Additional part 3 to

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 of Technical University Braunschweig

[b]Institute for Biometry of MH Hannover

1) Content

Part 3: Retesting after OOS Results (Retesting/Resampling) and Outlier Treatment

I Retesting after OOS results
(Retesting / Resampling)

If a test yields a narrow OOS result it is tempting to show by means of a higher data number that the deviations occurred purely at random. Is it still possible to meet the specification by generating extensive data material retroactively? What must be considered in such cases?

Example 9:

Performed are 100 analytical determinations (see Table 3); relevant is the mean value. As one can see, many of the values are above the specification. Mean value and standard deviation for all n = 100: 300.12 ± 1.20 (e.g. ppb of a toxic ancillary component); 300 ppb is the limit.

Figure 4 shows clearly that even random samples with a very high percentage of OOS values can simulate a WS result after selective cumulation of the mean values of high data numbers.

II Outlier treatment

II.1 General remarks

Outlier tests serve to identify data in a data series that are to be viewed as extremely improbable [2]. In such cases there will always be an error investigation. Even if the error cannot be clearly identified as a writing, dilution or weighing error, for example, it makes sense in many cases not to take into consideration measurement values that are likely erroneous (compare Example 5). On the other hand, the careless handling of outlier tests can quickly lead to manipulated, no longer reliable data sets. Especially for small data sets it is often difficult to say whether a measurement value severely deviating from the mean value is still representative of the measurement-value distribution or whether another systematic error, for example a dilution or weighing error, additionally influences individual values.

The rule states: If a single value is OOS, then the entire data set is OOS if n is small (see Chapter 2.2).

The previous paragraph warned specifically of the dangers of „testing into specification“ (compare Example 8). If a specific test result is desired, the analyst conducting the analysis will develop prejudices against a part of the measurement values: „These values can't possibly be correct...“. The arbitrary selection of measurement values, however, is strictly forbidden. There are many examples where extreme measurement values first interpreted as outliers, later turned out to be important indicators of major deviations in production or analysis. Outlier elimination thus always harbors the great danger of a **bias**, i.e. of intended and even more often unintended manipulation. The FDA therefore specifies that outliers, as a rule, may not be eliminated ([1], 4th paragraph: Outlier tests): “If no laboratory or statistical errors are identified, there is no scientific basis for invalidating initial OOS results in favor of passing retest results. All test results, both passing and suspect, should be reported and considered in batch release decisions.” **„**Outlier tests have no applicability in cases where the variability in the product is being assessed“. In the following paragraphs these statements are put into better perspective: In “biological tests,” outlier tests are useful if they are typical for a problem. Here, the number of sources for systematic errors is so great that single errors can rarely be investigated within a reasonable scope. Without outlier tests the spread of measurement-values is often extremely high. Useful analysis is no longer possible. Moreover, the often high data number in biological tests facilitates the use of outlier tests. It is reiterated in the conclusions: “Statistical treatments of data should not be used to invalidate a discrete chemical test result.” Chemical tests are to be understood as separate from biological tests. They include all non-biological tests such as HPLC, NIR, UV, CE.... The formulation reads as follows: “In very rare occasions and only after a full investigation has failed to reveal the cause of the OOS result, a statistical analysis may be valuable as one assessment of the probability of the OOS result as discordant.” The clear preference of error investigation is considered highly important – an outlier selection based solely on statistics is always unsatisfactory [1]. At best, the outlier test serves as a diagnostic tool in identifying suspect values, which can then be examined in targeted fashion to isolate laboratory errors.

The errors are then identified and documented. On the other hand there are trivial errors (weighing, spilling, ...), which cannot be identified after the fact. And there are data sets which obviously would be falsely judged without outlier elimination. Outlier tests are permitted in error investigations but may not replace them [7].

Example 10:

(Continuation of Example 5)

If the value 91.1 (possibly a writing error but no longer traceable) is disregarded, the standard deviation is not 2.46 but approximately 0.1. It is obvious that this one value is unnaturally far removed from the other data. The data probably do not follow a normal distribution and can therefore not be evaluated by statistical means based on normal distribution.

The preliminary remarks about the ever present danger of manipulation when using outlier tests clearly show that it is difficult to find generally acceptable criteria for obvious deviations.

II.2 Outlier tests

There is a series of calculation specifications for outlier tests and the use of statistical methods that do not take outliers into consideration (robust statistics). These specifications have been described in detail and comprehensive fashion in textbooks and articles [5, 9, 10]. A selection is introduced in Examples 11 and 12:

Dixon's test ([5], page 346 ff.) is based on normal distribution: Considered is the probability of having an extremely high or extremely low value in the data set. In an example demonstration, the distance from the lowest or highest value to the secondlowest or second-highest value is correlated to the total range x_n-x_1 of the values. A simple calculation rule results:

$$r_{10}^{1} = \frac{x_2 - x_1}{x_n - x_1} \qquad \text{(Equ. 6)}$$

The associated limits are cited in the table beginning with n = 6 (Table 5):

One way to obtain a quick and clear picture of the distribution of a data set and to assess the potential of outliers, is the Box Plot ([5], page 825 ff, esp. 835-838). First, the measurement values are entered in a list or coordinates system arranged by size. Next, the median m and the **hinges** are determined, i.e. the medians of the lower and upper half of the data set, divided by the median [5]:

Example 10 – continued (see Figure 5):

91.1 97.0 97.1 97.15 97.2 97.2

The median measures 97.125. In this case it is the arithmetic mean of the values in the middle of the data set. This divides the list into 2 parts (with marked hinges):

91.10 **97.00** 97.10 and 97.15 **97.20** 97.20

The h-spread (spread at the hinges), i.e. the difference h_o - h_u, is a measure for the distribution of the data in the center of the data set. Here, this value measures 0.2. A value 1.5x or 3x h-spread distant from the hinge is considered an outlier or a blatant outlier.

In the given example, h_u –1.5 h-spread = 96.7, h_u - 3 h-spread = 96.4.

At 91.1, we are dealing with an extremely blatant outlier.

Example 10 shows beautifully how the tests function. The obtained results are clear. In contrast, the following example is a typical borderline case:

Example 11:

A content determination is performed 8 times. Obtained are the values 93, 98, 98, 98, 96, 96, 98 and 99 (mean value 97.57; standard deviation 1.927). The data set does not meet the specification of 95 - 105.

Is the first value an outlier? It appears possible. If the first value is not taken into consideration, the standard deviation measures only 1.134, and the data set is WS.

It is difficult to evaluate this data set intuitively. The data number is small. Could the next value measured be, for example, 94? This does not seem improbable – and with an additional measurement value of 94, the value 93 would likely no longer constitute an outlier. And now to the described tests...

Dixon's test (see Equation 6) does not point to an outlier:

$$\frac{x_2 - x_1}{x_n - x_1} = \frac{96-93}{99-93} = \frac{3}{6} = 0.5 < 0.59 = Dixon\acute{}s \;\; r_{10,8,0.99}$$

The box plot could be drawn as shown in the example (Figure 6). The h-spread is 2; the lower limit for outliers is calculated at 96 - 1.5 * h-spread = 93; i.e. the value 93 lies just at the border. Since the data number is low, the result seems to indicate that the value 93 belongs to the data set. If 91 had been measured instead of 93, the indication would have been an outlier. Still, 93 is a suspect value which should be examined more closely for laboratory errors. If no errors are found, the value should remain in the data set. It is then evaluated as OOS.

II.3 Selection of distribution or outlier tests

Which outlier test delivers correct results? In borderline cases this can never be stated with certainty solely on the basis of the data. An error search is always crucially important. If outlier tests are to be used at all, it is particularly important to specify the planned procedure in advance in Standard Operating Procedures (SOPs).

If in fact several tests (for purely informational purposes) are used and only one states that outliers cannot be clearly confirmed, outliers are not present. Intuitively obvious outliers will yield a clear result in all outlier tests (compare Examples 10 and 11).

Which outlier tests are recommended for an SOP? The calculation specifications for the tests are often too abstract to allow a quick judgment of whether the tests will yield intuitively convincing assessments. It is therefore best to take the reverse approach. First, the available data sets are evaluated by means of borderline cases. A part of the data sets should (preferably in the opinions of as many experts as possible) contain outliers, another part should not. Once the experts have agreed on the assessment of the data sets, the next step consists of the specification of outlier tests, which (preferably by means of simple calculation specifications) should lead to the same results.

In order to arrive at a uniform assessment of outlier tests, a similar approach can be suggested. The following shows a few data sets with the subjective evaluation of the author. It is recommended to gather additional data sets that were evaluated by as many experts as possible. In the end it may be possible to specify conservative conditions for outlier tests. According to these conditions, not a single value should be allowed to be eliminated as an outlier that would have been kept in by even just one expert. This means, though, that some outliers may not be detected by the tests. For these cases, individual SOPs can be developed.

II.4 Discussion of example data sets

A few examples have already been introduced: Examples 6 and 10, resp.:
Measurement values 91.1 97.1 97.2 97.2 97.15 97.0
Specification limits 95 - 105; 91.1 outlier, result of the data sets: WS.

How big would the first value have to be at a minimum for it no longer to be recognized as an outlier? According to my intuition: 96.8. This also corresponds nicely to the assessment that would be obtained by using Dixon's or the box plot test (compare Example 10).

Example 12:

93, 98, 98, 98, 96, 96, 98, 99; 95 - 105 to be complied with: 93 not an outlier, the examined total population is OOS.

How small would the first value have to be at a minimum in order for it to be recognized as an outlier? According to my intuition: 91. This, too, fits nicely the assessment that would be obtained by using Dixon's or the box-plot test (compare Example 10): The Dixon r-value is (96 - 91) / (99 - 91) = 0.625. It is thus greater than the critical value for n = 8 (0.59, compare Table 5). The hinges remain 96 and 98, the h-spread remains 2, the limit for outliers is therefore 93 (see above), for blatant outliers 90. If the 1st value is 93, it is not considered an outlier, and the data set is therefore OOS. If 91 is measured instead of 93, the measured value is recognized as an outlier and the data set is WS. Is one rewarded for the worse measurement (91 instead of 93) on top of it? No, for it is not known before the measurement that the value will be especially low. If one were to measure badly on purpose in order to obtain an especially low value, it would not only affect the low values, but also all others. This would make the distribution great and the data set would likely be OOS.

Example 13:

Even 89, 107, 107%: 89% is not an outlier, as it is possible that the same value of 107% was hit twice at random and the next value could again be 89% (compare Example 3). Outlier tests require a minimum data number: n = 6 both for Dixon's test and for the evaluation by means of the box plot.

Example 14:

In a statement regarding [1] the German Pharmaceutical Society recommends in a working paper a sequential sampling plan [11]. Initially, three determinations are to be made. If all are WS, the test is considered completed (compare, however, Figure 2, cases D1 and D2). If one value is OOS and two are WS, then 3 additional determinations are made. If all others are WS, the one OOS measurement value is considered an outlier and the data set is considered WS.

The first 3 measurement values are: 89.0 91.9 92.2.
The specification limits are 90 and 110.

A The next measurement values are: 92.0 92.5 92.2.

The total data set is first arranged by size:

89.0 91.9 92.0 92.2 92.2 92.5.

If the data set is evaluated in accordance with Equation 5, the data set is OOS. Is 89.0 an outlier? The intuition of the author, the test according to Dixon (r^{1}_{10} = 0.83, crit. value $r_{10,6;0.99}$ = 0.70) and the box plot (h-spread = 0.3, h_u = 91.9) suggest it.

B The next measurement values are: 90.6 90.2 91.5.

The total data set is first arranged by size:

89.0 90.2 90.6 91.5 91.9 92.2.

There is no reason to assume in this case that 89.0 is an outlier (r^{1}_{10} = 0.375; h-spread = 1.7, h_u = 90.2).

The sequential approach recommended in [11] is pragmatic and will yield plausible results in most cases. The recommendation needs to be more concrete, however, to also allow proper assessment of the cases shown in Figure 2 / D1 and D2 and in Example 14 B. As stated in additional part3, chapter II.3, the rules for the treatment of especially low or high values are best specified in the form of example data sets to be uniformly evaluated by experts. Please send me such (if necessary disguised) data sets to: (see the address at the end of this article, or via e-mail: h.waetzig@tu-bs.de). These data sets could then be introduced to a larger readership for discussion in a follow-up article.

Figures:

Figure 4: Cumulative Mean Values for the Single Values from Table 4

Figure 5: Box Plot for Example 10

Legend: *krasser Ausreißer = blatant outlier

Figure 6: Box Plot for Example 11

Tables

Table 5: Critical Values of the Dixon Test for the 1% Level (from [5])

Table 5

n	$r_{10,n;0.99}$
6	0.70
7	0.64
8	0.59
9	0.56
10	0.53
11	0.50
12	0.48
13	0.47
14	0.45
15	0.44
20	0.39
25	0.36
30	0.34