Additional Chapter 2.4 to

Out of Specification Test Results from the Statistical Point of View Part 2: Calculations: distribution-free or on the basis of a normally distributed total population

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

[a]*Institute for Pharmaceutical Chemistry of the Technical University Braunschweig*

[b]*Institute for Biometry of MH Hannover*

2.4 Review of additional examples and calculations

2.4.1 Additions to Chapter 1.2.3 / Figure 3

In Case 2 the random sample consists of 10 values of which one is greater than 10 ppm and thus OOS. The other 9 values are smaller than 10 ppm and thus WS. In this case one can assume with a reliability of 90% that at least a percentage of 0.663 of the total population is WS, and with a reliability of 95% that at least a percentage of 0.606 is WS. The hypothesis that the percentage γ of WS results in the total population is greater than 0.6 can thus be adopted with a reliability of 95%.

In Case 3 the random sample consists of 30 measurement values of which 2 are greater than 10 ppm and thus OOS. For this result it can be stated with a reliability of 80% that in the total population at least a percentage of γ = 0.86 conforms to specification. With a reliability of 90% it can be stated that γ is at least 0.83, and with a reliability of 95% that at least a percentage of 0.805 conforms to specification. If the threshold is set at γ_0 = 0.80, the production unit can be considered WS with a reliability of 95%.

In Case 4 the random sample consists of 100 values of which 5 are OOS. The frequency of WS results thus amounts to 0.95. It can therefore be stated with a reliability of 80% that the percentage of values in the total population conforming to specification is at least 0.928, with a reliability of 90% that it amounts to at least 0.914, and with a reliability of 95% that it amounts to at least 0.901. With a reliability of 95% the production unit can therefore still be considered WS if a threshold value of $\gamma_0 = 0.9$ is specified.

2.4.2 Additions to Chapter 2.1.1

A) Measured are the 5 values: 96.8 97.1 97.2 97.2 97.0.

The specification range includes the values from 95.0 to 105.0. The prediction interval ranges from the smallest measured value to the largest measured value, i.e. from 96.8 to 97.2.

Table 2 states that for a selection of a confidence probability $1 - \gamma$ of 80% (for 5 measured values) it can only be guaranteed that at least 50% (0.5098) of all future values will lie within the limits of this prediction interval.

B) Accordingly, one can assume with a probability of 95%, for a random sample of the size n = 50 whose values lie all within the specification range, that at least 90% (0.9086) of all future values will lie within the range between the smallest and largest value of the random sample, i.e. within the prediction interval of this random sample.

2.4.3 Additions to Chapter 2.2.1

Example 4:

(Continuation of Example 3)

After having obtained the measurement values 94.5 and 96.0%, the following additional measurements are performed: a) 3, b) 8 and c) 50. The result for each of these additional measurements is 96.0%. Can any of these data sets be considered to conform to specification? The associated lower limits of the prediction intervals (calculated on the assumption of a t-distribution) lie at a) 94.1% (n = 5), b) 94.95% (n = 10) und c) 95.6% (n = 52). With a high data number the individual value no longer plays an important role: the data set as a whole is WS, although a single value is OOS. At n = 10, a single OOS result still has a strong influence on the total result. In this example the specification is just barely not yet met.

Example 5:

Measured are the 6 values: 91.1 97.1 97.2 97.2 97.0 97.15.

Here, too, one value has a strong influence on the prediction interval. But this interpretation is highly unsatisfactory. The distribution is low if the measurement value 91.1 is not taken into consideration – perhaps it was a writing or transcription error (91.1 instead of 97.1)? If this cannot be ascertained retroactively, Equation 5 cannot be used here. The random sample is obviously not normally distributed – one of the values is much too far removed from the others. The prediction interval calculated by means of Equation 5 is based on a normally distributed total population. We will return to this example later. In such cases, outlier tests can be used (see additional online part 3).

Is every data set considered WS if all single values are WS? Our intuition expects this. In most cases it is true, but there are also exceptions (data sets D1 and D2 in Figure 2). It is often surprising how close to the specification limit the prediction interval moves. Even in data set D3, in Figure 2, whose highest single value lies far below the specification limit, the upper limit of the prediction interval lies very close to the specification limit. Only in cases such as D4 can (at n = 5!) it be assumed without further calculation that the data set conforms to the specification. Here, however, the distance between the specification limit and the highest single value already amounts to approximately 2 standard deviations. It becomes quickly obvious that the prediction interval can overlap the specification limit in individual cases despite all single values being WS. This can happen if the distribution within the data set is great and the single values lie relatively close to the specification limit.

Example 6:

The mean value in each case is 99%, $\hat{\sigma}$ is 1%, n = 2 or n = 3, the specification: 95 - 105%. The product of t * root is 7.73 (n = 2) or 3.37 (n = 3; compare Table 1); 99% - $\hat{\sigma}$ * t * root: 91.27% or 95.62%, resp. It becomes apparent that in similar cases of dual determinations the standard deviation and thus the prediction interval is very difficult to estimate. Even a standard deviation of 1%, as given in this example, is still too unsatisfactory for n = 2. The t-factor takes into consideration the great uncertainty in determining the distribution (standard deviation) based on only two measurements. If additional information about the distribution of the measurement values is available (e.g. R-control cards) confirming constancy over lengthy measurement series, the additional information may possibly be included in the evaluation.

Example 7:

The mean value is again 99%, $\hat{\sigma}$ is 1.5, 2 and 2.5%, resp. For what data numbers is the lower limit of the prediction interval $prd_u(x)$ greater than the specification limit of 95%?

The term $(\bar{x} - x_{soll})/\hat{\sigma}$ assumes for the above cited values the numbers (99 - 95)/1.5 = 2.67, and 2 and 1.6, resp. The distance, standardized to the standard deviation, thus gets smaller. At 2.67, n = 4 suffices to fall below the value; at 2, n = 9 is necessary. 1.6 can never be reached.

Example 8:

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

Based on these results the specification is clearly not met. If only the first 5 values are considered, the result is the same. Nevertheless, a case can be construed in which the specification is seemingly fulfilled: by forming mean values in steps:

$\bar{x}$ (1) = 300.35 (the first value)
$\bar{x}$ (1...2) = 300.59 (the mean value of the first two)
$\bar{x}$ (1...3) = 301.19...

It becomes apparent that by random accumulation of relatively low values starting at k = 9 die specification is seemingly fulfilled. This happens quite often. At some point even the mean value falls randomly below the target value if these values are not too far apart from each other and if enough measurements are taken. In the long term the mean value will always move in the direction of the mean value of the total population, although short-term fluctuations usually occur at the same time.
Here an interesting paradox on this subject [8]: In a coin-tossing experiment (heads or tails?) each of the surfaces will be on top at an average of 50%. If one looks at the "advantage" (the difference) the one event has over the other, <u>every</u> value is reached after a certain (possibly extremely high) number of tosses – 100 times more frequently heads than tails after 1,000,000 tosses – which can happen quite easily.

Taking measurements until the desired mean value is reached is considered non-permitted data manipulation. It gets even more serious if „outliers“ are eliminated. If value No. 38 (see Table 3) is eliminated from the series, the average drops significantly. The removal of this value is not permitted as both higher and especially low values (see Table 3) occur among normally distributed data. It is not justifiable to leave those values in the data set that are most useful for the desired interpretation and to eliminate the most extreme values that are unsuitable for the interpretation.

In order to avoid conscious or inadvertent manipulation, the random-sample size n should be specified in advance in relationship to the problem at hand [7]. Meeting a specification by means of arbitrary multiple measurements (testing into specification) is not permitted. Previously specified expanded (sequential) sampling plans are often useful for still detecting small differences without conducting unnecessarily extensive measurements if the results are clear. This approach is not only permitted, but even expressly recommended (USP, Ph. Eur., [6]). In such cases it is precisely specified when a repeat measurement is permitted and how many additional samples are to be taken. Even if the specification is met after, for example, 17 measurements, measurement must continue until the specified data number is reached. Previously devised rules prevent unintended manipulation.

Tables

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

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

Table 3: 100 single values of the determinations (in ppb) for Example 8

Table 1:

	α	β	n	k_0
γ_0=0,5 γ_1=0,9	20%	20%	3	3
	10%	20%	5	4
	10%	10%	7	6
	5%	20%	8	7
	5%	10%	10	8
	5%	5%	11	9
γ_0=0,6 γ_1=0,9	20%	20%	5	4
	10%	20%	9	8
	10%	10%	12	10
	5%	20%	13	11
	5%	10%	16	13
	5%	5%	19	15
γ_0=0,7 γ_1=0,9	20%	20%	11	9
	10%	20%	18	16
	10%	10%	24	20
	5%	20%	26	23
	5%	10%	33	28
	5%	5%	39	33
γ_0=0,8 γ_1=0,9	20%	20%	35	30
	10%	20%	59	52
	10%	10%	81	70
	5%	20%	83	73
	5%	10%	109	95
	5%	5%	133	114

Table 2:

n	1-γ =0.8	1-γ =0.9	1-γ =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

Table 3:

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