Power and sample size calculations

Patrick Breheny

October 22

Planning a study

- One of the most important questions as far as planning and budgeting a study is concerned is: how many subjects do I need?
- The number of subjects tends to play a very large role in determining the cost of a study, so funding agencies generally want to know the number of subjects that a study will require before they make a decision about whether or not to pay for it
- But of course, the fewer subjects you have, the harder it is to distinguish a real phenomenon from chance

Power

- The probability that you will successfully distinguish the real phenomenon from chance (i.e., reject the null hypothesis) is captured in the notion of *power*
- Power is the probability of rejecting the null hypothesis given that it is in reality false
- Note that this is the complement of the type II error rate (β), which was defined the probability of failing to reject the null hypothesis given that it was false:

$$\text{Power} = 1 - \beta$$

Two important questions

There are two important, highly related questions here:

- If I plan a study with a certain number of subjects, what is my power going to be?
- If I want to achieve a certain power, how large does my sample size need to be?

Why is power important?

- Over the past 40 years, power calculations have played a very important role in advancing scientific rigor, particularly in medical studies
- Decades ago, it was common to carry out studies with small sample sizes and obtain non-significant results
- This is problematic for two important reasons:
  - People are very bad at interpreting non-significant results
  - There is often a publication bias against non-significant results, skewing the results that appear in print
- The goal of power calculations is to plan a study with a sufficiently large sample size to avoid these problems

The null and true distributions

- In order to calculate power, we will need to keep track of two sampling distributions for our statistic of interest:
  - Its distribution under the null hypothesis
  - Its actual distribution (note that this is different from the null distribution; otherwise we'd be calculating a Type I error rate)
- The power of any test depends on these distributions, which in turn depend on a number of factors, such as the size of the effect (the signal) and on the variability/standard error (the noise)
- In general, these are unknown parameters and we typically know very little about them – especially before we have conducted the study

Specifying unknown parameters

- So, in order to actually calculate power, we must make educated guesses about realistic/biologically plausible values for these quantities, and our calculated power will depend on the values that we choose
- Of course, if we specify values that are far away from reality, our power calculations are not going to be accurate
- Sometimes, reasonable values for certain quantities can be chosen on the basis of past studies or observations
- Other times, a small initial study called a "pilot study" is conducted in order to provide some data with which to estimate these quantities and help plan for a larger study that would take place in the future.

Outline

Depending on the complexity of the problem, there are three basic approaches one can take in calculating power:

- In simple settings (such as for t-tests), exact solutions are possible
- In moderately complex settings, it is often possible to (conservatively) approximate the power by considering a related simpler problem for which an exact solution is available
- In more complex settings, simulations are required

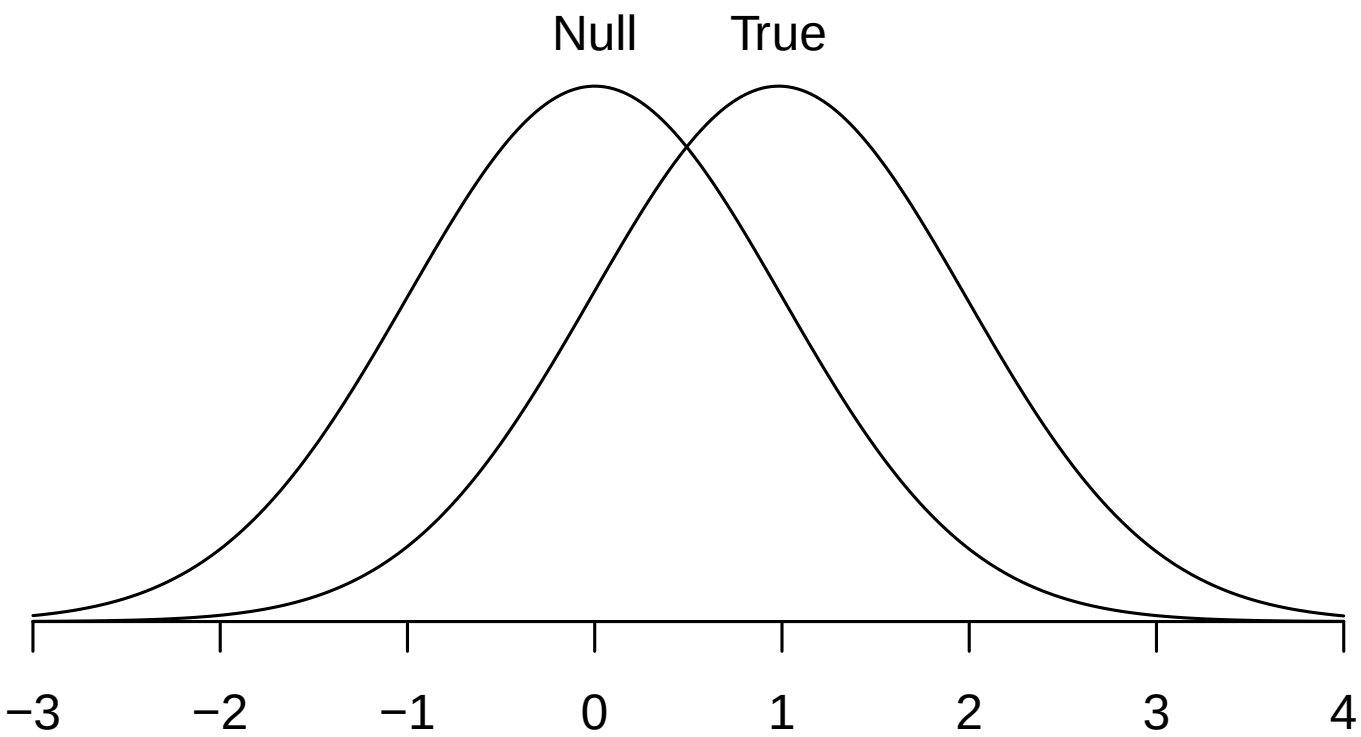
Hypothetical example

- Suppose we develop some intervention that may reduce LDL cholesterol levels
- Suppose we plan to conduct a study in which n individuals are randomized to receive this intervention and n individuals are randomized to receive a placebo, and we are going to look at the difference in mean LDL cholesterol levels at the end of the study between the two groups
- We propose to analyze this data using a two-sample t-test; to begin, however, let's assume that we know the true standard deviation σ

Example (cont'd)

- The power of our study depends on four factors, only two of which are under our control:
  - The sample size, n
  - The type I error rate we are willing to live with, α
  - How variable LDL cholesterol levels are in the population we are studying: σ, the variability
  - The amount by which our intervention actually reduces cholesterol: $\mu_1 - \mu_2$, the *effect size*
- For our calculations, we will assume a variability of $\sigma = 36$ mg/dL, an effect size of $\mu_1 - \mu_2 = 5$ mg/dL, a type I error rate of 5%, and start out with $n = 100$ in each group

Picture of sampling distributions



$$(\bar{x}_1 - \bar{x}_2)/\sigma\sqrt{1/n_1 + 1/n_2}$$

Critical values

- The key to calculating the power of a test is to determine its *critical value* – the value that the test statistic will have to lie outside in order to reject the test
- With this in mind, a power calculation is essentially a two-step procedure:
  - First, calculate the critical value based on the null distribution
  - Second, calculate the probability of lying outside the critical value based on the hypothesized true distribution

Calculating power for our cholesterol study

- In our hypothetical cholesterol study, the critical value (on the standardized scale) is 1.96
- On that scale, our true distribution has mean

$$\frac{\mu_1 - \mu_2}{\sigma\sqrt{1/n + 1/n}} = \frac{5}{36\sqrt{2/100}} = 0.982$$

- Thus, our power is

$$1 - \Phi(1.96 - 0.982) = 0.164;$$

with $n = 100$, our study is likely to be non-significant and lead to the problems mentioned at the outset of this lecture

Sample size calculation

- How large must n be in order to increase our power to, say, 80%?
- Let's let $z_{1-\alpha/2}$ and $z_{1-\beta}$ denote the $1-\alpha/2$ and $1-\beta$ quantiles of the standard normal distribution; we can easily solve the equation on the previous slide for n to obtain

$$n = 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{(\mu_1 - \mu_2)/\sigma}\right)^2$$

- Thus, in our example, we need

$$n = 2\left(\frac{1.96 + 0.84}{5/36}\right)^2 = 813.8$$

people in each group

Accounting for uncertainty concerning σ

- The preceding calculations were somewhat unrealistic in that we assumed we knew σ
- We can carry out more accurate calculations by basing everything on a t distribution instead of a normal distribution – conceptually this is a simple change, but the calculations get quite a bit messier
- For $n = 100$ in each group, the critical value is $t = 1.984$, but we must base power now off of the *noncentral t-distribution*:

$$1 - F(1.984|198, 0.982) = 0.163,$$

  where $F(x|\nu, \mu)$ is the CDF of the noncentral t distribution with noncentrality parameter μ and ν degrees of freedom
- As we would expect, this is slightly lower than the z-calculations, although the difference is negligible since n is reasonably large

Sample size calculation: t

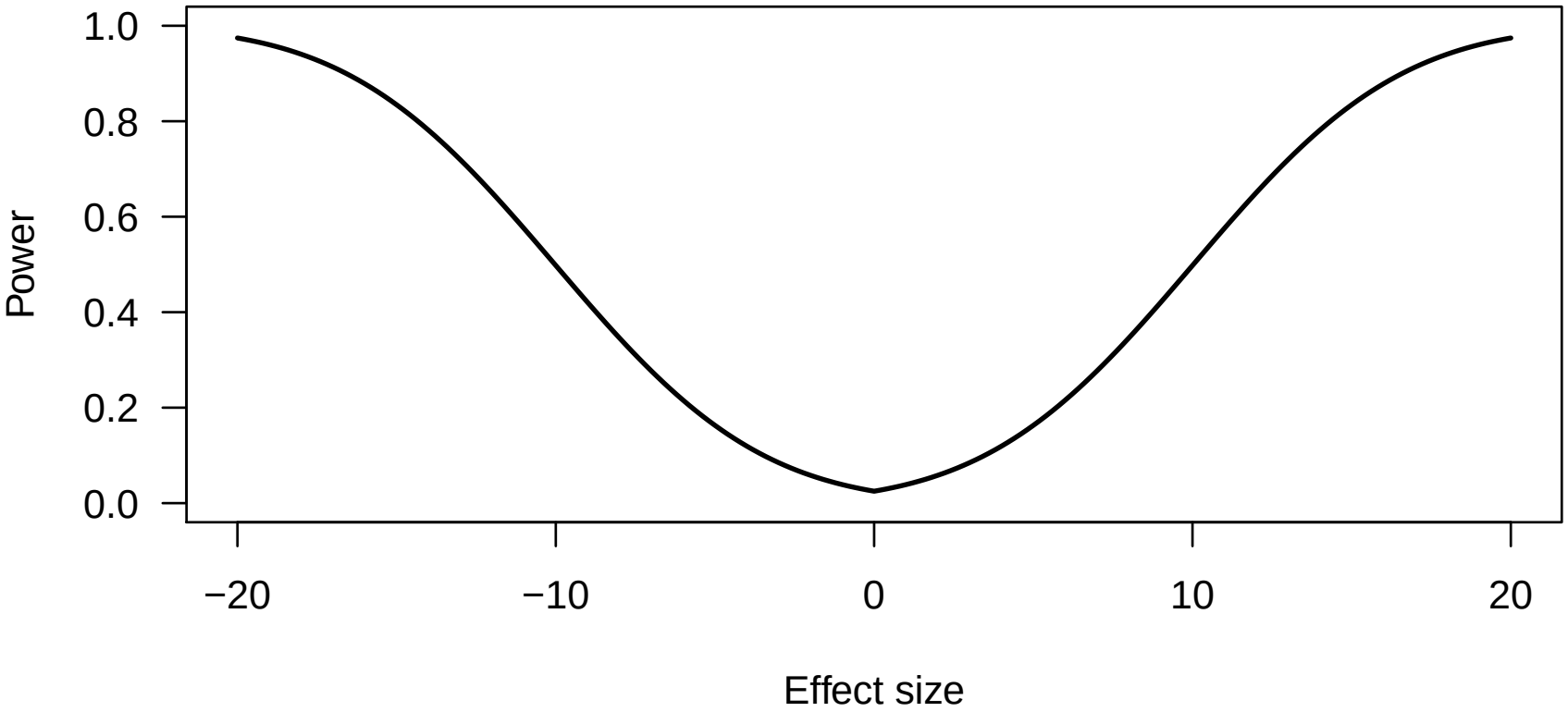
- Also, the complicated dependence of t quantiles on n prevents us from using a simple formula as we had in the z case
- Instead, we need to find the value n that satisfies

$$t_{2n-2,1-\alpha/2} = t_{2n-2,\beta,\mu},$$

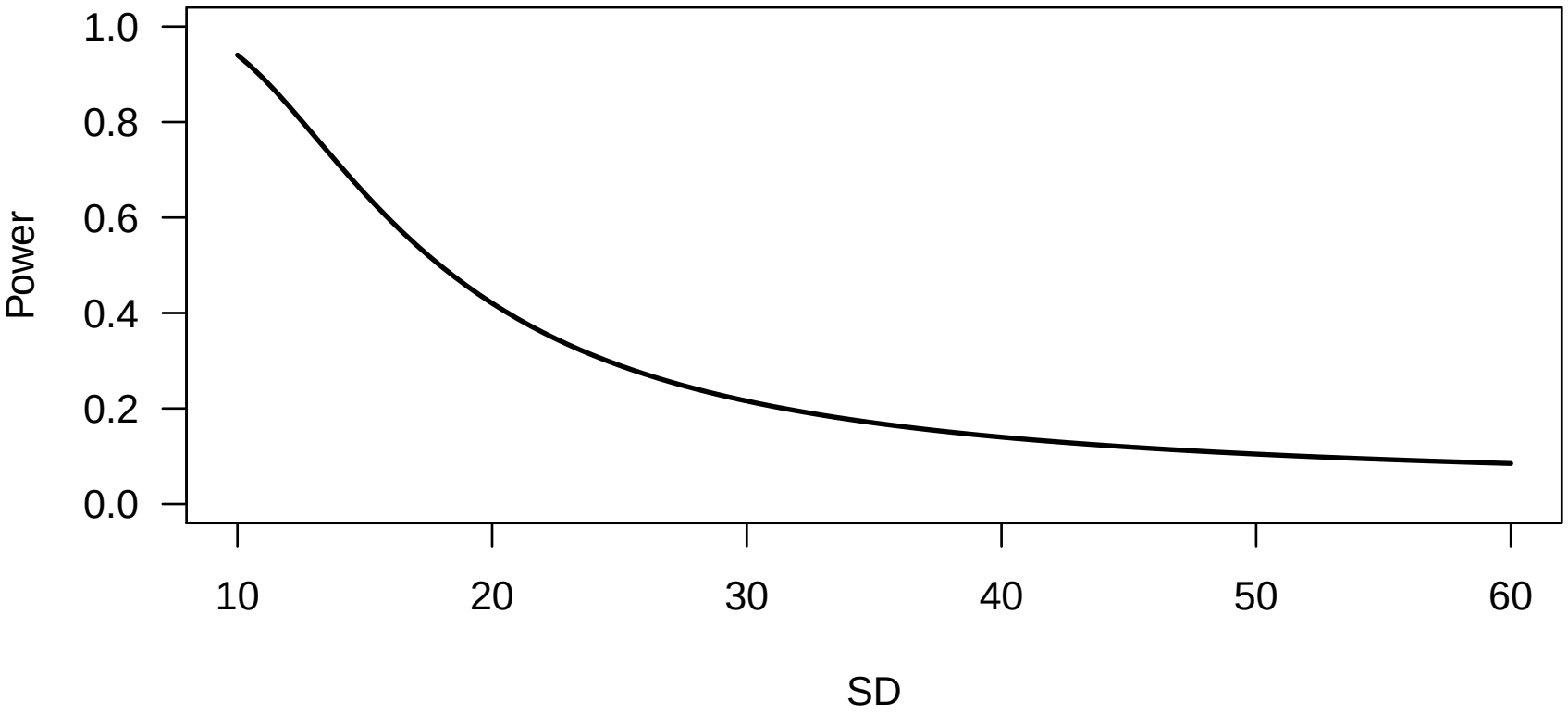
  where $t_{\nu,p,\mu}$ is the pth quantile of the noncentral t distribution with noncentrality parameter μ and ν degrees of freedom
- This, in turn, is a root-finding problem which requires a computer to solve; using one, we find that we need $n = 814.7$ in each group

Power curve: $n = 100$, $\sigma = 36$

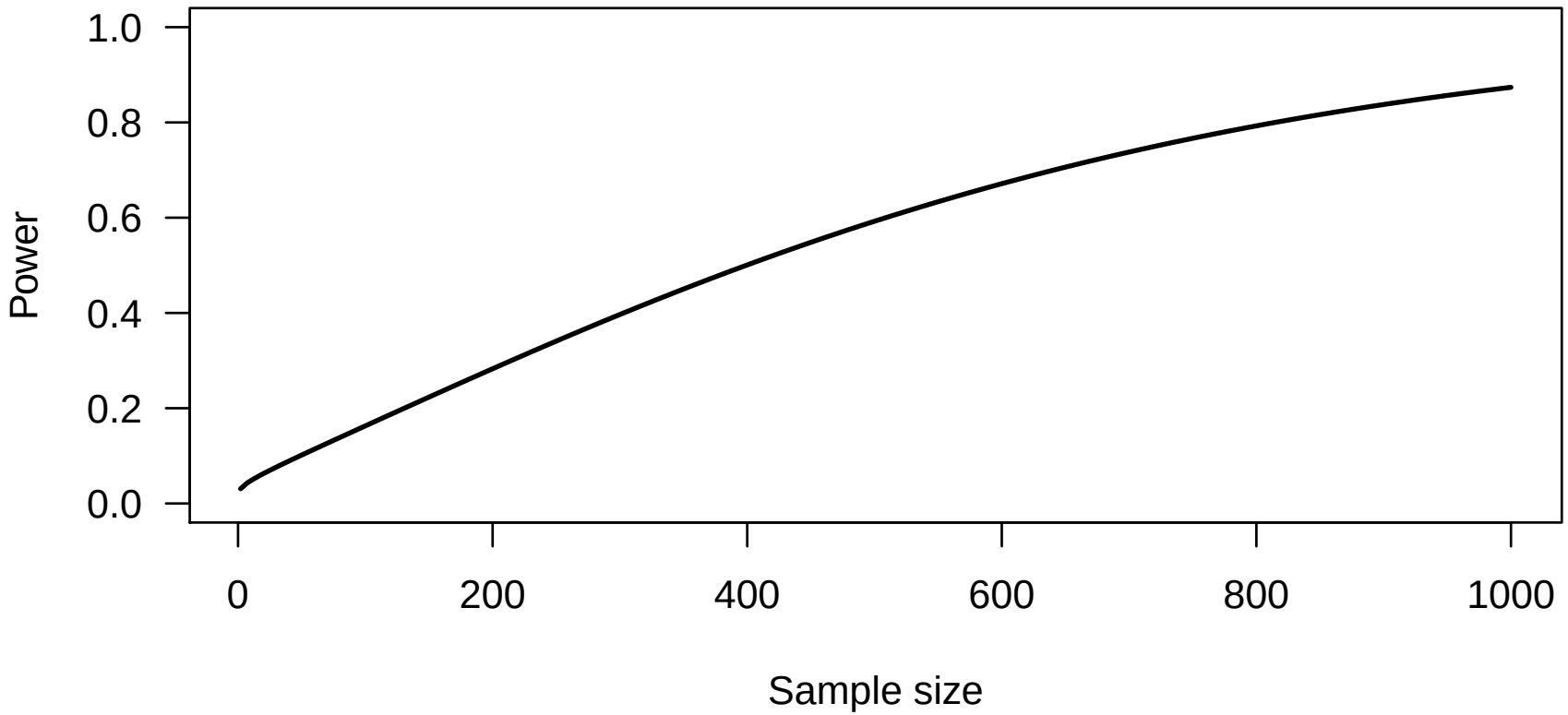
It is often instructive to look at power curves, which plot power as a function of various parameters



Power curve: $n = 100$, $\mu_1 - \mu_2 = 5$



Power curve: $\mu_1 - \mu_2 = 5$, $\sigma = 36$



A more complicated version of the cholesterol study

- In practice, our study might be more complicated than what we have described for the cholesterol example
- For example, rather than simply measuring cholesterol at the end of the study, the researchers might measure cholesterol repeatedly over the course of the study and analyze the trajectories of patients over their time on the study
- This would be a considerably more complicated analysis and difficult to carry out an exact power calculation

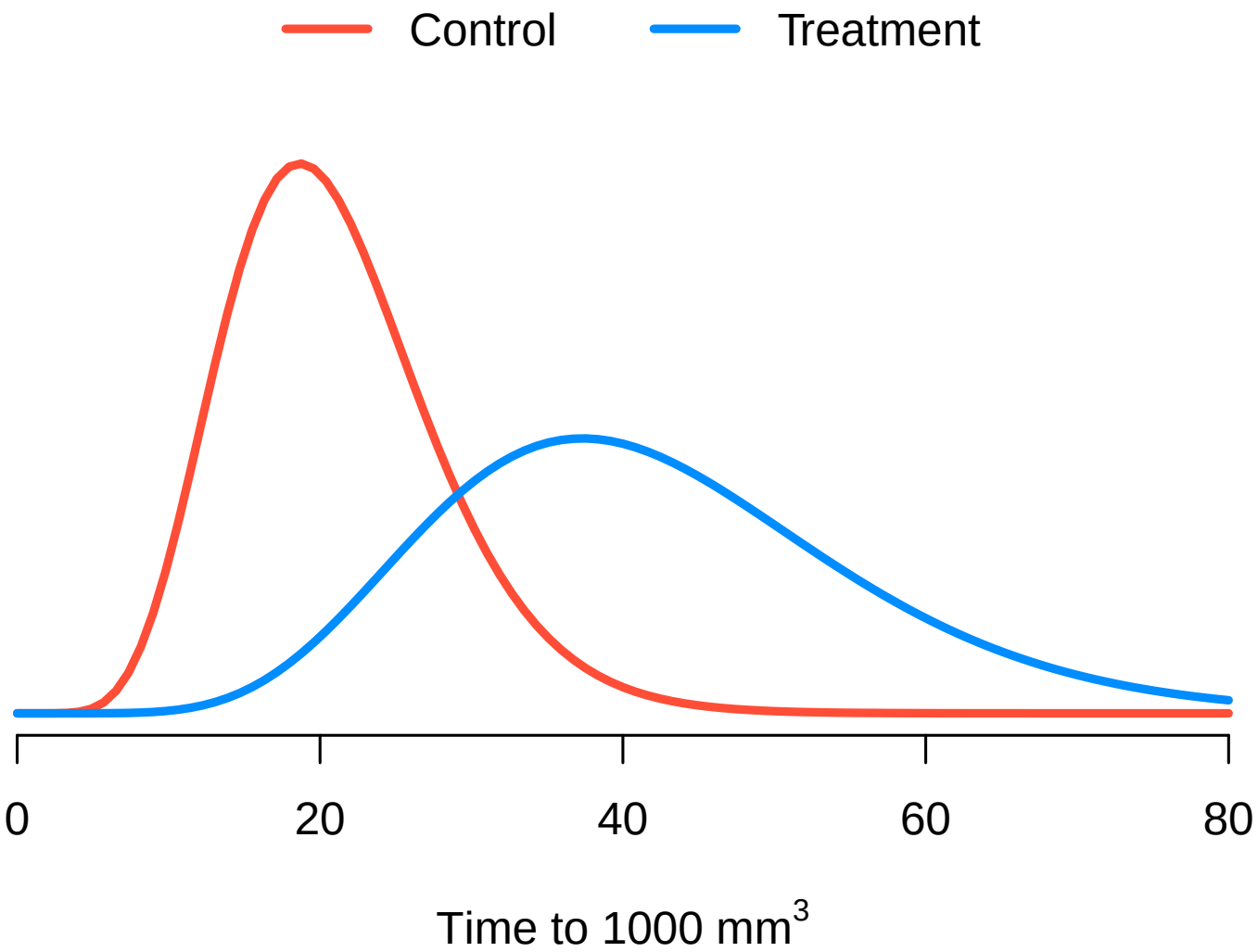
Using simpler studies to obtain a bound

- However, one can approximate (or perhaps more accurately, bound) the power of this more complicated study using the simpler study
- A more sophisticated analysis that takes into account the trajectories and repeated measurements on subjects should be more powerful than the simple endpoint analysis; therefore, if $n = 815$ yielded 80% power for the simple analysis, it should yield $> 80\%$ for a more powerful analysis
- Obviously, if the complex analysis is dramatically more powerful than the simple analysis, this approach will be rather conservative

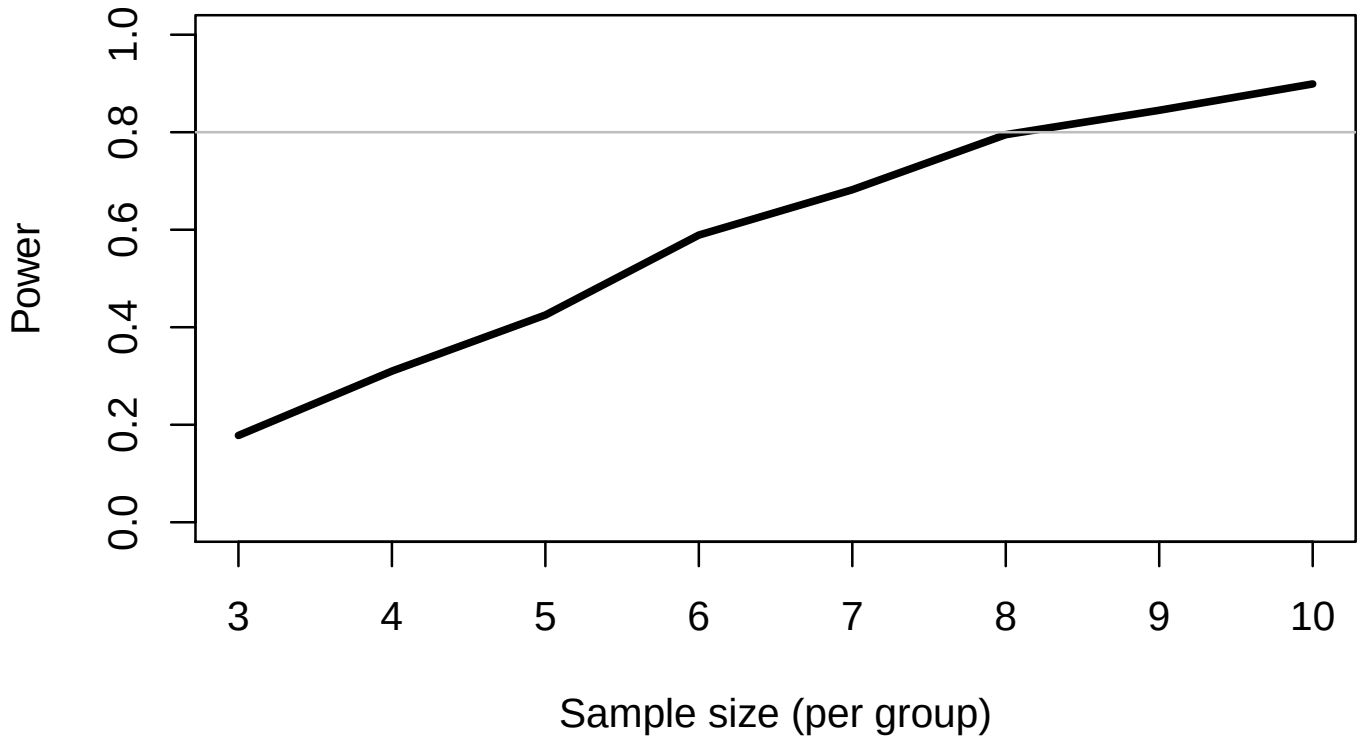
Simulating power: Growth curve example

- If one wants to avoid this conservatism, or is in a situation so complex that no simple version seems sufficient, one can use simulations to calculate the power
- For example, I recently performed a power calculation for a researcher at the UI cancer center who was investigating a treatment expected to slow (on average) the growth rate of a tumor by a factor of 2
- The proposal was to induce tumor growth in a sample of rats, then split them into a treatment and control group
- We would then analyze the effect of the treatment by measuring the growth of the tumor over time, fitting growth curves for each rat, and testing whether the average growth rate in the two groups was equal

Assumed distributions



Simulated power



We then proposed $n = 8$ rats per group in the grant

Sample size determination in practice

- Statistical calculations are essential to ensure a meaningful study adequately powered to answer the question of interest
- In reality, of course, lots of other things like money, time, resources, availability of subjects, etc., also influence the actual sample size of a study
- Finally, we may be interested in calculating the required sample size under a few different designs to see which way is the easiest/cheapest to conduct the study

Summary

- The power of a study depends on:
  - Sample size
  - Variability
  - Effect size
- Power calculations are a two-step process involving both the null and assumed true distributions
- Know how to calculate power and sample size for one- and two-sample z tests (and know how to use a computer to calculate the power of a t-test)