ATLANTIS PRESS
Atlantis Studies in Uncertainty Modelling, volume 3
Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th International Summer School on Aggregation Operators (AGOP)

Quantile Based Summation of Random Variables

Kadir Emir[a] and *__David Kruml__[b] and **Jan Paseka**[c] and **Iveta Selingerová**[d]
[a,b,c,d]Department of Mathematics and Statistics, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic
[a]emir@math.muni.cz, [b]kruml@math.muni.cz, [c]paseka@math.muni.cz, [d]selingerova@math.muni.cz

Abstract

A simplified calculus for aggregation of random variables is derived. The calculus provides a fast and simple estimation of a given quantile. It can be applied for testing of desired reliability of a production plan.

Keywords: Fuzzy probability, production plan, reliability in manufacturing, aggregation of random effects, order statistics, quantile estimation.

1 Introduction

In manufacturing or other kinds of complex activity, particular events perform with some amount of randomness. Then it is hard to model and guess the overall reliability of a given production plan or process. We face the problem of how to aggregate the effects of the random processes.

One of the critical objectives in manufacturing is to reduce excessive product variability [4, 6], and many people are devoted to this task. Since we have many passively collected data obtained in the manufacturing process, we can use variance components analysis to analyze received data to identify the key causes of variation and each cause's contribution to the total variance. This approach allows us to develop cost-effective improvement strategies.

Since we obtain our data using a measurement system and measurements are often inexact, we have to count also with the variability of our measurement system. Measurement system variability involves both the repeatability and the reproducibility of the measurement system. Hence the final observed variability includes both the true product-to-product variability and the measurement system variability.

From a theoretical point of view, we assume that random variables are precisely defined, and there is a precise calculus for their summation. However, the actual data are often insufficient to make a correct decision about probability distribution parameters or even about their type. Since there is a lack of information, it can not be significant to use precise but complex mathematical methods. Finally, a planner need not have the necessary skill in probability theory and statistics, access to mathematical software, or computational power. It makes sense to seek a simplified calculus for all of these reasons.

In [3], we have already suggested a simplified method for the aggregation of random effects based on quantiles' and moments' properties. Since the resulting calculus is approximate on inputs and outputs, it can contribute to "fuzzified" probability theory. The aggregation can be either serial or parallel. For example, in the former case, we can sum the effects of more processes on one product, or more products worked in one process. In the latter case, we can add output streams of more processes (machines). We distinguish two main tasks:

Summation of independent random variables. A particular case is that we sum multiple identical copies of a variable.

Change of coordinates. We need to model a dependent variable "orthogonal" to a given variable. A typical situation is a relation between time and processed material — we know when a given quantity is processed, and we need to know how many products are made at a given time.

Combining the two tasks allows to aggregate parallel effects (because the orthogonal variables are in a serial position) or estimate stock development.

It is well known that mean, variance, and the third central moment are additive, i. e. the sum of independent

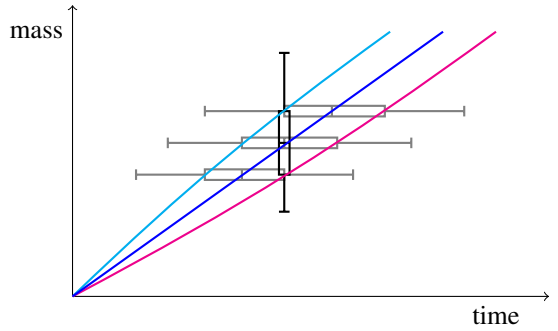

Figure 1: Point (t,N) in the picture represents a state that mass N is produced at time t. It is equivalent to say that some run of the process is slower because it produces less in the same time or takes more time to produce the same mass. Thus we can model the vertical (mass, black) distribution from development of quantiles of the horizontal (time, gray) distribution.

variables is parametrized by sums of the moments [7]. Thus these moments are promising characteristics for the simplified summation of random variables. We believe that they can model the actual probability distributions with accuracy sufficient for many practical applications.

On the other hand, the two variables mentioned in the task "change of coordinates" share quantiles — at any point, we can speak about slower or faster runs, and such meaning is the same for both "views" (see Figure 1).[1]

Thus we have used Tukey's approach [9] to model approximations of moments from quartiles $Q_{1/4}, Q_{1/2}, Q_{3/4}$ by rules

$$m = \frac{Q_{1/4}+Q_{3/4}}{2}, \qquad s = \frac{Q_{3/4}-Q_{1/4}}{2},$$

$$g = \frac{Q_{1/4}+Q_{3/4}}{2} - Q_{1/2}.$$

Then m, s^2, gs^2 are considered as replacements of the three moments. Notice that they have correct "physical dimensions" to fulfill such expectations.[2] The actual values of the mean, variance, or skewness may, of cause, differ from these guesses and there are many

[1] More precisely, the quantiles are complementary in the sense that a counterpart for Q_p in the orthogonal variable is Q_{1-p}.

[2] However, s and g should be understood just as measures of the standard deviation or skewness but not misplaced with them. For example, the ratio between s (half of the so-called interquartile range) and standard deviation σ of the normal distribution is about 0.6745. We do not need to use these conversion constants because we start and end with quantile parametrization, and the constants are cancellated.

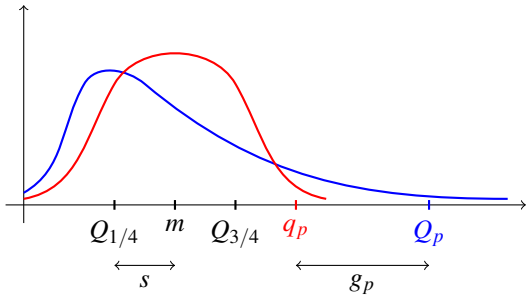

Figure 2: The reference distribution (red) shares quartiles $Q_{1/4}, Q_{3/4}$ with the modeled distribution (blue). Parameter g_p is the difference between the reference and the modeled p-quantiles.

works improving the formulas or methods of calculation from quartiles [1, 5, 8, 10] . Yet this is not in contradicion with the assuption that m, s^2, gs^2 behave linearly.

This paper adjusts the ideas mentioned above when an arbitrary quantile replaces the median (or eventually another of the quartiles). The reason for this is that we can directly estimate whether a production plan is realizable for reliability represented by the quantile. As we shall see, such change can provide more precise results, especially when the desired reliability is close to bounds 0 or 1.

2 Derivation

In our model, the median $Q_{1/2}$ participates only in the expression for g. The parameter g expresses the difference between the approximative mean and the median, and we can interpret it as a measure of skewness. Nevertheless, the median coincides with the mean in any symmetric distribution. Hence, we can solve g as a difference of means of the modeled distribution and its "embedded" symmetric version with the same values of m and s.

Now we can extend this idea for arbitrary quantile Q_p. We assume that there is some suitable symmetric distribution $S(m,s)$, called a *reference distribution* (see Figure 2). We denote the quantile for probability p in $S(m,s)$ by q_p, and we put

$$g_p = q_p - Q_p.$$

The above situation may resemble the Elo ranking system used in chess competitions [2]. There is a symmetric distribution defined on possible differences of players' rankings, and we can predict a result of a game as a value of cumulative distribution function. The difference between an actual result and this expected

result serves for an update of the players' rankings. In the past, the normal distribution was used as the reference distribution, but recently it was replaced by logistic distribution. Anyway, both distributions provide outstanding results because there are a few differences between values calculated for more games at once and summed values calculated separately for each game played by a player. Thus the differences between the actual and expected results preserve the additivity property. (The variance is constant here, so we need not weigh by it as we do for the third central moment.)

So, we suggest the following rules for the summation:

R1 Let X be a random variable with the 1st and 3rd quartiles $Q_{1/4}, Q_{3/4}$ and a p-quantile Q_p. We put

$$m = \frac{Q_{1/4} + Q_{3/4}}{2}, \qquad s = \frac{Q_{3/4} - Q_{1/4}}{2}.$$

Let $S(m,s)$ be a reference distribution for m,s with p-quantile q_p. We put

$$g_p = q_p - Q_p.$$

R2 Let X,Y be two independent random variables parametrized by triples (m_1,s_1,g_{p1}) and (m_2,s_2,g_{p2}), respectively. Then the sum $X+Y$ is parametrized by triple (m,s,g_p) where

$$m = m_1 + m_2, \qquad s^2 = s_1^2 + s_2^2,$$

$$g_p s^2 = g_{p1} s_1^2 + g_{p2} s_2^2.$$

Or explicitly,

$$s = \sqrt{s_1^2 + s_2^2}, \qquad g_p = \frac{g_{p1} s_1^2 + g_{p2} s_2^2}{s_1^2 + s_2^2}.$$

R3 Let $X_i, i = 1,\ldots,n$ be independent copies of the same random variable parametrized by triple (m_1,s_1,g_{p1}). Then the sum $\sum_{i=1}^n X_i$ is parametrized by triple (m,s,g_p) where

$$m = nm_1, \qquad s^2 = ns_1^2, \qquad g_p = g_{p1}.$$

Or explicitly,

$$s = \sqrt{n} s_1.$$

R4 Let X be a random variable parametrized by triple (m,s,g_p). Then we put

$$Q_{1/4} = m - s, \quad Q_{3/4} = m + s, \quad Q_p = q_p - g_p.$$

R5 Let (n,t) be a state that mass n is produced (recorded, stored, etc.) at time t. Let T be a random variable expressing time for producing mass n and N be a random variable expressing mass produced at time t. Then t is a p-quantile for T iff n is a $(1-p)$-quantile for N.[3]

Example. Let P_1, P_2 be two independent parallel production processes (machines). Assume that P_1 operates each product in time T_1 and P_2 in time T_2. Quantiles of both random variables T_1, T_2 are known. A plan states that P_1 works n_1 products and P_2 works n_2 products. We ask if the battery of processes finishes working of the $n = n_1 + n_2$ products before time t with reliability at least 95%. The solution consists of the following steps:

- Get quantiles $Q^{T_i}_{1/4}, Q^{T_i}_{3/4}, Q^{T_i}_{0.95}$ of $T_i, i = 1,2$ and convert them to moment measures by R1.
- Calculate sums of n_1 copies of T_1 and n_2 copies of T_2 by R3.
- Convert moment measures to quantiles by R4.
- Change coordinates by R5 at time t to mass random variables N_1, N_2, i. e. we get quantiles $Q^{N_i}_{1/4}, Q^{N_i}_{3/4}, Q^{N_i}_{0.05}$ for $N_i, i = 1,2$.
- Convert the quantiles of N_1, N_2 to moment measures by R1.
- Sum $N = N_1 + N_2$ by R2.
- Convert moment measures to quantiles by R4.
- Change coordinates by R5 for mass n. We get again quantiles $Q^T_{1/4}, Q^T_{3/4}, Q^T_{0.95}$ for time random variable T.
- If $Q^T_{0.95} \le t$, the plan is confirmed.

3 Experiments

We have performed several tests to verify the method. Selected simulations are displayed in Figures 3–6. Some selected natural examples of distributions X and Y were generated, each comprising 1000 observations. The quantile function for the sum $X+Y$ (empirical, blue) is compared with the one obtained from our method (estimated, black). For all of the experiments, the logistic and the normal distribution were used as the reference distributions and compared.

Example. Let us also demonstrate the aforementioned example of two parallel proceses with concrete parameters. Let the process P_1 work one product in time $T_1 \sim 1 + 2\mathrm{B}(2,4)$ and process P_2 work one product in

[3] The complementarity of probabilities is explained by Figure 1.

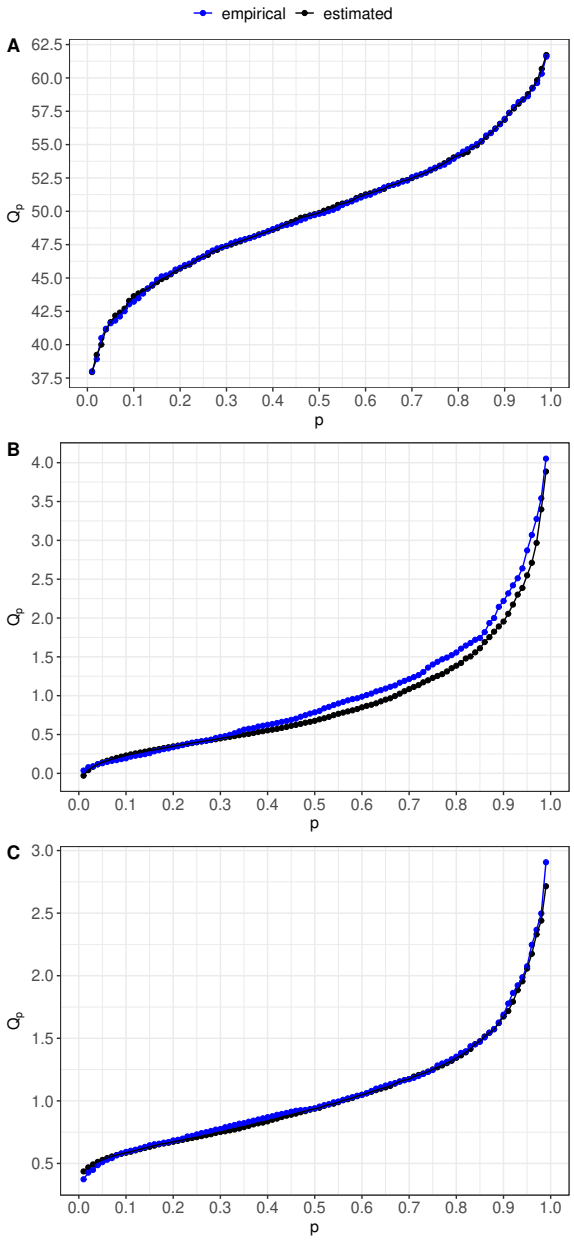

Figure 3: Reference distribution: logistic, (A) X: normal (mean=40, sd=5), Y: normal (mean=10, sd=1), (B) X: gamma (shape=0.5, rate=1), Y: gamma (shape=1, rate=2), (C) X: beta (shape1=6, shape2=4), Y: gamma (shape=1, rate=2).

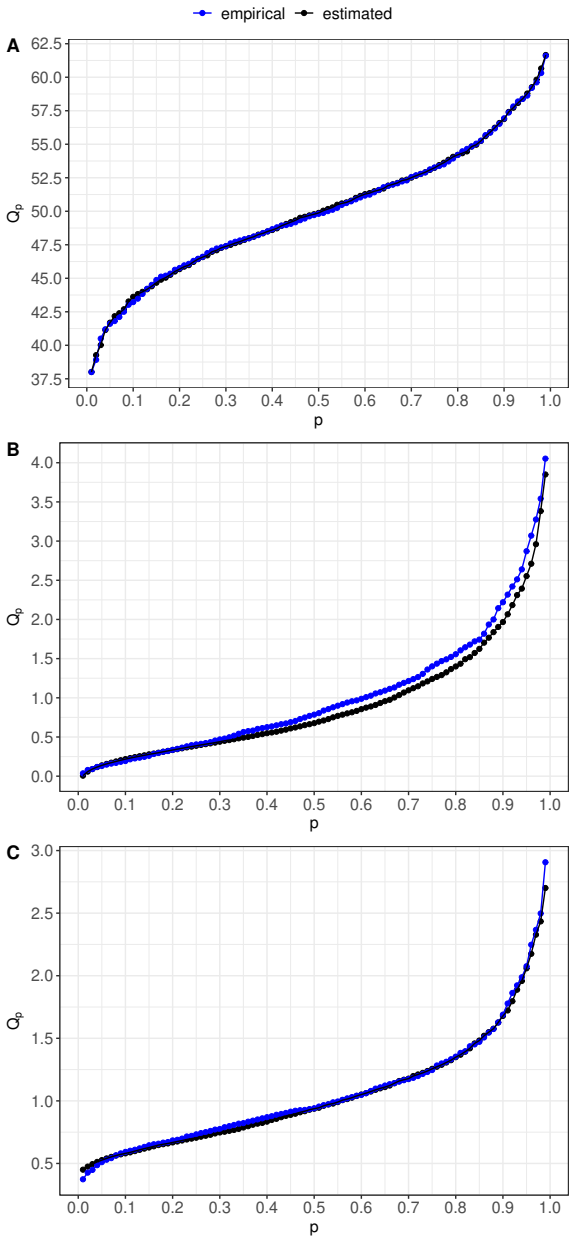

Figure 4: Reference distribution: normal, (A) X: normal (mean=40, sd=5), Y: normal (mean=10, sd=1), (B) X: gamma (shape=0.5, rate=1), Y: gamma (shape=1, rate=2), (C) X: beta (shape1=6, shape2=4), Y: gamma (shape=1, rate=2).

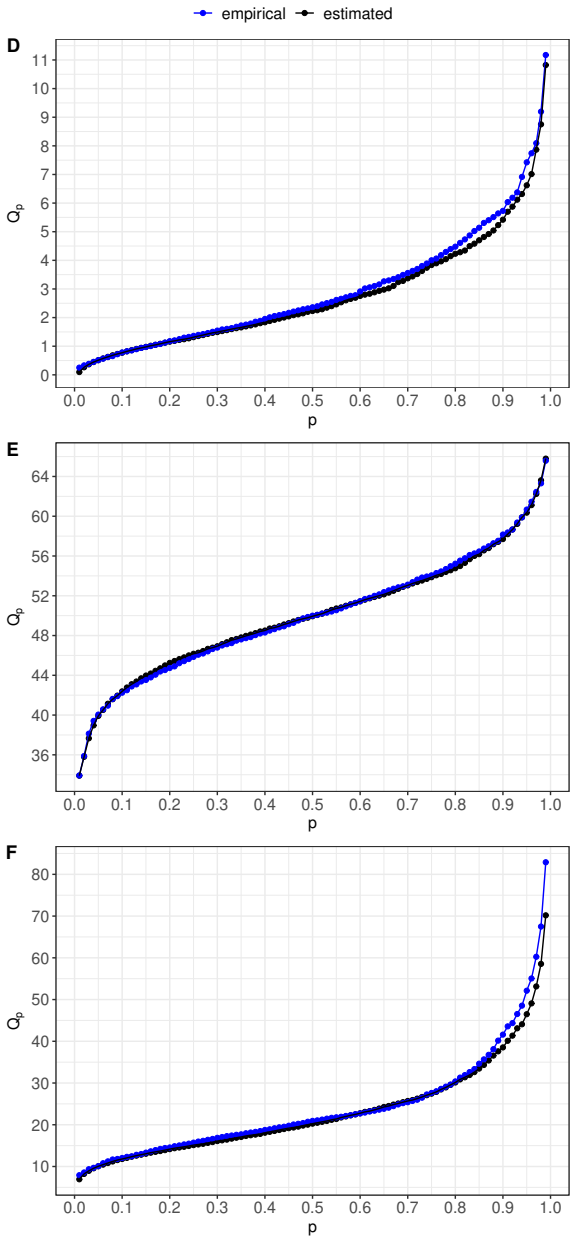

Figure 5: Reference distribution: logistic, (D) X: exponential (rate=1), Y: Weibull (shape=1, scale=2), (E) X: logistic (location=10, scale=2), Y: logistic (location=40, scale=3), (F) X: uniform (min=5, max=20), Y: lognormal (meanlog=2, sdlog=1).

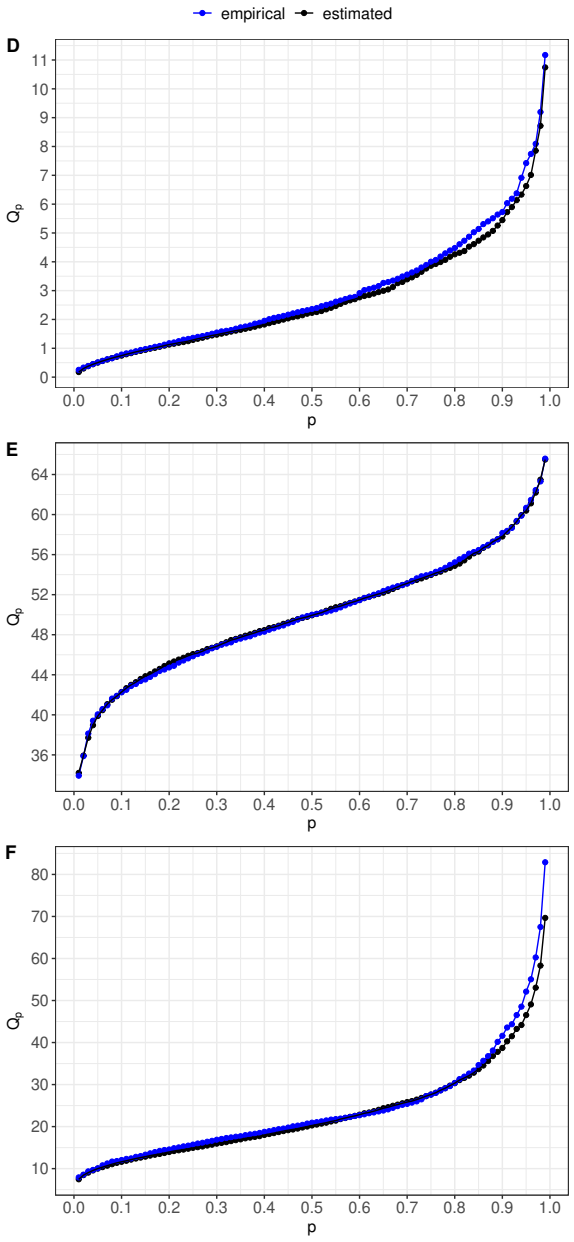

Figure 6: Reference distribution: normal, (D) X: exponential (rate=1), Y: Weibull (shape=1, scale=2), (E) X: logistic (location=10, scale=2), Y: logistic (location=40, scale=3), (F) X: uniform (min=5, max=20), Y: lognormal (meanlog=2, sdlog=1).

time $T_2 \sim 2+3\mathrm{B}(1.5,3)$ where $\mathrm{B}(\alpha,\beta)$ is a beta distribution.

The means of the variables are then $E(T_1) = 1+2\cdot\frac{2}{2+4} = \frac{5}{3}$ and $E(T_2) = 2+3\cdot\frac{1.5}{1.5+3} = 3$. Thus, in time $t = 10$, the processes "typically work" $6+3=9$ or $5+3=8$ products together but there is a quite large variance. Assume that (for a sake of planning) we are interested whether producting of at least $n = 8$ products in time $t = 10$ has a reliability at least 90%. Since there is no specification how the processes should contribute, we can skip the final conversion from mass to time and just check whether we produced enough at $t = 10$.

The distributions provide quantiles[4]

$$Q^{T_1}_{1/4} = 1.39, \quad Q^{T_1}_{3/4} = 1.91, \quad Q^{T_1}_{0.9} = 2.17,$$
$$Q^{T_2}_{1/4} = 2.52, \quad Q^{T_2}_{3/4} = 3.41, \quad Q^{T_2}_{0.9} = 3.87.$$

If we use the logistic distribution as the reference distribution, we expect

$$q_p = m + s\frac{\ln(p/(1-p))}{\ln 3}$$

which yields quantile guesses

$$q_{0.9} = m+2s, \qquad q_{0.1} = m-2s.$$

and the moment measures

$$m^{T_1} = 1.65, \quad s^{T_1} = 0.26, \quad g^{T_1}_{0.9} = -0.01,$$
$$m^{T_2} = 2.97, \quad s^{T_2} = 0.46, \quad g^{T_2}_{0.9} = -0.01.$$

According to R3, in repeated processes the quantiles develop as

$$Q_{1/4}(N) = Nm - \sqrt{N}s,$$
$$Q_{3/4}(N) = Nm + \sqrt{N}s,$$
$$Q_{0.9}(N) = Nm + 2\sqrt{N}s + g_{0.9}.$$

In our case, we substitute $t = 10$ for the quantiles and solve the resulting quadratic equations in $\sqrt{N}$ to obtain N (R5). These Ns represent mass of worked products in time $t = 10$ for the given quantile. We obtain quantity random variables N_1, N_2 at $t = 10$ with quantiles

$$Q^{N_1}_{1/4} = 5.69, \quad Q^{N_1}_{3/4} = 6.47, \quad Q^{N_1}_{0.1} = 5.34,$$
$$Q^{N_2}_{1/4} = 3.10, \quad Q^{N_2}_{3/4} = 3.66, \quad Q^{N_2}_{0.1} = 2.85,$$

and moment measures

$$m^{N_1} = 6.08, \quad s^{N_1} = 0.39, \quad g^{N_1}_{0.1} = 0.04,$$
$$m^{N_2} = 3.38, \quad s^{N_2} = 0.28, \quad g^{N_2}_{0.1} = 0.04.$$

[4]The values were obtained from generated dataset, thus they can slightly differ from exact values.

The sum N of variables N_1, N_2 can be calculated by R2 providing moment measures

$$m^N = 9.47, \qquad s^N = 0.48, \qquad g^N_{0.1} = 0.04,$$

and we finish with guess

$$Q^N_{0.1} = m^N - 2s^N + g^N_{0.1} = 8.54.$$

Since $Q^N_{0.1} > 8$, the model confirms the plan as reliable.

4 Discussion and conclusion

The choice of reference distribution between normal and logistic has a marginal effect in all of the performed experiments. Likewise, in other today's applications, the logistic distribution can be preferred by many users for a simpler analytic description and a more straightforward generalization.

The presented method provides perfect results if the summed distributions X, Y are normal, logistic, or perhaps just symmetric, which is not surprising. However, we also consider the other experiments entirely satisfactory, namely for the quantile estimations on the tales ($p \in (0, 0.2)$ or $p \in (0.8, 1)$) which are essential for reliability estimations in practice. We improved our previous method from [3] significantly.

Anyway, there are still many ways for further development. It is possible to store and calculate values for more quantiles Q_p than just one, and these can participate in a more precise estimation of the mean and the variance. For example, a better mean estimation $\frac{Q_{1/4}+2Q_{1/2}+Q_{3/4}}{4}$ puts more weight on the median [5]. In such a case, we could work with four quantiles $Q_{1/4}, Q_{1/2}, Q_{3/4}, Q_p$ and four moment measures $m, s, g_{1/2}, g_p$ and adjust the conversion formulas.

The example with two parallel processes also displayed another issue of the summation. Since we assumed and calculated the development of production as a continuous process, we have obtained more optimistic results (and a resolute confirmation of the plan) because we also summed "partially worked" products, e.g. when P_1 finishes six products and almost the 7th product and P_2 finishes two products and almost the 3rd product then continuous summation can assume that 9 products are finished but, in reality, it is only 8. In such situations, especially when the number of products is small or the number of processes in the battery is large, the error can be large and one should consider to convert resulting continuous distributions to discrete ones and combine the summation methods with usual addition (convolution) of discrete distributions.

The skew measure g^T_p can be small with respect to other measures and its role even marginalize with large

repetitions of a process. This can be the case of many applications where we can ignore the skewness and work just with the mean and the variance (what is a common technical practice). However, mass variable N need not be symmetric even that time variable T is. In some cases (e. g. large variance of T and low number of repetitions) the skewness of N and g_p^N can be important.

When we applied the Monte Carlo simulation on the example with 1000 runs in each sample, most of samples also approved the reliability at least 90% but there was a few of runs which did not reach it. In contrast, our method behaves deterministically and provides stable answers (even that they can be sometimes wrong).

Acknowledgement

Research of the second (corresponding), third and fourth authors was supported by the project Practical use of big data for intelligent production flow decision system financed by the Technology Agency of the Czech Republic: project TAČR FW03010296. Research of the first author was supported by the project Mathematical structures 10 (MUNI/A/1160/2020).

References

[1] M. Bland, Estimating mean and standard deviation from the sample size, three quartiles, minimum, and maximum, International Journal of Statistics in Medical Research 4 (1) (2015) 57–64.

[2] A. Elo, The Rating of Chessplayers, Past and Present, NY: Arco Publishing, 1978.

[3] K. Emir, D. Kruml, J. Paseka, I. Selingerová, A simplified probabilistic validation of production flows, in: Proceedings of the 18th International Conference on Modeling and Applied Simulation, 18-20 September 2019, Lisbon (Portugal), 2019, pp. 166–173.

[4] L. Fiondella, Y. Lin, P. Chang, System performance and reliability modeling of a stochastic-flow production network: A confidence-based approach, IEEE Transactions on Systems, Man, and Cybernetics: Systems 45 (11) (2015) 1437–1447.

[5] S. P. Hozo, B. Djulbegovic, I. Hozo, Estimating the mean and variance from the median, range, and the size of a sample, BMC Medical Research Methodology 5 (1) (2005) 13–22.

[6] M. Manitz, Queueing-model based analysis of assembly lines with finite buffers and general service times, Comput. Oper. Res. 35 (8) (2008) 2520–2536.

[7] A. Rényi, Foundations of probability. Reprint of 1970 edition, reprint of 1970 edition Edition, Mineola, NY: Dover Publications, 2007.

[8] J. Shi, D. Luo, X. Wan, Y. Liu, J. Liu, Z. Bian, T. Tong, Detecting the skewness of data from the sample size and the five-number summary, preprint.
URL `https://arxiv.org/abs/2010.05749`

[9] J. W. Tukey, Exploratory data analysis, Reading, Massachusetts: Addison-Wesley Publishing Company, 506 p. (1977).

[10] X. Wan, W. Wang, J. Liu, T. Tong, Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range, BMC Medical Research Methodology 14 (1) (2014) 135–147.