Now, coming back to the Beta prior for Binomial, note that what simplied the computation of the posterior density is that the prior and likelihood have the same functional form.

**Conjugate families of prior distributions.** Let $\mathcal{F}$ denote a class of density functions $f(x|\theta)$. A class $\mathcal{P}$ of prior densities is said to be a conjugate family for $\mathcal{F}$ if $\pi(.|x) \in \mathcal{P}$ for all $f \in \mathcal{F}$ and $\pi \in \mathcal{P}$.

**Example.** $X|\theta \sim \text{Binomial}(n, \theta)$. Then

$$\mathcal{F} = \{ \text{ all Binomial}(n, \theta),\ n = 1, 2 \ldots \},$$

$$\mathcal{P} = \{ \text{ all Beta}(a, b),\ a > 0,\ b > 0. \ \}.$$

If $\pi \in \mathcal{P}$ and $f \in \mathcal{F}$, then $\theta \sim \text{Beta}(a, b)$ for some $a > 0$, $b > 0$, and $X|f \sim \text{Binomial}(n, \theta)$ for some $n > 0$, so $\theta|X = x \sim \text{Beta}(x+a, n-x+b) \in \mathcal{P}$.

**Example.** $X|\theta \sim N(\theta, \sigma^2)$, $\sigma^2$ known. Consider $\theta \sim N(\mu, \tau^2)$, $\mu, \tau^2$ known. Then, $\theta|X = x \sim ?$ $h(x, \theta) = \frac{1}{2\pi}\frac{1}{\sigma\tau}\exp(-\frac{1}{2}[\frac{(x-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2}])$. One can complete the square for $\theta$, proceed using calculus to find $m(x)$ and then determine $\pi(\theta|x)$. We will use a property of the multivariate normal instead. Note that $X|\theta \sim N(\theta, \sigma^2)$ is equivalent to $X = \theta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ independent of $\theta$. Since $\theta \sim N(\mu, \tau^2)$, we can obtain the joint bivariate normal distribution for $X$ and $\theta$ as:

$$\begin{pmatrix} X \\ \theta \end{pmatrix} = \begin{pmatrix} \theta + \epsilon \\ \theta \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix}\right),$$

because $E(X) = E(\theta + \epsilon) = \mu$, $Var(X) = Var(\theta + \epsilon) = \sigma^2 + \tau^2$, $Cov(X, \theta) = Cov(\theta + \epsilon, \theta) = Cov(\theta, \theta) = Var(\theta) = \tau^2$. Therefore,

$$\begin{aligned} \theta|X = x &\sim N\left(\mu + \frac{\tau^2}{\sigma^2 + \tau^2}(x - \mu), \tau^2 - \frac{\tau^4}{\sigma^2 + \tau^2}\right) \\ &= N\left(\frac{\tau^2}{\sigma^2 + \tau^2}x + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2} = \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right). \end{aligned}$$

**Remark.** If instead $X_1, X_2, \ldots, X_n$ are i.i.d $N(\theta, \sigma^2)$, $\sigma^2$ known, in the above example, then $\bar{X}$ is sufficient for $\theta$ and $\bar{X}|\theta \sim N(\theta, \sigma^2/n)$. Therefore replace $X$ by $\bar{X}$ and $\sigma^2$ by $\sigma^2/n$ above.

**Question.** Are conjugate priors reasonable for expressing prior information? If they represent actual prior information, there is no problem. Otherwise they are easy to work with but not robust – prior and likelihood have the same functional form, so similar weight is given to prior and sample data. Mixtures of conjugate priors are much better, and computations are not too difficult because MCMC sampling methods are available.

### Noninformative or vague priors

**Example.** $X_1, X_2, \ldots, X_n$ are i.i.d $N(\theta, \sigma^2)$, $\sigma^2$ known. Inference on $\theta$ is of interest. Consider $\pi(\theta) \equiv 1$ as an expression of lack of prior information. This is not a probability density, but that of a limit of $N(0, \tau^2)$ as $\tau^2 \to \infty$. Since $\bar{X}|\theta \sim N(\theta, \sigma^2/n)$, $h(\bar{x}, \theta) = f(\bar{x}|\theta)\pi(\theta) = f(\bar{x}|\theta)$, we get

$$m(\bar{x}) = \int_{-\infty}^{\infty} f(\bar{x}|\theta)\, d\theta = \int_{-\infty}^{\infty} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp(-\frac{n}{2\sigma^2}(\theta - \bar{x})^2)\, d\theta = 1.$$

Therefore,

$$\pi(\theta|\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp(-\frac{n}{2\sigma^2}(\theta - \bar{x})^2),$$

so that $\theta|\bar{X} = \bar{x} \sim N(\bar{x}, \sigma^2/n)$. Why not then use $\pi(\theta) \equiv c$ whenever no prior information is available, or when a noninformative prior is needed? Observe the problem with this approach. Suppose $\theta > 0$ and consider $\pi(\theta) \equiv c$. Then $\int_0^\infty \pi(\theta)\, d\theta = \infty$. Consider the reparametrization: $\eta = \eta(\theta) = \exp(\theta)$. Then $\theta = \log(\eta)$, so $d\theta = d\eta/\eta$. Therefore, the prior density on $\eta$ is given by

$$\pi^*(\eta) = \pi(\log(\eta))\frac{1}{\eta} = \frac{c}{\eta},$$

which is not uniform as is the case with $\pi(\theta)$. If we did not have information about $\theta$ how did we get information about a transform of it? There are no strictly noninformative priors, there are default and reference priors for objective choice, for example, the Jeffreys' prior.

### Jeffreys' Prior

The idea is to employ a prior which contains the minimal amount of prior information needed to be able to conduct Bayesian analysis for the given experiment. Let $f(x|\theta)$ be the model density of $X|\theta$ for which $I(\theta)$ be the Fisher Information. Then the Jeffreys' prior in this case is defined to be

$$\pi(\theta) = (I(\theta))^{1/2} \text{ if } \theta \text{ is univariate; more generally } \pi(\theta) = |I(\theta)|^{1/2}.$$

**Example.** Suppose $X|\theta \sim N(\theta, 1)$. Then $I(\theta) = 1$ since $\frac{\partial}{\partial\theta} \log f(x|\theta) = x - \theta$ and $\frac{\partial^2}{\partial\theta^2} \log f(x|\theta) = -1$. Therefore, $\pi(\theta) \equiv C$ is the Jeffreys' prior in this case.

It may be verified that this prior is invariant with respect to any one-one differentiable transformations on $\theta$.

In practice, however, Jeffreys' employed a different group invariance argument. For any location family, his suggestion for the prior on the location parameter is the translation invariant (thus indicating lack of information) measure. Note that this agrees with the Jeffreys' prior above for the $N(\theta, 1)$. For a location-scale family $f(x|\theta, \sigma)$, his argument is as follows. If $\sigma$ is fixed, then the prior for the location $\theta$ is $\pi_1(\theta|\sigma) = 1$ as above. Now note that, if $\sigma$ is a scale parameter for a positive r.v. $Y$ then $\log\sigma$ is a location for $\log Y$. So, the prior for $\log\sigma$ is a constant which in turn gives $\pi_2(\sigma) = 1/\sigma$. Thus $\pi(\theta, \sigma) = \pi_1(\theta|\sigma)\pi_2(\sigma) = 1/\sigma$. This is the right invariant Haar measure for the affine group of transaformations whereas the Jeffreys' prior according to the formal definition above is the left invariant Haar measure which is $1/\sigma^2$. Most Bayesians prefer the former one, for various reasons including posterior consistency.

## Estimation

$\pi(\theta|$ data) is the probability density of $\theta$ having seen the data. It contains all the post-experimental information about $\theta$. Any Bayesian inference about $\theta$ must be based on it. Note the following in this context.
(i) We have an actual probability distribution on the unknown *parameters* to describe their uncertainty, namely $\pi(\theta|$ data).
(ii) We can readily make probability statements on where $\theta$ lies using this distribution.

A Bayesian can simply report the posterior distribution, or report some summary descriptive measures associated with the posterior distribution. For example, as mentioned previously, $\hat{\theta}_{\text{hpd}}$, is one such measure which is analogous to the MLE. If $\pi(\theta|x)$ is unimodal, this may be a reasonable estimate for $\theta$. However, the usual Bayes estimate of $\theta$ is $E(\theta|x)$, which is a measure of location or centre of $\pi(\theta|x)$. For this estimate, the precision may be measured by the posterior standard deviation, $s.d.(\theta|x)$, which is a standard measure of spread or dispersion. Note that (for a real valued $\theta$)

$$E(\theta|x) = \int_{-\infty}^{\infty} \theta\pi(\theta|x)\, d\theta$$

and the posterior variance

$$\begin{aligned} Var(\theta|x) &= E\{(\theta - E(\theta|x))^2|x\} \\ &= \int_{-\infty}^{\infty} (\theta - E(\theta|x))^2\pi(\theta|x)\, d\theta. \end{aligned}$$

Consider the binomial example again: $X|\theta \sim \text{Binomial}(n,p)$, for which the prior is $p \sim \text{Beta}(\alpha,\beta)$. Then $p|X = x \sim \text{Beta}(x+\alpha, n-x+\beta)$. Therefore, the posterior mean and variance are

$$\begin{aligned} E(p|x) &= (\alpha+x)/(\alpha+\beta+n), \\ Var(p|x) &= \frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}, \\ s.d(p|x) &= \sqrt{\frac{(\alpha+x)(\beta+n-x)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}}. \end{aligned}$$

The posterior mean can be rewritten as a weighted average of the prior mean and MLE.

$$\frac{(\alpha+\beta)}{(\alpha+\beta+n)}\frac{\alpha}{(\alpha+\beta)} + \frac{n}{(\alpha+\beta+n)}\frac{x}{n}.$$

Once again, the importance of both the prior and the data comes out, the relative importance of the prior and the data being measured by $(\alpha+\beta)$ and $n$. It will not escape one's attention that if $n$ is large then the posterior mean is approximately equal to the MLE, $\hat{p}_{\text{mle}} = x/n$ and the posterior variance is quite small, so the posterior distribution is concentrated around $\hat{p}_{\text{mle}}$ for large $n$. We can interpret this as an illustration of a fact that when we have lots of data, the data tend to wash away the influence of the prior.