Establish methodology for estimating process performance capability during the design phase for biopharmaceutical processes

by

Rashmeet Sangari

B.S. Chemical and Biochemical Engineering, Rutgers University, 2006

Submitted to the MIT Sloan School of Management and the Institute for Data, Systems, and Society (IDSS) in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Systems Engineering

In conjunction with the Leaders for Global Operations Program at the
Massachusetts Institute of Technology

September 2016

Signature of Author Signature redacted
MIT Sloan School of Management and Institute for Data, Systems and Society
Aug 5, 2016

Certified by Signature redacted
Daniel Whitney, Senior Research Scientist, Emeritus, MIT Institute for Data, Systems and Society

Certified by Signature redacted
Roy E. Welsch, Professor of Statistics and Management Science, MIT Sloan School of Management

Approved By Signature redacted
John N. Tsitsiklis, Clarence J. Lebel Professor of Electrical Engineering, IDSS Graduate Officer

Approved By Signature redacted
Maura Herson, Director, MBA Program, MIT Sloan School of Management

This page is intentionally left blank

Establish methodology for estimating process performance capability during the design phase for biopharmaceutical processes

by

Rashmeet Sangari

Submitted to the MIT Sloan School of Management and the Institute for Data, Systems, and Society (IDSS) on August 5, 2016 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Systems Engineering.

Abstract

It is highly desirable to apply a systematic methodology in order to measure, improve and ensure robust process performance during process development (PD) for biopharmaceutical processes. Designing a standardized approach to leverage process performance capability or Ppk during the process design phase will enable processes to consistently meet pre-defined targets for performance and facilitate continuous improvement. The project goal is to define the criteria, which includes recommendations on optimal sample sizes, techniques to account for measurement variation, and options to leverage platform variation knowledge, for determining Ppk during process design of a biopharmaceutical product. This can be especially challenging to ascertain in early phase for biopharmaceutical processes due to low sample sizes and long product cycle times. The scope of this project includes performance measurements in drug substance manufacturing from unit operations: vial thaw to final drug substance formulation. The assessment begins with selection of an upstream or downstream process parameter followed by evaluation of variation due to the analytical method and concludes with the determination of optimal sample size and appropriate next steps. A novel decision matrix using Intraclass Correlation Coefficient (ICC) and Precision to Tolerance (P/T) ratios is being recommended to determine the method variation. A procedure has also been designed to filter out method variation to calculate the Ppk of only the process. Additionally, the effect of sample size on Ppk has been established. Based on the results from bootstrapping, a multi-tiered approach taking into account the differing complexity of parameters to estimate Ppk at small sample sizes is being recommended. To elaborate, sample sizes greater than 11 independent batches typically provide sufficient confidence in Ppk estimates. Sample sizes from 8 to 11 batches will require uncertainty to be evaluated along with Ppk calculation, due to the width of confidence intervals. For sample sizes less than 8, the confidence intervals for the Ppk statistic are too large and not reliable. In this scenario, Bayesian analysis, a powerful tool for prediction of data, is conducted using data from other scales and molecules as historical database to estimate Ppk and the credible interval around it.

Thesis Supervisor: **Daniel Whitney**, Senior Research Scientist, Emeritus,
MIT Institute for Data, Systems and Society
Thesis Supervisor: **Roy E. Welsch**, Professor of Statistics and Management Science,
MIT Sloan School of Management

This page is intentionally left blank

Acknowledgements

First and foremost, I would to like express my gratitude to Amgen for sponsoring my LGO internship. I really enjoyed my experience within Amgen's process development organization and the internship provided me with a rich learning experience.

I especially want to thank my manager, Tom Mistretta for providing significant guidance and support and my director, Patrick Gammell for being a fantastic sponsor and providing me regular input as well as opportunities to showcase my work. I also want to thank Greg Naugle and Aine Hanly for being fully supportive of this project and communicating the organization's needs very clearly. Thanks to Dave Stadolnik for managing my internship experience so well.

A heartfelt thanks to Keith Britt, my core team member, who was amazing to work with and with whom brainstorming discussions on new ideas were very effective and fun! Thanks to my extended team members: Justin Ladwig from Pivotal Drug Substance, Grant Cairns from Drug Substance and Technology, Shekar Ganesa from Attribute Sciences, Sydney Rush from Process Development for providing me support and feedback through the project.

The successful completion of this thesis would not have been possible without the critical guidance from both my MIT faculty advisors Dr. Daniel Whitney and Dr. Roy E. Welsch, who amaze me with their intellect and experience. I am grateful for their insights and their time in guiding me through this project. I would also like to acknowledge the Leaders for Global Operations program administration, in particular, Ted Equi, Patricia Eames and Thomas Roemer for their support of this work.

Last but not the least, I would like to thank my family for their tremendous support and love: my husband, Dr. Satinder Singh, who has dedicated his life to serving patients and who inspires me to improve each and every day, my daughter, Simar Kaur, whom I want to inspire to serve and my parents, Drs. Santosh and Harpal Sangari, who have inculcated good values in me and have always been there for me irrespective of the challenge. I am forever thankful to have my family in my life.

This page is intentionally left blank

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

This page is intentionally left blank

1.0 Introduction

The content within this thesis summarizes the outcome of a six-month internship with Amgen Inc., a multinational biopharmaceutical company, from January 2016 to July 2016. This internship is being championed by the process development organization at Amgen.

The key deliverable of this thesis is to establish standardized a methodology for estimating process performance capability (Ppk) during the design phase of process development for biopharmaceutical processes. I am the team lead for this project and my core team includes my supervisor, a principal engineer, and an engineer in the Process Monitoring Analysis and Control (PMAC) group within process development. My extended team consists of scientists and engineers from Pivotal Drug Substance, PMAC, Drug Substance Technology and Engineering and Attribute Sciences.

1.1 Problem Statement

The role of process development at Amgen is to develop the processes that deliver Amgen's products to patients. The Process Monitoring Analytics and Control (PMAC) group, where this thesis has been conducted, provides process monitoring and remote commercial support for Amgen's drug substance manufacturing facilities and conducts process capability assessments during the development lifecycle for Amgen's pipeline molecules. These deliverables ensure process consistency, identify any excursion outside of normal process variability, and provide basis for corrective or lifecycle improvement opportunities.

Currently, in the biopharmaceutical industry, it is not routine to assess process capability during the process design phase, which is the first stage of process validation (reviewed in more detail in Chapter 2). Process capability data are typically assessed for commercial products during continued process verification (CPV), which is the third and final stage of the validation lifecycle. During this stage, the sample size becomes large enough for robust evaluations.

In the biopharmaceutical industry, performing Ppk analysis during the process design phase is especially challenging due to low sample sizes, long product cycle times, cost and resources to generate more data or information. Predicting process capability during process design can ensure robust and capable performance during the process performance qualification (PPQ) phase and post-launch during the CPV phase of the validation lifecycle. This can have significant impact on cost, resources and compliance. Early pilot studies have demonstrated the value to predicting capability during the process design phase when some low capability parameters were detected and risks were mitigated. This resulted in a successful PPQ campaign and avoided delays for an important pipeline program.

1.2 Project Goals

The over-arching goal of this project is to establish standardized methodology for estimating process performance capability during the design phase of process development. This project includes three sub-projects.

First and foremost, the criteria for performing process performance capability assessments specifically for small sample sizes during development or design phase needs to be defined. Based on the literature, a sample size of 30[1] is established as the current threshold to performing process capability assessments and in the process design phase, we typically have much smaller sample sizes than 30.

Second, this methodology also needs to include techniques to assess and filter out measurement system variation (i.e. measurement error) from overall process variation. This assessment is critical to ensure that the process capability we are focusing on is actually of the process and does not include measurement system variability.

Third, since we are working with a small data set in the process design phase, the techniques to improve estimates of process performance capability when the sample size is below a minimum threshold need to be recommended. Similarities in design and performance from other processes and molecules need to be taken into account to achieve this objective.

The results from the above three sub-projects will serve as an input to answering how process capability can inform process design.

1.3 Project Scope

The scope for the thesis includes looking at the drug substance process (vial thaw to drug substance formulation). Parameters from various unit operations from a drug substance manufacturing process have been evaluated to develop and test the recommended methodology. These include Seed bioreactor final viability (%), Column 1 step yield, Host cell protein after column 1, Final UF/DF protein concentration, Drug substance glycan content, Drug substance % relative potency and Drug substance pH. An explanation of these parameters is included in Section 5.2. The recommendations or the methods developed will also be evaluated for application to other phases of product life cycle e.g. Drug Product development in the future.

1.4 Project Approach

The key goal of this project is to establish a methodology to reliably estimate process capability in the design phase of process development. The problem has been separated into 3 research areas and the approach for these areas is outlined below:

I. **Process Capability Methodology:** Conduct Statistical analysis to determine the appropriate metric, sample size, and process for capability estimation.

- Selecting the appropriate Process capability metric to use: This includes evaluation of different process capability indices and determination of the most appropriate metric for process development and biopharmaceutical manufacturing. This topic is covered under a separate chapter (chap 4) in the thesis.

- Statistical analysis to establish the relationship between Ppk and sample size. Different statistical approaches have been evaluated to determine how this relationship should be established.

- Development and testing of a framework to determine optimal sample size for Process Performance capability (Ppk) calculations.

II. **Evaluation of Measurement Variation**: Evaluate and propose methods to quantify measurement system variation, criteria for when measurement system variation is unacceptable, and approaches for filtering it from process capability calculations when the variation exceeds pre-defined thresholds.

- Understanding sources of measurement variation: Sources of measurement variation were evaluated for selected measurement systems.

- Selecting the appropriate measurement capability metric to use: Measurement capability metrics have been evaluated based on different industry use cases to determine the most appropriate metric for process development and biopharmaceutical manufacturing.

- Propose a procedure to filter out measurement system variation: A procedure has been established to separate measurement system variation from process variation. Case studies have been performed to test this process.

III. **"Big" Data Analytics**: Leverage additional data from different scales and platforms to supplement small data sets to improve the reliable estimation of Ppk

- Improve estimates of average & standard deviation when datasets are below an optimal sample size: Conduct statistical tests to evaluate if variances are similar. If variances are determined to be similar, Bayesian statistical approaches are being used to incorporate additional data sources in the estimation of Ppk values and credible[1] intervals around the metric.

[1] A credible interval is an interval in the domain of a posterior probability distribution or predictive distribution used

1.5 Thesis Outline

This thesis is organized into eight chapters and the content of each chapter is summarized below.

Chapter One describes the context of the problem statement, the objective and scope of this project and details how the research was conducted.

Chapter Two discusses the biotechnology industry, explains the manufacturing process of a biologic and the different stages of process validation, topics that are applicable to this thesis. and ends with an overview of Amgen.

Chapter Three gives an overview of the literature that was consulted to understand the different methods, statistical tools and metrics used for Process capability, analytical method variation and making predictions based on historical data.

Chapter Four lays the groundwork by explaining the importance of process capability, describing the difference between Ppk and Cpk and discusses which metric is suitable to be used for biopharmaceutical process development.

Chapter Five describes the work done to determine the optimal sample size for Ppk calculation-bootstrapping analysis and development of a tiered approach.

Chapter Six describes the analysis performed to evaluate and propose methods for filtering out variation from analytical method (assay).

Chapter Seven discusses "Big data" analytics and how Bayesian analysis can be used to estimate Ppk values in the future.

Chapter Eight closes the thesis with the suggested methodology, key recommendations and potential future work in each research area.

2.0 Biotechnology Industry and Organizational Analysis

2.1 What is Biotechnology?

Biotechnology is the use of living systems, organisms or derivatives to develop or make products. Biotechnology has many industrial applications including food, health care, agriculture, biofuel and other environmental applications[2]. In medicine, modern biotechnology finds applications in areas such as pharmaceutical drug discovery, production, pharmacogenomics, and genetic testing (or genetic screening). Biotechnology has contributed to the discovery and manufacturing of traditional small molecule pharmaceutical drugs as well as drugs that are the product of biotechnology - biopharmaceutics.

Biologics can include a wide range of products including, but not limited to vaccines, blood and blood components, allergenics, somatic cells, gene therapy, tissues, and recombinant therapeutic proteins [3]. Biologics are medicinal products produced from biological sources (human, animal or microorganism). Biologics can be composed of sugars, proteins, or nucleic acids or complex combinations of these substances, or may be living entities such as cells and tissues. Most biologics are given via parenteral (or non-oral) routes [4] such as intravenous therapy (IV), subcutaneous (SC) or intramuscular injection (IM)[5]. Biologics are also much more complicated to manufacture than the traditional small molecule drugs since biologics are produced through a living organism compared to drugs which are made through a series of chemical synthesis steps. Additionally, biologics have very complex structures and modifications required for their desired functions that may also impact efficacy or other quality attributes.

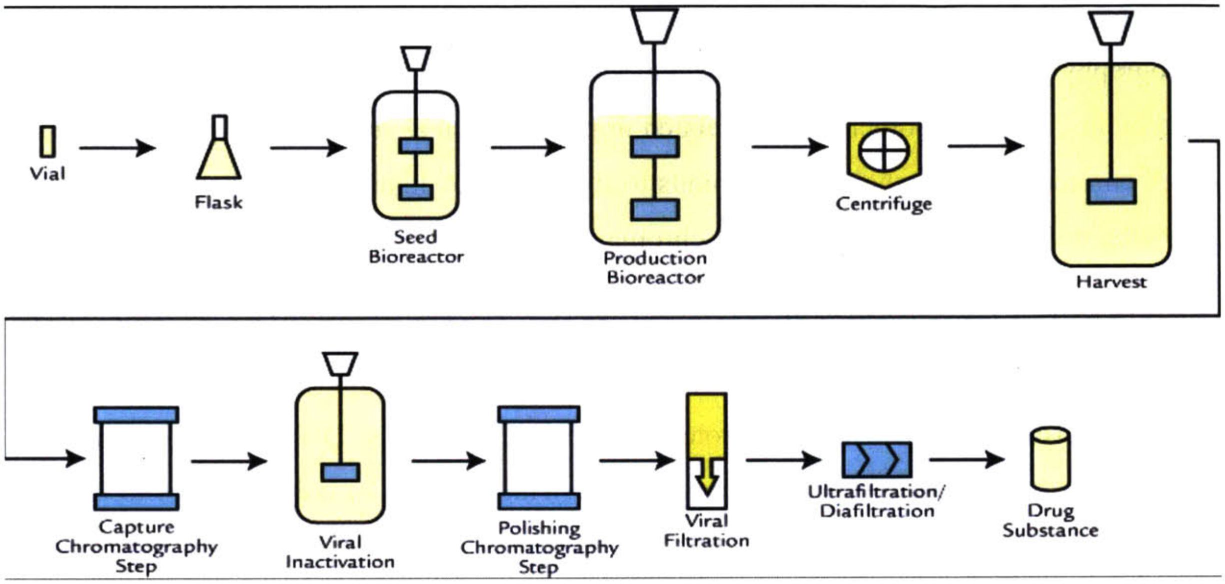
2.2 Recombinant Biologic Manufacturing Process

A recombinant biologic is a genetically engineered protein produced in living systems such as a microorganism (e.g. *E. coli*), plant or animal cells. Most biologics are very large, complex molecules or mixtures of molecules and are produced using recombinant DNA technology.

Recombinant monoclonal antibodies are a common form of biologic medicine and can be produced using Chinese hamster ovary (CHO) cells. The manufacturing process shown in Figure

1 is a typical process flow used for the commercial production of drug substance for a CHO based biologic.

Figure 1: Manufacturing process of a Biologic



As shown in Figure 1, the process can be split into upstream and downstream processes. Upstream processes consist of processes from thawing of cell bank vials with frozen, genetically modified CHO cells through cell expansion to harvesting of cells via centrifugation and filtration. Downstream processes consist of purification, filtration, formulation and filling drug substance into storage containers.

The upstream process begins when recombinant CHO cells that have been engineered to produce a target molecule are thawed. From the initial thaw, the cell culture is expanded through a series of flasks and expansion bioreactors to increase the cell mass. The expansion bioreactors can consist of bioreactors that mix the cell culture through a rocking motion or in stainless steel based stirred tank bioreactors. Disposable polymer based-bioreactors can also be used. The expanded culture is used to inoculate a production bioreactor, where protein expression occurs. Overall, the process from vial thaw to harvest can take up-to six weeks. The production bioreactor contents are typically harvested using centrifugation and filtration. The primary function of centrifugation is to remove host cells and cell debris from the cell containing production culture. The resulting centrate is further clarified using depth and membrane filtration.

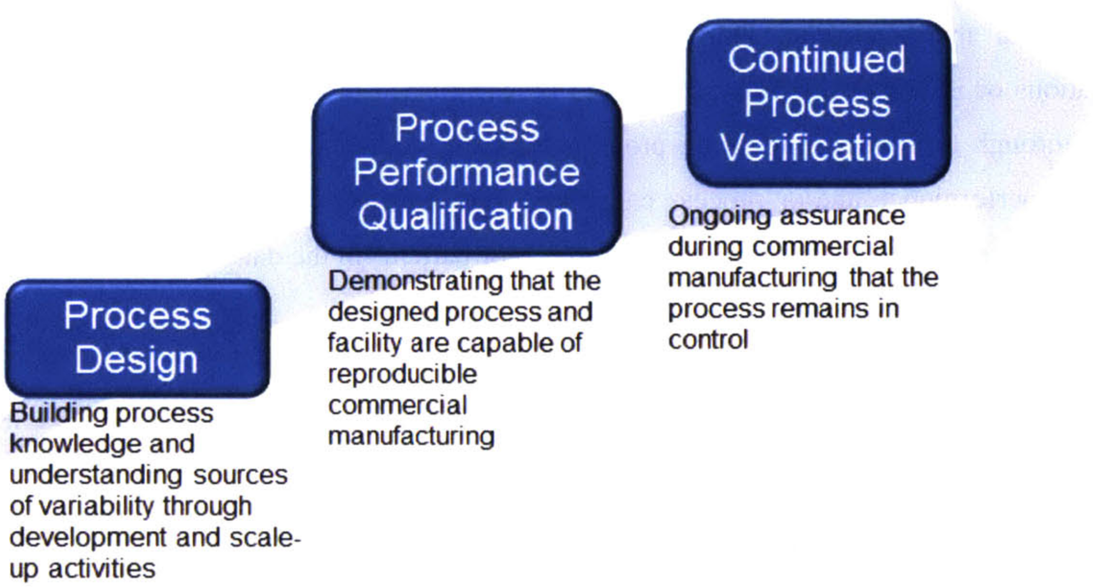
In the downstream process, the protein is purified from the harvest filtrate through a series of chromatography steps, a viral inactivation step, virus filtration, and formulation that typically utilizes tangential flow filtration. The first chromatography column typically utilizes an affinity chromatography column e.g. Protein A for antibodies. Protein A enables removal of host cell (CHO) proteins, DNA and other process related impurities such as medium components. During Protein A chromatography, the antibody binds to the Protein A ligand and is then eluted at low pH following multiple wash steps. Affinity chromatography can clear the majority of impurities.

Following initial chromatography, it is common to use a series of polishing chromatography steps, e.g. using ion exchange resins to specifically purify the desired product while clearing additional impurities, such as leached protein A, host cell proteins, DNA and other product-related impurities including higher order aggregates. This decision of column types to be used depends on the molecule or on specific company technology. The product is then concentrated and buffer exchanged into the Drug Substance formulation buffer using an ultrafiltration/diafiltration step. Finally, the Drug Substance is filtered into storage containers for long term storage (typically frozen).

2.3 Process Validation

Effective process validation contributes significantly to assuring drug quality. The Food and Drug Administration (FDA) requires that each step of a manufacturing process of a drug is controlled to assure that the finished product meets all quality attributes including specifications. The Process Validation effort ensures that a process, operated within established parameters, can perform effectively and reproducibly to produce an intermediate, active pharmaceutical ingredient (API), Drug substance, or Drug product, meeting predetermined specifications and quality attributes. Effective validation is achieved through a three stage Process Validation Lifecycle approach [6] as shown in Figure 2.

Figure 2: Three Stage Process Validation Lifecycle


Process Design
Building process knowledge and understanding sources of variability through development and scale-up activities
Process Performance Qualification
Demonstrating that the designed process and facility are capable of reproducible commercial manufacturing
Continued Process Verification
Ongoing assurance during commercial manufacturing that the process remains in control

Stage 1 is the process design phase, which is focused on designing the commercial process and the process control strategy to ensure product quality expectations are consistently met. In this stage, Product development activities and early process design experiments such as DOEs (Design of Experiments) are conducted to provide key input to the process design phase. Additionally, a strategy for establishing process controls is also set during this stage. Biopharmaceutical companies establish In-Process Control (IPC) parameters to monitor and control processes. These IPCs typically have acceptance criteria and/or action/rejection limits and are typically classified as critical or noncritical.

Stage 2 of the lifecycle approach to validation is process performance qualification (PPQ). PPQ combines the actual facility, utilities, equipment, and the trained personnel with the commercial scale manufacturing process, control strategy, and materials to produce commercial batches. A successful PPQ will confirm the process design and demonstrate that the commercial manufacturing process performs as expected in a reproducible manner. Success at this stage signals an important milestone in the product lifecycle as a manufacturer must successfully complete PPQ prior to gaining approval to manufacture and distribute product.

Stage 3 of the lifecycle approach to validation is CPV or the continued process verification stage. The goal of this stage is continued assurance that the process remains in control during long term commercial manufacturing. During this phase, a system for detecting unplanned process deviations or issues in the process needs to be established to assess process variability. This is done through a process monitoring program where process data is collected and evaluated for process performance and/or process capability. Based on the results, assessments are made to take action to correct or mitigate based on trends or patterns in the data.

2.4 Company Background

Amgen, previously known as Applied Molecular Genetics, is headquartered in Thousand Oaks, California. It is a multinational company founded in 1980 and has a presence in over 75 countries. Amgen's mission is to serves patients across the globe through the discovery, development, manufacturing and delivery of transformational medicines. With over 18,000 employees worldwide, the company has $20.1 Billion in revenue and spends approximately 20% of its revenue on R&D ($4.1 Billion) [7].

Amgen's products cover several therapeutic areas such as bone health, cardiovascular, oncology, inflammation, metabolic disorders, nephrology and neuroscience. Amgen has internally developed molecules such as Repatha® (evolocumab) , Xgeva® (denosumab) and has acquired products through external acquisition such as Enbrel® (etanercept) and Vectibix® (panitumumab).

Though a significant amount of Amgen's current drugs are monoclonal antibodies (mAb), Amgen's pipeline of drugs include 11 different [8] modalities including BiTe® (BiSpecific T Cell Engager) Antibody Constructs, fusion proteins, peptides, small molecules and others. Amgen is now also venturing into Biosimilars, which are biological molecules highly similar to an innovator product that has lost exclusivity. To commercialize its products, Amgen has to go through very strict regulatory approval process from US FDA (Food and Drug Administration) and other regulatory agencies around the world.

From an organizational structure standpoint, Amgen has a functional structure, which facilitates innovation and collaboration across teams. Key functions are led by Executive Vice Presidents (EVP), who report directly to the CEO, Bob Bradway. The five key functions represented by EVP's at Amgen are Operations, R&D, Global Commercial Operations, Finance, and Value Delivery. In addition, there are multiple Senior Vice Presidents (SVP), spanning functions from Business Development to Manufacturing.

3.0 Literature Review

A detailed literature review was carried out prior to developing the proposed techniques outlined within this thesis and prior to defining the recommendations to estimate process capability during the design phase of process development. This review encompasses literature to understand the different methods, statistical tools and metrics used to estimate process capability.

3.1 Resampling techniques: Bootstrapping

3.1.1 What is Bootstrapping and Why is it needed?

During analysis of data using a sample or samples, we have to estimate the uncertainty in the magnitude of an effect representing the relationship between predictor and dependent variables. This *uncertainty* refers to the fact that a different sample would give a different value for the magnitude [9]. Uncertainty is expressed as confidence limits or a confidence interval, representing the range of values within which we are reasonably certain the true magnitude of the relationship would fall. Here *True* refers to the value we would get if we had the luxury of a huge sample, and *reasonably certain* is the level of confidence, such as 90% or 99%. These confidence limits and probabilities are inferential statistics that help us make a probabilistic decision about the magnitude of the true effect.

Bootstrapping is an approach to generating the uncertainties (i.e. confidence limits and probabilities) about the true value of the effect or the statistic from a study of a sample. The idea behind the bootstrap is that the sample is an estimate of the population, so an estimate of the sampling distribution can be obtained by drawing many samples (with replacement).

Bootstrapping is the *only* approach when the sampling distribution is either not known or too difficult to quantify [9]. This feature makes this technique very applicable for small sample sizes where there is not enough data to make an accurate assessment on the distribution of data and interpret results statistically. Alternative and usual approaches include making assumptions on how the value of the effect statistics would vary, if we repeated the study again and again. These values make up the so-called *sampling distribution* of the statistic.

3.1.2 How do you conduct Bootstrapping analysis?

Bootstrapping is a numerical sampling technique where a dataset is first sampled with replacement randomly. Figuratively, this means that we pick a sample, place it back and select another sample several hundred or thousand times. The end result is the generation of many sample sets with each sample set having the same size as the original sample dataset. The value of a statistic is then calculated for each of the thousand or more sample datasets by treating each sample dataset as if it came from the repetition of the study. After this calculation, the sample distribution of the statistic is determined using the confidence limits given by appropriate percentiles of the values, and probabilities are determined by using the proportion of values falling above or below chosen magnitude thresholds. Note that the precision of the confidence interval is completely determined by the initial sample size.

In this manner, various descriptive statistics such as mean, median, mode, variance and correlation can be bootstrapped. Bootstrapping statistics allows the researcher to analyze any distribution and make inferences. The sampled data becomes the population and the resampled data are the samples.

3.1.3 Benefits and Challenges in using Bootstrapping

Below are some of the benefits as well as challenges in using bootstrapping.

Benefits include:

- Bootstrapping determines a confidence interval by generating a lot of possible samples so this can be very useful especially in the analysis of small data sets.
- No need to determine the underlying sampling distribution for any population quantity.
- Assumptions of normality and equality of variances for the population are not required to infer the results from bootstrapping [10][11].

Disadvantages include:

- Languages such as R or SAS might be required for conducting bootstrapping if the sample data is very large and Excel cannot handle.
- Randomness generated from the technique must be understood

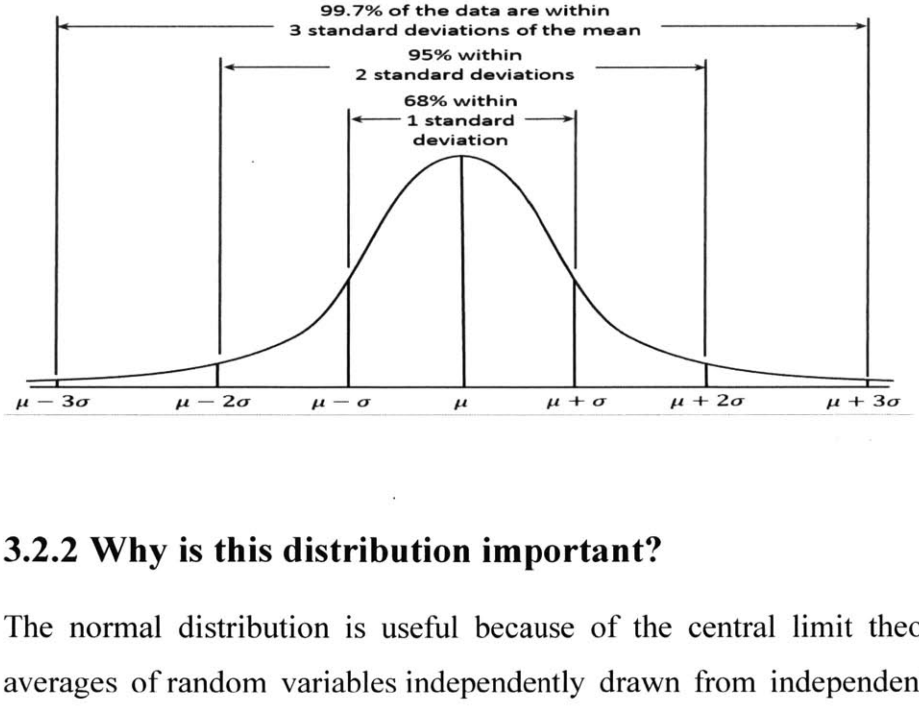
- Need to have at least 8 or more samples to be bootstrapped in order to have a rich enough estimate of the confidence interval [12].

3.2 Gaussian Distribution

3.2.1 What is Gaussian distribution?

The normal distribution, also known as bell curve, is the most important and most widely used distribution in statistics. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss. Normal distributions are defined by two parameters, mean (μ) and standard deviation (σ) and are symmetric around their mean. As we can see from Fig 3, about 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This fact is known as the **68-95-99.7 (empirical) rule, or the *3-sigma rule*.** The mean, median, and mode of a normal distribution are equal to each other and can be used as a measure of central tendency.

Figure 3: Normal Distribution



3.2.2 Why is this distribution important?

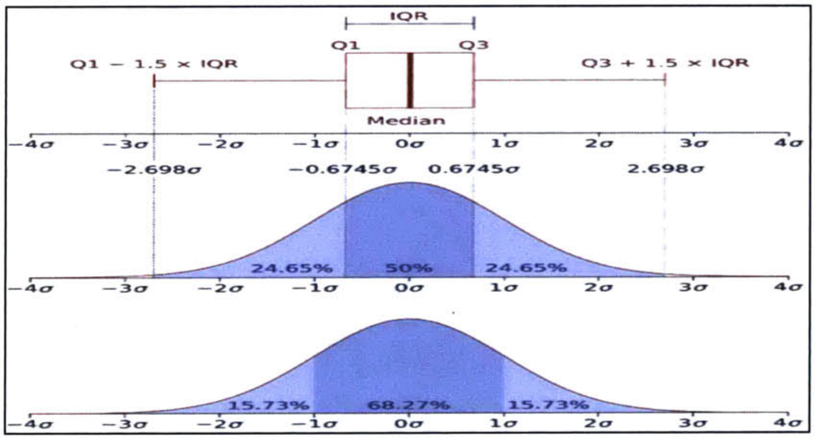
The normal distribution is useful because of the central limit theorem, according to which, averages of random variables independently drawn from independent distributions converge in distribution to the normal i.e. become normally distributed when the number of random variables is sufficiently large. Additionally, most of the statistical tests such as t-tests, linear regression, analysis of variance etc. assume that data is normally distributed. Therefore, it is important to understand whether or not the data is normally distributed before we begin using these tests otherwise we run the risk of getting inaccurate results.

3.3 Outliers and Statistical control

3.3.1 What is an outlier?

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Usually, this abnormal distance is specified as more than 1.5 times the interquartile range[2] above the third quartile or below the first quartile, which is more than +/- 2.698 σ (standard deviation) from the center i.e. mean or median in a normal distribution) as shown in the below Figure 4.[13]

Figure 4:Probability Density function of a Normal distribution



3.3.2 What kind of outlier tests are there?

Test for normal distribution of a data set can be done in the following ways:

- Review the distribution graphically (via histograms, boxplots, normal quantile plots)
- Analyze the skewness and kurtosis

[2] The interquartile range (IQR) is a measure of variability or statistical dispersion, based on dividing a data set into quartiles, which divide a rank-ordered data set into four equal parts.

- Employ statistical tests (esp. Shapiro Wilk, Chi-square, Kolmogorov-Smironov, Jarque-Barre, D'Agostino-Pearson)

3.3.3 Which statistical test is the best and what does it measure?

Shapiro-Wilk test (also known as the goodness of fit test) has been shown to be more powerful than alternative tests (such as Anderson Darling or Kolmogorov-Smironov tests) to check the distribution for a departure from a normal distribution [14]. The test checks whether or not our data is normally distributed as it checks the distribution of the data set that we provide against normally distributed data. The final result is given in the form of a p or a probability value. If the P value is less than 0.05, the null hypothesis that there is no difference between the data distribution and normal distribution is rejected. However, if the P value is greater than 0.05, we conclude that our data is normally distributed and ready to be used for further statistical tests.

Note that all such tests have low power with sample sizes less than 30, which are considered small for statistical analysis. Therefore, it is helpful to look at combinations of few different methods to evaluate outliers for small sample sizes. These methods could include analyzing the data graphically through outlier box plots and normal quantile plots as well as a goodness of fit test. This is the strategy employed for this project and it was very effective to help identify outliers. Note that outliers are an indication of non-normality but not a proof as the data might inherently have a non-normal distribution.

Lastly, a non-normal distribution can be transformed to a normal distribution through several statistical techniques if needed. We decided not to do the transformation for our project to avoid the overly complex situation of using logarithmic values for Ppk calculation that could have affected the accuracy of the results as well.

3.3.4 What is Statistical Control?

There are two common sources of variation within a process: common cause and special cause variation. Common cause variations are the many ever-present factors (i.e. process inputs or conditions such as common input material variation, equipment variation) that contribute in varying degree to relatively small, apparently random shifts in outcomes in the short term and long term [15]. This variation is inherent to a system and it is usually difficult, if not impossible, to link random, common cause variation to any particular source. Special cause variations, on the other hand, are factors that are induced by special effects not always present or built into the process. Some examples include: uncontrolled environmental factors, power surges, human error etc. Special causes are often referred to as assignable causes because the variation produced can be tracked down and assigned to an identifiable source.

A process is in statistical control or considered stable if only random variation or common cause variation is present. Control charts, which represent a picture of the process over time, are one of the most popular SPC (Statistical Process Control) tools used by manufacturers to determine whether a process is in a state of statistical control. Control charts include control limits, which are set at +/-3 standard deviations of the plotted statistic from the statistic's mean and they are used to detect signals in process data that indicate if a process is not in control and, therefore, not operating predictably. If the process goes outside the control limits or has a potential to fail these limits, the process will be considered out of control. This guidance is covered under a set of rules used across the industry such as Western Electric, which will be covered in Section 3.6.3a. Note that in a normal process, 3 standard deviations in the long term (assuming a 1.5 sigma shift) represent 99.87% values within range as shown in table 1 so 0.14% values will "naturally" fail the control limits in the long term.

It is also important to note that a process not in statistical control can still be within predefined specifications. Specifications and control limits are very different limits-Specification limits represent the voice of the customer and control limits represent the voice of the process. As mentioned above, control limits are calculated as +/- 3 standard deviations from the mean of the plotted statistic while specifications are defined based on patient requirements.

3.4 Relationship between Ppk and Yield

Process capability analysis is a prominent technique to predict how well a process will operate within its specification limits. This analysis is based on a sample of data taken from a process and often produces: an estimate of the dpmo (defects per million opportunities), one or more capability indices, an estimate of the sigma quality level at which the process operates and the yield or the % of values within the spec limits [16].

The relationship between Ppk, Sigma, DPMO and yield is given below for sigma levels from one through six[17].

Table 1: Relationship between Ppk, Sigma level, DPMO and yield

Ppk Value	Short term Sigma σst	Long term sigma σlt (assumes 1.5 sigma shift)	DPMO	% of total values within range	% of total values outside range
0.33	1	-0.5	690,000	31.00%	69%
0.667	2	0.5	308,537	69.14%	30.86%
1	3	1.5	66,807	93.319%	6.68%
1.33	4	2.5	6,210	99.379%	0.621%
1.5	4.5	3.0	1350	99.865%	0.135%
1.667	5	3.5	233	99.98%	0.02%
2	6	4.5	3.4	99.99966%	0.00034%

The reporting convention for sigma is short term so process capability is also reported in short term since it is difficult to predict how the process will change in the long term. This is the reason 1.5 sigma shift is assumed to account for special cause variations as we move from short term to long term[18]. As we can see from table 1, sigma level is decreased by 1.5 in the long term to represent the sigma drop that will fit between the process mean and the nearest specification limit over time, compared to an initial short-term study. For example a process that fits 6 sigma in the short term will typically fit to 4.5 sigma in the long term.

All of these outputs (Ppk, Sigma level, DPMO and yield) are inter-related and represent process capability. These inter-dependencies are explained below in detail.

First, for a process, based on the provided spec limits and the data, we calculate standard deviation and mean. These values can then be used in the below formula to calculate Ppk.

$$Ppk = \min[\frac{(USL - \overline{X})}{3 * \sigma}, \frac{(\overline{X} - LSL)}{3 * \sigma}]$$

Now we need to understand, what the Ppk value will tell us. It will help us understand 2 things- First, It will help us understand how close mean is to the center of the upper and lower spec limit represented by $USL - \overline{X}$and $\overline{X} - LSL$ respectively. Second, it will tell us the spread of the data based on σ in the denominator. If the Ppk value is 1, it basically means that the mean is 3 standard deviations away from the spec limits. With Ppk of 2, we have a 6 sigma process, which helps us understand how well a process is performing with respect to its specifications. These values are shown in Table 1.

Let's now discuss the DPMO or Defects per million opportunities.

Defects per million opportunities (DPMO) or Non conformities per million opportunities (NPMO) is a measure of process performance and is given by the below formula

$$DPMO = \frac{1,000,000 \times \text{number of defects}}{\text{number of units} \times \text{number of opportunities per unit}}$$

Here, a defect is defined as a nonconformance of a quality characteristic (e.g. strength, width, response time) to its specification and an opportunity is the total quantity of chances for defect. Highly capable processes experience very few defects per million units produced. Note that the DPMO can be drastically improved by a relatively small improvement in the capability index as shown in Table 1.

The relationship between Sigma and DPMO is calculated using the inverse of the cumulative probability distribution function[3]. This calculation is done in excel as NORMSINV(1-Defect Per Opportunity) + 1.5 (accounts for 1.5 sigma shift in the long term) [19]. The sigma level is raised here to calculate sigma level in the short term.

[3] A normal random variable is continuous so the Cumulative distribution is the integral of the density

Note that the defects Per Million Opportunities are put in the formula as Defect per opportunity. For e.g. 69000 defects per Million Opportunities will be entered as 0.69 Defects per opportunity.

Based on the above formula, the sigma level is calculated for the below defects.

Table 2: Relationship between Sigma level and Defects

Sigma	Defects Per Million Opportunities
1	690,000
2	308,537
3	66,807
4	6,210
5	233
6	3.4

Last, we move on to Yield, which is another measure of the quality level and can be understood as classic yield, first pass yield and Rolled throughput yield. Classic Yield (YC) is the ratio of units passed and final units tested, First time yield is the ratio of the units passed and units input for first time and Rolled Throughput Yield (Ytp) is a multiplication of different yields to calculate a final yield.

The relationship between Yield and DPMO is as follows [20]:

Yield= 1- (DPMO / 10^6)

Therefore, once we have a Ppk value, we can use it to get an understanding of a lot of different measures- sigma level, DPMO and % of total values within range. In general, a lot of companies strive for achieving 6 sigma level in the short term or 4.5 sigma in the long term (taking into account the 1.5 sigma shift in the long term) as 6 sigma level in short term it means that 99.9996% values are within range for normally distributed data.

3.5 Measurement variation

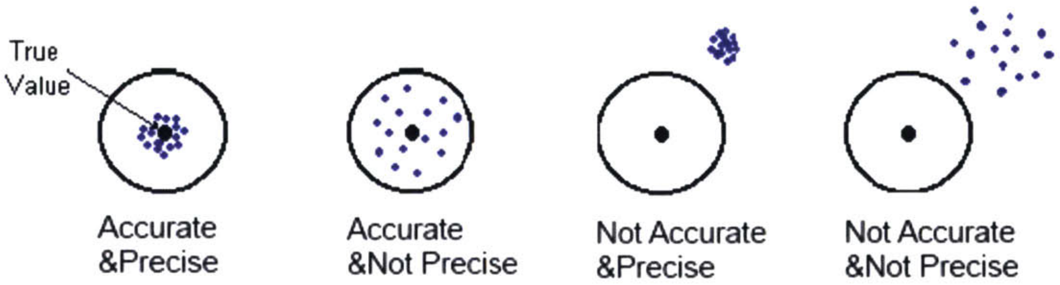
3.5.1 Types of Method or measurement system variation

In order to understand method variation, we first need to understand overall process variation. Overall variation in a process can be broken down in two parts: Actual Process Variation and Measurement Variation. Both these variations are independent of each other. This relationship is presented in the below formula where variation is represented by σ^2 or variance.

$$\sigma^2_{\text{Observed Process}} = \sigma^2_{\text{Actual Process}} + \sigma^2_{\text{Measurement System}}$$

The actual process variation consists of variation in the short term as well as in the long term to a certain extent. The measurement variation, which we are interested in understanding, can be classified into two categories: Accuracy and Precision. *Accuracy* refers to how close measurements are to the "true" value and characterizes the location of the measurement, while *Precision* refers to how close measurements are to each other and characterizes the actual variation[4] in the measurement system. Figure 5 demonstrates this difference between two categories through different systems [21].

Figure 5: Systems with different Accuracy and Precision



The accuracy of a measurement system can be understood by looking at 3 factors [21]: Bias, which is a measure of the distance between the average value of measurements and the true or

[4] The actual variation refers to the spread of the measured values whereas the overall variation mentioned earlier includes location of the values, also referred to as accuracy of the measurement system and the spread.

actual value of the sample; Linearity, which measures the consistency of bias (i.e. linear or non linear) over the range of use of the measurement system and Stability, which refers to the capacity of a measurement system to produce same values over time for the same sample. This also means the absence of special cause variation and the presence of only common cause variation (random variation).

A measurement system can have both accuracy and precision issues[5]. Generally, precision is the principal concern and measurement errors or variation measures the degree of precision [22]. Inaccuracy issues due to linearity can be typically be corrected through calibration. We will also be looking at Precision for this project to measure and filter out method variation from overall process variation.

3.5.2 Measuring Precision

Precision of a measurement system expresses the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of a homogeneous sample under a set of conditions. Precision is usually expressed as the variance, standard deviation or coefficient of variation, which is standard deviation divided by mean of a series of measurements. Precision can be considered at three levels: repeatability, intermediate precision and reproducibility.

Repeatability assesses whether the same appraiser can measure the same part/sample multiple times with the same measurement system over a short interval of time and get the same value [22]. Repeatability is also termed intra-assay precision and essentially provides us the variation due to the equipment itself with all operating conditions kept consistent.

Reproducibility conditions stipulate that the same method be conducted on identical test items in different laboratories, which necessarily involves different operators and equipment, as well as differences in other factors such as laboratory environment, management and quality control

[5] Together, Accuracy and Precision define the quality or the integrity of the measurement system. Accuracy helps us understand the location of measurements and Precision helps us understand the spread of the measured values

policies. Therefore, all measurement factors (working environment, appraiser, and apparatus) are changed to measure reproducibility.

Intermediate precision is determined under conditions that are intermediate between repeatability and reproducibility conditions, which represent the two extreme conditions for determining test method precision. It assesses whether different appraisers can measure the same part/sample with the same measurement system and get the same value in the **same lab or working environment**.

Table 3 lays out these differences and conditions for measuring precision [23].

Table 3: Conditions for Precision

Factors in Measurements	Repeatability Condition	Intermediate Precision Condition	Reproducibility Condition
Laboratory	Same	Same	Different
Operator	Same	Different	Different
Equipment	Same	Same[a]	Different
Time between Tests[6]	Short[b]	Multiple Days	Not Specified

[a]This situation can be different instruments meeting the same design requirement.

[b]Standard test method dependent, typically does not exceed one day.

[6] The conditions listed above for time between tests are observed in general. Overall, the time between test depends on the speed with which the process varies.

3.6 Measurement capability indices - P/T, %R&R and ICC

There are many indices to understand the relative utility or the capability of a measurement system. In this section, we will be covering some of the well-known indices such as precision to tolerance (P/T), %R&R and a very useful but less known index- Interclass Correlation Coefficient (ICC).

3.6.1 Measurement capability index – Precision/Tolerance Ratio

Precision to tolerance ratio is a very common measurement index and addresses what *percent of the tolerance or the difference between spec limits* is taken up by measurement error. The formula is given below

$$P/T = \frac{6 * \hat{\sigma}_{MS}}{Tolerance}$$

$$= \frac{6 * \hat{\sigma}_{MS}}{USL - LSL}$$

Note: Here, sigma hat represents the standard deviation of a sample data set.

The number 6 in the formula is an industry standard and denotes the standard deviations, which accounts for 99% of Measurement System (MS) variation given as $\hat{\sigma}_{MS}$ [7]. Prior to the 1990s, the number 5.15 was used rather than the number 6.00. Based on the guidelines, good measurement systems will have a *P/T* ratio that is less than 0.10, while adequate measurement systems will have a *P/T* ratio that is less than 0.30.

P/T ratio is also related to capability ratio. P/T ratio is an inverse of the capability ratio which is given by

$$\frac{USL - LSL}{6 * \hat{\sigma}}$$

[7] The techniques for finding the measurement system variation will be covered in sections 6.2 and 6.3

3.6.2 Measurement capability index – % R&R

%R&R is another very common measurement index and addresses what *percent of observed process variation* is taken up by measurement error. The formula is given below

$$\%R\,\&\,R = \frac{\hat{\sigma}_{MS}}{\hat{\sigma}_{Observed\,\Pr ocess\,Variation}}\,x\,100$$

Note: Here, sigma hat represents the standard deviation of a sample data set.

As a target, a good and capable measurement system has a %R&R less than 30%. There are several limitations with the %R&R ratio and the Gauge R&R study conducted to determine this ratio. Though the study starts out with a sound strategy for collecting data to conduct experiments (DOE), it overstates the impact of measurement error since the guidelines used to interpret these inflated ratios are very conservative. For example, the Gauge R&R study recommends that we need to modify our measurement system if the %R&R i.e. ratio of measurement variation and overall process variation is greater than 30%. There is neither an explanation given for this nor a logical solution given besides changing the measurement system, which might not be necessarily required or possible for many industries. Additionally, research studies done by noted Statistician Donald Wheeler have found out that the calculated ratios from the study are made to look like proportions when they really are trigonometric functions [24].

3.6.3a Western Electric Rules

Before we define the next and last measurement capability index i.e. ICC, let us first define the 4 Western Electric rules, which this index utilizes. The Western Electric rules were developed by the Western Electric company in 1950s and are widely used in statistical process control studies to assess process stability and detect any issues or out of control process by plotting the data on control charts. The rules, as listed below, interpret the points outside the three sigma limits as signals to detect process stability.

- Rule No.1 detects a signal when any single point plots outside the three-sigma limits.
- Rule No.2 detects a signal whenever two out of three successive values are on the same side of the central line on the control chart and are more than two sigma units away from the central line.
- Rule No.3 detects a signal whenever four out of five successive values are on the same side and are more than one sigma unit away from the central line.
- Rule No.4 detects a signal whenever eight successive values are on the same side of the central line.

These rules utilize historical data and look for a non-random pattern that can signify that the process is out of control, before reaching the normal +/-3 sigma limits. Therefore, these rules are extremely beneficial in understanding if a process is stable, though sometimes these rules can lead to false positives. For this reason, it is important to consider these rules together with other ratios to confirm our observations.

3.6.3b Intraclass Correlation Coefficient (ICC)

ICC is a traditional measure of relative utility introduced by Sir Ronald Fisher in 1921. ICC measures the proportion of variation in product measurements attributed to measurement system using well defined "Western Electric" to help understand process stability. The formula of ICC is given as:

$$ICC = (1 - \frac{\hat{\sigma}^2_{MS}}{\hat{\sigma}^2_{Observed\ \mathrm{Pr}ocess\ Variation}})$$

Note: Here, sigma hat represents the standard deviation of a sample data set.

As we can see, the ICC formula is similar to %R&R in that both are ratios of variation of measurement system and observed process variation. However, to measure variation, ICC uses variance rather than standard deviation. Variance is much easier to work with mathematically than standard deviations, which cannot be added or subtracted easily. Also, ICC lays out very clear guidelines to help us understand the capability of the measurement system based on the severity of the measurement issue.

The general ICC guideline is as follows: ICC>0.8 denotes minimal measurement variation due to the assay while ICC <0.2 means that we are working with an ineffective measurement system. When the ICC values are between 0.2 and 0.8, we can use the western electric rules to understand the capability of our measurement system as shown below [25]:

- *When the intraclass correlation exceeds 0.80, we are virtually certain to detect our three-sigma shift using rule No. 1 alone. In this case the measurement system will be said to provide a first-class monitor for the product being measured.*
- *When the intraclass correlation exceeds 0.50, we have better than a 90-percent chance of detecting our three-sigma shift using rule No. 1 alone. Therefore, when the intraclass correlation is between 0.80 and 0.50, the measurement system will be said to provide a second-class monitor.*
- *When the intraclass correlation exceeds 0.20, we have better than a 90-percent chance of detecting our three-sigma shift using rules No. 1, No. 2, No. 3, and No. 4 together. Therefore, when the intraclass correlation is between 0.50 and 0.20, the measurement system will be said to provide a third-class monitor.*
- *As the intraclass correlation falls below 0.20, the probabilities of detecting our three-sigma shift will rapidly vanish. When the intraclass correlation is below 0.20 the measurement system will be said to provide a fourth-class monitor.*

3.7 What is Statistical Hypothesis Testing?

A statistical hypothesis is a method of statistical inference to compare two statistical datasets or a data set obtained by sampling against a simulated data set. This is either to understand the effect of a variable or simply to understand if the two data sets are different statistically. An idealized null hypothesis is used in these tests, which means that there is no difference between the data sets. The alternate hypothesis proposes statistical significant difference.

To determine whether the difference between the two data sets is, in fact, statistically significant, a *p*-value, which is the probability of observing an effect given that the null hypothesis is true, is calculated. The null hypothesis is rejected if the *p*-value is less than the significance level of the test or α (which is usually set at 5% or 1%), thereby confirming that the datasets are statistically significantly different.

3.7.1 Different types of statistical inference tests

There are many different types of tests that can be used for hypothesis testing. The two most common hypothesis tests are t-test and one way ANOVA, which stands for Analysis of Variances or F test. Both these tests compare the mean of the two groups to determine statistical significant difference. ANOVA is preferred over a t-test because it can compare more than 2 groups effectively and because it has less chance of committing a type I error, which refers to incorrectly rejecting a null hypothesis (a false positive) [26]. However, ANOVA which is actually a misnomer, since it is not comparing variances, actually makes many assumptions before even conducting the test. One of them is that the variances of the two populations are same. This assumption is actually made by many other statistical tests including a type of t-test. Therefore, it is important to determine if the variances are in fact equal before performing the statistical test otherwise the results will not be accurate.

3.7.2 What is Homogeneity of Variance test? Why do we need it?

As we know, variance is the average squared deviations between a group of observations and their means and helps us understand the "spread" of the data. It is given by the below formula.

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - X_{avg})^2}{n-1}$$

Here, s represents the standard deviation of a sample.

It is important to test for homogeneity of variances for two key reasons: First, many tests which compare two datasets or analyze the impact of a change (such as t-tests or ANOVA) based on the means assume that the variances are equal as we saw above. Second, testing the homogeneity of variance is useful for testing meaningful hypothesis. For example, if we are comparing two datasets, which are very different in terms of their sources or other factors and trying to decide whether we can use them together, it will not make sense to compare their means. The means of these two very different datasets will more than likely be different. What we will need to compare is the variance of the spread of the data and is a reliable statistic to determine if the datasets are comparable.

3.7.3 Why use the Brown Forsythe Test?

There are four kinds of tests that can be done to test group variances. The first three tests of equal variances perform an analysis of variance on a new response variable constructed to measure the spread in each group. These tests are called *Brown-Forsythe, Levene* and *O'Brien test* [27]. The fourth test is *Bartlett's test*, which is similar to the likelihood-ratio test under normal distributions. Out of all these tests, Brown Forsythe test is most commonly used. We will be covering this test in detail and comparing it with the other tests.

The Brown and Forsythe Test **tests equal population variances.** It is a robust test based on the absolute differences within each group from the group median.

The Brown-Forsythe Test is relatively insensitive to departures from normality and it is not sensitive to skewed distributions (e.g., chi square or $\chi 2$) and extremely heavy-tailed distributions (e.g., Cauchy). In these cases, it is more robust than Bartlett's test, which requires a normal distribution and is sensitive to unequal sample sizes. Additionally, Brown Forsythe compares the datasets against median of the group unlike Levene's test, which uses the mean. Statistically, comparison against median is better especially for giving robust results for non-normal distributions. Therefore, Brown-Forsythe is much more commonly used and recommended over these tests.

To perform the Brown-Forsythe Test:

1. Calculate each $z_{ij} = |y_{ij} - y_{ei}|$ where y_{ei} is the median for the i th group, y_{ij} is the value in the dataset and z_{ij} represents the absolute differences within each group.

2. Run a one way Analysis of Variances test on the set of z_{ij} 's.

3. If p-value $\leq \alpha$, reject Ho or null hypothesis and conclude the variances are not all equal.

3.8 Bayesian Approach

3.8.1 What is Bayesian Approach?

Bayesian approach is a method based on Bayes theorem to statistically infer the true state of an event or a condition based on related conditions to the event. Bayesian analysis gives us information about unknown parameters by using probability statements. In Bayesian analysis, prior knowledge about model parameters is updated with the evidence from observed data to form the posterior distribution. Bayesian analysis uses this posterior distribution to form various summaries for the model parameters including point estimates such as posterior means, medians, percentiles, and interval estimates [28].

The relationship between posterior and prior distribution can be expressed as follows:

Posterior probability $\propto$ Likelihood $\times$ Prior probability.

Where, posterior density of the parameters is computed using Bayes theorem by multiplying the likelihood function of the data and the prior density of the parameters.

The interval estimates that define the posterior density or probability distribution are called credible intervals. Credible intervals are analogous to confidence intervals in frequentist statistics. Key difference is that Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value.

3.8.2 How do you conduct Bayesian analysis?

The first thing we need to be concerned with is obtaining the Prior [29]. Mathematically, Prior is written as $p(\theta)$, a probability distribution called “Prior distribution” and refers to the current knowledge about the parameters. This is usually developed based on historical information before any evidence is taken into account and forms informative prior. However, sometimes prior information is not available in which case, we want a prior with minimal influence on the inference [30]. We call such a prior a non-informative prior. Very commonly, a non-informative prior called Jeffrey’s prior is used in this case. Jeffrey’s prior takes the motivation from Fisher’s information $I(\theta)$, which is an indicator of the amount of information brought by the model (observations) about θ. As $I(\theta)$ becomes large, the influence of prior minimizes and thus, Jeffrey’s prior is very appropriate as a non-informative prior. It is proportional to the square root of the determinant of the Fisher information, $I(\theta)$ and is given by the below formula.

$$\pi(\theta) \propto I^{1/2}(\theta)$$

$I(\theta)$ or Fisher information and is given by the below formula

$$I(\theta) = E_\theta(\frac{\partial \log f(X \mid \theta)}{\partial \theta})^2$$

In case of informative prior, we use an inverse gamma function distribution [31] represented below.

$$IG(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-1-\alpha} \exp(\frac{-\beta}{x})$$

Here α is a shape parameter and β is a scale parameter. The larger the scale parameter, the more spread out the distribution is.

Moving on to the other factors in Bayesian analysis- when new data becomes available, the information of the model parameters is expressed in the "likelihood", which is proportional to the distribution of the observed data, given the model parameters. The likelihood is represented as p (y|θ) where y is the new data. This information is then combined with the Prior to produce an updated probability distribution called the "posterior distribution", on which all Bayesian inference is based. The expression then becomes [29]:

$$\boldsymbol{p(\theta \mid y) = \frac{p(\theta) \times p(y \mid \theta)}{\int_{\Theta} p(\theta) \times p(y \mid \theta)\, d\theta}.}$$

The denominator is termed the marginal likelihood or "model evidence" and can be expressed as an integral. It is the same for all possible hypotheses being considered and therefore, doesn't affect the relative probabilities of different hypothesis. The posterior distribution is more commonly written as follows:

$$p(\theta \mid y) = \frac{p(y \mid \theta) p(\theta)}{p(y)}$$

Bayesian inference computes the posterior probability according to Bayes' theorem, which is expressed as follows [32]:

$$\boldsymbol{P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}}$$

Bayes' theorem describes the probability of an event, based on conditions that might be related to the event and is a way to understand how the probability of an event is affected by a new evidence.

Here | denotes a conditional probability, A and B are events.

- *P*(*A*) and *P*(*B*) are the probabilities of events *A* and *B*
- *P*(*A* | *B*), a conditional probability, is the probability of observing event *A* given that *B* is true.
- *P*(*B* | *A*) is the probability of observing event *B* given that *A* is true.

Note that when Bayes approach is applied, A becomes the Hypothesis and B becomes Evidence.

3.8.3 Why use Bayesian analysis?

Bayesian analysis is a powerful and an important analytical tool for statistical modeling, interpretation of results, and prediction of data. Bayesian approach can be universal as it is based on a single rule of probability, the Bayes rule, which is applicable to all parametric models. This greatly facilitates its application. On the other hand, the more traditional frequentist approach relies on a variety of estimation methods designed for specific statistical problems and often inferential methods from one class of problems cannot be applied to another one[31].

Bayesian analysis also assumes that all model parameters are random quantities and thus can incorporate prior knowledge. We can therefore use prior information to get a more accurate response especially for unknown quantities. For example, we can use prior information to mitigate the effect of a small sample size. This assumption is in sharp contrast with the more traditional, also called frequentist, statistical inference where all parameters are considered unknown but fixed quantities. However, one needs to ensure that the prior distribution is accurate as possible so that we can get reliable results from the Posterior distribution.

4.0 Evaluating Process Capability Metrics

4.1 Why is Process Capability study important?

In any manufacturing process, the specifications or the tolerance limits based on customer requirements are developed to ensure that the product can meet the user requirements. The manufacturing process for the product must also meet or be able to achieve product specifications. Here is where the process capability study becomes critical. Process capability analysis is used to assess the following:

- Assess the potential capability of a process at a specific point or points in time in order to obtain values within a specification
- Predict the future potential of a process in order to create a value within specification with the use of meaningful metrics
- Identify improvement opportunities in the process by reducing or possibly eliminating sources of variability.

Understanding Process performance or capability is even more important for biologics where "the product is the process" [33]. A biologic is a very large complex molecule or set of complex molecules typically produced using recombinant DNA technology. It is critical that biologic manufacturers ensure product consistency, quality, and purity through control of the manufacturing process.

4.2 What is the difference between Ppk and Cpk?

Capability statistics such as Cpk, which is a statistical measure of process capability when the process is in control, and Ppk, which estimates the performance of a new process before a process is brought under control, are used to assess whether or not a process is capable of meeting user requirements. For our project, we have to decide whether to look at Cpk or Ppk for understanding process capability. The key assumption in conducting these calculations is that the data is normally distributed, which makes it possible to estimate the probability of an incident within any data set [34]. The most interesting values in Cpk and Ppk relate to the probability of data occurring outside of customer specifications, which is the data appearing below lower specification limit (LSL) or above upper specification limit (USL).

In order to understand the differences between Ppk and Cpk, we will first cover the formula of the two metrics, discuss what each of them measure and then talk about which metric has most utility during process development.

Let's first look at Cpk.

Cpk is given by the below formula:

$$Cpk = \min[\frac{(USL - \overline{X})}{3 * \hat{\sigma}}, \frac{(\overline{X} - LSL)}{3 * \hat{\sigma}}]$$

where *sigma hat* represents the standard deviation of the subgroups $\hat{\sigma} = \frac{\overline{R}}{d_2}$, *R-bar* is the average range (max-min) representing the average variation within the subgroup and d_2 is a control chart constant that depends on the subgroup size[35].

Each subgroup represents a group of measurements produced under the same set of conditions and hence, the snapshot of a process. Therefore, the measurements that make up a subgroup are most often taken from a similar point in time. For our analysis, one subgroup could be a single batch of a biologic as the batches are not made at the same time.

Moving on to Ppk, which is given by the below formula

$$Ppk = \min[\frac{(USL - \overline{X})}{3 * s}, \frac{(\overline{X} - LSL)}{3 * s}]$$

Where s is the standard deviation of the overall data and is calculated by the following formula

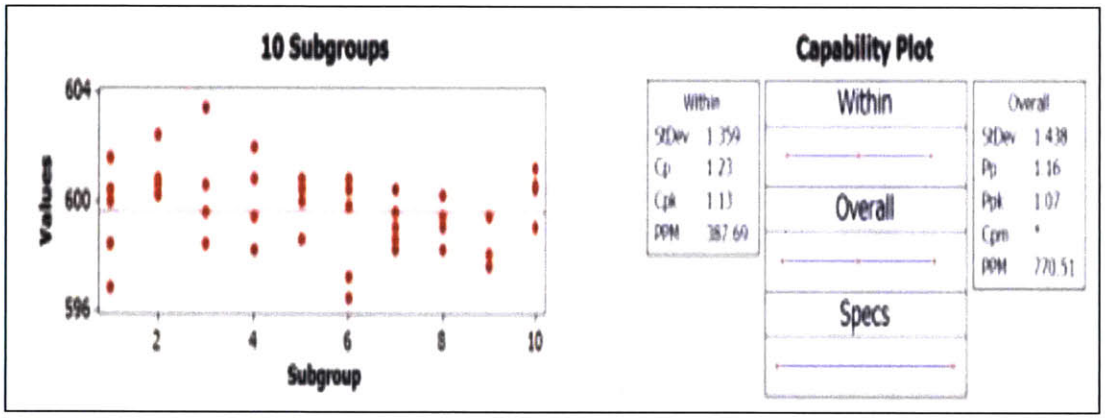
$$s = \sqrt{\frac{\Sigma (X_i - \overline{\overline{X}})^2}{N-1}}$$

Here, N is the number of data points, x bar is the mean of the data sets.

The numerators for Cpk and Ppk are exactly the same. The key difference lies in the method of estimating the statistical population standard deviation. $\hat{\sigma}$ in Cpk represents standard deviation of each subgroup in a sample whereas s in Ppk represents standard deviation of the overall sample dataset.

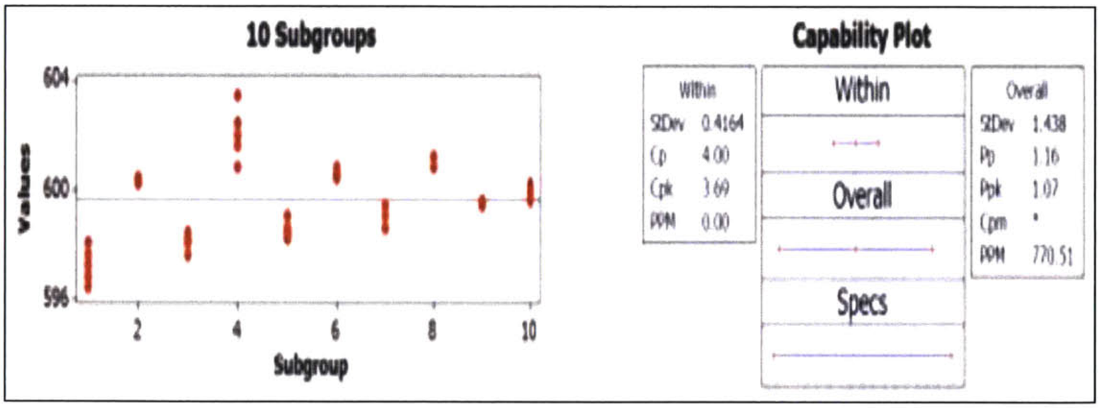
To further elaborate with an example [36], the below case represents 5 measurements taken every day for 10 days and shows the difference in calculation of Cpk and Ppk. Figure 6 shows Ppk and Cpk calculation when the variation within subgroups and between subgroups is similar while Fig 7 shows the calculation of the metrics when the variation between and within subgroups is different.

Figure 6: Similar Cpk and Ppk



When the variation between subgroups is similar to the variation within the subgroups themselves, 'within' and 'overall' standard deviations are similar, which means Cpk and Ppk are similar as well (at 1.13 and 1.07, respectively).

Figure 7: Different Cpk and Ppk



Compared to Fig 6, Figure 7 demonstrates the impact when there is a difference between the variation within subgroups and across subgroups. Where there is less variation within each subgroup we get a higher Cpk (3.69), however, we can see that the variation between the subgroups hasn't changed so Ppk remains 1.07.

Therefore, Cpk measures the short term capability (data within the same subgroup) and Ppk measures the long term capability of the process (data across different subgroups or batches)

Additionally, since Cpk considers standard deviation within a rational subgroup where samples are produced around the same time, the standard deviation is lower and this increases the Cpk value. On the other hand, the variation between subgroups allows the standard deviation in Ppk calculation to be higher, which results in more conservative estimates of Ppk. Hence, using the Ppk metric can also prevent us from overestimating the performance or capability of our process.

Last but not least, let's look at statistical control as a measure to evaluate the usage of the process capability and/or process performance metrics. Cpk is a process capability metric and is useful to measure only if the process is in statistical control so the mean and standard deviation of the data can be relied upon. As we know from Section 3.3.4, a process is in statistical control only when common cause variation is present [37]. This most often happens in a mature or stable process where special cause variation is absent. So, Cpk measures only common cause variation or variation present in subgroups, a set of values recorded at the same time.

Ppk, on the other hand, estimates the performance of usually a new process, which hasn't been brought in control yet. This means that the process includes both common cause and special cause variation. Therefore, Ppk is able to measure both types of variations, common cause present within subgroups and special cause variation, which is usually seen between subgroups, and is a very useful metric to assess new processes or processes for which, there is not sufficient data to determine exactly whether the process is in statistical control or not. Note: Cpk =Ppk when process is in statistical control.

4.3 Which metric should be used during Bio-manufacturing?

As we saw, Cpk measures the short term capability i.e. data within the same subgroup, which will mostly contain common cause variation, whereas Ppk measures the long term capability of the process i.e. data across different subgroups or batches, which will contain both common cause and special cause variation.

To manufacture a biologic, a lot of time and resources go into producing one single batch and the batches are not all manufactured at the same time. Additionally, there is a lot of interest within process development/biopharmaceutical manufacturing to understand the trends between different batches or subgroups i.e. overall variation containing special causes and common cause. The variation within subgroups or batches, which are practically produced at the same time is not as relevant for biologics, therefore, Ppk should be the metric of choice for biopharmaceutical manufacturing. Last but not least, Cpk is measured for a mature process that is in statistical control whereas Ppk doesn't require the process to be in statistical control. Therefore, for the purpose of this project, measuring Ppk is more suitable as the methodology being designed is primarily for processes that are still in development.

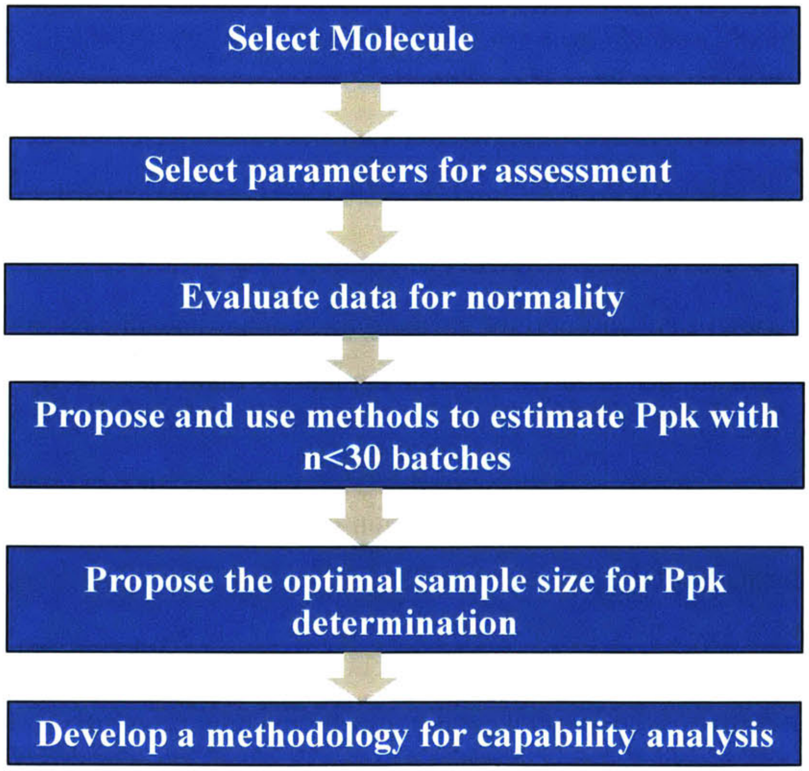
5.0 Establishing the relationship between Ppk and Sample Size

5.1 Methodology

The chart in Figure 8 shows the steps in a chronological order that have been followed to conduct the analysis for determining the optimal sample size for Ppk calculation. Note that the sample size refers to the batch size. For example, 1 sample point corresponds to data from 1 batch.

In this methodology, we first select the molecule and the parameters for assessment, evaluate data for each parameter for normality and test for outliers. This is followed by bootstrapping, which provides us confidence intervals. This analysis has been conducted for several parameters and based on the results, we recommend a multi-tier approach for determining the appropriate action based on the sample size available to use for calculations. In this chapter, each and every step of this methodology is covered in detail and the calculations within each step are also explained. Finally, the tiers are presented in section 5.5

Figure 8: Optimal Sample Size determination methodology


Select Molecule
Select parameters for assessment
Evaluate data for normality
Propose and use methods to estimate Ppk with n<30 batches
Propose the optimal sample size for Ppk determination
Develop a methodology for capability analysis

5.2 Molecule and Parameter Selection

The process selected for development of this methodology is representative of current manufacturing processes. The parameters for assessment have been selected from upstream and downstream unit operations and exhibited varying levels of complexity. These parameters include Seed bioreactor final viability (%), Column 1 step yield, Host cell protein after column 1, Final UF/DF protein concentration, Drug substance glycan content, Drug substance % relative potency and Drug substance pH.

Seed bioreactor final viability is an upstream parameter to measure consistent cell growth. The final viability is measured using an automated cell counter and is considered non-critical.

Process step yield at Column 1 is a downstream parameter to measure product recovery across a single unit operation. Percent step yield is defined as the ratio of the quantity of protein recovered from a unit operation (after sampling) relative to the quantity prior to that unit operation. It is considered non-critical and used to monitor process consistency. Step yield is calculated based on ultraviolet (UV) absorbance based measure of protein concentration.

Host cell protein (HCP) is a downstream parameter and quantifies host cell protein impurities generated as byproducts of the manufacturing process. **HCPs** are generated during cell culture and removed during purification.

Final UF/DF Protein Concentration: The UF/DF pool protein concentration is controlled through automation of the UF/DF step.

Glycan Content: Glycan content is an important attribute for biopharmaceuticals as glycans can impact efficacy and function of the product. Glycan content is quantified using hydrophilic interaction chromatography.

DS pH: pH is operationally controlled by the formulation of the buffer and through a targeted number of diafiltrations during UF/DF step.

Drug Substance % Relative Potency: In a biopharmaceutical process, Potency is a measure of drug activity expressed in terms of the amount required to produce an effect of given intensity.

Potency is a critical quality attribute and is assessed using cell based bioassays, which can determine the cȯncentration or biological activity of a substance by measuring an effect on tissues or cells.

5.3 Test for Normality and Outliers

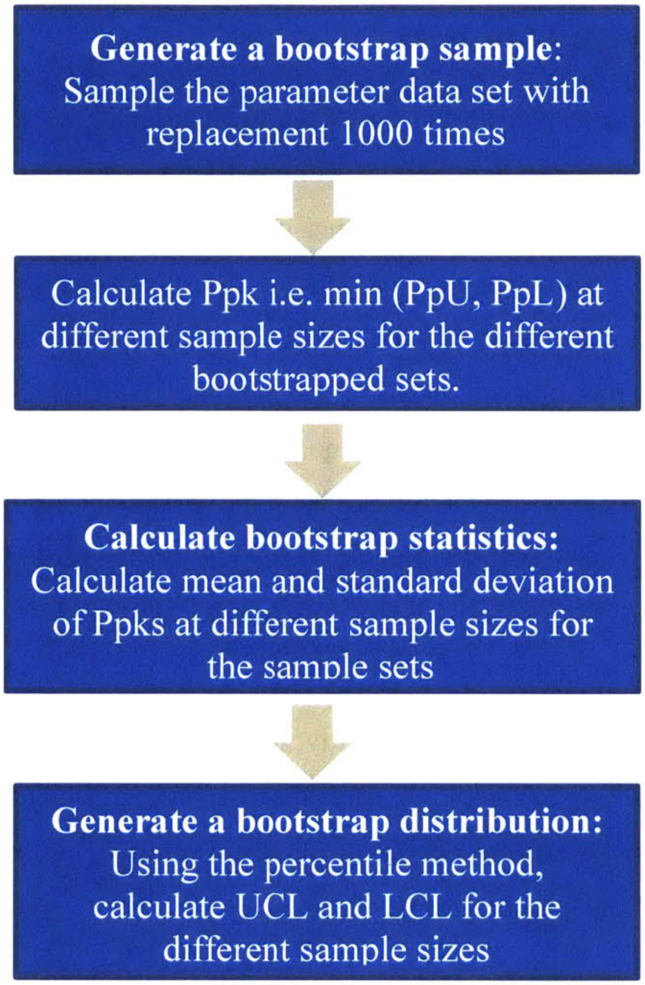
Since Ppk assumes that our data are normal, we must evaluate data for normality graphically using the outlier box plot and use the goodness of fit normality test. First, we look at the p value given by the goodness of fit test to determine whether it is greater or less than 0.05. A p value of less than 0.05 rejects the null hypothesis that the data are from a normal distribution. Therefore, if the p value is less than 0.05, it means that the data are not normal. We then evaluate the data graphically using the outlier box plot and check if the plot shows outlier data points. If it does, we look at the batch history of the outliers in the non- conformance system to determine if the outliers are caused by special cause in which case, we remove outliers. In case of no assignable event, we consult with the SME (subject matter expert) to understand if the outliers are a 'normal' or 'expected' part of the process even if infrequently observed. If the response is yes, we keep the 'outliers' in the data, otherwise we take them out.

As an example, in the case of seed bioreactor viability, there were 2 outliers and upon review, we established that these two outliers had a known special cause event associated with them (culture duration time had exceeded the normal operating range). Therefore, we removed the outliers from the analysis and passed the goodness of fit test.

5.4 Statistical analysis using Bootstrapping

Even when a distribution is known to be normal, bootstrapping is helpful in generating confidence intervals and helping us to understand the uncertainty about the true value of Ppk by incorporating the data from 1000 or more data sets. Figure 9 shows the steps in a chronological order that have been followed to conduct this analysis.

Figure 9: Bootstrap Methodology



Generate a bootstrap sample: Sample the parameter data set with replacement 1000 times

Calculate Ppk i.e. min (PpU, PpL) at different sample sizes for the different bootstrapped sets.

Calculate bootstrap statistics: Calculate mean and standard deviation of Ppks at different sample sizes for the sample sets

Generate a bootstrap distribution: Using the percentile method, calculate UCL and LCL for the different sample sizes

To elaborate, we first generate a bootstrap sample by taking a parameter data set of 100 batches and sample it with replacement 1000 times to produce 1000 data sets, each of which have the same size (100 batches) as the original data set. This gives us a sampling distribution. We then use the formula of Ppk (as listed below) to calculate Ppk for each sample size or batch size from 2 to 70 out of the data for 100 available batches. The Ppk calculation is done for each of the 1000 data sets at these sample sizes.

$$Ppk = \min[\frac{(USL - \overline{X})}{3 * \sigma}, \frac{(\overline{X} - LSL)}{3 * \sigma}]$$

We calculate bootstrap statistics for this data set i.e. the average and standard deviation of the Ppks at each sample size across all 1000 Ppk values. Once we have the required statistics, we have to generate a bootstrap distribution. We use the percentile method to generate this distribution. The data is sorted for each sample size and upper and lower 95% confidence intervals are calculated by picking top or bottom 2.5 percentile from the Ppk values. An example of this calculation is shown below for generating Upper confidence interval:

$$1000 \text{ data sets}*0.025 = 25^{th} \text{ value}$$

The Ppk value at 25th place (in ascending order) is chosen from the 1000 Ppk values generated for the data sets at each sample size.

The percentile method has been picked over other methods i.e. t or z distribution as this method yields a confidence interval irrespective of the distribution of the data and is appropriate for both large and small sizes. These features make the method very useful for our work on understanding relationship between Ppk and N at small sample sizes since there is not enough data to determine the most accurate distribution at small sample sizes, though we do check the data for normality.

5.4.1 Interpretation of results

The bootstrapping analysis has been conducted for all seven parameters across upstream and downstream processes. Below graphs show the results of Host Cell Protein after Column 1, one of the seven chosen parameters. The results of most other parameters are very similar.

Figure 10: Ppk and Sample Size variation for Host Cell Protein

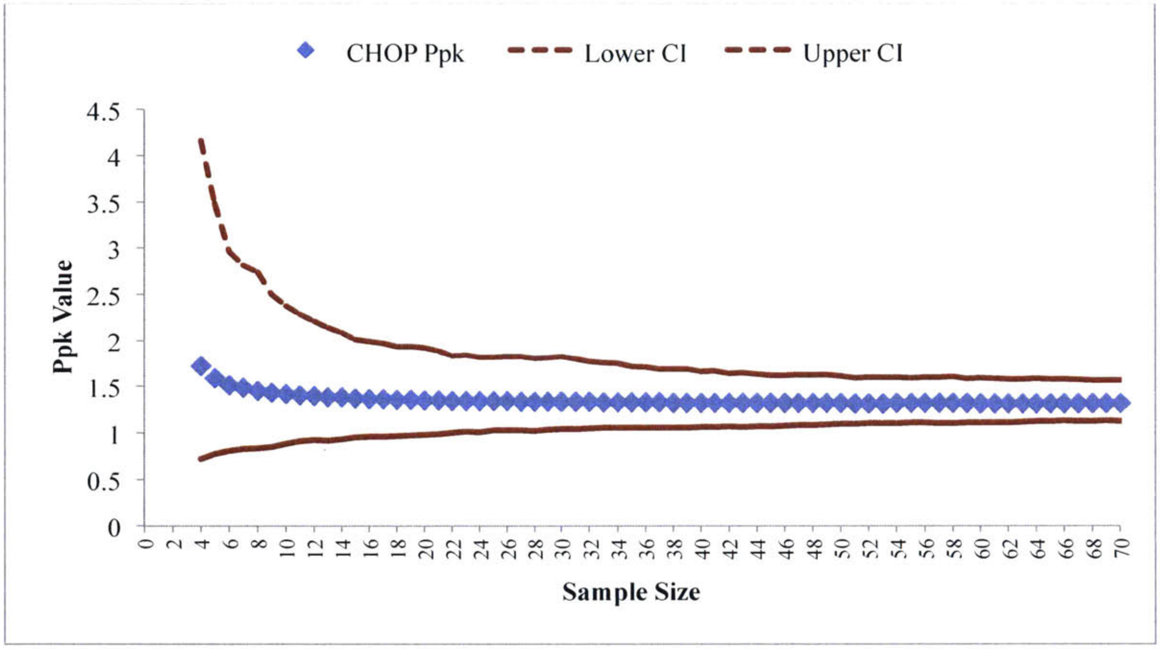
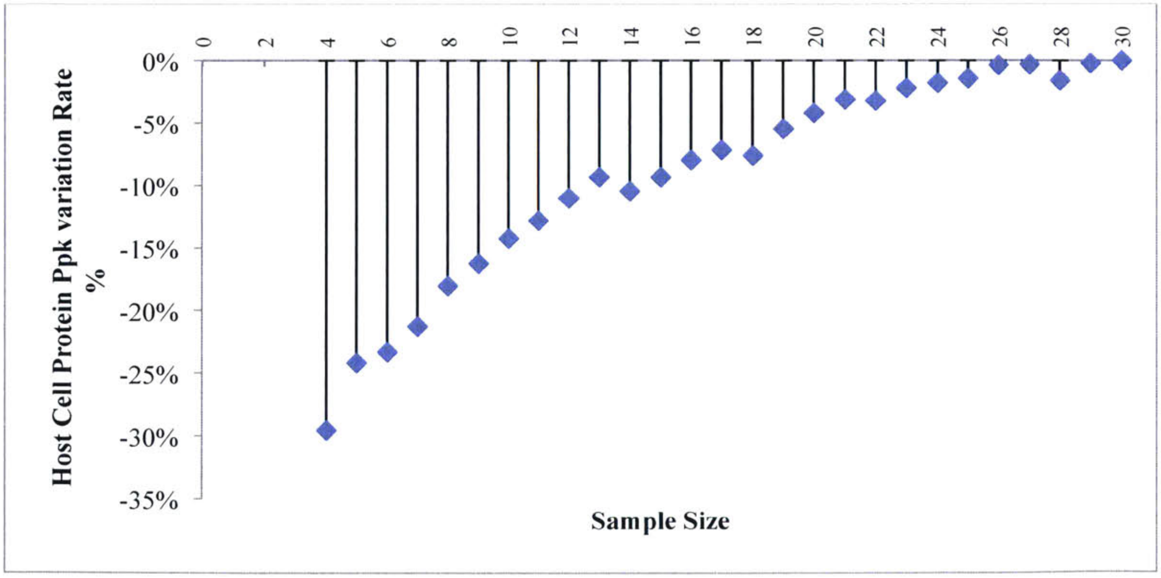


Figure 11: Rate of Host Cell Protein Ppk variation with sample size



The rate of Ppk variation shown above has been calculated by keeping Ppk at sample size (N) of 30 as a reference, which is also considered the gold standard for minimum amount of data required for statistical analysis[1] and is a typical threshold for when process capability should be evaluated.

As we can see from Figure 10, Ppk values start high at smaller sample size and become more stable as we increase the sample size. This makes intuitive sense as we have more data at a higher sample size, which brings more reliable or stable results. However, the key is to focus on rate of Ppk variation, which decreases significantly with increase in sample size as we can see from Figure 10. Additionally, as we look at Figure 11, we can make the following conclusions:

1. The rate of Ppk variation at sample sizes less than 5 is more than 30%,which is quite high.
2. Ppk values get more stable at higher sample size and the variation reduces to 18%- 25% at samples sizes between 5 and 8.
3. Rate of Ppk variation consistently reaches below 18% for sample sizes greater than 8.

Note: Rate of Ppk variation (18-30%) presented here takes into account the permitted 1.5 sigma shift [38] which occurs as we move from short term to long term due to special cause variations[8].

Most of the other parameters such as seed bioreactor final viability (%), Column 1 step yield, UF/DF protein concentration, Drug substance glycan content and Drug substance % relative potency show similar results.

However, for some parameters which are known to have a lot of variation in either the process or the assay, the reduction in the rate of variation of Ppk is seen to be slower (though not significantly different) than the reduction rate for the rest of the parameters as shown above. An example is Ppk for drug substance pH, which is known to have a lot of overall process variation mainly contributed by the measurement variation in its method and which will be covered in

[8] Reporting convention of Six Sigma requires the process capability to be reported in short-term sigma – without the presence of special cause variation. Long-term sigma is determined by adding 1.5 sigma from the short-term sigma calculation to account for the process shift that is known to occur over time.

detail in Chapter 6. The rate of Ppk variation for Drug Substance pH is depicted in the below Figures 12 and 13.

Figure 12: Ppk and Sample Size variation for DS pH

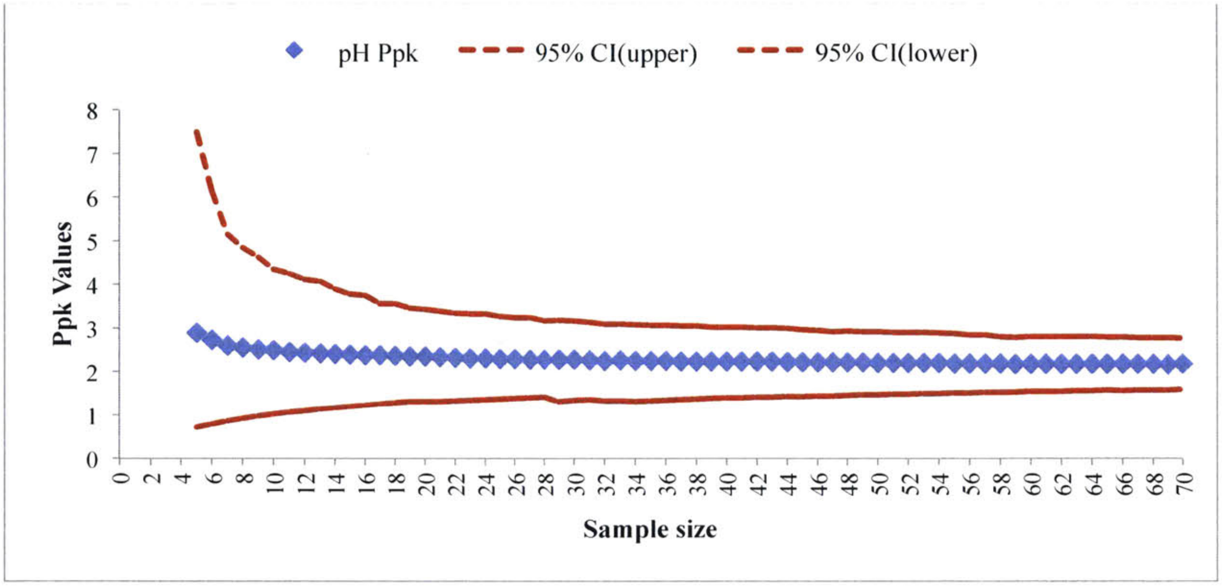
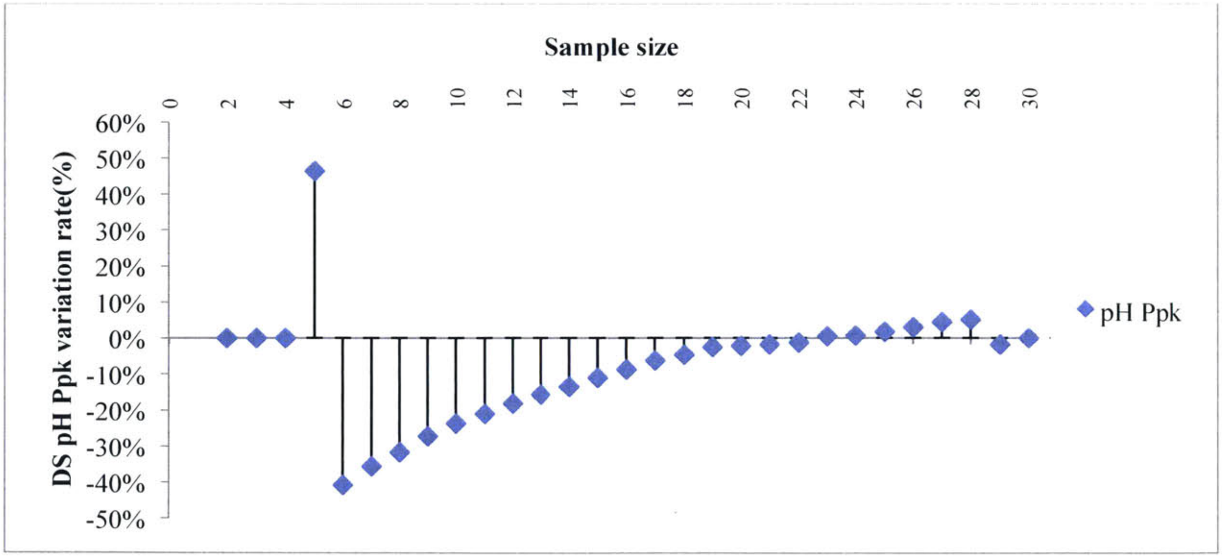


Figure 13: Rate of DS pH Ppk variation with sample size



Note: For sample sizes below 6, we can see that there is either no variation or positive compared to N=30. The reason is that bootstrapping is not giving a reliable value in this case for such small sample sizes as the resampled datasets are too similar too each other to produce a reliable standard deviation and in turn, a reliable confidence interval. This is further confirmed by the previous research done on bootstrapping as discussed in Section 3.3.1 which shows that bootstrapping results are reliable for sample sizes equal to and greater than 8 [39].

In summary, the rate of Ppk variation decreases significantly with increase in sample size. The key difference, though, is how the rate of Ppk variation decreases, particularly at small sample sizes. For example, as can be seen from Figure 13 the rate of Ppk variation is very high (more than 30%) for sample sizes less than 8. The variation reduces to under 30% (20-30%) for sample sizes between 8 and 11. Finally, Ppk variation becomes very stable and the rate consistently reaches below 20% for sample sizes greater than 11. Note the rate of Ppk variation presented here also takes into account the permitted 1.5 sigma shift as discussed in Section 3.4.

As we can see, the trend in variation of Ppk values is very similar to what was found earlier, except that the decrease in rate of Ppk variation is slightly slower in parameters, which are known to be more variable relative to spec ranges.

5.5 Selection of a Multi-tier approach

Based on the results from the analysis, we conclude that a multi-tiered approach is required for estimating Ppk at low sample sizes and determining the appropriate next steps. These steps include deciding whether to use the Ppk data as is or to calculate confidence intervals along with the Ppk value to understand the uncertainty in the Ppk value or to collect more data before conducting Ppk analysis. We also have to decide the right sample size for each tier taking into consideration the spectrum of complexity of parameters i.e. from parameters with low variation to parameters with very high variation and the reliability of the results from bootstrapping as mentioned in Section 5.4.1. In order to address this situation and make the final assessment of sample size parameter independent, the following tiers are recommended:

Tier 1 or Green Zone

Characteristic: Rate of Ppk variation is decreasing and is around <=15%-18% from Ppk of N=30 for all parameters including parameters with known significant variation. The Ppk values are getting stable.

Sample Size: Greater than 11

Recommendation: The data can be used for statistical analysis and for process capability calculations without any concerns

Tier 2 or Yellow Zone

Characteristic: Moderate variation in Ppk values. Ppk variation <=22%-30% from Ppk of N=30 for all parameters including parameters with known significant variation.

Sample Size: 8-11 for all parameters

Recommendation: Use the data for process capability analysis but include the uncertainty in the Ppk values defined by the confidence intervals.

Tier 3 or Red Zone

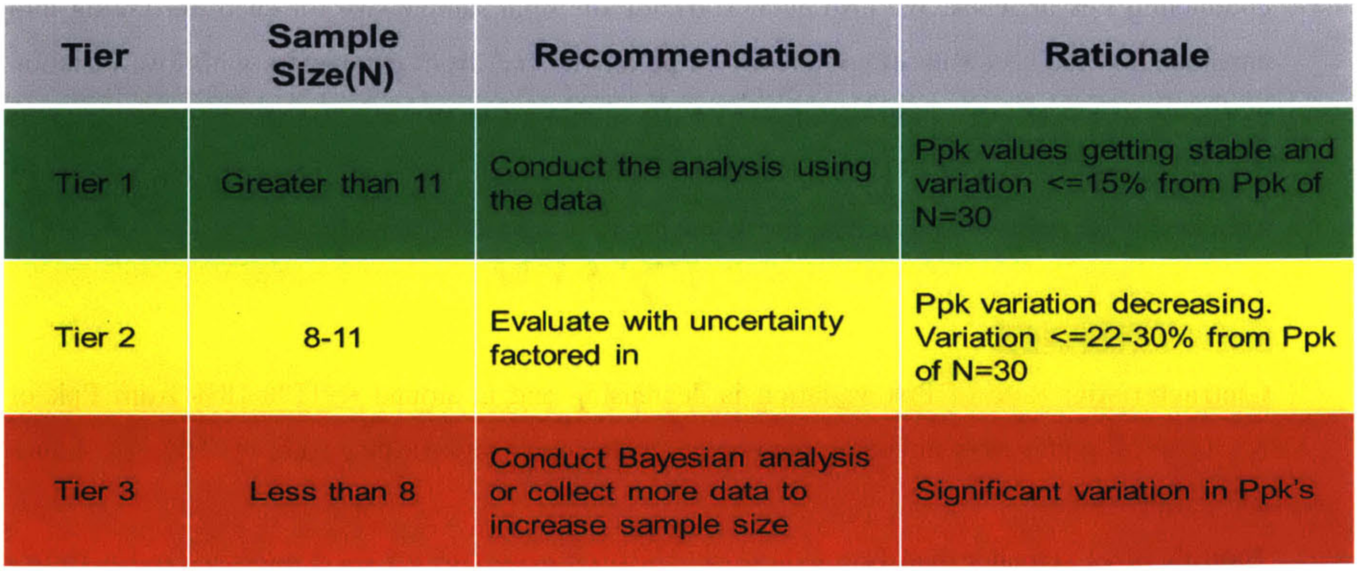
Characteristic: There is significant variation in Ppk values and Ppk is unreliable because of very few data points. Ppk variation >30%-40% from Ppk of N=30 for all parameters including parameters with known significant variation.

Sample Size: Less than 8

Recommendation: Collect more data or use data from different scales (lab scale or commercial scale) or other molecules before conducting statistical analysis. These techniques are presented as Bayesian approach in Chapter 7.

The tiers are also summarized in the table below.

Table 4: Tiered approach for Optimal Sample Size determination



Tier	Sample Size(N)	Recommendation	Rationale
Tier 1	Greater than 11	Conduct the analysis using the data	Ppk values getting stable and variation <=15% from Ppk of N=30
Tier 2	8-11	Evaluate with uncertainty factored in	Ppk variation decreasing. Variation <=22-30% from Ppk of N=30
Tier 3	Less than 8	Conduct Bayesian analysis or collect more data to increase sample size	Significant variation in Ppk's

6.0 Measuring Method or Assay Variation

6.1 Measuring Precision

Section 3.5.1 discusses that a measurement system is characterized by the location of its measured data, which detects the gap between the measured values and true value and the variation between the different measurements. Within location, *Accuracy* is measured to understand how close measurements are to the "true" value and within variation, *Precision* is measured to assess how close measurements are to each other.

We will be focusing on Precision for this project as we are trying to understand the variation generated by the methods. Precision is commonly understood at three levels-Repeatability, Intermediate Precision and Reproducibility. These terms are explained in section 3.5. Precision is often expressed as coefficient of variation, which is the ratio of standard deviation and mean of a set of observations. To understand the measurement variation due to equipment and the appraiser, Precision is measured by doing repeatability and intermediate precision studies, respectively. The working environment or lab is often not changed in these cases.

6.2 Formulae for Measurement variation

Overall Process variation consists of two parts: Actual Process Variation and Measurement Variation. This relationship is presented in the following manner:

$$\sigma^2_{\text{Observed Process}} = \sigma^2_{\text{Actual Process}} + \sigma^2_{\text{Measurement System}}$$

Variation is represented by variance or σ^2

We are interested in understanding the variation arising from the measurement system, which is a sum of the variation from repeatability and reproducibility. Repeatability provides variation due to gauge and reproducibility measures variation due to many different factors such as appraiser, working environment and gauge. The relationship between variation from measurement system, repeatability and reproducibility can be expressed below [19]:

$$\sigma^2_{\text{Measurement System}} = \sigma^2_{\text{Repeatability}} + \sigma^2_{\text{Reproducibility}}$$

When intermediate precision is measured instead of reproducibility, the equation changes to the below. This calculation helps us understand the variation created by the gauge and appraiser. The lab or the working environment is not being changed in testing the parameters.

$$\sigma^2_{\text{Measurement System}} = \sigma^2_{\text{Repeatability}} + \sigma^2_{\text{Intermediate Precision}}$$

6.3 Which measurement capability index is appropriate for Bio-manufacturing?

As we have seen in Chapter 3, P/T Ratio measures the percent of the tolerance being taken up by measurement error and is given by

$$P/T = \frac{6 * \sigma_{MS}}{Tolerance}$$

Based on the guidelines, good measurement systems will have a *P/T* ratio that is less than 0.10, while adequate measurement systems will have a *P/T* ratio that is less than 0.30.

% R&R measures the percent of Observed Process Variation being taken up by measurement error. The formula for this index is

$$\%R\&R = \frac{\hat{\sigma}_{MS}}{\hat{\sigma}_{Observed\ Process\ Variation}} x100$$

As a target, a good and capable measurement system has a %R&R less than 30%.

ICC (Intraclass Correlation Coefficient) measures the proportion of variation in product measurements attributed to measurement system and uses well defined "Western Electric" rules defined in Chapter 3 to help understand process stability. The formula for ICC is given by :

$$ICC = (1 - \frac{\hat{\sigma}^2_{MS}}{\hat{\sigma}^2_{Observed\ \ Process\ Variation}})$$

ICC uses variance rather than standard deviation, which is used in %R&R calculation. Variance is much easier to work with mathematically than standard deviations, which cannot be added or subtracted easily. ICC also lays out very clear guidelines to help us understand the capability of the measurement system based on the severity of the measurement issue.

As discussed in Section 3.6.3b, ICC>0.8 denotes minimal measurement variation while ICC <0.2 signifies an ineffective measurement system. For the ICC values between 0.2 and 0.8, we use the western electric rules to understand the capability of our measurement system.

Out of the three capability indices, we have to decide which measurement capability index to pick to measure the variation from our methods. There are 2 key criteria that are important to consider in making this decision.

1) We need an index, which could measure the variation with respect to specifications defined by <u>all</u> necessary customers. This includes final product specifications defined by the target product profile based on patient needs and total process variation defined by the process. Upon further research, we have found out that none of these ratios by themselves achieved this objective.

2) We need an index, which lays out clear guidelines on how to understand the measurement variation but isn't overly conservative. This would help avoid constraining or overstating the noise in the measurement system. As noted previously, %R&R guidelines can be overly conservative for understanding the variation in the measurement system.

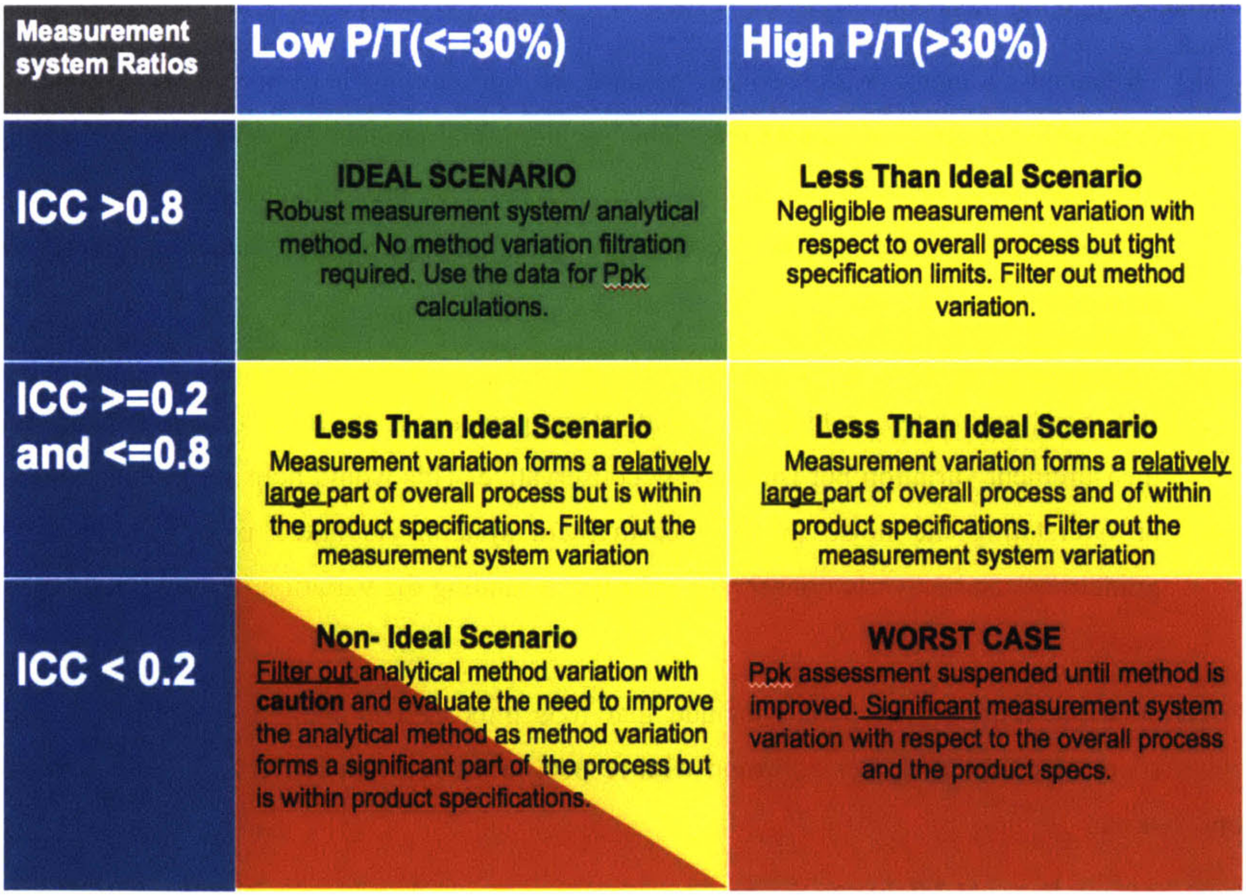
Based on these requirements, we conclude that it is important to look at a combination of different ratios, which evaluates variation with respect to the relevant limits and is not overly conservative.

Therefore, we have decided to measure the variation in methods by computing **both ICC and P/T ratio.** These well- defined multi-dimensional criteria can be very effective as ICC will help us understand the measurement variation with respect to overall process variation and P/T ratio will help us do the same with respect to product specification limits. Additionally, as noted above, ICC provides a nice set of clear guidelines in a tier format that take into account the severity of the variation in a rational manner.

6.4 Measurement System Decision Matrix

Based on the above defined criteria to measure and analyze assay variation, we have built the following Measurement system Decision Matrix as shown in table 5. Based on the computed values of ICC and P/T ratio, this matrix guides us on the next steps to address method or assay variation.

Table 5: Measurement System Decision Matrix



Measurement system Ratios	Low P/T(<=30%)	High P/T(>30%)
ICC >0.8	**IDEAL SCENARIO** Robust measurement system/ analytical method. No method variation filtration required. Use the data for Ppk calculations.	**Less Than Ideal Scenario** Negligible measurement variation with respect to overall process but tight specification limits. Filter out method variation.
ICC >=0.2 and <=0.8	**Less Than Ideal Scenario** Measurement variation forms a relatively large part of overall process but is within the product specifications. Filter out the measurement system variation	**Less Than Ideal Scenario** Measurement variation forms a relatively large part of overall process and of within product specifications. Filter out the measurement system variation
ICC < 0.2	**Non- Ideal Scenario** Filter out analytical method variation with **caution** and evaluate the need to improve the analytical method as method variation forms a significant part of the process but is within product specifications.	**WORST CASE** Ppk assessment suspended until method is improved. Significant measurement system variation with respect to the overall process and the product specs.

The below criteria provide an overview on how to utilize this matrix:

- **When P/T<=30% and ICC>0.8**

This is the ideal scenario as we have a highly capable measurement system. The measurement system variation is a very small part of the total process variation and of the product tolerance or specifications.[9]

The recommendation is to use the dataset for calculating process capability. There is no need to filter out the variation from analytical method.

- **When P/T>30% and ICC<0.2**

This is the worst case scenario as the measurement system variation is a significant component of overall process variation and of product tolerance in this case.

Recommendation is to improve the analytical method before conducting any analysis for process capability or Ppk as the original data contains too much variation from the method and removing it will cause the process capability to be artificially very high.

- **Special case: When P/T<30% and ICC<0.2**

This is an interesting case. Though the measurement system variation is a significant part of overall process variation, it doesn't contribute significantly to the specifications or the tolerance. For this reason, the recommendation is to evaluate if the analytical method needs to be improved based on the capability (Ppk) of the current process and the set specification limits. Process capability can also be calculated after filtering out the method variation, but with **caution** as Process capability (Ppk) of "only" the process will be high in this case.

[9] The total process variation comes from the true process variation and the continuous data measurement system.

- **Filter out the analytical method variation for the rest**

For the rest of the scenarios in the matrix, we need to filter out the method variation using the procedure shown in Section 6.5 as the measurement variation in these cases is neither low enough to use the original data for process capability nor high enough to stop the process capability analysis.

As we can see, removing the measurement variation from the original dataset containing overall process variation allows us to calculate process capability, which shows the actual capability of only the process. This is a huge step in helping us understand the variation coming from "only" the process.

6.5 Procedure to filter out method variation in Ppk calculation

We have created a procedure to filter out the method variation from the overall process variation. The following steps are required to filter out the noise from the assay or the method from the original dataset.

We know that Ppk of the original dataset of a parameter can be calculated using the below formula

$$Ppk = \min[\frac{(USL - \overline{X})}{3 * s}, \frac{(\overline{X} - LSL)}{3 * s}]$$

Where s= standard deviation of the original sample data set

- Step 1: Obtain % Coefficient of variation (CV) of repeatability and intermediate precision of the assays *(from the method development report)*

- Step 2: Calculate $\hat{\sigma}_{repeatability}$ and $\hat{\sigma}_{intermediate\,precision}$ of the measurement system using the formula $CV = \frac{\hat{\sigma}}{\overline{X}}$ where $\overline{X}$ is mean of original sample data set and CV is from method development report and $\hat{\sigma}$ is variance of the sample.

- Step 3: Calculate $\hat{\sigma}_{Measurement\ System} = \sqrt{(\hat{\sigma}^2 Repeatability) + (\hat{\sigma}^2 Intermediate\,Precision)}$

- Step 4: Calculate net standard deviation $\hat{\sigma}_{net} = S - \hat{\sigma}_{measurement\ system}$

  Note: The subtraction of standard deviation is possible in this case since S (standard deviation of the original data set and $\hat{\sigma}_{Measurement\ System}$ (standard deviation of measurement system) are based on the same mean $\overline{X}$ (mean of the original data set).

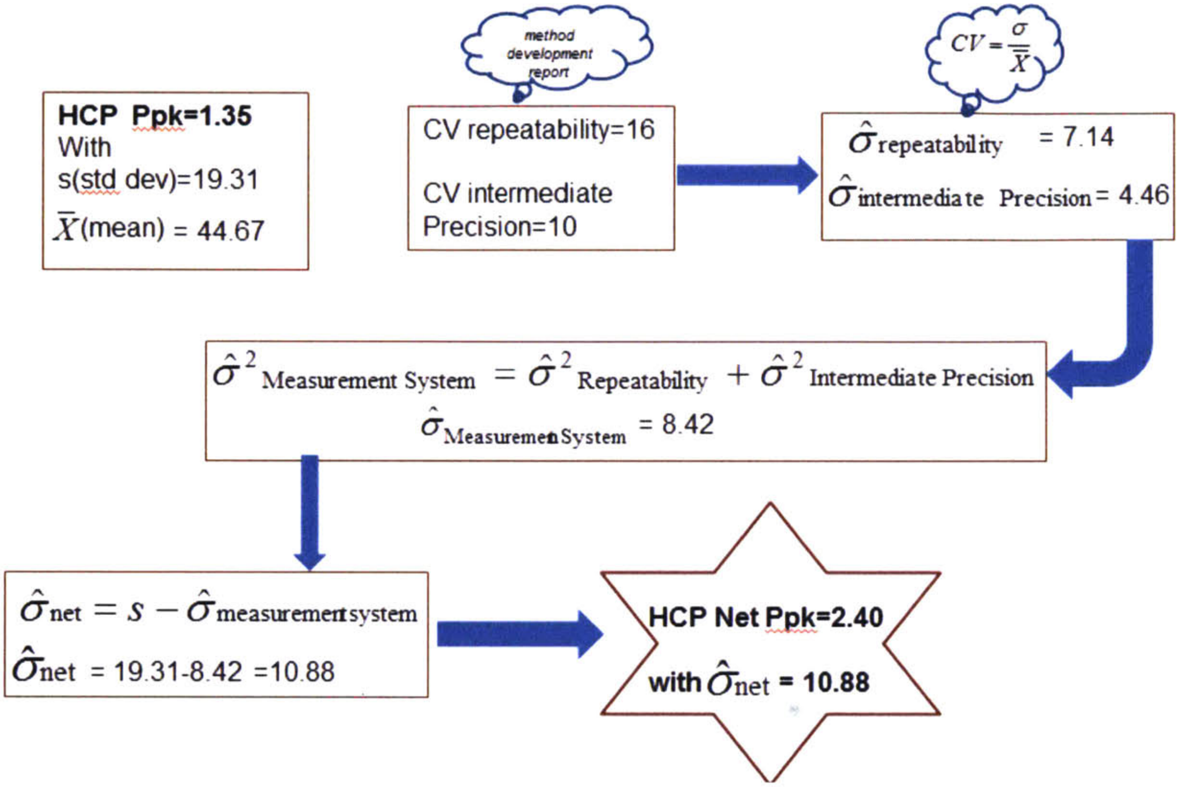
- Step 5: Use the net standard deviation or $\hat{\sigma}_{net}$ in the Ppk equation to calculate Ppk net as shown below

$$Ppk_{net} = \min[\frac{(USL - \overline{X})}{3 * \hat{\sigma}_{net}}, \frac{(\overline{X} - LSL)}{3 * \hat{\sigma}_{net}}]$$

In this manner, we are able to calculate the "true" process capability or Ppk of the process.

A case study demonstrating the use of this procedure on filtering out the method variation of host cell protein from the overall process variation has also been conducted (as shown below in Figure 14)

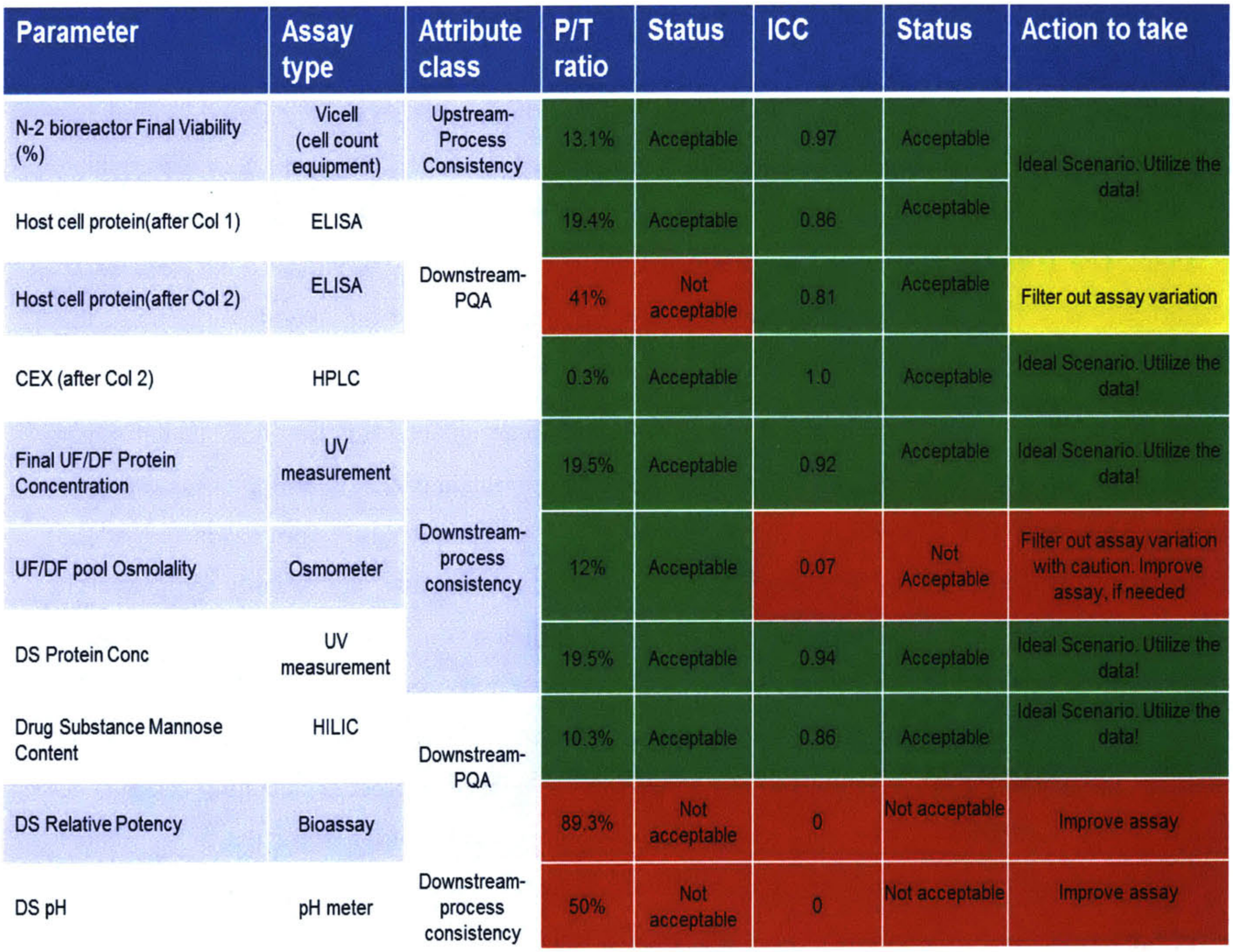
Figure 14: Filtration of method variation of Host Cell Protein


HCP Ppk=1.35
With
s(std dev)=19.31
$\bar{X}$(mean) = 44.67
method development report
CV repeatability=16
CV intermediate Precision=10
$CV = \frac{\sigma}{\bar{X}}$
$\hat{\sigma}_{\text{repeatability}}$ = 7.14
$\hat{\sigma}_{\text{intermediate Precision}}$ = 4.46
$\hat{\sigma}^2_{\text{Measurement System}} = \hat{\sigma}^2_{\text{Repeatability}} + \hat{\sigma}^2_{\text{Intermediate Precision}}$
$\hat{\sigma}_{\text{MeasuremetSystem}}$ = 8.42
$\hat{\sigma}_{\text{net}} = s - \hat{\sigma}_{\text{measurementsystem}}$
$\hat{\sigma}_{\text{net}}$ = 19.31-8.42 =10.88
HCP Net Ppk=2.40
with $\hat{\sigma}_{\text{net}}$ = 10.88

6.6 Expanded Analysis

The above analytical method variation analysis has been expanded to include many different parameters and assay types. Table 6 includes the results of these analyses and the recommendations.

Table 6: Analytical method variation analysis for several parameters



Parameter	Assay type	Attribute class	P/T ratio	Status	ICC	Status	Action to take
N-2 bioreactor Final Viability (%)	Vicell (cell count equipment)	Upstream-Process Consistency	13.1%	Acceptable	0.97	Acceptable	Ideal Scenario. Utilize the data!
Host cell protein(after Col 1)	ELISA	Downstream-PQA	19.4%	Acceptable	0.86	Acceptable	
Host cell protein(after Col 2)	ELISA		41%	Not acceptable	0.81	Acceptable	Filter out assay variation
CEX (after Col 2)	HPLC		0.3%	Acceptable	1.0	Acceptable	Ideal Scenario. Utilize the data!
Final UF/DF Protein Concentration	UV measurement	Downstream-process consistency	19.5%	Acceptable	0.92	Acceptable	Ideal Scenario. Utilize the data!
UF/DF pool Osmolality	Osmometer		12%	Acceptable	0.07	Not Acceptable	Filter out assay variation with caution. Improve assay, if needed
DS Protein Conc	UV measurement		19.5%	Acceptable	0.94	Acceptable	Ideal Scenario. Utilize the data!
Drug Substance Mannose Content	HILIC	Downstream-PQA	10.3%	Acceptable	0.86	Acceptable	Ideal Scenario. Utilize the data!
DS Relative Potency	Bioassay		89.3%	Not acceptable	0	Not acceptable	Improve assay
DS pH	pH meter	Downstream-process consistency	50%	Not acceptable	0	Not acceptable	Improve assay

Looking at this at a holistic level, some of the interesting cases to note are comparisons between host cell protein in col 1 and in col 2 and DS protein concentration and final UF/DF protein concentration. The results of UF/DF pool Osmolality and DS pH also need to be paid attention to get an understanding of special cases.

In the case of host cell protein (HCP), the measurement system variation in Column 1 is a small contributor to the tolerance (small P/T ratio) but in Column 2, it contributes significantly to the tolerance. (high P/T). The reason is that though both parameters are measured using the same assay(

Enzyme linked immunosorbent assay or ELISA) , the spec limits for host cell protein in Column 2 are set narrower than the specs for the host cell protein in Column 1, which leads to the high P/T ratio for host cell protein in Column 2.

On the other hand, in the case of UF/DF protein concentration and DS protein concentration measured immediately afterwards, both the P/T ratio and the ICC are almost similar. The reason is that both UF/DF and DS are measured against the same spec limits and are also measured through the same method of UV absorbance. Therefore, the P/T ratio for both the UF/DF and DS protein, which measures the contribution of measurement system variation towards the tolerance or specifications, is the same. The ICC for DS protein concentration is a little higher(0.94) than it is for UF/DF protein(0.92) since the measurement system becomes a smaller part of the overall variation at the drug substance stage and can increase the overall process variation.

In the case of UF/DF pool osmolality, the P/T ratio is acceptable but the ICC is definitely unacceptable. Here, the measurement system variation is a significant part of overall process variation but it doesn't contribute significantly to the specifications or the tolerance. The reason is that the measurement system has a very large allowable variation so it contributes significantly to the overall variation but the spec limits are wide enough so that the variation coming from measurement system is not a big contributor to the specifications. This conclusion has been confirmed both with the manufacturer of the equipment and lab scientists at Amgen. The suggestion is to evaluate if the method or the measurement system needs to be improved. The revised process capability can also be calculated but caution has to be applied as once the significant variation from the measurement system is taken out, the resulting Ppk value can be very high.

The fourth and the last case is of DS pH, which has P/T ratio and ICC much below acceptable limits. Here, measurement system variation is a significant part of overall process variation and of process tolerance. The recommendation is to improve the analytical method before conducting any analysis. On further research, this conclusion was confirmed as we found out that the pH meter is known to have a high allowable variation relative to the specifications for this parameter

As we can see, the use of this matrix is a powerful way to assess and filter out, if needed, the variation from the analytical method of any parameter. This further allows the calculation of the "actual" process capability without the variation coming from the method or assay.

7.0 Estimating Ppk

7.1 Proposed Methodology Overview

This chapter covers the last phase of this project, which includes estimating Ppk for molecules, where there is not enough data to calculate Ppk with the current values. This usually happens when the number of batches for a molecule is very low (i.e. the sample size is very low) because the product is very new and has had either very few commercial runs or in some cases, only lab scale or clinical runs. Note that the lab scale runs are done in the early phase of process development and are followed by clinical and commercial runs in that order.

Bayesian analysis is being recommended to estimate Ppk of the molecule in these cases as it leverages data from different scales i.e. small scale done at the lab or large scale done clinically or commercially and from different molecules for Ppk estimation.

7.2 Data Collection

The very first step in this approach is data collection. The molecule we have picked for the analysis has recently started running some commercial lots and has lab scale, clinical and some commercial scale data available. The clinical and commercial lots are just under 10 in number and including lab scale data increases the total sample size 30-40 depending on the parameter. Data from all scales has been incorporated for different parameters in order to have all the relevant data possible for more accurate calculations. The parameters of the molecule picked for this analysis are the same as the ones picked for bootstrapping as discussed in Section 5.2. The lab scale data is available for all parameters except for glycan content and potency.

After the data from different scales is gathered, the datasets are checked for normality as the Ppk calculations done later assume a normal distribution. Next, a type of analysis of variance test called Brown Forsythe test is done on the data to check if the small scale and large scale data can be combined. This test compares the spread of the different data sets and tests if the data sets are statistically significantly different as has been discussed in Section 3.7.3. Next, we have to decide the historical dataset of molecules to use for Bayesian approach. In order for the analysis to be effective, there needs to be a representative dataset that captures historical variation. The two key criteria we have determined for the historical datasets are:

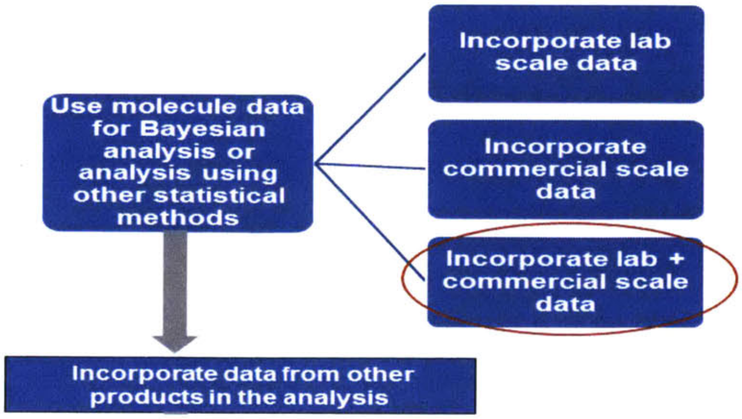
1. The molecule(s) should have similar unit operations as the molecule in question for comparison. For example, on a high level, we can combine monoclonal antibodies as they have similar manufacturing processes but not monoclonal antibodies and small molecules as they have very different manufacturing processes. Looking at the processes more closely, even within monoclonal antibodies, not all molecules can be combined together as the unit operations might be different between molecules.
2. The molecule should have at least 5 or more data points for each parameter that will be compared. This is important as the quality of the output produced by the Bayesian approach depends a lot on the input or the data provided in the analysis as discussed in Section 3.8. So, the more relevant data we can have, the more reliable our estimation of the Ppk value will be.

These key criteria are important mainly to ensure that we are picking the "right" combinations from the beginning, which can save us a lot of time later on when performing the testing. Note that after the molecules are selected, the normality test[10] and the analysis of variance test is again conducted. The analysis of variance test serves as a final check to verify the compatibility of the data.

The above explained data collection process (without the key criteria) is captured in Figure 15.

Figure 15: Data collection for Bayesian analysis

[10] Note that the normality test for different molecules is done separately before combining them as the datasets will very likely have different means and so cannot be tested together for normality.


Use molecule data for Bayesian analysis or analysis using other statistical methods
Incorporate lab scale data
Incorporate commercial scale data
Incorporate lab + commercial scale data
Incorporate data from other products in the analysis

7.3 Homogeneity of Variance tests

Now, we will discuss the Homogeneity of variance tests that were done to check if the data from different sources i.e. large scale, small scale, different molecules could be combined. Note that we could not check the compatibility of the data with commonly used tests such as t-tests or ANOVA. As discussed in section 3.7.1, these tests determine the statistical significant difference between the datasets based on the comparing the mean of the data sets and assume that the variances of the datasets are equal. Since the data sets we are comparing are very different in terms of scale and molecule, the means are going to be different in most cases and we definitely cannot assume that the variances of the data amongst the different datasets are equal.

Therefore, we conducted a Homogeneity of Variance test called the Brown Forsythe test first on datasets from different scales and then on datasets from different molecules. In order to determine whether the datasets are statistically significantly different, the test calculates a *p*-value, which is the probability of observing an effect given that the null hypothesis that there is no difference in the datasets is true. The null hypothesis is rejected if the *p*-value is less than the significance level of the test or α (which is usually set at 5% or 1%) and therefore, confirms that the datasets are statistically significantly different.

The results from the test are shown below in Figure 16 and 17 for 2 parameters- drug substance (DS) pH and column 1 step yield. Note that the data from clinical and commercial lots can be combined as they are made at on the same scale in this case.

Figure 16: Brown Forsythe test for DS pH lab data and Clinical data

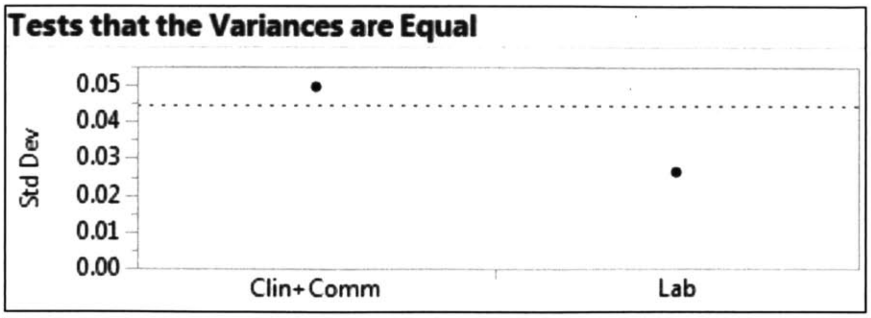


Figure 16 shows the standard deviation plotted for DS pH lab data and commercial data combined with clinical data. The P value calculated from the Brown Forsyth test based on the difference of standard deviation and the median was 0.4223, which is greater than 0.05, the significance level of the test. This shows that the two datasets are not statistically significantly different.

Figure 17: Brown Forsythe test for Column 1 step yield (1)

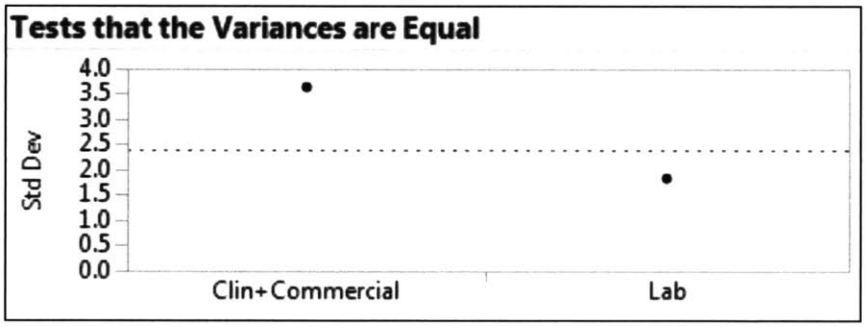
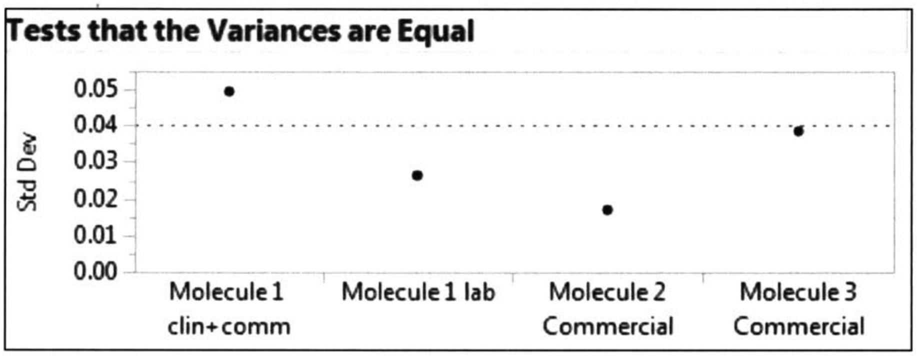


Figure 17 shows the standard deviation plotted for Column 1 step yield lab data and commercial data combined with clinical data. The P value in this case was 0.0041 from the Brown Forsyth test, which is much less than 0.05, the significance level of the test. This shows that the two datasets are statistically significantly different.

All parameters for which the lab scale data was available (host cell protein after Column 1, final UF/DF protein concentration, drug substance pH) except for Column 1 step yield passed the Homogeneity of Variance (HOV) test between lab and commercial scale. On review of the step yield data, it was confirmed that step yield can be impacted due to scale and sampling differences between lab and commercial scale. Hence, the data are not comparable.

The datasets for different molecules went through the same rigorous test once the molecules were selected. The results for these are shown below in Figure 18, 19 and 20.

Figure 18: Brown Forsythe Test for DS pH for Molecule 1,2 and 3



Note: Molecule 1 is the molecule we are comparing Molecule 2 and 3 against in this test.

Figure 18 shows the standard deviation plotted for DS pH data plotted for three different molecules. Molecule 1 has both lab data and commercial data included as the two datasets passed the HOV test as shown in Figure 16. These data are combined with commercial scale data for Molecule 2 and 3. The P value from the Brown Forsythe test in this case was 0.3474, which is greater than 0.05, the significance level of the test. This shows that the datasets of the 3 molecules can be combined as they are not statistically significantly different.

Figure 19:Brown Forsythe test for Column 1 Step Yield(2)

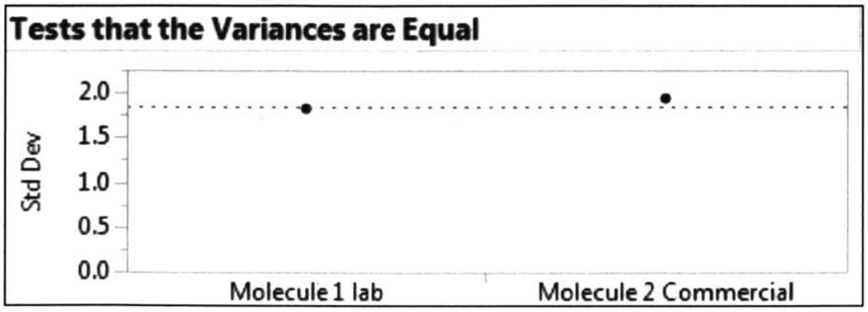
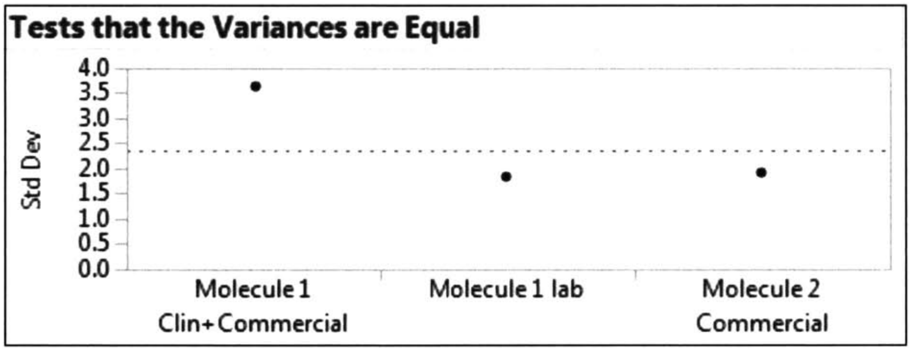


Figure 19 is another example of the comparison tests done between the different molecules for Column 1 step yield. Here, Molecule 1 lab and Molecule 2 commercial scale data have been combined. The P value in this case from the Brown Forsythe was 0.8187, which shows that the variances are comparable and the two datasets are not statistically significantly different. Note that, we didn't use the Molecule 1 commercial data in addition to the lab data since the two datasets failed the homogeneity of the variance test as we saw earlier in Figure 17. If we would have combined Molecule 1 commercial and lab scale data and compared it with Molecule 2 data, we would have definitely failed the HOV test. This is shown in Figure 20, which demonstrates the results from the Brown Forsythe test for Column 1 step yield for Molecule 1 lab and commercial scale data and Molecule 2 commercial scale data. The P value in this case is 0.0092, less than 0.05. Hence the datasets cannot be combined.

Figure 20: Brown Forsythe test for Column 1 Step Yield (3)



As we can see, this test has been very effective in analyzing the comparability of the datasets and helping us understand which datasets could or could not be combined. This is a crucial step in Bayesian approach as this approach takes into account historical information provided in order to present estimations of a particular statistic.

7.4 Estimation of Ppk

Bayesian analysis assumes that all model parameters are random quantities and thus can incorporate prior knowledge. This prior information can be used to get a more accurate response especially for unknown quantities and the prior can be both informative and non-informative. The algorithm, which is being used for Bayesian analysis, uses an informative prior represented by the inverse gamma function distribution represented below. This is also covered in Section 3.8.3.

$$IG(x \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-1-\alpha} \exp(\frac{-\beta}{x})$$

However, in order to make sure that the informative prior is utilizing the most accurate information possible, the algorithm during its first run inputs the information available, which in this case is of molecule 2 and molecule 3, in a non- informative Jeffrey's prior. As we saw in Section 3.8.3, Jeffrey's prior is proportional to the square root of the determinant of the Fisher information, I(θ) as shown below.

$$\pi(\theta) \propto \mathrm{I}^{1/2}(\theta)$$

Where I(θ) or Fisher information is given by the below formula

$$I(\theta) = E_{\theta}(\frac{\partial \log f(X \mid \theta)}{\partial \theta})^2$$

The non-informative prior, only used during the first run, provides information on the shape parameter α and scale parameter β of the inverse gamma distribution of the informative prior. Using this information, the informative prior is run for the rest of the runs. This process ensures that the "prior" is very accurate and reliable, which is important as Bayesian output very much relies on the input.

The "likelihood" function is also developed through the algorithm as the prior gets updated with new evidence, which is information of molecule 1 (molecule in question). This creates an updated probability distribution called the "posterior distribution" based on the below formula as shown in Section 3.8: $\mathrm{p}(\theta \mid \mathrm{y}) = \frac{\mathrm{p}(\mathrm{y} \mid \theta)\mathrm{p}(\theta)}{\mathrm{p}(\mathrm{y})}$

Where y is the new data, p(θ) is the probability based on historical information and p(y|θ) is the likelihood function.

Based on the "posterior" distribution or density, the Bayesian simulation uses a computer random-generator to produce a large data matrix or sample of the model parameters such as mean and standard deviation.[11] The algorithm we are using produces 20,000 samples of mean and standard deviation data and takes each individual independent sample of one mean and one standard deviation to find the quantiles for each case. Conducting this process for all 20,000 independent samples allows us to construct the distribution and therefore, the confidence intervals for the parameters as shown below in Figure 21.

Figure 21: Distribution of Standard deviations

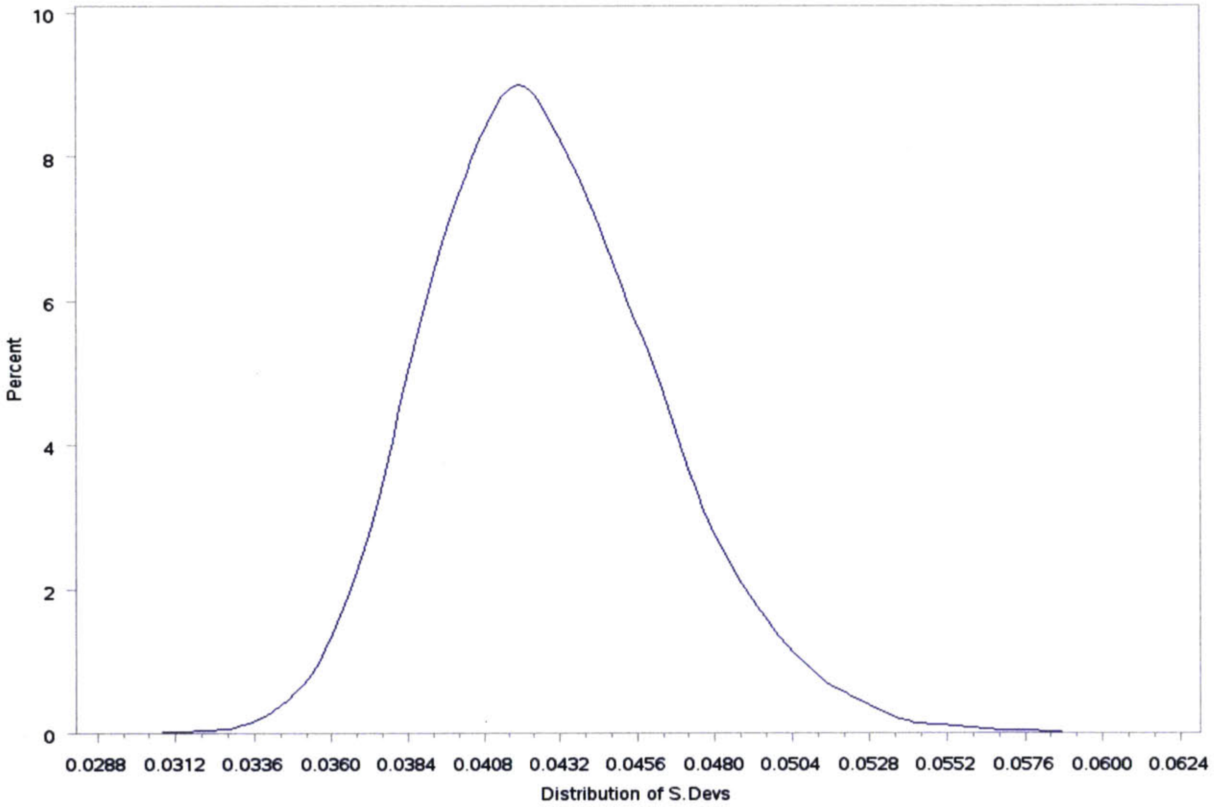


Figure 21 shows distribution of standard deviations. A very similar chart is created for the mean data through the Bayesian algorithm.

From these distributions, we take the mean of the standard deviation data (or s) and mean of the means (or X-bar) to calculate Ppk of the molecule in question using the formula

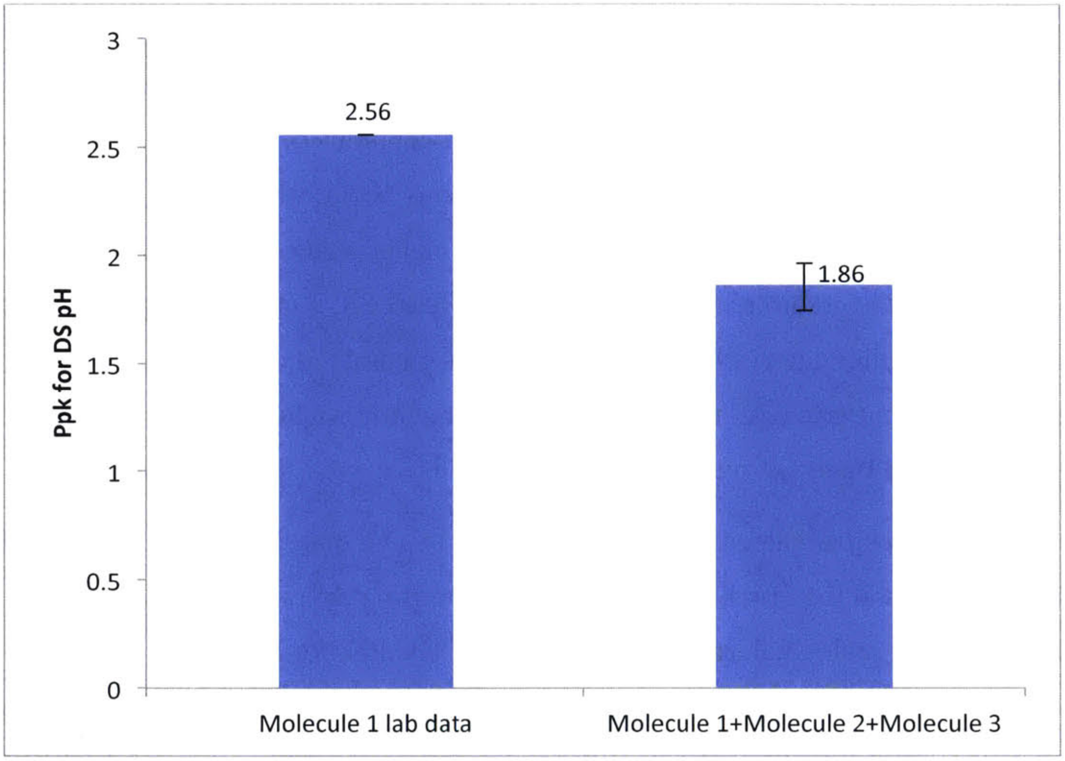
[11] Tolerance Intervals for Variance Component Models Using Bayesian Simulation By Russell Wolfinger

$$Ppk = \min[\frac{(USL - \overline{X})}{3 * s}, \frac{(\overline{X} - LSL)}{3 * s}]$$

It is also important to calculate a credible interval, which is analogous to confidence interval in Frequentist statistic, especially since the variability in the distribution of the location of the center of the distribution i.e. (the "mean") is most susceptible to changing with the sample size. With very small N, the central location (the mean) about which we place the credible intervals and tolerance intervals varies quite a lot. With higher N, the central location (the mean) about which we place the credible intervals varies much less. So, it is important to consider both the values and the credible intervals around it. The 90% credible intervals have been calculated in this case taking into account the 5th and 95th percentile values for mean and standard deviations given by the Bayesian output as shown in Fig 21.

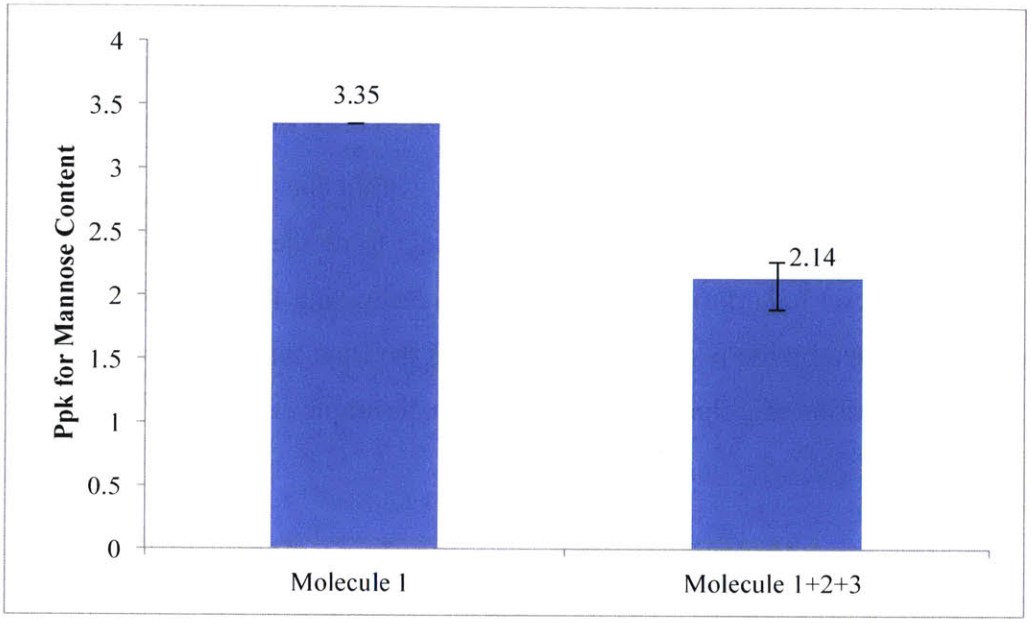
Figure 22 shows the Ppk calculated for parameter DS pH using only Molecule 1 lab DS pH data with sample size of 6 and the Ppk for the same parameter using data from Molecule 1, 2 and 3 at bench and commercial scale total sample size (N) of 39. The Ppk for Parameter at sample size of 6 has been calculated through the simple Ppk formula while the Ppk calculated using Molecule 1,2 and 3 data has been done through using the mean and standard deviation from the Bayesian output. As we can see, the algorithm is able to estimate a Ppk value of 1.86 for DS pH for Molecule 1 at an N of 39 using the data from other molecules. Additionally, we calculate a 90% credible interval of (1.74, 1.96) based on Bayesian output. This implies that 90% of the time the Ppk values for Molecule 1 at N=39 will be in this range of 1.74-1.96.

Figure 22: Ppk Estimation for DS pH from Bayesian analysis



Another example is shown below with the Ppk and the credible interval calculated for DS glycan content. Figure 23 shows the Ppk for glycan content calculated mathematically using only Molecule 1 data for sample size of 5 and the Ppk calculated using the Bayesian output for the same parameter using data from Molecule 1, 2 and 3 for a total sample size (N) of 40. The algorithm estimates a Ppk value of 2.24 for the DS glycan content of Molecule 1 at an N of 40 using the data from other molecules. Additionally, the 90% credible interval of (2.27, 1.90) for Molecule 1 DS glycan content has also been calculated based on Bayesian output.

Figure 23: Ppk Estimation for High Mannose Content



The reliability of this approach is further confirmed by a confirmation analysis (as shown in Figure 24) carried out for Molecule 2, which has more than 30 lots of commercial scale data.

Figure 24: Ppk Confirmation run for DS pH

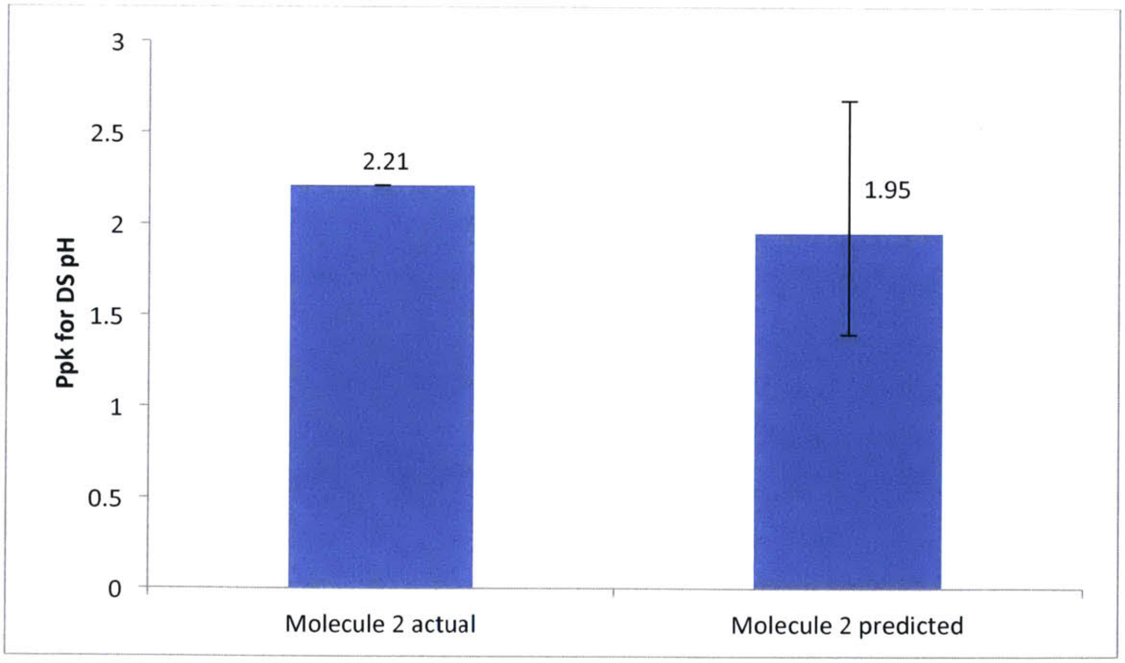


Figure 24 shows the actual and predicted Ppk at sample size of 30 for Molecule 2. Actual Ppk (2.21) for DS pH has been calculated using the available batch data of 30 batches of Molecule 2 as an input to the Ppk formula and the predicted Ppk (1.95) was calculated using the data from 6 batches of Molecule 2 and the rest of the 24 batches from Molecule 1 and 3 as an input to the Bayesian analysis. The credible interval estimated for Ppk from Bayesian analysis is (1.39, 2.68).

As we can see, the actual Ppk value of 2.21 falls within the predicted credible interval. In addition, the actual Ppk value of 2.21 is 16% away from the estimated Ppk of 1.95. This difference is minor and bound to occur since we are estimating Ppk based on information from other molecules. Based on the results, we can see that the Bayesian analysis is very effective and useful technique in estimating Ppk values for very small sample sizes.

8.0 Recommendations & Conclusions

8.1 Operational Recommendations

The process capability data are typically assessed for commercial products during the Continued Process Verification (CPV) phase of the validation lifecycle. Predicting process capability during process design will enable processes to meet pre-defined performance targets and facilitate continuous improvement.

The main goal of this thesis has been to establish standardized methodology for estimating process performance capability during the design phase of process development. Within this project, we have defined the criteria for performing predictive process capability assessments as a function of sample size, specifically for small sample sizes during development. We have also defined procedures to assess and filter out variation introduced by the measurement system from the overall process variation and recommended techniques to improve process capability estimates when the sample size is below a minimum threshold.

To define the criteria for performing predictive process capability for small sample sizes, we have evaluated different process capability indices, and determined Ppk to be the most appropriate metric for process development and biopharmaceutical manufacturing. Ppk measures the long term capability of the process which contains both common cause and special cause variation within and between different batches. This really facilitates the metric's use in biopharmaceutical manufacturing, where the batches are not all manufactured at the same time and we are interested in understanding the trends between different batches or subgroups with common cause and special cause variation. Additionally, measuring Ppk doesn't require the process to be in statistical control, which is ideal for our project as we are designing a methodology for processes that are still in development. Next, we have conducted a rigorous bootstrapping analysis to establish relationship between Ppk and sample size, particularly for small sample sizes. Based on the results, we are recommending a multi-tiered approach to estimate Ppk at low sample sizes. This approach takes into account the differing complexity of parameters, the reliability of the results from bootstrapping and produces a final assessment, which is parameter independent. To elaborate, sample sizes greater than 11 independent batches provide sufficient confidence in process performance capability estimates. Sample sizes (8-11 batches) will require uncertainty to be evaluated as part of the Ppk calculation, due to the width of confidence intervals. For sample sizes less than 8, the confidence intervals for the Ppk statistic are too large and not reliable. Therefore, Ppk cannot be used in practical application for such small sample sizes unless more batch data are collected or data from other scales and molecules is used to estimate Ppk through Bayesian analysis as shown in Chapter 7. To ensure the robustness of this bootstrapping analysis, a 1.5 sigma shift, which accounts for accepted variation in the mean over the long term has been considered.

Next, in order to evaluate and propose methods to quantify measurement system variation, we have focused on precision, which is commonly understood at three levels-repeatability, intermediate precision and reproducibility. We have assessed repeatability and intermediate precision instead of reproducibility since the lab or the working environment is not being changed in testing the parameters and we mainly want to understand the variation generated by equipment and appraiser.

A measurement capability index (or indices) has (have) been selected based on two key requirements: First, the metric or index needs to measure the variation with respect to specifications defined by all necessary internal and external customers. Second, the index needs to lay out clear guidelines to understand measurement variation but isn't overly conservative to avoid overstating the variation from the measurement system.

Based on the above requirements, we conclude that variation from the method needs to be understood by analyzing both ICC (Intraclass Correlation Coefficient) and P/T (Precision/Tolerance) ratio. A well-defined multi-dimensional criterion which takes into account both these ratios can be very effective as ICC will help us understand the measurement variation with respect to the variation of the overall process and P/T ratio will help us understand the measurement variation with respect to specification limits. ICC also provides a set of clear guidelines that take into account the severity of the variation in a rational manner. With these factors in consideration, we have established a decision matrix, a novel tool to guide us on when the variation from the measurement system is unacceptable and when is it appropriate to filter out assay variation.

To elaborate, when both P/T and ICC are well above threshold (i.e. P/T<=30% and ICC>0.8), we have a highly capable measurement system as the measurement system variation is a very small part of the total process variation and of the product tolerance or the specifications. The recommendation, in this case, is to use the dataset for calculating process capability without filtering out the variation from analytical method. On the other hand, when both P/T and ICC are below threshold (i.e. P/T>30% and ICC<0.2), we are presented with a worst case scenario, where there is significant measurement variation with respect to the overall process and the product tolerance or specifications. Here, we need to improve the analytical method before conducting any analysis for process capability or Ppk as the original data contains significant variation from the method. For the rest of the scenarios, we need to filter out the method variation using the procedure we have established. One interesting scenario to point out is when only the ICC is below threshold (ICC<0.2). The recommendation in this case is to check if the analytical method can be improved and calculate process capability after filtering out the method variation with **caution.**

The procedure we have established to separate measurement system variation from process variation uses Coefficient of variation (CV) values from the assays and the mean of the original dataset to calculate variation from repeatability and intermediate precision. The overall measurement system is then mathematically calculated by taking the square root of the sum of repeatability and intermediate precision variation. Variation from "only" the process is calculated by subtracting the variation of the measurement system from the overall process. Case studies have been performed to test this process.

Above analytical method variation analysis has been expanded to include many different parameters and assay types and the conclusions on the capability of measurement systems for the different parameters from the decision matrix have been verified through internal discussions with SME as well external discussions with the equipment suppliers, where applicable. This is discussed in Section 6.6. In summary, removing the measurement variation from the original dataset containing overall process variation allows us to calculate the actual capability of only the process. This is a huge step in helping us understand the variation coming from "only" the process.

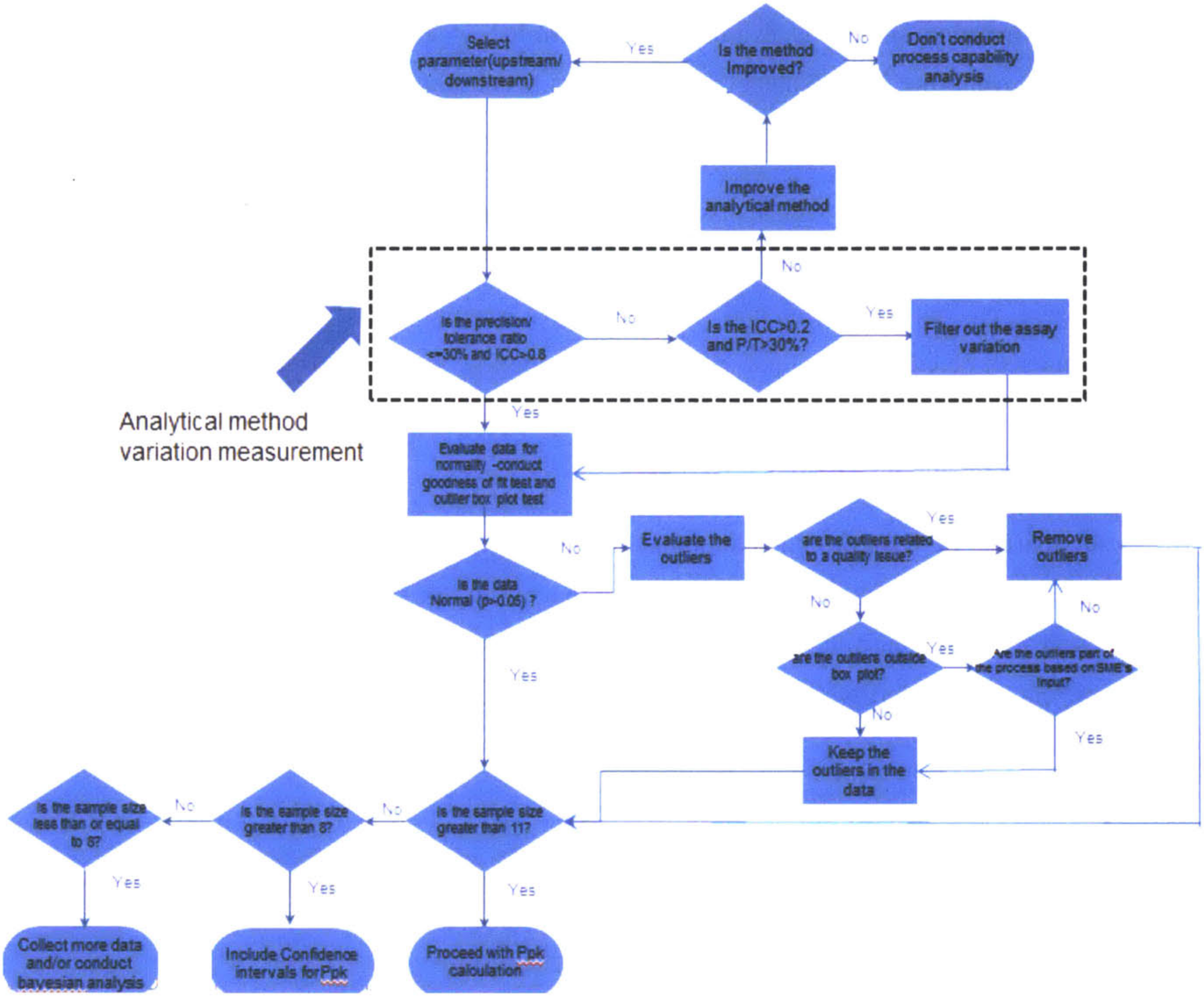
Last, in order to improve the estimation of Ppk for very small sample sizes (less than 8), additional data from different scales (small and large) and molecules have been leveraged for use in the Bayesian analyses. Before Bayesian analysis is conducted, we have analyzed the comparability of datasets by performing Homogeneity of Variance (HOV) test. This test helps us determine if the variances across different scales and different molecules are equal, in which case the datasets can be combined. The molecules to develop a historical database have been picked keeping in mind 2 key criteria. First, the molecule(s) should have similar unit operations as the molecule in question for comparison. Second, the molecule should have at least 5 or more data points for each parameter that will be compared.

This database of different molecules is then fed into the Bayesian algorithm, a powerful statistical tool for modeling and predictions, to form a posterior distribution. Accurate and careful selection of molecules to develop a historical dataset is important as the "posterior distribution" of the Bayesian analysis i.e. the output depends on the "prior" or the input provided for the analysis and the likelihood function, which is proportional to the distribution of the observed data. The Bayesian algorithm calculates a average standard deviation and a average mean after 20,000 iterations of the datasets. These point estimates such as standard deviation and mean are used to calculate Ppk. The credible intervals are also calculated based on high and low values of standard deviation and mean. Therefore, Bayesian analysis is crucial in estimating Ppk for small sample sizes.

8.2 Conclusions

This thesis proposes a methodology as shown in Figure 21 to predict process performance capability during the process design phase of a biopharmaceutical manufacturing process. Rigorous analysis and research has gone into determining the different techniques for measuring process and measurement method variation within this assessment whether it is bootstrapping analysis to determine confidence intervals for small sample sizes or a decision matrix to determine the variation from the method or the assay. Next steps include evaluating the applicability of the recommendations to phases beyond drug substance e.g. drug product development.

Figure 25: Ppk Assessment in Process Design Phase


Select parameter(upstream/ downstream)
Yes
Is the method Improved?
No
Don't conduct process capability analysis
Improve the analytical method
No
Is the precision/ tolerance ratio <=30% and ICC>0.8
No
Is the ICC>0.2 and P/T>30%?
Yes
Filter out the assay variation
Analytical method variation measurement
Yes
Evaluate data for normality -conduct goodness of fit test and outlier box plot test
No
Evaluate the outliers
are the outliers related to a quality issue?
Yes
Remove outliers
Is the data Normal (p>0.05) ?
No
No
are the outliers outside box plot?
Yes
Are the outliers part of the process based on SME's input?
No
Yes
Yes
Keep the outliers in the data
Is the sample size less than or equal to 8?
No
Is the sample size greater than 8?
No
Is the sample size greater than 11?
Yes
Yes
Yes
Collect more data and/or conduct bayesian analysis
Include Confidence intervals forPpk
Proceed with Ppk calculation

As shown in Figure 25, process performance capability (Ppk) assessment begins with choosing an upstream or downstream parameter. Analytical method variation analysis is conducted to assess the contribution of the method variation towards the overall process through a novel decision matrix using ICC and P/T ratios and a procedure we have created to calculate the Ppk of only the process. If the method variation is very significant, the method will need to be improved before conducting any process capability assessments.

Then, the data are checked for normality since Ppk calculations done later assume a normal distribution. The tests for normality are done graphically using the outlier box plot and mathematically using the p-values in the goodness of fit test. In case the data is not normal, the outlier box plot is checked for outliers and if there are any, the history of outlier batches in the non- conformance system is evaluated to determine special cause variation. In case of no assignable event, the SME (subject matter expert) is consulted to understand if the outliers are a normal part of the process.

Next, the sample size is checked based on the recommended multi-tiered approach to estimate Ppk at low sample sizes. Tier 1, where sample size is greater than 11 batches, provides sufficient confidence in Ppk estimates and Tier 2, where sample size from 8-11 batches requires uncertainty to be evaluated as part of the Ppk calculation. In case of Tier 3, when sample size is less than 8, Bayesian analysis needs to be conducted using data from different scales/molecules to calculate Ppk and the credible interval around it.

To summarize, we have designed a methodology to predict Ppk from small development datasets that leverages a filter for method variation and that allows for supplementation of small data sets. This will have immense value to enable prediction of process capability for important process parameters and will drive continuous improvement throughout the design phase of process development.

Bibliography

[1] M.A. Pett. *Nonparameteric Statistics for Healthcare Research: Statistics for Small Samples and Unusual Distributions*. Sage Publications,1997.

[2] "Biotechnology." *Wikipedia, the Free Encyclopedia*, June 22, 2016. **https://en.wikipedia.org/w/index.php?title=Biotechnology&oldid=726544535**.

[3] Research, Center for Biologics Evaluation and. "About the Center for Biologics Evaluation and Research - What Are." WebContent. Accessed April 26, 2016. **http://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CBER/ucm133077.htm**.

[4] Jin, Jing-fen, Ling-ling Zhu, Meng Chen, Hui-min Xu, Hua-fen Wang, Xiu-qin Feng, Xiu-ping Zhu, and Quan Zhou. "The Optimal Choice of Medication Administration Route Regarding Intravenous, Intramuscular, and Subcutaneous Injection." *Patient Preference and Adherence* 9 (July 2, 2015): 923–42. doi:10.2147/PPA.S87271.

[5] "Pharmaceutical Industry - Drug Discovery and Development | Britannica.com." Accessed July 6, 2016. https://www.britannica.com/topic/pharmaceutical-industry/Drug-discovery-and-development#ref925401.

[6] Division of Manufacturing and Product Quality, Center for Drug Evaluation and Research (CDER). "Guidance for Industry: Process Valiation General Principles and Practices," n.d. **http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm070336.pdf**.

[7] "About Amgen," 2015. **https://www.amgen.com/~/media/amgen/full/www-amgen-com/downloads/fact-sheets/fact_sheet_amgen.ashx/.**

[8]"Amgen Pipeline." Accessed May 20, 2016. **http://www.amgenpipeline.com/pipeline/**.

[9] "Bootstrapping Inferential Statistics with a Spreadsheet." Accessed Feb 10, 2016. **http://www.sportsci.org/2012/wghboot.htm**.

[10] John A. Rochowicz Jr, Alvernia University. "Bootstrapping Analysis, Inferential Statistics and EXCEL," Spreadsheets in Education (eJSiE), 4, no. 3 (2010). **http://epublications.bond.edu.au/cgi/viewcontent.cgi?article=1080&context=ejsie**.

[11] "Minimum Value of Sample Size to Bootstrap? - ResearchGate." Accessed March 10, 2016. **https://www.researchgate.net/post/Minimum_value_of_sample_size_to_bootstrap2**.

[12] Michael R. Chernick. “When Bootstrapping Fails Along with Some Remedies for Failures.” In *Bootstrap Methods: A Guide for Practitioners and Researchers, Second Edition*, 2007.

[13] “Interquartile Range.” *Wikipedia, the Free Encyclopedia*, June 9, 2016. **https://en.wikipedia.org/w/index.php?title=Interquartile_range&oldid=724451037**.

[14] Razali, Normadiah Modh, and Yap Bee Wah. “Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests,” Journal of Statistical Modeling and Analytics, Vol.2 No.1 (2011).

[15] “The Real Truth Behind Cpk and Ppk Capability and Potential Process Studies.” Accessed March 15, 2016. **http://www.gearsolutions.com/article/detail/6246/The-Real-Truth-Behind-Cpk-and-Ppk-Capability-and-Potential-Process-Studies**

[16] Özlem Şenvar and Hakan Tozan. “Process Capability and Six Sigma Methodology Including Fuzzy and Lean Approaches.” In *Products and Services; from R&D to Final Solutions*, n.d. **http://cdn.intechopen.com/pdfs-wm/12326.pdf**.

[17] “Six Sigma.” *Wikipedia, the Free Encyclopedia*, June 24, 2016. https://en.wikipedia.org/w/index.php?title=Six_Sigma&oldid=726748704.

[18] “Short Term Sample versus Long Term Sample.” Accessed April 15, 2016. **http://www.six-sigma-material.com/longtermshorttermsample.html**.

[19] “Converting DPMO & Sigma Level.” Accessed April 15, 2016. **http://qualityamerica.com/LSS-Knowledge** Center/leansixsigma/converting_dpmo_sigma_level.php.

[20] “How Is the Defects Per Million Opportunities (DPMO) to Sigma Level and Yield Conversion Chart Determined?” Accessed April 18, 2016. **https://www.isixsigma.com/ask-dr-mikel-harry/how-defects-million-opportunities-dpmo-sigma-level-and-yield-conversion-chart-determined/**.

[21] “Measurement System Analysis (MSA) Tutorial.” Accessed April 20, 2016. **https://www.moresteam.com/toolbox/measurement-system-analysis.cfm**.

[22] Raytheon Company. “Measurements System Analysis,” 2003-2007 http://www.raytheon.com/connections/rtnwcm/groups/public/documents/content/rtn_connect_msa_pdf.pdf.

[23] “ASTM International - Standards Worldwide.” Accessed April 25, 2016. **http://www.astm.org/SNEWS/ND_2010/datapoints_nd10.html**.

[24] “Problems With Gauge R&R Studies | Quality Digest.” Accessed May 1, 2016. **http://www.qualitydigest.com/inside/twitter-ed/problems-gauge-rr-studies.html**.

[25] Donald Wheeler. "The Intraclass Correlation Coefficient- Is Your Measurement System Adequate." *SPC Press* Manuscript No. 222, Dec 2, 2010 http://www.spcpress.com/pdf/DJW222.pdf.

[26] "Difference Between T-TEST and ANOVA | Difference Between | T-TEST vs ANOVA." Accessed June 1, 2016. **http://www.differencebetween.net/miscellaneous/difference-between-t-test-and-anova/**

[27] "Unequal Variances." Accessed June 2, 2016. **http://www.jmp.com/support/help/Unequal_Variances.shtml**.

[28] "Introduction to Bayesian Analysis." In *STATA BAYESIAN ANALYSIS REFERENCE MANUAL RELEASE 14*, 255. Stata Press, 2015. **http://www.stata.com/manuals14/bayes.pdf**.

[29]"What Is Bayesian Analysis? | International Society for Bayesian Analysis." Accessed June 2, 2016. **https://baycsian.org/Bayes-Explained**.

[30] ANNE RANDI SYVERSVEEN. "NONINFORMATIVE BAYESIAN PRIORS INTERPRETATION AND PROBLEMS WITH CONSTRUCTION AND APPLICATIONS," n.d.

[31] Russell Wolfinger. "Tolerance Intervals for Variance Component Models Using Bayesian Simulation," Journal of Quality Technology, VOL. 30 NO. 1 (January 1998): 18–32

[32] Scott M. Lynch. "Basics of Bayesian Statistics." In *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, 47–75, n.d.

[33] "How Do Drugs and Biologics Differ? - BIO." Accessed June 20, 2016. https://www.bio.org/articles/how-do-drugs-and-biologics-differ.

[34] "Cp, Cpk, Pp and Ppk: Know How and When to Use Them." Accessed Feb 12, 2016. **https://www.isixsigma.com/tools-templates/capability-indices-process-capability/cp-cpk-pp-and-ppk-know-how-and-when-use-them/**

[35] "Cpk vs Ppk: Who Wins? | BPI Consulting." Accessed Feb12, 2016. **https://www.spcforexcel.com/knowledge/process-capability/cpk-vs-ppk-who-wins**.

[36] "Process Capability Statistics: Cpk vs. Ppk | Minitab." Accessed Feb 20, 2016. **http://blog.minitab.com/blog/michelle-paret/process-capability-statistics-cpk-vs-ppk**

[37] Robert H. Mitchell, 3M Company. "Process Capability Indices," ASQ Statistics Division Newsletter, Vol 18, No.1

[38] “1.5 Sigma Process Shift Explanation.” Accessed March 15, 2016. **https://www.isixsigma.com/new-to-six-sigma/dmaic/15-sigma-process-shift/**

[39] Peter Hall. *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics, 1992.