Likelihood, Replicability and Robbins' Confidence Sequences

Luigi Pace[*]
Department of Economics and Statistics, University of Udine, Italy
and
Alessandra Salvan
Department of Statistical Sciences, University of Padova, Italy

Abstract

The widely denounced replicability crisis in science hints at revised standards of significance. The customary frequentist confidence intervals, calibrated through hypothetical repetitions of the experiment that is supposed to have produced the data at hand, have a feeble concept of replicability. In particular, contrasting conclusions may be reached when a substantial enlargement of the study is undertaken. To redefine statistical confidence in such a way that inferential conclusions are compatible, with large enough probability, under enlargements of the sample, we give a new reading of a proposal dating back to the 60's, Robbins' confidence sequences. Directly bounding the probability of reaching, in the future, conclusions incompatible with the current ones, Robbins' confidence sequences ensure a clear-cut form of replicability when inference is performed on accumulating data. Moreover, we show that they can be justified under various views of inference. They are likelihood based, can incorporate prior information, obey the strong likelihood principle. They are easily computable, even when inference is on a parameter of interest. Finally, their main frequentist property is easy both to understand and to prove.

Keywords: Bayes factor; Confidence region; Laplace expansion; Profile likelihood; Revision of standards; Statistical evidence.

[*] This research was supported by a grant from a Project of National Interest, Italian Ministry of Education, PRIN 2015 – 2015EASZFS 003.

1 Introduction

Announcing a result is a hazard when the supporting evidence is statistical in nature. In the long run, scientific credibility is jeopardized if discoveries are claimed or implied (or understood) to be more firmly established than they will eventually prove to be. The issue has become pressing, especially after the denounce of a replicability crisis in science by Ioannidis (2005) and many others on its wake, triggering the recent ASA statement, Wasserstein and Lazar (2016). To reduce failure to replicate, more strict evidential thresholds are propounded in the literature, see Johnson (2013), possibly variable by discipline (Goodman, 2016). Benjamin et al. (2017) advocate changing the standard threshold for significance from 0.05 to 0.005, while Lakens et al. (2017) recommend a case by case transparently justified choice, better if pre-registered. A general warning against the dichotomization of evidence is given in McShane and Gal (2017).

When interest lies in reporting effect sizes and related confidence intervals (see e.g. Nakagawa and Cuthill, 2007), a revision of standards for statistical significance would entail a parallel revision of standards for confidence levels, say from 0.95 to 0.995. These higher levels are not, however, linked to some formal replicability requirement. In this note we bring to the fore a simple but apparently new concept of replicability for inference based on confidence regions and explore its relations with a proposal dating back to the 60s, Robbins' *confidence sequences* (Robbins, 1970; see also Darling and Robbins, 1967a,b).

We interpret here replicability as assurance that compatible conclusions are reached when information increases, i.e. the sample is enlarged. To be specific, inferential conclusions from confidence regions for the same parameter are compatible if these regions overlap, incompatible if their intersection is empty. Compatible confidence regions are also said to be non-contradictory.

As a form of replicability, non-contradiction is especially compelling in experimental sciences when inference is performed on accumulating data. Early conclusions are susceptible to be falsified within the matter of years or months, and sometimes even earlier. When the true state of nature, or a much more reliable representation of it, becomes eventually available, reputational penalty ensuing from hasty wrong conclusions could be large. This denouement does not happen in hard sciences alone. Think for instance of estimating the result of an election from early reporting counting areas, where the estimate is made only hours before a winner is declared. Other contexts where coherence under sample enlargement seems to be cogent are long-term epidemiological studies and drugs surveillance.

We show that the use of confidence sets ensuing from Robbins' approach produces compatible confidence regions with large enough probability, at least in the idealized situation of i.i.d. sampling from a correctly specified parametric model. Robbins' papers are highly technical and reasearch on confidence sequences seems to have been neglected after the equally technical contributions Lai (1976) and Csenki (1979). We try here to give an accessible account and to highlight the bearing of Robbins' confidence sequences on principles of statistical inference.

Fixed level confidence regions, even with a higher revised level, fail to fulfill the compatibility requirement. Hence, they correspond to a feeble sense of replicability. As a simple example, consider i.i.d. sampling from a normal distribution with known variance σ_0^2. Let $\bar{Y}_n = \sum_{i=1}^n Y_i/n$ be the sample mean. Then

$$\bar{Y}_{n+m} - \bar{Y}_n \sim N\left(0, \sigma_0^2\left(\frac{1}{n} - \frac{1}{n+m}\right)\right)$$

and the probability that $(1-\alpha)$-level confidence intervals for the mean at sample sizes n and $n+m$ do not overlap is

$$\begin{aligned} &2P_\mu\left(\bar{Y}_n + \frac{\sigma_0}{\sqrt{n}} z_{1-\alpha/2} < \bar{Y}_{n+m} - \frac{\sigma_0}{\sqrt{n+m}} z_{1-\alpha/2}\right) \\ &= 2\Phi\left(-z_{1-\alpha/2}\left(\sqrt{1+\frac{n}{m}} + \sqrt{\frac{n}{m}}\right)\right) > 0\,. \end{aligned}$$

Therefore the probability is 1 of finding a couple of disjoint intervals, i.e. of observing a sequence of samples that gives rise to incompatible $(1-\alpha)$-level confidence intervals. When the realistically attainable sample size is very large but finite, though the usual confidence intervals shrink towards the true value of the parameter as the sample size increases, conflicting conclusions may be reported at various stages of the data acquisition process, with a probability that may be close to 1.

A side advantage of Robbins' confidence sequences is their justification under various views of inference. They are likelihood based, can incorporate prior information, have frequentist properties, have Bayesian properties under a proper prior, obey the strong likelihood principle. On top of that, Robbins' confidence sequences have great pedagogical benefits. They need virtually no sample space calculations (Fisherian distribution problems disappear) and require a fairly limited amount of parameter space calculations, the expectation of a scalar function, easily performed by simulation.

The outline of the paper is as follows. A new reading of Robbins' confidence sequences is given in Section 2. Their inferential properties are summarized in Section 3, with technical details provided in the Appendix. Section 4 presents two examples and illustrates the replicability properties of Robbins' confidence sequences through simulation results dealing with normal mean and binomial probability. Section 5 concludes.

2 Non-contradiction and Robbins' confidence sequences

Let us consider the highly idealized and simplified situation of a statistician who is potentially able to obtain any number n of i.i.d. observations $y^{(n)} = (y_1, \dots, y_n)$, realization of the random vector $Y^{(n)} = (Y_1, \dots, Y_n)$. Let P_θ denote the joint probability distribution of the sequence $Y^{(\infty)} = (Y_1, Y_2, \dots)$. We suppose that P_θ belongs to a statistical model with parameter space $\Theta \subseteq \mathbb{R}^p$. Let $p_n(y^{(n)}; \theta)$ denote the density of $Y^{(n)}$ under P_θ. Assume that, for every given n, all these densities are strictly positive on the same support, i.e., the support does not depend on θ.

A confidence region, based on $y^{(n)}$ and constructed according to a certain rule, is a subset of Θ denoted by $\hat{\Theta}_n = \hat{\Theta}(y^{(n)})$. A confidence sequence is a sequence of confidence regions. To avoid triviality, we consider only confidence sequences that are consistent, i.e. such that $\lim_{n\to\infty} P_\theta(\theta' \in \hat{\Theta}_n) = 0$ for every $\theta' \neq \theta$, where $\theta, \theta' \in \Theta$. Consistency implies that, for $\theta' \neq \theta$,

$$P_\theta\left(\theta' \in \cap_{n\geq 1}\hat{\Theta}_n\right) \leq \lim_{n\to\infty} P_\theta(\theta' \in \hat{\Theta}_n) = 0\,.$$

We will say that a confidence sequence is non-contradictory if no $\hat{\Theta}_n$ is contradicted by a $\hat{\Theta}_{n+m}$, for some $m > 0$. Contradiction happens when, for an $m > 0$, $\hat{\Theta}_n \cap \hat{\Theta}_{n+m} = \emptyset$. When a confidence sequence is non-contradictory, there are conclusions that are common to all confidence statements, i.e., $\cap_{n\geq 1}\hat{\Theta}_n \neq \emptyset$. Consistency ensures that non-contradictory sequences shrink towards the true parameter value. Indeed, for consistent confidence sequences, only the true θ may belong to $\cap_{n\geq 1}\hat{\Theta}_n$, even though with probability strictly less than 1.

Since $\cap_{n\geq 1}\hat{\Theta}_n = \emptyset$ implies $\theta \notin \cap_{n\geq 1}\hat{\Theta}_n$, we have

$$P_\theta\left(\cap_{n\geq 1}\hat{\Theta}_n = \emptyset\right) \leq P_\theta\left(\theta \notin \cap_{n\geq 1}\hat{\Theta}_n\right) = 1 - P_\theta\left(\theta \in \hat{\Theta}_n \ \text{ for } \ \text{every } \ n \geq 1\right) .$$

It follows that, if, for $0 < \varepsilon < 1$,

$$P_\theta\left(\theta \in \hat{\Theta}_n \ \text{ for } \ \text{every } \ n \geq 1\right) \geq 1 - \varepsilon\,, \tag{1}$$

then

$$P_\theta\left(\cap_{n\geq 1}\hat{\Theta}_n = \emptyset\right) \leq \varepsilon\,,$$

so that the probability of contradiction as evidence accumulates is held in check.

Confidence sequences satisfying (1) are obtained in Robbins (1970, see formula (3)). A heuristic argument for their consistency is outlined in the Appendix. Robbins' regions, denoted by $\hat{\Theta}_{1-\varepsilon}(Y^{(n)})$, with realization $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$, have the form

$$\hat{\Theta}_{1-\varepsilon}(y^{(n)}) = \left\{\theta \in \Theta \,:\, p_n(y^{(n)};\theta) > \varepsilon q_n(y^{(n)})\right\}\,, \tag{2}$$

where $q_n(y^{(n)})$ is the averaged, or marginal, density

$$q_n(y^{(n)}) = \int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta\,.$$

Above, the weight function $\pi(\theta)$ is a preset probability density over Θ with $\pi(\theta) > 0$ for every $\theta \in \Theta$. The value $1-\varepsilon$ will be called here the assured persistence level, or simply the persistence level, of the confidence sequence.

To illustrate the pedagogical virtues of the approach, the proof in Robbins (1970) that the sequence of regions $\hat{\Theta}_{1-\varepsilon}(Y^{(n)})$ satisfies (1) is sketched in the Appendix. The key argument is an inequality giving a bound on the probability of reaching strongly misleading evidence from the likelihood ratio statistic (Royall, 1997, page 7).

Robbins' confidence sequences are likelihood-based. Specifically, $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ is the region of θ values whose likelihood $L(\theta;y^{(n)}) = p_n(y^{(n)};\theta)$ is large, in particular larger than a preset fraction of the integrated likelihood $q_n(y^{(n)})$. Therefore, regions (2) are invariant under one-to-one transformations of y and one-to-one transformations of θ. The computation of $q_n(y^{(n)})$ incorporates prior information, notional or real. In either way, the importance of the choice of the prior is downplayed because property (1) holds for every preset specification of $\pi(\theta)$, provided $\pi(\theta)$ does not place probability 0 on a plausible region of Θ. Examples of the effects of different choices of the prior are given in Section 4. Confidence regions (2) are nested, i.e., $\hat{\Theta}_{1-\varepsilon'}(y^{(n)}) \subseteq \hat{\Theta}_{1-\varepsilon}(y^{(n)})$, when $1-\varepsilon' < 1-\varepsilon$. The maximum likelihood estimate $\hat{\theta}_n$ is always in $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$, being $p_n(y^{(n)}; \hat{\theta}_n) \geq q_n(y^{(n)})$.

When the parameter is partitioned as $\theta = (\psi, \lambda)$, where $\psi \in \Psi$ is a p_0-dimensional component of interest and λ is nuisance, in some cases inference on ψ can be based on a statistic $t^{(n)} = t(y^n)$ producing a marginal or conditional model free of λ. In these cases, $p_n(t^{(n)}; \psi)$ or $p_n(y^{(n)}|t^{(n)}; \psi)$ may replace $p_n(y^{(n)}; \theta)$ in (2) and $q(y^{(n)})$ is redefined accordingly. However, the implementation of Robbins' confidence sequences for a parameter of interest is also feasible when a reduction by marginalization or conditioning is not available, and does not require sample space calculations. The confidence sequence for ψ is the projection of $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ on Ψ,

$$\hat{\Psi}_{1-\varepsilon}(y^{(n)}) = \left\{ \psi \in \Psi \; : \; (\psi, \lambda) \in \hat{\Theta}_{1-\varepsilon}(y^{(n)}) \text{ for some } \lambda \right\} . \tag{3}$$

The confidence sequence (3) turns out to be based on the profile likelihood:

$$\hat{\Psi}_{1-\varepsilon}(y^{(n)}) = \left\{ \psi \in \Psi \; : \; p_n(y^{(n)}; \psi, \hat{\lambda}_\psi) > \varepsilon q_n(y^{(n)}) \right\} ,$$

where $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ in the model for $y^{(n)}$ with ψ fixed and

$$q_n(y^{(n)}) = \int_\Theta p_n(y^{(n)}; \theta) \pi(\theta) \, d\theta \, ,$$

as above, independently of $\psi = \psi(\theta)$.

3 Frequentist, pure likelihood and Bayesian properties

Robbins' confidence sequences have frequentist properties from their very inception. The usual asymptotics where

$$2 \left(\ell(\hat{\theta}_n; Y^{(n)}) - \ell(\theta; Y^{(n)}) \right) \xrightarrow{d} \chi^2_p$$

entails that the marginal asymptotic coverage of $\hat{\Theta}_{1-\varepsilon}(Y^{(n)})$ is one,

$$\lim_{n\to\infty} P_\theta(\theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(n)})) = 1\,. \tag{4}$$

Details are in the Appendix. This behaviour contrasts greatly with what is usually sought for in conventional frequentist inference, i.e., asymptotic coverage equal to the nominal level $1-\alpha$. Under this respect, a frequentist statistician willing to ensure her confidence regions to be non-contradictory with positive probability seems to have to pay a price in terms of overcoverage for fixed n.

A revival of Robbins' confidence sequences entails a novel concept of confidence, involving the current size n experiment and its future, hypothetical or not, enlargements. A persistence level $1-\varepsilon$ has the frequentist assurance that

$$P_\theta\left(\theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(m)}) \ \text{ for every } \ m \geq n\right) \geq 1-\varepsilon\,,$$

so that

$$P_\theta\left(\cap_{m\geq n}\hat{\Theta}_{1-\varepsilon}(Y^{(m)}) \neq \emptyset\right) \geq 1-\varepsilon\,.$$

In practice, we have high confidence that no contradiction with $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ would occur with larger sample sizes, even in settings where the sample enlargement is only hypothetical.

It is important to stress that what happened for sample sizes from 1 to $n-1$ does not matter. Moreover, although the sequence $\cap_{j\leq n}\hat{\Theta}_{1-\varepsilon}(Y^{(j)})$ satisfies (1) as well, it is not eligible as a sensible confidence sequence because $\cap_{j\leq n}\hat{\Theta}_{1-\varepsilon}(y^{(j)})$ could be empty, and therefore not consistent.

There is another way to express the frequentist assurance coming with $\hat{\Theta}_{1-\varepsilon}(Y^{(n)})$. From

$$\begin{aligned} P_\theta\left(\theta \in \cap_{m\geq n}\hat{\Theta}_{1-\varepsilon}(Y^{(m)})\right) \quad &= \quad P_\theta\left(\theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(n)})\right) \\ &\qquad P_\theta\left(\theta \in \cap_{m>n}\hat{\Theta}_{1-\varepsilon}(Y^{(m)}) \mid \theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(n)})\right) \end{aligned}$$

we get

$$P_\theta\left(\theta \in \cap_{m>n}\hat{\Theta}_{1-\varepsilon}(Y^{(m)}) \mid \theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(n)})\right) = \frac{P_\theta\left(\theta \in \cap_{m\geq n}\hat{\Theta}_{1-\varepsilon}(Y^{(m)}\right)}{P_\theta\left(\theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(n)})\right)} \geq 1-\varepsilon\,.$$

Therefore, if $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ covers the truth (an easily conceded premise if n is large enough, in view of (4)), then, with probability at least $1-\varepsilon$, no contradiction with $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ will be seen in future enlargements of the study.

Moreover, the reward for overcoverage in a fixed n perspective is that Robbins' confidence sequence (2) offers inference that rarely fails to reproduce even in a multiple investigation perspective. Let the sequences $Y^{(n)}$ and $Y^{*(n')}$ be independent with the same statistical model $\{P_\theta,\ \theta \in \Theta \subseteq \mathbb{R}^p\}$ and the same true parameter value. Statistician A will observe the initial part of the sequence $Y^{(n)}$, statistician B will observe the initial part of the sequence $Y^{*(n')}$. If both adopt and communicate publicly Robbins' confidence regions with the same ε, though with possibly different preset weight functions, they will be usually found in agreement, because

$$\begin{aligned}&P_\theta\left(\hat{\Theta}_{1-\varepsilon}(Y^{(n)}) \cap \hat{\Theta}_{1-\varepsilon}(Y^{*(n')}) \neq \emptyset\right)\\ &\geq P_\theta\left(\theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(n)}) \cap \hat{\Theta}_{1-\varepsilon}(Y^{*(n')}) \quad \text{for every} \quad n, n' \geq 1\right) \geq (1-\varepsilon)^2 .\end{aligned}$$

In fact, conventional inference, both Bayesian and frequentist, is contingent on the current sample or the generating mechanism of the current sample. Inference from Robbins' confidence sequences is in a sense absolute, it leads to conclusions that with reasonably high probability can withstand any further scrutiny under the same data generating model. From a practical stance, as traditional confidence regions give at best what they promise, the same is true of Robbins' confidence sequences. Both are vulnerable to model misspecification.

Regions $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$, depending on the data only through the likelihood function, agree with the strong likelihood principle. In particular, they obey both the sufficiency and the conditionality principles. For sufficiency, let $s^{(n)} = s(y^{(n)})$ be a sufficient statistic for $p_n(y^{(n)};\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$, so that

$$p_n(y^{(n)};\theta) = p_{S^{(n)}}(s^{(n)};\theta)\, p_n(y^{(n)}|s^{(n)}),$$

with $p_{S^{(n)}}(s^{(n)};\theta)$ the marginal density of $S^{(n)} = s(Y^{(n)})$ and $p_n(y^{(n)}|s^{(n)})$ the conditional density of $Y^{(n)}$ given $S^{(n)} = s^{(n)}$. Then, $\hat{\Theta}_{1-\varepsilon}(y^{(n)}) = \hat{\Theta}_{1-\varepsilon}(s^{(n)})$, as is easy to see. As to the conditionality, let $a^{(n)} = a(y^{(n)})$ be a distribution constant statistic, so that

$$p_n(y^{(n)};\theta) = p_{A^{(n)}}(a^{(n)})\, p_n(y^{(n)}|a^{(n)};\theta).$$

Then

$$\hat{\Theta}_{1-\varepsilon}(y^{(n)}) = \left\{\theta \in \Theta \,:\, p_n(y^{(n)}|a^{(n)};\theta) \geq \varepsilon \int_\Theta p_n(y^{(n)}|a^{(n)};\theta)\pi(\theta)d\theta\right\} = \hat{\Theta}_{1-\varepsilon}(y^{(n)}|a^{(n)})\,,$$

so that regions $\hat{\Theta}_{1-\varepsilon}(Y^{(n)})$ have probability of contradiction bounded by ε also conditionally on $a^{(n)}$.

Let us consider now Bayesian properties of the confidence sequence (2). The ratio

$$\frac{q_n(y^{(n)})}{p_n(y^{(n)};\theta)} = \frac{\int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta}{p_n(y^{(n)};\theta)} \tag{5}$$

is a Bayes factor, Kass and Raftery (1995). In fact, Kass and Raftery (1995, Section 3.2) suggest quantitative standards for interpretation of (5) as evidence against the hypothesis θ. For instance, decisive evidence requires (5) greater than 100. This corresponds $\varepsilon = 0.01$ in (2) and hence a persistence level equal to 0.99.

When $\pi(\theta)$ represents a prior distribution, with data $y^{(n)}$, the posterior is

$$\pi(\theta|y^{(n)}) = \frac{p_n(y^{(n)};\theta)\pi(\theta)}{\int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta}\,.$$

Definition (2) may be recast as

$$\hat{\Theta}_{1-\varepsilon}(y^{(n)}) = \left\{\theta \in \Theta\,:\,\pi(\theta|y^{(n)}) > \varepsilon\pi(\theta)\right\}\,. \tag{6}$$

The complementary set $\bar{\Theta}_{1-\varepsilon}(y^{(n)}) = \Theta \setminus \hat{\Theta}_{1-\varepsilon}(y^{(n)})$ has posterior probability

$$\int_{\bar{\Theta}_{1-\varepsilon}(y^{(n)})} \pi(\theta|y^{(n)})\,d\theta \le \varepsilon \int_{\bar{\Theta}_{1-\varepsilon}(y^{(n)})} \pi(\theta)\,d\theta \le \varepsilon\,.$$

Therefore, $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ has posterior probability at least $1-\varepsilon$.

Credible regions $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ have bounded probability of being contradictory even in a Bayesian sense. Indeed, let P be the joint probability model of θ and $Y^{(\infty)}$, where θ has marginal density $\pi(\theta)$ and, given θ, $Y^{(\infty)}$ has conditional distribution P_θ. In this setting, (1) is a conditional probability statement. With $\hat{\Theta}_n = \hat{\Theta}_{1-\varepsilon}(Y^{(n)})$ it implies that, marginally,

$$P\left(\theta \in \bigcap_{n\ge 1} \hat{\Theta}_{1-\varepsilon}(Y^{(n)})\right) \ge 1-\varepsilon\,.$$

Representation (6) shows that inference from Robbins' confidence sequences proceeds by subtraction, eliminating the most implausible values of θ. Viewing the prior as a posterior for an empty dataset, $\pi(\theta) = \pi(\theta|\emptyset)$, from (6) we conclude formally that $\hat{\Theta}_{1-\epsilon}(\emptyset) = \Theta$, which makes sense.

4 Examples

The implementation of Robbins' confidence sequences requires the specification of $\pi(\theta)$ and the choice of ε or a range of ε values. These issues are sketched through two simple examples.

Example 1. Normal population with known variance.
Suppose that Y_i, $i = 1, 2, \ldots$, are i.i.d. $N(\theta, \sigma_0^2)$, with unknown mean θ and known variance σ^2. Reduction by sufficiency produces the sequence of sample means $\bar{Y}_n = \sum_{i=1}^n Y_i/n$ with model $N(\theta, \sigma_0^2/n)$, $n = 1, 2, \ldots$. The density of $\bar{Y}_n$ given θ is

$$p_n(\bar{y}_n; \theta) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{n(\bar{y}_n - \theta)^2}{2\sigma_0^2} \right\} .$$

Let us consider as the weight function a conjugate prior $N(\mu_0, \tau_0^2)$ density, i.e.

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\left\{ -\frac{(\theta - \mu_0)^2}{2\tau_0^2} \right\} .$$

With this $\pi(\theta)$, the marginal distribution of $\bar{Y}_n$ is $N(\mu_0, \tau_0^2 + \sigma_0^2/n)$, so that

$$q_n(\bar{y}_n) = \frac{1}{\sqrt{2\pi}\sqrt{\tau_0^2 + \dfrac{\sigma_0^2}{n}}} \exp\left\{ -\frac{1}{2} \frac{(\bar{y}_n - \mu_0)^2}{\tau_0^2 + \dfrac{\sigma_0^2}{n}} \right\} .$$

After some algebra, Robbins' confidence sequence $\hat{\Theta}_{1-\varepsilon}(\bar{y}_n) = \{\theta \in \mathrm{I\!R} \, : \, p_n(\bar{y}_n; \theta) \geq \varepsilon q_n(\bar{y}_n)\}$ is seen to consist of the intervals $\bar{y}_n \pm d_n$, where

$$d_n = \frac{\sigma_0}{\sqrt{n}} \sqrt{\log n + \log \frac{\tau_0^2 + \sigma_0^2/n}{\sigma_0^2} + \frac{(\bar{y}_n - \mu_0)^2}{\tau_0^2 + \sigma_0^2/n} - 2 \log \varepsilon} \, .$$

A simulation study has been performed in order to shed some light on the practical bearing of using Robbins' confidence sequences vis-à-vis the customary confidence intervals. A sequence of confidence intervals $(\underline{\theta}_n, \bar{\theta}_n)$ shows a contradiction in the range $n_{min} \leq n \leq n_{max}$ whenever $\max(\underline{\theta}_n) > \min(\bar{\theta}_n)$, where min and max are over the n values of interest. Analogously, the sequence shows non-coverage of θ at some n in the range $n_{min} \leq n \leq n_{max}$ whenever, over the n values considered, $\max(\underline{\theta}_n) > \theta$ or $\min(\bar{\theta}_n) < \theta$. Contradictions and non-coverages have been monitored for 1,000 replications of enlarging samples of size n with $n_{min} = 1$ and $n_{max} = 40,000$.

Table 1: Normal population with known variance: empirical percentages of contradictions and non-coverages at some n for intervals for the mean with confidence level $1-\alpha$ in 1,000 sequences of samples with size from 1 to 40,000.

$100(1-\alpha)$	90	95	99	99.5
contradictions	76.9	51.1	12.8	7.4
non-coverages	90.3	68.9	26.0	15.7

In Table 1 the results for confidence intervals $\bar{y}_n \pm \sigma_0 z_{1-\alpha/2}/\sqrt{n}$, with confidence level $1-\alpha = 0.90, 0.95, 0.99, 0.995$, are shown. Contradictions and non-coverages are dominant for the levels 90% and 95%. They are both comparatively uncommon for the level 99.5%, but their relative frequency could be made as close to 1 as desired by letting n_{max} large enough. In the same scenario, all non-coverages become contradictions. The simulation has been performed by sampling standard normal deviates, the results, however, do not depend on the true value of the parameters of the normal population.

Table 2 displays the results for Robbins' confidence sequences with persistence levels $1-\varepsilon = 0.50, 0.80, 0.90, 0.95$ and various prior distributions on θ. When the prior is concentrated around the true θ, contradictions and non-coverages are comparatively abundant, but their relative frequency remains under the upper bound probability ε. When the prior is discrepant from the likelihood, that is μ_0 is far from θ, the conflict between the two is resolved in favour of the likelihood, through wider confidence regions. As is seen, this counterbalance increases conservativeness of the ε bound. Apart from these cases, the results in terms of observed contradictions and non-coverages for some n in the range 1–40,000 when $1-\varepsilon = 0.80$ are qualitatively comparable with those for the customary intervals with confidence level 0.995.

Example 2: Bernoulli population.

Suppose that Y_i, $i = 1, 2, \ldots$ are i.i.d. Bernoulli $Bi(1, \theta)$, with unknown mean $\theta \in (0, 1)$.

Table 2: Normal population with known variance 1: empirical percentages of contradictions and non-coverages at some n for Robbins' confidence sequences for the mean with persistence level $1-\varepsilon$ in 1,000 sequences of samples with size from 1 to 40,000 and various normal priors on the mean. The true value of the mean is 0.

prior	$100(1-\varepsilon)$	50	80	90	95
$\mu_0 = 0,\ \tau_0^2 = 0.1$	contradictions	29.9	8.3	4.0	1.2
	non-coverages	42.7	15.4	7.6	4.2
$\mu_0 = 0,\ \tau_0^2 = 1.0$	contradictions	26.0	9.4	4.5	1.8
	non-coverages	32.7	13.1	7.2	3.1
$\mu_0 = 0,\ \tau_0^2 = 10$	contradictions	13.0	5.0	2.4	1.3
	non-coverages	15.8	6.9	3.3	1.6
$\mu_0 = 1,\ \tau_0^2 = 1.0$	contradictions	21.5	7.0	3.2	1.6
	non-coverages	25.9	9.9	5.1	2.4
$\mu_0 = 2,\ \tau_0^2 = 1.0$	contradictions	12.6	4.3	2.2	1.2
	non-coverages	14.3	5.0	2.7	1.6
$\mu_0 = 5,\ \tau_0^2 = 1.0$	contradictions	0.6	0.4	0.1	0.0
	non-coverages	0.6	0.4	0.2	0.0

Reduction by sufficiency produces the sequence of sample sums $S_n = \sum_{i=1}^{n} Y_i$, whose model is $Bi(n, \theta)$. The density of S_n given θ is

$$p_n(s_n; \theta) = \binom{n}{s_n} \theta^{s_n} (1-\theta)^{n-s_n} .$$

Let us consider as a weight function the conjugate prior $Beta(\alpha, \beta)$ density

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} ,$$

where $\alpha, \beta > 0$ and $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. With this prior, the marginal distribution of S_n is beta-binomial, with density

$$q_n(s_n) = \binom{n}{s_n}\frac{B(s_n+\alpha, n-s_n+\beta)}{B(\alpha,\beta)} .$$

The choice $\alpha = \beta = 0.5$ is Jeffreys' prior. When $\alpha = \beta = 1$ the prior is a continuous uniform distribution on $[0, 1]$ and the marginal distribution of S_n is discrete uniform on $\{0, 1, \ldots , n\}$. Intervals that form the Robbins' confidence sequence

$$\hat{\Theta}_{1-\varepsilon}(\bar{y}_n) = \{\theta \in \mathrm{I\!R} \, : \, p_n(\bar{y}_n; \theta) \geq \varepsilon q_n(\bar{y}_n)\}$$

have not a closed-form expression but are easily computed numerically.

Intervals with asymptotic confidence level $1 - \alpha$ are obtained from the likelihood ratio statistic and have the form

$$\tilde{\Theta}_{1-\alpha}(y^{(n)}) = \left\{\theta \in (0, 1) \, : \, p_n(s_n; \theta) \geq p_n(s_n; \hat{\theta}_n) \exp\{-0.5\chi^2_{1,1-\alpha}\}\right\} ,$$

where $\hat{\theta}_n = s_n/n$ is the maximum likelihood estimate and $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-quantile of a chi-squared distribution with 1 degree of freedom. Even these intervals are easily computed numerically.

A small simulation study with various true θ values and various weight functions has been performed. In particular, contradictions and non-coverages for n in the range $n_{min} = 100$ and $n_{max} = 4,000$ have been enquired. The number of replications remains $1,000$.

Table 3 displays the results for the confidence intervals with confidence level $1 - \alpha = 0.90, 0.95, 0.99, 0.995$ obtained from the likelihood ratio statistic. Contradictions and non-coverages are important when $1 - \alpha = 0.90, 0.95$, even over such a restricted range of n values. The case $1 - \alpha = 0.995$ shows a marked improvement.

In table 4 results for Robbins' confidence sequences with persistence levels $1 - \varepsilon = 0.50$, 0.80, 0.90, 0.95 and various beta prior distributions on θ are shown. When the prior is centered at the true θ, non-coverages are comparatively abundant. Contradictions are rarely observed due to the limited range of n considered. As expected, conservativeness

Table 3: Bernoulli population: empirical percentages of contradictions and non-coverages at some n for likelihood ratio intervals for the mean with confidence level $1 - \alpha$ in 1,000 sequences of samples with size from 100 to 4,000.

	$100(1-\alpha)$	90	95	99	99.5
$\theta = 0.5$	contradictions	28.5	11.6	1.5	0.5
	non-coverages	65.7	41.0	13.5	8.6
$\theta = 0.7$	contradictions	30.4	14.2	1.2	0.2
	non-coverages	64.2	43.3	12.5	7.1
$\theta = 0.9$	contradictions	29.5	11.8	0.9	0.2
	non-coverages	65.0	42.6	12.1	7.1

increases as the prior moves away from the true parameter value. Again, the results when $1 - \varepsilon = 0.80$ are qualitatively comparable with those for the customary intervals with confidence level 0.995.

5 Conclusions

Herbert E. Robbins is mostly acknowledged in Statistics for his path-breaking introduction of empirical Bayes methods, stochastic approximation methods, and his contributions to sequential analysis (Lai and Siegmund, 1986). On the other hand, his proposal of confidence sequences has been largely neglected. One exception, at least in the statistical literature, is Gandy and Hahn (2016), where Robbins' confidence sequences provide a tool to keep in check stochastic simulations. Robbins' confidence sequences also inspired repeated confidence intervals (Jennison and Turnbull, 1989), where coverage of the true θ is required at a finite (typically small) number of interim analyses of a study. Notably, in the discussion of Jennison and Turnbull (1989), Whitehead (1989) points to situations, such as long-term epidemiological studies, where a fixed number of analyses "might become a barrier". Here, we have stressed the link between non-contradiction and coverage along the whole sequence as the basis for a novel interest in Robbins' confidence sequences. The price to pay for controlling for the probability of non-contradiction is that wider regions are needed. When data are scarce and difficult to obtain, this could appear as a serious drawback of Robbins' confidence sequences.

Robbins' confidence sequences offer durable inferences, satisfying coverage requirements along the whole sequence of samples. By contrast, inferences as those stemming from the usual statistical procedures satisfy coverage requirements separately for any given sample size. They may be called episodic inferences. The distinction between episodic and sequential environments appears in artificial intelligence, see Russel and Norvig (2010, Section 2.3.2). All in all, Robbins' confidence sequences strengthen the standards of confidence, and, thanks to their frequentist assurance, offer more compelling summarizations of evidence.

Appendix

Robbins' confidence sequences have the required persistent coverage (Robbins, 1970)

To see that, for regions of the form (2), inequality (1) holds for every $\theta \in \Theta$, consider that

$$P_\theta\left(\theta \in \hat{\Theta}_{1-\varepsilon}(Y^{(n)}) \ \text{ for every } \ n \geq 1\right) = 1 - P_\theta\left(\theta \notin \hat{\Theta}_{1-\varepsilon}(Y^{(n)}) \ \text{ for some } \ n \geq 1\right)$$

and

$$P_\theta\left(\theta \notin \hat{\Theta}_{1-\varepsilon}(Y^{(n)}) \ \text{ for some } \ n \geq 1\right) = P_\theta\left(\frac{q_n(Y^{(n)})}{p_n(Y^{(n)};\theta)} \geq \frac{1}{\varepsilon} \ \text{ for some } \ n \geq 1\right) .$$

The last probability does not exceed ε in force of a fundamental inequality for the likelihood ratio statistic.

Let P and Q denote the joint probability distribution of the sequence $Y^{(\infty)}$ when $Y^{(n)}$, $n = 1, 2, \ldots$, have density $p_n(Y^{(n)})$ and $q_n(y^{(n)})$, respectively. Then

$$P\left(\frac{q_n(Y^{(n)})}{p_n(Y^{(n)})} \geq k \ \text{ for some } \ n\right) \leq \frac{1}{k}\,, \tag{7}$$

for any $k > 0$. Robbins' proof of (7) is as follows. Define the stopping time

$$N = \min\left\{n \geq 1 \, : \, \frac{q_n(Y^{(n)})}{p_n(Y^{(n)})} \geq k\right\} ,$$

with $\min \emptyset = \infty$. Then

$$\begin{aligned} P\left(\frac{q_n(Y^{(n)})}{p_n(Y^{(n)})} \geq k \text{ for some } n\right) &= P(N < \infty) \\ &= \sum_{n \geq 1} P(N = n) = \sum_{n \geq 1} \int_{\{y^{(n)} \, : \, N=n\}} p_n(y^{(n)}) \, dy^{(n)} \\ &\leq \sum_{n \geq 1} \int_{\{y^{(n)} \, : \, N=n\}} \frac{1}{k} q_n(y^{(n)}) \, dy^{(n)} \\ &= \frac{1}{k} \sum_{n \geq 1} Q(N = n) = \frac{1}{k} Q(N < \infty) \\ &\leq \frac{1}{k} \, . \end{aligned}$$

Inequality (7) also follows from a well-known martingale inequality, see e.g. Jacod and Protter (2000, Theorem 26.1).

A heuristic argument for the consistency of confidence sequences

Rigorous proofs of consistency of $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ when the density of $Y^{(n)}$ belongs to an exponential family are given by Lai (1976) and Csenki (1979) for the one-parameter and the multiparameter case, respectively. For models whose likelihood function obeys the usual regularity conditions (see e.g. Severini, 2000, Section 3.4), consistency of $\hat{\Theta}_{1-\varepsilon}(y^{(n)})$ may be seen by the following heuristic argument.

Assume that $\hat{\theta}_n$ is the unique maximum of $L(\theta; y^{(n)})$ in an open neighborhood of the true θ. Let $\ell(\theta; y^{(n)}) = \log L(\theta; y^{(n)})$ be the log likelihood function and let $j_n(\theta) = j(\theta; y^{(n)}) = -\partial^2 \ell(\theta; y^{(n)})/\partial\theta\partial\theta^\top$ be the observed information. Assume moreover that, as under repeated sampling of size n, $\ell(\theta; Y^{(n)}) = O_p(n)$ and $j(\hat{\theta}_n; Y^{(n)})$ is positive definite and of order $O_p(n)$. Using Laplace expansion, see e.g. Barndorff–Nielsen and Cox (1989, Section 3.3), we have

$$q_n(y^{(n)}) = \int_\Theta p_n(y^{(n)}; \theta)\pi(\theta) \, d\theta = p_n(y^{(n)}; \hat{\theta}_n) \frac{\pi(\hat{\theta}_n)(2\pi)^{p/2}}{|j_n(\hat{\theta}_n)|^{1/2}} \{1 + O(n^{-1})\} , \tag{8}$$

so that

$$\hat{\Theta}_{1-\varepsilon}(y^{(n)}) = \left\{\theta \in \Theta \, : \, \ell(\theta; y^{(n)}) > \ell(\hat{\theta}_n; y^{(n)}) + \log\left(\frac{\varepsilon\pi(\hat{\theta}_n)(2\pi)^{p/2}}{|j_n(\hat{\theta}_n)|^{1/2}}\right) + O(n^{-1})\right\} .$$

Let $k_n = -\log\left(\varepsilon\pi(\hat{\theta}_n)(2\pi)^{p/2}/|j_n(\hat{\theta}_n)|^{1/2}\right)$. Then, for $\theta' \neq \theta$,

$$P_\theta(\theta' \in \hat{\Theta}_n) = P_\theta\left(\ell(\hat{\theta}_n; Y^{(n)}) - \ell(\theta'; Y^{(n)}) < k_n + O_p(1)\right) ,$$

where $\ell(\hat{\theta}_n; Y^{(n)}) - \ell(\theta'; Y^{(n)})$ is $O_p(n)$ and positive, while

$$k_n = \frac{p}{2}\log n + O_p(1) . \tag{9}$$

Therefore,

$$\lim_{n\to\infty} P_\theta(\theta' \in \hat{\Theta}_n) = 0 .$$

Proof of (4).

Using Laplace expansion (8) and the definition of k_n we see that

$$\hat{\Theta}_{1-\varepsilon}(y^{(n)}) = \left\{\theta \in \Theta \,:\, 2\left(\ell(\hat{\theta}_n; y^{(n)}) - \ell(\theta; y^{(n)})\right) < 2k_n + O(n^{-1})\right\} ,$$

so that, if

$$2\left(\ell(\hat{\theta}_n; Y^{(n)}) - \ell(\theta; Y^{(n)})\right) \xrightarrow{d} \chi^2_p ,$$

then (4) follows from (9).

References

Barndorff–Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics.* London, Chapman and Hall.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., et al. (2017). Redefine statistical significance. *Nature Human Behaviour*, 33, 175.

Csenki, A. (1979). A note on confidence sequences in multiparameter exponential families. *Journal of Multivariate Analysis*, 9, 337–340.

Darling, D.A. and Robbins, H. (1967a). Iterated logarithm inequalities. *Proc. Nat. Acad. Sci. U.S.A.*, 57, 1188–1192.

Darling, D. A. and Robbins, H. (1967b). Confidence sequences for mean, variance and median. *Proc. Nat. Acad. Sci. U.S.A.*, 58, 66–68.

Gandy, A. and Hahn (2016). A framework for Monte Carlo based multiple testing. *Scand. J. Statist.*, 43, 1046–1063.

Goodman, S.N. (2016). Aligning statistical and scientific reasoning. *Science*, 352, 1180–1181.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124.

Jacod, J. and Protter, P. (2000). *Probability Essentials.* Berlin, Springer.

Jennison, C. and Turnbull, B.W. (1989). Interim analyses: The repeated confidence interval approach. *J. R. Statist. Soc.* B, 51, 305–361.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the USA*, 110, 19313–19317.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Lai, T. L. (1976). On confidence sequences. *Ann. Statist.*, 4, 265–280.

Lai, T. L. and Siegmund, D. (1986). The contributions of Herbert Robbins to mathematical statistics. *Statistical Science*, 1, 276–284.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., van Assen, M.A.L. M. et al. (2017). Justify your alpha: A response to “Redefine statistical significance”. Retrieved from psyarxiv.com/9s3y6.

McShane, B. B., and Gal, D. (2017). Statistical significance and the dichotomization of evidence (with discussion). *Journal of the American Statistical Association*, 112, 885–908.

Nakagawa, S. and Cuthill, I.C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82, 591–605.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Statist.*, 41, 1397–1409.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm.* London, Chapman and Hall.

Russel, S. and Hall, P. N. E. P. (2010). *Artificial Intelligence: A Modern Approach*, Third Ed., Prentice Hall, NJ.

Severini, T.A. (2000). *Likelihood Methods in Statistics*, Oxford University Press, Oxford.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70, 129–133.

Whitehead, J. (1989). Discussion of: Interim analyses: The repeated confidence interval approach by C. Jennison and B.W. Turnbull, *J. R. Statist. Soc.* B, 51, 338.

Table 4: Bernoulli population: empirical percentages of contradictions and non-coverages at some n for Robbins' confidence sequences for the mean with persistence level $1 - \varepsilon$ in 1,000 sequences of samples with size from 100 to 4,000 and various priors on the mean.

true θ	prior on θ	$100(1 - \varepsilon)$	50	80	90	95
0.5	$Beta(.5, .5)$	contradictions	0.6	0.3	0.1	0.0
		non-coverages	8.5	4.1	2.3	1.0
0.5	$Beta(1, 1)$	contradictions	1.8	0.4	0.2	0.0
		non-coverages	10.5	5.9	3.3	1.7
0.5	$Beta(5, 5)$	contradictions	6.3	1.1	0.4	0.2
		non-coverages	20.5	10.0	6.0	3.4
0.7	$Beta(.5, .5)$	contradictions	0.7	0.1	0.0	0.0
		non-coverages	7.2	3.0	1.7	0.5
0.7	$Beta(1, 1)$	contradictions	1.2	0.2	0.0	0.0
		non-coverages	10.3	3.7	2.3	0.8
0.7	$Beta(5, 5)$	contradictions	2.7	0.3	0.0	0.0
		non-coverages	12.0	5.6	2.1	1.3
0.9	$Beta(.5, .5)$	contradictions	0.4	0.1	0.0	0.0
		non-coverages	6.9	2.4	1.3	0.7
0.9	$Beta(1, 1)$	contradictions	0.3	0.0	0.0	0.0
		non-coverages	6.1	2.3	1.2	0.7
0.9	$Beta(5, 5)$	contradictions	0.0	0.0	0.0	0.0
		non-coverages	1.1	0.4	0.1	0.1