

# Sample Size Estimation and Power Analysis

March 2008

**Ayumi Shintani, PhD, MPH**  
**Department of Biostatistics**  
**Vanderbilt University**

1



A researcher conducted a study comparing the effect of an intervention vs placebo on reducing body weight, and found 5 lbs reduction among the intervention group with  $P=0.01$ .



Another researcher conducted a similar study comparing the effect of the same intervention vs the same placebo on reducing body weight, and found the same 5 lbs reduction with the intervention group but could not claim that the intervention was effective because  $P=0.35$ .

What do you think the crying researcher did differently from the smiling one?

2

What impacts on p-value when comparing new drug v.s. placebo?

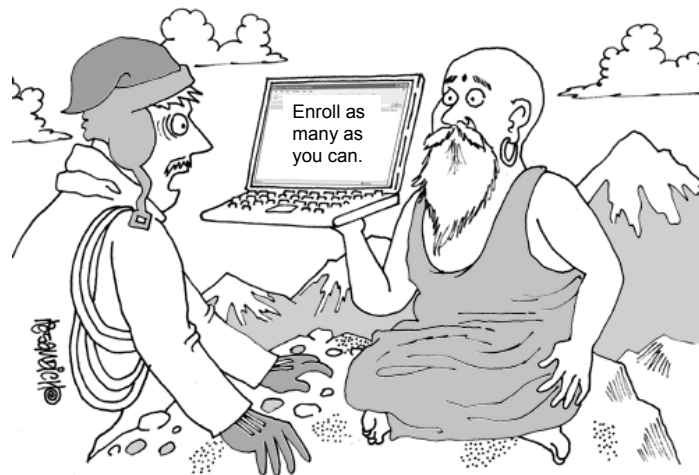
The effect of the new drug. ex: Larger reduction (10lbs) in weight by the new drug!

Variation of data: Larger variation can result in larger p-value.  
Source of variation:  
Between-subject variation  
Measurement error

And what else???????

3

Question: How can I make my P-value smaller.



"I climb all this way, and you tell me *THAT'S* the meaning of P-value!?"

4

What impacts on p-value when comparing new drug v.s. control?

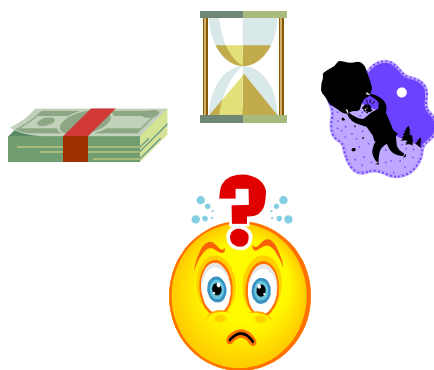
The effect of the new drug. ex: Larger reduction (10lbs) in weight by the new drug!

Variation of data: Larger SD can result in larger p-value,  
Source of variation:  
Between-subject variation  
Measurement error

Sample size. Larger sample size can make p-value smaller!  
↳ Even a tiny, clinically meaningless effect can become significant some day if you keep enrolling patients indefinitely.

5

As many as I can??????  
How many can I ??????



So you need to justify the enrollment of a minimum number of subjects enough to prove that the drug is effective.



Need Sample size Estimation!!!

Do I have a enough resource?  
Does NIH agree to pay me that much????  
Is it ethical to expose unnecessary large number of patient to a unproven drug?

### Example of reporting sample size estimation (1)

CONSORT statements provide a check list for required items for RCT, and used by many journals such as NEJM, LANCET, JAMA, Anals Int Med

- Title and Abstract
- Introduction
  - Background
- Methods
  - Participants
  - Interventions
  - Objectives
  - Outcomes
  - Randomization
  - Blinding
  - **Sample Size and Power**
  - Statistical methods
- Results
  - Recruitment
  - Baseline data
  - Numbers analyzed
  - Outcomes and estimation
  - Ancillary analyses
  - Adverse events
- Comments
  - Interpretation
  - Generalizability
  - Overall evidence

7

### Example of reporting sample size estimation (2)

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

## Premature Coronary-Artery Atherosclerosis in Systemic Lupus Erythematosus

Yu Asanuma, M.D., Ph.D., Annette Oeser, B.S., Ayumi K. Shintani, Ph.D., M.P.H.,  
Elizabeth Turner, M.D., Nancy Olsen, M.D., Sergio Fazio, M.D., Ph.D.,  
MacRae F. Linton, M.D., Paolo Raggi, M.D., and C. Michael Stein, M.D.

#### STATISTICAL ANALYSIS

Assuming the frequency of coronary-artery calcification is 15 percent among asymptomatic 40-year-old women,<sup>22</sup> the study required 65 patients and 65 controls to have 85 percent power to detect a minimal frequency of coronary-artery calcification of 35 percent among patients with lupus. Statistical analy-

8

Scientific approach of proving a hypothesis = Disproving a null hypothesis.

**Null Hypothesis: There is no difference between the new drug and control drug.**

P-value: The probability of observing a difference as large or larger just by chance alone when the null hypothesis is true.

Reject → The new drug is more effective than the control.

Fail to reject → No evidence to support that the new drug is more effective than the control.

When you make this inferential judgment, two type of errors can occur.

9

**Two critical errors involved in hypothesis testing:**

**Type 1 error ( $\alpha$ ):** falsely concluding that the drug is effective when the drug actually is not effective.

*Traditionally, you are allowed to make this error up to 5%  
{i.e., significance level ( $\alpha$ ) = 5% }*

**Type 2 error ( $\beta$ ):** falsely concluding that the drug has no effect when the drug is actually effective.

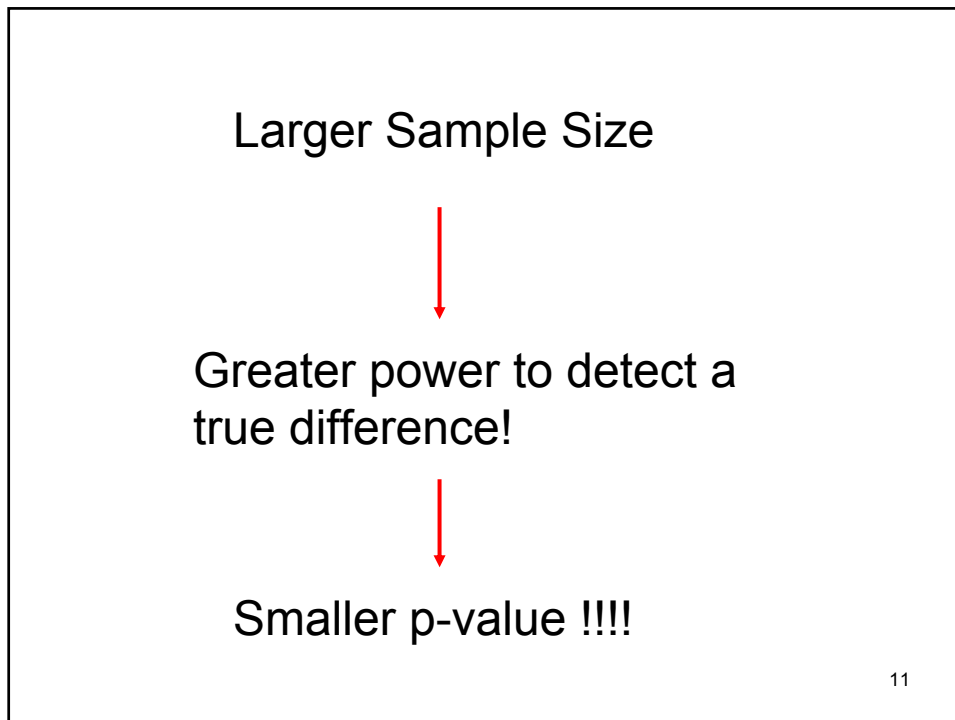
*Traditionally, you are allowed to make this error up to 20%*

**Power of statistical tests:**

**Power of the test ( $1-\beta$ ):** the probability of correctly concluding that the drug is effective, when the drug actually is effective.

A statistical test with a larger sample size can decrease both type I and II errors. We try to get a sample large enough to ensure that  $1-\beta$  is at a reasonable level (80% or more).

10




**When to conduct Sample Size and Power Analysis?**

You almost always need to estimate a required sample size or estimate analytical power given a sample size when you are planning a study. Only exception may be a pilot study (a smaller study to show feasibility, or to collect data to plan a larger study).

Through this process, you can avoid wasting your efforts and resources conducting studies that are hopeless to begin with.

Question: Can I keep enrolling patients into my study until I observe  $P < 0.05$ ?



Absolutely NOT

12

**Possible consequences of a study with small sample size  
(insufficient power):**

Observed difference between two treatment groups is clinically important, however it may not reject the null hypothesis indicating that the drug is not effective ( $p\text{-value} > 0.05$ ).



**Possible consequences of study with sample size being too large  
(excess power):**

Difference between two treatment groups is not clinically important, however it may reject the null hypothesis indicating that the drug is effective ( $p\text{-value} < 0.05$ ).

e.g., a new drug reduces body weight by a half pound in 1 year ( $P=0.001$ ).  
Is this worth publishing?

13

### Post-Hoc Power Analysis

Power Analysis after data-analysis results (calculating p-value) is called Post Hoc power analysis, and it is not necessary.

**Significant difference was detected with analysis lacking power —>  
OK, Go ahead to report your finding.**

**Significant difference was not detected with analysis lacking power  
—>**

**If the observed difference is clinically meaningful, too early to discard your study, increase samples, and re-analyzed you data.**

14

### Free Software for Sample Size and Power – PS Software

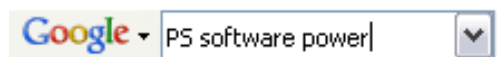
PS (Power and Sample Size) software

PS is an interactive program for performing power and sample size calculations and freely available on the Internet. This program was developed by my colleagues, Professor William Dupont and Dale Plummer. You can download it from our website of the department of Biostatistics, Vanderbilt University at:

<http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize>

15

How to download PS software (1): How to find the download site for PS



[PowerSampleSize < Main < Biostatistics TWiki](#)

PS is an interactive program for performing **power** and sample size calculations. ... To obtain this **software** on your computer click **PS** (5.2 MB). ...

[biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize](http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize) - 24k -

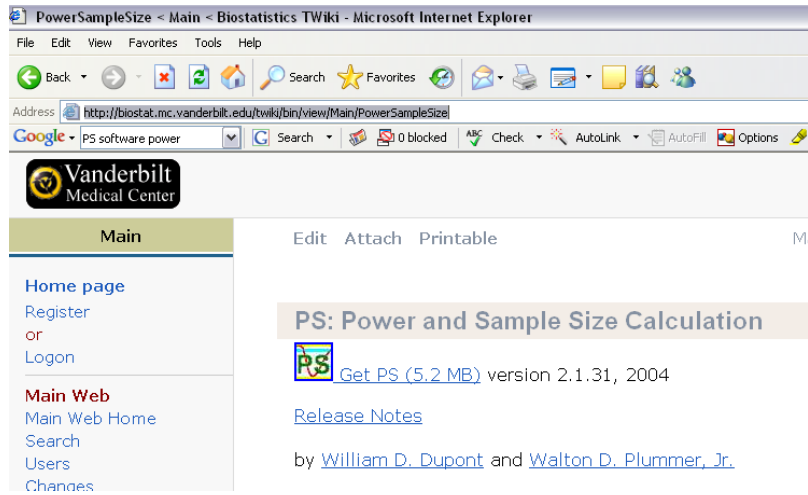
[Cached](#) - [Similar pages](#)



<http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize>

16

How to download PS software (2): Download site



17

**Factors Affecting the Sample Size:**

- |  |   |                      |   |
|--|---|----------------------|---|
| (1) Effect of treatment (effect size, $\delta$ ) | ↑ | Required sample size | ↓ |
| (2) Variation of data (SD, $\sigma$ )            | ↑ | Required sample size | ↑ |
| (3) Type I error ( $\alpha$ )                    | ↓ | Required sample size | ↑ |
| (e.g., Reject if $P < 0.025$ , rather than 0.05) |   |                      |   |
| (e.g., Use 2-sided rather than 1-sided)          |   |                      |   |
| (4) Power = 1-Type II error ( $\beta$ )          | ↑ | Required sample size | ↑ |
| (e.g. set to 90%, rather than 80%)               |   |                      |   |

Effect size, and SD are usually obtained through pilot studies, or published data.<sup>18</sup>

**Example: Estimation of sample size comparing 2 group means (independent sample t-test): Comparing post trial values (1)**

A clinical trialist wants to conduct RCT to assess the effect of an intervention to reduce HbA1c level among patients with type 2 diabetes. A pilot data suggests that mean HbA1c level among patients without this intervention is 8.7% with standard deviation of 2.2%. We believe that the intervention will decrease patient's HbA1c level by 1%. A total of 154 patients (77 patients in each group) are needed to achieve 80% power at two-sided 5% significance level.

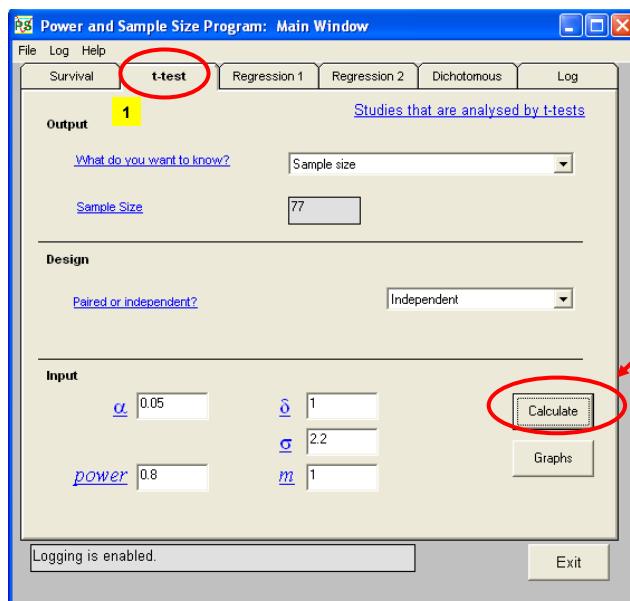
Parameters needed for sample size computation:

Power = 80%,  $\alpha$  = 5% (2 sided),  $\delta$  (delta)=1

$\sigma$  (sigma)=2.2, m (sample size ratio between the two groups) = 1

19

**Sample size estimation using PS software for Student's t-tests (1)**

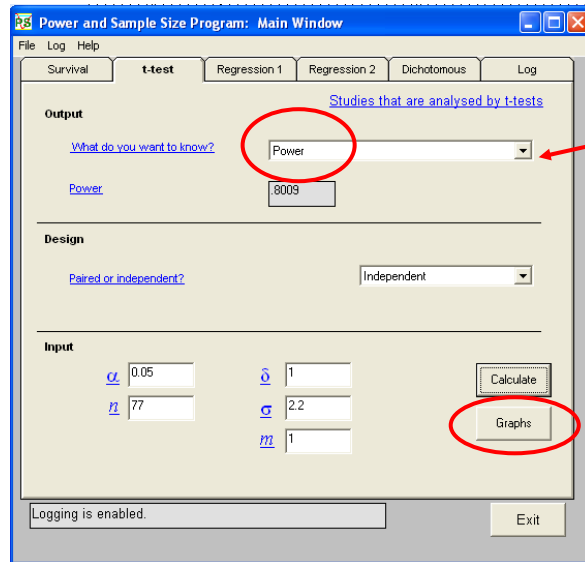


**2** Enter parameters

**3** After entering all parameters, click here

20

**Sample size estimation using PS software for Student's t-tests (2):  
Drawing a graph of statistical power by varying sample sizes (1)**

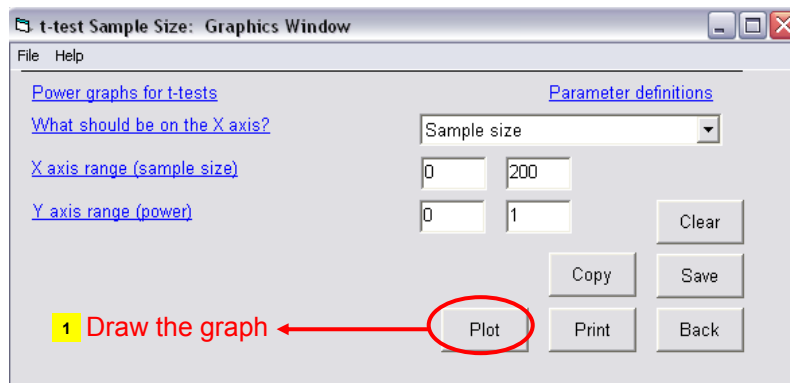


1 Change sample size to power to obtain a graph

2 Click here for the graph

21

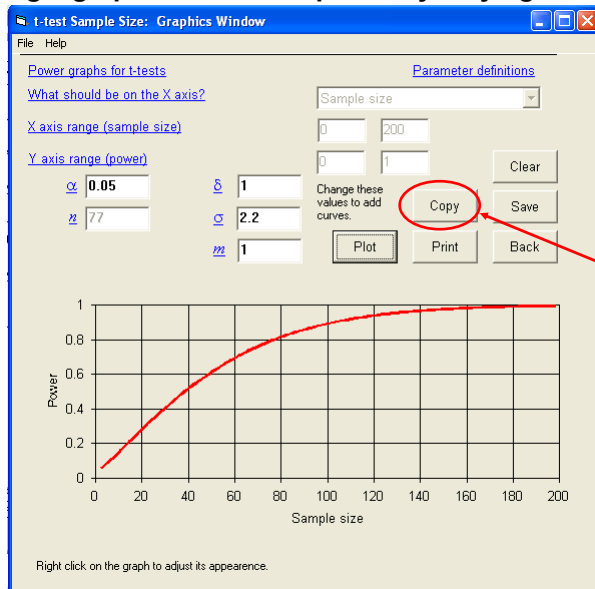
**Sample size estimation using PS software for Student's t-tests (3):  
Drawing a graph of statistical power by varying sample sizes (2)**



1 Draw the graph

22

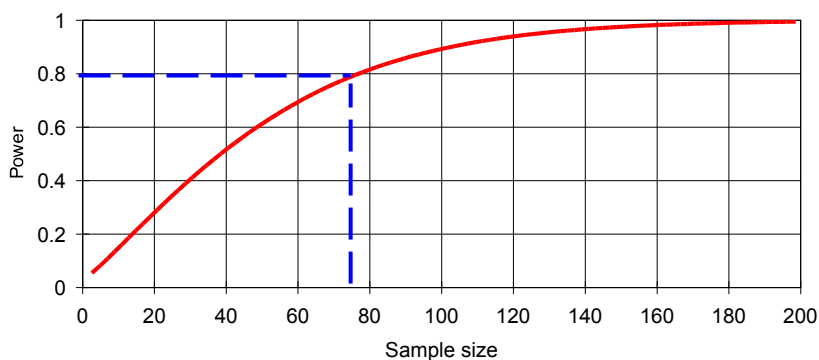
**Sample size estimation using PS software for Student's t-tests (4):  
Drawing a graph of statistical power by varying sample sizes (3)**



Copy and  
paste into your  
document

23

**Sample size estimation using PS software for Student's t-tests (5):  
Drawing a graph of statistical power by varying sample sizes (4)**

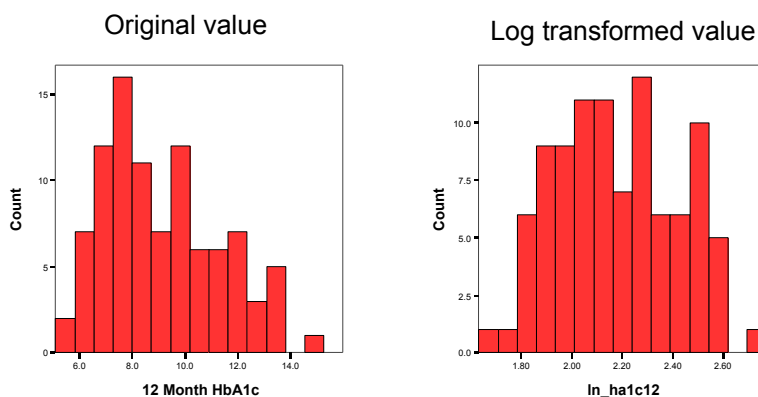


It requires about 77 patients in each group (total of 154) to achieve 80% power.

24

Improving analytical power (reducing required sample size):  
 Transformation

For skewed distribution, transformation may improve power.



25

Parameters needed for sample size computation:  
 Power = 80%,  $\alpha$  = 5% (2 sided),  $\delta$  (delta) =  $\ln(8.7) - \ln(7.7) = 0.122$   
 $\sigma$  (sigma) = 0.21,  $m$  (sample size ratio between the two groups) = 1  
 ————  $\leftarrow$  SD of  $\ln(\text{HbA1c})$

26

**Improving analytical power (reducing required sample size)  
Using change from baseline**

A clinical trialist wants to assess the effect of an intervention to reduce HbA1c level among patients with type 2 diabetes. A pilot data suggests that mean **reduction** in HbA1c level **from baseline** among patients without this intervention is 0.2% with standard deviation of 1.5%. We believe that the intervention will further decrease patient's HbA1c level by 1%. A total of **72** patients (**36** patients in each group) are needed to achieve 80% power at two-sided 5% significance level.

27

Sample size computation for the question on the previous page

Power and Sample Size Program: Main Window

File Log Help

Survival **t-test** Regression 1 Regression 2 Dichotomous Log

Studies that are analysed by t-tests

Output

What do you want to know? Sample size

Sample Size 36

Design

Paired or independent? Independent

Input

$\alpha$  0.05  $\delta$  1

power 0.8  $\sigma$  1.5  $\mu$  1

Calculate

Graphs

Required sample size reduced from 77 to 36 per arm.

SD of within patient change value (1.5%) is smaller than SD of post value (2.2%) resulting in greater power.

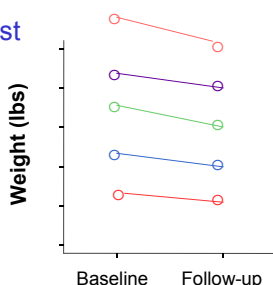
28

**Impact on power between comparing post value vs change value**

Usually in most studies, between patient variability is greater than within patient variability. Using within-patient change value can provide more precision/power to the analysis by removing between-patient variability (standard deviation for change value is much smaller than standard deviation for group means).

Between patients post intervention weight

(100, 113, 145, 186, 200lbs)



Within patient reduction in weight

(10, 6, 10, 8, 4lbs)

High ICC → Values within a patient (cluster) are similar to each other than to values of another patient (cluster)

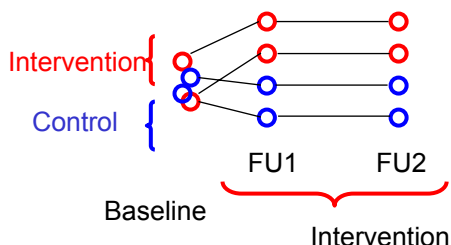
Intraclass correlation coefficient (ICC)

= Between patient variation / (Between and within patient variations) 29

**ICC (correlation among repeated measures) and power**

When estimating a **change** (post – baseline) or average rate of change (**slope**) over repeated measures, power **increases** with larger ICC, *smaller sample size required*.

When estimating an **average** over repeated measures, power **decreases** with larger ICC, *larger sample size required*.



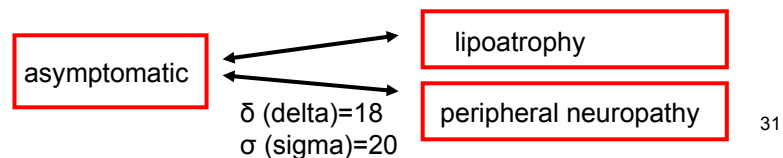
Power increases with larger ICC when comparing FU1 – Baseline, or FU2 – Baseline between 2 groups.

But Power decreases with larger ICC when comparing average(FU1, FU2) between 2 groups.

### Comparing means of 3 groups (Bonferroni correction)

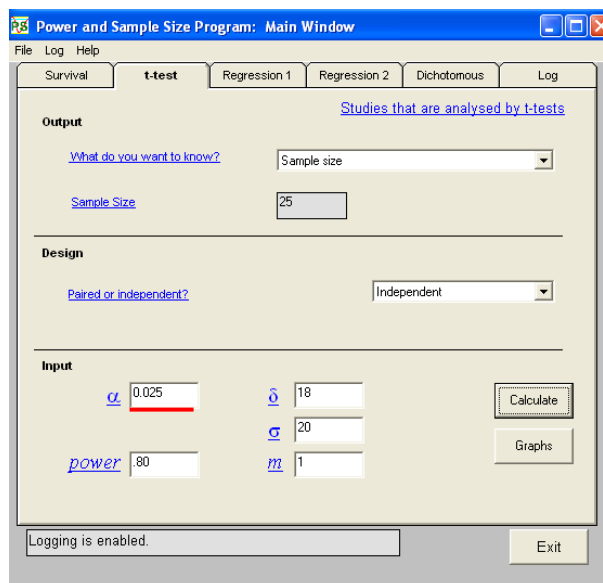
Primary Outcome: F2 isoprostane concentrations  
Exposure Variables: Three groups of symptoms of HIV patients

We want to conduct a study comparing mean F2 isoprostane concentrations among 3 groups (1) asymptomatic patients, (2) patients with lipoatrophy, and (3) patients with peripheral neuropathy. A previous study found that a mean F2-IsoP level of 60 pg/mL in subjects with lipoatrophy and 42 pg/mL in subjects without lipoatrophy. We assume a similar F2-IsoP level for patients with peripheral neuropathy. A common standard deviation for F2 Isoprostane level is 20 pg/mL. A total of 75 patients (25 patients in each group) are needed to 80% power with two-sided 2.5% significance level. Bonferroni adjustment was used as alpha level/number of comparisons =  $0.05/2 = 0.025$



31

Sample size computation for the question on the previous page



32

### Effect of Bonferroni adjustment on required sample size

If this trialist conducts 2-arm study,  $\alpha=0.05$  would be used.

↳ 20 patients per arm (total of 40)

If this trialist conducts 3-arm study and 2 comparisons are planned,  $\alpha=0.025$  would be used.

↳ 25 patients per arm (total of 75)  
↳ 27 patients per arm (total of 81) if 3 comparisons planned ( $\alpha=0.167$ )

Thus planning 3-arm study requires sample size more than 1.5 times of 2-arm study.

33

### Comparing 2 proportions (Pearson chi-square test)

Percent of patients who did not improve HbA1c level (defined as not achieving  $< 7\%$ ) was 80% among patients without an intervention. We anticipate 25% reduction with this intervention ( $80\% \times 0.75=60\%$ ). A total of 162 patients (81 patients in each group) are needed to achieve 80% power with two-sided 5% significance level.

34

Sample size computation for the question on the previous page

Power and Sample Size Program: Main Window

File Log Help

Survival t-test Regression 1 Regression 2 **Dichotomous** Log

Output

Studies that are analysed by chi-square or Fisher's exact test

What do you want to know? Sample size

Case sample size for uncorrected chi-squared test 81

Design

Matched or Independent? Independent

Case control? Prospective

How is the alternative hypothesis expressed? Two proportions

Uncorrected chi-square or Fisher's exact test? Uncorrected chi-square test

Input

$\alpha$  0.05  $p_0$  0.8

power 0.8  $p_1$  0.6

$m$  1

Calculate

Graphs

35

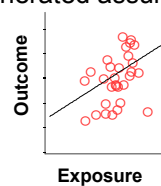
### Impact of data categorization on analytical power

General Rule: Greater granularity in Data leads to greater power.  
 Loss of Data often leads to loss of power.

Two continuous variables (N=30) were generated assuming correlation = 0.5

Pearson's correlation

→ Power > 80%

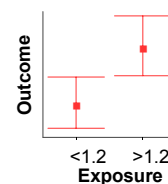


Outcome (Median=0.17)

Exposure (Median=1.2)

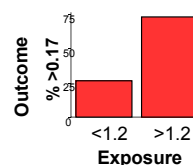
Independent Sample t-test

→ Power = 59%



Pearson chi-square test

→ Power = 12%



36

Sample size analysis for parameter estimation (Precision)

Specific Aim: To describe the incidence of cognitive dysfunction in survivors of medical ICU. We anticipate that proportion of patients with cognitive dysfunction among ICU survivors is 30% based on a relevant literature.

Since the analytical aim is to estimate an incidence of the outcome variable, the power computation will focus the precision of the estimate by showing estimated confidence interval (CI). With 240 patients, we will be able to construct 95% CI of an observed incidence (30%) of (24%, 36%).

$$95\% \text{ CI of } p = p \pm 1.96 \text{ SD} / \sqrt{n}$$

$$\text{SD for proportion} = \sqrt{p \times (1-p)}$$

37

Guideline for the maximum number of independent variables to be included in a multivariable model

Linear regression	# patients (samples) / 15 (10-20)
Logistic regression	Min(# events, # non-events) / 15 (10-20)
Cox regression	# events / 15 (10-20)
Proportional odds logistic regression	$n - \frac{1}{n^2} \sum_{i=1}^k n_i^3$ / 15 (10-20)

#, number of

K: number of categories, n: total sample size, n<sub>i</sub>: sample size in each category

References:

- \* Harrell FE, Jr. Regression Modeling Strategies. Springer Verlag. (2001).
- \* Peduzzi P et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996 Dec;49(12):1373-9.
- \* Peduzzi P et al. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995 Dec;48(12):1503-10.

38

**Example: Multivariable regression**

The purpose of this aim is to understand the association between delirium and severity of long-term cognitive impairment after controlling for a set of covariates. The statistical analysis will use multiple linear regression. The allowable number of sample size will be based on the general rule that a number of independent variables times 15 to allow for proper multivariable analysis. The main independent variable of the linear model will be “total delirium days” (continuous) plus the following 9 covariates: age (continuous), baseline comorbidity index (continuous), baseline cognitive impairment (continuous), severity of illness (continuous), sepsis (dichotomous), hypoxemia (continuous), total days of mechanical ventilation (continuous), total coma days (continuous) and apoE genotype (categorical with 3 levels). Therefore, the minimum number of patients required for the model will be  $10 \times 15=150$ .

39