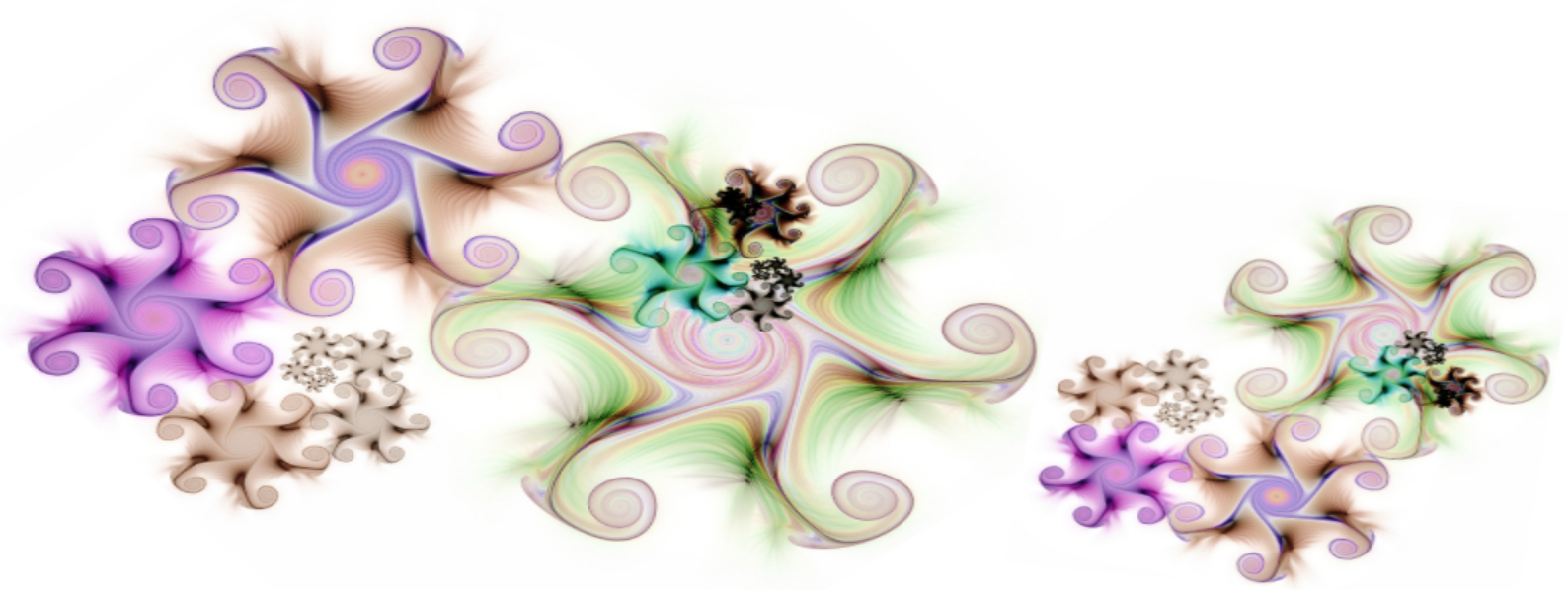
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Iztapalapa

Mixba'al
Revista Metropolitana de Matemáticas
www.doi.org/10.24275/uami/dcbi/mix/v16n1/asafmar ISSN: 2007-7874

Ciencias Básicas e Ingeniería CBI



# *Análisis de datos con un enfoque bayesiano*

Asael Fabian Martínez Martínez

## 7° Coloquio del Departamento de Matemáticas UAM-I

Del 27 al 31 de enero de 2025

Universidad Autónoma Metropolitana

Casa abierta al tiempo | UNIVERSIDAD AUTÓNOMA METROPOLITANA



MIXBA'AL. Vol. 16, No. 1, enero-diciembre de 2025, es una publicación anual de la Universidad Autónoma Metropolitana a través de la Unidad Iztapalapa, División de Ciencias Básicas e Ingeniería, Departamento de Matemáticas. Prolongación Canal de Miramontes 3855, Col. Ex Hacienda San Juan de Dios, Alcaldía Tlalpan, C.P. 14387, CDMX, México y Av. Ferrocarril San Rafael Atlixco, No. 186, Col. Leyes de Reforma 1a Sección, Alcaldía Iztapalapa, C.P. 09340, CDMX, México. Tel. 5804 4658. Página electrónica de la revista: http://mat.izt.uam.mx/mat/ index.php/revistamixba-al. Correos electrónicos: mixbaal2009@gmail.com, mixb@xanum.uam.mx. Coordinador Editorial Mario Pineda Ruelas. Certificado de Reserva de Derechos al Uso Exclusivo de Título No. 04-2023-07031 1572300-102, ISSN5 2007-7874, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este número Mario Pineda Ruelas, Departamento de Matemáticas, edificio AT, oficina 318. División de Ciencias Básicas e Ingeniería, Universidad Autónoma Metropolitana-Iztapalapa. Av. Ferrocarril San Rafael Atlixco No. 186,Colonia Leyes de Reforma 1a Sección, Alcaldía Iztapalapa, C.P. 09340, CDMX, México. Fecha de última modificación 30 de agosto de 2025. Tamaño del archivo 121.3 MB.


# 7º COLOQUIO DEL DEPARTAMENTO DE MATEMÁTICAS

del 27 al 31 de enero del 2025, Unidad Iztapalapa de la UAM, Ciudad de México

Departamento de Matemáticas
UAM IZTAPALAPA

Posgrado de Matemáticas
UAM IZTAPALAPA

## Taller 1: Análisis de datos con un enfoque bayesiano

Autor: Dr. Asael Fabian Martínez Martínez

## Taller 2: Un breve recorrido por la teoría de prerradicales y sus retículas

Autores: Dr. Rogelio Fernández-Alonso, Dra. Silvia Gavito Ticozzi, Dra. Martha Lizbeth Shaid Sandoval Miranda

## Taller 3: Resultados del cálculo y álgebra lineal relevantes en modelos y aplicaciones

Autor: Dr. Lorenzo Héctor Juárez

## Taller 4: Análisis geométrico de superficies: una introducción elemental

Autores: Dr. Josué Melendez y M. en C. Eduardo Rodríguez Romero

## Taller 5: Una introducción a los torneos y sus generalizaciones

Autores: Dr. Ilán Goldfeder Ortiz y Dra. Nahid Yelene Javier Nol

## Taller 6: Del cero al quantum: un viaje por el mundo de los códigos

Autores: Dr. Jorge Bolaños Servín, Dra. Yuriko Pitones Amaro y Dr. Josué Rios Cangas

## Taller 7: Introducción a la teoría de juegos epistémica

Autor: Dr. Rubén Becerril Borja

## Taller 8: Cómo contar más allá del infinito y para qué sirve. Inducción transfinita y algunas aplicaciones

Autor: Dr. Rodrigo Hernández Gutiérrez

UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Iztapalapa

Ciencias Básicas e Ingeniería CBI

7° Coloquio del Departamento de Matemáticas UAM-I 2025
Universidad Autónoma Metropolitana, Iztapalapa CDMX.

# Introducción

En muchas de nuestras actividades diarias debemos tomar decisiones. Estas pueden no ser simples de tomar debido a que vivimos rodeados de eventualidades que nos imposibilitan conocer con certeza el rumbo de las cosas. Dentro de las Matemáticas, existen dos ramas de estudio que fueron concebidas para tratar de afrontar estas situaciones: la Probabilidad y la Estadística. La Probabilidad provee las herramientas para modelar la «incertidumbre», mientras que la Estadística utiliza la evidencia disponible y, echando mano de las herramientas probabilísticas, nos ayuda a tomar decisiones informadas. Pero vamos más allá. En general, debido a que muchos fenómenos —naturales o sociales— se desarrollan bajo ambientes de incertidumbre y es necesario tener un mejor entendimiento de ellos; los interesados recurren a la Estadística para contestar sus preguntas acerca de estos.

Ahora bien, un elemento fundamental para lograr lo anterior son los «datos». Sin entrar en demasiados detalles, existe un proceso largo y complejo para determinar qué datos utilizar que nos sean útiles o nos permitan responder la pregunta o investigación de interés —formulada como una hipótesis—; es un proceso de abstracción en el cual intervienen diferentes disciplinas y cuyo propósito es definir las variables y los métodos a utilizar que vayan de acuerdo con la hipótesis y permitan confirmarla o refutarla. El siguiente paso consiste en hacer las mediciones de las variables, esto es —en un lenguaje estadístico—, registrar las observaciones. Cada observación es un dato, nuestra materia prima para utilizar cualquier procedimiento estadístico. Con base en los datos, obtendremos información que nos permitirá dar una respuesta a nuestra interrogante; en términos estadísticos, haremos inferencias.

Si bien la descripción anterior ha sido la forma de trabajo habitual para quienes hacen Estadística —aunque su participación en las diferentes etapas puede darse en mayor o menor medida— notemos que, desde otra perspectiva, la aplicación de este flujo de trabajo nos lleva a «aprender» acerca de cierto fenómeno de interés. Podríamos utilizar algún otro método —no estadístico— para analizar los datos, y aún así aprender; la clave es poder extraer información de los datos. De esta manera se tienen métodos computacionales —englobados en lo que se conoce como Inteligencia Artificial, y más cercano a nuestros propósitos, el Aprendizaje Automático (del inglés *Machine Learning*)— cuyo propósito es este mismo: lograr un «aprendizaje» sobre un fenómeno y que nos permita contestar nuestras preguntas de interés.

Este preámbulo sirve para presentar el tema principal de este documento: el análisis de datos bajo una perspectiva estadística. En particular, seguiremos un enfoque bayesiano no paramétrico. No entraremos en los detalles de cómo se obtienen los datos ni si son adecuados para el problema en cuestión. Nuestro propósito será dar las bases matemáticas para realizar una tarea de gran utilidad: detectar patrones en los datos de tal manera que nos permitan agruparlos; estamos hablando del «análisis de conglomerados», para los estadísticos, o del «aprendizaje no supervisado», para los computólogos. El análisis de conglomerados es de gran utilidad, ya que nos permite, por ejemplo, identificar distintos «comportamientos» dentro de un mismo fenómeno y gracias a algún análisis posterior por grupo, se pueden entender mejor esas diferencias.

De esta manera, en los siguientes capítulos presentaremos los fundamentos de la inferencia bayesiana —enfocándonos en los modelos «no paramétricos»—. Esta perspectiva de la Estadística se basa fuertemente en la Probabilidad; nosotros, en particular, nos fundamentaremos en el Teorema de representación de de Finetti para variables aleatorias intercambiables. Una ventaja de este teorema es que justifica matemáticamente los modelos no paramétricos. Esto se abordará en el Capítulo 1.

Un elemento primordial para estos modelos es la distribución inicial —que toma valores en un espacio de distribuciones de probabilidad—. El proceso Dirichlet ha sido la piedra angular para definir distribuciones de este tipo. Además, ha permitido el desarrollo de nuevas «distribuciones aleatorias» y también ha permitido establecer una conexión natural entre la inferencia bayesiana no paramétrica y el análisis de conglomerados. En el Capítulo 2, exploraremos algunas distribuciones aleatorias, mientras que su conexión con el análisis de conglomerados se estudiará en el Capítulo 3.

Asael Fabian Martínez Martínez
fabian@xanum.uam.mx
Departamento de Matemáticas
UAM-Iztapalapa

# Índice general

# Capítulo 1

# Inferencia Bayesiana

Como hemos dicho, la Estadística es una rama de las Matemáticas que nos permite estudiar fenómenos aleatorios o que se realizan bajo ambientes de incertidumbre. En algunas situaciones, una vez definidas las variables relevantes para el estudio, se obtiene una muestra de éstas —un conjunto de datos—; otras veces, los datos fueron registrados o generados previamente —no precisamente para nuestros fines— y sólo se utilizarán. En cualquier caso, la aleatoriedad inherente al fenómeno se verá reflejada en los datos y necesitamos poder trabajar con ella para obtener respuestas —o inferencias—, hablando estadísticamente.

Existen dos enfoques estadísticos para realizar inferencias, el más conocido se denomina «frecuentista» o «clásico», mientras que el segundo enfoque se denomina «bayesiano». Ambos utilizan a la Probabilidad en sus metodologías pero de diferente manera; nosotros estudiaremos el segundo enfoque, ya que la utiliza de manera integral para modelar la aleatoriedad del fenómeno. De esta forma, el propósito de este primer capítulo es presentar los elementos para realizar inferencias bajo el enfoque bayesiano. Nos concentraremos en los modelos denominados «no parámetricos», en donde la distribución inicial es una «distribución aleatoria». Comenzaremos con algunos conceptos básicos de Probabilidad.

Una explicación más detallada de la inferencia bayesiana «paramétrica» se puede encontrar en libros como [3], [20], [6] o [45] —este último siendo de un nivel más avanzado—; mientras que para modelos no paramétricos, se puede consultar [19] o [14].

## 1.1 Incertidumbre y Probabilidad

Existe una diferenciación —aunque un tanto informal— en la incertidumbre inherente a los fenómenos aleatorios, dividiéndola en las denominadas «incertidumbre aleatoria» e «incertidumbre epistémica» (véase, por ejemplo, [29] o [16]). La incertidumbre aleatoria se refiere a una variabilidad intrínseca del fenómeno; la incertidumbre epistémica, por otro lado, está relacionada con la falta de información. En los modelos de inferencia estadística, podemos relacionar a la incertidumbre aleatoria con los supuestos distribucionales sobre las variables (aleatorias) que conforman la muestra, mientras que la incertidumbre epistémica se ve reflejada en el parámetro de interés o los supuestos que hagamos acerca de él. En ambos casos, la incertidumbre queda descrita en términos de distribuciones de probabilidad, y es a través de ellas como tendríamos que obtener nuestras respuestas.

Podemos entender el proceso de inferencia estadística como sigue (una explicación más amplia en esta línea se puede ver en el Capítulo 4 de [6]). Supongamos que tenemos una colección de eventos *no observables* $B_1, \ldots, B_d$ que forman una partición del universo. Al asignar probabilidades a cada evento $B_j$, i.e. $\mathbb{P}(B_j)$, estamos cuantificando su ocurrencia de acuerdo con nuestras creencias. Ahora supongamos que ha ocurrido —hemos observado— un evento $A$.

Dado este evento $A$, nos es posible calcular la probabilidad de cada evento $B_j$ condicionado a $A$, la cual denotamos por $\mathbb{P}(B_j \mid A)$ para $j = 1, \ldots, d$. Si comparamos ambas probabilidades sobre los eventos $B_j$, $\mathbb{P}(B_j)$ no está influenciada por la ocurrencia de $A$ —por lo que se les dice probabilidades *a priori* o iniciales—, pero $\mathbb{P}(B_j \mid A)$ sí lo está —a éstas se les conoce como probabilidades *a posteriori* o finales—. Conceptualmente, nos parece sensato decir que ambas probabilidades están relacionadas para cualquier evento $B_j$ fijo; formalmente, ¿cómo lo podemos describir? Una manera es utilizando el Teorema de Bayes.

**Teorema 1.1** (Bayes). *Sean $A$ y $B$ dos eventos tales que $\mathbb{P}(A) \neq 0$. Entonces*

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

La probabilidad condicional $\mathbb{P}(A \mid B)$ recibe el nombre de *verosimilitud* del evento $B$, y es el peso asignado a $B$ determinado por la ocurrencia de $A$. La probabilidad (marginal) de $A$, $\mathbb{P}(A)$, se puede obtener a través de la ley de probabilidad total.

*Definición* 1.2 (Ley de probabilidad total). Sea $B_1, \ldots, B_d$ una partición del universo, entonces la probabilidad de cualquier evento $A$ se puede calcular como

$$\mathbb{P}(A) = \sum_{j=1}^{d} \mathbb{P}(A \cap B_j) = \sum_{j=1}^{d} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j).$$

Del Teorema de Bayes podemos notar la relación entre las probabilidades *a priori* y *a posteriori* de $B_j$: las segundas combinan la verosimilitud del evento con nuestras creencias iniciales acerca de su ocurrencia, en otras palabras, reflejan cómo fueron «actualizadas» nuestras creencias iniciales acerca de $B_j$ a la luz de la ocurrencia de $A$.

Ahora bien, esta explicación se puede trasladar a un contexto estadístico de inferencia. Supongamos que tenemos interés en alguna característica de la población, denotada por $\theta$; no es observada, por lo que modelamos nuestra incertidumbre acerca de ella utilizando alguna distribución de probabilidad —esto es, la consideramos como una variable aleatoria—. Por simplicidad, podemos suponer que toma un número finito de $d$ valores diferentes, $\vartheta_j$, $j = 1, \ldots, d$. Entonces, los eventos $B_j$ se pueden definir como $B_j = \{\theta = \vartheta_j\}$. Por otro lado, asumimos que existe alguna variable (aleatoria) $Y$ que está relacionada con $\theta$; podemos obtener varias *realizaciones* de ésta —una muestra, en términos estadísticos— y (esperamos que) nuestro conocimiento acerca del valor de $\theta$ mejore conforme aumente la cantidad de realizaciones. Englobamos todo esto en el evento $A$. De esta forma, hemos obtenido una metodología —basada en el Teorema de Bayes— para la estimación de parámetros. Las probabilidades *a posteriori* de $B_j$, condicionadas a $A$, contienen toda la información que disponemos acerca de $\theta$ una vez que se ha incorporado a la información *a priori* la evidencia proporcionada por la muestra.

# 1.2 Intercambiabilidad

Hasta el momento no hemos detallado la forma o estructura del evento observable $A$, únicamente mencionamos que tiene que ver con la muestra. Para contestar nuestra pregunta de interés, se determinó que existe una variable $Y$ que nos proporciona información sobre ella; además, ésta también se desarrolla bajo un ambiente de incertidumbre, por lo que podemos modelarla como variable aleatoria. La muestra queda definida, entonces, como la observación repetida de esta variable, es decir

$$A = \{Y_1 = y_1, \ldots, Y_n = y_n\} := \{Y_1 = y_1\} \cap \cdots \cap \{Y_n = y_n\}$$

para algún valor de $n > 0$. El subíndice nos sirve para diferenciar las repeticiones. Por el momento, asumimos que la variable puede tomar una cantidad numerable de valores —similarmente al parámetro $\theta$— pero ambos se pueden extender al caso continuo.

Lo que resta es determinar las probabilidades asociadas al evento observable $A$. Para ello, consideremos el siguiente escenario. En el contexto descrito anteriormente, supongamos que observamos otro evento $A^*$ bajo las mismas condiciones de $A$ y estos no tienen relación alguna —siendo más formales, estamos asumiendo que $A$ y $A^*$ son condicionalmente independientes dado cualquier evento no observable $B_j$, $j = 1, \ldots, d$—. Podemos definir a $A^*$ como la observación de una muestra *ampliada* de $Y$, i.e. $A^* = \{Y_{n+1} = y_{n+1}, \ldots, Y_{n+m} = y_{n+m}\}$ para algún valor de $m > 0$. Ahora, nuestro interés es poder decir algo acerca del comportamiento de la nueva muestra, $A^*$, después de haber observado $A$.

Probabilísticamente, nuestro interés lo podemos expresar como la probabilidad condicional de $A^*$ dado $A$, esto es

$$\mathbb{P}(A^* \mid A) = \frac{\mathbb{P}(A^* \cap A)}{\mathbb{P}(A)}. \tag{1.1}$$

Suena razonable pensar que lo que podamos decir de $A^*$ después de haber observado $A$ se vea influido –de alguna manera— justamente por la evidencia proporcionada por $A$. Si deseamos que así suceda, no es buena idea asumir que los eventos $A^*$ y $A$ son independientes, porque tendríamos que $\mathbb{P}(A^* \mid A) = \mathbb{P}(A^*)$. Por tanto, necesitamos relajar el supuesto de independencia para obtener algún «aprendizaje».

## 1.2.1 Teorema de representación

Para continuar con la presentación de ideas, es necesario que cambiemos el enfoque, de eventos a variables aleatorias. Los eventos observables —como ya se ha hecho— quedarán expresados en términos de muestras de variables aleatorias $Y_i$, y los eventos no observables se asociarán a un parámetro $\theta$ —que también se considerará aleatorio para fines de modelado—.

La distribución predictiva nos indica que asumir independencia entre las variables $Y_i$, $i = 1, 2, \ldots$ impide que aprendamos a partir de un conjunto inicial de ellas. Requerimos, por tanto, introducir alguna clase de dependencia estocástica. Nos concentraremos en uno conocido como «intercambiabilidad». (No profundizaremos en este tema; un tratamiento más completo y riguroso se puede encontrar, por ejemplo, en [1], [45] y [23].)

*Definición* 1.3 (Intercambiabilidad). Una sucesión de variables aleatorias $X_1, \ldots, X_n$ se dice que es *finitamente intercambiable* si —para cualquier permutación $\rho$ de $\{1, \ldots, n\}$,

$$(X_1, \ldots, X_n) \overset{\mathrm{d}}{=} (X_{\rho(1)}, \ldots, X_{\rho(n)}),$$

donde $\overset{\mathrm{d}}{=}$ denota la igualdad en distribución. Una sucesión infinita $X_1, X_2, \ldots$ de variables aleatorias se dice *intercambiable* si toda subsucesión finita lo es.

Al asumir intercambiabilidad, estamos diciendo que el orden en que se registran las observaciones es indistinto. Para muchas situaciones prácticas, esto tiene sentido ya que la información que nos proporcione una colección de observaciones no debería depender de cuál observación se registró antes o después, sino de todas en conjunto. Lo que resta es poder asignar probabilidades.

Una manera de definir probabilidades para sucesiones intercambiables es a través del conocido «teorema de representación». La primera versión de este teorema fue demostrada por el estadístico y actuario italiano Bruno de Finetti [7] utilizando variables aleatorias binarias —i.e. variables aleatorias Bernoulli—. Presentamos este caso a continuación y posteriormente estudiaremos su generalización.

**Teorema 1.4.** *Una sucesión de variables aleatorias $X_1, X_2, \ldots$ con valores en $\mathbb{X} = \{0, 1\}$ es intercambiable si y sólo si existe una distribución de probabilidad $\pi$ en $[0, 1]$, tal que, para toda $n \geq 1$,*

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^{s_n} (1 - \theta)^{n - s_n} \pi(\mathrm{d}\theta),$$

*donde $s_n = x_1 + \cdots + x_n$. Además, $\pi$ es la distribución de $\lim_{n \to \infty} \frac{s_n}{n}$.*

Una consecuencia inmediata de este teorema es que las variables $X_1, \ldots, X_n$ son condicionalmente independientes dado $\theta$. Si cada $X_i$, dado $\theta \in (0, 1)$, se distribuye Bernoulli de parámetro $\theta$, tenemos que $\mathbb{P}(X_i = x_i \mid \theta) = \theta^{x_i}(1 - \theta)^{1 - x_i}$, por lo que

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid \theta) = \theta^{s_n}(1 - \theta)^{n - s_n}.$$

Notemos que esta última expresión corresponde con la verosimilitud mencionada previamente, i.e. $\mathbb{P}(A \mid B_j)$. Similarmente, $\pi$ corresponde a la distribución *a priori* $\mathbb{P}(B_j)$, por lo que tenemos completos los ingredientes para obtener la distribución *a posteriori* del parámetro $\theta$ dada una muestra —de variables binarias, en este caso— de tamaño $n$, esto es

$$\mathbb{P}(\theta \mid Y_1 = y_1, \ldots, Y_n = y_n) = \frac{\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta)\pi(\theta)}{\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n)}.$$

El teorema de representación fue generalizado por Hewitt y Savage [18] para espacios medibles. Como explicaremos más adelante, este resultado constituye el fundamento teórico de la estadística bayesiana no paramétrica; o al menos, es una manera de justificar este modo de hacer inferencias.

**Teorema 1.5** (Teorema de Representación). *Sea $\mathscr{M}$ el espacio de todas las medidas de probabilidad en $(\mathbb{X}, \mathscr{X})$. Se dice que una sucesión de variables aleatorias $X_1, X_2, \ldots$ es* intercambiable *si*

*y sólo si existe una medida $Q$ en $\mathscr{M}$ tal que, para toda $n \geq 1$,*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathscr{M}} \prod_{i=1}^{n} P(A_i) Q(\mathrm{d}P), \tag{1.2}$$

*con $A_i \in \mathscr{X}$. Además, la distribución $P$ es igual al límite de la distribución empírica $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ y la distribución $Q$ es única.*

Similarmente al caso de variables aleatorias binarias, una consecuencia de este teorema es que las variables $X_1, \ldots, X_n$ son condicionalmente independientes dado $P$.

# 1.3 Modelos paramétricos de inferencia

A pesar de contar con un marco teórico para la inferencia bayesiana a través del Teorema de Representación —que se llegó a denominar «no paramétrico»—, por mucho tiempo se restringió la modelación a familias paramétricas. Uno de los principales problemas era la falta de modelos concretos para $Q$.

Analicemos brevemente el enfoque paramétrico de la inferencia bayesiana —lo que nos servirá para terminar de aterrizar ideas— comenzando con una versión *simplificada* del Teorema de Representación 1.5. Definamos una familia paramétrica de distribuciones

$$\mathscr{F}_\Theta := \{F_\theta : \theta \in \Theta\},$$

donde $\theta$ es el parámetro de $F$, de dimensión finita, y toma valores en $\Theta$, y supongamos que $Q(\mathscr{F}_\Theta) = 1$. La integral de (1.2) se simplifica en varios aspectos: ésta se hará sobre el espacio $\Theta$ y los posibles valores de $P$ son la misma distribución $F$ pero variando su parámetro, lo que modifica también a $Q(\mathrm{d}P)$. Por tanto, obtenemos la siguiente expresión:

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\Theta} \prod_{i=1}^{n} F_\theta(A_i) \pi(\mathrm{d}\theta),$$

y lo que se garantiza es la existencia de la distribución $\pi$ sobre el espacio $\Theta$. Un estudio más formal de esto se puede ver, por ejemplo, en [3].

Hasta este punto tenemos todos los elementos para explicar completamente la inferencia bayesiana —paramétrica, por el momento—. Al querer hacer inferencias sobre algún fenómeno aleatorio, nuestro interés debería de quedar reflejado en la distribución conjunta de las variables observables ya que es la que tiene relación más directa con lo que sí podemos medir y el fenómeno, además, es la que utilizaríamos para hacer predicciones —como quedó expresado en la Ecuación (1.1)—, lo cual nos indica que no podemos asumir independencia y recurrimos al supuesto de intercambiabilidad. Para fines prácticos, asumir intercambiabilidad sólo nos dice que no importa el orden del muestreo,

pero también nos permite obtener una manera alternativa para hacer predicciones, pues

$$\begin{aligned}
\mathbb{P}(Y_{n+1} \in A_{n+1} \mid Y_1 \in A_1, \ldots, Y_n \in A_n) &= \frac{\mathbb{P}(Y_1 \in A_1, \ldots, Y_{n+1} \in A_{n+1})}{\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n)} \\
&= \frac{\int_\Theta \prod_{i=1}^{n+1} \mathbb{P}(Y_i \in A_i \mid \theta) \pi(\mathrm{d}\theta)}{\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n)} \\
&= \int_\Theta \mathbb{P}(Y_{n+1} \in A_{n+1} \mid \theta) \frac{\prod_{i=1}^{n} \mathbb{P}(Y_i \in A_i \mid \theta) \pi(\mathrm{d}\theta)}{\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n)} \\
&= \int_\Theta \mathbb{P}(Y_{n+1} \in A_{n+1} \mid \theta) \frac{\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n \mid \theta) \pi(\mathrm{d}\theta)}{\mathbb{P}(Y_1 \in A_1, \ldots, Y_n \in A_n)} \\
&= \int_\Theta \mathbb{P}(Y_{n+1} \in A_{n+1} \mid \theta) \pi(\mathrm{d}\theta \mid Y_1 \in A_1, \ldots, Y_n \in A_n).
\end{aligned}$$

Esta última igualdad nos indica que la «distribución predictiva» queda expresada en términos de la verosimilitud de la nueva observación $Y_{n+1}$ y la distribución posterior de $\theta$ dada la muestra.

Por lo general —al hacer inferencias— el fenómeno de interés se caracteriza a través de un parámetro $\theta$, al cual se le asigna una distribución inicial y el modelo de los datos depende de éste, y nuestro conocimiento acerca del parámetro se actualiza a través del Teorema de Bayes. Operativamente, funciona. Sin embargo, el Teorema de Representación nos garantiza la existencia, y unicidad, de la distribución inicial $\pi$, es decir, este teorema nos proporciona todo el fundamento teórico para las inferencias.

Esquemáticamente, un modelo bayesiano se expresa especificando las distribuciones de probabilidad asignadas a cada elemento aleatorio involucrado, así como la dependencia entre ellos. Mínimamente se requiere indicarlo para las variables observables $Y_i$, $i = 1, \ldots, n$, y para el parámetro, $\theta$, de su distribución. Así, tenemos

$$\begin{aligned}
Y_i \mid \theta &\sim F_\theta(Y_i) \text{ [iid]} \quad i = 1, \ldots, n, \\
\theta &\sim \pi.
\end{aligned}$$

# 1.4 Inferencia no paramétrica

Un aspecto crucial en la modelación bayesiana es la selección de la distribución inicial para el parámetro $\theta$. Es claro que, por ejemplo, el soporte de la distribución posterior correspondiente depende fuertemente de aquel definido para la distribución inicial. Además, uno de los propósitos de la distribución inicial es incorporar alguna información sobre el fenómeno bajo estudio. En los modelos paramétricos, es relativamente fácil entender estas implicaciones y existen diversas propuestas para tratarlas. Sin embargo, su extrapolación a los modelos no paramétricos no ha sido trivial.

De manera muy general —en un método estadístico no paramétrico— los supuestos acerca de la distribución de las variables observables son mínimos. Bajo un enfoque bayesiano, esto significaría trasladar la incertidumbre del parámetro $\theta$ al modelo de los datos, i.e. la distribución $F$. Más aún, necesitaríamos ser capaces de definir una distribución inicial cuyas realizaciones sean distribuciones de probabilidad, para luego obtener la distribución posterior correspondiente.

La justificación teórica del enfoque bayesiano no paramétrico se tiene por el Teorema de Representación 1.5, ya que se garantiza la existencia y unicidad de una «distribución aleatoria» $Q$

—esto es, una distribución cuyas realizaciones son distribuciones de probabilidad sobre el espacio de las observaciones— que modela la distribución, marginal, de las variables observables, $P$. Esta descripción, comúnmente, se escribe

$$\begin{aligned} Y_i \mid P &\sim P \text{ [iid]} \quad i = 1, \ldots, n, \\ P &\sim Q. \end{aligned}$$

La distribución posterior se obtiene simplemente al aplicar el Teorema de Bayes. A pesar de esto, la falta de casos concretos para la distribución aleatoria $Q$ —y, posteriormente, la dificultad de obtener muestras de la distribución posterior respectiva— frenó el desarrollo de los modelos no paramétricos.

### 1.4.1 Proceso Dirichlet

Uno de los primeros trabajos sobre la construcción de distribuciones de probabilidad aleatorias fueron las distribuciones *tailfree*, presentadas en [12]. Sin embargo, el proceso Dirichlet, introducido por Ferguson [10], fue el primer ejemplo concreto y manejable analíticamente para hacer inferencias, y se considera la piedra angular de la estadística bayesiana no paramétrica. Este proceso se basa en la distribución Dirichlet.

*Definición* 1.6 (Distribución Dirichlet). Sea $X_1, \ldots, X_n$ una sucesión de variables aleatorias independientes con distribución Gamma con parámetros $(a_i, 1)$, donde $a_i \geq 0$ para toda $i$ y $a_i > 0$ para al menos una $i$, $i = 1, \ldots, n$. La *distribución Dirichlet* con parámetro $a = (a_1, \ldots, a_n)$ es la distribución del vector $(Z_1, \ldots, Z_n)$, donde

$$Z_j = \frac{X_j}{\sum_{i=1}^{n} X_i}, \quad j = 1, \ldots, n.$$

Si $a_i > 0$ para toda $i$, la distribución $(n-1)$-dimensional del vector $(Z_1, \ldots, Z_{n-1})$ es absolutamente continua con densidad

$$f(z_1, \ldots, z_{n-1} \mid a_1, \ldots, a_n) = \frac{\Gamma(a_1 + \cdots + a_n)}{\Gamma(a_1) \cdots \Gamma(a_n)} \left( \prod_{j=1}^{n-1} z_j^{a_j - 1} \right) \left( 1 - \sum_{j=1}^{n-1} z_j \right)^{a_n - 1},$$

para $(y_1, \ldots, y_{n-1}) \in \{(z_1, \ldots, z_{n-1}) : z_j \geq 0, \sum_{j=1}^{n-1} z_j \leq 1\}$.

Utilizando la distribución Dirichlet, Ferguson [10] define el «proceso Dirichlet» a partir de sus distribuciones *finito dimensionales* y demuestra su existencia a través de las condiciones de consistencia de Kolmogorov.

*Definición* 1.7 (Proceso Dirichlet). Sea $\alpha$ una medida finita, no nula sobre un espacio medible $(\mathbb{X}, \mathscr{X})$. Se dice que $P$ es un *proceso Dirichlet* en $(\mathbb{X}, \mathscr{X})$ con parámetro $\alpha$, si, para cada $k \geq 1$ y cada partición medible $B_1, \ldots, B_k$ de $\mathscr{X}$, la distribución del vector $(P(B_1), \ldots, P(B_k))$ es Dirichlet con parámetro $(\alpha(B_1), \ldots, \alpha(B_k))$.

Entre los atractivos del proceso Dirichlet, principalmente cuando se dio a conocer, está el hecho de que es modelo «conjugado» —esto significa que la distribución inicial y posterior pertenecen a la misma familia de distribuciones; para los modelos paramétricos, está la familia exponencial que tiene esta propiedad, por ejemplo—. Además, a partir de él, se han encontrado muchas maneras de caracterizarlo, por mencionar sólo algunas, están los procesos con incrementos independientes normalizados, el proceso Poisson–Dirichlet de dos parámetros y las medidas de probabilidad aleatorias inducidas mediante la representación stick-breaking; algunos de estas construcciones se estudiarán en el siguiente capítulo.

Otra característica de este proceso —y de las otras construcciones mencionadas— es que es una distribución discreta casi seguramente (c.s.); véase, [4]. Por un lado, esto y el siguiente resultado sentaron las bases para poder implementar computacionalmente modelos de inferencia, y por el otro, ha permitido que estos procesos sean utilizados en el análisis de conglomerados y en modelos de mezclas —temas que abordaremos en el último capítulo—.

Blackwell y McQueen en [5] caracterizan el proceso Dirichlet a través de un modelo de urnas, la conocida «urna de Pólya». Resumimos el resultado en la siguiente proposición.

**Proposición 1.8.** *Una sucesión de variables aleatorias $X_1, X_2, \ldots$ con valores en $\mathbb{X}$ es una* sucesión de Pólya con parámetro $\alpha$ *si*

$$\mathbb{P}(X_1 \in \cdot\,) = \frac{\alpha(\cdot)}{\alpha(\mathbb{X})} \quad y \quad \mathbb{P}(X_{i+1} \in \cdot \mid X_1, \ldots, X_i) = \frac{\alpha_i(\cdot)}{\alpha_i(\mathbb{X})}, \quad i \geq 1,$$

*donde* $\alpha_i(\cdot) = \alpha(\cdot) + \sum_{k=1}^{i} \delta_{X_k}(\cdot)$.

*Sea $X_1, X_2, \ldots$ una sucesión de Pólya con parámetro $\alpha$, entonces*

1. $\alpha_i(\cdot)/\alpha_i(\mathbb{X})$ *converge c.s., cuando $n \to \infty$, a una medida de probabilidad aleatoria discreta $\alpha^*$;*

2. $\alpha^*$ *corresponde al proceso Dirichlet con parámetro $\alpha$;*

3. *dado $\alpha^*$, las variables aleatorias $X_1, X_2, \ldots$ son independientes con distribución $\alpha^*$.*

En el siguiente capítulo continuaremos con el estudio de las distribuciones de probabilidad aleatorias, también conocidas como «medidas de probabilidad aleatorias» y de otros objetos aleatorios que son la base teórica para el análisis de conglomerados —las denominadas «particiones aleatorias».

# Capítulo 2

# Medidas de probabilidad aleatorias y particiones aleatorias

El proceso Dirichlet presentado en el capítulo anterior es un ejemplo de medida de probabilidad aleatoria. Como se dijo, a partir de la presentación de este proceso, han existido diversas generalizaciones e investigaciones en esta línea, las cuales —entre otras cosas— permitieron la expansión y utilización de los modelos de inferencia bayesianos no paramétricos. En este capítulo, comenzamos estudiando dos generalizaciones del proceso Dirichlet: los procesos con incrementos independientes normalizados y la representación stick-breaking. Posteriormente, veremos cómo se deriva un objeto aleatorio que constituye el fundamento teórico para el análisis de conglomerados —las denominadas «particiones aleatorias»—. Exploraremos en la segunda parte del capítulo cómo obtener distribuciones de probabilidad para particiones aleatorias basándonos en las medidas de probabilidad aleatorias de la primera parte.

*Definición* 2.1. Sean $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad, $(\mathbb{X}, \mathscr{X})$ un espacio completo y separable y $(\mathbb{M}, \mathscr{M})$ un espacio de medidas de probabilidad sobre $\mathbb{X}$, equipado con la topología de convergencia débil. Se dice que $P : \mathbb{M} \times \Omega \to \mathbb{X}$ es una *medida de probabilidad aleatoria* si

1. $P(m, \cdot)$ es una variable aleatoria para cada $m \in \mathbb{M}$;
2. $P(\cdot, \omega)$ es una medida de probabilidad para cada $\omega \in \Omega$.

Además, por convención, se dirá que una sucesión $\{y_1, \ldots, y_n\}$ es una *muestra de* $P$ si, dada $P = p$, las variables $y_i$ son condicionalmente independientes y con distribución común $p$.

## 2.1 Procesos con incrementos independientes normalizados

Una manera de construir distribuciones aleatorias es a través de la normalización de procesos estocásticos particulares —procesos con incrementos independientes— y fue presentada en [43]. Algo interesante de este enfoque es que —de acuerdo con los autores— representa una aproximación natural al problema de definir una medida de probabilidad aleatoria, ya que la idea es ir de distribuciones deterministas a aleatorias —definidas en $\mathbb{R}$— considerando sus incrementos en intervalos disjuntos como variables aleatorias independientes.

Comenzamos presentando brevemente una clase particular de procesos con incrementos independientes, los conocidos «procesos de Lévy crecientes»; mayores detalles se pueden encontrar en, por ejemplo, [44].

*Definición* 2.2 (Procesos de Lévy crecientes). Un proceso estocástico $\{\xi_t\}_{t\geq 0}$ —definido en un espacio de probabilidad $(\Omega, \mathscr{F}, \mathbb{P})$— es un *proceso de Lévy creciente* si

1. $\xi_0 = 0$ c.s.;
2. existe $\Omega_0 \in \mathscr{F}$ con $\mathbb{P}(\Omega_0) = 1$, tal que, para todo $\omega \in \Omega_0$, $\xi_t(\omega)$ es continua por la derecha en $t \geq 0$ y tiene límites por la izquierda en $t > 0$;
3. para cualquier $n \geq 1$ y $0 \leq t_0 < t_1 < \cdots < t_n$, las variables aleatorias $\xi_{t_0}, \xi_{t_1} - \xi_{t_0}, \ldots, \xi_{t_n} - \xi_{t_{n-1}}$ son independientes;
4. la distribución de $\xi_{s+t} - \xi_s$ no depende de $s$, esto es $(\xi_{s+t} - \xi_s) \overset{\mathrm{d}}{=} \xi_t$ para toda $s, t \geq 0$;
5. $\lim_{h\to 0} \mathbb{P}(|\xi_{t+h} - \xi_t| \geq \varepsilon) = 0$ para toda $t \geq 0$ y toda $\varepsilon > 0$;
6. $\xi_t(\omega)$ es creciente c.s. como función de $t$.

Estos procesos tienen una descomposición única —la conocida descomposición de Lévy-Itó— definida por la terna $(G, d, \nu)$. El término $G$ se conoce como término gaussiano —que define la varianza del componente gaussiano continuo—, $d$ es la deriva —la responsable del desarrollo promedio del proceso—, y $\nu$ es una medida en $\mathbb{R}$ que satisface

$$\nu(\{0\}) = 0 \quad \text{y} \quad \int_{\mathbb{R}} \min(|x|^2, 1)\nu(\mathrm{d}x) < \infty,$$

—llamada medida de Lévy y exhibe la frecuencia y magnitud de los brincos del proceso—. Adicionalmente, su exponente característico también está determinado por esta terna, como lo indica el siguiente teorema.

**Teorema 2.3** (Representación de Lévy-Khintchine). *Sea $\{\xi_t\}_{t\geq 0}$ un proceso de Lévy en $\mathbb{R}$ con terna $(G, d, \nu)$. Entonces*

$$\mathbb{E}\left[e^{-iz\xi_t}\right] = e^{t\psi(z)}, \quad z \in \mathbb{R},$$

*donde $\psi$ —llamado exponente característico—, está dado por*

$$\psi(z) = idz - \frac{1}{2}Gz^2 + \int_{\mathbb{R}} \left(e^{izx} - 1 - izx\mathbf{1}(|x| \leq 1)\right)\nu(\mathrm{d}x),$$

*con $\mathbf{1}(A)$ la función indicadora del evento $A$.*

Nuestro interés está en procesos de Lévy con terna $(0, 0, t\nu)$, ya que son procesos crecientes de brincos puros. Para estos, su exponente característico está dado por

$$\psi(z) = -\int_{\mathbb{R}^+} (1 - e^{zx})\nu(\mathrm{d}x),$$

además, resulta mejor utilizar la transformada de Laplace, obteniendo

$$\mathbb{E}\left[e^{-z\xi_t}\right] = \exp\left(-t\int_{\mathbb{R}^+}(1-e^{-zx})\nu(\mathrm{d}x)\right).$$

Finalmente, la condición sobre su medida de Lévy queda dada por

$$\nu(\{0\}) = 0 \quad \text{y} \quad \int_{\mathbb{R}^+} \min(x,1)\nu(\mathrm{d}x) < \infty.$$

Para la normalización de procesos de Lévy, consideraremos aquellos definidos en $\mathbb{R}$ y con terna $(0,0,t\nu)$ tal que $\nu(\mathbb{R}^+) = \infty$; esta última condición permitirá que la normalización esté bien definida. Veamos a continuación dos ejemplos de procesos de Lévy de esta clase.

*Ejemplo* 2.4 (Proceso Gamma). Sea $\{\xi_t\}_{t\geq 0}$ un proceso de Lévy con brincos Gamma —esto es, para toda $t > 0$, $\xi_t$ se distribuye Gamma de parámetros $(t,\beta)$ con esperanza $t/\beta$—. Entonces, su transformada de Laplace es

$$\mathbb{E}\left[e^{z\xi_t}\right] = \left(\frac{\beta+z}{\beta}\right)^{-t} = \exp\left(-t\int_0^\infty (1-e^{-zx})x^{-1}e^{-\beta x}\mathrm{d}x\right).$$

En este caso, tenemos que

$$\nu(\mathrm{d}x) = x^{-1}e^{-\beta x}\mathrm{d}x,$$

y su exponente característico es

$$\psi(z) = \log\left(\frac{\beta+z}{\beta}\right).$$

Además, $\nu(0+) = \infty$ y por tanto $\nu(\mathbb{R}^+) = \infty$.

*Ejemplo* 2.5 (Proceso $\sigma$-estable). Sea $\{\xi_t\}_{t\geq 0}$ un proceso de Lévy con brincos $\sigma$-estable, para alguna $\sigma \in (0,1)$. Este proceso es tal que para cada $a > 0$, $\{\xi_{at}, t \geq 0\} \overset{\mathrm{d}}{=} \{a^{1/\sigma}\xi_t, t \geq 0\}$; [25]. En general, no existe una forma analítica para su función de densidad.

La medida de Lévy de una distribución $\sigma$-estable es

$$\nu(\mathrm{d}x) = cx^{-(1+\sigma)}\mathrm{d}x, \quad c > 0,$$

tiene transformada de Laplace

$$\mathbb{E}\left[e^{z\xi_t}\right] = \exp\left(-t\frac{c\Gamma(1-\sigma)z^\sigma}{\sigma}\right),$$

y su exponente característico es

$$\psi(z) = \frac{c\Gamma(1-\sigma)z^\sigma}{\sigma}.$$

Al igual que el proceso Gamma, se tiene que $\nu(0+) = \infty$ y por tanto $\nu(\mathbb{R}^+) = \infty$.

Una vez seleccionado el proceso de Lévy, la medida de probabilidad aleatoria se obtiene al normalizarlo, como se indica en la siguiente definición.

*Definición* 2.6. Sea $\{\xi_t\}_{t\geq 0}$ un proceso de Lévy con terna $(0,0,t\nu)$ cuya medida de Lévy $\nu$ sobre $\mathbb{R}^+$ tal que satisface $\nu(\mathbb{R}^+)=\infty$. Sea $\alpha$ una medida finita, no nula, sobre $\mathbb{R}$ con masa total $a:=\alpha(\mathbb{R})$ y sea $t=\alpha(-\infty,x]$. Se dice que $P$ es un *proceso con incrementos independientes normalizado* con parámetros $(\alpha,\nu)$, si

$$P(\cdot)=\frac{\xi_{\alpha(\cdot)}}{\xi_a}.$$

De la misma manera que con una variable aleatoria, es posible calcular esperanzas, varianzas o cualquier otro momento para una medida de probabilidad aleatoria. Para estos procesos, la siguiente proposición los enuncia; en [22] se tienen más detalles.

**Proposición** 2.7. *Si $P$ es un proceso con incrementos independientes normalizado de parámetros $(\alpha,\nu)$ y con masa total $a$. Entonces, para $A$ y $B$ conjuntos medibles, se tiene que*

$$\begin{aligned}
\mathbb{E}[P(B)] &= \frac{\alpha(B)}{a},\\
\mathrm{Var}[P(B)] &= \frac{\alpha(B)(a-\alpha(B))}{a^2}I_a,\\
\mathrm{Cov}[P(A),P(B)] &= \frac{a\alpha(A\cap B)-\alpha(A)\alpha(B)}{a^2}I_a,
\end{aligned}$$

*donde*

$$I_a = a\int_{\mathbb{R}^+} ue^{-a\psi(u)}\int_{\mathbb{R}^+} x^2e^{-ux}\nu(\mathrm{d}x)\mathrm{d}u.$$

Es importante notar que el valor esperado de estas medidas de probabilidad aleatorias sólo depende de $\alpha$. En lo que sigue, utilizaremos $P_0(\cdot):=\alpha(\cdot)/a$. Otra característica común con el proceso Dirichlet es que su distribución predictiva admite una representación a través del esquema de urnas de Pólya.

**Proposición** 2.8. *Sea $X_1,\ldots,X_n$ una muestra de $P$, con $P$ un proceso con incrementos independientes normalizado de parámetros $(\alpha,\nu)$. Entonces*

$$\begin{aligned}
\mathbb{P}[X_2\in B\mid X_1] &= (1-I_a)P_0(B)+I_a\delta_{X_1}(B),\\
\mathbb{P}[X_{n+1}\in B\mid X_1,\ldots,X_n] &= w_nP_0(B)+\frac{1}{n}\sum_{j=1}^{k}w_{nj}\delta_{X_j^*}(B),
\end{aligned}$$

*donde $I_a$ es como en la proposición anterior, $X_1^*,\ldots,X_k^*$ denotan los $k$ valores distintos de la muestra $X_1,\ldots,X_n$ y $w_n$ y $w_{nj}$ son pesos determinados por la medida de Lévy de $P$.*

Las expresiones para el cálculo de $w_n$ y $w_{nj}$ se pueden consultar en [22, Corolario 1].

*Ejemplo* 2.4, continuación (Proceso Gamma normalizado). Del proceso Gamma se tiene

$$\begin{aligned}
\mathbb{E}[P(B)] &= P_0(B),\\
\operatorname{Var}[P(B)] &= \frac{P_0(B)(1-P_0(B))}{(1+a)},\\
\operatorname{Cov}[P(B_1),P(B_2)] &= \frac{P_0(B_1\cap B_2)-P_0(B_1)P_0(B_2)}{(1+a)},\\
\mathbb{P}[X_2\in B\mid X_1] &= \frac{a}{(1+a)}P_0(B)+\frac{1}{(1+a)}\delta_{X_1}(B),\\
\mathbb{P}[X_{n+1}\in B\mid X_1,\ldots,X_n] &= \frac{a}{(n+a)}P_0(B)+\frac{1}{(n+a)}\sum_{j=1}^{k} n_j\delta_{X_j^*}(B).
\end{aligned}$$

Este proceso corresponde al proceso Dirichlet y es la definición alternativa dada por [10, Sección 4].

*Ejemplo* 2.5, continuación (Proceso $\sigma$–estable normalizado). Del proceso $\sigma$–estable se tiene

$$\begin{aligned}
\mathbb{E}[P(B)] &= P_0(B),\\
\operatorname{Var}[P(B)] &= P_0(B)(1-P_0(B))(1-\sigma),\\
\operatorname{Cov}[P(B_1),P(B_2)] &= (P_0(B_1\cap B_2)-P_0(B_1)P_0(B_2))(1-\sigma),\\
\mathbb{P}[X_2\in B\mid X_1] &= \sigma P_0(B)+(1-\sigma)\delta_{X_1}(B),\\
\mathbb{P}[X_{n+1}\in B\mid X_1,\ldots,X_n] &= \frac{\sigma k}{n}P_0(B)+\frac{1}{n}\sum_{j=1}^{k}(n_j-\sigma)\delta_{X_j^*}(B).
\end{aligned}$$

## 2.2 Representación stick-breaking

Como se dijo, una característica del proceso Dirichlet —que comparten las medidas de probabilidad aleatorias anteriores— es que sus realizaciones son discretas c.s. Esto significa, que si $P$ se distribuye de acuerdo a alguna de estas medidas, se puede escribir de la siguiente manera:

$$P(\cdot)=\sum_{j=1}^{\infty} W_j\delta_{Z_j}(\cdot), \tag{2.1}$$

donde la sucesión de pesos aleatorios $W_1,W_2,\ldots$ es tal que $W_j>0$, $j=1,2,\ldots$, y suman uno c.s. y es independiente de la sucesión de localizaciones aleatorias $Z_1,Z_2,\ldots$. La forma, o distribución, de los pesos no siempre se puede encontrar para un proceso con incrementos independientes normalizado, es por esta razón que a continuación estudiaremos distribuciones aleatorias que se pueden escribir como (2.1).

Una manera de definir los pesos aleatorios es utilizando un procedimiento denominado «asignación residual», que es mejor conocido por su término en inglés «stick-breaking».

*Definición* 2.9 (Representación stick-breaking). Una sucesión de variables aleatorias $W_1,W_2,\ldots$

se dice que tiene una representación «stick-breaking» si tienen la siguiente forma

$$W_1 = V_1, \quad \text{y} \quad W_j = V_j \prod_{l<j} (1 - V_l), \quad j > 2,$$

para una sucesión de variables aleatorias $V_1, V_2, \ldots$ con valores en $(0,1)$.

Un caso de gran interés es el proceso Poisson-Dirichlet de dos parámetros —introducido en [38]— que está basado en la distribución de probabilidad con el mismo nombre, la cual surge del estudio de distribuciones asintóticas de frecuencias relativas aleatorias.

*Definición* 2.10 (Proceso Poisson-Dirichlet de dos parámetros). Se dice que una medida de probabilidad aleatoria $P$ de la forma (2.1) tiene distribución de acuerdo con un *proceso Poisson-Dirichlet de dos parámetros* $(\sigma, \theta)$ si

1. la sucesión de localizaciones aleatorias $Z_1, Z_2, \ldots$ son independientes y con distribución común $P_0$ no atómica,
2. los pesos aleatorias $W_1, W_2, \ldots$ tienen representación stick-breaking tal que las variables $V_j$, $j = 1, 2, \ldots$, son independientes y con distribución Beta de parámetros $(1 - \sigma, \theta + j\sigma)$ para algunas $0 \leq \sigma < 1$ y $\theta > -\sigma$.

Este proceso es de interés porque contiene a los dos ejemplos de procesos con incrementos independientes normalizados presentados anteriormente. Si $\sigma = 0$ y $\theta = \alpha$, se obtiene el proceso Gamma normalizado —o proceso Dirichlet—, mientras que si $\theta = 0$, se obtiene el proceso $\sigma$-estable normalizado.

Por otro lado, este proceso también admite una representación de urna para su distribución predictiva; véase, por ejemplo, [39, Ejemplo 16], para más detalles.

**Proposición 2.11.** *Sea $X_1, \ldots, X_n$ una muestra del proceso Poisson-Dirichlet de dos parámetros $(\sigma, \theta)$ y supóngase que $X_1$ tiene distribución $P_0$ no atómica. Entonces, la distribución predictiva para este proceso es*

$$\mathbb{P}[X_i \in \cdot \mid X_1, \ldots, X_{i-1}] = \frac{\theta + \sigma k}{\theta + i - 1} P_0(\cdot) + \sum_{j=1}^{k} \frac{n_j^* - \sigma}{\theta + i - 1} \delta_{X_j^*}(\cdot), \quad i = 2, \ldots, n, \tag{2.2}$$

*donde $X_1^*, \ldots, X_k^*$ es el conjunto de los $k$ valores únicos (o distintos) de $X_1, \ldots, X_{i-1}$, cada uno con frecuencia $n_j^*$, $j = 1, \ldots, k$.*

*Ejemplo* 2.12. El proceso Poisson-Dirichlet de dos parámetros $(0, \alpha)$ corresponde al proceso Dirichlet con parámetro $\alpha$ (Ejemplo 2.4). En este caso, su distribución predictiva está dada por

$$\mathbb{P}[X_n \in \cdot \mid X_1, \ldots, X_{n-1}] = \frac{\theta}{\theta + n - 1} P_0(\cdot) + \frac{1}{\theta + n - 1} \sum_{j=1}^{k} \delta_{X_j^*}(\cdot).$$

*Ejemplo* 2.13. El proceso Poisson-Dirichlet de dos parámetros $(\sigma, 0)$ corresponde al proceso $\sigma$-estable normalizado (Ejemplo 2.5) y su distribución predictiva está dada por

$$\mathbb{P}[X_n \in \cdot \mid X_1, \ldots, X_{n-1}] = \frac{\sigma k}{n-1} P_0(\cdot) + \frac{1}{n-1} \sum_{j=1}^{k} (n_j^* - \sigma) \delta_{X_j^*}(\cdot),$$

## 2.3 Particiones aleatorias

La representación de urna de Pólya de las distribuciones aleatorias presentadas en las secciones anteriores es otra manera de ver su naturaleza discreta. Supongamos que secuencialmente observamos la muestra, iniciando con $X_1$, que será un valor aleatorio de acuerdo con $P_0$; la segunda observación puede ser igual a $X_1$ o un nuevo valor, de acuerdo con $P_0$; así sucesivamente.

Si nos concentramos en los valores de la muestra completa, $X_1, \ldots, X_n$, es claro que existirán valores repetidos. Equivalentemente, únicamente habrá $k \leq n$ valores diferentes. Al obtener otra realización del mismo tamaño, sucederá lo mismo, pero la configuración y cantidad de valores diferentes cambiará. Si posteriormente agrupamos la muestra de acuerdo con el valor de cada $X_i$, tenemos que todos los posibles agrupamientos están en biyección con una clase combinatoria conocida como el «conjunto de particiones» —el término *partición* tiene el mismo significado que en teoría de conjuntos—.

La importancia de esta clase combinatoria es que codifica todas las posibles maneras en que un conjunto de elementos puede ser agrupado. En otras palabras, es el soporte para los problemas de análisis de conglomerados que estudiaremos en el siguiente capítulo. Por tanto, tenemos en las distribuciones aleatorias una herramienta probabilística para definir modelos de inferencia bayesiana, no paramétrica, útiles para el análisis de conglomerados.

Veamos un ejemplo concreto del conjunto de particiones y cómo se relacionan con las realizaciones de una medida aleatoria discreta c.s.

*Ejemplo* 2.14. Supongamos que tenemos tres elementos $e_1$, $e_2$, y $e_3$. Entonces, todas las posibles formas de agruparlos son las siguientes:

$$\{\{e_1, e_2, e_3\}\}, \{\{e_1\}, \{e_2, e_3\}\}, \{\{e_1, e_2\}, \{e_3\}\}, \{\{e_1, e_3\}, \{e_2\}\}, \{\{e_1\}, \{e_2\}, \{e_3\}\}.$$

Así, por ejemplo, la partición $\{\{e_1, e_2, e_3\}\}$ nos indica que los tres elementos se agruparon juntos en un mismo «bloque», mientras que $\{\{e_1\}, \{e_2\}, \{e_3\}\}$ nos indica que cada elemento está en un bloque diferente.

Si tomamos una muestra de tamaño tres de una medida aleatoria $P$, una realización de $(X_1, X_2, X_3)$ es $(x, x, x)$ —esto es, se observó el mismo valor siempre— y corresponde a la partición $\{\{e_1, e_2, e_3\}\}$. Otra realización es $(x, y, z)$ —donde todos los valores son diferentes— siendo $\{\{e_1\}, \{e_2\}, \{e_3\}\}$ la partición asociada.

Denotemos por $\mathscr{P}_{[n]}$ al conjunto de todas las particiones para un conjunto de $n$ elementos. Es claro que no importa qué elementos se agrupen, sino cómo lo hacen; por esta razón, se acostumbra utilizar al conjunto $[n] := \{1, \ldots, n\}$ cuando se estudia el conjunto de particiones.

*Definición* 2.15 (Partición aleatoria). Una *partición aleatoria* $\Pi_n$ es una variable aleatoria con valores en $\mathscr{P}_{[n]}$.

A continuación examinaremos algunas formas de obtener la distribución de probabilidad de una partición aleatoria.

## 2.3.1 Distribuciones basadas en sucesiones de variables aleatorias intercambiables

Supongamos que una muestra $(X_1,\dots,X_n)$ de tamaño $n$ se obtiene de una medida de probabilidad $P$ como en (2.1) —esto es, las variables son condicionalmente independientes e idénticamente distribuidas con distribución $P$—. Al ser $P$ una distribución discreta, la probabilidad de que existan observaciones repetidas es positiva, i.e. $\mathbb{P}[X_i = X_j] > 0$ para $i \neq j$. Denotemos por $K_n$ el número de valores diferentes en la muestra, y sean tales valores $(X_1^*,\dots,X_{K_n}^*)$; notemos que $K_n$ es una variable aleatoria. Sea $(N_1,\dots,N_{K_n})$ el vector de frecuencias de cada $X_i^*$. Tenemos, entonces, que $K_n \leq n$ y $\sum_k N_k = n$ c.s. Debido a que la muestra es una sucesión de variables aleatorias intercambiables, la distribución conjunta del vector $(K_n, N_1,\dots,N_{K_n})$ se puede calcular como

$$\mathbb{P}(K_n = k, N_1 = n_1,\dots,N_{K_n} = n_k) = \int_{\mathbb{X}^k} \mathbb{E}_Q\left(P^{n_1}(\mathrm{d}x_1)\cdots P^{n_k}(\mathrm{d}x_k)\right), \tag{2.3}$$

esto es, estamos calculando la probabilidad de observar $k$ valores distintos en una muestra de tamaño $n$ de $P$, donde cada uno de ellos aparece exactamente $n_j$ veces para $j = 1,\dots,k$.

Observemos que los valores diferentes en la muestra inducen una partición, $\{C_1,\dots,C_k\}$ del conjunto $[n]$, donde cada bloque es tal que $C_j = \{i : X_i = X_j^*\}$ para $j = 1,\dots,k$. Por tanto, existe una correspondencia uno a uno entre cada elemento de $\mathscr{P}_{[n]}$ y cada posible realización del vector aleatorio $(K_n, N_1,\dots,N_{K_n})$, y podemos definir una distribución de probabilidad para alguna partición aleatoria $\Pi_n$ —una vez que seleccionemos la medida $Q$—.

## 2.3.2 Distribuciones de probabilidad sobre particiones intercambiables

La distribución de probabilidad definida en la Ecuación (2.3) posee dos características especiales. La primera de ellas se conoce como «consistencia», que básicamente nos dice que la distribución no cambia conforme el tamaño de la muestra incrementa. Esta propiedad fue introducida por Kingman [26, 27] y en el caso particular de (2.3), [1] demuestra que esta clase de distribuciones es consistente.

La segunda característica de estas distribuciones se le conoce como «intercambiabilidad». No debe confundirse la intercambiabilidad de la Sección (1.2) —de hecho, un mejor nombre para la propiedad que estudiaremos a continuación es «simetría», pero la literatura adoptó el primer nombre—. Este término fue introducido por Kingman [28] y Aldous [1].

*Observación* 2.16. Para cualquier entero positivo $n$, una *composición del entero n* es una sucesión de enteros positivos $(n_1,\dots,n_k)$ tales que $n_1 + \cdots + n_k = n$. Sea

$$\Delta_{n,k} := \left\{ (n_1,\dots,n_k) : n_j \geq 1, \sum_{j=1}^{k} n_j = n \right\} \tag{2.4}$$

el conjunto de todas las composiciones del entero $n$ con exactamente $k$ términos.

*Definición* 2.17. Una partición aleatoria $\Pi_n$ es *intercambiable* si, para cualquier partición $\pi = \{\pi_1, \ldots, \pi_k\} \in \mathscr{P}_{[n]}$

$$\mathbb{P}(\Pi_n = \pi) = \mathbb{P}(\Pi_n = \rho(\pi)), \tag{2.5}$$

para toda permutación $\rho$ en el grupo simétrico actuando sobre el conjunto $[n]$ Equivalentemente,

$$\mathbb{P}(\Pi_n = \{\pi_1, \ldots, \pi_k\}) = \Pi_k^{(n)}(n_1, \ldots, n_k), \tag{2.6}$$

para alguna función simétrica $\Pi_k^{(n)} : \Delta_{n.k} \to [0,1]$, con $n_j := \#\pi_j$, $j = 1, \ldots, k$, y donde $\#\pi_j$ denota el número de elementos en $\pi_j$. Adicionalmente, la función $\Pi_k^{(n)}$ se llama *función de probabilidad sobre particiones intercambiables* (FPPI).

Las FPPIs constituyen una herramienta muy importante en el estudio de las particiones aleatorias. De hecho, mucha de la literatura sobre el tema se centra en particiones aleatorias intercambiables. También notemos que todas las particiones aleatorias inducidas por el Teorema de representación de de Finetti, Ecuación (2.3), son intercambiables, ya que es claro que esta ecuación es una función simétrica con respecto a $(n_1, \ldots, n_k)$. Por tanto, podemos definir una FPPI haciendo

$$\Pi_k^{(n)}(n_1, \ldots, n_k) := \mathbb{P}(K_n = k, N_1 = n_1, \ldots, N_{K_n} = n_k).$$

*Observación* 2.18. La propiedad de intercambiabilidad en las FPPIs es muy útil en los análisis bayesianos al utilizarla como distribución inicial, ya que (i) la probabilidad de cualquier partición no depende de los valores de la muestra —es decír, no depende de *qué* vamos a agrupar— sino sólo del tamaño de cada bloque, y (ii) la simetría no favorece ninguna partición particular entre aquellas que tengan el mismo número de bloques.

También es importante mencionar que —al trabajar con particiones aleatorias intercambiables— la propiedad de consistencia es equivalente a la denominada «regla de la adición». En otras palabras, una FPPI es consistente si satisface

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \Pi_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} \Pi_k^{(n+1)}(n_1, \ldots, n_j + 1, \ldots, n_k).$$

Notemos que no toda partición intercambiable es consistente.

### Familias de funciones de probabilidad sobre particiones intercambiables

A continuación presentaremos dos casos específicos de FPPIs. La primera familia se obtiene de los procesos con incrementos independientes normalizados. La segunda es una construcción diferente y tiene relación con el proceso Poisson-Dirichlet de dos parámetros.

**Procesos con incrementos independientes normalizados** Retomando los procesos con incrementos independientes normalizados de la Sección 2.1, podemos definir una FPPI como sigue

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{1}{\Gamma(n)} \int_0^\infty u^{n-1} e^{-\psi(u)} \prod_{j=1}^{k} \int_{\mathbb{X}} \tau_{n_j}(u) \alpha(\mathrm{d}x) \mathrm{d}u, \tag{2.7}$$

para $\alpha$ una medida positiva, finita y no atómica, $\psi(u)$ es el exponente característico del proceso, y $\tau_m(u)$ se define, para cualquier $m \geq 1$, como

$$\tau_m(u) = \int_{\mathbb{R}^+} s^m e^{-us} \nu(\mathrm{d}s),$$

con $\nu$ la medida de Lévy correspondiente.

*Ejemplo* 2.4, continuación. La FPPI asociada al proceso Gamma normalizado —o proceso Dirichlet— está dada por

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{\theta^k}{(\theta)_{n\uparrow}} \prod_{j=1}^{k} \Gamma(n_j),$$

donde $(x)_{n\uparrow}$ denota el símbolo de Pochhammer —definido por $(x)_{n\uparrow} := \prod_{j=0}^{n-1}(x+j)$ con $(x)_{0\uparrow} = 1$—. Esta expresión coincide con la presentada en [2].

*Ejemplo* 2.5, continuación. Para el proceso $\sigma$-estable normalizado, tenemos que su FPPI asociada es

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{\sigma^{k-1}\Gamma(k)}{\Gamma(n)} \prod_{j=1}^{k} (1-\sigma)_{n_j-1\uparrow},$$

para $0 < \sigma < 1$.

**Particiones aleatorias tipo Gibbs** Pitman [40] proporciona una construcción general para distribuciones de probabilidad aleatorias discretas c.s. Como caso particular existen las denominadas «iniciales tipo Gibbs» —introducidas por [15]—. Nosotros únicamente presentamos la forma de la FPPI para estas distribuciones.

*Definición* 2.19. Sea $P$ una medida de probabilidad aleatoria como en (2.1). Entonces, decimos que $P$ es una *medida aleatoria tipo Gibbs* si, para toda $1 \leq k \leq n$ y cualquier composición del entero $n$, $(n_1, \ldots, n_k)$, su FPPI se puede escribir como

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j-1\uparrow},$$

para algún $0 \leq \sigma < 1$.

Los pesos $V_{n,k}$ deben satisfacer la siguiente ecuación recursiva

$$V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1},$$

para cualquier $n \geq 1$ y $1 \leq k \leq n$, con $V_{1,1} = 1$.

*Ejemplo* 2.20. El proceso Poisson-Dirichlet de dos parámetros $(\sigma, \theta)$ es una inicial tipo Gibbs cuyos pesos $V_{n,k}$ están dados por

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1\uparrow}}, \tag{2.8}$$

para algunas $0 \leq \sigma < 1$ y $\theta > -\sigma$, obteniendo así la FPPI

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k}(1-\sigma)_{n_j-1\uparrow}.$$

**Distribuciones de producto de particiones** Las «distribuciones de producto de particiones» —o «modelos de producto de particiones»— fueron propuestas por Hartigan [17] para tratar problemas de clasificación bajo un enfoque bayesiano paramétrico. Sin embargo, resultan de interés porque comprende algunos de los enfoques ya estudiados para las FPPIs; véase [42] y [31] para un estudio más a fondo.

*Definición* 2.21. Sea $\Pi_n$ una partición aleatoria. Una *distribución de producto de particiones* es una distribución de probabilidad para $\Pi_n$ dada por

$$\mathbb{P}(\Pi_n = \{\pi_1,\ldots,\pi_k\}) = M \prod_{j=1}^{k} c(\pi_j),$$

para cualquier $\{\pi_1,\ldots,\pi_k\} \in \mathscr{P}_{[n]}$, y donde $c : [n] \to \mathbb{R}^+ \cup \{0\}$ es una *función cohesión* y $M$ es la constante de normalización, i.e.

$$M^{-1} = \sum_{\pi \in \mathscr{P}_{[n]}} \prod_{j=1}^{\#\pi} c(\pi_j),$$

con $\#\pi$ el número de bloques en la partición $\pi$.

Una distribución de producto de particiones será intercambiable o consistente de acuerdo con la función cohesion seleccionada. Por ejemplo, si $c(\pi_j) = \theta\Gamma(\#\pi_j)$, para alguna $\theta > 0$, obtenemos la FPPI inducida por el proceso Dirichlet.

### 2.3.3 Distribución de probabilidad sobre el número de bloques

Como he comentó, al trabajar con particiones aleatorias, se define directamente una variable aleatoria —que denotamos por $K_n$— la cual modela el número de bloques en la partición. Esta representa una herramienta muy valiosa tanto a nivel teórico como aplicado. Para las FPPIs presentadas anteriormente, daremos la distribución de esta variable; en el siguiente capítulo profundizaremos en su aplicación.

Para iniciales tipo Gibbs, la distribución de $K_n$ se puede escribir como

$$\mathbb{P}(K_n = k) = \frac{V_{n,k}}{\sigma^k} G(n,k,\sigma), \quad k = 1,\ldots,n,$$

donde $G(n,k,\sigma)$ denota el *número de Stirling generalizado* —o *coeficiente factorial generalizado*— definido como

$$G(n,k,\sigma) = \frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}(-1)^j(-j\sigma)_{n\uparrow}.$$

*Ejemplo* 2.4, continuación. La distribución de $K_n$ para el proceso Dirichlet de parámetro $\theta$ está dada por

$$\mathbb{P}(K_n = k) = \frac{c(n,k)\theta^k}{(\theta)_{n\uparrow}}, \quad k = 1,\ldots,n,$$

donde $c(n,k)$, $k = 1,\ldots,n$, denota el *número de Stirling de primera clase sin signo*. Su valor esperado es

$$\mathbb{E}(K_n) = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}.$$

Además, [30] obtuvieron que

$$\frac{K_n}{\log n} \to \theta, \quad n \to \infty.$$

*Ejemplo* 2.5, continuación. Para el proceso $\sigma$-estable normalizado, tenemos

$$\mathbb{P}(K_n = k) = \frac{\Gamma(k)}{\sigma\Gamma(n)} G(n,k,\sigma), \quad k = 1,\ldots,n,$$

y

$$\mathbb{E}(K_n) = \frac{(1+\sigma)_{n-1\uparrow}}{\Gamma(n)}.$$

*Ejemplo* 2.20, continuación. Para el proceso Poisson-Dirichlet de dos parámetros, la distribución de $K_n$ es

$$\mathbb{P}(K_n = k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma^k(\theta+1)_{n-1}} G(n,k,\sigma) \quad k = 1,\ldots,n.$$

Además,

$$\mathbb{E}(K_n) = \frac{(\theta+\sigma)_{n\uparrow}}{\sigma(\theta+1)_{n-1\uparrow}} - \frac{\theta}{\sigma},$$

Con relación a su crecimiento asintótico, [41] demuestra que

$$\frac{K_n}{n^\sigma} \to Y_{\theta/\sigma} \text{ a.s.}, \quad n \to \infty,$$

donde $Y_q$ tiene función de densidad dada por

$$f(y) = \frac{\Gamma(q\sigma+1)}{\sigma\Gamma(q+1)} y^{q-1-1/\sigma} f_\sigma(y^{-1/\sigma}),$$

para $q \geq 0$ y donde $f_\sigma(\cdot)$ es la función de densidad de una variable aleatoria positva con distribución estable de parámetro $\sigma$.

# Capítulo 3

# Análisis de conglomerados y modelos de mezclas

Uno de los propósitos principales de este documento es dar los elementos para el análisis de datos bajo un enfoque bayesiano no paramétrico. Adicionalmente —como se ha podido observar— en la teoría presentada, surge de manera directa el problema de agrupación y se le da un tratamiento probabilístico. Es por esto que en este capítulo abordaremos un problema muy importante en el análisis de datos —que tiene diferentes nombres, de acuerdo con la disciplina en donde se esté—; para los estadísticos, se conoce como «análisis de conglomerados». Explicaremos dos enfoques que se utilizan para trabajar este problema; el más conocido se basa en «modelos de mezclas», mientras que el segundo es una aplicación directa de la inferencia bayesiana donde el parámetro de interés es una partición aleatoria.

## 3.1 Modelos de mezclas

Los modelos de mezclas (de densidades de probabilidad) son de gran interés en diferentes contextos debido a su capacidad para modelar la heterogeneidad. De manera simple, si el histograma de los datos muestra diversas modas, entonces, un modelo de mezclas es un candidato para definir su distribución de probabilidad. Formalmente, decimos que una muestra aleatoria $Y_1, \ldots, Y_n$ se distribuye de acuerdo con un modelo de mezclas si su distribución de probabilidad admite la siguiente forma para su función de densidad

$$f(y) = \sum_{j=1}^{M} w_j g_j(y \mid x_j), \tag{3.1}$$

donde $w_1, \ldots, w_M$ son pesos no negativos que suman uno, y $g_j$ es una función *kernel* —e.g. una función de densidad— para cada $j$, y con parámetro $x_j$. Es frecuente asumir un número finito de componentes, $M < \infty$, aunque veremos más adelante que existen modelos con un número infinito. Además, se suele utilizar una misma función kernel para todas las componentes y únicamente se deja libre su parámetro, esto es $g_j(\cdot \mid x_j) = g(\cdot \mid x_j)$ con $x_j$ un parámetro de dimensión finita.

Uno de los primeros estudios sobre modelos de mezclas fue presentado por Pearson [37], y desde entonces se ha creado muchísima literatura al respecto. Algunas referencias con un tratamiento estadístico del tema son [46], [35], [34] y [13] sólo por mencionar algunas; además, estos modelos

también se encuentran en libros sobre aprendizaje automático, un disciplina dentro de las Ciencias de la Computación. Sin embargo, una de las interpretaciones de estos modelos —la cual muestra su atractivo en la práctica— fue dada por Feller [9].

Asumamos que una población está conformada por $M$ categorías —también llamadas «grupos» o «subpoblaciones»— cada una con proporción relativa $w_j$, para $j = 1, \ldots, M$. Por «categoría» se entiende que la característica de interés —representada por el parámetro $x_j$— es homogénea dentro del grupo pero heterogénea entre grupos diferentes. De esta manera, en la práctica, uno de los principales intereses es describir las subpoblaciones (latentes) en términos de su proporción relativa, dada por $w_j$, y su característica de interés, $x_j$, que las define.

A pesar de lo sencilla que resulta esta interpretación, en la práctica, ha sido muy difícil de lograr. Pero por el momento, no abordaremos estos problemas.

Existen muchas metodologías que hacen estimación de los parámetros de un modelo de mezclas con un número finito y fijo de componentes. Se puede revisar la literatura ya mencionada. Nosotros nos enfocaremos en el caso donde, potencialmente, se tiene un número infinito de ellas.

El primer trabajo que estudia modelos «infinitos» de mezclas bajo un enfoque bayesiano no paramétrico fue dado por Lo [32], quien utiliza el proceso Dirichlet para definir los pesos de la mezcla y los parámetros del kernel. Sin embargo, como ya se mencionó, tuvieron que pasar varios años antes de poder implementar para casos reales este enfoque. A partir del trabajo de Escobar y West [8], la popularidad de los modelos infinitos de mezclas creció.

El modelo de Lo se puede generalizar a cualquier medida de probabilidad aleatoria discreta c.s. Recordando de la Sección 2.2, estas medidas de probabilidad se pueden escribir como

$$P(\cdot) = \sum_{j=1}^{\infty} w_j \delta_{x_j}(\cdot), \tag{3.2}$$

donde $w_j$ son variables aleatorias con valores en $(0, 1)$, para toda $j \geq 1$, y son tales que suman uno c.s., las variables $x_j$ son variables aleatorias independientes y con la misma distribución que toman valores en un espacio $\mathbb{X}$, y ambas sucesiones son independientes entre sí. La distribución de las variables $x_j$, denotada por $P_0$, debe ser no atómica.

Cuando se utiliza un modelo de mezclas, se dice que se sigue un enfoque «basado en modelos» [11]. Debido a que cada función kernel, $g$, es una función de densidad, las observaciones que sean asociadas a una componente particular se modelaran probabilísticamente de acuerdo con esa distribución, digamos $G$. Si se elige una distribución $G$ diferente, es probable que sean otras observaciones las que se asocien a la misma componente. Por el momento, sólo asumamos que se ha elegido una función de densidad $g$ que será la función kernel del modelo de mezclas.

Si la función kernel $g$ tiene parámetro $x$, el cual se *mezcla* utilizando una distribución aleatoria como (3.2), obtenemos lo siguiente

$$f(y) = \int_{\mathbb{X}} g(y \mid x) P(\mathrm{d}x) = \sum_{j=1}^{\infty} w_j g(y \mid x_j). \tag{3.3}$$

Notemos que se obtiene una función de densidad como en (3.1), pero con un número infinito de componentes. En particular, si la distribución de $P$ es el proceso Dirichlet, recuperamos el modelo propuesto por Lo, el cual se conoce por el nombre de «modelo de mezclas del proceso Dirichlet».

Por tanto, tenemos en (3.3) una clase muy amplia de modelos «no paramétricos» de mezclas, sólo necesitamos cambiar la distribución aleatoria $P$.

# 3.2

## Análisis bayesiano de modelos de mezclas

Para entender mejor cómo hacer inferencias utilizando un modelo de mezclas como (3.3), lo escribimos de manera esquemática como sigue

$$\begin{aligned} Y_i \mid X_i &\sim g(Y_i \mid X_i), \ [\text{ind}] \quad i = 1, \ldots, n, \\ X_i \mid P &\sim P, \ [\text{iid}] \\ P &\sim Q. \end{aligned} \tag{3.4}$$

Contamos con una muestra $Y_1, \ldots, Y_n$ donde cada una, $Y_i$, se modela de acuerdo con una distribución $g$ de parámetro $X_i$; a esto nos referimos con un enfoque basado en modelos. La segunda y tercera líneas en el esquema es un modelo bayesiano no paramétrico como en el primer capítulo. Notemos que —al ser $P$ una distribución discreta y por lo que ya se explicó en el capítulo anterior— los valores de los parámetros $X_1, \ldots, X_n$ pueden estar repetidos, los cual ocasiona que distintas observaciones sean modeladas por la misma distribución, digamos $g(\cdot \mid X_j^*)$.

Si la distribución $P$ admite una representación de urna de Pólya, será posible integrar la distribución $P$, y se obtiene el siguiente modelo

$$\begin{aligned} Y_i \mid X_i &\sim g(Y_i \mid X_i), \ [\text{iid}] \quad i = 1, \ldots, n \\ X_1, \ldots, X_n &\sim \nu(X_1, \ldots, X_n), \end{aligned}$$

con $\nu$ una distribución conjunta —precisamente la obtenida por el esquema de Urnas asociada a la distribución aleatoria—. Este fue el modelo presentado por Escobar y West [8] utilizando el proceso Dirichlet.

La forma estándar de hacer inferencia bayesiana es a través de métodos numéricos, debido a la complejidad de las distribuciones posteriores resultantes. Los métodos más utilizados se conocen como métodos Monte Carlo vía cadenas de Markov. Un caso particular es el denominado muestreador de Gibbs.

Para obtener muestras de la posterior correspondiente, notemos que los parámetros son intercambiables, por lo que se pueden obtener las distribuciones condicionales completas para cada parámetro $X_i$, que tienen la misma forma para cualquiera, y es

$$p(X_i \mid X_{-i}, Y) = q_{i,0}^* p(X_i \mid Y_i) + \sum_{j=1}^{k_{i,n-1}} q_{i,j}^* \delta_{X_{i,j}^*}(X_i), \quad i = 1, \ldots, n,$$

con $X_{-i} := \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, x_n\}$, y donde $X_{-i}^* = \{X_{i,1}^*, \ldots, X_{i,k_{i,n-1}}^*\}$ son tales que $k_{i,n-1}$ valores distintos del vector $X_{-i}$; aquí, el subíndice negativo indica que al vector completo se le elimina la $i$-ésima.

Los pesos $q_{i,0}, \ldots, q_{i,k_{i,n-1}}$ quedan determinados por la distribución $P$ seleccionada. Por ejemplo, para el proceso Poisson-Dirichlet de dos parámetros $(\sigma, \theta)$, tenemos

$$\begin{aligned} q_{i,0} &\propto (\theta + k_{i,n-1}\sigma) \int_{\mathbb{X}} p(Y_i|x) P_0(\mathrm{d}x), \\ q_{i,j}^* &\propto (n_{i,j}^* - \sigma) p(Y_i|X_{i,j}^*), \quad j \geq 1, \end{aligned}$$

donde $P_0$ es la medida base de $P$, y $(n^*_{i,1}, \ldots, n^*_{i,k_{i,n-1}})$ son las frecuencias de los parámetros $X^*_{-i}$.

Una mejora para este método fue propuesta por Mac Eachern [33]. Debido a que la probabilidad de generar un nuevo valor para los parámetros $X^*_j$ disminuye considerablemente en el método original, podían pasar muchas iteraciones antes de obtener un *mejor valor*. La modificación propuesta consiste en introducir una serie de variables indicadoras $D_i$, $i = 1, \ldots, n$, las cuales identifican a qué componente de la mezcla la $i$-ésima observación está asociada —esto es, $D_i = j$ si y sólo si $X_i = X^*_j$, donde $\{X^*_1, \ldots, X^*_{k^*}\}$ son los $k^*$ valores diferentes en $\{X_1, \ldots, X_n\}$—. Así, el paso extra consiste en actualizar cada $X^*_j$, $j = 1, \ldots, k^*$, utilizando su distribución posterior

$$p(X^*_j \mid \cdots) \propto P_0(X^*_j) \prod_{D_i = j} g(Y_i \mid X^*_j), \quad j = 1, \ldots, k^*. \tag{3.5}$$

### 3.2.1 Estimación de la densidad

Una vez que se obtienen muestras de la distribución posterior del modelo en (3.4), podemos hacer diferentes *tipos* de estimaciones. La primera de ellas es acerca de la distribución de probabilidad de las observaciones, la cual es una mezcla de densidades. El muestreador de Gibbs es una clase particular de método Monte Carlo, por lo que sus estimadores se basan en promedios de ciertas cantidades que se simulan por un número determinado de iteraciones.

Supongamos que el método de muestreo iteró $T$ veces. Durante cada una de ellas, se obtuvo una muestra de $k^*$ valores de los parámetros $\{X^*_1, \ldots, X^*_{k^*}\}$ y se tiene también el número de observaciones $Y_i$ que se asocian a cada uno de ellos, que son las frecuencias $(n^*_1, \ldots, n^*_{k^*})$. Entonces, en cada iteración $t$, se puede construir una función de densidad —que es una mezcla— dada por

$$f^{(t)}(y) = w^{(t)}_1 g(y \mid X^{*(t)}_1) + \cdots + w^{(t)}_{k^{*(t)}} g(y \mid X^{*(t)}_{k^{*(t)}}),$$

donde las variables $w_1, \ldots, w_{k^*}$ son los pesos de cada componente y están dados por $w_j = n_j/n$, con $n$ el tamaño de la muestra. Hicimos explícito el número de la iteración con el superíndice $(t)$. Por tanto, el «estimador Monte Carlo» de la densidad está dado por

$$\tilde{f}(y) = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{k^{*(t)}} w^{(t)}_j g(y \mid X^{*(t)}_j),$$

Existen otros métodos de muestro para modelos de mezclas bajo esta perspectiva bayesiana no paramétrica que fueron desarrollados posteriormente; véase [21, 24, 36, 47] por mencionar algunos.

## 3.3 Análisis de conglomerados

Otra estimación de interés que se puede obtener del modelo anterior es la estructura de agrupamiento de la muestra $Y_1, \ldots, Y_n$. Debido a que los parámetros $X_1, \ldots, X_n$ son una muestra de una distribución aleatoria discreta c.s., $P$, se está induciendo una partición de estos de acuerdo con sus valores diferentes $X^*_j$, como ya se ha estudiado. Sin embargo, esta partición también

induce automáticamente un agrupamiento al nivel de las observaciones; denotemos por $\pi$ a este agrupamiento.

Este estimador $\pi$ es justamente el propósito en cualquier método de análisis de conglomerados, ya que particiona a la muestra de acuerdo a algún criterio —en nuestro caso, es que se modelan por la misma distribución de probabilidad—. Utilizando el método de urnas de Pólya, el agrupamiento de la muestra $\pi$ se obtiene haciendo

$$\pi_j = \{Y_i : X_i = X_j^*\}, \quad j = 1, \ldots, k^*,$$

donde $\pi = \{\pi_1, \ldots, \pi_{k^*}\}$.

Cabe resaltar que el modelo de urnas de Pólya aplicado al análisis de conglomerados proporciona información adicional valiosa. Por un lado —y es una característica muy importante— estos modelos no paramétricos infieren también el número de grupos en la muestra, que anteriormente fue definido por la variable aleatoria $K_n$ y en este capítulo se denotó por $k^*$ por simplicidad. Prácticamente cualquier otro método para el análisis de conglomerados requiere que se fije el número de grupos que se desean *detectar*, lo cual es una desventaja, principalmente cuando se da un valor muy pequeño. Si este parámetro se modela con una variable aleatoria, el método de simulación se puede complicar grandemente; para métodos bayesianos, esto por lo general requiere utilizar algún método Monte Carlo vía cadenas de Markov «transdimensional».

Por otro lado, el método de muestreo contiene también los parámetros $X_j^*$ que caracterizan a cada subpoblación. Así que, si condicionamos a una partición $\hat{\pi}$, podremos también dar una estimación de los parámetros de cada grupo $\hat{\pi}_j$.

### 3.3.1 Inferencia utilizando particiones aleatorias

Un enfoque alternativo para hacer inferencias sobre la estructura de agrupamiento, $\pi$, es utilizando directamente una partición aleatoria $\Pi$. Esquemáticamente, para una muestra de $n$ observaciones, $Y_1, \ldots, Y_n$, tenemos el modelo

$$\begin{aligned} Y_i \mid X_j^*, \Pi &\sim g(Y_i \mid X_j^*)\mathbf{1}(i \in \Pi_j), \ \text{[ind]} \quad i = 1, \ldots, n, \\ X_j^* \mid \Pi &\sim P_0, \ \text{[iid]} \quad j = 1, \ldots, k^*, \\ \Pi &\sim \mu_0, \end{aligned}$$

donde $\Pi_j$ es el $j$-ésimo bloque de la partición. Hacemos el mismo supuesto distribucional sobre las observaciones pertenecientes a un mismo grupo, esto es, que siguen cierta distribución $g$ de parámetro $X_j$. La distribución $P_0$ debe ser no atómica y $\mu_0$ es una distribución de probabilidad para particiones aleatorias. Esta última distribución puede ser una FPPI, aunque el esquema de simulación puede no ser tan fácil de obtener. Sin embargo, el esquema de urnas de Pólya de la sección anterior es prácticamente equivalente a este segundo modelo.

# Ejercicios

1. Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes. Demuestre que la sucesión es intercambiable.

2. Utilizando la representación stick-breaking para el proceso Dirichlet, realice lo siguiente.

   *a*) Sea $P_0$ una distribución normal estándar. Grafique muestras de la distribución aleatoria $P$ haciendo el parámetro de masa total $a = 0{,}1, 1, 10$. (Tendrá que truncar la medida a $m$ sumandos.)

   *b*) ¿Cómo influye el valor de $a$ en la distribución del número de bloques, $K_n$, que se induce?

3. Diseñe un algoritmo para obtener muestras de la distribución predictiva del proceso Poisson-Dirichlet de dos parámetros.

4. Implemente el algoritmo anterior y obtenga una estimación Monte Carlo de las probabilidades para la partición aleatoria inducida y del número de bloques.

5. Sea $\{\xi_t\}_{t\geq 0}$ un proceso de Lévy creciente con terna $(0, 0, \nu)$ donde

$$\nu(\mathrm{d}s) = \frac{s^{-(1+\sigma)}\exp(-\tau s)}{\Gamma(1-\sigma)}\mathrm{d}s,$$

   con $\Gamma$ la función gamma, y donde $\sigma \in (0,1)$ y $\tau \geq 0$. Calcule la FPPI inducida.

6. Para las siguientes FPPIs, indique si se cumple la regla de la adición.

   *a*) Sea $\theta > 0$,

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{\theta^k}{(\theta)_{n\uparrow}} \prod_{j=1}^{k} \Gamma(n_j). \tag{3.6}$$

   *b*) Sea $\Pi_k^{\prime(n)}$ una FPPI,

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{1}{k!}\binom{n}{n_1, \ldots, n_k}\Pi_k^{\prime(n)}(n_1, \ldots, n_k),$$

   con $n = n_1 + \cdots + n_k$.

7. Para la EPPI en (3.6), fije $n = 4$ y $\theta = 1$. ¿Cuál es la partición modal, $\tilde{\pi}$? ¿El número de bloques en $\tilde{\pi}$ coincide con la moda de $K_n$? Explique.

# Bibliografía


[1] D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, Lecture notes in mathematics. Springer-Verlag, 1985.

[2] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[3] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, Inc., 1994.

[4] D. Blackwell. Discreteness of Ferguson Selections. *The Annals of Statistics*, 1(2):356–358, 1973.

[5] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[6] W. M. Bolstad and J. M. Curran. *Introduction to Bayesian Statistics*. John Wiley & Sons, Ltd, tercera edition, 2016.

[7] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937.

[8] M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[9] W. Feller. On a General Class of "Contagious" Distributions. *Annals of Mathematical Statistics*, 14(4):389–400, 1943.

[10] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[11] C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[12] D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, 1963.

[13] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

[14] Ghosal, Subhashis and van der Vaart, Aad. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

[15] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. In *Representation Theory, Dynamical Systems, Combinatorial and Algorithmic Methods. Part 12*, volume 325 of *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, pages 83–102. PDMI, 2005.

[16] M. Goldstein. Observables and models: exchangeability and the inductive argument. In *Bayesian Theory and Applications*. Oxford University Press, 2013.

[17] J. A. Hartigan. Partition models. *Communications in Statistics - Theory and Methods*, 19:2745–2756, 1990.

[18] E. Hewitt and L. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501, 1955.

[19] N. L. Hjort, C. C. Holmes, and and, editors. *Bayesian Nonparametrics*. Cambridge University Press, 2010.

[20] P. D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[21] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[22] L. F. James, A. Lijoi, and Prünster. Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33(1):105–120, 2006.

[23] O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Probability and its Applications. Springer, New York, 2005.

[24] M. Kalli, J. E. Griffin, and S. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

[25] J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 37:1–22, 1975.

[26] J. F. C. Kingman. Random Partitions in Population Genetics. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 361(1704):1–20, 1978.

[27] J. F. C. Kingman. The Representation of Partition Structures. *Journal of the London Mathematical Society*, s2-18(2):374–380, 1978.

[28] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[29] A. D. Kiureghian and O. Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009.

[30] R. M. Korwar and M. Hollander. Contributions to the Theory of Dirichlet Processes. *The Annals of Probability*, 1(4):705–711, 1973.

[31] A. Lijoi, R. H. Mena, and I. Prünster. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):715–740, 2007.

[32] A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357, 1984.

[33] S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Theory and Methods*, 23(3):727–741, 1994.

[34] J.-M. Marin, K. L. Mengersen, and C. Robert. Bayesian modelling and inference on mixtures of distributions. In D. Dey and C. R. Rao, editors, *Handbook of Statistics: Volume 25*. Elsevier, 2005.

[35] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley-Interscience, 2000.

[36] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[37] K. Pearson. Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110, 1894.

[38] M. Perman, J. Pitman, and M. Yor. Size–biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92:21–39, 1992.

[39] J. Pitman. Some developments of the Blackwell–MacQueen urn scheme. In T. S. Ferguson, L. S. Shapley, and J. B. MacQueen, editors, *Statistics, Probability and Game Theory; papers in honor of David Blackwell*, volume 30 of *Lecture Notes–Monograph Series*, pages 245–267. Institute of Mathematical Statistics, 1996.

[40] J. Pitman. Poisson-Kingman partitions. In D. R. Goldstein, editor, *Statistics and science: a Festschrift for Terry Speed*, volume 40 of *Lecture Notes-Monograph Series*, pages 1–34. Institute of Mathematical Statistics, 2003.

[41] J. Pitman. *Combinatorial stochastic processes*. Ecole d'eté de probabilités de Saint-Flour XXXII - 2002. Springer, 2006.

[42] F. Quintana and P. L. Iglesias. Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574, 2003.

[43] E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585, 2003.

[44] K. Sato. *Lévy processes and infinitely divisible distributions*. Cambridge University Press, 1999.

[45] M. J. Schervish. *Theory of statistics*. Springer series in statistics. Springer, New York, Berlin, Heidelberg, 1995.

[46] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

[47] S. Walker. Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, 2007.