UCLA

UCLA Electronic Theses and Dissertations

Title

The Relative Density and Applications to Bayesian Analysis

Permalink

https://escholarship.org/uc/item/8qj392s1

Author

Fan, Haibo

Publication Date

2023

Supplemental Material

https://escholarship.org/uc/item/8qj392s1#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

The Relative Density
and Applications to
Bayesian Analysis

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Haibo Fan

2023

© Copyright by

Haibo Fan

2023

ABSTRACT OF THE THESIS

The Relative Density
and Applications to
Bayesian Analysis

by

Haibo Fan
Master of Science in Statistics
University of California, Los Angeles, 2023
Professor Mark S. Handcock, Chair

We introduce the relative distribution as a tool in Bayesian analysis to compare the posterior to the prior distribution. Two interpretations are given: one as a density ratio, and another as a reparametrized likelihood. Several important properties are reviewed, as well as notes on usage and connections to information theory and relative surprise inference. Explicit examples and derivations for several common models are given. The relative distribution focuses on the effect of the data and the resulting differences between the posterior and prior, emphasizing the role of the likelihood in connecting the two.

Keywords: Bayesian analysis, relative distribution, density ratio

The thesis of Haibo Fan is approved.

Oscar Madrid Padilla

Nicolas Christou

Mark S. Handcock, Committee Chair

University of California, Los Angeles

2023

To Destiny — I could not have done this without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Professor Handcock, for their patience and guidance throughout my master's program. Their expertise and enthusiasm were invaluable in helping me complete this research and write this thesis. They were always able to point me to new ideas or suggest different possiblities.

I would also like to thank Professors Nicolas Christou and Oscar Padilla for serving on my thesis committee and providing helpful feedback and suggestions.

Finally, I would like to thank my close friends for their support during this process: Sabih, Bryan, and Lyndal. Their companionship has been invaluable.

CHAPTER 1

Introduction

1.1 The Relative Distribution

Notationally, for this introductory section, all relevant functions of Y will have no subscript, and all relevant functions of Y_0 will have the subscript 0.

Given random variables Y (from a *comparison distribution*), and Y_0 (from a *reference distirbution*), the *relative density* from Y to Y_0 is defined as

$$R = F_0(Y)$$

We transform the comparison Y by the CDF of the reference Y_0. Literally, R is the probability that Y_0 is at most Y. To help interpret this, recall that with one random variable, applying its CDF to it would result in a standard uniform, as we map each value of the random variable to its appropriate quantile. With the relative distribution, we map each value of the comparison to where it would fall in the reference distribution — the values of Y are dispersed according to which quantile they would belong to if they actually came from Y_0.

We may also consider this manipulation:

$$\begin{aligned} P(a \leq R \leq b) &= P(a \leq F_0(Y) \leq b) \\ &= P(Q_0(a) \leq Y \leq Q_0(b)) \end{aligned}$$

The probability of $R \in [a, b]$ is the probability that Y is between the a^{th} and b^{th} quantiles of Y_0. This highlights the importance of thinking in terms of Y_0's quantiles when interpreting

R. Note that R is, at its core, a transformation of Y, the comparison distribution. Thus the randomness in R comes from Y alone; the use of the reference Y_0 is to establish quantiles.

We can find the CDF of R, which we denote with G, as

$$\begin{aligned} G(r) &= P(R \leq r) = P(F_0(Y) \leq r) \\ &= P(Y \leq F_0^{-1}(r)) \\ &= F(Q_0(r)) \end{aligned}$$

Then we can differentiate to find the PDF, which we denote as g. Recall the differentiation rule for derivatives of inverse functions: $\frac{d}{dx} f^{-1}(x) = 1/f'(f^{-1}(x))$.

$$\begin{aligned} g(r) &= \frac{d}{dr} G(r) \\ &= \frac{d}{dr} F(Q_0(r)) \\ &= f(Q_0(r)) \frac{1}{f_0(Q_0(r))} \\ &= \frac{f(Q_0(r))}{f_0(Q_0(r))} \end{aligned}$$

Note that the support of R is $(0,1)$. We can interpret its PDF as a ratio of the comparison to the reference densities, at each quantile r of the reference distribution. Generally, this interpretation is one of the easiest to work with and visualize, and so we shall use it often.

1.1.1 Example: Comparing Two Gaussians

Suppose we have $Y_0 \sim N(0,1)$ and $Y \sim N(1,4)$.

Looking at the graph of the densities, we can see that Y places much less mass between -1 and 0 compared to Y_0. It has more mass for values less than -2, and also values greater than approximately 1. Thus, thinking of the relative density as a density ratio, we should expect that it will be greater than 1 near the lower quantiles, fall down as it approaches the median of Y_0 ($r = 0.5$, at 0), but slowly rise after that, and become much greater than 1 near the upper quantiles.

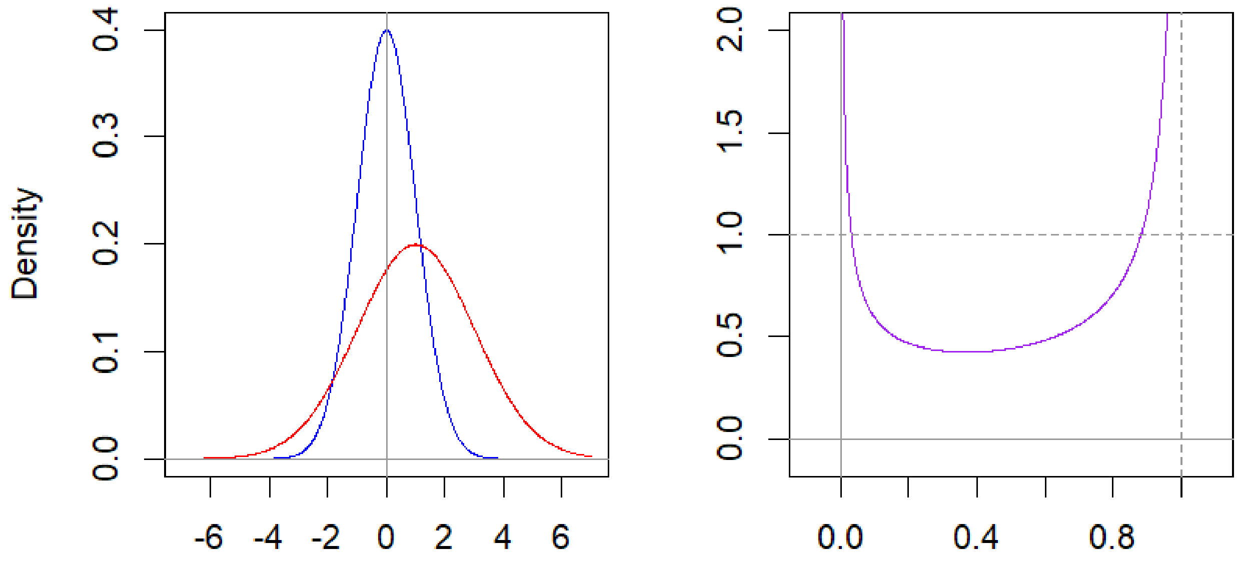

Figure 1.1: (Left) The density of Y (in red) and Y_0 (in blue). (Right) The relative density of Y to Y_0.

1.2 The Relative Distribution in a Bayesian Context

Notationally, we denote the data as X, a realization by x, and its likelihoods, which may be PDFs or PMFs, by $p(x)$; the CDFs and quantile functions will be represented by $F(x)$ and $Q(r)$ respectively. For parameters, we only consider those with PDFs, which we represent with π; CDFs and quantile functions will be represented by $\digamma$ and K, respectively. Generally, the prior will be denoted by θ, and the posterior by $\theta|x$.

Using the relative distribution, one natural pair of distributions to compare is in the Bayesian framework: the prior and posterior distributions of a (one-dimensional) parameter θ. Set the prior as the reference distribution, and the posterior as the comparison.

Hereon, we will use R to denote a random variable following this relative distribution, with G and g its CDF and PDF respectively. In this Bayesian case, the relative density is

$$\begin{aligned} g(r|x) &= \frac{\pi(K(r)|x)}{\pi(K(r))} \\ &= \frac{p(x|K(r))\pi(K(r)) \ / \ p(x)}{\pi(K(r))} \\ &= \frac{p(x|K(r))}{p(x)} \propto p(x|K(r)), \ 0 \leq r \leq 1 \end{aligned}$$

The relative distribution here is proportional to a reparametrization of the likelihood function of the data model, which scales it into the range of $(0, 1)$. We shall call this the *unitized likelihood* form of the relative density, which only appears different but is equivalent to the original form as a density ratio. Note that we generally keep our observed data x fixed, and the actual variable is r.

By the law of total probability, the marginal data probability is $p(x) = \int_{-\infty}^{\infty} p(x|\theta)\pi(\theta) \ d\theta$. Performing the change of variable $\theta = K(r)$ gives $p(x) = \int_0^1 p(x|K(r)) \ dr$, which we can recognize as the normalizing constant for the unitized likelihood to become a proper density.

Since the shape of the relative distribution is essentially the likelihood, this implies that all of the differences between the prior and posterior is based on the likelihood. When considering the form of Bayes' Rule and also the likelihood principle, this is fairly evident. The relative distribution should contain as much information about these differences as the likelihood.

CHAPTER 2

Properties of the Relative Distribution

We will often make use of the substitution $r = F(\theta) \implies \theta = K(r)$, with differentials $dr = \pi(\theta)\ d\theta$ and $d\theta = dr/\pi(K(r))$.

2.1 Properties of the Relative Distribution Itself

Here, we discuss properties of the relative distribution in general, which do not rely on the Bayesian context. We will consider Y as a random variable from the comparison distribution and Y_0 from the reference distribution.

2.1.1 Switching the Comparison and Reference

Typically, switching which distribution is considered the "comparison" and which the "reference" will not heavily impact analysis. Graphically, the relative density appears to have its values inverted. Note that this is not the true multiplicative inverse, as the scale used to compress the density into $[0, 1]$ is based on the quantiles of which distribution we consider the reference, thus some slight dissimilarities will arise.

Due to the nature of inverting values, one representation may be easier to visually grasp than another when using a standard linear axis. For example, in Figure 2.1, we can see that since the blue relative density does not approach infinity, it can be easier to see the full picture.

Additionally, we currently require that the prior has a quantile function in order to compute the relative distribution (see below for more details). In the Bayesian case however, we may sometimes have an improper prior (which does not have a quantile function) but with a proper posterior. We may somewhat circumvent this issue by instead setting the prior as the comparison and the posterior as the reference. If we denote this random variable as R^*, we have

$$\begin{aligned} g^*(r) &= \frac{\pi(K(r|\theta))}{\pi(K(r|\theta)|\theta)} \\ &= \frac{p(x)}{p(x|K(r|\theta))} \propto \frac{1}{p(x|K(r|\theta))} \end{aligned}$$

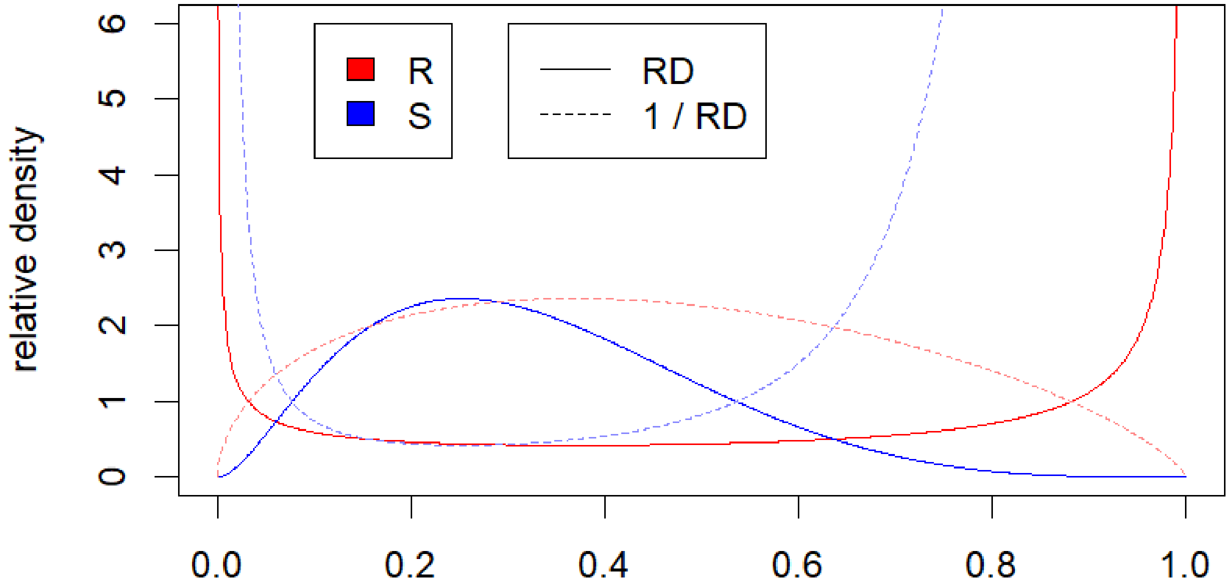

Figure 2.1: In red, the relative density of $Y \sim N(1, 2)$ to $Y_0 \sim N(0, 1)$. In blue, the relative density of Y_0 to Y. The dashed lines are the literal multiplicative inverses of the density values. These inverses do not exactly line up with simply switching the comparison and reference distribution, as different quantiles are used.

2.1.2 Invariance under Monotone Transformation

The relative distribution is *invariant* under a monotone transformation on both variables Y and Y_0 in the sense that the shape of its density does not change. For a monotonically increasing transformation, the relative density is the same. For a monotonically decreasing transformation, the relative density is reflected vertically about the prior median ($r = 0.5$).

Let h be a monotone transformation, with $Z = h(Y)$ and $Z_0 = h(Y_0)$ the transformed variables. First suppose h is monotonically increasing, i.e. $a \leq b \iff h(a) \leq h(b)$ for a, b in the domain of h. Then,

$$\begin{aligned} R_Z &= F_{Z_0}(Z) \\ &= P(Z_0 \leq Z) \\ &= P(h(Y_0) \leq h(Y)) \\ &= P(Y_0 \leq Y), \text{ by monotonicity} \\ &= F_{Y_0}(Y) = R_Y \end{aligned}$$

We can see that the relative distribution of the transformed variables is the same as the relative distribution of the original variables.

For h monotonically decreasing, i.e. $a \leq b \iff h(a) \geq h(b)$, we have

$$\begin{aligned} R_Z &= F_{Z_0}(Z) \\ &= P(Z_0 \leq Z) \\ &= P(h(Y_0) \leq h(Y)) \\ &= P(Y_0 \geq Y), \text{ by monotonicity} \\ &= 1 - P(Y_0 < Y) \\ &= 1 - F_0(Y) = 1 - R_Y \end{aligned}$$

From this, we can determine that $F_{R_Z}(r) = P(1 - R_Y \leq r) = 1 - F_{R_Y}(1 - r)$, which implies that $f_{R_Z}(r) = f_{R_Y}(1 - r)$. Thus the relative density — which we recall can be interpreted as the ratio of densities between Z and Z_0 — is the same as the original for R_Y, though mirrored along the median ($r = 0.5$).

One thing to note in this algebra is that we should treat $Z = g(Y)$ as the source of randomness. As we are including Z_0 (essentially, Y_0) only through its CDF, there is no randomness that stems from it, as the CDF is deterministic. Instead, the source of randomness is from Z (essentially, Y), and so we cannot collapse probabilities into F_Z, as that would make Z deterministic.

2.1.3 MLE and the Relative Distribution

The mode of the relative distribution is the argmax of its density, which occurs at the maximal value of the data likelihood, i.e. the MLE. With the density ratio view, this can be interpreted as the specific value of the parameter whose probability increases the most due to the data. Intuitively, it seems reasonable that the single parameter value which the data supports the most — the MLE — is exactly the one which would have been most likely to have generated that data.

The value of the relative density at the MLE can then help us infer how well the prior agrees with the data. When the value is very large, then the data must have impacted the posterior to place much more weight at the point than the prior did. On the other hand, when the value is small, it means that the posterior doesn't place much more weight than the prior did — the prior already found that point fairly likely.

The impact of the prior is demonstrated in Figure 2.2. Using a typical beta-binomial model, the relative densities of θ — the probability of success in the binomial — are plotted. We can see that the peaks — where the MLE would be — occur at different values of r, which are indicated by the vertical dashed lines.

Note that translating these quantiles r into actual values of θ based on the different priors would lead to the same estimate $\hat{\theta}^{MLE}$. The values of r are different as they are quantiles of

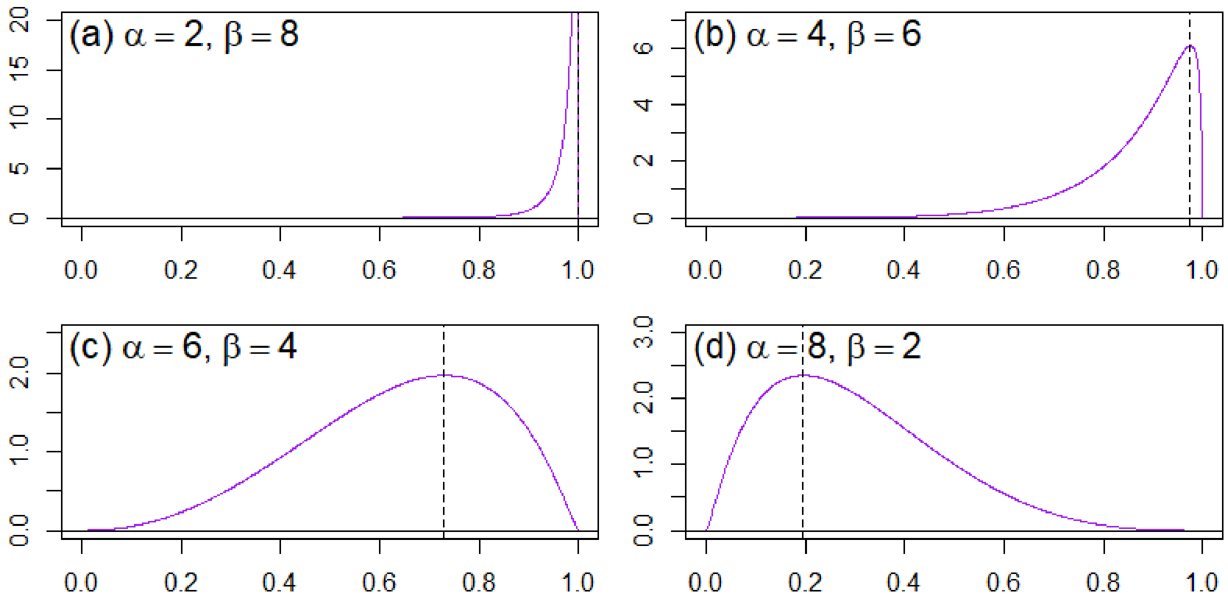

Figure 2.2: The resulting relative densities for the single trial probability in a beta-bernoulli model with various priors, all at the same strength of belief ($\alpha + \beta = 10$) but with different means: (a) $\alpha = 2$, (b) $\alpha = 4$, (c) $\alpha = 6$, (d) $\alpha = 8$. The data was simulated and had a sample mean of 0.7.

different priors, though they all correspond to the same actual value of $\hat{\theta}^{MLE}$.

2.1.4 Asymptotic Behavior

In the regular case of asymptotics, the relative density will approach a delta function in $[0, 1]$, whose peak will depend on the choice of prior.

Using the interpretation of the relative density as a density ratio, consider that under appropriate regularity conditions, the posterior distribution approaches a normal distribution, and its variance approaches 0. The vanishing variance explains the delta function, as the posterior approaches 0 at most points, and only peaks at its mode.

We can also understand this phenomenon by thinking of the relative density as the unitized likelihood. As the amount of data increases, the likelihood concentrates around the

MLE, which approaches a delta function. The location of the peak — the MLE — depends on the specific prior used.

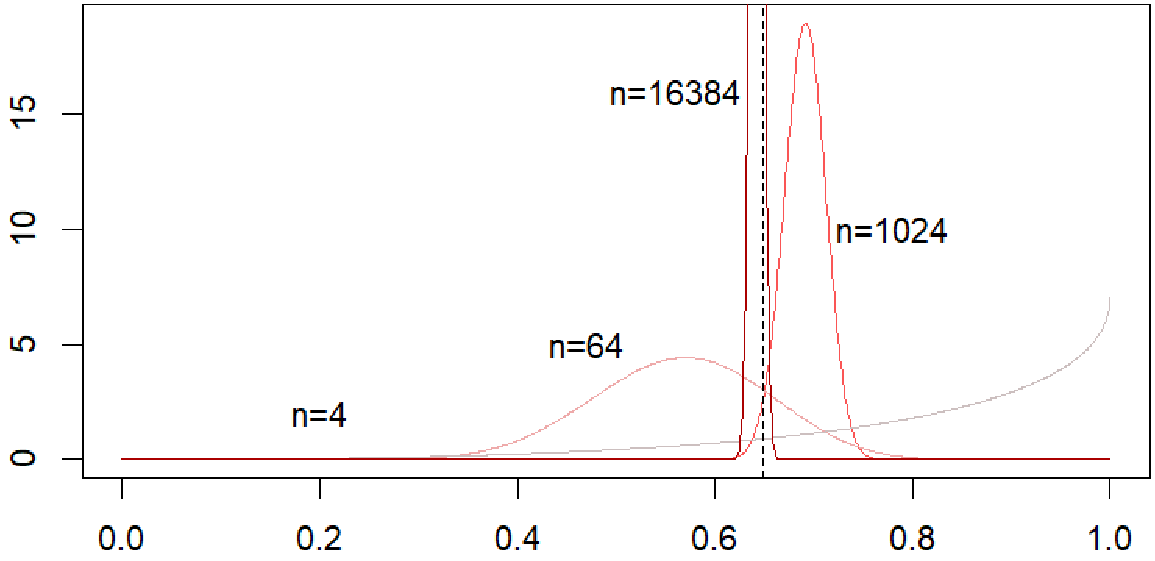

Figure 2.3: The relative density of a Beta$(2, 2)$ prior to a Bernoulli rate model (with true rate of 0.6) for various sample sizes. As the number of data points increases, the relative density concentrates more around the true rate, indicated by the dashed black line.

2.2 Properties of the Bayesian Relative Distribution

Here we present two interesting properties of the relative distribution when applied to the Bayesian context. We have a one-dimensional parameter θ, with a prior density of $\pi(\theta)$, giving CDF $F(\theta)$ and quantile function $K(r)$. Data is observed according to some density $p(x|\theta)$, which we use to update our parameter and determine a posterior density $\pi(\theta|x)$.

2.2.1 View of the Relative Distribution as a Reparametrization

Take a reparametrization of our data model by $r = F(\theta)$. While $\Omega_\theta \subseteq \mathbb{R}$, we have $\Omega_r \subseteq [0, 1]$.

Consider the integration of the data likelihood over the original parameter θ, that is $\int_{-\infty}^{\infty} p(x|\theta)\ d\theta$. This is generally not simplifiable. If we want to marginalize properly and find the marginal data probability $p(x)$, we need to weigh the likelihood by the prior $\pi(\theta)$. In contrast, when we directly integrate under r, in order to produce the marginal, we do not need to weigh the reparametrized likelihood when integrating, as $p(x) = \int_0^1 p(x|\theta = K(r))\ dr$.

Here, we actually implicitly weigh the likelihood by the standard uniform — a standard uniform prior over r is the same as $\pi(\theta)$ over θ. We can also see this by calculating the form of the reparametrized prior. Note that $\theta = K(r) \implies d\theta = \frac{1}{\pi(K(r))}\ dr$.

$$\begin{aligned}\pi(r) &= \pi(\theta)\left|\frac{d\theta}{dr}\right| \\ &= \pi(\theta = K(r))\left|\frac{1}{\pi(K(r))}\right| = 1\end{aligned}$$

Now recall the unitized likelihood form of the relative density in our Bayesian case, $g(r) = p(x|\theta = K(r))/p(x)$, which is the reparametrized data likelihood. Thus we can interpret this alternative perspective of the relative density as the likelihood, reparametrized such that when our prior places equal weight on each parameter value (of r), the marginal data probabilities do not change. Note that while under the reparametrization r, though we always have the implied uniform prior r, the *original* choice of prior is still important, as that is what controls the reparametrization itself, by the transformation $r = \digamma(\theta)$.

The existence of the transformation $r = \digamma(\theta)$ is crucial. Additionally, we have restricted ourselves here to the case where $\digamma$ is bijective, and so K is a proper inverse. The form of our likelihood under r, which involves K, suggests that the existence of K is necessary. We would find that we cannot directly apply these calculations to cases in which the transformation and its inverse do not exist; for example, to improper priors. Indeed, an earlier indication that such priors would be difficult to analyze using the relative density is that we require the marginal probability $p(x)$ as the normalization term for the unitized likelihood, which cannot be found when using improper priors.

Using this alternate perpsective, we can reduce our relative distribution analyses into the case of just a standard uniform prior. If we reparametrize our entire model from θ to r, then proceed with the analysis using a uniform prior on r, the resulting relative distribution would be the same had we continued using the original parametrization of θ

2.2.1.1 Coercion of the Relative Density's Shape

The unitized likelihood form $g(r|x) \propto p(x|\theta = K(r))$ suggests that we can manipulate the "shape" of the transformed data likelihood — that is, of the relative distribution — by carefully choosing the function K. Phrased more explicitly, for a nonnegative function $t(r)$, $r \in [0, 1]$, where $\int_0^1 t(r)\ dr = 1$, we may be able to find a function K (a prior) such that $t(r) = p(x|\theta = K(r))/p(x)$. Whatever "shape" t, we seek K so that the relative density looks like t.

Let us rewrite the unitized likelihood so that is more explicitly made a function of r, using $p(x|\theta = K(r)) = \mathcal{L}(\theta = K(r))$. One immediate option presents itself if we assume $\mathcal{L}$ has a *right inverse*, which we denote $\mathcal{L}^{-1_R}$. This one-sided inverse function satisfies $\mathcal{L}(\mathcal{L}^{-1_R}(r)) = r$. Now we simply let $K(r) = \mathcal{L}^{-1_R}(t(r))$, so that $\mathcal{L}(K(r)) = t(r)$.

Thus it seems may produce a fairly wide range of *relative* (as the relative density has a normalization constraint) ratios of the posterior to the prior in this fashion. Note that we are not actually affecting the amount of information in the data; instead, we are changing the *relative* impact the data has on the prior, by changing the prior itself. Of course, this further emphasizes the importance of choosing an appropriate prior.

As an example, consider n data points distributed as $N(\theta, \sigma^2)$, where θ is the unknown mean and σ^2 is the known variance. Then we can write the data likelihood as

$$\begin{aligned}
p(x|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right) \\
&= \frac{\exp\left(-\frac{\overline{x^2} - \bar{x}^2}{2\sigma^2/n}\right)}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{(\theta - \bar{x})^2}{2\sigma^2/n}\right) = \mathcal{L}(\theta|x)
\end{aligned}$$

See Appendix A.1.1 for a derivation. Denote $C = \exp\left(-\frac{\overline{x^2}-\bar{x}^2}{2\sigma^2/n}\right)/(2\pi\sigma^2)^{n/2}$. We may choose $\mathcal{L}^{-1_R}(r) = \sqrt{-\frac{2\sigma^2}{n}\log(r/C)} + \bar{x}$, and so $K(r) = \mathcal{L}^{-1_R}(t(r)) = \sqrt{-\frac{2\sigma^2}{n}\log(t(r)/C)} + \bar{x}$. Note that we are constrained by $t(r)/C \leq 1$, which may not always be possible.

In this specific case, we may be able to fix this constraint by adjusting the constant C. Suppose $t(r)$ has a finite maximum and call it A. Construct a "pseudo" right inverse, $L^{-1_R}(r) = \sqrt{-\frac{2\sigma^2}{n}\log(r/A)} + \bar{x}$. Then $\mathcal{L}(L^{-1_R}(t(r))) = \frac{C}{A}t(r)$. The resulting relative density, which must be a proper density, must then have normalizing constant $1/p(x) = \frac{A}{C}$, as we have assumed $t(r)$ has unit area. An example is given in Figure 2.4.

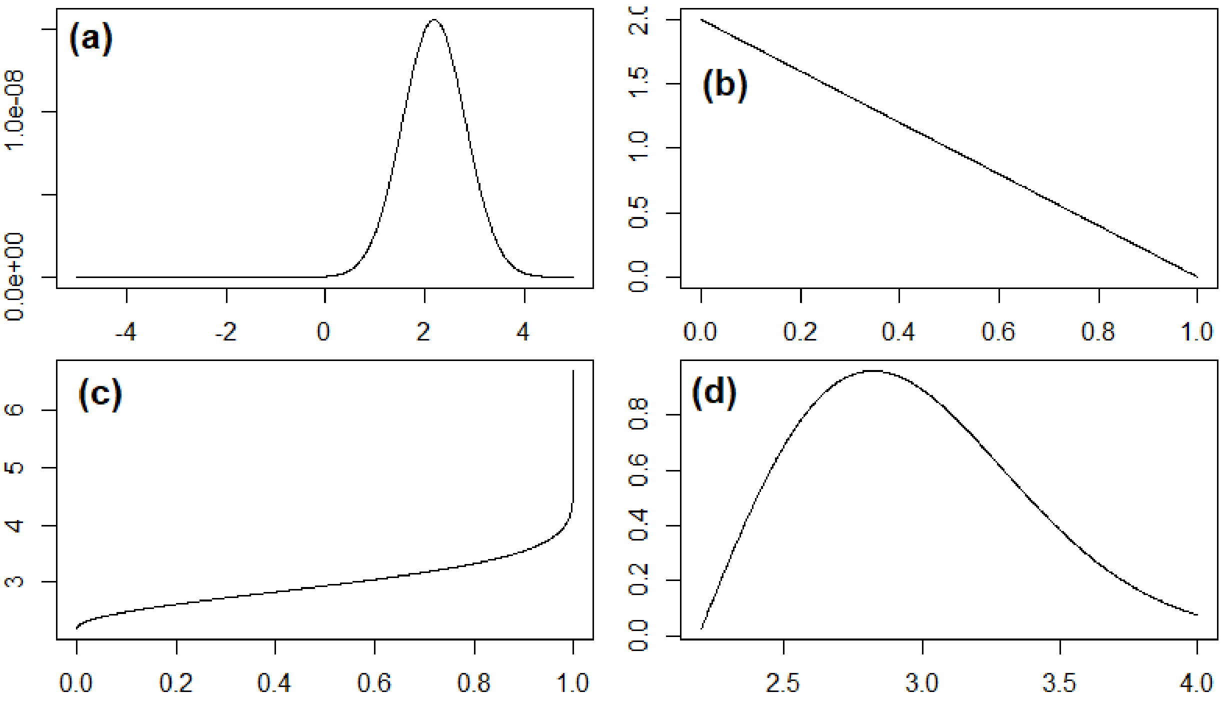

Figure 2.4: The data was simulated with $X \sim N(1.5, 4)$, with $n = 10$. (a) The likelihood function of the data, given as a function of θ. Note that this is not a proper density. (b) The desired functional form, $t(r) = 2 - 2r$, also the resulting relative density. (c) The coercive quantile function, $K(r) = \mathcal{L}^{-1_R}(t(r))$. (d) The density of the coercive quantile function.

While this could fix our specific case, in general we may not be able to find a workaround. There are a few considerations when using this method. As seen above, we must find a right inverse, which can be difficult, for example if we explicitly need all normalizing constants. Sometimes there are constraints on the domain, which limit the possibilities of $t(r)$, though as demonstrated this could be overcome to some extent. We also need to know the data already, as we require knowledge of $\mathcal{L}(\theta|x)$, which would not be possible if we were properly choosing a prior. Finally, we have to ensure that the resulting $K(r)$ is a valid quantile function, of which the strictest requirement is that it be monotonically increasing. These factors indicate that not just any arbitrary shape of the relative density is possible, though a wide variety of them are.

2.2.1.2 Connection to Jeffreys Priors

In mentioning reparametrization and priors, it is natural to be reminded of Jeffreys priors, which are typically formulated as "non-informative" priors that are invariant to reparametrization. Jeffreys priors, however, arise from a fundamentally different question than that of the relative distribution, which is of prior selection.

Jeffreys sought a *method* of choosing an "objective" prior. By his reasoning, such a method should lead to the same prior, regardless of the initial parametrization. Given the likelihood, using the method to find a prior and then reparametrizing it should lead to the same prior as if the likelihood was first reparametrized, and then the method was used to find a prior.

In using the relative distribution, we have not made any suggestions as to which prior to use or how to select one. Instead, we assume we have some prior to work with, and then use the relative distribution to analyze how that prior compares to the resulting posterior. Jeffreys priors are a result of seeking "objective" priors, while the unitized likelihood form of the relative distribution is just another way to view an already selected prior. Jeffreys priors are invariant — in the sense described above — to *any* reparametrization, while the relative distribution involves a very specific reparametrization based on the quantile function of the chosen prior.

2.2.2 Relation to Other Comparative Measures

We momentarily return to the general, not necessarily Bayesian case. The (differential) entropy of the relative distribution is

$$\begin{aligned} H(R) &= -\int_0^1 g(r) \log g(r)\ dr \\ &= -\int_0^1 g(r) \log \frac{g(r)}{1}\ dr \end{aligned}$$

This can be interpreted as the negative of the KL divergence of the relative distribution to the uniform distribution. The Kullback-Leibler divergence (KL divergence) of a distribution P to another distribution Q is defined as

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\ dx$$

Bringing the Bayesian context back, the relative distribution provides a different viewpoint of information-theoretic Bayesian statistics. Instead of focusing on the prior and posterior, it emphasizes the likelihood as the agent which connects the two. Supposing our relative distribution compares the posterior to the prior, let's perform the substitution $\theta = K(r)$. The entropy is

$$\begin{aligned} H(R) = -\int_0^1 g(r) \log g(r)\ dr &= \int_0^1 \frac{\pi(K(r)|x)}{\pi(K(r))} \log \frac{\pi(K(r)|x)}{\pi(K(r))}\ dr \\ &= -\int_{\Omega_\Theta} \frac{\pi(\theta|x)}{\pi(\theta)} \log \frac{\pi(\theta|x)}{\pi(\theta)} \pi(\theta)\ d\theta \\ &= -\int_{\Omega_\Theta} \pi(\theta|x) \log \frac{\pi(\theta|x)}{\pi(\theta)}\ d\theta \\ &= -D_{KL}(\pi(\theta|x)||\pi(\theta)) \end{aligned}$$

And we see that this is equal to the negative of the KL-divergence between the posterior and the prior. As the KL divergence is invariant under reparametrization, this result agrees with our view of the relative density as a reparametrization from posterior versus prior to relative density versus standard uniform

Now consider the *mutual information* between X and θ (data and prior), denoted $I(X;\theta)$. This can be interpreted as the amount of "information" — in an entropic sense — present in both X and θ; that is, how much information entropy can be learned from one of them by observing the other.

$$\begin{aligned}
I(X;\theta) &= \int\int p(x,\theta)\log\frac{p(x,\theta)}{p(x)\pi(\theta)}\ d\theta dx \\
&= \int\int p(x)\pi(\theta|x)\log\frac{\pi(\theta|x)p(x)}{p(x)\pi(\theta)}\ d\theta dx \\
&= \int p(x)\int \pi(\theta|x)\log\frac{\pi(\theta|x)}{\pi(\theta)}\ d\theta dx \\
&= \int p(x)D_{KL}(\pi(\theta|x)||\pi(\theta))dx \\
&= \mathbb{E}_X\left[D_{KL}(\pi(\theta|X)||\pi(\theta))\right]
\end{aligned}$$

If we take X as fixed data x, then the expectation reduces to just $D_{KL}(\pi(\theta|x)||\pi(\theta))$, the same quantity as above. Thus the "distance" between the posterior and the prior can also be seen as the amount of "information" that observing X will tell us about the prior.

This quantity is central to the idea of reference priors, as outlined in [BBS09]; see also [Ber79b], [Ber79a], and [BBM88] for more on reference priors. Essentially, a reference prior should maximize this mutual information, which would mean maximizing the KL-divergence between the prior and posterior. That is, the reference prior is the prior which results in the posterior being most "different" from it. As the posterior consists of the prior and the likelihood, maximizing this difference between the prior and posterior means letting the likelihood — the data — have the greatest impact on the final posterior probabilities.

The relative distribution gives us another view of this. Instead of maximizing the KL-divergence between the posterior and the prior, we could imagine instead *minimizing* the entropy of the relative distribution. The entropy should decrease as the relative entropy becomes more and more sharply spiked, which in the limiting case suggests a delta function.

While this may appear as an attractive alternative to finding reference priors, there are some considerations to keep in mind. The first is that the mutual information should be taken as an expectation over the marginal distribution of data $p(x)$, and not simply with assuming the data is constant. The relative density depends on the specific posterior, and will thus depend on the data as well. We would need to expand the definition of the relative distribution to allow for using the expected value of random data points. We could do so directly by substituting the KL-divergence with the entropy of the relative distribution in the formula for mutual information, giving $I(X;\theta) = \mathbb{E}_X\left[H(R)\right]$. Note that unlike before, $H(R)$ is a random quantity depending on the data X, which we no longer assume is a fixed realization x.

Another consideration is that the reference prior is found in the limit of infinite information, $n \to \infty$. This may be infeasible to compute directly, considering the relative density tends to approach a delta function under appropriate regularity conditions. We may have to borrow a similar idea from Bernardo's original approach to finding reference priors, by instead computing a prior based on a finite sample size k in the relative density, then taking the limiting form of that prior.

CHAPTER 3

Closed Forms

3.1 Closed Forms of the Relative Distribution

The relative distribution can be regarded as a function of the prior θ and the data X (giving the likelihood), which together automatically determine the posterior. We assume the data is a collection of n i.i.d. realizations, $x = (x_1, \ldots, x_n)$.

3.1.1 Jeffreys Priors

In many common cases, Jeffreys priors are improper, and thus cannot be analyzed with our current relative distribution methods.

A simple proper Jeffreys prior can be found for a binomial data model. Suppose $X \sim \text{Bin}(m, \theta)$; then the Jeffreys prior is a $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ distribution — the arcsine distribution — with quantile function $K(r) = \sin^2\left(\frac{\pi}{2}r\right)$.

This prior is conjugate for our data, leading to posterior $\text{Beta}(\frac{1}{2} + n\bar{x}, \frac{1}{2} + nm - n\bar{x})$. The relative distribution can be calculated as

$$\begin{aligned} g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\ &= \frac{\pi\Gamma(1+nm)}{\Gamma(\frac{1}{2}+n\bar{x})\Gamma(\frac{1}{2}+nm-n\bar{x})} \left(\sin^2\left(\frac{\pi}{2}r\right)\right)^{n\bar{x}} \left(1 - \sin^2\left(\frac{\pi}{2}r\right)\right)^{nm-n\bar{x}} \end{aligned}$$

See Appendix A.2.1 for a derivation. Below in Figure 3.1, an example of what the relative density looks like in this case is shown. In Figure 3.2, the effect of increasing n is demonstrated.

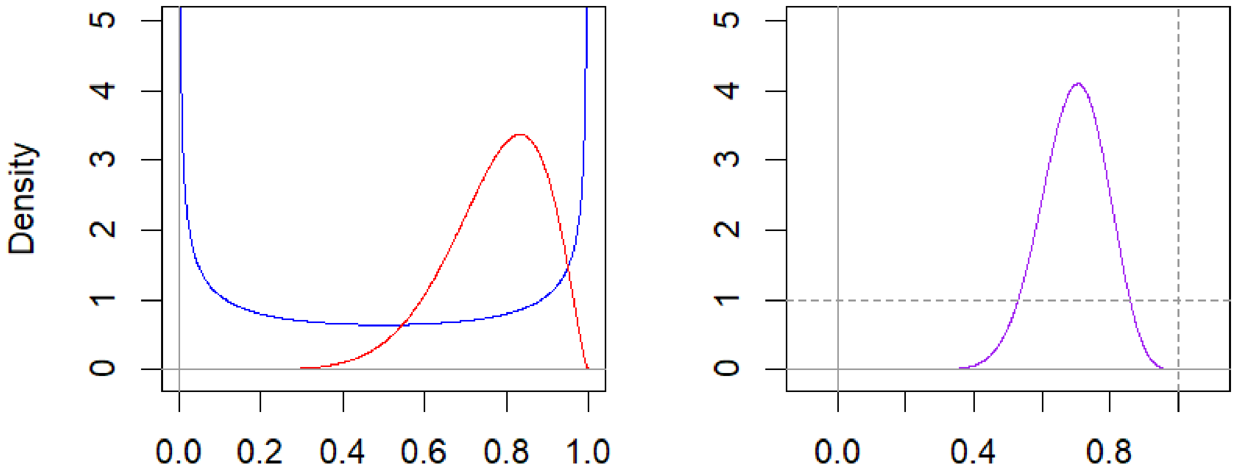

Figure 3.1: In blue, the Jeffreys prior. In red, the posterior. The data was generated with a "true" success rate of 0.8. The relative density is shown on the right, in purple.

3.1.2 With Exponential Families

In the case where data comes from a one-parameter exponential family, i.e.

$$\begin{aligned} p(x|\theta) &= \prod_{i=1}^{n} h(x) \exp\left(\eta(\theta) \cdot T(x_i) - A(\theta)\right) \\ &= h(x)^n \exp\left(\eta(\theta) \cdot \sum_{i=1}^{n} T(x_i) - nA(\theta)\right) \end{aligned}$$

the general conjugate prior distribution for θ as described in [DY79], with hyperparameters (n_0, x_0), is given by

$$\pi(\theta) = h_p(n_0, x_0) \exp\left(n_0 x_0 \eta(\theta) - n_0 A(\theta)\right)$$

The posterior distribution has hyperparameters $\left(n_0 + n, \frac{n_0 x_0 + \sum_{i=1}^{n} T(x_i)}{n_0 + n}\right)$.

The relative density, based on the unitized likelihood, is

$$g(\theta = K(r)) \propto \exp\left(\eta(\theta) \cdot \sum_{i=1}^{n} T(x_i) - nA(\theta)\right)$$

To find a general closed form of this, we must have a closed form for $K(r)$, the quantile function of the prior. As this does not have a general form, we would not be able to

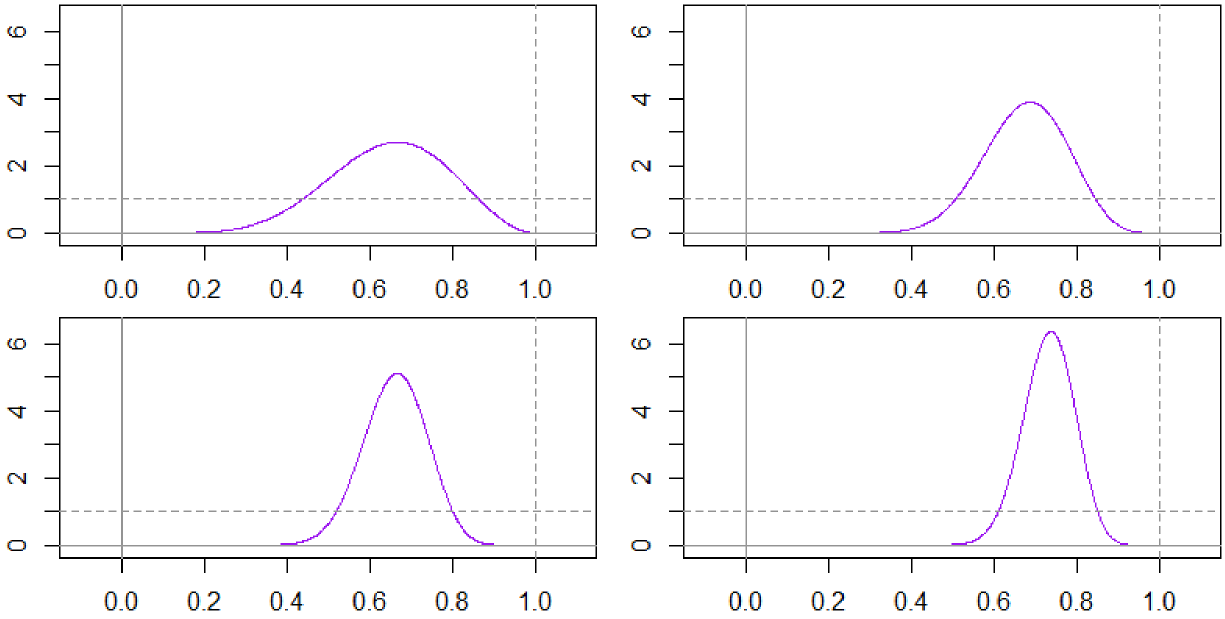

Figure 3.2: The effect of sample sizes on the relative density with a binomial model and Jeffreys prior.

easily find a general closed form. As explained in the previous chapter, the prior's quantile function is critical to the existence of the relative distribution. Below, we instead consider some specific cases where we do have a closed form of $K(r)$.

3.1.3 With a Uniform Prior

Suppose $\theta \sim \text{Unif}(a, b)$. Then $K(r) = a + (b - a)r$. Recall that the relative density has the property $g(r) \propto p(x|K(r))$. By Bayes' Theorem, since the prior is uniform, the posterior will also be proportional to just the likelihood; that is, the likelihood completely determines the shape of the posterior. Our entire belief of the posterior parameters arises from the data.

Consider Gaussian data with an unknown mean and known variance, $X \sim N(\theta, \sigma^2)$.

$$\begin{aligned} g(r) &\propto p(x|\theta = K(r)) \\ &\propto \exp\left(-\frac{(r - (\bar{x} - a)/(b - a))^2}{2\sigma^2/(n(b-a)^2)}\right) \end{aligned}$$

See Appendix A.2.2.1 for a derivation. This has the form of a Gaussian. Note however that r is constrained to lie in $[0, 1]$. Therefore this is actually a truncated Gaussian, $R \sim N\left(\frac{\bar{x}-a}{b-a}, \frac{\sigma^2}{n(b-a)^2}\right)$ with support only in $[0, 1]$.

As an aside, consider the case of $\theta \sim \text{Unif}(-c, c)$, where $c > 0$. As we take the limit $c \to \infty$, we approach a flat improper prior on the real line. From this, we can see that in the distribution of R, the variance approaches 0 while the mean gets relatively closer to the midpoint of 0. Thus the resulting relative density becomes a delta function centered at $r = 0.5$. However, we do not actually obtain any more information from the data, despite the extremely narrow relative density; it's just that as the prior allows more and more possible parameter values, what data we do have will suggest a *relatively* tighter posterior.

A table of several others potential data models is given below in Table 3.1. The exact normalizing constants are omitted for brevity; see the appendix for details.

Data Distribution	Relative Density Kernel	Derivation
$\text{Bin}(m, \theta)$	$r^{n\bar{x}}(1-r)^{1+n(m-\bar{x})}$	A.2.2.2
$\text{Po}(\theta)$	$(a + (b-a)r)^{n\bar{x}}e^{-n\theta}$	A.2.2.3
NegBin (m, θ)	$r^{nm}(1-r)^{n\bar{x}}$	A.2.2.4

Table 3.1: The kernel of the relative density with a uniform prior, under different data models. See the appendix for full normalizing constants and derivation.

3.1.3.1 With a Standard Uniform Prior

Recall that we can interpret the relative density as a reparametrization of the data likelihood, from $\theta \in \mathbb{R}$ to $r \in [0, 1]$. In doing so, applying a standard uniform prior to r gives the same

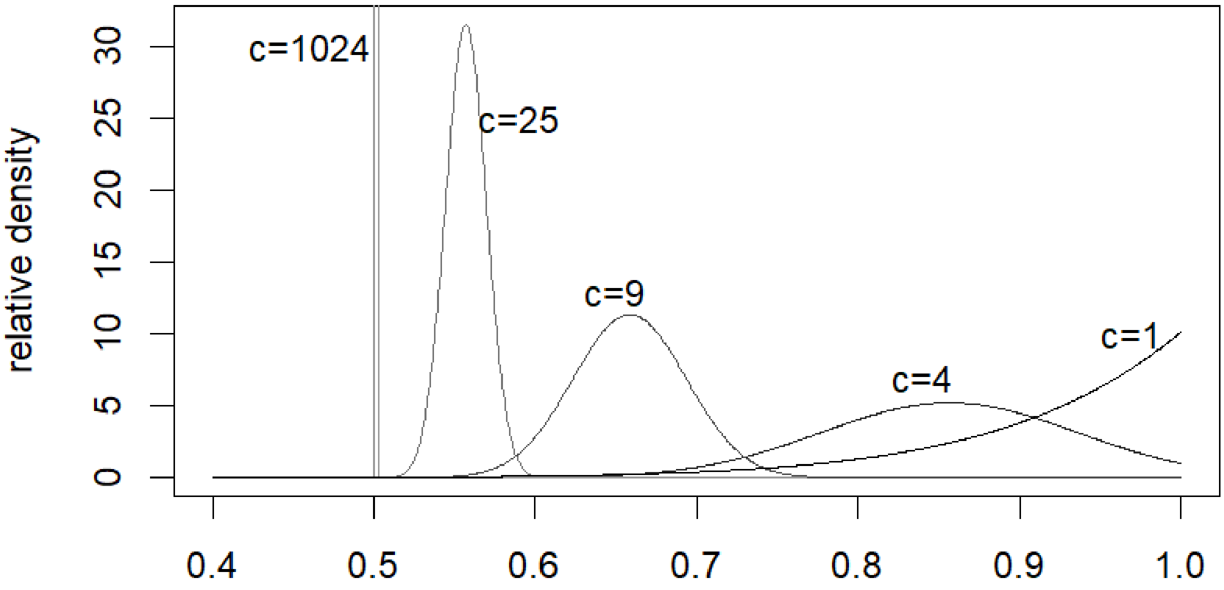

Figure 3.3: The resulting relative density as the uniform prior's support is expanded. A "true" mean value of 4 was used to generate the data, which is why the graphs lean toward upper quantiles; a negative mean would cause the graphs to lean toward lower quantiles. With any mean value, the densities approach the delta function at $r = 0.5$.

marginal data probabilities as with the original prior on θ. Therefore, the case of a standard uniform prior on r is of particular interest.

Of course, this case is also the simplest, as the relative density is directly proportional to the data likelihood itself and also the posterior density, with the data marginal probability as the proportionality constant.

3.1.4 With an Exponential Prior

If $\theta \sim \text{Exp}(\beta)$, then we have $F(\theta) = 1 - e^{-\beta\theta}$ and $K(r) = -\frac{1}{\beta}\log(1-r)$. We can use this as a Gamma conjugate prior ($\theta \sim \text{Gamma}(1, \beta)$).

Using data model $X \sim \text{Poisson}(\theta)$, the relative density has the form

$$\begin{aligned} g(r) &\propto p(x|K(r)) \\ &= \prod_{i=1}^{n} \frac{(K(r))^{x_i}}{x_i!} e^{-K(r)} \\ &\propto \left(-\frac{1}{\beta}\log(1-r)\right)^{n\bar{x}} e^{\frac{n}{\beta}\log(1-r)} \\ &= (-\log(1-r))^{n\bar{x}} (1-r)^{\frac{n}{\beta}} \end{aligned}$$

We can explicitly calculate the constant term using the original definition as well, giving

$$g(r) = \frac{\left(1+\frac{n}{\beta}\right)^{1+n\bar{x}}}{\Gamma(1+n\bar{x})} (-\log(1-r))^{n\bar{x}} (1-r)^{\frac{n}{\beta}}$$

As a side note, if we perform the change of variables $R = \digamma(S)$, we would find that the new variable S has the same distribution as the posterior $\theta|x$, confirming the transformation $R = \digamma(\theta|x)$. See Appendix A.2.3.2 for details.

The resulting relative density with other potential data models is given below in Table 3.2.

Data Distribution X	Relative Density Kernel	Derivation
$N(\mu, 1/\theta^2)$	$(-\log(1-r))^{n/2} (1-r)^{\frac{n}{2\beta}S^2}$	A.2.3.3
$\text{Pareto}(k, \theta)$	$(-\log(1-r))^{n} (1-r)^{\left(\sum_{i=1}^{n}\log\frac{x_i}{k}\right)/\beta}$	A.2.3.4
$\text{Gamma}(m, \theta)$	$(-\log(1-r))^{nm} (1-r)^{n\bar{x}/\beta}$	A.2.3.5

Table 3.2: The kernel of the relative density with an exponential prior, under different data models. See the appendix for full normalizing constants and derivation.

CHAPTER 4

Practical Applications of the Relative Distribution

We should note that it is difficult to draw conclusions about only the posterior distribution when using the relative distribution. It compares the posterior to the prior while implicitly using the prior's quantile scales, and so there is difficulty in completely disentangling the two distributions from each other.

4.1 Examples of Elementary Interpretation

Here we use the relative distribution on a simple example given in [GCS20]. We take a look at an early study of placenta previa births in Germany, of which 437 were female and 544 were male. We are interested in the proportion of female births, which we model as a binomial distribution, and whether it is less than 0.485. The conjugate prior beta distribution is used.

The relative densities for four different priors are shown in Figure 4.1. As the strength of belief of the prior increases, the relative density becomes less peaked, indicating that the data had a smaller impact on our beliefs in the parameter's distribution. Additionally, the relative density places more mass near 0, indicating that the data favors parameter values in the lower quantiles of the prior.

It appears that the relative densities lean towards the posterior intervals, suggesting that the data has more influence on the posterior than the prior does. We can also see that these intervals are either just outside or right on the cusp of the prior median ($r = 0.5$): the posteriors place most of their mass away from the prior medians. Such posteriors suggest

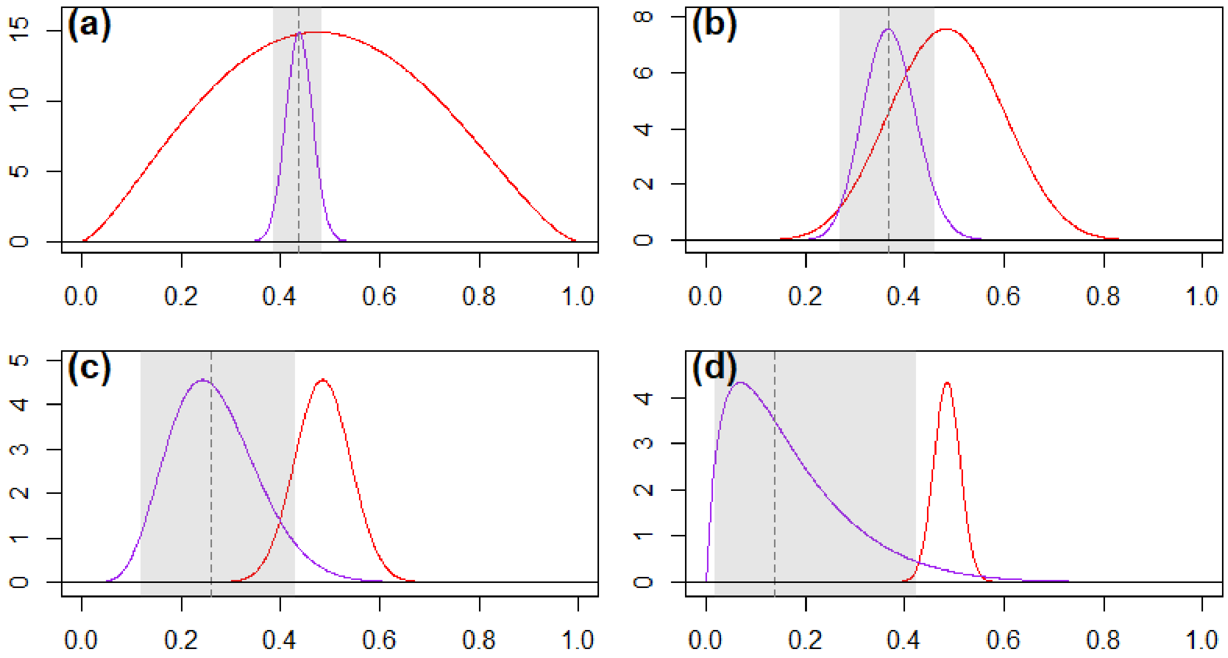

Figure 4.1: The relative density, in purple, of a Beta(2, 2) prior to a Bernoulli rate model for priors at varying strengths of belief but with the same mean 0.485: (a) $\alpha + \beta = 5$, (b) $\alpha + \beta = 20$, (c) $\alpha + \beta = 80$, (d) $\alpha + \beta = 360$. The prior is plotted in red, vertically stretched to fit in the same window as the relative density. The shaded region represents a 95% central posterior interval, and the dashed line represents the posterior median, both scaled based on the prior's quantiles to match the relative density's scale.

that the data *does* support the claim that the true rate is less than 0.485.

Let's zoom in on case (d), which has a strong prior belief on the true rate being near 0.485. The posterior central interval is $[0.4298245, 0.480]$. This, as noted before, does not contain 0.485. Compare this to a central 95% interval for the relative distribution itself, which is $[0.018, 0.472]$, which translates to $[0.430, 0.483]$. This is slightly different to the central posterior interval. There is a very slightly shift in the left endpoint which is not reflected in three decimal places; this leads to a more noticeable shift in the right endpoint because there is less mass at the right endpoint.

The interpretation of this interval as an interval for the posterior, as described in Section 1.1, is still valid. The reason this is different from the central posterior interval is simply because it is a central interval based on the relative distribution, not on the posterior. We note that the posterior interval is based on entirely on the posterior distribution, and thus its quantiles; while the relative distribution is based on the quantiles of the prior distribution.

4.2 Connection to Relative Surprise Inference

We make a quick connection to the *relative belief ratio*, a theory of Bayesian inference which has been developed over several papers by Evans:

$$RB_\theta(\theta|x) = \frac{\pi(\theta|x)}{\pi(\theta)}$$

This has a very similar form to the relative density, with some differences. The value RB_θ is calculated for a single value θ, i.e. at one possible parameter value, while the relative distribution gives a comparison of the entire posterior to the entire prior at corresponding quantiles (of the prior). While the use of the reparametrization $\theta = K(r)$ gives the relative distribution the interesting property of being a proper probability distribution, it may be more cumbersome and unnecessary to always work in the quantiles of the prior. On the other hand, RB_θ is a simple ratio,

As the relative belief ratio is invariant to reparametrizations, the relative density can be used in the same ways to make inferences. Perhaps one of the most immediate applications is as an alternative to Bayes Factors. As noted in [Eva16], if we consider a Bayes Factor distinguishing only in favor an event A and its complement A^C, then

$$BF(A|x) = \frac{RB(A|x)}{RB(A^C|x)}$$

Apart from having additional attractive properties, the RB cannot be expressed in terms of the BF alone, and so seems to be a more fundamental measure than the BF is. See also [Eva97] and [AAE23] for further discussion on the uses of the RB as a tool of inference.

4.3 Comparing Predictive Distributions

While the natural pair of distributions to compare is the posterior to the prior of the parameter itself, we may also consider comparing the posterior predictive to the prior predictive distribution of a new data point. Generally, the predictive distributions cannot be found explicitly.

Let us suppose we have data from an exponential family, and also a conjugate prior from the exponential family. This means

$$\pi(\theta; n_0, x_0) = h_p(n_0, x_0) \exp\left(n_0 x_0 \eta(\theta) - n_0 A(\theta)\right)$$
$$p(x) = h(x)^n \exp\left(\eta(\theta) \cdot \sum_{i=1}^{n} T(x_i) - nA(\theta)\right)$$

and the posterior has parameters (n_1, x_1) with $n_1 = n_0 + n$ and $x_1 = \left(n_0 x_0 + \sum_{i=1}^{n} T(x_i)\right) / (n_0 + n)$.

The prior predictive distribution, essentially the marginal data distribution of one sample, can be found as

$$\begin{aligned} p(\tilde{x}) &= \int_{\Omega_\theta} p(\tilde{x}|\theta) p(\theta)\ d\theta \\ &= \int_{\Omega_\theta} h(\tilde{x}) \exp\left(\eta(\theta) \cdot T(\tilde{x}) - A(\theta)\right) h_p(n_0, x_0) \exp\left(n_0 x_0 \eta(\theta) - n_0 A(\theta)\right)\ d\theta \\ &= h(\tilde{x}) h_p(n_0, x_0) \int_{\Omega_\theta} \exp\left(\eta(\theta) \cdot T(\tilde{x}) + n_0 x_0 \eta(\theta) - A(\theta) - n_0 A(\theta)\right)\ d\theta \end{aligned}$$

We can recognize the integrand as the kernel from the same exponential family as the prior, with parameters $(n_0 x_0 + 1, (n_0 x_0 + T(\tilde{x})) / (n_0 + 1))$, so we can replace the integral with its normalizing constant.

$$= h(\tilde{x}) \frac{h_p(n_0, x_0)}{h_p\left(n_0 x_0 + 1, \frac{n_0 x_0 + T(\tilde{x})}{n_0 + 1}\right)}$$

Similar calculations for the posterior predictive distribution gives

$$p(\tilde{x}|x) = h(\tilde{x}) \frac{h_p(n_1, x_1)}{h_p\left(n_1 x_1 + 1, \frac{n_1 x_1 + T(\tilde{x})}{n_1 + 1}\right)}$$

Then, we take their ratio, but now substitute $\tilde{x} = \tilde{Q}(r)$, the prior predictive (marginal data) quantile function.

$$\begin{aligned} g(\tilde{r}) &= \frac{p(\tilde{x}|x)}{p(\tilde{x})} \\ &= \frac{h_p(n_1, x_1) h_p\left(n_0 x_0 + 1, \frac{n_0 x_0 + T(\tilde{Q}(r))}{n_0 + 1}\right)}{h_p(n_0, x_0) h_p\left(n_1 x_1 + 1, \frac{n_1 x_1 + T(\tilde{Q}(r))}{n_1 + 1}\right)} \end{aligned}$$

Once again, without further information of h_p or $\tilde{Q}$, this is not simplifiable.

As an example, let us consider the normal-normal location model. Suppose the data is $X \sim N(\theta, \sigma^2)$, and the prior is $\theta \sim N(\mu_0, \tau_0^2)$. The prior predictive is known to be distributed as $\tilde{X} \sim N(\mu_0, \tau_0^2 + \sigma^2)$. The posterior distribution is $\theta|x \sim N(\mu_1, \tau_1^2)$, where

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

Thus the posterior predictive distribution is $\tilde{X}|x \sim (\mu_1, \tau_0^2 + \sigma^2/n)$.

An example is given in Figure 4.2. The most striking feature is that the predictive relative density seems less skewed than the parameter relative density, as well as having slightly greater variance. This can be explaind by the fact that predicting a single new data point will be less precise than making inferences about parameters. In predicting the new data point, we not only have the variance from our uncertainty in what the true mean is, but also the inherent variance in the data-generating process. We can see in both cases, however, that the data pulls our estimates towards larger values, which is to be expected, as the “true” mean of the data was greater than our prior’s mean.

Note that the predictive relative density is comparing the distribution of a new sample, and thus is based on the X scale. The original relative density compares the distribution of the parameters, and is based on the θ scale.

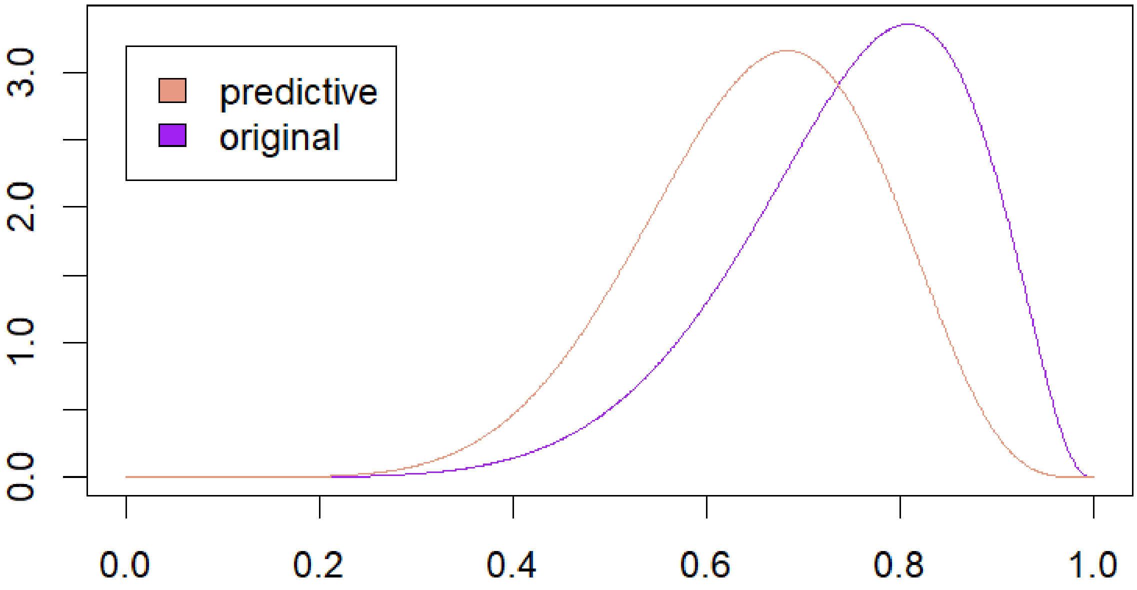

Figure 4.2: The relative density of the predictive distributions, in orange, and the relative density of the parameter posterior to the prior, in purple. For simulation purposes, $\theta \sim N(0, 1)$ and $X \sim N(1, 2)$.

CHAPTER 5

Summary and Discussion

The relative distribution gives us another way to compare two continuous univariate distributions. As a random variable itself, it maps values of a comparison distribution to the reference distribution's quantiles, as though that value came from the reference distribution. Its density is most readily interpretable as a ratio of the comparison and reference densities themselves, while still retaining the property that it integrates to 1.

While it can be used in general on any two distributions, here we focused on a Bayesian setting and compared the posterior to the prior. The relative density here reduces to essentially just the data likelihood, though compressed into a unit interval using the prior's quantiles. An interesting alternative perspective of the relative distribution is as a reparametrization of the model's parameter using the prior's CDF. A standard uniform prior on the reparametrization induces the same marginal data probabilities as the original prior on the original parameter.

There are several limitations to using the relative distribution. While thinking in terms of quantiles allows us to think relatively on the unit interval, it also necessitates the existence of a strictly monotonic CDF in order to use a proper quantile function of the prior, which is not easily obtained for most distributions. Modern computational tools can alleviate this issue to some extent. An extension to non-continuous distributions, such as purely discrete ones, would require some new idea to be used. Additionally, the relative distribution becomes less insightful as sampling variance decreases, such as when sample sizes become larger. The increased precision makes the relative error smaller, and differences are less noticeable when viewed on the relative distribution.

Potential applications of the relative distribution include its connections to KL-divergence and reference priors. The entropy of the relative density is the negative KL-divergence between the posterior and prior. We could leverage this fact as an alternative to deriving reference priors by minimizing the entropy of the relative density with respect to infinite sample sizes and expected data values. Another application is in direct connection to relative surprise inference, for which the relative belief ratio is simply another parametrization of the relative density, though typically evaluated at single points and not for all prior quantiles. Whether there are additional interesting properties of RB_θ to be found when considering that the relative distribution is a proper probability distribution is another potential avenue of research.

Further work could include alternatives to using the quantile function and applications to sensitivity analysis. If the limitation of using the quantile function can be overcome, then even non-continuous distributions could be analyzed. We could use improper priors as well. For sensitivity analysis, as the purpose of the relative distribution is to compare how the posterior differes from the prior, with some slight tweaking we should be able to observe how different options for the prior affect the resulting posterior. One potential complication is that all results are cast into the quantiles of the prior, and thus different options for the prior may not be immediately comparable to each other.

We make a quick note on Bayes factors. A general Bayes factor is essentially the ratio of the likelihoods under two different models, not necessarily complement to each other. The relative density is the transformed likelihood, and so a ratio of two relative densities from two different models could serve a similar role as a Bayes factor. However, since the priors for the two different models will differ, the relative densities will also be on a different quantile scale, and so a further refinement must be made to directly compare relative densities themselves.

APPENDIX A

Calculations

A.1 Chapter 2

A.1.1 Normal data likelihood

$$
\begin{aligned}
p(x|\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - 2\theta x_i + \theta^2\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{n}{2\sigma^2}\left(\theta^2 - 2\theta\bar{x} + \overline{x^2}\right)\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2/n}\left((\theta-\bar{x})^2 - \bar{x}^2 + \overline{x^2}\right)\right) \\
&= \frac{\exp\left(-\frac{\overline{x^2}-\bar{x}^2}{2\sigma^2/n}\right)}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{(\theta-\bar{x})^2}{2\sigma^2/n}\right) = \mathcal{L}(\theta|x)
\end{aligned}
$$

A.2 Chapter 3

A.2.1 Jeffreys Prior for Binomial Data

The Fisher Information can be found as

$$\begin{aligned}
I(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(x|\theta)\right] \\
&= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log\binom{m}{x_i}\theta^{x_i}(1-\theta)^{m-x_i}\right] \\
&= -\mathbb{E}\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\frac{x_i}{\theta} - \frac{m-x_i}{1-\theta}\right] \\
&= -\mathbb{E}\left[\sum_{i=1}^{n} -\frac{x_i}{\theta^2} - \frac{m-x_i}{(1-\theta)^2}\right] \\
&= \frac{mn}{\theta} + \frac{mn}{1-\theta} \\
&= \frac{mn}{\theta(1-\theta)}
\end{aligned}$$

which implies that the Jeffreys prior satisfies $\pi(\theta) \propto (\theta(1-\theta))^{-\frac{1}{2}}$, leading to a Beta$(\frac{1}{2}, \frac{1}{2})$ distribution, also known as the arcsine distribution.

We find the proportionality constant by taking the integral of the kernel on $[0, 1]$, and substitute $\theta = \frac{1}{2} + \frac{1}{2}x$.

$$\begin{aligned}
\int_0^1 \frac{1}{\sqrt{\theta(1-\theta)}}\, d\theta &= \int_{-1}^{1} \frac{1}{\sqrt{\left(\frac{1}{2}-\frac{1}{2}x\right)\left(\frac{1}{2}-\frac{1}{2}x\right)}}\frac{1}{2}\, dx \\
&= \int_{-1}^{1} \frac{1}{2\sqrt{\frac{1}{4}-\frac{1}{4}x^2}}\, dx \\
&= (\arcsin x|_{-1}^{1} = \pi
\end{aligned}$$

Then we can find the CDF using the substitution $t = \sin^2 x \implies dt = 2\sin x \cos x dx$.

$$\begin{aligned}
F(\theta) &= \int_0^\theta \frac{1}{\pi\sqrt{t(1-t)}}\ dt \\
&= \int_0^{\arcsin\sqrt{\theta}} \frac{1}{\pi\sqrt{\sin^2 x(1-\sin^2 x)}} 2\sin x \cos x\ dx \\
&= \int_0^{\arcsin\sqrt{\theta}} \frac{2}{\pi}\ dx \\
&= \frac{2}{\pi}\arcsin\sqrt{\theta}
\end{aligned}$$

which implies that the quantile function is $K(r) = \sin^2\left(\frac{\pi}{2}r\right)$.

For the relative distribution,

$$\begin{aligned}
g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\
&= \frac{\frac{\Gamma(1+nm)}{\Gamma(\frac{1}{2}+n\bar{x})\Gamma(\frac{1}{2}+nm-n\bar{x})}\theta^{\frac{1}{2}+n\bar{x}-1}(1-\theta)^{\frac{1}{2}+nm-n\bar{x}-1}}{\frac{1}{\pi}\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}} \\
&= \frac{\pi\Gamma(1+nm)}{\Gamma(\frac{1}{2}+n\bar{x})\Gamma(\frac{1}{2}+nm-n\bar{x})}\theta^{n\bar{x}}(1-\theta)^{nm-n\bar{x}} \\
&= \frac{\pi\Gamma(1+nm)}{\Gamma(\frac{1}{2}+n\bar{x})\Gamma(\frac{1}{2}+nm-n\bar{x})}\left(\sin^2\left(\frac{\pi}{2}r\right)\right)^{n\bar{x}}\left(1-\sin^2\left(\frac{\pi}{2}r\right)\right)^{nm-n\bar{x}}
\end{aligned}$$

A.2.2 Uniform prior

We have $\pi(\theta) = \frac{1}{b-a} \implies K(r) = a + (b-a)r$. Using the standard uniform is special case of Beta$(1,1)$.

With a standard uniform for a proportion parameter, the resulting relative distribution looks exactly like the likelihood. No change in scale, all information in posterior comes from the data, flat prior.

A.2.2.1 Normal data

$$\begin{aligned}
g(r) &\propto p(x|\theta = K(r)) \\
&\propto \prod_{i=1}^{n} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right) \\
&= \exp\left(\sum_{i=1}^{n} -\frac{(x_i-(a+(b-a)r))^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(n(b-a)^2r^2 + 2\left(na(b-a)-(b-a)\sum_{i=1}^{n} x_i\right)r\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2/(n(b-a))^2}\left(r^2 + 2\left(\frac{a}{b-a}-\frac{\bar{x}}{b-a}\right)r\right)\right) \\
&\propto \exp\left(-\frac{(r-(\bar{x}-a)/(b-a))^2}{2\sigma^2/(n(b-a)^2)}\right)
\end{aligned}$$

A.2.2.2 Binomial data

$X \sim \text{Bin}(m,\theta)$; set $a = 0, b = 1$.

$$\begin{aligned}
\pi(\theta|x) &\propto \prod_{i=1}^{n} \binom{m}{x_i} \theta^{x_i}(1-\theta)^{m-x_i} \\
&\propto \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{nm-\sum_{i=1}^{n} x_i} \\
\implies \theta|x &\sim \text{Beta}(1+n\bar{x}, 1+n(m-\bar{x}))
\end{aligned}$$

$$\begin{aligned}
g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\
&= \frac{\Gamma(2+nm)}{\Gamma(1+n\bar{x})\Gamma(1+n(m-\bar{x}))}\theta^{n\bar{x}}(1-\theta)^{1+n(m-\bar{x})} \times (1-0) \\
&= \frac{\Gamma(2+nm)}{\Gamma(1+n\bar{x})\Gamma(1+n(m-\bar{x}))}r^{n\bar{x}}(1-r)^{1+n(m-\bar{x})}
\end{aligned}$$

This is a $\text{Beta}(1+n\bar{x}, 1+n(m-\bar{x}))$ distribution.

A.2.2.3 Poisson data

$X \sim \text{Po}(\theta)$.

$$\begin{aligned}\pi(\theta|x) &\propto \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} e^{-\theta} \\ &\propto \theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}\end{aligned}$$

Form of a Gamma, but note that $\theta \in [a, b]$, so this is a truncated Gamma. Denote C as the mass of an untruncated Gamma in this interval, so that $\frac{1}{C}$ is the additional normalizing factor in the truncated Gamma.

$$\begin{aligned}g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\ &= \frac{1}{C} \frac{n^{n\bar{x}}}{\Gamma(n\bar{x})} (a + (b-a)r)^{\sum_{i=1}^{n} x_i} e^{-n\theta} \times (b-a) \\ &= \frac{b-a}{C} \frac{n^{n\bar{x}}}{\Gamma(n\bar{x})} (a + (b-a)r)^{\sum_{i=1}^{n} x_i} e^{-n\theta}\end{aligned}$$

A.2.2.4 Negative binomial data

$X \sim \text{NegBin}(m, \Theta)$; set $a = 0, b = 1$.

$$\begin{aligned}\pi(\theta|x) &\propto \prod_{i=1}^{n} \binom{x_i + m - 1}{x_i} (1-\theta)^{x_i} \theta^m \\ &\propto \theta^{nm} (1-\theta)^{n\bar{x}} \\ \implies \theta|x &\sim \text{Beta}(1 + nm, 1 + n\bar{x})\end{aligned}$$

$$\begin{aligned}g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\ &= \frac{\Gamma(2 + nm + n\bar{x})}{\Gamma(1 + nm)\Gamma(1 - n\bar{x})} \theta^{nm} (1-\theta)^{n\bar{x}} \times (1 - 0) \\ &= \frac{\Gamma(2 + nm + n\bar{x})}{\Gamma(1 + nm)\Gamma(1 - n\bar{x})} r^{nm} (1-r)^{n\bar{x}} \times (1 - 0)\end{aligned}$$

This is a Beta$(1 + nm, 1 + n\bar{x})$ distribution.

A.2.3 Gamma (Exponential) prior

Using a Gamma$(1, \beta) = \text{Exp}(\beta)$ prior. $K(r) = -\frac{1}{\beta}\log(1-r)$.

A.2.3.1 Poisson data

$X \sim \text{Po}(\theta)$. Posterior is $\theta | x \sim \text{Gamma}(1 + n\bar{x}, \beta + n)$.

$$\begin{aligned}
g(\theta = K(r)) &= \frac{\frac{(\beta+n)^{1+n\bar{x}}}{\Gamma(1+n\bar{x})} \left(K(r)\right)^{1+n\bar{x}-1} e^{-(\beta+n)K(r)}}{\beta e^{-\beta K(r)}} \\
&= \frac{(\beta+n)^{1+n\bar{x}}}{\beta\Gamma(1+n\bar{x})} \left(K(r)\right)^{n\bar{x}} e^{-nK(r)} \\
&= \frac{(\beta+n)^{1+n\bar{x}}}{\beta\Gamma(1+n\bar{x})} \left(-\frac{1}{\beta}\log(1-r)\right)^{n\bar{x}} e^{-n\left(-\frac{1}{\beta}\log(1-r)\right)} \\
&= \frac{\left(1+\frac{n}{\beta}\right)^{1+n\bar{x}}}{\Gamma(1+n\bar{x})} \left(-\log(1-r)\right)^{n\bar{x}} (1-r)^{\frac{n}{\beta}}
\end{aligned}$$

A.2.3.2 Change of variables $R = F(S)$

If we perform the change of variables $-\frac{1}{\beta}\log(1-R) = S \implies R = 1 - e^{-\beta S}$, the resulting density becomes

$$\begin{aligned}
f(s) &= g(r = 1 - e^{-s}) \left|\frac{dr}{ds}\right| \\
&= \frac{\left(1+\frac{n}{\beta}\right)^{1+n\bar{x}}}{\Gamma(1+n\bar{x})} \left(-\log(1-r)\right)^{n\bar{x}} (1-r)^{\frac{n}{\beta}} \times \left|\beta e^{-\beta s}\right| \\
&= \frac{\beta\left(1+\frac{n}{\beta}\right)^{1+n\bar{x}}}{\Gamma(1+n\bar{x})} (\beta s)^{n\bar{x}} e^{-ns} e^{-\beta s} \\
&= \frac{\beta^{1+n\bar{x}}\left(1+\frac{n}{\beta}\right)^{1+n\bar{x}}}{\Gamma(1+n\bar{x})} s^{n\bar{x}} e^{-(n+\beta)s} \\
&= \frac{(\beta+n)^{1+n\bar{x}}}{\Gamma(1+n\bar{x})} s^{n\bar{x}} e^{-(\beta+n)s}
\end{aligned}$$

Therefore $S \sim \text{Gamma}\left(1 + n\bar{x}, \beta + n\right)$, and R is a transformation of this.

A.2.3.3 Normal precision, known mean μ

Denote $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$. Posterior $\theta|x \sim \text{Gamma}(\alpha + n/2, \beta + nS^2/2)$.

$$\begin{aligned}
g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\
&= \frac{\frac{\left(\beta+\frac{1}{2}nS^2\right)^{1+n/2}}{\Gamma(1+n/2)}\theta^{n/2}e^{-\left(\beta+\frac{n}{2}S^2\right)\theta}}{\beta e^{-\beta\theta}} \\
&= \frac{\left(\beta+\frac{n}{2}S^2\right)^{1+n/2}}{\beta\Gamma(1+n/2)}\theta^{n/2}e^{-\left(\frac{n}{2}S^2\right)\theta} \\
&= \frac{\left(\beta+\frac{n}{2}S^2\right)^{1+n/2}}{\beta\Gamma(1+n/2)}\left(-\frac{1}{\beta}\log(1-r)\right)^{n/2} e^{-\left(\frac{n}{2}S^2\right)\left(-\frac{1}{\beta}\log(1-r)\right)} \\
&= \left(\frac{\beta+\frac{n}{2}S^2}{\beta}\right)^{1+n/2}\frac{1}{\Gamma(1+n/2)}\left(-\log(1-r)\right)^{n/2}(1-r)^{\frac{n}{2\beta}S^2}
\end{aligned}$$

A.2.3.4 Pareto shape, known minimum k

$p(x|\theta) = \frac{\theta k^\theta}{x^{\theta+1}}, x \geq k$. Posterior $\theta|x \sim \text{Gamma}\left(\alpha + n, \beta + \sum_{i=1}^{n}\log\frac{x_i}{k}\right)$.

$$\begin{aligned}
g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\
&= \frac{\frac{\left(\beta+\sum_{i=1}^{n}\log\frac{x_i}{k}\right)^{\alpha+n}}{\Gamma(\alpha+n)}\theta^{\alpha+n-1}e^{-\left(\beta+\sum_{i=1}^{n}\log\frac{x_i}{k}\right)\theta}}{\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}} \\
&= \frac{\left(\beta+\sum_{i=1}^{n}\log\frac{x_i}{k}\right)^{\alpha+n}\Gamma(\alpha)}{\beta^\alpha\Gamma(\alpha+n)}\theta^{n}e^{-\left(\sum_{i=1}^{n}\log\frac{x_i}{k}\right)\theta} \\
&= \frac{\left(\beta+\sum_{i=1}^{n}\log\frac{x_i}{k}\right)^{\alpha+n}\Gamma(\alpha)}{\beta^\alpha\Gamma(\alpha+n)}\left(-\frac{1}{\beta}\log(1-r)\right)^{n} e^{-\left(\sum_{i=1}^{n}\log\frac{x_i}{k}\right)\left(-\frac{1}{\beta}\log(1-r)\right)} \\
&= \left(\frac{\beta+\sum_{i=1}^{n}\log\frac{x_i}{k}}{\beta}\right)^{\alpha+n}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\left(-\log(1-r)\right)^{n}(1-r)^{\left(\sum_{i=1}^{n}\log\frac{x_i}{k}\right)/\beta}
\end{aligned}$$

A.2.3.5 Gamma rate

$X \sim \text{Gamma}(m, \Theta)$, the posterior is $\theta|x \sim \text{Gamma}(\alpha + nm, \beta + n\bar{x})$. Then the relative density is

$$\begin{aligned}
g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\
&= \frac{\frac{(\beta+n\bar{x})^{\alpha+nm}}{\Gamma(\alpha+nm)}\theta^{\alpha+nm-1}e^{-(\beta+n\bar{x})\theta}}{\frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}} \\
&= \frac{(\beta+n\bar{x})^{\alpha+nm}\Gamma(\alpha)}{\beta^{\alpha}\Gamma(\alpha+nm)}\theta^{nm}e^{-(n\bar{x})\theta} \\
&= \frac{(\beta+n\bar{x})^{\alpha+nm}\Gamma(\alpha)}{\beta^{\alpha}\Gamma(\alpha+nm)}\left(-\frac{1}{\beta}\log(1-r)\right)^{nm} e^{-(n\bar{x})\left(-\frac{1}{\beta}\log(1-r)\right)} \\
&= \left(\frac{\beta+n\bar{x}}{\beta}\right)^{\alpha+nm}\frac{\Gamma(\alpha)}{\Gamma(\alpha+nm)}\left(-\log(1-r)\right)^{nm}(1-r)^{n\bar{x}/\beta}
\end{aligned}$$

A.2.4 Misc priors

A.2.4.1 Pareto prior to uniform data

$\pi(\theta) = \frac{\alpha x_0^{\alpha}}{\theta^{\alpha+1}}\mathbf{1}\{\theta \geq x_0\}$, $f(x|\theta) = \prod_{i=1}^{n}\frac{1}{\theta}\mathbf{1}\{x_i \leq \theta\}$. $K(r) = \frac{x_0}{(1-r)^{\frac{1}{\alpha}}}$.

$$\begin{aligned}
\pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \\
&= \frac{1}{\theta^n}\mathbf{1}\{x_m \leq \theta\}\frac{\alpha x_0^{\alpha}}{\theta^{\alpha+1}}\mathbf{1}\{\theta \geq x_0\} \\
&\propto \frac{1}{\theta^{\alpha+n+1}}\mathbf{1}\{\theta \geq x_1\}
\end{aligned}$$

implies $\pi(\theta|x) = \frac{(\alpha+n)x_1^{\alpha+n}}{\theta^{\alpha+n+1}}\mathbf{1}\{\theta \geq x_1\}$.

$$\begin{aligned}
g(\theta = K(r)) &= \frac{\pi(\theta|x)}{\pi(\theta)} \\
&= \frac{\frac{(\alpha+n)x_1^{\alpha+n}}{\theta^{\alpha+n+1}}\mathbf{1}\{\theta \geq x_1\}}{\frac{\alpha x_0^{\alpha}}{\theta^{\alpha+1}}\mathbf{1}\{\theta \geq x_0\}} \\
&= \frac{\alpha+n}{\alpha}\frac{x_1^{\alpha+n}}{x_0^{\alpha}}\theta^{-n}\mathbf{1}\{\theta > x_1\} \\
&= \frac{\alpha+n}{\alpha}\frac{x_1^{\alpha+n}}{x_0^{\alpha}}\left(\frac{(1-r)^{\frac{1}{\alpha}}}{x_0}\right)^{n}\mathbf{1}\left\{\frac{x_0}{(1-r)^{\frac{1}{\alpha}}} > x_1\right\} \\
&= \frac{\alpha+n}{\alpha}\left(\frac{x_1}{x_0}\right)^{\alpha+n}(1-r)^{\frac{n}{\alpha}}\mathbf{1}\left\{r > 1-\left(\frac{x_0}{x_1}\right)^{\alpha}\right\}
\end{aligned}$$

REFERENCES

[AAE23] Luai Al-Labadi, Ayman Alzaatreh, and Michael Evans. "How to Measure Evidence: Bayes Factors or Relative Belief Ratios?" 2023.

[BBM88] J. Berger, Jose Bernardo, and Manuel Mendoza. "On priors that maximize expected information." 01 1988.

[BBS09] James O. Berger, José M. Bernardo, and Dongchu Sun. "The formal definition of reference priors." *The Annals of Statistics*, **37**(2), April 2009.

[Ber79a] Jose M. Bernardo. "Expected Information as Expected Utility." *The Annals of Statistics*, **7**(3), May 1979.

[Ber79b] Jose M. Bernardo. "Reference Posterior Distributions for Bayesian Inference." *Journal of the Royal Statistical Society: Series B (Methodological)*, **41**(2):113–128, January 1979.

[DY79] Persi Diaconis and Donald Ylvisaker. "Conjugate Priors for Exponential Families." *The Annals of Statistics*, **7**(2), March 1979.

[Eva97] Michael Evans. "Bayesian ikference procedures derived via the concept of relative surprise." *Communications in Statistics - Theory and Methods*, **26**(5):1125–1143, January 1997.

[Eva16] Michael Evans. "Measuring statistical evidence using relative belief." *Computational and Structural Biotechnology Journal*, **14**:91–96, 2016.

[GCS20] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis.* 2020.