Simulation of probability distributions using two quantile characteristics

M S Tokmachev
Yaroslav-the-Wise Novgorod State University, ul. B. St. Petersburgskaya, 41 173003 Veliky Novgorod, Russia
E-mail: Mikhail.Tokmachov@novsu.ru

Abstract. The paper presents a method for modeling probability distributions based on the base normal distribution. The transformation of the base distribution includes truncating the distribution, splitting it into three components in accordance with two specified quantiles, multiplying the density function in each of the intervals by a linear factor. The distribution obtained in this way has a complex nonstandard configuration, including mixtures of the normal distribution and the Rayleigh distribution. A computer program has been developed that makes it possible to model the distribution according to its parameters and implements the criterion for the agreement of sample data with the distribution of the specified type with automatic adjustment of the parameters. The construction algorithm allows for use of other base distributions.

1. Introduction
In real problems, statistical data are often enough to have distributions that are far from classical (normal, exponential, etc.), and they have a more complex structure. These non-standard probability distributions may be asymmetric, multi-vertex, at separate intervals described by various formulas, etc. At the same time, knowledge of the theoretical distribution is one of the most important conditions in the study of sample data.

There are many methods for modeling distributions. For example, truncation of distributions in various ways [1], use of mixtures of distributions in various combinations [2–5], application of functional transformations of random variables [6–8], etc.

The paper presents a method for modeling probability distributions with a specific structure of the density function. The basis is some basic distribution, in this case, normal with the parameters $m$ and $\sigma$. Then, the density function $f(x)$ of this basis distribution is transformed by truncating the distribution over a given interval, introducing two quantiles, $x_\alpha$ and $x_\beta$, at this interval and adding linear factors on each of the resulting intervals. As a result, the density function $g(x)$ of the new distribution is

$$
g(x) = \begin{cases} 
(k_1x + c_1)f(x) & \text{при } x \in [a; x_\alpha), \\
(k_2x + c_2)f(x) & \text{при } x \in [x_\alpha; x_\beta), \\
(k_3x + c_3)f(x) & \text{при } x \in [x_\beta; b], \\
0 & \text{при } x \notin [a; b].
\end{cases}
$$

(1)

In formula (1) the values $k_i, c_i (i = 1, 3)$ are calculated based on the input parameters $a, b, \alpha, \beta, x_\alpha, x_\beta$. 

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Distributions with a similar structure were considered by the author earlier [9, 10] in cases of a single quantile with basic normal and exponential distributions, as well as with quadratic factors.

Modeling distributions with two quantile characteristics complicates the structure of the model, introduces additional parameters. At the same time, the possibility of varying the parameters and selecting the theoretical distribution according to real data is significantly expanded.

The theoretical distribution with a density of the form (1) can be selected on the basis of the sample. With the help of a computer program, the distribution parameters are fitted, the values of $k_i, c_i (i = 1, 3)$ are calculated, and the compliance of the theoretical and selective distributions is verified using the agreement criterion.

2. The main part

Let us consider an algorithm for modeling probability distributions with density $g(x)$ of the form (1). The base distribution is assumed normal with the density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Next, we select the parameters of the new distribution. Position parameters: $a, b$ are truncation boundaries, $x_\alpha, x_\beta$ are the quantiles of the corresponding orders. Concentration parameters: $\alpha, \beta$, where $\alpha = P(X \in [a; x_\alpha]), \beta = P(X \in [x_\alpha; x_\beta)).$ Correspondingly, $P(X \in [x_\beta; b]) = 1 - \alpha - \beta \equiv \gamma$. That is, $\gamma$ is not an independent parameter, and its value is uniquely determined by the values of $\alpha$ and $\beta$. An obvious correlation is made:

$$\alpha + \beta + \gamma = 1.$$

To construct the function $g(x)$, the following values are also required: $g(a), g(x_\alpha), g(x_\beta), g(b)$.

Based on the indicated values, the density function $g(x)$ can be constructed differently, in particular, it can be continuous or discontinuous at the point $x = x_\alpha$ and $x = x_\beta$. Accordingly, some of the above parameters may be interrelated. Natural restrictions are also imposed, such as $\alpha, \beta, \gamma > 0, a < x_\alpha < x_\beta < b, g(x) \geq 0$.

With fixed parameters, the problem is reduced to calculating the coefficients $k_i, c_i (i = 1, 3)$ and checking the normalizing conditions for the function $g(x)$, ensuring that the function found can be a probability distribution density.

We define the coefficients so that the function $g(x)$ is continuous on $[a; b]$. In particular, we eliminate the possible discontinuity of continuity for $x = x_\alpha$ and $x = x_\beta$. Also, when determining the coefficients, it is necessary to take into account the values of the concentration parameters: $\alpha, \beta, \gamma$.

Based on this, with fixed non-negative $g(x_\alpha), g(x_\beta)$ we arrive at a system of linear algebraic equations for the coefficients $k_i, c_i (i = 1, 3)$.

In this system of seven linear algebraic equations, six unknowns should be found: $k_1, k_2, k_3, c_1, c_2, c_3$. Since the number of equations exceeds the number of unknowns, it may not be only solutions satisfying the formulated constraints, but no system solution at all. More specifically, the problem is that in system (2) for unknowns $k_2, c_2$, not two, but three linear equations are given. The easiest way is to drop one of the equations, for example, the third equation in system (2). But then the continuity of $g(x)$ at the point $x = x_\beta$ is not guaranteed.

You can go the other way: save the continuity condition $g(x)$ (leave all seven equations), but enter one of the parameters as unknown. Such an unknown parameter is $g(x_\beta)$, which is linearly included in the equations of system (2). Consequently, we obtain a system of seven linear algebraic equations with seven unknowns. In this case, the continuity of the function $g(x)$ is guaranteed at the point $x = x_\beta$. For arbitrary values of the parameters, the solution of system (2) is unique, but not necessarily probabilistic: the function $g(x)$ found can take negative values, i.e. not be a probability density.
Therefore, the non-specified, and the calculated value of \( g(x_\beta) \) must also be checked for non-negativity.

Let us introduce sufficient conditions for the non-negativity of \( g(x) \).

**Lemma 1.** In order for the function \( g(x) \) of the form (1), found according to the solution of system (2), to be non-negative for acceptable parameters, it suffices to satisfy the inequalities

\[
k_1 a + c_1 \geq 0, \quad k_3 b + c_3 \geq 0.
\]

**Proof.** Since the base density function \( f(x) \) is everywhere nonnegative, for the condition \( g(x) \geq 0 \) to be valid, the linear factors are to be nonnegative:

\[
\begin{align*}
&k_1 x + c_1 \geq 0, \text{ при } x \in [a; x_\alpha], \\
&k_2 x + c_2 \geq 0, \text{ при } x \in [x_\alpha; x_\beta], \\
&k_3 x + c_3 \geq 0, \text{ при } x \in [x_\beta; b].
\end{align*}
\]

The non-negativity of a linear function at two points \( x_0, x_1 \) guarantees its non-negativity on the entire interval \([x_0; x_1]\). Therefore, to be fair (4), the sign of the factors at the corresponding boundary points \( a, x_\alpha, x_\beta, b \) should be checked, namely:

\[
\begin{align*}
&k_1 a + c_1 \geq 0, \quad k_1 x_\alpha + c_1 \geq 0, \\
&k_2 x_\alpha + c_2 \geq 0, \quad k_2 x_\beta + c_2 \geq 0, \\
&k_3 x_\beta + c_3 \geq 0, \quad k_3 b + c_3 \geq 0.
\end{align*}
\]

The ratios at the points \( x_\alpha, x_\beta \) are a priori valid, based on the first four equations of system (2), the non-negativity of the density function \( f(x) \) and the initial conditions for (2) \( g(x_\alpha) \geq 0, \quad g(x_\beta) \geq 0 \). So, of the six inequalities in (5), only two remain as conditions: \( k_1 a + c_1 \geq 0, \quad k_3 b + c_3 \geq 0 \). We arrived at conditions (3).

Lemma is proved.

Comment. Note that the values of \( g(a), g(b) \), in contrast to \( g(x_\alpha), g(x_\beta) \), are absent in system (2), and the conditions for their non-negativity must be controlled.

**Lemma 2.** The integral of a form

\[
I = \int_a^b (kx + c) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx
\]
is expressed through the Laplace function $\Phi(x)$ and the density function of the standard normal law $\varphi(x)$, namely:

$$I = k \cdot \left\{ m \cdot \left[ \phi \left( \frac{b - m}{\sigma} \right) - \phi \left( \frac{a - m}{\sigma} \right) \right] - \sigma \cdot \left[ \varphi \left( \frac{b - m}{\sigma} \right) - \varphi \left( \frac{a - m}{\sigma} \right) \right] \right\} + c \cdot \left\{ \phi \left( \frac{b - m}{\sigma} \right) - \phi \left( \frac{a - m}{\sigma} \right) \right\},$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt,$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

This relation is obtained directly by integrating the sum of functions after changing the variable $t = \frac{x - m}{\sigma}$.

Further, for a complete definition of the function $g(x)$ for given $a, b, m, \sigma$, we turn to system (2). To preserve the continuity of the density function $g(x)$, we set $g(x_0)$ as an unknown non-negative variable.

**Theorem 1.** Let on the interval $[a; b]$ the base density function $f(x)$ of the normal distribution $N(m; \sigma)$ is set. Then the coefficients $k_i, c_i (i = 1, 3)$ and the value $g(x_0)$ of the probability density function $g(x)$ of the form (1) with the parameters $\alpha, \beta, x_\alpha, x_\beta, g(x_\alpha)$ can be found as a solution of a system of linear algebraic equations

$$\begin{cases}
    k_1x_\alpha + c_1 = \frac{g(x_\alpha)}{f(x_\alpha)}, \\
    k_2x_\alpha + c_2 = \frac{g(x_\alpha)}{f(x_\alpha)}, \\
    k_2x_\beta + c_2 = \frac{g(x_\beta)}{f(x_\beta)}, \\
    k_3x_\beta + c_3 = \frac{1}{f(x_\beta)} = 0, \\
    k_3A_1 + c_1B_1 = A, \\
    k_2A_2 + c_2B_2 = B, \\
    k_3A_3 + c_3B_3 = \gamma
\end{cases}
$$

with additional restriction $k_1a + c_1 \geq 0, k_3b + c_3 \geq 0, g(x_0) \geq 0$. Coefficients $A_i, B_i$ are

$$A_1 = m \cdot \left[ \phi \left( \frac{x_\alpha - m}{\sigma} \right) - \phi \left( \frac{a - m}{\sigma} \right) \right] - \sigma \cdot \left[ \varphi \left( \frac{x_\alpha - m}{\sigma} \right) - \varphi \left( \frac{a - m}{\sigma} \right) \right],$$

$$A_2 = m \cdot \left[ \phi \left( \frac{x_\beta - m}{\sigma} \right) - \phi \left( \frac{x_\alpha - m}{\sigma} \right) \right] - \sigma \cdot \left[ \varphi \left( \frac{x_\beta - m}{\sigma} \right) - \varphi \left( \frac{x_\alpha - m}{\sigma} \right) \right],$$

$$A_3 = m \cdot \left[ \phi \left( \frac{b - m}{\sigma} \right) - \phi \left( \frac{x_\beta - m}{\sigma} \right) \right] - \sigma \cdot \left[ \varphi \left( \frac{b - m}{\sigma} \right) - \varphi \left( \frac{x_\beta - m}{\sigma} \right) \right],$$

$$B_1 = \phi \left( \frac{x_\alpha - m}{\sigma} \right) - \phi \left( \frac{a - m}{\sigma} \right),$$

$$B_2 = \phi \left( \frac{x_\beta - m}{\sigma} \right) - \phi \left( \frac{x_\alpha - m}{\sigma} \right),$$

$$B_3 = \phi \left( \frac{b - m}{\sigma} \right) - \phi \left( \frac{x_\beta - m}{\sigma} \right).$$

**Proof.** System (6) with the required variables $k_1, k_2, k_3, c_1, c_2, c_3, g(x_0)$ and the indicated coefficients is obtained by an elementary transformation of system (2). In the last three equations, the
coefficients $A_i, B_i (i = 1, 3)$ are calculated based on Lemma 2. Conditions in the form of inequalities, according to Lemma 1, guarantee the non-negativity of the function $g(x)$.

Theorem 1 is proved.

The simulation of the density of the probability distribution $g(x)$ of the form (1) is computerized [11] and is carried out automatically according to the given parameters $m, \sigma, a, b, \alpha, \beta, x_\alpha, x_\beta, g(x_\alpha)$.

An example of the density function $g(x)$ of the simulated distribution is shown in figure 1.

![Figure 1. An example of a graph of the probability density function $g(x)$.](image)

Density function with the parameters $m = 46.2, \sigma = 12.961, a = 4.5, b = 79.5, \alpha = 0.793, \beta = 0.096, x_\alpha = 52.801, x_\beta = 57.801, g(x_\alpha) = 0.02$ and calculated $k_1, k_2, k_3, c_1, c_2, c_3, g(x_\beta)$ looks like (1)

$$
g(x) = \begin{cases} 
(-0.0307x + 2.363) \frac{1}{\sqrt{2\pi}12.961} e^{\frac{(x-46.2)^2}{212.961^2}} & \text{with } x \in [4.5; 52.801), \\
(0.0269x - 0.6805) \frac{1}{\sqrt{2\pi}12.961} e^{\frac{(x-46.2)^2}{212.961^2}} & \text{with } x \in [52.801; 57.801), \\
(-0.0398x + 3.1754) \frac{1}{\sqrt{2\pi}12.961} e^{\frac{(x-46.2)^2}{212.961^2}} & \text{with } x \in [57.801; 79.5], \\
0 & \text{with } x \notin [4.5; 79.5].
\end{cases}
$$

The specific values of the coefficients of the density function $g(x)$ are found as solutions of the system of equations (2). The calculated value is $g(x_\beta) = 0.018$. In figure 1, three parts of the graph are clearly distinguished, corresponding to the specified quantile characteristics: $\alpha, \beta, \gamma, x_\alpha, x_\beta$, where $\gamma = 1 - \alpha - \beta$. Note that this type of “refracted” graphs are characteristic of health indicators [12], considered depending on the age of the observed population. With different combinations in the ratio of parameters, the density graph configuration may differ significantly from that shown in figure 1.

The developed program allows not only to simulate the distribution for given parameters, but also checks the sample data for agreement with the theoretical distribution with the density of the form (1). At the same time, the parameters of the theoretical distribution, as well as the values of $m, \sigma, a, b$ are adjusted according to the sample. The Pearson approval criterion applies. The goal of fitting the parameters is to minimize the observed value of $\chi^2$.

3. Conclusion

The method of modeling non-standard probability distributions of the considered type:

- allows you to enter a new class of simulated distributions with two quantile characteristics;
expands the possibilities of selecting the theoretical distribution of selective data;
allows you to explore real data that do not meet the standard laws of distribution.

The developed computer program implements the presented theory.

According to the same scheme, other distributions can also be used as the baseline, with various forms of the density graph. For example, a uniform distribution on a segment and a one-parameter exponential distribution are quite simple in implementation. A more complicated case is the consideration of the hyperbolic cosine type as a basic three-parameter distribution [13–15], which has already been used in the case of specifying one quantile. In this case, the configuration of the simulated distribution turns out to be even more diverse.

Note that the input of linear factors to the base density function converts this function at each of the intervals into a mixture of distributions. In the case presented, the mixture consists of the normal distribution and the Rayleigh distribution [16] with the density function

\[ f(x) = \frac{x - m}{\alpha^2} e^{- \frac{(x-m)^2}{2\alpha^2}} \text{ with } x > m. \]

Thus, in modeling, both truncation, mixture, and functional transformations of distributions were used.

Also, linear factors for the density function \( g(x) \) can be replaced by quadratic ones. Fundamental changes in the algorithmization of computations will not occur. However, due to the increase in the number of variables, the number of equations in system (2) will also increase.

The presented method of modeling probability distributions with a complex structure turned out to be practically realizable and has a specific application in mathematical statistics in the study of sample data. The results of the study can find applications in other mathematical and applied disciplines.

References

[1] Tokmachev M S 2018 Modeling of truncated probability distributions *IOP Conf. Series: Materials Science and Engineering* Electronic edition C 012056
[2] Feller V 1984 *Vvedeniye v teoriyu veroyatnostey i yeyo prilozheniya* [Introduction to probability theory and its applications] (Moscow: Mir) Vol 2 p 752
[3] Lukach E 1979 *Kharakteristicheskiye funktii* [Characteristic functions] (Moscow: Nauka) p 424
[4] Medgyessy P 1977 *Decomposition of superpositions of density functions and discrete distributions* (Budapest: Akademia Kiado)
[5] Ayyazyan S A and Mkhitaryan V S 1998 *Prikladnaya statistika i osnovy ekonometriki* [Applied Statistics and Econometrics Basics] (Moscow: YUNITY) p 1022
[6] *Sbornik zadach po teorii veroyatnostey, matematicheskoy statistike i teorii sluchaynykh funktsiy* [Collection of problems on the theory of probability, mathematical statistics and the theory of random functions] ed A A Sveshnikov 2008 (Saint-Petersburg: Lan’) p 448
[7] Tokmachev M S 2004 *Funktsii sluchaynykh velichin* [Functions of random variables] (VelikiyNovgorod: NovSU) p 74
[8] Rao S R 1968 *Lineynyye statisticheskiye metody i ikh primeneniya* [Linear statistical methods and their applications] (Moscow: Nauka) p 548
[9] Tokmachev M S 2011 Modeling of truncated distributions with a given quantile *Bulletin of A N Tupolev Kazan National Research Technical University* 2 157–162
[10] Tokmachev M S 2011 Modeling truncated distributions with a given quantile based on an exponential distribution *Vestnik NovSU. Ser.: Tech. Sciences* 65 97–100
[11] Nosov D V and Tokmachev M S Simulation of probability distributions with non-standard structure. Certificate of state registration of computer programs No. 2017610154. Application No. 2016661738 dated 03.10.2016. Registered in the Register of computer programs 09.01.2017.
[12] Healthcare in Russia 2017 Stat. compil. (Moscow: Rosstat) p 170
[13] Tokmachev M 1995 Constancy of regression of quadratic statistics on linear statistics *Vestnik NovSU* 1 139–141
[14] Tokmachev M S 2017 Investigation of the probability distribution of the hyperbolic cosine type *SUSU Bulletin Ser.: “Mathematics. Mechanics. Physics”* Vol 9 3 18–26
[15] Tokmachev M S 2005 Applied aspect of the generalized distribution of hyperbolic cosine *Vestnik NovSU* 34 96–99
[16] Vadzinskiy R N 2001 *Spravochnik po veroyatnostnym raspredeleniyam* [Handbook of Probability Distributions] (Saint-Petersburg: Nauka) p 295