

Computing Percentage Using PROC TABULATE – from Simple to More Complex

Wende (Ted) Tian, Merck Sharpe & Dohme Corp., Rahway, NJ
Hong (Lily) Zhang, Merck Sharpe & Dohme Corp., Rahway, NJ

ABSTRACT

The TABULATE procedure can calculate the percentages of the total count or sum of analysis variables for individual cells. This paper will first review how to calculate basic percentage using PCTN, PCTSUM, ROWPCTN, ROWPCTSUM, COLPCTN and COLPCTSUM, and then focus on creating more complex percentages on concatenated tables and computing percentages of subtotals, etc.

INTRODUCTION

The TABULATE procedure displays descriptive statistics in tabular format using some or all of the variables in a data set. We can create a variety of tables ranging from simple to highly customized using the TABULATE procedure. The general syntax of PROC TABULATE is as follows:

```
PROC TABULATE <option(s)>;  
  BY <DESCENDING> variable-1  
  <...<DESCENDING> variable-n>  
  <NOTSORTED>;  
  CLASS variable(s) <| options>;  
  CLASSLEV variable(s) | STYLE=<style-element-name | PARENT> <[style-attribute-specification(s)] >;  
  FREQ variable;  
  KEYLABEL keyword-1='description-1'  
  <...keyword-n='description-n'>;  
  KEYWORD keyword(s) | STYLE=<style-element-name | PARENT> <[style-attribute-specification(s)] >;  
  TABLE <<page-expression,> row-expression,> column-expression<| table-option(s)>;  
  VAR analysis-variable(s) <| options>;  
  WEIGHT variable;
```

One of main applications of the TABULATE procedure is calculating percentages. This paper will next briefly review the percentage statistics used in TABULATE and then move to more complex calculations of percentage.

PERCENTAGE STATISTICS

In this paper, we use the following CDISC ADaM format test data to illustrate the application of the TABULATE procedure in calculating percentages:

SUBJID	SEX	SITE	TRT	AVISIT	PARAMCD	AVAL
51063	F	USA	Placebo	Week 0	ALT	9
51063	F	USA	Placebo	Week 0	AST	10
51063	F	USA	Placebo	Week 4	ALT	11
51063	F	USA	Placebo	Week 4	AST	10
51068	F	USA	Study drug	Week 0	ALT	11
51068	F	USA	Study drug	Week 0	AST	16
51068	F	USA	Study drug	Week 4	ALT	11
51068	F	USA	Study drug	Week 4	AST	13
51077	M	USA	Placebo	Week 0	ALT	13
51077	M	USA	Placebo	Week 0	AST	10
51077	M	USA	Placebo	Week 4	ALT	10
51077	M	USA	Placebo	Week 4	AST	9
51112	F	Ex-USA	Placebo	Week 0	ALT	22
51112	F	Ex-USA	Placebo	Week 0	AST	17
51112	F	Ex-USA	Placebo	Week 4	ALT	22
51112	F	Ex-USA	Placebo	Week 4	AST	18
51230	M	USA	Placebo	Week 0	ALT	18
51230	M	USA	Placebo	Week 0	AST	14
51230	M	USA	Placebo	Week 4	ALT	17
51230	M	USA	Placebo	Week 4	AST	14
51248	M	Ex-USA	Study drug	Week 0	ALT	25
51248	M	Ex-USA	Study drug	Week 0	AST	19
51248	M	Ex-USA	Study drug	Week 4	ALT	25
51248	M	Ex-USA	Study drug	Week 4	AST	17

Below is a summary of percentage statistics used in PROC TABULATE:

Statistic	Percentage Calculated	Require a denominator definition?
PCTN and PCTSUM	Display the percentage of the value in a single table cell in relation to the value (used in the denominator of the calculation of the percentage) in another table cell or to the total of the values in a group of cells. By default, PROC TABULATE summarizes the values in all N cells (for PCTN) or all SUM cells (for PCTSUM) and uses the summarized value for the denominator	Can specify a denominator definition other than the default
ROWPCTN and ROWPCTSUM	Display the percentage of the value in a single table cell in relation to the total of the values in the row.	No
COLPCTN and COLPCTSUM	Display the percentage of the value in a single table cell in relation to the total of the values in the column	No
REPPCTN and REPPCTSUM	Display the percentage of the value in a single table cell in relation to the total of the values in the report	No
PAGEPCTN and PAGEPCTSUM	Display the percentage of the value in a single table cell in relation to the total of the values in the page	No

In the above table, PCTN and XXXPCTN are used to calculate the percentage for class variables, and PCTSUM and XXXPCTSUM are used for analysis variables. Below are some of examples of these TABULATE procedure statistics:

PCTN AND PCTSUM

We will start with the simple SAS® code below to illustrate how to display percentage using PCTN and PCTSUM.

```
proc tabulate data=test.adlb (where=(paramcd='ALT' and avisit='Week 4'));
  class paramcd trt sex;
  var aval;
  table trt all, (sex='Gender (Count)' all)*n*f=4. (sex='Gender (%)' all)*PCTN
    /box='Example of using PCTN';

  table trt all, (sex='Gender (Count)' all)*aval='Lab Result'*sum*f=6.
    (sex='Gender (%)' all)*aval='Lab Result'*PCTSUM
    /box='Example of using PCTSUM';
run;
```

This code generates the following output:

Figure 1: Example of using PCTN

Example of using PCTN	Gender (Count)		All	Gender (%)		All
	F	M		F	M	
	N	N	N	PctN	PctN	PctN
Treatment Arm						
Placebo	2	2	4	33.33	33.33	66.67
Study drug	1	1	2	16.67	16.67	33.33
All	3	3	6	50.00	50.00	100.00

Figure 2: Example of using PCTSUM

Example of using PCTSUM	Gender (Count)		All	Gender (%)		All
	F	M		F	M	
	Lab Result	Lab Result	Lab Result	Lab Result	Lab Result	Lab Result
	Sum	Sum	Sum	PctSum	PctSum	PctSum
Treatment Arm						
Placebo	33	27	60	34.38	28.13	62.50
Study drug	11	25	36	11.46	26.04	37.50
All	44	52	96	45.83	54.17	100.00

Basically, PCTN is used to compute the percentages for frequency counts of the grand total and PCTSUM is used to compute the percentage for the grand total of the analysis variable. The formula is:

$$\text{Percent} = 100 * \text{value in given cell} / \text{sum of all cell values in the table}$$

The percentage number in Figures 1 and 2 are calculated using this formula.

When using PCTN and PCTSUM, we can also define the denominator in the table statement. The general expression is:

$$\text{PCTN} < \text{denominator} > \text{ or } \text{PCTSUM} < \text{denominator} >$$

The denominator within <> represents the class variables in the dimension across which you are summing. In this case, the output from PCTN<denominator> is the same as COLPCTN or ROWPCTN, and PCTSUM<denominator> is the same as COLPCTSUM or ROWPCTSUM, depending on the dimension of the denominator within <>. Below is SAS code and output to illustrate this.

```
proc tabulate data=test.adlb(where=(paramcd='ALT' and avisit='Week 4'));
  class sex site;
  table sex*
    (n='Number of Patients'*f=4.
    pctn<site>= 'Percentage of Row Totals (% of site)'
    pctn<sex>='Percentage of Column Totals (% of sex)'
    pctn='% of All Patients'),
    site/rts=60 box='Example of PCTN with specifying denominator';
run;
```

Figure 3. Example of PCTN with specifying denominator

Example of PCTN with specifying denominator		SITE	
		Ex-USA	USA
Gender			
F	Number of Patients	1	2
	Percentage of Row Totals (% of site)	33.33	66.67
	Percentage of Column Totals (% of sex)	50.00	50.00
	% of All Patients	16.67	33.33
M	Number of Patients	1	2
	Percentage of Row Totals (% of site)	33.33	66.67
	Percentage of Column Totals (% of sex)	50.00	50.00
	% of All Patients	16.67	33.33

In the above output, pctn<site> calculates percentage by summing the number of patients in all sites within the same sex as the denominator. For example, for female patients at Ex-USA site, the % of site is $1/(1+2)*100=33.33$.

Similarly, pctn<sex> defines the number of patients in all genders within the same site as the denominator. For example, for Ex-USA female patients, the % of sex is $1/(1+1)*100=50.00$. PCTN does not define the denominator, so the grand total of the number of patients in the table is the denominator. For example, % of all female patients at Ex-US site = $1/(1+2+1+2)*100=16.67$.

In another example, the following code calculates the total number of lab test values in each week by each test and the resulting three different percentages of PCTSUM.

```
proc tabulate data=test.adlb;
  class paramcd avisit;
  var aval;
  table paramcd=' '*
    (sum='Total Test Value'*f=5.
     n='Number of Patients'*f=4.
     pctsum<avisit>= 'Percentage of Row Totals (% by week)'
     pctsum<paramcd>='Percentage of Column Totals (% by test code)'
     pctsum='% of All Patients'),
     avisit*aval=''/rts=50 box='Example of PCTSUM with specifying denominator';
run;
```

Figure 4. Example of PCTSUM with specifying denominator

Example of PCTSUM with specifying denominator		Analysis timepoint description	
		Week 0	Week 4
ALT	Total Test Value	98	96
	Number of Patients	6	6
	Percentage of Row Totals (% by week)	50.52	49.48
	Percentage of Column Totals (% by test code)	53.26	54.24
	% of All Patients	27.15	26.59
AST	Total Test Value	86	81
	Number of Patients	6	6
	Percentage of Row Totals (% by week)	51.50	48.50
	Percentage of Column Totals (% by test code)	46.74	45.76
	% of All Patients	23.82	22.44

The above output shows the total value for the ALT test (98) at week 0. Pctsum<week> defines the total value of all weeks for a specific test as the denominator. For example % by week for ALT in week 0 is calculated as $98/(98+96)*100=50.52$. Pctsum<paramcd> defines the total value for all the tests in a specific week as the denominator. For example, % by test code for ALT in week 0 is calculated as $98/(98+86)*100=53.26$. Pctsum does not define the denominator. The default denominator is

the total value of all the tests in all weeks. For example, % of all patients for ALT in week 0 is calculated as $98/(98+96+86+81)*100=27.15$.

ROWPCTN AND ROWPCTSUM COLPCTN AND COLPCTSUM

ROWPCTN, ROWPCTSUM, COLPCTN and COLPCTSUM statistics for the TABULATE procedure simplify the code for calculating percentage. The general formula for computing percentage based on row or column totals is:

$$\text{Row (Column) percent} = 100 * \text{value in the given cell} / \text{Sum over all rows (columns) for that row (column)}$$

In this paper, we only use an example to show how to use ROWPCTN and ROWPCTSUM. The usage of COLPCTN and COLPCTSUM are similar. Below is the code that illustrates use of ROWPCTN and ROWPCTSUM

```
proc tabulate data=test.adlb (where=(paramcd='ALT' and avisit='Week 4'));
class paramcd trt sex;
var aval;
table trt all, (sex='Gender (Count)' all)*n*f=4. (sex='Gender (%)' all)
*ROWPCTN/box='Example of using ROWPCTN';

table trt all, (sex='Gender (Count)' all)*aval='Lab Result'*sum*f=6.
(sex='Gender (%)' all)*aval='Lab Result'*ROWPCTSUM
/box='Example of using ROWPCTSUM';
run;
```

Below is the resulting output.

Figure 5. Example of using ROWPCTN

Example of using ROWPCTN	Gender (Count)		All	Gender (%)		All
	F	M		F	M	
	N	N	N	RowPctN	RowPctN	RowPctN
Treatment Arm						
Placebo	2	2	4	50.00	50.00	100.00
Study drug	1	1	2	50.00	50.00	100.00
All	3	3	6	50.00	50.00	100.00

Figure 6. Example of using ROWPCTSUM.

Example of using ROWPCTSUM	Gender (Count)		All	Gender (%)		All
	F	M		F	M	
	Lab Result	Lab Result	Lab Result	Lab Result	Lab Result	Lab Result
	Sum	Sum	Sum	RowPctSum	RowPctSum	RowPctSum
Treatment Arm						
Placebo	33	27	60	55.00	45.00	100.00
Study drug	11	25	36	30.56	69.44	100.00
All	44	52	96	45.83	54.17	100.00

REPPCTN AND REPPCTSUM

The syntax and the output is the same as when you use pctn and pctsum.

CONCATENATED TABLES

Using the lab test data presented at the beginning of this paper, assume we want to create a table like Figure 7 with both Site and Treatment Arm in the same dimension. For a concatenated table, defining the denominator becomes more complicated. In the table statement, we must include an expression in the denominator definition for each expression crossed with PCTN or PCTSUM statistics.

Figure 7. Concatenated Tables

	Gender			
	F		M	
	Count	Percent of This Sex	Count	Percent of This Sex
Site				
Ex-USA	1	33.33	1	33.33
USA	2	66.67	2	66.67
Treatment Arm				
Placebo	2	66.67	2	66.67
Study drug	1	33.33	1	33.33

As a start, first we need to define the following TABLE statement,

```
Table site trt, sex*n;
```

Now we include the percentage that each cell is of the total for that particular value of SEX.

Table site trt, sex*(n, pctn);

This implies that crossing exists between dimension expressions:

site*sex*pctn and trt*sex*pctn

Therefore, we must include an expression in the denominator definition for each of the crossings. Because we want the denominator to be the total for each value of sex (total number of female in the first column, total number of male in second column), we do not want PROC TABULATE to sum across the value of the SEX variable when computing the denominator. We have to remove the sex variable in the above expressions. Now we have SITE and TRT left in the expressions. Below is the SAS code to illustrate this concept:

```
proc tabulate data=test.adlb (where=(paramcd='ALT' and avisit='Week 4')) format=8.2;
  class site trt sex;
  table site='Site' trt,
        sex*(n='Count'*f=8. pctn<site trt>='Percent of This Sex');
run;
```

At this point, if we want to include ALL in the table above to display column total which will add a new crossing ALL*SEX in table statement, we must add ALL in the denominator definition. Below is the code (Please note if we do not add ALL in the denominator definition, we will get following error message:

ERROR: PctN base is not in table. A PctN crossing has no denominator):

```
proc tabulate data=test.adlb (where=(paramcd='ALT' and avisit='Week 4')) ;
  class site trt sex;
  table site='Site' trt all='Total for Each Sex',
        sex*(n='Count'*f=8. pctn<site trt all>='Percent of This Sex');
run;
```

The resulting output table:

Figure 8. Concatenated Tables with Total

	Gender			
	F		M	
	Count	Percent of This Sex	Count	Percent of This Sex
Site				
Ex-USA	1	33.33	1	33.33
USA	2	66.67	2	66.67
Treatment Arm				
Placebo	2	66.67	2	66.67
Study drug	1	33.33	1	33.33
Total for Each Sex	3	100.00	3	100.00

COMPUTING PERCENTAGES OF SUBTOTALS

So far the percentage we calculated is of totals in the table (for example, column total, row total, page total, etc.). PROC TABULATE can also display percentage of subtotal which means you can define a denominator as any subtotal in your table following this rule: determine which class variable must be summed to form the subtotal you want. As an example, assume we want to see the number of female or male subjects at each visit under different treatment arms at either USA or Ex-USA site, we have following SAS code:

```
options missing=0;
proc tabulate data=test.adlb (where=(paramcd='ALT')) format=8.2;
  class site trt sex avisit;
  table site='Site' *(trt all), avisit*(sex all)*n='Count'*f=8.;
run;
```

The resulting output is:

Figure 9. Count of Subtotals

		Analysis timepoint description					
		Week 0			Week 4		
		Gender		All	Gender		All
		F	M		F	M	
		Count	Count	Count	Count	Count	Count
Site	Treatment Arm						
Ex-USA	Placebo	1	0	1	1	0	1
	Study drug	0	1	1	0	1	1
	All	1	1	<u>2</u>	1	1	<u>2</u>
USA	Placebo	1	2	3	1	2	3
	Study drug	1	0	1	1	0	1
	All	2	2	<u>4</u>	2	2	<u>4</u>

Now we want a percentage that indicates what portion of the subjects under each visit at a site (USA or Ex-USA) is represented by female or male subjects in each treatment arm at the site. This percentage is the number of subjects in each cell of the table divided by the total number of subjects at each site during each visit. These denominator values are the number underlined in Figure 9.

These cells represented subtotals that were obtained by summing the frequency counts over two class variables TRT and SEX. Please note this is done for each SITE and AVISIT. Because the values are not summed over SITE and AVISIT, we shall only include TRT and SEX in the denominator definition.

Since the example also involves a concatenated table – ALL is concatenated with both TRT and SEX in the table statement, we must include these concatenations in the denominator definition as well (The crossing ALL*ALL can be abbreviated as ALL). Here is the SAS code and the resulting output to illustrate this:

```
options missing=0;
proc tabulate data=test.adlb (where=(paramcd='ALT')) format=8.2;
  class site trt sex avisit;
  table site='Site' *(trt all),
        avisit*(sex all)*pctn<trt*sex trt*all all*sex all>='';
run;
```

Figure 10. Percentage of Subtotals

		Analysis timepoint description					
		Week 0			Week 4		
		Gender		All	Gender		All
		F	M		F	M	
Site	Treatment Arm						
	Ex-USA	50.00	0	50.00	50.00	0	50.00
	Study drug	0	50.00	50.00	0	50.00	50.00
	All	50.00	50.00	100.00	50.00	50.00	100.00
USA	Treatment Arm						
	Placebo	25.00	50.00	75.00	25.00	50.00	75.00
	Study drug	25.00	0	25.00	25.00	0	25.00
	All	50.00	50.00	100.00	50.00	50.00	100.00

CONCLUSION

1. The TABULATE procedure can calculate the percentages of the total count or sum of analysis variables for individual cells.
2. Use PCTN and PCTSUM to calculate percentage of the grand total; The output is the same as when you use REPPCTN and REPPCTSUM;
3. Use ROWPCTN and ROWPCTSUM to calculate the percentage of row total;
4. Use COLPCTN and COLPCTSUM to calculate the percentage of column total;
5. When calculating complex percentage for concatenated tables, in the table statement include an expression in the denominator definition for each expression crossed with PCTN or PCTSUM statistics.
6. When calculating percentage of subtotal, determine which class variable must be summed to form the subtotal we want.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Wende (Ted) Tian
Enterprise: Merck Sharpe & Dohme Corp
Address: 126 E Lincoln Ave
City, State ZIP: Rahway, NJ 07065
Work Phone: 732 594-2048
Fax: 732 594-6075
E-mail: Wende_tian@merck.com
Web:

Name: Hong (Lily) Zhang
Enterprise: Merck Sharpe & Dohme Corp
Address: 126 E Lincoln Ave
City, State ZIP: Rahway, NJ 07065
Work Phone: +1 732 594-5413
Fax: 732 594-6075
E-mail: lily_zhang2@merck.com
Web:

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.