

Sample Size Estimation and Power Analysis

January 2010

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?

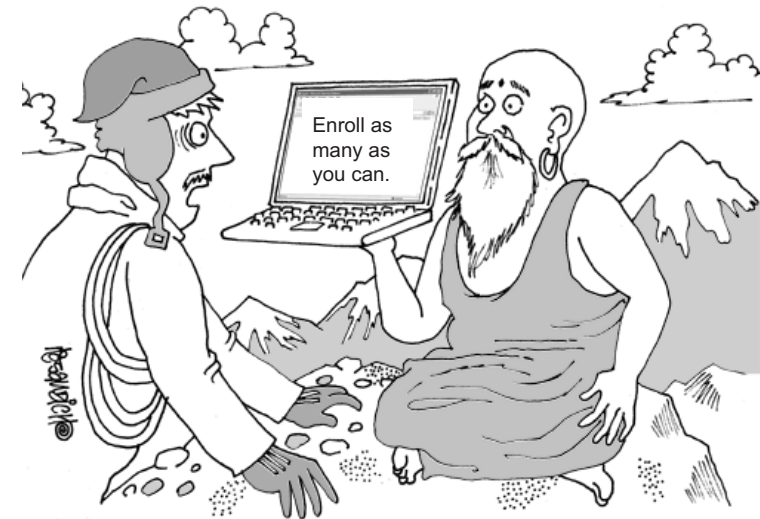
The effect of the new drug. ex: Larger reduction (10lbs) in weight by the new drug!

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.



4

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

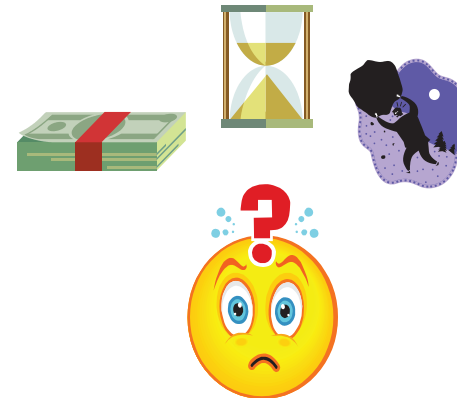
The effect of the new drug. ex: Larger reduction (10lbs) in weight by the new drug!

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.

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?

42% of R01 applications are criticized for sample size or power problems.

Inouye SK, Fiellin DA. An evidence-based guide to grant proposals for clinical research. Ann Intern Med. 2005;142:274-282

7

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

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 types of error can occur.

9

Two critical errors involved in hypothesis testing:

Type 1 error (α): 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 (α) = 5% }

Type 2 error (β): 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

Larger Sample Size



Greater power to detect a true difference!



Smaller p-value !!!!

11

Factors Affecting the Sample Size:

- (1) Effect of treatment (effect size, δ) ↑ Require sample size ↓
- (2). Variation of data (SD, σ) ↑ Require sample size ↑
- (3) Type I error (α) ↓ Require sample size ↑
(e.g., Reject if $P < 0.025$, rather than 0.05)
(e.g., Use 2-sided rather than 1-sided)
- (4) Power = $1 - \text{Type II error } (\beta)$ ↑ Require sample size ↑
(e.g. set to 90%, rather than 80%)

Effect size, and SD are usually obtained through pilot studies, or published data.

12

Free Software for Sample Size and Power – PS Software

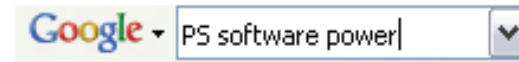
PS (Power and Sample Size) software

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>

13

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 - 24k -
[Cached](#) - [Similar pages](#)



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

14

How to download PS software (2): Download site

PowerSampleSize < Main < Biostatistics TWiki - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print Mail News RSS Feeds

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

Google PS software power Search 0 blocked Check AutoLink AutoFill Options


Vanderbilt Medical Center

Main Edit Attach Printable Mi

Home page
Register
or
Logon

Main Web
Main Web Home
Search
Users
Changes

PS: Power and Sample Size Calculation

 [Get PS \(5.2 MB\)](#) version 2.1.31, 2004

[Release Notes](#)

by [William D. Dupont](#) and [Walton D. Plummer, Jr.](#)

15

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

Study Design:

2-arm RCT to compare HbA1c level among patients with type 2 diabetes.

Sample size computation:

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 ? patients ? patients in each group) are needed to achieve 80% power at two-sided 5% significance level.

Parameters needed for sample size computation:

Power =

Significance level (α) =

Anticipated difference (δ : delta) =

Standard deviation (σ : sigma)=

m (sample size ratio between the two groups) =

16

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

2 Enter parameters

3 After entering all parameters, click here

17

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

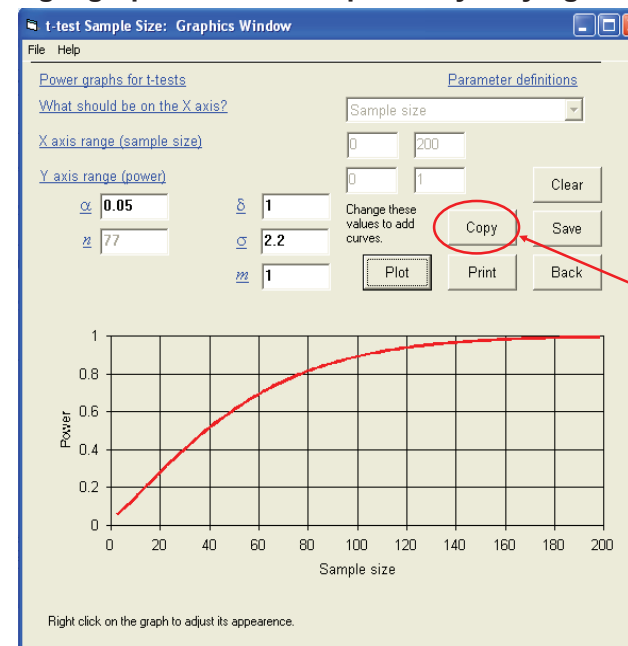
18

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

19

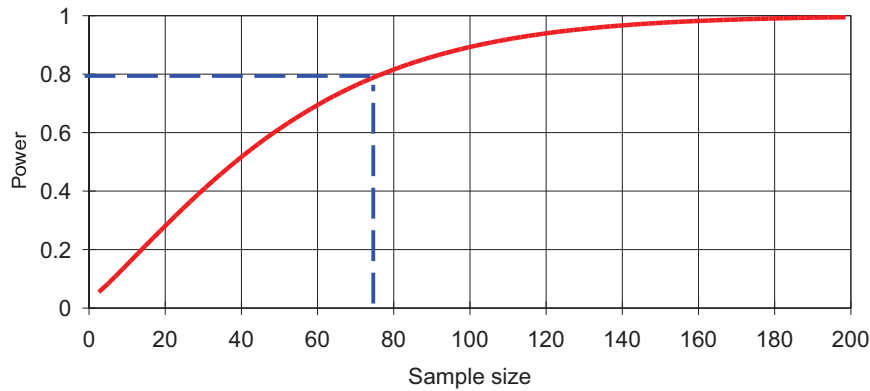
Sample size estimation using PS software for Student's t-tests (4): Drawing a graph of statistical power by varying sample sizes (3)



1 Copy and paste into your document

20

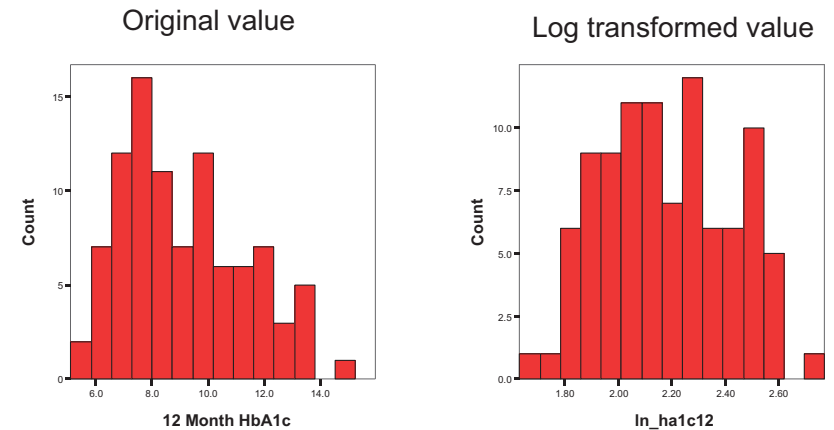
**Sample size estimation using PS software for Student's t-tests (5):
Drawing a graph of statistical power by varying sample sizes (4)**



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

**Improving analytical power (reducing required sample size):
Transformation**

For skewed distribution, transformation may improve power.

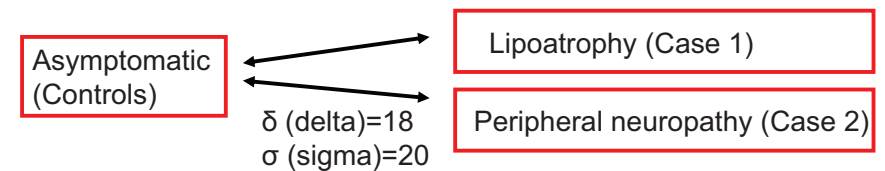


Parameters needed for sample size computation:

Power = 80%, α = 5% (2 sided), δ (delta)= $\ln(8.7)-\ln(7.7)=0.122$
 σ (sigma)=0.21, m (sample size ratio between the two groups) = 1

↑ SD of $\ln(\text{HbA1c})$

Comparing means of 3 groups (Bonferroni correction)



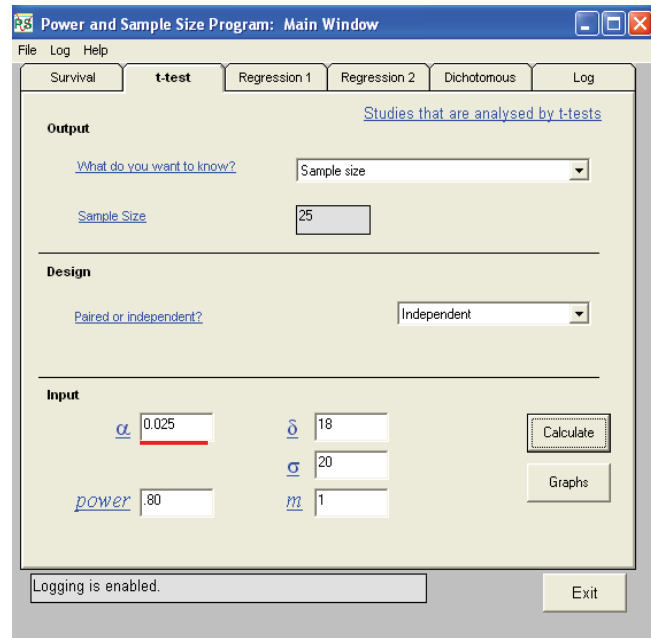
Study Design: Case – control study with 3 groups.

Sample size computation:

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 ? patients (? 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 = $0.05/2=0.025$

Sample size computation for the question on the previous page



25

Effect of Bonferroni adjustment on required sample size

# groups to compare	Number of pair-wise comparisons	Type I error rate	Required Sample Size per group	Total sample size
2 groups	1	0.05	20	40
3 groups	2	0.025	25	75
3 groups	3	0.0167	27	81

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

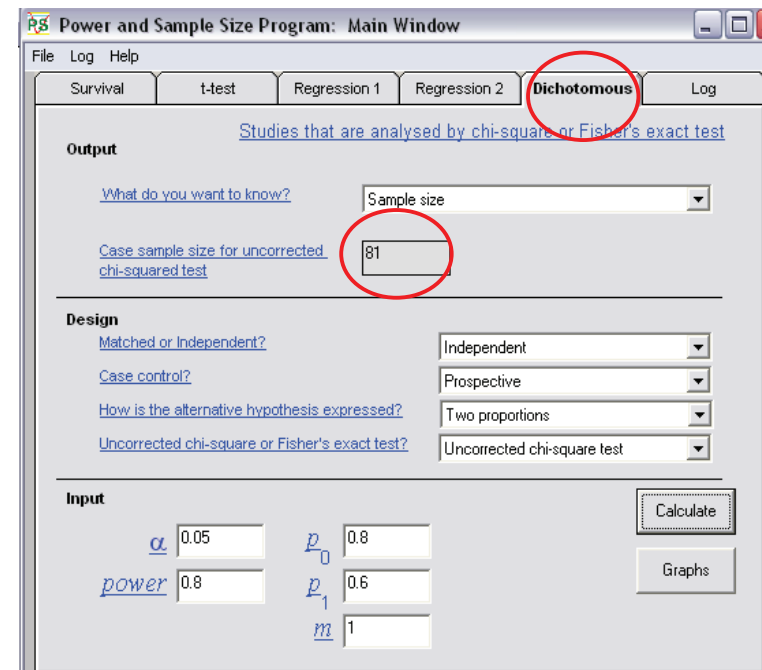
26

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% x 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.

27

Sample size computation for the question on the previous page



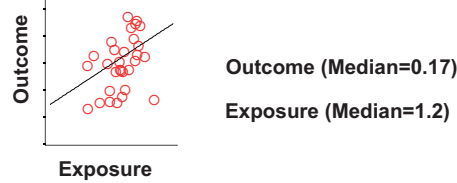
28

Impact of data categorization on analytical power

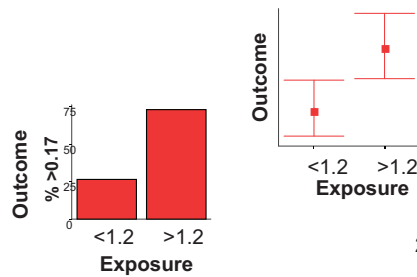
General Rule: Greater granularity in Data leads to greater power.
Loss of Data often leads to loss of power.

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

Pearson's correlation
→ Power > 80%



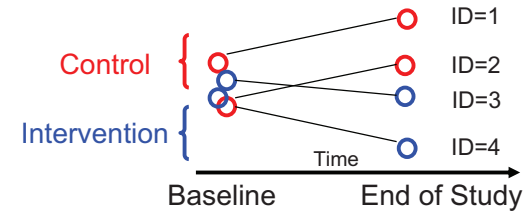
Independent Sample t-test
→ Power = 59%



Pearson chi-square test
→ Power = 12%

29

Power to compare effect of drug in 2-arm RCT with value of outcome is measured at baseline.

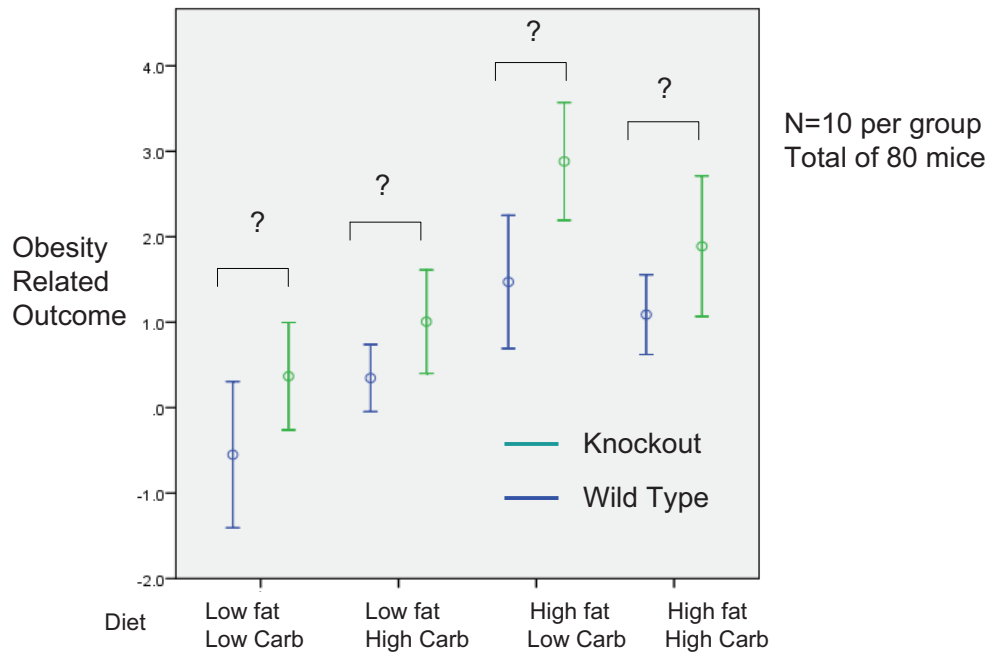


	Comparing End-of-Study values Only	Comparing Change from Baseline	Adjusting for baseline in regression

N=10 in each group, alpha=2-sided 5%
Effect size = 1 SD in outcome

30

Power gain with regression rather than conducting many t-tests (1).



Power gain with regression rather than conducting many t-tests (2).

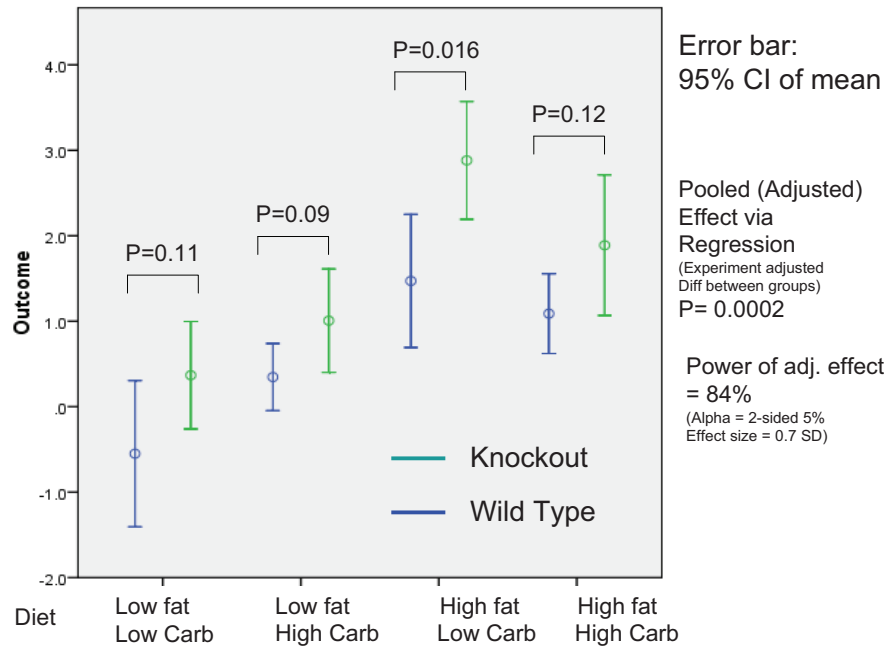
-Sample size and power computation (if many t-tests are planned)-

You can only afford 10 mice per group and based on a pilot data, we know plausible difference between knockout and wild-type mice is 0.7 SD. Power of this analysis is estimated being 30%, further with Bonferonni adjustment, we have only 12% power. I.e., to achieve 80% power to detect 0.7SD difference, total of 33 x 8 = 264 mice are required.

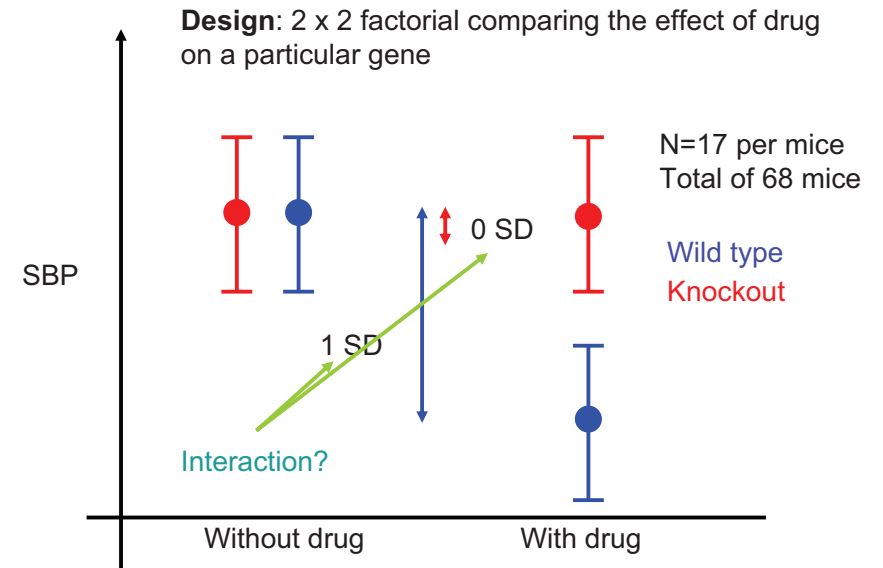
We can only afford 80 mice total, what should we do??

32

Power gain with regression rather than conducting many t-tests (3).



Power Loss to detect Interaction / Effect Modification.



Power to detect 1 SD difference between 2 group means is 80%.
However power to detect the differential effect of Drug (1 SD: 0 vs 1 SD₃₄ change by Gene is only 54%. Need 30 per group to detect this interaction.

Guideline for determining sample size for a multivariable model

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

#, number of

K: number of categories, n: total sample size, n_i : 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.

Question 1

- With a data including 100 patients, up to how many independent variables can you include in a linear regression model with a rule of 15 subjects per a variable rule?

- (A) 6
- (B) 10
- (C) 1

Question 2

- With a data including 100 patients where 90 patients had an event, how many independent variables can you include in a binary logistic regression model with a rule of 10 event per variable (EPV).
- (A) 10
- (B) 1

37

Question 3

- With a data including 100 patients where 90 patients had an event, up to how many independent variables can you include in a Cox regression model with a rule of 10 event per variable (EPV).?
- (A) 9
- (B) 10
- (C) 1

38