# How can I create crosstabs using PROC FREQ in SAS?

Authored by

**stats writer**

June 24, 2024

PROC FREQ is a SAS procedure that allows users to easily create crosstabs, also known as contingency tables, to analyze the relationship between two categorical variables. This procedure provides a comprehensive summary of the frequency distributions and percentages of the data in a table format, making it easier to identify any patterns or associations between the variables. By inputting the desired variables and options, users can generate customizable crosstabs, making it a powerful tool for data analysis in research and statistical studies. Overall, PROC FREQ in SAS provides a user-friendly and efficient method for creating crosstabs and understanding the relationship between categorical variables.

## Introduction

To describe a single categorical variable, we use frequency tables. To describe the relationship between two categorical variables, we use a special type of table called a *cross-tabulation* (or "crosstab" for short). Consider the following sets of tables, both of which summarize the categorical variables "Gender" and "Athlete":

Frequency tables of variables Gender and Athlete

The FREQ Procedure

| Gender | | | | |
|---|---|---|---|---|
| Gender | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| Male | 204 | 47.89 | 204 | 47.89 |
| Female | 222 | 52.11 | 426 | 100.00 |
| Frequency Missing = 9 | | | | |

| Are you an athlete? | | | | |
|---|---|---|---|---|
| Athlete | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| Non-athlete | 251 | 57.70 | 251 | 57.70 |
| Athlete | 184 | 42.30 | 435 | 100.00 |

Crosstab of Gender and Athlete

The FREQ Procedure

| Frequency<br>Percent | Table of Gender by Athlete | | |
|---|---|---|---|
| | Athlete(Are you an athlete?) | | |
| Gender(Gender) | Non-athlete | Athlete | Total |
| Male | 103<br>24.18 | 101<br>23.71 | 204<br>47.89 |
| Female | 143<br>33.57 | 79<br>18.54 | 222<br>52.11 |
| Total | 246<br>57.75 | 180<br>42.25 | 426<br>100.00 |
| Frequency Missing = 9 | | | |

In a cross-tabulation, the categories of one variable determine the rows of the table, and the categories of the other variable determine the columns. The cells of the table contain the number of times that a particular combination of categories occurred. The "edges" (or "margins") of the table typically contain the total number of observations for that category.

This type of table is also known as a:

Crosstab.Two-way table.Contingency table.

In this tutorial, we'll cover how to create crosstabs using the SAS procedure PROC FREQ, and how to interpret the frequencies and proportions in these tables.

## Describing a Crosstab

The dimensions of the crosstab refer to the number of rows and columns in the table (not including the row/column totals). The table dimensions are reported as as *R*x*C*, where *R* is the number of categories for the row variable, and *C* is the number of categories for the column variable.

Additionally, a "square" crosstab is one in which the row and column variables have the same number of categories. Tables of dimensions 2x2, 3x3, 4x4, etc. are all square crosstabs.

### Example 1: "Square" table

**Gender * Do you drink alcohol? Crosstabulation**

Count

| | | Do you drink alcohol? | | |
|---|---|---|---|---|
| | | No | Yes | Total |
| Gender | Male | 13 | 33 | 46 |
| | Female | 13 | 32 | 45 |
| Total | | 26 | 65 | 91 |

**Row variable**: *Gender* (2 categories: male, female)**Column variable**: *Alcohol* (2 categories: no, yes)**Table dimension**: 2x2 (square)

## Example 2: "Long" table

**Class Rank * Gender Crosstabulation**

Count

| | | Gender | | |
|---|---|---|---|---|
| | | Male | Female | Total |
| Class Rank | Freshman | 14 | 9 | 23 |
| | Sophomore | 15 | 13 | 28 |
| | Junior | 9 | 11 | 20 |
| | Senior | 12 | 14 | 26 |
| Total | | 50 | 47 | 97 |

**Row variable**: *Class Rank* (4 categories: freshman, sophomore, junior, senior)**Column variable**: *Gender* (2 categories: male, female)**Table dimension**: 4x2

## Example 3: "Wide" table

**Gender * Do you smoke cigarettes? Crosstabulation**

Count

| | | Do you smoke cigarettes? | | | |
|---|---|---|---|---|---|
| | | Never smoked | Past smoker | Current smoker | Total |
| Gender | Male | 32 | 1 | 14 | 47 |
| | Female | 25 | 3 | 16 | 44 |
| Total | | 57 | 4 | 30 | 91 |

**Row variable**: *Gender* (2 categories: male, female)**Column variable**: *Smoking* (3 categories: never smoked, past smoker, current smoker)**Table dimension**: 2x3

## Understanding Row, Column, and Total Percents

A typical 2x2 crosstab has the following construction:

| | Column 1 | Column 2 | Row totals |
|---|---|---|---|
| **Row 1** | *a* | *b* | *a* + *b* |
| **Row 2** | *c* | *d* | *c* + *d* |
| **Column totals** | *a* + *c* | *b* + *d* | *a* + *b* + *c* + *d* |

The letters *a*, *b*, *c*, and *d* represent what are called *cell counts*.

*a* is the number of observations corresponding to Row 1 AND Column 1.*b* is the number of observations corresponding to Row 1 AND Column 2.*c* is the number of observations corresponding to Row 2 AND Column 1.*d* is the number of observations corresponding to Row 2 AND Column 2.

By adding *a*, *b*, *c*, and *d*, we can determine the total number of observations in each category, and in the table overall.

*Row sum* of row 1 (i.e., total number of observations in Row 1): *a* + *bRow sum* of row 2 (i.e., total number of observations in Row 2): *c* + *dColumn sum* of column 1 (i.e., total number of observations in Column 1): *a* + *cColumn sum* of column 2 (i.e., total number of observations in Column 2): *b* + *d*Total sum (i.e., total number of observations in the table): *n* = *a* + *b* + *c* + *d*

The row sums and column sums are sometimes referred to as *marginal frequencies*. Note that if you were to make frequency tables for your row variable and your column variable, the frequency table should match the values for the row totals and column totals, respectively.

When you are describing the composition of your sample, it is often useful to refer to the proportion of the row or column that fell within a particular category. This can be achieved by computing the *row percentages* or *column percentages.*

| | Column 1 | Column 2 | Row totals |
|---|---|---|---|
| **Row 1**<br>**Row 1 %** | *a*<br>*a / (a + b)* | *b*<br>*b / (a + b)* | *a + b*<br>*(a + b) / (a+b) = 100%* |
| **Row 2**<br>**Row 2 %** | *c*<br>*c / (c + d)* | *d*<br>*d / (c + d)* | *c + d*<br>*(c + d) / (c + d) = 100%* |
| **Column totals**<br>**% of total** | *a + c*<br>*(a + c) / (a + b + c + d)* | *b + d*<br>*(b + d) / (a + b + c + d)* | *a + b + c + d*<br>*(a + b + c + d) / (a + b + c + d) = 100%* |

Notice that when computing row percentages, the denominators for cells *a*, *b*, *c*, *d* are determined by the row sums (here, *a* + *b* and *c* + *d*). This implies that the percentages in the "row totals" column must equal 100%.

| | Column 1 | Column 2 | Row totals |
|---|---|---|---|
| **Row 1<br>Column 1 %** | *a*<br>*a / (a + c)* | *b*<br>*b / (b + d)* | *a + b*<br>*(a + b) / (a + b + c + d)* |
| **Row 2<br>Column 2 %** | *c*<br>*c / (a + c)* | *d*<br>*d / (b + d)* | *c + d*<br>*(c + d) / (a + b + c + d)* |
| **Column totals<br>Percentage %** | *a + c*<br>*(a + c) / (a + c) = 100%* | *b + d*<br>*(b + d) / (b + d) = 100%* | *a + b + c + d*<br>*(a + b + c + d) / (a + b + c + d) = 100%* |

Notice that when computing column percentages, the denominators for cells *a*, *b*, *c*, *d* are determined by the column sums (here, *a* + *c* and *b* + *d*). This implies that the percentages in the "column totals" row must equal 100%.

| | Column 1 | Column 2 | Row totals |
|---|---|---|---|
| **Row 1<br>% of total** | *a*<br>*a / (a + b + c + d)* | *b*<br>*b / (a + b + c + d)* | *a + b*<br>*(a + b) / (a + b + c + d)* |
| **Row 2<br>% of total** | *c*<br>*c / (a + b + c + d)* | *d*<br>*d / (a + b + c + d)* | *c + d*<br>*(c + d) / (a + b + c + d)* |
| **Column totals<br>% of total** | *a + c*<br>*(a + c) / (a + b + c + d)* | *b + d*<br>*(b + d) / (a + b + c + d)* | *a + b + c + d*<br>*(a + b + c + d) / (a + b + c + d) = 100%* |

Notice that when total percentages are computed, the denominators for all of the computations are equal to the total number of observations in the table, i.e. *a* + *b* + *c* + *d*.

## Data Set-Up and Requirements

### Data Requirements

Your data must meet the following requirements:

At least two categorical variables.Each categorical variable should have two or more categories (groups).

Note that the choice of row/column variable is often dictated by space requirements or interpretation of the results. If your particular set of variables has what could be considered "independent" and "dependent" variables, it is conventional to put the "independent" variable as

the column variable, and the "dependent" variable as the row variable. However, if you plan to compute relative risk, it is conventional to put the "independent" variable as the row and the "dependent" variable as the column variable.

## Data Set-Up

Your dataset should have the following structure:

Each case (row) represents a subject, and each subject appears once in the dataset. That is, each row represents an observation from a unique subject.The dataset contains at least two nominal categorical variables (string or numeric). The categorical variables used in the test must have two or more categories; they should also not have too many categories.

| | ID Number | Class rank | Gender | Are you an athlete? |
|---|---|---|---|---|
| 1 | 20183 | . | Male | Non-athlete |
| 2 | 20230 | Freshman | Male | Athlete |
| 3 | 20243 | Junior | Female | Non-athlete |
| 4 | 20248 | Freshman | . | Non-athlete |
| 5 | 20255 | Sophomore | Female | Non-athlete |
| 6 | 20278 | . | Male | Non-athlete |
| 7 | 20389 | . | Male | Non-athlete |
| 8 | 20402 | Sophomore | Male | Non-athlete |
| 9 | 20531 | Freshman | Male | Athlete |
| 10 | 20615 | Freshman | Female | Non-athlete |
| 11 | 20626 | Sophomore | Female | Non-athlete |
| . . . | | | | |
| 425 | 49386 | Sophomore | Female | Non-athlete |
| 426 | 49445 | Sophomore | Male | Athlete |
| 427 | 49572 | Senior | Female | Non-athlete |
| 428 | 49688 | . | Male | Non-athlete |
| 429 | 49806 | Junior | Female | Non-athlete |
| 430 | 49821 | Junior | Female | Athlete |
| 431 | 49838 | Freshman | Male | Athlete |
| 432 | 49854 | Junior | Male | Non-athlete |
| 433 | 49879 | Sophomore | Male | Athlete |
| 434 | 49931 | Junior | Male | Athlete |
| 435 | 49947 | Freshman | Female | Athlete |



## Creating Cross-Tabulations using PROC FREQ

For crosstabs, the basic syntax of the FREQ procedure is:

```
PROC FREQ DATA=dataset <options>;
TABLES RowVar*ColVar / <options>;
RUN;
```

In the first line, `PROC FREQ` tells SAS to execute the FREQ procedure on the dataset given in the `DATA=` argument. If desired, additional options you can include on this line are:

`NLEVELS`
Adds a table to the output summarizing the number of levels (categories) for each variable named in the `TABLES` statement.

**Number of Variable Levels**

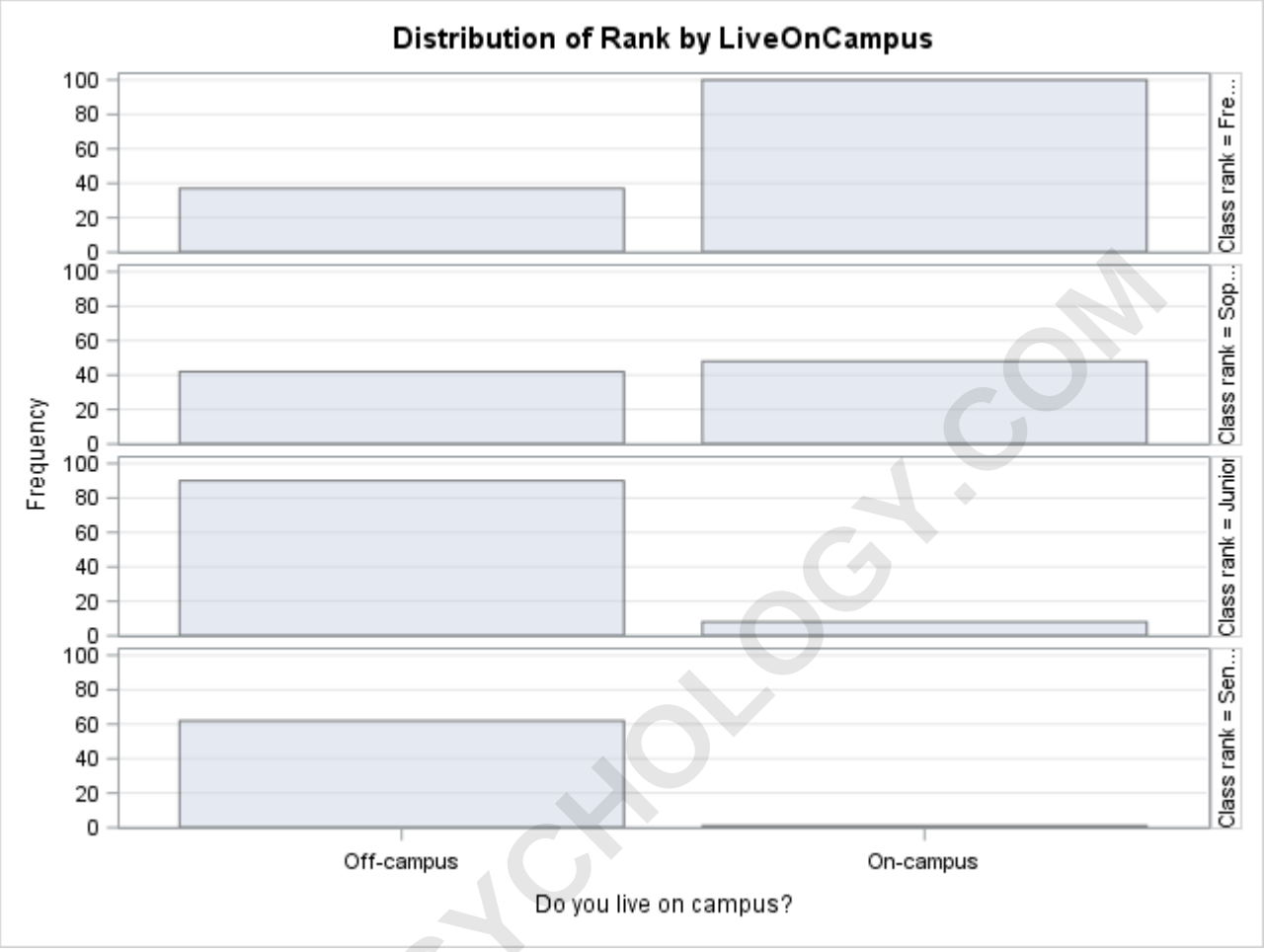
| Variable | Label | Levels | Missing Levels | Nonmissing Levels |
|---|---|---|---|---|
| Rank | Class rank | 5 | 1 | 4 |
| LiveOnCampus | Do you live on campus? | 3 | 1 | 2 |

`ORDER=data`
Sorts the rows and columns of the crosstab in the same order as they appear in the dataset.`ORDER=freq`
Sorts the rows and columns of the crosstab from most frequent to least frequent.

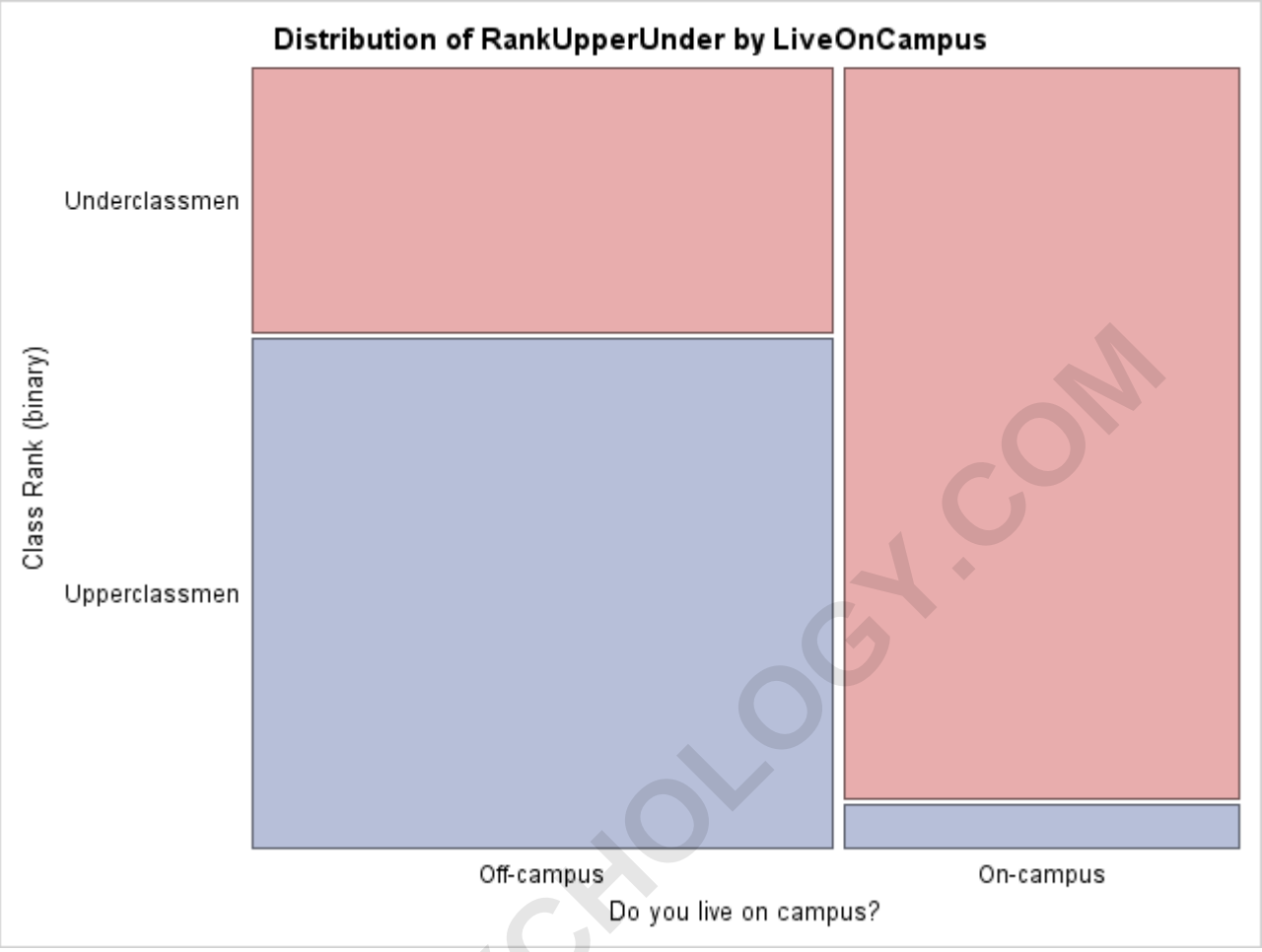
On the next line, the `TABLES` statement is where you put pairs of variables you want to produce crosstabs for. To create a basic cross-tab between two variables A and B, place an asterisk (*) between the names of the variables in the TABLES statement. You can list as many variables or variable pairs as you want, with each variable or variable pair separated by a space. This is the minimum that is required to produce a crosstab using PROC FREQ, but there are several important analysis options to be aware of, which you can add on this line after a slash (/) character:

`PLOTS=FREQPLOT`
Adds faceted barplots to the output for each crosstab (example shown below).`PLOTS=MOSAICPLOT`
Adds a mosaic plot to the output for each crosstab (example shown below).`MISSING`
Include missing values as a row in the frequency frequency tables. The missing category will be treated as if it were an observed category, so those cases will be included in the computation of the percents, cumulative frequencies, and cumulative proportions.`MISSPRINT`
Include missing values as a row in the frequency tables, but do not count those cases towards computing the percentages, cumulative frequencies, or cumulative proportions.`NOROW`, `NOCOL`, and `NOPERCENT`
Suppress the display of row proportions, column proportions, or overall proportions, respectively.

`PLOTS=FREQPLOT`




`PLOTS=MOSAICPLOT`






## Examples

### Problem Statement

Some universities in the United States require that freshmen live in the on-campus dormitories during their first year, with exceptions for students whose families live within a certain radius of campus. That is, certain freshmen whose families live close enough to campus are permitted to live off-campus. After completing their first or second year of school, students living in the dorms may choose to move into an off-campus apartment. How prevalent is this pattern?

In the sample dataset, there are several variables relating to this question:

Rank - Class rank (Freshmen, Sophomore, Junior, Senior)LiveOnCampus - Do you live on campus? (Off-campus, On-campus)

Let's use different aspects of PROC FREQ to investigate the relationship between class rank and living on campus.

## Part 1 - Simple Crosstab

Using the sample data, let's make crosstab of the variables Rank and LiveOnCampus. Let the row variable be Rank, and the column variable be LiveOnCampus.

### Syntax

```
PROC FREQ DATA=work.sample;
TABLE Rank*LiveOnCampus;
RUN;
```

In this syntax:

`Rank*LiveOnCampus` will create a crosstab of variable Rank (as the row variable) against LiveOnCampus (as the column variable). The table will include all of the default output for crosstabs.

### Output

The first table contains the actual crosstab:

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Table of Rank by LiveOnCampus | | |
|---|---|---|---|
| | LiveOnCampus(Do you live on campus?) | | |
| Rank(Class rank) | Off-campus | On-campus | Total |
| Freshman | 37<br>9.54<br>27.01<br>16.02 | 100<br>25.77<br>72.99<br>63.69 | 137<br>35.31 |
| Sophomore | 42<br>10.82<br>46.67<br>18.18 | 48<br>12.37<br>53.33<br>30.57 | 90<br>23.20 |
| Junior | 90<br>23.20<br>91.84<br>38.96 | 8<br>2.06<br>8.16<br>5.10 | 98<br>25.26 |
| Senior | 62<br>15.98<br>98.41<br>26.84 | 1<br>0.26<br>1.59<br>0.64 | 63<br>16.24 |
| Total | 231<br>59.54 | 157<br>40.46 | 388<br>100.00 |
| Frequency Missing = 47 | | | |



Notice the square to the left of the table: it contains the legend for how to read the cells of this crosstab. The legend here tells us that, in this example, each cell of the table has 4 numbers:

The first number is the frequency, i.e. the number of cases having that particular combination of the row and column variable.The second number is the "percent", i.e. the cell's proportion of the total; the denominator is the total number of nonmissing values, 388).The third number is the row percentage, i.e., the cell's proportion of that row; the denominator is the value in the 'Total' cell at the end of the row).The fourth number is the column percentage, i.e., the cell's proportion of that column; the denominator is the value in the 'Total' cell at the end of the column.

From this table, we can make several observations:

Many more freshmen lived on-campus (100) than off-campus (37)About an equal number of sophomores lived off-campus (42) versus on-campus (48)Far more juniors lived off-campus (90)

than on-campus (8)Only one (1) senior lived on campus; the rest lived off-campus (62)

Note the margins of the crosstab (i.e., the "total" row and column) give us the same information that we would get from frequency tables of Rank and LiveOnCampus, respectively:

The sample had 137 freshmen, 90 sophomores, 98 juniors, and 63 seniorsThere were 231 individuals who lived off-campus, and 157 individuals lived on-campus

Lastly, the outermost row of the table shows the total number of cases with missing values for either Rank, LiveOnCampus, or both (47).

## Part 2 - Row, column, and total percentages

Let's delve into the proportions from the previous table. Although the default table already contains all three types of proportions, it's a little overwhelming to see all the information at once, so let's use the `NOROW`, `NOCOL`, and `NOPERCENT` options to limit what type of percentages we see.

### Row proportions

If the row variable is Rank and the column variable is LiveOnCampus, then the row percentages will tell us what percentage of the freshmen, sophomores, juniors, and seniors live on campus. That is, variable Rank will determine the denominator of the percentage computations.

#### Syntax

```
/*Show row proportions only - suppress column and total proportions*/
PROC FREQ DATA=work.sample;
TABLE Rank*LiveOnCampus / NOCOL NOPERCENT;
RUN;
```

#### Output

| Frequency<br>Row Pct | Table of Rank by LiveOnCampus | | |
|---|---|---|---|
| | LiveOnCampus(Do you live on campus?) | | |
| Rank(Class rank) | Off-campus | On-campus | Total |
| Freshman | 37<br>27.01 | 100<br>72.99 | 137 |
| Sophomore | 42<br>46.67 | 48<br>53.33 | 90 |
| Junior | 90<br>91.84 | 8<br>8.16 | 98 |
| Senior | 62<br>98.41 | 1<br>1.59 | 63 |
| Total | 231 | 157 | 388 |
| Frequency Missing = 47 | | | |



#### Interpretation

The proportion of freshmen who live on campus is 72.99% (100/137).The proportion of sophomores who live on campus is 53.33% (48/90).The proportion of juniors who live on campus is 8.16% (8/98).The proportion of seniors who live on campus is 1.59% (1/63).

### Column proportions

If the row variable is Rank and the column variable is LiveOnCampus, then the column percentages will tell us what percentage of the individuals who live on campus are freshmen, sophomores, juniors, or seniors. That is, variable LiveOnCampus will determine the denominator of the percentage computations.

#### Syntax

```
/*Show column proportions only - suppress row and total proportions*/
PROC FREQ DATA=work.sample;
TABLE Rank*LiveOnCampus / NOROW NOPERCENT;
RUN;
```

#### Output

| Frequency<br>Col Pct | Table of Rank by LiveOnCampus | | |
|---|---|---|---|
| | LiveOnCampus(Do you live on campus?) | | |
| Rank(Class rank) | Off-campus | On-campus | Total |
| Freshman | 37<br>16.02 | 100<br>63.69 | 137 |
| Sophomore | 42<br>18.18 | 48<br>30.57 | 90 |
| Junior | 90<br>38.96 | 8<br>5.10 | 98 |
| Senior | 62<br>26.84 | 1<br>0.64 | 63 |
| Total | 231 | 157 | 388 |
| Frequency Missing = 47 | | | |



#### Interpretation

63.69% of the people living on campus are freshmen (100/157).30.57% of the people living on campus are sophomores (48/157).5.10% of the people living on campus are juniors (8/157).0.64% of the people living on campus are seniors (1/157).

### Overall proportions

If the row variable is Rank and the column variable is LiveOnCampus, then the total percentage tells us what proportion of the total is within each combination of Rank and LiveOnCampus. That is, the overall table size determines the denominator of the percentage computations.

#### Syntax

```
/*Show overall proportions only - suppress row and column proportions*/
PROC FREQ DATA=work.sample;
TABLE Rank*LiveOnCampus / NOROW NOCOL;
RUN;
```

#### Output

| Frequency<br>Percent | Table of Rank by LiveOnCampus | | |
|---|---|---|---|
| | LiveOnCampus(Do you live on campus?) | | |
| Rank(Class rank) | Off-campus | On-campus | Total |
| Freshman | 37<br>9.54 | 100<br>25.77 | 137<br>35.31 |
| Sophomore | 42<br>10.82 | 48<br>12.37 | 90<br>23.20 |
| Junior | 90<br>23.20 | 8<br>2.06 | 98<br>25.26 |
| Senior | 62<br>15.98 | 1<br>0.26 | 63<br>16.24 |
| Total | 231<br>59.54 | 157<br>40.46 | 388<br>100.00 |
| Frequency Missing = 47 | | | |



#### Interpretation

Freshmen living off-campus make up 9.54% of the sample (37/388).Freshmen living on-campus make up 25.77% of the sample (100/388).Sophomores living on-campus make up 10.82% of the sample (42/388).Sophomores living off-campus make up 12.37% of the sample (48/388).

## For More Information

A full list of options for the FREQ procedure can be found in the SAS Help and Documentation guide.

Base SAS 9.4 Procedures Guide: Statistical Procedures, Sixth Edition: The FREQ Procedure

## Tutorial Feedback