Eureka! I’ve Discovered C Major!

Replicating a Finding
Issues and Music and Sciences
Dr. David John Baker
HU Berlin, Winter 2020

Outline

I. The world and the data we collect is messy

II. We need tools to be able to discern if we’re looking at noise or something real

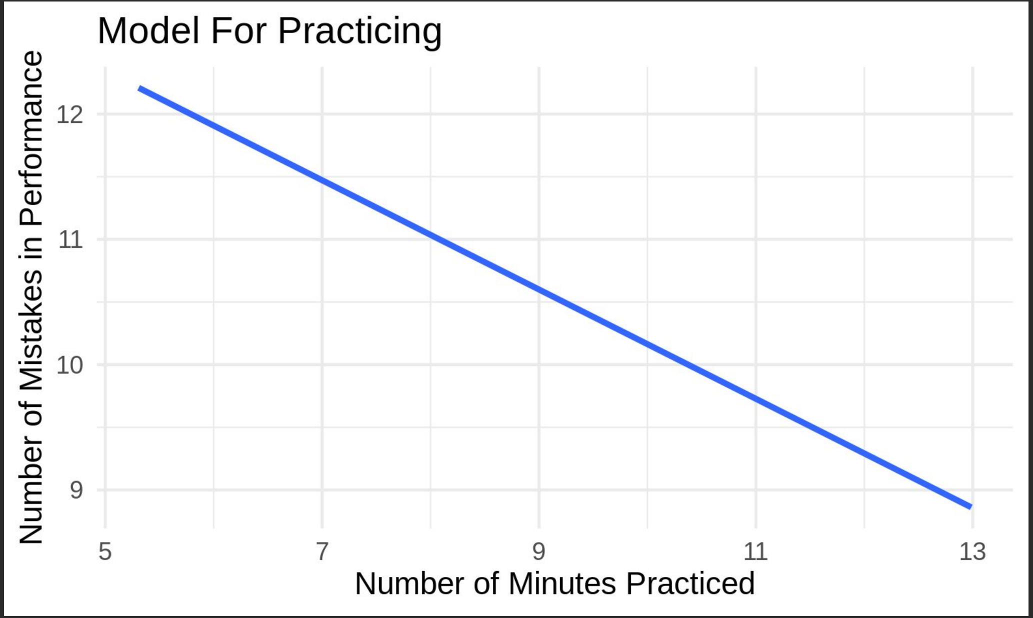
III. The tools of statistics are used and abused to help us quantify how much uncertainty we’re looking at when we look at data

Regression Review

Regression

→ We work with models to make problems easier to understand
→ PROBLEM: The world is not deterministic (same thing happens every time)
→ Need ability to be able to quantify how much uncertainty we have in what we observe

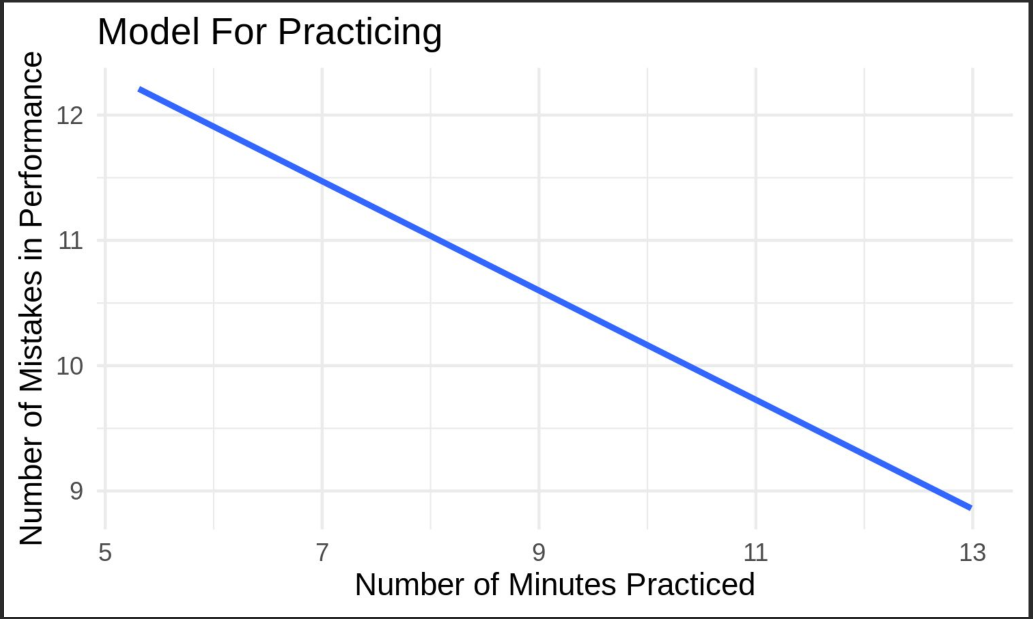
Practicing vs Errors Example
Errors = Minutes Practice * Beta



Mistakes = -0.43 * minutes + 14.5

→ Makes explicit assumption about how practicing works
→ States what direction we think the relationship will occur
→ Quantifies the magnitude of the relationship (how big = effect size)

If we collect data, how do we know if we have found something?



Mistakes = -0.43 * minutes + 14.5

→ Makes explicit assumption about how practicing works
→ States what direction we think the relationship will occur
→ Quantifies the magnitude of the relationship (how big = effect size)

Learning Goals

- How do we decide if we actually discover real relationships when we collect data?

- What could go wrong in this process!?

The Plan

→ Walk through of experiment

→ Note new terms as we encounter them

→ Highlight what might go wrong along the way

→ Notice points where to be skeptical when assessing data

SCENARIO

You are a music psychologist who is interested music learning. You have read some papers about the effects of practice in learning a new language and want to see if you can find a similar relationship in music. To do this, you plan to do an experiment where people are told to practice a piece of music on the piano, record how long they practice throughout the course of a week, then asses how many errors they make at the end of a week in a performance in your lab.

Let’s try to frame our research question in light of a larger scientific question...

Our Program of Research (Lecture 2)

- Framework → Phenomena we want to describe
- Theory → set of causal relations that attempt to describe, explain, or make predictions about the world
- Specification → (ignore for now)
- Implementation → (ignore for now)
- Hypothesis → A narrow, testable statement
- Data → Observations collected from the real world

Our Program of Research (Lecture 2)

- Framework → Phenomena we want to describe
    - ???

Our Program of Research (Lecture 2)

- Framework → Phenomena we want to describe
  - Learning
  - Specifically, music learning
  - Performance

Our Program of Research (Lecture 2)

- Theory → set of causal relations that attempt to describe, explain, or make predictions about the world
  - ???

Our Program of Research (Lecture 2)

- Theory → set of causal relations that attempt to describe, explain, or make predictions about the world
    - Hebbian Learning
    - “Fire together, wire together”
    - Create efficient neural pathways through repetition
    - Like a sled going down a fresh, new snow
    - Assume that this learning isn’t dependent on subject
    - Works with language, works with music

Our Program of Research (Lecture 2)

- Hypothesis → A narrow, testable statement
  - Weak ??
  - Strong ??

Our Program of Research (Lecture 2)

- Hypothesis → A narrow, testable statement
  - Weak : As time goes up, errors go down
  - Strong: Y = -0.43 * minutes + 14.5
    - Get STRONG hypothesis from previous research
    - Usually we’re not the first to ask a question...

Our Program of Research (Lecture 2)

- Data → Observations collected from the real world
    - ???

Our Program of Research (Lecture 2)

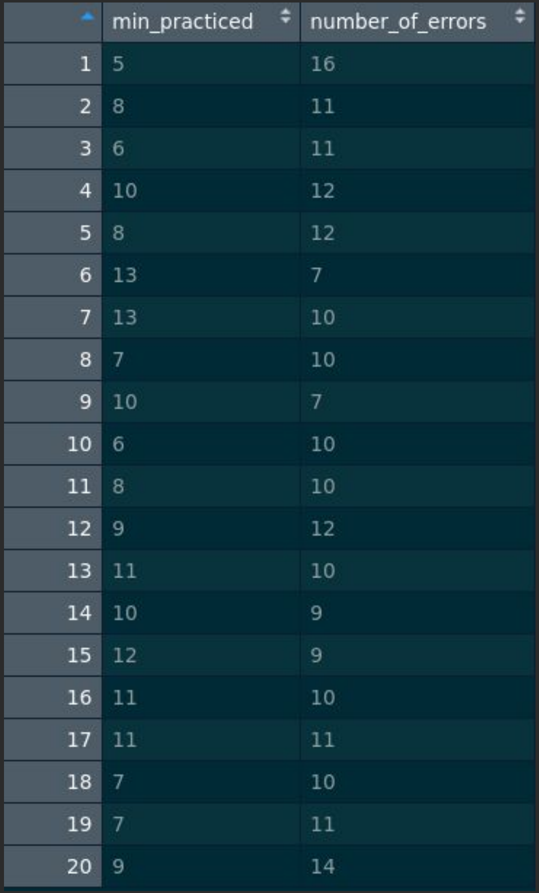
- Data → Observations collected from the real world
  - Create from design of experiment
  - Assume time will affect number of errors
  - Give piano piece to people
  - Ask to document number of minutes practiced
  - Count number of mistakes they make
    - Create list ahead of time ?
    - Two people judge and look for inter-rater reliability?

Our Program of Research (Lecture 2)

- Data → Observations collected from the real world
  - Minutes Practiced → INDEPENDENT VARIABLE
  - Number of Errors → DEPENDANT VARIABLE
- Independent Variable
  - What we think we can manipulate
- Dependent Variable
  - What we think will change based on manipulations

Plan

- Recruit 20 people
- Ask them to practice first 20 measures of a piece of music
- Participants asked to record themselves practicing
- Number of minutes of video used as Independent Variable
- Participants come into lab at end of week to perform their piece
- We count the number of errors they make in their performance
- Put all the data into a spreadsheet!

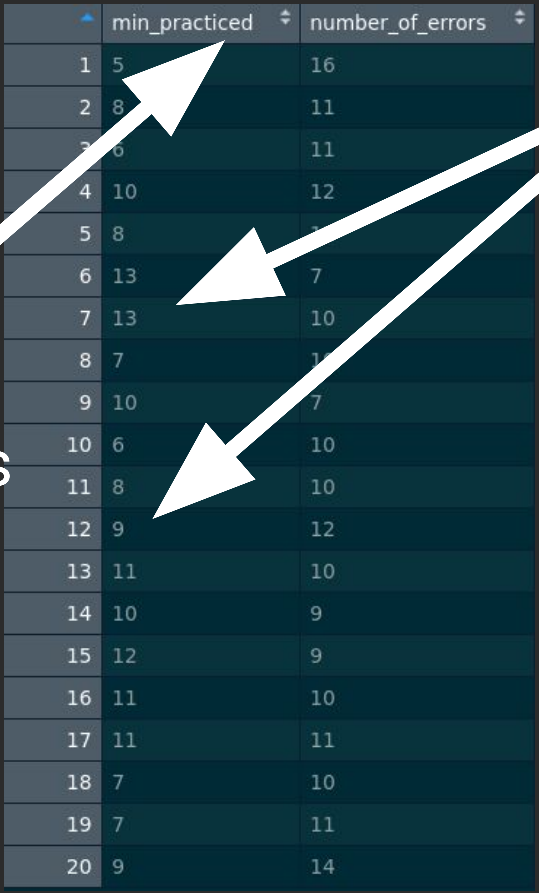


	min_practiced	number_of_errors
1	5	16
2	8	11
3	6	11
4	10	12
5	8	12
6	13	7
7	13	10
8	7	10
9	10	7
10	6	10
11	8	10
12	9	12
13	11	10
14	10	9
15	12	9
16	11	10
17	11	11
18	7	10
19	7	11
20	9	14

THE DATA

Each observation (participant) is a row

Each variable is a column

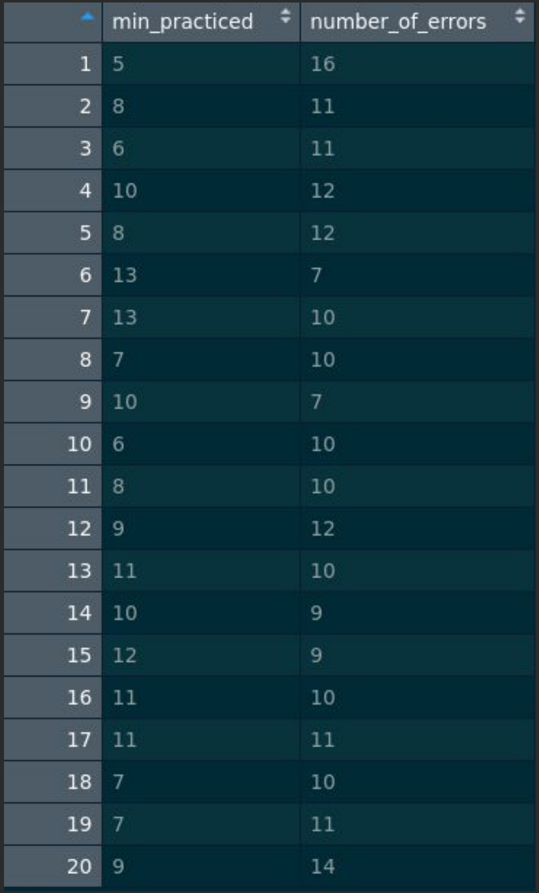


	min_practiced	number_of_errors
1	5	16
2	8	11
3	6	11
4	10	12
5	8	[illegible]
6	13	7
7	13	10
8	7	10
9	10	7
10	6	10
11	8	10
12	9	12
13	11	10
14	10	9
15	12	9
16	11	10
17	11	11
18	7	10
19	7	11
20	9	14

Only one type of data per cell!

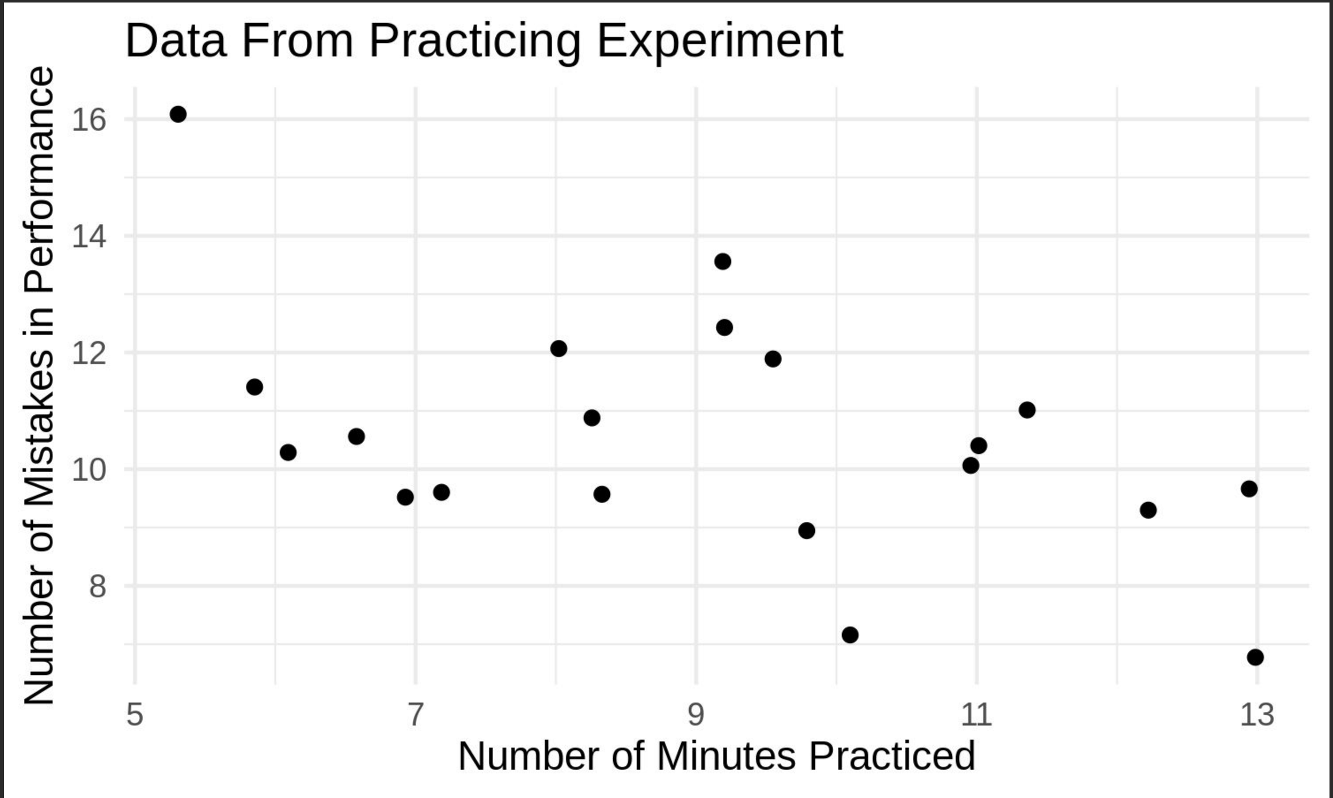
THE (Tidy) DATA

THE DATA



	min_practiced	number_of_errors
1	5	16
2	8	11
3	6	11
4	10	12
5	8	12
6	13	7
7	13	10
8	7	10
9	10	7
10	6	10
11	8	10
12	9	12
13	11	10
14	10	9
15	12	9
16	11	10
17	11	11
18	7	10
19	7	11
20	9	14

Do we have evidence of learning !?

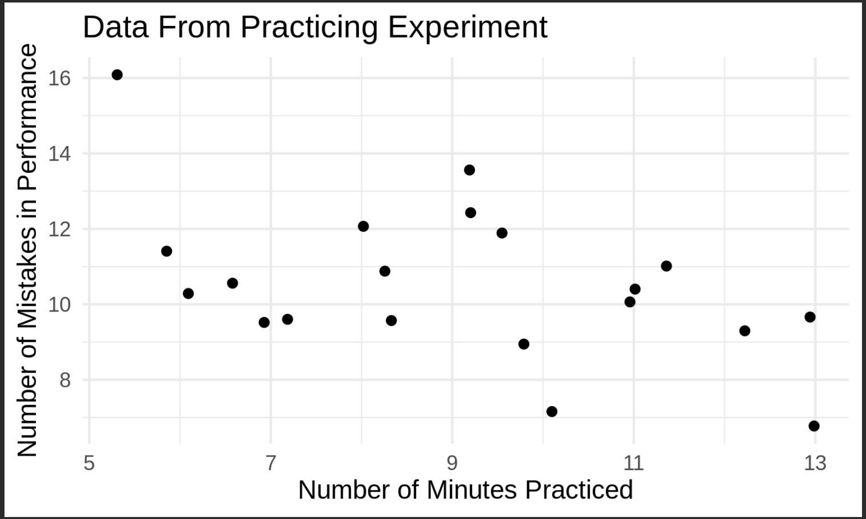


Each participant in our experiment is represented by a point

As number of minutes practiced goes up, what happens to the number of mistakes?

Is there evidence here of learning??

How do we know this pattern of data didn’t happen by chance?



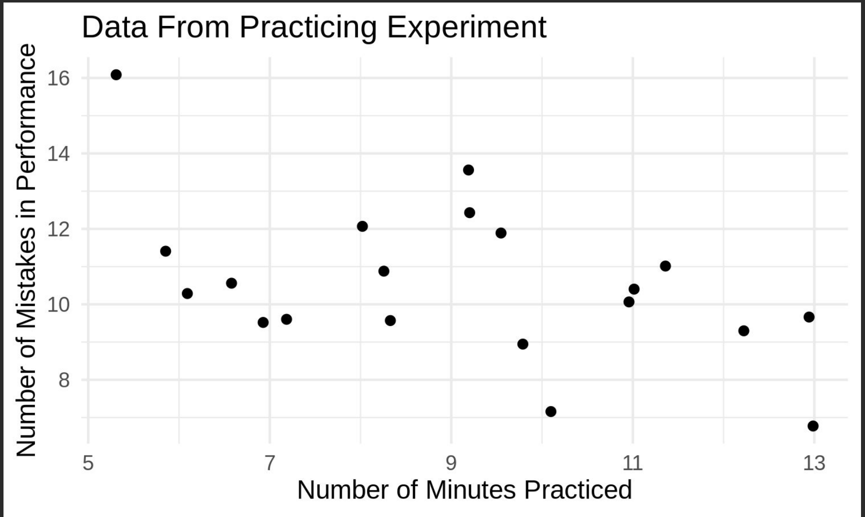
→ We need a way to quantify that something IS happening here

→ This is where statistics comes in

→ Few different ways of thinking about how statistics works

- Null Hypothesis Significance Testing (p values)
- Bayesian Statistics
- Likelihood

→ Don’t worry about them for now, going to focus on NHST (the world of p values)



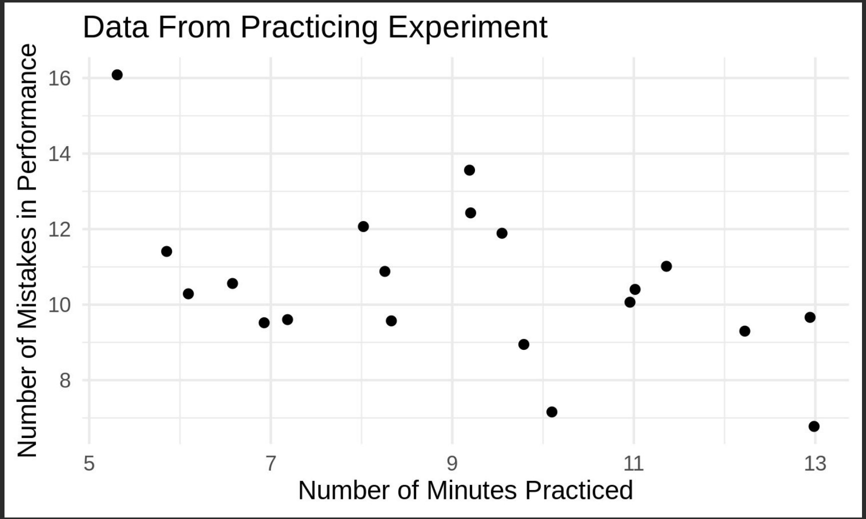
PROBLEM!!

→ We want to be able to say something EXISTS, but this runs into problem of induction!

- What if we got data where most people made no errors with little practice?
- Or had people who practiced a lot and never got better…?

→ Going to flip the problem on its head

→ Exploit asymmetry in that it's easier to take down a theory rather than accumulate evidence for one



PROBLEM!!

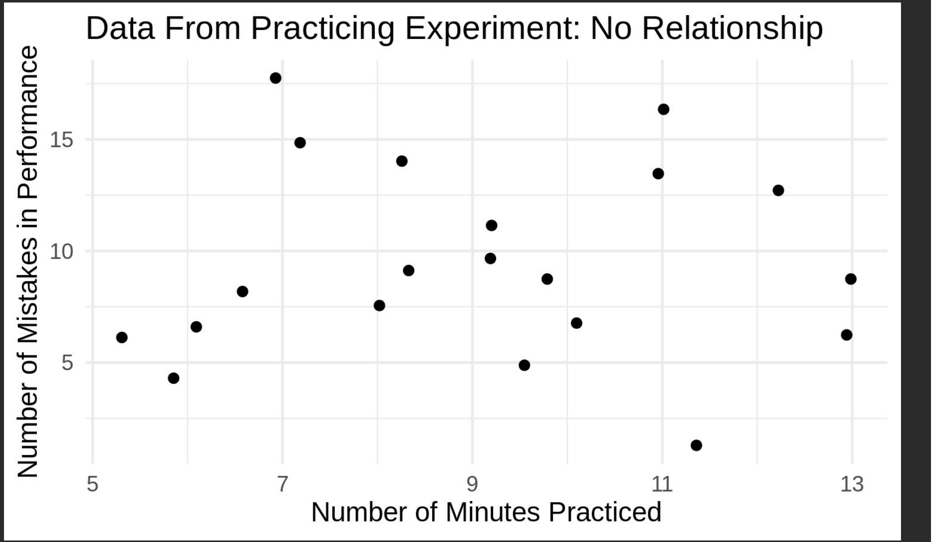
→ Logic with NHST: Instead of providing evidence FOR theory, we try to find evidence that nothing is NOT happening

→ If we take down the theory that nothing is happening, what are we left with?

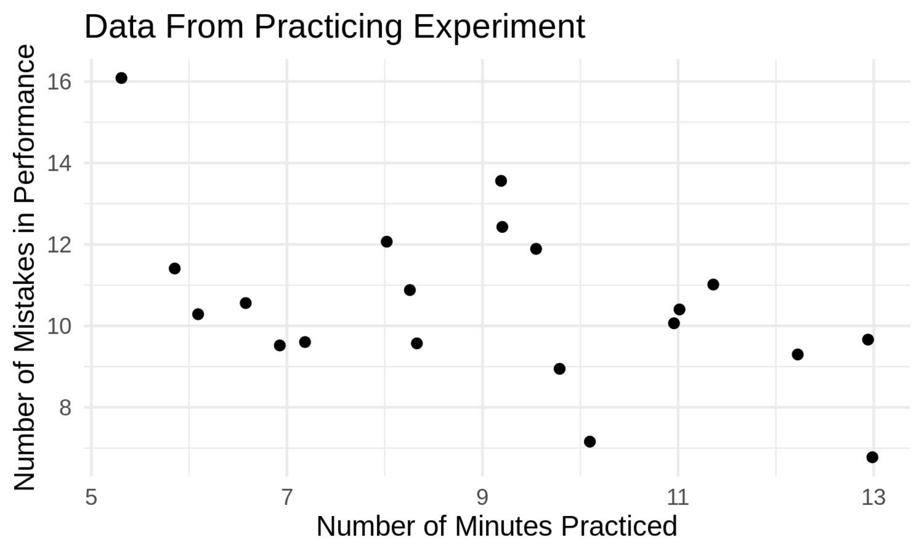
→ SOMETHING must be happening (statistical significance!)

→ This logic is difficult to grasp, will take a few attempts (or a career in science)

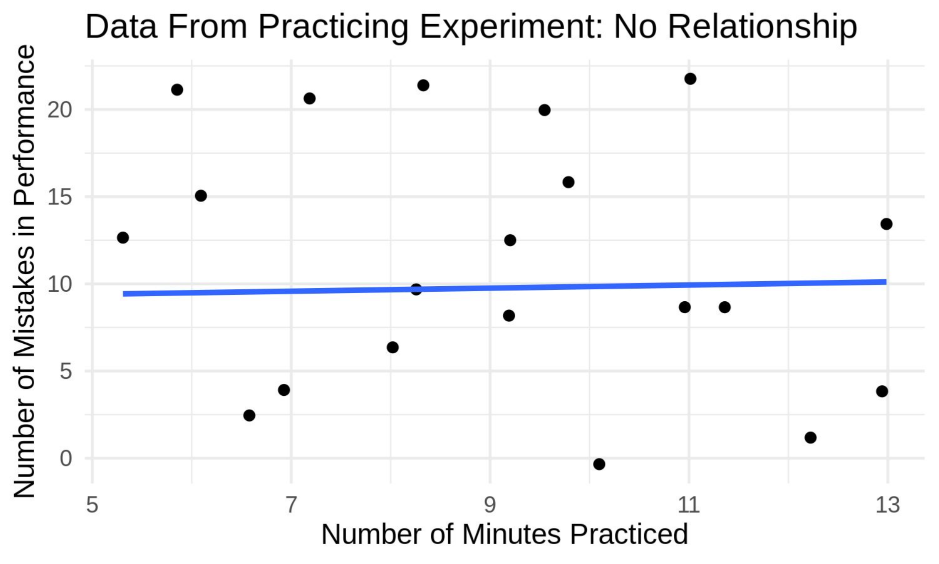
H0

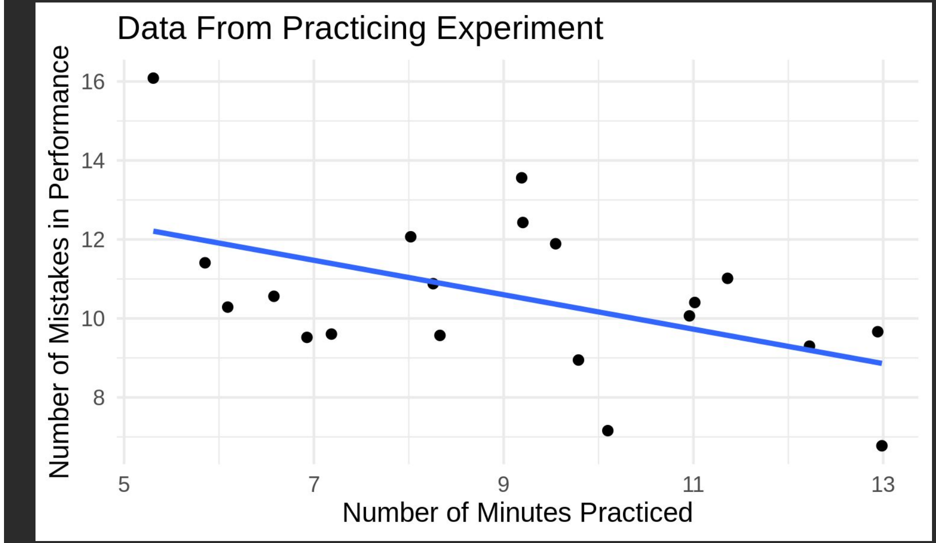

Data From Practicing Experiment: No Relationship
Number of Mistakes in Performance
5
10
15
5
7
9
11
13
Number of Minutes Practiced

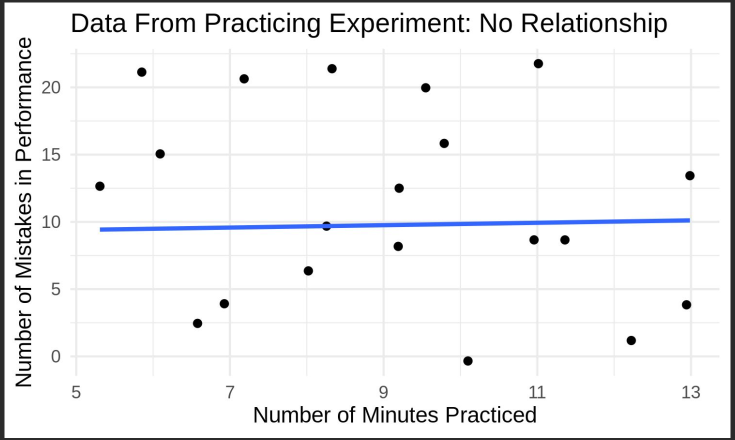
H1

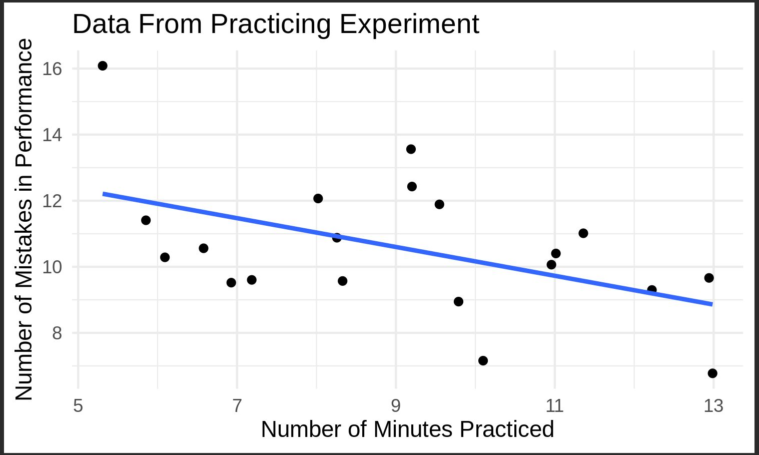

Data From Practicing Experiment
Number of Mistakes in Performance
8
10
12
14
16
5
7
9
11
13
Number of Minutes Practiced

Summarizing with a Line





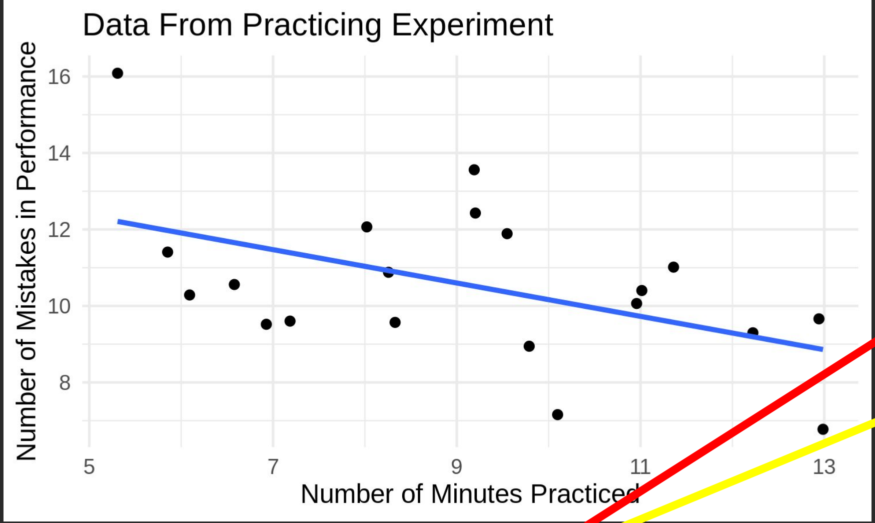




→ Run statistical test to show that the slope of the line of the data we collected (BOTTOM) is NOT zero

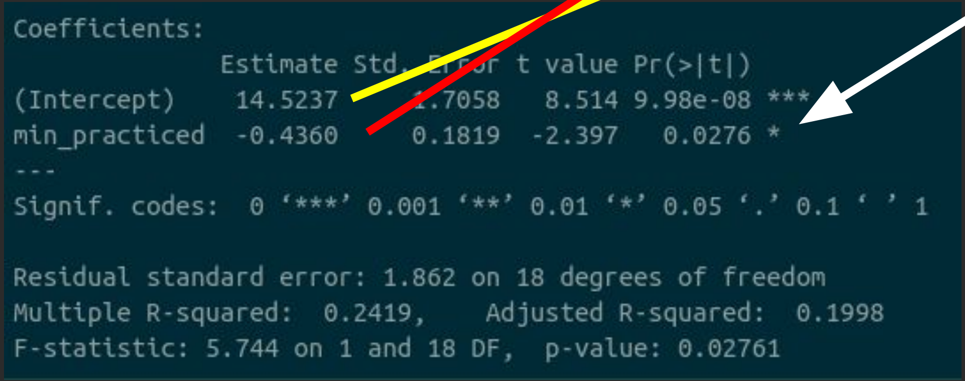
→ Compared with imaginary situation where nothing is happening (TOP)

→ Size of the slope will tell us how STRONG the relationship is (effect size)



# of Errors = -0.43 * min_practice + 14.52

P value (is the number NOT 0?!) , typically < .05 for   significant results



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    14.5237     1.7058   8.514 9.98e-08 ***
min_practiced  -0.4360     0.1819  -2.397   0.0276 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.862 on 18 degrees of freedom
Multiple R-squared:  0.2419,	Adjusted R-squared:  0.1998 
F-statistic: 5.744 on 1 and 18 DF,  p-value: 0.02761
```

Do we have evidence of learning yet?!

We have evidence to suggest that NO learning did NOT happen

Next steps!

- Excited!
- Have a statistical model that shows something is happening, in the direction we predicted, and very close to our predictions!!
- Write up a little study
- Send it to a journal !!

… wait a few weeks to hear back……?

Journal Response Back

- Reviewer 1: The authors provide evidence of a learning effect in practicing music after a week! They found a significant effect of time practiced on number of errors in a performance. The regression model looks great and I think this should get published!! **Accept as is!!**
- Reviewer 2: The authors provide some evidence for a learning effect, but they only did it with 20 people over the course of a week… how do I know that this isn't just a fluke? Please run a second experiment so I am more convinced about the results…. **Revise and resubmit!!!**
- Editor: After reading the reviewer comments, we would like to publish your study in our journal, but can only do this if you provide more evidence to satisfy one of our reviewers who remains skeptical of your finding given your sample size!!

Another Experiment!

- If you do this experiment again and this learning effect you found really does exist, do you think you will get the EXACT same results?

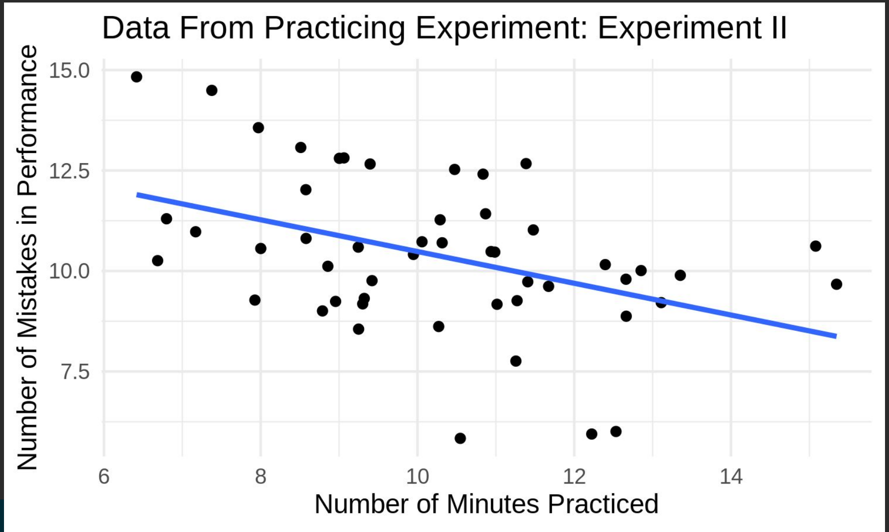
- What aspects of this experiment can you modify to satisfy Reviewer 2 and have more stable results?

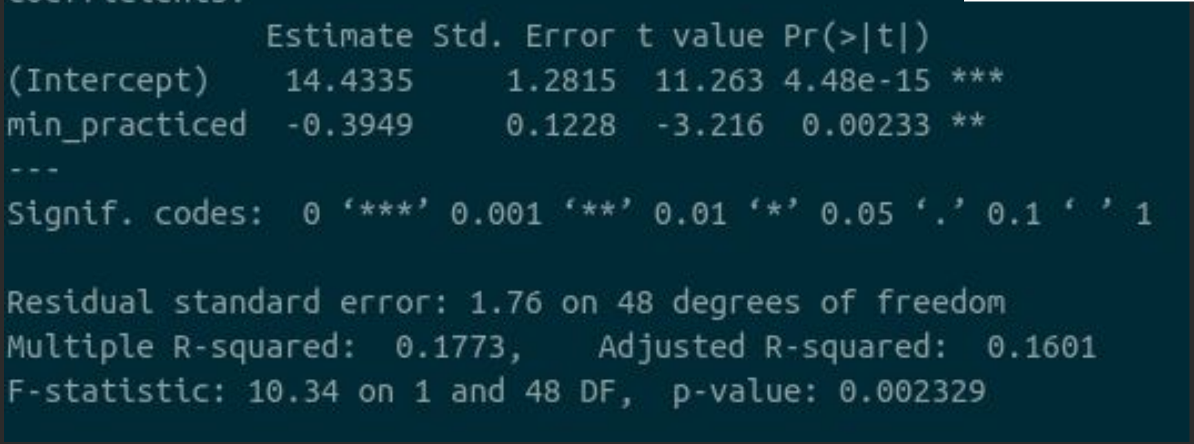
Another Experiment!

- If you do this experiment again and this learning effect you found really does exist, do you think you will get the EXACT same results?
  - No, but they should be close!!!
- What aspects of this experiment can you modify to satisfy Reviewer 2 and have more stable results?
  - Bigger sample!

→ **Run it again!!!**

Experiment II

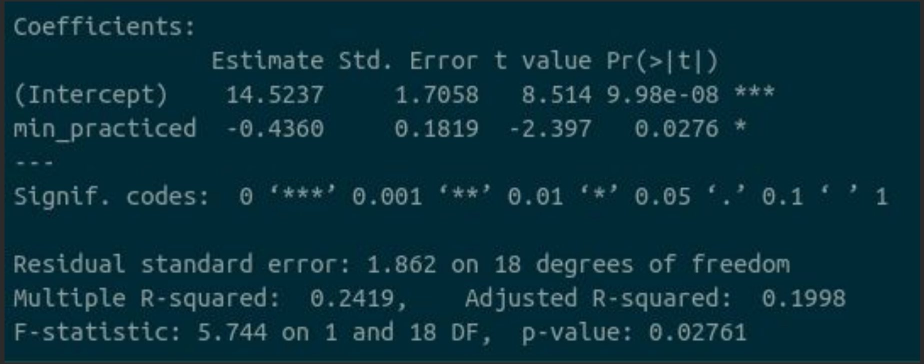




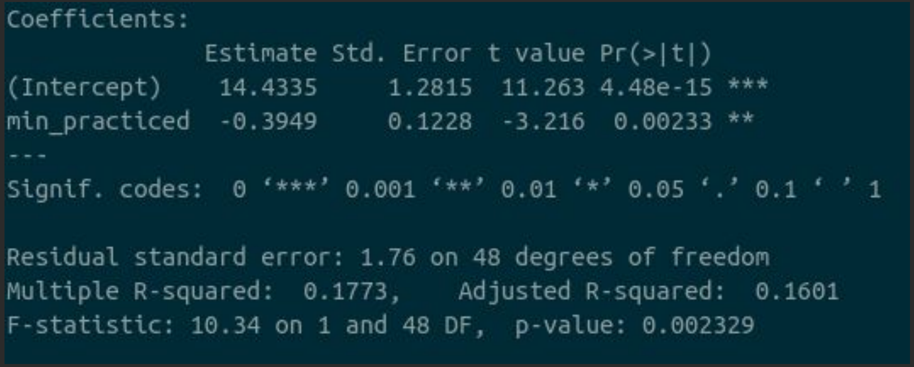
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.4335     1.2815  11.263 4.48e-15 ***
min_practiced  -0.3949     0.1228  -3.216  0.00233 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.76 on 48 degrees of freedom
Multiple R-squared:  0.1773,    Adjusted R-squared:  0.1601
F-statistic: 10.34 on 1 and 48 DF,  p-value: 0.002329
```

Did we replicate our results?

- What numbers are similar?
- What numbers are different? How so?
- Do you think the evidence from Experiment II will satisfy Reviewer 2?



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   14.5237     1.7058   8.514 9.98e-08 ***
min_practiced  -0.4360     0.1819  -2.397   0.0276 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.862 on 18 degrees of freedom
Multiple R-squared:  0.2419,	Adjusted R-squared:  0.1998 
F-statistic: 5.744 on 1 and 18 DF,  p-value: 0.02761
```



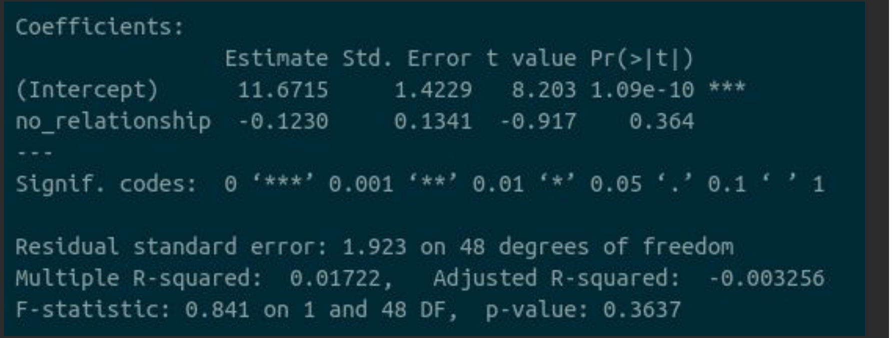
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   14.4335     1.2815  11.263 4.48e-15 ***
min_practiced  -0.3949     0.1228  -3.216  0.00233 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.76 on 48 degrees of freedom
Multiple R-squared:  0.1773,	Adjusted R-squared:  0.1601 
F-statistic: 10.34 on 1 and 48 DF,  p-value: 0.002329
```

Journal Response Back II

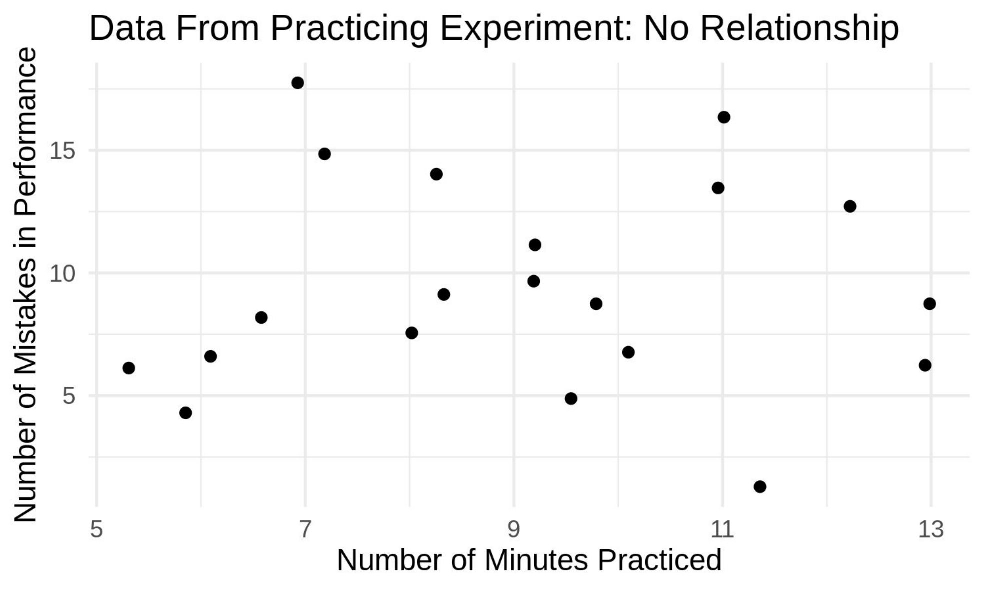
- Reviewer 1: Great, even more evidence for the effect! ACCEPT!
- Reviewer 2: Thanks for doing this, this makes me think that you indeed have found a robust effect of time practicing on the errors produced in piano performance! ACCEPT !!!
- Editor: ACCEPT!!

Break

What happens when the previous scenario doesn’t happen!?
Instead your first results looked like:



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      11.6715     1.4229   8.203 1.09e-10 ***
no_relationship  -0.1230     0.1341  -0.917    0.364    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.923 on 48 degrees of freedom
Multiple R-squared:  0.01722,   Adjusted R-squared:  -0.003256 
F-statistic: 0.841 on 1 and 48 DF,  p-value: 0.3637
```



Data From Practicing Experiment: No Relationship

What might have happened...

- Failed to find an effect, even though one existed!
- Maybe didn’t have enough participants to detect an effect?
- Just got a batch of data where it wasn’t there (could happen do to chance)
- Could be that there really is no effect of learning

There a specific names for these types of outcomes!

Claim to have found something, it really exists!
→ True Positive

Claim to have found something, it’s not really true…
→ Type I Error

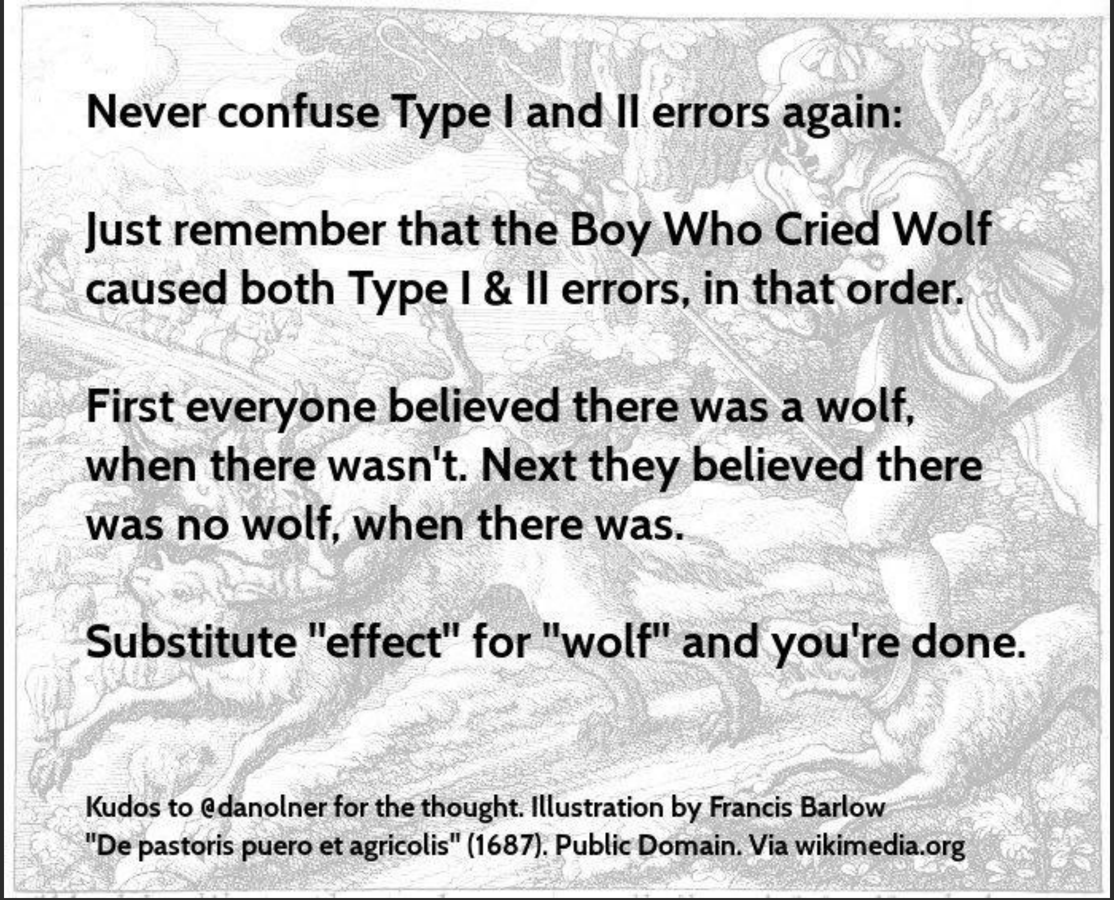
Claim to have not found something, it doesn’t exist
→ True Negative

Claim to have not found something, but it does exist…
→ Type II Error

Types of Errors

	H0 True	H1 True
Significant Finding	False Positive	True Positive
Non-Significant Finding	True Negative	False Negative

How to Remember Types of Errors



Never confuse Type I and II errors again:

Just remember that the Boy Who Cried Wolf caused both Type I & II errors, in that order.

First everyone believed there was a wolf, when there wasn't. Next they believed there was no wolf, when there was.

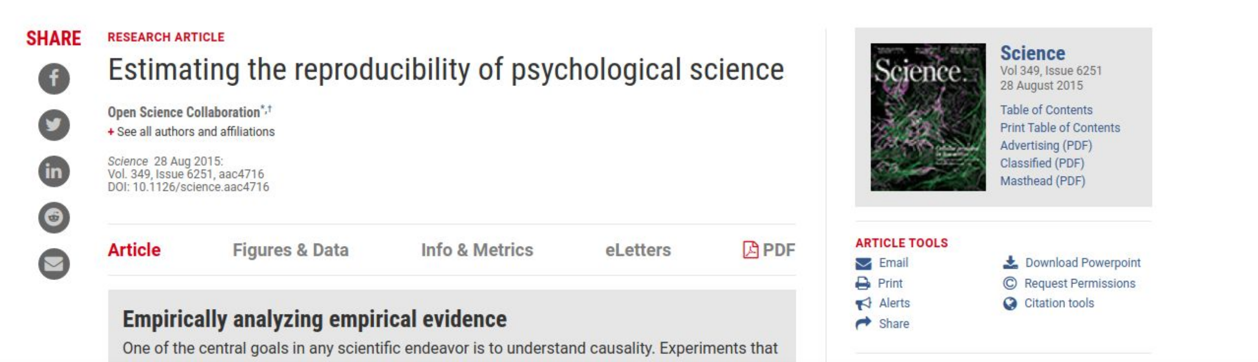
Substitute "effect" for "wolf" and you're done.

Kudos to @danolner for the thought. Illustration by Francis Barlow "De pastoris puero et agricolis" (1687). Public Domain. Via wikimedia.org

Replication in Psychological Sciences

Replication Crisis

- Questions of if findings are reproducible have been at forefront of psychological sciences for past 15 years
- Many “findings” that have been “discovered” have been found not to replicate
    - Type I errors!!
- Referred to as the Replication Crisis
- Other areas of science also experiencing this
    - Medicine
    - Artificial Intelligence
    - Food Science
    - Economics
    - Music Psychology….



RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration[*,†]

+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

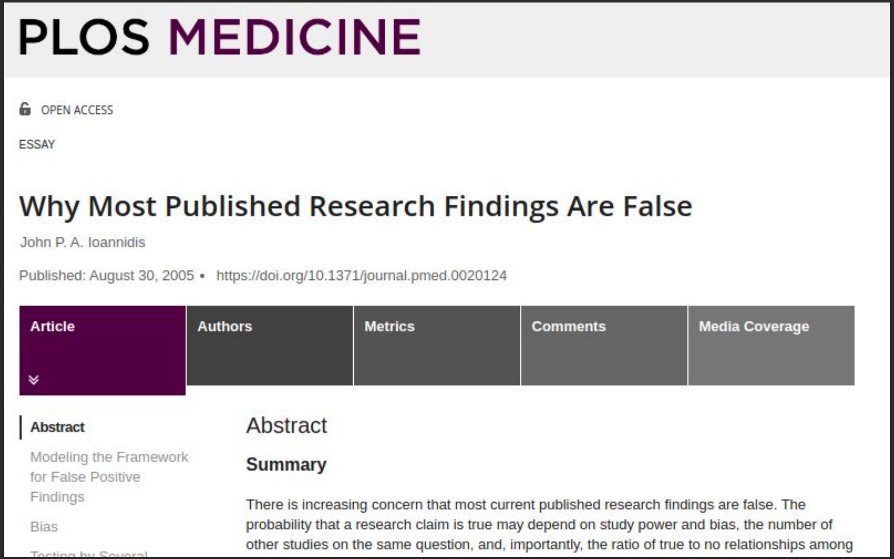
Empirically analyzing empirical evidence

One of the central goals in any scientific endeavor is to understand causality. Experiments that

Science
Vol 349, Issue 6251
28 August 2015

Food for Thought

- Replication Crisis has caused a lot of people think what do we want from our scientific questions and statistics
- Not only crisis of replication, but also generalizing
- What if the effect in our previous scenario only worked with one specific piece of piano music? Or only with teenagers?
- We want theories to generalize



PLOS MEDICINE

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among

Patterns Associated with Replication Studies

- Less significant findings
- Smaller effect sizes
- Often hard to exactly reproduce the experiment
- Stimuli lost
- People do not save their analysis scripts or data …

Why?!

Fraud

- With important discoveries comes clout, credibility, funding, career advancements
- People will fabricate to get ahead

Bias

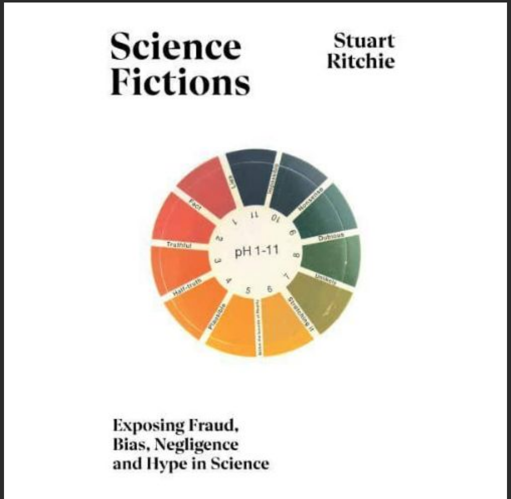
- People know what they want ahead of time
- Do a study, get null results, don’t publish

Negligence

- Sloppy in reporting numbers
- Will cut corners on statistical analysis to get results that can be published (p-hacking)

Hype

- Scientists want to pursue questions that are surprising and newsworthy! Will get research published that might be a fluke but sounds nice



Terms to Review

Deterministic

Null Hypothesis Significance Testing

Null Hypothesis (H0)

Alternative Hypothesis (H1)

Generalizability

Type I Error

Type II Error

True Positive

True Negative

Replication Crisis

Reproducibility

Takeaways

- Many ways for scientific research to not go well
- It’s important to be skeptical of research
- Point is to try to prove yourself wrong, if you can’t, good evidence for theory