Bayesian Methods for Clinical Trials

Lecture 2: Introduction to Bayesian Statistics

Libby Daniells & Pavel Mozgunov & Thomas Jaki

MRC Biostatistics Unit

February 24, 2025

- 15 people who reported a headache were given aspirin.
- 10 reported their headache vanishing within half an hour.

What inferences can we make about **probability of response**?

Conventionally, the probability of response could be given by **expectation** $\hat{p} = 10/15 \approx 0.67$.

- Frequentist estimate for the standard error could be $\mathrm{se} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = \sqrt{\frac{0.67(1-0.67)}{15}} = 0.12$
- 95% confidence interval for p is (0.428,0.905).
- Problems with what the resultant CI means and how to use it.

Bayesian methods offer a different approach

- Consideration of the parameter of interest, denoted by θ.
- In this case, θ is the probability of headache vanishing within half an hour given aspirin, is in itself a **random variable** that has a distribution (reflecting our uncertainty in it).
- Information about the nature of this random variable comes from the sample.

How do we go about calculating a suitable distribution for θ when θ is a continuous variable taking values between 0 and 1?

Notation

- Let θ be the parameter of interest with prior distribution $\pi(\theta)$.
- The observations are $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, and have marginal density function $f(\mathbf{x})$.
- The likelihood function $f(\mathbf{x}|\theta)$ is the distribution of $\mathbf{x}$ conditional on specific values of θ.

Then:

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

Note: π is used to denote functions of θ, and f functions of the data.

Data marginal distribution

- The function $f(\mathbf{x})$ is the marginal distribution of the data.
- For most problems, we are not that interested in inferences on $\mathbf{x}$, but in θ.

Because of this we can write the posterior distribution as:

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{1}{f(\mathbf{x})}\pi(\theta)f(\mathbf{x}|\theta) \\ &\propto \pi(\theta)L(\theta)\end{aligned}$$

as $f(\mathbf{x})$ can be regarded as a normalising constant.

Likelihood function

Critical to Bayesian inference are likelihood functions. These

- serve to link the **sampling** space to the **parameter** space.

The probability:

$$f(\mathbf{x}|\theta) = L(\theta) = \prod_i f(x_i|\theta)$$

is a general expression for a likelihood function given *iid* data.

Example: Bernoulli likelihood

- The observation of 10 responses from 15 patients is a series of Bernoulli trials with $\mathbf{x} = \{0, 1\}$. For any single trial i the probability mass function is

$$f(x_i | \theta) = \theta^{x_i}(1 - \theta)^{1 - x_i}$$

- The likelihood function is therefore:

$$L(\theta) = f(\mathbf{x} | \theta) = \prod_{i=1}^{n} \theta^{x_i}(1 - \theta)^{1 - x_i} = \theta^{\sum_i x_i}(1 - \theta)^{n - \sum_i x_i}$$

- As for a series of $\sum x_i$ where $\mathbf{x} = \{1, 0\}$ is the number of "successes", and n the number of trials, then $n - \sum x_i$ can be seen as the number of "failures". $\sum x_i$ can be denoted s.
- This is the functional form of the *Beta* distribution.
- θ is restricted such that $0 \leq \theta \leq 1$.

Prior: Beta Distribution

Consider the Beta prior distribution $\mathcal{B}(a, b)$ of the probability of response

$$\pi(\theta) = \frac{1}{\mathrm{Beta}(a, b)} \theta^{a-1} (1 - \theta)^{b-1}$$

where $\mathrm{Beta}(a, b)$ is Beta function.

The parameters of the prior distribution are called **hyperparameters** to distinguish them from the parameters of the sampling space.

Posterior Distribution

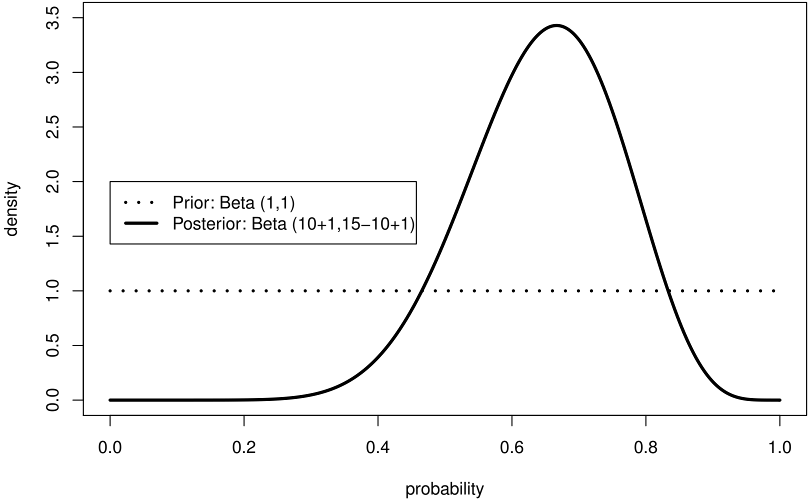
$$
\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto \pi(\theta) f(\mathbf{x}|\theta) \\
&\propto \left[\theta^{a-1}(1-\theta)^{b-1}\right] \times \left[\theta^{s}(1-\theta)^{n-s}\right] \\
&= \left[\theta^{a-1}\theta^{s}(1-\theta)^{b-1}(1-\theta)^{n-s}\right] \\
&= \theta^{a+s-1}(1-\theta)^{b+n-s-1}
\end{aligned}
$$

It follows that the posterior distribution is

$$
\theta|\boldsymbol{x} \sim \mathrm{Beta}\left(s+a, n-s+b\right)
$$

The chosen Beta prior is called **conjugate prior**: A prior is conjugate to likelihood if posterior is of the same family as prior.

Beta Posterior



- For the *Beta* distribution, there is a number of summaries which are readily available:
  1. $$E(\theta) = \frac{a+s}{a+b+n}$$
  2. $$\mathrm{mode} = \frac{a+s-1}{a+b+n-1}$$
  3. $$\mathrm{var}(\theta) = \frac{(a+s)(b+n-s)}{(n+a+b)^2(n+a+b+1)}$$

Posterior means and variances are useful but don't tell us everything.

The posterior distribution can give as various sort of information about the distribution of the parameter.

A specific region which contains a given area of the posterior density function known as a **Credible Interval.**

Interpretation of the CrI: there is a probability of $(1-\alpha) \times 100\%$ that θ falls within the region.

- **Highest density region (HDR)**: smallest interval for θ which contains $(1-\alpha) \times 100\%$ of area.
- **Equal-Tailed Interval:** the interval where the probability of being below the interval is as likely as being above it.

Now, we would like to make a prediction. Assume that 10 more patients had headaches and were assigned aspirin.

What is the probability that 7 of them will respond?

$E(\theta) \approx 0.65$, so should we use the binomial distribution to compute it?

$$\Pr(X = 7) = \frac{10!}{k!(10-7)!} 0.65^7 (1-0.65)^{(10-7)}$$

```
dbinom(x=7,size=10,prob=11/17)
[1] 0.2505501
```

This, does not take into account the **uncertainty** about θ.

Predictive distribution

In many cases, where the observations are assumed to be independent and identically distributed given θ, we have that

$$f_Y(y|\mathbf{x}) = \int f(y|\theta)\pi(\theta|\mathbf{x})\, d\theta.$$

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Posterior predictive distribution

The posterior predictive distribution is the distribution of possible unobserved values conditional on the observed values.

A posterior predictive distribution accounts for uncertainty about θ. So, instead of computing $\Pr(X = 7|\theta = 0.65)$, we compute

$$\Pr(X = 7|data) = \int_0^1 \Pr(X = 7|\theta) \times \pi(\theta|\mathbf{x})\mathrm{d}\theta.$$

In our example, it is the **beta-binomial** distribution

$$\pi(k \mid N, \alpha, \beta) = \frac{\Gamma(N+1)}{\Gamma(k+1)\Gamma(N-k+1)} \frac{\Gamma(k+\alpha)\Gamma(N-k+\beta)}{\Gamma(N+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

where $\alpha = s + a$ and $\beta = n - s + b$ (parameters of our Beta posterior) and N is the number of new patients.

Posterior predictive distribution

$$\pi_{\text{pred}}(X = 7 | \textit{data}) \approx 0.205.$$

```
 library("rmutil")
 dbetabinom(7, 10, 11/17, 17)
[1] 0.2050343
```

Using our previous estimate of the probability, we overestimate the probability, as we it does not take into account that a lot of probability mass lies below $\theta = 11/17$.

Normal distribution

One of the most important continuous distributions in statistics.

$$X \sim N(\mu, \sigma^2)$$

Defined in terms of two parameters, the mean μ and variance σ^2.

Often parameterised in terms of the mean μ and precision $\tau = 1/\sigma^2$. Thus:

$$X \sim N(\mu, \sigma^2) \equiv N(\mu, 1/\tau).$$

For $X \sim N(\mu, \sigma^2) \equiv N(\mu, 1/\tau)$, X has probability density function

$$f(x) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right) \qquad (x \in \mathbb{R}).$$

Therefore the likelihood satisfies

$$\begin{aligned} L(\mu, \tau) = f(\mathbf{x}|\mu, \tau) &= \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2}(x_i-\mu)^2\right) \\ &\propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right). \end{aligned}$$

Known variance

For known variance we are interested in $\pi(\mu|\mathbf{x})$.

Mean μ can take any value between $-\infty$ and $+\infty$.

Natural candidate for the prior on μ: $N(\alpha, 1/\beta)$

Then

$$\begin{aligned}
\pi(\mu) &= \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2}(\mu-\alpha)^2\right) \\
&\propto \exp\left(-\frac{\beta}{2}(\mu-\alpha)^2\right).
\end{aligned}$$

Here α and β are hyperparameters:
μ was estimated from observations with total precision β and with sample mean α.

Known variance - posterior

Writing down the likelihood and prior, one can obtain that the posterior distribution has the form

$$\mu|\mathbf{x} \sim N\left(\frac{n\tau\bar{x} + \alpha\beta}{n\tau + \beta}, \frac{1}{n\tau + \beta}\right).$$

$\alpha = \beta = 0$?

If we set $\alpha = \beta = 0$, this corresponds to $N(0, 1/0) = N(0, \infty)$ prior.

This is known as an **improper** prior.

In this case:

$$\mu|\mathbf{x} \sim N\left(\frac{n\tau\bar{x} + 0}{n\tau + 0}, \frac{1}{n\tau + 0}\right) = N(\bar{x}, \sigma^2/n).$$

The posterior distribution for μ has mean $\bar{x}$, which is the MLE and variance σ^2/n, which is the frequentist standard error of the mean.

Thus we return to something familiar!

General Normal (Gaussian) prior

For a general prior, we can write the posterior mean as a weighted mean between the sample mean $\bar{x}$ (MLE) and prior mean.

Let

$$w = \frac{\beta}{\beta + n\tau},$$

then

$$\begin{aligned} \frac{n\tau\bar{x} + \alpha\beta}{n\tau + \beta} &= \frac{n\tau}{n\tau + \beta}\bar{x} + \frac{\beta}{n\tau + \beta}\alpha \\ &= (1 - w)\bar{x} + w\alpha. \end{aligned}$$

Note that as n increases, w gets smaller and approaches 0. In other words, as we get more and more data, the data dominates the prior.

Observe that "posterior precision" = "prior precision" + $n\times$ "precision of each data item".

Summary

- Using a weakly-informative (also referred to as "non-information" or "vague") prior yields exactly equivalent result to classical approaches.

- An informative prior, in the case above another normal, gives us a posterior distribution which is some weighted sum of the parameters of the prior and data.

- Classical approaches are often (in some sense) a subset of Bayesian methods.

- Can use similar approaches for when the variance cannot be assumed known.