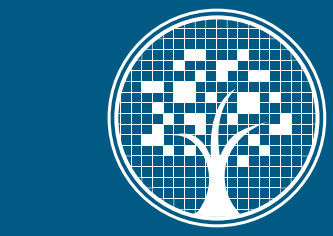

HCUP Methods Series

Inferences with HCUP State Databases Final Report

Report # 2010-05

U.S. Department of Health and Human Services
Agency for Healthcare Research and Quality

Contact Information:
Healthcare Cost and Utilization Project (HCUP)
Agency for Healthcare Research and Quality
540 Gaither Road
Rockville, MD 20850
http://www.hcup-us.ahrq.gov

For Technical Assistance with HCUP Products:

Email: hcup@ahrq.gov

or

Phone: 1-866-290-HCUP

Recommended Citation: Houchens, R. *Inferences with HCUP State Databases Final Report*. HCUP Methods Series Report # 2010-05. Online October 12, 2010. U.S. Agency for Healthcare Research and Quality. Available: http://www.hcup-us.ahrq.gov/reports/methods.jsp.

TABLE OF CONTENTS

INTRODUCTION

Purpose

This report addresses a recurring question from users and journal reviewers concerning whether it is appropriate to calculate standard errors, conduct statistical hypothesis tests, or produce confidence intervals for statistics calculated from HCUP State data sources, including the State Inpatient Databases (SID), the State Emergency Department Databases (SEDD), and the State Ambulatory Surgery Databases (SASD).

The issue

Each HCUP State database contains all or nearly all visits. For example, the 2008 California SID contains a record for every inpatient discharged during 2008 from every California hospital (community, non-rehabilitation, short-term hospitals). While any one record could contain erroneous information—a wrong diagnosis, for instance—the database contains all hospital discharges for 2008. It is a census of discharges, not a sample of discharges. As a result, the user can calculate, for example, the average length of stay (ALOS) for the population of inpatients discharged during 2008. There is no sampling error associated with this calculated ALOS. It is the actual ALOS with 100 percent certainty. Likewise, the analyst can calculate with certainty other statistics, such as the in-hospital mortality rate among discharges from California hospitals in 2008.

However, this view of the database as a census of discharges depends on the inferences the analyst wants to make.

If the user wants to make the following type of statement:

> Statement # 1: The ALOS for uncomplicated diabetes was 2.9 days for males and 3.7 days for females discharged from California hospitals in 2008.

Then it would not be appropriate to calculate standard errors or estimate confidence intervals for these values of ALOS. Assuming no data errors, the statement is true, accurate, and indisputable. The actual length of stay and diagnosis is known for every male and female discharged from every California hospital in 2008. No other estimates are possible because the calculation is based the on all of the stays for male and female diabetes patients. It is also possible to calculate other population statistics without error. For example, one could calculate the standard deviation of length of stay, which would be the true population standard deviation, also calculated without sampling error *for the exact set of patients discharged from California hospitals in 2008*.

However, if the user wants to make a statement about anything other than what actually occurred—anything other than the known history—then this census view has to be reconsidered. Consider the following statement:

> Statement # 2: On average, for uncomplicated diabetes patients in California hospitals, females can expect to stay in the hospital 0.8 days longer than males do.

While the difference of 0.8 days is based on statement # 1 (0.8 = 3.7 – 2.9), statement # 2 is not about what happened in 2008. It is about what to expect now or in the future for similar patients

treated in California hospitals. In this case, the 2008 calculations for ALOS are used to make inferences about another "infinite" population of inpatients, not the specific "finite" population of inpatients that was observed in 2008.

Consider another scenario, comparing male and female patients with a rare disease. Males had 10 complications among 100 patients at risk for a complication rate of 10 %. Females had 11 complications from 100 patients at risk for a complication rate of 11 %. These were the true complication rates for males and females during 2008. There is no sampling error. Does the analyst conclude that females had a higher complication rate than males in 2008? One could state that conclusion unequivocally. The female complication rate was higher than that for males.

However, is that really the statement the analyst wants to make? Perhaps the desired statement is that the underlying *predisposition* for complications differs between the genders. If so, then the user wants to generalize beyond the specific set of patients treated at the specific times by the specific clinicians rendering treatment.

We discuss this issue in more detail in the next section.

FINITE VERSUS INFINITE POPULATIONS

This section starts by explaining some basic statistical concepts concerning estimation that explain why it is inappropriate to calculate standard errors based on data from the entire population of interest.

Statistical concepts

To begin, analysts are interested in some quantities associated with a population. These quantities are called "population parameters." For example, the population mean and the standard deviation are population parameters, often denoted by the Greek letters μ and σ, respectively. Often, the user has information about these quantities only from a sample of individuals in the population, not the entire population (census). Analysts employ the sample information to estimate the population parameters. For example, the population ALOS might be estimated based on the ALOS calculated from the lengths of stay for a random sample of discharges. This sample estimate is labeled a "point estimate." The sample ALOS is a point estimate of the population ALOS. The sample mean and standard deviation are usually denoted by $\bar{x}$ and *s*, respectively.

For a given sample, the sample mean is just one point estimate out of many point estimates that would arise from different random samples. Therefore, the sample mean is a "random variable" that has a "sampling distribution" describing the distribution of sample means that would occur if we calculated sample means from many different random samples of the same size. The standard deviation of this (theoretical) distribution of means is called the "standard error" of the mean. For simple random sampling from an *infinitely large population*, the formula for the standard error of the mean from a sample of size *n* is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Clearly, the sample mean tends to become a more precise estimate of the population mean—the standard error of the mean gets smaller—as the sample size *n* gets larger. To estimate the

standard error from the sample data, one substitutes the sample estimate of the standard deviation *s* for σ in the equation above.

If the population is finite, containing *N* elements, then the formula for the standard error of the mean from a sample of size *n* is slightly different:

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}} = f \times \frac{\sigma}{\sqrt{n}}$$

The factor *f* is related to a quantity called the finite population correction factor. The standard error of the mean approaches zero as the sample size *n* approaches the population size *N*. In particular, when the sample size *n* equals the population size *N*, the sample becomes a census and the standard error of the mean is equal to zero. This is the situation with the HCUP State databases when users are interested in population parameters like the mean and standard deviation of ALOS in the 2008 California SID. The "sample" is the population. Therefore, the standard error of the mean is zero. The population mean is equal to the "sample" mean.

Is it the entire population?

As explained above, classical sampling theory is concerned with inference about finite population parameters based on a sample from that finite population. If the sample size is less than the population size then different samples are possible and the standard error of a sample estimate describes the range of error that is expected in estimating the finite population statistic of interest. If the data comprise the entire finite population then the standard error is zero.

However, often analysts are not interested in finite population parameters. For example, researchers might be interested in the relationship between age and in-hospital mortality rate for patients who underwent percutaneous transluminal coronary angioplasty (*PTCA) in California hospitals during 2008*. One could plot the observed mortality rate as a function of age and that would represent the observed relationship between age and mortality for all PTCA patients in California during 2008. There might be a sharp increase or decrease in the mortality rate between two consecutive years of age. That is acceptable because it describes what actually occurred in the population of California PTCA patients in 2008.

However, if there is interest in the *general relationship* between age and in-hospital mortality, and not in *the specific relationship* between age and in-hospital mortality for the 2008 PTCA patients, then the observed relationship should be considered a sample that is representative of a larger population of PTCA patients.

This is the simplest possible example of "superpopulation inference" (Korn and Graubard, 1999), where the analyst considers the 2008 PTCA patients to be a simple random sample from an infinite superpopulation of interest. In this case, the superpopulation is the collection of all *potential* 2008 PTCA patients in California, and the population model is the relationship between age and in-hospital mortality among *potential* PTCA patients. In that case, researchers might be less willing to accept a sharp increase or decrease in the mortality rate between two consecutive years of age because it might seem implausible for a general relationship between age and mortality. This view is consistent with the view that in-hospital PTCA mortality is subject to random factors. The outcome for a given patient might have been different if the procedure had been done at another time, at another hospital, or by another surgeon.

Whether researchers should calculate standard errors, perform hypothesis tests, or calculate confidence intervals depends on whether they consider the HCUP State database to be 1) the

entire population of interest or 2) a representative sample of the population of interest. The answer is not always easy.

The State database is unquestionably the entire population of interest when interest centers on population parameters germane only to the specific patients treated in the specific hospitals by the specific clinicians at the specific times of treatment represented by the units in the database. All of this "specificity" is not meant to discourage the analyst from considering the database as a census of the population. Rather, it is meant to emphasize the specificity of the population for which the statistics are calculated.

On the other hand, the State database is unquestionably a sample of the population when inferences go beyond the database. For example, the in-hospital mortality rates from the 2008 California SID are sample estimates if they are used to estimate 2009 in-hospital mortality rates in California. Likewise, if investigators publicly post 2008 outcome parameters to help prospective patients choose among alternative treatments in 2009, then it would be appropriate to calculate confidence intervals or standard errors if the prospective patients are assuming that the information is predictive of those parameters for 2009 or 2010. In that case, the 2008 information is a "sample" representing expectations for 2009 or 2010. Conversely, if the treatment statistics are posted as "historical information" meant only to inform prospective patients concerning things like the number of CABGs performed in 2008, then these should be regarded as population parameters, not sample statistics.

The situation described in the introduction concerning gender differences is more subtle. Males had a 10 % complication rate and females had an 11 % complication rate in 2008. We can state plainly that the female complication rate was higher than the male complication rate in 2008. However, the desired inferences probably go beyond the historical data. The researcher probably wants to say something about the comparative *underlying risk* of a complication for a random patient at risk of a complication. We might describe the underlying risk by phrases such as "the patient's long-run complication rate" or "the patient's predisposition for a complication" or the "patient's quality of care." All of these concepts are rooted in describing the complication rate that would have occurred in a very large sample of patients at risk. In that case, researchers are making inferences beyond the historical data. They are making inferences about a large hypothetical population of patients, and they should consider the State database to be a sample.

Examples

Regression of trends

Suppose an analyst calculates the ALOS for diabetes patients for every year from 2001 to 2008 using the SID data for one state. The investigator wants to estimate the 2009 ALOS for diabetes patients based on a regression using one *observation for each of the eight years with ALOS as the dependent variable and year as the pre*dictor variable. Should an estimate of the standard error for each year's ALOS be incorporated into the regression?

The researcher has to ask whether he or she is making inferences beyond the historical data. The answer is yes. Therefore, an estimate of the standard error should be incorporated into the regression.

Suppose the user was only interested in the historical trend. The analyst could plot the historical trend from 2001 to 2008 and assert that this represents the historical trend, without error. However, that might not be the trend of interest because that historical trend is for a different finite set of patients in each year. The trend is particular to those specific patients

discharged in each year. On the other hand, one might think of the patients each year as a sample from a "superpopulation" of patients with diabetes who might have been treated each year (this is essentially what one does when forecasting the 2009 ALOS from the 2001 to 2008 ALOS). In that case, conceptually the researcher is making inferences about an infinite population of potential patients and the observed patients represent a sample of potential patients.

Hypothesis tests

In the regression example, suppose the analyst wanted to test the (null) hypothesis that the 2007 ALOS equals the 2008 ALOS versus the (alternative) hypothesis that they differ, again based on the SID data from one State. If the finite population point of view is taken, then it is appropriate to calculate the population ALOS for each year and compares them. No statistical testing procedure is required. However, each ALOS is specific to the year for which it was calculated. The analyst would say that the ALOS for the *patients treated in 2007* is different from the ALOS for the *patients treated in 2008*.

If the point of the comparison is to say something about the difference in ALOS *for similar populations of patients*, then the analyst should calculate a standard error and perform a statistical significance test because he or she is assuming that patients from the two years are similar. Assuming that the patients are "similar" is another way of saying that they are representative of a larger, common, infinite population of potential patients, not the specific patients who were actually treated.

Point estimates of means or rates

Often, researchers are interested in a mean or a rate. Suppose an analyst is interested in the average charge and mortality rate for patients hospitalized for AMI in each California hospital during 2008. The analyst can calculate hospital-specific mean charges and mortality rates for AMI discharges using the entire 2008 California SID. A histogram can be constructed showing the range of average hospital charges and the range of hospital mortality rates for AMI. Some hospitals will have had a small number of AMI patients and other hospitals will have had a large number of AMI patients. That does not matter from the finite population viewpoint because the investigator only cares about the specific outcomes for patients hospitalized in 2008. However, if there is concern about the differences in observed mortality rates attributable to differences among the number of AMI patients across hospitals, then the finite population viewpoint is inappropriate. The researcher is thinking of the AMI patients as a "sample" of AMI patients who might have been hospitalized, and it would be appropriate to calculate standard errors for the estimates of average charges and mortality rates.

ANNOTATED REFERENCES

Brousseau, D.C., Owens, P.L., Mosso, A.L., Panepinto, J.A., and Steiner, C.A. (2010). Acute care utilization and rehospitalizations for sickle cell disease. *Journal of the American Medical Association* 303, 1288-1294.

> The authors use SID data from eight states to estimate utilization rates for sickle cell patients. The authors are interested in generalizing beyond the historical data in the eight states. Consequently, they calculate standard errors and estimate confidence intervals for the statistics they report.

Elliott, M.N., Zaslavsky, A.M., and Cleary, P.D. (2006). Are finite population corrections appropriate when profiling institutions? *Health Services and Outcomes Research Methodology* 6, 153-156.

This short article is highly recommended for researchers who consider adopting the finite population viewpoint. The authors argue that it is not appropriate to use finite sample theory for estimating hospital-specific parameters in order to compare quality among hospitals (profiling). Their discussion includes several citations for works by prominent statisticians who argue that the finite population sampling model is seldom defensible.

Gray, C.G., Knoke, J.D., Berg, S.W., Wignall, F.S., and Barrett-Connor, E. (1998). Counterpoint: Responding to suppositions and misunderstandings. *American Journal of Epidemiology* 148, 328-333.

This is a response to charges that the authors did not apply finite population correction factors in constructing confidence intervals. They argue that contemporary epidemiologic investigators frequently adopt the superpopulation model, the model they adopt as most appropriate for their analysis and the reason that finite population correction factors are inappropriate.

Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys.* New York: John Wiley & Sons, Inc.

This book is mostly concerned with the analysis of health survey data. However, section 5.7 entitled “Variance Estimation for Superpopulation Inference” has a nice introductory discussion about superpopulation models. It should be noted that the superpopulation model is most often associated with probability samples of various designs and the objective is to account for the sample design in estimating superpopulation parameters. However, researchers can generally consider the state data to be a simple random sample of the superpopulation, and standard statistical procedures for simple random samples from an infinite population apply.