5B Bayesian inference

About notation. The general and unambiguous notation for the density of whatever random variable X is $f_X(x)$. So if the random variable is Θ, we write $f_\Theta(\theta)$. Similarly, for a joint density we write $f_{X,Y}(x,y)$, and for a conditional density we write $f_{X \mid \Theta}(x \mid \theta)$.

The subscript clarifies which density function we are talking about. The arguments inside parentheses are where the function is being evaluated. So $f_X(5)$ is the density of X evaluated at point 5.

However, for brevity, we often leave out the subscripts if the argument makes it clear which f we are talking about. So we may write $f(\theta)$ or $f(\vec{x} \mid \theta)$ or $f(\theta \mid \vec{x})$. Do not be confused by this. The f's refer to different functions. This is the typical shorthand notation in Bayesian inference, because otherwise we would have so many subscripts. Look at the arguments to understand which density function it is.

Don't use the shorthand if it is not clear. For example, does $f(1)$ refer to the density of X or Θ?

Class problems

5B1 (Missing airplane) This problem is very loosely based on the (different but likewise Bayesian) method that was used to locate the airplane that disappeared during Malaysia Airlines flight MH370 in March 2014.

The search area is divided into four quadrants. From background information, we assess that the missing airplane has probabilities 0.5, 0.3, 0.1 and 0.1 for being in quadrants 1, 2, 3, 4 respectively. We also assess that whenever we search a quadrant, if the airplane really is there, we find it with probability 30% (and fail to find with probability 70%), independent of any previous search attempts. If the airplane is *not* in the quadrant we search, of course we do not find it.

(a) The airplane location is modelled as a random variable $\Theta \in \{1, 2, 3, 4\}$. Express its prior distribution as a table.

(b) The search crew first searches quadrant 1. Let X be an indicator variable: $X = 1$ if the airplane was now found, and $X = 0$ otherwise. Determine the likelihood function for Θ given the observation $X = 0$.

(c) The airplane was *not* found on the first attempt. Now determine the *posterior* distribution of the airplane location, based on this information. Do you think we should search quadrant 1 again?

(d) The search crew decides to search quadrant 1 again. Let Y be the indicator variable for the second search result. The result is again negative (airplane not found). Determine the posterior distribution of the airplane location, based on this information. Which quadrant do you think we should search next?

5B2 (Metro intervals.) The metro runs regularly with an unknown interval θ (minutes). As Thomas arrives at the station at random times each day, his waiting time X (in minutes) has uniform distribution over $[0, \theta]$, with density

$$f(x \mid \theta) \;=\; \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta, \\ 0, & \text{otherwise.} \end{cases}$$

The interval is unknown to Thomas, so he models it as a random variable Θ. By his prior knowledge he assumes the interval is at least 1 minutes; and that the interval is fairly short, so longer intervals have smaller probability density. To capture these ideas, he takes the prior distribution of Θ to have density

$$f_\Theta(\theta) \;=\; \begin{cases} 0.2\,\theta^{-1.2}, & \theta \ge 1, \\ 0, & \text{otherwise.} \end{cases}$$

Over five days Thomas has observed waiting times $x = (7, 3, 2, 9, 6)$.

(a) **Draw** Thomas's *prior* density function over the interval $[0, 20]$ (by hand or computer). **Verify** that it is is indeed a density function, by calculating its integral over the range of possible values. Hint: You can integrate it just like a polynomial, even though the exponent is not an integer. For drawing, it is enough to calculate the function at a few points to get the rough shape.

(b) **Calculate** the *posterior* density of Θ and **draw** it over $[0, 20]$ by hand or computer. Hint: prior times likelihood gives the unnormalized posterior, but you need to normalize it.

(c) Thomas decides to use the *posterior mean* (i.e. expected value of the posterior distribution) as a point estimate for Θ. **Calculate** his estimate.

(d) Using the posterior distribution, **calculate** the probability that $\Theta < 15$.

Home problems

5B3 (Leaf lengths) A botanist assumes that the leaf lengths (in cm) of a certain rose species have the normal distribution with an unknown mean Θ and a known standard deviation $\sigma = 2$. Each leaf length is independent from others (but they have this same distribution). Further, his prior distribution for Θ is normal with mean $\mu_0 = 10$ and standard deviation $\sigma_0 = 1$. The botanist has measured five leaf lengths as $\vec{x} = (9, 13, 14, 12, 17)$. After these observations, find:

(a) The posterior mean of Θ.

(b) An interval that contains Θ with probability 90%.

Hint. Under the given assumptions, the posterior distribution of Θ is also a normal distribution, whose mean and standard deviation have known formulas. See Lecture 5A and/or Ross's Section 7.8. Once you know the posterior distribution of Θ, you can use its CDF to find a suitable interval. You can use tables or a calculator. Check the R commands `qnorm` and `pnorm`.

5B4 (Dangerous road) The number of accidents in a month, on a certain strip of road, is assumed to follow the Poisson distribution

$$f(x \mid \theta) \ = \ e^{-\theta}\frac{\theta^x}{x!}, \quad x = 0, 1, 2, \dots,$$

with an unknown mean parameter $\theta > 0$. The numbers of different months are assumed independent and follow the same distribution. From previous experience, a bayesian engineer has estimated the unknown parameter Θ to have probabilities 0.25, 0.50 and 0.25 for the values 1, 2, 3 respectively.

(a) During the first month, there are two accidents on the road. Find the posterior distribution of Θ after these observations. In other words, find the posterior probabilities for $\Theta = 1, 2, 3$.

(b) During the second month, there are no accidents on the road. Find the posterior distribution of Θ, using the observations of both months.