

Bayesian Inference in Single Parameter Models

Mark Andrews

December 3, 2018

Inference in a Bernoulli Distribution

...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 Bernoulli random variables with parameter θ .
- ▶ In other words, our probabilistic model is

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

and we would like 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.

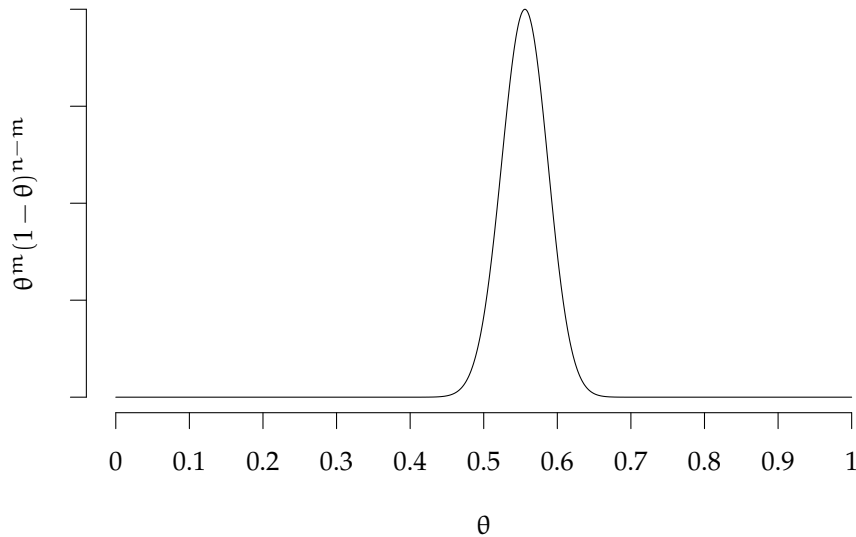
Inference in a Bernoulli Distribution

- 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 $x_1, x_2 \dots x_n$ 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 $\text{Bernoulli}(\theta)$.

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

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

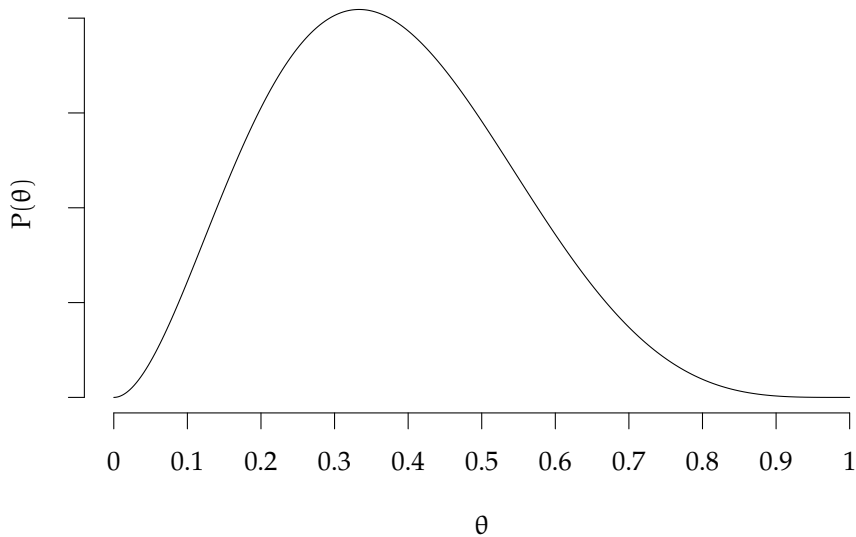
- The *normalizing constant* term is

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

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

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

where the normalizing constant is the reciprocal of the beta function

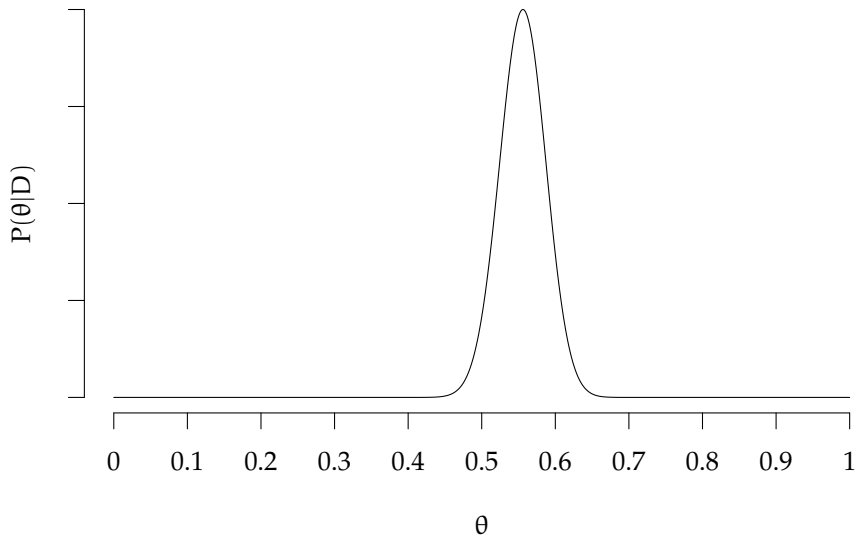
$$\begin{aligned} \frac{\Gamma(m + \alpha)\Gamma(n - m + \beta)}{\Gamma(n + \alpha + \beta)} &= \int \theta^{\alpha+m-1}(1-\theta)^{\beta+n-m-1} d\theta. \\ &= B(\alpha + m, \beta + n - m). \end{aligned}$$

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



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

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

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

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

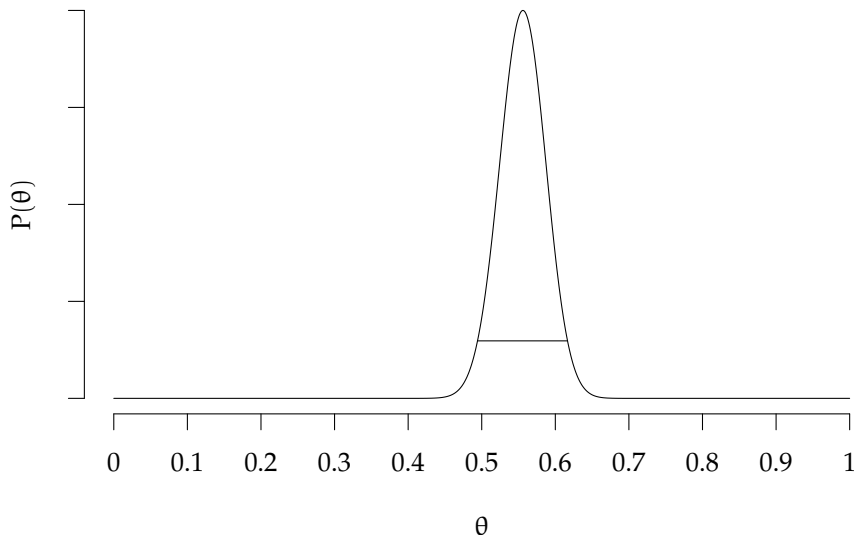
High posterior density (HPD) intervals

- ▶ 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

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

- ▶ 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

- ▶ 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 $\theta = \frac{1}{2}$.

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

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 $p = 0.0875$.

Inference of the Upper Bound of a Uniform Distribution

- ▶ Given a sample of K values from a uniform distribution, bounded between 0 and λ , i.e.

$$x_1, x_2 \dots x_K \sim \text{Uniform}(0, \lambda),$$

what are the probable values of λ ?

- ▶ This is related to the *German tank problem* whereby the total number of German tanks being produced during World War II was estimated from the serial numbers of captured tanks.
- ▶ In that problem, the tank serial numbers could be treated like random samples (without replacement) from a uniform distribution over the integers from 0 to T , where T is the total number of tanks.

Likelihood function of the Upper Bound

- The likelihood function over λ conditional on observations $x_1, x_2 \dots x_K$ is

$$P(x_1, x_2 \dots x_K | \lambda) = \prod_{k=1}^K P(x_k | \lambda)$$

where

$$P(x_k | \lambda) = \begin{cases} \frac{1}{\lambda}, & \text{for } \lambda \geq x_k, \\ 0, & \text{otherwise.} \end{cases}$$

- Therefore,

$$P(x_1, x_2 \dots x_K | \lambda) = \begin{cases} \frac{1}{\lambda}, & \text{for } \lambda \geq \max(x_{1:K}), \\ 0, & \text{otherwise.} \end{cases}$$

Posterior of the Upper Bound

- ▶ Assuming a uniform (improper) prior on λ , the posterior is obtained by normalizing the likelihood function.
- ▶ This leads to a posterior distribution

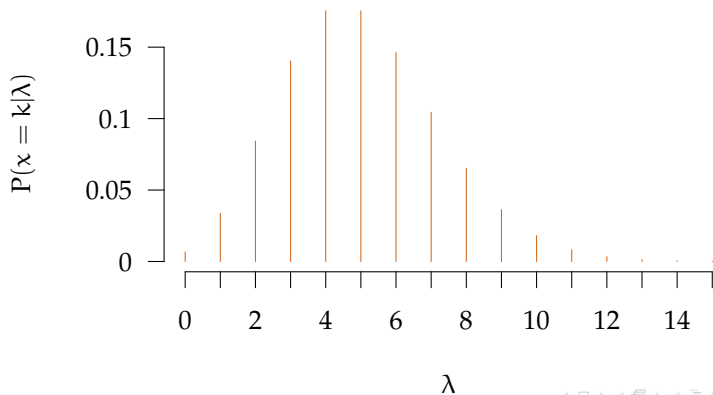
$$P(\lambda|x_1, x_2 \dots x_K) = \frac{\alpha x_0^\alpha}{\lambda^{\alpha+1}},$$

which is a *Pareto* distribution with shape parameter $\alpha = K - 1$ and scale parameter $x_0 = \max(x_{1:K})$.

Inference in Poisson models

- ▶ The Poisson distribution can be used to model the rare occurrences in fixed intervals.
- ▶ If x is a Poisson random variable with rate λ , then

$$P(x = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$



Inference in Poisson models

- ▶ Let's say that the number of emails I get per hour, in $n = 10$ hours, is

$$D = 6, 8, 12, 7, 11, 14, 10, 12, 11, 16.$$

- ▶ We can assume that these frequencies are generated according to a Poisson process with rate λ per hour.
- ▶ Given this assumption and this observed data, what is the probable value of λ ?
- ▶ In other words, what is

$$P(\lambda|D)?$$

Likelihood of a Poisson model

- Given a known value for λ , the probability of $x_1, x_2 \cdots x_n$ is

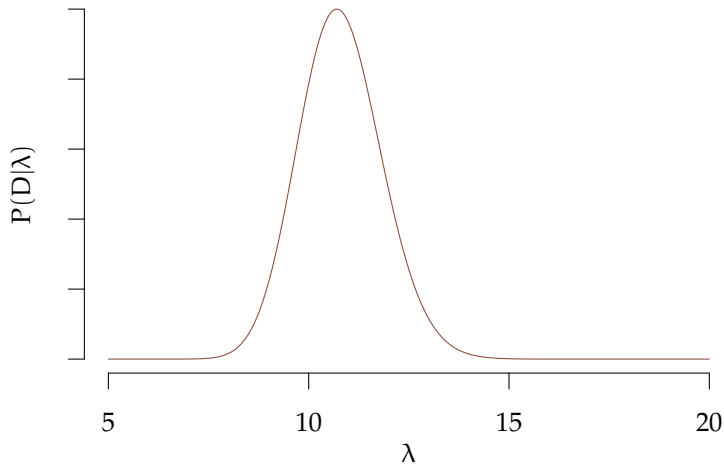
$$\begin{aligned} P(x_1, x_2 \cdots x_n | \lambda) &= \prod_{i=1}^n P(x_i | \lambda), \\ &= \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \\ &\propto e^{-n\lambda} \prod_{i=1}^n \lambda^{x_i} = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}. \end{aligned}$$

When $D = x_1, x_2 \cdots x_n = 6, 8 \cdots 16$, the likelihood of λ is

$$P(D | \lambda) = e^{-n\lambda} \lambda^S,$$

where $S = \sum_{i=1}^n x_i = 107$.

Likelihood of a Poisson model

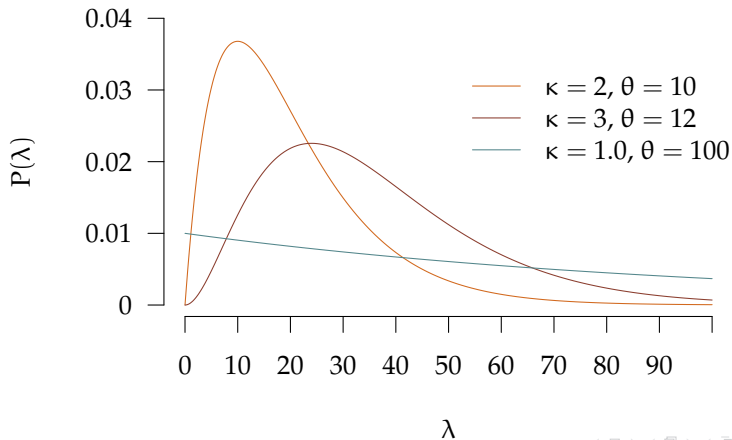


The likelihood of λ given the sufficient statistics $S = 107$ and $n = 10$.

Conjugate prior for the Poisson model

- The Gamma distribution with shape κ and scale θ is a conjugate prior for the Poisson model:

$$\text{Gamma}(\lambda|\kappa, \theta) = \frac{\lambda^{\kappa-1} e^{-\lambda/\theta}}{\theta^{\kappa} \Gamma(\kappa)}.$$



Posterior distribution

- With the Poisson likelihood, the Gamma prior leads to

$$\begin{aligned} P(\lambda|D, \kappa, \theta) &\propto e^{-n\lambda} \lambda^S \times \frac{\lambda^{\kappa-1} e^{-\lambda/\theta}}{\theta^{\kappa} \Gamma(\kappa)}, \\ &\propto e^{-\lambda\left(n + \frac{1}{\theta}\right)} \lambda^{S+\kappa-1}. \end{aligned}$$

Given that

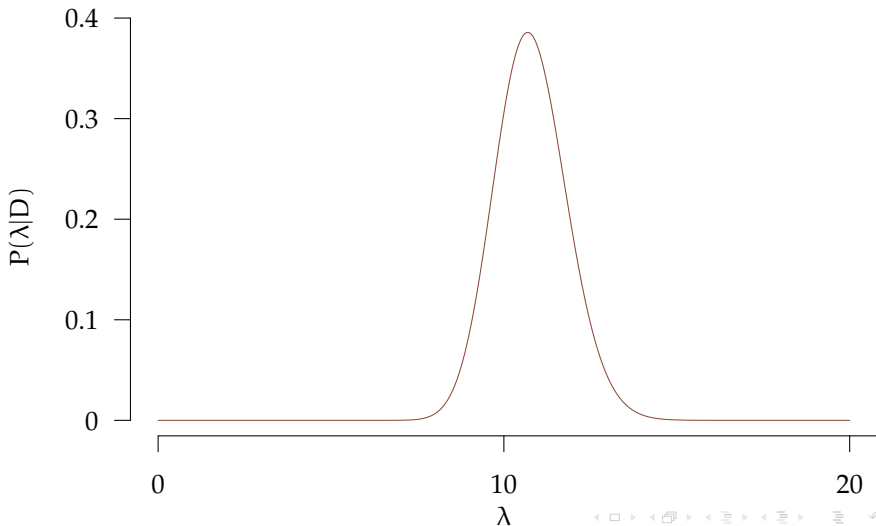
$$\int e^{-\lambda\left(n + \frac{1}{\theta}\right)} \lambda^{S+\kappa-1} d\lambda = \left(n + \frac{1}{\theta}\right)^{-(S+\kappa)} \Gamma(S+\kappa),$$

we have

$$P(\lambda|D, \kappa, \theta) = \text{Gamma}(\lambda|S + \kappa, (n + \frac{1}{\theta})^{-1}).$$

Posterior distribution

- With prior hyper-parameters of $\kappa = 1.0$, $\theta = 100.0$, and sufficient statistics of $S = 107$ and $n = 10$, we have Gamma distribution with shape 108 and scale ≈ 0.1



Summarizing the posterior

- The mean, variance and modes of any Gamma distribution with shape κ and scale θ are as follows:

$$\langle \lambda \rangle = \kappa \theta,$$

$$V(\lambda) = \kappa \theta^2,$$

$$\text{mode}(\lambda) = (\kappa - 1)\theta.$$

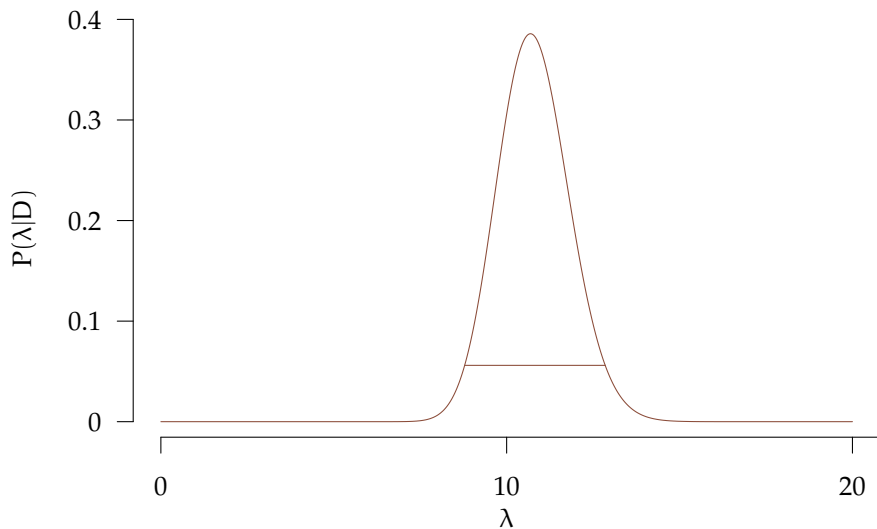
- Thus in our case, we have

$$\langle \lambda \rangle = 108.0,$$

$$V(\lambda) = 1.08, \quad \text{sd}(\theta) = 1.04,$$

$$\text{mode}(\lambda) = 10.69.$$

The 0.95 HPD interval



The 0.95 HPD interval is from 8.78 to 12.86.

Posterior predictive distribution

- Given that the number of emails every hour were

$$D = 6, 8, 12, 7, 11, 14, 10, 12, 11, 16,$$

what do we predict will be number of emails in the next hour?

- This is given by the posterior predictive distribution:

$$\begin{aligned} P(x_{\text{next}} = k | D, \kappa, \theta) &= \int P(x_{\text{next}} = k | \lambda) P(\lambda | D, \kappa, \theta) d\lambda, \\ &= \int \frac{e^{-k} \lambda^k}{k!} \times \frac{(n + \frac{1}{\theta})^{S+\kappa}}{\Gamma(S+\kappa)} e^{-\lambda(n + \frac{1}{\theta})} \lambda^{S+\kappa-1}, \\ &= \frac{\Gamma(S+\kappa+k)}{\Gamma(S+\kappa)\Gamma(k+1)} q^k (1-q)^{S+\kappa}, \\ &= \text{NegativeBinomial}(k | S+\kappa, q), \end{aligned}$$

with $q = (n + \frac{1}{\theta} + 1)^{-1}$.

Posterior predictive distribution

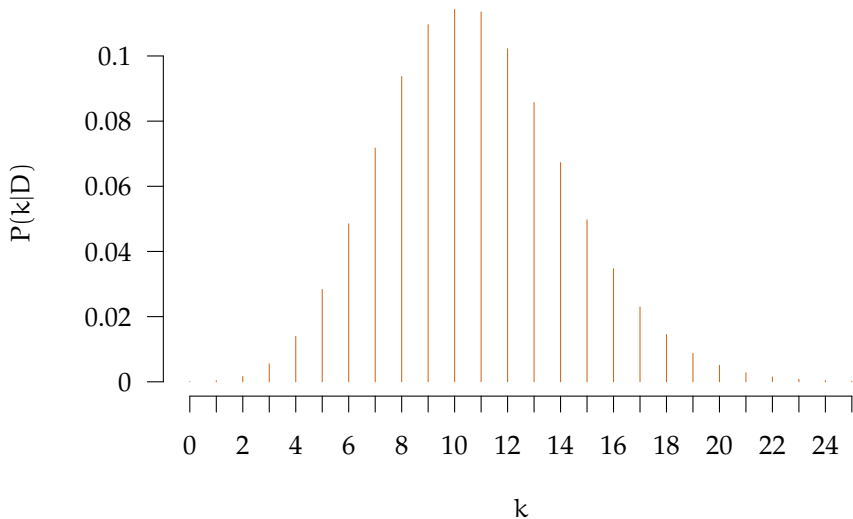
- ▶ The negative binomial distribution gives the number of “successes” until a predefined number r of “failures” have occurred, with the probability of a success being q .
- ▶ The mean and variance of the negative binomial are

$$\langle k \rangle = \frac{qr}{1-q}, \quad V(k) = \frac{qr}{(1-q)^2}.$$

- ▶ In our case, $r = S + \kappa = 108$ and $q = (n + \frac{1}{\theta} + 1)^{-1} = (10 + \frac{1}{100} + 1)^{-1} = 0.091$.
- ▶ The mean and variance of the negative binomial are

$$\langle k \rangle = \frac{S + \kappa}{n + \frac{1}{\theta}} = 10.79, \quad V(k) = \frac{S + \kappa}{n + \frac{1}{\theta}} (n + \frac{1}{\theta} + 1) = 11.87.$$

Posterior predictive distribution



The posterior predictive distribution $P(x_{\text{next}} = k|D, \kappa, \theta)$ is a negative binomial distribution with $r = 108$, $q = 0.091$.