Mixture distributions models: Application to multimodal data

Jornadas Andaluzas de Bioinformática
Granada (España), 24 – 26 June 2024

Andrea Ruiz Vega, Miguel Á. Montero Alonso, Juan de Dios Luna del Castillo

1. Statistics and O.I., Faculty of Education, Economics and Technology of Ceuta. University of Granada.. arvega@ugr.es
2. Statistics and O.I., Faculty of Medicine. University of Granada. mmontero@ugr.es
3. Statistics and O.I., Faculty of Medicine. University of Granada. jdluna@ugr.es

Introduction

Mixture of distributions models are a weighted combination of simpler distributions that help to analyze the heterogeneity of data sets with multimodality. The respective parameters will be estimated using the EM algorithm and the maximum likelihood method. A mixture of two components of normal distributions will be fitted. This result is supported by the goodness-of-fit study of the model together with the estimation of the model parameters. The R mixtools library is used, applied to a data set between the years 2002 and 2021 where the distribution of the variable of interest presents several peaks where most of the observations are collected. The objective is to identify subgroups that present specific behaviors or characteristics in order to obtain meaningful information and better understand the characteristics and trends within the studied population and the resulting subgroups.

Methodology

- Estimate models with 1 and more components (K = 1, 2, 3, 4).
- Calculate the Bayesian information criterion (BIC) in each model.
- Choose the model with the lowest BIC and/or the most efficient one.
- Estimation of the model parameters.
- Acceptance of the model.

Δ BIC	Rejection of the highest BIC
0 a 2	Simple mention
2 a 6	Positive
6 a 10	Strong
> 10	Very strong

Selection of the model

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

Bayesian Information Criteria - Men

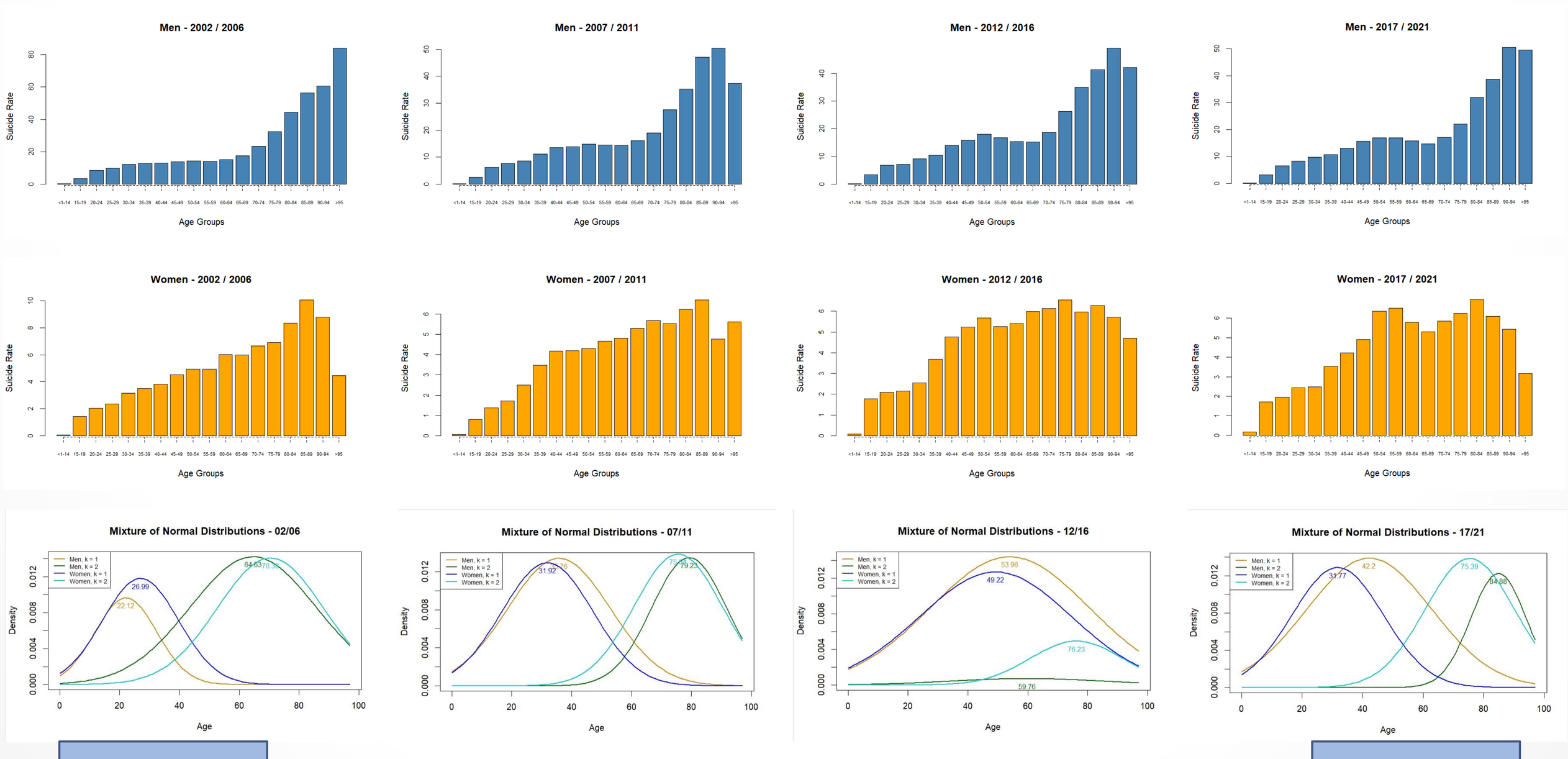
k	Muj.0206	Muj.0711	Muj.1216	Muj.1721
1	174.7263	174.7263	174.7263	174.7263
2	172.5895	172.3700	173.9798	172.3750
3	172.9248	172.2280	172.2113	172.5153
4	172.4758	172.3088	172.2887	171.8505

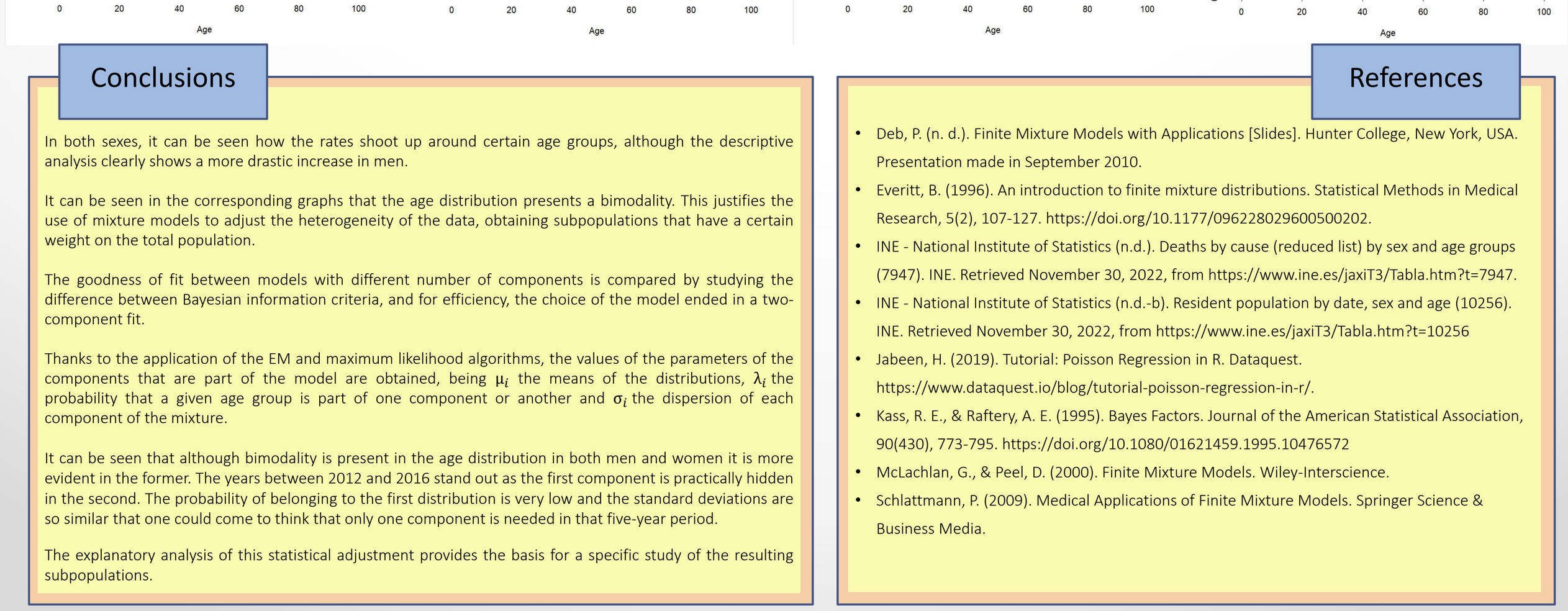
Bayesian Information Criteria - Women

Age groups have different weights in the distribution to be adjusted, since they are made up of different populations. Rates allow us to compare between different populations or groups that have different population sizes, i.e., that are not homogeneous.

There are several reasons why it is preferable to model rate data rather than direct count data:

- Normalization of exposure size. Allows comparison of groups of different sizes.
- Tendency to reduce variance. Useful when counts are low.
- Likelihood of meeting assumptions of normality of errors. Errors are assumed to follow a normal distribution.





Conclusions

In both sexes, it can be seen how the rates shoot up around certain age groups, although the descriptive analysis clearly shows a more drastic increase in men.

It can be seen in the corresponding graphs that the age distribution presents a bimodality. This justifies the use of mixture models to adjust the heterogeneity of the data, obtaining subpopulations that have a certain weight on the total population.

The goodness of fit between models with different number of components is compared by studying the difference between Bayesian information criteria, and for efficiency, the choice of the model ended in a two-component fit.

Thanks to the application of the EM and maximum likelihood algorithms, the values of the parameters of the components that are part of the model are obtained, being μ_i the means of the distributions, λ_i the probability that a given age group is part of one component or another and σ_i the dispersion of each component of the mixture.

It can be seen that although bimodality is present in the age distribution in both men and women it is more evident in the former. The years between 2012 and 2016 stand out as the first component is practically hidden in the second. The probability of belonging to the first distribution is very low and the standard deviations are so similar that one could come to think that only one component is needed in that five-year period.

The explanatory analysis of this statistical adjustment provides the basis for a specific study of the resulting subpopulations.

References

- Deb, P. (n. d.). Finite Mixture Models with Applications [Slides]. Hunter College, New York, USA. Presentation made in September 2010.
- Everitt, B. (1996). An introduction to finite mixture distributions. Statistical Methods in Medical Research, 5(2), 107-127. https://doi.org/10.1177/096228029600500202.
- INE - National Institute of Statistics (n.d.). Deaths by cause (reduced list) by sex and age groups (7947). INE. Retrieved November 30, 2022, from https://www.ine.es/jaxiT3/Tabla.htm?t=7947.
- INE - National Institute of Statistics (n.d.-b). Resident population by date, sex and age (10256). INE. Retrieved November 30, 2022, from https://www.ine.es/jaxiT3/Tabla.htm?t=10256
- Jabeen, H. (2019). Tutorial: Poisson Regression in R. Dataquest. https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/.
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. Journal of the American Statistical Association, 90(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572
- McLachlan, G., & Peel, D. (2000). Finite Mixture Models. Wiley-Interscience.
- Schlattmann, P. (2009). Medical Applications of Finite Mixture Models. Springer Science & Business Media.