Chapter 6
Analytical Procedures

Keywords Analytical target profile (ATP) • Bayesian analysis • Method transfer • Prediction intervals • Procedure validation • Procedure qualification • Ruggedness factors • Tolerance intervals

6.1 Introduction

Analytical chemistry is used across the pharmaceutical industry to quantify and identify the components in drug substance, drug product, and raw material to ensure that the final dosage form remains safe and efficacious from lot release throughout the product's shelf life. To understand any potential shifts in the components impacting safety and efficacy, laboratories require analytical procedures which are reliable, fit for purpose, and executed consistently over time. Analytical procedures provide the instructions used by the analyst to ensure consistent use of laboratory equipment, solution preparation, measurement recording, and documentation. As such, analytical procedures form a critical component in any quality system. This chapter considers statistical methods that ensure that these procedures are fit for their intended purpose.

Martin et al. (2013) describe a holistic view of the analytical procedure life cycle. It frames this problem using concepts consistent with Quality by Design (QbD), ICH Q8 (2009), the FDA method validation guidance (2015), and the FDA process validation guidance (2011). The performance requirements of a procedure are defined by the analytical target profile (ATP). The ATP defines the analyte to be measured, the concentration range, procedure performance criteria, and product specifications. The criteria and specifications are established to define the purpose of the analytical procedure.

The analytical procedure life cycle is presented in the following three stages:

1. Stage 1: Procedure development and preparation for Stage 2.
2. Stage 2: Procedure performance validation (Qualification).
3. Stage 3: Procedure performance verification (Transfer and Monitoring).

The detail of each stage is discussed in this chapter. Other references of interest not discussed in this chapter are USP <1030>, <1033>, and <1223>.

R.K. Burdick et al., *Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry*, Statistics for Biology and Health, DOI 10.1007/978-3-319-50186-4_6

6.2 Terminology

6.2.1 Description of an Analytical Procedure

An analytical procedure and relevant terms must be clearly defined in order to design an appropriate analytical study. Descriptors such as "replicates" or "preparations" without further explanation often lead to confusion. Table 6.1 reports terminology used to describe an analytical procedure.

Not all analytical procedures entail all descriptions shown in Table 6.1. For example, liquid laboratory samples that require no further manipulations employ only a test solution. Table 6.2 provides an example of an analytical procedure for a solid dosing form.

Table 6.1 Analytical procedure terminology

Terminology	Description
Laboratory sample	The material received by the laboratory
Analytical sample	Material created by any physical manipulation of the laboratory sample such as crushing or grinding
Test portion	The quantity (aliquot) of material taken from the analytical sample for testing
Test solution	The solution resulting from chemical manipulation of the test portion such as chemical derivatization of the analyte in the test portion or dissolution of the test portion
Reading (individual determination)	The measured numerical value from a single unit of test solution
Reportable value	A summary value of individual readings, such as an average, from one or more units of a test solution. Replication may also occur across any level of the study design

Table 6.2 An analytical procedure for solid dosage coated pills

Terminology	Description			
Laboratory sample	100 coated pills			
Analytical sample	20 pills are removed from the laboratory sample and are crushed together in a mortar and pestle (i.e., composted)			
Test portion	Replicate 1: 1 gram crushed powder aliquot from analytical sample		Replicate 2: 1 gram crushed powder aliquot from analytical sample	
Test solution	Replicate 1: Test portion is dissolved in 1 L solvent		Replicate 2: Test portion is dissolved in 1 L solvent	
Reading (individual determination)	Reading 1 of replicate 1: test solution	Reading 2 of replicate 1: test solution	Reading 1 of replicate 2: test solution	Reading 2 of replicate 2: test solution
Reportable value	Average value of four readings			

6.2.2 Measurement Error Models

In this chapter, we consider the reportable value to be the key output from an analytical procedure and the focus of any investigation. In many cases, a particular analytical procedure may be used for different applications, with a different definition for the reportable value in each application. However, for purposes of discussion in this chapter, the term "reportable value" is used with the understanding that it may not be unique to a particular analytical procedure. A model that is useful for representing a reportable value is

$$\text{Reportable Value} = \text{True Value} + \text{Systematic Bias} + \text{Random Error} \quad (6.1)$$

where the true value and the systematic bias are fixed constants and the random error assigns a different error value to each reportable value. This relationship is represented symbolically as

$$Y = \tau + \beta + E \quad (6.2)$$

where Y represents the reportable value, τ (tau) is the true value, β (beta) is the systematic bias, and E is a random error with mean 0 and variance σ^2. (Note that β is not to be confused with its use as a regression slope in Sect. 2.12.) Model (6.2) uses the convention described in Sect. 2.12.7 of representing constants with Greek letters and random effects with upper case Latin letters. In many applications, σ^2 may be further decomposed into components that represent the various causes of variability.

6.2.3 Accuracy

In this text, accuracy concerns the magnitude of the systematic bias, β. The bias is defined as the long-run average of the difference, $Y - \tau$. Note that bias can only be determined if the true value, τ, is known. USP <1225> notes that a reference standard or a well-characterized procedure can be used to assign the value of τ. For relative content procedures used for large molecules, accuracy cannot be defined in this manner. Relative content procedures, sometimes referred to as relative purity procedures, include such procedures as size exclusion and cation exchange chromatography. Generally, minor species observed in purity procedures are product related variants or degradants, and orthogonal procedures are typically not available to provide a value for τ. Thus, the accuracy of the measurement as defined in this context cannot be independently confirmed. In cases where τ is not available, ICH Q2 (2005) states accuracy may be inferred once precision, linearity, and specificity have been established.

6.2.4 Precision

Precision of an analytical procedure is the degree of agreement among reportable values when the procedure is applied repeatedly to multiple samplings (possibly under different conditions) of a homogeneous test solution. The precision of an analytical procedure is quantified by the magnitude of the variance σ^2, or alternatively in terms of the standard deviation, σ. The standard deviation is the preferable measure of precision because it has the same measurement units as Y. The lesser the value of σ, the better the precision. Precision of a test procedure may be influenced by factors that vary during the normal use of the analytical procedure. These are called ruggedness factors, and include factors such as analyst, day, and instrument.

6.3 Stage 1: Procedure Development (Pre-validation)

In order to maximize the likelihood of a successful validation, it is imperative that all aspects of the procedure be well understood prior to the validation. Pre-validation work allows one to best design the experiment employed in the procedure validation. Martin et al. note that pre-validation experiments can be leveraged to support the validation and may reduce work in the validation itself. A lack of pre-validation work will often lead to a failed validation and costly rework.

The following series of questions provided by the USP Statistics Expert Team (2016) should be considered during pre-validation in order to ensure a successful validation experiment.

1. What are the allowable ranges for operational parameters such as temperature and time that impact the performance of the analytical procedure?
   - Robustness of these ranges can be determined using statistical design of experiments (DoE) as described in Chap. 3.
2. Are there ruggedness factors that impact precision?
   - Factors such as analyst, day, and instrument that vary in routine use and impact the precision of a test procedure are called ruggedness factors. When ruggedness factors impact precision, reportable values within the same ruggedness grouping (e.g., analyst) are correlated. Depending on the strength of the correlation, this may necessitate a statistical analysis that appropriately accounts for this dependence. Ruggedness factors can be identified empirically during pre-validation or based on a risk assessment. This topic is addressed in more detail in Sect. 6.4.10.
3. Are statistical assumptions for data analysis reasonably satisfied?

- These assumptions typically include normality, homogeneity of variance, and independence of reportable values. It is useful during pre-validation to employ statistical tests or visual representations to help answer these questions. USP <1010> provides information on this topic as does Sect. 2.12.2.

4. What is the required analytical range for the procedure?
5. Do accepted reference values or results from an established procedure exist for validation of accuracy?
   - If not, ICH Q2 states accuracy may be inferred once precision, linearity, and specificity have been established.
6. How many individual readings will be averaged to form a reportable value?
   - To answer this question, it is necessary to understand the contributors to the procedure variance and the procedure's ultimate purpose. Estimation of variance components during pre-validation provides useful information for making this decision. A good rule of thumb is to replicate against the source representing the largest component of variance.
7. What are appropriate validation acceptance criteria?
   - We provide discussion on this topic throughout Sect. 6.4.
8. How large a validation experiment is necessary?
   - Validation experiments should be properly powered to ensure there are sufficient data to conclude accuracy and precision can meet pre-specified acceptance criteria. Computer simulation is a useful tool for performing power calculations as discussed in Sect. 6.4.8.

Based on the answers to these and similar questions, a suitable validation experimental protocol may be designed.

6.4 Stage 2: Procedure Performance Validation (Qualification)

As noted in ICH Q2, the objective of validation of an analytical procedure is to demonstrate that it is suitable for its intended purpose. Suitability for intended purpose can be expressed in several ways. For instance, when a reportable value is used to disposition a product batch, suitability may be expressed in terms of decision error rates (e.g., passing an unacceptable batch, or failing an acceptable batch). In other cases, it may be sufficient to define suitability by placing limits on the quality metrics of the analytical procedure itself (e.g., maximum bias or precision). The life cycle approach suggests that these suitability metrics be documented in an analytical target profile (ATP) statement that guides quality decision making at all stages of the analytical procedure life cycle.

As discussed in the introduction, the term validation aligns with the process described in the USP document <1225>. In the life cycle approach described by Martin et al., this validation process is referred to as qualification.

The validation experiment is the culmination of all the investigational work needed to determine the operational details of the procedure. These details include selected inputs, operating conditions, equipment, limits, ranges, replication strategy, and other factors thought to potentially influence the outcome. The validation experiment is the final check that a newly developed procedure is fit for use.

Traditionally, validation of accuracy and precision provide the essential evidence that a procedure meets the requirements for the intended analytical application. Accordingly, we focus on these two topics in this section. Other factors that are typically characterized in a validation experiment are more descriptive in nature (e.g., range, detection, and quantitation limits), or more internal to the analytical procedure (e.g., linearity). For example, the impact of linearity is captured during DoE, repeatability, and intermediate precision studies because each experiment requires a calibration that includes the impact of the linearity. This is important because the decision rule and ATP provide an overarching criterion for the validation study, and require identification and quantification of all potential uncertainty components. For a full understanding of other characteristics, the reader should consult USP <1225>.

6.4.1 *Experimental Design for Validation of Accuracy and Precision*

A single experimental design will allow validation of both accuracy and precision. As will be discussed, individual assessment of accuracy and precision is not generally an effective approach. Such an approach is first described, and then better approaches that address bias and precision together are presented.

An example is provided to demonstrate the statistical analysis that follows. This example considers validation of a test procedure using high performance liquid chromatography (HPLC). The measured drug substance (DS) is a USP compendial substance, so information concerning τ is available. Three different quantities of reference standard were weighed to correspond to three different percentages of the test concentrations: 50, 100, and 150%. The unit of measurement for the reportable value is the mass fraction of DS expressed in units of mg/g and $\tau = 1000$ mg/g for all three concentrations. The DS product specification is from 980 to 1020 mg/g (see Weitzel 2012). Similar experiments are often established with levels expressed as a percent of the API label claim for the drug (as opposed to the weight of the entire tablet). Table 6.3 presents the $n = 12$ reportable values and the computed statistics.

Table 6.3 Example data set for procedure validation

Test concentration (%)	Test solution (plate or run)	Reportable value (mg/g)
50	1	1000.57
50	2	996.93
50	3	1002.4
50	4	994.91
100	5	994.16
100	6	992.72
100	7	1000.03
100	8	1004.89
150	9	1002.53
150	10	1004.83
150	11	998.17
150	12	994.15

To begin, the following two assumptions are made:

1. Each row in Table 6.3 is independent. In Sect. 6.4.10 and Sect. 6.4.11, the addition of ruggedness factors that invalidate this assumption is discussed. For example, consider the ruggedness factor "day." Suppose the experiment had been run over four days. Each day a reportable value was obtained from each of the three concentration levels: 50, 100, and 150. If there is variation in the procedure across days, then reportable values made on the same day are correlated and the assumption of independence is violated.
2. The standard deviation of the reportable value is constant across all three concentration levels. Discussion of how to proceed if this assumption is not met is provided in Sect. 6.4.9.

6.4.2 *Confidence Intervals for Accuracy and Precision*

The model in Eq. (6.2) is used to represent the data in Table 6.3 as

$$\begin{aligned} Y_{ij} &= \tau_i + \beta_i + E_{ij} \\ i &= 1, \ldots, c \text{ (concentration level)}; \; j = 1, \ldots r; \end{aligned} \tag{6.3}$$

where Y_{ij} is the jth reportable value in the ith concentration level, τ_i is the known true value of the ith concentration level, β_i is the procedure bias in the ith concentration level, and E_{ij} is a random error specific to jth reportable value in the ith concentration level. The random error is assumed to have a normal distribution with mean 0 and variance σ^2. For the data in Table 6.3, $c = 3$, $r = 4$, and so the total sample size is $n = c \times r = 12$. We present results for computing confidence intervals on β_i and σ^2 that can be used for validation of accuracy and precision.

6.4.2.1 Case 1: Bias Is Constant Across Concentration Levels

In this case, it is assumed that $\beta_i = \beta$ across all c concentration levels. Note this does not require that τ_i be equal across concentration levels. Since there are an equal number of reportable values for each concentration level, the estimator for β is

$$\hat{\beta} = \frac{\sum_{i=1}^{c} (\bar{Y}_i - \tau_i)}{c}. \tag{6.4}$$

The bounds for a $100(1-\alpha)\%$ two-sided confidence interval for β are

$$\begin{aligned} L &= \hat{\beta} - t_{1-\alpha/2:n-1}\sqrt{\frac{S^2}{n}} \\ U &= \hat{\beta} + t_{1-\alpha/2:n-1}\sqrt{\frac{S^2}{n}} \\ \bar{Y}_i &= \frac{\sum_{j=1}^{r} Y_{ij}}{r} \\ \bar{Y} &= \frac{\sum_{i=1}^{c}\sum_{j=1}^{r} Y_{ij}}{n} \\ S^2 &= \frac{\sum_{i=1}^{c}\sum_{j=1}^{r} \left(Y_{ij} - \bar{Y}\right)^2}{n-1}. \end{aligned} \tag{6.5}$$

Validation for precision typically requires only an upper bound, since it is only problematic if the standard deviation is too large. A $100(1-\alpha)\%$ upper bound on σ is

$$U = \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha:n-1}}}. \tag{6.6}$$

Note that (6.6) can be calculated with no knowledge of τ_i. Thus, although the true value is required to estimate accuracy, it is not needed to estimate precision.

For the data shown in Table 6.3, $\tau_i = \tau = 1000$ mg/g for each concentration level. The calculated statistics are $\bar{Y}_1 = 998.70$, $\bar{Y}_2 = 997.95$, $\bar{Y}_3 = 999.92$, $\bar{Y} = 998.86$, and $S^2 = 18.55$. Equation (6.5) is now simplified since all τ_i are equal and provides the 90% confidence interval on β

$$
\begin{aligned}
L &= \bar{Y} - \tau - t_{1-\alpha/2:n-1}\sqrt{\frac{S^2}{n}} \\
L &= 998.86 - 1000 - 1.796\sqrt{\frac{18.55}{12}} = -3.38 \text{ mg/g} \\
U &= \bar{Y} - \tau + t_{1-\alpha/2:n-1}\sqrt{\frac{S^2}{n}} \\
U &= 998.86 - 1000 + 1.796\sqrt{\frac{18.55}{12}} = 1.09 \text{ mg/g}
\end{aligned}
\tag{6.7}
$$

Equation (6.6) provides the upper 95% confidence bound on σ

$$
\begin{aligned}
U &= \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha:n-1}}} \\
U &= \sqrt{\frac{(12-1)18.55}{4.57}} = 6.68 \text{ mg/g}
\end{aligned}
\tag{6.8}
$$

6.4.2.2 Case 2: Bias Changes Across Concentration Levels

In this case, it is necessary to estimate the bias separately for each concentration level. However, since the standard deviation is assumed equal across all concentration levels, it is still possible to use all the data to estimate σ^2. This is referred to as "pooling." In order to pool the variance estimates, an analysis of variance table is constructed as shown in Table 6.4 (see Sect. 2.12.7 for more on the analysis of variance).

where

$$
\begin{aligned}
S_C^2 &= \frac{r\sum_{i=1}^{c}(\bar{Y}_i - \bar{Y})^2}{c-1} \\
S_E^2 &= \frac{\sum_{i=1}^{c}\sum_{j=1}^{r}(Y_{ij} - \bar{Y}_i)^2}{c(r-1)}.
\end{aligned}
\tag{6.9}
$$

Table 6.4 Analysis of variance

Source of variation	Degrees of freedom	Mean square
Concentration	$c-1$	S_C^2
Error	$c(r-1)$	S_E^2

Table 6.5 Analysis of variance for example data

Source of variation	Degrees of freedom	Mean square
Concentration	2	$S_C^2 = 3.953$
Error	9	$S_E^2 = 21.796$

Since the bias is different across concentration levels, a separate confidence interval is needed for each concentration level. The bounds for a $100(1-\alpha)\%$ two-sided confidence interval for β_i are

$$\begin{aligned} L_i &= \bar{Y}_i - \tau_i - t_{1-\alpha/2:c(r-1)}\sqrt{\frac{S_E^2}{r}} \\ U_i &= \bar{Y}_i - \tau_i + t_{1-\alpha/2:c(r-1)}\sqrt{\frac{S_E^2}{r}} \end{aligned} \tag{6.10}$$

where $\bar{Y}_i$ is defined in (6.5) and S_E^2 is defined in (6.9).

A $100(1-\alpha)\%$ upper bound on σ is

$$U = \sqrt{\frac{c(r-1)S_E^2}{\chi^2_{\alpha:c(r-1)}}}. \tag{6.11}$$

The analysis of variance table for the data in Table 6.3 is shown in Table 6.5.

Using Eq. (6.10) a 90% two-sided confidence interval for the bias in the 100% concentration level is

$$\begin{aligned} L &= \bar{Y}_i - \tau_i - t_{1-\alpha/2:c(r-1)}\sqrt{\frac{S_E^2}{r}} \\ L &= 997.95 - 1000 - 1.833\sqrt{\frac{21.796}{4}} = -6.33 \text{ mg/g} \\ U &= \bar{Y}_i - \tau_i + t_{1-\alpha/2:c(r-1)}\sqrt{\frac{S_E^2}{r}} \\ U &= 997.95 - 1000 + 1.833\sqrt{\frac{21.796}{4}} = 2.23 \text{ mg/g}. \end{aligned} \tag{6.12}$$

The confidence intervals for the 50 and 150% concentration levels are made in a similar fashion.

A 95% upper bound on the (pooled) precision from (6.11) is

$$\begin{aligned} U &= \sqrt{\frac{c(r-1)S_E^2}{\chi^2_{\alpha:c(r-1)}}} \\ U &= \sqrt{\frac{9(21.796)}{3.33}} = 7.68 \text{ mg/g}. \end{aligned} \tag{6.13}$$

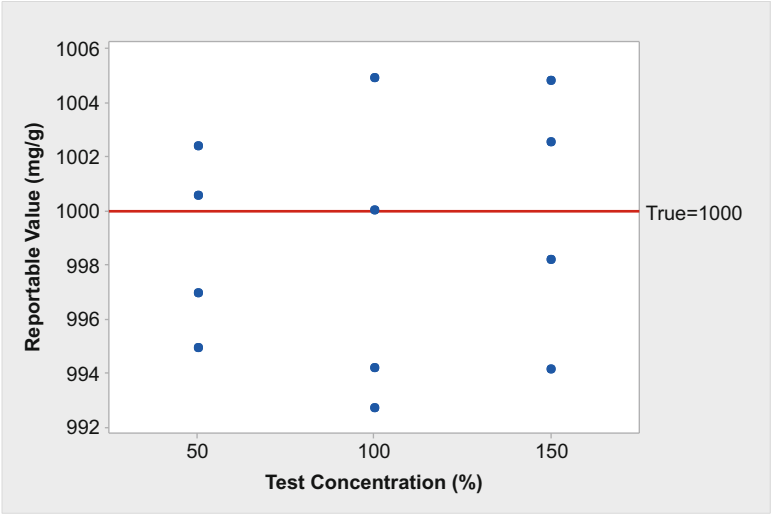

Fig. 6.1 Plot of data in Table 6.3

Note this upper bound is slightly greater than that computed in Case 1 because there are fewer degrees of freedom associated with the pooled estimate of σ (11 in Case 1 and 9 in Case 2). A plot of the data in Table 6.3 is shown in Fig. 6.1.

Note that the reportable values are centered at roughly the same value across levels of concentration. This suggests the bias is constant and that Case 1 is the more appropriate procedure. Also note that the spread appears constant across concentration. This is consistent with the second assumption noted in Sect. 6.4.1.

6.4.3 Using Confidence Intervals to Validate Accuracy and Precision

The confidence intervals provided in the previous section can be used to validate accuracy and precision individually. Because bias can be either positive or negative, it is customary to perform a statistical test of equivalence to validate accuracy. Tests of equivalence are discussed in Sect. 2.11. Assume Case 1 is appropriate and Eq. (6.5) is used to compute a 90% two-sided confidence interval on the bias. A pre-selected value of the equivalence acceptance criterion (EAC) to validate accuracy was selected to be 5 mg/g or 0.5% of the true value. (Section 6.4.4 discusses considerations in selecting an appropriate EAC.) Since the 90% confidence interval from $L = -3.38$ mg/g to $U = 1.09$ mg/g falls completely in the range from –5 to +5 mg/g, the statistical equivalence test is passed, and it can be claimed the procedure is validated for accuracy.

To validate precision, the upper bound on the standard deviation must be less than a pre-selected acceptance criterion. In our example, assume precision is validated if it can be shown that the standard deviation is less than 7 mg/g. Since the computed 95% upper bound in this example, U=6.68 mg/g, is less than the acceptance criterion of 7 mg/g, the procedure has been validated with respect to precision.

6.4.4 Validation Criteria for Accuracy and Precision

As discussed in Sect. 2.11, acceptance criterion should ideally be defined by the analytical scientist and not the statistician. The criterion must be meaningful in the sense that it must define what is meant by "fit for purpose." When validating accuracy and precision individually, this is the most difficult part of the analysis. This is because a procedure that has very small bias can accept a greater standard deviation than a procedure with a large bias. Similarly, a procedure with a relatively small standard deviation can accept a relatively large bias. For this reason, the criteria for these two attributes are linked, making it difficult to get a good assessment of individual criteria. Many companies, as well as industry standards organizations, have default limits that are used for all validations. These may be based on industry benchmarking, but it is arguable whether such an approach truly demonstrates "fit for purpose." For this reason, and to account for the relationship between accuracy and precision as it relates to overall performance, we recommend two other approaches for validation of accuracy and precision. We present these approaches in the next two sections.

6.4.5 Validation of Accuracy and Precision Using Statistical Intervals

Hubert et al. (2004, 2007a, b) proposed validation of both accuracy and precision simultaneously rather than individually as described in the previous section. The reasoning is to take advantage of the natural tradeoffs between these two characteristics. For example, a procedure with a relatively small standard deviation can accept a greater bias than a procedure with a larger standard deviation. Because the intended purpose of an analytical procedure is to provide accurate and precise measurements, one may consider that the procedure is validated if it is shown to provide a high degree of assurance that future measured values will be close to their true values. A criterion that can be used to simultaneously validate accuracy and precision seeks to ensure

$$\begin{aligned} &\Pr(-\lambda < Y - \tau < \lambda) \geq P, \text{ or} \\ &\Pr(-\lambda + \tau < Y < \lambda + \tau) \geq P \end{aligned} \tag{6.14}$$

where $\lambda > 0$ is an acceptable limit defined a priori to be consistent with the purpose of the procedure. The term P is a desired probability value (e.g., $P = 0.90$).

For example, a desired goal for a procedure that reports concentration in mg/mL may be written in the following manner: "The procedure must ensure that at least 90% of the time, measurement error (i.e., the difference between the reported value and the true value) is no greater than 0.5 mg/mL." In terms of Eq. (6.14), this means $\lambda = 0.5$ mg/mL and $P = 0.90$.

Equation (6.14) can be interpreted as either (1) the probability that *the next* reportable value falls in the range from $-\lambda + \tau$ to $\lambda + \tau$ is greater than or equal to P, or (2) the proportion of *all* future reportable values falling between $-\lambda + \tau$ and $\lambda + \tau$ is greater than or equal to P. Accordingly, two statistical intervals have been proposed for validating Eq. (6.14).

1. A prediction interval of reportable values (also referred to as an expectation tolerance interval) and
2. A tolerance interval of reportable values (also referred to as a content tolerance interval).

The prediction interval validates that (6.14) is true for the *next* reportable value, whereas the tolerance interval validates that (6.14) is true for *all future* reportable values with a specific level of confidence. Since the inference associated with the tolerance interval concerns a larger set of values, the tolerance interval is always wider than the prediction interval.

Both intervals can be used in the following manner to validate accuracy and precision simultaneously:

1. Compute the appropriate statistical interval using Eq. (2.21) for the prediction interval and Eq. (2.23) for the tolerance interval.
2. Compute a $100P$ % prediction interval or a 90% tolerance interval that contains $100P$ % of the population. A 90% confidence level for the tolerance interval will provide a statistical test with a type I error rate (probability of rejecting the null hypothesis when it is true) of approximately 5%.
3. If the computed interval falls completely in the range from $-\lambda + \tau$ to $\lambda + \tau$, criterion (6.14) is satisfied and the procedure is validated for both accuracy and precision.

When computed by classical statistical methods (as we do below in this section) the interpretation of these intervals is as follows. When the interval estimation methodology is applied repeatedly to many (i.e., an infinite number) of hypothetical future data sets of size n from possibly many different populations, each prediction interval obtained has a $100P$ % probability of containing the hypothetical next reportable value. Similarly, each $100(1-\alpha)$% tolerance interval has a $100(1-\alpha)$ % probability of containing at least $100P$ % of hypothetical future reportable values.

Our inference about the truth (or not) of relationship (6.14) is thus based on the properties of the statistical procedure. Whether (6.14) is true for any particular sampled population is unknown because it depends on parameters whose values are unknown. In Sect. 6.4.7, Bayesian interval estimation is introduced as a more direct alternative to validation using (6.14).

Huber et al. recommend the testing strategy based on the prediction interval. Yang and Zhang (2015) recommend the tolerance interval. The tolerance interval is the appropriate choice if one desires a statistical test in which the type I error rate is controlled. The tolerance interval is therefore more consistent with the approach described in Sects. 6.4.2 and 6.4.3.

To demonstrate, consider Case 1 (bias constant) and analyze the data in Table 6.3. Suppose (6.14) is defined so that $\lambda = 0.015 \times \tau = 15$ mg/g and $P = 0.90$. Thus, we seek to validate the claim

$$\begin{aligned} &\Pr(-\lambda + \tau < Y < \lambda + \tau) \geq P \\ &\Pr(-15 + 1000 < Y < 15 + 1000) \geq 0.90 \\ &\Pr(985 < Y < 1015) \geq 0.90. \end{aligned} \tag{6.15}$$

From Eq. (2.21) the 90% prediction interval is computed as

$$\begin{aligned} L &= \bar{Y} - t_{(1+P)/2:n-1}\sqrt{\left(1 + \frac{1}{n}\right) \times S^2} \\ L &= 998.86 - 1.796\sqrt{\left(1 + \frac{1}{12}\right) \times 18.55} = 990.8 \text{ mg/g} \\ U &= \bar{Y} + t_{(1+P)/2:n-1}\sqrt{\left(1 + \frac{1}{n}\right) \times S^2} \\ U &= 998.86 + 1.796\sqrt{\left(1 + \frac{1}{12}\right) \times 18.55} = 1006.9 \text{ mg/g}. \end{aligned} \tag{6.16}$$

From Eq. (2.23) using an exact K value of 2.414, the 90% tolerance interval that includes 90% of the future population of reportable values is

$$\begin{aligned} L &= \bar{Y} - K\sqrt{S^2} \\ L &= 998.86 - 2.414\sqrt{18.55} = 988.5 \text{ mg/g} \\ U &= \bar{Y} + K\sqrt{S^2} \\ U &= 998.86 + 2.414\sqrt{18.55} = 1009.3 \text{ mg/g}. \end{aligned} \tag{6.17}$$

Since both intervals (6.16) and (6.17) fall within the range from 985 to 1015 defined in (6.15), both intervals validate the procedure. As described, the tolerance

interval is wider than the prediction interval, since it makes an inference to a larger set of values. It is also true that the tolerance interval provides a statistical test with a type I error rate near 5%.

One final comment concerning application of this approach. When validated individually, each test has a type I error rate of 5% and the combined error rate can be as high as 10%. Thus, it is not unreasonable to apply a 10% type I error rate with the simultaneous methods described in this section. This means one could use an 80% confidence level for the two-sided tolerance interval. In the present application, the 80% tolerance interval that contains 90% of all future reportable values is from 989.6 to 1008.1 mg/g. This compares to the previously computed 90% tolerance interval of 988.5 mg/g to 1009.3 mg/ml.

6.4.6 *Validation of Accuracy and Precision Based on Out-of-Specification Rates*

A typical application for an analytical procedure is lot (batch) release. After a lot is manufactured, a reportable value of the product quality is obtained using the analytical procedure. If the reportable value falls within the lower specification limit (LSL) and upper specification limit (USL), it is deemed as satisfying the quality requirement. However, if it falls outside of this range, action must be taken to determine the lot disposition. Thus, an obvious criterion for procedure validation is the probability that a reported value is out-of-specification (OOS). If the process is operating as designed, then a reported OOS alarm in most cases is "false," and can lead to unnecessary time and expense in further examination of the lot. The probability statement in (6.14) can be adapted to consideration of the OOS rate by defining $-\lambda + \tau = LSL$ and $\lambda + \tau = USL$ where LSL and USL are the process lower and upper specifications, respectively, and it is assumed the process is symmetric about τ (i.e., $(LSL + USL)/2 = \tau$). Thus, (6.14) is rewritten as

$$\begin{aligned} &\Pr(-\lambda + \tau < Y < \lambda + \tau) \geq P \\ &\Pr(LSL < Y < USL) \geq P \\ &\pi \leq 1 - P \end{aligned} \tag{6.18}$$

where $\pi = 1 - \Pr(LSL < Y < USL)$ is the probability of an OOS signal. A 95% upper bound can be constructed on π using the process described in Sect. 2.6.5. If the upper bound is less than $1-P$, then (6.18) is satisfied and the analytical procedure is validated.

To demonstrate for the present Case 1 example, $LSL = 980$ mg/g, $USL = 1020$ mg/g, $\bar{Y} = 998.86$, $S^2 = 18.55$, and $P = 0.90$. Following the instructions from Eqs. (2.17) to (2.19) with $\alpha = 0.10$ in (2.19), we compute $K_{LSL} = 4.38$ and $K_{USL} = 4.91$. Since both of these values exceed $(n-1)/\sqrt{n} = (12-1)/\sqrt{12} = 3.175$, $K^* = \min(4.38, 4.91) = 4.38$, $\lambda_U = 9.51$, and $U = 0.003$. Since this upper

bound is less than $1-P = 1\text{-}0.90 = 0.10$, the procedure is validated against the OOS criterion.

There is an interesting relationship between the upper bound on π and the tolerance interval used to validate (6.14) when $-\lambda + \tau = LSL$ and $\lambda + \tau = USL$. To demonstrate, consider a situation where there is only an upper specification, USL. Let U_1 represent the 95% upper tolerance bound that exceeds $100P\%$ of the population computed with (2.28) and the exact value of K_1. Let U_2 represent the upper 95% confidence bound computed using K_{USL} in Eq. (2.14). When $U_1 = USL$, it must be true that $U_2 = 1 - P$. Thus, the two rules for validation are exactly the same, and have a type I error rate of 0.05. Although the situation with two-sided specifications involves approximations, the two approaches will also generally provide the same result, and the type I error rate is very close to 0.05.

One final adjustment is required for application of this approach. To this point, only the measurement error has been quantified. However, if specifications are used to define "fitness for purpose," it is necessary to also account for process variation. To do this, let σ_P^2 represent the variance of the manufacturing process and σ^2 the variance of the reportable value. (The statistic S^2 is an estimator for σ^2). Now define

$$\rho = \frac{\sigma_P^2}{\sigma_P^2 + \sigma^2}. \tag{6.19}$$

which represents the proportion of the total variance in the reportable value due to the process. An empirical estimate of σ_P^2 might be available from process data, but if not, a subject manner expert can generally provide a reasoned guess for ρ. For example, if the procedure is a bioassay, it is expected that variance due to the analytical procedure is greater than the process variance, and so a value of $\rho = 0.2$ might be reasonable. In contrast, a procedure with relatively little measurement error might employ $\rho = 0.8$.

With a known or well-informed value for ρ, the total of process and measurement variance is written as

$$\sigma_P^2 + \sigma^2 = \sigma^2\left[\frac{\sigma_P^2}{\sigma^2} + 1\right] = \sigma^2\left[\frac{\rho}{1-\rho} + 1\right] = \sigma^2\left[\frac{1}{1-\rho}\right]. \tag{6.20}$$

The confidence interval for π is now computed as before, but with S^2 replaced with $S^{2*} = S^2\left[\frac{1}{1-\rho}\right]$.

In the present example, assume that $\rho = 0.5$. Performing the previous calculations with $S^{2*} = S^2\left[\frac{1}{1-\rho}\right] = 18.55\left[\frac{1}{1-0.5}\right] = 37.1$, we obtain $K_{LSL} = 3.096$ and $K_{USL} = 3.471$, $K^* = \min(3.096, 3.471) = 3.096$, $\lambda_U = 6.545$, and $U = 0.029$. Since this upper bound is less than $1-P$=1-.90=0.10, the procedure is validated against the OOS criterion.

Burdick et al. (2005) provide an approach for estimating false failure rates and missed fault rates. As noted earlier, the false failure rate is generally very close to

the observed OOS rate. However, it is also of interest to know if the missed fault rate is acceptably low. By defining criteria based on misclassification rates, a procedure can be validated using upper confidence bounds on these rates.

Other approaches for establishment of criteria are provided by Chatfield and Borman (2009).

6.4.7 A Bayesian Approach

It is also possible to estimate $\Pr(-\lambda < Y - \tau < \lambda)$ directly using a Bayesian approach (see Sect. 2.13 for a discussion of Bayesian statistics). The validation criterion is thus satisfied if this estimated probability exceeds P. A Bayesian tolerance interval is provided in Wolfinger (1998) and can be computed using the statistical software package WinBUGS (Ntzoufras 2009 or Spiegelhalter et al. 1996).

The WinBUGS code required for Case 2 is shown below:

```
# data
Level[]    Y[]
1    1000.57
1    996.93
1    1002.4
1    994.91
2    994.16
2    992.72
2    1000.03
2    1004.89
3    1002.53
3    1004.83
3    998.17
3    994.15
END
# more data
list(n=12, c=3, tau=1000)
model{
  # Priors
  for(i in 1:c){ beta[i] ~ dnorm(0,0.000001) }
  sigma ~ dunif(0, 100)
  precision <- pow(sigma,-2)
  # Likelihood
```

(continued)

```
  for(obs in 1:n){
    Dif[obs] <- Y[obs] - tau
    Dif[obs] ~ dnorm (beta[ Level[obs] ],precision)
  }
}
```

Boxplots that compare the posterior distributions for beta[i] are shown in Fig. 6.2. The distributions are labeled with the concentration level index (i.e., [1] = 50, [2] = 100, [3] = 150). These boxplots are created by WinBUGS and are different than those described in Sect. 2.4. Boxes represent inter-quartile ranges and the solid black line at the (approximate) center of each box is the mean. The arms of each box extend to cover the central 95% of the distribution. The horizontal line behind the boxes is the overall mean of the posterior means.

The two-sided 90% credible interval for beta2 is −5.885 to +3.307 which may be compared to the classical frequentist two-sided 90% confidence interval computed in (6.12) of −6.33 to +2.23. The upper 95% credible bound for sigma is 8.49 which may be compared to the classical frequentist upper 95% confidence bound of 7.68 shown in (6.13). Figure 6.2 supports the conclusions from Fig. 6.1 that Case 1 (constant bias) provides a more appropriate model for these data.

The WinBUGS code required for Case 1 is shown below:

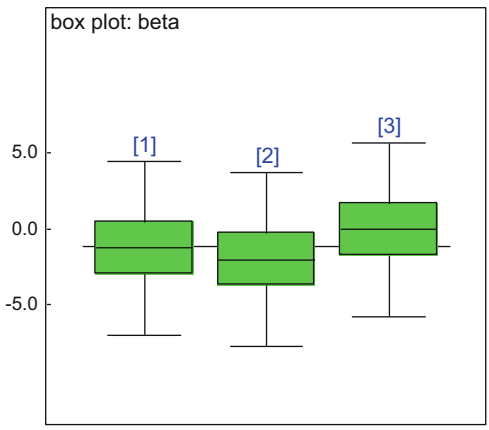

Fig. 6.2 Comparison of posterior distributions

```
# data
Y[]
1000.57
...
END
# more data
list(n=12, tau=1000)
# Let WinBUGS pick initials
model{
  # Priors
  beta ~ dnorm(0,0.000001)
  sigma ~ dunif(0, 100)
  precision <- pow(sigma,-2)
  # Likelihood
  for(obs in 1:n){
    Dif[obs] <- Y[obs] - tau
    Dif[obs] ~ dnorm (beta,precision)
  }
}
```

The posterior sample obtained from executing this code was imported into R for further analysis. This sample consists of 300,000 pairs of values for β and σ drawn from their joint posterior. A two-sided 90% central credible interval for β can be obtained using the R function

```
quantile(beta,c(0.05,0.95))
```

The interval is -3.50 to $+1.22$ mg/g which is comparable to the corresponding classical frequentist interval of -3.38 to $+1.09$ that was computed in (6.7).

A one-sided upper 95 credible bound for sigma can be obtained using the following R function:

```
quantile(sigma,c(0.95))
```

The resulting upper bound is 7.204 mg/g which is comparable to the classical frequentist upper bound of 6.68 mg/g computed in (6.8).

It is of interest to estimate the posterior probability that 90% of future values will be within $\tau \pm \lambda$ where $\tau = 1000$ and $\lambda = 15$. This can be obtained using the following two lines of R code:

```
Content <- pnorm(tau + lambda, tau + beta, sigma) - pnorm(tau
- lambda, tau + beta, sigma)
mean(Content >= 0.9)
```

The result is 0.989. Since this value is greater than $0.90 = 90\%$, the method is validated. The frequentist approach computes an upper bound on $\Pr(-\lambda < Y - \tau < \lambda)$ using results from Sect. 2.6.5. An example of this approach was given in (6.18) where $\lambda = 20$.

A sample of 100,000 draws from the posterior predictive distribution of future reportable results may be obtained using the following random normal R function:

```
Y.fut <- rnorm(100000,beta+tau,sigma)
```

The central two-sided 90% credible posterior predictive interval of future values may be obtained using the R function

```
quantile(Y.fut,c(0.05,0.95))
```

The resulting interval is from 990.3 to 1007.5 which may be compared to the classical prediction interval of 990.8–1006.9 computed in (6.16).

It is of interest to estimate the posterior predictive probability that future values will be within $\tau \pm \lambda$. This can be obtained using the following two lines of R code:

```
Y.fut.in <- (tau - lambda <= Y.fut) & (Y.fut <= tau + lambda)
mean(Y.fut.in)
```

The resulting probability is 0.989. There is no comparable classical frequentist estimate available.

It is of interest to estimate the posterior predictive probability that future values will be OOS (outside 980 to 1020). The following two lines of R code will provide this estimate:

```
OOS <- (Y.fut <= 980) & (1020 >= 1020)
mean(OOS)
```

The resulting probability is 0.00137 which can be compared to the result obtained previously using classical frequentist approaches of 0.003 shown in Sect. 6.4.6.

A Bayesian content tolerance interval to contain 90% of future values with 90% credibility can be obtained using Algorithm 11.2 of Krishnamoorthy and Mathew (2009). The R code is a bit more involved because it requires an iterative search. The implementation of this algorithm in R is given below.

```
P<-0.90; C<-0.90
L<- tau + beta + qnorm((1-P)/2,0,sigma)
U<- tau + beta + qnorm((1+P)/2,0,sigma)
ndraws<-length(L)
mu.bar <- mean(c(L,U))
intervals<-1000 # 10000 takes too long
mu.2.range <- seq(min(U),max(U), ( max(U)-min(U) )/intervals
)
PP <- matrix (rep(NA,3*length(mu.2.range)),ncol=3)
i <-0
for(mu.2 in mu.2.range){
  i<-i+1
  mu.1 <- -mu.2 + 2*mu.bar
      PP[i,]<-c(mu.1,mu.2,mean((U<=rep(mu.2,ndraws))&(L>=rep
(mu.1,ndraws))))
}
# select values in PP for which the proportion covered is
near C
# may need to adjust the factors of C to select a narrow range
of values close to C
PP.out<- PP[ (PP[,3]<=1.02*C) & (PP[,3]>=0.98*C),]
TI <- c(mean(PP.out[,1]),mean(PP.out[,2]))
TI
```

The resulting Bayesian content tolerance interval is 986.5 to 1011.3 which is comparable to the 988.5–1009.3 as computed in (6.17).

While the Bayesian results are quite close to those obtained using classical statistical methods, the interpretation of these intervals and probabilities are different from the frequentist interpretations. The Bayesian results were obtained using relatively uninformative prior distributions for β and σ. Had more informative distributions been available, the estimates could be much different, and arguably more informative. The other difference is that the Bayesian methodology replaces analytical solutions (some of which are necessarily approximate) with computer

algorithms for which there are no approximations. However, these computer algorithms have their own challenges related to Markov Chain Monte-Carlo (MCMC) convergence verification and the requirement for large MCMC samples to minimize simulation error. An additional advantage of the Bayesian approach is that it is easily extended to more complex models for which frequentist analytical approaches are intractable.

6.4.8 Power Considerations

As noted in Sect. 6.3, it is important to conduct pre-validation work to gain understanding of the procedure. Part of this work should include a power analysis to determine the probability of passing validation under given scenarios. Computer simulation is extremely useful for this purpose. Statistical power is defined as the probability that one meets the acceptance criterion given a true value of the parameter of interest. To demonstrate, a simulation program was written to determine the power of a validation test based on the requirement that the probability of an OOS (π) is less than $1-P=0.10$ with a type I error rate of 5%. The specifications are $LSL=980$ and $USL=1020$. The test is to be conducted as described in Sect. 6.4.6. The simulation was conducted with 100,000 iterations. Table 6.6 presents the results for validation designs with $n=6$, $n=12$, and $n=20$ using a 90% confidence coefficient (type I error rate of 0.05). Table 6.7 reports the same design with an 80% confidence coefficient (type I error rate of 0.10).

Note that the power in the last row where $\pi = 0.10$ is the estimated type I error rate. Since these values are all less than the desired rate (0.05 in Table 6.6 and 0.10 in Table 6.7), this provides an additional argument for applying an 80% confidence coefficient on the confidence interval for π. The simulation results also show that with the typical sample sizes used in a validation study, the criterion ($1 - P = 0.10$ in this case) must be much greater than the true value of π in order to provide a reasonable chance of passing the validation.

Table 6.6 Power for several designs with 90% confidence

True value of π	$n=6$	$n=12$	$n=20$
0.001	0.566	0.875	0.997
0.005	0.375	0.644	0.941
0.010	0.270	0.455	0.808
0.10	0.041	0.028	0.039

Table 6.7 Power for several designs with 80% confidence

True value of π	$n=6$	$n=12$	$n=20$
0.001	0.775	0.968	0.999
0.005	0.586	0.842	0.983
0.010	0.455	0.684	0.919
0.10	0.087	0.068	0.086

6.4.9 Violation of Homogeneity Across Concentration Levels

Procedures based on chemical or biological principles will sometimes demonstrate different variances as concentrations vary. This violates one of the assumptions made during discussions to this point. In such a situation, it is sometimes possible to transform the data so that the standard deviations can be assumed equal across the concentration range. The analyses that combine the data from all concentration levels as described above can then be performed using the transformed data, with appropriately transformed validation criteria.

It is extremely important that pre-validation work be used to determine necessary transformations that will allow the pooling of data across concentration levels. Failure to do so could lead to either unnecessary experimentation, or an underpowered validation experiment. Section 4.3 of USP chapter <1032> presents an excellent review of this topic. The normality transformations described in Sect. 2.6.10 also often stabilize the variance.

6.4.10 Experimental Designs to Incorporate Ruggedness Factors

In order to validate a procedure across the total environment in which it is expected to operate, it is sometimes necessary to manipulate ruggedness factors in the experimental design. Examples of ruggedness factors include analysts, equipment, and days. Table 6.8 reports the same data shown in Table 6.3, but with information concerning the analyst that performed the preparation work for the assay.

We again assume that the reportable value has a constant variance across all three concentration levels, and that bias is constant (Case 1) so that we may

Table 6.8 Example data set with ruggedness factors

Test concentration (%)	Analyst	Test solution (plate or run)	Reportable value (mg/g)
50	1	1	1000.57
50	1	2	996.93
50	2	3	1002.4
50	2	4	994.91
100	3	5	994.16
100	3	6	992.72
100	4	7	1000.03
100	4	8	1004.89
150	5	9	1002.53
150	5	10	1004.83
150	6	11	998.17
150	6	12	994.15

combine all concentration levels into a single data set. Unlike the previous analysis with Table 6.3, the 12 rows in Table 6.8 are not independent unless analysts do not impact the reportable value. For example, rows 1 and 2 were both prepped by analyst 1. If analyst impacts the reportable value, the values in rows 1 and 2 are more similar with each other than with the values in other rows of the table. In such a situation, we state the responses made by the same analyst are correlated. We described such a situation previously in Sects. 2.7 and 2.12.7.

In the present example, represent the reportable value with the statistical model

$$\begin{aligned} &Y_{ij} = \tau + A_i + E_{ij} \\ &i = 1, \ldots, a \text{ (analyst)}; \; j = 1, \ldots r; \end{aligned} \tag{6.21}$$

where Y_{ij} is the reportable value for the j^{th} replicate of the i^{th} analyst. The number of analysts in this example is $a = 6$, and each analysts performs $r = 2$ independent repetitions. The random error A_i represents between analyst variability. It is assumed to have a mean of zero and a variance σ_A^2. The random error E_{ij} is the within analyst variability which has an assumed mean of zero and variance σ_E^2. The parameter τ represents the true value (1000 mg/g for all concentration levels in our example).

The ANOVA table for model (6.21) is shown in Table 6.9. Formulas for S_A^2 and S_E^2 are provided in Table 2.16. The numerical values for the data in Table 6.8 are shown in Table 6.10.

The model shown in (6.21) assumes that analyst is a random effect. That is, a sample of six analysts used in the experiment is viewed as a random sample from a population of analysts that will perform the procedure in the future. In some situations, it may be reasonable to treat analyst as a fixed effect. This case is considered in the next section.

The total variance associated with the procedure is the sum of the variance components, $\sigma_{Total}^2 = \sigma_A^2 + \sigma_E^2 = (\theta_1 - \theta_2)/r + \theta_2 = \theta_1/r + (1 - 1/r)\theta_2$. This sum is called the intermediate precision. The estimator for the intermediate precision is

$$S_{Total}^2 = \frac{S_A^2}{r} + \frac{(r-1)S_E^2}{r}. \tag{6.22}$$

Table 6.9 ANOVA for model (6.21)

Source of variation	Degrees of freedom	Mean square	Expected mean square
Between analysts	$n_1 = a - 1$	S_A^2	$\theta_1 = \sigma_E^2 + r\sigma_A^2$
Within analysts	$n_2 = a(r-1)$	S_E^2	$\theta_2 = \sigma_E^2$

Table 6.10 ANOVA, for example

Source of variation	Degrees of freedom	Mean square
Between analysts	$n_1 = 5$	S_A^2 =29.165
Within analysts	$n_2 = 6$	S_E^2 =9.708

Using the Satterthwaite approximation given in (2.120), the degrees of freedom associated with this estimator is

$$m = \frac{S_{Total}^4}{\dfrac{S_A^4}{r^2 \times n_1} + \dfrac{(r-1)^2 \times S_E^4}{r^2 \times n_2}} \tag{6.23}$$

Using the data in Table 6.10,

$$S_{Total}^2 = \frac{29.165}{2} + \left(1 - \frac{1}{2}\right) 9.708 = 19.437 \tag{6.24}$$

and

$$m = \frac{(19.437)^2}{\dfrac{(29.165)^2}{2^2 \times 5} + \dfrac{(2-1)^2 \times (9.708)^2}{2^2 \times 6}} = 8.13 = 8 \text{ (rounded)} \tag{6.25}$$

Notice that $S_{Total}^2 = 19.437$ is greater than the estimate of the measurement variability obtained when assuming all 12 rows of the data table are independent $\left(S^2 = 18.55\right)$. This demonstrates the problem of not properly modeling the ruggedness effects to account for correlation. Namely, one will underestimate the true procedure variance, and possibly validate a procedure that is not truly fit for purpose.

The prediction and tolerance intervals used for validation in the previous section can be easily modified for this correlated condition. In particular, simply replace S^2 with S_{Total}^2 and the error degrees of freedom $(n-1)$ with m (rounded to the nearest integer). The formula for $\bar{Y}$ remains unchanged. Thus, the 90% prediction interval is computed as

$$\begin{aligned}
L &= \bar{Y} - t_{(1+P)/2:m} \sqrt{\left(1 + \frac{1}{a \times r}\right) \times S_{Total}^2} \\
L &= 998.86 - 1.86 \sqrt{\left(1 + \frac{1}{12}\right) \times 19.437} = 990.3 \text{ mg/g} \\
U &= \bar{Y} + t_{(1+P)/2:m} \sqrt{\left(1 + \frac{1}{a \times r}\right) \times S_{Total}^2} \\
U &= 998.86 + 1.86 \sqrt{\left(1 + \frac{1}{12}\right) \times 19.437} = 1007.4 \text{ mg/g}
\end{aligned} \tag{6.26}$$

The 90% tolerance interval that contains 90% of the population is computed as

$$
\begin{aligned}
L &= \bar{Y} - K\sqrt{S^2_{Total}} \\
L &= 998.86 - 2.59\sqrt{19.437} = 987.4 \text{ mg/g} \\
U &= \bar{Y} + K\sqrt{S^2_{Total}} \\
U &= 998.86 + 2.59\sqrt{19.437} = 1010.3 \text{ mg/g}
\end{aligned}
\tag{6.27}
$$

where

$$
\begin{aligned}
K &= \sqrt{\frac{\left(1 + \frac{1}{a \times r}\right) Z^2_{(1+P)/2} \times m}{\chi^2_{0.1:m}}} \\
K &= \sqrt{\frac{\left(1 + \frac{1}{6 \times 2}\right) (1.64)^2 \times 8}{3.49}} = 2.59.
\end{aligned}
\tag{6.28}
$$

Note that both of the computed intervals are wider than their counterparts computed earlier (990.8 to 1006.9 for the prediction interval and 988.5 to 1009.3 for the tolerance interval). This difference occurs for two reasons:

1. S^2_{Total} is generally greater than S^2 and
2. The error degrees of freedom, m, is generally less than $n-1$.

Thus, incorporation of ruggedness effects requires more experimental runs to obtain the same power as a completely independent design. If during pre-validation work a ruggedness factor has been discovered to not impact the intermediate precision, do not include it in the analysis. This will needlessly decrease statistical power.

The intervals in (6.26) and (6.27) can be recommended for validation as described in this chapter. The same substitutions can be applied to the formulas described in Sect. 6.4.6 to estimate OOS.

To finish the example, we now account for the process variation assuming $\rho = 0.5$ and compute the tolerance interval using $S^{2*}_{Total} = S^2_{Total}/(1-\rho) = 38.873$. The resulting tolerance interval from 986.1 to 1011.6 falls within the range from $LSL = 980$ to $USL = 1020$, and the procedure is validated as fit for purpose.

We have considered the case where there is only a single ruggedness factor. If more ruggedness factors are included, more power is lost for a fixed number of experimental runs due to additional partitioning of σ^2_{Total}. Again, one is reminded to not employ ruggedness factors unless they have a demonstrable impact on the intermediate precision.

If a ruggedness factor can be more properly considered as a fixed effect rather than a random effect, power will not be as dramatically impacted. This topic is discussed in the next section.

6.4.11 Incorporating Fixed Effect Ruggedness Factors

In some situations, ruggedness factors are more properly treated as fixed effects. For example, suppose that a major contributor to the intermediate precision of an analytical procedure is the instrument used in the procedure. Suppose there are four instruments in the laboratory, and these will be the only instruments used to perform the procedure in the foreseeable future. Since these are the only four instruments that will be used when performing the procedure, instrument is a fixed effect. Even though it is a fixed effect, differences among the instruments will contribute to the total variation, since only one instrument will be selected for a given application. Thus, it is still necessary to account for this component of variance in the intermediate precision.

As another example, Schwenke and O'Connor (2008) argue that in many cases, analysts can be considered a fixed effect. They argue an analyst is a trained professional proficient on the procedure through a lab-sponsored training program. As such, they are viewed as fixed effects since the training program has made them interchangeable. In many labs, only a small set of analysts perform a given procedure. If they are all used in the validation process, then assuming analyst to be fixed effect is a reasonable assumption.

Analysis of the data in Table 6.8 is now performed assuming analyst to be a fixed effect. The statistical model used to describe the fixed design is

$$\begin{aligned} Y_{ij} &= \tau + \alpha_i + E_{ij} \\ i &= 1, \ldots, a \text{ (analyst)}; \; j = 1, \ldots r; \end{aligned} \tag{6.29}$$

where Y_{ij} is the reportable value for the j^{th} replicate of the i^{th} analyst. The term α_i is a fixed unknown constant that represents the ith analyst and replaces the random variable A_i shown in the random model (6.21). The variance of the a values of α_i is defined as

$$\sigma_\alpha^2 = \frac{\sum_{i=1}^{a} \alpha_i^2}{a-1}. \tag{6.30}$$

The random error E_{ij} has an assumed mean of zero and variance σ_E^2. The parameter τ represents the true value (1000 mg/g for all concentration levels in our example). The total variance associated with the procedure is the sum of the variance components, $\sigma_{Total}^2 = \sigma_\alpha^2 + \sigma_E^2$.

Using an approximation described in Dolezal et al. (1998), all the formulas described in the previous section can be used with a fixed effect by simply replacing n_1 with n_1^* where

$$n_1^* = \frac{[n_1 + 2\lambda]^2}{n_1 + 4\lambda}$$
$$\lambda = \frac{n_1}{2}\left[\frac{S_A^2}{S_E^2}\left(\frac{n_2 - 2}{n_2}\right) - 1\right]. \tag{6.31}$$

In the present example,

$$\lambda = \frac{5}{2}\left[\frac{29.165}{9.708}\left(\frac{6-2}{6}\right) - 1\right] = 2.507$$
$$n_1^* = \frac{[5 + 2 \times 2.507]^2}{5 + 4 \times 2.507} = 6.673 \tag{6.32}$$
$$m = \frac{(19.437)^2}{\frac{(29.165)^2}{2^2 \times 6.673} + \frac{(2-1)^2 \times (9.708)^2}{2^2 \times 6}} = 10.55 = 11 \text{ (rounded)}.$$

Thus, the degrees of freedom used in the prediction and tolerance intervals has increased from $m = 8$ to $m = 11$, and length of the intervals will be properly reduced. In this problem, the 90% tolerance interval that contains 90% of the population computed in (6.27) assuming random analysts is from 987.4 to 1010.3 mg/g. If analysts are treated as fixed, then the interval that results replacing $m = 8$ with $m^* = 11$ is from 988.3 to 1009.5 mg/g. Although the difference is relatively modest in this example, this adjustment will have a major impact on results when $a = 2$ or 3.

This example demonstrates the importance of identifying whether ruggedness factors are fixed or random in the validation experiment. If the validation experiment includes all levels of a ruggedness factor that will be employed in the future, then properly treating it as a fixed effect will increase the likelihood of a successful validation.

6.5 Stage 3: Procedure Performance Verification and Analytical Procedure Transfer

Once the validation is done, it is important to continually monitor the performance of the analytical procedure. A useful statistical tool for this purpose is a control chart of measurements made with the reference standard (refer to Chap. 5 for information on control charts). It is also good practice to perform a system suitability test before every application. USP, ICH, and FDA all provide recommendations as to the need for system suitability tests. Procedures used for this purpose will vary by the procedure and the company.

The purpose of an analytical procedure transfer is to ensure that the receiving laboratory can perform an analytical procedure with the same ability as the transferring laboratory.

6.5.1 Objectives and Regulatory Guidance for Transfers

Some guidance for procedure transfers is provided in General USP Chapter <1224>. The purpose of <1224> is to summarize the types of transfers that may occur, including the possibility of waiver of any transfer, and to outline the potential components of a transfer protocol. However, the chapter does not provide any statistical methods.

A procedure transfer study requires a preapproved transfer protocol that includes details pertaining to the procedure, the sample types being tested, and predetermined acceptance criteria. The acceptance criteria often consider both bias and variability. The acceptance criteria must be satisfied in order to successfully demonstrate the receiving lab is qualified to perform the analytical procedure.

6.5.2 Experimental Designs for Transfers

USP <1224> refers to three types of studies employed in procedure transfer:

1. Comparative testing,
2. Covalidation, and
3. Re-validation.

As described in <1224>, comparative testing requires the analysis of a predetermined number of samples of the same lot by both the transferring and the receiving labs. (More than one lot can be employed if the measurements of the two labs are properly matched.) Covalidation occurs when more than one lab is involved in the initial procedure validation. Re-validation occurs when the receiving lab performs its own independent validation of the procedure.

Statistical designs are not provided in USP <1224>, and so the procedure transfer design is typically determined by the individual company. Consider the summary data in Table 6.11 where each lab is assigned $n = 10$ independent samples from the same lot of material. The response variable is the amount of active ingredient measured in mg.

Table 6.11 Summary of procedure transfer

Parameters	Point estimator	Computed estimate
μ_1—Mean of transferring lab	$\bar{Y}_1$	247.7
μ_2—Mean of receiving lab	$\bar{Y}_2$	249.4
σ_1^2—Variance of transferring lab	S_1^2	10.2
σ_2^2—Variance of receiving lab	S_2^2	27.1

6.5.3 An Equivalence Test for Bias

Bias between the labs is defined as the difference in the lab averages, $\mu_1 - \mu_2$. The equivalence test described in Sect. 2.11 based on the 90% confidence interval on the difference in means can be used for this purpose. Since the two samples are independent, the appropriate confidence interval on the mean difference is provided in Eq. (2.58) where we assume variances are not equal. Based on historic reference sample measurements in the transferring lab, the EAC is taken to be 10 mg. Thus, the 90% confidence interval on the difference $\mu_1 - \mu_2$ must fall entirely within the range from -10 to $+10$ mg.

We begin by computing the degrees of freedom for the confidence interval.

$$\begin{aligned} df &= \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}} \\ &= \frac{\left(\frac{10.2}{10} + \frac{27.1}{10}\right)^2}{\frac{(10.2)^2}{(10)^2(10-1)} + \frac{(27.1)^2}{(10)^2(10-1)}} = 14.9 = 15 \text{ (rounded)}. \end{aligned} \tag{6.33}$$

The lower and upper bounds of the confidence interval are now computed as

$$\begin{aligned} L &= \bar{Y}_1 - \bar{Y}_2 - t_{1-\alpha/2:df}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ &= 247.7 - 249.4 - 1.75\sqrt{\frac{10.2}{10} + \frac{27.1}{10}} = -5.1 \text{ mg} \\ U &= \bar{Y}_1 - \bar{Y}_2 + t_{1-\alpha/2:df}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \\ &= 247.7 - 249.4 - 1.75\sqrt{\frac{10.2}{10} + \frac{27.1}{10}} = 1.7 \text{ mg}. \end{aligned} \tag{6.34}$$

Since the 90% confidence interval falls entirely within the range from -10 to $+10$ mg, equivalence of means between the laboratories has been demonstrated.

Rugaiganisa (2016) has proposed an approach for setting the EAC for the equivalence test. As an alternative to this procedure, one may wish to establish transfer criteria using an ATP as described by Martin et al. (2013).

6.5.4 Tests for Precision

Precision of a procedure is described by the magnitude of the variance (or standard deviation). As discussed in Schwenke and O'Connor (2008), the necessity and form of an equivalence test for precision is not obviously apparent. If the receiving lab provides better precision, that is a good thing, even if the two procedure precisions are not equivalent. Thus, a difference testing approach as opposed to equivalence testing might be considered.

In particular, one might test the null hypothesis $\sigma_1^2 \geq \sigma_2^2$ versus the alternative hypothesis $\sigma_1^2 < \sigma_2^2$ and conclude the receiving lab is no worse than the transferring lab if one does not reject the null hypothesis. Such a test could be performed by computing an upper bound on the ratio σ_1^2/σ_2^2 as described in Sect. 2.8.3. If this upper bound is greater than 1, then one is unable to reject the null hypothesis, and will conclude the procedure transfer is successful. However, one would need to ensure that the power associated with the test is sufficiently high to discover a situation where $\sigma_1^2 < \sigma_2^2$.

Alternatively, one might compute a range of expected variances in the receiving lab based on the variance in the transferring lab. The receiving lab passes the transfer if the computed variance falls in this range. Such an approach is similar to using a control chart or a system suitability test. A 95% upper prediction bound based on n_1 observations to contain the variance of a future sample of size n_2 from the same normal population is

$$U = S_1^2 \times F_{0.95, n_2 - 1, n_1 - 1} \tag{6.35}$$

(see page 64 of Hahn and Meeker (1991)).

That is, if the receiving lab is performing the procedure in the same manner as the transferring lab, the transfer criteria are satisfied if $S_2^2 \leq U$ where U is defined in (6.35). To demonstrate, using the data in Table 6.11,

$$\begin{aligned} U &= S_1^2 \times F_{0.95, n_2 - 1, n_1 - 1} \\ &= 10.2 \times 3.18 = 32.4. \end{aligned} \tag{6.36}$$

Since $S_2^2 = 27.1$ is less than 32.4, the procedure transfer satisfies the precision requirement.

References

Burdick RK, Borror CM, Montgomery DC (2005) Design and analysis of gauge R&R experiments: making decisions with confidence intervals in random and mixed ANOVA models. ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia

Chatfield MJ, Borman PJ (2009) Acceptance criteria for method equivalency assessments. Anal Chem 81:9841–9848

Dolezal KK, Burdick RK, Birch NJ (1998) Analysis of a two-factor R&R study with fixed operators. J Qual Technol 30:163–170

Food and Drug Administration. Center for Drugs Evaluation Research (2011) Process validation: general principles and practices, guidance for industry

Food and Drug Administration. Center for Drugs Evaluation Research (2015) Analytical procedures and methods validation for drugs and biologics, guidance for industry

Hahn GJ, Meeker WQ (1991) Statistical intervals: a guide for practitioners. Wiley, New York

Hubert Ph, Nguyen-Huu J-J, Boulanger B, Chapuzet E, Chiap P, Cohen N, Compagnon P-A, Dewé W, Feinberg M, Lallier M, Laurentie M, Mercier N, Muzard G, Nivet C, Valat L (2004) Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal—part I. J Pharm Biomed Anal 36:579–586

Hubert Ph, Nguyen-Huu J-J, Boulanger B, Chapuzet E, Chiap P, Cohen N, Compagnon P-A, Dewé W, Feinberg M, Lallier M, Laurentie M, Mercier N, Muzard G, Nivet C, Valat L, Rozet E (2007a) Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal—part II. J Pharm Biomed Anal 45:70–81

Hubert Ph, Nguyen-Huu J-J, Boulanger B, Chapuzet E, Cohen N, Compagnon P-A, Dewé W, Feinberg M, Laurentie M, Mercier N, Muzard G, Valat L, Rozet E (2007b) Harmonization of strategies for the validation of quantitative analytical procedures: a SFSTP proposal—part III. J Pharm Biomed Anal 45:82–96

International Conference on Harmonization (2005) Q2 (R1) Validation of analytical procedures: text and methodology

International Conference on Harmonization (2009) Q8 (R2) Pharmaceutical development

Krishnamoorthy K, Mathew T (2009) Statistical tolerance regions. Wiley, Hoboken

Martin GP, Barnett KL, Burgess C, Curry PD, Ermer J, Gratzl GS, Hammond JP, Herrmann J, Kovacs E, LeBlond DJ, LoBrutto R, McCasland-Keller AK, McGregor PL, Nethercote P, Templeton AC, Thomas DP, Weitzel J (2013) Stimuli to the revision process: lifecycle management of analytical procedures: method development, procedure performance qualification, and procedure performance verification. Pharm Forum 39(5). http://www.usp.org/uspnf/notices/stimuli-article-lifecyclemanagement-analytical-proceduresposted-comment. Accessed 11 Mar 2014

Ntzoufras I (2009) Bayesian modeling in WinBUGS. Wiley, New York

Rugaiganisa A (2016) Comparative analytical method transfer. Presentation at the 2016 ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop, September

Schwenke JR, O'Connor DK (2008) Design and analysis of analytical method transfer studies. J Biopharm Stat 18:1013–1033

Spiegelhalter D, Thomas A, Best N, Gilks W (1996) BUGS 0.5 Examples Volume 1(version i). Accessed 21 April 2014. http://www.mrc-bsu.cam.ac.uk/bugs/

USP Statistics Expert Team (2016) In-process revision: <1210> Statistical tools for procedure validation. Pharm Forum 42(5). http://www.usppf.com/pf/pub/index.html

USP 39-NF 34 (2016) General Chapter <1010> analytical data—interpretation and treatment. US Pharmacopeial Convention, Rockville

USP 39-NF 34 (2016) General Chapter <1030> biological assay chapters—overview and glossary. US Pharmacopeial Convention, Rockville

USP 39-NF 34 (2016) General Chapter <1032> design and development of biological assays. US Pharmacopeial Convention, Rockville

USP 39-NF 34 (2016) General Chapter <1033> biological assay validation. US Pharmacopeial Convention, Rockville

USP 39-NF 34 (2016) General Chapter <1223> validation of alternative microbiological methods. US Pharmacopeial Convention, Rockville

USP 39-NF 34 (2016) General Chapter <1224> transfer of analytical procedures. US Pharmacopeial Convention, Rockville

USP 39-NF 34 (2016) General Chapter <1225> validation of compendial procedures. US Pharmacopeial Convention, Rockville

Weitzel MLJ (2012) The estimation and use of measurement uncertainty for a drug substance test procedure validated according to USP<1225> Accred Qual Assur 17:139–146

Wolfinger RD (1998) Tolerance intervals for variance component models using Bayesian simulation. J Qual Technol 30:18–32

Yang H, Zhang J (2015) Validation based on total error: a generalized pivotal quantity approach to analytical method. PDA J Pharm Sci Technol 69:725–735