

Table of Contents

List of Programs	ix
Introduction	xvii
Acknowledgments	xix

1

Checking Values of Character Variables

Introduction	1
Using PROC FREQ to List Values	1
Description of the File PATIENTS.TXT	2
Using a DATA Step to Check for Invalid Values	6
Using PROC PRINT with a WHERE Statement to List Invalid Values	11
Using Formats to Check for Invalid Values	13
Using Informats to Check for Invalid Values	17

2

Checking Values of Numeric Variables

Introduction	21
Using PROC MEANS, PROC TABULATE, and PROC UNIVARIATE to Look for Outliers	22
Using PROC PRINT with a WHERE Statement to List Invalid Data Values	32
Using a DATA Step to Check for Invalid Values	33
Creating a Macro for Range Checking	34
Using Formats to Check for Invalid Values	37
Using Informats to Check for Invalid Values	40
Using PROC UNIVARIATE to Look for Highest and Lowest Values by Percentage	43
Using PROC RANK to Look for Highest and Lowest Values by Percentage	48
Extending PROC RANK to Look for Highest and Lowest "n" Values	51

Finding Another Way to Determine Highest and Lowest Values	55
Checking a Range Using an Algorithm Based on Standard Deviation	58
Macros Based on the Two Methods of Outlier Detection	62
Demonstrating the Difference between the Two Methods	64
Checking a Range Based on the Interquartile Range	65
Checking Ranges for Several Variables	68

3 **Checking for Missing Values**

Introduction	73
Inspecting the SAS Log	73
Using PROC MEANS and PROC FREQ to Count Missing Values	76
Using DATA Step Approaches to Identify and Count Missing Values	78
Using PROC TABULATE to Count Missing and Nonmissing Values for Numeric Variables	82
Using PROC TABULATE to Count Missing and Nonmissing Values for Character Variables	83
Creating a General Purpose Macro to Count Missing and Nonmissing Values for Both Numeric and Character Variables	84
Searching for a Specific Numeric Value	88

4 **Working with Dates**

Introduction	93
Checking Ranges for Dates (Using a DATA Step)	94
Checking Ranges for Dates (Using PROC PRINT)	95
Checking for Invalid Dates	95
Working with Dates in Nonstandard Form	99
Creating a SAS Date When the Day of the Month Is Missing	101
Suspending Error Checking for Known Invalid Dates	103

5 Looking for Duplicates and "n" Observations per Subject

Introduction	105
Eliminating Duplicates by Using PROC SORT	105
Detecting Duplicates by Using DATA Step Approaches	110
Using PROC FREQ to Detect Duplicate ID's	113
Selecting Patients with Duplicate Observations by Using a Macro List and SQL	115
Identifying Subjects with "n" Observations Each (DATA Step Approach)	117
Identifying Subjects with "n" Observations Each (Using PROC FREQ)	119

6 Working with Multiple Files

Introduction	121
Checking for an ID in Each of Two Files	121
Checking for an ID in Each of "n" Files	124
A Simple Macro to Check ID's in Multiple Files	126
A More Complicated Multi-File Macro for ID Checking	129
More Complicated Multi-File Rules	131
Checking That the Dates Are in the Proper Order	134

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

Introduction	137
Conducting a Simple Comparison of Two Data Sets without an ID Variable	138
Using PROC COMPARE with an ID Variable	144
Using PROC COMPARE with Two Data Sets That Have an Unequal Number of Observations	146
Comparing Two Data Sets When Some Variables Are Not in Both Data Sets	149

8 **Some SQL Solutions to Data Cleaning**

Introduction	153
A Quick Review of PROC SQL	154
Checking for Invalid Character Values	155
Checking for Outliers	156
Checking a Range Using an Algorithm Based on the Standard Deviation	158
Checking for Missing Values	159
Range Checking for Dates	161
Checking for Duplicates	162
Identifying Subjects with "n" Observations Each	163
Checking for an ID in Each of Two Files	163
More Complicated Multi-File Rules	165

9 **Using Validation Data Sets**

Introduction	169
A Simple Example of a Validation Data Set	169
Making the Program More Flexible and Converting It to a Macro	174
Validating Character Data	180
Converting Program 9-7 into a General Purpose Macro	187
Extending the Validation Macro to Include Valid Character Ranges	191
Combining Numeric and Character Validity Checks in a Single Macro with a Single Validation Data Set	197
Introducing SAS Integrity Constraints (Versions 7 and Later)	207

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

Description of the Raw Data File PATIENTS.TXT	213
Layout for the Data File PATIENTS.TXT	214
Listing of Raw Data File PATIENTS.TXT	215
Program to Create the SAS Data Set PATIENTS	216
Listing of Raw Data File PATIENTS2.TXT	217
Program to Create the SAS Data Set PATIENTS2	217
Program to Create the SAS Data Set AE (Adverse Events)	218
Program to Create the SAS Data Set LAB_TEST	219

Index

