Let $\mathcal{F}$ be a Binomial model where n is fixed and $\theta \sim \text{Beta}(\alpha, \beta) = \frac{1}{\text{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$. It turns out that

$$\text{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$

and

$$\text{Var}[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Then

$$\begin{aligned}
\mathbb{P}(\theta \mid X) &= \frac{\mathbb{P}(X \mid \theta)\,\mathbb{P}(\theta)}{\mathbb{P}(X)} \\
&= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{1}{\text{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{1}{\text{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\,d\theta} \\
&= \frac{\theta^{x-\alpha-1}(1-\theta)^{n-x+\beta-1}}{\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\,d\theta} \\
&= \frac{1}{\text{B}(x+\alpha, n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\
&= \text{Beta}(x+\alpha, n-x+\beta)
\end{aligned}$$

What we have done here is that we went from $\theta \to \theta|X$. We went from $\text{Beta}(\alpha,\beta))$ to $\text{Beta}(x-\theta, n-x+\beta)$. The beta is the conjugate prior for the binomial likelihood model.

Note:

- $\hat{\theta}_{\text{MMSE}} = \text{E}[\theta|X] = \frac{x+\alpha}{n+\alpha+\beta}$
- $\hat{\theta}_{\text{MAP}} = \text{Mode}[\theta|X] = \frac{x+\alpha-1}{n+\alpha+\beta-2}$ if $x+\alpha > 1$ and $n-x+\beta > 1$
- $\hat{\theta}_{\text{MAE}} = \text{Med}[\theta|X]$ which is done by a computer

Let's look at X^*, a future observation. This means $n^* = 1$. Then

$$\begin{aligned}
\mathbb{P}(X^* \mid X) &= \int_{\Theta)} \mathbb{P}(X^* \mid \theta)\,\mathbb{P}(\theta \mid X)\,d\theta \\
&= \int_0^1 \underbrace{\theta^{x^*}(1-\theta)^{1-x^*}}_{PMF} \cdot \underbrace{\frac{1}{\text{B}(x+\alpha, n-x+\beta-1)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}_{PDF}\,d\theta \\
&= \frac{1}{\text{B}(x+\alpha, n-x+\beta)}\int_0^1 \theta^{x^*+x+\alpha-1}(1-\theta)^{-x^*+n-x+\beta}\,d\theta \\
&= \frac{\text{B}(x^*+x+\alpha, -x^*+n-x+\beta+1)}{\text{B}(\alpha+\beta, n-x+\beta-1)} \\
&= \frac{\Gamma(x^*+x+\alpha)\Gamma(-x^*+n-x+\beta+1)/\Gamma(n+\alpha+\beta+1)}{(\Gamma(x+\alpha)\Gamma(n-x+\beta))/\Gamma(n+\alpha+\beta)}
\end{aligned}$$

If we let $X^* = 1$:

$$\begin{aligned}
\mathbb{P}\left(X^* = 1 \mid X\right) &= \frac{\Gamma(1+x+\alpha)\Gamma(n-X+\beta)/\Gamma(n+\alpha+\beta+1)}{(\Gamma(x+\alpha)\Gamma(n-x+\beta))/\Gamma(n+\alpha+\beta)} \\
&= \frac{(x+\alpha)\Gamma(x+\alpha)/(n+\alpha+\beta)\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)/\Gamma(n+\alpha+\beta)} \\
&= \frac{x+\alpha}{n+\alpha+\beta}
\end{aligned}$$

Here we went from θ to $\theta|X$ using X, or Beta(α, β) to Beta$(x+\alpha, n-x+\beta)$ where x is the number of successes in the data and $n-x$ is the number of failures in the data. Thus we say α is the number of prior successes (pseudosuccesses) and β is the number of prior failures (pseudofailures) Together, α and β represent pseudocounts.

When we assumed $\theta \sim U(0,1)$, we assumed Beta$(\alpha, \beta) =$ Beta$(1, 1)$. Thus $\mathrm{E}[\theta] = \frac{1}{1+1} = \frac{1}{2}$. We think we assumed nothing but actually we assumed 0.5. This is a criticism of Bayesian inference.

In a conjugate model, the prior parameter α, β are "usually" interpreted as pseudocounts.

$$\begin{aligned}
\theta_{\text{MMSE}} = \mathrm{E}[\theta|X] = \frac{x+\alpha}{n+\alpha+\beta} &= \frac{n}{n} \cdot \frac{x}{n+\alpha+\beta} + \frac{\alpha+\beta}{\alpha+\beta} \cdot \frac{\alpha}{n+\alpha+\beta} \\
&= \frac{n}{n+\alpha+\beta}\hat{\theta}_{\text{MLE}} + \frac{\alpha+\beta}{n+\alpha+\beta}\mathrm{E}[\theta] \\
&= (1-\rho)\hat{\theta}_{\text{MLE}} + \rho(\mathrm{E}[\theta])
\end{aligned}$$

If n is high, then ρ is low and thus θ_{MLE} dominates. If n is low, then ρ is high and $\mathrm{E}[\theta]$ dominates. ($\lim_{n\to\infty} \rho = 0$).

$\mathrm{E}[\theta|X]$ is called a "shrinkage estimation" because it shrinks to $\mathrm{E}[\theta]$.

Let's say $n = 2, x = 0$, and $\theta \sim U(0,1)$, meaning $\alpha = \beta = 1$. Thus $\mathrm{E}[\theta] = 0.5$, as shown above, Then $\theta_{\text{MLE}} = 0$. If $\rho = 0.5$, then

$$\mathrm{E}[\theta|X] = (1-\rho)\theta_{\text{MLE}} + \rho\mathrm{E}[\theta] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Here we have shrunk $\mathrm{E}[\theta|X]$ closer to $\mathrm{E}[\theta]$. If α and β are bigger, it shrinks harder.

Wilson Estimate:

$$\mathrm{E}[\theta|X] = \frac{x+\alpha}{n+\alpha+\beta} = \frac{x+1}{n+2}$$

when $\alpha = \beta = 1$.

Confidence Interval:

$$CI_{\theta,1-\alpha} = \left[\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta}_{\text{MLE}})\right]$$

Let's say $x = 1, n = 2, \hat{\theta} = \bar{x} = 0.5$. Then the confidence interval at the 95% confidence level is

$$CI_{\theta,95\%} = \left[0.5 \pm 2\sqrt{\frac{0.5(1-0.5)}{2}}\right] = (-0.21, 1.21)$$

This is absurd because one value is negative and the other is more than 1. We can say $[0, 1]$ but that is just useless.

Let $\theta \sim U(0,1)$, then $\theta|X \sim \text{Beta}(x+1, n-x+1) = \text{Beta}(2,\ 2)$. Here we won't make a best guess but a range.

Credible Region (CR) for θ of size $1-\alpha$:

$$CR_{\theta,1-\alpha} = [\text{Quantile}[\theta|X, \frac{\alpha}{2}], \text{Quantile}(\theta|X, 1-\frac{\alpha}{2})]$$

For this example,

$$\begin{aligned} &= [\text{qbeta}(0.025, 2, 2), \text{qbeta}(0.975, 2, 2)] \\ &= [0.094, 0.906] \end{aligned}$$

Let's say we have a distribution such that there are three peaks. To find a credible region of it, we would have to find the the union of three different peaks, or the HDR (higher density region). This is a disadvantage because it is not plausible to have non contiguous regions and it is computationally expensive.