Chapter 9
Analytical Comparability and Similarity

Keywords Accelerated stability • Analytical comparability • Analytical similarity • Biosimilar products • Comparability criterion • Equivalence testing • Non-profile data • Power calculations • Profile data • Scale comparisons • Stability data • Tolerance intervals

9.1 Introduction

In all manufacturing settings, there is an inherent drive to improve product through the reduction in process variation, implementing new technology, increasing efficiency, optimizing resources, and improving customer experience through innovation. In the pharmaceutical industry, these improvements come with added responsibility to the patient such that product made under the post-improvement or post-change condition maintains the safety and efficacy of the pre-change product. As described in FDA comparability guidance (1996) and ICH Q5E (2004), regulatory agencies also recognize the importance in providing manufacturers the flexibility to improve their manufacturing processes. Agencies also acknowledge that some changes may not require additional clinical studies to demonstrate safety and efficacy so that implementation may be more efficient and expeditious to benefit patients. Activities performed when changes are made to the process include demonstration of comparability in product parameters. The actual timing of each activity and the statistical rigor required for the evaluation of pre- and post-change product is linked to the stage of the product development (e.g., clinical versus commercial material) and the scope of the change (e.g., process transfer with similar scale versus a new cell line or formulation).

To set the stage for this chapter, the requirements of comparing pre- and post-change product are reviewed. Comparability is defined by ICH Q5E as a demonstration that the quality attributes of the pre- and post-change product are highly similar and that the existing knowledge is sufficiently predictive to ensure that any differences in quality attributes have no adverse impact on safety or efficacy of the drug product. Guidance provides the manufacturer with flexibility to adjust study rigor based on the stage of development and prior knowledge.

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_9

9.2 Statistical Methods for Comparability

The FDA comparability guidance (1996) recognized the need for manufacturers to improve manufacturing processes and analytical methods without performing additional clinical studies to demonstrate product safety and efficacy. This guidance was extended in ICH Q5E to provide additional direction for comparing pre- and post-change manufacturing processes. The direction is related to the scope of the comparability exercise and the type of change under consideration. Major process changes should consider a larger array of testing than those of lesser scope. For example, a change of major scope might reasonably need to consider additional pharmacokinetic (PK) or clinical studies, whereas a change with lesser scope may rely only on analytical comparability for a set of critical quality attributes (e.g., biological activity, purity, and protein structure).

Although product comparability guidance does not cover the comparison of in-process parameters given similar process changes (e.g., site transfers, scale changes, and equipment improvements), these issues are addressed in FDA guidance (2011) and were discussed in Chaps. 3–5 of this book.

Across the regulatory documents, there are only high level recommendations for the design of a comparability study and for setting acceptance criteria to assess the impact of the change. These documents do not contain prescriptive rules for setting acceptance criteria, study design, or statistical methods for analysis. This chapter provides examples of how these issues might be addressed. The study design and statistics associated with clinical, PK, and animal studies are out of scope.

The design and scope of an analytical comparability study will vary depending on the product and process complexity, complexity of the change, and the stage of the clinical/commercial life cycle. The analytical methods used for analytical comparability minimally include lot release. In addition, non-routine methods may be used to further understand the impact of the change on the biochemical, biophysical, and biological properties of the product. Comparison of degradation rates and degradation profiles from select analytical methods may also enhance the understanding of the change on key product degradants. Typically, the conditions considered for evaluating degradation rates are harsher than recommended storage conditions. By design, the degradation observed for a product at recommended storage conditions is small. Given the short period of time typically available for implementing a change, evaluation of degradation rates under recommended storage conditions provides only minimal insight into how a post-change molecule degrades over time. Instead, the pre- and post-change products are held at stressed stability conditions. These conditions may be used to detect potential impurities and structural modifications not otherwise detected by lot release and in-process control testing of non-degraded material. Analytical procedures used during the assessment of drug substance and drug product comparability should be validated or qualified as appropriate for their intended purpose (refer to Chap. 6 for more on this topic).

Chatfield et al. (2011) provide a nice description of statistical techniques that are useful for demonstrating comparability. They differentiate between the statistical

Table 9.1 A summary of the comparability approaches

Comparability approach	Lot release	Stability at recommended storage conditions	Stability at stressed storage conditions	Characterization methods
Comparison of individual values				
Visual comparisons	X	X	X	X
Tolerance intervals	X			X
Specifications	X	X		
Limit evaluations	X	X		X
Comparison of summary measures				
Equivalence testing	X		X	X

equivalence tests described in Chap. 2 and other comparability approaches. Statistical equivalence testing provides a formal statistical approach in which statistical decision errors can be controlled. Equivalence testing is particularly desirable when

1. Summary measures such as means and slopes are relevant to the process change
2. Acceptance criteria that considers scientific importance can be defined a priori, and
3. Data are amenable to the statistical requirements of equivalence testing.

If these conditions are satisfied, then equivalence testing provides the strongest scientific evidence of comparability. Attributes that are not amenable to equivalence testing can be evaluated with alternative comparability approaches including graphical summaries and comparison of individual values to pre-specified ranges.

Table 9.1 presents typical comparability approaches categorized by application. Tolerance intervals and equivalence testing are discussed in Chap. 2 and will be demonstrated with examples in this chapter.

Once an approach has been selected, the comparability study is designed. Since equivalence testing involves a statistical test, concepts typically associated with hypothesis testing such as error rate and statistical power are employed to select an appropriate study design. With the other approaches, the design will be determined by availability of relevant pre-change data and heuristic rules used for defining comparability. The final step is to collect the post-change data and compare it against the pre-stated criteria. If all of the criteria are met, one declares the product manufactured under the post-change process is comparable to the product manufactured under the pre-change process. When one or more acceptance criteria are not met, further investigation is required to determine if the pre- and post-change product is comparable. This could include further characterization, analytical method improvement, or the performance of additional nonclinical or clinical studies.

9.2.1 Lot Release

Minimally, results from the post-change process are compared to the currently approved lot release specification limits regardless of where the molecule is in its development life cycle. As the complexity of the change increases, additional assessment criteria may be required.

It may be desirable to retest pre-change lots at the same time that post-change lots are tested. By recording these measurements in the same analytical run, uncertainty related to precision and changes to analytical methods over time will be mitigated. In order to use such a design, it must be assumed that product stability (freeze-thaw or storage stability) has been demonstrated to be negligible.

Lot release tests are implemented early in clinical drug development to assess safety and efficacy prior to product release. A subset of lot release tests may be selected for comparability assessment. Typically these methods provide a quantitative assessment of critical product quality attributes. As product progresses through clinical development, a data set of analytical test results is accumulated. These data allow an ongoing assessment of patient exposure to levels of product quality attributes that may vary from lot-to-lot. Specifications and comparability assessment criteria may be adjusted using these data during clinical development as patient exposure experience is gained. These limits should factor in the ranges of analytical test data as well as the statistical and operational components that influence the variability of the analytical method (determined during method validation) and the variability of the process (determined during process characterization).

9.2.2 Stability at Recommended and Stressed Storage Conditions

Stability at recommended storage condition is typically assessed by comparing post-change stability results to pre-change stability data. The appropriate stability indicating assays are identified and implemented as part of the normal GMP stability program during clinical and commercial development. As product progresses through clinical development, a data set of stability test results is accumulated. These data allow an ongoing assessment of patient exposure to drug substance/drug product stability profiles that may vary due to manufacturing and formulation variability. Generally speaking, there is typically little to no degradation of product observed under recommended storage conditions. Appropriate recommended storage stability comparability limits can be set using the specification, visual assessments (chromatographic overlays), or limit tests. Because the degradation profile estimated from the post-change data at recommended storage is generally not extrapolated to the established expiry, care must be taken in setting the acceptance criterion if an equivalence test is performed. This is because the pre-change product will have a slope which is estimated across the entire range of shelf life whereas the post-change product will have limited information to estimate the slope. This difference in range causes the variances for the two slope estimates to differ, even if the processes are identical.

Stability at a stressed storage condition is also assessed by comparing post-change stability results to the pre-change stability data. The stressed stability conditions may be conducted under elevated temperatures or other stressed conditions such as chemically induced oxidation. The selection of the stressed condition will be based on the primary degradation pathway. In some cases one or more analytical methods may detect degradation of the product. For stressed stability studies, it is desired to compare the slope of the post-change data to the slope of the pre-change data. Since the comparison of interest concerns the summary measure slope, a test of equivalence is an appropriate choice to assess comparability. Such a test is demonstrated in Sect. 9.4.2. It may also be appropriate to include a visual assessment such as chromatographic overlays at specified time points. Appropriate accelerated storage stability comparability criteria can be established and adjusted during clinical development as patient exposure experience is gained.

9.2.3 Characterization Methods

Biochemical, biological, and biophysical analyses are performed on new process lots as appropriate. These lots are compared side-by-side to representative pre-change lots as well as to the current reference standard. A side-by-side study is conducted when an analytical method is not used routinely. A predetermined number of batches are collected from both the pre-change and post-change process and placed on the assay at the same point in time. This way, any differences associated with the analytical method will not manifest itself as differences between the two processes.

9.3 Comparability Examples for Individual Post-change Values

In this section, several examples are provided where criteria for post-change individual values are represented as ranges based on pre-change expectations. Most typically, comparability is demonstrated if a defined percentage of the post-change individual values fall within these ranges. The range criteria are computed with pre-change data using tolerance intervals. In some cases, specification limits or an LOQ may be appropriate for defining such criteria.

The following examples demonstrate how to compute prediction and tolerance intervals for several types of data structures. Chapter 2 provides the formulas that are demonstrated in this section. The three-step comparability approach is as follows:

1. Plot the data and visually compare the two groups.
2. Compute a tolerance interval using pre-change data.
3. Assess the post-change data by determining the percentage of post-change values that fall in the tolerance interval.

Demonstration of comparability requires a pre-specified proportion of post-change observations falling within the tolerance interval. Therefore, the width of the computed interval is a key component in setting the interval-based acceptance criteria. Confidence levels and proportions contained in a tolerance interval are often based on the amount of both pre- and post-change data. Dong et al. (2015b) offer considerations when using tolerance intervals to define the quality of a pre-change process. In general, if the pre- and post-change data sets are small, confidence intervals and coverage proportions must be reduced to provide meaningful intervals. Use of 99% confidence or 99% coverage with small data sets will result in intervals that are too wide to be useful in assessing comparability. In such cases, specifications or other limit evaluations may be required to serve as criteria. The pre-change data used to compute tolerance intervals must be assessed against the statistical assumption of normality as described in Chap. 2.

9.3.1 Combining Pre-change Data Sets at Different Scales

This example considers a process transfer where pre-change data are available from two manufacturing scales: a clinical scale and a commercial scale from a licensed facility. The process in the licensed facility is to be transferred to a different commercial facility at the same commercial scale (i.e., the post-change facility). The parameter of interest is an in-process control parameter that measures yield in kilograms with a specification of 40.8–75.0 kg. Figure 9.1 presents a plot of the pre-change data. The $n_1 = 5$ lots on the left are from the clinical scale, and the $n_2 = 5$ lots on the right are from the commercial scale. It is clear from the plot that the yields differ between the clinical and commercial scale processes. The spread in the data for the clinical and commercial scale appears similar.

These data are now combined to construct a tolerance interval to provide a comparability criterion. Yields from the post-change process will be expected to fall in this range. Since the spreads of the two scales in Fig. 9.1 are comparable, it is desired to pool (combine) the two data sets for estimating the pre-change variance. Since the commercial scale best represents the expected average of the post-change facility, it is desired to center a tolerance interval on the commercial scale average.

The tolerance interval formula in Eq. (2.23) can be used to compute the desired interval with some slight modifications. In particular, $\bar{Y}$ now represents the sample mean of the commercial scale, and S^2 is replaced with the pooled variance estimate

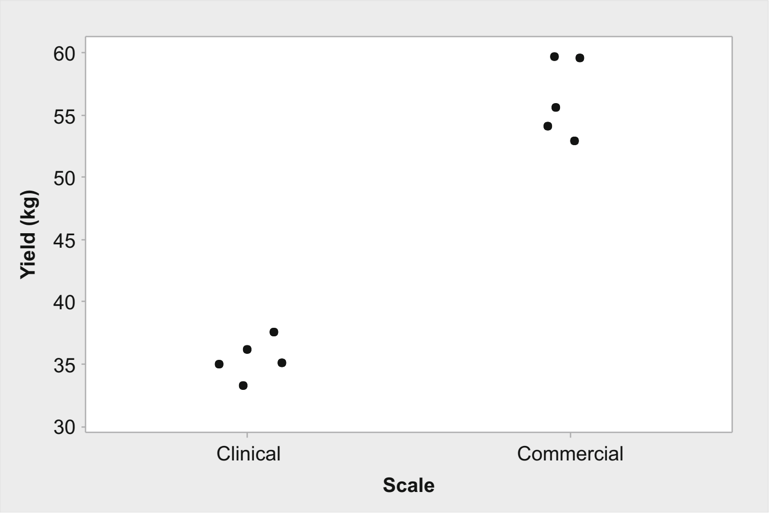


Fig. 9.1 Clinical and commercial scale pre-change data

Table 9.2 Values required to compute tolerance interval

Statistic description	Notation	Example values
Sample mean of commercial scale	$\bar{Y}$	56.36
Pooled variance	S_P^2 in (2.56)	6.21
Error degrees of freedom	$\nu = n_1 + n_2 - 2$	8
Effective sample size	$n_e = n_2$	5
Confidence level	$(1 - \alpha)$	0.95
Proportion contained	P	0.99

of the two scales. This pooled variance is denoted S_P^2 and is defined in Eq. (2.56). The value of K is defined in Eq. (2.25) with $n_e = n_2 = 5$ and $\nu = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$. The required values to compute a two-sided 95% tolerance interval with 99% coverage are shown in Table 9.2.

The computed interval using (2.23) with K defined in (2.25) is

$$L = \bar{Y} - K\sqrt{S^2}$$

$$U = \bar{Y} + K\sqrt{S^2}$$

$$K = \sqrt{\frac{\left(1 + \frac{1}{n_e}\right) Z^2_{(1+P)/2} \times \nu}{\chi^2_{\alpha:\nu}}} = \sqrt{\frac{\left(1 + \frac{1}{5}\right) 2.58^2 \times 8}{2.73}} = 4.83 \tag{9.1}$$

$$L = 56.36 - 4.83 \times \sqrt{6.21} = 44.3$$

$$U = 56.36 + 4.83 \times \sqrt{6.21} = 68.4$$

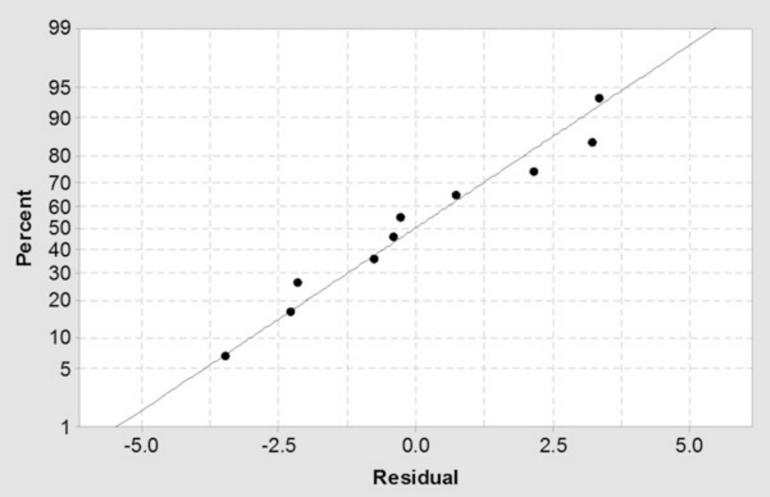


Fig. 9.2 Normal quantile plot of yield residuals

The 95% tolerance interval containing 99% of all future observation is 44.3–68.4 kg which falls within the in-process specification of 40.8–75.0 kg. The residuals formed by subtracting the appropriate scale mean from each observation are plotted in the normal quantile plot shown in Fig. 9.2. The plot suggests the normality assumption is reasonable.

The advantage of combining the two data sets to estimate the variance is seen by comparing the computed interval in (9.1) to an interval based solely on the commercial lots. This calculation is

$$\begin{aligned}
L &= \bar{Y} - K\sqrt{S^2} \\
U &= \bar{Y} + K\sqrt{S^2} \\
K &= \sqrt{\frac{\left(1+\frac{1}{n_e}\right)Z^2_{(1+P)/2} \times \nu}{\chi^2_{\alpha:\nu}}} = \sqrt{\frac{\left(1+\frac{1}{5}\right)2.58^2 \times 4}{0.71}} = 6.69 \\
L &= 56.36 - 6.69 \times \sqrt{9.90} = 35.3 \\
U &= 56.36 + 6.69 \times \sqrt{9.90} = 77.4
\end{aligned} \tag{9.2}$$

This interval is so wide that it exceeds the specification range of 40.8–75.0 kg and has no value as a comparability range.

Figure 9.3 presents the computed tolerance interval (dashed line) and the specifications (solid line) with the pre-change data used in the computations. The yields from the post-change facility are expected to fall in the tolerance interval.

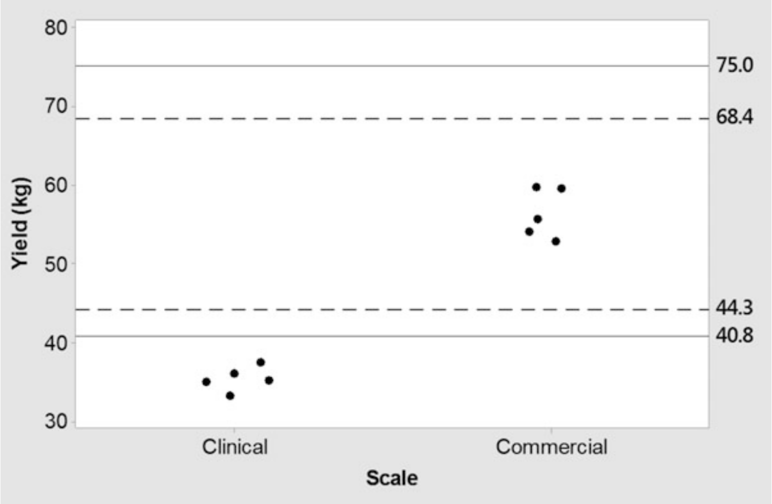


Fig. 9.3 Specifications and tolerance interval for yield data

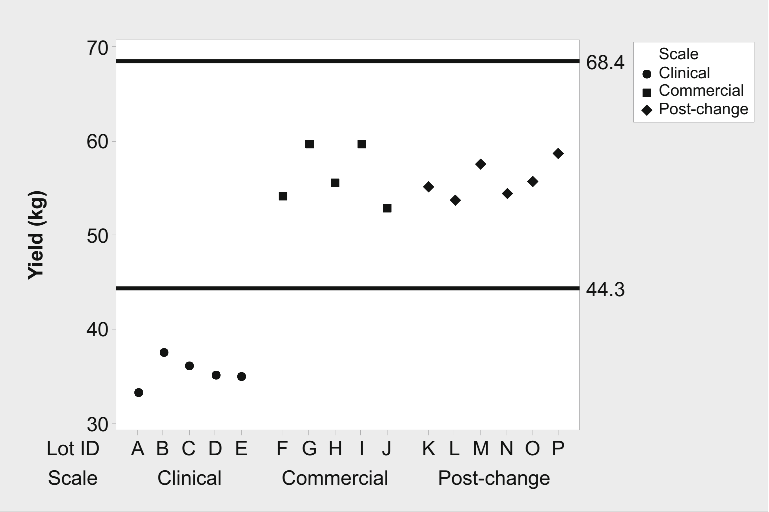


Fig. 9.4 Post-change yield data

Figure 9.4 presents the computed intervals using the pre-change data with the first six post-change yield values. Since all post-change values fall within the tolerance interval, comparability has been demonstrated.

9.3.2 Tolerance Intervals with Replicate Measures on Each Lot

Unlike the data presented in Sect. 9.3.1, this example considers a situation where there are replicate measurements taken on all or a portion of the batches. The appropriate formula for a tolerance interval for these data is provided in Eq. (2.52). Figure 9.5 presents a data set with $r = 5$ purity measurements taken on each of $a = 11$ lots in a pre-change data set. The lot release specification for this parameter is a one-sided specification of ≥40.0%. The data are plotted in time order. Calculation of the 95% tolerance interval that contains 99% coverage is shown in Table 9.3. The Hoffman and Kringle interval referenced in Sect. 2.7.4 is from 42.7 to 57.8 (calculations on spreadsheet at website).

Figure 9.6 is a plot of the pre-change data used to compute the tolerance intervals, the specification (solid line), and the two-sided tolerance interval (long dashed lines). Note that although the specification is one-sided on the lower end, the comparability tolerance interval is still two-sided. Comparability is a comparison of the two processes, apart from the specification. It is possible that two processes are not comparable, but are both capable of meeting specification.

As an alternative to the interval computed in Table 9.3, one may choose to compute the comparability interval using lot averages instead of individual values. In this case, the averages are independent across lots, and so the tolerance interval is computed using the independent formulas in Sect. 2.6.7. Figure 9.7 presents a plot of the lot averages for the data set in Fig. 9.5, and Table 9.4 provides the tolerance interval calculations based on formulas (2.23) and (2.25).

Figure 9.8 includes the tolerance intervals with the plot of lot averages.

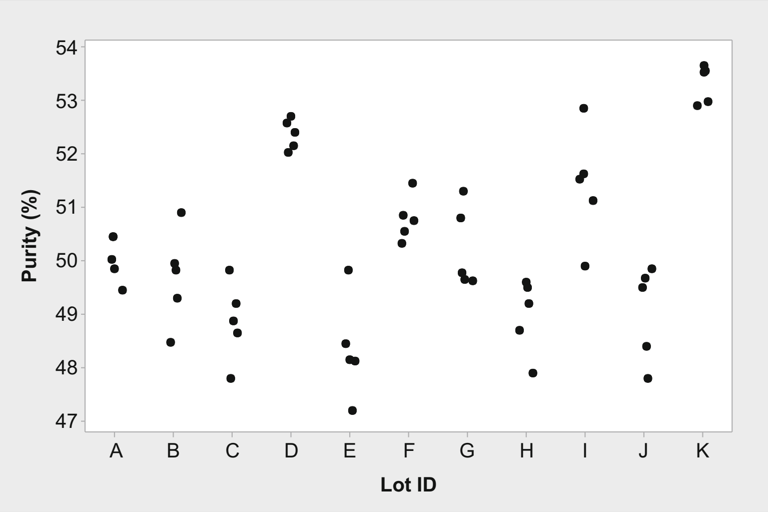


Fig. 9.5 Pre-change batch purity (%) plotted in time order

Table 9.3 Statistics needed to compute tolerance interval

Statistic	Value
Confidence level $100(1-\alpha)\%$	95%
Proportion covered $100P\ \%$	99%
a	11
r	5
$\bar{Y}$	50.261
S_A^2	12.260
S_E^2	0.529
S_{Total}^2 from Eq. (2.46)	2.875
m from Eq. (2.50)	13.65
m (rounded)	14
K	3.794
$Z_{\frac{1+P}{2}}$	2.576
L from (2.52)	43.8
U from (2.52)	56.7

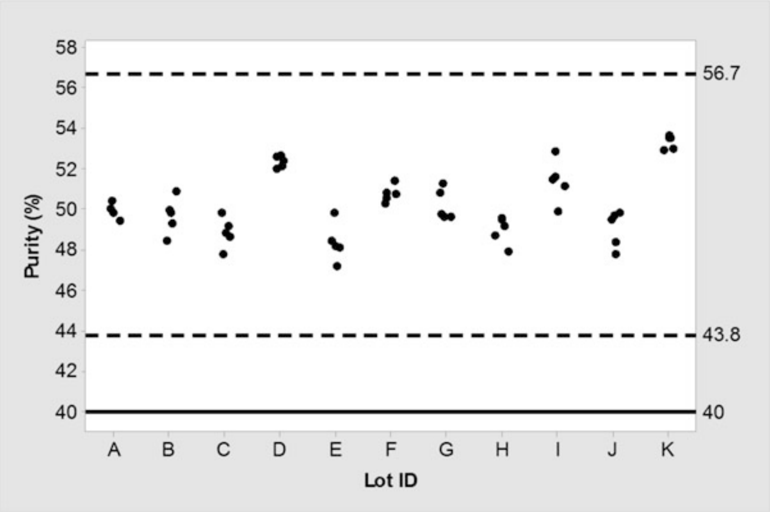


Fig. 9.6 Pre-change data acceptance criteria and specification

As expected, the tolerance interval in Fig. 9.8 is tighter than the tolerance interval in Fig. 9.6. This is because variability in lot means is smaller than variability in individual values. One may construct comparability limits in either manner as long as consistency is maintained between the limits and the post-change values begin compared (i.e., either individual values or lot averages).

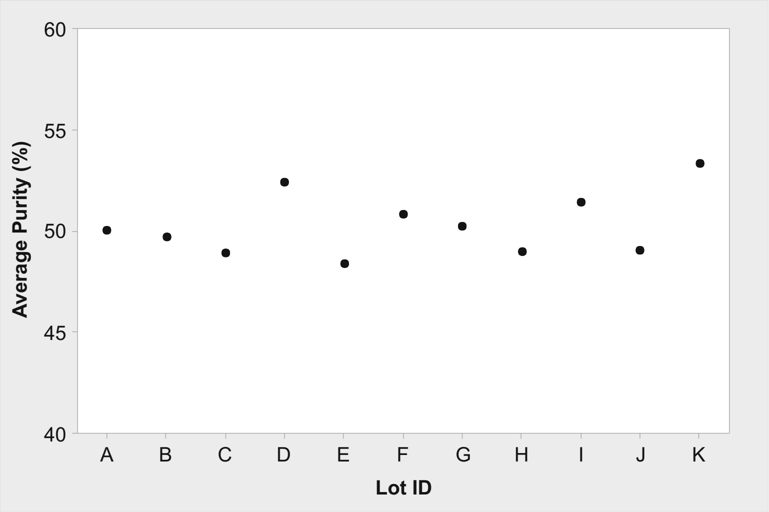


Fig. 9.7 Plot of lot averages for purity (%)

Table 9.4 Statistics needed to compute tolerance interval

Statistic	Value
Confidence level $100(1-\alpha)\%$	95%
Proportion covered $100P\,\%$	99%
n_e	11
$\bar{Y}$	50.261
S^2	2.452
ν	10
$\chi^2_{\alpha:\nu}$	3.940
$Z_{\frac{1+P}{2}}$	2.576
K from (2.25)	4.286
L from (2.23)	43.6
U from (2.23)	57.0

9.4 Equivalence Testing for Summary Parameters

When summary parameters are informative, equivalence testing provides the strongest statistical evidence of comparability. Equivalence testing is discussed in Sect. 2.11. Data used in an equivalence test can be either profile or non-profile. Non-profile data are collected at a single point in time. Examples of non-profile data include lot release or in-process control measurements. Profile data are collected over time. In the context of comparability, a stability profile is of interest when data are collected in this manner. Typically, non-profile data involve a comparison of averages, and profile data involve a comparison of slopes.

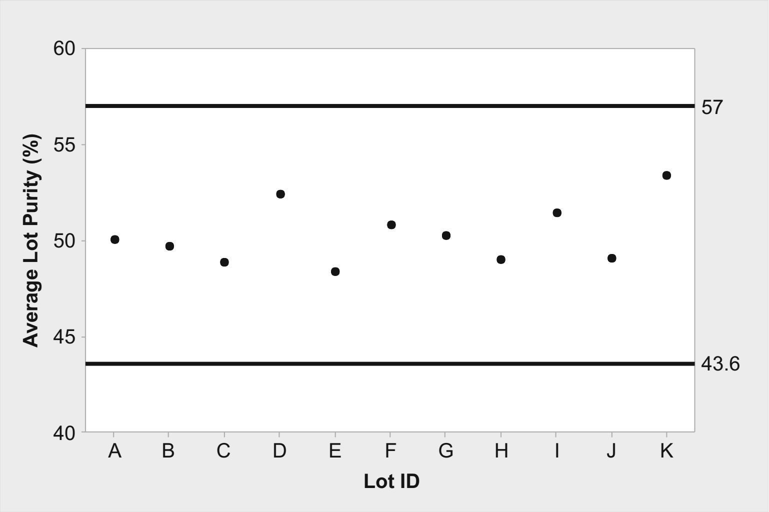


Fig. 9.8 Tolerance interval based on lot averages

As noted in Sect. 2.11, the most challenging part of an equivalence test is often establishment of the equivalence acceptance criterion (EAC). In the next two sections, guidance is offered for both non-profile and profile data.

9.4.1 Equivalence Acceptance Criterion for Non-profile Data

The equivalence hypotheses used to demonstrate comparability with non-profile data are

$$\begin{aligned} H_0 &: |\mu_{Pre} - \mu_{Post}| \geq \text{EAC} \\ H_a &: |\mu_{Pre} - \mu_{Post}| < \text{EAC} \end{aligned} \tag{9.3}$$

where the subscripts denote the pre- and post-change conditions, respectively. As discussed in Sect. 2.11, equivalence is assessed by constructing a two-sided 100 $(1 - 2\alpha)\%$ confidence interval on the difference $\mu_{Pre} - \mu_{Post}$. The null hypothesis H_0 in Eq. (9.3) is rejected and equivalence is demonstrated if the entire confidence interval falls in the range from $-$EAC to $+$EAC.

When evaluating product comparability, the EAC defines the maximum difference in means that has no practical scientific impact. It is ideal if a subject matter expert (SME) can define the EAC. In the absence of an SME definition, parameters defined by specifications or other decision making limits are used in the decision process.

To provide an example, consider a situation where the lot release specification for protein concentration is LSL = 58.5 mg/mL and USL = 71.5 mg/mL.

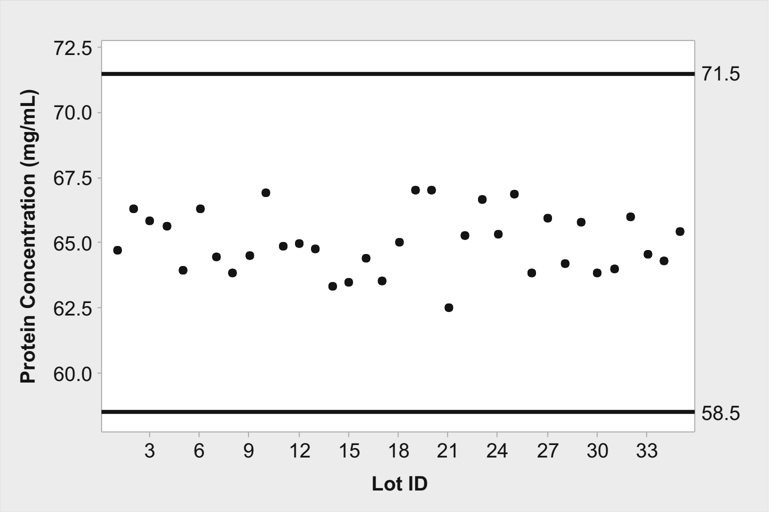


Fig. 9.9 Protein concentration (mg/mL) by lot

Table 9.5 Descriptive statistics for pre-change protein concentration

Statistic	Value (mg/mL)
Mean ($\bar{Y}$)	65.00
Standard deviation (S)	1.18
Minimum	62.47
Maximum	67.02
Lot count	35

The pre-change process data are plotted in Fig. 9.9, and the descriptive statistics are listed in Table 9.5. The data are independent with one value for each lot.

One might ask the question, "Given the pre-change process mean is 65.00 mg/mL and the standard deviation is 1.18 mg/mL, what is the maximum allowable shift in the post-change mean that would not cause an unacceptable probability for an out-of-specification (OOS) observation?" This question can be answered by using a process capability index as presented in Eq. (5.13). This capability measure is

$$\hat{C}_{pk} = \min\left[\frac{USL - \bar{Y}}{3S}, \frac{\bar{Y} - LSL}{3S}\right] \tag{9.4}$$

Using the values in Table 9.5, the computed capability measure is

$$\begin{aligned} C_{pk} &= \min\left[\frac{USL-\mu_Y}{3\sigma_{st}}, \frac{\mu_Y-LSL}{3\sigma_{st}}\right] \\ &= \min\left[\frac{71.5-65.00}{3\times 1.18}, \frac{65.00-58.5}{3\times 1.18}\right] \\ &= \min[1.84, 1.84] = 1.84 \end{aligned} \tag{9.5}$$

For this example, the two quantities within the parentheses are equal, implying that the process is centered within the specification. For this example, let's assume that a capability of 1.5 is acceptable. This corresponds to a 0.0007% chance of an individual value falling outside of the specification limits (see Montgomery 2013). The largest mean protein concentration for the post-change process that meets this requirement if the process shifts to the right is computed as follows:

$$\begin{aligned} 1.5 &= \frac{71.5-\mu_{Post}}{3\times 1.18} \\ \mu_{Post} &= 71.5 - 1.5\times 3\times 1.18 = 66.19\ \text{mg/mL}. \end{aligned} \tag{9.6}$$

Thus, the allowable shift from the present position is computed as 66.19-65.00 = 1.19 mg/mL, or the equivalence acceptance criterion is EAC = 1.19 mg/mL. Because the process is centered, a shift of the post-process change to the left would provide the same EAC.

Figure 9.10 presents a simulated post-change data set with a mean of 66.19 mg/mL and standard deviation of 1.18 mg/mL next to the pre-change data

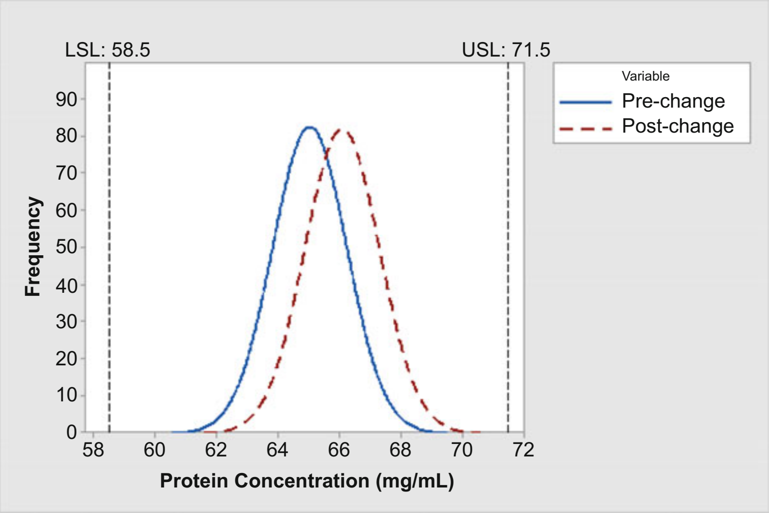


Fig. 9.10 Graphical representation of an acceptable process shift

distribution. This provides an effective representation of a process shift at the EAC. It is clear from Fig. 9.10 that the amount of tolerable shift is small in order to maintain a low percentage of observations exceeding the upper specification limit.

When no specification or SME defined EAC is available, one can define an EAC based on behavior of the pre-change process and a visual assessment as in Fig. 9.10. This approach describes the difference in means based on "expected" behavior of the pre-change process as opposed to "acceptable" behavior in terms of safety or efficacy. The notion of "expected" behavior is proposed by Hauk et al. (2008). Pre-change process behavior is described with a statistical model that incorporates the pre-change process average, lot-to-lot variation, and intermediate precision of the analytical method.

The EAC is defined as the acceptable shift in population means expressed in terms of the standard deviation of the pre-change response variance. This ratio is called the effect size (ES). The ES is defined as

$$\mathrm{ES} = \frac{|\mu_{Pre} - \mu_{Post}|}{\sigma_{Pre}} \tag{9.7}$$

and discussed in Sect. 2.8.2. An EAC describing the pre-change process is defined as a function of an acceptable value for ES. In particular,

$$\mathrm{EAC} = \mathrm{ES} \times \sigma_{Pre} \tag{9.8}$$

where σ_{Pre} is estimated based on a sample of pre-change lots. In some situations, it may be reasonable to replace σ_{Pre} with an upper bound to account for sampling error. This approach has been advocated by Limentani et al. (2005) using a confidence coefficient of 80%.

An acceptable value of ES will depend on the application and rigor required to demonstrate equivalence. For example, demonstration of analytical similarity of a biosimilar may have a smaller ES compared to a demonstration of comparability for a process transfer. Selection of ES in Eq. (9.8) is aided using SMEs and visual representations. By visually representing a variety of ES values, the SME can evaluate the overlap of the pre-change data and simulated post-change data.

Figure 9.11 represents four possible values of ES with the corresponding overlapping coefficient as defined by Inman and Bradley (1989). This overlapping coefficient is defined as

$$\mathrm{OVL} = 2 \times \Phi\left(-\frac{|\mu_{\mathrm{Pre}} - \mu_{Post}|}{2\sigma}\right) \tag{9.9}$$

where $\sigma_{\mathrm{Pre}} = \sigma_{Post} = \sigma$ and $\Phi(\bullet)$ is the cumulative function of a standard normal random variable. For example, if two distributions differ by one standard deviation, then

$$\begin{aligned}\mathrm{OVL} &= 2 \times \Phi\left(-\frac{|\mu_{\mathrm{Pre}} - \mu_{Post}|}{2\sigma}\right)\\ &= 2 \times \Phi\left(-\frac{\sigma}{2\sigma}\right) = 2 \times \Phi\left(-\frac{1}{2}\right) = 2 \times 0.309 = 0.62\end{aligned} \tag{9.10}$$

The pre-change population is represented by a red dashed line and the post-change population is represented by a blue solid line in Fig. 9.11. In the top-left panel, the effect size is zero which corresponds to 100% overlap between the pre- and post-change populations. As the effect size increases, the amount of overlap decreases. The most extreme case presented in Fig. 9.11 is an effect size of three. In this situation there is only 13% overlap between the pre- and post-change populations. There is an important consideration when using these plots to select an acceptable value for ES. When the mean shift is equal to the ES in each panel of Fig. 9.11, there is only a 5% chance that one will pass the statistical test of equivalence. Thus, in order to reasonably pass a test of equivalence, the true mean difference must be much less than the EAC.

Table 9.6 presents a summary table of eight pre-change process lots that are to be used to define a statistically based EAC.

The selected EAC using Eq. (9.8) is

$$\mathrm{EAC} = 2 \times 0.367 = 0.73 \tag{9.11}$$

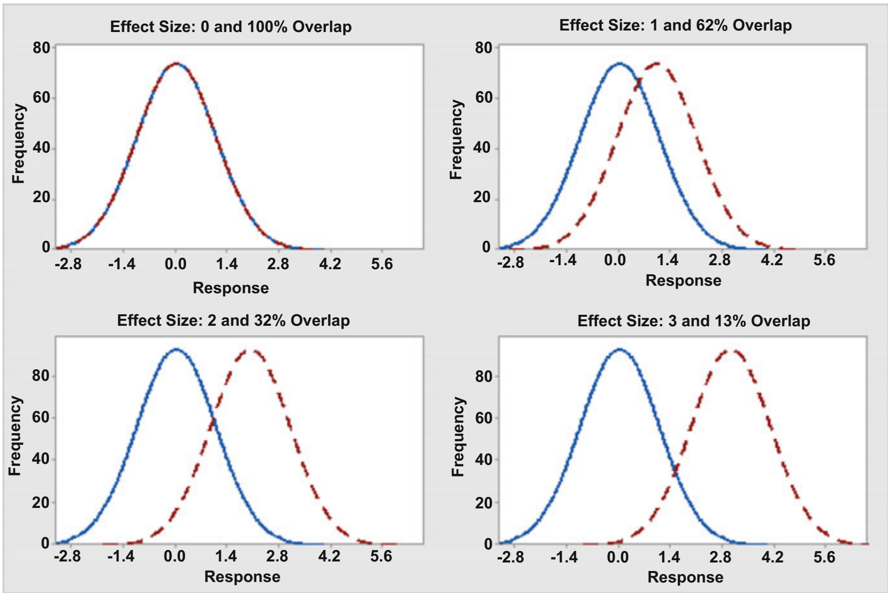


Fig. 9.11 Plots of effect sizes (ES) with overlapping pre-change (*dashed line*) and post-change (*solid line*) populations

Table 9.6 Values required to compute EAC

Description	Value
Number of pre-change lots	$n_{\mathrm{Pre}} = 8$
Pre-change sample standard deviation	0.367
Acceptable effect size	2

Using Eq. (2.10) to incorporate an 80% upper bound on the variance, the EAC is

$$\begin{aligned} \mathrm{EAC} &= \mathrm{ES}\sqrt{\frac{(n_{\mathrm{Pre}}-1)S_{Pre}^2}{\chi^2_{\alpha:n_{\mathrm{Pre}}-1}}} \\ &= 2 \times \sqrt{\frac{(8-1)\times(0.367)^2}{3.82}} = 0.99 \end{aligned} \tag{9.12}$$

At this point, the SME can help evaluate the reasonableness of the EAC. If there are repeated measures for each lot, one should consider working with lot averages as discussed in Sect. 9.3.2.

Another option when basing the EAC on effect size is to perform the equivalence test directly on the effect size. That is, change the hypotheses in Eq. (9.3) to

$$\begin{aligned} H_0 &: \frac{|\mu_{Pre} - \mu_{Post}|}{\sigma_{\mathrm{Pre}}} \geq \mathrm{EAC} \\ H_a &: \frac{|\mu_{Pre} - \mu_{Post}|}{\sigma_{\mathrm{Pre}}} < \mathrm{EAC} \end{aligned} \tag{9.13}$$

A confidence interval can be computed to test the effect size in Eq. (9.13) using results presented in Sect. 2.8.2. An example of such an application is provided in Sect. 9.7.

9.4.2 Equivalence Acceptance Criterion for Profile Data

The previous section considered the computation of an EAC using non-profile data. However, ICH Q5E also requires that the stability profiles of the pre- and post-change products be highly similar. Burdick and Sidor (2013) provide an approach for defining an EAC with profile data under stressed conditions.

For stability data, the stability profile of the pre-change product is compared to the post-change product in order to determine if the degradation rates are highly similar. The hypothesis test is focused on the difference between the pre- and post-change degradation rates (slopes). The hypotheses are

$$\begin{aligned} H_0 &: |\beta_{Pre} - \beta_{Post}| \geq EAC \\ H_a &: |\beta_{Pre} - \beta_{Post}| < EAC \end{aligned} \tag{9.14}$$

where β_{Pre} is the slope of the pre-change profile and β_{Post} is the slope of the post-change profile. As in the non-profile case, the EAC is a pre-selected constant that reflects the maximum allowable difference between two parameters such that they can be deemed equivalent. The challenge with profile data is that degradation at recommended storage conditions may be slow and differences are manifested over a very long time period. Typically, product is exposed to non-recommended storage conditions such as a higher temperature to accelerate degradation. The exposure of product to specific accelerated conditions allows the stability profiles to be compared in a more timely manner. When comparing degradation rates under either recommended storage conditions or accelerated conditions, it is assumed that reaction kinetics driving the stability properties are consistent between the pre-and post-change processes. With this assumption, the slopes can be compared using a statistical test of equivalence.

Under recommended storage conditions, the EAC may be directly linked to product safety and efficacy through the use of the product's specification. However, under non-recommended storage conditions, the linkage to specifications is not meaningful. For small molecules, it might be possible to establish EAC using Arrhenius kinetics to link acceptable degradation rates at accelerated conditions to product specifications at recommended conditions. However, such kinetics are difficult to apply to biological product degradation mechanisms, and thus is not considered a generally useful approach. Instead, with non-recommended storage conditions, the EAC can be expressed as an effect size in much the same manner described for non-profile data.

Assuming that the reaction kinetics driving the stability properties are consistent between the pre- and post-change processes, the random intercept mixed model in Eq. (2.115) is used to define the responses. The assumed model for establishing the preliminary EAC using the pre-change data when all lots are measured at the same time points is

$$\begin{gathered} Y_{ij} = \mu + L_i + \beta_{Pre} \times t_j + E_{ij} \\ i = 1, \ldots, n;\ j = 1, \ldots, T \end{gathered} \tag{9.15}$$

where Y_{ij} is a response measured for lot i at time point j, μ is the average y-intercept across all pre-change lots, β_{Pre} is the average slope across all pre-change lots, L_i is a random variable that allows the y-intercept to vary from μ for a given lot, L_i has a normal distribution with mean 0 and variance σ_L^2, t_j is the time point for measurement j of each lot, E_{ij} is a random normal error term created by measurement error and model misspecification with mean 0 and variance σ_E^2, n is the number of sampled lots, T is the number of time points obtained for each pre-change lot, and L_i and E_{ij} are jointly independent.

Once the model has been fit, the EAC is computed. The methodology used to compute the EAC for profile data is similar to the concept presented in Sect. 9.4.1. For the accelerated stability model in (9.15), consider the ordinary least squares estimator of the slope based on the ith lot, $\hat{\beta}_i$. The statistical test of equivalence is now based on comparing the distribution of the $\hat{\beta}_i$ for the pre- and post-change processes. For the pre-change process, the distribution of $\hat{\beta}_i$ is normal with mean β and variance

$$\begin{gathered} Var\left(\hat{\beta}\right) = \frac{\sigma_E^2}{\text{SST}} \\ \text{SST} = \sum_{j=1}^{T}\left(t_j - \bar{t}\right)^2 \\ \bar{t} = \frac{\sum_{t=1}^{T} t_j}{T}. \end{gathered} \tag{9.16}$$

Treating the pre-change process as the reference distribution, the EAC is

$$\text{EAC} = \text{ES} \times \sqrt{\frac{\sigma_E^2}{\text{SST}}}. \tag{9.17}$$

where an estimate σ_E^2 is based on the pre-change sample. Typically, most of the variability represented by σ_E^2 is due to the analytical method error. If the analytical method is well characterized, the intermediate precision may be used to estimate σ_E^2. Should the intermediate precision not be available, σ_E^2 may be obtained from the pre-change data collected at the storage condition of interest. In some cases, it may be reasonable to use a $100(1 - \alpha)\%$ upper bound on σ_E^2.

The next step in computing (9.17) is to determine an appropriate effect size, ES. Similar to the non-profile case, it is helpful to evaluate the effect size visually. Figure 9.12 displays plots of two processes for four values of ES. The figure presents 15 randomly generated individual slope estimates from each process. The pre-change slope estimates are represented by the solid lines and the post-change process estimates by the dashed lines. All lines are emanating from the same y-intercept in order to better focus on the differences in slopes. One can see from Fig. 9.12 that an ES of three provides essentially two distinct distributions. This suggests that an ES more extreme than three might be too great a separation to declare populations comparable. Overlap of the distributions can be defined as with non-profile data, suggesting a value of ES = 2 is reasonable. Recall that for a given EAC, when the true difference in slopes is equal to the EAC, there is only a 5% chance of passing the equivalence test.

To demonstrate, consider a pre-change data set collected for a purity assay over a 3 month time period. Samples are held at the stressed condition of 37°C for the entirety of the study. There are $n = 15$ lots in the pre-change data set and all lots have been evaluated at 0, 1, 2, and 3 months. Figure 9.13 consists of the individual predicted slopes fit through each lot where all regression lines are emanating from the average y-intercept of 86.0% to better visualize the range of slopes for the

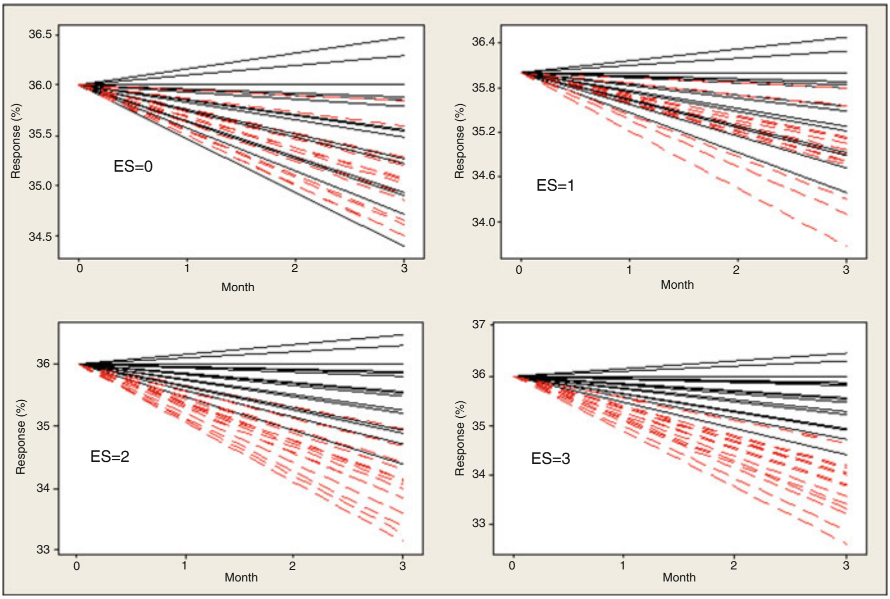


Fig. 9.12 Plots of effect sizes (ES) with stability profiles (*solid lines* are the pre-change process and *dashed lines* the post-change process)

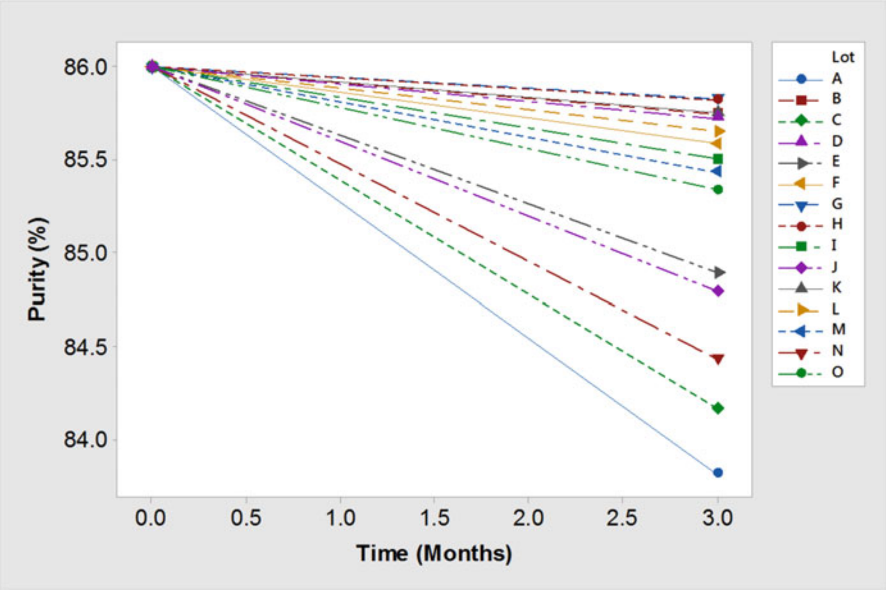


Fig. 9.13 Pre-change data with lot specific regression lines and common intercept

Table 9.7 Slope estimates of the pre-change lots

Lot ID	Slope	Lot ID	Slope
A	−0.727	I	−0.165
B	−0.090	J	-0.402
C	−0.610	K	−0.082
D	−0.092	L	−0.115
E	−0.368	M	−0.188
F	−0.137	N	−0.521
G	−0.056	O	−0.220
H	−0.059	Average	−0.255

pre-change data. The slopes range from −0.056%/month (lot G) to −0.727%/month (lot A). Table 9.7 lists the individual slope estimates for each lot.

The average slope is $\hat{\beta}_{Pre} = -0.255$. The value for the mean squared error is obtained by regressing Purity on Time and Lot (as a random effect), and using the estimate of σ_E^2. For this example, the mean squared error is $\hat{\sigma}_E^2 = 0.200$. The associated degrees of freedom are $n_{\text{Pre}} \times (T-1) - 1 = 15 \times (4-1) - 1 = 44$. Formula (9.17) is used to compute the EAC based on the pre-change data. Here

$$\bar{t} = \frac{\sum_{j=1}^{T} t_j}{T} = \frac{(0+1+2+3)}{4} = 1.5$$
$$SST = \sum_{j=1}^{T} \left(t_j - \bar{t}\right)^2 = (0-1.5)^2 + (1-1.5)^2 + (2-1.5)^2 + (3-1.5)^2 = 5 \tag{9.18}$$

and so

$$\begin{aligned} \text{EAC} &= \text{ES}\sqrt{\frac{\hat{\sigma}_E^2}{SST}} \\ &= 2 \times \sqrt{\frac{0.200}{5}} = 0.40\% \text{ per month} \end{aligned} \tag{9.19}$$

The test of the hypotheses

$$\begin{aligned} H_0 &: |\beta_{Pre} - \beta_{Post}| \geq 0.40\% \text{ per month} \\ H_a &: |\beta_{Pre} - \beta_{Post}| < 0.40\% \text{ per month} \end{aligned} \tag{9.20}$$

was performed by selecting six post-change lots, and subjecting them to the stressed condition of 37°C at 0, 1, 2, and 3 months. The slopes and average for these six lots are shown in Table 9.8. The mean squared error is 0.183 with $n_{Post} \times (T-1) - 1 = 17$ degrees of freedom.

Table 9.8 Slope estimates of the post-change lots

Lot ID	Slope
P	−0.547
Q	−0.576
R	−0.613
S	−0.460
T	−0.362
U	−0.196
Average	−0.459

Table 9.9 Summary of slopes and mean squared errors

Group	Slope average	Mean squared error	Error degrees of freedom
Pre-change	−0.255	0.200	44
Post-change	−0.459	0.183	17

By pooling the two mean squared errors in Table 9.9, the estimate of σ_E^2 is

$$\hat{\sigma}_E^2 = \frac{44 \times 0.200 + 17 \times 0.183}{61} = 0.195. \tag{9.21}$$

Because the same time points are used for each data set, $SST = 5$ for both groups, and a 90% two-sided confidence interval on the difference in slopes is

$$\begin{gathered}\hat{\beta}_{\text{Pre}} - \hat{\beta}_{Post} \pm t_{0.95:61}\sqrt{\frac{\hat{\sigma}_E^2}{SST}\left(\frac{1}{n_{\text{Pre}}} + \frac{1}{n_{Post}}\right)} \\ L = -0.255 - (-0.459) - (1.67)\sqrt{\frac{0.195}{5}\left(\frac{1}{15} + \frac{1}{6}\right)} = 0.045\% \text{ per month} \\ U = -0.255 - (-0.459) + (1.67)\sqrt{\frac{0.195}{5}\left(\frac{1}{15} + \frac{1}{6}\right)} = 0.363\% \text{ per month}\end{gathered} \tag{9.22}$$

As shown in Fig. 9.14, the confidence interval from 0.045 to 0.363% falls between the EAC range of −0.40 to +0.40%/month. This demonstrates equivalence of slopes.

9.5 Design and Power Considerations

Before performing any equivalence test, it is important to plan a design that has a good chance of passing when the groups are indeed equivalent. First, the collected samples for the study should be run in a random order to minimize the impact of bias. The randomization of the samples should be discussed with the laboratory

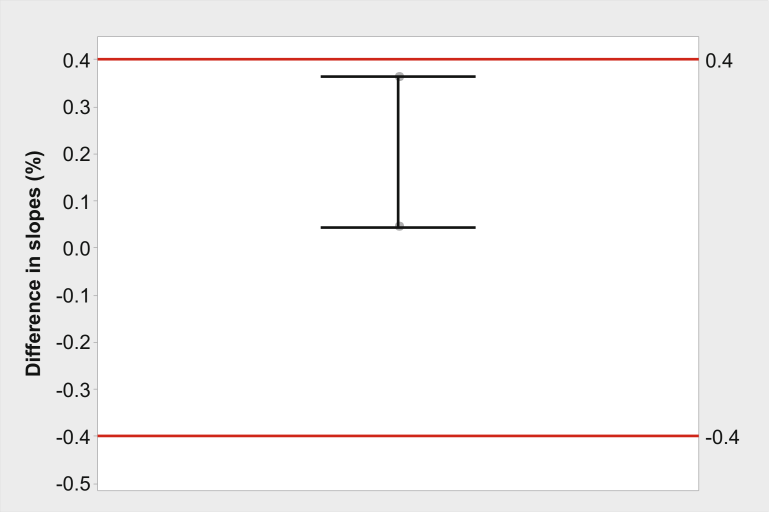


Fig. 9.14 Results of equivalence test on slopes

prior to sample collection so that any logistics with sample freezing and run order can be agreed upon. Second, the study owner should understand the format of the data set. This includes significant figures, labeling format, and administration of the data set. Data precision should align with the maximum precision allowed by the analytical method. If data are overly rounded, the parameter estimates will be over- or underestimated depending on the direction of the rounding. Section 2.3 provides more discussion on this topic.

Power is the probability of rejecting the null hypothesis for a given value of the parameter of interest. It is important to properly power a statistical test in order to ensure that equivalence can be demonstrated when it is present. Recommendations for determining the number of post-change lots in an equivalence study are discussed in the next two sections.

9.5.1 Non-profile Data

Recall the equivalence test used to demonstrate comparability with non-profile data is

$$\begin{aligned} H_0 &: |\mu_{Pre} - \mu_{Post}| \geq \text{EAC} \\ H_a &: |\mu_{Pre} - \mu_{Post}| < \text{EAC} \end{aligned} \tag{9.23}$$

As discussed in Sect. 2.11, equivalence is assessed by constructing a two-sided $100(1-2\alpha)\%$ confidence interval on the difference $\mu_{Pre} - \mu_{Post}$. The null

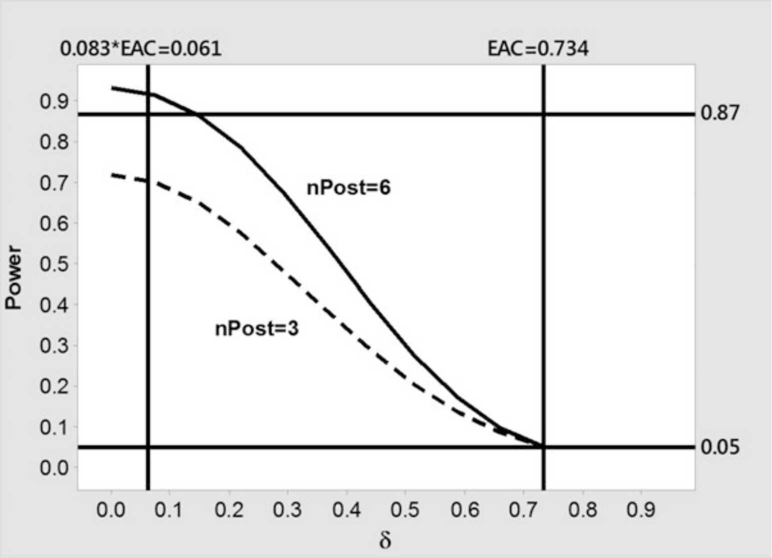


Fig. 9.15 Power curve with $\delta = |\mu_{Pre} - \mu_{Post}|$, $n_{\text{Pre}} = 8$, $\sigma = 0.367$, and EAC = 0.734

hypothesis H_0 in Eq. (9.23) is rejected and equivalence is demonstrated if the entire confidence interval falls in the range from −EAC to + EAC.

Power is defined as the probability of rejecting H_0 and claiming equivalence for a given value of $|\mu_{Pre} - \mu_{Post}|$. By definition, if the value $|\mu_{Pre} - \mu_{Post}| =$ EAC, then the power is equal to α. Typically, $\alpha = 0.05$ and one constructs a $100(1 - 2\alpha)\% = 90\%$ confidence interval on the difference $\mu_{Pre} - \mu_{Post}$. For values of $|\mu_{Pre} - \mu_{Post}|$ less than EAC, the power is greater than 0.05, and for values greater than EAC, it is less than 0.05. Figure 9.15 presents a power curve for the EAC = 0.734 with a standard deviation of 0.367.

As expected, increasing the number of post-change lots from 3 to 6 increases the power for any given value of $\delta = |\mu_{Pre} - \mu_{Post}|$. In order to determine an appropriate number of post-change lots, we recommend a power somewhere between at least 0.74–0.87 when $|\mu_{Pre} - \mu_{Post}| = 0.083 \times$ EAC. This assumes one is using a two-sided 90% confidence interval to conduct the test. Applying this rule to our example, a sample size of three post-change lots is minimally sufficient and six post-change lots provides more than adequate power.

The power curve in Fig. 9.15 was computed using the SAS program PROC POWER. This code is shown below for the case where $n_{Post} = 6$ post-change lots are tested against $n_{\text{Pre}} = 8$ pre-change lots assuming $|\mu_{Pre} - \mu_{Post}| = 0.083 \times \text{EAC} = 0.061$. The computed power from this code is 0.923.

```
proc power;
twosamplemeans test=equiv_diff
lower=−0.734
```

```
upper=0.734
meandiff=0.061
stddev=0.367
power=.
groupns=(8 6);
run;
```

If software is not available to perform this calculation, one can write a simple simulation code to compute the power. Consider the same example shown in Fig. 9.15. A simulation can be constructed using Excel by following these steps and using the pooled confidence interval shown in Eq. (2.56):

1. Select values for $\delta = |\mu_{Pre} - \mu_{Post}|$, σ, $n_{\mathrm{Pr}\,e}$, and n_{Post}. For our example select $\delta = 0.061$, $\sigma = 0.367$, $n_{\mathrm{Pre}} = 8$, and $n_{Post} = 6$.
2. Simulate a random value for $\bar{Y}_{\mathrm{Pre}} - \bar{Y}_{Post}$ using the formula

$$\begin{aligned}\bar{Y}_{\mathrm{Pre}} - \bar{Y}_{Post} &= \mu_{\mathrm{Pre}} - \mu_{Post} + Z \times \sqrt{\sigma^2\left(\frac{1}{n_{\mathrm{Pre}}} + \frac{1}{n_{Post}}\right)} \\ &= 0.061 + Z \times \sqrt{(0.367)^2\left(\frac{1}{8} + \frac{1}{6}\right)}\end{aligned} \tag{9.24}$$

where Z is a randomly simulated standard normal random variable.

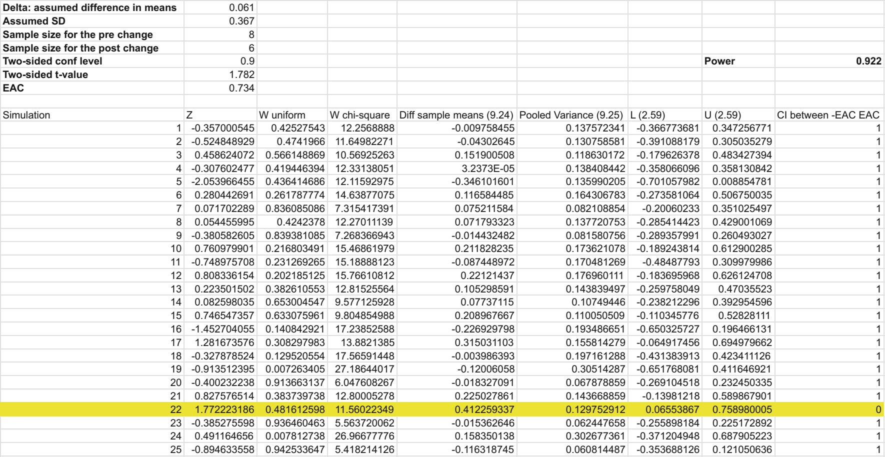
3. Simulate a random value for S_P^2 using the formula

$$S_P^2 = \frac{\sigma^2}{n_{\mathrm{Pre}} + n_{Post} - 2} \times W = \frac{(0.367)^2}{8 + 6 - 2} \times W \tag{9.25}$$

where W is a chi-squared random variable with $n_{\mathrm{Pre}} + n_{Post} - 2$ degrees of freedom.

4. Compute L and U using Eq. (2.56) to form a 90% confidence interval on $\mu_{Pre} - \mu_{Post}$.
5. If the confidence interval in step 4 falls between –EAC and +EAC, increase a counter by one, and simulate another iteration of steps 1–5. Repeat 10,000 times.

Figure 9.16 shows the first 25 rows of an Excel spreadsheet with 10,000 iterations of the simulation. (The entire spreadsheet is available at the website for this book.) Note that in Excel, W needs to be determined by first using the random uniform function since a chi-squared generator is not available. The percentage of the simulated 10,000 values that falls within the range –EAC to +EAC is 0.922. This matches to two decimal places the value computed using PROC POWER.



Delta: assumed difference in means	0.061							
Assumed SD	0.367							
Sample size for the pre change	8							
Sample size for the post change	6							
Two-sided conf level	0.9						**Power**	**0.922**
Two-sided t-value	1.782							
EAC	0.734							
Simulation	Z	W uniform	W chi-square	Diff sample means (9.24)	Pooled Variance (9.25)	L (2.59)	U (2.59)	CI between -EAC EAC
1	-0.357000545	0.42527543	12.2568888	-0.009758455	0.137572341	-0.366773681	0.347256771	1
2	-0.524848929	0.4741966	11.64982271	-0.04302645	0.130758581	-0.391088179	0.305035279	1
3	0.458624072	0.566148869	10.56925263	0.151900508	0.118630172	-0.179626378	0.483427394	1
4	-0.307602477	0.419446394	12.33138051	3.2373E-05	0.138408442	-0.358066096	0.358130842	1
5	-2.053966455	0.436414686	12.11592975	-0.346101601	0.135990205	-0.701057982	0.008854781	1
6	0.280442691	0.261787774	14.63877075	0.116584485	0.164306783	-0.273581064	0.506750035	1
7	0.071702289	0.836085086	7.315417391	0.075211584	0.082108854	-0.20060233	0.351025497	1
8	0.054455995	0.4242378	12.27011139	0.071793323	0.137720753	-0.285414423	0.429001069	1
9	-0.380582605	0.839381085	7.268366943	-0.014432482	0.081580756	-0.289357991	0.260493027	1
10	0.760979901	0.216803491	15.46861979	0.211828235	0.173621078	-0.189243814	0.612900285	1
11	-0.748975708	0.231269265	15.18888123	-0.087448972	0.170481269	-0.48487793	0.309979986	1
12	0.808336154	0.202185125	15.76610812	0.22121437	0.176960111	-0.183695968	0.626124708	1
13	0.223501502	0.382610553	12.81525564	0.105298591	0.143839497	-0.259758049	0.47035523	1
14	0.082598035	0.653004547	9.577125928	0.07737115	0.10749446	-0.238212296	0.392954596	1
15	0.746547357	0.633075961	9.804854988	0.208967667	0.110050509	-0.110345776	0.52828111	1
16	-1.452704055	0.140842921	17.23852588	-0.226929798	0.193486651	-0.650325727	0.196466131	1
17	1.281673576	0.308297983	13.8821385	0.315031103	0.155814279	-0.064917456	0.694979662	1
18	-0.327878524	0.129520554	17.56591448	-0.003986393	0.197161288	-0.431383913	0.423411126	1
19	-0.913512395	0.007263405	27.18644017	-0.12006058	0.30514287	-0.651768081	0.411646921	1
20	-0.400232238	0.913663137	6.047608267	-0.018327091	0.067878859	-0.269104518	0.232450335	1
21	0.827576514	0.383739738	12.80005278	0.225027861	0.143668859	-0.13981218	0.589867901	1
22	1.772223186	0.481612598	11.56022349	0.412259337	0.129752912	0.06553867	0.758980005	0
23	-0.385275598	0.936460463	5.563720062	-0.015362646	0.062447658	-0.255898184	0.225172892	1
24	0.491164656	0.007812738	26.96677776	0.158350138	0.302677361	-0.371204948	0.687905223	1
25	-0.894633558	0.942533647	5.418214126	-0.116318745	0.060814487	-0.353688126	0.121050636	1

Fig. 9.16 Simulated power spreadsheet

9.5.2 Power Considerations with Profile Data

The equivalence test used to demonstrate comparability with profile data as described in (9.14) is

$$\begin{aligned} H_0 &: |\beta_{Pre} - \beta_{Post}| \geq \mathrm{EAC} \\ H_a &: |\beta_{Pre} - \beta_{Post}| < \mathrm{EAC} \end{aligned} \tag{9.26}$$

Equivalence is demonstrated if a two-sided $100(1-2\alpha)\%$ confidence interval on $\beta_{Pre} - \beta_{Post}$ falls in the range from $-\mathrm{EAC}$ to $+\mathrm{EAC}$. A simulation to determine power can be constructed using Excel by following these steps and using the pooled confidence interval shown in Eq. (9.22):

1. Select values for EAC, $|\beta_{Pre} - \beta_{Post}|$, σ_E^2, T, SST, $n_{\mathrm{Pr}\,e}$, and n_{Post}. For our example assume we select $\mathrm{EAC} = 0.40\%$, $|\beta_{Pre} - \beta_{Post}| = 0.083 \times \mathrm{EAC} = 0.033$, $\sigma_E^2 = 0.20$, $T = 4$, $SST = 5$, $n_{\mathrm{Pre}} = 15$, and $n_{Post} = 6$. Assuming we will be pooling the data to estimate error, the error df is $(n_{\mathrm{Pre}} + n_{Post}) \times (T-1) - 2 = 61$.
2. Simulate a random value for $\hat{\beta}_{\mathrm{Pre}} - \hat{\beta}_{Post}$ using the formula

$$\begin{aligned} \hat{\beta}_{\mathrm{Pre}} - \hat{\beta}_{Post} &= \beta_{\mathrm{Pre}} - \beta_{Post} + Z \times \sqrt{\frac{\sigma_E^2}{SST}\left(\frac{1}{n_{\mathrm{Pre}}} + \frac{1}{n_{Post}}\right)} \\ &= 0.033 + Z \times \sqrt{\frac{0.20}{5}\left(\frac{1}{15} + \frac{1}{6}\right)} \end{aligned} \tag{9.27}$$

where Z is a randomly simulated standard normal random variable.

3. Simulate a random value for $\hat{\sigma}_E^2$ using the formula

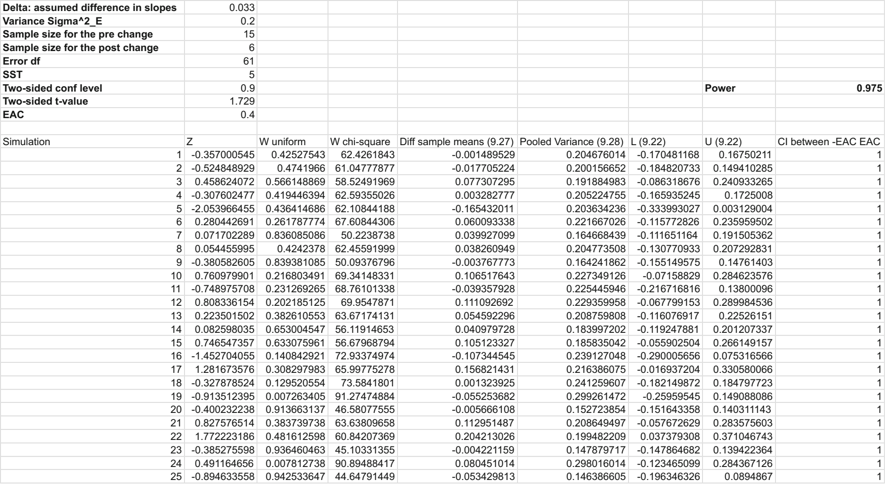
$$\begin{aligned} \hat{\sigma}_E^2 &= \frac{\sigma_E^2}{\text{Error df}} \times W \\ &= \frac{0.20}{61} \times W \end{aligned} \tag{9.28}$$

where W is a chi-squared random variable with 61 error degrees of freedom for this example.

4. Compute L and U using Eq. (9.22) to form a 90% confidence interval on $\beta_{\text{Pre}} - \beta_{Post}$.
5. If the confidence interval in step 4 falls between –EAC and +EAC, increase a counter by one, and simulate another iteration of steps 1–5. Repeat 10,000 times.

Figure 9.17 shows the first 25 rows of an Excel spreadsheet with 10,000 iterations of the simulation. The percentage of the simulated 10,000 values that falls within the range –EAC to +EAC is 0.975, which exceeds the target of 0.87.

It is important to note the impact of *SST* on power. In the previous example, time points were at 0, 1, 2, and 3 months. If the timeframe is increased, the power will increase for the same number of lots. For example, suppose we select the three time points 0, 3, and 6 months. Our range is now 6 months instead of only 3 months. Using Eq. (9.19) we have SST = 18 and given the same EAC, the power is increased to almost 1.0.



Delta: assumed difference in slopes	0.033							
Variance Sigma^2_E	0.2							
Sample size for the pre change	15							
Sample size for the post change	6							
Error df	61							
SST	5							
Two-sided conf level	0.9						**Power**	**0.975**
Two-sided t-value	1.729							
EAC	0.4							
Simulation	Z	W uniform	W chi-square	Diff sample means (9.27)	Pooled Variance (9.28)	L (9.22)	U (9.22)	CI between -EAC EAC
1	-0.357000545	0.42527543	62.4261843	-0.001489529	0.204676014	-0.170481168	0.16750211	1
2	-0.524848929	0.4741966	61.04777877	-0.017705224	0.200156652	-0.184820733	0.149410285	1
3	0.458624072	0.566148869	58.52491969	0.077307295	0.191884983	-0.086318676	0.240933265	1
4	-0.307602477	0.419446394	62.59355026	0.003282777	0.205224755	-0.165935245	0.1725008	1
5	-2.053966455	0.436414686	62.10844188	-0.165432011	0.203634236	-0.333993027	0.003129004	1
6	0.280442691	0.261787774	67.60844306	0.060093338	0.221667026	-0.115772826	0.235959502	1
7	0.071702289	0.836085086	50.2238738	0.039927099	0.164668439	-0.111651164	0.191505362	1
8	0.054455995	0.4242378	62.45591999	0.038260949	0.204773508	-0.130770933	0.207292831	1
9	-0.380582605	0.839381085	50.09376796	-0.003767773	0.164241862	-0.155149575	0.14761403	1
10	0.760979901	0.216803491	69.34148331	0.106517643	0.227349126	-0.07158829	0.284623576	1
11	-0.748975708	0.231269265	68.76101338	-0.039357928	0.225445946	-0.216716816	0.13800096	1
12	0.808336154	0.202185125	69.9547871	0.111092692	0.229359958	-0.067799153	0.289984536	1
13	0.223501502	0.382610553	63.67174131	0.054592296	0.208759808	-0.116076917	0.22526151	1
14	0.082598035	0.653004547	56.11914653	0.040979728	0.183997202	-0.119247881	0.201207337	1
15	0.746547357	0.633075961	56.67968794	0.105123327	0.185835042	-0.055902504	0.266149157	1
16	-1.452704055	0.140842921	72.93374974	-0.107344545	0.239127048	-0.290005656	0.075316566	1
17	1.281673576	0.308297983	65.99775278	0.156821431	0.216386075	-0.016937204	0.330580066	1
18	-0.327878524	0.129520554	73.5841801	0.001323925	0.241259607	-0.182149872	0.184797723	1
19	-0.913512395	0.007263405	91.27474884	-0.055253682	0.299261472	-0.25959545	0.149088086	1
20	-0.400232238	0.913663137	46.58077555	-0.005666108	0.152723854	-0.151643358	0.140311143	1
21	0.827576514	0.383739738	63.63809658	0.112951487	0.208649497	-0.057672629	0.283575603	1
22	1.772223186	0.481612598	60.84207369	0.204213026	0.199482209	0.037379308	0.371046743	1
23	-0.385275598	0.936460463	45.10331355	-0.004221159	0.147879717	-0.147864682	0.139422364	1
24	0.491164656	0.007812738	90.89488417	0.080451014	0.298016014	-0.123465099	0.284367126	1
25	-0.894633558	0.942533647	44.64791449	-0.053429813	0.146386605	-0.196346326	0.0894867	1

Fig. 9.17 Simulated power spreadsheet for profile data

9.6 Reporting Analytical Comparability Results

Once the EAC has been defined, it is time to collect the post-change data using the study design outlined in the analytical comparability protocol. Regardless of the approach, plots of the raw data and descriptive statistics should be part of the analysis. For non-profile data, plot the pre-change data and post-change data in time order along with descriptive statistics. For profile data, plot the raw data and the average slope for the pre- and post-change data and report the appropriate descriptive statistics. Additional plots and results are required if an equivalence test is performed. We now provide some examples.

9.6.1 Reports for Individual Post-change Values

When the acceptance criterion is based on pre-change data, the subsequent analysis consists of evaluating each post-change value relative to the acceptance criterion and the specification. It is recommended to plot the post-change and pre-change data by time-ordered batch ID (trend plot). This plot provides a visual assessment of any shift in the post-change mean along with changes in variability. Inclusion of reference lines for a specification or the acceptance criterion is a matter of personal preference. However, in cases where the data are far away from the specification, the excessive white space between the specification and the actual data may not be of value. In addition, the raw data are difficult to see because they are isolated to a narrow range of the y-axis. A rule of thumb for graphing data is to retain approximately one-third of the graph for white space.

In addition to the graphical presentation, descriptive statistics should be presented in the analysis. When the sample size of the post-change data set is at least four lots, the descriptive statistics (mean, standard deviation, minimum, and maximum) should be presented for both the pre- and post-change data sets. If there are fewer than four lots in the post-change data set, providing the minimum and maximum values is adequate. Table 9.10 reports descriptive statistics, types of

Table 9.10 Guidance on presentations for reporting comparability results for individual post-change values

		Sample size	
	Item	1–3 lots	≥4 lots
Descriptive statistics	Mean		✓
	Standard deviation		✓
	Variance		Optional
	Range (minimum and maximum value)	✓	✓
	Confidence interval on the mean		Optional
Plots	Trend plot	✓	✓
	Boxplot		Optional
	Individual value plot		Optional

plots, and guidance on sample size for presentation. Items with a check mark are strongly recommended. Optional plots and descriptive statistics are listed as such. If there is a blank, the use of that plot or statistic is not recommended.

9.6.2 Reports for Equivalence Testing with Non-profile Data

Once the data for the post-change process have been collected, the equivalence test is performed. This test is performed by computing a two-sided 90% confidence interval on the difference in means. Equivalence is demonstrated if both bounds fall within the range from –EAC to +EAC. The confidence intervals are computed using one of the intervals shown in Table 9.11.

Table 9.12 summarizes useful plots and descriptive statistics for reporting an equivalence test. Check marks denote recommendations and optional items are defined as such.

To demonstrate, consider a non-profile analysis of lot release data for protein concentration measured in mg/mL. There are $n_{\mathrm{Pre}} = 35$ lots of pre-change product and $n_{Post} = 3$ lots of post-change product. The equivalence hypotheses of interest are

Table 9.11 Confidence intervals used with equivalence tests of mean

Data structure	Compute the two-sided confidence interval with equation
Independent measurements with equal variances	(2.56)
Independent measurements with unequal variances	(2.58)
Dependent measurements	(2.71)

Table 9.12 Guidance on presentations for reporting comparability results for equivalence tests with non-profile data

	Item	Guidance
Descriptive statistics	Mean	✓
	Standard deviation	✓
	Variance	Optional
	Range (minimum and maximum value)	Optional
	Confidence interval on the mean	Optional
	Confidence interval on the difference in means	✓
Plots	Trend plot	✓
	Boxplot	Optional
	Individual value plot	✓
	Equivalence plot	✓

$$\begin{aligned} H_0 &: |\mu_{\text{Pre}} - \mu_{Post}| \geq 2.0\ \text{mg/mL} \\ H_a &: |\mu_{\text{Pre}} - \mu_{Post}| < 2.0\ \text{mg/mL}. \end{aligned} \tag{9.29}$$

Table 9.13 shows the statistics recommended in Table 9.12.

Figure 9.18 presents the trend plot recommended in Table 9.12. Figure 9.19 presents the individual value plot, and Fig. 9.20 the equivalence plot.

It is clear that there are no unexpected trends in the post-change data relative to the pre-change lots. For this example, the specification reference lines are added to the trend plot.

The plus signs in Fig. 9.19 represent the pre- and post-change process means. This plot is valuable as it gives a visual assessment of the two process means relative to each other along with the spread of the data.

Table 9.13 Recommended descriptive statistics in Table 9.12

Statistic	Protein concentration (mg/mL)
Pre-change mean	65.00
Post-change mean	65.29
Difference in means ($\bar{d}$)	−0.29
Pre-change standard deviation	1.18
Post-change standard deviation	0.48
90% margin of error (ME)	0.73
Lower bound of 90% CI on difference from (2.56)	−1.46
Upper bound of 90% CI on difference from (2.56)	0.88
EAC	2.0
Conclusion	Statistically equivalent

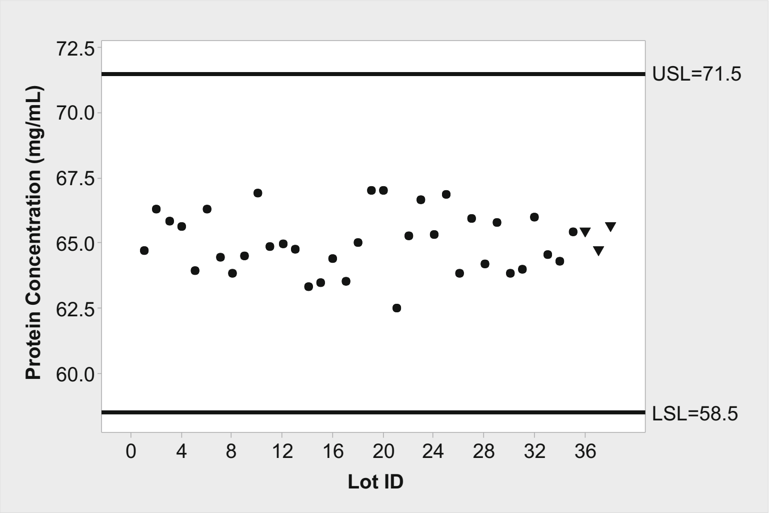


Fig. 9.18 Trend plot with pre-change data (*circle*) and post-change data (*triangle*)

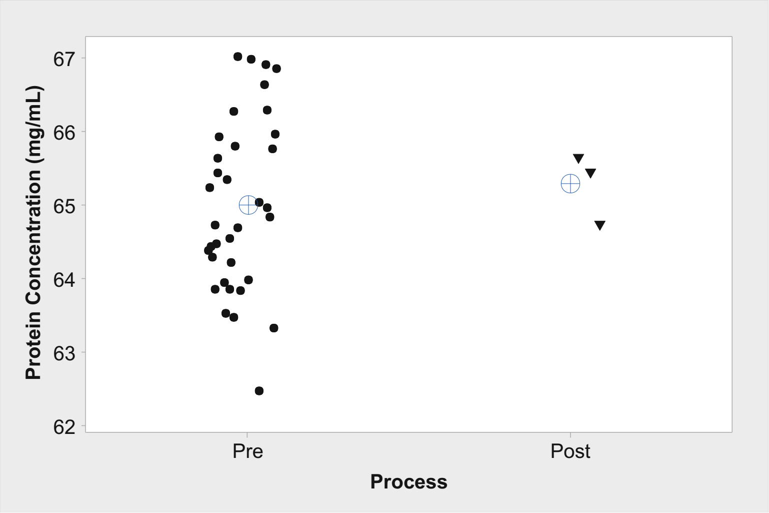


Fig. 9.19 Individual value plot recommended in Table 9.12

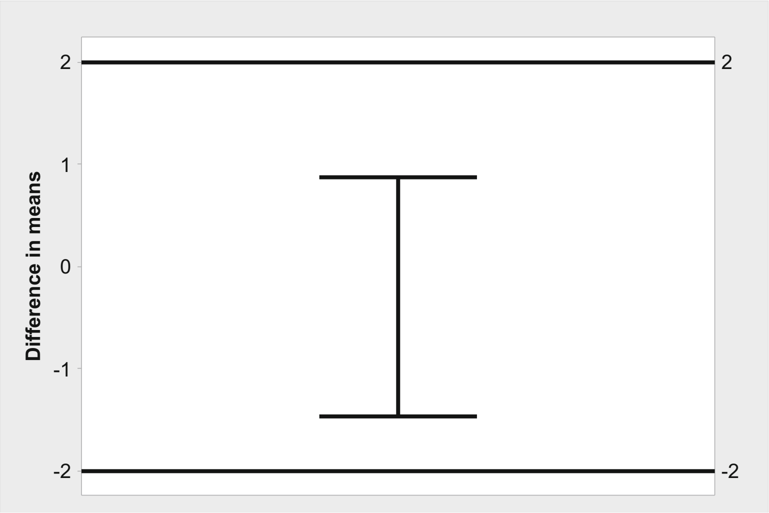


Fig. 9.20 Equivalence plot recommended in Table 9.12

The confidence interval in Fig. 9.20 assumes equal variances and is computed with Eq. (2.56). Since the confidence interval in Fig. 9.20 falls completely inside the EAC of ±2.0, evidence has been provided that the pre- and post-change processes are statistically equivalent.

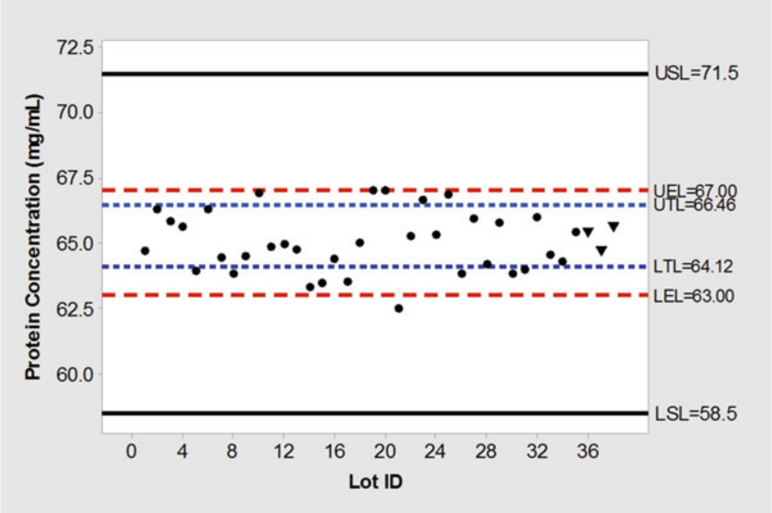


Fig. 9.21 Trend chart with equivalence test

Burdick et al. (2011) developed a trend chart that can be used to visually present the test of equivalence. This chart is shown in Fig. 9.21 for the data described in this section.

The lines in Fig. 9.21 represent algebraic re-expressions of the equivalence inequalities. In the present example, equivalence is demonstrated using Eq. (2.56) if

$$\begin{aligned}&\bar{Y}_{\text{Pre}} - \bar{Y}_{Post} - ME > -EAC \text{ and}\\&\bar{Y}_{\text{Pre}} - \bar{Y}_{Post} + ME < EAC \text{ where}\\&ME = t_{1-\alpha/2:n_{\text{Pre}}+n_{Post}-2}\sqrt{S^2_{Pooled}\left(\frac{1}{n_{\text{Pre}}}+\frac{1}{n_{Post}}\right)}\end{aligned} \tag{9.30}$$

Equation (9.30) can be rewritten as

$$\begin{aligned}&\bar{Y}_{Post} + ME < \bar{Y}_{\text{Pre}} + EAC, \text{ and}\\&\bar{Y}_{Post} - ME > \bar{Y}_{\text{Pre}} - EAC\end{aligned} \tag{9.31}$$

Define the lower test limit $LTL = \bar{Y}_{Post} - ME$, the upper test limit $UTL = \bar{Y}_{Post} + ME$, the lower equivalence limit $LEL = \bar{Y}_{\text{Pre}} - EAC$, and the upper equivalence limit $UEL = \bar{Y}_{\text{Pre}} + EAC$. Average equivalence is demonstrated when $UTL < UEL$ and $LTL > LEL$. Visually this translates into having both the UTL and LTL (short dashed lines) falling within UEL and LEL (long dashed lines).

Using the data in Table 9.13, $LTL = \bar{Y}_{Post} - ME = 65.29 - 1.17 = 64.12$, $UTL = \bar{Y}_{Post} + ME = 65.29 + 1.17 = 66.46$, $LEL = \bar{Y}_{\mathrm{Pre}} - EAC = 65.00 - 2.00 = 63.00$, and $UEL = \bar{Y}_{\mathrm{Pre}} + EAC = 65.00 + 2.00 = 67.00$. Since *LTL* and *UTL* are contained within *UEL* and *LEL*, equivalence is demonstrated. Note that when using this visualization technique, some individual values are likely to fall outside the limits, because the limits are based on means.

9.6.3 Reports for Equivalence Testing with Profile Data

Table 9.14 summarizes the recommended descriptive statistics and plots for profile data.

Recall the example to test the hypotheses in (9.20) presented in Sect. 9.4.2. Table 9.15 provides a tabular summary of the equivalence test. Figures 9.22 and 9.23 present the two recommended plots. Since the lower and upper confidence bounds fall within the range from –EAC to +EAC, the two processes are statistically equivalent.

Table 9.14 Guidance on presentations for reporting comparability results for equivalence tests with profile data

	Item	Guidance
Descriptive statistics	Slope for each process	✓
	Confidence interval on the difference in slopes	✓
	Standard deviation for each process slope	Optional
	Range (minimum and maximum slope for each process)	Optional
	Individual slopes for each lot	Optional
	Confidence interval on the slope	Optional
Plots	Regression plot	✓
	Equivalence plot	✓
	Regression plot with normalized y-intercept for each lot	Optional

Table 9.15 Tabular summary of equivalence test results

Statistic	Value
Pre-change slope	−0.255%/month
Post-change slope	−0.459%/month
Difference in slopes	0.204%/month
SST from (9.18)	5
Lower bound of 90% CI on difference from (9.22)	0.045%/month
Upper bound of 90% CI on difference from (9.22)	0.363%/month
EAC	0.40%/month
Conclusion	Statistically equivalent

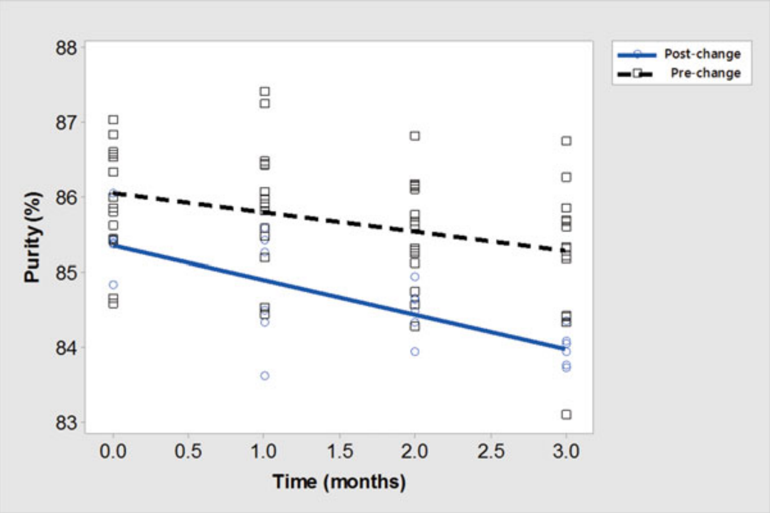


Fig. 9.22 Regression plot

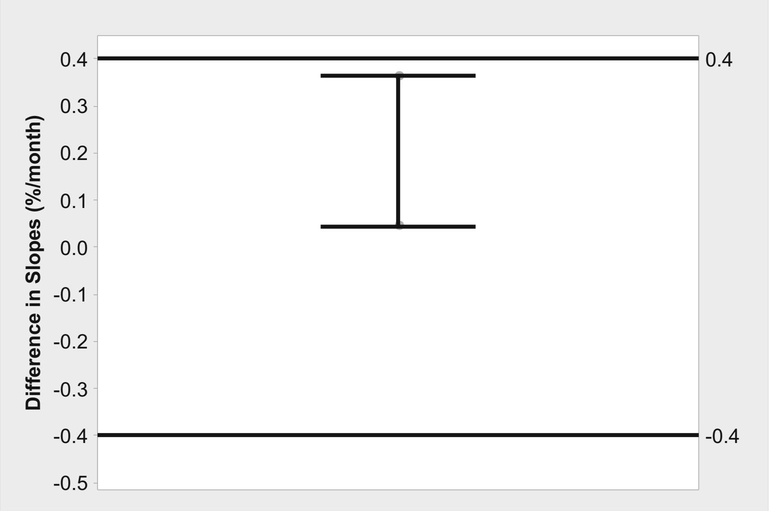


Fig. 9.23 Equivalence plot

9.7 Analytical Similarity for Biosimilar Products

In this section, we provide statistical methods for demonstrating analytical similarity between a proposed biosimilar product (BP) and its associated reference product (RP). The Biologics Price Competition and Innovation Act (BPCIA) of 2009 created an abbreviated licensure pathway for biological products shown to be highly similar to an FDA licensed biological product (also known as the reference product). Biosimilarity means that the biological product is highly similar to the RP notwithstanding minor differences in clinically inactive components and that there are no clinically meaningful differences between the BP and RP in terms of safety, purity, and potency of the product (FDA 2015b). A biosimilar sponsor can rely on existing scientific knowledge about the safety and effectiveness of the RP, and consequently enable a BP to be licensed based on less than a full complement of preclinical and clinical data typically required with a section 351(a) marketing application.

The underlying assumption of this abbreviated pathway is that if a molecule is shown to be analytically and functionally similar to an RP, it will behave like the RP in the clinic. FDA recommends that sponsors use a stepwise approach of data collection and the evaluation of residual uncertainty (FDA 2015b). This approach begins with the assessment of analytical similarity, which includes comparisons of structural and functional attributes between the BP and RP. Animal studies are conducted to address any remaining uncertainties concerning the proposed biosimilar product before initiation of clinical testing of the product in human subjects. The stepwise approach continues with clinical studies including assessment of immunogenicity and pharmacokinetics or pharmacodynamics to establish safety and efficacy equivalency as needed. Approval of biosimilar applications is based on the totality of the evidence and information submitted in the application. FDA guidance on this topic has been published including FDA (2015a, b, c). There is also a planned guidance on statistical methods to demonstrate analytical similarity due in 2016.

Although statistical approaches used to demonstrate similarity are generally the approaches used for comparability, there are some important differences that are now described.

9.7.1 Differences Between Comparability and Similarity

Burdick et al. (2016) have described several features that distinguish demonstration of similarity between an RP and a BP from an assessment of comparability. Key differences include the following:

1. Lack of RP knowledge in a similarity assessment relative to knowledge concerning the pre-change process in a comparability assessment. Lack of RP product knowledge includes such items as

    a. RP process changes which may make pooling of data inappropriate for statistical analysis.
    b. RP process deviations resulting in quality within permitted specifications but outside expected variability. For example, a sampled RP lot may have a measured value that is out of trend with respect to other RP values, even if

the release of the lot was justified based on impact to quality, safety, and efficacy.

   c. Linkage between drug substance (DS) and drug product (DP) lots is not identifiable from sampled RP lots in similarity assessments. If sampled DP lots were manufactured with the same DS, they are correlated, and the assumption of independence required in many statistical calculations is not appropriate.

2. RP target specifications and in-process control (IPC) limits are not known for the majority of the analytical methods in a similarity assessment. This lack of knowledge makes the selection of meaningful acceptance criteria more difficult.
3. The sampling process used to collect the RP lots has an inherent bias that leads to RP lots being generally older than newly manufactured TP lots. This bias is especially problematic for stability indicating methods.

These differences present some important limitations to the statistical methods that have been recommended for demonstration of analytical similarity. We now review a statistical approach suggested by FDA as described in an FDA ODAC briefing document (2015d), Chow (2014, 2015), Dong et al. (2015a), Dong (2015), Shen et al. (2015), Tsong (2015) and Tsong et al. (2015).

9.7.2 Risk Categories for Critical Quality Attributes

Demonstration of analytical similarity begins with the assessment of the relative criticality of quality attributes. Table 9.16 reports the three categories described by the FDA in the previously mentioned references.

Tier 1 attributes require the most statistically rigorous evidence of similarity. This evidence is provided using a statistical test of equivalence. Tier 2 quality attributes require a lesser level of statistical rigor. The recommended approach for Tier 2 attributes are quality ranges. Finally, Tier 3 attributes can be examined using graphical display. Each of these approaches is now described below.

Table 9.16 Risk categories

Risk category	Definition
Tier 1	High impact on activity, PK/PD, safety, or immunogenicity Where practical the attributes measured require a statistical test of equivalence between the proposed biosimilar product and the RP
Tier 2	Moderate impact on activity, PK/PD, safety, or immunogenicity Attributes measured are consistent with a statistical quality range
Tier 3	Low impact on activity, PK/PD, safety, or immunogenicity Descriptive raw data and graphical presentations of similarity

Table 9.17 Summary of adalimumab data

Product	Number of batches	Sample mean (%)	Sample standard deviation (%)	Min (%)	Max (%)
ABP 501(BP)	$n_B = 10$	$\bar{Y}_B = 104$	$S_B = 4.1$	98	110
US-licensed Humira (RP)	$n_R = 21$	$\bar{Y}_R = 105$	$S_R = 5.7$	95	114

9.7.3 Tier 1 Testing

Table 9.17 reports a data set from Table 6 of the FDA (2016) document presented as part of the Arthritis Advisory Committee Meeting for ABP 501, a proposed biosimilar to Humira (adalimumab). The quality attribute comes from an apoptosis inhibition bioassay and is measured in %. This assay measures the primary mechanism of action for the product.

Demonstration of statistical equivalence for a Tier 1 attribute requires testing the following set of hypotheses:

$$\begin{aligned} H_0 &: |\mu_B - \mu_R| \geq 1.5\sigma_R \\ H_1 &: |\mu_B - \mu_R| < 1.5\sigma_R \end{aligned} \tag{9.32}$$

where μ_B is the mean of BP, μ_R is the mean of the RP, and σ_R is the standard deviation for the RP. The value of 1.5 was established by the FDA based on numerous simulation studies and is described in Shen et al. (2015).

Equivalence testing is described in Sect. 2.11, with $\text{EAC} = 1.5\sigma_R$ in this application. A 90% confidence interval on the difference is computed assuming equal variances using the formula in (2.56),

$$\begin{aligned} S_P^2 &= \frac{(n_B-1)S_B^2 + (n_R-1)S_R^2}{n_B+n_R-2} = \frac{(10-1)(4.1)^2 + (21-1)(5.7)^2}{10+21-2} = 27.6 \\ L &= \bar{Y}_B - \bar{Y}_R - t_{1-\alpha/2:n_B+n_R-2}\sqrt{S_P^2\left(\frac{1}{n_B}+\frac{1}{n_R}\right)} \\ L &= 104 - 105 - 1.7\sqrt{27.6\left(\frac{1}{10}+\frac{1}{21}\right)} = -4.1 \\ U &= \bar{Y}_B - \bar{Y}_R + t_{1-\alpha/2:n_B+n_R-2}\sqrt{S_P^2\left(\frac{1}{n_B}+\frac{1}{n_R}\right)} \\ U &= 104 - 105 + 1.7\sqrt{27.6\left(\frac{1}{10}+\frac{1}{21}\right)} = 2.1 \end{aligned} \tag{9.33}$$

(Note the reported interval in the FDA report was computed to a greater decimal precision.) The EAC is determined by replacing σ_R with S_R to yield EAC $= 1.5 \times 5.7 = 8.6$. Since the interval from -4.1 to 2.1 falls entirely within the range from -8.6 to 8.6, equivalence has been demonstrated.

One problem with this approach discussed by Burdick et al. (2016) concerns the fact that σ_R is estimated using S_R to define EAC. The consequence of estimating EAC is that the confidence interval in Eq. (9.33) does not maintain the desired probability of rejecting H_0 when H_0 is true. This increases the risk of passing the test when the BP is not equivalent to the RP. This problem can be resolved by changing the hypotheses in (9.32) to

$$\begin{aligned} H_0 &: \left|\frac{\mu_B - \mu_R}{\sigma_R}\right| \geq 1.5 \\ H_a &: \left|\frac{\mu_B - \mu_R}{\sigma_R}\right| < 1.5 \end{aligned} \tag{9.34}$$

Note that (9.34) is equivalent to (9.32). The only difference is that σ_R has moved to the left-hand side of the equation, and EAC $= 1.5$ is now a known constant. The hypotheses in (9.34) are tested by constructing a 90% confidence interval on the effect size, $\frac{\mu_B -^{B} \mu_R{}^{R}}{\sigma_R}$, as described in Sect. 2.8.2.

To demonstrate, the confidence interval on the effect size for the data in Table 9.17 using the procedure described in Sect. 2.8.2 yields

$$\begin{aligned} t_{calc} &= \frac{\bar{Y}_B - \bar{Y}_R}{\sqrt{S_P^2\left(\frac{1}{n_B} + \frac{1}{n_R}\right)}} \\ &= \frac{104 - 105}{\sqrt{27.6\left(\frac{1}{10} + \frac{1}{21}\right)}} = -0.5 \end{aligned} \tag{9.35}$$

and the resulting confidence interval is from -0.8 to 0.4. Since the entire confidence interval falls in the range from -1.5 to +1.5, equivalence is demonstrated.

Yang et al. (2016) have studied the impact on the recommended Tier 1 test when the RP lots are correlated. They show that when RP lots are correlated, the probability of rejecting H_0 when H_0 is true (i.e., falsely concluding equivalence) will increase, and the probability of passing when the products are equivalent will decrease. As noted in Sect. 9.7.1, linkage between drug substance (DS) and drug product (DP) is often not identifiable from sampled RP lots. If sampled RP lots were manufactured with the same DS, they are correlated, and the Tier 1 equivalence test is impacted. Yang et al. describe approaches to mitigate this problem, but sponsors of biosimilar products are cautioned to avoid correlation by selecting a small number of lots at any given time, and spreading purchase of RP lots over as long a time period as feasible.

9.7.4 Tier 2 Testing

Tier 2 testing uses a quality range approach. The quality range is defined as

$$\mu_R \pm K \times \sigma_R \tag{9.36}$$

where K is appropriately justified. In practice, μ_R is estimated with $\bar{Y}_R$ and σ_R is estimated with S_R. The biosimilar product passes Tier 2 if a predefined proportion (e.g., 90%) of the measured assay responses from the biosimilar lots falls within the quality range. Yang et al. provide results that justify the use of $K = 3$, and demonstrate that correlation among RP lots will cause the quality range to be too tight, because the lot-to-lot variation will not be fully represented.

The quality interval for the data in Table 9.17 is computed as

$$\begin{aligned} &\bar{Y}_R \pm K \times S_R \\ &105 \pm 3 \times 5.7 \\ &L = 87.9\% \\ &U = 122.1\% \end{aligned} \tag{9.37}$$

To pass the Tier 2 test, nine of the ten biosimilar lots (90%) must fall in the range from 87.9 to 122.1%. Since the range of the biosimilar shown in Table 9.17 is from 98 to 110%, all ten lots fall in the quality range, and Tier 2 similarity is demonstrated.

More information on the strategy demonstrated in this example is presented by Velayudhan et al. (2016).

References

Burdick RK, Sidor L (2013) Establishment of an equivalence acceptance criterion for accelerated stability studies. J Biopharm Stat 23:730–743

Burdick RK, Pferdeort V, Sidor L, Tholudur A (2011) A graphical representation for a statistical test of average equivalence and variance comparison with process data. Qual Reliab Eng Int 27(6):771–780

Burdick R, Coffey T, Gutka H, Gratzl G, Conlon H, Huang C-T, Boyne M, Kuehne H (2016) Statistical approaches to assess biosimilarity from analytical data. AAPS J. doi: 10.1208/s12248-016-9968-0

Chatfield MJ, Borman PJ, Damjanov I (2011) Evaluating change during pharmaceutical product development and manufacture—comparability and equivalence. Qual Reliab Eng Int 27:629–640

Chow S-C (2014) On assessment of analytical similarity in biosimilar studies. Drug Des 3:119. doi:10.4172/2169-0138.1000e124, Accessed 28 Nov 2015

Chow S-C (2015) Challenging issues in assessing analytical similarity in biosimilar studies. Biosimilars 5:33–39

Dong X (2015) Equivalence test for biosimilar analytical assessment. Second Statistical and Data Management Approaches for Biotechnology Drug Development, USP Headquarters, Rockville

Dong X, Shen M, Tsong Y (2015a) EP2006 statistical equivalence testing for bioactivity and content. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM431118.pdf

Dong X, Shen M, Tsong Y, Zhong J (2015b) Using tolerance intervals for assessment of pharmaceutical quality. J Biopharm Stat 25:317–327

Food and Drug Administration. Center for Drugs Evaluation Research (1996) Demonstration of comparability of human biological products, including therapeutic biotechnology-derived products, guidance for industry

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 (2015a) Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product, guidance for industry

Food and Drug Administration. Center for Drugs Evaluation Research (2015b) Scientific considerations in demonstrating biosimilarity to a reference product, guidance for industry

Food and Drug Administration. Center for Drugs Evaluation Research (2015c) Questions and answers regarding implementation of the biologics price competition and innovation act of 2009, guidance for industry

Food and Drug Administration. Center for Drugs Evaluation Research (2015d) FDA briefing document for the oncologic drugs advisory committee (ODAC) meeting held on January 7, 2015. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM428781.pdf

Food and Drug Administration. Center for Drugs Evaluation Research (2016) FDA briefing document, arthritis advisory committee meeting, ABP 501, a proposed biosimilar to Humira (adalimumab) by Amgen. http://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/drugs/arthritisadvisorycommittee/ucm510293.pdf

Hauk W, Abernethy D, Williams R (2008) Metrological approaches to setting acceptance criteria: unacceptable and unusual characteristics. J Pharm Biomed Anal 48:1042–1045

International Conference on Harmonization (2004) Q5E Comparability of biotechnological/biological products subject to changes in their manufacturing process

Inman HF, Bradley EL Jr (1989) The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Commun Stat Theory Methods 18(10):3851–3874

Limentani G, Ringo M, Ye F, Bergquist M, McSorley E (2005) Beyond the t-test: statistical equivalence testing. Anal Chem 77(11):221A–226A

Montgomery DC (2013) Introduction to statistical quality control, 7th edn. Wiley, New York

Shen M, Dong X, Tsong Y (2015) Equivalence margin determination for analytical biosimilar assessment, 2015 FDA Industry Statistics Workshop, September 16–18, Washington, DC

Tsong Y (2015) Statistical strategies for determining biosimilarities, 2015 Nonclinical Biostatistics Conference, October 13–15, Villanova, PA

Tsong Y, Shen M, Dong X (2015) Development of statistical approaches for analytical biosimilarity evaluation, 2015 DIA/FDA Statistical Forum, April, 2015, Rockville, MD

Velayudhan J, Chen Y-F, Rohrbach A, Pastula C, Maher G, Thomas H, Brown R, Born T (2016) Demonstration of functional similarity of proposed biosimilar ABP 501 to adalimumab. BioDrugs. doi:10.1007/s40259-016-0185-2. Published online July 15, 2016

Yang H, Novick S, Burdick R (2016) On statistical approaches for demonstrating analytical similarity in the presence of correlation. PDA J Pharm Sci Technol 70:6