Safe Testing

CWI

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam
Mathematisch Instituut Universiteit Leiden

Universiteit Leiden

Partly based on joint work with
Stéphanie van der Pas, Rianne de Heide,
Wouter Koolen, Allard Hendriksen

Slate **Sep 10th: 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)

---

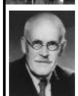
80 years and still unresolved...

- Standard method 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 does it...
  - .... most statisticians agree it's not o.k....
  - ...but still can't agree on what to do instead!

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

Jerzy Neyman: alternative exists, "inductive . . behaviour"

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



Sir Harold Jeffreys: **Bayesian**, alternative exists, inductive behaviour; compression interpretation

---

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
- **We will present a method in which 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

- 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!?**

Safe (i.e. adaptive) Testing

- **We aim for a 'safe' or adaptive method that better suits the real-life research world where obviously either you yourself or another research group wants to, and will, study more data given preliminary test results that are promising but inconclusive!**

Should we be Bayesian?

- These and several other problems with p-values attracted a lot of attention in the 1960s and...
- ...caused several people to become Bayesian
  - and right now there's a Bayesian revolution in psychology...
- As we will see though, **Bayesian methods don't fully resolve** the issues at hand
- We propose a new method that does: Safe Testing

Should we be Bayesian?

- These and several other problems with p-values attracted a lot of attention in the 1960s and...
- ...caused several people to become Bayesian
  - and right now there's a Bayesian revolution in psychology...
- As we will see though, **Bayesian methods don't fully resolve** the issues at hand
- We propose a new method: Safe Testing
  - for **simple H_0**, all Bayes factor tests are also Safe Tests
  - for **composite H_0**, Bayes factor tests are usually *not* safe (**T-Test, independence testing**)

Earlier Work

- 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, ...)

- Both Berger and Vovk inspired by the great Jack **Kiefer**

- The only thing that is really radically new here is the treatment of **composite H_0 and its relation to reverse-information projection**

Menu

1. Some of the problems with p-values
2. Safe Testing
   - ...solves the adaptivity 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

Menu

1. Some of the problems with p-values
2. Safe Testing
   - ...solves the adaptivity 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

Null Hypothesis Testing

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

  $\Theta_0 = \left\{\frac{1}{2}\right\}, \Theta_1 = [0,1] \setminus \left\{\frac{1}{2}\right\}$ 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 $X_1, X_2, \ldots$ 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

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

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

Null Hypothesis Testing

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - For simplicity, 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: **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

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

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

Safe Test: 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

- A **test** is a function $M : \bigcup_{n \geq 0} \mathcal{X}^n \rightarrow \mathbb{R}_0^+$
- A **safe test** for sample size n is a test such that for **all** $P_0 \in H_0$, we have

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

General Definition

- Let T be a positive-integer valued random variable
- A safe test for **stopping time** T is a test such that for all $P_0 \in H_0$, we have

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

First Interpretation: p-values

- Proposition: Let M be a safe test. Then $M^{-1}(X^T)$ is a nonstrict p-value, i.e. a p-value with **wiggle room**:
- for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{M(X^T)} \leq \alpha\right) \leq \alpha$$

First Interpretation: p-values

- Proposition: Let M be a safe test. Then $M^{-1}(X^T)$ is a nonstrict p-value, i.e. a p-value with **wiggle room**:
- for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{M(X^T)} \leq \alpha\right) \leq \alpha$$

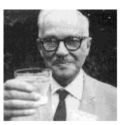
- Proof: just Markov's inequality!

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

First Interpretation: p-values

- Proposition: Let M be a safe test. Then $M^{-1}(X^T)$ is a nonstrict p-value, i.e. a p-value with **wiggle room**:
- for all $P \in H_0$, all $0 \leq \alpha \leq 1$,

$$P\left(\frac{1}{M(X^T)} \leq \alpha\right) \leq \alpha$$



- Hence if we reject H_0 iff $M^{-1}(X^T) < 0.05$, then we have **Type-I Error** Bound of 0.05

Safe Tests are Safe ('Adaptive')

- Suppose we observe data $(X_1, Y_1), (X_2, Y_2), \ldots$
  - Y_i: side information, independent of X_i's
- Let $M_1, M_2, \ldots, M_k$ be an arbitrarily large collection of (potentially identical) safe tests for sample sizes $n_1, n_2, \ldots, n_k$ respectively.
- Suppose we first perform test M_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 perform test M_2 ; otherwise we stop.

Safe Tests are Safe (‘Adaptive’)

- We first perform test M_1.
- If outcome is in certain range and Y_{n_1} has certain values then we perform test M_2 ; otherwise we stop.
- If outcome of test M_2 is in certain range and $Y_{n_1+n_2}$ has certain values then we perform M_3 ,else we stop.
- ...and so on

(note that sequentially performed tests may but need not be identical, but data must be different for each test!)

Safe Tests are Safe (‘Adaptive’)

- We first perform test M_1.
- If outcome is in certain range and Y_{n_1} has certain values then we perform test M_2 ; otherwise we stop.
- If outcome of test M_2 is in certain range and $Y_{n_1+n_2}$ has certain values then we perform M_3 ,else we stop.
- ...and so on

Main Result, Informally: any Meta-Test composed of Safe Tests in this manner is itself a safe test, irrespective of the stop/continue rule used!

Safe Tests are Safe

Formally (and a bit more generally):

Let $S : \bigcup_{n>0} \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{\texttt{stop}, \texttt{continue}\}$ represent an **arbitrary stop/continue strategy**, and:

Define $M := M_1(X^{n_1})$ and $T := n_1$ if $S(X^{n_1}, Y^{n_1}) = \texttt{stop}$

else

Define $M := M_1(X^{n_1}) \cdot M_2(X^{N_2}_{n_1+1})$ and $T := N_2$ if $S(X^{N_2}, Y^{N_2}) = \texttt{sto}$

else

Define $M := \prod_{j=1}^{3} M_j(X^{N_j}_{N_{j-1}+1})$ and $T := N_3$ if $S(X^{N_3}, Y^{N_3}) = \texttt{stop}$

and so on...

Safe Tests are Safe

Theorem:

Let $S : \bigcup_{n>0} \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{\texttt{stop}, \texttt{continue}\}$ represent an **arbitrary stop/continue strategy**, and let the combined test M with stopping time T be defined as before. Then :

If the $M_1, M_2, \ldots, M_k$ are safe tests, then so is M !

Safe Tests are Safe

Theorem:

Let $S : \bigcup_{n>0} \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{\texttt{stop}, \texttt{continue}\}$ represent an **arbitrary stop/continue strategy**, and let the combined test M with stopping time T be defined as before. Then :

If the $M_1, M_2, \ldots, M_k$ are safe tests, then so is M !

Corollary:

Suppose we combine safe tests with arbitrary stop strategy and reject H_0 whenever $M^{-1} \leq 0.05$. Then our Type-I Error is guaranteed to be below 0.05!

We solved the main problem with p-values!

Second, Main Interpretation:
Gambling!

Safe Testing = Gambling!

- At each time n there are k tickets for sale, all for 1$.
  - Ticket j pays off $M_j(X_n, \ldots, X_{n+n_j})$ $ after n_j steps.
  - 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 M_1 or you continue and buy M_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $M_1 \cdot M_2$ or you continue and buy $M_1 \cdot M_2$ 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 M_1 tickets of type 2. After $N_2 = n_1 + n_2$ outcomes you stop with end capital $M_1 \cdot M_2$ or you continue and buy $M_1 \cdot M_2$ tickets of type 3, and so on...
- **M is simply your end capital**
- Your expected gain for arbitrary M is at most 0, since none of the individual gambles M_k are strictly favorable to you
- Hence a **large value of M** indicates that something very unlikely has happened under H_0 ...

Safe Testing = Gambling!

- Your expected gain for arbitrary M is at most 0, since none of the individual gambles M_k are strictly favorable to you
- Hence a **large value of M** indicates that something has happened that is higly unlikely under H_0 ...
- **"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 many in the long run if H_0 were true!**

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 many in the long run if H_0 were true**
- ≈ Minibatch-wise- **Kelly Gambling**
- Also related to but different from **Wald's Sequential Testing Paradigm** (Balsubramani & Ramdas 2015)
- ≈ **Nonnegative supermartingales** introduced by Ville (1939) and Vovk's (1993) Test Martingales

every test martingale defines a safe test, but not vice versa!

Menu

1. Some of the problems with p-values
2. Safe Testing
   - ...solves the adaptivity 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

Safe Testing and Bayes

- **Bayes factor hypothesis testing** (Jeffreys '39)

with $H_0 = \{ p_\theta | \theta \in \Theta_0 \}$ vs $H_1 = \{ p_\theta | \theta \in \Theta_1 \}$:

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

where
$$\bar{p}(X_1, \ldots, X_n \mid H_1) := \int_{\theta \in \Theta_1} p_\theta(X_1, \ldots, X_n) w_1(\theta) d\theta$$
$$\bar{p}(X_1, \ldots, X_n \mid H_0) := \int_{\theta \in \Theta_0} p_\theta(X_1, \ldots, X_n) w_0(\theta) d\theta$$

Then "posterior probability of H_0" is $< 1/(K+1)$

Safe Testing and Bayes, simple H_0

- **Bayes factor hypothesis testing**

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

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

  where

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

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

Safe Testing and Bayes, simple H_0

- **Bayes factor hypothesis testing**

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

  Pick H_1 if $M(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{p_0(X_1, \ldots, X_n)} > K$

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

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

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

Safe Testing and Bayes, simple H_0

- **Bayes factor hypothesis testing**

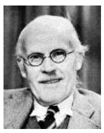
  between $H_0 = \{p_0\}$ and $H_1 = \{p_\theta | \theta \in \Theta\}$:

  Pick H_1 if $M(X^n) := \frac{\bar{p}(X_1, \ldots, X_n \mid H_1)}{p_0(X_1, \ldots, X_n)} > K$

  but note that

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

The Bayes Factor for Simple H_0 is a Safe Test!



Safe Test vs. Bayes Factor vs. MDL

Every Simple vs Composite Bayes Factor Hypothesis Test corresponds to a Safe Test

But not vice versa!

- sometimes 'non-Bayesian' definition of $\bar{p}(\cdot|H_1)$ is preferable ⟶ **MDL**
  - Normalized Maximum Likelihood/Sharkov distribution (Rissanen '96)
  - Prequential Plug-In Distribution (Dawid '84)
  - Switch Distribution (Van Erven et al., NIPS 2007)

Type II Error for Simple H_0

- Asymptotically, standard null hypothesis testing rejects H_0 whenever

$\|\hat{\theta}_n - \theta_0\| \gtrsim \sqrt{\frac{1}{n}}$

- Optimal Power
- Not Safe, Not Consistent

Type II Error for Simple H_0

- Asymptotically, standard null hypothesis testing rejects H_0 whenever

$\|\hat{\theta}_n - \theta_0\| \gtrsim \sqrt{\frac{1}{n}}$

- Optimal Power
- Not Safe, Not Consistent

- Bayes rejects H_0 whenever

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

- SubOptimal Power
- Safe, Consistent

Law of the Iterated Logarithm! VdPas, G. 2016

Type II Error for Simple H_0

- Asymptotically, standard null hypothesis testing rejects H_0 whenever

$$\|\hat{\theta}_n - \theta_0\| \gtrsim \sqrt{\frac{1}{n}}$$

- Optimal Power
- Not Safe, Not Consistent

- Bayes rejects H_0 whenever

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

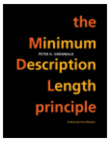
- SubOptimal Power
- Safe, Consistent

- Setting $\bar{P}(\cdot|H_1)$ to be **switch distribution** rejects H_0 whenever

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

- Almost Optimal Power
- Safe, Consistent

MDL Testing/Model Selection



MDL: Pick H_1 if $\frac{\bar{p}_1(x_1,\ldots,x_n)}{\bar{p}_0(x_1,\ldots,x_n)} > K$

where $\bar{p}_0$ and $\bar{p}_1$ are 'universal' distributions ("codes") relative to H_0 viz. H_1

= Single distributions (codes) that represent a whole set thereof

- For simple H_0, Safe Tests are essentially equivalent* to MDL Tests

Menu

1. Some of the problems with p-values
2. Safe Testing
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)
   - Allows for a general construction of Safe Tests
   - Examples: Safe t-Test, Safe Independence Test

Composite H_0: Bayes may not be Safe!

- Bayes picks H_1 if $M(X^n) := \frac{\bar{p}(X_1,\ldots,X_n \mid H_1)}{\bar{p}(X_1,\ldots,X_n \mid H_0)} > K$

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

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

Composite H_0: Bayes may not be Safe!

- Bayes picks H_1 if $M(X^n) := \frac{\bar{p}(X_1,\ldots,X_n \mid H_1)}{\bar{p}(X_1,\ldots,X_n \mid H_0)} > K$

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

Safe test requires that **for all** $P_0 \in H_0$:

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

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

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

Composite H_0: Bayes can be unsafe!

- ...for a Bayes test we can in general only guarantees

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

- 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 test

(and they lack several other nice properties as well)

Composite H_0: Bayes can be unsafe!

- ...for a Bayes test we can in general only guarantees

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

- 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 a Bayes test we can in general only guarantees

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

- In general Bayesian tests with composite H_0 are not safe
- ...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 tests become truly safe
- I will now show you how to make such priors!

RIPr: Reverse Information Projection

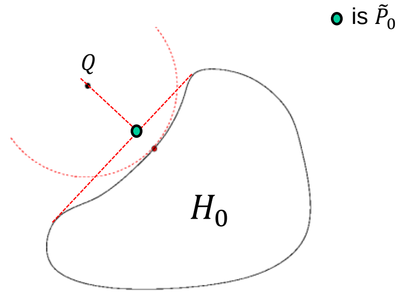
- For arbitrary sets H_0 of distributions on Z, and arbitrary distribution Q on Z, **the reverse I-projection of Q onto H_0** is defined as the density $\tilde{p}_0$ of the distribution achieving

$$\min_{P \text{ in convex hull of } H_0} D(Q\|P)$$

- Theorem (Li, Barron 1999): $\tilde{p}_0$ generally exists, is unique 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



Towards Main Result

- Associate H_1 with representing distribution $\bar{P}_1$ restricted to n outcomes, with density $\bar{p}_1(x^n)$
- By Barron-Li result: there exist a distribution $\tilde{P}_0$ of form

$$\tilde{p}_0(x^n) := \int_{\theta \in \Theta_0} p_\theta(x^n) dW(\theta)$$

i.e. a Bayes mixture, 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 H_1 with representing distribution $\bar{P}_1$ restricted to n outcomes, with density $\bar{p}_1(x^n)$
- By Barron-Li result: there exist a distribution $\tilde{P}_0$ of form

$$\tilde{p}_0(x^n) := \int_{\theta \in \Theta_0} p_\theta(x^n) dW(\theta)$$

i.e. a Bayes mixture, 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 (!!!):

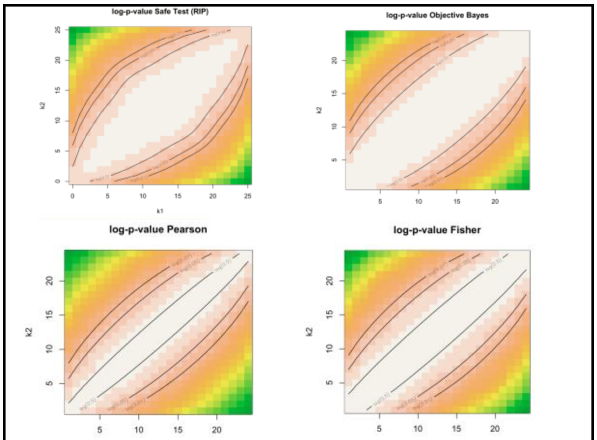
$$\mathbf{E}_{X^n \sim P_0} \left(\frac{\bar{p}_1(X^n)}{\tilde{p}_0(X^n)} \right) \leq 1$$

Main Result : A General Method for Safe Test construction with Composite H_0

- This shows that the reverse I-projection $\tilde{p}_0$ of $\bar{p}_1$ onto composite H_0 defines a safe test $\frac{\bar{p}_1}{\tilde{p}_0}$
- This works for completely arbitrary H_0 and H_1
  - May e.g. be nonparametric...
  - But practical implementation may be complicated...
  - For two of the most important (and simple) examples it works out fine though...

Example 1: Independence Testing

- $X_i \in \{0,1\}$; $Z_i \in \{1,2\}$
- H_0: $X_1, X_2, \ldots, X_n \mid Z_1, \ldots, Z_n$ i.i.d. Bernoulli(θ),
- H_1: $X_1, X_2, \ldots, X_n \mid Z_1, \ldots, Z_n$ independent, but $P(X_i = 1 \mid Z_i = 1) = \theta_1$
  $P(X_i = 1 \mid Z_i = 2) = \theta_2$
- Are **both populations the same or different?**



	10/10	20/20	25/25	50/50	75/75	90/90	5/10	10/15	10/20	10/30	20/50	20/75	20/100
Fisher	0.15	0.42	0.49	0.65	0.72	0.75	0.06	0.30	0.33	0.36	0.55	0.58	0.60
ChiSq	0.25	0.49	0.56	0.69	0.75	0.77	0.17	0.37	0.37	0.42	0.58	0.61	0.62
ObjBayes	0.12	0.29	0.36	0.52	0.60	NA	0.00	0.17	0.19	0.22	0.38	0.42	0.43
M1	0.07	0.28	0.32	0.50	0.58	0.62	0.00	0.11	0.15	0.21	0.37	0.41	0.42
RIP	0.20	0.35	0.39	0.53	0.60	0.63	0.12	0.23	0.25	0.29	0.42	0.45	0.46

In constrast to the safe test, the objective Bayes test *does not* handle dependent test combinations!
(Type-I error guarantee breaks down)

Example 2: Jeffreys' (1961) Bayesian t-test

t-test setting

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

- 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 best (**higher power**) safe test!

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

1. Combining (in)dependent tests, adding extra data
2. Results do not depend on counterfactuals
3. More than two decisions: not just "accept/reject"

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

1. Combining (in)dependent tests, adding extra data
2. Results do not depend on counterfactuals
3. 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. Results do not depend on counterfactuals
3. 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)

All Safe Tests have a gambling and MDL (data compression) interpretation
(with again, advantages over standard MDL codes)

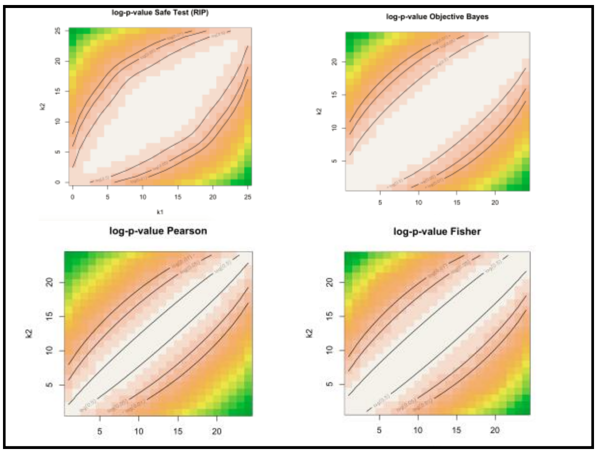
Safe Testing unifies yet improves the main testing paradigms

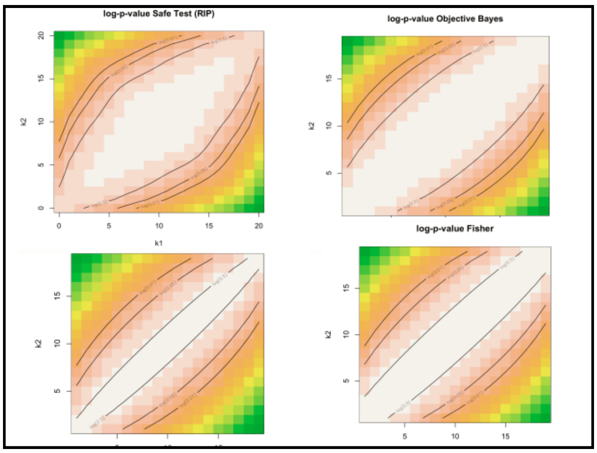
Read more?

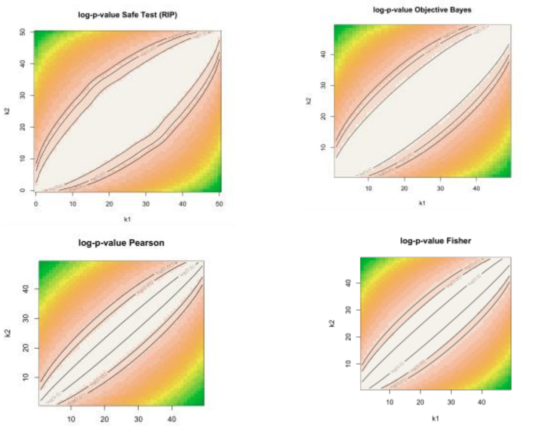
- S. van der Pas and G. ***Almost the Best of Three Worlds***. Accepted for Statistica Sinica
- G. ***Safe Probability***, Arxiv 2016
- Reversed I-Projection and Learning Theory: Van Erven, G., Mehta, Reed and Williamson, ***Fast Rates in Statistical and Online Learning***, JMLR 2015

Much more to come...

Additional Material







2. Standard p-values depend on counterfactuals, TM's do not

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result ($p = 0.03$ based on fixed $n=100$). But just to make sure I ask a statistician whether I did everything right.

2. Standard p-values depend on counterfactuals, TM's do not

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result ($p = 0.03$ based on fixed $n=100$). But just to make sure I ask a statistician whether I did everything right.
- Now the statistician asks: what *would* you have done if your result had been 'almost-but-not-quite' significant?

2. Standard p-values depend on counterfactuals, TM's do not

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result ($p = 0.03$ based on fixed $n=100$). But just to make sure I ask a statistician whether I did everything right.
- Now the statistician asks: what *would* you have done if your result had been 'almost-but-not-quite' significant?
- I say "Well I never thought about that. Well, perhaps, but I'm not sure, I would have asked my boss for money to test another 50 patients".

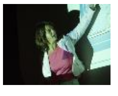
2. Standard p-values depend on counterfactuals, TM's do not

- Suppose I plan to test a new medication on exactly 100 patients. I do this and obtain a (just) significant result ($p = 0.03$ based on fixed $n=100$). But just to make sure I ask a statistician whether I did everything right.
- Now the statistician asks: what *would* you have done if your result had been 'almost-but-not-quite' significant?
- I say "Well I never thought about that. Well, perhaps, but I'm not sure, I would have asked my boss for money to test another 50 patients".
- Now the statistician has to say: ***that means your result is not significant any more!***

No Issues with Counterfactuals

- You can use martingale tests to find out who is the best weather forecaster!
- Use

$$M(X^n) = \prod_{t=1}^{n} \frac{P_{\text{Peter}}(X_t \mid X^{t-1})}{P_{\text{Margot}}(X_t \mid X^{t-1})}$$



Advantages of Martingale over Bayesian Testing

- In fact most arguments put forward in the 1960s in favor of Bayesian testing are as just shown and can just as well be used to argue in favor of martingale tests
- Yet you can do things with martingale tests that you cannot do with Bayes tests...
  - Ryabko 2005: compression test (MDL≈test martingale approach if H_0 simple)
  - switch distribution...