Sample Size Calculation

Rahim Moineddin
Department of Family and Community Medicine
February 2017

Sample size estimation

- Determination of the appropriate sample size is a crucial part of a study design.
- If we make a study too small we may produce inconclusive results.
- At the same time we cannot waste limited resources on a study which is too large.
- Before undertaking a study, the investigator should first determine the minimum number of subjects (i.e., sample size estimation) that must be enrolled in order that the null hypothesis can be rejected if it is false.
- The ethical, reasons pertain to the risks of enrolling either an inadequate number of subjects or more subject's than the minimum necessary to reject the null hypothesis.

Sample size estimation

- Simulation is a method for sample size calculation.
- Simulation is sometimes necessary when statistical method is very complex or sample size calculation formula is not yet available.
- Steps for sample calculations are as follows:

Simulation Steps

- Decide on
  - Null hypothesis
  - A statistical method
  - Type I error (5%)
  - Power (80%) or type II error(0.20)
  - Important difference of interest (clinically significant difference, or effect size)
  - Initial sample size

Simulation Steps

- Write a computer program to generate at least 1000 data sets of size ‘n’ according to distribution under alternative hypothesis
- For each data set calculate test statistics of null hypothesis
- Keep number of times that test was rejected
- Percent of rejected tests is power for sample size ‘n’

Comparing Two Means

- Population mean for one group is m_1 and for other group is m_2
- Standard deviation for both groups is σ
- Type I error is α and type II error is β
- Sample size for each group is

$$n = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{\Delta^2} + \frac{Z_{1-\alpha/2}^2}{4}; \quad \Delta = \frac{\left|m_1 - m_2\right|}{\sigma}$$

Comparing Two Means

```
%let m1 = 1.0;
%let m2 = 1.5;
%let s  = 2.0;
%let a  = 0.05;
%let b  = 0.2;
data sample;
za=probit(1-&a/2);
zb=probit(1-&b);
m=2*&s**2*(za+zb)**2/(&m1-&m2)**2;
n=round(m,1);
run;
title"sample size for comparing two means &m1 with &m2 with s=&s";
proc print data=sample noobs; var n; run;
```

Comparing Two Means

```
data dat;
seed = 1267384;
do it = 1 to 1000;
  do n = 100 to 300 by 10;
    do i = 1 to n;
      x = &m1 + &s*rannor(seed);
        group=1;
      output;
      end;
    do i = n+1 to n+n;
      x = &m2 + &s*rannor(seed);
        group=2;
      output;
      end;
  end;
end;
run;
```

Comparing Two Means

```
proc ttest data=dat;
ods listing close;
by it n;
class group;
var x ;
ods output Ttests=t;
run;
data t; set t; if method='Pooled';
if probt lt 0.05 then sig=1; else sig=0;
run;
ods listing;
proc freq data=t; tables sig*n / nopercent norow; run;
```

Comparing Two Means (N=251)

sig   n

Frequency Col Pct	200	210	220	230	240	250	Total
0	300 30.00	275 27.50	239 23.90	247 24.70	199 19.90	189 18.90	2283
1	700 70.00	725 72.50	761 76.10	753 75.30	801 80.10	811 81.10	8717
Total	1000	1000	1000	1000	1000	1000	11000

sig   n

Frequency Col Pct	260	270	280	290	300	Total
0	200 20.00	170 17.00	158 15.80	171 17.10	135 13.50	2283
1	800 80.00	830 83.00	842 84.20	829 82.90	865 86.50	8717
Total	1000	1000	1000	1000	1000	11000

Comparing Two Proportions

$$n = \frac{\left\{ Z_{1-\alpha/2}\sqrt{\left[2\overline{\pi}(1-\overline{\pi}\right]} + Z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)} \right\}^2}{\left|\pi_1 - \pi_2\right|^2}$$

$$\pi = \frac{\pi_1 + \pi_2}{2}$$

Comparing Two Proportions

```
%let p1 = 0.20;
%let p2 = 0.40;
data dat;
seed=1;
do it = 1 to 1000;
  do n = 10 to 200 by 5;
    group = 1;
    call ranbin(seed, n, &p1, x);
        outcome = 'y';
        weight = x;
    output;
        outcome = 'n';
        weight = n - x;
        output;
    group = 2;
    call ranbin(seed, n, &p2, x);
        outcome = 'y';
        weight = x;
    output;
        outcome = 'n';
        weight = n - x;
        output;
  end;
end;
run;
```

Comparing Two Proportions

```
proc freq data=dat;
by it n;
ods listing close;
tables group*outcome/chisq;
weight weight;
ods output Chisq=t;
run;
data chisq; set t;
if statistic='Chi-Square';
if prob lt 0.05 then chi_square = 1 ; else chi_square=0;
run;
proc freq data = chisq;
tables chi_square*n / nopercent norow;
run;
```

Comparing Two Proportions Chi-squared test (N=81)

chi_square   n

Frequency Col Pct	50	55	60	65	70	75	80	85	Total
0	388 38.80	350 35.00	309 30.90	268 26.80	288 28.80	214 21.40	205 20.50	168 16.80	8471
1	612 61.20	650 65.00	691 69.10	732 73.20	712 71.20	786 78.60	795 79.50	832 83.20	30528
Total	1000	1000	1000	1000	1000	1000	1000	1000	38999

chi_square   n

Frequency Col Pct	90	95	100	105	110	115	120	125	Total
0	151 15.10	149 14.90	116 11.60	106 10.60	95 9.50	80 8.00	79 7.90	77 7.70	8471
1	849 84.90	851 85.10	884 88.40	894 89.40	905 90.50	920 92.00	921 92.10	923 92.30	30528

Comparing Two Proportions Chi-squared continuity corrected (N=91)

chi_square n

Frequency Col Pct	50	55	60	65	70	75	80	85	Total
0	457 45.70	439 43.90	393 39.30	350 35.00	349 34.90	266 26.60	266 26.60	214 21.40	10204
1	543 54.30	561 56.10	607 60.70	650 65.00	651 65.10	734 73.40	734 73.40	786 78.60	28795
Total	1000	1000	1000	1000	1000	1000	1000	1000	38999

chi_square n

Frequency Col Pct	90	95	100	105	110	115	120	125	Total
0	191 19.10	187 18.70	149 14.90	133 13.30	122 12.20	104 10.40	98 9.80	90 9.00	10204
1	809 80.90	813 81.30	851 85.10	867 86.70	878 87.80	896 89.60	902 90.20	910 91.00	28795
Total	1000	1000	1000	1000	1000	1000	1000	1000	38999

Sample size for difference of two proportions

- Rate of outcome for both placebo and treatment group is 50%.
- Reduction in placebo group is estimated 5%.
- Reduction in treatment group 20%.
- Effect size is 15%.
- Required sample size for a 95% confidence interval of size 10%.

Difference of two proportions

```
%let pi0=0.50; /* baseline intervention group */
%let pi1=0.30; /* end of study intervention group */
%let pc0=0.50; /* baseline control group */
%let pc1=0.45; /* end of study control group */
data dat;
seed=1;
do it = 1 to 1000;
 do n = 300 to 2000 by 20;
    call ranbin(seed,n,&pi0,i0);
    call ranbin(seed,n,&pi1,i1);
    call ranbin(seed,n,&pc0,c0);
    call ranbin(seed,n,&pc1,c1);
    ri0=i0/n; ri1=i1/n; rc0=c0/n; rc1=c1/n;
    diff_int = ri1-ri0; diff_con = rc1-rc0;
    effect = diff_con - diff_int;
    output;
 end;
end;
run;
```

Difference of two proportions

```
proc univariate data=dat noprint;
class n;
var effect;
output out=pout mean=effect pctlpre=ci_ pctlpts=2.5  97.5 pctlname=lower
    upper;
run;
data pout; set pout; ci_size=ci_upper - ci_lower; run;
proc print data=pout noobs ; var n effect ci_lower ci_upper ci_size;
format effect ci_lower ci_upper ci_size f4.2;
run;
```

Calculated sample size

n	effect	ci_lower	ci_upper	ci_size
300	0.15	0.04	0.26	0.22
400	0.15	0.06	0.25	0.19
500	0.15	0.06	0.24	0.17
600	0.15	0.07	0.23	0.15
700	0.15	0.08	0.22	0.15
800	0.15	0.08	0.22	0.14
900	0.15	0.09	0.21	0.13
1000	0.15	0.09	0.21	0.12
1100	0.15	0.09	0.21	0.11
1200	0.15	0.09	0.21	0.11
1300	0.15	0.10	0.20	0.11
1400	0.15	0.10	0.20	0.10
1500	0.15	0.10	0.20	0.10
1600	0.15	0.10	0.20	0.09
1700	0.15	0.10	0.20	0.09
1800	0.15	0.11	0.19	0.09
1900	0.15	0.11	0.19	0.09
2000	0.15	0.11	0.19	0.09

Logistic Regression

- Several specialized statistical packages compute power and sample size for logistic regression under various scenarios:
  - PASS 2000,
  - nQuery,
  - EGRET SIZ.
- Eugene Demidenko. Sample size determination for logistic regression revisited. *Statist. Med.* 2007; **26**:3385–3397
- http://www.dartmouth.edu/~eugened/power-samplesize.php

Logistic Regression

- Covariate X: binary
  - $P(x=1)=0.5$
- Alpha 0.05
- $P(Y=1|X=0)=0.25$
- Odds Ratio: 2
- Power 80%
  - $N=310$
- Power 90%
  - $N=416$

Logistic Regression

```
%let px = 0.50; /* P(X=1)=0.5 */
%let py = 0.25; /* P(Y=1|X=0)=0.25 */
%let OR = 2.0;
%let a  = 0.05; /* Alpha=0.05 */
data dat;
seed = 0;
p = &py*&or/(1-&py+&py*&or); /* P(Y=1|X=1) using odds ratio */
do it = 1 to 1000;
  do n = 100 to 1000 by 10;
    do i = 1 to n;
      CALL RANBIN(seed,1,&px,x);
        if x = 0 then call RANBIN(seed,1,&py,y);
        if x = 1 then call RANBIN(seed,1,p,y);
      output;
      end;
  end;
end;
run;
```

Logistic Regression

```
proc logistic data=dat desc;
by it n;
ods listing close;
model y=x;
ods output ParameterEstimates=t;
run;
data t; set t; if variable='x';
if probchisq lt 0.05 then sig=1; else sig=0;
run;
ods listing;
proc freq data=t; tables sig*n/nopercent norow; run;
```

sig n

Frequency Col Pct	100	110	120	130	140	Total
0	638 63.80	638 63.80	608 60.80	544 54.40	521 52.10	5262
1	362 36.20	362 36.20	392 39.20	456 45.60	479 47.90	4738
Total	1000	1000	1000	1000	1000	10000

Frequency Col Pct	150	160	170	180	190	Total
0	500 50.00	511 51.10	463 46.30	418 41.80	421 42.10	5262
1	500 50.00	489 48.90	537 53.70	582 58.20	579 57.90	4738
Total	1000	1000	1000	1000	1000	10000

sig n

Frequency Col Pct	200	210	220	230	240	Total
0	379 37.90	326 32.60	340 34.00	329 32.90	315 31.50	2974
1	621 62.10	674 67.40	660 66.00	671 67.10	685 68.50	7026
Total	1000	1000	1000	1000	1000	10000

Frequency Col Pct	250	260	270	280	290	Total
0	299 29.90	281 28.10	263 26.30	232 23.20	210 21.00	2974
1	701 70.10	719 71.90	737 73.70	768 76.80	790 79.00	7026
Total	1000	1000	1000	1000	1000	10000

sig n

Frequency Col Pct	300	310	320	330	340	Total
0	195 19.50	190 19.00	166 16.60	163 16.30	166 16.60	1542
1	805 80.50	810 81.00	834 83.40	837 83.70	834 83.40	8458
Total	1000	1000	1000	1000	1000	10000

Frequency Col Pct	350	360	370	380	390	Total
0	149 14.90	128 12.80	137 13.70	126 12.60	122 12.20	1542
1	851 85.10	872 87.20	863 86.30	874 87.40	878 87.80	8458
Total	1000	1000	1000	1000	1000	10000

sig n

Frequency / Col Pct	400	410	420	430	440	Total
0	105 10.50	105 10.50	75 7.50	89 8.90	78 7.80	793
1	895 89.50	895 89.50	925 92.50	911 91.10	922 92.20	9207
Total	1000	1000	1000	1000	1000	10000

Frequency / Col Pct	450	460	470	480	490	Total
0	82 8.20	65 6.50	67 6.70	67 6.70	60 6.00	793
1	918 91.80	935 93.50	933 93.30	933 93.30	940 94.00	9207
Total	1000	1000	1000	1000	1000	10000

sig n

Frequency Col Pct	500	510	520	530	540	Total
0	56 5.60	55 5.50	49 4.90	35 3.50	40 4.00	409
1	944 94.40	945 94.50	951 95.10	965 96.50	960 96.00	9591
Total	1000	1000	1000	1000	1000	10000

sig n

Frequency Col Pct	550	560	570	580	590	Total
0	48 4.80	41 4.10	28 2.80	38 3.80	19 1.90	409
1	952 95.20	959 95.90	972 97.20	962 96.20	981 98.10	9591
Total	1000	1000	1000	1000	1000	10000

Survival analysis

- Hull Hypothesis
  - H_0: $S_1(t)=S_2(t)$ for all t
  - H_0: $h_1(t)=h_2(t)$ for all t
  - H_0: $h_1(t)/h_2(t)=1$ for all t
  - H_0: HR=1 for all t assuming proportional hazards
- Test statistics
  - Log-rank test
  - HR estimated from Cox regression model

Survival analysis

- Effect size
  - Assuming proportional hazards, effect size is measured as Hazard Ratio HR= $h_1(t)/h_2(t)$
- Required number of events:
  - $\# \; events = \frac{\left(Z_{1-\alpha/2}+Z_{\beta}\right)^2}{\pi_1\pi_2(\ln(HR))^2}$ where π_1 and $\pi_2=1-\pi_1$ are the proportions to be allocated to group 1 and group2. Balance study $\pi_1=\pi_2=0.5$
- N=#events/Pr(event)
- Pr(event)=1-($\pi_1 S_1(t)+\pi_2 S_2(t)$)

Survival Analysis

- Exponential failure times
  - T is a continuous random variable with pdf f(t)
  - Cumulative distribution function F(t)=Pr(T<t)
  - Survival function=S(t)=1-F(t)
  - Hazard function f(t)/S(t)
- For exponential distribution with constant hazard ‘lambda’ if the incidence rate is lambda*t and S(t)=exp(-lambda*t)

- For exponential failure times, Incidence Rate (IR) is Lambda times time. For example if IR is 10 events per 100 person rate then PR is 0.1 per one person year. So for one year (t=1) we S(1)=exp(-0.1)=0.905 therefore 1-90.5%=9.5% of participants will have an event within one year.

Example

- One year study with equal allocation
- Suppose the IR for control group is 10 events per 100 person years
- Let Hazard Ratio HR=0.5
- S1(1)=exp(-0.1)=0.904837418
- S2(1)=exp(-0.1*0.5)=0.951229425
- P(event)=1-(0.9048+0.9512)/2=0.072
- Events=88 (using formula for 90% power)
- N=88/0.072=1223 or 1223/2=612 per group

- With the same assumptions for 6 months study we need 1190 participants per group.
- For handling loss n_{adj}=n/(1-loss)
- For 10% loss we need 612/0.90=680 per group

Simulation

- * Incidence rate for contro group is 10 per 100 person year
- * HR is 0.5;

```
%macro randexp(lambda);
    ((&lambda)*rand("Exponential"));
%mend;
```

Simulation

```
%let N=680;
data dat;
call streaminit(345);
lambda1=10/100 ;/* lambda for control*/
HR=0.5;
lambda2=lambda1*hr; /* lambda for treatment */
censorrate=0.10 ; /* rate of loss */
endtime=1; /* one year study */
do iter =1 to 1000; /* number of iterations */
do patid=1 to &n;
group='Control  ';
tevent=%randexp(1/lambda1);
c=%randexp(1/censorrate);
t=min(tevent, c, endtime);
censored=1-(c<tevent | tevent>endtime);
output;
group='Treatment';
tevent=%randexp(1/lambda2);
c=%randexp(1/censorrate);
t=min(tevent, c, endtime);
censored=1-(c<tevent | tevent>endtime);
output;
end;
end;
run;
```

Power for HR=0.5 and 90% power

power for N=500

significant	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	227	22.70	227	22.70
1	773	77.30	1000	100.00

power for N=600

significant	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	156	15.60	156	15.60
1	844	84.40	1000	100.00

power for N=680

significant	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	104	10.40	104	10.40
1	896	89.60	1000	100.00

Bootstrap

```
options nocenter;
data test;
seed=12345;
b0=5.0; b1=0.2;
do i=1 to 20;
age=min(abs(round(50*rannor(seed),1)),100);
y=b0+b1*age+7*rannor(seed); output;
keep age y ;
end;
run;
proc reg data=test;
model y = age;
run;
```

Regression

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.22365	2.52863	2.07	0.0535
age	1	0.21962	0.05986	3.67	0.0018

```
data test; set test;
seed=43;
rand=ranuni(seed);
run;
proc sort data=test; by rand; run;
data temp; set test; if _N_ le 5;
run;
```

Bootstrap

```
proc reg data=temp;
model y = age;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.07602	5.30771	1.33	0.2747
age	1	0.15128	0.10420	1.45	0.2425

```
%macro sim(datain=temp);
data all; run;
%do n=5 %to 10 %by 2;
proc surveyselect data= &datain out=boot
    seed = 1347 method = urs outhits rep = 1000 n=&n;
run;
proc glm data=boot;
by Replicate;
freq NumberHits;
model y = age/solution;
ods listing close;
ods output parameterestimates=p;
run;
data p; set p; if parameter='age';
if probt lt 0.05 then sig=1; else sig=0;
n=&n;
keep n sig;
run;
data all; set all p; run;
proc datasets; delete boot p; run;
%end;
ods listing;
proc freq data=all;
tables sig*n/norow nopercent;
run;
%mend;
%sim();
```

Frequency Col Pct	5	7	9	11	13	15	Total
0	534 53.40	381 38.10	231 23.10	124 12.40	86 8.60	69 6.90	1425
1	466 46.60	619 61.90	769 76.90	876 87.60	914 91.40	931 93.10	4575
Total	1000	1000	1000	1000	1000	1000	6000

Frequency Col Pct	5	7	9	11	13	15	Total
0	540 54.00	503 50.30	468 46.80	433 43.30	404 40.40	397 39.70	2745
1	460 46.00	497 49.70	532 53.20	567 56.70	596 59.60	603 60.30	3255
Total	1000	1000	1000	1000	1000	1000	6000

Table of sig by n

sig n

Frequency Col Pct	5	6	7	8	9	10	11	12	Total
0	5545 55.45	4537 45.37	3882 38.82	3260 32.60	2671 26.71	2428 24.28	2134 21.34	1867 18.67	35706
1	4455 44.55	5463 54.63	6118 61.18	6740 67.40	7329 73.29	7572 75.72	7866 78.66	8133 81.33	124294
Total	10000	10000	10000	10000	10000	10000	10000	10000	160000

(Continued)

Table of sig by n

sig n

Frequency Col Pct	13	14	15	16	17	18	19	20	Total
0	1698 16.98	1453 14.53	1322 13.22	1191 11.91	1047 10.47	976 9.76	892 8.92	803 8.03	35706
1	8302 83.02	8547 85.47	8678 86.78	8809 88.09	8953 89.53	9024 90.24	9108 91.08	9197 91.97	124294
Total	10000	10000	10000	10000	10000	10000	10000	10000	160000

Longitudinal study

- A **longitudinal study** is a research study that involves repeated observations of the same items over time
- Longitudinal studies are used in medicine to uncover predictors of certain diseases.
- In a hypertension study diastolic blood pressure of placebo and treatment group are measured at baseline (time=0), then after 2 and 5 years.

Longitudinal study

- Variance of Y_{ij} is 100 and it is expected that the difference between trends of placebo and treatment group reaches 0.5 mmHg
- Type I error is 0.05 and power is 80%.
- Correlation among each subject measurements is estimated to be 0.5 (ICC)

Longitudinal Study

Generate data from a multivariate normal distribution

PURPOSE:

The %MVN macro generates multivariate normal data using the Cholesky root of the variance-covariance matrix.

REQUIREMENTS:

Version 6 or later of SAS/IML software.

%inc "<location of your file containing the MVN macro>";

Following this statement, you may call the %MVN macro.

The following parameters are required except for SEED=:

VARCOV= SAS data set that contains the variance-covariance matrix.

MEANS= SAS data set that contains the mean vector.

N= Number of observations to generate.

SEED= Starting seed value for the random number generator.

SAMPLE= SAS data set name for the resulting multivariate normal data.

LIMITATIONS: No error checking is done.

Longitudinal Study

```
data me; input means;
datalines;
0
0
0
;
run;
data vare; input v1-v3;
datalines;
100     50      50
50      100     50
50      50      100
;
run;
```

Longitudinal Study

```
%macro sim(vare=, rho=, n=);
%do it=1 %to 1000;
* simulate error terms for control group;
  %mvn(version, varcov=vare, means=me, n=&n, seed=0, sample=control);
  data control; set control;
    group='control  ';
    id=_N_;
    time=0;
    eij=col1;
    output;
    time=2;
    eij=col2;
    output;
    time=5;
    eij=col3;
    output;
    keep id group eij time ;
 run;
```

Longitudinal Study

```
* simulating error terms for treatment group;
 %mvn(version, varcov=vare, means=me, n=&n, seed=0,
     sample=treatment);
 data treatment; set treatment;
   group='treatment';
   id=_N_;
   time=0;
   eij=col1;
   output;
   time=2;
   eij=col2;
   output;
   time=5;
   eij=col3;
   output;
 keep id group eij time ;
 run;
```

Longitudinal Study

```
data dat; set control treatment;
   t=time;
   if group='control'  then yij=eij;
   else if group='treatment' then yij=0.5*time+eij;
run;
proc mixed data=dat;
class group t;
model yij = group | time/s;
repeated t / subject=id type=cs;
ods listing close;
ods output Tests3=test3;
*ods output SolutionF=e;
run;
```

Longitudinal Study

```
data test3; set test3;
   if effect='time*group' ;
   if probf lt 0.05 then significant=1;
   else significant=0;
run;
%if &it=1 %then %do; data p; set test3;run;%end;
%else %do; data p; set p test3; run; %end;
%end;
ods listing;
title"Power for n=&n";
proc freq data=p; tables significant; run;
%mend;
```

Longitudinal Study

```
*%sim(vare=100, rho=0.2, n=150);
%sim(vare=100, rho=0.5, n=135);
*%sim(vare=100, rho=0.8, n=40);
```

Longitudinal Study

Power for n=135

significant	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	163	16.30	163	16.30
1	837	83.70	1000	100.00

Power for n=55

significant	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	224	22.40	224	22.40
1	776	77.60	1000	100.00

Segmented Regression

- There are two regression line
  - $X_t = \beta_0 + \beta_1 t + e_t$ for $t < T$ and
  - $X_t = \alpha_0 + \alpha_1 t + e_t$ for $t \geq T$
  - Define It=0 for t<T and It=1 for t≥T
  - Regression $X_t = \beta_0 + \beta_1 t + \beta_2 I_t + \beta_3 I_t t\text{-after} + e_t$
- In this equation β_2 is difference between two intercept and β_3 is the difference between two slopes.

```
%let b0=1;
%let b1=0.5;
%let b2=0.0;
%let b3=0.3;
%let s=1.0;
```

```
data dat;
seed=1234;
do nsim=1 to 1000;
do n=5 to 15 by 2;
do t=1 to 2*n;
time=t;
it=(t>n);
time_after=(time-n)*it;
xt=&b0+&b1*time+&b2*it+&b3*time_after+&s*rannor(seed);
output;
end;
end;
end;
keep nsim n time time_after xt it;
run;
```

```
proc sort data=dat; by n nsim; run;
proc reg data=dat;
ods listing close;
by n nsim;
model xt=time it time_after;
ods output parameterestimates=p;
run;

data p; set p;
if variable in('it' 'time_after');
if probt lt 0.05 then significant=1; else significant=0;
run;
```

```
ods listing;
options nodate nonumber nocenter;
title"Power b0=&b0  b1=&b1  b2=&b2  b3=&b3";
proc freq data=p;
tables variable*significant*n/nopercent norow;
run;
```

Power b0=1 b1=0.5 b2=0.0 b3=0.0

Controlling for Variable=it

significant n

Frequency / Col Pct	5	7	9	11	13	15	Total
0	97 97.00	93 93.00	92 92.00	96 96.00	94 94.00	96 96.00	568
1	3 3.00	7 7.00	8 8.00	4 4.00	6 6.00	4 4.00	32
Total	100	100	100	100	100	100	600

Controlling for Variable=time_after

significant n

Frequency / Col Pct	5	7	9	11	13	15	Total
0	96 96.00	95 95.00	96 96.00	94 94.00	97 97.00	98 98.00	576
1	4 4.00	5 5.00	4 4.00	6 6.00	3 3.00	2 2.00	24
Total	100	100	100	100	100	100	600

Power b0=1 b1=0.5 b2=1.5 b3=0.0

Controlling for Variable=it

significant n

Frequency Col Pct	5	7	9	11	13	15	Total
0	90 90.00	85 85.00	67 67.00	64 64.00	59 59.00	50 50.00	415
1	10 10.00	15 15.00	33 33.00	36 36.00	41 41.00	50 50.00	185
Total	100	100	100	100	100	100	600

Controlling for Variable=time_after

significant n

Frequency Col Pct	5	7	9	11	13	15	Total
0	96 96.00	95 95.00	96 96.00	94 94.00	97 97.00	98 98.00	576
1	4 4.00	5 5.00	4 4.00	6 6.00	3 3.00	2 2.00	24
Total	100	100	100	100	100	100	600

Power b0=1 b1=0.5 b2=1.5 b3=0.4

Controlling for Variable=it

significant n

Frequency Col Pct	5	7	9	11	13	15	Total
0	90 90.00	85 85.00	67 67.00	64 64.00	59 59.00	50 50.00	415
1	10 10.00	15 15.00	33 33.00	36 36.00	41 41.00	50 50.00	185
Total	100	100	100	100	100	100	600

Controlling for Variable=time_after

significant n

Frequency Col Pct	5	7	9	11	13	15	Total
0	95 95.00	81 81.00	48 48.00	25 25.00	3 3.00	1 1.00	253
1	5 5.00	19 19.00	52 52.00	75 75.00	97 97.00	99 99.00	347
Total	100	100	100	100	100	100	600

A real example

- Annual data is available form 2001 to 2005 as the proportion of a given service
- There was an intervention in 2006
- Objective: Assess the impact of intervention
- Maximum available number of chart for review is 80 in each year
- Chart review is very time consuming
- How many charts should be reviewed ?

First service

Year	n	Stated in report
2001	1	1
2002	8	4
2003	2	1
2004	2	2
2005	4	3
Total	17	11

Second service

Year	n	Stated in report
2001	1	0
2002	8	4
2003	2	0
2004	2	0
2005	4	0
Total	17	4

SAS Codes

```
%macro sim(it=, b=0, effect=0.0, out=);
%do it=1 %to &it;
 data dat;
 seed=-1;
 do n=10 to 80 by 10;
   do p=0.1 to 0.5 by 0.1;
     do year = 0 to 4;
                  int=0;
*                 yy=year;
                  p0=min(p+&b*year,0.9);
       y = ranbin(seed, n, p0);
       output;
                end;
                year=5;
                int=1;
*               yy=year;
     p0=min(p+&b*year+&effect, 0.95);
                y = ranbin(seed, n, p0);
                output;
   end;
 end;
 drop seed;
 run;
```

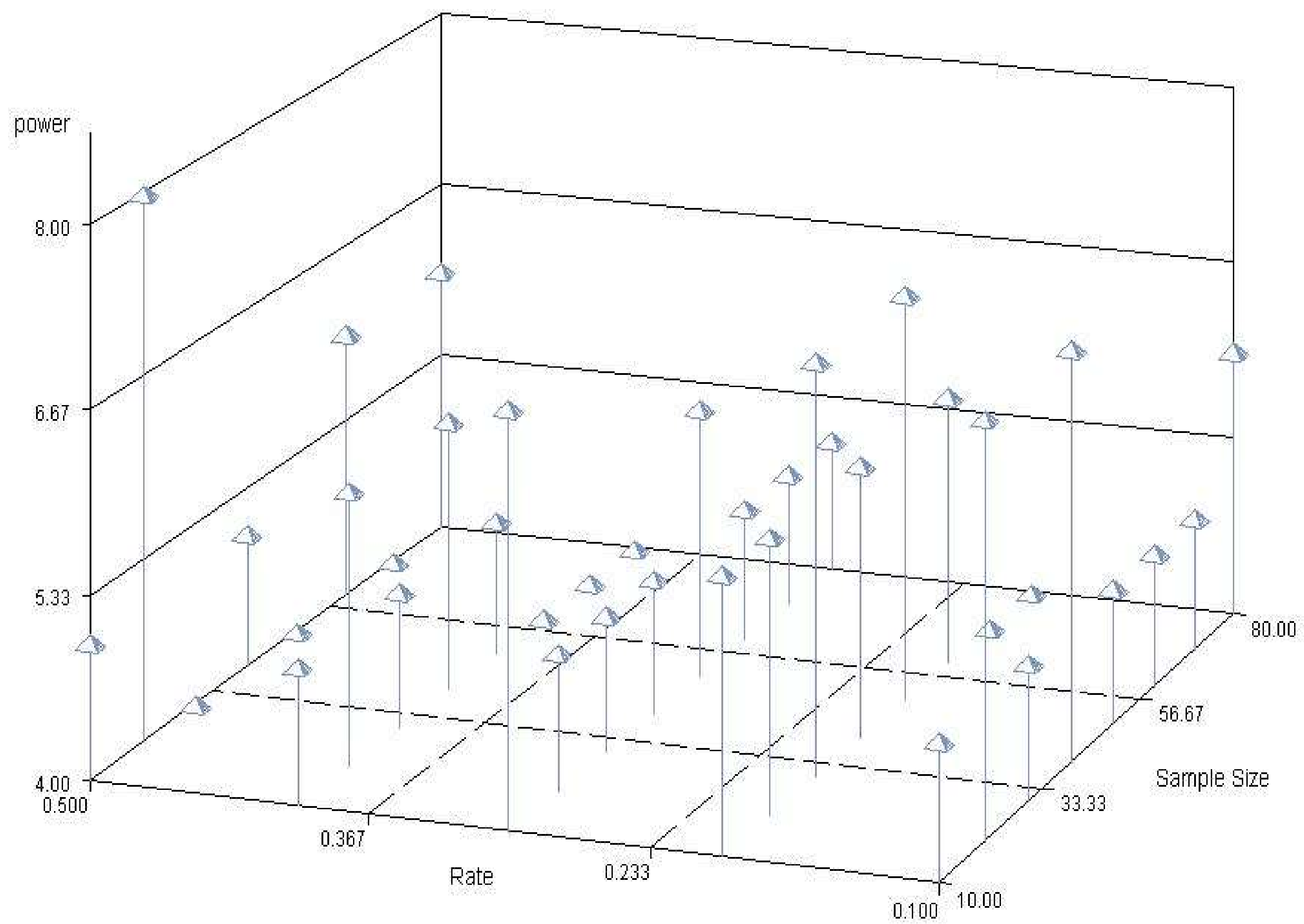
```
proc genmod data=dat desc;
  ods listing close;
  class yy;
  by n p;
  model y/n=year int /d=b link=logit type3;
*repeated subject=yy /type=ar(1);
  ods output Type3=e;
  run;
  data e; set e; it=&it;
  sig=0;
  if source='int';
  if probchisq lt 0.05 then sig=1;
  run;
  %if &it=1 %then %do; data o; set e; run; %end;
  %else %do; data o; set o e; run; %end;
%end;
```

```
ods listing close;
proc sort data=o; by p n; run;
ods listing ;
title"power for Effect Size=&effect";
proc freq data=o;
tables p*sig*n/nopercent norow out=&out;
run;
ods listing close;
proc freq data=o; by p n;
tables sig/nopercent norow out=&out;
run;
data &out; set &out; effect=&effect; power=percent/100;run;
%mend;
```

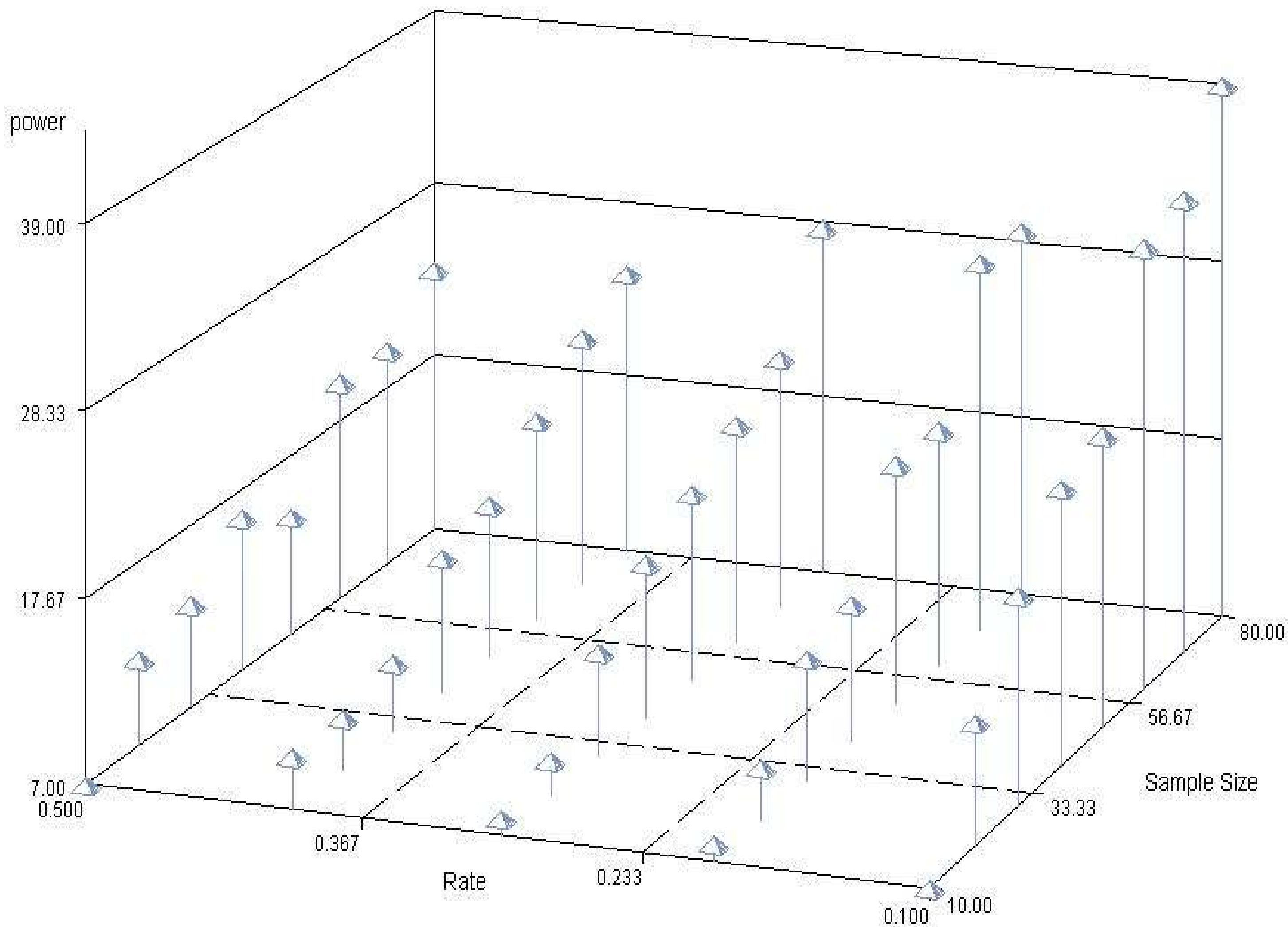
```
%sim(it=500, b=0.0, effect=0.0, out=out1);
%sim(it=500, b=0.0, effect=0.10, out=out2);
%sim(it=500, b=0.0, effect=0.20, out=out3);
%sim(it=500, b=0.0, effect=0.30, out=out4);
data out1; set out1; effect=0.0; power0=power;run;
data out2; set out2; effect=0.10; power10=power; run;
data out3; set out3; effect=0.20; power20=power;run;
data out4; set out4; effect=0.30; power30=power;run;
data out; set out1 out2 out3 out4;if sig=1; rate=p;run;
data pp; merge out1 out2 out3 out4; by p n sig; if sig=1;
keep p n power0 power10 power20 power30;
run;

proc sort data=out; by effect p n;run;
```

```
goptions reset=all border;
title "Power Plot";
*footnote j=r "GTDSURFA";
 proc g3d data=out;by effect;
   scatter rate*n=power/grid yaxis=axis1;
run;
quit;
```

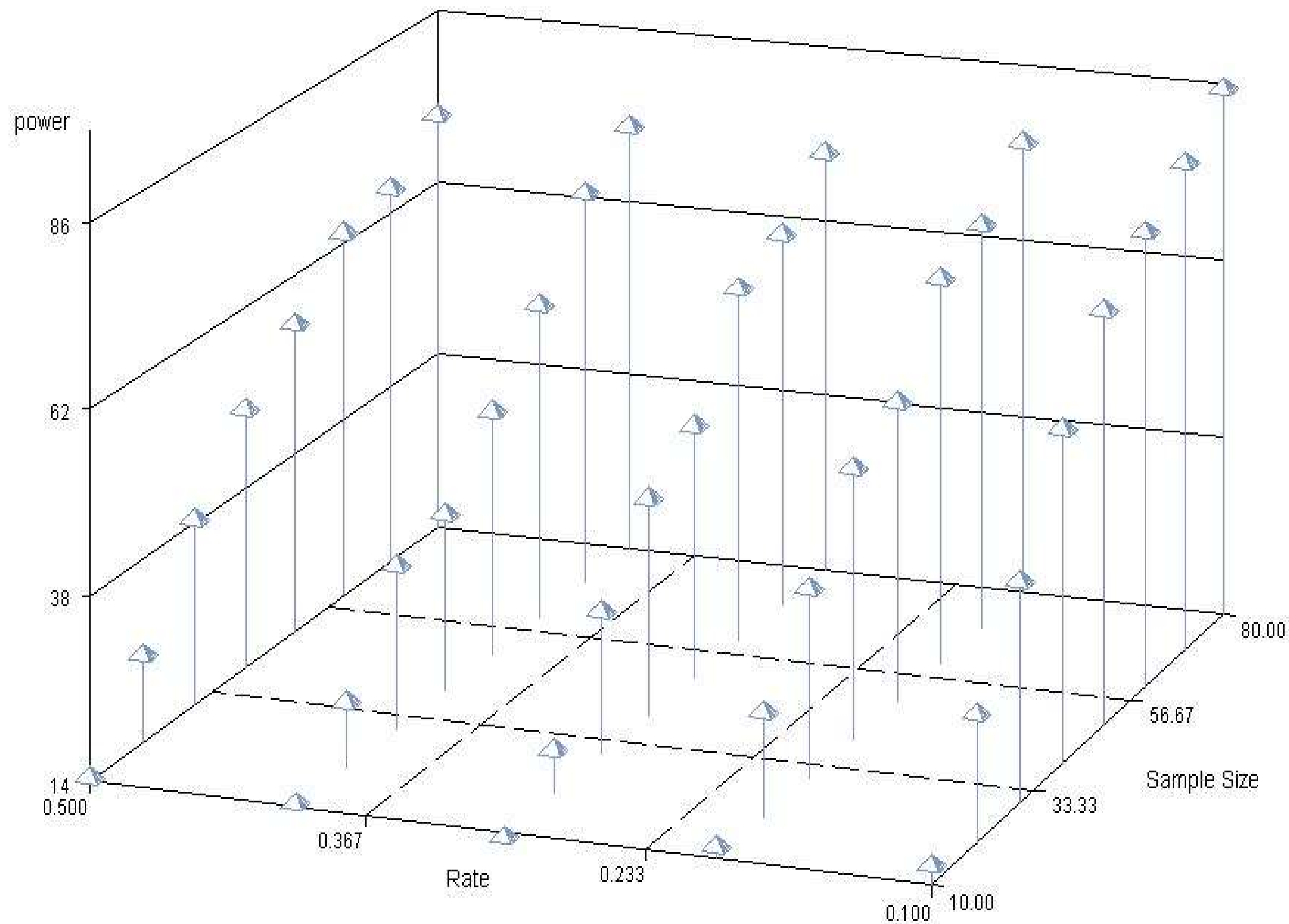
Power Plot for Effect Size 0%

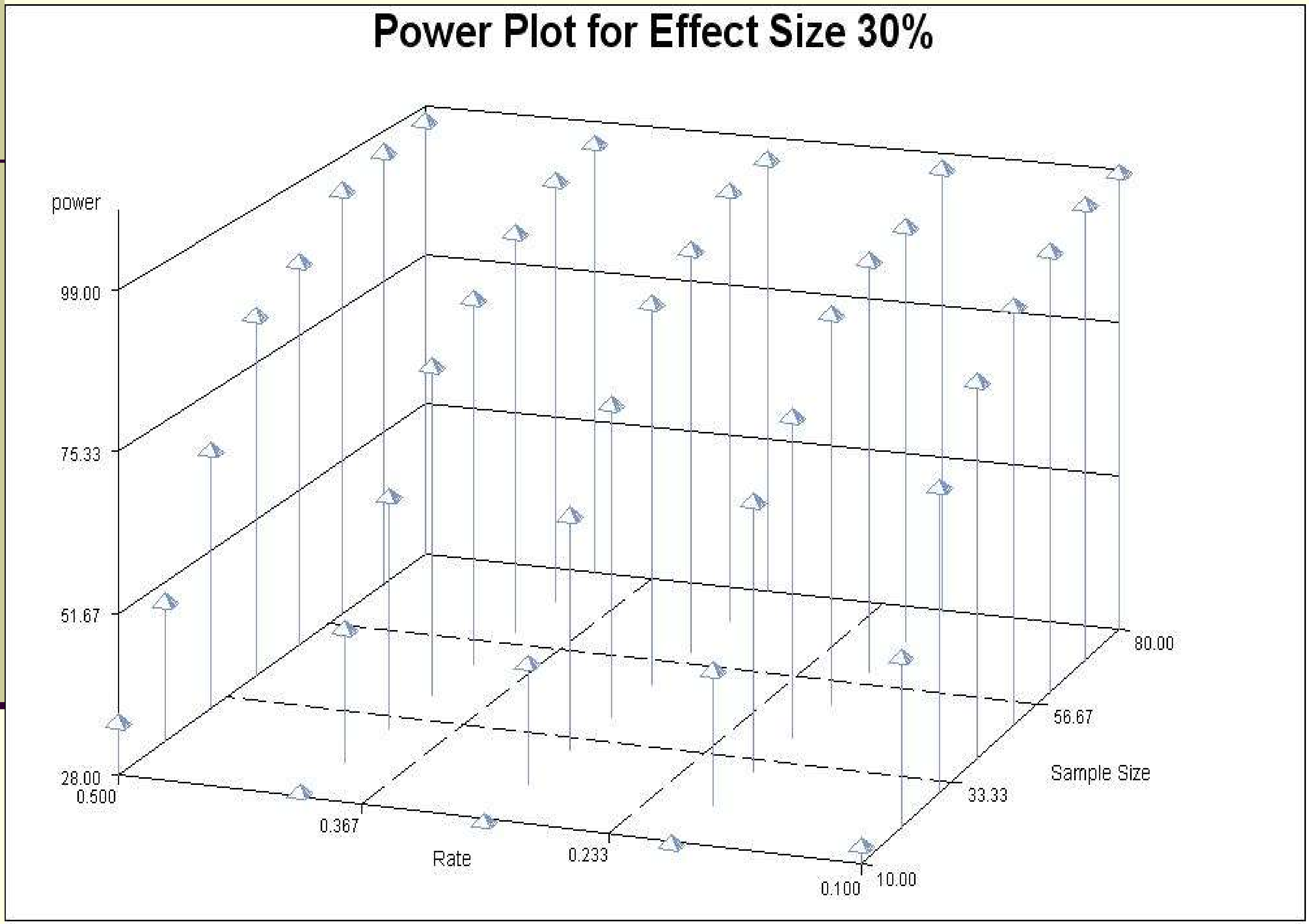


Power Plot for Effect Size 10%



Power Plot for Effect Size 20%




Power Plot for Effect Size 30%
power
99.00
75.33
51.67
28.00
0.500
0.367
Rate
0.233
0.100
10.00
33.33
Sample Size
56.67
80.00

Sample size and power analysis

- Consider two-level regression model

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$$

combined form

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} + e_{ij}$$

Sample size and power analysis

- Power of test for significance of individual level regression coefficients depend on total sample size (level 1 and level 2).
- Power of test for level 2 and cross-level interactions depends MORE strongly on the number of groups than on the total sample size.
- For accuracy and high power a large number of groups is more important than a large number of individuals per group.

Sample size and power analysis

- Increasing sample size at all levels make estimates and their standard errors more accurate
- 30/30 rule
  - 30 groups with 30 individuals pre group for being on the safe side
- 50/20 rule
  - If cross-level interaction is the parameter of interest
- 100/10 rule
  - If variance covariance components are the parameters of interests

Sample size and power analysis

- Simulation
  - The accuracy of the estimates for the fixed and random parameters for different sample sizes can be investigated by simulation.
  - For example we can assess the accuracy of the parameter estimates for two level logistic model with
    - Number of Groups: 30, 50, 100
    - Group Size: 5, 30, 50
    - ICC: 0.1, 0.2, 0.3

Sample size and power analysis

Steps

1. Obtain the initial estimates for fixed and random parameters
2. Run the simulations for 1000 replications
3. Calculate the percent bias for all parameter estimates
4. Flag whether the 95% confidence internal for each parameter contains the true parameter or not (coverage indicator)
5. Calculate the summary statistics for bias and non-coverage indicators

Sample size and power analysis

- Simple two level logistic model

$$\text{logit}(p_{ij}) = \pi_{0j} + \pi_{1j}\text{x}_{ij}$$

$$\pi_{0j} = \gamma_{00} + \gamma_{01} z_j + u_{0j}$$

$$\pi_{1j} = \gamma_{10} + \gamma_{11} z_j + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{bmatrix} \right)$$

Sample size and power analysis

```
%let g00 = -1;
%let g01 = 0.3 ;
%let g10 = 0.3 ;
%let g11 = 0.3 ;
%let s0 = 1.0;
%let s1 = 1.0;
```

Sample size and power analysis

```
%macro sim(nsim=, out=);
data test ;
seed=1;
do rep =1 to &nsim ;
  do nregion = 30, 50, 100 ;
    do region = 1 to nregion;
        zj = rannor(seed) ;
        u0j=&s0 * rannor(seed) ;
        u1j=&s1 * rannor(seed) ;
        p0j=&g00+&g01*zj+u0j ;
        p1j=&g10+&g11*zj+u1j ;
    do n = 5, 30, 50;
      do id = 1 to n ;
```

Sample size and power analysis

```
      xij = rannor(seed);
      eta=p0j + p1j * xij ;
       p=exp(eta)/(1+exp(eta));
       yij=ranbin(seed,1,p);
       output ;
      end;
     end;
   end;
  end;
end;
keep rep region n id zj yij xij nregion ;
run ;
```

Sample size and power analysis

```
proc sort data=test ; by rep nregion n; run;
proc nlmixed data=test ; by rep nregion n;
  ods listing close ;
  parms g00=&g00 g10=&g10 g01=&g01 g11=&g11 s0=&s0
     s1=&s1 s01=0;
  p0j = g00 + g01 * zj + u0j;
  p1j = g10 + g11 * zj + u1j;
  eta = p0j + p1j * xij ;
  p = exp(eta)/(1+exp(eta));
  model yij ~ binary(p);
```

Sample size and power analysis

```
random u0j u1j ~ normal([0,0],[s0, s01, s1]) subject=region;
  ods output ParameterEstimates = pest ;
  ods output ConvergenceStatus = conv ;
run ;
data &out ; merge pest conv ; by rep nregion n;
run ;
%mend sim;

%sim(nsim=1000, out=sout);
```

Sample size and power analysis

```
%macro ana(datain=, su0=, su1=, out=);
data &out; set &datain ;
if standarderror = . or status=1 then conv=0;
else conv=1;
if parameter ne 's01';
if parameter='g00' then p_rel_bias=100*estimate/&g00 - 100;
if parameter='g01' then p_rel_bias=100*estimate/&g01 - 100;
if parameter='g10' then p_rel_bias=100*estimate/&g10 - 100;
if parameter='g11' then p_rel_bias=100*estimate/&g11 - 100;
if parameter='s0 ' then p_rel_bias=100*estimate/&su0**2 - 100;
if parameter='s1 ' then p_rel_bias=100*estimate/&su1**2 - 100;
lowercl=estimate-1.96*standarderror;
uppercl=estimate+1.96*standarderror;
flag=1;
if parameter='g00' and lowercl <= &g00 <= uppercl then flag=0 ;
if parameter='g01' and lowercl <= &g01 <= uppercl then flag=0 ;
if parameter='g10' and lowercl <= &g10 <= uppercl then flag=0 ;
if parameter='g11' and lowercl <= &g11 <= uppercl then flag=0 ;
if parameter='s0 ' and lowercl <= &su0**2 <= uppercl then flag=0 ;
if parameter='s1 ' and lowercl <= &su1**2 <= uppercl then flag=0 ;
run;
%mend;
```

# of groups	group size	ICC	Rate of convergence	γ_{00}	γ_{01}	λ_{10}	γ_{11}	σ_{00}	σ_{11}
30	5	0.1	56	8.77	11.12	15.85	13.26	174.04	55.49
		0.2	68	4.75	11.56	10.70	14.93	24.25	54.55
		0.3	76	3.94	12.22	5.91	14.89	15.82	54.02
	30	0.1	94	0.07	1.09	-1.93	3.57	-7.07	-6.81
		0.2	100	-0.08	3.70	-1.60	3.44	-5.89	-7.02
		0.3	100	-0.18	5.74	-1.72	5.31	-3.55	-6.92
	50	0.1	99	-0.39	0.69	-0.18	5.32	-8.47	-7.71
		0.2	100	-0.43	2.65	-1.70	5.74	-6.25	-7.16
		0.3	100	-0.39	4.62	-2.85	4.49	-5.05	-7.30
50	5	0.1	71	3.82	9.32	4.93	5.88	110.80	35.84
		0.2	86	1.44	8.40	2.34	4.73	11.95	25.55
		0.3	90	1.00	5.96	3.50	6.68	6.80	28.11
	30	0.1	99	0.09	0.21	2.01	2.72	-5.16	-2.60
		0.2	100	-0.32	0.28	2.06	3.69	-2.90	-2.55
		0.3	100	-0.37	-0.56	1.89	3.46	-3.24	-2.81
	50	0.1	100	-0.23	0.08	1.25	2.31	-6.18	-3.54
		0.2	100	-0.33	0.51	1.39	2.56	-3.76	-3.59
		0.3	100	-0.53	0.83	1.29	1.73	-3.19	-3.23
100	5	0.1	87	1.64	0.14	2.42	1.55	47.87	7.64
		0.2	98	0.47	-0.50	1.11	0.89	1.84	3.46
		0.3	98	0.95	0.25	2.25	0.41	2.23	8.82
	30	0.1	100	-0.02	-0.12	0.31	0.41	-5.13	-1.25
		0.2	100	-0.11	0.36	-0.06	0.96	-2.14	-1.17
		0.3	100	-0.21	1.04	-0.96	1.07	-2.06	-1.30
	50	0.1	100	0.03	0.35	-0.09	0.48	-3.28	-2.01
		0.2	100	-0.06	0.71	0.05	0.63	-2.12	-1.34
		0.3	100	-0.14	0.62	0.50	1.18	-1.73	-1.52

# of groups	group size	ICC	γ_{00}	γ_{01}	λ_{10}	γ_{11}	σ_{00}	σ_{11}
30	5	0.1	0.029	0.048	0.032	0.032	0.142	0.045
		0.2	0.049	0.046	0.040	0.028	0.090	0.041
		0.3	0.042	0.046	0.041	0.055	0.081	0.030
	30	0.1	0.061	0.051	0.070	0.068	0.095	0.111
		0.2	0.060	0.067	0.066	0.064	0.117	0.104
		0.3	0.053	0.061	0.066	0.062	0.107	0.111
	50	0.1	0.053	0.061	0.064	0.071	0.104	0.127
		0.2	0.062	0.063	0.063	0.063	0.108	0.116
		0.3	0.054	0.059	0.071	0.071	0.113	0.114
50	5	0.1	0.031	0.045	0.050	0.048	0.117	0.038
		0.2	0.051	0.041	0.040	0.044	0.070	0.056
		0.3	0.062	0.049	0.040	0.046	0.076	0.064
	30	0.1	0.050	0.067	0.059	0.064	0.091	0.092
		0.2	0.060	0.055	0.064	0.060	0.088	0.083
		0.3	0.062	0.060	0.059	0.057	0.094	0.076
	50	0.1	0.065	0.058	0.056	0.069	0.091	0.091
		0.2	0.065	0.060	0.053	0.070	0.102	0.091
		0.3	0.067	0.064	0.066	0.062	0.087	0.088
100	5	0.1	0.042	0.037	0.038	0.043	0.172	0.052
		0.2	0.046	0.058	0.055	0.047	0.070	0.078
		0.3	0.056	0.052	0.048	0.048	0.082	0.084
	30	0.1	0.045	0.046	0.057	0.055	0.085	0.061
		0.2	0.046	0.064	0.053	0.055	0.086	0.064
		0.3	0.039	0.058	0.049	0.050	0.087	0.066
	50	0.1	0.063	0.059	0.055	0.042	0.086	0.075
		0.2	0.053	0.063	0.052	0.050	0.079	0.067
		0.3	0.053	0.056	0.051	0.051	0.069	0.076