

Determining the Number of Process Performance Qualification Batches Using Statistical Tools – Supplement to Prior Discussion Paper

This discussion paper is presented as supplement to the original discussion paper that was issued in August 2012, “Topic 1 – Stage 2 Process Validation: Determining and Justifying the Number of Process performance Qualification Batches (Version 2)” proposing ideas for answering the question “How many process performance qualification batches (PV stage 2) are needed to demonstrate a high degree of assurance in the manufacturing process and that the control strategy is sufficiently robust to support commercial release of the product”. The purpose of this paper is to present four (4) statistical tools that may be applied to determine the number of PPQ batches. This paper discusses the statistical approaches, their limitations, assumptions, and also presents simulated examples. Considerable input has already been received, considered, and/or incorporated. The author team is interested in hearing about other approaches that could be used, and lessons from use of the proposed approaches described in the discussion paper and additional examples. The paper may be modified or expanded sometime in the future to reflect additional input.

Please direct all feedback to pvpapers@ispe.org.

Determining the Number of Process Performance Qualification Batches Using Statistical Tools – Supplement to Prior Discussion Paper

Prior Discussion Paper: “Topic 1 – Stage 2 Process Validation: Determining and Justifying the Number of Process Performance Qualification Batches (Version 2)”

Authors: Christopher Breen (Eli Lilly), Dilip Ayyala Somayajula (Grand River Aseptic Manufacturing), Maneesha Altekar (AstraZeneca), Pritesh Patel (Pritesh Patel Consulting, LLC), Richard A. Lewis (GlaxoSmithKline)

1 Introduction

This is the second discussion paper written by the ISPE Process Validation (PV) Team on the topic of determining and justifying the number of initial process qualification batches (e.g., FDA Stage 2 [10], EudraLex Annex 15 process validation [13], EMA Guideline on process validation [14] [15], etc.) needed to demonstrate a high degree of assurance in the product manufacturing process and control strategy, and thereby support commercialization of the product. Most regulatory authorities agree on implementation of the lifecycle approach to PV. The science- and risk-based approach to determining the number of batches for the PV exercise is preferable, as opposed to defaulting to three batches. Differences of opinion and questions remain about the best way to translate the risk-assessment results into the actual number of batches, including acceptable statistical methodology options and the required level of statistical confidence, if applicable.

The Team’s original discussion paper, “Topic 1 – Stage 2 Process Validation: Determining and Justifying the Number of Process Performance Qualification Batches (Version 2),” by Bryder, et al., was issued in August 2012 and updated in 2014 [3]. It featured three approaches to assigning an appropriate number of PV batches to risk-assessment results in an attempt to address the expectation outlined in the FDA PV Guidance of 2011 [10], which stated that a high degree of assurance in the process must be obtained prior to release of commercial batches. This process assurance evaluation must also consider the result of process development (e.g., FDA Stage 1 [10], EudraLex Annex 15 planning for qualification and validation [13]) together with the results of the process-validation exercise (e.g., FDA Stage 2 [10], EudraLex Annex 15 process validation [13]). The three approaches presented in the first paper were useful in provoking thought and moving pharmaceutical industry understanding forward through various discussions, including discussions at scientific forums. As a result, several additional approaches have been developed, four of which are discussed in this paper.

2 Scope

The scope of this paper is limited to determining the number of initial PV, or Process Performance Qualification (PPQ) batches, using statistical tools identified in this paper. This paper supplements the “Topic 1 – Stage 2 Process Validation” Discussion Paper [3] which covers additional approaches in determination of number of PPQ batches.

This paper presents various statistical tools that can be used in determination of the number of PPQ batches. This paper will not attempt to explain basic statistics, but will present statistical tools with case studies or examples to demonstrate the use of the tool to arrive at the number of PV batches. Each statistical tool presented in this paper is intertwined with risk-based approaches, making it a hybrid approach. This paper assumes that a quality risk-management program is established, as described in ICH Q9, Quality Risk Management [6]. Alternatively, the risk-assessment methodology presented in the “Topic 1 – Stage 2 Process Validation” Discussion Paper may be used [3].

Strategies presented in this paper may be modified as needed by manufacturers to suit their established systems, processes, procedures, and/or policies. Suggestions and recommendations presented in this paper are intended to furnish a wide range of tools in determining the number of PV batches. They are not intended to represent an industry consensus, nor is it the paper’s intention to present one as more preferable than the others, as each approach may have its own limitations.

The statistical approaches, in combination with the risk ratings presented in this paper, can be applied to both Active Pharmaceutical Ingredients (APIs) and drug products. However, the risk assessment should consider the differences in the subject (i.e., API versus drug product).

3 Background

The “Topic 1 – Stage 2 Process Validation” Discussion Paper [3] explains FDA’s revised guidance on process validation and provides a risk-based framework for determining the number of batches for process validation. It discusses three risk-based approaches and outlines the pros and cons of each. Summaries of the three approaches are presented below:

Approach #1: Based on Rationales and Experience

This approach seems quite reasonable, but does not use statistics in the determination of the number of batches (see Table 1). It suggests using rationales based on historical information and process knowledge.

Approach #2: Target Process Confidence and Target Process Capability

This approach uses statistics, but it is immediately applicable only to normally distributed data and those processes that are in a state of statistical control. Although transformations from non-normal to normal distribution are sometimes possible, identifying a suitable transformation method can be a matter of trial and error. In addition, the user needs to draw conclusions consciously so as to apply them to the original process that generated non-normal data. This approach recommends implementing improvements to the process at the end of the PPQ batches if an expected process capability index (C_{pk}) range is not met.

Approach #3: Expected Coverage

This is a nonparametric approach based on order statistics—the minimum and maximum observed results are referred to as $X_{(1)}$ and $X_{(n)}$, respectively—and the properties of intervals based on these order statistics. In particular, the expected probability of a future observation being within the range defined

by $[x_{(1)}, x_{(n)}]$ is $(n - 1)/(n + 1)$. Users of this approach should select an appropriate level of coverage that is most suitable to their processes.

It is apparent that there is no single approach that fits all scenarios or processes, and that additional approaches may be useful.

4 Statistical Approaches for Determination of the Number of Batches

4.1 Tolerance Intervals

FDA guidance states that “The number of samples should be adequate to provide sufficient statistical confidence of quality both within a batch and between batches.” [10] This statement prompts two questions:

1. What is sufficient statistical confidence?
2. How will the between-batch quality be measured?¹

Both questions can be addressed by using tolerance intervals as PPQ protocol acceptance criteria for critical quality attributes (CQAs). Acceptance criteria in the PPQ protocol could be that the Tolerance Interval must be within the routine release criteria for the CQA (this may be the specification or a tighter internal requirement).

Tolerance intervals are intervals that cover a fixed proportion of the population (p) with a stated confidence (γ). A γ/p Tolerance Interval provides limits within which at least a specified proportion (p , also known as coverage) of future individual observations from a population is predicted to fall with a specified confidence level (γ). For example, a 95%/99% Tolerance Interval of 98.5 to 101.7 means that there is at least 95% confidence that at least 99% of future results will be between 98.5 and 101.7. For more information on tolerance intervals, refer to the *NIST/SEMATECH e-Handbook of Statistical Methods* [9].

All tolerance intervals discussed in this approach assume continuous normally distributed data, two-sided specifications, and k values approximated using the Howe method [5]. Additionally, it is assumed that the process is in a state of statistical control.

The confidence and coverage levels chosen should be based on a risk assessment, as in the “Topic 1 – Stage 2 Process Validation” Discussion Paper [3], where lower confidence and coverage levels are applied to a lower risk situation and higher confidence and coverage levels are applied to a higher risk situation. While all confidence and coverage selections are arbitrary, some minimum acceptable level should be stated. A level of 50% for either would be low. A floor of 80% is suggested, but this may be too

¹ An additional question might be “How will the within-batch quality be measured?” While between-batch criteria are relevant across all types of pharmaceutical production, within-batch criteria may be less impactful for scientifically homogenous products like monoclonal antibodies.

low for certain companies or applications. Table 1 provides an example of how these levels might be assigned:

Table 1: Residual Risk Level, Confidence and Coverage Assignment

Residual Risk Level	Confidence	Coverage
Severe (5)	N/A	N/A
High (4)	95%	99%
Moderate (3)	95%	95%
Low (2)	90%	90%
Minimal (1)	85%	85%

Many other level assignments are possible. The key feature of any assignment is that the levels be non-decreasing with risk. It is also suggested that the confidence and coverage levels be fixed well in advance of PPQ activities. In this way, the focus can be uniquely on the risk assessment process.

To determine the number of batches (N) needed for PPQ, one additional value should be set. Looking at the general equation of a tolerance interval:

$$\bar{Y} \pm k * S \text{ where } k = f(\gamma, p, N)$$

Where:

N = minimum number of PPQ batches: the total number of unique batch results included in the PPQ final report Tolerance Interval calculations. These results are **not** limited to batches run as part of PPQ, but can include previously identified stage 1 batches from comparable processes.

p = percent of the population to be contained in the Tolerance Interval.

γ = the statistical confidence that the Tolerance Interval contains p percent of the population.

f = a function where the inputs (γ, p, N) determine the output, k . This function is complicated and cannot be written in closed form.

k = output of $f(\gamma, p, N)$. It is a constant, which when multiplied by S , determines the γ/p Tolerance Interval width.

S = standard deviation of the N results.

\bar{Y} = the mean of the N results.

k is uniquely determined when three values (γ, p, N) are known. Similarly, when γ, p , and k are known, N is uniquely determined. It is this fact that will be used to determine the number of batches for PPQ in this approach. This fact can be summarized as:

$$k = f(\gamma, p, N) \Rightarrow N = f(\gamma, p, k)$$

Example: If $\gamma = 95\%$, $p = 95\%$, and $k = 3.5$, then $N = f(95\%, 95\%, 3.5) = 10$.

Interpretation: a 95%/95% Tolerance Interval needs a sample size (number of batches in this case) of at least 10 to have a k value no bigger than 3.5.

While choices for k are not extremely intuitive, reasonable values can be narrowed down based on common relationships. If $k = 3.0$, then the Tolerance Interval is the width of control chart limits ($\pm 3 \sigma$). If k were smaller than 3.0, then the Tolerance Interval would not “cover” the expected common cause variability. Therefore, k should be at least 3.0. At the other end, large values of k can be related to process capability measures. A k of 4.0 corresponds to a process performance index (P_{pk}) of 1.3, and a $k = 6.0$ to a $P_{pk} = 2.0$. While a $P_{pk} = 2.0$ (Six Sigma process) is a laudable goal, the final commercial specifications will likely be closer to 4σ . Additionally, the smaller k is restricted to, the more conservative the final sample size for validation will be. While there is no absolute answer, it seems reasonable that k should be restricted to between 3.0 and 4.0. In this document, $k = 3.5$ will be utilized.

Table 2: Number of Batches With Respect to Residual Risk Level

Residual Risk Level	Minimum Number of PPQ Batches	Readily Pass Tolerance Interval	Marginally Pass Tolerance Interval
Severe (5)	Not ready for PPQ	N/A	N/A
High (4)	24	95%/99%	80%/80%
Moderate (3)	10	95%/95%	80%/80%
Low (2)	5	90%/90%	80%/80%
Minimal (1)	4	85%/85%	80%/80%

“Minimum number of PPQ batches” refers to the total number of unique batch results included in the PPQ final report Tolerance Interval calculations. These results are **not** limited to batches run as part of PPQ, but can include previously identified stage 1 batches from comparable processes.²

Example: A new monoclonal antibody (drug substance) is being validated for the first time. The associated company has chosen (in advance) to use the confidence and coverage levels from this paper (along with a $k = 3.5$). A team of scientists go through a risk assessment. The residual risk level is determined to be moderate. Using this, the minimum number of batches for PPQ is determined to be 10. However, this can include previously identified stage 1 batches from comparable processes. It is determined that there are five such batches (three registration stability batches and two clinical trial batches). Therefore, at least five additional batches will need to be manufactured during PPQ. Criteria in the PPQ protocol for each quantitative CQA will be: 95%/95% Tolerance Interval on release results (all PPQ batches and five previously identified stage 1 batches) must fall within the acceptance criteria for the CQA.

² This is supported by FDA Guidance for Industry: “This conclusion should be based on a documented justification for the approval of the process, and release of lots produced by it to the market in consideration of the entire compilation of knowledge and information gained from the design stage through the process qualification stage.” [10]

If the 95%/95% Tolerance Interval is within the acceptance criteria for the CQA (i.e., readily pass criteria), then there is high confidence that the process can reproducibly comply with the CQA. Sampling and testing may be adjusted to a routine level in stage 3.

If the 80%/80% Tolerance Interval is within the acceptance criteria for the CQA (i.e., marginally pass criteria), then adequate confidence of quality in the CQA has been demonstrated between batches. Heightened monitoring and/or testing of the CQA is needed until a “readily pass” level of quality can be demonstrated for the CQA.

If the release results do **not** meet the “marginally pass” criteria, then the PPQ acceptance criteria has not been met and appropriate actions need to be taken before moving to stage 3.

4.2 Probability of Batch Success Approach

There are typically quite a few CQAs for each batch of drug substance or drug product, with specification limit(s) for each CQA. Examples of CQAs for drug substances include, but are not limited to, color, assay, related substances (impurities), moisture, residual solvents (such as ethanol or propanol), heavy metals, and particle size distribution. Examples of CQAs for drug products include but are not limited to appearance, assay, content uniformity, drug-related impurities, and tablet hardness and dissolution.

In this section we discuss using the probability that a batch will meet all of its specifications to determine the number of PPQ batches. This approach offers an intuitive and relatively simple methodology for determining the number of PPQ batches. In particular, it imposes no new process performance criteria (such as minimum values for process capability indexes) and naturally incorporates any number of specifications. It should be noted that this approach is particularly useful for high-volume processes with a large number of CQAs, all of which are likely to meet specification. The categorization of batches into successes and failures results in a loss of information that can increase sample size, and sample size may be further increased by batch failures.

Let p_i represent the probability of meeting the i^{th} specification, for $i = 1, 2, \dots, r$ specifications, and let p represent the overall probability of a batch meeting all of its specifications. The value of p is a complex function of p_1, p_2, \dots, p_r , where each p_i can involve both the true status of a CQA and measurement error. It may be tempting to think that $p = p_1 p_2 \cdots p_r$, but this expression ignores dependencies among the CQAs. For example, tablet hardness may be highly correlated with dissolution, and impurities may be correlated with one another. Fortunately, the approach used in this section does not need the individual values p_1, p_2, \dots, p_r . Instead, the binomial distribution is used to determine the probability of observing k successful batches in n independent trials:

$$P(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where k is the number of successful batches (i.e., the number of batches meeting all specifications), n is the number of batches produced, and p is the overall probability of a successful batch.

Suppose we have observed k successful batches in n independent trials and wish to make a statement about the overall probability of success, p . Statistical inference for p is a well-studied problem. A common estimator of p is the maximum likelihood estimator $\hat{p} = k/n$, and there are numerous methods for constructing confidence intervals for p [2]. Some commonly used confidence intervals for p are the Agresti–Coull interval [1], the Clopper–Pearson interval [4], the Jeffreys interval [2], the Wilson score interval [11], and the Wald interval.

The Wald interval is based on a normal approximation; it is easy to compute, but has poor coverage properties [2]. The other four intervals have reasonably good coverage properties, but the Clopper–Pearson interval is the only one that guarantees the stated level of confidence. This guarantee comes at a cost however; Brown et al. deem the Clopper–Pearson interval “wastefully conservative” [2]. In the remainder of this section we use either the Wilson score interval or the Jeffreys interval.

Bayesian statistical methodology can also be used to make inferences about the probability of success, p . Bayesian methodology allows one to formally include prior information about p in the analysis. A Bayesian approach that utilizes an informative prior distribution for p is discussed by Yang [12]. Here we focus on using a non-informative prior in conjunction with a risk analysis, but note that an informative prior may be very useful when making minor modifications to an existing process. The well-known conjugate prior distribution for p is a beta distribution with parameters (α_0, β_0) . The parameter α_0 can be thought of as the prior number of successful batches, and the parameter β_0 can be thought of as the prior number of failed batches. The prior mean of p is simply $\alpha_0/(\alpha_0 + \beta_0)$. Two popular non-informative priors are the Jeffreys prior ($\alpha_0 = \beta_0 = 1/2$) and a uniform prior ($\alpha_0 = \beta_0 = 1$). The ensuing posterior distribution for p is also a beta distribution, with parameters $(\alpha_1, \beta_1) = (\alpha_0 + k, \beta_0 + n - k)$. It can be seen that the conjugate prior adds α_0 successes and β_0 failures to the observed results. Credibility intervals for p , a Bayesian version of confidence intervals, can be obtained using percentiles of the posterior distribution. The Jeffreys prior leads to the Jeffreys interval previously discussed in this section.

The number of PPQ batches can be based on statistical inference for p , the overall probability of a batch meeting all of its specifications. We focus on the situation in which we have observed n successful batches in n independent trials. Situations in which one or more batch failures have been observed will depend on the results of a root cause analysis and the perceived effectiveness of corrective and preventive actions. It is possible that some previous batch failures can be ignored if they are considered no longer probable. Figure 1 shows one-sided 80% lower confidence limits (LCLs) for p , based on the Wilson score and Jeffreys intervals, and

Figure 2 shows one-sided 95% LCLs. Note that fairly large sample sizes are needed to obtain 95% confidence in a specific value of p , relative to 80% confidence. For example, using the Wilson score interval we are 80% confident that p is at least 0.93 if we observe 10 successful batches in 10 independent trials, while we are 95% confident that p is at least 0.92 if we observe 30 successful batches in 30 independent trials.

As an example of using the probability of batch success to determine the number of PPQ batches, suppose a project has completed 10 successful clinical batches with same process and formulation used in the regulatory filing, with no batch failures. Based on a risk assessment that results in a moderate level of residual risk, the project team has determined that they would like 90% confidence that the probability of batch success is 90% or greater. This requires a total of 15 batches with no batch failures, so the project team recommends five PPQ batches. If all five batches are successful, the process will be deemed validated and further information on process performance will be obtained in stage 3 (continued process verification). Sample sizes for additional levels of residual risk are illustrated in Table 3, which is provided for illustrative purposes only. Table 4 provides a Bayesian version based on the Jeffreys prior. These tables show that a large number of batches are needed to obtain high levels of confidence in p . Consequently, manufacturers should give considerable thought to the level of confidence that is needed for a specified residual risk, and develop their own versions of Table 3 and Table 4.

Figure 1: One-Sided 80% Lower Confidence Limits for p

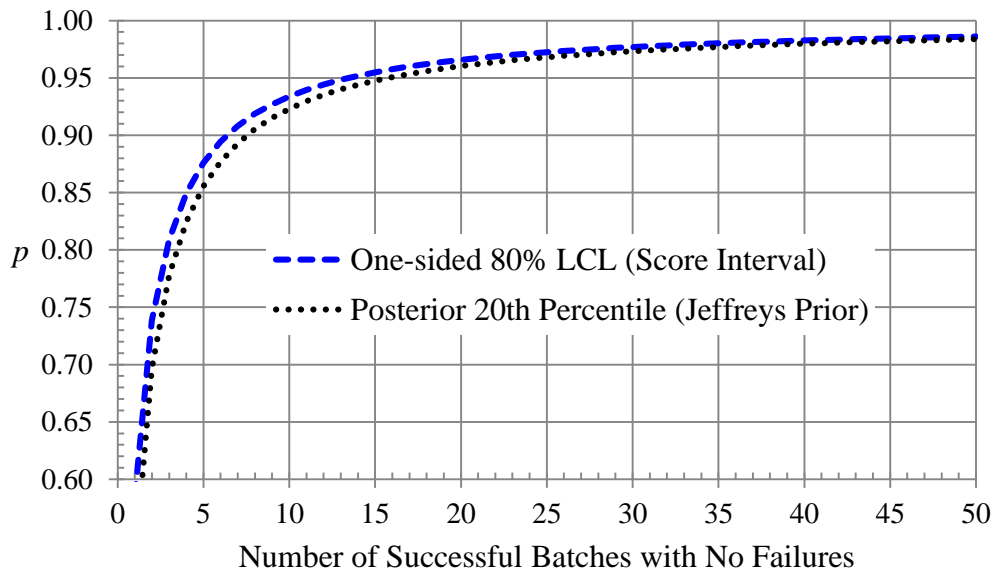


Figure 2: One-sided 95% Lower Confidence Limits for p

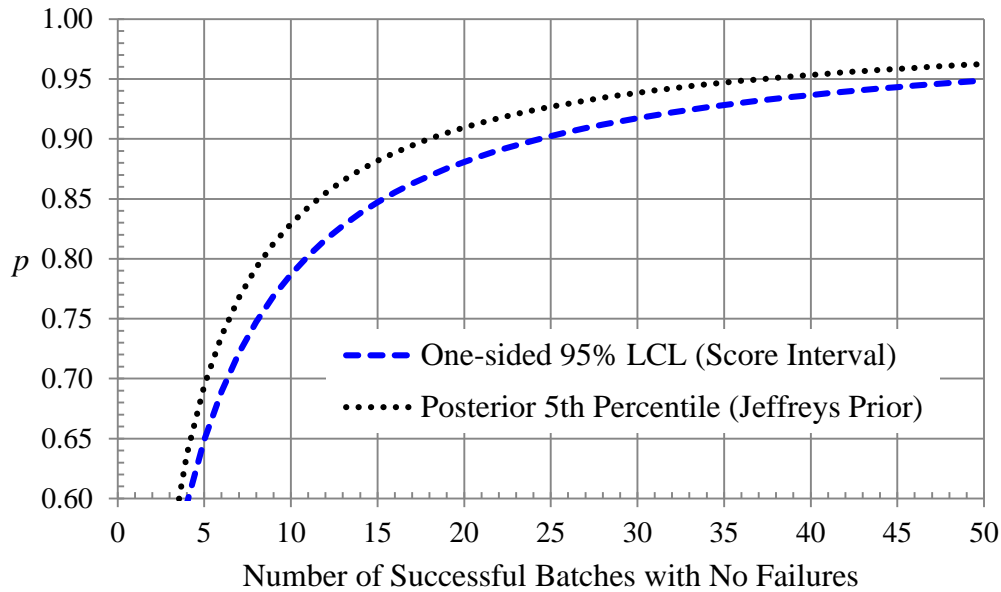


Table 3: A Straw Model of a Confidence-Based Approach (Wilson Score Interval)

Residual Risk Level	Total Number of Batches with No Failures	Confidence Level	One-sided LCL for p
Severe	N/A	N/A	Not ready for PPQ
High	24	95%	89.9%
Moderate	15	90%	90.1%
Low	10	85%	90.3%
Minimal	7	80%	90.8%

Table 4: A Straw Model of a Bayesian Approach (Jeffreys Beta ($\frac{1}{2}$, $\frac{1}{2}$) Prior)

Residual Risk Level	Number of PPQ Batches	Probability Level	One-sided LCL for p
Severe	N/A	N/A	Not ready for PPQ
High	18	95%	90.0%
Moderate	13	90%	90.3%
Low	10	85%	90.4%
Minimal	8	80%	90.5%

Note: If one or more of the CQAs poses a significant risk of not meeting specification, it might be more useful to examine those sub-processes in more detail and use an alternative approach to PPQ sample size, or perhaps to model the probability of meeting those specific specifications, as in LeBlond and Mockus [7].

4.3 Combinational Approach of Analysis of Variance with Risk Assessments

For this approach to determine the number of batches required, the observed variation from stage 1 batches is partitioned into the within-batch and between-batch components. Data from commercial scale engineering/demo batches would be the ideal data set to analyze. However, this may not be possible, and thus the user should use scientific and technical expertise when determining if smaller scale batch data can be used. The between-batch component is then used with a National Institute of Standards and Technology (NIST) based calculation to determine finally the number of batches required. Using just the observed variation could overly estimate the number of required batches, thus increasing costs and lengthening the time required to complete the validation study. FDA process validation guidance [10] states that between-batch and within-batch variability needs to be understood. This could be in terms of what can cause these variabilities and could include the ability to quantify them.

The main prerequisite needed is to understand what “shift” in the mean response of the parameter is considered to be of practical significance. The baseline for the shift consideration is the mean of the response of the attribute as seen in the demo or engineering runs. Because specification limits were used during the manufacturing of phase I and phase II batches, one suggestion would be to define this shift in mean as a percentage of the specification window, and leverage some form of risk analysis that looks at the criticality of the quality attribute in question, as well as the residual risks identified after a control strategy has been identified (via formal risk assessment) and implemented. Figure 3 and Figure 4 summarize this approach.

Figure 3: Risk Index Identified for Each Quality Attribute and Residual Risk Interaction

		Quality Attribute's criticality is...		
		High	Medium	Low
Residual Risk is...	High	A	A	B
	Moderate	A	B	C
	Low	A	C	C
	Minimal	B	C	C

Using the risk index from Figure 3, the respective shift can be determined as seen in Figure 4:

Figure 4: Recommended Shifts to be of Practical Significance

The Risk Index is...		
A	B	C
Shift to be no more than 10% of Spec Window	Shift to be no more than 15% of Spec Window	Shift to be no more than 20% of Spec Window

The values of 10%, 15%, and 20% are selected based on the risk of significantly changing the C_{pk} and hence the out of specification product being produced. Table 5 shows how out of specification (OOS), parts per million (PPM) failure, and C_{pk} change as the shifts get bigger. Table 5 was calculated using the Statistical software, JMP, which is a product of the SAS company using Avg. = 100; Std. Dev = 2, USL = 110, and LSL = 90.

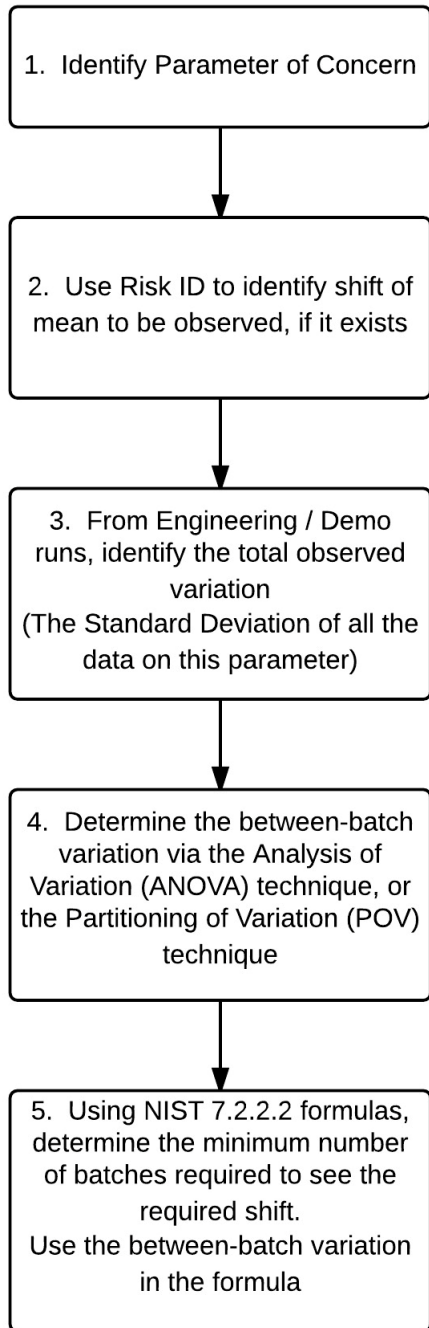
Table 5: Changes in PPM and OOS for Different Mean Shifts to Be Observed

% of Spec Window	Actual Shift	Cpk	OOS	PPM Failure	Change in Cpk
0.00%	0	1.67	0.0000287%	0	0.00%
10.00%	2	1.33	0.0031671%	32	-20.00%
15.00%	3	1.17	0.0232629%	233	-30.00%
20.00%	4	1.00	0.1349898%	1,350	-40.00%

As the criticality of the attribute goes up, the acceptable mean shift to be seen, if it exists, gets smaller.

Given this understanding of a mean shift that is practically significant, the flow map in Figure 5 can be constructed to determine the minimum number of batches to be run for validation, as well as minimum samples to be taken per batch.

Figure 5: Flow Map to Determine Minimum Number of Batches



Once the mean shift to look for has been identified, one can use section 7.2.2.2 of the *NIST/SEMATECH e-Handbook of Statistical Methods* [9] to calculate the minimum sample size required. The between-batch variation can be determined using the approach discussed by Little [8].

Example: Data for CQA1 for product “W” gave the graph, as seen in Figure 6, for five engineering batches. The specification for the CQA1 is 90 – 110.

These engineering batches were executed by including as much input (raw materials, operators, environmental conditions, non-key process parameters) variation as possible.

Figure 6: Plot of CQA1 from Five Batches of Product “W”

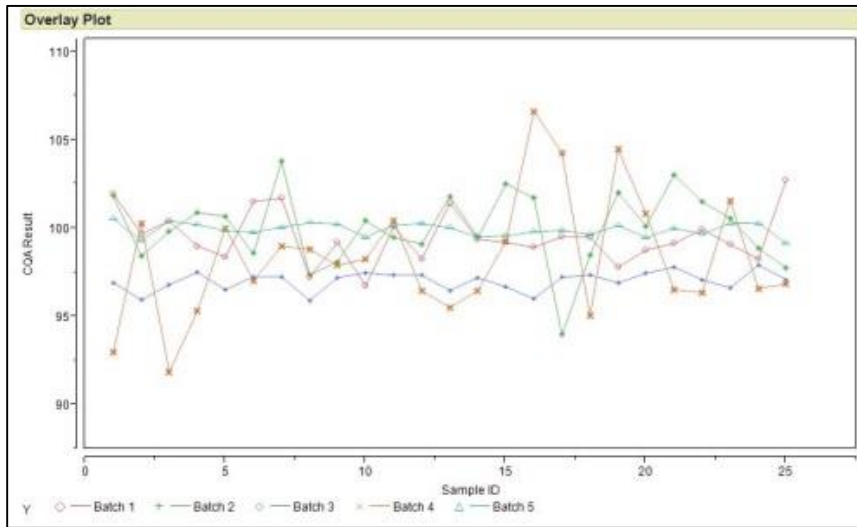
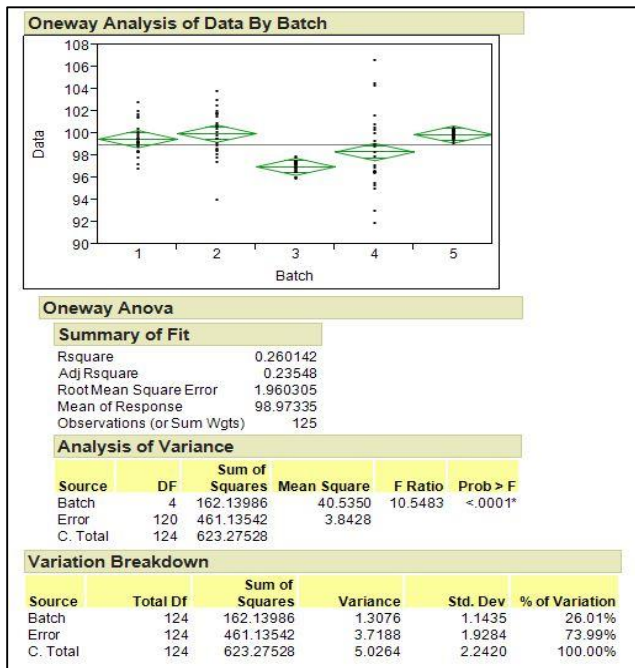


Figure 7: Analysis of Variance (ANOVA) and Process Output Variable (POV) Breakdown



From Figure 7, using the POV analysis (sum squares/total Df = variance) one can tell that 26% of the overall variation comes from between-batch variation, and the rest (74%) comes within-batch variation. In addition, one can determine that the total observed standard deviation is 2.242 and the total variance is 5.0264.

Now, to determine the shift in the mean response of the parameter, given that this is a CQA with a specification of 90 to 110, the desired shift in mean is determined to be no more than 10% of the specification window or $[0.1 \times (110 - 90)] = < 2.0$.

Table 6 shows the calculations based on the *NIST/SEMATECH e-Handbook of Statistical Methods* [9].

Table 6: Output of Sample Size Calculation based on the NIST Approach

USL	110									
LSL	90									
Tol	20									
Mean shift that is of Practical Significance	2									
Total Variation	5.026564									
% var (between batch)	26.00%									
Between Batch Var	1.30690664									
Between Batch S.D.	1.143200175									
alpha	0.05									
Beta	0.1									
1-beta =	0.9									
z(1-beta) =	1.28									
1-alpha/2 =	0.975									
z (1-alpha/2) =	1.96									
N1 =	4									
N	4	8	5	7	5	7	5	7	5	7
t(1-alpha/2)	3.182	2.365	2.776	2.447	2.776	2.447	2.776	2.447	2.776	2.447
t(1-beta)	1.638	1.415	1.533	1.440	1.533	1.440	1.533	1.440	1.533	1.440
New N	8	5	7	5	7	5	7	5	7	5
Number of batches required	6									

Using the NIST approach and the POV variation analysis method suggested by Little [8], one needs a minimum of six batches to clearly see a shift in mean of 10% of the specification window, if this shift exists.

A similar approach was conducted for the remaining two CQAs and the 2 parameters of medium criticality. Table 7 summarizes the results of each.

Table 7: Summary of Calculations for All Product Characteristics

CQA	Min. Number of Batches Required
CQA1	6
CQA2	5
CQA3	4
Med4	3
Med5	4

For this particular product, running a minimum of six batches for validation will allow seeing the required mean shifts, should they exist. The validation protocol can contain an acceptance criteria to specify “The mean of <<Quality Attribute ‘A’>> should be between $X \pm S$ ” where X is the mean value of the quality attribute as derived from the demo/engineering studies, and S is the shift in mean that is of practical significance. If the value of the mean of the combined batches is outside of this range, then an investigation should be conducted to understand why this is so and appropriate actions should be taken. It goes without saying that this should be one of several acceptance criteria to judge the success of the validation.

One of the limitations of this approach is that it is not possible to make confidence statements such as “ $X\%$ confidences that $Y\%$ will be within specification”. The intent of the outlined approach was never to show that the lots would be within specification; that should have been demonstrated as part of stage 1 development and engineering studies. Instead, the intent of this approach is to show a consistent performance of the process being validated and that there is no difference between the commercial scale process at stage 1 and stage 2. Another limitation of this approach is that the stage 1 batches are assumed to be at commercial scale, which may not be common for small-molecule manufacturing. One unintended consequence of this approach is that the shift in mean to detect could be widened. This would lead to fewer batches being needed for stage 2 purposes.

4.4 Variability Based Approach

One criterion for a successful validation is that the process exhibits low variability. In this approach, a predetermined estimate of variability is used as the basis for selecting the number of batches for validation.

This approach is applicable when validating a process following a significant change to the current process or when a process is being transferred from one site to another. In such cases, there is a well-established estimate of process variability based on the historical process data. The goal of the validation is to ensure that the process being validated continues to show variability similar to, or no worse than, the current process. Although it is important for the process average to remain on target as

well, there may be differences between the current and new process averages due to factors such as equipment, setup, and raw materials being used. This approach assumes that adjustments to achieve the target process average can be made more easily once the process variability is deemed acceptable.

If the current process has been stable prior to the PPQ, the historical estimate of variability from that process can then be thought of as a “known” parameter or a population value. If the data from validation batches can be shown to have the same variability as the current process, the estimate of variability from the current process can be used to determine the capability of the process being validated.

A power curve is used to select a sample size to test the hypothesis that the ratio of the sample standard deviation to the population value is no more than a specified value R . The hypothesis to be tested is

$$H_0: (S_{DV}/\sigma_p) \leq R$$

$$H_1: (S_{DV}/\sigma_p) > R$$

Where:

R = ratio of validation standard deviation to historical population standard deviation

H = null or alternative hypothesis

S_{DV} = the standard deviation associated with the validation batches

σ_p = the known standard deviation from the historical batches of the current process.

The confidence level $(1 - \alpha)$ and power $(1 - \beta)$ should be specified, where α is the probability of rejecting the null hypothesis when it is true and β is the probability of accepting the null hypothesis when it is false.

The choice of R depends on the process knowledge available, including the impact of the process change to the process being validated. A value of 2 can be used as a good rule of thumb to indicate similarity of variability, especially when the risk to the process is low. A lower value of R may be preferred when risk is high.

The choice of α and β will also depend on the identified residual risk. For example, for quality attributes that are associated with moderate risk, smaller α and β values may be chosen (e.g., 0.05 and 0.2 respectively), whereas for quality attributes that are associated with lower risk, higher α and β values may be chosen (e.g., 0.05 and 0.3 respectively). Higher values of β mean that there is greater probability that we would find the validation variability acceptable when it is actually unacceptably high. If the validation is being performed for a technology transfer or after changes to the current process, the residual risk is generally determined to be small due to the abundance of information available from the current process. The number of batches for a given residual risk is summarized in Table 8; the risk levels listed are consistent with those defined in the “Topic 1 – Stage 2 Process Validation” Discussion Paper

[3], as well as the topics discussed in this paper. Other choices of α and β are possible, and these should be determined by each organization during their risk assessment process.

Table 8: Number of Batches Determination

Residual Risk Level	Ratio	Target Confidence Level	Target Power	Number of Batches
Severe	N/A	N/A	N/A	N/A
High	2	95%	90%	11
Moderate	2	95%	80%	8
Low	2	95%	70%	6
Minimal	2	80%	70%	4

Note: There is a certain level of arbitrariness in selecting a particular confidence level or power for a given risk level. It is up to the organization to decide whether the chosen confidence level and/or power are consistent with their understanding of the risk level identified and impact to the process and/or product.

For example, if eight validation batches are run for a moderate risk level, there is an 80% chance of detecting a ratio exceeding 2.

If the product being validated has multiple variants, the total number of batches may be made up across the different variants, as appropriate. Based on the risk assessment, certain variants may be replicated more than others. The total number of batches may also include process development batches or any other batches made by the same manufacturing process as that used for stage 2. Statistical analysis should be performed on the data collected for the CQAs. A test of hypothesis should be performed on the sample S_{DV} to determine whether its ratio to the historical standard deviation (σ_p) does not exceed a specified value. If multiple variants have been used, then it may be possible to combine data for some CQAs across some or all variants, as these CQA values do not vary across the variants. This helps ensure appropriate power for the analysis.

If the null hypothesis is not rejected and the ratio does not exceed the specified value, the validation batches may be considered to be no different from the current process with respect to the standard deviation, and the historical estimate of the process standard deviation can be used for any further evaluation of the process, fe.g., in calculating process capability. Since a known estimate of variability is being used, there is less uncertainty in the estimate of process capability.

If the null hypothesis is rejected and ratio is greater than the specified value, the standard deviation of the validation batches is unacceptably higher than that of the current process. In this case, additional work may be needed to better understand and mitigate the sources of variability before validation can be considered complete.

Additional analyses should be performed to assess the capability of the process using the historical estimate of the standard deviation and the sample mean from the validation batches. These should be

determined when planning the process validation and will depend on the risk assessment performed. Various analyses and/or criteria for passing validation are possible; some of which are mentioned below:

- All validation batches are within specification
- Calculate C_{pk} and compare it to historical C_{pk}
- Calculate C_{pk} and compare against a stated acceptance criteria

If C_{pk} is being calculated, then the target C_{pk} values should be determined by the organization prior to validation, based on the product being validated. Acceptance criteria based on C_{pk} were previously proposed in the “Topic 1 – Stage 2 Process Validation” Discussion Paper [3] and may be used to judge the validation. These criteria use the confidence interval for the C_{pk} to determine the acceptability of the batch; this is necessary since very few batches are available for this calculation, leading to greater uncertainty in its estimate. An example is provided in Table 9 using the number of batches proposed in Table 8.

Table 9: Example Showing Acceptance Criteria Assignment With Respect to Residual Risk Level, Number of Batches

Residual Risk Level	Number of Batches	Target Lower Confidence Bound for C_{pk}	Acceptance Criteria		
			If $C_{pk} \geq 1.6$, readily pass validation	If $1.0 \leq C_{pk} < 1.6$, marginally pass/more work needed	If $C_{pk} < 1.0$, fail validation
High	11	95%			
Moderate	8	90%			
Low	6	85%			
Minimal	4	80%			

If the observed $C_{pk} \geq 1.6$, the true C_{pk} value is at least 1.0 in each case above. Thus, there is high confidence that the sample mean is also close to the target value; in this case, we can consider the validation to have passed easily, and enhanced monitoring may not be necessary. However, since the validation batches may not represent all sources of process variability, the process variability should continue to be monitored to ensure that it continues to meet the target set for validation. If the $1.0 \leq C_{pk} < 1.6$, there is less confidence that the process mean is at or close to target and further work is needed to a) better understand and improve the mean, and b) to reduce the variability further, as needed. The process should be evaluated further to identify and/or adjust process parameters that impact the mean. Depending on the residual risk associated with the process, this may be done before validation can be considered complete or during the enhanced monitoring stage. Finally, if $C_{pk} < 1.0$, the process is at high risk of producing OOS batches during manufacture and cannot be validated.

The approach provided assumes that the process is stable and the data are normally distributed. These assumptions should continue to be verified as more data become available during enhanced or routine monitoring.

Example: Validation was to be performed on an existing manufacturing process following a significant change to the process. The existing process had been used for a long time and there was a wealth of

information and data for this process. The product was made in two batch sizes, had three strengths, and was sold in two market segments with slightly different manufacturing requirements.

During the risk assessment process, which included analyses of nearly 2 years of historical data, the following was concluded:

1. Ten of the 12 combinations of product/process needed to be evaluated during validation; therefore, a minimum of 10 batches would need to be run. The remaining two combinations were not going to be manufactured.
2. Some CQAs could be combined across all combinations; these did not vary across markets or strengths, and were made to the same specification. Others could only be combined for markets or only for product strength.
3. Some CQAs show more variability for some combinations than others.
4. Not all CQAs are equally critical from a validation impact perspective; some were impacted by the change to the process while others were not.

The risk analysis resulted in the selection of a minimum power of 80% to test the ratio of the standard deviations for various CQAs. Thus, the smallest number of batches needed for any CQA was 8. In order to achieve this power, the following replicates were run (see Table 10):

Table 10: Number of Batches

Product Strength	Market 1		Market 2	
	Batch Size 1	Batch Size 2	Batch Size 1	Batch Size 2
1	x	2	x	2
2	2	2	1	1
3	2	2	2	2

Based on the risk assessment performed, for one CQA (assay), variability for product strength “3” was evaluated on its own using eight batches by combining markets and batch size, while variability for product strengths “1” and “2” was evaluated using 10 batches by combining strengths, markets, and batch size. For other CQAs, all 18 batches were used in the evaluation.

5 Conclusion and Further Discussion

The approaches presented in this discussion paper use a combination of statistics and risk assessments, making them hybrid approaches for determining the number of PPQ batches. Four possible approaches are presented with examples and/or case studies.

A Tolerance Interval approach is presented that recommends using tolerance intervals as the PPQ protocol acceptance criteria. Using risk assessments, the confidence and coverage levels may be determined, lower confidence and coverage levels may be applied to a lower risk situation, and vice versa.

The probability of batch success approach described in this paper provides a starting point that is particularly useful for a high-volume process with a large number of CQAs, none of which pose a significant risk of not meeting specification. This approach incorporates any number of specifications without imposing any additional process performance criteria, such as minimum values for process capability indexes.

The combinational approach of analysis of variance approach with risk assessments presented in this paper is based on the shift in the mean response of the key parameter or CQAs. The number of PPQ batches required is estimated by determining the between-batch variation using an ANOVA or POV technique and utilizing NIST 7.2.2.2 formulas [9] for the required shift in mean of a key parameter or CQA.

The variability-based approach presented in this paper is based on the historical variability and the selection of the target ratio, confidence level, and power appropriate with residual risk. This approach is applicable to processes that undergo a significant change or have been transferred from one site to another.

Other approaches may be used, and industry is encouraged to collaborate further with ISPE to support progress in the PV lifecycle approach.

References

1. Agresti, A., and B.A. Coull. "Approximate Is Better Than 'Exact' for Interval Estimation of Binomial Proportions," *American Statistician* 52 (1998): 119–126.
2. Brown, L. D., T. T. Cai, and A. Das Gupta, A. "Interval Estimation for a Binomial Proportion," *Statistical Science*, 16 (2001): 101–133.
3. Bryder, M., et al. "Topic 1—Stage 2 Process Validation: Determining and Justifying the Number of Process Performance Qualification Batches (Version 2)." ISPE Discussion Paper. August 2012, updated 2014. www.ispe.org/discussion-papers/stage-2-process-validation.pdf.
4. Clopper, C. J., and E.S. Pearson, E. S. "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika* 26(1934): 404–413.
5. Howe, W. G. "Two-Sided Tolerance Limits for Normal Populations—Some Improvements," *Journal of the American Statistical Association* 64, no. 326 (1969): 610–620.
6. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guideline. "Quality Risk Management: Q9" 9 November 2005. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q9/Step4/Q9_Guideline.pdf
7. LeBlond, D., and L. Mockus. "The Posterior Probability of Passing a Compendial Standard, Part 1: Uniformity of Dosage Units," *Statistics in Biopharmaceutical Research*, 6 (2014): 270–286.
8. Little, T. A. "Understanding and Modeling Product and Process Variation," *BioPharm International* 26, no. 12 (December 2013): 50–53, 56. <http://www.biopharminternational.com/understanding-and-modeling-product-and-process-variation>.
9. National Institute of Standards and Technology. *NIST/SEMATECH e-Handbook of Statistical Methods*. Updated 30 October 2013. <http://www.itl.nist.gov/div898/handbook>.
10. US Food and Drug Administration. Guidance for Industry, Process Validation: General Principles and Practices, January 2011. <http://www.fda.gov/downloads/Drugs/.../Guidances/UCM070336.pdf>.
11. Wilson, E. B. "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association* 22 (1927): 209–212.
12. Yang, H. "How Many Batches Are Needed for Process Validation under the New FDA Guidance?" *PDA Journal of Pharmaceutical Science and Technology* 67 (2013): 53–62.
13. EU Guidelines for Good Manufacturing Practice for Medicinal Products for Human and Veterinary Use, Annex 15: Qualification and Validation, 30 March 2015.
14. EMA Guideline on process validation for the manufacture of biotechnology-derived active substance and data to be provided in the regulatory submission, 25 April 2014.
15. EMA Guideline on process validation for finished products – information and data to be provided in regulatory submissions, 27 February 2014.