Results in Nonlinear Analysis 6 (2023) No. 4, 40–46
https://doi.org/10.31838/rna/2023.06.04.005
Available online at www.nonlinear-analysis.com

Results in Nonlinear Analysis
ISSN 2636-7556

Peer Reviewed Scientific Journal

Three vital steps to validating a probability distribution model

Showkat Ahmad Lone[a], Gilbert Chambashi[b]

[a]Department of Basic Sciences, College of Science and Theoretical Studies, Saudi Electronic University, Riyadh, 11673-KSA; [b]School of Business Studies, Unicaf University, Longacres, Lusaka, Zambia.

Abstract

This study presents a process of three important steps leading to the validation of a probability distribution model. It discusses how a family of probability distribution models can be selected to stand as candidate models to fit a given set of quantitative data, and then it discusses the methods of classifying these candidate models. The last part presents hypothesis testing as the final step in the process of probability model validation.

Key words: Candidate Model Selection, Model Fitting, Model Classification, Hypothesis Testing, Model Validation

Mathematics Subject Classification (2020): 62D05

1. Introduction

While mathematics provides a variety of mathematical methods for modelling quantitative data, one aspect that can never be avoided is the process of validating the best fitting model from among the candidate models [1]. In essence, the process of validating is simply a process of finding out if the best-fitting model is valid for use. Xiaomo and Sankaran [2] defined model validation as a decision-making process that involves deciding which model is valid to use by comparing the original data set (observed data) and the data set that gets produced by a fitted model (expected data) while taking into account the uncertainty that exists during the whole process.

Email addresses: s.lone@seu.edu.sa (Showkat Ahmad Lone)*

Received August 25, 2023; Accepted September 11, 2023; Online October 2, 2023

Several mathematical models exist. However, they can all be classified into two groups: deterministic models and stochastic (probabilistic) models. Stochastic modelling takes into consideration the random nature of data being modelled and seeks to measure this randomness [3]. The measure or quantification of the randomness (uncertainty) of data during modelling is also called probability [4, 5]. So, we can say that probability distribution models are part of the class of stochastic models.

One of the most interesting models is a group of probability distribution models, as well as the way to fit them to data [6], how to classify them [7], and the way to test which model fits the data the best [1, 8].

This paper brings, into the field of modeling, an organized and documented step by step procedure of probabilistic modeling from the first stage where we ask "which probability fits which data?" to the last stage of proving the validity of best selected and fitted probability distribution models to data.

Without leaving out the taking note of the three steps as being the candidate models' selection, the candidate models' classification, and then the validating hypothesis testing, this paper is structured as follows: First, in Section 2, it presents the discussion of how to select candidate probability models for the given data. Secondly, in section 3, the paper discusses the methods of how to classify these candidate probability models, and finally, in section 4, the paper discusses the process of proving the validity of the best fitting model(s) through hypothesis testing.

2. Candidate Model Selection

2.1. Selecting Probability Distributions' Families Using Histograms

The process of model validation is just an aftermath of the fitting of the available data to the probability distributions, which is a very important process of modelling. Fitting the data to the probability distributions would go well by first identifying the probability or the probability family to which the data's real distribution is very closely related [1, 8, 9, 19]. This is so because the data's real distribution is, in most cases, very difficult, if not impossible, to find, and so we attempt to associate the real distribution with its closely related distribution among those available. This is the essence of modelling. Research on discovering and constructing probability distributions that may be very closely related to most available data is in development. For example, insurance data's histograms show that it is always right-skewed and thick-tailed [1], hence the research that has developed closely related diverse corresponding distributions. Identifying a probability distribution family to which the real distribution of the data belongs can be done by computationally creating the data's histogram and observing to see which of the available distributions' density graphs the data's histogram resembles. This requires knowing various varieties of distributions and their density graphs. Figures 2.1.1–2.1.4 are examples of histograms [10, 12].

The first histogram is right-skewed (positively skewed); this takes the shape of density functions of probability distributions such as Pareto, Gamma, Lognormal, Inverse Gaussian, and Weibull [1, 19]. Among others, insurance claims data histograms take the shape of the first histogram. The second histogram is left skewed, the shape of which takes after the density functions of the distributions of returns on investments, daily stock market returns, age of deaths, etc. The second histogram is somehow symmetric and data which produces this shape is usually modelled by a distribution such as a

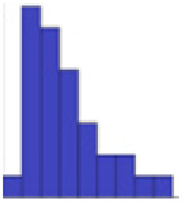

Figure 2.1.1

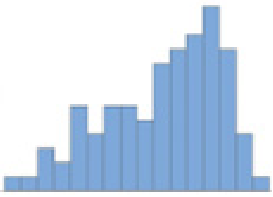

Figure 2.1.2

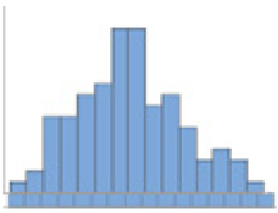

Figure 2.1.3

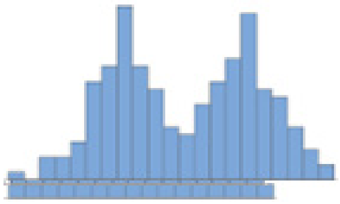

Figure 2.1.4

normal distribution, although depending on the shape of being symmetric, distributions such as uniform, Cauchy, logistic, etc. can also be considered to model data with symmetric histograms [13]. The fourth figure indicates a mixture of distributions that are somehow symmetric, although histograms of mixed distributions are not always of symmetrically shaped distributions. A mixed distribution can be right-skewed, symmetric, or any other possible mixture [9].

As is the case in modelling, no probability model is “a best fit it all” model, hence the reason for modelling. But the histograms of insurance data always show that it fits best with right-skewed distributions.

Based on what’s been said, it seems like it would be best to make a histogram of the data first so that it can be linked to a closely related probability distribution model in terms of histograms before fitting the model to the given data.

2.2. Selection by the Q-Q plots/P-P plots

The Q-Q plots and P-P plots simply connote Quantile-Quantile plots and Probability-Probability plots, respectively [9]. The term “Q-Q plot” simply refers to plotting the quantiles of a given set of data (observed data) and data estimated from the fitted probability distribution model (expected data). In the same vein, the P-P plot denotes the comparison of the observed and expected data’s probabilities on a plot by plotting the probabilities of the respective sets of data [9]. Check out the figure below, which shows the three P-P plots of Algeria’s GAM General insurance company’s 2015 (plots 1 and 2) and 2016 (plot 3) claims data fitted to Lognormal-Pareto, Lognormal-Logistic, and Lognormal-Burr composite distribution models [8].

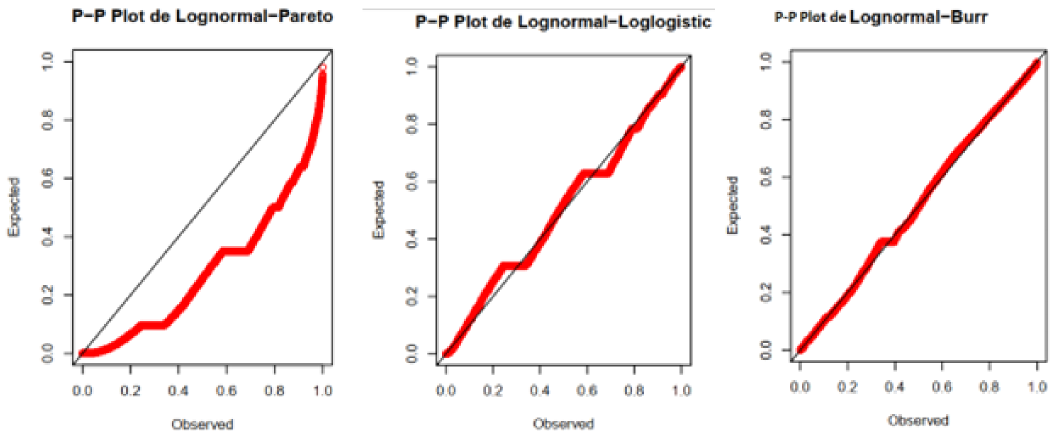

Figure 2.2.1: P-P Plots of Lognormal-Pareto, Lognormal-Loglogistic and Lognormal-Burr.

It should be noted that with P-P plots, the more the thick red line aligns with the thin dark line, the better the data set fits the probability distribution. Where the thick red line perfectly fits the thin dark line, it implies the data set (observed data) perfectly fits the probability distribution or it is simply of the same probability distribution. The first P-P plot indicates the data set that clearly doesn’t come from the Lognormal-Pareto probability distribution model, and the third P-P plot indicates that the Lognormal-Burr probability distribution can be considered as one of the candidate models for fitting the data set as the red line almost perfectly fits the dark, thin, straight line.

The Q-Q Plots are applied the same way the P-P Plots have been applied above.

3. Classification of Candidate Probability Models

This is the immediate process that comes after fitting the data to the candidate models. Fitting the data to the candidate models involves estimating the models’ parameters by the maximum likelihood estimation method [2, 6].

3.1. Classification by NLL, AIC and SBC

The Negative Log-likelihood (NLL): The NLL is simply the negative of the maximized log-likelihood. The estimation of parameters by maximum likelihood estimation methods produces a maximized log-likelihood value along with estimated parameters [6]. The maximized log-likelihood value is by itself meaningless, but it is derived (determined or calculated) from the maximum likelihood, which is simply a measure that is given for estimated parameters, at their maximum probability, of having to have their model produce the observed data [7]. Where two or more models are being fitted to data, the negative log-likelihood can be a very helpful value to select a best-fitting model, except that the models have to be of the same number of parameters [7]. The model with the biggest negative log-likelihood value can be picked as the best fitting model.

The Akaike Information Criterion (AIC): The AIC comes in to cover the limitation of the NLL where it only works for models with the same number of parameters. The AIC has the benefit of being able to compare fitted models with different numbers of parameters [7]. By having $NLL = -L(\theta)$, where θ is a set of a model's parameters, the AIC is given by

$$AIC = 2k - 2L(\theta)$$

Where k represents the number of parameters in the set θ. This simply means that, by subtracting 2k from $2\,L(\theta)$, AIC penalises a model for having more parameters [7].

The Schwarz's Bayesian Criterion (SBC): Also known as the Bayesian Information Criterion (BIC), it is also a remedy for NLL's incapability of comparing models with varying numbers of parameters [7]. It is given by

$$SBC = kln(n) - 2L(\theta)$$

Here, as in AIC, k is the number of the model's parameters and is the size of the data being modelled. From SBC, it can be stated that the bigger the data size and the greater the number of parameters being estimated, the more the model gets penalized.

As with NLL, models with the smallest values for AIC and SBC are the ones that fit the data best [1].

3.2. Classification by KS, AD and CvM Test Statistics

Unlike the NLL, AIC, and SBC, which look at the maximum likelihood of the fitted model to produce the observed data, the KS, AD, and CvM aim at giving the value of the 'distance' between the fitted model's cumulative distribution function (CDF) and the observed data's empirical distribution function (EDF) [1]. The best-fitting model is the one that has the smallest difference between its CDF and its EDF [10]. From Calderin-Ojeda and Kwok work [1], for $\hat{F}$ being the fitted model's CDF, $x_1, x_2, \ldots, x_N$ being the observed data and $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$ being the observed data in increasing order, we have

- Kolmogorov-Smirnov (KS) test statistics given by $D = \max(D^+, D^-)$, where $D^+ = \max_{1\le j\le N}\left\{\frac{j}{N} - \hat{F}(x_{(j)})\right\}$ and $D^- = \max_{1\le j\le N}\left\{\hat{F}(x_{(j)}) - \frac{j-1}{N}\right\}$
- Cramer-von Mises (CvM) test statistic given by $W^2 = \sum_{j=1}^{N}\left[\hat{F}(x_{(j)}) - \frac{2j-1}{2N}\right]^2 + \frac{1}{12N}$
- Anderson-Darling (AD) test statistic given by

$$A^2 = -N - \frac{1}{N}\sum_{j=1}^{N}[(2j-1)\log(\hat{F}(x_{(j)})) + (2n+1-2j)\log(1-\hat{F}(x_{(j)}))]$$

4. Proving the Validity of Best Fitting Model by Hypothesis Testing

A best-fitting probability model can be incorporated into diverse relevant formulae and be used to calculate various statistical measures such as a business's maximum expected loss, insurance premiums, reinsurance premiums, etc. However, before a model can be used, it has to be tested to see if it qualifies. A way to do this is by hypothesis testing [2].

Hypothesis testing involves statistically finding out whether we approve a defined null hypothesis (H_0) and, by default, disapprove the alternative hypothesis (H_a) or vice versa [15, 11]. This can be given by;

H_0: The best fitting model is valid to be used
vs
H_{a}: The best fitting model is not valid to be used

4.1. Approving/Disapproving H_0 by Comparing a Test Statistic and a Critical Value

For a given level of significance (α), we can determine its corresponding critical value by some methods, such as the most basic one, which is the use of the probability distribution model's table of values [15].

Statistical Hypothesis: For a better understanding, we give an example of a hypothesis test dependent on a parameter γ_0. In this case, the H_0 and H_a for each test can, as a couple, be defined in one of the following three ways [9], depending on the suitable prevailing situation:

$$\begin{array}{lll} H_0 : \gamma \geq \gamma_0 & \qquad H_0 : \gamma = \gamma_0 & \qquad H_0 : \gamma \leq \gamma_0 \\ H_a : \gamma < \gamma_0 & \qquad H_a : \gamma \neq \gamma_0 & \qquad H_a : \gamma > \gamma_0 \end{array}$$

In the first case, the test is called a left-tailed test; in the second case, the test is called a two-tailed test; and in the third case, the test is called a right-tailed test. The first and the third tests can be classified under one name: one-tailed tests [15, 17].

Graphically, using the probability model's density, a critical value is a value dividing the density figure into a rejection and a non-rejection region. And therefore, in the left-tailed test's null hypothesis, $\gamma \geq \gamma_0$ means that we do not reject (we approve) the null hypothesis if the test statistic is greater than or equal to the critical value, otherwise we reject (we disapprove). In the right-tailed test, $\gamma \leq \gamma_0$ means that we approve the null hypothesis if the test statistic is less than or equal to the critical value. And in the two-tailed test, we accept the null hypothesis if the test statistic is between the negative and positive values of the critical value found at half the level of significance (α). Otherwise, we reject the null hypothesis and accept the alternative hypothesis [17].

4.2. Approving/Disapproving H_0 by Comparing a Level of Significance and a p-value

Here, the approving and disapproving are facilitated by two highly interrelated statistical measures, namely the level of significance, also known as alpha (α), and the p-value. The level of significance (α) is simply the probability of disapproving the true null hypothesis, and the p-value is the smallest value of the level of significance at which we disapprove the null hypothesis [17]. This can be summarized by the following mathematical statement.

If p-value ≥ alpha (α) then approve the H_0 otherwise disapprove H_0

The p-value can also be defined as a direct measure of the likelihood of the given set of data being in the candidate probability model [17]. This being the case, sometimes it is not necessary to determine the level of significance and compare it with the p-value. Instead, we take the p-value as the probability of having our given set of data come from the candidate probability model. In this way, we take it that the closer the probability is to one (1), the more likely it is that the data is of the candidate probability model and vice-versa. In most strict cases, how close to one (1) a probability is can be judged by its being able to be rounded off to one (1).

4.3. Determining the Level of Significance and the p-value

The level of significance can be determined by choice, depending on which would best fit the kind of test being carried out; it can be 0.01, 0.1, or the standard one, 0.05 [18]. However, the p-value can, among other methods, be determined by the bootstrap method using the following steps [8, 1]

- For the best fitting model, calculate the test statistics such as t_{KS}, t_{CvM} and t_{AD} as given in section 3.2.Using the model providing the best fit to $x_1, x_2, \ldots, x_N$ as data,
  - generate M sets of resampled data and denote them as $\hat{x}_1^{(i)}, \hat{x}_2^{(i)}, \ldots, \hat{x}_N^{(i)}$ for $i = 1, \ldots, M$.
  - refit it to each set of the resampled data and then compute the test statistics $t_{KS}^{(i)}, t_{CvM}^{(i)}$ and $t_{AD}^{(i)}$ for $i = 1, \ldots, M$.
- Then, finally, determine the p-values by $\frac{\#\left\{i : t_{KS}^{(i)} \geq t_{KS}\right\}}{M}$, $\frac{\#\left\{i : t_{AD}^{(i)} \geq t_{AD}\right\}}{M}$ and $\frac{\#\left\{i : t_{CvM}^{(i)} \geq t_{CvM}\right\}}{M}$

5. Conclusion

In the event that no probability distribution model fits the data, probability as a field provides the possibility that one can be created to fit the data, although this would take more work than just selecting it from the available ones [14, 16]. In section 2, although histograms and Q-Q/P-P plots can be used separately to select candidate probability distributions, Q-Q/P-P plots can also be used to confirm candidate models selected by histograms. To put it more sequentially, candidate models may be selected using histograms, histograms may be confirmed using Q-Q/P-P Plots, and thereafter, fitted models can be confirmed by Q-Q/P-P Plots and can be ordered/classified from the best-fitting to the worst-fitting ones by using NLL, AIC, SBC, KS, AD, and CvM [8]. The hypothesis testing method, which is proposed in this paper as the final stage in the probability model validation process, can then be used to test the best-fitting model(s) to see if they are valid.

References

[1] Calderín-Ojeda, E., Kwok, C.F., *Modeling claims data with composite Stoppa models*, Scandinavian Actuarial Journal, (2015). doi:10.1080/03461238.2015.1034763

[2] Xiaomo, J., Sankaran, M., '*Bayesian inference method for model validation and confidence extrapolation*', Journal of Applied Statistics, 36 (6), (2009), 659–677. doi:10.1080/02664760802499295

[3] Renard, P., Alcolea, A., Ginsbourger, D.D., *Stochastic versus Deterministic Approaches*. In: Environmental Modelling: Finding Simplicity in Complexity. Willey-Blackwell Publishers. (2009). doi:10.1002/9781118351475.ch8.

[4] Loucks, D.P., Beek, E.V., *An Introduction to Probability, Statistics, and Uncertainty*. In: Water Resource Systems Planning and Management. Springer, Cham. (2017). doi:10.1007/978-3-319-44234-1_6

[5] Brownlee, J., *A Gentle Introduction to Uncertainty in Machin Learning*, (2019), accessed 22 September 2022. https://machinelearningmastery.com/uncertainty-in-machine-learning.

[6] Nadarajah, S., Bakar, S., *CompLognormal: An R Package for Composite Lognormal Distributions*. R Journal. 5 (2), (2013) 98–104.

[7] Bakar, S.A.A., Hamzah, N., Maghsoudi, M., Nadarajah, S., *Modeling loss data using composite models*. Insurance: Mathematics and Economics. 61, 146–154. (2015). doi:10.1016/j.insmatheco.2014.08.008

[8] Chambashi, G., Mushala, W., Mwaanga, C., Mayondi, C., Kolosa, B., Matindih, L.K., Moyo, E., *Computation of Reinsurance Premiums by Incorporating a Composite Lognormal Model in a Risk-Adjusted Premium Principle*. Journal of Mathematical Finance, (2023), 13, 1–16. doi:10.4236/jmf.2023.131001

[9] Bakar, S.A.A, Nadarajah, S., Adzhar, Z.A.K.A, Mohamed, I., *Gendist: An R Package for Generated Probability Distribution Models*. PLoS ONE 11 (6), (2016), e0156537. doi:10.1371/journal.pone.0156537

[10] Boels, L., Bakker, A., Dooren, W.V., Drijvers, P. *Conceptual difficulties when interpreting histograms: A review*, Educational Research Review, 28, (2019), 100291, ISSN 1747-938X. doi:10.1016/j.edurev.2019.100291

[11] Frost, J., *Hypothesis Testing: An Intuitive Guide for Making Data Drive Decisions*. (2020) ISBN: 9781735431161. (1st ed.). Jim Frost.

[12] Frost, J., *Using Histograms to Understand your Data*, (n.d.), accessed 15 June 2022. https://statisticsbyjim.com/basics/histograms.

[13] Fang, K., Symmetric Multivariate and Related Distributions. Chapman & Hall/CRC. (2018).

[14] Klugman, S.A., Panjer H.H., Willmot, G.E., *Loss Models, From Data to Decisions* 2nd edition. A John Wiley & Sons Inc Publication, New Jersey, USA. (2004).

[15] Black, K., *Business Statistics: For Contemporary Decision Making*, Edition: 8th Wiley. (2014).

[16] Nadarajah, S., Bakar, S., *New composite models for the Danish fire insurance data*. Scandinavian Actuarial Journal, (2012), 97–101. doi:10.1080/03461238.2012.695748

[17] Doane, D.P., Seward, L.E. *Applied Statistics in Business and Economics*. (5th ed.). McGraw-Hill Higher Education-USA. (2016).

[18] François, D., *Les probabilités et la statistique de A a Z. Dunod*. (2007), ISBN 978-2-10-051403-8. www.dunod.com, www.biblio-scientifique.net

[19] Dickson, D.C.M., *Insurance Risk and Ruin*. Cambridge University Press, (2005).