

## 1 MLE for the Bernoulli/ binomial model

$$X_i \sim Ber(\theta) \quad (1)$$

$$p(D|\theta) = \theta^{N_1} (1 - \theta)^{N_0} \quad (2)$$

$$\begin{aligned} \ln(p(D|\theta)) &= \ln(\theta^{N_1} (1 - \theta)^{N_0}) \\ &= \ln(\theta^{N_1}) + \ln(1 - \theta)^{N_0} \\ &= N_1 \ln \theta + N_0 \ln(1 - \theta) \\ \theta \ln p(D|\theta) &= \frac{N_1}{\theta} - \frac{N_0}{1 - \theta} \end{aligned}$$

The log-likelihood will take a maximum when the derivative equals 0.

$$\begin{aligned} 0 &= \frac{N_1}{\theta} - \frac{N - N_1}{1 - \theta} \\ 0 &= N_1(1 - \theta) - \theta(N - N_1) \\ 0 &= N_1 - \theta N_1 - \theta N + \theta N_1 \\ 0 &= N_1 - \theta(N_1 + N - N_1) \\ 0 &= N_1 - \theta N \\ \hat{\theta} &= \frac{N_1}{N} \end{aligned}$$

## 2 Marginal likelihood for Beta-Bernoulli model

$$p(X_{1:N}) = p(x_1)p(x_2|x_1)p(x_3|x_{1:2})...p(x_N|x_{N-1}) \quad (3)$$

$$p(X = k|D_{1:N}) = \frac{N_k + \alpha_k}{\sum_i N_i + \alpha_i} \quad (4)$$

$$(\alpha - 1)! = \Gamma(\alpha) \quad (5)$$

Given  $D = H, T, T, H, H \triangleq 1, 0, 0, 1, 1$

$$\begin{aligned}
p(X = 1|\alpha) &= \frac{\alpha_1}{\alpha} \\
p(X = 0|\alpha, D_1) &= \frac{\alpha_0}{\alpha + 1} \\
p(X = 0|\alpha, D_{1:2}) &= \frac{\alpha_0 + 1}{\alpha + 2} \\
p(X = 1|\alpha, D_{1:3}) &= \frac{\alpha_0 + 1}{\alpha + 3} \\
p(X = 1|\alpha, D_{1:4}) &= \frac{\alpha_0 + 2}{\alpha + 4}
\end{aligned}$$

$$\begin{aligned}
p(D) &= p(D_{1:5}) \\
&= p(D_1) \cdot p(D_2|D_1) \cdot p(D_3|D_{1:2}) \cdot p(D_4|D_{1:3}) \cdot p(D_5|D_{1:4}) \quad \text{by (3)} \\
&= \frac{\alpha_1}{\alpha} \cdot \frac{\alpha_0}{\alpha + 1} \cdot \frac{\alpha_0 + 1}{\alpha + 2} \cdot \frac{\alpha_1 + 1}{\alpha + 3} \cdot \frac{\alpha_1 + 2}{\alpha + 4} \quad \text{by (4)} \\
&= \frac{[\alpha_1(\alpha_1 + 1)(\alpha_1 + 2)] [\alpha_0(\alpha_0 + 1)]}{\alpha(\alpha + 1)(\alpha + 2)(\alpha + 3)(\alpha + 4)} \\
&= \frac{[(\alpha_1) \dots (\alpha_1 + N_1 - 1)] [(\alpha_0) \dots (\alpha_0 + N_0 - 1)]}{(\alpha) \dots (\alpha + N - 1)} \\
&= \frac{(\alpha_1 + N_1 - 1)!}{(\alpha_1 - 1)!} \cdot \frac{(\alpha_0 + N_0 - 1)!}{(\alpha_0 - 1)!} \cdot \frac{(\alpha - 1)!}{(\alpha + N - 1)!} \\
&= \frac{\Gamma(\alpha_1 + N_1)}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0)} \cdot \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \\
&= \frac{\Gamma(\alpha_1 + N_1) \Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_1) \Gamma(\alpha_0)} \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0 + \alpha_1 + N)}
\end{aligned}$$

### 3 Posterior predictive for a Beta-Binomial model

$$\begin{aligned}
p(x|n, D) &= Bb(x|\alpha'_0, \alpha'_1, n) \\
&= \frac{B(x + \alpha'_1, n - x + \alpha'_0)}{B(\alpha'_1, \alpha'_0)} \binom{n}{x}
\end{aligned}$$

Given  $n = 1$

$$\begin{aligned}
Bb(1|\alpha_0, \alpha_1, 1) &= \frac{B(1 + \alpha_1, \alpha_0)}{B(\alpha_1, \alpha_0)} \binom{1}{1} \\
&= \frac{\Gamma(1 + \alpha_1)\Gamma(\alpha_0)}{\Gamma(\alpha_0 + \alpha_1 + 1)} \cdot \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \\
&= \frac{\alpha_1\Gamma(\alpha_1)\Gamma(\alpha_0)}{(\alpha_0 + \alpha_1)\Gamma(\alpha_0 + \alpha_1)} \cdot \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \\
&= \frac{\alpha_1}{\alpha_0 + \alpha_1} \\
&= \frac{\alpha_1}{\alpha}
\end{aligned}$$

## 4 Beta updating from censored likelihood

Let  $n$  represent the number of coin tosses. Let  $X$  represent the number of heads. Given  $n = 5$  and  $X < 3$ , we need to compute the posterior  $p(\theta|X < 3)$  under a  $B(1, 1)$  prior up to normalization constants.

$$\begin{aligned}
P(\theta) &= \frac{P(\theta)P(D|\theta)}{P(D)} \\
&= \frac{P(\theta) \cdot P(X < 3|\theta)}{P(X < 3)} \\
P(\theta) &\propto P(\theta) \cdot P(X < 3) \\
&\propto B(1, 1) \cdot \sum_{k=0}^2 P(k|\theta, 5) \\
&\propto \sum_{k=0}^2 \binom{5}{k} \theta^k (1 - \theta)^{5-k}
\end{aligned}$$

## 5 Uninformative prior for log-odds ratio

Let  $\phi = \log \frac{\theta}{1-\theta}$ .  $p(\phi) = 1$  is equivalent to  $p(\phi) = k$ , where  $k$  is a constant and  $0 < k < 1$ .

$$\begin{aligned}
\int_{\phi} p(\phi) d\phi &= \int_{\phi} k d\phi = 1 \\
d\phi &= \frac{d\phi}{d\theta} d\theta \\
\frac{d\phi}{d\theta} &= \frac{d}{d\theta} \left( \ln \frac{\theta}{1-\theta} \right) \\
&= \frac{d}{d\theta} (\ln \theta - \ln(1-\theta)) \\
&= \frac{1}{\theta} - \frac{1}{1-\theta} \cdot -1 \\
&= \frac{1}{\theta} + \frac{1}{1-\theta} \\
\int_{\phi} k d\phi &= \int_{\theta} k \left( \frac{1}{\theta} + \frac{1}{1-\theta} \right) d\theta \\
1 &= k \int_{\theta} \theta^{-1} (1-\theta)^{-1} d\theta
\end{aligned}$$

We recognize the final integral as the normalization constant for a  $Beta(\theta|0, 0)$  distribution.

## 6 MLE for the Poisson distribution

$$P(X = k|\lambda) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for } k \in \{0, 1, 2, \dots\}$$

$$\begin{aligned}
p(x|\lambda) &= e^{-\lambda} \cdot \frac{\lambda^{x_1}}{x_1!} \cdot e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!} \cdot \dots \\
&= \prod_{i=1}^N e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}
\end{aligned}$$

Now we take the log-likelihood and find its maximum.

$$\begin{aligned}
\ell(\lambda) &= \ln p(x|\lambda) \\
&= \ln \left( \prod_{i=1}^N \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \\
&= \ln \left( \frac{e^{-\lambda} \lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda} \lambda^{x_2}}{x_2!} \right) \dots \\
&= \sum_{i=1}^N \ln \left( \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \right) \\
&= \sum_{i=1}^N \ln(e^{-\lambda} \lambda^{x_i}) - \ln(x_i!) \\
&= \sum_{i=1}^N \ln(e^{-\lambda}) + \ln(\lambda^{x_i}) - \ln(x_i!) \\
&= \sum_{i=1}^N -\lambda + x_i \ln \lambda - \ln(x_i!) \\
&= -N\lambda + (\ln \lambda) \sum_{i=1}^N x_i - \sum_{i=1}^N \ln(x_i!)
\end{aligned}$$

$$\begin{aligned}
\ell'(\lambda) &= \frac{1}{\lambda} \sum_{i=1}^N x_i - N = 0 \\
\frac{\sum_{i=1}^N x_i}{\lambda} &= N \\
\lambda_{MLE} &= \frac{\sum_{i=1}^N x_i}{N}
\end{aligned}$$

## 7 Bayesian analysis of the Poisson distribution

7.1 a.

$$\begin{aligned} p(\lambda|D) &= \frac{p(D|\lambda)p(\lambda)}{p(D)} \\ &\propto p(D|\lambda)p(\lambda) \\ &= \frac{e^{-N\lambda} \cdot \lambda^{\sum_{i=1}^N x_i}}{\prod_{i=1}^N x_i!} \cdot \frac{\lambda^{a-1} e^{-\lambda b}}{k} \\ &\propto e^{-N\lambda - \lambda b} \cdot \lambda^{a-1 + \sum_{i=1}^N x_i} \\ &= e^{-\lambda(N+b)} \cdot \lambda^{[a + \sum_{i=1}^N x_i] - 1} \\ &= Ga(\lambda|a + \sum_{i=1}^N x_i, N + b) \end{aligned}$$

7.2 b.

$$\begin{aligned} \frac{a + \sum_{i=1}^N x_i}{N + b} a &\rightarrow 0, b \rightarrow 0 \\ &= \frac{\sum_{i=1}^N x_i}{N} \\ &= \lambda_{MLE} \end{aligned}$$

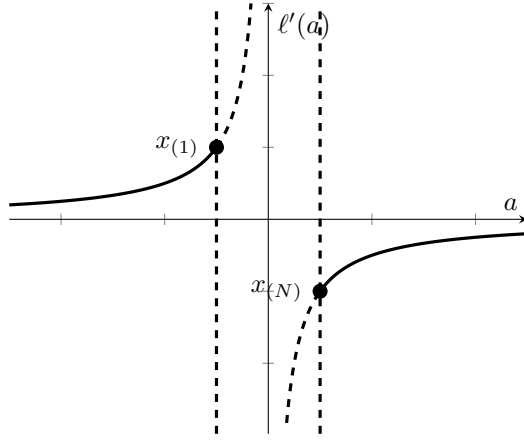
## 8 MLE for the uniform distribution

8.1 a.

$$\begin{aligned} p(D|a) &= \prod_{i=1}^N p(x_i) \\ &= \left(\frac{1}{2a}\right)^N I(x_1, x_2, \dots, x_N \in [-a, a]) \\ &= \begin{cases} \left(\frac{1}{2a}\right)^N & \text{if } -a \leq x_i \leq a, \forall i \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Now we take the derivative of the log-likelihood.

$$\begin{aligned}
\ell(a) &= \ln \left( \frac{1}{2a} \right)^N \\
&= \ln 1 - \ln(2a)^N \\
&= -N \ln(2a) \\
\ell'(a) &= -N \frac{1}{2a} \cdot 2 \\
&= -\frac{N}{a}
\end{aligned}$$



For  $a < 0$ , the likelihood is increasing, so it will be maximized where  $a = x_{(1)}$ , assuming that  $|x_{(1)}| \geq |x_{(N)}|$ . Similarly, for  $a > 0$ , the likelihood is decreasing, so it is maximized where  $a = x_{(n)}$ , assuming that  $|x_{(N)}| \geq |x_{(1)}|$ . This function is only defined when  $\max_i |x_i| \leq a$ . This means that  $\ell$  is maximized at  $a = \max(|x_{(1)}|, |x_{(n)}|)$ .

## 8.2 b.

$$p(x_{n+1}) = \frac{1}{b-a} = \frac{1}{2a}$$

## 8.3 c.

Our approach is not Bayesian, so we will assign zero probability to  $x_{n+1} > a$  and  $x_{n+1} < -a$ . A better solution would be to derive  $\hat{a}_{MAP}$  and give a plug-in approximation.

# 9 Bayesian analysis of the uniform distribution

We must derive the posterior,  $p(D|\theta)$  given the following:

$$p(D, \theta) = \frac{Kb^K}{\theta^{N+K+1}} I(\theta \geq \max(D, b)) \quad (6)$$

Let  $m = \max(D)$ .

$$\begin{aligned} p(D) &= \int_m^\infty \frac{Kb^K}{\theta^{N+K+1}} d\theta \\ &= \begin{cases} \frac{K}{(N+K)b^N} & \text{if } m \leq b \\ \frac{Kb^K}{(N+K)m^{N+K}} & \text{if } m > b \end{cases} \end{aligned} \quad (7)$$

$$\begin{aligned} p(\theta|D) &= \frac{p(\theta, D)}{p(D)} \\ &= \begin{cases} \frac{Kb^K}{\theta^{N+K+1}} \cdot \frac{(N+K)b^N}{K} & \text{if } m \leq b \leq \theta \\ \frac{Kb^K}{\theta^{N+K+1}} \cdot \frac{(N+K)m^{N+K}}{Kb^K} & \text{if } b < m \leq \theta \end{cases} \\ &= \begin{cases} (N+K) \cdot b^{N+K} \cdot \theta^{-(N+K+1)} & \text{if } m \leq b \leq \theta \\ (N+K) \cdot m^{N+K} \cdot \theta^{-(N+K+1)} & \text{if } b < m \leq \theta \end{cases} \\ &\propto \begin{cases} \text{Pareto}(\theta|N+K, b) & \text{if } m \leq b \leq \theta \\ \text{Pareto}(\theta|N+K, m) & \text{if } b < m \leq \theta \end{cases} \\ &= \text{Pareto}(\theta|N+K, \max(m, b)) \end{aligned}$$

## 10 Taxicab (tramcar) problem

$$\text{Pareto}(\theta|N+K, \max(m, b)) \quad (8)$$

### 10.1 a.

Given a non-informative prior,  $\text{Pareto}(\theta|0, 0)$ :

$$\begin{aligned} p(\theta|D) &= \text{Pareto}(\theta|1+0, \max(100, 0)) \\ &= \text{Pareto}(\theta|1, 100) \end{aligned}$$

### 10.2 b.

$E[\theta|D] = \frac{km}{k-1}$  and  $k = 1$ , so the posterior mean is not defined.  
The mode is  $\max(D) = 100$ .



$$\begin{aligned}
\int_{100}^x km^k \theta^{-(k+1)} d\theta &= \frac{1}{2} \\
100 \int_{100}^x \theta^{-2} d\theta &= \frac{1}{2} \\
\left[ -\frac{1}{\theta} \right]_{100}^x &= \frac{1}{200} \\
-\frac{1}{x} + \frac{1}{100} &= \frac{1}{200} \\
x &= 200
\end{aligned}$$

The median is 200.

### 10.3 c.

$$\begin{aligned}
p(D'|D, \alpha) &= \int_{\theta} p(D'|\theta) p(\theta|D, \alpha) d\theta \\
&= \int_{\theta} \frac{1}{\theta} I(x \leq \theta) \cdot N \cdot m^N \cdot \theta^{-(N+1)} d\theta \\
&= Nm^N \int_{\theta} \theta^{-(N+2)} d\theta \\
&= Nm^N \left[ -\frac{1}{N-1} \theta^{-N-1} \right]_{\max(x, m)}^{\infty} \\
&= Nm^N \left[ 0 - \left( -\frac{1}{N-1} \max(x, m)^{-N-1} \right) \right] \\
&= \frac{Nm^N}{(N+1) \max(x, m)^{N+1}} \\
&= \begin{cases} \frac{N}{m(N+1)} & \text{if } x < m \\ \frac{Nm^N}{x^{N+1}(N+1)} & \text{if } x \geq m \end{cases}
\end{aligned}$$

### 10.4 d.

$$\begin{aligned}
p(x = 100|D, \alpha) &= \frac{1 \cdot 100^1}{(1+1)100^{1+1}} = \frac{1}{200} \\
p(x = 50|D, \alpha) &= \frac{1}{(1+1)100} = \frac{1}{200} \\
p(x = 150|D, \alpha) &= \frac{1 \cdot 100^1}{(1+1)150^{1+1}} = \frac{1}{450}
\end{aligned}$$

## 10.5 e.

To improve the accuracy we could use a more informative prior. For example, we could estimate the number of cabs based on the city's population. Accuracy will also improve with more observations.

## 11 Bayesian analysis of the exponential distribution

### 11.1 a.

$$p(x|\theta) = \theta e^{-\theta x} \quad \text{for } x \geq 0, \theta \geq 0$$

$$p(D|\theta) = \prod_{i=1}^N p(x_i|\theta) = \prod_{i=1}^N \theta e^{-\theta x_i}$$

$$\begin{aligned} \ln(p(D|\theta)) &= \sum_{i=1}^N \ln \theta e^{-\theta x_i} \\ &= \sum_{i=1}^N \ln \theta + \ln e^{-\theta x_i} \\ &= \sum_{i=1}^N \ln \theta + -\theta x_i \end{aligned}$$

$$\frac{d}{d\theta} \left( \sum_{i=1}^N \ln \theta + -\theta x_i \right) = \sum_{i=1}^N \frac{1}{\theta} - x_i$$

$$0 = \frac{N}{\theta} - \sum_{i=1}^N x_i$$

$$\hat{\theta} = \frac{N}{\sum_{i=1}^N x_i}$$

$$\hat{\theta} = \frac{1}{\frac{1}{N} \sum_{i=1}^N x_i}$$

11.2 b.

$$\begin{aligned}\hat{\theta} &= \frac{1}{\frac{1}{N} \sum_{i=1}^N x_i} \\ &= \frac{1}{\frac{1}{3}(5+6+4)} \\ &= \frac{1}{5}\end{aligned}$$

11.3 c.

$$\begin{aligned}p(\theta) &= \text{Expon}(\theta|\lambda) \\ &= \lambda e^{-\lambda\theta} \\ &= \frac{\lambda^1}{\Gamma(1)} \theta^{1-1} e^{-\lambda\theta} \\ &= \text{Ga}(\theta|1, \lambda)\end{aligned}$$

The mean of the Gamma distribution is  $\frac{1}{\lambda}$ , so  $\hat{\lambda} = 3$

11.4 d.

$$p(\theta|D, \hat{\lambda}) \propto p(D|\theta)p(\theta|\hat{\lambda})$$

$$\begin{aligned}p(D|\theta) &= \prod_{x=1}^N \theta e^{-\theta x_i} \\ &= \theta e^{-\theta x_1} \cdot \theta e^{-\theta x_2} \cdot \theta e^{-\theta x_3} \dots \\ &= \theta^N e^{(-\theta x_1 - \theta x_2 - \theta x_3 \dots)} \\ &= \theta^N e^{-\theta \sum x_i}\end{aligned}$$

$$p(\theta|\hat{\lambda}) = \hat{\lambda} e^{-\theta \hat{\lambda}}$$

$$\begin{aligned}p(\theta|D, \hat{\lambda}) &\propto \theta^N e^{-\theta \sum x_i} \hat{\lambda} e^{-\theta \hat{\lambda}} \\ &\propto \theta^N e^{-\theta(\hat{\lambda} + \sum x_i)} \\ &= \text{Ga}(\theta|N+1, \hat{\lambda} + \sum_{i=1}^N x_i)\end{aligned}$$

### 11.5 e.

Yes, the prior is equivalent to a Gamma distribution and the posterior is also a Gamma distribution.

### 11.6 f.

The mean of a  $\text{Ga}(\theta|a, b)$  distribution is  $\frac{a}{b}$ . The mean of  $\text{Ga}(\theta|N + 1, \hat{\lambda} + \sum_{i=1}^N x_i)$   
 $= \frac{N+1}{\hat{\lambda} + \sum_{i=1}^N x_i}$

### 11.7 g.

The posterior accounts for the exponential prior. The prior accounts for expert knowledge, thus it is more reasonable.

## 12 MAP estimate for the Bernoulli with non-conjugate priors

### 12.1 a.

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta) \cdot p(\theta) \\ &\propto \theta^{N_1} (1 - \theta)^{N_0} \cdot \begin{cases} .5 & \text{if } \theta = .4 \text{ or } .5 \\ 0 & \text{else} \end{cases} \\ &\propto \begin{cases} \theta^{N_1} (1 - \theta)^{N - N_1} & \text{if } \theta \in \{.4, .5\} \\ 0 & \text{else} \end{cases} \\ &= \begin{cases} .4^{N_1} (.6)^{N - N_1} & \text{if } \theta = .4 \\ .5^N & \text{if } \theta = .5 \\ 0 & \text{else} \end{cases} \end{aligned}$$

### 12.2 b.

If  $N$  is small,  $\theta = .4$  will lead to a better estimate since the prior is close to the true value. As  $N$  grows, the data will overwhelm the prior.

### 13 Posterior predictive distribution for a batch of data with the Dirichlet-multinomial model

$$\begin{aligned}
p(D'|D, \alpha) &= \int_{\theta} p(D'|\theta) \cdot p(\theta|D) d\theta \\
&= \int_{\theta} \text{Mu}(N_{new_1}, N_{new_2}, \dots | \theta) \cdot \text{Dir}(\theta | N_{old_1} + \alpha_1, N_{old_2} + \alpha_2, \dots) \\
&= \frac{1}{B(\alpha + N_{old})} \binom{N_{new}}{N_{new_1}! \dots N_{new_k}!} \int_{\theta} \prod_k^k \theta_k^{N_{new_k}} \cdot \prod_k^k \theta_k^{\alpha_k + N_{old_k} - 1} d\theta \\
&= \frac{1}{B(\alpha + N_{old})} \binom{N_{new}}{N_{new_1}! \dots N_{new_k}!} \int_{\theta} \underbrace{\prod_k^k \theta_k^{\alpha_k + N_{new_k} + N_{old_k} - 1}}_{\text{normalization constant for Dir}(\vec{\alpha} + N_{new}^{\rightarrow} + N_{old}^{\rightarrow})} d\theta \\
&= \binom{N_{new}}{N_{new_1}! \dots N_{new_k}!} \frac{B(\alpha + N_{old} + N_{new})}{B(\alpha + N_{old})} \\
&= \frac{N'_0!}{\prod_k N'_k!} \cdot \frac{\prod_k \Gamma(\alpha_k + N_k + N'_k)}{\Gamma(\alpha_0 + N_0 + N'_0)} \cdot \frac{\Gamma(\alpha_0 + N_0)}{\prod_k \Gamma(\alpha_k + N_k)} \quad \text{where } \alpha_0 = \sum_{i=1}^k \alpha_i, N_0 = \sum_{i=1}^k N_i, \text{ and } N'_0 = \sum_{i=1}^k N'_i \\
&= \frac{N'_0!}{\prod_k N'_k!} \cdot \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N'_0)} \cdot \frac{\prod_k \Gamma(\alpha_k + N_k + N'_k)}{\prod_k \Gamma(\alpha_k + N_k)}
\end{aligned}$$

Notice that this looks like the formula we derived in exercise ??.

### 14 Posterior predictive for Dirichlet-multinomial

#### 14.1 a.

$$\begin{aligned}
p(X = j|D) &= E[\theta_j|D] \quad \text{Equation (3.51) in the textbook} \\
&= \frac{\alpha_j + N_j}{\alpha_0 + N} \\
&= \frac{10 + 260}{(10 \cdot 27) + 2000} \\
&= .119
\end{aligned}$$

#### 14.2 b.

We use the equation from ??.

$$p(D'|D, \alpha) = \frac{N'_0!}{\prod_k N'_k!} \cdot \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N'_0)} \cdot \frac{\prod_k \Gamma(\alpha_k + N_k + N'_k)}{\prod_k \Gamma(\alpha_k + N_k)}$$

For a single trial,  $N'_0 = N'_j = 1$ , so the first term is equal to 1. Now we examine the last term.

$$\begin{aligned}
p(x = j|D, \alpha) &= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N'_0)} \cdot \frac{\prod_{k \neq j} \Gamma(\alpha_k + N_k + N'_k)}{\prod_{k \neq j} \Gamma(\alpha_k + N_k)} \cdot \frac{\Gamma(\alpha_j + N_j + N'_j)}{\Gamma(\alpha_j + N_j)} \\
&= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N'_0)} \cdot \prod_{k \neq j} \frac{\Gamma(\alpha_k + N_k + 0)}{\Gamma(\alpha_k + N_k)} \cdot \frac{\Gamma(\alpha_j + N_j + N'_j)}{\Gamma(\alpha_j + N_j)} \\
&= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + N'_0)} \cdot 1 \cdot \frac{\Gamma(\alpha_j + N_j + N'_j)}{\Gamma(\alpha_j + N_j)} \\
&= \frac{\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_0 + N_0 + 1)} \cdot \frac{\Gamma(\alpha_j + N_j + 1)}{\Gamma(\alpha_j + N_j)} \\
&= \frac{\Gamma(\alpha_0 + N_0)}{(\alpha_0 + N_0)\Gamma(\alpha_0 + N_0)} \cdot \frac{(\alpha_j + N_j)\Gamma(\alpha_j + N_j)}{\Gamma(\alpha_j + N_j)} \\
&= \frac{\alpha_j + N_j}{\alpha_0 + N_0}
\end{aligned}$$

For independent samples

$$\begin{aligned}
p(x_{2001} = a, x_{2002} = p|D, \alpha) &= p(x_{2001} = a|D, \alpha) \cdot p(x_{2002} = p|D', \alpha) \\
&= \frac{10 + 100}{270 + 2000} \cdot \frac{10 + 87}{270 + 2001} \\
&= .0021
\end{aligned}$$