Statistics for Biology and Health

Richard K. Burdick
David J. LeBlond
Lori B. Pfahler
Jorge Quiroz
Leslie Sidor
Kimberly Vukovinsky
Lanju Zhang

Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry

Springer

Statistics for Biology and Health

More information about this series at http://www.springer.com/series/2848

Richard K. Burdick • David J. LeBlond
Lori B. Pfahler • Jorge Quiroz • Leslie Sidor
Kimberly Vukovinsky • Lanju Zhang

Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry

Springer

Richard K. Burdick
Elion Labs
Louisville, Colorado, USA

David J. LeBlond
CMC Statistics
Wadsworth, IL, USA

Lori B. Pfahler
Merck & Co., Inc.
Telford, PA, USA

Jorge Quiroz
Merck & Co., Inc.
Kenilworth, NJ, USA

Leslie Sidor
Biogen
Cambridge, Massachusetts, USA

Kimberly Vukovinsky
Pfizer
Old Saybrook, Connecticut, USA

Lanju Zhang
Nonclinical Statistics, Abbvie Inc.
North Chicago, IL, USA

ISSN 1431-8776 ISSN 2197-5671 (electronic)
Statistics for Biology and Health
ISBN 978-3-319-50184-0 ISBN 978-3-319-50186-4 (eBook)
DOI 10.1007/978-3-319-50186-4

Library of Congress Control Number: 2016962314

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction

Keywords Chemistry, Manufacturing, and Controls (CMC) statistics • Clinical statistics • Pharmaceutical activities • Regulatory guidance • Statistical methods

1.1 Objectives

The motivation for this book came from an American Association of Pharmaceutical Scientists (AAPS) short course on statistical methods applied to Chemistry, Manufacturing, and Controls (CMC) applications presented by four of the authors. One of the course participants asked us for a good reference book, and the only book we could recommend was written over 20 years ago by Chow and Liu (1995). We agreed that a more recent book would serve a need in our industry. This book presents statistical techniques that are critically important to CMC activities.

Statistical methods are presented with a focus on applications unique to the CMC pharmaceutical industry. The target audience consists of statisticians and other scientists who are responsible for performing statistical analyses within a CMC environment. Basic statistical concepts are addressed in Chap. 2 followed by applications to specific topics related to development and manufacturing. The mathematical level assumes an elementary understanding of statistical methods. The ability to use Excel or statistical packages such as Minitab, JMP, or *R* will provide more value to the reader.

Since we began this project, an edited book has been published on the same topic by Zhang (2016). The chapters in Zhang discuss statistical methods for CMC as well as drug discovery and nonclinical development. We believe our book complements Zhang by providing more detailed statistical analyses and examples.

1.2 Regulatory Guidance for CMC Applications

Persons responsible for statistical analyses in CMC applications should be familiar with guidance and regulations that pertain to the pharmaceutical industry. The legality of CMC issues is covered in the Code of Federal Regulations (CFR),

R.K. Burdick et al., *Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry*, Statistics for Biology and Health, DOI 10.1007/978-3-319-50186-4_1

Title 21, Food and Drugs Administration (FDA). Several relevant sections of this code are reported in Table 1.1.

In addition to the CFR, regulatory agencies have produced a number of useful documents that direct the approaches used in statistical analysis. Tables 1.2, 1.3, 1.4, and 1.5 report documents referenced and discussed in this book.

1.3 Use of Statistical Tools in Pharmaceutical Development and Manufacturing

This book focuses on statistical methods used in the development and manufacturting of pharmaceutical products. An excellent description of this area is presented by Peterson et al. (2009). Pharmaceutical products are developed over five parallel activities:

1. Clinical trials,
2. Preclinical assessment,
3. Active pharmaceutical ingredient (API) development,
4. Drug product (DP) formulation, and
5. Analytical method development.

Table 1.1 Important sections of 21 CFR

Source	Title
Code of Federal Regulations, Title 21, Food and Drugs Administration (FDA), Part 210 (21 CFR 210)	Current good manufacturing practice in manufacturing, processing, packing, or holding of drugs
21 CFR 211	Current good manufacturing practice for finished pharmaceuticals
21 CFR 600	Biological products: general
21 CFR 820	Quality system regulations

Table 1.2 Useful regulatory statistical guidance ASTM international

Title	Chapter
E29: Standard practice for using significant digits in test data to determine conformance to specifications	2
E2281: Standard practice for process capability and performance measurement	5
E2475: Standard guide for process understanding related to pharmaceutical manufacture and control	4
E2587: Standard practice for use of control charts in statistical process control	5
E2709: Standard practice for demonstrating capability to comply with an acceptance procedure	7
E2810: Standard practice for demonstrating capability to comply with the test for uniformity of dosage units	7

Table 1.3 Useful regulatory statistical guidance Food and Drug Administration, Center for Drugs Evaluation Research (FDA,CDER)

Title	Chapter
Guidance for industry: immediate release solid oral dosage forms, scale-up and postapproval changes: chemistry, manufacturing and controls, in vitro dissolution testing, and in vivo bioequivalence documentation (1995)	7
Guidance for industry: demonstration of comparability of human biological products, including therapeutic biotechnology-derived products (1996)	9
Guidance for industry: SUPAC-MR modified release solid oral dosage forms, scale-up and postapproval changes: chemistry, manufacturing and controls, in vitro dissolution testing, and in vivo bioequivalence documentation (1997a)	7
Guidance for industry: dissolution testing of immediate release solid oral dosage forms (1997b)	7
Guidance for industry: ANDAs: blend uniformity analysis (1999 withdrawn 2002)	7
Guidance for industry: powder blend and finished dosage units—stratified in-process dosage unit sampling and assessment (October 2003 withdrawn 2013)	7
Guidance for industry: process validation: general principles and practices (2011)	3, 5, 6, 9
Guidance for industry: quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product (2015a)	9
Guidance for industry: scientific considerations in demonstrating biosimilarity to a reference product (2015b)	9
Guidance for industry: biosimilars: questions and answers regarding implementation of the biologics price competition and innovation act of 2009 (2015c)	9
Guidance for industry: analytical procedures and methods validation for drugs and biologics (2015d)	6

Figure 1.1 from Peterson et al. displays the timeline for these activities.

While the most common area for statisticians to work is in the clinical area (activities 1 and 2), the focus of this book is on paths 3–5 in Fig. 1.1. Key research questions and statistical methods used to help answer them are shown in Table 1.6.

Statistical quality control methods are applied throughout all activities in Phase IV. These methods are discussed in Chap. 5.

1.4 Differences Between Clinical and CMC Statisticians

To better understand the nature of CMC statisical analysis, it is useful to contrast this work to that of the clinical statistician. The role of a clinical statistician is well established. It is required and integrated into regulations and internal business processes. Often, these predefined roles and responsibilities are outlined in company procedures. Given the key role they play in the clinical drug development process, the clinical statistician is well linked into clinical project teams with strong management support. Among their many responsibilities, clinical statisticians are responsible for statistical design of clinical trials and statistical analysis plans included in protocols which are sent to the FDA for review. These protocols are

Table 1.4 Useful regulatory statistical guidance International Conference on Harmonization (ICH)

Title	Chapter
Q5C stability testing of biotechnological/biological products (1995)	8
Q1B photostability testing of new drug substances and products (1996)	8
Q1C stability testing for new dosage forms (1997)	8
Q6A specifications: test procedures and acceptance criteria for new drug substances and new drug products: chemical substances (1999a)	3, 7, 8
Q6B specifications: test procedures and acceptance criteria for biotechnological/biological products (1999b)	7, 8
Q7 good manufacturing practice guide for active pharmaceutical ingredients (2000)	4
Q1D bracketing and matrixing designs for stability testing of new drug substances and products (2002)	8
Q1A(R2) stability testing of new drug substances and products (2003a)	8
Q1E evaluation for stability data (2003b)	7, 8
Q3A impurities in new drug substances (2003c)	8
Q3B (revised) impurities in new drug products (2003d)	8
Q5E comparability of biotechnological/biological products subject to changes in their manufacturing process (2004)	2, 9
Q2(R1) validation of analytical procedures: text and methodology (2005a)	6, 8
Q9 quality risk management (2005b)	3–5
Q10 pharmaceutical quality system (2008)	3, 5
Q8(R2) pharmaceutical development (2009)	3, 5
Q11 development and manufacture of drug substances (chemical entities and biotechnological/biological entities) (2012)	3

Table 1.5 Useful regulatory statistical guidance United States US Pharmacopeial (USP)

Title	Chapter
⟨905⟩ Uniformity of dosage units	7
⟨1010⟩ Analytical data—interpretation and treatment	2, 6
⟨1030⟩ Biological assay chapters—overview and glossary	6
⟨1032⟩ Design and development of biological assays	6
⟨1033⟩ Biological assay validation	6
⟨1160⟩ Pharmaceutical calculations in prescription compounding	8
⟨1223⟩ Validation of alternative microbiological methods	6
⟨1224⟩ Transfer of analytical procedures	6
⟨1225⟩ Validation of compendial procedures	6
General notices 3.10: conformance to standards, applicability of standards	7
General notices 7.20: rounding rules	2, 7

very detailed and provide clear articulation of the exact analyses to be followed and the specific endpoints that must be met for clinical success. These protocols typically are based on regulatory requirements. The FDA has a team of statistical reviewers that evaluates the protocols and the definitive pass/fail nature of these

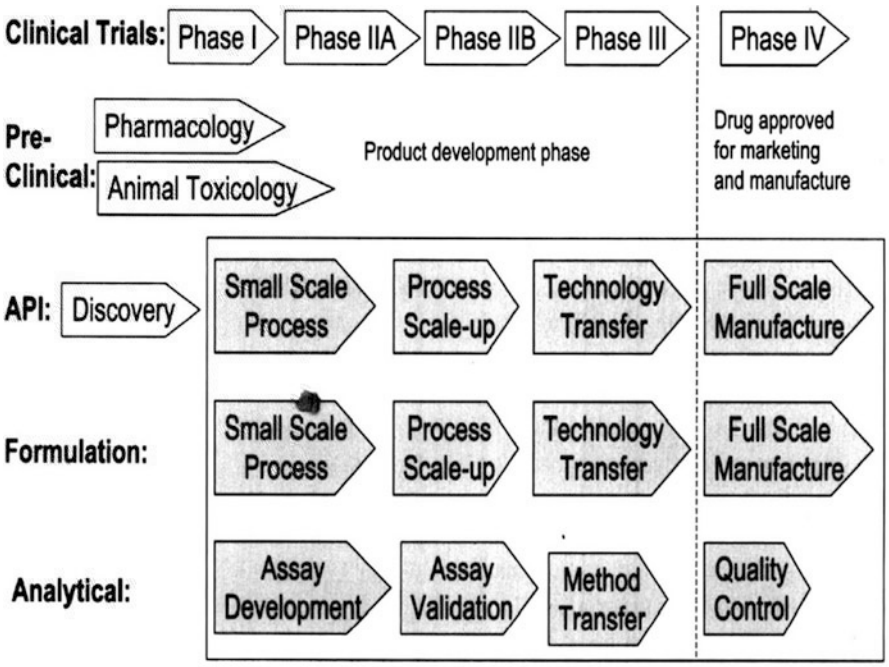

Fig. 1.1 Pharmaceutical activities

Table 1.6 Research questions and statistical methods

Activity	Research questions	Book chapter with application
API development	How can API be scaled to a level required for commercial market?	4
API development	How do we effectively characterize our product and develop a knowledge base for the API?	3, 4
DP formulation	What shelf life limits and specifications need to be established to ensure DP is safe and efficatious?	7, 8
DP formulation	If our process is transferred to another manufacturing site, how do we ensure product safety and efficacy are not impacted?	9
Analytical method development	How do we determine if analytical methods are fit for use?	6
Analytical method development	How do we know if a method will perform the same in two different labs?	6

criteria heightens the criticality of the statistician on the cross-functional team. There is much external guidance that must be followed and the clinical statisticians must follow strict documentation expectations. Data systems are well developed

and SAS and R are common data analysis tools. Other software packages are rarely used except for occasional exploratory work. Upon completion of the clinical study, the clinical statistician co-authors reports with clinical colleagues which are included in the regulatory submissions. Given the existing system, the clinical statistician does not spend a significant portion of time training their clinical colleagues to perform their own statistical analyses. Most of their interaction with colleagues involves discussions to help them understand and correctly interpret the statistical analyses completed by the statistician. In order to be the most successful, clinical statisticians must understand the science of the disease so that they can add value to the project team.

Although there are similarities between the roles of the CMC and clinical statisticians, there are differences in both type and degree. CMC statisticians work with scientists to help develop new drug substance and drug product manufacturing processes, develop and validate analytical procedures, improve existing processes and products, and troubleshoot systems when issues arise. Unlike the clinical statistician, the role of the CMC statistician does not have a regulatory requirement and as such, the nature of the role can vary both within and across companies. Factors impacting these differences include the technical and interpersonal skills of the statistician, the nature of management support, and the strength of the partnership with individual collaborators. Small, relatively short duration studies are common as opposed to large clinical trials and access to large data sets created to satisfy Good Manufacturing Practice (GMP) or regulatory requirements are the exception rather than the rule. Unlike the large and well-defined statistical departments in the clinical organization, CMC statisticians often work alone or in very small groups. The CMC statistician may report to management in the area they support, or to the broader clinical organization. Good documentation practices are important for the CMC statistician to adhere to GMP requirements but there are no statistical protocols sent to regulatory agencies for review. In the CMC area, studies are performed and documented internally so as to be available if requested by regulatory agencies. Documentation in these reports must be clear so that analyses can be explained and reproduced when necessary. Given that data sets are often small, data analysis packages such as Minitab and JMP are commonly used to perform calculations. SAS and R are also employed with larger data sets, or when requests are made from regulatory agencies. Because the CMC statistical workforce is relatively small, CMC statisticians spend time teaching their scientific collegues how to peform their own statistical analyses. This is one reason why statistical packages that do not require written code (e.g., JMP and Minitab) are often selected for analysis. Similar to the clinical statistician, the CMC statistician often contributes to the contents of a regulatory submission. CMC statisticians help write sections describing process and formulation development, stability, justification of specifications, process and product comparability, and analytical method validations. Similar to clinical statisticians, the CMC statistician must understand the science and engineering concepts of their collaborators in order to be successful.

1.5 How to Use This Book

It is possible to gain a working understanding of the methods in this book with no advanced statistical training. In fact, one objective of this book is to make these methods available for scientists who do not possess a degree in statistics. Professional statisticians will also find it helpful to have these methods in a single source for their own use and for training others.

Chapter 2 provides statistical methods that are useful for performing the analyses required to address research questions in the CMC manufacturing environment. We recommend that the reader begin by reading Sects. 2.1–2.5. This provides both a high level view of statistical applications and some specific examples for simple data sets. After reading this material, the reader may complete Chap. 2, or jump to any particular application of interest in Chaps. 3–9. Where needed, Chaps. 3–9 refer back to statistical methods in Chap. 2 where the reader is provided a more thorough understanding of the statistical method. Worked numerical examples are provided in all chapters.

The reader may use any number of statistical packages to help work the examples. Many of the examples can be performed using Excel. Some require user-friendly statistical packages, such as Minitab and JMP. Additionally, we have provided data sets and program codes written in SAS and R at the website for many of the examples. Discussions of the examples will focus on the output rather than specific code used to generate the output.

References

ASTM E29 (2013) Standard practice for using significant digits in test data to determine conformance with specifications. ASTM International, West Conshohocken

ASTM E2281 (2015) Standard practice for process capability and performance measurement. ASTM International, West Conshohocken

ASTM E2475 (2010) Standard guide for process understanding related to pharmaceutical manufacture and control. ASTM International, West Conshohocken

ASTM E2587 (2016) Standard practice for use of control charts in statistical process control. ASTM International, West Conshohocken

ASTM E2709 (2014) Standard practice for demonstrating capability to comply with an acceptance procedure. ASTM International, West Conshohocken

ASTM E2810 (2011) Standard practice for demonstrating capability to comply with the test for uniformity of dosage units. ASTM International, West Conshohocken

Chow S-C, Liu J-P (1995) Statistical design and analysis in pharmaceutical science: Validation, process controls, and stability. Marcel Dekker, Inc., New York

Code of Federal Regulations (CFR) (2003)Title 21, Food and Drugs Administration (FDA), Part 210, http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=210. Accessed 21 Sept 2015

Food and Drug Administration. Center for Drugs Evaluation Research (1995) Immediate release solid oral dosage forms, scale-up and postapproval changes: chemistry, manufacturing and controls, in vitro dissolution testing, and in vivo bioequivalence documentation, guidance for industry

Food and Drug Administration. Center for Drugs Evaluation Research (1996) Demonstration of comparability of human biological products, including therapeutic biotechnology-derived products, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (1997a) SUPAC-MR modified release solid oral dosage forms, scale-up and postapproval changes: chemistry, manufacturing and controls, in vitro dissolution testing, and in vivo bioequivalence documentation, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (1997b) Dissolution testing of immediate release solid oral dosage forms, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (1999) withdrawn 2002 ANDAs: blend uniformity analysis, guidance for industry.
Food and Drug Administration. Center for Drugs Evaluation Research (2003) withdrawn 2013 Powder blend and finished dosage units—stratified in-process dosage unit sampling and assessment, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (2011) Process validation: general principles and practices, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (2015a) Quality considerations in demonstrating biosimilarity of a therapeutic protein product to a reference product, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (2015b) Scientific considerations in demonstrating biosimilarity to a reference product, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (2015c) Questions and answers regarding implementation of the biologics price competition and innovation act of 2009, guidance for industry
Food and Drug Administration. Center for Drugs Evaluation Research (2015d) Analytical procedures and methods validation for drugs and biologics, guidance for industry
International Conference on Harmonization (1995) Q5C stability testing of biotechnological/biological products
International Conference on Harmonization (1996) Q1B photostability testing of new drug substances and products
International Conference on Harmonization (1997) Q1C stability testing for new dosage forms
International Conference on Harmonization (1999a) Q6A specifications: test procedures and acceptance criteria for new drug substances and new drug products: chemical substances
International Conference on Harmonization (1999b) Q6B specifications: test procedures and acceptance criteria for biotechnological/biological products
International Conference on Harmonization (2000) Q7 good manufacturing practice guide for active pharmaceutical ingredients
International Conference on Harmonization (2002) Q1D bracketing and matrixing designs for stability testing of new drug substances and products
International Conference on Harmonization (2003a) Q1A (R2) stability testing of new drug substances and products
International Conference on Harmonization (2003b) Q1E evaluation for stability data
International Conference on Harmonization (2003c) Q3A impurities in new drug substances
International Conference on Harmonization (2003d) Q3B (Revised) impurities in new drug products
International Conference on Harmonization (2004) Q5E comparability of biotechnological/biological products subject to changes in their manufacturing process
International Conference on Harmonization (2005a) Q2 (R1) validation of analytical procedures: text and methodology
International Conference on Harmonization (2005b) Q9 quality risk management
International Conference on Harmonization (2008) Q10 pharmaceutical quality system
International Conference on Harmonization (2009) Q8 (R2) pharmaceutical development

International Conference on Harmonization (2012) Q11 development and manufacture of drug substances (chemical entities and biotechnological/biological entities)
Peterson JJ, Snee RD, McAllister PR, Schofield TL, Carella AJ (2009) Statistics in pharmaceutical development and manufacturing. J Qual Technol 41(2):111–147
USP 39-NF 34 (2016a) General chapter <905> uniformity of dosage units. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016b) General chapter <1010> analytical data—interpretation and treatment. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016c) General chapter <1030> biological assay chapters—overview and glossary. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016d) General chapter <1032> design and development of biological assays. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016e) General chapter <1033> biological assay validation. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016f) General chapter <1160> pharmaceutical calculations in prescription compounding. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016g) General chapter <1223> validation of alternative microbiological methods. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016h) General chapter <1224> transfer of analytical procedures. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016i) General chapter <1225> validation of compendial procedures. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016j) General notices 3.10: conformance to standards, applicability of standards. US Pharmacopeial Convention, Rockville
USP 39-NF 34 (2016k) General notices 7.20: rounding rules. US Pharmacopeial Convention, Rockville
Zhang L (ed) (2016) Nonclinical statistics for pharmaceutical and biotechnology industries. Springer, Heidelberg

Chapter 2
Statistical Methods for CMC Applications

Keywords Analysis of variance • Bayesian analysis • Confidence intervals • Data reporting • Data rounding • Data transformations • Dependent measures • Equivalence testing • Hypothesis testing • Interaction effects • LOQ values • Mixed models • Multiple regression • Nonlinear models • Non-normal data • Prediction intervals • Quadratic effects • Regression analysis • Residual analysis • Statistical consulting • Statistical intervals • Tolerance intervals • Visualization of data

2.1 Introduction

In this chapter, we provide statistical methods that are useful in CMC applications. Our goal is to provide a description of these methods without delving deeply into the theoretical aspects. References are provided for the reader who desires a more in depth understanding of the material.

2.2 Statistical Analysis in a CMC Environment

As described in Chap. 1, CMC statisticians work directly with individual subject matter experts (SME) from other areas of science. In this section, we describe a typical example of this interaction from the viewpoint of the statistician. Here "statistician" is defined as the person who is responsible for performing the required statistical analyses. This need not be a person with a terminal degree in statistics. The term "client" is used to represent the SME requiring help in performing the statistical analysis. The statistical analysis process consists of the following four steps:

1. Initial client meeting.
2. Planning of statistical analysis.
3. Data analysis.
4. Communication of results to the client.

Each of these steps is now described in the context of an example.

R.K. Burdick et al., *Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry*, Statistics for Biology and Health, DOI 10.1007/978-3-319-50186-4_2

2.2.1 Initial Client Meeting

The initial meeting with a client is where everything begins. The goal for the statistician is to listen to the client and learn about the problem. The statistician should listen for at least 80% of the conversation, and ask clarifying questions for the other 20%. This is not the time to overwhelm the client with statistical jargon and tales of the sophisticated power of a pet statistical procedure. One of the benefits of being a statistician is that you have the opportunity to work with clients in a variety of fields. So take advantage of these sessions and learn something new. You will have an opportunity later in the process to share your expertise in statistics.

An example of a typical initial meeting is shared below. The statistician, Tom, has met the client, Noel, in Noel's office. Noel is a product quality leader who is responsible for all quality issues related to a particular product.

Tom: Good morning, Noel. How have you been doing?
Noel: Fine. Just trying to keep up with all the work.
Tom: Tell me about it! (Note: There is always the obligatory greeting that suggests both parties are the hardest working people in the company.) What can I help you with today?
Noel: As you know, we are transferring our manufacturing process to Ireland, and are in the middle of the comparability phase of the transfer. We need to demonstrate to regulatory agencies that once the process is operating in Ireland, it will be manufacturing product of a similar quality to our present process with no risks to patient safety or product efficacy.
Tom: Right. I have helped on these types of projects in the past. What are some of the details?
Noel: We have produced eight lots in Ireland and want to compare the lot release values for some of our quality attributes with those we have collected from our process here in the USA. The two processes don't have to be exactly the same, but as I have already mentioned, we cannot compromise patient safety or product efficacy. One thing we do to ensure our present process continues to provide safe and efficacious product is to demonstrate that the average purity is no less than 93%. Anything less than 93% would be sufficiently different from our present process that I would be concerned.
Tom: Thanks. So if our process in Ireland has a mean of 93% or greater, we believe the new process is operating as expected. Right?
Noel: Exactly.
Tom: What data do we have available?
Noel: Purity is measured using the reversed-phase high performance liquid chromatography (RP-HPLC) main peak. I have these values for each of the eight lots produced at the Ireland plant. Do you think eight values is a large enough data set to draw any meaningful conclusion?
Tom: Well, eight lots is better than seven lots but not as good as nine lots. We can perform calculations with the available eight lots, but the uncertainty may be too

large to provide meaningful results. You will have to make that determination. If there is too much uncertainty to be useful, you will need more data.

Noel: Right. We would like to know now if there are any apparent issues at this point, so let's see what the eight lots tell us.

Tom: Sounds good.

Noel: I will get the data to you as soon as possible. I know you like the data in a certain format. Can you remind me how to prepare it for you?

Tom: Thanks, Noel. It lessens the opportunity for errors if you get me the data in the proper format. I would like the data placed in two columns of an Excel spreadsheet. The first column will be the lot number and the second column will be the purity value. The first row of the spreadsheet will have the label for each column and then there will be eight rows of data. Make sure to report the recorded values for your measurements.

Noel: That doesn't sound too hard. I will get the data verified and then place the spreadsheet in the company information system for you to access.

Tom: That sound's great. What kind of timeframe do we have?

Noel: It should take me about a week to get everything verified, approved, and into the system. If you could do your magic within a week after receiving the data, it would give us time to assess the results and move on to the next step.

Tom: That sounds workable. Please send me a note when the data are ready. Talk to you later.

Noel: Thanks, Tom. I look forward to talking to you again in a couple of weeks.

2.2.2 Planning of Statistical Analysis

After meeting with the client, it is time to formulate a strategy for answering the research question. To do this, it is often necessary to make a statistical inference. Statistical inference concerns the ability to answer a question based on a collected data set. The research question concerns a collection of items called the population or process output. For purposes of our discussion, a population is a finite collection of items, such as all vials within a manufactured lot of drug product. A process is a series of actions or operations used to convert a set of inputs into a set of outputs. The process output over a fixed interval of time constitutes a population of items. When thinking of a process, we often want to make a statement about future items that will result if the process operates in a manner consistent with the observed data. In CMC applications, we are often interested in process outputs where one set of outputs is created with an existing process and the other set of output is created with a new or improved process. The planning of a statistical analysis should consider three components:

1. Statement of the study objective.
2. Data acquisition.
3. Selection of a statistical tool.

2.2.2.1 Statement of the Study Objective

One result of the initial client meeting is an understanding of the research question. In order to develop an effective statistical strategy, it is necessary for the statistician and client to agree on a clear research objective. In working with the client, it is sometimes helpful to ask them how they intend to use the results of the statistical analysis. For example, ask the client what action she would take if presented with a certain outcome of the statistical analysis. Similarly, ask the client what action he would take based on an alternative outcome. If such questions cannot be answered, then the research question is not well-defined, and more discussion is needed to clarify the study objective. The research question in our present example is "Does the Ireland manufacturing process operate at an acceptable capability?" One piece of information that will be used to help answer this question is the process mean for purity. Once the research question has been clearly defined, the next step is to determine the most appropriate data to answer the question.

2.2.2.2 Data Acquisition

The following questions are worth consideration in any discussion of data acquisition.

1. What is the population or process of interest, and how can we ensure the collected data are representative of this group?
   Require an exact definition of the population/process including the time period of interest. In the present example, Noel wants to know something about long-run behavior of a process for which eight items presently exist. If it is planned to run the same process in the future, then the eight available lots are likely representative of future output. The assessment of whether a data set is representative is *not* a statistical question. It requires the judgment of an expert in the field of application (i.e., Noel in our example).
2. Do the data already exist (observational), or will we have to create it experimentally?
   Although the answer to this question is often based on convenience or timeliness, it is important to understand the different types of inference that can be made with each type of data. Generally speaking, inferences of a causal nature require experimental data. That is, if one wishes to provide evidence that changes in factor X cause change in factor Y, then an experimental data set is required in order to properly isolate the relationship of Y and X and protect it from other factors. Although observational data cannot directly demonstrate causality, it does provide a description of how variables relate to each other in the "real world." To better explain in the context of our example, the lot release data is observational since it was collected as part of the manufacturing process. It does not require a separate set of experimental studies to generate the data. In the analysis that follows, we are able to estimate the mean purity of the Ireland

process. However, if it is lower than desired, we have not learned anything to help us determine how to increase the mean and improve the process. To do this, we would likely need a set of experimental studies where inputs of the process are systematically changed to determine the impact on the mean purity. Data generated in this manner is an example of experimental data. In the CMC world, experimental data is generally required in process development, whereas observational data are used in quality operations.

3. What sampling method is used to collect the data?
   It is necessary that the data used for analysis be collected using a valid statistical sampling procedure. There are many different types of sampling procedures, and the best approach in any given situation will depend on the structure and availability of the data. Natural groupings of population items often impact the manner in which a statistical sample is selected. Consider the following example. As part of a method validation study, four plates are prepared, each one containing three aliquots. Measurements made on the three aliquots in the same plate are typically more similar than measurements of aliquots in different plates. For this reason, the statistical analysis must account for this relationship. Had the 12 measurements been collected using 12 plates with one aliquot each, a different statistical analysis would be performed. The larger physical units (plate in this example) are called experimental units, and the smaller units (aliquots) are called observational units. Observational units are what are actually measured. In many situations, the experimental unit is the observational unit. As noted, it is always important to identify any grouping of observational units in order to perform the correct statistical analysis.
4. How are the variables defined and measured?
   A variable provides information of interest about each individual item in a population. Variables can provide a number (quantitative variable) or a category (qualitative variable). It is important that all variable definitions be included in the statistical report. In our present example, the variable assigns a number that represents lot purity as measured by the RP-HPLC main peak. We might label this variable "Purity." Each variable in a study must be measured for each sampled item.
   When one measures the value of a continuous variable, the measured value is never *exactly* equal to the true value. The difference between the measured value and the true value is called measurement error. We say that a measurement procedure is unbiased if there is no systematic tendency for the procedure to underestimate or overestimate the true value. For example, if the weight of 100 people were measured using an unbiased scale, we would expect the reported weight to be greater than the true value for some people and less than the true value for others. On average, the error in measurement would be essentially zero. However, if the scale is not calibrated properly, it might consistently yield readings that are lower (or higher) than the true weights. In such a case, the scale is said to be biased.
   In addition to having a small amount of bias, a good measuring device is precise. Precision concerns the reliability of a measurement procedure. If one measures

the same item several times, it is hoped that the measurements are more or less the same each time and do not fluctuate wildly. Measurement procedures that exhibit little variability when measuring the same item are said to be precise. One of the reasons that analytical procedure validations are performed is to ensure that the bias and the precision are acceptable. Chapter 6 concerns the validation of analytical procedures.

Another question to ask concerning quantitative variables is whether the numerical values are recorded to enough decimal places to allow a meaningful analysis. It is recommended to always perform calculations with recorded data, rather than data that have been reported to a lesser degree of accuracy. More on this topic is provided in Sect. 2.3.

5. What sample size is required to provide a useful analysis?
   This is probably the most frequent question asked of a statistician. The statistician can never answer this question in isolation. It always requires interaction between the client and the statistician. The thing to remember is that as sample size increases, then uncertainty in the statistical inference will decrease. At some point, one must balance the cost and time of selecting additional sample items against the risks associated with an uncertain decision. In many CMC applications, sample sizes are small because a single item, such as a production lot, is produced infrequently or is extremely expensive. Valid statistical analyses can be performed with relatively small data sets, but it is important to report the level of uncertainty in order to fairly represent the usefulness of the conclusions.

2.2.2.3 Selection of a Statistical Tool

Once the research question is well defined and the relevant data have been identified, it is time to select a statistical procedure to help answer the research question. For the Ireland manufacturing problem discussed earlier, it is necessary to estimate the mean of the new process using the data set of eight values. We will use the sample mean of eight production runs to provide a point estimate of the true process mean. A point estimate is useful, but we need to recognize that it is based on only eight process lots. Thus, a point estimate has uncertainty associated with it. One quantifies this uncertainty by computing a statistical interval that provides a range of possible values. The various types of statistical intervals are introduced in Sect. 2.5. For the present example, a confidence interval is the appropriate statistical interval. A confidence interval contains the true unknown value of the purity mean with an associated level of confidence (e.g., 95%). The confidence level of 95% describes the ability of the confidence interval to correctly capture the true value of the purity mean. More discussion is provided on the interpretation of the confidence coefficient in Sect. 2.5.1.

In our example, if all values in the confidence interval exceed 93%, we will conclude that the new process has attained the desired mean. Now that our strategy has been determined, we wait for the data.

2.2.3 *Data Analysis*

Data analysis consists of both graphical representation of the data and numerical description. Data analysis consists of the following operations:

1. Obtain the data.
2. Plot the data.
3. Estimate the unknown quantities of interest.
4. Quantify uncertainty in point estimates using a statistical interval.

We demonstrate this process within the context of the previous example.

2.2.3.1 Obtain the Data

The big day arrives and Tom receives an email that the data are ready for analysis. The data are presented in Table 2.1.

True to his word, Noel provided the data in a format that is amenable for performing the statistical analysis. Each column in the table represents a variable. The name for each variable is found in the first row. For example, the first column in Table 2.1 reports values for the variable "Lot." The variable "Lot" is a qualitative variable because the values assigned are categories. The variable "Purity" in the second column is quantitative or numerical. Since the unit of measurement for purity is a percentage, it can assume any numerical value between 0 and 100%, inclusive. Each lot is represented by a single purity value as reported by the lab.

2.2.3.2 Plot the Data

Lynne Hare is a well-known statistician who established a rule that should be followed in every data analysis. The rule is termed the "Always, always, always--without exception rule" (AAAWER). The AAAWER is simple—Plot the data. Some years after creation of the AAAWER, he added the corollary—"and look at it." It is an old adage that a picture is worth a thousand words, but it is really true in

Table 2.1 Purity measures (%) from the Ireland process

Lot	Purity (%)
A	94.20
B	92.68
C	94.47
D	94.14
E	95.17
F	94.47
G	94.14
H	95.17

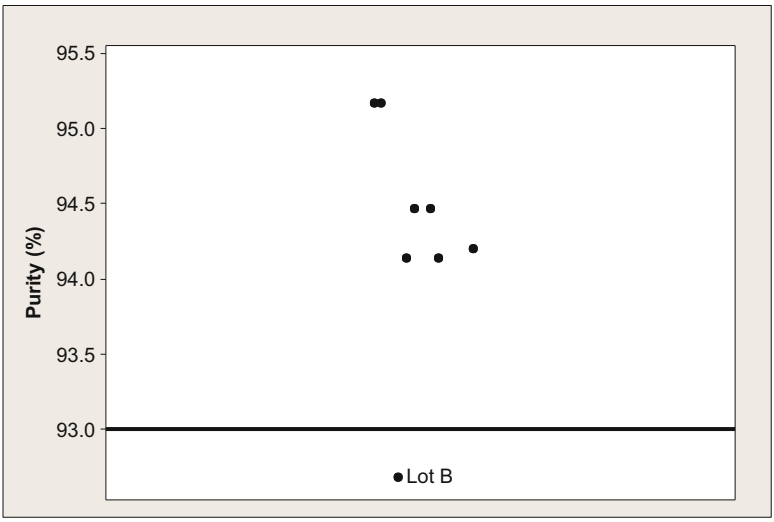

Fig. 2.1 Plot of purity data (%)

data analysis. Most learnings are obtained by simply plotting the data (and looking at the plot). We begin the analysis of the purity data by constructing a plot. The plot is shown in Fig. 2.1.

Recall that the objective of the study is to determine if the process mean in Ireland is at least 93%. Looking at the plot, all of the purity values except one exceed 93%, so it is clear that the mean of the eight values exceeds 93%.

Note that the purity values are all different. It is useful when looking at data to ask yourself the question, "Why aren't these values all the same?" There are two primary reasons why these eight values are not all equal.

1. There is expected variation from lot-to-lot for any manufacturing process.
2. There is measurement error in the data.

As in this example, most variables encountered in CMC applications are impacted by variation from both the manufacturing process and the analytical (measurement) method.

Another aspect to note about the plot is that Lot B is somewhat removed from the other seven lots. Points that are inconsistent with the rest of the data are called outliers. Outliers may arise due to coding errors, or perhaps due to some unique event that occurred during the manufacturing of a particular lot. The SME should be consulted immediately to discuss possible reasons for any detected outlier. Sections 2.4 and 2.12.2 provide more discussion on the statistical definition and identification of outliers.

Table 2.2 Definition of parameters

Parameter symbol	Interpretation
μ (Lower case Greek letter "mu")	Process mean at Ireland site
σ^2 (Lower case Greek letter "sigma" squared	Process variance at Ireland site
σ (Lower case Greek letter "sigma")	Process standard deviation at Ireland site

2.2.3.3 Estimate the Unknown Quantities of Interest

As discussed earlier, one objective of a statistical analysis is to make an inference (conclusion) about an unknown characteristic of a population or process. Both populations and processes are described by summary measures called parameters. The parameter of interest in the example is the process mean. The values of the eight completed lots represent a sample of all process outcomes. One can compute the mean for the eight lots in Table 2.1, but since this is only a sample, one cannot know with certainty the exact value of the true process mean.

It is a convention in the statistical literature to represent an unknown parameter value with a Greek letter. Table 2.2 shows Greek notation used to represent three parameters in the example. Think of the Greek symbol as you would an acronym. It is simply a shorthand notation to replace all the words in column 2 of Table 2.2.

Let the measured value of the variable "Purity" for lot i be represented as Y_i. Since there are eight lots, the index i goes from 1 to 8, inclusive. The sample of eight measured values is represented as $Y_1, Y_2, \ldots, Y_8$. The sample mean is computed by adding all eight values and dividing by the number of values (8). Representing as a formula we have

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{94.20 + \cdots + 95.17}{8} = 94.305\% \tag{2.1}$$

where n is the sample size. Summary measures computed from a sample such as $\bar{Y}$ are referred to as statistics. A point estimator is a function of one or more statistics used to provide a "best guess" of the true (unknown) parameter value. For example, the sample mean $\bar{Y}$ is the point estimator of the true and unknown process mean, μ. The realized value of $\bar{Y}$, 94.305, is the point estimate for μ.

A point estimator for the process standard deviation, σ, is the sample standard deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{(94.20 - 94.305)^2 + \cdots + (95.17 - 94.305)^2}{8-1}}$$
$$= 0.780\%. \tag{2.2}$$

The square of the standard deviation is called the variance, and S^2 is the point estimator for σ^2.

The point estimate for the process mean, 94.305 %, exceeds the required value of 93%. However, we must ask the question, "How close to the true value of μ is the estimate?"

2.2.3.4 Quantify the Uncertainty in the Point Estimates Using a Statistical Interval

Point estimates are useful, but they do not tell us anything about the uncertainty associated with the estimates. Recall we are sampling only eight lots from the process, and we wish to make an inference concerning future lots. The best way to quantify the uncertainty in this example is to compute a statistical confidence interval. For this example, the following formula provides a 95% lower bound on the true process mean (see Eq. (2.8) later in this chapter).

$$L = \bar{Y} - 1.895 \times \frac{S}{\sqrt{n}}. \tag{2.3}$$

In our example,

$$L = 94.305 - 1.895 \times \frac{0.780}{\sqrt{8}} = 93.78\%. \tag{2.4}$$

The 95% lower bound exceeds the 93% criterion, providing statistical evidence that the process mean will exceed 93% if the process continues to operate in the future as it has in the past.

2.2.4 Communication of Results to Client

The data analysis is complete and now it is time for another visit with the client. Unlike the first visit where the statistician took the role of listener, the statistician now assumes the role of teacher.

Tom: Good morning, Noel. How have you been doing?
Noel: Great. Good to see you again, Tom. What can you tell me about the analysis?
Tom: As you will recall, we wanted to see if the new process has a mean greater than or equal to 93%. I looked at the sample of eight lots you sent me and computed an estimate of the new process mean. My conclusion is that the new process is expected to have a mean no less than 93.78% with 95% confidence. Since 93.78% exceeds your requirement of 93%, we can expect the new process to deliver the desired level of quality.