# Contents

About This Book ................................................................. vii  
About the Author ................................................................. xi  
Acknowledgments ............................................................... xiii  

Chapter 1 / Preface ................................................................ 1  
Preface ............................................................................... 1  

Chapter 2 / Introduction ....................................................... 3  
Data Analysis Plans ............................................................... 3  

Chapter 3 / Before the Plan: Develop an Analysis Policy ....... 5  
Summary ........................................................................... 5  
The Analysis Policy ............................................................ 6  
Analysis Policy Components ............................................. 7  
Analysis Policies for Large Projects ................................... 11  
Example Project Analysis Policy ......................................... 11  
References ........................................................................ 14  

Chapter 4 / Outline of a Plan ................................................ 17  
Summary ........................................................................... 17  
Outline of a Typical Plan ...................................................... 18  

Chapter 5 / The Plan Introduction ........................................ 21  
Summary ........................................................................... 21  
Objectives, Hypotheses or Questions, and Aims ................... 22  
Identify the Project Team .................................................... 24  
Identify the Target Audience .............................................. 25  

Chapter 6 / Hypotheses, Questions, and Study Design .......... 27  
Summary ........................................................................... 27  
Conceptualizing the Problem .............................................. 28  
Hypotheses ......................................................................... 29  
Research Questions ........................................................... 29  
Study ................................................................................ 29  
Identifying Confounders ..................................................... 31  
Communicating with the Study Team .................................. 32
## Contents

### Chapter 7 / Data Description
- Summary .................................................. 33
- Data Sources ................................................. 33
- Data Definitions ............................................ 35
- Data Summary ............................................... 36
- Reviewing the Data with the Study Team .................. 37

### Chapter 8 / Data Exploration
- Summary .................................................. 39
- Descriptive Statistics for Major Variables ................. 40
- Exploring Outliers and Problematic Data Distributions .... 42
- Distribution Problems ....................................... 46
- Determining the Source of Variability ...................... 49
- Producing Baseline Tables ................................ 55
- Communicating with the Project Team .................... 57

### Chapter 9 / Analysis
- Summary .................................................. 59
- General Approach .......................................... 60
- Analysis of Questions/Hypotheses .......................... 63
- References .................................................. 69

### Chapter 10 / Potential Conclusions, Study Weaknesses, and a Timeline
- Summary .................................................. 71
- Potential Conclusions ...................................... 71
- Limitations and Weaknesses ................................ 72
- Timeline .................................................... 72

### Chapter 11 / Revising, Producing and Sharing the Plan
- Summary .................................................. 73
- Communicating and Revising the Plan ...................... 73
- Turning the Plan into a Report ............................. 74
- Developing Tabular Presentations ......................... 74
- Cross-References .......................................... 75
- Unplanned Analysis ........................................ 75
- Closing Remarks ............................................ 76

### Chapter 12 / Example Analysis Plan: The TIAD Study
- Summary .................................................. 77
- Introduction ............................................... 78
- Questions ................................................... 79
- Data Description ............................................ 79
The Plan Introduction

Summary

Objectives, Hypotheses or Questions, and Aims

Introduction

Objectives

Hypotheses

Questions

Aims

Identify the Project Team

Identify the Target Audience

Summary

Set the stage for the analysis plan by stating the objective or purpose, hypotheses or questions to be addressed, and the specific aims of the investigation. In addition, the introduction identifies who will work on the study and in what capacity, and where the study results will ultimately be presented if public presentation is appropriate and expected. As an analyst, you may have already discussed these matters with the project team. But putting them in writing and in your own words is an opportunity to clarify any misunderstandings. Often members of the team have different ideas about the scope of the study, so it is important to agree on the aims and objectives up front. Moreover, this gives the analyst an opportunity to step foot into the realm of the investigative team’s field. Much of this information may have been presented somewhat vaguely in a proposal, but re-stating it here in concise but specific language will help to focus the team and the upcoming work.
Objectives, Hypotheses or Questions, and Aims

Introduction

Objectives are the broader reasons or purposes for the undertaking. Hypotheses are proposed explanations for observable phenomena. Research questions pertain to exploratory studies where the outcome is not necessarily predictable. Aims are the general methods or approaches that will be used to test the hypotheses or explore the research questions.

Objectives

Open with a brief statement about the purpose of the inquiry. The objective or purpose of the study is its potential significance and impact on the field in the case of scientific investigation, or on the client’s strategy or operations if you are working for a commercial or advocacy group. Here are some examples of objectives:

- Test a new drug intervention to reduce blood-sugar levels in patients with diabetes.
- Determine which is the best among common methods for eradicating the mosquito vector for dengue fever.
- Determine the most effective outreach method for encouraging residents of impoverished villages in sub-Saharan African countries to register for a new United Nations assistance program.
- Determine which states are best candidates for federal-state partnerships to pilot a new military veterans benefits program intended to replace several existing programs.

Follow the objective with a statement of the significance of the problem noting the relevance of this work for the field in which the study is being conducted. A summary of the current knowledge with references provides a foundation for the rationale for conducting the investigation. Identify any gaps in knowledge that this study might address, and convey how the work will help to fill such gaps.

Hypotheses

Studies are set up to address either hypotheses, which are based on existing theories, or research questions, which are purely exploratory in nature. It is more common for research projects to be based on hypotheses because they are more closely tied to established science, which means they are driven by previously established knowledge or best practices. They are more likely to contribute to a body of knowledge, and a statistical probability can be estimated for their findings.
Most people tend to think of hypotheses as being associated only with “scientific” studies. However, explicitly stated hypotheses are necessary and appropriate for any study where there is an expected outcome that can be refuted by evidence, even if the domain of investigation is not a classic field of science. For example, your client may be an advocacy group focused on resettlement of immigrant refugees. Studies of other populations have suggested to them that a new method of outreach through social media is more effective than traditional media in encouraging refugees to register for available services. With proper data collection and analysis, this may be a refutable claim and thus qualifies as a hypothesis. Some example hypotheses are as follows:

- Drug A can safely and effectively reduce blood sugar in diabetic patients.
- Pamphlets distributed at food markets are more effective than broadcast or print media advertisements in increasing the rate of pre-registration for a new vaccine in impoverished villages.

**Questions**

Not all endeavors have an expected outcome that is refutable. Some are simply exploratory in nature, and research questions are more suitable than hypotheses. Some examples are as follows:

- In lower income neighborhoods, which intervention is most effective at reducing the mosquito vector for dengue fever: 1) an educational program aimed at preventing standing pools of water, 2) subsidies for mosquito nets and window screen maintenance, or 3) traditional chemical fogging programs?
- Which states have combined military veteran populations that are most representative of the national veteran population, and whose combined veteran populations suggest that pilot participation of those states would approach but not exceed the federal dollars allotted for a five-year pilot program?

**Aims**

An aim describes how the study team will address the hypothesis or question. There may be more than one aim per question. However, each aim should be directly linked to a specific hypothesis or question. The following aims might be appropriate for the example hypotheses and questions, above:

- We will conduct a randomized clinical trial of drug A versus currently accepted treatments with drugs B and C to determine if A is equally safe yet more effective in reducing blood-sugar levels in pre-diabetic patients.
- We will conduct a randomized trial of 25 remote villages, clustered by tribal affiliation, to determine which of three outreach methods result in the highest rate of pre-registration for a new vaccine: 1) pamphlets distributed by community merchants at food markets, 2) local radio advertisements, or 3) print media advertisements.
We will conduct a prospective cohort study of lower income neighborhoods to determine which of three interventions is most effective at reducing the mosquito vector for dengue fever: 1) an educational program aimed at preventing standing pools of water, 2) subsidies for mosquito nets and window screen maintenance, or 3) traditional chemical fogging programs?

We will conduct an iterative series of observational analyses and actuarial projections to identify the set of U.S. states most suitable for a five-year federal-state partnership pilot for a veterans benefit program. Observational statistics from existing veterans databases will serve as baseline data for the actuarial projections of total veterans who meet the demographic and socioeconomic eligibility criteria and are who are most likely to enroll in the new program. Iterative analyses will proceed until the set of states is identified that best represents the national veteran population and whose projected pilot program costs will not exceed allocated funds.

### Identify the Project Team

In a table, list the full name, degrees, and affiliations of each member. Also list the role that each one is expected to fulfill. In most circumstances, it is best to identify the project lead, who is usually the main author if publications are to ensue. Stating the names and roles of each member at the outset avoids problems down the road.

It is surprising how common are misunderstandings about roles. An illustration in academia is assumptions that are made about authorship of publications resulting from the work. For example, junior statisticians who assist in analysis may hope for or expect to be credited as authors while other team members may think differently. Less experienced clinicians may not envision authorship for any of the statisticians. Senior team members may not plan to credit the data collection and processing staff. Journals sometimes provide guidance for defining authorship-worthy contributions. If so, the guidance should be consulted at this time.

Beyond credits, the project team table should define responsibilities for each team member. The responsibilities included in the table should be as comprehensive as possible. For example, if a data collection and processing staff is identified as well as the statisticians/analysts, the two groups may assume that the other will take responsibility for data cleaning unless arrangements are made in advance and indicated in the table.

The project team table can be expanded to include other information as appropriate. This might be contact information, estimates of level of effort (e.g., full-time equivalencies), and the identity of back-up or support staff for some or all roles.

The table below presents some common roles and responsibilities to consider including. Small projects may have only two or three persons listed and only a handful of roles and responsibilities dispersed among them. Large projects may list many persons and include roles, responsibilities, and other details not listed here. The responsibilities listed here are only notional and may be redistributed among roles or changed as appropriate for your situation.
<table>
<thead>
<tr>
<th>Role</th>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal investigator/project director</td>
<td>Principal subject matter expert; responsible for the overall project and final product(s); primary author of publications; responsible for adherence to human subjects research requirements</td>
</tr>
<tr>
<td>Co-investigator</td>
<td>Additional subject matter expert; assists principal in all responsibilities</td>
</tr>
<tr>
<td>Lead statistician/analyst</td>
<td>Develops study design; drafts analysis plan; leads or conducts all quantitative analyses; interprets findings for investigators; writes methods and findings sections of reports and publications</td>
</tr>
<tr>
<td>Junior statistician</td>
<td>Assists lead statistician in all matters; performs final data cleaning efforts</td>
</tr>
<tr>
<td>Data collection roles</td>
<td>Collects raw data (e.g., laboratory, survey, or clinical data) or acquires secondary data (e.g., public or commercial files for secondary data analysis)</td>
</tr>
<tr>
<td>Data preparation and processing roles</td>
<td>Converts data to a form usable by programmers and/or statisticians; responsible for initial data cleaning</td>
</tr>
<tr>
<td>Programmer</td>
<td>May write and execute computer programs for any project stage or task, including data capture, processing, cleaning, analysis, or reporting.</td>
</tr>
<tr>
<td>Editor/illustrator</td>
<td>Edits manuscripts and reports; assists principal investigator with publication</td>
</tr>
</tbody>
</table>

**Identify the Target Audience**

An ultimate goal of any investigation is to disseminate the results to some audience. However, just as unstated assumptions about roles can lead to trouble downstream, so too can misunderstandings about the targeted audience. The target audience may be a small group internal to your organization, the management staff of the client’s commercial enterprise, academics or clinicians in a particular branch of study, public policy decision makers, or even the general public.

While disagreements regarding audience scope are rare, the details of dissemination are often subject to differences. For example, the client or principal investigator may want to
exercise complete control over whether and how results are disseminated. He or she may wish to bury negative findings or findings that are contrary to his or her beliefs. Alternatively, if the findings are auspicious he or she may want to attempt publication in a highly prestigious journal that is not likely to take interest in a modest study. Analysts and statisticians, on the other hand, may wish to publish any findings in any journal in order to grow their curriculum vitae and support academic job security.

These can be difficult matters to discuss, and hence are too often put aside. We strongly recommend that they be openly addressed in the introductory section of the analysis plan, even to the extent of identifying appropriate journals or other venues, reviewing the requirements of those venues, and summarizing why the results might be of interest to the targeted audience.
Adolescent Medicine Trials Network (ATN) for HIV/AIDS Intervention

aims 22, 23
analysis about 59
general approach to 60
of questions/hypotheses 19, 63, 83
sensitivity 69
subgroup 69
unplanned 75
analysis plan example 78
analysis policy about 6
components of 7
example project 11
for large projects 11
for TIAD study 82
ANOVA procedure 60

baseline tables
for TIAD study 80
producing 55
BY processing 74

cases 30
Chi-square test 80
co-investigator, responsibilities for 25
COLLINOINT option 68

communications, importance of 73
conclusions, potential 71
confounding variables about 67
identifying 31
CONTENTS procedure 35
Cook's D plot 53
covariates 67
cross-references 75

data reading 34
reviewing with study team 37
data analysis plans about 3
as a component of analysis policies 8
types of 3
data analysis proposals, as a component of analysis policies 8
data collection roles 25
data definitions 35
data description about 33
as a typical plan component 18
data definitions 35
data in hand 34
data not in hand 33
data sources 33
data summary 36
for TIAD study 79
reading data 34
reviewing data with study team 37
data exploration about 39
as a typical plan component 18
communicating with project team 57

descriptive statistics for major variables 40
determining source of variability 49
distribution problems 46
outliers 42
problematic distributions 42
producing baseline tables 55
data in hand 34
data not in hand 33
data policies, as a component of analysis policies 8
data preparation and processing roles 25
data sources 33
data summary 36
data tables, for TIAD study 80
descriptive statistics, for major variables 40
detailed analysis plan 3
distribution problems 46

editor, responsibilities for 25
example project, analysis policy 11
experimental studies 29

Fisher's exact test 80
FREQ procedure 36, 41, 55

Gantt chart 72
GANTT procedure 72
general statistical approach
as a typical plan component 18
for TIAD study 82
GLIMMIX procedure 63
GLM procedure 52, 60

GROUPMEANS option 65

HbA1c response, for TIAD study 83
HOLM option 61
HOVTEST option 51
hypotheses/questions about 22, 23, 27, 29
analysis of 63, 83
as a typical plan component 18
conceptualizing the problem 28
for TIAD study 79
primary 63
research 29
secondary 69

illustrator, responsibilities for 25
inclusion 30
intention to treat (ITT) analyses 10, 79, 83
Internal Review Board (IRB) 9
introduction about 21
aims 22, 23
as a typical plan component 18, 21
hypotheses 22
objectives 22
questions 22, 23

junior statistician, responsibilities for 25

large projects, analysis policy for 11
Index 89

lead statistician/analyst, responsibilities for 25
limitations, of studies 72
longitudinal data analysis, for TIAD study 82

M

mean 48
MEANS procedure 36, 40, 55, 62
MEANS statement 50, 51
median 48
MI procedure 63
missing values 62
mission statement, as a component of analysis policies 7
MIXED procedure 63
MULTTEST procedure 60

N

National Institute of Child Health and Human Development (NICHD) 11
NMISS option, MEANS procedure 62
NORMAL option 43
NPERG= option 65
Nurses Health Study 11

O

objectives 22
observational studies 29
ODS statement 74
outable= option 36
outcomes for TIAD study 79
specifying 31
outliers, exploring 42
Output Delivery System (ODS) 36, 74

P

parameters, statistical 60
plans about 17
outline of 18
revising 73
turning into reports 74
plot=diagnostics option, GLM procedure 52
potential conclusions about 71
as a typical plan component 19
power 64
POWER procedure 64
primary hypotheses 63
primary outcomes about 31
for TIAD study 79
principal investigator, responsibilities for 25
PRINT procedure 55
problematic distributions, exploring 42
problems, conceptualizing 28
programmer, responsibilities for 25
project director, responsibilities for 25
project initiation, as a component of analysis policies 7
project team communicating with 57
identifying 24
roles and responsibilities for 25
proposal analysis plan 3

Q

quantile-quantile (Q-Q) plot 48
quasi-experimental study 30
questions/hypotheses about 22, 23, 27, 29
analysis of 63, 83
as a typical plan component 18
conceptualizing the problem 28
for TIAD study 79
primary 63
research 29
secondary 69

SUMMARY procedure 36, 40

T
t-test 80
tabular presentations, developing 74
TABULATE procedure 40, 55, 74
target audience, identifying 25
templates 72, 81
TIAD study example 78
timeline
  about 72
  as a typical plan component 19
TWOSAMPLEMEANS statement 65

U
UNIVARIATE procedure 36, 42, 46
unplanned analysis 75

V
values, missing 62
variability, determining source of 49
variable distributions, for TIAD study 80
variables
  confounding 31, 67
descriptive statistics for major 40

W
weaknesses
  as a typical plan component 19
  of studies 72
WELCH option 51
Wilcoxon rank sum test 80

randomized controlled trial (RCT) 30
REG procedure 68
REPORT procedure 36, 40, 55
reporting guidelines, as a component of
  analysis policies 10
reports, turning plans into 74
research questions 29
revising the plan 73
rich text format (RTF) 36

sample size 64
secondary hypotheses 69
secondary outcomes
  about 31
    for TIAD study 79
sensitivity analysis 69
SHEET_INTERVAL option 74
skewness 48
SQL procedure 36
SQUARE option 43
statistical parameters 60
statistical policies, as a component of
  analysis policies 9
studies
  limitations of 72
  timeline for 72
types of 29
  weaknesses of 72
study team
  communicating with the 32
    for TIAD study 78
  reviewing data with 37
subgroup analyses 31, 69

About This Book

Purpose

*Data Analysis Plans: A Blueprint for Success Using SAS®* is a getting started guide to building an effective data analysis plan with a solid foundation for planning and managing your analytics projects. Data analysis plans are critical to the success of analytics projects. When implemented effectively, they facilitate team focus and communication through a mutual understanding of goals, data, and methods. In this way, workflow is improved and client expectations are well managed. This book provides step-by-step instructions to write, implement, and update your data analysis plan. It emphasizes the concept of an analysis plan as a working document that is updated throughout the life of a project, and may evolve into a sound and comprehensive draft for a final report or publication.

Is This Book for You?

This book was written primarily for applied statisticians and data analysts in the early stages of their career, when they are building foundational knowledge for performing data analyses in a consultative role. Analysts in environments such as university research support centers, corporate or government strategic planning and analysis departments, and private endeavors such as think tanks and management consulting firms will find this book helpful. However, this book also provides more experienced SAS users with the opportunity to adjust how they plan and conduct analytic projects. For them, the book can serve as a guide for establishing more fruitful and productive relationships with collaborators and clients. Subject areas include medical research, public health research, social studies, educational testing and evaluation, and environmental studies.
Prerequisites

While a basic knowledge of SAS is assumed, SAS code and program examples are given. It is assumed that the reader has some experience with the SAS data step and SAS procedure syntax.

Scope of This Book

The book explains how analysis planning is comparable to a home construction blueprint. Just as a blueprint adheres to local building codes, an analytics support group should have analysis policies to guide all of their analysis planning. The book shows you how to develop these policies. Guidance is then provided for identifying project team players and roles, and developing clear questions or hypotheses that directly address project goals and are suitable given the available data. The ensuing chapters cover the importance of describing the data sources and fields in the shared plan, and conducting and reporting data explorations to reveal and resolve potential problems such as outliers and unsuitable data value distributions. Next, the book demonstrates how to present the analyses that are appropriate for the agreed upon questions or hypotheses. The final chapters explain why and how to document potential conclusions and study weaknesses, and how to implement the plan with the goal of turning it into final report or publication.

The book is not a primer on inferential statistical methods required for most analyses. It is assumed that the statistician or analyst has sufficient training to choose statistical methods that are appropriate for their project, and understands the mathematical assumptions required of those methods.

About the Examples

Software Used to Develop the Book's Content

SAS BASE 9.4
Example Code and Data

You can access the example code and data for this book by linking to its author page at http://support.sas.com/publishing/authors. Select the name of the author. Then, look for the cover thumbnail of this book, and select Example Code and Data to display the SAS programs that are included in this book.

If you are unable to access the code through the Web site, send e-mail to saspress@sas.com.

Using SAS® University Edition? You can download SAS® University Edition for free, directly from SAS, then use the code and data sets provided with this book. Get started by visiting http://support.sas.com/publishing/import_ue.data.html.

Additional Resources

Although this book illustrates many analyses regularly performed in businesses across industries, questions specific to your aims and issues may arise. To fully support you, SAS Institute and SAS Press offer you the following help resources:

- For questions about topics covered in this book, contact the author through SAS Press:
  - Send questions by email to saspress@sas.com; include the book title in your correspondence.
  - Submit feedback on the author’s page at http://support.sas.com/author_feedback.

- For questions about topics in or beyond the scope of this book, post queries to the relevant SAS Support Communities at https://communities.sas.com/welcome.

- SAS Institute maintains a comprehensive website with up-to-date information. One page that is particularly useful to both the novice and the seasoned SAS user is its Knowledge Base. Search for relevant notes in the “Samples and SAS Notes” section of the Knowledge Base at http://support.sas.com/resources.

- Registered SAS users or their organizations can access SAS Customer Support at http://support.sas.com. Here you can pose specific questions to SAS Customer Support; under Support, click Submit a Problem. You will need to provide an email address to which replies can be sent, identify your organization, and provide a customer site number or license information. This information can be found in your SAS logs.
Keep in Touch

We look forward to hearing from you. We invite questions, comments, and concerns. If you want to contact us about a specific book, please include the book title in your correspondence.

To Contact the Author through SAS Press

   By e-mail: saspress@sas.com
   Via the Web: http://support.sas.com/author_feedback

SAS Books

For a complete list of books available through SAS, visit http://support.sas.com/bookstore.
   Phone: 1-800-727-3228
   Fax: 1-919-677-8166
   E-mail: sasbook@sas.com

SAS Book Report

Receive up-to-date information about all new SAS publications via e-mail by subscribing to the SAS Book Report monthly eNewsletter. Visit http://support.sas.com/sbr.

Publish with SAS

SAS is recruiting authors! Are you interested in writing a book? Visit http://support.sas.com/saspress for more information.
About the Authors

Kathleen Jablonski  Kathleen Jablonski, PhD, is an applied statistician in the Milken School of Public Health at George Washington University where she mentors students in research design and methods using SAS for data management and analysis. Her interests include study design, statistics, and epidemiology. She has served as principal investigator, as well as lead statistician, on several multicenter NIH funded studies and networks. She received a PhD in biological anthropology and a Master of Science in Applied Statistics.

Mark Guagliardo  Mark Guagliardo, PhD, is an analyst with the Veterans Health Administration, U.S. Department of Veterans Affairs and has used SAS for 35 years on a wide variety of data management and analysis tasks. He has been the principal investigator or co-investigator for dozens of federally funded grants and contracts, primarily in the area of health services research. His peer reviewed publications and presentations have been in the areas of health services, public health, pediatrics, geography, and anthropology. He is a certified GIS (geographic information systems) professional, and specializes in studies of access to health care services locations.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more:

http://support.sas.com/publishing/authors/jablonski.html

http://support.sas.com/publishing/authors/guagliardo.html
About the Authors
Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore for additional books and resources.