Testing errors and human errors

Matti Sarvimäki

Principles of Empirical Analysis
Lecture 6

Priming

- Let’s talk a few minutes about the Nature news article on priming

Priming

- Let's talk a few minutes about the Nature news article on priming
- Here is part of Daniel Kahneman's response to a blog post going through the articles he referred to in "Thinking Fast and Slow"
  - "*What the blog gets absolutely right is that I placed too much faith in* **underpowered studies**. *As pointed out in the blog, and earlier by Andrew Gelman, there is a special irony in my mistake because the first paper that Amos Tversky and I published was about the belief in the "law of small numbers," which allows researchers to trust the results of underpowered studies with unreasonably small samples. [...] Our article was written in 1969 and published in 1971, but I failed to internalize its message.*"

Significance, power and common mistakes

- Today we focus on things that often go wrong in statistical reasoning
  - again, we do this in the context of randomized experiments
  - ... but these issues are important also for other types of statistical work

Significance, power and common mistakes

- Today we focus on things that often go wrong in statistical reasoning
  - again, we do this in the context of randomized experiments
  - ... but these issues are important also for other types of statistical work
- Learning objectives. You will understand the following concepts:
  1. false positives and negatives (a.k.a. type I and II errors)
  2. multiple hypothesis problem
  3. publication bias, file-drawer effect and p-hacking
  4. pre-registration and replication files
  5. power
  6. minimum detectable effect size

  and become able to use them to interpret basic empirical results

Testing errors

		Reality	
		Effect	No effect
Result of an experiment	Effect	True positive	**False positive**
	No efect	**False negative**	True negative

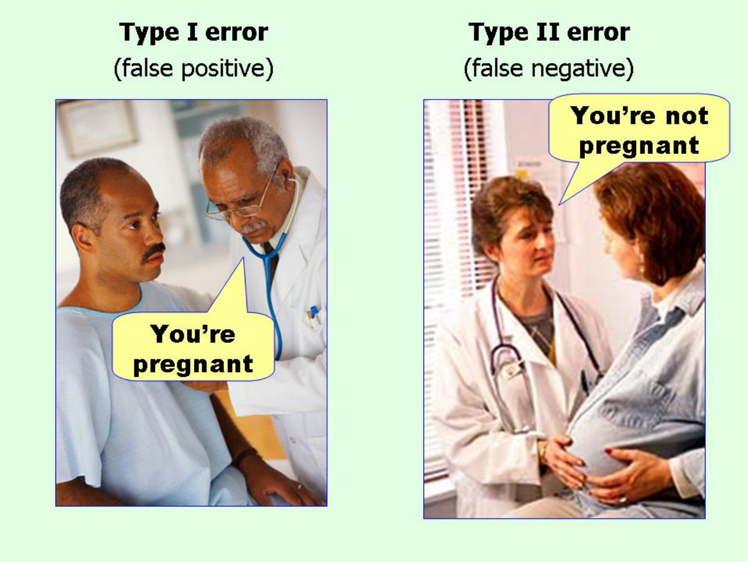
- False positive: Claiming an effect when it does not exist
  - also known as "type I error" or "acceptance error"

Testing errors

		Reality	
		Effect	No effect
Result of an experiment	Effect	True positive	**False positive**
	No efect	**False negative**	True negative

- False positive: Claiming an effect when it does not exist
  - also known as ”type I error” or ”acceptance error”
- False negative: Not finding an effect when it does exist
  - a.k.a. ”type II error” or ”rejection error”
- Power: the probability of finding an effect when it exists

Testing errors



Source: Effect size FAQs

Statistical significance and testing errors

- Statistical significance testing is build to avoid false positives
  - we typically call estimates ”statistically significant” if $p < .05$
  - i.e. if there was no effect, differences as extreme as the one we observed between treated/control would occur less than 1 out of 20 times
- Trade off between false positives and false negatives
  - efforts to reduce one type of error increase the other type of error

Statistical significance and testing errors

- The convention of dividing results to "statistically significicant" and "statistically insignificant" often leads to severe misunderstandings
  - treatment is thought to have been "proven to be effective" when $p < .05$ or "proven to have no effect" when $p > .05$

Statistical significance and testing errors

- The convention of dividing results to "statistically significicant" and "statistically insignificant" often leads to severe misunderstandings
    - treatment is thought to have been "proven to be effective" when $p < .05$ or "proven to have no effect" when $p > .05$
- The prevalence of such misconceptions has led to demands for abandoning the whole concept of statistical significance
    - even if this would eventually happen, you will have to understand and interpret lots of research where statistical significance is used

Statistical significance and testing errors

- The convention of dividing results to "statistically significicant" and "statistically insignificant" often leads to severe misunderstandings
  - treatment is thought to have been "proven to be effective" when $p < .05$ or "proven to have no effect" when $p > .05$
- The prevalence of such misconceptions has led to demands for abandoning the whole concept of statistical significance
  - even if this would eventually happen, you will have to understand and interpret lots of research where statistical significance is used
- No-one demands abandoning p-values and confidence intervals!
  - rather, the debate is about the misleading and unnecessary dichotomy between "significant" and "insignificant" results

A simulation exercise

- Let's illustrate these issues with the following simulation using one year of the FLEED teaching data
  1. draw a random sample of n persons

A simulation exercise

- Let's illustrate these issues with the following simulation using one year of the FLEED teaching data
  1. draw a random sample of n persons
  2. assign half of the sample into treatment and half into control groups

A simulation exercise

- Let's illustrate these issues with the following simulation using one year of the FLEED teaching data
  1. draw a random sample of n persons
  2. assign half of the sample into treatment and half into control groups
  3. replace everyone's income in the treatment group with $y_i + \beta$, where y_i is individual i's true income and β is the simulated treatment effect

A simulation exercise

- Let's illustrate these issues with the following simulation using one year of the FLEED teaching data
  1. draw a random sample of n persons
  2. assign half of the sample into treatment and half into control groups
  3. replace everyone's income in the treatment group with $y_i + \beta$, where y_i is individual i's true income and β is the simulated treatment effect
  4. calculate difference in average income between treatment and control groups and test for its statistical signficance

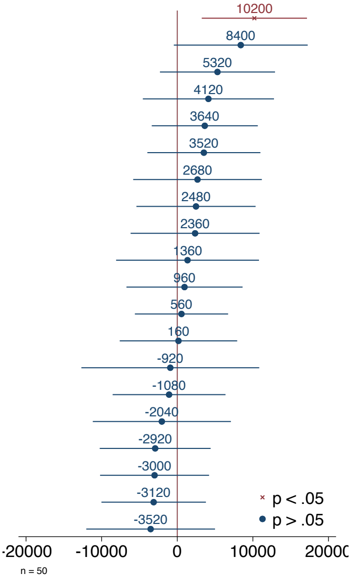
A simulation exercise

- Let's illustrate these issues with the following simulation using one year of the FLEED teaching data
  1. draw a random sample of n persons
  2. assign half of the sample into treatment and half into control groups
  3. replace everyone's income in the treatment group with $y_i + \beta$, where y_i is individual i's true income and β is the simulated treatment effect
  4. calculate difference in average income between treatment and control groups and test for its statistical signficance
  5. repeat many times and summarize the results

A simulation exercise

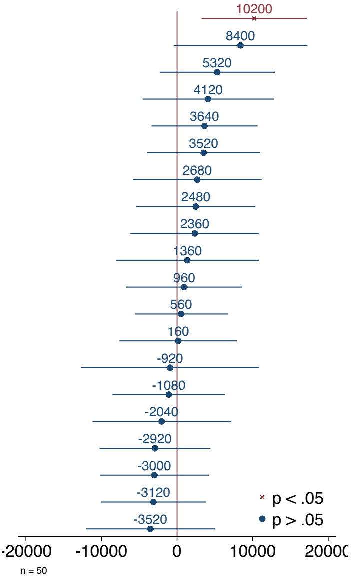
- Let's illustrate these issues with the following simulation using one year of the FLEED teaching data
  1. draw a random sample of n persons
  2. assign half of the sample into treatment and half into control groups
  3. replace everyone's income in the treatment group with $y_i + \beta$, where y_i is individual i's true income and β is the simulated treatment effect
  4. calculate difference in average income between treatment and control groups and test for its statistical signficance
  5. repeat many times and summarize the results
- Let's start with the case where the treatment has no impact ($\beta = 0$)
  - question: among the false positives, how should we expect the estimated size of the effect to vary with sample size?

False positives in small samples



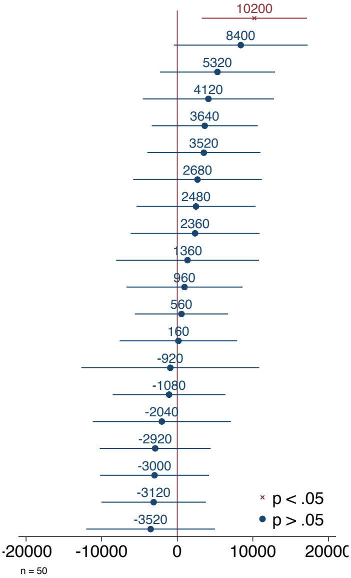
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control

False positives in small samples



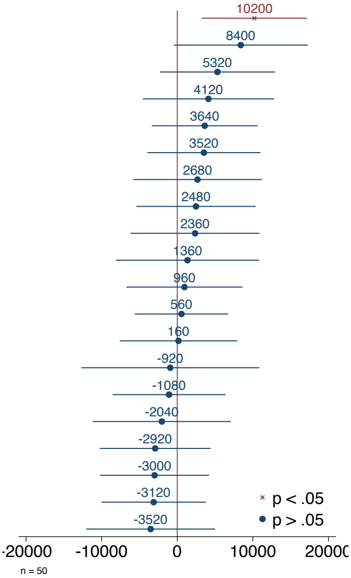
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control
- 1 out of 20 is a false positive
  - exactly what one should expect when using $p < .05$ as the criterion for significance

False positives in small samples



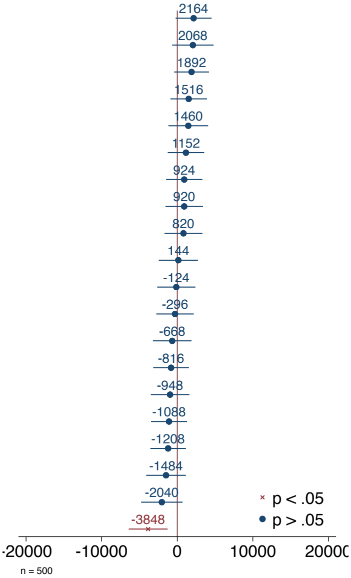
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control
- 1 out of 20 is a false positive
  - exactly what one should expect when using $p < .05$ as the criterion for significance
- By construction, the point estimate for the false positive is spectacularly large
  - given such large standard errors, it *has* to be large in order to be significant!
  - the false positive result suggests that this "treatment" increased income by 10,200 euros or 0.7 standard deviations

False positives in small samples



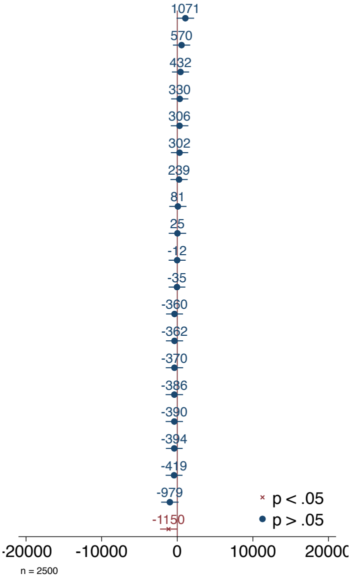
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control
- 1 out of 20 is a false positive
  - exactly what one should expect when using $p < .05$ as the criterion for significance
- By construction, the point estimate for the false positive is spectacularly large
  - given such large standard errors, it *has* to be large in order to be significant!
  - the false positive result suggests that this "treatment" increased income by 10,200 euros or 0.7 standard deviations
- All confidence intervals include large effects
  - 95%CI average width is 16,000 euros!

False positives with larger samples



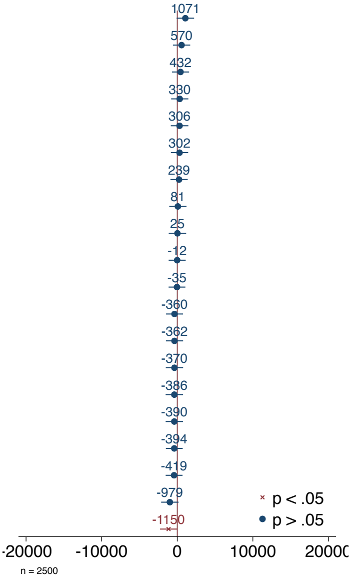
- 20 simulations with $n = 500$
  - again, one happens to be a false positive
- Now, the point estimate for the false positive is less spectacular
  - none of the estimates is close to 10,000
  - CI average width is 5,000 euros

False positives with larger samples

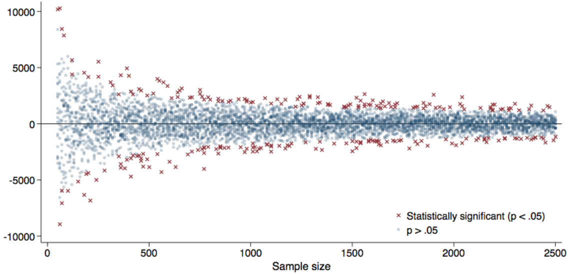


- 20 simulations with $n = 2500$
  - even less spectacular false positive
  - and still tighter confidence intervals (CI average width is 2,300 euros)

False positives with larger samples



- 20 simulations with $n = 2500$
  - even less spectacular false positive
  - and still tighter confidence intervals (CI average width is 2,300 euros)
- More simulations
  - 20 rounds for 50,60,....,2500 observations
  - 0–5 false positives per round
  - overall 5.2% of simulations false positive



Take-aways from the first simulation

- The likelihood of a false positive does not vary with sample size
  - by definition, depends only on the p-value required for calling the esimate statistically significant (significance level)

Take-aways from the first simulation

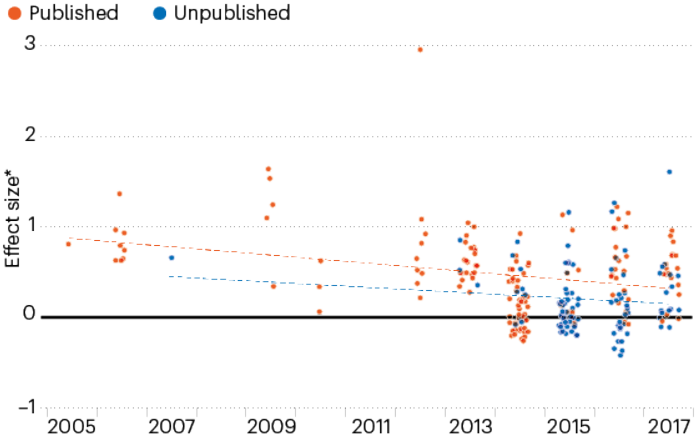
- The likelihood of a false positive does not vary with sample size
  - by definition, depends only on the p-value required for calling the esimate statistically significant (significance level)
- Small samples lead to large point estimates for false positives
  - small sample → wide CI → only large estimates significant
  - thus false positives from small samples may cause more damage
    - policy mistakes more likely if the effects are believed to be large
    - sadly, few people understand the dangers of underpowered studies

Take-aways from the first simulation

- The likelihood of a false positive does not vary with sample size
    - by definition, depends only on the p-value required for calling the esimate statistically significant (significance level)
- Small samples lead to large point estimates for false positives
    - small sample → wide CI → only large estimates significant
    - thus false positives from small samples may cause more damage
        - policy mistakes more likely if the effects are believed to be large
        - sadly, few people understand the dangers of underpowered studies
    - results from small samples sometimes get huge media attention
        - unfortunately, editors and referees of scientific journals may also like spectacular and statistically signficant results

WANING EFFECT

A meta-analysis of 246 experiments that exposed people to money-related stimuli found that early studies reported larger priming effects on behaviour, emotions and attitudes than did later ones. It also revealed larger effects in published work than in unpublished experiments provided by authors of the original studies.



*Effect size measured by a value known as Hedges' g, where '1' indicates that primed and control groups differed by 1 standard deviation.

Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments

Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"

Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"
  - **file-drawer effect**: researchers never finish papers with statistically insignificant results, because they would not be published anyways
    - less likely in large RCTs (funding agencies require to publish something)

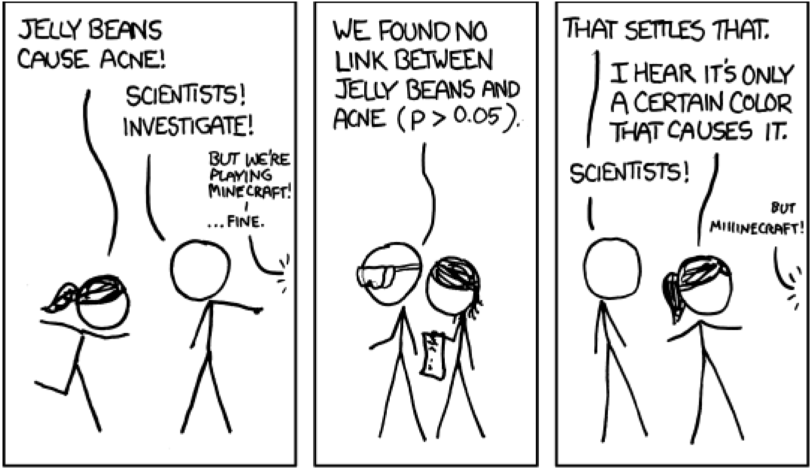
Publication bias, file-drawer effect and p-hacking

- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"
  - **file-drawer effect**: researchers never finish papers with statistically insignificant results, because they would not be published anyways
    - less likely in large RCTs (funding agencies require to publish something)
  - **p-hacking**: researcher reports only a specification with $p < .05$

Publication bias, file-drawer effect and p-hacking

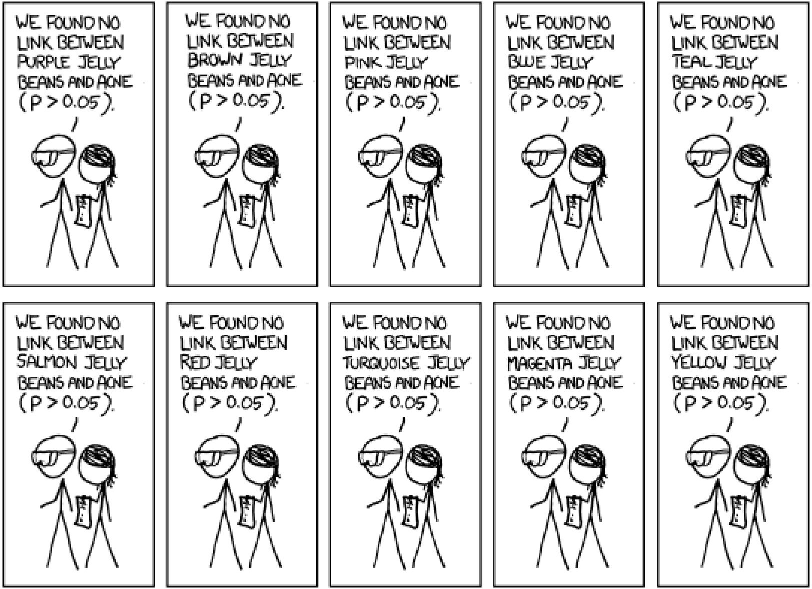
- For treatments with no impact, we should expect to see 5% significance for every 20th experiment
  - we can take this into account if we see results from all experiments
- The problem is that we may get to see only the "significant" ones
  - **publication bias**: academic journals may be more likely to publish statistically significant results than insignificant "imprecise zeros"
  - **file-drawer effect**: researchers never finish papers with statistically insignificant results, because they would not be published anyways
    - ▶ less likely in large RCTs (funding agencies require to publish something)
  - **p-hacking**: researcher reports only a specification with $p < .05$
- No-one needs to be neferious for these problems to arise
  - people who farbricate results rarely want to be researchers
  - but: honest researchers may "follow the data" into wrong conclusions

Multiple comparisons problem

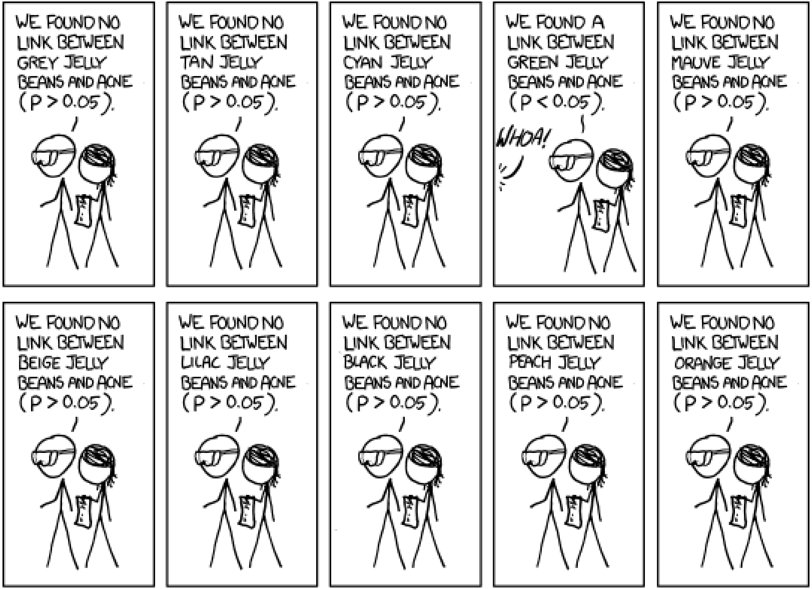

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

xkcd 882

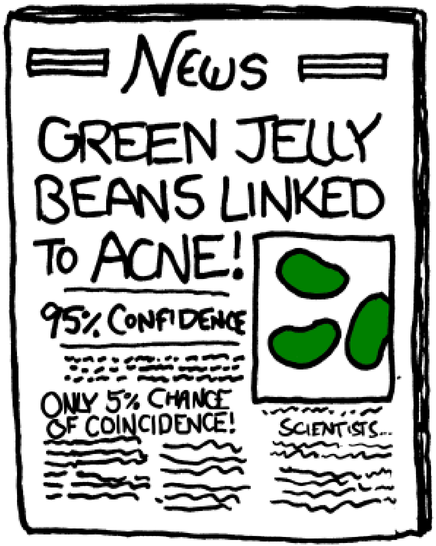
Multiple comparisons problem


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

Multiple comparisons problem


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Multiple comparisons problem


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

- **Multiple comparisons problem** occurs when many comparisons are performed, but this is not taken into account in hypothesis testing
- A *human* error that can happen even with the best intentions
  - "the Garden of Forking Paths"
  - can take also other forms (e.g. subsample analysis)
- Tests taking into account the number of comparisons exist
  - you'll learn some of them in the more advanced courses

Approaches for human mistakes

- Pre-registration of RCTs
  - researchers can ”tie their hands” by documenting their primary outcomes and specifications before seeing the data
  - long tradition in medicine; now also required in economics
- Replication files
  - top economics journals require researchers to post their code and data (or details about accessing the data) of published papers
  - allows other researchers to analyze the robustness of the results
- Running larger experiments

RCTs to Scale: Comprehensive Evidence from Two Nudge Units*

Stefano DellaVigna
UC Berkeley and NBER

Elizabeth Linos
UC Berkeley

April 2021

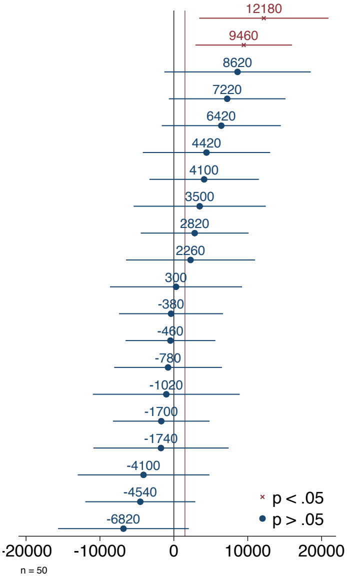
Abstract

Nudge interventions have quickly expanded from academic studies to larger implementation in so-called Nudge Units in governments. This provides an opportunity to compare interventions in research studies, versus at scale. We assemble a unique data set of 126 RCTs covering 23 million individuals, including all trials run by two of the largest Nudge Units in the United States. We compare these trials to a sample of nudge trials in academic journals from two recent meta-analyses. In the Academic Journals papers, the average impact of a nudge is very large—an 8.7 percentage point take-up effect, which is a 33.4% increase over the average control. In the Nudge Units sample, the average impact is still sizable and highly statistically significant, but smaller at 1.4 percentage points, an 8.0% increase. We document three dimensions which can account for the difference between these two estimates: (i) statistical power of the trials; (ii) characteristics of the interventions, such as topic area and behavioral channel; and (iii) selective publication. A meta-analysis model incorporating these dimensions indicates that selective publication in the Academic Journals sample, exacerbated by low statistical power, explains about 70 percent of the difference in effect sizes between the two samples. Different nudge characteristics account for most of the residual difference.

False negatives

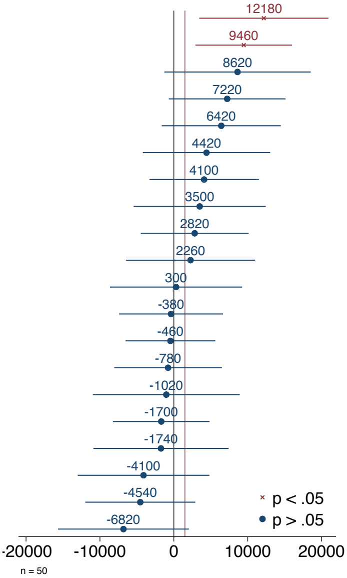
- Statistical error of not detecting an effect when it exists
  - getting $p > .05$ when there is an effect
- Let's demonstrate this with another simulation
  - identical to the one before except that now the treatment increase annual income of the treated by 1,500 euros

False negatives in small samples



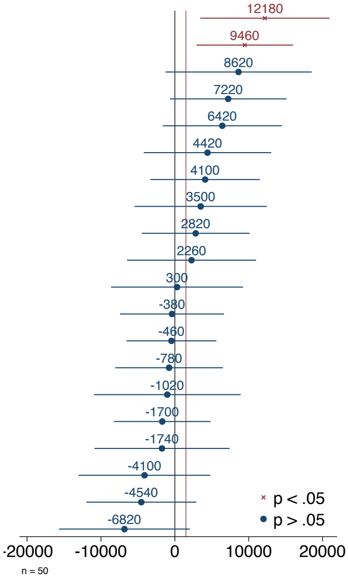
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control

False negatives in small samples



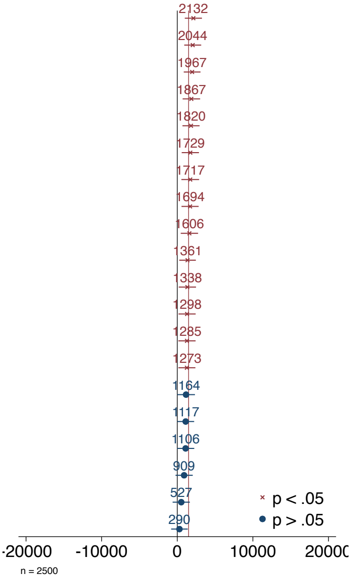
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control
- 2 out of 20 is statistically significant
  - but they are also wrong in the sense of being 6–8 times larger than the truth!

False negatives in small samples



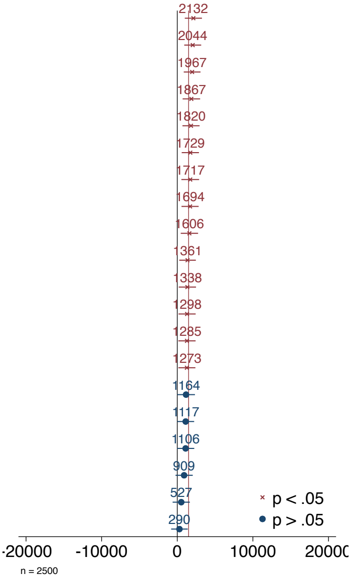
- Here are 20 simulations with $n = 50$
  - 25 persons in treatment, 25 in control
- 2 out of 20 is statistically significant
  - but they are also wrong in the sense of being 6–8 times larger than the truth!
- 18 out of 20 are false negatives
  - 5 some of them are larger with the wrong sign than the true effect!
- Take-away: these estimates contain very little information

False negatives with larger samples

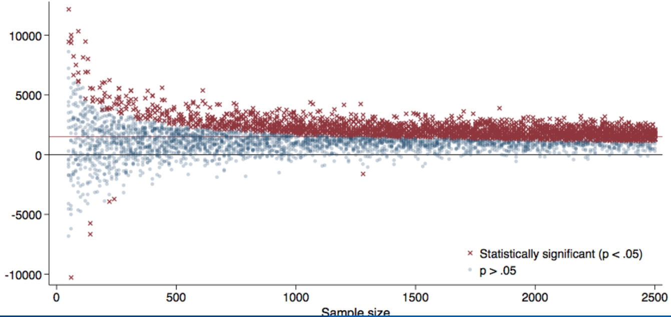


- 20 simulations with $n = 2500$
  - 12 out of 20 statistically significant
  - all relatively close to to the truth

False negatives with larger samples



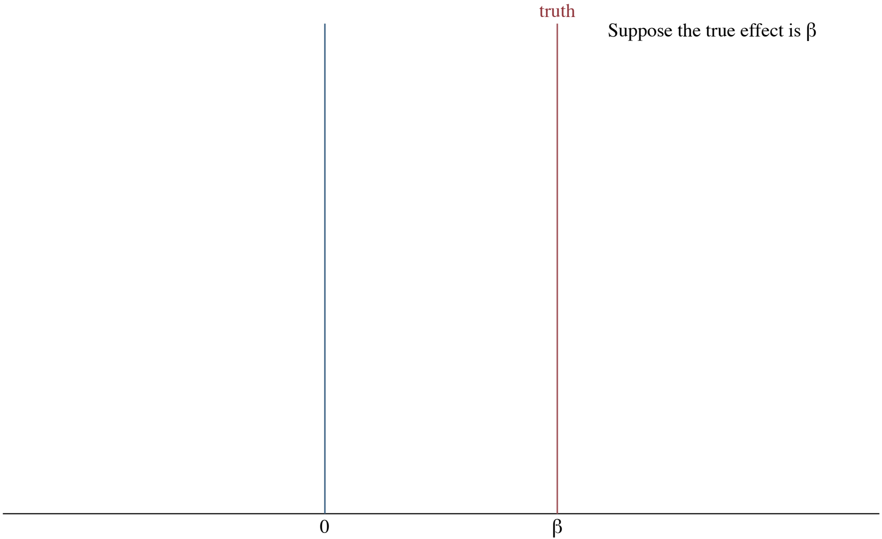
- 20 simulations with $n = 2500$
  - 12 out of 20 statistically significant
  - all relatively close to to the truth
- More simulations
  - 20 rounds for 50,60,....,2500 observations
  - as n increases, share of false negatives and wild point estimates decrease



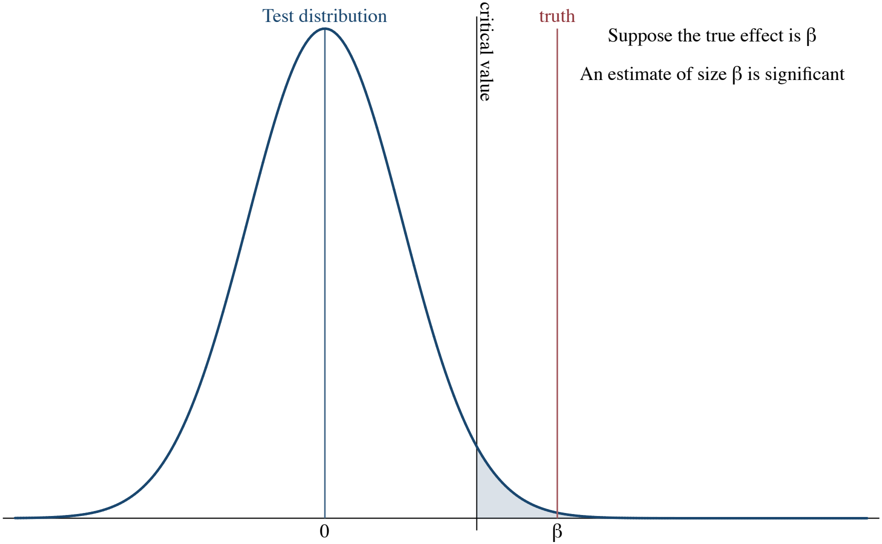
Power

- **Power** $= Pr(\text{reject } H_0 | H_1 \text{ is true})$
  - in our context: how likely are we to conclude that a treatment has an impact, when it truly has an impact
- Power depends on
  - true effect size
  - sample size
  - variability of the outcome variable
  - statistical significance level
- Next: a graphical illustration of power

Power



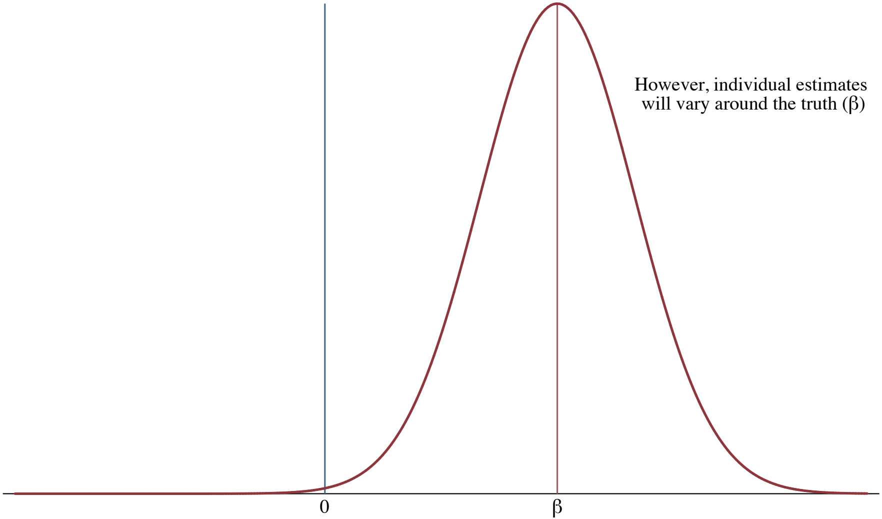
Power



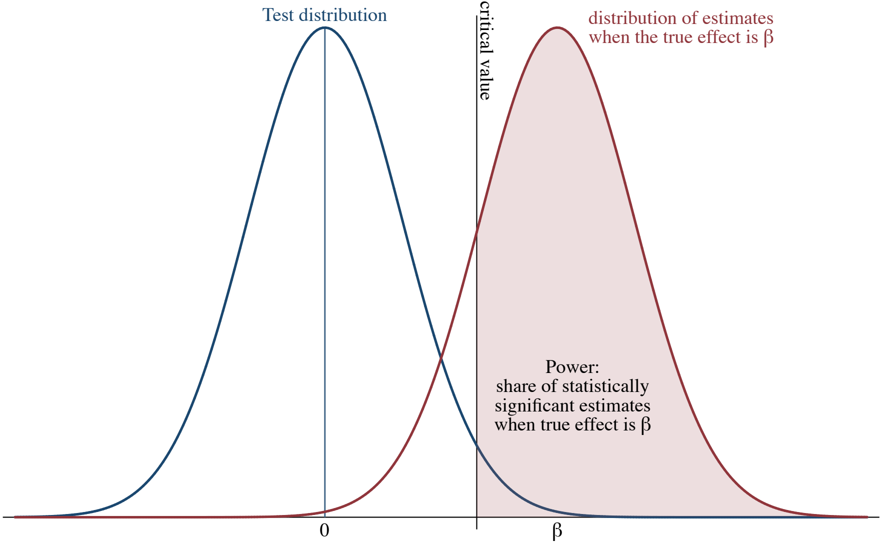
Suppose the true effect is β

An estimate of size β is significant

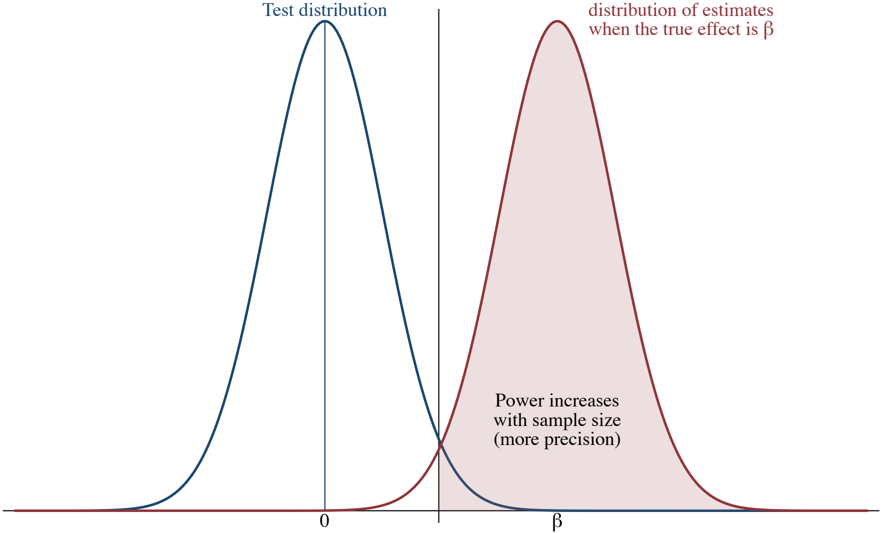
Power


However, individual estimates
will vary around the truth (β)
0
β

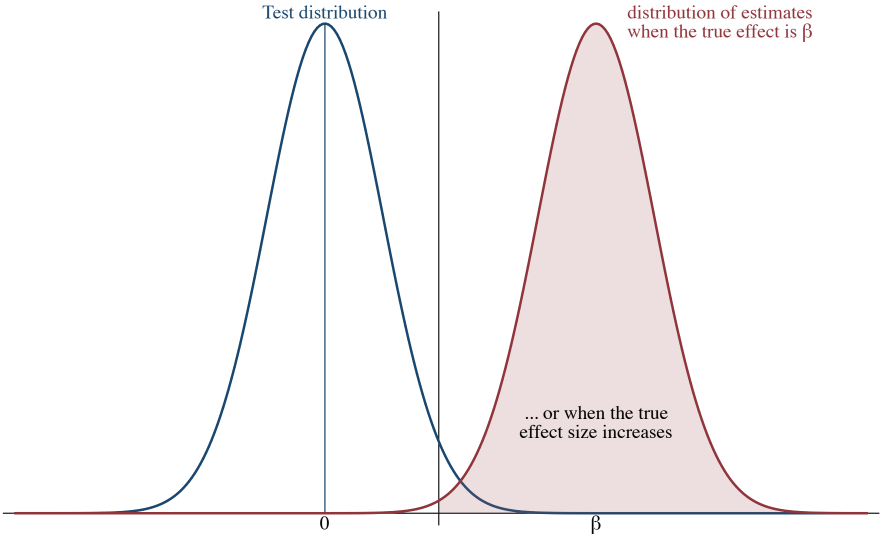
Power


Test distribution
critical value
distribution of estimates
when the true effect is β
Power:
share of statistically
significant estimates
when true effect is β
0
β

Power


Test distribution
distribution of estimates
when the true effect is β
Power increases
with sample size
(more precision)
0
β

Power


Test distribution
distribution of estimates
when the true effect is β
... or when the true
effect size increases
0
β

Minimum detectable effect size (MDE)

- Often helpful to ask: How large would the true effect need to be in order for us to have sufficient power?
  - "sufficient" typically defined as 80% power with 5% for significance
  - but, again, this is just a convention

Minimum detectable effect size (MDE)

- Often helpful to ask: How large would the true effect need to be in order for us to have sufficient power?
  - "sufficient" typically defined as 80% power with 5% for significance
  - but, again, this is just a convention
- This **minimum detectable effect size** is given by

$$MDE = (t_{(1-\kappa)} + t_\alpha) \times \sqrt{\frac{1}{P(1-P)} \frac{\sigma^2}{n}}$$

  - $t_{(1-\kappa)}$ is a critical value for power (0.84 for 80% power)
  - t_α is the critical value for signifiance (1.96 for 5% significance)
  - P is the share of sample assigned to the treatment group
  - σ^2 is the variance of the outcome variable
  - n is sample size

Minimum detectable effect size (MDE)

- To make sense of this, note that

$$\sqrt{\frac{1}{P(1-P)}\frac{\sigma^2}{n}} = S(y_i)\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}$$

i.e estimator for **standard error** that we used in the previous lecture

Minimum detectable effect size (MDE)

- To make sense of this, note that

$$\sqrt{\frac{1}{P(1-P)}\frac{\sigma^2}{n}} = S(y_i)\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}$$

  i.e estimator for **standard error** that we used in the previous lecture

- How to get from one expression to the other?
  1. $\sqrt{\sigma^2} = S(y_i)$ (just different notation in different sources)

Minimum detectable effect size (MDE)

- To make sense of this, note that

$$\sqrt{\frac{1}{P(1-P)}\frac{\sigma^2}{n}} = S(y_i)\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}$$

  i.e estimator for **standard error** that we used in the previous lecture

- How to get from one expression to the other?
  1. $\sqrt{\sigma^2} = S(y_i)$ (just different notation in different sources)
  2. n observations in the full sample, n_1 observations in the treatment group, n_0 observations in the control group, and P is the share of the sample allocated to the treatment group. Thus:

$$\frac{1}{n_1}+\frac{1}{n_0}=\frac{1}{Pn}+\frac{1}{(1-P)n}=\frac{1-P}{P(1-P)n}+\frac{P}{P(1-P)n}=\frac{1}{P(1-P)n}$$

Minimum detectable effect size (MDE)

- Helpful rule-of thumb:

$$MDE \approx 2.8 \times \hat{SE}$$

for 80% power and 5% significance.

Minimum detectable effect size (MDE)

- Helpful rule-of thumb:

$$MDE \approx 2.8 \times \hat{SE}$$

  for 80% power and 5% significance.
- "How large would the true effect have to be in order for there to be a reasonable chance of finding a statistically significant effect?"
  - you only need to know the standard error to answer this!
  - remembering this rule-of-thumb will reveal many misleading statements of the form "we have shown that X does not affect Y"

Minimum detectable effect size (MDE)

- Helpful rule-of thumb:

$$MDE \approx 2.8 \times \hat{SE}$$

  for 80% power and 5% significance.
- "How large would the true effect have to be in order for there to be a reasonable chance of finding a statistically significant effect?"
  - you only need to know the standard error to answer this!
  - remembering this rule-of-thumb will reveal many misleading statements of the form "we have shown that X does not affect Y"
- Always ask: "Can we rule out an **economically significant** effect?"

Minimum detectable effect size (MDE)

- Take-aways from the MDE formula

$$MDE = (t_{(1-\kappa)} + t_{\alpha}) \times \sqrt{\frac{1}{P(1-P)} \frac{\sigma^2}{n}}$$

- MDE is smaller when
  - the experiment has more participants (larger n)
  - outcome variable is less variable (smaller σ^2)
  - P is closer to 50%

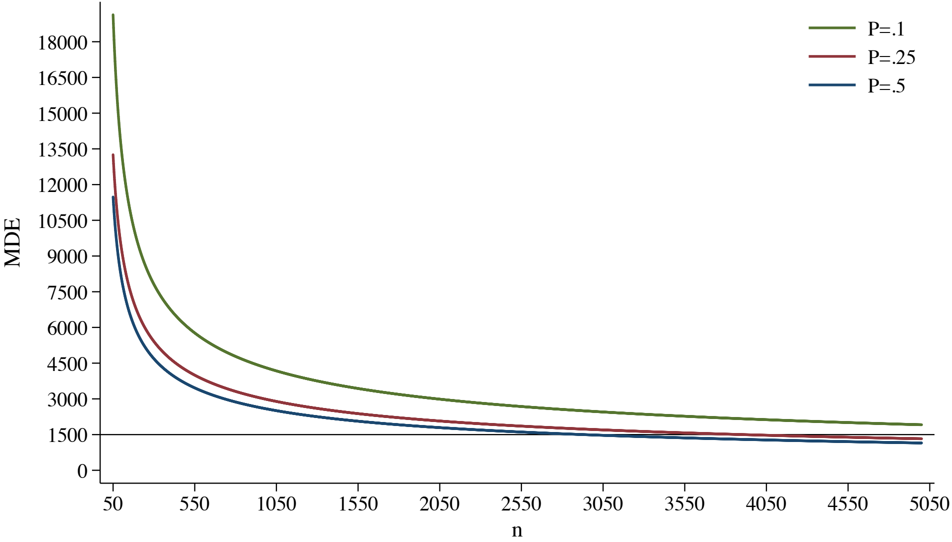
Minimum detectable effect size (MDE)

- Take-aways from the MDE formula

$$MDE = (t_{(1-\kappa)} + t_\alpha) \times \sqrt{\frac{1}{P(1-P)}\frac{\sigma^2}{n}}$$

- MDE is smaller when
  - the experiment has more participants (larger n)
  - outcome variable is less variable (smaller σ^2)
  - P is closer to 50%
- MDE formula also implicitly answers: "How large an experiment do we need, in order to be able to detect an effect of a certain size?"
  - note that the SE estimator used here is based on specific assumptions
  - often you need to relax those assumption and use other SE estimators (discussed in later courses)

MDE by n and P for our simulation example



Summary

- Today, we discussed two kinds of errors
  - statistical: well-defined properties of statistical tests
  - human: messy reality of how people (mis)use/interpret statistics

Summary

- Today, we discussed two kinds of errors
    - statistical: well-defined properties of statistical tests
    - human: messy reality of how people (mis)use/interpret statistics
- Key concepts to understand
    - false negative, false positive
    - power, minimum detectable effect size

Summary

- Today, we discussed two kinds of errors
  - statistical: well-defined properties of statistical tests
  - human: messy reality of how people (mis)use/interpret statistics
- Key concepts to understand
  - false negative, false positive
  - power, minimum detectable effect size
- Ways to avoid human errors
  - being alert and suspicious (particularly regarding your own results)
  - tying one's hands: pre-registration, replication, machine learning...