METRON (2018) 76:251–268
https://doi.org/10.1007/s40300-018-0139-1

Confidence, credibility and prediction

Murray Aitkin[1] · **Charles Liu**[2]

Received: 25 September 2017 / Accepted: 2 April 2018 / Published online: 19 June 2018

Abstract This paper examines, from a historical and model-based Bayesian perspective, two inferential issues: (1) the relation between confidence coverage and credibility of interval statements about model parameters; (2) the prediction of new values of a random variable. Confidence and credible intervals have different properties. This worries some statisticians, who want them to have the same properties, in frequentist repeated sampling coverage and Bayesian credibility content. Research continues on the conditions under which intervals have the same repeated sampling coverage and credibility content. We conclude that these two inferential approaches generally have incommensurable properties, which converge only asymptotically in sample size, and in a restricted class of samples. The repeated confusion, by new students of statistics, of the coverage of a confidence interval with the credibility of the observed interval, suggests that in general the credibility of the interval is the more important inferential aim. We show how the credibility of a confidence interval can be assessed generally. Bayesian prediction of new values is well established using the posterior predictive distribution. This has some curious features, known since Laplace but not well understood. The prediction of new Bernoullis highlights these features. We suggest that the credibility of predictive intervals needs to be reassessed.

Keywords Likelihood · Confidence · Credible · Prediction intervals

1 History of confidence and credibility

1.1 The central limit theorem

In the nineteenth century there was no general theory of statistical inference. The repeated sampling distribution of means was known to be normal from the Central Limit Theorem,

✉ Murray Aitkin
murray.aitkin@unimelb.edu.au

[1] School of Mathematics and Statistics, University of Melbourne, Parkville, Australia

[2] Cytel Inc., Boston, USA

first discovered by de Moivre in 1733 and extended by Laplace in 1812. So the mean and variance of a normal sampling distribution could be estimated from data, and the precision of the sample mean estimate expressed through multiples of the sampling standard deviation. This was initially expressed through the *probable error* (giving 50% probability between the sample mean $\pm$ one probable error), later changed to the *standard error*, the square root of the sampling variance. The absence of any assessment of goodness-of-fit of the normal distribution to sample data allowed a general belief that all, or nearly all, variables were normally distributed, and so estimation of *any* population mean could be based on the sampling distribution of the sample mean, with its sampling standard deviation.

1.2 Bayes's theorem

Bayes's theorem of [4] had shown how to use "inverse probability" to draw conclusions about competing hypotheses for data, but this needed a prior distribution. The general assumption of normal distributions for variables allowed a uniform prior distribution for the mean, widely recommended by Laplace [18], under his *Principle of Insufficient Reason*, which led to the same posterior conclusions as the repeated sampling distribution, since the standard deviation was assumed known. So in the nineteenth century data analysis was mixed: some analysts used Bayes's theorem, others used the repeated sampling distribution, with no serious conflict among them.

1.3 Karl Pearson

From 1894, Karl Pearson made a dramatic change to this comfortable state. To do this, he developed three new concepts and tools:

- a method for assessing the agreement between empirical data and the normal distribution;
- a set of alternative distributions which could represent departures from normality;
- a criterion for fitting these distributions to empirical data.

Over an 8-year period 1894–1902, he achieved all three of these aims:

- the X^2 test for goodness-of-fit [24] was the first such test, based on the comparison of the sample frequencies in each histogram bin with those expected under the normal distribution, computed as normal bin probabilities from the fitted normal distribution multiplied by the sample size. He showed that the sum of squared differences divided by the expected frequencies was distributed asymptotically as χ^2, with degrees of freedom which he took as the number of histogram bins.[1]
- Pearson [25,26] proposed a large system of non-normal distributions, based on the solutions to a set of differential equations. These distributions had up to four parameters, allowing for skewness and kurtosis in the densities of these distributions.
- He had already [23] proposed the *method of moments* for estimating the parameters of distributions, by equating the population moments (easily computed for his new family of distributions) to the sample moments of the data, and solving the resulting set of simultaneous equations.

With his new goodness-of-fit test Pearson was able to show that many published data sets which were put forward as examples of the normal distribution were in fact definitely non-normal, with minute tail-area goodness-of-fit probabilities from the χ^2 distribution. Pearson's

[1] The small expected frequencies in the extreme bins affected the validity of the asymptotic distribution, and so he suggested *pooling* of the extreme bins.

examples of (grouped) continuous variables which were found to be clearly non-normal began to discredit the almost universal assumption of normality, and the *Pearson system* of non-normal distributions became an important tool for analysis.

1.4 Fisher and likelihood

Fisher's development of likelihood [11–13] was a revolution in statistical inference. It depended on a probability model specification, but did *not* depend on prior information about the model parameters.

Maximum likelihood (ML) became a standard tool in non-Bayesian inference, and the first two derivatives of the log-likelihood function became the standard tools for estimation and precision for nearly a century. The key to this success was that the log-likelihood was generally a *linear function* of the responses y_i, and so the Central Limit Theorem would provide the *asymptotic* distributions of any linear functions involved in the MLE. In some probability distributions the *exact* distribution of the MLE could be found. Fisher dismissed Pearson's method of moments as it did not use the likelihood, so did not give *efficient* estimates. An inefficient estimate could be improved by using it as a starting value in the scoring algorithm; after one iteration the new estimate would be effectively efficient.

A troubling feature of using the first two derivatives of the log-likelihood was that monotone transformations of the model parameter would give different precisions, and so the precision quoted for the ML estimate depended on the parametrization adopted. Fisher's theory depended on the Central Limit Theorem for the distribution of the sufficient statistics, but he did not accept the repeated sampling principle, though the interpretation of the precision of the estimates *required* this principle. He referred to the "unique sample", and developed his own fiducial theory of how to transform from a precision statement about the MLE to a precision statement about the parameter, without using the repeated sampling principle. This theory was very close to the Bayesian theory, but without the prior: Fisher was fiercely anti-Bayesian.

Neyman formulated the repeated sampling approach through *confidence intervals*: the *confidence coefficient* of a confidence interval for a model parameter is the coverage probability of the interval in repeated sampling. For the normal distribution this gave the same intervals as Fisher's fiducial approach, and Fisher at first thought that the theories were equivalent, but more complex examples showed that they were different, and Fisher then rejected confidence intervals. Since Fisher's fiducial theory did not produce acceptable general results, confidence intervals became the standard expression of precision.

Early studies of maximum likelihood showed that it could give biased estimators in small samples, so was only *asymptotically* efficient, and so other estimators, like moment estimators, might have better small-sample properties than ML estimators. The repeated sampling principle does not specify which estimator should be used, and this stimulated a search for alternative estimators, which would be compared by their sampling variances or mean squared errors, if biased estimators were to be allowed.

1.5 Jeffreys and Bayesian views of likelihood

Bayesians were already using the likelihood, in Bayes's theorem. So in the view of many Bayesians, Fisher was reproducing with maximum likelihood the Bayesian results achieved with flat or noninformative priors. In Bayesian analysis there was no need to rely on the first two derivatives of the log-likelihood: the whole posterior distribution was available. The problem was computational rather than theoretical: except in simple models with flat priors,

posterior distributions were not analytic, and computation was slow and tedious on electric calculators.

Bayesian theory was developed further by Jeffreys [17]. He argued from first principles for the Bayesian inferential theory, and stimulated by Fisher's likelihood developments, he gave the Bayesian versions of several of Fisher's innovations, using flat priors on normal means and regression model parameters and the log standard deviation. He showed that the Bayesian and Fisherian analyses were very close. He also argued strongly for non-informative priors, following Bayes and Laplace. Bayesian contributions from 1925 to 1960 were modest, and Fisher and Neyman dominated the statistical world, with the FNP—Fisher–Neyman–Pearson—paradigm developing rapidly and fruitfully.

In his 1961 edition Jeffreys hinted at the development of more general priors which would expand the Bayesian analysis possibilities. This development, of conjugate priors, by Raiffa and Schlaifer [27] greatly expanded the Bayesian toolkit. The idea was a simple one: to make the prior "fit"—*conjugate with*—the likelihood by replicating its structure with additional prior parameters. So for the Bernoulli model with likelihood $c \cdot p^r(1-p)^{n-r}$ the prior has the same form, of a *Beta* distribution:

$$\pi(p) = p^{a-1}(1-p)^{b-1}/B(a,b),$$

where $B(a,b)$ is the complete Beta function.

The posterior distribution is then another Beta distribution

$$\pi(p \mid r,n) = p^{r+a-1}(1-p)^{n-r+b-1}/B(r+a, n-r+b),$$

in which the prior parameters are simply added to the success and failure counts in the likelihood. By varying the prior parameters, the prior information could be represented by the full variety of shapes of the Beta distribution in both location and variation. An additional valuable feature was that the prior information could be thought of as coming from an *auxiliary experiment* which gave $a-1$ successes and $b-1$ failures. In particular, when $a = b = 1$ (no auxiliary experiment has actually been performed), and there is clearly *no prior information*, the prior reduces to the *uniform* prior, giving another justification for this prior as the *noninformative* prior for the Bernoulli case.

The term *credible interval*—an interval statement about the parameter from the posterior—was coined to replace "Bayesian confidence interval". Although in normal models credible intervals had the same formal probability content as confidence intervals, this was not so in other simple models like the binomial, which we discuss below.

1.6 The MCMC revolution

The EM algorithm of Dempster, Laird and Rubin [9] was a remarkable contribution to the maximum likelihood analysis of complex data. It enabled full ML analysis of models with latent variables, latent structure, mixtures and missing data. However the reliance on the first two derivatives of the log likelihood for estimates and precisions showed the difficulties of the Fisherian paradigm: the more latent or missing data structures there were in the model, the less reliable was the second derivative information matrix for the precisions of the parameter estimates. Missing data produced *skewness and kurtosis* in the likelihood which could not be identified in the first two derivatives.

The Bayesian version of EM, first published by Tanner and Wong [29], solved this by alternating between *random draws of the parameters from their conditional distribution given random draws of the missing data*, analogous to the M-step computation of ML estimates, and *random draws of the missing data from their conditional distribution given the random*

draws of the parameters, analogous to the E-step conditional expectation of the missing data, in the EM algorithm. This provided the full joint posterior distribution of the model parameters, which EM could not: the *random draws* of the parameters were based on the full likelihood (plus priors), and so allowed correctly for the shape of the likelihood.

A further great advantage of MCMC was that it allowed the posterior computation of *any function of the model parameters and observed data*, by simple substitution of the random parameter draws into the function.

This development has reinvigorated the Bayesian paradigm, which now has the tools for full analysis, which the FNP paradigm does not. This has important implications for standard data analyses, which we illustrate below.

2 Bernoulli trials I

An experimenter performs $n = 10$ Bernoulli trials with success probability p. $r = 3$ successes are observed. What interval statement can be made about the success probability p?

Frequentists and Bayesians approach this question from different philosophical positions: the *repeated sampling principle* for frequentists and the *likelihood principle* (combined with a prior distribution) for Bayesians. (Extensive discussions of these principles can be found in [7].)

For frequentists, in single parameter models, inference is based on the repeated sampling distribution of the maximum likelihood estimate (MLE), or of the score function. For the Bernoulli model, the likelihood, log-likelihood and score functions are

$$\begin{aligned} L(p) &= \binom{n}{r} p^r (1-p)^{n-r} \\ \ell(p) &= c + r\log(p) + (n-r)\log(1-p) \\ s(p) &= \frac{r}{p} - \frac{n-r}{1-p} \\ &= \frac{r-np}{p(1-p)}. \end{aligned}$$

The MLE is $\widehat{p} = r/n$, with sampling mean p and variance $p(1-p)/n$, while the score has sampling mean 0 and variance $1/[np(1-p)]$. These lead to the same conclusion, invoking the Central Limit Theorem for the asymptotic normal distribution of $\widehat{p}$: $N(p, p(1-p)/n)$.

Replacing the unknown p in the variance by its MLE $\widehat{p}$ gives the Wald asymptotic (approximate) 95% confidence interval for p: $\widehat{p} \pm 2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$, giving 0.3 ± 0.29, or $[0.01, 0.59]$. A better approximation has the p confidence limits as the roots of the quadratic inequality obtained from the interval in which p is *not* replaced by $\widehat{p}$ in the variance:

$$\begin{aligned} &(p-\widehat{p})^2 \le 4p(1-p)/n \\ &(n+4)p^2 - 2p(n\widehat{p}+2) + n\widehat{p}^2 \le 0, \end{aligned}$$

which are

$$\left[n\widehat{p} + 2 \pm 2\sqrt{n\widehat{p}^2 + n\widehat{p} + 4}\right]/(n+4),$$

or

$$\left[r + 2 \pm 2\sqrt{r^2/n + r + 4}\right]/(n+4).$$

For the example this gives [0.16, 0.56]. Other approximate intervals can be constructed, but these two are sufficient for our purposes. Asymptotically, the sets of intervals (hypothetically) constructed in this way will cover the true value of p in 95% of the intervals. Whether the confidence interval from the *observed* sample covers the true value is unknown: the coverage probability refers to a hypothetical ensemble of such samples, not to the sample on which the interval is based.

For a Bayesian, a prior is needed to answer the question. Throughout this paper we use the uniform prior, given no other information about the experiment. The posterior distribution of p is then $Beta(4, 8)$, and the central 95% credible interval for p is [0.11, 0.61]. Figure 1 shows the cdf of the Beta(4, 8) distribution.

The 95% credible and asymptotic 95% confidence intervals are different, and the Wald confidence interval is longer, while the quadratic interval is shorter, than the credible interval. The *posterior credibility of any interval* is easily computed from the cdf of the Beta(4, 8) distribution. For the Wald interval this is 0.966, and for the quadratic interval it is 0.867. Credibility increases with length, not surprisingly. If our aim is 95% credibility, the Wald interval is unnecessarily long, and the quadratic interval is too short.

Many improvements on the Wald interval have been suggested, mostly based on *monotone transformations* of the success probability p. These were investigated by Anscombe [2] and extended by Diaz-Francés [10]. As noted above, this dependence on parametrization is a difficulty with the Fisherian paradigm. For the purposes of this paper, we focus on the proposal of Agresti and Caffo [1], who suggested improving the Wald interval by adding two successes and two failures to the observed data; they noted that this was equivalent

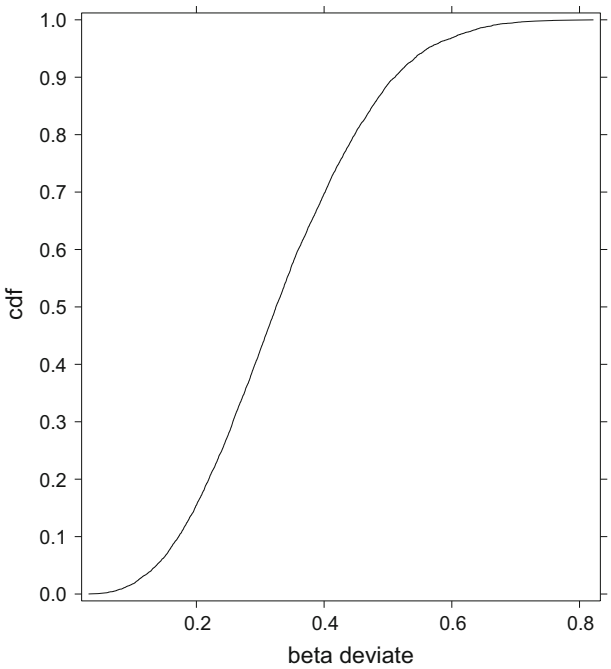

Fig. 1 Beta(4, 8) cdf

to an informative $Beta(3, 3)$ prior on p. Under this prior, the prior probability is 0.5 that $p \in [0.36, 0.64]$.

For a Bayesian, it is difficult to see why this informative prior could be *generally* appropriate—what is the relevant external information on which it is based? Are values of p around 0.5 *universally* more frequent than values near 0 or 1? A Bayesian using this prior with the trial data would give a central 95% credible interval of [0.16, 0.62], compressed slightly and shifted slightly to the right compared to the uniform prior, as would be expected from its informativeness.

2.1 The relation of credibility to coverage

Many Bayesians *and* frequentists would like credibility and confidence coverage to agree, at least asymptotically. Rubin [28] argued that credible intervals should be *well calibrated* in the sense that 95% credible intervals should have repeated-sampling coverage near 95%, and suggested posterior predictive checks as a way of using frequentist repeated sampling procedures to assess Bayesian model assumptions. Gelman et al. [15] extended and formalised this approach. Little [19–21] discussed it at length, in arguing for a broader acceptance of Bayesian procedures.

Frequentists saw the problem differently. For example, in their review of *probability matching priors* (PMPs), Datta and Sweeting [8] commented:

> From a Bayesian point of view it can be argued that a PMP is a suitable candidate for a nonsubjective Bayesian prior [for p], since the repeated sampling property of the associated posterior regions provides some assurance that the Bayesian results will make some inferential sense, at least on average, whatever the true value of [p].

Datta and Sweeting took it as given that Bayesians have to assure frequentists that their credible intervals "will make some inferential sense", presumably in the sense of having the same confidence coverage as their credibility. The implication is that this is to be achieved by *getting the prior right*. Datta and Sweeting's review paper however showed that the agreement between confidence coverage and credibility could not be achieved in general by choice of the prior. Since coverage and credibility refer to different inferential processes, it is not surprising that in general they give different intervals (for the same credibility and nominal confidence). Datta and Sweeting [8] summarised the connections.

Exact agreement occurs in models with normal likelihoods and flat priors in the parameter θ: if the log-likelihood is quadratic in θ then central confidence and central credible intervals are exactly equivalent. This result extends *asymptotically* (with flat priors) to models with internal log-likelihood maxima in the parameter space, where the cubic and higher order terms in the Taylor expansion about the MLE go to zero with sample size, relative to the quadratic term.

The binomial case is a simple example. If the MLE is near 0.5, the likelihood will be close to normal even in small samples. If it is near 0 or 1, the likelihood will not be close to normal even in large samples. The credible interval automatically adjusts for any skew in the likelihood; its validity as a measure of precision is unaffected by the sample size or the location of the MLE.

The Datta and Sweeting results have recently been upgraded by Müller and Noretz [22]. They showed that under "mild" conditions (excluding continuous parameter spaces) there existed a *coverage inducing prior* for which the $1 - \alpha$ highest posterior density or equal-tailed credible interval gave exact confidence coverage of $1 - \alpha$. This prior however had to be determined iteratively as the solution of a fixed-point problem, involving adjusting an

initial guess for the prior by iteratively updating the (discrete) prior probabilities of parameter values which were found to undercover in repeated sampling with these values. This could require a vast amount of computing, just to obtain a prior.

For Bayesians, this approach raises the same question as that of Agresti and Caffo: why should the choice of prior be determined by the frequentist coverage of the credible interval? For frequentists, the Bayesian analysis depends on the prior assumption. Many Bayesians emphasize the importance of *getting the prior right*. To these Bayesians the prior is fundamentally subjective—"right" for the analyst who specifies it. Many frequentists have no idea of the reproducibility of this kind of inference, unlike the confidence interval which is based on the repeated sampling process and the properties of the intervals in this process. Simulation studies follow exactly this process in evaluating the properties of interval procedures. Fraser [14] argued that Bayes is just *quick and dirty confidence*.

Bayesians *are* in conflict over the role of priors. The Jeffreys recommendation for the uniform prior is supported by the "reference" school of Bayesians [5] which requires a default non-informative (or minimally informative) analysis so that the data can "speak" through the likelihood, unaffected by an informative prior. Then we can assess the effect of an informative, subjective or personal prior on the default analysis.

To most Bayesians, the reproducibility of the confidence interval statement is irrelevant. We know that students constantly misinterpret confidence intervals as credible intervals, despite the endless emphasis by even the best teachers that the confidence coefficient is *not* the probability that the parameter lies in the interval, and that we can say *nothing* about this probability—either the interval covers it, or it doesn't. The replication of the interval construction process is hypothetical, and its claimed coverage depends on unverifiable assumptions about the repeated sampling distribution of the sufficient statistic or MLE in the hypothetical unobserved samples.

However it is in principle straightforward to determine the credibility content of a confidence interval by direct analysis, or by sampling from the model parameter posterior, though determining the confidence coverage of a credible interval procedure would require large-scale simulations. For most experimenters or survey analysts, the precision question of interest is: *how precise was my estimation in this sample*, not *how precise would this form of estimation be in an ensemble of such samples?*

We argue that the concern of Datta and Sweeting for the reassurance of frequentists by Bayesians is misplaced: it is rather for frequentists to *reassure themselves* of their precision statements by establishing that their confidence intervals have *acceptable credibility for the given sample*.

We now consider prediction.

3 History of prediction

Statistical prediction is a relatively recent development of statistical modelling. The first example was given by Laplace, in his famous *Law of Succession*, discussed below. Jeffreys (1961, p. 143) obtained the predictive distributions for the mean and variance of a second sample from a normal distribution given a first sample and a flat prior distribution of the mean and log variance. The first comprehensive book, by Aitchison and Dunsmore [3] followed the Bayesian approach, which we describe below. Fourteen non-Bayesian likelihood prediction approaches are discussed at length in Bjornstad [6]. We do not deal with them here as our concern is with the fundamentals of the Bayesian approach.

4 Bernoulli trials II

4.1 Predicting a binary

The previous experimenter wants to *predict*—make a predictive statement about—the outcome of the 11-th trial, given the three successes in the first ten trials. How should this be done? We use the non-informative uniform prior throughout this section. The posterior distribution of p is $Beta(4, 8)$. We note that the posterior mode is 0.3, the posterior median is 0.324, the posterior mean is 0.333, and the posterior standard deviation is 0.13.

The prediction of a new *binary* Y is of the form $p^Y(1-p)^{1-Y}$, which is equivalent to a prediction of p, the probability that the new $Y = 1$.

The Bayesian approach has been standard to Bayesians (Gelman et al. [16], p. 7) since Laplace's Law of Succession. The probability of success at the 11th trial, given the trial data, is expressed as

$$\begin{aligned}\Pr[Y_{11} = 1 \mid r, n] &= \int_0^1 \Pr[Y_{11} = 1 \mid p] \cdot \pi(p \mid r, n)\,\mathrm{d}p \\ &= \int_0^1 p \cdot p^3(1-p)^7\,\mathrm{d}p / B(4, 8) \\ &= B(5, 8)/B(4, 8) \\ &= 4/12 = 0.333.\end{aligned}$$

This is called the *posterior predictive probability* of success at the next trial. It can be extended to more complex predictions, like *the probability of s successes in a second set of m trials*. This is

$$\begin{aligned}\Pr[s \mid m, r, n] &= \binom{m}{s} \int_0^1 p^s(1-p)^{m-s} \cdot \pi(p \mid r, n)\,\mathrm{d}p \\ &= \binom{m}{s} \int_0^1 p^{s+3}(1-p)^{m-s+7}\,\mathrm{d}p / B(4, 8) \\ &= \binom{m}{s} B(s+4, m-s+8)/B(4, 8).\end{aligned}$$

For example, the posterior predictive probability of $s = 3$ successes in a second set of $m = 10$ trials is

$$\binom{10}{3} B(7, 15)/B(4, 8) = 0.195.$$

This approach, though widely used, has some strange features:

- it gives a precise probability for any prediction event, however small the sample size;
- it gives a precise probability for any prediction event, *without any data*.

On the second point, the predictive argument can be applied with the prior distribution instead of the posterior, if there are no data, to give a *prior predictive distribution* (Gelman et al. [16], p. 7).

For the prior predictive probability of success at the outcome of the *first* trial, we have

$$\Pr[Y_1 = 1] = \int_0^1 \Pr[Y_1 = 1 \mid p] \cdot \pi(p)\,\mathrm{d}p$$

$$= \int_0^1 p \cdot 1 \, \mathrm{d}p$$
$$= 0.5.$$

For the prior predictive probability of three successes in the first 10 trials, we have

$$\Pr[s = 3 \mid m = 10] = \int_0^1 \binom{10}{3} p^3 (1-p)^7 \cdot 1 \, \mathrm{d}p$$
$$= \binom{10}{3} B(4, 8)$$
$$= 1/11 = 0.091.$$

It seems very strange that, with no data and a non-informative prior, we can make precise informative probability statements about future values. The explanation is simple, though rarely emphasized:

> … the posterior predictive distribution [is] an *average* of conditional predictions over the posterior distribution of θ (Gelman et al. [16] p. 7, emphasis added).

So the posterior predictive probability of success at the 11-th trial is simply the *mean* of the posterior distribution of p. *It is not the probability of success at the 11th trial*,—that remains at p—*but a one-number summary of its posterior distribution*, like the posterior mode or median, or the plug-in MLE. In the same way, the prior predictive probability of success at the first trial is not the probability of success, but its prior mean: the choice of prior determines the prior predictive probability, but not the probability of success.

What predictive statements can be made then, priorly or posteriorly, about new values? Exactly those inferences from the prior or the posterior.

Without any data, what can be said about the probability of success at the first trial? With no data, the only statement is that the probability of success has the uniform prior distribution. Given three successes in the first 10 trials, the probability of success at the 11th trial has the $Beta(4, 8)$ distribution. The probability of three more successes in the next 10 trials does not have an analytic posterior distribution, but it can be simulated easily:

- make M = 10,000 random draws $p^{[m]}$ of p from its $Beta(4, 8)$ posterior distribution;
- substitute them into the binomial probability of three successes in 10 trials:

$$b^{[m]}(3, 10) = b(3, 10; p^{[m]})$$
$$= \binom{10}{3} p^{[m]3} (1 - p^{[m]})^7,$$

and sort the binomial draws into increasing order. Figure 2 shows the (severely skewed) cdf of the 10,000 draws of the binomial probability. The 95% central credible interval is [0.028, 0.267], the posterior mean is 0.195 (the posterior predictive probability value) and the posterior median is 0.218.

With continuous distributions, the problem is more algebraically complex. We give a general definition.

4.2 General theory

Given a probability model $f(y \mid \theta)$ for a random variable Y with cdf $F(y \mid \theta)$, a random sample $\mathbf{y}$ of size n already drawn from f, and a prior distribution $\pi(\theta)$ for θ, what can be said (as a 95% central probability interval statement) about the next value of Y to be drawn?

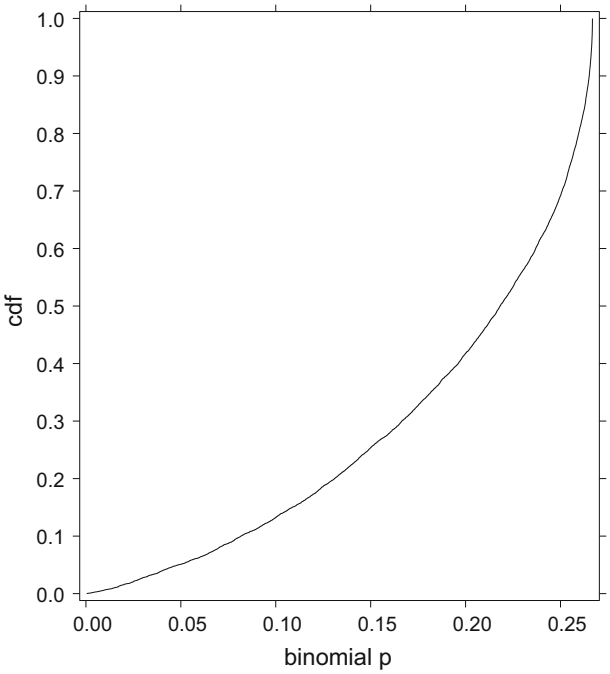

Fig. 2 Binomial (3, 7) cdf

The two methods to be considered both use the posterior distribution $\pi(\theta \mid \mathbf{y})$ of θ given the sample data.

- Method 1 (traditional):
  - Compute the posterior predictive distribution of Y_{n+j}: $g(y_{n+j} \mid \mathbf{y}) = \int f(y_{n+j} \mid \theta)\pi(\theta \mid \mathbf{y})\mathrm{d}\theta$, with cdf G;
  - find the appropriate percentiles α and β of the posterior predictive distribution: $G(\beta \mid \mathbf{y}) = 0.975,\ G(\alpha \mid \mathbf{y}) = 0.025$.
- Method 2 (alternative):
  - Find the appropriate percentiles $a(\theta)$ and $b(\theta)$ of the *conditional* distribution of Y_{n+j} given θ: $F(b(\theta)) = 0.975,\ F(a(\theta)) = 0.025$;
  - compute the *conditional credibility* $\gamma(a, b \mid \theta) = b(\theta) - a(\theta)$;
  - make M draws $\theta^{[m]}$ from the posterior distribution of θ and substitute them in γ to give M draws from the *marginal credibility distribution*: $\gamma^{[m]}(a, b) = \gamma(a, b \mid \theta^{[m]}) = b(\theta^{[m]}) - a(\theta^{[m]})$;
  - sort the M (typically 10,000) values of γ; and find the 2.5 and 97.5 percentiles of the posterior of γ.

These define the central 95% *credible interval* for γ.

4.3 Normal example

The experimenter draws a sample $y_1, \ldots, y_n$ with $n = 10$ from the normal distribution $N(\mu, \sigma^2)$. The sample mean $\bar{y}$ is 1 and the (unbiased) sample variance s^2 is 4. What inferential statement can be made about the next random value Y_{11} to be drawn from the distribution?

The traditional approach is very well known:

With reference flat priors on μ and $\log\sigma$, we have the standard posterior distributions:

- $\nu s^2/\sigma^2 \sim \chi^2_\nu$;
- $\mu \mid \sigma \sim N(\bar{y}, \sigma^2/n)$.

So

$$f(Y_{11} \mid \bar{y}, s^2) = \int f(Y_{11} \mid \mu, \sigma)\pi(\mu, \sigma \mid \bar{y}, s^2)\,\mathrm{d}\mu\,\mathrm{d}\sigma$$

$$Y_{11} \sim \bar{y} + s\sqrt{1 + 1/n} \cdot t_\nu,$$

and the $100(1-\alpha)\%$ central prediction interval for Y_{11} is

$$\bar{y} - t_{1-\alpha/2,n-1}s\sqrt{1+1/n} < Y_{11} < \bar{y} + t_{1-\alpha/2,n-1}s\sqrt{1+1/n}.$$

(This result is also obtained as one of the 14 frequentist likelihood-based predictive distributions.) For the example with $\bar{y} = 1$, $s = 2$, $n = 10$, the 95% central predictive interval is $[-3.75, 5.75]$.

For the alternative approach, we assess the credibility of this prediction interval. The probability that the new value Y_{11} lies in the interval (a, b), here $[-3.75, 5.75]$, given the parameters μ and σ, is

$$\Pr[Y_{11} \in (a, b) \mid \mu, \sigma] = \Phi[(b-\mu)/\sigma] - \Phi[(a-\mu)/\sigma] = \gamma(a, b \mid \mu, \sigma).$$

So the *credibility* $\gamma(a, b \mid \mu, \sigma)$ of the *conditional predictive interval* (a, b) is *a random function* of μ and σ. To compute its posterior distribution, we make 10,000 random draws $(\mu^{[m]}, \sigma^{[m]})$ from the joint posterior distribution of μ and σ, and substitute them into γ to give 10,000 random draws $\gamma^{[m]}$ from its posterior distribution.

It can be shown easily that the posterior predictive interval is the *mean* (with respect to the posterior distribution of (μ, σ)) of the conditional predictive interval given (μ, σ). Figure 3 shows the 10,000 values of $\mu^{[m]}$ graphed against $\sigma^{[m]}$, and Fig. 4 shows the cdf of the credibility distribution of γ for this example.

The 2.5%, median, mean and 97.5% values of the credibility γ are 0.782, 0.972, 0.951 and 0.999. The median is close to, and the mean is equal to, the predictive coverage, and the central 95% credible interval is [0.782, 0.999]. The uncertainty in μ and σ from the small sample is reflected in the variation in the credibility content of the 95% prediction interval.

Credibility can be improved by increasing the predictive coverage. Changing this to 99%, with the same sample data, gives the 2.5%, median, mean and 97.5% values of the credibility as 0.926, 0.999, 0.990 and 1.000. Figure 5 shows the credibility distribution.

4.4 Relation between approaches

These two approaches have different *foci*. The traditional approach focuses on the predictive *distribution*—we have to find this by integrating out the parameters to make an interval, or other, statement from it. Integrating out eliminates some of the random variability, since the result is the posterior mean—the posterior variance and higher moments are ignored. The

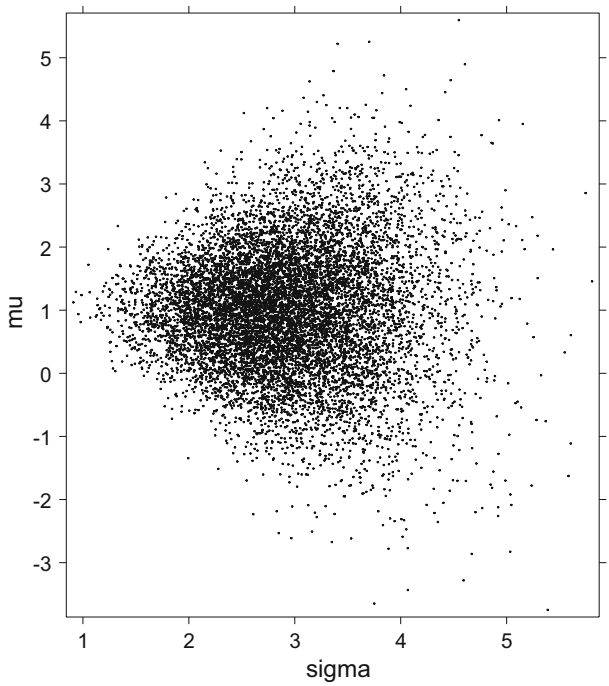

Fig. 3 10,000 posterior draws from $\pi(\mu, \sigma|\bar{y}, s)$

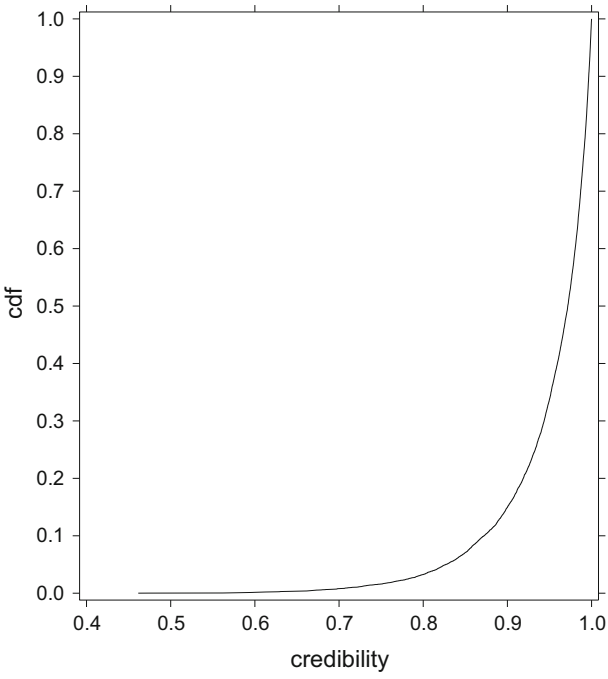

Fig. 4 Cdf of draws from the credibility distribution, 95% predictive

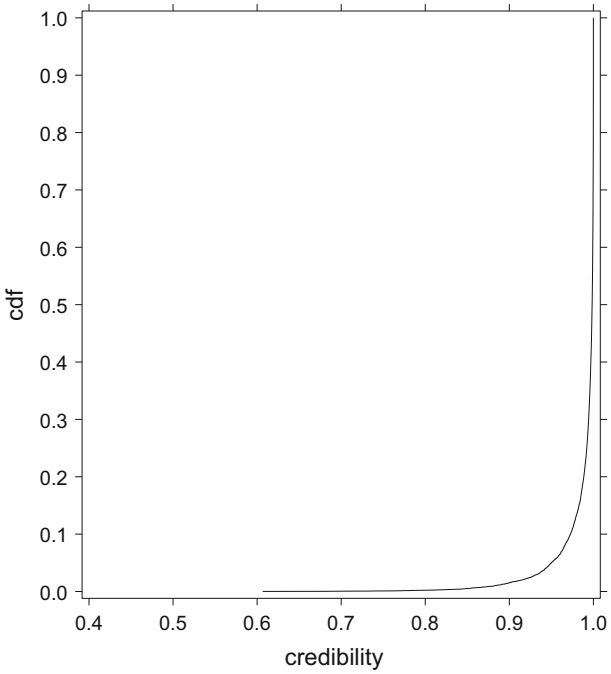

Fig. 5 Cdf of draws from the credibility distribution, 99% predictive

alternative approach focuses on the *interval*: we integrate out the parameters, not from the distribution of the new Y but from the conditional interval probability. The approaches differ by when *marginalisation* across the posterior distribution occurs.

The advantage of the traditional approach is that a full probability interval statement can be made; the advantage of the alternative approach is that it assesses the credibility of the traditional (or any other) interval. Its disadvantage is that it cannot give a predictive interval with exact credibility: *all* intervals will have a credibility *distribution*, though this converges to the predictive coverage with increasing sample size.

A small simulation illustrates this. The effect of sample size on the credibility of the 95% predictive interval is shown in two graph panels. Figure 6 shows 10,000 posterior draws of μ and σ for sample sizes n = 5, 10, 20 and 50, with the same data mean of 1 and SD of 2. Figure 7 shows the posterior cdfs of the 10,000 draws of γ from the parameter draws. Table 1 gives the medians and central 95% intervals for the credibility of the 95% predictive interval from the predictive distribution, extended to $n = 2$ and 100.

For samples of 50 or more the median credibility is close to the predictive coverage, and the credibility distribution converges to it as n increases and the posterior for μ and σ concentrates on the sample estimates. The strange appearance of the interval for $n = 2$ is a consequence of the very long tails of the Cauchy and χ^2_1 distributions. From a sample of 2 we learn almost nothing about a new Y.

Table 1 Median and 95% credibility intervals

n	2.5%	Median	97.5%
2	0.303	1.000	1.000
5	0.659	0.991	1.000
10	0.782	0.972	0.999
20	0.848	0.961	0.995
50	0.893	0.954	0.984
100	0.912	0.952	0.977

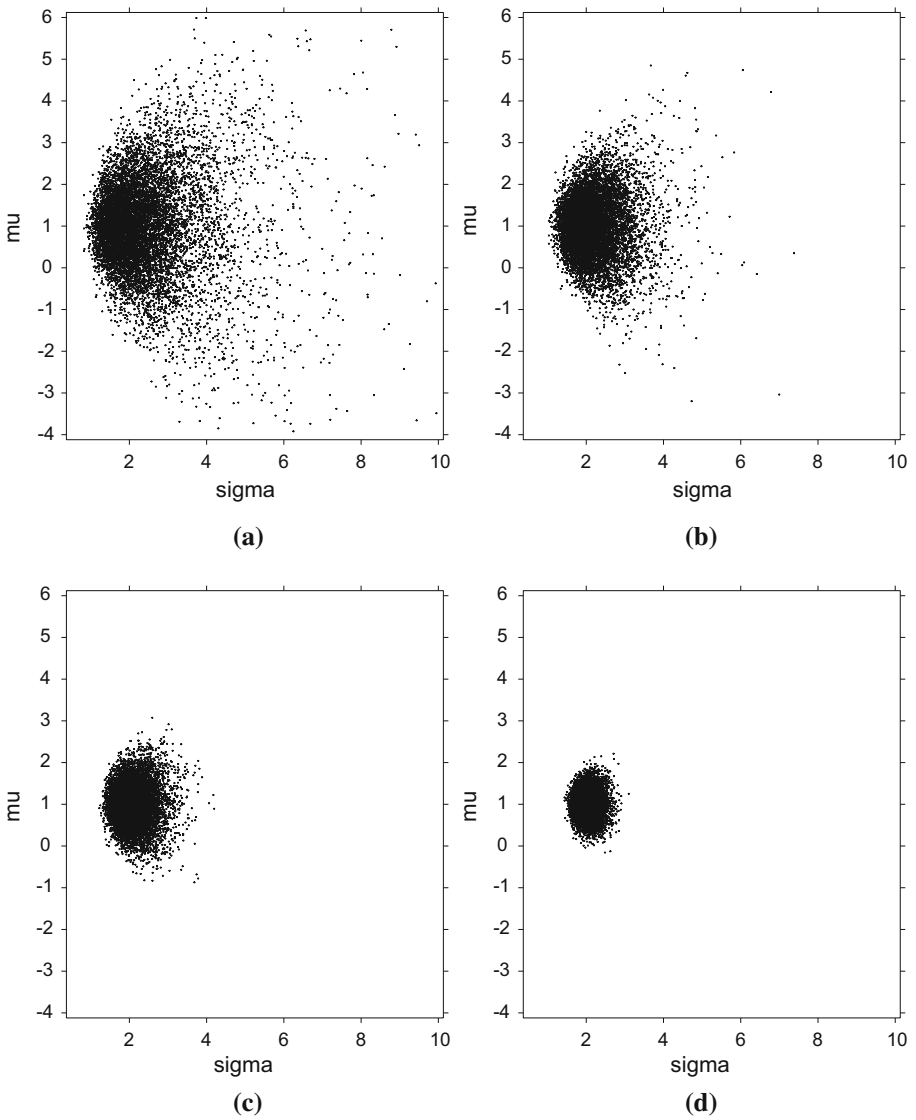

Fig. 6 10,000 posterior draws of μ and σ. **a** $n = 5$. **b** $n = 10$. **c** $n = 20$. **d** $n = 50$

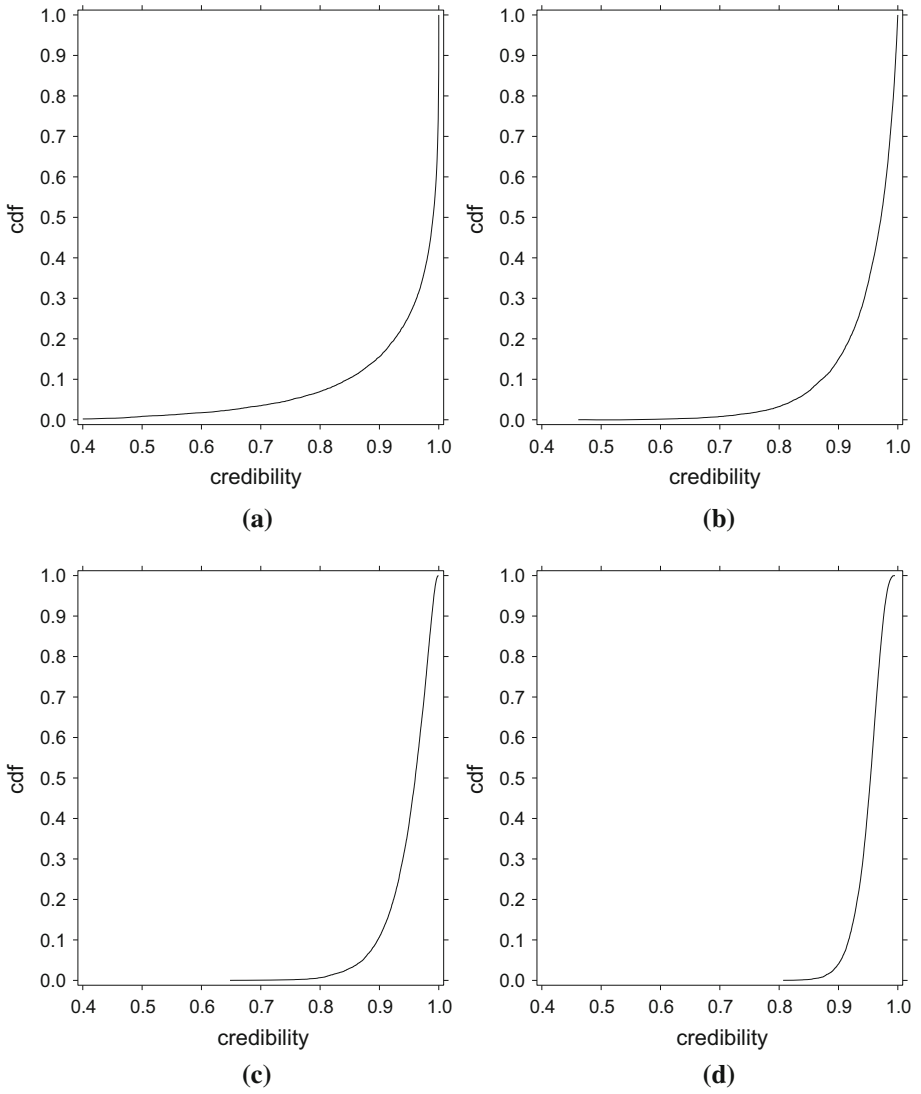

Fig. 7 Cdfs of 10,000 posterior draws from the posterior credibility distribution. **a** $n = 5$. **b** $n = 10$. **c** $n = 20$. **d** $n = 50$

5 Conclusion

As heterodox Bayesians, we see the claimed need for at least asymptotic coherence of Bayesian and frequentist interval procedures from a different angle. The "need" is clearly felt by frequentists: most Bayesians are not concerned about the repeated sampling properties of their credible intervals since these are irrelevant to the interpretation of the precision with the given sample. What concerns the heterodox Bayesians is that confidence coefficients may be much higher than credibility, and therefore that scientists relying on confidence interval procedures for their precisions need to ensure that the credibility of their quoted intervals

is adequate for their purposes: credibility is the scientifically more important expression of precision.

Since Laplace's Law of Succession it has become standard Bayesian practice to predict new values from the posterior predictive distribution. The simple binomial case shows the difficulty with this approach: the summarisation of the posterior by its mean. The prediction of continuous variables is more complicated, but the same issue appears: the posterior predictive distribution is again the posterior mean of the conditional predictives, and so the posterior predictive interval is the mean of the conditional predictive intervals.

Bayesians have been rightly critical of one-number mean or median summaries of posterior distributions of model parameters: their criticism needs to be extended to this aspect of prediction. Jeffreys (1961, p.53) long ago emphasised this point:

> Incorrect results have often been obtained by taking an expectation as a prediction of an actual value; this can be done only if it is also shown that the probabilities of different actual values are closely concentrated about the expectation. …

References

1. Agresti, A., Caffo, B.: Simple and effective confidence intervals for proportions and differences of proportions. Am. Stat. **54**, 280–288 (2000)
2. Anscombe, F.J.: Normal likelihood functions. Ann. Inst. Stat. Math. **16**, 1–19 (1964)
3. Aitchison, J., Dunsmore, I.R.: Statistical Prediction Analysis. Cambridge University Press, Cambridge (1975)
4. Bayes, T.: An essay towards solving a problem in the doctrine of chances. Philos. Trans. R. Soc. Lond. **53**, 370–418 (1764)
5. Berger, J.O., Bernardo, J.M., Sun, D.: The formal definition of reference priors. Ann. Stat. **37**, 905–938 (2009)
6. Bjornstad, J.F.: Predictive likelihood: a review. Stat. Sci. **5**, 242–265 (1990)
7. Cox, D.R.: Principles of Statistical Inference. Cambridge University Press, Cambridge (2006)
8. Datta, G.S., Sweeting, T.J.: Probability matching priors. In: Dey, D.K., Rao, C.R. (eds.) Handbook of Statistics, 25: Bayesian Thinking: Modeling and Computation, pp. 91–114. Elsevier, Amsterdam (2005)
9. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**, 1–38 (1977)
10. Diaz-Francés, E.: Simple estimation intervals for Poisson, exponential and inverse Gaussian means obtained by symmetrizing the likelihood function. Am. Stat. **70**, 171–180 (2016)
11. Fisher, R.A.: On an absolute criterion for fitting frequency curves. Messenger Math. **41**, 155–160 (1912)
12. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. Lond. A **222**, 309–368 (1922)
13. Fisher, R.A.: Theory of statistical estimation. Proc. Camb. Philos. Soc. **22**, 700–725 (1925)
14. Fraser, D.A.S.: Is Bayes just quick and dirty confidence? (with discussion). Stat. Sci. **26**, 299–316 (2011)
15. Gelman, A., Meng, X.L., Stern, H.S.: Posterior predictive assessment of model fitness via realised discrepancies (with discussion). Stat. Sin. **6**, 733–807 (1996)
16. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian date analysis, 3rd edn. CRC Press, Boca Raton (2014)
17. Jeffreys, H.: Theory of Probability (3rd edn. 1961, reissued 1998). Oxford University Press, Oxford (1939)
18. Laplace, P.-S.: Essai philosophique sur les probabilités (1814). English translation A Philosophical Essay on Probabilities. Wiley (1902) and Dover, New York (1950)
19. Little, R.J.A.: Calibrated Bayes: a Bayes/frequentist roadmap. Am. Stat. **60**, 213–223 (2006)
20. Little, R.J.A.: Calibrated Bayes, for statistics in general, and missing data in particular. Stat. Sci. **26**, 162–174 (2011)
21. Little, R.J.A.: Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. Stat. J. IAOS **31**, 555–563 (2015)
22. Müller, U.K., Noretz, A.: Coverage inducing priors in nonstandard inference problems. J. Am. Stat. Assoc. **111**, 1233–1241 (2016)
23. Pearson, K.: Contribution to the mathematical theory of evolution. Philos. Trans. R. Soc. Lond. A **185**, 71–110 (1894)

24. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. **50**, 157–175 (1900)
25. Pearson, K.: On the systematic fitting of curves to observations and measurements, Parts I. Biometrika **I**, 265–303 (1902)
26. Pearson, K.: On the systematic fitting of curves to observations and measurements, Parts II. Biometrika **II**, 1–23 (1902)
27. Raiffa, H., Schlaifer, R.: Applied statistical decision theory. Harward Business School, Boston (1961)
28. Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. **12**, 1151–1172 (1984)
29. Tanner, M., Wong, W.: The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**, 528–550 (1987)