

Modelos Bayesianos Jerárquicos

José Luis Baroja

Universidad Nacional Autónoma de México

Nota del Autor

José Luis Baroja, Facultad de Psicología, Universidad Nacional Autónoma de México

Trabajo realizado gracias al Programa de Apoyo a Proyectos de Investigación e Innovación

Tecnológica (PAPIIT) de la Universidad Nacional Autónoma de México, proyecto IN307214.

Correo electrónico: j.luis.baroja@gmail.com

Modelos Bayesianos Jerárquicos

Supón que a lo largo del semestre registramos el número de canciones que cierta persona baila en diferentes fiestas. Al final del curso el registro puede verse así:

$$c : 5 \ 3 \ 4 \ 7 \ 4 \ 5 \ 5 \ 4 \ 5 \ 7 \ 7 \ 6$$

en donde cada posición c_f del vector c corresponde con el número de canciones que la persona bailó en la fiesta f .

Supón también que estamos interesados en utilizar los datos que recolectamos para averiguar qué tanto le gusta bailar a la persona en observación. Una aproximación que permite utilizar las observaciones registradas para aprender sobre nuestro sujeto consiste en construir un modelo que especifica cómo se relaciona el rasgo *gusto por bailar* con el número de canciones que la persona bailó en cada fiesta del semestre.

En este capítulo mostraremos cómo podemos construir dicho modelo y cómo podemos extenderlo para aprender no sólo sobre una persona sino también sobre poblaciones de individuos. La aproximación y modelos que presentamos pueden utilizarse para inferir un amplio conjunto de características psicológicas en diferentes poblaciones de organismos.

El plan del capítulo es el siguiente: primero construiremos un modelo para inferir el gusto por bailar de nuestra primera persona. Después mostraremos cómo extender el modelo para aprender sobre varias personas. Posteriormente ampliaremos el modelo para varias personas de tal manera que también nos permita llegar a conclusiones sobre la población de la que las personas forman parte. Finalmente presentaremos una extensión adicional que permite aprender al mismo tiempo no sólo sobre cada persona y sobre la población de personas, sino también

sobre cada fiesta y sobre la población de fiestas.

Para inferir el gusto por bailar de la primera persona comenzaremos con un supuesto central: Asumiremos que la probabilidad de bailar cierto número de canciones c_f es una función de cierto valor λ que representa el gusto por bailar de la persona. Nuestro modelo debe capturar la intuición de que si una persona tiene una valor λ alto es altamente probable que baile muchas canciones. Por el contrario, si la persona tiene un valor λ bajo entonces deberíamos esperar que baile pocas canciones por fiesta. Una función que cumple estas características es la distribución Poisson. Esta distribución está construida para modelar conteos de eventos que ocurren con cierta tasa a lo largo del tiempo: si sabemos que los taxis pasan aproximadamente cada minuto y medio en cierta calle, por ejemplo, la distribución Poisson especifica qué tan probable es observar 0, 1, 2, o cada posible número de taxis en cierto intervalo de tiempo arbitrario. En este tipo de escenarios, entre menor sea la tasa de ocurrencia del evento, menor es el número esperado de ocurrencias por intervalo. La Figura 1 ilustra cómo se comporta esta distribución. Cada curva en la Figura 1 representa a una persona con un valor λ particular. A la primera persona le gusta bailar poco, lo cual representamos con un valor $\lambda = 2$. Dado que la persona tiene un valor λ pequeño es altamente probable que baile pocas canciones y poco probable que baile muchas, como lo indica la curva que une los puntos blancos. La segunda persona, representada por la curva de puntos grises, tiene un valor de $\lambda = 5$. Como consecuencia, es altamente probable que esta persona baile entre 3 y 6 canciones por fiesta y poco probable que baile más de 8 o 9. La última persona tiene un valor $\lambda = 10$. Los puntos negros representan dicha distribución y muestran que es altamente probable que la tercera persona baile un número elevado de canciones en la siguiente fiesta.

*** FIGURA 1 ***

En otras palabras, la distribución Poisson permite calcular qué valores de c_f son más probables dado un valor λ fijo. Podemos representar la relación probabilística entre cada posición c_f y λ con la siguiente notación:

$$Pr(c_f | \lambda) \sim \text{Poisson}(\lambda). \quad (1)$$

que indica que los posibles números de conteos c_f se distribuyen Poisson con parámetro λ .

Recapitulando, asumiremos que la variable aleatoria c_f , que representa el número de canciones que una persona baila en cierta fiesta, se distribuye de acuerdo con una distribución Poisson en función de λ , donde λ representa el gusto por bailar de la persona. Llamaremos a este primer conjunto de supuestos modelo M_0 .

La idea del modelo M_0 , que es la idea general de cualquier modelo probabilístico, es que, incluso si el parámetro λ es un valor fijo y estable, el número de canciones c_f que la persona baila en cada fiesta es variable. Sin embargo, la variación de c_f está sujeta a ciertas regularidades que dependen del valor λ : de acuerdo con la distribución Poisson, deberíamos esperar que una persona con un valor λ bajo baile pocas canciones por fiesta, y que una persona con un valor λ alto baile un número de canciones elevado.

Si conocemos el valor λ de una persona, el modelo M_0 permite calcular cuántas canciones es probable que la persona baile en la siguiente fiesta. El problema es que el gusto por bailar no es una característica observable y no podemos medir su valor directamente. Sin embargo, podemos inferir cuáles son los valores más probables de λ utilizando las observaciones que

recolectamos del individuo y la relación probabilística entre λ y cada observación que compone al vector c . Este proceso de inferencia está basado en una identidad conocida como *Regla de Bayes* (Griffiths & Yuille, 2006):

$$Pr(\lambda | c) = Pr(c | \lambda)Pr(\lambda) / Pr(c), \quad (2)$$

en donde $Pr(\lambda | c)$ se conoce como *distribución posterior* y especifica qué valores de λ son más probables dado el conjunto de observaciones que hemos recolectado. De acuerdo con la Regla de Bayes, la distribución posterior depende de la función de verosimilitud, de la distribución a priori, y de la verosimilitud marginal. La *función de verosimilitud*, $Pr(c | \lambda)$, especifica qué tan probable es observar el vector de datos c bajo cada posible valor de λ y en este caso corresponde con la distribución Poisson (ver Ecuación (1)). La *distribución a priori*, $Pr(\lambda)$, especifica qué tan probable es cada valor del parámetro λ antes de tomar en cuenta el vector de datos c y se utiliza para expresar supuestos previos sobre cada posible valor paramétrico en el modelo bajo estudio. Finalmente, la *verosimilitud marginal*, $Pr(c)$, especifica qué tan probable es observar el vector c bajo todos los posibles valores del parámetro λ , y está completamente determinada una vez que se conoce la función de verosimilitud y que se especifica una distribución a priori.

En otras palabras, la Regla de Bayes permite calcular qué valores de λ son más probables dado cierto conjunto de datos c , o bien, qué deberíamos concluir sobre el gusto por bailar de la persona después de observar el número de canciones que bailó en cada fiesta del semestre. En este sentido, la Regla de Bayes es una herramienta formal que especifica cómo actualizar nuestro conocimiento a priori sobre el gusto por bailar de la persona utilizando las observaciones que recolectamos en las fiestas.

Existen varios métodos para calcular distribuciones posteriores. En este capítulo aproximaremos las distribuciones posteriores de todos nuestros modelos utilizando un programa llamado JAGS (Just Another Gibbs Sampler; Plum- mer, 2003) y el paquete de cómputo estadístico R (R Core Team, 2015). Para inferir una distribución posterior con estos programas es necesario especificar los tres componentes principales del proceso de inferencia probabilística:

- Un conjunto de datos. Como primer ejemplo utilizaremos el vector de observaciones c con el que comenzamos el capítulo.
- Un modelo que especifica una relación probabilística entre los datos y ciertos parámetros o nodos desconocidos. Como primer ejemplo utilizaremos el modelo M_0 , en donde λ es el único parámetro desconocido.
- Una distribución de probabilidad sobre cada parámetro desconocido del modelo. Como primer ejemplo, asumiremos que cualquier valor de λ entre 0 y 50 es igualmente probable.

Una forma común de presentar estos tres componentes y las relaciones entre ellos es utilizando notación gráfica (Vincent, en prensa; Lee, 2008). La Figura 2 presenta al modelo M_0 escrito en esta notación: los nodos blancos representan variables o parámetros desconocidos; los grises a las variables observadas o conocidas. Las variables discretas son representadas como nodos cuadrados y las continuas como nodos circulares. Cuando un nodo tiene más de un elemento, como en el caso del vector c en este ejemplo, se coloca dentro de un plato para indicar que cada elemento de dicho nodo es una replicación independiente del mismo proceso probabilístico, o bien, que el número de canciones c_f que el participante bailó en cada fiesta es independiente del resto de las fiestas, aunque siempre es aleatorio y depende de un parámetro λ constante que caracteriza a la persona. Aparte de la relación entre cada elemento c_f y el nodo

desconocido λ , el modelo incluye la distribución prior sobre λ , que captura los supuestos iniciales sobre este parámetro antes descritos.

*** FIGURA 2 ***

Una vez especificado el conjunto de datos y el modelo probabilístico, el paso siguiente para hacer inferencia Bayesiana consiste en traducir el modelo gráfico a código y dejar que los programas calculen las distribuciones posteriores del modelo¹. El resultado que JAGS y R devuelven es una serie de muestreros que aproximan la distribución posterior de cada nodo desconocido en el modelo. En el caso del modelo M_0 , la distribución posterior sobre λ calculada por JAGS se presenta en la Figura 3. La zona gris corresponde con la densidad² posterior sobre el nodo λ , que especifica qué tan probable es cada posible valor del parámetro dado el conjunto de datos c , de acuerdo con los datos, el modelo M_0 , y la Regla de Bayes. Podemos resumir cualquier distribución posterior utilizando diferentes medidas descriptivas. En la Figura 3 hemos incluido la *media posterior*, cuyo valor es igual a 5.25 y está señalada por la línea punteada, y el intervalo

¹ Los datos y todo el código necesario para implementar los análisis de este capítulo están disponibles en: <https://gist.github.com/JLBaroja/0d047481975c01c453a5>

² Al discutir los resultados de los modelos siguientes, utilizaremos los términos *distribución de probabilidad*, y *densidad* indistintamente. De igual manera, los términos *nodo* y *parámetro* son intercambiables.

de máxima densidad posterior (MDP) al 95%³. Este intervalo indica que podemos estar 95% seguros de que el valor del parámetro λ de la primera persona en observación se encuentra entre 4.02 y 6.69, de acuerdo este modelo.

*** FIGURA 3 ***

En resumen, el modelo M_0 supone que existe una variable latente λ que refleja el gusto por bailar de una persona. De acuerdo con este modelo, existe una relación probabilística entre λ y el número de canciones c_f que la persona puede bailar en una fiesta. Utilizando las observaciones que recolectamos a lo largo del semestre, el modelo M_0 , y la Regla de Bayes, calculamos los valores del gusto por bailar λ más probables de la primera persona en nuestro estudio.

Ahora extenderemos el modelo M_0 para inferir el gusto por bailar de varias personas. Como estamos interesados en aprender sobre varios individuos necesitamos obtener observaciones de cada uno. La Figura 4 presenta la cantidad de canciones que 15 personas bailaron en las 12 primeras fiestas del semestre.

³ Presentamos la zona de MDP con 95% de credibilidad para fomentar la comparación de los resultados obtenidos con métodos Bayesianos con los obtenidos con técnicas clásicas, en donde conclusiones con un “nivel de confianza al 95%” son comúnmente reportadas, aunque es posible calcular el intervalo correspondiente a cualquier nivel de credibilidad deseado utilizando los muestreos de las distribuciones posteriores que devuelven los programas.

Una inspección rápida de la Figura 4 sugiere que las personas son diferentes respecto al número de canciones que bailaron en cada fiesta. Los participantes 4 y 5, por ejemplo, bailaron pocas canciones en las fiestas del semestre, mientras que los sujetos 7 y 1 bailaron una cantidad de canciones elevada. Aunque cada persona bailó un número diferente de canciones en cada fiesta, la cantidad de canciones que cada persona bailó a lo largo del semestre parece relativamente estable, es decir, las personas que bailaron pocas canciones en las primeras fiestas también bailaron pocas canciones en las últimas, y quienes bailaron muchas canciones al inicio del semestre también bailaron muchas al final.

***** FIGURA 4 *****

Podemos extender el modelo M_0 para inferir el valor del gusto por bailar de cada persona p , que denotamos como λ_p . La extensión simplemente consiste en suponer que cada persona tiene un valor λ_p particular, potencialmente diferente al de las demás personas. Esta extensión, a la que llamaremos M_1 , aparece en notación gráfica en la Figura 5. Como puede apreciarse al examinar los modelos gráficos correspondientes, la única diferencia entre los modelos M_0 y M_1 es un nuevo plato que indexa personas. En palabras, el modelo M_1 conserva la misma relación probabilística entre cada nodo λ_p y la columna de la matriz c correspondiente a la persona p , pero a diferencia del modelo M_0 , el modelo M_1 supone que hay varias personas, cada una con un valor λ_p propio.

Los resultados del modelo M_1 aparecen en la Figura 6. En el panel superior de la figura

presentamos las distribuciones posteriores sobre el parámetro λ_p de cada participante. En tanto que el modelo M_1 asume que hay tantas personas como columnas en la matriz c , el modelo devuelve tantas distribuciones posteriores sobre λ como columnas en la matriz. Al examinar las distribuciones posteriores sobre los nodos λ_p podemos apreciar que el modelo M_1 concluye que, aunque hay algunos participantes que se parecen entre sí, la mayoría de los participantes son diferentes respecto a su gusto por bailar, lo cual se refleja en la variación de las diferentes distribuciones posteriores sobre λ_p . Al asumir que pueden existir diferencias individuales en el parámetro λ_p , el modelo M_1 permite identificar a los participantes extremos P4 y P7, que resaltamos en el panel superior.

En el panel central y en el inferior presentamos en detalle las distribuciones posteriores sobre λ_p de los participantes 4 y 7, quienes, de acuerdo con M_1 , tienen el menor y el mayor gusto por bailar, respectivamente. La línea punteada en cada distribución corresponde a la media posterior, y la línea gruesa nuevamente señala el intervalo de MDP. Las conclusiones de M_1 sobre los participantes 4 y 7 parecen consistentes con los datos de ambos individuos: estos participantes fueron quienes bailaron menos y más canciones durante el semestre, respectivamente.

*** FIGURA 6 ***

Una forma útil de evaluar la capacidad descriptiva y predictiva de un modelo consiste en analizar la *distribución posterior predictiva* del mismo. La distribución posterior predictiva es

una distribución de probabilidad que especifica los datos que el modelo espera observar con base en los que ya han sido observados. En la Figura 7 mostramos la distribución posterior predictiva del modelo M_1 . El tamaño de cada círculo negro corresponde exactamente con el número de canciones que la persona p bailó en la fiesta f (ver Figura 4). Esta representación visual permite identificar rápidamente a los participantes que bailan más, a los que bailan menos, y a los que bailan una cantidad de canciones intermedia, y también permite tener una idea clara del tamaño de las diferencias entre ellos.

Los círculos grises en la Figura 7 señalan el intervalo de MDP de la distribución posterior predictiva de M_1 sobre cada nodo c_{fp} ⁴. Para evaluar la capacidad descriptiva de M_1 podemos comparar cada círculo negro contra el círculo gris en cada posición c_{fp} . Desde la perspectiva Bayesiana, un modelo describe adecuadamente el conjunto de datos recolectados en la medida que los datos observados (círculos negros) se ubican dentro del intervalo de MDP de la distribución posterior predictiva (círculos grises). Al comparar los datos observados contra los esperados por M_1 aparecen algunas deficiencias importantes. Específicamente, parece que algunos datos de cada participante se ubican en algún extremo de los círculos grises. En el caso del participante 1, por ejemplo, el modelo describe adecuadamente las observaciones en las fiestas 6 y 9 porque los círculos negros se ubican en el centro de las zonas grises correspondientes. Sin embargo, M_1 espera ver que el participante 1 baile más canciones que las

⁴ Al implementar cualquier modelo dentro del marco Bayesiano, el resultado es una distribución posterior sobre cada nodo del modelo. En este caso, el modelo termina con una distribución completa sobre los posibles números de canciones bailadas para cada persona p en cada fiesta f , y por lo tanto es posible calcular diferentes características de cada una, como el intervalo de mayor densidad posterior.

de hecho observadas en las fiestas 7 y 8 porque los círculos negros se ubican en la frontera interior de las zonas grises en esas fiestas, y espera ver menos que las observadas en las fiestas 10, 11 y 12 porque en estas fiestas el círculo negro se ubica en la frontera exterior de la distribución posterior predictiva. Casos similares aparecen en la mayoría de participantes. En otras palabras, aunque en promedio las predicciones de M_1 se ajustan a las observaciones de cada persona, el modelo espera ver más canciones bailadas que las observadas en algunas fiestas, y espera ver menos canciones bailadas que las observadas en otras.

***** FIGURA 7 *****

Un problema todavía más grave del modelo M_1 es su limitada capacidad predictiva. En la Figura 7 hemos incluido una columna adicional, p_n , en la que presentamos el número de fiestas bailadas que el modelo M_1 espera ver en un participante nuevo. Las distribuciones posteriores predictivas de dicha columna sugieren que M_1 considera igualmente probable que el participante nuevo baile pocas o muchas canciones, incluso en rangos que ninguno de los participantes observados ha presentado hasta el momento. El tamaño de los círculos grises en la columna p_n , en relación a la escala de la figura, sugiere que el modelo considera probable que el participante no observado hubiera bailado desde 0 hasta 30 o quizás 40 canciones en cada fiesta del semestre (ver Figura 4). ¿Por qué deberíamos esperar que un nuevo participante baile 30 o 40 canciones cuando ninguno de los observados ha bailado más de 20? M_1 hace esta extraña predicción porque, aunque ha aprendido algo sobre cada participante, no ha aprendido nada sobre *la población* de participantes. Como consecuencia, cuando M_1 tiene que predecir cómo se

comportará una persona nueva no puede utilizar el conocimiento que ha ganado sobre todas las personas que ha observado anteriormente y la predicción resulta poco informativa. Idealmente, un buen modelo debería aprender sobre cada persona y también sobre la población de personas para predecir adecuadamente cómo lucirá un sujeto no observado en las situaciones observadas e idealmente también en situaciones nuevas (p. ej., la siguiente fiesta).

Una posible extensión que permite aprender sobre individuos y poblaciones de individuos al mismo tiempo, consiste en suponer que, aunque cada persona tiene un valor λ_p particular, los valores λ_p de todas las personas provienen de una población común. La notación gráfica de este nuevo modelo, denominado M_2 , se presenta en la Figura 8.

*** FIGURA 8 ***

Similar a sus predecesores, el modelo M_2 conserva el supuesto central que relaciona cada nodo λ_p con las observaciones c_{fp} correspondientes, y también supone de que cada persona tiene un valor λ_p propio. Sin embargo, M_2 adicionalmente supone que todos los parámetros λ_p provienen de una distribución poblacional común, caracterizada por los parámetros μ^λ y σ^λ , que corresponden con la media y la desviación estándar poblacionales⁵, respectivamente. Este

⁵ Utilizamos la distribución Lognormal como distribución jerárquica porque los valores de λ_p tienen que ser forzosamente mayores que cero para garantizar consistencia con la distribución Poisson que depende de ellos. Esta restricción vuelve inadecuado modelar la variación en λ_p con

supuesto adicional vuelve al modelo M_2 un *modelo jerárquico*.

En general, un modelo jerárquico asume que los nodos desconocidos en cierto nivel provienen de una distribución definida en un nivel superior. Las extensiones jerárquicas pueden incluir varios niveles y, como mostraremos más adelante, pueden definirse no sólo respecto a participantes sino también respecto a estímulos o condiciones experimentales.

La Figura 9 permite comparar los datos observados contra la distribución posterior predictiva del modelo M_2 . En la figura podemos observar que el ajuste de M_2 es similar al ajuste de M_1 , en el sentido de que ambos modelos muestran una relación similar entre los datos observados y sus predicciones. Sin embargo, la predicción que M_2 hace sobre un participante nuevo parece más razonable que la del modelo anterior. M_2 puede hacer una predicción sensible sobre el nuevo participante porque tiene información sobre cómo se comporta la población de participantes observados y puede usar dicha información para inferir cuál será el valor λ_p de un participante nuevo proveniente de la misma población de gustos por bailar.

*** FIGURA 9 ***

También podemos examinar las conclusiones de M_2 sobre el gusto por bailar de cada participante en la muestra. La Figura 10 muestra estos resultados. En el panel superior

distribuciones definidas sobre todos los números reales (p. ej., la distribución Normal) (ver Limpert et al., 2001).

nuevamente presentamos las distribuciones posteriores sobre λ_p de cada participante. A primera vista parece que el modelo M_2 concluye algo similar al modelo M_1 : la mayoría de los participantes tienen valores λ_p diferentes, distribuidos aproximadamente en el mismo rango de valores bajo ambos modelos. Aparte, el modelo M_2 también identifica a los participantes 4 y 7 como participantes extremos.

Sin embargo, al estudiar en detalle las conclusiones de M_2 sobre los participantes extremos aparece una diferencia importante entre ambos modelos. En el panel central y en el inferior presentamos nuevamente las distribuciones posteriores sobre λ_p de los participantes 4 y 7, respectivamente. En cada panel, la línea de densidad negra identifica a la distribución posterior inferida por M_2 y la gris a la densidad posterior inferida por M_1 (ver Figura 6). En el caso del participante 4, el modelo M_2 infiere una distribución posterior sobre valores de λ_4 mayores respecto del modelo M_1 : tanto la media posterior como los intervalos de la zona de MDP de λ_4 de acuerdo con M_2 aparecen recorridos a la derecha respecto de los de M_1 . En otras palabras, aunque el modelo M_2 también concluye que el participante 4 tiene el gusto por bailar menor en la muestra de participantes, el valor de λ_4 inferido por M_2 no es tan pequeño como el inferido por M_1 . Algo similar ocurre en el participante 7, pero en sentido contrario: la distribución posterior sobre λ_7 calculada por M_2 aparece recorrida a la izquierda con respecto a la de M_1 , es decir, aunque al participante 7 le gusta bailar más que al resto de la muestra, el modelo M_2 concluye que el gusto por bailar de este participante no es tan grande como sugiere el modelo anterior.

Es importante resaltar que los modelos M_1 y M_2 llegan a conclusiones diferentes sobre los participantes extremos incluso cuando ambos modelos observaron exactamente los mismos datos de cada participante. Si ambos modelos observan los mismos datos de cada participante, ¿por

qué llegan a conclusiones diferentes? La razón es que en el modelo M_1 la única información relevante para inferir cada nodo λ_p son las observaciones del participante correspondiente, mientras que bajo el modelo M_2 los valores λ_p inferidos en cada participante dependen de las observaciones del participante y también de las observaciones de todos los participantes en la muestra. Cuando M_2 calcula el gusto por bailar del participante 4, por ejemplo, concluye algo como: *las observaciones de P4 sugieren que este participante tiene un gusto por bailar pequeño, pero dado que el participante proviene de una población de individuos que tienen un gusto por bailar intermedio, debería creer que su gusto por bailar no es tan pequeño después de todo*, y algo similar ocurre respecto al participante 7 en la dirección opuesta. En general, en cualquier modelo jerárquico las conclusiones sobre cada sujeto, cada ítem o cada condición experimental dependen no sólo de las observaciones asociadas directamente a cada elemento sino también de las observaciones de la población correspondiente. Este efecto es una característica importante de los modelos jerárquicos y se conoce como *contracción jerárquica* (Rouder et al., 2017).

*** FIGURA 10 ***

En el modelo M_2 los parámetros μ^λ y σ^λ corresponden con la media y la desviación estándar de la distribución poblacional de gustos por bailar. Podemos examinar las distribuciones posteriores sobre ambos parámetros para averiguar qué valores son más probables en cada uno. La Figura 11 presenta las distribuciones posteriores sobre los parámetros poblacionales inferidas por el modelo M_2 .

***** FIGURA 11 *****

Recapitulando, el modelo M_2 asume que las variables latentes λ_p , que corresponden con el gusto por bailar de cada persona en la muestra, provienen de la misma distribución poblacional. Este supuesto permite a M_2 inferir los valores paramétricos poblacionales y utilizar dicha información para predecir cuántas canciones hubiera bailado otro participante proveniente de la misma población en cada fiesta del semestre. La capacidad de predecir a un nuevo participante es una ventaja importante de M_2 respecto de su antecesor.

Sin embargo, tanto M_1 como M_2 comparten una deficiencia descriptiva más sutil. En concreto, aunque las predicciones de los dos modelos se acercan a los datos observados en cada participante, en algunas fiestas ambos modelos esperan que los participantes bailen más canciones que las que de hecho bailaron (p. ej., en la fiesta 5), mientras que en otras los dos modelos esperan ver que los participantes bailen menos de lo que bailaron (p. ej., fiesta 10; ver Figuras 7 y 9). En las secciones siguientes sugerimos cómo mejorar estas deficiencias.

Los modelos que hemos construido y evaluado hasta el momento comparten un supuesto central: suponen que la cantidad de canciones que cada persona bailó en cada fiesta sólo depende del gusto por bailar de la persona. Aunque este supuesto es razonable, parece incompleto, como lo sugiere una inspección detallada de ciertas tendencias en los datos recolectados. En particular, parece que hay fiestas en las que todas las personas tienden a bailar más canciones. La fiesta 7 y la 4 son ejemplos de este tipo de fiestas. Por el contrario, en otras fiestas la mayoría de la

muestra de participantes bailó pocas canciones, como en la fiesta 8 o en la 5. Estos patrones sugieren que la cantidad de canciones que una persona baila en una fiesta no sólo depende del gusto por bailar de la persona sino también de alguna característica de la fiesta, como el número de canciones que tocaron en ella. Si en una fiesta tocan pocas canciones esperamos observar que una persona baile pocas canciones incluso si tiene un gusto por bailar alto; o bien, si en una fiesta tocan muchas canciones esperamos observar que incluso las personas a las que les gusta bailar poco bailen más canciones que de costumbre.

Los modelos que presentamos a continuación formalizan estas intuiciones y las ponen a prueba.

El modelo M_3 , que aparece en notación gráfica en la Figura 12, supone que la cantidad de canciones que la persona p bailó en la fiesta f , c_{fp} , depende del gusto por bailar de la persona, esta vez anotado como θ_p , y también del número de canciones que tocaron en la fiesta, n_f .

*** FIGURA 12 ***

Es importante destacar que tanto el gusto por bailar de la persona θ_p como el número de canciones que tocaron en cada fiesta, n_f , son nodos no observados, y que afectan condicionalmente a la variable observada c_{fp} . Por lo tanto, podemos inferir el valor de ambos parámetros desconocidos utilizando el mismo conjunto de herramientas de inferencia que hemos presentado previamente. Como el parámetro θ_p es una característica de la persona que suponemos se mantiene constante entre fiestas, el nodo θ_p aparece indexado en el plato de

personas pero fuera del plato de fiestas. Por su parte, el nodo n_f está dentro del plato de fiestas pero fuera del plato de personas para reflejar el supuesto de que la cantidad de canciones que tocaron en la fiesta f es una característica propia de cada fiesta, constante para todas las personas.

De acuerdo con M_3 la cantidad de canciones que la persona p bailó en la fiesta f es una variable aleatoria con distribución Binomial con parámetros θ_p y n_f :

$$Pr(\text{cfp} | \theta_p, n_f) \sim \text{Binomial}(\theta_p, n_f). \quad (3)$$

Elegimos la distribución Binomial porque esta distribución tiene dos propiedades que parecen reflejar el escenario de las fiestas. Primero, entre más grande es el parámetro n en una distribución Binomial el valor esperado de la variable aleatoria aumenta, al margen del valor del parámetro θ . Segundo, entre más grande es el valor del parámetro θ el valor esperado de la variable aleatoria también aumenta, al margen del valor de n . En tanto que el parámetro θ en una distribución Binomial está restringido al rango $0 \leq \theta \leq 1$ necesitamos especificar una distribución jerárquica que esté definida para variables aleatorias en dicho rango. La distribución Beta cumple con esta característica. Como hemos hecho explícito en el modelo gráfico, en M_3 parametrizamos la distribución Beta en términos de la media μ^θ y desviación estándar σ^θ de la población de personas. Utilizamos distribuciones prior uniformes sobre los nodos del modelo M_3 ; en el caso de los nodos n_f y σ^θ los límites de la distribución prior fueron elegidos arbitrariamente, buscando que cubrieran rangos intuitivamente razonables, pero en el caso del nodo μ^θ el rango de valores válidos está restringido por la distribución que depende de dicho nodo: como la distribución Beta está definida únicamente dentro del intervalo $(0,1)$, los valores de la media de dicha distribución también tienen que ubicarse en dicho intervalo.

Los resultados del modelo M_3 se resumen en la Figura 13. Al comparar la distribución

posterior predictiva de M_3 contra los datos registrados observamos que esta vez los círculos negros se ubican en medio de las zonas grises con mayor frecuencia, lo cual indica que en la mayoría de los casos la predicción de M_3 se acerca a las observaciones recolectadas de cada persona. Aparte, M_3 también parece predecir adecuadamente cómo se comportará un participante nuevo. M_3 predice razonablemente a un nuevo participante porque conserva la estructura jerárquica sobre participantes del modelo M_2 . Es decir, incluso si cambiamos la distribución jerárquica específica (de Log-normal a Beta, en este ejemplo), el hecho de suponer que todos los participantes provienen de una distribución poblacional común es suficiente para predecir el desempeño de un participante nuevo utilizando el conocimiento adquirido al observar a la muestra de personas.

En la misma Figura hemos incluido un renglón adicional que representa la siguiente fiesta del semestre. Un buen modelo sobre el fenómeno bajo estudio debería predecir no sólo cómo se comportará un participante nuevo en las fiestas conocidas, sino también cómo se comportarán los participantes conocidos en una fiesta nueva.

Al examinar la distribución posterior predictiva de M_3 en la fiesta nueva aparecen conclusiones sospechosas. En concreto, aunque M_3 predice que las personas que han bailado más canciones en las fiestas observadas también bailarán más en la fiesta nueva, la cantidad de canciones que M_3 espera ver en la fiesta nueva parece ingenua para ciertos participantes. En el caso del participante 7, por ejemplo, M_3 espera ver que este participante baile cualquier número de canciones entre 0 y 25, aproximadamente. ¿Por qué esperarías que un participante que siempre ha bailado más de 9 canciones en las fiestas del semestre baile 1 o 2 en la fiesta siguiente? O bien, ¿por qué el mismo participante bailaría hasta 25 canciones, si en ninguna fiesta ha bailado más de 20?

***** FIGURA 13 *****

Si planteamos estas preguntas a M_3 podemos intuir su respuesta: *aunque he aprendido sobre el participante 7 y sobre la población de la que este participante proviene, no puedo hacer buenas predicciones sobre cuánto bailará este participante en la siguiente fiesta porque no sé cuántas canciones tocarán en ella.* En otras palabras, aunque M_3 infiere cuántas canciones tocaron en cada fiesta observada, no puede utilizar dicho conocimiento para predecir el número de canciones que tocarán en una fiesta nueva. Para utilizar el conocimiento adquirido en las fiestas observadas y predecir adecuadamente cuántas canciones tocarán en la siguiente fiesta es conveniente suponer que todas las fiestas tienen algo en común, o bien, que provienen de la misma distribución jerárquica. El modelo M_4 , que presentamos en la Figura 14, es una extensión del modelo M_3 que adicionalmente supone que la cantidad de canciones que tocaron en cada fiesta, n_f , proviene de una distribución jerárquica.

***** FIGURA 14 *****

Cuando el modelo M_4 observa los datos que recolectamos infiere cuántas canciones tocaron en cada fiesta, de manera similar a M_3 , pero M_4 también infiere el valor del parámetro λ_n , que representa la media (y la varianza) de la distribución poblacional de fiestas. En

consecuencia, M_4 puede utilizar este conocimiento para predecir cuántas canciones tocarán en la fiesta siguiente (y por lo tanto cuántas de ellas bailará cada persona). Como resultado, las predicciones de M_4 sobre cómo lucirá una fiesta nueva, que presentamos en la Figura 15, parecen mucho más razonables que las de M_3 .

*** FIGURA 15 ***

Resumiendo, el suponer una distribución jerárquica sobre participantes y una distribución jerárquica sobre fiestas permite al modelo M_4 predecir cuánto hubiera bailado un participante nuevo en cada fiesta conocida, cuánto bailarán los participantes conocidos en una fiesta nueva, y cuánto bailará un participante nuevo en una fiesta nueva.

En este capítulo hemos desarrollado diferentes modelos que sirven para inferir características individuales y poblacionales con base en cierto conjunto de datos. Las extensiones jerárquicas respecto a personas y respecto a fiestas mejoran el ajuste y la capacidad predictiva de nuestros modelos, o bien, de nuestras explicaciones sobre el mundo.

Los modelos Bayesianos jerárquicos son una poderosa herramienta de inferencia que tiene un rango enorme de aplicación en psicología: lo único necesario para implementarlos es contar con un modelo que especifique una relación probabilística entre cierto rasgo psicológico y cierto conjunto de observaciones. Por ejemplo, si suponemos que la probabilidad de recordar un ítem depende de qué tan buena es la memoria de una persona y de qué tan difícil es recordar el estímulo, podemos inferir la capacidad de retención personal y la dificultad del memorización

del ítem. En caso de registrar el desempeño de varios participantes en la tarea, cada uno recordando varios ítems que potencialmente difieren en dificultad, es pertinente asumir que todos los participantes provienen de una población común respecto del rasgo capacidad de retención y que todos los ítems provienen de su propia distribución jerárquica respecto de la dificultad de memorización. El suponer distribuciones jerárquicas sobre participantes y sobre ítems permite computar estimaciones más precisas sobre ambos rasgos, como lo sugieren algunos resultados reportados por Jeff Rouder y colaboradores (Rouder y Lu, 2005; Rouder et al., 2007, 2017).

Para profundizar sobre el tema a continuación se enlistan algunas lecturas adicionales, entre las que destaca el tutorial de Shiffrin et al. (2008), donde se presenta un ejemplo con estructura similar a la de este capítulo acompañado por numerosas técnicas adicionales para el estudio y la evaluación de modelos probabilísticos en cognición y conducta.

Lecturas Recomendadas

- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15*, 1-15.
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.
- Rouder, J. N., Morey, R. D. & Pratte, M. S. (2017). Bayesian hierarchical models of cognition. En W. H. Batchelder, H. Colonius, E. Dzhafarov, y J. I. Myung, (Eds.), *The New Handbook of Mathematical Psychology, Volume 1: Measurement and Methodology*. Cambridge University Press.
- Shiffrin, R. M., Lee, M. D., Kim, W. & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248-1284.

Referencias

- Griffiths, T. L. & Yuille, A. (2006). Technical introduction: A primer on probabilistic inference.
<http://dx.doi.org/doi:10.1016/j.tics.2006.05.007>
- Limpert, E., Stahel, W. A. & Abbt, M. (2001). Log-normal distributions across the sciences:
Keys and clues. *BioScience*, 51, 341-352.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 20-22.
- R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rouder, J. N. & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12, 573-604.
- Rouder, J. N., Lu, J., Sun, D., Speckman, P. L., Morey, R. D. & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, 72, 621-642.
- Vincent, B. T. (2015). A tutorial on Bayesian models of perception. *Journal of Mathematical Psychology*, 66, 103-114.

Pies de Figura

Figura 1. Distribuciones Poisson para tres valores λ diferentes.

Figura 2. Modelo M_0 expresado en notación gráfica.

Figura 3. Distribución posterior del parámetro λ de acuerdo con el modelo M_0 .

Figure 4. Datos de todas las personas. Cada columna corresponde con una persona y cada renglón con una fiesta. Cada posición c_{fp} de la matriz c es el número de canciones que la persona p bailó en la fiesta f .

Figura 5. Modelo M_1 expresado en notación gráfica.

Figura 6. Distribuciones posteriores de cada nodo λ_p de acuerdo con el modelo M_1 . Cada curva en la gráfica superior corresponde con la densidad posterior del parámetro λ_p de cada participante. La gráfica central y la inferior detallan la distribución posterior sobre λ de los participantes extremos (P4 y P7, respectivamente).

Figura 7. Datos comparados contra la distribución posterior predictiva del modelo M_1 . Los círculos negros y delgados representan los datos observados (comparar con la Figura 4). Los círculos grises demarcan las zonas de mayor densidad posterior de la distribución posterior predictiva del modelo M_1 en cada fiesta de cada persona. El ajuste entre los datos y las predicciones de M_1 es cuestionable, pero la deficiencia principal de este modelo es que no puede hacer predicciones razonables sobre un participante nuevo, representado en la columna p_n .

Figura 8. Modelo M_2 expresado en notación gráfica.

Figura 9. Datos comparados contra la distribución posterior predictiva del modelo M_2 . La capacidad de ajuste de este modelo parece similar a la de M_1 , pero a diferencia del

modelo anterior, M_2 puede hacer predicciones más razonables sobre un participante nuevo.

Figura 10. Distribuciones posteriores sobre los nodos λ_p de acuerdo con el modelo M_2 . Aunque ambos modelos observan exactamente el mismo conjunto de datos, M_1 y M_2 llegan a conclusiones diferentes, sobre todo acerca de los participantes extremos. En el modelo M_2 las distribuciones posteriores sobre de los participantes P4 y P7 no son tan extremas como en el modelo M_1 .

Figura 11. Distribuciones posteriores sobre los parámetros μ^λ y σ^λ y poblacionales de acuerdo con el modelo M_2 .

Figura 12. Modelo M_3 expresado en notación gráfica.

Figura 13. Datos comparados contra la distribución posterior predictiva del modelo M_3 . Aunque M_3 hace predicciones razonables sobre un participante nuevo, la predicción sobre una fiesta nueva parece poco informada.

Figura 14. Modelo M_4 expresado en notación gráfica.

Figura 15. Datos comparados contra la distribución posterior predictiva del modelo M_4 . Al suponer una distribución jerárquica sobre participantes y otra sobre fiestas, este modelo puede hacer predicciones razonables sobre un participante nuevo y sobre una fiesta nueva.

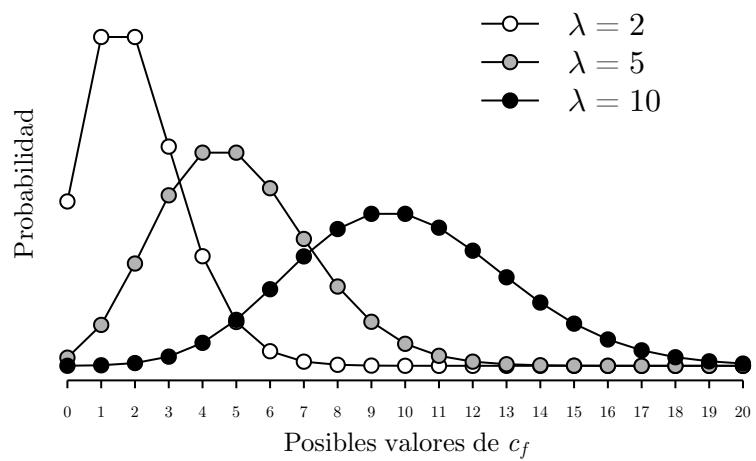
Figura 1

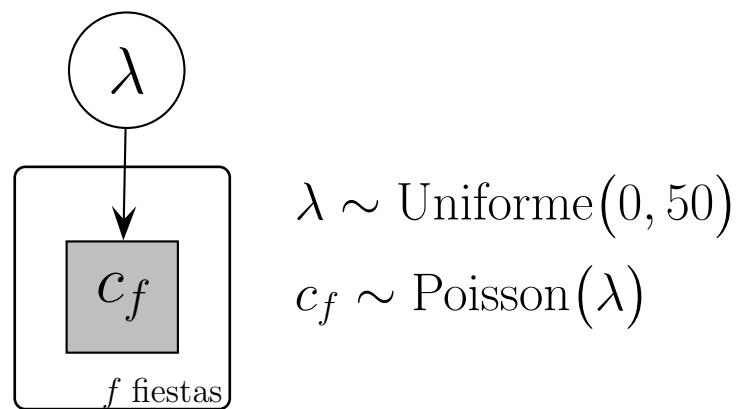
Figura 2

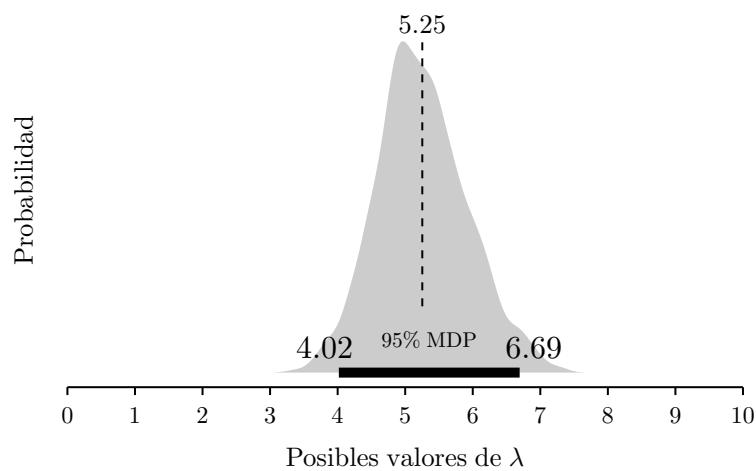
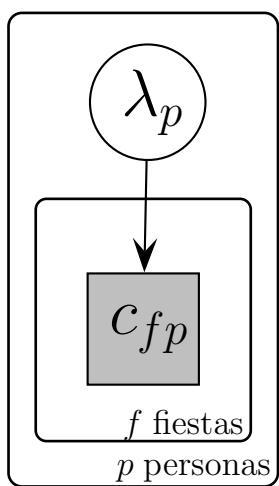
Figura 3

Figura 4

		Personas														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Fiestas	1	6	8	2	2	3	1	10	3	3	2	6	10	7	5	2
	2	8	8	6	1	1	5	12	4	5	3	9	2	4	3	3
	3	8	9	4	1	2	4	17	6	7	1	10	14	5	4	3
	4	9	13	10	2	2	5	17	9	3	3	13	8	8	7	4
	5	7	8	4	0	1	4	9	7	2	1	5	5	0	4	3
	6	11	7	5	1	2	3	11	8	6	2	5	6	3	5	0
	7	6	11	9	2	2	4	18	11	6	2	10	11	3	5	1
	8	6	9	5	1	2	2	10	3	3	2	9	6	6	4	3
	9	10	8	8	3	5	5	13	7	4	5	11	10	5	5	1
	10	13	12	7	1	2	8	20	8	7	6	16	8	10	7	3
	11	13	10	9	5	4	3	17	10	5	4	9	6	5	7	9
	12	15	8	7	1	0	5	16	11	6	5	13	11	7	6	6

Figura 5

$$\lambda_p \sim \text{Uniforme}(0, 50)$$

$$c_{fp} \sim \text{Poisson}(\lambda_p)$$

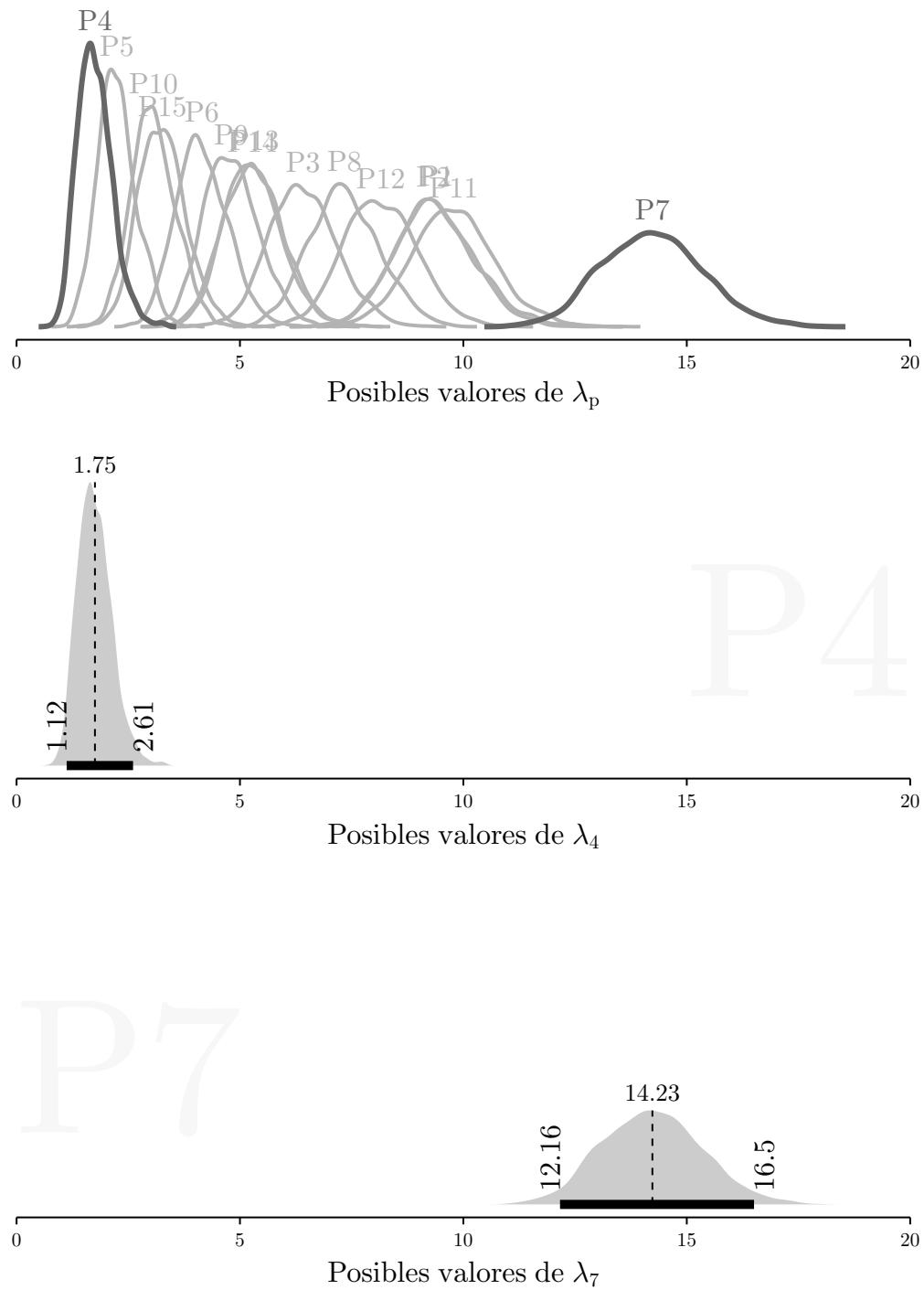
Figura 6

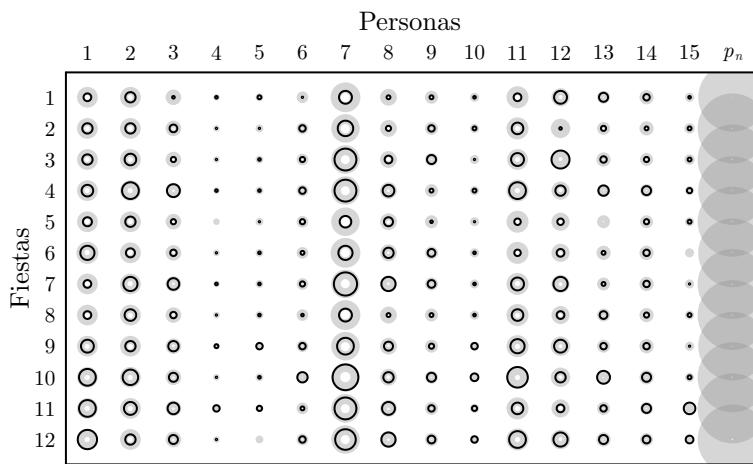
Figura 7

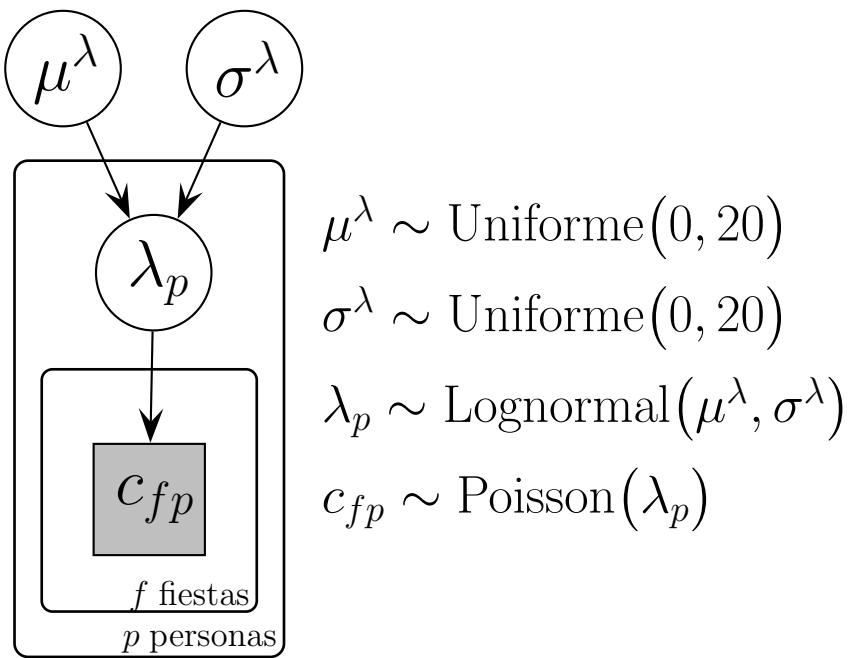
Figura 8

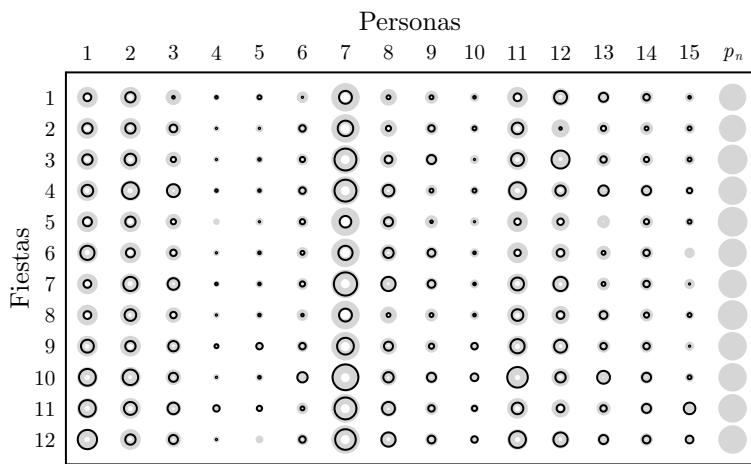
Figura 9

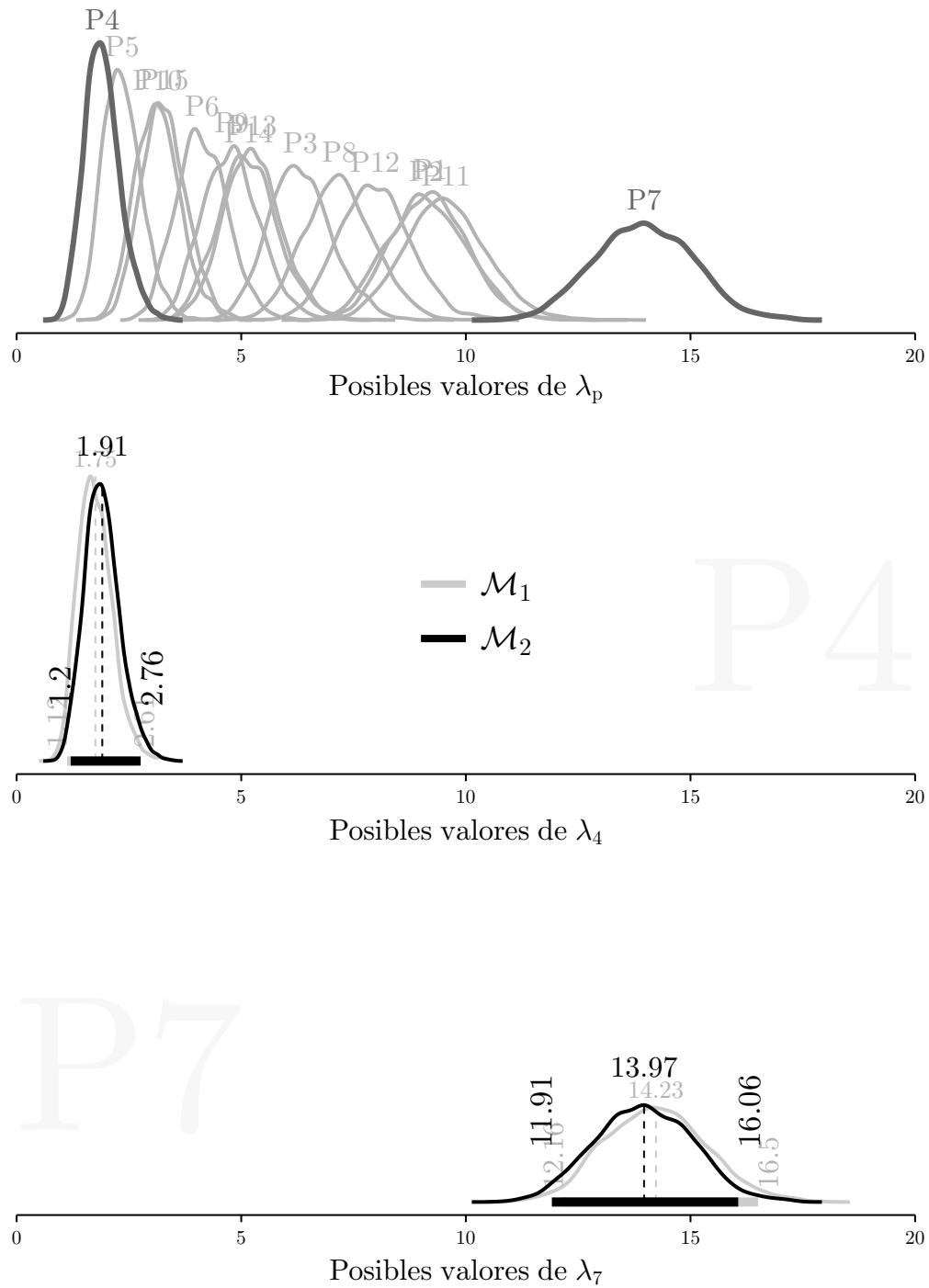
Figura 10

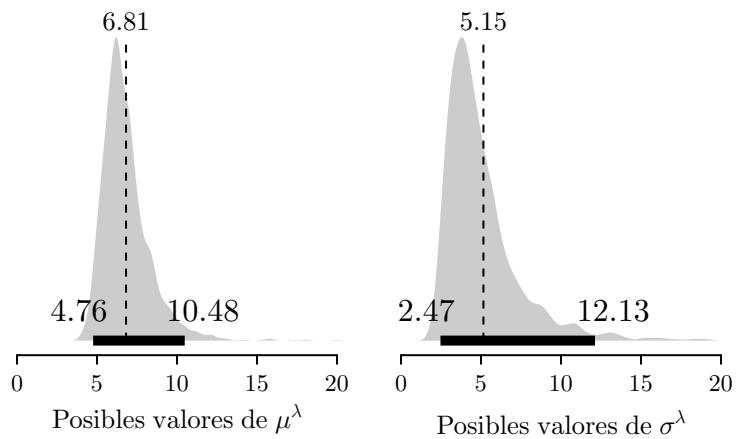
Figura 11

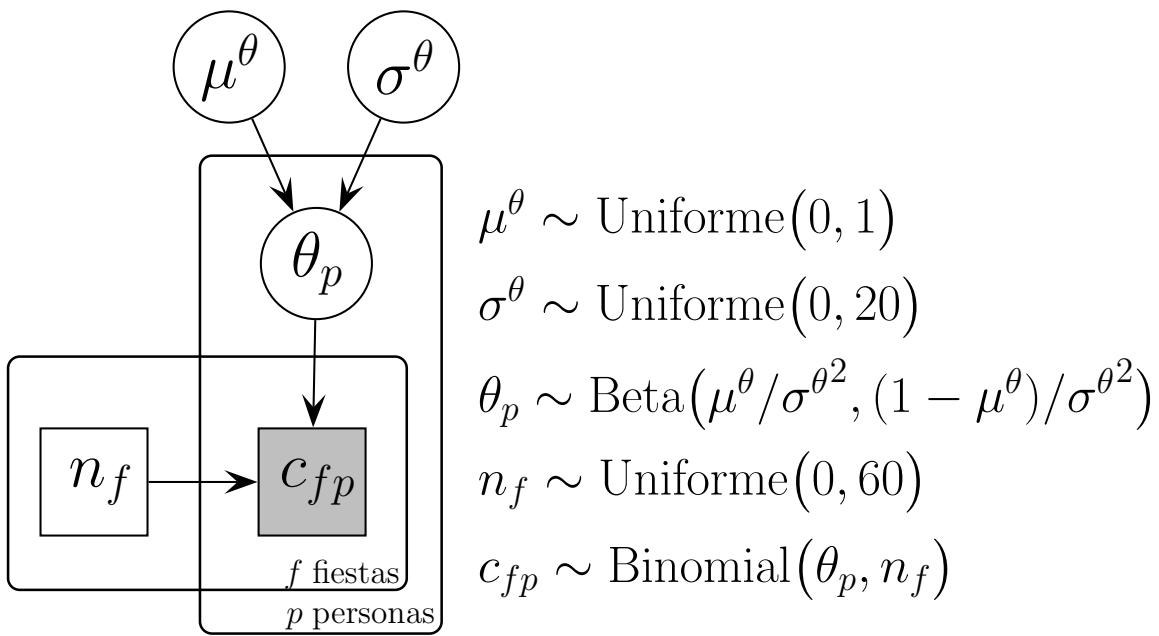
Figura 12

Figura 13

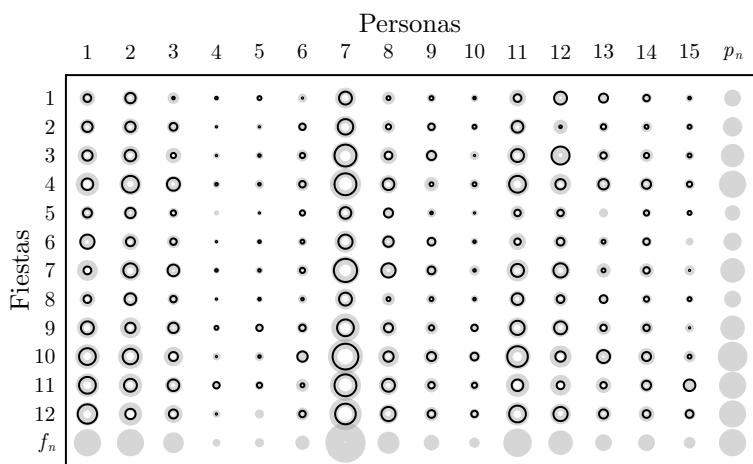


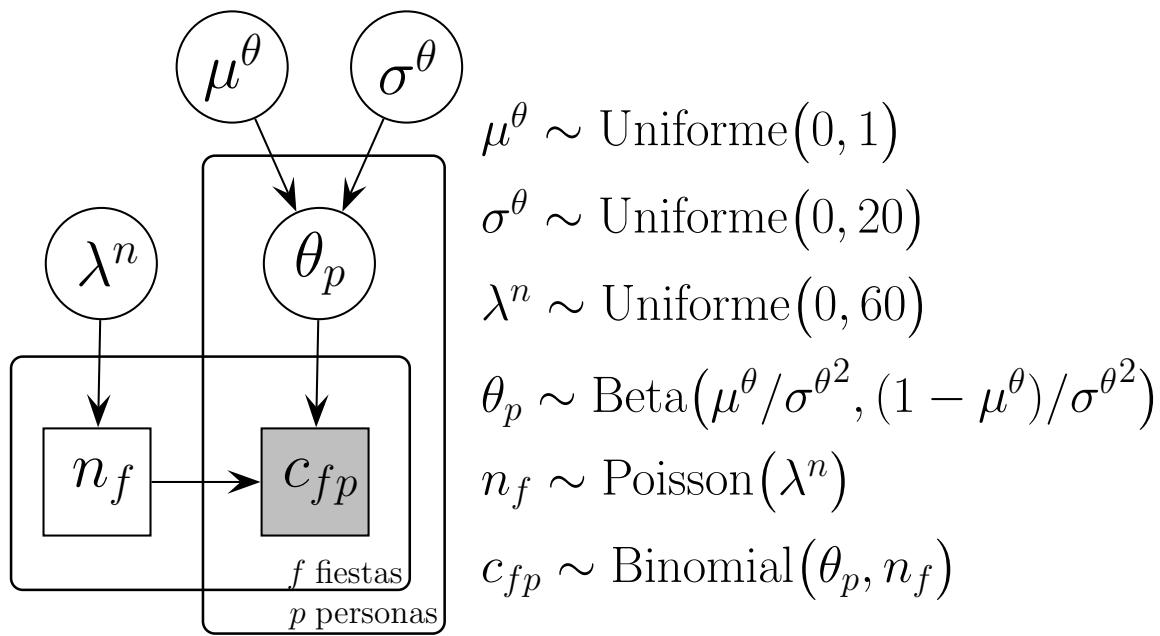
Figura 14

Figura 15

