Paper SA-13

Bootstrap power analysis using SAS ®

Doug Thompson, Thompson Research Consulting LLC, Chicago, IL

Nort Holschuh, Bell Institute of Health and Nutrition, General Mills Inc., Minneapolis, MN

Bruce Barton, University of Massachusetts Medical School, Worcester, MA

Ann Albertson, Bell Institute of Health and Nutrition, General Mills Inc., Minneapolis, MN

Abstract

Secondary data analyses are frequently done to address questions in epidemiology and public health. Often investigators conduct such studies without first determining whether the sample size is adequate to test the study hypotheses. As a case study, this paper uses an analysis that examined the association between whole grain intake and body mass index (BMI) utilizing data from the National Health and Nutrition Examination Survey (NHANES), published in Journal of Pediatrics. We show how power analysis for a study like this could be conducted using bootstrapping in SAS. The target audience is biostatisticians, epidemiologists and other investigators seeking to address research questions using secondary data analyses.

Introduction

Epidemiology and public health studies often utilize available datasets that were not designed to test the study hypotheses. Statistical tests are used to test the significance of group differences observed in the data. For example, a study might use data from the National Health and Nutrition Examination Survey (NHANES) or the National Longitudinal Survey of Youth (NLSY), to test whether body mass index differs significantly between groups of children with higher vs. lower whole grain intake, or test whether lipid levels differ significantly between individuals who frequently eat fast food vs. those who do not.

If the dataset used in the analysis was not designed to test the study hypotheses, the sample size may be too small to support credible statistical tests of group differences. At the extremes, it may be obvious that the data are adequate (or inadequate) to support credible statistical tests. If there are, say, 10,000 observations per group, it may be fairly obvious that the sample size is sufficient (assuming a reasonably large effect size, an adequate study design, and so forth). If there are, say, 5 observations per group, it may be fairly obvious that the test of group differences will not be credible (although there are exceptions, e.g., when a very large and consistent group difference is expected). The greatest problem is the gray area between the extremes – what if the sample size per group is 50, 100, or 200? Sometimes these are adequate sample sizes, and sometimes they are not, depending on the study design, sampling approach, test of group difference to be used and other considerations. Unfortunately these gray areas are very common in secondary data analyses that have been reported in the literature.

Power analysis is one approach to estimating the sample size that is needed to support credible statistical tests of group differences in secondary data analyses. Power analysis should be conducted before the secondary data analysis is undertaken. If the data are insufficient to support credible statistical tests, the study should be abandoned, or a dataset with a sufficient sample should be sought.

Available datasets can be utilized in power analyses. Bootstrapping is one approach to power analysis (Efron & Tibshirani, 1998). This assumes that the available data represents the population. One keeps drawing random samples of a fixed size n with replacement from the dataset. The assumed population difference between the groups is simulated within each sample. Power is the proportion of samples where a significant group difference was found (Walters, 2004). Power analysis using bootstrapping is relatively easy to implement in SAS ®.

This paper describes an example of power analysis using bootstrapping and shows how this can be implemented with SAS. The example is in the context of a secondary data analysis. The intention is to show how power analysis can be used to determine whether available data are sufficient support credible statistical tests of group differences. However, the techniques described here are general; they can be used for power analysis in any type of study (e.g., clinical trials and other experimental designs), provided that there is relevant available data to support the bootstrapping approach.

Example of bootstrap power analysis for a secondary data analysis

A publication by Zanovec and co-authors (2010) in Journal of Pediatrics described the association between whole grain intake and measures of adiposity, for example z-scores of body mass index relative to specific age and gender groups ("BMI-for-age"). The study focused on 6-18-year-olds and utilized data from the National Health and Nutrition Examination Surveys (NHANES). NHANES was designed to estimate measures of health and nutrition in the United States, with a specific degree of precision. NHANES data have been used to study a variety of health and nutrition questions. Although NHANES includes measures relevant to a study looking at the relationship between whole grain intake and body weight, NHANES was not specifically designed to test this relationship, raising the question of whether the sample size is sufficient (Thompson & Barton, 2011).

In the study by Zanovec et al, NHANES participants were categorized in 4 groups based on their whole grain intake, as indicated by 24-hour dietary intake recalls (Group A, ≥ 0 to < 0.6 servings of whole grain per day; Group B, ≥ 0.6 to < 1.5 servings/d; Group C, ≥ 1.5 to < 3 servings/d; and Group D, ≥ 3 servings/d). Analyses were conducted separately for 6-12- and 13-18-year-olds. The sample sizes in each group, by age category, were ages 6-12: Group A, n=2,675; Group B, n=712; Group C , n=328; Group D, n=153. Ages 13-18: Group A, n=3426; Group B, n=814; Group C, n=477; Group D, n=214. The groups with the greatest whole grain intakes (Group D) had very small sample sizes (n=153 and 214 for ages 6-12 and 13-18, respectively). Zanovec et al reported mean BMI-for-age and other weight-related measures by whole grain intake group and age category. Their statistical tests showed no BMI-for-age difference between whole grain intake Groups C (≥ 1.5 to < 3 servings/d) and D (≥ 3 servings/d). This is an example of the type of sample size "gray area" that is common in epidemiology and public health research – the sample size seems small (<300 per group), but it is unclear whether or not it is sufficient to support credible statistical tests of group differences.

Zanovec et al did not state whether or not they did power analyses prior to conducting the study. If they had done power analysis and reported the results, the reader would have less doubt regarding the adequacy of the sample size for the statistical tests that were conducted.

In this case, it is possible to illustrate power analysis utilizing the same data as Zanovec et al, because NHANES is public data and can be readily obtained. These data were utilized to conduct a bootstrap-based power analysis in SAS. This illustrates how investigators in future, similar studies can determine the necessary sample size prior to conducting the study.

Methods

In this section, we walk through the steps used to implement the bootstrap power analysis in SAS.

The data were the same NHANES sample data that Zanovec et al used in their analysis, based on our understanding of the sample as it was described in their article in Journal of Pediatrics. Bootstrap-based power analysis consisted of three basic steps: 1) draw 1,000 bootstrap samples of size n (where n can be varied to gauge the impact on power) with replacement from the NHANES sample data; 2) test the Group C (≥ 1.5 to < 3 servings whole grain/day) vs. Group D (≥ 3 servings/d) difference within each of the 1,000 bootstrap samples; and 3) summarize the test results across all of the bootstrap samples -- power is the proportion of bootstrap samples with a significant test result. Separate power analyses were conducted for 6-12- and 13-18-year-olds in the NHANES sample data, because Zanovec et al analyzed these age groups separately.

Power analysis was conducted via a SAS macro. The macro executes the following steps, in sequence.

1. Identify age and whole grain intake groups in the data – do the steps below separately for each age group (which involves a separate macro call for each age group).

Arguments passed to the macro are age group, sample size (n) for each whole grain intake group (n_grp3 is for Zanovec et al's Group C and n_grp4 is for Group D), and seeds for random number generation.

```
%macro powermac_zanovec_n_bmiz(
agegrp=,
n_grp3=,
n_grp4=,
seed_grp3=0,
seed_grp4=0);
```

The outcome variable is BMI-for-age (called bmiz in the data).

```
%let outcome = bmiz;
```

An example macro call:

```
* Macro call for ages 6-12 (agegrp=1 in the data);
%powermac_zanovec_n_bmiz(agegrp=1,
n_grp3=328,
n_grp4=153,
seed_grp3=34797,
seed_grp4=25039);
```

To determine power with different sample sizes, n_grp3 and n_grp4 can be modified. In the results reported below, we assumed that the relative proportions of individuals in each group would stay the same. Therefore, the sample sizes reported by Zanovec et al were multiplied by a constant. For example, we calculated power with sample sizes of 328 and 153 (the sample sizes reported by Zanovec et al), twice this number (i.e., 656 and 306), three times this number (i.e., 984 and 459), and so forth.

2. Create separate datasets for each whole grain group (C and D).

The NHANES sample dataset used in this analysis is named kids9904. This consists of NHANES data from the 1999-2004 waves of data collection, including only data for 6-18-year-olds. Whole grain intake groups were coded in a data step (not shown) prior to the macro call (WGgroupP=3 is Zanovec et al's Group C and WGgroupP=4 is Zanovec et al's Group D).

```
* Limit the dataset to the age group that will be used in;
* the analysis.;
* Split data into Groups C (WGgroupP_3) and D (WGgroupP_4);
* (Group D data are not actually used in the power analysis, for;
* reasons described below).;
data _kids9904 WGgroupP_3 WGgroupP_4;
set kids9904(where=(ageGroup=&agegrp));
if WGgroupP=3 then output WGgroupP_3;
if WGgroupP=4 then output WGgroupP_4;
output _kids9904;
run;
```

3. Count n in Group C. This is needed to keep the "pointer" used to select random observations within the range available in the data.

```
* Count the number of subjects in subset WG Group 3 (Zanovec et als Group C);
ods listing close;
data _null_;
set WGgroupP_3 end=eof;
retain count 0;
count+1;
if eof=1 then call symput('nobs_3',count);
run;
%put sample size in WG group 3 = &nobs_3.;
```

4. Draw 1000 random bootstrap samples of size n1 with replacement from the Group C data in the NHANES sample. (To be clear, we will use n1 to refer to the sample size in Group C and n2 to refer to the sample size in Group D.) This treats the observed Group C data as if it were the population. Repeated samples are drawn from this, with replacement. Note that PROC SURVEYSELECT also provides a convenient syntax for executing this step.

```
* Draw repeated samples of size n_grp3 with replacement from WG Group 3;
data bootsamp_3;
do sampnum = 1 to 1000;
do i = 1 to &n_grp3;
x = round(ranuni(&seed_grp3) * &nobs_3);
set WGgroupP_3
point = x;
output;
end;
end;
stop;
run;
```

5. Draw another 1000 random samples of size n2 with replacement from the Group C data in NHANES, to represent group D. The assumption is that the Group C and Group D data are drawn from exactly the same population, which is here represented by the NHANES Group C data. The sample size for Group D (n2) was passed to the macro in the macro call; it is the macro variable &n_grp4.

```
* Draw repeated samples of size n_grp4 with replacement, to represent WG;
* Group 4 (Group D in Zanovec et al).;
* These are drawn from the NHANES group 3 (i.e., C) data, and an incremental;
* difference is simulated, representing;
* the assumed true population difference between Groups C and D.;
* This simulates what would be observed if group 4 was really like group 3;
* (that is drawn from exactly the;
* same population), except for the specified specified difference in the;
* outcome.;
* Sample size is held constant as that reported by Zanovec et al for Grp D.;
data bootsamp_4;
do sampnum = 1 to 1000;
do i = 1 to &n_grp4;
x = round(ranuni(&seed_grp4) * &nobs_3);
set WGgroupP_3
point = x;
output;
end;
end;
stop;
run;
```

6. Take BMI-for-age from the data generated after Step 5, and shift it down by specified degree d (in this example, we use the same degree of difference in BMI-for-age between Groups C and D as reported by Zanovec et al), simulating the assumed true whole grain effect on BMI-for-age – this assumes that the mean differences reported by Zanovec et al are the true population differences. Then the question becomes, what is the power to detect the group difference if the true group difference is d and sample sizes are n1 and n2? Z-scores are the unit of measurement for BMI-for-age, so differences in z-scores between groups can be expressed in standard deviation units.

   Group C vs. D differences in BMI-for-age are specified for each age group. The Group C vs. D difference in BMI-for-age found by Zanovec et el was 0.38 SD for 6-12-year-olds (0.86 in Group C vs. 0.48 in Group D) and 0.13 (0.68 in Group C vs. 0.55 in Group D) for 13-18-year-olds. For power analysis purposes, the group differences reported by Zanovec et al were assumed to be the true population differences. However, this is not necessary for power analysis. There are other approaches; for example, if 0.5 SD is considered to be a practically important difference, the power analysis could assume this degree of difference.

```
* In WG group 4, similate specific percentile-point reductions in BMI;
* z-scores.;
* Two possible differences are simulated -- reduce BMI z-scores by;
* 0.13, 0.38 SD;
* (z-scores are in standard deviation or SD units).;
* This similates drawing samples from the same population as;
* Group 3 (C), except that;
* BMI z-scores are shifted down to a specified degree.;
data
bootsamp_4_05(drop=sim_outcome_10 sim_outcome_20 sim_outcome_30
rename=(sim_outcome_05=sim_outcome))
bootsamp_4_10(drop=sim_outcome_05 sim_outcome_20 sim_outcome_30
rename=(sim_outcome_10=sim_outcome));
set bootsamp_4;
sim_outcome_05=&outcome-0.13;
sim_outcome_10=&outcome-0.38;
run;

* In WG group 3, just need to rename the outcome.;
data bootsamp_3;
set bootsamp_3;
sim_outcome=&outcome;
run;
```

7. Combine the simulated data for Groups C and D. The simulated data for each group consists of 1000 bootstrap samples.

```
* Combine simulated WG Groups 3 and 4, where the population difference;
* is 0.13 SD (actual observed difference in 13-18 year olds).;
data bootsamp_05;
set bootsamp_3 bootsamp_4_05(in=a);
WGgroup4=a;
run;

* Combine simulated WG Groups 3 and 4, where the population difference;
* is 0.38 SD (actual observed difference in 6-12 year olds).;
data bootsamp_10;
set bootsamp_3 bootsamp_4_10(in=a);
WGgroup4=a;
run;

* Get rid of temporary datasets that will not be used after this point.;
proc datasets library=work nolist;
delete
```

```
_kids9904 WGgroupP_3 WGgroupP_4
bootsamp_3 bootsamp_4
bootsamp_4_05
bootsamp_4_10;
run;
quit;

* Sort the bootstrap datasets by bootstrap sample (each has 1000 samples).;
proc sort data=bootsamp_05;
by sampnum;
run;

proc sort data=bootsamp_10;
by sampnum;
run;
```

8. Conduct the statistical test of group C vs group D in each of the 1000 bootstrap samples (use PROC SURVEYREG to take the complex survey design of NHANES into account, similar to Zanovec et al). Many of the datasets that are frequently used in public health and epidemiology studies have complex survey designs, therefore the SAS survey procedures may be very useful in this step.

```
* Statistical test of WG group within each bootstrap sample.;
* Do this for each sample scenario.;
* Note that a limitation is that these are subgroup analyses rather than;
* domain analyses, but that typically has only a minor effect on the;
* results. The effect is that the standard errors might be a little off, but
* we should still get in the right ballpark with respect to power.;

proc surveyreg data=bootsamp_05;
by sampnum;
cluster sdmvpsu;
strata sdmvstra;
weight DRweights;
model sim_outcome = WGgroup4;
ods output parameterestimates=parm_05;
run;

proc surveyreg data=bootsamp_10;
by sampnum;
cluster sdmvpsu;
strata sdmvstra;
weight DRweights;
model sim_outcome = WGgroup4;
ods output parameterestimates=parm_10;
run;
```

9. Flag bootstrap samples with a significant test result (sig=1).

```
* Flag bootstrap samples with a significant test result.;
data _parm_05;
set parm_05(where=(Parameter='WGgroup4'));
sig=(probt<0.05);
run;

data _parm_10;
set parm_10(where=(Parameter='WGgroup4'));
sig=(probt<0.05);
run;
```

10. Compute the percentage of bootstrap samples with a significant result – this is the estimated power to detect the Group C vs. D difference.

    Note that the power analysis in this example is conducted for the assumed Group C vs. D differences for both 6-12-year-olds (d = 0.38 SD) and 13-18-year-olds (d = 0.13 SD). When the macro is run for 6-12-year-olds, one should only use the result assuming a 0.38 SD difference between groups, and when it is run for 13-18-year-olds one should only use the result assuming a 0.13 SD difference.

```
* Count the number of samples with a significant test result (WG group 3;
* vs 4). This is the estimate of power.;
proc freq data=_parm_05;
tables sig / out=sig_05(where=(sig=1) keep=sig percent);
run;

data sig_05(rename=(percent=power));
retain outcome effect_size;
set sig_05;
format outcome $30.;
format effect_size $15.;
effect_size='0.13 SD';
outcome="&outcome";
drop sig;
run;

proc freq data=_parm_10;
tables sig / out=sig_10(where=(sig=1) keep=sig percent);
run;

data sig_10(rename=(percent=power));
retain outcome effect_size;
set sig_10;
format outcome $30.;
format effect_size $15.;
effect_size='0.38 SD';
outcome="&outcome";
drop sig;
run;

* Save and print results.;

data fiberlet.power4_age&agegrp._&outcome;
set sig_05 sig_10;
run;

ods listing;
proc print data=fiberlet.power4_age&agegrp._&outcome;
run;

%mend powermac_zanovec_n_bmiz;
```

Results

For the 6-12-year-old group, assuming a Group C vs. D difference of 0.38 SD in BMI-for-age, a sample size approximately twice that reported by Zanovec et al would be needed to achieve a conventionally acceptable degree of power (>80%). For the 13-18-year-olds, assuming a Group C vs. D difference of 0.13 SD in BMI-for-age, a sample size approximately 10 times that reported by Zanovec et al would be needed to achieve a conventionally acceptable degree of power. If Zanovec et al had done power analyses prior to conducting their study, perhaps they would have decided to abandon the study due to insufficient sample size, or they may have elected to find a dataset with a larger sample size.

Conclusions

Analyses of available datasets are common in epidemiology, public health and other fields. Often, such analyses use datasets that were not designed to test the hypotheses of the study. In these cases, it is worthwhile to do power analysis prior to conducting the study, to demonstrate that the sample size is large enough to test the study hypotheses. If power is inadequate, the data cannot be used to credibly test the study hypotheses – the investigator should abandon the study, seek out a dataset that is large enough to support a credible test of the study hypotheses, or modify the hypotheses so that the data are adequate to test them (for example, collapse groups in the analysis). It is worthwhile for researchers to address this issue proactively, otherwise study reviewers may (and should) take issue with insufficient sample size when the work is under review.

SAS makes bootstrap power analysis relatively easy. SAS provides a great deal of flexibility for conducting such analyses. Bootstrapping allows a realistic simulation because the analyst uses distributions in the data itself, therefore it is unnecessary to specify theoretical distributions. This paper provided one example. The code described here can be adapted to other situations.

References

Efron B, Tibshirani RJ. An introduction to the bootstrap. 1998. New York: Chapman & Hall/CRC.

Thompson D, Barton B. Concerns regarding power and other analysis issues. J Pediatr. 2011 Apr;158(4):689.

Walters, SJ. Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36. Health and Quality of Life Outcomes 2004, 2:26.

Zanovec M, O'Neil CE, Cho SS, Kleinman RE, Nicklas TA. Relationship between whole grain and fiber consumption and body weight measures among 6-to-18-year-olds. J of Ped 2010; 157(4):578-583.

Contact Information

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

Doug Thompson
Thompson Research Consulting LLC
151 North Michigan Avenue #3308
Chicago, IL 60601
(518) 248-3886
Doug.thompson@ThompsonResearchConsulting.com