

LECTURE 03

21CSE295T

Statistical Data Analytics And Integration

Unit 02: Data Analysis Using SAS

2.1 Basic Reports
2.2 Selecting Variables
2.3 Identifying Observations
2.4 Subsetting the Data
2.5 Limiting the Observations
2.6 SORT Procedure
2.7 Generating Column Totals
2.8 Specifying Titles and Footnotes in Procedure Output
2.9 Assigning Descriptive Labels and Permanent Labels
2.25 PROC PRINT Procedure
2.26 WHERE Statement
2.27 Comparison Operators (EQ, NE, GT, LT, GE, LE)
2.28 AND / OR / CONTAINS

Virupakshan K

+91-9840847872 | VIRU@KART.LA

SYLLABUS

21CSE295T Statistical Data Analytics And Integration

VirupakshanK@gmail.com

Unit 1 Introduction to Statistical Analysis

9 hours

Computational Statistics- Statistical Pattern Recognition, Nonparametric Regression, Exploratory Data Analysis, Monte Carlo Methods for Inferential Statistics-Basic concepts of statistics, Types of data (categorical, numerical), Populations and samples, Measures of central tendency (mean, median, mode), Measures of variability (range, variance, standard deviation), Descriptive vs. inferential statistics.

Tutorial:

- SAS Statements, Structure, Libraries, Rules, Datasets, Representation of missing values, Assigning Libraries, PROC CONTENTS, visualize the descriptor and datasets attributes

Unit 2 Data Analysis Using SAS

9 hours

Basic Reports – Selecting Variables – Identifying observations – Subsetting the Data – Limiting the observations-SORT Procedure – Generating Column totals – Specifying titles and footnotes in procedure output – Assigning Descriptive labels and permanent label -How SAS Processes Programs – Compilation Phase – PDV – Descriptor Portion-Execution Phase – Data Portion – Debugging a Data Step – Testing Programs-BY-Group Processing Definitions – Preprocessing Data – Determine Whether the Data Requires Preprocessing – Sorting Observations for BY-Group Processing-FIRST, and LAST, DATA Step Variables – How the DATA Step Identifies BY Groups – Assignment Statements – Operators in SAS Expressions

Tutorial:

- Introduction of report and PRINT Procedure-Hands on practice of VAR, ID, WHERE statement, where expression[eq, ne, gt, lt, ge, le, AND, OR, CONTAINS]

Unit 3 Advanced Statistical Techniques in SAS

9 hours

Determining the Structure and Contents of Data Sets – Testing Program – Methods of Combining SAS Data Sets-One-to-One Reading – Concatenating – Match-Merging – Renaming Variables – Excluding Unmatched Observations -Selecting Matching Observations -Basics of DO Loops – Executing DO Loops -OUTPUT Statements – Nesting DO Loops-Iteratively Processing Observations from a Data Set -DO UNTIL Statement – DO WHILE Statement Defining and Processing Arrays –Arrays -Applying SAS Formats and Informats – Specifying SAS Formats

Tutorial:

- One-to-One Reading to Combine Data Sets-Concatenating to Combine Data Sets-Match-Merging to Combine Data Sets-Renaming Variables-Processing Iterative DO Loops-Indenting and Nesting DO Groups

Unit 4 SAS Programming and Automation

9 hours

Definitions – Reading Date and Time with Informats – Displaying Date and Time Values with Formats-Basics of SAS Functions - SAS Functions Syntax – Converting Data with Functions- The MEANS Procedure – Specifying Descriptive Statistics – Limiting Decimal Places – Specifying Variable Using VAR Statement – Summarized Data Set Using OUTPUT Statement – The FREQ Procedure-Specifying Variable Using TABLES Statements- Two-Way and N-Way Tables – LIST Options-The Output Delivery System (ODS) – HTML Outputs – PDF Outputs-RTF Outputs – EXCEL Outputs – The EXPORT Procedure

Tutorial:

- Reading Date and Time with Formats and Informats- Execution of Data Values Using Numeric Functions-Creating a Descriptive Statistics by Using PROC MEANS-Creating a Descriptive Statistics by Using PROC FREQ-Creating Outputs with HTML, PDF, RTF, EXCEL Formats-Exporting a External File

Unit 5 Advanced SAS Topics

9 hours

Creating a Default List Report – Windowing/Non Windowing Mode - Define Statement-Defining Variables – Column Statement – Selecting Observations – Defining the Variables-Group Processing the Variables by Using Different Options – Computing the Variable-Bar Chart – Box Plot – Mean Measurement Over the Time using Line Plot – Spaghetti Plot – Survival Plot – Waterfall Plot – Forest Plot-Introduction to Macro Facility – SAS Programs and Macro Processing – Macro Variables – Macro Processing – Scopes of Macro Variables – Macro Expressions-Scopes of Macro Variables – Macro Expressions-Macro Quoting – Interfaces with Macro Facility -Storing and Reusing Macros.

Tutorial:

- Creating Graphs by Using Bar Chart -Creating Graphs Using Various Types of Plots-Generating SAS Code Using Macros-Defining User-Defined Macro Variables-Creating a Tables Using PROC SQL with different Clauses- Generating a Cartesian Product -Combining Tables Using Multiple Joins

VirupakshanK@gmail.com

VirupakshanK@gmail.com

Unit 02: Data Analysis Using SAS

VirupakshanK@gmail.com

2.1 Basic Reports

2.2 Selecting Variables

2.3 Identifying Observations

2.4 Subsetting the Data

2.5 Limiting the Observations

2.6 SORT Procedure

2.7 Generating Column Totals

2.8 Specifying Titles and Footnotes in Procedure Output

2.9 Assigning Descriptive Labels and Permanent Labels

2.25 PROC PRINT Procedure

2.26 WHERE Statement

2.27 Comparison Operators (EQ, NE, GT, LT, GE, LE)

2.28 AND / OR / CONTAINS

2.10 SAS Process Programs

2.11 Compilation Phase

2.12 Program Data Vector (PDV)

2.13 Descriptor Portion

2.14 Execution Phase

2.15 Data Portion

2.18 BY-Group Processing Definitions

2.19 Preprocessing Data

2.20 Determine Whether the Data Requires Preprocessing

2.21 Sorting Observations for BY-Group Processing

2.22 FIRST. and LAST. DATA Step Variables

2.16 Debugging a DATA Step

2.17 Testing Programs

2.23 Assignment Statements

2.24 Operators in SAS Expressions

Lecture 03 Topics

VirupakshanK@gmail.com

2.1 Basic Reports

2.2 Selecting Variables

2.3 Identifying Observations

2.4 Subsetting the Data

2.5 Limiting the Observations

2.6 SORT Procedure

2.7 Generating Column Totals

2.8 Specifying Titles and Footnotes in Procedure Output

2.9 Assigning Descriptive Labels and Permanent Labels

2.25 PROC PRINT Procedure

2.26 WHERE Statement

2.27 Comparison Operators (EQ, NE, GT, LT, GE, LE)

2.28 AND / OR / CONTAINS

2.10 SAS Process Programs

2.1 Basic Reports

2.2 Selecting Variables

2.3 Identifying Observations

2.4 Subsetting the Data

2.5 Limiting the Observations

2.6 SORT Procedure

2.7 Generating Column Totals

2.8 Specifying Titles and Footnotes in Procedure Output

2.9 Assigning Descriptive Labels and Permanent Labels

2.25 PROC PRINT Procedure

2.26 WHERE Statement

2.27 Comparison Operators (EQ, NE, GT, LT, GE, LE)

2.28 AND / OR / CONTAINS

2.1 Basic Reports

A **basic report** is a simple printed output that displays observations and variables from a SAS dataset.

The most commonly used procedure for creating basic reports is:

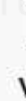
```
sas  
  
PROC PRINT DATA=dataset-name;  
RUN;
```

Diagram: Basic Report Flow

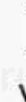
Raw Data



SAS Dataset



PROC PRINT



Printed Report

2.2.1 Selecting Variables

Selecting variables means choosing only specific columns from a dataset for display or processing.

In SAS, variables can be selected using:

- VAR statement in procedures
- KEEP statement or option
- DROP statement or option

Using VAR Statement

```
PROC PRINT DATA=students;  
  VAR Name Marks;  
RUN;
```

This displays only Name and Marks.

Using KEEP Statement

```
DATA student_marks;  
  SET students;  
  KEEP Name Marks;  
RUN;
```

Using DROP Statement

```
DATA student_details;  
  SET students;  
  DROP Gender;  
RUN;
```

2.2.2 Selecting Variables: KEEP vs DROP

Statement	Meaning
KEEP	Keeps only selected variables
DROP	Removes selected variables

```
DATA student_marks;  
SET students;  
KEEP Name Marks;  
RUN;
```

```
DATA student_details;  
SET students;  
DROP Gender;  
RUN;
```

2.3 Identifying Observations

An **observation** is a row in a SAS dataset.

Identifying observations means selecting or recognizing specific rows based on conditions.

Example:

```
PROC PRINT DATA=students;  
WHERE Marks > 80;  
RUN;
```

This identifies students who scored more than 80.

Output

ID	Name	Gender	Marks	Age
102	Anu	F	85	19
104	Meena	F	92	20

2.4 Subsetting the Data

Subsetting means creating a smaller dataset from a larger dataset by selecting certain observations or variables.

Types of Subsetting

Type	Description
Variable subsetting	Selecting specific columns
Observation subsetting	Selecting specific rows
Combined subsetting	Selecting both rows and columns

Example

```
DATA high_scores;  
SET students;  
WHERE Marks >= 80;  
KEEP Name Marks;  
RUN;
```

This creates a dataset with only students having marks greater than or equal to 80 and keeps only Name and Marks.

2.5 Limiting the Observations

Limiting observations means displaying or processing only a certain number of rows.

Using OBS= Option

```
PROC PRINT DATA=students(OBS=2);  
RUN;
```

This prints only the first 2 observations.

Using FIRSTOBS= Option

```
PROC PRINT DATA=students(FIRSTOBS=2 OBS=4);  
RUN;
```

This prints observations from row 2 to row 4.

2.6 SORT Procedure

PROC SORT is used to arrange observations in ascending or descending order.

Syntax

```
PROC SORT DATA=input-dataset OUT=output-dataset;  
    BY variable-name;  
RUN;
```

Example: Ascending Order

```
PROC SORT DATA=students OUT=sorted_students;  
    BY Marks;  
RUN;
```

This sorts students by marks in ascending order.

Example: Descending Order

```
PROC SORT DATA=students OUT=sorted_students;  
    BY DESCENDING Marks;  
RUN;
```

Important Point

If OUT= is not used, SAS overwrites the original dataset.

```
PROC SORT DATA=students;  
    BY Age;  
RUN;
```

2.7 Generating Column Totals

Column totals are totals calculated for numeric variables.

In PROC PRINT, use the SUM statement.

Example

```
PROC PRINT DATA=students;  
  VAR Name Marks;  
  SUM Marks;  
RUN;
```

This prints student marks and gives the total of the Marks column.

Formula

If marks are: 78, 85, 67, 92

Then total marks: $78+85+67+92=322$

2.8 Specifying Titles and Footnotes in Procedure Output

Titles and footnotes improve the readability and presentation of SAS output.

TITLE Statement

```
TITLE "Student Marks Report";
```

FOOTNOTE Statement

```
FOOTNOTE "Prepared by Department of Statistics";
```

Example

```
TITLE "Student Marks Report";  
FOOTNOTE "End of Report";  
  
PROC PRINT DATA=students;  
RUN;  
  
TITLE;  
FOOTNOTE;
```

Important Point

Use empty TITLE; and FOOTNOTE; statements to clear previous titles and footnotes.

2.9.1 Assigning Descriptive Labels and Permanent Labels

A **label** is a descriptive name assigned to a variable.

Variable names are often short, but labels can be more meaningful.

Example:

Variable Name	Label
ID	Student Identification Number
Marks	Examination Marks

Temporary Label in PROC PRINT

```
PROC PRINT DATA=students LABEL;  
  LABEL ID = "Student ID"  
        Marks = "Exam Marks";  
RUN;
```

Permanent Label in DATA Step

```
DATA students_labeled;  
  SET students;  
  LABEL ID = "Student Identification Number"  
        Marks = "Examination Marks";  
RUN;
```

2.9.2 Assigning Descriptive Labels and Permanent Labels

Temporary Label in PROC PRINT

```
PROC PRINT DATA=students LABEL;  
    LABEL ID = "Student ID"  
           Marks = "Exam Marks";  
RUN;
```

Permanent Label in DATA Step

```
DATA students_labeled;  
    SET students;  
    LABEL ID = "Student Identification Number"  
           Marks = "Examination Marks";  
RUN;
```

Difference Between Temporary and Permanent Labels

Type	Description
Temporary label	Used only for current output
Permanent label	Stored in dataset descriptor portion

NOTE: The following topics will be in “Lecture 04”

2.10	SAS Process Programs
2.11	Compilation Phase
2.12	Program Data Vector (PDV)
2.13	Descriptor Portion
2.14	Execution Phase
2.15	Data Portion
2.16	Debugging a DATA Step
2.17	Testing Programs
2.18	BY-Group Processing Definitions
2.19	Preprocessing Data
2.20	Determine Whether the Data Requires Preprocessing
2.21	Sorting Observations for BY-Group Processing
2.22	FIRST. and LAST. DATA Step Variables
2.23	Assignment Statements
2.24	Operators in SAS Expressions

2.25 PROC PRINT Procedure

Basic Syntax

```
PROC PRINT DATA=dataset-name;  
RUN;
```

PROC PRINT displays the contents of a SAS dataset.

Common PROC PRINT Statements

Statement	Purpose
VAR	Select variables
WHERE	Select observations
ID	Use a variable instead of Obs column
SUM	Generate totals
BY	Group output
LABEL	Display labels

Example

```
PROC PRINT DATA=students LABEL;  
ID ID;  
VAR Name Gender Marks;  
WHERE Marks >= 70;  
SUM Marks;  
LABEL Marks = "Exam Marks";  
RUN;
```

2.26.1 WHERE Statement

The WHERE statement selects observations that satisfy a condition.

Syntax `WHERE condition;`

Example

```
PROC PRINT DATA=students;  
    WHERE Marks > 80;  
RUN;
```

WHERE in DATA Step

```
DATA high_scores;  
    SET students;  
    WHERE Marks > 80;  
RUN;
```

2.26.2 WHERE Statement: WHERE vs IF

WHERE vs IF

Feature	WHERE	IF
Used before reading into PDV	Yes	No
Works with existing variables only	Yes	No
Can use newly created variables	No	Yes
More efficient for large datasets	Usually yes	Usually no

Example Difference

```
DATA test;  
SET students;  
Grade = Marks / 10;  
IF Grade >= 8;  
RUN;
```

Here, IF is used because Grade is newly created.

2.27 Comparison Operators (EQ, NE, GT, LT, GE, LE)

Comparison operators compare two values and return true or false.

Symbol	Mnemonic	Meaning
=	EQ	Equal to
^= or ~=	NE	Not equal to
>	GT	Greater than
<	LT	Less than
>=	GE	Greater than or equal to
<=	LE	Less than or equal to

Examples:

```
WHERE Marks EQ 85;  
WHERE Marks NE 67;  
WHERE Marks GT 80;  
WHERE Marks LT 70;  
WHERE Marks GE 75;  
WHERE Marks LE 90;
```

Equivalent Statements

```
WHERE Marks > 80;
```

is the same as:

```
WHERE Marks GT 80;
```

2.28 AND / OR / CONTAINS

AND Operator : The AND operator selects observations where all conditions are true.

```
PROC PRINT DATA=students;  
    WHERE Gender = "F" AND Marks > 80;  
RUN;
```

This selects female students with marks greater than 80.

OR Operator: The OR operator selects observations where at least one condition is true.

```
PROC PRINT DATA=students;  
    WHERE Gender = "F" OR Marks > 80;  
RUN;
```

This selects students who are female or scored more than 80.

CONTAINS Operator:

The CONTAINS operator searches for a substring inside a character variable.

```
PROC PRINT DATA=students;  
    WHERE Name CONTAINS "a";  
RUN;
```

This selects names containing the letter a.

Important Note:

Character comparisons are case-sensitive in SAS.

Lecture 03 Topics

VirupakshanK@gmail.com

2.1 Basic Reports

2.2 Selecting Variables

2.3 Identifying Observations

2.4 Subsetting the Data

2.5 Limiting the Observations

2.6 SORT Procedure

2.7 Generating Column Totals

2.8 Specifying Titles and Footnotes in Procedure Output

2.9 Assigning Descriptive Labels and Permanent Labels

2.25 PROC PRINT Procedure

2.26 WHERE Statement

2.27 Comparison Operators (EQ, NE, GT, LT, GE, LE)

2.28 AND / OR / CONTAINS

2.10 SAS Process Programs

2.1 Basic Reports

2.2 Selecting Variables

2.3 Identifying Observations

2.4 Subsetting the Data

2.5 Limiting the Observations

2.6 SORT Procedure

2.7 Generating Column Totals

2.8 Specifying Titles and Footnotes in Procedure Output

2.9 Assigning Descriptive Labels and Permanent Labels

2.25 PROC PRINT Procedure

2.26 WHERE Statement

2.27 Comparison Operators (EQ, NE, GT, LT, GE, LE)

2.28 AND / OR / CONTAINS

LECTURE 03

21CSE295T

Statistical Data Analytics And Integration

Doubts?

THANK YOU!

Virupakshan K

+91-9840847872 | VIRU@KART.LA