

Statistical Comparisons of Disease Prevalence Rates Using the Bootstrap Procedure

Matthew Dutton, Florida A&M University, Charlotte Baker, Florida A&M University

ABSTRACT

Disease prevalence is one of the most basic measures of the burden of disease in the field of epidemiology. As an estimate of the total number of cases of disease in a given population, prevalence is a standard in public health analysis. The prevalence of diseases in a given area is also frequently at the core of governmental policy decisions, charitable organization funding initiatives, and countless other aspects of everyday life. However, all too often, prevalence estimates are restricted to descriptive estimates of population characteristics when they could have a much wider application through the use of inferential statistics. As an estimate based on a sample from a population, disease prevalence can vary based on random fluctuations in that sample rather than true differences in the population characteristic. Statistical inference uses a known distribution of this sampling variation to perform hypothesis tests, calculate confidence intervals, and perform other advanced statistical methods. However, there is no agreed-upon sampling distribution of the prevalence estimate. In cases where the sampling distribution of an estimate is unknown, statisticians frequently rely on the bootstrap re-sampling procedure first given by Efron in 1979. This procedure relies on the computational power of software to generate repeated pseudo-samples similar in structure to an original, real data set. These multiple samples allow for the construction of confidence intervals and statistical tests to make statistical determinations and comparisons using the estimated prevalence. In this paper, we use the bootstrapping capabilities of SAS[®] 9.4 to compare statistically the difference between two given prevalence rates. We create a bootstrap analog to the two-sample t test to compare prevalence rates from two states despite the fact that the sampling distribution of these estimates is unknown using SAS[®].

INTRODUCTION

Traditional statistical inference is frequently limited by the underlying distributional assumptions of our tests and procedures. In instances where the underlying distributions are unknown, or the necessary distributional assumptions are not met, the bootstrap procedure originally given by Efron (1979) can be a valuable tool for statistical comparisons and confidence intervals. This paper focuses on the specific example of the comparison of disease prevalence rates. These rates are often based on sample data, yet the distributional theory needed for their formal statistical comparison is currently limited. Applications will focus on the use of PROC SURVEYSELECT to generate bootstrap samples to allow for the statistical comparison of two prevalence estimates.

Readers will gain a basic understanding of the bootstrap procedure, its application to prevalence rates, and the completion of bootstrap analysis using PROC SURVEYSELECT. It is assumed that the reader will have some familiarity the concepts of statistical inference and disease prevalence and a moderate skill level using SAS[®], but the paper is meant to be introductory in nature. SAS[®] 9.4 is used in all applications.

STATISTICAL INFERENCE

Traditional statistical inference is the process by which sample statistics are used to make some determination about population parameters and characteristics. This process is generally dependent upon the selection of a specific statistic, the sampling distribution of that statistic, and some statistical procedure built around that sampling distribution. The sampling distribution of a statistic can be thought of as the distribution of the statistic when calculated for all possible samples of a given size from the same population.

Figure 1 is an illustration of the sampling distribution of a statistic. To demonstrate this concept, imagine a population of 100 individuals. From this population, take a random sample of 10 individuals and

calculate the desired statistic. If those 10 individuals were placed back into the population and a new sample of ten was randomly selected,

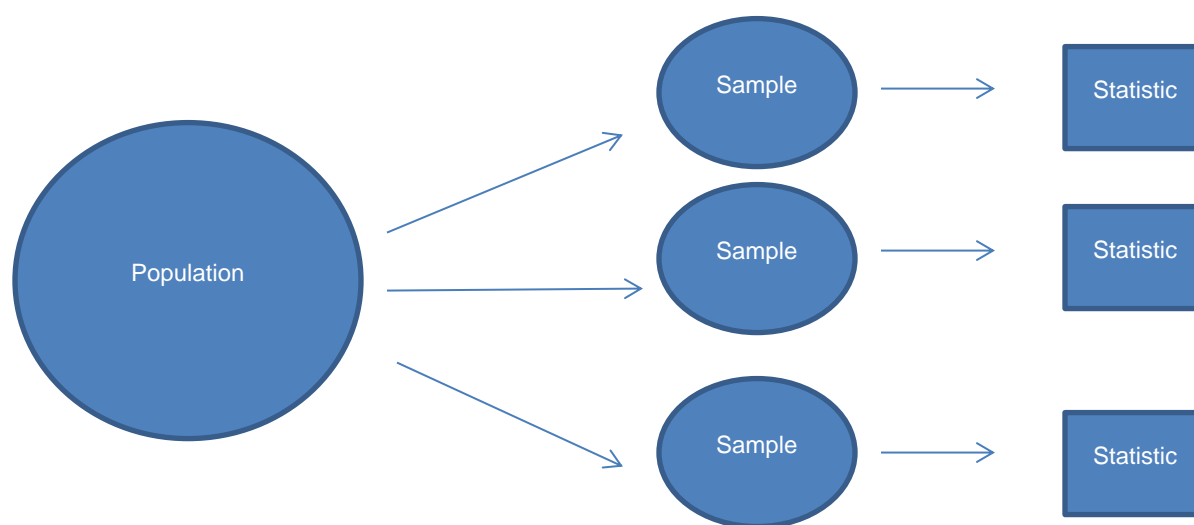


Figure 1. Given a population of individuals take a random sample of individuals and calculate the desired statistic. If those individuals were placed back into the population and a new sample of ten was randomly selected, a new value of the statistic could be calculated. If all possible samples of the same size were taken from the population, then all possible values of the statistic could be calculated and the entire sampling distribution would be given. Knowing the exact distribution of all possible values of the statistic would give all of the necessary information to develop hypothesis tests, confidence intervals, and the other traditional components of statistical inference.

The most familiar sampling distribution is that of the sample mean, \bar{x} . In this instance, the Central Limit Theorem states that assuming a large enough sample size, \bar{x} has a normal distribution with a mean equal to the population mean and a standard deviation given by $\frac{\sigma}{\sqrt{n}}$, the population standard deviation divided by the square root of the sample size. In instances when the Central Limit Theorem applies, we can assume a normal distribution when conducting our inference and bypass the process of generating all possible values of the statistic. When we are dealing with the sample mean, and the Central Limit Theorem applies, statistical inference is a basic procedure in statistical analysis.

THE BOOTSTRAP PROCEDURE IN SAS®

In situations where the Central Limit Theorem does not apply, statistical inference can be more difficult. If we are working with a statistic other than the mean, or our sample size is too small, we cannot depend on the assumption of normality to complete our statistical inference. It is these situations, when the sampling distribution of our statistic is unknown, that the bootstrap method is most valuable.

In general, statistical bootstrapping can refer to any procedure that requires random sampling with replacement from a given data set. Bootstrapping can be used to estimate statistical bias, give hypothesis tests, and create confidence intervals when the traditional statistical inference is not possible. Rather than resampling from the study population, which is most often unknown and unmeasurable, the bootstrap procedure approximates the sampling distribution of a statistic by resampling from the data set itself. In order to complete a statistical bootstrap, the user must sample, with replacement, from the original data set to create a bootstrap sample with the same number of observations as the original set. If the user repeats this process a large number of times, the bootstrap samples will act as surrogates for the repeated samples from the population illustrated in Figure 1. We can then calculate the statistic of interest for each of these bootstrap samples. The distribution of the statistics calculated for each of the bootstrap samples will be an approximation to the true sampling distribution of the statistic based on the population.

Figure 2 demonstrates this approximation in relation to the sampling distribution.

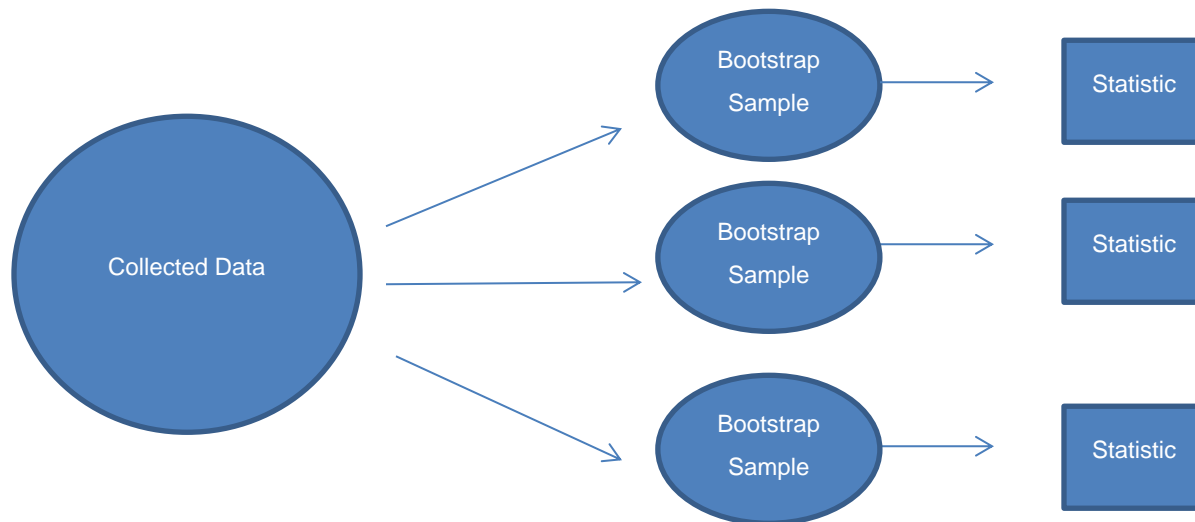


Figure 2. Bootstrapping approximates the sampling distribution of a statistic by sampling, with replacement, from the original data set. If this process is repeated a sufficiently large number of times, the resulting distribution of bootstrap statistics can be used as a surrogate for the true sampling distribution of the statistic.

Although there are several macros, and other procedures, for conducting random resampling in SAS®, the bootstrap can be easily accomplished using PROC SURVEYSELECT. Using this PROC, often used to generate a sample from a known sampling frame, we can select a sample, with replacement, from our original data set. When this process is repeated a large number of times, we will produce the bootstrap replications needed to approximate the sampling distribution of any statistic.

The necessary SAS® code to complete a bootstrap resampling procedure using PROC SURVEYSELECT is as follows:

```
proc surveyselect data=Original out=Bootstrap method=urs samprate=1 outhits
rep=1000;
run;
quit;
```

Assuming the working library has the initial data set, named “Original”, the given procedure will result in a new data set, called “Bootstrap”, which will contain all the necessary information for calculation of the sampling distribution of any statistic. By setting the method to “urs” or “Unrestricted Random Sampling” we allow for sampling with replacement from the initial data set. When we use the option “samprate=1” we make sure that each bootstrap sample data set contains the same number of observations as the input data set. We allow for some large number of replications, here 1000 (this number need only be sufficiently large and may change given each situation), to produce a large number of bootstrap samples. It is these bootstrap samples that will allow us to calculate repeated estimates of our statistic and create our sampling distribution.

If the data set “Original” contains n observations, then the data set “Bootstrap” will contain $n \times (\text{number of replications})$ observations consisting of the resampled data observations identified by their replication number in the variable “replicate”. Any statistical calculation can then be performed on the “Bootstrap” data set specifying “by replicate” in order to obtain the bootstrap distribution of our statistic.

USING THE BOOTSTRAP TO COMPARE DISEASE PREVALENCE

Keep in mind that prevalence is defined as the number of cases of a disease that are present in a particular population at a given time. We can calculate the estimated prevalence of any disease by collecting an unbiased representative sample from a population, counting the frequency of occurrence of our disease, and multiplying the relative frequency of disease in our data by the size of our total population.

The following illustration will use PROC SURVEYSELECT and PROC UNIVARIATE to demonstrate how to compare two diabetes prevalence estimates. This simplified example is based on the minimum amount of information needed to calculate the prevalence for two different states. Real data will likely be more complex than this, but for the purposes of illustration we stick to this simple approach. The data set will contain the diabetes status of twenty patients, ten from the state of Florida and ten from Georgia. The DATA step is as follows:

```
data Original;
input state $ diabetes;
datalines;
  Florida 1
  Florida 1
  Florida 0
  Florida 0
  Florida 0
  Florida 1
  Florida 0
  Florida 0
  Florida 1
  Florida 0
  Georgia 0
  Georgia 1
  Georgia 1
  Georgia 0
  Georgia 0
  Georgia 1
  Georgia 0
  Georgia 0
  Georgia 0
  Georgia 0
  ;
run;
```

We can use PROC UNIVARIATE to calculate the disease prevalence of each of these small samples by state. The code to complete this procedure is:

```
proc univariate data=Original;
class state;
var diabetes;
by state;
run;
quit;
```

Using the PROC UNIVARIATE code we can easily see that 40% of the Florida group and 30% of the Georgia group has diabetes. We could then use overall population estimates for each state to estimate how many people within each state have diabetes. However, what if we want to compare the estimates between the two states? We only have two numbers and no distributional theory that will allow us to perform statistical analysis. It is here that the bootstrap procedure is applied. The code for the bootstrap in this example is given by:

```
proc surveyselect data=Original out=Boot seed=5454 method=urs samprate=1
  outhits rep=1000;
  strata state;
run;
proc sort data=Boot;
  by replicate state;
run;
```

By including the STRATA statement, we have ensured that each of the bootstrap samples will have ten resampled observations from each state, just like the original data set. We can then use a PROC SORT to sort by replicate and state for our future calculations. We now need to calculate the prevalence estimate for each state across all the replications using the statement given below:

```
proc univariate data=Boot;
  var diabetes;
  by replicate state;
  output out=Results sum=Prevalence;
run;
```

We now have 1000 estimates for the prevalence rate of disease for each of the two states as an approximation of the sampling distributions of these statistics. We can perform all of our customary statistical inference procedures based on these sampling distributions. We could find the average prevalence estimate for each state, 95% confidence intervals by state, or other distribution measures for our estimates. If we want to test whether or not these two states' distributions are statistically equal, one method would be to find the difference between the Florida estimate and the Georgia estimate by replicate and determine whether or not zero is within the central 95% of the distribution of these differences. If it is, then the two distributions can be considered statistically equal. If it is not, then we can conclude that the two estimates are statistically different. This process can be completed using PROC UNIVARIATE:

```
proc univariate data=prevalence;
  var difference;
run;
```

A histogram of the differences between the prevalence estimates is given in Figure 3. We can see that our bootstrap procedure indicates that the two prevalences are likely similar.

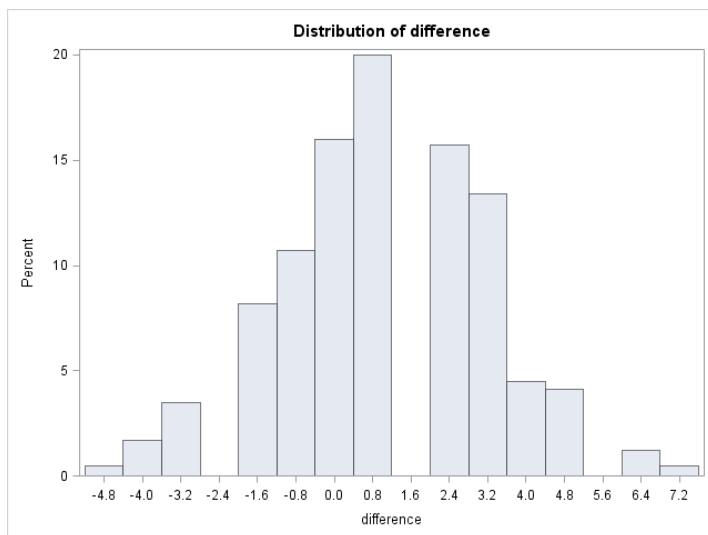


Figure 3. The above histogram represents the distribution of the differences between the Florida prevalence estimate and the Georgia prevalence estimate in our bootstrap procedure. Since the value zero is within the

center 95% of this distribution, we can conclude that the two estimates are statistically equivalent.

CONCLUSION

Bootstrapping is a useful tool when we do not know the sampling distribution of a statistic we have calculated. It can be applied to most of the customary inference techniques we use in statistical analysis when we are trying to make conclusions about an unstudied population. Given that prevalence estimates are frequently based on samples of data rather than populations, and that we typically do not have a sampling distribution for prevalence estimates, the bootstrap can be helpful when we are trying to compare prevalence between two populations. The simple example given in the text demonstrates how PROC SURVEYSELECT can generate all the information we need to complete a statistical analysis using bootstrap replications. The process given here can easily be generalized to more complex scenarios and problems.

REFERENCES

Efron, Brad. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics*, Vol 7, No. 1:1-26.

ACKNOWLEDGMENTS

I would like to thank my colleagues Dr. Charlotte Baker and Dr. C. Perry Brown for their help, guidance, and support in the development of this paper.

CONTACT INFORMATION

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

Matthew Dutton
Florida A&M University
matthew.dutton@famu.edu

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 trademarks of their respective companies.