Error bars; Power and sample size

Patrick Breheny

March 22

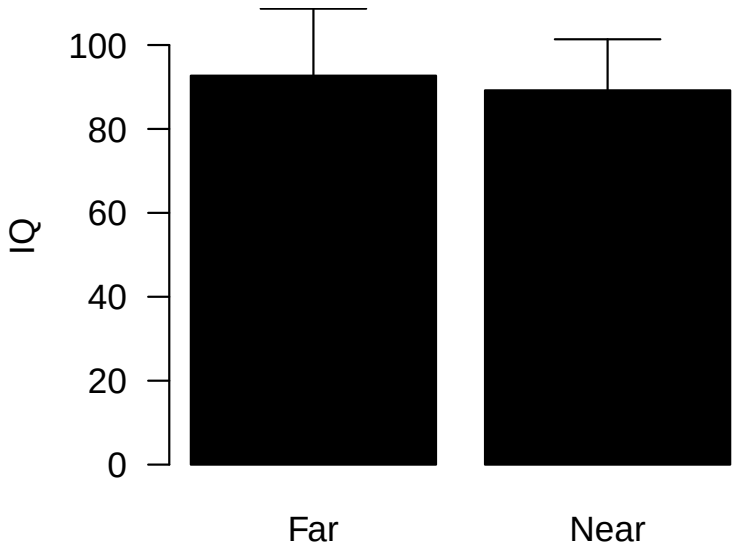
Introduction

In this lecture, we'll discuss two topics that come up in one-sample studies (as well as other types of studies):

- One is the presentation of results, and specifically, the decision when making figures and tables to include error bars and whether the error bar should be based on the standard deviation or the standard error
- The other is the issue of planning a study, and determining beforehand the power of the study for a given sample size

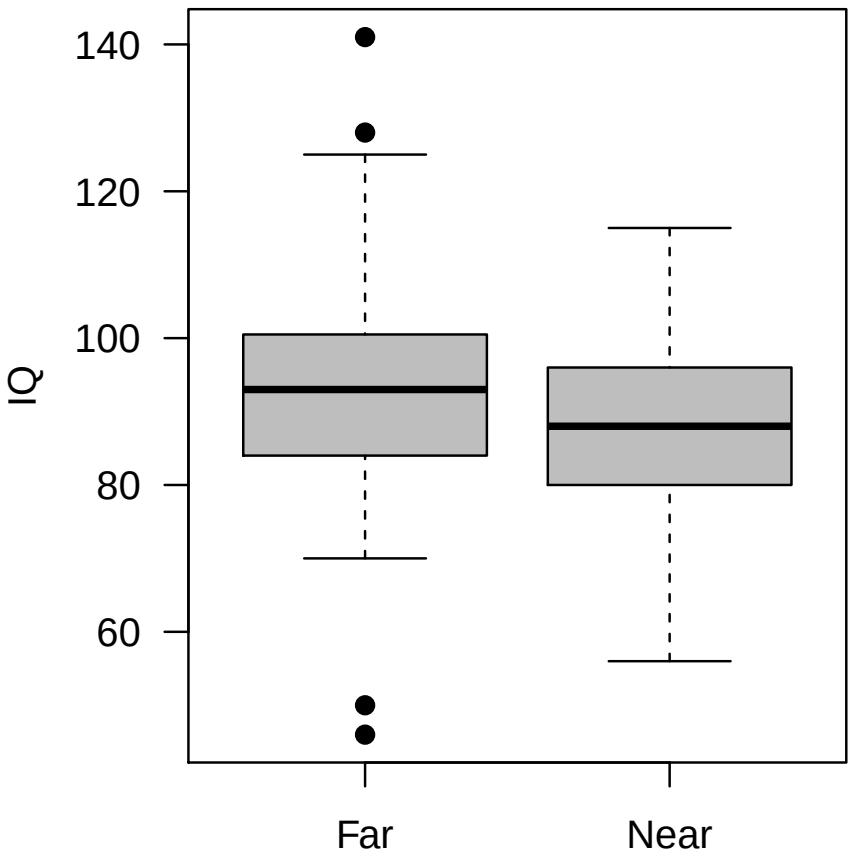
A dynamite plot

Figures like the following are incredibly common:



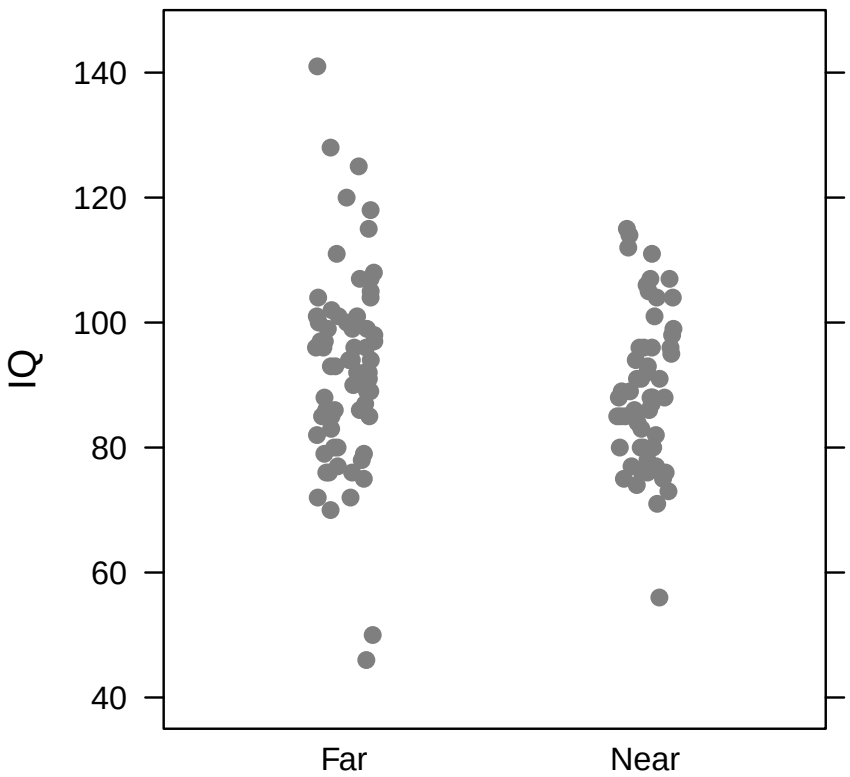
- There are a number of possible variations, but the general idea is to have a bar or a dot or a line that shows the mean, then *error bars* that show . . . something
- These particular error bars are one SD, so they illustrate something about variability of individuals' IQs in the two samples

Box plot



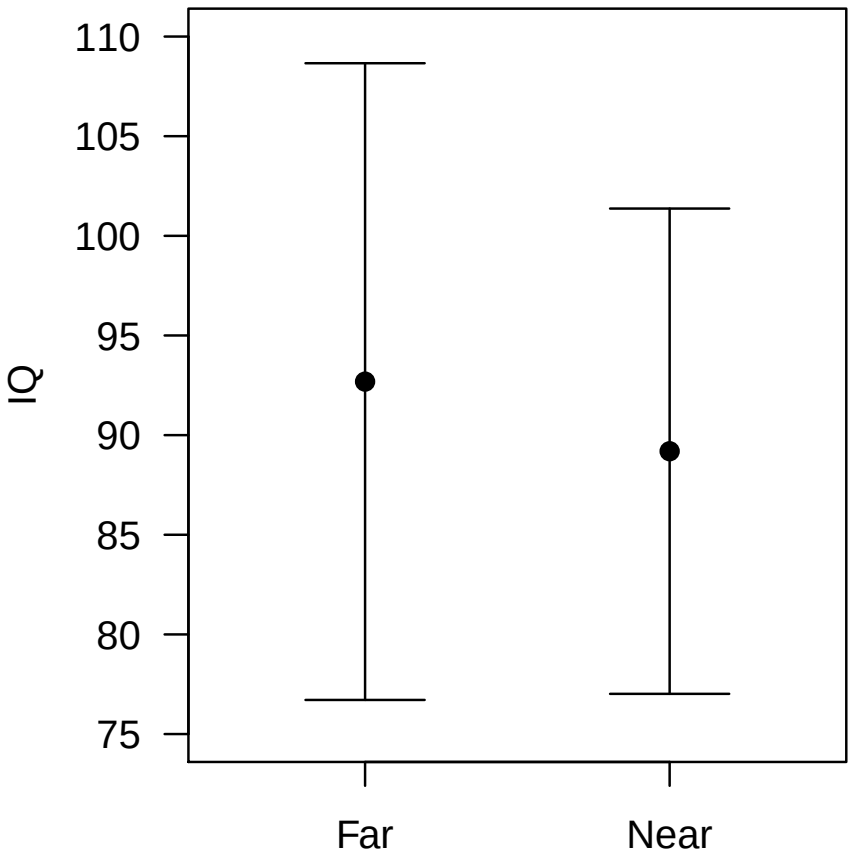
- Keep in mind, however, that the "dynamite" plot is just showing two numbers, the mean and SD
- This has the potential to misrepresent the actual distribution of data, and more informative sorts of plots, such as the box plot, are alternatives worth considering

Strip plot



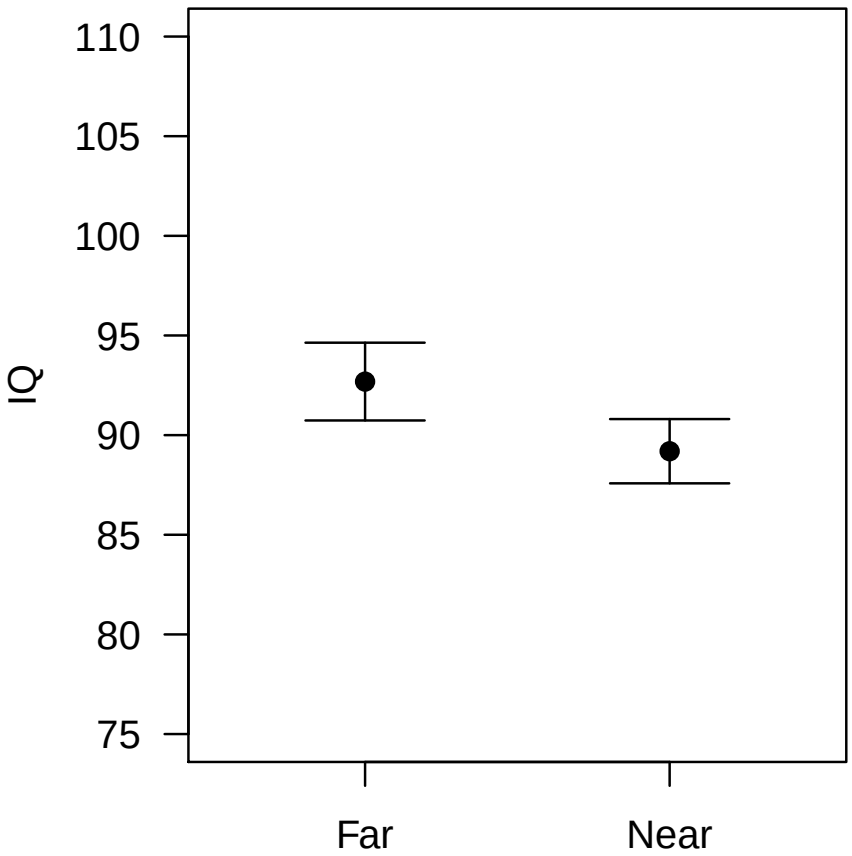
- We could also just show all the data
- This type of plot is known as a *strip plot*

$\pm 1\,\mathrm{SD}$ again



- This is the same as the first plot, although with dots replacing the bars; this has the advantage that the bar doesn't block the lower error bar, so we can see the $\pm 1\,\mathrm{SD}$ region more clearly
- Recall that the mean $\pm 1\,\mathrm{SD}$ typically contains about the middle two-thirds of the data

$\pm 1\,\mathrm{SE}$

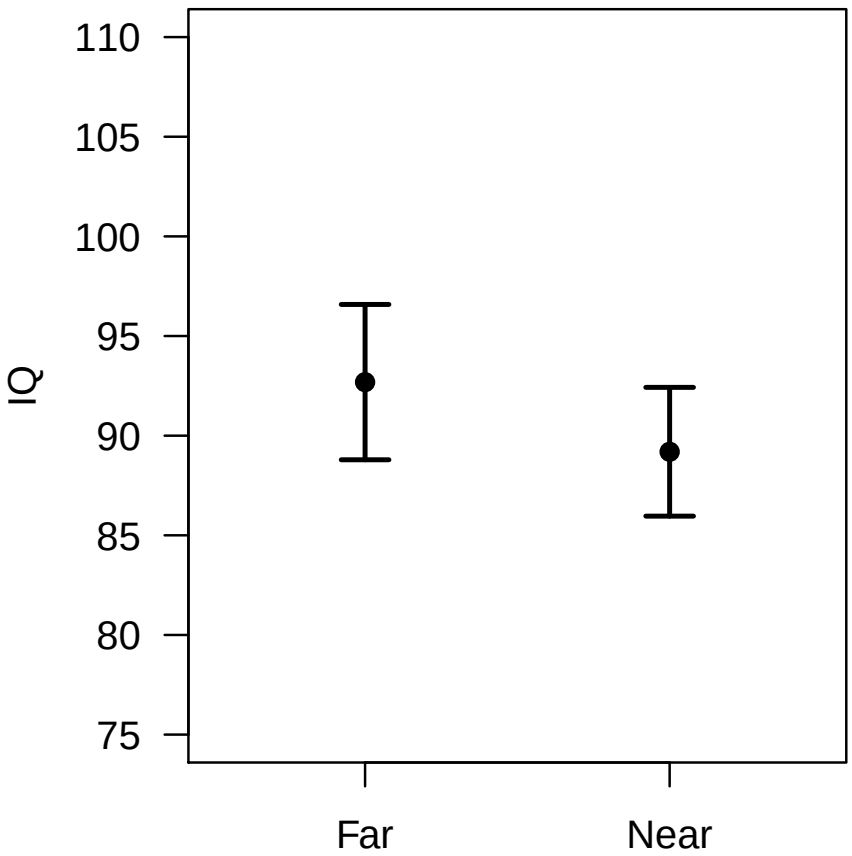


- Here, we're plotting $\pm 1\,\mathrm{SE}$ instead of $\pm 1\,\mathrm{SD}$
- The error bars are a lot smaller here, of course, since $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$, and now say something about the variability of the sample mean or, if you prefer, the error with which we have measured the sample mean

Error bars: SD vs. SE

- So, the $\pm\mathrm{SD}$ error bar plot emphasized the *variability* of data
- This doesn't really have anything to do with any sort of "error" at all, and furthermore, other kinds of plots are usually better for showing the distribution of your data
- On the other hand, the $\pm\mathrm{SE}$ error bar plot emphasizes uncertainty (or possible error) about the mean
- However, this isn't a particularly meaningful plot either, since $\pm\mathrm{SE}$ is only a 68% CI, and really, it isn't even that, since as we learned in the previous lecture, the coverage of the CI depends on the degrees of freedom with which we've measured the SD

CI plot

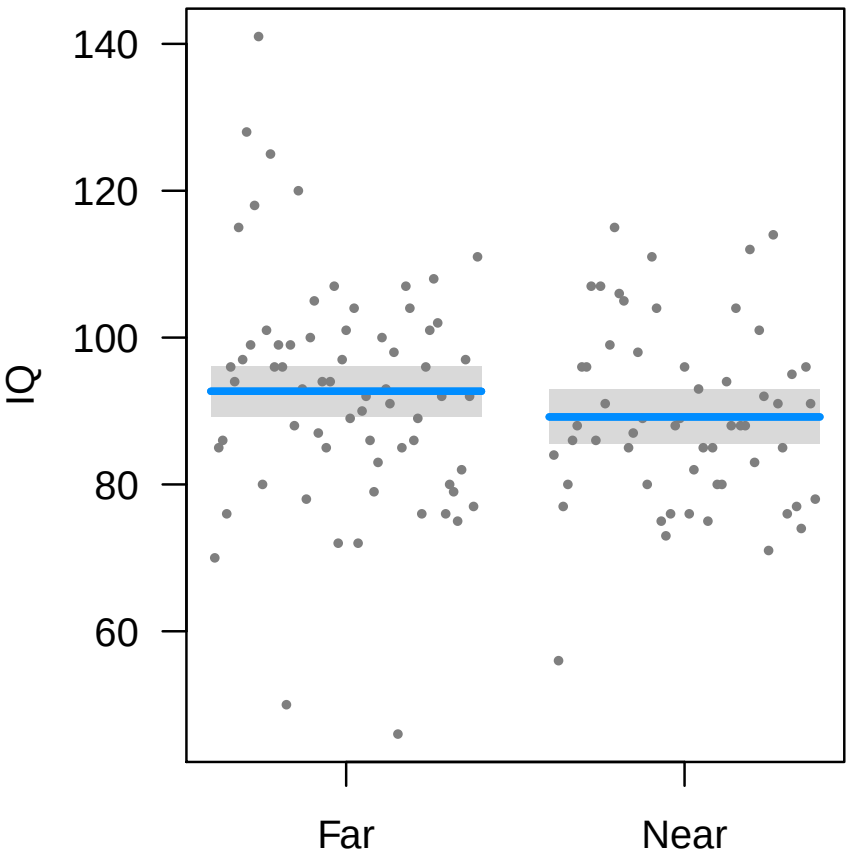


- So typically, the most meaningful route is to plot the confidence interval
- For example, the plot on the left suggests that it isn't unreasonable to think that the average IQ for both groups is 90

Variability and uncertainty are both of interest

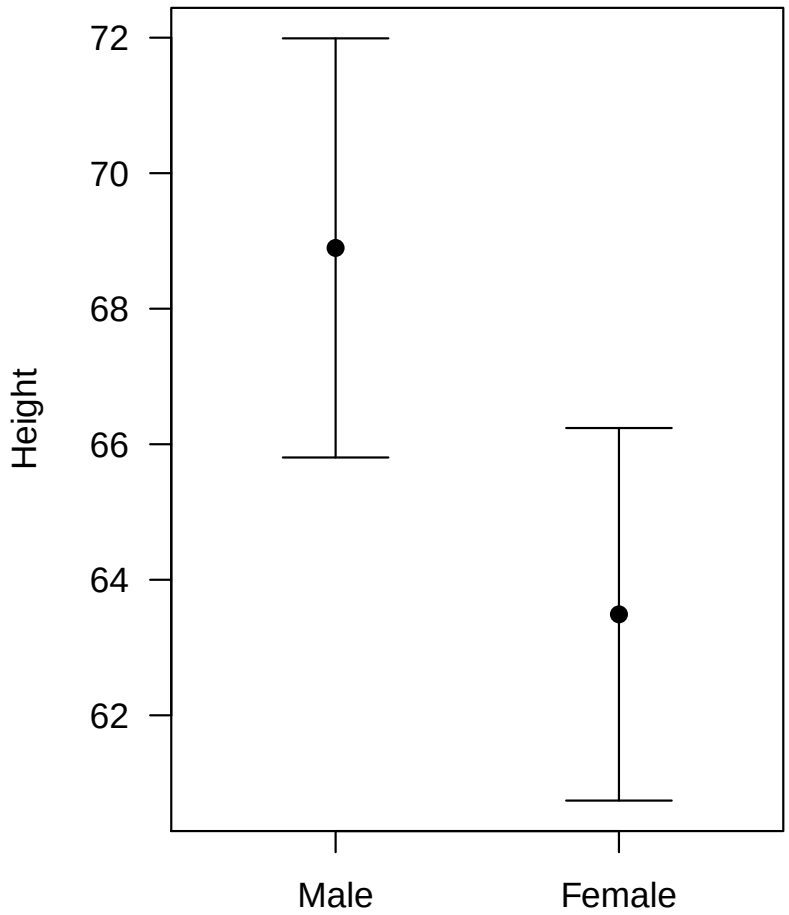
- It should be noted, however, that the confidence interval plot, although very useful and informative, doesn't really tell us anything about the distribution of the data, which is also interesting
- In principle, this is the choice one makes in deciding on a figure, whether to emphasize uncertainty about the average or to emphasize variability/diversity, although there are ways to illustrate both in a single figure

Illustrating both variability and uncertainty



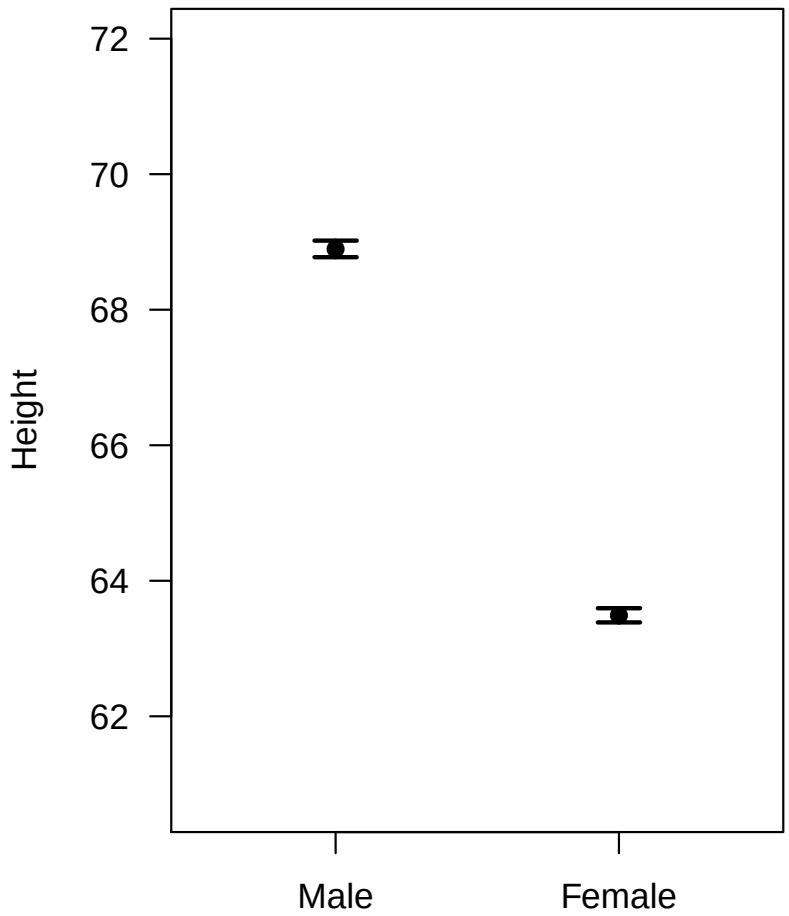
In this plot, the shaded region denotes the 95% confidence interval, and we can get a sense of the variability among individuals as well as the uncertainty about the population means

$\pm 1\,\mathrm{SD}$ for NHANES height



Just to make sure the point is clear, let's look at the NHANES height data; here are the means $\pm\mathrm{SD}$

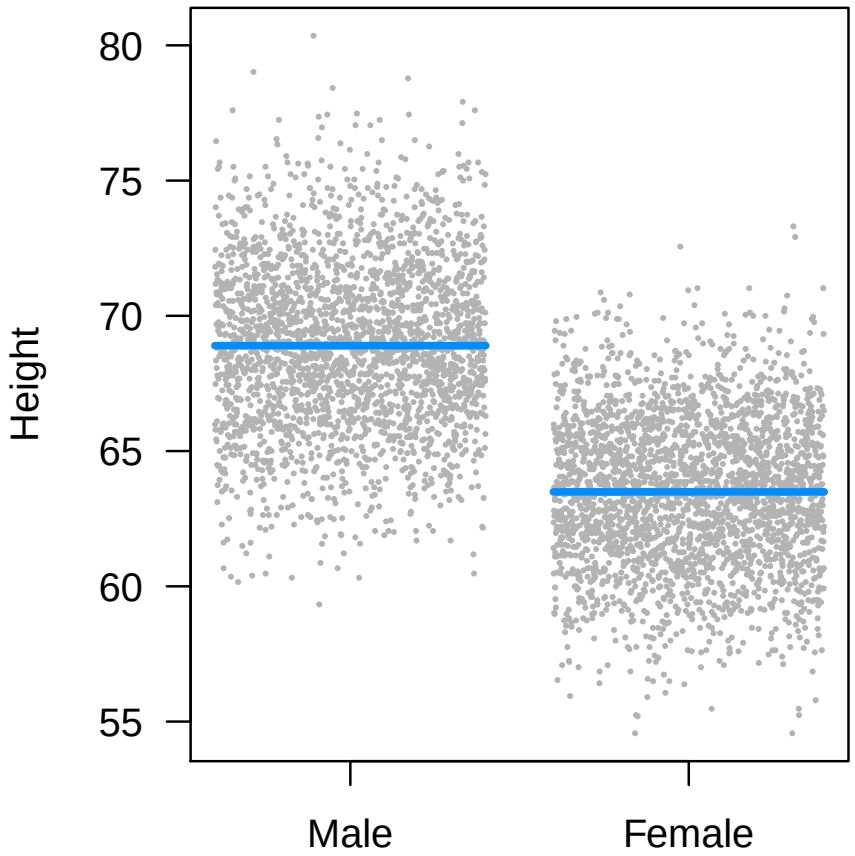
CI plot for NHANES height



And here the error bars denote the 95% confidence interval; both this plot and the previous one tell us different things

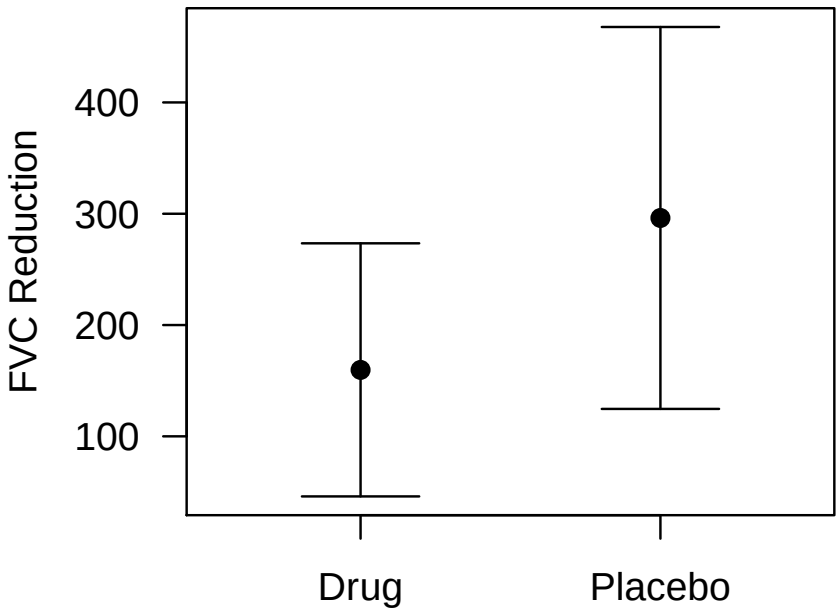
- It is obvious that men are, on average, taller than women
- At the same time, it is just as obvious that there is plenty of overlap in the distributions – i.e., that it is quite common for a woman to be taller than a man

NHANES height data

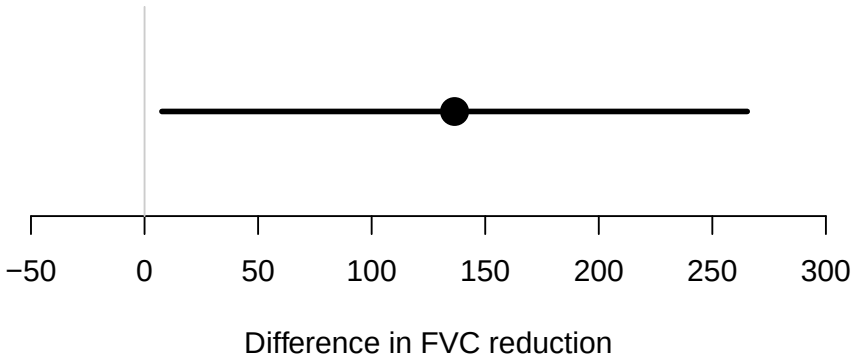


Cystic Fibrosis

One last comment: plotting separate confidence intervals and looking for overlap is not the same as plotting a confidence interval for the difference



The disparity is particularly drastic for paired studies, but also occurs for two-sample studies, as we will see in a week or two



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 opposite of the type II error we discussed back at the beginning of the semester
- Power is the probability of rejecting the null hypothesis given that it is in reality false; the type II error rate (β) was the probability of failing to reject the null hypothesis given that it was false
- Thus, by the compliment rule:

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

Two important questions

- With the time remaining in today's lecture, I want to address two highly related questions:
  - 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?

Two important questions (cont'd)

- These are important questions for any kind of data, and each type of study has its own formulas and procedures for calculating power
- We won't get into the details of power calculations in this class (which can be quite complicated)
- Instead, we will focus on the main concepts, which are generally similar for any type of study

Power: Basement analogy

- Consider the following analogy[1]: you send a child into the basement to find an object
- What's the probability that she actually finds it?
- This depends on three things:
  - How long does she spend looking?
  - How big is the object?
  - How messy is the basement?

[1] This analogy comes from *Intuitive Biostatistics*, which in turn credits John Hartung for the original idea

Power: Basement analogy (cont'd)

- If the child spends a long time looking for a large object in a clean, organized basement, then she'll probably find it
- If the child spends a short time looking for a small object in a messy basement, then there's a good chance she won't find it
- All three of these questions have statistical analogs:
  - How long does she spend looking? = How big was the sample size?
  - How big is the object? = How large is the effect size?
  - How messy is the basement? = How noisy/variable is the data?

Specifying effect size and variability

- In general, one does not know the effect size or the variability – especially before we have conducted the study
- So, in order to calculate power, we are going to essentially make up 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.

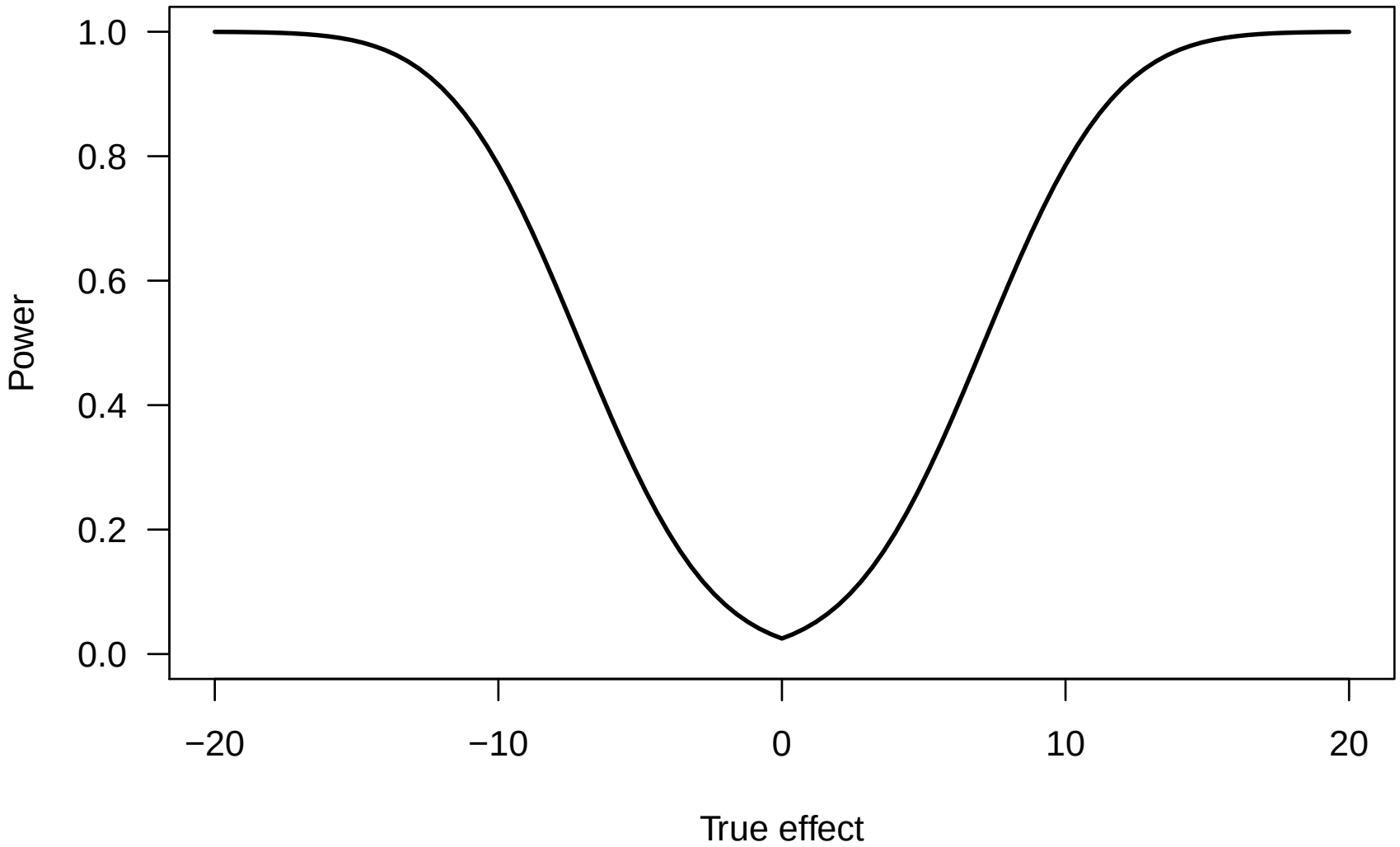
Example

- Suppose we develop some intervention that we think can reduce the LDL cholesterol levels of individuals who participate in it
- Suppose we plan to conduct a study in which individuals try both the intervention and a control, and we are going to look at the difference in each individual's LDL cholesterol levels on and off the intervention
- The power of our study (the probability that we will get a p-value under .05) depends on:
  - The sample size (how many people we enroll in our study)
  - The variability (how much variability there is in a person's LDL cholesterol levels)
  - The effect size (the amount by which our intervention actually reduces cholesterol)

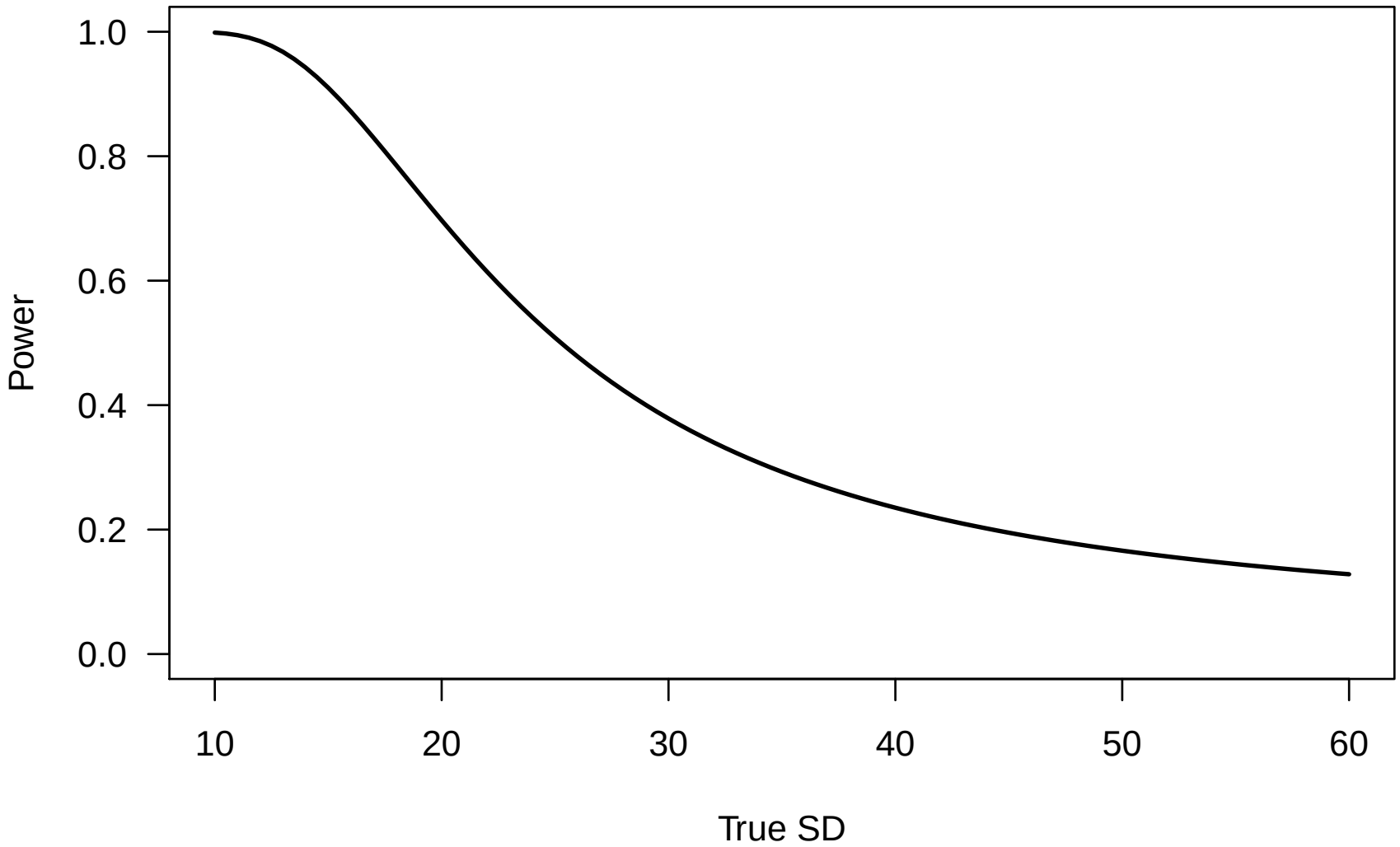
Power curves

- The main concepts behind power and sample size can be illustrated in *power curves* – graphs of what happens to power as we change one of sample size/variability/effect size
- In the curves that follow, I will start with:
  - A sample size of 100
  - A variability of 36 mg/dL
  - An effect size of 5 mg/dL
- And we will see what happens to power as we vary each of them

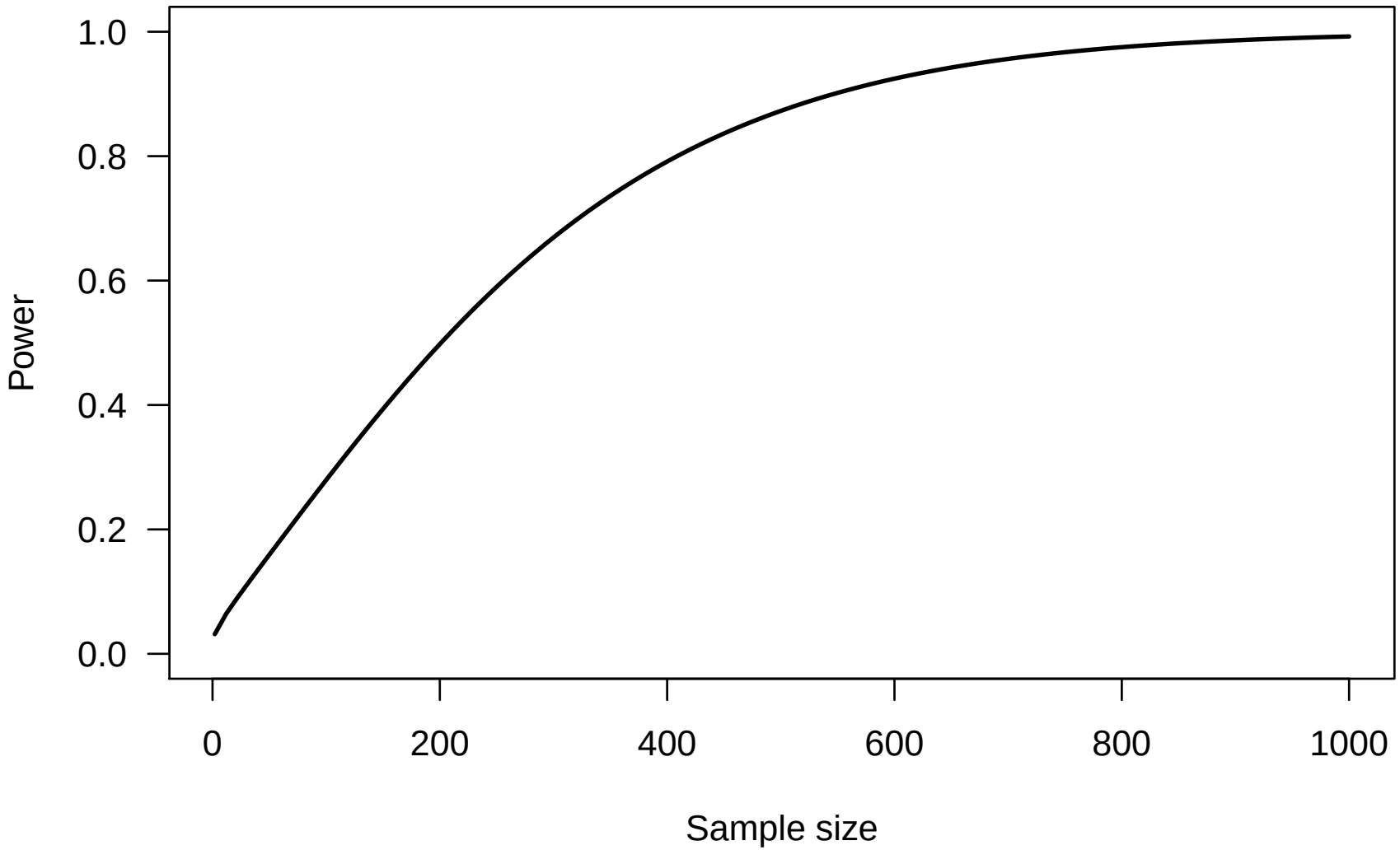
Power curve #1



Power curve #2



Power curve #3



Sample size determination

- Thus, we can determine our sample size by looking at the power curve
- For instance, in the previous example, if we want a power of 80%, we would need a sample size of about 400
- In reality, of course, lots of other things like money, time, resources, availability of subjects, etc., influence the actual sample size of a study
- Also, 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

- Displaying $\pm\text{SD}$ in a figure emphasizes variability, while displaying $\pm\text{SE}$ emphasizes uncertainty, although ask yourself: Wouldn't I be better off plotting the confidence interval?
- The power of a study depends on:
  - Sample size
  - Variability
  - Effect size