Safe Testing

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

with Rianne de Heide, Wouter Koolen, Judith ter Schure, Alexander Ly, Rosanne Turner





Slate **Sep 10th 2016: yet another classic finding in psychology—that you can smile your way to happiness—just blew up…**

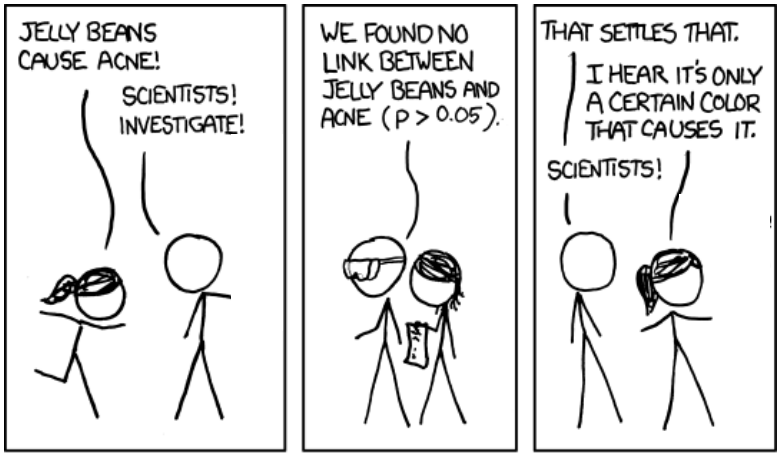
"at least 50% of highly cited results in medicine is irreproducible"
J. Ioannidis, PLoS Medicine 2005

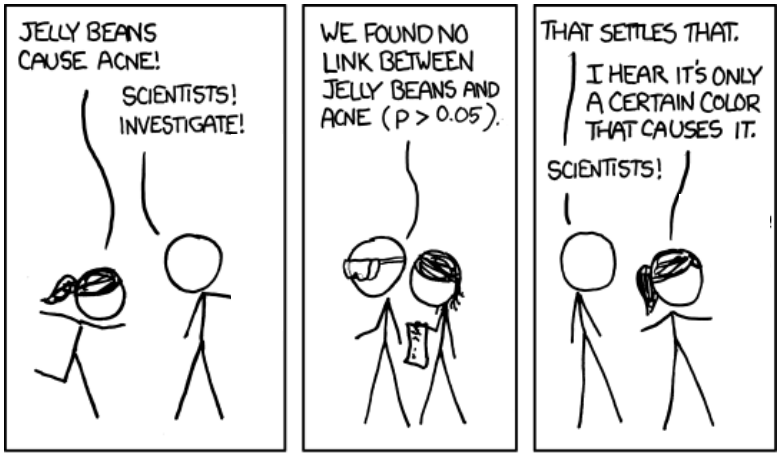
Reproducibility Crisis
Cover Story of Economist (2013),
Science (2014)

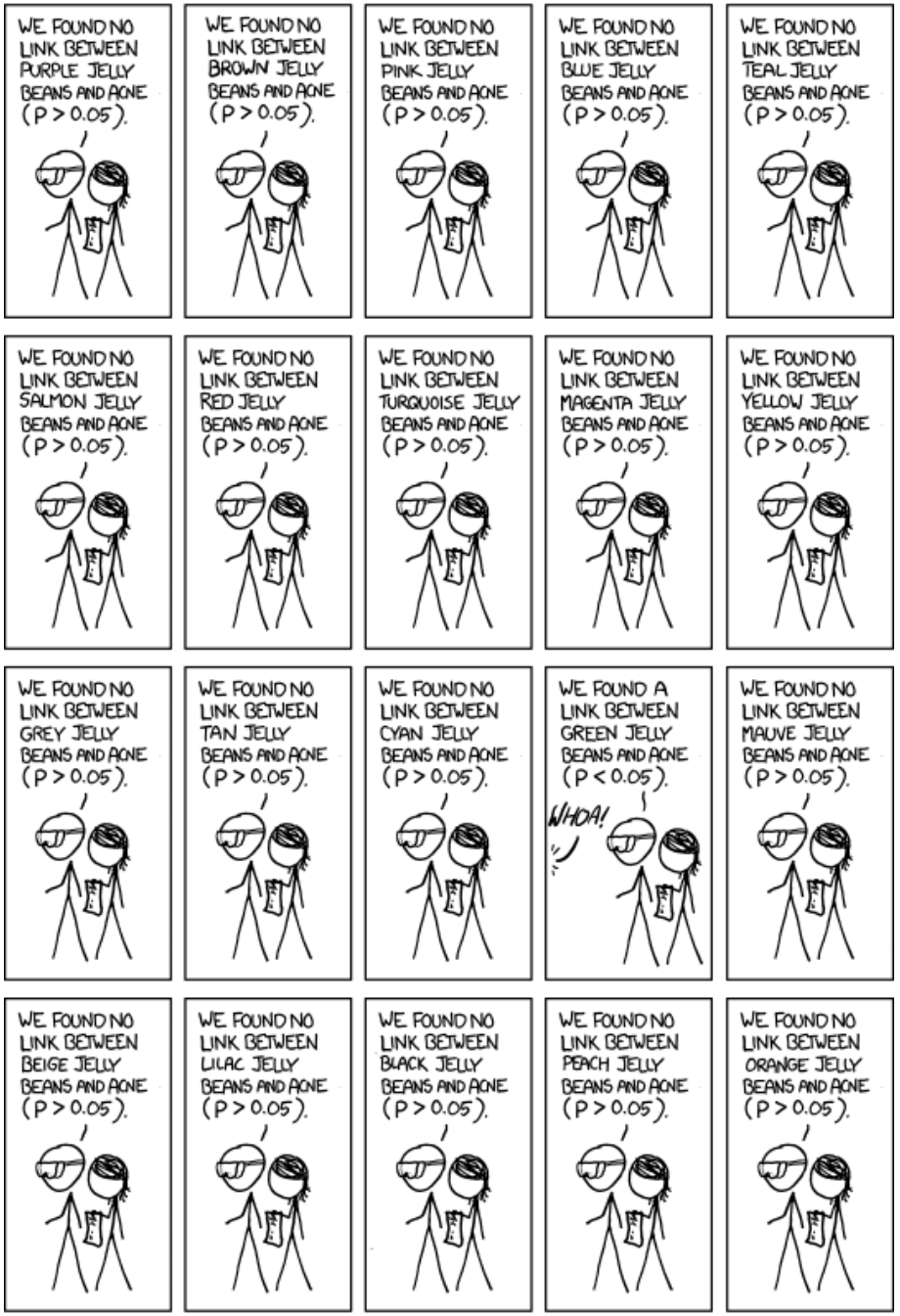
Reasons for Reproducibility Crisis

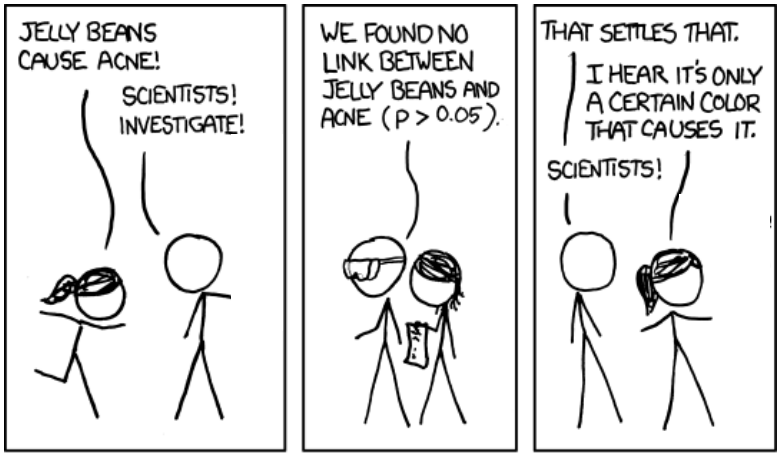
1. **Publication Bias**

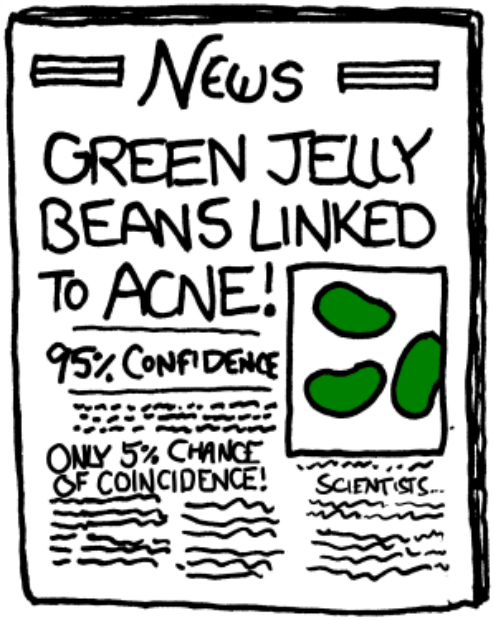
2. Problems with Hypothesis Testing Methodology

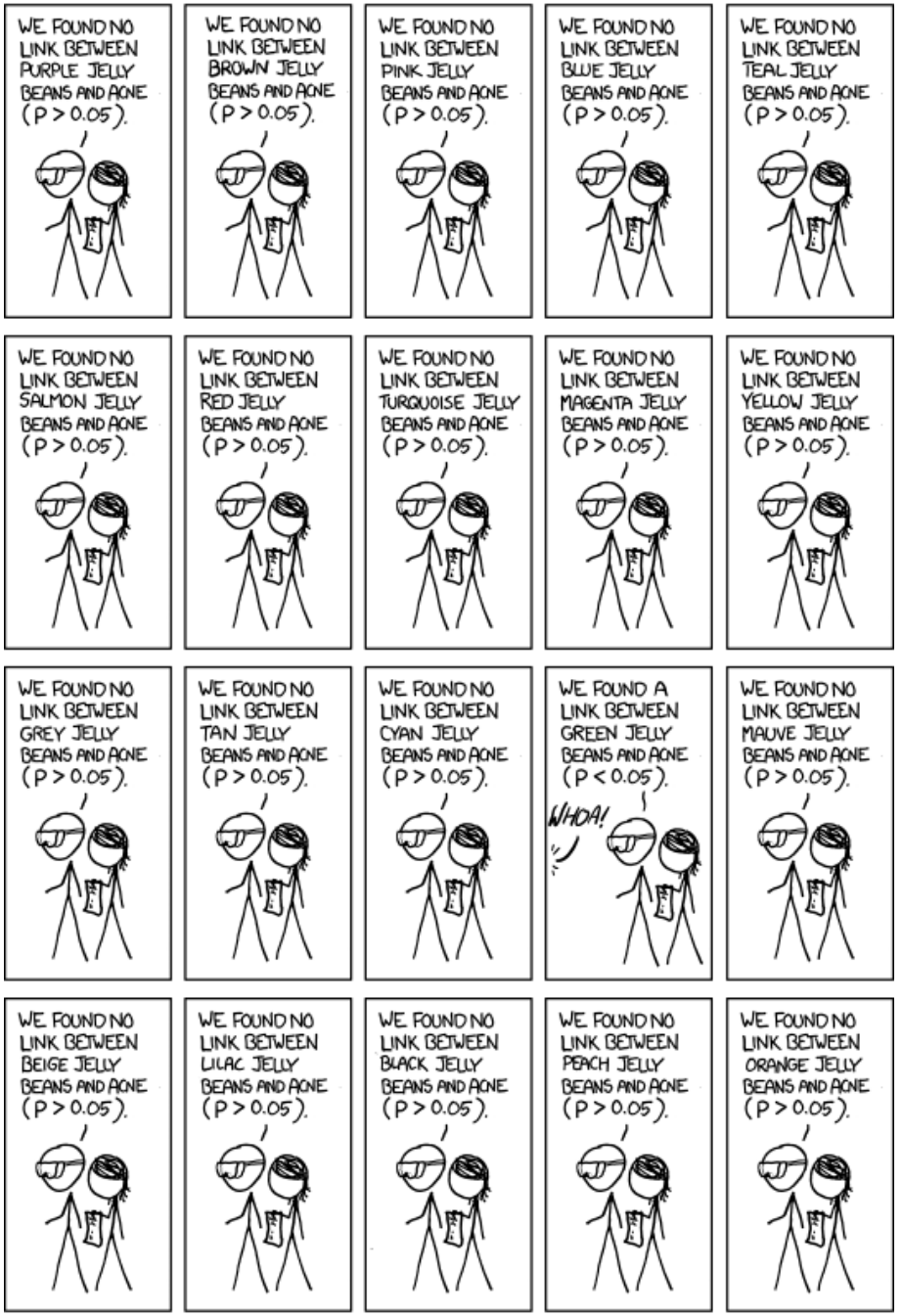

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($P > 0.05$).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!


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


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


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


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


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Reasons for Reproducibility Crisis

1. Publication Bias

2. **Problems with Hypothesis Testing Methodology**

Replication Crisis in Science

somehow related to use of p-values and significance testing…

Replication Crisis in Science

somehow related to use of p-values and significance testing…

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XlOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

Replication [illegible]s in Science

somehow [illegible] p-values and signi[illegible]ng...

Redefine Statistical Significance (to $p < 0.005$): Benjamin et al. 2017, incl. some of the most famous statisticians

ASA news
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
[illegible] North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XlOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

...lication... ...s in ...

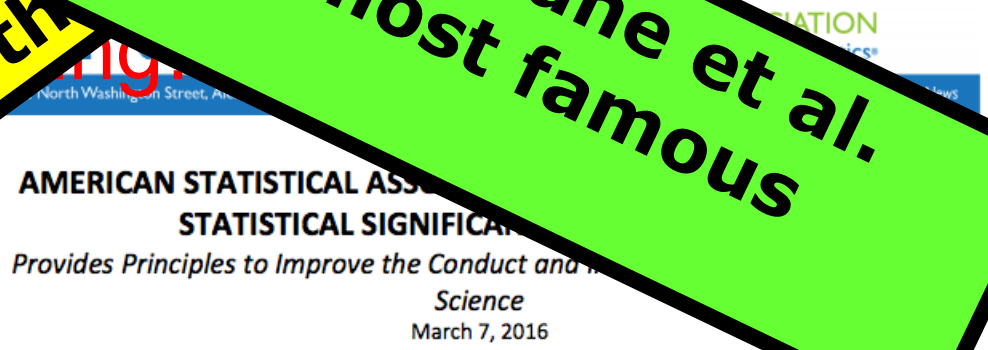
somehow ... and

sign... ...ng...

Redefine Sta... (to p < 0.005 ... incl. some of th...

Abandon Significance: McShane et al. (including some of the most famous statisticians) 2019

...Significance ...nin et al. 2017, ...famous statisticians

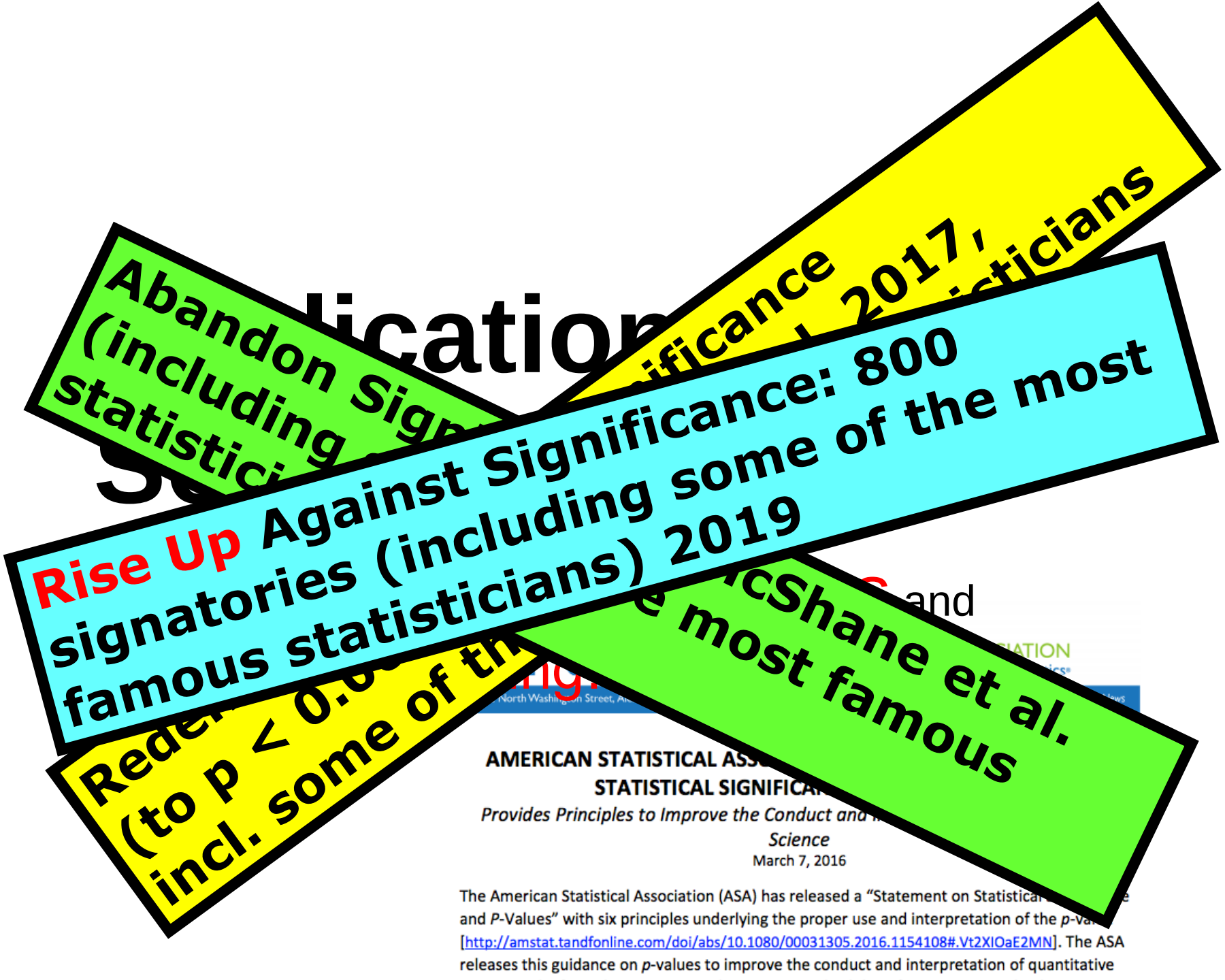


...ATION ...ics...

...North Washington Street, A... ...laws

AMERICAN STATISTICAL ASS...
STATISTICAL SIGNIFICA...
Provides Principles to Improve the Conduct and ...
Science
March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical ...e and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-va... [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XlOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative


Abandon Sig
(including
statistic
ificance
2017,
icians
Rise Up Against Significance: 800
signatories (including some of the most
famous statisticians) 2019
Rede
(to p < 0.
incl. some of th
cShane et al.
most famous
and
AMERICAN STATISTICAL ASS
STATISTICAL SIGNIFICA
Provides Principles to Improve the Conduct and
Science
March 7, 2016
The American Statistical Association (ASA) has released a "Statement on Statistica
and P-Values" with six principles underlying the proper use and interpretation of the p-va
releases this guidance on p-values to improve the conduct and interpretation of quantitative

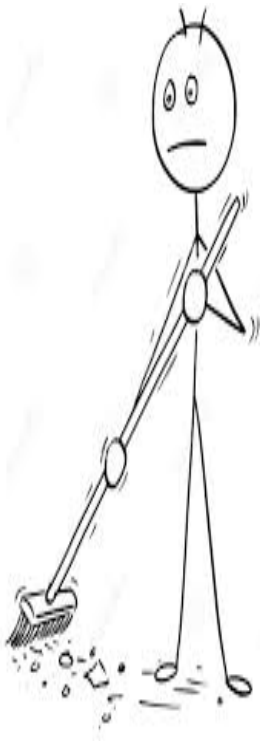
A Central P-value Problem

- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?

A Central P-value Problem

- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?
- **Current method:**

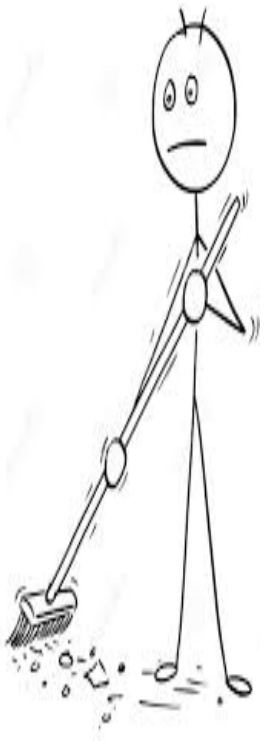
sweep data together and re-calculate p-value



A Central P-value Problem

- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?
- **Current method:**

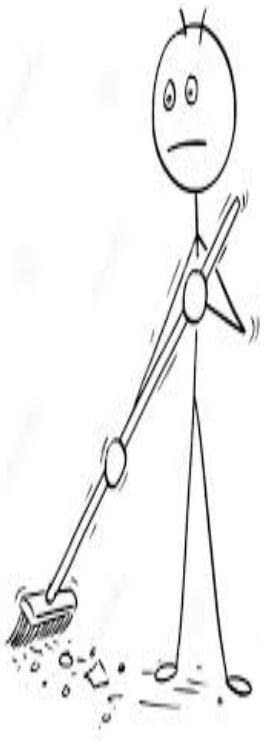
  sweep data together and re-calculate p-value
- **Is this p-hacking?** **YES**



A Central P-value Problem

- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?
- **Current method:**

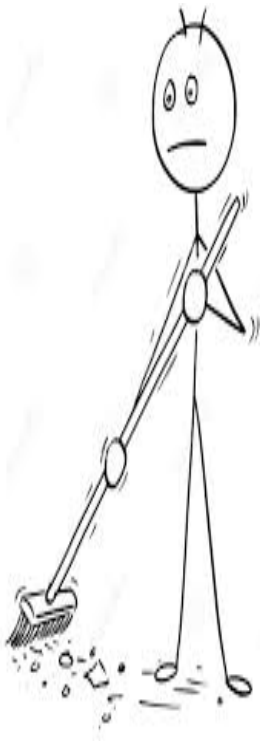
  sweep data together and re-calculate p-value
- **Is this p-hacking?** **YES**
- **Is this meta-analysis, a respected branch of science?** **YES**



A Central P-value Problem

- Suppose reseach group A tests medication, gets ‘promising but not conclusive’ result.
- ...**whence** group B tries again on new data.
- ...hmmm...still would like to get more evidence. Group C tries again on new data
- How to combine their test results?
- **Current method:**

  sweep data together and re-calculate p-value
- **Is this p-hacking?** **YES**
- **Is there a good way of doing this?** **NO**



CORONA-BCG Trials

- 750 hospital workers get BCG (anti-tuberculosis) vaccin, 750 get placebo vaccine.
- Our Role (originally): can we stop early if results very convincing? can we continue experiment with more (> 1500) subjects if results promising but inconclusive?
- **As of today: combine results with similar clinical trials done in Australia, Egypt and UK and …**
  - Rosanne Turner, Alexander Ly, Judith ter Schure



S is the new P

- We propose a generic replacement of the p-value that we call the S-value
- S-values handle **optional continuation** (to the next test (and the next, and ..)) without any problems

(can simply multiply S-values of individual tests, despite dependencies)

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, today we assume data X_1, X_2, ... are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{\frac{1}{2}\right\}$, $\Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **testing whether a coin is fair**

  Under P_θ , data are i.i.d. Bernoulli(θ)

  $\Theta_0 = \left\{\frac{1}{2}\right\}$, $\Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$ **Simple H_0**

  Standard test would measure frequency of 1s

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs.

  H_1 : $X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  σ^2 unknown ('nuisance') parameter

  $H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$

  $H_1 = \left\{ P_{\sigma,\mu} \middle| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \right\}$

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis

- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- Example: **t-test (most used test world-wide)**

  H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs.

  $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

  Composite H_0

  σ^2 unknown ('nuisance') parameter

  $H_0 = \{ P_\sigma | \sigma \in (0, \infty) \}$

  $H_1 = \left\{ P_{\sigma,\mu} \middle| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \right\}$

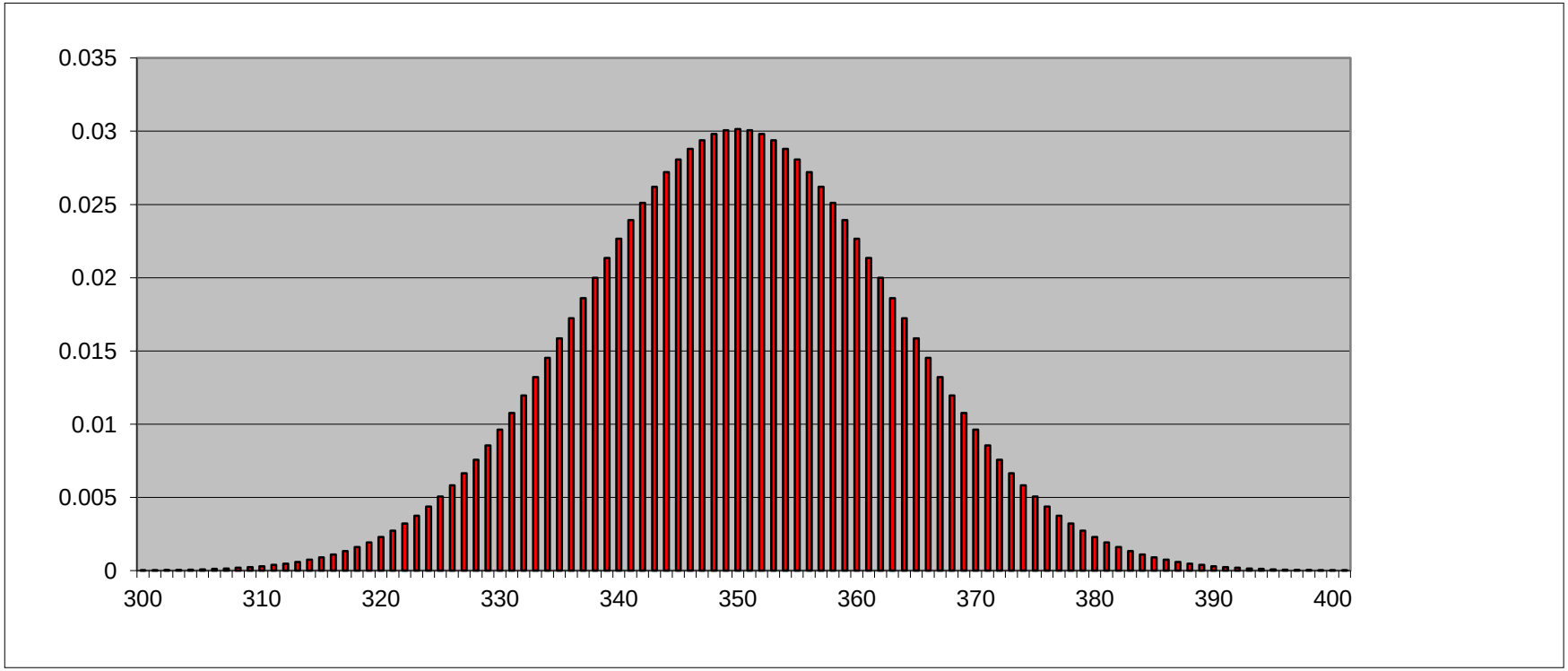
Standard Method: p-value, significance

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
- A ("nonstrict") **p**-value is a random **variable** (!) such that, for all $\theta \in \Theta_0$,

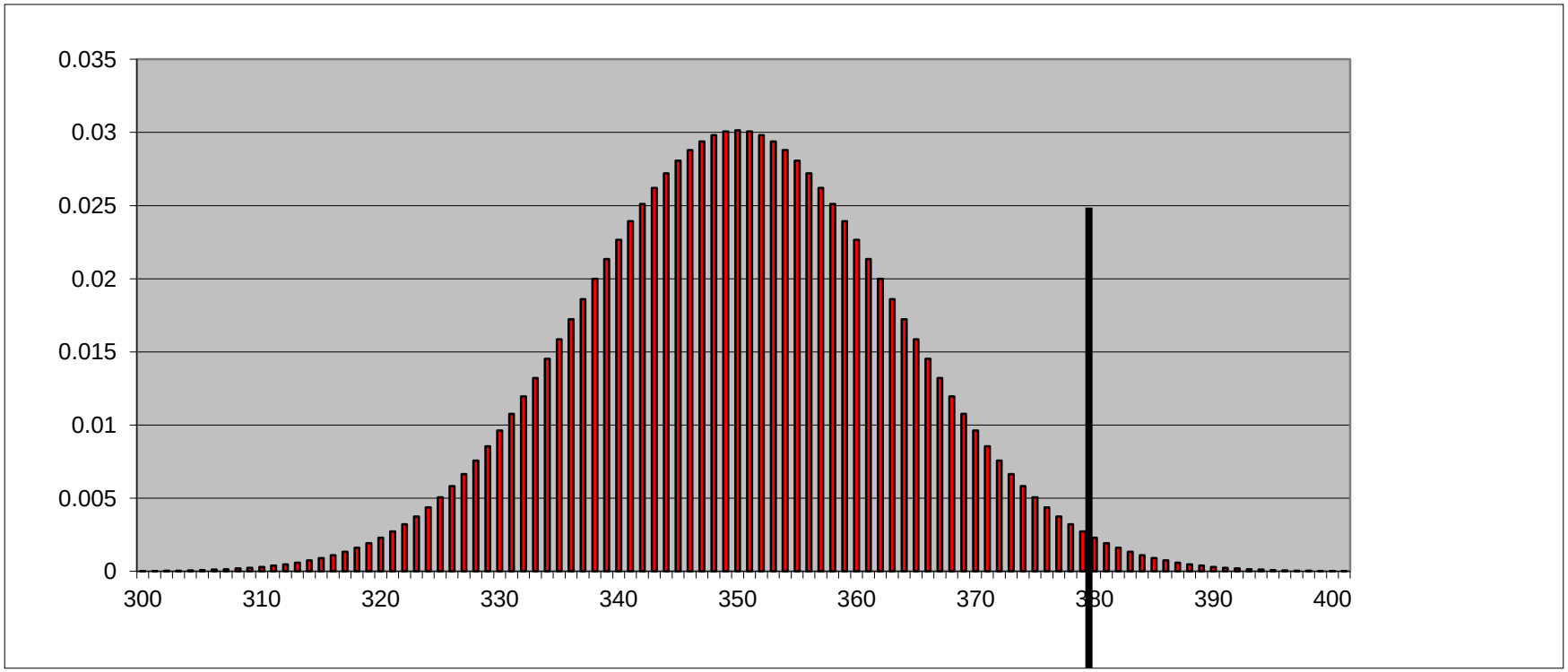
$$P_{\theta_0} (\mathbf{p} \leq \alpha) \leq \alpha$$

Coin Tossing Example, $n = 700$

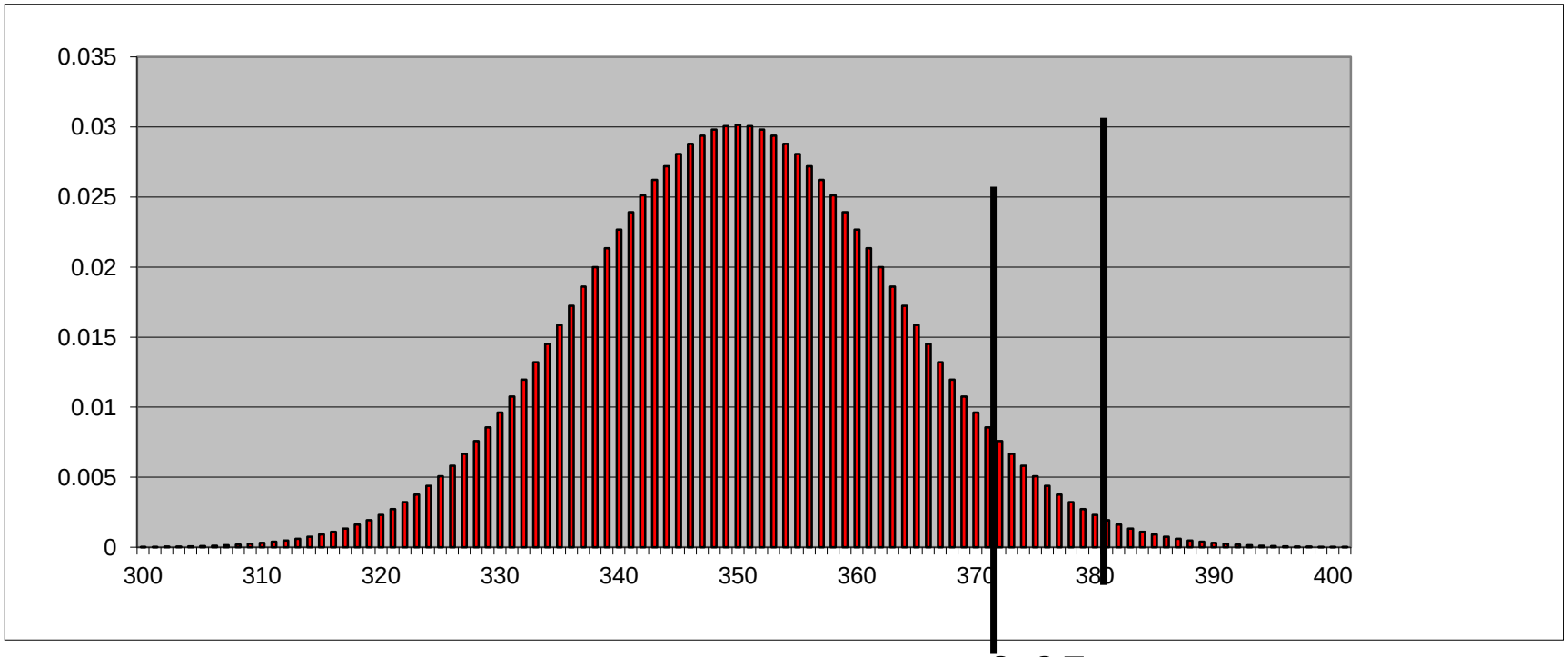
According to H_0 : $T := \sum_{i=1}^{700} X_i \sim \text{Bin}(0.5, 700)$



- We now do an experiment and we observe *T=380*. **The p-value is the probability that we would get this value, or an even smaller one**
- ≈ total probability mass right from black line. We find, for *T = 380,* that *p = 0.02*



- We determine (before experiment!) a ***significance level*** α and we 'reject' the null hypothesis iff $p \leq \alpha$
- This gives a **Type-I Error Probability bound** α
- **If we follow this decision rule consistently throughout our lives, then in long run we reject the null while it is correct at most 5% of the time**



P-value Problem: Combining Dependent Tests

- Suppose reseach group A tests medication, gets ‘almost significant’ result.
- ...whence group B tries again on new data. How to combine their test results?
  - **Standard methods for combining p-values (Fisher’s and Stouffer’s) require independence hence cannot be applied**
- **In “our” method, despite dependence, evidences can still be safely multiplied**

P-value Problem (b): Extending Your Test

- Suppose reseach group A tests medication, gets ‘almost significant’ result.
- **Sometimes group A can’t resist to test a few more subjects themselves...**
  - A recent survey revealed that **55% of psychologists** have succumbed to this practice
- But isn’t this just **cheating?**
  - **Not clear: what if you submit a paper and the *referee* asks you to test a couple more subjects? Should you refuse because it invalidates your p-values!?**

Menu

1. A problem with/limitation of p-values

2. S-Values and Safe Tests
   - ...solves the stop/continue problem
   - gambling interpretation

3. **The New Work:** Safe Testing for Composite H_0

S-Values: General Definition

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - Assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$.
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- An **S-value** for sample size n is a function $S : \mathcal{X}^n \to \mathbb{R}_0^+$ such that for **all** $P_0 \in H_0$, we have

$$\mathbf{E}_{X^n \sim P_0} \left[S(X^n) \right] \leq 1$$

First Interpretation: p-values

- Proposition: Let S be an S-value. Then $S^{-1}(X^n)$ is a conservative p-value, i.e. p-value with **wiggle room**:
- for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{S(X^n)} \leq \alpha \right) \leq \alpha$$

- Proof: just Markov's inequality!

$$P\left(S(X^n) \geq \alpha^{-1} \right) \leq \frac{\mathbf{E}[S(X^n)]}{\alpha^{-1}} \leq \alpha$$

Safe Tests

- The Safe Test against H_0 at level α based on S-value S is defined as the test which rejects H_0 if $S(X^n) \geq \frac{1}{\alpha}$
- Since for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{S(X^n)} \leq \alpha\right) \leq \alpha$$

- ....the safe test which rejects H_0 iff $S(X^n) \geq 20$, i.e. $S^{-1}(X^n) \leq 0.05$, has **Type-I Error** Bound of 0.05

Interpretation 1(b): Type-I Error

- The Safe Test against H_0 at level α based on S-value S is defined as the test which rejects H_0 if $S(X^n) \geq \frac{1}{\alpha}$
- Since for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{S(X^n)} \leq \alpha\right) \leq \alpha$$

- ....the safe test which rejects H_0 iff $S(X^n) \geq 20$, i.e. $S^{-1}(X^n) \leq 0.05$, has **Type-I Error** Bound of 0.05

First Example

1. H_0 and H_1 are point hypotheses:

$$S(X^n) = \frac{p_1(X^n)}{p_0(X^n)}$$

...is an S-value.

First Example

1. H_0 and H_1 are point hypotheses:

$$S(X^n) = \frac{p_1(X^n)}{p_0(X^n)}$$

...is an S-value, since

$$\mathbf{E}_{X^n \sim P_0}\left[\frac{p_1(X^n)}{p_0(X^n)}\right] = \sum_{x^n \in \mathcal{X}^n} p_0(x^n) \cdot \frac{p_1(x^n)}{p_0(x^n)} = \sum_{x^n \in \mathcal{X}^n} p_1(x^n) = 1.$$

...can be extended to general stopping times τ, densities, Radon-Nikodym derivatives etc...

Safe Tests are Safe under optional continuation

- Suppose we observe data $(X_1, Y_1), (X_2, Y_2), \ldots$
  - Y_i: side information, independent of X_i's

  ...coming in batches of size $n_1, n_2, \ldots, n_k$. Let $N_j := \sum_{i=1}^{j} n_i$
- We first evaluate some S-value S_1 on $(X_1, \ldots, X_{n_1})$.
- If outcome is in certain range (e.g. promising but not conclusive) and Y_{n_1} has certain values (e.g. 'boss has money to collect more data') then....
  we evaluate some S-value S_2 on $\left(X_{n_1+1}, \ldots, X_{N_2}\right)$,
  otherwise we **stop.**

Safe Tests are Safe

- We first evaluate S_1.
- If outcome is in certain range and Y_{n_1} has certain values then we evaluate S_2 ; otherwise we **stop.**
- If outcome of S_2 is in certain range and Y_{N_2} has certain values then we compute S_3 , else we **stop**.
- ...and so on
- ...when we finally stop, after say K data batches, we report as final result the product $S := \prod_{j=1}^{K} S_j$
- **First Result, Informally: any S composed of S-values in this manner is itself an S-value, irrespective of the stop/continue rule used!**

Safe Tests are Safe

Let S_1 be S-value on $\mathcal{X}^{n_1}$. For $j = 1,2, \ldots,$ let

$\mathcal{S}_{j+1}$ be any collection of S-values defined on $\mathcal{X}^{n_{j+1}}$

Let $g_j : \mathcal{X}^{N_j} \times \mathcal{Y}^{N_j} \to \{\texttt{stop}\} \cup \mathcal{S}_{j+1}$ be **arbitrary stop/continue strategy**, and:

Define $S := S_1(X^{n_1})$ **if** $g_1(X^{n_1}, Y^{n_1}) = \texttt{stop}$

else

Define $S := S_1(X^{n_1}) \cdot S_{g_1(X^{n_1}, Y^{n_1})}(X_{n_1+1}^{N_2})$ **if** $g_2(X^{N_2}, Y^{N_2}) = \texttt{stop}$

else

Define $S := S_1 \cdot \prod_{j=2}^{3} S_{g_{j-1}}$ **if** $g_3 = \texttt{stop}$

and so on...

Safe Tests are Safe

Theorem:

S , the end-product of all employed S-values S_1, S_{g_1}, S_{g_2}, ... is **itself an S-value**

Safe Tests are Safe

Theorem:

S , the end-product of all employed S-values S_1, S_{g_1}, S_{g_2}, ... is **itself an S-value**

Corollary: Type-I Error Guarantee Preserved under Optional Continuation

Suppose we combine S-values with arbitrary stop/continue strategy and reject H_0 when final S has $S^{-1} \leq 0.05$. Then resulting test is a safe test and our Type-I Error is guaranteed to be below 0.05!

Safe Tests are Safe

Theorem:

S , the end-product of all employed S-v
S_1, S_{g_1}, S_{g_2}, ... is **itself an S-value**

Corollary: Type-I Error Guara reserved under Optional Continu

Suppose we combi ues with arbitrary stop/continue and reject H_0 when final S has $S^{-1} \leq 0.0$ resulting test is a safe test and our Type-I is guaranteed to be below 0.05!

We solved a central problem of p-values!

Second, Main Interpretation:
Gambling!

Safe Testing = Gambling!

Kelly (1956)

- At time 1 you can buy ticket 1 for 1$. It pays off $S_1(X_1, \ldots, X_{n_1})$ $ after n_1 steps
- At time 2 you can buy ticket 2 for 1$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ $ after n_2 further steps.... and so on.

You may buy multiple and fractional nrs of tickets.

Safe Testing = Gambling!

- At time 1 you can buy ticket 1 for 1$. It pays off $S_1(X_1, \ldots, X_{n_1})$ $ after n_1 steps
- At time 2 you can buy ticket 2 for 1$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ $ after n_2 further steps.... and so on.

  You may buy multiple and fractional nrs of tickets.
- You start by investing 1$ in ticket 1.

Safe Testing = Gambling!

- At time 1 you can buy ticket 1 for 1$. It pays off $S_1(X_1, \ldots, X_{n_1})$ $ after n_1 steps
- At time 2 you can buy ticket 2 for 1$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ $ after n_2 further steps.... and so on.

  You may buy multiple and fractional nrs of tickets.
- You start by investing 1$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2.

Safe Testing = Gambling!

- At time 1 you can buy ticket 1 for 1$. It pays off $S_1(X_1, \ldots, X_{n_1})$ $ after n_1 steps
- At time 2 you can buy ticket 2 for 1$. It pays off $S_2(X_{n_1+1}, \ldots, X_{N_2})$ $ after n_2 further steps.... and so on.

  You may buy multiple and fractional nrs of tickets.
- You start by investing 1$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $S_1 \cdot S_2$ or you continue and buy $S_1 \cdot S_2$ tickets of type 3, and so on..

Safe Testing = Gambling!

- You start by investing 1$ in ticket 1.
- After n_1 outcomes you either stop with end capital M_1 or you continue and buy S_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $S_1 \cdot S_2$ or you continue and buy $S_1 \cdot S_2$ tickets of type 3, and so on...
- **S is simply your end capital**
- You don't expect to gain money, no matter what the stop/continuation rule since **none of individual gambles S_k are strictly favorable to you**

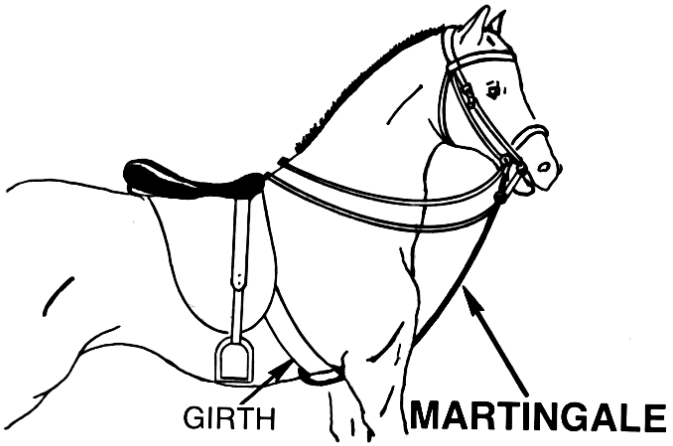
$$\mathbf{E}_{P_0}[S_1] \leq 1, \mathbf{E}_{P_0}[S_2] \leq 1, \ldots \Rightarrow \mathbf{E}_{P_0}[S] \leq 1$$

Safe Testing = Gambling!

- You start by investing 1$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $S_1 \cdot S_2$ or you continue and buy $S_1 \cdot S_2$ tickets of type 3, and so on...
- **S is simply your end capital**
- Your don't expect to gain money, no matter what the stop/continuation rule since **none of individual gambles S_k are strictly favorable to you**
- Hence a **large value of S** indicates that something very unlikely has happened under H_0 ...

Technical Aside

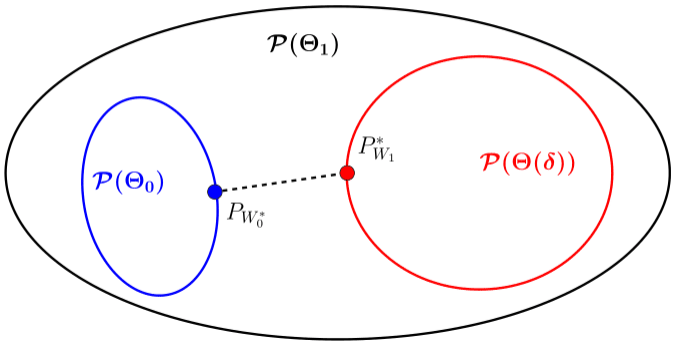
- Technically, we can view the process $(S_1, S_1 \cdot S_2, S_1 \cdot S_2 \cdot S_3, \ldots)$ as a nonnegative supermartingale.
- The Type-I Error Probability result is then Ville's (1939) Inequality, and the Proof is Immediate by Doob's Optional Stopping Theorem



Menu

1. S-Values and Safe Tests
   - solve the optional continuation problem
   - gambling interpretation
2. Safe Testing: relation to Bayes
3. Safe Testing: **composite H_0**, main theorem:

The JIPr (Joint Information Projection) provides the optimal S-Value



Safe Testing and Bayes

- **Bayes factor hypothesis testing** **(Jeffreys '39)**
  with $H_0 = \{ p_\theta | \theta \in \Theta_0 \}$ vs $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:
  Evidence in favour of H_1 measured by

$$\frac{p_{W_1}(X_1, \ldots, X_n)}{p_{W_0}(X_1, \ldots, X_n)}$$

where

$$p_{W_1}(X_1, \ldots, X_n) := \int_{\theta \in \Theta_1} p_\theta(X_1, \ldots, X_n) dW_1(\theta)$$

$$p_{W_0}(X_1, \ldots, X_n) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) dW_0(\theta)$$

Safe Testing and Bayes, simple H_0

Bayes factor hypothesis testing

between $H_0 = \{p_0\}$ and $H_1 = \{p_\theta | \theta \in \Theta_1\}$:

Bayes factor of form

$$M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)}$$

Note that (no matter what prior W_1 we chose)

$$\mathbf{E}_{X^n \sim P_0}[M(X^n)] =$$

$$\int p_0(x^n) \cdot \frac{\bar{p}_{W_1}(X^n)}{p_0(x^n)} dx^n = \int \bar{p}_{W_1}(x^n) dx^n = 1$$

Safe Testing and Bayes, simple H_0

Bayes factor hypothesis testing

between $H_0 = \{p_0\}$ and $H_1 = \{p_\theta | \theta \in \Theta_1\}$:

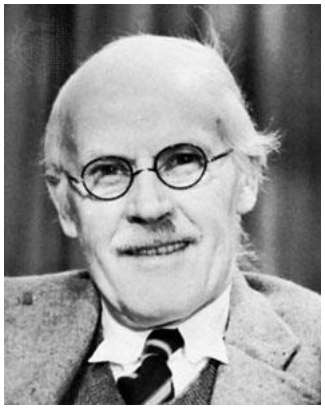
Bayes factor of form

$$M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)}$$

Note that (no matter what prior W_1 we chose)

$$\mathbf{E}_{X^n \sim P_0}[M(X^n)] = 1$$

The Bayes Factor for Simple H_0 is an S-value!



Composite H_0: Bayes may not be Safe!

Bayes factor given by $M(X^n) := \frac{p_{W_1}(X_1, \ldots, X_n)}{p_{W_0}(X_1, \ldots, X_n)}$

S-value requires that **for all** $P_0 \in H_0$:

$$\mathbf{E}_{X^n \sim P_0}\left[M(X^n)\right] \leq 1$$

...but for a Bayes factor we can only guarantee that

$$\mathbf{E}_{X^n \sim P_{W_0}}\left[M(X^n)\right] \leq 1$$

Composite H_0: Bayes may not be Safe!

Bayes factor given by $M(X^n) := \dfrac{p_{W_1}(X_1, \ldots, X_n)}{p_{W_0}(X_1, \ldots, X_n)}$

- In general Bayes factors with composite H_0 are not S-values
- ...but there do exist *very special priors* W_1^* , W_2^* **(sometimes highly unlike priors that "Bayesian" statisticians tend to use!)** for which Bayes factors become S-values
- I will now show you how to construct such priors!

First: How to design S-Values?

- Suppose we are willing to admit that we'll only be able to tell H_0 and H_1 apart if $P \in H_0 \cup H_1'$ for some $H_1' \subset H_1$ that excludes points that are 'too close' to H_0 e.g.

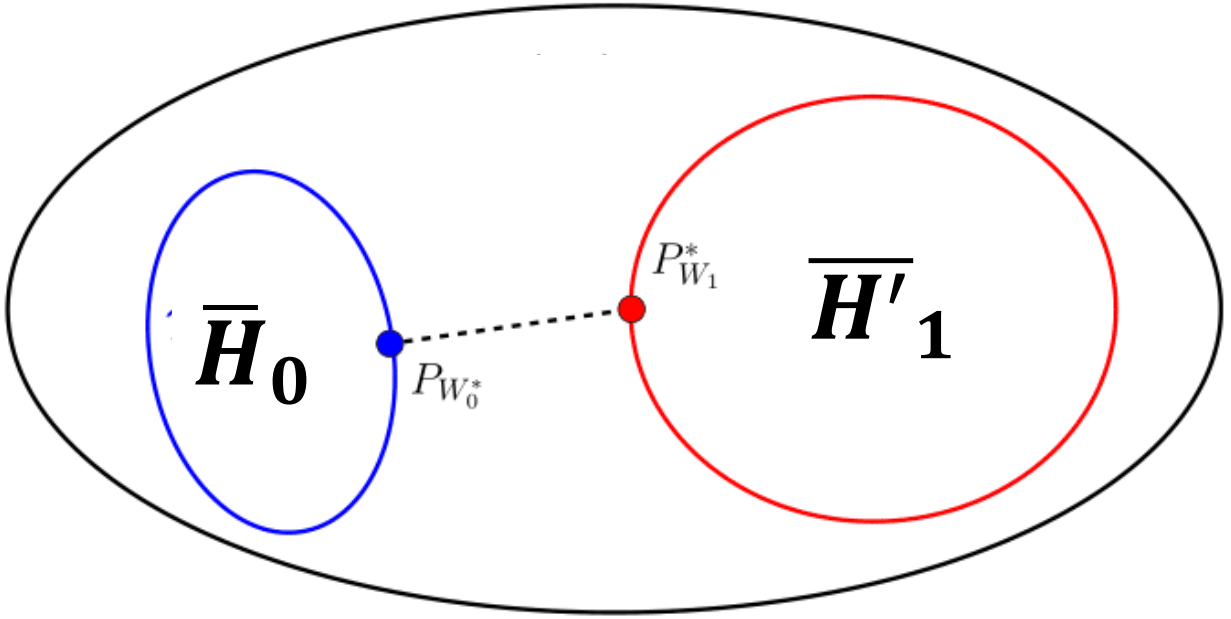
$$H_1' = \{P_\theta : \theta \in \Theta_1'\}, \Theta_1' = \{\theta \in \Theta_1 : \inf_{\theta_0 \in \Theta_0} \|\theta - \theta_0\|_2 \geq \delta\}$$

The best S-Value is given by the Joint Information Projection (JIPr)

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$\mathcal{W}_1$ set of all priors (prob distrs) on Θ_1'

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$



Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Here D is the relative entropy or Kullback-Leibler divergence, the central divergence measure in information theory and large deviations

$$D(P\|Q) := \mathbf{E}_{X^n \sim P} \left[\log \frac{p(X^n)}{q(X^n)} \right]$$

(can give measure-theoretic definition making it well-defined even if P and Q not abs. cont.)

Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Suppose (W_1^*, W_0^*) exists. Then $S^* := \frac{p_{W_1^*}(X^n)}{p_{W_0^*}(X^n)}$

is (a) an S-value relative to H_0. (b)….

Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Suppose (W_1^*, W_0^*) exists. Then $S^* := \dfrac{p_{W_1^*}(X^n)}{p_{W_0^*}(X^n)}$

is (a) an S-value. (b) In fact it is the **GROW** S-value, i.e.

$$\inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S^*] = \sup_S \inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S]$$

Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Suppose (W_1^*, W_0^*) exists. Then $S^* := \dfrac{p_{W_1^*}(X^n)}{p_{W_0^*}(X^n)}$

is (a) an S-value. (b) In fact it is the **GROW** S-value, i.e.

$$\inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S^*] = \sup_S \inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S]$$

and (c) ,

$$= \min_{W_1 \in \mathcal{W}_1} \min_{W_0} D(P_{W_1} \| P_{W_0})$$

GROW: an analogue of Power

- The GROW (growth-optimal in worst-case) S-value relative to $H_{1,\delta}$ is the S-value achieving

$$\sup_{S} \inf_{\theta \in \Theta_1'} \mathbf{E}_{X^n \sim P_\theta}[\log S]$$

where the supremum is over all S-values relative to H_0

- ...so we don't expect to gain anything when investing in S under H_0
- ...but among all such S we pick the one(s) that make us rich fastest if we keep reinvesting in new gambles under H_1

Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0} D(P_{W_1} \| P_{W_0})$$

This is really a minimax Theorem in disguise.

It does not follow from standard minimax theorems such as Sion's or Fan's

is (a) an S-value. (b) In fact it is the **GROW** S-value, i.e.

$$\inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S^*] = \sup_S \inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S]$$

and (c) ,

$$= \min_{W_1 \in \mathcal{W}_1} \min_{W_0} D(P_{W_1} \| P_{W_0})$$

Example: Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

σ^2 unknown ('nuisance') parameter

$$H_0 = \{ P_\sigma | \sigma \in (0, \infty) \} \quad H_1 = \left\{ P_{\sigma,\mu} \middle| \sigma \in (0, \infty), \mu \in \mathbb{R} \setminus \{0\} \right\}$$

- In general Bayes factors are not S-values
- But lo and behold, Jeffreys' uses very special priors and his Bayes factor is an S-value, so his Bayesian t-test is a Safe Test!

Example: Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

- Jeffreys uses improper right-Haar prior $w(\sigma) = 1/\sigma$ within both models, and uses Cauchy on $\frac{\mu}{\sigma}$
- In fact, for right-Haar prior combined with arbitrary prior on effect size μ/σ we get that S has same distr. under all $P \in H_0$, and $\mathbf{E}_{X^n \sim P}(S) = 1$

Example: Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

- Jeffreys uses improper right-Haar prior $w(\sigma) = 1/\sigma$ within both models, and uses Cauchy on $\frac{\mu}{\sigma}$
- In fact, for right-Haar prior combined with arbitrary prior on effect size μ/σ we get that S has same distr. under all $P \in H_0$, and $\mathbf{E}_{X^n \sim P}(S) = 1$
- But the GROW S-value under the constraint that $|\mu/\sigma| \geq \delta_0$ is given by right-Haar + 2-point prior on μ/σ with probability ½ on δ_0 and ½ on - δ_0

Example:
Jeffreys' (1961) Bayesian t-test

H_0: $X_i \sim_{i.i.d.} N(0, \sigma^2)$ vs. $H_1 : X_i \sim_{i.i.d.} N(\mu, \sigma^2)$ for some $\mu \neq 0$

For general composite testing with nuisance parameters admitting a group structure (scale, location, rotation invariance), the improper right Haar prior always gives S-values

prior on effect size μ/σ we get that S has same distr. under all $P \in H_0$, and $\mathbf{E}_{X^n \sim P}(S) = 1$

- But the GROW S-value under the constraint that $|\mu/\sigma| \geq \delta_0$ is given by right-Haar + 2-point prior on μ/σ with probability ½ on δ_0 and ½ on - δ_0

Empirical Results

- We compare ourselves to a standard use of the t-test:
- We fix $\alpha = 0.05$ and power $\beta = 0.8$ and 'minimum clinically relevant effect size' δ_0
- We determine the n_{St}, n_{Jef} and n_{Safe} at which
  1. the standard t-test
  2. Jeffreys' Bayesian t-test (Haar+Cauchy)
  3. The GROW t-test (Haar+ (1/2,1/2) on δ')

  has power at least 0.8 under H_1 if effect size $|\delta| \geq \delta_0$

Empirical Results

- We compare ourselves to a standard use of the t-test:
- We fix $\alpha = 0.05$ and power $\beta = 0.8$ and 'minimum clinically relevant effect size' δ_0
- We determine the n_{St}, n_{Jef} and n_{Safe} at which
  1. the standard t-test
  2. Jeffreys' Bayesian t-test (Haar+Cauchy)
  3. The GROW t-test (Haar+ (1/2,1/2) on δ')

  has power at least 0.8 under H_1 if effect size $|\delta| \geq \delta_0$

n_{St} about 30% smaller than n_{Jef} 20% smaller than n_{Safe}

BUT: we can do optional stopping with S-Values!

Empirical Results

- We compare ourselves to a standard use of the t-test:

Some S-Values preserve Type-I Errors under optional stopping (stop at each sample point you want) a much stronger property than preservation under optional 'continuation' (stop after each trial you want).

Open Question 3: Characterize when this is possible!

BUT: we can do optional stopping with S-Values!

Example 2: Independence Testing/2x2 tables

- $X_i \in \{0,1\}$; $Z_i \in \{m, f\}$

- H_0: $X_1, X_2, \ldots, X_n \mid Z_1, \ldots, Z_n$ iid Bernoulli(θ),
- H_1: $X_1, X_2, \ldots, X_n \mid Z_1, \ldots, Z_n$ independent but
  $P(X_i = 1 \mid Z_i = m) = \theta_m$
  $P(X_i = 1 \mid Z_i = f) = \theta_f \neq \theta_m$
- Are **both populations same or different?**

2x2 Contingency Tables

- Comparable Results (for fixed sample size, need about 20% more data points than with Fisher's exact test) but now the JIPr gives us a prior which does not even remotely resemble common priors
- For $\Theta_1 = \{(\theta_{f,\theta_m}) : \theta_f \geq \theta_m + \delta\}$

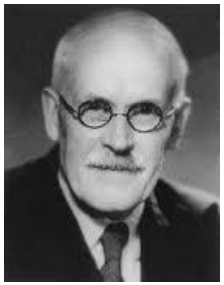
$$S^* := \frac{p_{(1/2+\delta,1/2-\delta)}(x^n \mid z^n)}{p(1/2)(z^n)}$$

- Major surprise: The JIPr priors W_0^* and W_1^* are both point priors, and again we can do **full optional stopping.**

Three Philosophies of Testing

Jerzy Neyman: alternative exists, “inductive behaviour”, ‘significance level’ and power

Sir Ronald Fisher: test statistic rather than alternative, p-value indicates “unlikeliness”



Sir Harold Jeffreys: **Bayesian**, alternative exists, absolutely no p-values

J. Berger (2003, IMS Medaillion Lecture): *Could Neyman, Fisher and Jeffreys have agreed on testing?*

...we think we have a unification/correction of the central ideas

Read/Do more?

- G., De Heide, Koolen. ***Safe Testing,*** *Arxiv 2019*
- **R-PACKAGES** for safe t-test, 2x2 tables
- Related work: Howard, **Ramdas** et al.'s *Uniform, Nonparametric, Nonasymptotic Confidence Sequences*

- *"Safe Confidence Intervals" are also possible...*
- ***G. Shafer and V. Vovk****:* Game-Theoretic Probability (2019)

Experimental Results/Conclusion

- With the GROW safe t-test you need to reserve about 20% more data points to obtain the same power at the same effect size, compared to the standard t-test
- …but you are allowed to do *optional stopping*: stop as soon as $S \geq 20$!
- Then **on average** you need about the same amount of data as with the standard t-test
- I wonder: is there a good excuse *not* to use the Safe t-test?

Safe Tests are Safe

- S_j may be same function as S_{j-1}, e.g. (simple H_0)

$$S_1 = \frac{\int_{\Theta_1} p_\theta(X_1, \ldots, X_{n_1}) dW(\theta)}{p_0(X_1, \ldots, X_{n_1})} \qquad S_2 = \frac{\int_{\Theta_1} p_\theta(X_{n_1+1}, \ldots, X_{N_2}) dW(\theta)}{p_0(X_{n_1+1}, \ldots, X_{N_2})}$$

- But choice of jth S-value S_j may also depend on previous X^{N_j}, Y^{N_j} , e.g.

$$S_2 = \frac{\int_{\Theta_1} p_\theta(X_{n_1+1}, \ldots, X_{N_2}) dW(\theta | X_1, \ldots, X_{n_1})}{p_0(X_{n_1+1}, \ldots, X_{N_2})}$$

and then (full compatibility with Bayesian updating)

$$S_1 \cdot S_2 = \frac{\int p_\theta(X_1, \ldots, X_{N_2}) dW(\theta)}{p_0(X_1, \ldots, X_{N_2})}$$