Paper SD-91

Using SAS to Create a p-value Resampling Distribution for a Statistical Test

Peter Wludyka, University of North Florida; Carmen Smotherman, University of Florida

ABSTRACT

One starts with data to perform a statistical test of a hypothesis. A p-value is associated with a particular test and this p-value can be used to decide whether to reject a null hypothesis in favor of some alternative. Since the data (the sample) is usually all a researcher knows factually regarding the phenomenon under study, one can imagine that by sampling (resampling) with replacement from that original data that additional information about the hypothesis and phenomenon/study can be acquired. One way to acquire such information is to repeatedly resample for the original data set (using, for example, PROC SURVEYSELECT) and at each iteration (replication of the data set) perform the statistical test of interest and calculate the corresponding p-value. At the end of this stage one has R p-values (R is typically greater than 1,000), one for each performance of the statistical test. The resampling distribution of p-values allows one to retrospectively assess the power of the test by finding the proportion of the p-values that are less than a specified level of significance (alpha). By creating a p-value resampling distribution for a selection of sample sizes one can create a power curve which can be used prospectively to gather sample size information for follow up studies. In addition to the p-value distributions bootstrap tests will be presented with the idea that these should be compared to the tests initially applied to the data.

INTRODUCTION

Typical of resampling (bootstrap) methods the basic idea is that the sample represents the population from which the sample arose; that is, one treats the sample as if it were a population. Given that a statistical test is performed on the sample there is a p-value associated with that test. One might wonder how this p-value might change were a "different sample" selected from the same population. The surrogate for this "different sample" is a sample drawn with replacement from the initial sample. Repeated resampling and performance of the statistical test on each of these resamples gives rise to the p-value resampling distribution. The basic idea is described in the schematic (Figure 1). The research question as well as the sampling methodology and nature of the measurements collected determine (influence) the choice of statistical test. The p-value resampling distribution describes tests results that might be associated with repeating the study with the same population.

In this paper the general method for creating the p-value resampling distribution using SAS will be presented. Two examples will be presented which illustrate how one might present the p-value resampling distribution, as well as how one might calculate and present both retrospective and prospective power estimates. In addition to the p-value distributions bootstrap tests will be presented with the idea that these should be compared to the tests initially applied to the data.

THE P-VALUE RESAMPLING DISTRIBUTION AND BOOTSTRAPING

The following will be illustrated through the examples:

- Step1: Perform statistical test and assess power retrospectively from the original sample
- Step2: Generate the p-value resampling distribution and estimate power using bootstrapping
- Step3: Assess the appropriateness of a particular statistical test

EXAMPLE 1: ONE SAMPLE T-TEST

We illustrate with a simple one sample problem with hypotheses:

$H0: \mu_Y = 140$ (1)

$H1: \mu_Y \neq 140.$

Step1. The analysis is performed using SAS program 1 in which the sample (1A) is a pseudo random sample of size n=10 from $Y \sim N(144, 7^2)$, a normal population with mean 144 and standard deviation 7. The actual sample is in Table 1; the summary statistics and output from PROC TTEST are in Output 1 (SAS program1, line 1). Based on the p-value = 0.0252 (t = 2.68) one rejects the null hypothesis at the 5% level of significance. In decision theoretic terms the story is over and one concludes that the population mean is not 140. This test assumes (approximate) normality and under that assumption the power of the test is 66.6% (See Output 2, which describes the scenario for the power analysis from PROC POWER, line 2). This assessment of retrospective power is based on notion that were the true

population mean and standard deviation equal to the sample values (Output 1) then the power of a one sample *t*-test of hypothesis (1) is 66.6%.

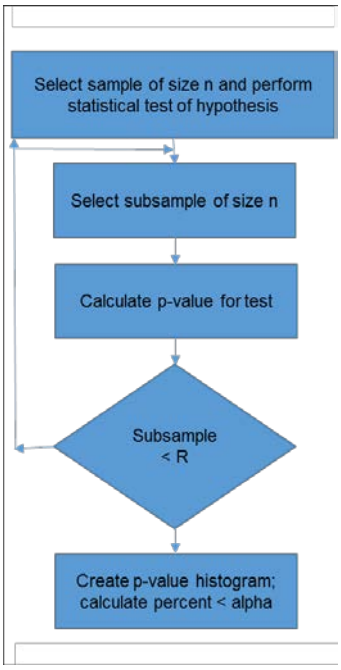


Figure 1. The Process of Finding the P-Value Distribution

Sample 1A	
150.06	139.45
137.46	146.36
151.10	141.94
140.18	144.25
141.10	150.64

Table 1. Sample 1A from Normal

N	Mean	Std Dev	Std Err	Minimum	Maximum
10	144.3	5.0195	1.5873	137.5	151.1

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
144.3	140.7	147.8	5.0195	3.4526	9.1637

DF	t Value	Pr > |t|
9	2.68	0.0252

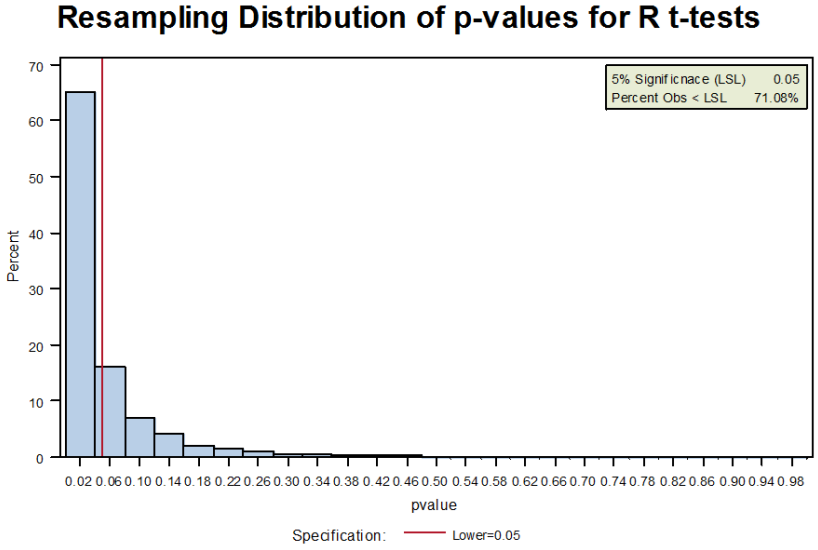
Output 1. Output from PROC TTEST: Summary Statistics for the Original Sample (SAS program 1, line 1)

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Mean	4.2538
Standard Deviation	5.0195
Total Sample Size	10
Number of Sides	2
Null Mean	0
Alpha	0.05

Computed Power
Power
0.666

Output 2. Output from a PROC POWER: Power of *T*-Test based on observed sample 1A (SAS Program 1, line 2)

Step 2. From the p-value resampling distribution (Figure 2) the estimated power is 71.1% (95% CI: 69.8%, 72.3%, Output 3); that is, 71.1% of the resampled p-values are less than 0.05. The p-value resampling distribution (and corresponding power estimates) makes no assumptions regarding the population from which the sample was selected (other than that of "representativeness" of the sample). The p-value distribution power statement basically says that were the researcher to perform a t-test based on a sample from the population as represented by the initial sample, the likelihood that the null hypothesis will be rejected is about 71.1%. This does not address the appropriateness of the *t*-test in this application. Note that based on Figure 2 one can judge the power associated with other values for the level of significance by estimating the area to the left of the supposed value of alpha.



Fiigure 2. P-value Resampling Distribution for R = 5000 *t*-test based on Sample 1A (PROC CAPABILITY, line 4).

Binomial Proportion for reject = 0	
Proportion	0.7108
ASE	0.0064
95% Lower Conf Limit	0.6982
95% Upper Conf Limit	0.7234
Exact Conf Limits	
95% Lower Conf Limit	0.6980
95% Upper Conf Limit	0.7233

Output 3. Output from PROC FREQ: Confidence Interval of Power Based on Sample 1A (SAS program 1, line 3)

Step 3. One way to assess the appropriateness of a particular statistical test (in this case the one sample *t*-test applied to this data/population) is to test the hypothesis using bootstrap methods. Two general approaches to this are:

1. Construct a bootstrap confidence interval for the parameter of interest and check to see whether the hypothesized value is in the interval (rejecting if the hypothesize value is not in the interval).

   The bootstrap confidence interval for data set 1A is $141.3 < \mu_Y < 147.1$, which can be seen in Figure 3 along with resampling distribution of R = 5000 sample means. Since this confidence interval does not contain 140, one rejects hypothesis (1), $\mu_Y = 140$. This confidence interval was found using PROC UNIVARIATE (SAS program 1, line 5) and Figure 3 was created using PROC CAPABILITY (SAS program 1, line 6). One can compare this confidence interval in Output 1.

2A. Transform the data to match the null distribution (denote this sample as the "null distribution") (SAS program 1, line 7) and then compare the sample mean of interest calculated from the initial sample to the distribution of bootstrap values of the mean from the null distribution. An unusual value from the initial sample argues against the null hypothesis.

   Since the sample mean Y=144.254 for sample 1A differs from the hypothesized mean by $4.254 = 144.254 - 140$, the null distribution is created by transforming each Y by subtracting 4.254. The resulting data set has mean zero but retains the sample standard deviation. One simple bootstrap test is to select samples from this null distribution (Noreen, 1989) and locate cutoffs: lower cutoff equals 140-4.254= 135.746 and upper cutoff equals 140+4.254 = 144.254. The p-value for the bootstrap test is the proportion of sample means from the null distribution below the lower cutoff plus the proportion above the upper cutoff. See Figure 4, in which the p-value = 0.0038. One can compare this with the p-value in Output 1.

2B. Resampling from the original data, at each iteration create a t statistic ($i = 1, \ldots R$)

$$t_i = \frac{\bar{Y}_i - \bar{Y}_0}{S_{Y_i}/\sqrt{n}}$$

where $\bar{Y}_0$ is the mean of the original sample and the denominator of t_i is the standard error of the mean calculated from the i^{th} resample. Then compare the *t*-statistic of interest calculated from the initial sample, t_0 ,to the distribution of bootstrap t statistics. An unusual value from the initial sample argues against the null hypothesis.

   The t- statistic is calculated for each sample obtained by bootstrapping (SAS program 1, line 9) and then it is compared to the t-statistic of the original sample 1A. The p-value of the bootstrap test is the proportion of tests statistics more extreme than the one observed (2.68). See Figure 5, in which p=0.0348.

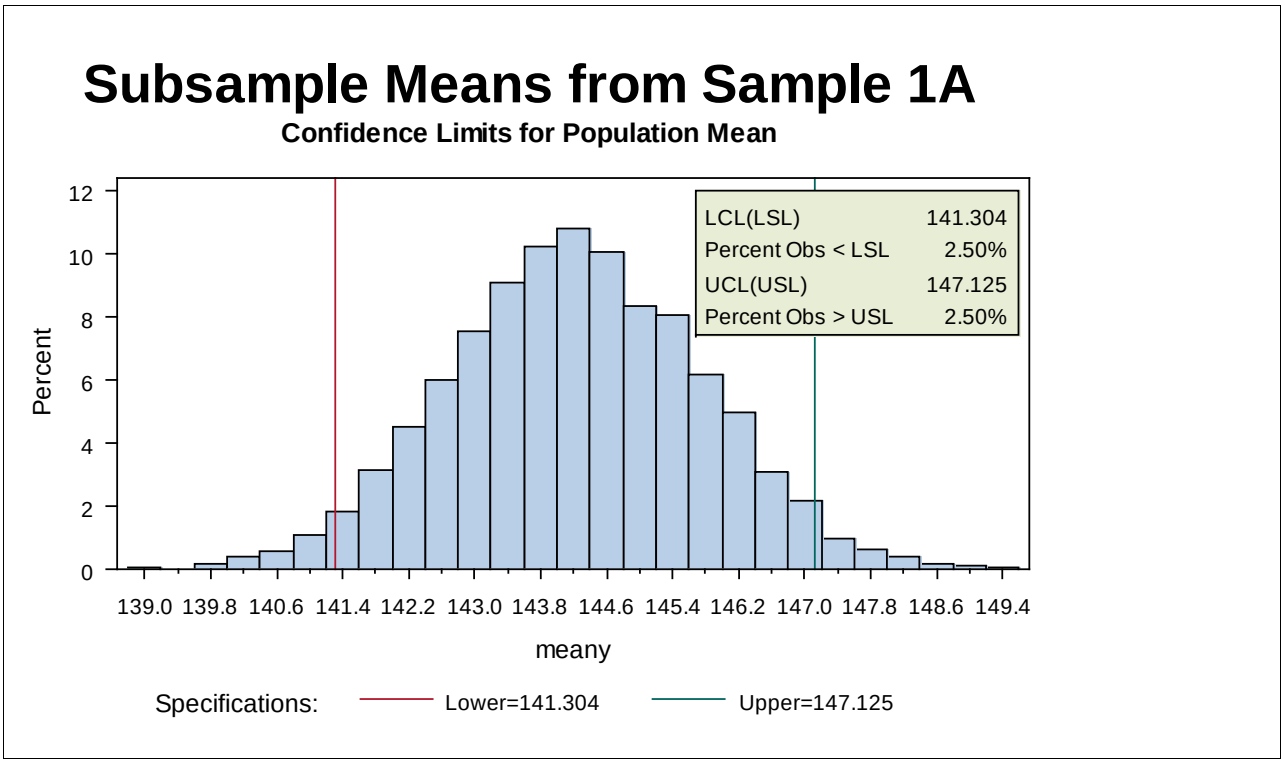


FIGURE 3. Resampling Distribution for R = 5000 Sample Means from Sample 1A with 95% Bootstrap Confidence Interval (LCL, UCL) for Population Mean (PROC CAPABILITY, line 6)

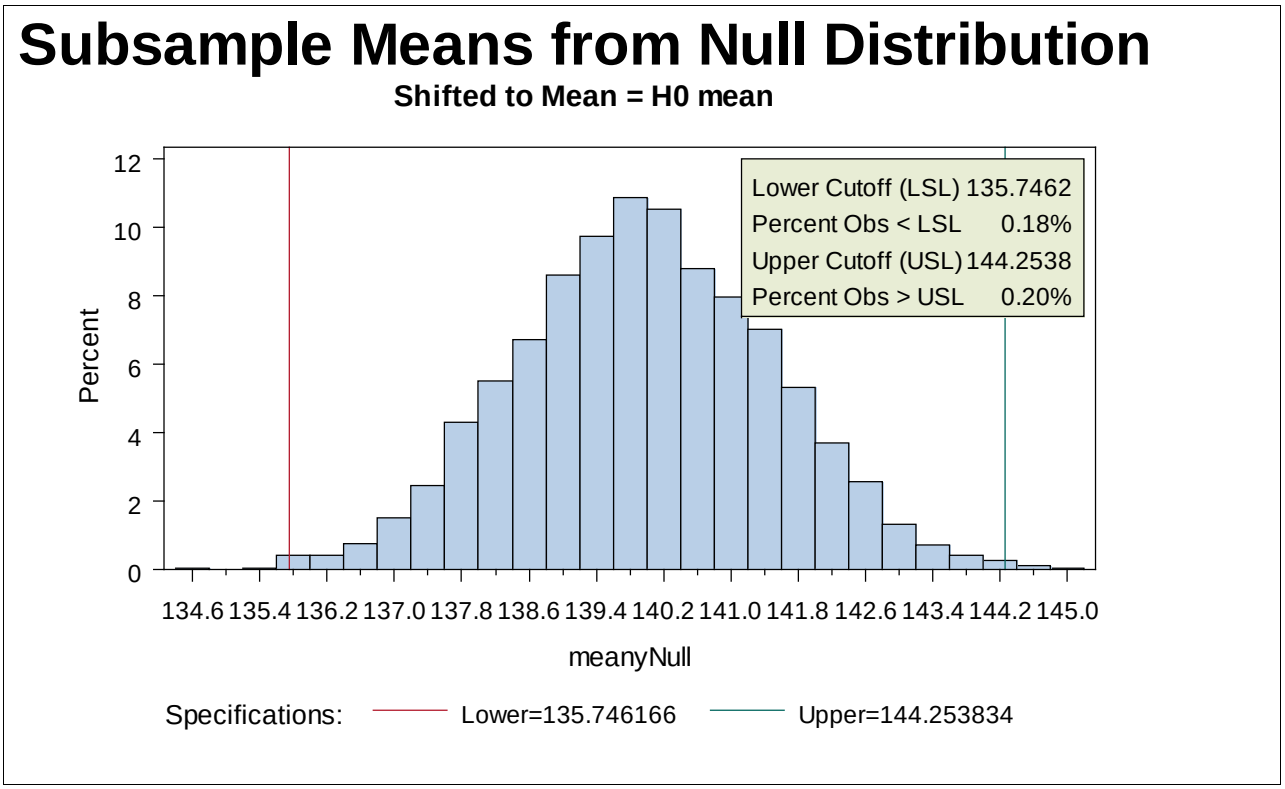


FIGURE 4. Resampling Null Distribution with Cutoffs and Tail Areas for 5% Bootstrap Test of Hypothesis of Hypothesis (1) based on Data Set 1A. (PROC CAPABILIT, line 8).

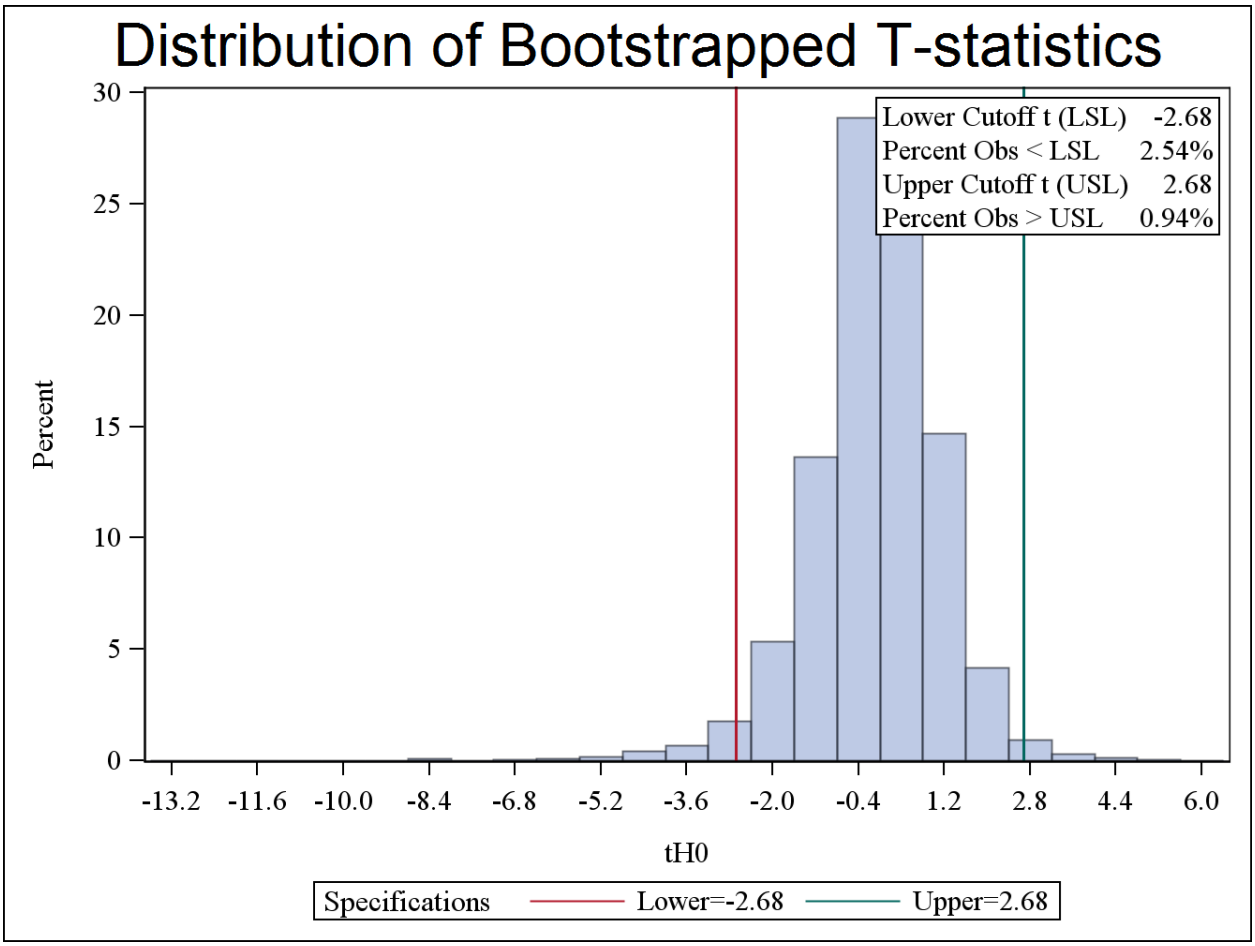


FIGURE 5. Distribution of Bootstrapped Test Statistic. (PROC CAPABILIT, line 10).

The SAS program 1 (for Example 1) is listed below.

```
 ***Generate the original sample 1A;
data Ydata;
       seed = 12345679;
       mu = 144;
       sigma = 7;
       H0mu = 140;
       n = 10;
       do i = 1 to n;
         call rannor(seed, norm01);
         y = mu + sigma*norm01;
       output;
       end; run;
proc means data = Ydata; var Y;
       title 'Summary Stats for Original Data Set';
       output out = YdataSummary n = n mean = mean
       stddev = stddev var = var stderr = stderr; run;
data YdataSummary; set YdataSummary;
       do s = 1 to n;
       n = n ;mean = mean; stddev = stddev; var = var; stderr = stderr;
       output;
```

```
      end; run;
data Ydata; merge Ydata YdataSummary;
      title 'Data Set with Appended Summary Stats';
      df = n-1;
      t = (mean - H0mu)/stderr;
      p = (1-probt(abs(t),df))*2;
      replicate = 5000; run;
***perform t test;
proc ttest data = Ydata h0 = 140;                                          /*1*/
      title 't test on Mean';
      var y; run;
***calculate power;
proc power ;                                                               /*2*/
      title 'power for t test based on normality';
      onesamplemeans
        mean   = 4.2538
        ntotal = 10
        stddev = 5.0195
         power  = .; run;
*********************************
Create subsamples
*********************************/;
proc surveyselect data=Ydata out=outbootB
      seed=30459585
      method= urs
      sampsize =10
      outhits
      rep=4999;
      title 'select bootstrap samples'; run;
data outbootB; set outbootB ;
      source = 2; /*'Boot';*/ run;
/*****************************************
Append original sample to bootstrap samples
*****************************************/
data outbootBplus; set outbootB Ydata;run;
proc sort data = outbootBplus;
      by replicate; run;
proc means data =  outbootBplus noprint;
      var y; by replicate;
      output out = ysummary n=ny mean = meany stddev = stddevy var = vary;run;
data  ysummary; set ysummary;
      df = ny-1;
      ObsDiff = meany - 140;
      SEM = stddevy/(ny**0.5);
      tstat= obsdiff/SEM;
      pvalue =(1-probt(abs(tstat),df))*2;
      if pvalue < 0.05 then reject = 0; else reject = 1; /* reject = 0*/
      popVar = vary*(ny-1)/ny;
      popSD = popVar**0.5; run;
proc freq  data = ysummary ;                                               /*3*/
      title 'Rejection rate across replicates: power';
      table reject/ binomial; run;
proc capability data =     ysummary noprint;                               /*4*/
      title 'Resampling Distribution of p-values for R t-tests';
      spec lsl=0.05 ;
      var pvalue ;
      histogram pvalue ;
      inset lsl='5% Significance (LSL)' lslpct / cfill = ywh pos=ne; run;

/*********************************
Bootstrap 95% CI for Mean
*********************************/
proc univariate data=ysummary noprint;
```

```
      title 'Distribution of Bootstrap Means';
      title2 'Bootstrap: CI for mean given by percentiles';
      var meany;
      histogram meany / normal;
      output out=finalA pctlpts=2.5, 97.5 pctlpre=ci; run;                    /*5*/
data CIspecs; set finalA;
      _LSL_ = ci2_5;
      _USL_ = ci97_5;
      _VAR_ = 'meanY'; run;
proc capability data =     ysummary spec = CIspecs noprint;                   /*6*/
      /*spec lsl=141.304 usl = 147.125;*//* One may enter these values manually*/
      var meanY;
      histogram meany;
      inset lsl='LCL(LSL)' lslpct usl = 'UCL(USL)' uslpct / cfill = ywh pos=ne;
      title 'Subsample Means from Sample 1A';
      title2 'Confidence Limits for Population Mean'; run;

/*************************************
Bootstrap test of hypothesis: Efron 1989
*************************************/
title 'Bootstrap test on mean';
proc means data = Ydata;var y;
      output out = realdataOUT mean = mean; run;
data realdataoutn; set realdataout;
      do m = 1 to 10;
      mean = mean;
      output;
      end; run;

***Create Null Data Set: transform to null mean;
data NullData; merge realdataOUTn Ydata;
      H0mean = 140;
      ShiftedY = y - (mean - H0mean);                                          /*7*/
      Shift = abs(mean - H0mean);
      LowerCutoff = H0mean -Shift;
      UpperCutoff = H0mean + Shift; run;
data NullDataShell; set realdataOUT;
      do ss = 1 to 5000;
      end; run;
proc surveyselect data=nulldata out=outbootBShifted
      seed=30459585
      method= urs
      sampsize =10
      outhits
      rep=5000;
      title2 'select bootstrap samples from null distribution'; run;
proc means data =  outbootBShifted  noprint;
      var ShiftedY; by replicate;
      output out = ysummaryS n=ny mean = meanyNull stddev = stddevy var = vary; run;
proc means data =  outbootBShifted  noprint;
      var H0mean; by replicate;
      output out = ysummaryS1 mean = H0mean; run;
proc means data =  outbootBShifted  noprint;
      var LowerCutoff; by replicate;
      output out = ysummaryS2 mean = LowerCutoff; run;
proc means data =  outbootBShifted  noprint;
      var UpperCutoff; by replicate;
      output out = ysummaryS3 mean = UpperCutoff; run;
data  ysummaryS ; merge ysummaryS  ysummaryS1 ysummaryS2 ysummaryS3;
      reject1 = 1; /* reject = 0*/
      if meanYnull < LowerCutoff then reject1 = 0;
      if meanYnull > UpperCutoff then reject1 = 0;
      title2 'P-Value for Bootstrap test on mean'; run;
```

```
proc freq data = ysummaryS ;
      table reject1 / chisq binomial; run;
data h0testSpecs ; set realdataOUT;
      H0mean = 140;
      ShiftedY = y - (mean - H0mean);
      Shift = abs(mean - H0mean);
      LowerCutoff = H0mean -Shift;
      UpperCutoff = H0mean + Shift;
      _LSL_ = LowerCutoff;
      _USL_ = UpperCutoff;
      _var_ = 'meanYnull'; run;
proc capability data =     ysummaryS spec = h0testSpecs noprint;                  /*8*/
      /*spec lsl=136.544 usl = 143.456;*//* One may enter these values manually*/
      var meanYnull;
      histogram meanynull ;
      inset lsl='Lower Cutoff (LSL)' lslpct usl = 'Upper Cutoff (USL)' uslpct / cfill
      = ywh pos=ne;
      title 'Subsample Means from Null Distribution';
      title2 'Shifted to Mean = H0 mean'; run;

/*****************************
t test boostrap:  Efron & Tibshirani (1993), pages 224-227.
*****************************/
title 'T test boostrap';
data ysummaryH0; set ysummary;
      BaseMean =  144.2538337; BaseT = 2.68;
      tH0 =  (meany-basemean)/ SEM;                                            /*9*/
      rejectT = 0;
      if tH0 > BaseT then rejectT = 1;
      if tH0 < -BaseT then rejectT = 1;run;
proc freq data = ysummaryH0;
      tables rejectT/ binomial; run;
proc capability data =     ysummaryH0 noprint;                                 /*10*/
      spec lsl= -2.68 usl = 2.68;      /* One may enter these values manually*/
      var tH0;
      histogram tH0;
      inset lsl='Lower Cutoff t (LSL)' lslpct usl = 'Upper Cutoff t (USL)' uslpct /
      cfill = ywh pos=ne;
      title 'Subsample t ';
      title2 '(meanSUB - meanData)/[SEM SUB]';
run;
quit;
```

EXAMPLE 2: SIMPLE LOGISITC REGRESSION

The second example illustrates how to estimate power for different sample sizes in a scenario with one predictor X (continuous variable) and a dichotomous dependent variable Y (values 0 or 1). The hypotheses are:

H0:$\beta_1=0$ (1)

H1: $\beta_1\neq0$,

The null hypothesis states that the line describing the relationship between the independent variable and the probability of the dependent variable has a slope of zero.

The original sample for this example has 66 observations, and the SAS MACRO Power is presented below.

To estimate the power, one directly re-samples the existing data without any assumption about the underlying distribution of the sampled population using PROC SURVEYSELECT (SAS MACRO Power, line 1). Then, simple logistic regression models are built for each sample generated (SAS MACRO Power, line 2). Wald Chi-square tests are used on each sample to test the null hypothesis that X has no effect Y (SAS MACRO Power, line 3). To estimate the power, the percentage of times the null hypothesis is rejected is calculated (SAS MACRO Power, line 4). Power is estimated at different sample sizes then the power is plotted against the corresponding sample size (SAS MACRO Power, line 6). The minimum size with an estimated power that satisfies the given requirements will then be chosen.

For the selected number of repetitions in this example (SAS MACRO Power, line 5), the power curve is presented in Figure 6.

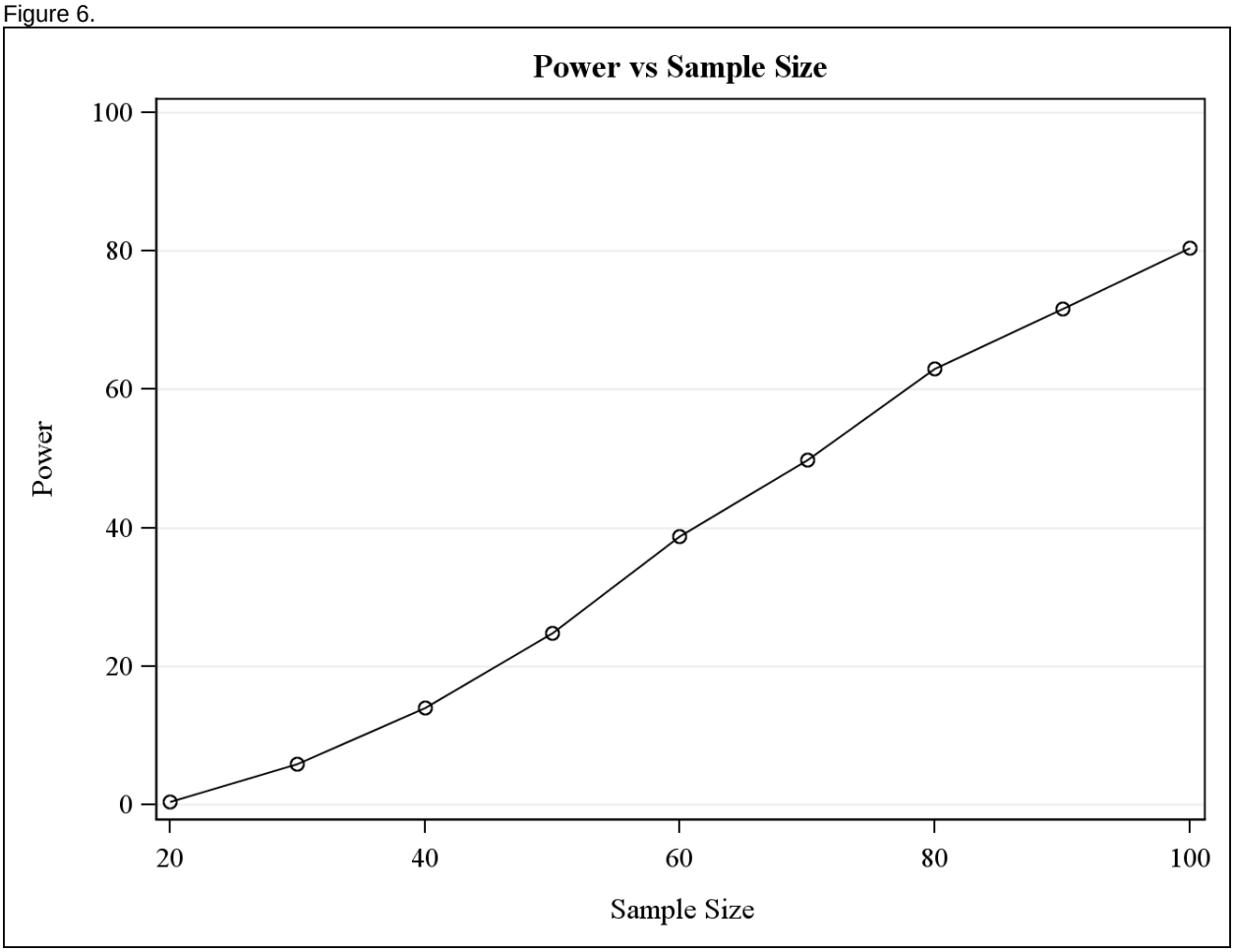


FIGURE 6. Power Curve by Sample Size

The SAS Macro Power (for Example 2) is listed below.

```
data sample;
       input id y x;
       datalines;
       1      0      7.24
       2      0      15.52
       …
       66     0      28.8
;run;

/*****************************************************************************/
/* MACRO Power name Macro, creates a table and a graph for Power by sample size
/* orig_dsn    data set name
/* seed        the initial seed for random number generation
/* method=urs  requests unrestricted random sampling, which is selection  with equal
               probability and with replacement
/* startsize   minimum sample size, which is the number of units to select for the
               sample
```

```
/* endsize     maximum sample size, which is the number of units to select for the
               sample
/* incr        the increment for sample size
/* rep         the number of repetitions
/* dep_var     the outcome variable, y
/* pred_var    the predictor variable, x
/*************************************************************************************/

%MACRO
Power(orig_dsn=,seed=,method=urs,startsize=,endsize=,incr=,rep=,dep_var=,pred_var=);
   PROC DATASETS;
      DELETE Power_est Power_est2 Power_Set;
   QUIT;

   %DO i=&startsize %TO &endsize %BY &incr;
      PROC SURVEYSELECT DATA=&orig_dsn OUT=outbootB                              /*1*/
         SEED=&seed
         METHOD=urs
         SAMPSIZE=&i
         OUTHITS
         REP=&rep;
      run;

      PROC LOGISTIC DATA=outbootB OUTEST=out1;                                   /*2*/
         BY replicate;
         MODEL  &dep_var(event='1')=&Pred_var;
                  ODS OUTPUT ParameterEstimates= work.ParameterEstimates1;;
       RUN;

      DATA ParameterEstimates2;
         SET ParameterEstimates1;
         reject=0;
         IF probchisq<.05 THEN reject=1;
         IF VARIABLE ='Intercept' THEN DELETE;                                   /*3*/
      RUN;

      PROC FREQ DATA=ParameterEstimates2;
         TABLES reject;
         ODS OUTPUT OneWayFreqs=Power_est;
      RUN;
      DATA Power_est2;
         SET power_est;
         IF reject ne 1 then DELETE;
         Samp_Size=&i;
         Power=Percent;                                                         /*4*/
         KEEP Samp_Size Power;
      RUN;

      PROC APPEND BASE=Power_set DATA=Power_est2;
      RUN;
   %END;

   ODS HTML;
   TITLE 'Power by sample size';
   PROC REPORT DATA=Power_set NOWD;
      COLUMN Samp_Size Power;
      DEFINE Samp_Size / DISPLAY "Sample Size";
      DEFINE Power / DISPLAY FORMAT=4.1;
   RUN;
   ODS HTML CLOSE;
%MEND RepSample;

%Power(orig_dsn=sample,                                                         /*5*/
```

```
            seed=30459585,
            startsize=10,
            endsize=100,
            incr=10,
            rep=1000,
            dep_var=y,
            pred_var=x)

ODS RTF;
title 'Power vs Sample Size';
PROC SGPLOT DATA=Power_SET;                                                    /*6*/
   SERIES X=Samp_Size Y=Power / markers;
   yaxis min=0 max=100 grid;
   xaxis label='Sample Size';
RUN;
ODS RTF CLOSE;
```

CONCLUSION

The resampling distribution of p-values can be easily constructed for a statistical test or procedure using bootstrap methods. It allows one to retrospectively assess the power of the test by finding the proportion of the p-values that are less than a specified level of significance (alpha). By creating a p-value resampling distribution for a selection of sample sizes one can create a power curve which can be used prospectively to gather sample size information for follow up studies. In addition to the p-value distributions, bootstrap tests can be successfully used when compared these to the tests initially applied to the data.

REFERENCES

Noreen, Eric. 1989. Computer Intensive Methods for Testing Hypotheses. P66-69. New York, NY, John Wiley & Son.

Efron, Bradley, and Tibsirani, R. 1993 Introdcution to the Bootstrap, Chapman and Hall.

CONTACT INFORMATION

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

Name: Peter Wludyka
Enterprise: University of North Florida / Department of mathematics and Statistics
Address:
City, State ZIP:
E-mail: pwludyka@unf.edu