

Introducing Bayesian Inference

Mark Andrews

Psychology, Nottingham Trent University

✉ mark.andrews@ntu.ac.uk, ✉

🐦 @xmjandrews,

March 25, 2018

Inference in a Bernoulli Distribution

e.g., inferring a coin's bias

... Polish mathematicians Tomasz Gliszczynski and Wacław 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, 2002¹

- ▶ A sample of $n = 250$ coin tosses can be modelled as n independent and identically distributed (iid) Bernoulli random variables with parameter θ , which represents the coin's bias.
- ▶ In other words, our *probabilistic generative model* of the data is

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

where $x_i \in \{0, 1\}$ represents the outcome of the coin flip i .

- ▶ Our aim is to infer the probable values of θ given an observation of $m = 139$ (or $m = 140$, etc.).

¹See <http://bit.ly/1B0Ku9b> for original story and <http://bit.ly/1B0Kx4Q> for discussion.

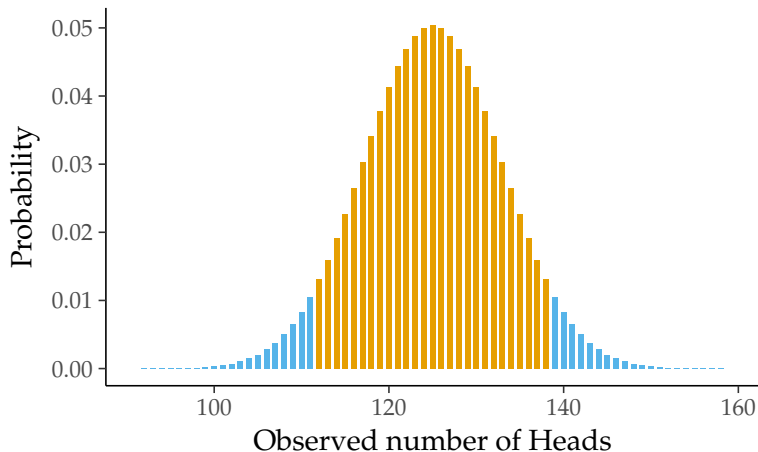
Sampling theory based inference

In a nutshell

- ▶ The traditional sampling theory based approach to inference begins by considering the *sampling distribution* of outcomes according to some hypothetical value of θ , which we signify with θ_0 .
- ▶ In our case, the sampling distribution is a binomial distribution with parameters $n = 250$ and $\theta = \theta_0$.
- ▶ We then calculate the probability of observing data *as or more extreme* than the result we obtained, i.e. $m = 139$ Heads.
- ▶ This probability is exactly the p-value, and effectively tells us how typical or unsurprising an outcome of $m = 139$ Heads would be if the true value of θ was θ_0 .
- ▶ Furthermore, the set of values of θ_0 that we would not reject at the α level of significance is exactly the $1 - \alpha$ *confidence interval*.

Sampling theory based inference

A binomial distribution with $n = 250$ and $\theta = 0.5$.



- Areas shaded in blue are values *as or more extreme* than $m = 139$.
- For the interactive version:
https://lawsofthought.shinyapps.io/binomial_test/.

Inference using the likelihood function

- ▶ The likelihood function is a function over the parameter space, i.e. over θ .
- ▶ It gives the probability² of the observed data, i.e. $m = 139$ Heads in $n = 250$ coin flips, for all possible values of θ .
- ▶ As such, it tells us how well any possible value of θ predicts the observed data.
- ▶ We usually denote the likelihood function as follows:

$$L(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta),$$

where \mathcal{D} signifies the observed data, i.e. $\mathcal{D} = \{n = 250, m = 139\}$.

²Strictly speaking, it gives the probability times an arbitrary positive constant c .

The likelihood function versus sampling distribution

- The sampling distribution is the binomial distribution:

$$P(m|\theta, n) = \binom{n}{m} \theta^m (1 - \theta)^{n-m}.$$

This is always a function of m , with θ and n fixed.

- The corresponding likelihood function treats the same function

$$P(m|\theta, n) = \binom{n}{m} \theta^m (1 - \theta)^{n-m},$$

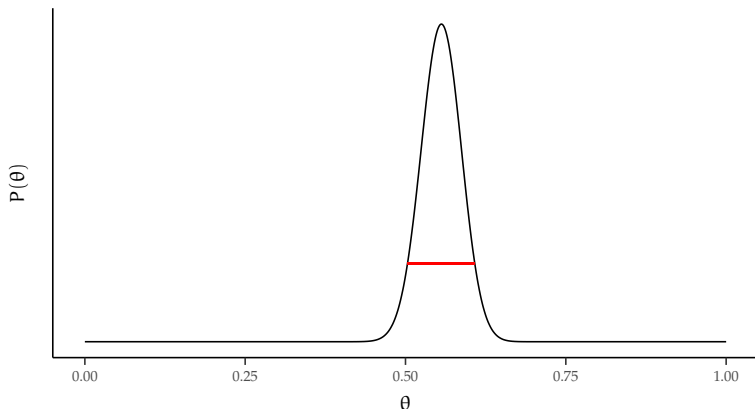
as a function of θ , but now with m and n fixed.

- As such, the binomial likelihood is

$$L(\theta|\mathcal{D}) \propto \theta^m (1 - \theta)^{n-m}.$$

Binomial likelihood

Given $n = 250$ and $m = 139$



- ▶ The red interval gives the values of θ that are more than $\frac{1}{8}$ the value of the maximum likelihood estimator.
- ▶ For the interactive version:
https://lawsofthought.shinyapps.io/binomial_likelihood/.

The likelihood principle

Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data.

Edwards *Likelihood* (1972, p.30)

- ▶ In any statistical model, all information provided by the observed data that are relevant to the parameters is given by the likelihood function.
- ▶ In our case, all information in the observed data concerning the true value of θ is given by $\theta^m(1 - \theta)^{n-m}$.

The law of the likelihood

Within the framework of a statistical model, a particular set of data supports one statistical hypothesis better than another if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis.

Edwards *Likelihood* (1972, p.30)

- In our case, how well the observed data support the hypothesis that $\theta = \theta_1$ versus $\theta = \theta_0$ is given by

$$\frac{L(\theta_1|\mathcal{D})}{L(\theta_0|\mathcal{D})},$$

and the data support the hypothesis $\theta = \theta_1$ more than $\theta = \theta_0$ if

$$\frac{L(\theta_1|\mathcal{D})}{L(\theta_0|\mathcal{D})} > 1.0.$$

The likelihood interval

- ▶ The likelihood interval gives the set of possible values of θ for which the data provide non-trivial evidence.
- ▶ The $\gamma \in (0, 1)$ likelihood interval is

$$\left\{ \theta: \frac{L(\theta|\mathcal{D})}{L(\theta_{\text{MLE}}|\mathcal{D})} > \gamma \right\},$$

where θ_{MLE} is the maximum likelihood estimator.

- ▶ All values within the γ likelihood interval have evidential support that is greater than γ times the support of θ_{MLE} .
- ▶ Typical values of γ are $\frac{1}{8}$, $\frac{1}{32}$, etc.

Bayesian inference of a coin's bias

- ▶ 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 $P(\theta)$:

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

- ▶ In other words, we assume that a value for θ was randomly drawn from $P(\theta)$ and then n binary variables were sampled from $\text{Bernoulli}(\theta)$.
- ▶ We refer to $P(\theta)$ as the *prior* distribution on θ .

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 parametric 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

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

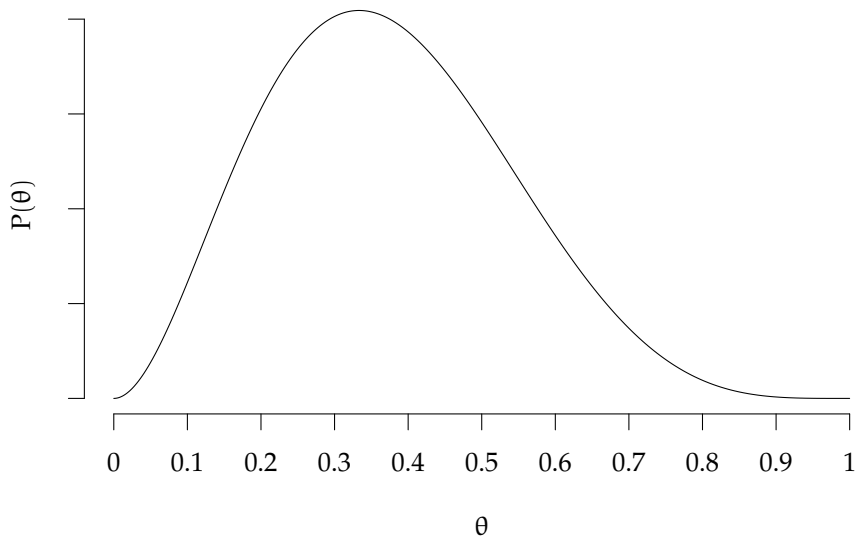
- The *normalizing constant* term is

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

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

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

where the normalizing constant is the reciprocal of the beta function

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

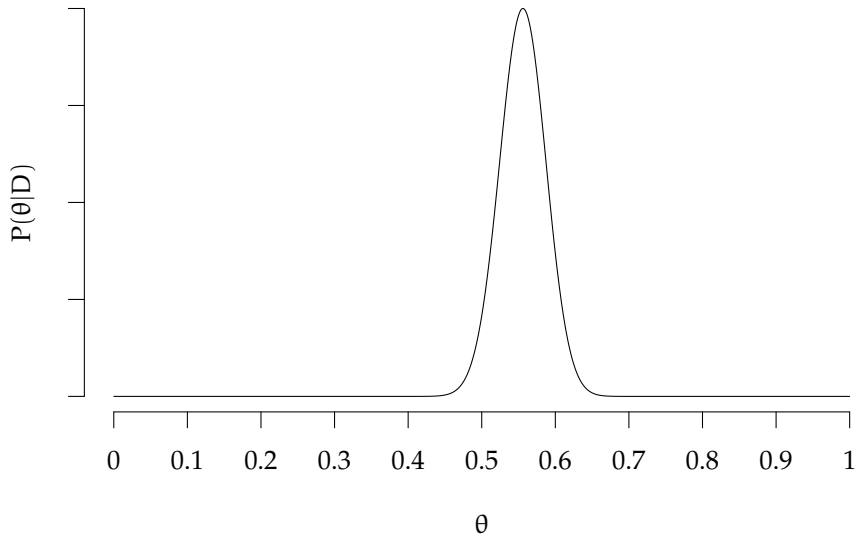
Posterior distribution

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

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

Posterior distribution

when $n = 250$, $m = 139$, $\alpha = 1$ and $\beta = 1$



For an interactive version:

https://lawsofthought.shinyapps.io/bayesian_coin_inference/

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

$$\langle \theta \rangle = 0.5556,$$

$$V(\theta) = 0.001, \quad \text{sd}(\theta) = 0.0312,$$

$$\text{mode}(\theta) = 0.556.$$

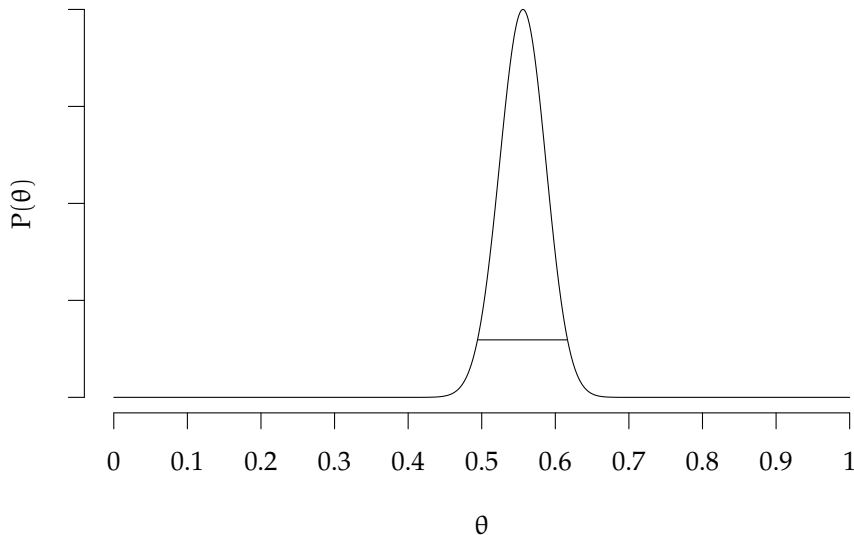
High posterior density (HPD) intervals

- ▶ 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

$$P(\{x: P(x) \geq p^*\}) = \varphi.$$

- ▶ In other words, we find the value p^* such that the probability mass of the set of points whose density is greater than 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 with Bayes factors

- ▶ 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 $\theta = \frac{1}{2}$.

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

Model comparison with Bayes factors

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

$$\begin{aligned}\frac{P(D|M_1)}{P(D|M_0)} &= \frac{\int P(D|\theta)P(\theta|\alpha=1, \beta=1) d\theta}{\int P(D|\theta)\delta(\theta - \frac{1}{2}) d\theta}, \\ &= \frac{\Gamma(\alpha+m)\Gamma(\beta+n-m)}{\Gamma(\alpha+\beta+n)} \Big/ \frac{1}{2}^m (1 - \frac{1}{2})^{n-m}, \\ &= \frac{m!(n-m)!}{(n+1)!} \Big/ \frac{1}{2^n}.\end{aligned}$$

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 $p = 0.0875$.