

**ESSnet Big Data II**

**Grant Agreement Number:**

<https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata>

[<https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en>](https://ec.europa.eu/eurostat/cros/content/essnetbigdata_en)

**Workpackage WPB**

**Implementation – Online Job Vacancies**

**Methodological description on fitting discrete probability distributions to a national dataset combining jobs survey data with online job advertisements data**

**Prepared by:**

**Marian NECULA**

**(National Institute of Statistics, Romania)**

**Date of present version: August 20 2020**

Workpackage Leader:

Tomaž Špeh (SURS, SI)

e-mail address: tomaz.speh@gov.si

mobile phone: +38651672116

**Contents**

[Introduction 3](#_Toc48740583)

[Theoretical background for the estimation procedure. 4](#_Toc48740584)

[Dataset description 5](#_Toc48740585)

[Estimation and software tools 6](#_Toc48740586)

[Conclusions 9](#_Toc48740587)

[References 9](#_Toc48740588)

# **List of figures**

[Figure 1 Simplified overview on how to fit a distribution to a dataset. 4](#_Toc48812173)

[Figure 2 Cullen- Frey graph for job vacancies. Dataset: Romanian NIS Job Vacancies Survey. 7](#_Toc48812174)

[Figure 3 Cullen-Frey graph for job advertisements. Dataset: Romanian National Agency for Job Vacancies. 7](#_Toc48812175)

**Introduction**

In contemporary society, based on observations made so far, we can make the following axiomatic assumption: data comes in different ”shapes and sizes”. Here by data we denote one or more random variables. If this is the case, a fundamental problem in statistics is derived from the natural question: “What data generation process, i.e. mathematical function, optimally approximates our data?”. This is a very hard question, either if we look at the type of the problem from the perspective of a search problem on an infinite space of solutions, more so if we look into at the mathematical theory which makes this type of problem tractable. In general, there is no silver bullet to this type of problems, and a variety of heterogenous strategies are used from many fields of mathematics, e.g. ranging from computing some measure on the sample space to solving some partial differential equation. A secondary problem which can be derived from finding the functions, is given by the question: “if two random variables are approximated by distributions within the same family, can we approximate the random variable for which we don’t have observations at a certain time instant, with another variable for which we have observations?”. This last question is of interest in the case of estimating the unobserved number of job vacancies in terms of the observed number of online jobs advertisements.

Although it is not necessary to grasp the mathematical reasoning behind the notions in order to apply the algorithms to generate feasible solutions, it is generally recommended to have at least some basic notions of probabilities and mathematical distributions. An excellent primer in theoretical probabilities and statistics is the Cassela & Berger textbook [1] in which most of the mathematical statistical concepts employed in this work are thoroughly explained, with rich examples.

One important tool in the belt of a practitioner statistician is the procedure of fitting a known family of theoretical probability distribution to a dataset, i.e. one or more random variables, generated by a process which entails, or not, independence between observations and/or variables. This is mainly a three-step iterative algorithm and consists in making an informed/educated guess about the theoretical family which best fits the data, estimating the parameters for the theoretical distribution and testing the null hypothesis which states that the parameters of the theoretical distribution optimally approximate the parameters of the empirical distribution, i.e. the differences between the two types of parameters asymptotically converge to 0. If the theoretical probability distribution optimally fits the empirical one, then different types of statistics about the dataset can be estimated and/or used in mathematical models which assume a certain type of probability distribution. This whole procedure is, more or less, abstracted/simplified in figure 1.

The first step in applying the distribution fitting procedures to a dataset is to make some sort of educated guess regarding what family of distribution, i.e. analytical form, best describes the generation process of the dataset. Here, two observations are in order to understand the difficulty of this process. Firstly, finding a good approximation for the empirical distribution is very hard, in large part due to the intractable number of probability distributions. Secondly, the problem is domain specific, requiring domain knowledge input. The second step involves estimating the parameters of the theoretical distribution based on the observations. At this step, a procedure called maximum likelihood estimation is used. For the third step some sort of measure for the goodness-of-fit is applied to test if the differences between the parameters for the empirical probability distribution function and the estimated parameters of the theoretical probability distribution function are statistically different from 0. If this is the case, then population parameters can be estimated, otherwise the procedure jumps to step 1, in which a different selection is made.

In the following sections, each of the steps will be presented along with examples using datasets collected as result of the project (national collected datasets) and already published national statistics on job vacancies. Also, we will provide a custom R function which will try to fit some theoretical distributions to a dataset comprising two categorical variables (count data), the official statistics regarding the number of job vacancies broken down by some type of classification rule, and an accurately classified according to the same rule. As an output this function will generate a pdf (portable document format) report.

Figure 1 Simplified overview on how to fit a distribution to a dataset.

# **Theoretical background for the estimation procedure**.

The procedure used to fit a dataset to a probability distribution function, through estimation of the parameters is known as maximum likelihood estimation.

Maximum likelihood estimation (MLE) [1] is a statistical technique for estimating the parameters of probability density distribution based on the likelihood function, which in a frequentist interpretation postulates that all the relevant information in a sample is contained in the likelihood function, i.e. “data are interpreted as evidence, and the strength of the evidence is measured by the likelihood function”.

Main steps in MLE:

1. Make an educated guess about the probability distribution function (p.d.f.) which optimally fits the data.
2. Compute the likelihood function given by the formula:

**,**

Where

- pdf

**x** – random variable or sample data,

Given the nature of this function in practice is simpler to compute the log-likelihood function, therefore the former can be expressed

**,**

Where

1. Estimate the parameter(s) θ which maximize , formally

This is usually achieved by solving the Langrage multipliers system of equations.

# **Dataset description**

The input dataset consists of three variables, namely Romanian NACE Section codes (main types of industries) [2] consisting in uppercase letters, denoted by NACE, the number of job vacancies from national survey by NACE Section code [3], denoted by JVS, and the number of jobs advertisements collected online from the Romanian National Agency for Job Vacancies [4], denoted by JVA. The JVS consists of quarterly aggregated data by NACE codes (sections, divisions, etc.), for the first quarter of 2020, which was aggregated in a set of 19 observations by NACE Section codes. The JVA consist in 17 observations (given the period of the year, there is missing – NA - data on some NACE Section codes, e.g. agriculture) from data collected on 28 March 2020, also aggregated on NACE Section codes. The JVA data is a tidy dataset by design, so this estimation exercise assumes that the dataset is optimally classified, already classified by NACE codes. The two datasets are joined in a common structure by NACE Section codes, resulting in 19 observations. Some descriptive statistics for our dataset are presented in table 1. Descriptive statistics for input dataset.

Table 1 Descriptive statistics for input dataset

|  |  |  |
| --- | --- | --- |
| **Descriptive statistics** | **JVS** | **JVA** |
| Minimum | 56 | 66 |
| 1st quartile | 397 | 160 |
| Median | 1008 | 446 |
| Mean | 2198 | 1126 |
| 3rd quartile | 2910 | 1355 |
| Maximum | 9230 | 6267 |
| Missing | - | 2 |
| Standard deviation | 2498.039 | 1677.894 |
| Inter-quartile range | 2513 | 1195 |
| Coefficient of variation | 1.1365 | 1.4895 |

# **Estimation and software tools**

For testing the optimal fit of a statistical distribution, we used fitdistrplus R package [5], which provides an easy to use interface for testing the goodness-of-fit for several popular probability distributions families. The easiest way of selecting a probability distribution is through *descdist* function, which compares the squared skewness, the third statistical moment, and kurtosis, the fourth statistical moment, of the empirical with those of theoretical distributions, for continuous or discrete probability distributions, and computes parameters of the empirical distribution, minimum, maximum, median, sample mean and standard deviation. From the output of this function a Cullen-Frey [6] graph is produced in which is easily identifiable membership of the empirical probability distribution to one of the classical families. In figures 2 and 3 Cullen-Frey plots are presented for job vacancies from the national survey and for job vacancies advertisements collected from the Romanian National Agency for Job Vacancies.

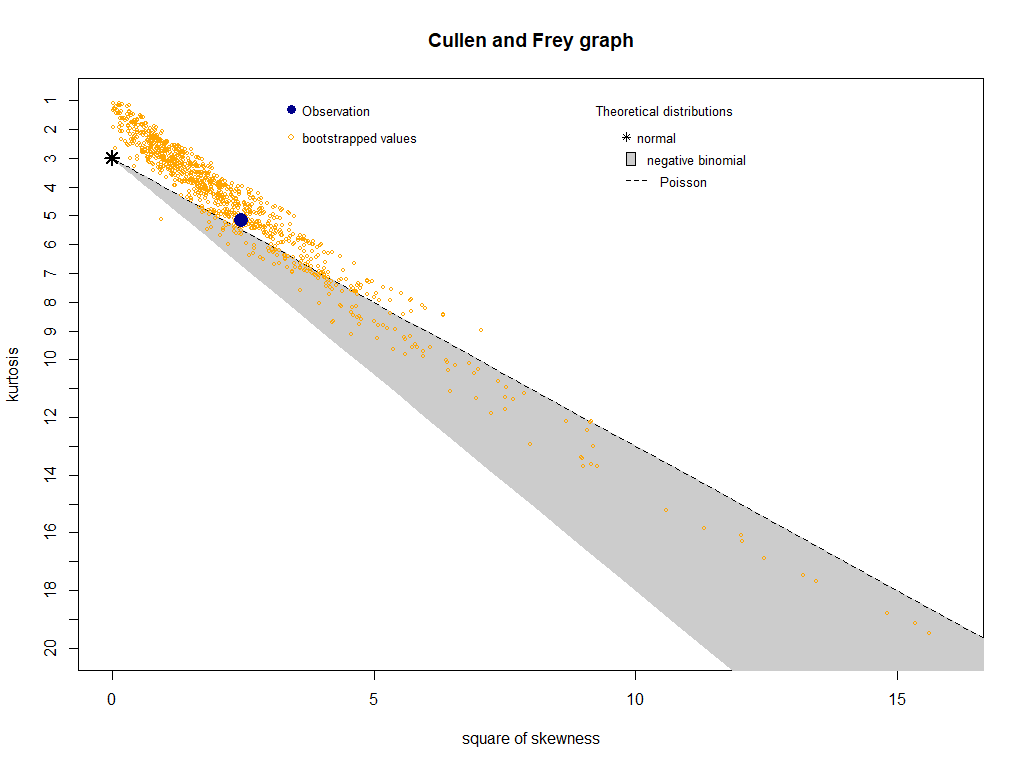


Figure 2 Cullen- Frey graph for job vacancies. Dataset: Romanian NIS Job Vacancies Survey.

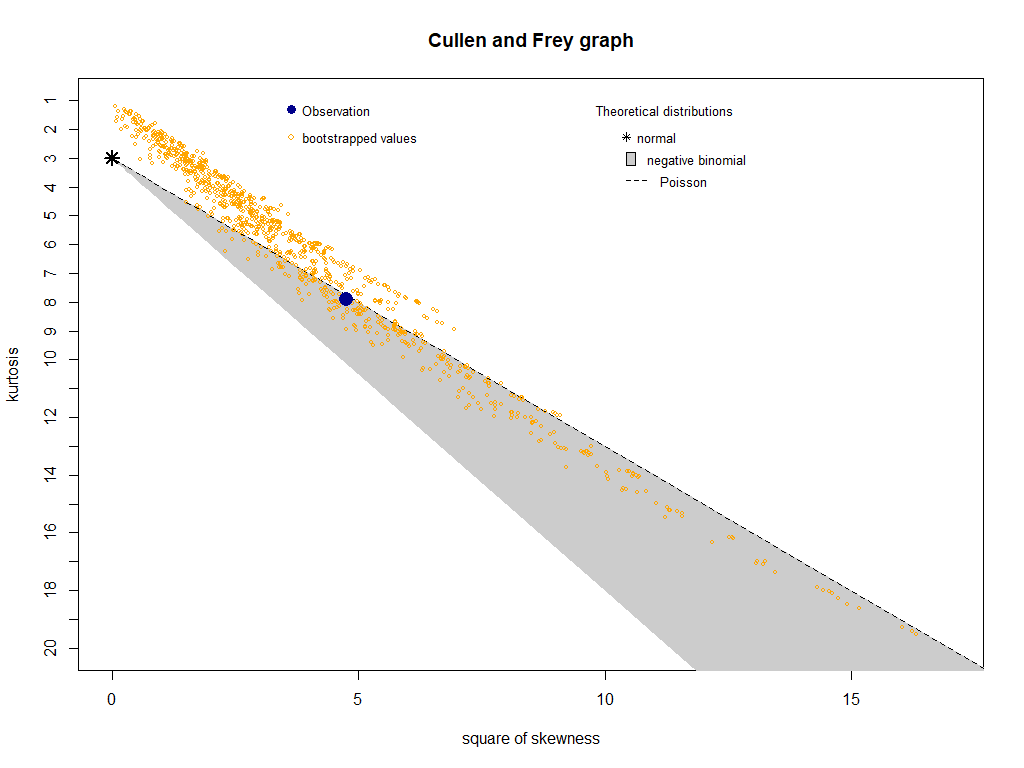


Figure 3 Cullen-Frey graph for job advertisements. Dataset: Romanian National Agency for Job Vacancies.

The two graphs describe how “close” the empirical probability distributions are to a few discrete probability distributions. The blue point represents the coordinates of the squared skewness and kurtosis computed by 1000 bootstrapped samples.

Based on the two graphs, we can try to fit a negative binomial distribution to our dataset. The negative binomial distribution has the following probability mass function:

where,

for which we can compute skewness (third-order moment):

where,

and the kurtosis (fourth-order moment)

upon which the probability distribution family was selected. The negative binomial distribution can be regarded a Poisson mixture with the rate parameter, λ, generated by a gamma probability distribution. In this case, the parameter of interest is r, number of successes or dispersion parameter. This solves the problem of constant rate λ and under/over dispersion can be modelled with ease by a negative binomial probability distribution.

The next step involves estimating the parameters through MLE. The parameters for the negative binomial distribution are the number of trials until the *r* success and the probability *p* for success.

In this case, the estimated parameters are presented in table 2. Estimated parameters for theoretical probability distribution fitting, where we can observe a greater variability of the JVS compared to JVA, which can entail an under coverage present in the data coming from the Romanian National Agency for Job Vacancies

Table 2 Estimated parameters for theoretical probability distribution fitting.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Parameter** | **Estimate** | **Standard Error** |
| JVS | Size = r | 0.8015 | 0.2246 |
| Mean | 2198 | 563 |
| JVA | Size = r | 0.6415 | 0.1862 |
| Mean | 1127 | 341 |

Using the estimates output we are now able to generate probabilities for each observation in our two variables, JVS and JVA.

# **Conclusions**

Estimating the probability density functions, under the simplest assumptions, i.e. data is generated by a univariate process, is a rather difficult problem. There are a lot of probability distribution functions, with different types of parametrisation and/or interpretations, given a problem from a certain domain. The standard procedures for fitting a probability distribution function to a dataset involves estimating the parameters of the theoretical probability distribution which optimally approximates our dataset and compare them to the parameters of the empirical probability distribution. If the differences between the two converge to 0, we can approximate our dataset with a known probability distribution. The main mathematical procedure used to compute the parameters is by maximum likelihood estimation.

The two data sources deal with count data. For count data we have a lot of probability distribution from which we can select an optimal one. Thankfully this process is quite easy using the R package fitdistrplus. By comparing the squared skewness and kurtosis of our empirical probability distribution with the same moments computed for a few classical probability distributions, an educated guess can be made.

Acknowledging that we barely scratched the surface on this type of inference problems, we are optimistic that new statistics can be developed by incorporating new data sources through statistical inference procedures. If this will be the case, then it can be quite easily to generate multiple types of statistical measures of uncertainty in the estimation procedures. One possible future research problem would be to estimate the joint probability distribution of the two variables and compute the conditional expectation: given a number of observations from our JVA for a certain NACE section, what is the expected value of JVS for that NACE section, formally stated as:

where,

To compute the conditional probability, we simply use the conditional probability formula:

# **References**

|  |  |
| --- | --- |
| [1] | G. Cassela en R. L. Berger, Statistical Inference, Pacific Groove, California, USA: Wadsworth Group, 2002. |
| [2] | Romanian National Institute for Statistics, „SENIN server,” [Online]. Available: http://colectaredate.insse.ro/senin/classifications.htm?selectedClassification=CAEN\_Rev.2&action=download. [Geopend 30 July 2020]. |
| [3] | Romanian National Institute of Statistics, „LMV102B - Vacancies by macroregions, development regions, activity of national economy at level of CANE Rev.2 section,” 2020. [Online]. Available: http://statistici.insse.ro:8077/tempo-online/. [Geopend 30 July 2020]. |
| [4] | Romanian National Agency for Job Vacancies, „Job Vacancies Ads,” 2020. [Online]. Available: https://www.anofm.ro/lmvw.html?agentie=&categ=3&subcateg=1. [Geopend 30 July 2020]. |
| [5] | M. L. a. D. C. Delignette-Muller, „fitdistrplus: An R Package for Fitting Distributions,” *Journal of Statistical Software,* vol. 64, nr. 4, pp. 1-34, 2015. |
| [6] | A. Cullen en H. Frey, Probabilistic Techniques in Exposure Assessment, New York: Plenum Publishing Co, 1999. |