

# Table of Contents

---

List of Programs	ix
Preface	xv
Acknowledgments	xvii

## 1

### Checking Values of Character Variables

---

Introduction	1
Using PROC FREQ to List Values	1
Description of the Raw Data File PATIENTS.TXT	2
Using a DATA Step to Check for Invalid Values	7
Describing the VERIFY, TRIM, MISSING, and NOTDIGIT Functions	9
Using PROC PRINT with a WHERE Statement to List Invalid Values	13
Using Formats to Check for Invalid Values	15
Using Informats to Remove Invalid Values	18

## 2

### Checking Values of Numeric Variables

---

Introduction	23
Using PROC MEANS, PROC TABULATE, and PROC UNIVARIATE to Look for Outliers	24
Using an ODS SELECT Statement to List Extreme Values	34
Using PROC UNIVARIATE Options to List More Extreme Observations	35
Using PROC UNIVARIATE to Look for Highest and Lowest Values by Percentage	37
Using PROC RANK to Look for Highest and Lowest Values by Percentage	43
Presenting a Program to List the Highest and Lowest Ten Values	47
Presenting a Macro to List the Highest and Lowest "n" Values	50
Using PROC PRINT with a WHERE Statement to List Invalid Data Values	52
Using a DATA Step to Check for Out-of-Range Values	54
Identifying Invalid Values versus Missing Values	55

Listing Invalid (Character) Values in the Error Report	57
Creating a Macro for Range Checking	60
Checking Ranges for Several Variables	62
Using Formats to Check for Invalid Values	66
Using Informats to Filter Invalid Values	68
Checking a Range Using an Algorithm Based on Standard Deviation	71
Detecting Outliers Based on a Trimmed Mean and Standard Deviation	73
Presenting a Macro Based on Trimmed Statistics	76
Using the TRIM Option of PROC UNIVARIATE and ODS to Compute Trimmed Statistics	80
Checking a Range Based on the Interquartile Range	86

### 3

#### **Checking for Missing Values**

---

Introduction	91
Inspecting the SAS Log	91
Using PROC MEANS and PROC FREQ to Count Missing Values	93
Using DATA Step Approaches to Identify and Count Missing Values	96
Searching for a Specific Numeric Value	100
Creating a Macro to Search for Specific Numeric Values	102

### 4

#### **Working with Dates**

---

Introduction	105
Checking Ranges for Dates (Using a DATA Step)	106
Checking Ranges for Dates (Using PROC PRINT)	107
Checking for Invalid Dates	108
Working with Dates in Nonstandard Form	111
Creating a SAS Date When the Day of the Month Is Missing	113
Suspending Error Checking for Known Invalid Dates	114

**5****Looking for Duplicates and "n" Observations per Subject**

---

Introduction	117
Eliminating Duplicates by Using PROC SORT	117
Detecting Duplicates by Using DATA Step Approaches	123
Using PROC FREQ to Detect Duplicate ID's	126
Selecting Patients with Duplicate Observations by Using a Macro List and SQL	129
Identifying Subjects with "n" Observations Each (DATA Step Approach)	130
Identifying Subjects with "n" Observations Each (Using PROC FREQ)	132

**6****Working with Multiple Files**

---

Introduction	135
Checking for an ID in Each of Two Files	135
Checking for an ID in Each of "n" Files	138
A Macro for ID Checking	140
More Complicated Multi-File Rules	143
Checking That the Dates Are in the Proper Order	147

**7****Double Entry and Verification (PROC COMPARE)**

---

Introduction	149
Conducting a Simple Comparison of Two Data Sets	150
Using PROC COMPARE with Two Data Sets That Have an Unequal Number of Observations	159
Comparing Two Data Sets When Some Variables Are Not in Both Data Sets	161

**8****Some PROC SQL Solutions to Data Cleaning**

---

Introduction	165
A Quick Review of PROC SQL	166
Checking for Invalid Character Values	166
Checking for Outliers	168

Checking a Range Using an Algorithm Based on the Standard Deviation	169
Checking for Missing Values	170
Range Checking for Dates	172
Checking for Duplicates	173
Identifying Subjects with "n" Observations Each	174
Checking for an ID in Each of Two Files	174
More Complicated Multi-File Rules	176

## 9

### Correcting Errors

---

Introduction	181
Hardcoding Corrections	181
Describing Named Input	182
Reviewing the UPDATE Statement	184

## 10

### Creating Integrity Constraints and Audit Trails

---

Introducing SAS Integrity Constraints	187
Demonstrating General Integrity Constraints	188
Deleting an Integrity Constraint Using PROC DATASETS	193
Creating an Audit Trail Data Set	193
Demonstrating an Integrity Constraint Involving More than One Variable	200
Demonstrating a Referential Constraint	202
Attempting to Delete a Primary Key When a Foreign Key Still Exists	205
Attempting to Add a Name to the Child Data Set	207
Demonstrating the Cascade Feature of a Referential Constraint	208
Demonstrating the SET NULL Feature of a Referential Constraint	210
Demonstrating How to Delete a Referential Constraint	211

**11****DataFlux and dfPower Studio**

---

Introduction	213
Examples	215

**Appendix****Listing of Raw Data Files and SAS Programs**

---

Programs and Raw Data Files Used in This Book	217
Description of the Raw Data File PATIENTS.TXT	217
Layout for the Data File PATIENTS.TXT	218
Listing of Raw Data File PATIENTS.TXT	218
Program to Create the SAS Data Set PATIENTS	219
Listing of Raw Data File PATIENTS2.TXT	220
Program to Create the SAS Data Set PATIENTS2	221
Program to Create the SAS Data Set AE (Adverse Events)	221
Program to Create the SAS Data Set LAB_TEST	222
Listings of the Data Cleaning Macros Used in This Book	222

**Index**

239

