Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 5, Issue Special (2012): 37-49

ON RELIABILITY APPROACH AND STATISTICAL INFERENCE IN DEMOGRAPHY

Léo Gerville-Réache, Mikhail Nikulin and Ramzan Tahir

University Bordeaux, IMB, UMR 5251, F-33400 Talence, France
Email: ramzantahir7@gmail.com

Abstract

In demography, Gompertz and Makeham models have significant role in modeling and in analysis of mortality and ageing. Till the end of the 20th century, researchers have used the tables of mortalities (also called life tables) for demographic analysis but in the end of the 20th century due to the development in statistical methods of survival analysis and reliability one can treat the individuals data even with the information of censoring. The Gompertz, Makeham, and Weibull models are compared with respect to the goodness-of-fit to the table of mortality and to the individuals data in the presence of censoring. For data from the table of mortality, the test statistic considered by Gerville-Reache and Nikulin (2000) is used. For censored individual data the chi-squared type test proposed by Bagdonavicius et al. (2010) is used.

Key words: Demography, Chi-square test, Mortality table, Composite hypothesis, Censoring, Gompertz model, Makeham model, Weibull model, ML estimators, NRR statistic.

1. Introduction

In reliability and demography model selection for some specific data is vital for further analysis and decision making. Testing the two-parameter Gompertz distribution (Gompertz (1825)) to model the rate of mortality has been used for a long time, where the rate of mortality increases with the age. Gompertz-Makeham model (William Makeham (1860)) with one additional parameter covers the mortality independent of age. The researchers have used the life and mortality tables to find the force of mortality. Gerville-Reache and Nikulin (2000) gave a chi-square type goodness-of-fit test for Makeham model using the table of mortality (grouped data). In section 3 we briefly discuss their proposed statistic and also we compare Makeham model with Gompertz and Weibull models for different age groups. But now with the advanced data collection techniques, one can have the individual's information (ungrouped data) also with censoring mechanism. Gompertz and Makeham models are frequently used in demography but in reliability Weibull model is considered the alternative for Gompertz model (Juckett and Rosenberg, (1993)).

Most researchers compare Gompertz model with the Weibull model due to its flexible parameters (Gavrilov and Gavrilova (2001)). Logistic distribution can be another alternative for Gompertz (Wilson (1994)). The Gompertz function is a better choice for all causes of mortality and combined disease categories while the Weibull model has been shown to be a better choice over Gompertz model for a specific cause of mortality (Juckett and Rosenberg (1993)). Nikulin et al. (2011) presented several models in demography but here we consider the Gompertz-Makeham and Weibull models for censored data.

For individual censored data the test is based on the Nikulin-Rao-Robson statistic known as NRR statistic Y_n^2: a modified chi-squared test which is based on the differences between two estimators of the probabilities in each interval. One estimator is based on the empirical distribution function and the other is on the ML estimators of unknown parameters of the tested model from ungrouped data (see Nikulin (1973), Rao and Robson (1974), Drost (1988), LeCam et al. (1983)). The test is based on the vector $Z = (Z_1, \cdots, Z_k)'$, where $Z_j = \frac{1}{\sqrt{n}}(U_j - e_j)$, $j = 1, \cdots, k$, where U_j and e_j are the number of observed and expected failures in each interval. In literature some other modifications of chi-square goodness-of-fit tests for censored data have been proposed (see for example Akritas (1988), Hjort (1990), Kim (1993), Van der Vaart (1998)).

Some details on the Gompertz, Makeham and Weibull models are given in section 2. The NRR statistic for composite hypotheses and application of the test for Gompertz and Weibull distributions are given in section 4. Non-parametric estimation of survival function in demography and actuaries is given in section 5.

2. Gompertz-Makeham and Weibull Models

Gompertz model of aging is widely used in demography and other scientific disciplines e.g. medical sciences, survival analysis, actuarial sciences and reliability. Gompertz (1825) gave the first mathematical model to explain the exponential increase in mortality rate with age. He explained that the law of geometric progression pervades in mortality after a certain age. Gompertz mortality rate can be presented as

$$\mu_x = \theta e^{\nu x}, \quad (\theta, \nu) > 0, \quad x > 0, \tag{1}$$

where θ is known as the baseline mortality and ν is the age specific growth rate of the force of mortality.

Mortality rate μ_x in demographic notation is equivalent to the failure rate $\mu(x)$ in reliability or hazard rate $\lambda(x)$ in survival analysis. The Gompertz law has been the main demographic model since its discovering to fit the human mortality (see for example Gavrilov and Gavrilova (2001), Ricklefs and Scheuerlein (2002)).

Since Gompertz model gives the rate of mortality only related to age and does not take into account the other factors independent of age, other researchers tried to modify this model to fulfill the requirement of real data. William Makeham (1860) modified the Gompertz model considering some other causes of death independent of age by proposing the so called *Gompertz-Makeham* law of mortality as

$$\mu_x = \gamma + \theta e^{\nu x}, \quad \text{where} \quad (\gamma, \theta, \nu) > 0, \quad x > 0. \tag{2}$$

Here the first term γ (Makeham parameter) is a constant and non-aging component of failure rate (e.g. accidents, independent of age) and the second term $\theta e^{\nu t}$ is the Gompertz function depending on age (aging factor).

The *Weibull* distribution is one of the most widely used distributions in survival analysis and reliability due to the characteristics of its shape parameter ν. The mortality rate or hazard function is

$$\mu_x = \frac{\nu}{\theta^{\nu}} x^{\nu - 1}, \quad \text{for} \quad x \geq 0, \quad (\theta, \nu) > 0. \tag{3}$$

The hazard function of the Weibull distribution can be decreasing, constant or increasing according to the value of its shape parameter i.e. three Weibull models can

make a bathtub shape, but now there are some models like the generalized Weibull model which can have bathtub shape (Bagdonavicius and Nikulin (2002)). The Weibull law is more commonly applicable for technical devices while the Gompertz law is more common for biological systems (Gavrilov & Gavrilova (1991)). When the Gompertz law fails to follow some biological failure mechanism, the best alternative is Weibull law due to its basis on reliability theory. If the probability of failure at the start of the system is almost zero, the failure rate increases with the power function with age i.e. Weibull law and if the system has defects at the beginning, the failure rate increases exponentially with age i.e. Gompertz law. So, to apply the Weibull law in demography, the biological population should be independent of initial deaths. Logistic distribution is considered as the other alternative for Gompertz distribution (Vanfleteren et al. (1998)).

3. Test Statistic for the Table of Mortality

Consider $t = 0$ as the origin of time for an individual of age x, and T_x is a random variable for its residual life from this origin. The probability of death is

$$ {}_tq_x = \mathbf{P}\{0 < T_x \le t\}, \quad t > 0, x > 0. $$

So the annual rate of mortality for the people having age x can be defined as

$$ q_x = \mathbf{P}\{0 < T_x \le 1\}, \quad x > 0. $$

A relation between the rate of mortality and the instantaneous rate of mortality μ_x is

$$ q_x = 1 - \exp\left(-\int_x^{x+1} \mu_y dy \right), \quad x > 0. $$

The theoretical annual rate of mortality in the case of Gompertz model can be written as

$$ q_x = 1 - \exp\left(-\frac{\theta}{\nu} e^{\nu x}(e^{\nu} - 1) \right), \quad \theta, \nu > 0. \tag{4} $$

In the same way we can find the theoretical annual rate of mortality for Makeham, Weibull and other parametric models.

We observe the n persons independent of mortality and we regroup them in the same age, say ω groups, where ω is the maximum age in years. The group G_x contains ℓ_x persons of age x $(x = 0, \cdots, \omega - 1)$ and q_x is the probability of death of each individual in the year. Let denote by D_x the number of deaths in the group G_x.

Using the data D_x and ℓ_x from the table of mortality, we can obtain the empirical annual rate of mortality observed at age x, such that

$$ Q_x = \frac{D_x}{\ell_x}, $$

which follows the binomial law with parameters ℓ_x and q_x. According to the central limit theorem if $min_x(\ell_x) \to \infty$ when $n \to \infty$, then $Q = (Q_0, \cdots, Q_{\omega-1})' \overset{as}{\sim} N_\omega(q, P)$, where $q = (q_0, \cdots, q_{\omega-1})'$ and P is the diagonal matrix of the elements $\frac{q_x(1-q_x)}{\ell_x}$ for $x = 0, 1, \cdots, \omega - 1$. So we can write that

$$ \frac{(D_x - \ell_x q_x)^2}{\ell_x q_x (1 - q_x)} \overset{as}{\sim} \chi_1^2. $$

As it is shown in Gerville-Reache and Nikulin (2000),

$$X_{\omega}^2 = \sum_{x=0}^{\omega-1} \frac{(D_x - \ell_x q_x)^2}{\ell_x q_x (1-q_x)} \overset{as}{\sim} \chi_{\omega}^2.$$

One can use this statistic for testing simple hypotheses, as one uses the classical Pearson statistic for testing simple hypotheses (see Greenwood and Nikulin (1996)).

3.1 Estimation of Parameters in Composite Hypothesis

Consider the composite hypothesis

$$H_0 : q_x = q_x(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} = (\theta_1, \cdots, \theta_s)' \in \Theta \subseteq R^s, s < \omega.$$

We estimate the parameters by the maximum likelihood method using the data from the table of mortality. We have the random variable D_x which follows the binomial law with parameters ℓ_x and q_x. The likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{x=0}^{\omega-1} \binom{\ell_x}{D_x} [q_x(\boldsymbol{\theta})]^{D_x} [1 - q_x(\boldsymbol{\theta})]^{\ell_x - D_x}.$$

We take the estimator $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood function, i.e. $\hat{\boldsymbol{\theta}} = \operatorname{argmax} L(\boldsymbol{\theta})$.

One can find the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ by solving the following score vector

$$\frac{\partial \ln L}{\partial \theta_i} = 0, \forall\, i = 1, \cdots, s.$$

Let consider the statistic

$$X_{\omega}^2(\hat{\boldsymbol{\theta}}) = \sum_{x=1}^{\omega-1} \frac{\left(D_x - \ell_x q_x(\hat{\boldsymbol{\theta}})\right)^2}{\ell_x q_x(\hat{\boldsymbol{\theta}})(1 - q_x(\hat{\boldsymbol{\theta}}))} \overset{as}{\sim} \chi_{\omega-s}^2.$$

Gerville-Reache and Nikulin (2000) showed that under the hypothesis H_0, $X_{\omega}^2(\hat{\boldsymbol{\theta}})$ asymptotically follows a chi-square statistic with $\omega - s$ degrees of freedom, where s is the number of parameters to be estimated, from where it follows that we may use this statistic for testing H_0. One can see that the statistic $X_{\omega}^2(\hat{\boldsymbol{\theta}})$ is different from the classical Pearson statistic.

3.2. Example – Data Analysis from the Table of Mortality

The data in Table 1 is from INSEE Aquitaine-France and give the number of deaths D_x in 1990 for each 5-year age group, where ℓ_x is the number of habitants for each age group on January 1st, 1990. This data is used for the validity of three models i.e. *Gompertz*, *Makeham*, and *Weibull*. The rate of mortality for these three models is adjusted with maximum likelihood estimators and then the value of chi-square is calculated. In case of the adjustment between 5 and 84 years of age, the annual rate of

Table 1: Table of Mortality (INSEE, Gironde, 1990)

age	ℓ_x	D_x	*age*	ℓ_x	D_x
5-9	75498	14	45-49	64575	195
10-14	77284	16	50-54	57974	247
15-19	90337	45	55-59	61871	384
20-24	102544	91	60-64	62473	622
25-29	91339	92	65-69	61122	958
30-34	90769	128	70-74	36425	944
35-39	93324	156	75-79	37124	1341
40-44	96692	226	80-84	29541	2020

mortality follows neither the Gompertz and Makeham nor the Weibull model. But when the adjustment is made for the age groups between 30 and 74 years, the Makeham model is accepted. The Gompertz model also becomes valid with Makeham when the annual rate of mortality is adjusted for the age between 50 and 79 years. This means that the Gompertz model is validated in the older age and it coincides with the theory regarding Gompertz model as discussed in the previous section. The Weibull model gets close but still it does not fit the data significantly. The calculated values of the test statistic with corresponding p-values are shown in Table 2 and the fitted models are presented in Figures 1-3.

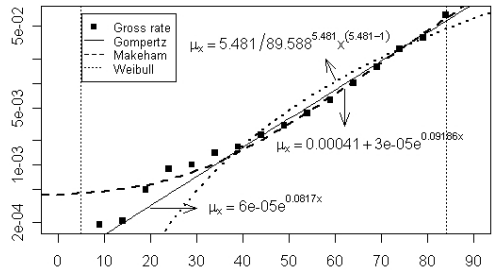

Figure 1: Models fitted for age between 5 and 84 years (log scale)

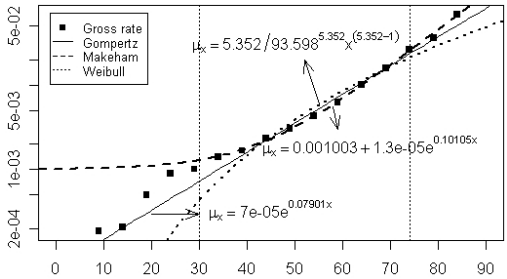

Figure 2: Models fitted for age between 30 and 74 years (log scale)

Table 2: Results from the Table of Mortality

	Gompertz		Makeham		Weibull	
Age Groups	$X^2_\omega(\hat{\boldsymbol{\theta}})$	p-value	$X^2_\omega(\hat{\boldsymbol{\theta}})$	p-value	$X^2_\omega(\hat{\boldsymbol{\theta}})$	p-value
5-84	214.19	≈ 0	99.98	≈ 0	2363.98	≈ 0
30-74	45.62	≈ 0	3.70	0.72	158.93	≈ 0
50-79	9.01	0.11	8.48	0.08	25.44735	0.0001

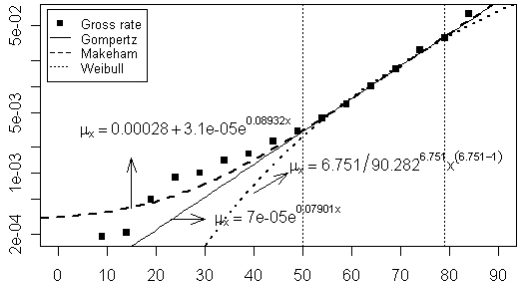

Figure 3: Models fitted for age between 50 and 79 years (log scale)

4. Goodness-of-fit Test for Right Censored Data

Here we apply the survival analysis methods in demography when we have individuals' information with right censoring. For this purpose Bagdonavicius et al. (2010) proposed a chi-squared type goodness of fit test based on the NRR statistic (Bagdonavicius and Nikulin (2011)). We give a chi-squared type test for testing composite parametric hypothesis when individual data are right censored.
Let us consider the composite hypothesis

$$H_0 : F(x) = F(x,\boldsymbol{\theta}), \quad x \in R^1, \qquad \boldsymbol{\theta} = (\theta_1,\cdots,\theta_s)' \in \Theta \subset R^s$$

i.e. the distribution of failure times *T* belongs to the given parametric class. Here we consider *Gompertz*, *Makeham*, and *Weibull* as parametric families.
Suppose we have right censored individual data as

$$(X_1,\delta_1),\ldots,(X_n,\delta_n), \quad X_i = T_i \wedge C_i, \quad \delta_i = \mathbf{1}_{\{T_i \le C_i\}}, \tag{5}$$

where $T_1,\cdots,T_n$ are the failure times which are absolutely continuous i.i.d. random variables and $C_1,\cdots,C_n$ are the censoring times which are independent. The probability density function of the random variable T_1 belongs to a parametric family $\{f(\cdot,\boldsymbol{\theta}),\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^s\}$. Denote by

$$S_t(\boldsymbol{\theta}) = S(t,\boldsymbol{\theta}) = \mathbf{P}_{\boldsymbol{\theta}}\{T_1 > t\} \;\&\; \Lambda_t(\boldsymbol{\theta}) = -lnS_t(\boldsymbol{\theta}) = \int_0^t \boldsymbol{\mu}_y(\boldsymbol{\theta})dy, \;\; \boldsymbol{\theta} \in \Theta,$$

the survival function and the cumulative hazard function, respectively. In demographic literature *t* is used in the subscript. With *non-informative* random censoring mechanism the loglikelihood function can be written as,

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \ln f_{X_i}(\boldsymbol{\theta}) + \sum_{i=1}^{n} S_{X_i}(\boldsymbol{\theta}) \quad \boldsymbol{\theta} \in \Theta \tag{6}$$

ML estimators can be found by equating the score vector $\dot{\ell}(\boldsymbol{\theta})$ to zero and the Fisher's information matrix is $\mathbf{I}(\boldsymbol{\theta}) = -\mathbf{E}_{\boldsymbol{\theta}} \ddot{\ell}(\boldsymbol{\theta})$. Consistency and asymptotic normality of the ML estimators $\hat{\boldsymbol{\theta}}$ hold under some regularity sufficient conditions (Hjort (1990), Bagdonavicius et al. (2010)).

To construct the test we introduce two counting processes and write the censored sample (5) as

$$(N_1(t), Y_1(t), t \geq 0), \cdots, (N_n(t), Y_n(t), t \geq 0), \tag{7}$$

where $N_i(t) = \mathbf{1}_{\{X_i \leq t, \delta_i = 1\}}$, $Y_i(t) = \mathbf{1}_{\{X_i \geq t\}}$,

$$N(t) = \sum_{i=1}^{n} N_i(t), \quad \text{and} \quad Y(t) = \sum_{i=1}^{n} Y_i(t).$$

Suppose that the processes N_i, Y_i are observed at finite time τ (time of experiment). Using these data one can calculate immediately the non-parametric Nelson-Aalen estimator and the Kaplan-Meier estimator

$$\hat{\Lambda}(t) = \int_0^t \frac{dN(u)}{Y(u)}, \qquad \hat{S}(t) = \hat{S}(t-)\left(1 - \frac{\Delta N(t)}{Y(t)}\right)$$

for the unknown cumulative hazard function Λ, and for the survival function *S(t)* respectively, when the censored data are ungrouped, where *S(t-)* is the value of survival function just before time *t*.

Let us divide the interval $[0, \tau]$ into $k > s$ smaller intervals

$$I_j = (a_{j-1}, a_j], \quad j = 1, 2, ..., k, \quad a_0 = 0, \quad a_k = \max(X_{(n)}, \tau).$$

We denote the number of observed failures in the j-th interval by

$$U_j = N(a_j) - N(a_{j-1}) = \sum_{i: X_i \in I_j} \delta_i.$$

The choice of random grouping intervals $\hat{a}_j$ is made to overcome the problem of very small expected number of events for some interval. This can happen in demography because the number of deaths at early age is very small. Set

$$b_i = (n-i)\Lambda_{X_{(i)}}(\hat{\boldsymbol{\theta}}) + \sum_{l=1}^{i} \Lambda_{X_{(l)}}(\hat{\boldsymbol{\theta}})$$

where $X_{(i)}$ is the i[th] element in the ordered statistics $(X_{(1)}, \cdots, X_{(n)})$. If *i* is the smallest natural number satisfying $E_j \in [b_{i-1}, b_i]$, $j = 1, \cdots, k-1$, then

$$(n-i+1)\Lambda_a(\hat{\boldsymbol{\theta}}) + \sum_{l=1}^{i-1} \Lambda_{X_{(l)}}(\hat{\boldsymbol{\theta}}) = E_j$$

and

$$\hat{a}_j = \Lambda^{-1}\left([E_j - \sum_{l=1}^{i-1} \Lambda_{X_{(l)}}(\hat{\boldsymbol{\theta}})] / (n-i+1), \hat{\boldsymbol{\theta}}\right), \qquad \hat{a}_k = \max(X_{(n)}, \tau),$$

where Λ^{-1} is the inverse of the cumulative hazard function Λ. We have $0 < \hat{a}_1 < \hat{a}_2 < \cdots < \hat{a}_k = \max(X_{(n)}, \tau)$. With this choice of intervals the expected number of failures is $e_j = E_k / k$, for any j where $E_k = \sum_{i=1}^{n} \Lambda_{X_{(i)}}(\hat{\boldsymbol{\theta}})$. Bagdonavicius et al. (2010) and Greenwood and Nikulin (1996) give some recommendations for the choice of intervals. If there is no alterative hypothesis, the number of intervals k is such that $n / k > 5$.

For testing H_0, Bagdonavicius et al. (2010) considered the following statistic

$$Y_n^2 = \mathbf{Z}' \widehat{\mathbf{V}}^{-} \mathbf{Z},$$

where

$$\widehat{\mathbf{V}}^{-} = \widehat{\mathbf{A}}^{-1} + \widehat{\mathbf{A}}^{-1} \widehat{\mathbf{C}}' \widehat{\mathbf{G}}^{-} \widehat{\mathbf{C}} \widehat{\mathbf{A}}^{-1}, \quad \widehat{\mathbf{G}} = \hat{\mathbf{i}} - \widehat{\mathbf{C}} \widehat{\mathbf{A}}^{-1} \widehat{\mathbf{C}}',$$

is a consistent estimator of a generalized inverse $\mathbf{V}^{-}$ of the asymptotic variance-covariance matrix $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta})$ of the statistic $\mathbf{Z} = (Z_1, ..., Z_k)'$, where $Z_j = \frac{1}{\sqrt{n}}(U_j - e_j), \quad j = 1, ..., k.$ So the test statistic can be written in a simple form as

$$Y_n^2 = \sum_{j=1}^{k} \frac{(U_j - e_j)^2}{U_j} + Q, \tag{8}$$

where

$$Q = \mathbf{W}' \widehat{\mathbf{G}}^{-1} \mathbf{W}, \quad \mathbf{W} = \widehat{\mathbf{C}} \widehat{\mathbf{A}}^{-1} \mathbf{Z} = (\mathrm{W}_1, ..., \mathrm{W}_s)',$$

$$\widehat{\mathbf{G}} = [\hat{g}_{ll'}]_{s \times s}, \quad \hat{g}_{ll'} = \hat{i}_{ll'} - \sum_{j=1}^{k} \hat{C}_{lj} \hat{C}_{l'j} \hat{A}_j^{-1}, \quad W_l = \sum_{j=1}^{k} \hat{C}_{lj} \hat{A}_j^{-1} Z_j.$$

$$\hat{\mathbf{i}}_{ll'} = \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\partial \ln \mu_{X_i}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_l} \frac{\partial \mu_{X_i}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}_{l'}}, \quad \widehat{\mathbf{C}}_{lj} = \frac{1}{n} \sum_{i: X_i \in I_j} \delta_i \frac{\partial}{\partial \boldsymbol{\theta}} \ln \mu_{X_i}(\hat{\boldsymbol{\theta}}),$$

$$\hat{A}_j = U_j / n, \quad U_j = \sum_{i: X_i \in I_j} \delta_i, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j),$$

$i = 1, \cdots, n, \quad j = 1, \cdots, k, \quad l, l' = 1, \cdots, s.$

Denote by $\hat{g}^{ll'}$ the elements of $\widehat{\mathbf{G}}^{-}$. The quadratic form Q can be written as follows:

$$Q = \sum_{l=1}^{s} \sum_{l'=1}^{s} W_l g^{ll'} W_{l'}.$$

The limiting distribution of the statistic Y_n^2 is chi-square with $r = rank(\mathbf{V}^{-}) = Tr(\mathbf{V}^{-}\mathbf{V})$ degrees of freedom. If G is non-degenerate then $r = k$.

Statistical inference for the hypothesis H_0 : The hypothesis is rejected with approximate significance level α if $Y_n^2 > \chi_\alpha^2(r)$.

4.1. Goodness-of-Fit Tests for Gompertz and Weibull Models

Let consider the hypothesis that under H_0 the distribution of the failure times is *Gompertz* with hazard function and cumulative hazard function given by

$$\mu_x = \theta e^{\nu x}, \quad \Lambda_x = \frac{\theta}{\nu}(e^{\nu x} - 1) \quad x > 0, \quad (\theta, \nu) > 0.$$

The loglikelihood function is

$$\ell(\theta, \nu) = \sum_{i=1}^{n} \left\{ \delta_i [\ln \theta + \nu X_i] - \frac{\theta}{\nu}(e^{\nu X_i} - 1) \right\}.$$

We denote by $\hat{\theta}$ and $\hat{\nu}$ the ML estimators of θ and ν.

Since the matrix *G* is found to be degenerated, the quadratic form can be written as:

$$Q = \frac{W_2^2}{\hat{g}_{22}},$$

where

$$\hat{g}_{22} = \hat{i}_{22} - \sum_{j=1}^{k} \hat{C}_{2j}^2 \hat{A}_j^{-1}, \quad \hat{i}_{22} = \frac{1}{n}\sum_{i=1}^{n} \delta_i X_i^2, \quad \hat{C}_{2j} = \frac{1}{n} \sum_{i: X_i \in I_j} \delta_i X_i,$$

$$\hat{A}_j = \frac{U_j}{n}, \quad W_2 = \sum_{j=1}^{k} \hat{C}_{2j} \hat{A}_j^{-1} Z_j, \quad Z_j = \frac{1}{\sqrt{n}}(U_j - e_j).$$

Choice of $\hat{a}_j$: Set

$$b_i = (n-i)\frac{\hat{\theta}}{\hat{\nu}}(e^{\hat{\nu} X_{(i)}} - 1) + \frac{\hat{\theta}}{\hat{\nu}} \sum_{l=1}^{i} (e^{\hat{\nu} X_{(l)}} - 1), \quad i = 1, \cdots, n.$$

If *i* is the smallest natural number satisfying the inequalities

$$b_{i-1} \le E_j \le b_i, \qquad E_j = \frac{j}{k} b_n,$$

then for $j = 1, \cdots, k-1$

$$\hat{a}_j = \frac{1}{\hat{\nu}} \ln \left\{ 1 + \frac{\hat{\nu}}{\hat{\theta}} \left(\frac{j}{k} b_n - \frac{\hat{\theta}}{\hat{\nu}} \sum_{l=1}^{i-1} (e^{\hat{\nu} X_{(l)}} - 1) \right) / (n - i + 1) \right\}, \quad \hat{a}_k = \max(X_{(n)}, \tau).$$

For such choices of intervals we have $e_j = E_k / k$ for any *j*.

Example: This example is taken from the book of Bagdonavicius et al. (2010). n = 120 electronic devices were observed for time τ = 5.54 (years). The number of failures is δ =113:

1.7440, 1.9172, 2.1461, 2.3079, 2.3753, 2.3858, 2.4147, 2.5404, 2.6205, 2.6471,
2.8370, 2.8373, 2.8766, 2.9888, 3.0720, 3.1586, 3.1730, 3.2132, 3.2323, 3.3492,

3.3507, 3.3514, 3.3625, 3.3802, 3.3855, 3.4012, 3.4382, 3.4438, 3.4684, 3.5019, 3.5110, 3.5297, 3.5363, 3.5587, 3.5846, 3.5992, 3.6540, 3.6574, 3.6674, 3.7062, 3.7157, 3.7288, 3.7502, 3.7823, 3.8848, 3.8902, 3.9113, 3.9468, 3.9551, 3.9728, 3.9787, 3.9903, 4.0078, 4.0646, 4.1301, 4.1427, 4.2300, 4.2312, 4.2525, 4.2581, 4.2885, 4.2919, 4.2970, 4.3666, 4.3918, 4.4365, 4.4919, 4.4932, 4.5388, 4.5826, 4.5992, 4.6001, 4.6324, 4.6400, 4.7164, 4.7300, 4.7881, 4.7969, 4.8009, 4.8351, 4.8406, 4.8532, 4.8619, 4.8635, 4.8679, 4.8858, 4.8928, 4.9466, 4.9846, 5.0008, 5.0144, 5.0517, 5.0898, 5.0929, 5.0951, 5.1023, 5.1219, 5.1223, 5.1710, 5.1766, 5.1816, 5.2441, 5.2546, 5.3353, 5.4291, 5.4360, 5.4633, 5.4842, 5.4860, 5.4903, 5.5199, 5.5232, 5.5335.

Suppose the failure times have a *Gompertz* distribution. The maximum likelihood estimators of *Gompertz model* are: $\hat{\theta} = 0.0051$, $\hat{\nu} = 1.1586$. We take 10 intervals i.e. k=10. Further results to calculate Y_n^2 are shown below:

J	1	2	3	4	5	6	7	8	9	10
$\hat{a}_j$	2.70	3.33	3.74	4.07	4.34	4.57	4.78	5.00	5.25	5.54
U_j	10	9	23	12	9	6	7	13	13	11
e_j	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3

$$\hat{i}_{22} = 16.7779, \quad \hat{g}_{22} = 0.0141, \quad W_2 = -0.3737.$$

The matrix G is degenerate, so $r = k - 1 = 9$. The value of the test statistic is $Y_n^2 = X^2 + Q = 15.1130 + 9.8867 = 24.9997$ and the $p\text{-}value = P\{\chi_9^2 > 24.9997\} =$ 0.0053. So from the result we can say that failure times don't follow the Gompertz distribution.

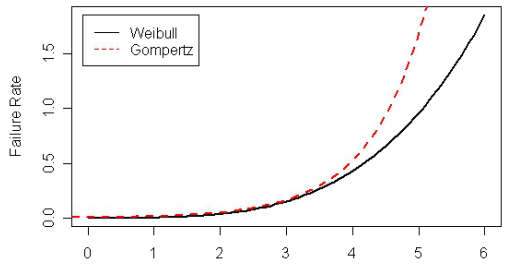

Figure 4: The failure rate of electronic devices

Suppose that the failure times follow a *Weibull model*. The maximum likelihood estimators of the Weibull model are: $\hat{\theta} = 4.6078$, $\hat{\nu} = 4.9554$. We take 10 intervals i.e. *k=10*. Further results to calculate Y_n^2 are shown below:

J	1	2	3	4	5	6	7	8	9	10
$\hat{a}_j$	2.89	3.36	3.70	3.98	4.24	4.47	4.68	4.90	5.16	5.54
U_j	13	9	17	12	7	8	8	13	11	15
e_j	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3	11.3

$$\hat{i}_{22} = 0.0618, \quad \hat{g}_{22} = 0.0027, \quad W_2 = -0.0545.$$

The matrix G is degenerate, so $r=k-1=9$. The value of test statistic is $Y_n^2 = X^2 + Q = 9.2692 + 1.0845 = 10.3536$ and the $p\text{-}value = P\{\chi_9^2 > 10.3536\} = 0.3226$. So from the result we have no reason to reject that the failure times follow the Weibull distribution. In the same way we can apply the test for Makeham model.

Here the Weibull model gives the better fit which is expected since the data refer to technical devices and according to Gavrilov and Gavrilova (2001) technical devices fail according to the Weibull law. Also from Figure 4 one can observe the behavior of the Gompertz model, according to which in later times the failure rate increases very fast.

5. Non-Parametric Estimation of the Survival Function with Data from the Table of Mortality

If there is no information about the model, one can estimate the survival function by using a non-parametric estimation method. In the case of right censored individual sample (5), it is easy to use the well known Kaplan-Meier estimation method for estimating the survival function S_t and consequently the distribution function $1-S_t$ of the failure times.

One can estimate the survival function S_t from the grouped data, for example from the table of mortality with censoring in the following way.

Suppose we observe n_0 individuals and the time scale is divided in k intervals

$$[a_0, a_1[, [a_1, a_2[, ..., [a_{j-1}, a_j[, ..., [a_{k-1}, a_k[. \tag{9}$$

Consider the j[th] interval $I_j = [a_{j-1}, a_j[$, $j = 1, ..., k$, $a_0 = 0$, $a_k = +\infty$.

We observe d_j - the number of deaths in the interval I_j, $j = 1, ..., k$, c_j - the number of individuals censored in the interval I_j, $j = 1, ..., k$, and n_j - the number of individuals at risk (not died and not censored) at time a_j, (the number of individuals who enter in the I_{j+1}-th interval). So

$$n_j = n_{j-1} - d_j - c_j. \tag{10}$$

Let r_j - the number of individuals at death risk in the interval I_j. If all censoring is at the start of the interval I_j, then $r_j = n_{j-1} - c_j$. If all censored are at the end of the interval I_j, then $r_j = n_{j-1}$. But the censoring times are unknown. Therefore we suppose that the censored data are uniformly distributed in the interval I_j and hence we take

$$r_j = n_{j-1} - c_j / 2. \tag{11}$$

Let us denote by

$$q_j = \mathbf{P}\{T > a_j \mid T > a_{j-1}\} = \frac{\mathbf{P}\{T > a_j\}}{\mathbf{P}\{T > a_{j-1}\}},$$

the conditional probability of being alive at a_j given that it was alive at a_{j-1}, and hence

$$\mathbf{P}\{T > a_j\} = q_j \cdot \mathbf{P}\{T > a_{j-1}\} = q_j \cdot q_{j-1} \mathbf{P}\{T > a_{j-2}\} = \cdots = \prod_{i=1}^{j} q_i .$$

So we have

$$S_j = \mathbf{P}\{T > a_j\} = \prod_{i=1}^{j} q_i .$$

Similarly

$$p_j = 1 - q_j = \mathbf{P}\left[T \le a_j \mid T > a_{j-1}\right]$$

is the conditional probability of death in the interval I_j given that it was alive at a_{j-1}. So we can write

$$S_j = \prod_{i=1}^{j} (1 - p_i). \tag{12}$$

The number of deaths d_i in the interval I_i follows approximately binomial law with parameters r_i and p_i, so $\mathbf{E}d_i \approx r_i p_i$ and the probability p_i is estimated by $\hat{p}_i = d_i / r_i$. So the survival function $S_j = S(t_j)$ is estimated by

$$\hat{S}_j = \prod_{i=1}^{j} (1 - \hat{p}_i) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{r_i}\right) = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i - c_i / 2}\right). \tag{13}$$

This is the analog of Kaplan-Meier estimator for grouped data and it is also called product limit estimator.

References

1. Akritas, M.G. (1988). Pearson-type goodness of fit tests: the univariate case. J. Amer. Statist. Assoc., 83, p. 222-230.
2. Bagdonavicius, V. and Nikulin, M. (2002). Accelerated Life Models. Chapman and Hall/CRC, Boca Raton.
3. Bagdonavicius, V., Kruopis, J. and Nikulin M. (2010). Nonparametric tests for Censored Data. ISTE and J.Wiley.
4. Bagdonavicius, V. and Nikulin, M. (2011). Chi-square goodness-of-fit test for right censored data. Int. J. Appl. Math. Stat. Vol.24, Issue No.SI-11A, p. 30-50.
5. Drost, F. (1988). Asymptotics for generalized chi-squared goodness-of-fit tests. CWI Tracs, Amsterdam: Centre for Mathematics and Computer Sciences, V.48.
6. Gavrilov, L.A. and Gavrilova, N. S. (1991). The Biology of Life Span: A Quantitative Approach. New York: Harwood Academic Publisher.

7. Gavrilov L.A. and Gavrilova N.S. (2001). The reliability theory of aging and longevity. Journal of Theoretical Biology, 213(4), p. 527-545.
8. Gerville-Réache L. and Nikulin, M. (2000). Analyse statistique du modele de Makeham. Revue Roumaine Math. Pure et Appl., V.45, No.6, p. 947-957.
9. Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality. Philosophical Transactions of the Royal Society of London, Ser. A, 115, 513-585, reprint: Haberman, S. and Sibbett, TA (1995), Vol. II, p. 119-191.
10. Greenwood, P. and Nikulin, M.S. (1996). A Guide to Chi-squared Testing. Wiley, New York.
11. Hjort, N.L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates. The Annals of Statistics, v.18,\#3, p. 1221-1258.
12. Juckett DA. and Rosenberg B. (1993). Comparison of the Gompertz and Weibull functions as descriptors for human mortality distributions and their intersections. Mech Ageing Dev., 69, p. 1–31.
13. Kim, J.H. (1993). Chi-square goodness-of-fit tests for randomly censored data. Annals of Statistics, V.21, p. 1621-39.
14. LeCam, L., Mahan,C. and Singh, A. (1983). An extension of a Theorem of H. Chernoff and E.L. Lehmann. In: Recent advances in statistics, Academic Press, Orlando, p. 303-332.
15. Nikulin, M.S. (1973). On a chi-square test for continuous distribution. Theory of Probability and its Application, 19, No.3, p. 638-639.
16. Nikulin, M.S. (2011). Modeles statistiques en demographie, gerontologie et assurance sur la vie. Lecture Notes for IMB Groupe de Fiabilité et Statistique, Université Victor Segalen Bordeaux 2, p 67.
17. Nikulin, M.S., Anisimov, V.N. and Nikulin, A.M. (2011). Statistical Models of Longevity, Aging and Degradation in Demography, Gerontology and Oncology. Advances in Gerontology, V.24, No.3, p. 366-379.
18. Rao, K.C. and Robson, D.S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. Communication in Statist., 3, p. 1139-1153.
19. Ricklefs, R.E. and A. Scheuerlein (2002). Biological implications of the Weibull and Gompertz model of aging. Journals of Gerontology: Biological Sciences 57A, 2 B69-B76.
20. Van der Vaart, A.W. (1998). Asymptotic Statistics. Cambridge: UK.
21. Vanfleteren, J.R., De Vreese, A. and Braeckman, B. P. (1998). Two-parameter logistic and Weibull equations provide better fits to survival data from isogenic populations of Caenorhabditis elegans in axenic culture than does the Gompertz model. J. Gerontol. Ser. A 53, B 393-403.
22. Wilson, D.L. (1994). The analysis of survival (mortality) data: Fitting Gompertz, Weibull, and logistic functions. Mechanisms of Ageing and Development. 74, p. 15-33.