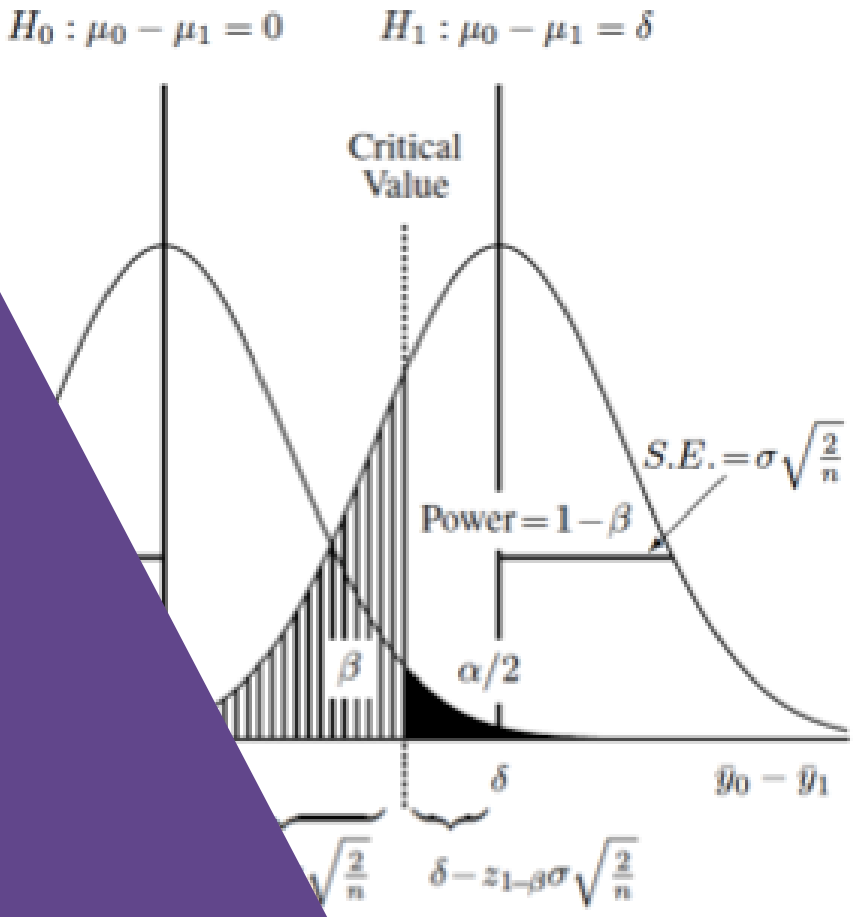
Northwestern Medicine®
Feinberg School of Medicine

Statistical Power and Sample Size: what you need and how much.

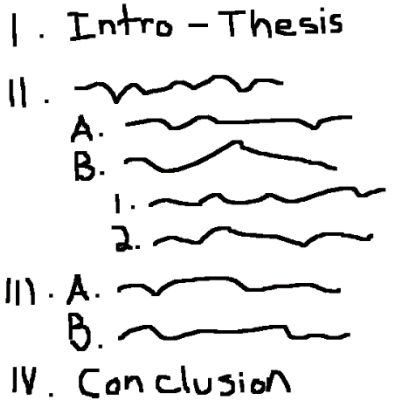
Mary J. Kwasny, ScD



Power is the most persuasive rhetoric (Fredrich Schiller), *but the greater the power, the more dangerous the abuse* (Edmund Burke)

Outline

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



Why?

- Most granting agencies (and some journal editors) now require some sort of justification of sample size.

- A study with too much power will usually be costly, and will often claim statistically significant results that are not clinically relevant.
- *Big data can lead to many "false hopes" that certain associations show promise*

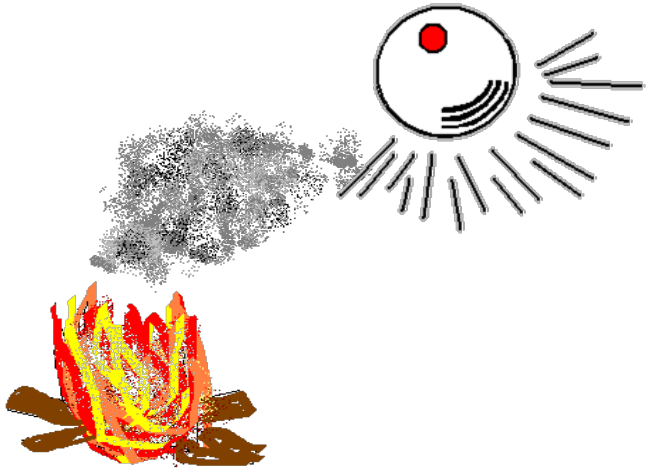
- A study that lacks power may not be statistically significant – even if results are clinically meaningful.
- *There is a known publication bias against studies with negative findings*.

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.

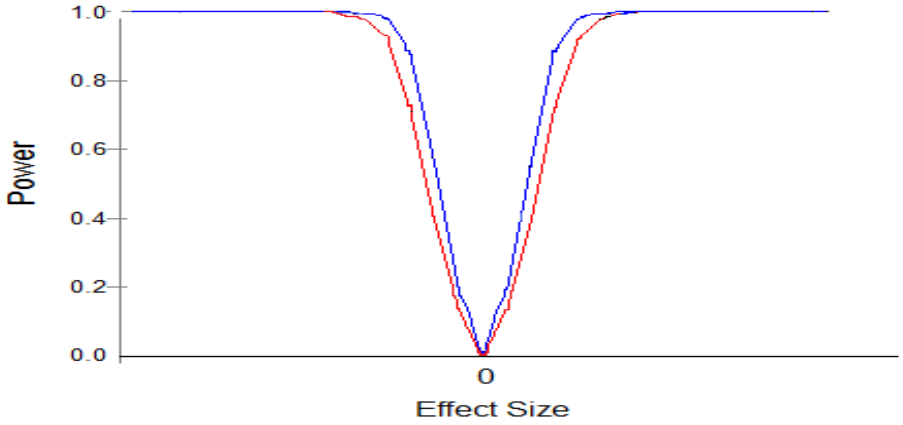
“testing” quick review

		Reality	
		No difference in Groups/Treatment *(Ho true)* All clear!	There is some Group Effect *(Ho not true)* FIRE!!
Test Result	Reject Ho (p < 0.05) ALARM!	Type I Error (α) α= 0.05 (5%)	Power 0.80 (80%)
	Fail to reject Ho (p > 0.05) All clear!	Confidence 0.95 (95%)	Type II Error (β) 0.20 (20%)

Power = conditional probability
= Pr(Reject Ho | There is some Effect)

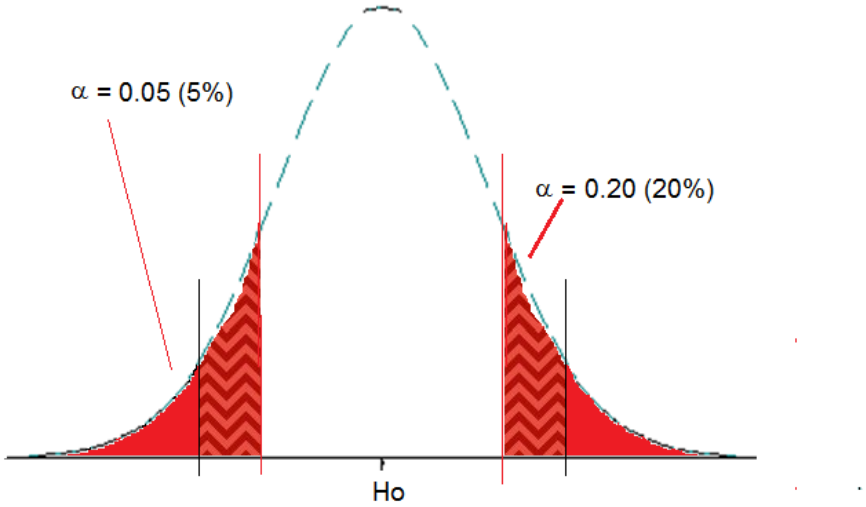
Note

- Power is vague (conditional on what, exactly?- an unknown reality!).
- 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 based on study design, etc. Those statistics have probability distributions (AUC=1), whose exact shape depends on certain parameters and N *(today we assume normality)*.
- Next we consider type I errors – how extreme does the statistic need to be to be "different" (when do we reject Ho)? – conditional on NO true difference



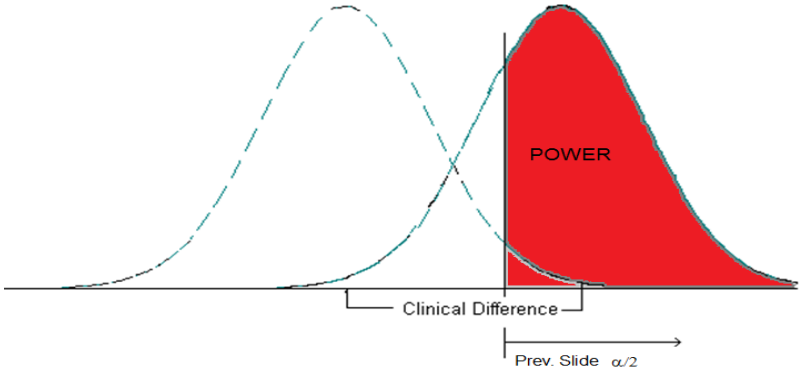
1. How "sensitive" a fire alarm?

large α (ok with "false alarms")=> always blaring

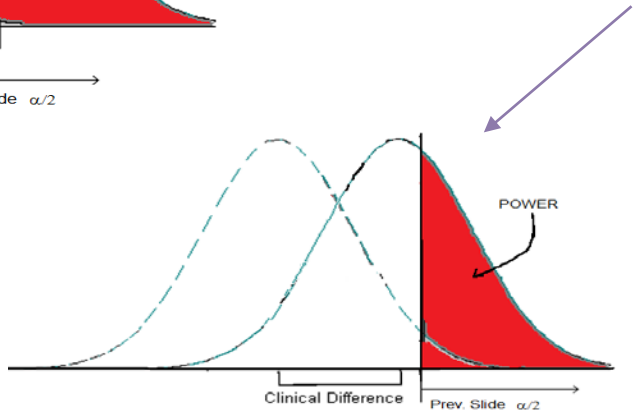
small α (never want a "false alarm") => never blaring

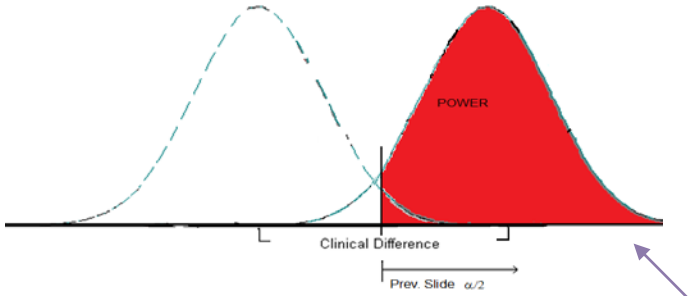
Then…

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



Low power to detect a small flame





High power to detect a conflagration

2. How big or small an <u>effect</u>?

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!

- *Although, based on the question, statistics can help with study design to have an efficient way to detect said differences.*

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.
- If you have multiple primary hypotheses, you should consider adjusting (lowering) your type I error.
- 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.
  - *NOT helpful if the design entails cluster randomization or other independent observation issues…*
- 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)
  - *Helpful for designs involving repeated measures over time or composite scores…*
- 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 AND EXAMPLES ARE NOT 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

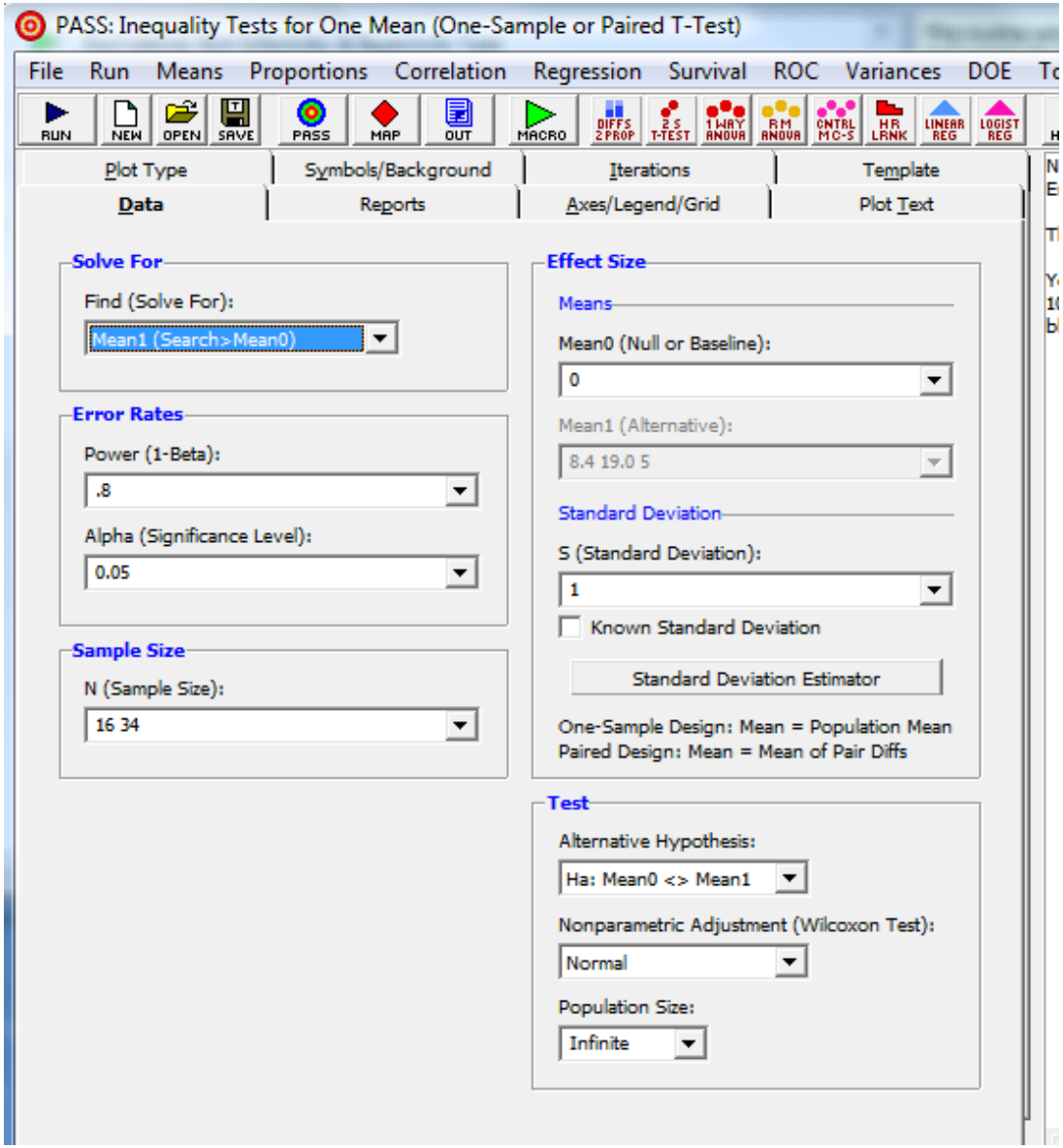
- To directly calculate power, you would have to either review your calculus books (area under the curve), or rely on tables that do that for you (in this case, t-tables).
- Power = pr(reject Ho|Effect size (Δ/σ))

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

Note: this formula is derived from probability models assuming SRS, independence of observations, etc…

…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)

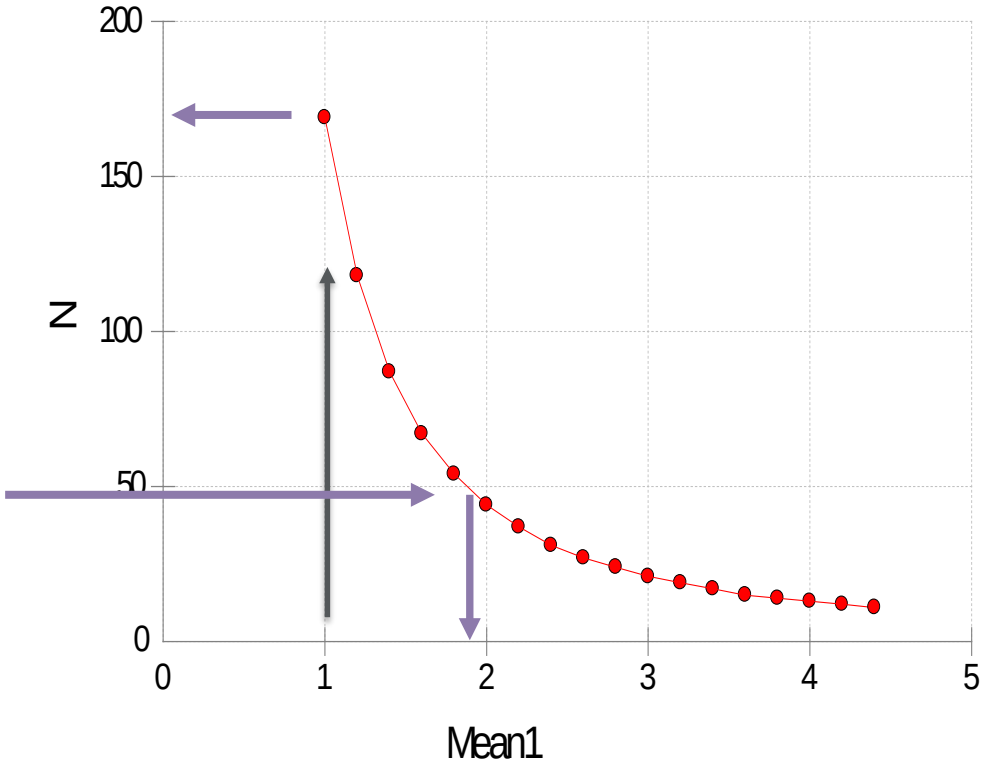


Example: paired t-test

- Lets say you want to examine systolic blood pressure (SBP) for women who have started oral contraceptives (OC). You saw a recent article in the literature for a small study that saw a mean change of 4.8mmHg and standard deviation (SD) of the *change* =4.6mmHg when women started OC use.
- Consider that pilot data.
- The thing we would be most interested in here is their estimate of the standard deviation of the change in SBP. (SD=4.6mmHg)
- … here you do need to be careful if they are reporting the standard deviation or the standard error.

Sample PASS Output for paired t-test

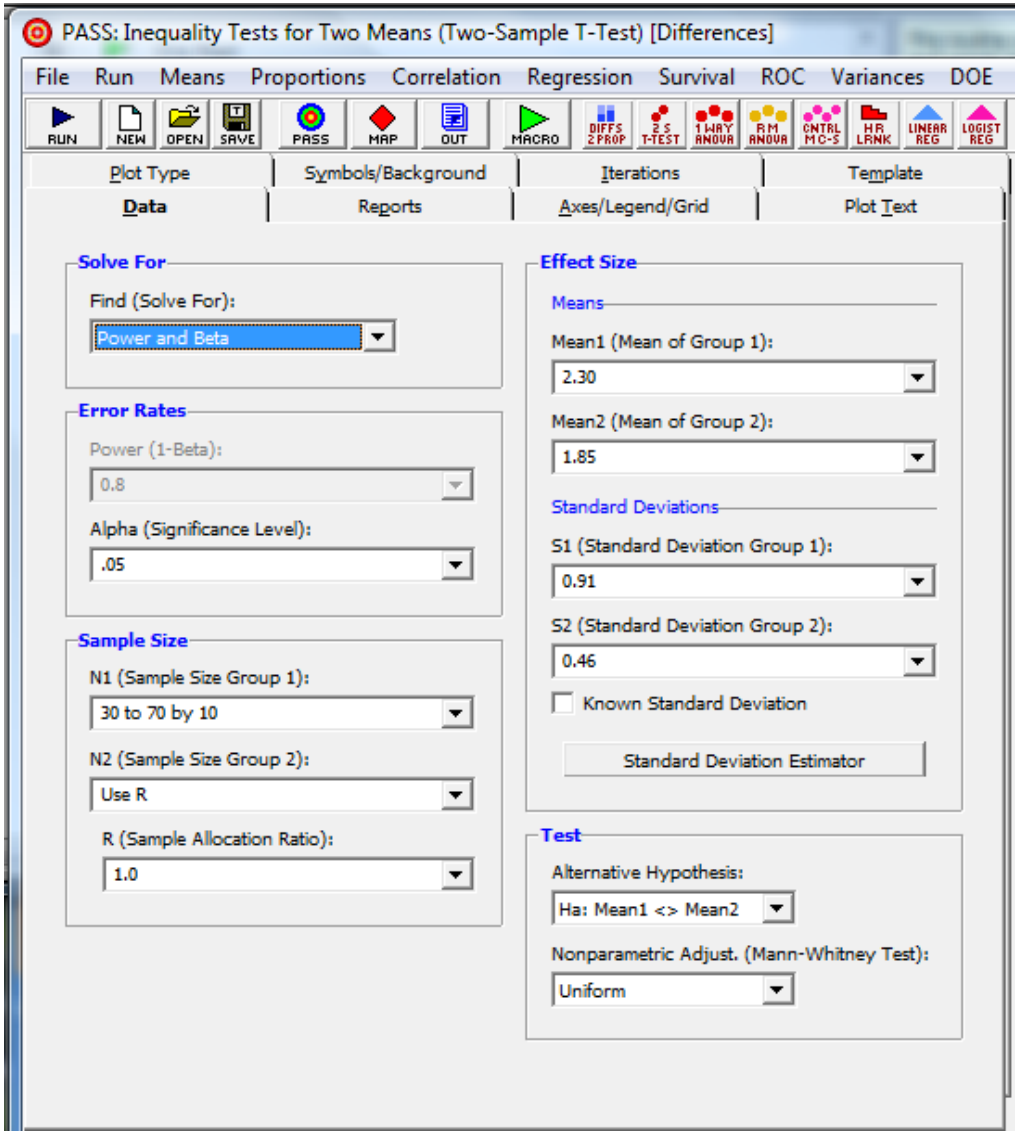
BASED on power = 80%,
SD of change = 4.6, and $\alpha = 0.05$.
Using a one sample (two sided) t-test.



- 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

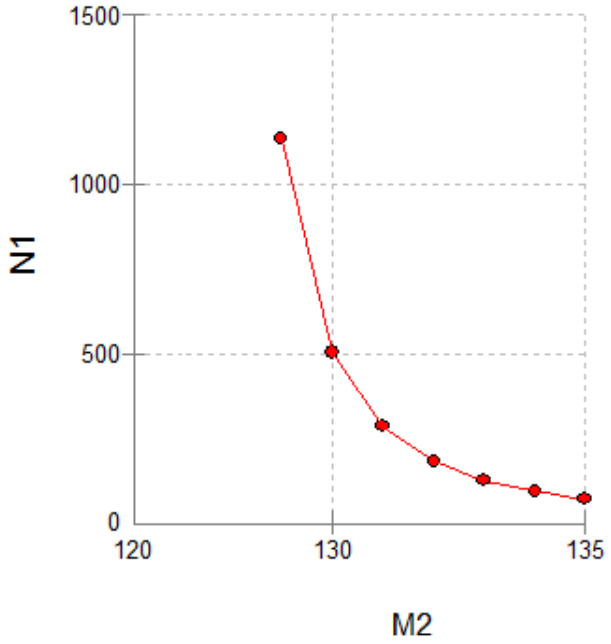
- Comparing 2 groups is a little more complex than looking at just one – there are more parameters involved.
- Here you can solve for means, either sd or n, power, or alpha – but you will have to estimate the others.



Example: 2 independent means

- Let's say we want to compare SBP was between OC users and OC non-users.
- A literature search for a similar population, but perhaps looking at different outcomes presented means (SDs) of 133(15) and 127(18) for the two groups (perhaps in a table 1).
- Since there is no reason to believe that the SDs could be different, let's assume they are and use SD=17 as an estimate; additionally assume a SBP of 127 for the non-users.

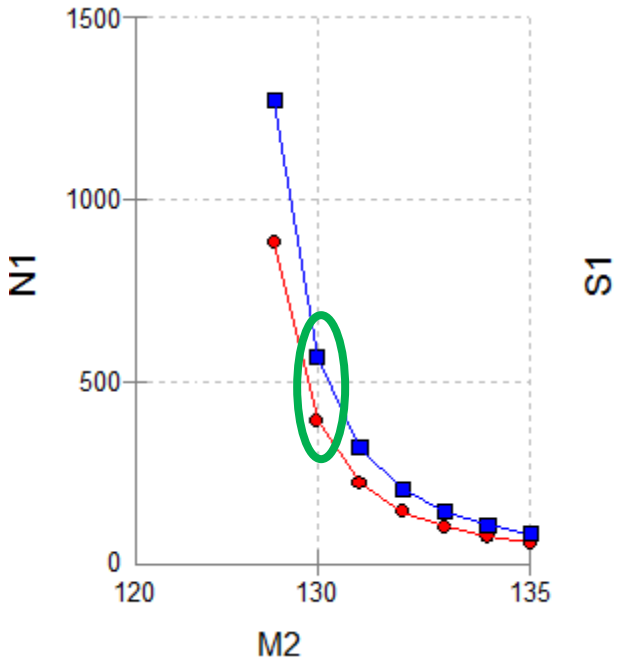
PASS output– 2 independent means



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.

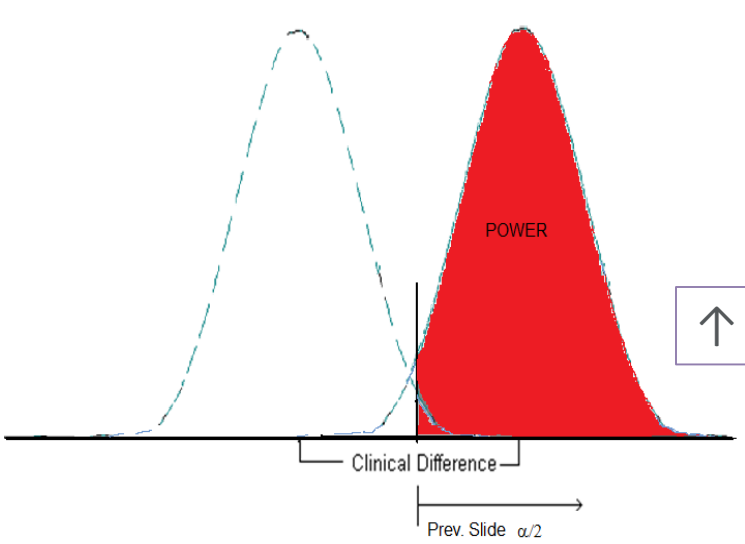
Sensitive to estimates!



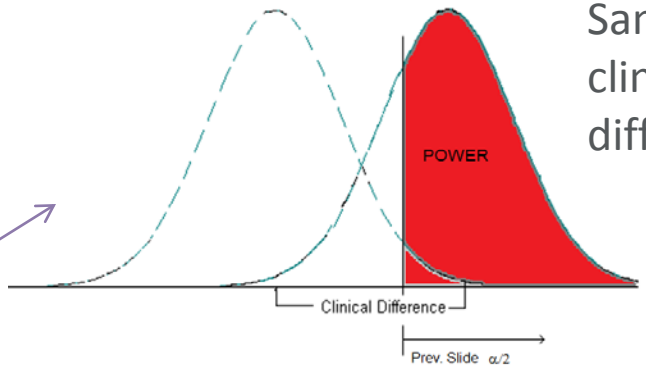
- Statisticians always want to know about variability.
- If the difference that we want to detect was 3mmHg –
- assume a common SD of 15, you would need 393/group.
- assume a common SD of 18, you would need 566/group!

WHY?

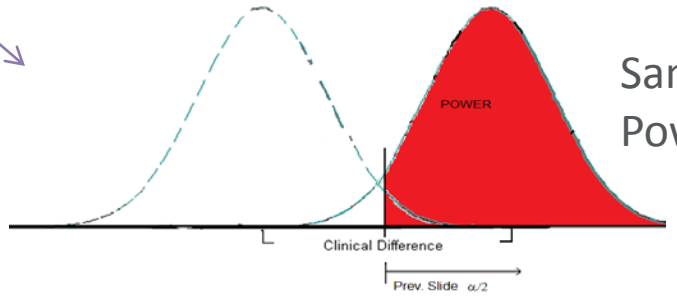
- Again, the distribution of the test statistic depends on the standard error (σ/√n)



↑ σ or ↓n



Same clinical difference



Same Power

Power – comparing proportions

- Power for proportions is not quite as complicated as power for means.
- The reason is that the variability of the proportion is a function of the proportion and sample size, so we don't need to estimate an extra parameter.

- HOWEVER the wording can get confusing…
- 10% increase…

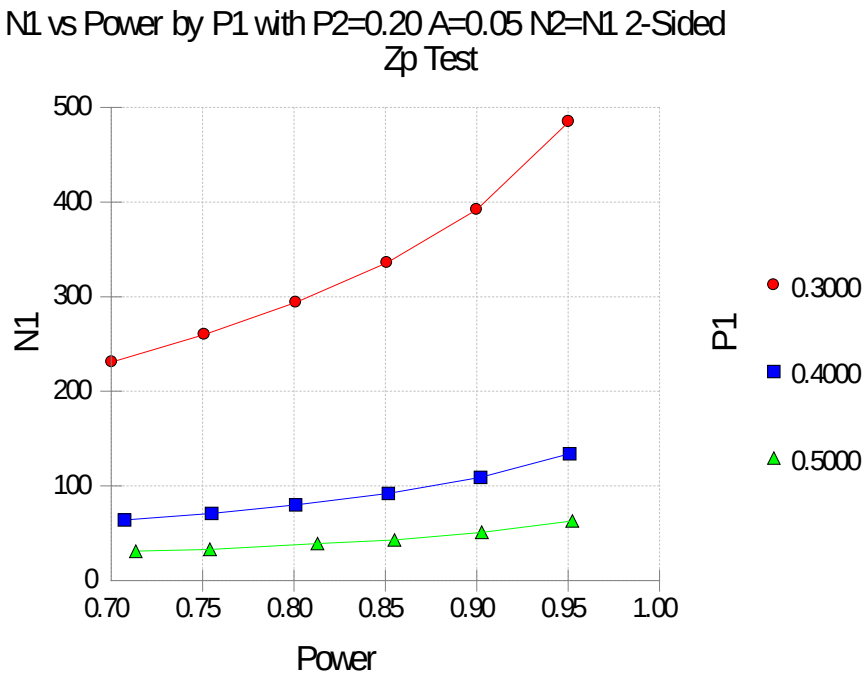
20% to 30% OR 20% to 22%

???

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 (α=0.05)



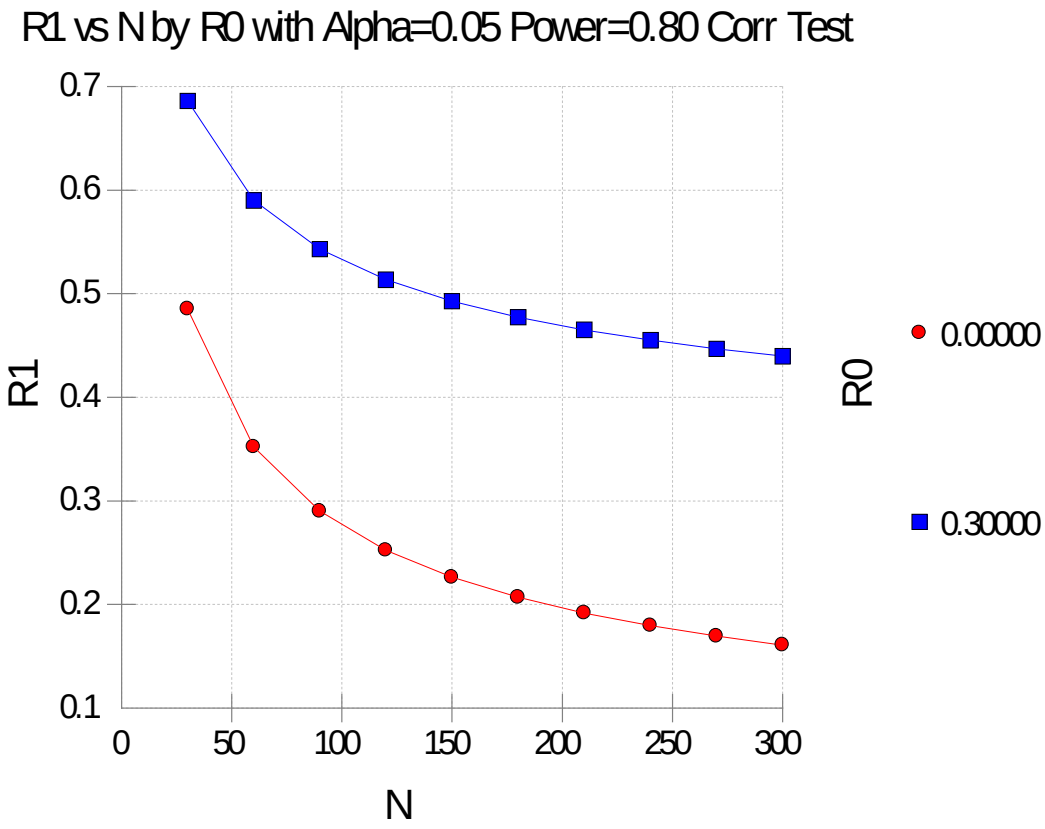
- Assuming a response rate of 20% in the control group,
- 80% power:

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)

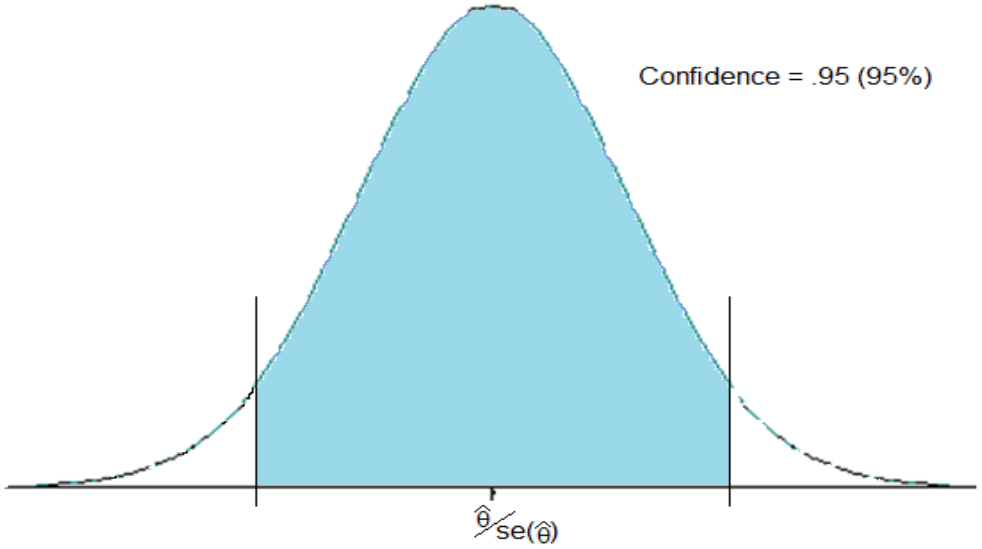
Sample PASS output



- 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

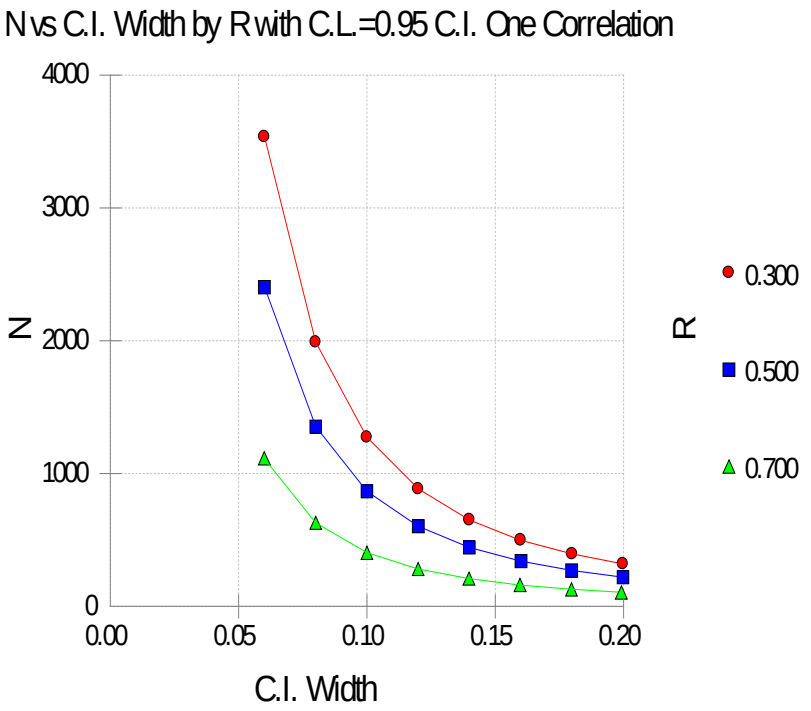
- Correlations are usually estimated for descriptive studies rather than trials; thinking in terms of a "margin of error" may be more appealing.
- Margin of error reflects how precise you would like to estimate an effect (or correlation)
- Conf. Interval: Est. ± $t_{\alpha/2,\, n-1}$SE(est.)



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

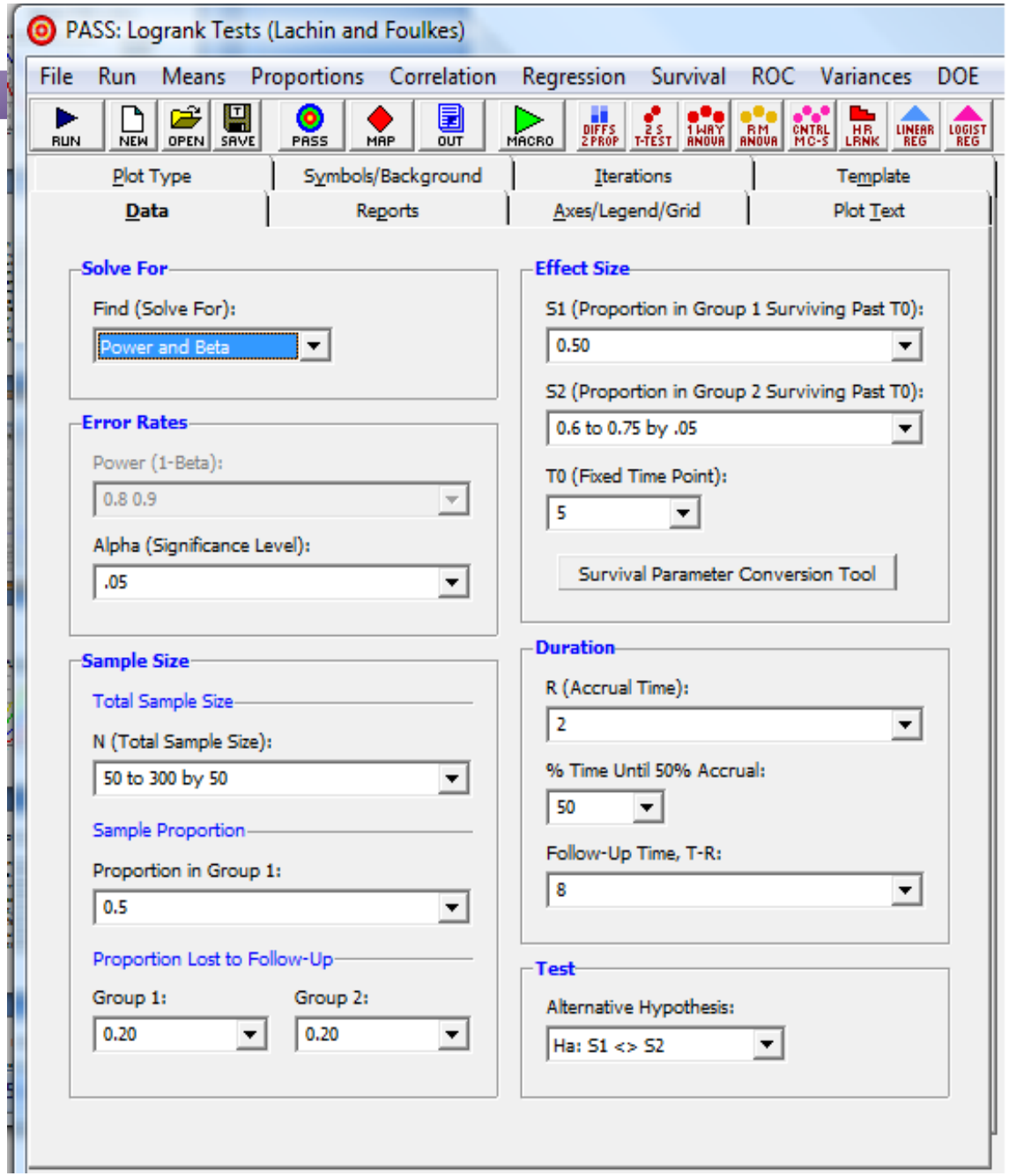
How wide?

- 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 ± .1 (width of .2)



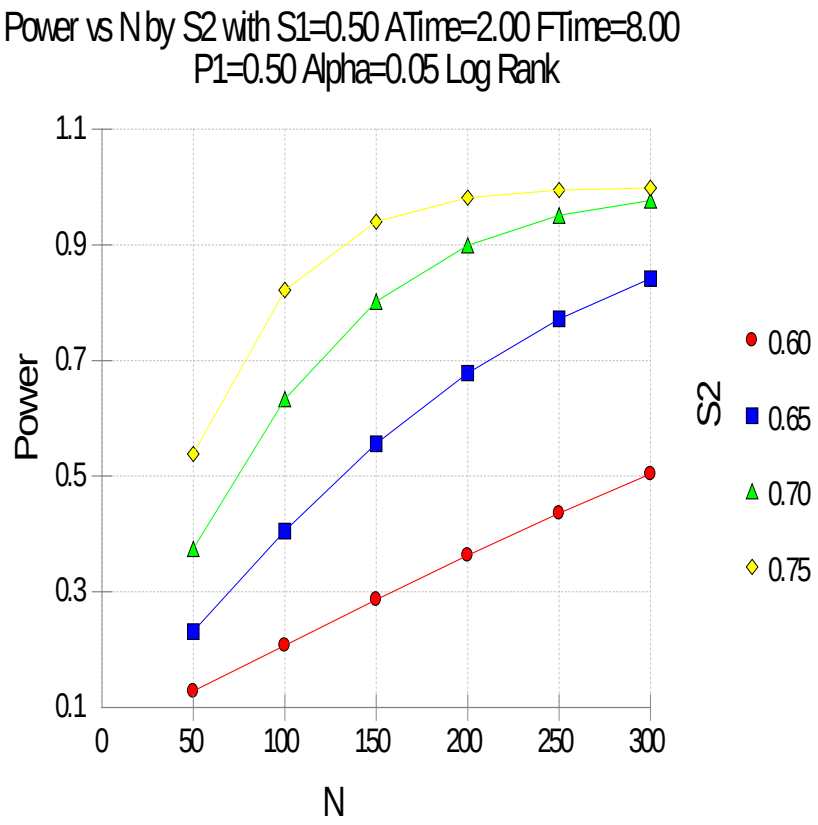
Survival data

- 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.



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

PASS for Logrank Test

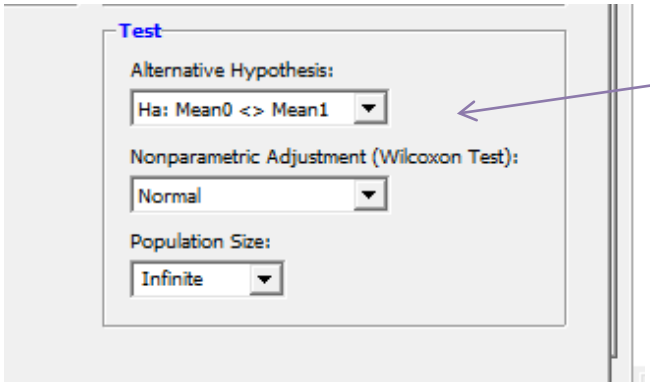


- 50% 5-year survival in group 1, 60-75% 5-year survival in group 2.
- 20% censoring in both groups
- Accrual time 2 years (50% accrued after 1 year); study duration 8 years.

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 for a DSMB 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, many 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). You may want to consider similar outcomes in different populations, or similar populations with different outcomes…
- If you have potential sample size limitations, that may guide the power calculation as well!

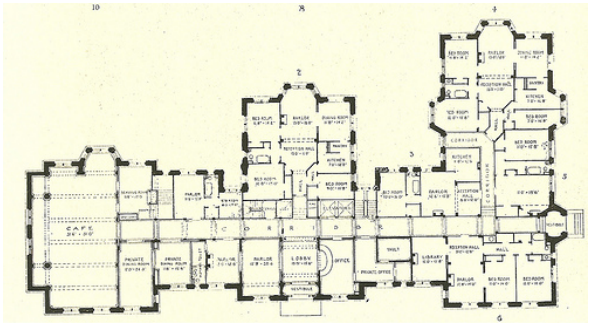
It may take a while…

- Some statistical analysis plans cannot be easily simplified.
- In those cases, it may be necessary for a statistician to run a simulation to determine the best estimate of power/sample size.
- While we appreciate that budgets tend to revolve around sample size, these simulations can take time.
- The earlier you seek assistance, the better!

Warnings for DIYs

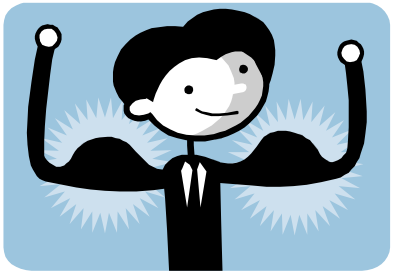
- There are several programs (free or relatively inexpensive) and websites available that will run calculations for you.
- Be aware that not all websites/packages have been validated nor is it clear what assumptions some programs use (For example, some packages default to an exact test for small n – but some do not).
- Also be aware that some packages may default to different quantities –a power calculation for 2 proportions may compare p1 and p2, OR it may compare p1 and p2-p1.
- Also be aware that some programs default to α and β (not α and power = 1-β).
- ALWAYS report the program or website that you used to calculate your sample size! (Statistical reviewers should be able to replicate your calculations)

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. (In a future lecture controlling type I error rate for studies with extraordinary levels of power will be discussed!)

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.
- Get a second opinion?
  - For DIYers… it never hurts getting a second opinion on your power analysis. Initial consults with the BCC are free, and it may save you in the long run!

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!

BCC: Biostatistics Collaboration Center

Who We Are

Leah J. Welty, PhD
Assoc. Professor
BCC Director

Joan S. Chmiel, PhD
Professor

Jody D. Ciolino, PhD
Asst. Professor

Kwang-Youn A. Kim, PhD
Asst. Professor

Masha Kocherginsky, PhD
Assoc. Professor

Mary J. Kwasny, ScD
Assoc. Professor

Julia Lee, PhD, MPH
Assoc. Professor

Alfred W. Rademaker, PhD
Professor

Hannah L. Palac, MS
Senior Stat. Analyst

Gerald W. Rouleau, MS
Stat. Analyst

Amy Yang, MS
Senior Stat. Analyst

Not Pictured:
1. David A. Aaby, MS
Senior Stat. Analyst

2. Tameka L. Brannon
Financial | Research
Administrator

Biostatistics Collaboration Center |680 N. Lake Shore Drive, Suite 1400 |Chicago, IL 60611

Northwestern Medicine®
Feinberg School of Medicine

BCC: Biostatistics Collaboration Center

What We Do

Our mission is to support FSM investigators in the conduct of high-quality, innovative health-related research by providing expertise in biostatistics, statistical programming, and data management.

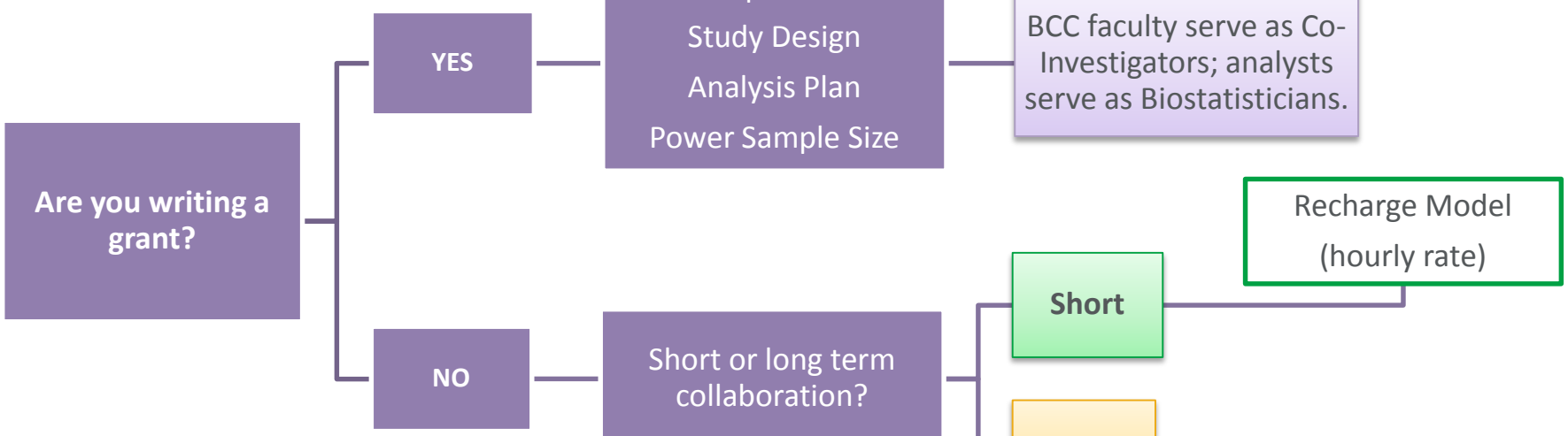
Northwestern Medicine®
Feinberg School of Medicine

BCC: Biostatistics Collaboration Center

How We Do It

The BCC recommends requesting grant support at least **6 -8 weeks** before submission deadline

Statistical support for **Cancer-related** projects or **Lurie Children's** should be triaged through their available resources.



Every investigator is provided a **FREE** initial consultation of up to 2 hours with BCC faculty of staff

BCC: Biostatistics Collaboration Center

How can you contact us?

- Request an Appointment
  - http://www.feinberg.northwestern.edu/sites/bcc/contact-us/request-form.html
- General Inquiries
  - bcc@northwestern.edu
  - 312.503.2288
- Visit Our Website
  - http://www.feinberg.northwestern.edu/sites/bcc/index.html

Biostatistics Collaboration Center |680 N. Lake Shore Drive, Suite 1400 |Chicago, IL 60611

Statistically Speaking …

What's next?

Friday, October 21	**Clinical Trials: Highlights from Design to Conduct** Masha Kocherginsky, PhD, Associate Professor, Division of Biostatistics, Department of Preventive Medicine
Tuesday, October 25	**Finding Signals in Big Data** Kwang-Youn A. Kim, PhD, Assistant Professor, Division of Biostatistics, Department of Preventive Medicine
Friday, October 28	**Enhancing Rigor and Transparency in Research: Adopting Tools that Support Reproducible Research** Leah J. Welty, PhD, BCC Director, Associate Professor, Division of Biostatistics, Department of Preventive Medicine

All lectures will be held from noon to 1 pm in Hughes Auditorium, Robert H. Lurie Medical Research Center, 303 E. Superior St.