sas®
THE POWER TO KNOW®

SAS® Certification Prep Guide

Statistical Business Analysis Using SAS®9

Joni N. Shreve
Donna Dea Holland

The correct bibliographic citation for this manual is as follows: Shreve, Joni N. and Donna Dea Holland . 2018. *SAS® Certification Prep Guide: Statistical Business Analysis Using SAS®9*. Cary, NC: SAS Institute Inc.

SAS® Certification Prep Guide: Statistical Business Analysis Using SAS®9

Copyright © 2018, SAS Institute Inc., Cary, NC, USA

978-1-62960-381-0 (Hardcopy)
978-1-63526-352-7 (Web PDF)
978-1-63526-350-3 (epub)
978-1-63526-351-0 (mobi)

All Rights Reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

December 2018

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to **http://support.sas.com/thirdpartylicenses**.

Contents

About This Book

What Does This Book Cover?

The *SAS® Certification Prep Guide: Statistical Business Analysis Using SAS®9* is written for both new and experienced SAS programmers intending to take the SAS® Certified Statistical Business Analyst Using SAS®9: Regression and Modeling exam. This book covers the main topics tested on the exam which include analysis of variance, linear and logistic regression, preparing inputs for predictive models, and measuring model performance.

The authors assume the reader has some experience creating a SAS program consisting of a DATA step and PROCEDURE step, and running that program using any SAS platform. While knowledge of basic descriptive and inferential statistics is helpful, the authors provide several introductory chapters to lay the foundation for understanding the advanced statistical topics.

Requirements and Details

Exam Objectives

See the current exam objectives at https://www.sas.com/en_us/certification/credentials/advanced-analytics/statistical-business-analyst.html. Exam objectives are subject to change.

Take a Practice Exam

Practice exams are available for purchase through SAS and Pearson VUE. For more information about practice exams, see https://www.sas.com/en_us/certification/resources/sas-practice-exams.html.

Registering for the Exam

To register for the official SAS® Certified Statistical Business Analyst Using SAS®9: Regression and Modeling exam, see the SAS Global Certification website at www.sas.com/certify (https://www.sas.com/en_us/ certification.html).

Syntax Conventions

In this book, SAS syntax looks like this example:

```
DATA output-SAS-data-set
     (DROP=variables(s) | KEEP=variables(s));
   SET SAS-data-set <options>;
   BY variable(s)
RUN;
```

Here are the conventions used in the example:

- DATA, DROP=, KEEP=, SET, BY, and RUN are in uppercase bold because they must be spelled as shown.
- *output-SAS-data-set*, *variable(s)*, *SAS-data-set*, and *options* are in italics because each represents a value that you supply.
- <*options*> is enclosed in angle brackets because it is optional syntax.
- DROP= and KEEP= are separated by a vertical bar (|) to indicate that they are mutually exclusive.

The example syntax shown in this book includes only what you need to know in order to prepare for the certification exam. For complete syntax, see the appropriate SAS reference guide.

What Should You Know about the Examples?

This book includes tutorials for you to follow to gain hands-on experience with SAS.

Software Used to Develop the Book's Content

To complete examples in this book, you must have access to Base SAS, SAS Enterprise Guide, or SAS Studio.

Example Code and Data

You can access all example code and data sets for this book by linking to the author pages at https://support.sas.com/shreve or https://support.sas.com/dholland. There you will also find directions on how to save the data sets to your computer to ensure that the example code runs successfully. The author pages also include appendices which contain detailed descriptions of the two main data sets used throughout this book: (1) the Diabetic Care Management Case, and (2) the Ames Housing Case.

You can also refer to the section "Getting Started with SAS" in Chapter 1, "Statistics and Making Sense of Our World," for a general description of the two main data sets, a list of all data sets by chapter, and a sample program which illustrates how to access the data within the SAS environment.

SAS University Edition

This book is compatible with SAS University Edition. In order to download SAS University Edition, go to https://www.sas.com/en_us/software/university-edition.html.

Where Are the Exercise Solutions?

Exercise solutions and Appendices referenced in the book are available on the author pages at https://support.sas.com/shreve or https://support.sas.com/dholland.

We Want to Hear from You

Do you have questions about a SAS Press book that you are reading? Contact us at saspress@sas.com.

SAS Press books are written *by* SAS Users *for* SAS Users. Please visit sas.com/books to sign up to request information on how to become a SAS Press author.

We welcome your participation in the development of new books and your feedback on SAS Press books that you are using. Please visit sas.com/books to sign up to review a book

Learn about new books and exclusive discounts. Sign up for our new books mailing list today at https://support.sas.com/en/books/subscribe-books.html.

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:
https://support.sas.com/shreve
https://support.sas.com/dholland

Chapter 1: Statistics and Making Sense of Our World

Introduction

The goal of this book is to prepare future analysts for the SAS statistical business analysis certification exam.[1] Therefore, the book aims to validate a strong working knowledge of complex statistical analyses, including analysis of variance, linear and logistic regression, and measuring model performance. This chapter covers the basic and fundamental information needed to understand the foundations of those more advanced analyses. We begin by explaining what statistics is and providing definitions of terms needed to get started.

The chapter continues with a birds-eye view of the data analytics process including defining the purpose, data preparation, the analysis, conclusions and interpretation. Special consideration is given to the data preparation phase--with such topics as sampling, missing data, data exploration, and outlier detection--in an attempt to stress its importance in the validity of statistical conclusions. Where necessary we refer you to additional sources for further readings.

This chapter includes a road map detailing the scope of the statistical analyses covered in this book and how the specific analyses relate to the purpose. Finally, the chapter closes with a description of the data sets to be used throughout the book and provides you the first opportunity to access the data using sample SAS code before proceeding to subsequent chapters.

In this chapter you will learn about:

- statistics' two branches, descriptive statistics and inferential statistics, data mining, and predictive analytics
- variable types and how SAS distinguishes between numeric and character data types
- the data analytics process, including defining the purpose, data preparation, analysis, conclusions and interpretation
- exploratory analysis versus confirmatory analysis
- sampling and how it relates to bias
- selection bias, nonresponse bias, measurement error, confounding variables
- the importance of data cleaning
- the role of data cleaning to identify data inconsistencies, to account for missing data, and to create new variables, dummy codes, and variable transformations
- terms such as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR), and conditions for imputation
- data exploration for uncovering interesting patterns, detecting outliers, and variable reduction
- the roles of variables as either response or predictors
- the analytics road map used for determining the specific statistical modeling approach based upon the business question, the variable types, and the variable roles

- the statistical models to be tested on the certification exam, including two-sample t-tests, analysis of variance (ANOVA), linear regression analysis, and logistic regression analysis
- the use of the training data set and the validation data set to assess model performance
- both the Diabetic Care Management Case and the Ames Housing Case to be used throughout the book, their contents, and the sample SAS code used the read the data and produce an output of contents.

What Is Statistics?

We see and rely on statistics every day. Statistics can help us understand many aspects of our lives, including the price of homes, automobiles, health and life insurance, interest rates, political perceptions, to name a few. Statistics are used across many fields of study in academia, marketing, healthcare, treatment regimes, politics, housing, government, private businesses, national security, sports, law enforcement, and NGOs. The extensive reliance on statistics is growing. Statistics drive decisions to solve social problems, guide and build businesses, and develop communities. With the wealth of information available today, business persons need to know how to use statistics efficiently and effectively for better decision making. So, what is statistics?

Statistics is a science that relies on particular mathematical formulas and software to derive meaningful patterns and extrapolate actionable information from data sets. Statistics involves the use of plots, graphs, tables, and statistical tests to validate hypotheses, but it is more than just these. Statistics is a unique way to use data to make improvements and efficiencies in virtually any business or organization that collects quality data about their customers, services, costs, and practices.

The Two Branches of Statistics

Before defining the two branches of statistics, it is important to distinguish between a population and a sample. The **population** is the universe of all observations for which conclusions are to be made and can consist of people or objects. For example, a population can be made up of customers, patients, products, crimes, or bank transactions. In reality, it is very rare and sometimes impossible to collect data from the entire population. Therefore, it is more practical to take a **sample**--that is, a subset of the population.

There are two branches of statistics, namely descriptive statistics and inferential statistics. **Descriptive statistics** includes the collection, cleaning, and summarization of the data set of interest for the purposes of describing various features of that data. The features can be in the form of numeric summaries such as means, ranges, or proportions, or visual summaries such as histograms, pie charts, or bar graphs. These summaries and many more depend upon the types of variables collected and will be covered in Chapter 2, "Summarizing Your Data with Descriptive Statistics" and Chapter 3, "Data Visualization."

Inferential statistics includes the methods where sample data is used to make predictions or inferences about the characteristics of the population of interest. In particular, a summary measure calculated for the sample, referred to as a **statistic**, is used to estimate a population **parameter**, the unknown characteristic of the population. Inferential methods depend upon both the types of variables and the specific questions to be answered and will be introduced in Chapter 4, "The Normal Distribution and Introduction to Inferential Statistics" and covered in detail in Chapter 5, "Analysis of Categorical Variables" through Chapter 7, "Analysis of Variance."

Another goal of this book is to extend the methods learned in inferential statistics to those methods referred to as predictive modeling. **Predictive modeling**, sometimes referred to as **predictive analytics**, is the use of data, statistical algorithms and machine learning techniques to predict, or identify, the likelihood of a future outcome, based upon historical data. In short, predictive modeling extends conclusions about what has happened to predictions about what will happen in the future. The methods used for predictive modeling will be covered in Chapter 8, "Preparing the Input Variables for Prediction" through Chapter 11, "Measure of Model Performance" and provide a majority of the content for successfully completing the certification exam.

Finally, all content in this book falls under the larger topic, referred to as **data mining**, which is the process of finding anomalies, patterns, and correlations within large data sets to predict outcomes (SAS Institute).

Variable Types and SAS Data Types

All structured data sets are composed of rows and columns, where the rows represent the observations to be studied and the columns represent the variables related to the question or questions of interest. As stated earlier, in order to conduct either descriptive or inferential statistics, it is imperative that the analyst first define the **variable types**. Here we will also distinguish variable types from **data types**.

Variable Types

There are two types of variables, qualitative and quantitative (Anderson, et al., 2014; Fernandez, 2010). A **qualitative variable** is a variable with outcomes that represent a group or a category to which the observation is associated, and is sometimes referred to as a **categorical variable**. A **quantitative variable** is a variable with outcomes that represent a measurable quantity and are numeric in nature. Quantitative variables can be further distinguished as either discrete or continuous. A **discrete variable** is a numeric variable that results from counting; discrete variables can be infinite and do not necessarily have to be whole numbers. A **continuous variable** is a numeric variable that can theoretically take on infinitely many values within an interval and is, therefore, uncountable.

Let's consider an excerpt from a data set collected on patients related to the study of diabetes, as shown in Table 1.1 Data for the Study of Diabetes. It is evident that the variable, GENDER, is categorical, having values M and F, corresponding to males and females, respectively; major adverse event (AE1) is categorical as well. Notice that these variables are made up of textual data.

Table 1.1 Data for the Study of Diabetes

			Controlled	Hemoglobin		Syst	Diast			
Patient_ID	**Gender**	**Age**	**_Diabetic**	**_A1c**	**BMI**	**_BP**	**_BP**	**Cholesterol**	**NAES**	**AE1**
85348444	F	73	1	4.24	23.12	94.0	69.0	99.57	0	
507587021	F	82	0	11.49	24.82	101.2	75.0	211.66	4	Itching
561197284	F	76	1	0.16	28.70	69.0	45.0	252.33	0	
618214598	M	69	1	0.02	27.95	105.0	89.0	201.21	1	Nausea
1009556938	M	82	0	7.35	29.28	87.0	63.0	275.56	3	Nausea

The clinical data including A1c (Hemoglobin_A1c), BMI, systolic blood pressure (SYST_BP), diastolic blood pressure (DIAST_BP), and cholesterol has quantitative, continuous values. The variable AGE, as measured in years, is a continuous quantitative variable because it measures fraction of a year; although when asked our age, we all report it to the nearest whole number. The number of adverse events (NAES) is quantitative discrete because the values are the result of counting. Note that PATIENT ID is recorded as a number, but really acts as a unique identifier and serves no real analytical purpose.

Finally, it should be noted that a patient's diabetes is controlled if his or her A1c value is less than 7. Otherwise, it is not controlled. For example, patient 1 has an A1c value of 4.24 which is less than 7, indicating that patient 1's diabetes is controlled (CONTROLLED_DIABETIC=1); whereas patient 2 has an A1c value of 11.49 which is greater than or equal to 7, indicating that patient 2's diabetes is not controlled (CONTROLLED_DIABETIC=0). In short, CONTROLLED_DIABETIC is a categorical variable represented by a numeric value.

SAS Data Types

When you are using SAS software, data is distinguished by its data type, either numeric or character. A variable is numeric if its values are recorded as numbers; these values can be positive, negative, whole, integer, rational, irrational, dates, or times. A character variable can contain letters, numbers, and special characters, such as #, %, ^, &, or *.

The three variable types previously discussed overlap with these two data types utilized by SAS. In particular, categorical variables may be character or numeric data types; however, discrete and continuous quantitative variables must be numeric data types. So consider the diabetes data in Table 1.1 Data for the Study of Diabetes. The variable, CONTROLLED_DIABETIC, is categorical with a numeric data type. Although not shown here, the three condition variables, HYPERTENSION, STROKE, and RENAL_DISEASE, also have numeric data type to represent categorical variables. GENDER is a categorical variable with character data type. All quantitative variables discussed previously have numeric data type. While PATIENT_ID is numeric, it makes no sense to perform arithmetic operations, so it is used solely for identifying unique patients, and could have been easily formatted as a character type.

The Data Analytics Process

The process of business analytics is composed of several stages: Defining the Purpose, Data Preparation, Analysis, Conclusions, and Interpretation.

Defining the Purpose

All statistical analyses have a purpose and, as stated previously, the statistical methods depend upon that purpose. Furthermore, the purpose of data analysis can be for either exploratory or confirmatory reasons. In exploratory data analysis, the purpose is strictly to summarize the characteristics of a particular scenario and relies on the use of descriptive statistics. In confirmatory data analysis, there is a specific question to be answered and relies on the use of inferential statistics. Table 1.2 Examples of Analyses by Purpose for Various Industries gives some examples of how statistical analyses are used to answer questions relative to both exploratory and confirmatory analyses in various industries.

Table 1.2 Examples of Analyses by Purpose for Various Industries

INDUSTRY	PURPOSE
Retail	Identify the advertising delivery method most effective in attracting customers
	Describe the best selling products and the customers buying those products
Healthcare	Identify the factors associated with extended length of stay for hospital encounters
	Predict healthcare outcomes based upon patient and system characteristics
Telecommunication	Identify customer characteristics and event triggers associated with customer churn
	Describe revenues collected for various products across various geographic areas
Banking	Identify transactions most likely to be fraudulent
	Predict those customers most likely to default on a personal loan
Education	Describe student enrollment for purposes of budgeting, accreditation, and resource planning
	Identify factors associated with student success
Government	Describe criminal activity in terms of nature, time, location for purposes of resource planning
	Predict tax revenue based upon the values of commercial and residential properties
Travel & Hospitality	Predict room occupancy based upon historical industry occupancy measures
	Describe customer needs and preferences by location and seasonality
Manufacturing	Predict demand for goods based upon price, advertising, merchandising, and seasonality
	Describe brand image and customer sentiment after product launch

Data Preparation

Once the purpose has been confirmed, the analyst must then obtain the data related to the question at hand. Many organizations have either a centralized data warehouse or data marts from which to access data, sometimes requiring the analyst to merge various databases to get the final data set of interest. For example, to study customer behavior, the analyst may need to merge one data set containing the customer's name, address, and other demographic information with a second data set containing purchase history, including products purchased, quantities, costs, and dates of purchases. In other cases, the analysts may have to collect the data themselves. In any event, care must be taken to ensure the quality and the validity of the data used. In order to do this, special consideration should be given to such things as sampling, cleaning the data, and a preliminary exploring of the data to check for outliers and interesting patterns.

Sampling

Sometimes it is either impractical or impossible to collect all data pertinent to the question of interest. That's where sampling comes into play! As soon as the analyst decides to take a sample, extreme care must be given to reduce any sources of bias. **Bias** occurs when the statistics obtained from the sample are not a 'good' representation of the population parameters. Obviously, bias exists when the sample is not representative of the target population, therefore, giving results that are not generalizable to the population. To ensure a representative sample, the analyst must employ some kind of probability sampling scheme. If a probability sample is not taken, the validity of the results should be questioned.

One such example of a probability sample is a **simple random sample** in which all observations in the population have an equal chance of being selected. The statistical methods used in this book assume that a simple random sample is selected. For a more thorough discussion of other probability sampling methods, we suggest reading *Survey Methodology, Second Edition* (Groves, R.M., et al., 2009).

There are other sources of bias and the analyst must pay close attention to the conditions under which the data is collected to reduce the effects on the validity of the results. One source of bias is **selection bias**. Selection bias occurs when subgroups within the population are underrepresented in the sample. For example, suppose the college administration is interested in studying students' opinions on its advising procedures and it uses an 'old' list of students from which to select a random sample. In this case, the sample would include those who have already graduated and not include those who are new to the college. In other words, the sample is not a good representation of the current student population.

Another type of bias is **nonresponse bias**. Nonresponse bias occurs when observations that have data values differ from those that do not have values. For example, suppose a telecommunications company which supplies internet service wants to study how those customers who call and complain differ from those customers who do not complain. If the analyst wants to study the reason for the complaint, there is information only for those who complain; obviously, no reason exists for those who do not call to complain. As a result, an analysis of the complaints cannot be inferred to the entire population of customers. See the section on data cleaning for more details on missing data.

Variable values can also be subjected to **measurement error**. This occurs when the variable collected does not adequately represent the true value of the variable under investigation. Suppose a national retailer provides an opportunity to earn a discount on the next purchase in return for completing an online survey. It could be that the customer is only interested in completing the survey in order to get the discount code and pays no attention to the specifics by answering yes to all of the questions. In this case, the actual responses are not a representation of the customer's true feelings. Therefore, the responses consist of measurement error.

Finally, the analyst should be aware of confounding. A **confounding variable** is a variable external to the analysis that can affect the relationship between the variables under investigation. Suppose a human production manager wants to investigate the effects of background music on employee performance as measured by number of units produced per hour but does not account for the time of day. It could be that the performance of employees is reduced when exposed to background music A as opposed to B. However, background music A is played at the end of the shift. In short, the performance is related to an extraneous variable, time of day, and time of day affects the relationship between performance and type of background music.

Cleaning the Data

Once the analyst has the appropriate data, the cleaning process begins. Data cleaning is one of the most important and often time-consuming aspects of data analysis. The information gleaned from data analysis is only as good as the data employed. Furthermore, it is estimated that data cleaning usually takes about 80% of a project's time and effort. So what is involved in the data cleaning process? Data cleaning involves various tasks, including checking for data errors and inconsistencies, handling missing data, creating new or transforming existing variables, looking for outliers, and reducing the number of potential predictors.

First, the analyst should check for data errors and inconsistencies. For example, certain variables should fall within certain ranges and follow specific business rules--the quantity sold and costs for a product should always be positive, the number of office visits to the doctor should not exceed 31 in a single month, and the delivery date should not fall before the date of purchase.

Then the question is what to do once you find these data inconsistencies. Of course, every effort should be made to find the sources of those errors and correct them; however, what should the analyst do if those errors cannot be fixed? Obviously, values that are in error should not be included in the analysis, so the analyst should replace those values with blanks. In this case, these variables are treated as having missing values. So what are missing values?

A missing value, sometimes referred to as **missing data**, occurs when an observation has no value for a variable. SAS includes for analysis only observations for which there is complete data. If an observation does not have complete data, SAS will eliminate that observation using either listwise or pairwise deletion. In **listwise deletion**, an observation is deleted from the analysis if it is missing data on any one variable used for that analysis. In **pairwise deletion**, all observations are used in analysis; however, only pairs of variables with missing values are removed from analyses. By default, most SAS procedures use listwise deletion, with the exception of the correlation procedure (PROC CORR) which uses pairwise deletion. It is important that the analyst know the sample size for analysis and the deletion method used at all times and to be aware of the effects of eliminating missing data.

So what should the analyst do when there is missing data? Schlomer, Bauman, and Card (2010) cite various suggestions on the percentage of missing observations where the analyst could proceed with little threat to bias; however, they further suggest, instead, looking at the 'pattern of missingness' and why data is missing so that imputation methods may be employed.

Some missing values occur because of a failure to respond or to provide data; others are due to data collection errors or mistakes, as mentioned previously. If the observations are **missing completely at random (MCAR)**, that is, if there are no systematic reasons related to the study for the missing values to exist, then the analysis can proceed using only the complete data without any real threats to bias (Little and Rubin, 2002). In short, it is believed that the observations with missing values make up a random sample themselves and, if deleted, the remaining observations with complete data are representative of the population.

While it is possible for data to be MCAR, that situation is very rare. It is more likely the case that data is **missing at random (MAR)**; MAR occurs if the reason for missing is not related to the outcome variable, but instead, related to another variable in the data set (Rubin, 1976). In either case, MCAR or MAR, there are imputation methods that use the known data to derive the parameter estimates of interest. When these methods are employed, all data will be retained for analyses. See Schlomer et al. (2010) for a description of non-stochastic and stochastic approaches to imputation.

If neither MCAR nor MAR exists, then the data is **not missing at random (NMAR)**. In this case, the reason that data is missing is precisely related to the variable under study. When data is NMAR, imputation methods are not valid. In fact, when observations are NMAR and missing data is omitted from analyses, results will be biased and should not be used for descriptive nor inferential purposes.

While there are various ways to handle missingness in data, we describe one method in particular. In Chapter 8, "Preparing the Input Variables for Prediction", we address this problem by introducing a **dummy variable**, or **missing value indicator**, for each predictor where missing data is of concern. The missing value indicator is coded as '1' for an observation if the variable under investigation is missing for that observation, or '0' otherwise. You are directed to Schwartz and Zeig-Owens (2012) for further discussion, a list of questions to facilitate the understanding of missing data, and the Missing Data SAS Macro as an aid in assessing the patterns of missingness.

In any event, when analyses involving missing data, it is critical to report both (1) the extent and nature of missing data and (2) the procedures used to manage the missing data, including the rationale for using the method selected (Schlomer, Bauman, and Card, 2010).

Another aspect of data cleaning involves creating new variables that are not captured naturally for the proposed analysis purpose. For example, suppose an analyst is investigating those factors associated with hospital encounters lasting more than the standard length of time. One such factor could be whether or not the encounter is considered a readmission. The patient data may not have information specifically indicating if the encounter under investigation is a readmission; however, the hospital admission data could be used to determine that. In other words, the analyst could create a new variable, called READMIT, which has a value of YES if the current encounter has occurred within 30 days of the discharge date of the previous hospital encounter, or NO otherwise.

In another example, suppose a retailer wants to know how many times a customer has made a purchase in the last quarter. Retailers probably don't collect that data at the time of each purchase--in fact, if surveyed, the customer may not correctly recall that number anyway. However, counting algorithms can be applied to transactional data to count the number of purchases for a specific customer ID within a defined period of time.

Many times, the analyst will create **'dummy' variables**, which are coded as '1' if an attribute about the observation exists or '0' if that attribute does not exist. For example, a churn variable could be coded as '1' if the customer has churned or '0' if that customer has been retained.

Next, the analyst may need to transform data. As you will see later in this book, some statistical analyses require that certain assumptions about the data are met. When those assumptions are violated, it may require transforming variables to ensure the validity of results. For example, the analyst may create a new variable representing the natural log of a person's salary as opposed to the salary value itself. Data transformations will be covered in Chapters 8 and 9. In Chapter 8, "Preparing the Input Variables for Prediction", methods to detect non-linearities are discussed in the context of logistic regression. In Chapter 9, "Linear Regression Analysis," we illustrate how to transform predictors for purposes of improving measures of fit in the context of linear regression analysis.

Finally, the analyst should check for **outliers**, that is, observations that are relatively 'far' in distance from the majority of observations; outliers are observations that deviate from what is considered normal. Sometimes outliers are referred to as **influential observations**, because they have undue influence on descriptive or inferential conclusions. Like missing values, the analyst must investigate the source of the outlier. Is it the result of data errors and how can it be fixed? If the observation is a legitimate value, is it influential and how should it be handled? Is there any justification for omitting the outlier or should it be retained? Sometimes outliers are detected during the data cleaning process, but ordinarily outliers are detected when specifically exploring the data, as discussed in the next section.

The data analyst must understand that data cleaning is an iterative process and must be handled with extreme care. For more in-depth information on data cleaning see Cody's *Data Cleaning Techniques Using SAS, Third Edition*.

Exploring the Data

Once the data is cleaned, the analyst should explore the data to become familiar with some basic data attributes--in general, what is the sample size, what products are included in data and which products account for a majority of the purchases, what types of drugs are administered based upon disease type, what geographic areas are represented by your customers, what books are purchased across various age groups.

The analyst should slice the data across groups and provide summary statistics on the variable of interest (such as the mean, median, range, minimum, and maximum or frequencies) or data visualizations (such as the histogram or bar chart) for comparative purposes, to look for various patterns, and to generate ideas for further investigation as it relates to the ultimate purpose. Many of these descriptive tools will be discussed in Chapter 2, “Summarizing Your Data with Descriptive Statistics” and Chapter 3, “Data Visualization.” Inferential analyses for confirming relationships between two variables will be discussed in Chapter 5, “Analysis of Categorical Variables,” ~~and~~ Chapter 6, “Two-Sample T-Test,” and Chapter 7, “Analysis of Variance (ANOVA).”

The analyst can provide scatter diagrams for pairs of variables to establish whether or not linear relationships exist. In situations where there are hundreds of predictors and inevitably correlations among those predictors exist, data reduction methods can be employed so that a few subsets of predictors can be omitted without sacrificing predictive accuracy. In Chapter 8, methods for detecting redundancy will be discussed for purposes of data, or dimension, reduction.

Finally, the analyst should explore the data specifically for detecting outliers. An observation can be an outlier with respect to one variable; methods of detecting these univariate outliers will be covered in both Chapters 2 and 3. Or an observation can be an outlier in a multivariate sense with respect to two or more variables. Specifically, a scatter diagram is a first step in detecting an outlier on a bivariate axis. Methods of detecting multivariate outliers will be covered in Chapter 9, “Linear Regression Analysis.”

Analyzing the Data and Roadmap to the Book

Once the data have been prepared, the goal of the analyst is to make sense of the data. The first step is to review the purpose and match that purpose to the analysis approach. If the purpose is explanatory, then the analyst will employ descriptive statistics for purposes of reporting, or describing, a particular scenario.

For example, in Chapter 2, “Summarizing Your Data with Descriptive Statistics,” you will learn about ways to describe your numeric data with measures of center (mean, median, mode), variation (range, variance, and standard deviation), and shape (skewness and kurtosis). In Chapter 3,” Data Visualization,” you will learn how to describe your categorical data using frequencies and proportions. Chapter 3 will illustrate how to employ data visualization techniques to get pie charts and bar graphs for categorical data and histograms, Q-Q plots, and box plots for numeric data. These data visualizations and numeric summaries, when used together, provide a powerful tool for understanding your data and describing what is happening now.

If the purpose of the analysis is confirmatory, then you as analyst will employ inferential statistics for the purposes of using sample data to make conclusions about proposed models in the population. It is when hypotheses about organizational operations--whatever those may be--are confirmed that decision makers are able to predict future outcomes or effect some change for increased operational performance. This book emphasizes the specific statistical models needed to pass the certification exam, as listed in Table 1.3 Summary of Statistical Models for Business Analysis Certification by Variable Role.

Table 1.3 Summary of Statistical Models for Business Analysis Certification by Variable Role

	TYPE of Predictor Variables	
TYPE of Response Variable	**CATEGORICAL**	**CONTINUOUS**
CONTINUOUS	t-Tests (Chapter 6) or Analysis of Variance (Chapter 7)	Linear Regression (Chapter 9)
CATEGORICAL	Logistic Regression (Chapter 10)	Logistic Regression (Chapter 10)

As we discuss each model throughout Chapters 5 through 7, 9 and 10, you will begin to associate a specific type of question with a specific type of statistical model; and with each type of model, the variables take on specific roles--either as response or predictor variables. A **response variable** is the variable under investigation and is sometimes referred to as the dependent variable, the outcome variable, or the target variable. A **predictor variable** is a variable that is thought to be related to the response variable and can be used to predict the value of the response variable. A predictor variable is sometimes referred to as the independent variable or the input variable.

So, for example, when the analyst is interested in determining if the categorical response variable--whether or not a customer will churn--is related to the categorical predictor variable—rent or own, the appropriate type of analysis is logistic regression. If the analyst wants to further research churn and includes continuous predictors such as monthly credit card average and mortgage amount, then the appropriate analysis is logistic regression as well. These statistical methods will be covered in Chapter 10, “Logistic Regression Analysis,” as illustrated in Table 1.3 Summary of Statistical Models for Business Analysis Certification by Variable Role.

If the analyst is interested in studying how crime rate is related to both poverty rate and median income (where the response variable, crime rate, is continuous and the predictors, poverty rate and median income, are both continuous), then the appropriate analysis in linear regression analysis. This statistical method will be covered in Chapter 9, “Linear Regression Analysis.”

Finally, suppose a retailer was interested in testing a promotion type (20% off of any purchase, buy-one-get-one-half-off, or 30% off for one-day-only) and the promotion site (online only purchase or in-store only purchase). If the analyst is interested in studying how sales are related to the promotion type and/or promotion site, then the appropriate method is analysis of variance (ANOVA) where the response variable is continuous and the predictors are categorical. This type of analysis will be covered in Chapter 7, “Analysis of Variance (ANOVA).” Note that when the question about a continuous response variable is restricted to the investigation of one predictor composed of only two groups, then the analyst would use the t-test, as described in Chapter 6, “Two-Sample T-Test.”

It is critical to note that if the purpose of data analysis is confirmatory, the analyst must also employ descriptive statistics for exploring the data as a way of becoming familiar with its features. Conducting confirmatory analyses without exploring the data is like driving to your destination without a map.

Finally, when the purpose of the analysis is classification, or predicting a binary categorical outcome using logistic regression analysis, the analyst must incorporate an assessment component to the modeling. In particular, the data is partitioned into two parts, the training data set and the validation data set. The best predictive models are developed and selected using the **training data set**. The performance of those models is tested by applying those methods to the **validation data set**. That model which performs or predicts best when applied to the validation data is the model selected for answering the proposed business question. This and other topics related to measures of model performance will be covered in Chapter 11, “Measure of Model Performance.”

Conclusions and Interpretation

As with the other parts of the research process, the conclusion and interpretation are essential. You may have heard that “the numbers speak for themselves.” No, they don’t! All statistical numbers must be interpreted. Your interpretation should always relate the analytic results back to the research question. If the purpose of the analysis is descriptive, report the findings and use those findings to describe the current state of affairs.

If the purpose of the analysis is confirmatory, or inferential, in nature, state the analytical conclusions and provide interpretations in terms of how an organization can be proactive to effect some improvement in operations. Always consider whether there is an alternative way to interpret the results. When two or more possible interpretations of the results exist, it is the analyst's job to follow each possible explanation and provide detailed reasons for interpreting one outcome in ~~a~~ one particular way or another way. Reliance on the subject matter expert is imperative to ensure proper interpretation.

Getting Started with SAS

Throughout the book, we introduce various business questions to illustrate which statistical analyses are used to generate the corresponding answers. Specifically, we define the problem relative to the chapter content, construct-the necessary SAS code-for generating output, and provide an interpretation of the results for purposes of answering the question.

In order to provide a context for questions, we use various data sets that accompany the book. The two main data sets, and variants of those data sets, are (1) the Diabetic Care Management Case, and (2) the Ames Housing Case. Those two data sets are described in this section.

Diabetic Care Management Case

The data file provided with this book, DIABETICS, contains demographic, clinical, and geo-location data for patients who have been diagnosed with diabetes. The observation under investigation is the patient, each having variables that fall into the following categories:

1. Demographic information, such as patient ID, gender, age, and age range.
2. Date of the last doctor's visit and the general state of the patient, including height, weight, BMI, systolic and diastolic blood pressure, type of diabetes, if the diabetes is controlled, medical risk, if the patient has hypertension, hyperlipidemia, peripheral vascular disease (PVD), renal disease, and if the patient has suffered a stroke.
3. The results of 57 laboratory tests, including those tests from the comprehensive metabolic panel (CMP) which are used to evaluate the how the organs function and to detect various chronic diseases.
4. Information related to prescription medicine, including type of medication, dosage form, and the number and nature of adverse events with duration dates.
5. Geo-location data including the City and State where the patient resides, along with longitude and latitude.

In some cases, a random sample of 200 patients, in a file called DIAB200, is used for analysis. A complete data dictionary of the full data set with detailed descriptions is found in the Appendix B.

Ames Housing Case

The second major data set used for this book is the Ames Housing data, created by Dean deCock as an alternative to the Boston housing data (deCock, 2011). The original data was collected from the Ames Assessor's Office and contains 2,920 properties sold in Ames, IA, from 2006 through 2010. The data includes 82 variables on each of the houses.

The observation under investigation is the house, each having data on the following types of variables:

1. Quantitative measures of area for various parts of the house (above ground living area, basement, lot area, garage, deck, porch, pool, etc.).
2. Counts of various amenities (number of bedrooms, kitchens, full baths above ground and in basement, half baths above ground and in basement, fireplaces, number of cars the garage will hold).
3. Ratings--from excellent to very poor--for various house characteristics (overall quality, overall condition, along with the quality and condition of the exterior, basement, kitchen, heating, fireplace, garage, fence, etc.).
4. Descriptive characteristics, including year built, type of road access to property, lot shape and contour, lot configuration, land slope, neighborhood, roof style, roof material, type of exterior, type of foundation, basement exposure, type of heating and air, type of electrical system, garage type, whether or not driveway is paved, etc. Go to http://ww2.amstat.org/publications/jse/v19n3/Decock/DataDocumentation.txt to see the original documentation.

For this book, we consider a specific group of properties; in particular, the population of interest is defined as all single-family detached, residential-only houses, with sale conditions equal to 'family' or 'normal.' The sale condition allows for excluding houses that were sold as a result of a foreclosure, short sale, or other conditions that may bias the sale price.

As a result, the data set used in this book, called AMESHOUSING, contains 1,984 houses. After extensive exploration and purposes related to topics in this book, we created additional variables, resulting in a total of 103 variables, as defined in Appendix A. For the chapters covering topics related to predictive modeling, twenty-nine (29) total numeric and binary input variables are considered in the modeling process. The book does reference variations of the Ames housing data, along with other data sets, as listed in Table 1.4 List of Data Sets Used in the Book by Chapter.

Table 1.4 List of Data Sets Used in the Book by Chapter

Chapter	Data Set Name	Chapter	Data Set Name
1	ameshousing, diabetics	7	cas
2	all, diab200	8	ames300miss, ames70
3	diabetics, diab200, sunglasses	9	amesreg300, revenue
4	diabetics, diab25f	10	ames300, ames70, amesnew
5	ames300	11	ameshousing, ames70, ames30
6	ames300, alt40		

Accessing the Data in the SAS Environment

As stated earlier, we are assuming that you have a basic understanding of the SAS environment and the components of the SAS program, namely the DATA step and the procedure or PROC step. Recall that in order to access a SAS data set using the DATA step, the analyst must first use a LIBNAME statement pointing to where the data set is located. In this book, all SAS code references data sets located in the SASBA folder on the C drive.

Each data set is saved in its own subfolder within the SASBA parent folder. So, for example, the Ames housing data set is saved in the AMES subfolder, and the LIBNAME statement used to point to the data location has the form:

```
libname SASBA ‘c:\sasba\ames’;
```

The diabetes data used in the Diabetic Care Management Case is saved in the HC subfolder and is accessed using the following LIBNAME statement:

```
libname SASBA ‘c:\sasba\hc’;
```

In order to ensure that all readers are able to run the code found in subsequent chapters, we start with a very simple SAS program so that you can both access the data for the Diabetes Care Management Case and run a basic CONTENTS procedure for purposes of reviewing the specific details of the data set. Consider the Program 1.1 PROC CONTENTS of the Diabetes Care Management Case Data Set.

Program 1.1 PROC CONTENTS of the Diabetes Care Management Case Data Set

```
libname SASBA ‘c:\sasba\hc’;
data patient;
   set sasba.diabetics;
run;

proc contents data=patient;
run;
```

First, you can see from Program 1.1 that the LIBNAME statement defines a library called SASBA which points to the C:\SASBA\HC directory for accessing data. The permanent data set, DIABETICS located in the SASBA library, is placed into the temporary data set, PATIENT, and PROC CONTENTS is then applied to the data set, PATIENT. When the SAS code is run, the analyst should get the SAS LOG 1.1 PROC CONTENTS of the Diabetes Care Management Case Data Set.

SAS Log 1.1 PROC CONTENTS of the Diabetes Care Management Case Data Set

```
1   libname SASBA ‘c:\sasba\hc’;
NOTE: Libref SASBA was successfully assigned as follows:
      Engine:        V9
      Physical Name: c:\sasba\hc
2  data patient;
3     set sasba.diabetics;
```

```
NOTE: There were 200 observations read from the data set SASBA.DIABETICS.
NOTE: The data set WORK.PATIENT has 200 observations and 125 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds


4   proc contents data=patient;
5   run;

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time           0.07 seconds
      cpu time            0.06 seconds
```

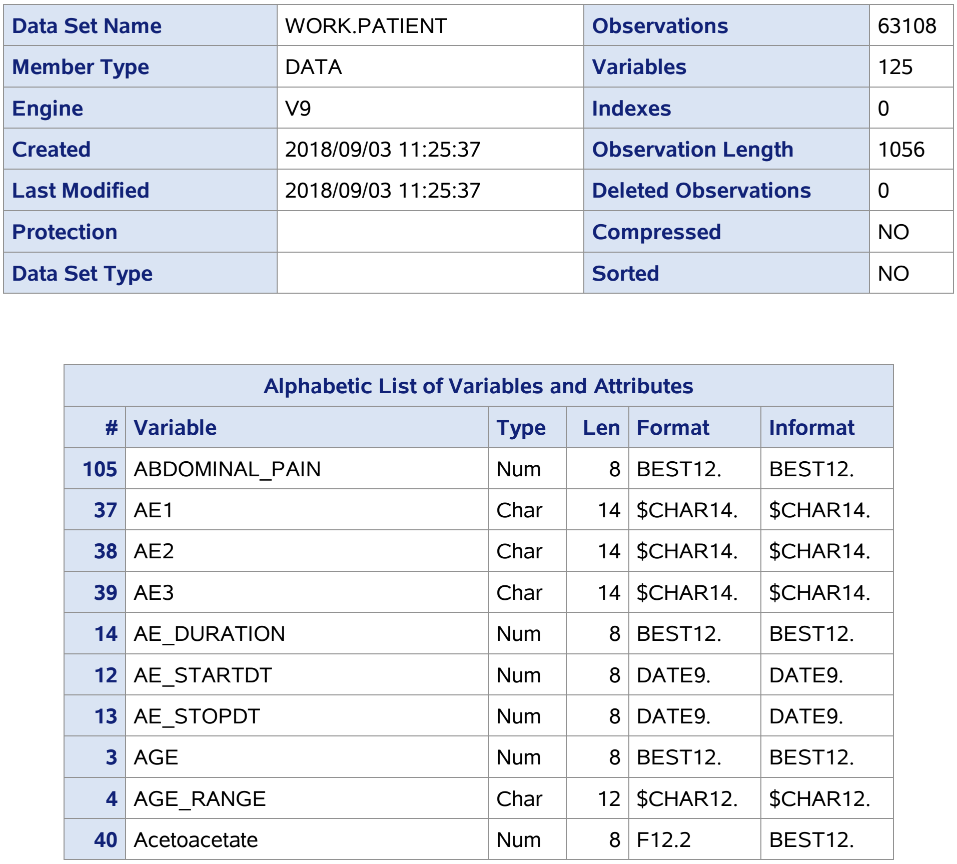
Remember that the LOG file documents everything you do when running a SAS session. The lines in the LOG beginning with numbers are the original SAS statements in your program. The remaining lines begin with a SAS message--either NOTE, INFO, WARNING, ERROR, or an error number--and provide the analyst with valuable information as to the accuracy of the output.

From the LOG file, you can see that the library reference was successfully assigned. You can then see that 63,108 observations were read from the permanent SAS data set, DIABETICS, and then read into the temporary data set, PATIENT, having 125 variables, followed by the CONTENTS procedure. Included in the LOG is total process time as well.

It should be noted that it is very important to review the LOG file after every program execution for errors and warnings. Keep in mind that executing a SAS program and getting output does not necessarily mean that the results are correct. While there may be no run-time errors, there may be logical errors, many of which can be detected by checking the LOG file for what the analyst thinks is reasonable given the task at hand.

Once the analyst has checked the LOG file and has reasonable certainty that the program has run successfully, he or she can review the output as illustrated in Output 1.1 PROC CONTENTS of the Diabetes Care Management Case Data Set.

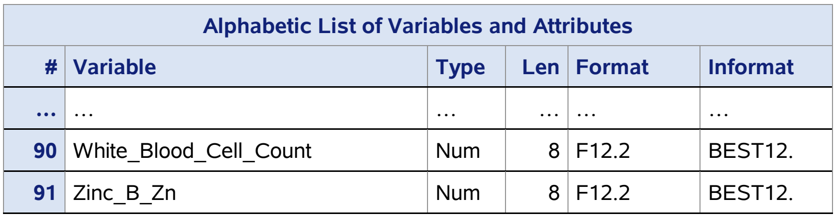
Output 1.1 PROC CONTENTS of the Diabetes Care Management Case Data Set



Data Set Name	WORK.PATIENT	Observations	63108
Member Type	DATA	Variables	125
Engine	V9	Indexes	0
Created	2018/09/03 11:25:37	Observation Length	1056
Last Modified	2018/09/03 11:25:37	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
105	ABDOMINAL_PAIN	Num	8	BEST12.	BEST12.
37	AE1	Char	14	$CHAR14.	$CHAR14.
38	AE2	Char	14	$CHAR14.	$CHAR14.
39	AE3	Char	14	$CHAR14.	$CHAR14.
14	AE_DURATION	Num	8	BEST12.	BEST12.
12	AE_STARTDT	Num	8	DATE9.	DATE9.
13	AE_STOPDT	Num	8	DATE9.	DATE9.
3	AGE	Num	8	BEST12.	BEST12.
4	AGE_RANGE	Char	12	$CHAR12.	$CHAR12.
40	Acetoacetate	Num	8	F12.2	BEST12.



Alphabetic List of Variables and Attributes					
#	**Variable**	**Type**	**Len**	**Format**	**Informat**
…	…	…	…	…	…
90	White_Blood_Cell_Count	Num	8	F12.2	BEST12.
91	Zinc_B_Zn	Num	8	F12.2	BEST12.

From the output, you can see that the first table summarizes information about the data set. Specifically, you can see that the temporary data set, PATIENT, has 63,108 observations with 125 variables, along with the creation data. The second table, representing an excerpt of the output, summarizes information about each individual variable; namely, the number (#) indicating the column location in the data set, the variable name, the variable type (numeric or character), the storage size in bytes (Len), the format for printing purposes, and the informat for input. If the variables had labels, those were included as well.

Key Terms

bias
categorical variable
character variable
confirmatory data analysis
confounding variable
continuous variable
data mining
data types
descriptive statistics
discrete variable
dummy variables
exploratory data analysis
inferential statistics
influential observations
listwise deletion
measurement error
missing at random (MAR)
missing completely at random (MCAR)
missing data
missing value indicator
nonresponse bias
not missing at random (NMAR)
numeric variable
outliers
pairwise deletion
parameter
population
predictor variable
predictive modeling
qualitative variable
quantitative variable
response variable
sample
selection bias
simple random sample
statistic
statistics
training data set
validation data set
variable types

[1] Officially, the name is the SAS Statistical Business Analysis Using SAS®9 Regression and Modeling exam.

Ready to take your SAS® and JMP® skills up a notch?

Be among the first to know about new books,
special events, and exclusive discounts.
support.sas.com/newbooks

Share your expertise. Write a book with SAS.
support.sas.com/publish

sas.com/books
for additional books and resources.

sas®

THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. © 2017 SAS Institute Inc. All rights reserved. M1588358 US.0217