TRABAJO FINAL DEL MÁSTER EN BIOESTADÍSTICA

Distribuciones previas objetivas para el modelo Dirichlet-multinomial: una aplicación en la agricultura

Danilo Alvares da Silva

VNIVERSITAT DE VALÈNCIA

Tutoras: Carmen Armero i Cervera — Departament d'Estadística i I.O. Universitat de València

Anabel Forte Deltell — Departament d'Estadística i I.O. Universitat de València

Burjassot (València), Mayo de 2015

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

La chufa, *Cyperus sculentus*, es un tubérculo que se utiliza principalmente para la elaboración de horchata. Se cultiva en la comarca de l'Horta Nord, en Valencia, y tiene una gran importancia socioeconómica en la zona. La mancha negra en los tubérculos de chufa es una enfermedad de origen desconocido, que produce un ennegrecimiento de la piel que conlleva a su depreciación comercial, pues un cierto porcentaje de tubérculos deben ser desechados para su venta.

El objetivo de este trabajo es analizar si la selección de tubérculos sin mancha negra, para su utilización como simiente, supone una mejora en la cosecha en cuanto a una menor incidencia de la enfermedad.

Los datos analizados proceden de un experimento en invernadero en el que se utilizaron simientes de chufas asintomáticas y con manchas negras. El análisis estadístico se ha realizado utilizando la metodología Bayesiana. Utilizamos modelos de regresión multinomial, estudiando su sensibilidad a la elección de la distribución previa, en concreto a la elección de los parámetros en una distribución Dirichlet. A la combinación del modelo multinomial con la distribución previa Dirichlet se le conoce como modelo Dirichlet-multinomial.

En conclusión, se identifica que la selección de simientes asintomáticas produce menos tubérculos con síntomas de mancha negra. Además, estos resultados son robustos en relación a elección de los parámetros de las distribuciones previas Dirichlet propuestas en la literatura como opciones objetivas o mínimo informativas.

CAPÍTULO 1

INTRODUCCIÓN

1.1. Contexto

Los primeros registros del cultivo de la chufa, *Cyperus sculentus*, se han encontrado en las tumbas y sarcófagos de las primeras dinastías egipcias (Serrallach, 1927). A través de las palabras de Teofrasto, en el libro *Historia plantarum* citado por Negbi (1992), tenemos evidencia de que la chufa era un alimento presente en la vida de los antiguos egipcios:

> "*...en tierras arenosas, no lejos del cauce del río, crece sobre la tierra la llamada Malniathalle, redonda de forma, sin hueso y sin piel. Los habitantes recogen los tubérculos y los cuecen, con lo que se vuelven muy dulces, saboreándose entonces como un postre.*"

Desde aquella época hasta nuestros días, la chufa ha sido consumida y referenciada por muchos pueblos (Dragendorff, 1898), donde la principal forma de consumo ha sido a través de una una bebida refrescante, preparada con agua, azúcar y chufas majadas (o molidas), actualmente denominada horchata (Figura 1.1). No obstante, fue en la Edad Media cuando su cultivo se expandió por el Norte de África, península Ibérica y Sicilia. Dos factores fueron esenciales para la proliferación de la chufa en este período, el primero está relacionado con la prohibición del vino por la religión mahometana, lo que produjo que

los refrescos y las bebidas no alcohólicas tuvieran un aumento de consumo, entre ellas, la horchata. Otro punto importante fue el descubrimiento de las propiedades medicinales de la horchata, especialmente para combatir algunas inflamaciones respiratorias y molestias estomacales (Cavanilles, 1797).

Figura 1.1: Chufas y horchata.

La chufa llegó a zonas mediterráneas y, en particular, a la actual Comunidad Valenciana. Los registros históricos del siglo XIII ya citan el consumo de una bebida refrescante llamada *llet de xufes* (leche de chufas). Existen además informes sobre el cultivo de la chufa en grandes extensiones de tierra, en las regiones de Alboraya y Almácera (Cavanilles, 1797), dentro de la comarca conocida como L'Horta Nord.

Desde entonces, el cultivo de la chufa se fue extendiendo por toda esta comarca, generando puestos de trabajo y poder económico en el entorno agrario. El mayor reconocimiento se obtuvo el 25 de septiembre de 1995 a través de la Denominación de Origen Protegida, que protege e identifica el cultivo de la chufa en dieciséis localidades (Albalat dels Sorells, Alboraya, Albuixech, Alfara del Patriarca, Almácera, Bonrepós y Mirambell, Burjasot, Foyos, Godella, Meliana, Moncada, Paterna, Rocafort, Tabernes Blanques, Valencia y Vinalesa).

L'Horta Nord posee condiciones óptimas para el cultivo de la chufa, cumpliendo las exigencias climáticas y edáficas para el buen desarrollo de los tubérculos. Además, la elaboración de la horchata es un factor clave para

el éxito de la chufa. La horchata, aparte de ser una bebida refrescante, contribuye a una alimentación saludable, tiene buenas propiedades digestivas y astringentes, mejora la salud cardiovascular, tiene un cierto efecto probiótico, ayuda al aumento de la inmunidad, es estimulante, nutritiva y energética y de fácil digestión (Morell y Barber, 1983 y Luján y Perucho, 2003).

El cultivo de la chufa, hasta llegar a los tubérculos utilizados para la horchata es, básicamente: plantación, recolección, lavado, secado, limpieza y clasificación. En este trabajo, nuestra atención se centra en el lavado de la cosecha, en el que las chufas pierden sus raices, se limpia su piel y se eliminan aquellos tubérculos que son clasificados como inadecuados para el consumo por la presencia de la mancha negra, que es una enfermedad que produce un ennegrecimiento en la piel de los tubérculos.

Los primeros casos de mancha negra en tubérculos de chufas en la Comunidad Valenciana fueron detectados en 2010. Desde entonces, su incidencia no ha dejado de crecer, así como los estudios sobre el tema. Pero lo cierto es que aún no se ha determinado con exactitud el origen de la mancha negra ni se ha encontrado una solución eficaz. Con la existencia de este problema en el cultivo de las chufas, los agricultores necesitan conocer con urgencia cuáles son los tratamientos o medidas preventivas que podrían aplicar para mantener a raya esta enfermedad.

Aunque no hay evidencias científicas sobre el efecto del uso de los tubérculos con mancha negra en la elaboración de la horchata, éstos se eliminan como medida de precaución, depreciando el valor comercial de la cosecha que no se aprovecha en su totalidad.

1.2. Experimento y datos

El experimento está relacionado con el impacto socioeconómico de la chufa en la Comunidad Valenciana y, en particular, con la finalidad de aumentar la calidad de los productos procedentes de la chufa. En concreto, consideramos un posible factor preventivo capaz de promover mejoras en la cosecha en cuanto a una menor incidencia de la mancha negra en los tubérculos.

El experimento se realizó en invernadero, en el Instituto Valenciano de Investigaciones Agrarias (IVIA) durante 2014. Consistió en separar los tubérculos utilizados como simientes en asintomáticos (a partir de ahora lote `asintomático`) y con manchas negras (lote `enfermo`). Cada uno de estos lotes contenían 7

macetas (réplicas) y en cada una de ellas se plantaron 5 simientes. En cada maceta, y tras la cosecha, se recogieron las siguientes informaciones acerca de los tubérculos:

1) Número de tubérculos sin síntomas de mancha negra (`asintomático`);

2) Número de tubérculos con síntomas leves de mancha negra (`leve`);

3) Número de tubérculos con síntomas graves de mancha negra (`grave`).

Los datos recogidos se presentan en la Tabla 1.1 y en la Figura 1.2.

Tabla 1.1: Datos del experimento.

Lote	Tubérculos		
	`asintomático`	`leve`	`grave`
`asintomático`	56	153	83
	272	48	8
	162	112	13
	296	2	1
	232	37	16
	278	88	31
	156	35	2
`enfermo`	0	35	213
	0	78	237
	0	43	199
	0	25	223
	0	0	196
	0	87	144
	0	68	161

Tanto en la Tabla 1.1 como en la Figura 1.2 se puede observar un desequilibrio a favor de los tubérculos asintomáticos en el lote `asintomático` y a favor de los tubérculos con síntomas graves en el lote `enfermo`.

En este análisis preliminar, también se puede apreciar un efecto positivo (en el sentido de cosechar más tubérculos asintomáticos) al plantar simientes asintomáticas. Tras este breve análisis descriptivo iniciaremos la fase de modelización, pero antes vamos a destacar los principales objetivos de este estudio.

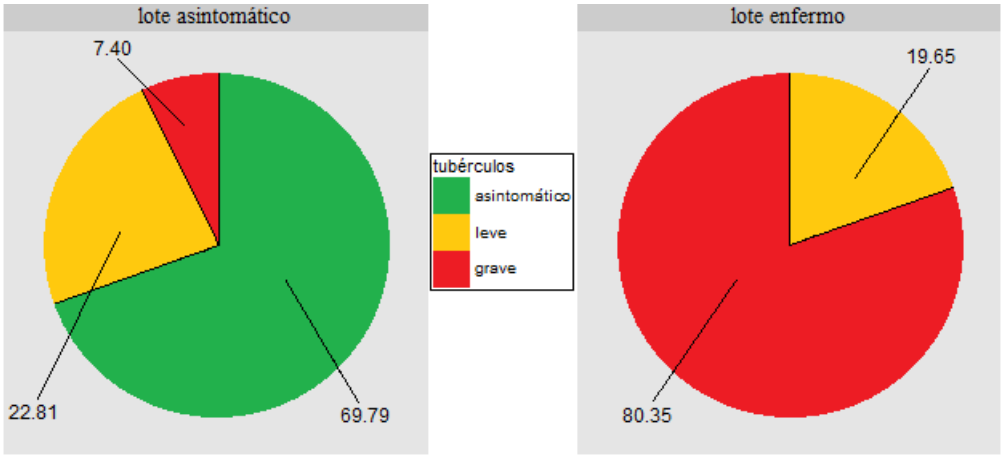

Figura 1.2: Proporción de tubérculos asintomáticos, con síntomas leves y graves obtenidos a partir de simientes asintomáticas (izquierda) y con manchas negras (derecha).

1.3. Objetivos del estudio

Este trabajo tiene dos objetivos fundamentales. El primero es de carácter aplicado y está dirigido a analizar si la selección de tubérculos asintomáticos, para su utilización como simiente, supone una mejora en la cosecha en cuanto a una menor incidencia de la mancha negra en los tubérculos. En este caso, es fácil ver que, desde el punto de vista agronómico, los resultados descriptivos son muy informativos.

El segundo objetivo tiene características más teóricas. En el enfoque Bayesiano es muy interesante el estudio de la sensibilidad de los resultados a la elección de la distribución previa de los parámetros del modelo considerado. En concreto, nuestro interés es analizar el grado de información añadido en la distribución posterior utilizando las principales distribuciones previas objetivas de la literatura para el modelo Dirichlet-multinomial.

CAPÍTULO 2

FORMULACIÓN

2.1. Enfoque Bayesiano

La estadística Bayesiana se basa en la idea de que las incertidumbres pueden ser representadas en términos matemáticos, basados en la teoría de probabilidades (Ditlevsen y Madsen, 1996). En concreto, este enfoque permite asignar probabilidades no sólo a variables aleatorias sino a cualquier evento, parámetro, hipótesis o modelo (Hoijtink *et al.*, 2008).

La base de cualquier procedimiento Bayesiano es el teorema de Bayes (Stuart y Ord, 1994), cuya expresión matemática está dada por

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}, \tag{2.1}$$

donde A y B son los eventos de interés, $P(A)$ y $P(B)$ son las probabilidades marginales de A y de B, $P(A|B)$ es la probabilidad de A dado que B ha ocurrido y $P(B|A)$ es la probabilidad de B dado que A ha ocurrido.

En este contexto, la inferencia sobre los parámetros de un determinado modelo se realiza a través de $P(A|B)$ (denominada distribución posterior) como consecuencia de $P(B|A)$ (función de verosimilitud del modelo estadístico para los datos observados), $P(A)$ (distribución previa) y $P(B)$ (verosimilitud marginal o evidencia del modelo).

Esto conduce a la teoría Bayesiana como herramienta apropiada para el uso de toda la información disponible, sea objetiva, proporcionada por los resultados de los análisis anteriores, o subjetiva, dictada por la experiencia del investigador (Bolstad, 2007).

Hay que destacar que las técnicas Bayesianas ganaron popularidad con la evolución de los ordenadores en cuanto a la capacidad de procesamiento, puesto que la construcción de la distribución posterior, en muchos casos, involucra integrales no analíticas. Sin embargo, con el uso de simulaciones (en especial, los métodos de Monte Carlo via cadenas de Markov, en inglés MCMC) esto ya no es un problema en la mayoría de los casos, y por esta razón la metodología Bayesiana se ha diseminado a diversos campos de la estadística y otras ciencias (Gamerman y Lopes, 2006).

2.2. Modelo Dirichlet-multinomial

Después de esta breve e informal introducción a la estadística Bayesiana, iniciamos la construcción del modelo Dirichlet-multinomial Bayesiano a partir de la descripción de las distribuciones multinomial y Dirichlet.

Distribución multinomial

Existen situaciones en las que nuestra variable de interés toma un número finito de valores, a los cuales nos referimos como categorías o clases. Una opción natural para analizar los datos con esta estructura son los modelos lineales generalizados (en inglés GLM) con respuesta multinomial, que es una extensión de los modelos binomiales para más de dos categorías (Nelder y Wedderburn, 1972; McCullagh y Nelder, 1989).

Consideramos un experimento con n ensayos idénticos e independientes, en el que los resultados pueden tomar J valores distintos $(r_1, \ldots, r_J)$. La probabilidad para cada valor r_j se denota por θ_j, para $j = 1, \ldots, J$, de modo que $0 < \theta_j < 1$ y $\sum_{j=1}^{J} \theta_j = 1$. De esta manera, si Y_j es la variable aleatoria que describe el numero de veces que se observa el valor r_j en los n ensayos, para $j = 1, \ldots, J$, entonces, el vector $\mathbf{Y} = (Y_1, \ldots, Y_J)^\top$ sigue una distribución multinomial con parámetros n y $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$:

$$(\mathbf{Y}|n, \boldsymbol{\theta}) \sim \text{Multinomial}(n, \boldsymbol{\theta}) \tag{2.2}$$

La función de probabilidad de dicha distribución viene dada por

$$f_{\mathbf{Y}}(n_1, \ldots, n_J) = \left(\frac{n!}{n_1!\, n_2! \ldots n_J!} \right) \theta_1^{n_1} \theta_2^{n_2} \ldots \theta_J^{n_J}, \tag{2.3}$$

y las distribuciones marginales son binomiales con los parámetros n y θ_j, para cada j, donde

$$E(Y_j) = n\theta_j, \quad Var(Y_j) = n\theta_j(1 - \theta_j), \tag{2.4}$$

siendo n_j es el número de veces que se observa el valor R_j, para $j = 1, \ldots, J$, con $\sum_{j=1}^{J} n_j = n$.

Distribución Dirichlet

La distribución Dirichlet es una extensión multivariante de la distribución beta. Se trata de una distribución ampliamente utilizada en el contexto Bayesiano como una distribución previa conjugada para un modelo multinomial (Connor y Mosimann, 1969; Agresti, 2013).

Sea $\mathbf{Y} = (Y_1, \ldots, Y_J)^\top$ un vector con J componentes, entonces diremos que este vector sigue una distribución Dirichlet de orden $J \geq 2$ con vector de parámetros $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)^\top$, es decir,

$$(\mathbf{Y}|\boldsymbol{\alpha}) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{2.5}$$

y su función densidad viene dada por

$$f_{\mathbf{Y}}(y_1, \ldots, y_J) = \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} y_j^{\alpha_j - 1}, \; 0 < y_j < 1, \tag{2.6}$$

siendo $\alpha_0 = \sum_{j=1}^{J} \alpha_j$, $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ la función gamma y $\sum_{j=1}^{J} y_j = 1$. En este caso también tenemos una distribución conocida para las marginales, que son distribuciones betas con parámetros $a = \alpha_j$ y $b = (\alpha_0 - \alpha_j)$, para cada j, donde

$$E(Y_j) = \frac{\alpha_j}{\alpha_0}, \quad Var(Y_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}. \tag{2.7}$$

El retorno de la ecuación (2.6) es un vector J-dimensional, en el que los elementos se interpretan como las probabilidades de tener las proporciones $(y_1, \ldots, y_J)^\top$ (Balakrishnan y Nevzorov, 2003).

El uso de (2.6) es recurrente como una distribución previa en varias modelizaciones, así que vale la pena destacar que su forma puede ser propia ($\alpha_j > 0$, para todo $j = 1, \ldots, J$) o impropia ($\alpha_j = 0$, para al menos uno de los j).

Proceso Dirichlet-multinomial

La combinación del modelo probabilístico multinomial con una distribución previa Dirichlet fue propuesta por Lindley (1964) y Good (1965), generalizando así los resultados obtenidos con el modelo binomial y distribución previa beta. Una revisión completa de esta modelización y varias otras que involucran datos discretos se presentan en Congdon (2005).

El modelo Dirichlet-multinomial tiene la siguiente estructura jerárquica:

$$\begin{array}{rcl}(Y_1,\ldots,Y_J|n,\boldsymbol{\theta}) & \sim & \text{Multinomial}(n,\theta_1,\ldots,\theta_J)\\ (\theta_1,\ldots,\theta_J|\boldsymbol{\alpha}) & \sim & \text{Dirichlet}(\alpha_1,\ldots,\alpha_J),\end{array} \tag{2.8}$$

con función de probabilidad y densidad conjunta definidas en (2.3) y (2.6), respectivamente. Utilizando la regla de Bayes (2.1), la distribución posterior para $\boldsymbol{\theta}$ viene dada por

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathscr{D}) &= \frac{\left(\frac{n!}{n_1!n_2!\ldots n_J!}\right)\prod_{i=1}^{J}\theta_i^{n_i}\frac{\Gamma\left(\sum_{j=1}^{J}\alpha_j\right)}{\prod_{j=1}^{J}\Gamma(\alpha_j)}\prod_{j=1}^{J}\theta_j^{\alpha_j-1}}{\int_{\boldsymbol{\theta}}\left(\frac{n!}{n_1!n_2!\ldots n_J!}\right)\prod_{i=1}^{J}\theta_i^{n_i}\frac{\Gamma\left(\sum_{j=1}^{J}\alpha_j\right)}{\prod_{j=1}^{J}\Gamma(\alpha_j)}\prod_{j=1}^{J}\theta_j^{\alpha_j-1}d\boldsymbol{\theta}}\\ \pi(\boldsymbol{\theta}|\mathscr{D}) &\propto \left(\frac{n!}{n_1!n_2!\ldots n_J!}\right)\prod_{i=1}^{J}\theta_i^{n_i}\frac{\Gamma\left(\sum_{j=1}^{J}\alpha_j\right)}{\prod_{j=1}^{J}\Gamma(\alpha_j)}\prod_{j=1}^{J}\theta_j^{\alpha_j-1}\\ \pi(\boldsymbol{\theta}|\mathscr{D}) &\propto \prod_{j=1}^{J}\theta_j^{n_j+\alpha_j-1},\end{aligned}$$

por tanto

$$\boldsymbol{\theta}|\mathscr{D} \sim \text{Dirichlet}(\alpha_1^* = n_1+\alpha_1,\ldots,\alpha_J^* = n_J+\alpha_J), \tag{2.9}$$

siendo $\mathscr{D} = (n_1,\ldots,n_J,\alpha_1,\ldots,\alpha_J)^\top$.

2.3. Distribuciones previas objetivas

La elección de los valores del vector de hiperparámetros $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_J)^\top$ establece el grado de información de la distribución previa. Las distribuciones previas objetivas más utilizadas en la literatura son:

- Dirichlet($\alpha_1 = 0, \ldots, \alpha_J = 0$): distribución previa Dirichlet propuesta por Haldane (1948) (distribución previa de Haldane de ahora en adelante). Aunque se trata de una distribución impropia, esto no implica que la distribución posterior también lo sea, situación altamente indeseable que invalidaría el proceso de estimación.

- Dirichlet($\alpha_1 = 1/J, \ldots, \alpha_J = 1/J$): distribución previa Dirichlet propuesta por Perks (1947) (distribución previa de Perks de ahora en adelante). Es la única distribución previa objetiva, de entre las que analizamos en este trabajo, que considera el número de categorías, J, como hiperparámetro. Esta distribución se utiliza también en el reciente artículo de Berger *et al.* (2015);

- Dirichlet($\alpha_1 = 1/2, \ldots, \alpha_J = 1/2$): distribución previa Dirichlet propuesta por Jeffreys (1946; 1961) (distribución previa de Jeffreys de ahora en adelante). Utiliza uno de los procedimientos más comunes para obtener una distribución previa no informativa (regla de Jeffreys) y también es una de las distribuciones previas objetivas globales (Berger *et al.* 2015) para el modelo probabilístico multinomial, sin embargo, utilizando como base teórica la distribución previa de referencia (Bernardo, 1979);

- Dirichlet($\alpha_1 = 1, \ldots, \alpha_J = 1$): distribución previa Dirichlet propuesta por Bayes (1763) y Laplace (1812) (distribución previa de Bayes-Laplace de ahora en adelante). Es conocida por simplificar la distribución Dirichlet a una constante y también puede ser tratada como una distribución uniforme.

A continuación, realizaremos un estudio analítico de cada una de las distribuciones previas propuestas y cómo se comportan sus distribuciones marginales.

Distribución previa de Haldane

La propuesta de Haldane (1948) para los parámetros de la distribución previa Dirichlet, define $\alpha_j = 0$ para todo $j = 1, \ldots, J$, donde la distribución posterior (2.9) se convierte en

$$\pi(\boldsymbol{\theta}|\mathscr{D}) \propto \prod_{j=1}^{J} \theta_j^{n_j - 1}. \tag{2.10}$$

La construcción de dicha distribución puede presentar problemas computacionales, pues la distribución Dirichlet (2.6) con vector de hiperparámetros $\boldsymbol{\alpha} = \mathbf{0}$ no está definida, es decir, se trata de una distribución previa impropia expresada por

$$\pi(\theta_1, \ldots, \theta_J) \propto \frac{1}{\prod_{j=1}^{J} \theta_j}, \; J > 2, \tag{2.11}$$

donde $\int \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \infty$ (Liu *et al.*, 2012). La Figura 2.1 muestra la forma de las distribuciones previas marginales de Haldane.

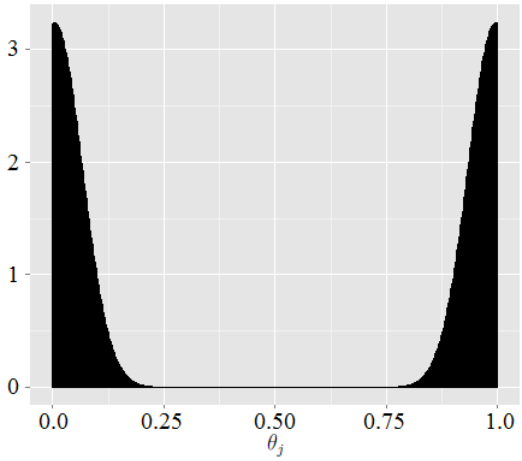

Figura 2.1: Simulación de una distribución previa marginal de Haldane Dirichlet$(\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0)$.

Las distribuciones marginales beta procedentes de una Dirichlet de Haldane tienen una forma casi dicotómica, es decir, la densidad asociada a cada θ_j, $j = 1, \ldots, J$, está cerca de 0 o 1 y cómo dicha distribución incorpora la restricción $\sum_{j=1}^{J} \theta_j = 1$, tenemos que si una componente está cercana a 1, las otras estarán cercanas a 0.

La visualización de la distribución de Haldane, a pesar de ser una distribución objetiva, es bastante contraria a la intuición. Sin embargo, la transformación log-odds, logit(θ_j), hace que la distribución sea uniforme en la recta real (Zellner, 1996). También podemos utilizar la distribución posterior (2.10) para demostrar que la distribución previa (2.11) es no informativa (Jaynes, 2003), donde la media a posteriori viene dada por (2.7) y puede ser reescrita como $E(\theta_j|\mathscr{D}) = n_j/n$, $j = 1, \ldots, J$. Puesto que se trata del estimador máximo verosímil, la distribución previa realmente no está agregando información al

modelo (2.8). La Figura 2.2 también ilustra la simulación de la distribución previa de Haldane.

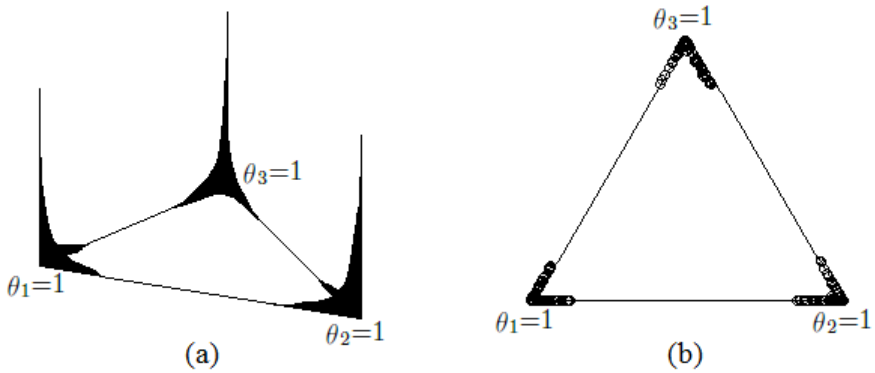

Figura 2.2: Simulación de la distribución previa de Haldane (Dirichlet($\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0$)) (a) y su proyección en el triángulo simplex (b).

La interpretación de la Figura 2.2, (a) y (b), es sencilla, como los θ's suman 1, cada vértice del triángulo indica qué componente del vector aleatorio Dirichlet tiene valor uno y en contraposición los otros componentes son cero. Cuando nos alejamos dicho vértice el valor de su respectivo θ disminuye.

Distribución previa de Perks

La distribución previa de Perks (Perks, 1947) utiliza $1/J$ como valor para los hyperparámetros α_j, $j = 1, \ldots, J$, donde, recordemos, J es la dimensión del vector aleatorio de la distribución Dirichlet. La función de densidad de la distribución previa de Perks viene dada por:

$$\pi(\theta_1, \ldots, \theta_J) = \frac{1}{\prod_{j=1}^{J} \Gamma(1/J)} \prod_{j=1}^{J} \theta_j^{1/J-1}, \; J > 2. \tag{2.12}$$

El aumento del número de categorías J implica el aumento del denominador de (2.12), puesto que la función gamma, para valores positivos, es una función convexa con el punto mínimo aproximadamente 0.8856. Por eso, en los casos que J es muy grande la distribución previa de Perks puede tener los mismos problemas que la distribución previa de Haldane, sin embargo, en la práctica J tiene un tamaño reducido. La distribución posterior (2.9) será una Dirichlet con parámetros $(n_j + 1/J)$ para todo $j = 1, \ldots, J$. La Figura 2.3 muestra la forma de las distribuciones previas marginales de Perks.

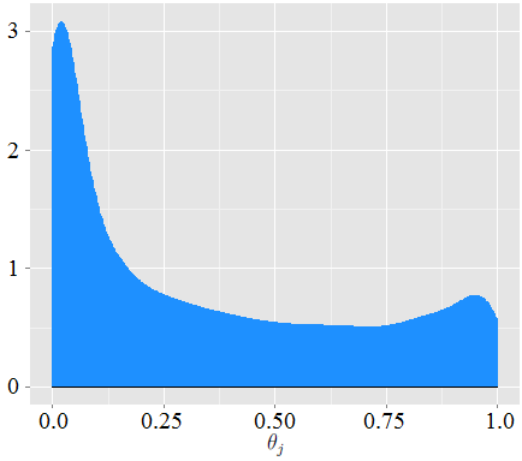

Figura 2.3: Simulación de una distribución previa marginal de Perks: Dirichlet($\alpha_1 = 1/3, \alpha_2 = 1/3, \alpha_3 = 1/3$).

Es fácil ver que esta distribución marginal es asimétrica positiva y ligeramente bimodal. Otro punto importante es que el aumento del número de categorías J implicará un desplazamiento de las probabilidades centrales hacia los extremos, como sucede en la Figura 2.1.

Otra forma de verificar el comportamiento de la distribución previa de Perks es a través de la Figura 2.4, donde se presenta la simulación de dicha distribución.

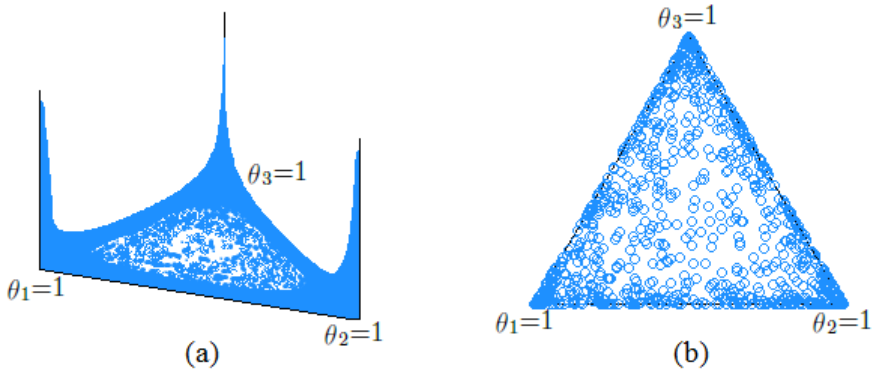

Figura 2.4: Simulación de la distribución previa de Perks (Dirichlet($\alpha_1 = 1/3, \alpha_2 = 1/3, \alpha_3 = 1/3$)) (a) y su proyección en el triángulo simplex (b).

A partir de las dos gráficas de la Figura 2.4 es posible notar que los valores de los θ's se concentran en los bordes de los triángulos. Sin embargo, la región

central comienza a "llenarse", lo que representa una mayor homogeneidad entre los valores posibles para el vector $\boldsymbol{\theta}$. Esta distribución previa también se obtiene, de forma alternativa, en el artículo de Berger *et al.* (2015).

Distribución previa de Jeffreys

Sea cual sea el modelo utilizado, si se quiere seguir un enfoque Bayesiano objetivo lo más común es recurrir a la distribución previa de Jeffreys (Jeffreys, 1946; Jeffreys, 1961), que en el contexto de modelo multinomial, es la distribución Dirichlet con $\alpha_j = 1/2$ para todo $j = 1, \ldots, J$. En este caso la densidad (2.6) se convierte en

$$\pi(\theta_1, \ldots, \theta_J) = \frac{\Gamma(J/2)}{\pi^{J/2}} \prod_{j=1}^{J} \theta_j^{-1/2}, \; J > 2. \tag{2.13}$$

En contraste a las distribuciones previas de Haldane y Perks, en este caso tenemos una mayor estabilidad con respecto a posibles problemas computacionales. La Figura 2.5 muestra la forma de las distribuciones previas marginales de Jeffreys.

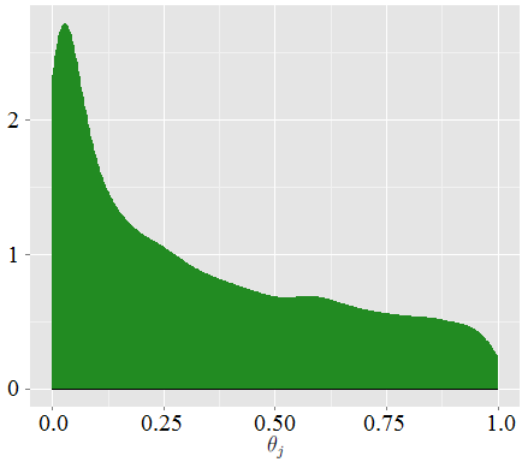

Figura 2.5: Simulación de una distribución previa marginal de Jeffreys: Dirichlet$(\alpha_1 = 1/2, \alpha_2 = 1/2, \alpha_3 = 1/2)$.

Se puede observar que la Figura 2.5 se asemeja a la Figura 2.3, donde los valores más probables siguen siendo los cercanos a cero, sin embargo, hay

una mayor uniformidad entre la región central y el límite superior. La Figura 2.6 también ilustra la distribución Dirichlet de Jeffreys, pero ahora con una representación conjunta.

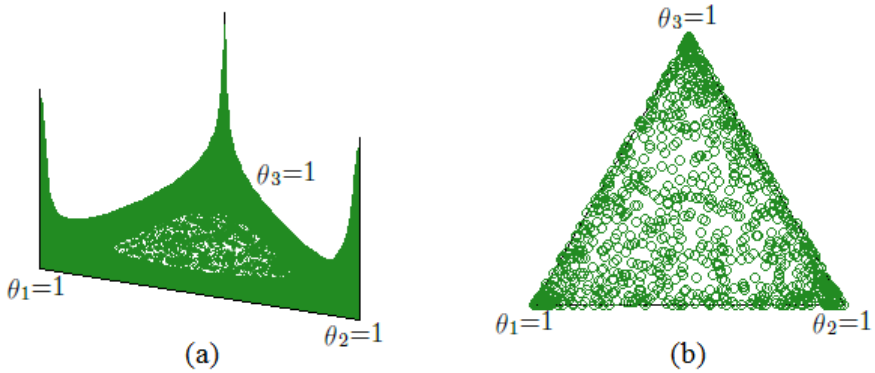

Figura 2.6: Simulación de la distribución previa de Jeffreys (Dirichlet($\alpha_1 = 1/2, \alpha_2 = 1/2, \alpha_3 = 1/2$)) (a) y su proyección en el triángulo simplex (b).

Análogo a la distribución previa de Perks, la Figura 2.6 muestra los triángulos (a) y (b) rellenos casi por completo. Sin embargo, en este caso la región central es más homogénea, con una distribución más equilibrada entre las posibles combinaciones de los valores del vector $\boldsymbol{\theta}$. Esta distribución previa también se obtiene, de forma alternativa, en el artículo de Berger *et al.* (2015).

Distribución previa de Bayes-Laplace

La distribución previa de Bayes-Laplace (Bayes, 1763 y Laplace, 1812) es otra opción bien establecida en la literatura. En este caso la distribución Dirichlet (2.6) se simplifica considerablemente con $\alpha_j = 1$ para todo $j = 1, \ldots, J$, asumiendo un valor constante:

$$\pi(\theta_1, \ldots, \theta_J) = (J-1)!, \; J > 2. \tag{2.14}$$

Se trata por tanto de una distribución uniforme. La distribución posterior (2.9) se mantiene como una Dirichlet, pero con vector de parámetros $\boldsymbol{\alpha}^* = [(n_1 + 1), \ldots, (n_J + 1)]^\top$. La Figura 2.7 muestra la forma de las distribuciones previas marginales de Bayes-Laplace.

La Figura 2.7 muestra una formato monótono decreciente entre las probabilidades marginales de uno de los componentes del vector aleatorio Dirichlet.

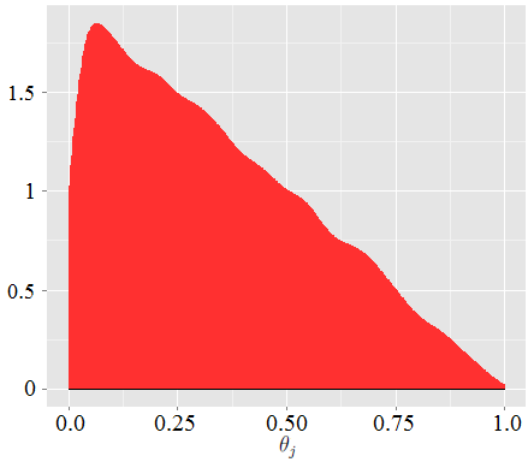

Figura 2.7: Simulación de una distribución previa marginal de Bayes-Laplace: Dirichlet($\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1$).

Sin embargo, la restricción $\sum_{j=1}^{J} \theta_j = 1$ es la conexión entre dicha estructura y la distribución conjunta uniforme (2.14). A través de la Figura 2.8 se puede identificar más fácilmente el comportamiento de una distribución equiprobable.

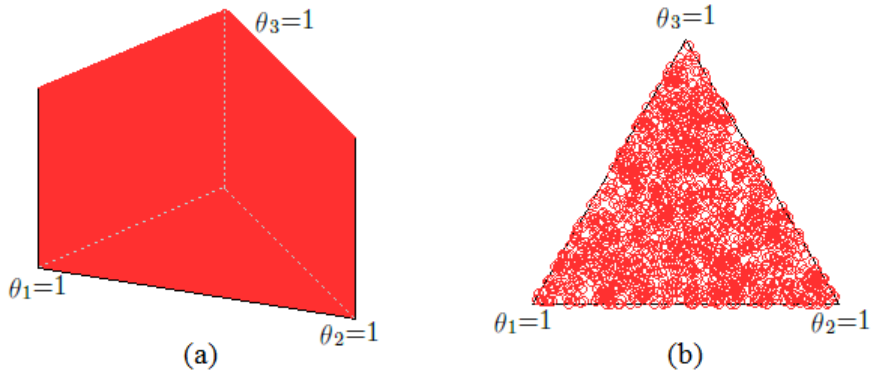

Figura 2.8: Simulación de la distribución previa de Bayes-Laplace (Dirichlet($\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1$)) (a) y su proyección en el triángulo simplex (b).

Nótese que en el caso de la Dirichlet de Bayes-Laplace, las simulaciones presentadas en la Figura 2.8 representan una estructura homogénea y, consecuentemente, no informativa.

Distribución previa Dirichlet informativas

Para explorar la sensibilidad en la forma de la distribución (2.6) con los cambios de sus parámetros, la Figura 2.9 muestra la distribución marginal Dirichlet para algunos casos en que los α's asumen valores superiores a los casos estudiados.

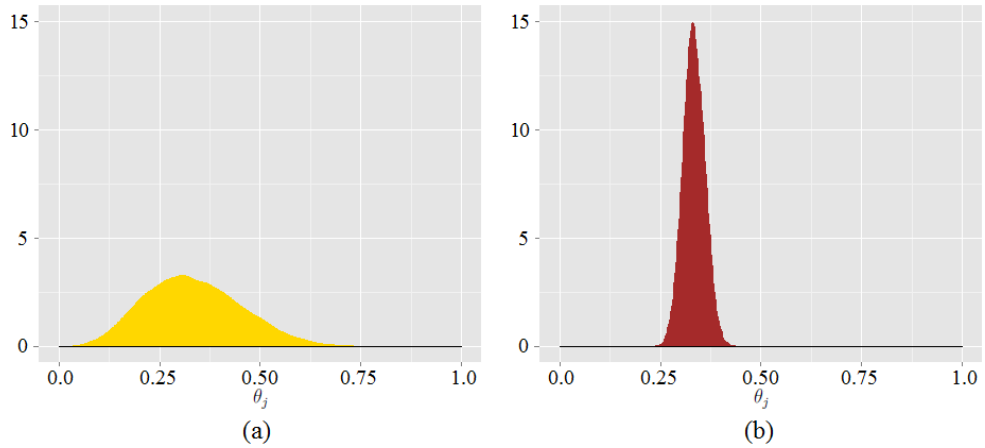

Figura 2.9: Simulación de una distribución marginal Dirichlet($\alpha_1 = 5, \alpha_2 = 5, \alpha_3 = 5$) (a) y Dirichlet($\alpha_1 = 100, \alpha_2 = 100, \alpha_3 = 100$) (b).

Las evidencias transmitidas desde la Figura 2.9 están relacionadas con el comportamiento informativo de la distribución previa Dirichlet con valores elevados para los parámetros. La Figura 2.9 (a) ya no tiene los extremos más densos y su rango está entre 0 y 0.75 con una media de 0.3. Siguiendo el mismo patrón, pero con una variabilidad más pequeña, en la Figura 2.9 (b) hay una redución drástica del intervalo y la media pasa a ser 0.33.

A continuación está representada en la Figura 2.10 la distribución Dirichlet para estos dos casos.

La Figura 2.10 refuerza la idea de una mayor precisión según crecen positivamente los valores de los α's y, otra vez, se puede ver que los extremos no están contenidos en la distribución conjunta (o con probabilidad casi nula), en otras palabras, la tendencia es que cada componente del vector $\boldsymbol{\theta}$ asuma el mismo valor $1/3$ y $1/J$ para el caso general de J categorías.

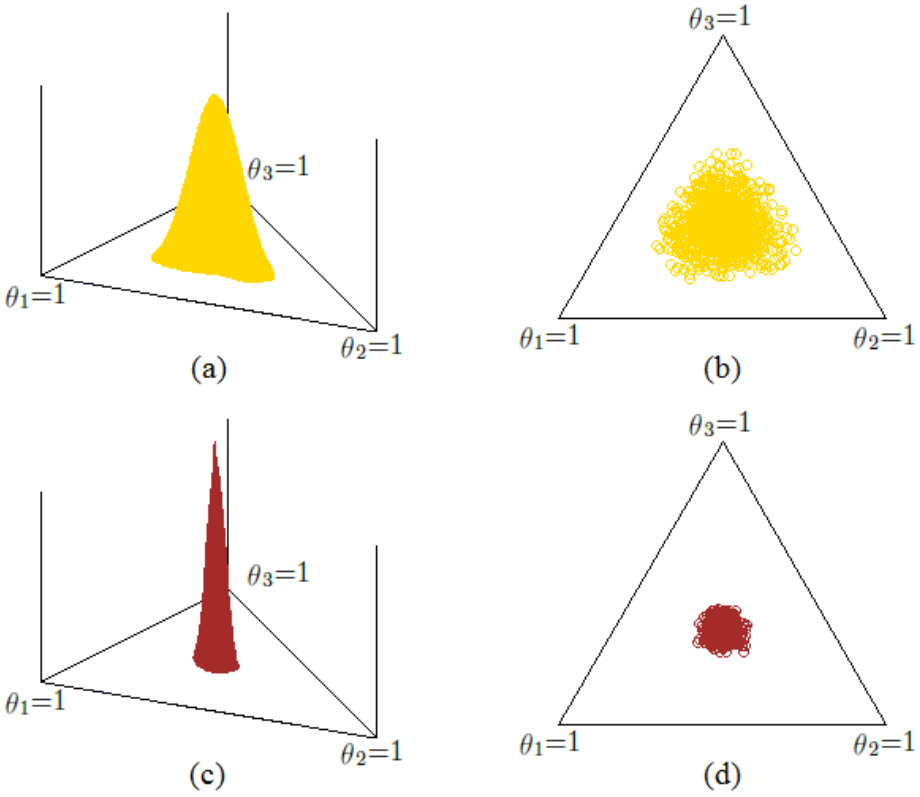

Figura 2.10: Simulación de las distribuciones previas Dirichlet($\alpha_1 = 5, \alpha_2 = 5, \alpha_3 = 5$)) (a) y Dirichlet($\alpha_1 = 100, \alpha_2 = 100, \alpha_3 = 100$)) (c) y sus proyecciones, (b) y (d) respectivamente, en el triángulo simplex.

CAPÍTULO 3

RESULTADOS Y DISCUSIÓN

Después de introducir algunas propuestas de distribuciones previas para el modelo Dirichlet-multinomial, iniciaremos la modelización de los datos de la enfermedad de la mancha negra de la chufa, presentados en el Capítulo 1. Nuestro modelo base se describe por

$$\begin{array}{rcl} (Y_{1ik}, Y_{2ik}, Y_{3ik}|n_{ik}, \boldsymbol{\theta}_k) & \sim & \text{Multinomial}(n_{ik}, \theta_{1k}, \theta_{2k}, \theta_{3k}) \\ (\theta_{1k}, \theta_{2k}, \theta_{3k}|\boldsymbol{\alpha}_k) & \sim & \text{Dirichlet}(\alpha_{1k}, \alpha_{2k}, \alpha_{3k}), \end{array} \tag{3.1}$$

donde Y_{jik} es el número de tubérculos del tipo j ($j = 1$: `asintomático`, $j = 2$: `leve` y $j = 3$: `grave`) en la maceta i ($i = 1, \ldots, 7$: número de réplicas) en el lote k ($k = 1$: lote `asintomático` y $k = 2$: lote `enfermo`); n_{ik} es el número total de tubérculos en la observación i en el lote k; θ_{jk} es la probabilidad de obtener un tubérculo del tipo j en el lote k; y el parámetro α_{jk} se refiere al componente θ_{jk} del vector aleatorio Dirichlet.

La estimación del modelo (3.1) se ha realizado para cada lote de forma independiente utilizando las cuatro distribuciones previas presentadas en el capítulo anterior. Además, también se ha utilizado una distribución previa Dirichlet extremadamente informativa ($\alpha_j = 100$ para todo $j = 1, \ldots, J$), presentada en las Figuras 2.9 y 2.10, para ilustrar los efectos sobre la distribución posterior.

La distribución posterior de los parámetros del modelo se ha aproximado por métodos de Monte Carlo basados en cadenas de Markov (MCMC) a través

del *software* estadístico `WinBUGS` (Lunn *et al.*, 2000). El algoritmo MCMC ha corrido tres cadenas de Markov con 100 000 iteraciones, con un periodo de calentamiento (*burn-in*) de 10 000 iteraciones, durante el cual esperamos que la simulación converja a la distribución posterior. Además, con el fin de disminuir la auto-correlación en la muestra, hemos establecido un intervalo de adelgazamiento (*thinnig*) de 5, es decir, sólo guardamos 1 de cada 5 iteraciones.

A través del análisis gráfico y utilizando el criterio de convergencia de Gelman y Rubin (Gelman y Rubin, 1992), donde el estadístico de contraste está cercano a 1, hemos comprobado que todas las cadenas convergieron. La Tabla 3.1 muestra algunas medidas de las probabilidades asociadas a las categorías de tubérculos asintomáticos, leves y graves procedentes de simientes asintomáticas obtenida con la distribución posterior del modelo con la distribución previa de Haldane.

Tabla 3.1: Resumen de la distribución posterior de la probabilidad de cosechar un tubérculo asintomático, leve y grave procedente de simientes asintomáticas utilizando la distribución previa de Haldane.

probabilidad	media	desviación típica	EMC ($\times 10^{-5}$)	2.5 %	mediana	97.5 %
θ_{11}	0.698	0.010	3.30	0.678	0.698	0.717
θ_{21}	0.228	0.009	3.16	0.210	0.228	0.247
θ_{31}	0.074	0.006	1.86	0.063	0.074	0.086

EMC: error de Monte Carlo.

Téngase en cuenta que las probabilidades (θ_{11}, θ_{21} y θ_{31}) muestran que la siembra de simientes asintomáticas proporciona una mayor probabilidad a los tubérculos asintomáticos (0.698), mientras que los tubérculos con manchas negras leve y grave tienen medias asociadas de 0.228 y 0.074, respectivamente. Para el lote `enfermo` no se puede hacer inferencia, puesto que el hecho de no tener observaciones de tubérculos asintomáticos combinado con $\alpha_j = 0$ (para todo $j = 1, \ldots, J$) produce una distribución posterior impropia.

Continuando con el análisis, la Tabla 3.2 muestra los principales resultados de la estimación del modelo Dirichlet-multinomial con la distribución previa de Perks. Óbservese que los valores del lote `asintomático` son sumamente similares a los obtenidos con la distribución previa de Haldane (Tabla 3.1) y, sin embargo, en este caso la distribución posterior es propia. El resultado del lote `enfermo` es opuesto al del lote `asintomático`, pues los tubérculos con síntomas graves tienen probabilidad predominante (0.803), seguido de los

tubérculos con síntomas leves de mancha negra (0.197) y los asintomáticos (0.000).

Tabla 3.2: Resumen de la distribución posterior de la probabilidad de cosechar un tubérculo asintomático, leve y grave procedente de simientes asintomáticas (filas superiores) y graves (filas inferiores) utilizando la distribución previa de Perks.

probabilidad	media	desviación típica	EMC ($\times 10^{-5}$)	2.5 %	mediana	97.5 %
θ_{11}	0.698	0.010	3.11	0.678	0.698	0.717
θ_{21}	0.228	0.009	3.02	0.210	0.228	0.247
θ_{31}	0.074	0.006	1.85	0.063	0.074	0.086
θ_{12}	0.000	0.003	9.31	0.000	0.000	0.001
θ_{22}	0.197	0.010	3.29	0.178	0.197	0.216
θ_{32}	0.803	0.010	3.39	0.784	0.803	0.821

EMC: error de Monte Carlo.

Además, hay diferencias relevantes entre cada tipo de tubérculos en los dos lotes analizados, que muestran que la selección de simientes asintomáticas contribuye a la reducción de los tubérculos con manchas negras.

El siguiente análisis se ha realizado con la distribución previa de Jeffreys, donde la Tabla 3.3 muestra un resumen de las características del modelo estimado.

Tabla 3.3: Resumen de la distribución posterior de la probabilidad de cosechar un tubérculo asintomático, leve y grave procedente de simientes asintomáticas (filas superiores) y graves (filas inferiores) utilizando la distribución previa de Jeffreys.

probabilidad	media	desviación típica	EMC ($\times 10^{-5}$)	2.5 %	mediana	97.5 %
θ_{11}	0.698	0.010	3.31	0.678	0.698	0.717
θ_{21}	0.228	0.009	3.14	0.211	0.228	0.247
θ_{31}	0.074	0.006	1.93	0.063	0.074	0.086
θ_{12}	0.000	0.003	9.31	0.000	0.000	0.001
θ_{22}	0.197	0.010	2.97	0.178	0.197	0.216
θ_{32}	0.803	0.010	3.06	0.784	0.803	0.822

EMC: error de Monte Carlo.

Una vez más tenemos resultados similares a los dos casos anteriores (Ta-

blas 3.1 y 3.2) para el lote `asintomático`, mientras que para el lote `enfermo` los resultados con las distribuciones previas de Jeffreys y de Perks son equivalentes.

La última estimación con una distribución previa objetiva se presenta en la Tabla 3.4 a través de la distribución previa de Bayes-Laplace.

Tabla 3.4: Resumen de la distribución posterior de la probabilidad de cosechar un tubérculo asintomático, leve y grave procedente de simientes asintomáticas (filas superiores) y graves (filas inferiores) utilizando la distribución previa de Bayes-Laplace.

probabilidad	media	desviación típica	EMC ($\times 10^{-5}$)	2.5 %	mediana	97.5 %
θ_{11}	0.697	0.010	3.19	0.677	0.697	0.717
θ_{21}	0.228	0.009	3.10	0.210	0.228	0.247
θ_{31}	0.074	0.006	1.82	0.063	0.074	0.086
θ_{12}	0.001	0.003	9.64	0.000	0.000	0.002
θ_{22}	0.197	0.010	3.22	0.178	0.197	0.216
θ_{32}	0.803	0.010	3.24	0.783	0.803	0.821

EMC: error de Monte Carlo.

Análogamente al caso anterior, la estimación del modelo multinomial (3.1) con distribución previa de Bayes-Laplace también presenta resultados casi idénticos a los dos casos anteriores (Perks y Jeffreys).

Aunque las distribuciones previas objetivas analizadas tienen diferentes características matemáticas (por ejemplo, la simetría y la uniformidad), desde el punto de vista aplicado son casi todas equivalentes. La Figura 3.1 muestra una visualización gráfica de la distribución posterior obtenida con cada una de las distribuciones previas objetivas en las diferentes categorías de tubérculos y simientes.

La Figura 3.1 ilustra de manera muy clara, mediante solapamiento, la equivalencia de la distribución posterior con las distribuciones previas de Perks, Jeffreys y Bayes-Laplace, y para el lote `asintomático` también la distribución previa de Haldane. Es fácil ver que la diferencia entre las distribuciones previas ocurre en la Figura 3.1 (d), donde no tenemos observaciones de tubérculos asintomáticos. En este caso, utilizando la distribución previa de Bayes-Laplace tenemos más probabilidades para los valores positivos, mientras que la distribución previa de Perks es la menos informativa (probabilidades concentradas cerca de cero), aunque en nuestro estudio estas diferencias son poco relevantes,

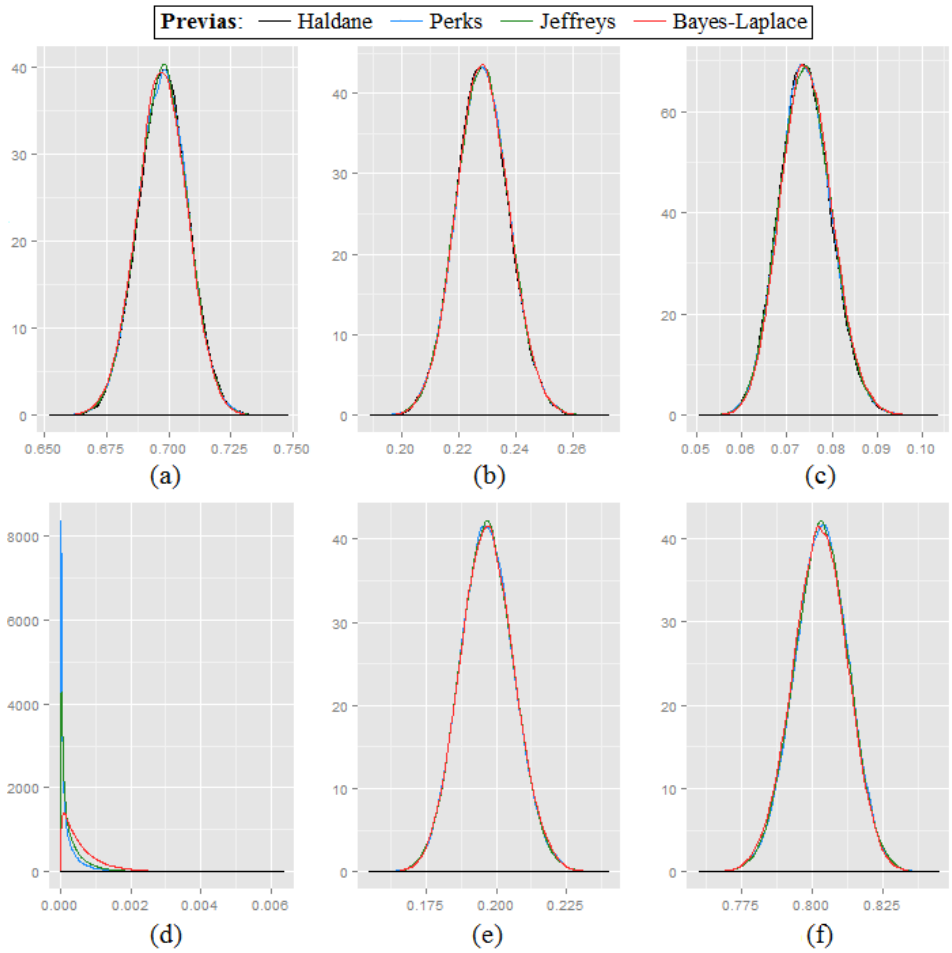

Figura 3.1: Distribución posterior marginal de la probabilidad de cosechar tubérculos asintomáticos [(a) y (d)], con síntomas leves [(b) y (e)] y graves [(c) y (f)] de mancha negra en lotes de simientes asintomáticas y con manchas negras, respectivamente, utilizando distribuciones previas Dirichlet objetivas.

pues la escala de valores es muy pequeña.

Para comparar la sensibilidad en los resultados obtenidos con estas distribuciones previas objetivas y una muy informativa, utilizamos una distribución previa Dirichlet$(100, 100, 100)$ cuya densidad aparece representada en las Figuras 2.9 (d) y 2.10 (d). La Tabla 3.5 muestra un resumen de los resultados de la distribución posterior con dicha distribución previa informativa.

Observe que los resultados globales no se has visto afectados por la información añadida de la distribución previa informativa, que está relacionada con

Tabla 3.5: Resumen de la distribución posterior de la probabilidad de cosechar un tubérculo asintomático, leve y grave procedente de simientes asintomáticas (filas superiores) y graves (filas inferiores) utilizando la distribución previa informativa Dirichlet$(100, 100, 100)$.

probabilidad	media	desviación típica	EMC ($\times 10^{-5}$)	2.5 %	mediana	97.5 %
θ_{11}	0.652	0.010	3.24	0.633	0.652	0.671
θ_{21}	0.241	0.009	2.95	0.224	0.241	0.259
θ_{31}	0.107	0.006	2.10	0.095	0.107	0.119
θ_{12}	0.050	0.005	1.63	0.041	0.050	0.060
θ_{22}	0.217	0.009	3.06	0.199	0.217	0.235
θ_{32}	0.733	0.010	3.43	0.713	0.733	0.752

EMC: error de Monte Carlo.

el equilibrio entre las probabilidades, probablemente debido a que los datos (a través de la función de verosimilitud) tienen carácter altamente informativo. Sin embargo, los valores medios e intervalos de credibilidad de 95 % se han modificado: se aprecia en el lote `asintomático` un aumento en la probabilidad de obtener un tubérculo con síntomas graves de mancha negra y en el lote `enfermo` el incremento ha sido para la probabilidad de obtener un tubérculo asintomático.

Aunque, aparentemente, las diferencias entre este modelo y los que contienen las distribuciones previas objetivas no son tan grandes, la Figura 3.2 muestra que, analizando en pequeña escala, las distribuciones posteriores son extremadamente diferentes en la mayoría de los casos.

En la Figura 3.2, para una visualización más limpia, sólo se ha utilizado la posterior Dirichlet obtenida con la distribución previa de Perks, pero la interpretación es la misma con las otras distribuciones previas objetivas (excepto la distribución previa de Haldane para el lote `enfermo`), como hemos visto en la Figura 3.1.

Además, podemos inferir que las distribuciones posterior de la probabilidad de obtener tubérculos con síntomas leves de mancha negra, Figura 3.2 (b) y (e), son las menos afectadas en ambos lotes. Sin embargo, el hecho de que la teoría Bayesiana pondere las informaciones de los datos (vía verosimilitud) y de la distribución previa, hace que la inclusión de información en la distribución previa cambie la forma de la distribución posterior, como se puede ver en las gráficas de la Figura 3.2.

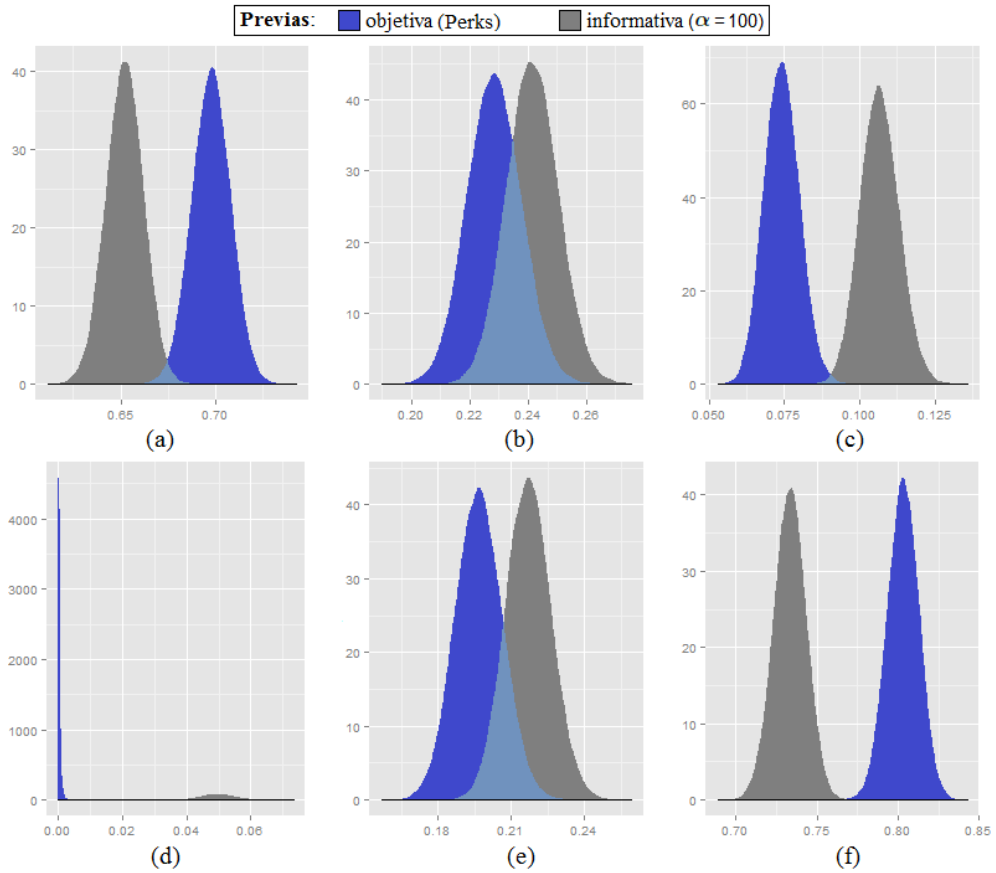

Figura 3.2: Distribución posterior marginal de la probabilidad de cosechar tubérculos asintomáticos [(a) y (d)], con síntomas leves [(b) y (e)] y graves [(c) y (f)] de mancha negra en lotes `asintomático` y `enfermo`, respectivamente, utilizando la distribución previa de Perks y de la distribución previa Dirichlet$(100, 100, 100)$.

También en el aspecto comparativo, la Figura 3.3 muestra las diferencias entre la probabilidad media posterior para cada tipo de tubérculo en cada lote.

La Figura 3.3 deja claro el intento, mediante la información añadida por la distribución previa informativa, de equilibrar las probabilidades de cada tipo de tubérculo. Sin embargo, una vez más, vale la pena destacar que el hecho de que los datos sean muy informativos no permite que la distribución previa tenga un peso determinante en la construcción de la distribución posterior.

El análisis comparativo también se puede hacer a través del análisis gráfico de la distribución posterior Dirichlet. La Figura 3.4 ilustra las diferencias para esta distribución posterior, en el lote `asintomático`, cuando utilizamos la

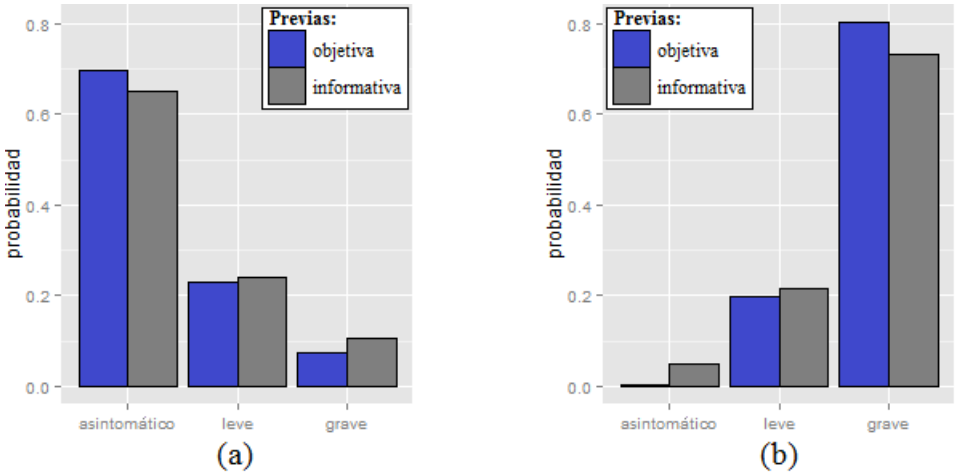

Figura 3.3: Comparación de la probabilidad media posterior de cosechar un tubérculo asintomático, con síntomas leves y graves de mancha negra obtenida a partir de la distribución previa de Perks y de la distribución previa Dirichlet$(100, 100, 100)$. Resultados de los lotes `asintomático` y `enfermo` en las gráficas (a) y (b), respectivamente.

distribución previa de Perks o la distribución previa Dirichlet$(100, 100, 100)$.

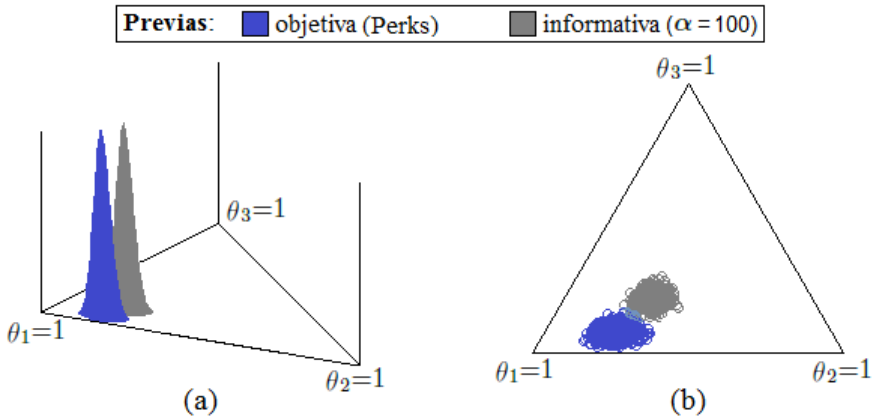

Figura 3.4: Comparación de la distribución posterior Dirichlet (a) y su proyección en el triángulo simplex (b), para el lote `asintomático`, de cosechar un tubérculo asintomático, con síntomas leves y graves de mancha negra obtenida a partir de la distribución previa de Perks y la distribución previa Dirichlet$(100, 100, 100)$.

Es fácil ver que las diferencias identificadas en la distribución posterior marginal (Figura 3.2) se destacan de forma integrada en la Figura 3.4. Además

de la influencia de la distribución previa informativa Dirichlet$(100, 100, 100)$ en el intento de centralizar la distribución posterior, estos gráficos dejan claro que en el lote de simientes asintomáticas el tubérculos con mayor probabilidad de ser cosechado es el tubérculo asintomático (representado por θ_1).

Análogamente, la Figura 3.5 muestra el mismo análisis para el lote `enfermo`.

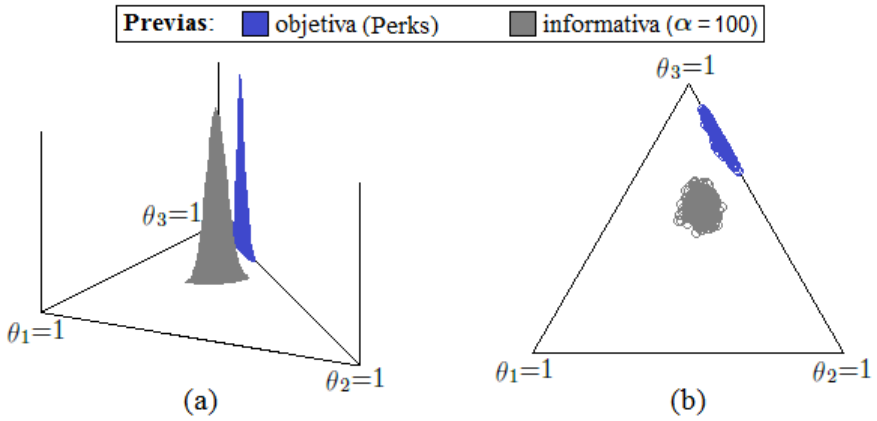

Figura 3.5: Comparación de la distribución posterior Dirichlet (a) y su proyección en el triángulo simplex (b), para el lote `enfermo`, de cosechar un tubérculo asintomático, con síntomas leves y graves de mancha negra obtenida a partir de la distribución previa de Perks y de la distribución previa Dirichlet$(100, 100, 100)$.

Una vez más las diferencias son bastante visibles entre la distribución posterior Dirichlet obtenida a partir de cada una de las dos distribuciones previas analizadas. Además, también podemos reforzar la idea de que al sembrar simientes con manchas negras tenemos mayor propensión a cosechar tubérculos con síntomas graves de mancha negra (representado por θ_3).

CAPÍTULO 4

CONCLUSIONES

Desde el punto de vista agronómico, este estudio ha aportado fuertes evidencias de que la selección de simientes asintomáticas es una buena decisión para mejorar la producción de tubérculos de chufa comercializables y puede considerarse como una medida preventiva inicial en la reducción de tubérculos con mancha negra.

Desde un punto de vista estadístico, hemos explorado el modelo Dirichlet-multinomal a través de las distribuciones previas objetivas – Dirichlet de Haldane, de Perks, de Jeffreys y de Bayes-Laplace – que han mostrado ser equivalentes en términos aplicados. La excepción es la distribución previa de Haldane que ha producido una distribución posterior impropia en una etapa del estudio. Además, también se ha realizado un análisis informativo para comparar con los resultados de las distribuciones previas objetivas.

En suma, los modelos Dirichlet-multinomial Bayesianos con distribuciones previas objetivas son adecuados para el problema propuesto y nos han permitido obtener información relevante acerca de las mejoras en el cultivo de chufas mediante selección de simientes asintomáticas.

Líneas futuras de investigación

Este trabajo es un primer estudio estadístico para identificar algunas características de las manchas negras en tubérculos de la chufa. La estructura de

los datos ha favorecido la exploración de la sensibilidad de las distribuciones previas Dirichlet objetivas para el modelo multinomial. Sin embargo, dada la relevancia del tema estudiado, tanto metodológico como aplicado, creemos que sería interesante continuar el estudio y explorar algunas cuestiones interesantes para trabajos futuros:

- Incorporación de covariables;
- Análisis de coste-efectividad de posibles tratamientos para la reducción o eliminación de las manchas negras;
- Inferencia predictiva en escenarios realistas;
- Sensibilidad de los modelos Dirichlet-multinomial jerárquicos.

BIBLIOGRAFÍA

A. Agresti. *Categorical data analysis*. Probability and Statistics. John Wiley and Sons, 3rd edition, 2013.

N. Balakrishnan and V. Nevzorov. *A primer on statistical distributions*. John Wiley and Sons, 1st edition, 2003.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370 – 418, 1763.

J. O. Berger, J. M. Bernardo, and D. Sun. Overall objective priors. *Bayesian Analysis*, 10(1):189–221, 2015.

J. Bernardo. Reference posterior distributions for Bayesian inference. *Journal Royal Statistical Society B*, 41:113–147, 1979.

W. Bolstad. *Introduction to Bayesian statistics*. John Wiley and Sons, 2nd edition, 2007.

A. Cavanilles. *Observaciones sobre la historia natural, geografia, agricultura, poblaciones y frutos del Reyno de Valencia*. Imprenta Real, Madrid, 1797.

P. Congdon. *Bayesian models for categorical data*. Probability and Statistics. John Wiley and Sons, 1st edition, 2005.

R. Connor and J. Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64:194 – 206, 1969.

O. Ditlevsen and H. Madsen. *Structural reliability methods*. John Wiley and Sons, 1996.

G. Dragendorff. *Die heilplanzen der verschiedenen völker und zeiten: ihre anwendung, wesentlichen bestandtheile und geschichte (The medical plants of various peoples and times)*. Verlag von Ferdinand Enke, Stuttgart, 1898.

D. Gamerman and H. Lopes. *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*. Texts in Statistical Science. Chapman and Hall/CRC, 2nd edition, 2006.

A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457 – 472, 1992.

I. Good. *The estimation of probabilities: an essay on modern Bayesian methods*. MIT Press, 1st edition, 1965.

J. Haldane. The precision of observed values of small frequencies. *Biometrika*, 35:297 – 300, 1948.

H. Hoijtink, I. Klugkist, and P. Boelen. *Bayesian evaluation of informative hypotheses*. Statistics for Social and Behavioral Sciences. Springer, New York, 2008.

E. T. Jaynes. *Probability theory: the logic of science*. Cambridge University Press, 2003.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, 186:453 – 461, 1946.

H. Jeffreys. *Theory of probability*. Clarendon Press, Oxford, 1961.

P. Laplace. *Théorie analytique des probabilités*. Courcier, Paris, 1812.

D. Lindley. The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, 35:1622 – 1643, 1964.

Q. Liu, J. Peng, and A. T. Ihler. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems 25*, pages 692 – 700. Curran Associates, Inc., 2012.

N. Luján and J. Perucho. *El libro de la cocina española*. Tusquets Editores, 2nd edition, 2003.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. Winbugs – A Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325 – 337, 2000.

P. McCullagh and J. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 2nd edition, 1989.

J. Morell and S. Barber. Chufa y horchata: características físicas, químicas y nutritivas. Technical report, Institute of Agrochemistry and Food Technology, Valencia, Spain, 1983.

M. Negbi. A sweetmeat plant, a perfume plant and their weedy relatives: a chapter in the history of cyperus esculentus l. and cyperus rotundus l. *Economic Botany*, 46:64 – 71, 1992.

J. Nelder and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society A*, 135:370 – 384, 1972.

W. Perks. Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, 73:285 – 334, 1947.

J. Serrallach. *Die wurzelknolle von Cyperus esculentus L. (The root tuber of Cyperus esculentus L.)*. PhD thesis, University Frankfurt am Main, 1927.

A. Stuart and K. Ord. *Kendall's advanced theory of statistics*, volume 1 of *Distribution Theory*. John Wiley and Sons, 6th edition, 1994.

A. Zellner. *An introduction to Bayesian inference in econometrics*. John Wiley and Sons, 1996.

APÉNDICE

MODELOS EN WINBUGS

Modelo multinomial con la distribución previa de Haldane

```
model{
# Modelo Multinomial
    for(j in 1:J){
        for(i in 1:I){
            N[j,i] <- sum(y[j,i,1:K])
            y[j,i,1:K] ~ dmulti(theta.hal[j,1:K],N[j,i])
        }
    }
# Previa de Halane
    for(j in 1:J){
        theta.hal[j,1:K] ~ ddirch(alpha.hal[j,1:K])
        for(l in 1:K){
            alpha.hal[j,l] <- 0
        }
    }
}
```

Modelo multinomial con la distribución previa de Perks

```
model{
# Modelo Multinomial
    for(j in 1:J){
        for(i in 1:I){
            N[j,i] <- sum(y[j,i,1:K])
            y[j,i,1:K] ~ dmulti(theta.per[j,1:K], N[j,i])
        }
    }
# Previa de Perks
    for(j in 1:J){
        theta.per[j,1:K] ~ ddirch(alpha.per[j,1:K])
        for(l in 1:K){
            alpha.per[j,l] <- 1/K
        }
    }
}
```

Modelo multinomial con la distribución previa de Jeffreys

```
model{
# Modelo Multinomial
    for(j in 1:J){
        for(i in 1:I){
            N[j,i] <- sum(y[j,i,1:K])
            y[j,i,1:K] ~ dmulti(theta.jef[j,1:K], N[j,i])
        }
    }
# Previa de Jeffreys
    for(j in 1:J){
        theta.jef[j,1:K] ~ ddirch(alpha.jef[j,1:K])
        for(l in 1:K){
            alpha.jef[j,l] <- 1/2
        }
    }
}
```

Modelo multinomial con la distribución previa de Bayes-Laplace

```
model{
# Modelo Multinomial
    for(j in 1:J){
        for(i in 1:I){
            N[j,i] <- sum(y[j,i,1:K])
            y[j,i,1:K] ~ dmulti(theta.bl[j,1:K], N[j,i])
        }
    }
# Previa de Bayes-Laplace
    for(j in 1:J){
        theta.bl[j,1:K] ~ ddirch(alpha.bl[j,1:K])
        for(l in 1:K){
            alpha.bl[j,l] <- 1
        }
    }
}
```

Modelo multinomial con la distribución previa Dirichlet informativa

```
model{
# Modelo Multinomial
    for(j in 1:J){
        for(i in 1:I){
            N[j,i] <- sum(y[j,i,1:K])
            y[j,i,1:K] ~ dmulti(theta.inf[j,1:K], N[j,i])
        }
    }
# Previa informativa (alpha=100)
    for(j in 1:J){
        theta.inf[j,1:K] ~ ddirch(alpha.inf[j,1:K])
        for(l in 1:K){
            alpha.inf[j,l] <- 100
        }
    }
}
```