İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

İstanbul Commerce University Journal of Science

http://dergipark.org.tr/ticaretfbd

Research Article / Araştırma Makalesi

A COMPARISON OF STATISTICAL DISTRIBUTIONS FOR THE CRUDE BIRTH RATE DATA

DOĞUM ORANININ MODELLENMESİ İÇİN İSTATİSTİKSEL DAĞILIMLARIN KARŞILAŞTIRILMASI

Ceren ÜNAL[1] **Gamze ÖZEL**[2]

https://doi.org/10.55071/ticaretfbd.1277633

Corresponding Author / Sorumlu Yazar
cerenunal@hacettepe.edu.tr

Received / Geliş Tarihi
05.04.2023

Accepted / Kabul Tarihi
03.08.2023-

Abstract

Population statistics and demographic are important indicators to show the country's quality of life, social and health, the status of the population, the change in the population structure, and its effect on economic life. In demography, the crude birth rate is used to measure the growth of a population. In this study, we perform the crude birth rate values by statistical regions in Türkiye with some statistical distributions. When the results are compared, the Gumbel distribution provides the best fit to model the crude birth rate values than Normal, Log-Normal, Exponential, Gamma, and Weibull distributions. Among compared distributions, the Gumbel model is the best model to present the CBR data since log likelihood (logL), Akaike information criteria (AIC) and Bayesian information criteria (BIC) values of the Gumbel distribution are the lowest among all models. In addition, the results with the compared distributions are supported by graphs.

Keywords: Crude birth rate, extreme value distributions, Gumbel distribution, population estimates, Turkish population.

Öz

Nüfus istatistikleri ve demografi, ülkenin yaşam kalitesini, sosyal ve sağlık durumunu, nüfusun durumunu, nüfus yapısındaki değişimi ve bunun ekonomik hayata etkisini gösteren önemli göstergelerdir. Demografide, bir nüfusun büyümesini ölçmek için kaba doğum hızı kullanılmaktadır. Bu çalışmada, Türkiye'deki istatistiki bölge sınıflandırmasına göre kaba doğum hızı değerlerinin istatistiki dağılımlara uyumu araştırılmaktadır. Elde edilen sonuçlar karşılaştırıldığında, Türkiye'de kaba doğum hızı değerlerini modellemek için Normal, Log-Normal, Exponential, Gamma ve Weibull dağılımlarına göre en iyi uyumu Gumbel dağılımı sağlamaktadır. Karşılaştırılan dağılımlar arasında, Gumbel dağılımının log olabilirlik (logL), Akaike bilgi kriterleri (AIC) ve Bayes bilgi kriterleri (BIC) değerleri tüm modeller arasında en düşük olduğundan, CBR verilerini en iyi sunan model Gumbel modelidir. Ayrıca, karşılaştırılan dağılımlar ile sonuçlar grafiklerle desteklenmiştir.

Anahtar Kelimeler: Gumbel dağılımı, kaba doğum hızı, nüfus tahminleri, Türkiye nüfusu, uç değer dağılımları.

[1] Hacettepe University, Faculty of Science, Department of Statistics, Ankara, Türkiye.
cerenunal@hacettepe.edu.tr, Orcid.org/0000-0002-9357-1771.

[2] Hacettepe University, Faculty of Science, Department of Statistics, Ankara, Türkiye.
gamzeozl@hacettepe.edu.tr, Orcid.org/0000-0003-3886-3074.

1. INTRODUCTION

The birth rate represents the development of the population in demography. It is used to determine the nature of the mass of the population, age, and gender distribution. Demographers use various indicators to measure the birth rate, one of which is the crude birth rate (CBR), which represents the number of live births per one thousand individuals within a population (Hamilton et al., 2009). A CBR of more than 30 per 1000 is considered high, while a rate of less than 18 per 1000 is considered low. India has the highest CBR in the world, with a CBR of 13.496,25 births per thousand individuals, accounting for 16,71% of the world's CBR as of 2020. The top five countries, including Nigeria, China, Pakistan, and Ethiopia, accounting for 38,57% of the world's CBR. The global CBR was estimated to be 80.754,21 births per thousand individuals in 2020.

The CBR is gradually decreasing in Türkiye. The social and social environment in which Türkiye is located economic transformation and human capital investments affect people's fertility tendencies. It is possible to classify the factors affecting fertility as biological, social, and economic. In 2001, while the CBR in Türkiye was 20,3 per thousand, it became 13,3 per thousand in 2020. In other words, there were 20,3 births per thousand population in 2001 and 13,3 births in 2020. This indicates that the CBR has decreased from 20,3 to 12,8 per thousand population over two decades from 2001 to 2021 (TurkStat, 2022a).

Demographic models aim to present demographic processes through mathematical functions that relate measurable demographic variables. The main goal of modeling is to simplify complex numerical data into a few easily understood basic parameters or to provide an approximate representation of reality without its intricacies (Abd Ellatif, 2017). Deterministic models are frequently used to describe the population size dynamics during growth (Hannon, 1997; Brauer and Castillo-Chavez, 2013; Anderson, 2014). These models typically describe population size as a continuous variable, and its temporal dynamics are governed by an ordinary differential equation. However, many of these models are nonlinear, making analytical progress challenging and sometimes limited (Tsoularis & Wallace, 2002; Marrec et al., 2022).

The logistic differential equation was originally developed to account for the self-regulating nature of population growth and to amend the Malthusian exponential growth model. Nevertheless, relying solely on deterministic models to make predictions can lead to inaccurate estimation of important parameters. It is essential to recognize instances when a deterministic equation fails to accurately depict the average dynamics of stochastic population growth and to comprehend the underlying causes for such discrepancies (Marrec et al., 2022).

In this work, we perform and compare the results of some statistical distributions to model the CBR values by statistical regions in Türkiye. As far as we know, these distributions have not been used to model the CBR values in Türkiye before. The data set used in the study was taken from the statistical data portal of TurkStat (2022b). It includes CBR data information between 2015-2020 years by statistical regions for Turkey.

The paper is organized as follows: In Section 2, we give the methodology of the study by explaining statistical distributions. The data set is described in Section 3 and the results are compared in Section 4. The concluding remarks are presented in Section 5.

2. METHODOLOGY

A probability distribution describes the values and probabilities for a random event to occur. Normal distribution, Log-normal distribution, exponential distribution, gamma distribution, and Weibull distribution are well-known continuous distributions (Demirci Biçer & Atakan, 2012). Gamma, Weibull, and Gumbel, distributions are commonly used distributions for the analysis of skewed data. This distribution has often been caused by industrial reliability issues and human is used as a model for survival (Lee & Wang, 2003). In this study, we perform these distributions to model CBR data. Now, we explain some important statistical characteristics of these distributions.

The normal distribution is one of the most important distributions. It is also called the Gaussian distribution after the German mathematician Carl Friedrich Gauss. It fits the probability distribution of many events. For example, IQ rating, and heart rate. The normal distribution, the Gaussian distribution, or the bell-curved distribution is abundant in physical nature and is often used in practically applied statistics. The reason for this is the central limit theorem. Normal distribution not skewed is a symmetrical distribution. If a random variable X has a normal distribution with mean μ and variance σ^2 $(\sigma^2 > 0)$, then probability density function (pdf) of the form is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{1}$$

Logarithmic conversion to the random variable as y = ln (x) is applied if the distribution of the transformed variable Y is normal. The distribution of X the variable is lognormal. Unlike the normal distribution, negative values are not used in the lognormal distribution. Its distribution is used more in modelling economic data (Jafari & Abdollahnezhad, 2017). The pdf of the log-normal distribution is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{(lnx-\mu)^2}{2\sigma^2}\right)}. \tag{2}$$

The gamma distribution is an extension of the exponential distribution, and it is a model by Brown and Flood (1947). The gamma distribution is one of the extensively used distributions for modelling skewed data in various fields such as hydrology, and finance, especially for reliability or lifetime (Basak & Balakrishnan, 2012; Hirose, 1995; Vaidyanathan & Lakshmi, 2015; Yonar & Yapıcı Pehlivan, 2022). The pdf of the gamma distribution with a shape parameter (θ) and scale parameter (λ) is given by

$$f(x) = \frac{\lambda^\theta}{\Gamma(\theta)} x^{\theta-1} exp(-\lambda x), \tag{3}$$

where $x > 0, \theta > 0, \lambda > 0$ and $\Gamma(\theta)$ is the gamma function.

The Weibull distribution was first modelled by Weibull (1951) for the modelling of material properties. It is well known that the Weibull distribution is the most popular and the most widely used distribution in reliability and in an analysis of lifetime data (Almalki & Nadarajah, 2014). The Weibull distribution is characterized by its shape parameter (θ) and scale parameter (λ). The pdf of the Weibull distribution is given by

$$f(x) = \frac{\theta}{\lambda}\left(\frac{x}{\lambda}\right)^{\theta-1} exp\left(-\frac{x}{\lambda}\right)^{\theta}. \tag{4}$$

The Gumbel distribution is also known as the extreme value distribution of type I in the literature. The Gumbel distribution has found application in various scientific areas, including but not limited to hydrology, meteorology, climatology, insurance, finance, and geology, among numerous other fields of study (Gómez et al., 2019). The pdf of the Gumbel distribution is given by

$$f(x)=\frac{1}{\lambda}exp\left(-\frac{x-\mu}{\lambda}-exp\left(-\frac{x-\mu}{\lambda}\right)\right), \tag{5}$$

where $x \geq 0$ and λ $(\lambda > 0)$is scale parameter.

3. MATERIAL

The CBR refers to the number of live births occurring in a population per thousand individuals during a particular year. It is calculated as follows:

$$\text{CBR= (B/N) x 1000.} \tag{6}$$

This formula includes the variables B, representing the number of births, and N, representing the midyear population. The denominator in the CBR formula is typically an average population size for a given period and is often expressed as a mid-year population estimate. This estimate is calculated by taking the average of the population at the beginning and end of the period. CBR is commonly expressed as the number of live births per 1,000 population (Spoorenberg, 2015). Figure 1 shows the CBR of Türkiye between 2001 and 2022.

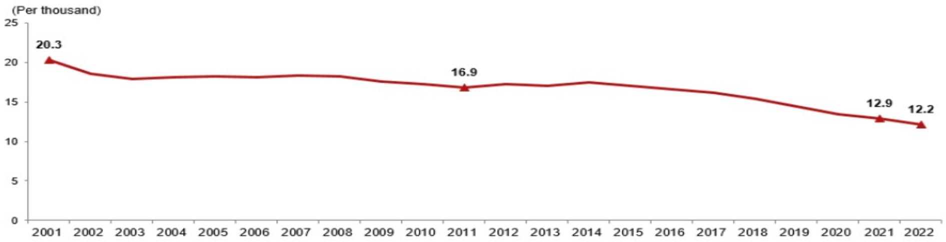

Figure 1. CBR Values of Türkiye between 2001 and 2022 (TurkStat, 2022a)

Figure 1 presents a decreasing trend seen in the CBR as a result of indicators such as the decrease in the number of births, the delayed age of marriage, and the effective use of birth control methods. Especially, it can be mentioned that there has been a continuous decrease in the CBR since 2014. In Figure 2, the first ten provinces with the highest and lowest CBR in 2022 are presented.

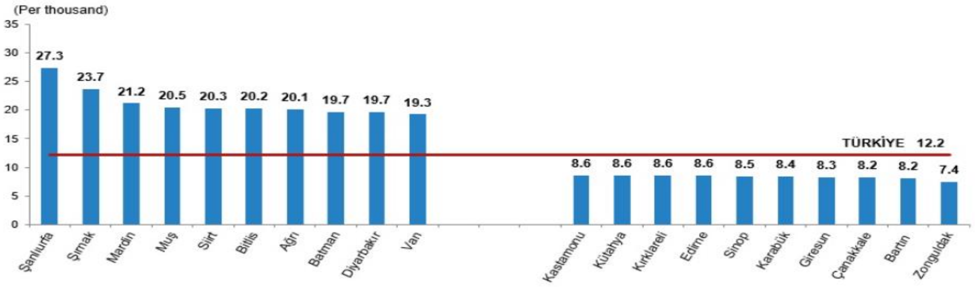

Figure 2. The First 10 Provinces with the Highest and Lowest CBR in 2022 (TurkStat, 2022a)

It is seen in Figure 2 that the province having the lowest CBR was Zonguldak with 7,4 per thousand and the highest CBR was Şanlıurfa with 27,3 per thousand. Bartın and Çanakkale with 8,2 per thousand, Giresun with 8,3 per thousand, Karabük with 8,4 per thousand, Sinop with 8,5 per thousand and Edirne, Kırklareli, Kütahya, and Kastamonu with 8,6 per thousand followed by Zonguldak. The CBR map of Türkiye with all provinces is presented in Figure 3. When the CBRs in the provinces in Eastern and Southeastern Anatolia are examined in Figure 3, it is seen that they are higher than the Western and Central Anatolian regions.

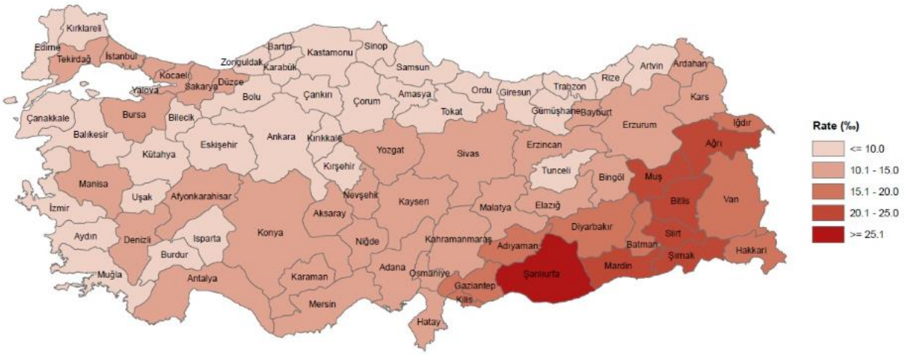

Figure 3. Map of the CBR by Provinces of Türkiye in 2022 (TurkStat, 2022a)

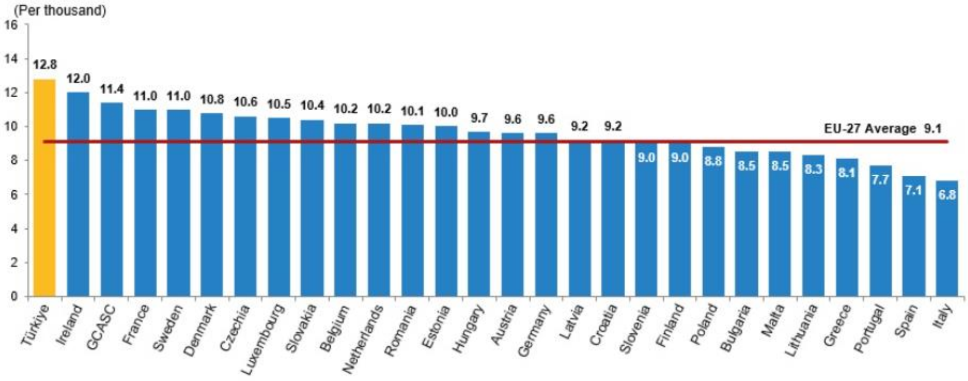

Figure 4. Comparison of the CBR with the EU Member Countries in 2021 (TurkStat, 2022a)

The CBR in Türkiye was found to be higher than that of the European Union (EU) member countries. An analysis of the CBRs of 27 EU member countries revealed that Türkiye's CBR was higher than those of all 27 countries. An analysis of the CBRs of 27 EU member countries revealed that Ireland had the highest CBR of 12,0 per thousand, while Italy had the lowest CBR of 6,8 per thousand in 2021 (The World Bank, 2022; TurkStat, 2022a).

4. RESULTS

In this study, we used the CBR values by statistical regions in Türkiye, which are taken from the statistical data portal of TurkStat (2022b). In order to collect and develop regional statistics, make socioeconomic analyzes of the regions, determine the framework of regional policies, and create a comparable statistical database in accordance with the European Union Regional Statistical System, the data and information produced are presented within the scope of the Classification of Statistical Regional Units 1st Level, 2nd Level, and 3rd Level. Level 2 Statistical Regions are constituted as 26 regions which are defined by “Level 3” Statistical Region Units grouped according to the neighboring provinces. Level 2 regions were used among these levels in this study. Since the data set is statistical regions based data, it consists of CBR values per thousand between 2015-2020 years. Descriptive statistics of the CBR data set are given in Table 1.

Table 1. Descriptive Statistics of the CBR Values (Per Thousand) by Statistical Regions in Türkiye

Min.	Q_1	Median	Mean	Q_3	Max	Variance	St. Dev.
8,7	12,2	14,1	15,6	16,9	30,3	25,1	5,01

When Table 1 was examined, it was concluded that the lowest CBR value was 8,7 and the highest CBR value was 30,3. In addition, it is determined that the average CBR value was 15,6 and the variance value was 25,05. When the quantile values are examined, it can be said that 25% of the CBR values are less than 12,2; 50% of the CBR values are less than 14,1, and 75% of the CBR values are less than 16,9.

Skewness of the CBR data set is obtained as S = 1,168> 0, hence the data set is skewed to the right. Kurtosis of the CBR data is K = 3,517 > 3, the data set is leptokurtic. Leptokurtic distribution can be defined as skinny in the center, it also features a fat tail. Then, the boxplot of the CBR data set is presented in Figure 5. Figure 5 also supports the skewness and fat tails of the data set.

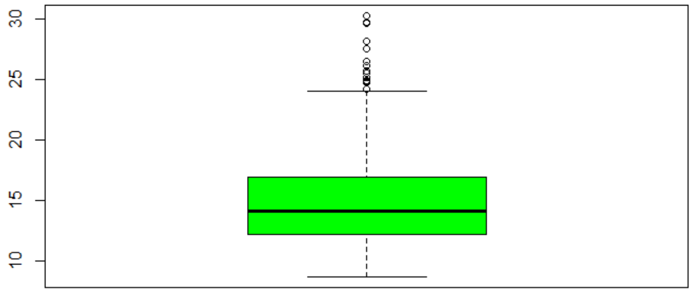

Figure 5. Box Plot of the CBR Values for Statistical Regions in Türkiye

A Histogram of the CBR data is given in Figure 6. Looking at the histogram graph in Figure 6, it is also seen that the CBR data is skewed to the right.

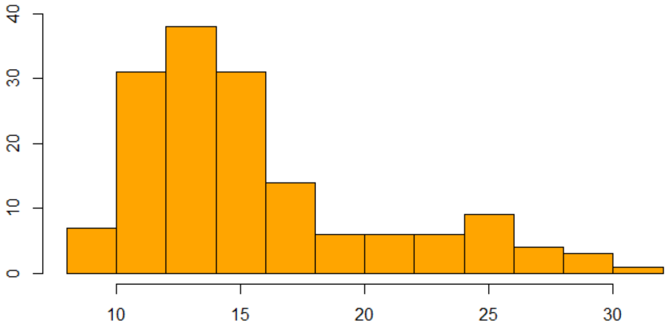

Figure 6. Histogram Graph of the CBR Values for Statistical Regions in Türkiye

In Table 2, parameter estimations of models are given with standard errors in parentheses, and model selection test results are also provided. Table 2 indicates that the Gumbel model is the best model to present the CBR data since log likelihood (logL), Akaike information criteria (AIC) and Bayesian information criteria (BIC) values of the Gumbel distribution are the lowest among all models.

Table 2. Parameter Estimations with Standard Errors in Parenthesis and Model Selection Criteria Test Results

Model	**a**	**b**	**logL**	**AIC**	**BIC**
Normal	15,554010 (0,3994355)	4,988948 (0,2824435)	-472,0815	948,163	954,2627
Log-normal	2,699335 (0,02332403)	0,291317 (0,01649170)	-450,0491	904,0981	910,1979
Exponential	0,0642921 (0,005146242)	-	-584,1137	1170,227	1173,277
Gamma	11,2820804 (1,25897418)	0,7253603 (0,08276918)	-455,7635	915,5271	921,6268
Weibull	3,195582 (0,1835044)	17,345840 (0,4624722)	-472,6722	949,3444	955,4441
Gumbel	13,387516 (0,2862654)	3,419526 (0,2281790)	-446,6194	897,2388	903,3386

The suitability of statistical models is based on some parametric can be analyzed by the goodness of fit tests (Oseni & Ayoola, 2013). Among tests, one-sample Kolmogorov-Smirnov test results are given in Table 3. These results also support the superiority of the Gumbel distribution. Since the p-value (0,22) is greater than 0,05, it can be said that the Gumbel distribution fits well with the CBR data.

Table 3. Results of Asymptotic One-Sample Kolmogorov-Smirnov Test

Model	**D**	**p-value**
Normal	0,16625	0,0004
Log-normal	0,10199	0,078
Exponential	0,43952	0,000
Gamma	0,12403	0,0165
Weibull	0,16608	0,0004
Gumbel	0,084089	0,22

After deciding the suitable distribution as Gumbel, we provide plots for the empirical pdf of data and theoretical pdf of Gumbel, quantile-quantile (Q-Q) plot, empirical cdf of data and theoretical cdf, and probability-probability (P-P) plot in Figure 7. Plots in Figure 7 also support the suitability of the Gumbel distribution to the data.

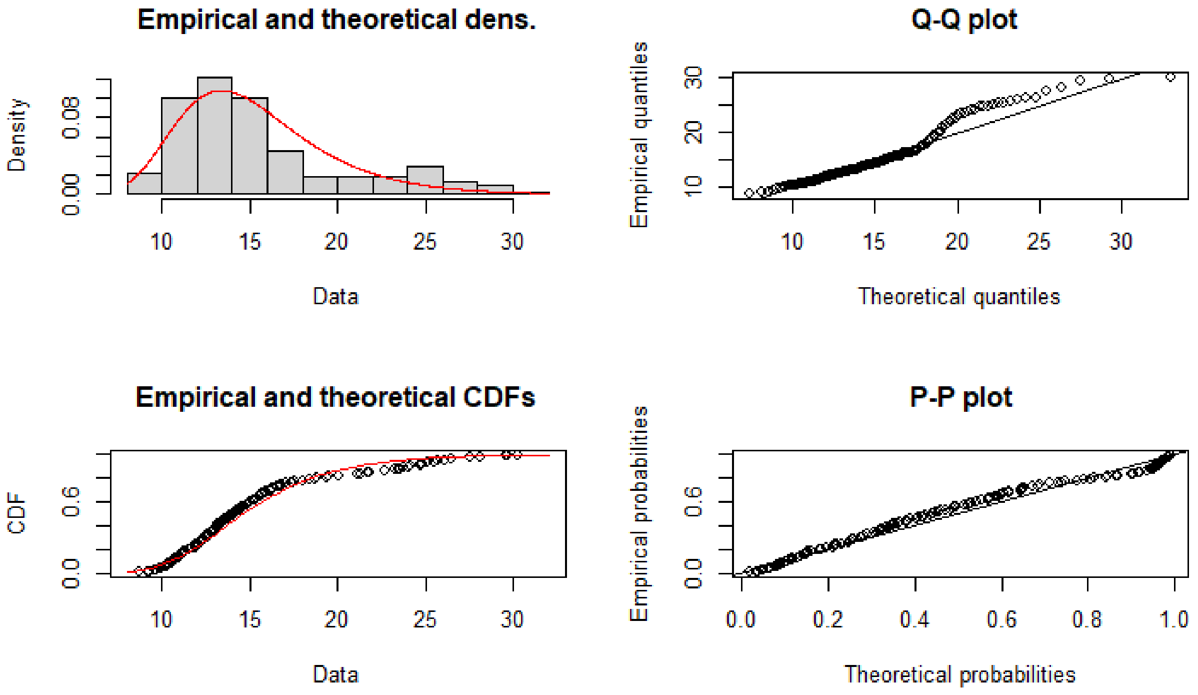

Figure 7. Plots for Q-Q, P-P, Empirical and Theoretical Pdfs and Cdf's of Gumbel Distribution

In Figure 8, histogram and theoretical densities of the competitive distribution are presented. Figure 8 shows that the Gumbel distribution is the best model among all models.

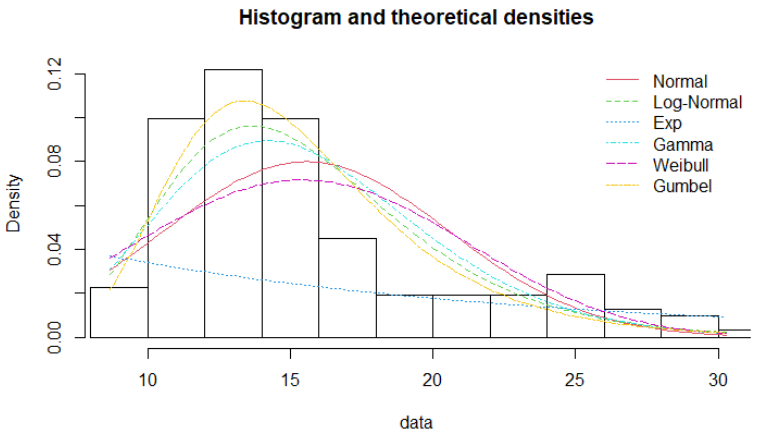

Figure 8. Comparison of Histogram and Theoretical Densities for the CBR Data

A comparison of the empirical cdf of data and theoretical cdf of the competitive distributions is presented in Figure 9. It can be seen in Figure 9 that the cdf of the Gumbel model fits the empirical cdf well.

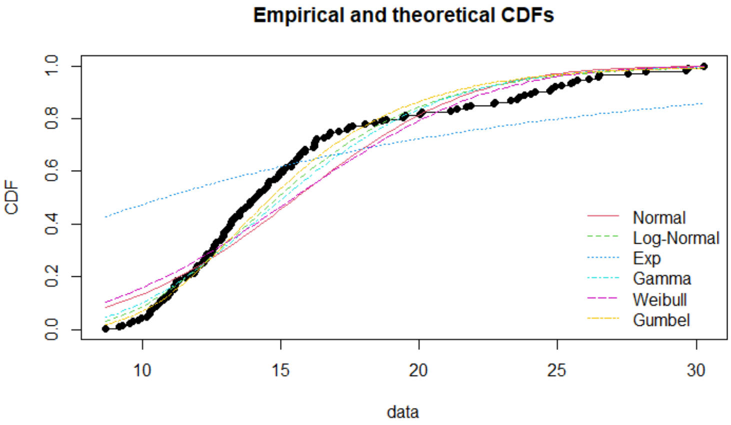

Figure 9. Comparison of Empirical and Theoretical Cdf’s for the CBR Data

Figures 10 and 11 present the P-P and Q-Q plots of the alternative distributions for the CBR values, respectively. As seen in Figures 10 and 11, the Gumbel distribution fits the data well.

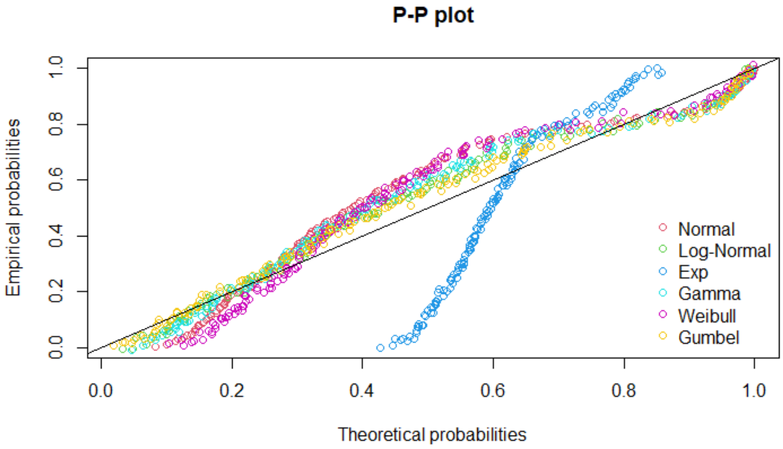

Figure 10. P-P Plot of the Alternative Distributions for the CBR Data

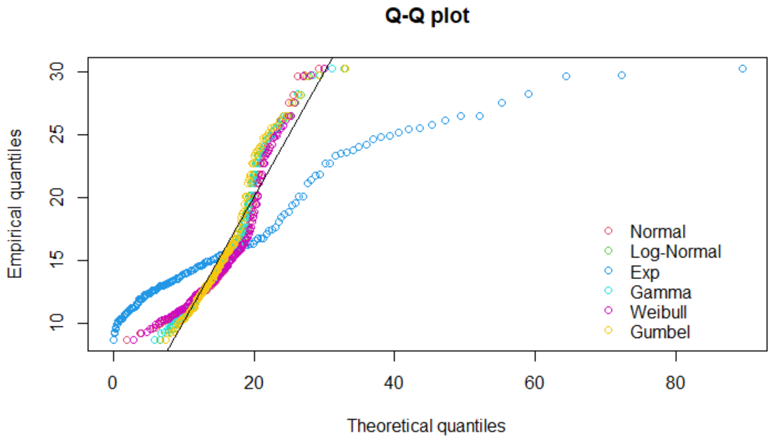

Figure 11. Q-Q Plot of the Alternative Distributions for the CBR Data

5. CONCLUSION

In this study, the crude birth rate values by statistical regions in Türkiye are performed with some statistical distributions. These values are taken from the statistical data portal of TurkStat and are given descriptive statistics in detail. According to the results, the Gumbel distribution provides a best fit than Normal, Log-Normal, Exponential, Gamma, and Weibull distributions to model the crude birth rate values by statistical regions in Türkiye. Because of the log-likelihood (logL), Akaike information criteria (AIC), and Bayesian information criteria (BIC) values, the Gumbel distribution is the lowest among all models. The plots for Q-Q, P-P, empirical and theoretical pdfs and cdf's of Gumbel distribution is given. In addition, the results with the compared distributions are supported by graphs.

Contribution of The Authors

The authors confirm that they equally contributed to this paper.

Conflict of Interest

The authors declare that there is no conflict of interest.

Statement of Research and Publication Ethics

Research and publication ethics were observed in the study.

REFERENCES

Abd Ellatif, S.M.A.E. (2017). *A comparative study to estimate and forecasting mortality using demographic and statistical models* [Ph.D. thesis]. Sudan University of Technolegy & Sciences, Sudan.

Almalki, S.J. & Nadarajah, S. (2014). Modifications of the Weibull distribution: A review. *Reliability Engineering & System Safety*, 124, 32-55.

Anderson, R.M. (2014). The population dynamics of infectious diseases: Theory and applications. *Springer*, New York.

Basak, I. & Balakrishnan, N. (2012). Estimation for the three-parameter gamma distribution based on progressively censored data. *Statistical Methodology*, 9(3), 305-319.

Brauer, F. & Castillo-Chavez, C. (2013). Mathematical models in population biology and epidemiology. Texts in applied mathematics. *Springer,* New York.

Brown, G.W. & Flood, M.M. (1947). Tumbler mortality. *Journal of the American Statistical Association*. 42(240), 562-574.

Demirci Biçer, H. & Atakan, C. (2012). Gamma, Weibull ve Log-Normal dağılımlarının doğru seçim olasılıklarına göre ayrıştırılması. *İstatistik Araştırma Dergisi*, 9(1), 11-20.

Gómez, Y. M., Bolfarine, H. & Gómez, H. W. (2019). Gumbel distribution with heavy tails and applications to environmental data. *Mathematics and Computers in Simulation*, 157, 115-129.

Hamilton, B.E., Martin, J.A. & Ventura, S.J. (2009). Births: Preliminary data for 2007. *National Vital Statistics Reports*, 57(12), 1-23.

Hannon, B., Ruth, M. & Levin, S.A. (1997). Modeling dynamics biological systems. Modeling Dynamic Systems. *Springer*, New York.

Hirose, H. (1995). Maximum likelihood parameter estimation in the three-parameter gamma distribution. *Computational Statistics & Data Analysis*, 20(4), 343-354.

Jafari, A.A. & Abdollahnezhad, K. (2017). Testing The equality means of several log-normal distributions. *Communications in Statistics - Simulation and Computation*, 46(3), 2311-2320.

Lee, E.T. & Wang, J. (2003). Statistical methods for survival data analysis, 476. *John Wiley & Sons, Inc.*

Marrec, L., Bank, C. & Bertrand, T. (2022). Solving the stochastic dynamics of population growth. *BioRxiv*. 1-15.

Oseni, B.A. & Ayoola, F.J. (2013). Fitting the Statistical Distribution for Daily Rainfall in Ibadan, Based On Chi-Square and Kolmogorov-Smirnov Goodness-Of-Fit Tests. *West African Journal of Industrial and Academic Research*, 7(1), 93-100.

Spoorenberg, T. (2015). Evaluation and analysis of fertility data. *Regional Workshop on the Production of Population Estimates and Demographic Indicators. Addis Ababa. United Nations, Department of Economic and Social Affairs*.

The World Bank, (2022). Retrieved July 21, 2023 from https://data.worldbank.org/indicator/SP.DYN.CBRT.IN?end=2020&locations=EU&most_recent_value_desc=true&start=2020.

Tsoularis, A. & Wallace, J. (2002). Analysis of logistic growth models. *Mathematical Biosciences*, 179, 21– 55.

TurkStat, Birth Statistics (2022a). Retrieved July 21, 2023 from https://data.tuik.gov.tr/Bulten/Index?p=Birth-Statistics-2022-49673&dil=2#:~:text=Crude%20birth%20rate%20was%2012.2,12.2%20per%20thousand%20in%202022.

TurkStat, (2022b). Retrieved July 21, 2023 from https://data.tuik.gov.tr/Bulten/Index?p=Birth-Statistics-2020-37229&dil=2.

Vaidyanathan, V. & Lakshmi, R.V. (2015). Parameter Estimation in Multivariate Gamma Distribution. *Statistics, Optimization Information Computing*, 3(2), 147-159.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 9(1951), 293-297.

Yonar, A. & Yapıcı Pehlivan, N. (2022). Evaluation and comparison of metaheuristic methods to estimate the parameters of gamma distribution. *Nicel Bilimler Dergisi*, 4(2), 96-119.