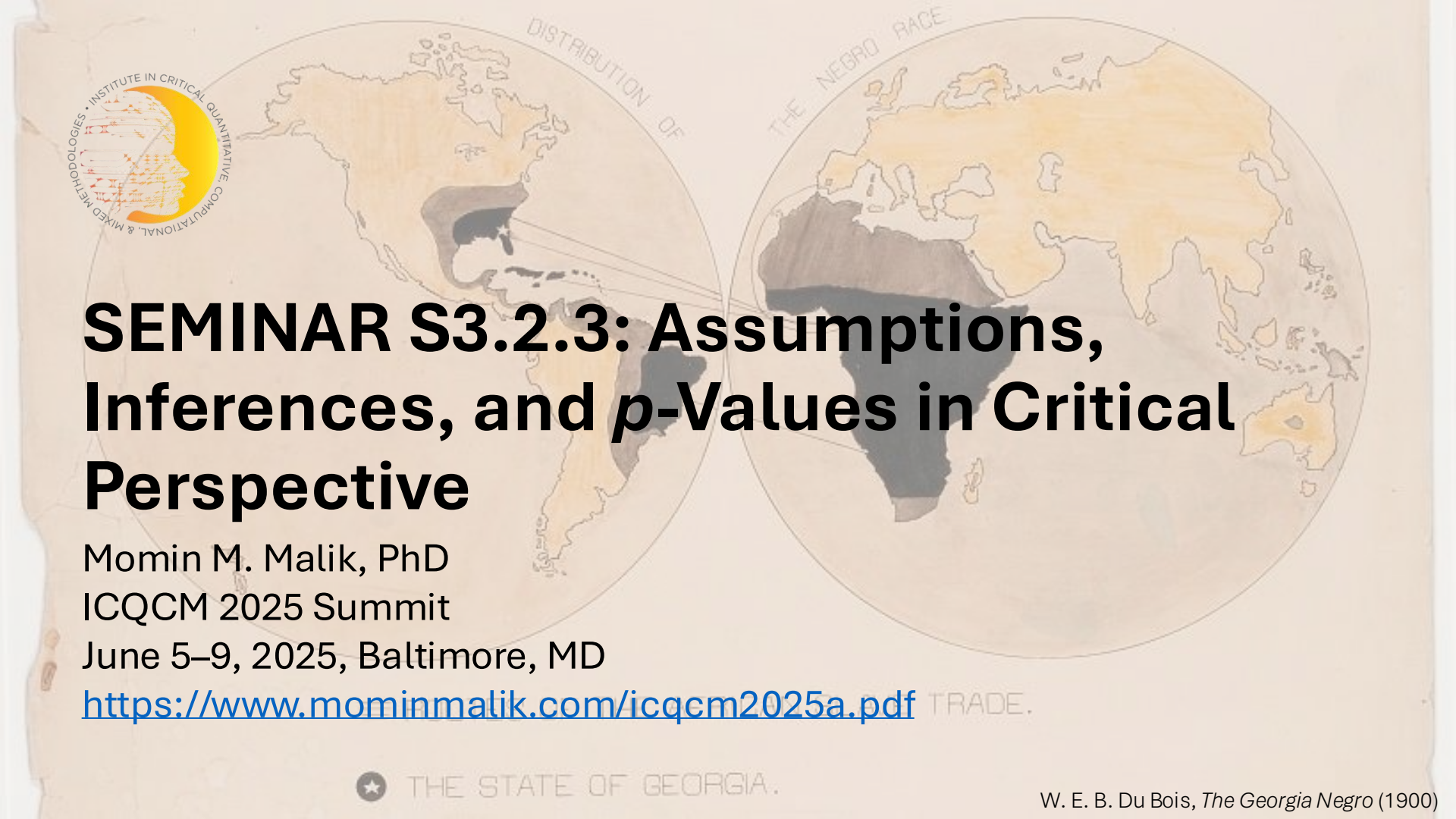


SEMINAR S3.2.3: Assumptions, Inferences, and *p*-Values in Critical Perspective

Momin M. Malik, PhD

ICQCM 2025 Summit

June 5–9, 2025, Baltimore, MD

https://www.mominmalik.com/icqcm2025a.pdf

W. E. B. Du Bois, *The Georgia Negro* (1900)

Outline

- What are *p*-values?
  - "Lady tasting tea"
- What is statistical inference?
  - Central limit theorem
  - Bayesian inference
- More about *p*-values
  - Distribution under the null
  - *p*-hacking
  - Multiple comparisons
- What are we assuming?

tl;dr

- Everything is terrible and nothing works. Statistics at makes no sense, and is maybe just one big con job (and its legitimacy is a conspiracy)
- Your options are:
  - Just do what others do, misusing methods in the accepted ways to get on with doing quantitative analysis
  - Engage in the messiness of the underlying theory and experience utter despair
  - Give up and do qualitative analysis in a world that doesn't respect it

Inference vs causality

- All of this is only about inferring whether or not there is an association, and not whether or not that association is causal!
- Causality is another layer that is just as big of a mess: nothing works.
  - In observational data, not even the most mathematically sophisticated techniques can ever overcome the possibility of unobserved confounders (Hu, 2021)
  - Experiments can get "causality" may not represent what happens in the world (ecological validity)

Activity: Standard deviation vs. standard error

- Turn to 1-2 people sitting close to you
- What is the difference between a *standard deviation* and a *standard error*?
- (This is something you will supposedly have learned, but it took me years to understand and appreciate it!)

One possible answer

- Standard *deviation* is a measure of "haphazard variability"
- Standard error is an expression of our uncertainty

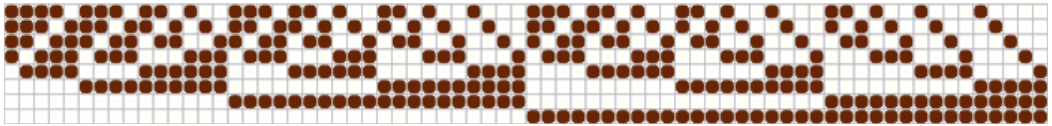
- The two are linked via the Central Limit Theorem; the standard error is often calculated as the *sample* standard deviation divided by the square root of (n – 1).

What are *p*-values?

"Lady tasting tea": *p*-value origin story

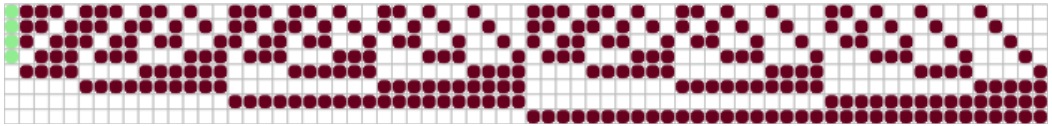
- R. A. Fisher (statistician and geneticist; but also eugenicist and misogynist and tobacco apologist) gave this example in *The Design of Experiments* (1935), inspired by real events
- He was at a party, where phycologist (person who studies algae) Muriel Bristol claimed to be able to tell whether the tea or the milk was added first to a cup
- Bristol's future husband, William Roach, suggested testing her: give her four cups with milk added first, and four cups with tea added first

“Lady tasting tea”: *p*-value origin story



- There are 70 different ways the cups could be arranged
- How many does Bristol need to correctly guess to justify believing her claim?
- Let us say she knows it’s 4 cups in each condition, and let us say she will always guess 4 of each

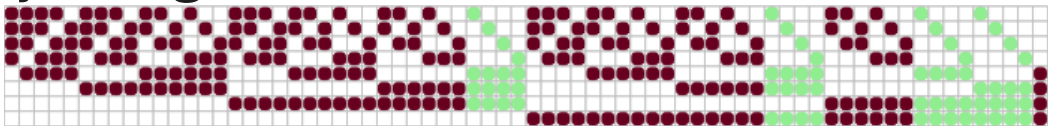
“Lady tasting tea”: *p*-value origin story



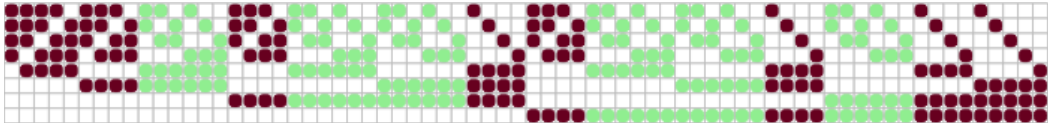
- Let’s say the true pattern is the first column; 4x tea first, then 4x milk first (doesn’t matter which is correct; probabilities are the same, I just choose a specific one to help illustrate)
- One possible set of guesses (the first column) corresponds to getting them all correct; if guessing randomly, 1/70 chance of getting it correct, or 1.43%
- One possible set of guesses (the last row) corresponds to getting them all incorrect. Same chance; 1.43%
- Otherwise, either get 2 correct, 4 correct, or 6 correct

“Lady tasting tea”: *p*-value origin story

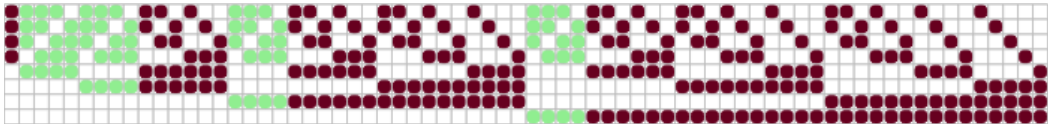
- 16 ways to get 2 correct: 16/70 = 22.86% chance



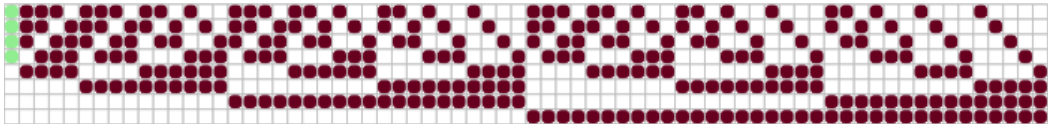
- 36 ways to get 4 correct: 36/70 = 51.43% chance



- 16 ways to get 6 correct: 16/70 = 22.86% chance

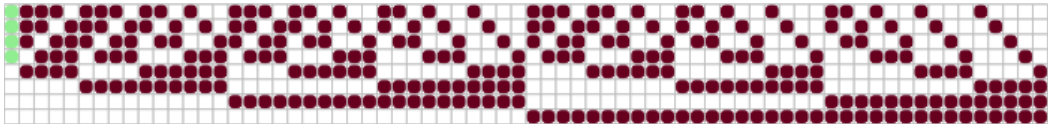


“Lady tasting tea”: *p*-value origin story



- How many cups does the lady need to correctly guess for us to believe her?
- Fisher said, *all of them*; there’s only a 1.43% chance of that happening at random, because there’s a 1.43% + 22.86% chance of guessing at least 6 out 8 correctly.
- **That 1.43% is the p-value!** 0.0143 < 0.05 (but we could choose a different cutoff for the *p*-value)

"Lady tasting tea": *p*-value origin story



- This is how probability is used as a way of finding *how likely it is that what we observe is just noise/randomness*
- This is also what *p*-values mean: *the probability that the lady correctly labeled all 8 cups even though she doesn't actually have the skill to tell*
- Formally, *p*-values are "the probability that a test statistic at least as extreme as the one observed would occur if the null hypothesis were true". **It cannot be simplified from that.**

“Lady tasting tea”: *p*-value origin story

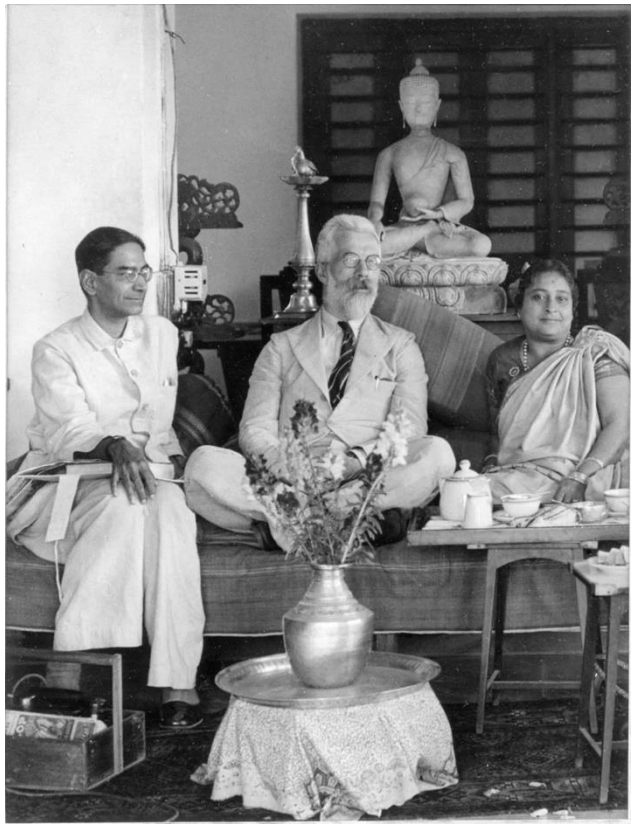
- Unfortunately, this is a misleadingly simple example, because here we can enumerate all possible combinations, and usually that doesn’t make sense
- We can’t, for example, enumerate all possible ways that hundreds of children’s test scores could come out based on whether the children do or do not get benefits from a government assistance program
- So we use more complex math (the normal distribution) to represent what “just noise” looks like

What *p*-values are NOT

- A *p*-value is NOT the probability that the hypothesis is true given the data
- A *p*-value is the probability that we would observe the test statistic (the data) we did if the hypothesis were true
- It is $P(D \mid H)$, not $P(H \mid D)$. Will return to later

What is statistical inference?

What is statistics?



"briefly, and in its most concrete form, the object of statistical methods is the reduction of data."

- R. A. Fisher, 1922, "On the mathematical foundations of theoretical statistics"

Fisher: Committed eugenicist (even until his death in 1962), raging misogynist, paid shrill for the tobacco industry… and one of if not the greatest inventors of statistical practice, (and supporter of the elite statistics community in India)

P.C. Mahalanobis Memorial Museum and Archives, Indian Statistical Institute, Kolkata, and Rare Books and Manuscripts, University of Adelaide Library

What is statistics?

"A quantity of data, which usually ***by its mere bulk is incapable of entering the mind****, is to be replaced by* ***relatively few quantities*** *which shall* ***adequately represent the whole****, or… as much as possible… of the* ***relevant information*** *contained in the original data."*

– R. A. Fisher, 1922, "On the mathematical foundations of theoretical statistics"

What is statistics?

- A “statistic” (singular) is defined as *a function of the data*. (Historically, statistics comes from “political arithmetic”, think censuses, and “stat” comes from “state”)
- The discipline of Statistics is about defining “relevant information,” and finding functions to capture it.
- *How* does it do so?

How statistics reduces data

Statistics: *The use of probability as a model of variability in the world*.

"Probability is used in two distinct, although interrelated, ways in statistics, phenomenologically to describe haphazard variability arising in the real world and epistemologically to represent uncertainty of knowledge."

– D. R. Cox, 1990, "Role of models in statistical analysis"

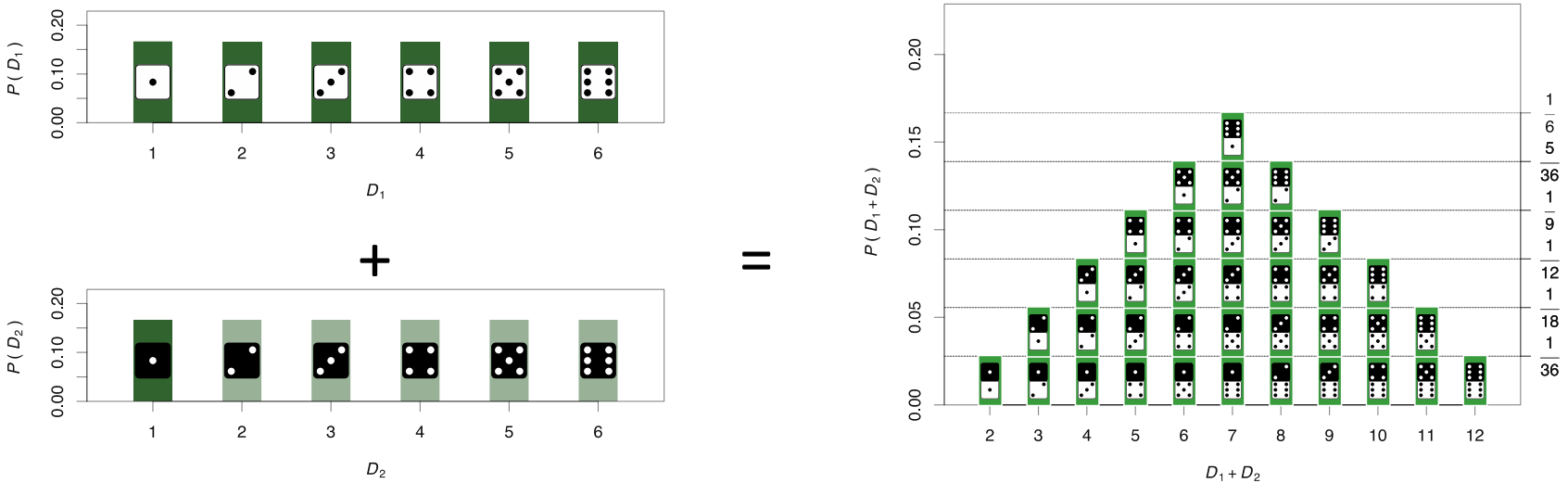
(See also: "aleatory" vs. "epistemic" uncertainty)

Connection to probability

- Distributions of data show us the “haphazard variability” of the world around a central tendency (Cox, 1990)
- Statistics sees them as representing an underlying *probability distribution*

Probability distribution: “shape” of randomness

- A fair 6-sided die can have 1 of 6 outcomes, each equally likely. It has a “flat” (uniform) distribution.
- But when we add two such dice, we get a different “shape” of the randomness!



Probability is an abstraction of *gambling*

- "It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge."
  - Pierre-Simon Laplace, *Théorie Analytique des Probabilitiés* (1812)

Likelihood principle and inference

- The “likelihood principle” is what connects probability (which is a field of mathematics) to data
- Instead of saying, what the probabilities are of a set of outcomes, we say: given a set of realizations, what is the most “likely” underlying distribution (or numerical summaries of that distribution)?
- Statistical inference is the process of using data to make ‘inferences’ about a hypothetical, unobservable, “true” underlying process (whether or not such a thing exists), represented with the language of probability

All the different moving pieces

Motivation

Interpretation

Solution

$\mathbf{X}\boldsymbol{\beta} = \log(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}})$, so $\boldsymbol{\mu} = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$. We can write this as $\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})$.

Probability, linear algebra

$$\mathbb{P}(\mathbf{y}|\mu) = \prod_{i=1}^{n} \mu^{y_i}(1-\mu)^{1-y_i}$$

Likelihood principle

$$\mathbb{P}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \mu(\mathbf{x}_i, \boldsymbol{\beta})^{y_i}(1-\mu(\mathbf{x}_i, \boldsymbol{\beta}))^{1-y_i}$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \mu(\mathbf{x}_i, \boldsymbol{\beta})^{y_i}(\mathbf{1}-\mu(\mathbf{x}_i, \boldsymbol{\beta}))^{1-y_i}$$

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^T \log(\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})) + (\mathbf{1}-\mathbf{y})^T \log(1-\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}))$$

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^T \log\left(\frac{\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})}{\mathbf{1}-\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})}\right) + \mathbf{1}^T \log(\mathbf{1}-\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}))$$

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^T \log\left(\frac{\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}}{1-\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}}\right) + \mathbf{1}^T \log\left(1-\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}\right)\Big]$$

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{1}^T \log(\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta}))$$

Solution

We need both the gradient and Hessian for the IRLS updates.

Linear algebra, convex optimization

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\left(\mathbf{y} - \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}\right) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta})$$

Standard errors and inference

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \frac{-\partial(\mathbf{X}^T\boldsymbol{\mu}(\mathbf{X}, \boldsymbol{\beta}))}{\partial \boldsymbol{\beta}^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X}$$

where $\mathbf{W}$ is a diagonal matrix whose ith element is $\mu(\mathbf{x}_i, \boldsymbol{\beta})(1-\mu(\mathbf{x}_i, \boldsymbol{\beta})$, which I am getting from *The Elements of Statistical Learning*. Then, the IRLS update is derived from Newton's method,

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} \leftarrow & \boldsymbol{\beta}^{(t)} + (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mu(\mathbf{X}, \boldsymbol{\beta}^{(t)})) \\
&= (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{X})\boldsymbol{\beta}^{(t)} + (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mu(\mathbf{X}, \boldsymbol{\beta}^{(t)})) \\
&= (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{W}^{(t)}\mathbf{X}\boldsymbol{\beta}^{(t)} + (\mathbf{y} - \mu(\mathbf{X}, \boldsymbol{\beta}^{(t)}))\right) \\
&= (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{W}^{(t)}\mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{W}^{(t)}\mathbf{W}^{(t)^{-1}}(\mathbf{y} - \mu(\mathbf{X}, \boldsymbol{\beta}^{(t)}))\right) \\
&= (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}\left(\mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{W}^{(t)^{-1}}(\mathbf{y} - \mu(\mathbf{X}, \boldsymbol{\beta}^{(t)}))\right) \\
&= (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}
\end{aligned}$$

where $\mathbf{z} = \left(\mathbf{X}\boldsymbol{\beta}^{(t)} + \mathbf{W}^{(t)^{-1}}(\mathbf{y} - \mu(\mathbf{X}, \boldsymbol{\beta}^{(t)}))\right)$ is the adjusted response.

Software & Code

```
loglik <- function(y, X, beta) t(y) %*% X %*% beta - sum(log(1 + exp(X %*% beta)))
logistic <- function(x) exp(x)/(1 + exp(x))
weight <- function(mu) diag(c(mu * (1 - mu)), nrow(mu), nrow(mu))
adjust <- function(y, X, beta, mu, W) X %*% beta + solve(W) %*% (y - mu)
update <- function(y, X, beta, mu, W, z) solve(t(X) %*% W %*% X) %*% t(X) %*% W %*% z

niter <- 20
beta <- rep(0, ncol(X))
objective <- rep(NA,niter)
ptm <- proc.time()
for (i in 1:niter) {
  b <- beta
  mu <- logistic(X%*%b)
  W <- weight(mu)
  z <- adjust(y, X, b, mu, W)
  beta <- update(y, X, b, mu, W, z)
  objective[i] <- loglik(y, X, beta)
  if (i > 1) if (objective[i] - objective[i-1] < 1e-6) break
}
```

Estimation (algorithm)

Data: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in [0,1]^n$, and boundary for convergence $\varepsilon > 0$.
Result: Estimated $\widehat{\boldsymbol{\beta}}$.
Add bias term $\mathbf{X} \leftarrow \begin{bmatrix} \mathbf{1} & \mathbf{X} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$;
Initialize $\boldsymbol{\beta}^{new} \leftarrow \mathbf{0} \in \mathbb{R}^{p+1}$;
while $\ell(\boldsymbol{\beta}^{new}) - \ell(\boldsymbol{\beta}^{old}) > \varepsilon$ **do**
  $\boldsymbol{\beta}^{old} \leftarrow \boldsymbol{\beta}^{new}$;
  $\mathbf{W} \leftarrow \text{diag}\left(\mu(\mathbf{x}_i, \boldsymbol{\beta}^{old})\left(1 - \mu(\mathbf{x}_i, \boldsymbol{\beta}^{old})\right)\right)$;
  $\mathbf{z} \leftarrow \left(\mathbf{X}\boldsymbol{\beta}^{old} + \mathbf{W}^{-1}(\mathbf{y} - \mu(\mathbf{X}, \boldsymbol{\beta}^{old})\right)$;
  $\boldsymbol{\beta}^{new} \leftarrow (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$;
end

Standard errors and inference

Statistical inference (Kass, 2011)

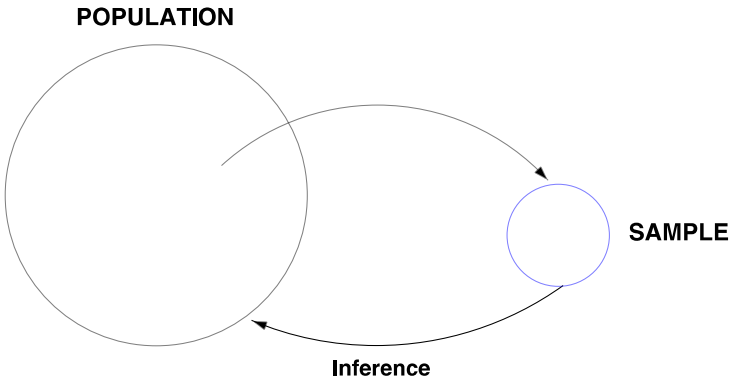


FIG. 3. *The big picture of statistical inference according to the standard conception. Here, a random sample is pictured as a sample from a finite population.*

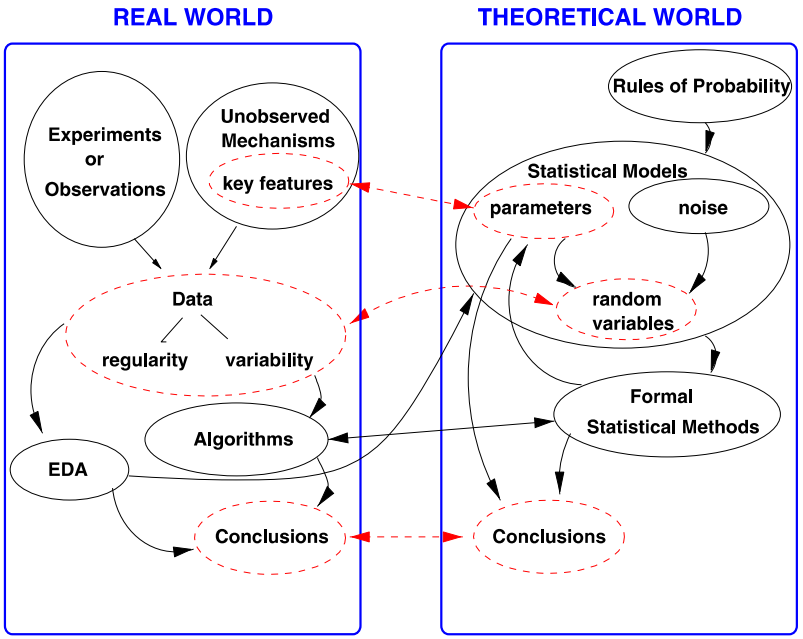


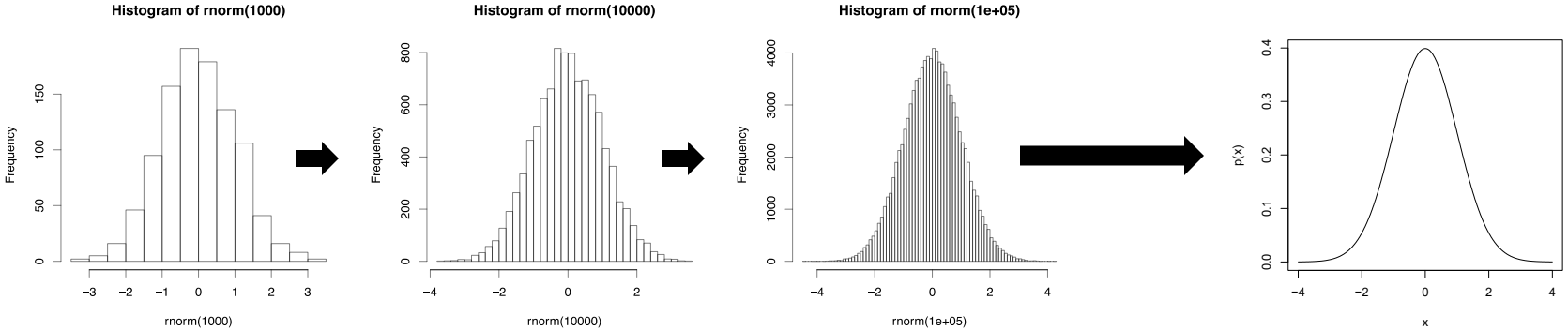
FIG. 4. *Amore elaborate big picture, reflecting more detail the processof statistical inference. As in Figure 1, there is a hypothetical link between data and statistical models but here the data are connected more explicitly to their representation as random variables*

Inference is more than just sampling from a population

- Sampling from a larger population, and using samples to make inferences about the larger population, is how inference is typically taught: but it is only *one* use of inference
- Even if we have the entire population, we still do inference: to a *hypothetical underlying process*
- (Historically: Fisher talked about underlying processes as "hypothetical infinite populations." But his rivals Jerzy Neyman and Karl Pearson's son Egon Pearson, hated the metaphysics of this. They tried to say that the only source of uncertainty was sampling. That's how it's usually taught today, but not how statisticians themselves think about it. See Kass, 2011).

Probability distributions are hypothetical

- There is no such thing as a normal (or any other) probability distribution in the world: they are mathematical objects
- We imagine more and more data as eventually "converging" to an underlying distribution. But we can never get to infinite data, and seldom even get more data. It's all ultimately a "heuristic" that we use to justify, develop, and test statistical methods



The “Great Intellectual Fraud”?

- Nasim Nicholas Taleb calls the normal distribution the “Great Intellectual Fraud”
- Perhaps accurate for the way it is taught in intro stats courses… but:
- Statistics does NOT say that everything is normally distributed (“haphazard variability” is not always normal)
- The Central Limit Theorem says that, using statistics, our our *knowledge* of means (epistemic uncertainty of statistical inference) will be normally distributed

Central Limit Theorem (CLT)

- The law of large numbers says that as you get more data from the underlying phenomenon, the closer the sample mean is to the true mean
- The central limit theorem says (among other things) that not only is the sample mean close to the true mean, but it is distributed normally around the true mean
  - Key point: mean is "distributed" is over multiple data sets
  - We seldom have multiple distinct data sets (and if we did, why wouldn't we just combine them??)
  - The set of "other data sets we could have drawn" is a *theoretical construct* we appeal to in frequentist inference

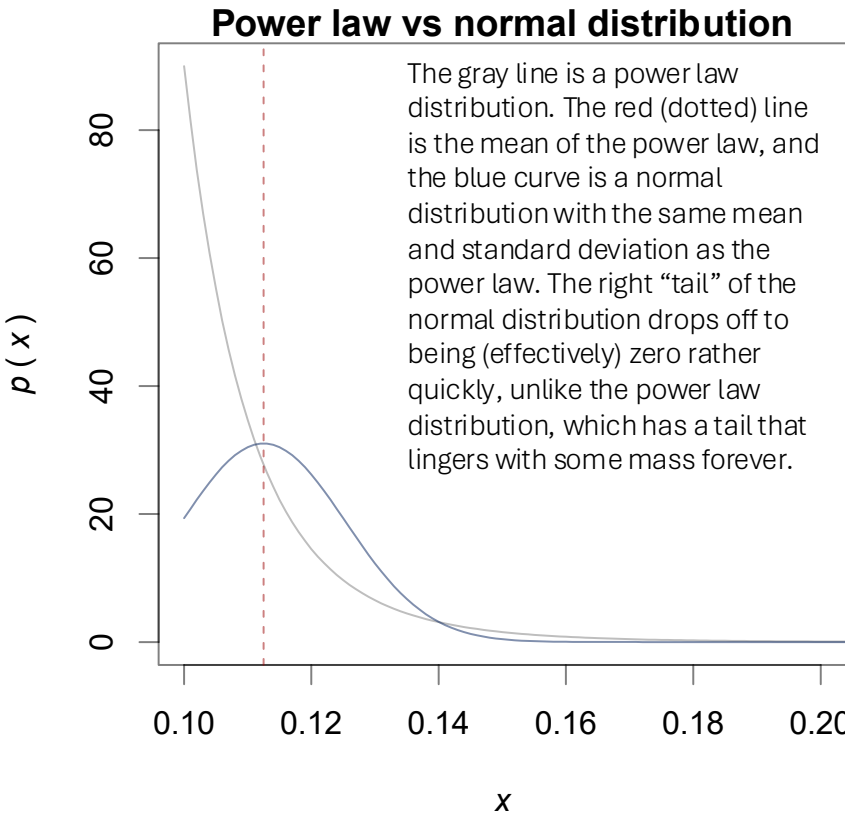
Illustration with power laws

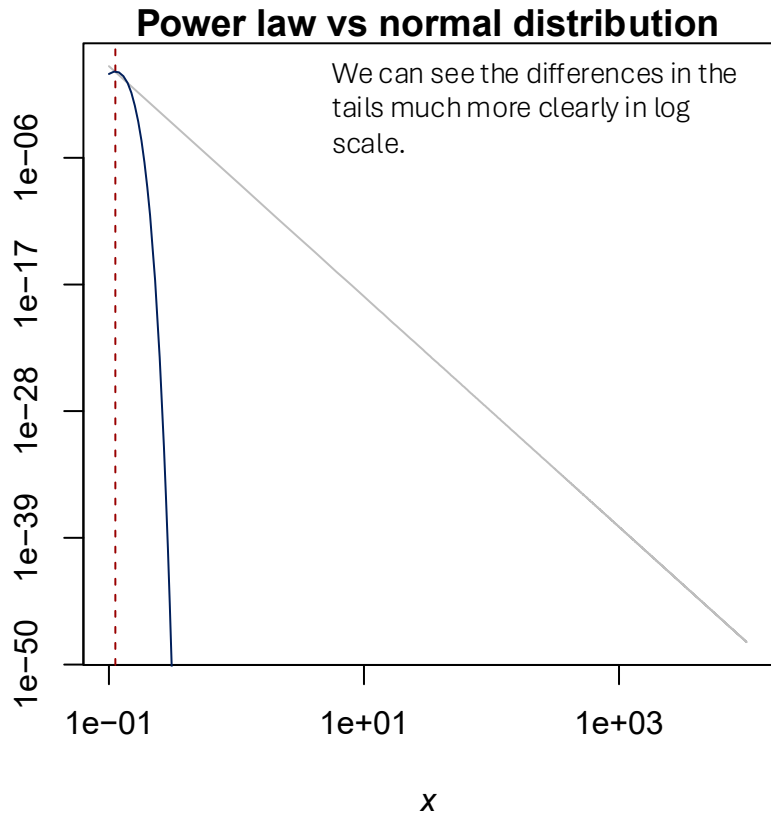
- To illustrate: take one the most “non-normally” distributed distributions, a power law distribution
- (Caution: *heavy-tailed distributions* are not automatically “power laws”!)
- Power laws can have a mean, it’s just that it’s not informative
- The sample mean, over multiple samples, is still normally distributed around the true mean!

(Note on simulations)

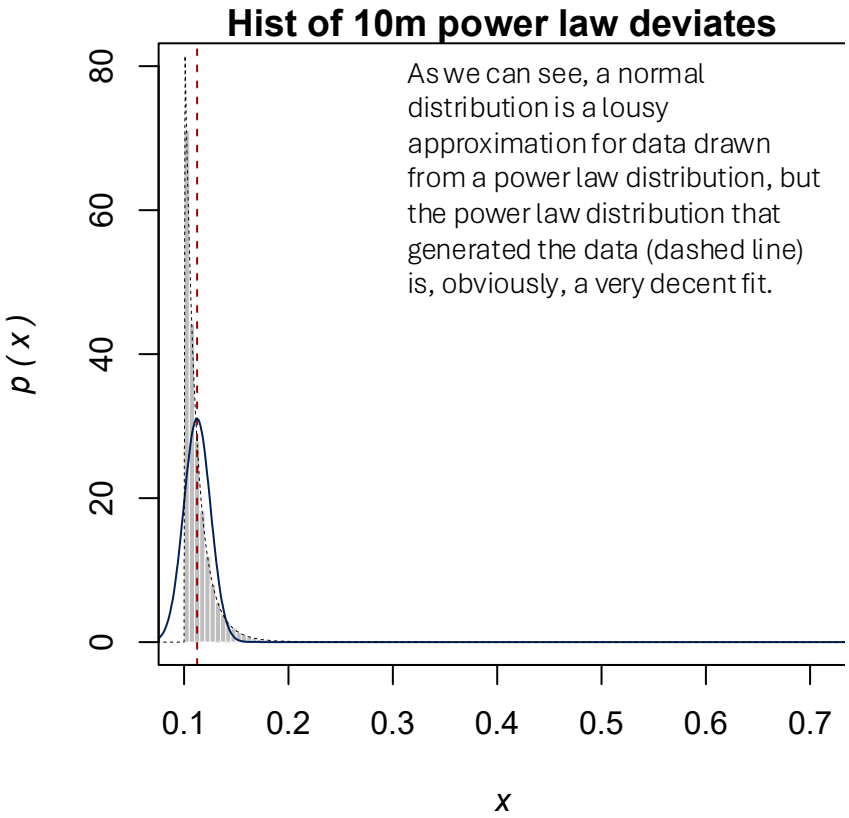
- Simulations (producing "synthetic data"/"toy data") are very useful in statistics
- It is to investigate, "if the world works the way I say it does, do my claims hold?" Low bar, but useful, and more for understanding/exploration than argument
- Important to note: Usually, we have no idea if we were successful from methods themselves
- When there is no way to independently validate a model, we can *simulate the world working a specific way*
- **For frequentist claims especially, we can simulate long-run frequencies**
- (This is distinct from simulation *modeling*, which is about modeling the world rather than exploring the performance of statistical techniques)

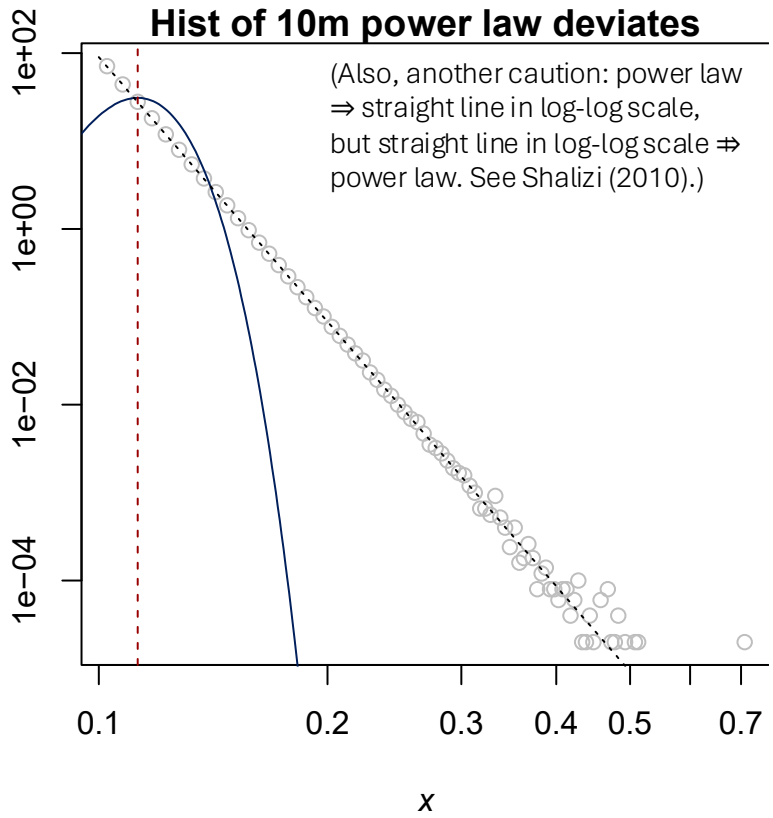
Theoretical distributions: Power law vs. normal



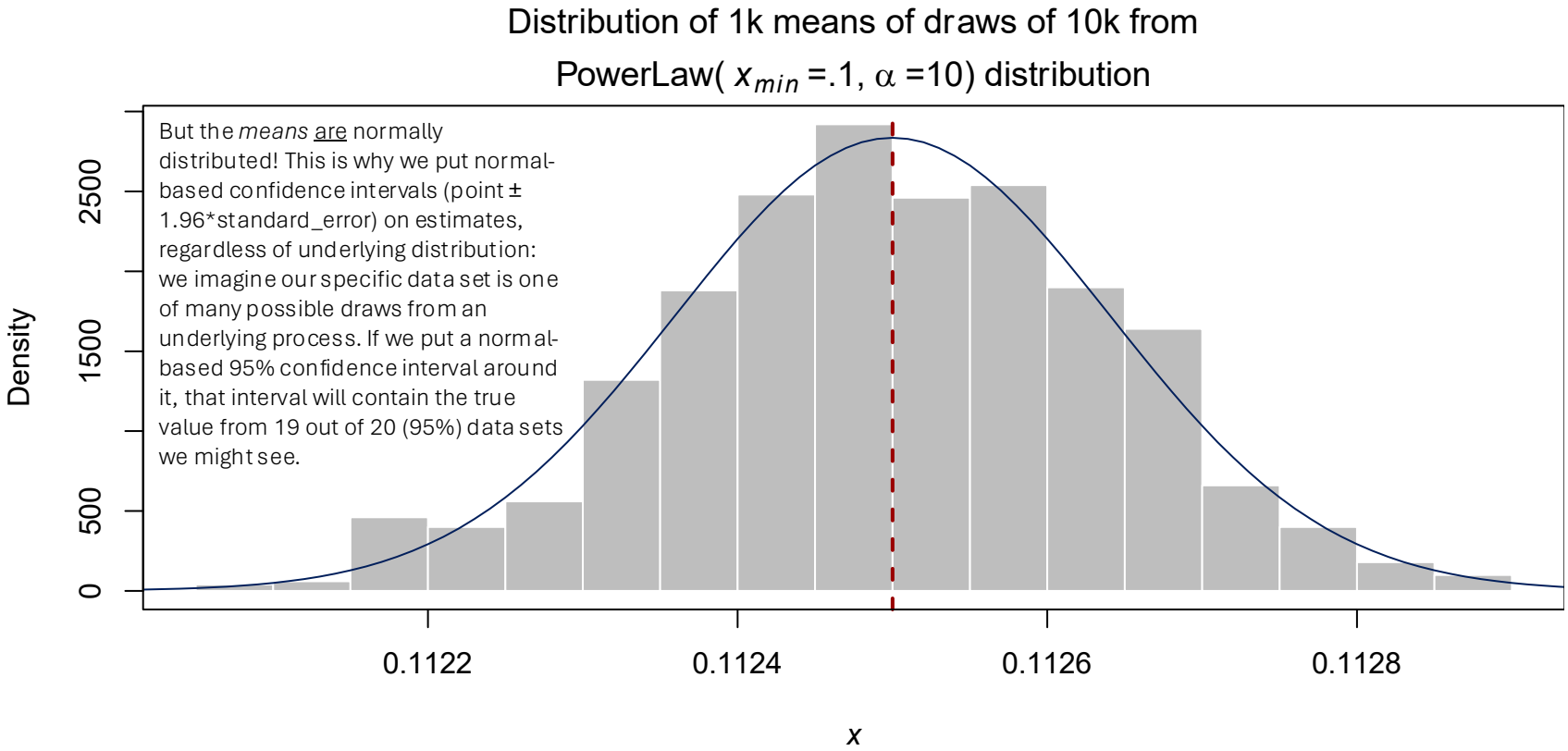


Distribution of *draws* (with normal reference)

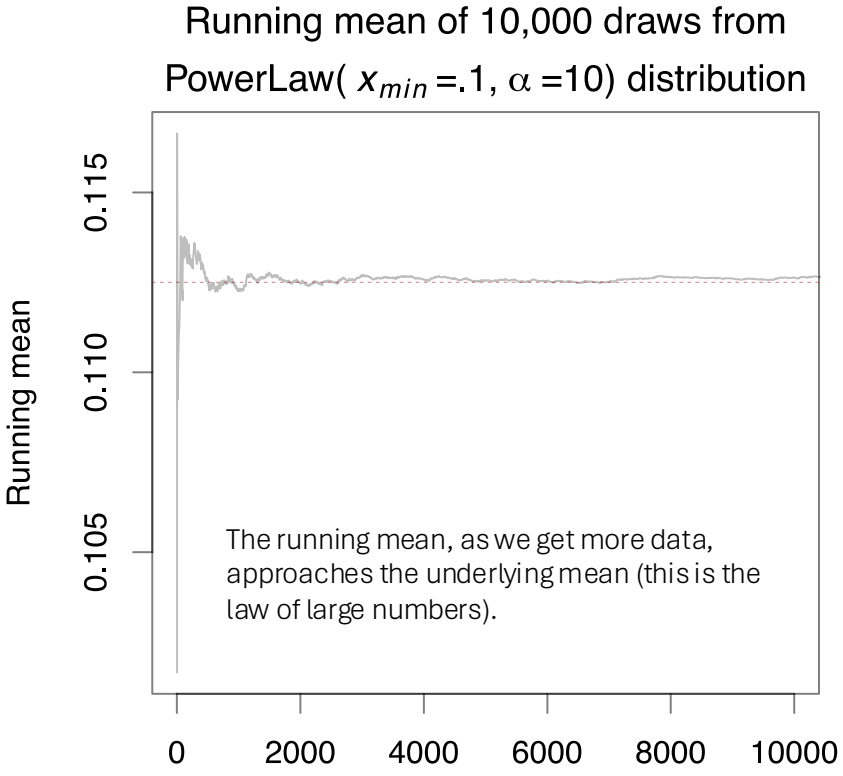


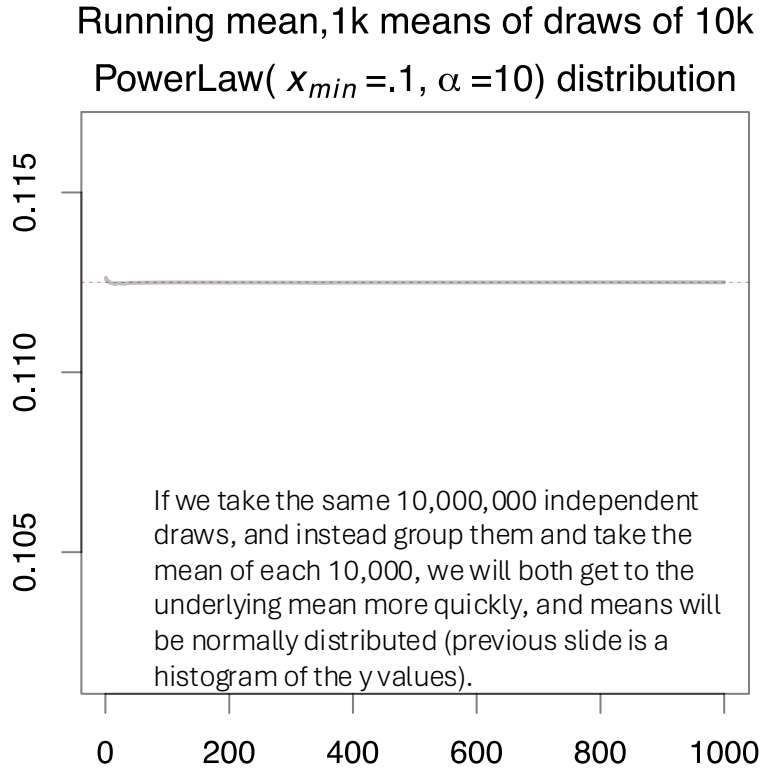


Distribution of *means* (grouped draws)

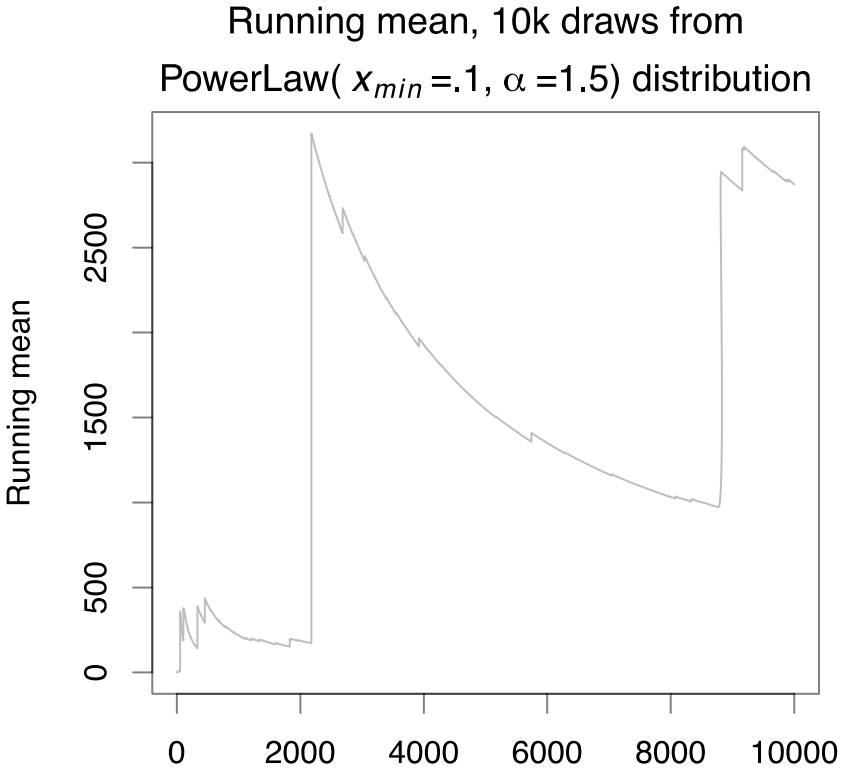


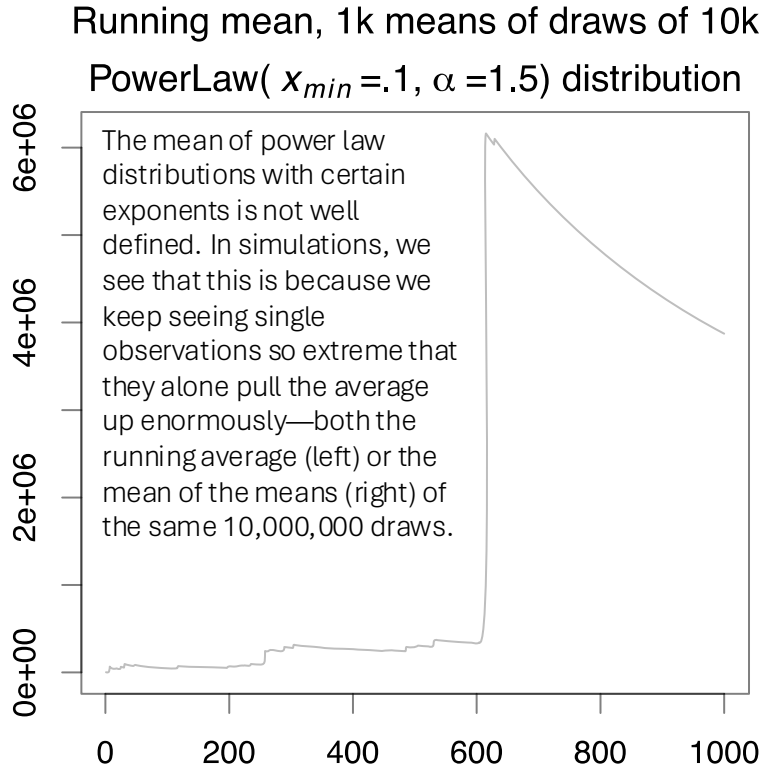
Alternative view: Mean as $n \to \infty$





But the mean may not exist, in which case, no CLT





The mean of power law distributions with certain exponents is not well defined. In simulations, we see that this is because we keep seeing single observations so extreme that they alone pull the average up enormously—both the running average (left) or the mean of the means (right) of the same 10,000,000 draws.

CLT: Take-aways

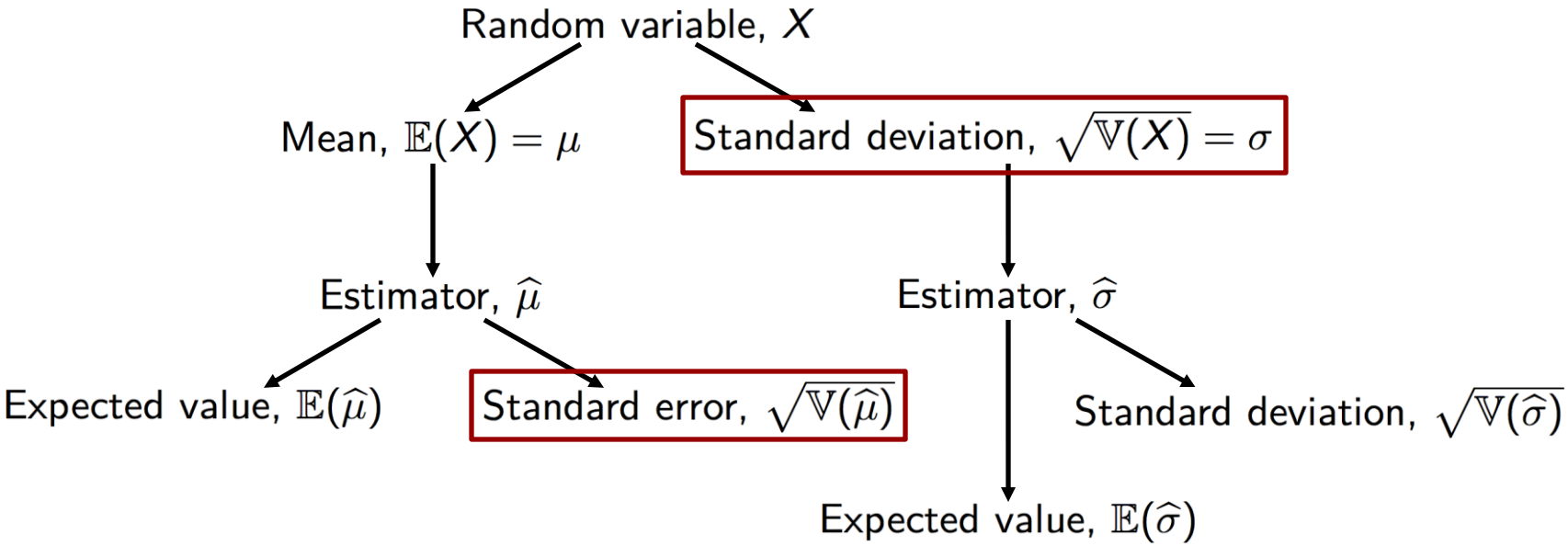
- *If* the mean exists: the *sample* average is normally distributed around the "true" mean (over multiple data sets), regardless of the underlying distribution
  - This is regardless of whether the mean is *useful* or not: for power laws (where the mean exists), it is <u>not</u> useful.
  - Other cases: bimodal distributions, log-normal and other skewed distributions (where the log-mean, or *geometric* mean, is the more useful "relevant information")

Revisiting: Standard deviation vs. standard errors

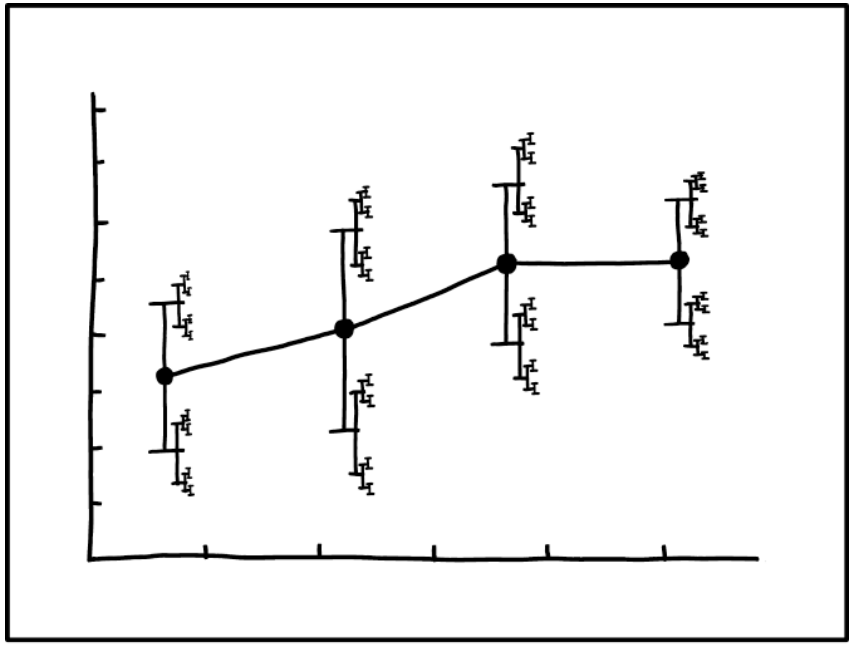
- The difference between the two confused me for a long time
- But, understanding the difference between the distribution of "haphazard variability", and distribution of our *knowledge* of the mean of that haphazard variability (over hypothetical multiple "draws", forming separate data sets) hopefully explains the difference
  - Standard deviation describes a distribution of *variability* of the underlying process
  - Standard errors describe a distribution of *uncertainty*, how close the sample mean (or other quantity) is to the true mean

In the notation of statistical theory

- Statistical theory studies the properties of each of these. And we could go on *ad infinitum*



This actually has a lot of wisdom...



- ...but statistical theory generally stops at the "second" set of error bars (and we only plot the first)

https://xkcd.com/2110/

What about Bayesian statistics?

- (Disclaimer: I've only read a bit into what is a deep controversy)
- What we want: *P*(Hypothesis is true | Data). Frequentist statistics can only give us *P*(Data | Hypothesis is false)
- Bayesian statistics gives us what we want, using Bayes rule to invert *P*(*D* | *H*):

$$P(H \mid D) = P(D \mid H)\, P(H) \,/\, P(D)$$

  ...but unfortunately it doesn't *work* the way we want.
  - If we repeat a simulation 1000 times, a Bayesian 95% credible interval may not contain the true value approximately 95% of the time
  - The problem is that we need a *prior P*(*H*), to invert the probability; and there's no choice of prior that consistently gives frequentist properties (the denominator, *P*(*D*), turns out to usually not matter as much)

What about Bayesian statistics?

- So we either get the quantity we we want, but it doesn't work the way we want; or we get something that works, but isn't what we want
- If we really have an empirical prior (e.g., base rate of a disease), then Bayesian inference can work in a frequentist way. I don't know if that fixes anything; I suspect the frequentist "base" makes everything frequentist, so what we don't "really" get $P(H \mid D)$
- 100+ years of varying levels of mathematical and philosophical sophistication hasn't found a way out of this impasse. **It's unlikely that we will ever have a way of doing statistical analysis that gives us what we want *and* will work the way we want**

Bayesian inference

- Bayesian inference: *updating priors* for our beliefs to be in better accordance with data.
- The "purest" form of Bayesian inference is *subjective* Bayes: expressing our subjective beliefs in the form of a probability distribution. All we can do is update our beliefs
- Possibly a way to have no singular underlying "truth" we are making inferences about. But in practice: subjective Bayes is <u>not</u> inherently anti-positivist. It can believe in an objective reality, just that our subjective beliefs help us get there quicker (Andrew Gelman: "Bayesians are frequentists").

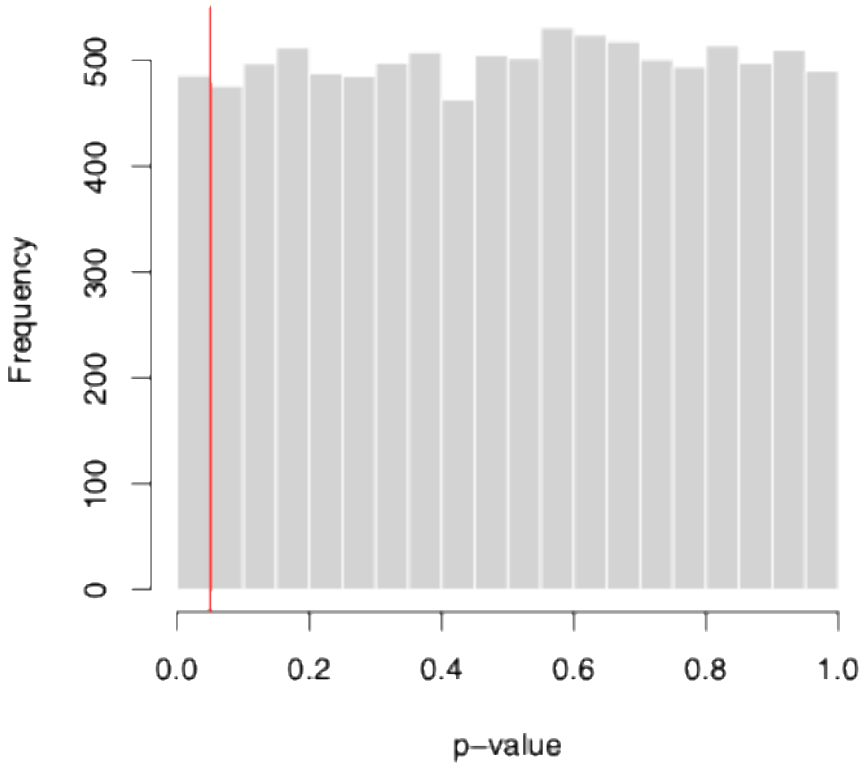
“Priors” are not ways to do critique

- Priors are *not* a good way to capture social theory. For example, a prior that “racism exists” doesn’t really work out, because in principle, enough data showing that racism does not exist could “wash out” that prior. But that’s not the nature of the social theory, which is about how we see and understand the basic building blocks of the world, prior to data and measurement
- Conversely, with not enough data, our conclusions may just be our priors. In which case, why bother with quantitative analysis?

More about *p*-values

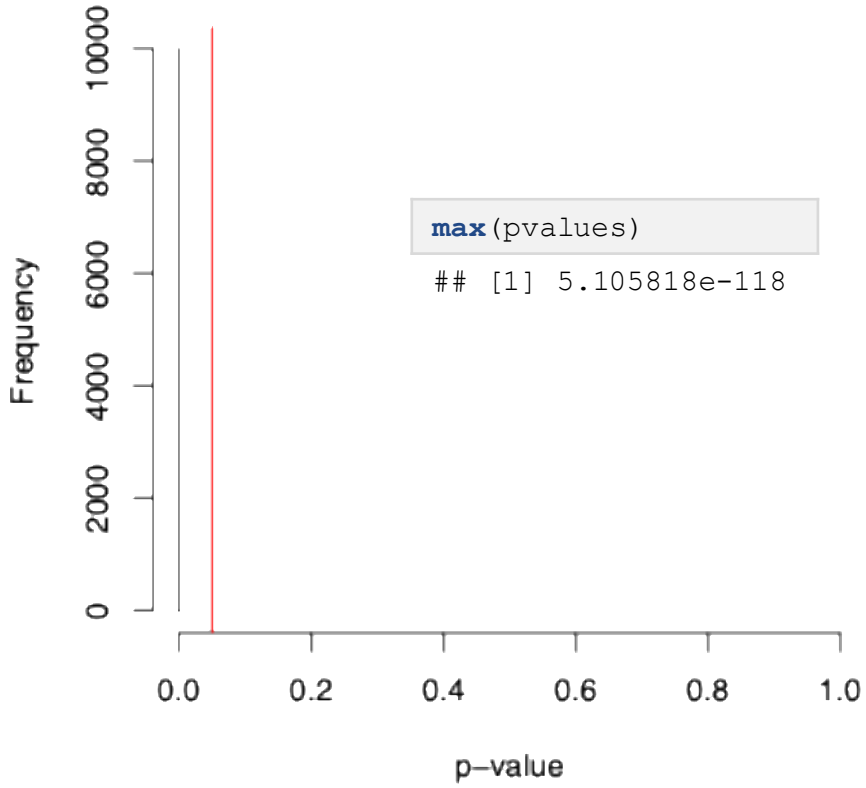
Useful property of *p*-values: Uniform distribution if the null hypothesis is true

```r
set.seed(2025)
nsim <- 10000
n <- 1000
pvalues <- rep(NA, nsim)
for (i in 1:nsim) {
  x <- rnorm(n)
  y <- rnorm(n)
  fstats <- summary(lm(y ~ x))$fstatistic
  pvalues[i] <- pf(fstats[1],
                   fstats[2],
                   fstats[3],
                   lower.tail = FALSE)
}

hist(pvalues,
     breaks = 20,
     main = NA,
     border = "white",
     xlab = "p-value")
abline(v = 0.05, col = "red")
```



Useful property of *p*-values: Uniform distribution if the null hypothesis is true

```r
set.seed(2025)
nsim <- 10000
n <- 1000
pvalues <- rep(NA, nsim)
for (i in 1:nsim) {
  x <- rnorm(n)
  y <- rnorm(n = n, mean = x, sd = 1)
  fstats <- summary(lm(y ~ x))$fstatistic
  pvalues[i] <- pf(fstats[1],
                   fstats[2],
                   fstats[3],
                   lower.tail = FALSE)
}

hist(pvalues,
     breaks = 20,
     main = NA,
     xlim = c(0, 1),
     xlab = "p-value")
abline(v = 0.05, col = "red")
```




```r
max(pvalues)
```

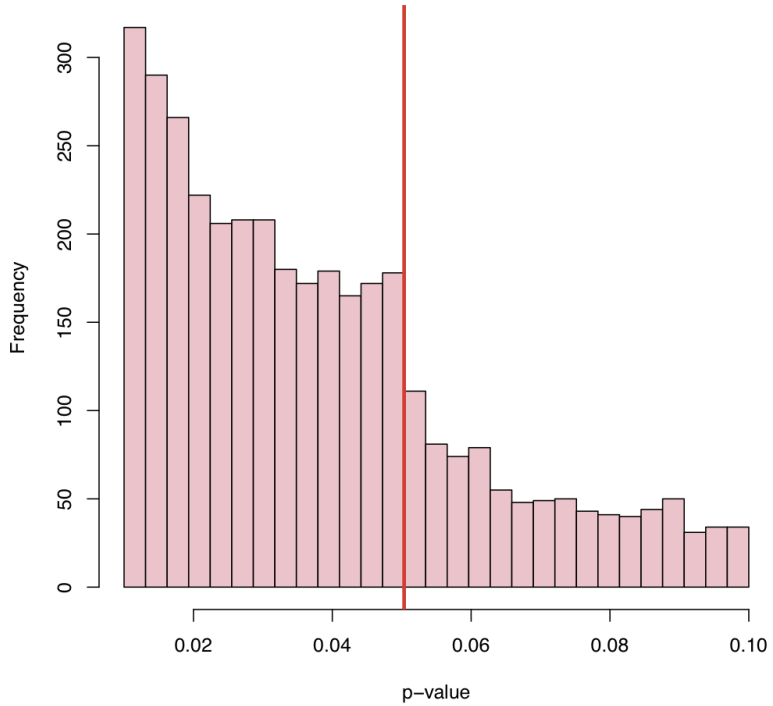
```
## [1] 5.105818e-118
```

Only care about above/below a preselected cutoff!

- If the null hypothesis is true, then a *p*-value of 0.050001 is just as likely as a *p*-value of 0.999999
  - This means that “almost significant”, “barely significant”, etc., shouldn’t be taken to mean anything
- “Extremely significant” is still a misuse of what *p*-values are meant to do
- There is a proposal for an “*s*-value”, for “surprise”: the log of the *p*-value, for substantive interpretation. I haven’t yet decided if I am convinced or not

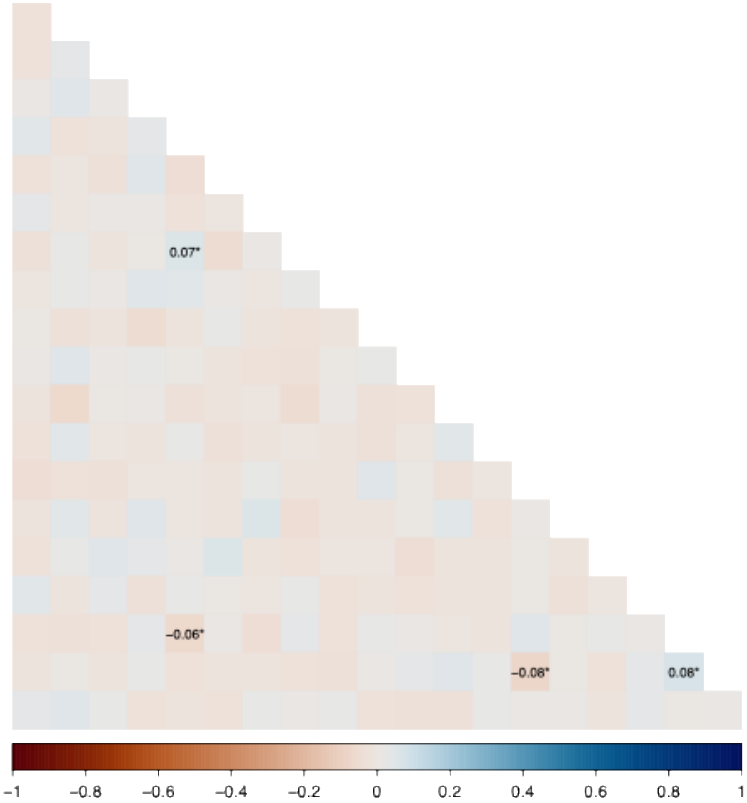
Distribution of *p*-values in published literature

- Empirical evidence of *p*-hacking/*p*-fishing: if we test enough hypotheses, by chance some false ones will have $p < 0.05$
- *p*-hacking ruins the guarantees of *p*-values, guarding against chance results



Data: Masicampo & Lalande (2012). Plot: Wasserman (2012)

The multiple comparisons problem



- Let's say you have 20 variables, and want to know which are correlated
- There are $\binom{20}{2}/2 = 95$ unordered pairs
- If none of them are actually correlated, and you chose a level of 0.05, then on average about 95/20 = 4.5 will have $p < 0.05$ (left: printed coefficient)
- This is *p*-values working correctly!

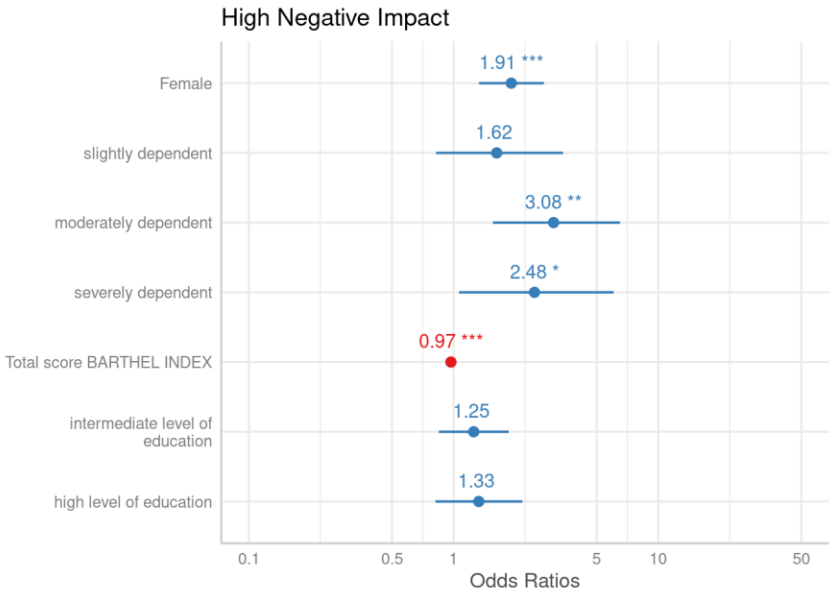
Addressing multiple comparisons

- The simplest is the "Bonferroni correction": divide the critical value by the number of comparisons. So if you are making 95 comparisons, you would want $p < 0.05/95$, or $p < 0.000526$
- This is overly conservative, but other corrections (e.g., "False Discovery Rate" method) are more complicated
- We maybe should always be doing this, e.g., dividing our critical value by the number of regression coefficients across *all* models we consider… but in practice we don't bother
- We can have *joint* (rather than pointwise) confidence intervals (also called confidence *bands*), but those are much harder to calculate

Alternatives to *p*-values?

- *p*-values are a way of controlling for the possibility of incorrect conclusions, using variability to estimate uncertainty
- One thing that is worse than *p*-values is ignoring *p*-values (contra Stephen Ziliak; Ziliak & McCloskey, 2008)
- Machine learning ignores uncertainty; every claimed success (AUC, precision/recall, F1, etc.) really has an associated *p*-value, and many of them are probably $p > 0.05$ in the sense that they don't generalize (see Kapoor et al., 2024).
  - Overfitting can be a form of multiple comparisons: if we try enough models, and eventually one of them improves on a baseline, it might be by chance. But machine learning is not even quantifying uncertainty, it is just going by "effect size"

Confidence intervals



- Confidence intervals (CIs) are mathematically equivalent: the narrowest CI that includes include the null hypothesis value (e.g., 0) is 1 minus the *p*-value. E.g., if a 91.25% CI includes 0 but a 91.26% CI does not, then the *p*-value is 0.0875.
- CIs can be visualized with error bars, which allow for visual interpretation

Lüdecke, D. (2024, November 19). Plotting Estimates (Fixed Effects) of Regression Models. *sjPlot Vignette*. https://strengejacke.github.io/sjPlot/articles/plot_model_estimates.html

What are we assuming?

Way statistics see the world

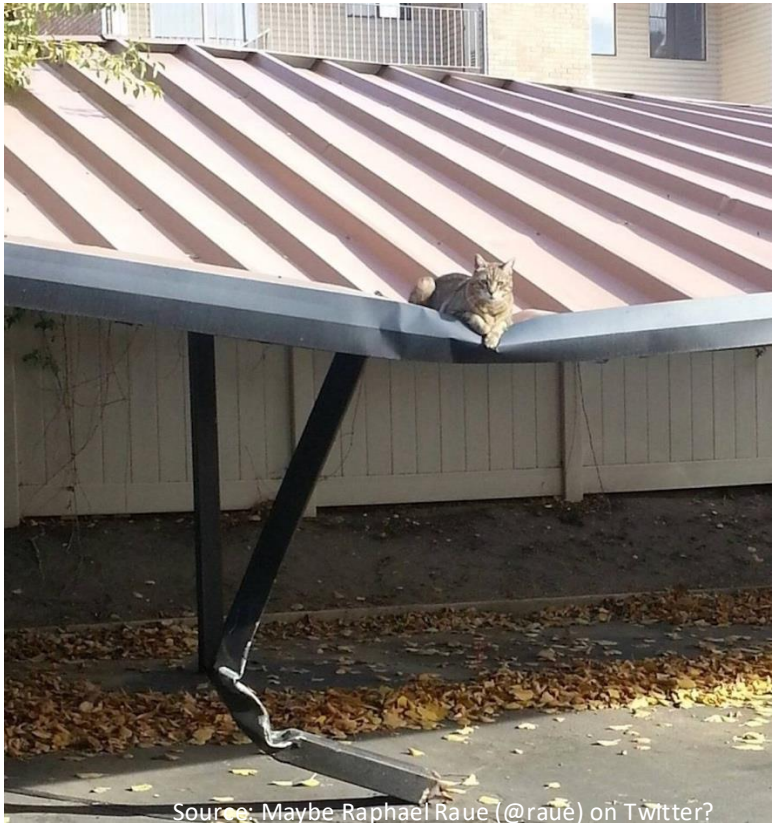
- Statistical modeling: "…assumes that the social world consists of fixed entities (the units of analysis) that have attributes (the variables)" (Abbott, 1988)
- **"…it is striking how absolutely these assumptions contradict those of the major theoretical traditions of sociology**. Symbolic interactionism rejects the assumption of fixed entities and makes the meaning of a given occurrence depend on its location—within an interaction, within an actor's biography, within a sequence of events. Both the Marxian and Weberian traditions deny explicitly that a given property of a social actor has one and only one set of causal implications…" (Ibid.)

Way statistics see the world

- The use of distributions of aggregates, and central tendency, implies a philosophical commitment:
  There are distinct entities in the world that, despite being different, are similar in some way.
- As a corollary, we can thus learn about one thing by studying other things (and eventually, make statements about not-yet-seen entities based on the study of seen entities).
- Treating the world via aggregates is a logic of domination, preceding eugenics but certainly exemplified by it, that is woven into the fundamental nature of statistics (see also: Clayton, 2020)

In contrast: non-statistical correlation

- This is a *non-statistical* correlation
- (It has meaning to us based on tons of background information/ intuition on physics, structures, and cats, that statistical models don't have)
- A statistical correlation is a property of *multiple observations*. Any meaning it has requires the idea of an "aggregate trend"



Source: Maybe Raphael Raue (@raue) on Twitter?

Caution! Understanding a person…

Table from Barbara Kiviat; see Kiviat (2023)

	As a case	**In narrative**
Context/circumstance	Stripped away	Key
Mental states	Absent (for the most part)	Crucial; constitutive
Relevant features	Determined in advance	Emergent
Orientation to time	Atemporal	Chronological
Ordering of features	Unimportant	Meaningful
Other actors	Invisible	Often present
Causal logic	Mathematical	Theoretical
To boost predictive validity	Add cases	Know person better

"Bowker and Star 2000; Bruner 1986; Desrosières 1998; Espeland 1998; Espeland and Stevens 1998, 2008; Fourcade and Healy 2017; Hacking 1990; Porter 1994, 1995; Ricouer 1998; White 1980, 1984". I would add: Patton 2005; Abbott 1988

What are the implications of the origins of statistics?

- Its *content*: treating people as exchangeable, and describing the world in terms of central tendency
- Its *uses*: a particular form of abstract, mathematically expressed discourse is what is seen as legitimate
  - Using statistics to refute racism risks reifying the methods: agreeing that statistics and data are how we determine if (for example) racism or sexism are or aren't "real"
- Its *creators*: only one Black person among those celebrated as creating modern statistics, David Blackwell, and no Black women
- Contra Zuberi (Bonilla-Silva & Zuberi, 2008), it's not that math is neutral but stats isn't. Even "2 + 2 = 4" encodes racial history: Galton has a story about Africans "struggling" with that idea. Why should we value some abstractions inherently, versus whether they are meaningful within an environment, and based on meaningfulness towards what purpose?

"Paradigms of inquiry"

Issue	Positivism	Post-positivism	Critical theory et al.	Constructivism	Participatory
Ontology	Naïve realism: Reality independent of and prior to human conception of it, apprehensible	Critical realism: Reality independent of and prior to human conception, but imperfectly and approx. apprehensible	Disenchantment theory: reality is secret/hidden, shaped by power structures and solidified over time	Relativism: multiple realities, constructed in history through social processes	Participative: multiple realities, co-constructed through interactions between specific people and environments
Epistemology	Reality knowable. Findings are singular, neutral, perspective-independent, atemporal, universally true	Findings provisionally true; multiple descriptions can be valid but are probably equivalent; findings can be affected/distorted by social + cultural factors	How we come to know something, or who knows it, matters for how meaningful it is	Relativistic: no neutral perspective to adjudicate competing claims	We come to know things, create new understandings, & transform world by involving other people in process of inquiry
Methodology	Hypotheses can be verified as true. Quant methods, math.	Falsification of hypotheses; primacy of quant, but some qual and mixed methods	Dialogic (conversation + debate) or dialectical ($thesis_1$ → $antithesis_1$ → $synthesis_2$:= $thesis_2$...)	Dialetical, or exegetical (reading between the lines)	Collaborative, action-focused; flattening hierarchies; engaging in self- and collective reflection, action
Axiology	Quant knowledge-holders have access to truth, and responsibility from it	Quant knowledge valuable but can be distorted; qual can help find and correct	Marginalization provides unique insights, knowledge of marginalized valuable	Understanding construction is valuable; value relative to given perspective	Reflexivity, co-created knowledge, and non- western ways of knowing are valuable and combat erasure and dehumanization

Malik & Malik (2021), via Guba and Lincoln (2005)

How it was (Campbell, 1975)

"In fields such as sociology and social psychology, **many of our ablest and most dedicated graduate students are increasingly opting for the qualitative, humanistic mode**. In political science, there has been a continuous division along these lines. Only economics and geography seem relatively immune… The critics taking what I am calling the humanistic position are often **well-trained in quantitative-experimental methods**. Their specific criticisms are often well-grounded in the experimentalist's own framework: experiments implementing a single treatment in a single setting are profoundly ambigous [sic] as to what caused what; there is a precarious rigidity in the measurement system, limiting recorded out-comes to those dimensions anticipated in advance;…

How it was (Campbell, 1975)

"...process is often neglected in an experimental program focused on the overall effect of a complex treatment, and thus knowing such effects has only equivocal implications for program replication or improvement; broad-guage [sic] programs are often hopelessly ambiguous as to goals and relevant indicators; changes of treatment program during the course of an ameliorative experiment, while practically essential, make input-output experimental comparisons uninterpretable; social programs are often implemented in ways that are poor from an experimental design point of view; even under well-controlled situations, experimentation is a profoundly tedious and equivocal process; experimentation is too slow to be politically useful; etc."

Quantification as a distraction

“The function, the very serious function of racism is distraction. It keeps you from doing your work. It keeps you explaining, over and over again, your reason for being. Somebody says you have no language and you spend twenty years proving that you do. Somebody says your head isn’t shaped properly so you have scientists working on the fact that it is. Somebody says you have no art, so you dredge that up. Somebody says you have no kingdoms, so you dredge that up. **None of this is necessary. There will always be one more thing.”** (Morrison, 1975)

“The starving fellah, (or the jobless inner city N.H.I., the global New Poor or *les dannés*), Fanon pointed out, does not have to *inquire into the truth*. He *is*, they *are*, the Truth. It is we who constitute this ‘Truth.’ **We must now undo their narratively condemned status.”** (Wynter, 1994)

- Is quantification the way to undo this “narrative condemnation”, or just a reification of it? Statistics as a technology of suspicion—that experience is anecdotal and unreliable, people are deluded or lie (see also Porter, 2012)

Conversely: Sins of qual

- “we are suggesting that anthropological analyses (of pain and passion and power), when they are experience—distant, are at risk of delegitimating their subject matter's human conditions. **The anthropologist thereby constitutes a false subject;** she can engage in a professional discourse **every bit as dehumanizing** as that of colleagues who unreflectively draw upon the tropes of biomedicine or behaviorism to create their subject matter.” (Kleinman & Kleinman, 1991)
- Linda Tuhiwai Smith’s (2012) *Decolonizing methodologies: Research and indigenous peoples* is about qualitative research
- “Methods are like people: if you focus only on what they can’t do, you will always be disappointed” (Shapiro, 2014)
- Fun words: methodolotry, methography

References

Abbott, A. (1988). Transcending general linear reality. *Sociological Theory*, *6*(2), 169–186. https://doi.org/10.2307/202114

Bonilla-Silva, E., & Zuberi, T. (2008). Toward a definition of white logic and white methods. In T. Zuberi & E. Bonilla-Silva (eds.) *White logic, white methods* (pp. 3–30). New York, NY: Rowman & Littlefield Publishers.

Campbell, D. T. (1975). Assessing the planned impact of social change. In G. M. Lyons, (Ed.), *Social Research and Public Policies*. Hanover, New Hampshire: University Press of New England.

Clayton, A. (2020, October 7). How eugenics shaped statistics: Exposing the damned lies of three science pioneers. *Nautilus*.

Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science*, *5*(2), 169–174. https://doi.org/10.1214/ss/1177012165

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*, 309–368. https://doi.org/10.1098/rsta.1922.0009

Hu, L. (2021, May 6). Race, policing, and the limits of social science. *Boston Review*. https://www.bostonreview.net/articles/race-policing-and-the-limits-of-social-science/

Kapoor, S., Cantrell, E., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., & Narayanan, A. (2024). REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, *10*(18), eadk3452. https://doi.org/10.1126/sciadv.adk3452. [Science link]

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science*, *26*(1), 1–9. https://dx.doi.org/10.1214/10-STS337

Kleinman, A., & Kleinman, J. (1991). Suffering and its professional transformation: Towards an ethnography of interpersonal experience. *Culture, Medicine and Psychiatry*, *15*(3), 275-301. https://doi.org/10.1007/BF00046540

Kiviat, B. (2023). The moral affordances of construing people as cases: How algorithms and the data they depend on obscure narrative and noncomparative justice. *Sociological Theory*, *41*(3). https://doi.org/10.1177/07352751231186797

Lanius, C. (2015, January 12). Fact check: Your demand for statistical proof is racist. *Cyborgology* [blog]. https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271-2279. https://doi.org/10.1080/17470218.2012.711335

Malik, M., & Malik, M. M. (2021). Critical technical awakenings. *Journal of Social Computing*, *2*(4), 365–384. https://doi.org/10.23919/JSC.2021.0035

Morrison, T. (1975, 30 May). A humanist view. Black Studies Center public dialogue, Pt. 2. Portland State University, Oregon Public Speakers Collection. https://mackenzian.com/blog/2014/07/07/transcript-morrison-1975/

Porter, T. M. (2012). Thin description: Surface and depth in science and science studies. Osiris, 27(1), 209–226. https://www.jstor.org/stable/10.1086/667828

Shalizi, C. R. (2010). So, you think you have a power law, do you? Well isn't that special. http://www.stat.cmu.edu/~cshalizi/2010-10-18-Meetup.pdf

Shapiro, I. (2014). Methods are like people: If you focus only on what they can't do, you will always be disappointed. In D. L. Teele (Ed.), *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* (pp. 228–241). New Haven & London: Yale University Press.

Smith, L. T. (2012). *Decolonizing methodologies: Research and indigenous peoples* (2nd ed.). London: Zed Books.

Wasserman, L. (2012, August 16). P-values gone wild and multiscale madness. *Normal Deviate: Thoughts on Statistics and Machine Learning* [blog]. https://normaldeviate.wordpress.com/2012/08/16/p-values-gone-wild-and-multiscale-madness/

Wynter, S. (1994). 'No Humans Involved:' An open letter to my colleagues. In *Forum N. H. I: Knowledge for the 21st Century*, *1*(1), 42–71. Institute N. H. I. https://lib.com.org/files/Wynter5.pdf

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: University of Michigan Press.