

## IT FREQUENTLY MAKES SENSE TO USE WEIGHTS

Christopher Toppe, Ph.D.

Atlantic Research Corporation, Rockville, MD 20850

Assume for a moment that you have taken a sample of 10 people from a population of 100, and that you have given these 10 people a questionnaire to complete. You used these 10 questionnaires to build a SAS® data base. You now want to analyze that data base using a SAS procedure such as PROC MEANS. You want the results from the procedure give you estimates of the population you sampled -- the 100 people -- not just the 10 in the sample data base. That is, since you have a sample of 10 people out of a population of 100, you want your results to show you (estimates of) the means and totals for the population, not just for the 10 in the sample. Does it make a difference if you run PROC MEANS using a FREQ or WEIGHT Statement?

Absolutely. To use a FREQ statement would be wrong. Only a WEIGHT statement would give you the results you wanted. The FREQ and WEIGHT statements are different in theory, application, and effect. They are not interchangeable. Each has a specific purpose. Each operates on SAS data sets in different ways.

Part of the problem with the FREQ and WEIGHT statements is that we tend to call them the same thing: a weighting variable, and we even tend to give them the same variable name, like WGT. In some sense, this is correct in that we intend for the WGT variable to somehow change our data base, but to confuse them can lead to mistakes.

### A Basic Distinction

FREQ works on observations, WEIGHT works on variables. When you use a FREQ statement, you are telling the SAS system to duplicate an observation *freq* number of times. That is, you are telling SAS that the data base you are using should really be made up of *freq* number of identical observations. The values of the variables in those *freq* number of identical observations are all identical. In the above example, if you used a FREQ statement you would be saying that the 100 people were really made up of 10 sets of 10 identical observations; that your analysis data base is really 100 observations large.

A WEIGHT statement, on the other hand, works on variables. It changes the contribution each variable makes to the total across observations. This change, then, impacts on the statistics the SAS procedure computes. Again going back to the example, using a WEIGHT statement says that your 10 observations represent 100 people, but your analysis data base is still only 10 observations large.

The implications of this basic distinction will be discussed later in this paper.

### Decimals in the FREQ Statement

Given that the FREQ statement really means to duplicate an observation *freq* number of times, it makes sense that the FREQ statement ignores decimals in SAS procedures. After all, you can't have a partial observation. You might have 10 observations, you might have 11. But you can't have 10.5 observations.

Since FREQ is only intended to change the number of observations undergoing analysis, only the integer portion of the FREQ variable's value is used. The decimal portion is ignored. It's as if you wrote a SAS statement that said:

WGT = INT(WGT);

Further, since the FREQ statement tells SAS to duplicate the observation *freq* number of times, FREQ ignores observations with *freq* values less than 1, including missing values. That is, you can't say 'this data base has 0 or -5 observations that look like this'. (See Table 1)

### Decimals in the Weight Statement

Since the weight statement operates on variable values, which are numbers, it makes sense that values can change by a non-integer amount. For example, you can increase a value by a factor of 10.5 simply by multiplying it by 10.5.

When the value of the *weight* variable is zero or less, the weight is treated as a zero. This means that the variables in the observation will contribute nothing to the statistics, but that the observation will still be counted as existing. Therefore, you should always make sure that you don't have any missing values for your *weight* variable.

### The Purpose of Sampling

In order for the differences between FREQ and WEIGHT to make sense, it is necessary to have a brief discussion of sampling theory because you will only use FREQ and WEIGHTS when you have a partial data base. That is, if your data base contains an observation for every member of the group under study, you should never need to use a FREQ or WEIGHT statement. Your data base already gives you information on the group you are studying. But when your data base is from a subset of the group you are studying -- when the data base is a sample from the population -- you will need to FREQ or WEIGHT the data base to get the results to reflect the total group.

There are four main reasons people collect sample data. It may be impossible to collect data on the entire population. If it's not impossible, collecting data on the population may be too expensive. Even if it's not too expensive, it may take too long. Finally, the thing you are testing may be potentially dangerous (a new drug) or destructive (you subject a water pipe to pressure until it breaks). In these cases you work from a sample.

A sample is drawn from the population in a random way, meaning that each member of the population has an equal chance of being selected into the sample. You then assume that the sample data base "looks like" the

population, that it is simply a smaller group with the same characteristics as the larger group. You then examine the characteristics of this smaller group and conclude that you would find the same traits in the larger group. You use the results from the sample to draw conclusions about the population.

This is the first place where it is appropriate to weight the data base. Since we are working with a sample of 10 from a population of 100, we create a sample weighting variable that tells us the frequency of our sampling design:

$$SWGT = \text{Population } N / \text{Sample } n;$$

In our case, SWGT (sample weight) has the value of 10.0: each observation in the data base represents 10 people in the population.

However, these observations are *not* identical. Not only do we expect that the people are different in socio-demographic characteristics, we also expect that their responses to our questionnaire are different. As a matter of fact, if they were not different then we wouldn't have had any reason to do the study. Therefore we do not want to duplicate these observations *freq* number of times by using a FREQ statement. We have to use a WEIGHT statement to get the SAS procedure to make this sample data base 'look like' the population.

But before we do that, let's look at another common occurrence when you work with sample data: the sample is inaccurate in an identifiable way and has to be corrected. Going on with our sample of 10 out of 100, assume that we know that the population we are studying has 50 females and 50 males, but our sample has 6 males and 4 females. That is, we know the population is 50-50

Table 1

### COMPARISON OF FREQ AND WEIGHT STATEMENTS

| <u>Condition of Weighing Variable</u>   | <u>FREQ Statement Uses</u> | <u>WEIGHT Statement Uses</u> |
|-----------------------------------------|----------------------------|------------------------------|
| Integer (eg, 10)                        | Integer (10)               | Integer (10)                 |
| Whole Number (eg, 8.67)                 | Integer portion (8)        | Whole Number (8.67)          |
| Positive Decimal (eg, .88)              | Observation ignored        | Decimal Weight (.88)         |
| 0 or Negative Decimal (-1.5) or Missing | Observation ignored        | Weight of 0 (0)              |

male-female, but our sample data is telling us that the ratio is 60-40. If we don't correct this imbalance, our results would be inaccurate. (As a matter of statistical fact, there is a 75% chance that our sample would not contain 5 people of each gender and would need to be corrected.)

To correct this, we need to create another weighting variable that corrects this known defect in the sample. Since our sample has too many males in it, we have to decrease the importance of each male record. Likewise, we have to increase the importance of each female record because our sample has too few female records. So, unlike the SWGT variable which had only one value across the entire data base, this new weighting variable will have two values, one for males and one for females. One formula we can use is to divide the population proportion by the sample proportion, by gender:

```
IF GENDER = 'M' then CWGT = 50 / 60;
ELSE CWGT = 50 / 40;
```

The correction weight, CWGT, for males, then, is .833 while CWGT for females is 1.25.

Finally, we can combine the two weighting variables into one overall weight by multiplication:

```
FWGT = SWGT * CWGT;
```

This leads to a final weight, FWGT, of 8.33 for the males and 12.5 for the females. As a reality check, we can see that the number of males in the sample, 6, times the final male weight, 8.33, equals 49.98 -- much closer to the 50 we know the population contains. Likewise, the number of females in the sample, 4, times the final female weight, 12.5, equals 50. So the combination of the two weighting variables (SWGT and CWGT) into a single weight (FWGT) has brought the sample in line with the population. We can now proceed with analyses on this data base.

#### What We Expect to Find

We expect that the application of either a FREQ or WEIGHT statement will have an impact on the results we get from the SAS procedure, in this case PROC MEANS. In our example case, we expect that somehow we will get estimates for a population that has 50 males and 50 females, and that the other statistics computed for the

analysis variables will reflect this 50-50 distribution. Further, as a statistician, it is important to be sure that you know how many raw (unweighted) observations you have at each level of analysis so that you can be sure you have enough observations on which to base a conclusion.

Given these goals, let's see how the FREQ and WEIGHT statements behave, and how to get the desired results.

#### The Sample Data Base

In order to show the differential impact of the FREQ and WEIGHT statements, the following data were created. These 10 observations are intended to be the sample we have been discussing, the sample which represents the population of 100.

```
DATA TEMP;
INPUT
@1 GENDER $1. @3 INCOME 3.1;
CARDS;
M 111
M 333
M 555
M 777
M 999
M 222
F 444
F 666
F 888
F 111
;
RUN;
```

Next the weighting variables were added:

```
DATA TEMP;
SET TEMP;

SWGT = 10;

IF GENDER = 'M' THEN CWGT = .833;
ELSE CWGT = 1.25;

FACTOR = SWGT * CWGT;

RUN;
```

The data were sorted by gender and the following PROC MEANS statement was run five (5) different ways:

```
PROC MEANS N MEAN SUM SUMWGT;  
  BY GENDER;  
  VAR INCOME;  
RUN;
```

The five runs were:

1. No WEIGHT or FREQ
2. FREQ SWGT;
3. WEIGHT CWGT;
4. FREQ SWGT;  
 WEIGHT CWGT;
5. WEIGHT FWGT;

The first run gave the raw data totals. This is useful in actually seeing where the subsequent FREQ or WEIGHT statement changed the results.

The second run applied the FREQ statement only, while the third applied the WEIGHT statement only.

The fourth run applied both the FREQ and WEIGHT statements, something the SAS procedures allow.

Finally, the fifth statement applied the WEIGHT statement using the combined weight variable, FWGT.

The results of these five runs are examined below and are displayed in Table 2.

The raw data run provides the baseline for uncovering the impacts of the FREQ and WEIGHT statements and variables. The raw data show, correctly, that the analysis data base contains 10 observations (6 male and 4 female), and that the data are unweighted because SUMWGT = N. The SUM is the raw sum, and the MEAN is SUM / SUMWGT.

#### Impacts of the FREQ Statement

The FREQ statement, using SWGT, shows that PROC MEANS did exactly what it was told to do: it duplicated each observation SWGT number of times before it calculated the statistics. It therefore shows that the analysis data base contains 100 observations (60 male and 40 female). The SUM has also been increased by SWGT, as has SUMWGT. Note that the relationship between the SUM for males and females did not change. Basically, SAS has given exactly what it was asked to give, but the answers are misleading. The sample data base does not

contain 100 observations. It instructs PROC MEANS to behave as if it does, and this is incorrect.

#### Impacts of the WEIGHT Statement

In the third column, the changes caused by applying the CWGT variable in a WEIGHT statement can be seen. Note that SAS did not change the number of observations in the data base: it's still 6 males and 4 females. However, look at the impact on the SUM and the SUMWGT. The WEIGHT statement lowered the importance of each male record and increased the importance of each female record -- just as we intended it to do. The relationship between the male and female records was changed. And the SUMWGT results show what we intended: that the statistics were calculated as if the data base contained 4.998 males and 5.0 females. The male-female proportion is correct. Now we need to get SAS to give us results for the 100 in the population.

#### Impacts of Using Both FREQ and WEIGHT Statements

The FREQ and WEIGHT statements each provided part of what we were looking for, so what happens if we use both statements in one PROC MEANS? This was the third run and can be examined in Column 4 of Table 2. Note that again the FREQ statement caused the number of observations to be duplicated SWGT number of times, giving the incorrect impression that the analysis data base contained 100 observations. But the SUM and SUMWGT statistics are now correct. The total income (SUM) has been both proportionally adjusted by the WEIGHT statement and increased in magnitude by the FREQ statement. And the SUMWGT statistic shows what we expected, that the data base was treated as if the population was 49.98 males and 50 females. We are clearly closer to statistics we can use. We just need to eliminate the misleading N that the FREQ statement causes without eliminating the correct FREQ statement impact on SUM and SUMWGT.

#### Impacts of WEIGHT using the Combined Weight

The final weight variable, FWGT, was created to be the product of the sample weight (SWGT) and the correction weight (CWGT). It was then applied to the data base in

**PROC MEANS with a WEIGHT statement.** The results show us that the data base contained 6 male and 4 female observations (N) that represented 49.98 males and 50.0 females (SUMWGT). The total income for each group was adjusted (impact of the decimal portion of FWGT) and inflated (impact of the integer portion of FWGT). The results of this way of weighting the data base show that this is the only method that always produces the correct results.

#### **Summary**

This paper has shown that the **FREQ** and **WEIGHT** statements are different in more ways than just how they handle decimals. They do different things. They are not interchangeable. Table 3 shows the results of this analysis in generic form. It shows that when you are working with a sample data base that needs to be weighted, use only one weight variable and use it in a **WEIGHT** statement. Do not use the **FREQ** statement unless you intend to **DUPLICATE** observations.

**SAS®** is a registered trademark of The SAS Institute, Cary, NC USA.

Table 2

SUMMARY OF OUTPUT

In this example, the FREQ variable (SWGT) is the sample weight and is an integer. The WEIGHT variable (CWGT) is the correction weight and is non-integer. Both are non-zero positive numbers. The combination of the two weighting variables is called FWGT, the final weight.

| <u>SAS Calculated</u> | <u>Unweighted</u>           | <u>SWGT in<br/>FREQ<br/>Statement</u> | <u>CWGT in<br/>WEIGHT<br/>Statement</u> | <u>SWGT in<br/>FREQ and<br/>CWGT in<br/>WEIGHT<br/>Statements</u> | <u>FWGT *<br/>CWGT in<br/>WEIGHT<br/>Statement</u> |
|-----------------------|-----------------------------|---------------------------------------|-----------------------------------------|-------------------------------------------------------------------|----------------------------------------------------|
| <i>N</i>              | 6 (M)<br>4 (F)              | 60 (M)<br>40 (F)                      | 6 (M)<br>4 (F)                          | 60 (M)<br>40 (F)                                                  | 6 (M)<br>4 (F)                                     |
| <i>SUM</i>            | \$299.7 (M)<br>\$210.9 (F)  | \$2997 (M)<br>\$2109 (F)              | \$249.65 (M)<br>\$263.625 (F)           | \$2496.5 (M)<br>\$2636.25 (F)                                     | \$2496.5 (M)<br>\$2636.25 (F)                      |
| <i>SUMWGT</i>         | 6 (M)<br>4 (F)              | 60 (M)<br>40 (F)                      | 4.998 (M)<br>5.000 (F)                  | 49.98 (M)<br>50.0 (F)                                             | 49.98 (M)<br>50.0 (F)                              |
| <i>MEAN</i>           | \$49.95 (M)<br>\$52.725 (F) | \$49.95 (M)<br>\$52.725 (F)           | \$49.95 (M)<br>\$52.725 (F)             | \$49.95 (M)<br>\$52.725 (F)                                       | \$49.95 (M)<br>\$52.725 (F)                        |

Table 3

SUMMARY OF FREQ and WEIGHT STATEMENTS  
Example Using Proc Means

ONLY the calculations in **BOLD** are always correct. The other calculations ignore at least one piece of critical data.

| <u>SAS<br/>Calculated:</u> | <u>Raw Data</u> | <u>SWGT in<br/>FREQ<br/>Statement</u> | <u>CWGT in<br/>WEIGHT<br/>Statement</u> | <u>SWGT in<br/>FREQ, CWGT<br/>in WEIGHT<br/>Statements</u> | <u>FWGT *<br/>CWGT in<br/>WEIGHT<br/>Statement</u> |
|----------------------------|-----------------|---------------------------------------|-----------------------------------------|------------------------------------------------------------|----------------------------------------------------|
| <i>N</i>                   | Raw N           | N *<br>FREQ                           | Raw N                                   | N *<br>FREQ                                                | Raw N                                              |
| <i>SUM</i>                 | Raw Total       | SUM *<br>FREQ                         | SUM*<br>WEIGHT                          | SUM *<br>(WEIGHT *<br>FREQ)                                | SUM *<br>WEIGHT                                    |
| <i>SUMWGT</i>              | Raw Total       | N *<br>FREQ                           | N *<br>WEIGHT                           | N *<br>(WEIGHT *<br>FREQ)                                  | N *<br>WEIGHT                                      |
| <i>MEAN</i>                | SUM /<br>SUMWGT | SUM /<br>SUMWGT                       | SUM /<br>SUMWGT                         | SUM /<br>SUMWGT                                            | SUM /<br>SUMWGT                                    |