Statistical Significance and Power

November 17

Clair

Big Picture – What are we trying to estimate?

- Causal effect of some treatment

$$E(Y_i \mid T_i=1) - E(Y_i \mid T_i=0)$$

- In words, we're comparing the average outcome among treated group to the average outcome among control group
  - Might want to estimate effects on other summary statistics, too (median, spread of distribution, etc.) but that's more complicated

Why our estimated treatment effect is only part of the story…

- Well, we *estimated* it, right?

- How much do we trust our estimate?
  - What makes a "good estimate"?

  - What makes a precise estimate?

Why our estimated treatment effect is only part of the story…

- Well, we *estimated* it, right?

- How much do we trust our estimate?
  - What makes a “good estimate”?
    - Unbiased
    - No spill-overs
    - High quality data

  - What makes a precise estimate?

Precision – Estimating average height of facilitators

- It matters which one(s) of us you sample!
  - True average height is

    (G) 5’4” + (C) 5’6’’ + (K) 6’2” = 5’8”
  - If you sampled only one of us, your estimate of the average would range from 5’4” to 6’2”
  - If you sampled two of us, your estimate of the average would be one of the following:

    (G+C) 5’5”, (G+K) 5’9”, (C+K) 5’10”

In general, sample of 2/3 gets you closer to the truth

Same deal with our estimates of treatment effect

- As long as we're sampling (not using the whole population), our sample estimate of the mean isn't going to be the same as the truth (the population mean)

- Every sample we draw would give us a different estimate of the population mean

- Disturbing, isn't it?

Is our estimate an outlier?

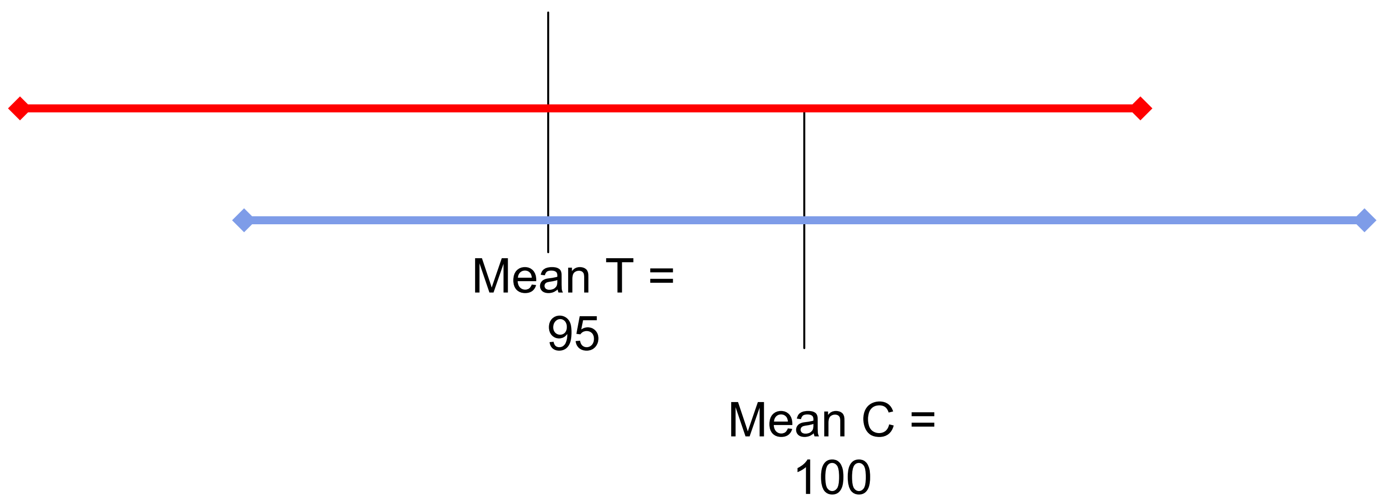
- We'd like to have a sense for whether or not the estimate we got is close to the actual truth – did we accidentally get an outlier?
- The problem is that we don't actually know anything about the truth.
- Solution: assume our estimate is the truth, and consider what the outliers would be in that case

Confidence Intervals

- Start with our estimate as the middle
- Consider how much bigger or smaller our estimate was likely to be
  - Depends on how much variation there is in the data
  - Which depends on how large our sample is
- Ultimately, we should only compare these confidence intervals, not specific estimates
  - Our estimate is just one of many – how much do they have in common?

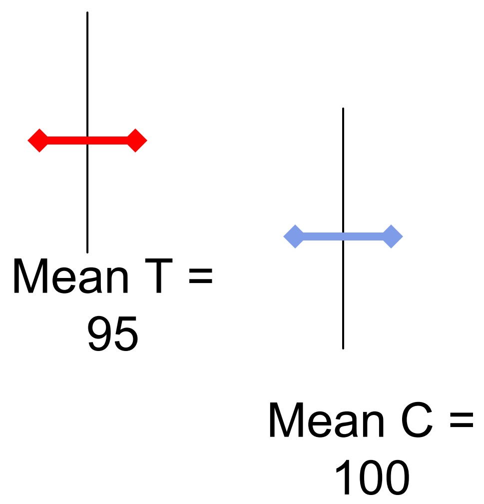
So did the treatment have an effect?

- Compare only the confidence intervals – do they overlap? If so how much?



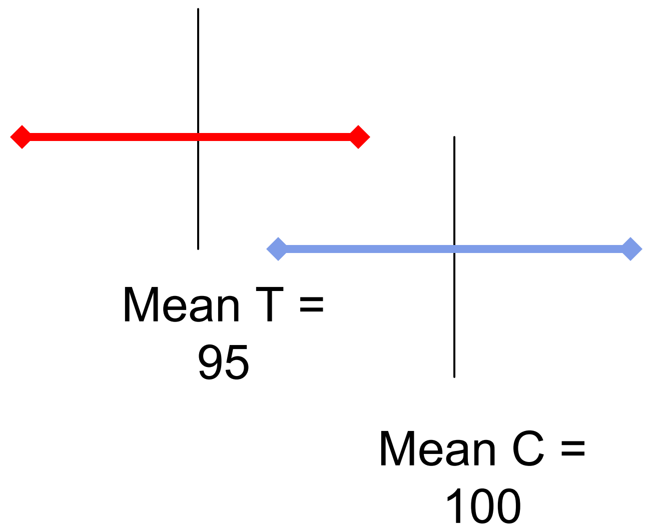
So did the treatment have an effect?

- Compare only the confidence intervals – do they overlap? If so how much?



So did the treatment have an effect?

- Compare only the confidence intervals – do they overlap? If so how much?



OK – so where do these confidence intervals come from?

- First, how “confident” do you want to be?
  - Choose your confidence level based on how many outliers you want to exclude
  - Exclude 5% of the data in the tails for a 95% C.I.
    - 95% of the time, this C.I. includes the truth
  - Exclude 10% of the data in the tails for a 90% C.I.
  - Exclude 1% of the data in the tails for a 99% C.I. if you really want to be sure you’re considering all the options for what the truth might really be

  - The less you exclude, the more conservative you’re being

Calculating the C.I.

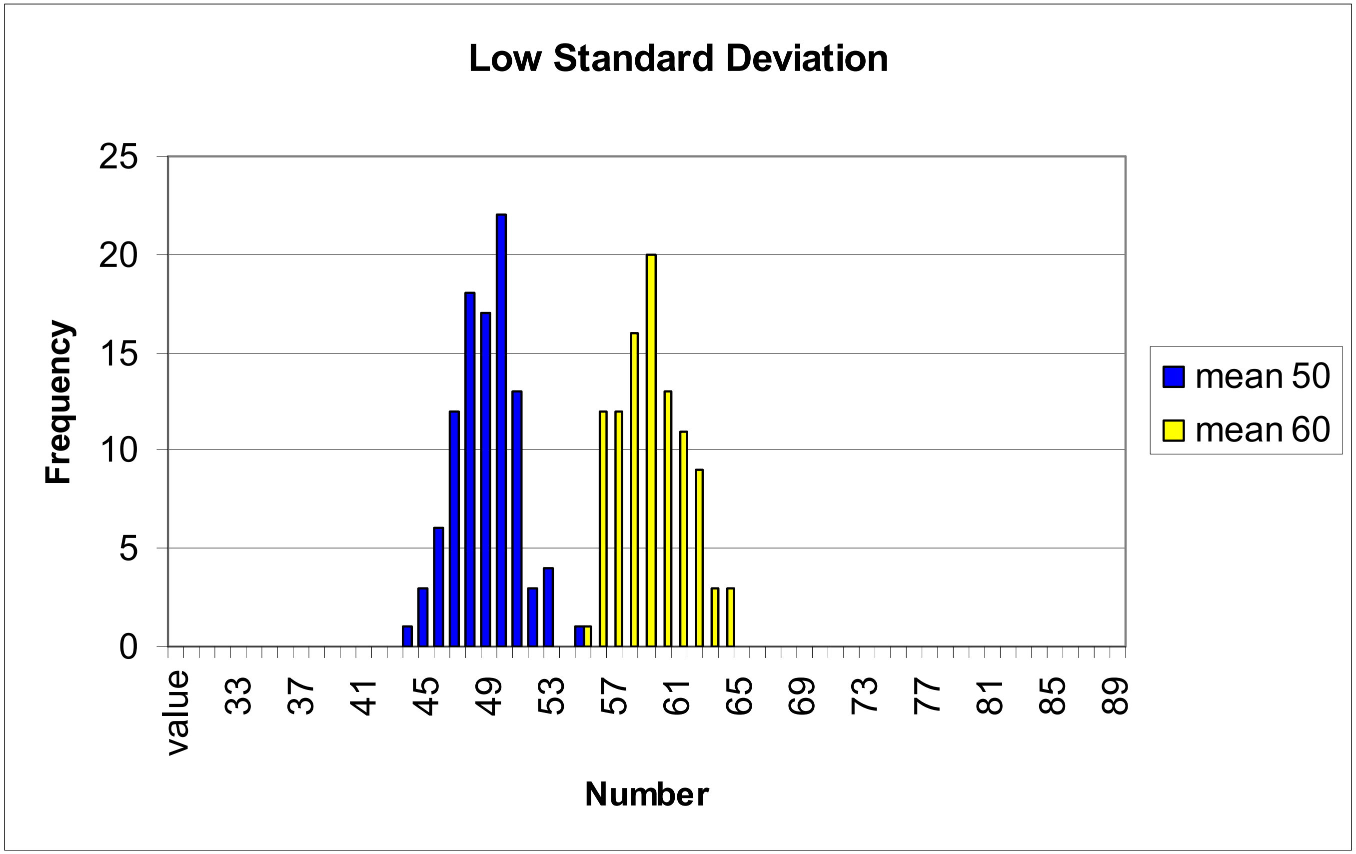
- Choose the confidence level (5%, etc.)
  - That determines your critical value ("2" is a good rule of thumb – about right for 95% C.I.)
- Calculate the sample mean and its standard error (=standard deviation/square root of n)

Sample mean ± 2 × (stddev / square root of n)

- Low variation in outcome (stddev) and large sample both lead to smaller confidence intervals (more precise estimates)

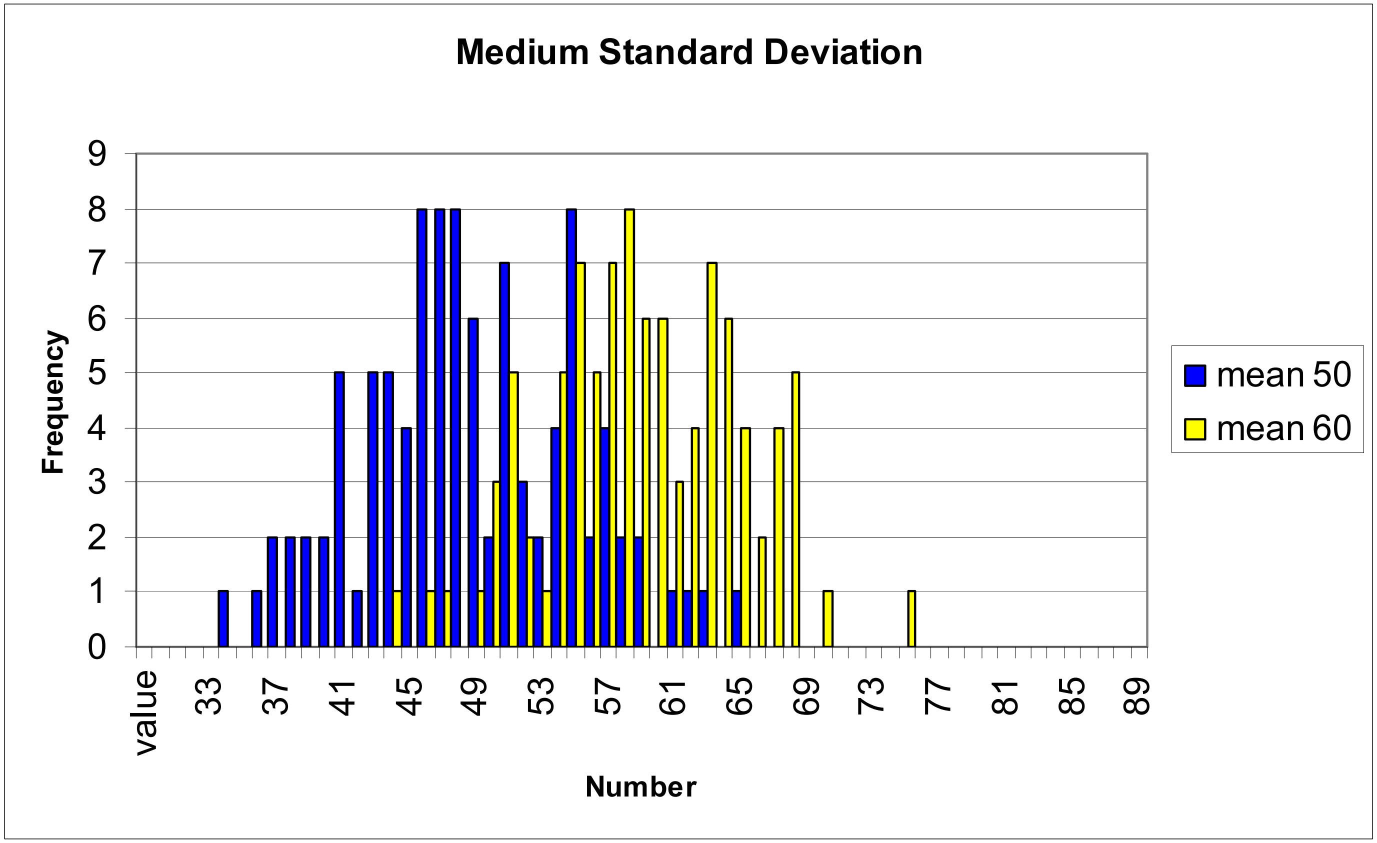
Effect of variation in Y

Graphs by Esther Duflo



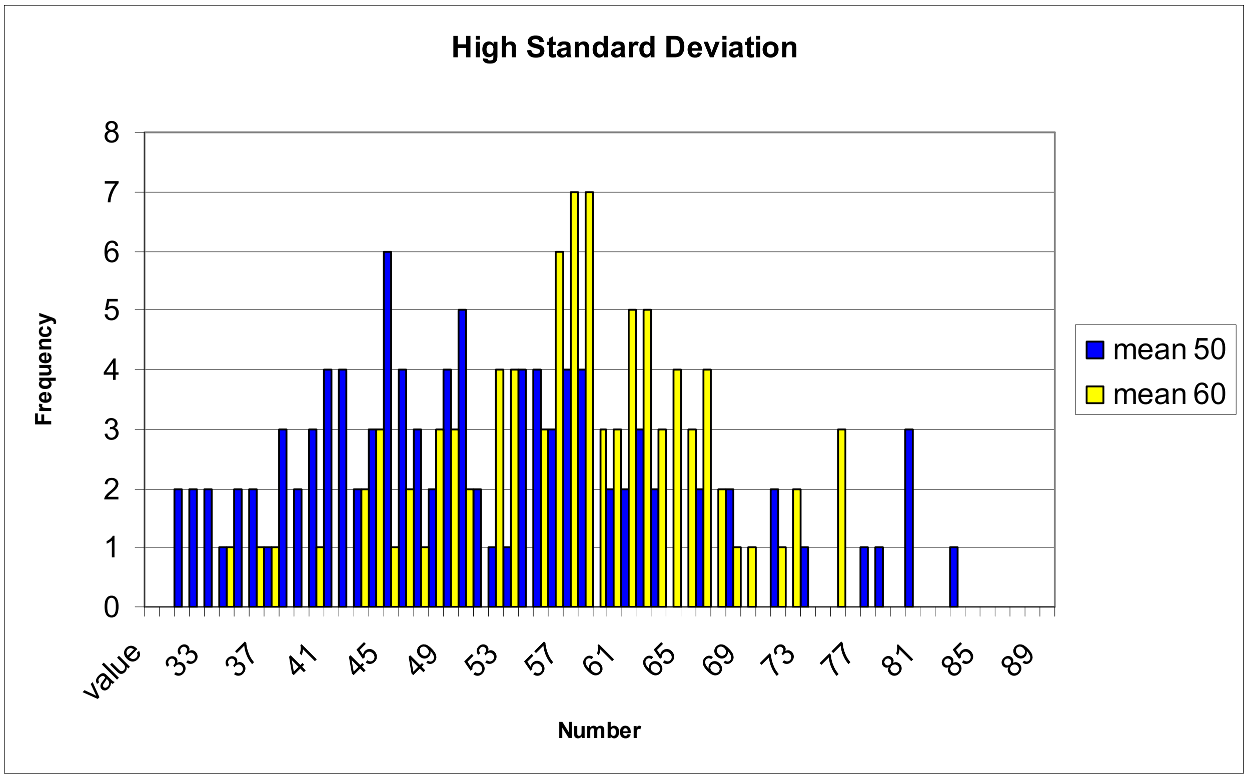
Effect of variation in Y

Graphs by Esther Duflo



Effect of variation in Y

Graphs by Esther Duflo



Data Exercise - Background

- Kenya Rural Water Project (Ted Miguel et. al.)
- Diarrhea is a leading cause of childhood mortality and morbidity
- Kids get sick in part because of dirty water
- Two ways to improve water quality
  - Spring protection
  - Dilute chlorine


FRY

- Same idea as chlorination in developed countries
- Households do it themselves
- One capful per bucket
- Strong taste & smell initially
- Cheap, but requires habit formation

Data Exercise

- Differences-in-Differences method
  - Remember, interested in average outcome for treated versus average outcome for control

- Confidence Interval

$$\text{Sample mean} \pm 2 \times \underbrace{(\text{stddev / square root of n})}_{\text{Standard error}}$$

Data Exercise – Smaller Samples

- What happens to our confidence intervals when the sample size is smaller?
- See for yourselves…

Data Exercise - Summary

(see Excel workbook)

Two types of mistakes

- Conclude that there is an effect, when in fact there is not
  - Confidence (significance) level is the probability that you will make this type of mistake (want it to be low, so usually work with 1-10%)
- Fail to find an effect when in fact there was one
  - Power is the probability you will find an effect

How big should the sample size be?

- What hypothesis are you trying to test?
  - Treatment has no effect; difference between two treatments
- What confidence level do you want?
  - More confidence requires larger sample for given power
- How much variation is there in the comparison group?
  - More variation in comparison group requires larger sample for given power
- How big do you think the effect will be?
  - Smaller effect size requires larger sample for given power

What effect size do you want to detect?

- The smallest one that justifies program adoption
  - Cost of program versus benefits
  - Cost of program versus alternative uses of money

- Careful – if you're too optimistic about what the effect size will be, you might end up with a sample that is too small to detect a difference between T & C

Is this all guesswork?

- Sort of
- Other related studies or baseline data can help with the ingredients of your power calculations
- But there are no guarantees

How are power calculations useful?

- Avoid starting an evaluation that is doomed from the start – no power to detect impacts (waste of time & money)

- Spend enough, but only that much, on the studies you really need

- Can set all but one of the ingredients to power calculation and figure out what that last one would have to be:
  - For 80% power, 95% significance, you can only detect effects of X or more…

Clustering

- If groups of observations are correlated in some way (go to the same clinic/school/spring), need to account for this in estimates – not as much variation as if observations were independent
  - Result is that confidence intervals will be wider
- Number of observations per group might not matter as much as number of groups
  - Be sure to randomize over enough groups!

A good resource for power calculations

- Optimal Design software from UMich
  http://sitemaker.umich.edu/group-based/optimal_design_software
- Plug in confidence (significance) level, group correlation, standardized effect size and see plot of trade-off between number of clusters and power

Summary: Two take-home points

- Consumers of research:
  - Size of the estimate is not enough, also need to consider precision
    - Confidence (significance) level is the probability you incorrectly conclude there was an effect
- Producers of research:
  - Is it worth doing your study? What will the power of your test be?
    - Power is the ability to detect an effect

Moral of the story

- Larger sample size increases both confidence (significance) and power

- Larger effects will be easier to detect (statistically speaking)

- Variation in the outcome variable makes it more difficult to detect program effects