Memorias de la XVII Semana Regional de
Investigación y Docencia en Matemáticas.
Departamento de Matemáticas, Universidad de Sonora, México,
Mosaicos Matemáticos, No. 20, Agosto, 2007, pp. 171–176.

LA FUNCIÓN GENERADORA DE MOMENTOS EN LAS DISTRIBUCIONES DE MEZCLAS

Enrique Villa Diharce **Luis Escobar Restrepo**
Centro de Investigación en Matemáticas
Louisiana State University

Resumen

Los modelos de mezclas son importantes en la teoría y la aplicación de la estadística. Los cursos de estadística matemática en los ultimos semestres de la licenciatura y en el primer año de la maestría, deberían exponer a los estudiantes con algunas propiedades importantes de los modelos de mezclas. En los libros de texto comunes, el enfoque predominante para obtener las distribuciones de mezclas se basa en la marginalización de la distribución conjunta definida por el modelo de mezclas. Este artículo propone el uso de la función generatriz de momentos (FGM) para obtener la distribución de algunas mezclas. Se da un ejemplo para el caso de funciones características. El enfoque de la FGM tiende a ser simple y usa algunas de las herramientas previamente construidas en el curso. Se dan varios ejemplos para ilustrar la metodología propuesta.

1 Introducción

1.1 Motivación

Los modelos de mezclas son muy importantes en la teoría y las aplicaciones estadísticas en un gran número de áreas. La razón por la que han tenido una aplicación frecuente estos modelos, es porque permiten introducir variabilidad extra en los modelos. Lindsay (1995) presenta una excelente discusión del problema de mezclas y muestra que hay una gran cantidad de problemas en estadística que tienen una estructura de mezclas. Por esta razón, es común encontrar que los problemas de mezclas han sido discutidos en la literatura bajo diferentes denominaciones, como "modelos compuestos" (Panjer y Willmot 1992, Capítulo 7), "modelos jerárquicos" (Casella y Berger 2002, p.153) y "modelos de sobre-dispersión" (Gelman, et al. 1995, p. 350). Con frecuencia una mezcla de distribuciones modela la heterogeneidad no observable de una población. Supongamos que modelamos el fenómeno de interés con una densidad $f(x;\lambda)$ que contiene sólo un parámetro λ. Cuando la población es homogénea, el parámetro de la población está dado por un número λ. Sin embargo, cuando la población no es homogénea, este modelo resulta muy restrictivo y debemos considerar que el parámetro λ es variable y que sigue alguna distribución, de acuerdo al tipo de heterogeneidad de la población. Un ejemplo de esto lo tenemos en confiabilidad, cuando modelamos el tiempo de vida de los componentes con una distribución exponencial cuya tasa de falla es λ y por la variabilidad entre los componentes, modelamos la tasa de falla de los componentes con una distribución Gama.

1.2 El Problema

En la mayoría de los libros de texto, cuando se discute una mezcla de la forma $(X|Y = y) \sim f(x|y)$ y $Y \sim g(y)$, una gran parte de esos textos marginalizan la distribución con-

junta $f(x|y)g(y)$, para obtener la distribución $f_X(x)$ de la mezcla X. Entonces $f_X(x) = \int f(x|y)g(y)dy$ cuando $g(y)$ es continua y $f_X(x) = \sum f(x|y)g(y)$ cuando $g(y)$ es discreta.

Estos cálculos resultan algunas veces complicados y tediosos. El propósito de este artículo es mostrar que para algunas mezclas importantes, la determinación de la mezcla puede simplificarse obteniendo la FGM de X como $E[M_{X|Y}(t)]$ donde $M_{X|Y}(t)$ es la FGM de $X|Y$ y la esperanza se toma sobre la distribución de Y.

2 El Modelo de Mezclas

2.1 El modelo

En este trabajo discutimos modelos de mezclas en dos etapas, continuos y discretos, de la forma

$$f_X(x) = \int f(x|y)g(y)dy \tag{1}$$

donde $g(y)$ es una función de densidad continua, $f(x|y)$ es la distribución condicional de $X|Y$ y la integral es calculada sobre el soporte de $g(y)$, esto es, valores de y tales que $g(y) > 0$. Para una mezcla discreta, $g(y)$ es una función de probabilidad y la integral es reemplazada por una suma. A $g(y)$ se le conoce como la distribución mezclante (o latente). En términos de funciones de distribución acumuladas, la mezcla en (1) puede escribirse como $F_X(x) = \int F(x|y)g(y)dy$. La distribución de la mezcla, $f_X(x)$ y $F_X(x)$, puede depender de algunos parámetros fijos pero esa dependencia no es explícitamente indicada aquí. El modelo de mezclas indicado en (1) se denotará por la notación jerárquica

$$\left.\begin{array}{c} X|Y \sim f(x|Y) \\ Y \sim g(y) \end{array}\right\} \Longrightarrow X \sim f_X(x) \tag{2}$$

Algunos autores como Casella y Berger (2002, Sección 4.4) y Gelman et al. (1995, p. 350) usan una notación similar a (2) para un modelo de mezclas.

2.2 La FGM de una mezcla

Es bien conocido que si $(X, Y) \sim f(x|y)g(y)$ y X tiene esperanza finta, entonces $E(X) = E[E(X|Y)]$. De forma similar, si X tiene FGM $M_X(t)$, entonces $M_X(t) = E[\exp(tX)] = E[M_{X|Y}(t)]$, donde $M_{X|Y}(t)$ es la FGM de $X|Y$ y la esperanza $E[M_{X|Y}(t)]$ se calcula con respecto a $g(y)$.

Enseguida tenemos un resultado simple que se utiliza repetidamente en los ejemplos de la Sección 3.

Resultado 1. Considere el modelo de mezclas dado en (2)

1. Suponga que existe un $\delta > 0$ tal que para t en $(-\delta, \delta)$

$$M_{X|Y}(t) = C_1(t) \exp[C_2(t)Y],$$

donde $C_1(t)$ y $C_2(t)$ son funciones finitas de t que pueden depender de algunos parámetros fijos, pero no dependen de Y.

2. Suponga que la FGM, $M_Y(\cdot)$, de Y existe y que $M_Y[C_2(t)]$ es finita para t en $(-\delta, \delta)$.

Entonces la FGM de X esta dada por

$$M_X(t) = C_1(t) M_Y[C_2(t)],\ -\delta < t < \delta.$$

Demostración: Haciendo evaluaciones directas, tenemos que

$$\begin{aligned} M_X(t) &= E[M_{X|Y}(t)] = E[C_1(t)\exp(C_2(t)Y)] \\ &= C_1(t)E[\exp(C_2(t)Y)] = C_1(t)M_Y[C_2(t)]. \end{aligned}$$

3 Ejemplos

En esta sección describimos algunos ejemplos de modelos de mezclas en los que el uso de las FGMs dan alguna simplificación sobre el enfoque tradicional de marginalización de la distribución conjunta $f(x|y)g(y)$ para obtener $f_X(x)$.

3.1 Mezcla Gama-Poisson

La mezcla Gama de variables aleatorias Poisson es una mezcla de uso frecuente en casos de sobredispersión. Esta mezcla pertenece a una familia grande de distribuciones Poisson mezcladas, que han sido ampliamente estudiadas por Karlis y Xekalaki (2005) y otros. Greenwood y Yule (1920) propusieron esta mezcla para modelar la susceptibilidad a accidentes. Aquí probamos que

$$\left.\begin{array}{l} X|Y \sim POI(Y) \\ Y \sim GAM(\alpha, \beta) \end{array}\right\} \Longrightarrow X \sim BINNEG(\alpha, \frac{1}{1+\beta})$$

Demostración: Inicialmente tenemos que $M_{X|Y}(t) = \exp\{Y[\exp(t) - 1]\}$. Usando la FGM para la distribución Gama, $M(t) = (1 - \beta t)^{-\alpha}, t < 1/\beta$ y el **Resultado 1** con $C_1(t) = 1$ y $C_2(t) = \exp(t) - 1$ tenemos

$$\begin{aligned} M_X(t) &= M_Y[\exp(t) - 1] \\ &= \left(\frac{1}{1 - \beta[\exp(t) - 1]}\right)^{\alpha} = \left(\frac{1/(1+\beta)}{1 - [\beta/(1+\beta)]\exp(t)}\right)^{\alpha}. \end{aligned}$$

Esta FGM corresponde a la distribución Binomial Negativa, luego $X \sim BINNEG[\alpha, 1/(1+\beta)]$.

Cuando $\alpha = 1$ tenemos una mezcla exponencial de variables aleatorias Poisson que tiene una distribución geométrica (ver Casella y Berger 2002, p. 156). La demostración de este resultado es muy sencilla al usar el enfoque de la FGM.

3.2 Mezcla Normal-Normal

El uso de la mezcla de distribuciones normales se remonta a la época de Pearson (1894). Él usó una mezcla discreta de dos normales para modelar la mezcla de dos poblaciones de

cangrejos. Actualmente se utilizan con frecuencia mezclas discretas de distribuciones para modelar la mezcla de diferentes cohortes de peces en pesquerías. Este ejemplo presenta una mezcla continua de normales con varianza constante. Esta mezcla corresponde a un modelo de efectos aleatorios en la media para datos normales con varianza homogénea. Se prueba que:

$$\left.\begin{array}{c} X|Y \sim NOR(Y, \sigma^2) \\ Y \sim NOR(\mu, \tau^2) \end{array}\right\} \Longrightarrow X \sim NOR(\mu, \tau^2 + \sigma^2)$$

Demostración: La FGM de $X|Y$ es

$$M_X(t) = \exp\left(tY + \frac{\sigma^2 t^2}{2}\right) = \exp\left(\frac{\sigma^2 t^2}{2}\right) \exp\left(tY\right).$$

Aplicando el **Resultado 1** con $C_1(t) = \exp(\sigma^2 t^2/2)$ y $C_2(t) = t$ tenemos la FGM

$$M_X(t) = \exp\left[t\mu + \frac{(\sigma^2 + \tau^2)t^2}{2}\right],$$

que corresponde a la distribución $NOR(\mu, \tau^2 + \sigma^2)$.

Este desarrollo es más simple que el que se sigue para marginalizar la distribución conjunta en algunos textos, como puede verse en Mukhopadhyay (2000, p. 484).

3.3 La Logística como una mezcla SEV-LEV

La distribución de valores extremos para mínimos (SEV por sus siglas en inglés) y la distribución de valores extremos para máximos (LEV por sus siglas en inglés) son muy importantes en la modelación de datos de valores extremos (ver Lawless 2003, Capítulo 1 o Meeker y Escobar 1998, Capítulo 4). Las distribuciones SEV y LEV son distribuciones de cola izquierda pesada y cola derecha pesada respectivamente. El siguiente ejemplo muestra que la distribución logística es una mezcla SEV-LEV. Esto explica que esta distribución tenga colas relativamente mas pesadas que la distribución normal. Se prueba que:

$$\left.\begin{array}{c} X|Y \sim LEV(Y, \sigma) \\ \frac{Y}{\sigma} \sim SEV(\xi, 1) \end{array}\right\} \Longrightarrow X \sim LOGIS(\mu, \sigma),$$

donde $\mu = \sigma\xi$. Demostración: Las distribuciones LEV y SEV son distribuciones de localización y escala, con FGMs dadas por $M(t) = \exp(\mu t)\Gamma(1 - \sigma t), |t| < 1/\sigma$ y $M(t) = \exp(\mu t)\Gamma(1+\sigma t), t > -1/\sigma$, donde μ y σ son los parámetros de localización y escala, respectivamente, en ambas distribuciones. La FGM de $X|Y$ es $M_{X|Y}(t) = \exp(tY)\Gamma(1-\sigma t), t < 1/\sigma$. Usando la FGM de las distribuciones SEV y LEV y el **Resultado 1** con $C_1(t) = \Gamma(1-\sigma t)$ y $C_2(t) = t$, obtenemos la FGM de la distribución logística, $M_X(t) = \exp(\mu t)\Gamma(1-\sigma t)\Gamma(1+\sigma t)$, donde $\mu = \sigma\xi$ y $\Gamma(z)$ es la función gama evaluada en z.

3.4 Extensiones

Existen ejemplos adicionales en los que el uso de las mezclas de distribuciones es de utilidad para la obtención de distribuciones. En ecología y entomología se utiliza usualmente la distribución Neyman Tipo A que resulta de una mezcla Poisson de distribuciones Poisson. Para

distribuciones que no tienen FGM, podemos utilizar la función característica para extender el procedimiento propuesto en este artículo. Un ejemplo de esto lo es la distribución t de Student que también puede obtenerse de una mezcla de distribuciones normales $NOR(0, p/Y)$, con Y distribuida como χ–cuadrada con p grados de libertad. Los detalles de este desarrollo pueden encontrarse en Villa y Escobar (2006). Existen otras distribuciones de interés como la distribución F (Casella y Berger 2002, p. 259) y la distribución Pareto (Cox y Oakes 1984, p. 19) que surgen como mezclas pero que no tienen FGM. Es fácil obtener las funciones características de estas distribuciones utilizando el método condicional aquí propuesto.

3.5 Conclusiones

Existen muchas mezclas de distribuciones que tienen una FGM. La derivación de la mezcla de distribuciones utilizando FGMs resulta más sencilla y breve que cuando se obtiene marginalizando la distribución conjunta, debido a que nos apoyamos en FGMs que se obtuvieron previamente y para el uso de este método sólo requerimos conocer las FGMs de las distribuciones. Sin embargo no debería sorprendernos que haya algunos ejemplos en donde ocurre lo contrario y por lo tanto es mejor usar el enfoque de marginalización. Desde un punto de vista pedagógico, el enfoque de la FGM para obtener distribuciones de mezclas debería enseñarse como una herramienta útil para un gran número de modelos de mezclas. Cuando los estudiantes tienen un conocimiento operativo del tema de variables complejas, podemos usar el método general basado en funciones características.

Bibliografía

[1] Casella, G., and Berger, R. L. (2002), *Statistical Inference*, Second edition. Pacific Grove. CA: Duxbury.

[2] Cox, D. R. and Oakes, D. (1984), *Analysis of Survival Data*, New York, NY: Chapman and Hall Ltd.

[3] Greenwood, M. and Yule, G. (1920), "An Inquire into the Nature of Frequency Distributions representative of Multiple Happenings with Particular reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents", Journal of the Royal Statistical Society, A, **83**, 255-279.

[4] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysys* , New York, NY: Chapman & Hall.

[5] Karlis, D. , and Xekalaki, E. (2005), "Mixed Poisson Distributions", International Statistical Review, **73**, 35-58.

[6] Lawless, J. F. (2003), *Statistical Models and Methods for Lifetime Data*, Second Edition. New York NY: John Wiley & Sons.

[7] Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5, Hayward, CA: Institute of Mathematical Statistics.

[8] Meeker, W. Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York NY: John Wiley & Sons.

[9] Mukhopadhyay, N. (2000), *Probability and Statistical Inference*, New York: Marcel Dekker.

[10] Panjer, H. H. and Willmot, G. E. (1992), *Insuranse Risk Models*. Schaumburg IL: Society of Actuaries.

[11] Pearson, K. (1894), "Contrbutions to the mathematical theory of evolution" , Phil. Trans. Roy. Soc. London (A), **185**, 71-100.

[12] Villa, E. R., and Escobar, L. A. (2006) "Using Moment Generating Functions to Derive Mixture Distributions", The American Statistician, February 2006, Vol. 60, Num. 1, pp. 75-80.