

# Lab proficiency testing: Proposals for limits that balance risks of triggering false alerts and lack of true alerts

Louis-Jean Hollebecq  
Scientific and technical Manager

## Table of content

1	Introduction .....	2
2	Symbols and abbreviations .....	3
3	Basics for the determination of limits of signals.....	4
3.1	Introduction.....	4
3.2	Determination of an appropriate level for nominal $\alpha$ and $\theta$ risks.....	5
3.3	Determination of an appropriate level of confidence for the determination of limits of alerts corresponding to this nominal risk.....	5
3.4	Shape of distribution .....	5
3.5	Basics used to compute the alerts.....	6
3.6	Basics about the impact of outliers and about the method used to reduce it.....	6
3.7	Basics used to determine reference parameters for the PT .....	6
3.8	Basics about the use of the Monte-Carlo method .....	7
4	Results for the assessment of bias.....	7
4.1	Introduction.....	7
4.2	Results of determinations .....	8
4.3	Conclusions.....	11
5	Results for the assessment of repeatability.....	11
5.1	Introduction.....	11
5.2	Results of determinations .....	12
5.3	Conclusions.....	15
6	Results for assessments using non-parametric methods .....	15
6.1	Introduction.....	15
6.2	Results for the first basics .....	17
6.3	Results for the second basics .....	17
6.4	Adaptation of the levels of risk and of IC to the cases where non-parametric methods need to be used.....	18
6.5	Use of this method for non-numerical test results .....	18
6.6	Conclusions for limits determined with non-parametric methods .....	19
7	Conclusions .....	19
8	References .....	20



## Abstract:

PT is based on scores that shall not overcome limits, (typically 2 and 3 for bias) checking whether participants pertain the main population. These limits are always conventional and associated theoretical  $\alpha$  risks are not always same (limits 2 and 3 correspond to 2,275% and 0,135% while ISO 5725-2 considers 1% and 5% risks). Usual practice using 2 warning levels enables to distinguish doubtful and bad performances. However, probabilities to fail to declare participants results as outliers ( $\beta$  risk) are not considered with these limits. In this study, we defined a “doubtful” zone as where both  $\alpha$  and  $\beta$  risks are low and balanced rather than ignoring the  $\beta$  risk. This avoids the usual situations where  $\beta$  is very large, i.e. PT with very low power. We determined corresponding limits for assessing bias and repeatability with  $\alpha=\beta=1\%$  at a level of confidence of 90%. For bias, these “bands of doubt” are close to usual ones when  $n=110$ , enlarged for lower values of  $n$  and vice-versa. We also determined limits using non-parametric methods, then expressed as ranks rather than scores. Unsurprisingly, this is less efficient and powerful and should be used only when parametric methods cannot be used.

## 1 Introduction



Performing proficiency testing of labs requests to fix limits for triggering alerts for scores that are computed to evaluate the performance of the participants. These limits are usually related to a theoretical  $\alpha$  risk of triggering an alert for a participant that is part of the population fulfilling the requirements for the test method. De facto, results of a participant that actually fulfils the requirements for the test method but, by chance, are part of one of the tails of the distribution will be regarded as an “outlier” even if they are not on a theoretical point of view.

Two different traditions can be distinguished concerning this issue:

-  ISO 5725-2 [1] provides tables including values of  $\alpha$  risk of 1% and 5%;
-  ISO 13528 [2] and ISO 17043 [3] recommend limits of 2 and 3 for z-scores concerning the bias, that implicitly refers to a gaussian distribution with bilateral values for  $\alpha$  risk equal to 0,135% and 2,275% respectively. It can also be noticed that, in other parts of ISO 13528 [2] (i.e. its § 10.6 concerning a combined score of bias and repeatability), examples are provided with  $\alpha$  values of 1% and 5%.

It can be concluded that the choice of  $\alpha$  values is always conventional. 0,135% or 1% are usually selected for action limits and 2,275% or 5% for warning limits, but other limits could make sense. For example, it could make sense to adopt a limit of 20% when the risk of non-detecting a malfunction is very critical. In all cases, the PT provider should make clear which  $\alpha$  values it is using and justify them when they are not the usually adopted ones.

In any cases, no consideration is given to the counterpart  $\beta$  risk of not triggering an alert for a participant that is not part of the population fulfilling the requirements for the test method, i.e. a true outlier. Several causes may lead to such situations:

-  When the deviation of the participant to the requirements for the test method does not have a very significant impact on its results, the induced bias may be too small to be detected during a PT performance;
-  When the number of participants or of repetitions is small, the effects of estimation of the reference values for the PT may allow an acceptance of many true outliers.

$\alpha$  risk is then a characterisation of the effectiveness of the PT while  $\beta$  risk is a characterisation of the power of the PT. This second issue was extensively studied in [4] and [5]. One of the main conclusions of these studies was that the actual  $\alpha$  risk is always much lower than the theoretical values, while the  $\beta$  risk is usually quite huge, up to almost 100 % when the PT conditions are not good (i.e. when 1- the number of repetitions is too small to cancel the effect

of internal SD of the participating labs or 2- when the number of participants is too small to get a proper assessment of the participants).

The current usual practice of using 2 levels of warning (typically signals of alert and signals of action) enables to distinguish cases where the performance is doubtful from those where the performance is likely to be bad. On the other hand, the willing to maintain alerts for doubtful performances can be explained by the willing to alert participants even in doubtful cases, that is to say, to keep in mind the existence of the  $\beta$  risk and avoid too many situations where the participant's results are actually outlying without triggering any alert, knowing that the  $\beta$  risk obviously decreases when the  $\alpha$  risk increases.

Then the idea comes that we could define a "doubtful" zone as where both  $\alpha$  and  $\beta$  risk are not small, what enables us to balance these risks, rather than ignoring the  $\beta$  risk as usually made, what leads to a massive imbalance between them in practice ( $\alpha \ll \beta$ ) and consequently, a lack of power of the PT.

To achieve this, the following steps need to be carried out:

1. Decide an equal and appropriate level of risk for  $\alpha$  and  $\beta$ , that we can call the "nominal risk";
2. Decide an appropriate level of confidence for the determination of limits of alerts corresponding to this nominal risk;
3. Determine these limits as the limits that may take a computed score during a PT for a participant which true score corresponds to the nominal risk, as function of the upper and of the number of participants

The aim of this study is to detail these steps for 3 different situations:

- For the case of a normal distribution, corresponding to the usual assessment of bias (typically z-scores) and of uncertainties (typically  $\zeta$ -scores as detailed in ISO 13528 [2]);
- For the case of a  $\chi^2$  based distribution corresponding to the assessment of repeatability, typically  $z_r$ -scores, as described in [6];
- For the case of an unknown distribution, using non-parametric statistics.

## 2 Symbols and abbreviations






The symbols used in this document are listed in Table 1.

Table 1. List of symbols used in this document.

Symbol	Designation and comments
$2u$	Enlarged uncertainty with $k = 2$
$k$	Enlargement coefficient for the computation of IC
<b>Limit-</b>	Lower limit of "doubt zones", i.e. of IC on scores of a participant exactly located at the nominal limit
<b>Limit+</b>	Upper limit of "doubt zones", i.e. of IC on scores of a participant exactly located at the nominal limit
$n$	Number of participants to a PT
$r$	Number of repetitions by a same participant during a PT
$sr_i$	Estimate of $\sigma_{ri}$
$sr_{ref}$	Estimate of $\sigma_r$ taken as reference for a PT
$X_{pt}$	Reference value used to assess bias during a PT
<b>Z-score</b>	Parameter that characterises the bias of a participant, as defined in ISO 13528

<i>Symbol</i>	<b>Designation and comments</b>
<b><i>z-score</i></b>	Estimate of the Z-score during a PT
<b><i>ZR-score</i></b>	Parameter that characterises the repeatability of a participant, defined as the ratio $S_{ri}/S_{ref}$
<b><i>zr-score</i></b>	Estimate of the <i>zr</i> -score of a participant during a PT
<b><math>\alpha</math></b>	Risk of triggering an alert for a participant that does not deserve it
<b><math>\beta</math></b>	Risk of failing to trigger an alert for a participant that deserves it
<b><math>\sigma_L</math></b>	Interlaboratory standard deviation
<b><math>\sigma_{pt}</math></b>	Reference value for standard deviations of participants biases during a PT
<b><math>\sigma_r</math></b>	Standard deviation of repeatability
<b><math>\sigma_{ri}</math></b>	Standard deviation of repeatability of participant <i>i</i>

Abbreviations:







-  IC: bilateral interval of confidence. For example, IC95% means the bilateral interval of confidence [2,5%;97,5%]
-  ILC: interlaboratory comparison
-  MCM: Monte-Carlo method
-  PT: proficiency tests
-  SD: standard deviation

## 3 Basics for the determination of limits of signals

### 3.1 Introduction

With regards to statements of § 1, the computation of limits of alert requests to consider the distribution of the computed scores for a participant exactly corresponding to the nominal risk, and consider the centiles of this distribution to determine an appropriate IC for the estimation of this nominal risk.

This involves the following issues:

-  The determination of an appropriate level for nominal  $\alpha$  and  $\beta$  risks, in order to determine a nominal Z-score of reference of a hypothetical participant exactly corresponding to the nominal risk;
-  The determination of an appropriate level of confidence for the determination of limits of alerts corresponding to this nominal risk;
-  The shape of the distribution of test results;
-  The basics used to determine limits of alerts;
-  The impact of outliers and of the method used to reduce it;
-  The basics used to determine reference parameters for the PT.

All these issues are detailed here after.

As the computation of limits involves some issues that can hardly be properly handled with theoretical tools (typically the robust algorithms for determining the reference parameters), the Monte-Carlo method (MCM) was used to determine our proposals for limits of signals, see for example [7] to get explanations about it.

### 3.2 Determination of an appropriate level for nominal $\alpha$ and $\beta$ risks

As reminded in § 1, the choice of  $\alpha$  values is always conventional. 0,135% or 1% are usually selected for action limits and 2,275% or 5% for warning limits, but other limits could make sense.

An appropriate value for  $\alpha$  and  $\beta$  risks should then be such that the “doubtful” area is when both  $\alpha$  and  $\beta$  risks is not reached, with a specified probability. It is then obvious that this level should have something to do with the traditional values selected for determining the warning signals and the action signals.

On our own, we chose a nominal risk of 1% bilateral (i.e. 0,5% unilateral), which leads for Gaussian distributions to a central limit of alert of 2,576, that can stand as an average value between the traditional values of 2 and 3 for warning signals and action signals respectively.

Such a choice lets us expect that the limits of alerts computed from the IC will be distributed at roughly the same distance from this nominal value, i.e. will be not far from the traditional values for the limits of alerts (i.e. 2 and 3) for a certain number of participants, not far from usual situations encountered in PT exercises.

Obviously, other values could make sense, but we will see further than this choice was an appropriate one.

### 3.3 Determination of an appropriate level of confidence for the determination of limits of alerts corresponding to this nominal risk

For each number of participants, the uncertainty on parameters used to compute scores leads to an uncertainty on the values of these computed scores. This uncertainty obviously decreases when the number of participants increases. The issue is then to determine a level of confidence for the determination of limits of alerts that:

- ✚ Makes sense, i.e. corresponds to a reasonable risk with regard with usual practices;
- ✚ Is not too far from current practices, i.e. triggers a reasonable number of alerts, not too few, not too many.

We founded out that IC90% (i.e. bilateral level of confidence of 90%) make sense with respect to these 2 requests.

### 3.4 Shape of distribution

Most of PT are intended to assess the bias of the results of participants, for which a Gaussian distribution is implicitly supposed in reference standards. ISO 13528 [2] states that normality of the distribution is not a requirement but recommends verifying that the distribution is symmetric and, if not, recommends using a change of variable to make it symmetric.

However, some types of assessment obviously request to consider other types of distributions. In particular, the assessment of repeatability of participants requests to consider  $\chi^2$  based distributions (see [6]) linked to the distribution of estimates of a SD.

We may also consider cases where no correct assumption can be made about the shape of the distribution. We have then tried to deal with such situations using nonparametric statistics, keeping in mind that, obviously, these methods can never be as powerful as parametric ones.

This study has then considered 3 cases:

- ✚ Test results following a Gaussian distribution;
- ✚ Standard deviations of test results, following a  $\chi^2$  based distributions, see [6];
- ✚ Data following an unknown distribution.

### 3.5 Basics used to compute the alerts

Even if this is not explicitly dealt with in reference standards (ISO 5725-2 [1], ISO 13528 [2] and ISO 17043 [3]), at least 2 basics are used to compute alerts:

- ✚ The first one, used to assess bias (typically z-scores), considers that outliers are those located at the tails of the distribution. The alerts are then based on a check whether the participant's results are located within or beyond a given distance from the reference value;
- ✚ The second one, used to check repeatability (typically Cochran ratios and Mandel k-scores), considers that all participants should show the same repeatability SD (hypothesis of homoscedasticity). The alerts are then based on a check whether the participant's results are significantly different from the reference value.

The first type of assessment requests to determine 2 parameters:

- ✚ A central reference value (typically  $X_{pt}$  of ISO 13528 [2]);
- ✚ A reference value for the acceptable distance from the reference value (typically  $\sigma_{pt}$  of ISO 13528 [2]).

The second type of assessment requests the determination of only 1 parameter: the central reference value (typically  $\sigma_r$  for the assessment of repeatability). It shall be noted that if used for bias, this option would lead to consider all participants that do not meet the IC on the reference value as outliers, i.e. most of the participants. It would then be a method that cannot be used in practice, because it would trigger too many alerts.

This study has then considered:

- ✚ The first type for assessments using Gaussian distributions;
- ✚ The second type for SD assessments;
- ✚ Both types for unknown distributions.

### 3.6 Basics about the impact of outliers and about the method used to reduce it

All reference methods for ILC (ISO 5725-2 [1], ISO 13528 [2] and ISO 17043 [3]) warn against the impact of outliers on the results of comparisons because, during ILC exercises, the organiser has low control over the quality of the data. In particular, when some figures seem strange, it is not possible to check whether some technical or practical reason can explain them and make possible to decide on a technical ground that these figures are not part of the main population of results (i.e., the population of test results obtained in accordance with the requirements for the test method). Several robust statistical methods are described in these standards to cope with the presence of outliers.

Moreover, when an almost outlying test result is present in the series of test results (what is, by construction, always the case in our MCM series), this affects the determination of the parameters and consequently affects the results of assessment. This effect is particularly strong when the PT exercise involves a low number of participants.

To take all of this into account, our computations were conducted using Algo A and Algo S robust methods (as described in ISO 5725-2 [1] and ISO 13528 [2]) to compute the reference parameters for the distributions, which are the most commonly used throughout the PT providers.

### 3.7 Basics used to determine reference parameters for the PT

Even if, for practical reasons, consensus of participants is usually used to determine the reference parameters for the PT, several other good solutions exist (see ISO 13528 [2]) that, according to the cases, can be better adapted to the situation than the usual method of consensus of participants.

Obviously, no statistical method may be used to determine limits when the reference value is not determined with the “consensus” method. For example, for assessment of bias against ISO 13528 [2], when the central value  $X_{pt}$  is fixed by formulation and the acceptable deviation  $\sigma_{pt}$  comes from an external source, some assumptions need to be made about the accuracy with which these parameters can be determined. These assumptions are related to the basics of the test method rather than statistical issues, and can never be applicable to all PT, i.e. become general rules for PT performance.

That is why we decided to focus our study about the consensus method. As a matter of fact, the consensus method is usually less efficient than the others and consequently, using the limits that we have computed in this study is likely to lead to IC with better confidence than our nominal 90%. Consequently, in most cases, they will make sense whatever the way with which the reference values are computed, even if they were computed only with the consensus method.

### 3.8 Basics about the use of the Monte-Carlo method

Using MCM requests to use a model that reasonably fits the situations encountered in the real world. This was easy for this study, for which all models are available in reference standards.

Using the Monte-Carlo methods also requests to use random input values. When several random values are necessary to produce one Monte-Carlo result and when correlations between them apply in real life, these correlations must be incorporated in the input values of the computations. In our case, we had to produce only one random value, either biases or random SD so that this was not an issue for our study.

To assure the validity of the conclusions, the random series need to be numerous enough, depending on many factors. In our study, we computed series from  $10^7$  to  $10^8$  series of results for each situation. Each of them was divided in 50 to 1000 sub-groups enabling us to check how repeatable the computed parameters and percentage of alerts were within these sub-groups and compute a related IC (with  $k = 2$ ).

## 4 Results for the assessment of bias

### 4.1 Introduction

With respect to statements of § 3, the determinations were carried out as follows:






-  Series of participants from 5 to 250 were considered;
-  Participants results were produced using a Gaussian distribution;
-  In each series of participants, one of them was set at the nominal limit of alert (i.e.  $Z = 2,576$ , corresponding to nominal risks  $\alpha$  and  $\beta$  of 1% bilateral) instead of being random;
-  Algo A was used to determine the reference value and the reference standard deviation of the distribution, then used to compute the estimated z-score of the participant with true value  $Z = 2,576$ );
-  The centiles 5% and 95% of these estimated z-scores were then computed to determine the corresponding IC90%.

Figure 1 shows an example of distribution of a sub-series of 400 estimates of z-scores for a participant whose true Z-score is 2,576, with a number of participants is 25. The corresponding centiles 5% and 95% are respectively 1,711 and 3,743 in this case, not far from the values 1,673 and 3,927 determined with 100 repetitions of such curves.



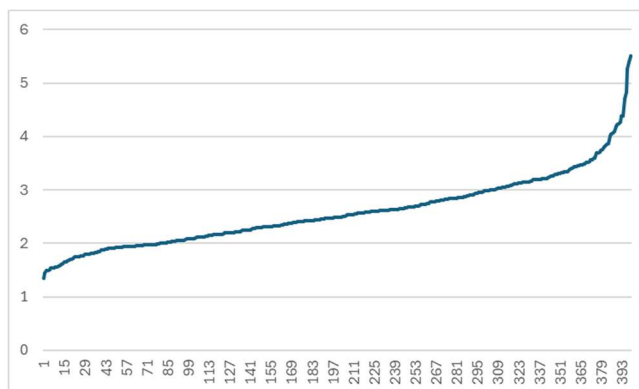


Figure 1: Example of distribution of a sub-series of 400 estimates of z-scores for a participant whose true Z-score is 2,576, with a number of participants equal to 25. The corresponding centiles 5% and 95% are respectively 1,711 and 3,743 in this case.

It shall be noted that these limits apply to the apparent value of  $\sigma_{pt}$ , which is usually greater than  $\sigma_L$  because an additional contribution of the inner SD of the lab can never be avoided, see [4] and [5]. This effect reduces the  $\alpha$  risk and increases the  $\beta$  risk. References [4] and [5] also demonstrated that when a  $\lambda$  ratio defined as  $\lambda = \sigma_r / (r \cdot \sigma_L)$  (where  $\sigma_r$  is the SD of repetitions within a same participant,  $r$  is the number of repetitions within each participant and  $\sigma_L$  is the interlaboratory SD) is less than 0,17, this effect becomes negligible and the  $\alpha$  and  $\beta$  risks only depend on the number of participants.

## 4.2 Results of determinations

Table 2 provides the results of determination of limits for alerts corresponding nominal risks  $\alpha$  and  $\beta$  equal to 1% (bilateral), corresponding to true Z-scores equal to 2,576. Of course, these values are opposite and inverted with negative signs for true Z-scores equal to -2,576 (for example, [-4,96;-1,24] for  $n = 10$ ).

Table 2. Results of determination of limits for alerts corresponding nominal risks  $\alpha$  and  $\beta$  equal to 0,5%.

n	Results				Proposals	
	Limit-	2u	Limit+	2u	Limit-	Limit+
3	0,6743	0,0000	13,468	0,021	<b>0,67</b>	<b>13,5</b>
4	0,7943	0,0002	6,6109	0,0064	<b>0,79</b>	<b>6,6</b>
5	0,8194	0,0004	8,6169	0,0085	<b>0,82</b>	<b>8,6</b>
6	0,9184	0,0004	5,9168	0,0047	<b>0,92</b>	<b>5,9</b>
7	0,9952	0,0004	6,2009	0,0053	<b>0,995</b>	<b>6,2</b>
8	1,1128	0,0004	5,3178	0,0041	<b>1,11</b>	<b>5,3</b>
9	1,1680	0,0005	5,6239	0,0047	<b>1,17</b>	<b>5,6</b>
10	1,2391	0,0005	4,9617	0,0035	<b>1,24</b>	<b>4,96</b>
11	1,2773	0,0006	5,0483	0,0038	<b>1,28</b>	<b>5,05</b>
12	1,3395	0,0006	4,6708	0,0034	<b>1,34</b>	<b>4,67</b>
13	1,3685	0,0006	4,7614	0,0035	<b>1,36</b>	<b>4,76</b>
14	1,4186	0,0006	4,4737	0,0032	<b>1,42</b>	<b>4,475</b>
15	1,4408	0,0006	4,5188	0,0031	<b>1,44</b>	<b>4,52</b>
16	1,4837	0,0007	4,3083	0,0029	<b>1,48</b>	<b>4,31</b>
17	1,5015	0,0007	4,3511	0,0032	<b>1,5</b>	<b>4,35</b>
18	1,5392	0,0007	4,1857	0,0029	<b>1,54</b>	<b>4,185</b>
19	1,5539	0,0007	4,2120	0,0030	<b>1,55</b>	<b>4,21</b>
20	1,5866	0,0007	4,0773	0,0027	<b>1,59</b>	<b>4,08</b>

n	Results				Proposals	
	Limit-	2u	Limit+	2u	Limit-	Limit+
21	1,5994	0,0007	4,0984	0,0027	<b>1,6</b>	<b>4,1</b>
22	1,6278	0,0007	3,9897	0,0026	<b>1,63</b>	<b>3,99</b>
23	1,6381	0,0007	4,0054	0,0027	<b>1,64</b>	<b>4,005</b>
24	1,6638	0,0007	3,9138	0,0025	<b>1,66</b>	<b>3,915</b>
25	1,6731	0,0008	3,9266	0,0027	<b>1,67</b>	<b>3,925</b>
26	1,6954	0,0008	3,8504	0,0025	<b>1,695</b>	<b>3,85</b>
27	1,7031	0,0007	3,8587	0,0025	<b>1,7</b>	<b>3,86</b>
28	1,7240	0,0008	3,7917	0,0024	<b>1,72</b>	<b>3,79</b>
29	1,7305	0,0008	3,7970	0,0024	<b>1,73</b>	<b>3,8</b>
30	1,7493	0,0008	3,7384	0,0024	<b>1,75</b>	<b>3,74</b>
31	1,7556	0,0008	3,7461	0,0024	<b>1,755</b>	<b>3,745</b>
32	1,7726	0,0008	3,6956	0,0024	<b>1,77</b>	<b>3,695</b>
33	1,7788	0,0008	3,7039	0,0023	<b>1,78</b>	<b>3,705</b>
34	1,7941	0,0008	3,6538	0,0023	<b>1,795</b>	<b>3,655</b>
35	1,7992	0,0008	3,6599	0,0023	<b>1,8</b>	<b>3,66</b>
36	1,8139	0,0008	3,6169	0,0021	<b>1,81</b>	<b>3,62</b>
37	1,8179	0,0009	3,6239	0,0023	<b>1,82</b>	<b>3,625</b>
38	1,8321	0,0008	3,5832	0,0022	<b>1,83</b>	<b>3,58</b>



n	Results				Proposals	
	Limit-	2u	Limit+	2u	Limit-	Limit+
39	1,8357	0,0008	3,5870	0,0022	<b>1,84</b>	<b>3,59</b>
40	1,8494	0,0009	3,5529	0,0022	<b>1,85</b>	<b>3,55</b>
45	1,8830	0,0009	3,5006	0,0022	<b>1,88</b>	<b>3,5</b>
50	1,9186	0,0009	3,4331	0,0021	<b>1,92</b>	<b>3,43</b>
55	1,9428	0,0009	3,3965	0,0021	<b>1,94</b>	<b>3,395</b>
60	1,9714	0,0010	3,3456	0,0019	<b>1,97</b>	<b>3,345</b>
65	1,9893	0,0009	3,3181	0,0020	<b>1,99</b>	<b>3,32</b>
70	2,0122	0,0009	3,2810	0,0019	<b>2,01</b>	<b>3,28</b>
75	2,0268	0,0010	3,2599	0,0019	<b>2,03</b>	<b>3,26</b>
80	2,0459	0,0010	3,2296	0,0018	<b>2,045</b>	<b>3,23</b>
85	2,0594	0,0010	3,2113	0,0020	<b>2,06</b>	<b>3,21</b>
90	2,0724	0,0010	3,1894	0,0019	<b>2,07</b>	<b>3,19</b>
95	2,0833	0,0010	3,1726	0,0018	<b>2,08</b>	<b>3,17</b>
100	2,0962	0,0010	3,1527	0,0019	<b>2,095</b>	<b>3,15</b>
105	2,1056	0,0010	3,1401	0,0019	<b>2,105</b>	<b>3,14</b>
110	2,1169	0,0010	3,1232	0,0017	<b>2,12</b>	<b>3,125</b>
115	2,1254	0,0010	3,1129	0,0018	<b>2,125</b>	<b>3,11</b>

n	Results				Proposals	
	Limit-	2u	Limit+	2u	Limit-	Limit+
120	2,1357	0,0010	3,0971	0,0017	<b>2,135</b>	<b>3,1</b>
125	2,1425	0,0010	3,0895	0,0016	<b>2,14</b>	<b>3,09</b>
130	2,1517	0,0010	3,0746	0,0017	<b>2,15</b>	<b>3,075</b>
135	2,1574	0,0010	3,0663	0,0017	<b>2,16</b>	<b>3,065</b>
140	2,1657	0,0010	3,0556	0,0018	<b>2,165</b>	<b>3,055</b>
145	2,1714	0,0010	3,0468	0,0017	<b>2,17</b>	<b>3,045</b>
150	2,1789	0,0010	3,0369	0,0018	<b>2,18</b>	<b>3,035</b>
160	2,1911	0,0011	3,0205	0,0017	<b>2,19</b>	<b>3,02</b>
170	2,2015	0,0010	3,0070	0,0017	<b>2,2</b>	<b>3,01</b>
180	2,2115	0,0011	2,9948	0,0015	<b>2,21</b>	<b>2,995</b>
190	2,2212	0,0010	2,9801	0,0017	<b>2,22</b>	<b>2,98</b>
200	2,2286	0,0011	2,9712	0,0017	<b>2,23</b>	<b>2,97</b>
210	2,2370	0,0011	2,9588	0,0017	<b>2,24</b>	<b>2,96</b>
220	2,2446	0,0011	2,9504	0,0017	<b>2,245</b>	<b>2,95</b>
230	2,2519	0,0012	2,9399	0,0016	<b>2,25</b>	<b>2,94</b>
240	2,2579	0,0010	2,9335	0,0016	<b>2,26</b>	<b>2,93</b>
250	2,2645	0,0011	2,9244	0,0016	<b>2,265</b>	<b>2,925</b>

It can be seen from these figures that the classical limits for z-scores, i.e. 2 and 3 are not far from the situation where  $\alpha$  and  $\beta$  are equal to 1% bilateral with an IC90% when  $n = 110$ .

Figure 2 shows the “Limit+” values for alerts corresponding nominal risks  $\alpha$  and  $\beta$  equal to 1% bilateral for  $3 \leq n \leq 25$ . It can be seen in this figure that odd values of  $n$  show higher values for limits than the neighbouring even values of  $n$ . The same phenomenon also occurs for “Limit-” values but with less importance. Unsurprisingly, the importance of the phenomenon decreases when  $n$  increases.

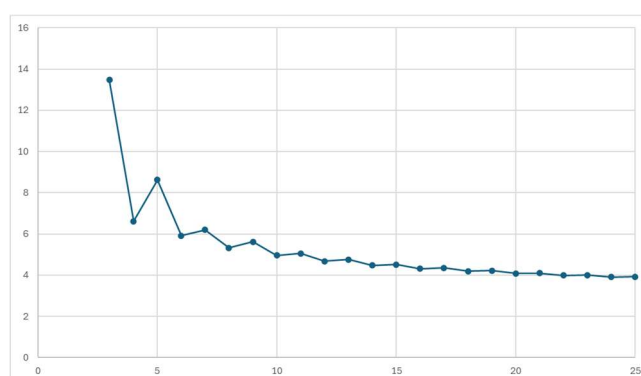
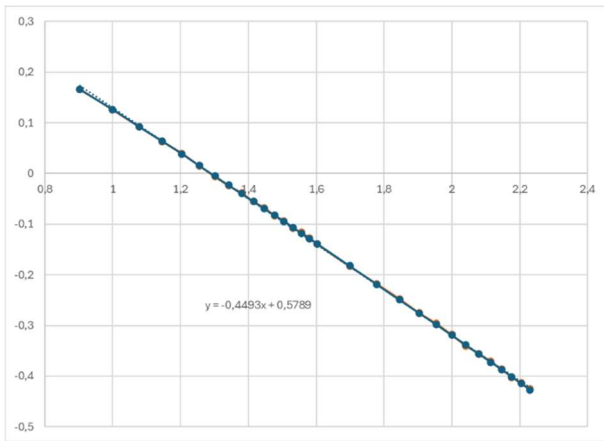
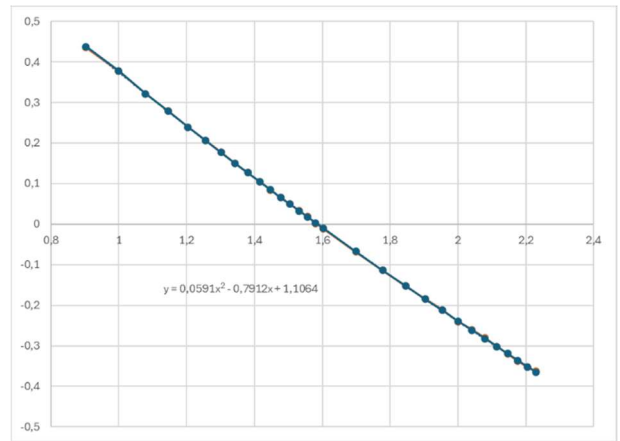


Figure 2: Limits+ for alerts corresponding nominal risks  $\alpha$  and  $\beta$  equal to 0,5% for  $3 \leq n \leq 25$ .

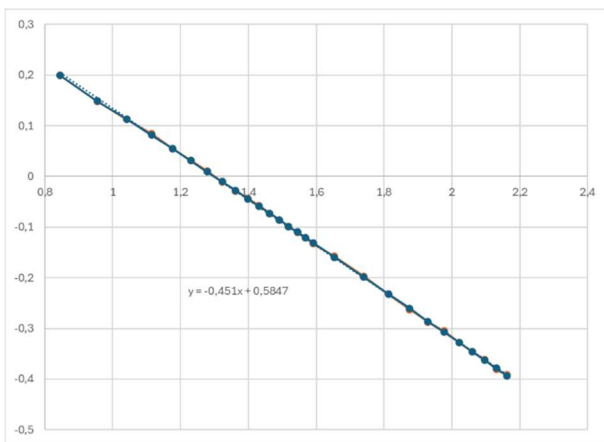
For this reason, we separated even numbers and odd numbers when finding out empirical equations that enable us to compute intermediate values of limits. To achieve this, we plotted the log values of gaps between the alerts and the nominal value (i.e. 2,576) as function of the log of the number of participants, see Figure 3.



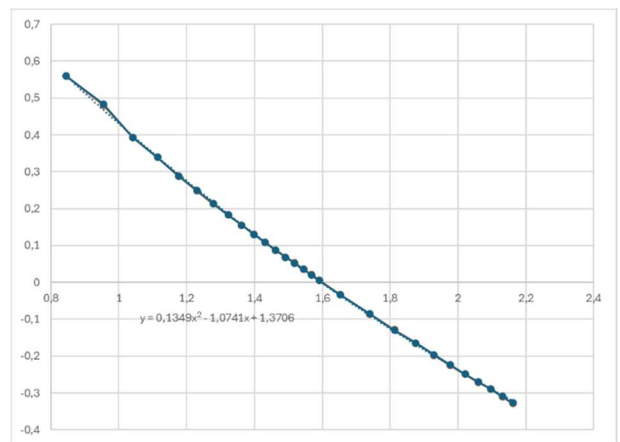
Lower limits for even numbers of participants



Upper limits for even numbers of participants



Lower limits for odd numbers of participants



Upper limits for odd numbers of participants

Figure 3: Logs of gaps between limits of alerts and nominal value (2,576) as function of log of number of participants for nominal risks  $\alpha$  and  $\beta$  equal to 0,5% and for  $7 \leq n \leq 250$ .

It can be seen from these figures that linear regressions enable to compute accurately enough the values of lower limits. For upper limits, we need to use a polynomial of degree 2 if we want to get an accuracy better than 0,01, even if the curve is not far from being straight.

The following Equation (1) enables to compute the limits of alerts for  $10 \leq n \leq 250$  to the nearest 0,02 for Limit+ for odd values of  $n$  and 0,01 for the other limits.

$$Lim = 2,576 \pm 10^a \quad (1)$$

Where

$\pm$  is “-” for lower limits and “+” for upper limits,

$a = -0,45 \cdot \log(n) + 0,58$  for lower limit when  $n$  is even,

$a = -0,45 \cdot \log(n) + 0,585$  for lower limit when  $n$  is odd,

$a = 0,059 \cdot (\log(n))^2 - 0,791 \cdot \log(n) + 1,106$  for upper limit when  $n$  is even,

$a = 0,135 \cdot (\log(n))^2 - 1,075 \cdot \log(n) + 1,37$  for upper limit when  $n$  is odd,

and  $n$  is the number of results used to compute  $a$ .

### 4.3 Conclusions

The classical limits for z-scores, i.e. 2 and 3 are not far from the situation where  $\alpha$  and  $\beta$  are equal to 1% bilateral with an IC90% when  $n = 110$ .

The IC90% around the nominal value 2,576 is higher for odd values of  $n$  than neighbouring even values of  $n$ .

Empirical formulas could be determined for the computation of the limits, as function of the number of participants.

These limits apply to the apparent value of  $\sigma_{pt}$ , that is usually greater than  $\sigma_L$  because an additional contribution of the inner SD of the lab can never be avoided. This effect reduces the  $\alpha$  risk and increases the  $\beta$  risk. When a  $\lambda$  ratio defined as  $\lambda = \sigma_r / (n_r \cdot \sigma_L)$  is less than 0,17, this effect becomes negligible and no difference between  $\alpha$  and  $\beta$  risk then applies.

When the  $\lambda$  ratio is significantly greater than 0,17, a solution to balance  $\alpha$  and  $\beta$  risks would be to use  $\sigma_{pt} = \sqrt{s_{pt}^2 - s_r^2 / r}$  (where  $s_{pt}$  is the SD of the mean value of the participants and  $s_r$  is the average SD of the repetitions within participant (that can be computed with Algo S). However, the IC on these  $\sigma_{pt}$  implying a nested variance is usually quite larger than on  $s_{pt}$  (see [8]) and consequently, our proposals of Table 2 are not valid (too narrow) for such a use.

## 5 Results for the assessment of repeatability

### 5.1 Introduction

With respect to statements of § 3, the determinations were carried out as follows:

- ✚ Series of from 3 to 250 participants and from 2 to 25 repetitions per participant were considered;
- ✚ Participants results were produced using a  $\sqrt{\chi_{r-1}^2 / (r - 1)}$  distribution, that describes the estimations of repeatability SD, see [6];
- ✚ As proposed in usual reference standards, the calculation of limits was made using the hypothesis of homoscedasticity, see § 3.5 and [6];
- ✚ zr-scores, i.e. ratios  $sr_i / sr_{ref}$  (where  $sr_i$  is the repeatability SD of participant  $i$  and  $sr_{ref}$  is the repeatability SD of reference) as proposed in [6] and that are similar to those proposed in ISO 5725-2 [1], were used to characterize the participants results;
- ✚ In each series of participants, one of them was set at the nominal limit of alert depending on  $r$  (i.e.  $ZR = \sqrt{\chi_{0,995,r-1}^2 / (r - 1)}$ , corresponding to nominal risks  $\alpha$  and  $\beta$  of 1% bilateral) instead of being random;
- ✚ Only upper limits of alerts were considered, because lower limits are usually not relevant for alerting, see [6]. Consequently, the nominal risks  $\alpha$  and  $\beta$  become unilateral and are taken equal to 0,5%;
- ✚ Algo S was used to determine the reference standard deviation of the distribution, then used to compute the estimated zr-scores of the participants with true value  $ZR = \sqrt{\chi_{0,995,r-1}^2 / (r - 1)}$ ;
- ✚ The centiles 5% and 95% of these estimated zr-scores were then computed to determine the corresponding IC90% (in the same way than done for assessment of bias, see Figure 1).

Contrarily to the case of assessment of bias (see § 4.1), thanks to the use of an hypothesis of homoscedasticity, no parasite additional SD is to be feared in the determination of limits, so that the here computed limits truly represent those for which nominal risks  $\alpha$  and  $\beta$  are equal to 0,5% (only the upper side of the distribution is considered).

## 5.2 Results of determinations

Table 3 provides the ZR values corresponding to nominal risks  $\alpha$  and  $\beta$  equal to 0,5%, as function of  $r$ , number of repetitions per participant.

Table 3. ZR values corresponding to nominal risks  $\alpha$  and  $\beta$  equal to 0,5%.

r	2	3	4	5	6	8	10	12	16	20	25
<b>Nominal Limits</b>	2,807	2,302	2,069	1,927	1,830	1,702	1,619	1,560	1,479	1,425	1,378

Table 4 provides the results of determination of limits for alerts corresponding nominal risks  $\alpha$  and  $\beta$  equal to 0,5% (unilateral, upper side), corresponding to true ZR-scores equal to nominal limits of Table 3.

Table 4. Results of determination of limits for alerts corresponding nominal risks  $\alpha$  and  $\beta$  equal to 0,5%.

n	r	Results				Proposals	
		Limit-	2u	Limit+	2u	Limit-	Limit+
3	2	1,1901	0,0002	7,7845	0,0083	<b>1,19</b>	<b>7,8</b>
3	3	1,1936	0,0002	3,8816	0,0022	<b>1,194</b>	<b>3,88</b>
3	4	1,1845	0,0001	2,9915	0,0014	<b>1,184</b>	<b>2,99</b>
3	5	1,1741	0,0001	2,5828	0,0009	<b>1,174</b>	<b>2,58</b>
3	6	1,1658	0,0001	2,3441	0,0007	<b>1,166</b>	<b>2,34</b>
3	8	1,1522	0,0001	2,0681	0,0005	<b>1,152</b>	<b>2,07</b>
3	10	1,1409	0,0001	1,9075	0,0004	<b>1,141</b>	<b>1,91</b>
3	12	1,1309	0,0001	1,7908	0,0004	<b>1,131</b>	<b>1,79</b>
3	16	1,1146	0,0001	1,6288	0,0003	<b>1,115</b>	<b>1,63</b>
3	20	1,1032	0,0001	1,5289	0,0002	<b>1,103</b>	<b>1,53</b>
3	25	1,0924	0,0001	1,4459	0,0002	<b>1,092</b>	<b>1,45</b>
4	2	1,2985	0,0002	5,9487	0,0054	<b>1,299</b>	<b>5,95</b>
4	3	1,2859	0,0002	3,4591	0,0017	<b>1,286</b>	<b>3,46</b>
4	4	1,2669	0,0002	2,7808	0,0011	<b>1,267</b>	<b>2,78</b>
4	5	1,2493	0,0002	2,4475	0,0008	<b>1,249</b>	<b>2,45</b>
4	6	1,2355	0,0001	2,2448	0,0006	<b>1,236</b>	<b>2,245</b>
4	8	1,2139	0,0001	2,0036	0,0004	<b>1,214</b>	<b>2,00</b>
4	10	1,1971	0,0001	1,8603	0,0004	<b>1,197</b>	<b>1,86</b>
4	12	1,1823	0,0001	1,7551	0,0003	<b>1,182</b>	<b>1,755</b>
4	16	1,1593	0,0001	1,6087	0,0002	<b>1,159</b>	<b>1,61</b>
4	20	1,1432	0,0001	1,5170	0,0002	<b>1,143</b>	<b>1,52</b>
4	25	1,1286	0,0001	1,4403	0,0002	<b>1,128</b>	<b>1,44</b>
5	2	1,4443	0,0005	6,9233	0,0081	<b>1,444</b>	<b>6,9</b>
5	3	1,3892	0,0005	3,6544	0,0032	<b>1,39</b>	<b>3,655</b>
5	4	1,3530	0,0004	2,8712	0,0020	<b>1,353</b>	<b>2,87</b>
5	5	1,3253	0,0003	2,5038	0,0013	<b>1,326</b>	<b>2,5</b>
5	6	1,3036	0,0003	2,2867	0,0011	<b>1,304</b>	<b>2,29</b>
5	8	1,2727	0,0003	2,0305	0,0007	<b>1,27</b>	<b>2,03</b>
5	10	1,2493	0,0002	1,8795	0,0006	<b>1,249</b>	<b>1,88</b>
5	12	1,2284	0,0002	1,7697	0,0005	<b>1,228</b>	<b>1,77</b>
5	16	1,1953	0,0002	1,6184	0,0003	<b>1,195</b>	<b>1,62</b>
5	20	1,1736	0,0001	1,5252	0,0003	<b>1,174</b>	<b>1,53</b>
5	25	1,1550	0,0001	1,4473	0,0003	<b>1,155</b>	<b>1,45</b>
6	2	1,5768	0,0007	5,9903	0,0057	<b>1,58</b>	<b>6,0</b>
6	3	1,4744	0,0007	3,4356	0,0027	<b>1,47</b>	<b>3,44</b>
6	4	1,4197	0,0005	2,7591	0,0015	<b>1,42</b>	<b>2,76</b>

n	r	Results				Proposals	
		Limit-	2u	Limit+	2u	Limit-	Limit+
6	5	1,3816	0,0005	2,4298	0,0012	<b>1,38</b>	<b>2,43</b>
6	6	1,3546	0,0004	2,2302	0,0009	<b>1,355</b>	<b>2,23</b>
6	8	1,3156	0,0003	1,9934	0,0007	<b>1,32</b>	<b>1,99</b>
6	10	1,2873	0,0003	1,8513	0,0005	<b>1,29</b>	<b>1,85</b>
6	12	1,2609	0,0002	1,7476	0,0005	<b>1,26</b>	<b>1,75</b>
6	16	1,2205	0,0002	1,6050	0,0003	<b>1,22</b>	<b>1,605</b>
6	20	1,1955	0,0002	1,5160	0,0003	<b>1,20</b>	<b>1,52</b>
6	25	1,1739	0,0001	1,4422	0,0002	<b>1,17</b>	<b>1,44</b>
8	2	1,7734	0,0008	5,7634	0,0057	<b>1,77</b>	<b>5,76</b>
8	3	1,6058	0,0007	3,3491	0,0025	<b>1,61</b>	<b>3,35</b>
8	4	1,5249	0,0005	2,7045	0,0015	<b>1,525</b>	<b>2,70</b>
8	5	1,4715	0,0005	2,3895	0,0011	<b>1,47</b>	<b>2,39</b>
8	6	1,4340	0,0004	2,1989	0,0009	<b>1,43</b>	<b>2,20</b>
8	8	1,3812	0,0003	1,9700	0,0007	<b>1,38</b>	<b>1,97</b>
8	10	1,3439	0,0003	1,8329	0,0005	<b>1,34</b>	<b>1,83</b>
8	12	1,3119	0,0003	1,7329	0,0005	<b>1,31</b>	<b>1,73</b>
8	16	1,2610	0,0002	1,5949	0,0003	<b>1,26</b>	<b>1,595</b>
8	20	1,2289	0,0002	1,5094	0,0003	<b>1,23</b>	<b>1,51</b>
8	25	1,2022	0,0001	1,4378	0,0003	<b>1,20</b>	<b>1,44</b>
10	2	1,9113	0,0009	5,5468	0,0050	<b>1,91</b>	<b>5,55</b>
10	3	1,6961	0,0008	3,2654	0,0023	<b>1,7</b>	<b>3,27</b>
10	4	1,5951	0,0006	2,6559	0,0015	<b>1,595</b>	<b>2,66</b>
10	5	1,5299	0,0005	2,3549	0,0010	<b>1,53</b>	<b>2,35</b>
10	6	1,4849	0,0004	2,1712	0,0009	<b>1,485</b>	<b>2,17</b>
10	8	1,4229	0,0003	1,9501	0,0006	<b>1,42</b>	<b>1,95</b>
10	10	1,3799	0,0003	1,8178	0,0005	<b>1,38</b>	<b>1,82</b>
10	12	1,3440	0,0003	1,7195	0,0004	<b>1,34</b>	<b>1,72</b>
10	16	1,2884	0,0002	1,5860	0,0003	<b>1,29</b>	<b>1,59</b>
10	20	1,2525	0,0002	1,5027	0,0003	<b>1,25</b>	<b>1,50</b>
10	25	1,2218	0,0001	1,4334	0,0002	<b>1,22</b>	<b>1,43</b>
13	2	2,0547	0,0011	5,3939	0,0056	<b>2,05</b>	<b>5,4</b>
13	3	1,7855	0,0009	3,1969	0,0024	<b>1,79</b>	<b>3,2</b>
13	4	1,6653	0,0007	2,6106	0,0014	<b>1,67</b>	<b>2,61</b>
13	5	1,5892	0,0006	2,3203	0,0010	<b>1,59</b>	<b>2,32</b>
13	6	1,5362	0,0004	2,1441	0,0008	<b>1,54</b>	<b>2,14</b>
13	8	1,4650	0,0004	1,9303	0,0007	<b>1,465</b>	<b>1,93</b>



<i>n</i>	<i>r</i>	Results				Proposals	
		Limit-	2 <i>u</i>	Limit+	2 <i>u</i>	Limit-	Limit+
13	10	1,4162	0,0003	1,8010	0,0005	1,42	1,80
13	12	1,3760	0,0002	1,7064	0,0004	1,38	1,71
13	16	1,3159	0,0002	1,5760	0,0003	1,32	1,58
13	20	1,2763	0,0002	1,4954	0,0003	1,28	1,50
13	25	1,2423	0,0002	1,4280	0,0002	1,24	1,43
16	2	2,1658	0,0011	5,0829	0,0045	2,16	5,1
16	3	1,8509	0,0009	3,1059	0,0022	1,85	3,105
16	4	1,7140	0,0006	2,5581	0,0014	1,71	2,56
16	5	1,6290	0,0006	2,2826	0,0010	1,63	2,28
16	6	1,5711	0,0005	2,1139	0,0008	1,57	2,11
16	8	1,4925	0,0004	1,9092	0,0006	1,49	1,91
16	10	1,4399	0,0003	1,7844	0,0005	1,44	1,78
16	12	1,3971	0,0003	1,6924	0,0004	1,40	1,69
16	16	1,3338	0,0002	1,5667	0,0003	1,33	1,57
16	20	1,2923	0,0002	1,4885	0,0002	1,29	1,49
16	25	1,2563	0,0002	1,4229	0,0002	1,26	1,42
20	2	2,2726	0,0012	4,8840	0,0046	2,27	4,9
20	3	1,9119	0,0010	3,0377	0,0022	1,91	3,04
20	4	1,7601	0,0007	2,5151	0,0013	1,76	2,515
20	5	1,6670	0,0005	2,2514	0,0009	1,67	2,25
20	6	1,6040	0,0005	2,0880	0,0008	1,60	2,09
20	8	1,5195	0,0004	1,8916	0,0006	1,52	1,89
20	10	1,4621	0,0003	1,7693	0,0005	1,46	1,77
20	12	1,4170	0,0003	1,6808	0,0004	1,42	1,68
20	16	1,3503	0,0002	1,5582	0,0003	1,35	1,56
20	20	1,3070	0,0002	1,4816	0,0002	1,31	1,48
20	25	1,2693	0,0002	1,4177	0,0002	1,27	1,42
25	2	2,3670	0,0012	4,7362	0,0042	2,37	4,735
25	3	1,9666	0,0013	2,9788	0,0030	1,97	2,98
25	4	1,8008	0,0010	2,4798	0,0019	1,80	2,48
25	5	1,7000	0,0007	2,2244	0,0013	1,70	2,22
25	6	1,6330	0,0008	2,0675	0,0011	1,63	2,07
25	8	1,5419	0,0006	1,8753	0,0007	1,54	1,875
25	10	1,4822	0,0005	1,7572	0,0007	1,48	1,76
25	12	1,4344	0,0004	1,6699	0,0005	1,43	1,67
25	16	1,3649	0,0003	1,5505	0,0004	1,375	1,55
25	20	1,3195	0,0003	1,4754	0,0003	1,32	1,48
25	25	1,2805	0,0002	1,4130	0,0003	1,28	1,41
32	2	2,4681	0,0020	4,5297	0,0054	2,47	4,5
32	3	2,0223	0,0017	2,9112	0,0028	2,02	2,91
32	4	1,8406	0,0011	2,4372	0,0018	1,84	2,44
32	5	1,7335	0,0009	2,1944	0,0014	1,73	2,19
32	6	1,6605	0,0007	2,0451	0,0012	1,66	2,045
32	8	1,5641	0,0006	1,8579	0,0008	1,56	1,86
32	10	1,5008	0,0004	1,7423	0,0006	1,50	1,74
32	12	1,4509	0,0004	1,6582	0,0005	1,45	1,66
32	16	1,3783	0,0003	1,5417	0,0005	1,38	1,54
32	20	1,3315	0,0003	1,4692	0,0003	1,33	1,47
32	25	1,2913	0,0003	1,4082	0,0003	1,29	1,41
40	2	2,5508	0,0020	4,3943	0,0056	2,55	4,4
40	3	2,0653	0,0017	2,8600	0,0029	2,065	2,86

<i>n</i>	<i>r</i>	Results				Proposals	
		Limit-	2 <i>u</i>	Limit+	2 <i>u</i>	Limit-	Limit+
40	4	1,8736	0,0010	2,4060	0,0017	1,87	2,41
40	5	1,7594	0,0009	2,1700	0,0014	1,76	2,17
40	6	1,6825	0,0008	2,0245	0,0010	1,68	2,03
40	8	1,5820	0,0005	1,8436	0,0008	1,58	1,84
40	10	1,5152	0,0005	1,7317	0,0006	1,52	1,73
40	12	1,4631	0,0005	1,6491	0,0005	1,46	1,65
40	16	1,3891	0,0003	1,5349	0,0004	1,39	1,535
40	20	1,3409	0,0003	1,4634	0,0003	1,34	1,46
40	25	1,2994	0,0002	1,4037	0,0003	1,30	1,40
50	2	2,6262	0,0023	4,2680	0,0055	2,63	4,3
50	3	2,1049	0,0018	2,8153	0,0027	2,11	2,815
50	4	1,9014	0,0012	2,3780	0,0015	1,90	2,38
50	5	1,7817	0,0008	2,1497	0,0013	1,78	2,15
50	6	1,7018	0,0007	2,0075	0,0011	1,70	2,01
50	8	1,5969	0,0005	1,8307	0,0009	1,60	1,83
50	10	1,5284	0,0005	1,7213	0,0007	1,53	1,72
50	12	1,4750	0,0004	1,6400	0,0005	1,48	1,64
50	16	1,3986	0,0003	1,5285	0,0004	1,40	1,53
50	20	1,3490	0,0003	1,4584	0,0003	1,35	1,46
50	25	1,3066	0,0002	1,3998	0,0003	1,31	1,40
63	2	2,6975	0,0023	4,1640	0,0050	2,70	4,2
63	3	2,1410	0,0017	2,7747	0,0025	2,14	2,78
63	4	1,9281	0,0012	2,3530	0,0017	1,93	2,35
63	5	1,8030	0,0009	2,1306	0,0013	1,80	2,13
63	6	1,7197	0,0007	1,9915	0,0009	1,72	1,99
63	8	1,6110	0,0006	1,8193	0,0008	1,61	1,82
63	10	1,5403	0,0005	1,7112	0,0007	1,54	1,71
63	12	1,4854	0,0004	1,6329	0,0005	1,49	1,63
63	16	1,4067	0,0003	1,5226	0,0004	1,41	1,52
63	20	1,3563	0,0003	1,4540	0,0004	1,36	1,45
63	25	1,3131	0,0003	1,3956	0,0003	1,31	1,40
80	2	2,7648	0,0022	4,0659	0,0048	2,77	4,1
80	3	2,1736	0,0016	2,7355	0,0026	2,18	2,74
80	4	1,9517	0,0012	2,3280	0,0015	1,95	2,33
80	5	1,8222	0,0010	2,1123	0,0012	1,82	2,11
80	6	1,7363	0,0009	1,9775	0,0009	1,74	1,98
80	8	1,6237	0,0006	1,8077	0,0007	1,62	1,81
80	10	1,5511	0,0005	1,7033	0,0007	1,55	1,70
80	12	1,4943	0,0005	1,6251	0,0005	1,49	1,625
80	16	1,4143	0,0004	1,5169	0,0004	1,41	1,52
80	20	1,3629	0,0003	1,4496	0,0003	1,36	1,45
80	25	1,3188	0,0002	1,3921	0,0003	1,32	1,39
100	2	2,8187	0,0026	3,9818	0,0045	2,82	4,0
100	3	2,2036	0,0025	2,7046	0,0028	2,21	2,705
100	4	1,9728	0,0018	2,3073	0,0027	1,97	2,31
100	5	1,8373	0,0015	2,0966	0,0015	1,84	2,10
100	6	1,7495	0,0014	1,9652	0,0014	1,75	1,965
100	8	1,6345	0,0010	1,7997	0,0011	1,635	1,80
100	10	1,5597	0,0008	1,6961	0,0008	1,56	1,70
100	12	1,5021	0,0005	1,6183	0,0007	1,50	1,62
100	16	1,4203	0,0005	1,5123	0,0005	1,42	1,51

<i>n</i>	<i>r</i>	Results				Proposals	
		Limit-	2 <i>u</i>	Limit+	2 <i>u</i>	Limit-	Limit+
100	20	1,3683	0,0004	1,4456	0,0004	<b>1,37</b>	<b>1,45</b>
100	25	1,3236	0,0003	1,3890	0,0004	<b>1,32</b>	<b>1,39</b>
125	2	2,8739	0,0043	3,9109	0,0064	<b>2,88</b>	<b>3,9</b>
125	3	2,2272	0,0024	2,6754	0,0042	<b>2,23</b>	<b>2,675</b>
125	4	1,9889	0,0017	2,2928	0,0026	<b>1,99</b>	<b>2,29</b>
125	5	1,8521	0,0011	2,0826	0,0016	<b>1,85</b>	<b>2,08</b>
125	6	1,7616	0,0012	1,9551	0,0013	<b>1,76</b>	<b>1,955</b>
125	8	1,6443	0,0008	1,7897	0,0011	<b>1,64</b>	<b>1,79</b>
125	10	1,5677	0,0007	1,6900	0,0008	<b>1,57</b>	<b>1,69</b>
125	12	1,5083	0,0005	1,6130	0,0007	<b>1,51</b>	<b>1,61</b>
125	16	1,4259	0,0005	1,5076	0,0006	<b>1,43</b>	<b>1,51</b>
125	20	1,3730	0,0004	1,4421	0,0005	<b>1,37</b>	<b>1,44</b>
125	25	1,3278	0,0003	1,3862	0,0004	<b>1,33</b>	<b>1,39</b>
125	2	2,8699	0,0034	3,9060	0,0059	<b>2,87</b>	<b>3,9</b>
125	3	2,2282	0,0022	2,6734	0,0037	<b>2,23</b>	<b>2,67</b>
125	4	1,9885	0,0017	2,2913	0,0022	<b>1,99</b>	<b>2,29</b>
125	5	1,8515	0,0012	2,0828	0,0017	<b>1,85</b>	<b>2,08</b>
125	6	1,7613	0,0012	1,9542	0,0013	<b>1,76</b>	<b>1,95</b>
125	8	1,6440	0,0009	1,7914	0,0012	<b>1,64</b>	<b>1,79</b>
125	10	1,5669	0,0007	1,6888	0,0008	<b>1,57</b>	<b>1,69</b>
125	12	1,5087	0,0006	1,6134	0,0008	<b>1,51</b>	<b>1,61</b>
125	16	1,4265	0,0005	1,5083	0,0006	<b>1,43</b>	<b>1,51</b>
125	20	1,3730	0,0004	1,4421	0,0005	<b>1,37</b>	<b>1,44</b>
125	25	1,3279	0,0004	1,3863	0,0004	<b>1,33</b>	<b>1,39</b>
160	2	2,9259	0,0031	3,8356	0,0059	<b>2,93</b>	<b>3,84</b>
160	3	2,2510	0,0026	2,6432	0,0033	<b>2,25</b>	<b>2,64</b>
160	4	2,0074	0,0019	2,2718	0,0021	<b>2,01</b>	<b>2,27</b>
160	5	1,8650	0,0014	2,0718	0,0016	<b>1,865</b>	<b>2,07</b>
160	6	1,7739	0,0010	1,9435	0,0013	<b>1,77</b>	<b>1,94</b>

<i>n</i>	<i>r</i>	Results				Proposals	
		Limit-	2 <i>u</i>	Limit+	2 <i>u</i>	Limit-	Limit+
160	8	1,6523	0,0009	1,7827	0,0010	<b>1,65</b>	<b>1,78</b>
160	10	1,5755	0,0007	1,6823	0,0007	<b>1,58</b>	<b>1,68</b>
160	12	1,5151	0,0005	1,6074	0,0008	<b>1,52</b>	<b>1,61</b>
160	16	1,4313	0,0005	1,5036	0,0005	<b>1,43</b>	<b>1,50</b>
160	20	1,3775	0,0004	1,4383	0,0005	<b>1,38</b>	<b>1,44</b>
160	25	1,3313	0,0004	1,3833	0,0004	<b>1,33</b>	<b>1,38</b>
200	2	2,9669	0,0043	3,7847	0,0060	<b>2,97</b>	<b>3,78</b>
200	3	2,2736	0,0026	2,6274	0,0033	<b>2,27</b>	<b>2,63</b>
200	4	2,0217	0,0019	2,2572	0,0021	<b>2,02</b>	<b>2,26</b>
200	5	1,8778	0,0011	2,0608	0,0017	<b>1,88</b>	<b>2,06</b>
200	6	1,7823	0,0010	1,9341	0,0013	<b>1,78</b>	<b>1,93</b>
200	8	1,6607	0,0007	1,7761	0,0010	<b>1,66</b>	<b>1,78</b>
200	10	1,5810	0,0008	1,6763	0,0008	<b>1,58</b>	<b>1,68</b>
200	12	1,5207	0,0006	1,6036	0,0007	<b>1,52</b>	<b>1,60</b>
200	16	1,4358	0,0004	1,5002	0,0006	<b>1,44</b>	<b>1,50</b>
200	20	1,3812	0,0005	1,4356	0,0005	<b>1,38</b>	<b>1,44</b>
200	25	1,3345	0,0004	1,3806	0,0004	<b>1,33</b>	<b>1,38</b>
250	2	3,0066	0,0039	3,7339	0,0054	<b>3,01</b>	<b>3,73</b>
250	3	2,2910	0,0027	2,6066	0,0028	<b>2,29</b>	<b>2,61</b>
250	4	2,0331	0,0018	2,2460	0,0020	<b>2,03</b>	<b>2,25</b>
250	5	1,8867	0,0018	2,0508	0,0018	<b>1,89</b>	<b>2,05</b>
250	6	1,7902	0,0009	1,9272	0,0012	<b>1,79</b>	<b>1,93</b>
250	8	1,6670	0,0008	1,7710	0,0010	<b>1,67</b>	<b>1,77</b>
250	10	1,5865	0,0009	1,6727	0,0007	<b>1,59</b>	<b>1,67</b>
250	12	1,5253	0,0008	1,5985	0,0007	<b>1,525</b>	<b>1,60</b>
250	16	1,4394	0,0005	1,4970	0,0005	<b>1,44</b>	<b>1,50</b>
250	20	1,3847	0,0005	1,4332	0,0004	<b>1,39</b>	<b>1,43</b>
250	25	1,3376	0,0004	1,3788	0,0004	<b>1,34</b>	<b>1,38</b>

These results are displayed in Figure 4, that represents:

-  The gaps between the lower limits and the nominal values of Table 3 (in blue);
-  The gaps between the upper limits and the nominal values of Table 3 (in orange);

for alerts corresponding nominal risks  $\alpha$  and  $\beta$  equal to 0,5% and  $2 \leq r \leq 25$  as function of  $\log(r)$ , for  $n = 3 - 4 - 5 - 6 - 8 - 10 - 13 - 16 - 20 - 25 - 32 - 40 - 50 - 63 - 80 - 100 - 125 - 160 - 200$  and 250. Each line represents one value of  $n$ . Lower and upper limits become increasingly closer when  $n$  increases (i.e. outer curves correspond to  $n = 3$  while curves closest to 0 correspond to  $n = 250$ ).



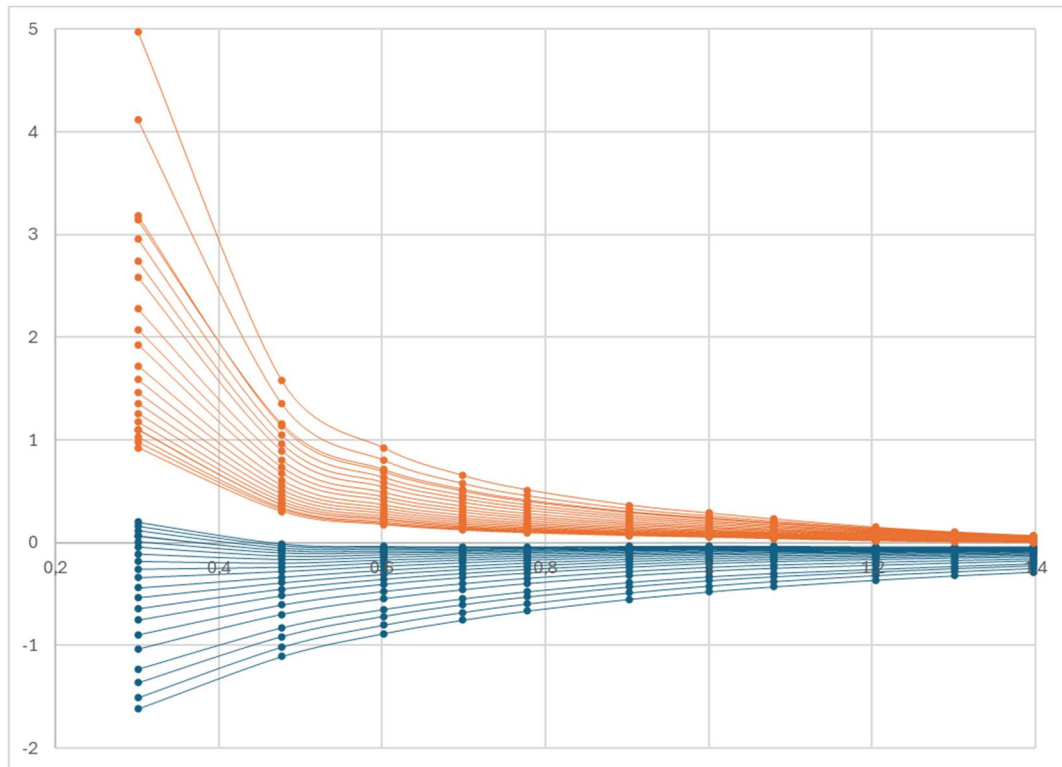


Figure 4: Gaps between lower (in blue) and upper (in orange) limits and nominal values for alerts corresponding nominal risks  $\alpha$  and  $\theta$  equal to 0,5% and  $2 \leq r \leq 25$  as function of  $\log(r)$ , for  $n = 3 - 4 - 5 - 6 - 8 - 10 - 13 - 16 - 20 - 25 - 32 - 40 - 50 - 63 - 80 - 100 - 125 - 160 - 200 - 250$ . (Each line represents one value of  $n$ , lower and upper limits become closer and closer when  $n$  increases)

It can be seen from this figure that, for large values of  $n$  and low values of  $r$ , the IC90% does not include the nominal value of ZR (some dots of some blue curves lie above 0). For example, for  $n = 250$  and  $r = 2$ , the nominal value of ZR is 2,807 while the IC90% is [2,97;3,78]. This means that, in those cases, the assessment of repeatability using our proposal of [6] is probably slightly too severe. In those cases, a proposal not based on the hypothesis of homoscedasticity could also make sense, alternatively to the present proposal. However, the main concern remains the choice of the level of the nominal risk, which is always conventional (see § 1) and which leads to quite higher differences in the determination of limits (for example, choosing a nominal risk of 0,135% corresponding to the usual  $z = 3$  limit for bias, would lead to a nominal ZR limit of 3,205 instead of 2,807).

Outside these extreme cases, the curves of limits envelop quite well the nominal ZR values, represented by the line of ordinate "0" in Figure 4.

### 5.3 Conclusions

The limits that were determined in this study make sense for the assessment of repeatability.

## 6 Results for assessments using non-parametric methods

### 6.1 Introduction

Most cases of PT can be dealt with using parametric methods, because in most cases, a distribution law can be reasonably assumed to represent adequately the distribution of test results (typically the Gaussian distribution for biases and a derivation of the  $\chi^2$  law for SD). When test results obviously do not follow a gaussian distribution, a log



transformation or (for proportions) a transformation  $\log(p/(1-p))$  (where  $p$  is the proportion) usually makes the transformed results follow a Gaussian distribution.

However, it might happen cases where such transformations are uneasy because of 0 values or where finding an adequate transformation of results is not obvious, see for example a hypothetical untypical distribution of test results in Figure 5.

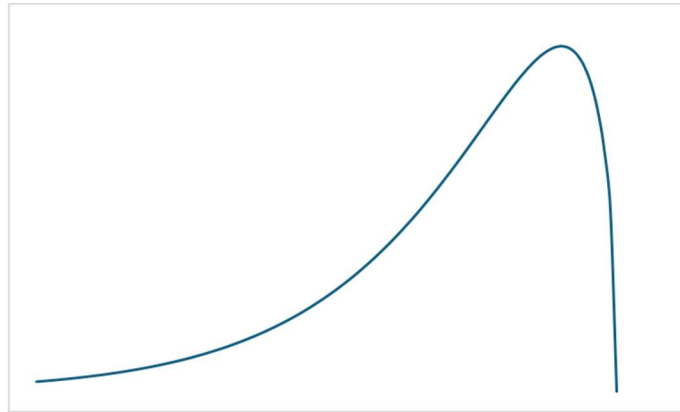


Figure 5: Example of untypical distribution of test results.

Such situations may come from particularities of the test method. For example, a law of distribution as shown in Figure 5 may appear when deviations to the test method are more likely to produce too small test results than too large test results. The PT provider should consider this for deciding how to determine adequate alert limits.

There also are situations where no law of distribution exists at all, for example for tests with results expressed in terms of categories, even when these categories are expressed as numbers (i.e., for example, category 1 to 5), because in those cases, these numbers shall be regarded as names of category rather than numerical values (see warning of ISO 13528 [2]).

In such cases, non-parametric methods usually based on ranks may be considered. However, ISO 13528 [2] states that methods that declare a fixed number of outliers should not be used, which implicitly makes all non-parametric methods not recommended. As a matter of fact, non-parametric methods usually use Binomial distributions to determine the probabilities attached to a ranked value. Applying this to the extreme values enables to determine their probabilities to happen and consequently, whether they should be regarded as outliers or not. A rank can then be associated to the nominal risk and to each side of an IC90%.

Obviously, we cannot expect the non-parametric methods to be as efficient as the parametric ones, so that they should be used only when other methods cannot be applied.

In our study, we considered 2 basics to compute alerts as exposed in § 3.5:

- ✚ The first one considers that outliers are those located at the tails of the distribution (option usually selected for the assessment of bias, see § 4.1);
- ✚ The second one considers that all participants show the same value (option usually selected for the assessment of SD, see § 5.1). If used for bias, this option would lead to declare most of the participants as outliers (see § 3.5).

In the first basics, we then used  $\alpha$  and  $\beta$  risks of 1% bilateral (i.e. 0,5% on each side of the distribution). In such conditions, the Binomial law can be approximated with the Poisson law, with  $p = 0,005$  and  $n$  taken equal to the number of participants ( $n.p$  is usually less than 1). The calculations consist then in computing ranks for which the

results of the cumulative Poisson law are less than 5% or more than 95% to determine the corresponding IC90%, as function of  $n$ .

In the second basics, we considered that all participants are supposed to have the same value, i.e. each of them has a probability equal to 0,5 to be higher or lower than the reference value. The calculations consist then in computing ranks for which the results of the cumulative Binomial law are less than 5% or more than 95% to determine the corresponding IC90%, as function of  $n$ .

In both cases, no use of the MCM is needed.

Obviously, the same number of participants get signals at each tail of the distribution, so that the critical  $n$  values for alerts actually trigger a signal for 2 more participants: one at the lower tail of the distribution and one at the upper tail of the distribution.

Another tricky situation that can be encountered using this method is when, due to rounding, same test results should be regarded as outlying and not outlying. For example, if the 5 largest results are equal and 2 of them should be regarded as outlying, what should be decided for these 5 equal largest results? An option is to compare the number of equal results and the number of signals to declare among these results to decide whether all or neither of them should trigger a signal. For example, in the here upper case, compare 2 “outlying” results and 3 “not outlying results” and then decide that the 5 of them are not outlying. In case of balance (for example when 2 “outlying” results and 2 “not outlying results” need to be compared), it should be decided to consider all values as not outlying because of the rules of rounding that were used to compute the values of Table 5, Table 6 and Table 7.

## 6.2 Results for the first basics

Table 5 provides the results of the determination of the number of participants at each tail of the distribution that should receive a signal of alert or a signal of action, corresponding nominal risks  $\alpha$  and  $\beta$  to belong to the 1% (bilateral) tails of the distribution.

*Table 5. Results of determination of the number of participants at each tail of the distribution that should receive a signal of alert or a signal of action, corresponding nominal risks  $\alpha$  and  $\beta$  to belong to the 1% (bilateral) tails of the distribution.*

Signals	$2 \leq n \leq 10$	$11 \leq n \leq 71$	$72 \leq n \leq 163$	$164 \leq n \leq 273$	$274 \leq n \leq 394$	$395 \leq n \leq 460$	$461 \leq n \leq 522$	$523 \leq n \leq 657$
Alert	0	1	2	3	4	5	5	6
Action	0	0	0	0	0	0	1	1

For example, in a PT involving 100 participants, the 2 lowest and the 2 largest test results get a signal of alert, and no signal of action is triggered.

Unsurprisingly, we can see that this method is not very efficient, we need 461 participants to be able to trigger one signal of action. Then, it can never be used in practice.

## 6.3 Results for the second basics

Table 6 provides the results of the determination of the number of participants at each tail of the distribution that should receive a signal of alert or a signal of action, corresponding nominal risks  $\alpha$  and  $\beta$  equal to 1% (bilateral) to be different from the reference value.

Table 6. Results of determination of the number of participants at each tail of the distribution that should receive a signal of alert or a signal of action, corresponding nominal risks  $\alpha$  and  $\beta$  equal to 1% (bilateral) to be different from the reference value.

	$2 \leq n \leq 4$	$5 \leq n \leq 7$	$8 \leq n \leq 10$	$11 \leq n \leq 12$	$13 \leq n \leq 15$	$16 \leq n \leq 17$	$18 \leq n \leq 20$	$21 \leq n \leq 22$
<b>Alerts</b>	All	All	All	All	All	All	All	All
<b>Actions</b>	0	1	2	3	4	5	6	7

Only results for  $n \leq 22$  are provided in this table, but we could verify that all participants get an alert for all values of  $n$  usually encountered during PT performances. This is then a method that can never been used in practice.

## 6.4 Adaptation of the levels of risk and of IC to the cases where non-parametric methods need to be used

To cope with the situations described in § 6.2 and § 6.3, we determined another series of limits using nominal risks  $\alpha$  and  $\beta$  to belong to the 10% (bilateral) tails of the distribution (instead of 1%) and an IC80% (instead of 90%). Table 7 provides the results of these determinations.

Table 7. Results of determination of the number of participants at each tail of the distribution that should receive a signal of alert or a signal of action, corresponding nominal risks  $\alpha$  and  $\beta$  to belong to the 10% (bilateral) tails of the distribution.

	Alert	Action		Alert	Action		Alert	Action
$n \leq 2$	0	0	$64 \leq n \leq 77$	6	1	$141 \leq n \leq 156$	11	4
$3 \leq n \leq 10$	1	0	$78 \leq n \leq 93$	7	2	$157 \leq n \leq 159$	12	4
$11 \leq n \leq 22$	2	0	$94 \leq n \leq 106$	8	2	$160 \leq n \leq 172$	12	5
$23 \leq n \leq 34$	3	0	$107 \leq n \leq 108$	8	3	$173 \leq n \leq 185$	13	5
$35 \leq n \leq 46$	4	0	$109 \leq n \leq 124$	9	3	$186 \leq n \leq 206$	13	6
$47 \leq n \leq 48$	4	1	$125 \leq n \leq 133$	10	3	$207 \leq n \leq 210$	14	6
$49 \leq n \leq 63$	5	1	$134 \leq n \leq 140$	10	4			

These results show that this method can be used when the number of participants is large enough. For low number of participants, a large amount of signals of alert but no signals of action are triggered, reflecting the lack of power of the method that can hardly decide whether a result is correct or not.

## 6.5 Use of this method for non-numerical test results

Non numerical test results (i.e. test results expressed in terms of categories) can be classified into 3 types:

1. Categories that are ordered (for example sweetness of wine);
2. Categories that cannot be ordered (with respect to the property that it is supposed to represent), an example is provided in § E15 of ISO 13528 [2];
3. Binary results (for example Pass/fail), that can be regarded as type 1 with only 2 categories.

For the type 2, ISO 13528 [2] proposes to build up a ranking by frequency of occurrence in the test results (for example, a series of test results like A: 5 – B: 13 – C: 3 – D: 27, the categories can then be ranked as follows: D – B – A – C).

The method exposed in § 6.4 can then be applied to the ordered test results. Obviously, unilateral checks should be used for types 2 and 3, because it does not make sense to consider the most frequent test results as outliers. Figures of § 6.4 can then be used on only one side of the distribution, with associated  $\alpha$  and  $\beta$  risks of 5% instead of 10%.

For example, for the not ordered series of tests mentioned here upper (i.e. A: 5 – B: 13 – C: 3 – D: 27), the following procedure should apply:

1. The total number of test results is 48, that should trigger 4 signals of alert and 1 signal of action (see Table 7);
2. The signal of action should be granted to one “C” test result;
3. However, there are 3 “C” test results. With regard to the statements of § 6.1 (i.e. 1 is less than or equal to the half of 3), no signal of action should be triggered. The signal of action shall then be converted into a signal of alert;
4. Then, a total of  $4 + 1 = 5$  signals of alert should be triggered;
5. Then, the 3 “C” test results trigger a signal of alert;
6. Then, it remains 2 signals of alert to grant to “A” test results, that are following “C” in the ranking;
7. With regard to the statements of § 6.1 (i.e. 2 is less than or equal to the half of 5), no signal of alert should be triggered to “A” test results;
8. As a global conclusion, “A”, “B” and “D” test results can be accepted while “C” test results trigger a signal of alert.

## 6.6 Conclusions for limits determined with non-parametric methods

Non-parametric methods can also be used to determine limits for alerts, that are then expressed in terms of ranks rather than in terms of scores.

Unsurprisingly, these methods are less efficient and powerful than parametric ones. That is why they are not recommended by the reference standards. They should be used only when parametric methods cannot be used, because no law of distribution can be reasonably assumed or even exists, provided that an enough number of participations is available.

## 7 Conclusions

We could determine limits for alerts as function of the number of participations for the assessment of bias as well as for the assessment of repeatability.

For bias, the classical limits 2 and 3 for z-scores are not far from the situation where  $\alpha$  and  $\beta$  risks equal to 1% bilateral with an IC90% when  $n = 110$ . Lower values of  $n$  request a larger “band of doubt” around the nominal value 2,576 while larger values of  $n$  enable a lower one. Empirical formulas could be determined for the computation of the limits, as function of the number of participants. These limits apply to the apparent value of  $\sigma_{pt}$ , that is usually greater than  $\sigma_L$ , what reduces the  $\alpha$  risk and increases the  $\beta$  risk. When a  $\lambda$  ratio defined as  $\lambda = \sigma_r / (r \cdot \sigma_L)$  is less than 0,17, this effect becomes negligible and no difference between  $\alpha$  and  $\beta$  risk then applies.

For repeatability assessments, for large values of  $n$  and low values of  $r$ , the IC90% does not include the nominal value of ZR. This means that, in those cases, the assessment of repeatability using the proposal of [6] is probably slightly too severe. In those cases, a proposal not based on the hypothesis of homoscedasticity could also make sense, alternatively to the present proposal. However, the main concern remains the choice of the level of the nominal risk, which is always conventional (see § 1) and that leads to quite higher differences in the determination of limits. Outside these extreme cases, the determined limits envelop quite well the nominal ZR values.

Non-parametric methods can also be used to determine limits for alerts, that are then expressed in terms of ranks rather than in terms of scores. Unsurprisingly, these methods are less efficient and powerful than parametric ones.

That is why they should be used only when parametric methods cannot be used, because no law of distribution can be reasonably assumed or even exists, provided that an enough number of participations is available.

## 8 References

- [1] 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
- [2] ISO 13528:2022, Statistical methods for use in proficiency testing by interlaboratory comparison
- [3] ISO/IEC 17043:2023, Conformity assessment — General requirements for the competence of proficiency testing providers
- [4] L.J. Hollebecq, “Beta risk in proficiency testing in relation with the number of participants”, CompaLab technical publications, December 2022,  
DOI: <https://www.compalab.org/medias/files/publication-interne-risque-beta-en.pdf>
- [5] L.J. Hollebecq, “Beta risk in proficiency testing in relation with the number of participants”, ACTA IMEKO, Vol. 12, Nr 3, 1-9, September 2023  
DOI: <https://doi.org/10.21014/actaimeko.v12i3.1433>
- [6] L.J. Hollebecq, “Proficiency testing of repeatability”, CompaLab technical publications, July 2025,  
DOI: <https://www.compalab.org/medias/files/publication-ea-et-en.pdf>
- [7] 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>
- [8] L.J. Hollebecq, “Intervals of confidence on nested standard deviations”, CompaLab technical publications, January 2025,  
DOI: <https://www.compalab.org/medias/files/publication-interne-estimation-et-composes-en.pdf>