Introduction to Modern Inverse Probability

Ryan Burn *ryan.burn@gmail.com*

April 9, 2024

In 1654 Pascal and Fermat worked together to solve the problem of the points [1] and in so doing developed an early theory for deductive reasoning with direct probabilities. Thirty years later, Jacob Bernoulli worked to extend probability theory to solve inductive problems. He recognized that unlike in games of chance, it was futile to *a priori* enumerate possible cases and find out "how much more easily can some occur than the others":

> But, who from among the mortals will be able to determine, for example, the number of diseases, that is, the same number of cases which at each age invade the innumerable parts of the human body and can bring about our death; and how much easier one disease (for example, the plague) can kill a man than another one (for example, rabies; or, the rabies than fever), so that we would be able to conjecture about the future state of life or death? And who will count the innumerable cases of changes to which the air is subjected each day so as to form a conjecture about its state in a month, to say nothing about a year? Again, who knows the nature of the human mind or the admirable fabric of our body shrewdly enough for daring to determine the cases in which one or another participant can gain victory or be ruined in games completely or partly depending on acumen or agility of body? (Bernoulli, 1713, p. 18)

The way forward, he reasoned, was to determine probabilities *a posteriori*

See Dale (1991) for a detailed history of inverse probability

> Here, however, another way for attaining the desired is really opening for us. And, what we are not given to derive a priori, we at least can obtain a posteriori, that is, can extract it from a repeated observation of the results of similar examples. (Bernoulli, 1713, p. 18)

To establish the validity of the approach, Bernoulli proved a version of the law of large numbers for the binomial distribution. Let X_n represent a sample from a Bernoulli distribution with parameter r/t (r and t integers). Then if c represents some positive integer, Bernoulli showed that for N large enough

$$P\left(|\frac{X_1 + \cdots + X_N}{N} - \frac{r}{t}| < \frac{2}{t}\right) > c \cdot P\left(|\frac{X_1 + \cdots + X_N}{N} - \frac{r}{t}| > \frac{2}{t}\right).$$

In other words, the probability the sampled ratio from a binomial distribution is contained within the bounds $\frac{r-1}{t}$ to $\frac{r+1}{t}$ is at least c times more likely than the the probability it is outside the bounds

Thus, by taking enough samples from a binomial, "we determine the [parameter] *a posteriori* almost as though it was known to us a prior".

Bernoulli, additionally, derived lower bounds, given r and t, for how many samples would be needed to achieve a desired levels of accuracy. For example, if $r = 30$ and $t = 50$, he showed

[1] Suppose two players A and B each contribute an equal amount of money into a prize pot. A and B then agree to play repeated rounds of a game of chance, with the players having an equal probability of winning any round, until one of the players has won k rounds. The player that first reaches k wins takes the entirety of the prize pot. Now, suppose the game is interrupted with neither player reaching k wins. If A has w_A wins and B has w_B wins, what's a fair way to split the pot?

> having made 25550 experiments, it will be more than a thousand times more likely that the ratio of the number of obtained fertile observations to their total number is contained within the limits 31/50 and 29/50 rather than beyond them (Bernoulli, 1713, p. 30)

This suggested an approach to inference, but it came up short in several respects. 1) The bounds derived were conditional on knowledge of the true parameter. It didn't provide a way to quantify uncertainty when the parameter was unknown. And 2) the number of experiments required to reach a high level of confidence in an estimate, moral certainty in Bernoulli's words, was quite large, limiting the approach's practicality. Abraham de Moivre would later improve on Bernoulli's work in his highly popular textbook *The Doctrine of Chances*. He derive considerably tighter bounds, but again failed to provide a way to quantify uncertainty when the binomial distribution's parameter was unknown, offering only this qualitative guidance

"*moral certain* is that whose probability is almost equal to complete certainty so that the difference is insensible." (Bernoulli, 1713, p. 9)

> if after taking a great number of Experiments, it should be perceived that the happenings and failings have been nearly in a certain proportion, such as of 2 to 1, it may safely be concluded that the Probabilities of happening or failing at any one time assigned will be very near that proportion, and that the greater the number of Experiments has been, so much nearer the Truth will the conjectures be that are derived from them. (De Moivre, 1756, p. 242)

INSPIRED BY de Moivre's book, Thomas Bayes took up the problem of inference with the binomial distribution. He reframed the goal to

> Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named. (Bayes, 1763, p. 4)

Recognizing that a solution would depend on prior probability, Bayes sought to give an answer for

> the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it (Bayes, 1763, p. 11)

He reasoned (incorrectly) that knowing nothing was equivalent to a uniform prior distribution [2]. Using the uniform prior and a geometric analogy with balls, Bayes succeeded in approximating integrals of posterior distributions of the form

$$\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \int_a^b \theta^y (1-\theta)^{n-y} d\theta$$

and was able to answer questions like "if I observe y success and $n - y$ failures from a binomial distribution with unknown parameter θ, what is the probability that θ is between a and b?".

[2] See (Stigler, 1990, p. 184–188) for a detailed account of Bayes' reasoning.

Despite Bayes' success answering inferential questions, his method was not widely adopted and his work, published posthumously in 1763, remained obscure up until De Morgan renewed attention to it over fifty years later. A major obstacle was Bayes' geometric treatment of integration; as mathematical historian Stephen Stigler writes,

> Bayes essay 'Towards solving a problem in the doctrine of chances' is extremely difficult to read today–even when we know what to look for. (Stigler, 1990, p. 179)

A DECADE after Bayes' death and likely unaware of his discoveries, Laplace pursued similar problems and independently arrive at the same approach. Laplace revisited the famous problem of the points, but this time considered the case of a skilled game where the probability of a player winning a round was modeled by a Bernoulli distribution with unknown parameter p. Like Bayes, Laplace assumed a uniform prior, noting only

> because the probability that A will win a point is unknown, we may suppose it to be any unspecified number whatever between 0 and 1. (Laplace, 1774)

Unlike Bayes, though, Laplace did not use a geometric approach. He approached the problems with a much more developed analytical toolbox and was able to derive more usable formulas with integrals and clearer notation.

Following Laplace and up until the early 20th century, using a uniform prior together with Bayes' theorem became a popular approach to statistical inference. In 1837, De Morgan introduced the term *inverse probability* to refer to such methods and acknowledged Bayes' earlier work

> De Moivre, nevertheless, did not discover the inverse method. This was first used by the Rev. T. Bayes, in Phil. Trans. liii. 370.; and the author, though now almost forgotten, deserves the most honourable remembrance from all who read the history of this science. (De Morgan, 1838, p. vii)

IN THE EARLY 20TH CENTURY, inverse probability came under serious attack for its use of a uniform prior. Ronald Fisher, one of the fiercest critics, wrote

> I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet that is quite exactly the position in respect of inverse probability (Fisher, 1930)

Fisher was not the first to criticize inverse probability, and he references the earlier works of Boole, Venn, and Chrystal. See Zabell (1989) for a detailed account of inverse probability criticism leading up to Fisher.

Fisher criticized inverse probability as "extremely arbitrary". Reviewing Bayes' essay, he pointed out how naive use of a uniform prior leads to solutions that depend on the scale used to measure probability. He gave a concrete example (Fisher, 1922): Let p denote the unknown parameter for a binomial distribution. Suppose that instead of p we parameterize by

$$\theta = \arcsin\left(2p - 1\right), \quad -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2},$$

and apply the uniform prior. Then the probability that θ is between a and b after observing S successes and F failures is

$$\frac{1}{\pi}\int_a^b \left(\frac{\sin\theta + 1}{2}\right)^S \left(\frac{1 - \sin\theta}{2}\right)^F d\theta.$$

A change of variables back to p shows us this is equivalent to

$$\frac{1}{\pi}\int_{(\sin a+1)/2}^{(\sin b+1)/2} \left(p\right)^{S-1/2} \left(1 - p\right)^{F-1/2} dp.$$

Hence, the uniform prior in θ is equivalent to the prior $\frac{1}{\pi}p^{-1/2}(1 - p)^{-1/2}$ in p. As an alternative to inverse probability, Fisher promoted maximum likelihood methods, p-values, and a frequentist definition for probability.

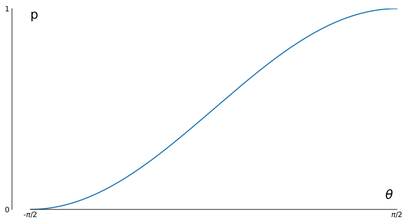

Figure 1: Fisher's alternate parameterization for the probability of the binomial distribution.

While Fisher and others advocated for abandoning inverse probability in favor of frequentist methods, Harold Jeffreys worked to put inverse probability on a firmer foundation. He acknowledged previous approaches to inverse probability had lacked consistency, but he agreed with their goal of delivering statistical results in terms of degree of belief and thought frequentist definitions of probability to be hopelessly flawed:

> frequentist definitions themselves lead to no results of the kind that we need until the notion of reasonable degree of belief is reintroduced, and that since the whole purpose of these definitions is to avoid this notion they necessarily fail in their object. (Jeffreys, 1961, p. 34)

Jeffreys pointed out that inverse probability needn't be tied to the uniform prior:

> There is no more need for [the idea that the uniform distribution of the prior probability was a necessary part of the principle of inverse probability] than there is to say that an oven that has once cooked roast beef can never cook anything but roast beef. (Jeffreys, 1961, p. 103)

Seeking to achieve results that would be consistent under reparameterizations, Jeffreys proposed priors based on the Fisher information matrix,

$$\pi(\boldsymbol{\theta}) \propto |\mathcal{I}(\boldsymbol{\theta})|^{1/2}$$
$$\mathcal{I}(\boldsymbol{\theta})_{st} = \mathbb{E}_{\mathbf{y}} \left\{ \left(\frac{\partial}{\partial \theta_s} \log P(\mathbf{y} \mid \boldsymbol{\theta}) \right) \left(\frac{\partial}{\partial \theta_t} \log P(\mathbf{y} \mid \boldsymbol{\theta}) \right) \mid \boldsymbol{\theta} \right\},$$

writing [3]

> If we took the prior probability density for the parameters to be proportional to $[|\mathcal{I}(\boldsymbol{\theta})|^{1/2}]$ ... any arbitrariness in the choice of the parameters could make no difference to the results, and it is proved that for this wide class of laws a consistent theory of probability can be constructed. (Jeffreys, 1961, p. 159)

Twenty years later, Welch and Peers (1963) investigated priors from a different perspective. They analyzed one-tailed credible sets from posterior distributions and asked how closely probability mass coverage matched frequentist coverage. They found that for the case of a single parameter [4], the prior Jeffreys proposed was asymptotically optimal, providing further justification for the prior that aligned with how intuition suggests we might quantify Bayes criterion of "knowing absolutely nothing".

In an unfortunate turn of events, mainstream statistics mostly ignored Jeffreys approach to inverse probability to chase a mirage [5] of objectivity that frequentist methods seemed to provide [6] ; but much as Jeffreys had anticipated with his criticism that frequentist definitions of probability couldn't provide "results of the kind that we need", a majority of practitioners filled in the blank by misinterpreting frequentist results as providing belief probabilities. Goodman (1999) introduced the term P-value fallacy to refer to this common error and described just how prevalent it is

> In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

James Berger and Thomas Sellke established theoretical and simulation results that show how spectacularly wrong this notion is

> it is shown that actual evidence against a null (as measured, say, by posterior probability or comparative likelihood) can differ by an *order of magnitude* from the P value. For instance, data that yield a P value of .05, when testing a normal mean, result in a posterior probability of the null of *at least* .30 for *any* objective prior distribution. (Berger and Sellke, 1987)

They concluded

[3] Unlike the uniform prior, Jeffreys prior is invariant to reparameterization. If Θ denotes a region of the parameter space and $\boldsymbol{\varphi}(\mathbf{u})$ is an injective continuous function whose range includes Θ, then applying the change-of-variables formula will show that

$$\int_{\Theta} P(\mathbf{y} \mid \boldsymbol{\theta}) |\mathcal{I}(\boldsymbol{\theta})|^{1/2} d\boldsymbol{\theta} =$$
$$\int_{\boldsymbol{\varphi}^{-1}(\Theta)} P(\mathbf{y} \mid \boldsymbol{\varphi}(\mathbf{u})) |\mathcal{I}^{\boldsymbol{\varphi}}(\mathbf{u}))|^{1/2} d\mathbf{u},$$

where $\mathcal{I}^{\boldsymbol{\varphi}}$ denotes the Fisher information with respect to the reparameterization.

[4] Deriving good priors in the multi-parameter case is considerably more involved. Jeffreys himself was dissatisfied with the prior his rule produced for multi-parameter models and proposed an alternative known as *Jeffreys independent prior* but never developed a rigorous approach. José-Miguel Bernardo and James Berger would later develop *reference priors* as a refinement of Jeffreys prior. Reference priors provide a general mechanism to produce good priors that works for multi-parameter models and cases where the Fisher information matrix doesn't exist. See (Berger et al., 2009) and (Berger et al., 2024, part. 3).

[5] see Question 4 in Discussion

[6] Development of inverse probability in the manner Jeffreys suggested would continue under the name *Objective Bayesian Analysis*; but it hardly occupies the center stage of statistics, and many people mistakenly think of Bayesian analysis as more of a subjective theory.

> for testing "precise" hypotheses, p values should not be used directly, because they are too easily misinterpreted. The standard approach in teaching–of stressing the formal definition of a p value while warning against its misinterpretation–has simply been an abysmal failure.
> (Selke et al., 2001)

In this paper, we'll look closer at how priors for inverse probabilities can be justified by matching coverage; and we'll reexamine the problems Bayes and Laplace contemplated to see how they might be solved with a more modern approach.

Priors and Frequentist Matching

The idea of matching priors intuitively aligns with how we might think about probability in the absence of prior knowledge. We can think of the frequentist coverage matching metric as a way to provide an answer to the question "How accurate are the Bayesian credible sets produced with a given prior?".

For more background on frequentist coverage matching and its relation to inverse probability, see Berger et al. (2022) and (Berger et al., 2024, ch. 5).

Consider a probability model with a single parameter θ. If we're given a prior, $\pi(\theta)$, how do we test if the prior reasonably expresses Bayes' requirement of knowing nothing? Let's pick a size n, a value θ_{true}, and randomly sample observations $\mathbf{y} = (y_1, \ldots, y_n)^\top$ from the distribution $P(\cdot|\theta_{\text{true}})$. Then let's compute the two-tailed credible set $[\theta_a, \theta_b]$ that contains 95% of the probability mass of the posterior,

$$\pi(\theta \mid \mathbf{y}) \propto P(\mathbf{y} \mid \theta) \times \pi(\theta),$$

and record whether or not the credible set contains θ_{true}. Now suppose we repeat the experiment many times and vary n and θ_{true}. If $\pi(\theta)$ is a good prior, then the fraction of trials where θ_{true} is contained within the credible set will consistently be close to 95%. [7]

Example 1 *Suppose we observe n normally distributed values,* $\mathbf{y}$, *with variance* 1 *and unknown mean,* μ. *Let's consider the prior*

$$\pi(\mu) \propto 1.$$

(Note: In this case Jeffreys prior and the constant prior in μ *are the same.) Then*

$$\begin{aligned} P(\mathbf{y} \mid \mu) &\propto \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \mu\mathbf{1}\right)'\left(\mathbf{y} - \mu\mathbf{1}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(n\mu^2 - 2\mu n\bar{y}\right)\right\} \\ &\propto \exp\left\{-\frac{n}{2}\left(\mu - \bar{y}\right)^2\right\}. \end{aligned}$$

[7] The coverage experiment expressed as an algorithm:

```
function COVERAGE-TEST($\theta_{\text{true}}, \alpha$)
    $cnt \leftarrow 0$
    $N \leftarrow$ a large number
    for $i \leftarrow 1$ to $N$ do
        $\mathbf{y} \leftarrow$ sample from $P(\cdot \mid \theta_{\text{true}})$
        $t \leftarrow \int_{-\infty}^{\theta_{\text{true}}} \pi(\theta \mid \mathbf{y}) d\theta$
        if $\frac{1-\alpha}{2} < t < 1 - \frac{1-\alpha}{2}$ then
            $cnt \leftarrow cnt + 1$
        end if
    end for
    return $\frac{cnt}{N}$
end function
```

Thus,

$$\int_{-\infty}^{t} \pi(\mu \mid \mathbf{y}) d\mu = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{t-\bar{y}}{\sqrt{2/n}}\right)\right].$$

I ran a 95% *coverage test with* 10000 *trials and various values of μ and n. As Table 5 shows the results are all close to* 95%*, indicating the constant prior is a good choice in this case. [source code for experiment]*

μ_{true}	$n=5$	$n=10$	$n=20$
0.1	0.9502	0.9486	0.9485
0.5	0.9519	0.9478	0.9487
1.0	0.9516	0.9495	0.9519
2.0	0.9514	0.9521	0.9512
5.0	0.9489	0.9455	0.9497

Table 1: Frequentist coverages for the mean of a normal distribution with known variance and constant prior. Values close to 0.95 indicate a good prior.

Example 2 *Now suppose we observe n normally distributed values, $\mathbf{y}$, with unknown variance and zero mean, μ. Let's test the constant prior and Jeffreys' prior,*

$$\pi_C(\sigma^2) \propto 1 \quad \textit{and} \quad \pi_J(\sigma^2) \propto \frac{1}{\sigma^2}.$$

We have

$$P(\mathbf{y} \mid \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\}$$

where $s^2 = \frac{\mathbf{y}'\mathbf{y}}{n}$. Put $u = \frac{ns^2}{2\sigma^2}$. Then

$$\begin{aligned}\int_0^t \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{ns^2}{2\sigma^2}\right\} d\sigma^2 &\propto \int_{\frac{ns^2}{2t}}^{\infty} u^{n/2-2} \exp\{-u\}\, du \\ &= \Gamma(\frac{n-2}{2}, \frac{ns^2}{2t}).\end{aligned}$$

$\Gamma(\cdot)$ denotes the Gamma function,

$$\Gamma(s) = \int_0^\infty t^{s-1} \exp(-t) dt,$$

and $\Gamma(\cdot,\cdot)$ denotes the incomplete Gamma function,

$$\Gamma(s,x) = \int_x^\infty t^{s-1} \exp(-t) dt.$$

Thus,

$$\int_0^t \pi_C(\sigma^2 \mid \mathbf{y}) d\sigma^2 = \frac{1}{\Gamma(\frac{n-2}{2})} \Gamma(\frac{n-2}{2}, \frac{ns^2}{2t}).$$

Similarly,

$$\int_0^t \pi_J(\sigma^2 \mid \mathbf{y}) d\sigma^2 = \frac{1}{\Gamma(\frac{n}{2})} \Gamma(\frac{n}{2}, \frac{ns^2}{2t}).$$

Table 2 shows the coverage results for the constant prior. We can see that for smaller values of σ^2_{true} and n, the coverages are notably smaller than 95%*. In comparison, Jeffreys prior (Table 3) performs well for all values of σ^2_{true} and n. [source code for experiment]*

σ^2_{true}	$n=5$	$n=10$	$n=20$
0.1	0.9014	0.9288	0.9445
0.5	0.9035	0.9309	0.9429
1.0	0.9048	0.9303	0.9417
2.0	0.9079	0.9331	0.9418
5.0	0.9023	0.9295	0.9433

Table 2: Frequentist coverages for the variance of a zero-mean normal distribution with the constant prior.

σ^2_{true}	$n=5$	$n=10$	$n=20$
0.1	0.9516	0.9503	0.9533
0.5	0.9501	0.9490	0.9537
1.0	0.9505	0.9511	0.9519
2.0	0.9480	0.9514	0.9498
5.0	0.9506	0.9497	0.9507

Table 3: Frequentist coverages for the variance of a zero-mean normal distribution with Jeffreys prior.

The Binomial Distribution

Let's revisit the binomial distribution with a modern approach to inverse probability.

SUPPOSE WE OBSERVE n values from the binomial distribution. Let y denote the number of successes and θ denote the probability of success. The likelihood function is given by

$$L(\theta; \mathbf{y}) \propto \theta^y (1-\theta)^{n-y}.$$

Taking the log and differentiating, we have

$$\begin{aligned}\frac{\partial}{\partial\theta}\log L(\theta;\mathbf{y}) &= \frac{y}{\theta} - \frac{n-y}{1-\theta}\\ &= \frac{y-n\theta}{\theta(1-\theta)}.\end{aligned}$$

Thus, the Fisher information for the binomial distribution is [8]

$$\begin{aligned}\mathcal{I}(\theta) &= \mathbb{E}_y\left\{\left(\frac{\partial}{\partial\theta}\log L(\theta;\mathbf{y})\right)^2 \mid \theta\right\}\\ &= \mathbb{E}_y\left\{\left(\frac{y-n\theta}{\theta(1-\theta)}\right)^2 \mid \theta\right\}\\ &= \frac{n\theta(1-\theta)}{\theta^2(1-\theta)^2}\\ &= \frac{n}{\theta(1-\theta)},\end{aligned}$$

and Jeffreys prior is

$$\begin{aligned}\pi(\theta) &\propto \mathcal{I}(\theta)^{1/2}\\ &\propto \theta^{-1/2}(1-\theta)^{-1/2}.\end{aligned}$$

Normalizing gives us

$$\pi(\theta) = \frac{1}{\pi}\theta^{-1/2}(1-\theta)^{-1/2}.$$

The posterior is then

$$\pi(\theta \mid \mathbf{y}) \propto \theta^{y-1/2}(1-\theta)^{n-y-1/2},$$

which we can recognize as the beta distribution with parameters $y+1/2$ and $n-y+1/2$.

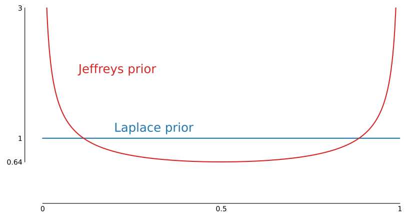

Figure 2: Jeffreys prior for binomial distribution together with the uniform prior. We can see that Jeffreys prior distributes more probability mass towards the extremes 0 and 1.

To TEST FREQUENTIST coverages, we can use an exact algorithm.

```
function BINOMIAL-COVERAGE-TEST($n$, $\theta_{\text{true}}$, $\alpha$)
    $cov \leftarrow 0$
    for $y \leftarrow 0$ to $n$ do
        $t \leftarrow \int_0^{\theta_{\text{true}}} \pi(\theta \mid y)d\theta$
        if $\frac{1-\alpha}{2} < t < 1 - \frac{1-\alpha}{2}$ then
            $cov \leftarrow cov + \binom{n}{y}\theta_{\text{true}}^{y}(1-\theta_{\text{true}})^{n-y}$
        end if
    end for
    return $cov$
end function
```

The tables below show frequentist coverages for the Bayes-Laplace uniform prior (left) and Jeffreys prior (right) using various values of

[8] Here we apply this formula for the mean of the binomial distribution, $n\theta$, and the variance, $n\theta(1-\theta)$.

n and θ_{true}.

	Coverges with Laplace Prior					Coverges with Jeffreys Prior			
θ_{true}	$n = 5$	$n = 10$	$n = 20$	$n = 100$	θ_{true}	$n = 5$	$n = 10$	$n = 20$	$n = 100$
0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.9995	0.9990	0.9980	0.9900
0.0010	0.0000	0.0000	0.0000	0.9048	0.0010	0.9950	0.9900	0.9802	0.9048
0.0100	0.9510	0.9044	0.8179	0.9206	0.0100	0.9510	0.9044	0.9831	0.9816
0.1000	0.9185	0.9298	0.9568	0.9364	0.1000	0.9914	0.9872	0.9568	0.9557
0.2500	0.9844	0.9803	0.9348	0.9513	0.2500	0.9844	0.9240	0.9348	0.9513
0.5000	0.9375	0.9785	0.9586	0.9431	0.5000	0.9375	0.9785	0.9586	0.9431

We can see coverage is identical for many table entries. For smaller values of n and θ_{true}, though, the uniform prior gives no coverage while Jeffreys prior provides decent results. [source code for experiment]

Applications From Bayes and Laplace

Let's now revisit some applications Bayes and Laplace studied. Given that the goal in all of these problems is to assign a belief probability to an interval, I think that we can make a strong argument that Jeffreys prior is a better choice than the uniform prior since it has asymptotically optimal frequentist coverage performance. This also addresses Fisher's criticism of arbitrariness.

In each of these problems, I'll show both the answer given by Jeffreys prior and the original uniform prior that Bayes and Laplace used. One theme we'll see is that many of the results are not that different. A lot of fuss is often made over minor differences in how objective priors can be derived, and they can be important; but often the data dominates and different reasonable choices will lead to nearly the same result.

Example 3 *In an appendix Richard Price added to Bayes' essay, he considers the following problem:*

> *Let us then first suppose, of such an event as that called M in the essay, or an event about the probability of which, antecedently to trials, we know nothing, that it has happened once, and that it is enquired what conclusion we may draw from hence with respct to the probability of it's happening on a second trial. (Bayes, 1763, p. 16)*

TL;DR: Suppose we observe the value 1 from a Bernoulli distribution with unknown parameter θ. What can we say about the value of θ?

Specifically, Price asks, what's the probability that θ is greater than $\frac{1}{2}$?. Using the uniform prior in Bayes' essay, we derive the posterior distribution

$$\pi_B(\theta \mid y) = 2x.$$

Integrating gives us the answer

$$\int_{\frac{1}{2}}^{1} 2x dx = x^2\Big|_{\frac{1}{2}}^{1} = \frac{3}{4}.$$

Using Jeffreys prior, we derive a beta distribution for the posterior,

$$\pi_J(\theta \mid y) = \frac{\Gamma(2)}{\Gamma(3/2)\Gamma(1/2)}\theta^{1/2}(1-\theta)^{-1/2},$$

and the answer

$$\int_{\frac{1}{2}}^{1} \pi_J(\theta \mid y)dx = \frac{1}{2} + \frac{1}{\pi} \approx 0.818$$

Price then continues with the same problem but supposes we see two 1*s, three* 1*s, etc. To the right, I show the result for Bayes prior and Jeffreys prior up to ten* 1*s. [source code for experiment]*

1s observed	$\int_{\frac{1}{2}}^{1} \pi_B(\theta \mid y)d\theta$	$\int_{\frac{1}{2}}^{1} \pi_J(\theta \mid y)d\theta$
1	0.7500	0.8183
2	0.8750	0.9244
3	0.9375	0.9669
4	0.9688	0.9850
5	0.9844	0.9931
6	0.9922	0.9968
7	0.9961	0.9985
8	0.9980	0.9993
9	0.9990	0.9997
10	0.9995	0.9998

Example 4 *Price also considers a lottery with an unknown chance of winning:*

> *Let us then imagine a person present at the drawing of a lottery, who knows nothing of its scheme or of the proportion of Blanks to Prizes in it. Let it further be supposed, that he is obliged to infer this from the number of blanks he hears drawn compared with the number of prizes; and that it is enquired what conclusions in these circumstances he may reasonably make. (Bayes, 1763, p. 19–20)*

He asks this specific question:

> *Let him first hear ten blanks drawn and one prize, and let it be enquired what chance he will have for being right if he gussses that the proportion of blanks to prizes in the lottery lies somewhere between the proportions of 9 to 1 and 11 to 1. (Bayes, 1763, p. 20)*

With Bayes prior and θ representing the probability of drawing a blank, we derive the posterior distribution

$$\pi_B(\theta \mid y) = \frac{\Gamma(13)}{\Gamma(11)\Gamma(2)}\theta^{10}(1-\theta)^{1},$$

and the answer

$$\int_{\frac{9}{10}}^{\frac{11}{12}} \pi_B(\theta \mid y)d\theta \approx 0.0770.$$

Using Jeffreys prior, we get the posterior

$$\pi_J(\theta \mid y) = \frac{\Gamma(12)}{\Gamma(21/2)\Gamma(3/2)}\theta^{19/2}(1-\theta)^{1/2}$$

and the answer

$$\int_{\frac{9}{10}}^{\frac{11}{12}} \pi_J(\theta \mid y)d\theta \approx 0.0804.$$

Price then considers the same question (what's the probability that θ lies between $\frac{9}{10}$ and $\frac{11}{12}$) for difference cases where an observer of the lottery sees w prizes drawn and $10 \times w$ blanks. Table 4 shows some of the possible results. [source code for experiment]

Blanks	Prizes	$\int_{\frac{9}{10}}^{\frac{11}{12}} \pi_B(\theta \mid y)d\theta$	$\int_{\frac{9}{10}}^{\frac{11}{12}} \pi_J(\theta \mid y)d\theta$
10	1	0.0770	0.0804
20	2	0.1084	0.1107
40	4	0.1527	0.1541
100	10	0.2390	0.2395
1000	100	0.6628	0.6618

Table 4: Bayesian credible sets for the probability of losing a lottery where various prizes and blanks are observed

Example 5 *Let's now turn to a problem that fascinated Laplace and his contemporaries: The relative birth rate of boys-to-girls. Laplace introduces the problem as follows:*

> *The consideration of the [influence of past events on the probability of future events] leads me to speak of births: as this matter is one of the most interesting in which we are able to apply the Calculus of probabilities, I manage so to treat with all care owing to its importance, by determining what is, in this case, the influence of the observed events on those which must take place, and how, by its multiplying, they uncover for us the true ratio of the possibilities of the births of a boy and of a girl. (Laplace, 1778, p. 1)*

Boys	Girls	$\int_0^{1/2} \pi_L(\theta \mid y) d\theta$	$\int_0^{1/2} \pi_J(\theta \mid y) d\theta$
0	0	0.5000	0.5000
749	751	0.5206	0.5206
1511	1489	0.3440	0.3440
2263	2237	0.3492	0.3492
3081	2919	0.0182	0.0182
3810	3690	0.0829	0.0829
4514	4486	0.3839	0.3839
5341	5159	0.0379	0.0379
6139	5861	0.0056	0.0056
6792	6708	0.2349	0.2349
7608	7392	0.0389	0.0389
8308	8192	0.1833	0.1832
9145	8855	0.0153	0.0153
9957	9543	0.0015	0.0015
10618	10382	0.0517	0.0517

Table 5: Probability that the birth rate of boys-to-girls is less than 0.5 for various samples drawn from a binomial distribution where the true parameter is set to $B/(B+G) \approx 0.5097$. This shows how Laplace's answer could evolve as more data is collected. I show probabilities using both Laplace's uniform prior and Jeffreys prior. We can see that the result is nearly the same

Like Bayes, Laplace approaches the problem using a uniform prior, writing

> *When we have nothing given a priori on the possibility of an event, it is necessary to assume all the possibilities, from zero to unity, equally probable; thus, observation can alone instruct us on the ratio of the births of boys and of girls, we must, considering the thing only in itself and setting aside the events, to assume the law of possibility of the births of a boy or of a girl constant from zero to unity, and to start from this hypothesis into the different problems that we can propose on this object. (Laplace, 1778, p. 26)*

Using data collection from Paris between 1745 and 1770, where 251527 boys and 241945 girls had been born, Laplace asks, what is "the probability that the possibility of the birth of a boy is equal or less than $\frac{1}{2}$"?

With a uniform prior, $B = 251527$, $G = 241945$, and θ representing the probability that a boy is born, we obtain the posterior

$$\pi_L(\theta \mid y) = \frac{\Gamma(B+G+2)}{\Gamma(B+1)\Gamma(G+1)} \theta^B (1-\theta)^G$$

and the answer [9]

$$\int_0^{1/2} \pi_L(\theta \mid y) d\theta \approx 1.1460 \times 10^{-42}.$$

With Jeffreys prior, we similarly derive the posterior

$$\pi_J(\theta \mid y) = \frac{\Gamma(B+G+1)}{\Gamma(B+1/2)\Gamma(G+1/2)} \theta^{B-1/2} (1-\theta)^{G-1/2}$$

and answer

$$\int_0^{1/2} \pi_J(\theta \mid y) d\theta \approx 1.1458 \times 10^{-42}.$$

[source code for experiment]

Discussion

Q1. *Where do inverse probabilities belong in statistics?*

[9] This matches up quite closely with the answer Laplace gets: "we will have, for the probability that x is equal or less than 21 , a fraction of which the numerator is little different from unity and equal to 1,1521, and of which the denominator is the seventh power of one million",

$$1.1521/(10^6)^7 = 1.152 \times 10^{-42}.$$

A1. I think Jeffreys was right and standard statistical procedures should deliver "results of the kind we need". While Bayes and Laplace might not have been fully justified in their choice of a uniform prior, they were correct in their objective of quantifying results in terms of degree of belief. Inverse probabilities of the kind Jeffreys outlined give us a pathway to provide "results of the kind we need" while addressing the arbitrariness of the Bayes-Laplace approach. Jeffreys approach isn't the only way to get to results as degrees of belief, and a more subjective approach can also be valid if the situation allows, but his approach give us good answers for the common situation "of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it" and can be used as a drop-in replacement for frequentist methods.

To answer more concretely, I think when you open up a standard introduction-to-statistics textbook and look up a basic procedure such as a hypothesis test of whether the mean of normally distributed data with unknown variance is non-zero, you should see a method built on objective priors and Bayes factor like Berger and Mortera (1999) rather than a method based on P values.

Q2. *But aren't there multiple ways of deriving good priors in the absence of prior knowledge?*

A2. I highlighted frequentist coverage matching as a benchmark to gauge whether a prior is a good candidate for objective analysis, but coverage matching isn't the only valid metric we could use and it may be possible to derive multiple priors with good coverage. Different priors with good frequentist properties, though, will likely be similar, and any results will be determined more by observations than the prior. If we are in a situation where multiple good priors lead to significantly differing results, then that's an indicator we need to provide subjective input to get a useful answer. Here's how Berger (2006) addresses this issue:

> Inventing a new criterion for finding "the optimal objective prior" has proven to be a popular research pastime, and the result is that many competing priors are now available for many situations. This multiplicity can be bewildering to the casual user.
>
> I have found the reference prior approach to be the most successful approach, sometimes complemented by invariance considerations as well as study of frequentist properties of resulting procedures. Through such considerations, a particular prior usually emerges as the clear winner in many scenarios, and can be put forth as the recommended objective prior for the situation.

Q3. *Doesn't that make inverse probability subjective, whereas frequentist methods provide an objective approach to statistics?*

A3. It's a common misconception that frequentist methods are objective. Berger and Berry (1988) provides this example to demonstrate: Suppose we watch a researcher study a coin for bias. We see the researcher flip the coin 17 times. Heads comes up 13 times and tails comes up 4 times. Suppose θ represents the probability of heads and the researcher is doing a standard P-value test with the null hypothesis that the coin is not bias, $\theta = 0.5$. What P-value would they get? We can't answer this question because the researcher would get remarkably different results depending on their experimental intentions. If their intention was to collect 17 sample coin flips, we would get a P-value of 0.049. But if their intention was to continue flipping the coin until at least 4 heads and 4 tails were observed, we would get a P-value of 0.021. The result is dependent on not just the data but also on the hidden intentions of the researcher. As Berger and Berry (1988) argue, "objectivity is not generally possible in statistics and ... standard statistical methods can produce misleading inferences". [source code for example]

If the experimeter's intention was to flip a coin 17 times, then the probability of seeing a value *less extreme* than 13 under the null hypothesis is given by summing binomial distribution terms representing the probabilities of getting 5 to 12 heads,

$$\binom{17}{5} \times 0.5^{17} = 0.047$$
$$\binom{17}{6} \times 0.5^{17} = 0.094$$
$$\binom{17}{7} \times 0.5^{17} = 0.148$$
$$\binom{17}{8} \times 0.5^{17} = 0.185$$
$$\binom{17}{9} \times 0.5^{17} = 0.185$$
$$\binom{17}{10} \times 0.5^{17} = 0.148$$
$$\binom{17}{11} \times 0.5^{17} = 0.094$$
$$\binom{17}{12} \times 0.5^{17} = 0.047,$$

which gives us 0.951 and hence a P-value of $1 - 0.951 = 0.049$

If, however, the experimenter's intention was to continue sampling until they got at least 4 heads and 4 tails, then the probability of seeing a value *less extreme* than 17 total flips under the null hypothesis is given by summing negative binomial distribution terms representing the probabilities of getting 8 to 16 total observations,

$$2\binom{7}{3} \times 0.5^{8} = 0.273$$
$$2\binom{8}{3} \times 0.5^{9} = 0.219$$
$$2\binom{9}{3} \times 0.5^{10} = 0.164$$
$$2\binom{10}{3} \times 0.5^{11} = 0.117$$
$$2\binom{11}{3} \times 0.5^{12} = 0.081$$
$$2\binom{12}{3} \times 0.5^{13} = 0.054$$
$$2\binom{13}{3} \times 0.5^{14} = 0.035$$
$$2\binom{14}{3} \times 0.5^{15} = 0.022$$
$$2\binom{15}{3} \times 0.5^{16} = 0.014,$$

which gives us 0.979 and a P-value of $1 - 0.978 = 0.021$

Q4. *If subjectivity is unavoidable, why not just use subjective priors?*

A4. When subjective input is possible, we should incorporate it. But we should also acknowledge that Bayes' "event concerning the probability of which we absolutely know nothing antecedently" is an important fundamental problem of inference that needs good solutions. As Edwin Jaynes writes

> To reject the question, [how do we find the prior representing "complete ignorance"?], as some have done, on the grounds that the state of complete ignorance does not "exist" would be just as absurd as to reject Euclidean geometry on the grounds that a physical point does not exist. In the study of inductive inference, the notion of complete ignorance intrudes itself into the theory just as naturally and inevitably as the concept of zero in arithmetic.
>
> If one rejects the consideration of complete ignorance on the grounds that the notion is vague and ill-defined, the reply is that the notion cannot be evaded in any full theory of inference. So if it is still ill-defined, then a major and immediate objective must be to find a precise definition which will agree with intuitive requirements and be of constructive use in a mathematical theory. Jaynes (1968)

Moreover, systematic approaches such as reference priors can certainly do much better than pseudo-Bayesian techniques such as choosing a uniform prior over a truncated parameter space or a

The term *pseudo-Bayesian* comes from Berger (2006). See that paper for a more detailed discussion.

> vague proper prior such as a Gaussian over a region of the parameter space that looks interesting. Even when subjective information is available, using reference priors as building blocks is often the best way to incorporate it. For instance, if we know that a parameter is restricted to a certain range but don't know anything more, we can simply adapt a reference prior by restricting and renormalizing it (Berger et al., 2024, p. 256)

Conclusion

The common and repeated misinterpretation of statistical results such as P values or confidence intervals as belief probabilities shows us that there is a strong natural tendency to want to think about inference in terms of inverse probability. [10] It's no wonder that the method dominated for 150 years.

Fisher and others were certainly correct to criticize naive use of a uniform prior as arbitrary, but this is largely addressed by reference priors and adopting metrics like frequentist matching coverage that quantify what it means for a prior to represent ignorance. As Berger puts it,

> We would argue that noninformative prior Bayesian analysis is the *single most powerful method of statistical analysis*, in the sense of being the *ad hoc* method most likely to yield a sensible answer for a given investment of effort. And the answers so obtained have the added feature of being, in some sense, the most "objective" statistical answers obtainable (Berger, 1985, p. 90)

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions of the Royal Society of London 53*, 370–418.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis 1*(3), 385–402.

Berger, J., J. Bernardo, and D. Sun (2022). Objective bayesian inference and its relationship to frequentism.

Berger, J., J. Bernardo, and D. Sun (2024). *Objective Bayesian Inference*. World Scientific.

[10] We don't have to look hard to find examples of the P value fallacy. Here's one I came across just the other day reading a book on the history of risk from a major publisher:

> Epidemiologists–the statisticians of health–observe the same convention as that used to measure the performance of investment managers. They usually define a result as statistically significant if there is no more than a 5% probability that an outcome was the result of chance. (Bernstein, 1998, p. 285)

In light of (Berger and Sellke, 1987) and (Selke et al., 2001), the statement would be more accurate if we replaced 5% with 28.9%.

Berger, J. and J. Mortera (1999). Default bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association 94*(446), 542–554.

Berger, J. and T. Sellke (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association 82*(397), 112–22.

Berger, J. O., J. M. Bernardo, and D. Sun (2009). The formal definition of reference priors. *The Annals of Statistics 37*(2), 905 – 938.

Berger, J. O. and D. A. Berry (1988). Statistical analysis and the illusion of objectivity. *American Scientist 76*(2), 159–165.

Bernoulli, J. (1713). *On the Law of Large Numbers, Part Four of Ars Conjectandi*. Translated by Oscar Sheynin.

Bernstein, P. (1998). *Against the Gods: The Remarkable Story of Risk*. Wiley.

Dale, A. (1991). *A History of Inverse Probability: From Thomas Bayes to Karl Pearson*. Springer-Verlag.

De Moivre, A. (1756). *The Doctrine of Chances*.

De Morgan, A. (1838). *An Essay On Probabilities: And On Their Application To Life Contingencies And Insurance Offices*.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222*, 309–368.

Fisher, R. (1930). Inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society 26*(4), 528–535.

Goodman, S. (1999, June). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine 130*(12), 995–1004.

Jaynes, E. T. (1968). Prior probabilities. *Ieee Transactions on Systems and Cybernetics* (3), 227–241.

Jeffreys, H. (1961). *Theory of Probability* (3 ed.). Oxford Classic Texts in the Physical Sciences.

Laplace, P. (1774). Memoir on the probability of the causes of events. Translated by S. M. Stigler.

Laplace, P. (1778). Mémoire sur les probabilités. Translated by Richard J. Pulskamp.

Selke, T., M. J. Bayarri, and J. Berger (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician 855*(1), 62–71.

Stigler, S. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press.

Welch, B. L. and H. W. Peers (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society Series B-methodological 25*, 318–329.

Zabell, S. (1989). R. A. Fisher on the History of Inverse Probability. *Statistical Science 4*(3), 247 – 256.