

Sample Size and Power

Paula Diehr

5-6-2004

Before the study, estimate the approximate number of subjects required to achieve a specified goal.

Approximate because:

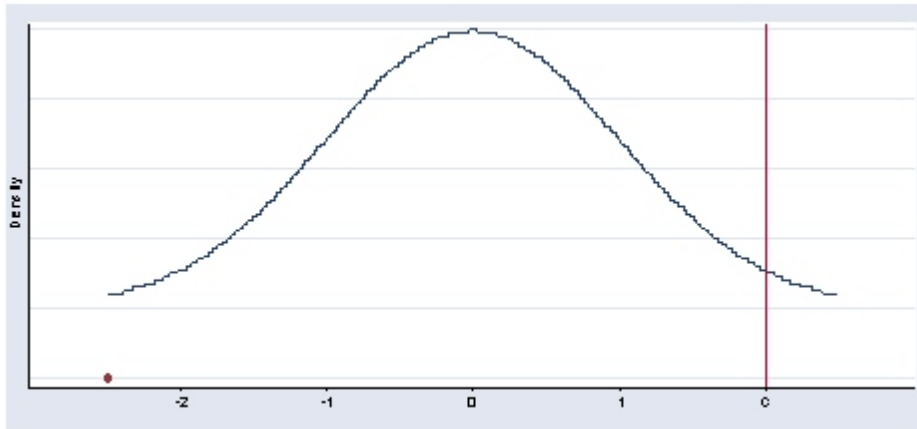
- Poor data for estimates of parameters
- Normality assumptions (rarely a problem)
- Murphy's Law
- etc.

Outline

- I Easy part: the formulas
 - Hypothesis Testing
 - Estimation
 - Software
 - Simulation
- II Harder part: the data to put into the formulas
- III Sample size for Cluster Randomized Trials
- IV Summary

Simple part, the formulas

Review of the normal distribution.



The area to the left

of Z_c is c

The area to the left of $Z_{1-\alpha}$ is $1-\alpha$

$$Z_{.975} = 1.96$$

$$Z_{.95} = 1.645$$

$$Z_{.80} = .84$$

$$Z_c = -Z_{1-c}$$

Review of Hypothesis Testing

H_0 Null Hypothesis $\mu_1 = \mu_2$

H_1 Alternative Hypothesis $\mu_1 \neq \mu_2$

ERRORS:

	Truth= H_0	Truth = H_1
Conclusion:		
H_0	-	Type II (β)
H_1	Type I (α)	-

Pr (Type I error) = α (we can select, .05)

(multiple comparisons?)

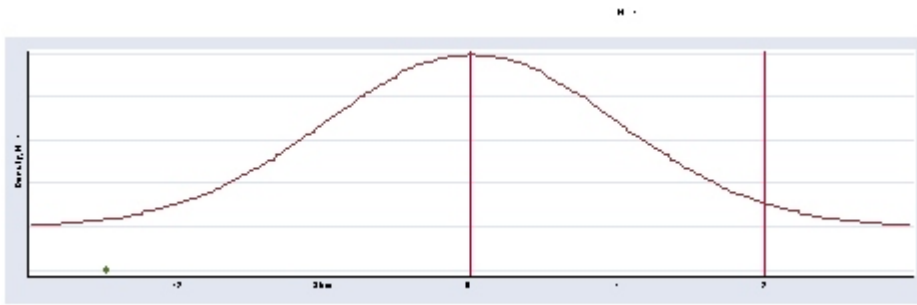
Pr (Type II error) = β (we can select by making N large enough, .8)

$1 - \beta$ = power = probability that we detect an effect (reject H_0) when there really is one.

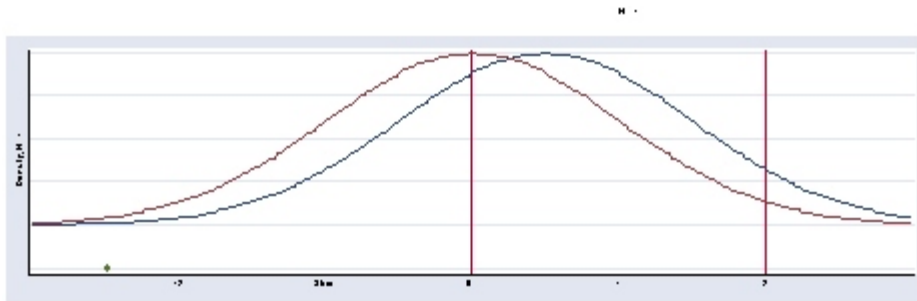
This is why sample size is important

Cohen says mean power ~ .4 in studies ($\beta = .6$), so most were hopeless from the start.

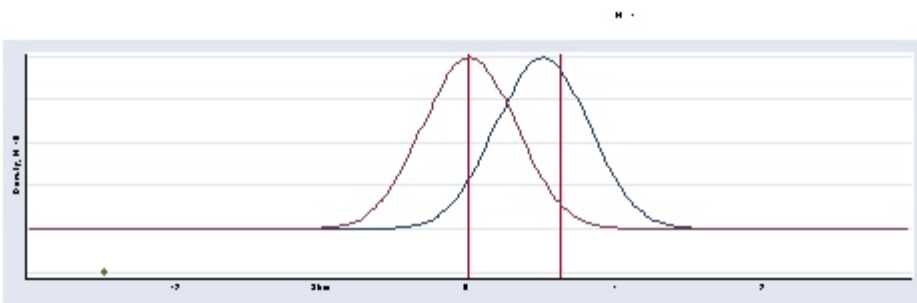
The idea



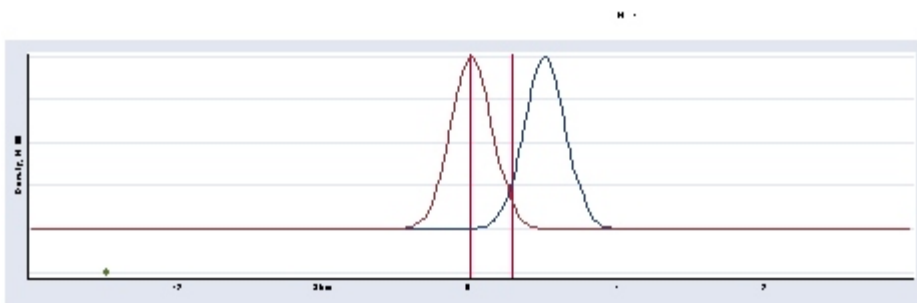
$N(0,1)$



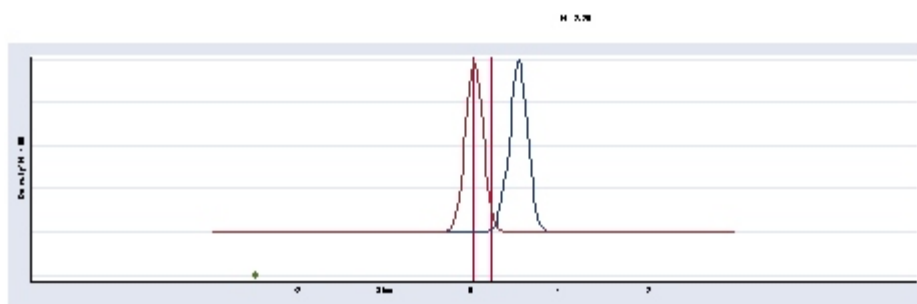
$N(0,1)$ Null
 $N(.5,1)$ Alternative
 (N=1)



dist'n of
 $\frac{\bar{x}_1 - \bar{x}_2}{s / \sqrt{N}}$
 (N=10)



(N=50)



(N=100)

Magic Formula:

2 groups

$$D = \mu_1 - \mu_2$$

$$(Z_{1-\alpha} + Z_{1-\beta})^2 = \frac{D^2}{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

(use $1-\alpha/2$ for 2-tailed tests)

Example Data

Number of **visits in 14 months** in a combined population of an HMO (GHC) and an indemnity plan (KCM) in the early 1970's. (1689 subjects)

		<u>Pop</u>	<u>Sample</u>
mean =	4.5	μ	\bar{x}
sd =	6.3	σ	s

Proportion of people hospitalized in one year

variance = $p(1-p)$

GHC	.05	Var=.05*.95, sd=.22
KCM	.10	Var=.10*.90, sd=.30

Examples

Continuous Variable: Plan a new study to look for differences in Visits/year between a new HMO and a new indemnity plan. Want enough power to detect small but important differences in visits between the two plans.

What's an "important" difference?

1 visit per year?

$D=1$

GH vs KC
 $\mu = 5$ $\mu=4$
 $\sigma=6.3$ (estimated from old data)

Magic Formula (one-tailed test):

$$(Z_{1-\alpha} + Z_{1-\beta})^2 = \frac{D^2}{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

If $N_1 = N_2 = N$, (Equal #'s in each group), solve for N as:

$$N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 (s_1^2 + s_2^2)}{D^2}$$

Want 80% power $\beta=.2$ (why 80%?)

$$Z_{1-\beta} = Z_{.80} = .8416$$

For **equal sample sizes**, with $N_1=N_2=N$,
 if we also assume $s_1=s_2=6.3$, solve for N

$$N = [(1.645+.84)^2(2)(6.3)^2] / D^2$$

(1-tail) [use 1.96 instead of 1.645 for 2-tail]

D	--N needed per group---	
	<u>1-tail</u>	<u>2-tail</u>
.1	49,018	62,234
.5	1,961	2,489
1.	490	622

"Gospel"?
 (consider the source)

Inflate to adjust for potential drop-outs

For a 2-tailed test with $\alpha=.05$, $\beta=.2$, power = .80, a simplified **generic sample size formula** from VanBelle is:

$$\text{let effect size} = \Delta = D/s = 1/6.3 = .16$$

$$N \sim 16/\Delta^2 = \mathbf{635}$$

For **unequal sample sizes**,
 suppose $N_1 = c N_2$ where "c" is known from some other source.

Magic Formula:

$$(Z_{1-\alpha} + Z_{1-\beta})^2 = \frac{D^2}{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

Replace N_1 with cN_2 , solve for N_2 ,
 results in

$$N_2 = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \left(\frac{s_1^2}{c} + s_2^2 \right)}{D^2}$$

$$N_2 = [(1.645 + .84)^2 (6.3)^2 (1/c + 1)] / D^2$$

Let $D = 1$.

$$N_2 = 245.09(1/c + 1)$$

if $c=1$, [equal sample sizes] then $(1/c+1)=2$, $N_2=N_1= 490$ (as above), or
980 total

if $c=2$, $(1/c+1)=1.5$,

$$N_2=367.5, N_1=(2)(N_2)=735$$

$$735+368= \mathbf{1103} > \mathbf{980}$$

Rule of thumb, don't gain much if $c > 3$

Other cases: N_1 is fixed (say, 100 is all the cases there are), solve for N_2
 (there isn't always a solution - you may need to settle for lower power)

Examples: Binary Variable

Prob(admission) GHC \sim .05 $p_1 = .05$
 KCM \sim .10 $p_2 = .10$

$$s_1^2 = p_1(1-p_1) = .0475$$

$$s_2^2 = .09$$

If $N_1 = N_2 = N$,

$$N = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 (s_1^2 + s_2^2)}{D^2}$$

$$N = (1.645 + .84)^2 (.0475 + .09) / (.05)^2 = \mathbf{339.64} \text{ per group}$$

If we'd assumed $s_1^2 = s_2^2 = .075(.925)$, $N = \mathbf{342.72}$

Central limit theorem guarantees that \hat{p} is normally distributed for large N , so this formula works fine.

Magic Formula has Four parameters:

- D Minimum Clinically Important Difference
- σ s.d. of dependent variable
- β probability of Type II error
- N Sample Size

N is often known (budgetary, only a certain number of subjects available)

Given N, solve for D? (**minimum detectable difference**)

$$N = (Z_{1-\alpha} + Z_{1-\beta})^2 2 s^2/D^2$$

$$D = \text{MDD} = (Z_{1-\alpha} + Z_{1-\beta}) (2 s^2/N)^{.5}$$

Minimum detectable effect size

$$\Delta = D/s = (Z_{1-\alpha} + Z_{1-\beta})(2/N)^{.5}$$

Given N, D, solve for β ?

$$Z_{1-\beta} = + (D/s)(N/2)^{.5} - Z_{1-\alpha}$$

if D=1 visit, s=6.3, $Z_{1-\alpha} = 1.645$,

If N=100 per group,

$$Z_{1-\beta} = -.5227, \quad 1-\beta = .3015 \text{ (from normal table).}$$

If N=490,

$$Z_{1-\beta} = .828, \quad 1-\beta = .80, \text{ as above.}$$

Everything is a t-test

compare **two means**

magic formula

compare **two proportions**

proportions are normally distributed, magic formula

logistic regression ---- sample size needed to detect an odds ratio of 1.5

-reformulate as a problem in comparing two proportions—

make a guess at “p” in one group, choose the other p such that the relative risk is 1.5, then use the magic formula for proportions to get the sample size

OR: Logistic Regression: Hsieh. Sample size tables for logistic regression. *Statistics in Medicine* 8:795-802. 1989.

OR: special software (see below)

survival analysis —

-reformulate as a problem comparing two proportions – make a guess at the proportion who will die in each group with the “average” follow-up, use the magic formula

OR: special software (see below)

Other:

Cohen J. t , r , r_1-r_2 , p , p_1-p_2 , χ^2 for contingency tables, anova, ancova, regression

OR: special software (see below)

Misplaced precision.

- Poor quality of the numbers that go into the estimate,
- Calculate power for a very simple (but related) analysis (t-test, comparing proportions).
- Claim that the more complex analysis planned will probably have even more power. (E.g., will eventually do logistic regression, but do power calculations as a difference in proportions or chi-square).

Multiple Regression.

-Calculate sample size as if you were doing a t-test on the outcome variable (perhaps comparing high to low), then claim that power of the regression analysis will be “even larger” because you will use covariates, which will probably (but not necessarily) reduce variability and make it easier to get significant results

-If there are similar regressions in the literature, including values of R^2 : if s^2 is the estimated variance of Y, then $(1-R^2)*s^2$ is the estimated variance of Y after controlling for the other X's. Use square root of $(1-R^2)*s^2$ instead of s in the magic formula.

OR: special software

Software:

Spreadsheet

program the magic formula (t, z, F all available as functions)

Palm Pilot

caslrt / palm stat (magic formula)

Stata (help, search for sample size)

sampsi (magic formula)

case-control studies

kappa

survival

cluster samples

case/parent, matching designs (genetics)

etc.

Freeware program, wdist95 can be downloaded at

<http://ourworld.compuserve.com/homepages/MSVonTress/wdist.htm>

UCLA has a user-friendly simple free power calculator at

<http://calculators.stat.ucla.edu/powercalc/>

normal, binomial, exponential, poisson, correlation coefficient, power given sample size, sample size given power.

[chapter 2 of statistical rules of thumb, http://www.vanbelle.org/.](http://www.vanbelle.org/)

<http://www.vanbelle.org>

<http://www.zoology.ubc.ca/%7Ekrebs/power.html>

Clinical trials (Southwest Oncology Group):

<http://www.swogstat.org/statoolsout.html>

Estimation versus Testing?

Estimation Problem: How big a sample is needed to estimate a parameter to a given accuracy?

1-sample.

95% confidence interval for μ is

$$\bar{X} \pm 1.96 \frac{s}{\sqrt{N}}$$

To estimate μ to within $\pm L$,

$$L = 1.96 s/\sqrt{N}$$

$$N = (1.96)^2 s^2 / L^2$$

μ = mean visits/yr, $\sigma=6.3$

$L=.1$, $N = 15,247$ needed to estimate # of visits to within ± 0.1 !

.5	610
1.0	152

2-sample, 95% ci is

$$\bar{X}_1 - \bar{X}_2 \pm 1.96 \sqrt{s_1^2 / N_1 + s_2^2 / N_2}$$

With equal values of N and s , $N = 2 (Z_{1-\alpha})^2 s^2 / L^2$

$L=.1$ $N = 30,494$ per group (Twice as many as one-sample c.i.)

.5	1,220
1.0	305

Proving the Null Hypothesis (Equivalence Trials)

Goal is to show "no difference"

Can always do that (fail to achieve significance) by using a small N

H_0 probably not true. With big enough N, even tiny differences will be significant. "How big a difference" is more relevant.

Better to design study to estimate $D = \mu_1 - \mu_2$ to a certain amount of accuracy. Then, with $N = 1,220$ per group, estimated difference in visits does not differ by more than .5 visits per year.

$$\bar{X}_1 - \bar{X}_2 \pm 1.96 \sqrt{s_1^2 / N_1 + s_2^2 / N_2}$$

Suppose in the actual data, at the end of the trial, the estimate of D is .3, and $s=6.4$, $N=1220$. The 95% CI is:
 $.3 \pm 1.96 (2(6.4^2)/1220) = (-.207, +.808)$.

We can conclude that there is not a significant difference between the two plans (the confidence interval includes zero).

You can also state that any difference is probably not larger than .808 or smaller than -.207. This is a stronger statement. Since we had previously decided that differences smaller than 0.5 were unimportant, we have not demonstrated equivalence, since $.808 > .500$. (Choose "N" to make the eventual c.i. small enough).

Hard part, the data

What if σ is **unknown**? [the usual situation]

- 1 **Pilot study** to estimate σ
- 2 **Literature Search (or data)**
a-this variable, this setting
b-this variable, similar setting
c-similar variable (VA, SIP)
- 3 Re-formulate the problem with respect to **proportions**. (example- instead of mean number of visits, use proportion with more than 5 visits). Investigators often have insight about what these proportions would be in the two groups. Once the p's are guessed at, the variance is just $p(1-p)$, use formulas for binomial sample. Revised question may be just as meaningful.
- 4 Re-formulate with respect to "**Effect Size**", $\Delta = D/\sigma$.
$$N = (Z_{1-\alpha} + Z_{1-\beta})^2 \frac{2 (s^2/D^2)}{\Delta^2}$$

$$= (Z_{1-\alpha} + Z_{1-\beta})^2 \frac{2}{\Delta^2}$$

So, don't necessarily need to know σ if we "know" the effect size.

Cohen denotes effect sizes as:

	<u>Δ</u>	<u>Overlap of 2 distns.</u>	<u>Similar to Difs in IQs</u>
small	.2	85%	ident twins
medium	.5	65%	prof vs mgr
large	.8	51%	frosh vs PhD

"80% power to detect a medium effect size" (Desperation)

Data should fit the situation

In data set mentioned above, there were 218 hospital admissions for 1689 people. This would suggest using

$$p=218/1689 \text{ and } s^2 = p(1-p)=.1124.$$

But, in fact, only 167 people had admissions, with an average of 1.3 admissions per person admitted. That is, the dependent variable (at the person level) is not binomial.

$$\text{mean}=.1291, s^2 =.21437$$

The sample sizes calculated from those two different estimates of s^2 differ by a factor of

$$.214376/.1124 = 1.91.$$

The sample size would differ by a factor of almost 2.

Based on this, it might be unwise to analyze admissions per person. It might be better to code each person as "ever admitted, yes/no", and treat this as a binomial variable.

This sample size calculation would have led to a change in the planned analysis.

Hard to find data for parameter estimates:

variance of changes over time

variances of cluster data

variances of changes over time in cluster data

Some Data Don't Exist Yet

Effect on mental health status of the "managedness" of the depressed patient's insurance plan. [Grembowski].

Some data are available on dist'n of mental health scores, possibly even on changes in scores, probably not for the group we want to study.

The "managedness" index for a plan was "to be developed" as a complex factor score from many variables, and we had no idea what it would be like.

We instead formulated the problem as a 2-group problem: the percent who improved in the "highly managed" vs "not-so-managed" plans, cutpoint chosen to ensure equal numbers in each group. Divide the 2000 available patients into 1000 high, 1000 low. The worst case is that 50% improve over time under the null hypothesis ($p=.5$). Given this amount of information we can solve for D . With $\beta = .20$, $D = MDD = .055$. That is, we have 80% power to detect a significant difference if the improvement in one group is about .50 and the other is .555 or more. That is, a 5.5 percentage point difference in improvement rates (or more) can be detected with 80% power (or more).

DHEP Example.

Goal: Improve HbA1c in dually eligible (Medicare+Medicaid) diabetics.
What sample size do we need?

No data available at all:

Assume the worst (most variable) case, that 50% of controls had HbA1c >7, design study to detect a difference of Δ in the groups at follow-up. (2 proportions)

$$D = .1, .2, .3 \quad N = 407, 103, 45$$

Probably 60% have HbA1c > 7 (2 proportions)

$$D = .1, .2, .3 \quad N = 407, 106, 48$$

Mean and sd of HbA1c in the literature (Compare means at f/u).

Mean	SD	D = 1.0	D = .5	D = .25	D = .1
7.4	2.2	76 = N	304	1216	7611
7.5	1.6	39	157	627	3920
7.8	1.9	58	234	936	5849

Choose study with mean or population description closest to what is expected. Or choose largest s.d. to be conservative. No covariates.

Mean and sd of DIFFERENCE in HbA1c in the literature.

A) same blood sent to two labs; (r = .998)

B) two measurements 1 year apart; (r = .57) (study data)

Mean	SD	D = 1	D = .5	D = .25	D = .1
A) -.07	.15	1 = N	1	3	18
B) -.25	1.91	58	234	936	5849

$\text{Var}(X_1 - X_2) = 2 \sigma^2 (1-\rho)$; no point using change score unless $\rho > .5$ — if not, might as well do “post-test only”

Summary

TESTING?

[is it different from ...?]

OR

ESTIMATION?

[give an accurate estimate of difference]

(confidence interval)

Power, β

What do you know?

D,

s,

$D/s = \Delta$ = effect size

N_1, N_2

β

What do you want to estimate? Use the Magic Formula, solve for what you don't know:

$$(Z_{1-\alpha} + Z_{1-\beta})^2 = \frac{D^2}{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

Assume: test statistic normally distributed, which is true for reasonably high N. If $N < 20$, say, do something to make it more conservative. More “exact” methods available for special cases, but not usually needed.

Cluster Randomized Trials

Example: “VA facility” Intervention (ACQUIP trial, Fihn et al.) To improve quality of care.

Assign K VA clinics at random to treatment, K to control. The SF-36 will be measured over time for a sample of patients in each clinic, over time. (Actual trial was firm-based, 2 per VA, this is preliminary work).

How many clinics and how many patients per clinic and how many measures per patient?

Sample size:

of VA's

of people sampled per VA [all?]

of times they are surveyed

K clusters (clinics)

N observations per cluster (patients)

Variance among cluster means is σ_c^2

Variance among people (within cluster) is σ_E^2

If there are K clusters with N patients per cluster, then the grand mean has variance $(\sigma_c^2 + \sigma_E^2/N)/K$.

Design depends on both K and N. If σ_c^2 is large, then increasing n does not help much. Need to increase K (get more clusters).

$$\text{Intraclass correlation} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_E^2}$$

$$\text{Design Effect (Deff)} = 1 + \text{ICC} * (\text{B}-1)$$

Calculate # of people needed as a person-level study, multiply N by the design effect.

“Similar data” from trial of frail veterans using the SIP.

σ^2_c = true variance among clusters. (But for SIP, estimated $\sigma^2_c = 0$).

$$\sigma^2_c = 0 \quad \text{power} = .80 \quad \text{n surveys} = 5$$

MINIMUM DETECTABLE DIFFERENCE (Koepsell, Donner, Murray)

Clinics

PER GROUP -----# OF PATIENTS PER Firm PER TIME-----

	250	500	1000	2000	4000
4	3.22*	2.28	1.61	1.14	.81
6	2.40	1.70	1.20	.85	.60
8	2.00	1.42	1.00	.71	.50
10	1.76	1.24	.88	.62	.44
16	1.35	.96	.67	.48	.34

* With 4 clinics per group, 250 patients per clinic, can detect a difference between groups as small as 3.22 SIP points with 80% power and 5 survey occasions.

$$\sigma^2_c = 6.6$$

Clinics

PER GROUP -----# OF PATIENTS PER Firm PER TIME-----

	250	500	1000	2000	4000
4	8.52	8.20	8.06	7.96	7.91
6	6.34	6.10	5.98	5.93	5.88
8	5.29	5.10	4.99	4.94	4.92
10	4.64	4.47	4.38	4.33	4.31
16	3.56	3.44	3.37	3.33	3.32

Effect must be bigger than in previous table.

Not much improvement by increasing number of patients

C3P example (new VA trial).

Goal: improve management of chronic stable angina (in patients with ischemic heart disease)

Intervention: randomize patients or doctors (?) to tx/control, try to improve care.

Data: from patients, criterion = Seattle Angina Questionnaire (SAQ). We **did have similar data** from an earlier study in different sites!! “N” is known approximately from VA data for the new sites, so the question is how big a difference can be detected with this design.

Patient-Level Minimum Detectable Difference Calculation:

Site	N of Patients	N of Providers
Site 1	15000	61
Site 2	10847	65
Site 3	12755	77
Site 4	11852	95
Number of Patients in 4 sites		N= 50454
45 percent have IHD		N= 22704.3
72 percent will participate		N= 16347.1
32 percent will have SAQfreq<70		N= 5231.071
50 % Attrition		N= 2615.535
50 percent in Tx, control Group		N= 1307.768

Outcome Variables

Change in SAQ frequency score in 1 year (for those whose initial score was <70)

Mean change (ACQUIP)	10.85	MDD (80% power)	2.503035	points
SD (ACQUIP)	22.8591			
N (new study) (yr 0-1 or half of yr 0 - 2).	1308	MDD (90% power)	2.898157	points

Randomizing providers. Provider must have at least 2 eligible patients, to allow for attrition, non-participation.

Sample size and power for new study– Provider level

date 10/23/2002

TX group only, measures 3,4,5 used to calculate slope.

Site	N of Providers
Site 1	61
Site 2	65
Site 3	77
Site 4	95

Number of Providers in 4 sites N= 298

# docs	ACQUIP		NEW STUDY
402		(Eligible means first frequency score < 70 and has a slope estimate)	
	191 47.51244	#, % with >= 1 eligible patient	N= 141.5871
		50 percent in Tx, control Group	N= 70.79353
	124 64.92147	#, % with >= 2 eligible patients	N= 45.9602

Outcome Variables

Change in SAQ frequency score in 1 year (for those whose initial score was <70)

Mean change per provider (ACQUIP)	-11	MDD (80% power)	9.298762	points
SD among providers (ACQUIP)	15.92			
N docs (in new study)	46	MDD (90% power)	10.76664	points

Patient-randomized trial seemed more powerful, since MDD is smaller (2 vs 10). If every doctor had exactly the same # of patients and there was no “intracluster correlation”, the MDD would have been identical in the two analyses.

The “**design effect**” is $[15.92^2/46] / [22.86^2 / 1308]$
 $= 5.50 / .40 = 13.8 = (9.298/2.50)^2$ (could use to design a new study)
 $= deff = 1+ICC (B-1)$

$B = 1308/46 = 28.3$ patients per doc on average.

$ICC = .486$ (assuming all docs had the same # of patients).

Simulation Approach

Create a statistical model that includes the important features of your data. Generate "1000" different samples. Analyze each sample the way you plan to analyze the real data. See what % of the samples give a significant result. That is the estimated power for this design.

For example, if we didn't have the ACQUIP database, but knew only some summary statistics, we could simulate data as follows:

Generate 300 records, one for each provider in the new study.

Each provider has a # of patients distributed $\sim N(55,109) = \underline{N}$
_____ (Or log normal, 3.2, 1.61), not Poisson, maybe neg expon

45% have IHD

50% will participate

If $\underline{n} = \underline{N} * .45 * .50 < 2$, then this provider is not in the study, drop record.

Each provider has an over-all effect on change $\sim N(0,20) = \delta$

Their patients' mean change is $\sim N(11, 23/\underline{n}) = \underline{C}$

Random number 1/0 to determine tx/control status, = **TX**

The provider's mean score is $\delta + \underline{C} + D * \underline{TX}$

Do a t-test on the mean scores, see if it's significant.

Repeat "1000" times

Count the % of times that t-test was significant

This is the estimated power of the proposed analysis.

Repeat for different values of D or whatever else needs to be varied.

```

*CRUDE SIMULATION IN STATA

*****start of program
program sampsize, rclass
version 8.0
args delta
drop _all
*start with 300 providers
set obs 300

*N is the total number of patients ~(MEAN=55, SD=109)
*gen N = invnorm(uniform())*109+55
**log normal, mean 55 and sd 109.
gen N = exp( invnorm(uniform())*1.27+3.2)

*n is the number of eligible and willing patients
gen n = .45*.5 * N
*** PROVIDER MUST HAVE 2 OR MORE PATIENTS
drop if n < 2
*D is the provider effect ~ n(0, 20)
gen D = invnorm(uniform())*20+ 0
** here is the mean change for the patients ~ n(11, 23/ROOT n)
gen C = invnorm(uniform())*23/n^.5 + 11
gen TX = uniform() < .5
*gen delta = 5
gen score = D + C + `delta' * TX

summ

ttest score, by(TX) unequal
return list
return scalar p = r(p)

end
*****end of program

*****runs the program
***100 reps

**here is the simulation
**delta = 5, rule is drop n<2
set seed 1235
simulate "sampsize 5" mean = r(p) , reps (100)
describe
gen pct = mean<.05
summ
*** REJECTS NULL HYPOTHESIS 38% OF THE TIME, so 38% power if delta = 5

**delta = 7
set seed 1235
simulate "sampsize 7" mean = r(p) , reps (100)
*describe
gen pct = mean<.05
summ
*** REJECTS NULL HYPOTHESIS 73% OF THE TIME

**and so on.
8, 76% OF THE TIME
10, 93% OF THE TIME
15, 100% of the time

plot the power curve

```

Another Simulation Example. [Diehr, work in progress]

Question: What sample size is needed for a test-retest to obtain a “good” estimate of the reliability coefficient for a health status instrument?

The Models for Z (True Health) and Y (Health Status Instrument)

Z = True Health

$$\begin{array}{l}
 Z_0 \\
 Z_1 \quad Z_0 \\
 Z_2 \quad Z_1 \quad +
 \end{array}
 \begin{array}{l}
 N(\mu_z = 50, \sigma_z = 10) \\
 \\
 N(\mu_{\text{trend}} = 1, \sigma_{\text{trend}} = 1) \\
 \text{Secular Trend}
 \end{array}
 \begin{array}{l}
 \\
 \\
 + \Delta = \text{Treatment effect} = 3 \\
 \text{Treatment}
 \end{array}$$

Y = INSTRUMENT

$$\begin{array}{l}
 y_0 \quad Z_0 \quad + \quad \epsilon ; \quad \epsilon \sim N(M, SD) \quad \mathbf{M=0; SD = 0,1,2,5,10} \\
 y_1 \quad Z_1 \quad + \quad \epsilon ; \quad \epsilon \sim N(M,SD) \\
 y_2 \quad Z_2 \quad + \quad \epsilon ; \quad \epsilon \sim N(M, SD)
 \end{array}$$

“For the simulations we generated data described by the model in Table 1, with varying values of SD (1,2,5,10) and N per treatment group (20, 40, 100, 200,500,1000,2000), with 150 replicates for each set of parameters. We estimated Reliability from each sample. We calculated the standard error of the Reliability estimates. We regressed the log of the sample size on the standard error and the true value of the parameter. The N needed to achieve a specified accuracy for the estimated Reliability can be estimated as:

$$N = \exp[69.4 - 1.75 * \ln(s_{\text{Reliability}}) - 75.1 * \text{Reliability} + 45.6 * \ln(\text{Reliability}) + .619^2/2].$$

For example, suppose the instrument is expected to have Reliability about .8, and we wish the estimate of Reliability to fall between .7 and .9 with 95% probability, or the confidence interval to have length 0.20. Assuming normality of the estimates, about 95% of the estimated Reliability values will fall in the range Reliability $\pm 2 * s_{\text{Reliability}}$, meaning that the value of $s_{\text{Reliability}}$ must be 0.05. From the equation, the required N is 100; that is, about 100 people would be needed for a test-retest study to estimate the Reliability to this level of accuracy.”

Summary

Is Estimating Sample Size really important?

- 1 N is often the only parameter the investigator does know.
- 2 Results may be sobering
 - MAY change the design
 - OR, just change D or β to make the sample size you can afford look sufficient.
(lower your expectations)
“Game” the system?
- 3 Solving for D or β for an affordable value of N may be more useful than calculating sample size.
- 4 Everything is approximate
- 5 At least it forces you to think about the design of the study in explicit terms.
- 6 "Calling card"

References.

- Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Assoc., Publishers. New Jersey.
- Donner A. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol* 114:906-14. 1981. Requires data not often available.
- Donner A, Klar A. *Design and analysis of cluster randomization trials in health research*. Arnold. London. 2000.
- Fihn SD, McDonnell MB, Diehr P, Anderson SM, Bradley KA, Au DH, Spertus JA, Burman M, Reiber GE, Kiefe CI, Cody M, Sanders KM, Whooley MA, Rosenfeld K, Baczek LA, Sauvigne A. Effects of sustained audit/feedback on self-reported health status of primary care patients. 2004. *Am J Med*. Feb 116:241-8.
- Grembowski D, Diehr P, Novak L , Roussel A , Martin DP, Patrick DL, Williams B, Ulrich CB. Measuring the “managedness” and covered benefits of health plans. *Health Services Research*. 2000. 35:707-734.
- Koepsell T, Martin D, Diehr P, Psaty B, Wagner E, Perrin E, Cheadle A: Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: A mixed-model analysis of variance approach. *Journal of Clinical Epidemiology* 44:701-713, 1991.
- Koepsell T, Wagner E, Cheadle A, Patrick D, Kristal A, Allan-Andrilla CH, Dey L, Martin DC, Diehr P. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annual Review of Public Health* 1992. 13:31-57
- Murray DM. *Design and analysis of group-randomized trials*. Oxford University Press New York. 1998.