



NORTHWESTERN
UNIVERSITY

Biostatistics

Collaboration Center

Biostatistics in Medical Research

Lecture #4: Statistical Power

Mary J. Kwasny, ScD

Biostatistics Collaboration Center

Department of Preventive Medicine

Outline

- Importance
- Terminology
- Examples
 - Means
 - Proportions
 - Correlation coefficients
 - Time to Event (Survival)
- Take home messages

Why?

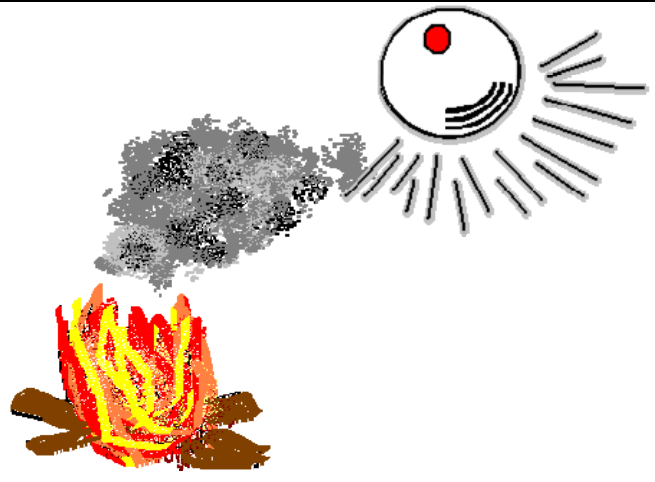
- Most granting agencies now require some sort of justification of sample size.
- A study with too much power will usually be costly, and will occasionally claim “significant” results that are not clinically relevant.
- A study that lacks power will not be “significant” – even if results are clinically meaningful. There is a known publication bias against studies with negative findings.

Fundamental point

- [Studies] should have sufficient statistical power (usually 80%) to detect differences considered to be of clinical interest between groups.
- To be assured of this without compromising levels of significance, a sample size calculation should be considered early in the planning stages.

Friedman, L.M., Furberg, C.D., and DeMets, D.L. Fundamentals of Clinical Trials, 3rd Edition. New York: Springer-Verlag, 1998.

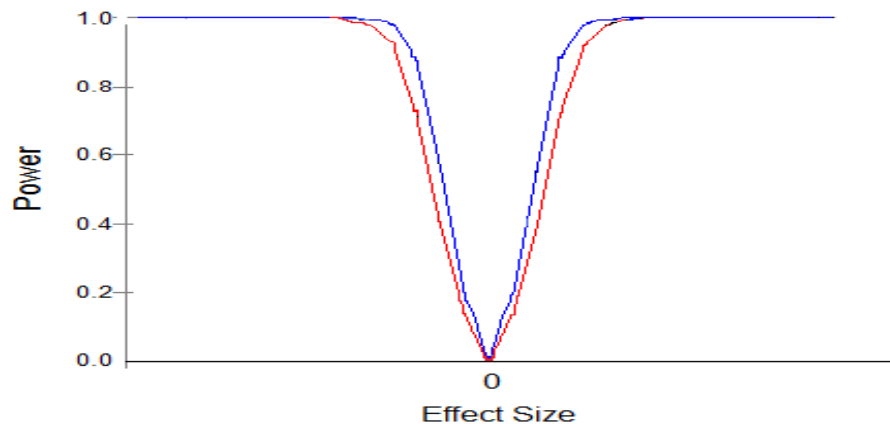
“testing” quick review

		Reality	
		No difference in Groups/Treatment (<i>H₀ true</i>)	There is some Group Effect (<i>H₀ not true</i>)
Test Result	Reject <i>H₀</i> ($p < 0.05$)	Type I Error (α) $\alpha = 0.05$ (5%)	Power 0.80 (80%)
	Fail to reject <i>H₀</i> ($p > 0.05$)	Confidence 0.95 (95%)	Type II Error (β) 0.20 (20%)

Power = conditional probability
 = $\Pr(\text{Reject } H_0 \mid \text{There is some Effect})$

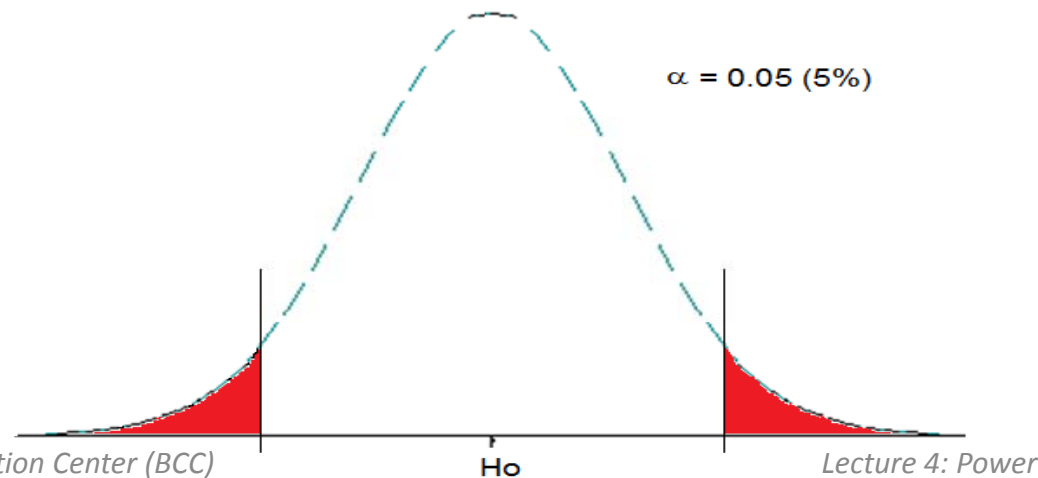
Note

- Power is vague (conditional on what, exactly?).
- In defining a “reality” we have either no effect (the null) or some effect (the alternative)
- This is OK, but makes the investigator decide some specific alternative under which to estimate power.



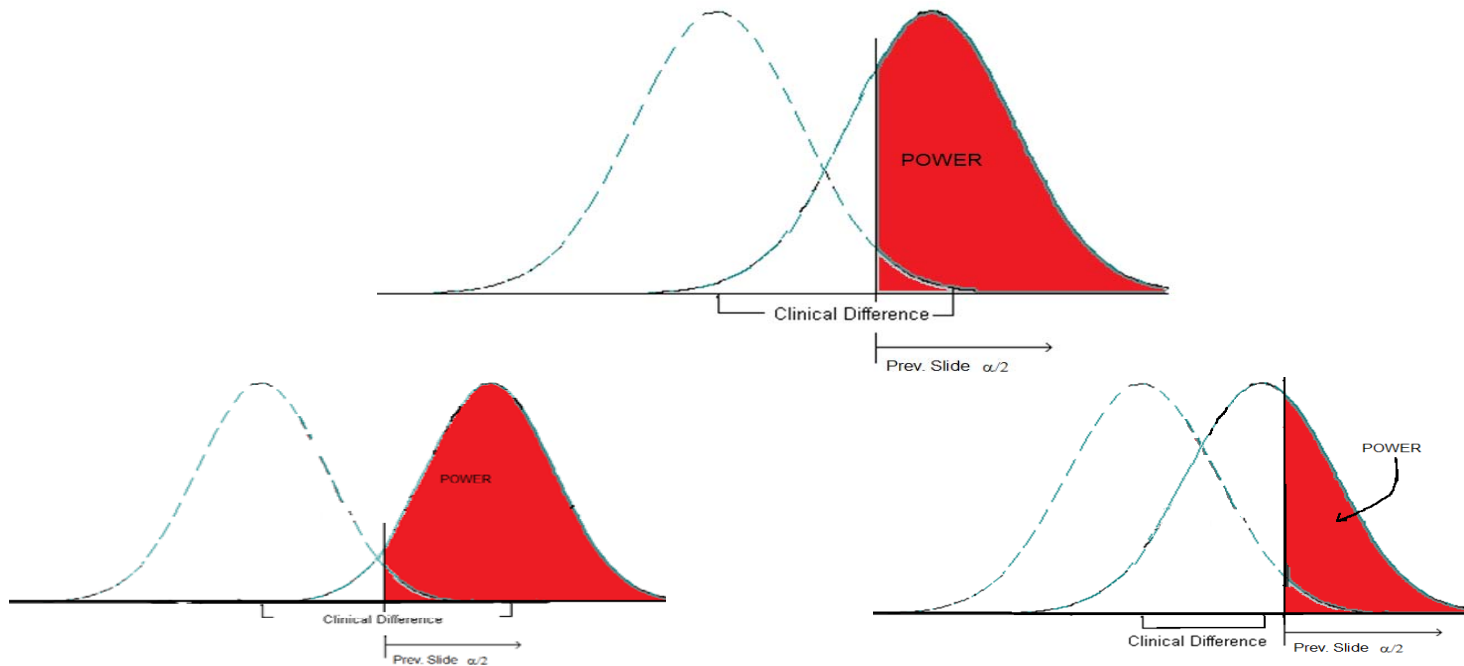
In pictures...

- Once your primary question/hypothesis is identified, a statistic that will be used to test that hypothesis is chosen. Those statistics have probability distributions (AUC=1), whose exact shape depends on sd, n. Next we consider type I errors – how extreme does the statistic need to be to be “different” (when do we reject H_0)?



Then...

- A similar statistical distribution is considered that is centered on the effect size that is expected or wanted to be detected (our specific alternative!).



Effect sizes?

Clinically relevant differences?

- In order to calculate power (or sample size), an investigator needs to have a question in mind, AND some difference (in means, rates, or median survival) in mind that would be meaningful to detect.
- Those differences might not be the same for every study. Combination of “what has an impact” (other studies?) and “how intense is intervention” (personal bias?)
- Key Point: NOT STATISTICAL!

Power (N) based on Primary outcome

- The sample size calculation should ALWAYS be based on the primary hypothesis if possible. Since that main question drives your research, you want to be sure that you can answer it.
- Sometimes, sample size calculations are based on the primary hypothesis, BUT also powered in a subset of the study (example: if you are studying the effect of aspirin on CVD, it may be of interest to power the study so that you can detect changes in gender subsets)

Power (sample size) calculations

- Usually it is possible to pare down your research question/design to a simple statistical method or a variation thereof.
- If not, there are instances where you can design a study using a more complex plan, but run power on a simpler analysis (why later)
- So, most things can be tested by comparing means, proportions, or time to event data; or examining correlations

Warning!



- The sample size calculations/presentations in this presentation all assume Simple Random Sampling.
- If the study design implements other techniques (stratified, cluster, systematic), then these formulae may not be accurate!

Power – comparing related Means

- Usually comparing related means involves a longitudinal study where we look at the change of some continuous measure.
- Looking for an overall change in the mean value is akin to looking at (after-before).
- Your study might be more interested in changes over time perhaps by treatment or group, but to calculate power for those more complicated designs, you need to assume more about your data (which may lead to inaccuracies)

Related means

- To directly calculate power, you would have to either review your calculus books (area under the curve), or rely on tables that do that for you (in this case, t-tables).
- Power = $\text{pr}(\text{reject } H_0 \mid \text{Effect size } (\Delta/\sigma))$

$$\frac{\left(t_{\alpha/2, n-1} + t_{\beta, n-1} \right) \sigma}{\Delta} \approx \sqrt{n}$$

...better yet

- Computer programs (PASS, n-Query, R, SAS...)
- You specify parameters, it will solve for others.
- This will solve for (mean, power, alpha, sigma, or n)

The screenshot shows the PASS software interface for 'Inequality Tests for One Mean (One-Sample or Paired T-Test)'. The window title is 'PASS: Inequality Tests for One Mean (One-Sample or Paired T-Test)'. The menu bar includes File, Run, Means, Proportions, Correlation, Regression, Survival, ROC, Variances, DOE, and T. The toolbar contains icons for RUN, NEW, OPEN, SAVE, PASS, MAP, OUT, MACRO, DIFFS 2PROP, 2 S T-TEST, 1 WAY ANOVA, R M ANOVA, CNTRL MC-S, HR LRNK, LINEAR REG, and LOGIST REG. The main panel is divided into several sections:

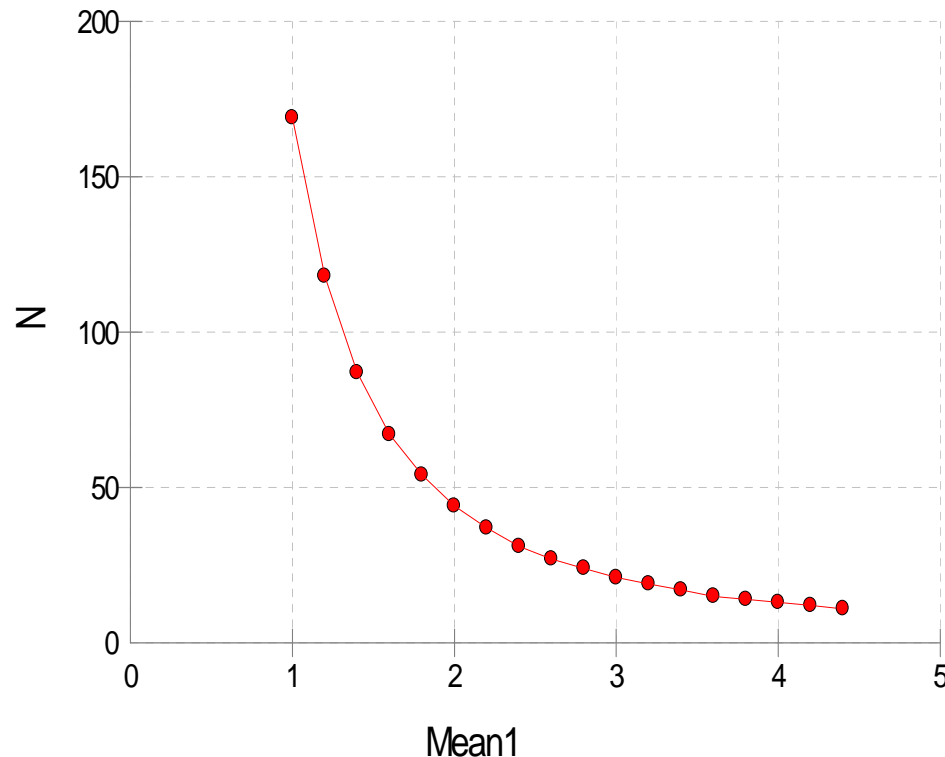
- Solve For:** Find (Solve For): Mean1 (Search > Mean0)
- Error Rates:** Power (1-Beta): .8; Alpha (Significance Level): 0.05
- Sample Size:** N (Sample Size): 16 34
- Effect Size:** Means: Mean0 (Null or Baseline): 0; Mean1 (Alternative): 8.4 19.0 5; Standard Deviation: S (Standard Deviation): 1; Known Standard Deviation; Standard Deviation Estimator
- Test:** Alternative Hypothesis: Ha: Mean0 <> Mean1; Nonparametric Adjustment (Wilcoxon Test): Normal; Population Size: Infinite

At the bottom of the window, it says 'One-Sample Design: Mean = Population Mean' and 'Paired Design: Mean = Mean of Pair Diffs'. The footer of the window displays 'Lecture 4: Power' and the number '15'.

Example: paired t-test

- If we go back a few lectures, you recall the example of SBP and OC use for $n=10$. Let's consider that pilot data. They saw a mean change of 4.8; $sd=4.6$.

Sample PASS Output for paired t-test



BASED on power = 80%,
SD of change = 4.6, and $\alpha = 0.05$.
Using a one sample (two sided) t-test.

- So, if you thought that a 1mmHg change in SBP was relevant, then you would need 169 women in your study.
- On the other hand, if you thought you could recruit 50 women, you could detect a mean change of 1.9mmHg.

Power - comparing independent means

- Comparing 2 groups is a little more complex than looking at just one – there are more parameters involved.
- Here you can solve for means, either sd or n, power, or alpha – but you will have to estimate the others.

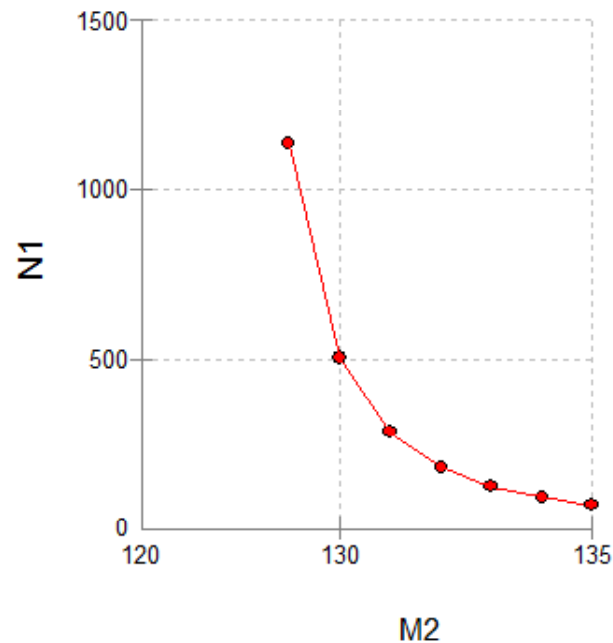
The screenshot shows the PASS software interface for 'Inequality Tests for Two Means (Two-Sample T-Test) [Differences]'. The interface is divided into several sections:

- Solve For:** Find (Solve For): Power and Beta
- Error Rates:** Power (1-Beta): 0.8; Alpha (Significance Level): .05
- Sample Size:** N1 (Sample Size Group 1): 30 to 70 by 10; N2 (Sample Size Group 2): Use R; R (Sample Allocation Ratio): 1.0
- Effect Size:** Means: Mean1 (Mean of Group 1): 2.30; Mean2 (Mean of Group 2): 1.85; Standard Deviations: S1 (Standard Deviation Group 1): 0.91; S2 (Standard Deviation Group 2): 0.46; Known Standard Deviation; Standard Deviation Estimator
- Test:** Alternative Hypothesis: Ha: Mean1 <> Mean2; Nonparametric Adjust. (Mann-Whitney Test): Uniform

Example: 2 independent means

- Again, from the lecture comparing 2 groups, recall the study design where SBP was compared between OC users and OC non-users (consider that preliminary data).
- They saw: mean (SD) of 133(15) and 127(18).
- Since there is no reason to believe that the SDs could be different, let's assume $SD=17$.

PASS output– 2 independent means

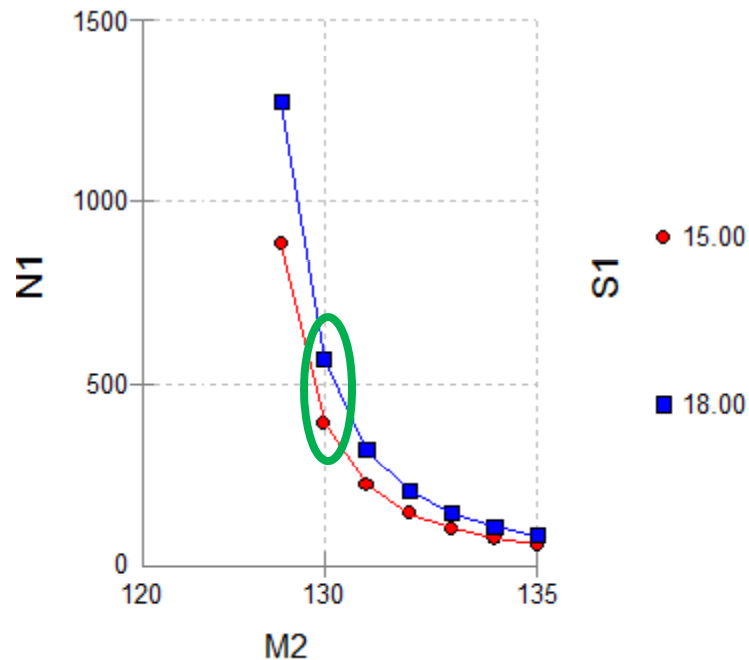


Assumes: power = 80%,
mean of non-users of 127,
equal sd=17, $\alpha = 0.05$, and
equal sample sizes.

Uses a two-sample t-test.

- You would need 183/group (total N=366) to have 80% power to detect a 5mmHg difference in groups.
- If you could recruit 100 total, you would only have power to detect differences larger than 9.6mmHg.

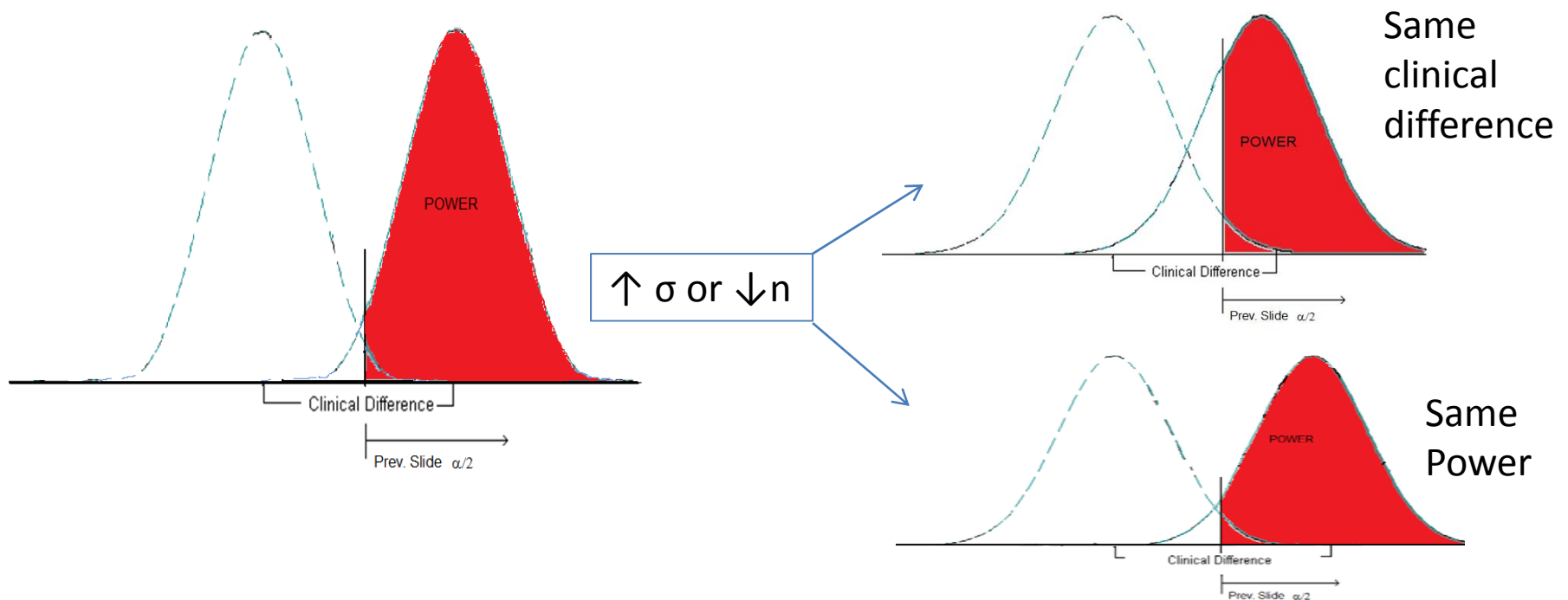
Sensitive to estimates!



- Statisticians always want to know about variability.
- Here if the difference that we want to find was 3mmHg – if you assume a common SD of 15, you would need 393/group. If you assume a common SD of 18, you would need 566/group!

WHY?

- Again, the distribution of the test statistic depends on the standard error (σ/\sqrt{n})



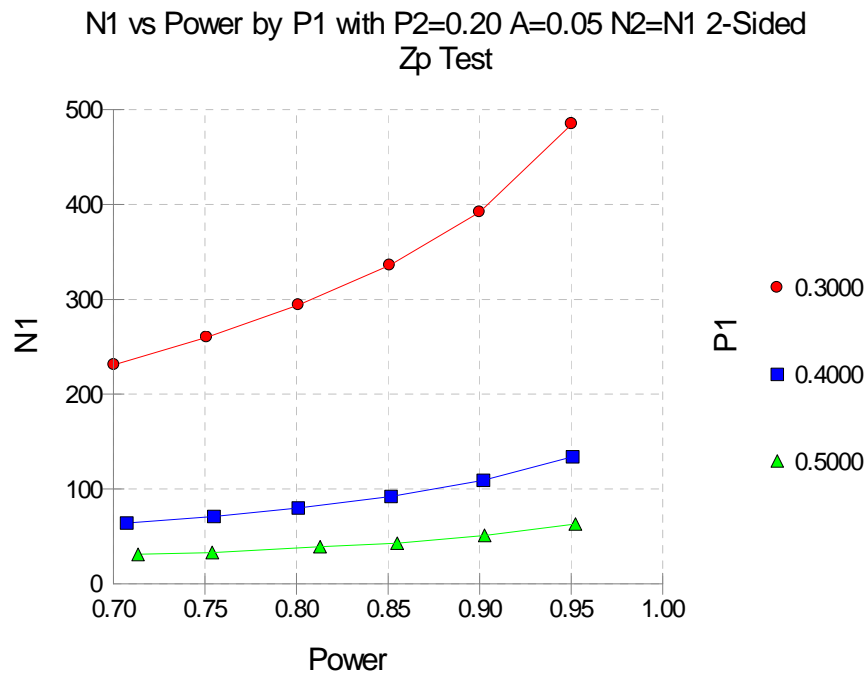
Power – comparing proportions

- Power for proportions is not quite as complicated as power for means. As you may recall from the first lecture, the variability of the proportion is a function of the proportion and sample size, so we don't need to estimate an extra parameter.

Example

- Although dependent on many factors, some randomized double-blind trials report a placebo response rate of 20%
- An investigator (filled out an IND to study) an off-label use of antidepressants for pain.
- Outcome of interest: Was your pain relieved?

Example: Chi-sq test ($\alpha=0.05$)



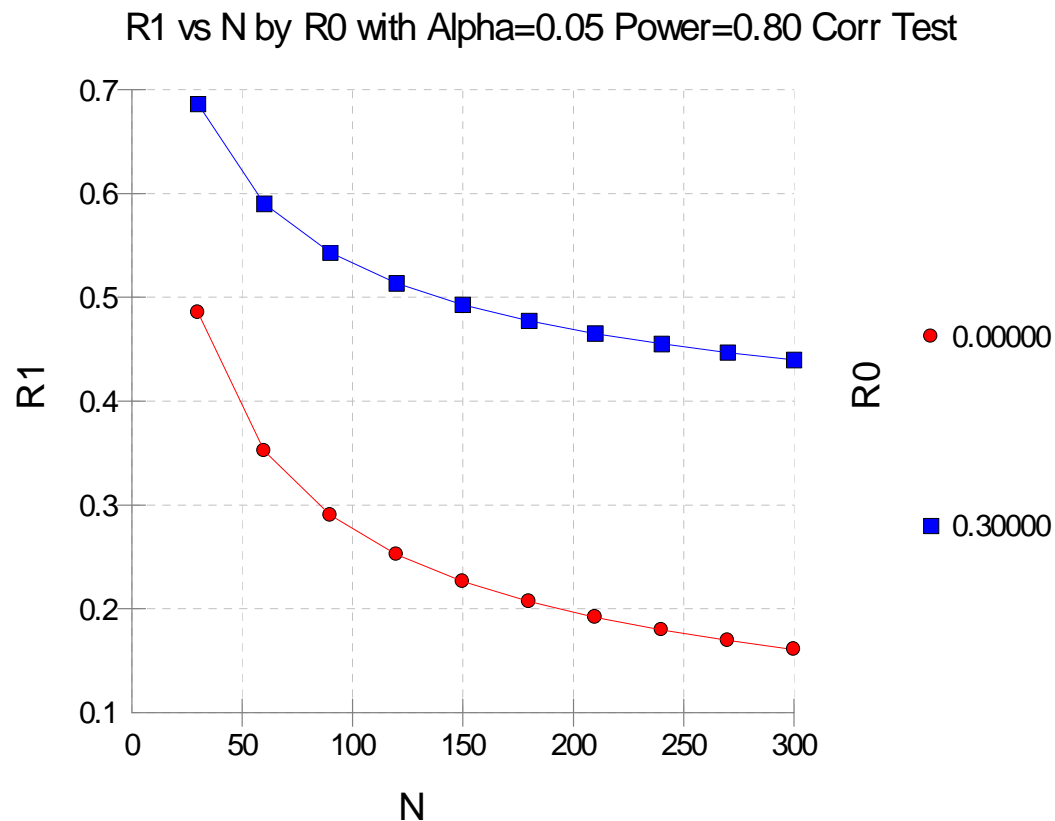
- Assuming a response rate of 20% in the control group,
- 80% power:

Treatment Response	Total N (balanced)
25% (not shown)	2188
30%	588
40%	160
50%	78

Correlations

- Pearson Correlation coefficients are used to quantify the amount of linear association between two continuous (normally distributed) variables.
- While debate rages in statistical circles if it is appropriate... usually correlation coefficients are tested against a null value of 0 (no linear association). The squared Pearson's correlation coefficient can be interpreted as the amount of variability in one factor that explains the other. (correlation of .3 => 9% of the variability can be explained)

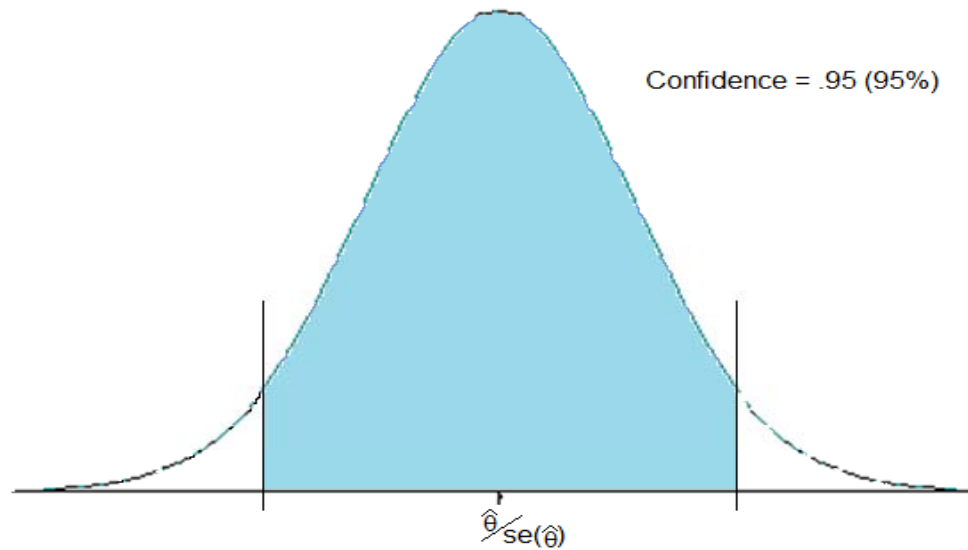
Sample PASS output



- N=300 would have 80% power to “detect” correlations as small as 0.161
- If you wanted 80% power to detect an $r = 0.5$, you would only need 29

Margin of Error

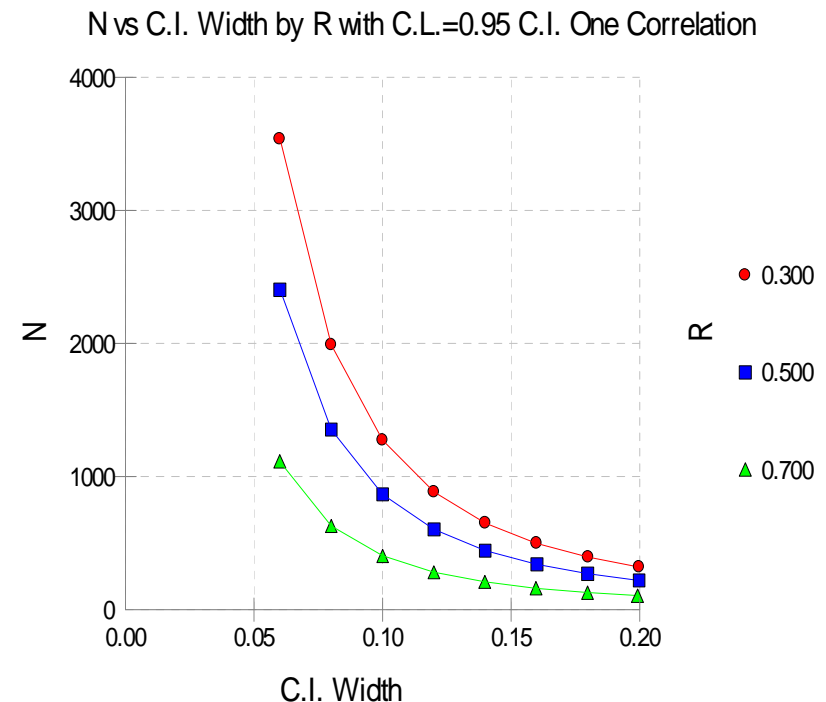
- Correlations are usually estimated for descriptive studies rather than trials; thinking in terms of a “margin of error” may be more appealing.
- Margin of error reflects how precise you would like to estimate an effect (or correlation)
- Conf. Interval: $\text{Est.} \pm t_{\alpha/2, n-1} \text{SE}(\text{est.})$



$$SE(\hat{r}) = \sqrt{\frac{1-r^2}{n-2}}$$

How wide?

- Confidence intervals “describe” what is going on in your data (without specifying a null hypothesis)
- If you think your correlation is about .5, you would need $n=219$ to estimate it $\pm .1$ (width of .2)



Survival data

- Stay tuned next week... but
- Survival data is also known as time-to-event data. Usually these data also have a chance of being “censored” – that is, the subject does not have a chance of experiencing event of interest after some time (e.g. breast cancer recurrence; patient is followed for some time, but then moves (lost to follow-up), or patient dies of non-related causes (car accident))
- Statistics of interest: median time to event (or median survival) or survival compared at a specific time.

Survival

- Log rank test compares “survival” at specific time.
- The more complicated the analysis is, the more parameters in the model need to be estimated.

Solve For
Find (Solve For):
Power and Beta

Error Rates
Power (1-Beta):
0.8 0.9
Alpha (Significance Level):
.05

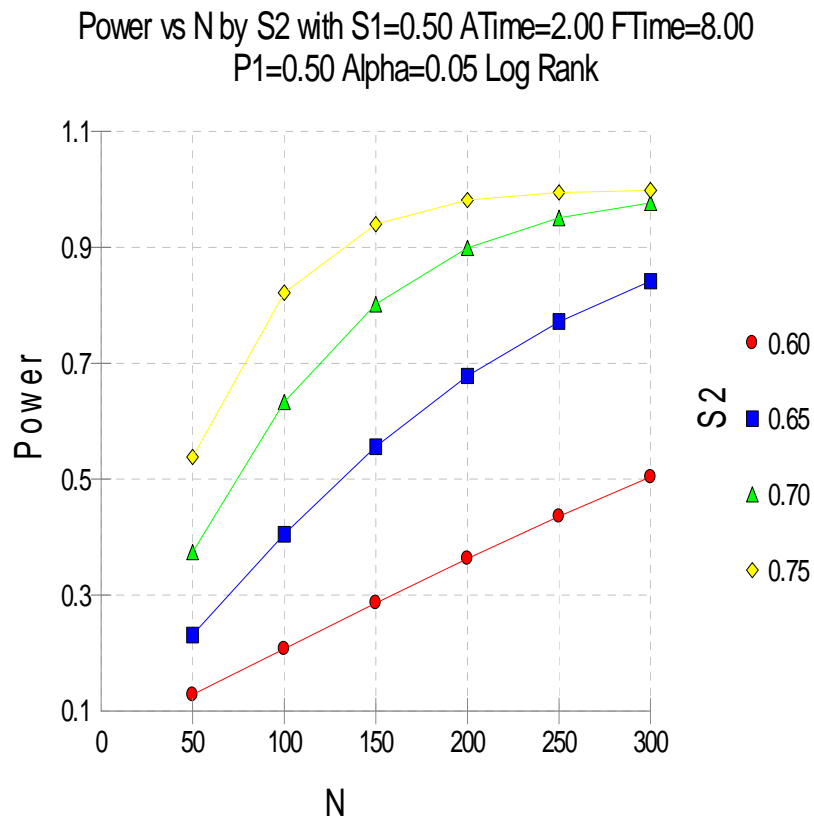
Effect Size
S1 (Proportion in Group 1 Surviving Past T0):
0.50
S2 (Proportion in Group 2 Surviving Past T0):
0.6 to 0.75 by .05
T0 (Fixed Time Point):
5
Survival Parameter Conversion Tool

Sample Size
Total Sample Size
N (Total Sample Size):
50 to 300 by 50
Sample Proportion
Proportion in Group 1:
0.5
Proportion Lost to Follow-Up
Group 1: 0.20
Group 2: 0.20

Duration
R (Accrual Time):
2
% Time Until 50% Accrual:
50
Follow-Up Time, T-R:
8

Test
Alternative Hypothesis:
Ha: S1 <> S2

PASS for Logrank Test

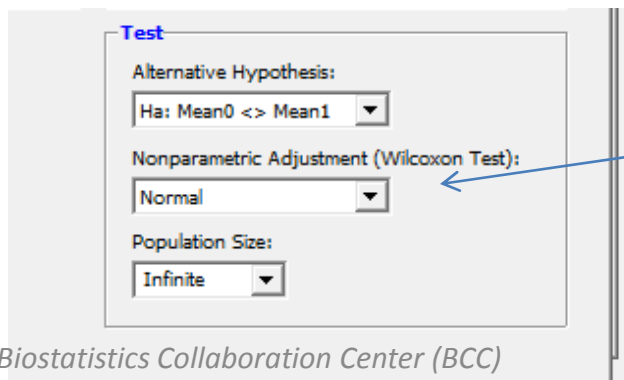


- 50% 5-year survival in group 1, 60-75% 5-year survival in group 2.
- 20% censoring in both groups
- Accrual time 2 years (50% accrued after 1 year); study duration 8 years.

Other Issues in Power or Sample Size calculations....

Non-parametric tests? (medians)

- Non-parametric tests have power calculations, too. Although those usually are done based on corresponding parametric tests and “adjusted.”
- If assumptions regarding parametric tests are met, usually non-parametric tests will have about 95% as much power as parametric ones. If assumptions are not met, non-parametric methods may actually have more power.



From PASS - two sample t-test data sheet

Other “N” calculations

- Number needed to screen?
 - If you need to confirm eligibility in the study, you need to adjust for “eligibility” rates
 - Not everyone eligible may want to be in the study
- Drop out/compliance?
 - If you are planning a study that involves follow-up or compliance, you should account for drop out rates (and compliance if planning anything other than “intent to treat” analysis)

Interim or Post-hoc power?

- Occasionally a study will conduct a post-hoc power calculation (most usually on a “negative” study – or if recruitment is not going well)
- These are power calculations done using information (effect sizes) obtained in the study to see how much larger a sample they needed to get to “achieve statistical significance”
- While some journals may still request something like this, some statisticians will not even consider it. There is a strong relationship between p-values and power, and while these might make sense for “future research” usually the goal is to explain “insignificant” findings.

Maximizing your (power) interaction with a statistician

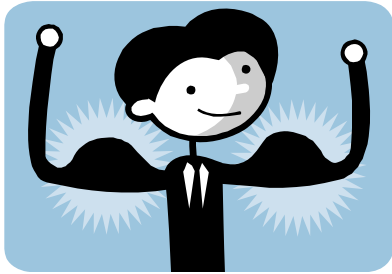
- Simple designs and questions need less preliminary information (although the more you have the better – e.g. some ballpark estimate for spread, effect sizes, “control” rates, etc.)
- More complex studies may require extensive literature review to get estimates of sample parameters (if not preliminary data)
- If you have potential sample size limitations, that may guide the power calculation as well!

Warnings for DIYs

- There are several programs (free or relatively inexpensive) and websites available that will run calculations for you.
- Be aware that not all websites/packages have been validated nor is it clear what assumptions some programs use (For example, some packages default to an exact test for small n – but some do not).
- Also be aware that some packages may default to different quantities –a power calculation for 2 proportions may compare p_1 and p_2 , OR it may compare p_1 and $p_2 - p_1$.
- Also be aware that some programs default to α and β (not α and power = $1 - \beta$).

But I really want to run...

- Although an analysis plan might call for generalized linear models or other more complicated analyses, and adjusted for many other factors, a simple power calculation is better than none, and may be much better than one based on speculation! (e.g. a power analysis for RM-ANOVA needs to specify means and within and between errors, as well as correlations and autocorrelations for the within-person factor.)
- If your field is so advanced such that advanced techniques are the norm (fMRI, genetics), then you need to really search the literature (or your databases) to justify assumptions that you will need to run power calculations; although in many of those areas, available “n” may be more influential.



Final thoughts...

- With great power comes great responsibility. (Stan Lee, FDR?)
 - Large power = large n and/or huge effect sizes. Either could lead to [statistical] abuses. It is always vital to keep clinical relevance in the picture.
- Simplify, simplify, simplify! (HDT)
 - Although tempting to write up a complicated analysis with power, usually those calculations are speculative at best.

References

- Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd Ed. Hillsdale, NJ: Erlbaum.
- Hoenig and Heisey. (2001) The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician*, 55,19-24.
- Goodman, SN and Berlin, JA (1994), The use of Predicted Confidence Intervals when Planning Experiments and the Misuse of Power when interpreting results. *Annals of Internal Medicine*. 121, 200-206.
- Lang and Secic. (1997) How to Report Statistics in Medicine. Philadelphia: American College of Physicians.
- Your friendly neighborhood Biostatisticians at the BCC!

Next Week:
SURVIVAL
(Analysis of Time to Event Data)