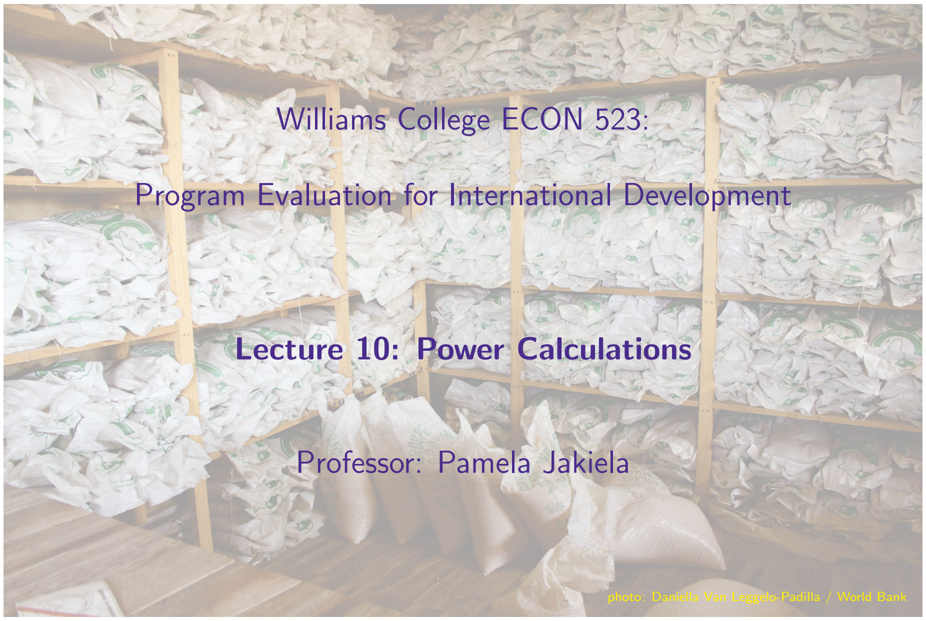


Williams College ECON 523:

Program Evaluation for International Development

Lecture 10: Power Calculations

Professor: Pamela Jakiela

photo: Daniella Van Leggelo-Padilla / World Bank

Hypothesis Testing

Every research design involves:

1. A null hypothesis and either one or several alternative hypotheses
2. A statistic related to our hypotheses that we can calculate from our data
3. A rule mapping values of our statistic into decisions about whether to reject the null

Type I and Type II Errors

Any statistic is a random variable:

- We don't know for sure that the statistic we calculate will be very close to the "true" value
- If we draw a random sample, the sample mean could be far from the population mean
- If we randomly assign treatment, most of the male/female/rural/urban/young/old/etc people could happen to end up in the treatment group rather than the control group

Our hypothesis testing procedure could lead to a mistake:

- We might see what looks like a large impact when there is no impact (Type I error)
- We might see what looks like no impact when there actually is an impact (Type II error)

Power Calculations and Sample Size

When we do **power calculations**, we are:

1. Choosing a maximum acceptable probability of Type I error, called our **test size**
   - We always choose $\alpha = 0.05$ (ever since Fisher)
2. Choosing a minimum acceptable probability of avoiding a Type II error, called our **power**
3. Figuring out how large a sample we need to achieve (1) and (2), which will depend on:
   - The properties of our outcome variable and chosen test statistic
   - Our beliefs about the likely impact of our policy intervention

Type I Errors and Test Size: Example

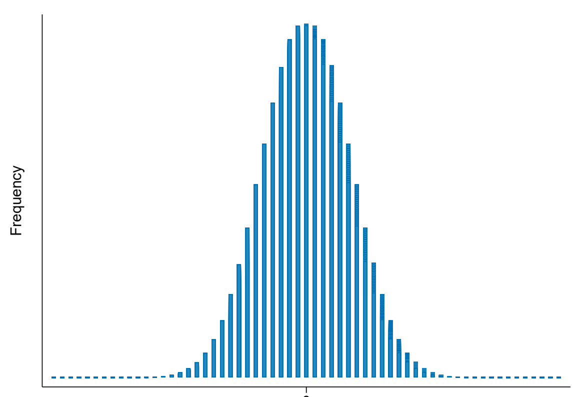
Randomly assign treatment in a sample of 1,000 people so that 500 are treatment, 500 control

- There are a lot of different possible random assignments
- Sometimes more tall people end up in treatment (or comparison) group

Test the hypothesis that average height is higher in treatment than in control

- Remember: we haven't implemented any type of treatment
- Average heights should be similar in both groups, on average
- Difference in means could still be large, depending on treatment assignment

Type I Errors and Test Size: Example

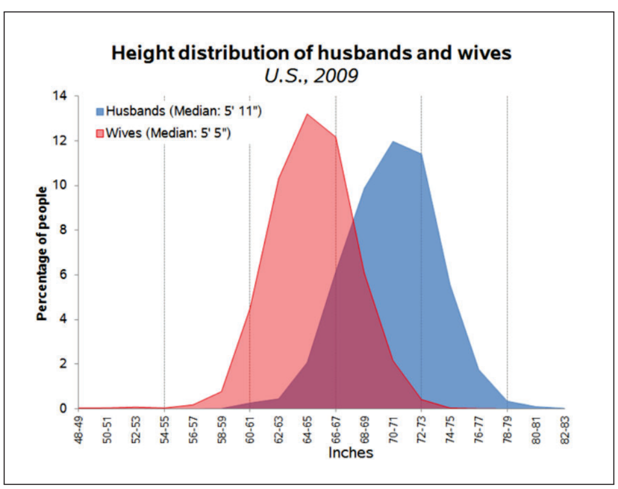


Test Size: The Probability of a Type I Error

The standard practice is to reject the null hypothesis when our test statistic translates into a p-value below 0.05, so statistics that "large" occur less than 5% of the time under the null

- Statistically significant differences occur by chance about 5% of the time
  - Why we stratify when randomly assigning treatment
  - Can lead to "publication bias" because findings get published (non-findings don't)
- Depends on sample size: larger sample $\Rightarrow$ smaller variance of chosen sample statistic
  - Test size is essentially fixed, but what $t = 1.96$ means in terms of outcome units depends on sample size, careful measurement of outcomes, ability to include meaningful control variables

Statistical Power



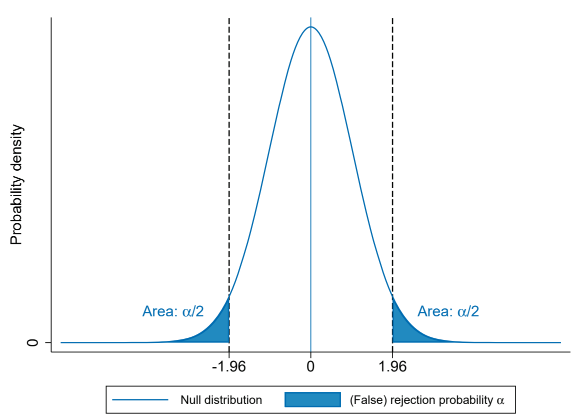
Source: Cohen (2013)

Statistical Power

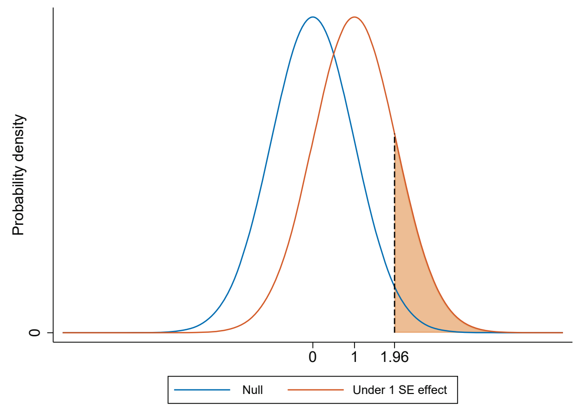
Wives (women) are shorter than husbands (men), on average, but distributions overlap

- If I measure the heights of N wives and N husbands, how likely is it that the difference in means will be large enough to reject the null hypothesis (given test size of $\alpha = 0.05$)?
- If I drew hundreds of different samples of N wives and N husbands, sometimes I'd see a large difference in average height, and sometimes I'd see a small difference in height
  - Estimated difference in height **when there is actually a difference** will also be normally distributed around the true (population) difference between male and female heights
  - If I knew the true difference in mean heights and its variance, I could tell you – for a fixed sample size – how often I'd expect to observe a statistically significant height difference
  - Depends on the variance of height in population, and on the sample size

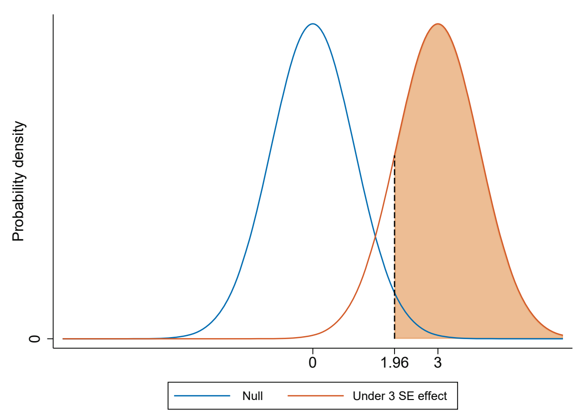
How Often Will We Reject the Null?



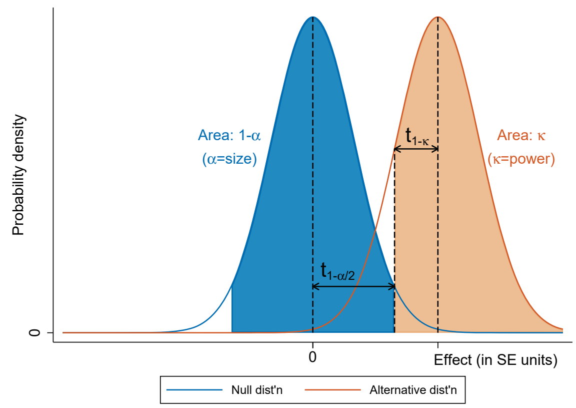
Power with a 1 SD Effect



Power with a 3 SD Effect



Statistical Power



The Minimum Detectable Effect

The **minimum detectable effect** (or MDE) is the smallest effect size that we can detect with power of 0.8 (i.e. the probability of a Type II error, failing to reject a false null, is 0.2)

$$\begin{aligned} \text{MDE} &= \left(t_{\alpha/2} + t_{1-\kappa}\right)\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \\ &\approx 2.8\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \end{aligned}$$

where:

- P is the proportion of the sample assigned to treatment
- N is the sample size
- σ^2 is the variance of the outcome

Implications of the Minimum Detectable Effect Formula

1. MDE is decreasing in N: a larger sample means great statistical power
2. MDE is maximized when $P = 1/2$ (though other factors such as costs may come into play)
3. MDE can be expressed in SD units, or converted into outcome variable units
   - Prior studies can suggest plausible effect sizes
   - Existing data sets can be used to calculate variance of some outcomes
   - For binary outcomes, variance only depends on mean and sample size

Practice Problem 1

You have a sample size of $N = 200$, but you can only offer the program to 50 people.

Find the MDE in standard deviation units.

Practice Problem 2

Half of your sample is allocated to treatment and the other half to control.

Find the sample size needed to have an MDE of 0.25 standard deviations.

Practice Problem 3

You would like to know whether vocational training improves the profits of self-employed youth. Your main outcome of interest is income, which has a mean of 968 and a standard deviation of 11,842. If you assign half of your sample to treatment and half to control, how large of a sample would you need to have an MDE equivalent to 50% of average income?