PharmaSUG2011 - Paper SP08

Estimating Sample Size through Simulations

Wuchen Zhao, University of Southern California, Los Angeles, CA
Arthur X. Li, City of Hope National Cancer Center, Duarte, CA

ABSTRACT

Determining sample size is one critical and important procedure for designing an experiment. The sample size for most statistical models can be easily calculated by using the POWER procedure. However, the PROC POWER cannot be used for a complicated statistical model. This paper reviews a more generalized method to estimate the sample size through a simulation approach by using SAS® software. The simulation approach not only applies to the simple but also to a more complex statistical design.

INTRODUCTION

One of the important components of planning a statistical study is determining the sample size. An adequate size is critical to produce a statistically-significant result. The larger the sample size, the more accurate the estimate of the parameters of the population. The law of large numbers tells us that the average of the results calculated from a large number of trials converges to the expected value. However, having a large sample size will definitely increase your research budget. If the experiment involves humans, the unnecessary number of human subjects will be potentially exposed to possible harmful treatment. Thus, we want the sample size to be large enough to detect the association of the research interest.

Estimating the sample size is closely related to understanding the type I error, type II error, and the power of the statistical test, which is summarized in the table below:

		Null Hypothesis (H_o)	
		TRUE	FALSE
Decision	Accept H_o	Correct (1 – α)	Type II Error (β)
	Reject H_o	Type I Error (α)	Correct (1- β)

The type I error (α), or false positive, is the probability of rejecting a null hypothesis when the null hypothesis is true. On the other hand, the type II error (β), or false negative, is the probability of failing to reject the null hypothesis when the null hypothesis should be rejected. Power is the probability of correctly rejecting the null hypothesis when the null hypothesis is false.

The sample size is closely related to the power because an experiment with a larger sample size will have more power than an experiment with a smaller sample size.

CALCULATING SAMPLE SIZE BY USING THE POWER PROCEDURE

Suppose that you are asked to estimate the sample size for testing a new diet pill for weight loss for a half-year period by a pharmaceutical company. For this experiment, overweight patients with mean weight equaling 265 pounds are recruited and they are assigned randomly to take either the new drug or a placebo. To simplify the experiment, you would like to have the same number of patients taking the drug and the placebo. You would like to have 80% power to detect a weight loss of at least 15 pounds for patients taking the new diet pill compared to the ones taking the placebo. Based on previous research, the standard deviation for the patient's weight is 50. To use PROC POWER to calculate the sample size, you can use the following program:

```
proc power;
    twosamplemeans test=diff
    groupmeans = 250 | 265
    stddev = 50
    alpha = 0.05
    sides = 1
    npergroup = .
    power = 0.8;
run;
```

Output:

```
The POWER Procedure
Two-sample t Test for Mean Difference

     Fixed Scenario Elements

Distribution                Normal
Method                       Exact
Number of Sides                  1
Alpha                         0.05
Group 1 Mean                   250
Group 2 Mean                   265
Standard Deviation              50
Nominal Power                  0.8
Null Difference                  0


Computed N Per Group

Actual    N Per
 Power    Group

 0.802      139
```

Based on the SAS output above, you will need 139 patients for each group.

CALCULATING SAMPLE SIZE THROUGH SIMULATION

The equation calculated from the two-sample t-test is based on the following formula:

$$N = \frac{\sigma^2(1/Q_e + 1/Q_c)(Z_\alpha + Z_\beta)^2}{d^2} \qquad where\ d = \mu_1 - \mu_2 \tag{1}$$

Q_e is the proportion of N subjects allocated to the experimental treatment (e.g. drug) and Q_c = 1 - Q_e
For the case of equal allocation, like in the example above, the calculation for N can be simplified by the following formula:

$$N = \frac{2\sigma^2(Z_\alpha + Z_\beta)^2}{d^2} \tag{2}$$

N in the equation (2) represents the sample size of each group. The equation (2) can also be arranged as the following:

$$N\left(\frac{\mu_1 - \mu_2}{2\sigma}\right)^2 = \frac{(Z_\alpha + Z_\beta)^2}{2} \tag{3}$$

or

$$N\left(E(T)\right)^2 = \frac{(Z_\alpha + Z_\beta)^2}{2} \tag{4}$$

since $\frac{\mu_1 - \mu_2}{2\sigma}$ can be considered the expected value of the two-sample t-test. Based on equation (4), if the expected value of the test statistics can be calculated, you can find N. The expected value can be calculated through simulation by using the following steps:

1. Simulate a large data set based on the design parameters
2. Analyze the simulated data set to obtain the test statistics
3. Calculate the expected test statistics depending on the distribution of the test statistics

CALCULATING SAMPLE SIZE FOR TWO-SAMPLE T-TEST THROUGH SIMULATION

The previous example for estimating the sample size for testing the diet pills can also use the simulation approach.

In this experiment, you need to simulate a large data set, perhaps two million subjects, and divide the subjects into two groups: one taking the diet pill and the other taking the placebo. Since you are expecting patients in the diet pill group to lose 15 pounds compared to the patients in the placebo group, you can set the mean weight in the diet pill group to 250 (losing 15 pounds) and 265 for the placebo group.

```
%let sim_n = 2000000;
%let mu_drug = 250;
%let mu_placebo = 265;
%let sigma = 50;

data sim (drop = i seed);
    retain seed 1;
    length group $ 7.;
    do i = 1 to &sim_n;
        if ranuni(seed) < 0.5 then do;
            group = 'drug';
            weight = &mu_drug + &sigma * rannor(seed);
        end;
        else do;
            group = 'placebo';
            weight = &mu_placebo + &sigma * rannor(seed);
        end;
        output;
    end;
run;

title "The first 5 observations of simulated data";
proc print data = sim (obs = 5) noobs;
run;
```

Output:

```
The first 5 observations of simulated data

group         weight

drug         240.039
drug         269.829
placebo      318.472
drug         218.788
placebo      376.986
```

The next step will be calculating the t-statistics based on the simulated data, which can be done by using the TTEST procedure.

```
proc ttest data = sim;
    class group;
    var weight;
    ods output ttests = stats;
run;

title 'The T-test Result';
proc print data = stats noobs;
run;
```

Output:

```
The T-test Result

 Variable     Method           Variances     tValue        DF     Probt

weightloss    Pooled            Equal       -211.37       2E6    <.0001
weightloss    Satterthwaite     Unequal     -211.37       2E6    <.0001
```

The expected values for the T statistics can be estimated by dividing t-value based on the simulated data set by the square root of simulated sample size, which is 2 million in our example ($E(T) = T / \sqrt{M}$). This is the average contribution of each person to the total test statistics. For calculating the expected value of F and chi-square statistics, the expected value can be calculated by dividing the F or chi-square statistics by the simulated size *(E(F) = F / M)*. The following SAS code illustrates the calculation for the expected value and the sample size:

```
%let alpha = 0.05;
%let power = 0.80;
%let sides = 1;
data size_n(keep = alpha power z_alpha z_beta tvalue expected_t n);
    set stats;
    if method = 'Pooled';
    z_alpha = probit(1 - &alpha/&sides);
    z_beta = probit(&power);
    expected_t = abs(tvalue)/sqrt(&sim_n);
    n=((z_alpha+z_beta)**2/expected_t**2)/2;
run;

title 'The required sample size for each group';
proc print data=size_n noobs;
run;
```

Output:

```
The required sample size for each group


                                 expected_
 tValue    z_alpha     z_beta        t           n

-211.37    1.64485    0.84162     0.14946     138.378
```

The sample size that is calculated from the simulation above is the same as the result from the POWER procedure. Sometimes the result may not be exactly the same, but it should be very close from the results that are based on the theoretical formula.

CALCULATING SAMPLE SIZE FOR INTERACTION THROUGH SIMULATION

For estimating the sample size for a two-sample t-test, you don’t need to use the simulation approach because the POWER procedure can handle almost all types of experimental design. Using simulation to estimate the sample size is more useful for a more-complicated statistical model.

To expand our previous example, suppose that you would like to estimate the sample size for detecting the interaction effect between caloric intake and whether or not the diet pill or the placebo was taken. We assume that there is a negative correlation between caloric intake and weight loss. That is to say, the more you eat, the less weight you will lose. However, if you are taking the diet pill, the rate of weight loss decreases slower for the amount of increases in food consumption compared to people who did not take the diet pill. How large the sample size do you need if you desire 80% power to detect the interaction at a 5% significant level? Here’s the statistical model for this experiment:

$$Y_{Weightloss} = \alpha + \beta_1 \cdot X_{Drug} + \beta_2 \cdot X_{Calorie} + \beta_3 \cdot X_{Drug} \cdot X_{Calorie} + \varepsilon \qquad (5)$$

In Equation (5), X_{Drug} equals 1 for a patient allocated in the diet pill group; otherwise X_{Drug}= 0 for the placebo group. $X_{Calorie}$ is a continuous variable for caloric intake. This equation (5) is equivalent to the following two equations:

Diet Pill Group: $Y_{Weightloss} = \alpha + \beta_1 + (\beta_2 + \beta_3) X_{Calorie}$

Placebo Group: $Y_{Weightloss} = \alpha + \beta_2 X_{Calorie}$

Based on the equation above, you can see that the only parameter that influences the interaction term will be β_3. You can set the arbitrary number of the rest of the parameters. Suppose that based on previous studies, you are setting the parameters as following: α = 10, β_1=5, β_2= -0.004, and β_3= 0.0025; you will then have the following two equations:

Diet Pill Group: $Y_{Weightloss} = 15 - 0.0015 X_{Calorie}$

Placebo Group: $Y_{Weightloss} = 10 - 0.004 X_{Calorie}$

These two equations tell us that for each additional 100 calories consumed per day, the weight loss will decrease 0.15 pounds in the diet pill group. On the other hand, weight loss will decrease 0.4 pounds for each additional 100 calories consumed per day in the placebo group.

Furthermore, assume the number of calories consumed per day follows the normal distribution with mean equaling 2500 and standard deviation equaling 1000. The residual (*ε*) is assumed to be normally distributed with means equaling 0 and standard deviation equaling 5. Lastly, only 30% of the subjects are taking the diet pill. Here's the SAS code for the simulation step:

```
%let sim_n=1000000;
%let p_drug=0.3;
%let mean_cal=2500;
%let sd_cal=1000;
%let alpha=10;
%let beta1=5;
%let beta2=-0.004;
%let beta3=0.0025;
%let sd_error=5;

data sim (drop = seed i);
    retain seed 1;
    do i=1 to &sim_n;
        if ranuni(seed) < &p_drug then drug=1;
        else drug=0;
        calorie = &mean_cal + &sd_cal * rannor(seed);
        weightloss =&alpha + &beta1*drug + &beta2* calorie  +&beta3*drug*calorie
             + &sd_error*rannor(seed);
        output;
    end;
run;

title 'The frist 5 observations of the simulated data';
proc print data=sim (obs=5) noobs;
run;
```

```
The frist 5 observations of the simulated data

drug    calorie    weightloss

  1     2300.78      18.7863
  0     1416.68      15.5247
  0     3187.84       8.4472
  1     1905.82      12.3007
  0     1258.70      -2.8413
```

Based on the simulated data, you can run the linear regression by using the GLM procedure.

```
proc glm data = sim;
    model weightloss = drug calorie drug*calorie/solution;
    ods output parameterestimates=pe;
quit;

title 'The result from the linear regression';
proc print data=pe;
run;
```

SAS output:

```
The result from the linear regression

Dependent     Parameter           Estimate          StdErr      tValue     Probt

weightloss    Intercept        9.999851985      0.01605913      622.69    <.0001
weightloss    drug             5.032458567      0.02931489      171.67    <.0001
weightloss    calorie         -0.003996279      0.00000597     -669.75    <.0001
weightloss    drug*calorie     0.002481769      0.00001089      227.91    <.0001
```

When calculating the sample size for the two-sample t-test in the previous example, we used the equation (4) that is based on a one-sided test. For calculating the sample size for the interaction term, you should use the following equation:

$$N\left(E(X)\right)^2 = (Z_\alpha + Z_\beta)^2$$

Here's the final SAS code for calculating the sample size:

```
%let alpha=0.05;
%let power=0.80;
%let sides=2;

data size_n(keep=z_alpha z_beta tvalue expected_t n);
    set pe;
    if parameter='drug*calorie';
    z_alpha = probit(1 - &alpha/&sides);
    z_beta = probit(&power);
    expected_t = abs(tvalue)/sqrt(&sim_n);
    n = (z_alpha + z_beta)**2/expected_t**2;
run;

title 'The Sample Size for Testing the Interaction';
proc print data=size_n noobs;
run;
```

SAS Output:

```
The Sample Size for Testing the Interaction


                                  expected_
 tValue    z_alpha     z_beta        t           n

 227.91    1.95996    0.84162     0.22791     151.112
```

Based on the calculation above, you need to have at least 152 people to detect the interaction between the drug and caloric intake.

CONCLUSION

Calculating the sample size is one of the most important tasks for statisticians. When the study design is complex and cannot be calculated by the POWER procedure, you can use the simulation method to solve the problem.

ACKNOWLEDGMENTS

We would like to thank Dr. Jim Gauderman from the Department of Preventive Medicine at U.S.C. for his guidance when we took his linear regression class. This paper is inspired by one of his lab handouts.

CONTACT INFORMATION

Wuchen Zhao
University of Southern California
Department of Preventive Medicine
Division of Biostatistics
1540 Alcazar Street
CHP 222
Los Angeles, CA 90089-9010
E-mail: zhaowuchen@gmail.com