Bayesian Inference in Single Parameter Models

Mark Andrews & Thom Baguley

September 25, 2018

Inference in a Bernoulli Distribution

> *...Polish mathematicians Tomasz Gliszczynski and Waclaw Zawadowski...spun a Belgian one euro coin 250 times, and found it landed heads up 140 times ...When tossed 250 times, the one euro coin came up heads 139 times and tails 111. ...*

The Guardian, January 4, 2002[1]

- A sample of $n = 250$ coin tosses can be modelled as n independent and identically dsibution Bernoulli random variables with parameter θ.
- In other words, our probabilitic model is

$$x_i \sim \text{Bernoulli}(\theta), \quad \text{for } i \in \{1, 2 \ldots n\}.$$

and we would like to inference the probable values of θ given an observation of $m = 139$ (or $m = 140$, etc.).

[1]See `http://bit.ly/1BOKu9b` for original story and `http://bit.ly/1BOKx4Q` for discussion.

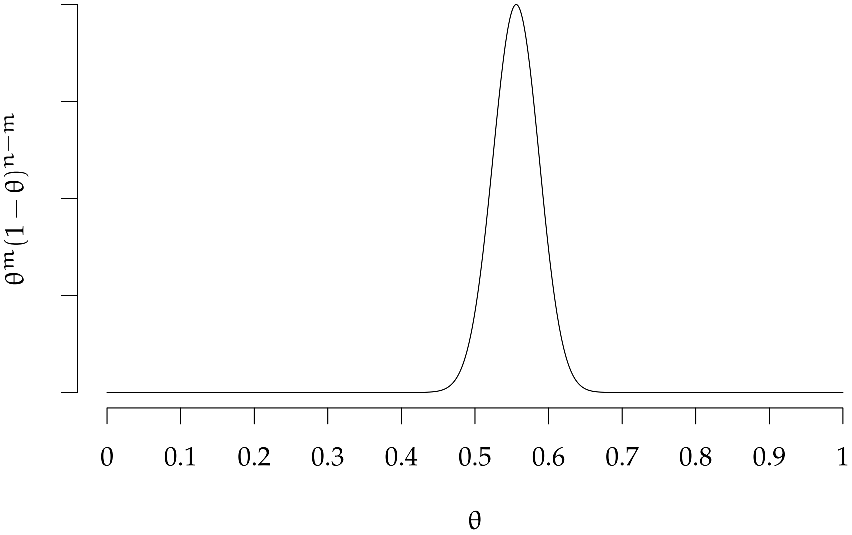
Inference in a Bernoulli Distribution

- The likelihood of the observed outcomes of the n coins is

$$\begin{aligned} P(x_1, x_2 \cdots x_n|\theta) &= \prod_{i=1}^{n} P(x_i|\theta), \\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}, \\ &= \theta^m(1-\theta)^{n-m}. \end{aligned}$$

- This is identical to a binomial distribution likelihood function.

Inference in a Bernoulli Distribution



The likelihood function for $n = 250$ and $m = 139$.

Inference in a Bernoulli Distribution

- The probabilistic generative model of the Euro coin toss data is as follows:
  - The coin's bias corresponds to the fixed but unknown value of the parameter θ of a Bernoulli random variable.
  - The observed outcomes $x_1, x_2 \cdots x_n$ are n iid samples from $\text{Bernoulli}(\theta)$.
- This generative model can be extended by assuming that θ is itself drawn from a prior distribution $P(\theta)$:

$$\begin{aligned} \theta &\sim P(\theta), \\ x_i &\sim \text{Bernoulli}(\theta), \quad \text{for } i \in \{1, 2 \ldots n\}. \end{aligned}$$

- In other words, we assume that a value for θ was randomly drawn from $P(\theta)$ and then n binary variables were sampled from $\text{Bernoulli}(\theta)$.

Conjugate Priors

- For a given likelihood function, a *conjugate prior* distribution is a prior probability distribution that leads to a posterior distribution of the same parameteric family.
- Using conjugate priors allows Bayesian inference and other probabilistic calculations to be performed analytically.
- Only a small subset of probabilistic models have conjugate priors.
- However, conjugate priors play a vital role in Monte Carlo methods like Gibbs sampling even in complex models.

The beta distribution

- For the binomial likelihood function

$$\theta^m(1-\theta)^{n-m}$$

a conjugate prior is the beta distribution

$$\text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

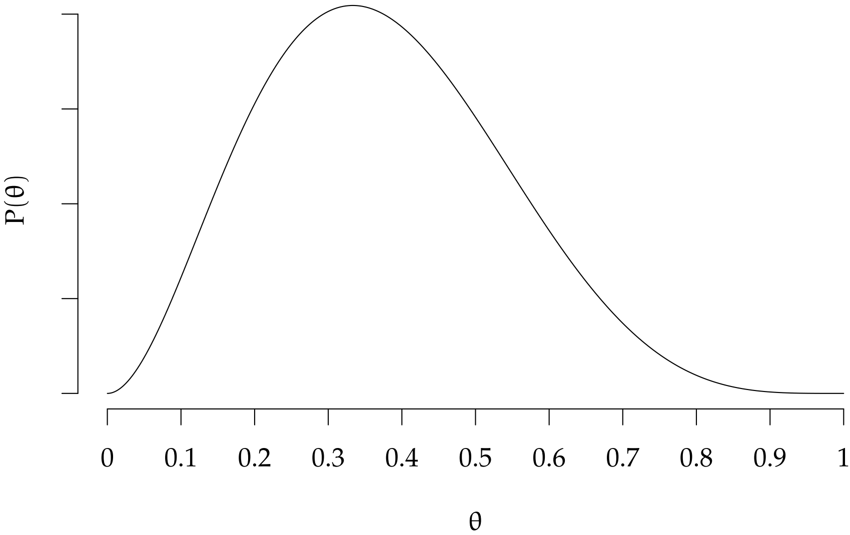
- The *normalizing constant* term is

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{1}{B(\alpha,\beta)},$$

where $B(\alpha, \beta)$ is the beta function:

$$B(\alpha,\beta) = \int \theta^{\alpha}(1-\theta)^{\beta-1}\, d\theta.$$

The beta distribution



The beta distribution with $\alpha = 3$ and $\beta = 5$.

Posterior distribution

- Denoting the observed data by $D = (n, m)$, with the beta prior, the posterior distribution is

$$
\begin{aligned}
P(\theta|D, \alpha, \beta) &= \frac{P(D|\theta)P(\theta|\alpha, \beta)}{\int P(D|\theta)P(\theta|\alpha, \beta)\, d\theta}, \\
&\propto \overbrace{\theta^{m}(1-\theta)^{n-m}}^{\text{likelihood}} \times \overbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}^{\text{prior}}, \\
&\propto \theta^{m+\alpha-1}(1-\theta)^{n-m+\beta-1}, \\
&= \mathrm{Beta}(m+\alpha, n-m+\beta).
\end{aligned}
$$

where the normalizing constant is the reciprocal of the beta function

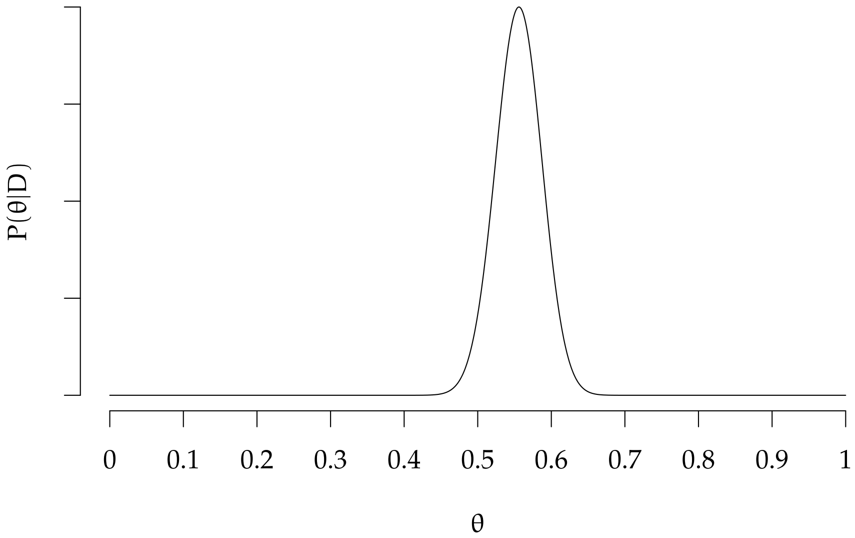
$$
\begin{aligned}
\frac{\Gamma(m+\alpha)\Gamma(n-m+\beta)}{\Gamma(n+\alpha+\beta)} &= \int \theta^{\alpha+m-1}(1-\theta)^{\beta+n-m-1}\, d\theta. \\
&= B(\alpha+m, \beta+n-m).
\end{aligned}
$$

Posterior distribution

- For our Euro coin example, our observed data are $n = 250$ and $m = 139$.
- A noninformative uniform prior on θ is $\mathrm{Beta}(\alpha = 1, \beta = 1)$.
- With this prior, the posterior distribution is

$$\begin{aligned}\mathrm{Beta}(m + \alpha, n - m + \beta) &= \mathrm{Beta}(139 + 1, 250 - 139 + 1), \\ &= \mathrm{Beta}(140, 112)\end{aligned}$$

Posterior distribution



The posterior distribution when $n = 250$, $m = 139$, $\alpha = 1$ and $\beta = 1$.

Summarizing the posterior distribution

- The mean, variance and modes of any beta distribution are as follows:

$$\langle\theta\rangle = \frac{\alpha}{\alpha+\beta},$$
$$V(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$
$$\text{mode}(\theta) = \frac{\alpha-1}{\alpha+\beta-2}.$$

- Thus, in our case of $\text{Beta}(140, 112)$, we have

$$\langle\theta\rangle = 0.5556,$$
$$V(\theta) = 0.001, \quad \text{sd}(\theta) = 0.0312,$$
$$\text{mode}(\theta) = 0.556.$$

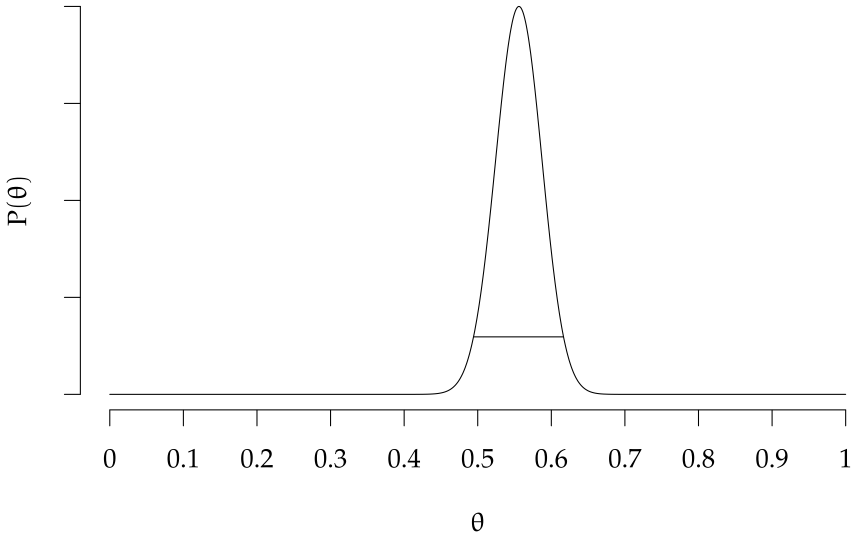
High posterior density (HPD) *intervals*

- HPD intervals provide ranges that contain specified probability mass. For example, the 0.95 HPD interval is the range of values that contain 0.95 of the probability mass of the distribution.
- The φ HPD interval for the probability density function $P(x)$ is computed by finding a probability density value p^* such that

$$P(\{x\colon P(x) \geqslant p^*\}) = \varphi.$$

- In other words, we find the value p^* such that the probability mass of the set of points whose density is greater than than p^* is exactly φ.
- In general, the HPD is not trivial to compute but in the case of symmetric distributions, it can be easily computed from the cumulative density function.

The 0.95 HPD *interval*



The posterior distribution, with its 0.95 HPD, when $n = 250$, $m = 139$, $\alpha = 1$ and $\beta = 1$. In this case, the HPD interval is $(0.494, 0.617)$.

Posterior predictive distribution

- Given that we have observed m heads in n coin tosses, what is the probability that the *next* coin toss is heads.
- This is given by the *posterior predictive* probability that $x = 1$:

$$
\begin{aligned}
P(x = 1|D, \alpha, \beta) &= \int P(x = 1|\theta) \overbrace{P(\theta|D, \alpha, \beta)}^{\text{Posterior}} \, d\theta, \\
&= \int \theta \times P(\theta|D, \alpha, \beta) \, d\theta, \\
&= \langle \theta \rangle, \\
&= \frac{\alpha + m}{\alpha + \beta + n}.
\end{aligned}
$$

- Thus, given 139 heads in 250 tosses, the predicted probability that the next coin will also be heads is ≈ 0.5556.

Marginal likelihood

- The posterior distribution is

$$P(\theta|D,\alpha,\beta) = \frac{\overbrace{P(D|\theta)}^{\text{Likelihood}}\ \overbrace{P(\theta|\alpha,\beta)}^{\text{Prior}}}{\underbrace{\int P(D|\theta)P(\theta|\alpha,\beta)\ d\theta}_{\text{Marginal likelihood}}}.$$

  where the *marginal likelihood* gives the likelihood of the model given the observed data:

$$\int P(D|\theta)P(\theta|\alpha,\beta)\ d\theta \stackrel{\text{def}}{=} P(D|\alpha,\beta).$$

- In this example, it has a simple analytical form:

$$P(D|\alpha,\beta) = B(\alpha+m,\beta+n-m) = \frac{\Gamma(m+\alpha)\Gamma(n-m+\beta)}{\Gamma(n+\alpha+\beta)}.$$

Model comparison

- Given D, we can compare the probability of model M_1 relative to model M_0 as follows:

$$\frac{P(M_1|D)}{P(M_0|D)} = \underbrace{\frac{P(D|M_1)}{P(D|M_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(M_1)}{P(M_0)}}_{\text{Priors odds}} .$$

- When both models are equally probable a priori, then the relative posterior probabilities is determined by the Bayes factor.
- We can compare our model M_1, i.e. with $\alpha = \beta = 1$, with the M_0 model that $\theta = \frac{1}{2}$.

$$\frac{P(D|M_1)}{P(D|M_0)} = \frac{\int P(D|\theta)P(\theta|\alpha = 1, \beta = 1)\ d\theta}{\int P(D|\theta)\delta(\theta - \frac{1}{2})\ d\theta.}$$

- This effectively compares a model that assumes a completely random coin machine, i.e., coin biases are generated according to $\text{Beta}(\alpha = 1, \beta = 1)$, to a completely perfect coin machine, i.e., coin biases are generated according to $\delta(\theta - \frac{1}{2})$.

Model comparison

- We can compare our model M_1, i.e. with $\alpha = \beta = 1$, with the M_0 model that $\theta = \frac{1}{2}$.

$$\frac{P(D|M_1)}{P(D|M_0)} = \frac{\int P(D|\theta)P(\theta|\alpha = 1, \beta = 1)\ d\theta}{\int P(D|\theta)\delta(\theta - \frac{1}{2})\ d\theta.},$$

$$= \frac{\Gamma(\alpha + m)\Gamma(\beta + n - m)}{\Gamma(\alpha + \beta + n)} \Big/ \frac{1}{2}^m (1 - \frac{1}{2})^{n-m},$$

$$= \frac{m!(n-m)!}{(n+1)!} \Big/ \frac{1}{2^n}.$$

If $n = 250$, $m = 139$, then

$$= \frac{139!111!}{251!} \Big/ \frac{1}{2^{250}} = 0.38.$$

- This is a factor of 2.65 in favour of the unbiased coin hypothesis.
- Note that the classical statistics null hypothesis test gives a p-value of $p = 0.0875$.

If n is a positive integer $\Gamma(n) = (n-1)!$.