1. To understand the effect of a potential carcinogen, a study was undertaken where $n = 23$ rats were treated with the potential carcinogen, and the time to tumor occurrence (in months) was recorded. Of particular biological relevance was whether a tumor developed within 6 months, and hence the data was collected over a period of 6 months. Let z_i denote the time to tumor development (in months) for the ith rat, with $z_i \in \{1, \ldots, k+1\}$ and $k = 6$. For all the rats which didn't develop a tumor within the first 6 months, we set $z_i = 7$.
The simplest version of a *continuation ratio* model for the time to tumor occurrence can be expressed as

$$\text{pr}(z_i = j) = p(1-p)^{j-1}, \quad j = 1, \ldots, k,$$

independently for $i = 1, \ldots, n$. The parameter p is commonly referred to as the *discrete hazards*, with the interpretation that $p = \text{pr}(z_i = j \mid z_i \geq j)$, that is, p is the probability of the occurrence of the tumor during month j *given that* it has not occurred within the first $(j-1)$ months.

(a) Assuming a Beta$(1/2, 1/2)$ prior on p, calculate the posterior distribution of p.

(b) Out of the 23 rats in the study, 11 developed tumor within the first month, 5 during the second month, 2 during the third month, 2 during the fourth month, 1 during the fifth month and 1 during the sixth month. One rat did not develop tumor within the first 6 months. Use this information to calculate the posterior mean of p under the Beta$(1/2, 1/2)$ prior.

Solution. First, $\text{pr}(z_i = k+1) = 1 - \sum_{j=1}^{k} p(1-p)^{j-1} = 1 - \{1 - (1-p)^k\} = (1-p)^k$. The joint likelihood given p is

$$L(\mathbf{z} \mid p) = \prod_{i=1}^{n} \prod_{j=1}^{k+1} \pi_j^{\mathbb{1}(z_i = j)},$$

where $\pi_j = \text{pr}(z_i = j \mid p)$ and $\mathbf{z} = (z_1, \ldots, z_n)$. Let $n_j = \sum_{i=1}^{n} \mathbb{1}(z_i = j)$. Substituting the expression for π_{k+1}, we have

$$\begin{aligned} L(\mathbf{z} \mid p) &= \left[\prod_{i=1}^{n} \prod_{j=1}^{k} \{p(1-p)^{j-1}\}^{\mathbb{1}(z_i = j)} \right] \times \left[\prod_{i=1}^{n} \{(1-p)^k\}^{\mathbb{1}(z_i = k+1)} \right] \\ &= p^{\sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}(z_i = j)} \, (1-p)^{\sum_{i=1}^{n} \sum_{j=1}^{k+1} (j-1) \mathbb{1}(z_i = j)} \\ &= p^{\sum_{j=1}^{k} n_j} \, (1-p)^{\sum_{j=2}^{k+1} (j-1) n_j}. \end{aligned}$$

Hence, the posterior of p is Beta$\left(\sum_{j=1}^{k} n_j + 0.5, \sum_{j=2}^{k+1} (j-1) n_j + 0.5 \right)$.

For part (b), use $n_1 = 11, n_2 = 5, n_3 = n_4 = 2, n_5 = n_6 = n_7 = 1$ to get $a = \sum_{j=1}^{k} n_j + 0.5 = 22.50$ and $b = \sum_{j=2}^{k+1} (j-1) n_j + 0.5 = 30.50$, so that $E(p \mid \mathbf{z}) = a/(a+b) = 0.4245$.

2. Let f and g be two probability density functions on $\mathbb{R}$ with $f(\theta)/g(\theta) \neq 0$ for all $\theta \in \mathbb{R}$. The Kullback–Leibler (KL) divergence between f and g, denoted $\mathrm{KL}(f||g)$, is defined as

$$\mathrm{KL}(f||g) = \int_{\theta\in\mathbb{R}} f(\theta) \log\left[\frac{f(\theta)}{g(\theta)}\right] d\theta.$$

Like the total variation distance, $\mathrm{KL}(f||g)$ is a "measure of distance" between densities f and g, though KL is not a distance metric. If f and g respectively have $\mathrm{N}(\mu_1, \tau_1^2)$ and $\mathrm{N}(\mu_2, \tau_2^2)$ distributions, we often write $\mathrm{KL}\big[\mathrm{N}(\mu_1, \tau_1^2) \;||\; \mathrm{N}(\mu_2, \tau_2^2)\big]$ instead of $\mathrm{KL}(f||g)$.

Suppose $x \mid \theta \sim \mathrm{N}(\theta, 1/n)$ and θ is assigned a $N(0,1)$ prior. Let θ_n and σ_n^2 respectively denote the posterior mean and variance of θ, so that the posterior distribution of $\theta \mid x$ is a $\mathrm{N}(\theta_n, \sigma_n^2)$ distribution.

Suppose the true data generating parameter is θ_0; let $\mathbb{E}_0$ denote an expectation under a $\mathrm{N}(\theta_0, 1/n)$ distribution. Let

$$T_n = \mathrm{KL}\big[\mathrm{N}(\theta_n, \sigma_n^2) \;||\; \mathrm{N}(x, 1/n)\big].$$

Find T_n. Does $\mathbb{E}_0 T_n \to 0$ as $n \to \infty$? Interpret the result.

Hint: Work out $\mathrm{KL}\big[\mathrm{N}(\mu_1, \tau_1^2) \;||\; \mathrm{N}(\mu_2, \tau_2^2)\big]$ first.

Solution: First let us find $\big[\mathrm{N}(\mu_1, \tau_1^2) \;||\; \mathrm{N}(\mu_2, \tau_2^2)\big]$. We have

$$\frac{f(\theta)}{g(\theta)} = \sqrt{\frac{\tau_2^2}{\tau_1^2}} \exp\left[-\frac{1}{2}\left\{\frac{(\theta-\mu_1)^2}{\tau_1^2} - \frac{(\theta-\mu_2)^2}{\tau_2^2}\right\}\right].$$

Thus,

$$\log\frac{f(\theta)}{g(\theta)} = \frac{1}{2}\log\frac{\tau_2^2}{\tau_1^2} - \frac{1}{2}\left[\frac{(\theta-\mu_1)^2}{\tau_1^2} - \frac{(\theta-\mu_2)^2}{\tau_2^2}\right].$$

Hence,

$$\begin{aligned}\mathrm{KL}(f||g) = \mathbb{E}_f \log\frac{f}{g} &= \frac{1}{2}\log\frac{\tau_2^2}{\tau_1^2} - \frac{1}{2}\left[1 - \frac{\tau_1^2 + (\mu_1-\mu_2)^2}{\tau_2^2}\right] \\ &= \frac{1}{2}\log\frac{\tau_2^2}{\tau_1^2} + \frac{1}{2}\left[\frac{\tau_1^2}{\tau_2^2} - 1\right] + \frac{1}{2}\frac{(\mu_1-\mu_2)^2}{\tau_2^2}.\end{aligned}$$

Returning to our case, we have $\theta_n = nx/(n+1)$ and $\sigma_n^2 = 1/(n+1)$. Using the above formula, we have

$$T_n = \frac{1}{2}\left[\log(1+1/n) - \frac{1}{n+1} + \frac{nx^2}{(n+1)^2}\right].$$

Thus,

$$\lim_{n\to\infty} \mathbb{E}_0 T_n = \lim_{n\to\infty} \frac{1}{2}\left[\log(1+1/n) - \frac{1}{n+1} + \frac{n(\theta_0^2 + 1/n)}{(n+1)^2}\right] = 0,$$

since $\lim_{x\to 0}\log(1+x) = 0$.

3. A coin with probability p of turning heads is independently flipped 10 times. Assume a $U(0,1)$ prior on p. We are told that 7 out of the 10 flips landed in tails. The results for the remaining three flips are not disclosed. Based on this data, calculate the *posterior distribution* and the *posterior mean* of p; can you identify what distribution the posterior is?

Solution: Let x denote the number of heads out of the $n = 10$ flips, so that $x \mid p \sim \text{Binomial}(n, p)$. We are given that $x \leq 3$, and we have to find the posterior distribution of $p \mid x \leq 3$. Since $\pi(p) \propto 1$, we have

$$\begin{aligned}\pi(p \mid x \leq 3) \propto P(x \leq 3 \mid p) &= \sum_{j=0}^{3} \binom{n}{j} p^j (1-p)^{n-j} \\ &= \sum_{j=0}^{3} \binom{n}{j} \text{Beta}(j+1, n-j+1)\, \pi_j(p) \\ &= \sum_{j=0}^{3} w_j \pi_j(p),\end{aligned}$$

where

$$w_j = \binom{n}{j} \text{Beta}(j+1, n-j+1) = \frac{n!}{j!(n-j)!} \frac{\Gamma(j+1)\Gamma(n-j+1)}{\Gamma(n+2)} = \frac{1}{n+1}.$$

and

$$\pi_j(p) = \frac{p^j (1-p)^{n-j}}{\text{Beta}(j+1, n-j+1)}, \quad p \in (0,1),$$

is the density of a beta$(j+1, n-j+1)$ distribution. Thus, the posterior distribution of p is a mixture of $\pi_j, j = 0, 1, 2, 3$. Further, since the w_js do not depend on j, all the mixture weights are equal. Thus,

$$\pi(p \mid x \leq 3) = \frac{1}{4}\pi_0(p) + \frac{1}{4}\pi_1(p) + \frac{1}{4}\pi_2(p) + \frac{1}{4}\pi_3(p) = \sum_{j=0}^{3} \frac{1}{4}\pi_j(p).$$

The posterior mean

$$\int p\ \pi(p \mid x \leq 3) dp = \sum_{j=0}^{3} \frac{1}{4} \frac{j+1}{n+2} = \frac{5}{24}.$$