Visualizing uncertainty; Power and sample size

Patrick Breheny

March 26, 2026

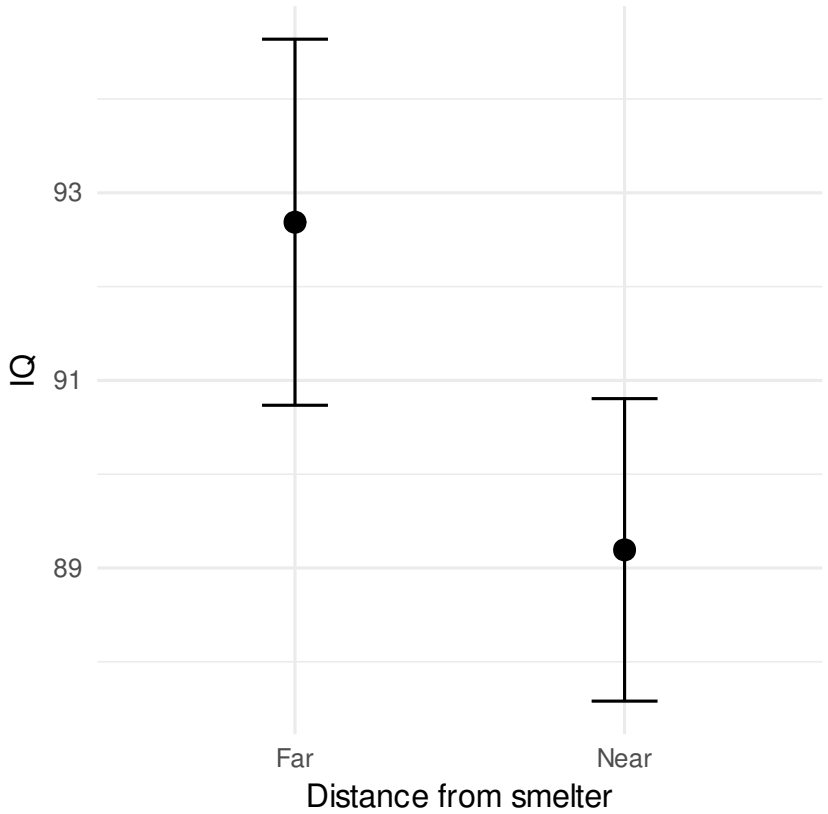
Introduction

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

- One is the presentation of results — specifically, revisiting our earlier lecture in which figures described the data; how to visualize results of statistical analyses?
- The other is the issue of planning a study, and determining beforehand the power of the study for a given sample size

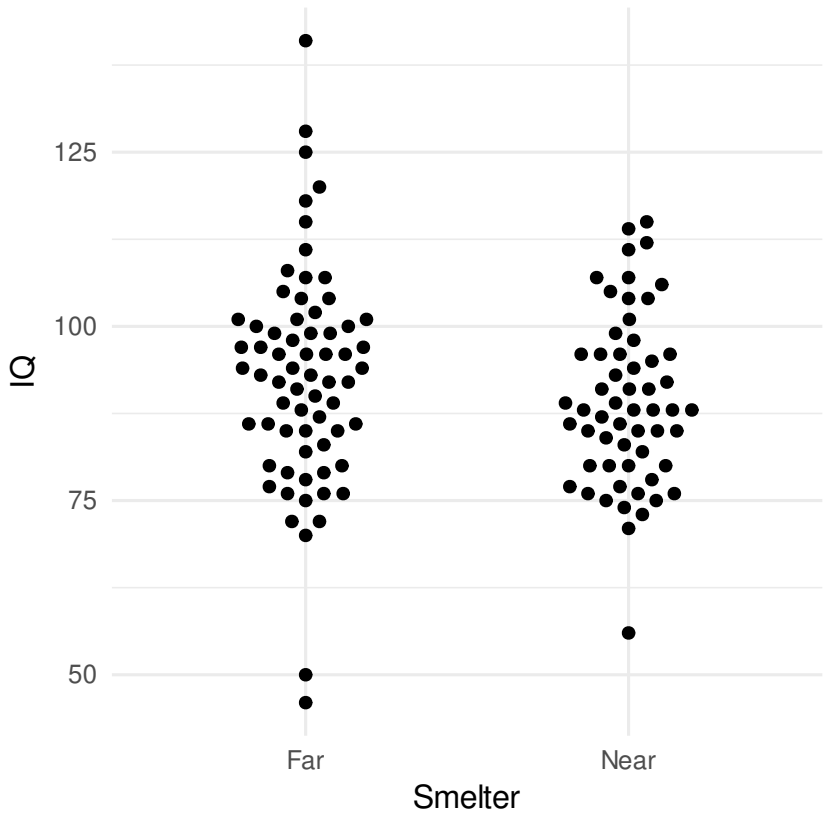
Figures and uncertainty

Consider the following figure:



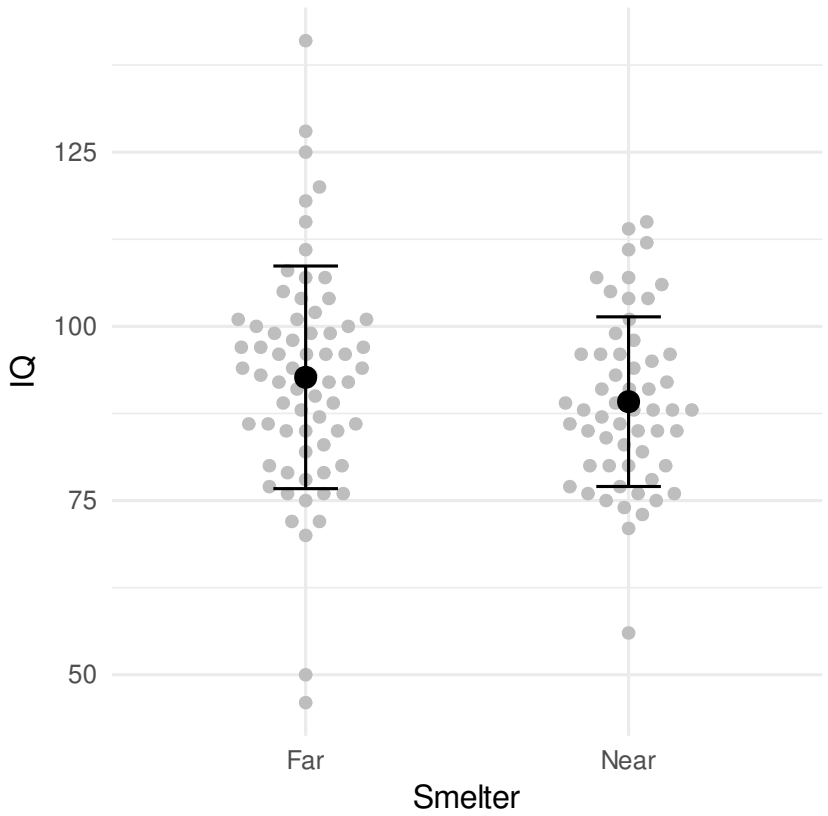
- Unlike our earlier figures, this does not show a distribution of data — the points and bars represent summaries from our analysis (here, the mean $\pm$ SE)
- There are many variations on this basic idea, however:
  - What should the error bars represent?
  - How to visually represent uncertainty? (error bars are one way, but not the only way)

Strip plot



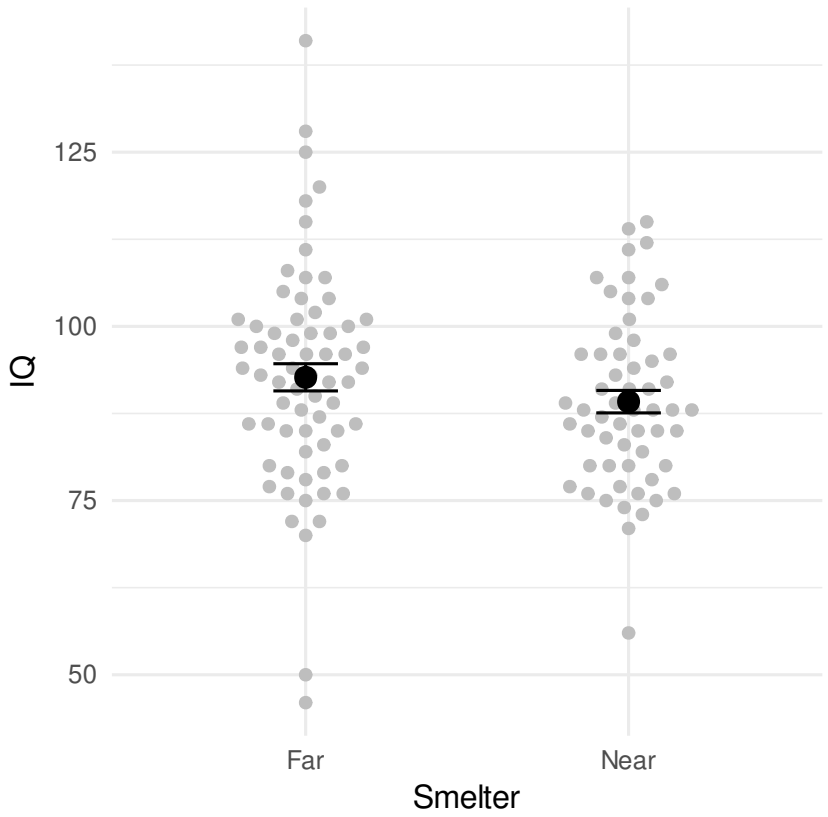
- Let's start by looking at the actual, raw data — no summarization whatsoever
- This type of plot is known as a *strip plot* or *beeswarm plot*

Mean ± SD



- As discussed in that earlier lecture, the mean ± SD is a way of summarizing the distribution of the data
- Keep in mind that the range here represents the *variability* of individuals, not our *uncertainty* about the mean — these are very different things

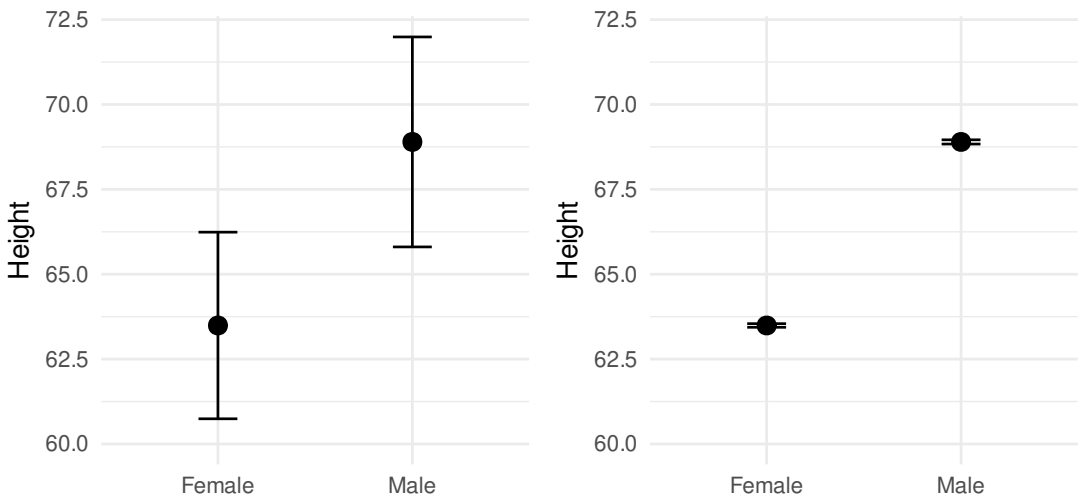
Mean ± SE



- 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 $\mathrm{SE} = \mathrm{SD}/\sqrt{n}$
- The reason is that they represent something completely different — our uncertainty about the *average*, not how individuals vary

Another example: NHANES height

To see a more extreme example, consider these plots of the NHANES height data

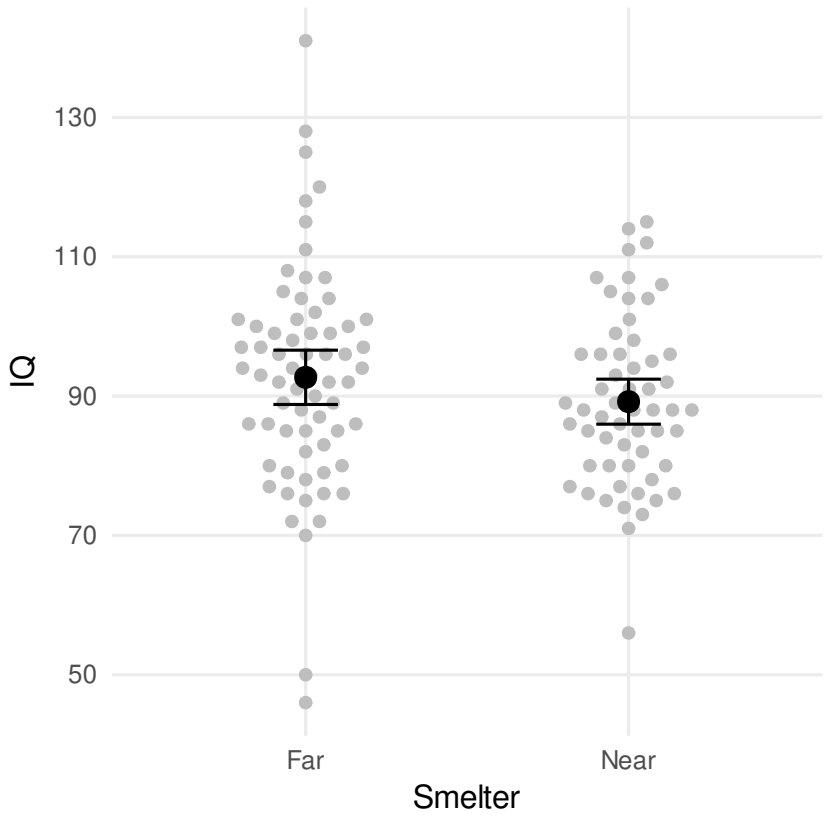


- On average, men are taller than women — there is no doubt about that
- It is just as obvious that the distributions of individuals overlap — it is very common for a woman to be taller than a man

Error bars: SD vs. SE

- Plotting the mean $\pm\text{SD}$ is frankly not very useful:
  - Other plots, like histograms and box plots, are better at representing the distribution of data
  - It's also a little misleading, as the "error" bars don't represent error
- Plotting the mean $\pm\text{SE}$ isn't really ideal either
  - As we learned in the previous lecture, our SE is an estimate, and the smaller n is, the less accurate that estimate is
  - If you're trying to describe uncertainty, the most meaningful route is usually to plot the confidence interval

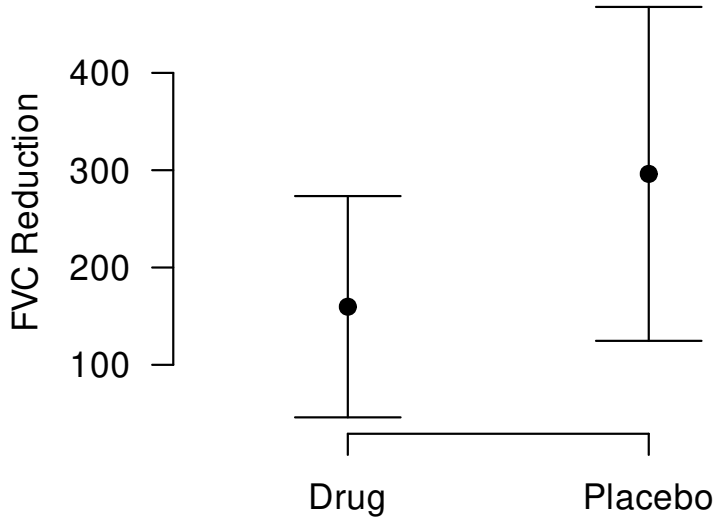
CI plot



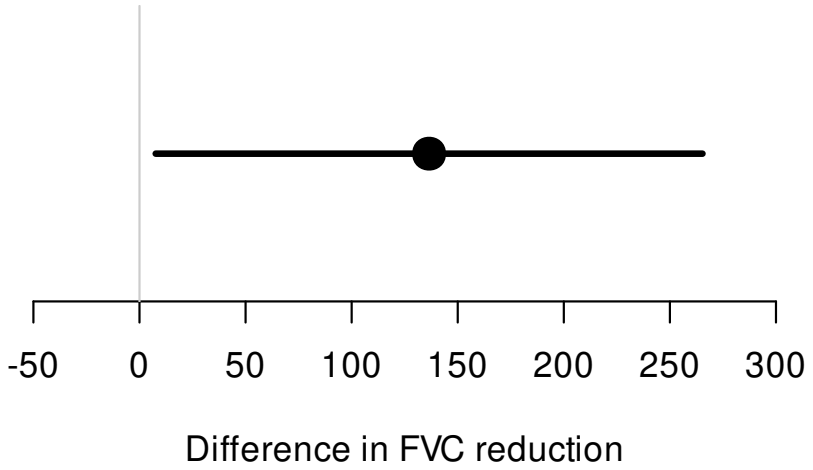
We've been emphasizing the choice between plotting variability or uncertainty, but note that is possible to represent both in the same figure, as we're doing here

Are overlapping CIs meaningful?

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



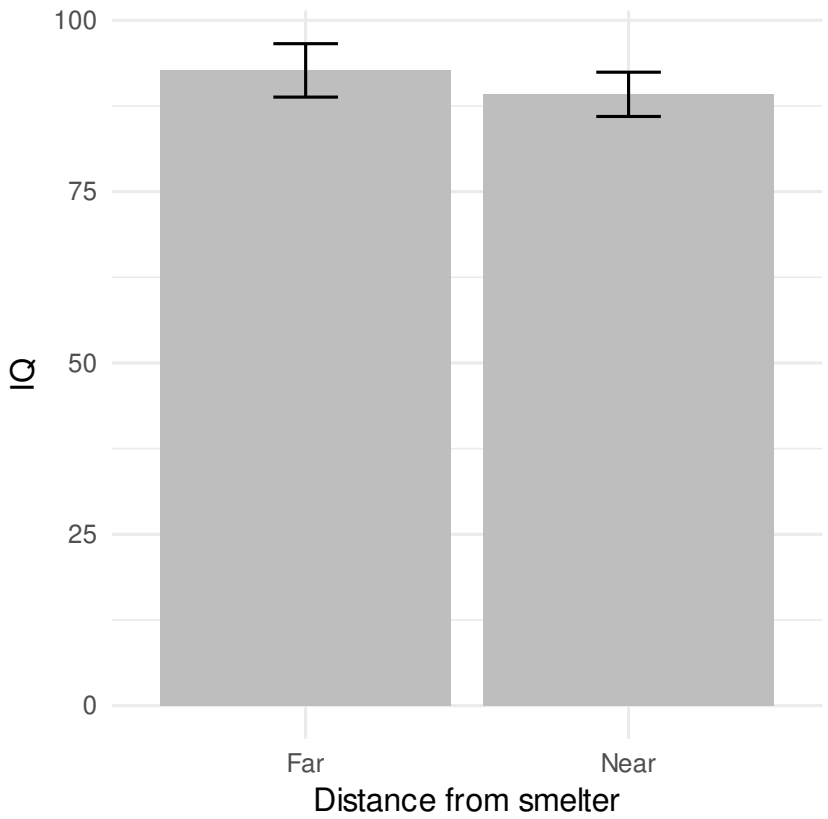
The disparity is particularly drastic for paired studies, but also occurs for two-sample studies, as we will see in a week or two



How to represent uncertainty

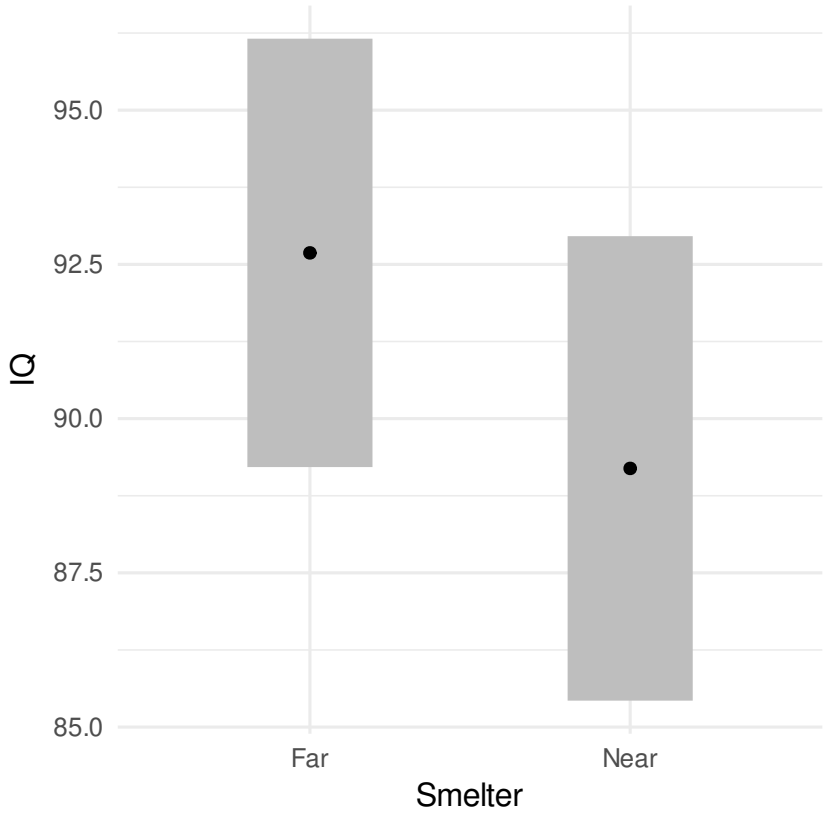
- Error bars are the most common way of plotting uncertainty, but they are not the only way, and not necessarily the best way either
- On the next several slides, I'm going to show a series of ways to plot confidence intervals, in order from worst to best (visualization researchers have conducted many studies of this, so it's not just my subjective opinion about which plot is the prettiest)

The dynamite plot



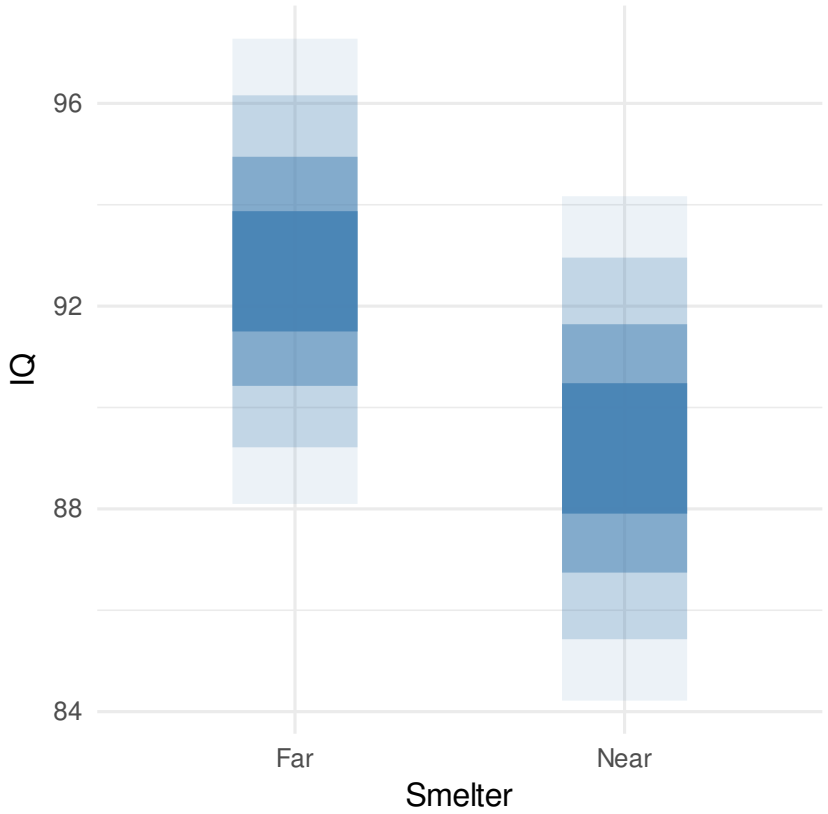
Very common in lab sciences, but low "ink-to-information" ratio; if you do go this route, make sure we can see the lower half of the error bar

Error band



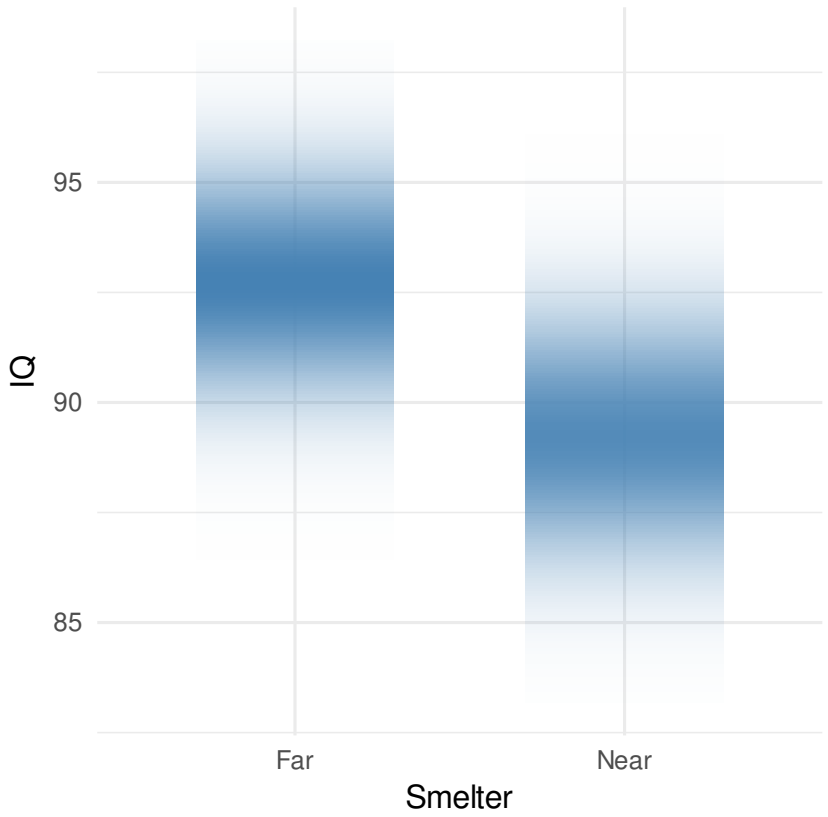
An error band is essentially the same as the error bar, except a shaded region is used to denote the confidence interval

Overlapping bands (fan)



This concept can be extended to show multiple overlapping confidence intervals: here, we see the 50%, 80%, 95%, and 99% intervals

Gradient chart (blur)



- Taking this idea out to its extreme, we can represent our uncertainty regarding the mean with a blur or gradient of colors
- Studies have indicated that people shown these gradient visualizations tend to make more accurate inferential conclusions, although they are not (yet?) widespread in science

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.

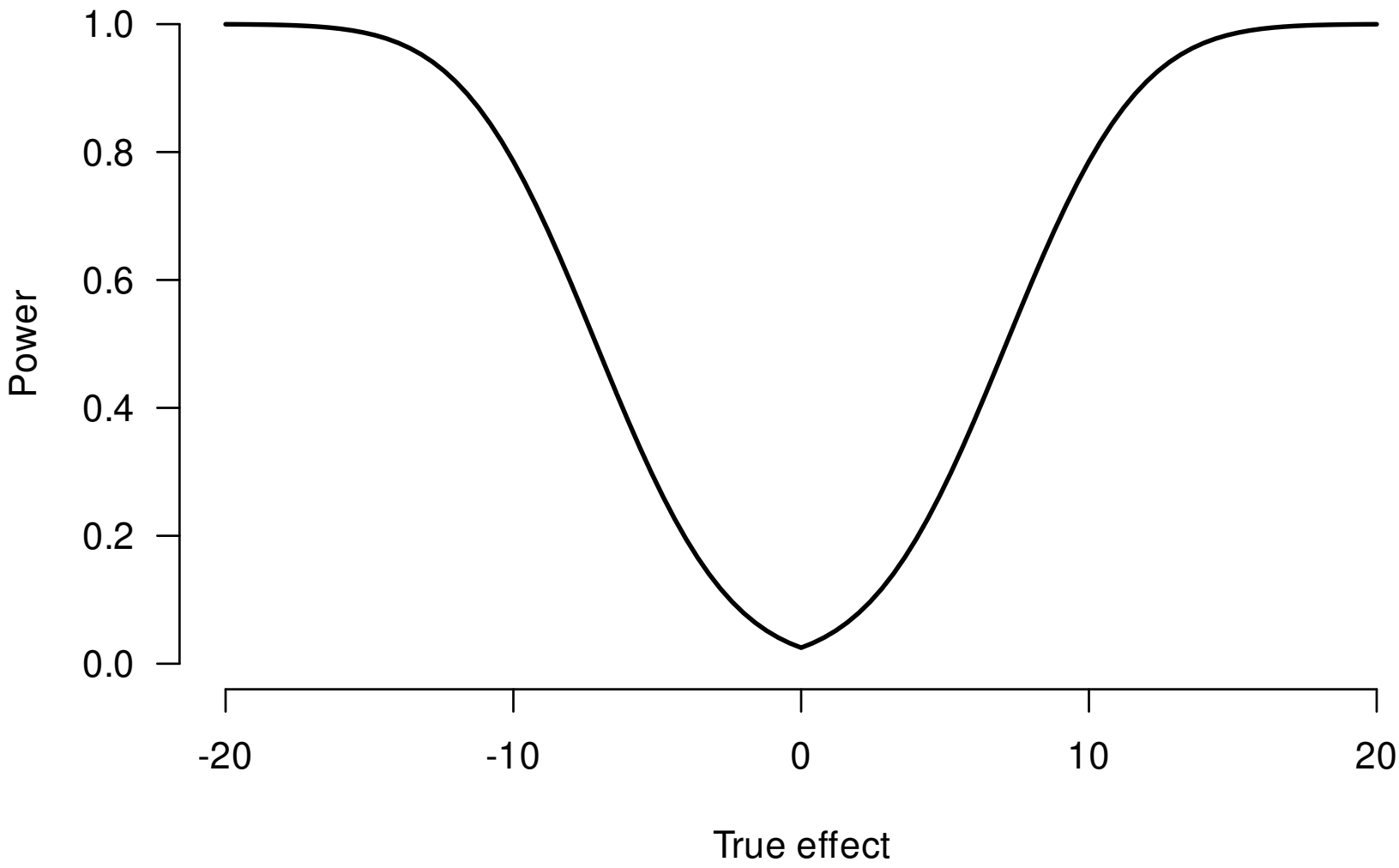
Example

- 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)

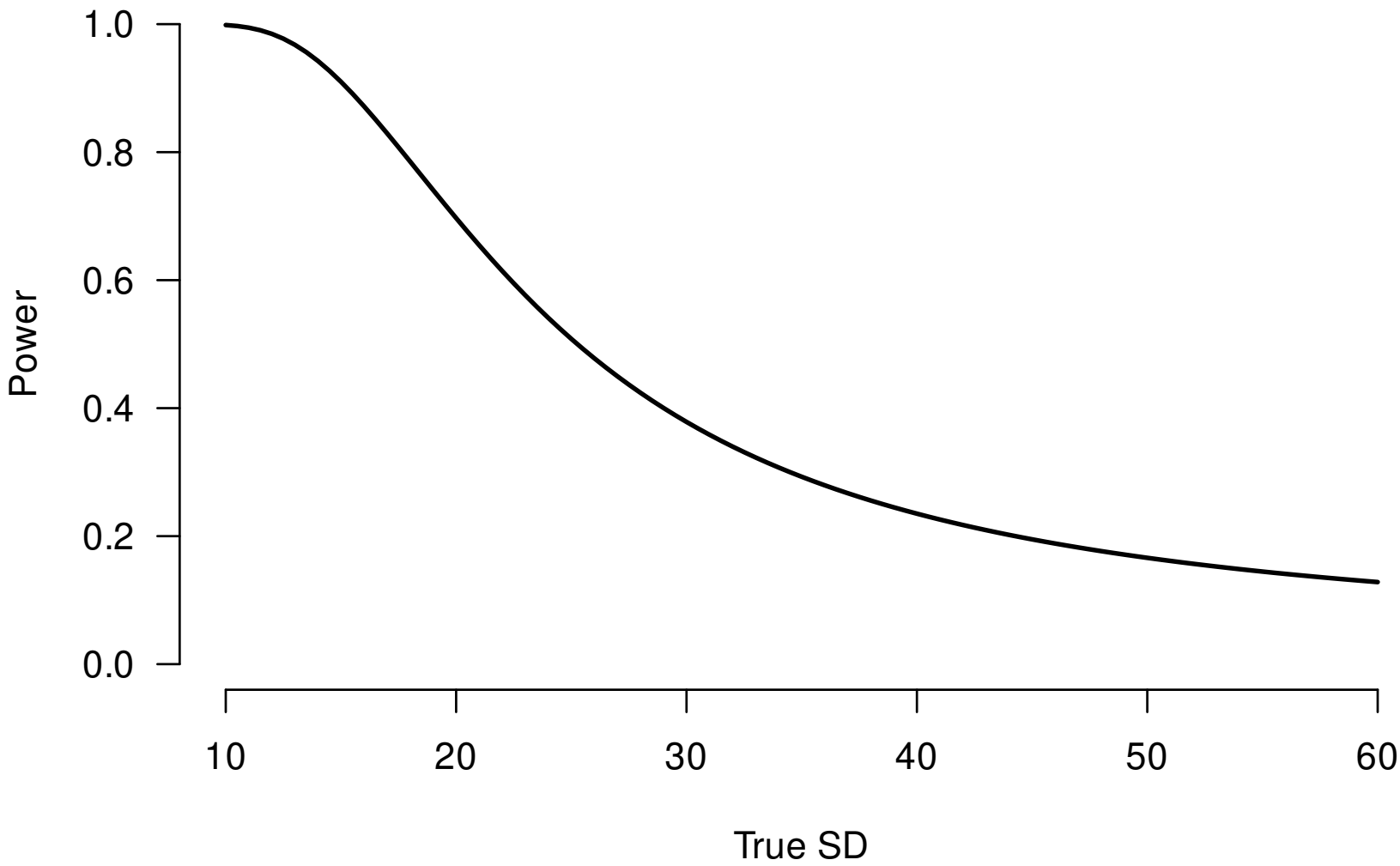
Power curves

- 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

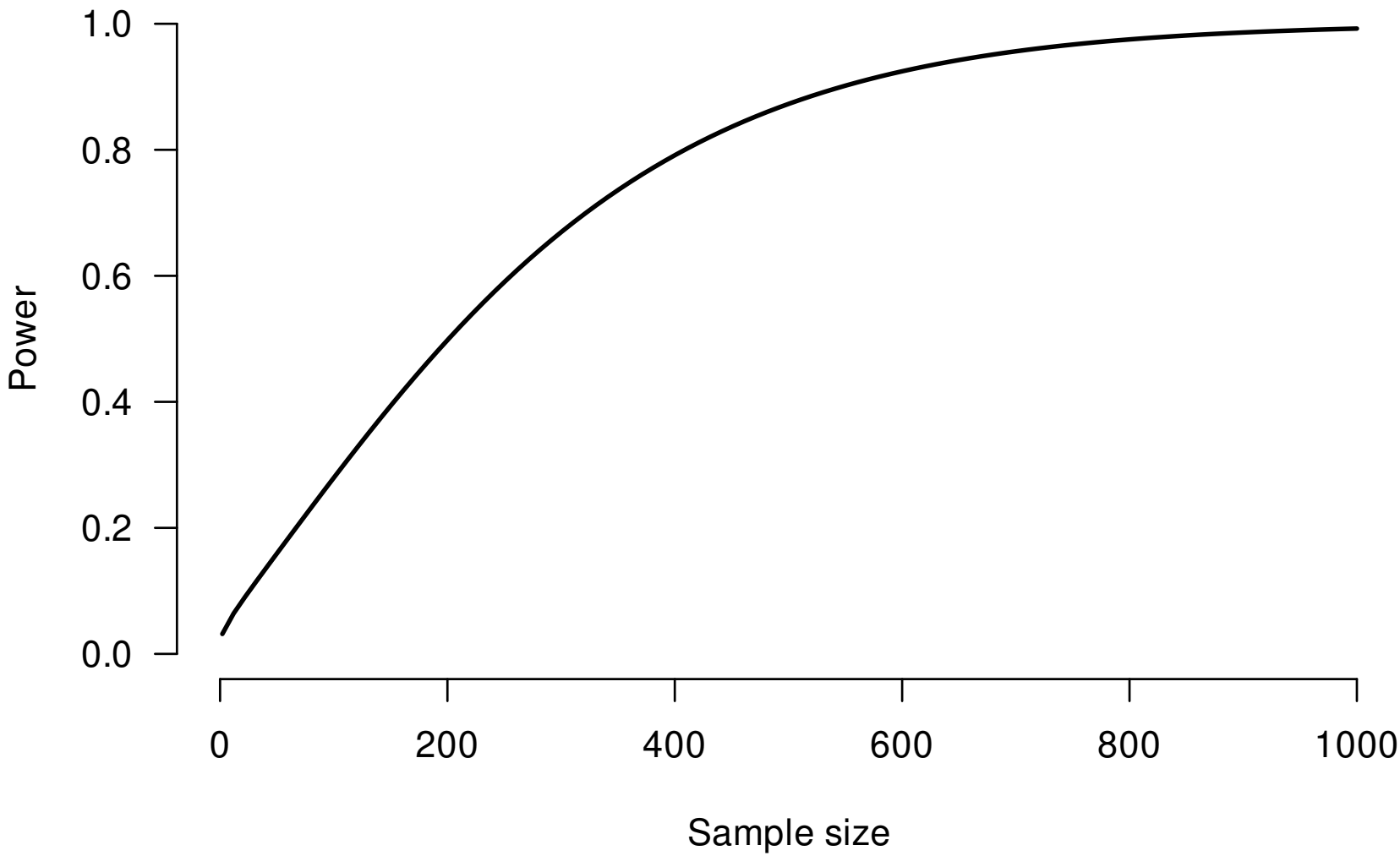
Power curve #1



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