International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Computation of Human Lifespan with a Weibull Distribution

Jonah Ringy Mahevahaja[1], Tovohery Josoa Michel[2]

Department of Mathematics and Computer Science, Ecole Normale Supérieure pour l'Enseignement Technique (ENSET) of the University of Antsiranana
[1]Email: *jonahringy[at]gmail.com*
[2]Email: *josoamicheltovohery[at]gmail.com*

Abstract: ***This article proposes to calculate the lifespan of a given human population after making an adjustment with a three-parameter Weibull distribution. A new approach to calculating Weibull distribution parameters is illustrated with real data processing. The latter is then compared with the methods proposed by Cran and David.***

Keywords: Human lifespan, Weibull distribution, Kolmogorov – Smirnov goodness-of-fit test.

1. Introduction

Human lifespan is defined as the maximum number of years a human being can live (Travis, David, & Xiao-Dong, 2019). It is an indicator of the level of development and a measure of a state's human development index (HDI) (Debbagh & Yousfi, 2020). Economists and demographers compute it to model a population's evolution and movements (Blanpain, 2018). These many important factors raise our minds to a study on the mortality of the population of the Antsiranana district – Madagascar.

In general, the method of calculating lifespan is illustrated in demography. However, we want to estimate it using a probabilistic method, i. e. using a probability distribution. Indeed, the literature illustrates usual probability distributions for calculating lifespan, such as the exponential distribution, the Gompertz distribution and the Weibull distribution (Trifon, Adam, Laszlo, Vladimir, & James, 2015), but the Weibull distribution is the most widely used. So the question we're going to answer is, how do we adjust the human lifespan in Weibull distribution?

The rest of this article is organised as follows: the second section presents a literature survey on the estimation of the parameters of the three-parameter Weibull distribution and our proposed new approach; the third section presents the results found during this study and illustrates the latter with interpretations and a comparison of the methods used; the fourth section concludes this analysis with a conclusion and some perspectives.

2. Literature Survey

We have considered the Weibull law with three parameters to make the adjustment. Cran's method and David's method were used to estimate the parameters of law (Cran, 1988), (David, 1975). The graphical method is also used to correct the estimates found. The Kolmogorov – Smirnov goodness-of-fit test is done to prove the relevance of the results found.

2.1 Three-parameter Weibull distribution

Definition 1: The Cumulative Distribution Function of a random variable T following a three-parameter Weibull distribution is defined by:

$$\forall t \in [0, +\infty[, P(T < t) = F_T(t) = 1 - \exp\left[-\left(\frac{t-\gamma}{\eta}\right)^{\beta}\right] \quad (1)$$

and its density function is defined by:

$$\forall t \in [0, +\infty[, f_T(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} \exp\left[-\left(\frac{t-\gamma}{\eta}\right)^{\beta}\right], \quad (2)$$

Where β is a shape parameter, η is the scale parameter, and γ is a location parameter (Kappenman, 1985).

These parameters can be interpreted as follows:

1) Interpretation of β:
   a) If 0<β<1, then the mortality rate is decreasing. This characterizes mortality during the youth period (example infant mortality).
   b) If β=1, then the mortality rate is constant. We therefore have an exponential distribution with parameter λ=1/η.
   c) If $\beta > 0$, then the mortality rate is increasing. That shows the impacts of aging.
2) Interpretation of: This parameter is used to match the real (observed) values to the theoretical values predicted by the Weibull distribution. It gives the order of magnitude of the mean lifespan. ηis always positive.
3) Interpretation of γ:
   a) If γ<0, then the system under study may be defective before birth.
   b) If γ=0, then the studied system may fail at birth.
   c) If γ>0, then the probability of failing at birth is zero. In this case, the probability of stillbirth is very low. The sign of the parameter γ can be detected by the shape of the point cloud (see Figure 1).

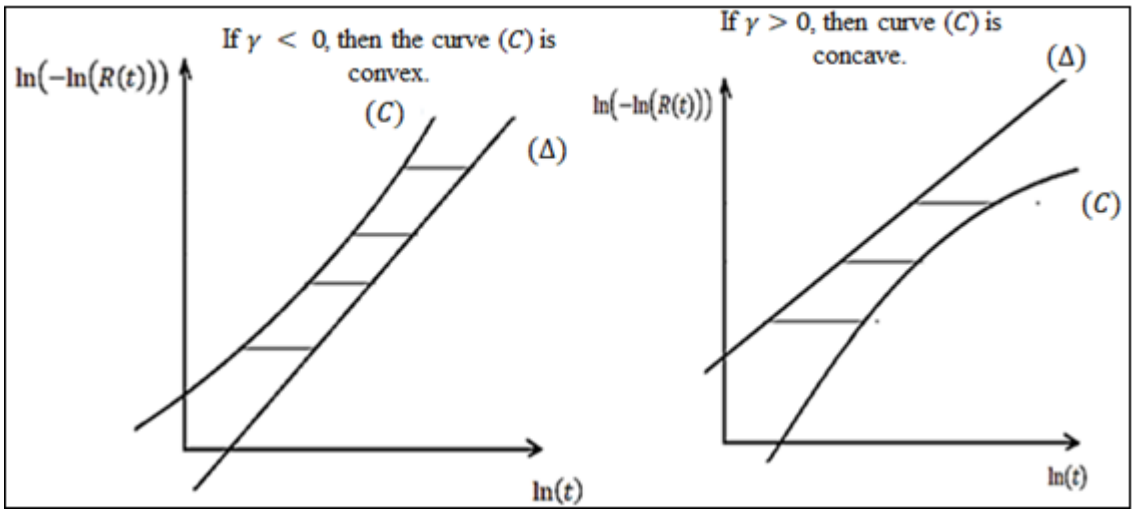

Figure 1: Curve (C) and sign of parameter γ (source: (Bellaouar & Beleulmi, 2013))

Definition 2: The mathematical expectation of a random variable T following the Weibull distribution with three parameters is defined by:

$$E(T) = \int_0^{+\infty} t.f_T(t)\,dt = A\eta + \gamma \qquad (3)$$

where $A = \Gamma\left(1+\frac{1}{\beta}\right)$ and $\Gamma(a) = \int_0^{+\infty} t^{a-1}e^{-t}\,dt$

Its variance is given by:

$$V(T) = \int_0^{+\infty} \left(t - E(T)\right)^2.f_T(t)\,dt = \eta^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma^2\left(1+\frac{1}{\beta}\right)\right] \qquad (4)$$

and the standard deviation by $\sigma(T) = \sqrt{Var(T)}$

2.2 Estimation of the parameters of a three-parameter Weibull distribution

In this section we present three methods for estimating the parameters of the Weibull distribution, such as the moment method (Cran, 1988), the graphical method by taking three values of the curve respecting David's condition (David, 1975) and the old method known as the "trial-and-error method" by gradually varying γ until the points $M_i(\ln(t_i); \ln(-\ln(R_i))$ are aligned (Bellaouar & Beleulmi, 2013).

Cranmethod:

The Cran method consists of estimating the parameters γ, βand η by the method of moments. Consider a series of lifetime values $S = \{t_1; t_2; \ldots; t_n\}$. Cran's method is organized in three steps:

Step 1: Establishing order statistics.

The values in the series S are arranged in ascending order $S = \{t_{(1,n)}\,;\, t_{(2,n)}\,;\ldots;\, t_{(n,n)}\}$ such as $t_{(1,n)} \le t_{(2,n)} \le \cdots \le t_{(n,n)}$. Here, $t_{(r,n)}$ denotes the r-th value of the sample of size n.

Step 2: Calculation of the moments of order $k = 1, 2$ and 4.

Let r be the rank of the value $t_{(r,n)}$. The moment of order k is estimated by:

$$m_k = \sum_{r=0}^{n-1}\left(1-\frac{r}{n}\right)^k\left(t_{(r+1,n)} - t_{(r,n)}\right), \quad t_{(0,n)} = 0. \qquad (5)$$

Step 3: Estimation of Weibull distribution parameters.

The parameters γ, β and η of the Weibull distribution are estimated by :

$$\hat{\gamma} = \frac{m_1 m_4 - m_2^2}{m_1 + m_4 - 2m_2}, \qquad (6)$$

$$\hat{\beta} = \frac{\ln(2)}{\ln(m_1 - m_2) - \ln(m_2 - m_4)}, (7)$$

$$\hat{\eta} = \frac{m_1 - \hat{\gamma}}{\Gamma\left(1+\frac{1}{\hat{\beta}}\right)}. \qquad (8)$$

Note: If the data are grouped into d-classes, then we need to make a slight adaptation to Cran's method. Consider the following data table:

Table 1: Table of data grouped in d-classes

T	$[0,t_1]$	$[t_1,t_2]$	$[t_2,t_3]$	$[t_3,t_4]$	...	$[t_{d-1},t_d]$	Total
Frequency	n_1	n_2	n_3	n_4	...	n_d	N

The moments of order k are estimated by:

$$m_k = \sum_{i=0}^{d-1}\left(1 - \frac{\sum_{j=1}^{i+1} n_j}{N}\right)^k (t_{i+1} - t_i),\ t_0 = 0. \qquad (9)$$

David Method:

Let the random variable T denote the lifetime. The reliability function R is defined by:

$$\forall t \in [0, +\infty[, R(t) = 1 - F_T(t) \qquad (10)$$

David's method involves taking three points A, B and C on the curve (C) defined by points M_i with coordinates $(x_i = \ln(t_i); y_i = \ln(-\ln(R(t_i))$, such that $y_C - y_B = y_B - y_C = 1$. Under this condition, the parameter γ is estimated by:

$$\hat{\gamma} = \frac{t_B^2 - t_A.t_C}{2t_B - t_A - t_C}. \qquad (11)$$

After transforming the abscissas t_i into $t_i' = t_i - \gamma$, the points A', B' and C' are aligned on a straight line with the equation: : $y = \hat{\beta}x + b$ and $\hat{\eta} = \exp(-b/\beta)$.

In the case of values grouped into d-classes, we apply the following principle:

Table 2: Estimation of the reliability function and coordinates of points $M_i\ (x_i, y_i)$

T	$[0, t_1[$	$[t_1, t_2[$	...	$[t_{d-2}, t_{d-1}[$	$[t_{d-1}, t_d[$	Total
Frequency	n_1	n_2	...	n_{d-1}	n_d	N
$F_i = \sum_{k=1}^{i} n_i/N$	F_1	F_2	...	F_{d-1}	$F_d = 1$	–
$R_i = 1 - F_i$	R_1	R_2	...	R_{d-1}	$R_d = 0$	–
$x_i = \ln(t_i)$	$x_1 = \ln(t_1)$	x_2		x_{d-1}	x_d	
$y_i = \ln(-\ln(R_i))$	y_1	y_2		y_{d-1}	–	–

So we just have to find three points $A(x_A, y_A), B(x_B, y_B)$ and $C(x_C, y_C)$ that satisfy the condition $\ln(-\ln(R_C)) - \ln(-\ln(R_B)) = \ln(-\ln(R_C)) - \ln(-\ln(R_A)) = 1$

For the values that are grouped in d classes, we interpolate to obtain the exact values. For a fixed point $A(x_A = \ln(t_A), y_A = \ln(-\ln(R_A)))$, we determine the interval $[y_i, y_{i+1}[$ containing y_B, such as $y_A + 1 = y_B$. Therefore, we estimate t_B by taking its logarithm, which is defined by:

$$\ln t_B = \ln t_i + \frac{(y_B - y_i)(\ln t_{i+1} - \ln t_i)}{y_{i+1} - y_i} \qquad (12)$$

We use the same method to calculate t_C.

Note: The estimation of the reliability function F by $F_i = \sum_{k=1}^{i} n_i/N$ is known as the estimation by the raw rank method (Tovohery, 2022).

Graphical method:

The graphical method consists in straightening the curve (C) by a translation of all the points by adding or subtracting from the abscissas t the same value γ in order to obtain a straight line (Δ) (see Figure 1) (Bellaouar & Beleulmi, 2013).

2.3 Kolmogorov – Smirnovgoodness-of-fit test

Kolmogorov – Smirnov goodness-of-fit test consists of comparing the maximum deviation between the cumulative distribution function of Weibull F with three parameters $\hat{\gamma}$, $\hat{\beta}$ and $\hat{\eta}$, and the empirical cumulative distribution functionF_i with a critical value. We note $D_{max} = |F_{(\hat{\gamma},\hat{\beta},\hat{\eta})}(t_i) - F_i|$. The critical value of the Kolmogorov-Smirnov test at the 95% confidence level is given by $1{,}36/\sqrt{N}$ (Frank & Massey, 1951). Thus, if $D_{max} < 1{,}36/\sqrt{N}$, then we accept at the 95% confidence level that the variable T follows the Weibull distribution with three parameters $\hat{\gamma}$, $\hat{\beta}$ and $\hat{\eta}$.

2.4 Problem definition

Parameter estimation is a kind of optimization problem. For example, James Evans mentioned that the maximum likelihood estimation requires an iterative procedure (Evans, et al., 2019). We therefore propose a simple approach based on the graphical method and the Kolmogorov-Smirnov goodness-of-fit test.

Formally, the estimation of the parameter γ consists in finding the value $\hat{\gamma}$, such that: $\max(|F_i - F_{(\hat{\gamma},\hat{\beta},\hat{\eta})}(t_i)| < 1{,}36/\sqrt{N}$ and for any very small $\varepsilon > 0$, $\max(|F_i - F_{(sign(\hat{\gamma})\times\delta,\ \hat{\beta},\ \hat{\eta})}(t_i)| \geq 1{,}36/\sqrt{N}$, with $\delta = |\hat{\gamma}| - \varepsilon$. That is, $\hat{\gamma}$ is the limit value that separates the

validation region from the Kolmogorov-Smirnov test with the rejection region at the 95% confidence level. The function $F_{(\hat{\gamma},\hat{\beta},\hat{\eta})}$ denotes the function F_T given in equation (1), but with parameters $\hat{\gamma}, \hat{\beta}$ and $\hat{\eta}$.

2.5 Methodology / Approach

Our approach is illustrated by the following algorithm:

1) Consider the variables $X = \ln(t_i)$ and $Y = \ln(-\ln(R_i))$
2) Draw the curve (C) of the points $M_i(x_i; y_i)$.
3) Determine the sign of the parameter γ :
   a) $\gamma < 0$ if (C) is convex ;
   b) $\gamma = 0$ if (C) is a straight line ;
   c) $\gamma > 0$ if(C) is concave.
4) If $\gamma = 0$, then $\hat{\beta} = \frac{Cov(X,Y)}{V(X)}$ and $\hat{\eta} = \exp\left(-\frac{\bar{Y}-\hat{\beta}.\bar{X}}{\hat{\beta}}\right)$ then END; otherwise go to step 5. We denote by $cov(X,Y)$ the covariance of X and Y, and $\bar{X}$ is the mean of X.
5) Estimate of γ by an iterative method :
   a) Initialization: $\hat{\gamma}_i = signe(\gamma) \times \delta$ such that δ is a sufficiently large positive value.
   b) Calculation of other parameters : $\hat{\beta}_i = \frac{Cov(\ln(t_i - \hat{\gamma}_i)\ ; Y)}{V(\ln(t_i - \hat{\gamma}_i))}$ and $\hat{\eta}_i = \exp\left(-\frac{\overline{\ln(t_i-\hat{\gamma}_i)}-\hat{\beta}.\bar{X}}{\hat{\beta}}\right)$
   c) Kolmogorov-Smirnov test: If $D_{max} = \max(|F_i - F_{(\hat{\gamma}_i,\hat{\beta}_i,\hat{\eta}_i)}(t_i)| < 1{,}36/\sqrt{N}$ is false, then $\hat{\gamma}_{i+1} = sign(\gamma).(|\hat{\gamma}_i| + w_i)$, where w_i is a sequence of positive random values. Return to step (b).
   d) If the test in step (c) gives a TRUE value, then we try to progressively reduce the value $|\gamma|$ until the limit value is found.

3. Results & Discussion

The data used in this study comes from a projection based on INSTAT census data and an analysis of the structure of Antsiranana's population data. In this section, we illustrate the results found from these three methods, such as the results from Cran, David and our new method. We then list the parameter values selected and, above all, the average lifespan of the Antsiranana population.

Table 3: Processing of mortality data for the Antsiranana district-year 2020 – 2022

Classes	t_i	n_i	F_i	R_i	$\ln(-\ln(R_i))$	$F_T(t_i) = 1 - \exp\left[-\left(\frac{t_i - 3329.11}{3393.40}\right)^{151.25}\right]$
[0 - 5[	5	208	0,0773	0,9227	-2,5206	0,0672
[5 - 10[	10	41	0,0925	0,9075	-2,3325	0,0835
[10 - 15[	15	26	0,1022	0,8978	-2,2279	0,1036
[15 - 20[	20	76	0,1304	0,8696	-1,9682	0,1281
[20 - 25[	25	103	0,1686	0,8314	-1,689	0,1578
[25 - 30[	30	106	0,208	0,792	-1,4558	0,1936
[30 - 35[	35	123	0,2537	0,7463	-1,2288	0,2362
[35 - 40[	40	101	0,2912	0,7088	-1,0664	0,2864
[40 - 45[	45	157	0,3496	0,6504	-0,8437	0,3444
[45 - 50[	50	173	0,4138	0,5862	-0,6271	0,4103
[50 - 55[	55	152	0,4703	0,5297	-0,4535	0,4834
[55 - 60[	60	198	0,5438	0,4562	-0,2422	0,5621
[60 - 65[	65	231	0,6296	0,3704	-0,0067	0,6438
[65 - 70[	70	247	0,7214	0,2786	0,2453	0,7246
[70 - 75[	75	232	0,8076	0,1924	0,4996	0,8002
[75 - 80[	80	176	0,873	0,127	0,7243	0,8661
[80 - 85[	85	162	0,9331	0,0669	0,9951	0,9187
[85 - 90[	90	99	0,9699	0,0301	1,2538	0,9564
[90 - 95[	95	58	0,9915	0,0085	1,5608	0,9799
[95 - 100[	100	23	1	0	-	0,9923
	TOTAL	2692				$\max(\lvert F_i - F_T(t_i)\rvert) = 0.0183$

First, mortality data for the population of Antsiranana district with some treatments for the years 2020, 2022 and 2022 are shown in Table 3. The shape of the scatter plot in Figure 2 tells us that the sign of γ is negative.

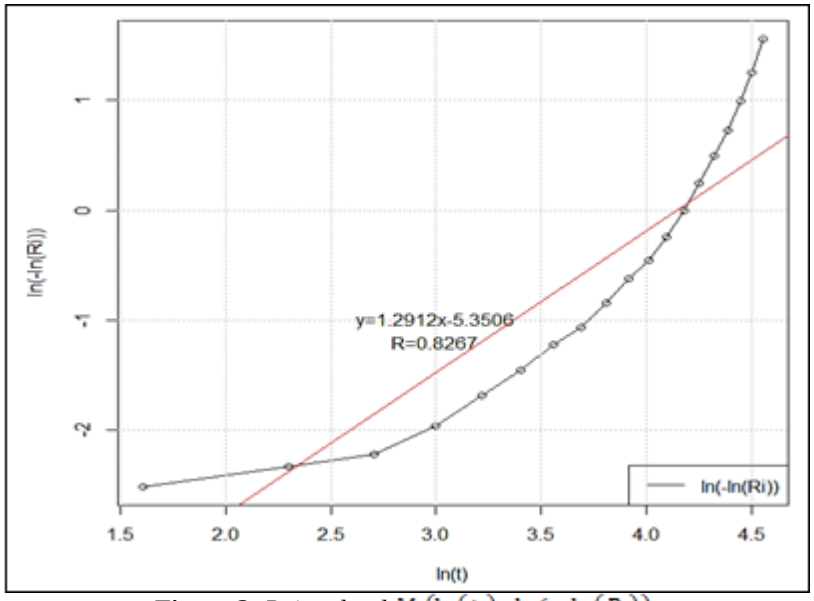

Figure 2: Point cloud $M_i\,(\ln(t_i)\,;\ln(-\ln(R_i))$.

Table 4 shows the values estimated by analytical methods such as Cran and David.

Table 4: Estimation results using Cran's method and David's method

Cran method		David's method					
$m_1 = 49{,}8626$	$\hat{\gamma} = -161{,}74$	$t_A = 10$	$\hat{\gamma} = -189{,}12$	$t_A = 20$	$\hat{\gamma} = -312{,}27$	$t_A = 30$	$\hat{\gamma} = +189{,}59$
$m_2 = 35{,}7189$	$\hat{\eta} = 222{,}41$	$t_B = 32{,}62$	$\hat{\eta} = 251{,}39$	$t_B = 42{,}13$	$\hat{\eta} = 374{,}91$	$t_B = 54{,}93$	$\hat{\eta} = N.D$
$m_4 = 22{,}5207$	$\hat{\beta} = 10{,}02$	$t_C = 57{,}81$	$\hat{\beta} = 10{,}57$	$t_C = 65{,}74$	$\hat{\beta} = 16{,}12$	$t_C = 75{,}96$	$\hat{\beta} = N.D$
$R(x,y) = 0{,}99$	$D_{max} = 0{,}07$	$R(x,y) = 0{,}99$	$D_{max} = 0{,}053$	$R(x,y) = 0{,}99$	$D_{max} = 0{,}04$	$R(x,y) = N.D$	$D_{max} = N.D$

By iteration according to our approach, we find $\gamma \in [3329{,}11\,;3329{,}21]$. For $\hat{\gamma} = -3329.11$, we have: $\hat{\beta} = 151.25$, $\hat{\eta} = 3393.40$, $R(X,Y) = 0.999$ and $D_{max} = 0.018$. The critical value of the Kolmogorov-Smirnov test at the confidence level 95% is given by $1.36/\sqrt{2692} = 0.0262$. This gives a mean life equal to: $E(T) = A\eta + \gamma = 51{,}48$ years and a standard deviation $\sigma(T) = 28{,}53$ years.

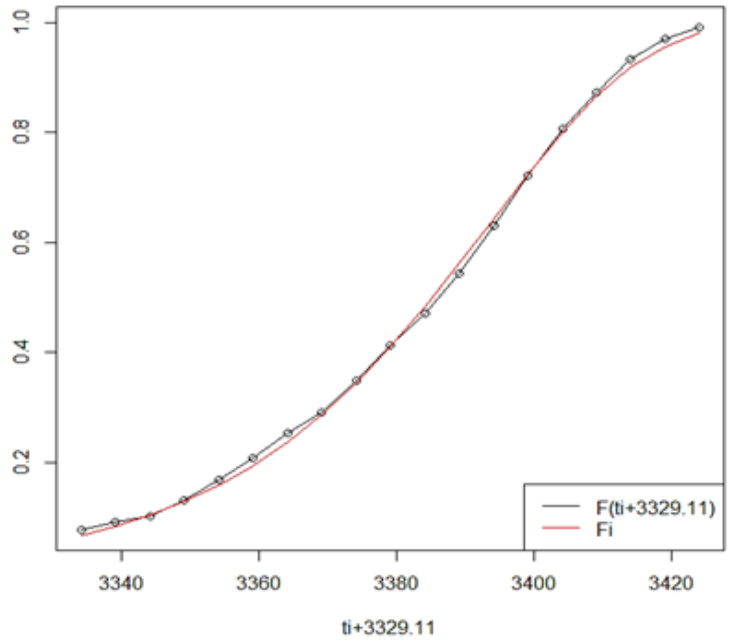

Figure 3: Curves of the empirical distribution function Fi and the Weibull distribution function with $\hat{\gamma} = -3329.11$, $\hat{\beta} = 151.25$ and $\hat{\eta} = 3393.40$.

Table 4 shows that Cran's method gives an estimate of γ which transforms the curve (C) into a straight line, but the model found is rejected by the Kolmogorov-Smirnov goodness-of-fit test. David's method gives different values depending on the coordinates of the first point A considered. This can result in an erroneous value (cf. last column of table 4). So, using David's method requires another iterative optimization algorithm to select the best point. A.

Our method offers an estimate of γ which transforms the curve (C) into a straight line, giving an acceptable model according to the Kolmogorov-Smirnov test. Indeed, it is the improved graphical method. However, the search for the best estimate of γ can be very time-consuming. It depends on the sequence of positive random values w_i that the analyst has taken into account. We therefore conclude our study with the following proposal:

Proposition: Let $X = \ln(t_i)$, $Y = \ln(-\ln(R_i))$. For any estimate of the parameter γ of the Weibull distribution, a coefficient of determination $R^2(X,Y) \geq 0.9$ (sufficiently high) does not lead to validation of the model according to the Kolmogorov-Smirnov test.

It should be noted that both Cran's and David's methods are initially used to estimate γ when this parameter is positive (David, 1975) and (Cran, 1988). However, some authors quote them to estimate γ regardless of its sign (Bellaouar & Beleulmi, 2013).

Finally, the result we have just found states that the inhabitants of the Antsiranana district expect to be alive in half a century.

4. Conclusion

In conclusion, it is possible to fit the human lifespan distribution with a three-parameter Weibull distribution. We have also seen that the Weibull distribution parameter estimation methods proposed by Cran and David do not quite give the best estimate whose model is validated by the Kolmogorov goodness-of-fit test. However, they do give us a starting point for finding γ. For this reason, we have chosen to estimate the parameter γ parameter using the graphical method, with validation of the Kolmogorov-Smirnov goodness-of-fit test as a constraint. Currently, several iterative algorithms are available for estimating Weibull parameters, but they do not take into account the Kolmogorov-Smirnov criterion. So, proposing an algorithm to estimate the parameters of the Weibull distribution is still an open problem.

References

[1] Bellaouar, A., & Beleulmi, S. (2013). Fiabilité, maintenabilité, disponibilité. Université Constantine 1.
[2] Blanpain, N. (2018). L'espérance de vie par niveau de vie méthode et principaux résultats. INSEE.
[3] Cran, G. W. (1988, October). Moment Estimators for the 3-Parameter Weibull Distribution. IEEE transactions on reliability, 37 (4).
[4] David, J. (1975). Détermination sanstâtonnement du coefficient gamma de la loi de Weibull. Révue de statistique appliqué, 23 (3), 81-85.
[5] Debbagh, B., & Yousfi, F. Z. (2020). Le mouvement coopératif féminin dans le milieu rural au Maroc: quelle contribution au développement humain ? Moroccan Journal of Entrepreneurship, Innovation and Management (MJEIM), 5 (1).
[6] Frank, J., & Massey, J. (1951, Mars). The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association, 46 (253), 68-78.
[7] Kappenman, R. F. (1985). Estimation for the three-parameter Weibull, lognormal, and gamma distributions. Computational Statistics & Data Analysis, 3, 11-23.
[8] Tovohery, J. M. (2022). Coefficient de variation et ses applications en régression linéaire et en extraction des règles d'association sur les variables quantitatives. Université de Toamasina-Madagascar.
[9] Travis, J. B., David, D. D., & Xiao-Dong, C. (2019). Use of MSCs in antiaging. In TravisJ. B., A Roadmap to Non-Hematopoietic Stem Cell-based Therapeutics.
[10] Trifon, I. M., Adam, L., Laszlo, N., Vladimir, C., & James, W. V. (2015, May 20). The Gompertz force of mortality in terms of the modal age at death. Demographic Research, 32, 1031–1048.