential sample weight

1. Introduction:

In most statistical settings an unusually high or unusually low observation can substantially influence results. In a survey setting each sample person is assigned a sample weight which can be interpreted as the number of persons in the population represented by the individual. Consequently an unusually high observation coupled with an unusually high sample weight – or an unusually low observation - coupled with an unusually high sample weight – can severely affect estimates of population parameters and yield improbable results. The magnitude of the impact is inversely proportional to the sample size, i.e. the smaller the sample size the greater the impact.

In the past NHANES was designed to provide estimates after 3 to 6 years of data collection. The yearly samples for the current NHANES conducted by the National Center for Health Statistics/Centers for Disease Control (NCHS/CDC) are designed to provide nationally representative samples with sufficient sample size to make estimates for many subgroups using two years of data collection. However, the fact that the sample for the 1999-2000 NHANES is a smaller sample than previous NHANES surveys (current sample=9965) presents some methodogic challenges.

In this paper components of the influential weights (stage-specific probabilities of selection), are discussed, examples of influential observations are presented and approaches to handling influential observations are suggested.

2. Methods:

The NHANES 1999-2000 is the most recent in a series of national health and nutrition examination surveys conducted by NCHS/CDC and the first in a series of ongoing annual health examination surveys. Through these surveys a wide variety of health related data were collected on a nationally representative sample of the United States non-institutionalized population. National estimates of various health characteristics such as the prevalence of obesity of adults, the prevalence of current smokers, and mean serum vitamin B12 can be obtained.

There were 53 race/ethnic (Mexican American, black, white and other), gender and age sampling domains in 1999. Beginning in 2000 white and others were further stratified by income making a total of 76 sampling domains.

The sampling plan of NHANES 1999-2000 was similar to that of prior NHANES surveys. It involved the selection of primary sampling units (PSUs) – counties or groups of contiguous counties, segments within PSUs, dwelling units within segments and sample persons within dwelling units. There were 26 PSUs – 1 certainty PSU and 25 non-certainty PSUs selected with probability proportional to a measure of size, which depended upon the number of Mexican Americans within PSU. PSUs were randomly selected from 2 of 4 National Health Interview Survey Panels, i.e. nationally representative , mutually exclusive PSUs. Subgroups oversampled included Mexican Americans, blacks, adolescents, the elderly and in 2000 low-income whites and pregnant women.

Information was collected in three phases – screener, interview and examination. During the screener, sample persons were identified and basic demographic data such as age and sex were collected. Household, family and person level data on health characteristics, e.g. currently taking medication for high blood pressure and cigarette smoking, were collected in the home interview. The standardized examination, conducted in mobile examination centers (MEC), consisted of e.g. anthropometric measurements such as height and weight, blood pressure measurements and the collection of blood and urine samples.

3. Sample weight

For NHANES 1999-2000 the basic components of the sample weight include 1) the adjusted base weight, 2) non-response adjustment factors 3) trimming factors and 4) poststratification factors

factors to independent population estimates.

In most related surveys the calculation of the sample weight involves an unadjusted base weight, which is defined as the inverse of the probability that a sample person is selected. The selection probability reflects the probability of selection at each stage of the sample design – PSU, segment within PSU, dwelling unit within segment and sample person within dwelling unit. In forming the adjusted base weight for NHANES, 1999-2000, this unadjusted base weight was inflated by two factors – a subsample factor and a deselection factor.

The subsample factor reflecting the number of dwelling units released to the interviewer for screening was applied in instances where the actual number of sample persons per PSU exceeded the expected number. This value was typically 1.5. Its largest value was 2 and occurred in a low income white PSU.

The deselection factor reflecting the proportion of released dwelling units deselected from the sample was applied in order to keep the number of persons in a given PSU to a manageable number. The largest value for this factor – and the only value different from 1 – was 1.86 occurring in a PSU with a high proportion of Mexican Americans. These two factors can contribute substantially to an influential sample weight. Of the 10 largest sample weights, 6 have a subsample factor of 2 and 1 has a deselection factor of 1.86 (Table 1).

Non-response adjustments were carried out at each phase of data collection – screener , interview and examination. An overall non-response adjustment factor is the product of the non-response adjustment at these 3 phases. The total nonresponse adjustment factor can be substantial even when the individual components are not substantially different from unity as shown in Table 1.

Poststratification also occurred at each phase of data collection to adjust for undercoverage. Here the total postrastification for the top 10 sample weights is not substantially different from unity as shown in Table 1.

In instances where 1) the product of the base weight and the screener non-response adjustment factor or 2) the product of the final interview weight and the examination non-response exceeds certain threshold values, the weights are trimmed. However, even after trimming, the sample weight can be large. For example, the 2 largest sample weights shown in Table 1 have been trimmed. They are 15 percent larger than the next 2 largest weights. Overall the sample weights ranged from 980- 261,361.

4. Examples:

We first consider the impact of influential observations on the mean of a continuous variable. Vitamin B12, measured on blood samples, provides an example of this type of variable. One way of detecting an outlier heuristically is to plot this laboratory variable against the sample weight for the subdomains of interest. Figure 1 provides an example for Mexican American women 20-39 years of age where an observation with a large vitamin B12 value and a large sample weight is circled.

To determine the impact of this observation on the mean vitamin B12 for this small subdomain (n=254), the estimated mean and its standard error with the influential observation included, are compared to the corresponding estimates excluding the observation. In each case the reliability of the mean is assessed using the relative standard error (RSE) defined as the ratio of standard error of the mean to the mean times 100. The impact of this influential observation on the mean and its standard error is also presented for two larger subdomains i.e. all women 20-39 years (n=899) and Mexican American women 20 years of age and older (n=657). The results are shown in Table 2a.

For all women 20-39 years there is a substantial reduction in the estimated standard error of the mean when the outlier is excluded. The estimate with the outlier included is 156 and only 27.5 when it is excluded (a ratio of 5.6:1). Excluding the outlier also substantially reduces the relative standard error – 24 percent with the outlier included vs. 5.3 percent with the outlier excluded. There is also a substantial difference in the estimated mean – a difference that is 22.8 percent of the estimate with the outlier included but these differences are larger for the two Mexican American subdomains – 67.3 percent for 20 years and older and 74.5 percent for 20-39 years . For these subdomains the estimates with the outlier included are extremely unreliable (RSE 53% for Mexican American women 20 years and older and 76% for Mexican American women 20-39 years) whereas with the outlier removed the relative standard errors are well below 20 percent.

This extremely high vitamin B12 value could be a valid observation and therefore should be retained. Measures of central tendency – robust to outliers- provide a means of including outlying values. Two robust measures are the median and the mean of the

log-transformed value. Table 2b shows the median together with the corresponding 95% confidence limits. There is virtually no difference between the estimated medians. Table 3c shows the mean log-transformed value where there is virtually no difference in the estimated medians excluding the influential observation for all women 20-39 years and only a slight difference for Mexican Amercian women 20 year and older and Mexican American women 20-39 years. Each estimate is within the 95 percent confidence interval of the other.

The second set of examples deal with prevalences based on dichotomous variables defined using cutpoints of body mass index – a continuous variable. Here two examples are given – 1) obesity of adults 20 years of age and older and 2) "overweight or at risk of overweight" of children and youths ages 6-19 years. Body mass index (BMI) is defined as weight in kilograms divided by height in meters squared. Pregnant women and pregnant girls are excluded.

An adult with a BMI value of at least 30 kg/m^2 is defined to be obese.[1/] Influential observations for this discrete variable (obese vs nonobese) can also be detected heuristically by plotting BMI against the sample weight. Here a threshold line differentiating obese from non-obese sample persons is drawn parallel to the x axis at the point corresponding to a BMI value of 30 kg/m^2 on the y axis. Figure 2 gives an example for non-Hispanic black women 20-39 years . An observation with a large sample weight is circled. Although the BMI value is not unusually large, it is above the threshold value indicating that it corresponds to an obese individual. A similar observation occurs for Mexican American women 40-59 years.

To examine the impact of each observation on prevalence estimates, estimates including the influential observation are compared to corresponding estimates excluding them. This is done for the two subgroups mentioned above and for larger subgroups containing them, i.e. non-Hispanic black women 20 years of age and older and Mexican American women 20 years of age and older. Results are shown in Table 3. Both estimates of obesity (with the outlier included as well as those with the outlier excluded) are statistically reliable, having a relative standard error well below 20 percent. What is important, however, from a subject matter point of view is the difference between corresponding prevalence estimates. Large differences can translate into a substantial number of estimated persons in the population. In this example the largest absolute difference in corresponding percentages occurs for Mexican American women ages 40-59 years – an absolute difference of 2.1 percent. These absolute differences are larger in the subdomains with smaller sample sizes - for non-Hispanic black women - 0.4 percent for ages 20 years and older vs. 1.1 percent for ages 20-39 years and for Mexican Americans - 0.9 percent for ages 20 years and older vs. 2.1 percent for 40-59 years.

For children and adolescents, "overweight or at risk of overweight " is defined as a BMI value at or above the gender and age specific 85^{th} percentile of BMI^2. For Mexican American girls 12-19 years if BMI is plotted against the sample weight, there is one sample person with a final exam weight that is much larger than those of any other in this subdomain. This individual is classified as being overweight or at risk of overweight. The difference between the estimated percent of overweight or at risk of overweight for this subdomain when the influential observation is included and when it is excluded is 4 percent.

The final example addresses a prevalence based on a discrete variable – cigarette-smoking status. During NHANES 1999-2000 sample persons were asked "Have you ever smoked at least 100 cigarettes in your lifetime?". Those who answered yes were then asked "Do you now smoke cigarettes?" . On the basis of these two questions sample persons can be classified into three groups. Those who answered " no" to the first question were classified as non-smokers. Those who responded " yes" to the first but " no" to the second were classified as past smokers and those who responded "yes" to both questions were classified as current smokers.

Outliers for this type of variable can be identified heuristically through a box plot obtained using SAS PROC UNIVARIATE. Figure 3 illustrates this type of plot for Mexican American women 40-59 years. One influential sample weight is identified for non- smokers. An influential sample weight can also be identified for non-Hispanic black women 20-39 years who are past smokers.

The prevalence of smoking status with the observation included is compared to the corresponding prevalence with the influential observation excluded for these two subdomains. Including the influential observation for non-Hispanic whites yields a prevalence estimate of 8.2 percent as contrasted with 6.8 when it is excluded – a difference of 1.4 percent. For Mexican American non-smokers the corresponding estimates are 67.0 vs 65.6 - again a difference of 1.4 percent.

5. Discussion:

Examples of influential observations based on data from NHANES 1999-2000 have been presented . They include a mean as a measure of central tendency for a continuous variable together with its corresponding standard error and three examples of prevalences estimates – two prevalences formed by dichotomizing body mass index – a continuous variable – and one prevalence based on a discrete variable.

For the example of a continuous variable - vitamin B12 – one observation was unusually high and had an unusually high sample weight compared to all other observations for Mexican American women 20-39 years. The alternative of presenting measures of central tendency other than the mean which are robust to outliers, namely the median and the mean log transformed, was explored. Other transformations such as the square root could also be used. In applying these transformations statistical properties should be addressed such as consistency and bias of the estimated standard errors and confidence limits of these estimated measures of central tendency.

For the prevalence examples presented in this paper, the greatest impact occurred for Mexican American girls 12-19 years (n=460). An estimate of the prevalence of overweight or at risk of overweight was 43.5 with a standard error of 4.2 percent when the influential observation was included vs. 39.6 with a standard error of 2.3 when it was excluded.

In approaching the issue of outliers the bias variance tradeoff should be addressed. An unusually high or low observation may occur by chance if a sample different from the other possible ones generated from the sampling design is (randomly) selected. Deleting the observation may introduce bias but hopefully decrease variance and thus reduce the width of the confidence intervals (around the estimate). Another possible alternative would be to reweight the data. This option has the same effect: possibly an introduction of bias but hopefully a decrease in variance .

The issue of comparability of estimates (both including and excluding influential observations) with other data sources should also be considered in the decision process and addressed in any published document.

If the discussion is made to publish means or percentages with influential observations deleted, the target population must be clearly defined. For example, in estimating the prevalence of obesity of adults, pregnant women are excluded. Therefore, the estimates are for non-pregnant adults rather than for all adults. For vitamin B12 the Mexican American woman with an unusually high vitamin B12 value had been taking injections for vitamin B12. The estimated mean vitamin B12 with this individual deleted would be for Mexican American women ages 20-39 years who were not taking vitamin B12 injections.

In instances where the prevalence of an adverse condition such as obesity of adults is to be estimated it is important to consider the implications of deleting influential observations (when they occur) from a public health point of view. If influential observations are to be deleted, it is possible that the number of individuals with the condition and who therefore need treatment for it may be underestimated. Thus, insufficient provisions for treatment of such individuals may result. One alternative might be to publish both estimates – with the influential observation included and with it excluded – together with the corresponding standard errors.

Because the NHANES 1999-2000 survey is based on a relatively small sample size, influential observations based on data from this survey are more likely to impact estimated means and percentages as well as their standard errors. Exploration of influential observations and approaches to addressing them should be undertaken in the context of subject matter considerations.

The authors wish to thank Jeffery Hughes, Orkand Corportation, Falls Church, Virginia for his assistance in producing the graphs for this paper.

1/ WHO Expert Committee Physical Status The Use, Interpretation of Anthropometry. Report of a WHO Expert Committee (WHO Technical Report Series 854) Geneva The Organization, 1995.

2/ Kuczmarski RJ, Ogden, CL, Guo, SS et al. 2000 CDC growth charts for the United States: Methods and development. Vital Health Stat 11(246): National Center for Health Statistics, 2002.

Table 1. Sample persons with the top 10 mec examined weights
THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY, 1999-2000

Unadjusted base weight	ADJ1[1]	ADJ2[2]	NON-RESPONSE at the level: Screener	Interview	Exam	Total Non-response[3]	POST STRATIFICATION at the level: Screener	Interview	Exam	Total Post-stratification[4]	MEC examined weight
89404.1	2	1	1	1.364	1.129	1.54	1.03	0.98	1.01	1.02	261361.3[5]
89404.1	2	1	1	1.364	1.108	1.51	1.03	0.98	1.01	1.02	261361.3[5]
89404.1	1.59	1	1	1.364	1.129	1.54	1.03	0.98	1.01	1.02	225470.5
89404.1	1.59	1	1	1.364	1.124	1.53	1.03	0.98	1.01	1.02	224546.8
89404.1	1.59	1	1	1.364	1.118	1.52	1.03	0.98	1.01	1.02	223395.6
89404.1	2	1	1	1.18	1	1.18	1.03	0.98	1.01	1.02	217056
40004.1	2	1.86	1	1.306	1.032	1.35	1.06	1.03	1	1.09	212358.5
89404.1	1.54	1	1	1.364	1.051	1.43	1.03	0.98	1.01	1.02	202639.5
69101.485	2	1	1	1.22	1.089	1.33	1.07	1	1	1.07	196501.9
69101.485	2	1	1	1.22	1.089	1.33	1.07	1	1	1.07	196501.9

1/ Subsampling factor for the stand
2/ Dwelling unit deselection factor
3/=screener*interview*exam non-response
4/=screener*interview*exam post-stratification
5/ Trimmed weights

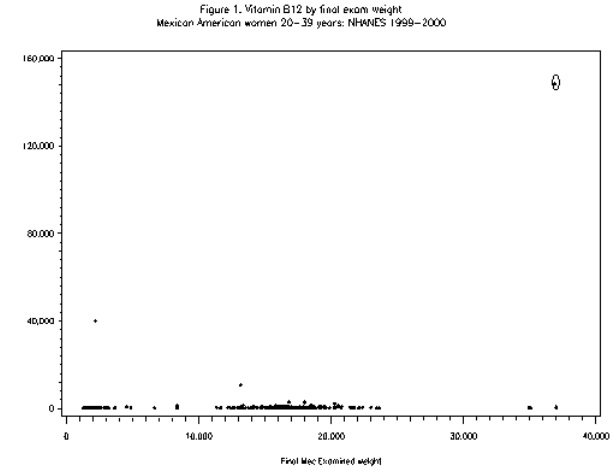

Table 2a. Mean vitamin B12 of women in selected subdomains based on data from the Natiional Health and Nutrition Examination Survey, 1999-2000

	Outlier included			***Outlier deleted***		
	Mean	Sem[1/]	RSE[2/]	Mean	Sem[1/]	RSE[2/]
20-39 years	661	156	24	510	27.5	5.3
Mexican Americans						
20 years and older	2040	1087	53	667	38.1	5.7
20-39 years	2550	2028	76	676	60.1	8.9

Table 2b. Median vitamin B12 and corresponding 95% confidence limits of women in selected subdomains based on data from the National Health and Nutrition Examination Survey, 1999-2000

	Outlier included			*Outlier deleted*		
	Median	95 % Confidence limits		Median	95 % Confidence limits	
		lower	upper		lower	upper
20-39 years	441	425	458	441	425	457
Mexican American						
20 years and older	513	481	545	511	476	545
20-39 years	510	447	572	505	439	570

Table 3c. Vitamin B12 log transformed of women in selected subdomans of the National Health and Nutrtion Examination Survey, 1999-2000

	Outlier included			*Outlier deleted*		
	Mean	Sem[1/]	RSE[2/]	Mean	Sem[1/]	RSE[2/]
20-39 years	436	9.2	2.1	434	8.8	2.0
Mexican American						
20 years and older	554	25.8	4.7	529	15.4	2.9
20-39 years	553	45.0	8.1	518	21.0	4.0

1/ Standard error of the mean
2/ Relative standard error=(sem/mean)*100

Figure 2. Body mass index vs final MEC examined weight
non-Hispanic black women 20-39 years
NHANES 1999-2000

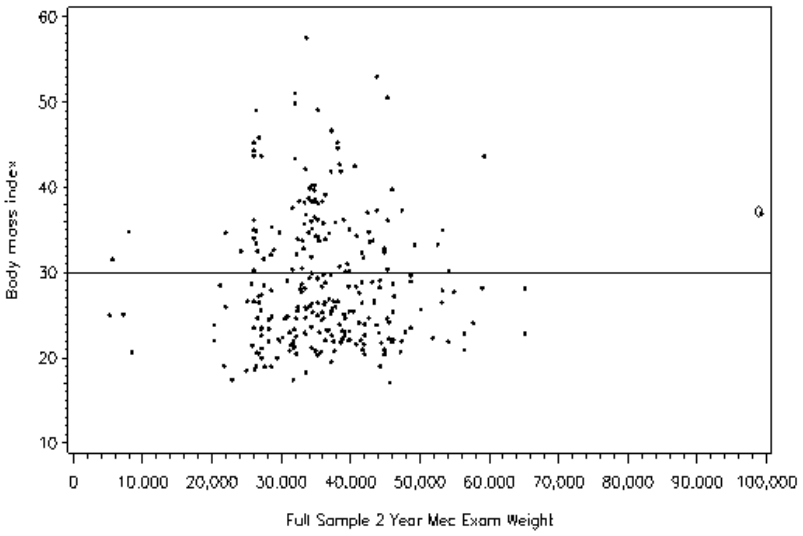

Figure 3. Smoking status of Mexican American
women 40-59 years: NHANES 1999-2000

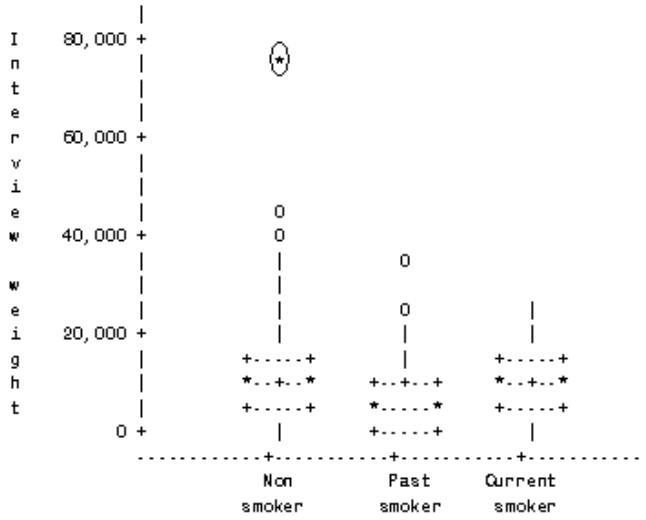

Table 3. Prevalence of obesity of women in selected subdomains of the National Health and Nutrition Examiation Survey, 1999-2000

	Outlier included			*Outlier deleted*			Δ^1
	%[2]	**SE %[3]**	**RSE[4]**	**%[2]**	**SE %[3]**	**RSE[4]**	
Non-Hispanic black							
20 years and older	49.7	2.74	5.5	49.3	2.74	5.6	0.4
20-39 years	45.8	4.08	8.9	44.7	4.04	9.0	1.1
Mexican American							
20 years and older	37.9	4.11	10.8	37.0	4.03	10.9	0.9
40-59 years	48.5	5.29	10.9	46.4	4.86	10.5	2.1

1/Absolute value in the change in the estimated prevalence as a result of excluding the outlier.
2/ Estimated prevalence of obesity
3/ Standard error of the prevalence
4/ Relative standard error=(SE %/%)*100

Table 4. Prevalence of smoking status for selected subgroups of the National Health and Nutrition Examination Survey, 1999-2000

Women	*OUTLIER INCLUDED*			*OUTLIER DELETED*			$\Delta^{1/}$
	Percent	SEP[2/]	RSE[3/]	Percent	SEP[2/]	RSE[3/]	
Non-Hispanic black 20-39 years							
Non-smokers	71.8	3.5	4.8	72.8	3.4	4.6	1.0
Past smokers	8.3	2.3	27.7	6.8	1.8	26.3	1.5
Current smokers	20.0	2.6	13	20.3	2.7	13.1	0.3
Mexican American 40-59 years							
Non-smokers	67.0	4.9	7.3	65.6	4.8	7.2	1.4
Past smokers	15.5	3.2	20.6	16.2	3.2	18.8	3.3
Current smokers	17.5	3.6	20.6	18.2	3.6	20.0	0.7

1/ Absolute value of the change in the estimate as a result of excluding the outlier
2/ Standard error in the percent
3/ Relative standard error