Eclectic Lectures

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

for all $P \in \mathcal{H}_0$:

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

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

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

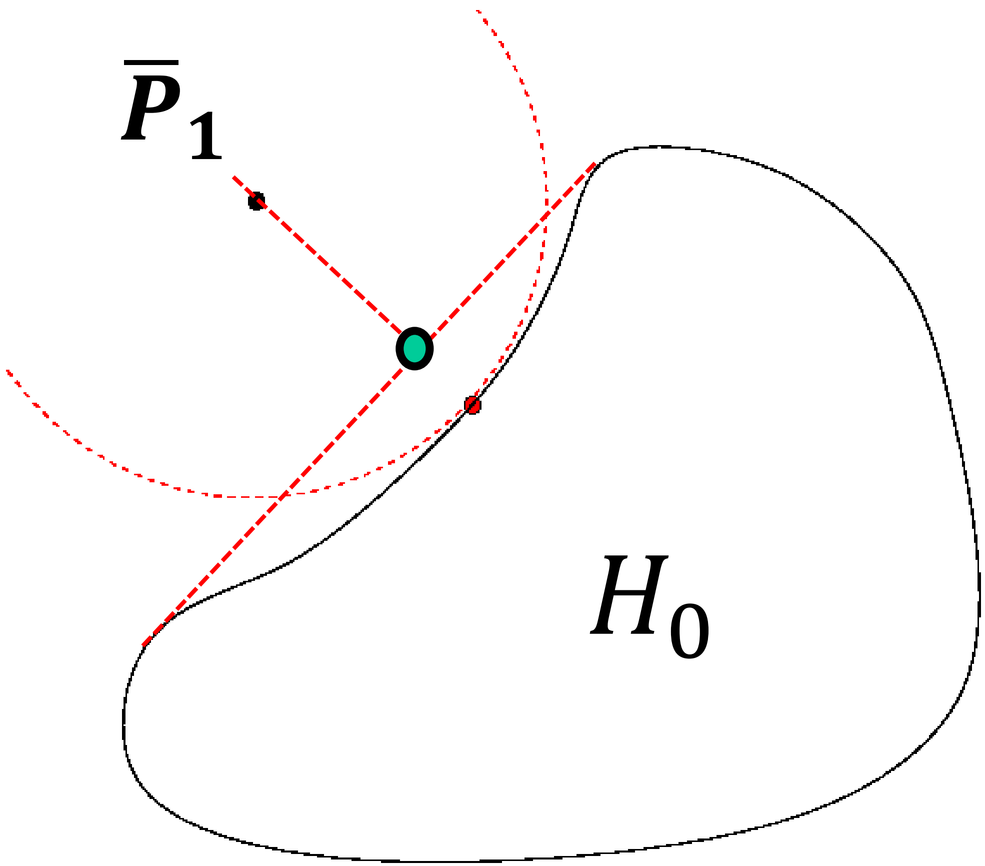
$$\sup_S \inf_{P \in H_{1,\delta}} \mathbf{E}_{X^n \sim P}[\log S]$$

- **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^*

Reverse Information Projection



is $\tilde{P}_0$

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^*

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 $p = \bar{p}_1$

Crucial Idea for Proof

Proper scoring rule

- 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 $p = \bar{p}_1$

GROW S-Value for simple H_0:

- 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 S-Value for simple $\boldsymbol{H_0}$

- The GROW S-value relative to $H_{1,\delta}$ achieves

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

- In case we are 'also' a classical frequentist, we are given an α and may want to pick $H_{1,\delta} \subset H_1$ such that power is maximized
- $H_0 = \{ P_0 \},\ H_1 = \{ P_\theta :\ \theta > 0 \}$ 1-dim exponential family: solution is to put point prior putting mass 1 on θ_n^* such that $D(P_0 || P_{\theta_n^*}) = \mathrm{n}^{-1} \cdot\ \log\left(\frac{1}{\alpha}\right)$
- ....so that $S = p_{\theta_n^*}(X^n)/p\_0(X^n)$

GROW S-Value for simple H_0

- The GROW S-value relative to $H_{1,\delta}$ achieves

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

- In case we are 'also' a classical frequentist, we are given an α and may want to pick $H_{1,\delta} \subset H_1$ such that power is maximized
- $H_0 = \{ P_0 \}, \; H_1 = \{ P_\theta : \; \theta > 0 \}$ 1-dim exponential family: solution is to put point prior putting mass 1 on θ_n^* such that $D(P_0 || P_{\theta_n^*}) = \mathrm{n}^{-1} \cdot \log \left(\frac{1}{\alpha}\right)$
- ....so that $S = p_{\theta_n^*}(X^n) / p\_0(X^n)$ (depends on n !)

Rejection Regions for Simple H_0

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

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

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

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

Close to Optimal Power
Safe, Not Consistent

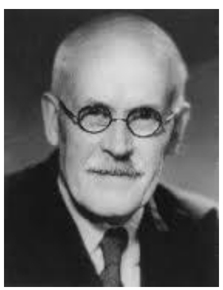
Menu

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
   - RIPr (Reverse Information Projection)
   - JIPR (Joint Information Projection)
5. **Historical Perspective**
6. S-Values and Test Martingales

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

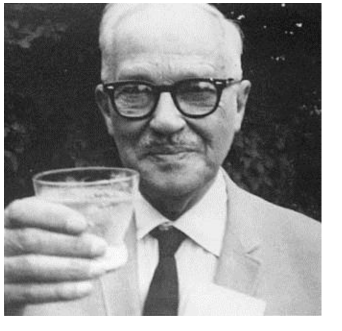
- *Before* experiment is done, state *significance level* α (e.g. $\alpha = 0.05$)
- **Reject** H_0 iff $\mathrm{p} < 0.05$
- This gives **Type-I Error** Guarantee of α
- If statisticians would follow this procedure for fixed α in all their experiments, the fraction of times in which the null hypothesis 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!

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

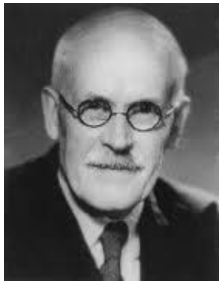
- To some extent, S-values *do* allow us to combine the features of Fisherian and Neymanian testing!
- S-value measures ‘unlikeliness’, even without alternative, just like p-value
- ...but behaves much better under optional continuation
- S-value leads to Type-I error/loss guarantees, even under optional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

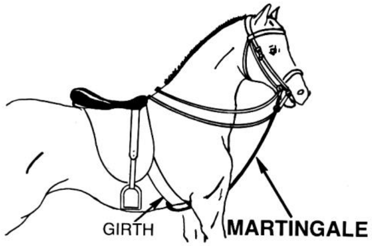
Earlier Work on S-Values

- 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**
- 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

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
   - RIPr (Reverse Information Projection)
   - JIPR (Joint Information Projection)
5. Historical Perspective
6. **S-Values and Test Martingales**
   - **Optional Stopping vs Optional Continuation**

Optional Stopping

- S-values defined as functions on data X^n of fixed size n (or X^τ for fixed stopping rule τ)
- After each minibatch $X_{n_{j-1}}, \ldots X_{n_{j+1}}$, can decide to stop or continue and do new test (and multiply results): optional continuation
- What if we want to be able to stop at each n and not just at the end of each minibatch? (optional stopping)
- First idea: take mini-batches of size 1 !

Simple $\boldsymbol{H_0}$, i.i.d.

Mini-Batches of size-1 idea works:

- start with prior w on Θ_1
- $\bar{p}_w(X^n) = \int_{\Theta_1} p_\theta(X^n) w(\theta) d\theta$
- $S_1 = \bar{p}_w(X_1) / p_0(X_1)$
- $S_2 = \bar{p}_w(X_2 \mid X_1) / p_0(X_2)$
- ....$S_n = \bar{p}_w(X_n \mid X^{n-1}) / p_0(X_n)$

Each S_k is an S-value, and $S_1 \cdot ... \cdot S_k$ is equal to the single S-value $S_{\langle k \rangle}$ we would have obtained if we had considered $X_1, \dots, X_k$ as a single minibatch

Simple H_0, i.i.d.

Mini-Batches of size-1 idea works:

- Each S_k is an S-value, and $S_1 \cdot \ldots \cdot S_k$ is equal to the single S-value $S_{\langle k \rangle}$ we would have obtained if we had considered $X_1, \ldots, X_k$ as a single minibatch
- Thus, our earlier optional continuation implies that we can actually stop at any time we like (e.g. as soon as $S_1 \cdot \ldots \cdot S_k \geq 20$ and the Type-I error guarantee will still be valid!
- **For simple H_0, testing with S-values is safe not just for 'optional continuation' but also for 'optional stopping'**

Simple H_0, i.i.d.

For simple H_0, testing with S-values is safe not just for 'optional continuation' but also for 'optional stopping'

But wait: what if we work with a 'power optimizing prior' that depends on n , as before?

Rejection Regions for Simple H_0

- Bayes with standard prior rejects H_0 whenever

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

- Bayes with GROW-prior chosen so as to maximize power **at sample size n^*** rejects H_0 at 5% when

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

but only if $n = n^*$

Rejection Regions for Simple H_0

- Bayes with standard prior rejects H_0 whenever

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

Safe for Optional Stopping, bound holds for all n

- Bayes with GROW-prior chosen so as to maximize power **at sample size n^*** rejects H_0 at 5% when

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

but only if $n = n^*$

Safe for OS, but no good power properties if $n^* \neq n$

Rejection Regions for Simple H_0

- Bayes with standard prior rejects H_0 whenever

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

Safe for Optional Stopping, bound holds for all n

- Bayes with GROW-prior chosen so as to maximize power **at sample size n^*** rejects H_0 at 5% when

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

but only if $n = n^*$

Safe for OS, but no good power properties if $n^* \neq n$

- Q: can we get an S-value that is safe for Optional Stopping but with a $\sqrt{1/n}$ rejection region (hence good power) for all n? A: **NO (LIL!)**

Rejection Regions for Simple H_0

- Q: can we get an S-value that is safe for Optional Stopping but with a $\sqrt{1/n}$ rejection region (hence good power) for all n ? A: **NO (Lille!)**
- **...but we can get an S-value that is safe for OS and satisfies, for all n:**

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

(still 'better' than Bayes)

- ...this is obtained by replacing $\bar{p}_1$ with the switch distribution (Van Erven et al., NIPS 2007, G. and Van der Pas, Stat. Sinica 2018)

What about composite H_0?

- Optional Stopping (with interesting little caveat) is still possible for S-values that are Bayes factors with right Haar priors (Bayes t-test etc.)
  - Minibatch of size 1 idea still works
  - (Hendriksen, De Heide & G., 2018)

What about composite $\boldsymbol{H_0}$?

- ...yet in general, ‘minibatch of size 1’ idea does not work any more...
- 2x2 contingency table test: take arbitrary prior w_1 on Θ_1, define $\bar{p}_1(X^n) = \int p_\theta(X^n)\, w_1(\theta) d\theta$
- Create S-value for $n = 1$ by doing reverse information projection. This gives $\bar{p}_0(X_1)$ such that $S = \bar{p}_1(X_1)/\bar{p}_0(X_1)$ is S-value
- Surprisingly, however, we find that $S = 1$ (it doesn’t listen to the data...)
- “All Bayes marginals for $n = 1$ relative to H_1 are also Bayes margonals relative to H_0”

What about composite $\boldsymbol{H_0}$?

- Many open questions:
- Can we use ‘minibatches of size 2’?
- Can we obtain S-values that allow OS at all?
- If so, can we make sure they have rejection regions of size

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

Test Martingales vs S-Values

- Suppose we are given a sequence of S-Values $S_1, S_2, \ldots$ for data $(X_1, \ldots, X_{n_1})$, $(X_{n_1+1}, \ldots, X_{n_2})$, ...
- The random process $\left(S^{\langle 1 \rangle}, S^{\langle 2 \rangle}, \ldots\right)$, $S^{\langle k \rangle} := \prod_{j=1..k} S_j$ is a nonnegative supermartingale
- Our earlier 'optional continuation' theorem is instance of Doob's optional stopping theorem for martingales
- In situations in which the 'minibatch of size 1' idea works, we have S_j a function of X_j only.
- ...then we can indeed stop at any n we like. For such cases, $S^{\langle k \rangle}$ has been called test martingale

(gambling at each n rather than each minibatch)

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

NP philosophy depends heavily on counterfactuals, S-values a little, TMs 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.

The Counterfactual Issue

- 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?

The Counterfactual Issue

- 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”.

The Counterfactual Issue

- 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!***

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

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

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 $L(0, a_1) \leq \dfrac{1}{\alpha}$

In terms of Loss Functions:

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

Frequentist Type-I Error Guarantee:

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

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:

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

Just 2 actions:

$$L(0, a_0) = 0 \qquad\qquad L(0, a_2) = 100$$

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:

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

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:

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

doesn’t work!

Many actions:

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

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

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

Many actions:

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

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

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

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:

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

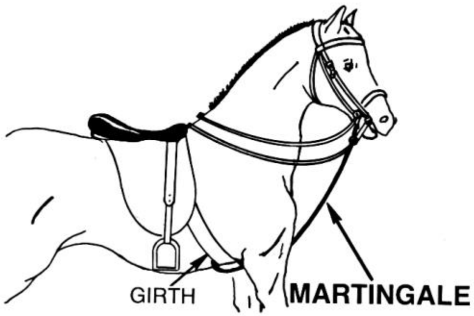
(works also with countably ∞ many actions)

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-value is better...



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!