Photo credit: Dave DiCello

17-803 Empirical Methods

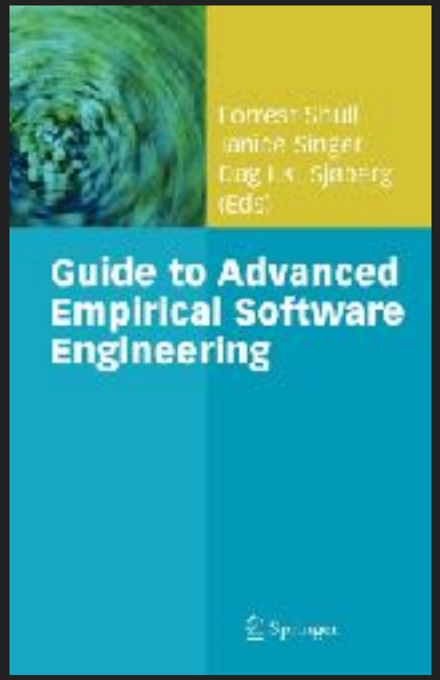
Bogdan Vasilescu, S3D

# Designing Experiments (III)

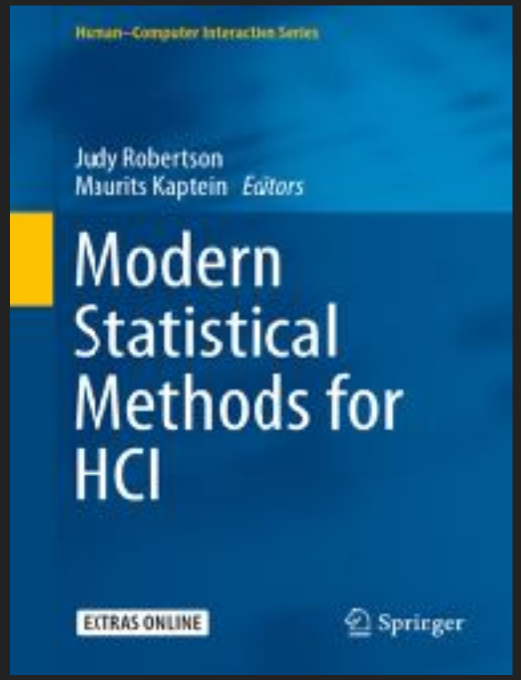
Thursday, March 14, 2024

# Readings

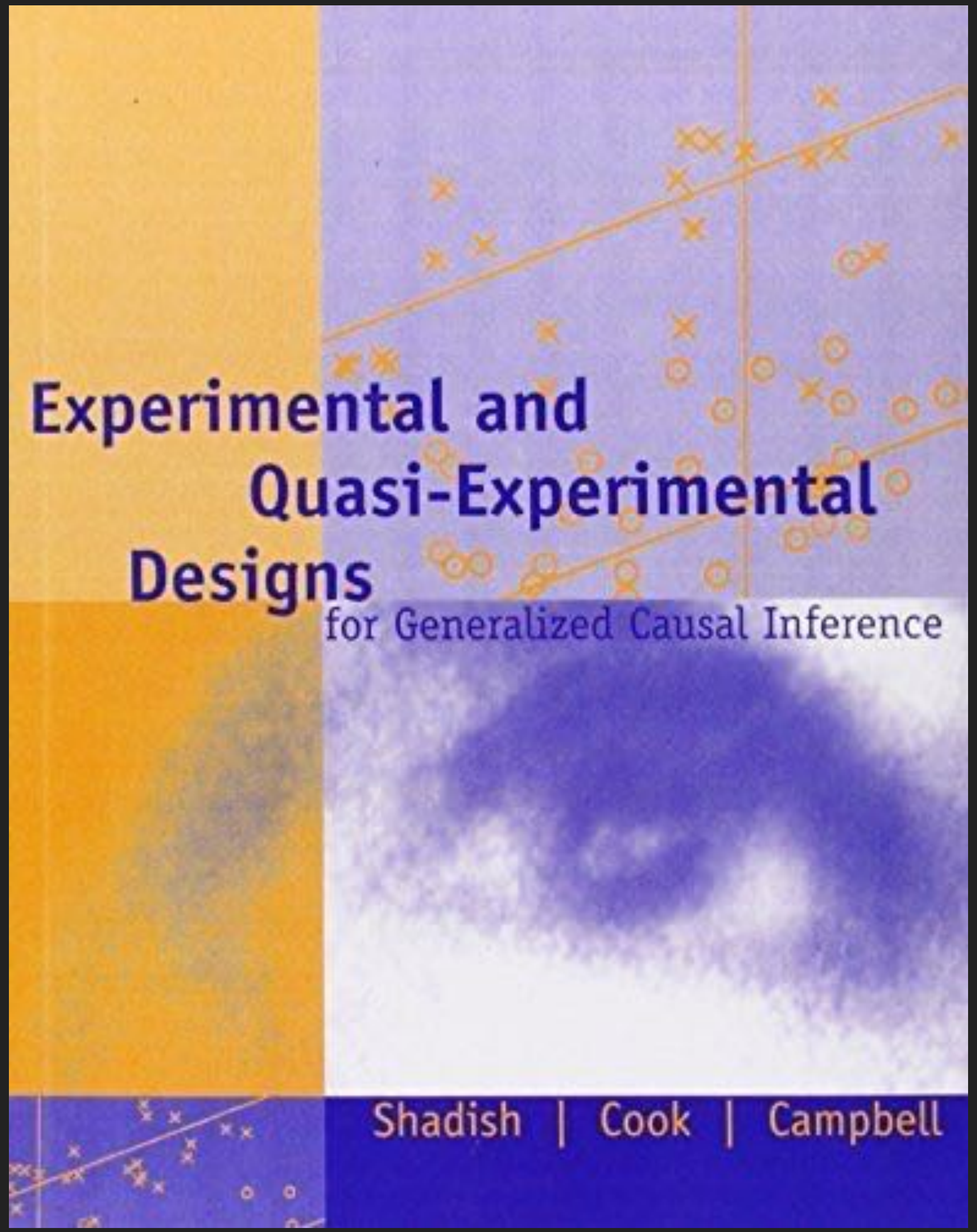

Ch 10 (Analysis and interpretation)



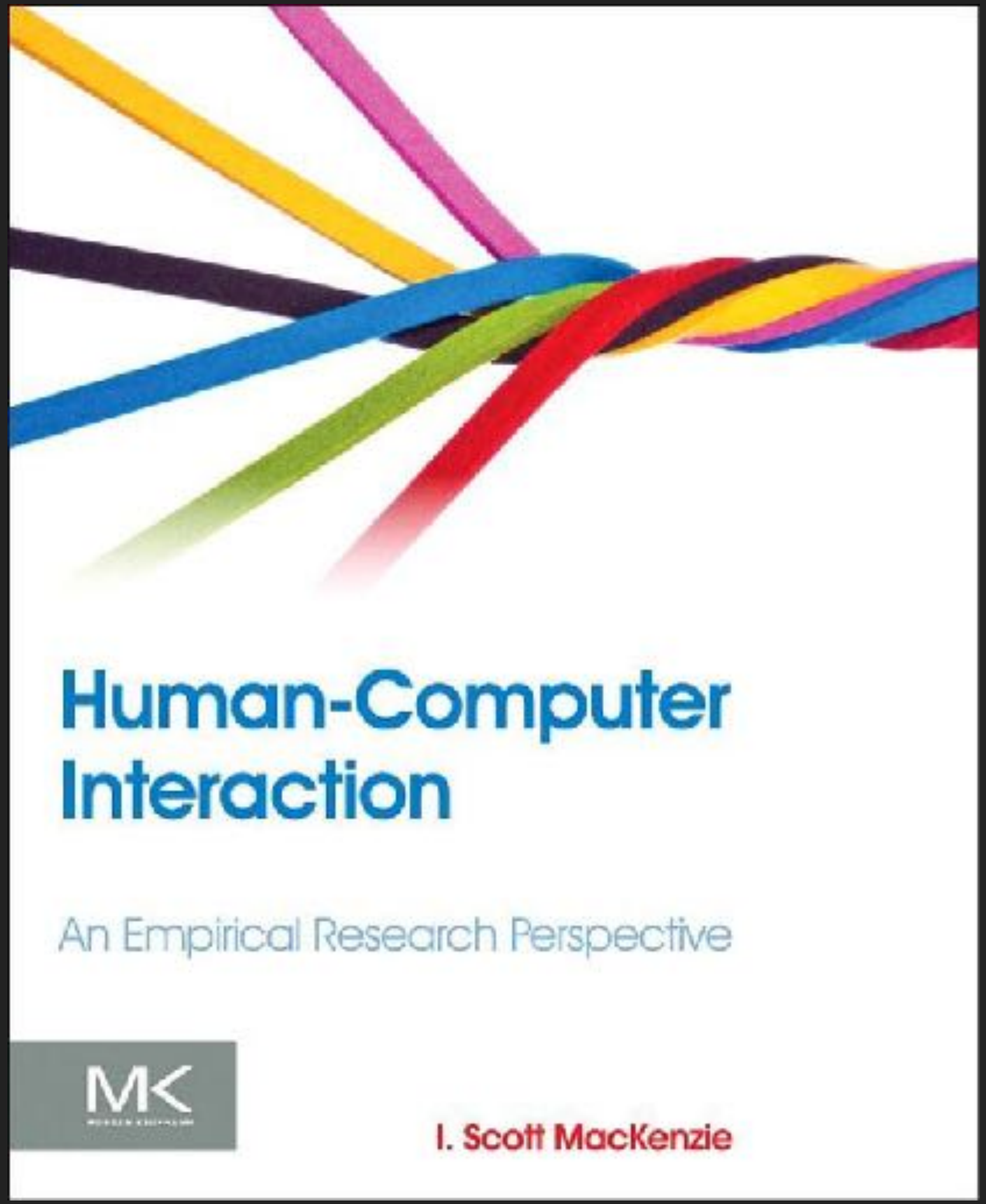

Ch 6 (Statistical methods and measurement)



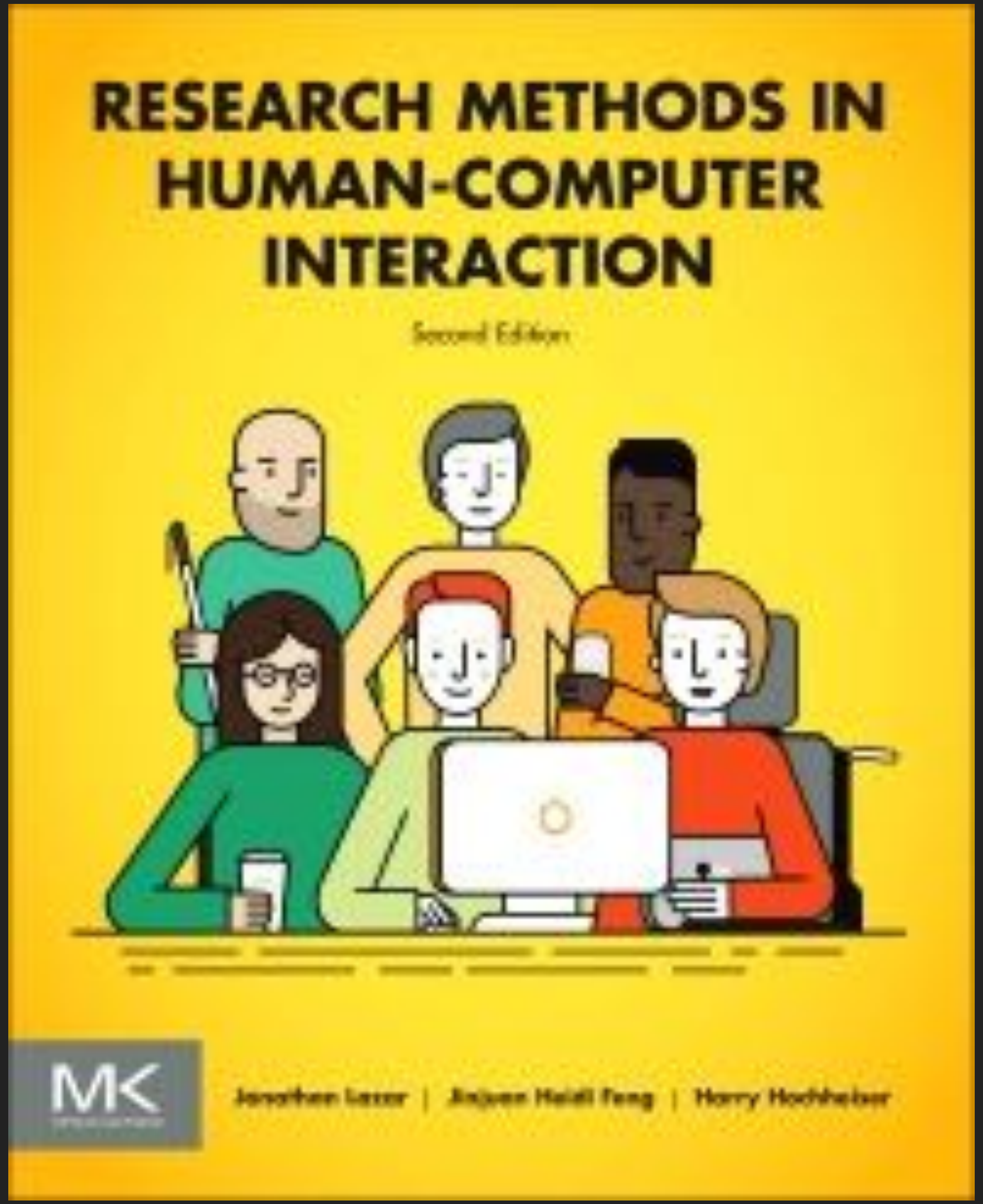

Ch 5 (Effect sizes and power analysis)
Ch 13 (Fair statistical communication)
Ch 14 (Improving statistical practice)




Ch 1 (Experiments and causality)
Ch 2 & 3 (Validity)
Ch 8 (Randomized experiments)




Ch 5 (Designing HCI Exp.)
Ch 6 (Hypothesis testing)




Ch 3 (Experimental design)
Ch 4 (Statistical analysis)

# The generalization of causal connections

# Four Types of Validity

## Statistical Conclusion Validity

The validity of inferences about the correlation (covariation) between treatment and outcome.

## Internal Validity

The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.

## Construct Validity

The validity of inferences about the higher order constructs that represent sampling particulars.

## External Validity

The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

# Construct Validity

See book

- Can we generalize results to the theoretical constructs that the units, treatments, observations, and settings are supposed to represent?
- E.g., whether
  - patient education (the target cause)
  - promotes physical recovery (the target effect)
  - among surgical patients (the target population of units)
  - in hospitals (the target universe of settings)
- Do the actual manipulations and measures used in the experiment really tap into the specific cause and effect constructs specified by the theory?

# External Validity

See book

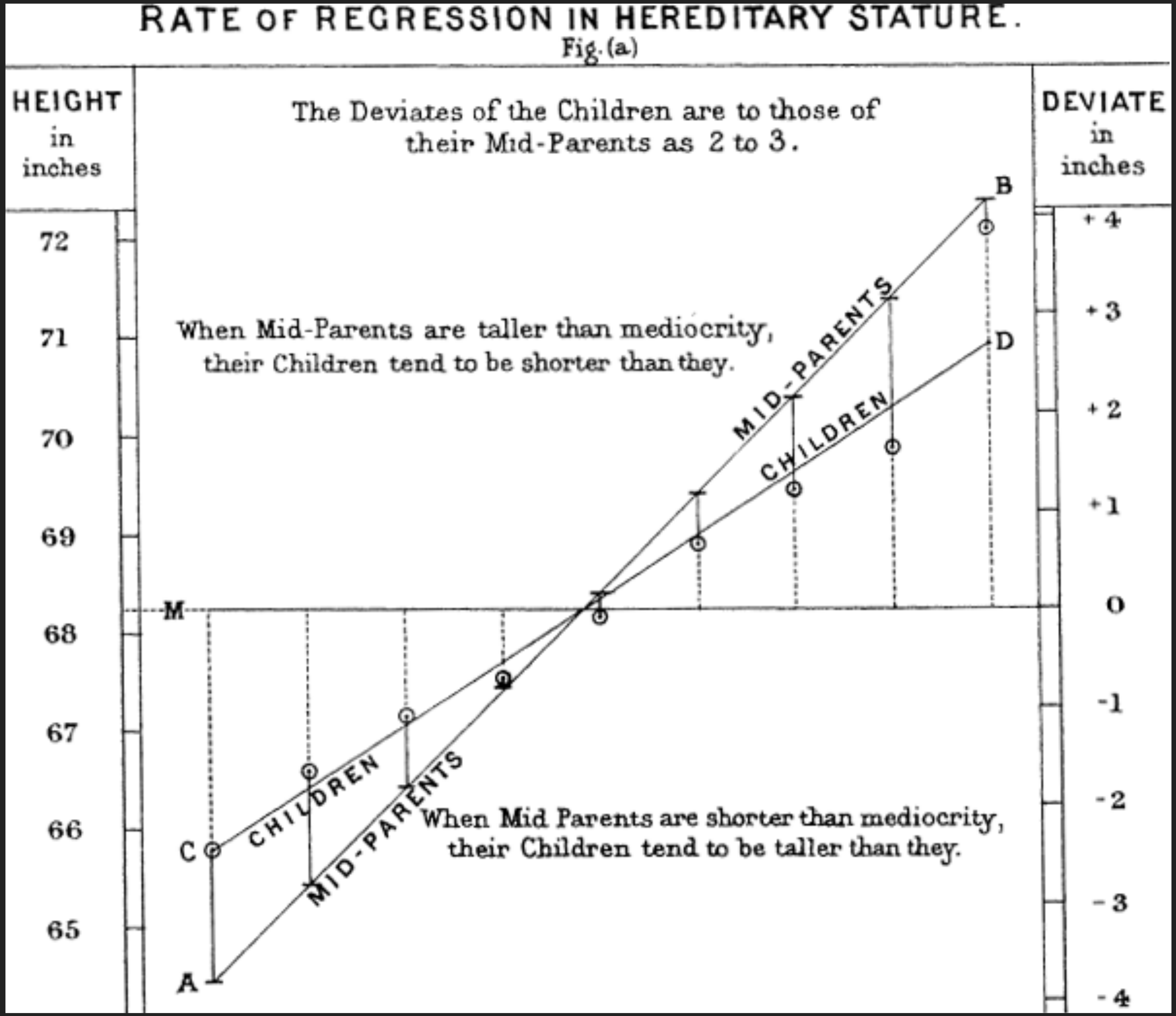
- Does the causal relationship hold over variations in persons, settings, treatments, and outcomes?
  - Narrow to broad?
  - Broad to narrow?
  - Across units at the same level of aggregation?

# A Few Threats to Internal Validity

- Ambiguous Temporal Precedence:
  - Which variable occurred first?
- Selection:
  - Systematic differences over conditions in respondent characteristics.
- History:
  - Events occurring concurrently with treatment.
- Maturation:
  - Naturally occurring changes over time confused with a treatment effect.
- Regression:
  - When units are selected for their extreme scores, they will often have less extreme scores on other variables.

# Regression to the Mean

- Phenomenon involving successive measurements on a given variable.
- Extreme observations tend to be followed by more central ones.
  - E.g., the children of extremely tall men tend not to be as tall as their father [Galton-1886].



# A Few Threats to Internal Validity

- Ambiguous Temporal Precedence:
  - Which variable occurred first?
- Selection:
  - Systematic differences over conditions in respondent characteristics.
- History:
  - Events occurring concurrently with treatment.
- Maturation:
  - Naturally occurring changes over time confused with a treatment effect.
- Regression:
  - When units are selected for their extreme scores, they will often have less extreme scores on other variables.
- Attrition:
  - Loss of respondents to treatment or to measurement.
- Testing:
  - Exposure to a test can affect scores on subsequent exposures to that test.
- Instrumentation:
  - The nature of a measure may change over time or conditions.

# Statistical Conclusion Validity

- Two related statistical inferences that affect the covariation component of causal inferences:
  - whether the presumed cause and effect covary.
  - how strongly they covary.
- Type I error:
  - incorrectly conclude that cause and effect covary when they do not.
- Type II error:
  - incorrectly conclude that they do not covary when they do.

# A Few Threats to Statistical Conclusion Validity

See book

- Low Statistical Power:
  - → Type II errors
- Violated assumptions of statistical tests:
  - Either over- or underestimate the size and significance of an effect.
- Fishing:
  - Repeated tests can inflate statistical significance.
- Unreliability of measures
- Restriction of range on variable:
  - Typically weakens the relationship between it and another variable.
  - E.g., don't dichotomize.

# Hypothesis Tests

- Aka "significance tests"
- Purpose:
  - Could random chance be responsible for an observed effect?
- Null hypothesis ($H_0$):
  - The hypothesis that chance is to blame.
  - e.g., "There is no difference in the mean time to complete a task using NL2Code vs. writing code from scratch."
- Alternative hypothesis ($H_a$):
  - Counterpoint to the null (what you hope to prove).
  - e.g., "It takes less time on average to complete a task using NL2Code rather than by writing code from scratch."

# Aside: Why Do We Need a Hypothesis? Why Not Just Look at the Outcome of the Experiment and Go With Whichever Treatment Does Better?

- Experiment: invent a series of 50 coin flips.
  - Write down a series of random 1s and 0s: [1, 0, 1, 0, 1, 0, …]

# Aside: Why Do We Need a Hypothesis? Why Not Just Look at the Outcome of the Experiment and Go With Whichever Treatment Does Better?

- Experiment: invent a series of 50 coin flips.
  - Write down a series of random 1s and 0s: [1, 0, 1, 0, 1, 0, …]
- Humans have a tendency to underestimate randomness.
- Computer-generated "real" coin flip results vs made-up human results:
  - the real ones will have longer runs of 1s or 0s.
  - median length of subsequences of 1s in a row:
    - 5 for the computer-generated sequences
    - only 4 for the human-generated set
- When most of us are inventing random coin flips and we have gotten three or four 1s in a row, we tell ourselves that, for the series to look random, we had better switch to 0.

# Aside: How Do You Interpret the P-Value?

- $H_0$: "There is no difference in the mean time to complete a task using NL2Code vs. writing code from scratch."
- $H_a$: "It takes less time on average to complete a task using NL2Code rather than writing code from scratch."

- You run some statistical test (e.g., t-test) and obtain a p-value.

# Aside: P-Value Controversy

- What we would like the p-value to convey:
  - (We hope for a low value, so we can conclude that we've proved something.)

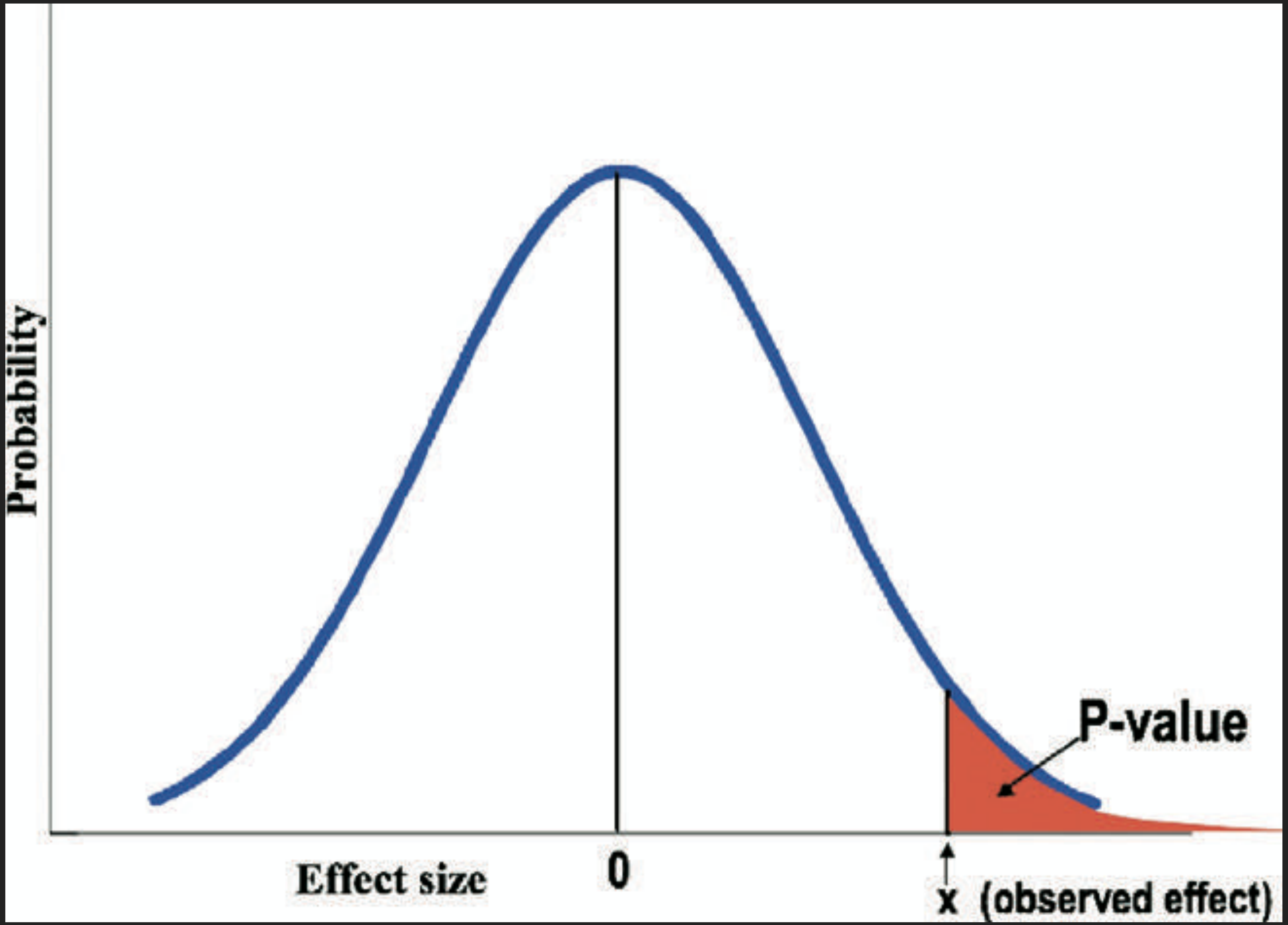
The probability that the result is due to chance: $P(H_0|D)$

- What the p-value actually represents:

The probability that, given a chance model, results as extreme as the observed results could occur: $P(D|H_0)$

Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI.
In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

# The P Value Is the Probability of the Observed Outcome (X) Plus all "More Extreme" Outcomes




Graphical depiction of the definition of a (one-sided) P value. The curve represents the probability of every observed outcome under the null hypothesis.

# The P Value Is the Probability of the Observed Outcome (X) Plus all "More Extreme" Outcomes

- Not the probability that the null hypothesis is true!
- Example: Is a coin fair or not?
  - $H_0$: The coin is fair: P(Heads) = P(Tails) = 1/2
  - $H_a$: The coin is biased: P(Heads) ≠ 1/2

# Consider Four Consecutive Coin Flips:

Probability

- First toss: 


?

# Consider Four Consecutive Coin Flips:

Probability

- First toss: 

0.5

- Second toss: 

?

# Consider Four Consecutive Coin Flips:

Probability

- First toss:  0.5
- Second toss:  0.25
- Third toss:  0.125
- Fourth toss:  0.0625

# Is Coin Fair?

- Two-sided P = 0.125.


0.0625


0.0625

- This does not mean that the probability of the coin being fair is only 12.5%!

# Aside: P-Value Controversy

- What we would like the p-value to convey:
  - (We hope for a low value, so we can conclude that we've proved something.)

The probability that the result is due to chance: $P(H_0|D)$

- What the p-value actually represents:

The probability that, given a chance model, results as extreme as the observed results could occur: $P(D|H_0)$

Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI.
In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

# Is Coin Fair?

- Two-sided P = 0.125.


0.0625


0.0625

- This does not mean that the probability of the coin being fair is only 12.5%!

$$P(H_0|D) = \frac{P(D|H_0)\,P(H_0)}{P(D)}$$

Common false belief that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to external evidence or the plausibility of the underlying mechanism.

# Twelve P-Value Misconceptions

**Table 1** Twelve *P*-Value Misconceptions

| | |
|---|---|
| 1 | *If P = .05, the null hypothesis has only a 5% chance of being true.* |
| 2 | *A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.* |
| 3 | *A statistically significant finding is clinically important.* |
| 4 | *Studies with P values on opposite sides of .05 are conflicting.* |
| 5 | *Studies with the same P value provide the same evidence against the null hypothesis.* |
| 6 | *P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.* |
| 7 | *P = .05 and P ≤.05 mean the same thing.* |
| 8 | *P values are properly written as inequalities (eg, "P ≤.02" when P = .015)* |
| 9 | *P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.* |
| 10 | *With a P = .05 threshold for significance, the chance of a type I error will be 5%.* |
| 11 | *You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.* |
| 12 | *A scientific conclusion or treatment policy should be based on whether or not the P value is significant.* |

Goodman, S. (2008, July). A dirty dozen: twelve p-value misconceptions.
In Seminars in hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.

# Type I and Type II Errors

| | | Study conclusion | |
|---|---|---|---|
| | | No difference | Using NL2Code is faster |
| Reality | No difference | ✓ | Type I error |
| | Using NL2Code is faster | Type II error | ✓ |

# Type I and Type II Errors

- In assessing statistical significance, two types of error are possible:
  - Type I: you mistakenly conclude an effect is real, when it is really just due to chance
    - False positives
  - Type II: you mistakenly conclude that an effect is due to chance, when it actually is real
    - False negatives
- The basic function of hypothesis tests is to protect against being fooled by random chance; thus they are typically structured to minimize Type I errors.

# Controlling the Risks of Type I and Type II Errors

- The probability of making a Type I error is called alpha.
  - (or “significance level”, “P-value”)
- The probability of making a Type II error is called beta.
- The statistical power of a test, defined as $1 - \beta$, refers to the probability of successfully rejecting a null hypothesis when it is false and should be rejected.
- To reduce errors:
  - Type I: $P < 0.05$
  - Type II: large sample size

# Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the overall probability of Type I errors (false positives)?

# Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the overall probability of Type I errors (false positives)?

- The probability that one will incorrectly test significant is ...?

# Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the overall probability of Type I errors (false positives)?

- The probability that one will incorrectly test significant is 0.05

# Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the overall probability of Type I errors (false positives)?

- The probability that one will correctly test nonsignificant is …?

# Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the overall probability of Type I errors (false positives)?

- The probability that one will correctly test nonsignificant is …?

| | | Study conclusion | |
|---|---|---|---|
| | | No difference | Difference |
| Reality | No difference | ? | 0.05 |
| | Difference | | |

# Aside: Torture the Data Long Enough, and It Will Confess.

- Imagine you have 20 predictor variables and one outcome variable, all randomly generated.
- You do 20 significance tests at the alpha = 0.05 level (one per variable).
- What's the overall probability of Type I errors (false positives)?

- The probability that one will correctly test nonsignificant is 0.95
- The probability that all 20 will correctly test nonsignificant is:
  - $0.95 \times 0.95 \times 0.95...$, or $0.95^{20} = 0.36$
- The probability that at least one predictor will (falsely) test significant:
  - 1 – (probability that all will be nonsignificant) = 0.64 😱

# Numbers and Nonsense

# Drink Hot Cocoa Before Bed?

- “99.9% caffeine-free”
- 20-ounce Starbucks coffee:
  - 415 milligrams of caffeine.
  - ~21 mg caffeine per ounce.
  - 1 fl oz water weighs ~28 grams.
  - Thus, Starbucks drip coffee is ~0.075% caffeine by weight.
- Strong coffee is also 99.9% caffeine free!

# Tweet About Your Work?

Monday, June 29, 2020 3:40 PM

**Tweeting about research results in three times more citations**

*Social media is proven to help share new science with the public*

**Dori Grijseels**
Neuroscience
University of Sussex

Unsplash

An important part of science is sharing the findings, both with the general public, and with fellow scientists. The main method of sharing science is done by writing articles that are published in academic journals. However, most people are not subscribed to the *Annals of Thoracic Surgery*, and thus may not be aware of the latest articles that came out. This means that a lot of articles never reach the general public, or sometimes even fellow scientists. A new study by the Thoracic Surgery Social Media Network shows that tweeting might be the solution.

https://massivesci.com/notes/tweet-science-communication-research-public/

Monday, June 29, 2020 3:40 PM

**Tweeting about research results in three times more citations**

*Social media is proven to help share new science with the public*

**Dori Grijseels**
Neuroscience
University of Sussex

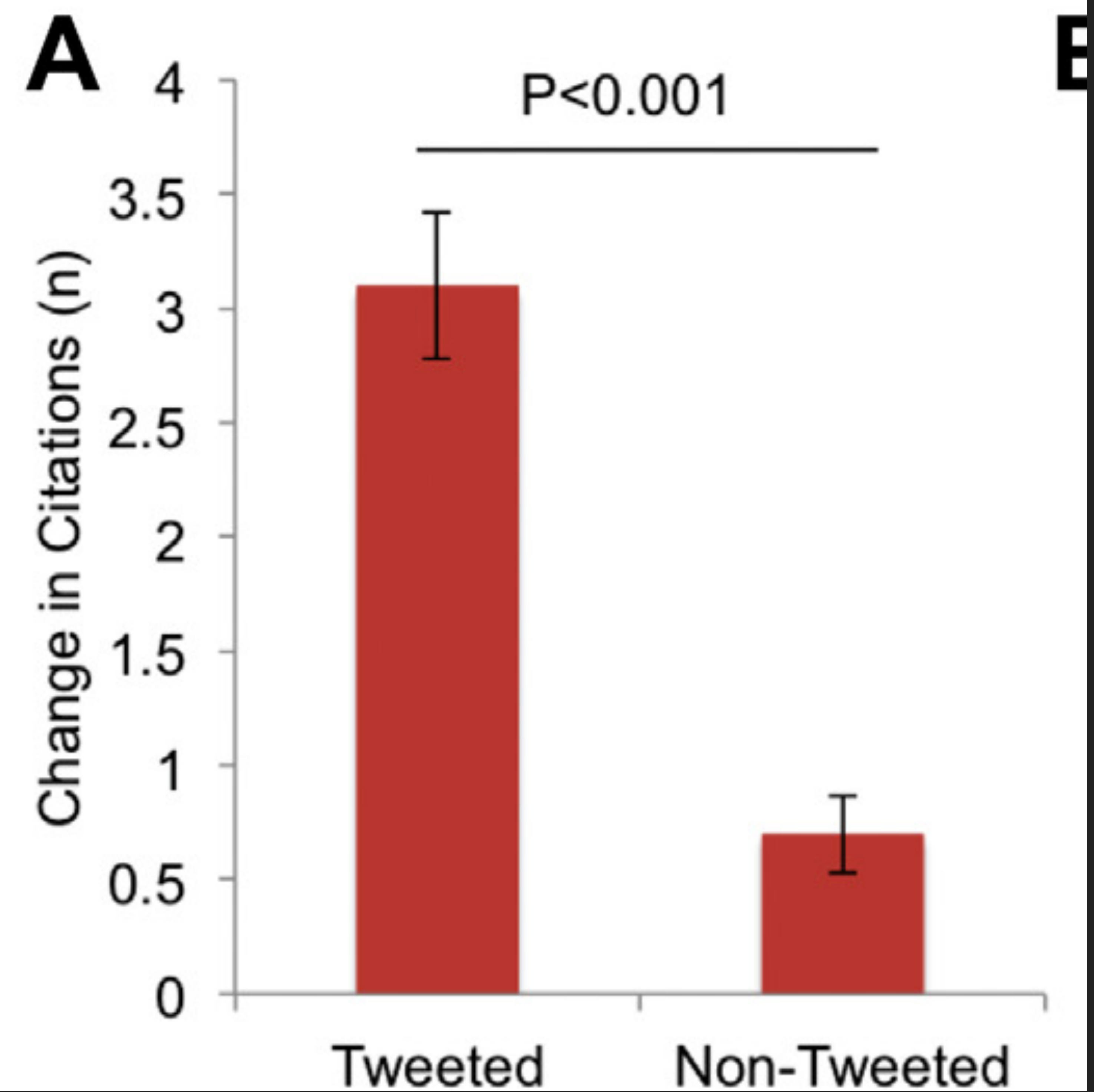
Unsplash

An important part of science is sharing the findings, both with the general public, and with fellow scientists. The main method of sharing science is done by writing articles that are published in academic journals. However, most people are not subscribed to the *Annals of Thoracic Surgery*, and thus may not be aware of the latest articles that came out. This means that a lot of articles never reach the general public, or sometimes even fellow scientists. A new study by the Thoracic Surgery Social Media Network shows that tweeting might be the solution.

https://massivesci.com/notes/tweet-science-communication-research-public/

# Tweet About Your Work?

## Meanwhile:




Luc, J. G., Archer, M. A., Arora, R. C., Bender, E. M., Blitz, A., Cooke, D. T., ... & Antonoff, M. B. (2021). Does tweeting improve citations? One-year results from the TSSMN prospective randomized trial. The Annals of Thoracic Surgery, 111(1), 296-300.

# Selection Bias

# The Friendship Paradox

😱 **Most likely, the majority of your friends have more friends than you do**

# The Friendship Paradox

😱 **Most likely, the majority of your friends have more friends than you do**

- Suppose you follow Rihanna and 499 other people on Twitter.
- Rihanna has over one hundred million followers.
- The 500 people you follow will average at the very least 100,000,000 / 500 = 200,000 followers—far more than you have.

Most people have fewer friends than their average (mean) friend has.

# The Friendship Paradox

😱 **Most likely, the majority of your friends have more friends than you do**

- 84 percent of Facebook users have fewer friends than the median friend count of their friends.

🤯 Most people also have fewer friends than their <u>median</u> friend has.

# Do You Often Have To Wait a Surprisingly Long Time for the Next Bus To Arrive?

- Suppose that buses leave a bus stop at regular ten minute intervals.
- If you arrive at an arbitrary time, how long do you expect to wait, on average?

# Do You Often Have To Wait a Surprisingly Long Time for the Next Bus To Arrive?

- Suppose that buses leave a bus stop at regular ten minute intervals.
- If you arrive at an arbitrary time, how long do you expect to wait, on average?

5 minutes?

# Do You Often Have To Wait a Surprisingly Long Time for the Next Bus To Arrive?

- What if buses leave every ten minutes on average – but traffic forces the buses to run somewhat irregularly?
- Sometimes the time between buses is quite short; other times it may extend for fifteen minutes or more.
- Now how long do you expect to wait?

# Do You Often Have To Wait a Surprisingly Long Time for the Next Bus To Arrive?

- What if buses leave every ten minutes on average – but traffic forces the buses to run somewhat irregularly?
- Sometimes the time between buses is quite short; other times it may extend for fifteen minutes or more.
- Now how long do you expect to wait?

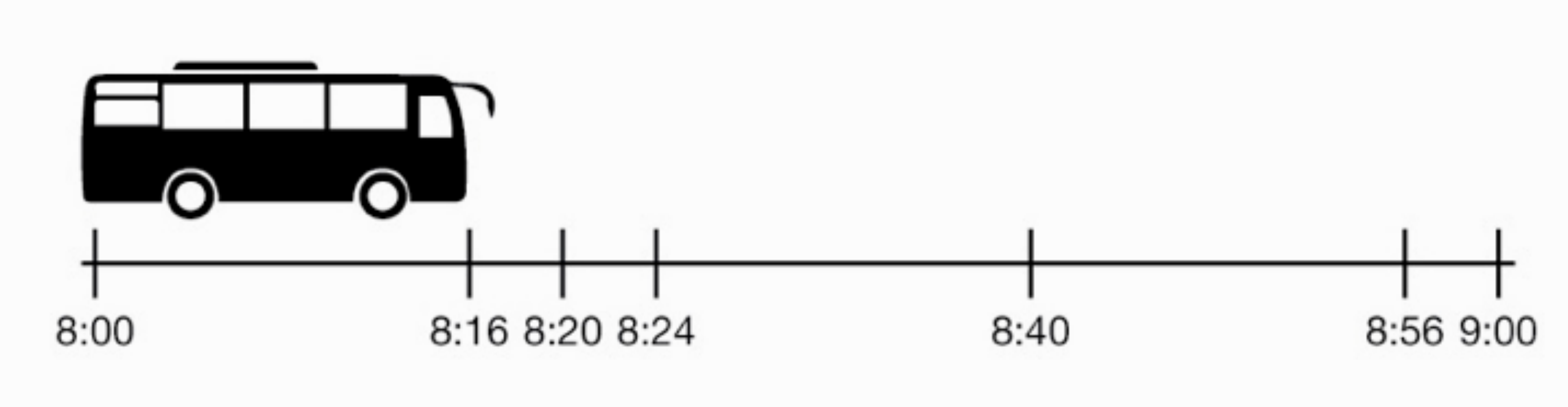
5 minutes?

# Do You Often Have To Wait a Surprisingly Long Time for the Next Bus To Arrive?

- You are more likely to arrive during one of the long intervals than during one of the short intervals.
- As a result, you end up waiting longer than five minutes, on average.

~~5 minutes?~~

# Do You Often Have To Wait a Surprisingly Long Time for the Next Bus To Arrive?

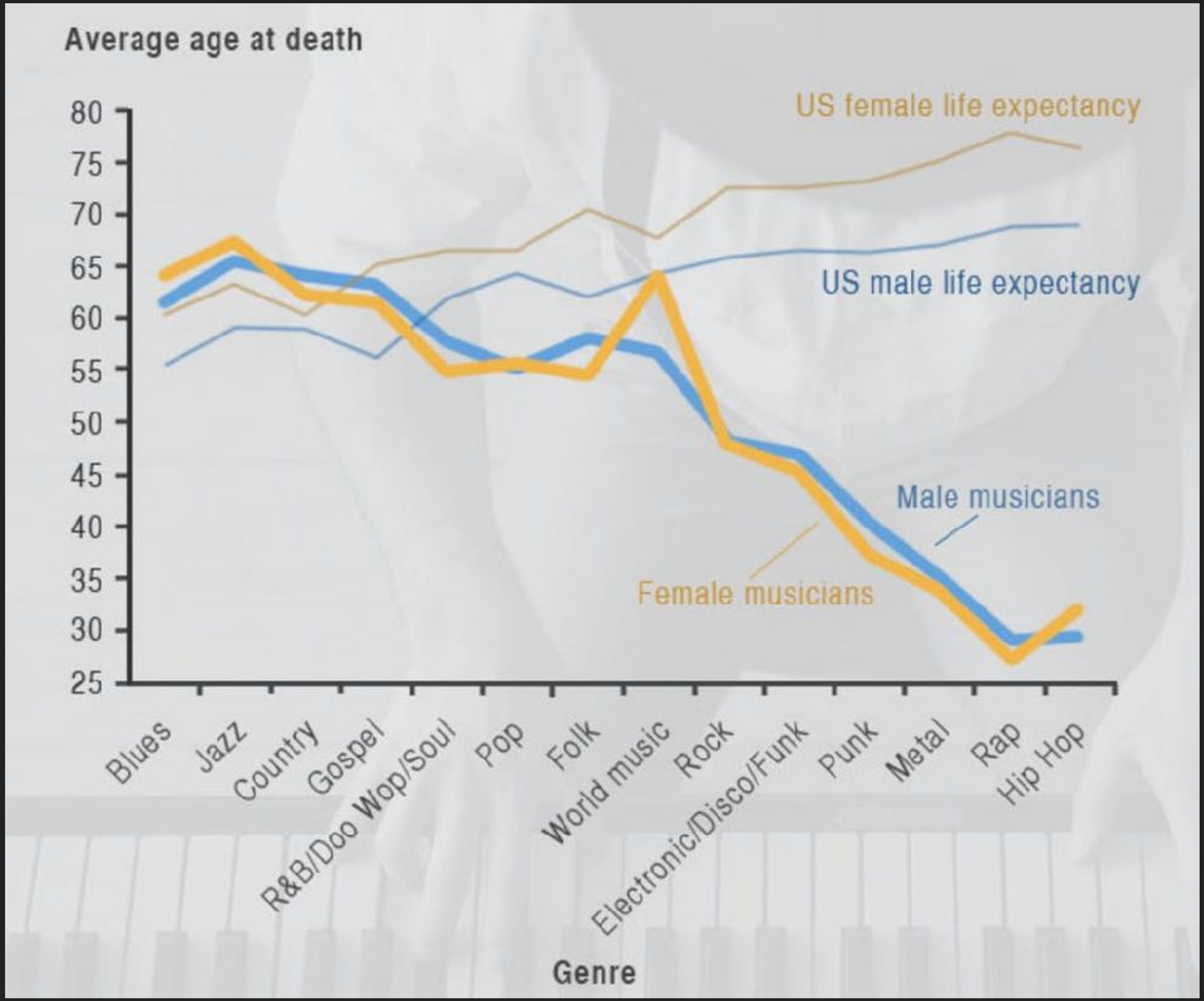



- 80% chance of arriving during one of the long intervals
  - wait 8 minutes on average.
- 20% chance of arriving during one of the short intervals
  - wait 2 minutes on average.
- Average overall wait time: $(0.8 \times 8) + (0.2 \times 2) = 6.8$ mins

# Observation Selection Effect

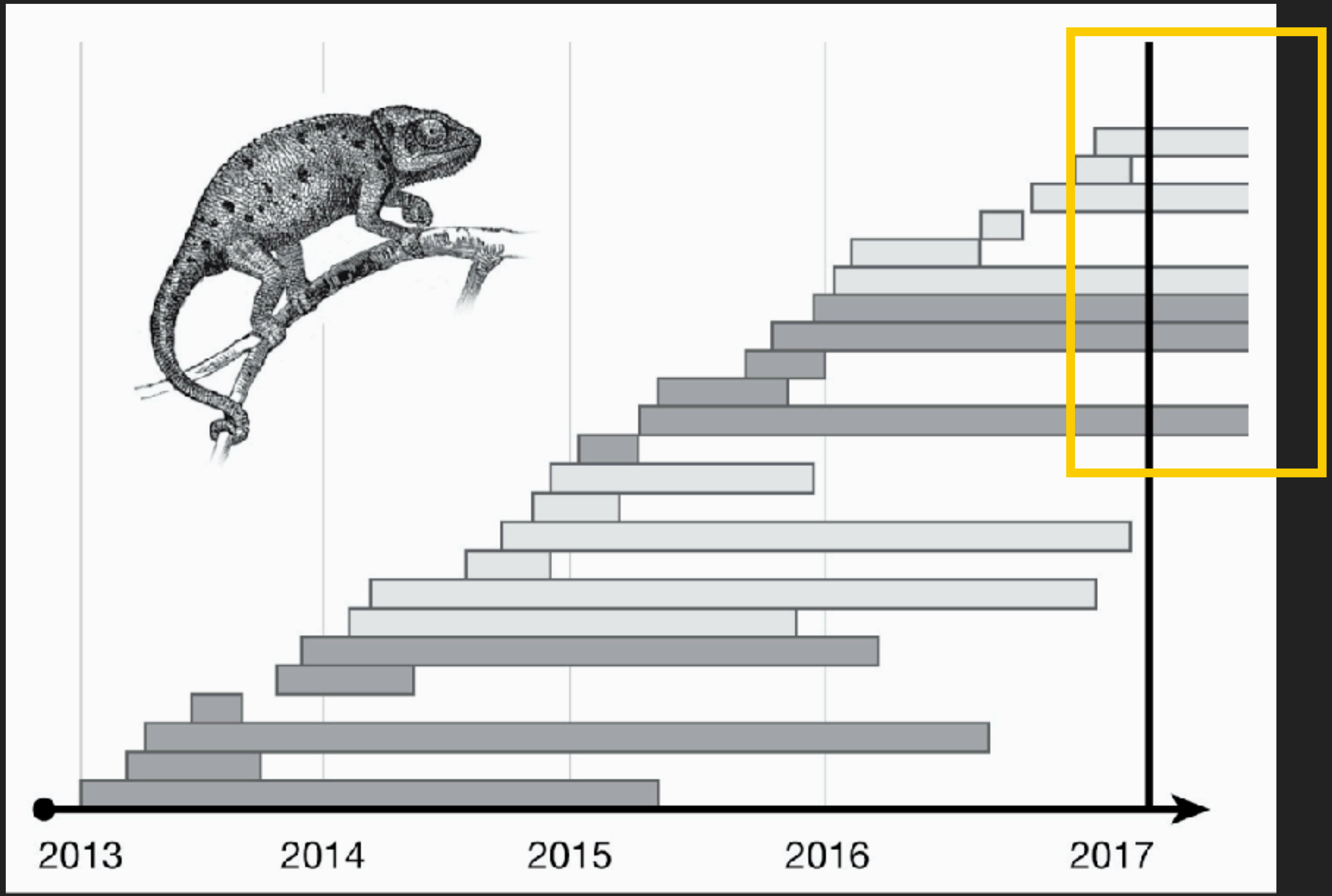
- Driven by an association between the very presence of the observer and the variable that the observer reports.

# Age of Death and Musical Genre



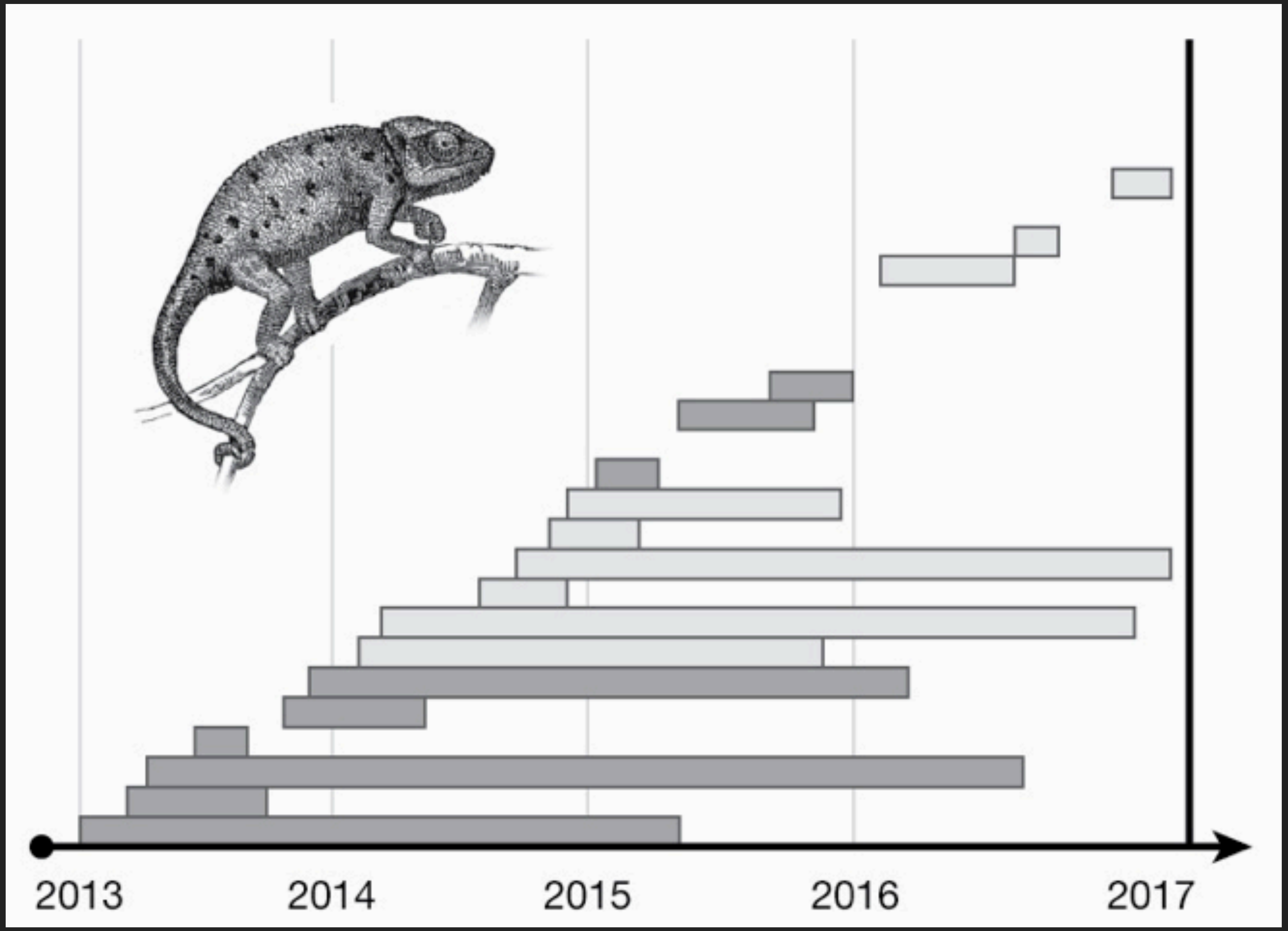

Rap and hip-hop musicians die at about half the age of performers in some other genres?

# Imagine You Are Tracking the Life Cycle of a Rare Chameleon on Madagascar




What to do about individuals not yet dead at the end of the study period?

# Imagine You Are Tracking the Life Cycle of a Rare Chameleon on Madagascar

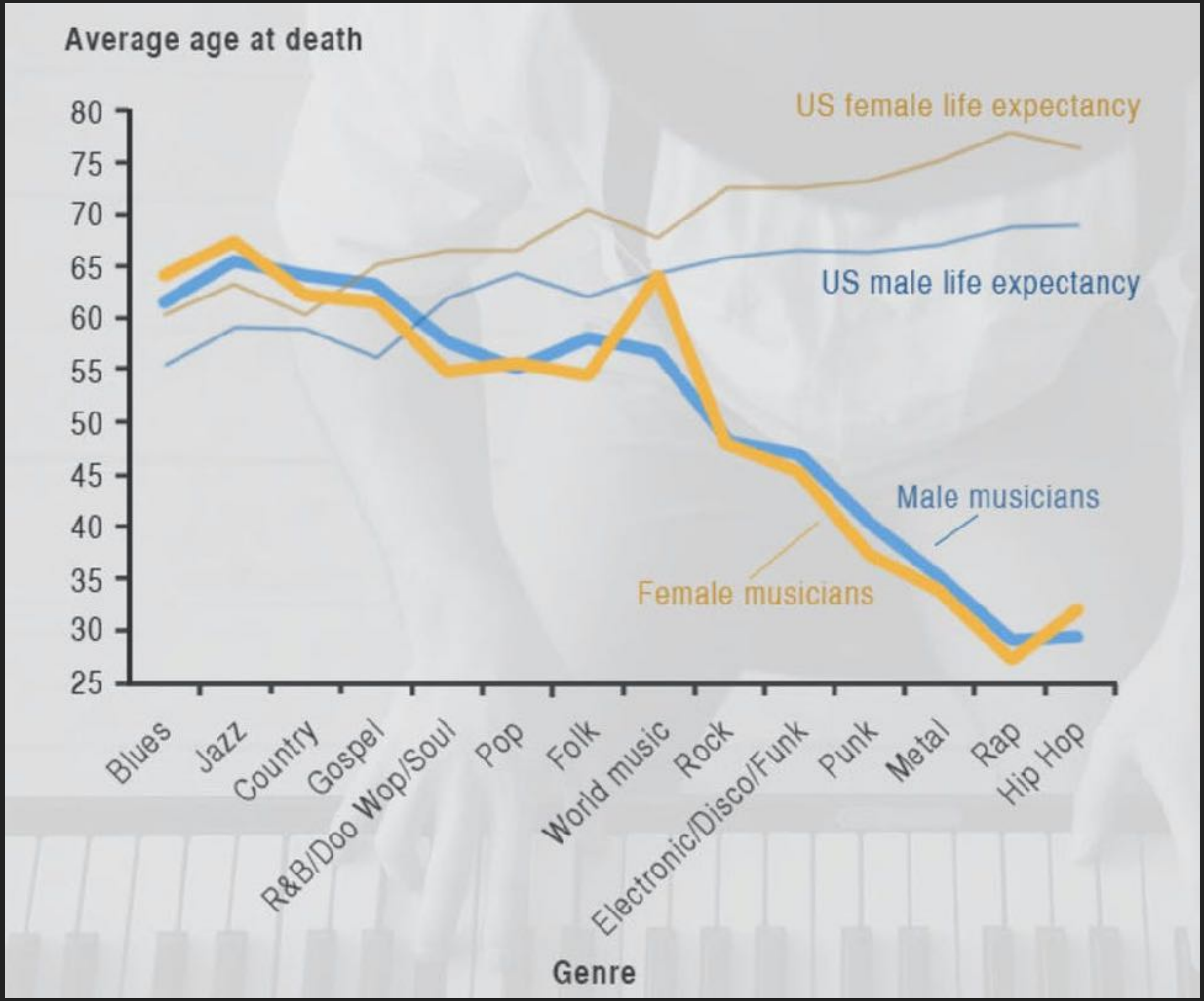



Maybe the safest thing to do is to throw out those individuals from your data set entirely?

–> Right-censoring your data

Misleading impression of mortality patterns.

# Age of Death and Musical Genre




Rap and hip-hop are new genres.

Most rap and hip-hop stars are still alive today, and thus omitted from the study.

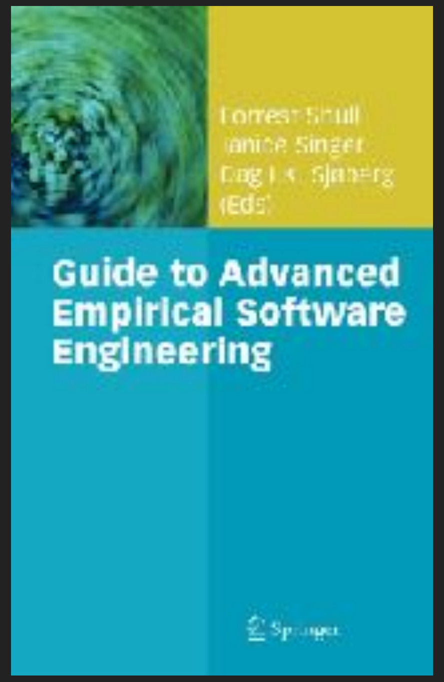
The only rap and hip-hop musicians who have died already are those who have died prematurely.

# Credits

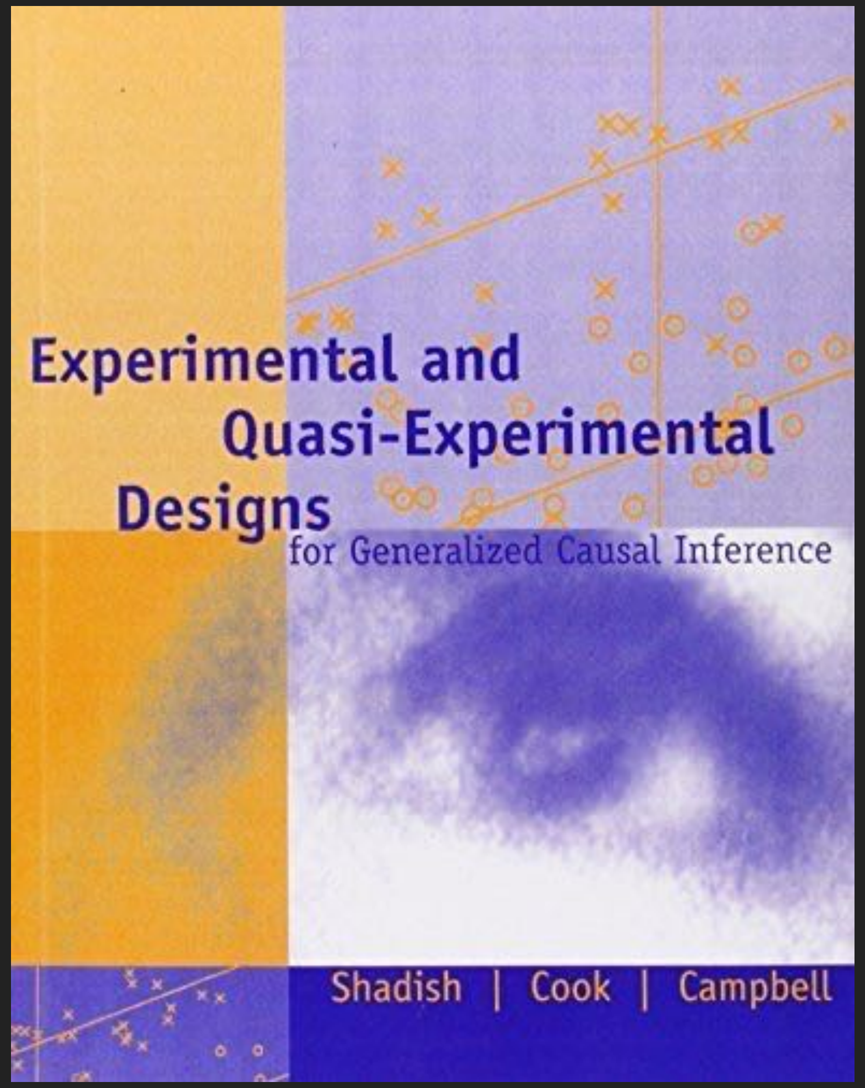
- Graphics: Dave DiCello photography (cover)
- Chapters from Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Publishing
  - Ch1: Experiments and generalized causal inference
  - Ch2: Statistical conclusion validity and internal validity
  - Ch3: Construct validity and external validity
  - Ch8: Randomized experiments
- Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media.
- Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (2007). Statistics.
- Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. In Seminars in Hematology (Vol. 45, No. 3, pp. 135-140). WB Saunders.
- Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann.
  - Ch 3: Experimental design
  - Ch 4: Statistical analysis
- MacKenzie, I. S. (2012). Human-computer interaction: An empirical research perspective.
  - Ch 6: Hypothesis testing
- Robertson, J., & Kaptein, M. (Eds.). (2016). Modern statistical methods for HCI. Cham: Springer.
  - Ch 5: Effect sizes and power analysis
  - Ch 13: Fair statistical communication
  - Ch 14: Improving statistical practice
- Kaptein, M., & Robertson, J. (2012). Rethinking statistical analysis methods for CHI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1105-1114).

# Read

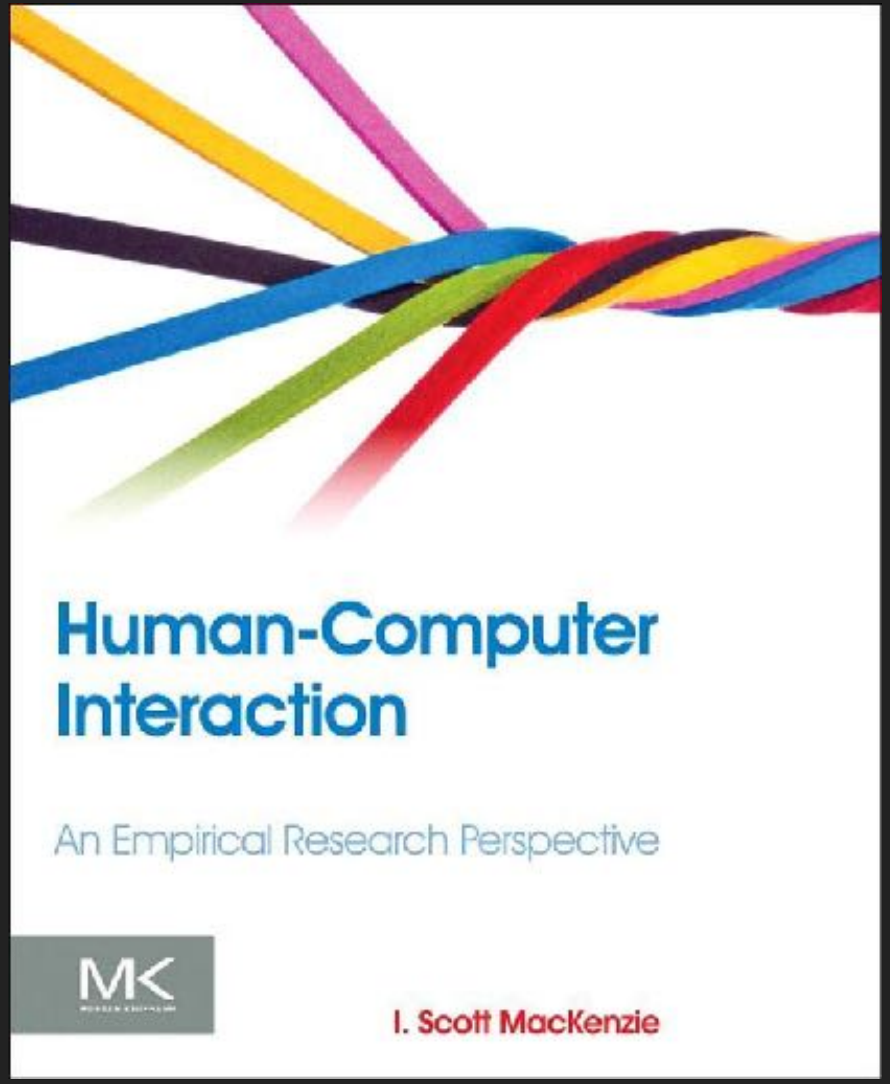

Ch 10 (Analysis and interpretation)



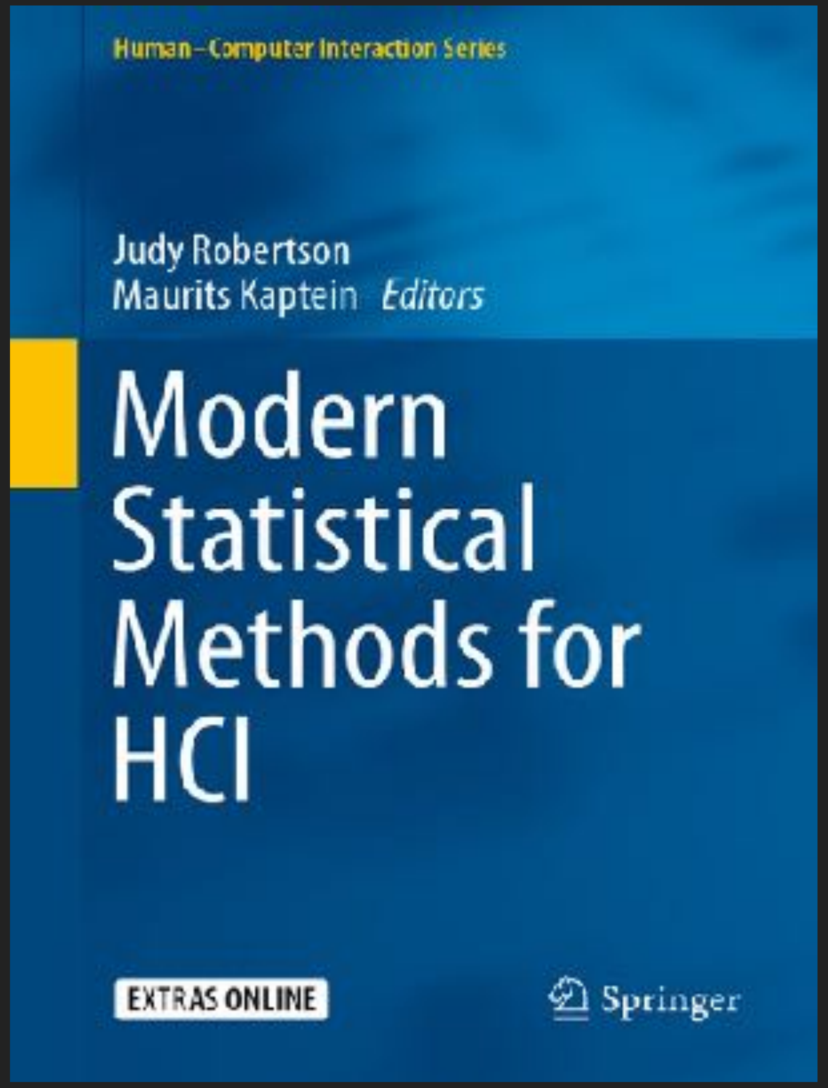

Ch 6 (Statistical methods and measurement)



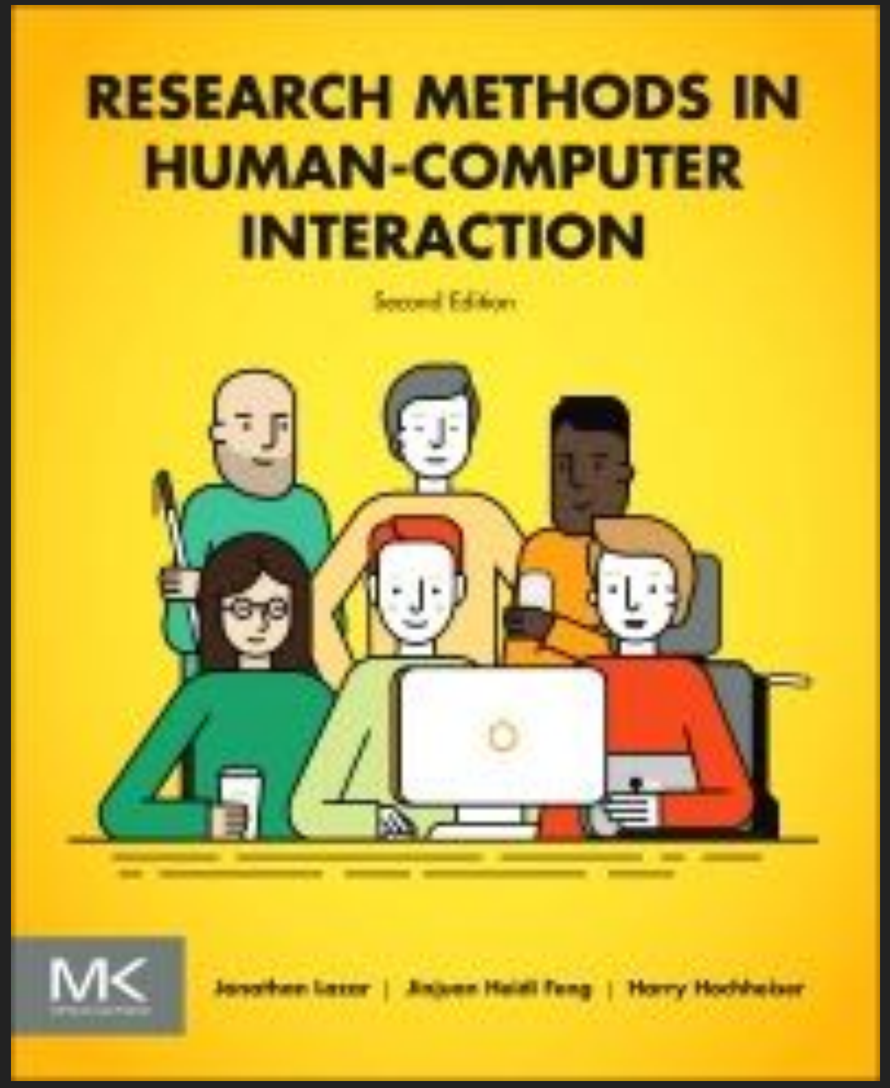

Ch 1 (Experiments and causality)
Ch 2 & 3 (Validity)
Ch 8 (Randomized experiments)




Ch 6 (Hypothesis testing)




Ch 5 (Effect sizes and power analysis)
Ch 13 (Fair statistical communication)
Ch 14 (Improving statistical practice)




Ch 3 (Experimental design)
Ch 4 (Statistical analysis)