MIXTURE OF DISTRIBUTION MODELS: APPLICATION TO THE AGE DISTRIBUTION OF PEOPLE WHO COMMITTED SUICIDE

Andrea Ruiz Vega[1], Miguel Ángel Montero Alonso, Juan de Dios Luna del Castillo
Dpto. de Estadística e Investigación Operativa. Universidad de Granada (España)

ABSTRACT
Mixture of distribution models are useful tools for analyzing data sets derived from diverse subpopulations, modeled through a weighted combination of simpler distributions. Various parameters, including mean or shape parameters, may be estimated using different methods like the EM algorithm or maximum likelihood method. These mixtures aid in modeling the heterogeneity of multimodal data distributions. The age distribution based on the number of suicides between 2002 and 2020 demonstrates several peaks where the majority of observations are clustered; hence, this modeling approach is employed for its examination. These data can be represented using a mixture of two-component normal distributions. The goodness of fit of the model and the estimation of its parameters will support such a result. In conclusion, the aim is to apply mixture models to the age distribution of people who have committed suicide and identify subgroups that exhibit specific behaviors or characteristics. This will provide meaningful information to create effective prevention treatments and aids for different age groups of people at risk.

KEY WORDS: Mixture Models, components, subgroups, algorithm, suicides.

MSC: 62P10

RESUMEN
Los modelos de mezcla de distribuciones son herramientas útiles para analizar conjuntos de datos derivados de diversas subpoblaciones, modelados mediante una combinación ponderada de distribuciones más simples. Se pueden estimar diversos parámetros, como la media o los parámetros de forma, utilizando distintos métodos, como el algoritmo EM o el método de máxima verosimilitud. Estas mezclas ayudan a modelizar la heterogeneidad de las distribuciones de datos multimodales. La distribución por edades basada en el número de suicidios entre 2002 y 2020 muestra varios picos en los que se agrupa la mayoría de las observaciones; por lo tanto, se emplea este enfoque de modelización para su examen. Estos datos pueden representarse mediante una mezcla de distribuciones normales de dos componentes. La bondad del ajuste del modelo y la estimación de sus parámetros corroborarán tal resultado. En conclusión, el objetivo es aplicar modelos de mezcla a la distribución por edades de las personas que se han suicidado e identificar subgrupos que presenten comportamientos o características específicas. Esto proporcionará información significativa para crear tratamientos de prevención y ayudas eficaces para los diferentes grupos de edad de personas en riesgo.

PALABRAS CLAVES: Modelos de mezcla, componentes, subgrupos, algoritmo, suicidios.

1. INTRODUCCIÓN

En la sociedad actual, el suicidio y los factores que llevan a ello se han convertido en un tema de máxima importancia. Es fundamental comprender las características demográficas, sociales y personales de los individuos que se han suicidado para poder desarrollar estrategias de prevención efectivas mejorando el apoyo y la ayuda a quienes se encuentran en situaciones que puedan llevarlos a cometer tal acto. En España, diariamente mueren por esta causa una media de diez personas. Aunque se encuentre entre los países de Europa con tasas más bajas, el suicidio sigue siendo la primera causa de muerte externa en este país (Ministerio de Sanidad - Gabinete de Prensa - Notas de Prensa, 2020). En el contexto estadístico, se plantea la cuestión de que la edad sea un factor influyente sobre el suicidio, por lo que un análisis de su distribución desempeña un papel crucial.
La edad puede proporcionar información necesaria sobre las diferentes situaciones en las que se puedan encontrar las personas como momentos de vulnerabilidad relacionados con posibles desencadenantes que los lleve a suicidarse. Ya que hay muchos factores que pueden estar relacionados, la distribución de la edad puede ser heterogénea (McLachlan & Peel, 2000), lo que supone una mayor dificultad a la hora de realizar un enfoque estadístico. El problema recae en que hay veces en las que un modelo simple no se ajusta de manera tan fiel como puede hacerlo un modelo más complejo (McLachlan & Peel, 2000).

1 1 rvandrea@correo.ugr.es, 2 mmontero@ugr.es, 3 jdluna@ugr.es

Muchos investigadores proponen abordar esta cuestión mediante modelos de mezcla de distribuciones, los cuales permiten analizar poblaciones compuestas por subgrupos que presenten distintos comportamientos o incluso sigan diferentes distribuciones.
El primer estudio conocido sobre mezcla de distribuciones se atribuye al matemático inglés Thomas Bayes en el siglo XVIII. Aunque no utilizó el término "mezcla de distribuciones", asentó las bases teóricas para su futuro desarrollo. Abordó el problema de la estimación de probabilidades cuando se desconoce parte de la información y se tienen distintas fuentes de datos. Propuso la hipótesis de que los datos provengan de distribuciones variadas que se puedan mezclar de manera ponderada (Bayes & Price, 1763). A lo largo de los años son muchos los investigadores que se enfrentan a esta parte de la estadística y se ha podido comprobar que, hoy en día, junto con los avances computacionales, es útil en muchos estudios resultando fácil hacer uso de ello.
Una de las características de los datos relacionados con la mezcla de distribuciones es que puedan presentar una multimodalidad, la cual sería más difícil de ajustar con un modelo simple. Una moda indica un punto en la distribución donde se aglomeren una gran cantidad de observaciones. Como indica el mismo nombre, una multimodalidad implica varios picos en la distribución de los datos, lo que puede suponer que la población esté dividida en subgrupos con distintos comportamientos. De esta manera, un modelo simple no se ajustará completamente a esta característica, por lo que no es extraño que se recurra a los modelos de mezcla.
La distribución sobre la que se partirá será de Poisson ya que se suele utilizar para modelizar datos que midan el número de eventos que ocurren en un periodo de tiempo (Salinas-Rodríguez, 2009), que es lo que sucede con los datos del presente trabajo, el número de suicidios en España entre 2002 y 2021. A medida que se avance en el estudio se verá que es mejor hacer uso de la mezcla de distribuciones normales, ya que la distribución de Poisson se puede aproximar a la normal si se cumplen determinadas características. Esto hace que estén estrechamente relacionadas.
Teniendo como objetivos estudiar la multimodalidad que presentan los datos y ajustarles las distribuciones pertinentes mediante modelos de mezcla exponiendo sus resultados, se abordará la cuestión teórica que engloba los modelos de mezcla de distribuciones de variables aleatorias continuas y discretas. Se mencionarán sus características, sus funciones de distribución y de densidad, y distintos métodos para estimar el número de componentes y el valor de sus parámetros. Se completará el capítulo con el estudio de la bondad de los ajustes mediante criterios de información como el BIC y el AIC.
Se abordará la aplicación práctica del estudio describiendo los datos, el origen de éstos y la limpieza realizada para poder tratar con ellos de una manera más simple. Haciendo distinción entre mujeres y hombres, se expondrán mediante gráficas el número de suicidios quinquenales en grupos de edad, y la distribución de las tasas de mortalidad también por grupos de edad en los intervalos de años establecidos. Para esta aplicación práctica se usará el *software* estadístico de R en la versión 2023.06.0+421 de RStudio. Se modelará una mezcla de distribuciones normales a cada quinquenio tanto para los hombres como para las mujeres. Se expondrán los métodos usados para estimar el número de componentes de los modelos y el valor de los parámetros que componen las distribuciones resultantes. Los resultados obtenidos motivan la propuesta de la creación de terapias especializadas para los distintos subgrupos que conforman la población de estudio.

2. MEZCLA DE DISTRIBUCIONES DE VARIABLES ALEATORIAS CONTINUAS Y DISCRETAS

Este apartado se basa en su gran mayoría en el estudio de McLachlan y Peel (2000) sobre modelos de mezcla finita, el cual reflejaron en el libro que publicaron de manera conjunta *Finite Mixture Models* y, por otro lado, se usa el libro de Schlattmann (2009) *Medical Applications of Finite Mixture Models*. Ambos libros son piezas fundamentales en la modelización de mezcla de distribuciones finitas. Muchos investigadores hacen referencia a ellos en sus estudios, ya que se enfocan en el desarrollo y la aplicación de estos modelos resultantes ofreciendo también una descripción exhaustiva de los métodos usados en estimación de parámetros de dichos modelos.
En la última década el potencial de las aplicaciones de estos modelos ha crecido de manera considerable gracias a su flexibilidad. Los modelos de mezcla finita tienen una gran aplicación en el análisis de datos ya que representan la distribución de probabilidad de una variable aleatoria como combinación de otras distribuciones más simples. A cada una de las distribuciones simples se le llama componente de mezcla y se entiende que los datos vienen de forma aleatoria de una de las distribuciones, ya sea continua o discreta (McLachlan & Peel, 2000). Eligiendo el número de componentes de manera adecuada — representando las áreas de la distribución correctamente — un modelo de mezcla puede proporcionar un modelo adecuado para las variaciones locales de los datos.

Uno de los primeros análisis de modelos de mezcla fue realizado por Pearson, donde ajustó dos funciones de densidad normales con diferentes parámetros en proporciones π_1 y π_2 (McLachlan & Peel, 2000). Everitt (1996), en el artículo *An introduction to finite mixture distributions*, también habla de los modelos de mezcla finita de la misma manera que McLachlan y Schlattmann. Los autores coinciden en que son una clase de modelos estadísticos flexibles usados para establecer un modelo sobre datos ya sean continuos, categóricos o de conteo, que provengan de distintos grupos o subpoblaciones. Estos modelos de mezcla finita pueden usarse también para estimar el número de subpoblaciones y los parámetros correspondientes (Everitt, 1996). Gracias a ello, se pueden aplicar en una gran cantidad de áreas de la ciencia, desde la ingeniería hasta la psicología y la medicina.

Son muchas las diferentes distribuciones de probabilidad que se pueden aplicar en estos modelos, incluyéndose así la distribución normal, la exponencial y la ***t*** de Student, entre otras (Everitt, 1996). Las distribuciones binomiales negativas y de Poisson son las más comúnmente utilizadas en datos de tipo conteo. La diferencia entre los datos que siguen ambas distribuciones es que la distribución binomial negativa se utiliza para modelar eventos en los que se repiten ensayos hasta que se alcance un número de éxitos deseados; mientras que la distribución de Poisson se usa para eventos discretos que ocurren de manera aleatoria en un intervalo de tiempo (Salinas-Rodríguez, 2009). Como el número de suicidios en un determinado grupo de edad son datos de tipo conteo, eventos discretos independientes, el presente estudio se enfoca en los modelos que se ajusten mediante la distribución de Poisson y su aproximación correspondiente. Un modelo simple puede llegar a ser muy estricto para llegar a describir los datos reales, por lo que una extensión heterogénea de estos modelos, en el que se supone que la población de estudio está dividida en varias subpoblaciones ($\lambda_1, \lambda_2, \dots, \lambda_k$), será más adecuada (Schlattmann, 2009).

Por tanto, como características de los modelos de mezcla de distribución finita (MMDF) se tiene que ayudan a describir poblaciones heterogéneas que sean difíciles de ajustar con distribuciones simples. Están compuestas de subpoblaciones que se ajustan a una distribución de probabilidad específica, por lo que están compuestas por una combinación ponderada de distribuciones más simples en la que cada subpoblación tiene un peso o proporción en la población total.

Por otro lado, siendo $F_i(x)$ la función de distribución acumulada, $f_i(x)$ la función de densidad, μ_i, σ_i^2 y π_i la media, la varianza y la proporción sobre la población total correspondiente a la distribución de la i-ésima subpoblación, y cumpliéndose los requisitos $\sum_{i=1}^{k} \pi_i = 1$ y $0 \leq \pi_i \leq 1$, la función de distribución y la función de densidad de los MMD se expresan como una combinación ponderada de la función de distribución y la función de densidad, respectivamente, de las subpoblaciones (McLachlan & Peel, 2000). Fijando que se parte de *K* subpoblaciones, la función de distribución y la función de densidad del modelo de mezcla son las siguientes:

$$F(x) = \sum_{i=1}^{k} \pi_i * F_i(x)$$

$$f(x) = \sum_{i=1}^{k} \pi_i * f_i(x)$$

La media y la varianza, de la misma manera, también se pueden calcular a partir de las subpoblaciones. Vendrán representadas como:

$$E(x) = \sum_{i=1}^{k} \pi_i * \mu_i(x)$$

$$var(x) = \sum_{i=1}^{k} \pi_i * (\sigma_i^2 + (\mu_i - E(x))^2)$$

La varianza del modelo de mezcla se descompone en dos partes. El primer término, $\pi_i * \sigma_i^2$, representa la varianza dentro de cada subpoblación de manera individual. El segundo término, $\pi_i * (\mu_i - E(x))^2$, representa la varianza debida a las diferencias entre las medias de las subpoblaciones individuales (π_i) y la media del modelo de mezcla ($E(x)$) (Deb, s. f.).

Los datos pueden venir de un proceso con un comportamiento más complejo de lo que puede modelar una única distribución. Cada subpoblación puede verse tratada como una aproximación a la distribución de Poisson, por lo que resulta en una mezcla de distribuciones normales (Jacobs, 2022) al ser aproximaciones de las de Poisson. Pueden representar de manera más fiel la variabilidad y la heterogeneidad de los datos, y proporcionan una gran flexibilidad a la hora de modelizar, adaptándose de mejor forma a los datos observados.

La cuestión entonces es contrastar la hipótesis de ajustar una distribución de un único componente frente a la alternativa de una mezcla con varios, por lo que se necesita estimar este número de componentes *k*. Para poder llevar a cabo esta modelización, se deben seguir determinados pasos para que la bondad de

dicho ajuste sea adecuada. Esta elección debe verse apoyada por el análisis del modelo resultante, a lo que Everitt (1996) presenta algunos enfoques de selección del número de subpoblaciones, como el criterio de información bayesiano (BIC) y el criterio de información de Akaike (AIC). Para los criterios de información BIC y AIC, cuanto menor sea el valor mejor será el ajuste del modelo (Amat Rodrigo, 2020). Ambas métricas solo sirven para comparar la calidad entre distintos modelos, no indican si el ajuste es realmente bueno o malo.

Especificar el número de componentes termina siendo complicado, por lo que en la parte práctica de este estudio se crea un proceso relativamente sencillo (Deb, s.f., Diapositiva número 40) que consiste en estimar modelos de uno y más componentes, calcular para cada uno de ellos el Criterio de Información Bayesiano (BIC) y terminar seleccionando el modelo con menor valor BIC.

$BIC = -2\log(L) + k * \log(N)$,

siendo L el estimador máximo-verosímil (*likelihood*), k el número de parámetros y N el número de observaciones.

Entonces, el primer paso para ajustar cada uno de los modelos de mezcla es estimar los parámetros de las subpoblaciones que los componen, es decir, encontrar el valor de los parámetros con los que dichas distribuciones pueden haber generado con mayor probabilidad los datos observados. McLachlan y Peel (2000) y muchos otros investigadores hacen referencia a distintos métodos de estimación para hallar esos parámetros. Los más comunes son el método de máxima verosimilitud (ML) y el algoritmo expectativa - maximización (EM).

El método de máxima verosimilitud pretende encontrar los valores de los parámetros que maximizan la función de verosimilitud de los datos. Esta función se construye multiplicando las funciones de densidad de probabilidad de cada componente y representa la probabilidad conjunta de obtener los datos observados con los parámetros establecidos. Los valores que maximicen esta función serán considerados las estimaciones de máxima verosimilitud de los parámetros del modelo (Schlattmann, 2009).

El algoritmo EM es un método iterativo para hallar los estimadores máximo-verosímiles con datos incompletos o variables latentes, con el objetivo de encontrar los valores más probables para estas variables u observaciones que faltan. Consta de dos pasos, el *E-step* (paso de expectativa) y el *M-step* (paso de maximización). Primero se calculan los valores esperados de las variables latentes o de los datos faltantes usando los valores actuales de los parámetros del modelo, los cuales se eligen de manera arbitraria o mediante un proceso de inicialización, y en el segundo paso se estiman los nuevos valores de los parámetros maximizando la función de verosimilitud utilizando las expectativas calculadas anteriormente (Schlattmann, 2009). Estos pasos se repiten hasta que los parámetros converjan a un valor estable.

La elección entre los distintos métodos depende de varios factores, como la disponibilidad de datos completos o incompletos, la complejidad del modelo y los objetivos específicos de la estimación. Mientras que el algoritmo EM calcula de manera iterativa estimaciones de los parámetros, también puede converger a máximos locales o incluso tener una velocidad de convergencia lenta. Por otro lado, el método ML, a pesar de ser eficiente y teniendo las estimaciones que convergen al valor correcto a media que la muestra aumenta de tamaño, tiene la desventaja de que requiere datos completos (McLachlan & Peel, 2000).

En resumen, muchos investigadores han tratado la cuestión de modelizar mezclas de distribuciones finitas siendo variables tanto aleatorias como continuas. La construcción del modelo constará de una serie de pasos para asegurar la bondad de este. Primero, se determina un modelo teórico estableciendo las variables de interés y la relaciones entre ellas. Luego, se estima el número de componentes, utilizando distintos criterios de información, y también el valor de los parámetros mediante métodos de estimación adecuados. Tras ello, se considera si aceptar el modelo o no, valorando la diferencia entre los datos observados y los ajustados. Una vez esté aprobado, solo quedará interpretarlo con respecto a la variable de respuesta.

3. MULTIMODALIDAD DE LA EDAD EN DIFERENTES EJEMPLOS DE LA MEDICINA

Se puede encontrar la presencia de multimodalidad en los datos de distintos estudios en la rama de la medicina. La multimodalidad en las distribuciones se refiere a la presencia de más de una moda en los datos de una población. En otras palabras, es una distribución que presenta más de un valor que se repite con frecuencia. Esto puede deberse a varios factores que influirán en las como pueden ser socioeconómicos, ambientales, relacionados con la genética o incluso con los distintos estilos de vida.

Se puede comprobar fácilmente cómo la edad puede influir en distintos estudios, tanto como factor de riesgo, como factor beneficioso o de protección. Por ejemplo, el cáncer de mama, a pesar de que es más probable que se manifieste en mujeres de mayor edad, también puede presentarse a edad temprana de la misma manera que los hombres con cáncer de próstata. Las enfermedades cardiovasculares son otro buen

ejemplo. La hipertensión arterial puede ocurrir en cualquier grupo de edad, aunque sea más frecuente en personas adultas. Estas diferencias en la edad están influenciadas por distintos factores como pueden ser genéticos o medioambientales. Tratando con esta multimodalidad es probable que un modelo simple no se ajuste correctamente a la heterogeneidad que puedan presentar los datos. Un gran desafío con esta cuestión es encontrar los puntos de corte que delimitan el cambio entre las distribuciones. Esto es muy importante para analizar los subgrupos que han dado lugar a este fenómeno. Pueden tener distintas características o comportamientos que pueden ser significativos en la interpretación del estudio. Al conocer estos puntos de corte se permite establecer criterios de clasificación entre un grupo u otro.
Gong et al (2017) en su artículo *Bimodal distribution of fasting plasma glucose in the Uyghur and Han populations of Xinjiang (China)*, con el objetivo de describir la distribución de la glucosa plasmática en ayunas en las dos poblaciones de Xinjiang, detectaron la presencia de bimodalidad en la edad, por lo que querían estimar los puntos de corte de esta distribución. Contrastaron la bondad del ajuste de una distribución unimodal con la bimodal. Sin embargo, los puntos de corte estimados no fueron significativos biológicamente y concluyeron que la distribución bimodal no era útil para diagnosticar diabetes en Xinjiang.
Un estudio de los autores Louis y Dogu donde comprueban si la edad de inicio del temblor esencial (TE), definida como la presencia de temblor de amplitud moderada en la cabeza o en los brazos (Louis, E. D., & Dogu, O., 2007), sigue también una distribución bimodal. En este caso estudian dos entornos distintos, uno basado en un centro de referencia terciario y otro basado en la población general. En la población se vio un pequeño pico a los 30 años, que correspondía a un 14.1% de los casos, mientras que en edades avanzadas se observó un pico más marcado que suponía un 85.9% de los casos. En el grupo de referencia terciario, también concluyeron con una distribución bimodal con un gran pico a los 40 años y otro pico más adelante, que suponían un 42.2% y un 57.8% de los datos respectivamente.
Mobbs et al (1993) estudia en el artículo *Evidence for bimodal distributuon of breast carcinoma ER and PgR values quantitated by enzyme inmunoassay* la distribución de los receptores de estrógeno (RE) y de progesterona (PgR) del carcinoma de mama. Pudieron ver que los valores de los carcinomas de mama mostraban una bimodalidad en los histogramas de frecuencias. No obstante, en el caso de los receptores de estrógenos, la bimodalidad era más acentuada que en el caso de los receptores de progesterona. Detectaron que la edad, el estado menopáusico y la variación del kit de preparación de citosol podían afectar al ajuste de la distribución, sin embargo, vieron que ninguno era significativo para la distribución.
En el artículo *Age at oneset of first suicide attempt: Exploring the utility of a potential candidate variable to subgroup attempters,* Menon et al (2018) tuvo como objetivo evaluar la utilidad de la edad como factor de identificación de subgrupos en personas que han cometido suicidio. Se basaron en 895 historias clínicas de pacientes de la Clínica de Intervención en Crisis (CIC) del Departamento de Psiquiatría de un hospital de atención terciaria en Puducherry, India del Sur. En lugar de formar los subgrupos a partir de los diagnósticos psiquiátricos, vieron que los datos variaban en función de la edad por lo que dividieron los subgrupos en base a ese factor. La edad terminó siendo un marcador útil para delinear estos subgrupos. Las subpoblaciones resultantes fueron “inicio temprano” con la edad al primer intento menor o igual a los 31 años y el “inicio tardío” con la edad superior a los 31.
En resumen, la multimodalidad de la edad es más común de lo que se puede llegar a pensar, por ello, es importante considerar que esa variable sea un factor significativo. Su utilidad, por tanto, recae en este punto, así se puede comprender de una manera más exacta la influencia de los factores y establecer estrategias para prevenir o tratar las enfermedades en cuestión.

4. DESCRIPCIÓN DE LOS DATOS DE MORTALIDAD POR SUICIDIO DE LOS AÑOS 2002 A 2021 QUE FIGURAN EN EL INE

Los datos que se usan en este trabajo son los datos de mortalidad por suicidio a nivel nacional entre los años 2002 y 2021 proporcionados por el Instituto Nacional de Estadística de España (INE - Instituto Nacional de Estadística, 2022a). Los datos están dispuestos por sexo y por grupos de edad divididos en intervalos de 5 años. Se han agrupado los grupos menos de 1 año, de 1 a 4 años y de 5 a 9 años, con el grupo de 10 a 14 años ya que prácticamente todos los datos eran 0 salvo un par de personas en el grupo de 5 a 9 años, una niña en 2005 y un niño en 2020. El fichero de datos resultante está dispuesto como un *dataframe* en el que cada fila contiene los datos para el año, el sexo, el grupo de edad, la frecuencia de suicidios y la población correspondiente a cada categoría, la cual también ha sido obtenida de la misma fuente (INE - Instituto Nacional de Estadística, 2022b) . Como vista principal, se muestra en la siguiente tabla un resumen de la estructura de este fichero.
Tabla 1: Datos Nacionales.

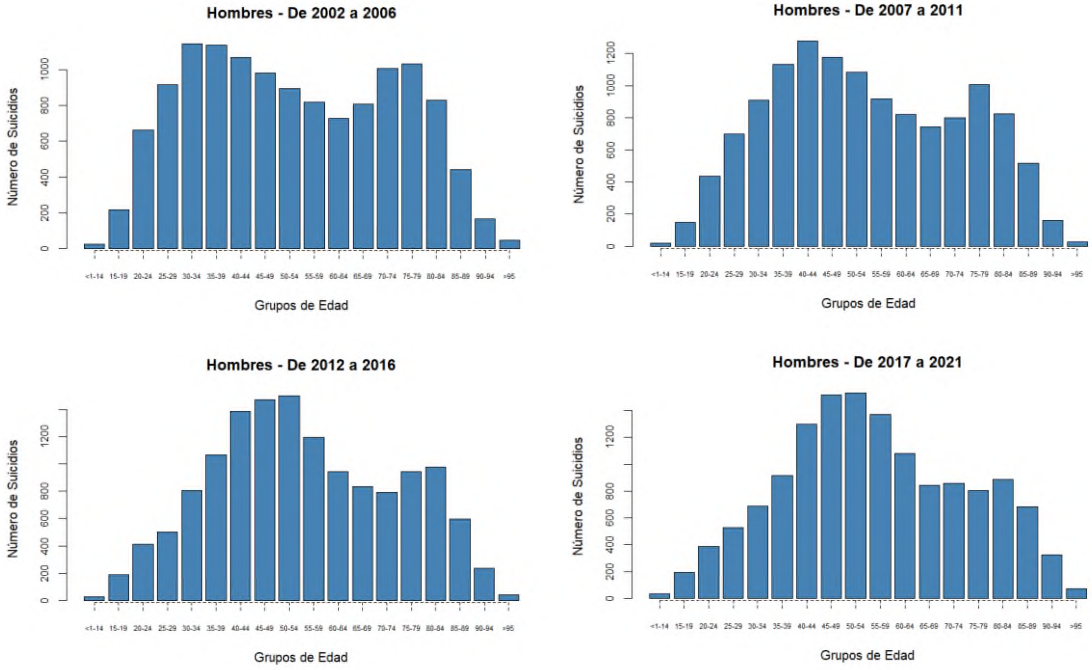
Año	Sexo	Edad	Recuento	Población
2021	Hombre	De < 1 a 14 años	14	3.483.338

2021	Hombre	De 15 a 19 años	28	1.259.328
…	…	…	…	…
2002	Hombre	De < 1 a 14 años	7	3.056.810
…	…	…	…	…
2002	Mujer	95 y más años	1	32.753

Descripción de los datos

Según los datos, la mortalidad por suicidio ha experimentado un ligero ascenso desde 2002 a 2021. En 2002 se registraron 3.371 fallecimientos por esta causa, mientras que en 2021 se registraron 4.003. El año con mayor número de suicidios fue para ambos sexos fue el último con 2.982 hombres y 1.021 mujeres, mientras que el año con menor número de suicidios fue 2011 para los hombres con 2.435 muertes y 2010 para las mujeres con 690. Por otro lado, cabe destacar que como promedio al año mueren por esta causa 2.679 hombres y 862 mujeres. Más adelante se verá que es importante tener en cuenta la población de cada año, ya que al no ser las mismas, el número de suicidios de cada población representa un porcentaje mayor o menor de la población.

El objetivo del estudio es ver el cambio en la distribución de la edad con el paso del tiempo. Sin embargo, si se quisiera observar los grupos de edad año a año, se estaría trabajando continuamente con 20 gráficas correspondientes a los años del estudio, por lo que la extensión de este trabajo sería muy grande. Por ello, se ha querido representar los datos en intervalos de cinco años, teniendo así tan solo 4 gráficas para ambos sexos. Para tener una primera visual descriptiva, mostramos en las siguientes gráficas el número de suicidios tanto para hombres como para mujeres en cada quinquenio de años por grupos de edad.



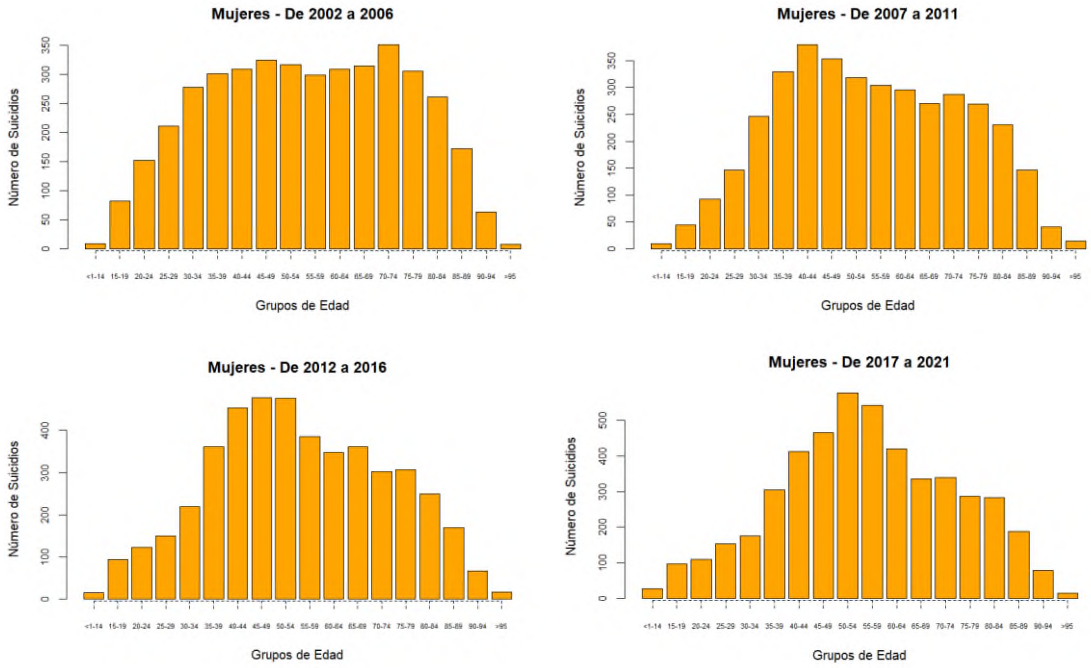


Figura 1: Número de suicidios para hombres y mujeres en cada quinquenio por grupos de edad.

La distribución de los datos no tiene una forma característica representativa de una distribución de Poisson. De hecho, presenta una bimodalidad clara en algunos de los quinquenios. Específicamente se ve más claro en los hombres, sobre todo en los primeros cinco años donde las dos modas están en los grupos de 30 a 34 y de 75 a 79. A medida que pasan los años, estas modas se van difuminando, como en el último lustro en el que la primera moda se sitúa entre 50 y 54 años, y la segunda moda se percibe muy levemente en el grupo de 80 a 84. Por otro lado, para las mujeres esta bimodalidad se aprecia menos. Es entre 2007 y 2011 cuando mejor se aprecia e incluso se podría intuir en los cinco años siguientes. Por ello, ya que estas modas pueden estar más o menos ocultas a simple vista en las mujeres, se asume que presentan la misma bimodalidad que los hombres. Esta característica indica una amplia dispersión de los datos (Jabeen, 2019). Como las gráficas muestran varios picos o modas, es probable que los datos se puedan clasificar en varias subpoblaciones o categorías.

Cálculo de tasas de suicidio

Como se mencionó al principio de este punto, los grupos de edad tienen distinto peso en la distribución que se ajuste, ya que tienen distinta población. Como los grupos no son homogéneos, si se modela directamente los datos de conteo se estaría induciendo a un error a la hora de comparar los resultados. Las tasas permiten comparar entre diferentes poblaciones o grupos que tengan tamaños de población distintos. Normalizan el tamaño de la exposición y tienen en cuenta el tiempo de riesgo asociado o la cantidad de exposición (McLachlan & Peel, 2000). También se tiende a reducir la varianza, que es útil cuando los conteos son bajos, como ocurre en este estudio. Por último, pero no menos importante, en muchos modelos estadísticos se asume que los errores siguen una distribución normal, por lo que, al modelar tasas, las cuales se basan en estos errores, es más probable que se cumplan los supuestos de normalidad conduciendo así a una mejor estimación (Jabeen, 2019).

Las tasas de mortalidad por suicidio por cada 100.000 habitantes indican el número de muertes ocurridas por este suceso durante un período determinado en una población específica, que son los distintos grupos de edad para hombres y mujeres. Se calculan las tasas mediante la siguiente fórmula.

$$Tasa\ de\ suicidio = \frac{Número\ de\ fallecidos\ por\ suicidio}{Población\ en\ el\ grupo\ de\ edad\ y\ sexo} * 100.000$$

Por tanto, cuanto mayor sea esa tasa por 100.000 habitantes, mayor será la mortalidad en esa población. Observar estas tasas a lo largo del tiempo ayudará a identificar tendencias o patrones en la distribución de la edad. Una disminución de las tasas puede implicar, por ejemplo, una mejora en la eficacia de la terapia psicológica o en estilos de vida, mientras que un aumento puede implicar una crisis socioeconómica como la que se vivió en España hace varios años, por lo que se podrá evaluar el progreso de la salud mental de la población. En resumen, la importancia del cálculo de tasas reside en que proporcionan una medida estandarizada para evaluar la mortalidad de la población permitiendo así comparar a lo largo del tiempo y entre diferentes subpoblaciones.

Según los datos, la tasa de mortalidad por suicidio en España en 2002 fue de 822 fallecidos por cada 100.000 habitantes, mientras que en 2021 fue de 845. Este último año corresponde a la tasa más alta de mortalidad por suicidio que se ha registrado. La tasa menor fue en 2010 con 680 fallecidos. Por otro lado, si se observan las tasas dispuestas en los intervalos de cinco en cinco años, el grupo con la mayor tasa de suicidios fue el último quinquenio, entre 2017 y 2021, situándose en 802 fallecidos. En el primer quinquenio, la tasa se sitúa en 800 suicidios. El intervalo con menor tasa corresponde a los años entre el 2007 y 2011 con 718 fallecidos por cada 100.000 habitantes.
Se expone en las siguientes gráficas el diagrama de barras de las tasas de suicidio por grupos de edad e intervalos de años haciendo distinción entre las dos categorías de la variable 'Sexo'.

- Hombres

El suicidio, por lo visto, es más frecuente en hombres que en mujeres. Se puede ver que la frecuencia de suicidarse aumenta drásticamente en personas de edad avanzada. A partir de los 75 años, las tasas crecen exponencialmente en cada quinquenio. La mayor tasa de suicidios en hombres corresponde a las personas mayores de 95 años entre los años 2002 y 2006 con una tasa de 8404 suicidios por cada 100.000 habitantes. Es en el primer quinquenio cuando más marcada está esta subida en comparación con el resto de los intervalos. Lo más interesante del análisis descriptivo se muestra en las ondas que presenta la distribución, este es el patrón que presenta la bimodalidad. Es entre los años 2012 y 2016, seguido de los últimos cinco años, donde mejor se aprecia, siendo por otro lado más suave entre 2002 y 2006.

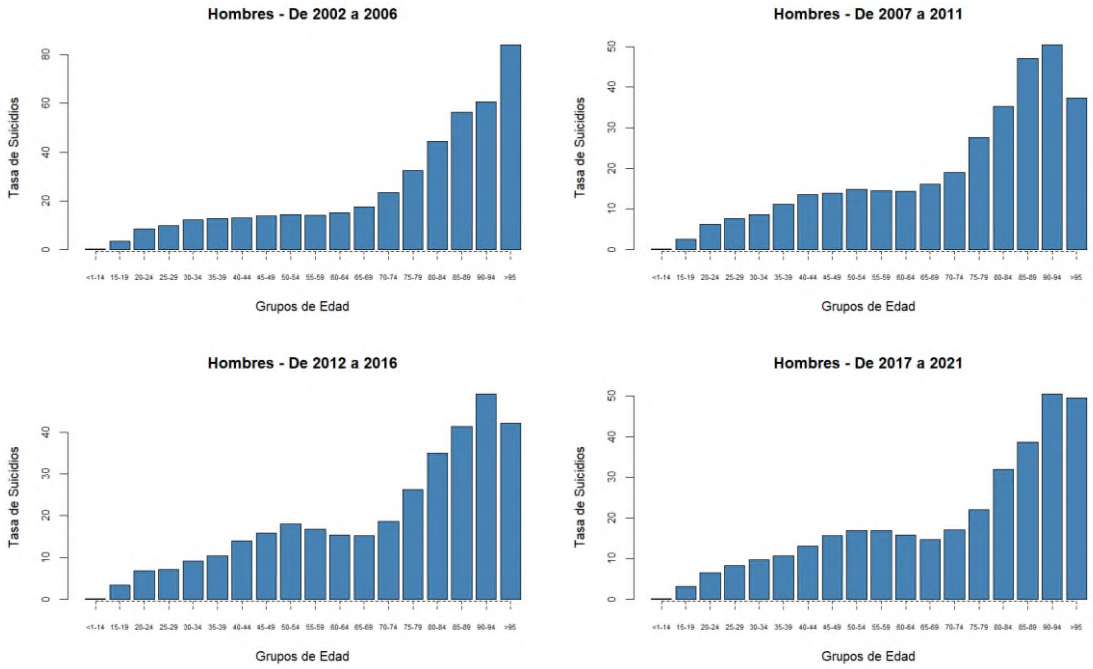


Figura 2: Tasa de suicidios en hombres en cada quinquenio por grupos de edad.

- Mujeres

Para las mujeres, el crecimiento de las tasas a medida que sube la edad no es tan exagerado aunque sea igualmente creciente. Al igual que los hombres, ocurre que en el primer quinquenio es cuando más marcada está esta subida, mientras que entre los años 2012 y 2016 las tasas están más equilibradas. La mayor tasa corresponde al grupo entre 85 y 89 años también entre 2002 y 2006 con una tasa de 1008 mujeres fallecidas por cada 100.000. Como se mencionaba al mostrar la distribución de la edad en función del número de suicidios, para las mujeres la bimodalidad está menos marcada, menos para el último quinquenio en el que se puede distinguir con claridad.

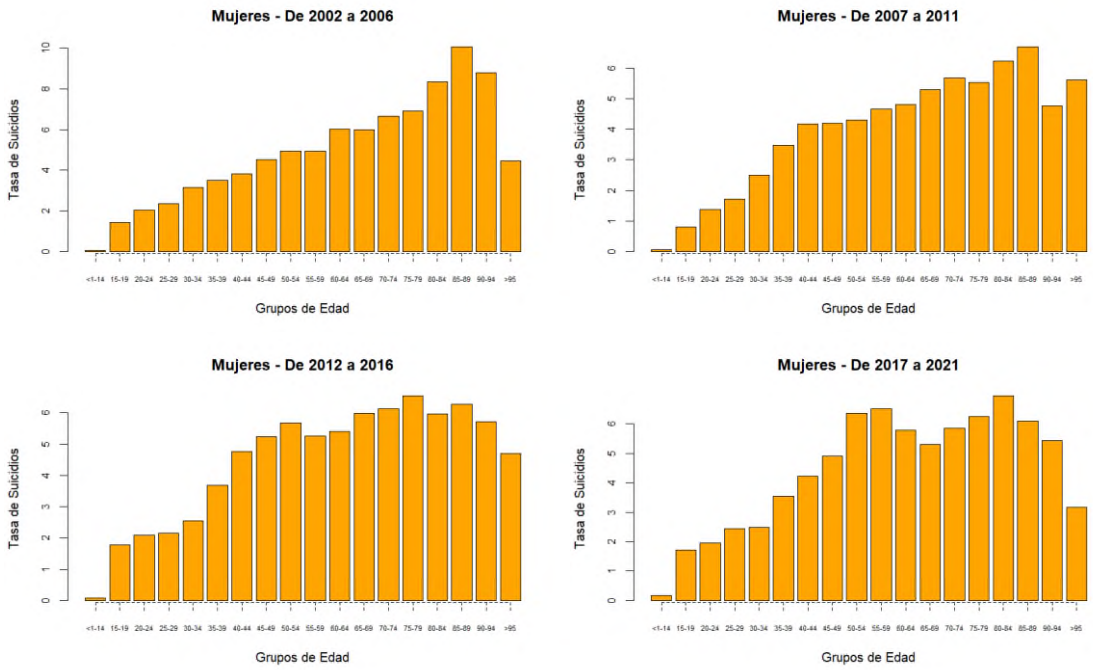


Figura 3: Tasa de suicidios en mujeres en cada quinquenio por grupos de edad.

Para ambos sexos, la menor tasa ocurre entre los años 2007 y 2011 en el primer grupo de edad, personas menores de 14 años a las que se les pueden atribuir pocos suicidios, por lo que tiene bastante sentido que sea la menor. Concretamente 12 hombres y 6 mujeres por cada 100.000. En las dos categorías el patrón de la bimodalidad es más claro conforme cambia el quinquenio.
La prevención del suicidio sigue siendo un desafío para la salud pública, y se requieren medidas de prevención, detección temprana y tratamiento adecuado para las personas que componen la población de riesgo. Conforme avanza el tiempo el suicidio se dispara hacia arriba, por lo que se puede asumir que una edad lejana tiene un componente fuerte en la distribución que se ajuste a los datos de este estudio. Esto puede estar relacionado con un factor social, es decir, cómo se construye la vida de las personas, los hábitos que mantengan, las dificultades que hayan experimentado a lo largo de su vida, etc. Se espera encontrar patrones y factores de riesgo asociados que permitan plantear distintas estrategias y tratamientos para prevenir que más personas cometan un primer intento de suicidio de manera efectiva.

5. MEZCLA DE DISTRIBUCIONES DE LA EDAD Y OBTENCIÓN DEL NÚMERO DE COMPONENTES

Como ya se mencionó, el número de suicidios es una variable de recuento. Estas variables determinan la cantidad de eventos ocurridos en una unidad de tiempo definido, que en este caso son los suicidios en un grupo de edad dentro del quinquenio que se esté observando. Son datos discretos y no negativos. En temas de salud pública son frecuentes estas variables, como el número de visitas al médico, número de días de estancia hospitalaria o número de fármacos prescritos. No obstante, aunque en términos generales este tipo de datos se modelaría por una distribución de Poisson, el patrón bimodal de los datos en las gráficas de barras lleva a pensar que es posible que se tenga que tratar con modelos de mezcla, concretamente distribuciones normales, las cuales son aproximaciones de los modelos de Poisson. Para ello se trabajará con el programa estadístico R, usando distintos paquetes para poder ajustar la distribución necesaria. En el *software* se ajusta primero una distribución normal de una componente a la variable edad con la función *fitdistr* del paquete *MASS* (Ripley, 2023) con el objetivo de compararlas con los modelos de mezcla creados después. Al ser la edad una variable de tipo intervalo, se ha calculado la marca de clase para cada uno de ellos, pudiendo así evitar que el programa de error. Entre los valores de los modelos resultantes se obtiene estimador máximo-verosímil con el cual se calculará manualmente el valor del criterio de información BIC correspondiente.
Tras ello, se carga el siguiente paquete *mixtools* con el que se creará un bucle usando la función *normalmixEM* para cada subgrupo para obtener todos los posibles modelos de mezcla de distribuciones normales (Young, 2022). Es lógico que, si el objetivo es modelizar la distribución de la edad, si se termina

con un modelo de 18 componentes, que son el número de grupos de edad, es probable que se esté sobre ajustando los datos lo que llevaría a una redundancia ya que puede haber grupos con patrones de comportamiento muy cercanos entre sí que no aporten información adicional. En esta ocasión solo se ha podido programar que las mezclas de distribuciones tengan como máximo 6, ya que la misma función avisaba de que el modelo no era convergente si se seguía añadiendo componentes. Esta función también ofrece el estimador máximo-verosímil con el que se volverán a calcular los criterios de información.
La construcción del modelo de mezcla de distribuciones atendiendo a los criterios numéricos del BIC, es una construcción que matemáticamente está cargada de razón. Sin embargo, el número de contrastes necesarios para llevarlo a cabo es tan elevado que provoca una inflación en el número de componentes a considerar. Este es un ejemplo claro del problema: por ejemplo, en la distribución del número de suicidios según la edad de hombres entre 2002 y 2006, se aprecia una fuerte bimodalidad y, sin embargo, el modelo de seis componentes tiene menor BIC. Para elegir entre varios modelos, se prefiere el que tenga menor BIC. Esta medida crece con el aumento de la variación de la variable dependiente y del número de variables explicativas. Por tanto, un BIC más bajo implica menos variables explicativas, mejor ajuste o ambos (Kass & Raftery, 1995). En las tablas siguientes (Tabla 2 y 3) se muestran los valores BIC de las mezclas de distribuciones de Poisson tanto para hombres como para mujeres. Se ve que no hay gran diferencia entre el BIC del modelo de dos componentes y los modelos superiores. Solo se encuentra una diferencia mencionable de dos unidades con el modelo simple de una distribución.

Tabla 2: Valores BIC de las mezclas de distribuciones de Poisson para hombres.

K	Hom.0206	Hom.0711	Hom.1216	Hom.1721
1	174.7263	174.7263	174.7263	174.7263
2	172.9334	172.2710	174.7256	172.2155
3	172.2891	172.2186	171.2964	171.3919
4	171.6963	171.4161	172.0876	171.4088
5	171.4192	171.5794	172.9107	171.1849
6	171.0784	170.7760	171.6682	172.1691

Tabla 3: Valores BIC de las mezclas de distribuciones de Poisson para mujeres.

K	Muj.0206	Muj.0711	Muj.1216	Muj.1721
1	174,7263	174,7263	174,7263	174,7263
2	172,5895	172,3700	173,9798	172,3750
3	172,9248	172,2280	172,2113	172,5153
4	172,4758	172,3088	172,2887	171,8505
5	171,9475	171,2380	171,3186	171,1080
6	171,7966	171,0596	170,8105	170,6731

Así se puede comprobar que los modelos tienen una bondad muy parecida, ya solo difieren en un par de unidades o incluso en décimas. Además, como el número de iteraciones crece a medida que se aumenta el número de componentes, el tiempo de ejecución del programa se dispara. En paralelo a esta cuestión, se tiene el factor económico porque si el objetivo de esta investigación es realizar terapias especializadas para cada división resultante de la población, un modelo muy complejo encarecería el proyecto. Por todo lo dicho anteriormente, se lleva a pensar que un modelo con tan solo dos componentes es suficientemente adecuado para acercarse a los datos observados. Se puede concluir, apoyándose en los resultados, que esta sea la distribución que se buscaba. Se exponen los parámetros de los modelos en las siguientes tablas.

- Parámetro μ

Al tener un ajuste de dos componentes, hay dos valores para μ. Representan el promedio de cada componente de la mezcla, determinando así la ubicación central de cada distribución.

Tabla 4: Valores μ para hombres.

μ_i	Hom0206	Hom0711	Hom1216	Hom1721
μ_1	22,12	35,76	53,96	42,20
μ_2	64,63	79,23	59,76	84,88

Tabla 5: Valores μ para mujeres.

μ_i	Muj0206	Muj0711	Muj1216	Muj1721
μ_1	26,99	31,92	49,22	31,77
μ_2	70,38	75,54	76,23	75,39

- Parámetro λ

De la misma manera se tienen dos valores distintos para cada λ. Son valores no negativos y deben sumar uno, es decir, para cada quinquenio $\lambda_1 + \lambda_2 = 1$. Representa la probabilidad de que un grupo de edad pertenezca al primer componente de la mezcla o al segundo.

Tabla 6: Valores λ para hombres.

$\boldsymbol{\lambda_i}$	**Hom0206**	**Hom0711**	**Hom1216**	**Hom1721**
$\boldsymbol{\lambda_1}$	0,250	0,576	0,954	0,718
$\boldsymbol{\lambda_2}$	0,750	0,424	0,046	0,282

Tabla 7: Valores λ para mujeres.

$\boldsymbol{\lambda_i}$	**Muj0206**	**Muj0711**	**Muj1216**	**Muj1721**
$\boldsymbol{\lambda_1}$	0,376	0,491	0,808	0,487
$\boldsymbol{\lambda_2}$	0,624	0,509	0,192	0,513

- Parámetro σ

Como tercer parámetro, σ representa la dispersión o la desviación estándar de cada componente. Indica la variabilidad de las observaciones pertenecientes a cada distribución.

Tabla 8: Valores σ para hombres.

$\boldsymbol{\sigma_i}$	**Hom0206**	**Hom0711**	**Hom1216**	**Hom1721**
$\boldsymbol{\sigma_1}$	10,346	17,153	26,428	20,623
$\boldsymbol{\sigma_2}$	21,052	12,569	25,521	9,177

Tabla 9: Valores σ para mujeres.

$\boldsymbol{\sigma_i}$	**Muj0206**	**Muj0711**	**Muj1216**	**Muj1721**
$\boldsymbol{\sigma_1}$	12,718	15,138	25,369	15,063
$\boldsymbol{\sigma_2}$	17,692	14,684	15,509	14,769

Por ejemplo, para los hombres entre los años 2002 y 2006, se ha obtenido un modelo de mezcla con parámetro de medias 22,12 y 64,63. Esto implica que el primer grupo está compuesto principalmente por hombres jóvenes con una concentración alrededor de los 22 años, mientras que el otro grupo está formado por hombres mayores alrededor de los 64 años. La proporción relativa de cada componente en la mezcla y la dispersión de las edades en cada grupo están determinadas por los parámetros λ y σ. En este caso, sugieren que el primer componente — de media de edad en 22 años — representa aproximadamente el 25% de la distribución total, mientras que el segundo componente — de media 64 años — representa aproximadamente el 75%. Los valores de la desviación estándar indican la dispersión de los datos en cada distribución. Un valor más bajo sugiere una concentración más estrecha alrededor de la media, mientras que si es más alto indica una mayor dispersión de edades. Se puede visualizar el efecto de un valor de σ más o menos elevado si se compara con el ajuste realizado a los hombres de los siguientes cinco años. Este grupo tiene como parámetros un vector de medias (35,76, 79,23) con unas probabilidades de ocurrencia del 57.6% y 42.4%. Las distribuciones que conforman este ajuste tienen desviaciones estándar de 17,153 y 12,569 años. Estos valores están más cercanos entre ellos que las desviaciones del primer quinquenio, que eran $\sigma_1 = 10.346$ y $\sigma_2 = 21.052$, lo cual se ve reflejado en la diferencia del área que ocupan las distribuciones en las gráficas correspondientes. Hay mayor diferencia entre el tamaño del área de los componentes del primer quinquenio que en las del segundo, que implica, como se reflejaba numéricamente, que hay mayor variabilidad entre los subgrupos resultantes en los años 2002 y 2006 que en los resultantes de 2007 y 2011. Entre los años 2017 y 2021 ocurre lo mismo. La primera componente con una media alrededor de los 42 años, con una desviación de 20,623 y una probabilidad de 71,8% se distingue claramente de la segunda situada alrededor de los 85 años, con una desviación de 9,177 y una probabilidad de 28,2%. La primera curva de la gráfica ocupará más espacio tanto de manera vertical, por tener mayor probabilidad de ocurrencia, como de manera horizontal por tener mayor variabilidad en los datos.

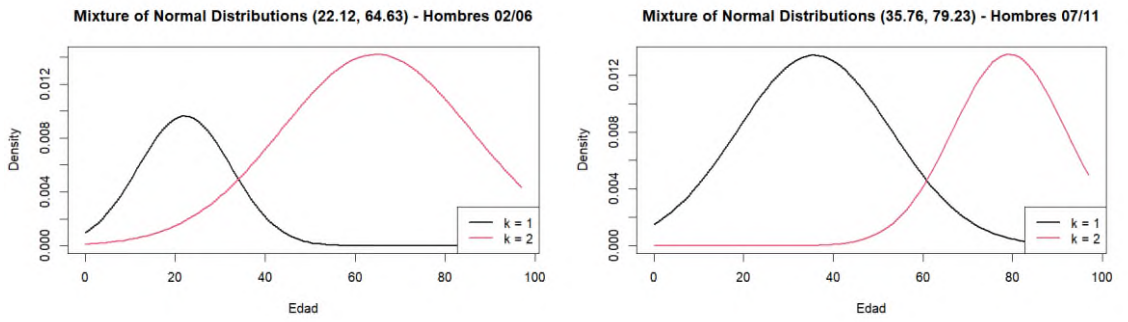


Figura 4: Mezcla de distribuciones normales para hombres, periodo 2002-2006 y 2007-2011.
Un caso interesante ocurre entre los años 2012 y 2016 para los hombres. Tenemos las medias $\mu = (53.96, 59.76)$ muy cercanas entre sí, unas probabilidades de ocurrencia $\lambda = (0.046, 0.954)$ y desviaciones $\sigma = (26.428, 25{,}521)$. Al tener la primera componente una probabilidad muy baja, un 4,6%, y una variación muy parecida a la segunda, se ve reflejado de manera que la primera distribución quede dentro de la segunda. Se puede interpretar como dos subpoblaciones superpuestas prácticamente iguales.

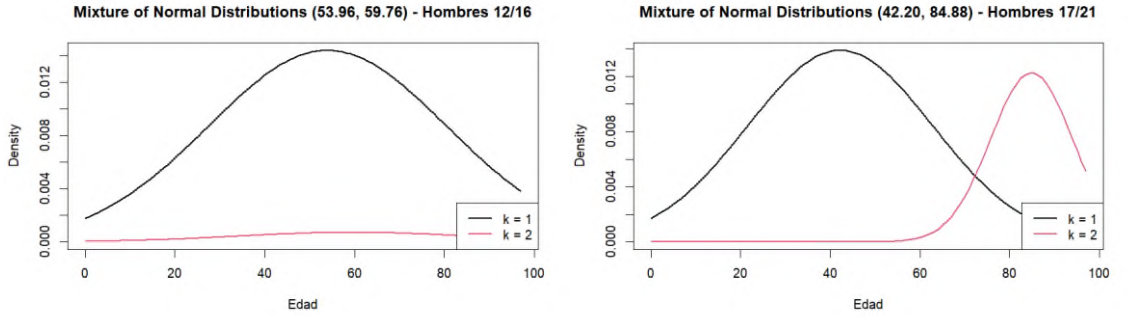
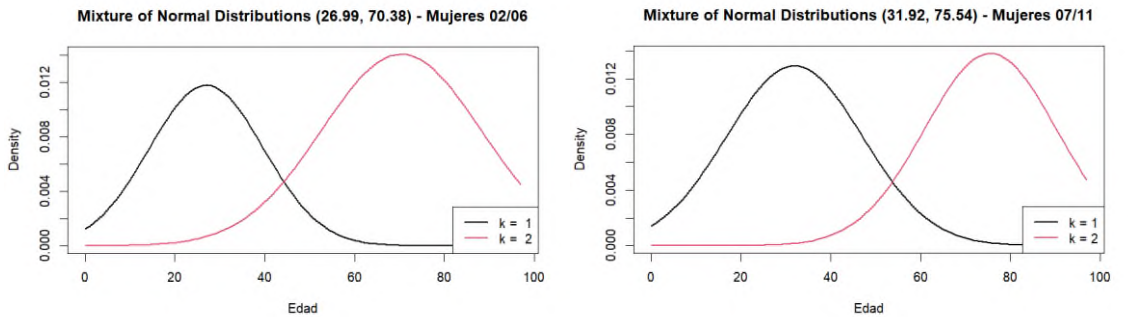


Figura 5: Mezcla de distribuciones normales para hombres, periodo 2012-2016 y 2017-2021 .
Las gráficas de las mujeres tienen la misma explicación que la de los hombres. Sin embargo, la diferencia visual entre los componentes para cada quinquenio no es tan drástica como en los hombres, lo cual se venía arrastrando desde antes, donde la bimodalidad se veía muy difuminada en las gráficas de barras. Entre los primeros años, de 2002 a 2006, es cuando más se puede apreciar dos modas distintas. Teniendo una concentración de los datos alrededor de los 27 y 70 años, la diferencia del área que ocupan las distribuciones viene dada por las probabilidades de pertenencia a ellas del 37,6% y del 62,4% junto con las varianzas, algo más cercanas, con valores 12,718 y 17,692 respectivamente. Al igual que los hombres, el caso interesante ocurre entre 2012 y 2016. Vuelve a ocurrir que la primera componente se fusiona con la segunda. A pesar de tener las medias (49,22 y 76,23) y las desviaciones (15,509 y 25,369) más alejadas entre ellas, la probabilidad de que la población pertenezca a la primera componente es tan solo de 0,192, lo que resulta en que una quede dentro de la otra.



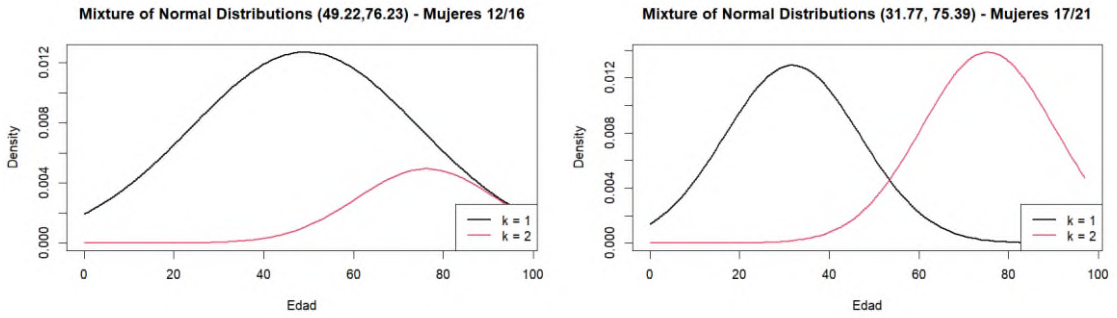


Figura 6: Mezcla de distribuciones normales para mujeres en los 4 periodos estudiados.

Por tanto, se puede ver que en gran parte las medias y desviaciones típicas son muy distintas entre sí. La interpretación que se puede dar a esta cuestión es que la población se pueda dividir en dos subpoblaciones distintas según la edad en la que se suiciden. Se puede conjeturar que los individuos jóvenes se deciden suicidar por motivos diferentes a los de los individuos pertenecientes al grupo de mayor edad, ya que al ser grupos o subpoblaciones distintas, no tienen comportamientos similares o pueden verse afectados por diversos factores que los divida.

6. CONCLUSIONES

Este trabajo partía con el objetivo de estudiar la distribución de la edad en la que las personas se han suicidado en España entre los años 2002 y 2021. Gracias al análisis descriptivo que se ha realizado sobre los datos, se ha podido ver un aumento en la mortalidad por esta causa. A pesar de que en hombres se observa un crecimiento más evidente, en ambos sexos se distingue claramente cómo se disparan las tasas alrededor de ciertos grupos de edad. Se pudo que la distribución de la edad presenta una bimodalidad, lo cual dio una idea de por qué se iba a necesitar un modelo de mezcla de distribuciones.

Para ello, se abordó de manera teórica las características de los modelos de mezcla y la aparición de varias modas en la distribución de la edad en distintas áreas de la medicina. Se justificó así el uso de los modelos de mezcla para ajustar de manera más precisa la heterogeneidad de los datos. El siguiente objetivo fue ajustar una mezcla de componentes normales, comparando la bondad entre los modelos con distinto número de componentes. Tras haber estudiado la diferencia entre los criterios de información bayesianos, y por cuestión de eficiencia, la elección del modelo terminó en un ajuste de dos componentes. Para finalizar, gracias a la aplicación de los algoritmos EM y máxima verosimilitud, se obtienen los valores de los parámetros de los componentes que forman parte del modelo. Se ha podido ver que la bimodalidad en la distribución de la edad es más patente en los hombres. Se destacan los años entre 2012 y 2016 al quedar la primera componente prácticamente oculta en la segunda. La probabilidad de pertenecer a la primera distribución es muy baja y las desviaciones estándar son tan parecidas que se podría llegar a pensar que solo se necesite una componente en ese quinquenio.

Un estudio de este estilo se puede aplicar muchos campos. Debido a su magnitud, solo se ha realizado el análisis explicativo de este ajuste estadístico. No obstante, sería interesante seguir con esta cuestión en un futuro, pudiendo así prevenir suicidios y estableciendo terapias especializadas en base a las distintas subpoblaciones resultantes.

RECEIVED: NOVEMBER 2023.
REVISED: MARCH, 2024.

REFERENCIAS

AMAT RODRIGO, J. (2020). Estadistica-con-R. **https://github.com/JoaquinAmatRodrigo/Estadistica-con-R/blob/master/PDF_format/55_ajuste_distribuciones_con_r.pdf**. Consultado 2, 5, 2023.

BAYES, M., & PRICE, S. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. **Philosophical Transactions of the Royal Society of London**, *53*, 370-418.

DEB, P. (2010). Finite Mixture Models with Applications [Diapositivas]. **Hunter College, New York, Estados Unidos. Presentación realizada en septiembre de 2010.**

EVERITT, B. (1996). An introduction to finite mixture distributions. **Statistical Methods in Medical Research**, *5*, 107-127.

GONG, H., PA, L., WANG, K., MU, H., DONG, F., YA, S., XU, G., TAO, N., PAN, L., WANG, B., HUANG, S., & SHAN, G. (2017). Bimodal distribution of fasting plasma glucose in the Uyghur and Han populations of Xinjiang, China. **Asia Pacific Journal of Clinical Nutrition**, *26,* 708–712.

INE - Instituto Nacional de Estadística. (2022a). Defunciones por causas (lista reducida) por sexo y grupos de edad (7947**). https://www.ine.es/jaxiT3/Tabla.htm?t=7947** Consultado 30, 11, 2022.
INE - Instituto Nacional de Estadística. (2022b). Población residente por fecha, sexo y edad (10256). **https://www.ine.es/jaxiT3/Tabla.htm?t=10256**. Consultado 30, 11, 2022.
JABEEN, H. (2023). Tutorial: Poisson Regression in R. ***Dataquest*. https://www.dataquest.io/blog/ tutorial-poisson-regression-in-r/.** Consultado 3, 12, 2022.
JACOBS, M. (2022). Mixture, Component, Zero-Inflated, and Hurdle models. *Medium*. **https://blog.devgenius.io/mixture-component-zero-inflated-and-hurdle-models-44c5e6fe5d7f.** Consultado 15, 1, 2023.
KASS, R. E., & RAFTERY, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90(*430*)*, 773-795.
LOUIS, E. D., & DOGU, O. (2007). Does age of onset in essential tremor have a bimodal distribution? Data from a tertiary referral setting and a population-based study. **Neuroepidemiology***, 29,* 208–212.
McLACHLAN, G., & PEEL, D. (2000). *Finite Mixture Models*. Wiley-Interscience.
MENON, V., KATTIMANI, S., SARKAR, S., SATHYANARAYANAN, G., SUBRAMANIAN, K., & VELUSAMY, S. K. (2018). Age at onset of first suicide attempt: Exploring the utility of a potential candidate variable to subgroup attempters. **Asian Journal of Psychiatry***, 37,* 40–45.
MINISTERIO DE SANIDAD - GABINETE DE PRENSA - NOTAS DE PRENSA. (2020). **https://www.sanidad.gob.es/gabinete/notasPrensa.do?id=5006**. Consultado 5, 5, 2023.
MOBBS, B. G., CHAPMAN, J. A., SUTHERLAND, D. J., RYAN, E., TUSTANOFF, E. R., OOI, T. C., & MURTHY, P. V. (1993). Evidence for bimodal distribution of breast carcinoma ER and PgR values quantitated by enzyme immunoassay. **European Journal of Cancer** *(Oxford, England: 1990), 29A*(9), 1293–1297.
RIPLEY, B. (2023). MASS package - RDocumentation. **https://www.rdocumentation.org/packages/ MASS/versions/ 7.3-58.3.** Consultado 4, 12, 2022.
SALINAS-RODRIGUEZ, A., MANRIQUE-ESPINOZA, B. & SOSA-RUBI, S.G. (2009). Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. **Salud pública Méx**. 51, 397-406.
SCHLATTMANN, P. (2009). **Medical Applications of Finite Mixture Models**. Springer Science & Business Media, Berlin.
YOUNG, D.S. (2022). mixtools package - RDocumentation. **https://www.rdocumentation.org/packages/ mixtools/versions**/2.0.0 Consultado 5, 3, 2023.