# PHIL 6334 - Probability/Statistics Lecture Notes 6: An Introduction to Bayesian Inference

Aris Spanos [Spring 2019]

# 1 Introduction to Bayesian Inference

The main objective is to introduce the reader to *Bayesian inference* by comparing and contrasting it to the *frequentist inference*. To avoid unnecessary technicalities the discussion focuses mainly on the simple Bernoulli model.

## 1.1 The Bayesian inference framework

**Bayesian inference** begins with a statistical model:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \text{for } \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^m,\ m<n,$$

where $f(\mathbf{x};\boldsymbol{\theta})$ is the *distribution of the sample* $\mathbf{X}:=(X_1,...,X_n)$, $\mathbb{R}^n_X$ is the sample space and $\Theta$ the parameter space. Bayesian inference modifies the frequentist inferential set up in two crucial respects:

(A) It views the unknown parameter(s) $\boldsymbol{\theta}$ as *random variables* with their own distribution, known as the **prior distribution**:

$$\pi(.):\ \Theta\rightarrow[0,1],$$

which represents one's *a priori* assessment of how likely the various values of $\boldsymbol{\theta}$ in $\Theta$ are, which amounts to *ranking* the different models $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, for all $\boldsymbol{\theta}\in\Theta$. In frequentist $\boldsymbol{\theta}$ is viewed as a set of unknown constants indexing $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}^n_X$.

(B) It re-interprets the distribution of the sample as **conditional** on the unknown parameters $\boldsymbol{\theta}$, and denoted by $f(\mathbf{x}|\boldsymbol{\theta})$.

Taken together these modifications imply that for Bayesians the *joint distribution* of the sample is now defined by:

$$f(\mathbf{x},\boldsymbol{\theta})=f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta,\ \forall\mathbf{x}\in\mathbb{R}^n_X,$$

where $\forall$ denotes 'for all'. In terms of the above distinguishing criteria:

[a] The Bayesian approach to statistical inference interprets probability as the **degrees of belief** [subjective, logical or rational].

[b] In the context of Bayesian inference, **relevant information** includes:

(i) the data $\mathbf{x}_0:=(x_1,x_2,...,x_n)$, and
(ii) the prior distribution $\pi(\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta$.

[c] The **primary aim** of the Bayesian approach is to revise the inital ranking $\pi(\boldsymbol{\theta})$ in light of the data $\mathbf{x}_0$, as summarized by $L(\boldsymbol{\theta}|\mathbf{x}_0)$, to derive the *updated ranking* in terms of the **posterior distribution**:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0)\ =\frac{f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}}f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}\propto L(\boldsymbol{\theta}|\mathbf{x}_0)\cdot\pi(\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\boldsymbol{\Theta},\tag{1}$$

where $L(\boldsymbol{\theta}|\mathbf{x}_0) \propto f(\mathbf{x}_0|\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\boldsymbol{\Theta}$, denotes the likelihood function, as *re-interpreted* by Bayesianism.

A famous Bayesian, Savage (1954) summarized Bayesian inference succinctly by:

'Inference means for us the change of opinion induced by evidence on the application of Bayes' theorem." (p. 178)

O'Hagan (1994) is more specific:

"Having obtained the posterior density $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$, the final step of the Bayesian method is to derive from it suitable inference statements. The most usual inference question is this: After seeing the data $\mathbf{x}_0$, what do we now know about the parameter $\boldsymbol{\theta}$. The only answer to this question is to present the entire posterior distribution." (p. 6)

In this sense, learning from data in the context of the Bayesian perspective pertains to how the original beliefs $\pi(\boldsymbol{\theta})$ being revised in light of data $\mathbf{x}_0$, the **revision** coming in the form of the posterior: $\pi(\boldsymbol{\theta}|\mathbf{x}_0),\ \forall\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

According to O'Hagan (1994):

"The objective [of Bayesian inference] is to extract information concerning $\boldsymbol{\theta}$ from the posterior distribution, and to present it helpfully via effective summaries. There are two criteria in this process. The first is to identify interesting features of the posterior distribution. ... The second criterion is good communication. Summaries should be chosen to convey clearly and succinctly all the features of interest." (p. 14)

## 1.2 The Bayesian approach to statistical inference

In order to avoid any misleading impressions it is important to note that there are numerous variants of Bayesianism; more than 46656 varieties of Bayesianism according to Good (1971)! In this section we discuss some of the elements of the Bayesian approach which are shared by most variants of Bayesianism.

Bayesian inference, like frequentist inference, begins with a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, but modifies the inferential set up in two crucial respects:

(i) the unknown parameter(s) $\boldsymbol{\theta}$ are now viewed as *random variables* (not unknown constants) with their own distribution, known as the *prior distribution*:

$$\pi(.):\ \Theta \to [0,1],$$

which represents the modeler's assessment of how likely the various values of $\boldsymbol{\theta}$ in $\Theta$ are *a priori*, and

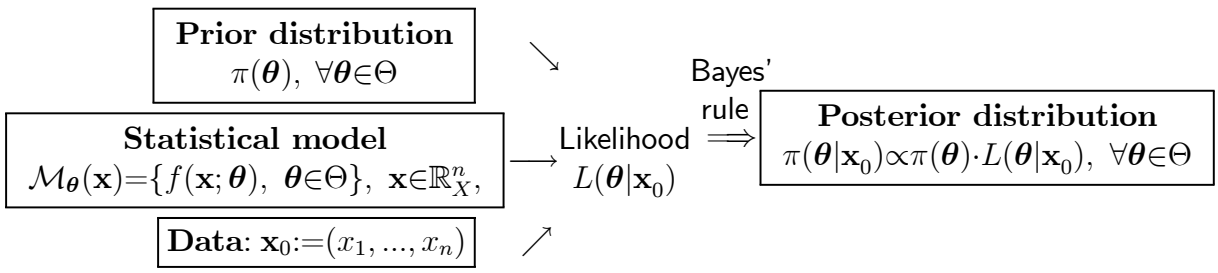
(ii) the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$ is re-interpreted by Bayesians to be defined as *conditional* on $\boldsymbol{\theta}$, and denoted by $f(\mathbf{x}|\boldsymbol{\theta})$.

Taken together these modifications imply that there exists a *joint distribution* relating the unknown parameters $\boldsymbol{\theta}$ and a sample realization $\mathbf{x}$:

$$f(\mathbf{x},\boldsymbol{\theta}){=}f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta.$$

Bayesian inference is based exclusively on the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ which is viewed as the revised (from the initial $\pi(\boldsymbol{\theta})$) *degrees of belief* for different values of $\boldsymbol{\theta}$ in light of the summary of the data by $L(\boldsymbol{\theta}|\mathbf{x}_0)$.

**Table 1: The Bayesian approach to statistical inference**




**Example 10.9**. Consider the *simple Bernoulli model* (table 10.2), and let the **prior** $\pi(\theta)$ be $\mathsf{Beta}(\alpha,\beta)$ distributed with a density function:

$$\pi(\theta)=\tfrac{1}{\mathsf{B}(\alpha,\beta)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1},\ \alpha>0,\ \beta>0,\ 0<\theta<1. \tag{2}$$

Combining the likelihood in (**??**) with the prior in (2) yields the **posterior** distribution:

$$\begin{aligned}\pi(\theta|\mathbf{x}_0)\ &\propto\left(\tfrac{1}{\mathsf{B}(\alpha,\beta)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1}\right)\left[\theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})}\right]=\\ &=\tfrac{1}{\mathsf{B}(\alpha,\beta)}\left[\theta^{n\overline{x}+(\alpha-1)}(1-\theta)^{n(1-\overline{x})+\beta-1}\right].\end{aligned} \tag{3}$$

In view of the formula in (2), (3) as an 'non-normalized' density of a $\mathsf{Beta}(\alpha^*,\beta^*)$, where:

$$\alpha^*=n\overline{x}+\alpha,\quad \beta^*=n(1-\overline{x})+\beta. \tag{4}$$

As the reader might have suspected, the choice of the prior in this case was not arbitrary. The Beta prior in conjunction with a Binomial-type LF gives rise to a Beta posterior. This is known in Bayesian terminology as a *conjugate pair*, where $\pi(\theta)$ and $\pi(\theta|\mathbf{x}_0)$ belong to the same family of distributions.

Savage (1954), one of the high priests of modern Bayesian statistics, summarizes Bayesian inference succinctly by asserting that: 'Inference means for us the change of opinion induced by evidence on the application of Bayes' theorem." (p. 178).

In the terms of the main grounds stated above, the Bayesian approach:

[a] Adopts the degrees of belief interpretation of probability introduced via $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$.

[b] The relevant information includes both (i) the data $\mathbf{x}_0:=(x_1,x_2,...,x_n)$, and (ii) prior information. Such prior information comes in the form of a prior distribution $\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$, which is assigned *a priori* and represents one's degree of belief in ranking the different values of $\boldsymbol{\theta}$ in $\Theta$ as more probable and less probable.

[c] The primary aim of the Bayesian approach is to **revise** the original *ranking* based on $\pi(\boldsymbol{\theta})$ in light of the data $\mathbf{x}_0$ by updating in the form of the *posterior distribution*:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0)=\tfrac{f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}}f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}\propto L(\boldsymbol{\theta}|\mathbf{x}_0)\cdot\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\boldsymbol{\Theta}, \tag{5}$$

where $L(\boldsymbol{\theta}|\mathbf{x}_0) \propto f(\mathbf{x}_0|\boldsymbol{\theta})$ denotes a *re-interpreted* likelihood function as being conditional on $\mathbf{x}_0$. The Bayesian approach is depicted in table 10.8. Since the denominator $m(\mathbf{x}_0)=\int_{\theta\in(0,1)} \pi(\boldsymbol{\theta})f(\mathbf{x}_0|\boldsymbol{\theta})d\boldsymbol{\theta}$, known as the **predictive** distribution, derived by integrating out $\boldsymbol{\theta}$, can be absorbed into the constant of proportionality in (5) and ignored for most practical purposes. The only exception to that is when one needs to treat $\pi(\theta|\mathbf{x}_0)$ as a proper density function which integrates to one, $m(\mathbf{x}_0)$ is needed as a normalizing constant.

**Learning from data**. In this context, learning from data $\mathbf{x}_0$ takes the form revising one's degree of belief for different values of $\boldsymbol{\theta}$ [i.e. different models $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta}\in\Theta$], in light of data $\mathbf{x}_0$, the learning taking the form $\pi(\boldsymbol{\theta}|\mathbf{x}_0)-\pi(\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$.. That is, the learning from data $\mathbf{x}_0$ about the phenomenon of interest takes place in the head of the modeler. In this sense, the underlying inductive reasoning is neither *factual* nor *hypothetical*, it's *all-inclusive* in nature: it pertains to *all* $\boldsymbol{\theta}$ in $\Theta$, as ranked by $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$. Hence, Bayesian inference does not pertain directly to the real world phenomenon of interest per se, but to one's beliefs about $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta}\in\Theta$.

### 1.2.1 The choice of prior distribution

Over the last two decades the focus of disagreement among Bayesians has been the choice of the prior. Although the original justification for using a prior is that it gives a modeler the opportunity incorporate substantive information into the data analysis, the discussions among Bayesians in the 1950s and 1960s made computational convenience the priority for the choice of a prior distribution, and that led to *conjugate priors* that ensure that the prior and the posterior distributions belong to the same family of distributions; see Berger (1985). More recently, discussions among Bayesians shifted the choice of the prior question to 'subjective' vs. 'objective' prior distributions. The concept of an 'objective' prior was pioneered by Jeffreys (1939) in an attempt to address Fisher's (1921) criticisms of Bayesian inference that routinely assumed a Uniform prior for $\theta$ as an expression of ignorance. Fisher's criticism was that if one assumes that a Uniform prior $\pi(\theta),\ \forall\theta\in\Theta$ expresses ignorance because all values of $\theta$ are assigned the same prior probability, then a reparameterization of $\theta$, say $\phi=h(\theta)$, will give rise to a *very informative* prior for $\phi$.

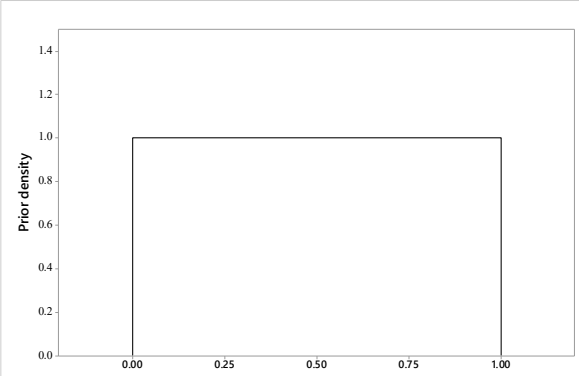



Fig. 10.1: Uniform prior density of $\theta$

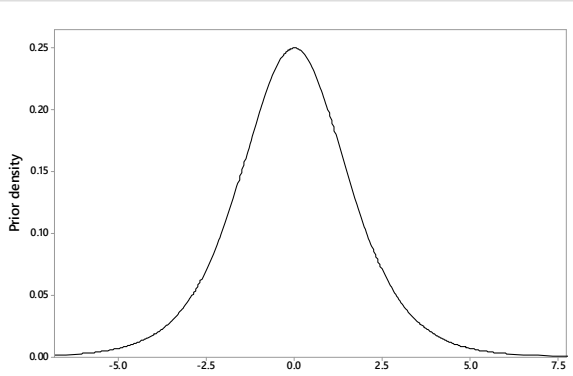



Fig. 10.2: Logistic prior density of $\phi$

**Example 10.10**. In the context of the simple Bernoulli model (table 10.2), let the prior be $\theta \backsim \mathsf{U}(0,1)$, $0 \leq \theta \leq 1$ (figure 10.1). Note that $\mathsf{U}(0,1)$ is a special case of the $\mathsf{Beta}(\alpha,\beta)$, for $\alpha{=}\beta{=}1$.

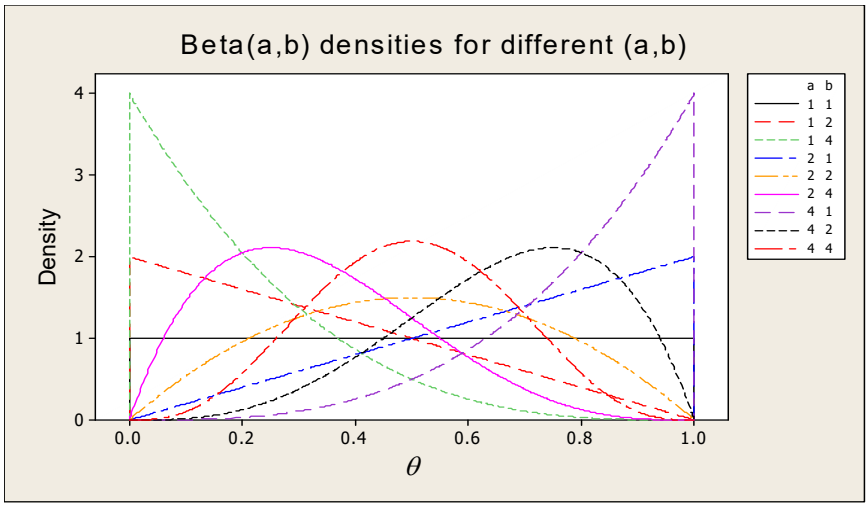



Fig. 10: $\mathsf{Beta}(\alpha,\beta)$ for different values of $(\alpha,\beta)$

Reparameterizing $\theta$ into $\phi{=}\ln(\frac{\theta}{1-\theta})$, implies that $\phi \backsim \mathsf{Logistic}(0,1)$, $-\infty < \phi < \infty$. Looking at the prior for $\phi$ (figure 10.2) it becomes clear that the ignorance about $\theta$ has been transformed into substantial knowledge about the different values of $\phi$.

In his attempt to counter Fisher's criticism, Jeffreys proposed a form of prior distribution that was invariant to such transformations. To achieve the reparameterization invariance Jeffreys had to use Fisher's information associated with the score function and the Cramer-Rao lower bound; see chapters 11-12.

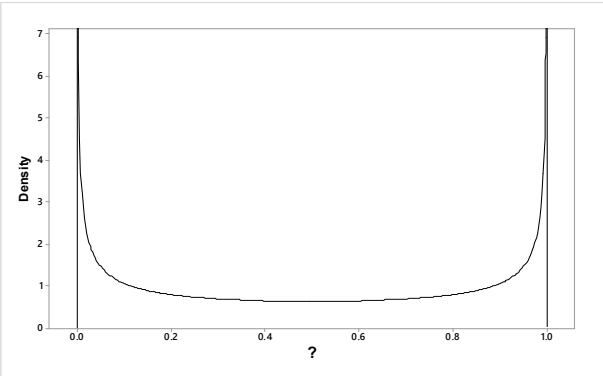



Fig. 10.3: Jeffreys prior

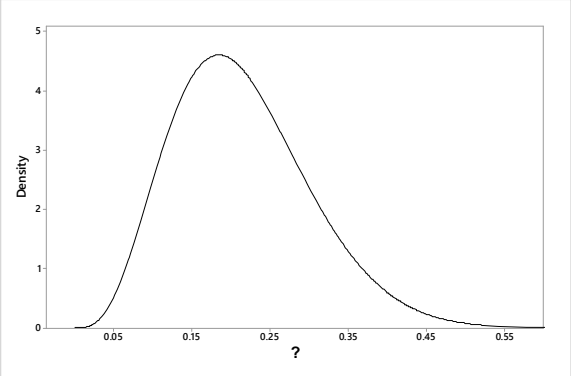



Fig. 10.4: Jeffreys posterior

**Example 10.11**. In the case of the simple Bernoulli model (10.2), the Jeffreys' prior is:

$$\pi_J(\theta) \backsim \mathsf{Beta}(.5,.5),\ \ \pi_J(\theta){=}\frac{\theta^{-.5}(1-\theta)^{-.5}}{\mathsf{B}(.5,.5)}$$

In light of the posterior in (3), using $n\overline{x}{=}4, n{=}20$ gives rise to:

$$\pi_J(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(4.5, 16.5).$$

For comparison purposes, the Uniform prior $\theta \backsim \mathsf{U}(0,1)$ in example 10.10 yields:

$$\pi(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(5,17).$$

Attempts to extend Jeffreys prior to models with more than one unknown parameter initiated a variant of Bayesianism that uses what is called *objective* (default, reference) priors because they minimize the role of the prior distribution and maximize the contribution of the likelihood function in deriving the posterior; see Berger (1985), Bernardo and Smith (1994).

## 1.3 Bayesian Estimation

Given the **posterior** distribution $\pi(\theta|\mathbf{x}_0)$ the Bayesian point estimator is often chosen to be its **mode** $\widetilde{\theta}_B$**:**

$$\pi(\widetilde{\theta}_B|\mathbf{x}_0) = \sup_{\theta\in\Theta} \pi(\theta|\mathbf{x}_0)$$

For $Z \backsim \mathsf{Beta}(\alpha,\beta)$, mode of $f(z)$ is $m{=}\frac{\alpha-1}{\alpha+\beta-2}$, the Bayesian estimator is:

$$\widetilde{\theta}_B{=}\tfrac{\alpha^*-1}{\alpha^*+\beta^*-2}{=}\tfrac{(n\overline{x}+\alpha-1)}{(n+\alpha+\beta-2)}. \tag{6}$$

If we compare this with the MLE $\widehat{\theta}(\mathbf{X}){=}\overline{X}_n{=}\frac{1}{n}\sum_{k=1}^{n} X_k$, the two coincide *algebraically* only when $\alpha{=}\beta{=}1 \rightarrow \widetilde{\theta}_B = \overline{x}_n$. The restrictions $\alpha{=}\beta{=}1$ imply that the prior $\pi(\theta)$ is Uniformly distributed.

Another "natural" choice (depending on the implicit loss function) for the Bayesian point estimator is the **mean of the posterior** distribution. For $Z \backsim \mathsf{Beta}(\alpha,\beta)$, $E(Z){=}\frac{\alpha}{\alpha+\beta}$, and thus:

$$\widehat{\theta}_B{=}\tfrac{\alpha^*}{\alpha^*+\beta^*}{=}\tfrac{(n\overline{x}+\alpha)}{(n+\alpha+\beta)}. \tag{7}$$

**Example**. Let $\pi(\theta) \backsim \mathsf{Beta}(.5,.5)$.

(a) $n\overline{x}{=}4$, $n{=}20$, $\alpha^*{=}n\overline{x}+\alpha{=}4.5$, $\beta^*{=}n(1\text{-}\overline{x})+\beta{=}16.5$,

$$\widetilde{\theta}_B{=}\tfrac{3.5}{21-2}{=}.184, \qquad \widehat{\theta}_B{=}\tfrac{4.5}{4.5+16.5}{=}.214.$$

(b) $n\overline{x}{=}12$, $n{=}20$, $\alpha^*{=}n\overline{x}{+}\alpha{=}12.5$, $\beta^*{=}n(1{-}\overline{x}){+}\beta{=}8.5$,

$$\widetilde{\theta}_B{=}\tfrac{11.5}{19}{=}.605, \qquad \widehat{\theta}_B{=}\tfrac{12.5}{21}{=}.595.$$

When Bayesians claim that **all the relevant information for any inference** concerning $\theta$ is given by $\pi(\boldsymbol{\theta}|\mathbf{x}_0)$ they only admit to half the truth. The other half is that for selecting a Bayesian 'optimal' estimator of $\theta$ one needs to invoke additional information like a loss (or utility) function $\mathcal{L}(\widehat{\theta}(\mathbf{X}),\theta)$. Using different loss functions gives rise to different choices of Bayes's estimate.

For example:

(i) When $\mathcal{L}(\widehat{\theta},\theta){=}(\widehat{\theta}-\theta)^2$ the resulting Bayes estimator is the *mean* of $\pi(\theta|\mathbf{x}_0)$,

(ii) when $\mathcal{L}(\widetilde{\theta},\theta){=}|\widetilde{\theta}-\theta|$ the Bayes estimator is the *median* of $\pi(\theta|\mathbf{x}_0)$, and

(iii) when $\mathcal{L}(\overline{\theta},\theta){=}\delta(\overline{\theta},\theta)$, where $\delta(.){=}\begin{cases} 0 & \text{for } \overline{\theta}{=}\theta \\ 1 & \text{for } \overline{\theta}{\neq}\theta \end{cases}$, the Bayes estimator $\overline{\theta}$ is the *mode* of $\pi(\theta|\mathbf{x}_0)$.

## 1.4 Bayesian Credible Intervals

A Bayesian $(1-\alpha)$ credible interval for $\theta$ is constructed by ensuring that the area between $a$ and $b$ is equal to $(1-\alpha)$:

$$\pi(a \leq \theta < b) = \int_a^b \pi(\theta|\mathbf{x}_0)d\theta = 1-\alpha,$$

In practice one can define an infinity of $(1-\alpha)$ credible intervals using the same posterior $\pi(\theta|\mathbf{x}_0)$. To avoid this indeterminacy one needs to impose additional restrictions like the interval with the **shortest length** or one with **equal tails**, i.e. $\int_a^1 \pi(\theta|\mathbf{x}_0)d\theta=(1-\frac{\alpha}{2})$, $\int_b^1 \pi(\theta|\mathbf{x}_0)d\theta=\frac{\alpha}{2}$; see Robert (2007).

**Example**. For the simple (one parameter - $\sigma^2$ is known) Normal model, the sampling distribution of $\overline{X}_n=\frac{1}{n}\sum_{k=1}^n X_k$ and the posterior distribution of $\mu$ derived on the basis of an improper uniform prior [$\pi(\mu)=1$ for all $\mu\in\mathbb{R}$] are:

$$\overline{X}_n \overset{\text{TSN}}{\backsim} \mathsf{N}(\mu^*, \tfrac{\sigma^2}{n}), \qquad (\mu|\mathbf{x}_0) \backsim \mathsf{N}(\overline{x}_n, \tfrac{\sigma^2}{n}). \tag{8}$$

The two distributions can be used, respectively, to construct $(1-\alpha)$ Confidence and Credible Intervals:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \leq \mu \leq \overline{X}_n + c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}); \mu=\mu^*\right) = 1-\alpha, \tag{9}$$

$$\pi\left(\overline{x}_n - c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}}) \leq \mu \leq \overline{x}_n + c_{\frac{\alpha}{2}}(\tfrac{\sigma}{\sqrt{n}})|\mathbf{x}_0\right) = 1-\alpha, \tag{10}$$

The two intervals might appear the same, but they are drastically different.

*First*, in (9) the r.v. is $\overline{X}_n$ and its sampling distribution $f(\overline{x}_n;\mu)$ is defined over $\mathbf{x}\in\mathbb{R}^n_x$, but in (10) the r.v. is $\mu$ and its posterior $\pi(\mu|\mathbf{x}_0)$ is defined over $\mu\in\mathbb{R}$.

*Second*, the reasoning underlying (9) is factual (TSN), but that of (10) is All Possible States of Nature (APSN).

Hence, the $(1-\alpha)$ Confidence Interval (9) provides the shortest random upper $U(\mathbf{X})=\overline{X}_n+c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$ and lower $L(\mathbf{X})=\overline{X}_n-c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$ bounds that cover the true $\mu$ with probability $(1-\alpha)$. In contrast, the $(1-\alpha)$ Credible Interval (10) provides the *shortest interval* defined by two non-random, lower $L(\mathbf{x}_0)=\overline{x}_n-c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$ and upper $U(\mathbf{x}_0)=\overline{x}_n+c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$, values, such that $(1-\alpha)\%$ of the posterior $\pi(\mu|\mathbf{x}_0)$ lies within, i.e. (10) is the highest posterior non-random interval of length $2c_{\frac{\alpha}{2}}(\frac{\sigma}{\sqrt{n}})$; it includes $(1-\alpha)\%$ of the highest ranked values of $\mu\in\mathbb{R}$.

This raises a pointed question:

▶ is a Bayesian $(1-\alpha)$ Credible Interval an inference that pertains to the "true" $\mu$? If not,

▶ what does a Bayesian $(1-\alpha)$ Credible interval say about the process that generated

data $\mathbf{x}_0$?

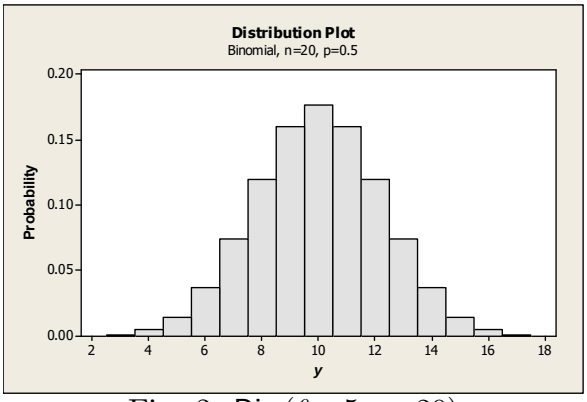



Fig. 2: $\mathsf{Bin}\,(\theta{=}.5; n{=}20)$

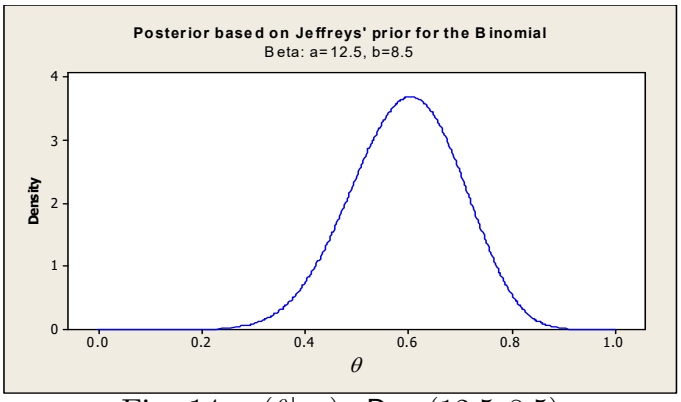



Fig. 14: $\pi(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(12.5, 8.5)$

The contrast between the sampling distribution of $\overline{X}_n$ and the posterior distribution of $\theta$ in the case of the simple Bernoulli model brings the difference out more starkly; one is discrete the other is continuous. For example, in the case of $n\overline{x}_n{=}12$, $n{=}20$, the **sampling distribution** of $\overline{X}_n$ ($\theta_0 = .5$) is given in fig. 2 and the **posterior distribution** based on a Jeffreys prior, centered at $\overline{x}_n{=}.6$, is given in fig. 14.

**Example**. For the simple Bernoulli model, the end points of an *equal-tail* credible interval can be evaluated using the F tables and the fact that:

$$Z \backsim \mathsf{Beta}(\alpha^*, \beta^*) \Rightarrow \quad \tfrac{\beta^* Z}{\alpha^*(1-Z)} \backsim \mathsf{F}(2\alpha^*, 2\beta^*).$$

Denoting the $\frac{\alpha}{2}$ and $(1{-}\frac{\alpha}{2})$ percentiles of the $\mathsf{F}(2\alpha^*, 2\beta^*)$ distribution, by $\mathsf{f}(\frac{\alpha}{2})$ and $\mathsf{f}(1{-}\frac{\alpha}{2})$, respectively, the Bayesian $(1{-}\alpha)$ credible interval for $\theta$ is:

$$\left(1 + \tfrac{\beta^*}{\alpha^* \mathsf{f}(1-\frac{\alpha}{2})}\right)^{-1} \leq \theta \leq \left(1 + \tfrac{\beta^*}{\alpha^* \mathsf{f}(\frac{\alpha}{2})}\right)^{-1}.$$

**Example**. Let $\pi(\theta) \backsim \mathsf{Beta}(.5, .5)$.
(a) $n\overline{x}{=}2$, $n{=}20$, $\alpha{=}.05$,

$$\alpha^*{=}n\overline{x} + \alpha{=}2.5, \;\; \beta^*{=}n(1-\overline{x}) + \beta{=}18.5,$$

$$f(1{-}\tfrac{\alpha}{2}){=}.163, \;\; f(\tfrac{\alpha}{2}){=}2.93,$$

$$\left(1{+}\tfrac{18.5}{2.5(.163)}\right)^{-1} \leq \theta \leq \left(1{+}\tfrac{18.5}{2.5(2.93)}\right)^{-1} \Leftrightarrow \quad (.0216 \leq \theta \leq .284)\,.$$

(b) $n\overline{x}{=}18$, $n{=}20$, $\alpha{=}.05$,

$$\alpha^*{=}n\overline{x} + \alpha{=}18.5, \;\; \beta^*{=}n(1\text{-}\overline{x}) + \beta{=}2.5,$$

$$\widehat{\theta}_B{=}\tfrac{18.5}{21}{=}.881, f(1\text{-}\tfrac{\alpha}{2}){=}.341, \;\; f(\tfrac{\alpha}{2}){=}6.188,$$

$$\left(1{+}\tfrac{2.5}{18.5(.341)}\right)^{-1} \leq \theta \leq \left(1{+}\tfrac{2.5}{18.5(6.188)}\right)^{-1} \Leftrightarrow \quad (.716 \leq \theta \leq .979)\,.$$

One can also use the *asymptotic approximation* in (**??**) to construct an approximate Credible interval for $\theta$ :

$$\pi\left(\widehat{\theta}_B - c_{\frac{\alpha}{2}}\frac{\sqrt{\widehat{\theta}_B[1-\widehat{\theta}_B]}}{\sqrt{(n+\alpha+\beta+1)}} \leq \theta < \widehat{\theta}_B + c_{\frac{\alpha}{2}}\frac{\sqrt{\widehat{\theta}_B[1-\widehat{\theta}_B]}}{\sqrt{(n+\alpha+\beta+1)}}\right) = 1-\alpha, \tag{11}$$

where $c_{\frac{\alpha}{2}}$ denotes the Normal $\frac{\alpha}{2}$ percentile.

**Example**. Let $\pi(\theta) \backsim \mathsf{Beta}(.5, .5)$.

(a) $n\overline{x}$=2, $n$=20, $\alpha$=.05, $\alpha^*$=2.5, $\beta^*$=18.5, $\widehat{\theta}_B$=$\frac{2.5}{21}$=0.119 :

$$\widehat{\theta}_B \pm c_{\frac{\alpha}{2}}\frac{\sqrt{\widehat{\theta}_B[1-\widehat{\theta}_B]}}{\sqrt{(n+\alpha+\beta+1)}} = (-0.0163,\ 0.254)\,.$$

(b) $n\overline{x}$=18, $n$=20, $\alpha$=.05, $\alpha^*$=18.5, $\beta^*$=2.5, $\widehat{\theta}_B$=$\frac{18.5}{21}$=.881 :

$$\widehat{\theta}_B \pm c_{\frac{\alpha}{2}}\frac{\sqrt{\widehat{\theta}_B[1-\widehat{\theta}_B]}}{\sqrt{(n+\alpha+\beta+1)}} = (0.746,\ 1.016)\,.$$

It important to emphasize that this approximation can be very crude in practice, when $n$ is small and/or the posterior distribution is skewed. The approximation will be better for $\phi = \ln\left(\frac{\theta}{1-\theta}\right)$.

For additional numerical examples see Appendix.

## 1.5 Bayesian Testing

Bayesian **testing of hypotheses** is not as easy to handle using the posterior distribution, as it is for Credible Intervals, especially for point hypotheses, because of the *technical difficulty* in attaching probabilities to particular values of $\boldsymbol{\theta}$ since the parameter space $\Theta \subset \mathbb{R}^m$ is usually *uncountable*. Indeed, the only prior distribution that makes technical sense is:

$$\pi(\theta) = 0, \text{ for each value } \theta \in \Theta.$$

In their attempt to deflect attention away from this technical difficulty, Bayesians criticized the use of point hypotheses such as $\theta = \theta_0$ in frequentist testing as nonsensical because they can never be exactly true! This is a nonsensical argument because the notion of *exactly true*, has no place in statistics.

### 1.5.1 Point null alternative hypotheses

There have been several attempts to address the difficulty with point hypotheses, but no agreement seems to have emerged; see Roberts (2007). Let us consider one such attempt for testing of the hypotheses:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \text{ vs. } H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1.$$

Like all Bayesian inferences, the basis is the posterior distribution. Hence, an obvious way to assess their respective degrees of belief is the **posterior odds**:

$$\frac{\pi(\boldsymbol{\theta}_0|\mathbf{x}_0)}{\pi(\boldsymbol{\theta}_1|\mathbf{x}_0)} = \frac{L(\boldsymbol{\theta}_0|\mathbf{x}_0)\cdot\pi(\boldsymbol{\theta}_0)}{L(\boldsymbol{\theta}_1|\mathbf{x}_0)\cdot\pi(\boldsymbol{\theta}_1)} = \left(\frac{\pi(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_1)}\right)\left(\frac{L(\boldsymbol{\theta}_0|\mathbf{x}_0)}{L(\boldsymbol{\theta}_1|\mathbf{x}_0)}\right), \tag{12}$$

where the factor $\frac{\pi(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_1)}$ represents the **prior odds**, and $\frac{L(\boldsymbol{\theta}_0|\mathbf{x}_0)}{L(\boldsymbol{\theta}_1|\mathbf{x}_0)}$ the **likelihood ratio**. In light of the fact that technical problem stems from the prior $\pi(\boldsymbol{\theta})$ assigning probabilities to particular values of $\boldsymbol{\theta}$, an obvious way to sidestep the problem is to cancel the prior odds factor, by using the ratio of the posterior to the prior odds to define the **Bayes Factor** (BF):

$$BF(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1|\mathbf{x}_0)= \left(\frac{\pi(\boldsymbol{\theta}_0|\mathbf{x}_0)}{\pi(\boldsymbol{\theta}_1|\mathbf{x}_0)}\right) / \left(\frac{\pi(\boldsymbol{\theta}_0)}{\pi(\boldsymbol{\theta}_1)}\right) \quad = \frac{L(\boldsymbol{\theta}_0|\mathbf{x}_0)}{L(\boldsymbol{\theta}_1|\mathbf{x}_0)}. \tag{13}$$

This addresses the technical problem because the likelihood function is definable for particular values of $\theta$.

For this reason Bayesian testing is often based on the BF combined with certain **rules of thumb**, concerning the *strength of the degree of belief against* $H_0$ as it relates to the magnitude of $BF(\mathbf{x}_0;\theta_0)$ (Robert, 2007):

- $0 \leq BF(\mathbf{x}_0;\theta_0) \leq 3.2$, the degree of belief against $H_0$ is *poor*,
- $3.2 < BF(\mathbf{x}_0;\theta_0) \leq 10$, the degree of belief against $H_0$ is *substantial*,
- $10 < BF(\mathbf{x}_0;\theta_0) \leq 100$, the degree of belief against $H_0$ is *strong*, and
- $BF(\mathbf{x}_0;\theta_0) > 100$, the degree of belief against $H_0$ is *decisive*.

**The Likelihoodist approach.** It is important to note that the *Law of Likelihood* defining the likelihood ratio:

$$LR(\theta_0, \theta_1|\mathbf{x}_0) = \frac{L(\theta_0|\mathbf{x}_0)}{L(\theta_1|\mathbf{x}_0)},$$

provides the basis of the Likelihoodist approach to testing, but applies only to tests of point vs. point hypotheses. In contrast, the Bayes Factor can be extended to composite hypotheses.

Notice that like point estimation and credible intervals, the claim that Bayesian inference relies exclusively on the posterior distribution for inference purposes is only half the truth. The other half for Bayesian testing is the use of rules of thumb to go from the BF to evidence for or against the null, that have been called into questioned as largely ad hoc; see Kass and Raftery (1995).

### 1.5.2 Composite hypotheses

Consider the following hypotheses:

$$H_0\text{: } \theta \leq \theta_0 \quad \text{vs.} \quad H_1\text{: } \theta > \theta_0,\ \theta_0 = .5,$$

in the context of the simple Bernoulli case with a Jeffreys invariant prior, with data $n\overline{x}$=12, $n$=20.

An obvious way to evaluate the posterior odds for these two interval hypotheses is as follows:

$$\pi(\theta \leq \theta_0|\mathbf{x}_0)=\tfrac{\Gamma(21)}{\Gamma(12.5)\Gamma(8.5)} \textstyle\int_0^{.5} \left(\theta^{11.5}(1\text{-}\theta)^{7.5}\right) d\theta=.186,$$

$$\pi(\theta > \theta_0|\mathbf{x}_0)=1\text{-}\pi(\theta \leq \theta_0|\mathbf{x}_0)=.814$$

One can then employ the **posterior odds** criterion:

$$\frac{\pi(\theta\leq\theta_0|\mathbf{x}_0)}{\pi(\theta>\theta_0|\mathbf{x}_0)}=\frac{.186}{.814}=.229,$$

which indicates that the degree of belief against $H_0$ is *poor*.

### 1.5.3 Point null but composite alternative hypothesis

**Pretending that point hypotheses are small intervals**. A 'pragmatic' way to handle point hypotheses in Bayesian inference is to sidestep the technical difficulty in handling hypotheses of the form:

$$H_0\text{: } \theta=\theta_0 \quad \text{vs.} \quad H_1\text{: } \theta\neq\theta_0,$$

by *pretending* that $\theta$=$\theta_0$ is actually a small interval:

$$H_0\text{: } \theta\in\Theta_0:=(\theta_0-\epsilon,\ \theta_0+\epsilon),$$

and attaching a **spiked prior** of the form:

$$\pi(\theta=\theta_0)=p_0, \quad p_1=\textstyle\int_0^1 \pi(\theta\neq\theta_0)d\theta=1-p_0. \tag{14}$$

i.e. attach a prior of $p_0$ to $\theta$=$\theta_0$, and then distribute the rest $1-p_0$ to all the other values of $\theta$; see Berger (1985).

**Using Credible Intervals as surrogates for tests**. Lindley (1965) suggested an adaptation of a frequentist procedure of using the duality between the acceptance region and Confidence Intervals as surrogates for tests, by replacing the latter with Credible Intervals. His Bayesian adaptation to handle point null hypotheses, say:

$$H_0\text{: } \theta=.8 \quad \text{vs.} \quad H_1\text{: } \theta\neq.8,$$

is to construct a $(1-\alpha)$ Credible Interval using an "uninformative" prior and reject $H_0$ if it lies outside that interval.

**Example**. For $n\overline{x}$=12, $n$=20, the likelihood function is:

$$L(\theta;\mathbf{x}_0)\propto\theta^{12}(1-\theta)^8,\ \theta\in[0,1],$$

when combined with $\pi(\theta)\backsim$Beta$(1,1)$ yields:

$$\pi(\theta|\mathbf{x}_0)\backsim \text{Beta}(13,9),\ \theta\in[0,1].$$

A .95 credible interval for $\theta$ is:

$$\pi(.384\leq\theta<.782)=.95,$$

$$\tfrac{\Gamma(22)}{\Gamma(13)\Gamma(9)}\textstyle\int_{.384}^1\theta^{12}(1-\theta)^8d\theta=0.975, \quad \tfrac{\Gamma(22)}{\Gamma(13)\Gamma(9)}\textstyle\int_{.7817}^1\theta^{12}(1-\theta)^8d\theta=0.025.$$

This suggests that the null $\theta_0=.8$ should be rejected because it lies outside the credible interval.

## 2 The large $n$ problem and Bayesian testing

*The large n problem* was initially raised by Lindley (1957) in the context of the simple Normal model (**??**) where the variance $\sigma^2 > 0$ is assumed known, by pointing out:

**[a] the large $n$ problem**: frequentist testing is susceptible to the fallacious result that there is always a large enough sample size $n$ for which any point null, say $H_0$: $\theta{=}\theta_0$, will be rejected by a frequentist $\alpha$-significance level test.

Lindley claimed that this result is *paradoxical* because, when viewed from the Bayesian perspective, one can show:

**[b] the Jeffreys-Lindley paradox***:* for *certain choices* of the prior, the posterior probability of $H_0$, given a frequentist $\alpha$-significance level rejection, will approach one as $n{\to}\infty$.

Claims [a] and [b] contrast the behavior of a frequentist test (p-value) and the posterior probability of $H_0$ as $n{\to}\infty$, that highlights a potential for conflict between the frequentist and Bayesian accounts of evidence:

**[c] Bayesian charge 1**: "The Jeffreys-Lindley paradox shows that for inference about $\theta$, P-values and Bayes factors may provide contradictory evidence and hence can lead to opposite decisions." (Ghosh et. al, 2006, p. 177)

**[d] Bayesian charge 2**: a hypothesis that is well-supported by Bayes factor can be (misleadingly) rejected by a frequentist test when $n$ is large; see Berger and Sellke (1987), pp. 112-3.

**A paradox**? No! From the error statistical perspective:

(i) There is nothing fallacious about a small p-value, or a rejection of $H_0$, when $n$ is large [it is a feature of a consistent frequentist test], **but** there *is* a problem when such results are detached from the test itself, and are treated as providing the same evidence for a particular alternative $H_1$, regardless of the generic capacity (the power) of the test in question, which depends crucially on $n$.

▶ Hence, the real problem does not lie with the p-value or the accept/reject rules as such, but with how such results are transformed into *evidence for* or *against* a particular $H$.

The large $n$ problem can be circumvented by using the *post-data* **severity assessment**.

How does the Bayesian approach explain why the result $\pi(\theta_0|\mathbf{x}_0){\to}1$ as $n{\to}\infty$ (irrespective of the truth or falsity of $H_0$), is conducive to a more sound evidential account?

**Example**. Consider the following example in Stone (1997):

"A particle-physics complex plans to record the outcomes of a large number of independent particle collisions of a particular type, where the outcomes are either type A or type B. ... the results are to be used to test a theoretical prediction that the proportion of type A outcomes, $h$, is precisely 1/5, against the vague alternative that $h$ could take any other value. The results arrive: 106298 type A collisions out of 527135." (p. 263)

## 2.1 Bayesian testing

Consider applying the *Bayes factor* procedure to the hypotheses (**??**) using a *uniform prior*:

$$\theta \backsim \mathsf{U}(0,1), \text{ i.e. } \pi(\theta)=1 \text{ for all } \theta\in[0,1]. \tag{15}$$

This gives rise to the Bayes factor:

$$\begin{aligned} BF(\mathbf{x}_0;\theta_0)=\tfrac{L(\theta_0;\mathbf{x}_0)}{\int_0^1 L(\theta;\mathbf{x}_0)d\theta} &= \tfrac{\binom{527135}{106298}(.2)^{106298}(1-.2)^{527135-106298}}{\int_0^1\left(\binom{527135}{106298}\theta^{106298}(1-\theta)^{527135-106298}\right)d\theta}= \\ &= \tfrac{.000015394}{.000001897}=8.115. \end{aligned} \tag{16}$$

NOTE that the same Bayes factor (16) arises in the case of the *spiked prior* (14) with $p_0$=.5, where $\theta$=$\theta_0$ is given prior probability of .5 and other half is distributed equally among the remaining values of $\theta$; for $p_0$=.5 the ratio $(p_0/[1-p_0])$ =1 and will cancel out from $BF(\mathbf{x}_0;\theta)$.

► A Bayes factor result $BF(\mathbf{x}_0;\theta_0) > 8.115$, indicates that data $\mathbf{x}_0$ *favor* the null against all other values of $\theta$ *substantially*; see Robert (2007).

**Is the result as clear cut as it appears?** No, because, on the basis of same data $\mathbf{x}_0$, the Bayes factor $BF(\mathbf{x}_0;\theta_0)$ 'favors', not only $\theta_0$=.2, but each individual value $\theta_1$ inside a certain interval around $\theta_0$=.2:

$$\Theta_{BF}:=[.199648,\ .203662]\subset\Theta_1:=\Theta-\{.2\}, \tag{17}$$

where the square bracket indicates inclusion of the end point, in the sense that, for each $\theta_1\in\Theta_{BF}$, $BF(\mathbf{x}_0;\theta_1) > 1$, i.e.

$$L(\theta_1;\mathbf{x}_0) > \int_0^1 L(\theta;\mathbf{x}_0)d\theta, \text{ for all } \theta_1\in\Theta_{BF}. \tag{18}$$

Worse, certain values $\theta^{\ddagger}$ in $\Theta_{BF}$ are favored by $BF(\mathbf{x}_0;\theta^{\ddagger})$ *more strongly* than $\theta_0$=.2:

$$\theta^{\ddagger}\in\Theta_{LR}:=(.2,\ .20331]\subset\Theta_{BF}. \tag{19}$$

It is important to emphasize that the subsets $\Theta_{LR}\subset\Theta_{BF}\subset\Theta$ exist for every data $\mathbf{x}_0$, and one can locate them by trial and error. However, there is a much more efficient way to do that. As shown below, $\Theta_{LR}$ can be defined as a subset of $\Theta$ around the Maximum Likelihood Estimate (MLE):

$$\widehat{\theta}_{MLE}(\mathbf{x}_0)=\tfrac{106298}{527135}=0.20165233.$$

Is this a coincidence? NO, as Mayo (1996), p. 200, pointed out, $\theta^{\blacklozenge}$=$\widehat{\theta}_{MLE}(\mathbf{x}_0)$ is always the *maximally likely alternative,* irrespective of the null or other substantive values of interest. In this example, the Bayes factor for $H_0$: $\theta$=$\theta^{\blacklozenge}$ vs. $H_1$: $\theta\neq\theta^{\blacklozenge}$ yields:

$$\begin{aligned} BF(\mathbf{x}_0;\theta^{\blacklozenge}) &= \tfrac{\binom{527135}{106298}(.20165233)^{106298}(1-.20165233)^{527135-106298}}{\int_0^1\left(\binom{527135}{106298}\theta^{106298}(1-\theta)^{527135-106298}\right)d\theta}= \\ &= \tfrac{.0013694656}{.000001897}=721.911, \end{aligned} \tag{20}$$

indicating extremely decisive evidence for $\theta{=}\theta^{\blacklozenge}{=}\widehat{\theta}_{MLE}(\mathbf{x}_0)$:

$$\theta^{\blacklozenge} \text{ is favored by } BF(\mathbf{x}_0;\theta^{\blacklozenge}) \text{ more than}$$
$$89{\simeq}\tfrac{7}{8}\tfrac{21.911}{.115} \text{ times stronger than } \theta_0{=}.2!$$

Indeed, if one were to test the point hypotheses:

$$H_0\text{: } \theta{=}.2 \text{ vs. } H_0\text{: } \theta{=}\theta^{\blacklozenge},$$

$$LR(\theta_0,\theta^{\blacklozenge};\mathbf{x}_0) = \frac{\binom{527135}{106298}(.2)^{106298}(1-.2)^{527135-106298}}{\binom{527135}{106298}(.20165233)^{106298}(1-.20165233)^{527135-106298}} = \frac{.000015394}{.001369466}{=}.011241,$$

the result *reverses* the original Bayes factor result and suggests that the degree to which data $\mathbf{x}_0$ favor $\theta{=}\theta^{\blacklozenge}$ over $\theta_0{=}.2$ is much stronger ($89{\simeq}\frac{1}{.011241}$) as in (20).

■ This result is an instance of the **fallacy of acceptance:** the Bayes factor $BF(\mathbf{x}_0;\theta_0) > 8$ is misinterpreted as providing evidence for $H_0$: $\theta_0 = .2$ against any value of $\theta$ in $\Theta_1{:=}\Theta - \{.2\}$, when in fact $BF(\mathbf{x}_0;\theta^{\ddagger})$ provides much stronger evidence for certain values of $\theta^{\ddagger}$ in $\Theta_1$; in particular all $\theta^{\ddagger}{\in}\Theta_{LR} \subset \Theta_1$.

**What is the source of the problem?** The key problem is that the Bayes factor is invariant to the sample size $n$, i.e. it is irrelevant whether $\overline{x}_n$ results from $n{=}10$ or $n{=}10^{10}$, when going from $BF(\mathbf{x}_0;\theta_0) > 8$ to claiming that data $\mathbf{x}_0$ provide strong evidence for $H_0 : \theta{=}.2$. Hence, going from:

$$\textit{step 1:}\quad LR(\theta_0,\theta_1;\mathbf{x}_0){=}\tfrac{L(\theta_0;\mathbf{x}_0)}{L(\theta_1;\mathbf{x}_0)} > k, \tag{21}$$

indicating that $\theta_0$ is $k$ times more likely than $\theta_1$, to:

$$\textit{step 2}\text{: fashioning } k > 0 \text{ into the strength of evidence } \textit{for } \theta_0,$$

the Bayesian interpretation goes astray by ignoring $n$.

Where does the above post-data severity perspective leave the Bayesian (and likelihoodist) inferences?

▶ Both approaches are plagued by two key problems:

(i) the *maximally likely alternative* problem (Mayo, 1996), in the sense that the value $\widehat{\theta}_{MLE}(\mathbf{x}_0){=}\theta^{\blacklozenge}$ is always favored against every other value of $\theta$, irrespective of the substantive values of interest, and

(ii) the invariance of $LR(\theta_0,\theta_1;\mathbf{x}_0)$ to the sample size $n$.

In contrast, the severity of the inferential claim $\theta > \theta^{\blacklozenge}$ is always low, being equal to .5 (table S), calling into question the appropriateness of such a choice.

In addition, the severity assessment in table S calls seriously into question the results associated with the two intervals $\Theta_{BF}{:=}[.199653,\ .203662]$ and $\Theta_{LR}{:=}(.2,\ .20331]$, because these intervals include values $\theta^{\ddagger}$ of $\theta$ for which the severity of the relevant inferential claim $\theta > \theta^{\ddagger}$ is very low, e.g. $SEV(T_{\alpha};\theta > .2033) \simeq .001$.

**Conclusion**. Any *evidential account* aiming to provide a sound answer the question:

'when do data $\mathbf{x}_0$ provide evidence for or against a hypothesis (or a claim)?'
can ignore the **generic capacity** of a test at its peril!

## 2.2 Nonsense Bayesians utter about the frequentist approach

**Frequentist inference, in general**

"Non-Bayesians, who we hereafter refer to as *frequentists*, argue that situations not admitting repetition under essentially identical conditions are not within the realm of statistical enquiry, and hence 'probability' should not be used in such situations. Frequentists define the probability of an event as its long-run relative frequency. ... that definition is nonoperational since only a finite number of trials can ever be conducted.' (p. 2)

Koop, G. D.J. Poirier and J.L. Tobias (2007), Bayesian Econometric Methods, Cambridge University Press, Cambridge.

**About p-values**

Bayesians often point to the closeness of the p-value and the posterior probability of $H_0$ to make two misleading claims:

**First**, Bayesian testing often enjoys good 'error probabilistic' properties, whatever that might mean.

**Second**, the closeness of the posterior probability and the p-value indicates the superiority of the former because it provides what modelers want:

"... the applied researcher would really like to be able to place a degree of belief on the hypothesis." (Press, 2003, p. 220)

**Really????**

This interpretation is completely at odds with the proper frequentist interpretation of the p-value which is firmly attached to the testing procedure and is used as a measure discordance between $\theta^*$ and the null value $\theta_0$. What is often ignored in such Bayesian discussions is that the p-value is susceptible to the fallacies of rejection and acceptance, rendering the posterior probability for $H_0$ equally vulnerable to the same fallacies; see Spanos (2013).

The above comparison raises several interesting questions:

▶ Is there a connection between the p-value and the posterior of $H_0$? More generally,

▶ is there a connection between frequentist error probabilities and posterior probabilities associated with the null and alternative hypotheses? If yes,

▶ what does that imply about the relationship between the frequentist and Bayesian approaches to testing?

**Example - simple (one parameter) Normal model**.

Consider the simple Normal Model, with $\sigma^2$=1. Returning to (9), the two distributions:

$$\overline{X}_n \overset{\mu=\mu^*}{\backsim} \mathsf{N}(\mu^*, \tfrac{1}{n}), \qquad (\mu|\mathbf{x}_0) \backsim \mathsf{N}(\overline{x}_n, \tfrac{1}{n}).$$

are different because one can easily draw $\pi(\mu|\mathbf{x}_0)$ because $\overline{x}_n$ is a known number, but $\mu$ in $\mathsf{N}(\mu, \frac{1}{n})$ is unknown. For instance, the evaluation of $E(\overline{X}_n^2-\mu^2)$ gives two different answers:

$$\underset{X}{E}(\overline{X}_n^2 - \mu^2)=\tfrac{1}{n}, \quad \text{but} \quad \underset{\mu}{E}(\overline{x}_n^2 - \mu^2)= -\tfrac{1}{n},$$

where $\underset{X}{E}(.)$ and $\underset{\mu}{E}(.)$ indicate expectations with respect to $f(\overline{x}_n;\mu)$ and $\pi(\mu|\mathbf{x}_0)$, respectively.

### 2.3 Bayesian Prediction

The best Bayesian predictor for $X_{n+1}$ is based on its posterior predictive density, which is defined by:

$$f(x_{n+1}|\mathbf{x}_0)=\int_0^1 f(x_{n+1}|\mathbf{x_0};\theta)f(\mathbf{x}_0|\theta)\pi(\theta)d\theta.$$

Note that $f(x_{n+1},\mathbf{x}_0,\boldsymbol{\theta})=[f(x_{n+1}|\mathbf{x_0};\theta)\cdot f(\mathbf{x}_0|\theta)\cdot\pi(\theta)]$ defines the joint distribution of $(x_{n+1},\mathbf{x}_0,\boldsymbol{\theta})$. Since $X_{n+1}\backsim\mathsf{Ber}(\theta(1-\theta))$, integrating out $\theta$ yields:

$$f(x_{n+1}|\mathbf{x}_0)=\begin{cases}\frac{(n\overline{x}+\alpha)}{(n+\alpha+\beta)}, & \text{if } x_{n+1}=1,\\ \frac{(n(1-\overline{x})+\beta)}{(n+\alpha+\beta)}, & \text{if } x_{n+1}=0,\end{cases}$$

which is a Bernoulli density with $\theta^*=\frac{(n\overline{x}+\alpha)}{(n+\alpha+\beta)}$. The Bayesian predictor, based on the mode of $f(x_{n+1}|\mathbf{x})$ is:

$$\widetilde{X}_{n+1}=\begin{cases}1, & \text{if } \max(\theta^*,[1-\theta^*])=\theta^*,\\ 0, & \text{if } \max(\theta^*,[1-\theta^*])=[1-\theta^*].\end{cases}$$

The posterior expectation predictor is given by:

$$E(X_{n+1}|\mathbf{x}_0) \quad =\frac{(n\overline{x}+\alpha)}{(n+\alpha+\beta)},$$

which in the case where $\alpha=\beta=1$, $\pi(\theta)\backsim\mathsf{Beta}(1,1)=U(0,1)$, the posterior expectation predictor gives rise to **Laplace's law of succession:**

$$E(X_{n+1}|\mathbf{x}_0) \quad =\frac{(n\overline{x}+1)}{(n+2)}.$$

# 3 Fisher's criticisms of Bayesian inference

In "The Design of Experiments" (1935), pp. 6-7, R. A. Fisher gave three criticisms of Bayesian inference.

The **first** of Fisher's criticisms concerns the degrees of belief interpretation of probability being unscientific, in contrast to the objective relative frequency interpretation. Modern Bayesians proposed a twofold counter-argument: there is nothing problematic about their interpretation for scientific reasoning purposes, but the frequentist interpretation of probability is problematic (circular). Despite their rhetoric, there is a clear move away from subjective 'informative' priors towards priors relating to the likelihood function; reference or 'objective' priors.

Fisher's **second** criticism of Bayesian inference is based on the fact that the reliability of scientific research and inference is never justified by invoking Bayesian reasoning, despite its long history going back to the 18th century. In particular, Fisher (1955) objected vehemently to viewing statistical inference as a 'decision problem under uncertainty' based on arbitrary **loss** or **utility functions**. In his writings

Fisher argued that notions like *inductive inference*, *evidence* and *learning from data* differ crucially from notions like decisions, behavior and actions with losses and gains. He considered Bayesian inference and decision-theoretic formulations highly artificial and not in tune with learning from data and scientific reasoning.

In his **second** criticism Fisher called into question the **Bayes formula**:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0) = \frac{\pi(\boldsymbol{\theta})\cdot f(\mathbf{x}_0|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}\in\Theta}\pi(\boldsymbol{\theta})\cdot f(\mathbf{x}_0|\boldsymbol{\theta})d\boldsymbol{\theta}},\ \boldsymbol{\theta}\in\Theta,$$

being viewed as an *axiom* whose truth can be taken for granted. His criticisms focused primarily on the fact that this formula treats the existence of the prior distribution $\pi(\boldsymbol{\theta})$ as **self-evident** and straightforward.

In particular, Fisher (1921) criticized the notion of **prior ignorance** widely used by Bayesians since the 1820s as the basis of their inference. The claim going back to Laplace that a **uniform prior:**

$$\pi(\theta) \backsim \mathsf{U}(0,1),\ \text{all } \theta\in\Theta,$$

can be used to quantify a **state of ignorance** about the unknown parameter $\theta$, was been vigorously challenged by Fisher (1921) on *non-invariance* to reparameterization grounds: one is ignorant about $\theta$ but very informed about $\varphi{=}h(\theta)$; see section 1.2.1 above.

Fisher's **third criticism** raised a fundamental question pertaining to Bayesian inference which has not been answer in a satisfactory way to this day. The question is:

| How does one choose the prior distribution $\pi(\theta)$? |
|---|

The answer 'from a priori substantive information' is both inaccurate and misleading. It is inaccurate because substantive information almost never comes in the form of a prior $\pi(\theta)$ defined over statistical parameter(s) $\theta$. It often comes in the form of the sign and magnitude of unknown parameters. It is also misleading because when no information pertaining different values of $\theta$ is available, one has a difficult task of framing that; falling into the above trap!

# 4 Where do prior distributions come from?

In what follows a number of different strategies for choosing the prior are discussed.

## 4.1 Conjugate prior and posterior distributions

The above Bayesian inference was based on a particular choice of a prior known as conjugate pairs. This is the case where the product of the **prior** $\pi(\theta)$ and the **posterior:**

$$\pi(\theta|\mathbf{x}_0) \propto \pi(\theta) \cdot L(\theta; \mathbf{x}_0), \text{ for all } \theta \in \Theta,$$

belongs to the same family of distributions; $L(\theta; \mathbf{x}_0)$ is family preserving.

**Table 2 - Conjugate pairs** $(\pi(\theta),\ \pi(\theta|\mathbf{x}_0))$

| **Likelihood** | $\pi(\theta)$ |
|---|---|
| Binomial (Bernoulli) | $\mathsf{Beta}(\alpha, \beta)$ |
| Negative Binomial | $\mathsf{Beta}(\alpha, \beta)$ |
| Poisson | $\mathsf{Gamma}(\alpha, \beta)$ |
| Exponential | $\mathsf{Gamma}(\alpha, \beta)$ |
| Gamma | $\mathsf{Gamma}(\alpha, \beta)$ |
| Uniform | $\mathsf{Pareto}(\alpha, \beta)$ |
| Normal for $\theta = \mu$ | $\mathsf{N}(m, \tau^2),\ m \in \mathbb{R}, \tau^2 > 0$ |
| Normal for $\theta = \sigma^2$ | $\mathsf{Inverse\ Gamma}(\alpha, \beta)$ |

**Example**. Let the prior distribution be: $\pi(\theta) \backsim \mathsf{Beta}(\alpha, \beta)$,

when combined with the likelihood function: $L(\theta; \mathbf{x}_0) \propto \theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})},\ \theta \in [0, 1]$,

as shown above, gives rise to the posterior: $\pi(\theta|\mathbf{x}_0) \backsim \mathsf{Beta}(\alpha^*, \beta^*)$.

Table 2 presents some examples of conjugate pairs of prior and posterior distributions, as they combine with different likelihood forms.

Conjugate pairs make mathematical sense, but does it make 'modeling' sense? The various justifications in the Bayesian literature vary from, 'these help the objectivity of inference' to 'they enhance the allure of the Bayesian approach as a black box' and these claims are often contradictory!

## 4.2 Jeffreys *invariant* prior for $\theta$

In respond to Fisher's second criticism that a uniform (proper or improper) prior distribution for $\theta$ is **not invariant to reparameterizations** of the form $\phi = h(\theta)$, e.g. $\phi = \ln \theta$, Jeffreys proposed a new class of priors which satisfy this property. This family of invariant priors proposed by Jeffreys was based on Fisher's average information:

$$I(\theta; \mathbf{x}) = \underset{\mathbf{x}}{E}\left(\frac{1}{n}\left[\frac{d \ln L(\theta;\mathbf{x})}{d\theta}\right]^2\right) = \int_{\mathbf{x}\in\mathbb{R}_x^n} \cdots \int \frac{1}{n}\left(\frac{d \ln L(\theta;\mathbf{x})}{d\theta}\right)^2 d\mathbf{x}. \tag{22}$$

Note that the above derivation involves some hand-waving in the sense that if the likelihood function $L(\theta;\mathbf{x}_0)$ is viewed, like the Bayesians do, as only a function of the data $\mathbf{x}_0$, then taking expectations outside the brackets makes no sense; the expectation is with respect to the distribution of the sample $f(\mathbf{x};\theta)$ for all possible values of $\mathbf{x}\in\mathbb{R}^n_x$. As we can see, the derivation of $I(\theta;\mathbf{x})$ runs afoul to the **likelihood principle** since all possible values of the sample $\mathbf{X}$, not just the observed data $\mathbf{x}_0$, are taken into account. Note that in the case of a random (IID) sample, the *Fisher information* $I_n(\theta;\mathbf{x})$ for the sample $\mathbf{X}:=(X_1,X_2,...,X_n)$ is related to the above average information via:

$$I_n(\theta;\mathbf{x})=nI(\theta;\mathbf{x}).$$

In the case of a single parameter, **Jeffreys invariant prior** takes the form:

$$\pi(\theta)\propto\sqrt{I(\theta;\mathbf{x})}. \tag{23}$$

That is, the *likelihood function* determines the *prior* distribution.

The crucial property of this prior is that it is invariant to reparameterizations of the form $\phi=h(\theta)$. This follows from the fact that:

$$\tfrac{d\ln L(\phi;\mathbf{x})}{d\theta} \quad =\tfrac{d\ln L(\theta;\mathbf{x})}{d\theta}(\tfrac{d\theta}{d\phi}), \tag{24}$$

which implies that:

$$I(\phi;\mathbf{x})=I(\theta;\mathbf{x})(\tfrac{d\theta}{d\phi})^2\Rightarrow \quad \pi(\phi)\propto\sqrt{I(\phi;\mathbf{x})}=\pi(\theta)(\tfrac{d\theta}{d\phi}), \tag{25}$$

using the change-of-variable rule with the Jacobian $(\frac{d\theta}{d\phi})$. Consequently:

$$\sqrt{I(\theta;\mathbf{x})}d\theta=\sqrt{I(\phi;\mathbf{x})}d\phi.$$

The intuition underlying this result is that the prior $\pi(\theta)$ inherits the **invariance** of MLE estimators to reparameterizations. That is, the MLE $\widehat{\phi}$ of $\phi=h(\theta)$ can be derived by a simple replacement of $\theta$ with its MLE $\widehat{\theta}$:

$$\widehat{\phi}=h(\widehat{\theta}).$$

**The simple Bernoulli model**. In view of the fact that the log-likelihood takes the form:

$$\ln L(\theta;\mathbf{x})=n\overline{x}\ln(\theta)+n(1-\overline{x})\ln(1-\theta),$$

$$\tfrac{d\ln L(\theta;\mathbf{x})}{d\theta}=\tfrac{n\overline{x}}{\theta}-\tfrac{n(1-\overline{x})}{1-\theta}, \quad \tfrac{d^2\ln L(\theta;\mathbf{x})}{d\theta^2}=-(\tfrac{n\overline{x}}{\theta^2})-\tfrac{n(1-\overline{x})}{(1-\theta)^2},$$

From the second derivative, it follows that:

$$E\left(\tfrac{1}{n}\left[\tfrac{d\ln L(\theta;\mathbf{x})}{d\theta}\right]^2\right)=E\left(\left[-\tfrac{1}{n}\tfrac{d^2\ln L(\theta;\mathbf{x})}{d\theta^2}\right]\right)=\tfrac{1}{\theta(1-\theta)}. \tag{26}$$

This follows directly from $E(\overline{X})=\theta$, since:

$$E\left(\left[-\tfrac{1}{n}\tfrac{d^2\ln L(\theta;\mathbf{x})}{d\theta^2}\right]\right)=\tfrac{\theta}{\theta^2}+\tfrac{n(1-\theta)}{(1-\theta)^2}=\tfrac{1}{\theta}+\tfrac{1}{1-\theta}=\tfrac{1}{\theta(1-\theta)}. \tag{27}$$

From the definition of *Jeffreys invariant prior* we can deduce that for $\theta$ :

$$\pi(\theta) \propto \sqrt{I(\theta;\mathbf{x})}=\sqrt{\frac{1}{\theta(1-\theta)}} = \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}},\ 0<\theta<1, \tag{28}$$

which is an 'unnormalized' $\mathsf{Beta}(\frac{1}{2},\frac{1}{2})$ distribution; it does *not* integrate to one as it stands since:

$$\int_0^1 \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}d\theta = \pi.$$

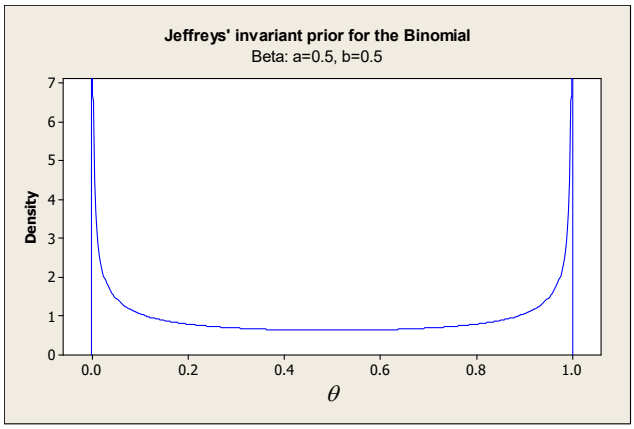



Fig. 11: Jeffreys prior
$\pi(\theta)=\frac{1}{B(.5,.5)}\theta^{-.5}(1-\theta)^{-.5}$

This is a special case of the $\mathsf{Beta}(\alpha,\beta)$ distribution:

$$\pi(\theta)=\frac{1}{B(\alpha,\beta)}\theta^{(\alpha-1)}(1-\theta)^{\beta-1},\ \alpha>0,\ \beta>0. \tag{29}$$

Jeffreys invariance prior (28) is also the reference prior; Bernardo and Smith (1994).

## 4.3 Examples of alternative Priors for $\theta$

**A. Uniform prior:** $\alpha=\beta=1,\ \pi(\theta) \backsim \mathsf{Beta}(1,1)=U(0,1)$

This is a proper prior used by both Bayes (1763) and Laplace (1774) in conjunction with the simple Bernoulli model.

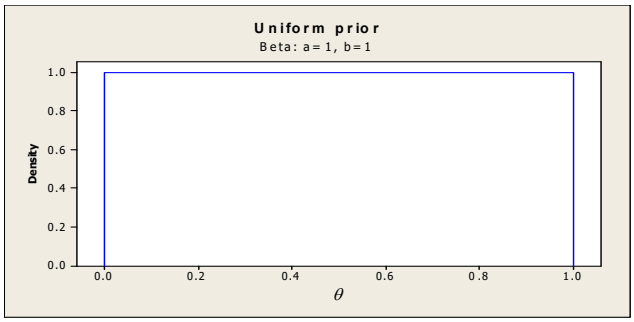



Fig. 15: Uniform prior

For this prior the posterior $\pi(\theta|\mathbf{x})$ and the likelihood function coincide:

$$\pi(\theta|\mathbf{x}_0)=L(\theta;\mathbf{x}_0) \propto \theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})},\ \theta\in[0,1]. \tag{30}$$

Note that in this case the posterior distribution *coincides* with the likelihood function!

**B. Jeffreys prior for the Negative Binomial**:

$\alpha$=0, $\beta$=$\frac{1}{2}$, $\pi(\theta) \backsim \mathsf{Beta}(0, .5)$

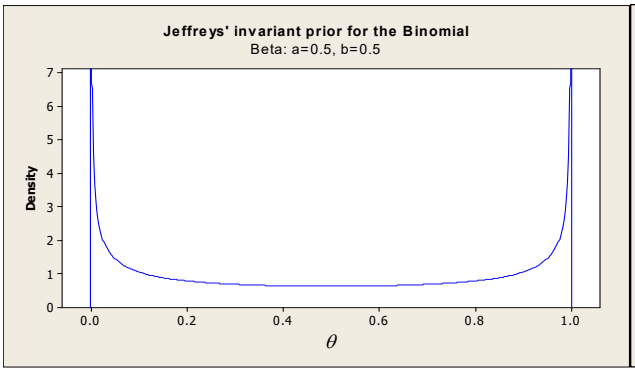



Fig. 11: Jeffreys prior $\pi(\theta) \backsim \mathsf{Beta}(.5, .5)$

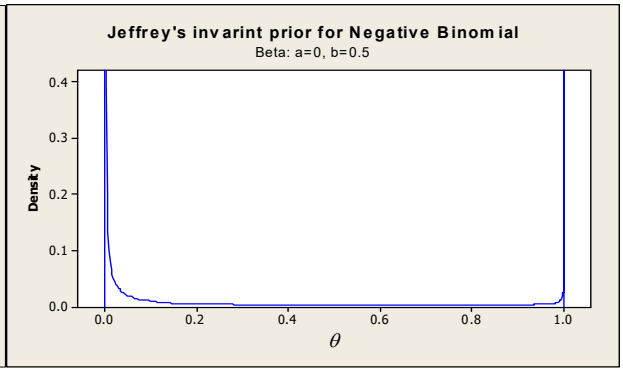



Fig. 17: Jeffreys prior: $\pi(\theta) \backsim \mathsf{Beta}(0, .5)$

The Negative Binomial case arises as a re-interpretation of the simple Bernoulli model where the likelihood function:

$$L(\theta; \mathbf{x}_0) \propto \theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})}, \ \theta \in [0,1], \tag{31}$$

is interpreted, not as arising from a sequence of IID Bernoulli trials, but as a sequence of trials until a pre-specified number of successes $r$ is achieved. In this case the distribution of the sample is **Negative Binomial** of the form:

$$\binom{z+r-1}{z}\theta^r(1-\theta)^y, \ y=0,1,2,...,$$

where the random variable $z$ denotes the *number of failures* before the $r$-th success. Assuming that it took $n$ trials to achieve $r=n\overline{x}$ successes, then $z=n-r=n-n\overline{x}$, and the above density function becomes: $\binom{z+r-1}{z}\theta^{n\overline{x}}(1-\theta)^{n(1-\overline{x})}$, giving rise to the same likelihood function (31) as the Binomial density; they differ only in the combinatorics term which is absorbed in the proportionality constant.

However, when one proceeds to derive Fisher's information by taking expectations over the random variable of interest, the answer is different because for a Negative Binomial:

$$E(Z)=\frac{r(1-\theta)}{\theta}.$$

Hence, the derivation of Fisher's information yields:

$$E\left(\left[-\frac{1}{n}\frac{d^2 \ln L(\theta;\mathbf{x})}{d\theta^2}\right]\right) = \frac{1}{n}E\left(\frac{r}{\theta^2}+\frac{z}{(1-\theta)^2}\right) = \frac{1}{n}\left(\frac{r}{\theta^2}+\frac{r(1-\theta)}{(1-\theta)^2}\right) = \\ = \frac{1}{n}\left(\frac{r}{\theta^2}+\frac{r}{(1-\theta)}\right) = \frac{r}{n}\left(\frac{1}{\theta^2(1-\theta)}\right).$$

This gives rise to the Jeffreys invariant prior:

$$\pi(\theta) \propto \sqrt{\frac{1}{\theta^2(1-\theta)}} = \theta^{-1}(1-\theta)^{-\frac{1}{2}} \backsim \mathsf{Beta}(0, \tfrac{1}{2})$$

This derivation shows most clearly that when one takes expectations of the derivatives of the log-likelihood function to derive Fisher's information, by definition one returns to the *distribution of the sample* which is a function of the random variables comprising the sample $\mathbf{X}$ given $\theta$; the likelihood function treats $\mathbf{x}$ as fixed at $\mathbf{x}_0$.

As shown above, Jeffreys invariance prior for the Binomial distribution is:

$$\pi(\theta) \backsim \mathsf{Beta}(\tfrac{1}{2}, \tfrac{1}{2}).$$

In order to get some idea as to the relative weights the two prior distributions attach to different values of $\theta$, let us evaluate the prior probability for $\theta$ at there different intervals:

$$\theta\in[.1, .15],\ \theta\in[.5, .55],\ \theta\in[.9, .95]$$

in the two cases.

**Prior for Binomial**. $\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$ :

$$\int_{.1}^{.15} \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}d\theta{=}.152, \quad \int_{.5}^{.55} \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}d\theta{=}.1, \quad \int_{.9}^{.95} \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}d\theta{=}.192.$$

**Prior for Negative Binomial**: $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-\frac{1}{2}}$ :

$$\int_{.1}^{.15} \theta^{-1}(1-\theta)^{-\frac{1}{2}}d\theta{=}.433, \quad \int_{.5}^{.55} \theta^{-1}(1-\theta)^{-\frac{1}{2}}d\theta{=}.138, \quad \int_{.9}^{.95} \theta^{-1}(1-\theta)^{-\frac{1}{2}}d\theta{=}.2.$$

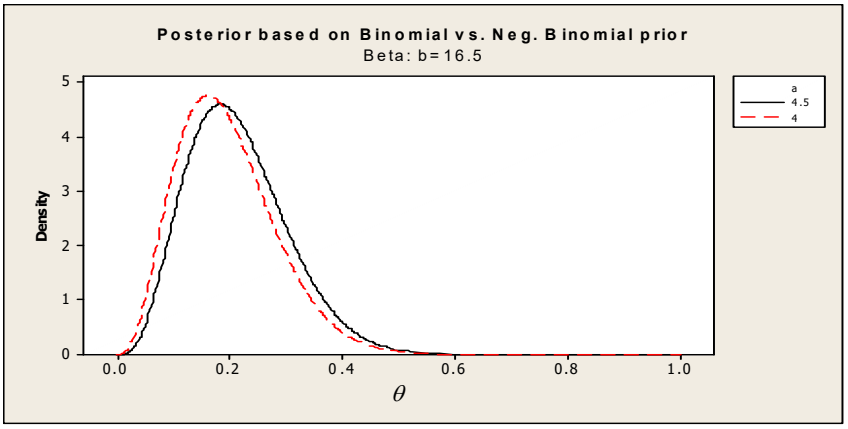



Fig. 18: The Posterior distribution:
Jeffrey's prior for Bin. vs. Neg. Bin.

The effect on the posterior can also be seen in fig. 18 for the case where $n\overline{x}$=4, $n$=20.

**C. Haldane's (improper) prior:** $\alpha{=}\beta{=}0,\ \pi(\theta) \backsim \mathsf{Beta}(0, 0)$.

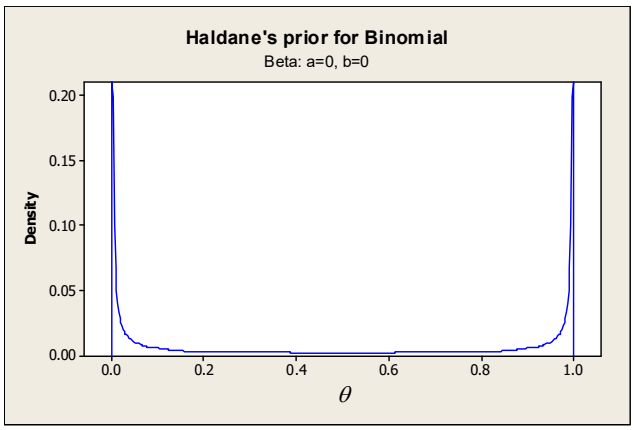



Fig. 19: Haldane's prior:
$\pi(\theta) \backsim \mathsf{Beta}(0,0)$

This is an *improper* prior because: $\int_0^1 \theta^{-1}(1-\theta)^{-1}d\theta=\infty$.

## 4.4 Choosing Prior distributions

The prior distribution $\pi(\theta)$ is the key to Bayesian inference and its determination is therefore the most important step in drawing inferences. In practice there is usually insufficient (not precise enough) information to help one specify $\pi(\theta)$ uniquely.

### 4.4.1 Subjective determination and approximations for priors

**A. Axiomatic approach**. This approach is similar to the way utility functions are constructed from preferences. The basic argument is that people who make consistent choices in uncertain situations behave *as if* they had subjective prior probability distributions over the different states of nature.

**B. Approximating the prior in the case where the parameter space $\Theta$ is finite**. When $\Theta$ is uncountable, say $\Theta$:=$[0,1]$, the approach is vulnerable to the *partition paradoxes*.

**C. Maximum entropy priors**. This involves maximizing:

$$E(\pi) = -\sum_{i=1}^{m} \pi(\theta_i)\ln(\pi(\theta_i)), \tag{32}$$

subject to some side conditions: $E(g_k(\theta)) = \omega_k,\ k$=$1,2,...,K$, giving rise to:

$$\pi^*(\theta) = \frac{\exp\left\{\sum_{k=1}^{K}\lambda_k g_k(\theta_i)\right\}}{\sum_{i=1}^{m}\exp\left\{\sum_{k=1}^{K}\lambda_k g_k(\theta_i)\right\}}. \tag{33}$$

Note that these priors belong to the *Exponential family* of distributions.

**D. Parametric approximations**. One chooses a prior, say $\pi(\theta) \backsim \mathsf{Beta}(\alpha,\beta)$, and then uses sample information, such as additional data, to choose $\alpha$ and $\beta$.

**E. Empirical Bayes**. An extension of D with a fully frequentist estimation of the parameters of the prior distribution.

**F. Hierarchical Bayes**. A variation on D with a priori specification of hyper-prior parameters over a certain narrower range of plausible values.

**Example**. In the case where the prior distribution is $\pi(\theta) \backsim \mathsf{Beta}(\alpha, \beta)$, one could assume that:

$$\alpha \backsim \mathsf{U}(0,2), \qquad \beta \backsim \mathsf{U}(0,2), \tag{34}$$

giving rise to a new prior distribution:

$$\pi^{\dagger}(\theta) = \tfrac{1}{4} \int_0^2 \int_0^2 (\tfrac{1}{\mathsf{B}(\alpha,\beta)}) \theta^{(\alpha-1)} (1-\theta)^{\beta-1} d\alpha d\beta. \tag{35}$$

This prior distribution can then be approximated using numerical integration, say Simpson's rule. As shown by Welsh (1996) this will give rise to a prior with $\alpha < 1,\ \beta < 1$, which looks like the Jeffreys invariance prior with flatter middle section.

### 4.4.2 'Objective' prior distributions

**A**. **Laplace's prior:** Uniform over the relevant parameter space.

**B**. **Data invariant priors**: the parameters obey the same transformations as the data; e.g. translation and scaling invariant priors.

**C**. **Jeffreys invariant priors**: the prior is invariance to reparameterizations $\phi = h(\theta)$.

**D**. **Reference priors**: an extension of Jeffreys priors to more than one parameters by adopting a priority list for the parameters in order to define their conditional prior distributions sequentially.

Consider the case where the distribution of the random variable $X$ depends on two unknown parameters, say $f(x; \theta_1, \theta_2)$, where $\theta_1$ is the parameter of interest. The reference prior $\pi_r(\theta_1)$ is obtained in three steps. Step one: derive $\pi(\theta_2 \mid \theta_1)$ as the Jeffreys prior associated with $f(x \mid \theta_1, \theta_2)$ assuming that $\theta_1$ is *fixed*. Step two: derive the marginal distribution:

$$f(x \mid \theta_1) = \int f(x \mid \theta_1, \theta_2) \pi(\theta_2 \mid \theta_1) d\theta_2,$$

by integrating out the nuisance parameter $\theta_2$. Step three: derive the Jeffreys prior $\pi(\theta_1)$ associated with $f(x \mid \theta_1)$; Bernardo and Smith (1994). Note that by reversing the order of the parameters $(\theta_1, \theta_2)$ both the reference priors change!

**E**. **Matching priors**: the choice of priors that achieve approximate (asymptotic) matching between posterior and frequentist 'error probabilities', such as coverage probabilities of Bayesian credible intervals with the corresponding frequentist CI coverage probabilities.

# 5 Appendix: Examples based on Jeffreys prior

For the simple Bernoulli model, consider selecting Jeffreys invariant prior:

$$\pi(\theta)=\tfrac{1}{B(.5,.5)}\theta^{-.5}(1-\theta)^{-.5},\ \theta\in[0,1].$$

This gives rise to a posterior distribution of the form:

$$\pi(\theta|\mathbf{x}_0)\backsim \mathsf{Beta}(n\overline{x}+.5,\ n(1-\overline{x})+.5),\quad \theta\in[0,1].$$

■ (a) For $n\overline{x}$=2, $n$=20, the likelihood function is:
$L(\theta;\mathbf{x}_0)\propto\theta^2(1-\theta)^{18},\ \theta\in[0,1]$,
and the posterior density is: $\pi(\theta|\mathbf{x}_0)\backsim\mathsf{Beta}(2.5,18.5)$, $\theta\in[0,1]$.
The Bayesian point estimates are: $\widetilde{\theta}_B=\frac{1.5}{19}=.0789$, $\widehat{\theta}_B=\frac{2.5}{21}=.119$.
A .95 credible interval for $\theta$ is: $\pi(.0214\leq\theta<.3803)=.95$,

$\frac{1}{\mathsf{B}(2.5,18.5)}\int_{a=.0214}^{1}\theta^{1.5}(1-\theta)^{17.5}d\theta=.975$, $\frac{1}{\mathsf{B}(2.5,18.5)}\int_{b=.3803}^{1}\theta^{1.5}(1-\theta)^{17.5}d\theta=.025$,

■ (b) For $n\overline{x}$=18, $n$=20, the likelihood function is:
$L(\theta;\mathbf{x}_0)\propto\theta^{18}(1-\theta)^2$, $\theta\in[0,1]$,
and the posterior density is: $\pi(\theta|\mathbf{x}_0)\backsim\mathsf{Beta}(18.5,2.5)$, $\theta\in[0,1]$.
The Bayesian point estimates are: $\widetilde{\theta}_B=\frac{17.5}{19}=.921$, $\widehat{\theta}_B=\frac{18.5}{21}=.881$.
A .95 credible interval for $\theta$ is: $\pi(.716\leq\theta<.97862)=.95$,

$\frac{1}{\mathsf{B}(18.5,2.5)}\int_{a=.716}^{1}\theta^{17.5}(1-\theta)^{1.5}d\theta=0.975$, $\frac{1}{\mathsf{B}(18.5,2.5)}\int_{b=.979}^{1}\theta^{17.5}(1-\theta)^{1.5}d\theta=0.025$.

■ (c) For $n\overline{x}$=72, $n$=80, the likelihood function is:
$L(\theta;\mathbf{x}_0)\propto\theta^{72}(1-\theta)^8$, $\theta\in[0,1]$,
and the posterior density is: $\pi(\theta|\mathbf{x}_0)\backsim\mathsf{Beta}(72.5,8.5)$, $\theta\in[0,1]$.
The Bayesian point estimates are: $\widetilde{\theta}_B=\frac{71.5}{79}=.905$, $\widehat{\theta}_B=\frac{72.5}{81}=.895$.
A .95 *credible interval* for $\theta$ is: $\pi(.82\leq\theta<.9515)=.95$,

$\frac{1}{\mathsf{B}(72.5,8.5)}\int_{a=.82}^{1}\theta^{71.5}(1-\theta)^{7.5}d\theta=0.975$, $\frac{1}{\mathsf{B}(72.5,8.5)}\int_{b=.9515}^{1}\theta^{71.5}(1-\theta)^{7.5}d\theta=0.025$.

■ (d) for $n\overline{x}$=40, $n$=80, the likelihood function is:
$L(\theta;\mathbf{x}_0)\propto\theta^{40}(1-\theta)^{40}$, $\theta\in[0,1]$,
and the posterior density is: $\pi(\theta|\mathbf{x}_0)\backsim\mathsf{Beta}(40.5,40.5)$, $\theta\in[0,1]$.
The Bayesian point estimates are: $\widetilde{\theta}_B=\frac{39.5}{79}=.5$, $\widehat{\theta}_B=\frac{40.5}{81}=.5$.
A .95 credible interval for $\theta$ is: $\pi(.3923\leq\theta<.6525)=.95$,

$\frac{1}{\mathsf{B}(40.5,40.5)}\int_{a=.392}^{1}\theta^{39.5}(1-\theta)^{39.5}d\theta=.975$, $\frac{1}{\mathsf{B}(40.5,40.5)}\int_{b=.6525}^{1}\theta^{39.5}(1-\theta)^{39.5}d\theta=.025$.

In view of the symmetry of the posterior distribution, even the asymptotic Normal credible interval (11) should give a good approximation. Given that $\widehat{\theta}_B = \frac{(n\overline{X}+\alpha)}{(n+\alpha+\beta)} = 0.5$, the *approximate credible interval* is:

$$\pi\left([.5-1.96\frac{\sqrt{.5(1-.5)}}{\sqrt{80}}] = .390 \leq \theta < .610 = [.5+1.96\frac{\sqrt{.5(1-.5)}}{\sqrt{80}}]\right) = 1-\alpha,$$

which provides a reasonably good approximation to the exact one.

## 5.1 A litany of misleading Bayesian claims concerning frequentist inference

"Broadly speaking, some of the arguments in favour of the Bayesian approach are that it is fundamentally sound, very flexible, produces clear and direct inferences and makes use of all the available information. In contrast, the classical approach suffers from some philosophical flaws, has restrictive range of inferences with rather indirect meaning and ignores prior information." (O'Hagan, 1994, p. 16)

[**1**] Bayesian inference is fundamentally sound because it can be given an axiomatic foundation based on coherent (rational) decision making, but frequentist inference suffers from several philosophical flaws.

[**2**] Frequentist inference is not very flexible and has a restrictive range of applicability.

According to Koop, Poirier and Tobias (2007):

"Non-Bayesians, who we hereafter refer to as *frequentists*, argue that situations not admitting repetition under essentially identical conditions are not within the realm of statistical enquiry, and hence 'probability' should not be used in such situations. Frequentists define the probability of an event as its long-run relative frequency. ... that definition is nonoperational since only a finite number of trials can ever be conducted.' (p. 2)

[**3**] Bayesian inference produces clear and direct inferences, in contrast to frequentist inference producing unclear and indirect inferences, e.g. credible intervals vs. confidence intervals.

[**4**] Bayesian inference makes use of all the available a priori information, but frequentist inference does not.

[**5**] A number of counter-examples, introduced by Bayesians, show that frequentist inference is fundamentally flawed.

[**6**] The subjectivity charge against Bayesians is misplaced because:

"All statistical methods that use probability are subjective in the sense of relying on mathematical idealizations of the world. Bayesian methods are sometimes said to be especially subjective because of their reliance on a prior distribution, but in most problems, scientific judgement is necessary to specify both the 'likelihood' and the prior' parts of the model." (Gelman, et al. (2004), p. 14)

"... likelihoods are just as subjective as priors." (Kadane, 2011, p. 445)

[**7**] For inference purposes, the only relevant point in the sample space $\mathbb{R}_X^n$ is just the data $\mathbf{x}_0$ as summarized by the *likelihood function* $L(\boldsymbol{\theta}|\mathbf{x}_0),\ \boldsymbol{\theta}\in\Theta$.

This feature of Bayesian inference is formalized by the *Likelihood Principle*.

**Likelihood Principle**. For inference purposes the only relevant sample information pertaining to $\boldsymbol{\theta}$ is contained in the likelihood function $L(\mathbf{x}_0|\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta$. Moreover, if $\mathbf{x}_0$ and $\mathbf{y}_0$ are two sample realizations contain the same information about $\boldsymbol{\theta}$ if they are proportional to one another (Berger and Wolpert, 1988, p. 19).

Frequentist inference procedures, such as estimation (point and interval), hypothesis testing and prediction inference procedure invoke other realizations $\mathbf{x}\in\mathbb{R}_X^n$, beyond the observed data $\mathbf{x}_0$, contravening the LP.

Indeed, two generations of Bayesian statisticians take delight in poking fun at frequentist testing by quoting Jeffreys's (1939) remark about the 'absurdity' of invoking the quantifier 'for all $\mathbf{x}\in\mathbb{R}_X^n$':

"What the use of P [p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a *remarkable* procedure." (p. 385) [ha, ha, ha!]