I N S T I T U T D E

S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N

P A P E R

0611

RISK MANAGEMENT FOR ANALYTICAL METHODS BASED ON THE TOTAL ERROR CONCEPT: CONCILIATING THE OBJECTIVES OF THE PRE-STUDY AND IN-STUDY VALIDATION PHASES

B. BOULANGER, W. DEWE, A. GILBERT, B. GOVAERTS and M. MAUMY

http://www.stat.ucl.ac.be

Risk management for analytical methods based on the total error concept: conciliating the objectives of the pre-study and in-study validation phases.

Bruno Boulanger[a], Walthère Dewé[a], Aurélie Gilbert[b],
Bernadette Govaerts[b] and Myriam Maumy[c]

[a]*Eli Lilly, European Early Phase Statistics, Belgium.*
[b] *Université Catholique de Louvain, Institut de Statistique, Belgium.*
[c]*Université Louis Pasteur, Laboratoire de Statistique, France.*

SUMMARY

In industries that involve either chemistry or biology, analytical methods are necessary to keep an eye on all the material produced. If the quality of an analytical method is doubtful, then the whole set of decisions based on those measures is questionable. For this reason, being able to assess the quality of an analytical method is far more than a statistical challenge; it is a matter of ethics and good business practices.

The validity of an analytical method must be assessed at two levels. The "pre-study" validation aims to show, by an appropriate set of designed experiments, that the method is able to achieve its objectives. The "in-study" validation is intended to verify, by inserting QC samples in routine runs, that the method remains valid over time. At these two levels, the total error approach considers a method as valid if a sufficient proportion of analytical results are expected to lie in a given interval around the (unknown) nominal value.

This paper discusses two methods, based on this total error concept, of checking the validity of a measurement method at the pre-study level. The first checks whether a tolerance interval for hypothetical future measurements lies within given acceptance limits; the second calculates the probability of a result lying within these limits and computes whether it is greater than a given acceptance level. For the "in-study" validation, the paper assesses the properties of the s–n–λ rule recommended by the FDA. The properties and respective advantages and limitations of these methods are investigated.

A crucial point is to ensure that the decisions taken at the pre-study stage and in routine use are coherent. More precisely, a laboratory should not see its method rejected in routine use when it has been proved to be valid and remains so. This paper shows how this goal may be achieved by choosing compatible validation parameters at both pre- and in-study levels.

Correspondance to:
Bernadette Govaerts, Institut de Statistique, 20 voie du roman pays, 1348 Louvain-la-Neuve, Belgium, Govaerts@stat.ucl.ac.be, Phone: +32-10-47.43.13.

1. INTRODUCTION

In industries that involve either chemistry or biology, such as the pharmaceutical, chemical and food industries, analytical methods are the eyes and ears for all the material produced and used. If the quality of an analytical method is doubtful, the decisions based on the measures obtained with this procedure may become questionable and possibly even the final product. For this reason, being able to assess the quality of an analytical method is far more than a statistical challenge; it's a matter of ethics and good business practices. Many regulatory organisations have addressed this issue in the chemical and pharmaceutical industry (e.g. the ICH (International Conference on the Harmonisation), the FDA (Food and Drug Administration) or EuraChem [1, 2, 3]).

The objective of validation is to give both the laboratory and the regulatory bodies a guarantee that every single routine measurement that will be performed will yield results close enough to the unknown "true" value of the sample [4]. The conformity of a given analytical method to this objective is usually assessed in two stages [5, 6, 7, 8]. First, a "pre-study" phase is conducted to prove, on the basis of a designed experiment, that the method can deliver quality results. Then, at a routine level, the laboratory must verify that the analytical method remains valid over time, and that each run provides trustworthy measures. This is usually achieved by inserting quality control (QC) samples in the unknown sample runs.

At these two stages, it is essential to have a way of quantifying the quality of a measure in terms of its closeness to the "true" value of the property of interest. Traditionally this is achieved by examining two main performance criteria of an analytical method: the bias or "trueness", and the precision of the method. Both should be small, and they are usually quantified separately [3, 9, 10, 11]. This approach focuses on the method itself, assuming that if the method is "good" then the measures it provides will also be "good". However this is not always the case [12]. The concept of "total error" [8, 13, 14, 15, 16, 17] puts the emphasis on the results themselves and tackles the problem globally by estimating the proportion π of the measurements which are expected to lie within a fixed interval $(+/- \lambda)$ around the true value. The assumption underlying this approach is that, if the results produced are "good", then the method that produces them is necessarily "good". This paper presents procedures to check the pre- and in-study validity of an analytical method based on this total error concept, by examining the quality of the results it produces.

At the pre-study level, the validation procedure consists of measuring a given set of samples for which the nominal values are known and arranged according to an adequately-designed experiment. The design should enable measurement bias and precision to be estimated for different nominal levels and, if necessary, it should provide a decomposition of the global precision in various components of variances (repeatability, between-run, and between laboratory). Two statistical procedures are discussed to assess the validity of the method on the basis of such an experiment. The first consists of estimating a tolerance interval in which a great proportion of "future" measurements are expected to lie and verifying that this interval is included in predefined acceptance limits. The second estimates the quality level i.e. the probability of getting a measure within these acceptance limits directly, and checks that it is greater than a given minimum acceptance level (on the basis of the lower limit of a maximum likelihood confidence interval for this probability).

In routine use, budgets and simplicity requirements usually lead to the use of validation rules that do not fully protect either the client or the laboratory. This paper studies the properties of the 4–6–15 rule (generalised here as the s–n–λ rule) recommended by the FDA [3] in this context. It consists of inserting a set of n QC samples into routine unknown samples and

checking that at least s of the measurements obtained from these samples are no more distant than λ from their nominal (true) value.

In the practical organisation of an industrial laboratory, pre- and in-study validation studies are often conducted separately by different people, especially if the method has been developed and validated in one place (e.g. a research laboratory) and is then routinely used in another (e.g. production plan). The compatibility of the decisions taken at these two stages is not obvious and not necessarily even well understood by the analysts. A laboratory that has declared the validity of a method in a pre-study phase would not appreciate (economically speaking) seeing its method rejected for use if it is still valid. On the other hand, if a valid method is subject to a significant total error increase, the in-study validation rule should be able to detect it rapidly. Conciliating pre- and in-study objectives is then crucial and may be achieved through an appropriate choice of validation rule parameters to align the associated risks.

This paper is organised as follows: Section 2 gives a precise definition of the validation method based on the concept of total error and introduces related notations. Section 3 introduces two procedures for "pre-study" validation, the β-expectation tolerance interval approach and a maximum likelihood approach aimed at estimating the quality level. Those two procedures are illustrated on a real example and their performances are compared using simulations. Section 4 discusses the properties of the s–n–λ rule in terms of client and laboratory risks. Finally, Section 5 shows how pre- and in-study validation parameters may be conciliated to achieve coherent properties for validation decisions.

2. EVALUATION OF ANALYTICAL METHOD BASED ON TOTAL ERROR

The objective of a good analytical method is to quantify accurately each of the unknown quantities that the laboratory will have to determine. In other words, the analytical method is expected to give results X's for which the difference from the unknown "true" value (μ_T) of the sample is sufficiently small, for example less than a predefined acceptance limit, λ, i.e.

$$-\lambda < X - \mu_T < \lambda \quad \Leftrightarrow \quad \left|X - \mu_T\right| < \lambda .$$

Two components may influence this difference: the bias or trueness of the method, and its precision. As illustrated in Figure 1, a biased method provides results that deviate "in the mean" or systematically from the true value μ_T: $\delta = E(X) - \mu_T = \mu - \mu_T$. The precision expresses how results vary around the mean value $\mu=E(X)$ when the measure is repeated. To quantify this precision, let σ denote the standard deviation available. The "closeness" of a result X to the unknown true value of the sample μ_T is directly linked to the size of the bias δ and precision σ of the method.

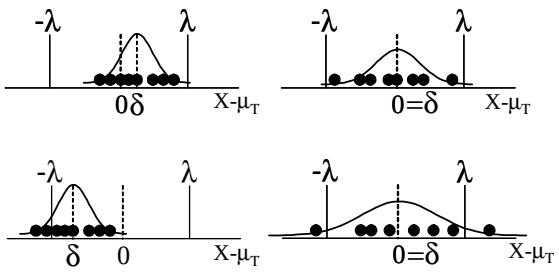

Figure 1: Comparison of four possible validation situations

Classical method validation and quality control tools usually check the size of these two components separately (t and χ^2 tests in validation, or $\overline{X}$ – R control charts in routine use) but this approach has the drawback that a very low value of one component may not compensate for a large value in the other.

The total error approach [4, 8, 13, 14, 15, 16, 17] suggests a global approach in considering a procedure acceptable if it is "very likely" that the difference between each measurement X of a sample and its "true" value (μ_T) is within the acceptance limits [–λ,+λ] predefined by the analyst. The notion of "very likely" can be translated to the probabilistic equation

$$\pi = P\left(\left|X - \mu_T\right| < \lambda\right) \geq \pi_{\min},$$

where $\pi_{\min}$ is called here the acceptance level and π the quality level. The acceptance limit λ can be expressed either in absolute or in relative (%) terms. In the latter case, the equation is redefined as

$$\pi = P\left(\left|\frac{X - \mu_T}{\mu_T}\right| < \lambda\right) \geq \pi_{\min}. \qquad (1)$$

All the results presented in this paper are applicable to both cases. Below, we use the first formulation without loss of generality.

The value of λ must be chosen according to the intended use of the results. The objective is linked to the requirements of the application area of the user (e.g. 1 or 2% on bulk, 5% on pharmaceutical specialties, 15% for biological samples, 30% for ligand-binding assays such as RIA or ELISA, and so on). The probability $\pi_{\min}$ must also be fixed by the analyst, according to cost, consumer and analytical domain requirements. The key is to ensure coherence between the $\pi_{\min}$ and λ values targeted in the pre-study and in-study phases. This issue is discussed in more detail in Section 5.

Under the assumption of normality for the measurement results it is easy to establish the relationship between the quality level π, the bias (systematic error) δ, and the precision (random error) σ as

$$\pi = P\left(\left|X - \mu_T\right| < \lambda\right) = P\left(\frac{-\lambda - \delta}{\sigma} < Z < \frac{\lambda - \delta}{\sigma}\right),$$

where *Z* is a standard normal random variable. This leads to a definition of the "acceptance region", i.e. the set of (δ,σ)'s such that the quality level π is greater than $\pi_{\min}$. Figure 2 shows, below the curves, the acceptance region for various values of $\pi_{\min}$ (99%, 95%, 90%, 80% and 66.7%) when the acceptance limits are fixed at [–15%,+15%] as recommended by FDA [3] for bioanalytical methods. Note that, in this graph, δ and σ must be interpreted as relative bias and relative standard deviation. Logically, as can be seen from Figure 1, the greater the variance of the measure or the greater the bias, the less likely it is that a result will fall within the measurement acceptance limits.

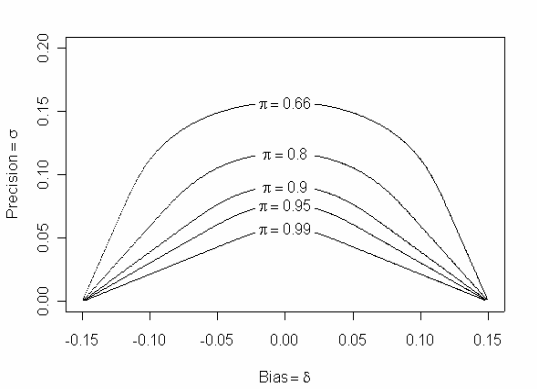

Figure 2: Acceptance region of analytical methods as a function of the method bias and precision when $\lambda = 15\%$.

3. PRE-STUDY METHOD VALIDATION

Before an analytical method is used routinely on unknown samples, it is normal practice to perform a more or less extensive set of experiments to evaluate whether it will be able to meet the criteria described above. Those experiments are usually called "pre-study validation" as opposed to the "in-study validation" experiments.

Since the bias δ and the precision σ, the intrinsic performance parameters of the analytical procedure, are unknown, experiments are required so that the user can obtain estimates of these quantities before using the method routinely. The objective of the pre-study validation phase is then to evaluate whether, given the estimates of the bias $\hat{\delta}$ and standard deviation $\hat{\sigma}$ obtained, the proportion π of measures of new unknown samples that will fall within the acceptance limits is greater than a predefined acceptance level, say π_{min} (see (1)).

However, there is neither an exact solution nor an easy way to answer the question, even for very simple validation experimental designs. Two approximate methods are discussed below. The first is based on the notion of a tolerance interval, and the second consists of calculating the lower limit of a maximum likelihood one-tailed confidence interval for π, using the delta method.

3.1 β-expectation tolerance interval method

The first method has already been introduced in [14, 4]. This paper discusses its properties more formally. It consists of computing the β-expectation tolerance interval [18]

$$E_{\hat{\delta},\hat{\sigma}}\left\{P_X\left(\hat{\delta} - k\hat{\sigma} < X - \mu_T < \hat{\delta} + k\hat{\sigma} \,\middle|\, \hat{\delta},\hat{\sigma}\right)\right\} = \beta \,,$$

where the factor k is determined so that the expected proportion of the population falling within the interval is equal to β. β is defined as the acceptance level π_{min} in this context and the value of k depends on the experimental design used for validation. In the simplest case, where a sample with nominal value μ_T is measured n times in repeatability conditions, $\hat{\delta}, \hat{\sigma}$ and k are calculated from the measurement results $X_1, X_2, \ldots X_n$ as follows, assuming that X is normally distributed:

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu_T = \overline{X} - \mu_T \qquad \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \text{ and } \quad k = t_{n-1;(1+\pi_{\min})/2}\sqrt{1+\frac{1}{n}}$$

where $t_{n-1;(1+\pi_{\min})/2}$ is the quantile $(1+\pi_{\min})/2$ of a (n–1) t distribution.

Mee [19] discussed how to calculate $\hat{\sigma}$ and k in a balanced one-way ANOVA random model where the within- and between-run variabilities are taken into account. Hoffman and Kringle [17] extended this method to more general random effect models for the β-content tolerance intervals (rather than the β-expectation tolerance intervals suggested here).

The decision rule proposed is then: if the β-expectation tolerance interval is within the acceptance limits [–λ,+λ], i.e. if ($\hat{\delta} - k\hat{\sigma} > -\lambda$ and $\hat{\delta} + k\hat{\sigma} < +\lambda$) then there is high strong evidence that the method is valid. As a matter of fact, if this condition is verified, the expected proportion of measurements within the acceptance limits is greater or equal to $\pi_{\min}$, i.e. equation (1) is also verified, on average. Note that the opposite statement is not true, i.e. either $\hat{\delta} - k\hat{\sigma} < -\lambda$ or $\hat{\delta} + k\hat{\sigma} > +\lambda$ does not imply that the expected proportion is smaller than $\pi_{\min}$. This is illustrated on simulations below.

3.2 Maximum likelihood one-tailed confidence interval on π using the delta method

Another approach to validating the method consists of deriving, from $\hat{\delta}$ and $\hat{\sigma}$, the lower bound of a one-tailed confidence interval on the quality level π and checking whether or not it is larger than $\pi_{\min}$. This is not easy as far as the estimation of a probability is concerned. No exact solution exists even in the simple sampling scheme case. However different statistical approaches may be used to attack the problem: searching for a mathematical approximation to the exact solution, asymptotic approximation by maximum likelihood, bootstrap or Bayesian modelling. A maximum likelihood (ML) solution for the simple sampling scheme, under normal-distribution assumption, is presented here.

Let $X_1, X_2, \ldots X_n$ be the measurement results of the validation experiment and suppose that X_i is normally distributed with unknown bias δ and variance σ^2. The maximum likelihood estimators of these parameters are given by

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu_T = \overline{X} - \mu_T \qquad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad = \frac{n-1}{n}\hat{\sigma}^2 = w\hat{\sigma}^2 .$$

By the invariance property, a maximum likelihood estimator of π can be defined as

$$\hat{\pi} = \phi\left(\frac{\lambda - \hat{\delta}}{\tilde{\sigma}}\right) - \phi\left(\frac{-\lambda - \hat{\delta}}{\tilde{\sigma}}\right),$$

where $\phi(.)$ is the distribution function of the standard normal distribution [19]. The delta method can be used to derive an asymptotic approximation for the variance of the estimator $\hat{\pi}$ as

$$Var(\hat{\pi}) \cong \frac{1}{n}(\varphi_L - \varphi_U)^2 + \frac{1}{2nw\sigma^2}\left((-\lambda - \delta)\varphi_L - (\lambda - \delta)\varphi_U\right)^2$$

where φ_L and φ_U are given by

$$\varphi_U = \varphi\left(\frac{\lambda-\delta}{\sigma}\right) \text{ and } \varphi_L = \varphi\left(\frac{-\lambda-\delta}{\sigma}\right)$$

and $\varphi(.)$ is the density of the standard normal distribution. The asymptotic lower bound of a one-tailed 1–α confidence interval on the quality level π can then be calculated as

$$\hat{\pi}_{\text{inf}} = \hat{\pi} - z_{1-\alpha}\hat{\sigma}_{\hat{\pi}}$$

where $z_{1-\alpha}$ is the quantile 1–α of a standard normal variable and $\hat{\sigma}_{\hat{\pi}}$ is calculated by replacing δ by $\hat{\delta}$ and σ by $\tilde{\sigma} = \sqrt{w}\hat{\sigma}$ in $Var(\hat{\pi})$ above. The analytical method is then declared valid if $\hat{\pi}_{\text{inf}} > \pi_{\text{min}}$. This approach can be generalised to more general variance components models.

3.3 Example

To illustrate the methodology described here, pre-study validation data from a bioanalytical procedure [20, 21] are used to illustrate the statistical methods described in Sections 3.1 and 3.2.

The design consisted of three runs with four replications per run at each concentration level. However the run factor was ignored in this illustration so as to be able to use the formulae detailed in Sections 3.1 and 3.2. This results in an overall estimate of the total variance without estimating its within- and between-run components. Therefore the sample size is n = 12 at each level of concentration. A measurement process was considered valid if the proportion of measurements within a range of +/– 15% (λ=0.15) around the target value μ_T was greater than 80% (acceptance level π_{min}=0.8). A 90% one-tailed confidence interval was calculated to get a lower bound for π using the ML method.

The results are summarised in Table 1. The β-expectation tolerance interval limits were all within the acceptance limits, whatever the concentration level. The bioanalytical method of interest can therefore be considered valid over the range investigated using the tolerance interval rule. Also, the 90% lower limits of the one-tailed confidence interval for π estimated by the maximum likelihood method were larger than 0.80 over the range of concentration levels investigated. Thus, the bioanalytical method can also be considered valid according to the maximum likelihood method.

Table 1: Validation results obtained by β-expectation tolerance interval and ML methods

μ_T	$\hat{\delta}$	$\hat{\sigma}$	Tolerance Interval	Acceptance Limits	$\hat{\pi}$	$\hat{\pi}_{\text{inf}}$
25.4	0	1.4	[–2.1, 2]	[–3.8, 3.8]	0.994	0.982
48.2	–2.7	2.7	[–6.5, 1.2]	[–7.2,7.2]	0.959	0.908
437.8	–9.3	19.7	[–37.2, 18.6]	[–65.7, 65.7]	0.999	0.995
838.6	11.8	45	[–52, 75.7]	[–125.8,125.8]	0.995	0.984

3.4 Validation method comparison on the basis of simulations

This section compares the two pre-study validation procedures on the basis of simulations. Four valid and two non valid hypothetical measurement processes or analytical methods were chosen, as shown in Table 2 and Figure 3, and normally distributed samples of sizes ranging from 5 to 200, were randomly generated. The acceptance limit λ was fixed at 0.15,

i.e. [–15%, +15%], the acceptance level π_{min} at 80%, and the confidence level for the ML one-tailed confidence interval for π at 90%.

Table 2: Six scenarios used to compare the performances of the two validation methods

δ_T	σ_T	π_T		δ_T	σ_T	π_T	
0	0.0765	0.95	Valid	0.05	0.0605	0.95	Valid
0	0.104	0.85	Valid	0.05	0.091	0.85	Valid
0	0.13	0.75	Non Valid	0.05	0.12	0.75	Non Valid

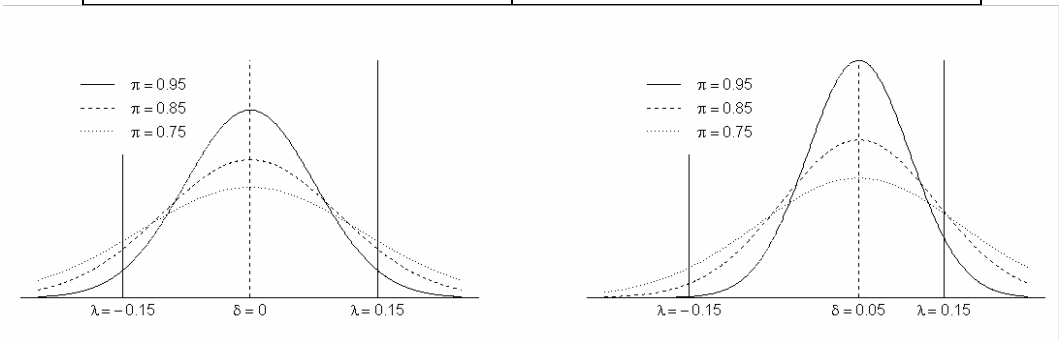

Figure 3: Six scenarios used to compare the performances of the two validation methods

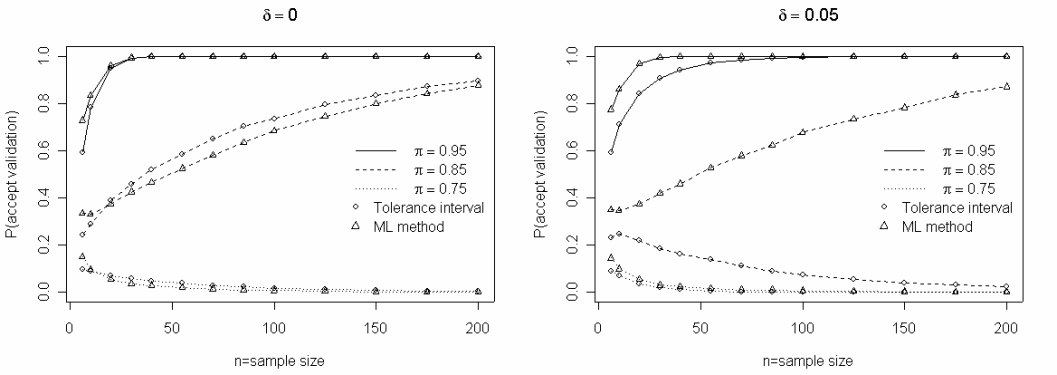

Figure 4: Simulation results: $\delta = 0$ on the left and $\delta = 0.05$ on the right

Figure 4 presents, for the six possible scenarios and the two validation methods, the proportion of cases, γ, for which the validation over 5000 simulated samples was accepted. It shows that:

- When the measurement process is not valid ($\pi = 0.75$), the behaviours of the two methods are quite similar and protect the client very well, since the probability of accepting the result is very small. The ML method performs slightly better than the tolerance interval method for centred processes, the reverse being true for biased processes.
- When the measurement process is well centred ($\delta=0$) and valid, the probability of accepting the result behaves as expected – it increases with n and with π. For $\pi = 0.85$, the tolerance interval method is more powerful than the maximum likelihood method.
- However, when the process is biased, the behaviour of the tolerance interval method is less attractive: for $\pi = 0.95$, the measurement process is less often accepted as valid than with the ML method.

- In addition, for $\pi = 0.85$, the probability of accepting the measurement process with the tolerance interval method tends to 0 as n increases. This is not a desirable result from a statistical point of view, as for $\pi = 0.85$ the process is valid. It can be shown that this arises (asymptotically) for the points in the acceptance region such that:

$$\lambda - |\delta| < \sigma z_{\frac{1+\pi_{\min}}{2}} .$$

  This result has already been roughly emphasised in [22].

This last point merits further investigation. Figure 5 illustrates it graphically: the area between the triangle and the acceptance region is a zone where, asymptotically, the measurement process is rejected by the tolerance interval method when the process is actually valid. Note that, asymptotically, this acceptance region coincides with the region where the process is asymptotically accepted by the ML method, so all valid methods should be accepted. The controversial point (δ=0.05, σ=0.091) is in this zone.

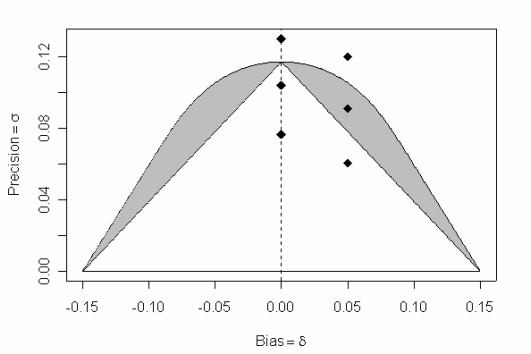

Figure 5: Representation of the acceptance region where the measurement method is (asymptotically) wrongly rejected with the tolerance interval method while it is valid and accepted with the ML method. The six simulated scenarios are marked.

For small samples, the scenario is different and the "gap" between the tolerance interval and ML methods is less important. Figure 6 gives the results of new simulations of the two validation methods behaviours in the δ, σ space. The three curves delimit the "real" acceptance region for π_{min}= 0.8 and two regions (one for each method) for which the power is higher or equal to γ=0.75. Results for two sample sizes, n=12 and n=200, are shown.

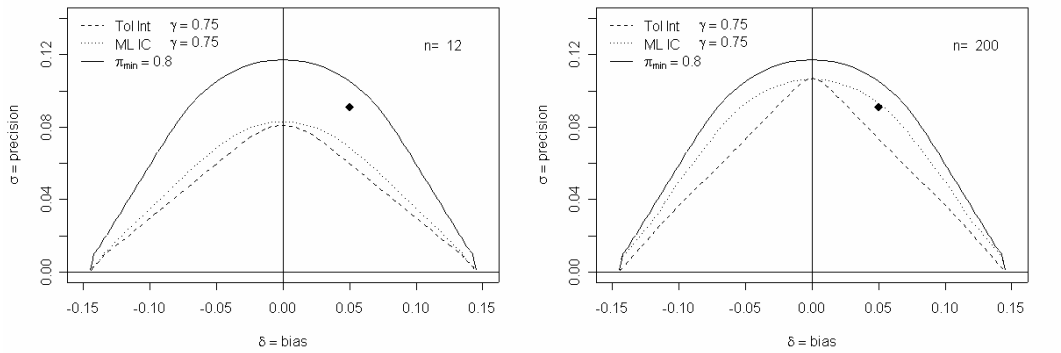

Figure 6: Iso-power line in the δ, σ space of the tolerance interval and maximum likelihood validation methods for two samples sizes (n=12 and n=200). The outer (solid) line delimits the true acceptance region for π_{min}= 0.8.

As it can be seen from Figure 6, for small sample sizes (n=12) there is very little difference in power between the two validation methods and the differences mainly occur when the analytical methods are slightly biased. The practical consequence is that, with limited sample size, there is a small region of good measurement processes that may be rejected by the tolerance interval procedure but accepted by the ML approach. This can increase the cost for the laboratories very slightly, but cannot increase the risk for the client because it does not occur outside the acceptance region where both methods are very conservative.

Once the sample size increases (n=200), it can be observed that the ML method converges to the "true" acceptance region while the tolerance interval acceptance region retains its "triangular" shape. The gap between the two procedures – the difference between the "bell" shape and the "triangular" shape – then becomes larger. This highlights the fact that the use of a one-tailed tolerance interval ignores the other side of the distribution once the tolerance limit goes outside the acceptance limit. This does not occur when the method is unbiased.

4. (S–N–λ) METHOD FOR ROUTINE FOLLOW UP

Once a method has been validated and is being used in routine analysis, it should be monitored regularly to check that the method remains valid over time. Unlike the pre-study validation phase, where expensive and cautious practices are usually envisaged, validation rule used in routine must be simple and cheap. An in-study rule that is largely accepted in the bioanalytical community, is the "4–6–15" rule and is defined in the FDA guidance [3] as: "…At least four of every six QC samples should be within 15% of their respective nominal values...". This rule provides a simple and practical guide for routine follow up; its properties are analysed below.

In general terms, the "s–n–λ" rule is applied as follows:

1. n QC samples with known nominal values are integrated into a daily run;
2. the number Y of QC samples such that the (absolute or relative) difference between the measured value X_i and the nominal value μ_T is lower than λ is counted;
3. if $Y \geq s$, the run is accepted and can be delivered to the laboratory client.

The properties of such a decision rule with respect to the laboratory's and the clients' interests depend crucially on the choice of s and n [21]. They are well represented by a power function which gives, for a given s-n-λ rule, the probability γ of accepting a run with respect to the quality level π, or the bias and precision δ and σ.

As Y is a binomial distribution with parameters n and π, the power γ is calculated as follows:

$$\gamma = P(\text{accept the run}|\pi) = P(Y \geq s|\pi) = P(Bi(n,\pi) \geq s) = \sum_{i=s}^{n} C_n^i \pi^i (1-\pi)^{n-i} \ .$$

Figure 7 illustrates this power function for the 4–6–λ rule with respect to π and iso-power curves for the 4–6–15 rule in the δ and σ space. These plots show (as expected) that the more valid the method (π large or δ and/or σ small) the higher the probability γ of accepting a run.

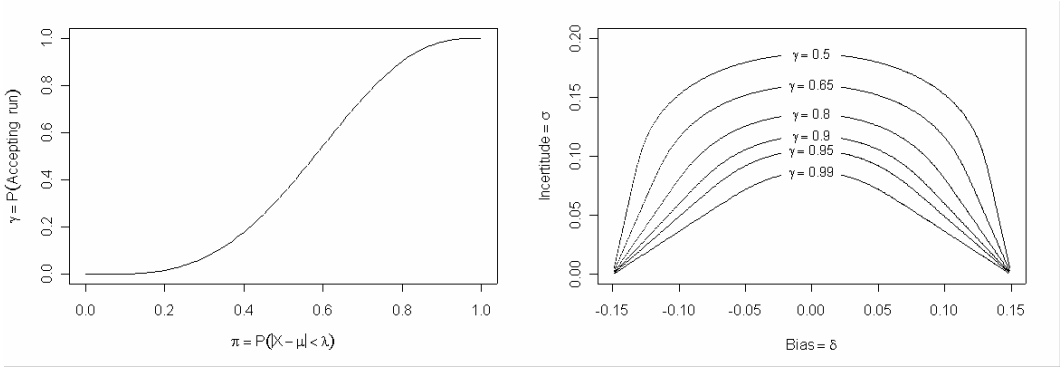

Figure 7: Power of 4–6–15 decision rule

It is instructive to study the power of the s–n–λ rule for other values of s and n. Figure 8 shows the evolution of the power of the s–6–λ rule for s = 2, 4 and 6 and of the s–n–λ rule for n = 6, 12, 24 and 96 with s/n = 2/3. These results are intuitively meaningful: increasing s (keeping n constant) decreases the client risk while increasing n (keeping s/n constant) simultaneously decreases the client and laboratory risks with respect to the compromise $\pi^* = s/n$ value. When n increases, the laboratory will have a high probability of seeing a run for which $\pi > \pi^*$ accepted, and the run will have a high probability of being rejected if $\pi < \pi^*$. Of course, the simultaneous protection of the clients' and laboratory's interests has a cost: the number n of QC samples required.

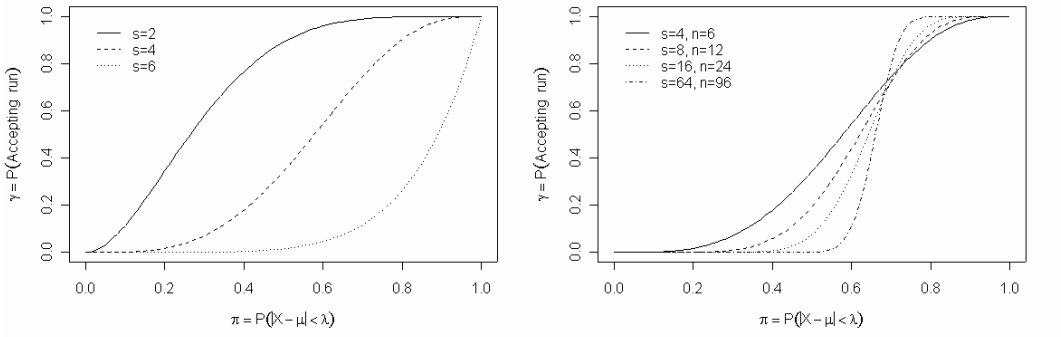

Figure 8: Power of n–s–λ as a function of s and n

Such a discussion is common in the context of lot acceptance sampling plans (see [24]) or ISO norms ([25]). In this framework, the problem is considered in the opposite direction, i.e. the requirements of both client and supplier are first fixed and then the optimal values for n and s are calculated to meet their requirements. In practice, the client has first to choose a quality level π_C under which the probability of accepting a run (or lot) is small (γ_C say). On the other side, the supplier (laboratory) has to choose a quality level π_L, above which the probability of accepting a run (or lot) is high enough (γ_L say). This leads to a system of two inequalities to be solved with two unknown values, n and s:

$$P(Y \geq s | \pi < \pi_C) < \gamma_C \qquad \text{client equation}$$
$$P(Y \geq s | \pi > \pi_L) > \gamma_L \qquad \text{supplier equation}$$

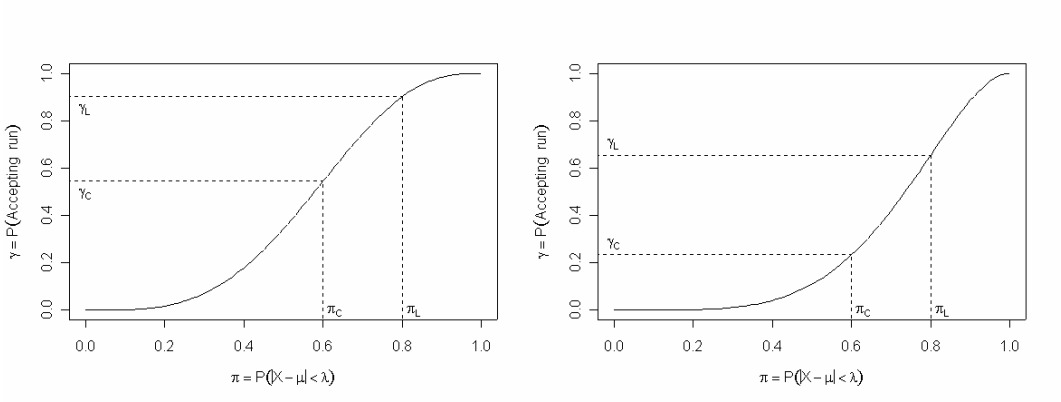

Figure 9: A comparison of client and laboratory risks for the 4–6–λ and 6–6–λ rules

This is illustrated in Figure 9 for the 4–6–λ and 6–6–λ rules. The figure shows that the 4–6–λ rule is highly protective for the laboratory and favours the laboratory over the client.

Table 3 presents the optimal n and s values for different combinations of client and supplier parameters. It illustrates clearly the impossibility of achieving acceptable client and supplier risks with a reasonable cost by using a classical s–n–λ rule. The FDA recommendations should therefore be interpreted as favouring laboratory risks over consumer risks.

Table 3: Optimal sampling plans for different combinations of client and laboratory requirements

π_C	γ_C	π_L	γ_L	n	s	π_C	γ_C	π_L	γ_L	n	s
0.6	0.2	0.8	0.8	19	14	0.6	0.1	0.8	0.9	36	26
0.7	0.2	0.8	0.8	55	42	0.7	0.1	0.8	0.9	127	96
0.7	0.2	0.9	0.8	14	12	0.7	0.1	0.9	0.9	25	21
0.8	0.2	0.9	0.8	39	34	0.8	0.1	0.9	0.9	86	74

5. CONCILIATING VALIDATION AND ROUTINE DECISION RULES

The basic aim, when applying pre-study and in-study validation procedures to a measurement method, is to conciliate the objectives of the two validation procedures. When the total error approach is used in the pre-study and the s–n–λ method in the in-study there is a common objective: to control the parameter π or the proportion of measurement results $(X–\mu_T)$ expected to lie within the acceptance limit $[–\lambda,+\lambda]$. The shape of the acceptance region of the two methods is then equivalent, as shown in Figures 2 and 5.

The ability to conciliate the pre- and in-study procedures depends then on the adequacy of the parameters chosen. These parameters should ensure that a laboratory which has proved a method to be valid in a pre-study experiment will see most of the runs using this analytical method accepted in routine analysis if the performance of the method remains stable over time. This is essential: it would be counterproductive to maintain an analytical method that frequently leads to runs being rejected when the method is still valid.

This requirement can be reformulated in terms of the test method's parameters in two ways:

1. If the parameters n and s of the s–n–λ rule are fixed, the value of π_{min} should be chosen so as to ensure that, if the method remains valid, the s–n–λ rule is accepted in most cases (e.g. with a minimum probability γ_{min}).
2. On the other hand, for a pre-study validation scheme (π_{min} and λ), the value of s (for a given n) should guarantee that most of the runs will be accepted if the method remains valid.

Note that this formulation focuses on the laboratory. On the client side, as seen in Section 4, it is difficult to really protect the client at a reasonable cost with an s–n–λ rule. The parameters should then protect the client as much as possible given the budget available and the laboratory's economic rationale.

Let us now consider the particular case of the 4–6–λ rule. As stated above, a good pre-study validation rule should work with a π_{min} value which ensures that the routine test will be accepted in most cases (say γ_{min}=90%) if the method is valid:

$$P(Y \geq s \mid \pi > \pi_{min}) > \gamma_{min}$$

For a given s and n, this π_{min} is obtained by inverting the binomial(n, π) distribution function in π, as shown in Figure 10.

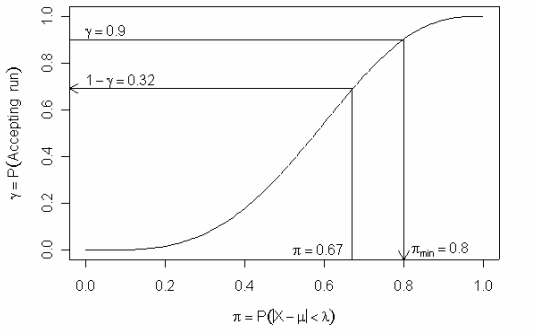

Figure 10: Conciliating pre-study validation π_{min} value with 4–6–λ rule

For the 4–6–λ rule we have π_{min} = 0.8. This means that, in the pre-study validation experiment, the laboratory should demonstrate that at least 80% of the measurements $X-\mu_T$ are expected to lie within the acceptance limits [–λ,+λ]. This will ensure that the 4–6–λ rule will accept 90% of cases in routine use if the process remains valid ($\pi > \pi_{min}$). This contrasts with the (intuitive) proposal frequently encountered in the literature [8, 17] that 4/6 or 66.7% of the results should lie within the acceptance limits. Adopting 66.7% as the value for π_{min} can lead to up to 32% of the "valid" runs being rejected, as can be seen in Figure 9.

On the other hand, when, as we recommend, the s–n–λ rule is not fixed in advance, it is easy to calculate the best value of s for given π_{min}, n and γ_{min}. The procedure consists of finding the maximum value of s ($0 \leq s \leq n$) such that

$$P(Y \geq s \mid \pi = \pi_{min}) > \gamma_{min} \ .$$

Figure 11 shows this optimal value of s for different values of n and π_{min}, when γ_{min} is fixed at 0.9.

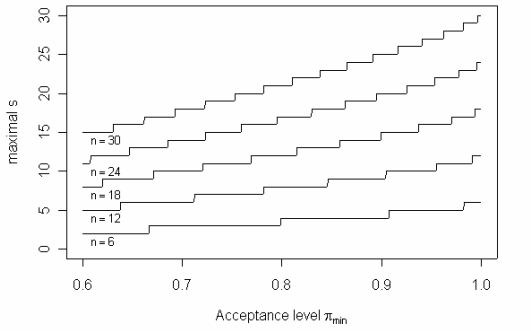

Figure 11: Optimal s values for a fixed n and pre-study acceptance level $\boldsymbol{\pi_{min}}$

6. CONCLUSION

Throughout this paper test methods, based on the total error concept, have been proposed to decide about the acceptance or rejection of an analytical method at a pre- and in-study stage so as to minimise both the consumer and the laboratory risks.

The tolerance interval and the maximum likelihood method are the two alternative methods proposed for the pre-study validation. For limited sample sizes, often encountered in laboratories, simulations have shown that they behave similarly in the sense that they have comparable power curves over the (δ,σ) space. The tolerance interval approach, however, has asymptotical unsuitable behaviour because the acceptance region of this test does not converge, as the maximum likelihood method does, to the true region. In practice, the difference is negligible, but the maximum likelihood method should nevertheless be preferred, not only for its good statistical properties, but also because it directly answers the fundamental question: what is the expected proportion of measures that will fall within the acceptance limits in the future?

The s-n-λ rule is presented as a possible approach for in-study validation. It is shown that, with a limited sample size, this rule is unfortunately not able to protect simultaneously the laboratory and the client interests. In particular, the 4-6-λ rule recommended by FDA in the pharmaceutical industry favours the laboratory to the detriment of the client.

This paper also discusses the necessity to conciliate pre-study validation criteria and routine-run acceptance rules so as to minimise both consumer and laboratory risks. In other words, depending on the rule to be used routinely for accepting runs, what should the minimal probability π_{min}, as estimated from validation experiments, be, to ensure that in routine use most valid runs are accepted keeping in mind the client risks.

The results show, that if the 4–6–λ rule is retained, the minimum proportion π_{min} of measures that should be expected to fall within the acceptance limits $[-\lambda,+\lambda]$ to guarantee that at least 90% of the runs are accepted when the measurement process remains valid, is 80%. Taking 80% as π_{min} value allows the pre-study and in-study decision rules to be made consistent. But, as also shown, the 4–6–λ rule lacks power and favours the laboratory over the consumer. The only way to minimise both the consumer and the laboratory risks is to improve the in-study rule, by increasing both the number of QC samples (n) and the number of successful samples (s), and adapting the π_{min} value accordingly for the pre-study validation. But, as shown in Table 3, this requires a very large and impracticable number of QC

samples. This is why more advanced techniques, based directly on the quantitative measures and/or taking into account the history of the method's results, are recommended. Sampling plans for measures [26], scan statistics [27] and moving-type control charts (Cusum, Ewma) [28] are all attractive solutions. They would merit to be more emphasised by international regulation texts.

REFERENCES

[1] International Conference on Harmonization (ICH) of Technical Requirements for the Registration of Pharmaceuticals for Human Use, *Validation of Analytical Procedures*, ICH–Q2A, Geneva 1995.

[2] International Conference on Harmonization (ICH) of Technical Requirements for the Registration of Pharmaceuticals for Human Use, *Validation of Analytical Procedures: Methodology*, ICH–Q2B, Geneva 1996.

[3] FDA Guidance for Industry, *Bioanalytical Methods Validation*, U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), May 2001.

[4] Hubert P., Nguyen J.J., Boulanger B., Chapuzet E., Chiap P., Cohen N., Compagnon P.A., Dewe W., Feinberg M., Lallier M., Laurentie M., Mercier N., Muzard G., Nivet C. and Valat L. Harmonization of approaches for validation of quantitative analytical procedures: a SFSTP proposal. *Journal of Pharmaceutical and Biomedical Analysis* (2004) 36, 579–586.

[5] Smith W.C., Sittampalam G.S. Conceptual and statistical issues in the validation of analytical dilution assays for pharmaceutical applications. *Journal of Biopharmaceutical Statistics* (1998), 8(4), 509–532.

[6] Findlay J.W.A., Smith W.C., Lee J.W., Nordblom G.D., Das I., Desilva B.S., Khan M.N. and Bowsher R.R. Validation of immunoassays for bio analysis: a pharmaceutical industry perspective. *Journal of Pharmaceutical and Biomedical Analysis* (2000), 21, 1249–1273.

[7] Miller K.J., Bowsher R.R., Celniker A., Gibbons J., Gupta S., Lee J.W., Swanson S.J., Smith W.C. and Weiner R. Workshop on bioanalytical methods validation for macromolecules: Summary report. *Pharmaceutical Research* (2001), 18(9), 1373–1383.

[8] Desilva B., Smith W., Weiner R., Kelley M., Smolec J, Lee B., Khan M., Tacey R., Hill H. and Celniker A. Recommendations for bioanalytical method validation of ligand-binding assays to support pharmacokinetic assessments of macromolecules, *Pharmaceutical Research* (2003), 20, 1885–1900.

[9] Caporal-Gautier J., Nivet J.M., Algranti P., Guilloteau M., Histe M., Lallier M., N'guyen-Huu J.J. and Russotto R. Guide de validation analytique – rapport d'une Commission SFSTP – Méthodologie. *STP Pharma Pratiques* (1992), 2(4), 205–226.

[10] Hartmann C., Smeyers-Verbeke J., Massart D.L. and McDowall R.D. Validation of bioanalytical chromatographic methods. *Journal of Pharmaceutical and Biomedical Analysis* (1998), 17, 193–218.

[11] Shah V.P., Midha K.K., Findlay J.W.A., Hill H.M., Hulse J.D., McGilveray I.J., McKay G., Miller K.J., Patnaik R.N., Powell M.L., Tonelli A. Viswanathan C.T. and Yacobi A. Bioanalytical method validation – a revisit with a decade of progress. *Pharmaceutical Research* (2000), 17(12), 1551–1557.

[12] Boulanger B., Dewe W., Hubert P., Rozet E., Moonen F., Govaerts B. and Maumy M.. Conciliating objectives of analytical methods and objectives of validation: a statistical perspective. *Inst. Validation Tech.* December 7, 2005, Philadelphia.

[13] Hubert P., Chiap P., Crommen J., Boulanger B., Chapuzet E., Mercier N., Bervoas-Martin S., Chevalier P. Grandjean D., Lagorce P., Lallier M., Laparra M.C., Laurentie

M. and Nivet C. The SFSTP guide on the validation of chromatographic methods for drug analysis: from the Washington Conference to the laboratory. *Analytica Chimica Acta* (1999), 391, 135–148.
[14] Boulanger B., Hubert P., Chiap P. and Dewe W. Objectives of pre-study validation and decision rules. AAPS APQ Open forum, Washington, 2000.
[15] Boulanger B., Hubert P., Chiap P., Dewe W. and Crommen J. Analyse statistique des résultats de validation de méthodes chromatographiques, *Journées GMP*, Bordeaux, 2000.
[16] Boulanger B., Chiap P., Dewe W., Crommen J. and Hubert P. An analysis of the SFSTP guide on validation of chromatographic bioanalytical methods: progresses and limitations. *Journal of Pharmaceutical and Biomedical Analysis* (2003) 32, 753–765.
[17] Hoffman D. and Kringle R. Two-sided tolerance intervals for balanced and unbalanced random effects models. *Journal of Biopharmaceutical Statistics* (2005), 15 (2) 283–293.
[18] Mee R., W. β-expectation and β-content tolerance limits for balanced one-way ANOVA random model. *Technometrics* (1984), 26(3), 251–254.
[19] Mee R., W. Estimation of the percentage of a normal distribution lying outside a specified interval. *Communication in Statistics – Theory and Methods* (1988), 17(5), 1465–1479.
[20] Chapuzet E., Mercier N., Bervioas-Martin S., Boulanger B., Chevalier P., Chiap P., Grandjean D., Hubert P., Lagorce P., Lallier M., Laparra M.C., Laurentie M. and Nivet J.C., Méthodes chromatographiques de dosage dans les milieux biologiques: stratégie de validation. Rapport d'une commission SFSTP. *STP Pharma Pratiques* (1997), 7, 169–194.
[21] Chapuzet E., Mercier N., Bervioas-Martin S., Boulanger B., Chevalier P., Chiap P., Grandjean D., Hubert P., Lagorce P., Lallier M., Laparra M.C., Laurentie M. and Nivet J.C. Méthodes chromatographiques de dosage dans les milieux biologiques: stratégie de validation, exemple d'application de la stratégie de validation. *STP Pharma Pratiques* (1997), 8, 81–107.
[22] Castaneda-Mendez K., Medical utility frequency, *Clinical Chemistry*, 33/2 (1987) 212–222.
[23] Kringle R.O. An assessment of the 4–6–20 rule for acceptance of analytical runs in bioavailability, bioequivalence, and pharmacokinetic studies. *Pharmaceutical Research* (1994), 11, 556–560.
[24] Ducan A., *Quality Control and Industrial Statistics*, Fifth Edition, Irwin, Boston, 1986.
[25] International Organization for Standardization. Sampling procedures for inspection by attributes. *ISO/DIS* (1995) 2859 (Part 0).
[26] International Organization for Standardization. Sampling procedures for inspection by variables. *ISO/DIS* (2005) 3951 (Part 0).
[27] Glaz J., Naus J. and Wallenstein S. *Scan Statistics*, Springer, New York, 2001.
[28] Montgomery D.C., *Introduction to Statistical Quality Control*, 4th edition, Wiley, New York, NY, 2004.