Parametric models. A family of distributions $f(x \mid \theta)$ and prior distribution $\pi(\theta)$ for ϑ. The posterior distribution of ϑ at $\vartheta = \theta$, given x, is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)d\theta}$$

Thus

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta).$$

The same form for θ continuous ($\pi(\theta \mid x)$ a pdf) or discrete ($\pi(\theta \mid x)$ a pmf) posterior $\propto$ prior $\times$ likelihood

Likelihood principle Notice that, if we base all inference on the posterior distribution, then we respect the likelihood principle. If two likelihood functions are proportional, then any constant cancels top and bottom in Bayes rule, and the two posterior distributions are the same.

Example 2. $X \sim \text{Bin}(n, \vartheta)$ for known n and unknown ϑ. Suppose our prior knowledge about ϑ is represented by a Beta distribution on $(0, 1)$, and θ is a trial value for ϑ.

Prior probability density

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)},\ 0 < \theta < 1.$$

Likelihood

$$f(x \mid \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0, \ldots, n$$

Posterior probability density

$$\pi(\theta \mid x) \propto \theta^{a+x-1}(1-\theta)^{n-x+b-1}$$

Here posterior has the same form as the prior with updated parameters a, b replaced by $a + x, b + n - x$, so

$$\pi(\theta \mid x) = \frac{\theta^{a+x-1}(1-\theta)^{n-x+b-1}}{B(a+x, b+n-x)}$$

For a Beta distribution with parameters a, b

$$\mu = \frac{a}{a+b}, \; \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

The posterior mean and variance are

$$\frac{a+X}{a+b+n}, \; \frac{(a+X)(b+n-X)}{(a+b+n)^2(a+b+n+1)}$$

For large n, the posterior mean and variance are approximately

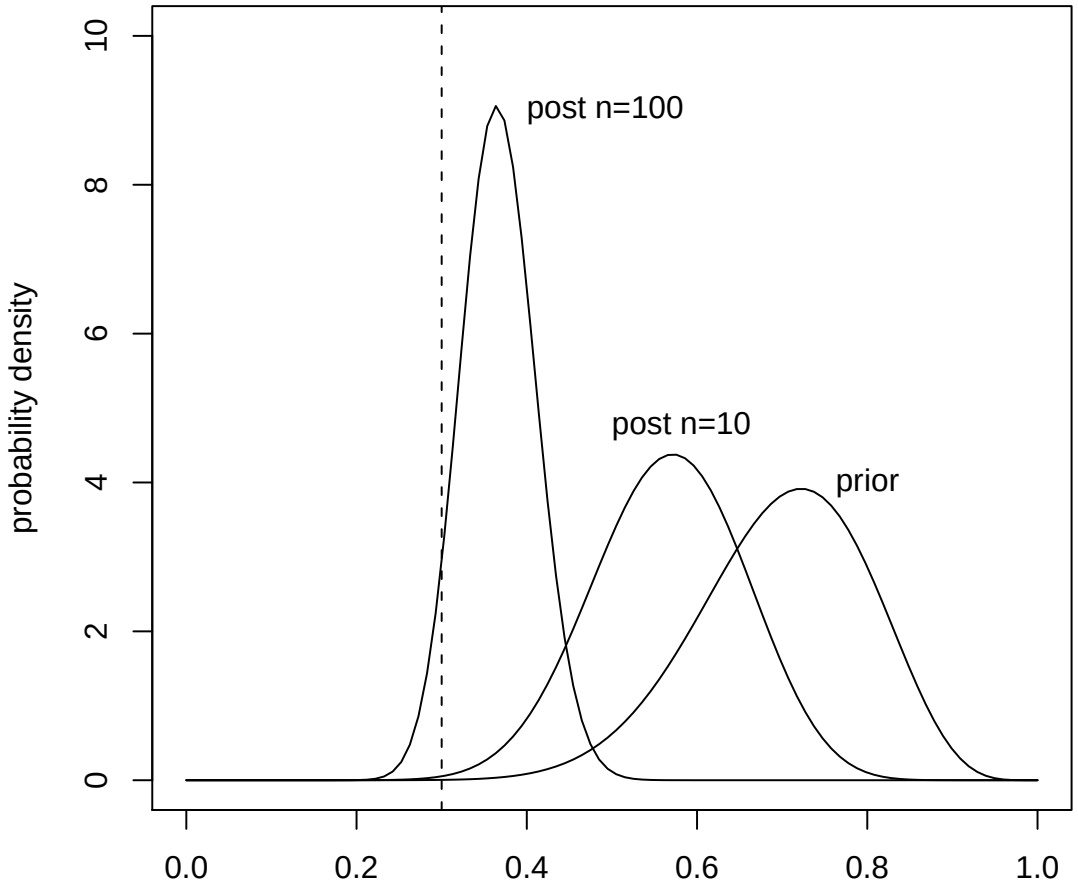
$$\frac{X}{n}, \; \frac{X(n-X)}{n^3}$$

In classical statistics

$$\widehat{\theta} = \frac{X}{n}, \; \frac{\widehat{\theta}(1-\widehat{\theta})}{n} = \frac{X(n-X)}{n^3}$$

As sample size n in Binomial increases, information in data overwhelms information in prior.

Example. Suppose $\vartheta = 0.3$ but prior mean is 0.7 with std 0.1. Suppose data $X \sim \text{Bin}(n, \vartheta)$ with $n = 10$ (yielding $X = 3$) and then $n = 100$ (yielding $X = 30$, say).



Asymptotic Normality of Posterior Distribution

Suppose the true parameter value is $\vartheta = \theta_0$. Under regularity conditions (likelihood $L(\theta; x)$ three times differentiable wrt θ, prior $\pi(\theta)$ continuous in a neighborhood of θ_0) the posterior converges to a normal density with mean θ_0.

In the iid case $X = (X_1, ..., X_n)$, with $X_i \sim f(x_i; \theta_0)$ and $X \in \chi^n$. The posterior density for real scalar $\theta \in R$ is

$$\pi(\theta|x) \propto \exp(n\overline{\ell}_n + \log(\pi(\theta)))$$

with $\overline{\ell}_n = n^{-1} \sum_{i=1}^n \ell(\theta; x_i)$. Let $\overline{\ell}(\theta) = \mathbb{E}_{\theta_0} \ell(\theta; X_1)$,

$$\overline{\ell}(\theta) = \int_\chi \ell(\theta; y) f(y; \theta_0) dy.$$

We have $\overline{\ell}_n \to \overline{\ell}(\theta)$ by the Law of Large Numbers. Also, $\partial\overline{\ell}/\partial\theta = 0$ when $\theta = \theta_0$ (by differentiation under the integral). Expand $\overline{\ell}(\theta)$ in a Taylor series about θ_0:

$$\overline{\ell}(\theta) = \overline{\ell}(\theta_0) + (\theta - \theta_0)^2 \frac{1}{2} \left.\frac{\partial^2 \overline{\ell}}{\partial \theta^2}\right|_{\theta=\theta_0} + O[(\theta - \theta_0)^3]$$

For θ close to θ_0 (ie, $|\theta - \theta_0| = O(1/\sqrt{n})$),

$$n\overline{\ell}_n + \log(\pi(\theta)) \to n\overline{\ell}(\theta_0) - (\theta - \theta_0)^2/(2v/n) + O(1/\sqrt{n})$$

where $v = \left[-\left.\frac{\partial^2 \overline{\ell}}{\partial \theta^2}\right|_{\theta=\theta_0} \right]^{-1}$. The posterior close to θ_0 is normal, $N(\theta_0, v/n)$. However the standard deviation is $O(1/\sqrt{n})$ so all states of non-negligible probability density are 'close' to θ_0.

Exercise Show that v is non-negative and $v = I_{\theta_0}$.

Conjugate Priors

In Example 2. above the parametric form of the posterior is like the prior. $f(x \mid \theta)$ and $\pi(\theta)$ form a conjugate prior family.

A family $\mathcal{P}$ of prior distributions for θ is closed under sampling from a model $f(x \mid \theta)$ if for every prior distribution $\pi(\theta) \in \mathcal{P}$, the posterior distribution

$$\pi(\theta \mid x) \propto \pi(\theta) f(x \mid \theta)$$

is also in $\mathcal{P}$. This is also then true conditional on a sample of n.

Example 3. Normal with a known variance and a prior normal distribution for θ.
Example 4. Normal with a unknown mean and variance, a normal prior for θ and a Gamma distribution for $1/\sigma^2$.

Example 3. $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$, where θ is unknown and σ^2 is known. The prior distribution for θ is $N(\mu_0, \sigma_0^2)$, with μ_0, σ_0^2 known. The posterior is proportional to

$$\pi(\theta) f(x \mid \theta) \propto \exp\left\{ -\frac{(\theta - \mu_0)^2}{2\sigma_0^2} - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2\sigma^2} \right\}$$

Complete the square:

$$\begin{aligned} \frac{(\theta - \mu_0)^2}{\sigma_0^2} + \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{\sigma^2} &= \theta^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\theta \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2} \right) \\ &= \frac{1}{\sigma_1^2}(\theta - \mu_1)^2 + C_2, \end{aligned}$$

where

$$\sigma_1^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$$\mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$$

$\pi(\theta \mid x)$ is $N(\mu_1, \sigma_1^2)$. As $n \to \infty$ and $\mu_1 \approx \bar{x}$, $\sigma_1^2 \approx \frac{\sigma^2}{n}$

Example 4 Normal distribution when the mean and variance are unknown. Let $\tau = 1/\sigma^2$, $\theta = (\tau, \mu)$. The prior τ has a Gamma distribution with parameters $\alpha, \beta > 0$, and conditional on τ, μ has a distribution $N(\nu, \frac{1}{k\tau})$ for some $k > 0$, $\nu \in \mathbb{R}$.

The prior is

$$\pi(\tau, \mu) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \cdot (2\pi)^{-1/2} (k\tau)^{1/2} \exp\left\{-\frac{k\tau}{2}(\mu - \nu)^2\right\}$$

or

$$\pi(\tau, \mu) \propto \tau^{\alpha-1/2} \exp\left[-\tau\left\{\beta + \frac{k}{2}(\mu - \nu)^2\right\}\right]$$

The likelihood is

$$f(x \mid \mu, \tau) = (2\pi)^{-n/2} \tau^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

Thus

$$\pi(\tau, \mu \mid x) \propto \tau^{(\alpha+n)/2-1/2} \exp\left[-\tau\left\{\beta + \frac{k}{2}(\mu-\nu)^2 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\right]$$

Complete the square to see that

$$\begin{aligned} & k(\mu-\nu)^2 + \sum(x_i-\mu)^2 \\ & \quad = (k+n)\left(\mu - \frac{k\nu + n\bar{x}}{k+n}\right)^2 + \frac{nk}{n+k}(\bar{x}-\nu)^2 + \sum(x_i-\bar{x})^2 \end{aligned}$$

Thus the posterior is

$$\pi(\tau, \mu \mid x) \propto \tau^{\alpha'-1/2} \exp\left[-\tau\left\{\beta' + \frac{k'}{2}(\nu'-\mu)^2\right\}\right]$$

where

$$\begin{aligned}
\alpha' &= \alpha + \frac{n}{2} \\
\beta' &= \beta + \frac{1}{2} \cdot \frac{nk}{n+k}(\bar{x} - \nu)^2 + \frac{1}{2}\sum(x_i - \bar{x})^2 \\
k' &= k + n \\
\nu' &= \frac{k\nu + n\bar{x}}{k+n}
\end{aligned}$$

This is the same form as the prior, so the class is conjugate prior.

Marginal posterior distribution and Nuisance parameters If θ is multivariate $\theta = (\theta_1, ...\theta_k)$, $\theta \in \Theta_1 \times \Theta_2 \times ...\Theta_k$ with posterior density $\pi(\theta|x)$, we may be interested in the marginal prior or posterior distribution for θ_1 say,

$$\pi(\theta_1) = \int_{\Theta_2 \times ...\Theta_k} \pi(\theta_1, \theta_2, ..., \theta_k) d\theta_2 d\theta_3 .... d\theta_k$$

$$\pi(\theta_1|x) = \int_{\Theta_2 \times ...\Theta_k} \pi(\theta_1, \theta_2, ..., \theta_k|x) d\theta_2 d\theta_3 .... d\theta_k.$$

Nuisance parameters $\theta = (\nu_1, \nu_2)$. ν_2 nuisance parameters (needed to define likelihood but not otherwise of interest). Given the data integrate out ν_2 in the posterior distribution

$$\pi(\nu_1 \mid x) = \int \pi(\nu_1, \nu_2 \mid x) d\nu_2$$

Example 4 (cont). If $\tau = 1/\sigma^2$ is a nuisance parameter we need the prior and posterior of μ alone

$$\begin{aligned}
\pi(\mu) &= \int_0^\infty \pi(\tau, \mu) d\tau \\
&\propto \int_0^\infty \tau^{\alpha - 1/2} \exp\left[-\tau\left\{\beta + \frac{k}{2}(\nu-\mu)^2\right\}\right] d\tau \\
&= \frac{\Gamma(\alpha + 1/2)}{\left(\beta + \frac{k}{2}(\nu-\mu)^2\right)^{\alpha+1/2}} \qquad \propto \qquad \frac{1}{(1 + Y^2/r)^{\frac{r+1}{2}}}
\end{aligned}$$

if $r = 2\alpha$ and $Y = \sqrt{\alpha k/\beta}(\nu - \mu)$, ie, proportional to the density of a Student's t-distribution with r dof, $t(r)$. It follows that

$$\sqrt{\frac{\alpha k}{\beta}}(\nu - \mu) \sim t(2\alpha)$$

marginally in the prior.

Since the prior is conjugate, the posterior is Gamma, so the calculations will be the same, and

$$\begin{aligned}\pi(\mu|x) &= \int_0^\infty \pi(\tau,\mu|x)d\tau \\ &\propto \frac{\Gamma(\alpha'+1/2)}{\left(\beta'+\frac{k'}{2}(\nu'-\mu)^2\right)^{\alpha'+1/2}}\end{aligned}$$

so that

$$\sqrt{\frac{\alpha' k'}{\beta'}}(\nu'-\mu) \sim t(2\alpha')$$

marginally in the posterior. This is convenient if we want to calculate the evidence for a hypothesis such as $a \leq \mu \leq b$.

Example 5 Conjugate prior for an exponential family

$$f(\mathbf{x} \mid \theta) = \exp\left\{\sum_{j=1}^{k} A_j(\theta) \sum_{i=1}^{n} B_j(x_i) + \sum_{i=1}^{n} C(x_i) + nD(\theta)\right\}$$

The prior distribution based on sufficient statistics

$$\pi(\theta) \propto \exp\{\tau_0 D(\theta) + \sum_{j=1}^{k} A_j(\theta)\tau_j\}$$

(where $(\tau_0, \ldots, \tau_k)$ are constant prior parameters) is conjugate.

The posterior density is proportional to

$$\begin{aligned} & f(x \mid \theta)\pi(\theta \mid \tau_0, \ldots, \tau_k) \\ & \propto \exp\left\{\sum_{j=1}^{k} A_j(\theta)\left[\sum_{i=1}^{n} B_j(x_i) + \tau_j\right] + (n + \tau_0)D(\theta)\right\} \end{aligned}$$

This is an updated form of the prior with

$$\begin{aligned} B'_j(x) &= \sum_{i=1}^{n} B_j(x_i) + \tau_j \\ n' &= n + \tau_0 \end{aligned}$$

Predictive distributions

$X_1, \ldots, X_n$ are observations from $f(x;\theta)$ and the predictive distribution of a further observation X_{n+1} is required. $x = (x_1, \ldots, x_n)$. The posterior predictive distribution is

$$g(x_{n+1} \mid x) = \int f(x_{n+1};\theta)\pi(\theta \mid x)d\theta$$

Predictive distributions are useful for ... prediction.

They are used also for model checking. Divide the data in two groups, $Y = (X_1, ..., X_a)$ and $Z = (X_{a+1}, ..., X_n)$. If we fit using Y and check that the 'reserved data' Z overlap $g(x_{n+1} \mid x)$ in distribution.

Example Data $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$ with σ^2 known and θ-prior $\vartheta \sim N(\mu_0, \sigma_0^2)$. Predict X_{n+1}.

We saw that $\pi(\theta \mid x) = N(\theta; \mu_1, \sigma_1^2)$ with μ_1 and σ_1 calculated in example 3 above. In order to calculate the posterior predictive density for X_{n+1} we need to evalaute

$$g(x_{n+1} \mid x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}} d\theta$$

We could complete the square. Alternatively, think how X_{n+1} is built up. If $Y, Z \sim N(0,1)$ then

$$\begin{aligned} \vartheta &= \mu_1 + \sigma_1 Z \qquad \text{(posterior)} \\ X_{n+1} &= \vartheta + \sigma Y \qquad \text{(observation model)} \\ &= \mu_1 + \sigma_1 Z + \sigma Y. \end{aligned}$$

It follows that $X_{n+1} \sim N(\mu_1, \sigma^2 + \sigma_1^2)$ is the posterior predictive density for $X_{n+1} | X_1, ..., X_n$.

Hierarchical models These are models where the prior has parameters which again have a probability distribution.

1. Data x have a density $f(x;\theta)$.
2. The prior distribution of θ is $\pi(\theta;\psi)$.
3. ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$.
4. We can work with posterior $\pi(\theta,\psi|x) \propto f(x;\theta)\pi(\theta;\psi)g(\psi)$, or

$$\begin{aligned}\pi(\theta|x) &= \int_\Psi \pi(\theta,\psi|x)d\psi \\ &\propto f(x;\theta)\int_\Psi \pi(\theta;\psi)g(\psi)d\psi \\ &\propto f(x;\theta)\pi(\theta)\end{aligned}$$

The hierarchical model simply specifies the prior for θ indirectly

$$\pi(\theta) = \int \pi(\theta;\psi)g(\psi)d\psi.$$

An example is when $\theta = (\theta_1, \ldots, \theta_k)$ and the parameters are each associated with a subpopulation and have an exchangeable distribution (the labels $i = 1, 2, ..., k$ can be permuted without changing the prior for θ).

Example For $i = 1, 2, ..., k$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\vartheta_i, \sigma^2)$. The ϑ_i are the unknown means for observations on the i'th population but σ^2 is known.

Suppose the prior model for the ϑ_i is iid normal, $\vartheta_i \sim N(\varphi, \tau^2)$.
If $\psi = (\varphi, \tau^2)$

$$\pi(\theta_1, \ldots, \theta_k; \psi) = \prod_{i=1}^{k} (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\tau^2}(\theta_i - \varphi)^2\right\},$$

Now we need a prior for φ and τ^2. For an uninformative prior choose independence so that $g(\varphi, \tau^2) = \text{cost}$, all possible φ, τ^2 equally likely *a priori*. [careful!]

The joint posterior of the parameters is

$$\pi(\theta, \psi|x) \propto f(x;\theta)\pi(\theta|\psi)g(\psi)$$

Integration with respect to ψ gives the posterior density of θ.

Here $g(\psi)$ is constant, so the joint posterior distribution is

$$\begin{aligned}\pi(\theta, \varphi, \tau^2|x) \;\propto\; & \left[\prod_{i=1}^{k} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(x_{ij}-\theta_i)^2\right\}\right] \\ \times\; & \left[\prod_{i=1}^{k} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_i-\varphi)^2\right\}\right]\end{aligned}$$

Integrate out wrt φ and τ^2 to obtain the marginal posterior distribution of θ. Integrating the last factor wrt φ gives a term proportional to

$$\tau^{1-k} \exp\left\{-\frac{1}{2\tau^2}\sum(\theta_i - \bar{\theta})^2\right\}$$

Then the integral wrt τ gives a term proportional to

$$\left[\sum(\theta_i - \bar{\theta})^2\right]^{1-k/2}$$

Thus the posterior distribution of θ is

$$\begin{aligned}
\pi(\theta|x) &\propto \left[\prod_{i=1}^{k} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(x_{ij} - \theta_i)^2\right\}\right] \cdot \left[\sum(\theta_i - \bar{\theta})^2\right]^{1-k/2} \\
&\propto \left[\prod_{i=1}^{k} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k} n_i(\theta_i - \bar{x}_i)^2\right\}\right] \cdot \left[\sum(\theta_i - \bar{\theta})^2\right]^{1-k/2}
\end{aligned}$$

where $\bar{x}_i = \sum_j x_{ij}/n_i$. Let $\widehat{\theta}_j$ be the MAP estimate for θ_j (posterior mode) and put $\widehat{\theta}^* = \sum \widehat{\theta}_j/k$. Differentiate $\pi(\theta|x)$ wrt θ_j and set to equal zero to get

$$\widehat{\theta}_j = (\omega_1 \bar{x}_j + \omega_2 \widehat{\theta}^*)/(\omega_1 + \omega_2),$$

where $\omega_1 = (\sigma^2/n_j)^{-1}$ and $\omega_2 = \left[\sum(\widehat{\theta}_i - \widehat{\theta}^*)^2/(k-2)\right]^{-1}$. If the θ_j were unrelated $\bar{x}_j$ would be the estimate of θ_j. The model modifies the estimate by pulling it towards the mean of the estimated θ_is.