Estimation of simple characteristics of samples from skewed and heavy-tailed distributions

Zdeněk Fabián

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2
Prague 18200 Czech Republic
(e-mail: zdenek@cs.cas.cz)

Abstract. We present new characteristics of the central tendency and dispersion of data samples. They are constructed from estimates of parameters of underlying distributions and make possible an easy comparison of results obtained under different assumptions.

Keywords: scalar inference function, generalized moment method.

1 Introduction

Johnson score - a scalar inference function - was introduced by [1,2] for a large class of continuous probability distributions. It was shown that the Johnson score moments exist under mild regularity conditions even in cases of distributions without mean and variance. The first moment describes the central tendency of the distribution and the reciprocal value of the second moment the dispersion of the values around the central point. It seems that whereas the mean \( m = \int x f(x) \) and variance \( \sigma^2 = \int x^2 f(x) - m^2 \) compare the properties of distribution \( F \) with the standard (with the normal distribution), the new characteristics have an ability to compare distributions within parametric families even when distributions are skewed and/or heavy-tailed.

Usually, having an idea about the type of the underlying distribution, it is to estimate the parameters. We argued for a slight change of view: there are the sample Johnson mean and sample Johnson variance, which are to be estimated as characteristics of data samples taken from the distribution under consideration. They make possible to compare results of estimation for various assumed distribution families parametrized by arbitrary ways.

2 Johnson score

Let us define the basic concept.

Definition 1. Let \( F \) be distribution with support \( X = (a, b) \subseteq \mathbb{R} \) and density \( f \) continuously differentiable according to the variable. Let mapping
\( \eta : \mathcal{X} \to \mathbb{R} \) be defined by

\[
\eta(x) = \begin{cases} 
  x & \text{if } (a, b) = \mathbb{R} \\
  \log(x - a) & \text{if } -\infty < a < b = \infty \\
  \log \frac{x - a}{b - x} & \text{if } -\infty < a < b < \infty \\
  \log(b - x) & \text{if } -\infty = a < b < \infty,
\end{cases}
\]

(1)

let function \( T(x) \) be given by

\[
T(x) = \frac{1}{f(x)} \frac{d}{dx} \left( -\frac{1}{\eta'(x)} f(x) \right), \quad x \in \mathcal{X}
\]

(2)

and the solution \( x^* \) of equation

\[
T(x) = 0
\]

(3)

be unique. Function

\[
S(x) = \eta'(x^*) T(x)
\]

(4)

will be called a Johnson score of distribution \( F \).

Since \( \eta'(x) > 0 \), \( x^* \) is the solution of equation

\[
S(x) = 0
\]

(5)

as well. The philosophy behind Definition 1 is the following. Any distribution \( F \) with interval support \( \mathcal{X} \neq \mathbb{R} \) is viewed as a transformed prototype \( G \) supported by \( \mathbb{R} \), that is,

\[
F(x) = G(\eta(x)), \quad x \in \mathcal{X}.
\]

(6)

Mapping \( \eta \) given by (1) is the Johnson transformation [3] adapted for arbitrary interval support. Denoting by \( g \) the density of \( G \), the density of the transformed prototype (6) is

\[
f(x) = g(\eta(x)) \eta'(x), \quad x \in \mathcal{X}
\]

(7)

where \( \eta'(x) \) is the Jacobian of the transformation. Let \( Q \) be the score function of \( G \),

\[
Q(x) = -\frac{g'(x)}{g(x)}.
\]

(8)

While the score function can be taken as a suitable inference function of prototypes, the transformed score function of the prototype,

\[
T(x) = Q(\eta(x)),
\]

(9)

was found to be a relevant inference function of (6. It is termed by [4] a core function of \( F \). Formula (2) follows from (9) by using (8) and (7) and shows
that the core function can be determined without reference to the prototype by a special type of differentiating of the density according to the variable.

Let \( G(y - \mu) \) be a prototype with location parameter \( \mu \in \mathbb{R} \) (expressing its central tendency). Consider distribution which is the transformed \( G(y - \mu) \) on \((a, b) \neq \mathbb{R}\). Set

\[
t = \eta^{-1}(\mu),
\]

(10) is called a Johnson parameter and will be considered even in multi-parameter cases to express the central tendency of the transformed distribution, the density and core function of which are

\[
f(x; t) = g(\eta(x) - \eta(t))\eta'(t)
\]

and

\[
T(x; t) = Q(\eta(x) - \eta(t)).
\]

It was shown by [4] that

\[
\frac{\partial}{\partial t} \log f(x; t) = \eta'(t)T(x; t).
\]

Johnson score \( S(x) = \eta'(t)T(x; t) \) of the transformed distribution with Johnson parameter thus equals to the likelihood score for this parameter.

However, there are distributions without Johnson parameter and without parameters at all. Johnson score (4) is a generalization of relation (11) for these distributions. The generalization consists of replacing \( t \) in term \( \eta'(t) \) by the zero of the core function, which actually means a replacing the location parameter of the prototype \( G \) by the mode of \( G \). Let us note that the solution of (3) is unique if \( G \) is unimodal; in cases of multimodal distributions we consider that \( x^* \) is the image of the global maximum of the density. Hence \( x^* \) is the most important point of the transformed distribution, characterizing its central tendency, and Johnson score can be interpreted as the ‘score’ for this point, which may or may not be a parameter of the distribution.

Densities and Johnson scores of distributions used throughout the paper are given in Table 1.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( \mathcal{X} )</th>
<th>( f(x) )</th>
<th>( S(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>( \mathbb{R} )</td>
<td>( \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} )</td>
<td>( \frac{x-\mu}{\sigma} )</td>
</tr>
<tr>
<td>lognormal</td>
<td>((0, \infty))</td>
<td>( \frac{1}{\sqrt{2\pi^2 x}} e^{-\frac{1}{2} \log^2(\frac{x}{t})} )</td>
<td>( \frac{1}{t} \log(x/t) )</td>
</tr>
<tr>
<td>Weibull</td>
<td>((0, \infty))</td>
<td>( \frac{c}{x} (\frac{x}{t})^{c-1} e^{-\left(\frac{x}{t}\right)^c} )</td>
<td>( \frac{1}{t} \left[\left(\frac{x}{t}\right)^c - 1\right] )</td>
</tr>
<tr>
<td>Fréchet</td>
<td>((0, \infty))</td>
<td>( \frac{c}{x} (\frac{x}{t})^{c-1} e^{-\left(\frac{x}{t}\right)^c} )</td>
<td>( \frac{1}{t} \left[\left(\frac{x}{t}\right)^c - 1\right] )</td>
</tr>
<tr>
<td>gamma</td>
<td>((0, \infty))</td>
<td>( \frac{1}{\Gamma(p/q)} x^{\alpha-1} e^{-\gamma x} )</td>
<td>( \frac{\Gamma(p/q) x^{\alpha-1} e^{-\gamma x}}{\Gamma(p/q) x^{\alpha-1} e^{-\gamma x}} )</td>
</tr>
<tr>
<td>beta-prime</td>
<td>((0, \infty))</td>
<td>( \frac{1}{B(p,q)} x^{p-1} )</td>
<td>( \frac{1}{B(p,q)} x^{p-1} )</td>
</tr>
</tbody>
</table>

Table 1. Densities and Johnson scores of some model distributions. \( \Gamma \) is the gamma function, \( B \) the beta function.

The use of the modified Johnson transformation (1) in Definition 1 leads to Johnson scores expressed by simple formulas for many commonly used
probability distributions. Moreover, it is the only transformation under which the prototype of the lognormal distribution is the normal distribution and the inference function of the uniform distribution is linear.

3 Johnson mean and Johnson variance

The Johnson score moments of distribution $F$ with Johnson score $S$ are defined by

$$ES^k = \int_X S^k(x) \, dF(x), \quad k = 1, 2, \ldots$$

(12)

It can be easily seen that $ES = 0$. From the Cramér-Rao regularity conditions imposed on $F$ it follows that $0 < ES^2 < \infty$.

**Definition 2.** Let $F$ be distribution regular in the Cramér-Rao sense with Johnson score $S$. The solution $x^*$ of equation $S(x) = 0$ will be called a Johnson mean and value

$$\omega^2 = (ES^2)^{-1}$$

(13)

a Johnson variance of $F$.

By (4), (13) turns for distributions with support $X = (0, \infty)$ into

$$\omega^2 = (x^*)^2/E_T^2.$$ 

(14)

Let us discuss Johnson characteristics of distributions from Table 1.

The zero of $S(x) = Q(x)$ of the normal distribution is $x^* = \mu$. Since $EQ^2 = 1/s^2$, Johnson mean and Johnson variance are equal to the mean and variance.

Parameter $t$ of the lognormal, Weibull and Fréchet distributions is the Johnson parameter. Fig. 1 shows densities and Johnson scores of three particular cases of Weibull distribution with $c = 1$ (exponential distribution), $c = 2$ (Rayleigh distribution) and $c = 3$ (Maxwell distribution). Johnson means of all three distributions are $x^* = 1$, the means (denoted by stars) have similar values, $m(1) = x^*$.

Fréchet distribution is a heavy-tailed distribution and its mean $m = t\Gamma(1 - 1/c)$ and variance $\sigma^2 = t^2[\Gamma(1 - 2/c) - \Gamma^2(1 - 1/c)]$ exist only if $c > 1$ and $c > 2$, respectively. Fig. 2 shows the densities and Johnson scores of Fréchet distributions with Johnson mean $x^* = 1$. Neither of the means (the stars) describes the position of the distribution on $x$-axis. $m(1)$ does not exist.

Fig. 3 shows densities of Fréchet distributions with various Johnson means. The variability of values around Johnson mean is apparently similar to all four distributions. Actually, they have the same Johnson variance $\omega^2 = 1$.

Gamma and beta-prime distributions are examples of distributions without Johnson parameter. By (2), the core function of the gamma distribution is $T(x) = \gamma x - \alpha$ so that $x^* = \alpha/\gamma$. Since $ET^2 = \alpha$, by (14)
Fig. 1. Densities (a) and Johnson scores (b) of Weibull distributions with $x^* = 1, c = 1, 2, 3$. The means $m(c)$ are denoted by stars. $m(1) = 1, m(2) = 0.885, m(3) = 0.893$.

\[ \omega^2 = (x^*)^2/\alpha = \alpha/\gamma^2. \]  

Johnson mean and Johnson variance of the gamma distribution are thus equal to the mean and variance. On the other hand, the mean $m = p/(q - 1)$ and variance

\[ \sigma^2 = \frac{p(p + q + 1)}{(q - 1)^2(q - 2)} \]  

of the beta-prime distribution exist only if $q > 1$ and $q > 2$, respectively. By (2), $T(x) = (qx - p)/(x + 1)$, by (3) $x^* = p/q, ET^2 = pq/(p + q + 1)$ and Johnson variance is, by (14),

\[ \omega^2 = \frac{(x^*)^2}{ET^2} = \frac{p(p + q + 1)}{q^3}. \]  

Note that (16) looks like (15) with a 'corrected denominator'.

Standard deviation and the square root of Johnson variance of the beta-prime distribution with $p = q$ as functions of $1/q$ are compared on Fig. 4. $\sigma$ blows up at $1/q = 1/2$ whereas 'Johnson deviation' $\omega$ is comparable with the simulated average median absolute deviation.
Fig. 2. Densities (a) and Johnson scores (b) of Fréchet distributions with $x^* = 1$, $c = 1, 1.5, 2$. The means $m(c)$ are denoted by stars. $m(1)$ do not exist, $m(1.5) = 2.68, m(2) = 1.77$.

Fig. 3. Densities of Fréchet distributions, $t = 1, 2, 3, 4, \omega = 1$.

4 Estimates

Let $X_1, \ldots, X_n$ be random variables i.i.d. according to $F_\theta, \theta \in \Theta, \Theta \subset \mathbb{R}^m$ with unknown $\theta$ and $x_1, \ldots, x_n$ their observed values. The structure of parameters of various distributions, the consequence of the historical development of mathematical statistics, exhibit a chaotic picture. It is difficult to compare results of the estimation for distribution families parametrized by different ways.
Fig. 4. Beta-prime distribution. 1 - σ, 2 - ω, 3 - simulated MAD

Both the Johnson mean $x^* : S(x; \theta) = 0$ and Johnson variance $\omega^2 = 1/ES^2(\theta)$ are functions of $\theta$ and can be constructed from the maximum likelihood estimate $\hat{\theta}_{ML}$ of $\theta$. In what follows $AN$ means 'asymptotically normal'. Since $ES^2(\theta) > 0$, numbers $\hat{x}_{ML}^* = x^*(\hat{\theta}_{ML})$ and $\hat{\omega}^2_{ML} = \omega^2(\hat{\theta}_{ML})$ characterize a 'center' and dispersion of the sample. Their asymptotic behavior can be easily established by using the delta method theorem, saying that if $\hat{\theta}$ is $AN(\theta, \sigma^2)$ and $\varphi(\theta)$ is differentiable at $\theta$ with $\varphi'(\theta) \neq 0$, $\varphi(\hat{\theta})$ is $AN(\varphi(\theta), [\varphi'(\theta)]^2\sigma^2)$ (Corollary to Theorem A, [5], pp.122).

Another possibility to estimate $x^*$ and $\omega$ is discussed by [1]. Unlike the usual moments, the sample versions of Johnson score moments cannot be determined without an assumption about the underlying distribution family. On the other hand, by substituting the empirical distribution function into (12), a system of equations

$$\frac{1}{n} \sum_{i=1}^{n} S^k(x_i; \theta) = ES^k(\theta), \quad k = 1, ..., m, \quad (17)$$

appears to be an alternative to the maximum likelihood equations in the whole range of the parameters. The estimates from (17) were shown to be asymptotically normal and, in cases of families with bounded Johnson scores, robust and with relative efficiencies near to one.

The first equation of (17) for distributions with Johnson parameter is the maximum likelihood equation for this parameter. New results are obtained for distributions without this parameter. Thus, the first equation (17) for gamma distribution is $\sum_{i=1}^{n} (\gamma x_i - \alpha) = 0$, from which the estimate of $x^* = \alpha/\gamma$ is $\hat{x}^*_n = n^{-1} \sum_{i=1}^{n} x_i$. From the first equation of (17) for the beta-prime distribution, $\sum_{i=1}^{n} (qx_i - p)/(x_i + 1) = 0$, the estimate of $x^* = p/q$ is

$$\hat{x}_n^* = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} \frac{1}{1 + x_i}}. \quad (18)$$

In a simulation study, samples of length 100 were generated consecutively from each distribution listed in rows of Table 3, each with values of $\theta$ giv-
ing $x^*(\theta) = 1$ and $\omega(\theta) = 1.118$. Both $x^*$ and $\omega$ were estimated under the assumption of either distribution listed in headlines of columns in Table 3, where the average values over 5000 samples are summarized. It is apparent that erroneous assumptions often lead to unacceptable estimates (note, however, the similar results obtained under assumptions of the lognormal and beta-prime distributions). By the use of the estimates of the Johnson mean and Johnson variance, it is easy to compare the results of estimation for various assumed distribution families parametrized by arbitrary ways.

\begin{table}[h]
\centering
\begin{tabular}{l|cccc}
\hline
 & Weibull & gamma & lognormal & beta-prime & Fréchet \\
\hline
$x^*$ & 1.003 & 0.94 & 1.66 & 1.77 & 4.05 \\
Weibull & 1.06 & 0.998 & 1.65 & 2.01 & 22.9 \\
gamma & 0.53 & 0.49 & 1.013 & 1.01 & 1.93 \\
lognormal & 0.67 & 0.64 & 1.01 & 1.006 & 1.53 \\
beta-prime & 0.26 & 0.24 & 0.62 & 0.59 & 1.016 \\
Fréchet & & & & & \\
\hline
$\omega$ & 0.925 & 0.85 & 1.745 & 1.63 & 2.55 \\
Weibull & 1.16 & 1.080 & 1.97 & 3.41 & 127. \\
gamma & 1.43 & 1.49 & 0.997 & 1.11 & 1.42 \\
lognormal & 0.83 & 0.81 & 0.99 & 1.103 & 2.09 \\
beta-prime & 0.18 & 0.15 & 0.66 & 0.55 & 0.931 \\
Fréchet & & & & & \\
\hline
\end{tabular}
\caption{Estimates of Johnson mean and Johnson deviation}
\end{table}

Acknowledgements. The work was supported by GA ASCR under grant No. 1ET 40030013.

References