Available online at www.sciencedirect.com

ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis 44 (2007) 718–729

JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS

www.elsevier.com/locate/jpba

Out-of-specification test results from the statistical point of view

Heidi Köppel[a], Berthold Schneider[b], Hermann Wätzig[a,*]

[a] Institute for Pharmaceutical Chemistry, TU Braunschweig, Braunschweig, Germany
[b] Institute for Biometry, MH Hannover, Germany

Received 15 September 2006; received in revised form 21 February 2007; accepted 25 February 2007
Available online 3 March 2007

Abstract

Although a generally accepted procedure has now been established for the organizational handling of out-of-specification test results, the uncertainty surrounding their statistical evaluation persists. Two statistical equations, the prediction and the confidence interval, are sufficient to examine whether data numbers indicate out-of-specification (OOS) results or not. This is demonstrated by means of 10 examples. These equations are usually sufficient to specify limit values as well. A number of consequences have been derived from a discussion of borderline cases:

(A) If only one measured value is OOS, the same is true for the whole result (there are three exceptions: high data numbers, outliers, or the reportable result is not the single value but e.g. the mean).
(B) The result is not automatically within specification, if this holds true for all measurements. If all measurements are close to the specification limit and the measurement error is high, an OOS results is still possible.
(C) If it is clear that the obtained data will be close to the limit, a precisely working method and a relatively high data number is required. In order to obtain future measurements that remain within specification, the difference between the limit and the mean value must not become smaller than 1.65 times the standard deviation, even if very high numbers of measurements are provided.

Procedures to deal with extreme values, so-called outliers, are not straightforward. The statistical evaluation is troublesome, because the probability distribution cannot be determined. This problem is discussed by another four examples. In several cases the outlier can be detected without doubt, for example, using Dixon's test or the box plot. However, there are a number of borderline cases, when a value is suspected to be an outlier, but this cannot be proven by statistics [7,9].

Keywords: Statistics; Out-of-specification results; Outliers

1. Preamble

1.1. Importance and definitions

The handling of out-of-specification test results (OOS results) has been a hot topic since the rendering of the Barr Decision in 1993. On the one hand, the intended or unintended manipulation of data can feign within-specification test results (WS results) even if quality limits are violated. On the other hand, there is the danger of further complicating control mechanisms. The latter would result in a disproportionate increase of the work scope and thus to higher production costs and ultimately be reflected in unnecessarily high drug prices [11].

First, the definition given by the FDA for purposes of the document "Guidance for Industry" [1]: "The term *OOS results* includes *all* test results that fall outside the specifications or acceptance criteria established in drug applications, drug master files (DMFs), official compendia, or by the manufacturer. The term also applies to all in-process laboratory tests that are outside of established specifications."

This definition should be amended as it concerns only "OOS results" for single measurements. The reportable result however can also be calculated from several single values. For example, sometimes the mean of a measurement series is the reportable

* Corresponding author at: Institute for Pharmaceutical Chemistry, TU Braunschweig, Beethovenstr. 55, 38106 Braunschweig, Germany.
Tel.: +49 531 3912764.
E-mail address: h.waetzig@tu-bs.de (H. Wätzig).
URL: http://www.pharmchem.tu-bs.de/forschung/waetzig (H. Wätzig).

0731-7085/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2007.02.035

result. In principle the considerations below remain valid if one treats the reportable results as single values. The reportable results however may provide some additional information, as information about the data distribution may be available.

Whether single measurements are OOS or WS can be easily determined by comparison of the results with the specifications and requires no statistical methods. Of crucial importance is the question whether entire production units, for example, production batches, must be considered WS or OOS. Generally, production units cannot be tested as a whole. Samples are taken from the unit whose pertinent parameters are then measured. The measurements obtained in this fashion will generally show random variation. A part of the single results can be WS, another OOS. The question here is: When should an entire production unit be declared OOS, when WS?

In order to formulate this question precisely and understand the solutions, one has to draw upon some of the basic terms of the theory of probability and statistics. Imagine that instead of taking only a few samples from the production unit, sampling is constantly repeated and measurements are taken each time. The samples are taken at random and independent of each other. The result is a series of measurements whose variation is random. This series is referred to as the total population (R. v. Mises has also introduced the term "collective"). The relative frequency at which measurements within a certain interval (from x to $x + \mathrm{d}x$) occur in this total population is the probability of finding a measurement within this interval in a sample randomly selected from the production unit. The distribution of the probabilities across the possible value range of the measurements is characterized by the distribution function $F(x)$. It is the probability of obtaining a measurement value from a randomly selected sample that is smaller than or equal to x, where x covers the entire value range. This distribution function characterizes in completely this total population of measurements. It therefore makes sense to define the specification conformity of a production unit (batch) by means of this distribution function. For example, if x_{lsl} represents the lower specification limit (i.e. the measurements $x \geq x_{\mathrm{lsl}}$ are WS and the measurements $x < x_{\mathrm{lsl}}$ are OOS), the production unit will be considered WS if the probability γ for WS results is greater than a prescribed threshold γ_0 (i.e. $1 - F(x_{\mathrm{lsl}}) \geq \gamma_0$), where γ_0 is to be set at a minimum of 50%. With two-sided specification limits x_{lsl} and x_{usl} (i.e. measurements in the range of $x_{\mathrm{lsl}} < x \leq x_{\mathrm{usl}}$ are WS and values outside this range are OOS) the production unit is considered WS if the probability γ for WS results is greater than the threshold γ_0 (i.e. $F(x_{\mathrm{usl}}) - F(x_{\mathrm{lsl}}) > \gamma_0$), otherwise it is considered OOS.

Thus, in order to decide whether a production unit is to be declared OOS or WS, it is necessary to specify a threshold value γ_0 for the probability of WS results in the total population. By means of the measurements obtained from a finite number of samples taken from the production unit it can then be determined whether the percentage of WS results in the total population is smaller than or equal to γ_0 (the production unit is then classified as OOS), or greater than γ_0 (the production unit is to be classified as WS). The number of samples taken is usually expressed as n and the n obtained measurements $x_1, x_2, \ldots, x_n$ are called a random sample. It is assumed that n measurements were taken from the total population at random and independently of each other. As the random sample usually constitutes only a very small part of the total population, it follows that it does not fully suffice to represent the total population. The determinations made based on the random-sample values and thus even the decision whether a production unit is WS or OOS, can therefore be incorrect. To formulate the method accurately it is necessary to also specify the permissible probability of incorrect decisions. The probability with which a sufficiently good production unit is falsely rejected as OOS is generally referred to as α, the probability with which an OOS production unit is considered WS as β.

The error probabilities α and β depend on:

1. The distribution $F(x)$ of the total population.
2. The employed method for the decision.
3. The size n of the random sample on which the decision is based.

They do not describe the total population but merely the method upon which the decision is made. In order to define α and β more precisely, one has to imagine that the decision between OOS and WS is based on every possible random sample from the total population. Again and again, n samples are taken from the production unit and measured. Based on the results of these measurements the same method is employed to decide whether the production unit is WS or OOS, with the actual quality of the production unit (i.e. either WS or OOS) being known.

The error probability α is the relative frequency with which an OOS decision is made using this method despite the production unit being WS. Accordingly, β is the probability with which the production unit is considered WS despite in reality being OOS.

The law of large numbers, established by and named after the Swiss mathematician Jakob Bernoulli (1654–1705), describes the correlation between the size n of the random sample and the error probability in evaluating a production unit based on a random sample. The larger the random sample n, the more precise a statement can be made in regard to the total population on the basis of this random sample. And therefore, the smaller the probability of rendering an incorrect decision based on the random sample. The size n of the random sample also affects the accuracy of the conclusions and decisions derived from the random sample. The complementary value $1 - \alpha$ accordingly expresses the reliability with which the correct decision of "WS" is rendered for a production unit, and the value $1 - \beta$ the reliability with which the correct decision of "OOS" is made for a production unit.

In evaluating the specification conformity of production units (production batches), it therefore does not suffice to formulate the specification limits or acceptance criteria for individual measurement values. One also has to specify the WS percentage γ_0 of the total population to be exceeded if the production unit is to be declared WS, as well as the error probability α or the reliability $1 - \alpha$ with which the specification conformity of the production unit is to be determined based on n random-sample values.

The handling of OOS results presents in large part also an organizational problem. In many cases the causes of errors can be quickly contained, detected and prevented in the future. Important for this are:

- documentation,
- error investigation,
- statistical evaluation,
- batch processing,
- process adaptation/change,
- standard operating procedures (SOPs) and their establishment,
- preparation of audits.

Specify an overview of procedures in the event of suspected out-of-specification quality [1–4].

This article discusses statistical methods to determine whether a production unit is OOS or WS by means of examples based on the proposed guidelines [1] and the above-cited additional observations. From the statistical point of view it also provides suggestions on devising tests aimed at yielding reliable results.

1.2. Introductory examples

When does a random sample of (analytical) measurements from a production unit (batch) provide information on the violation of the specification in this production unit? Conversely, in which cases can it be assumed based on the available analytical results that the specifications are met? Let us discuss a few examples (compare Fig. 1).

1.2.1. Various options for the correlation between limit, mean value and measurement range

The amount of a secondary component should not exceed a specified limit (e.g. 0.1%). Horizontal lines depict the highest and lowest measurement value, respectively, forming the measurement range, and the mean value of the random sample. Cases A and B are clear: even the mean value lies above the limit; the investigated samples are OOS (out-of-specification, i.e. they do not conform to the requirements). In case C the mean value lies below the specified limit but single values lie above it. Can the test result still be within specification (WS)? Even, if the reportable result is not the mean?

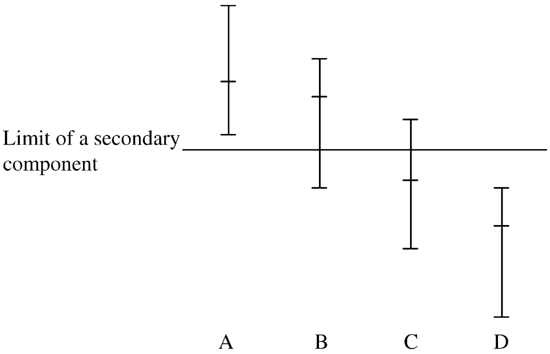

Fig. 1. Various options for the correlation between limit, mean value and measurement range.

In case D all values of the random sample lie above the specified limit. Does this offer a guaranty for meeting the specification or can the test result be OOS nonetheless? If yes: with what probability? Each of these questions correlates to the statistic inference drawn from a random sample to the respective total population.

1.2.2. Are borderline cases still within specification?

Fig. 1 shows that in borderline cases it is not always possible to decide clearly between OOS and WS results. Fig. 2 depicts some of these discussed borderline cases in more detail. Depicted are five measurement values each from random samples A, B, C1, . . ., D4. They represent hypothetical but realistic data sets based on normally distributed random numbers. The dashed line marks the specification limit, the amount of a secondary component in this example should not be smaller than 0.1. Cases A and B are not borderline cases: the mean values of the random samples all lie above the specification limit, as does a large part of the prediction range. What is the case in

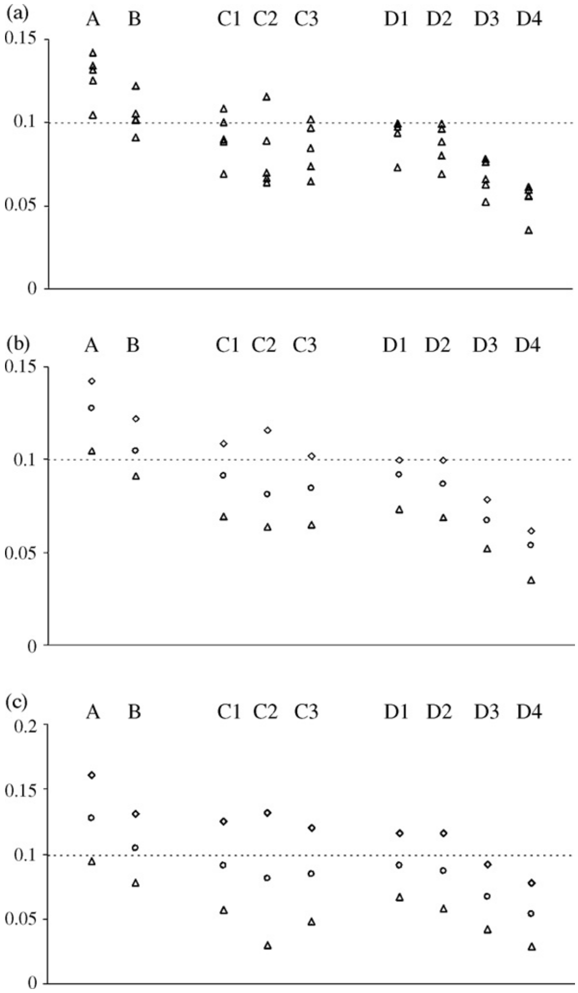

Fig. 2. (a–c): Detailed discussion of scenarios A–D (compare Fig. 1). The limit in all examples is 0.1. (a) Single values. (b) Mean value and range. (Circle) Mean; (diamond and triangle) highest and lowest single value. (c) Prediction interval. (Circle) Mean; (diamond and triangle) upper and lower limit of the prediction interval.

examples C1–C3, however? Can a single value be OOS if an investigation reveals the total result to be WS? Quite so! With very high data numbers it is even to be expected that single values will lie outside the specification range, as the normal distribution is generally unlimited. However, for the frequently used lower data numbers ($n < 10$), even a single OOS value points strongly to an overall OOS result, provided that the single values and thus the prediction range are relevant for the evaluation of the random sample ("individual OOS results indicate nonconformance," compare [1]). If only the mean value is required to conform to specification for the batch to be WS, a single value in smaller random samples may sometimes be OOS (e.g. Fig. 2, C3) without the total batch having to be judged OOS [8]. It is impossible, however, in cases of small data sets (cases C1–C3) for the total data set to be WS if a single sample is OOS and the prediction range is used for the evaluation (compare Sections 2.1.3 and 2.2.1). This interval must be relied upon for the assessment if it is crucial for the individual elements of a batch to conform to a given specification.

This statement is generally valid if the *t*-distribution is used for single values. Even if all other measurement values lie clearly within the specification and only one value just outside, the standard deviation increases significantly. This, in turn, increases the prediction interval, which then overlaps the specification limit. Each single value of a random sample thus strongly influences both the position and the size of the prediction interval, especially for the customary small data sets. The rule therefore is: if one value is OOS, the *entire* data set is OOS if the individual values are crucial and the number of data is small.

1.2.3. Assessment of production units based on the results obtained from a random sample n

Fig. 3 depicts the measurement values from four random samples of varying sizes, obtained by simulation from a normally distributed total population having a mean value of 6.2 ppm and a standard deviation of 2.2. A one-sided, upper specification limit of 10 ppm has been specified for this example. This means, values greater than or equal to 10 are considered OOS, smaller values WS. 95% of the values of this total population conform to specification, 5% are OOS. The dashed line marks the specification limit. Any values lying above this limit are OOS. The specification limit for this example is set to be 10 ppm and corresponds to the maximum permitted concentration of an impurity in a pharmaceutical. In case 1 the random sample consists of $n = 5$ values, all of which are smaller than 10 ppm, i.e. within specification. One would therefore be inclined to consider the production unit specification-conform. But what can indeed be said about the percentage γ of specification-conform measurement values in the total population based on these five values?

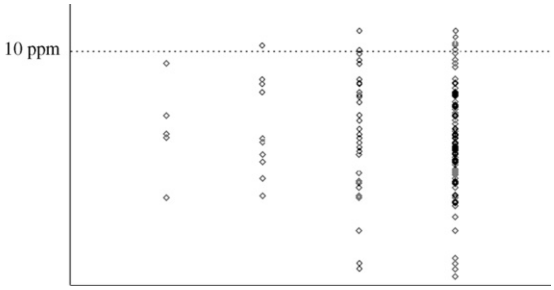

Fig. 3. Four random samples from the same total population ($N = 100$) with $\gamma = 95\%$ WS results. Case 1: $n = 5$; case 2: $n = 10$; case 3: $n = 30$; case 4: $n = 100$.

As we will show later, based on five random-sample values that are all WS one can state with a reliability (probability) $1 - \alpha$ of 80% that at least the percentage $\gamma_0 = 0.725$ (72.5%) of the total population is specification-conform (WS). At this reliability of 80%, the production unit can be considered WS if the specified threshold $\gamma_0 = 0.70$.

If one wishes to make the statement with a reliability $1 - \alpha$ of 90%, then the lower limit for γ_0 can be set at only 0.631. If the reliability is to be 95%, the lower limit can only be 0.549. The reliability $1 - \alpha$ and the minimum percentage γ_0 of WS results in the total population able to be assumed based on the random-sample result, move in opposite directions for a given random-sample size n. The greater the reliability the smaller the lower limit that can be maintained for the percentage γ_0.

In the second case the random sample consists of 10 values of which one is greater than 10 ppm and therefore OOS. In case 3 the random sample comprises 30 measurement values of which 2 are OOS. In case 4 the random sample comprises 100 values of which 5 are greater than 10 ppm. All of these cases show the type of inference one can draw on the specification conformity of production units based on the results obtained from random samples: at a specified threshold γ_0, it can be determined with the error probability α or the reliability $1 - \alpha$ whether a production unit is WS or OOS. Furthermore, limits for the percentage γ of WS results in the total population can be maintained with a reliability of $1 - \alpha$.

The calculations for these special cases can be viewed in the additional online Section 2.4.1 which the interested reader will find at following web site: http://www.pharmchem.tu-bs.de/forschung/waetzig/support/.

2. Statistical methods to determine the OOS status of production units Statistical inferences from the random sample to the total population

2.1. Distribution-free calculations

2.1.1. Test methods to determine whether the percentage of WS results in the total population is greater than a threshold γ_0

As stated in the preamble, the aim is to state to a certain reliability $1 - \alpha$, based on the results of random sampling, on the probability γ with which in a production unit results conforming to specification can be expected. In particular, it is to be determined based on this reliability whether γ is greater than a specified threshold γ_0 or smaller than or equal to γ_0. With given specification limits, the actual percentage γ depends on the distribution function $F(x)$. It is important to point out that the real distribution function is normally not known. This function, however, is not required to determine the percentage γ. It suffices to determine how many of n random-sample values $x_1, \ldots, x_n$ are WS and how many are OOS. In the first case of the five random-sample values, for example, all five values were WS (Fig. 3). In case 2, 9 of 10 values were 9 WS (0.9), in case 3, 28 of 30 (0.933) were WS and in case 4, 95 of 100 values (0.95) (Fig. 3). By means of the number k of WS results among n random-sample values (or by means of the frequency $h = k/n$) it is to be determined whether the percentage γ of WS results in the total population is greater than γ_0 or whether it does not exceed this threshold. If $\gamma > \gamma_0$ one can assume that the production unit conforms to specification. In the statistical test theory the assumption "the production unit meets the specifications" is called the null hypothesis and the assumption "the production unit does not meet the specifications" as alternative hypothesis. This assignment is arbitrary and may be reversed. The rejection of the null hypothesis despite its being correct should occur with the probability α at the highest. α is the error probability of the 1st kind, also referred to as "producer's risk." The reverse case, where one decides in favor of the null hypothesis although the alternative hypothesis applies, is referred to as error of the 2nd kind or as "consumer's risk" and occurs with the probability β.

The test method implies that for the number k of WS results in the random sample of the size n, a threshold k_0 is specified. If the number k of WS results in the random sample of the size n reaches or exceeds this threshold, the null hypothesis is confirmed and the production unit considered WS. If k does not reach the threshold k_0 the null hypothesis is abandoned and the production unit considered OOS.

If n is not too small, Eq. (1) should apply (for detailed information see the online attachment to this publication)

$$n \cdot \gamma_0 \cdot (1 - \gamma_0) > 1 \qquad (1)$$

the smallest natural number may be used as threshold k_0 for which the following is true:

$$k_0 \geq n \cdot \gamma_0 + z_{1-\alpha}\sqrt{n \cdot \gamma_0 \cdot (1 - \gamma_0)}; \quad k_0 \in N \qquad (2)$$

k_0 is the threshold value required to be reached for the percentage of WS values in the random sample in order for the production unit to be considered; WSγ_0 is minimum threshold value required to be reached for the percentage of WS values in the production unit in order for the production unit to be considered WS. $z_{1-\alpha}$ is the $(1 - \alpha)$ quantile of the standard normal distribution (mean value 0, variance 1); n is the size of the random sample.

Here, $z_{1-\alpha}$ is the $(1 - \alpha)$-quantile of the standard normal distribution (mean value 0, variance 1). It describes the value for which $n \cdot \alpha$ observed values are smaller or equal and $n \cdot (1 - \alpha)$ values greater. The values for $z_{1-\alpha}$ can be found in the table "Quantiles of the Standard Normal Distribution $N(0, 1)$," cited in every statistics text book or manual. For example, one finds for:

$$\alpha = 0.2 \quad z_{1-\alpha} = 0.8416 \quad \alpha = 0.05 \quad z_{1-\alpha} = 1.6449$$
$$\alpha = 0.1 \quad z_{1-a} = 1.2816 \quad \alpha = 0.025 \quad z_{1-\alpha} = 1.960$$

For example, for $n = 5$, $\gamma_0 = 0.5$ and $\alpha = 0.1$, the value on the right side of the formula is equal to 3.93. Rounded to the next greater whole number, the value becomes 4. The null hypothesis is also acceptable and the production unit considered WS if at least four of the five measurement results are WS. If one wishes to make a decision using $\alpha = 0.05$, then the figure on the right of the formula is equal to 4.34. Rounded to the next greater whole number it becomes 5. The production unit can be considered WS with a reliability of 95% only if all five measurement results are WS. In case 1 all five measurement values are WS. Conversely, it can be stated with a reliability of 95% that the percentage γ of WS results in the total population is greater than or equal to 0.5 (da $\gamma_0 = 0.5$). If the percentage 0.7 is specified as threshold γ_0, then the value on the right side for $\alpha = 0.1$ is 4.81, i.e. 5. Accordingly, the production unit is considered WS with a reliability of 90% $(1 - \alpha)$ if all five measurement values are WS. If at $\gamma_0 = 0.7$ one wishes to make a determination with a reliability of 95% $(\alpha = 0.05)$, then the number on the right of the formula is equal to 5.19, i.e. 6! This means that of five measurement values, 6 must be WS, which, of course, is impossible. With this reliability it is therefore no longer possible to determine for five random samples whether γ_0 is greater than 0.7. This shows that in cases of small random-sample sizes, determinations with a specified reliability $1 - \alpha$ can be made for limited values of γ_0 only. If all random-sample results are WS and one wishes to determine, based on a given α, whether the percentage of WS results in the total population is greater than or equal to γ_0, the following applies to the size n of the random sample as a first approximation:

$$n \geq \frac{\log \alpha}{\log \gamma_0}; \quad n \in N. \qquad (3)$$

For $\alpha = 0.05$ and $\gamma_0 = 0.7$, therefore, n should be greater than 8.40, i.e. at least 9.

A useful result can still be obtained from Eq. (2) for a random-sample size of $n = 7$. Here, all seven values of the random sample must be WS in order for the batch to be considered WS, i.e. $n = k_0 = 7$. For smaller random-sample sizes, even Eq. (2) is unsuitable for making a determination for the given parameters. The deviation during the calculation of n originates from the fact that Eq. (3) allows only a rough estimate of the minimum required random-sample size n. By inputting the values for α and γ_0 one obtains the value 9. Eq. (2) allows for a more exact calculation but is more cumbersome. By equating k_0 and n and inputting the values for α and γ_0 one arrives relatively quickly at the wanted number even here in solving the equation. If the number k of WS results in the random sample does not reach the threshold k_0 and the null hypothesis is therefore abandoned, it does not follow that the WS percentage γ in the production unit cannot be greater than γ_0. For a given random-sample size n the null hypothesis just cannot be maintained with the specified reliability $1 - \alpha$.

The probability with which the null hypothesis is abandoned for a WS percentage $\gamma < \gamma_0$ (the total population thus is OOS), i.e. the production unit is subsequently rejected as OOS, is called the "power" of the test method at the value γ. This power is designated as $1 - \beta$. It is the complement to the β-error, i.e. the error of the 2nd kind, which expresses the probability with which an OOS production unit is accepted as WS. The power of the test characterizes the precision of the test, i.e. the reliability with which an OOS result is recognized as such. In addition to

γ and the given α, it also depends on the random-sample size n, as β, in turn, depends on α and n. The question is, how large must the random-sample size n be at a minimum for the null hypothesis, at a given reference value $\gamma_1 < \gamma_0$ to be abandoned with the power $1 - \beta$ and the production unit therefore to be considered OOS.

As deduced in the online attachments, n must be at least as great as the smallest whole number for which the following applies:

$$n \geq \frac{\left(z_{1-\alpha}\sqrt{\gamma_0 \cdot (1-\gamma_0)} + z_{1-\beta}\sqrt{\gamma_1 \cdot (1-\gamma_1)}\right)^2}{(\gamma_1 - \gamma_0)^2} \qquad (4)$$

Calculation example:

$$\alpha = 0.05; \quad \beta = 0.05; \quad \gamma_0 = 0.7; \quad \gamma_1 = 0.8.$$

$$n \geq \frac{\left(1.645 \cdot \sqrt{0.7 \cdot 0.3} + 1.645 \cdot \sqrt{0.8 \cdot 0.2}\right)^2}{(0.8 - 0.7)^2} \Rightarrow n \geq 200$$

For the given parameters the size of the random sample therefore should comprise at least 200 values in order for the batch, at $\gamma_1 = 0.8$ and the threshold $\gamma_0 = 0.7$, to be significantly recognized as WS. For $\gamma_1 = 0.96$, for example, a random-sample size of $n = 18$ (17.2) suffices in order to consider the batch to be WS with the required reliability and for $\gamma_1 = 0.99$, n is only 11 (10.01). This means, the better the quality of the batch (the greater the difference between the required minimum quality and the verified quality), the smaller n becomes at the same reliability! This is relevant in pharmaceutical production insofar as due to the consistently high production quality of today's production lines, the required quality is usually far exceeded.

Table 1 (see Section 2.2.2) cites the required random-sample sizes n and the thresholds k_0 for different combinations of γ_0 and γ_1 and different values of α and β to allow a test determination to be made. These data show that small random samples allow for only relatively inexact statements. It should also be pointed out that γ_0 and γ_1 are merely indicators for the actual percentage γ of WS results in the total population. If the null hypothesis is assumed, it can be stated with a reliability of $1 - \alpha$ that γ is greater than γ_0. If the null hypothesis is abandoned, however, it can be stated with a reliability of $1 - \beta$ that γ is smaller than γ_0. For five random-sample values of which at least four are WS, it can thus be stated with a reliability of 90% ($\alpha = 0.1$) that γ is greater than 0.5. If only three of the five values are WS, it can stated with a reliability of 80% ($1 - \beta = 0.8$) that γ is smaller than 0.9. More cannot be expressed by means of this test method for a random-sample size of 5.

Table 1
Required random-sample sizes n and the thresholds k_0 for different combinations of γ_0 and γ_1 and different values of α and β to allow a test determination to be made

	α (%)	β (%)	n	k_0
	20	20	3	3
	10	20	5	4
$\gamma_0 = 0.5$,	10	10	7	6
$\gamma_1 = 0.9$	5	20	8	7
	5	10	10	8
	5	5	11	9
	20	20	5	4
	10	20	9	8
$\gamma_0 = 0.6$,	10	10	12	10
$\gamma_1 = 0.9$	5	20	13	11
	5	10	16	13
	5	5	19	15
	20	20	11	9
	10	20	18	16
$\gamma_0 = 0.7$,	10	10	24	20
$\gamma_1 = 0.9$	5	20	26	23
	5	10	33	28
	5	5	39	33
	20	20	35	30
	10	20	59	52
$\gamma_0 = 0.8$,	10	10	81	70
$\gamma_1 = 0.9$	5	20	83	73
	5	10	109	95
	5	5	133	114

If the null hypothesis is assumed, it can be stated with a reliability of $1 - \alpha$ that γ is greater than γ_0. If the null hypothesis is abandoned, it can be stated with a reliability of $1 - \beta$ that γ is smaller than γ_0.

2.1.2. *Confidence intervals for the percentage γ of WS results in the total population*

The test determination may be important, but it is not fully satisfactory. In addition, one wants to make more-precise statements with a given reliability on the actual percentage γ of WS results in the total population. This can be done by determining a confidence interval for the unknown value γ based on the number k of WS results among n random-sample values at a given confidence probability $1 - \alpha$. A distinction is drawn between a one-sided upper, a one-sided lower and a two-sided confidence interval. The one-sided upper rank is specified by the value $\gamma_{\min}$, which is selected in a way to be able to state with the reliability $1 - \alpha$ that γ is greater than $\gamma_{\min}$. This reliability refers to the hypothetical, indefinitely repeated random sampling and determination of $\gamma_{\min}$. In this indefinite series of $\gamma_{\min}$ values the percentage $1 - \alpha$ is smaller than the actual percentage γ of WS results in the total population and only the percentage α is greater than γ. Analogously, the one-sided lower confidence interval for the confidence $1 - \alpha$ is specified as a value $\gamma_{\max}$, which is selected in a way to allow to state with the reliability $1 - \alpha$ that the actual value γ is smaller than $\gamma_{\max}$. Both values together define a two-sided confidence interval for γ. This interval overlaps with a given reliability the actual value γ. The reliability of this statement is $1 - 2\alpha$, as the error probability for the specification of $\gamma_{\min}$ and $\gamma_{\max}$, respectively, is α.

Tables 6–16, which are cited very often in the following stanzae can be viewed as well as explaining formulas in respect to better understanding in the additional online attachment, see therefore http://www.pharmchem.tu-bs.de/forschung/waetzig/dokumente/courtesy_translation.pdf or http://www.pharmchem.tu-bs.de/forschung/waetzig/support/.

Tables 7–10 (see the online attachment, A–D) cite the lower limits $\gamma_{\min}$ of the upper confidence level for the different values of α, n and k, and Tables 11–14 the upper limits $\gamma_{\max}$ of the lower confidence interval. The formulas used to calculate these values are given in the online attachment, too. Tables 7–10 show that with small random-sample sizes, only relatively imprecise statements can be made on the percentage of WS results to be expected in the total population. For a random sample comprising $n = 3$ values, of which all are WS ($k = 3$), for example, it can be stated with a reliability of 80% that the percentage γ is greater than 0.585, with a reliability of 90% that it is greater than 0.464, and with a reliability of 95% that it is greater than 0.368. For a random-sample size of $n = 10$ one can make more precise statements. If all random samples are WS, one can state with a reliability of 80% that the percentage γ is greater than 0.851, with a reliability of 90% that it is greater than 0.794, and with a reliability of 95% that it is greater than 0.741. This means that if one wishes to make relatively reliable statements ($1 - \alpha$ as high as possible), one has to accept that the statement itself becomes less and less precise.

If none of the values in a random sample is WS, then the lower limit $\gamma_{\min} = 0$, as it cannot be ruled out that all of the values in the total population are OOS as well. Analogously, in a random sample showing only WS results the upper limit $\gamma_{\max} = 1$, as all of the results in the total population could be WS as well.

If the number k of WS results in a random sample of the size n is not 0 or n, then $\gamma_{\min}$ and $\gamma_{\max}$ form the lower and upper limits of a two-sided confidence interval for γ for the confidence probability $1 - 2\alpha$. Tables 8 and 12 show that an actual percentage γ between 0.035 and 0.804 can be asserted with a reliability of 90% if one of three random-sample values is WS and two are OOS; with a reliability of 97.5% an interval of 0.008–0.906 can be asserted for γ (compare Tables 10 and 14). For 10 random-sample values of which 9 are WS and 1 OOS, it can be asserted with a reliability of 97.5% that the percentage γ amounts to between 0.555 and 0.997 (compare Tables 10 and 14). With an increasing random-sample size n the confidence interval for the probability γ becomes narrower, making the statement about the actual value for γ more exact.

The lower limit $\gamma_{\min}$ can serve to test the null hypothesis $\gamma \geq \gamma_0$ against the alternative $\gamma < \gamma_0$. One determines (e.g. from Tables 7–10 or by means of the formulas given in the online attachment) the lower limit $\gamma_{\min}$ for the number k of WS results observed the random sample of n values and for the specified α. The null hypothesis is accepted if $\gamma_{\min}$ is greater than γ_0, or else it is abandoned. It becomes clear that this test is equivalent to the test discussed in Section 2.1.1, i.e. that it leads to the same result. For example, for $n = 3$ and $k = 3$, with $\alpha = 20\%$, the lower limit $\gamma_{\min} = 0.585$. The null hypothesis $\gamma \geq 0.5$ is to be accepted. With α being 10%, WS can be asserted if all of four random-sample values are WS ($\gamma_{\min} = 0.562$). In order to make a determination with α amounting to 5%, at least five random-sample values of which all are WS would have to be available. This corresponds to the test method cited in Section 2.1.1.

2.1.3. Tolerance intervals for the total population with a given γ

In the treatment of OOS problems to date, a percentage γ_0 of WS results was specified for the total population, which had to be exceeded if the production unit was to be considered WS. The measurement values of a random sample served to test, with a specified error probability α or reliability $1 - \alpha$, respectively, whether this is correct. One can also take a different approach by determining the rank in which a given percentage of the total population lies. This type of rank is called the tolerance interval. Analogous to the confidence intervals (which must not be confused with the tolerance intervals), tolerance intervals can be defined as one-sided or two-sided. A one-sided upper tolerance interval for a given percentage of the total population, designated as γ (but not to be confused with the percentage of WS results in the total population), is bordered at the bottom by a value x_{lsl}, making the probability of measurement values $>x_{\text{lsl}}$ equal to γ. Accordingly, a one-sided lower tolerance interval for the percentage γ of the total population is bordered at the top by a value x_{usl}, making the probability of measurement values $\leq x_{\text{usl}}$ equal to γ. A two-sided tolerance interval for the probability γ is bordered at the bottom by x_{lsl} and at the top by x_{usl}, making the probability of measurement values lying between these two limits equal to γ (more precisely, the demand is for the probability of measurement values $\leq x_{\text{lsl}}$ and measurement values $>x_{\text{usl}}$ to be identical $(1 - \gamma)/2$). The limits of the tolerance intervals correspond to the quantiles ξ_{q} of the distribution function $F(x)$, which are defined by: $q = F(\xi_q)$. For the one-sided upper tolerance interval applies: $\gamma = 1 - F(x_{\text{lsl}})$, from which follows $x_{\text{lsl}} = \xi_{1-\gamma}$; for the one-sided lower tolerance interval applies: $\gamma = F(x_{\text{usl}})$, from which follows $x_{\text{usl}} = \xi_{\gamma}$; and for the tolerance interval applies $\gamma = F(x_{\text{usl}}) - F(x_{\text{lsl}})$ (or, more precise: $F(x_{\text{lsl}}) = 1 - F(x_{\text{usl}}) = (1 - \gamma)/2$), from which follows that $x_{\text{lsl}} = \xi_{(1-\gamma)/2}$ and $x_{\text{usl}} = \xi_{(1+\gamma)/2}$.

The distribution function is normally unknown and the limits of the tolerance interval have to be estimated based on the random-sample results $x_1, \ldots, x_n$. Since the random-sample values, ranked in order of size, contain all of the information on the distribution function $F(x)$ that can be derived from the random sample (they constitute a "sufficient" statistic for $F(x)$), it suggests itself to determine the limits of a tolerance interval based on these ranked random-sample values. If arranged from the lowest to the highest value, the rank of a random-sample value is designated as l (left rank), if arranged from the highest to the lowest value, as r (right rank). The random-sample value belonging to the left rank l is designated as $x_{[l]}$ and the value belonging to the right rank r as $x_{(r)}$. Any random-sample value $x_{[l]}$ having a low left rank number l can serve as the lower limit x_{lsl} of a two-sided tolerance interval, for example the lowest random-sample value $x_{[1]}$. According to the definition of tolerance intervals, this value is not part of the tolerance interval itself, it merely limits it at the bottom. Any random-sample value having a low right rank number r can serve as the upper limit x_{usl}, e.g. the highest random-sample value $x_{(1)}$. As these borders are determined based on random samples, the percentage of the total population included within these limits can only be stated with an error probability α or reliability $1 - \alpha$. The question is, what minimum percentage of the total population falls within the tolerance interval formed by $x_{[l]}$ and $x_{(r)}$ and how reliably can this percentage be asserted?

The question can be answered by means of the values $\gamma_{\min}$ in Tables 7–10 or the appropriate equation (see herefor the online attachment). It can be stated with the reliability $1-\alpha$ that the percentage of the total population in interval $x_{[l]} < x \leq x_{(r)}$ amounts to at least the value $\gamma_{\min}$, which for a given α and n corresponds to the value $k = n+1-(l+r)$. For a one-sided upper tolerance interval, $r = 0$ and the value $x_{[l]}$ is taken as the lower limit. For a one-sided lower tolerance interval, $l = 0$ and the value $x_{(r)}$ serves as the upper limit.

Example 1. The method is to be demonstrated based on the following data.

Random sample	0.13	0.49	0.69	0.39	0.50	0.09	0.24	0.13	0.56	0.57
Arranged values	0.09	0.13	0.13	0.24	0.39	0.49	0.50	0.56	0.57	0.69
Left rank l	1	2	3	4	5	6	7	8	9	10
Right rank r	10	9	8	7	6	5	4	3	2	1

The lowest value $x_{[1]}$ is 0.09. If it serves as the lower limit of a one-sided upper tolerance level, one can state with a reliability of 95% that the percentage of the total population exceeding 0.09 is at least 0.741. This is the value $\gamma_{\min}$ found in Table 9 (online attachment) for $n = 10$ and $k = 10$ ($= n+1-(l+r)$ for $l = 1$ and $r = 0$). If the second-lowest value $x_{[2]} = 0.13$ serves as the lower limit of the tolerance interval, only the minimum percentage $\gamma_{\min} = 0.606$, which can be found in Table 9 at $n = 10$ and $k = 9$, can be asserted with a reliability of 95%. The same $\gamma_{\min}$ can be asserted with a reliability of 95% for the percentage of the total population lying between the two-sided tolerance interval formed by the lowest value $x_{[1]}$ and the highest value $x_{(1)}$ (as in this case $l = 1$ and $r = 1$ and thus $n+1-(l+r) = n-1$). The highest random-sample value is 0.69. It can therefore be stated with a reliability of 95% that the probability of measurement values from the production unit being higher than 0.09 but not higher than 0.69, is at least 0.606 (Table 2).

The selection of the ranked values l and r to form the tolerance interval depends on the random size n. For small random-sample sizes n (smaller than or equal to 10) one will select $l = 1$ and $r = 1$, i.e. the lowest and highest random-sample value as the lower and upper limit of the tolerance interval, respectively. However, these extreme values can fluctuate greatly from random sample to random sample. They can also represent "outliers". If such outliers are identified as measuring errors or other errors, they are to be eliminated from the random sample. If they do not represent errors, they can account for very wide tolerance limits. The impact of outliers on the determination of the tolerance limits is reduced if the tolerance interval is not formed by the lowest and highest random-sample values but by values tending more toward the middle of the random sample. For random-sample sizes between 10 and 50, the second- or third-lowest and second- or third-highest random-sample values can serve as the tolerance limits; for larger random-sample sizes even higher l and r ranks. When selecting l and r, one has to weigh the robustness of the estimated values against outliers and the precision with which the tolerance probability can be stated.

Table 2
Minimum percentage of future values in the prediction interval in relationship to the random-sample size n and the confidence probability $1-\gamma$

n	$1-\gamma = 0.8$	$1-\gamma = 0.9$	$1-\gamma = 0.95$
5	0.5098	0.4160	0.3426
10	0.7290	0.6631	0.6058
15	0.8133	0.7644	0.7206
20	0.8576	0.8190	0.7839
30	0.9035	0.8764	0.8514
40	0.9270	0.9062	0.8868
50	0.9413	0.9244	0.9086
60	0.9509	0.9367	0.9234
70	0.9578	0.9456	0.9340
80	0.9630	0.9522	0.9421
90	0.9671	0.9575	0.9484
100	0.9704	0.9617	0.9534

How can a prediction regarding the percentage γ of WS results in the total population be made based on the tolerance intervals? One possibility is to form the tolerance interval taking the lowest and highest values and to assume the production unit as WS if this tolerance interval lies completely within the specification range and if for the share of the total population lying within the interval a value of at least $\gamma_{\min} > \gamma_0$ can be asserted with a reliability $1-\alpha$. This means that the production unit can be declared to be WS only if all random-sample values lie within the specification range and the random-sample size n is large enough to be able to assert a value $\gamma_{\min} > \gamma_0$ with a reliability $1-\alpha$ This is a conservative rule of decision. There is a high risk for batches that are actually WS to be erroneously rejected. The alternative is to form a tolerance interval only for the WS results of the total population. If only a one-sided lower specification limit is given, a one-sided upper tolerance interval is formed using the lowest value $x_{[l]}$ that still conforms to the specification. The minimum share $\gamma_{\min}$ of the total population of WS results that can be asserted for this interval with a reliability $1-\alpha$ is the value belonging to α, n and $k = n+1-l$ found in Tables 7–10. This follows from the fact that of n random-sample values, $l-1$ are OOS and $n+1-l$ are WS if the lowest random sample value within specification has the left rank number l. If only one one-sided upper specification limit is given, a one-sided lower tolerance interval is formed using the highest value $x_{(r)}$ that is still within the specification. For this interval it is possible to assert with a reliability $1-\alpha$ at least the probability $\gamma_{\min}$ that can be found in Tables 7–10 for the given α and n for $k = n+1-r$, as in this case $r-1$ random-sample values are OOS and $n+1-r$ are WS. If a specification interval is given, a tolerance interval is formed for WS results with the lower limit formed by the smallest random-sample value $x_{[l]}$ that is smaller than the lower specification limit and thus OOS, and whose upper limit is formed by the highest random-sample value $x_{(r)}$ that is either smaller than or equal to the upper specification limit and thus still. If the lowest random-sample value is greater than lower specification limit, this limit serves as the lower limit of the tolerance interval and $l = 0$. The minimum probability that can be asserted for this tolerance interval with a reliability $1-\alpha$ is the value $\gamma_{\min}$ that can be found in Tables 7–10 (see online section) for a given α for $k = n+1-(l+r)$. The production unit is considered WS if $\gamma_{\min}$ is greater than γ_0.

Re Example 1. It is assumed that for the random sample of Example 1 a one-sided upper specification range with a lower limit of 0.1 is specified. Of the 10 random-sample values, the second-lowest value $x_{[2]} = 0.13$ is just higher than this specification limit. This value serves as the lower limit of a one-sided upper tolerance interval. For $n = 10$, $k = 10 + 1 - 2 = 9$ and $1 - \alpha = 95\%$, the $\gamma_{\min}$ value depicted in Table 9 is $\gamma_{\min} = 0.606$. It can thus asserted with a reliability of 95% that at least the percentage 0.606 of the total population conforms to specification. If the production unit is considered to be within specification if the probability of WS results is greater than 0.6, the production unit from which the random sample was taken can be classified as being WS with a reliability $1 - \alpha$ of 95%.

2.2. Calculations on the basis of a normally distributed or t-distributed total population, respectively

2.2.1. Analytical examination of homogeneity—examination using the prediction range of the t-distribution

How can information from a random sample be used to evaluate the total population? Let us examine another example:

Example 2. 10,000 infusion solutions are produced of which the contents of 5 are each to be examined once during an in-process control. Based on the results of the random sample of these 5 solutions the question of how many of these 10,000 are OOS is to be answered.

The question is answered by the prediction range prd(x):

$$\mathrm{prd}(x) = \bar{x} \pm t_{\alpha,n_1+n_2-2} \cdot \hat{\sigma}_{\mathrm{ges}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (5)$$

Here $\bar{x}$ is the mean value of the random sample n_2 and $\hat{\sigma}_{\mathrm{ges}}$ the standard deviation estimated from it. The number n_2 stands for the size of the random sample for which $\bar{x}$ and $\hat{\sigma}_{\mathrm{ges}}$ were determined. n_1 indicates the size of a future random sample whose result is to be predicted. If n_1 future measurements are performed, their mean value lies in the prediction range prd(x) with a probability of $1 - \alpha$. The correction factor t_{α,n_1+n_2-2} takes into consideration that the standard deviation for small data numbers can be estimated with only low certainty. Therefore it is not the normal distribution itself that is used to calculate the prediction interval but the highly similar but somewhat broader t-distribution [5,6].

As the degree of freedom for the calculation of the quantile of the t-distribution, the sum of the random-sample numbers reduced by 1 is used here, i.e. of the already taken and of the future random sample, since the size of the future random sample also influences the size of the prediction range. The prediction range indicates the position of the mean value of a future random sample from the same total population and becomes narrower the greater $n = n_1 + n_2$ is. It does not matter whether random sample n_1 or n_2 will be greater.

As this question deals with the prediction of the properties of an individual infusion solution, $n_1 = 1$ (compare [1]). The prediction range indicates in which area almost all of the values of the total population can be found, except for a very small share α based on error probability. A high α results in a narrow prediction range, a low α in a wide prediction range. The error probability α can be selected to assure that only very few elements of the total population will lie outside the prediction range. Prerequisite for the calculation of the prediction range in accordance with Eq. (5) are normally distributed or, for smaller data numbers, t-distributed data.

Why can $n_1 = 2$ not be asserted if dual determinations of each examined solution are performed later on? From a purely mathematical point of view this is permitted. The prediction range predicts where the mean values of the respective dual determinations will lie. The thus obtained prediction range of the mean values is smaller than the prediction range of the single values. In this case, however, the dual determinations are made for one solution. The variability to be examined is not only due to the measurement error but also by the compounding. If dual determinations for two solutions were made and their values reported, the problem would not be properly addressed, as the objective is to evaluate potential deviations of individual products.

If the prediction range overlaps the specification limit (Section 1.2.2, Fig. 2c), a part of the elements of the total population lies within, another outside the specification range. Thus, if the mean value is WS and the prediction range does not overlap the specification limit, one can assume that most of the elements of the total population are WS. “Most” is concretized by the error probability α.

Example 3. The release specification for a content determination is prescribed as 95.0–105.0% of the declared content. A determination from two different infusion bottles yields the measurement values 94.5 and 96.0%. Without needing to perform a calculation it becomes apparent that the result is OOS, if the reportable result is the single measurement. Even if the mean value is still slightly above the lower specification limit, the prediction range clearly overlaps the specification range. It is clear that another, i.e. third, measurement value could again lie at 94.5% with a very high probability. What is the situation if many additional WS values are measured? When would the batch be considered WS? (Continued as Example 4 in the additional online Section 2.4.3).

2.2.2. Comparison of the calculations using non-parametric test methods

Eq. (5) (Section 2.2.1) is easy to deal with and allows a quick estimate of future values. The share of measurement values within the interval, however, is only correct in respect of the mean. If $\alpha = 0.05$ (5%) is selected, the calculated share of measurement values expected to be in the prediction range $1 - 2\alpha = 0.9$ or 90% (2α is selected if the problem is two-sided, i.e. if extreme measurement values can lie above or below the mean value). This percentage of 90% is only correct in respect of the mean; in half of the cases the percentage will be slightly exceeded (it can be 92%, for example), in the other half it will not be reached. The deviations to be expected depend on the random-sample numbers, whose size is unknown. It is not possible to determine on the basis of the t-distribution how improbable a great deviation of 90% (e.g. 69%) is.

It is therefore more elegant to cite a prediction range containing, with a given reliability $1-\gamma$, at least the percentage $1-2\alpha$ of the distribution (compare Section 2.1). If one then selects α and $\gamma=0.05$, one can state: in 95% of the cases the prediction range will overlap at least 90% of the measurement values.

In order to be able to actually use the optimized assertion possibilities of such empirical prediction intervals, larger random-sample sizes than those used in t-based calculations are required. If one wishes to guarantee with a confidence probability of 80% that at least 90% of all future values will be within the interval limits, 29 measurements are needed (see Table 1). If the confidence probability is to measure 90%, as many as 37 measurements are needed. These large random-sample numbers are in part necessary because no assumptions regarding density distribution of the measurement values are made during such rank tests, i.e. potential density distribution information cannot be relied upon. But even if the classical prediction interval is selected, which is based on the t-distribution (Section 2.2.1, Eq. (5)), the assertion is only correct in respect of the mean (Tables 3 and 4).

This discussion shows that for the typical random-sample sizes used and having stood the test in practice for years, only assertions with a relatively low confidence probability can be made, which, however, is not viewed as problematic. If large random-sample sizes can be easily processed, for example by rapid analysis methods, the gain in information in the area of confidence probability would be enormous.

2.3. Discussion of the employed methods

The results of the distribution-free calculations (Section 2.1) and of the calculations based on a t-distribution (Section 2.2) are very similar if the data number is high ($n \geq 30$; compare Table 1) or if the data number is low and a part of the measurement values lies close to the specification limit (compare data sets C1–C3, D1 and D2 in Fig. 2). This may be the reason why false conclusions based on the t-distribution (greater deviations from the safety of 90% applying only on average, compare Section 2.2.2) are relatively rare in pharmaceutically relevant data sets. For very small data numbers (e.g. $n=2$ or $n=3$) a reasonable statement is not possible for either of the two methods. The additionally existing problem of outliers is separately discussed within the online addition to this publication [7,9,10].

Table 3
Hundred single values of the determinations (in ppb) for Example 8

n	$y(n)$	$y(n+20)$	$y(n+40)$	$y(n+60)$	$y(n+80)$
1	300.35	302.17	298.59	300.69	300.36
2	302.83	298.83	300.88	301.18	298.75
3	300.39	301.62	299.39	301.60	300.61
4	301.36	300.26	298.68	299.78	301.26
5	299.72	300.40	301.07	299.47	301.24
6	297.47	300.29	301.55	300.01	300.95
7	298.99	300.69	301.42	301.93	300.90
8	299.57	299.97	300.35	301.27	298.18
9	298.92	301.80	301.20	298.22	300.25
10	298.36	300.03	301.66	300.22	300.71
11	300.22	300.69	299.77	300.09	302.53
12	298.50	299.67	298.01	300.12	299.27
13	298.92	298.07	299.47	300.99	299.77
14	299.37	301.08	300.06	300.03	299.89
15	299.69	299.54	298.81	300.91	299.34
16	300.29	300.09	301.85	297.64	298.48
17	299.81	301.56	302.10	299.08	299.68
18	301.35	303.13	300.19	300.21	299.68
19	299.00	298.51	297.49	300.07	300.95
20	298.28	300.36	299.88	301.72	300.11

Table 4
Cumulative mean values for Example 8

k	Mean value (1. . .k)
1	300.35
2	301.59
3	301.19
4	301.23
5	300.93
6	300.35
7	300.16
8	300.09
9	299.96
10	299.80
11	299.83
12	299.72
13	299.66
14	299.64
15	299.64
16	299.68
17	299.69
18	299.78
19	299.74
20	299.67

If all measurement values are WS and they lie far from the limit value, and if the data number is small (e.g. $n=5$, compare the data set D4 in Fig. 2), then the two discussed calculation methods yield different results. With the distribution-free calculation the measurement values are reduced to their qualitative information of WS or OOS, resulting in loss of information. With this method, therefore, there is no influence on the calculated maximum OOS percentage if a result lies only slightly or very far outside the specification limit. This is unsatisfactory and speaks for the use of t-based calculations in such cases. These t-based calculations, on the other hand, do not provide concrete data on the reliability of the information, which is also unsatisfactory. Unfortunately there is no better method at the present time to evaluate such data sets. It is therefore necessary to look for ways to optimize the methods. It may be possible to gain additional information from the status of the total population. The t-based calculation may possibly be used to develop a preliminary estimate of γ_1, which would allow the use of Eq. (4) instead of Eq. (2) (compare Section 2.1), and statements with clearly defined statistical reliability would be possible even for smaller data numbers. This will be a subject of future exploration.

This article is supplemented with further online attachments, with the additional Section 2.4 “Review of additional examples and calculations” and also a complete additional part 3 “Retesting after OOS Results (Retesting/Resampling) and Outlier Treatment”. All those very meaningful additions complete the publication in respect to establish better understanding and can be found at: http://www.pharmchem.tu-bs.de/forschung/waetzig/support/.

3. Conclusion and outlook

3.1. Conclusion

Not every measurement value outside the specification limit constitutes an OOS result for the entire data set. OOS results can occur at random. For very large data numbers it is even expected that single values will sharply deviate from the mean value and possibly lie outside the specification limits, as the normal distribution is generally unlimited. This fundamental statement applies initially to large data sets. The mean value often meets the specification even in smaller random samples if single values are OOS.

If the homogeneity of the random sample is of critical importance, i.e. even if the single values have to lie within given specification limits, it is still possible for a single value to be unusually high or low. For the frequently used small data numbers ($n < 10$) a single OOS result points strongly to an overall OOS result. In other words: a “random” OOS result spoils the entire random sample in such cases.

Multiple measurements are often worthwhile, especially if measurement values and specification lie close to each other. Dual determinations lead to very large prediction intervals. Here, it is often impossible to assure compliance with the specification, even if either measurement value on its own meets the specification. OOS results can usually be prevented in the forefront by using suitable data numbers during the planning phase of the tests.

Taking measurements until the desired result is achieved (testing into specification) is not permitted. Sequential random-sampling plans, however, make sense in many cases. Important is the random-sampling plan, which should be previously specified in a standard operation procedure (SOP). As problematic as multiple measurements with arbitrary termination is the non-permitted elimination of extreme values. When dealing with extreme values it must be kept in mind that the error search cannot be replaced with statistical methods—no matter how sound and established they may be. Still, outlier tests serve as an important tool, though mainly of a diagnostic nature. When relying on outlier tests, SOPs are especially important. The uniform evaluation of extreme values in data sets is permitted but requires the consensus of all departments involved.

3.2. Outlook and additional important aspects regarding the problem of OOS results

This article does not address in greater detail the variability in the spread of measurement data. Although the uncertainty in determining the standard deviation has been implicitly considered through the used *t*-factors, it is also important to question whether the spread of a process or measurement method is acceptable at all. The long-term documentation of results is an important quality assurance measure by which OOS results can often be prevented in the early stages. Control cards and trend analyses can serve as aides in this effort. All considerations regarding *t*-statistics are based on a normally distributed total data population. It is therefore interesting to examine the cases in which great deviations from the normal distribution can be expected. An important case group is certainly data sets containing outliers. In-depth discussion of outlier problems will remain important into the future. It would be beneficial to conduct this discussion in a broad forum and with the aid of many example data sets.

Acknowledgements

We wish to thank I. Astner (GAA Braunschweig) and H. Gifhorn, Dr. M. Graf (Amcapharm), Dr. J. Grünefeld (TU Braunschweig) and Dr. U. Schepers (Bayer Healthcare) as well as Dr. H.-U. Amschler (Ministry of Labor and Health, Kiel) and Dr. H. Häusler (Boehringer Ingelheim, Ingelheim), who through their competent review and critical comments greatly contributed to the generation of this article. Special thanks also to Siglinde Rach for her big support in improving the language. Further we appreciate Dr. Graf’s assistance in generating the Figures. Last but not least we are very grateful to Dr. B. Renger (Vetter Pharma), who encouraged us to write this article.

References[1]

[1] U.S. Department of Health and Human Services, Food and Drug Administration (FDA), Center of Drug Evaluation and Research (CDER), Guidance for Industry, Investigating Out-of-Specification (OOS) Test Results for Pharmaceutical Production, http://www.fda.gov/cder/guidance/3634fnl.pdf.

[2] H. Häusler, M. Niehörster, K.P. Wörns, Pharm. Ind. 61 (1999) 935–939.

[3] Out-of-specification issues (OOS)—Special Report, Institute of Validation Technology, http://www.ivtconferences.com/pdf/0906_oos.pdf.

[4] R. Schmidt, FDA-/GMP-konforme Bearbeitung von OOS-Ergebnissen, Presentation in Course Nr. 405 der Arbeitsgemeinschaft für Pharmazeutische Verfahrenstechnik (APV), http://www.apv-mainz.de.

[5] J. Hartung, Statistik, Oldenbourg Verlag München, 8. Aufl. 1991.

[6] H. Wätzig, in: E., Nürnberg, P., Surmann (Eds.), Hagers Handbuch der Pharmazeutischen Praxis, vol. 2, Springer Verl., Berlin/Heidelberg/New York, 5. Aufl., 1991, pp. 1048–1084.

[7] B. Renger, Pharm. Ind. 61 (1999) 1053–1055.

[8] I. Stewart, Irrfahrt zum Mean value, Spektrum der Wissenschaft (1999) 112–114.

[9] K. Baumann, Process Control Qual. 10 (1997) 75–112.

[10] P.J. Rousseuw, A. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.

[11] Expert Group Pharmaceutical Analysis/Quality Control of the German Pharmaceutical Society, Comments and Suggestions Regarding the FDA/CDER Draft Guidance for Industry, Investigating Out-of-Specification (OOS) Test Results for Pharmaceutical Production, see http://www.fda.gov/cder/guidance/3634fnl.pdf.

[1] The author has copies of citations listed as Web addresses in Ref. [1]. If the Web page no longer exists, the documents can be requested from the author via e-mail. The interested reader will find continuative attachments and varied complementing examples, as well as the additional online Section 2.4 and an additional part 3 at following web site http://www.pharmchem.tu-bs.de/forschung/waetzig/dokumente/waetzig-publ3.pdf.

Index of symbols and abbreviations

α: probability of an error of the 1st kind, i.e. to discard the null hypothesis despite its applicability; producer risk

β: probability of an error of the 2nd kind, i.e. to discard the alternative hypothesis despite its applicability; consumer risk

γ_0: threshold value for the percentage of WS values in the production unit required to be reached at a minimum in order for the production unit to be considered to conform to specification

Bias: difference between the true value and the analysis result

Collective: see total population; term according to R. v. Mises

Confidence interval: range in which the (unknown) value for the actual percentage of WS results in the total population can be found with the probability $1-\alpha$

k_0: threshold value for the percentage of WS values in the random sample required to be reached at a minimum in order for the production unit to be considered to be within specification

Measurement range: range from the lowest measured value to the highest measured value

n: size of the random sample

N: size of the total population

OOS result: out-of-specification result; result lying outside the specification limits

Outlier: (extreme) values due to errors

Power: "test power" of a method; indicates the certainty with which an OOS result is recognized as such, $1-\beta$

Precision: measure for the spread and the congruence of the measurement values after the multiple performance of an analysis (within one measurement series)

Prediction interval: range in which the mean value of a future random sample will lie with a certain probability ($1-2\alpha$ for 2-sided intervals)

Production unit: production batch, lot

Range: see measurement range

Random sample: measurement values obtained for a sample of the size *n* taken from a total population; $x_1, x_2, \ldots, x_n$

Robustness: measure for the independence of the analysis results of a method after minor changes to the measuring system

Test power: see power

Tolerance interval: measurement range in which a specified percentage of the total population can be found

Total population: all of the elements of a quantity to be examined, e.g. all tablets in a batch, number of all possible measurement values obtained by one method

WS values: within-specification values; values lying within the specification limits

$z_{1-\alpha}$: $(1-\alpha)$ quantile of the standard normal distribution (mean value 0, variance 1)