TODAY: Safe Testing

1. Reproducibility Crisis/Problems with p-values
2. The S-Value
3. Optional Continuation
4. Optional Stopping vs Optional Continuation
5. Gambling Interpretation, Again
6. Types of S-Values / GROW S-Values

[Next Week No Lecture – Homework due Mo May 11]

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

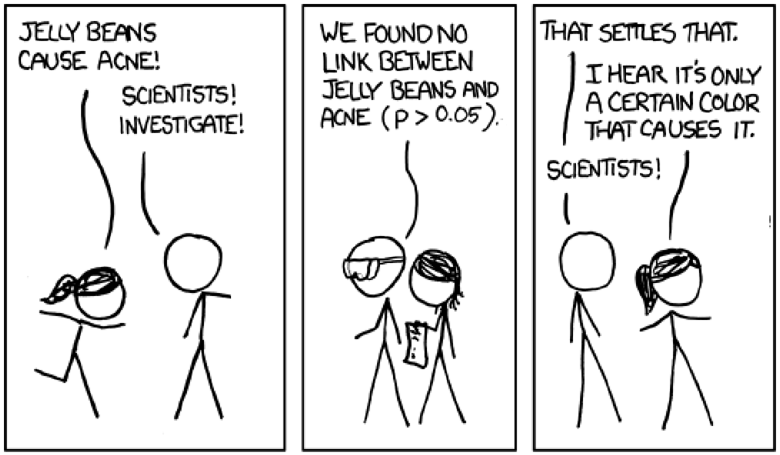
Reproducibility Crisis

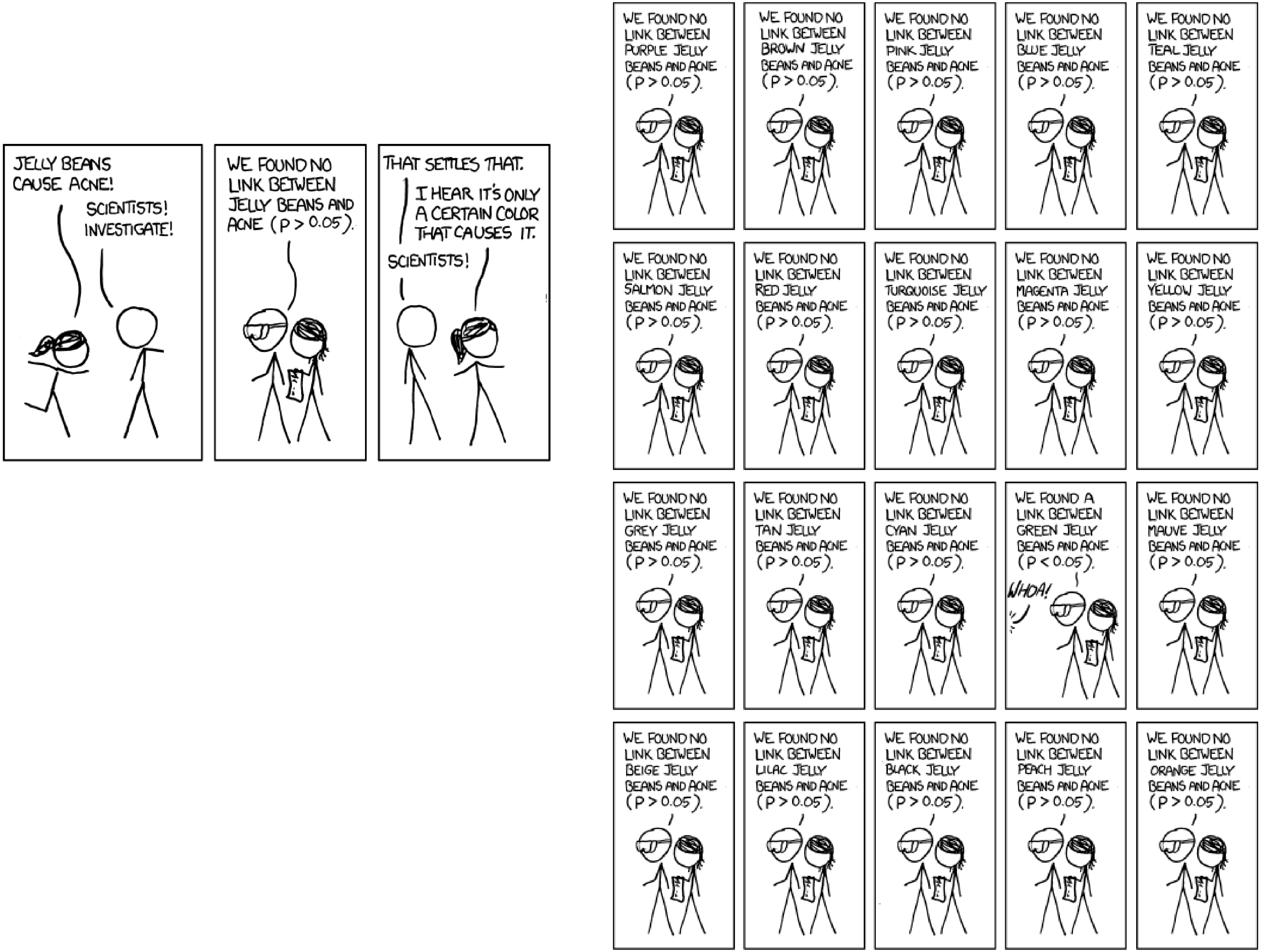
Cover Story of Economist (2013), Science (2014)

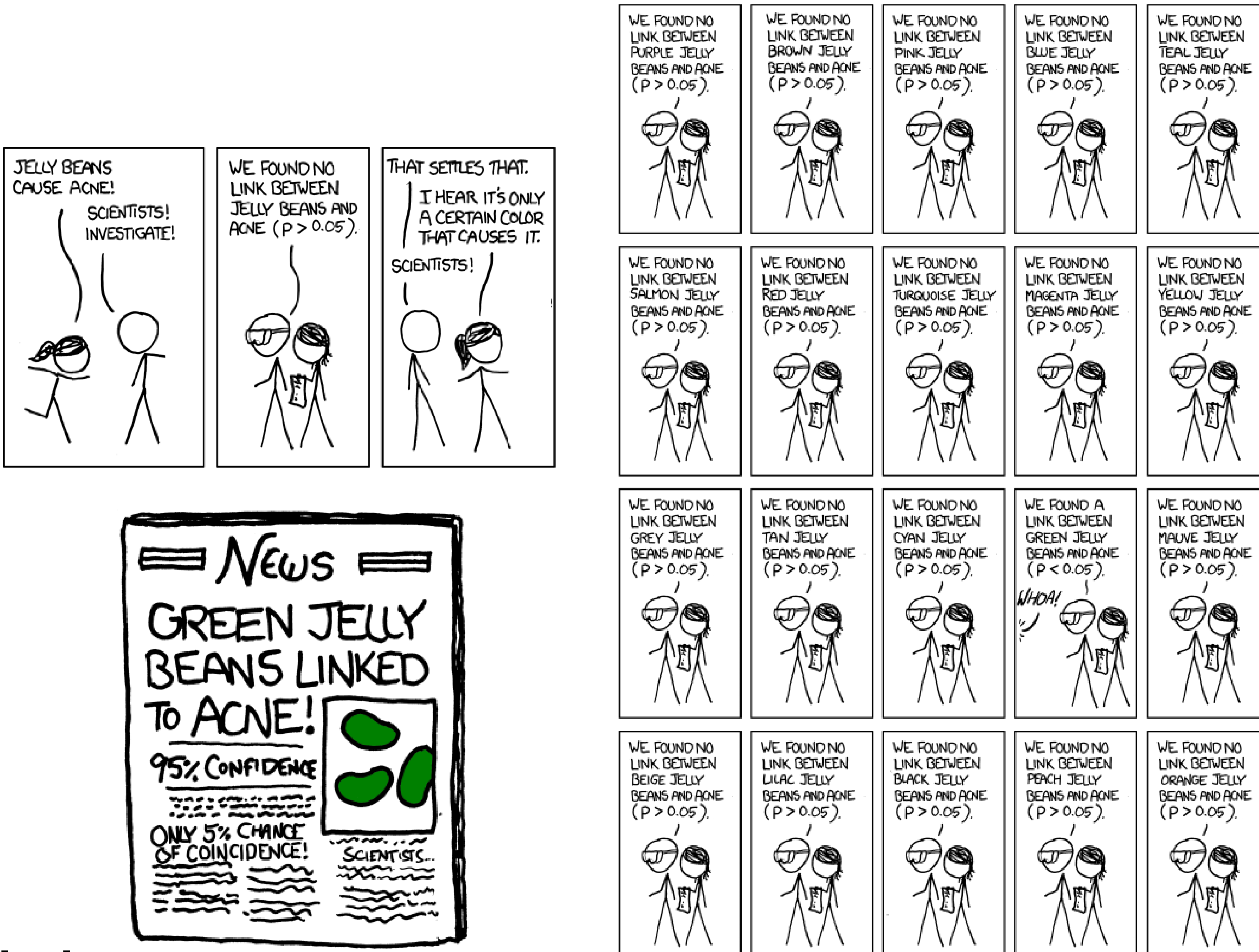
Reasons for Reproducibility Crisis

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

Xkcd.org

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…

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 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

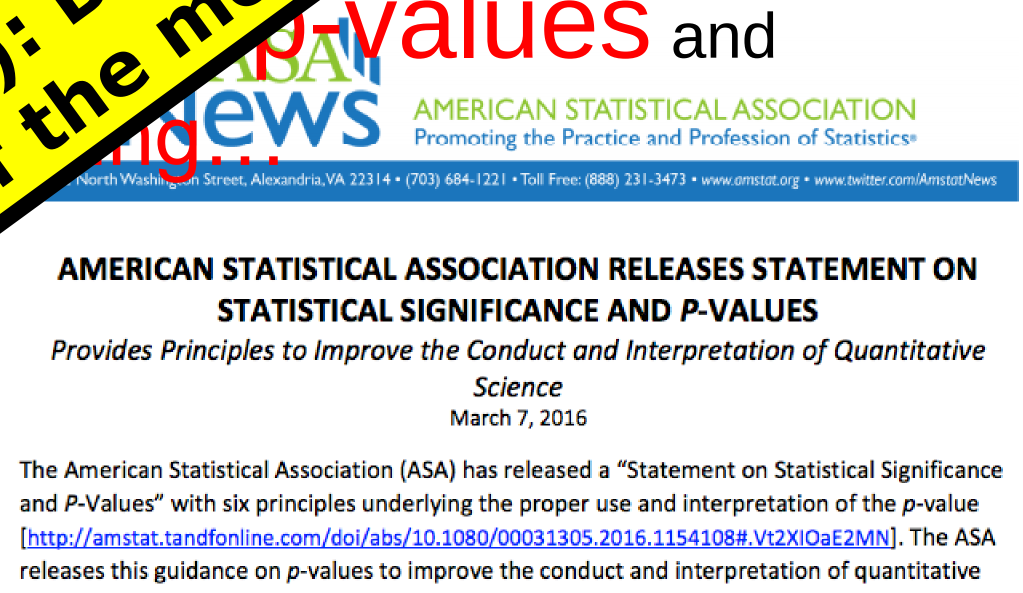
March 7, 2016

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#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

Replication … s in Science

somehow … p-values and signi… ing…

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®

… 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#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

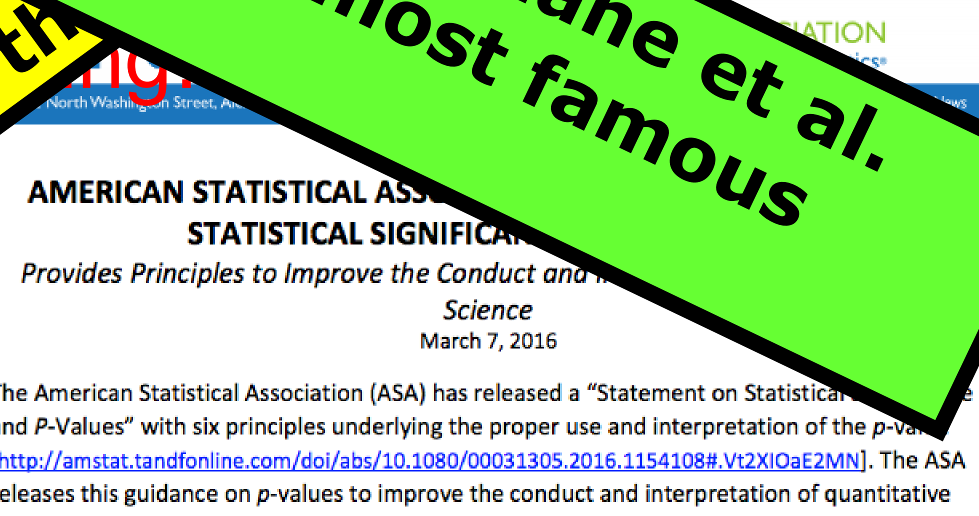
Abandon Significance: McShane et al. (including some of the most famous statisticians) 2019

Redefine Sta... Significance (to p < 0.005) ...in et al. 2017, incl. some of th... famous statisticians

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

somehow ... and

sig... ing...


...IATION

North Washington Street, Al...

AMERICAN STATISTICAL ASS...

STATISTICAL SIGNIFICA...

Provides Principles to Improve the Conduct and I...

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#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

...lication...

S...

...ificance ... 2017, ...ticians

Redef... (to p < 0.0... incl. some of th...

Abandon Sign... (including ... statistici... ...cShane et al. ...e most famous

Rise Up Against Significance: 800 signatories (including some of the most famous statisticians) 2019

...and

...IATION

...ics

...ing!

North Washington Street, Al...

...News

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 ... 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#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

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 Method 1: sweep data together, recompute p-value. *This is not correct; type-I error guarantee does not hold any more*
  - **Standard method 2: use Fisher’s method for combining p-values. Again not correct, since tests cannot be viewed as independent**
  - **Standard method 3: multiply p-values. Just plain wrong – a mortal sin!**
- **With the type of “p-value” introduced here, 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 (and then treat data as if large sample size was determined in advance)
- 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!?**

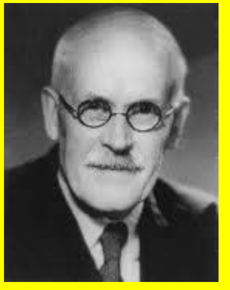
S is the new P

- 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)

S is the new P

S-values have Fisherian, Neymanian and Bayes-Jeffreys' aspects to them, all at the same time



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

individual tests, despite dependencies)

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

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

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!

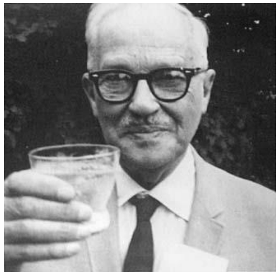
$$P\left(S(X^n) \geq \alpha^{-1} \right) \leq \frac{\mathbf{E}[S(X^n)]}{\alpha^{-1}} = \alpha$$

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 S^{-1} is a conservative p-valuue...

- ....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



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)

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

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

Safe Testing and Bayes, simple H_0

Bayes factor hypothesis testing

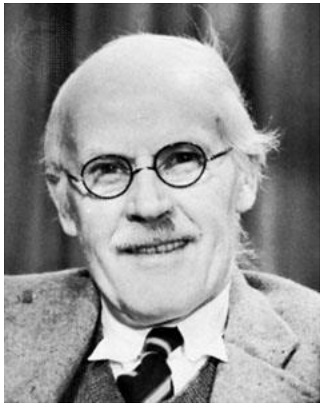
between $H_0 = \{ p_0 \}$ and $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

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!

Default S-Value $\neq$ Neyman

1. H_0 and H_1 are point hypotheses – then default S-value is:

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

... the safe test based on S looks a bit like, but is *not* a standard Neyman-Pearson test. **more conservative**

Safe Test: reject if $S(X^\tau) \geq 1/\alpha$

NP: reject if $S(X^\tau) \geq 1/B$ with B s.t. $P_0(S(X^\tau) \geq B) = \alpha$

Safe Tests are Safe under optional continuation

- Suppose we observe data $(X_1, Y_1), (X_2, Y_2), \ldots$
  - Y_i: side information

  ...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

- 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)

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

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 $g_j : \mathcal{X}^{N_j} \times \mathcal{Y}^{N_j} \rightarrow \{\texttt{stop}\} \cup \mathcal{S}_{j+1}$ 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 $S := S_1(X^{n_1}) \cdot S_{g_1(X^{n_1}, Y^{n_1})}(X^{N_2}_{n_1+1})$ **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:

Suppose that for all i, $X_i \perp \left(X^{i-1}, Y^{i-1}\right)$. Then S, the end-product of all employed S-values $S_1, S_{g_1}, S_{g_2}, \ldots$ is **itself an S-value**

- Technically, the process $(S_1, S_1 \cdot S_{g_1}, S_1 \cdot S_{g_1} \cdot S_{g_2}, \ldots)$ is a **nonnegative supermartingale** (Ville '39) and the theorem is proved using Doob's optional stopping theorem

Safe Tests are Safe

Theorem:

S , the end-product of all employed S-values $S_1, S_{g_1}, S_{g_2}, \ldots$ 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[illegible]
$S_1, S_{g_1}, S_{g_2}, \ldots$ is **itself an S-value**

Corollary: Type-I Error Guara[illegible]reserved under Optional Continu[illegible]

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

We solved a central problem of p-values!

Generalizing the Result

Theorem says:

- Let $S_{\langle j+1 \rangle} := S_{g_j(Z^{N_j})}(X_{N_j+1}^{N_{j+1}})$ where $g_j : \mathcal{Z}^{N_j} \to \{\texttt{stop}\} \cup \mathcal{S}_{j+1}$ s.t. $\forall j, \forall k \in \mathcal{S}_j$, $S_k : \mathcal{X}^{n_j} \to \mathbb{R}_0^+$ is S-value: $\forall P \in H_0 : \mathbf{E}_P[S_k] \leq 1$

- Suppose that for all i, $X_i \perp \left(X^{i-1}, Y^{i-1}\right)$.
- Let τ be smallest j such that $g_j\left(Z^{N_j}\right) = stop$
- Then $S := \prod_{j=1}^{\tau} S_{\langle j \rangle}$ is an S-value

Generalizing the Result

Theorem says:

- ~~Let $S_{\langle j+1\rangle} := S_{g_j(Z^{N_j})}(X^{N_{j+1}}_{N_j+1})$ where $g_j : \mathcal{Z}^{N_j} \to \{\texttt{stop}\} \cup \mathcal{S}_{j+1}$~~
  For all *j*, let $g_j : \mathcal{Z}^{N_j} \to \{\texttt{continue}, \texttt{stop}\}$ and let …
  ~~$\forall j, \forall k \in \mathcal{S}_j, S_k : \mathcal{X}^{n_j} \to \mathbb{R}_0^+$ is S value: $\forall P \in H_0 : \mathbf{E}_P[S_k] \leq 1$~~
- $S_{\langle j+1\rangle} : \mathcal{Y}^{N_j} \times \mathcal{X}^{N_{j+1}} \to \mathbb{R}_0^+$ s.t. $\forall P \in H_0 : \mathbf{E}_P[S_{\langle j+1\rangle} \mid Z^{N_j}] \leq 1$
- ~~Suppose that for all i, $X_i \perp (X^{i-1}, Y^{i-1})$.~~
- Let τ be smallest j such that $g_j(Z^{N_j}) = stop$
- Then $S := \prod_{j=1}^{\tau} S_{\langle j\rangle}$ is an S-value

Generalizing the Result

Theorem says: for all j,

- Let $X_{\langle j \rangle} = (X_{N_{j-1}+1}, \ldots, X_{N_j})$; $X^{\langle j \rangle} = (X_1, \ldots, X_{N_j})$; similarly for Y, Z
- Let $S_{\langle j+1 \rangle} : \mathcal{Y}^{\langle j \rangle} \times \mathcal{X}^{\langle j+1 \rangle} \to \mathbb{R}_0^+$ be a function such that $\forall P \in H_0 : \mathbf{E}_P[S_{\langle j+1 \rangle} \mid Z^{\langle j \rangle}] \leq 1$
- Let τ be a **stopping time** for the process $(Z_{\langle 1 \rangle}, Z_{\langle 2 \rangle}, \ldots)$
- Then $S := \prod_{j=1}^{\tau} S_{\langle j \rangle}$ is an S-value

Optional Stopping vs Continuation

- Let $X_{\langle j \rangle} = (X_{N_{j-1}+1}, \ldots, X_{N_j})$; $X^{\langle j \rangle} = (X_1, \ldots, X_{N_j})$
- Let $S_{\langle j+1 \rangle} : \mathcal{Y}^{\langle j \rangle} \times \mathcal{X}^{\langle j+1 \rangle} \to \mathbb{R}_0^+$ be a function such that $\forall P \in H_0 : \mathbf{E}_P[S_{\langle j+1 \rangle} \mid Z^{\langle j \rangle}] \leq 1$
- Let τ be a **stopping time** for the process $(Z_{\langle 1 \rangle}, Z_{\langle 2 \rangle}, \ldots)$
- Then $S := \prod_{j=1}^{\tau} S_{\langle j \rangle}$ is an S-value

Optional Continuation: we have *batches* of data $Z_{\langle 1 \rangle}, Z_{\langle 2 \rangle}, \ldots$ and we can do optional stopping at the batch-level (and obtain an S-Value and preserve Type I error guarantees)

Traditional Optional Stopping: we take $Z_{\langle j \rangle} = Z_j$ for all j.

Optional Stopping vs Continuation

- In many but certainly not all cases, we can also do optional stopping based on S-values.
- Suppose that $H_0 = \{P_0\}$, $S_j = \frac{\bar{p}_1(X_j | X^{j-1})}{p_0(X_j)}$, no Y_i 's
- Then $\forall P \in H_0 : \mathbf{E}_P[S_{j+1} \mid X^j] \leq 1$ and we can do optional stopping at each j and not just optional continuation between 'blocks'
- …but if H_0 composite then sometimes the only conditional S-value satisfying $\forall P \in H_0 : \mathbf{E}_P[S_{j+1} \mid X^j] \leq 1$ is given by the trivial S-value $S_{j+1} = 1$. Then OS impossible (but OC with batches of size $n_j \gg 1$ still possible)

(Again:) 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 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**

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

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 ...

Default S-Value $\neq$ Neyman

1. H_0 and H_1 are point hypotheses – then default S-value is:

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

... the safe test based on S looks a bit like, but is *not* a standard Neyman-Pearson test. **more conservative**

Safe Test: reject if $S(X^\tau) \geq 1/\alpha$

NP: reject if $S(X^\tau) \geq 1/B$ with B s.t. $P_0(S(X^\tau) \geq B) = \alpha$

SafeTests & Neyman-Pearson, again

- Let p be a strict p-value: for all $P \in H_0$, $P(p \leq \alpha) = \alpha$
- Let $S = \frac{1}{\alpha}$ if $p \leq \alpha$, and $S =$0 otherwise
- Then for all $P \in H_0$,

$$\mathbf{E}_P[S] = P(p \leq \alpha) \cdot \frac{1}{\alpha} + P(p > \alpha) \cdot 0 = 1$$

...so S is an S-value, and obviously, the safe test based on S rejects iff $p \leq \alpha$. It thus implements the Neyman-Pearson test at significance level α.

SafeTests & Neyman-Pearson, again

- Let p be a strict p-value: for all $P \in H_0$, $P(p \leq \alpha) = \alpha$
- Let $S = \frac{1}{\alpha}$ if $p \leq \alpha$, and S =0 otherwise
- Then for all $P \in H_0$,

$$\mathbf{E}_P[S] = P(p \leq \alpha) \cdot \frac{1}{\alpha} + P(p > \alpha) \cdot 0 = 1$$

...so S is an S-value, and obviously, the safe test based on S rejects iff $p \leq \alpha$. t thus implements the Neyman-Pearson test at significance level α.

...but it is a very silly S-value to use! With probability α, you loose all your capital, and you will never make up for that in the future!

Safe Tests and Neyman-Pearson, again

- **The Safe Test based on an S-Value that is a likelihood ratio is *not* a Neyman-Pearson test (it is more conservative)**
- **Neyman-Pearson tests (that only report ‘reject’ and ‘accept’, and not the p-value) are (other) Safe Tests, but useless ones corresponding to irresponsible gambling...**

Some S-Values are Better than Others

- The Trivial S-Value $S = 1$ is valid, but useless
- The Neyman-Pearson S-value is valid, but extremely dangerous to use!
- We need some idea of ‘optimal S-value’

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.

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

- We can then look for the GROW (growth-optimal in worst-case) S-value achieving

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

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

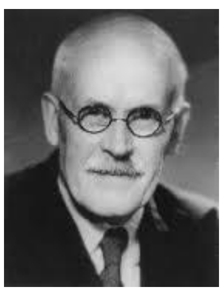
Main Theorem (will be made precise in 3 weeks)

- Under ‘hardly any conditions’ on H_0 and H_1 a GROW S-value exists! [G., De Heide, Koolen, 2019]

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?*

... Using S-Values we can unify/correct the central ideas

“Fisherian” Example

2. Ryabko & Monarev’s (2005)

Compression-based randomness test

R&M checked whether sequences generated by famous random number generators can be compressed by standard data compressors such as gzip and rar

Answer: yes! 200 bits compression for file of 10 megabytes

$$S(X^n) = 2^{\text{nr of bits compressed}} \quad (!!)$$

Additional Background Slides

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0\}$ represent the null hypothesis
  - For simplicity, today we assume data $X_1, X_2, \ldots$ 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 $\mu \neq 0$

  σ^2 unknown ('nuisance') parameter

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

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

Composite H_0

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 $\theta \in \Theta_0$,

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