Math 775: Homework 2

Alex Dewey

February 23, 2016

1 Exercises

Exercise 1.

i.

$$P(y|\theta) = \binom{y+r-1}{y}\theta^r(1-\theta)^y, y = 0, 1, ...$$

$$\log P(y|\theta) = \log\binom{y+r-1}{y} + r\log\theta + y\log(1-\theta)$$

$$\frac{\delta \log P(y|\theta)}{\delta\theta} = \frac{r}{\theta} - \frac{y}{1-\theta}$$

$$\frac{\delta^2 \log P(y|\theta)}{\delta\theta^2} = -\frac{r}{\theta^2} - \frac{y}{(1-\theta)^2}$$

r is a given constant and $E(y|\theta)$ is just the expectation of a negative binomial distribution $E(y|\theta) = \frac{r\theta}{1-\theta}$ (we can derive this by using the infinite sum

$$\theta + \theta^2 + ... = \frac{\theta}{1-\theta}$$

for r = 1 and treating the spaces between negative trials as r IID draws).

Anyway, the Fisher information follows from this:

$$J^{\text{Fisher}}(\theta) = -E\big[\frac{\delta^2 \log P(y|\theta)}{\delta\theta^2}\big|\theta\big] = \frac{r}{\theta^2} + \frac{\frac{r\theta}{1-\theta}}{(1-\theta)^2} = \frac{r}{\theta^2} + \frac{r\theta}{(1-\theta)^3}$$

ii.

$$P(y|\theta) = \frac{1}{y!}\theta^y e^{-\theta}$$

$$\log P(y|\theta) = -\log y! + y\log\theta - \theta$$

$$\frac{\delta \log P(y|\theta)}{\delta\theta} = \frac{y}{\theta} - 1$$

$$\frac{\delta^2 \log P(y|\theta)}{\delta\theta^2} = -\frac{y}{\theta^2}$$

Now, the expectation of y given θ is θ, because by definition y is a Poisson distribution with mean θ. So we have:

$$J^{\text{Fisher}}(\theta) = -E\big[\frac{\delta^2 \log P(y|\theta)}{\delta\theta^2}\big|\theta\big] = -E(-\frac{y}{\theta^2}|\theta) = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

iii.

Since σ^2 is known, we can use the standard normal density:

$$P(y|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

.

$$\log P(y|\mu) = \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(y-\mu)^2$$

$$\frac{\delta \log P(y|\mu)}{\delta\mu} = \frac{1}{\sigma^2}(y-\mu)$$

$$\frac{\delta^2 \log P(y|\mu)}{\delta\mu^2} = -\frac{1}{\sigma^2}$$

So

$$J^{\text{Fisher}}(\mu) = E(\frac{1}{\sigma^2}|\mu) = \frac{1}{\sigma^2}$$

So the Fisher information is the constant, known value $1/\sigma^2$.

iv.

By *Bayesian Data Analysis* (Gelman, pg. 43), we can write the likelihood for a normal model with known mean μ and unknown variance σ^2 (with y an iid vector of n observations) as:

$$p(y|\sigma^2) \propto \sigma^{-n} e^{-\frac{n}{2\sigma^2}v}$$

with

$$v = \frac{1}{n}\sum_i^n (y_i - \mu)^2$$

.

$$p(y|\sigma^2) = C(\sigma^2)^{-n/2} e^{-nv/2\sigma^2}$$

$$\log P(y|\sigma^2) = \log C + -\frac{n}{2}\log\sigma^2 + -\frac{nv}{2\sigma^2}$$

$$\frac{\delta \log P(y|\sigma^2)}{\delta\sigma^2} = -\frac{n}{2\sigma^2} + \frac{nv}{2(\sigma^2)^2}$$

$$\frac{\delta^2 \log P(y|\mu)}{\delta(\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{nv}{(\sigma^2)^3}$$

$(nv|\sigma^2) \sim \sigma^2\chi^2_n$ and $E(\sigma^2\chi^2_n) = n\sigma^2$ by definition (note that v is centered around the population mean μ, not around the sample mean of y, so we have n degrees of freedom). So:

$$J^{\text{Fisher}}(\mu) = -\frac{n}{2(\sigma^2)^2} + \frac{n\sigma^2}{(\sigma^2)^3} = \frac{n}{2(\sigma^2)^2}$$

.

Exercise 2.

Suppose $p(y|\theta) \sim Bin(n = 50, \theta)$ where $p(\theta) = 0.5 \times Beta(10, 20) + 0.2 \times Beta(15, 15) + 0.3 \times Beta(20, 10)$

We want to compute and then plot $p(\theta|y = 14)$.

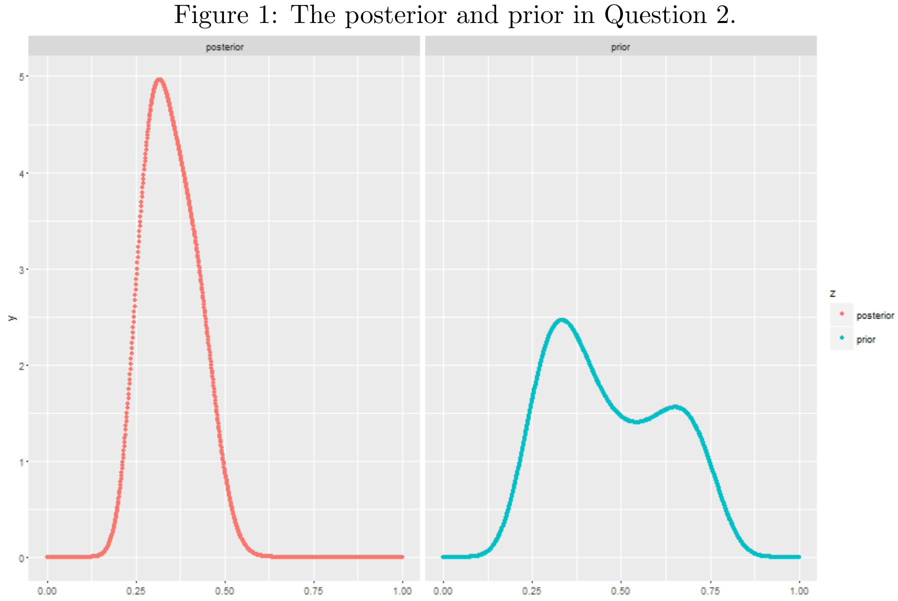
First thing to note is that we don't have to integrate anything to get the answer. The binomial distribution is just the result of n iid (hence exchangeable) Bernoulli trials, and the beta distribution is a conjugate prior with respect to exchangeable Bernoulli data.

Therefore, once we know the prior and data, we can immediately get a posterior for each of the distributions. And we can treat each beta distribution in θ separately (since integration over our Prior $\times$ Likelihood is linear).

So we can simply add y=14 successes and 36 failures to each of our beta distributions to get the form

$p(\theta|y = 14) = 0.5 \times Beta(24, 56) + 0.2 \times Beta(29, 51) + 0.3 \times Beta(34, 46)$.

Figure 1: The posterior and prior in Question 2.



Exercise 3.

Because the W and B follow independent Poisson distributions, we can just multiply Poisson($\lambda + \mu$) and Poisson(λ) in order to get the likelihood. The derivations are straightforward and the functions I used for each observed value of (W, B) are below.

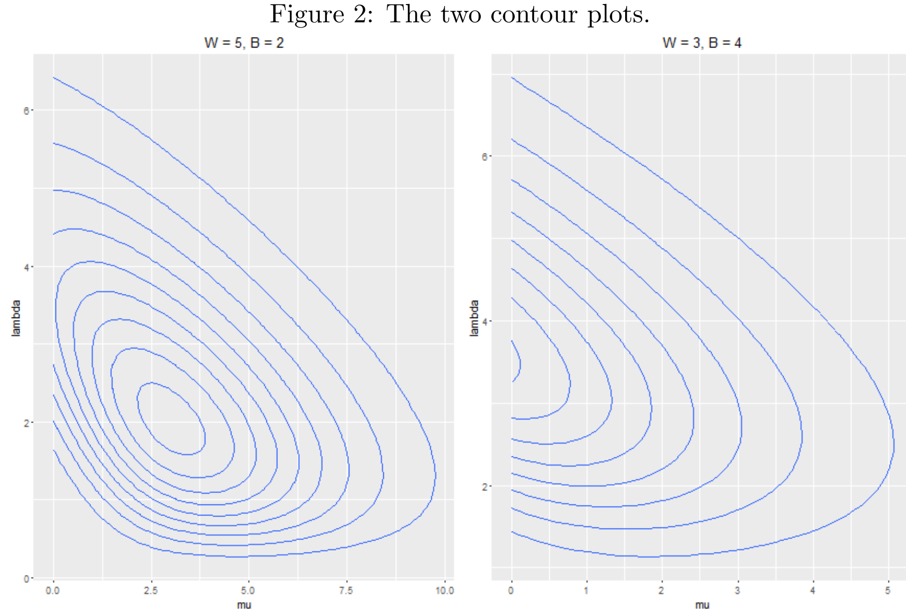
$$p(W = 5, B = 2 | \Theta = (\mu, \lambda)) = \frac{1}{5!2!}(\mu + \lambda)^5 \lambda^2 e^{-(\mu+2\lambda)}$$
$$p(W = 3, B = 4 | \Theta = (\mu, \lambda)) = \frac{1}{3!4!}(\mu + \lambda)^3 \lambda^4 e^{-(\mu+2\lambda)}$$

The contour plots are below. The first plot shows that the maximum likelihood value for $(W = 5, B = 2)$ is $\Theta = (\mu = 3, \lambda = 2)$.

This fits with intuition, since W is distributed with mean $\mu + \lambda$ and B is distributed with mean λ.

For $(W = 3, B = 4)$, the situation is more complicated. The likelihood appears to take on a maximum outside our domain, at $\Theta = (\mu = -1, \lambda = 4)$. Because these are meant to be signals which have, by construction, $\mu, \lambda \geq 0$, we can't draw from $\mu = -1$. But it's clear that the signals would hit a maximum likelihood there if we could draw from it.

Figure 2: The two contour plots.



Exercise 4.

For (y_1, y_2), there is only one non-trivial case in which permuting the indices $p(y_1, y_2)$ is even possible.

In particular, the exchangeability condition holds if and only if $p(1,0) = p(0,1)$. For this distribution, that is clearly true.

$$p(y) = \int_0^1 \theta^{\sum y_i}(1-\theta)^{2-\sum y_i} p(\theta)\delta\theta \tag{1}$$

To get equation (1), we would need the following four conditions to hold:

$$p(0,0) = \int_0^1 (1-\theta)^2 p(\theta)\delta\theta = 0$$

$$p(0,1) = \int_0^1 \theta^1(1-\theta)^1 p(\theta)\delta\theta = 0.5$$

$$p(1,0) = \int_0^1 \theta^1(1-\theta)^1 p(\theta)\delta\theta = 0.5$$

$$p(1,1) = \int_0^1 \theta^2 p(\theta)\delta\theta = 0$$

Note that in the first and fourth equations, $\theta^2 > 0$ and $(1-\theta)^2 > 0$ for all values of θ in the interior of the interval [0,1].

Without writing out the Riemann sums for the first and fourth integrals, it's clear that if we require our prior $p(\theta) \geq 0$ to be a probability distribution, then we must have $p(\theta) = 0$ for all but a subset of measure zero over $[0, 1]$.

Therefore the second and third equations, which are being multipled by the same $p(\theta)$, over the same subset, must also integrate to zero. But they don't, so such a $p(\theta)$ cannot exist. So the condition analogous to De Finetti's Theorem cannot hold for this distribution.

Exercise 5.

a.

When $\alpha = \beta = 1$, the beta prior $B(\alpha, \beta)$ is uniform. So it's natural to take this as the form of the prior and incorporate the data $\hat{\theta} = y/n$ as n Bernoulli trials.

$p(y = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}\theta^k(1-\theta)^{n-k}$.

Integrating this function over θ, we note that it's a beta distribution with parameters $n+1, k+1$ (which integrates to 1) but multiplied by $\frac{\Gamma(n+1)}{\Gamma(n+2)} = \frac{1}{n+1}$.

So our predictive prior is $p(y = k) = \frac{1}{n+1}$ for all k.

b.

The posterior distribution is Beta$(\alpha + k, \beta + (n-k))$ with mean

$$\frac{\alpha + y}{\alpha + \beta + n}$$

.

We can rewrite this posterior mean as a weighted average of the prior mean and the sample mean:

$$\frac{\alpha + y}{\alpha + \beta + n} = \frac{y}{n}\frac{n}{\alpha + \beta + n} + \frac{\alpha}{\alpha + \beta}\frac{\alpha + \beta}{\alpha + \beta + n}$$

.

Since, by definition, we have $\alpha, \beta, y, n \geq 0$ and $\frac{\alpha+\beta}{\alpha+\beta+n} + \frac{n}{\alpha+\beta+n} = 1$, the weighed average above is a convex sum and it must lie between the prior and the sample means.

c.

The variance of Beta(1,1) (our uniform prior) for θ is

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{2^2 * 3} = 1/12$$

.

The variance of our posterior distribution $p(\theta|y = k) = \text{Beta}(1+k, 1+(n-k))$ is

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{(1+k)(1+(n-k))}{(n+2)^2(n+3)}$$

.

The numerator is the only part that depends on k, and (if we allow any real k), it attains a maximum at $k = n/2$ by a standard argument (suppose not: then we could write it as $((1 + n/2) + a)((1 + n/2) - a) = (1 + n/2)^2 - a^2$ for some $a > 0$. Then we would have $a^2 \leq 0$, a contradiction). So we want to show that

$(n/2+1)^2/(n+2)^2(n+3) > 1/12$ for any $n > 0$. Note that $2(n/2+1) = (n+2)$, so this reduces to $\frac{1}{4(n+3)} < \frac{1}{12}$.

So the posterior variance must be less for any observed data than that of the uniform prior.

d.

Start with a prior of Beta(20, 1), $n = 19, k = 0$. So our posterior ends up as Beta(20,20).

The variance of our prior is $\frac{20}{21^2 * 22} \approx 0.00206$. The variance of our posterior is $\frac{400}{40^2 * 41} \approx 0.00610$, or nearly 3 times higher.

Exercise 8

By Gelman (pg. 42), the posterior is just $N(\theta|\mu_n, \tau_n^2)$ with

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

and

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

, with prior $N(\mu_0, \tau_0^2)$ and observed y and known σ^2.

a.

We have $\bar{y} = 150, \sigma^2 = 20^2, \mu_0 = 180, \tau_0 = 40$.

So our posterior distribution is $N(\theta|\mu_n, \tau_n^2)$ with:

$$\frac{1}{\tau_n^2} = \frac{1}{1600} + \frac{n}{400}$$

and

$$\mu_n = \frac{\frac{180}{1600} + \frac{150}{400}}{\frac{1}{1600} + \frac{n}{400}}$$

b.

The predictive posterior distribution for $\tilde{y}$ is a normal distribution with $N(\theta|\mu_n, \tau_n^2 + \sigma^2)$, where μ_n, τ_n are defined as in part (a.). We get extra variance from the fact that we're predicting a single observation rather than the underlying parameter θ.

c-d.

For the posterior and posterior predictive intervals, I used R to implement the above discussion (using $\mu_n \pm 1.960 * \tau_n$ for θ and $\mu_n \pm 1.960 * \sqrt{\tau_n^2 + \sigma^2})$ for $\tilde{y}$).

The results I found were:

Interval	$n = 10$	$n = 100$
θ	(138.488, 162.976)	(146.160, 153.990)
$\tilde{y}$	(109.664, 191.799)	(110.680, 189.470)

Exercise 10

a.

The first thing to note is that for $n < 203$, we must have $p(y = 203|N = n) = 0$.

For $n \geq 203$, we're dealing with a uniform likelihood. $p(y = 203|N = n) = \frac{1}{n}$.

So the posterior distribution is $p(N = n|y = 203) \propto \frac{1}{n}(\frac{99}{100})^n$ for $n \geq 203$, where constants have been removed.

b.

Analytically (after summing from 203 to 50000 to get the normalizing constant), I got a mean of 279.088 and a standard deviation of 79.96.

c.

I tried the Poisson non-informative prior $\sqrt{\frac{1}{\lambda}}$ with $\lambda = 0.01$.

The mean didn't converge as I increased n, because the normalizing constant is given by the harmonic series from 203 to n, and this series diverges. Therefore, the mean and standard deviation continued to increase with n.