QUALITATIVE ASPECTS OF THE MIN PARETO BINOMIAL DISTRIBUTION

Bogdan Gheorghe MUNTEANU

"Henri Coandă" Air Force Academy, Braşov, Romania (munteanu.b@afahc.ro)

DOI: 10.19062/1842-9238.2017.15.2.8

Abstract: *Centered upon statistical models relating to qualitative aspects, the following paper sets out to demonstrate that by means of the Akaike information criterion (AIC) a statistical selection of the MinParB distribution for different parameter values can be obtained based on the statistical simulation algorithm of the power series distribution, called the Min Pareto Binomial [4], and the EM algorithm for the statistical estimation of the parameters of the MinParB distribution. The determining of the MinParB distribution from a unitary perspective regarding the class of the power series distributions [2] has also been taken into consideration.*

Keywords: *power series distributions, Pareto distribution, Binomial distribution, distribution of the minimum, EM algorithm, information criterion.*

1. INTRODUCTION

According to the Pareto principle, also known as the "80/20 rule", in the case of events, about 80% of the effects is generated by 20% of the causes. Management consultant Joseph M. Juran was the first to suggest this principle, which he named after the Italian economist Vilfredo Pareto, who identified the well-known 80/20 ratio. Basically, Pareto demonstrated that about 80% of the land in Italy was owned by 20% of the population. In business, the same basic rule applies (for example, 80% of sales come from 20% of clients) [1]. Similarly, for a given set of parameters, in the case of natural phenomena, the existence of an empirically obtained Pareto distribution has been observed [2].

The Pareto distribution is particularly used in situations in which there is a high probability of paying large sums in compensation, namely liability insurance.

Let $(X_i)_{i\geq 1}$ be a sequence of a number of independent and identically distributed random variables, $X_i \sim Par(\mu,\alpha),\ \mu,\alpha>0$ with the cumulative distribution function $F_{X_i}(x)=F_{Par}(x)=1-\left(\frac{\mu}{x}\right)^{\alpha},\ x\geq\mu$ and the probability density function $f_{X_i}(x)=f_{Par}(x)=\frac{\alpha\mu^{\alpha}}{x^{\alpha+1}},\ x\geq\mu$.

Also, we denote by $V_{Par}=\min\{X_1,X_2,...,X_Z\}$, where random variable $Z\sim Binom(n,p),\ n\in\{1,2,...\},\ p\in(0,1)$, and $(X_i)_{i\geq 1}$ are independent and Pareto distributed random variables. The cumulative distribution function, the probability density function and some reliability characteristics of the random variable V_{Par} are given in the paper [4].

The random variable V_{Par} generates two events:

- The event to minimize the amounts claimed $(X_i)_{i=\overline{1,Z}}$ regarding the civil liability insurance;
- The number of the claimed amounts Z represents the number of successes out of the n independent events with the probability of success p.

Therefore, we discuss the distribution $MinParB(\mu, \alpha, n, p)$, $\mu, \alpha > 0$, $p \in (0,1)$, $n \in \{1,2,...\}$. The numerical characteristics of this distribution are presented in the paper [4].

2. INFORMATION CRITERION

The common approach to model selection involves choosing a model that minimizes one or several information criteria applied to a set of statistical models [1],[5].

The commonly used information criteria are: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC) and consistent Akaike information criteria (CAIC).

Each criterion is a sum of two terms: the first term characterizes the entropy rate or model prediction error, whereas the second one describes the number of the free parameters estimated based on the model [2].

2.1. Akaike Information Criterion. The Akaike Information Criterion (AIC) is a criterion for selecting from nested (overlapping) economic models. Basically, AIC is a measure estimating the quality of each studied economic model, since they relate to one another for a given set of data. Therefore, AIC is an ideal method for selecting the model.

Discovered and put forward by Professor Hirotugu Akaike in 1971, respectively in 1974, the AIC was defined as a measure of matching the statistical model.

AIC is an associated number for each separate model, as follows:

$$AIC = -2L(x, \hat{\Theta}) + 2q, \tag{1}$$

where $L(x, \hat{\Theta})$ represents the maximum likelihood function, $\hat{\Theta} = (\hat{\alpha}, \hat{p})$ the parameter vector estimated by applying the EM algorithm [4], and q represents the number of parameters of the statistical model. In our case, $q = 2$.

Therefore, for the set of AIC values corresponding to each particular economic model, the preferred one in terms of relative quality is the model with the minimum value. ($AIC_{\min}$).

The loss of information when the statistical model that has been studied and analysed in relation to the best estimated model is given by:

$$\Delta_i = AIC_{\min} - AIC_i, \tag{2}$$

where i is the number of statistical models to which AIC has been applied, and $AIC_{\min}$ stands for the minimal value AIC out of the values' vector.

2.2. Bayesian Information Criterion. Bayesian information criterion (BIC) is also a mathematical tool applied to statistical models from the economic field. It is a criterion similar to AIC. The BIC or the Schwartz criterion (1978) is a number characterized by the relation:

$$BIC = -2L(x,\hat{\Theta}) + q\ln(m), \tag{3}$$

where $L(x,\Theta)$ represents the maximum likelihood function, $\hat{\Theta} = (\hat{\alpha}, \hat{p})$ is the parameter vector estimated as a result of applying the EM algorithm [4], q represents the number of parameters of the statistical model ($q = 2$), and m characterizes the volume of the statistical data.

2.3. Hannan-Quinn Information Criterion. The Hannan-Quinn Information Criterion (HQIC) is an information criterion that is used alternatively with AIC and BIC. The criterion is represented by the number:

$$HQIC = -2L(x,\hat{\Theta}) + 2q\ln(\ln(m)), \tag{4}$$

where $L(x,\hat{\Theta})$, q and m have the same interpretations as the AIC and BIC criteria.

2.4. Consistent Akaike Information Criteria. Consistent Akaike Information Criteria (CAIC) is, essentially, a correction to the Akaike Information Criterion (AIC), this being characterized by the relation:

$$CAIC = AIC + \frac{2q(q-1)}{m-q-1} \tag{5}$$

or

$$CAIC = -2L(x,\hat{\Theta}) + \frac{2qm}{m-q-1}, \tag{6}$$

where $L(x,\hat{\Theta})$, q and m have the same interpretations as the AIC and BIC criteria.

3. APPLICATIONS

According to the studies in paper [4], the logarithm maximum likelihood function is defined as follows:

$$\ln L(x,\hat{\Theta}) = m\left(\ln n + \ln \hat{p} + \ln \hat{\alpha} + \hat{\alpha}\ln \mu\right) - m\ln\left[1-(1-\hat{p})^n\right] + \sum_{j=1}^{m}\left\{(n-1)\ln\left[1-\hat{p}+\hat{p}\left(\frac{\mu}{x_j}\right)^{\hat{\alpha}}\right] - (\hat{\alpha}+1)\ln x_j\right\}, \tag{4}$$

where $\hat{\Theta} = (\hat{\alpha}, \hat{p})$ the parameter vector estimated as a result of applying the EM algorithm.

The step-by-step description of the algorithm is included in paper [4] and implemented in the GUI Octave 1.5.4 programming environment.

The values of the estimated parameters, as well as the AIC values are shown in Tables 1 and 2 for sample values $m = 100$, the parameters of the Pareto distribution $\mu = 1$ and $\alpha \in \{0,5;1;3;10\}$, and for Binomial distribution parameters $n \in \{4;40\}$ and $p \in \{0,2;0,5;0,9\}$.

Also, in Tables 1 and 2 the values of the AIC, BIC, HQIC, CAIC are expressed. These values have been obtained by means of the EXCEL computing environment.

Based on the numerical values in Tables 1 and 2, the following situations are represented:

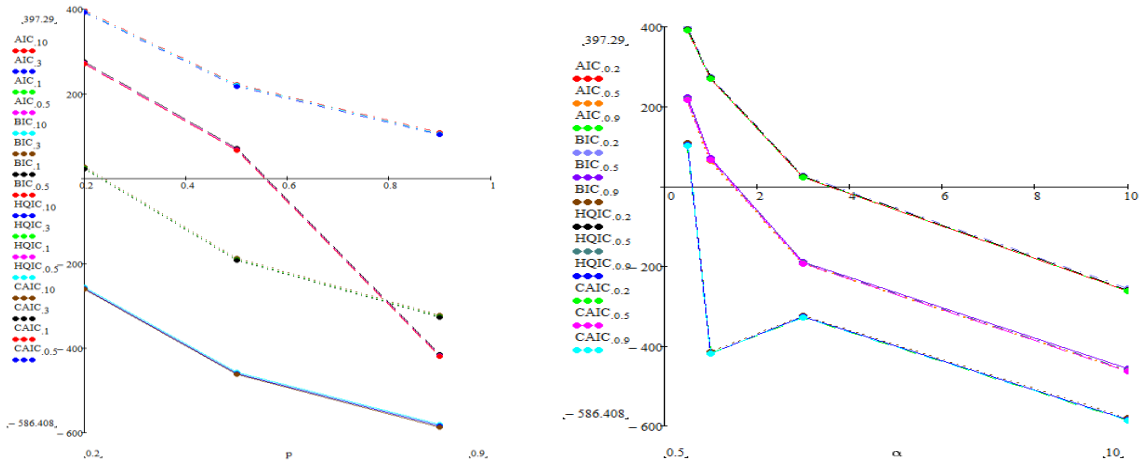
- The values of the information criteria (AIC, BIC, HQIC, CAIC) according to the estimated parameters $\hat{\Theta} = (\hat{\alpha}, \hat{p})$, (Fig. 1);
- The comparative graphical analysis (based on the value categories of α) of the information criteria values in relation to the values of the parameter p (Fig. 2).

Table1. Estimated parameter values and the AIC, BIC, HQIC, CAIC values for $MinParB(1, \alpha, 4, p)$

$(\alpha; p)$	$\hat{\alpha}$	$\hat{p}$	$\hat{h}$	*AIC*	*BIC*	*HQIC*	*CAIC*
(10;0,2)	10,633	0,025	1159	-261,302	-256,091	-259,193	-261,178
(10;0,5)	10,625	0,446	712	-462,254	-457,044	-460,146	-462,131
(10;0,9)	11,033	0,724	457	-586,408	-581,197	-584,299	-586,284
(3;0,2)	3,188	0,025	855	22,376	27,586	24,485	22,500
(3;0,5)	3,187	0,446	588	-193,921	-188,711	-191,813	-193,798
(3;0,9)	3,310	0,724	392	-328,444	-323,233	-326,335	-328,320
(1;0,2)	1,062	0,027	646	269,888	275,098	271,996	270,011
(1;0,5)	1,062	0,446	474	65,457	70,668	67,566	65,581
(1;0,9)	1,103	0,724	333	-420,047	-414,837	-417,938	-419,923
(0,5;0,2)	0,531	0,027	646	392,080	397,290	394,189	392,204
(0,5;0,5)	0,530	0,447	426	217,308	222,518	219,417	217,432
(0,5;0,9)	0,551	0,724	302	102,860	108,070	104,968	102,983

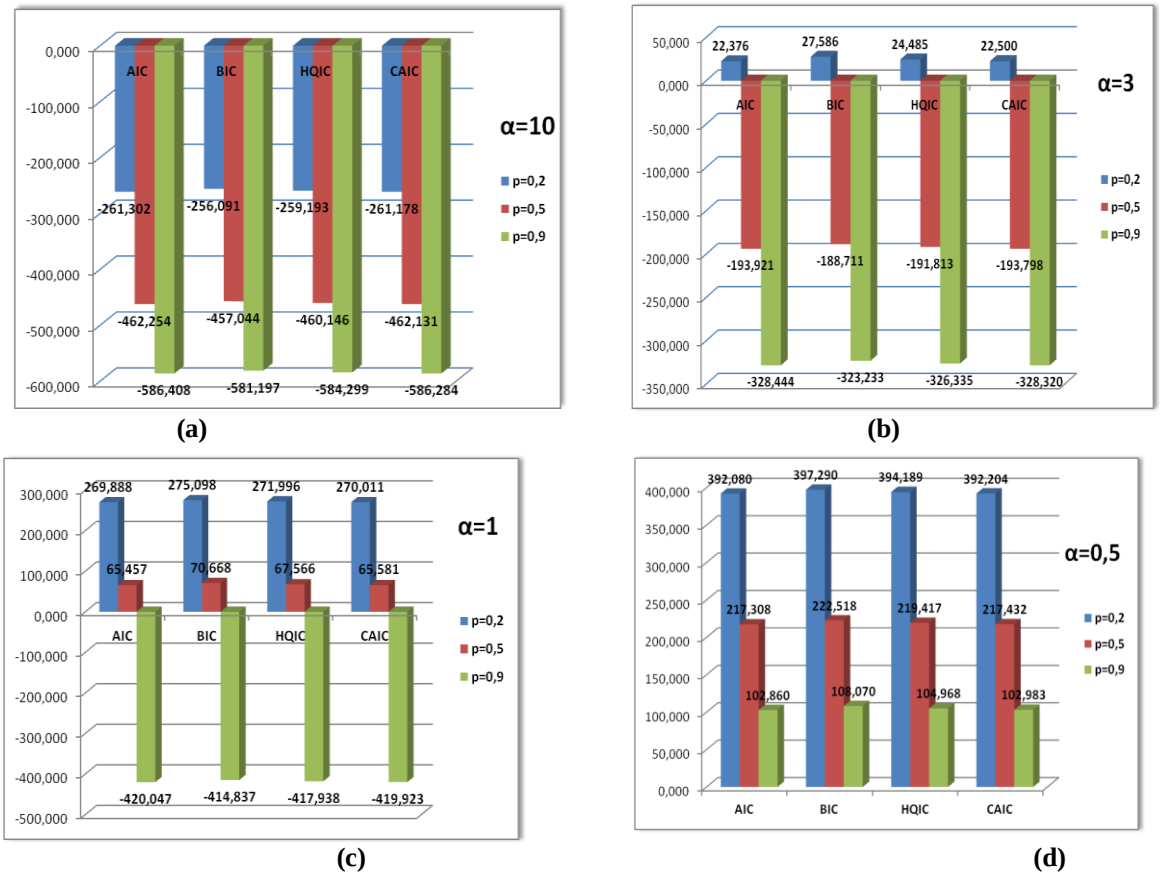
Table 2. Estimated parameter values and AIC, BIC, HQIC, CAIC values for $MinParB(1, \alpha, 40, p)$

$(\alpha; p)$	$\hat{\alpha}$	$\hat{p}$	$\hat{h}$	*AIC*	*BIC*	*HQIC*	*CAIC*
(10;0,2)	14,575	0,142	1339	-1,078E+71	-1,078E+71	-1,078E+71	-1,078E+71
(10;0,5)	8,831	0,585	5001	-7,134E+50	-7,134E+50	-7,134E+50	-7,134E+50
(10;0,9)	10,116	0,832	5001	-2,937E+53	-2,937E+53	-2,937E+53	-2,937E+53
(3;0,2)	4,371	0,142	1149	-584,960	-579,750	-582,852	-584,837
(3;0,5)	2,649	0,585	5001	-809,318	-804,108	-807,210	-809,195
(3;0,9)	3,035	0,832	5001	-910,978	-905,768	-908,869	-910,854
(1;0,2)	1,456	0,142	975	-345,206	-339,996	-343,097	-345,082
(1;0,5)	0,883	0,585	5001	-583,025	-577,815	-580,916	-582,901
(1;0,9)	1,101	0,832	5001	-703,705	-698,494	-701,596	-703,581
(0,5;0,2)	0,727	0,142	865	-184,485	-179,275	-182,376	-184,361
(0,5;0,5)	0,442	0,585	5001	-434,522	-429,312	-432,413	-423,398
(0,5;0,9)	0,506	0,832	5001	-576,236	-571,025	-574,127	-576,112



(a) **(b)**

FIG.1. (a) AIC depending on the probability p and $\alpha \in \{0,5;1;3;10\}$ established; (b) AIC depending on α and $p \in \{0,5;1;3;10\}$ for the distribution $MinParB(1,\alpha,4,p)$



(c) **(d)**

FIG.2. The values of the AIC, BIC, HQIC, CAIC depending on the probability p and $\alpha \in \{0,5;1;3;10\}$ established in situations (a), (b), (c), respectively (d) for the distribution $MinParB(1,\alpha,4,p)$ **.** Comparative graphic analysis

CONCLUSIONS

The main objective of this study is to perform a quantitative analysis of the statistical model as described in paper [4] in a unitary manner and from the perspective of the power series distribution class [3]. The values of the main information criteria (AIC, BIC HQIC, CAIC) as described in Section 2 have been determined. The values of the information criteria are closely related to the existence of the maximum likelihood function and the presence of the estimated parameters by means of the EM algorithm [4].

The findings of our analysis are, as follows: according to the representations in Fig. 1(a) a decrease in the parameter α determines an increase in the values of the information criteria AIC, BIC, HQIC, CAIC; this dependence decreases when the values of p increase. Also, it has been noted that compared to the threshold value $p = 0,5$ the values AIC, BIC HQIC, CAIC are equidistant, except when $\alpha = 1$. The following can be observed based on Fig. 1 (b): the higher the probability p, the smaller the values of AIC, BIC, HQIC, CAIC. Compared to the threshold value of the parameter $\alpha = 3$, the values of the information criteria are equidistant.

It can also be noted that the lowest values are characterized by the AIC information criterion in all the analyzed situations (Table 1, Table 2, Fig. 2). For example, based on Fig. 2(a), it can be concluded that the statistical model $MinParB(1,10,4,0.9)$ is selected as the best, providing us the best information, whereas Fig. 2(d) shows that the distribution $MinParB(1,0.5,4,0.9)$ is the model selected as being the best. From Fig. 2, for high probabilities (for example $p = 0,9$) we have low values for all the information criteria, and from Table 2, the values of the information criteria AIC, BIC HQIC, CAIC for the distribution $MinParB(1,10,4,p)$, $p \in \{0.2;0.5;0,9\}$ are very small, therefore a qualitative analysis of this distribution cannot be made in relation to the other distributions.

REFERENCES

[1] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19 (6), pp. 716–723, 1974.

[2] K. P. Burnham, D. R. Anderson, *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach* (2nd ed.), Springer-Verlag, 2002.

[3] A. Leahu, B. Gh. Munteanu, S. Cataranciuc, On the lifetime as the maximum or minimum of the sample with power series distributed size, *Romai J.*, vol. 9, no. 2, pp. 119-128, 2013;

[4] B. Gh. Munteanu, The Min-Pareto power series distributions of lifetime, *Appl. Math. Inf. Sci.*, vol. 10, no. 5, pp. 1673-1679, 2016;

[5] N. Enache-David, *On an application on entropy*, Bulletin of the "Transilvania" University of Brasov, Vol 6(55), No. 2 – 2013, Series III: Mathematics, Informatics, Physics, ISSN 2065-2151, pp. 87-94, 2013.