P-values and statistical significance

Mohn Nutrition Research Laboratory
Solstrand, February 2020

Vegard Lysne
Clinical Dietitian, PhD
vegard.lysne@uib.no
www.vegardlysne.no

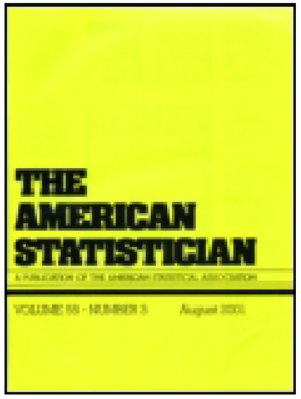
Q: Why do so many colleges and grad schools teach p = 0.05?

A: Because that's what the scientific community and journal editors use.

Q: Why do so many people still use p = 0.05?

A: Because that's what they were taught in college or grad school

Wasserstein & Lazar, The ASA' Statement on p-values, Am Stat 2016



Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

The ASA's Statement on p-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

ASA2016: Principle 1

”P-values can indicate how incompatible the data are with a specified statistical model”

The most common statistical model is the null hypothesis, postulating the absence of an effect, together with some assumptions

The P-value is a continuous measure of compatibility between your data and this model

Lower P = less compatibility

Wasserstein & Lazar, The ASA’ Statement on p-values, Am Stat 2016

ASA2016: Principle 2

”P-values does not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone”

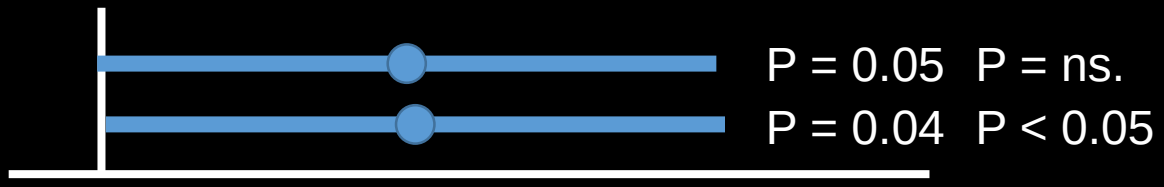
The P-value is a statement about the data in relation to a hypothetical explanation, not a statement about the explanation itself

Lower P = the data would be less likely to occur, assuming the model is true

Wasserstein & Lazar, The ASA’ Statement on p-values, Am Stat 2016

ASA2016: Principle 3

”Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold”



Bring other contextual factors into play, such as study design, quality of measurements, external evidence, validity of assumptions etc.

Wasserstein & Lazar, The ASA’ Statement on p-values, Am Stat 2016

ASA2016: Principle 4

”*Proper inference requires full reporting and transparency*”

Selective reporting = BAD

“Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis.”

Wasserstein & Lazar, The ASA’ Statement on p-values, Am Stat 2016

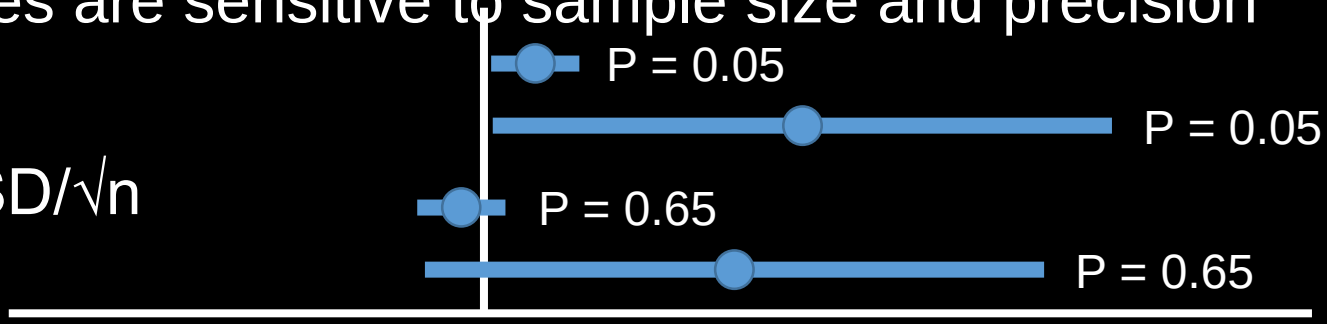
ASA2016: Principle 5

”A *p-value, or statistical significance, does not measure the size of an effect or the importance of a result*”

Statistical significance ≠ clinical relevance

P-values are sensitive to sample size and precision

SE = SD/√n



Wasserstein & Lazar, The ASA’ Statement on p-values, Am Stat 2016

ASA2016: Principle 6

”By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis”

Without context or other evidence, the P-value provide limited information

The calculation of P-values should not be the final goal of data analysis

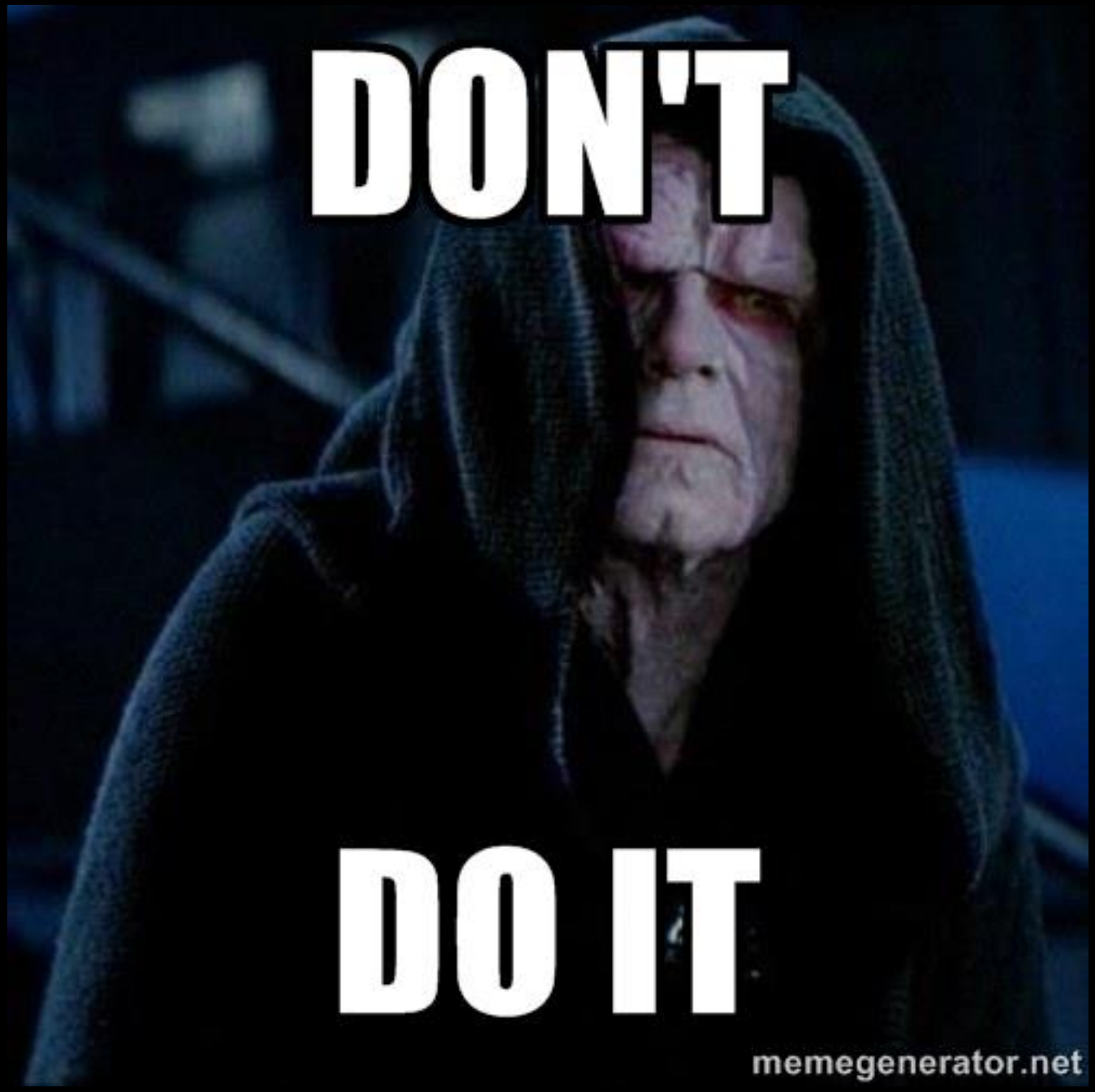
Wasserstein & Lazar, The ASA’ Statement on p-values, Am Stat 2016

ASA 2016: Conclusion

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.

No single index should substitute for scientific reasoning.

Wasserstein & Lazar, The ASA' Statement on p-values, Am Stat 2016


DON'T
DO IT
memegenerator.net

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Moving to a World Beyond "$p < 0.05$"

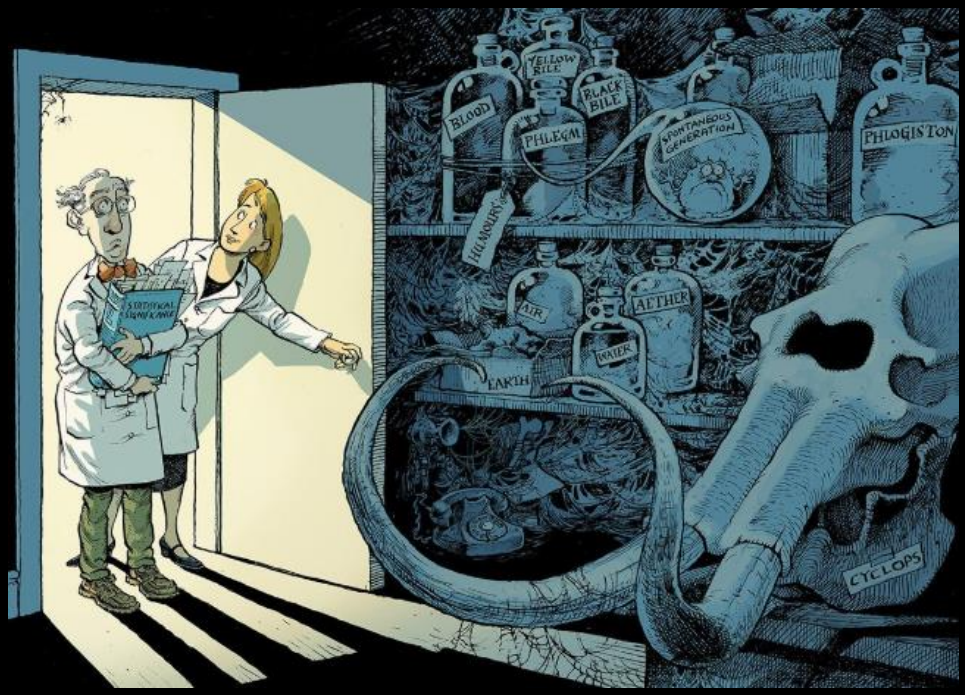
Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

The final don’t

Stop using the term “Statistically significant”!

Including all variants of:

- (not) statistically different
- Significant/nonsignificant
- Borderline significant/nonsignificant
- “P < 0.05” / “P > 0.05”
- *, **, ***, ns., etc.



Wasserstein, Shirm & Lazar, Moving to a World Beyond “p<0.05”, Am Stat 2019

The final don’t

Stop using the term “Statistically significant”!

1. Statistical and ordinary meaning of the word is different.
2. A label of ”statistcal significance” adds nothing to what is conveyed by the P-value already
3. As the P-value doesn’t reveal the plausibility, presence, truth, or importance of an effect, neither does “statistical significance”
4. Dichotomization into “significant” and “not significant” contributes to selective reporting and publication bias

Wasserstein, Shirm & Lazar, Moving to a World Beyond “p<0.05”, Am Stat 2019

The final don’t

Stop using the term “Statistically significant”!

The problem is the arbitrary categorization, not the number of categories. Categorizing into any number of groups is problematic

We must avoid arbitrarily categorizing of other statistical measures

“In sum: Statistically significant – don’t say it and don’t use it.”

Wasserstein, Shirm & Lazar, Moving to a World Beyond “p<0.05”, Am Stat 2019

What is a world beyond “p<0.05”?

- A world where researchers are free to treat “p=0.051” and “p=0.049” as not being categorically different
- A world where authors are not constrained to selectively publish their results based on a single magic number
- A world where studies with “P<0.05” and “P>0.05” are not automatically in conflict
- A world with fewer false alarms and fewer overlooked discoveries
- A world where researchers are free to communicate all their findings in all their glorious uncertainty

Wasserstein, Shirm & Lazar, Moving to a World Beyond “p<0.05”, Am Stat 2019

Case 1: D-vitamin supplements

Aim:

Does D-vitamin supplements influence the risk of cancer?

Results:

Unadjusted Cox-regression: **HR 0.70 (0.47 – 1.02), p = 0.06**

Conclusion:

"...supplementation with vitamin D3 and calcium compared with placebo **did not** result in a significantly lower risk of all-type cancer at 4 years."

Lappe J et al. Effect of Vitamin D and Calcium Supplementation on Cancer Incidence in Older Women. JAMA 2017

Case 1: D-vitamin supplements

Inappropriate Reliance on P-values in Medical Research

To the editor:

Lappe et al. (1) reported that women receiving vitamin D and calcium supplementation had 30% lower cancer risk than women receiving placebo after four years (hazard ratio (HR)=0.70, 95% confidence interval (CI): 0.47 to 1.02). Remarkably, they interpreted this result as indicating no effect. So did the authors of the accompanying editorial (2), who described the 30% lower risk for cancer as "the absence of a clear benefit," because the P-value was 0.06. Given the expected bias toward a null result in a trial that comes from non-adherence coupled with an intent-to-treat analysis (3), the interpretation of the authors and editorialists is perplexing. The warning issued last year by the American Statistical Association (ASA) (4) about this type of misinterpretation of data should be embraced by researchers and journal editors. In particular, the ASA stated: "Scientific conclusions …should not be based only on whether a p-value passes a specific threshold." Editors in particular ought to guide their readership and the public at large to avoid such mistakes and foster more responsible interpretation of medical research.

Case 2: Antidepressant and autism risk

Aim:

Does antidepressant use during pregnancy influence the risk of autism spectrum disorder of the child?

Results:

Multivariate Cox-regression:

HR 1.59 (1.17 – 2.17)

High-dimensional propensity score adjustment:

HR 1.61 (0.997 – 2.59)

Brown HK et al. Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. JAMA 2017

Case 2: Antidepressant and autism risk

Results:

Multivariate Cox-regression: **HR 1.59 (1.17 – 2.17)**

High-dimensional propensity score: **HR 1.61 (0.997 – 2.59)**

Conclusion:

"...exposure compared with no exposure **was not** associated with autism spectrum disorder in the child. Although a causal relationship cannot be ruled out, the previously observed association may be explained by other factors."

Brown HK et al. Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. JAMA 2017

Case 2: Antidepressant and autism risk

Medscape Wednesday, February 12, 2020

Antidepressants in Pregnancy: No Link to Autism, ADHD

Batya Swift Yasgur, MA, LSW

April 21, 2017

2 Read Comments

Use of antidepressants before and during pregnancy does not cause autism, or attention-deficit/hyperactivity disorder (ADHD) new research shows.

Not significant ≠ not different

Absence of evidence is not evidence of absence

The p-value is the probability of **data**, given the null hypothesis

- How likely are my data if the null hypothesis is true?

Every P<1 indicate some degree of incompatibility

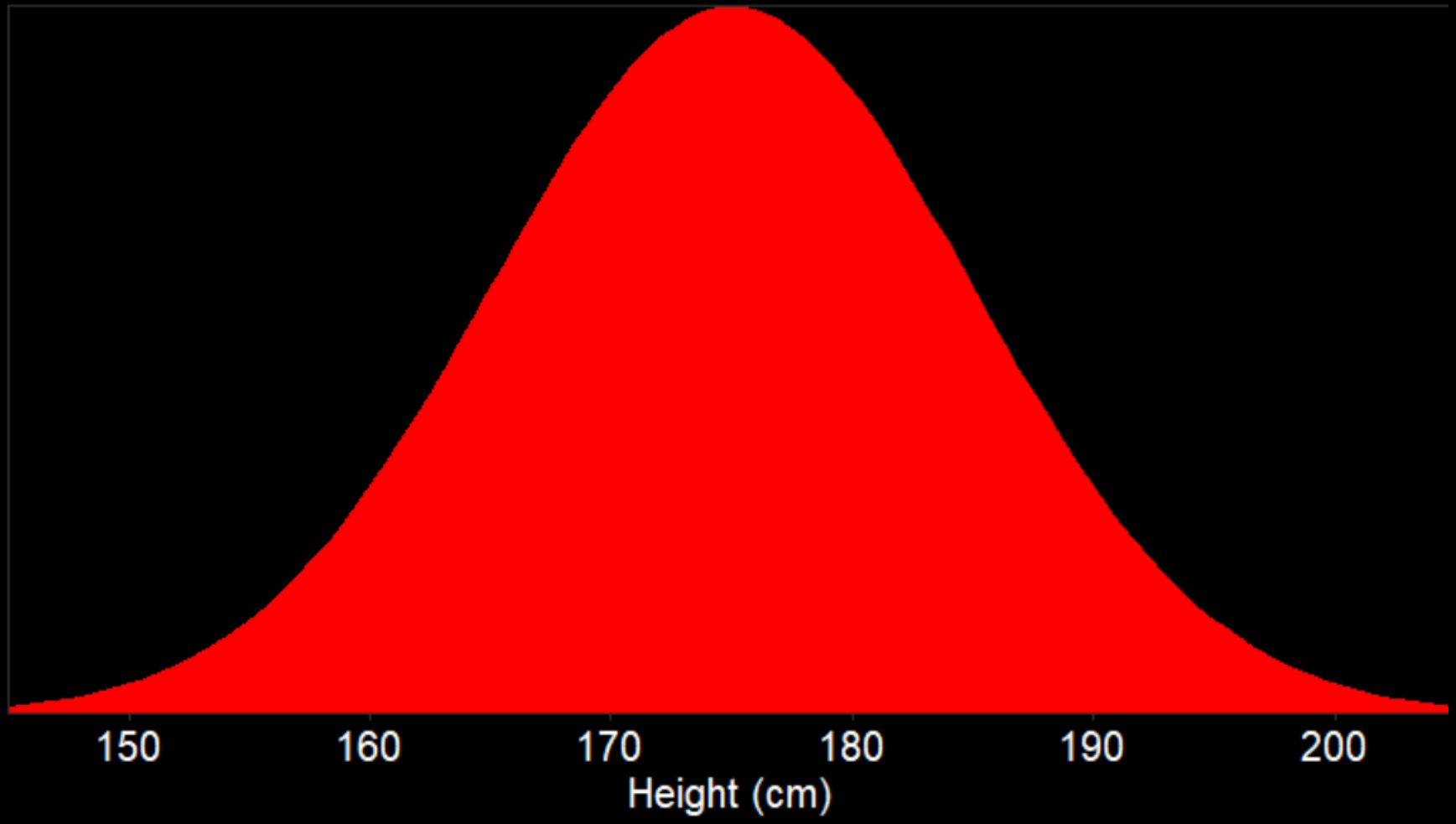
- A high P-value indicate that the data is compatible with the null hypothesis
- It does not indicate that the null hypothesis is the best – or the correct – model

True null hypothesis: The data was generated by random

Null hypotheses are specific random number generators

When the null hypothesis is true, the groups belong to the same distribution of the outcome, and all observed differences are generated by chance

Simulation




WAIT, I'M CONFUSED...
WHAT DO WE DO NOW???
makeameme.org

Zad Chow & Sander Greenland

Compatibility and Surprise, Not Confidence and Significance arXiv: 1909.08579 | Version: 21 Sept. 2019

APPLIED STATISTICS

Semantic and Cognitive Tools to Aid Statistical Inference:
Replace Confidence and Significance by Compatibility and Surprise

Zad R. Chow[1] | Sander Greenland[2]

Emphasize Unconditional Descriptions arXiv: 1909.08583 | Version: 21 Sept. 2019

APPLIED STATISTICS

To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics

Sander Greenland[1] | Zad R. Chow[2]

Suggestions

1. We should replace overconfident terms like “(non)significance” and “confidence interval” with more modest descriptions like “high (low) compatibility” and “compatibility interval”
2. We should use alternative ways of looking at the P-value, such as the S-value
3. Single P-values, S-values or intervals should be replaced with tables or graphs showing the results for relevant test hypoteses
4. Usual interpretations of statistical outputs is misleading because it condition on background assumptions. Replace with unconditional descriptions.

Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

2. S-values

S = -log(p) or log(1/p)

Other names: Shannon information, surprisal, logworth

Addresses a scaling problem with P-values

- P is restricted between 0 and 1
- Small changes in P near 1 and near 0 mean different things

S is measured on an absolute scale, calibrated to an intuitive physical mechanism

Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

2. S-values

$-\log_2(p)$ = bits of information

Calibrated towards a mechanism producing a binary outcome, e.g. a coin toss

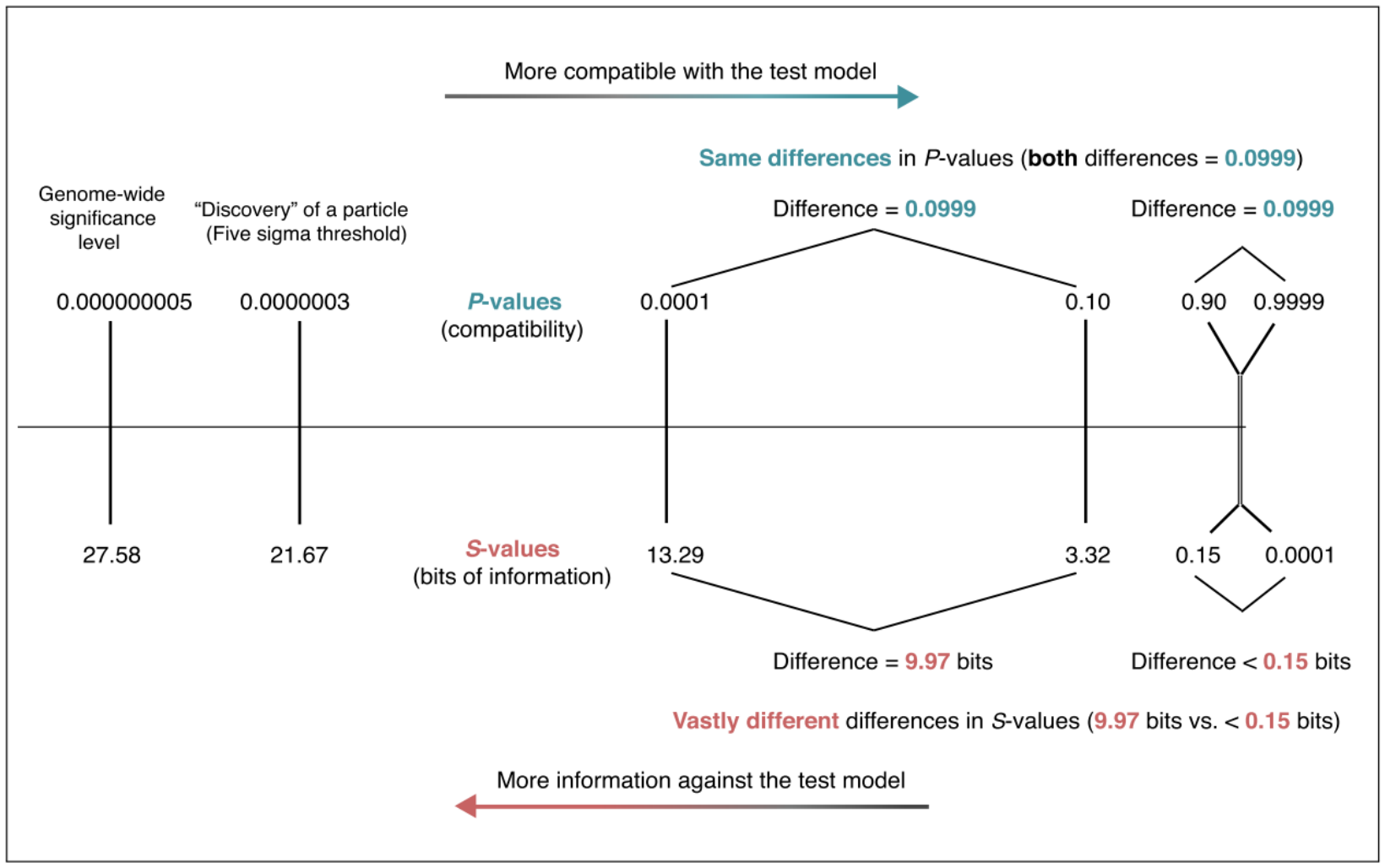
Heads = success, Tails = failure

X bits = X consecutive heads

P=0.05, S = $-\log_2(0.05)$ = 4.32

Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

S-values



Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

1. Semantics

Significance → compatibility

- P-values is a measure of compatibility

Confidence Intervals → Compatibility Intervals?

- Our data can be tested towards hypotheses other than the null
- A 95% CI display the range of test hypotheses which would yield P>0.05, or S<4.32
- The CI contains a range of values more compatible with the data than the values outside the CI

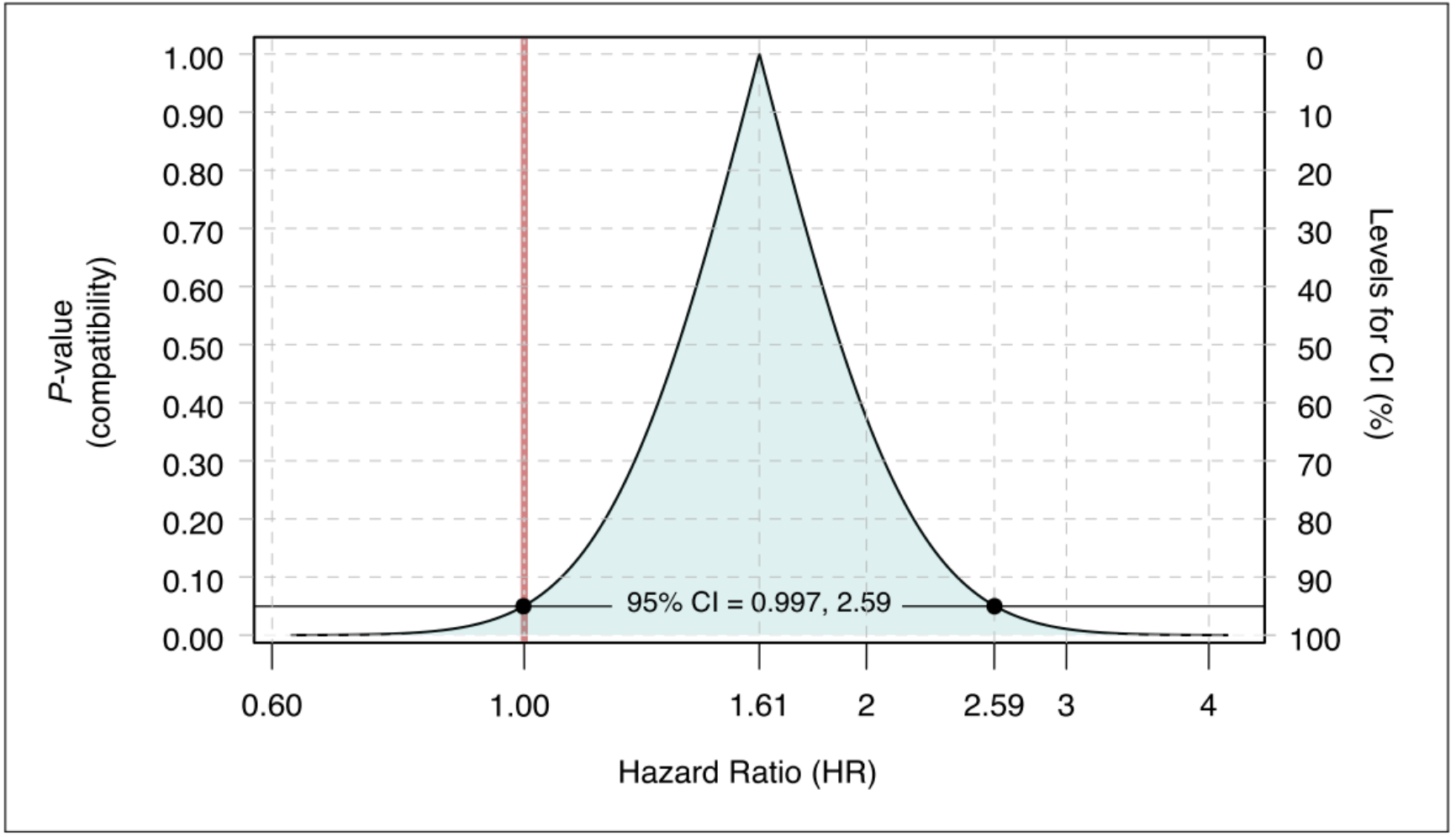
Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

3. Show a range of values

Test Hypothesis (H)	*P*-value (compatibility)	*S*-value (bits)
Halving of hazard (HR = 0.5)	1.6×10^{-6}	19.3
No association (null) (HR = 1)	0.05	4.31
Point estimate (HR = 1.61)	1	0.00
Doubling of hazard (HR = 2)	0.37	1.42
Tripling of hazard (HR = 3)	0.01	6.56
Quintupling of hazard (HR = 5)	3.3×10^{-6}	18.2

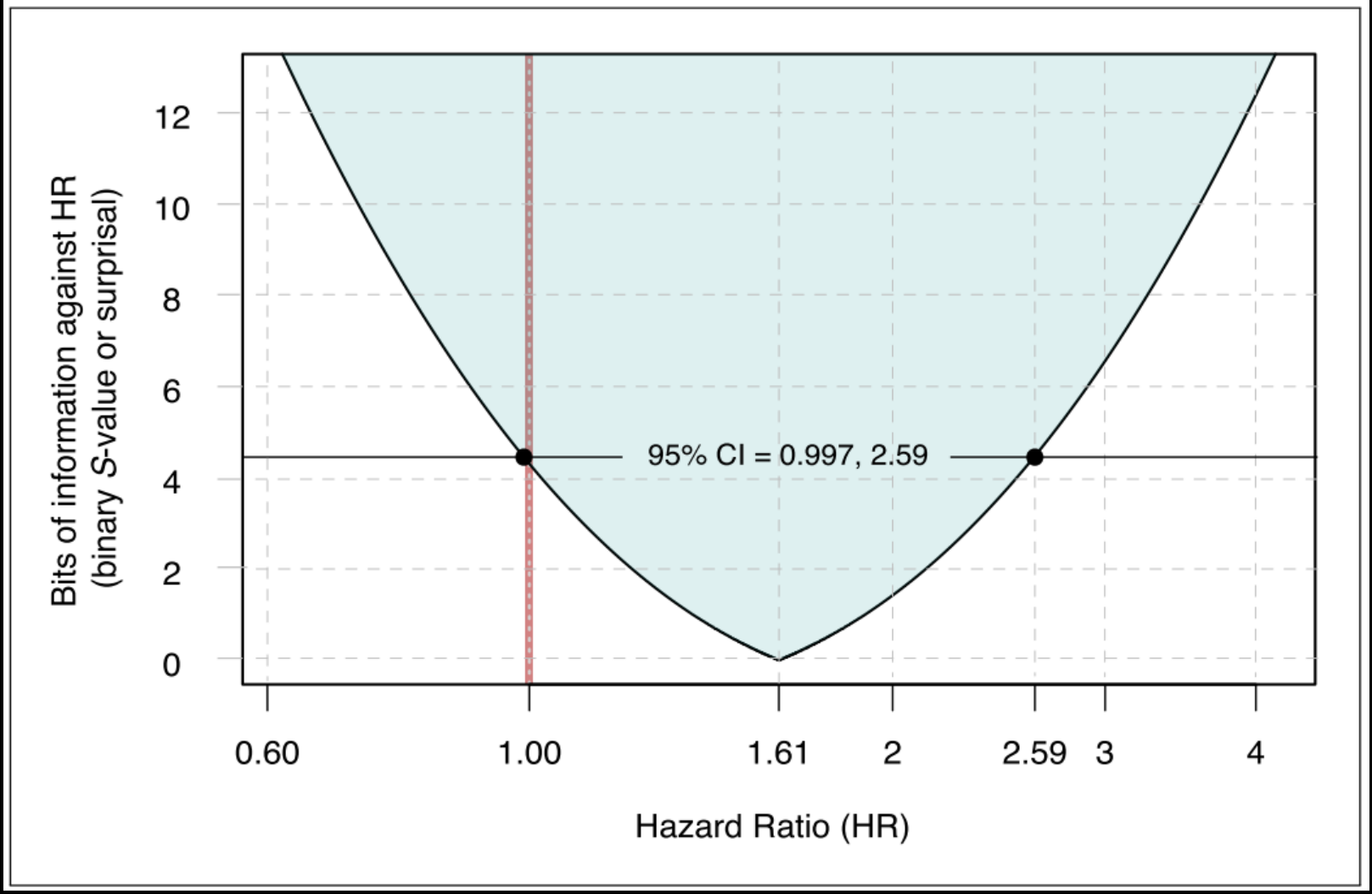
Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

3. Show a range of values



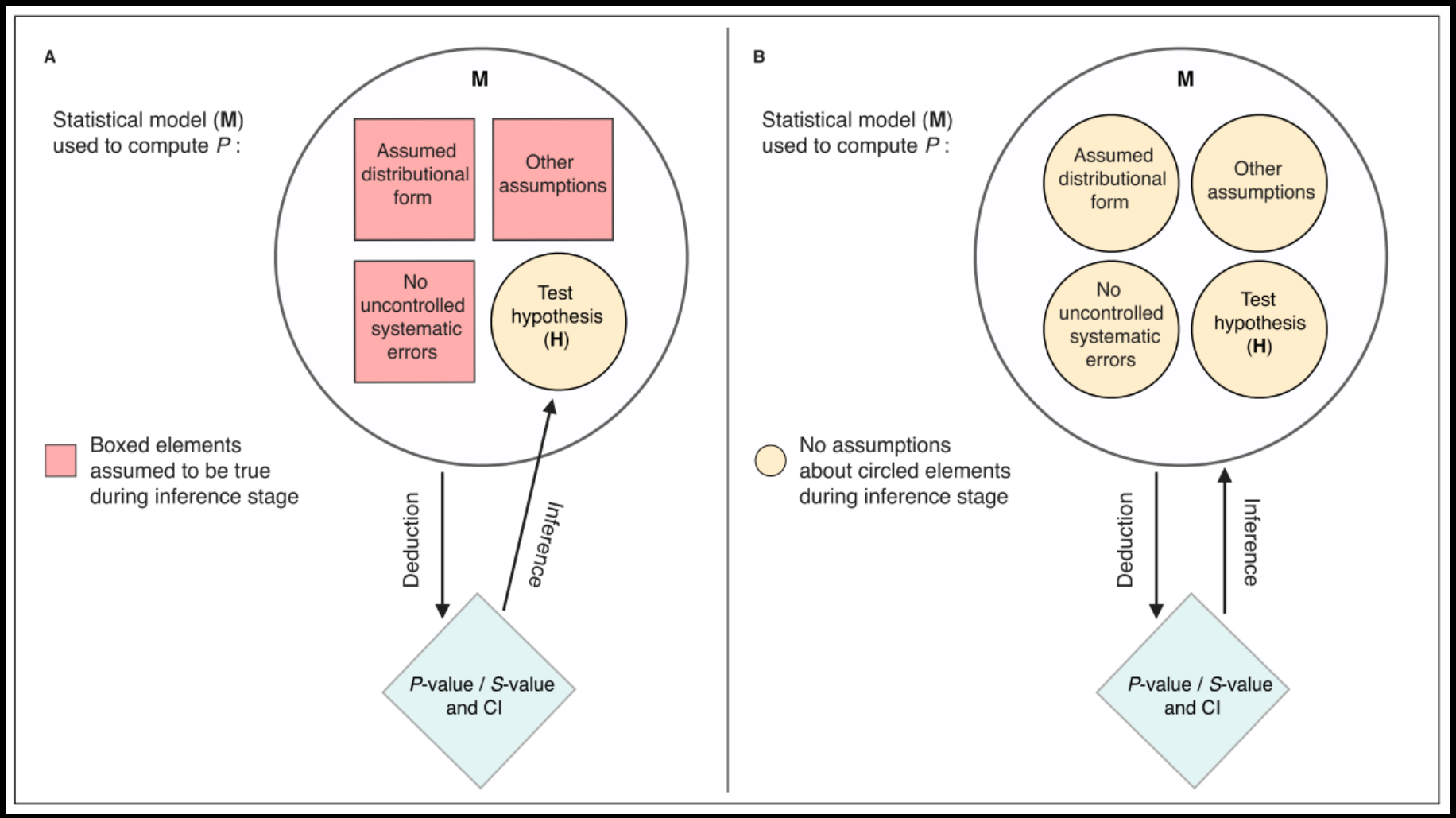
Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

3. Show a range of values



Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

4. Unconditional description of statistics



Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

Case 2: Antidepressant and autism risk

Results:

Multivariate Cox-regression: **HR 1.59 (1.17 – 2.17)**

High-dimensional propensity score: **HR 1.61 (0.997 – 2.59)**

Suggested reporting:

"After HDPS adjustment for confounding, a 61% hazard elevation remained; however, under the same model, every hypothesis from no elevation up to a 160% hazard increase had $p \geq 0.05$; Thus, while quite imprecise, these results are most consistent with previous observations of a positive association between antidepressant exposure and subsequent ASD (although the association may be partially or wholly due to uncontrolled biases)."

Chow Z & Greenland S. Semantic and Cognitive Tools to Aid Statistical Inference. https://arxiv.org/abs/1909.08579
Greenland S & Chow. To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics. http://arxiv.org/abs/1909.08583

1. Wasserstein & Lazar. **The ASA's Statement on p-Values: Context, Process, and Purpose.** Am Stat 2016
2. Wasserstein, Schirm & Lazar. **Moving to a World Beyond "p < 0.05"**. Am stat 2019
3. Greenland S & Chow Z. **To Aid Statistical Inference, Emphasize Unconditional Descriptions of Statistics**. http://arxiv.org/abs/1909.08583
4. Chow Z & Greenland S. **Semantic and Cognitive Tools to Aid Statistical Inference: Replace Confidence and Significance by Compatibility and Surprise**. http://arxiv.org/abs/1909.08579
5. Greenland S, Senn S, Rothman K et al. **Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations**. Eur J Epidemiol 2016
6. Gelman A & Loken E. **The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.** http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
7. Lash T. **The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing.** Am J Epidemiol 2017
8. Lakens D & Etz A. **Too True to be Bad: When Sets of Studies With Significant and Nonsignificant Findings Are Probably True.** Soc Psychol Personal Sci 2017
9. Gelman A & Stern H. **The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant.** Am Stat 2006
10. Tong C. **Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science.** Am Stat 2019
11. McShane B, Gal D, Gelman A et al. **Abandon Statistical Significance.** Am Stat 2019
12. Amrhein V, Trafimow D & Greenland S. **Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication.** Am Stat 2019
13. Colin-Jagerman R & Cumming G. **The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known**. Am Stat 2019
14. Hubbard R, Haig B & Parsa R. **The Limited Role of Formal Statistical Inference in Scientific Inference**. Am Stat 2019
15. www.datamethods.org, www.andrewgelman.com, www.lesslikely.com, #statstwitter