Lesson 2: Summarizing Data

Summary

Main Points

Creating an Accumulating Variable Using the RETAIN Statement

RETAIN variable-name <initial-value> …;

- An accumulating variable accumulates the value of another variable and keeps its value from one observation to the next. You can use the RETAIN statement to create an accumulating variable.
- The RETAIN statement is a compile-time-only statement that prevents SAS from reinitializing the variable at the top of the DATA step. Because the variable is not reinitialized, it retains its value across multiple iterations of the DATA step.
- The RETAIN statement starts with the keyword RETAIN followed by the name of the variable whose values you want to retain. You can optionally specify an initial value for the variable. If you don't specify an initial value, the RETAIN statement initializes the variable to missing before the first execution of the DATA step.

Creating an Accumulating Variable Using the Sum Statement

variable+expression;

- As an alternative to using the RETAIN statement with an assignment statement, you can use the sum statement. By default, the sum statement initializes the variable to 0 and retains the variable. It also ignores missing input values from the expression.

Using BY-Group Processing: Summarizing Data by Groups

DATA output-SAS-data-set;
SET input-SAS-data-set;
BY BY-variable …;
<additional SAS statements>
RUN;

- When you need to accumulate totals for a group of data, (for example, if you need to see total salaries allocated to special projects by department), the input data set needs to be sorted on the BY-variable.
- You can then use a BY statement in the DATA step to process the data in groups.
Lesson 2: Summarizing Data

**FIRST. BY-variable**
**LAST. BY-variable**

- The BY statement creates two temporary variables (FIRST.variable and LAST.variable) for each BY variable listed. These variables identify the first and last observation in each BY group.
- You can use the FIRST. and LAST. variables in a DATA step to summarize the grouped data.
  - First, set the accumulating variable equal to 0 at the start of each BY group.
  - Second, increment the accumulating variable with a sum statement.
  - Third, output only the last observation of each BY group.

Using BY-Group Processing: Summarizing Data by Multiple Groups

- When you need to accumulate totals for multiple groups, you specify two or more BY variables. The first variable listed becomes the primary sort variable, and the second variable listed is the secondary sort variable.
- The BY statement creates two temporary variables for each BY variable listed.
- If the last observation for a value of the primary sort variable is encountered, it sets LAST. to 1 for all subsequent BY variables.

Sample Code

**Creating an Accumulating Variable Using the RETAIN Statement**

```sas
data mnthtot2;
   set orion.aprsales;
   retain Mth2Dte 0;
   Mth2Dte=Mth2Dte+SaleAmt;
run;
```

**Creating an Accumulating Variable Using the Sum Statement**

```sas
data mnthtot2;
   set orion.aprsales;
   Mth2Dte+SaleAmt;
run;
```
Lesson 2: Summarizing Data

Using BY-Group Processing: Summarizing Data by Groups

```sas
data deptsals (keep=Dept DeptSal);
   set SalSort;
   by Dept;
   if First.Dept then DeptSal=0;
   DeptSal+Salary;
   if Last.Dept;
run;
```

Using BY-Group Processing: Summarizing Data by Multiple Groups

```sas
data deptsals (keep=Proj Dept DeptSal NumEmps);
   set projsort;
   by Proj Dept;
   if First.Dept then
      do;
         DeptSal=0;
         NumEmps=0;
      end;
   DeptSal+Salary;
   NumEmps+1;
   if Last.Dept;
run;
```