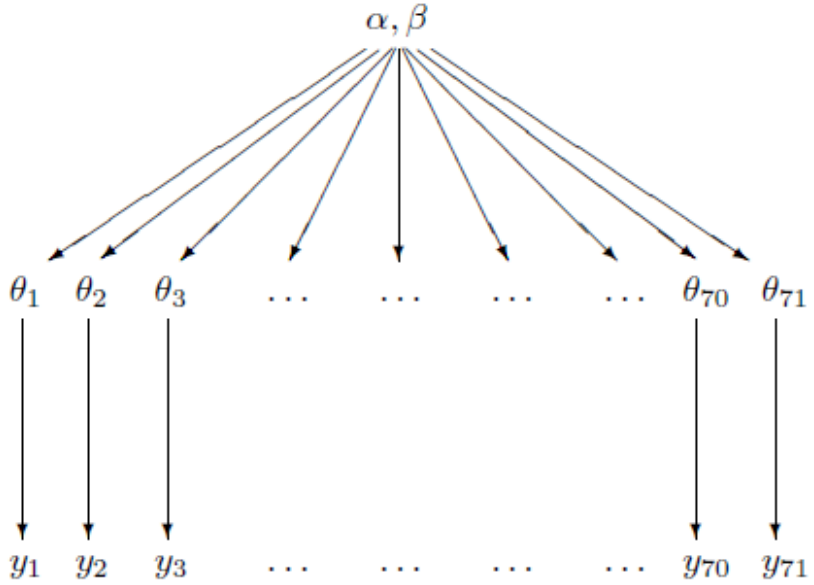
MAT703.Deber Seminario Investigación. Ejemplos y ejercicios Capítulo 5 "Bayesian Data Analysis"

Fausto Fabian Crespo Fernandez

Junio 2016

0.1. Ejemplo: estimando el riesgo de tumor en ratas

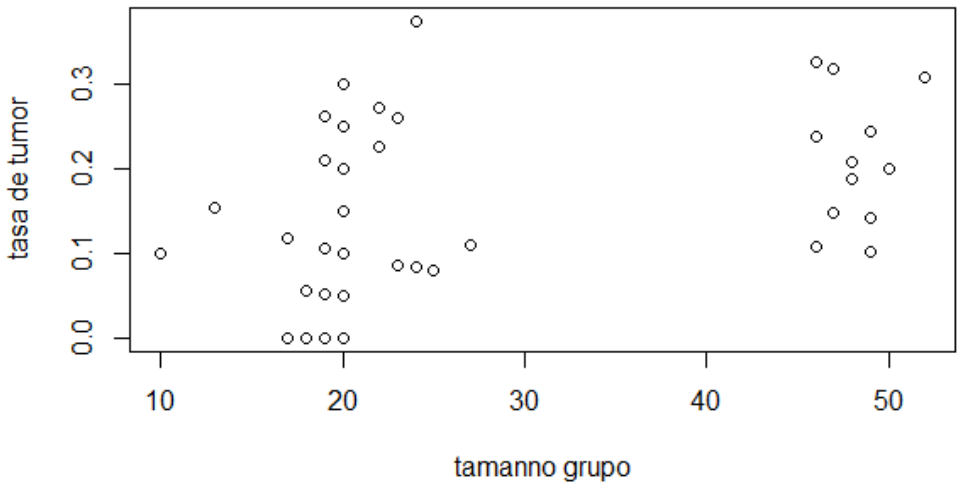
Supongamos que queremos estimar θ que es la probabilidad de tumor en población de ratas hembras que reciben dosis 0 de una droga(grupo control). La a priori para θ es $Beta(\alpha, \beta)$(conjugada).Tenemos 70 datos previos de grupos conde tenemos en número de muertes y_j y el numero de ratas en el grupo n_j y queremos modelar el experimento 71.
Se usa modelo jerárqico porque los valores θ_j de cada experiento sale de una distribución $Beta(\alpha, \beta)$ pero luego de tener el valor θ_j la distribución condicional $y_j|\theta_j$ es binomial:



El número de tumores y_j dado θ es $Bin(n_j, \theta)$partir de los datos históricos sabemos que θ sigue una $Beta(\alpha, \beta)$ y de los datos podemos calcular la media y la varianza de los valores de $\hat{\theta} = \frac{y_j}{n_j}$ con lo cual podemos hallar los valores α, β. En R
$datos = rat[-71,]$
$proporciones = datos[,1]/datos[,2]$
$esperado.theta = mean(proporciones)$
$varianza.theta = var(proporciones)$
$normaDiferencia <- function(vectorparametros)\{$
$alpha = vectorparametros[1]$

$beta = vectorparametros[2]$
$denominador = alpha + beta$
$return((alpha/denominador - esperado.theta)^2 + (alpha * beta/(denominador^2 *$
$(denominador + 1)) - varianza.theta)^2)$
$\}$
$library(optimx)$
$optimos = optimx(par = c(2, 80), normaDiferencia)$
$print(optimos)$
$alpha1 = optimos[1, 1]$
$beta1 = optimos[1, 2]$

Lo que da $(\alpha, \beta) = (1{,}356929, 8{,}62021)$ Luego asumiendo una a priori para θ de $Beta(\alpha, \beta)$ con los valores α, β anteriores obtenemos la a posteriori para θ es $Beta(\alpha + 4, \beta + 10) = Beta(5{,}356929, 18{,}62021)$ y la media a posteriori es $5{,}356929/(5{,}356929 + 18{,}62021)$. En R:
$alpha71 = alpha1 + 4$
$beta71 = beta1 + 10$
$theta71 = alpha71/(alpha71 + beta71)$
$print(theta71)$ Lo que da $\theta_{71} = 0{,}2234182$ que es menor que la proporción cruda $4/14 = 0{,}2857143$ lo que indica en que basado en la información anterior este último experimento tiene un valor elevado.
Los valores θ_i con $i = 1 \ldots 70$ provienen de una $Beta(\alpha, \beta$ y son independientes en la a priori, pero son dependientes en la a aposteriori y no deben ser analizados independientemente.
Si supieramos que ciertos lotes de experimentos fueron realizados en diferentes laboratorios podriamos modelar con intercambiabilidad parcial o sea un modelo jerárquico de dos niveles para modelar variación entre los laboratotios y dentro de los laboratorios.
Lo único que tenemos para distinguir entre los experimentos es el número de ratas en el experimento n_j y no parece razonable usar n_j para modelar tasa de tumor pero para verificar esto podríamos usar un modelo conjunto intercambiable de los pares $(n, y)_j$ y podemos plotear $\frac{y_j}{n_j}$ contra n_j.Esto por ejemplo es importante si suponemos que los experimentos con n_j mayores ocurrieron porque se suponía que θ_j era menor. En R:
$plot(datos[, 2], proporciones, xlab = "tamannogrupo", ylab = "tasadetumor")$



Se observa que no existe relación entre $\frac{y_j}{n_j}$ y n_j
El hecho de que los experiementos se hayan realizado en diferentes tiempos, diferentes ratas y quizás en diferentes laboratorios no invalida la intercambiabilidad Si los experimentos difieren implica que los θ_j son distrintos pero se puede asumir que provienen de una misma a priori. Además a falta de informacion lo más lógico es asumir intercambiailidad.
Tenemos $y_j \sim Bin(n_j, \theta_j)$ y $\theta_j \sim Beta(\alpha, \beta)$. La distribución conjunta a posteriori es:
$p(\theta, \alpha, \beta|y) \propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta)$
$\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}$
Dado α, β las componentes de θ tienen densidades a posteriori independientes de la forma $\theta_j^A(1-\theta_j)^B$ o sea $Betas$ y la conjunta a posteriori es
$p(\theta|\alpha, \beta|y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+n_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+n_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$
La marginal a posteriori de α, β es
$p(\alpha, \beta|y) = \frac{p(\theta,\alpha,\beta|y)}{p(\theta|\alpha,\beta|y)}$
$= p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+n_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$
que no se puede simplificar analíticamente pero se calcular con R para α, β específicos.
Por la falta de conocimiento escogemos una hiper apriori(la distribución de los hiperparámetros α, β)difusa.Primero se reparametriza en función de

$logit(\frac{\alpha}{\alpha+\beta}) = ln\frac{\alpha}{\beta}$ y $ln(\alpha+\beta)$ que son el logit de la media de la distribución Beta de θ y el logaritmo natural del "tamaño de muestra". Podemos asignar hiper ditribuciones a priori independientes para media a priori $\frac{\alpha}{\beta}$ y para $\alpha+\beta$ y se usa las transformaciones para llevar estos valores a la escala $(-\infty, \infty)$. Pero una distribución uniforme para estos nuevos hiperparámetros transformados da una a posteriori impropia ya que si $x = (\alpha, \beta)$ y $y = (ln\alpha/\beta, ln\alpha + \beta)$ y $g(x) = (g_1(x), g_2(x))$ con $g_1(x) = ln(\alpha/\beta)$ y $g_1(x) = ln(\alpha + \beta)$ entonces $g^{-1}(y) = (\frac{e^{y_1+y_2}}{1+e^{y_1}}, \frac{e^{y_2}}{1+e^{y_1}})$ y

$$p(y) = p(ln(\alpha/\beta), ln(\alpha + \beta)) = p(\alpha, \beta)|det \begin{pmatrix} \frac{d\alpha}{dy_1} & \frac{d\alpha}{dy_1} \\ \frac{d\beta}{dy_1} & \frac{d\beta}{dy_2} \end{pmatrix}| =$$

$p(\alpha, \beta)|det \begin{pmatrix} \frac{e^{y_1+y_2}}{(1+e^{y_1})^2} & \frac{e^{y_1+y_2}}{1+e^{y_1}} \\ -\frac{e^{y_1+y_2}}{(1+e^{y_1})^2} & \frac{e^{y_2}}{1+e^{y_1}} \end{pmatrix}| = p(\alpha, \beta)\frac{e^{y_1+2y_2}}{(1+e^{y_1})^2}$ de donde

$p(\alpha, \beta) = p(ln(\alpha/\beta), ln(\alpha + \beta))\frac{(1+e^{y_1})^2}{e^{y_1+2y_2}}$

$= p(ln(\alpha/\beta), ln(\alpha + \beta))(1/(\alpha\beta))$ $(*)$

$\propto 1/(\alpha\beta)$

y entonces la a posteriori

$p(\alpha, \beta|y) \propto (1/(\alpha\beta))\prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+n_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$ es impropia.

Una a hiper priori difusa uniforma es $y = (\frac{\alpha}{\alpha+\beta}, (\alpha + \beta)^{-1/2})$ tenemos que :

$p(\alpha, \beta) = p(\frac{\alpha}{\alpha+\beta}, (\alpha + \beta)^{-1/2})|det \begin{pmatrix} \frac{dy_1}{d\alpha} & \frac{dy_1}{d\beta} \\ \frac{dy_2}{d\alpha} & \frac{dy_2}{d\beta} \end{pmatrix}|$

$\propto |det \begin{pmatrix} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ (-1/2)(\alpha + \beta)^{-3/2} & (-1/2)(\alpha + \beta)^{-3/2} \end{pmatrix}| = (1/2)(\alpha + \beta)^{-5/2}$

y usando el resultado $(*)$ $p(ln(\alpha/\beta), ln(\alpha+\beta)) = p(\alpha, \beta)\alpha\beta = \alpha\beta(\alpha+\beta)^{-5/2}$

El gráfico de contorno de la marginal posterior se puede obtener en R:

$y1 = seq(-2{,}5, -1, 0{,}01)$

$y2 = seq(1{,}3, 3, 0{,}01)$

$marginal.posterior = function(y1, y2)\{$

$alpha2 = exp(y1 + y2)/(1 + exp(y1))$

$beta2 = exp(y2)/(1 + exp(y1))$

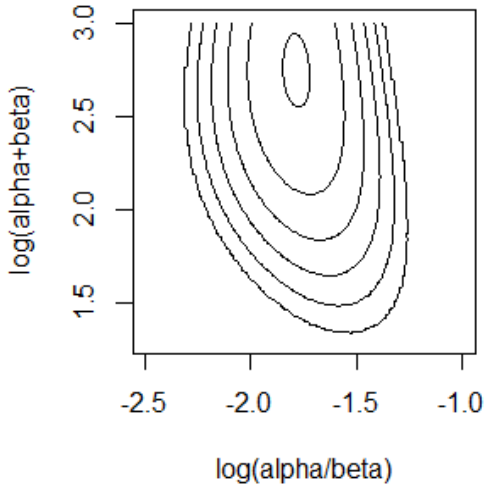
$prod = 1$

$for(iin1 : 71)\{$

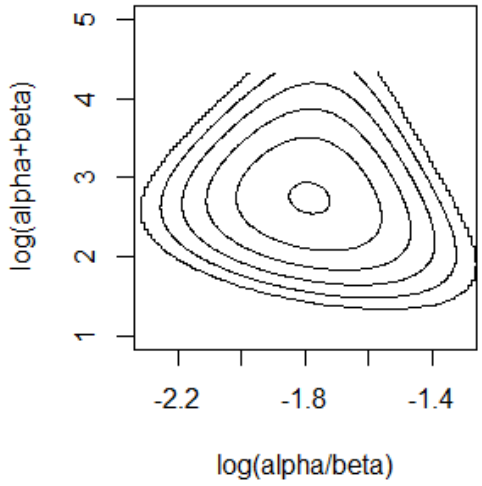
$prod = prod$

$* (gamma(alpha2 + beta2) * gamma(alpha2 + rat[i, 1]) * gamma(beta2 + rat[i, 2] - rat[i, 1]))/(gamma(alpha2) * gamma(beta2) * gamma(alpha2 + beta2 + rat[i, 2]))$

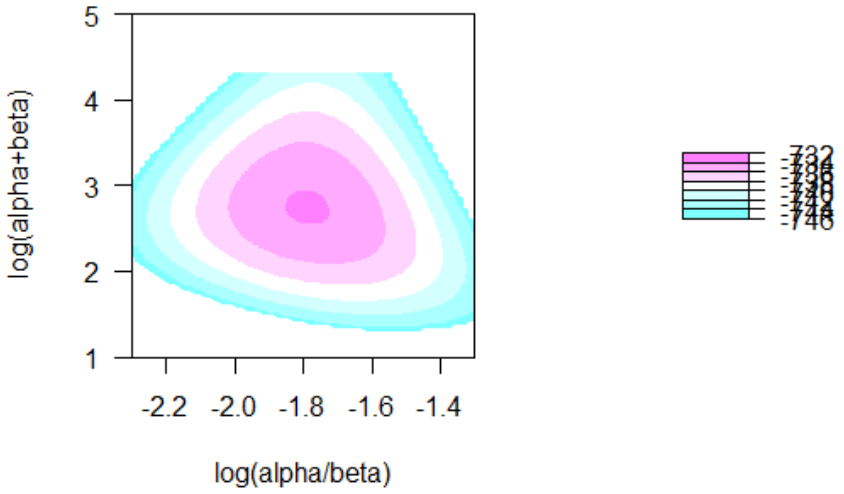
$$
\}
$$
$$
return(log(alpha2 * beta2 * prod * (alpha2 + beta2)^(-5/2)))
$$
$$
\}
$$
$$
z = outer(y1, y2, marginal.posterior);
$$
$$
par(pty = "s")
$$
$$
contour(y1, y2, z, nlevels = 6, drawlabels
$$
$$
= FALSE, xlab = "log(alpha/beta)", ylab = "log(alpha + beta)",
$$
$$
ylim = c(1,3, 3),
$$
$$
xlim = c(-2,5, -1))
$$



Y se ve que la moda no esta lejos de la estimación puntual de (1,8, 2,3) o
$(\alpha, \beta) = (1,4, 8,6)$
Si escogemos el intervalo $(ln(\alpha/\beta), ln(\alpha + \beta)) = ([-2,3, -1,3], [1, 5])$ En R :
$$
par(pty = "s")
$$
$$
contour(y1, y2, z, nlevels = 6, drawlabels = FALSE, xlab = "log(alpha/beta)", ylab =
$$
$$
"log(alpha + beta)",
$$
$$
ylim = c(1, 5),
$$
$$
xlim = c(-2,3, -1,3))
$$



O con :
$y1 = seq(-2{,}5, -1, (1{,}5)/100)$
$y2 = seq(1, 5, 4/100)$
$z = outer(y1, y2, marginal.posterior)$;
$par(pty = "s")$
$filled.contour(y1, y2, z, nlevels = 6, drawlabels$
$= FALSE, xlab = "log(alpha/beta)", ylab = "log(alpha + beta)",$
$ylim = c(1, 5),$
$xlim = c(-2{,}3, -1{,}3)$
)



El gráfico es simétrico alrededor de (1,75, 2,8) que corresponde a $(\alpha, \beta) =$ (2,4, 14,0) que difiere un poco del estimado crudo anterior

0.2. Ejemplo intercambiabilidad de muestreo

Supongamos que hemos seleccionado 8 estados de los Estados Unidos, y en cada uno medimos la tasa de divorcio y_i por 1000 habitantes en 1981.Como no podemos distinguir un estado del otro podemos usar modelo intercambiable.La a priori para la tasa de divorcios puede ser alguna distribución en $[0, 1]$ como $Beta$,$logit - normal$,etc. Ahora seleccionamos aleatoriamente 7 de los 8 estados y las tasas de divorcio fueron 5,8, 6,6, 7,8, 5,6, 7,0, 7,1, 5,4, basado en esto una a posteriori predictiva para el estado restante y_8 sería una distribución centrada en la media de los datos 6,471429 y el 95 % de los datos en [4,715744, 8,227113 (o sea aumienndo una aproximación normal en $[media - 1{,}96\sqrt{(var)}, media + 1{,}96\sqrt{(var)}]$. Los índices son intercambiables sin afectar la distribución conjunta, pero los y_j no son independientes a posteriori porque dados los 7 datos se espera que el octavo sea similar a los anteriores.SI los 7 estados son Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah and Wyoming pero aún sin saber que tasa de divorcio co-

rresponde a que estado, se espera que Utah que es mayormente de población mormona tenga tasa inferior y Nevada tenga tasa superior al resto. Estos valores se pueden modelar con una distribucón de colas pesadas.Al ver los valores y observar que son cercanos podemos asumir que el estado restante es Utah o Nevada y esto lleva a una distribución a posteriori bimodal o trimodal(o sea de varias máximos). Todavía es intercambiable pues no se sabe que estado corresponde a que tasa.Si se sabe que el estado no selecionado es Nevada ya no se puede escoger una a apriori intercambiable pues se espera que la tasa en Nevada sea mayor que en los otros 7 estados.Las observaciones pueden no ser intercambiables pero pueden ser condicionalmente o parcialmente intercambiables. Esto esporque se pueden agrupar las observaciones podemos usar un modelo jerárquico o sea cada grupo tienen sus propios parámetros y si los grupos son intercambiables podemos asumir que los parámetros de los grupos vienen de una sola a priori.En este caso podemos asumir 3 grupos: Utah, Nevada y el resto. También puede ocurrir que y_i no es intercambiable pero tenemos la información adicional x_i y (x_i, y_i) son intercambiables, con lo que podemos hacer un modelo conjunto para (x_i, y_i) o un modelo condional $y_i|x_i$.
Si supieramos la tasa de divorcio del año anterior en los 8 estados x_j pero no que indice corresponde a que estado, entonces podríamos identificar los valores de y_j pero los probabilidad conjunta a priori $p(x_j, y_j)$ sería la misma para cada estado. Para los estdo que tienen la misma tasa de divorcio el año anterior podemos agrupar esos datos y asumir intercambiabilidad parcial de los distintos grupos o también podemos asumir intercambiailidad condicional y agregar x_j como covariante en la regresión.

0.3. Ejercicio 5.1

1. Exchangeability with known model parameters: For each of the following three examples, answer: (i) Are observations y_1 and y_2 exchangeable? (ii) Are observations y_1 and y_2 independent? (iii) Can we act *as if* the two observations are independent?
   (a) A box has one black ball and one white ball. We pick a ball y_1 at random, put it back, and pick another ball y_2 at random.
   (b) A box has one black ball and one white ball. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2.
   (c) A box has a million black balls and a million white balls. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2 at random.

Solución

- (a) Si, y_1 y y_2 son intercambiables e independientes
- (b) y_1 y y_2 no son intercambiables y tampoco independientes: y_2 depende de y_1 y no podemos actuar como si fueran independientes
- (c) y_1 y y_2 no son intercambiables y tampoco independientes: y_2 depende de y_1 pero en este caso podemos actuar como si fueran independientes porque el número de bolas blancas y negras es grande 1 millón

0.4. Ejercicio 5.2

2. Exchangeability with known model parameters: For each of the following three examples, answer: (i) Are observations y_1 and y_2 exchangeable? (ii) Are observations y_1 and y_2 independent? (iii) Can we act *as if* the two observations are independent?
   (a) A box has n black and white balls but we do not know how many of each color. We pick a ball y_1 at random, put it back, and pick another ball y_2 at random.
   (b) A box has n black and white balls but we do not know how many of each color. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2 at random.
   (c) Same as (b) but we know that there are many balls of each color in the box.

Solución

(a) Si, y_1 y y_2 son intercambiables e independientes

(b) y_1 y y_2 no son intercambiables(la distribución conjunta de y_1 y y_2 no es invariante con permutaciones en los indices) y tampoco independientes: y_2 depende de y_1(la probabilidad de y_2 es blanca o negra depende de lo que fue y_1) y no podemos actuar como si fueran independientes

(c) y_1 y y_2 no son intercambiables y tampoco independientes: y_2 depende de y_1 pero en este caso podemos actuar como si fueran independientes porque el número de bolas blancas y negras es grande (el efecto de una bola no es importante cuando el número de bolas es grande)

0.5. Ejercicio 5.3

3. Hierarchical models and multiple comparisons:
   (a) Reproduce the computations in Section 5.5 for the educational testing example. Use the posterior simulations to estimate (i) for each school j, the probability that its coaching program is the best of the eight; and (ii) for each pair of schools, j and k, the probability that the coaching program in school j is better than that in school k.
   (b) Repeat (a), but for the simpler model with τ set to ∞ (that is, separate estimation for the eight schools). In this case, the probabilities (i) and (ii) can be computed analytically.
   (c) Discuss how the answers in (a) and (b) differ.
   (d) In the model with τ set to 0, the probabilities (i) and (ii) have degenerate values; what are they?

Solución

- (a) El número de distribuciones a posteriori fue de 1000 y el intervalos de τ fue de $[0, 40]$. En las simulaciones de [2] se obtuvo

		Pr(better than school)							
School	Pr(best)	A	B	C	D	E	F	G	H
A	0.25	–	0.64	0.67	0.66	0.73	0.69	0.53	0.61
B	0.10	0.36	–	0.55	0.53	0.62	0.61	0.37	0.49
C	0.10	0.33	0.45	–	0.46	0.58	0.53	0.36	0.45
D	0.09	0.34	0.47	0.54	–	0.61	0.58	0.37	0.47
E	0.05	0.27	0.38	0.42	0.39	–	0.48	0.28	0.38
F	0.08	0.31	0.39	0.47	0.42	0.52	–	0.31	0.40
G	0.21	0.47	0.63	0.64	0.63	0.72	0.69	–	0.60
H	0.12	0.39	0.51	0.55	0.53	0.62	0.60	0.40	–

- (b) En el modelo con $\tau \to \infty$, los efectos de la escuela $\theta_j$son independientes en la distribución a posteriori con $\theta_j|y \sim N(y_j, \sigma_j^2)$ y entonces $Pr(\theta_i > \theta_j|y) = \Phi((y_i - y_j)/\sqrt{(}\sigma_i^2 + \sigma_j^2))$. La probabilidad de que θ_i es el más grande de los efectos de las escuelas se puede expresar como $Pr(\theta_i es el más grande) = \int_{-\infty}^{\infty} \prod_{j \neq i} \Phi(\frac{\theta_i - y_j}{\sigma_j}) \phi(\theta_i|y_i, \sigma_i) d\theta_i$ La integral se puede evaluar numéricamente, los resultados de 2] fueron :

		Pr(better than school)							
School	Pr(best)	A	B	C	D	E	F	G	H
A	0.556	–	0.87	0.92	0.88	0.95	0.93	0.72	0.76
B	0.034	0.13	–	0.71	0.53	0.73	0.68	0.24	0.42
C	0.028	0.08	0.29	–	0.31	0.46	0.43	0.14	0.27
D	0.034	0.12	0.47	0.69	–	0.70	0.65	0.23	0.40
E	0.004	0.05	0.27	0.54	0.30	–	0.47	0.09	0.26
F	0.013	0.07	0.32	0.57	0.35	0.53	–	0.13	0.29
G	0.170	0.28	0.76	0.86	0.77	0.91	0.87	–	0.61
H	0.162	0.24	0.58	0.73	0.60	0.74	0.71	0.39	–

- (c) El modelo con $\tau \to \infty$ tiene más probabilidades extremas, Por ejemplo en la primera columna la probabilidad que la escuela A sea la mejor se incrementa de 0.25 a 0.56. También esto es cierto en comparaciones por pares de escuelas, por ejemplo la probabilidad de que la escuela A sea mejor que la escuela E bajo modelo jerárquico completo es 0.73, mientras que es 0.95 en el modelo con $\tau \to \infty$. Las respuestas más conservativas en el caso del modelo jerárquico completo reflejan la evidencia en los datos de que los programas de entrenamiento son casi iguales en efectividad. También la escuela preferida en una comparación entre dos escuelas puede cambiar, por ejemplo la escuela E es mejor que la escuela C cuando $\tau \to \infty$ mientras que la escuela C es mejor que la escuela E cuando promediamos por la distribución a posteriori de τ. Esto ocurre porque los errores estándares σ_j difieren .
- (d) Si $\tau = 0$ todos los efectos en las escuelas son iguales y no hay escuelas mejor que otras.

0.6. Ejercicio 5.6

6. Exchangeable models:

EXERCISES 135

(a) In the divorce rate example of Section 5.2, set up a prior distribution for the values $y_1, \ldots, y_8$ that allows for one low value (Utah) and one high value (Nevada), with independent and identical distributions for the other six values. This prior distribution should be *exchangeable*, because it is not known which of the eight states correspond to Utah and Nevada.

(b) Determine the posterior distribution for y_8 under this model given the observed values of $y_1, \ldots, y_7$ given in the example. This posterior distribution should probably have two or three modes, corresponding to the possibilities that the missing state is Utah, Nevada, or one of the other six.

(c) Now consider the entire set of eight data points, including the value for y_8 given at the end of the example. Are these data consistent with the prior distribution you gave in part (a) above? In particular, did your prior distribution allow for the possibility that the actual data have an outlier (Nevada) at the high end, but no outlier at the low end?

Solución

0.7. Ejercicio 5.7

7. Continuous mixture models:

(a) If $y|\theta \sim \text{Poisson}(\theta)$, and $\theta \sim \text{Gamma}(\alpha, \beta)$, then the marginal (prior predictive) distribution of y is negative binomial with parameters α and β (or $p = \beta/(1+\beta)$). Use the formulas (2.7) and (2.8) to derive the mean and variance of the negative binomial.

(b) In the normal model with unknown location and scale (μ, σ^2), the noninformative prior density, $p(\mu, \sigma^2) \propto 1/\sigma^2$, results in a normal-inverse-χ^2 posterior distribution for (μ, σ^2). Marginally then $\sqrt{n}(\mu - \overline{y})/s$ has a posterior distribution that is t_{n-1}. Use (2.7) and (2.8) to derive the first two moments of the latter distribution, stating the appropriate condition on n for existence of both moments.

Solución

0.8. Bibliografía

[1] http://www.stat.columbia.edu/ gelman/book/solutions2.pdf
[2] http://www.stat.columbia.edu/ gelman/book/solutions3.pdf