Bayesian Analysis

Contents

Configuring R

Functions from these packages will be used throughout this document:

```
library(conflicted) # check for conflicting function definitions
# library(printr) # inserts help-file output into markdown output
library(rmarkdown) # Convert R Markdown documents into a variety of formats.
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggfortify) # help with graphics
library(dplyr) # manipulate data
library(tibble) # `tibble`s extend `data.frame`s
library(magrittr) # `%>%` and other additional piping tools
library(haven) # import Stata files
library(knitr) # format R output for markdown
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
library(haven) # import Stata files
library(latex2exp) # use LaTeX in R code (for figures and tables)
library(fs) # filesystem path manipulations
library(survival) # survival analysis
library(survminer) # survival analysis graphics
library(KMsurv) # datasets from Klein and Moeschberger
```

```r
library(parameters) # format model output tables for
library(webshot2) # convert interactive content to static for pdf
library(forcats) # functions for categorical variables ("factors")
library(stringr) # functions for dealing with strings
library(lubridate) # functions for dealing with dates and times
library(broom) # Summarizes key information about statistical objects in tidy tibbles
library(broom.helpers) # Provides suite of functions to work with regression model 'broom::tidy()' t
```

Here are some R settings I use in this document:

```r
rm(list = ls()) # delete any data that's already loaded into R

conflicts_prefer(dplyr::filter)
ggplot2::theme_set(
  ggplot2::theme_bw() +
        # ggplot2::labs(col = "") +
    ggplot2::theme(
      legend.position = "bottom",
      text = ggplot2::element_text(size = 12, family = "serif")))

knitr::opts_chunk$set(message = FALSE)
options('digits' = 6)

panderOptions("big.mark", ",")
pander::panderOptions("table.emphasize.rownames", FALSE)
pander::panderOptions("table.split.table", Inf)
conflicts_prefer(dplyr::filter) # use the `filter()` function from dplyr() by default
legend_text_size = 9
run_graphs = TRUE
```

This appendix introduces the **Bayesian** approach to statistical inference. The structure parallels the development in (Dobson and Barnett 2018, chap. 12): we contrast the frequentist and Bayesian paradigms, state Bayes' theorem as an inference rule for parameters, discuss the role of the prior distribution, and outline hierarchical models and the software used to fit them. Computation by simulation is deferred to the next appendix[1], and fully worked analyses appear in the appendix of example analyses[2].

1 Frequentist and Bayesian paradigms

Throughout these notes we have treated an unknown parameter θ as a fixed but unknown constant, and we have quantified uncertainty using the sampling distribution of an estimator $\hat{\theta}$ — that is, the distribution of $\hat{\theta}$ over hypothetical repetitions of the data-generating process. This stance is the **frequentist** paradigm: probability describes long-run frequencies, and a parameter, being a constant, has no probability distribution.

The **Bayesian** paradigm instead treats the unknown parameter as a **random variable** with its own probability distribution (Dobson and Barnett 2018, chap. 12, p. 271). Probability here measures *degree of belief*: before seeing data we encode our uncertainty about θ in a **prior distribution** $p(\theta)$, and after observing data $\tilde{y}$ we update that belief to a **posterior distribution** $p(\theta \mid \tilde{y})$. All Bayesian inference flows from this single posterior distribution.

The two paradigms answer subtly different questions. A frequentist asks, "for what parameter values would these data be unsurprising?"; a Bayesian asks, "given these data, what should I now believe about the parameter?". Neither question is wrong; they are different, and they call for different machinery.

[1] mcmc-methods.qmd#sec-mcmc-methods
[2] bayesian-examples.qmd#sec-bayes-examples

1.1 Two readings of an interval estimate

The distinction is sharpest in how each paradigm interprets an interval estimate (Dobson and Barnett 2018, chap. 12, p. 271). A frequentist 95% **confidence interval** $(l(\tilde{y}), r(\tilde{y}))$ is a random interval whose *coverage* — the probability, over repeated sampling, that the interval contains the fixed true θ — is 0.95. For one realized data set the interval either contains θ or it does not; the 0.95 refers to the procedure, not to the particular interval.

A Bayesian **credible interval** makes the statement many users *want* to make about a confidence interval but cannot:

Definition 1.1 (Credible interval)**.** A $100(1-\alpha)\%$ **credible interval** for θ is an interval $(l(\tilde{y}), r(\tilde{y}))$ such that, under the posterior distribution,

$$p[l(\tilde{y}) \leq \theta \leq r(\tilde{y}) \mid \tilde{y}] \overset{\text{def}}{=} 1-\alpha.$$

Here the data $\tilde{y}$ are fixed at their observed values and θ is random, so the probability statement is about θ directly.

2 Bayes' theorem for parameters

The engine of Bayesian inference is **Bayes' theorem**, applied to the parameter and the data. Writing $p(\theta)$ for the prior and $\mathcal{L}(\theta) = p(\tilde{y} \mid \theta)$ for the likelihood, the posterior is

Definition 2.1 (Posterior distribution)**.**

$$p(\theta \mid \tilde{y}) \overset{\text{def}}{=} \frac{p(\tilde{y} \mid \theta)\, p(\theta)}{p(\tilde{y})}, \qquad p(\tilde{y}) = \int p(\tilde{y} \mid \theta)\, p(\theta)\, d\theta. \tag{1}$$

The denominator $p(\tilde{y})$ — the **marginal likelihood** or *evidence* — does not depend on θ; it is the constant that makes the posterior integrate to one. Consequently the posterior is determined up to proportionality by the numerator alone (Dobson and Barnett 2018, chap. 12, p. 272):

$$\underbrace{p(\theta \mid \tilde{y})}_{\text{posterior}} \propto \underbrace{p(\tilde{y} \mid \theta)}_{\text{likelihood}} \cdot \underbrace{p(\theta)}_{\text{prior}}. \tag{2}$$

This proportionality is the workhorse of applied Bayesian analysis: it lets us recognize the posterior from the shape of likelihood $\times$ prior without ever computing the integral $p(\tilde{y})$, and it is what makes the simulation methods of the next appendix[3] possible.

2.1 A Gaussian mean with a Gaussian prior

When prior and likelihood combine to give a posterior in the same family as the prior, the prior is called **conjugate**. The normal-mean model is the cleanest example.

Exm

Example 2.1 (Posterior for a normal mean)**.** Suppose $X_1, \ldots, X_n \sim_{\text{iid}} \mathrm{N}(\mu, 1)$ with the variance known, and place a $\mathrm{N}(0,1)$ prior on the mean: $M \sim \mathrm{N}(0,1)$. Collecting the terms of $p(M=\mu \mid X=x) \propto p(X=x \mid M=\mu)\, p(M=\mu)$ that involve μ and completing the square:

[3] mcmc-methods.qmd#sec-mcmc-methods

$$
\begin{aligned}
p(M=\mu \mid X=x) &\propto p(X=x \mid M=\mu)\, p(M=\mu) \\
&\propto \exp\Big\{-\frac{1}{2}(n\mu^2 - 2\mu n\bar{x})\Big\} \cdot \exp\Big\{-\frac{1}{2}\mu^2\Big\} \\
&= \exp\Big\{-\frac{1}{2}((n+1)\mu^2 - 2\mu n\bar{x})\Big\} \\
&\propto \exp\Big\{-\frac{1}{2}(n+1)\Big(\mu - \frac{n}{n+1}\bar{x}\Big)^2\Big\}.
\end{aligned}
$$

The last line is the kernel of a normal density, so the posterior is

$$
M \mid X = x \;\sim\; \mathrm{N}\Big(\frac{n}{n+1}\bar{x},\ \frac{1}{n+1}\Big).
$$

The posterior mean $\frac{n}{n+1}\bar{x}$ is the sample mean shrunk toward the prior mean 0, and the shrinkage vanishes as $n \to \infty$.

3 The parameter space

Because the Bayesian treats θ as random, the prior and posterior are distributions over the **parameter space** Θ — the set of values θ can take (Dobson and Barnett 2018, chap. 12, p. 274). The parameter space must be respected when choosing a prior: a probability $\pi \in (0,1)$ calls for a prior supported on the unit interval (such as a Beta distribution), a variance $\sigma^2 > 0$ calls for a prior on the positive half-line, and a vector of regression coefficients calls for a multivariate prior on $\mathbb{R}^p$. A prior that places no mass on part of Θ forces the posterior to do the same, no matter what the data say — so the support of the prior is itself a modeling assumption.

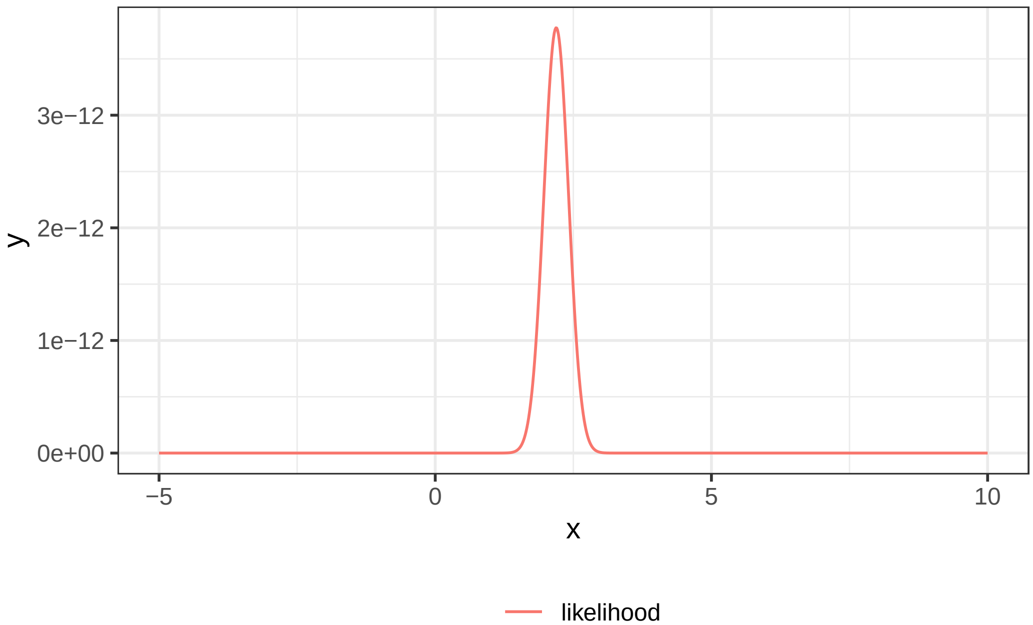
Exm

Example 3.1 (Credible vs. confidence interval for a normal mean)**.** We compare the two interval estimates from Section 1.1 on simulated data from the model of Example 2.1. First, the frequentist interval:

```
set.seed(1)
mu <- 2
sigma <- 1
n <- 20
x <- rnorm(n = n, mean = mu, sd = sigma)
xbar <- mean(x)
se <- sigma / sqrt(n)
CI_freq <- xbar + se * qnorm(c(.025, .975))
print(CI_freq)
#> [1] 1.75226 2.62879
```

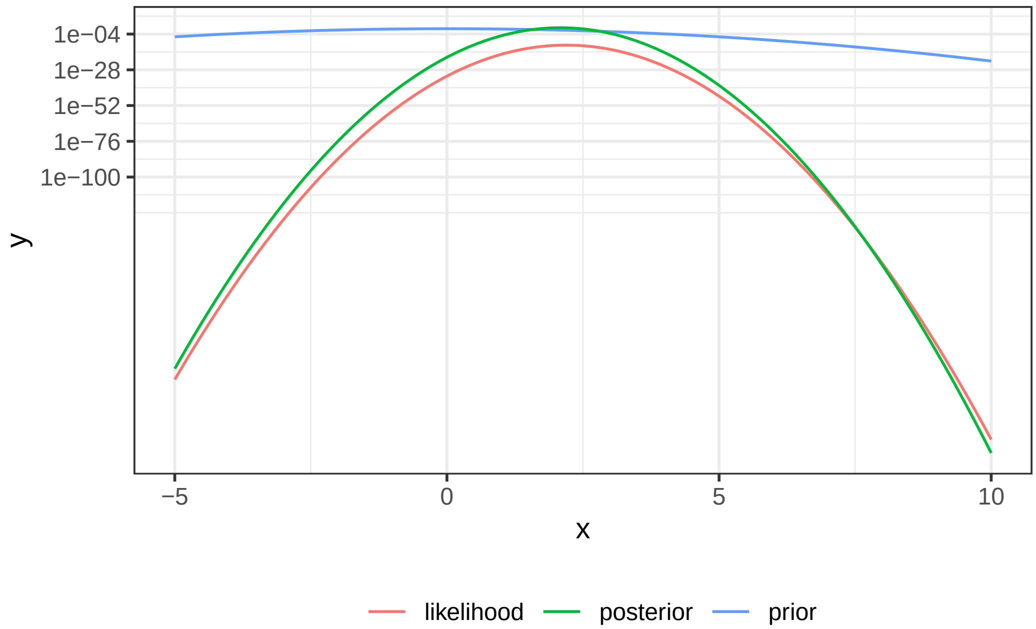
The likelihood as a function of μ:

```r
lik <- function(mu) {
  (2 * pi * sigma^2)^(-n / 2) *
    exp(
      -1 / (2 * sigma^2) *
        (sum(x^2) - 2 * mu * sum(x) + n * (mu^2))
    )
}
ngraph <- 1001
plot1 <- ggplot() +
  geom_function(fun = lik, aes(col = "likelihood"), n = ngraph) +
  xlim(c(-5, 10)) +
  theme_bw() +
  labs(col = "") +
  theme(legend.position = "bottom")
print(plot1)
```



Next, the Bayesian credible interval, using the closed-form posterior $\mathrm{N}\left(\frac{n}{n+1}\bar{x},\, (n+1)^{-1}\right)$ from Example 2.1:

```r
mu_prior_mean <- 0
mu_prior_sd <- 1
mu_post_mean <- n / (n + 1) * xbar
mu_post_var <- 1 / (n + 1)
mu_post_sd <- sqrt(mu_post_var)
CI_bayes <- qnorm(
  p = c(.025, .975),
  mean = mu_post_mean,
  sd = mu_post_sd
)
print(CI_bayes)
#> [1] 1.65851 2.51391
prior <- function(mu) dnorm(mu, mean = mu_prior_mean, sd = mu_prior_sd)
posterior <- function(mu) dnorm(mu, mean = mu_post_mean, sd = mu_post_sd)
plot2 <- plot1 +
  geom_function(fun = prior, aes(col = "prior"), n = ngraph) +
  geom_function(fun = posterior, aes(col = "posterior"), n = ngraph)
print(plot2 + scale_y_log10())
```



Finally, the quantity a credible interval reports — the posterior probability that M lies in the *frequentist* interval — is well defined for the Bayesian but not for the frequentist:

```r
pr_in_CI <- pnorm(
  CI_freq,
  mean = mu_post_mean,
  sd = mu_post_sd
) |> diff()
print(pr_in_CI)
#> [1] 0.930583
```

4 Priors

The prior distribution $p(\theta)$ encodes what is known about θ before the current data are seen. Choosing it is the step that most distinguishes Bayesian practice from frequentist practice, and it is where most of the controversy and most of the craft live (Dobson and Barnett 2018, chap. 12, p. 281).

4.1 Conjugate priors

A prior is **conjugate** to a likelihood when the resulting posterior belongs to the same parametric family as the prior. Conjugate priors make the posterior available in closed form, as in Example 2.1. The Beta family is conjugate to the Bernoulli/binomial likelihood:

Exm

Example 4.1 (Beta-Bernoulli updating)**.** Let $Y_1, \dots, Y_n \sim_{\text{iid}}$ Bernoulli(π) with $r = \sum_i y_i$ successes, and place a Beta(a, b) prior on π. By Equation 2,

$$p(\pi \mid \tilde{y}) \propto \underbrace{\pi^r (1-\pi)^{n-r}}_{\text{likelihood}} \cdot \underbrace{\pi^{a-1}(1-\pi)^{b-1}}_{\text{prior}} = \pi^{a+r-1}(1-\pi)^{b+n-r-1},$$

which is the kernel of a Beta($a + r,\ b + n - r$) density. The prior parameters a and b act as **pseudo-counts** of prior successes and failures: the posterior simply adds the observed r successes and $n - r$ failures to them. With a uniform Beta$(1, 1)$ prior, $n = 91$, and $r = 55$, the posterior is Beta$(56, 37)$:

```r
a_prior <- 1
b_prior <- 1
n_obs <- 91
r_obs <- 55
a_post <- a_prior + r_obs
b_post <- b_prior + n_obs - r_obs
round(c(
  post_mean = a_post / (a_post + b_post),
  post_lower = qbeta(0.025, a_post, b_post),
  post_upper = qbeta(0.975, a_post, b_post)
), 4)
#>  post_mean post_lower post_upper
#>     0.6022     0.5014     0.6988
```

4.2 Informative, weakly informative, and flat priors

Priors range along a spectrum of how much they constrain θ (Dobson and Barnett 2018, chap. 12, p. 281):

- An **informative** prior concentrates mass in a region supported by external knowledge (previous studies, biological limits, expert judgment). It pulls the posterior toward that region, most strongly when the data are sparse.
- A **weakly informative** prior rules out implausible values (for example, an odds ratio of 10^6) while remaining flat across the range of scientifically plausible values.
- A **flat** (or **noninformative**) prior aims to "let the data speak," approximating constant density over Θ. Flat priors can be *improper* (not integrating to one) and are not invariant to reparameterization, so they require care.

4.3 A skeptical prior

A useful special case is a **skeptical prior**, centered on "no effect" and deliberately tight, used to ask how strong the data must be to overturn a default of no effect (Dobson and Barnett 2018, chap. 12, p. 281).

Exm

Example 4.2 (A skeptical prior for a normal mean)**.** Reusing the normal-mean model of Example 2.1 but with a skeptical $\mathrm{N}(0, \tau^2)$ prior of standard deviation τ, the posterior mean is the precision-weighted average

$$\mathrm{E}[M \mid X = x] = \frac{n\,\bar{x}}{n + 1/\tau^2},$$

which shrinks toward 0 more aggressively as the prior becomes more skeptical (smaller τ). We tabulate the posterior mean across a range of prior skepticism, holding the data fixed:

```
xbar_obs <- 2
n_obs2 <- 20
tau_grid <- c(0.1, 0.25, 0.5, 1, 2, 10)
post_mean <- (n_obs2 * xbar_obs) / (n_obs2 + 1 / tau_grid^2)
round(
  data.frame(prior_sd = tau_grid, posterior_mean = post_mean),
  3
)
#> # A tibble: 6 x 2
#>   prior_sd posterior_mean
#>      <dbl>          <dbl>
#> 1     0.1           0.333
#> 2     0.25          1.11
#> 3     0.5           1.67
#> 4     1             1.90
#> 5     2             1.98
#> 6    10             2.00
```

A very skeptical prior ($\tau = 0.1$) drags the posterior mean almost to zero despite $\bar{x} = 2$; a diffuse prior ($\tau = 10$) leaves it essentially at the sample mean.

5 Distributions and hierarchies

Many models have parameters that are themselves naturally described by a distribution with further unknown parameters (Dobson and Barnett 2018, chap. 12, p. 281). A **hierarchical** (or **multilevel**) model builds the prior in stages: a **hyperprior** governs **hyperparameters**, which govern the group-level parameters, which govern the data.

For example, with groups $j = 1, \dots, J$, each with its own mean θ_j, a two-level model might specify

$$\begin{aligned} Y_{ij} \mid \theta_j &\sim \mathrm{N}(\theta_j,\, \sigma^2), \\ \theta_j \mid \mu, \tau &\sim \mathrm{N}(\mu,\, \tau^2), \\ \mu, \tau &\sim \text{hyperprior}. \end{aligned}$$

The middle level lets the groups **borrow strength** from one another: the posterior for each θ_j is pulled toward the overall mean μ, by an amount governed by the between-group spread τ that the data themselves estimate. This random-effects (multilevel) *model structure* is the same one fit by maximum likelihood in (Dobson and Barnett 2018, chap. 11); the Bayesian *inference method* differs only in placing a hyperprior on μ and τ and returning a full posterior for them, rather than point estimates of the variance components. The random-effects example[4] fits the same structure by simulation.

6 Software for Bayesian computation

Outside conjugate special cases, the posterior rarely has a closed form, and the normalizing constant $p(\tilde{y})$ in Equation 1 is an intractable integral. Practical Bayesian analysis therefore relies on simulation — **Markov chain Monte Carlo (MCMC)** — implemented in dedicated software (Dobson and Barnett 2018, chap. 12, p. 281). The Dobson and Barnett text uses WinBUGS; in these notes we use **JAGS** ("Just Another Gibbs Sampler") through R, which follows the same model-specification style.

[4] bayesian-examples.qmd#sec-bayes-examples-random-effects

Related tools include **Stan** (Hamiltonian Monte Carlo) and the R package **`brms`**, which compiles familiar formula syntax to Stan.

The mechanics of MCMC — why it is needed, how the samplers work, and how to check that they have converged — are the subject of the next appendix[5]. Worked analyses of real data sets follow in the appendix of example analyses[6].

References

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. https://doi.org/10.1201/9781315182780.

[5] mcmc-methods.qmd#sec-mcmc-methods
[6] bayesian-examples.qmd#sec-bayes-examples