Bayesian modelling

Léo Belzile

Table of contents

Welcome

This book is a web complement to MATH 80601A *Bayesian modelling*, a graduate course offered at HEC Montréal.

The objective of the course is to provide a hands on introduction to Bayesian data analysis. The course will cover the formulation, evaluation and comparison of Bayesian models through examples and real-data applications.

1 Bayesics

The Bayesian paradigm is an inferential framework that is used widespread in data science. Numerical challenges that prevented it's widespread adoption until the 90's, when the Markov chain Monte Carlo revolution allowed models estimation.

Bayesian inference, which builds on likelihood-based inference, offers a natural framework for prediction and for uncertainty quantification. The interpretation is more natural than that of classical (i.e., frequentist) paradigm, and it is more easy to generalized models to complex settings, notably through hierarchical constructions. The main source of controversy is the role of the prior distribution, which allows one to incorporate subject-matter expertise but leads to different inferences being drawn by different practitioners; this subjectivity is not to the taste of many and has been the subject of many controversies.

The Bayesian paradigm includes multiples notions that are not covered in undergraduate introductory courses. The purpose of this chapter is to introduce these concepts and put them in perspective; the reader is assumed to be familiar with basics of likelihood-based inference. We begin with a discussion of the notion of probability, then define priors, posterior distributions, marginal likelihood and posterior predictive distributions. We focus on the interpretation of posterior distributions and explain how to summarize the posterior, leading leading to definitions of high posterior density region, credible intervals, posterior mode for cases where we either have a (correlated) sample from the posterior, or else have access to the whole distribution. Several notions, including sequentiality, prior elicitation and estimation of the marginal likelihood, are mentioned in passing. A brief discussion of Bayesian hypothesis testing (and alternatives) is presented.

1.1 Probability and frequency

In classical (frequentist) parametric statistic, we treat observations $\boldsymbol{Y}$ as realizations of a distribution whose parameters $\boldsymbol{\theta}$ are unknown. All of the information about parameters is encoded by the likelihood function.

The interpretation of probability in the classical statistic is in terms of long run frequency, which is why we term this approach frequentist statistic. Think of a fair die: when we state that values $\{1,\dots,6\}$ are equiprobable, we mean that repeatedly tossing the die

should result, in large sample, in each outcome being realized roughly $1/6$ of the time (the symmetry of the object also implies that each facet should be equally likely to lie face up). This interpretation also carries over to confidence intervals: a $(1-\alpha)$ confidence interval either contains the true parameter value or it doesn't, so the probability level $(1-\alpha)$ is only the long-run proportion of intervals created by the procedure that should contain the true fixed value, not the probability that a single interval contains the true value. This is counter-intuitive to most.

In practice, the true value of the parameter $\boldsymbol{\theta}$ vector is unknown to the practitioner, thus uncertain: Bayesians would argue that we should treat the latter as a random quantity rather than a fixed constant. Since different people may have different knowledge about these potential values, the prior knowledge is a form of **subjective probability**. For example, if you play cards, one person may have recorded the previous cards that were played, whereas other may not. They thus assign different probability of certain cards being played. In Bayesian inference, we consider $\boldsymbol{\theta}$ as random variables to reflect our lack of knowledge about potential values taken. Italian scientist Bruno de Finetti, who is famous for the claim "Probability does not exist' ', stated in the preface of Finetti (1974):

> Probabilistic reasoning — always to be understood as subjective — merely stems from our being uncertain about something. It makes no difference whether the uncertainty relates to an unforseeable future, or to an unnoticed past, or to a past doubtfully reported or forgotten: it may even relate to something more or less knowable (by means of a computation, a logical deduction, etc.) but for which we are not willing or able tho make the effort; and so on [...] The only relevant thing is uncertainty — the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense *determined*, or known by other people, and so on, is of no consequence.

On page 3, de Finetti continues (Finetti 1974)

> only subjective probabilities exist — i.e., the degree of belief in the occurrence of an event attributed by a given person at a given instant and with a given set of information.

1.2 Posterior distribution

We consider a parametric model with parameters $\boldsymbol{\theta}$ defined on $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$. In Bayesian learning, we adjoin to the likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) \equiv p(\boldsymbol{y} \mid \boldsymbol{\theta})$ a **prior** function $p(\boldsymbol{\theta})$ that reflects the prior knowledge about potential values taken by the p-dimensional parameter vector, before

observing the data $\boldsymbol{y}$. The prior makes $\boldsymbol{\theta}$ random and the distribution of the parameter reflects our uncertainty about the true value of the model parameters.

In a Bayesian analysis, observations are random variables but inference is performed conditional on the observed sample values. By Bayes' theorem, our target is therefore the posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, defined as

$$\underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{\theta})}^{\text{likelihood}} \times \overbrace{p(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}_{\text{marginal likelihood } p(\boldsymbol{y})}}. \tag{1.1}$$

The posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is proportional, as a function of θ, to the product of the likelihood and the prior function.

For the posterior to be **proper**, we need the product of the prior and the likelihood on the right hand side to be integrable as a function of $\boldsymbol{\theta}$ over the parameter domain $\boldsymbol{\Theta}$. The integral in the denominator, termed marginal likelihood or prior predictive distribution and denoted $p(\boldsymbol{y}) = \mathsf{E}_{\boldsymbol{\theta}}\{p(\boldsymbol{y} \mid \boldsymbol{\theta})\}$. It represents the distribution of the data before data collection, the respective weights being governed by the prior probability of different parameters values. The denominator of Equation 1.1 is a normalizing constant, making the posterior density integrate to unity. The marginal likelihood plays a central role in Bayesian testing.

If $\boldsymbol{\theta}$ is low dimensional, numerical integration such as quadrature methods can be used to compute the marginal likelihood.

To fix ideas, we consider next a simple one-parameter model where the marginal likelihood can be computed explicitly.

Example 1.1 (Binomial model with beta prior)**.** Consider a binomial likelihood with probability of success $\theta \in [0, 1]$ and n trials, $Y \sim \mathsf{Bin}(n, \theta)$. If we take a beta prior, $\theta \sim \mathsf{Be}(\alpha, \beta)$ and observe y successes, the posterior is

$$\begin{aligned} p(\theta \mid y = y) &\propto \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\overset{\theta}{\propto} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

and is

$$\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \mathrm{d}\theta = \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)},$$

a Beta function. Since we need only to keep track of the terms that are function of the parameter θ, we could recognize directly that the posterior distribution is $\mathsf{Be}(y+\alpha, n-y+\beta)$ and deduce the normalizing constant from there.

If $Y \sim \mathsf{Bin}(n,\theta)$, the expected number of success is $n\theta$ and the expected number of failures $n(1-\theta)$ and so the likelihood contribution, relative to the prior, will dominate as the sample size n grows.

Another way to see this is to track moments (expectation, variance, etc.) The Beta distribution, whose density is $f(x;\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$, has expectation $\alpha/(\alpha+\beta)$ and variance $\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$. The posterior mean is

$$\mathsf{E}(\theta \mid y) = w\frac{y}{n} + (1-w)\frac{\alpha}{\alpha+\beta}, \qquad w = \frac{n}{n+\alpha+\beta},$$

a weighted average of the maximum likelihood estimator and the prior mean. We can think of the parameter α (respectively β) as representing the fixed prior number of success (resp. failures). The variance term is $\mathrm{O}(n^{-1})$ and, as the sample size increases, the likelihood weight w dominates.

Figure 1.1 shows three different posterior distributions with different beta priors: the first prior, which favors values closer to 1/2, leads to a more peaked posterior density, contrary to the second which is symmetric, but concentrated toward more extreme values near endpoints of the support. The rightmost panel is truncated: as such, the posterior is zero for any value of θ beyond 1/2 and so the posterior mode may be close to the endpoint of the prior. The influence of such a prior will not necessarily vanish as sample size and should be avoided, unless there are compelling reasons for restricting the domain.

Remark (Proportionality). Any term appearing in the likelihood times prior function that does not depend on parameters can be omitted since they will be absorbed by the normalizing constant. This makes it useful to compute normalizing constants or likelihood ratios.

Remark. An alternative parametrization for the beta distribution sets $\alpha = \mu\kappa$, $\beta = (1-\mu)\kappa$ for $\mu \in (0,1)$ and $\kappa > 0$, so that the model is parametrized directly in terms of mean μ, with κ capturing the dispersion.

Remark. A density integrates to 1 over the range of possible outcomes, but there is no guarantee that the likelihood function, as a function of $\boldsymbol{\theta}$, integrates to one over the parameter domain $\boldsymbol{\Theta}$.

For example, the binomial likelihood with n trials and y successes satisfies

$$\int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \mathrm{d}\theta = \frac{1}{n+1}.$$

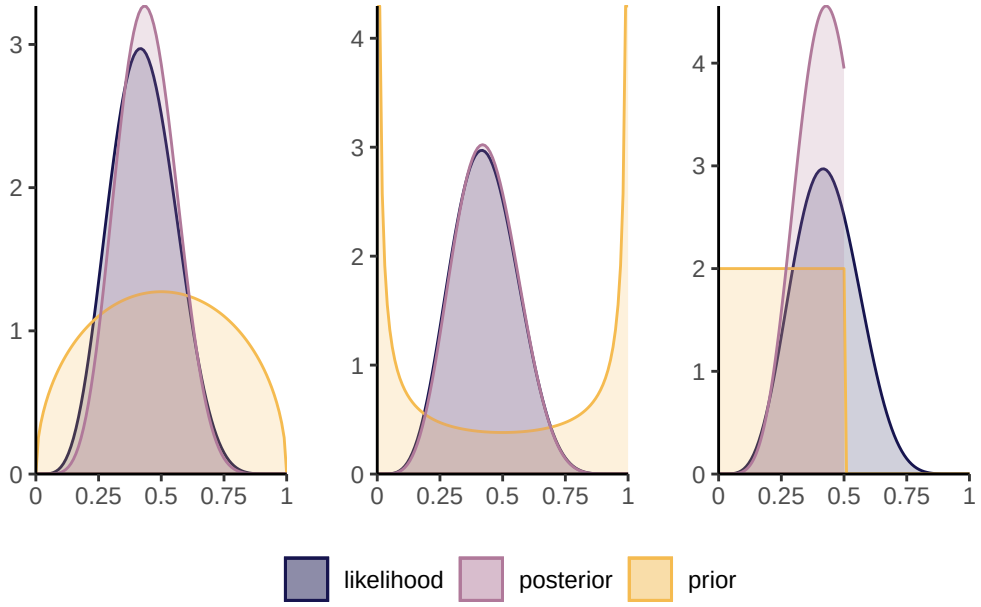

Figure 1.1: Scaled binomial likelihood for six successes out of 14 trials, with $\mathsf{Beta}(3/2, 3/2)$ prior (left), $\mathsf{Beta}(1/4, 1/4)$ (middle) and truncated uniform on $[0, 1/2]$ (right), with the corresponding posterior distributions.

Moreover, the binomial distribution is discrete with support $0, \ldots, n$, whereas the likelihood is continuous as a function of the probability of success, as evidenced by Figure 1.2

Proposition 1.1 (Sequentiality and Bayesian updating)**.** *The likelihood is invariant to the order of the observations if they are independent Thus, if we consider two blocks of observations* $\boldsymbol{y}_1$ *and* $\boldsymbol{y}_2$

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = p(\boldsymbol{\theta} \mid \boldsymbol{y}_1) p(\boldsymbol{\theta} \mid \boldsymbol{y}_2),$$

so it makes no difference if we treat data all at once or in blocks. More generally, for data exhibiting spatial or serial dependence, it makes sense to consider rather the conditional (sequential) decomposition

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = f(\boldsymbol{y}_1; \boldsymbol{\theta}) f(\boldsymbol{y}_2; \boldsymbol{\theta}, \boldsymbol{y}_1) \cdots f(\boldsymbol{y}_n; \boldsymbol{\theta}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n-1})$$

where $f(\boldsymbol{y}_k; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1})$ *denotes the conditional density function given observations* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}$.

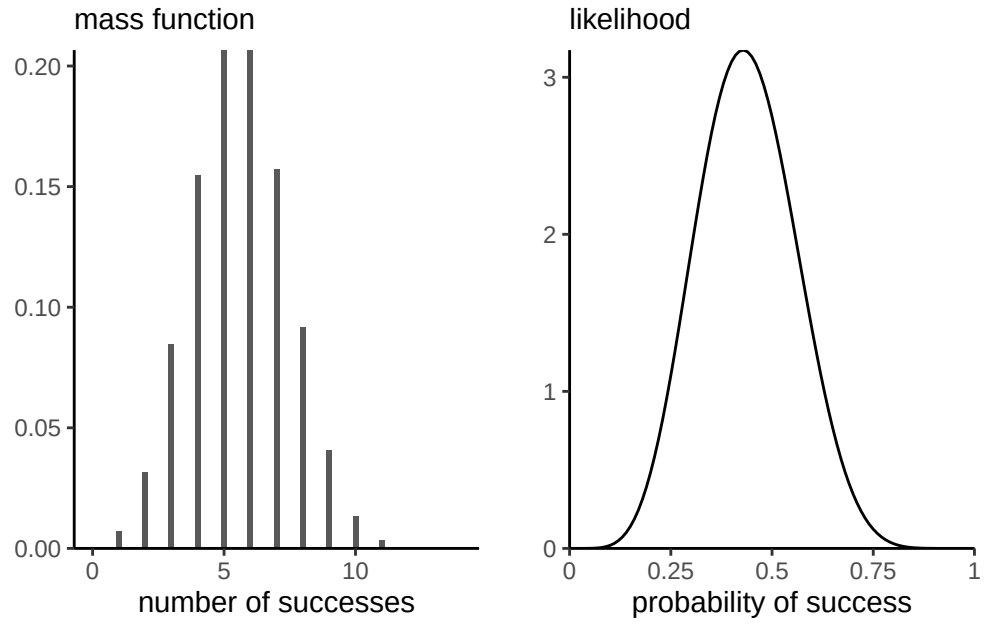

Figure 1.2: Binomial mass function (left) and scaled likelihood function (right).

By Bayes' rule, we can consider updating *the posterior by adding terms to the likelihood, noting that*

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) \propto p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)$$

which amounts to treating the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)$ *as a prior. If data are exchangeable, the order in which observations are collected and the order of the belief updating is irrelevant to the full posterior. Figure 1.3 shows how the posterior becomes gradually closer to the scaled likelihood as we increase the sample size, and the posterior mode moves towards the true value of the parameter (here 0.3).*

Example 1.2. While we can calculate analytically the value of the normalizing constant for the beta-binomial model, we could also for arbitrary priors use numerical integration or Monte Carlo methods in the event the parameter vector $\boldsymbol{\theta}$ is low-dimensional.

While estimation of the normalizing constant is possible in simple models, the following highlights some challenges that are worth keeping in mind. In a model for discrete data (that is, assigning probability mass to a countable set of outcomes), the terms in the likelihood are probabilities and thus the likelihood becomes smaller as we gather more observations

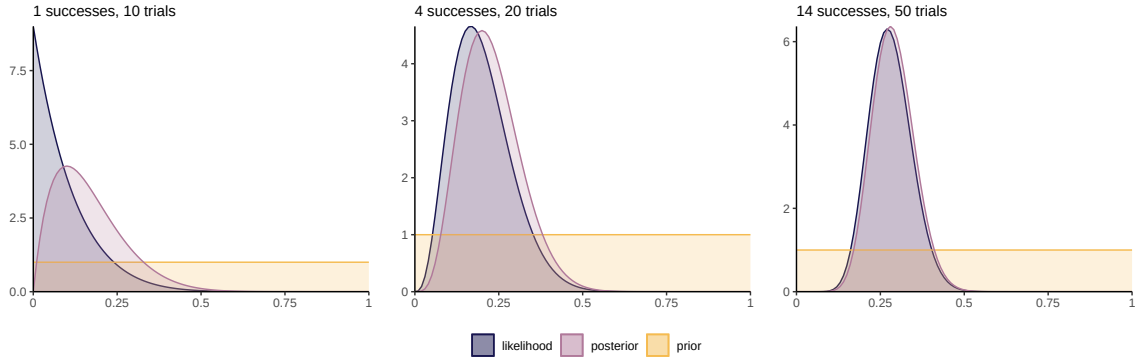

Figure 1.3: Beta posterior and binomial likelihood with a uniform prior for increasing number of observations (from left to right) out of a total of 100 trials.

(since we multiply terms between zero or one). The marginal likelihood term becomes smaller and smaller, so it's reciprocal is big and this can lead to arithmetic underflow.

```
y <- 6L # number of successes
n <- 14L # number of trials
alpha <- beta <- 1.5 # prior parameters
unnormalized_posterior <- function(theta){
  theta^(y+alpha-1) * (1-theta)^(n-y + beta - 1)
}
integrate(f = unnormalized_posterior,
          lower = 0,
          upper = 1)
```

```
1.066906e-05 with absolute error < 1e-12
```

```
# Compare with known constant
beta(y + alpha, n - y + beta)
```

```
[1] 1.066906e-05
```

```
# Monte Carlo integration
mean(unnormalized_posterior(runif(1e5)))
```

```
[1] 1.064067e-05
```

When $\boldsymbol{\theta}$ is high-dimensional, the marginal likelihood is intractable. This is one of the main challenges of Bayesian statistics and the popularity and applicability has grown drastically with the development and popularity of numerical algorithms, following th publication of Geman and Geman (1984) and Gelfand and Smith (1990). Markov chain Monte Carlo methods circumvent the calculation of the denominator by drawing approximate samples from the posterior.

1.3 Posterior predictive distribution

Prediction in the Bayesian paradigm is obtained by considering the *posterior predictive distribution*,

$$p(y_{\text{new}} \mid \boldsymbol{y}) = \int_{\Theta} p(y_{\text{new}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

Given draws from the posterior distribution, say $\boldsymbol{\theta}_b$ $(b = 1, \ldots, B)$, we sample from each a new realization from the distribution appearing in the likelihood $p(y_{\text{new}} \mid \boldsymbol{\theta}_b)$. This is different from the frequentist setting, which fixes the value of the parameter to some estimate $\widehat{\boldsymbol{\theta}}$; by contrast, the posterior predictive, here a beta-binomial distribution $\mathsf{BetaBin}(n, \alpha + y, n - y + \beta)$, carries over the uncertainty so will typically be wider and overdispersed relative to the corresponding binomial model. This can be easily seen from the left-panel of Figure 1.4, which contrasts the binomial mass function evaluated at the maximum likelihood estimator $\widehat{\theta} = 6/14$ with the posterior predictive.

```
npost <- 1e4L
# Sample draws from the posterior distribution
post_samp <- rbeta(n = npost, y + alpha, n - y + beta)
# For each draw, sample new observation
post_pred <- rbinom(n = npost, size = n, prob = post_samp)
```

Example 1.3 (Posterior predictive distribution of univariate Gaussian with known mean)**.** Consider an n sample of independent and identically distributed Gaussian, $Y_i \sim \mathsf{No}(0, \tau^{-1})$ $(i = 1, \ldots, n)$, where we assign a gamma prior on the precision $\tau \sim \mathsf{Ga}(\alpha, \beta)$. The posterior is

$$p(\tau \mid \boldsymbol{y}) \overset{\tau}{\propto} \prod_{i=1}^{n} \tau^{n/2} \exp\left(-\tau \frac{\sum_{i=1}^{n} y_i^2}{2}\right) \times \tau^{\alpha-1} \exp(-\beta\tau)$$

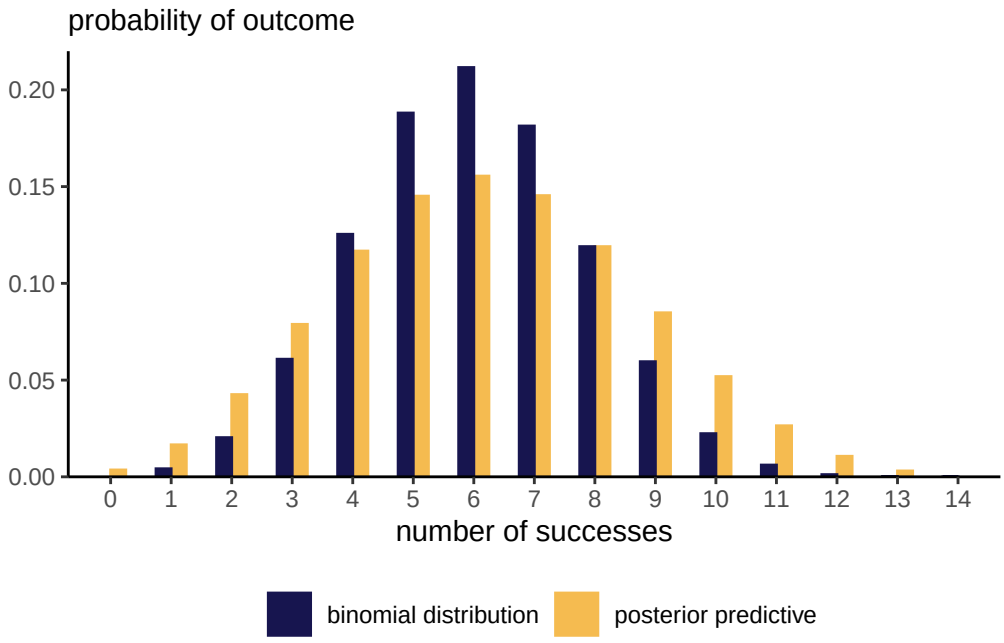

Figure 1.4: Beta-binomial posterior predictive distribution with corresponding binomial mass function evaluated at the maximum likelihood estimator.

and rearranging the terms to collect powers of τ, etc. we find that the posterior for τ must also be gamma, with shape parameter $\alpha^* = \alpha + n/2$ and rate $\beta^* = \beta + \sum_{i=1}^{n} y_i^2/2$.

The posterior predictive is

$$
\begin{aligned}
p(y_{\text{new}} \mid \boldsymbol{y}) &= \int_0^\infty \frac{\tau^{1/2}}{(2\pi)^{1/2}} \exp(-\tau y_{\text{new}}^2/2) \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \tau^{\alpha^*-1} \exp(-\beta^*\tau) \mathrm{d}\tau \\
&= (2\pi)^{-1/2} \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \int_0^\infty \tau^{\alpha^*-1/2} \exp\left\{-\tau(y_{\text{new}}^2/2 + \beta^*)\right\} \mathrm{d}\tau \\
&= (2\pi)^{-1/2} \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \frac{\Gamma(\alpha^* + 1/2)}{(y_{\text{new}}^2/2 + \beta^*)^{\alpha^*+1/2}} \\
&= \frac{\Gamma\left(\frac{2\alpha^*+1}{2}\right)}{\sqrt{2\pi}\Gamma\left(\frac{2\alpha^*}{2}\right) \beta^{*1/2}} \left(1 + \frac{y_{\text{new}}^2}{2\beta^*}\right)^{-\alpha^*-1/2} \\
&= \frac{\Gamma\left(\frac{2\alpha^*+1}{2}\right)}{\sqrt{\pi}\sqrt{2\alpha^*}\Gamma\left(\frac{2\alpha^*}{2}\right) (\beta^*/\alpha^*)^{1/2}} \left(1 + \frac{1}{2\alpha^*} \frac{y_{\text{new}}^2}{(\beta^*/\alpha^*)}\right)^{-\alpha^*-1/2}
\end{aligned}
$$

which entails that Y_{new} is a scaled Student-t distribution with scale $(\beta^*/\alpha^*)^{1/2}$ and $2\alpha + n$ degrees of freedom. This example also exemplifies the additional variability relative to the distribution generating the data: indeed, the Student-t distribution is more heavy-tailed than the Gaussian, but since the degrees of freedom increase linearly with n, the distribution converges to a Gaussian as $n \to \infty$, reflecting the added information as we collect more and more data points and the variance gets better estimated through $\sum_{i=1}^n y_i^2/n$.

1.4 Summarizing posterior distributions

Most of the field of Bayesian statistics revolves around the creation of algorithms that either circumvent the calculation of the normalizing constant (notably using Monte Carlo and Markov chain Monte Carlo methods) or else provide accurate numerical approximation of the posterior pointwise, including for marginalizing out all but one parameters (integrated nested Laplace approximations, variational inference, etc.) The target of inference is the whole posterior distribution, a potentially high-dimensional object which may be difficult to summarize or visualize. We can thus report only characteristics of the the latter.

The choice of point summary to keep has it's root in decision theory.

Definition 1.1 (Loss function)**.** A loss function $c(\boldsymbol{\theta}, \boldsymbol{\upsilon})$ is a mapping from $\boldsymbol{\Theta} \to \mathbb{R}^k$ that assigns a weight to each value of $\boldsymbol{\theta}$, corresponding to the regret or loss arising from choosing this value. The corresponding point estimator $\widehat{\boldsymbol{\upsilon}}$ is the minimizer of the expected loss,

$$\widehat{\boldsymbol{\upsilon}} = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \int_{\boldsymbol{\Theta}} c(\boldsymbol{\theta}, \boldsymbol{\upsilon}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

For example, in a univariate setting, the quadratic loss $c(\theta, \upsilon) = (\theta - \upsilon)^2$ returns the posterior mean, the absolute loss $c(\theta, \upsilon) = |\theta - \upsilon|$ returns the posterior median and the 0-1 loss $c(\theta, \upsilon) = \mathrm{I}(\upsilon \neq \theta)$ returns the posterior mode. All of these point estimators are central tendency measures, but some may be more adequate depending on the setting as they can correspond to potentially different values, as shown in the left-panel of Figure 1.5. The choice is application specific: for multimodal distributions, the mode is likely a better choice.

If we know how to evaluate the distribution numerically, we can optimize to find the mode or else return the value for the pointwise evaluation on a grid at which the density achieves it's maximum. The mean and median would have to be evaluated by numerical integration if there is no closed-form expression for the latter.

If we have rather a sample from the posterior with associated posterior density values, then we can obtain the mode as the parameter combination with the highest posterior, the

median from the value at rank $\lfloor n/2 \rfloor$ and the mean through the sample mean of posterior draws.

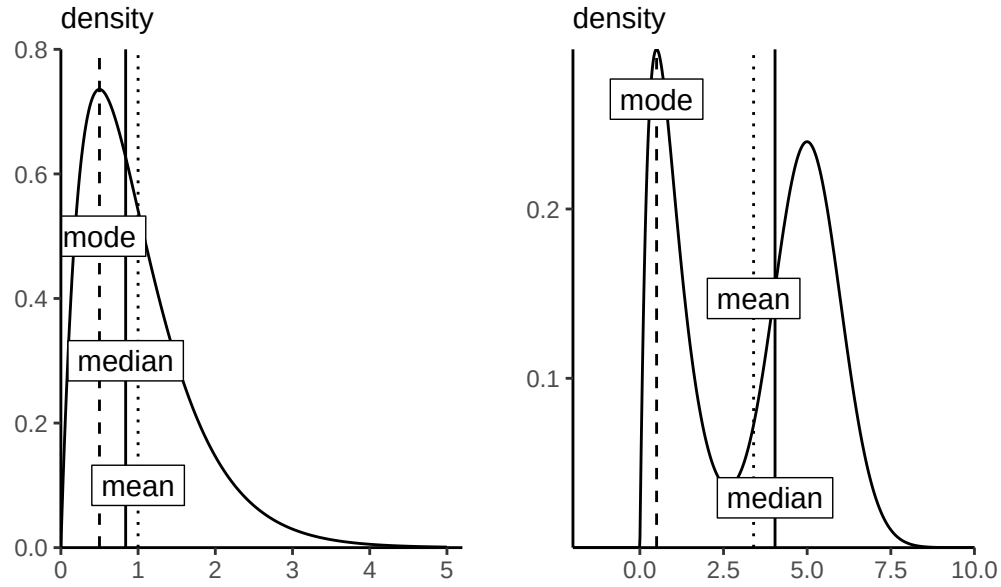

Figure 1.5: Point estimators from a right-skewed distribution (left) and from a multimodal distribution (right).

The loss function is often a functional (meaning a one-dimensional summary) from the posterior. The following example shows how it reduces a three-dimensional problem into a single risk measure.

Example 1.4 (Danish insurance losses)**.** In extreme value, we are often interested in assessing the risk of events that are rare enough that they lie beyond the range of observed data. To provide a scientific extrapolation, it often is justified to fit a generalized Pareto distribution to exceedances of $Z = Y - u$, for some user-specified threshold u which is often taken as a large quantile of the distribution of Y. The generalized Pareto distribution function is

$$F(z;\tau,\xi) = 1 - \begin{cases} (1+\xi/\tau z)_+^{-1/\xi}, & \xi \neq 0 \\ \exp(-z/\tau), & \xi = 0. \end{cases}$$

The shape ξ governs how heavy-tailed the distribution is, while τ is a scale parameter.

Insurance companies provide coverage in exchange for premiums, but need to safeguard themselves against very high claims by buying reinsurance products. These risks are often communicated through the value-at-risk (VaR), a high quantile exceeded with probability p. We model Danish fire insurance claim amounts for inflation-adjusted data collected from January 1980 until December 1990 that are in excess of a million Danish kroner, found in the `evir` package and analyzed in Example 7.23 of McNeil, Frey, and Embrechts (2005). These claims are denoted Y and there are 2167 observations.

We fit a generalized Pareto distribution to exceedances above 10 millions krones, keeping 109 observations or roughly the largest 5% of the original sample. Preliminary analysis shows that we can treat data as roughly independent and identically distributed and goodness-of-fit diagnostics (not shown) for the generalized Pareto suggest that the fit is adequate for all but the three largest observations, which are (somewhat severely) underestimated by the model.

The generalized Pareto model only describes the n_u exceedances above $u = 10$, so we need to incorporate in the likelihood a binomial contribution for the probability ζ_u of exceeding the threshold u. Provided that the priors for (τ, ξ) are independent of those for ζ_u, the posterior also factorizes as a product, so ζ_u and (τ, ξ) are a posteriori independent.

Suppose for now that we set a $\mathsf{Be}(0.5, 0.5)$ prior for ζ_u and a non-informative prior for the generalized Pareto parameters. The `post_samp` matrix contains exact samples from the posterior distribution of (τ, ξ, ζ_u), obtained using a Monte Carlo algorithm. Our aim is to evaluate the posterior distribution for the value-at-risk, the α quantile of Y for high values of α and see what point estimator one would obtain depending on our choice of loss function. For any $\alpha > 1 - \zeta_u$, the q_α is

$$\begin{aligned} 1 - \alpha &= \Pr(Y > q_\alpha \mid Y > u) \Pr(Y > u) \\ &= \left(1 + \xi \frac{q_\alpha - u}{\tau}\right)_+^{-1/\xi} \zeta_u \end{aligned}$$

and solving for q_α gives

$$q_\alpha = u + \frac{\tau}{\xi}\left\{\left(\frac{\zeta_u}{1-\alpha}\right)^{\xi} - 1\right\}.$$

To obtain the posterior distribution of the α quantile, q_α, it suffices to plug in each posterior sample and evaluate the function: the uncertainty is carried over from the simulated values of the parameters to those of the quantile q_α. The left panel of Figure 1.6 shows the posterior density estimate of the $\mathsf{VaR}(0.99)$ along with the maximum a posteriori (mode) of the latter.

Suppose that we prefer to under-estimate the value-at-risk rather than overestimate: this could be captured by the custom loss function

$$c(q, q_0) = \begin{cases} 0.5(0.99q - q_0), & q > q_0 \\ 0.75(q_0 - 1.01q), & q < q_0. \end{cases}$$

For a given value of the value-at-risk q_0 evaluated on a grid, we thus compute

$$r(q_0) = \int_{\boldsymbol{\Theta}} c(q(\boldsymbol{\theta}), q_0) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}$$

and we seek to minimize the risk, $\widehat{q} = \mathrm{argmin}_{q_0 \in \mathbb{R}_+} r(q_0)$. The value returned that minimizes the loss, shown in Figure 1.6, is to the left of the posterior mean for q_α.

```
# Compute value at risk from generalized Pareto distribution quantile fn
VaR_post <- with(post_samp,   # data frame of posterior draws
            revdbayes::qgp(   # with columns 'probexc', 'scale', 'shape'
  p = 0.01/probexc,
  loc = 10,
  scale = scale,
  shape = shape,
  lower.tail = FALSE))
# Loss function
loss <- function(qhat, q){
    mean(ifelse(q > qhat,
           0.5*(0.99*q-qhat),
           0.75*(qhat-1.01*q)))
}
# Create a grid of values over which to estimate the loss for VaR
nvals <- 101L
VaR_grid <- seq(
  from = quantile(VaR_post, 0.01),
  to = quantile(VaR_post, 0.99),
  length.out = nvals)
# Create a container to store results
risk <- numeric(length = nvals)
for(i in seq_len(nvals)){
  # Compute integral (Monte Carlo average over draws)
 risk[i] <- loss(q = VaR_post, qhat = VaR_grid[i])
}
```

The output of the Bayesian learning problem will be either of:

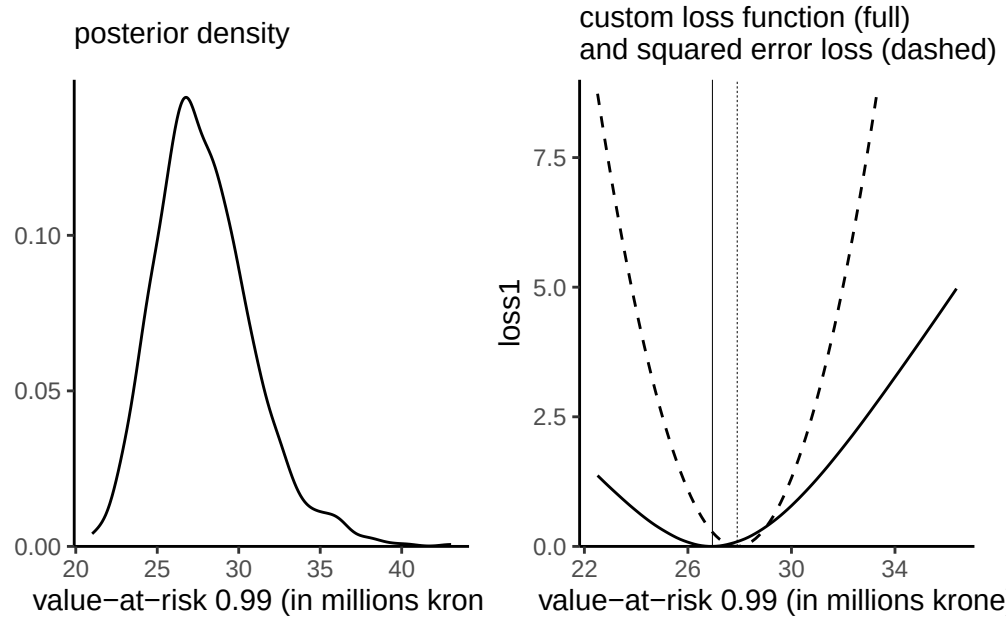

Figure 1.6: Posterior density (left) and losses functions for the 0.99 value-at-risk for the Danish fire insurance data. The vertical lines denote point estimates of the quantiles that minimize the loss functions.

1. a fully characterized distribution
2. a numerical approximation to the posterior distribution (pointwise)
3. an exact or approximate sample drawn from the posterior distribution

In the first case, we will be able to directly evaluate quantities of interest if there are closed-form expressions for the latter, or else we could draw samples from the distribution and evaluate them via Monte-Carlo. In case of numerical approximations, we will need to resort to numerical integration or otherwise to get our answers.

Often, we will also be interested in the marginal posterior distribution of each component θ_j in turn ($j = 1, \ldots, J$). To get these, we carry out additional integration steps,

$$p(\theta_j \mid \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}_{-j}.$$

With a posterior sample, this is trivial: it suffices to keep the column corresponding to θ_j and discard the others.

To communicate uncertainty, we may resort to credible regions and intervals.

Definition 1.2. A $(1-\alpha)$ **credible region** (or credible interval in the univariate setting) is a set $\mathcal{S}_\alpha$ such that, with probability level α,

$$\Pr(\boldsymbol{\theta} \in \mathcal{S}_\alpha \mid \boldsymbol{Y} = \boldsymbol{y}) = 1 - \alpha$$

These intervals are not unique, as are confidence sets. In the univariate setting, the central or equi-tailed interval are the most popular, and easily obtained by considering the $\alpha/2, 1-\alpha/2$ quantiles. These are easily obtained from samples by simply taking empirical quantiles. An alternative, highest posterior density credible sets, which may be a set of disjoint intervals obtained by considering the parts of the posterior with the highest density, may be more informative. The top panel Figure 1.7 shows the distinction for a bimodal mixture distribution, and a even more striking difference for 50% credible intervals for a symmetric beta distribution whose mass lie near the endpoints of the distribution, leading to no overlap between the two intervals.

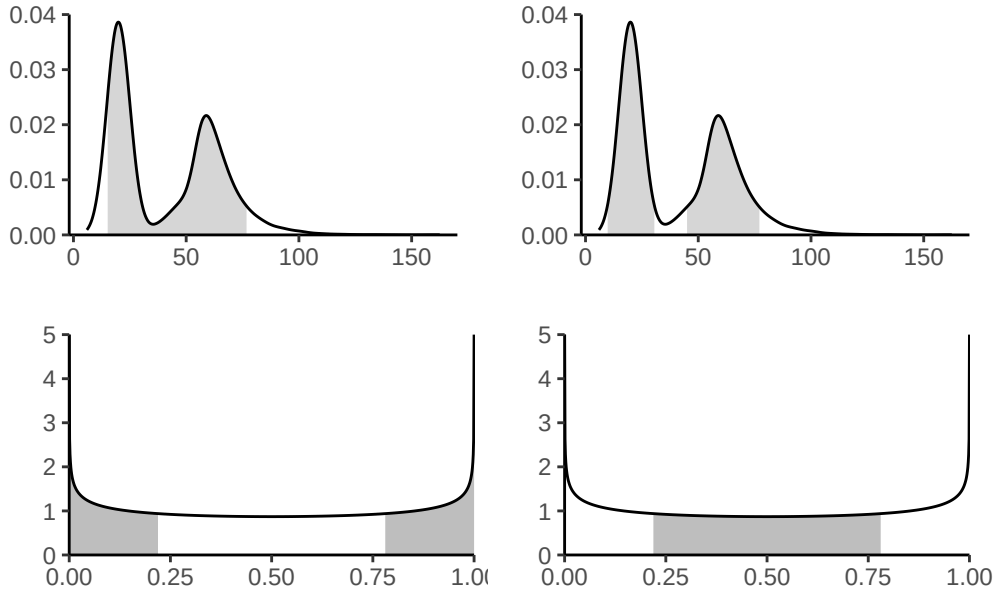

Figure 1.7: Density plots with 89% (top) and 50% (bottom) equitailed or central credible (left) and highest posterior density (right) intervals for two data sets, highlighted in grey.

2 Priors

The posterior distribution combines two ingredients: the likelihood and the prior. If the former is a standard ingredient of any likelihood-based inference, prior specification requires some care. The purpose of this chapter is to consider different standard way of constructing prior functions, and to specify the parameters of the latter: we term these hyperparameters.

2.1 Conjugate priors

In very simple models, there may exists prior densities that result in a posterior distribution of the same family. We can thus directly extract characteristics of the posterior. Conjugate priors are chosen for computational convenience and because interpretation is convenient, as the parameters of the posterior will often be some weighted average of prior and likelihood component.

Definition 2.1. A prior density $p(\boldsymbol{\theta})$ is conjugate for likelihood $L(\boldsymbol{\theta}; \boldsymbol{y})$ if the product $L(\boldsymbol{\theta}; \boldsymbol{y})p(\boldsymbol{\theta})$, after renormalization, is of the same parametric family as the prior.

Exponential families (including the binomial, Poisson, exponential, Gaussian distributions) admit conjugate priors[1]

[1]A distribution belongs to an exponential family with parameter vector $\boldsymbol{\theta} \in \mathbb{R}^D$ if it can be written as

$$f(y; \boldsymbol{\theta}) = \exp\left\{\sum_{k=1}^{K} Q_k(\boldsymbol{\theta}) t_k(y) + D(\boldsymbol{\theta})\right\}$$

and in particular, the support does not depend on unknown parameters. If we have an independent and identically distributed sample of observations $y_1, \ldots, y_n$, the log likelihood is thus of the form

$$\ell(\boldsymbol{\theta}) = \sum_{k=1}^{K} \phi_k(\boldsymbol{\theta}) \sum_{i=1}^{n} t_k(y_i) + nD(\boldsymbol{\theta}),$$

where the collection $\sum_{i=1}^{n} t_k(y_i)$ $(k = 1, \ldots, K)$ are sufficient statistics and $\phi_k(\boldsymbol{\theta})$ are the canonical parameters. The number of sufficient statistics are the same regardless of the sample size. Exponential families play a prominent role in generalized linear models, in which the natural parameters are modeled as linear

Example 2.1 (Conjugate prior for the binomial model)**.** The binomial log density with y successes out of n trials is proportional to

$$y \log(p) + (n-y) \log(1-p) = y \log\left(\frac{p}{1-p}\right) + n \log(1-p)$$

with canonical parameter $\mathrm{logit}(p)$.[2] The binomial distribution is thus an exponential family.

Since the density of the binomial is of the form $p^y(1-p)^{n-y}$, the beta distribution $\mathsf{Be}(\alpha, \beta)$ with density

$$f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$$

is the conjugate prior.

The beta distribution is also the conjugate prior for the negative binomial, geometric and Bernoulli distributions, since their likelihoods are all proportional to that of the beta. The fact that different sampling schemes that result in proportional likelihood functions give the same inference is called likelihood principle.

Example 2.2 (Conjugate prior for the Poisson model)**.** The Poisson distribution with mean μ has log density proportional to $f(y;\mu) \propto y \log(\mu) - \mu$, so is an exponential family with natural parameter $\log(\mu)$. The gamma density,

$$f(x) \propto \beta^\alpha / \Gamma(\alpha) x^{\alpha-1} \exp(-\beta x)$$

with shape α and rate β is the conjugate prior for the Poisson. For an n-sample of independent observations $\mathsf{Po}(\mu)$ observations with $\mu \sim \mathsf{Ga}(\alpha, \beta)$, the posterior is $\mathsf{Ga}(\sum_{i=1}^n y_i + \alpha, \beta + n)$.

Knowing the analytic expression for the posterior can be useful for calculations of the marginal likelihood, as Example 2.3 demonstrates.

Example 2.3.

function of explanatory variables. A log prior density with parameters $\eta, \nu_1, \ldots, \nu_K$ that is proportional to

$$\log p(\boldsymbol{\theta}) \propto \eta D(\boldsymbol{\theta}) + \sum_{k=1}^{K} Q_k(\boldsymbol{\theta}) \nu_k$$

is conjugate.

[2] The canonical link function for Bernoulli gives rise to logistic regression model.

Negative binomial as a Poisson mixture

One restriction of the Poisson model is that the restriction on its moments is often unrealistic. The most frequent problem encountered is that of **overdispersion**, meaning that the variability in the counts is larger than that implied by a Poisson distribution.

One common framework for handling overdispersion is to have $Y \mid \Lambda = \lambda \sim \mathsf{Po}(\lambda)$, where the mean of the Poisson distribution is itself a positive random variable with mean μ, if Λ follows a conjugate gamma distribution with shape $k\mu$ and rate $k > 0$, $\Lambda \sim \mathsf{Ga}(k\mu, k)$, the posterior $\Lambda \mid Y = y \sim \mathsf{Ga}(k\mu + y, k + 1)$.

Since the joint density of Y and Λ can be written

$$p(y, \lambda) = p(y \mid \lambda)p(\lambda) = p(\lambda \mid y)p(y)$$

we can isolate the marginal density

$$\begin{aligned}
p(y) &= \frac{p(y \mid \lambda)p(\lambda)}{p(\lambda \mid y)} \\
&= \frac{\frac{\lambda^y \exp(-\lambda)}{\Gamma(y+1)} \frac{k^{k\mu}\lambda^{k\mu-1}\exp(-k\lambda)}{\Gamma(k\mu)}}{\frac{(k+1)^{k\mu+y}\lambda^{k\mu+y-1}\exp\{-(k+1)\lambda\}}{\Gamma(k\mu+y)}} \\
&= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} k^{k\mu}(k+1)^{-k\mu-y} \\
&= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} \left(1 - \frac{1}{k+1}\right)^{k\mu} \left(\frac{1}{k+1}\right)^{y}
\end{aligned}$$

and this is the density of a negative binomial distribution with probability of success $1/(k+1)$. We can thus view the negative binomial as a Poisson mean mixture.

By the laws of iterated expectation and iterative variance,

$$\begin{aligned}
\mathsf{E}(Y) &= \mathsf{E}_\Lambda\{\mathsf{E}(Y \mid \Lambda\} \\
&= \mathsf{E}(\Lambda) = \mu \\
\mathsf{Va}(Y) &= \mathsf{E}_\Lambda\{\mathsf{Va}(Y \mid \Lambda)\} + \mathsf{Va}_\Lambda\{\mathsf{E}(Y \mid \Lambda)\} \\
&= \mathsf{E}(\Lambda) + \mathsf{Va}(\Lambda) \\
&= \mu + \mu/k.
\end{aligned}$$

Negative binomial as a Poisson mixture

The marginal distribution of Y, unconditionally, has a variance which exceeds its mean, as

$$\mathsf{E}(Y) = \mu, \qquad \mathsf{Va}(Y) = \mu(1 + 1/k).$$

In a negative binomial regression model, the term k is a dispersion parameter, which is fixed for all observations, whereas $\mu = \exp(\boldsymbol{\beta}\mathbf{X})$ is a function of covariates $\mathbf{X}$. As $k \to \infty$, the distribution of Λ degenerates to a constant at μ and we recover the Poisson model.

Example 2.4 (Posterior rates for A/B tests using conjugate Poisson model)**.** Upworthy.com, a US media publisher, revolutionized headlines online advertisement by running systematic A/B tests to compare the different wording of headlines, placement and image and what catches attention the most. The Upworthy Research Archive (Matias et al. 2021) contains results for 22743 experiments, with a click through rate of 1.58% on average and a standard deviation of 1.23%. The `clickability_test_id` gives the unique identifier of the experiment, `clicks` the number of conversion out of `impressions`. See Section 8.5 of Alexander (2023) for more details about A/B testing and background information.

Consider an A/B test from November 23st, 2014, that compared four different headlines for a story on Sesame Street workshop with interviews of children whose parents were in jail and visiting them in prisons. The headlines tested were:

1. Some Don't Like It When He Sees His Mom. But To Him? Pure Joy. Why Keep Her From Him?
2. They're Not In Danger. They're Right. See True Compassion From The Children Of The Incarcerated.
3. Kids Have No Place In Jail ... But In This Case, They *Totally* Deserve It.
4. Going To Jail *Should* Be The Worst Part Of Their Life. It's So Not. Not At All.

At first glance, the first and third headlines seem likely to lead to a curiosity gap. The wording of the second is more explicit (and searchable), whereas the first is worded as a question.

We model the conversion rate λ_i for each headline separately using a Poisson distribution and compare the posterior distributions for all four choices. Using a conjugate prior and selecting the parameters by moment matching yields approximately $\alpha = 1.64$ and $\beta = 0.01$ for the hyperparameters.

Table 2.1: Number of views, clicks for different headlines for the Upworthy data.

headline	impressions	clicks
H1	3060	49
H2	2982	20

headline	impressions	clicks
H3	3112	31
H4	3083	9

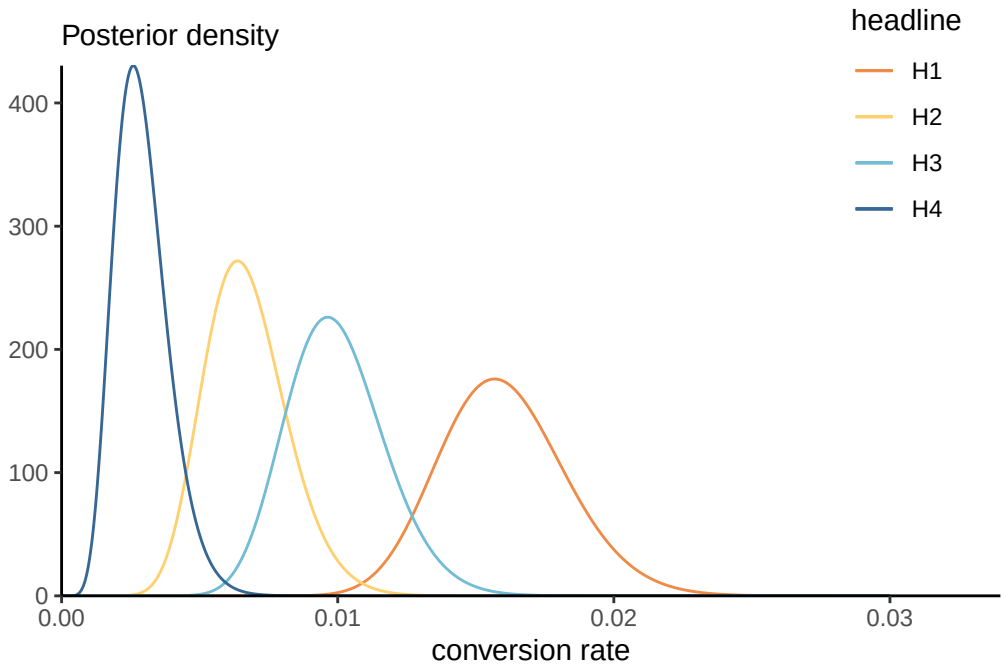

Figure 2.1: Gamma posterior for the Upworthy Sesame street headline.

We can visualize the posterior distributions. In this context, the large sample size lead to the dominance of the likelihood contribution $p(Y_i \mid \lambda_i) \sim \mathsf{Po}(n_i\lambda_i)$ relative to the prior. We can see there is virtually no overlap between different rates for headers H1 (preferred) relative to H4 (least favorable). The probability that the conversion rate for Headline 3 is higher than Headline 1 can be approximated by simulating samples from both posteriors and computing the proportion of times one is larger: we get 1.7% for `H3` relative to `H1`, indicating a clear preference for the first headline `H1`.

Example 2.5 (Should you phrase your headline as a question?)**.** We can also consider aggregate records for Upworthy, as Alexander (2023) did. The `upworthy_question` database contains a balanced sample of all headlines where at least one of the choices featured a question, with at least one alternative statement. Whether a headline contains a question or not is determined by querying for the question mark. We consider aggregated counts for all such headlines, with the `question` factor encoding whether there was a question, `yes` or

`no`. For simplicity, we treat the number of views as fixed, but keep in mind that A/B tests are often sequential experiments with a stopping rule.[3]

We model first the rates using a Poisson regression; the corresponding frequentist analysis would include an offset to account for differences in views. If λ_j $(j = 1, 2)$ are the average rate for each factor level (yes and no), then $\mathsf{E}(Y_{ij}/n_{ij}) = \lambda_j$. In the frequentist setting, we can fit a simple Poisson generalized linear regression model with an offset term and a binary variable.

```
data(upworthy_question, package = "hecbayes")
poismod <- glm(
  clicks ~ offset(log(impressions)) + question,
  family = poisson(link = "log"),
  data = upworthy_question)
coef(poismod)
```

```
(Intercept)  questionno
-4.51264669  0.07069677
```

The coefficients represent the difference in log rate (multiplicative effect) relative to the baseline rate, with an increase of 6.3 percent when the headline does not contain a question. A likelihood ratio test can be performed by comparing the deviance of the null model (intercept-only), indicating strong evidence that including question leads to significatively different rates. This is rather unsurprising given the enormous sample sizes.

Consider instead a Bayesian analysis with conjugate prior: we model separately the rates of each group (question or not). Suppose we think apriori that the click-rate is on average 1%, with a standard deviation of 2%, with no difference between questions or not. This would translate, using moment matching, into a gamma prior distribution $p(\lambda_j)$ with rate $\beta = 0.04 = \mathsf{Var}_0/\mathsf{E}_0$ and shape $\alpha = 2.5$ $(j = 1, 2)$. If λ_j is the average rate for each factor level (yes and no), then $\mathsf{E}(Y_{ij}/n_{ij}) = \lambda_j$ so the log likelihood is proportional, as a function of λ_1 and λ_2, to

$$\ell(\boldsymbol{\lambda}; \boldsymbol{y}, \boldsymbol{n}) \overset{\boldsymbol{\lambda}}{\propto} \sum_{i=1}^{n}\sum_{j=1}^{2} y_{ij} \log \lambda_j - \lambda_j n_{ij}$$

and we can recognize that the posterior for λ_i is gamma with shape $\alpha + \sum_{i=1}^{n} y_{ij}$ and rate $\beta + \sum_{i=1}^{n} n_{ij}$. For inference, we thus only need to select hyperparameters and calculate

[3]The stopping rule means that data stops being collected once there is enough evidence to determine if an option is more suitable, or if a predetermined number of views has been reached.

the total number of clicks and impressions per group. We can then consider the posterior difference $\lambda_1 - \lambda_2$ or, to mimic the Poisson multiplicative model, of the ratio λ_1/λ_2. The former suggests very small differences, but one must keep in mind that rates are also small. The ratio, shown in the right-hand panel of Figure 2.2, gives a more easily interpretable portrait that is in line with the frequentist analysis.

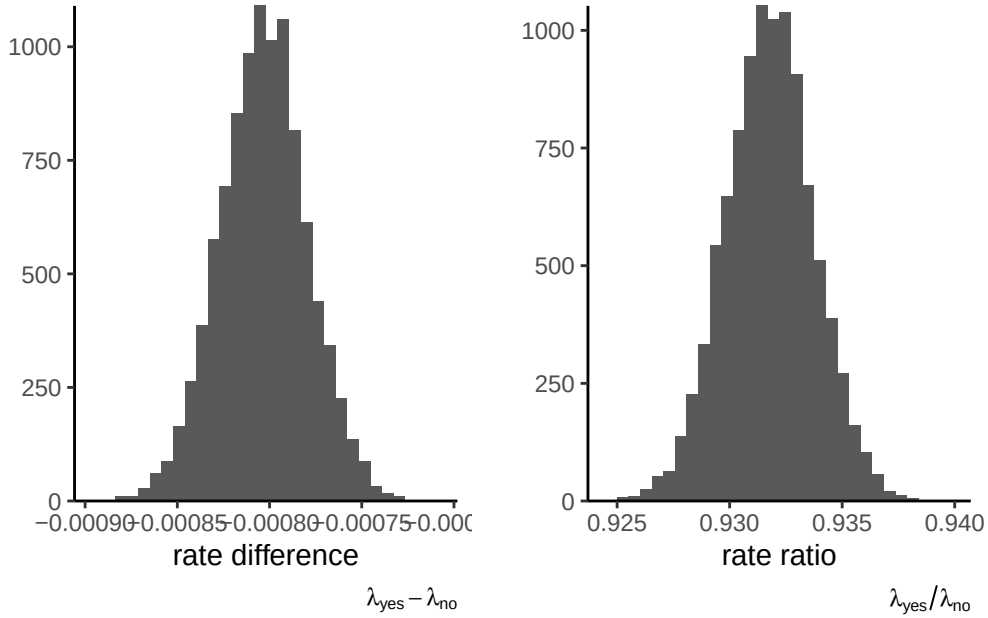

Figure 2.2: Histograms of posterior summaries for differences (left) and rates (right) based on 1000 simulations from the independent gamma posteriors.

To get an approximation to the posterior mean of the ratio λ_1/λ_2, it suffices to draw independent observations from their respective posterior, compute the ratio and take the sample mean of those draws. We can see that the sampling distribution of the ratio is nearly symmetrical, so we can expect Wald intervals to perform well should one be interested in building confidence intervals. This is however hardly surprising given the sample size at play.

Example 2.6 (Conjugate prior for Gaussian mean with known variance)**.** Consider an n simple random sample of independent and identically distributed Gaussian variables with mean μ and standard deviation σ, denoted $Y_i \sim \mathsf{No}(\mu, \sigma^2)$. We pick a Gaussian prior for the location parameter, $\mu \sim \mathsf{No}(\nu, \tau^2)$ where we assume μ, τ are fixed hyperparameter values. For now, we consider only inference for $p(\mu \mid \sigma)$: discarding any term that is not a function

of μ, the conditional posterior is

$$\begin{aligned} p(\mu \mid \sigma) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}\exp\left\{-\frac{1}{2\tau^2}(\mu-\nu)^2\right\} \\ &\propto p(\sigma)\sigma^{-1}\exp\left\{\left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}+\frac{\nu}{\tau^2}\right)\mu-\left(\frac{n}{2\sigma^2}+\frac{1}{2\tau^2}\right)\mu^2\right\}. \end{aligned}$$

The log of the posterior density conditional on σ is quadratic in μ, it must be a Gaussian distribution truncated over the positive half line. This can be seen by completing the square in μ, or by comparing this expression to the density of $\mathsf{No}(\mu,\sigma^2)$,

$$f(x;\mu,\sigma) \overset{\mu}{\propto} \exp\left(-\frac{1}{2\sigma^2}\mu^2+\frac{x}{\sigma^2}\mu\right)$$

we can deduce by matching mean and variance that the conditional posterior $p(\mu \mid \sigma)$ is Gaussian with reciprocal variance (precision) $n/\sigma^2+1/\tau^2$ and mean $(n\overline{y}\tau^2+\nu\sigma^2)/(n\tau^2+\sigma^2)$. The precision is an average of that of the prior and data, but assigns more weight to the latter, which increases linearly with the sample size n. Likewise, the posterior mean is a weighted average of prior and sample mean, with weights proportional to the relative precision.

The exponential family is quite large; Fink (1997) *A Compendium of Conjugate Priors* gives multiple examples of conjugate priors and work out parameter values.

In general, unless the sample size is small and we want to add expert opinion, we may wish to pick an *uninformative prior*, i.e., one that does not impact much the outcome. For conjugate models, one can often show that the relative weight of prior parameters (relative to the random sample likelihood contribution) becomes negligible by investigating their relative weights.

2.2 Uninformative priors

Definition 2.2 (Proper prior)**.** We call a prior function *proper* if it's integral is finite over the parameter space; such prior function automatically leads to a valid posterior.

The best example of prior priors arise from probability density function. We can still employ this rule for improper priors: for example, taking $\alpha,\beta \to 0$ in the beta prior leads to a prior proportional to $x^{-1}(1-x)^{-1}$, the integral of which diverges on the unit interval $[0,1]$. However, as long as the number of success and the number of failures is larger than 1, meaning $k \geq 1, n-k \geq 1$, the posterior distribution would be proper, i.e., integrable. To find the posterior, normalizing constants are also superfluous.

Many uninformative priors are flat, or proportional to a uniform on some subset of the real line and therefore improper. It may be superficially tempting to set a uniform prior on a large range to ensure posterior property, but the major problem is that a flat prior may be informative in a different parametrization, as the following example suggests.

Example 2.7 (Transformation of flat prior for scales)**.** Consider the parameter $\log(\tau) \in \mathbb{R}$ and the prior $p(\log \tau) \propto 1$. If we reparametrize the model in terms of τ, the new prior (including the Jacobian of the transformation) is τ^{-1}

Some priors are standard and widely used. In location scale families with location ν and scale τ, the density is such that

$$f(x; \nu, \tau) = \frac{1}{\tau} f\left(\frac{x-\nu}{\tau}\right), \qquad \nu \in \mathbb{R}, \tau > 0.$$

We thus wish to have a prior so that $p(\tau) = c^{-1}p(\tau/c)$ for any scaling $c > 0$, whence it follows that $p(\tau) \propto \tau^{-1}$, which is uniform on the log scale.

The priors $p(\nu) \propto 1$ and $p(\tau) \propto \tau^{-1}$ are both improper but lead to location and scale invariance, hence that the result is the same regardless of the units of measurement.

One criticism of the Bayesian approach is the arbitrariness of prior functions. However, the role of the prior is often negligible in large samples (consider for example the posterior of exponential families with conjugate priors). Moreover, the likelihood is also chosen for convenience, and arguably has a bigger influence on the conclusion. Data fitted using a linear regression model seldom follow Gaussian distributions conditionally, in the same way that the linearity is a convenience (and first order approximation).

Definition 2.3 (Jeffrey's prior)**.** In single parameter models, taking a prior function for θ proportional to the square root of the determinant of the information matrix, $p(\theta) \propto |\imath(\theta)|^{1/2}$ yields a prior that is invariant to reparametrization, so that inferences conducted in different parametrizations are equivalent.[4]

To see this, consider a bijective transformation $\theta \mapsto \vartheta$. Under the reparametrized model and suitable regularity conditions[5], the chain rule implies that

$$\begin{aligned} i(\vartheta) &= -\mathsf{E}\left(\frac{\partial^2 \ell(\vartheta)}{\partial^2 \vartheta}\right) \\ &= -\mathsf{E}\left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right)\left(\frac{\mathrm{d}\theta}{\mathrm{d}\vartheta}\right)^2 + \mathsf{E}\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)\frac{\mathrm{d}^2\theta}{\mathrm{d}\vartheta^2} \end{aligned}$$

[4]The Fisher information is linear in the sample size for independent and identically distributed data so we can derive the result for $n = 1$ without loss of generality.

[5]Using Bartlett's identity; Fisher consistency can be established using the dominated convergence theorem.

Since the score has mean zero, $\mathsf{E}\left\{\partial\ell(\theta)/\partial\theta\right\} = 0$ and the rightmost term vanishes. We can thus relate the Fisher information in both parametrizations, with

$$\imath^{1/2}(\vartheta) = \imath^{1/2}(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\vartheta}\right|,$$

implying invariance.

Most of the times, Jeffrey's prior is improper. For the binomial model, it can be viewed as a limiting conjugate beta prior with $\alpha, \beta \to 0$). Unfortunately, in multiparameter models, the system isn't invariant to reparametrization if we consider the determinant of the Fisher information.

Example 2.8 (Jeffrey's prior for the binomial distribution)**.** Consider the binomial distribution $f(y;\theta,n) \propto \theta^y(1-\theta)^{n-y}\mathsf{I}_{\theta\in[0,1]}$. The negative of the second derivative of the log likelihood with respect to p is

$$\jmath(\theta) = -\partial^2\ell(\theta;y)/\partial\theta^2 = y/\theta^2 + (1-y)/(1-\theta)^2$$

and since $\mathsf{E}(Y) = n\theta$, the Fisher information is

$$\imath(\vartheta) = \mathsf{E}\{\jmath(\theta)\} = n/\theta + n/(1-\theta) = n/\{\theta(1-\theta)\}$$

Jeffrey's prior is thus $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$.

2.3 Jeffrey's prior for the normal distribution

Check that for the Gaussian distribution $\mathsf{No}(\mu,\sigma^2)$, the Jeffrey's prior obtained by treating each parameter as fixed in turn, are $p(\mu) \propto 1$ and $p(\sigma) \propto 1/\sigma$, which also correspond to the default uninformative priors for location-scale families.

Example 2.9 (Jeffrey's prior for the Poisson distribution)**.** The Poisson distribution with $\ell(\lambda) \propto -\lambda + y\log\lambda$, with second derivative $-\partial^2\ell(\lambda)/\partial\lambda^2 = y/\lambda^2$. Since the mean of the Poisson distribution is λ, the Fisher information is $\imath(\lambda) = \lambda^{-1}$ and Jeffrey's prior is $\lambda^{-1/2}$.

2.4 Prior simulation

Oftentimes, expert elicitation is difficult and it is hard to grasp what the impacts of the hyperparameters are. One way to see if the priors are reasonable is to sample values from them and generate new observations, resulting in prior predictive draws.

The prior predictive is $\int_{\boldsymbol{\Theta}} p(y \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$: we can simulate outcomes from it by first drawing parameter values from the prior, then sampling new observations from the distribution in the likelihood and keeping only the latter.

Example 2.10. Consider the daily number of Bixi bike sharing users for 2017–2019 at the Edouard Montpetit station next to HEC: we can consider a simple linear regression with log counts as a function of temperature,[6]

$$\log(\texttt{nusers}) \sim \mathsf{No}_+\{\beta_0 + \beta_1(\texttt{temp} - 20), \sigma^2\}.$$

The β_1 slope measures units in degree Celsius per log number of person.

The hyperparameters depend of course on the units of the analysis, unless one standardizes response variable and explanatories: it is easier to standardize the temperature so that we consider deviations from, say 20°C, which is not far from the observed mean in the sample. After some tuning, the independent priors $\beta_0 \sim \mathsf{No}(\overline{y}, 0.5^2)$, $\beta_1 \sim \mathsf{No}(0, 0.05^2)$ and $\sigma \sim \mathsf{Exp}(3)$ seem to yield plausible outcomes and relationships.[7]

We can draw regression lines from the prior, as in the left panel of Figure 2.3: while some of the negative relationships appear unlikely after seeing the data, the curves all seem to pass somewhere in the cloud of point. By contrast, a silly prior is one that would result in all observations being above or below the regression line, or yield values that are much too large near the endpoints of the explanatory variable. Indeed, given the number of bikes for rental is limited (a docking station has only 20 bikes), it is also sensible to ensure that simulations do not return overly large numbers. The maximum number of daily users in the sample is 68, so priors that return simulations with more than 200 (rougly 5.3 on the log scale) are not that plausible. The prior predictive draws can help establish this and the right panel of Figure 2.3 shows that, expect for the lack of correlation between temperature and number of users, the simulated values from the prior predictive are plausible even if overdispersed.

2.5 Informative priors

One strength of the Bayesian approach is the capability of incorporating expert and domain-based knowledge through priors. Often, these will take the form of moment constraints, so one common way to derive a prior is to perform moment matching to related ellicited quantities with moments of the prior distribution. It may be easier to set priors on a different scale than those of the observations, as Example 2.11 demonstrates.

[6] If counts are Poisson, then the log transform is variance stabilizing.

[7] One can object to the prior parameters depending on the data, but an alternative would be to model centered data $y - \overline{y}$, in which case the prior for the intercept parameter β_0 would be zero.

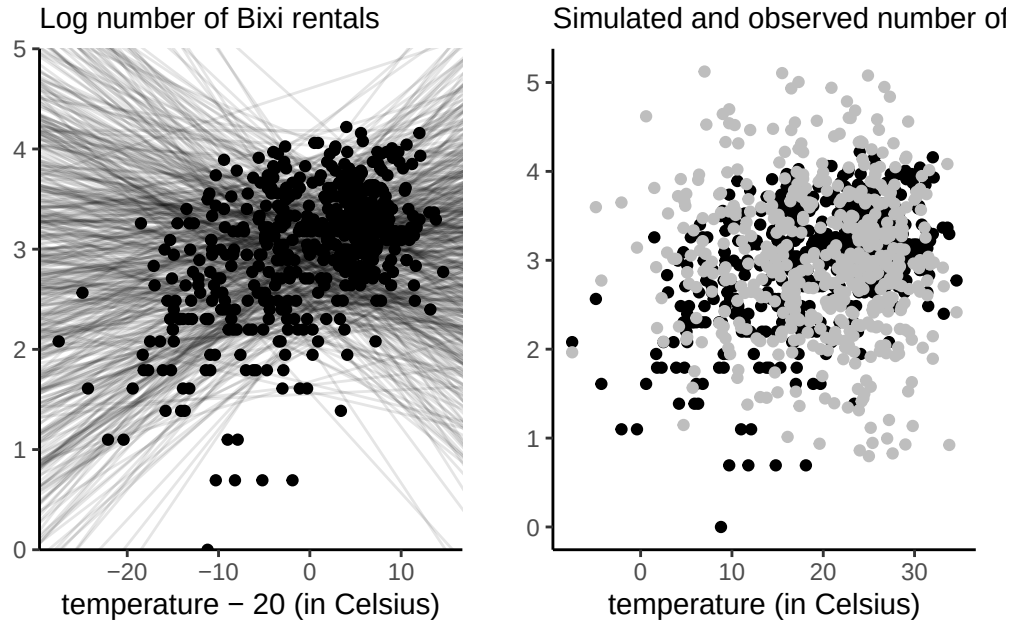

Figure 2.3: Prior draws of the linear regressions with observed data superimposed (left), and draws of observations from the prior predictive distribution (in gray) against observed data (right).

Example 2.11 (Gamma quantile difference priors for extreme value distributions)**.** The generalized extreme value distribution arises as the limiting distribution for the maximum of m independent observations from some common distribution F. The $\mathsf{GEV}(\mu, \sigma, \xi)$ distribution is a location-scale with distribution function

$$F(x) = \exp\left[-\{1+\xi(x-\mu)/\sigma\}_+^{-1/\xi}\right]$$

where $x_+ = \max\{0, x\}$.

Inverting the distribution function yields the quantile function

$$Q(p)\mu + \sigma\frac{(-\log p)^{-\xi}-1}{\xi}$$

In environmental data, we often model annual maximum. Engineering designs are often specified in terms of the k-year return levels, defined as the quantile of the annual maximum exceeded with probability $1/k$ in any given year. Using a GEV for annual maximum, Coles and Tawn (1996) proposed modelling annual daily rainfall and specifying a prior on the

quantile scale $q_1 < q_2 < q_3$ for tail probabilities $p_1 > p_2 > p_3$. To deal with the ordering constraints, gamma priors are imposed on the differences $q_1 - o \sim \mathsf{Ga}(\alpha_1, \beta_1)$, $q_2 - q_1 \sim \mathsf{Ga}(\alpha_2, \beta_2)$ and $q_3 - q_2 \sim \mathsf{Ga}(\alpha_3, \beta_3)$, where o is the lower bound of the support. The prior is thus of the form

$$p(\boldsymbol{q}) \propto q_1^{\alpha_1 - 1} \exp(-\beta_1 q_1) \prod_{i=2}^{3} (q_i - q_{i-1})^{\alpha_i - 1} \exp\{\beta_i (q_i - q_{i-1})\}.$$

where $0 \leq q_1 \leq q_2 \leq q_3$. The fact that these quantities refer to moments or risk estimates which practitioners often must compute as part of regulatory requirements makes it easier to specify sensible values for hyperparameters.

As illustrating example, consider maximum daily cumulated rainfall in Abisko, Sweden. The time series spans from 1913 until December 2014; we compute the 102 yearly maximum, which range from 11mm to 62mm, and fit a generalized extreme value distribution to these.

For the priors, suppose an expert elicits quantiles of the 10, 50 and 100 years return levels; say 30mm, 45mm and 70mm, respectively, for the median and likewise 40mm, 70mm and 120mm for the 90% percentile of the return levels. We can compute the differences and calculate the parameters of the gamma distribution through moment-matching: this gives roughly a shape of $\alpha_1 = 18.27$ and $\beta_1 = 0.6$, etc. Figure 2.4 shows the transfer from the prior predictive to the posterior distribution. The prior is much more dispersed and concentrated on the tail, which translates in a less peaked posterior than using a weakly informative prior (dotted line): the mode of the latter is slightly to the left and with lower density in the tail.

What would you do if we you had prior information from different sources? One way to combine these is through a mixture: given M different prior distributions $p_m(\boldsymbol{\theta})$, we can assign each a positive weight w_m to form a mixture of experts prior through the linear combination

$$p(\boldsymbol{\theta}) \propto \sum_{m=1}^{M} w_m p_m(\boldsymbol{\theta})$$

2.6 Priors for regression models

Gaussian components are widespread: not only for linear regression models, but more generally for the specification of random effects that capture group-specific effects, residuals spatial or temporal variability. In the Bayesian paradigm, there is no difference between fixed effects $\boldsymbol{\beta}$ and the random effect parameters: both are random quantities that get assigned priors.

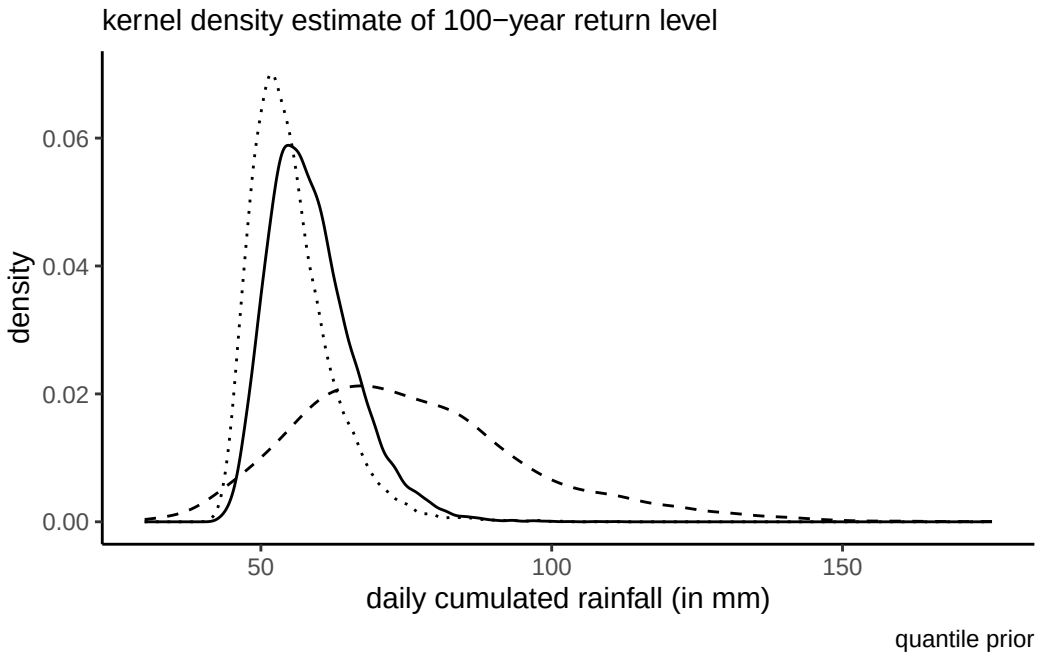

Figure 2.4: Kernel density estimates of draws from the posterior distribution of 100 year return levels with a Coles–Tawn quantile prior (full line) and from the corresponding prior predictive (dashed). The dotted line gives the posterior distribution for a maximum domain information prior on the shape with improper priors on location and scale.

It is generally good advice to center and scale explanatory variables and response vectors so they have approximately mean zero and unit variance, as this facilitates prior specification.

Andrew Gelman uses the following taxonomy for various levels of prior information: uninformative priors are generally flat or uniform priors with $p(\beta) \propto 1$, vague priors are typically nearly flat even if propor, e.g., $\beta \sim \mathsf{No}(0, 100)$, weakly informative priors provide little constraints $\beta \sim \mathsf{No}(0, 10)$), and informative prior are typically application-specific, but constrain the ranges. Uninformative and vague priors are not recommended.

If Gaussian priors are ubiquitous for the mean parameters β, priors for the scale are more contentious. Gelman (2006) recommends a Student-t distribution truncated below at 0, with low degrees of freedom.

The rationale for this choice comes from the simple two level model:

$$\begin{aligned} Y_{ij} &\sim \mathsf{No}(\mu + \alpha_j, \sigma^2), \qquad i = 1, \ldots, n_j; j = 1, \ldots, J \\ \alpha_j &\sim \mathsf{No}(0, \tau_\alpha^2), \qquad j = 1, \ldots, J \end{aligned}$$

Given α, μ, σ and the data $\boldsymbol{y}$, the conditionally conjugate prior is inverse gamma. Standard inference with this parametrization is complicated, because there is strong dependence between parameters.

To reduce this dependence, one can consider an overparametrization in which $\alpha_j = \xi\eta_j$ and $\eta_j \sim \mathsf{No}(0, \tau_\eta^2)$, where now $\tau_\alpha = |\xi|\tau_\eta$ so there is an additional parameter. Consider the likelihood conditional on μ, η_j: we have that $(y_{ij} - \mu)/\eta_j \sim \mathsf{No}(\xi, \sigma^2/\eta_j)$ so conditionally conjugate priors for ξ and τ_η are respectively Gaussian an inverse-gamma. This translates into a prior distribution for τ_α which is that of the absolute value of a noncentral Student-t with location, scale and degrees of freedom ν. If we set the location to zero, the prior puts high mass at the origin, but is heavy tailed with polynomial decay. We recommend to set degrees of freedom so that the variance is heavy-tailed, e.g., $\nu = 3$. While this prior is not conjugate, it compares favorably to the $\mathsf{IGa}(\epsilon, \epsilon)$ that used to be widespread with $\epsilon > 0$ typically set to 0.01 or 0.001, approaching an improper prior. Posterior inference is unfortunately sensitive to the value of ϵ in hierarchical models when the random effect variance is close to zero, and more so when there are few levels for the groups since the relative weight of the prior relative to that of the likelihood contribution is then large.

Example 2.12 (Poisson random effect models)**.** We consider data from an experimental study conducted at Tech3Lab on road safety. In Brodeur et al. (2021), 31 participants were asked to drive in a virtual environment; the number of road violation was measured for different type of distractions (phone notification, phone on speaker, texting and smartwatch). The data are balanced, with each participant exposed to each task exactly once.

We model the data using a Poisson mixed model to measure the number of violations, `nviolation`, with a fixed effect for `task`, which captures the type of distraction, and a random effect for participant `id`. The hierarchical model fitted for individual i and distraction type j is

$$\begin{aligned} Y_{ij} &\sim \mathsf{Po}\{\mu = \exp(\beta_j + \alpha_i)\}, \qquad i = 1, \ldots, 31; j = 1, \ldots, 4 \\ \beta_j &\sim \mathsf{No}(0, 100) \qquad j = 1, \ldots, 4 \\ \alpha_i &\sim \mathsf{No}(0, \kappa^2); \qquad i = 1, \ldots, 31 \kappa \sim \mathsf{St}_+(3) \end{aligned}$$

so observations are conditionally independent given hyperparameters α and β.

In frequentist statistics, there is a distinction made in mixed-effect models between parameters that are treated as constants, termed fixed effects and corresponding in this example

to β, and random effects, equivalent to α. There is no such distinction in the Bayesian paradigm, except perhaps for the choice of prior.

We can look at some of posterior distribution of the 31 random effects (here the first five individuals) and the fixed effect parameters β, plus the variance of the random effect κ: there is strong evidence that the latter is non-zero, suggesting strong heterogeneity between individuals. The distraction which results in the largest number of violation is texting, while the other conditions all seem equally distracting on average (note that there is no control group with no distraction to compare with, so it is hard to draw conclusions).

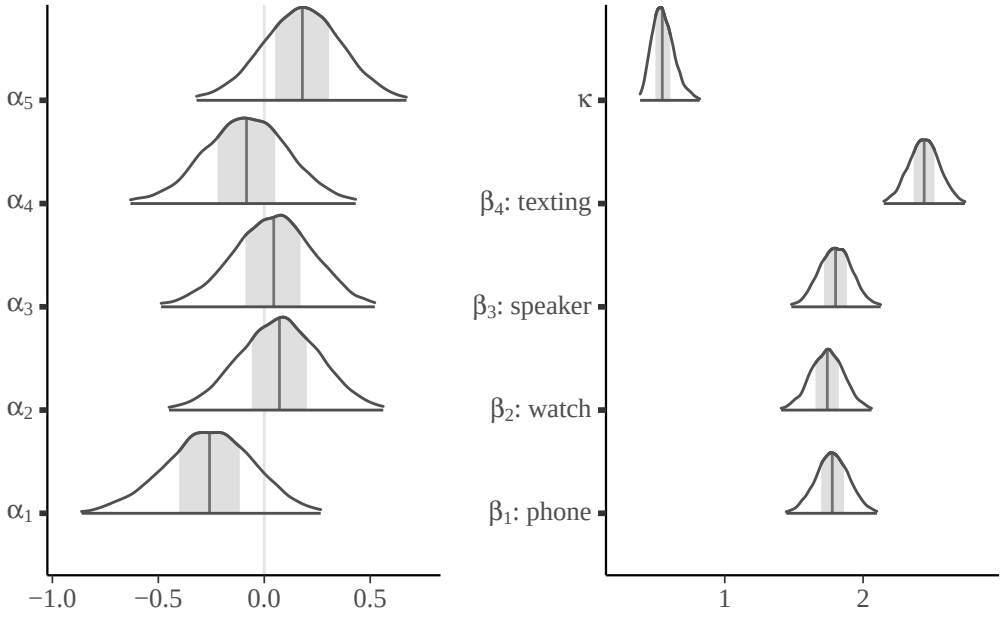

Figure 2.5: Posterior density plots with 50% credible intervals and median value for the random effects of the first five individuals (left) and the fixed effects and random effect variance (right).

2.7 Penalized complexity priors

Oftentimes, there will be a natural family of prior density to impose on some model component, $p(\boldsymbol{\theta} \mid \zeta)$, with hyperparameter ζ. The flexibility of the underlying construction leads itself to overfitting. Penalized complexity priors (Simpson et al. 2017) aim to palliate this by penalizing models far away from a simple baseline model, which correspond to a fixed

value ζ_0. The prior will favour the simpler parsimonious model the more prior mass one places on ζ_0, which is in line with Occam's razor principle.

To construct a penalized-complexity prior, we compute the Kullback–Leibler divergence between the model $p_\zeta \equiv p(\boldsymbol{\theta} \mid \zeta)$ relative to the baseline with ζ_0, $p_0 \equiv p(\boldsymbol{\theta} \mid \zeta_0)$; the Kullback–Leibler divergence is

$$\mathsf{KL}(p_\zeta \| \, p_0) = \int p_\zeta \log\left(\frac{p_\zeta}{p_0}\right) \mathrm{d}\boldsymbol{\theta}.$$

The distance between the prior densities is then set to $d(\zeta) = \{2\mathsf{KL}(p_\zeta \mid\mid p_0)\}^{1/2}$. which is zero at the model with ζ_0. The PC prior then constructs an exponential prior on the distance scale, which after back-transformation gives $p(\zeta \mid \lambda) = \lambda \exp(-\lambda d(\zeta)) \, |\partial d(\zeta)/\partial \zeta|$. To choose λ, the authors recommend elicitation of a pair (U, α) such that $\Pr(\lambda > U) = \alpha$.

Example 2.13 (Penalized complexity prior for random effects models)**.** Simpson et al. (2017) give the example of a Gaussian prior for random effects $\boldsymbol{\alpha}$, of the form $\boldsymbol{\alpha} \mid \zeta \sim \mathsf{No}_J(\mathbf{0}_J, \zeta^2 \mathbf{I}_J)$ where $\zeta_0 = 0$ corresponds to the absence of random subject-variability. The penalized complexity prior for the scale ζ is then an exponential with rate λ,[8] with density $p(\zeta \mid \lambda) = \lambda \exp(-\lambda\zeta)$. Using the recommendation for setting λ, we get that $\lambda = -\ln(\alpha/U)$ and this can be directly interpreted in terms of standard deviation of ζ; simulation from the prior predictive may also be used for calibration.

Example 2.14 (Penalized complexity prior for autoregressive model of order 1)**.** Sørbye and Rue (2017) derive penalized complexity prior for the Gaussian stationary AR(1) model with autoregressive parameter $\phi \in (-1, 1)$, where $Y_t \mid Y_{t-1}, \phi, \sigma^2 \sim \mathsf{No}(\phi Y_{t-1}, \sigma^2)$. There are two based models that could be of interest: one with $\phi = 0$, corresponding to lack of autocorrelation, and a static mean $\phi = 1$ for no change in time, which is not stationary. For the former, the penalized complexity prior is

$$p(\phi \mid \lambda) = \frac{\lambda}{2} \exp\left[-\lambda\left\{-\ln(1-\phi^2)\right\}^{1/2}\right] \frac{|\phi|}{(1-\phi^2)\left\{-\ln(1-\phi^2)\right\}^{1/2}}.$$

One can set λ by considering plausible values by relating the parameter to the variance of the one-step ahead forecast error.

2.8 Sensitivity analysis

Do priors matter? The answer to that question depends strongly on the model, and how much information the data provides about hyperparameters. While this question is easily

[8] Possibly truncated above if the support of ζ has a finite upper bound.

answered in conjugate models (the relative weight of hyperparameters relative to data can be derived from the posterior parameters), it is not so simple in hierarchical models, where the interplay between prior distributions is often more intricate. To see the impact, one often has to rely on doing several analyses with different values fr the prior and see the sensitivity of the conclusions to these changes, for example by considering a vague prior or modifying the parameters values (say halving or doubling). If the changes are immaterial, then this provides reassurance that our analyses are robust.

Example 2.15. To check the sensitivity of the conclusion, we revisit the modelling of the `smartwatch` experiment data using a Poisson regression and compare four priors: a uniform prior truncated to $[0, 10]$, an inverse gamma $\mathsf{IG}(0.01, 0.01)$ prior, a penalized complexity prior such that the 0.95 percentile of the scale is 5, corresponding to $\mathsf{Exp}(0.6)$. Since each distraction type appears 31 times, there is plenty of information to reliably estimate the dispersion κ of the random effects α: the different density plots in Figure 2.6 are virtually indistinguishable from one another.

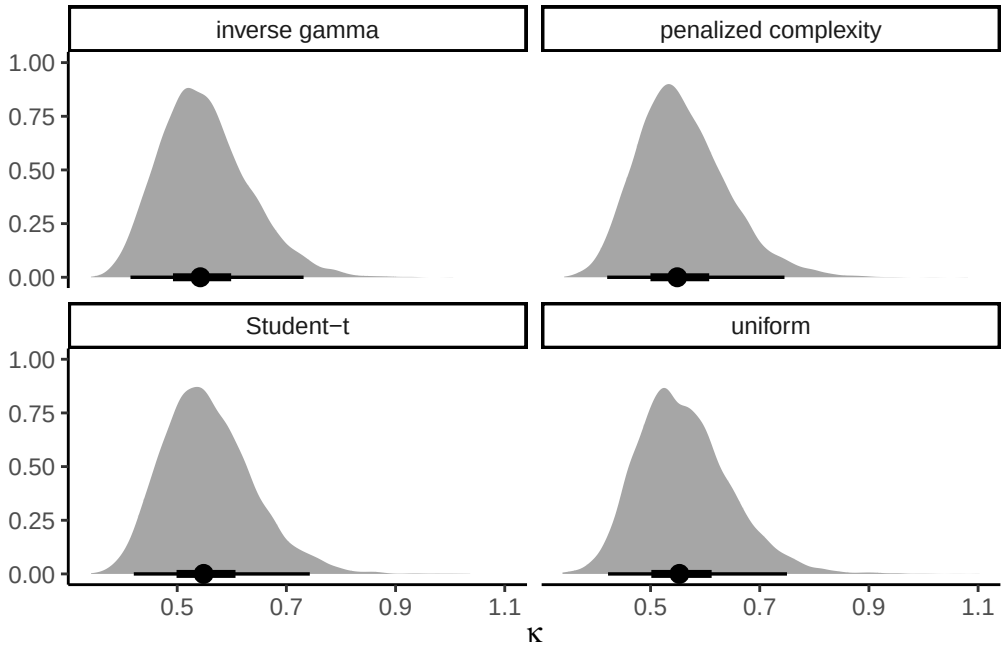

Figure 2.6: Posterior density of the scale of the random effects with uniform, inverse gamma, penalized complexity and folded Student-t with three degrees of freedom. The circle denotes the median and the bars the 50% and 95% percentile credible intervals.

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in R*. Boca Raton, FL: CRC Press.

Brodeur, Mathieu, Perrine Ruer, Pierre-Majorique Léger, and Sylvain Sénécal. 2021. "Smartwatches Are More Distracting Than Mobile Phones While Driving: Results from an Experimental Study." *Accident Analysis & Prevention* 149: 105846. https://doi.org/10.1016/j.aap.2020.105846.

Coles, Stuart G., and Jonathan A. Tawn. 1996. "A Bayesian Analysis of Extreme Rainfall Data." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 45 (4): 463–78. https://doi.org/10.2307/2986068.

Finetti, Bruno de. 1974. *Theory of Probability: A Critical Introductory Treatment*. Vol. 1. New York: Wiley.

Gelfand, Alan E., and Adrian F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (410): 398–409. https://doi.org/10.1080/01621459.1990.10476213.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis* 1 (3): 515–34. https://doi.org/10.1214/06-BA117A.

Geman, Stuart, and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 721–41. https://doi.org/10.1109/TPAMI.1984.4767596.

Matias, J. Nathan, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. 2021. "The Upworthy Research Archive, a Time Series of 32,487 Experiments in U.S. Media." *Scientific Data* 8 (195). https://doi.org/10.1038/s41597-021-00934-7.

McNeil, A. J., R. Frey, and P. Embrechts. 2005. *Quantitative Risk Management: Concepts, Techniques, and Tools*. 1st ed. Princeton, NJ: Princeton University Press.

Simpson, Daniel, Håvard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye. 2017. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science* 32 (1): 1–28. https://doi.org/10.1214/16-STS576.

Sørbye, Sigrunn Holbek, and Håvard Rue. 2017. "Penalised Complexity Priors for Stationary Autoregressive Processes." *Journal of Time Series Analysis* 38 (6): 923–35. https://doi.org/10.1111/jtsa.12242.