Journal of Statistical Research *https://doi.org/10.3329/jsr.v59i2.88068*
2025, Vol. 59, No. 2, pp. 249-257 *ISSN 0256 - 422 X*

SEQUENTIAL BAYESIAN APPROACH TO ESTIMATE THE PROPORTION OF PEOPLE IN RETIREMENT CONDITIONS IN ARGENTINA

MELINA GUARDIOLA
Instituto de Matemática de Bahía Blanca (INMABB), Departamento de Matemática
Universidad Nacional del Sur (UNS) - CONICET, Bahía Blanca, Argentina
Email: guardiol@uns.edu.ar

FERNANDA VILLARREAL*
Instituto de Matemática de Bahía Blanca (INMABB), Departamento de Matemática
Universidad Nacional del Sur (UNS) - CONICET, Bahía Blanca, Argentina
Email: fvillarreal@uns.edu.ar

MILVA GERI
Instituto de Investigaciones en Ciencias de la Salud
Departamento de Ciencias de la Salud (UNS-CONICET), Departamento de Matemática
Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina
Email: milva.geri@uns.edu.ar

SUMMARY

One of the dimensions of a pension system's performance is coverage. In most countries, pension systems require a minimum number of years of contribution as part of their eligibility requirements. For example, in Argentina, a minimum of 30 years of contributions is required. Therefore, it is not enough to study what percentage of the target population is covered at a given time; it is also necessary to study employment histories over a considerable period of time. Sometimes, developing countries do not have sufficient information for this purpose, but they incorporate new information as they digitize their records. The Bayesian approach can be useful in these cases where information is limited but regularly updated. The objective of this paper is to demonstrate the usefulness of the sequential Bayesian approach for estimating the proportion of workers eligible to retire in Argentina year after year. It is observed that as more information is incorporated, the proportion of people who remain active contributors (and, therefore, eligible for a pension) decreases. This implies that, for any individual, the probability of meeting the contribution requirement decreases as time passes from the first month of contribution. At the end of the process (once all available information has been incorporated), the Bayesian proportion estimate is different from that obtained with a frequentist approach, which is explained by the importance of a priori information provided by prior knowledge about the phenomenon. This type of sequential estimation exercise may be of interest to social security decision-makers.

Keywords and phrases: Bayesian approach, estimated proportion, pension system, Argentina

AMS Classification: 62F15

* Corresponding author

1 Introduction

According to the International Labour Organization (ILO), one of the dimensions of a pension system's performance is its coverage, (ILO, 2011), that is, what percentage of the target population is covered by the system. In this sense, the target population of a pension system is not only older people, but also people of working age, since in order to be entitled to a pension there are generally contributory requirements (having contributed for a minimum period of time). In Argentina, for example, to be entitled to a pension, 30 years of contributions are required, in addition to having reached a minimum age (60 years for women and 65 for men). Thus, it is not only important to evaluate coverage at a given point in time (what proportion of active people contribute to the pension system at time t), but also what is the contributory density to the pension system (what proportion of people have managed to accumulate sufficient contributions to be entitled to a pension). In this sense, although there are numerous studies that study the performance of the Argentine pension system in terms of its coverage (Arza, 2012, 2016; Bertranou and Bonari, 2005; Bertranou and Rofman, 2001; Bertranou and Casanova, 2011; Rofman and Ourens, 2009), few studies focus on contributory density (Rofman and Oliveri, 2012; Geri and Villarreal, 2025).

The objective of this work is then to estimate the proportion of people eligible to retire in Argentina using a Bayesian approach. This approach has the strength of not assuming repeated experiments as the classic frequentist approach assumes, (Johnson et al., 2022). This is interesting for most studies in social sciences where there are no repeated experiments under the same conditions, but rather there are events that are more or less plausible.

In turn, this approach admits that decision makers usually have at their disposal a considerable set of knowledge (a priori information), in addition to the specific observations that are analyzed at a given time. Such information is completely discarded under a classic frequentist approach, (Moore, 1966). In contrast, the Bayesian approach allows the probability of an event occurring to be estimated based on a priori knowledge about the phenomenon and current information (data set). In fact, some frequentist results can be seen as particular cases of Bayesian analysis when the distribution of prior information is non-informative, (Luque and Sosa, 2024). Finally, in studies such as these, the information (data set) is usually updated over time as new generations enter the labor market. Thus, the entity that administers the system is interested in knowing how the parameters evolve over time as the years go by, without losing past knowledge about the phenomenon. For this reason, it is useful to adopt a sequential approach that allows us to estimate what proportion of people will be able to retire next year, updating our a priori knowledge with new data about the effective contributors during the last years.

2 Data and Methods

2.1 Data

The Bayesian approach requires a priori knowledge and a set of data. As a priori knowledge, we will rely on previous studies. In this sense, Rofman and Oliveri (2012), claim that only 20% of

pension system members had sufficient regularity of contributions to be entitled to receive a pension in Argentina between 1994 and 2003.

The database from which we will extract information is the Longitudinal Sample of Registered Employment (MLER by its acronym in Spanish) which is updated periodically by the Ministry of Labor, Employment and Social Security (MTEySS by its acronym in Spanish). The first version of this database was published in 2018 and contained information on the work histories of about 500 thousand individuals between 1996 and 2015. In 2022, it was updated to extend the period and currently covers 1996-2021. The sample is representative of the population of pension system members who have the following characteristics: they are employees and belong to the private sector. This means that this data does not represent independent workers or public sector employees. In addition, the database only records people who have made at least one month of contributions, which is why it does not represent people who have worked their entire lives under informal conditions. The key variables of database are the gender, de age, the month and year of contribution and the contributor status. With this information, the variables of interest are constructed for each analysis window: $i)$ N: "Total number of contributors" and $ii)$ Y: "Number of contributors who contributed during at least 80% of the months that make up the window".

Since the objective would then be to estimate the proportion of workers (private sector employees) who will be able to retire next year and knowing that the database is updated with a certain periodicity, we work with successive windows that go from the moment t to moment t_v (to simulate what happens in reality as the information is updated). Each window adds an additional year of observation to the previous window: in the first window we observe only 1996, in the second 1996-1997 and so on until we end up observing the entire period 1996-2021. In this way, in all the windows t_0 corresponds to January 1996 and t_v is December of every year between 1996 and 2021 (26 windows).

To select the N individuals in each window, we filter out those workers who will reach the minimum retirement age in the year after the upper limit of each window (t_v) (men aged 65 and women aged 60) and we observe what proportion of them have sufficient regularity of contributions. In this sense, following Rofman and Oliveri (2012), we understand that sufficient regularity implies having contributed to the system for at least 80% of the months of the window. Thus, for example, having contributed at least 10 months is sufficient to have sufficient regularity of contributions during the first window, while in the last window at least 250 months are required. This will make the requirement of regularity of contributions more difficult to achieve as the successive windows go by (since each time a longer period is considered and it is more likely that people will suffer some interruption of work).

We focus on estimating the proportion of individuals reaching contribution regularity right before attaining retirement age, assuming that these individuals will become eligible in the subsequent year. In this sense, the pension system administrator is interested in knowing what the system's implicit debt will be year after year (the flow of new workers eligible for retirement).

2.2 Methods

We define Y_i as a binary random variable that takes the value 1 if worker i has sufficient regularity of contributions and 0 otherwise. Thus, $Y_i \sim Ber(\theta)$, with probability function

$$P(Y_i = y|\theta) = \theta^y(1-\theta)^{1-y}, \tag{2.1}$$

being $\theta = P(Y_i = 1)$.

Since the proportion θ will be subject to fluctuations that are explained by changes in the characteristics of individuals and also by events that affect the economic situation (public policies, national and international crises, commodity prices, etc.), we can consider it as a random variable that can take any value between 0 and 1.

The first step is to propose a distribution that is flexible enough to model our beliefs about θ, taking into account the a priori information and some credibility in it. One of the models that is often used in problems related to proportions is the $Beta$ distribution, which, like θ, is continuous and assumes values in the interval $[0, 1]$.

We then model the variability in $\theta \in [0, 1]$ with a $Beta$ distribution with parameters $\alpha > 0$ and $\beta > 0$:

$$\theta \sim Beta(\alpha, \beta),$$

whose probability density function (p.d.f.) is:

$$f(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I_{[0,1]}, \tag{2.2}$$

where $\Gamma(z)$ is the $Gamma$ function defined by $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$.

To solve for the parameters α and β of the prior $Beta$ distribution, we use the moment-matching method:

$$\mu = \frac{\alpha}{\alpha+\beta}, \tag{2.3}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{2.4}$$

where μ is the expected value and σ^2 the variance of the prior distribution for θ.

Given the expected value and variance derived from expert knowledge or previous studies, equations (2.3) and (2.4) can be solved simultaneusly to obtain values for α and β:

$$\alpha = \mu\left(\frac{\mu\,(1-\mu)}{\sigma^2} - 1\right) \quad \text{and} \quad \beta = (1-\mu)\left(\frac{\mu\,(1-\mu)}{\sigma^2} - 1\right). \tag{2.5}$$

The p.d.f. of $Beta(\alpha, \beta)$ takes different forms depending on the values of its parameters α and β. Thus, for example, the uniform distribution is a particular case of the beta distribution with $\alpha = \beta = 1$. Furthermore, when $\alpha = \beta$, the distribution is symmetric around 0.5 and the larger α and β are, the more the probability density will be concentrated around a point. Finally, whenever $\alpha < \beta$ the highest probability density is concentrated on the left side and the opposite occurs when

$\beta < \alpha$. Thus, the values we assume for α and β must be based on beliefs about the behavior of θ. For example, if we believe few people have regular contributions, $\alpha < \beta$, and if we believe most do, $\alpha > \beta$. A symmetric prior ($\alpha = \beta$) reflects uncertainty or neutrality.

Second, the problem data are used to collect new information about θ. When modeling the dependence of Y given a value of θ that we assume to be constant, we assume that the observations are independent, then

$$Y|\theta \sim Bin(N, \theta),$$

with conditional probability distribution function $f(y|\theta)$ defined for $y \in \{0, 1, \ldots, N\}$

$$f(y|\theta) = P(\text{Y=y}|\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y}. \tag{2.6}$$

This function allows us to answer the following question: given our belief about θ, how many of the N formal workers in the sample are expected to have sufficient regularity of contributions?

Looking at the data $Y = y$, where $y \in \{0, 1, \ldots, N\}$, the likelihood function of θ, $L(\theta|y)$ is then obtained by replacing the value of y into the binomial distribution function given by equation (2.6), thus obtaining a mechanism for determining how likely a given value of θ is, given our data.

The final step is to update our beliefs about θ by applying Bayes' theorem as follows:

$$f(\theta|y) = \frac{f(\theta)L(\theta|y)}{f(y)} \propto f(\theta)L(\theta|y), \tag{2.7}$$

where $f(y)$ is the normalization constant that does not depend on θ. Note that the probability of an unknown parameter θ assuming a certain value if the data (y) are given is proportional to the product of the initial probability of θ by the probability of θ given y. Proportionality here means proportionality when considering y as fixed; if a different data y were observed, there would naturally be a different proportionality constant, (Moore, 1966).

Thus, since the pdf is proportional to the product of the prior function and the likelihood function, it is not necessary to calculate $f(y)$. Then, replacing in equation (2.7) the expression for $f(\theta)$ and $L(\theta|y)$ given in equations (2.2) and (2.6), respectively

$$f(\theta)L(\theta|y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{N}{y}\theta^y(1-\theta)^{N-y}.$$

Thus, $f(\theta|y) \propto \theta^{(\alpha+y)-1}(1-\theta)^{[\beta+(N-y)]-1}$ is represented by its structural dependence on θ; that is, its kernel. In this sense, the posterior distribution can be described by a $Beta(\alpha+y, \beta+(N-y))$ model that reveals the influence of the prior function (α and β) and the data (N and y). In addition, the expectation and variance are obtained as

$$E(\theta|\text{Y=y}) = \frac{\alpha+y}{\alpha+\beta+N},$$
$$Var(\theta|Y=y) = \frac{(\alpha+y)(\beta+N-y)}{(\alpha+\beta+N)^2(\alpha+\beta+N+1)}.$$

And thus the Bayesian $Beta - Binomial$ model is defined for the proportion $\theta \in [0, 1]$:

$$Y|\theta \sim Bin(N, \theta),\ \theta \sim Beta(\alpha, \beta),\ \ \theta|Y = y \sim Beta(\alpha + y, \beta + (N - y)).$$

This model is applied sequentially as many times as there are observation windows. In the first window, a priori information provided by previous studies is used; in the second window, the posterior distribution obtained in the first window is used as a priori information, and so on until the 26 windows are complete. For clarity, Figure 1 includes the following flow chart illustrating the process:

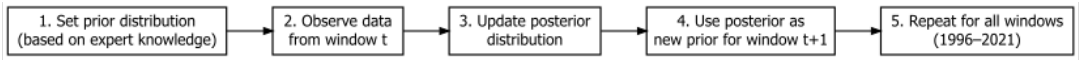

Figure 1: Flow chart of the Bayesian sequential approach

Finally, credibility intervals can be constructed in each window. A credibility interval provides a range of plausible posterior values of θ and, therefore, a summary of both the posterior central tendency and the variability. For example, a 95% credibility interval for θ is constructed using the 2.5th and 97.5th percentiles of the posterior distribution. Thus, there is a 95% posterior probability that θ lies in the interval found, i.e.

$$P(\theta \in (\theta_{0.025}, \theta_{0.975}|Y = y) = \int_{0.025}^{0.975} f(\theta|y)d\theta = 0.95.$$

3 Results

We can assume that the expected value of θ will take a value close to 0.2; that is, $E(\theta) = \frac{\alpha}{\alpha+\beta} \approx 0.2$, which is equivalent to $\alpha \approx \frac{1}{4}\beta$. Regarding the variance of θ, Arroyo Bravo et al. (2022), analytically demonstrate that for the parameters α and β to belong to the interval $(0, \infty)$, the following must be fulfilled:

$$\frac{\mu(1-\mu)}{\sigma^2} - 1 > 0, \sigma^2 < \mu(1-\mu).$$

So, for α and β to be positive, μ can take any value within the interval $[0, 1]$. However, the parameter σ^2 cannot take any value in that interval, since, if the imposed condition is not met, α and β could have negative values. This leads to the conclusion that $\sigma^2 \in [0, \mu(1-\mu)]$ and then σ^2 depends on the values that μ can take.

In this case, a standard deviation of 0.05 is chosen; that is,

$$SD(\theta) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \approx 0.05.$$

With these assumptions, we conclude that a reasonable a priori model to model θ in the first window is $\theta \sim Beta$(12.6,50.4), with p.d.f

$$f(\theta) = \frac{\Gamma(63)}{\Gamma(12.6)\Gamma(50.4)}\theta^{12.6-1}(1-\theta)^{50.4-1}I_{[0,1]}.$$

The posterior distribution resulting from the first window functions as the prior distribution for the second window, and so on. Figure 2 shows the three distributions in the first four windows. It can be observed that as the windows become larger (they cover longer periods), the posterior distribution becomes closer to the prior distribution because the information provided by the data begins to have greater relative weight.

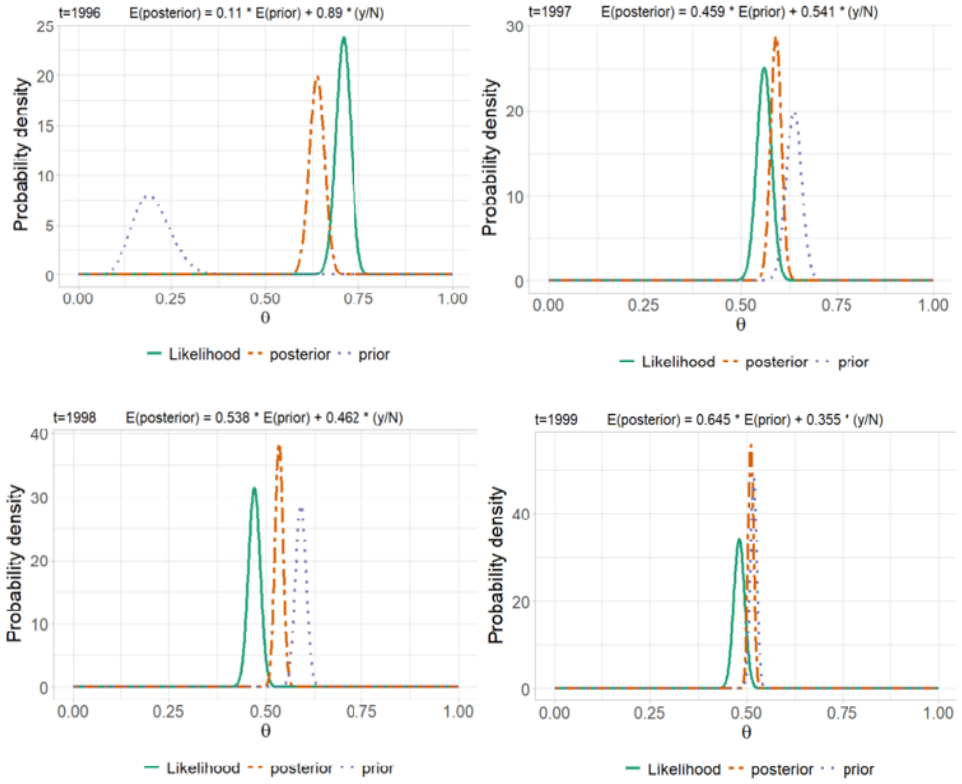

Figure 2: Beta-Binomial Model (first four windows)

As stated before, in each of these windows it is possible to construct credibility intervals. For example, in the first window the credibility interval is [0.60, 0.67], indicating that the proportion of private sector workers eligible for retirement would be between 0.60 and 0.67 if we only had information from the first window. The sequential analysis allows us to update our knowledge sequentially until we obtain the credibility interval for the last of our windows: [0.23, 0.24]. As can be seen, at the end of the process we obtain more precise intervals that, in our case, are more similar to the a priori information coming from the knowledge of experts on this phenomenon.

Figure 3 shows the evolution of the subsequent estimate of the mean at the end of each window. Although the proportion estimated a posteriori using the Bayesian approach is approaching the proportion estimated using a frequentist approach as the data set grows, it is important to mention that

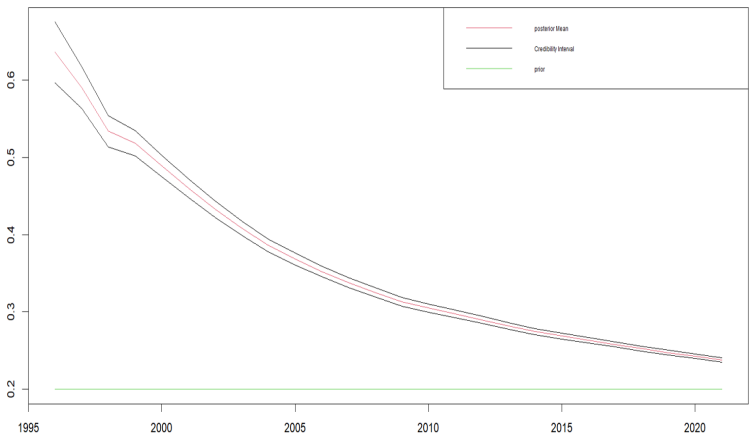

Figure 3: Posterior average proportion of workers eligible for retirement and its credibility interval at each t

both estimates differ at the end of the window (the frequentist estimate would be 7 percentage points lower than the Bayesian one). This means that, although the a priori information is weighing less and less in the estimate, it still has some influence.

4 Conclusions

The Bayesian approach is more appropriate for some social science problems where there are no events that result from repeated experiments, but simply events that are more or less plausible.

In this paper, a sequential Bayesian approach was adopted to estimate the proportion of private sector workers eligible to retire next year in Argentina, given a priori knowledge and a set of updated data on the phenomenon. It was found that as the information is updated and accumulated (increasingly wider observation windows), the posteriori proportion gets closer to the a priori information from expert knowledge. In turn, the credibility intervals become increasingly precise. In addition, although the a priori information loses weight in the estimate as the information from data increases, it still has an impact to some extent, yielding values different from those that would be obtained with the frequentist approach. These data analysis tools would be extremely useful for decision-makers in the field of social security, as they would allow knowledge about the phenomenon to be updated without losing past knowledge.

References

Arroyo Bravo, L. G., Lasso Balanta, F. A., and Tovar Cuevas, J. R. (2022), "Propuesta para obtener distribuciones previas para los parámetros de la distribución beta," *Investigación Operacional*, 43,

51–63.

Arza, C. (2012), "Extending coverage under de Argentinian pension system: distribution of Access and prospects for universal coverage," *International Social Security Review*, 65, 29–49.

— (2016), "Non-contributory benefits, pension re-reforms and the social protection of older women in Latin America," *Social Policy and Society*, 16, 361–375.

Bertranou, F.; Grushka, C. and Rofman, R. (2001), "Evolución reciente de la cobertura previsional argentina, en: Cobertura previsional en Argentina, Brasil y Chile," *Organización Internacional del Trabajo, Chile*.

Bertranou, F. and Bonari, D, C. (2005), "Protección social en Argentina: Financiamiento, cobertura y desempeño (1990-2003)," *Organización Internacional del Trabajo, Chile*.

Bertranou, F.; Cetrángolo, O. G. C. and Casanova, L. (2011), "Encrucijadas en la seguridad social argentina: reforma, cobertura y desafíos para el sistema de pensiones," *Organización Internacional del Trabajo y Comisión Económica Para América Latina y El Caribe, Argentina*.

Geri, M. and Villarreal, F. (2025), "Determinants of Contribution Density to the Argentine Pension System: An Analysis by Cohorts, 1996–2021," *Latin American Policy*, 16, 1–16.

ILO (2011), "Social security for social justice and a fair globalization," *International Labour Organization*, 62–117.

Johnson, A. A., Ott, M. Q., and Dogucu, M. (2022), *Bayes rules!: An introduction to applied Bayesian modeling*, Chapman and Hall/CRC.

Luque, C. and Sosa, J. (2024), "Bayesian analysis for social science research," *Model Assisted Statistics and Applications*, 19, 173–195.

Moore, P. (1966), "The bayesian approach to statistics," *Journal of the Institute of Actuaries*, 92, 326–339.

Rofman, R., L. L. and Ourens, G. (2009), "Pension systems in Latin America: concepts and measurements of coverage," *World Bank Discussion pape*.

Rofman, R. and Oliveri, M. L. (2012), "Un repaso sobre las políticas de protección social y la distribución del ingreso en Argentina," *Económica*, 58.

Received: April 25, 2025

Accepted: September 17, 2025