MDL , E-Values, Evidence

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University

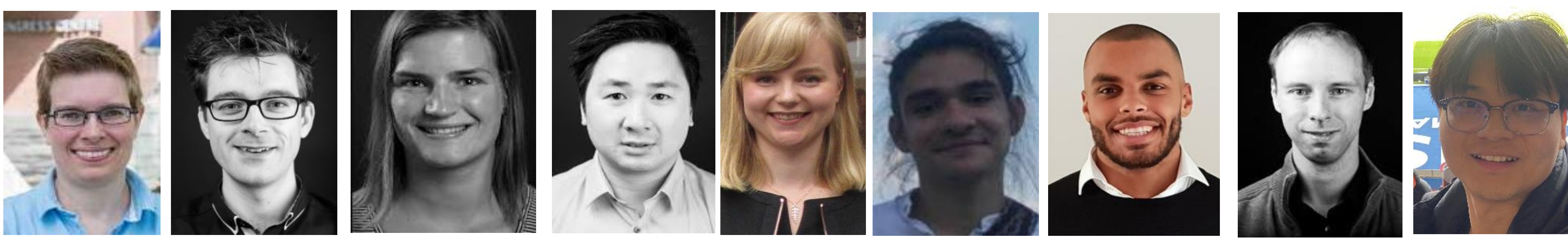
        

Anytime Valid Methods, MDL, perhaps e-values

Peter Grünwald

Centrum Wiskunde & Informatica – Amsterdam

Mathematical Institute – Leiden University



Uniform Tests of Randomness (1976)

Доклады Академии наук СССР
1976. Том 227, № 1

УДК 519.211+517.11 *МАТЕМАТИКА*

Л. А. ЛЕВИН

РАВНОМЕРНЫЕ ТЕСТЫ СЛУЧАЙНОСТИ

(Представлено академиком А. Н. Колмогоровым 14 X 1975)

1. В теории сложности даются определения ряда понятий: сложности, случайности, количества информации априорной вероятности (см., например, ([1–9])). В настоящей работе предлагается единый подход к такого рода

…then almost nothing happened until **2019** when Levin's concept was given a name:

E-Value

…then almost nothing happened until **2019** when Levin's concept was given a name:

E-Value

…serve as an alternative to **p-values**

…essential to do **testing** and **confidence intervals** in

Anytime-Valid

context

…e-values were given a **name** only in 2019 when the following breakthrough papers appear on arXiv:

…e-values were given a **name** only in **2019** when the following breakthrough papers appear on arXiv:

Safe Testing

(Grünwald, De Heide, Koolen, now *Journal of the Royal Statistical Society*)

…e-values were given a **name** only in **2019** when the following breakthrough papers appear on arXiv:

Safe Testing

(Grünwald, De Heide, Koolen, now *JRSS*)

E-Values: Calibration, Combination and Applications

(V. Vovk, R. Wang, now *Annals of Statistics)*

…e-values were given a **name** only in **2019** when the following breakthrough papers appear on arXiv:

Safe Testing

(Grünwald, De Heide, Koolen, now *JRSS*)

E-Values: Calibration, Combination and Applications
(V. Vovk, R. Wang, now *Annals of Statistics)*

Testing by Betting

(G. Shafer, now *JRSS*)

…e-values were given a **name** only in **2019** when the following breakthrough papers appear on arXiv:

Safe Testing

(Grünwald, De Heide, Koolen, now *JRSS*)

E-Values: Calibration, Combination and Applications

(V. Vovk, R. Wang, now *Annals of Statistics)*

Testing by Betting

(G. Shafer, now *JRSS*)

Universal Inference

(L. Wasserman, A. **Ramdas**, S. Balakrishnan, now *Proceedings National Academy of Sciences USA*)

2023: 100s of papers...eg in *Annals, JRRS, Biometrika, Neurips, JASA, Statistical Science*

Ramdas' group at CMU and my group at CWI the most active…

2023: 100s of papers...eg in *Annals, JRRS, Biometrika, Neurips, JASA, Statistical Science*

Ramdas' group at CMU and my group at CWI the most active…

NEWSFLASH:

Aaditya Ramdas just received the 2023 Institute of Mathematical Statistics **Early Career Prize** "for significant contributions in […] **reproducibility** in science […] active, **sequential** decision-making and **assumption-light uncertainty quantification**. ….. the prize recognizes Dr. Ramdas' outstanding **potential to shape the future of statistics.**"

Brittleness of Classical, “Frequentist” Testing and Confidence Intervals

Null Hypothesis Testing

- Null Hypothesis: **status quo**
- "Coin is Fair"
- No Difference between **Treatment** and **Control**

Null Hypothesis Testing

Prototypical case: ***z*-test:**

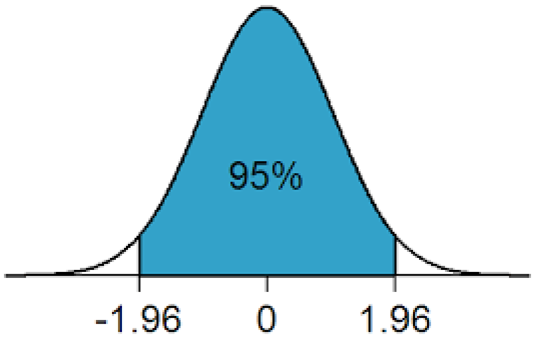
$X_1, X_2, \ldots$ independently identically distributed (i.i.d.),
Gaussian, variance 1

H_0: mean is some μ_0 (usually 0)

H_1: mean $\neq \mu_0$

Classical, p-value based testing

- I test new medication on n patients at level α
  n decided upon in advance
- p_n : p-value for null hypothesis H_0 at n,
  i.e. data $X^n = (X_1, \ldots, X_n)$
- If $p_n \leq \alpha$ I “reject” the null, otherwise I “accept” it
- z-test,standard $\alpha = 0.05 \Leftrightarrow$ “reject iff $|\bar{X} - \mu_0| \geq \frac{1.96}{\sqrt{n}}$ “



Issues with classical tests and CIs

- p_n : p-value for null hypothesis H_0, n data points
- I test new medication on $n = 100$ patients at level α
- At $n = 50$ boss says: let's **peek** at data.
  Perhaps we can reject null already!

Issues with classical tests and CIs

- p_n : p-value for null hypothesis H_0, n data points
- I test new medication on $n = 100$ patients at level α
- At $n = 50$ boss says: let's **peek** at data.
  Perhaps we can reject null already!
- …I find that $p_{50} \leq \alpha$ so I reject the null
- Is this OK?

Issues with classical tests and CIs

- p_n : p-value for null hypothesis H_0, n data points

- I tes[illegible]w medic[illegible]n on $n = 100$ patients [illegible]vel α
- At n [illegible]boss s[illegible]: let's peek at data.
  Perh[illegible] can [illegible]ect null [illegible]
- …I f[illegible] tha[illegible]$_0$ ≤ [illegible] so I [illegible]ect the [illegible]
- Is th[illegible]K?

No!

Issues with classical tests and CIs

- I test new medication on $n = 100$ patients
- At $n = 50$ boss says: let's peek at data.
- …I find that $p_{50} \leq \alpha$ so I already reject the null
- Is this OK?

NO!…because then you violate **Type-I error guarantee**, the method's cornerstone

You will conclude "there is an effect" (far) too often!

Type- I Error Guarantee: if the null is true then the probability I reject it (claim a nonexisting effect) is $\leq \alpha$

Issue runs deeper

- p_n : p-value for null hypothesis H_0, n data points

- I test new medication on $n = 100$ patients at level α
- At $n = 50$ boss says: let's peek at data.
  Perhaps we can reject null already!
- ..I find that $p_{50} > \alpha$ so I **simply wait** until $n = 100$ to make a decision
- Is this OK?

Issue runs deeper

- p_n : p-value for null hypothesis H_0, n data points

- Test new medication on $n = 100$ patients at level α
- At n ...0 boss s...: let's peek at data.
  Perh... ...e can ...ect null already!
- ..we ...t p_{50} ... α so w... **wait** u... $n =$
  100 ... ma... de...ion
- Is th... OK?

No!

classical stats: almost as spooky as quantum mechanics…

- By **merely peeking** at the data we destroy validity of accept/reject conclusion, even if we don’t act upon the data we actually saw
- …only if we can guarantee that no matter what data we will see upon early peeking, we will **never** act upon it can we guarantee validity of our conclusion
  - …but then we are indistinguishable from an agent who does **not** peek at the data…

Replication Crisis in Science

“at least 50% of highly cited results in medicine is irreproducible”

J. Ioannidis, PLoS Medicine 2005

Replicatio… s in Scienc…

somehow … p-values and

signi… …ing…

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

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

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

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

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

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

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

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

somehow ... and

sign... ...ing!

...IATION

...North Washington Street, Al...

AMERICAN STATISTICAL ASS...
STATISTICAL SIGNIFICA...

Provides Principles to Improve the Conduct and ...
Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical ... and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-va... [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

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

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

...ificance ... 2017,
...ticians
Rede...
(to $p < 0.0$...
incl. some of th...

...lication...

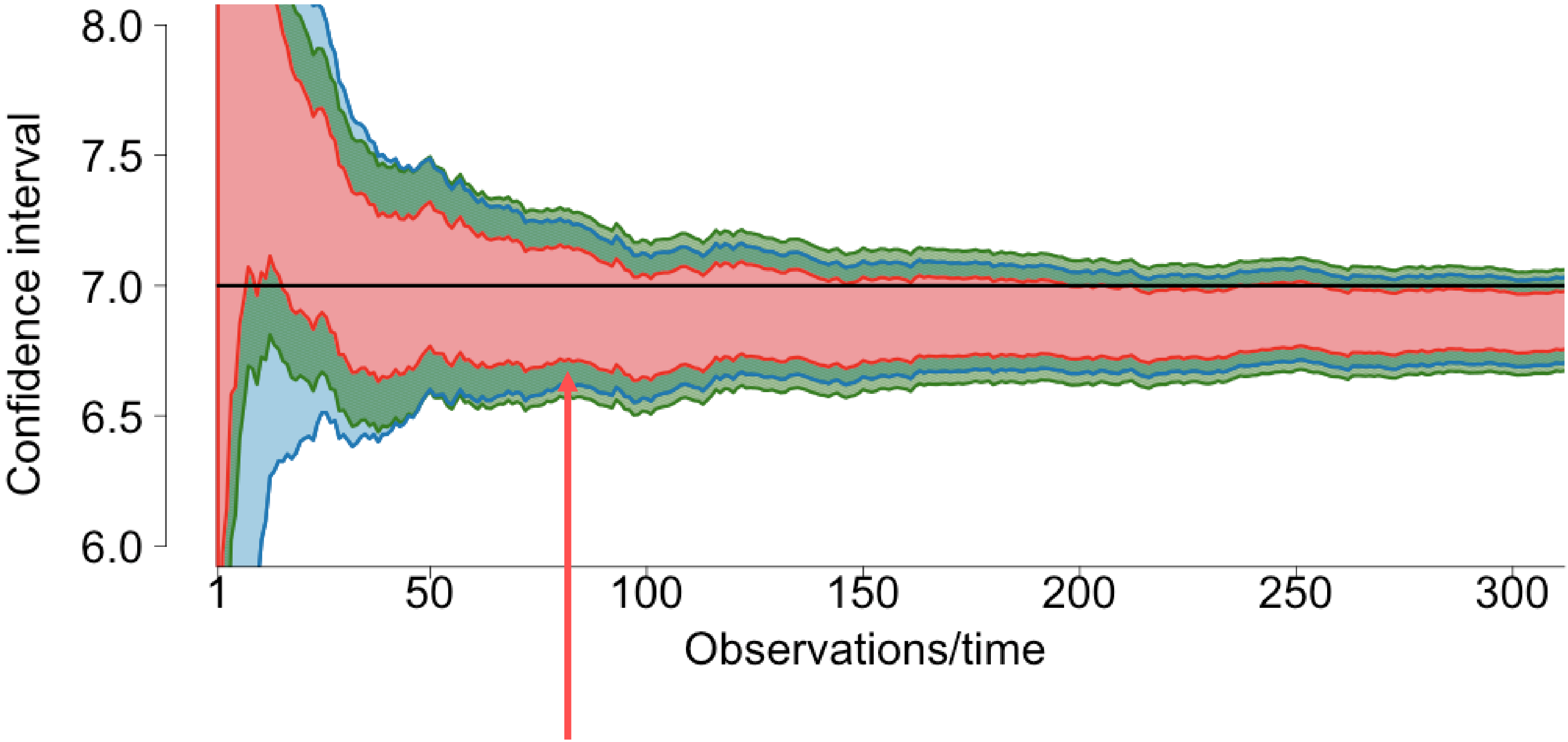
...and

AMERICAN STATISTICAL ASS...
STATISTICAL SIGNIFICA...

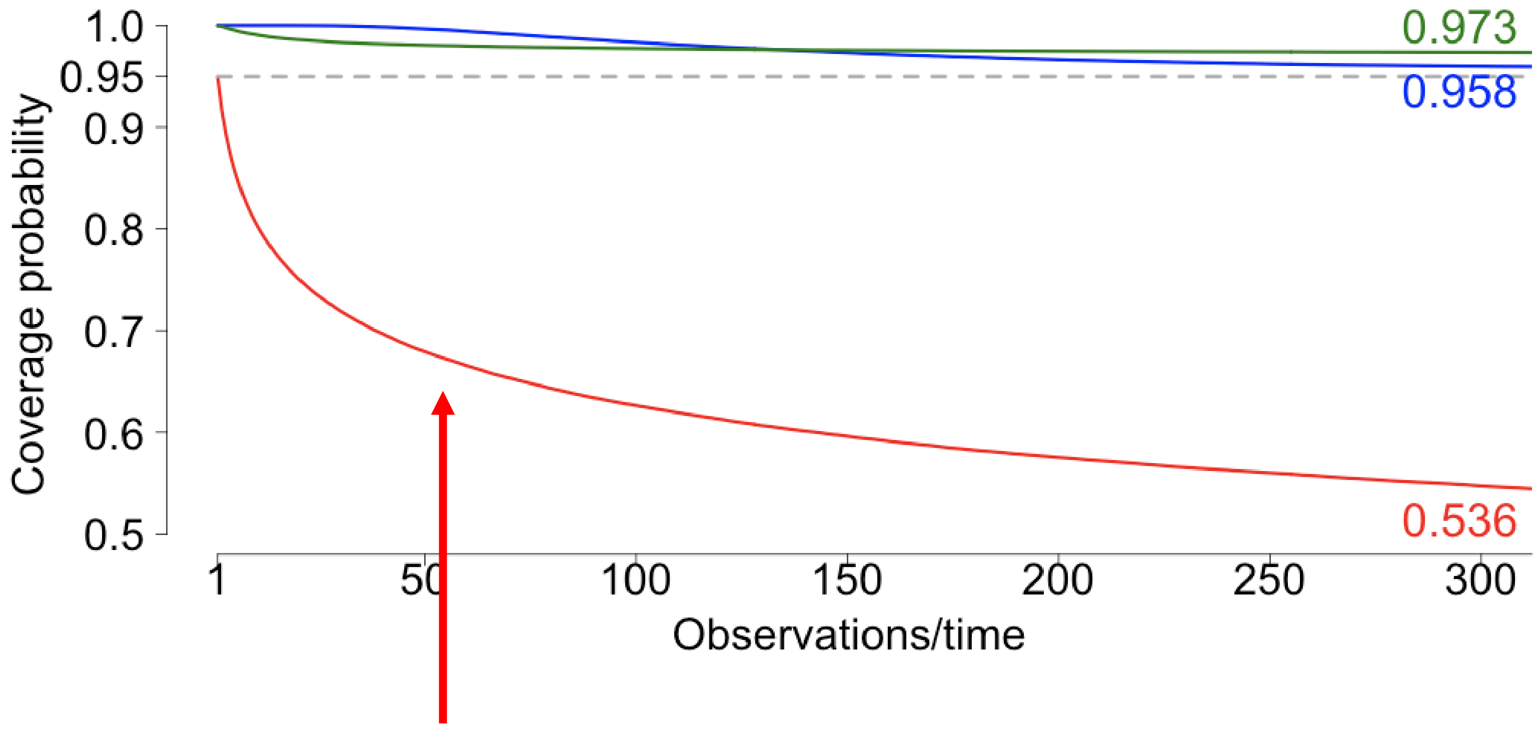
Provides Principles to Improve the Conduct and ...
Science
March 7, 2016

The American Statistical Association (ASA) has released a “Statement on Statistical ... and *P*-Values” with six principles underlying the proper use and interpretation of the *p*-va... [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

Z-test ⇒ Z-Confidence Interval

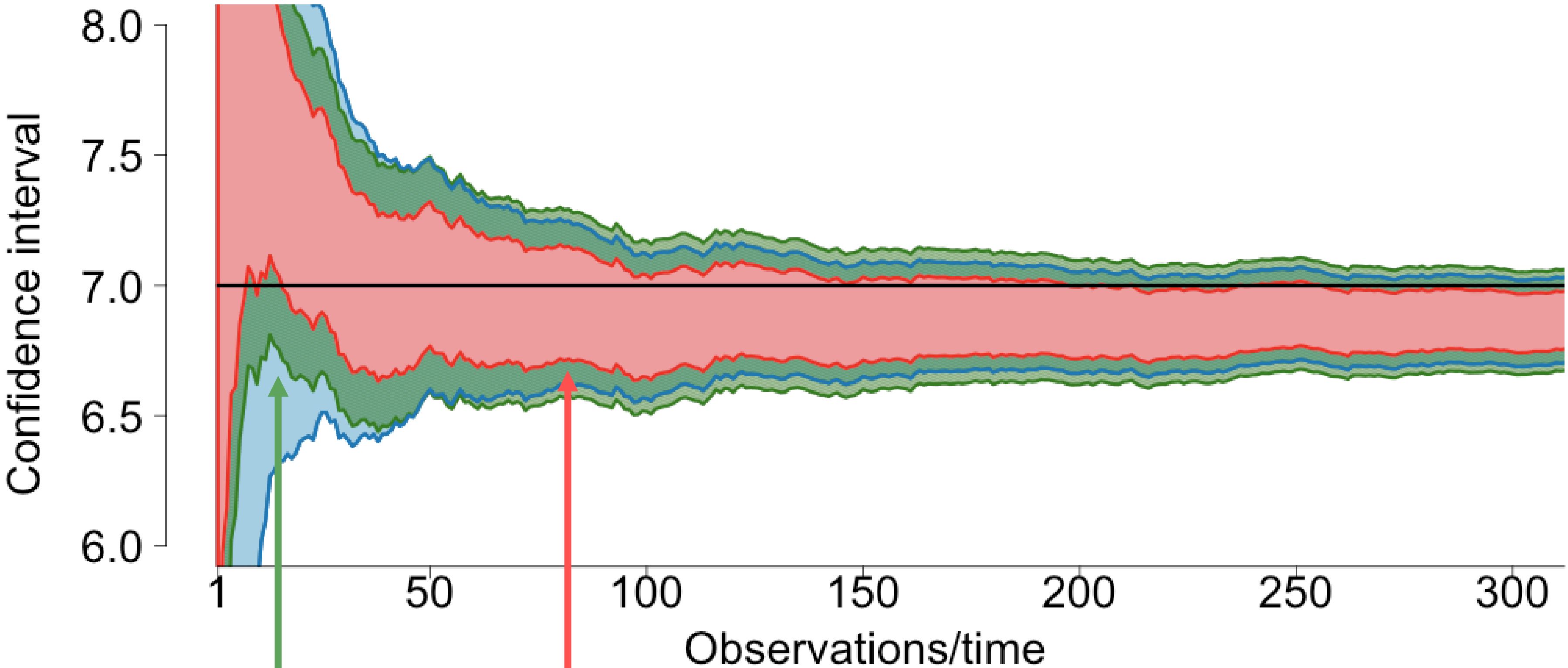


standard 95% CI: $\bar{X} \pm 1.96/\sqrt{n}$



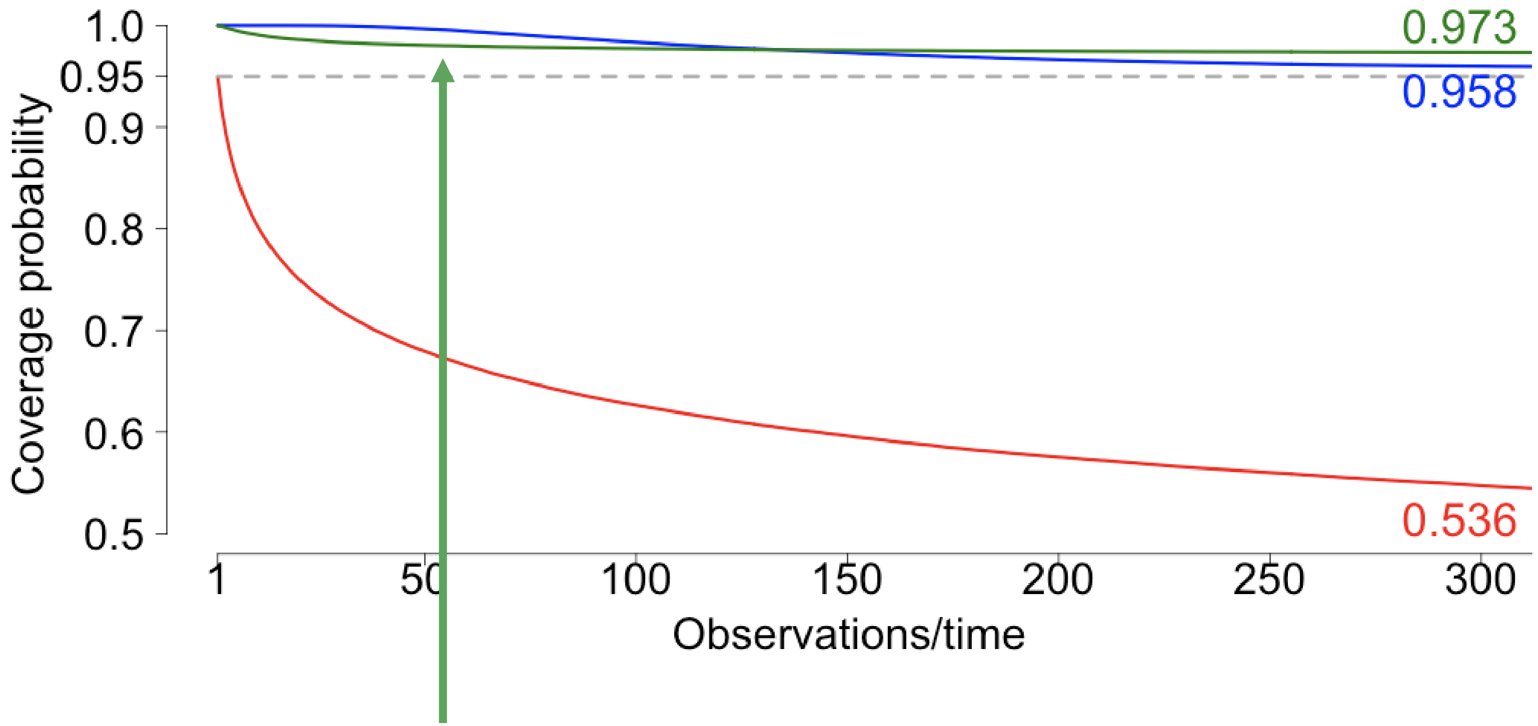
Suppose $H_0: \mu = 7$ is **true** yet you keep sampling until H_0 can be rejected (falls outside of the CI) or some $n_{\max}$ has been achieved. We plot the probability that θ is contained in your CI at time $n_{\max}$ as function of $n_{\max}$

Anytime-Valid Confidence Interval (“Confidence Sequence”)

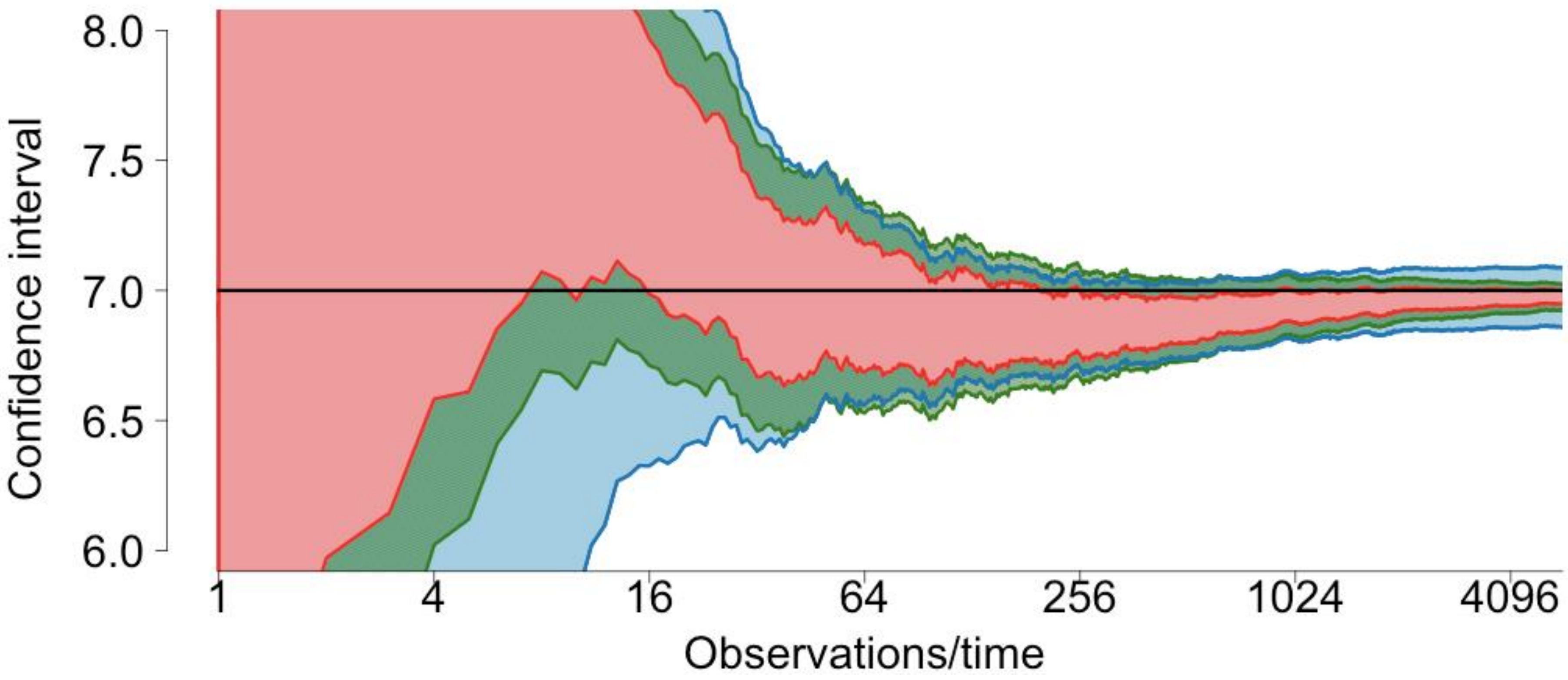


standard CI: $\bar{X} \pm 1.96/\sqrt{n}$

anytime-valid CI based on “non-informative” prior distribution



Suppose $H_0\!: \mu = 7$ is **true** yet you keep sampling until H_0 can be rejected (falls outside of the CI) or some $n_{\max}$ has been achieved. We plot the probability that θ is contained in your CI at time $n_{\max}$ as function of $n_{\max}$



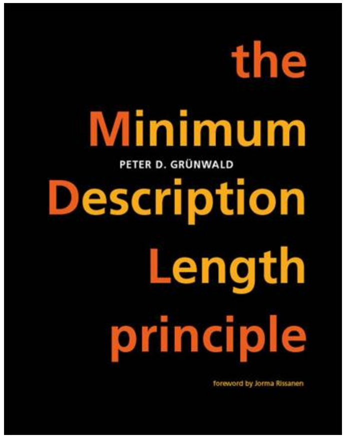
standard CI: $\bar{X} \pm 1.96/\sqrt{n}$

AV CI, “non-informative” prior: $\bar{X} \pm \sqrt{\frac{6+\log n}{n}}$

When repeated experiments are not possible…

- Outside of medical and psychological sciences, the very concept of error guarantees may not make sense
- …data may exhibit patterns, that may even have predictive value, but data cannot seriously be thought of as repeated realizations of a random process
- …then classical tests/CIs just make no sense at all. AV methods may make some sense…at least they can handle the fact that we might be just *given* some data…without precise knowledge of the underlying sampling plan

Part 2: Minimum Description Length Principle



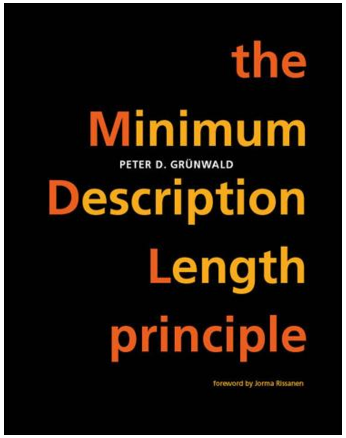
Any regularity in a sequence of data can be used to compress this data, i.e. describe it using less bits than would be used to describe the data literally…

0001000100010001….00010001

- **Pick the model that allows for the most (lossless) compression of the data**

(e.g. Alex' setting when we are NOT yet in the asymptotic regime! – that's also when you want to have uncertainty estimates)

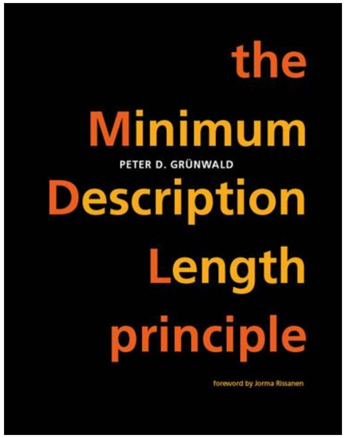
Part 2: Minimum Description Length Principle



- **Pick the model that allows for the most (lossless) compression of the data**
- Also originally (Rissanen 1978) intended as an approach towards statistics that remains meaningful even if "true distributions" don't really exist

We never want to make the false assumption that the observed data were actually generated by a distribution of some kind, say Gaussian, and then analyze the consequences. Our deductions may be entertaining, but quite irrelevant from the task at hand, which is to learn useful properties of the data (Rissanen 1990)

Part 2:
Minimum Description Length Principle

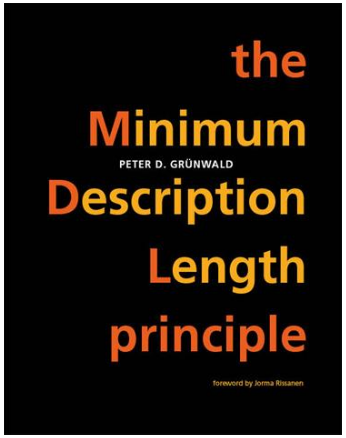


- **Pick the model that allows for the most (lossless) compression of the data**

- To make the informal idea well-defined, we must associate each model under consideration with a lossless, “universal” code

- Now probabilities inevitably come in after all

MDL Principle



- Associate each model under consideration with a lossless, "universal" code
- If a 'model' $H_0 = \{P_0\}$ really stands for just 1 distribution, this is the idealized **Shannon-Fano** code with lengths $L_{H_0}(X_1, \ldots, X_n) = -\log p_0(X_1, \ldots, X_n)$
- The Shannon-Fano code minimizes expected codelength under P_0 among all lossless codes:

$$\min_L \; E_{P_0}[\, L(X^n)] = E_{P_0}[-\log p_0\,(X^n)]$$

…minimum runs* over all uniquely decodable codes

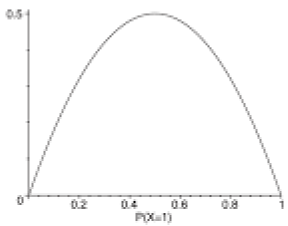
ASIDE: Interpretation of Shannon Entropy

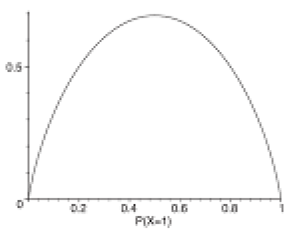
- The Shannon-Fano code minimizes expected codelength under P_0 among all lossless codes:

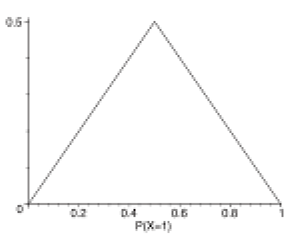
$$\min_L \; E_{P_0}[\, L(X^n)] = E_{P_0}[-\log p_0\,(X^n)] = \boldsymbol{H(P_0)}$$

…minimum runs* over all uniquely decodable codes

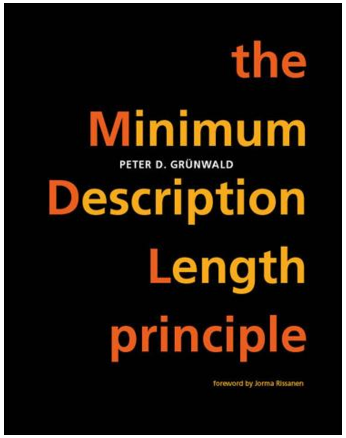
- Shannon Entropy: expected amount of bits needed to code your data if you use the **best** code i.e. the one minimizing this expected codelength
- There are many other entropies corresponding to different types of “prediction”
- Grünwald/Dawid Game Theory, Maximum Entropy,… Annals of Statistics 2004







MDL Principle

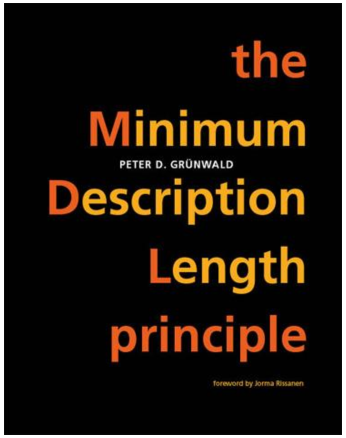


- If model $H_0 = \{P_0\}$ is simple, take **Shannon-Fano** code with lengths $L_{H_0}(X_1, \ldots, X_n) = -\log p_0(X_1, \ldots, X_n)$

- If model H $=\{P_\theta : \theta \in \Theta\}$ is larger (even nonparametric), we take code such that $\mathrm{E}_{P_\theta}[L_H(X^n) - [-\log p_\theta(X^n)]]$ is small for all $\theta \in \Theta$

no matter what P_θ actually obtains, we will not need many more bits to encode our data than we would need if we would actually know P_θ **(=universality)**

Minimum Description Length Principle



- If model H $=\{P_\theta : \theta \in \Theta\}$ is large,take code s.t. $\mathrm{E}_{P_\theta}[L_H(X^n) - [-\log p_\theta(X^n)]]$ is small for all $\theta \in \Theta$

For parametric models, universal codes can be designed based on two-part techniques…

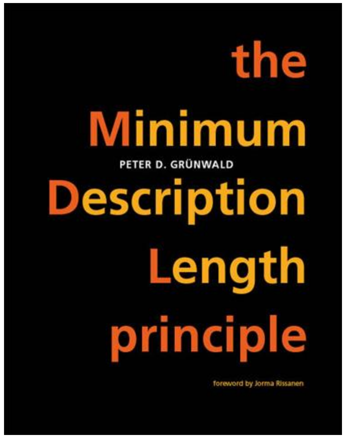
$$L_H(X^n) = \min_{\theta \in \ddot{\Theta}} \underbrace{L(\theta)}_{= -\log w(\theta)} - \log p_\theta(X^n)$$

"explicit regularization"

Misspecification: use $\beta(-\log w(\theta))$, correction factor β

Optimal prior: Jeffreys' based on Fisher information

Minimum Description Length Principle



- If model H $=\{P_\theta : \theta \in \Theta\}$ is large,take code s.t.
  $\mathrm{E}_{P_\theta}[L_H(X^n) - [-\log p_\theta(X^n)]]$ is small for all $\theta \in \Theta$

For parametric models, universal codes can be designed based on two-part techniques…

$$L_H(X^n) = \min_{\theta \in \ddot{\Theta}} L(\theta) - \log p_\theta(X^n)$$

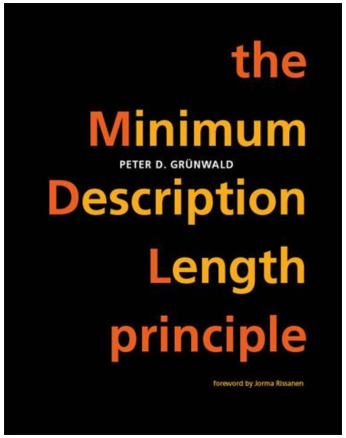
or as pseudo-Bayesian marginal likelihoods

$$L_H(X^n) = -\log \int p_\theta(X^n) w(\theta) d\theta$$

or in terms of sequential prediction errors

$$L_H(X^n) = \sum_{i=1..n} -\log p_{\hat{\theta}(X^{i-1})}(X_i)$$

Minimum Description Length Principle



- If model H $=\{P_\theta : \theta \in \Theta\}$ is large,take code s.t. $\mathrm{E}_{P_\theta}[L_H(X^n) - [-\log p_\theta(X^n)]]$ is small for all $\theta \in \Theta$

For parametric models with k parameters, all universal codes asymptotically achieve

$$L_H(X^n) = \frac{k}{2}\log n - \log p_{\hat{\theta}}(X^n) + \mathrm{const.}$$

MDL and BIC

- Suppose we compare two models H_0 and H_1
- We pick H_j for which $L_{H_j}(X^n)$ is the smallest
- If we keep dimensionality k_j of each model fixed, and model is sufficiently regular, then for sufficiently large n we pick H_j with minimal $-\log p_{\hat{\theta}_j}(X^n) + \frac{k_j}{2} \log n$

MDL and BIC

- Suppose we compare two models H_0 and H_1
- We pick H_j for which $L_{H_j}(X^n)$ is the smallest
- If we keep dimensionality k_j of each model fixed, and model is sufficiently regular, then for sufficiently large n we pick H_j with minimal $-\log p_{\hat{\theta}_j}(X^n) + \frac{k_j}{2} \log n$
- …which is also the model selected by **BIC**
- **This led some to believe that MDL=BIC, but that is an incredibly misleading statement**
- **…for small n, model "complexity" depends on so much more than the number of parameters…**

MDL and Bayes

- Similar remarks apply to MDL and Bayes factor approaches…
- …they are similar in low-dimensional cases, but (completely) diverge in some other cases

G. The E-Posterior. Phil. Trans. Roy. Society of London, 2023

MDL and Cross/Forward Validation

- It is often thought that such an information-theoretic approach is at odds with cross-validation…

- …but it is not: **it can be re-interpreted in terms of forward (or “prequential”) validation, a variation of cross-validation**

  (Rissanen ‘84, Dawid ‘84)

MDL and anytime-valid methods

- If null is simple, then $S(X^1), S(X^2), \ldots$ with

  $$S(X^n) \ := \exp\left(L_{H_0}(X^n) - L_{H_1}(X^n)\right)$$

. is an instance of what is called an "e-process"

- E-processes are the basis of anytime-valid tests, since…

E-Processes and Anytime-Valid Testing

$$\mathrm{P}_0\left(\max_{n=1,2,\ldots} S(X^n) \geq \frac{1}{\alpha}\right) \leq \alpha.$$

so the procedure that rejects the null iff e-value $S \geq 1/\alpha$, has Type-I error guarantee α **no matter when one stops sampling**

- One can always stop for **any** reason, and always continue for **any** reason

MDL and anytime-valid methods

- If null is simple, then $S(X^1), S(X^2)$, ... with

$$S(x^n) \coloneqq \exp\left(L_{H_0}(X^n) - L_{H_1}(X^n)\right)$$

. is an instance of what is called an “e-process”

- If one rejects the null if $S(X^n) \geq 1/\alpha$, one obtains a procedure with Type-I error α , no matter how the sample size n was arrived at
- …one gets Type-I error bounds under optional stopping and continuation – anytime validity!
- …and based on these one can make anytime-valid confidence intervals

MDL and AV

- Any e-process can be reinterpreted in terms of a codelength difference and thus e-value based testing and CIs are really* also MDL methods

- The probability, if the null hypothesis is true, that with any fixed code you will ever compress the data more than $\log_2 20 = 4.3$ bits extra compared to how much you compress with L_{H_0} is at most $\frac{1}{20} = 0.05$.

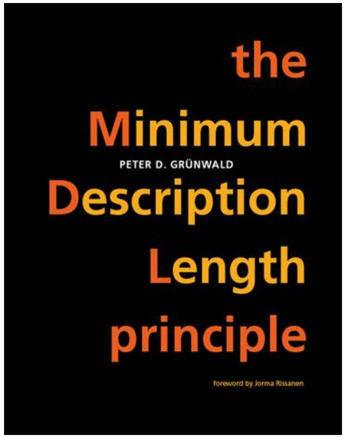
MDL and AV

- Any e-process can be reinterpreted in terms of a codelength difference and thus e-process based testing and CIs are really* also MDL methods
- The converse is not true – if H_0 is composite (large), we need an extra (but very natural!) condition on the code associated with H_0, with lengths L_{H_0}, to make exponentiated codelength difference an e-process
- …otherwise we can get, as also happens with Bayesian approaches to high-dimensional learning, "**overconfidence**"

MDL and AV

- Any e-process can be reinterpreted in terms of a codelength difference and thus e-process based testing and CIs are really* also MDL methods
- The converse is not true – if H_0 is composite (large), we need an extra (but very natural!) condition on the code associated with H_0, with lengths L_{H_0}, to make exponentiated codelength difference an e-process
- …basically condition ensures that we do not just have a valid **codelength interpretation** but also a **betting** interpretation

An Unusual Example with simple null



1. 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
   - The probability that this would ever, no matter at what length one cuts off the file, happen, is thus bounded by 2^{-200} . Really a strong refutation of the randomness hypothesis. (note that the 10 megabytes total length play no role in this computation)

Take Home

- Anytime-Valid Tests and Confidence Intervals
  - a bit less power/wider than standard CIs
  - …but much more robust: Type-I validity preserved under any-time evaluation
  - in practice always built on “e-values”
- MDL methods:
  - want to use in realms where traditional statistical assumptions do not necessarily apply
  - …but as a sanity check, if these assumptions do apply, they should give valid results! This requires special conditions and then MDL approaches become equivalent to anytime-valid approaches

(Additional Slides in Case of Questions)

E-Variables: Building Blocks of AV tests

An **e-variable** for data $X_1, X_2, \ldots, X_n$ is any **nonnegative** statistic $S = S(X_1, \ldots, X_n)$ such that if the null hypothesis holds, we have:

$$\mathbf{E}[S] \leq 1$$

if null is simple, Bayes factors are e-values

- Suppose $H_0 = \{p_{\theta_0}\}$ is just a single distribution
  - …as in our running example
- Then for any set of distributions $H_1 = \{p_\theta : \theta \in \Theta_1\}$, and any “prior” distribution $w(\theta)$,

$$S_{[\theta_0]} := \frac{\int p_\theta(X_1,\ldots,X_n) w(\theta) d\theta}{p_{\theta_0}(X_1,\ldots,X_n)}$$ is an e-variable,

…since $\mathbf{E}\left[S_{[\theta_0]}\right] =$

$$\int p_{\theta_0}(x_1, \ldots, x_n) \cdot \frac{\int p_\theta(x_1, \ldots, x_n) w(\theta) d\theta}{p_{\theta_0}(x_1, \ldots, x_n)} \mathrm{d}x_1 \ldots \mathrm{d}x_n = 1$$

$$\mathbf{E}\left[S(X_1,\ldots,X_n)\right]\leq 1$$

- E-values are nonnegative. If the null is true we expect them to be small, so:
- …if they turn out **large** this provides **evidence against the null**
- In fact, if the null is true, then for any $0 < \alpha \leq 1$:

$$\mathbf{P}\left(S(X_1,\ldots,X_n)\geq \frac{1}{\alpha}\right)\leq \alpha.$$

E-Values and Classical Testing

$$\mathbf{P}\left(S(X_1,\ldots,X_n) \geq \frac{1}{\alpha}\right) \leq \alpha.$$

so the procedure that rejects the null iff e-value $S \geq 1/\alpha$, has Type-I error guarantee α

E-values can be used for classical testing!

Optional Continuation

- …but now suppose we decide to do a second test, because the results look promising…
- based on additional data $X_{n+1}, \dots X_{n_2}$ we calculate a new e-value $S'(X_{n+1,} \dots, X_{n_2})$

Fundamental Insight:

if we multiply both e-values, we get a new e-value, which can still be used for testing

…and we can multiply in a third, and a fourth…

Optional Continuation Theorem

Let $S_1, S_2, \ldots$ be a sequence of e-variables:

$S_1 = s_1\left(X^{(1)}\right), S_2 = s_2\left(X^{(2)}\right), \ldots$

with $X^{(1)}, X^{(2)}, \ldots$ independent samples, yet definition s_j of S_j allowed to depend on all past data $X^{(1)}, \ldots X^{(j-1)}$

Then for any random stopping time τ , $S^{(\tau)} = \prod_{j=1..\tau} S_j$ is an e-variable. As a consequence, if the null is true:

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

Optional Continuation Theorem

"**Theorem**". Let $S_1, S_2, \ldots$ be a sequence of e-variables $S_1 = s_1\left(X_{(1)}\right), S_2 = s_2\left(X_{(2)}\right), \ldots$

with $X_{(1)}, X_{(2)}$ independent samples, but definition s_j of S_j allowed to depend on all past data $X_{(1)}, \ldots X_{(j-1)}$

Then for any stopping time τ , $S^{(\tau)} = \prod_{j=1..\tau} S_j$ is an e-variable. As a consequence, if the null is true, **even**:

The probability that $S^{(n)}$ will **ever** grow larger than $1/\alpha$, is bounded by α : we have our Type-I error guarantee

Optional Stopping

Suppose the null is true.

The probability that $S^{(n)}$ will **ever** grow larger than $1/\alpha$, is then bounded by α

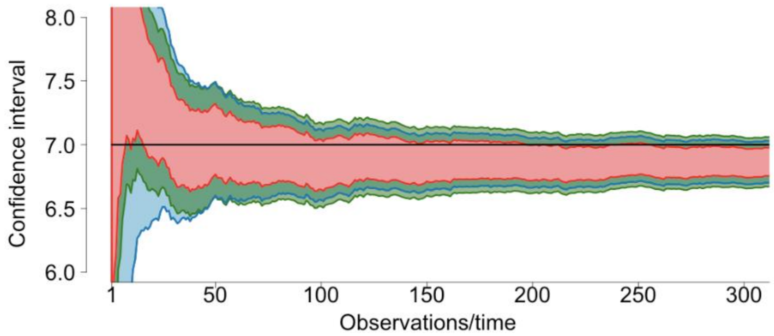
Similarly, under some further conditions it holds that the probability that there will **ever** be an n for which $S(X_1, \dots, X_n)$ is larger than $1/\alpha$, is bounded by α

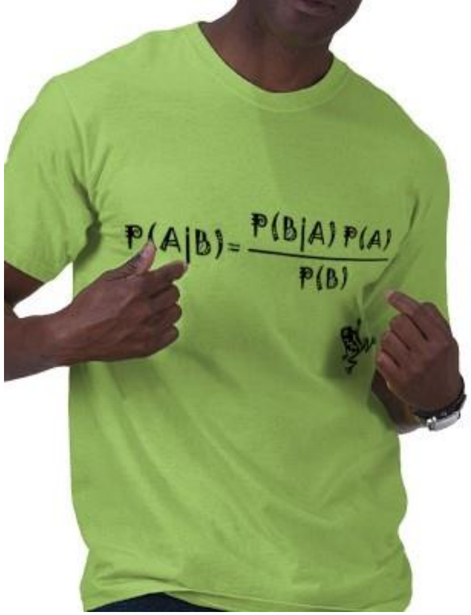
From tests to AV CIs

For every value θ of parameter of interest, let $S_{[\theta]}$ be an e-variable relative to H_0: θ represents ground truth

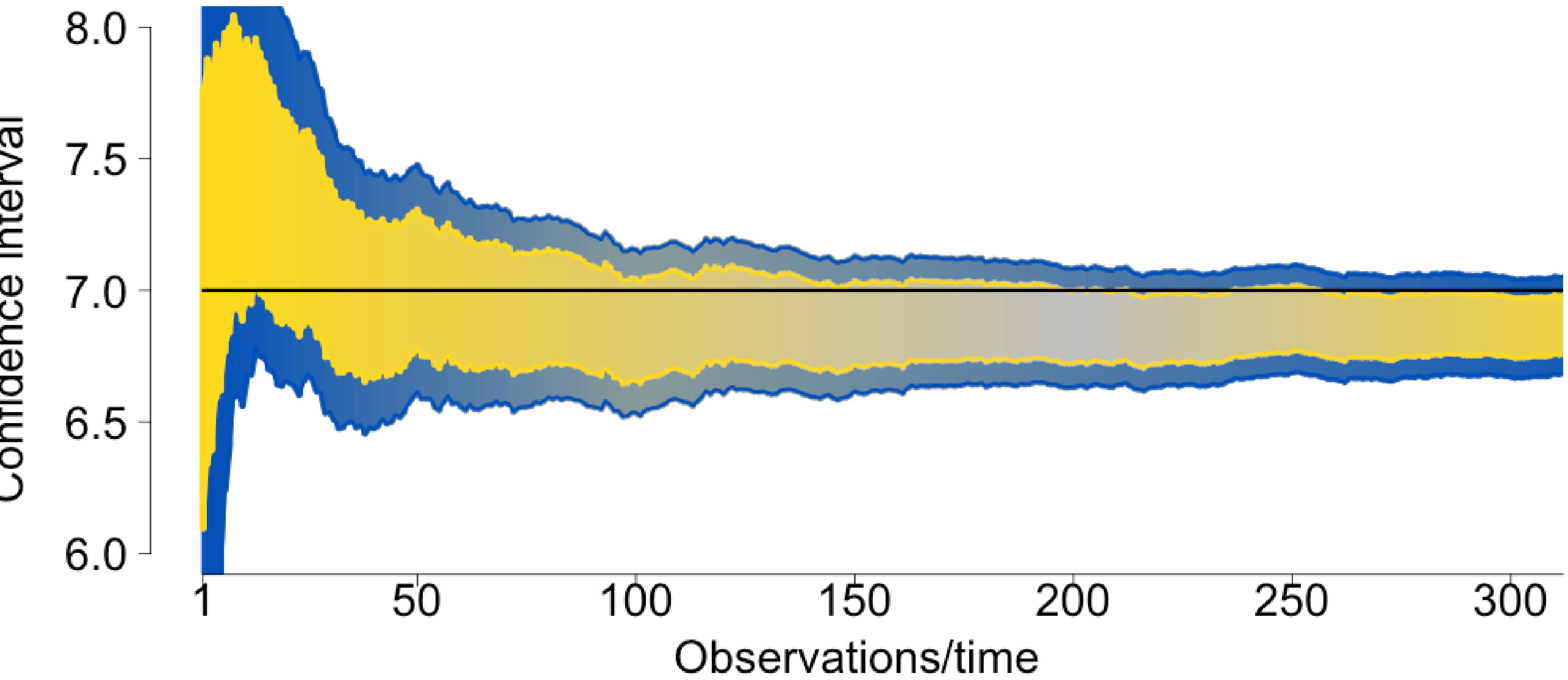
Our 95% AV CI at sample size n is now simply defined to be the set of θ such that $S_{[\theta]}(X_1, \ldots, X_n) < 20$

“The $\boldsymbol{\theta}$ we cannot reject at $\boldsymbol{n}$”





1. **Brittleness of Classical Testing and Confidence Intervals**
2. **Brittleness of Bayesian Testing and Credible Intervals**



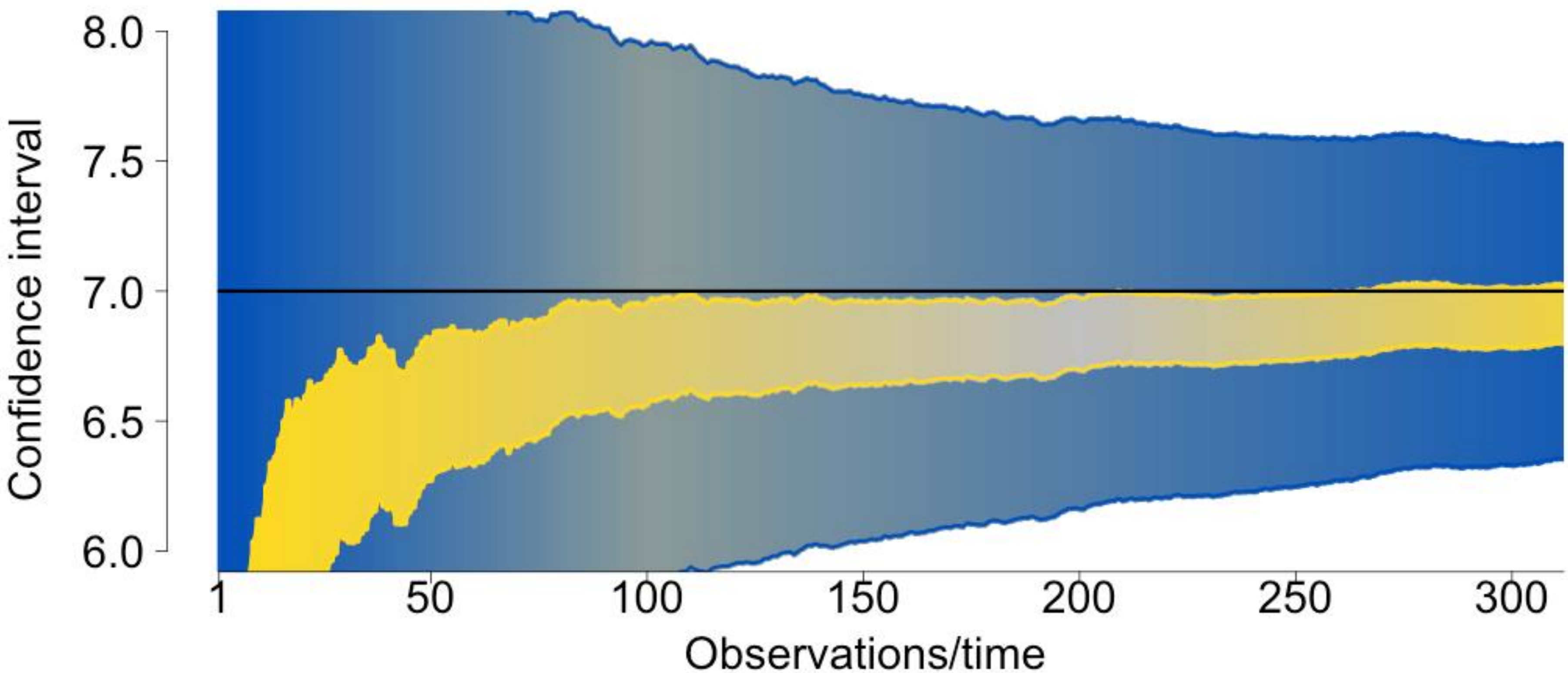
Yellow: Bayes 95% credible interval based on noninformative prior ≈ $\bar{X} \pm 1.96/\sqrt{n}$

Blue: 95% AV interval based on same prior:

Subjective and Objective, at the same time

- E-Posteriors and the AV CIs they induce rely on a prior, just like Bayesian posteriors…
- …but they **remain valid** irrespective of prior you use

…suppose for example you have a **pretty mistaken prior belief** that $\theta = 0$, with variance 0.5 …


Confidence interval
8.0
7.5
7.0
6.5
6.0
1
50
100
150
200
250
300
Observations/time

Subjective and Objective, at the same time

- E-Posteriors and the AV CIs they induce rely on a prior, just like Bayesian posteriors…

  …but they **remain valid** irrespective of prior you use

with a bad prior, the e-posterior gets wide rather than wrong

Main Interpretation: Betting

- 1-to-1 correspondence between **testing with e-values** and **betting in a casino**
- product of e-values can be interpreted as amount of money you made so far in a game in which, at each time n, you don't expect to gain any money if H_0 is true, and you re-invest all your earnings so far

Evidence against null ⇔ getting rich

- Different betting strategies ⇔ different e-variables
- Multiply e-values ⇔ reinvest **all** your money
- Anytime validity ⇔in real casino, **you don't expect to get rich - no matter what is your rule for stopping and going home**

Optimal E-Values

- Optimal e-values are those **that make you rich as fast as possible** if the null hypothesis is wrong
- This has been called **growth rate optimality:** use e-values such that $\mathbf{E}[\log S(X_1, \ldots, X_n)]$ is large under alternative
    - good reasons for taking logarithm…
    - (log) **growth rate** replaces **power**
    - related to minimum cross-entropy, data compression


Vested
p-interests
Better
use e -values!