

## Anyone Can Learn PROC TABULATE

Lauren Haworth, Kaiser Permanente Center for Health Research, Portland, OR

### ABSTRACT

SAS provides hundreds of ways you can analyze your data. You can use the DATA step to slice and dice your data, and there are dozens of procedures that will process your data and produce all kinds of statistics. But odds are that no matter how you organize and analyze your data, you'll end up producing a report in the form of a table.

This is why every SAS user needs to know how to use PROC TABULATE. While TABULATE doesn't do anything that you can't do with other PROCs, the payoff is in the output. TABULATE computes a variety of statistics, and it neatly packages the results in a single table.

Unfortunately, TABULATE has gotten a bad rap as being a difficult procedure to learn. This paper will prove that if you take things step by step, anyone can learn PROC TABULATE.

### INTRODUCTION

This paper will start out with the most basic one-dimensional table. We will then go on to two-dimensional tables, tables with totals, and finally, three-dimensional tables. By the end of this paper, you will be ready to build most basic TABULATE tables.

### ONE-DIMENSIONAL TABLES

To really understand TABULATE, you have to start very simply. The simplest possible table in TABULATE has to have three things: a PROC TABULATE statement, a TABLE statement, and a CLASS or VAR statement. In this example, we will use a VAR statement. Later examples will show the CLASS statement.

The PROC TABULATE statement looks like this:

```
PROC TABULATE DATA=TEMP;
```

The second part of the procedure is the TABLE statement. It describes which variables to use and how to arrange the variables. This first table will have only one variable, so you don't have to tell TABULATE where to put it. All you have to do is list it in the TABLE statement. When there is only one variable, you get a one-dimensional table.

```
PROC TABULATE DATA=TEMP;  
TABLE INCOME;  
RUN;
```

If you run this code as is, you will get an error message because TABULATE can't figure out whether the variable INCOME is intended as an analysis variable, which is used to compute statistics, or a classification variable, which is used to define row or column categories in the table.

In this case, we want to use income as the analysis variable. We will be using it to compute a statistic. To tell TABULATE that INCOME is an analysis variable, you use a VAR statement. The syntax of a VAR statement is simple: you just list the variables that will be used for analysis. So now the syntax for our PROC TABULATE is:

```
PROC TABULATE DATA=TEMP;  
VAR INCOME;  
TABLE INCOME;  
RUN;
```

The result is the table shown below. It has a single column, with the header INCOME to identify the variable, and the header SUM to identify the statistic. There is just a single table cell, which contains the value for the sum of INCOME for all of the observations in the dataset TEMP.

| INCOME       |
|--------------|
| SUM          |
| 166411827.00 |

### ADDING A STATISTIC

The previous table shows what happens if you don't tell TABULATE which statistic to use. If the variable in your table is an analysis variable, meaning that it is listed in a VAR statement, then the statistic you will get by default is the sum. Sometimes the sum will be the statistic that you want. Most likely, sum isn't the statistic that you want.

To add a statistic to a PROC TABULATE table, you modify the TABLE statement. You list the statistic right after the variable name. To tell TABULATE that the statistic MEAN should be applied to the variable INCOME, you use an asterisk to link the vari-

## Core Skill Development

able name to the statistic keyword. The asterisk is a TABULATE operator. Just as you use an asterisk as an operator when you want to multiply 2 by 3 (2\*3), you use an asterisk when you want to apply a statistic to a variable.

```
PROC TABULATE DATA=TEMP;  
  VAR INCOME;  
  TABLE INCOME*MEAN;  
RUN;
```

The output with the new statistic is shown in below. Note that the variable name at the top of the column heading has remained unchanged. However, the statistic name that is shown in the second line of the heading now says "MEAN." And the value shown in the table cell has changed from the sum to the mean.

| INCOME   |
|----------|
| MEAN     |
| 25065.80 |

### ADDING ANOTHER STATISTIC

Each of the tables shown so far was useful, but the power of PROC TABULATE comes from being able to combine several statistics and/or several variables in the same table. TABULATE does this by letting you specify a series of "tables" within a single large table. We're going to add a "table" showing the number of observations to our table showing the mean income.

The first part of our combined table is the code we used before to compute mean income.

```
PROC TABULATE DATA=TEMP;  
  VAR INCOME;  
  TABLE INCOME*MEAN;  
RUN;
```

Next, we can add similar code to our TABLE statement to get the number of observations. To add this statistic to the first table, all you do is combine the code for the mean ("INCOME\*MEAN") with the code you would use to get the number of observations ("INCOME\*N"). The code for the two "tables" is combined by using a space between the two statements. The space operator tells TABULATE that you want to add another column to your table.

```
PROC TABULATE DATA=TEMP;  
  VAR INCOME;  
  TABLE INCOME*N INCOME*MEAN;  
RUN;
```

The resulting table is shown in below.

| INCOME  | INCOME   |
|---------|----------|
| N       | MEAN     |
| 6639.00 | 25065.80 |

Note that the additional statistic is shown as an additional column in the table. When SAS is creating a one-dimensional table, additional variables, statistics, and categories are always added as new columns.

### ADDING A CLASSIFICATION VARIABLE

After seeing the tables we've built so far in this chapter, you're probably asking yourself, "Why use PROC TABULATE? Everything I've seen so far could be done with a PROC MEANS."

One answer to this question is classification variables. By specifying a variable to categorize your data, you can produce a concise table that shows values for various subgroups in your data. For example, wouldn't it be more interesting to look at mean income if it were broken down by the education level of the survey respondent?

To break down income by education level, we will use education as a classification variable. Just as we used a VAR statement to identify our *analysis* variable, we use a CLASS statement to identify a *classification* variable. By putting a variable in a CLASS statement, we are telling TABULATE that the variable will be used to identify categories of the data. In this case we will add EDUC as a class variable.

The other thing we have to do to our code is tell TABULATE where to put the classification variable EDUC in the table. We do this by again using the asterisk operator. By adding another asterisk to the end of the TABLE statement, and following it with the variable name EDUC, TABULATE knows that EDUC will be used to categorize the mean values of INCOME.

```
PROC TABULATE DATA=TEMP;  
  CLASS EDUC;  
  VAR INCOME;  
  TABLE INCOME*MEAN*EDUC;  
RUN;
```

The resulting table is shown below. Now the column headings have changed. The variable name INCOME and the statistic name MEAN are still there, but under the statistic label there are now three columns. Each column is headed by the variable label "EDUCATION" and the category name "< HS," "HS," and "COLLEGE." The values shown in the table cells now represent subgroup means.

| INCOME    |          |          |
|-----------|----------|----------|
| MEAN      |          |          |
| EDUCATION |          |          |
| <HS       | HS       | COLLEGE  |
| 11937.52  | 19581.80 | 32710.58 |

## TWO-DIMENSIONAL TABLES

You probably noticed that our example table is not very elegant in appearance. That's because it only takes advantage of one dimension. It has multiple columns, but only one row. It is much more efficient to build tables that have multiple rows and multiple columns. You can fit more information on a page, and the table looks better, too.

The easiest way to build a two-dimensional table is to build it one dimension at a time. First we'll build the columns, and then we'll add the rows.

For this first table, we'll keep things simple. This is the table we built in a previous example. It has two columns: one showing the number of observations for INCOME and another showing the mean of INCOME.

```
PROC TABULATE DATA=TEMP;
  VAR INCOME;
  TABLE INCOME*N INCOME*MEAN;
RUN;
```

This table is shown below.

| INCOME  | INCOME   |
|---------|----------|
| N       | MEAN     |
| 6639.00 | 25065.80 |

To turn this table into a two-dimensional table, we will add another variable to the TABLE statement. In this case, we want to add rows that show the N and MEAN of INCOME for each level of education.

To add another dimension to the table, you use a comma as an operator. All you do is put a comma between the old part of the TABLE statement and the new part of the TABLE statement.

If a TABLE statement has no commas, then it is assumed that the variables and statistics are to be created as columns. If a TABLE statement has two parts, separated by a comma, then TABULATE builds a two-dimensional table using the first part of the TABLE statement as the rows and the second part of the TABLE statement as the columns.

So to get a table with income as the columns and education as the rows, we just need to add a comma and the variable EDUC. Since we want to add EDUC as a row, we list it before the rest of the TABLE statement. If we wanted to add it as a column, we'd add it to the end of the TABLE statement.

```
PROC TABULATE DATA=TEMP;
  VAR INCOME;
  CLASS EDUC;
  TABLE EDUC, INCOME*N INCOME*MEAN;
RUN;
```

This table is shown below.

|           | INCOME  | INCOME   |
|-----------|---------|----------|
|           | N       | MEAN     |
| EDUCATION |         |          |
| <HS       | 1001.00 | 11937.52 |
| HS        | 2282.00 | 19581.80 |
| COLLEGE   | 3356.00 | 32710.58 |

By the way, there is a limit to which variables you can use in a two-dimensional table. You can't have a cross-tabulation of two analysis variables. A two-dimensional table must have at least one classification variable (i.e., you must have a CLASS statement). If you think about it, this makes sense. A table of mean income by mean years of education would be meaningless, but a table of mean income by categories of education makes perfect sense.

## ADDING CLASSIFICATION VARIABLES ON BOTH DIMENSIONS

The previous example showed how to reformat a table from one to two dimensions, but it did not show the true power of two-dimensional tables. With two dimensions, you can classify your statistics by two different variables at the same time.

To do this, you put one classification variable in the row dimension and one classification in the column dimension. The previous example had education displayed in rows, and income as the column variable. In this new table, we will add employment status as an additional column variable. Instead of just displaying the MEAN of the variable INCOME for each level of education, we will display the statistic broken down by part-time versus full-time employment status.

So in the following code, we leave EDUC as the row variable, and we leave INCOME in the column dimension. The only change is to add EMPLOY to the column dimension using the asterisk operator. This

## Core Skill Development

tells TABULATE to break down each of the column elements into categories of employment status.

```
PROC TABULATE DATA=TEMP;
  VAR INCOME;
  CLASS EDUC EMPLOY;
  TABLE EDUC, INCOME*EMPLOY*MEAN;
RUN;
```

This table is shown below. Notice how the analysis variable INCOME remains as the column heading, and MEAN remains as the statistic, but now there are additional column headings to show the two categories of EMPLOY.

|           | INCOME    |           |
|-----------|-----------|-----------|
|           | EMPLOY    |           |
|           | Full-time | Part-time |
|           | MEAN      | MEAN      |
| EDUCATION |           |           |
| <HS       | 19267.53  | 19755.32  |
| HS        | 24798.92  | 23363.48  |
| COLLEGE   | 38179.60  | 35092.72  |

### ADDING ANOTHER CLASSIFICATION VARIABLE

The previous example showed how to add a classification variable to both the rows and columns of a two-dimensional table. But you are not limited to just one classification per dimension. This next example will show how to display additional subgroups of the data.

In this case, we're going to add gender as an additional row classification. The variable is added to the CLASS statement and to the row dimension of the TABLE statement. It is added using a space as the operator, so we will get rows for education followed by rows for gender.

```
PROC TABULATE DATA=TEMP;
  VAR INCOME;
  CLASS EDUC EMPLOY SEX;
  TABLE EDUC SEX,
    INCOME*EMPLOY*MEAN;
RUN;
```

In the results shown below, you can see that we now have two two-dimensional "mini-tables" within a single table. First we have a table of income by employment status and education, and then we have a table showing income by employment status and gender.

|           | INCOME    |           |
|-----------|-----------|-----------|
|           | EMPLOY    |           |
|           | Full-time | Part-time |
|           | MEAN      | MEAN      |
| EDUCATION |           |           |
| <HS       | 19267.53  | 19755.32  |
| HS        | 24798.92  | 23363.48  |
| COLLEGE   | 38179.60  | 35092.72  |
| SEX       |           |           |
| FEMALE    | 23821.25  | 24048.34  |
| MALE      | 40656.44  | 35598.25  |

This ability to stack multiple mini-tables within a single table can be a powerful tool for delivering large quantities of information in a user-friendly format.

### NESTING THE CLASSIFICATION VARIABLES

So far all we have done is added additional "tables" to the bottom of our first table. We used the space operator between each of the row variables to produce stacked tables.

By using a series of row variables, we can explore a variety of relationships between the variables. In previous table, we could see how income varies by education and employment status, and we could see how income varies by gender and employment status, but we could not see how education and gender interacted to affect income for each employment status.

But we can learn even more from our data. The power of TABULATE comes from being able to look at combinations of categories within a single table. In the following example, we will build a table to look at income by employment status for combinations of education and gender.

This code is the same as we used for the last example. The only change is that in the row definition, the asterisk operator is used to show that we want to nest the two row variables. In other words, we want to see the breakdown of income by gender within each category of education.

```
PROC TABULATE DATA=TEMP;
  VAR INCOME;
  CLASS EDUC EMPLOY SEX;
  TABLE EDUC*SEX,
    INCOME*EMPLOY*MEAN;
RUN;
```

As you can see in the table below, this code produces nested categories within the row headings. The row headings are now split into two columns. The first column shows education and the second shows gender. It is now easier to interpret the interaction of gender and education.

|           |        | INCOME    |           |
|-----------|--------|-----------|-----------|
|           |        | EMPLOY    |           |
|           |        | Full-time | Part-time |
|           |        | MEAN      | MEAN      |
| EDUCATION | SEX    |           |           |
| <= HS     | FEMALE | 14850.22  | 14449.73  |
|           | MALE   | 22625.90  | 25060.91  |
| HS        | FEMALE | 18189.30  | 17574.26  |
|           | MALE   | 31085.59  | 28307.08  |
| COLLEGE   | FEMALE | 27745.10  | 28767.81  |
|           | MALE   | 47960.27  | 42817.40  |

You can also reverse the order of the row variables to look at education within gender, instead of gender within education. All you do is move SEX so that it comes before EDUC. TABULATE always produces the nested rows in the order the variables are listed on the TABLE statement.

#### ADDING TOTALS TO THE ROWS AND COLUMNS

One trick that will definitely make your tables more readable is row and column totals. Totals are quite easy to generate in TABULATE because you can use the ALL variable. This is a built in classification variable supplied by TABULATE that stands for "all observations." You do not have to list it in the CLASS statement because it is a classification variable by definition.

The following code produces a table similar to the previous example, but with the addition of row totals. The statistic is changed to N so that you can see how the totals work.

```
PROC TABULATE DATA=TEMP;
  CLASS EDUC SEX EMPLOY;
  TABLE EDUC*SEX, EMPLOY*N ALL*N;
RUN;
```

As you can see from the table below, the table now has row totals.

|       |        | EMPLOY    |           |         |
|-------|--------|-----------|-----------|---------|
|       |        | Full-time | Part-time | ALL     |
|       |        | N         | N         | N       |
| EDUC  | SEX    |           |           |         |
| <= HS | FEMALE | 695.00    | 98.00     | 793.00  |
|       | MALE   | 760.00    | 111.00    | 871.00  |
| > HS  | FEMALE | 1092.00   | 149.0     | 1241.00 |
|       | MALE   | 1165.00   | 122.00    | 1287.00 |

Not only can you use ALL to add row totals, but you can also use ALL to produce column totals. What you do is list ALL as an additional variable in the row definition of the TABLE statement. No asterisk is needed because we just want to add a total at the bottom of the table.

```
PROC TABULATE DATA=TEMP;
  CLASS EDUC2 SEX EMPLOY;
  TABLE EDUC2*SEX ALL, EMPLOY*N;
RUN;
```

The resulting table is shown below. Now there are overall totals for each employment status.

|       |        | EMPLOY    |           |
|-------|--------|-----------|-----------|
|       |        | Full-time | Part-time |
|       |        | N         | N         |
| EDUC  | SEX    |           |           |
| <= HS | FEMALE | 695.00    | 98.00     |
|       | MALE   | 760.00    | 111.00    |
| > HS  | FEMALE | 1092.00   | 149.00    |
|       | MALE   | 1165.00   | 122.00    |
| ALL   |        | 3712.00   | 480.00    |

#### THREE-DIMENSIONAL TABLES

Now that you have mastered two-dimensional tables, let's add a third dimension. You may be asking yourself: Three dimensions? How do you print a table shaped like a cube?

Actually, a three-dimensional table is not shaped like a cube, it just looks like a two-dimensional table, except that it spans multiple pages. A one-dimensional table just has columns. A two-dimensional table has both columns and rows. A three-dimensional table is just a two-dimensional table that is repeated across multiple pages. Basically,

## Core Skill Development

you print a new page for each value of the page variable.

The hardest part about three-dimensional tables is making sense of the TABLE statement. So the best way to start is with the first two dimensions: the rows and columns. Once you've got that set up correctly, it's relatively easy to add the page variable to expand the table to multiple pages.

For our example, we're going to build a table of income by occupation and gender, and then we're going to add sector as the page variable. We'll end up with two pages of tables, the first page will have the first sector ("Private"), and the second page will have the second sector ("Public").

Ignoring the third dimension for now, let's build the basic table. This table has rows showing occupation, and columns showing income by gender. The code is as follows:

```
PROC TABULATE DATA=TEMP;
  CLASS OCCUP GENDER;
  VAR INCOME;
  TABLE OCCUP,
    (GENDER ALL)*INCOME*MEAN;
RUN;
```

At this point you should run the code and look the table over carefully to be sure you've got exactly what you want to see in your final table. The only difference between this table and our final three-dimensional table is that right now, the table is showing results for both SECTOR="Private" and SECTOR="Public" combined. In the final tables, each page will have the results for just one of the two sectors.

Assuming the two-dimensional table looks correct, we'll go on to adding the third dimension. Just as when we converted a one-dimensional table to a two-dimensional table, we add a new dimension with a comma operator in the TABLE statement. We just add the new variable SECTOR to the existing TABLE statement with a comma to separate it from the row and column definitions.

You might think that the following code is the correct way to add the third dimension to the table statement:

```
TABLE OCCUP,
  (GENDER ALL)*INCOME*MEAN,
  SECTOR;
```

However, what this would generate is a table with GENDER as the rows, SECTOR as the columns, and OCCUP as the pages! In order to add the third dimension to a table, you add it at the beginning of the table statements. Remember that this was also true

when we added the second dimension to the table. We added rows to the columns by adding a row definition before the column definition. The correct code for our table is:

```
PROC TABULATE DATA=TEMP;
  CLASS OCCUP GENDER SECTOR;
  VAR INCOME;
  TABLE SECTOR, OCCUP,
    (GENDER ALL)*INCOME*MEAN;
RUN;
```

The resulting tables are shown below. To save space, both tables are shown on a single page. In reality, TABULATE puts a page break between each of the tables, and the table for SECTOR="Public" would appear on a second page. Notice that each table was automatically given a title that defines the sector it represents.

**SECTOR Private**

|              | GENDER   |          |          |
|--------------|----------|----------|----------|
|              | Male     | Female   | ALL      |
|              | INCOME   | INCOME   | INCOME   |
|              | MEAN     | MEAN     | MEAN     |
| OCCUPATION   |          |          |          |
| Managerial   | 55948.92 | 32337.14 | 46075.54 |
| Professional | 51926.11 | 31693.72 | 42098.03 |
| Technical    | 35336.78 | 25162.63 | 30209.65 |

**SECTOR Public**

|              | GENDER   |          |          |
|--------------|----------|----------|----------|
|              | Male     | Female   | ALL      |
|              | INCOME   | INCOME   | INCOME   |
|              | MEAN     | MEAN     | MEAN     |
| OCCUPATION   |          |          |          |
| Managerial   | 54836.21 | 32044.57 | 44390.04 |
| Professional | 41492.10 | 33635.86 | 36029.82 |
| Technical    | 29510.75 | 26975.44 | 28168.53 |

### CONTACTING THE AUTHOR

Please direct any questions or feedback to the author at:

Kaiser Permanente Center for Health Research  
3800 N. Kaiser Center Drive  
Portland, OR 97227-1098

E-mail: [haworthla@chr.mts.kpnw.org](mailto:haworthla@chr.mts.kpnw.org)