Distribution Overview

Falko Bause, Philipp Gerloff, Jan Kriege, Daniel Scholtyssek
Informatik IV, TU Dortmund
profido@ls4.cs.uni-dortmund.de

April 07, 2014

This documents collects a few relevant information of each distribution supported by node `DistFit` of ProFiDo (Processes Fitting Toolkit Dortmund).

1 Basic Definitions and Notation

In the following sections we summarize properties and fitting methods for distributions supported by ProFiDo. Starting from a trace $T = (x_1, \cdots, x_n)$ the fitting methods adjust the parameters of the distributions such that the distributions approximate characteristics of the trace.

We define the following measures from the trace. Estimators for the mean and variance are given by

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$S^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}(n)\right)^2,$$

respectively. Furthermore the skewness and kurtosis are used in some of the presented estimators. We use the following estimators for the skewness

$$\overline{v}(n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}(n)}{S(n)}\right)^3$$

and for the kurtosis

$$\overline{w}(n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{X}(n)}{S(n)}\right)^4$$

respectively.

From the above estimators one can derive estimators for the parameters of a distribution. If a parameter is denoted by λ we will denote its estimator by $\hat{\lambda}$.
A measure for the quality of the estimated parameter is the likelihood. The likelihood is defined for a trace T and a distribution. Using the probability density function $f(x|\Theta)$ with parameters Θ it is computed as

$$\mathcal{L}(\Theta, T) = \prod_{i=1}^{n} f(x_i|\Theta).$$

For computational reasons often the log-likelihood is used instead:

$$\log \mathcal{L}(\Theta, T) = \sum_{i=1}^{n} \log f(x_i|\Theta).$$

2 Exponential distribution

2.1 Properties

Property	Value
Parameter	λ
pdf	$f(x) = 1/\lambda e^{-x/\lambda}$
CDF	$F(x) = 1 - e^{-x/\lambda}$
Mean	$E[X] = \lambda$
Variance	$VAR[X] = \lambda^2$
MLE	$\hat{\lambda} = \overline{X}(n)$

Further details can be found in [3, p. 300].

2.2 Fitting Approach

MLE[1] can be used directly.

2.2.1 Fitting Approach supported by node DistFit

DistFit uses the MLE approach.

3 Normal distribution

3.1 Properties

Property	Value
Parameter	μ, σ
pdf	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)\right]$
Mean	$E[X] = \mu$
Variance	$VAR[X] = \sigma^2$
MLE	$\hat{\mu} = \overline{X}(n), \hat{\sigma} = \left[\frac{n-1}{n} S^2(n)\right]^{1/2}$

$$\operatorname{erf}(x) \;=\; \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, dt \tag{1}$$

Further details can be found in [3, p. 305].

[1] MLE = Maximum Likelihood Estimator

3.2 Fitting Approach

MLE can be used directly.

3.2.1 Fitting Approach supported by node DistFit

DistFit uses the MLE approach.

4 Lognormal distribution

4.1 Properties

Property	**Value**
Parameter	μ, σ
pdf	$f(x) = \frac{1}{x\sqrt{2\pi}\sigma}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$
CDF	$F(x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left[\frac{\ln x-\mu}{\sqrt{2}\sigma}\right]$
Mean	$E[X] = e^{\mu+\sigma^2/2}$
Variance	$VAR[X] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$
MLE	$\hat{\mu} = \left(\sum_{i=1}^{n} \ln x_i\right)/n,$ $\hat{\sigma} = \left[\left(\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2\right)/n\right]^{1/2}$

$$\mathrm{erf}(x) \;=\; \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, dt \tag{2}$$

Further details can be found in [3, p. 307].

4.2 Fitting Approach

MLE can be used directly.

4.2.1 Fitting Approach supported by node DistFit

DistFit uses the MLE approach.

5 Johnson Family of Distributions

The Johnson Translation System defines a family of distributions related to the normal distribution. The family consists of four classes: lognormal S_L, unbounded S_U, bounded S_B and normal S_N.

5.1 Properties

Property	**Value**
Parameter	$\gamma, \delta, \xi, \lambda$
CDF	$F(x) = \Phi\left(\gamma + \delta f\left(\frac{x-\xi}{\lambda}\right)\right)$

where $\Phi()$ is the standard normal cdf and $f()$ is defined as follows

$$f(y) = \begin{cases} \log(y) & \text{lognormal } S_L \\ \log(y + \sqrt{y^2+1}) & \text{unbounded } S_U \\ \log\left(\frac{y}{1-y}\right) & \text{bounded } S_B \\ y & \text{normal } S_N \end{cases}$$

5.2 Fitting Approach

Fitting is possible according to moments [1], using least squares [5] or according to quantile estimators [7] or likelihood [6]. For a comparison of fitting methods see [4].

5.2.1 Fitting Approach supported by node DistFit

DistFit performs a moment fitting approach as described in [2] and uses the given code translated to C/C++ by the unix tool f2c (Fortran to C/C++).

6 Uniform distribution

6.1 Properties

Property	Value
Parameter	a, b
pdf	$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a,b) \\ 1 & \text{for } x \geq b \end{cases}$
Mean	$E[X] = \frac{1}{2}(a+b)$
Variance	$VAR[X] = \frac{1}{12}(b-a)^2$
MLE	$\hat{a} = \min\limits_{1 \leq i \leq n} x_i, \hat{b} = \max\limits_{1 \leq i \leq n} x_i$

Further details can be found in [3, p. 299].

6.2 Fitting Approach

MLE can be used directly.

6.2.1 Fitting Approach supported by node DistFit

DistFit uses the MLE approach.

7 Triangular distribution

7.1 Properties

Property	Value
Parameter	a, b, c
pdf	$f(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \le x \le c, \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x \le b, \\ 0 & \text{for } b < x. \end{cases}$
CDF	$F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{(x-a)^2}{(b-a)(c-a)} & \text{for } a \le x \le c, \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)} & \text{for } c < x \le b, \\ 1 & \text{for } b < x. \end{cases}$
Mean	$E[X] = \frac{a+b+c}{3}$
Variance	$VAR[X] = \frac{a^2+b^2+c^2-ab-ac-bc}{18}$

Further details can be found in [3, p. 317].

7.2 Fitting Approach

Set $\hat{a} = \min\limits_{1 \le i \le n} X_i$ and $\hat{b} = \max\limits_{1 \le i \le n} X_i$. c is the mode of the distribution, i.e. the point with the largest density. This can be estimated from a histogram of the trace by choosing the largest bin. A further possibility is to equate the mean with the mean of the trace giving $\hat{c} = 3\overline{X}(n) - (\hat{a} + \hat{b})$.

7.2.1 Fitting Approach supported by node DistFit

DistFit uses the described fitting approach, i.e. the parameters are determined by $\hat{a} = \min\limits_{1 \le i \le n} X_i, \hat{b} = \max\limits_{1 \le i \le n} X_i$ and $\hat{c} = 3\overline{X}(n) - (\hat{a} + \hat{b})$. In order to avoid zero probability for the minimum and maximum values of the trace, a small relative margin is applied to the estimators leading to the final parameter estimate as $\hat{a} = \hat{a} - \frac{|\hat{a}-\hat{b}|}{100}$ and $\hat{b} = \hat{b} + \frac{|\hat{a}-\hat{b}|}{100}$

8 Erlang distribution

8.1 Properties

Property	Value
Parameter	k, λ
pdf	$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$
CDF	$F(x) = \frac{\gamma(k, \lambda x)}{(k-1)!} = 1 - \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\lambda x} (\lambda x)^n$
Mean	$E[X] = \frac{k}{\lambda}$
Variance	$VAR[X] = \frac{k}{\lambda^2}$

Further details can be found in [3].

8.2 Fitting Approach

Fitting can be performed using the relationship between Gamma and Erlang distributions.

8.2.1 Fitting Approach supported by node DistFit

DistFit uses a moment fitting approach for the Gamma dstribution and adjusts the parameters in order to determine a valid Erlang distribution.

Let $\hat{\alpha}$ and $\hat{\beta}$ be the parameters obtained from fitting a gamma distribution from the given trace data (cf. Sect. 9). Then the parameters $\hat{\lambda}$ and $\hat{k}$ are determined as follows:

$$\hat{k} = \max(1, round(\hat{\alpha})) \tag{3}$$

$$\hat{\lambda} = \frac{\hat{k}}{\overline{X}(n)} \tag{4}$$

9 Gamma distribution

9.1 Properties

Property	**Value**
Parameter	α, β
pdf	$f(x) = \begin{cases} \frac{\beta^{-\alpha} x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)} & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases}$
CDF	$F(x) = \begin{cases} 1 - e^{-x/\beta} \sum_{j=0}^{\alpha-1} \frac{(x/\beta)^j}{j!} & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases}$
Mean	$E[X] = \alpha\beta$
Variance	$VAR[X] = \alpha\beta^2$
MLE	Equation 5 and Equation 6 must be satisfied.

Further details can be found in [3, p. 302].

9.2 Fitting Approach

The following equations must be satisfied for MLE estimation [3, p. 303]:

$$\ln \hat{\beta} + \Psi(\hat{\alpha}) = \frac{\sum\limits_{i=1}^{n} \ln x_i}{n}, \tag{5}$$

$$\hat{\alpha}\hat{\beta} = \overline{X}(n), \tag{6}$$

with

$$\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}. \tag{7}$$

Both equations can be solved numerically.

9.2.1 Fitting Approach supported by node DistFit

DistFit uses a different approach based on moment fitting:

$$\hat{\beta} = \frac{S^2(n)}{\overline{X}(n)} \tag{8}$$

$$\hat{\alpha} = \frac{\overline{X}(n)}{\hat{\beta}} \tag{9}$$

References

[1] David J. DeBrota, Stephen D. Roberts, James J. Swain, Robert S. Dittus, James R. Wilson, and Sekhar Venkatraman. Input Modeling with the Johnson System of Distributions. In *Proceedings of the 20th Winter Simulation Conference (WSC '88)*, pages 165–179, New York, NY, USA, 1988. ACM.

[2] D. Hill, R. Hill, and R. L. Holder. Algorithm AS 99. Fitting Johnson curves by moments. *Appl. Statist.*, 25(2):180–189, 1976.

[3] Averill M. Law and David W. Kelton. *Simulation Modelling and Analysis*. McGraw-Hill Education - Europe, 2000.

[4] Robert H. Storer, James J. Swain, Sekhar Venkatraman, and James R. Wilson. Comparison Methods for Fitting Data Using Johnson Translation Distributions. In *Proceedings of the 20th Winter Simulation Conference (WSC '88)*, pages 476–481, New York, NY, USA, 1988. ACM.

[5] James J. Swain, Sekhar Venkatraman, and James R. Wilson. Least-Squares Estimation of Distribution Functions in Johnson's Translation System. *Journal of Statistical Computation and Simulation*, 29(4):271–297, 1988.

[6] Efthymios G. Tsionas. Likelihood and Posterior Shapes in Johnson's S_B System. *Sankhya: The Indian Journal of Statistics*, 63(1):3–9, 2001.

[7] R. E. Wheeler. Quantile Estimators of Johnson Curve Parameters. *Biometrika*, 67(3), 1980.