# 6A Bayesian inference II

## Class problems

**6A1** (Predicting coin tosses) A coin has an unknown probability $P$ for turning up heads (1), and $1-P$ probability for tails (0), independently on each toss. A priori we assume the unknown probability parameter $P$ to have the uniform distribution over $[0, 1]$.

(a) If the unknown probability has value $P = p$, what is then the probability that 10 consecutive tosses are all heads?

(b) Before seeing any data, working from the prior distribution of $P$, what is the (prior) predictive probability that the first 10 tosses will be all heads? Hint: By the law of total probability,

$$f(\vec{x}) = \int_0^1 f_{\vec{X}|P}(\vec{x} \mid p)\ f_P(p)\ dp.$$

One of the things in the integrand is the prior, and the other you get from (a). Then you need to do the integral.

(c) After 20 tosses, of which 20 were heads, what is the posterior distribution for $P$? You can report it either by its density function, or its name and parameters.

(d) After the 20 observed heads in (c), what is now the (posterior) predictive probability that the *next* 10 tosses will be all heads? Hint: Again use the law of total probability, but apply the posterior distribution of $P$. (Or consult slides of lecture 5B.)

(e) Compare the numerical results from (b) and (d) to the alternative scenario: A coin is *known* to be fair ($p = 0.5$), we toss it ten times, and we ask for the probability of obtaining ten heads. Can you explain, by common sense, the relative order of these three numbers?

**Solution.**

(a) $p^{10}$.

(b) Here $\vec{X}$ are the first ten tosses, and $\vec{1}$ denotes a sequence of ten ones.

$$\begin{aligned} f_{\vec{X}}(\vec{1}) &= \int_0^1 f_{\vec{X}|P}(\vec{1} \mid p) f_P(p) dp \\ &= \int_0^1 p^{10} \cdot 1\ dp \\ &= \frac{1}{11} \approx 0.091. \end{aligned}$$

(c) Here $\vec{X}$ are the first twenty tosses, and $\vec{1}$ denotes twenty ones. If we work out the posterior density as in Lecture 5B slides, we find it is

$$f_{P|\vec{X}}(p \mid \vec{1}) = 21p^{20}$$

for $0 \le p \le 1$. In other words, the posterior distribution of $P$ is Beta$(21, 1)$.

(d) Let us denote the first twenty tosses by $\vec{X}$ and the next ten tosses by $\vec{Y}$. In both cases we denote "all ones" by $\vec{1}$, hopefully causing no confusion. We work like in (b), but use the posterior density from (c). Our posterior predictive probability for $\vec{Y} = \vec{1}$ is

$$\begin{aligned} f_{\vec{Y}|\vec{X}}(\vec{1} \mid \vec{1}) &= \int_0^1 f_{\vec{Y}|P}(\vec{1} \mid p) f_{P|\vec{X}}(p \mid \vec{1}) \, dp \\ &= \int_0^1 p^{10} \cdot 21p^{20} \, dp \\ &= 21 \int_0^1 p^{30} \, dp \\ &= \frac{21}{31} \approx 0.677. \end{aligned}$$

(e) For the known fair coin, the probability of ten heads is $(\frac{1}{2})^{10} = \frac{1}{1024} \approx 0.001$.

Common-sense explanation:

- If we know that the coin is fair, the extreme result "ten heads" is rather unlikely.
- In (a), based on the prior, it was quite possible that the coin might be highly upbiased, like $p = 0.90$ or $p = 0.95$, so a result of ten heads would not be too surprising.
- In (d), we have seen some evidence suggesting that the coin seems up-biased, so we have even stronger reasons to believe it might produce ten heads. However, from 20 tosses we are still not very sure about the strength of the bias, so we are still assigning *only* a probability 0.677 to the extreme event of ten heads (although the predictive probability for *one* heads result is already $21/22 \approx 0.954$).

**6A2** (Waiting times) In this exercise we practice working with a new distribution family. Keep calm, read carefully, look at the known facts and think which fact(s) could be useful in your task.

If a continuous random variable $\Lambda$ has density function

$$f_\Lambda(\lambda) = \begin{cases} c\,\lambda^{\alpha-1}\,e^{-\beta\lambda} & \text{when } \lambda > 0, \\ 0 & \text{otherwise,} \end{cases}$$

we say that $\Lambda$ has the **gamma distribution** with parameters $\alpha > 0$ and $\beta > 0$. We denote this by $\Lambda \sim \text{Gam}(\alpha, \beta)$. The $c$ is a constant that ensures that $\int_0^\infty f(\lambda) = 1$, so that $f$ is indeed a density function. It is known that if $\Lambda \sim \text{Gam}(\alpha, \beta)$, then

$$\text{E}(\Lambda) = \alpha/\beta.$$

It is also known that if $\alpha$ is a positive integer, then $c = \beta^\alpha/(\alpha-1)!$, where $\alpha!$ is the factorial.

(a) Particle decays happen at random intervals, independently, at an unknown rate of $\Lambda$ decays per second. For the unknown parameter, we assume a prior density $\Lambda \sim \text{Gam}(2, 10)$. Write down its density function. Calculate its mean (applying the known facts above it should be easy). Based on the mean, explain roughly what kind of decay rate we are expecting for the particles.

(b) If the decay rate is $\Lambda = \lambda$, then each decay interval $X_i$ has the exponential distribution with rate $\lambda$. Thus it has density

$$f_{X_i|\Lambda}(x_i \mid \lambda) = \lambda e^{-\lambda x_i},$$

for $x_i > 0$. This is the likelihood. Write down the unnormalized posterior density

$$f_\Lambda(\lambda)\; f_{\vec{X}|\Lambda}(\vec{x} \mid \lambda),$$

when three decay intervals $\vec{x} = (x_1, x_2, x_3) = (3.0, 12.2, 16.1)$ have been observed, measured in seconds. Apply the fact that the three observations are independent.

(c) Using the unnormalized posterior density, find the MAP estimate of $\Lambda$. Hint: Logarithm and derivate?

(d) From the form of the unnormalized posterior density, you should recognize that it is the density of a gamma distribution, up to a normalizing constant. What are its parameters $\alpha$ and $\beta$? Hint: Look separately at the exponents of $\lambda$ and $e$.

(e) Now that you know the posterior distribution of $\Lambda$, calculate its mean (again use the formula to make it easy). Based on the mean, explain roughly what kind of decay rate are we now expecting for the particles.

(f) (Optional, requires computer.) The $q$-quantile of the $\text{Gam}(\alpha, \beta)$ distribution can be computed with the R command `qgamma(q, alpha, beta)`. Find a 95% credible interval for the parameter $\Lambda$, based on its posterior distribution.

**Solution.**

(a) Plugging in $\alpha = 2$ and $\beta = 10$, the density function is

$$f(\lambda) = 100\ \lambda\ e^{-10\lambda}$$

for $\lambda > 0$. The mean of a $\text{Gam}(2, 10)$ distribution is simply $2/10 = 0.2$. If the parameter $\lambda$ is near this value, then we are expecting a rate of about 0.2 decays per second (or one decay in five seconds).

(b) Multiplying prior and likelihood (for three data points), we obtain

$$\begin{aligned} & 100\ \lambda\ e^{-10\lambda}\ \lambda e^{-\lambda x_1}\ \lambda e^{-\lambda x_2}\ \lambda e^{-\lambda x_3} \\ =& 100\ \lambda^4\ e^{-\lambda\,(10+x_1+x_2+x_3)} \\ =& 100\ \lambda^4\ e^{-\lambda\,(10+3.0+12.2+16.1)} \\ =& 100\ \lambda^4\ e^{-\lambda\,41.3}, \end{aligned}$$

for $\lambda > 0$.

(c) Logarithm of the unnormalized posterior is

$$\log(100) + 4\log(\lambda) - 41.3\lambda,$$

its derivative is

$$4/\lambda - 41.3$$

whose only zero point is

$$\lambda = 4/41.3 \approx \mathbf{0.097}.$$

By inspecting the shape of the function we note that this is the maximum point, that is, the MAP estimate.

(d) If we multiply the expression by a suitable normalizing constant, it will become

$$c\ \lambda^4\ e^{-41.3\lambda},$$

which indeed has the form of a gamma distribution. We observe that the exponent of $\lambda$ is $\alpha - 1 = 4$, so $\alpha = 5$. Also we observe that the exponent of $e$ is $-\beta\lambda = -41.3\lambda$, so $\beta = 41.3$. So this is the density of a $\text{Gam}(5, 41.3)$ distribution.

(e) As hinted, the posterior mean of $\Lambda$ is $5/41.3 \approx \mathbf{0.121}$. If $\lambda$ is near this value, then we are expecting a rate of about 0.121 decays per second (or one decay in $1/0.121 \approx 8.3$ seconds).

(f) The posterior distribution of $\Lambda$ is $\text{Gam}(5, 41.3)$. In R, we can choose the two endpoints as `qgamma(0.025, 5, 41.3)` and `qgamma(0.975, 5, 41.3)`. The credible interval is approximately $[0.039, 0.248]$.

After the three observations whose sample mean was about 10.4, we have 95% probability for the particle decay rate being within $[0.039, 0.248]$ decays per second. Or, since they follow an

exponential distribution, that the mean decay interval would be within $[1/0.248, 1/0.039] \approx [4.0, 25.4]$ seconds. The credible interval is wide, but it will become narrower if more data is accumulated. Compare to exercise 5A4 where we could not really work with 3 observations, because their sample mean would not be normal.

## Home problems

**6A3** (Multinomial DNA model) The human DNA is can be modelled as a string of length $3 \times 10^9$, made of the four letters A, C, G, T. Only about 1.5% of the string consists of *protein-coding* regions (direct recipes for building proteins by concatenating amino acids). Other parts of the DNA have other duties (partially unknown).

Within protein-coding regions, based on some studies, the four letters occur at relative frequencies $(0.17, 0.29, 0.33, 0.21)$ respectively. Outside these regions, the letters occur at relative frequencies $(0.25, 0.25, 0.25, 0.25)$. Our simplified model states that each letter occurs randomly with these probabilities, independent of the other letters.

We are looking at a randomly chosen location $i$ in the string, and are trying to find whether it is *inside* a protein-coding region ($\Theta = 1$), or *outside* them ($\Theta = 0$). We look at a sequence of 100 letters around the location $i$, and observe a string AACTG...TGA, that contains 16 A's, 26 C's, 38 G's and 20 T's, in some order. We assume, for simplicity, that the whole 100 letters is either completely within a protein-coding region, our completely outside them.

Since $\Theta$, the indicator for our location being inside a protein-coding region, is unknown, we treat it as the unknown parameter, whose value determines the letter probabilities. Note that here we are not estimating the multinomial probability parameters; they are given, the only unknown thing is which of the two distributions applies.

(a) What is the prior distribution of $\Theta$? What does it mean?

(b) If $\Theta = 0$, what is the probability of observing exactly the 100-letter sequence we observed? Hint: Think of the ordered three-party samples in Lecture 5B. Do not worry that the probability is very small. Express it with at least three significant digits. *Alternatively*, you can use the likelihoods of the *counts*.

(c) If $\Theta = 1$, what is the probability of observing exactly the 100-letter sequence we observed?

(d) Calculate the posterior distribution of $\Theta$, and explain what it means.

This is a grossly simplified model of the DNA, and the numbers are somewhat made up. However, similar methods are have been used for finding which regions of the DNA are protein-coding.

**Grading.** 0.5 p for each item.

**Solution.**

(a) $\mathrm{P}(\Theta = 0) = 0.985$ and $\mathrm{P}(\Theta = 1) = 0.015$. Since the protein-coding regions cover 1.5% of the DNA, pointing at a random location has this probability of hitting one of those regions.

(b) Let $\vec{X} = (X_1, X_2, \ldots, X_{100})$, where each $X_j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Note that A, C, G, T are the possible *values* of $X_i$. (If you wish, you could encode them as the integers $1, 2, 3, 4$.)

The probability for observing our particular sequence $\vec{x}$ outside the protein-coding regions is

$$f_{\vec{X}|\Theta}(\vec{x} \mid 0) = 0.25^{16} \cdot 0.25^{26} \cdot 0.25^{38} \cdot 0.25^{20} = 0.25^{100} \approx 6.22 \cdot 10^{-61}.$$

(c) The probability for observing our particular sequence $\vec{x}$ inside a protein-coding region is

$$f_{\vec{X}|\Theta}(\vec{x} \mid 1) = 0.17^{16} \cdot 0.29^{26} \cdot 0.33^{38} \cdot 0.21^{20} \approx 7.20 \cdot 10^{-59}.$$

(d) Since $\Theta$ has only two values, we can express the whole thing in a table.

| $\theta$ | Prior $f(\theta)$ | Likelihood $f(\vec{x} \mid \theta)$ | Unnormalized posterior | Posterior $f(\theta \mid \vec{x})$ |
|---|---|---|---|---|
| 0 | 0.985 | $6.22 \cdot 10^{-61}$ | $6.13 \cdot 10^{-61}$ | 0.362 |
| 1 | 0.015 | $7.20 \cdot 10^{-59}$ | $1.08 \cdot 10^{-60}$ | 0.638 |
| $\sum$ | 1.000 | | $1.69 \cdot 10^{-60}$ | 1.000 |

Based on the prior and the observations, the location has (posterior) probability of 63.8% of being in a protein-coding region.

**6A4** (Coin shaking) Professor Abel has made a coin-tossing machine. Initially, the coin is placed tails up (0) on the bottom of a drinking glass. Then a machine shakes the glass for a while, and the coin is inspected to see whether the top face is tails (0) or heads (1). This is done repeatedly to generate a sequence of random numbers $X_1, X_2, X_3, \ldots$, where $X_i \in \{0, 1\}$ is the coin position after $i$ shaking rounds. The initial position $X_0 = 0$ is known.

After 50 shaking rounds, the coin positions were

$$\vec{x} = (x_1, x_2, \ldots, x_{50}) = (0000110111\ 1111100000\ 0000000100\ 0111110011\ 1111111111)$$

(We have dropped the commas between digits for brevity, and added a small space after each ten digits, but really this is a sequence of 50 integers, each zero or one.)

By our physical understanding, we assume that on each round, the coin will *flip* with probability $\theta$, independent of its current position and of what happened before. So to each round we associate a random variable $F_i \in \{0, 1\}$, where 0 indicates no flip, and 1 indicates flip. If the coin does not flip, then $X_i = X_{i-1}$. If it flips, then $X_i = 1 - X_{i-1}$.

(a) Look at the data $\vec{x}$. Does it seem like the coin is flipping with probability $\theta = 0.5$ every time?

(b) We observe that the coin flipped 9 times in our experiment. Treat the unknown flipping probability as a continuous random variable $\Theta$, with uniform prior over the interval $[0, 1]$. Find the posterior distribution of $\Theta$, and also a posterior point estimate of your choice. (Posterior mean and posterior mode would be easy choices.)

Hint. Think of the sequence of *flips*, not of the sequence of *coin positions*. The flips are independent. You can use either the actual flip sequence, or the number of flips as your data; the result will be the same. Recall Lectures 5A and 5B. Apply the fact that in the binary model, the prior $\text{Beta}(1, 1)$ becomes the posterior $\text{Beta}(1 + a, 1 + b)$ when $a$ ones and $b$ zeros are observed. You do *not* need to find the normalization constant manually. You can also use the fact that the mean of $\text{Beta}(a, b)$ is $a/(a + b)$.

(c) Find a 95% credible interval for $\Theta$, that is, an interval that contains $\Theta$ with 95% probability (conditional on the observed flips). Hint: use a computer. For example, the R command `qbeta(q, a, b)` gives the $q$-quantile of the $\text{Beta}(a, b)$ distribution.

(d) Professor Abel observes that the coin landed heads 28 times, and tails 22 times. Because the observed relative frequencies were approximately half and half, he claims that his machine manages to toss coins very randomly, with each result being heads with 50% probability, independent of other results. What do you think of his reasoning?

(e) Suppose that after a long series of experiments, we have become convinced that $\theta = 0.2$ (to a high precision). Consider a series of ten shakings. What is the probability that after ten shakings, the coin is in the same position as before them? Report with three decimals. Hint: Consider the number of flips that occurred. What is the probability that it is even?

**Grading.** 0.4 points for each item, total rounded up. In (b), points given if found beta distribution with roughly correct parameters. In (d), points given for observing (in any meaningful

way) that Abel's idea of independent coin results does not hold water. In (e), points given for the correct numerical result.

**Solution.**

(a) The coin seems to be staying in the same position for a long time. It does not seem to be flipping often enough.

(b) The prior of $\Theta$ is uniform, or $f(\theta) = 1$, or $\text{Beta}(1, 1)$.

In each shaking, the likelihood for flip is $f(1 \mid \theta) = \theta$, and the likelihood for no flip is $f(0 \mid \theta) = 1 - \theta$. So the likelihood for a sequence that contains 9 flips and 41 no-flips is $\theta^9(1-\theta)^{41}$. The posterior density is proportional to the unnormalized posterior $1 \cdot \theta^9(1-\theta)^{41}$. It would be possible to find the normalizing constant (by doing the integral), but as hinted, we will simply observe that this is the density of a $\text{Beta}(10, 42)$ distribution.

The two suggested posterior point estimates are:

- posterior mean $= \frac{10}{10+42} \approx 0.192$, applying the hint.
- posterior mode $= \frac{9}{50} = 0.180$, which can be found by maximizing the posterior (either normalized or not).

(c) In R, computing `qbeta(0.025, 10, 42)` and `qbeta(0.975, 10, 42)` we get the endpoints of such an interval. The interval is $[0.098, 0.309]$.

You can use your favorite software to calculate it. In Matlab or Octave, the command would be `betainv`. Working out the CDF by manually integrating the density would be possible, because it is a polynomial, but quite laborious.

(d) Abel may be right about the heads and tails being half and half, but there is no support for his claim on the independence of the coin results. On the contrary, the physical model makes it plausible that the results are *not* independent. In (b) and (c) we worked out that the flipping probability is roughly around 0.2 (or within $[0.1, 0.3]$), which makes sense because 9 flips were observed in 50 shakings. If Professor Abel was right, we would have expected something like 25 flippings, more or less.

(e) Since the *flippings* are independent (even though the *coin positions* are not), we can consider this as a binary experiment with $n = 10$ and success probability $p = 0.2$. The number of flips has the binomial distribution with these parameters. The coin lands in the same position where it started exactly if the number of flips is even, that is, has one of the values $0, 2, 4, 6, 8, 10$. By additivity, this happens with probability

$$b(0) + b(2) + b(4) + b(6) + b(8) + b(10) \approx 0.503,$$

where $b(k) = \binom{10}{k} 0.2^k 0.8^{10-k}$ is the density of the binomial distribution at point $k$. An easy way to compute this is the following R command:
`sum(dbinom(c(0,2,4,6,8,10), 10, 0.2))`

Professor Abel was partially right in the following sense: If we look at coin positions with *many* shakings in between, for example $X_0$ and $X_{10}$, then there are roughly equal chances that $X_{10}$

equals or does not equal $X_0$. So if we consider the results only after each 10 shakings, we get a fairly independent random sequence $X_{10}, X_{20}, X_{30}, \ldots$ Such *thinning* is one part of modern Markov Chain Monte Carlo (MCMC) methods, by which complicated posterior densities are studied. Such methods are beyond this course.