B-RISK IN PROFICIENCY TESTING IN RELATION TO THE NUMBER OF PARTICIPANTS

Louis-Jean Hollebecq[1]
[1] CompaLab, Rosny-sous-Bois, France, ljh@compalab.org

Abstract:
The Monte Carlo method was applied to PT schemes to investigate their efficiency. Probabilities that the computed z values are over 3 while the true value is less than 2 and that the computed z values are less than 2 while the true values are over 3 are computed for a series of situations: number of participants from 5 to 30, various ratios of repeatability over reproducibility and number of test results per participant, introduction or not of outliers with k from 3 to 10. For each situation, the probabilities of detecting true outliers and to trigger false alerts are discussed.

Keywords: Laboratory proficiency testing, Monte Carlo methods, efficiency of assessment.

1. INTRODUCTION

Proficiency tests (PT) are widely used to assess the performance of laboratories. Participating to such programs is required by ISO 17025 [1], which is the standard of reference for accreditation of laboratories. Reference standards for interlaboratory comparisons (ILC), ISO 5725-2 [2], ISO 13528 [3] and ISO 17043 [4], define limits for computing the alerts, corresponding to theoretical risks of 5% and 0,3%.

Note: [2] deals with ILC to assess test methods. [3] deals with ILC for PT of labs. [4] is the reference for accreditation of PT providers. Limits in [2] are intended to assure the reliability of the assessment. Limits in [3] and [4] are for proficiency checking of participants. [2] is referred here because it is the "historical one" and it is still widely used by PT providers, even if [3] is obviously better adapted to proficiency testing of labs.

These risks are of α-type (risk to trigger a warning that should not). Another risk actually occurs, usually called β-type (risk not to trigger a warning when it should). However, even if this question is of main importance, this β-type risk is quite hard to compute, and for this reason, is almost always just ignored, including in the reference standards [2] and [3]. Everybody knows that an enough number of participants is necessary to ensure the efficiency of the PT, but there is no clear consensus of what should be "an enough number". On the other hand, test methods for which there are very few potential participants to a PT are quite numerous. There is then no opportunity for them to get the advantages of a participation to a PT. This paper proposes to overcome the difficulty of computing the β-risk by using the Monte-Carlo method and to provide a beginning of answer to the question: does it make sense or not to organise PTs with 5 or 8 or 12 participants, especially when the number of potential participants is quite low?

To do so, the following issues are dealt with:

1. How α-type and β-type risks can be computed and what hypotheses to do it were taken into account in the present study;
2. What are the principles of the Monte-Carlo method, in which conditions it can be used and how it was implemented in the present study;
3. What is the impact of the use of robust statistics that are usually used to avoid the deleterious impact of outlying results on the so-called assigned values;
4. What is the impact of the number of test results by each participant, with regard to interlaboratory and repeatability standard deviations.

2. DESIGN OF EXPERIMENTS

2.1. Calculation of α-type and β-type risks

Computing α-type and β-type risks requests to define underlying alternate hypotheses usually designated as H_0 and H_1. α is the probability to reject the H_0 hypothesis while it is actually true and β is the probability to reject the H_1 hypothesis while it is actually true, as shown in Table 1.

Table 1: α and β-risks with regard to H_0 and H_1 hypotheses

	H_0 is true	**H_1 is true**
H_0 is accepted	Right decision (p=1-α)	Wrong decision (p=β)
H_1 is accepted	Wrong decision (p=α)	Right decision (p=1-β)

The issue of α-type and β-type risks have been extensively discussed for a very long time because they address many practical decision problems, notably the assessment of conformity of products to specifications, see for example ISO 2859-1. In all cases:

1. α and β-risks decrease when the available number of test results increases;
2. For a given number of test results, α-risk increases when β-risk decreases, and vice-versa.

In the context of PT organisation, the H_0 hypothesis can be quite obviously defined as "The results of the participant belong to the general population of expected results". In the same way, H_1 can be defined as "The results of the participant belong to a [5] other than the one of the expected results".

It is needed then to define how conclusions about H_0 and H_1 shall be carried out. The decision rules described in the reference standards [2] and [3], i.e. the calculations of z-scores obviously apply to H_0. On the contrary, the distribution of H_1 is not known (other populations of results than the expected one can practically be very

different ones, including gross errors, different types of deviations to the method, etc. …). One way to solve this problem is to construct "power curves" in function of parameters of the problem, and especially the number of results and the distance to H_0. This principle was used to build up the design of experiments for this study.

In details:

1. We considered that the α-risk occurs when $|z_{calc}|>3$ and $|z_{true}|<2$ (as recommended in [2] and [3]), and that the β-risk occurs when $|z_{calc}|<2$ and $|z_{true}|>3$;

2. We computed these α and β-risks on populations of test results without any true outlier, i.e. to a whole Gaussian population of expected results. This implicitly includes 5% of corresponding z-scores outside the [-2;+2] interval and 0,3% outside the [-3;+3] interval;

3. We also computed the α and β-risks on populations of test results including one true outlier with various z values from 3,5 to 10. These computations of α and β-risks were carried out separately for the main population and for the outlier, enabling to check the impact of the outlier on both categories of participant results.

It should be kept in mind that the computed β-risks fully depend on the definition of H_1 (see here upper) and that other ways to define H_1 would also make sense, leading to other meaningful values of β.

To deal with the upper, we have built up a design of experiments pursuing the following goals:

2.2. Impact of the number of participants

[1] recommends that at least 12 participants are present and [2] recommends not to use robust statistics when the number of participants is less than 18. On the other hand, our computations showed that α and β-risks do not significantly change when the number of participants goes over 30. We then did not investigate higher values and decided to compute the α and β-risks for a number of participants varying from 5 to 30. This enabled us to investigate areas that are not recommended by the standards.

2.3. Principles of the Monte-Carlo method

The Monte-Carlo methods are a large category of algorithms that use random numerical realisations of a given model. They are often used to solve mathematical or physical problems, difficult or impossible to solve by other methods. For a survey of the history and applications of the Monte-Carlo methods, see for example [6].

In our problem, using the Monte-Carlo methods enables us to create series of "true values" of test results that cannot be known in real life. In practice, we always know whether H_0 and H_1 are accepted or not (i.e. whether an alert was sent to the participant), but we can never know whether H_0 and H_1 are actually true or not. Using Monte-Carlo methods enables us to control at the same time for each series of random results whether H_0 and H_1 are accepted or not and whether H_0 and H_1 are true or not. Having this whole information is necessary to compute both α and β-risks.

However, using Monte-Carlo methods requests to use a model that reasonably fits the situations encountered in the real world. In this study, we used the model of ISO 5725-1 [7] widely used to cope with problems of precision of test results:

$$y = m + B + e\ . \qquad (1)$$

where "m" is the general mean value, "B" is the bias of the lab and/or the method, and "e" is the random error.

In this model, we used m=0, a Gaussian distribution with 0 as mean value and 1 as standard deviation for B and another Gaussian distribution with 0 as mean value and a varying s_r as standard deviation for "e" (see at 2.5 how s_r was chosen).

Using the Monte-Carlo methods also requests to use random inputs. Moreover, when correlations between them apply in real life, these correlations must be incorporated in the inputs of the computations. That can be a bit difficult to do properly. In our case, we can reasonably rule it out, assuming that there is no correlation between the results of the different participants and between the results of a same participant. As a matter of fact [3] requests PT providers to care about it (no collusion between participants), because it is a condition to ensure the validity of the statistical treatment.

To assure the validity of the conclusions, the random series need to numerous enough, depending on many factors. In our study, we computed series of 500000 to 4000000 z-scores for each situation (i.e. for each combination of number of participants, number of test results per participant and s_r/s_L ratio) in 40 groups of each. This grouping enabled us to check how repeatable were the computed α and β within the 40 groups and compute a related interval of confidence (IC). This IC always happened to be less than ±2% (with enlarging coefficient k=2) and in all cases significantly lower than of the computed α and β.

2.4. Impact of the type of statistics used to compute the so-called assigned values

Results with gross errors often occur during the organisation of PT. They are usually caused by typing errors, by misunderstanding of instructions for participation or by using wrong units. In most cases, gross errors are due to necessary deviations to routine procedures of the labs when they participate to a PT. Typically, typing errors usually never occur in real life because data transfer is nowadays never performed manually, contrarily to the cases of participations to PT.

However, gross errors are a big problem for the statistical data processing, because they strongly impact the estimation of statistical parameters. In particular, they strongly increase the computed standard deviations. Hence they strongly increase the β-risk. On the other hand, just ignoring the suspicious results might lead to underestimate the standard deviations of reference, increasing then the α-risk.

To face this problem, [2] and [3] recommend to detect outliers and/or use so-called robust statistics. These robust statistics generally consist in replacing outlying results by softened virtual ones, using algorithms specifically designed for that. Full information about this can be found in particular [3] Annex C and [8]. These robust methods tend to produce mean values and standard deviations resisting to a certain proportion of outliers (called breaking point) but also to decrease the speed of convergence of the estimates towards their central values. [3] Annex D provides a comparison of the breaking

points and speeds of convergence of the different algorithms it proposes.

Because of the decrease of their speed of convergence [3] and [9] recommend not to use robust statistics for a low number of participants. However, [10] and [11] went quite deeper in studying the issue and both showed that using robust statistics considerably improve the estimation of central value and scatter of the distribution in presence of outliers and consequently improve the assessment of the performance of PT with low number of participants. They both compared the different robust methods usually used but they could not make a definitive conclusion that would be valid for all types and proportions of outlying results.

On our own, as PT provider, our line of action has always been to use robust statistics, even for low number of participants, preferring running the risk of a day-to-day slightly lower efficiency of assessment than a risk of completely misleading one, even sporadically.

For the sake of this study (which is not to compare the efficiency of the different available robust methods), we chose to compute α and β-risks without robust statistics and with the so-called A algorithm described in [2] and [3], which is the most widely used by PT providers. This enables to check the impact of using robust statistics or not without increasing to much the number of necessary calculations.

In order to check the impact of outliers, we produced series of test results without outliers and with one outlier which true z-score varies from z=3,5 to z=10. It follows that the proportion of outliers depends on the number of participants p, from 20% for p=5 to 3,3% for p=30. This option does not necessarily represent faithfully what happens in practice (see [10] and [11] for that), but 1. it does not request any modelling of outlying and 2. it provides information about the impact of outliers easier to handle.

2.5. Impact of the number of repetitions by each participant, with regard to interlaboratory and repeatability standard deviations

In almost all cases, PT providers use z-scores or equivalents to assess the performance of the participants. According to [3] and [4], z-scores can be computed according to the equation (2):

$$z = \frac{x_i - X_{pt}}{\sigma_{pt}} \quad (2)$$

where x_i is the result of the participant "i",
X_{pt} is the central value
and σ_{pt} is the standard deviation assigned for the PT.

The performance is regarded as satisfactory when $z \in [-2; +2]$ and not satisfactory when $z \notin]-3; +3[$.

These limits implicitly refer to the idea that the probabilities for these events to occur are respectively 95% and 0,3%. Consequently, the theoretical α-risk is 0,3%. In other words, the probability to decide that the results are unsatisfactory while in fact they do belong to the main population is 0,3%.

In fact, this would be true if σ_{pt} had exactly represented σ_{BL}, standard deviation of the biases of all the participating laboratories, what is never true. In most cases, σ_{pt} is computed as s (or s* when a robust algorithm is used), defined in [2] and [3] as the standard deviation of the results of all participants. Then, in practice, σ_{pt} can be computed with the equation (3):

$$\sigma_{pt}^2 = \sigma_{BL}^2 + \sigma_{iL}^2 + \frac{\sigma_r^2}{N_r} + \frac{\sigma_H^2}{N_s} \quad (3)$$

where σ_{BL} is the standard deviation of the biases of the participating laboratories,

σ_{iL} is the standard deviation due to internal scatter of the laboratory results other than repeatability (differences between operators, machines of the lab, variations of environmental conditions within the lab along the time),

σ_r is the repeatability standard deviation,

N_r is the number of test results per lab,

σ_H is the standard deviation representing the homogeneity of samples

and N_s is the number of samples provided to each lab.

In other words, the test results that a given lab sends to the PT provider are not only governed by their bias, but also by which combination of equipment – operator – testing conditions that are used to perform the tests for PT, by the repeatability of tests and by chance with regard to inhomogeneities of samples.

In any cases, σ_{pt} is then always greater than σ_{BL}, what leads to α-risk lower than the expected 0,3%, but also and consequently to increased β-risk.

In some cases, for example when $\sigma_r >> \sigma_{BL}$ and only one test result is sent by each lab, the PT can become completely inefficient (see 3.3 here after).

In practice:

1. σ_{iL} usually cannot be computed, because each lab provides results obtained by only one operator, one test equipment set, performed in a short period of time. Consequently, when each lab provides several test results, their standard deviation is s_r and does not include any contribution of σ_{iL};

2. In most cases, labs are requested to perform a few tests on a same sample or one test on each of a few samples. In these conditions, σ_r and σ_H cannot be computed separately.

Consequently, in most cases only two standard deviations are governing the assessment:

1. An interlaboratory standard deviation that we call σ_L in our study, and that includes σ_{BL}, σ_{iL} and, when only one sample is provided, σ_H;

2. A repeatability standard deviation that we call σ_r in our study, and that includes σ_r and σ_H when several samples are provided and one test per sample is performed.

When only one test result from only one sample is provided per each participant (what happens in fact quite often), σ_{pt} is then the reproducibility standard deviation σ_R.

By the way, we see here that PT providers could strongly improve their scheme and use ANOVA to separate all these standard deviations, but this goes far beyond the scope of this study and is not dealt with here.

To come back to our study, we computed α and β-risks for σ_r/σ_L from 0,1 to 3 (corresponding to σ_r/σ_R from 0,1 to 0,95 that encompass the ratios actually encountered in practice) and for N_r (number of test results per lab) from 1 to up to 48. This so high number of repetitions is

in practice never encountered. However, including it in our scheme made possible to investigate whether there could be of some benefit in some cases.

3. RESULTS AND DISCUSSIONS

3.1. Pertinence of a ratio relating repeatability, interlaboratory standard deviation and number of test results per participant

To deal with the issue exposed at chapter 2.5, we defined a parameter λ defined as follows:

$$\lambda = \frac{\sigma_r}{\sigma_L \times \sqrt{N_r}} \quad (4)$$

where σ_r, is repeatability,
σ_L is interlaboratory standard deviation,
and N_r is the number of test results per lab.

This parameter reflects the idea that each participant's test results are distributed on a Gaussian distribution with its bias as mean value and $\sigma_r/\sqrt{N_r}$ as standard deviation.

We found out that this parameter is valid to describe the full effect described in chapter 2.5, see Figure 1 here after.

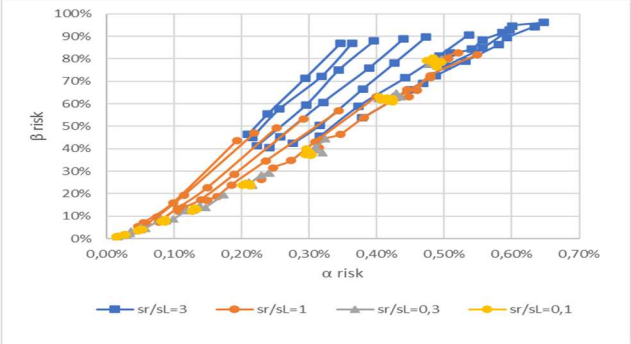

Figure 1: α and β-risks for participants without outlier in function of λ and number of participants (from left to right, p = 5 – 6 – 8 – 10 – 13 – 16 – 20 – 25 – 30)

The figure clearly shows that, for each number of participants, the s_r/s_L curves are in extension of each other, so that a merge of these curves make sense, as shown in Figure 2 here after.

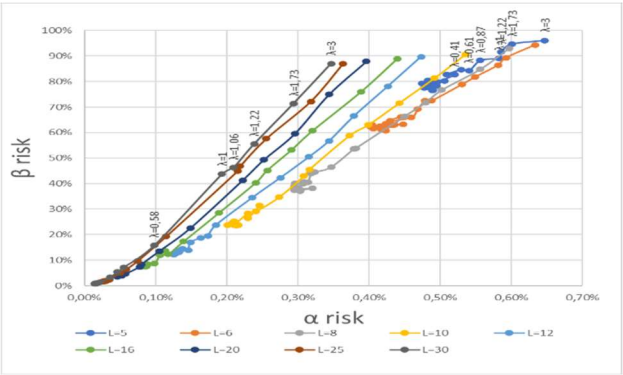

Figure 2: α and β-risks for participants without outlier in function of the number of participants

3.2. Impact of the use of robust statistics

[2] and [3] recommend not to use robust statistics when the number of participants is low. However, this recommendation was made with consideration to the loss of efficiency when doing so. However, [10] and [11] did not confirm that robust statistics should not be used with little number of participants.

In fact, our computations confirmed that:

1. The α-risk is slightly increased when using robust statistics, what is consistent with the related loss of efficiency;
2. The β-risk is significantly reduced when using robust statistics, what is consistent with the better robustness of the assigned values.

In details, three cases were considered:

1. Comparison of risks for participants when no outlier is present;
2. Comparison of risks for not outlying participants when one outlier is present;
3. Comparison of risks β for a outlier (by definition, in that case, the α-risk does not exist).

In the first case, illustrated in Figure 3, we observed that α-risk slightly increases while the β-risk slightly decreases. However, both evolutions are not significant compared to the impact of the other factors (λ and N_p).

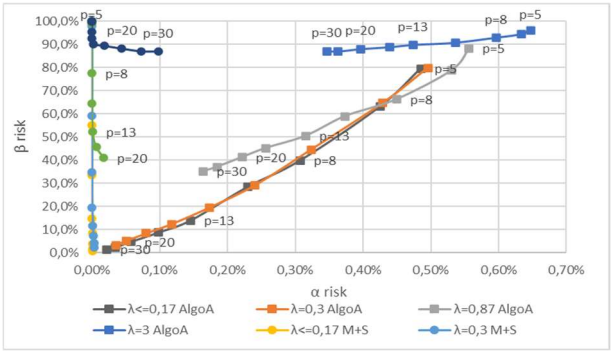

Figure 3: Comparison of α and β-risks obtained with Algorithm A and without robust statistics (m and s), for participants without any outlier in function of λ

In the second case, illustrated in Figure 4, we observed that α-risk slightly increases while the β-risk significant decreases when the λ factor is adverse (i.e. when λ>1). In particular, we can see that even with 30 participants and λ>1, statistics not robust completely fail to detect participants with z>3.

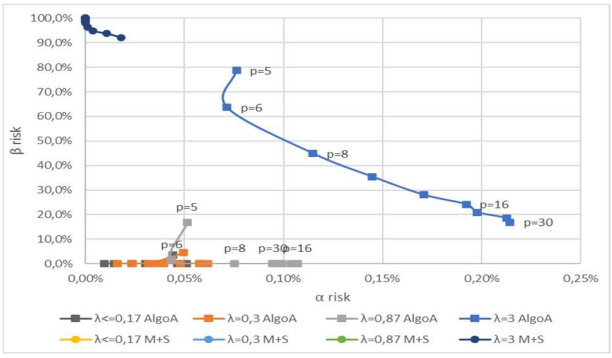

Figure 4: Comparison of α and β-risks obtained with Algorithm A and with statistics not robust (m and s), for main participants when an outlier with z=10 is present, in function of λ

In the third case, illustrated in Figure 5, we observed that AlgoA is significantly more efficient to detect outliers when PT conditions are adverse (i.e. when λ>1 and/or when N_p<13).

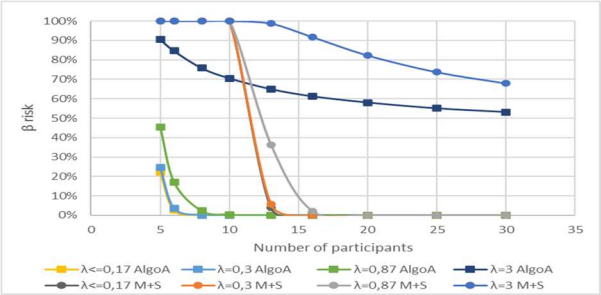

Figure 5: β-risks obtained with Algorithm A and with not robust statistics, for an outlier with z=10, in function of λ

3.3. Impact of λ ratio

Figure 2 clearly shows that both α and β-risks decrease with λ until a certain value of λ that we evaluated to be 0,17, whatever the number of participants.

This also occurs for other cases (i.e. when an outlier is present) as shown in Table 2.

Table 2: Lower λ limits under which α and β-risks decrease anymore according to the number of participants (α and β in %, computed with Algo A)

		No outlier		**Main participant with one outlier**		**Outlier**	
N_p	**λ**	**α**	**β**	**α**	**β**	**α**	**β**
5	0,17	0,5	80	0,55	90	-	22
6	0,17	0,45	65	0,53	80	-	2
8	0,17	0,30	40	0,44	65	-	0
10	0,17	0,2	23	0,38	59	-	0
13	0,17	0,12	12	0,32	50	-	0
16	0,17	0,10	10	0,25	45	-	0
20	0,17	0,05	5	0,22	40	-	0
25	0,17	0,03	3	0,18	38	-	0
30	0,17	0,01	1	0,16	34	-	0

PT providers do not control σ_r and σ_L, which depend on the test method, but they control N_r (the number of test results per lab). Hence, they control λ (increasing N_r decreases λ, see Equation 4). They should use their historical data or literature to determine, s_r/s_R for each test method proposed for PT and use Table 3 to determine the minimum N_r values to optimize the PT programs. However, practical reasons may limit N_r (costs or impossibilities of production or transportation of the samples, or for laboratories to perform the tests).

Table 3: Optimal number of repetitions for PTs, according to the s_r/s_L and s_r/s_R ratios.

s_r/s_L	**s_r/s_R**	**N_r**
≤0,17	≤0,17	1
0,3	0,29	3
0,42	0,39	6
0,59	0,51	12
1	0,71	35
3	0,95	310

As a conclusion, when N_r is chosen equal or superior than the value of Table 3, the best α and β-risks of Table 2 can be reached, according to the number of participants.

Further experiments are requested to understand the undergrounds of this λ=0,17 constant. In particular, it should be studied how it varies with the definitions of H_0 and H_1 (see 2.1).

3.4. Discussion about α-risks

Theoretical α-risk with our definition of H_0 is 0,0027/0,95 = 0,28% (probability that $|z_{calc}|>3$ while $|z_{true}|<2$). This risk is reduced by the impact of the repeatability, especially when the λ value is high (see 2.5). When the PT conditions are bad (i.e. λ>1 and/or $N_p<13$) the use of robust algorithms tend to increase α-risk while the use of mean value and standard deviation tend to decrease α-risk.

On the other hand, the comparison of Figure 3 and Figure 4 shows that the presence of outliers tends to decrease α-risk. Indeed, in those cases the σ_{pt} standard deviation is significantly over estimated, what decreases the z-scores of all participants including those of the opposite side of the distribution of results.

In any cases, even in very bad PT conditions (i.e. λ=3 and/or N_p=5) α-risk always remains very low (less than 0,7%), see Figure 3.

3.5. Discussion about β-risks

Whatever the situation (with or without presence of an outlier), β is mainly governed by 1. the λ ratio and 2. the number of participants.

Without any outlier, using λ≤0,3 and N_p≥13, is needed to get a β-risk less than 20%, see Figure 3.

When an outlier whose z=10 is present:

1. The β-risk for the main population is very close to 0 in almost all cases for which λ≤0,9, whatever N_p, see Figure 4;

2. The β-risk for the outlier is under control as soon as λ≤0,3, whatever the number of participants, see Figure 5

Figure 6 and Figure 7 show the β-risks respectively for the main participants and to the outlier in function of the outlier's z-score.

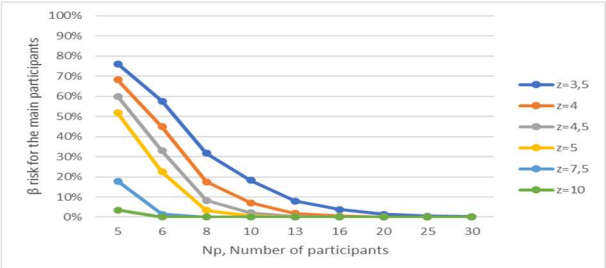

Figure 6: β-risks obtained with Algorithm A and λ=0,17 for the main participants when an outlier is present, in function of the outlier's z-score.

It is reminded that 0,3% of the participants of the main population get z-scores z<-3 or z>+3. However, the H_0 hypothesis considers them as outliers, so that the H_1 hypothesis can be checked, i.e. a β-risk can be computed.

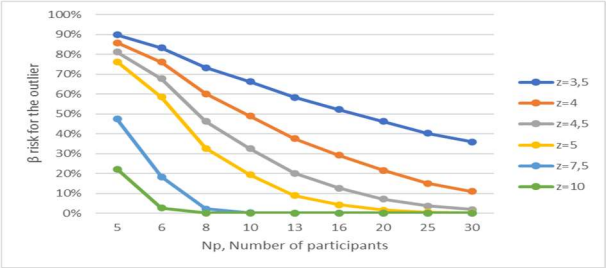

Figure 7: β-risks obtained with Algorithm A and λ=0,17 for an outlier in function of its z-score.

These figures show that 6 participants are enough to detect a strongly outlying participant (whose z-score is 10), while 30 participants are not enough to detect a slightly outlying participant (whose z-score is 3,5).

3.6. General conclusions

This study enabled to show that:

1. The ratio $\lambda = \sigma_r/(\sigma_L \times \sqrt{N_r})$ is of main importance to control the efficiency of a PT scheme, even more than the number of participants. The PT providers should then care N_r, number of test results per participant that they request;

2. Robust algorithms improve the efficiency of the PT program (i.e. β-risk) at a slight expense on α-risk. This comes from a significantly better estimation of the standard deviation of reference when an outlier is present among the participants;

3. Even in adverse conditions, the α-risk is always very low (less than 0,7%);

4. A number of 6 participants is large enough to detect a strongly outlying participant provided that good PT conditions (i.e. low value of λ) are present.

Reference standards [2] and [3] recommend not to organise an ILC with less than 12 participants. This makes sense for [2], which goal is to determine the performance of a test method. It makes less sense for [3], which goal is to check the performance of a lab. Obviously, when no PT is organised, the β-risk is 100%, as any lab having a problem cannot at all realise it! Consequently, for test methods that are performed by a little number of labs, it is probably better to organise PT with 6 participants than nothing. In those cases, the PT provider should specially care the N_r it requests, to ensure a proper λ value and consequently an efficiency as good as possible.

4. SUMMARY

A minimal number of participants is requested by reference standards to organise a PT program, in order to ensure a minimal efficiency of it. However, these recommendations mainly care the so-called α-risk to trigger an alert where it should not, rather than the so called β-risk not to trigger an alert where it should.

This paper shows:

1. How an alternative hypothesis for computing the β-risk can be defined;

2. How the Monte-Carlo method can be used to overcome the difficulties of calculation of β-risk;

3. How a ratio $\lambda = \sigma_r/(\sigma_L \times \sqrt{N_r})$ mainly governs the efficiency of the PT, even more than the number of participants;

4. That a limit value 0,17 exists for λ, ensuring a max efficiency of the PT for each number of participants;

5. Confirm that using the so-called robust statistical methods as described in the reference standards should always be used, even for low numbers of participants;

6. Confirm that in all cases, the α-risk remains always low, even with a very low number of participants;

7. Provide numerical values of β-risks for main participants as well as for outliers, in function of the λ ratio, number of participants and the z-score of the outlying participant;

8. Provide guidance to improve the β-risk of PT programs, whatever the number of participants;

9. Confirm that a PT with a low number of participants is (almost) always better than nothing.

5. REFERENCES

[1] ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories

[2] ISO 5725-2:2019 Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method

[3] ISO 13528:2015 Statistical methods for use in proficiency testing by interlaboratory comparison

[4] ISO 17043:2010 General requirements for proficiency testing

[5] ISO 2859-1:1999 Sampling procedures for inspection by attributes — Part 1: Sampling schemes indexed by acceptance quality limit (AQL) for lot-by-lot inspection

[6] David Luengo, Luca Martino, Mónica Bugallo, Víctor Elvira and Simo Särkkä, "A survey of Monte Carlo methods for parameter estimation" EURASIP Journal on Advances in Signal Processing, Article 25, May 2020 DOI: https://doi.org/10.1186/s13634-020-00675-6

[7] ISO 5725-1:1994 Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions

[8] ISO 16269-4:2010 Statistical interpretation of data — Part 4: Detection and treatment of outliers

[9] Maria Belli, Stephen, L. R. Ellison, Ales Fajgelj, Ilya Kuselman, Umberto Sansone, Wolfhard Wegscheider, "Implementation of proficiency testing schemes for a limited number of participants", Accreditation and Quality Assurance vol. 12 pp. 391–398, February 2007 DOI:https://link.springer.com/article/10.1007/s00769-006-0247-0

[10] Isao Kojima, Kakutoshi kakita, "Comparative study of robustness of statistical methods for laboratory proficiency testing", Analytical sciences, The journal of the Japanese Society for Analytical Chemistry, vol. 30, December 2014 DOI: https://doi.org/10.2116/analsci.30.1165

[11] Dimitris Tasmatsoulis, "Comparing the Robustness of Statistical Estimators of Proficiency Testing Schemes for a Limited Number of Participants", Computation, 2022 10(3) 44, February 2022 DOI: https://doi.org/10.3390/computation10030044