University of Coimbra

Doctoral Program in Information Science and Technology

Statistics

Assignment #2 - Distribution Fitting

Joaquim Pedro Bento Gonçalves Pratas Leitão - 2011150072

January 3, 2017

1 Introduction

The current document is framed in the scope of the second assignment of the Statistics course, taught for the Doctoral Program in Information Science and Technology at the University of Coimbra, during the academic year of 2016/2017.

In the current assignment a distribution fitting task was proposed. In this task, a probability distribution was intended to be fitted to a collected, and previously provided, dataset.

The dataset in question corresponds to concentrations of Strontium-90 (in micromicrocouries) recorded in 50 samples of milk prepared for consumption by a given company. Taking into account that no prior knowledge regarding the nature or the distribution of the collected data, the approach followed in the current work can be divided in the following steps:

1. Identification of potential probability distributions that justify the collected data. At this point, this was accomplished by means of visual inspection, comparing of the histogram of the data with the probability density plots of candidate distributions.

2. Estimation of the parameters of the probability distributions identified in the previous point, based on information extracted from the collected data: At this point, *Maximum Likelihood Estimations* of the parameters will be performed.

3. For each of the probability functions identified in the first point perform the *Kolmogorov-Smirnov* Test, with the parameters estimations obtained in the previous point. The probability distribution chosen to justify the collected data will be the one with the higher *p-value* in the *Kolmogorov-Smirnov* Test.

In the remainder of this document each of these steps will be addressed in an individual section (sections 2 to 4). In addition, the document will conclude with a brief analysis of the results achieved in this work (section 5).

2 Candidate Probability Distributions

The current section covers the first step of the adopted approach, where potential probability distribution functions that justify the observed data were identified and selected.

Such functions were selected by comparing the histogram of the observed data with the corresponding probability density functions. Therefore, this section will consist of a very brief presentation of the observed dataset and its corresponding histogram, followed by its analysis and comparison with candidate probability distribution functions.

2.1 The Dataset

As already stated in section 1, the dataset provided for this work consisted of concentrations of Strontium-90 (in micromicrocuries) recorded in 50 samples of milk prepared for consumption by a given company. The dataset is presented in table 1:

Concentration
7,974
9,592
7,792
8,366
8,816
7,015
8,605
7,681
9,042
6,368
6,477
8,637
8,617
6,796
7,73
7,412
8,128
11,055
4,881
7,298
11,314
10,861
8,342
6,939
5,842
10,188
6,698
6,183
8,623
10,479
7,425
9,031
4,039
9,207
7,143
8,559
5,501
11,725

6,191
5,904
5,931
7,151
9,638
8,832
7,76
8,011
7,637
7,488
8,514
9,309

Table 1: Measured Strontium-90 concentrations.

Figure 1 presents the histogram of the observed data:

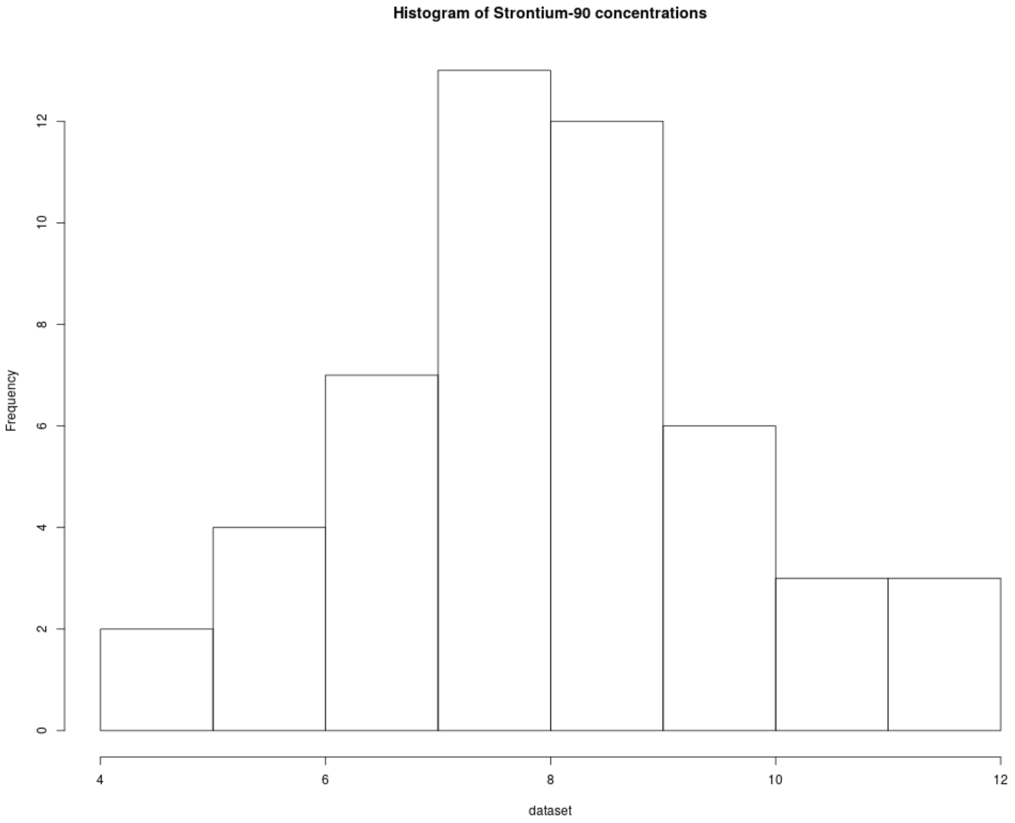


Figure 1: Histogram of the observed data.

2.2 Histogram Analysis

Analysing the histogram presented in figure 1, a symmetry of the data relative to a value close to 8 is suggested.

Based on this property, symmetric probability distributions were further studied and compared with the obtained histogram. As a result, density functions of several distributions were plotted and compared with this histogram, of which the following distributions were highlighted: *Cauchy*, *Gamma*, *Logistic* and *Normal*.

Figures 2 to 5 presents the histogram of the data overlapped with the density function for each of the mentioned distributions. In order to obtain estimations for the parameters of these distributions the *method of moments* was used.

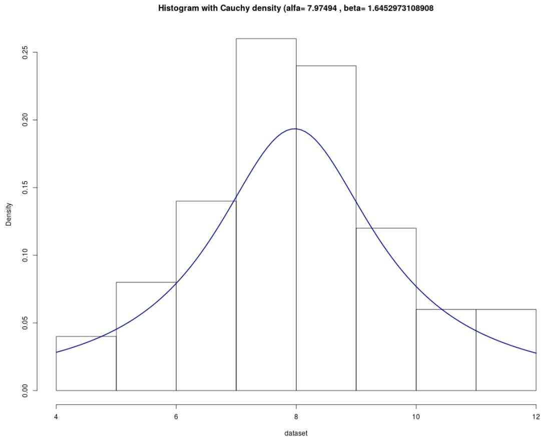


Figure 2: Cauchy Distribution.

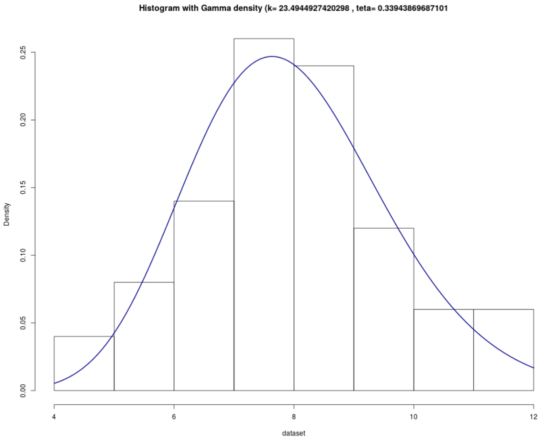


Figure 3: Gamma Distribution.

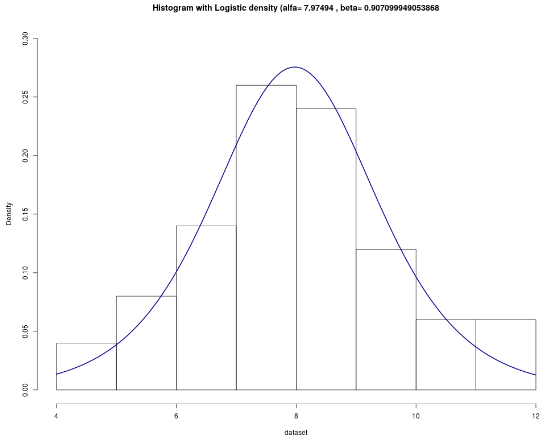


Figure 4: Logistic Distribution.

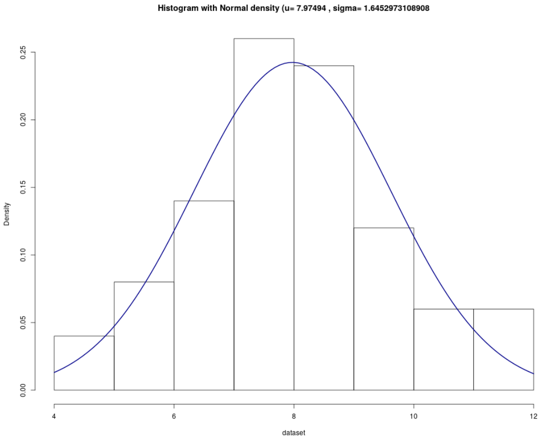


Figure 5: Normal Distribution.

From the presented figures it becomes clear that the Cauchy distribution cannot accurately justify the observed data as well as the remaining three distributions.

Subsequently, only the Gamma, Logistic and Normal distributions will be considered in the remaining steps of this work.

3 Parameter Estimation

Once candidate distributions that seem to justify the observed data were identified, the next logical step to perform is to compute more robust estimations for their parameters, to be used in appropriate statistical tests.

In this work such estimations were obtained by computing *Maximum Likelihood Parameter Estimations* for the three distributions identified in the previous section.

3.1 Maximum Likelihood Estimations

The *Maximum Likelihood Estimations* were computed using the *R*[1] software environment. More precisely, the *mle2* function available in the *bbmle* package was used.

The following estimations were obtained for the considered distributions:

- **Gamma Distribution:** A value of 23.49449 was estimated for the *shape* parameter k and a value of 0.3394412 was estimated for the *scale* parameter θ.
- **Logistic Distribution:** A value of 7.946271 was estimated for the *location* parameter α and a value of 0.9224431 was estimated for the *scale* parameter β.
- **Normal Distribution:** A value of 7.97494 was estimated for the *location* parameter μ and a value of 1.628762 was estimated for the *scale* parameter σ.

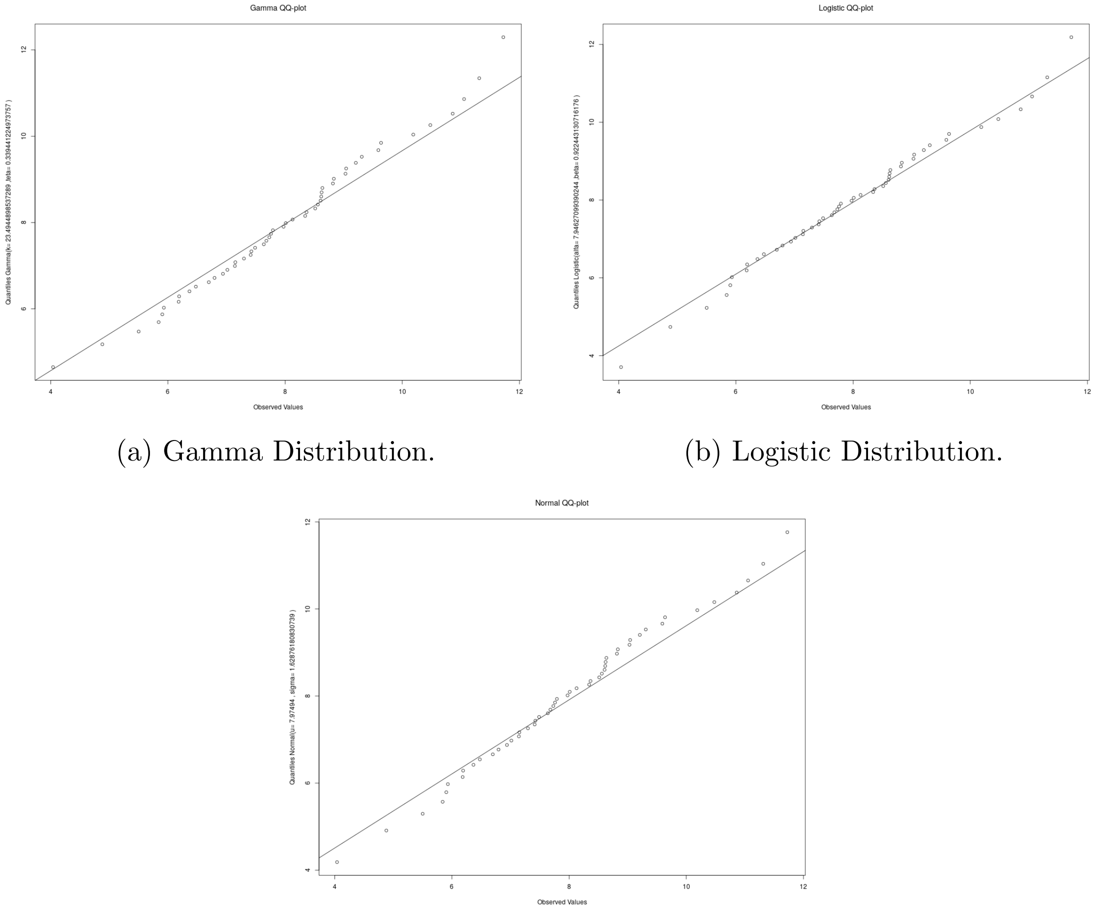
3.2 QQ-Plots

A *QQ-Plot* is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the QQ-Plot will approximately lie on the line $y = x$.

The inclusion of *QQ-Plots* at this stage of the analysis provides additional information regarding the similarity between the observed data and each of the considered distributions.

Figure 6 presents the obtained *QQ-Plots* for the three distributions, making use of the *maximum likelihood* parameter estimations.

[1] `https://www.r-project.org/`



(a) Gamma Distribution.

(b) Logistic Distribution.

(c) Normal Distribution.

Figure 6: QQ-Plots with Maximum Likelihood Parameter Estimations.

Observing the three *QQ-Plots* presented the *Logistic* distribution appears to be the one which better justifies the observed data, as the points in the plot appear to be the ones closer to the line $y = x$.

4 Kolmogorov-Smirnov Test

In the final stage of the work, the *Kolmogorov-Smirnov Test* (KS Test) was performed for the three distributions considered. The KS Test is a non-parametric test of equality of continuous, one-dimensional probability distributions, used to compare a sample with a reference probability distribution (one-sample KS Test) or to compare two samples (two-sample KS test).

Considering the data available in the current work, three one-sample KS tests were performed. In these tests the *maximum likelihood parameter estimations* computed in section 3.1 were used. The following *p-value* results were registered:

- **Gamma Distribution:** $p - value = 0.9974$

- **Logistic Distribution:** $p - value = 0.9993$
- **Normal Distribution:** $p - value = 0.9838$

Indeed, in the KS Test, a high *p-value* signals a higher change of the null hypothesis being true, meaning that the higher the *p-value* the more likely the data is to be drawn from the corresponding distribution.

5 Result Analysis

Based on the results of the *Kolmogorov-Smirnov Test* presented in the previous section, and at the significance level of 0.05, all of the three considered null hypothesis[2] cannot be rejected. Nevertheless, the test for the Logistic probability distribution registers the higher *p-value*, meaning that this is the most likely distribution that justifies the observed data.

In addition, the Logistic distribution that justifies the observed data, at the significance level of 0.05 has a *location* parameter $\alpha = 7.946271$ and a *scale* parameter $\beta = 0.9224431$.

Once a probability distribution has been fitted to the observed data, further questions can be formulated and answered about such data. In the current work the following two questions were raised and, based on the results presented so far in this document, answered:

1. *What is the probability of the concentration of Strontium-90 in milk ranging between 9 and 10 micromicrocuries?*
2. *What is the concentration of Strontium-90 in milk that has a probability of 0.05 of being exceeded?*

In the first question we want to compute the value of $P(9 \leq X \leq 10)$, where X is a continuous random variable representing the concentration of Strontium-90 in milk (in micromicrocuries). Making use of the cumulative distribution function the previous expression can be written as: $P(9 \leq X \leq 10) = F(10) - F(9)$, yielding a probability of 0.1444894.

In the second question a value $c \in \mathbb{R}$ is intended to be determined such that $P(X > c) = 0.05$. This expression is equivalent to $P(X \leq c) = 0.95$, which means that we want to find out what is the quantile of order 95 of our continuous random variable X. Making use of the *qlogis* function of the *R* software environment, this value can be computed to 10.66235.

[2] Which state that the observed data follows the Gamma, Logistic and Normal distributions.