Applied Microeconometrics

Lecture 2: Statistical power and post-study probability

Manuel Bagues

August 2022

Lecture Slides

Previous lecture

- Reproducibility crisis
- False positives and multiple testing
  - The garden of forking paths (Andrew Gelman)

- Note that even if researchers do not fabricate the data and they do not undertake any fishing expeditions, we still may have a problem.
- Researchers might in good faith select the specification that gives a significant result
- But we need to look at the entire garden of forking paths and recognize how each path can lead to statistical significance in its own way
- Otherwise we might end up being too confident about the statistical significance of our results.

Today:

- How does the lack of statistical power affect the probability of:
  1. a false positive?
  2. a false negative?
- How should we think about empirical estimates that are statistically significant but imprecise?
- Post-study probability that a significant finding is true

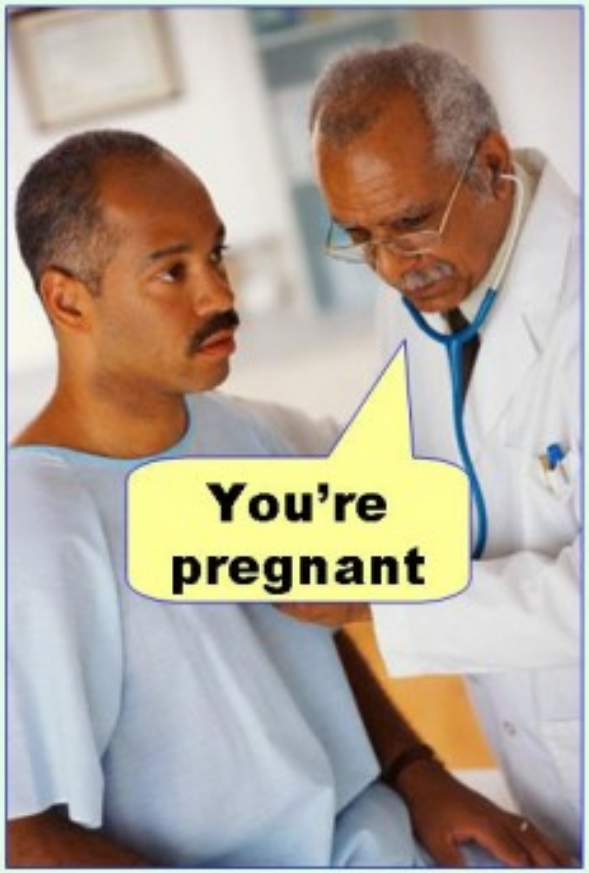
Reminder of basic statistics

- In economics we typically we follow the so-called frequentist or classical approach to hypothesis testing
- Usually our null hypothesis is that the treatment has no effect
  $H_0 : \beta = 0$
  $H_1 : \beta \neq 0$
- The type-I error or false-positive (α) is the incorrect rejection of the null hypothesis
  - In economics we usually consider as a standard level of significance 5% (this goes back to Fisher 1935)
- The type-II error or false-negative (β) is the incorrect non-rejection of the null hypothesis
  - It depends on the size of effect, sample size and the variability in the data
  - Power=1-β $\rightarrow$ probability of detecting the effect of the treatment (i.e. reject the null hypothesis)
  - Useful to do power calculations before you conduct your study (more on this below)

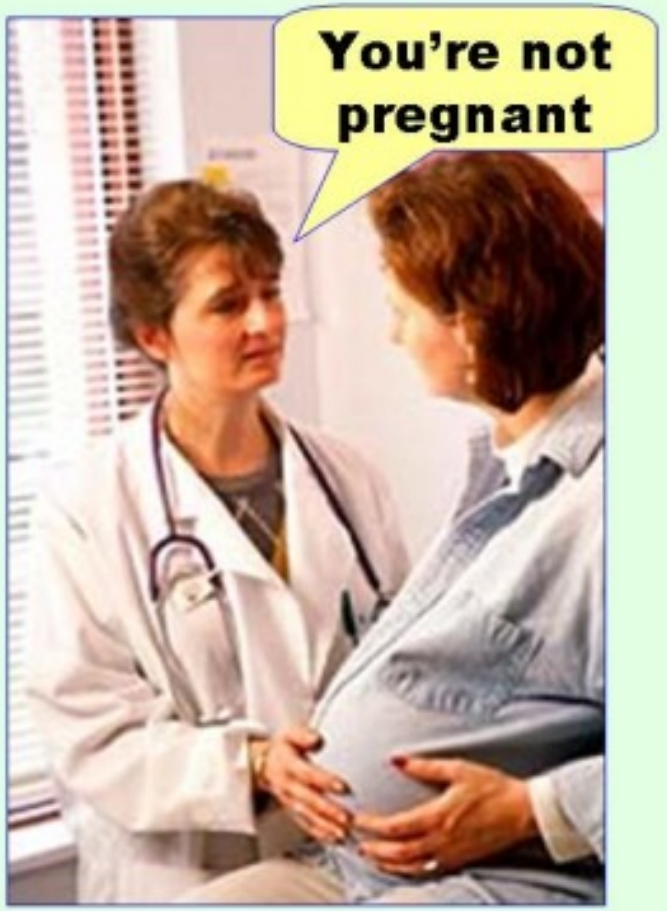
Type I and Type II errors

H_0 : not pregnant; H_1 : pregnant

Type I error
(false positive)



Type II error
(false negative)



The quest for (statistical) power in Economics

- Power! Unlimited power!

The quest for (statistical) power in Economics

- Unfortunately, very often we lack statistical power (e.g. sample size is small relative to the size of the effect that we are trying to detect and the variability of the data)
  - Small samples
    - RCTs
    - Observational data: limited number of regions, countries, etc.
  - Small effects
    - Most of the low hanging fruits have already been picked (e.g. discrimination against African-Americans)
- Most studies in Economics are underpowered (Ioannidis, Stanley and Doucouliagos 2017)
  - 6,700 empirical studies
  - median statistical power is 18%
  - only 10.5% of papers with power$\geq$80%

Lack of power and false positives

- First, let us discuss how the lack of statistical power affects the probability of a false positive
- Would larger samples help to decrease the probability of false positives?

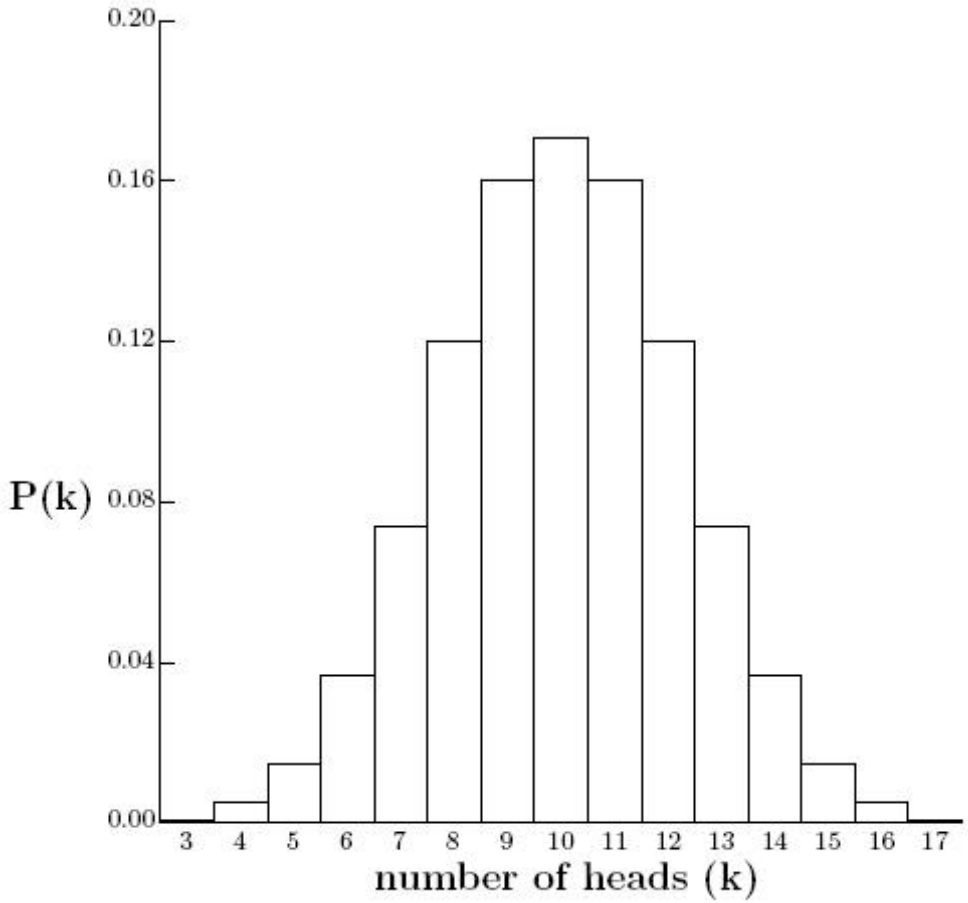
Lack of power and false positives

- First, let us discuss how the lack of statistical power affects the probability of a false positive
- Would larger samples help to decrease the probability of false positives?
  - In a context where the true effect is zero, the magnitude of the estimate will tend to be lower...
  - ... but the probability of this estimate being significantly different from zero does not vary with sample size.

Simple example: coin tossing

- Fair coins: 50% probability of heads or tails
- H_0 : coin is fair; H_1 : coin is unfair
- Type-I error: likelihood of falsely concluding that a coin is unfair
- If we fix the significance level to be equal to 5%, what is the likelihood to reject that a coin is fair if we throw the coin:
  - N=20
  - N=100

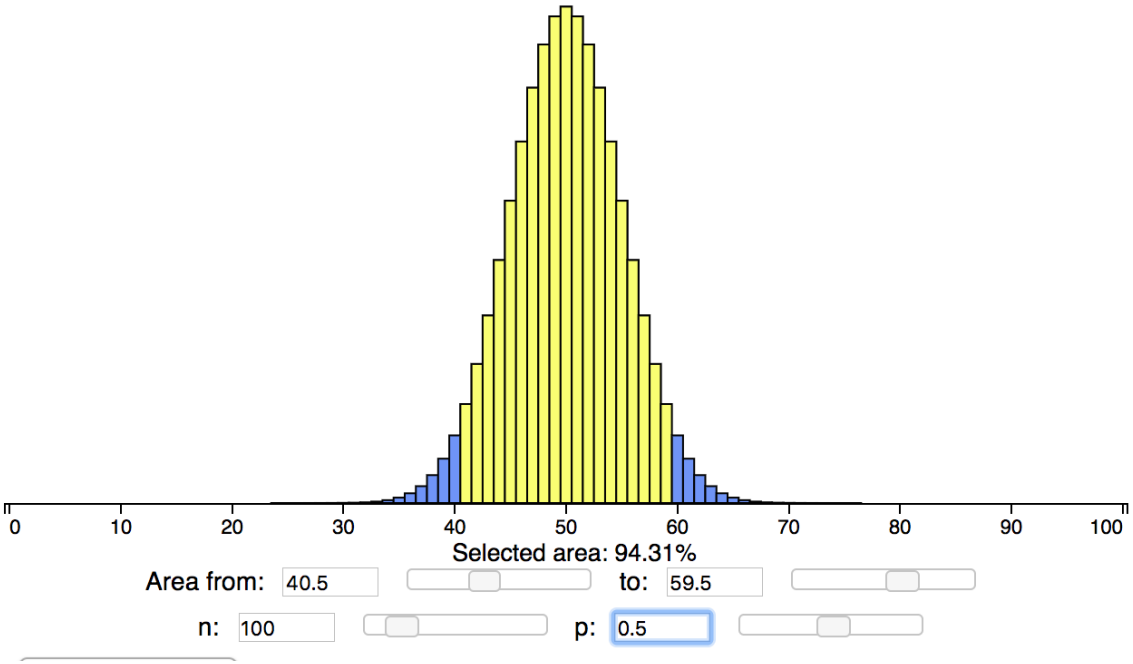
Histogram - Number of heads in 20 coin tosses



Type-I error

- N=20
- If we throw a fair coin 20 times, around 95% of the time we will get between 6 and 14 heads
- We reject that the coin is fair whenever the number of tails/heads is below 6, which happens (roughly) 5% of the time if the coin is fair (4% to be precise)

Histogram - Number of heads in 100 coin tosses



Would larger samples help to decrease the probability of false positives?

- N=100
- When N is larger, we adjust our rejection criteria in such a way that we always reject 5% of the time (assuming that the coin was fair)
- If we throw a fair coin 100 times, around 95% of the time we will get between 41 and 59 heads
- We reject that the coin is fair whenever the number of tails/heads is below 41, which happens (roughly) 5% of the time
- N=20 vs N=100
  - N=20 $\rightarrow$ reject if heads/coins less than 30% (6 out 20)
  - N=100 $\rightarrow$ reject if heads/coins less than 41% (41 out of 100)
  - This happens 5% of the time in both cases

Lack of power and false negatives

- Would larger samples help to decrease the probability of false negatives?

Lack of power and false negatives

- Would larger samples help to decrease the probability of false negatives?

$\rightarrow$ Of course!

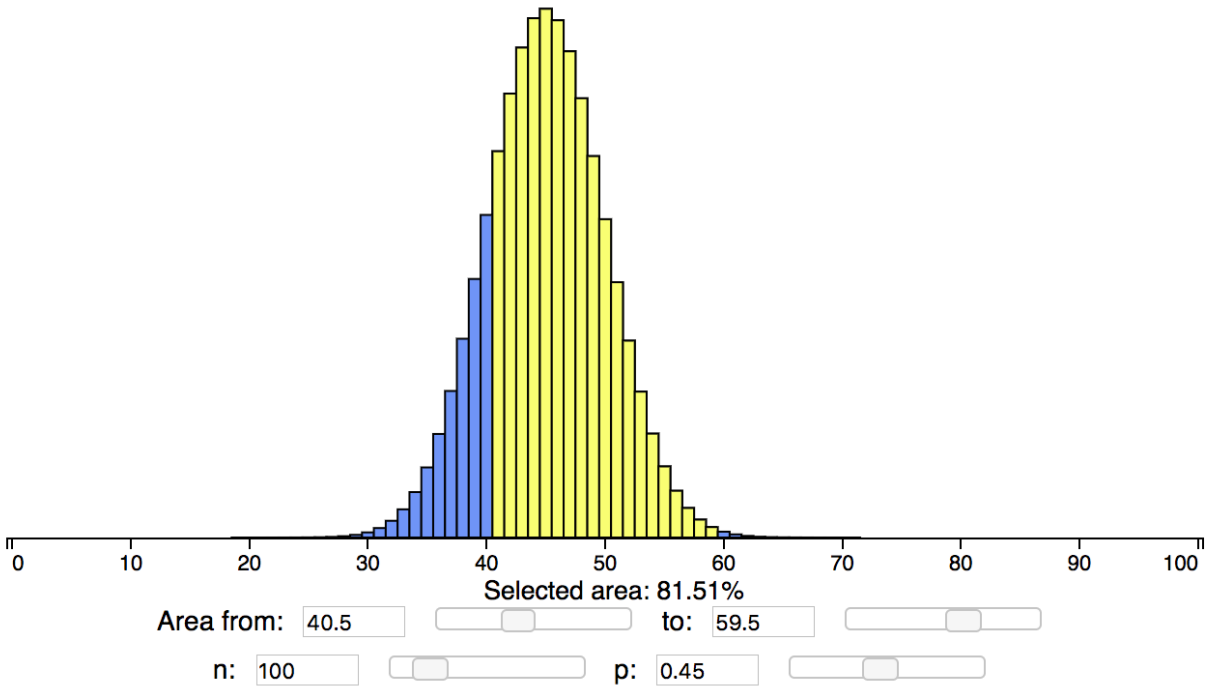
Coin tossing example

- Imagine that there are two possible types of coins:
  - Fair coins: 50% probability of heads or tails
  - Coins with small imperfections: 45% heads
- How likely are you to detect an imperfect coin [prob(tails)=0.45] if you toss the coin:
  - 100 times
  - 20 times

Coin tossing example

- Imagine that there are two possible types of coins:
  - Fair coins: 50% probability of heads or tails
  - Coins with small imperfections: 45% heads
- How likely are you to detect an imperfect coin [prob(tails)=0.45] if you toss the coin:
  - 100 times
  - 20 times
- When N=100, we would reject whenever we obtain less than 41 heads/tails. How often would you say that it occurs if the coin is imperfect?

Histogram - Imperfect coin [prob(tails)=0.45] - Number of heads in 100 coin tosses



Coin tossing example

- How likely are you to detect an imperfect coin [prob(tails)=0.45] if you toss the coin 100 times? $\rightarrow$ 18.5% of the time
- And if N=20?
  - When N=20, we would reject if we obtain less than 6 tails or heads
  - How often is this happening (assuming the coin is imperfect)? $\rightarrow$ 6%
- Note also that the magnitude of any significant estimates will be 'too' large
  - When there is a 0.05 imperfection (from 0.45 to 0.50), we get a significant result only when in the sample we observe an "effect" of at least 0.25 (25% of heads/tails)

Type-II error

- How large should be the sample in order to have power=80%?
  - N=780 (stata command: *power oneproportion 0.45 0.50, power(0.80) alpha(0.05)*)

What are the consequences of the lack of power?

- How does it affect:
  1. the probability of a false positive?
     - No effect
  2. the probability of a false negative?
     - Most likely we will be unable to detect the effect, even if it exists
- How should we think about empirical estimates that are statistically significant but imprecise?

What are the consequences of the lack of power?

- How does it affect:
  1. the probability of a false positive?
     - No effect
  2. the probability of a false negative?
     - Most likely we will be unable to detect the effect, even if it exists
- How should we think about empirical estimates that are statistically significant but imprecise?
  - When power is low, if we find significant results, their magnitude will be implausibly large (or even with the wrong sign)
  - This problem is exacerbated by journals' demand for papers with significant results (instead of precise estimates)

What are the consequences of the lack of power?

- How does it affect:
  1. the probability of a false positive?
     - No effect
  2. the probability of a false negative?
     - Most likely we will be unable to detect the effect, even if it exists
- How should we think about empirical estimates that are statistically significant but imprecise?
  - When power is low, if we find significant results, their magnitude will be implausibly large (or even with the wrong sign)
  - This problem is exacerbated by journals' demand for papers with significant results (instead of precise estimates)
- Let us illustrate this point with two examples:
  - Gelman and Weakliem (2009), 'On Beauty, Sex and Power'
  - Oster (2005), 'Hepatitis B and the Case of the Missing Women'

Of beauty, sex, and power: Statistical challenges in estimating small effects

How should we think about empirical estimates that are statistically significant but imprecise?

- Gelman and Weakliem (2009) discuss an article titled 'Beautiful parents have more daughters' (Kanazawa, Journal of Theoretical Biology 2007)
- Generalized Trivers-Willard hypothesis:
  - parents who possess a heritable trait which increases the female (male) reproductive success relatively more will have more daughters (sons)
- Kanazawa focuses in this paper on the role of physical attractiveness
  - beauty increases relatively more the reproductive success of daughters $\rightarrow$ physically attractive parents tend to have more daughters

- Survey data: N≈3,000
- Attractiveness was measured on a 1-5 scale (very unattractive to very attractive)
  - 327 individuals classified as very attractive (category 5)
  - 2645 individuals in other groups (categories 1-4)
- Gender of children
  - 56% of children of parents in category 5 were girls
  - 48% of children of parents in categories 1-4 were girls
- Statistically significant (2.44 s.e.'s from zero, p = 1.5%)
- What should we make out of this result?

- Levitt at the Freakonomics blog:
  - A new study by Satoshi Kanazawa, an evolutionary psychologist at the London School of Economics, shows that (. . .) good-looking parents are 36 percent more likely to have a baby daughter as their first child than a baby son - which suggests, evolutionarily speaking, that beauty is a trait more valuable for women than for men. The study was conducted with data from 3,000 Americans, derived from the National Longitudinal Study of Adolescent Health, and was published in the Journal of Theoretical Biology.

- Several potential problems
  1. Observational evidence
     - Maybe beauty correlates with other things?
     - Incidentally: Gelman also complains about including as a control the number of children, why?
  2. Multiple-testing problem
  3. Small sample (or small effects)

Background on sex ratios

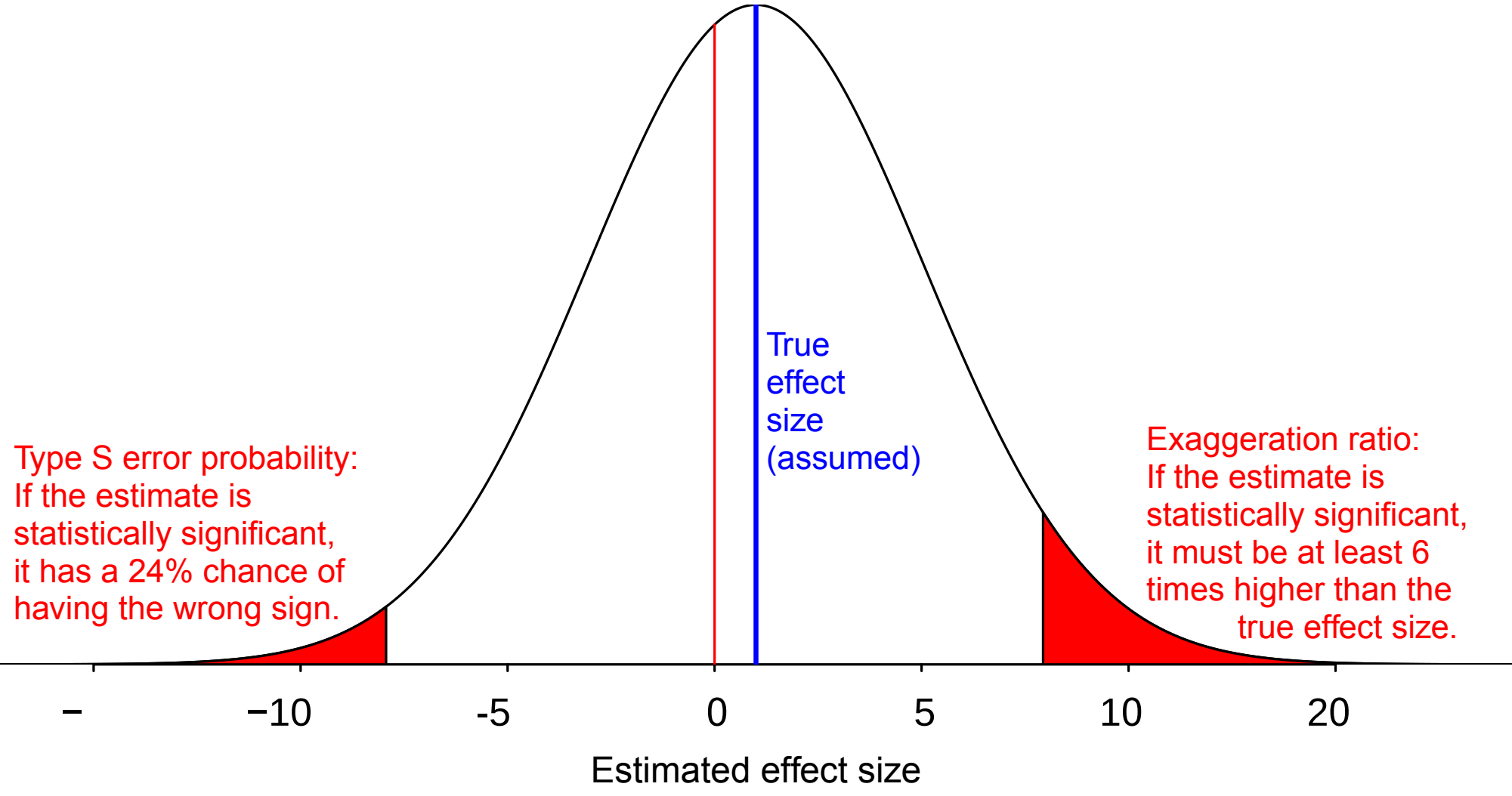
- Pr(boy birth)=51.5%
  - Evolutionary story: Fisher's principle $\rightarrow$ 1:1 ratio at reproductive age
  - Boys (used to) die at a higher rate than girls
  - At reproductive age, the number of boys and girls was about the same
- What can affect Pr(boy births)?
  - Race, parental age, birth order, maternal weight, season of birth: effects of about 1% or less
  - Extreme poverty and famine: effects as high as 3%
- We expect any effects of beauty to be less than 1%

Ex-ante power calculations

- The author could have calculated the power of the analysis beforehand.
- Power calculations require:
  - fix $\alpha \rightarrow 5\%$
  - sample size $\rightarrow$ n_1=327 and n_2=2645
  - some assumption about the 'variabilitiy' in the data
    - in this case the outcome variable is a dummy $\rightarrow$ variance = p * (1-p)
    - otherwise we need some estimate
  - some assumption about the potential magnitude of the effect
- If the true effect was equal to 0.3 p.p.:
  - *Stata: power twoproportions 0.515 0.518, n1(327) n2(2645)*
  - power=0.0512 (slightly above the 5% of false positives)
  - We only observe significant results with a 5.12% probability
- If the true effect was equal to 1 p.p.:
  - *Stata: power twoproportions 0.515 0.525, n1(327) n2(2645)*
  - power=0.0635
- If the true effect was equal to 2 p.p.:
  - *Stata: power twoproportions 0.515 0.535, n1(327) n2(2645)*
  - power=0.105

- Let us consider the hypothesis that the effect = 1 p.p.
- Type 2 error: Most of the time (94%) results will not be significant
- Type M error: Any statistically significant finding is necessarily a huge overestimate!
  - If estimates are statistically significant, the estimate must be at least 2 standard errors away from 0 ... (above 6 p.p.)
- Type S error
  - And moreover the sign of the estimate may be even in the wrong direction
- You can find a detailed discussion of the following graph here:
  - https://alexanderetz.com/2015/05/21/type-s-and-type-m-errors

This is what "power = 0.06" looks like. Get used to it.

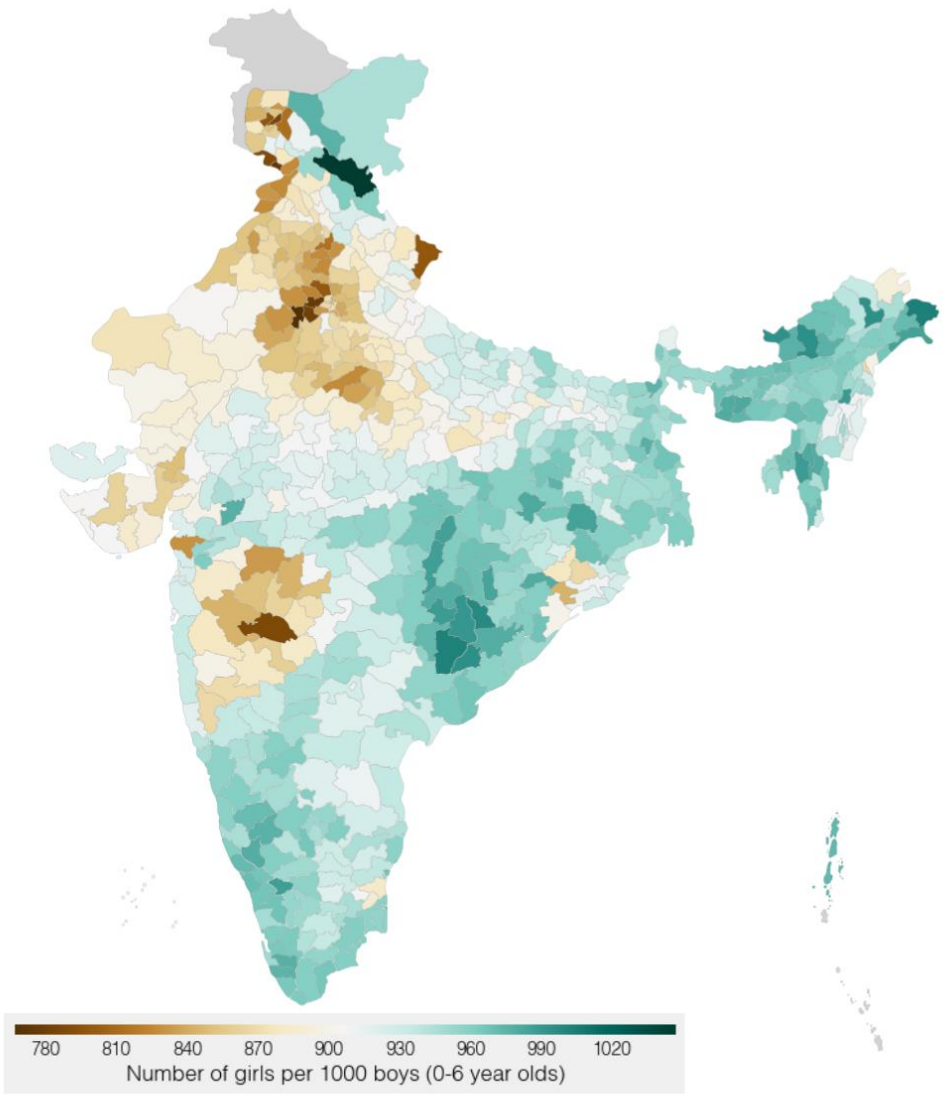


Economists also make similar mistakes

Hepatitis B and the Case of the Missing Women (Oster 2005)

- More Than 100 Million Women Are Missing (Amartya Sen 1990)
- His subject was the wildly off-kilter sex ratios in India, China and elsewhere in the developing world.
- Sen invoked the “neglect” of third-world women, citing disparities in health care, nutrition and education.

Girls per 1000 boys among 0-6 year olds (2011)



Source: India Census 2011 (devinfo.org/indiacensuspopulationtotals2011/libraries/aspx/Home.aspx)
Author: Rishabh Srivastava (sriv.org) | The Numerate Indian (TheNumerateIndian.com)

Economists come up with the explanation!

Hepatitis B and the Case of the Missing Women (Oster 2005)

- According to some studies, among women with the HBV, the gender ratio is around 150 (boys per 100 girls)
  - Note: these studies include only a few hundred observations!
- Hepatitis B is common in many Asian countries, especially China, where some 10-15 percent of the population is infected.
- Oster (2005) concludes that, *(u)sing data on prevalence of the virus by country and estimates of the effect of hepatitis on the sex ratio, I argue that hepatitis B can account for about 45 percent of the 'missing women'*

However, other authors suggested additional testable implications that did not match the evidence

- Das Gupta (2005) pointed out whether or not females 'go missing' is determined by the existing sex composition of the family into which they are conceived. Girls with no older sisters have similar chances of survival as boys. However, girls conceived in families that already have a daughter experience steeply higher probabilities of being aborted or of dying in early childhood. This indicates that cultural factors still provide the overwhelming explanation for the 'missing' females.
- Ebenstein (2007) questioned Oster's conclusion based on the fact that among first born children the sex ratio is close to the natural one. It is the skewed female-male ratios among second and third born children that account for the bulk of the disparity.
- To her credit, Oster published eventually a paper titled 'Hepatitis B does not explain male-biased sex ratio in China'

And the small-sample studies documenting a strong relationship between hepatitis-B and children's gender turned out to be unreliable

- Let us ignore the potential problem that hepatitis is not exogenous
- Oster's analysis was based on several previous studies with small samples and large (barely significant) estimates
  - It could not be otherwise - papers with insignificant and imprecise estimates would not be published (or even written)
  - therefore, we should be cautious about giving face value to this type of evidence
- In a study published in the American Economic Review, Lin and Luoh (2008) use data on almost 3 million births in Taiwan over a long period of time
  - The effect of maternal Hepatitis B infection on the probability of male birth is, at best, small, at most one quarter of one percent.

Why was this not obvious to Oster?

- Statistical theory and education are focused on estimating one effect at a time
- Statistical significance is a useful idea, but it doesn't work when studying very small effects
- There are methods for including prior knowledge of effect sizes, but these methods are not well integrated into statistical practice

How should we deal with this problem?

- Ex-ante: Make power calculations
  - To detect with a 80% probability a 1 p.p. difference between two hypothesized proportions p1 and p2 with a study of size n, equally divided between the two groups, a conservative sample size is
    $n = [2.8/(p1 - p2)]^2 = [2.8/0.01]^2 = 78400$
  - or if you prefer using stata type:
    power twoproportions 0.505 0.515, power(.80)
- Ex-post: Calculate post-study probability that the hypothesis is true
  - Post-Study Probability (PSP): the probability that a declaration of a research finding, made upon reaching statistical significance, is true.
- Two ways:
  - Go fully bayesian: prior distribution + signal $\rightarrow$ posterior distribution
  - Simpler way: assume degenerate probability distributions (Maniadis, Tufano and List 2014)

Post-Study Probability (Maniadis, Tufano and List 2014)

One Swallow Doesn't Make a Summer

- Imagine that we give face value to Kanazawa's results
- How much would you update your priors about the probability that beautiful parents have more daughters?
- Let us follow the notation used by Maniadis, Tufano and List 2014:
  - π represents the prior probability that a certain hypothesis is true (e.g. beautiful parents have are 1 p.p. more likely to have a daughter)
  - α is the significance level of a result (e.g. 5%)
  - $1 - \beta$ denotes the power of the experimental design

Post-study probability (PSP)

- Probability of a true positive=P_t=$(1-\beta)*\pi$
- Probability of a false positive=P_f=$\alpha*(1-\pi)$
- Probability of a significant result=P_t+P_f
- PSP=$\frac{P_t}{P_t+P_f}$=$\frac{(1-\beta)*\pi}{(1-\beta)*\pi+\alpha*(1-\pi)}$
- Factors that lead to a low PSP:
  - low π
  - low $1-\beta$
  - high α

- Kanazawa's results:
  - $\alpha = 0.05$
  - $1 - \beta$=0.063 (assuming the effect, if true, is equal to 1%)
  - $\pi \rightarrow$ consider different possibilities
- Imagine π=10%, then PSP=$\frac{(1-\beta)*\pi}{(1-\beta)*\pi+\alpha*(1-\pi)}$=$\frac{0.063*0.10}{0.063*0.10+0.05*0.90}$= =$\frac{0.0063}{0.0063+0.045} = 12\%$
- Given a 10% prior, ex-post there is still only a 12% probability that beauty has an impact on the gender ratio!
- Corollary: surprising findings from underpowered studies should not move much our priors

- What happens if we have a larger power?
- For instance, with power=0.80, a significant finding increases the belief that the hypothesis is true from 10% to 64%
- Other prior probabilities (π):
  - $1\% \rightarrow 14\%$
  - $2\% \rightarrow 25\%$
  - $5\% \rightarrow 46\%$
  - $10\% \rightarrow 64\%$
  - $20\% \rightarrow 80\%$
- The posterior becomes much larger if there are successful independent replications:
  - $10\% \rightarrow 64\% \rightarrow 97\%$
- Corollary: one or two successful independent replications may be very convincing

Post-study probability

Example: Coin tossing

- Let us go back to the previous example of coin tossing.
- We are trying to detect imperfect coins (probability=0.45).
  - N=100
  - Power=18%
- Let us imagine that our prior that the probability that a given coin is imperfect is equal to 0.1%
- If we toss a coin 100 times and we get less that 41 heads (or tails), according to the frequentist approach we would reject the possibility that the coin is fair
- But if we move to the bayesian approach, what would be the posterior probability that this coin is actually imperfect?
- PSP=$\frac{(1-\beta)*\pi}{(1-\beta)*\pi+\alpha*(1-\pi)}=\frac{0.18*0.001}{0.18*0.001+0.05*0.999}=$
  $=\frac{0.00018}{0.00018+0.0499}=0.36\%$

Post-study probability

Example: Serological test

- Serology tests are blood-based tests that can be used to identify whether people have been exposed to a particular pathogen by looking at the presence of antibodies.
- The performance of these tests is described by their "sensitivity," or their ability to identify those with antibodies to SARS-CoV-2 (true positive rate), and their "specificity," or their ability to identify those without antibodies to SARS-CoV-2 (true negative rate).
- If I take a serology test that has 95% specificity rate (1-α) and a 99% sensitivity rate (1-β) and I get a positive result, how certain should I be about the possibility of being infected?

Post-study probability

Example: Serological test

- Post-study probability:
  - $\alpha = 0.05$
  - $1 - \beta$=0.99
  - $\pi \rightarrow$ consider different possibilities
- Imagine π=10%, then PSP=$\frac{(1-\beta)*\pi}{(1-\beta)*\pi+\alpha*(1-\pi)}$=$\frac{0.99*0.10}{0.99*0.10+0.05*0.90}$= =$\frac{0.099}{0.099+0.045} = 69\%$
- In a context where only 10% of people are infected, around 1/3 of people that test positive with this test are false positives

Summary: power calculations and post-study probability

- Run power calculations for your RCT
  - You want to make sure that you have enough sample size to detect the effect, if it exists
  - The power calculation requires some assumptions about the potential magnitude of the effect and the variability in the data
  - Remember that standard errors are proportional to $\frac{1}{\sqrt{n}}$
- Power calculations are also very useful if you are going to use observational data
  - collecting data may be costly...
  - ...and you do not want to end up with imprecise and uninformative estimates
- Rules of thumb to understand the power of an existing study:
  - If the standard error is equal to X, this implies that the power to detect an effect of magnitude 2*X is around 50%
- You may calculate the post-study probability (I personally like the concept, but please note that so far it is rarely used)