Eclectic Lectures

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

We shall mostly study **nonnegative random variables *S*** satisfying:

$$\mathbf{E}_{S\sim P}[S] \leq 1$$

for all $P \in \mathcal{H}$:

$$\mathbf{E}_{S \sim P}[S] \leq 1$$

for all $f \in \mathcal{F}$:

$$\mathbf{E}_{S \sim P}[S_f] \leq 1$$

for all $f \in \mathcal{F}$:

$$\mathbf{E}_{S \sim P}[S_f] \leq 1$$

Invariably,
S **nonnegative**

Rough Plan of Lectures

1. Safe Testing (Statistics/AB Testing)
2. Safe Testing (Information Theory!)
3. Safe and Generalized Bayes
4. Fast Rate Conditions in Statistical (stochastic) and Online (nonstochastic) Learning
5. Safety and Luckiness – A Philosophy of Learning and Inference

First Lectures: Statistics, Testing

We will call a nonnegative random variable S satisfying

$$\text{for all } P \in H_0\text{:}$$

$$\mathbf{E}_{S \sim P}[S] \leq 1$$

an ***S*-value**. It is a better-behaved alternative to a p-value (large S roughly corresponding to small p)

From Stats to Information Theory

- Let H_0 be a set of prob distrs, and let Q be a prob distr
- **The reverse I-projection of Q onto H_0** is the prob. measure $\tilde{P}_0$ achieving

$$D(Q\|\tilde{P}_0) = \inf_{P \text{ in convex hull of } H_0} D(Q\|P)$$

- Theorem (Li, Barron 1999): $\tilde{P}_0$ generally exists, is unique, has density*, and satisfies, for all $P_0 \in H_0$,

$$\mathbf{E}_{Z\sim Q}\left(\frac{p_0(Z)}{\tilde{p}_0(Z)}\right) \leq 1$$

Generalized and Safe Bayes

- Let $\{ p_f : f \in \mathcal{F} \}$ be a set of probability densities and let π_0 be a prior density on $\mathcal{F}$
- The standard Bayesian posterior

$$\pi(f \mid Z^n) \propto \prod_{i=1}^{n} p_f(Z_i) \cdot \pi_0(f)$$

can behave very badly under misspecification, i.e. if the model is wrong but useful

- However, if we consider the tempered posterior

$$\pi(f \mid Z^n, \eta) \propto \prod_{i=1}^{n} p_f(Z_i)^{\eta} \cdot \pi_0(f)$$

for $\eta < \bar{\eta}$, then everything works just fine again.

Generalized and Safe Bayes

- If we consider the tempered posterior

$$\pi(f \mid Z^n, \eta) \propto \prod_{i=1}^{n} p_f(Z_i)^\eta \cdot \pi_0(f)$$

for $\eta < \bar{\eta}$, then everything works just fine, even under misspecification

Here $\bar{\eta}$ is the **critical** $\bar{\eta}$,defined as the largest $\bar{\eta} > 0$ satisfying, for all $f \in \mathcal{F}$

$$\mathbf{E}_{Z \sim P} \left(\frac{p_f(Z)}{p_{\tilde{f}}(Z)} \right)^{\bar{\eta}} \leq 1$$

with $\tilde{f}$ achieving $\min_{f \in \mathcal{F}} D(P \| P_f)$

Fast Rate Conditions in Statistical and Online Learning

- $\mathcal{F}$ set of predictors, $\ell_f \colon \mathcal{Z} \to \mathbb{R}$ loss function for f

.

- We say that $(P, \mathcal{F}, \ell_f)$ satisfies the **strong central condition** if for some $\eta > 0$, for all $f \in \mathcal{F}$,

$$\mathbf{E}_{Z \sim P}\left(e^{\eta(\ell_{f^*}(Z) - \ell_f(Z))}\right) \leq 1$$

- ...allows fast learning ($O\left(\frac{1}{n}\right)$ convergence rates)
- Generalizes existing conditions such as Bernstein's, exp-concavity, mixability

Rough Plan of Lectures

1. **Safe Testing (Statistics/AB Testing)**
2. Safe Testing (Information Theory!)
3. Safe and Generalized Bayes
4. Fast Rate Conditions in Statistical (stochastic) and Online (nonstochastic) Learning
5. Safety and Luckiness – A Philosophy of Learning and Inference

Part I: Safe Testing

- Classical Hypothesis Testing, A/B Testing

Partly based on joint work with Rianne de Heide, Wouter Koolen, Allard Hendriksen

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

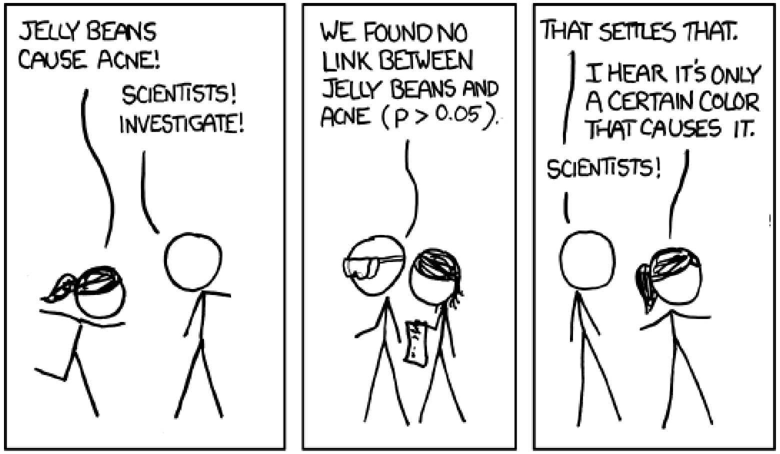
"at least 50% of highly cited results in medicine is irreproducible"
J. Ioannidis, PLoS Medicine 2005

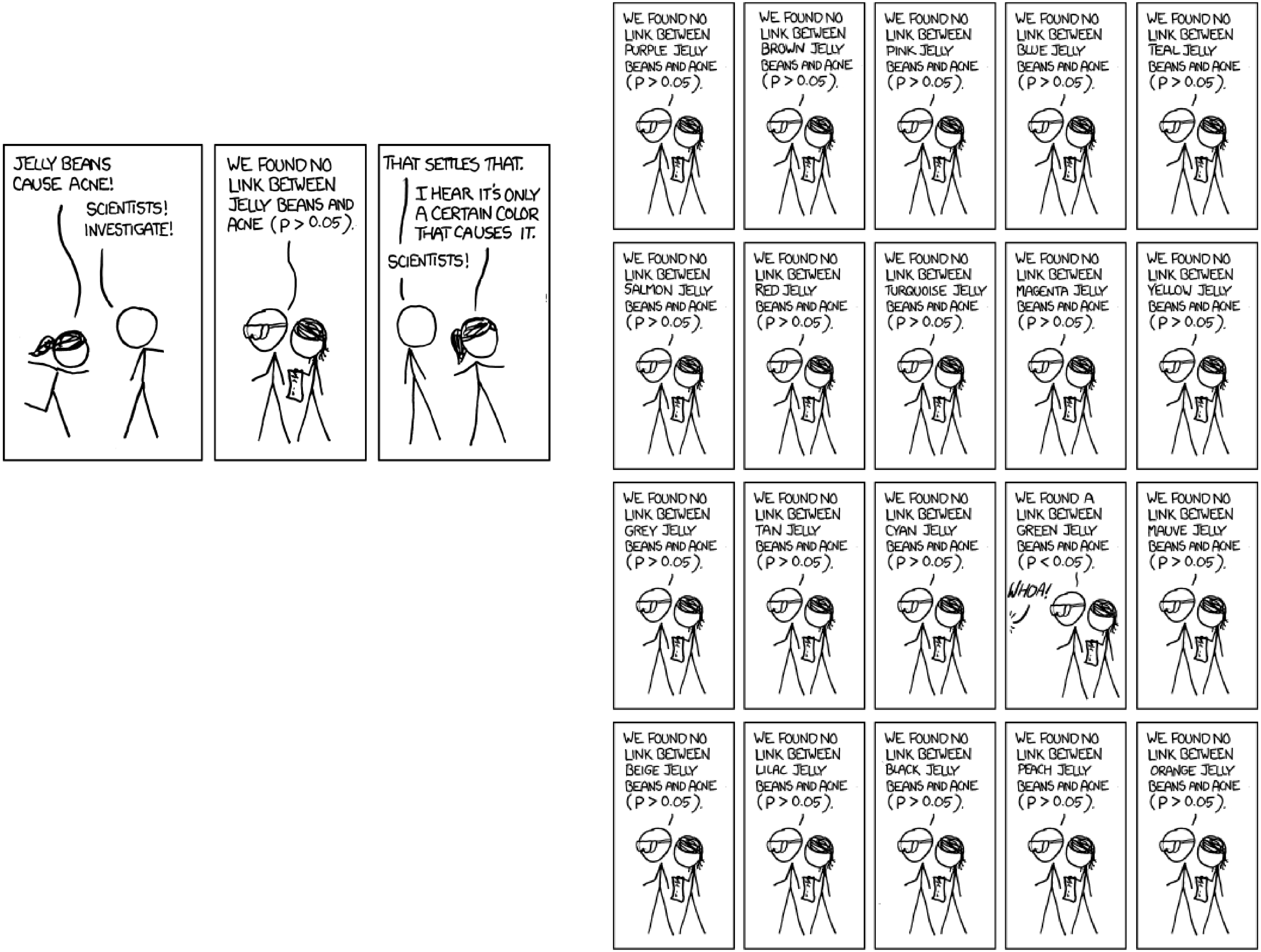
Reproducibility Crisis
Cover Story of Economist (2013), Wall Street Journal, Science (2012)

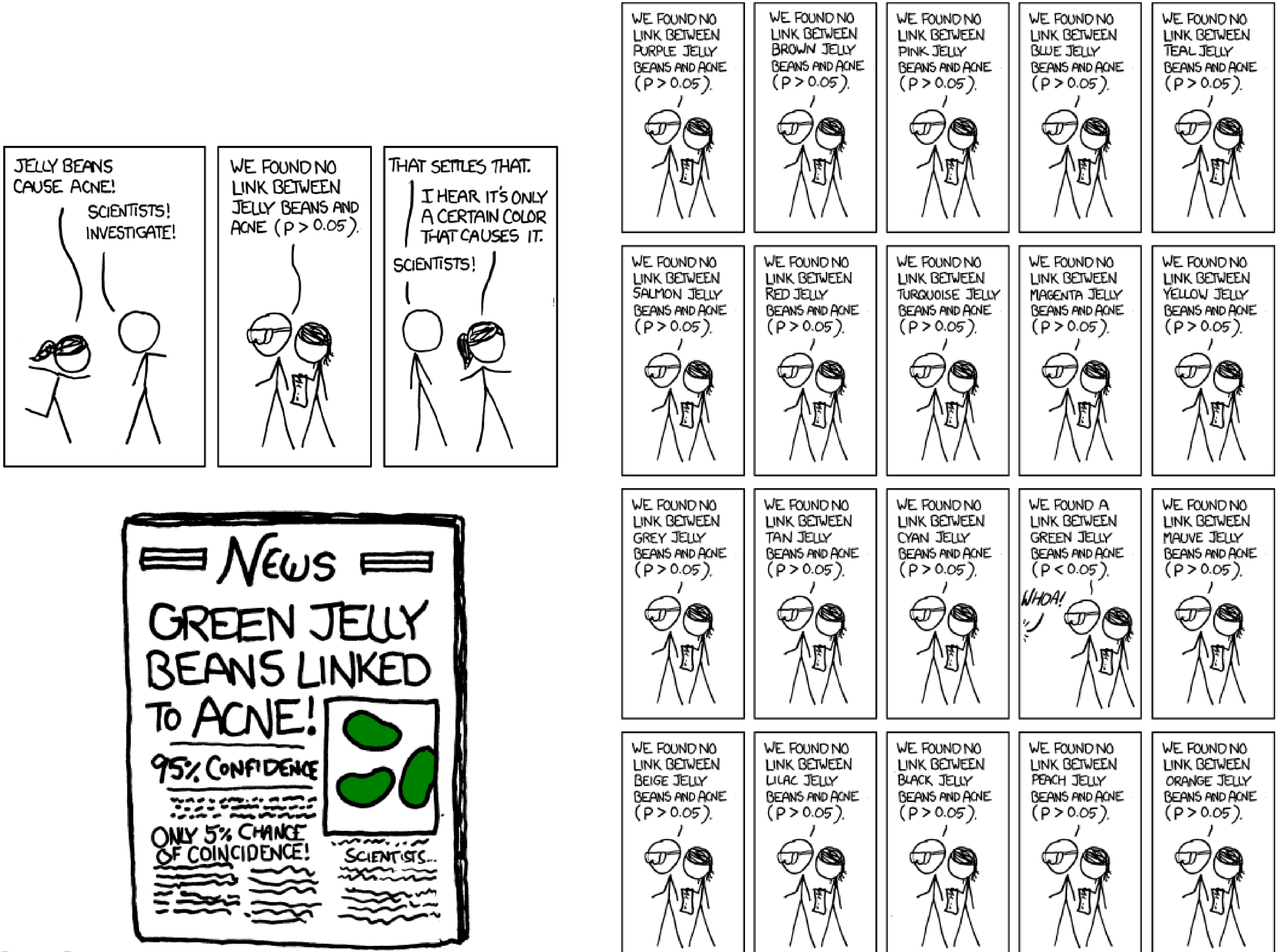
Reasons for Reproducibility Crisis

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

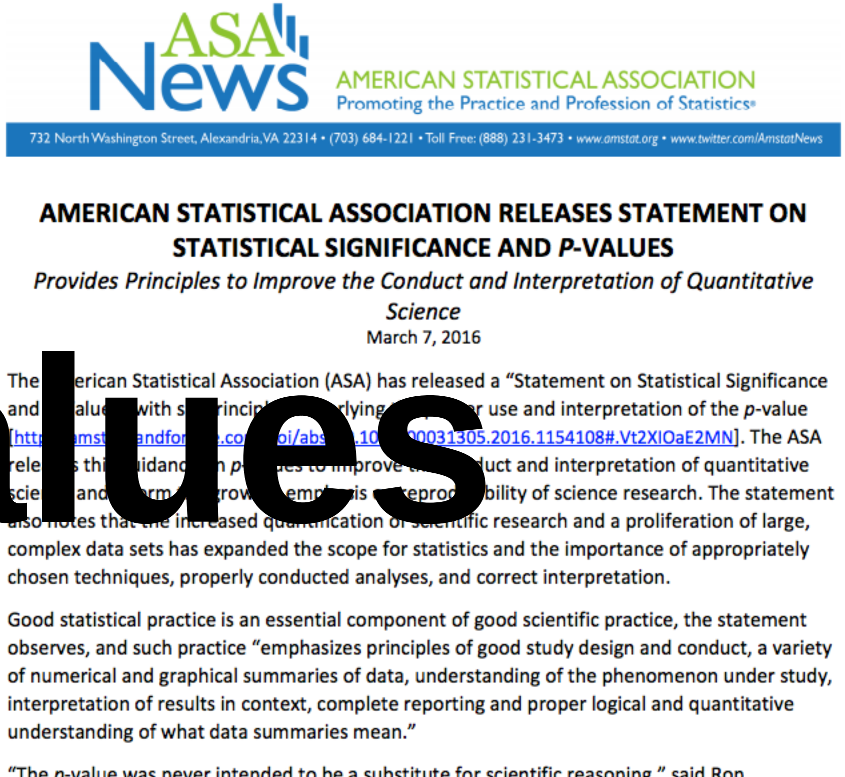
2. **Problems with Hypothesis Testing Methodology**

Reasons for Reproducibility Crisis

1. Publication Bias

2. **Problems with...**

p-values



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 [illegible]erican Statistical Association (ASA) has released a "Statement on Statistical Significance and [illegible]alue[illegible] with s[illegible]rincipl[illegible]rlying [illegible]r use and interpretation of the *p*-value [http[illegible]amst[illegible]andfor[illegible].co[illegible]oi/abs[illegible].10[illegible]00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA rele[illegible]s thi[illegible]uidanc[illegible]n *p*[illegible] to improve [illegible]duct and interpretation of quantitative scie[illegible] and [illegible]rm [illegible]grow[illegible] emph[illegible]is [illegible]eprod[illegible]bility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

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

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron

80 years and still unresolved...

- Standard method for testing is still

  p-value-based
  null hypothesis significance testing

  ...an amalgam of Neyman-Pearson’s and Fisher’s 1930s methods

  - everybody in psychology and medical sciences (and even in A/B testing) does it...
  - .... most statisticians agree it’s not o.k....
  - ...but still can’t agree on what to do instead!

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
  - For simplicity, assume $X_1, X_2, \ldots$ 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)

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

  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
  - For simplicity, 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: **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
  - For simplicity, 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: **t-test (most used test world-wide)**

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

  Composite H_0

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

  σ^2 unknown ('nuisance') parameter

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

P-value Problem #1: Combining Independent Tests

- Suppose two different research groups tested the same new medication. How to combine their test results?
- **You can't multiply p-values!**
  - **This will (wildly) overestimate evidence against the null hypothesis!**
  - Different valid p-value combination methods exist (Fisher's; Stouffer's) but give different results
- **In "our" method evidences can be safely multiplied**

P-value Problem #2: 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?
  - **Now Fisher’s and Stouffer’s method don’t work anymore – need complicated methods!**
- **In “our” method, despite dependence, evidences can still be safely multiplied**

P-value Problem #2b: Extending Your Test

I WILL NOT CUT CORNERS. I WILL NOT CUT CORNERS
NOT CUT CORNERS. I WILL NOT CUT CORNERS

- Suppose reseach group A tests medication, gets ‘almost significant’ result.
- **Sometimes group A can’t resist to test a few more subjects themselves...**
  - In a recent survey **55% of psychologists** admit to have succumbed to this practice [L. John et al., *Psychological Science*, 23(5), 2012]
- **In “our” method, despite dependence, evidences can still be safely multiplied**

P-value Problem #2b: 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 with p-values
2. S-Values and Safe Tests
   - ...solves the stop/continue problem
   - gambling interpretation
3. Safe Testing, simple (singleton) H_0
   - relation to Bayes
   - relation to MDL (data compression)
4. Safe Testing, Composite H_0
   - Magic: RIPr (Reverse Information Projection)
   - Examples: Safe t-Test, Safe Independence Test

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

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

S-Values: General Definition

I promised you!

- 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

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

General Definition

- An S-Value for **stopping time $\boldsymbol{\tau}$** is a fn S with nonnegative range such that for all $P_0 \in H_0$, we have

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

First Interpretation: p-values

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

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

- Proof: just Markov's inequality!

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

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 $S(X^\tau) \geq \frac{1}{\alpha}$
- Since for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

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

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

Second Interpretation: 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 $S(X^\tau) \geq \frac{1}{\alpha}$
- Since for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

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

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

First Examples

1. H_0 and H_1 are point hypotheses:

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

...is an S-value.

First Examples

1. H_0 and H_1 are point hypotheses:

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

...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...

First Examples: Safe ≠ Neyman

1. H_0 and H_1 are point hypotheses:

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

...note: one might think ‘the Neyman-Pearson paradigm tells us to use a LR ratio test here, and this is an LR ratio test, so safe testing is NP testing”

...but the safe test based on S is *not* a standard NP test.

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

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

First Examples: Safe ≠ Neyman

1. H_0 and H_1 are point hypotheses:

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

...note: one might think ‘the Neyman-Pearson paradigm tells us to use a LR ratio test here, and this is an LR ratio test, so safe testing is NP testing”

...but the safe test based on S is *not* a standard NP test.

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

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

First Examples

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

First Examples

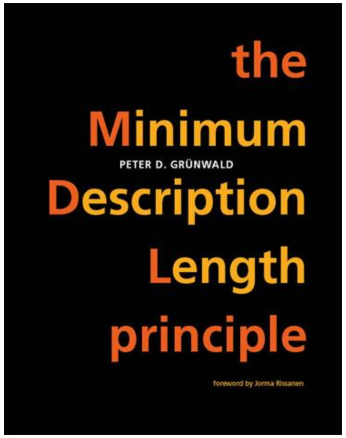
2. Ryabko & Monarev's (2005)

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

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



First Examples

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

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

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
- Let $S_1, S_2, \ldots, S_k$ be an arbitrarily large collection of (potentially "identical") S-values for sample sizes $n_1, n_2, \ldots, n_k$ respectively. Let $N_j := \sum_{i=1}^{j} n_i$
- We first evaluate S_1 on data $(X_1, \ldots, X_{n_1})$.
- 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 S_2 on data $\left(X_{n_1+1}, \ldots, X_{N_2}\right)$,

  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 on new batch of data; otherwise we **stop.**
- If S_2 is in certain range and Y_{N_2} has certain values then we perform S_3 , else we **stop**.
- ...and so on

(note that sequentially computed S-values may but need not have identical definitions, but data must be different for each test!)

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

Formally (and a bit more generally):

Let $g : \bigcup_{n \in \{n_1, n_2, \ldots\}} \mathcal{X}^n \times \mathcal{Y}^n \to \{\texttt{stop, continue}\}$ represent **arbitrary stop/continue strategy**, and:

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

else

Define $S := S_1(X^{n_1}) \cdot S_2(X^{N_2}_{n_1+1})$ **if** $g(X^{N_2}, Y^{N_2}) = \texttt{stop}$

else

Define $S := \prod_{j=1}^{3} S_j(X^{N_j}_{N_{j-1}+1})$ **if** $g(X^{N_3}, Y^{N_3}) = \texttt{stop}$

and so on...

Safe Tests are Safe

Theorem:

Let $g : \bigcup_{n \in \{n_1, n_2, \ldots, n_k\}} \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{\texttt{stop, continue}\}$ represent an **arbitrary stop/continue strategy**, and let the combined S be defined as before. Then :

If the $S_1, S_2, \ldots, S_k$ are S-values, then so is S !

Safe Tests are Safe

Theorem:

Let $g : \bigcup_{n \in \{n_1, n_2, \ldots, n_k\}} \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{\texttt{stop, continue}\}$ represent an **arbitrary stop/continue strategy**, and let the combined S be defined as before. Then :

If the $S_1, S_2, \ldots, S_k$ are S-values, then so is S !

- Can extend to:
  - choices between several tests at each time
  - tests that each have their own local stopping rule
  - Potentially infinite nr of tests (as long as stop/continue strategy stops eventually almost surely)
- Technically, the process $(S_1, S_1 \cdot S_2, \prod_{j=1}^{3} S_j, \ldots)$ is a **nonnegative supermartingale** (Ville '39)

Safe Tests are Safe

Theorem:

Let $g: \bigcup_{n \in \{n_1, n_2, \ldots, n_k\}} \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{\texttt{stop, continue}\}$ represent an **arbitrary stop/continue strategy**, and let the combined S be defined as before. Then :

If the $S_1, S_2, \ldots, S_k$ are S-values, then so is S !

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:

Let $g : \bigcup_{n \in \{n_1, n_2, \ldots, n_k\}} \mathcal{X}^n \times \mathcal{Y}^n \to \{\texttt{stop, conti}$ represent an **arbitrary stop/continue st** and let the combined S be defined as befor .

If the $S_1, S_2, \ldots, S_k$ are S-valu n so is S !

Corollary: Type-I Erro antee Preserved under Optional Contin

Suppose w ne S-values with arbitrary stop/c strategy and reject H_0 when final S has $S^{-1} \leq$ 05 . Then resulting test is a safe test and our Type-I Error 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**
- 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!

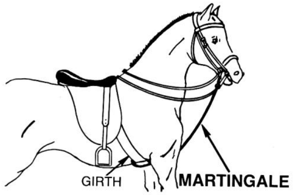
- You start by investing 1$ in ticket 1.
- After n_1 outcomes you either stop with end capital S_1 or you continue and buy S_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $S_1 \cdot S_2$ or you continue and buy $S_1 \cdot S_2$ tickets of type 3, and so on...
- ***S***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

Safe Testing = Gambling!

- Your don’t expect to gain money with S since none of individual gambles S_k are strictly favorable to you
- Hence a **large value of $\boldsymbol{S}$** indicates that something has happened that is higly unlikely under H_0 ...
- **“Amount of evidence against $\boldsymbol{H_0}$” is thus measured in terms of how much money you gain in a game that would allow you not to make money in the long run if $\boldsymbol{H_0}$ were true!**

Safe Testing = Gambling!

- Your don’t expect to gain money with S since none of individual gambles S_k are strictly favorable to you
- Hence a **large value of $\boldsymbol{S}$** indicates that something has happened that is higly unlikely under H_0 ...
- **“Amount of evidence against $\boldsymbol{H_0}$” is thus measured in terms of how much money you gain in a game that would allow you not to make money in the long run if $\boldsymbol{H_0}$ were true!**



relation to martingales will be considered later!

SafeTests & Neyman-Pearson, again

- Let p be a 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.

SafeTests & Neyman-Pearson, again

- Let p be a 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...**

Menu

1. Some of the problems with p-values
2. Safe Testing with S-values
   - ...solves the optional continuation problem
   - gambling interpretation
   - Neyman-Pearson tests are useless safe-tests...
3. Safe Testing, simple (singleton) H_0
   - relation to Bayes
4. Safe Testing, Composite H_0
   - Magic: RIPr (Reverse Information Projection)
   - Examples: Safe t-Test, Safe Independence Test

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

$$\frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$$

where

$$\bar{p}(X_1, \ldots, X_n \mid H_1) := \int_{\theta \in \Theta_1} p_\theta(X_1, \ldots, X_n) w_1(\theta) d\theta$$

$$\bar{p}(X_1, \ldots, X_n \mid H_0) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) w_0(\theta) d\theta$$

Safe Testing and Bayes, simple $\boldsymbol{H_0}$

Bayes factor hypothesis testing

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

Evidence measured by

$$\frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$$

where

$$\bar{p}(X_1, \ldots, X_n \mid H_1) := \int_{\theta \in \Theta_1} p_\theta(X_1, \ldots, X_n) w_1(\theta) d\theta$$

$$\bar{p}(X_1, \ldots, X_n \mid H_0) := p_0(X_1, \ldots, X_n)$$

Safe Testing and Bayes, simple H_0

Bayes factor hypothesis testing

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

Take $S(X^n) := \dfrac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{p_0(X_1, \ldots, X_n)}$

and note that (no matter what prior w_1 we chose)

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

$$\int p_0(x^n) \cdot \frac{\bar{p}(x^n \mid H_1)}{p_0(x^n)} dx^n = \int \bar{p}(x^n \mid H_1) dx^n = 1$$

Safe Testing and Bayes, simple H_0

Bayes factor hypothesis testing

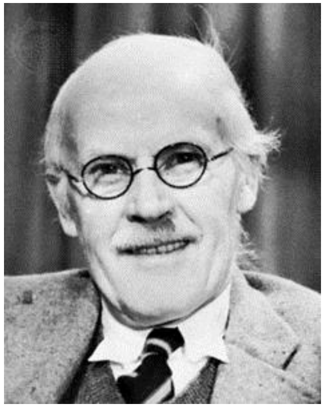
between $H_0 = \{ p_0 \}$ and $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

Take $$S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{p_0(X_1, \ldots, X_n)}$$

and note that (no matter what prior w_1 we chose)

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

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



Menu

1. Some of the problems with p-values
2. Safe Testing
3. Safe Testing, simple (singleton) H_0
   - relation to Bayes
4. **Safe Testing, Composite $\boldsymbol{H_0}$**
   - **Magic**: RIPr (Reverse Information Projection)
   - Allows for a general construction of Safe Tests
   - Examples: Safe t-test, Safe independence test

Composite H_0: Bayes may not be Safe!

Bayes factor given by $S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$

where $\bar{p}(X_1, \ldots, X_n \mid H_0) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) w_0(\theta) d\theta$

Composite H_0: Bayes may not be Safe!

Bayes factor given by $S(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{\bar{p}(X_1, \ldots, X_n \mid H_0)}$

where $\bar{p}(X_1, \ldots, X_n \mid H_0) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) w_0(\theta) d\theta$

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

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

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

$$\mathbf{E}_{X^n \sim \bar{P}(\cdot | H_0)} [S(X^n)] \leq 1$$

Composite H_0:
Bayes can be unsafe!

- ...for Bayes factor we can in general only guarantee

$$\mathbf{E}_{X^n \sim \bar{P}(\cdot|H_0)}\left[S(X^n)\right] \leq 1$$

- In general Bayesian tests with composite H_0 are not safe ...which means that they loose their Type-I error guarantee interpretation when we combine (in)dependent Bayes factors
- (and they lack several other nice properties as well)

Composite $\boldsymbol{H_0}$:
Bayes can be unsafe!

- ...for Bayes factor we can in general only guarantee

$$\mathbf{E}_{X^n \sim \bar{P}(\cdot|H_0)}\left[S(X^n)\right] \leq 1$$

- Bayesian tests with composite H_0 ***are*** safe if you really believe your prior on H_0
- I usually don't believe my prior, so no good for me!

Composite H_0: Bayes can be unsafe!

- ...for Bayes factor we can in general only guarantee

$$\mathbf{E}_{X^n \sim \bar{P}(\cdot|H_0)} \left[S(X^n) \right] \leq 1$$

- Bayesian tests with composite H_0 ***are*** safe if you really believe your prior on H_0
- I usually don't believe my prior, so no good for me!

Bayesian statisticians often claim

Optional Stopping: No Problem for Bayesians (Rouder, '14)

...but that only works if you believe your prior – viz.

Why Optional Stopping is a Problem for Bayesians

(G. & De Heide, '18)

Composite H_0:
Bayes can be unsafe!

- ...for Bayes factor we can in general only guarantees

$$\mathbf{E}_{X^n \sim \bar{P}(\cdot|H_0)} [S(X^n)] \leq 1$$

- In general Bayesian factors with composite H_0 are not S-values
- ...but there do exist *very special priors* **(in general dependent on $\bar{P}(\cdot\,|H_1)$, and highly unlike the priors that people tend to use!)** for which Bayes factors become S-values
- I will now show you how to construct such priors!

Reverse Information Projection

- Let $\bar{H}_0$ be a convex set of prob distrs, and let Q be a prob distr, such that Q and all $P \in \bar{H}_0$ have densities relative to the same underlying measure.
  The reverse I-projection of Q onto P_0 is the prob. measure $\tilde{P}_0$ achieving

$$D(Q\|\tilde{P}_0) = \inf_{P \in \bar{H}_0} D(Q\|P)$$

Reverse Information Projection

- Let $\bar{H}_0$ be convex set of prob distrs, and let Q be a prob distr, such that Q and all $P \in \bar{H}_0$ have densities relative to the same underlying measure. **The reverse I-projection of $\boldsymbol{Q}$ onto $\boldsymbol{\bar{H}_0}$** is the prob. measure $\tilde{P}_0 \in \bar{H}_0$ achieving

$$D(Q\|\tilde{P}_0) = \inf_{P \in \bar{H}_0} D(Q\|P)$$

Here $D(Q\|P) = \mathbf{E}_{Z\sim Q}\left[\log \frac{q(Z)}{p(Z)}\right]$

is Kullback-Leibler divergence between P and Q

Reverse Information Projection

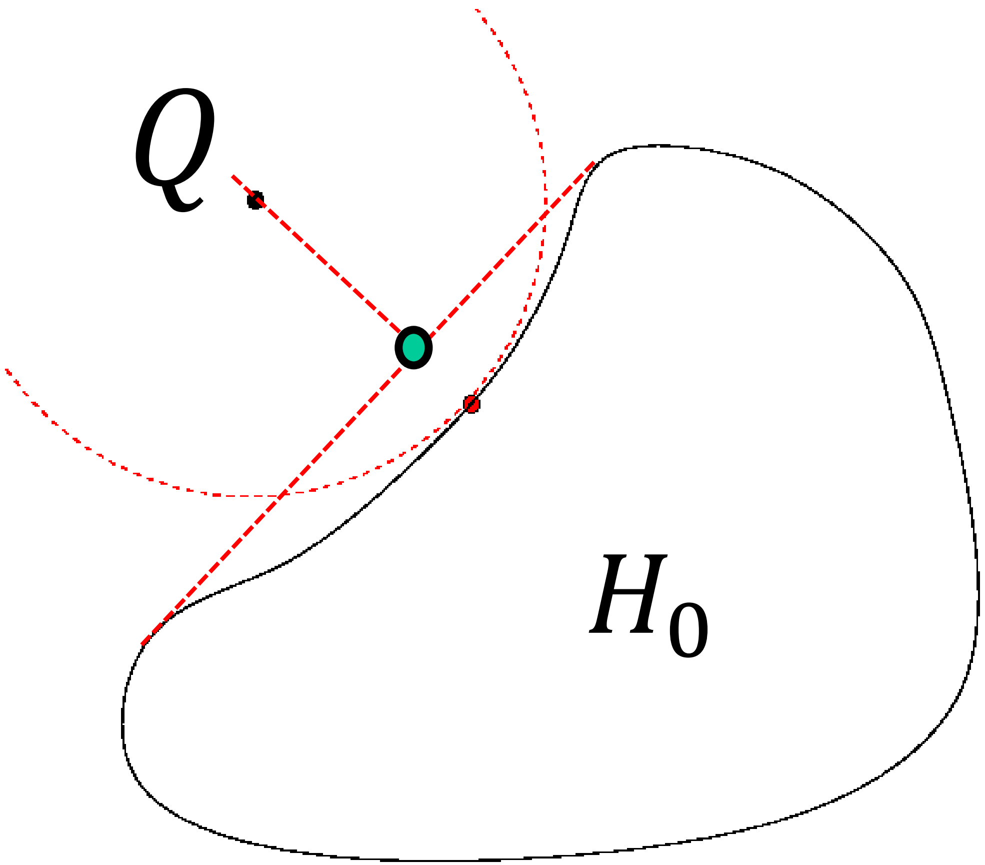
- Let $\bar{H}_0$ be convex set of prob distrs, and let Q be a prob distr, such that Q and all $P \in \bar{H}_0$ have densities relative to the same underlying measure.
  The reverse I-projection of Q onto $\bar{H}_0$ is the prob. Measure $\tilde{P}_0 \in \bar{H}_0$ achieving

$$D(Q \| \tilde{P}_0) = \inf_{P \in \bar{H}_0} D(Q \| P)$$

- Theorem (Li, Barron 1999): $\tilde{P}_0$ generally exists, is unique, has density*, and satisfies, for all $P_0 \in H_0$,

$$\mathbf{E}_{Z \sim Q} \left(\frac{p_0(Z)}{\tilde{p}_0(Z)} \right) \leq 1$$

Reverse Information Projection



● is $\tilde{P}_0$

Proof (Easy but Crucial Part)

- **Suppose I-projection of Q onto $\overline{H}_0$** exists, i.e. there is a prob. measure $\tilde{P}_0 \in \overline{H}_0$ achieving

$$D(Q\|\tilde{P}_0) = \inf_{P \in \overline{H}_0} D(Q\|P)$$

- Let $P_0 \in \overline{H}_0$ with density p_0. Calculate

$$\frac{d}{d\alpha} D(Q\|(1-\alpha)\tilde{P}_0 + \alpha P_0) = \frac{d}{d\alpha} \mathbf{E}_{X^n \sim Q}\left[-\log(1-\alpha)\tilde{p}_0 + \alpha p_0)\right]$$

Proof (Easy but Crucial Part)

- **Suppose I-projection of $\boldsymbol{Q}$ onto $\boldsymbol{\overline{H}_0}$** exists, i.e. there is a prob. measure $\tilde{P}_0 \in \overline{H}_0$ achieving

$$D(Q\|\tilde{P}_0) = \inf_{P \in \overline{H}_0} D(Q\|P)$$

- Let $P_0 \in \overline{H}_0$ with density p_0. Calculate

$$\frac{d^2}{d\alpha^2} D(Q\|(1-\alpha)\tilde{P}_0+\alpha P_0) = \frac{d^2}{d\alpha^2} \mathbf{E}_{X^n \sim Q}\left[-\log(1-\alpha)\tilde{p}_0 + \alpha p_0)\right]$$

Proof (Easy but Crucial Part)

- **Suppose I-projection of $\boldsymbol{Q}$ onto $\boldsymbol{\bar{H}_0}$** exists, i.e. there is a prob. measure $\tilde{P}_0 \in \bar{H}_0$ achieving

$$D(Q\|\tilde{P}_0) = \inf_{P \in \bar{H}_0} D(Q\|P)$$

- Let $P_0 \in \bar{H}_0$ with density p_0. Calculate

$$\frac{d^2}{d\alpha^2} D(Q\|(1-\alpha)\tilde{P}_0 + \alpha P_0) = \frac{d^2}{d\alpha^2} \mathbf{E}_{X^n \sim Q}\left[-\log(1-\alpha)\tilde{p}_0 + \alpha p_0)\right]$$

- This is > 0 at all $0 \leq \alpha \leq 1$ so fn is convex
- Since $(1-\alpha)\tilde{P}_0 + \alpha\, P_0 \in \bar{H}_0$, first derivative must be ≥ 0 at $\alpha = 0$

Proof (Easy but Crucial Part)

- **Suppose I-projection of $\boldsymbol{Q}$ onto $\boldsymbol{\bar{H}_0}$** exists, i.e. there is a prob. measure $\tilde{P}_0 \in \bar{H}_0$ achieving

$$D(Q\|\tilde{P}_0) = \inf_{P \in \bar{H}_0} D(Q\|P)$$

- Let $P_0 \in \bar{H}_0$ with density p_0. First dervative

$$\frac{d}{d\alpha} D(Q\|(1-\alpha)\tilde{P}_0 + \alpha P_0) = \frac{d}{d\alpha} \mathbf{E}_{X^n \sim Q} [- \log(1-\alpha)\tilde{p}_0 + \alpha p_0)]$$

at $\alpha = 0$ is given by

$$1 - \mathbf{E}_{Z \sim Q} \left(\frac{p_0(Z)}{\tilde{p}_0(Z)} \right)$$

Towards Main Result

- Associate composite H_1 with single “representing” distribution $\bar{P}_1$ restricted to n outcomes, e.g.

$$\bar{p}_1(x^n) = \int_{\theta \in \Theta_1} p_\theta(x^n) dW(\theta)$$

for some prior W over Θ_1

- Let $\bar{H}_0$ be set of Bayes marginals over H_0, i.e. all distributions with densities of form

$$p(x^n) = \int_{\theta \in \Theta_0} p_\theta(x^n) dW(\theta)$$

... for some distribution W on Θ_0. Note $\bar{H}_0$ is convex!

Reverse Information Projection

- Let $\bar{H}_0$ be convex set of prob distrs, and let Q be a prob distr, such that Q and all $P \in \bar{H}_0$ have densities relative to the same underlying measure.
  The reverse I-projection of Q onto $\bar{H}_0$ is the prob. Measure $\tilde{P}_0 \in \bar{H}_0$ achieving

$$D(Q \| \tilde{P}_0) = \inf_{P \in \bar{H}_0} D(Q \| P)$$

- Theorem (Li, Barron 1999): $\tilde{P}_0$ generally exists, is unique, has density*, and satisfies, for all $P_0 \in H_0$,

$$\mathbf{E}_{Z \sim Q} \left(\frac{p_0(Z)}{\tilde{p}_0(Z)} \right) \leq 1$$

Towards Main Result

- Associate composite H_1 with single “representing” distribution $\bar{P}_1$ restricted to n outcomes
- **For now** we will be Bayesian about H_1 (but not H_0) and assume that we can come up with a prior W on Θ_1 such that we can simply set

$$\bar{p}_1(x^n) = \int_{\theta \in \Theta_1} p_\theta(x^n) dW(\theta)$$

Towards Main Result

- Associate composite H_1 with single “representing” distribution $\bar{P}_1$ restricted to n outcomes
- Let $\overline{H}_0$ be set of Bayes marginals over H_0, i.e. all distributions with densities of form

$$p(x^n) = \int_{\theta \in \Theta_0} p_\theta(x^n) dW(\theta)$$

... for some distribution W on Θ_0. Note $\overline{H}_0$ is convex!

Hence by Barron-Li result, there exists* $\tilde{P}_0 \in \overline{H}_0$ with for all $P_0 \in \overline{H}_0$,

$$\mathbf{E}_{X^n \sim \bar{P}_1} \left(\frac{p_0(X^n)}{\tilde{p}_0(X^n)} \right) \leq 1$$

Towards Main Result

- Associate composite H_1 with single “representing” distribution $\bar{P}_1$ restricted to n outcomes

- By Barron-Li result: there exists* distribution $\tilde{P}_0$ with density $$\tilde{p}_0(x^n) := \int_{\theta \in \Theta_0} p_\theta(x^n) dW(\theta)$$

- i.e. a Bayes marginal, such that for all $P_0 \in H_0$,

$$\mathbf{E}_{X^n \sim \bar{P}_1} \left(\frac{p_0(X^n)}{\tilde{p}_0(X^n)} \right) \leq 1$$

Towards Main Result

- Associate composite H_1 with single “representing” distribution $\bar{P}_1$ restricted to n outcomes

- By Barron-Li result: there exists* distribution $\tilde{P}_0$ with density $$\tilde{p}_0(x^n) := \int_{\theta \in \Theta_0} p_\theta(x^n) dW(\theta)$$

- i.e. a Bayes marginal, such that for all $P_0 \in H_0$,

$$\mathbf{E}_{X^n \sim \bar{P}_1} \left(\frac{p_0(X^n)}{\tilde{p}_0(X^n)} \right) \leq 1$$

or equivalently (!!!):

$$\boxed{\mathbf{E}_{X^n \sim P_0} \left(\frac{\bar{p}_1(X^n)}{\tilde{p}_0(X^n)} \right) \leq 1}$$

First Main Result : A General Method for S-Value construction with Composite $\boldsymbol{H_0}$

- This shows that reverse I-projection $\tilde{P}_0$ of $\bar{P}_1$ onto composite $\bar{H}_0$ defines an S-value $S^* = \frac{\bar{p}_1}{\tilde{p}_0}$
- Moreover, among all S-values S against H_0 this S^* is **optimal** in the sense that it maximizes the $\bar{P}_1$- **expected capital growth rate**

$$\mathbf{E}_{X^n \sim \bar{P}_1} (\log S(X^n))$$

- This works for **completely arbitrary $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$**

Example 1:
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

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

- In general Bayes factor tests are *not* safe
- But lo and behold, Jeffreys' uses very special priors and his Bayesian t-test is a Safe Test!
  - ...but not the "frequentist best" (**highest power/captital growth**) safe test!

Example 1:
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 factor tests are *not* safe
- 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 1: 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

Jeffreys uses improper right-Haar prior $w(\sigma) = 1/\sigma$ within both models, and uses Cauchy on μ/σ

$$\bar{p}(X^n \mid H_0) := \int_{\sigma > 0} w(\sigma) p_\sigma(X^n) d\sigma = \int \frac{1}{(\sqrt{2\pi}\sigma)^n} \cdot \frac{1}{\sigma} \cdot \text{exp}\left(- \frac{\sum X_i^2}{2\sigma^2} \right) d\sigma$$

$$S := \frac{\bar{p}(X^n \mid H_1)}{\bar{p}(X^n \mid H_0)}$$

- With this choice S has same distribution under all $P \in H_0$, and $\mathbf{E}_{X^n \sim P}(S) = 1$

Example 1:
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

Jeffreys uses improper right-Haar prior $w(\sigma) = \ 1/\sigma$ within both models, and uses Cauchy on μ/σ

In fact, for improper right-Haar prior combined with every 0-symmetric prior on effect size μ/σ we get that

S has same distribution under all $P \in H_0$, and

$$\mathbf{E}_{X^n \sim P}(S) = 1$$

Nuisance Parameters with Group Structure

- In many practical problems, only free parameter in H_0 is nuisance parameter (vector) (like σ in scale families such as in t-test, or (μ, σ) in location-scale families) such that
  - nuisance parameter also part of H_1
  - nuisance parameter/distributions satisfy appropriate group structure
- Berger et al. ‘98, Dass & Berger, ‘03 give many examples

Nuisance Parameters with Group Structure

- In many practical problems, only free parameter in H_0 is nuisance parameter (vector) (like σ in scale families such as in t-test, or (μ, σ) in location-scale families) such that
  - nuisance parameter also part of H_1
  - nuisance parameter/distributions satisfy appropriate group structure
  - In all such cases, the Bayes factor based on the improper right Haar prior is also an S-value!
- But what if the ‘nuisance’ parameter has no group structure?

Example 2: Independence Testing

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

- H_0: $X_1, X_2, \dots, X_n \mid Z_1, \dots, Z_n$ iid Bernoulli(θ),
- H_1: $X_1, X_2, \dots, X_n$ iid Bernoulli(θ) , 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?**

- **...can calculate RIPr numerically, encouraging results**

How to design S-Values?

- The RIPr gives us an S-value for every given $\bar{P}_1$ representing H_1.
- If we want to be Bayesian about H_1 can pick

$$\bar{p}_1(x^n) = \int_{\theta \in \Theta_1} p_\theta(x^n) dW(\theta)$$

....and we're done

- (as Berger et al. (2016) argue, many frequentists are in fact secretly Bayesian about H_1)

How to design S-Values?

- The RIPr gives us an S-value for every given $\bar{P}_1$ representing H_1.
- If we want to be Bayesian about H_1 can pick

$$\bar{p}_1(x^n) = \int_{\theta \in \Theta_1} p_\theta(x^n) dW(\theta)$$

....and we're done

- (as Berger et al. (2016) argue, many frequentists are in fact secretly Bayesian about H_1)
- **...but what if we don't know how to pick prior W_1 on Θ_1?**

How to design S-Values?

- The RIPr gives us an S-value for every given $\bar{P}_1$ representing H_1...but what if we don't know how to pick $\bar{P}_1$, prior W_1 on Θ_1?
- ...suppose we are willing to admit that we'll only be able to tell H_0 and H_1 apart if $P \in H_0 \cup H_{1,\delta}$ for some $H_{1,\delta} \subset H_1$ that excludes points that are 'too close' to H_0 (e.g. $H_1 = \left\{ P_\theta : \middle| | \theta - \theta_0 ||_2 \geq \frac{c}{\sqrt{n}} \right\}$)
- We can then look for GROW (growth-optimal in worst-case) S-value achieving

$$\sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

The GROW S-Value

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

$$\sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

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

The GROW S-Value and the JIPr

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

$$\sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

- **Second Main Theorem:** under conditions on H_0, $H_{1,\delta}$:

$$\inf_{P \in \bar{H}_{1,\delta}} \inf_{Q \in \bar{H}_0} D(P \| Q) = \sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

...and $S^* = p^* \, / \, \lfloor p^* \rfloor_{H_0}$ where $(p^*, \lfloor p^* \rfloor_{H_0})$ achieves the minimum on the left and $\lfloor p^* \rfloor_{H_0}$is the RIPr for p^*

The GROW S-Value and the JIPr

JIPr = Joint Information Projection

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

$$\sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

- **Second Main Theorem:** under conditions on H_0, $H_{1,\delta}$:

$$\inf_{P \in \bar{H}_{1,\delta}} \inf_{Q \in \bar{H}_0} D(P \| Q) = \sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

...and $S^* = p^* / \lfloor p^* \rfloor_{H_0}$ where $(p^*, \lfloor p^* \rfloor_{H_0})$ achieves the minimum on the left and $\lfloor p^* \rfloor_{H_0}$ is the RIPr for p^*

Crucial Idea for Proof

- For any fixed $\bar{P}_1$,

$$\max_{S: S\text{-val rel. to } H_0} \mathbf{E}_{X^n \sim \bar{P}_1} [\log S]$$

...given by $S = \bar{p}_1 / \lfloor \bar{p}_1 \rfloor_{H_0}$ where $\lfloor \bar{p}_1 \rfloor_{H_0}$ is RIPr of $\bar{p}_1$

(this is surprising because the $\bar{p}_1$ inside logarithm is not fixed here!)

- Hence

$$\min_{p:\text{density}} \mathbf{E}_{X^n \sim \bar{P}_1} \left[-\log \frac{p(X^n)}{\lfloor p \rfloor_{H_0}(x^n)} \right]$$

...is achieved for $p = \bar{p}_1$

The GROW S-Value and the JIPr

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

$$\sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

- **Second Main Theorem:** under conditions on H_0 and $H_{1,\delta}$ we have:

$$\inf_{P \in \bar{H}_{1,\delta}} \inf_{Q \in \bar{H}_0} D(P \| Q) = \sup_{S} \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

...and $S^* = p^*$ / $\lfloor p^* \rfloor_{H_0}$ where $(p^*, \lfloor p^* \rfloor_{H_0})$ achieves the minimum on the left and $\lfloor p^* \rfloor_{H_0}$is the RIPr for p^*

The GROW S-Value and the JIPr

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

$$\sup_S \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

- **Second Main Theorem:** under conditions on H_0, $H_{1,\delta}$:

$$\inf_{P \in \bar{H}_{1,\delta}} \inf_{Q \in \bar{H}_0} D(P \| Q) = \sup_S \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

...and $S^* = p^*$ / $\lfloor p^* \rfloor_{H_0}$ where $(p^*, \lfloor p^* \rfloor_{H_0})$ achieves the minimum on the left and $\lfloor p^* \rfloor_{H_0}$ is the RIPr for p^*

GROW Safe T-Test:

- Jeffreys sets $\bar{p}(X^n \mid H_1) := \int_{\sigma > 0} w(\sigma) w(\mu \mid \sigma) p_{\mu,\sigma}(X^n) d\mu d\sigma$

- where $p_{\mu,\sigma}$ is density of n i.i.d. N(μ, σ) RVs and **$w(\mu \mid \sigma)$ is a standard Cauchy with scale σ**
- Instead we want to pick the GROW S-value under the constraint that $|\mu/\sigma| \geq \delta_0$ for some 'minimally clinically relevant effect size'
- It turns out that this S-value is given by the Bayes factor with the right Haar prior and a 2-point prior on μ/σ with probability ½ on δ_0 and ½ on - δ_0

GROW Safe T-Test:

- Jeffreys sets $\bar{p}(X^n \mid H_1) := \int_{\sigma>0} w(\sigma)w$ [illegible] $\sigma)p_{\mu,\sigma}(X^n)d\mu d\sigma$

- where $p_{\mu,\sigma}$ is density of n [illegible] $\sigma)$ RVs and **$w(\mu \mid \sigma)$ is a standar** [illegible] **with scale σ**

- Instead we want t [illegible] GROW S-value under the constraint that [illegible] for some 'minimally clinically re [illegible] ct size'

- It turns [illegible] is S-value is given by the Bayes factor w [illegible] right Haar prior and a 2-point prior on μ/σ with probability ½ on δ_0 and ½ on - δ_0

Everything fits!

Type II Error for Simple H_0

- Neyman-Pearson null hypothesis testing rejects H_0 at 5% level whenever (asymptotically)

$$\|\widehat{\theta}_n - \theta_0\| \geq 1.96 \cdot \sqrt{\frac{\mathrm{var}(P_{\theta_0})}{n}} \asymp \sqrt{\frac{1}{n}}$$

Optimal Power

Not Safe, Not Consistent

- Bayes with standard prior rejects H_0 whenever

$$\|\widehat{\theta}_n - \theta_0\| \gtrsim \sqrt{\frac{\log n}{n}}$$

SubOptimal Power

Safe, Consistent

- Bayes with JIPr-prior chosen so as to maximize power rejects H_0 at 5% whenever

$$\|\widehat{\theta}_n - \theta_0\| \geq 2.45 \cdot \sqrt{\frac{\mathrm{var}(P_{\theta_0})}{n}} \asymp \sqrt{\frac{1}{n}}$$

Close to Optimal Power

Safe, Not Consistent

What about power?

- Fixed n at small sample sizes: need about 30% more data to achieve same power as with classical Neyman-Pearson test
- But: for subclass of safe tests, we are allowed to do **optional stopping** (stronger requirement than optional continuation, which is always possible)
  - possible for t-test, but not for independence test
- ...with optional stopping sometimes need less data than with classical approach!

Menu

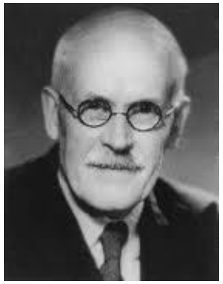
1. Some of the problems with p-values
2. Safe Testing
3. Safe Testing, simple (singleton) H_0
   - relation to Bayes
4. Safe Testing, Composite H_0
   - **Magic**: RIPr (Reverse Information Projection)
   - JIPR (Joint Information Projection) Allows for a general construction of Safe Tests
   - Examples: Safe t-test, Safe independence test
5. **Historical Perspective**

Some Historical Perspective

The Three Classical Approaches to Testing

Jerzy Neyman (1930s): alternative exists, “inductive behaviour”, p-value vs ‘significance level’

Sir Ronald Fisher (1920s): test statistic rather than alternative, p-value indicates “unlikeliness”



Sir Harold Jeffreys (1930s): **Bayesian**, alternative exists, absolutely no p-values

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

Sir Ronald's view on testing

Sir Ronald Fisher: a statistical test should just report a "p-value". This is a measure of evidence that indicates "unlikeliness" ; no explicit alternative H_1 needs to be formulated

- "Goodness-of-Fit, Randomness Test"

Safe Tests comply: they can be formulated without clear alternatives (think of Ryabko-Monarev GZIP-test for randomness). But the p-value gets replaced by the more robust S-value!

Neyman’s View on Testing

- *Before* experiment is done, state *significance level* α (e.g. $\alpha = 0.05$)
- **Reject** H_0 iff $\mathrm{p} < 0.05$
- This gives **Type-I Error** Guarantee of α
- If statisticians would follow this procedure for fixed α in all their experiments, the fraction of times in which the null hypothesis would be true but they would reject, would be at most α
- alternative H_1 is crucial: among all p-values, pick one maximizing power (minimizing Type-II error)
- ...actual p-value is of lesser (no!?!?) concern!

A Big Issue with Testing as currently practiced / p-values

- The standard way of doing null hypothesis testing is an amalgam of Fisher’s and Neyman’s ideas
- We reject if $p \leq \alpha$ but we do report p, and claim that we have ‘a lot more evidence’ if $p \ll \alpha$
- But how to interpret an observation like $p < 0.01$ when we a priori set $\alpha = 0.05$?

A Big Issue with Testing as currently practiced / p-values

- The standard way of doing null hypothesis testing is an amalgam of Fisher's and Neyman's ideas
- We reject if $p \leq \alpha$ but we do report p, and claim that we have 'a lot more evidence' if $p \ll \alpha$
- But how to interpret an observation like $p < 0.01$ when we a priori set $\alpha = 0.05$?

"in those cases where we observe $p < 0.01$, we will only make a Type I error (false reject) 1% of the time"

NO! We might make a Type I error in fact in 100% of the time in those cases!

A Big Issue with Testing as currently practiced / p-values

- How to interpret an observation like $p < 0.01$ when we a priori set $\alpha = 0.05$?
- Perhaps Wald’s reinterpretation of NP tests in terms of loss functions can come to the rescue?

Neyman-Pearson Decision Theory

$\delta : X^n \to \{a_0, a_1\}$ decision rule

In a Classical Null-Hypothesis test we fix some α and set:

$$\delta(X^n) := \begin{cases} a_1 : \text{reject!} & \text{if p-val}(X^n) \leq \alpha \\ a_0 : \text{accept!} & \text{otherwise} \end{cases}$$

In terms of Loss Functions:

$L(i, a_j)$:

Loss you make when H_i is the case, yet a_j is what you decide

Now decision rule better interpreted as:

$$\delta(X^n) = \begin{cases} a_0 : \text{“do nothing”} \\ a_1 : \text{“do something!”} \end{cases}$$

In terms of Loss Functions:

- For simplicity assume $L(0, a_0) = L(1, a_1) = 0$

Frequentist Type-I Error Guarantee:

$$P_0(\delta(X^n) = a_1) \leq \alpha$$

where

$$\delta(X^n) := \begin{cases} a_1 & \text{if p-val}(X) \leq \alpha \\ a_0 & \text{otherwise} \end{cases}$$

In terms of Loss Functions:

For simplicity assume $L(\theta_0, a_0) = L(\theta_1, a_1) = 0$

Frequentist Type-I Error Guarantee:

$$P_0(\delta(X^n) = a_1) \leq \alpha$$

In terms of Loss Functions:

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq 1$$

as long as $L(0, a_1) \leq \frac{1}{\alpha}$

In terms of Loss Functions:

For simplicity assume $L(\theta_0, a_0) = L(\theta_1, a_1) = 0$

Frequentist Type-I Error Guarantee:

$$P_0(\delta(X^n) = a_1) \leq \alpha \quad = 0.05$$

In terms of Loss Functions:

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq 1$$

as long as $L(0, a_1) \leq \dfrac{1}{\alpha} \quad = 20$

What if there are more than 2 actions?

$$\delta(X) = \begin{cases} a_0 : \text{“do nothing”} \\ a_1 : \text{“do a second, more expensive investigation”} \\ a_2 : \text{“start an expensive anti-meat eating campaign”'} \\ a_3 : \text{“ban meat right away”} \end{cases}$$

$$L(0, a_0) = 0$$
$$L(0, a_1) = 10$$
$$L(0, a_2) = 100$$
$$L(0, a_3) = 1000$$

We want procedure that guarantees:

$$E_{X^n \sim P_0}[L(\theta_0, \delta(X^n))] \leq \text{bound } (\text{say, } 1)$$

Just 2 actions:

$L(0, a_0) = 0$

$L(0, a_2) = 100$

We want procedure that guarantees:

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq 1$$

We achieve this by setting

$$\delta(X^n) := \begin{cases} a_2 & \text{if p-val}(X) \leq \frac{1}{100} \\ a_0 & \text{otherwise} \end{cases}$$

3 actions:

$$L(0, a_0) = 0 \quad L(0, a_1) = 10 \quad L(0, a_2) = 100$$

We want procedure that guarantees:

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq 1$$

It seems we achieve this by setting:

$$\delta(X^n) := \begin{cases} a_2 & \text{if p-val} \leq \frac{1}{100} \\ a_1 & \text{if } \frac{1}{100} < \text{ p-val } \leq \frac{1}{10} \\ a_0 & \text{otherwise} \end{cases}$$

3 actions:

$$L(0, a_0) = 0 \quad L(0, a_1) = 10 \quad L(0, a_2) = 100$$

We want procedure that guarantees:

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq 1$$

It seems we achieve this by setting:

$$\xcancel{\delta(X^n) := \begin{cases} a_2 & \text{if p-val} \leq \frac{1}{100} \\ a_1 & \text{if } \frac{1}{100} < \text{p-val} \leq \frac{1}{10} \\ a_0 & \text{otherwise} \end{cases}}$$

doesn’t work!

3 actions:

$$L(0, a_0) = 0 \quad L(0, a_1) = 10 \quad L(0, a_2) = 100$$

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] =$$

$$\frac{1}{100} \cdot 100 + \left(\frac{1}{10} - \frac{1}{100}\right) \cdot 10 = 2 - \frac{1}{10} \approx 2$$

It seems we achieve this by setting:

$$\xcancel{\delta(X^n) := \begin{cases} a_2 & \text{if p-val} \leq \frac{1}{100} \\ a_1 & \text{if } \frac{1}{100} < \text{p-val} \leq \frac{1}{10} \\ a_0 & \text{otherwise} \end{cases}}$$

doesn’t work!

Many actions:

$$L(0, a_k) = 10^k \text{ for } k = 0 \ldots k_{\mathsf{max}}$$

We want procedure that guarantees:

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq \text{const.}$$

But “natural” decision rule based on p-value gives

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \approx k_{\mathsf{max}} \to \infty$$

Many actions:

$$L(0, a_k) = 10^k \text{ for } k = 0 \ldots k_{\mathsf{max}}$$

We want procedure that guarantees:

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq \text{const.}$$

But “natural” decision rule based on p-value gives

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \approx k_{\mathsf{max}} \rightarrow \infty$$

Yet “natural” decision rule based on S-value does give

$$E_{X^n \sim P_0}[L(0, \delta(X^n))] \leq 1$$

3 actions:

$$L(0, a_0) = 0 \quad L(0, a_1) = 10 \quad L(0, a_2) = 100$$

$$\mathbf{E}_{X^n \sim P_0}[L(0, \delta(X^n))] =$$

$$= \mathbf{E}\left[\mathbf{1}_{S \geq 100} \cdot 100 + \mathbf{1}_{10 \leq S < 100} \cdot 10 + \mathbf{1}_{S < 10} \cdot 0\right] \leq \mathbf{E}[S] \leq 1$$

Everything works fine if we set:

$$\delta(X^n) := \begin{cases} a_2 & \text{if } S^{-1} \leq \frac{1}{100} \\ a_1 & \text{if } \frac{1}{100} < S^{-1} \leq \frac{1}{10} \\ a_0 & \text{otherwise} \end{cases}$$

3 actions:

$$L(0, a_0) = 0 \quad L(0, a_1) = 10 \quad L(0, a_2) = 100$$

$$\mathbf{E}_{X^n \sim P_0}[L(0, \delta(X^n))] =$$

$$= \mathbf{E}\left[\mathbf{1}_{S \geq 100} \cdot 100 + \mathbf{1}_{10 \leq S < 100} \cdot 10 + \mathbf{1}_{S < 10} \cdot 0\right] \leq \mathbf{E}[S] \leq 1$$

Everything works fine if we set:

$$\delta(X^n) := \begin{cases} a_2 & \text{if } S^{-1} \leq \frac{1}{100} \\ a_1 & \text{if } \frac{1}{100} < S^{-1} \leq \frac{1}{10} \\ a_0 & \text{otherwise} \end{cases}$$

(works also with countably ∞ many actions)

A Big Issue with Testing as currently practiced / p-values

- The standard way of doing null hypothesis testing is an amalgam of Fisher’s and Neyman’s ideas
- We reject if $p \leq \alpha$ but we do report p, and claim that we have ‘a lot more evidence’ if $p \ll \alpha$
- But how to interpret an observation like $p < 0.01$ when we a priori set $\alpha = 0.05$?

...I claim: interpretation with p-values is terribly unclear!

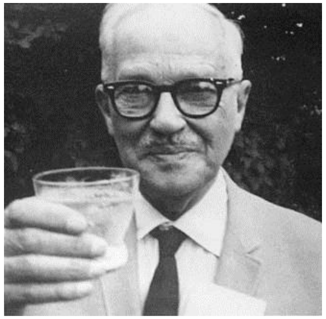
S-values resolve this issue!

Neyman’s View on Testing

- *Before* experiment is done, state *significance level* α (e.g. $\alpha = 0.05$*)*
- **Reject** H_0 iff $\mathrm{p} < 0.05$
- This gives **Type-I Error** Guarantee of α
- If statisticians would follow this procedure for fixed α in all their experiments, the fraction of times in which the null hypothesis would be true but they would reject, would be at most α
- alternative H_1 is crucial: among all p-values, pick one maximizing power (minimizing Type-II error)
- ...actual p-value is of lesser (no!?!?) concern!

Neyman and Fisher together

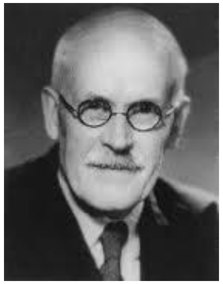
- To some extent, S-values *do* allow us to combine the features of Fisherian and Neymanian testing!
- S-value measures ‘unlikeliness’, even without alternative, just like p-value
- ...but behaves much better under optional continuation
- S-value leads to Type-I error/loss guarantees, even under optional continuation, and even if there are more than 2 actions



The Three Classical Approaches to Testing

Jerzy Neyman (1930s): alternative exists, “inductive behaviour”, p-value vs ‘significance level’

Sir Ronald Fisher (1920s): test statistic rather than alternative, p-value indicates “unlikeliness”



Sir Harold Jeffreys (1930s): **Bayesian**, alternative exists, absolutely no p-values

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

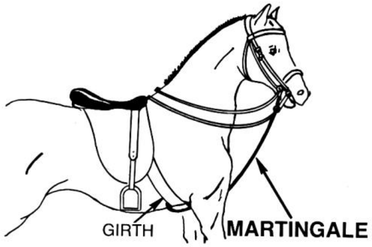
Earlier Work on S-Values

- The simple H_0 case (and related developments) was essentially covered in work by Volodya **Vovk** and collaborators (1993, 2001, 2011,...)
  - see esp. Shafer, Shen, Vereshchagin, Vovk: Test Martingales, Bayes Factors and p-values, 2011

- Also Jim **Berger** and collaborators have earlier ideas in this direction (1994, 2001, ...)

- In particular Berger was inspired by the great Jack **Kiefer**

- What is really radically new here is interpretation & the general treatment of **composite H_0 and its relation to reverse/joint-information projection**



Vovk's Work on S-Values

- S-Value is natural weakening of the concept of a **test martingale** (more about this next lecture)
- Test martingales go back to Ville (1939), in the paper that introduced the modern concept of a martingale
- In fabulous 2011 paper, Shafer, Vovk et al. compare test martingales, p-values and S-values
  - Very confusingly, they call S-values 'Bayes factors' (this is because they focus on simple H_0)
- A lot more on S-values vs p-values in forthcoming book by Vovk and Shafer on game-theoretic probability

Conclusion First Part

Safe Testing has a frequentist (type-I error) interpretation. Advantages over Standard frequentist testing:

1. Combining (in)dependent tests, adding extra data
2. More than two decisions: not just “accept/reject”

Bayes tests with very special priors are SafeTests. Advantages over Standard Bayes priors/tests:

1. **Combining (in)dependent tests, adding extra data**
2. **Possible to do pure ‘randomness test’ (no clear alternative available)**

Safe Testing has a frequentist (type-I error) interpretation. Advantages over Standard frequentist testing:

1. Combining (in)dependent tests, adding extra data
2. More than two decisions: not just “accept/reject”

Bayes tests with very special priors are SafeTests, even in composite case. Advantages over Standard Bayes priors/tests:

1. **Combining (in)dependent tests, adding extra data**
2. **Possible to do pure ‘randomness test’ (no clear alternative available)**

All Safe Tests have a gambling and MDL (data compression) interpretation

(with again, advantages over standard MDL tests)

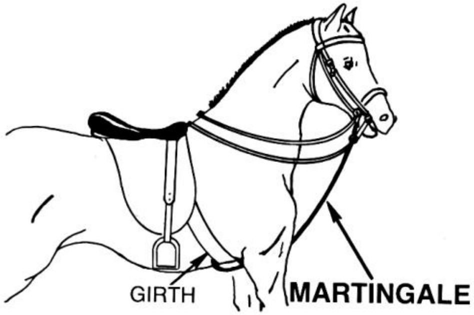
Additional Material

Read more?

safe tests!

- G. Shafer, A. Shen, N. K. Vereshchagin, and V. Vovk. Test martingales, Bayes Factors and p-values. *Statistical Science,* 2011
- G. **Safe Probability**. *Journal of Stat. Planning and Inference*, 2018
- Reversed I-Projection: G. & Mehta, **Fast Rates for Unbounded Losses: from ERM to Generalized Bayes,** *arXiv*, 2017
- G., De Heide, Koolen. ***Safe Testing****.* In preparation.

•*More to come...*



Safe Testing and...

- **"Amount of evidence against H_0" is thus measured in terms of how much money you gain in a game that would allow you not to make money in the long run if H_0 were true**
- $\approx$ **Nonnegative supermartingales** introduced by Ville (1939) and Vovk's (1993) Test Martingales

every test martingale defines an S-value, but not vice versa!

Undiscovered Gems

- Jonathan Li’s (1999) Ph.D. Thesis supervised by Andrew Barron – establishes basic properties of reverse information projection, shows that they generally exist*
- Shafer, Shen, Vovk, Vereshchagin (2011)
- Shafer & Vovk (2001, 2018): Probability and Finance, it’s only a game!