JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Methods in Sample Surveys
140.640
3rd Quarter, 2009

Sample Size and Power Estimation

Saifuddin Ahmed, PHD

Biostatistics Department

School of Hygiene and Public Health

Johns Hopkins University

Sample size and Power

"When statisticians are not making their lives producing confidence intervals and p-values, they are often producing power calculations"

Newson, 2001

”In planning of a sample survey, a stage is always reached at which a decision must be made about the size of the sample. The decision is important. **Too large** a sample implies a waste of resources, and **too small** a sample diminishes the utility of the results.“

Cochran, 1977

Sample size estimation: Why?

- Provides validity of the clinical trials/intervention studies – in fact any research study, even presidential election polls
- Assures that the intended study will have a desired power for correctly detecting a (clinically meaningful) difference of the study entity under study if such a difference truly exists

Sample size estimation

- ONLY two objectives:
  - Measure with a precision:
    - Precision analysis
  - Assure that the difference is correctly detected
    - Power analysis

First objective: measure with a precision

- Whenever we propose to estimate population parameters, such as, population mean, proportion, or total, we need to estimate with a specified level of precision
- **We like to specify a sample size that is sufficiently large to ensure a high probability that errors of estimation can be limited within desired limits**

Stated mathematically:

- we want a sample size to ensure that we can estimate a value , say, ***p*** from a sample which corresponds to the population parameter, ***P***.

- Since we may not guarantee that p will be exact to ***P***, we allow some error

- Error is limited to certain extent, that is this error should not exceed some specified limit, say ***d***.

- We may express this as:

  p - P = $\pm$ ***d***,

  i.e., the difference between the estimated p and true ***P*** is not greater than ***d*** (allowable error: margin-of-error)

- But do we have any confidence that we can get a ***p***, that is not far away from the error of $\pm$***d***?

- In other words, we want some confidence limits, say 95%, to our error estimate d.

  That is ***1-*** α = 95%

It is a common practice: α-error = 5%

In probability terms, that is,

$$prob\ \{-d \le p - P \le d\} \ge 1 - \alpha$$

In English, we want our estimated proportion *p* to vary between *p-d* to *p+d*, and we like to place our confidence that this will occur with a *1-*α probability.

From our basic statistical course, we know that we can construct a confidence interval for p by:

$$p \pm z_{1-\alpha/2} * se(p)$$

where z_α *denotes a value on the abscissa of a standard normal distribution (from an assumption that the sample elements are normally distributed) and se(p) =* σ_p is the standard error.

$$p \pm d = p \pm z_{1-\alpha/2}\sigma_p$$

Hence, we relate p ± d in probabilities such that:

$$d = Z_{1-\alpha/2}\sigma$$

$$= Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

If we square both sides,

$$d = Z_{1-\alpha/2}\sigma$$
$$= Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$d^2 = Z^2_{1-\alpha/2} \frac{p(1-p)}{n}$$

$$n = \frac{Z^2_{1-\alpha/2}}{d^2} p(1-p)$$

$$n = \frac{Z^2_{1-\alpha/2} p(1-p)}{d^2}$$

For the above example:

$$n = \frac{(1.96)^2 * 0.4 * 0.6}{(.10)\wedge 2} = 92.2 \approx 93$$

Note that, the sample size requirement is highest when p=0.5. It is a common practice to take p=0.5 when no information is available about p for a conservative estimation of sample size.

As an example, p = 0.5, d = 0. 05
(5% margin-of-error), and α-error = 0.05:

$$n = \frac{(1.96)^2 * 0.5 * 0.5}{(.05)\text{^}2} = 384.16 = 385 \approx 400$$

```
. di 1.96^2*.5*(1-.5)/(.05^2)
384.16
. di (invnorm(.05/2))^2*.5*(1-.5)/(.05^2)
384.14588
```

Stata

```
. sampsi .5 .55, p(.5) onesample

Estimated sample size for one-sample comparison of proportion
  to hypothesized value

Test Ho: p = 0.5000, where p is the proportion in the population

Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.5000
 alternative p =   0.5500

Estimated required sample size:

             n =      385
```

Sample Size Estimation for Relative Differences

If d is relative difference,

$$n = \frac{t^2 p(1-p)}{(d^* p)2}$$

$$= \frac{t^2 (1-p)}{d^2 p}$$

Consider that 10% change is relative to p=.40 in the above example.

Then, d = 0.4*0.10=0.04, that is, p varies between 0.36 to 0.44. Now,

$$\mathrm{n} = \frac{(1.96)^2 * 0.4 * 0.6}{(.10 * 0.4)\hat{}2} = 576.24 \approx 577$$

Note, *d* is very sensitive for sample size calculation.

Change the variance

Sample Size for Continuous Data

$$n = \frac{t^2 \sigma^2}{d^2}$$

Sources of variance information:

- Published studies
  - (Concerns: geographical, contextual, time issues – external validity)
- Previous studies
- Pilot studies

Study design and sample size

- Sample size estimation depends on the study design – as variance of an estimate depends on the study design
- The variance formula we just used is based on “simple random sampling” (SRS)
- In practice, SRS strategy is rarely used
- Be aware of the study design

Sample Size Under SRS Without Replacement

We know that under SRSWOR,

$$V(\bar{y}) = \frac{\sigma^2}{n} \frac{N-n}{N-1}$$

So, under SRSWOR:

$$d^2 = t_\alpha^2 \frac{p(1-p)}{n} \frac{N-n}{N-1}$$

$$n = \frac{NP(1-P)}{(N-1)D^2 + P(1-P)}$$

where,

$$D = d / t_\alpha$$

For continuous data,

$$n = \frac{N\sigma^2}{(N-1)D^2 + \sigma^2}$$

Alternative Specification (in two-stages):

This n, say n', is estimated under simple random sampling with replacement (SRSWR). When sampling is without replacement, we adjust the n by

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

or,

$$n = \frac{1}{\frac{1}{n'} + \frac{1}{N}}$$

(n is adjusted for finite population correction factor, 1- n/N).

Example: For exercise study, 93 samples are needed.

Say, the population size is 200.

Under SRSWOR:

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

$$n = \frac{93}{1 + \frac{93}{200}} = \frac{93}{1 + 0.465} = \frac{93}{1.465} = 63.48$$

$$n \approx 64$$

Smaller sample size is needed when population size is small, but opposite is not true

Derivation (alternative two-stage formula):

Remember the relationship between $\frac{S^2}{n'}....vs....\frac{N-n}{N}\frac{S^2}{n}$

$$=> \frac{1}{n'} = (1 - \frac{n}{N})\frac{1}{n}$$

$$=> \frac{1}{n'} = \frac{1}{N} - \frac{1}{n}$$

$$=> \frac{1}{n} = \frac{1}{n'} + \frac{1}{N}$$

$$=> \frac{1}{n} = \frac{1}{n'} + \frac{n'}{N}\frac{1}{n'}$$

$$=> \frac{1}{n} = \frac{1 + \frac{n'}{N}}{n'}$$

$$=> n = \frac{n'}{1 + n'/N}$$

Sample Size Based on Coefficient of Variation

- In the above, the sample size is derived from an *absolute* measure of variation, σ^2.
- Coefficient of variation (cv) is a *relative* measure, in which units of measurement is canceled by dividing with mean.
- Coefficient of variation is useful for comparison of variables.

Coefficient of variation is defined as,

$C_Y = \frac{S_y}{\bar{Y}}$, and is estimated by $c_y = \frac{s_y}{\bar{y}}$

Coefficient of variation (CV) of mean is

$$\boldsymbol{CV = \frac{SE}{\bar{y}} = \frac{s/\sqrt{n}}{\bar{y}} = \frac{s}{\bar{y}}\frac{1}{\sqrt{n}}}$$

So, $$n = \frac{1}{CV^2}\frac{s^2}{\bar{y}^2}$$

For proportion p,

$$n = \frac{1}{CV^2}\frac{p(1-p)}{p^2} = \frac{1}{CV^2}\frac{(1-p)}{p}$$

Caution about using coefficient of variation (CV)

- If mean of a variable is close to zero, *CV* estimate is large and unstable.

- Next, consider *CV* for binomial variables. For binary variables, the choice of P and Q=1-P does not affect P(1-P) estimate, but *CV* differs. So, the choice of P affects sample size when CV method is used.

Cost considerations for sample size

How many samples you may afford to interview, given then budget constraints?

$C(n)$ = cost of taking n samples

c_o = fixed cost

c_1 = cost for each sample interview

then,

$$C(n) = c_o + c_1 x\ n$$

Example:

C(n)=$10000 - your budget for survey implementation

c_o = $3000 - costs for interviewer training, questionnaire prints, etc

c_1 = $8.00 - cost for each sample interview

10000=3000+8*n

So, n=875

Objective 2: Issues of Power Calculation

POWER
The power of a test is the probability of rejecting the null hypothesis if it is incorrect.

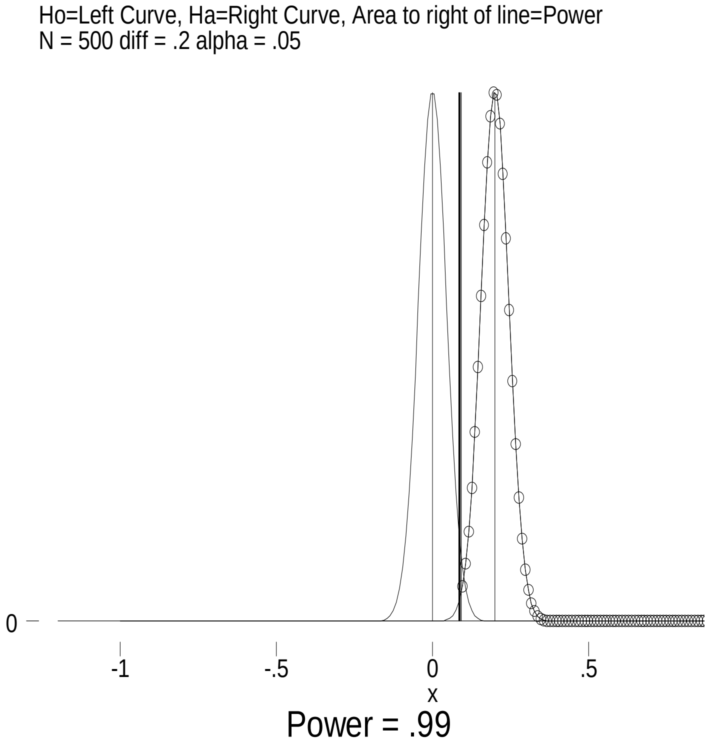
TRICKS to REMEMBER:
R, T: Reject the null hypothesis if it is true - Type I error (alpha error) { one stick in R, T}
A, F: Accept the null hypothesis if it is false - Type II error (beta error) {two sticks in A, F}
POWER: 1- type II error

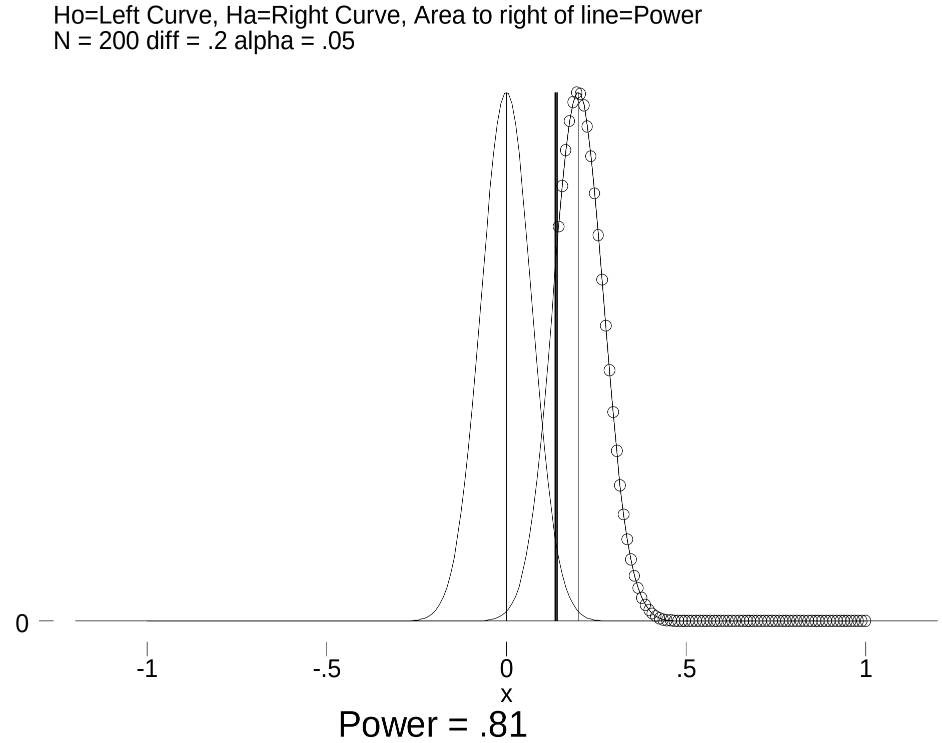
Power: Reject the null hypothesis if it is false.

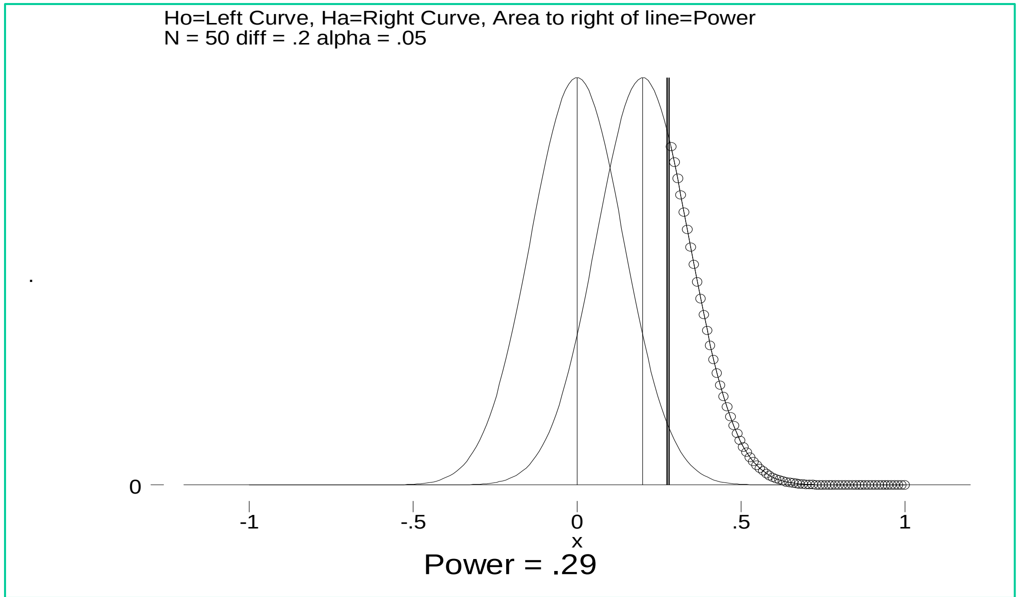
Another way:

False Positive (YES instead of NO) ?

False Negative (NO instead of YES) ?


Ho=Left Curve, Ha=Right Curve, Area to right of line=Power
N = 500 diff = .2 alpha = .05
0
-1
-.5
0
.5
x
Power = .99


Ho=Left Curve, Ha=Right Curve, Area to right of line=Power
N = 200 diff = .2 alpha = .05
0
-1
-.5
0
.5
1
x
Power = .81


Ho=Left Curve, Ha=Right Curve, Area to right of line=Power
N = 50 diff = .2 alpha = .05
0
-1
-.5
0
.5
1
x
Power = .29

We take power into consideration when we test hypotheses.

Example:

Consider following study questions:

1. What proportions of pregnant women received antenatal care?

There is no hypothesis.

b) Whether 80% of women received antenatal care?

There is a hypothesis:
To test that the estimated value is greater, less, or equal to a pre-specified value.

c) Do women in project (intervention) area more likely to utilize antenatal care, compared to control area?

There is a hypothesis:
To test that that P_1 is greater than P_2.

In terms of hypothesis:

Null hypothesis:	*Ho:P1=P2, i.e., P1-P2=0*
Alternative hypothesis:	*Ha:P1 > P2 (one-sided)*
	Ha:P1$\neq$P2 (two-sided) i.e., P1-P2 $\neq$ 0

Issues:

One-sided vs. two-sided tests.

One-sided: sample size will be smaller.
Two-side: sample size will be larger.

Always prefer "two-sided" - almost a mandatory in clinical trials.

Why? Uncertainty in knowledge (*a priori*).

How to incorporate "power" in sample size calculations?

1. Proportions:

$$n = \frac{(t_{\alpha} + t_{\beta})^2 \, \overline{p}(1 - \overline{p})}{d^2}$$

where $\overline{p}$ is $(p1 + p2)/2$

Note: **n for each group.**

Alternative:

$$n_1 = n_2 = \left[\frac{Z_{\alpha} + Z_{\beta}}{2\,arcsin\sqrt{p_1} - 2\,arcsin\sqrt{p_2}} \right]^2$$

Why?

Arcsin provides normal approximation to proportion quantities.

For continuous variables:

$$n = \frac{(Z_{1-\alpha/2} + Z_{\beta})^2 s^2}{d^2}$$

Values of $Z_{1-\alpha/2}$ and Z_{β} corresponding to specified values of significance level and power

	Values	Two-sided	One-sided
Level	1%	2.576	2.326
	5%	1.960	1.645
	10%	1.645	1.282
Power	80%	0.84	
	90%	1.282	
	95%	1.645	
	99%	2.326	

How to incorporate "power" in sample size calculations?

a) Proportions:

$$n = \frac{(z_{(1-\alpha/2)} + z_{1-\beta})^2 \, variance\ of\ difference[\mathrm{var}(p_1 - p_2)]}{(p_1 - p_2)^2}$$

How to estimate variance of difference?

$$\sigma^2_{(p_1-p_2)} = \sigma^2_d = \sigma^2_{p_1} + \sigma^2_{p_2} - 2\sigma_{p_1}\sigma_{p_2}$$

$$Under\ the\ assumption\ of\ independence, \mathrm{cov}(p1, p2) = \sigma_{p_1}\sigma_{p_2} = 0$$

$$If\ we\ also\ assume\ that\ \mathrm{var}(p1) = \mathrm{var}(p2) = \mathrm{var}(p),\ i.e.,\ have\ common\ variance$$

$$\sigma_d = \sigma^2_p + \sigma^2_p = 2\sigma^2_p$$

So,

$$n = \frac{(z_{(1-\alpha/2)} + z_{1-\beta})^2 \, variance \;\; of \;\; difference[v(p_1 - p_2)]}{(p_1 - p_2)^2}$$

$$= \frac{(z_{(1-\alpha/2)} + z_{1-\beta})^2 \, 2 * \overline{p}\overline{q}}{(p_1 - p_2)^2} \;\; where \;\; \overline{p} = (p_1 + p_2)/2$$

under the assumption of common variance

The sample size formula for testing two proportions under independence without the assumption of common variance is then:

$$n_1 = n_2 = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 [p_1(1-p_1) + p_2(1-p_2)]}{(p_1 - p_2)^2}$$

Note that Fleiss (1981) suggested more precise formula:

$$n' = \frac{\left\{ z_{1-\alpha/2}\sqrt{2\overline{p}(1-\overline{p})} + z_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)} \right\}^2}{(p_1 - p_2)^2}$$

$$where,\ p = (p_1 + p_2)/2$$

When n_1 and n_2 is not equal and related by a ratio, say by r, the formula is:

$$n' = \frac{\left\{ z_{1-\alpha/2}\sqrt{(r+1)\overline{p}(1-\overline{p})} + z_{1-\beta}\sqrt{rp_1(1-p_1) + p_2(1-p_2)} \right\}^2}{r(p_1 - p_2)^2}$$

The final formula (using normal approximation with continuity correction [without the correction, the power is considered low than expected] with proportions) is:

$$n_1 = \frac{n'}{4}\left\{ 1 + \sqrt{1 + \frac{2(r+1)}{n'r\,|\,p_1 - p_2\,|}} \right\}^2$$

$$n_2 = rn_1$$

The STATA has implemented this formula in SAMPSI command.

Stata implementation

NO Hypothesis

. sampsi .5 .55, p(.5) onesample

Estimated required sample size:

n = 385

Study has a hypothesis, but comparing with a hypothesized value

. sampsi .5 .55, p(.8) onesample

Estimated sample size for one-sample comparison of proportion to hypothesized value

Estimated required sample size:

n = 783

Study has a hypothesis, and comparing between two groups

. sampsi .5 .55, p(.8)

Estimated sample size for two-sample comparison of proportions

n1 = 1605

n2 = 1605

```
. di 783/385
2.0337662

. di
(1.96+.84)^2/1.96^2
2.0408163
```

Stata implementation

. sampsi .5 .55, p(.8) <u>nocontinuity</u>

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
and p2 is the proportion in population 2

Assumptions:

alpha = 0.0500 (two-sided)
power = 0.8000
p1 = 0.5000
p2 = 0.5500
n2/n1 = 1.00

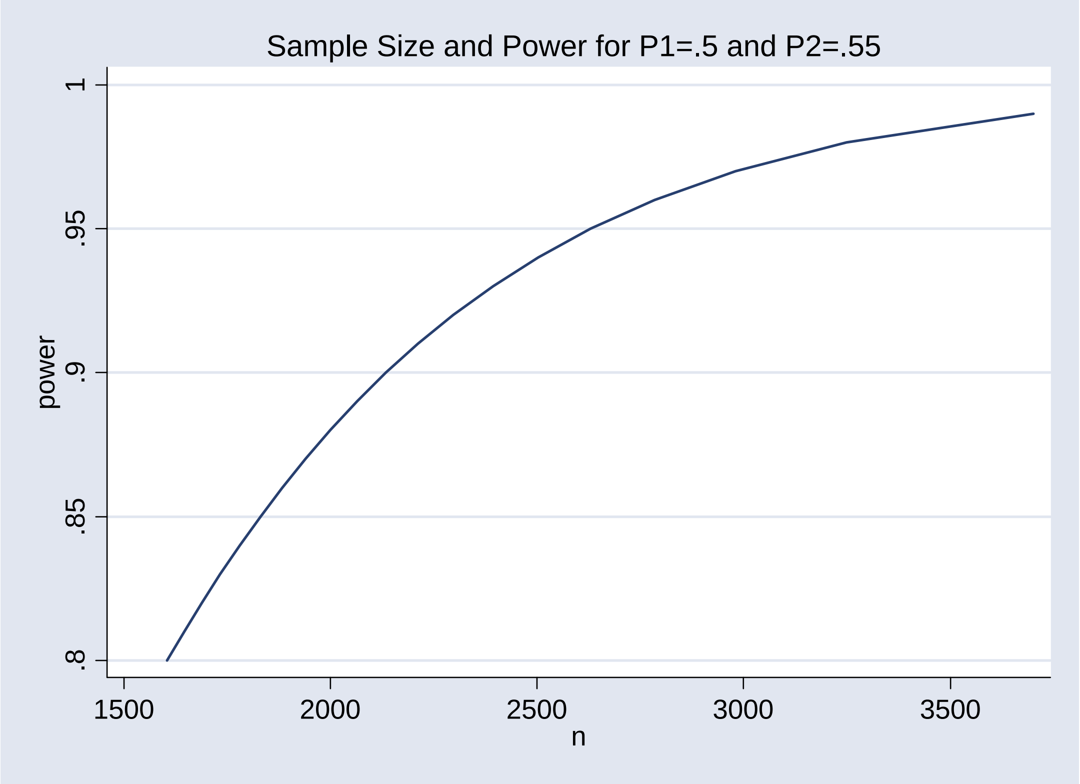
Estimated required sample sizes:

n1 = 1565
n2 = 1565

```
. di 783*2
1566

In each group,
sample size is
doubled
```

Power graph in Stata



```
*Calculate and plot sample size by power from .8 to .99

***************************************************************************
args p1 p2 type

clear
set obs 20
gen n=.
gen power=.
local i 0

while `i' <_N {
    local i = `i' +1

    local j=.79 +`i'/100

    quietly sampsi `p1' `p2', p(`j') `type'
    replace n=r(N_1) in `i'
    replace power=r(power) in `i'

    }

noisily list power n

graph twoway line power n, t1("Sample Size and Power for P1=`p1' and P2=`p2' `type'")
*****************************************************************************
```

Save the above commands as “do” file (e.g., sample_graph.do).

Execute the above file by:

run sample_graph

Sample size determination when expressed in "relative risk"

In epidemiological studies, often the hypothesis is expressed in
relative risk or odds ratio, e.g, H0:R=1.

A sample size formula given in Donner (1983) for Relative Risk (p. 202) is:

$$n = \left\{ Z_{\alpha}\sqrt{2\overline{P}_R(1-\overline{P}_R)} + Z_{\beta}\sqrt{P_c\left\{1+R-P_c(1+R^2)\right\}} \right\}^2 / [P_c(1-R)]^2$$

$$Where \quad \overline{P}_R = [P_c(1+R)]/2 \quad and \quad R = P_E / P_c$$

Nothing but the Fleiss' formula:

$$n' = \frac{\left\{ z_{1-\varepsilon}\sqrt{2\overline{p}(1-\overline{p})} + z_{1-\beta}\sqrt{p_E(1-p_E) + p_C(1-p_C)} \right\}^2}{(p_C - p_E)^2}$$

$$where,\ p = (p_E + p_C)/2$$

Note, $R = \frac{P_E}{P_C} = P_E = RP_C$

Solution: Replace all P_E with RP_C and apply Fleiss' formula

How Donner's formula was derived:

$P=(P_E+P_C)/2=(RP_C+P_C)/2=[P_C(R+1)]/2=[P_C(1+R)]/2$

$P_E(1-P_E)+P_C(1-P_C)=RP_C(1-RP_C)+P_C(1-P_C)$

$=RP_C-R^2P_C^2+P_C-P_C^2$

$=P_C(R-R^2P_C+1-P_C)$

$=P_C\ (1+R-P_C\ (1+R^2)$

and, $(P_C-P_E)^2=(P_C-RP_C)^2=[P_C(1-R)]^2$

Sample size for odds-ratio (OR) estimates:

$$OR = \frac{\dfrac{P_2}{1-P_2}}{\dfrac{P_1}{1-P_1}} = \frac{P_2 Q_1}{P_1 Q_2}$$

$$P_2 Q_1 = OR * P_1 Q_2 = OR * P_1 (1 - P_2) = OR * P_1 - OR * P_1 * P_2$$

$$P_2 (Q_1 + OR * P_1) = OR * P_1$$

$$P_2 = \frac{OR * P_1}{OR * P_1 + Q_1}$$

Convenient to do in two stages:

1. Estimate P_2 from odds-ratio (OR)
2. Apply "proportion method" (of Fleiss)

An example

Suppose we want to detect an OR of 2 using an ratio of 1:1 cases to controls in a population with expected exposure proportion in non-cases of 0.25 while requiring a α= 0.05 and power = 0.8.

How to estimate SS?

EpiTable calculates m1 = m2 = 165. (Total sample size = 330).

So, P1=.25,

P2= (2*.25)/(2*.25+.75) = 0.4

In Stata:

. sampsi .25 .40, p(.8)

Estimated required sample sizes:

n1 = 165

n2 = 165

SAMPLE SIZE determination for Logistic Regression Models

Consider a logistic regression,

$$\log\left(\frac{p}{1-p}\right) = logit(p) = \alpha + \beta x$$

We want to estimate sample size needed to achieve certain power for testing null hypothesis

Ho:β=0. Recall that null hypothesis testing depends on the **variance of β**. In logistic regression, the **effect size** is expressed as "log odds ratio" (η).

Hsieh(1989) suggested the following formula for one-sided test:

$$n = \left[z_\alpha + z_\beta \exp(-\eta^2/4)\right]^2 \left(1 + 2\hat{p}\delta\right)/(\hat{p}\eta^2)$$

where,

$$\delta = [1 + (1+\eta^2)\exp(5\eta^2/4)]/[1+\exp(-\eta^2/4)]$$

Say, you want to examine the **effect size** of **log odds ratio** of 1.5 = log(1.5)= .40546511 = ~0.41

See, the implementation of formula in STATA:

```
. clear
. set obs 1
obs was 0, now 1

. *Enter "odds-ratio"
. gen p=0.10
. gen or=1.5

. gen beta=log(or)

. di beta
.4054651
. gen delta= (1+(1+beta^2)*exp(5*beta^2/4))/(1+exp(-beta^2/4))
.
. di delta
1.2399909
.
. di " n = "  (1.645+1.282*exp(-
beta^2/4))^2*(1+2*p*delta)/(p*beta^2)
 n = 627.61987
```

So, **n = 628 = ~ 630**

Sample Size for Multiple logistic Regression

Multiple logistic regression requires larger n to detect effects. Let R denote the multiple correlation between the independent variable of interest, *X,* and the other covariates.

Then, sample size:

$n` = n/(1-R^2)$

Say, if R=0.25, then n` = 630/(1-0.25^2)= 672

Stata’s add-on programs for sample size estimation

- STPOWER: Survival studies
- Sampsi_reg: Linear regression
- Sampclus: Cluster sampling
- ART: randomized trials with survival time or binary outcome
- XSAMPSI: Cross-over trials
- Samplesize: Graphical results
- MVSAMPSI: multivariate regression

- STUDYSI: Comparative study with binary or time-to-event outcome
- SSKAPP: Kappa statistics measure of inter-rater aggrement
- CACLSI: log-rank/binomial test

Additional topics to be covered

- Sample allocation – stratified sampling
- Sample size corrected for design-effect(DEFF)
- Optimal sample size per cluster
- Sample size for clusters
- Sample size and power for pre-post surveys in program evaluation