Anytime-valid confidence sequences

Alexander Ly[1]

Machine Learning Group

Centrum Wiskunde & Informatica

18 July 2023

[1]Supported by the Dutch Research Council (NWO): VI.Veni.211G.040

Take home message

(i) 95% Anytime-valid confidence sequences cover the true value with at least 95% chance at any moment in time
   - 95% Classical confidence/(Bayesian) credible intervals do not, see here + practical.

(ii) AV Confidence sequences invert safe tests/e-variables.
   - "Tests and intervals are the same thing"

(iii) In the `safestats` package:
   - Safe z-tests
   - Safe t-tests
   - Safe 2x2 tables
   - Safe logrank tests
   - More safe tests coming to your home soon

Overview

1. Anytime-valid

2. E-variables inversions $\Rightarrow$ anytime-valid confidence sequence

Section 1

Anytime-valid

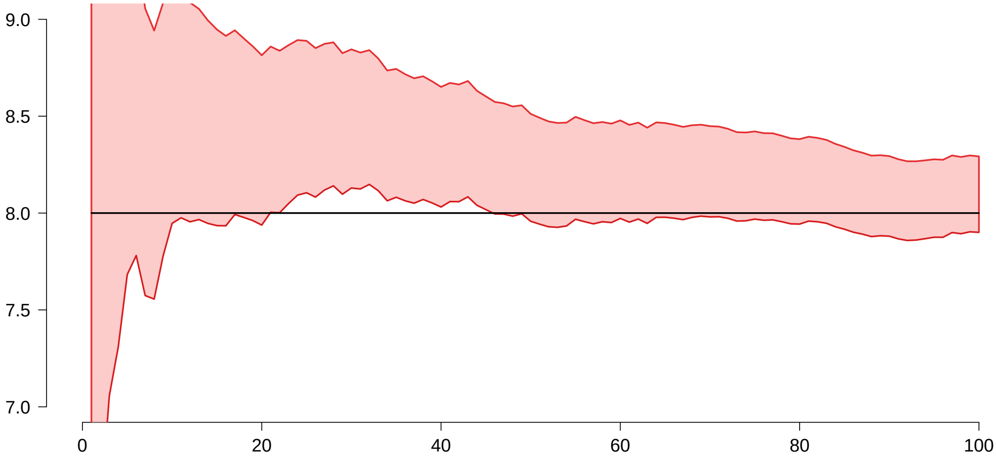
Example: Normal mean

Set-up

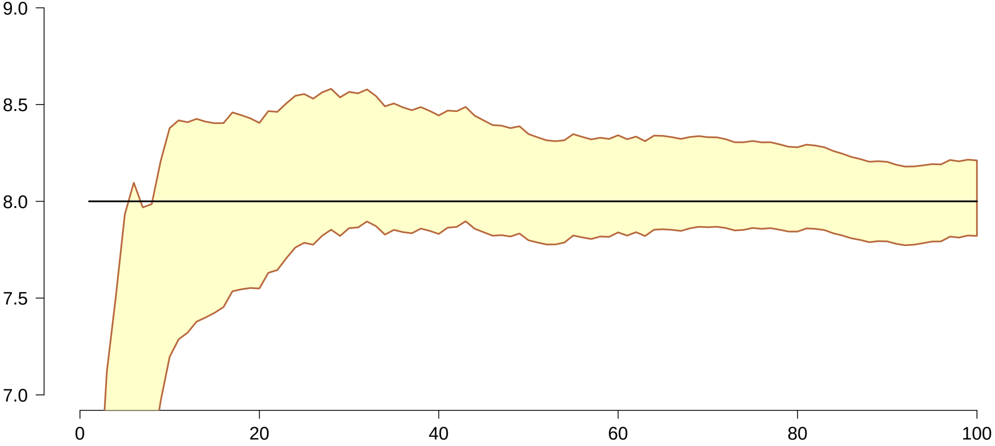
$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1) \tag{1}$$

- Goal 1: Cover μ^* with an interval, where μ^* can be any true value with 95% chance
  - Special case: $\mu^* = 8$
- Goal 2: At every n with 95% chance

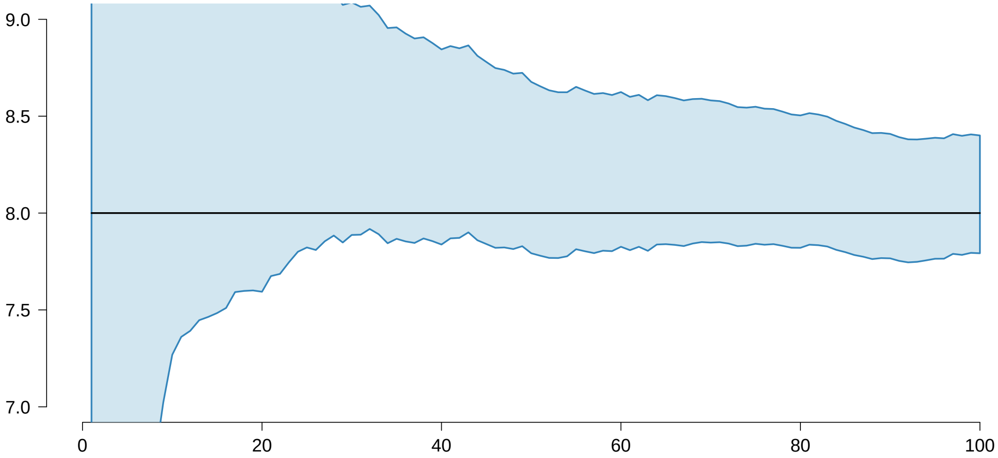
Example: Classical confidence intervals



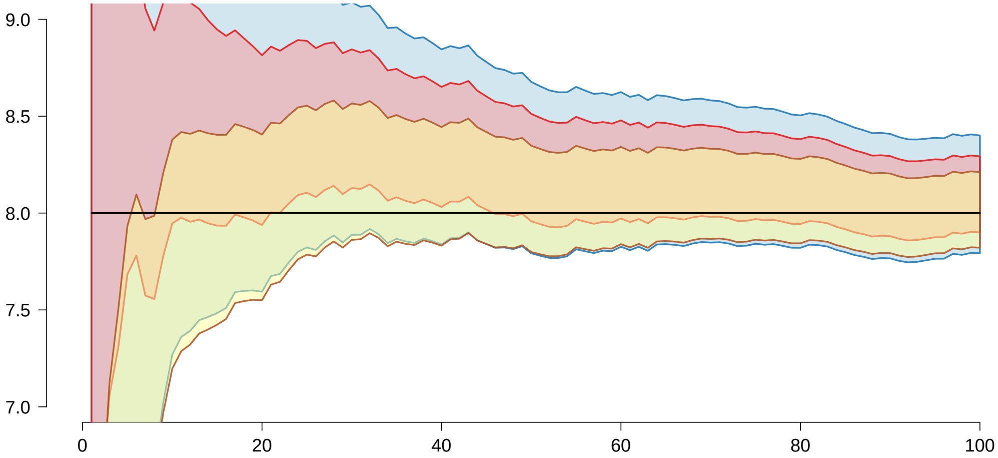
Example: Bayesian credible intervals



Example: Anytime-valid confidence sequence



Example: All the intervals



Tests not enough?

CIs at any moment in time?

Motivation: Sample size vs effect size

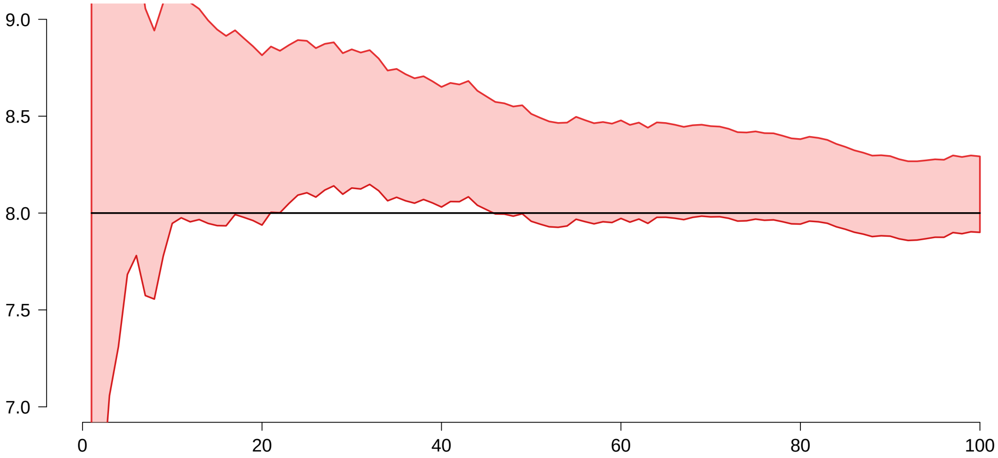
E-value: Does the effect exist?

- Questions:
  - Enough evidence against $\mathcal{H}_0 : \delta = \delta_0$?
  - When to stop your experiment? How many samples needed?
- Answers:
  - Stop (smallest) n with $E > 1/\alpha$, but you can go on if you'd like to

Confidence sequences: Likely values of the effect?

- Question:
  - What are the viable values of δ?
- Answer:
  - Any δ within the interval

Example: Classical confidence intervals



Section 2

E-variables inversions $\Rightarrow$ anytime-valid confidence sequence

Anytime-valid confidence sequences are inversions of E-variables/safe tests

E-variables and Ville's inequality

(i) A non-negative random variable $S^n := (S_1, \ldots, S_{n-1}, S_n)$ is called a conditional *E*-variable wrt to a null model and its past, if

$$\text{For each } P \in \mathcal{M}_0 : \mathbb{E}_P[S_n \mid X^{n-1}] \leq 1 \qquad (2)$$

At each small increment wrt its past data $X^{n-1} = (X_1, \ldots, X_{n-1})$, *expect* a small value if the null is true.

(ii) Direct consequence: Ville's inequality

$$\text{For each } P \in \mathcal{M}_0 : P(\text{ There exists an } n : S^n > 1/\alpha) < \alpha \qquad (3)$$

(iii) Example:

- Any postulated real number μ_0 st $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, 1)$

Inverting Ville's inequality/the E-variable

(i) Ville's inequality

$$\text{For each } P \in \mathcal{M}_0 : P(\text{ There exists an } n : S^n > 1/\alpha) < \alpha \tag{4}$$

(ii) Inversion

$$\text{For each } P \in \mathcal{M}_0 : P(\text{ For all } n \text{ simultaneously} : S^n \leq 1/\alpha) \geq 1 - \alpha \tag{5}$$

(iii) Example:

- Let $P = P_{\mu_0} \in \mathcal{M}_0$ and $S^n = S^n(\mu_0)$, e.g.
  - $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, 1)$, $S^n = f(Z)$, $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$
- If μ_0 is true, then for that μ_0, high chance $S^n(\mu_0) \leq 1/\alpha$.
- Collect all those μ_0 for which $S^n(\mu_0) \leq 1/\alpha$

Explicit inverses for z-test

$$\text{For each } P \in \mathcal{M}_0 : P(\text{ For all } n \text{ simultaneously} : S^n \leq 1/\alpha) \geq 1 - \alpha \tag{6}$$

(i) Z statistic depends on the null hypothesis

$$Z(\mu_0) = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \tag{7}$$

(ii) E-variable, Gaussian mixture $\mathcal{N}(0, g)$, fixed g

$$S^n(z, n) = (1 + ng)^{-\frac{1}{2}} \exp\left(\frac{ngz^2}{2(1 + ng)}\right). \tag{8}$$

(iii) Solve μ for $S^n(z, n) \leq 1/\alpha$, with $\sigma = 1$ we get

$$CS(1 - \alpha) = \left[\bar{X} - \sigma\sqrt{\frac{1 + ng}{n^2 g} \log\left(\frac{1 + ng}{\alpha^2}\right)}, \bar{X} + \sigma\sqrt{\frac{1 + ng}{n^2 g} \log\left(\frac{1 + ng}{\alpha^2}\right)}\right]. \tag{9}$$

Compare to frequentist

(i) AV confidence sequence: Width about $\frac{\sigma}{\sqrt{n}}\sqrt{\log(C+n)}$

$$\text{For each } P \in \mathcal{M}_0 : P(\text{ For all } n \text{ simultaneously} : S^n \leq 1/\alpha) \geq 1-\alpha \qquad (10)$$

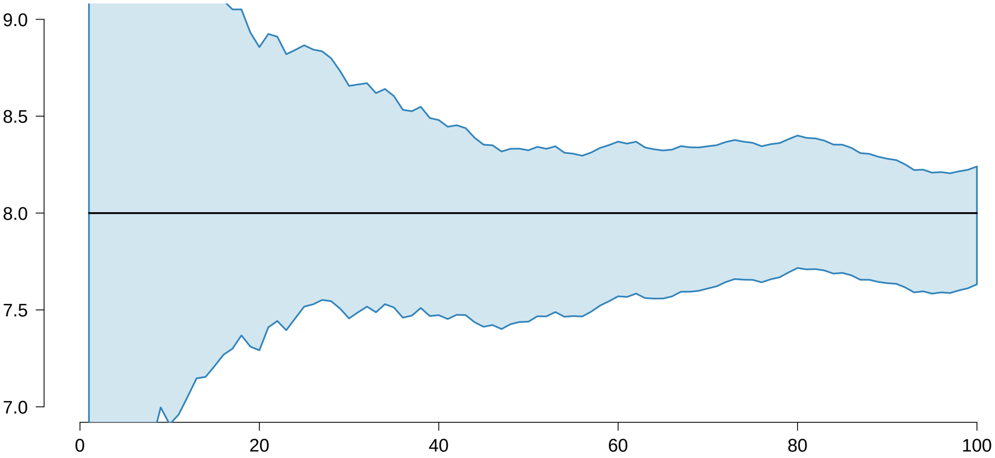
(ii) Frequentist: Width $\frac{\sigma}{\sqrt{n}}$, e.g., $[\bar{x} \pm \frac{1.96}{\sqrt{n}}]$

$$\text{At each } n \text{ separately, and for each } P \in \mathcal{M}_0 : P(|T_n| \leq q_\alpha) = 1-\alpha \qquad (11)$$

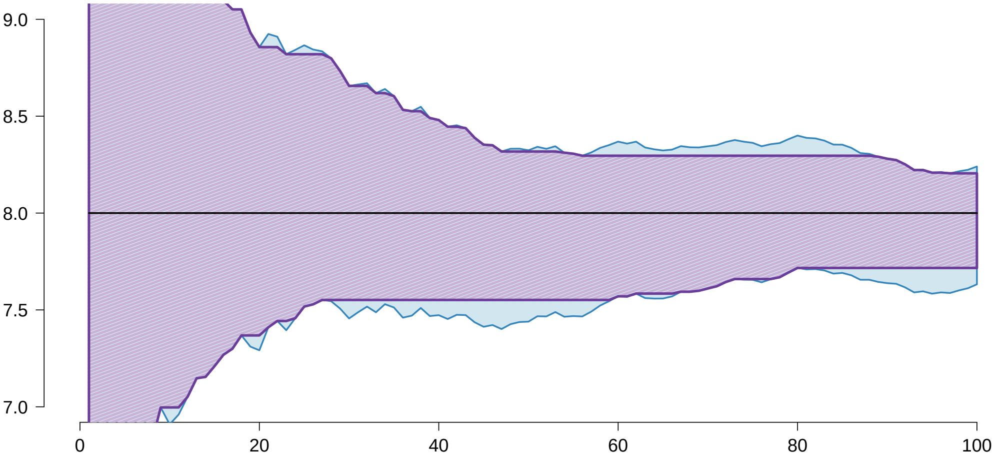
(iii) See R tutorial + play with the code

(iv) Clarification on 95% chance anytime-valid. If we sample 1,000 experiments, each with a confidence sequence until ∞, then of these 1,000 experiments about 50 will have at least one time point that the interval does not cover the true mean. Hence, 950 of these 1,000 confidence sequence will capture the true mean all the time.

Bonus: Running intersection



Bonus: Running intersection



Take home message

(i) 95% Anytime-valid confidence sequences cover the true value with at least 95% chance at any moment in time
   - 95% Classical confidence/(Bayesian) credible intervals do not, see here + practical.

(ii) AV Confidence sequences invert a safe test/e-variable.
   - "Tests and intervals are the same thing"

(iii) In the `safestats` package:
   - Safe z-tests
   - Safe t-tests
   - Safe 2x2 tables
   - Safe logrank tests
   - More safe tests coming to your home soon