Downloading and distribution via your company's intranet of the following article in accordance with the terms and conditions hereinafter set forth is authorized by SAS Institute Inc. Each article must be distributed in complete form with all associated copyright, trademark, and other proprietary notices. No additional copyright, trademark, or other proprietary notices may be attached to or included with any article.

THE ARTICLE CONTAINED HEREIN IS PROVIDED BY SAS INSTITUTE INC. "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. RECIPIENTS ACKNOWLEDGE AND AGREE THAT SAS INSTITUTE INC. SHALL NOT BE LIABLE FOR ANY DAMAGES WHATSOEVER ARISING OUT OF THEIR USE OF THIS MATERIAL. IN ADDITION, SAS INSTITUTE INC. WILL PROVIDE NO SUPPORT FOR THE MATERIALS CONTAINED HEREIN.

Using the TABULATE Procedure for Computing Event Rates and Describing Variability in Safety and Efficacy Assessments

David Olaleye and Jim Goin

David Olaleye is a Pharmaceutical Industry Technologist at SAS Institute. Prior to joining the Institute, David worked as a Senior Statistician/Statistical Programming Manager with DataMedix Corporation located in Wayne, Pennsylvania. He has been a SAS user for the past 9 years, spending the last 4 years in consulting for pharmaceutical industries and academia. His areas of SAS expertise include base SAS software, the SAS Macro Language, SAS/STAT software and application development using SAS/FSP and SAS/AF software with FRAME entries. He also uses SAS to generate statistical analysis plans for clinical trials protocols, efficacy and safety tables/listings for New Drug Application submission. He has a Ph.D. in Demography (with concentration in Statistics) from the University of Pennsylvania, Philadelphia. David is a member of North-East SAS User Group.

James E. Goin is President and Founder of DataMedix Corporation, which is a contract research organization and biomedical consulting firm. He is a statistical and regulatory consultant to numerous pharmaceutical, biotechnology, device, and imaging companies, and has been a SAS user for many years. He currently uses SAS in the design and analysis of phase I through phase IV clinical research studies for evaluating the efficacy and safety of new drugs, biologicals, devices and therapeutic interventions. He received his Ph.D. in biostatistics from the University of Oklahoma in 1975.

Abstract

The objective of this article is two-fold: (1) to illustrate the use of PROC TABULATE for computing event rates using the life table method and (2) to show how the TABULATE procedure can be used to describe variability in failure times (or any other efficacy measure) for patients exposed to different treatment doses. The latter objective displays summary statistics along with percentages that describe the failure times distribution in terms of quartiles. Equivalent DATA steps for computing event rates as demonstrated for the first objective are also shown.

Contents

Introduction

The TABULATE procedure is a powerful tool for creating tabular reports and displaying hierarchical relationship among two or more variables in a single table. In most instances, the TABULATE procedure meets the needs of a clinical data analyst or statistical programmer when data summaries need only be presented as proportions and percentages, or as simple descriptive statistics such as the mean and standard deviation.

This article describes how PROC TABULATE can be extended to compute event rates such as failure, mortality, and adverse event rates using the life table method (Elandt-Johnson and Johnson, 1980). Tabulation of event rates can be helpful in complying with U.S. Food and Drug Administration (FDA) and other regulatory agencies' requirements for an application to market a new drug. Event rates can be used to summarize safety data, for example, by describing the overall event profile of the drug. Event rates can also be employed to compare efficacy assessments across treatment groups, (for example, for comparison of failure time or time to virus positivity between treatments).

This article also describes how to use a combination of DATA steps with PROC TABULATE to generate and display summary statistics for failure times by treatment and study period. The summary statistics include means, standard deviations, and information derived from the ranked distribution of patients' failure times. Reports generated in this manner may facilitate interpretation of the results of a clinical trial, and may be included in a clinical study report or in an integrated summary of safety or efficacy data.

Part I: Using PROC TABULATE to Compute Event Rates

This article utilizes two sample data sets to show how PROC TABULATE can be used to compute different types of event rates.

First Example Data Set

The first sample data came from a randomized clinical trial comparing the Halstead mastectomy with the quadrantectomy plus axillary dissection and radiotherapy in early breast cancer patients (Veronesi et al. 1981). A sample of 701 patients diagnosed as having a breast cancer less than 2cm in diameter without palpable nodes were enrolled in the trial. The researcher was interested in studying time free from different types of competing events such as distant disease (metastates), ipsilateral breast tumor, controlateral tumor, and death. For illustration purposes, 10 individual records randomly selected from the entire data set are used. The date of surgery (SURG_DT) represents the individual patient's entry into the study, while the date of death (DEATH_DT) or date of last contact (CONT_DT) represents the day that the patient died or was last observed. The observation time (YEARS), measured in years, is the difference between SURG_DT and CONT_DT or DEATH_DT. The raw data and the SAS DATA step that generated the SAS data set are given below:

```
data breast (keep=id years positive);
   input @1 id @4 surg_dt ddmmyy8. @14 cont_dt ddmmyy8. @24 death_dt ddmmyy8.
         @34 years @40 positive;
   format  surg_dt cont_dt death_dt ddmmyy8.;
   label    id='Observation ID'
       surg_dt='Surgery Date'
       cont_dt='Date of Last Contact'
      death_dt='Date of Death'
         years='Observation time (years)'
      positive='Indicator of failure: 1=event 0=censored';
cards;
1  27/06/73  .         14/04/84  10.81 1
2  17/07/73  27/01/92     .      18.49 0
3  28/01/74  18/11/91     .      17.82 0
4  29/04/76  08/01/91     .      14.70 0
5  17/01/77  .         28/09/85   8.70 1
6  19/04/78  .         15/12/88  10.67 1
7  24/07/78  12/05/92     .      13.45 0
8  26/07/78  08/06/92     .      13.43 0
9  23/04/80  .         08/03/85   4.88 1
10 30/04/80  11/03/92     .      11.68 0
;
run;
```

Second Example Data Set

The second sample data set comes from the AIDS virus data reported in Wei et al. (1989). The data were collected in a randomized clinical trial that sought to evaluate the effectiveness of the drug ribavirin among patients with acquired immune deficiency syndrome (AIDS) who were assigned to one of three treatment groups: placebo, low dose, and high dose of ribavirin. Blood samples were collected at weeks 4, 8, and 12, and each sample was observed each day to evaluate HIV viral load until virus positivity was detected via a surrogate marker. The day that virus positivity was detected is the failure time. As an example, if a sample was negative on days 1, 2, and 3, and on day 4 became positive, then the number of days to virus positivity is 4 and the failure time is 4 days. Since only AIDS

patients were enrolled in the trial, potentially each patient in the study should have three such failure times. Censored observations occurred when the culture required more days to register as positive than were achievable in the laboratory, or when the sample was contaminated before positivity was detected. The raw data and the DATA step that generated the SAS data set are given below:

```
data virus (keep=id sample treat days positive);
  array mth month1-month3;
  retain id sample 0;
  input treat month1-month3;
  id + 1;
  do over mth;
    if mth ^= . then
       do;
         if mth < 0 then do; days= -mth; positive= 0; end;
         else            do; days=  mth; positive= 1; end;
         sample= _i_;
         output;
       end;
  end;
cards;
 1 9 6 7
 1 4 5 10
 1 6 7 6
 1 10 . -21
 1 15 8 .
 1 3 . 6
 1 4 7 3
 1 9 12 12
 1 9 19 -19
 1 6 5 6
 1 9 . 18
 1 9 -20 -17
 2 6 4 5
 2 16 17 -21
 2 31 -19 -21
 2 -27 -19 .
 2 7 16 -23
 2 -28 7 -19
 2 -28 3 16
 2 15 12 16
 2 18 -21 22
 2 8 4 7
 2 4 -21 7
 3 21 9 -8
 3 13 7 -21
 3 16 6 20
 3 3 8 6
 3 21 . -25
 3 7 19 3
 3 11 13 -21
 3 -27 -18 9
 3 14 14 6
 3 8 11 15
 3 8 4 7
 3 8 3 9
 3 -19 10 -17
;
run;
```

Caution

The analysis of mortality and failure times demonstrated with PROC TABULATE in this article does not attempt to establish or evaluate the effectiveness of either the breast cancer treatment therapy or ribavirin. Therefore, no conclusion should be drawn from the results presented. The demonstration is simply for illustration purposes and to show that the function of PROC TABULATE can be extended to compute event rates.

Deriving Variables for Calculation of Event Rates

Computation of event rates such as mortality and virus positivity rates involves two factors: (1) identifying the number of observed events, either deaths or failures (numerator), and (2) the total observation time for all subjects in person-time units (denominator). Counting the number of events in the numerator term is straightforward. Computation of the denominator term requires the following: (1) the total duration time for subjects who experienced the events; (2) the total duration period up to the last day of interval for other subjects who did not experience the event; and (3) the duration period up to the last evaluated day for subjects who are lost to follow-up or who discontinued the study.

Reasons for this include the fact that the longer a subject is observed, the greater likelihood there is for failure, and the greater the ”weight” to be assigned to a subject in the evaluation (Marubini and Valsecchi, 1995).

Using the first sample data set, the SAS code below shows how to process the data and derive the variables needed to calculate event rates:

```
proc format;
 value positive     1='Event'  0='Censored';
 value yearfmt      1='[0,5)'  2='[5,10)'  3='[10,15)'  4='15+';
run;

data _temp2_;
  set breast;
  if 0 <= YEARS<=5  then group=1; else
  if 5 <  YEARS<=10 then group=2; else
  if 10<  YEARS<=15 then group=3; else
  if 15<  YEARS<=20 then group=4;

  do i=1 to group;
     if group<=i then
        do;
            interval=i;
            n_years=years-(5*(group-1));
            if      positive=0 then do; censor=1; fail=.; end;
            else if positive=1 then do; censor=.; fail=1; end;
        end;
     if group > i then
        do; interval=i;  n_years=5; end;
     n_intv=interval;    /*Create a duplicate variable for Interval*/
     output;
   end;
   drop i group;
run;
```

First, a GROUP variable is created to classify the failure or observed duration times (YEARS) into 5-year intervals. It is assumed that the rate is constant during each of the intervals but not necessarily across them. The choice of a 5-year interval is arbitrary since other meaningful intervals can be chosen. Within a given interval, the DO-loop assigns values to INTERVAL, N_YEARS, POSITIVE, CENSOR, and FAIL variables that PROC TABULATE uses for computation. The INTERVAL variable indicates the time interval, and for each interval the N_YEARS variable indicates the computed person-years for each subject at risk. The CENSOR and FAIL variables indicate whether the

observation is censored or is an event. As demonstrated in the following PROC TABULATE statements, the FAIL variable is used in creating the numerator term, and N_YEARS is used in creating the denominator term.

Case 1: Calculating Mortality Rates

Figure 1 presents output from our extension of PROC TABULATE that computes mortality rates. The following PROC TABULATE statements show how to request and obtain the numerator and denominator terms needed to compute the breast cancer mortality rate. The CLASS statement identifies the INTERVAL variable as the classification variable. The VAR statement lists the analysis variables that PROC TABULATE uses to compute the following totals: number of subjects at risk within a given interval (N_INTV, column 2 in TABULATE procedure output), number of subjects censored due no follow-up or end of observation period (CENSOR, column 3), number experiencing the event (FAIL, column 4), and the total person-years attributable to subjects in the at-risk set (N_YEARS, column 5).

The event (mortality) rate shown in column 6 in the TABULATE procedure output is calculated as the ratio of the total number of observed deaths to the total time all subjects in a given interval have been observed (that is, column 4/column 5). The term **fail=' ' *pctsum<n_years>='Mortality Rate (%)'** contained in the TABLE statement tells TABULATE to summarize the values of the FAIL variable within each given interval and to divide it by the sum of the values of the N_YEARS variable in that same interval multiplied by 100 to obtain the mortality rate for the interval. For example, during the first five years of follow-up, there was one death (which occurred at 4.88 years) and the remaining 9 subjects survived through the entire interval. Thus, the rate in this first interval is calculated as 0.02 or [1/(4.88+9*5) * 100 =2%]. Similarly, for the next interval, the estimate of mortality rate for the 9 subjects at risk is 0.023 or [1/(3.70+8*5) * 100 =2.3%]. For the entire sample, the estimated mortality rate is given in column 6 of the last row [4/(50+44+25+6)*100=3.21]. The term **misstext='0'** is used to replace all cells having missing values with 0.

```
proc sort data=_temp2_; by interval; run;
options center nonumber nodate ps=54 ls=80;
proc tabulate data=_temp2_ order=data format=10. missing;
 class interval;
 var n_years censor fail n_intv;
 table (interval='Time Interval' all='Overall'),
(n_intv='Number Interval'*N=' '*f=8.) (censor='Number Censored'*sum=' '*f=8.)
 (fail='Number Failed'*sum=' '*f=7.) (n_years='Person-Years'*sum=' '*f=8.)
 (fail=' ' * pctsum<n_years>='Mortality Rate (%)'* F=9.2)
 / misstext='0' rts=22 condense box='Breast Cancer Sample';
format interval yearfmt. ;

title1 "Figure 1. Calculate Event Rates for Early Breast Cancer Patients in";
title2 "Veronesi et al.'s (1981) study";
run;
```

Deriving Variables for Calculation of Virus Positivity Rates

The second sample data is used to demonstrate calculation of virus positivity rates for blood samples collected at different time points in an AIDS clinical trial study. The SAS code below shows how to process the data and derive the variables needed to calculate the event rates. Explanation of the code follows the same manner that was used for processing the data set for breast cancer mortality rate. Here, the unit measurement for observation period is days. A GROUP variable is created to classify the time to virus positivity or end of observation period (DAYS) into 5-day intervals. Another variable, SAMPLE, is used to stratify the blood samples by week of collection (stratum variable). For efficacy assessment of ribavirin, the purpose in this case is to facilitate comparison of the estimated positivity rates for the different dose groups of ribavirin across blood sample collection weeks.

```
proc format;
   value sample   1='Week 4'  2='Week 8'   3='Week 12' 9='All Blood Samples';
   value treat    1='Placebo' 2='Low Dose' 3='High Dose'  9='Overall';
   value positive 1='Event'   0='Censored';
   value interval 1='[0,5)'   2='[5,10)'  3='[10,15)' 4='[15,20)'
                  5='[20,25)' 6='[25,30)' 7='30+';
run;

data _temp2b_;
  set virus;
  if 1<=sample<=3;
  if 0 <= DAYS <=5  then group=1; else
  if 5 <  DAYS <=10 then group=2; else
  if 10<  DAYS <=15 then group=3; else
  if 15<  DAYS <=20 then group=4; else
  if 20<  DAYS <=25 then group=5; else
  if 25<  DAYS <=30 then group=6; else
  if      DAYS > 30 then group=7;

  do i=1 to group;
     if group<=i then
        do;
            interval=i;
            n_days=days-(5*(group-1));
            if      positive=0 then do; censor=1; fail=.; end;
            else if positive=1 then do; censor=.; fail=1; end;
        end;
     if group > i then
        do; interval=i;  n_days=5;  end;
     n_intv=interval;    /*Create a duplicate variable for Interval*/
     output;
  end;
  drop i group;
run;
```

Case 2: Calculating Virus Positivity Rates

Figure 2 presents output from our extension of PROC TABULATE that computes virus positivity rates. From the PROC TABULATE statements, the CLASS statement identifies SAMPLE, TREAT, and INTERVAL variables as classification variables. The VAR statement lists the analysis variables that PROC TABULATE uses to compute the following totals for each treatment group: number of subjects' blood samples at risk in a given interval (N_INTV, column 3 in TABULATE procedure output); number of censored observations due to non-availability of sample or contamination of sample before positivity was detected or when the culture required more days to register as positive than were achievable in the laboratory (CENSOR, column 4); number of samples that experienced the event, that is, registered as virus positive (FAIL, column 5); and the total culture-days attributable to blood samples in the at-risk set (N_DAYS, column 6).

The event (virus positivity) rate shown in column 7 of the table in Figure 2 is calculated as the ratio of the total number of blood samples that registered as virus positive to the total number of culture-days observed for the blood samples within that given interval (that is, column 5/column 6). The term **fail=' ' *pctsum<n_days>='Event Rate (%)'** contained in the TABLE statement tells PROC TABULATE to summarize the values of the FAIL variable within each given interval and to divide it by the sum of the values of the N_DAYS variable in that same interval multiplied by 100 to obtain the virus positivity rate. For example, during the first five days, and for blood samples randomly assigned to the placebo group, 10 subjects' blood samples were at risk during the interval in which 3 blood samples registered as virus positive and none were censored; thus, the rate in this first interval is calculated as 5.36 or [3/(56) *100 =5.36%]. The term **sample=' ' ALL='All Blood Samples'** in the TABLE statement indicates the *page-expression* and tells the TABULATE procedure to compute positivity rates for each week of sample

collection and the combined blood samples. The term **misstext='0'** is used to replace all cells having missing values with 0.

```
proc sort data=_temp2b_; by sample treat interval; run;

options center nonumber nodate ps=54 ls=80;
proc tabulate data=_temp2b_ order=data format=10. missing ;
 class sample treat interval;
 var n_days censor fail n_intv;
 table (sample=' ' ALL='All Blood Samples'),
 (treat='Treatment Group') * (interval='Time Interval'),
 (n_intv='Number Interval'*N=' '*f=8.) (censor='Number Censored'*sum=''*f=8.)
 (fail='Number Failed' * sum=' '*f=8.) (n_days='Culture-Days'*sum=' '*f=8.)
 (fail=' ' * pctsum<n_days>='Event Rate (%)'* F=14.2)
 / misstext='0' rts=22 condense box=_page_;
 format sample sample. interval interval. treat treat.;
title1 "Figure 2: Calculate Virus Positivity Rates for Each Treatment Group
for";
title2 "Blood Samples at Weeks 4, 8, 12 and All Combined Blood Samples";
run;
```

Validation of Event Rate Computations

SAS DATA step solution that computes virus positivity rates similar to those obtained from the code in the preceding section is given below. The output of this validation code is the same as that presented in Figure 2 and has been omitted in this article for reasons of brevity.

```
data _temp2b_;
  set virus;
  if 1<=sample<=3;
  if 0 <= DAYS <=5  then group=1; else
  if 5 <  DAYS <=10 then group=2; else
  if 10<  DAYS <=15 then group=3; else
  if 15<  DAYS <=20 then group=4; else
  if 20<  DAYS <=25 then group=5; else
  if 25<  DAYS <=30 then group=6; else
  if      DAYS > 30 then group=7;
  do i=1 to group;
     if group<=i then
        do;
            interval=i;
            n_days=days-(5*(group-1));
            if      positive=0 then do; censor=1; fail=.; end;
            else if positive=1 then do; censor=.; fail=1; end;
        end;
     if group > i then
        do; interval=i;  n_days=5;  end;
     output;
  end;
  drop i group;
run;

proc sort data=_temp2b_; by sample treat interval id; run;
data _temp3_;
  set _temp2b_ (drop=n_intv) end=last;
  by sample treat interval id;
  if first.interval then do; n_intv=0; p_days=0; n_cens=0; n_fail=0; end;
```

```
  n_intv+1;
  p_days+n_days;
  if censor=1 then n_cens+censor;
  if fail=1 then n_fail+fail;
  rate=n_fail/p_days*100;
  if last.interval then output;
  drop id positive censor fail days n_days;
 run;

proc sort data=_temp2b_; by treat interval id; run;
data _temp4_;
  set _temp2_(drop=n_intv) end=last;
  by treat interval id;
  sample=9;            /*Category for Total of all blood samples combined*/

  if first.interval then do; n_intv=0; p_days=0; n_cens=0; n_fail=0; end;
  n_intv+1;
  p_days+n_days;
  if censor=1 then n_cens+censor;
  if fail=1 then n_fail+fail;
  rate=n_fail/p_days*100;
  if last.interval then output;
  drop id positive censor fail days n_days;
run;

data _temp34_;
  set _temp3_ _temp4_;
  proc fsview data=_temp34_;
run;

options center nonumber nodate ps=54 ls=80;
proc tabulate data=_temp34_ order=data format=10. missing ;
 class sample treat interval;
 var n_intv p_days n_cens n_fail rate;
 table sample=' ',(treat='Treatment Group') * (interval='Time Interval'),
 (n_intv='Number  Interval'*sum='  '*f=8.)  (n_cens='Number  Censored'*sum='
'*f=8.)
 (n_fail='Number Failed'*sum=' '*f=8.) (p_days='Culture-Days'*sum=' '*f=8.)
 (rate='Event Rate(%)' * sum=' '*f=14.2)
 / misstext='0' rts=22 condense box=_page_;
   format sample sample. interval interval. treat treat.;

title1 "Using DATA Step and By-Group Statements to Validate and Reproduce";
title2 "Life Table Estimates Obtained with Proc Tabulate in Figure 2";
title3 "Compute Event Rates for Each Treatment Group by Blood Sample and
Combined Blood Sample";
run;
```

The SAS DATA step solution yields the same results obtained with the previous SAS code. Although both SAS programs employ the TABULATE procedure to report the results, this particular program uses more lines of SAS code, computes the event rates within the SAS DATA step, and computes the values for the combined totals (that is, all blood samples) in a different DATA step. On the other hand, the previous SAS program makes maximum use of the functions and computing power of the TABULATE procedure.

Part II: Using PROC TABULATE to Present Percentile Information

The second part of this article describes the use of PROC TABULATE for generating percentile information that can be used to describe the distribution of an efficacy variable of interest or make comparisons across different sets of data or samples. This example is illustrated by comparing the failure times (time to detection of virus positivity) of blood samples obtained from patients exposed to different doses of ribavirin or placebo at different times.

Presently, PROC TABULATE computes different forms of summary statistics which include the mean, range, and standard deviation for any given data set. Details on how to obtain such statistics can be found in the *SAS Guide to TABULATE Processing, Second Edition.* However, there are instances when one may want to show additional information about the behavior of an efficacy variable of interest to complement the statistics provided by PROC TABULATE. For example, one may wish to show how the values of the variable are distributed across different treatment groups or strata. Descriptive statistics to adopt in this circumstance might include quartiles, deciles and percentiles. These statistics are referred to as measures of location and they are useful for comparing values or describing the distribution of values of a given variable. Quartiles divide the ranked scores into four equal parts. (Q_1, Q_2, and Q_3 are the values that divide the parts.) Q_1 separates the bottom 25% of the ranked scores from the top 75%; Q_2 (known as the median) separates the ranked scores into two equal parts; and Q_3 separates the top 25% of the ranked scores from the bottom 75%. In the same way, deciles partition the data into 10 groups with about 10% of the data in each group, while percentiles partition the data into 100 groups with about 1% of the scores in each group. In SAS, these measures are produced by PROC UNIVARIATE.

Deriving Variables for Calculating Percentages of Failure Times

The following SAS code was used to process the data set needed for deriving variables to calculate percentages of failure times. The code is summarized in the following steps. First, using the second example SAS data set (VIRUS), a new data set (STATS) was created using the OUT= option in PROC UNIVARIATE. The data set contained the quartiles (or if desired, deciles or percentiles) and the minimum and maximum values of the efficacy variable. In this example, the efficacy variable is time to virus positivity detection (DAYS).

Second, in the next DATA step the values of the quartiles and other statistics are passed to macro variables using the CALL SYMPUT statement. Also in this DATA step, another macro variable (TOTCAT) was created to hold the value that will be used to represent the 'Total' category in the PROC FORMAT definition for the QUARTILE variable. Third, PROC FORMAT was used to create formatted labels for categories of QUARTILE. [*Caution*: The placement of PROC FORMAT statements used for creating formatted labels of QUARTILE is very important. The format definition must come after the DATA step used to create the macro variables. Otherwise, SAS will report error messages to the *LOG* window for not being able to resolve the macro variables used in the PROC FORMAT statements.] The last DATA step creates a SAS data set (_TEMP3_) that will be used in PROC TABULATE. In addition, the values of the macro variables are used to create categories for the QUARTILE variable.

In this example, summarization of data using information on quartiles computed from the patients' failure times serves two purposes: (1) to compare percentages of failure times both across treatment groups and blood sample collection weeks and (2) to facilitate detection of whether the variable of interest has the same or skewed distribution across the treatment groups.

```
proc univariate data=virus noprint;
  var days;
  output out=stats min=min max=max median=med q1=q1 q3=q3;
run;

data stats;
   set stats;
   call symput('min',left(min));
   call symput('max',left(max));
   call symput('med',left(med));
   call symput('med_',left(med+1));
   call symput('q1',left(q1));
```

```
   call symput('q3',left(q3));
   call symput('q1_',left(q1+1));
   call symput('q3_',left(q3+1));
   call symput('totcat',left(max+10)); /*Macro variable to represent the*/
                                       /*value for the Total Category*/
run;

proc format;
  value quartile  1="%left(&min) to &q1"  2="%left(&q1_) to &med"
                  3="%left(&med_) to &q3" 4="%left(&q3_) to &max"
                  &totcat='Total';
  value sample    1='Week 4'  2='Week 8'   3='Week 12'    9='Total';
  value treat     1='Placebo' 2='Low Dose' 3='High Dose'  9='Overall';
run;

data _temp5_;
   set virus;
   if &min<=days<=&q1  then quartile=1; else
   if &q1  <days<=&med then quartile=2; else
   if &med <days<=&q3  then quartile=3; else
   if &q3  <days<=&max then quartile=4;
run;
```

Creating a Summary Report with PROC TABULATE

Figure 3 presents output showing the summary report that was generated using PROC TABULATE. The PROC TABULATE statements in this case concatenate the usual summary statistics (mean and standard deviation) and descriptive information on quartiles (that is, the proportion of cases falling between the lower and upper boundaries of quartiles). This type of reporting may serve as a useful validation guide for 1) verifying the consistency and meaningfulness of the summary statistics, 2) detecting outliers among the values of the efficacy variable, and 3) checking for strange values or errors in computation.

The CLASS statement identifies SAMPLE, TREAT, POSITIVE and QUARTILE as the classification variables. The term **sample='Sample' ALL='Total'** in the TABLE statement indicates the *page-expression* term and tells the TABULATE procedure to compute the descriptive summary statistics for the different weeks of sample collection and combined blood samples.

For each treatment group, the TABLE statement computes the percentages of failure times in each quartile, POSITIVE (whether a patient experienced the event or censored), and then concatenates the results with those of the descriptive statistics. From the output, it was observed that in the placebo group and for blood samples collected in Week 4, all placebo patients experienced the event with over 90% having time to virus positivity less than or equal to the overall median time to virus positivity. On the other hand, for that same week, it was noted that 3 blood samples were censored in the last quartile interval in the low dose group, and 1 blood sample was censored in each of the last two quartiles for the high dose group.

```
proc sort data=_temp5_; by treat sample quartile; run;
options center nonumber nodate ps=43 ls=126;
proc tabulate data=_temp5_ order=data format=6.1 missing;
  class sample treat positive quartile;
  var days;
  table (sample='Sample' ALL='Total'),
     treat='Treatment' * (positive='Event Status' ALL='Total'),
    (quartile='Failure Times (Quartiles)' * (N*f=4. pctn<quartile>)
     all='Total'*(N*f=6. pctn<all>))
    days='Failure Times (Days)' * (N*f=4. Mean std Min Max)
    / misstext='0' rts=22 condense box=_page_;
```

```
  format treat treat. sample sample. quartile quartile. positive positive.;
title1 "Figure 3: Descriptive Summary Statistics of Time to Virus
Positivity";
title2 "Detection by Treatment, Event Status and Blood Sample Collection
     Week";
run;
```

Output from another example of PROC TABULATE is presented in Figure 4. The output gives the descriptive summary statistics of time to virus positivity by treatment group and blood sample collection week. Here, the rows of the table show a hierarchical relationship, with respect to time to virus positivity, by the patient's treatment group assignment and week of blood sample collection. The columns display the descriptive aggregate information on a patient's blood sample time to virus positivity.

```
options center nonumber nodate ps=43 ls=126;
proc tabulate data=_temp_ order=data format=6.1 missing;
  class sample treat quartile;
  var days;
  table (treat='Treatment' ALL='All Group')* (sample='Sample' ALL='Total'),
   (quartile='Failure Times (Quartiles)' * (N*f=4. pctn<quartile>)
             all='Total'*(N*f=6. pctn<all>))
   days='Failure Times (Days)' * (N*f=4. Mean std Min Max)
         / misstext='0' rts=22 condense box=_page_;
   format treat treat. sample sample. quartile quartile. positive positive.;
title1 "Figure 4: Descriptive Summary Statistics of Time to Detection of
Virus";
title2 "Positivity by Treatment and Blood Sample Collection Week";
run;
```

Summary

This article describes two different applications of PROC TABULATE for summarizing and reporting data. The article also shows that the functions of the TABULATE procedure can be extended to compute event rates such as mortality rate or virus positivity rates. The beauty of using PROC TABULATE as a reporting tool lies in its simplicity and parsimony. A reporting task that requires many lines of code with SAS DATA steps and PUT statements can be easily accomplished with one or two PROC TABULATE statements. The SAS macro programs presented in this article can be generalized to compute other similar event rates.

This paper also highlights the use of PROC TABULATE for producing customized tables and descriptive summary information on an efficacy variable of interest. Two examples show a concatenated output of summary statistics and percentages derived from quartiles of the ranked distribution of time to virus positivity.

References

Bernholc, Alissa (1997), "A Tip for Using the TABULATE Procedure: Generating Tables with Percentages in Only One Category," *Observations*, The Technical Journal for SAS Software Users, Cary, NC: SAS Institute Inc.

Elandt-Johnson, R.C. and Johnson, N.L. (1980), "Survival Models and Data Analysis," New York: John Wiley & Sons, Inc.

Kent, Paul (1995), "SQL Joins - The Long and The Short of It." *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., 206-215.

L.J. Wei, D.Y. Lin, and L. Weissfeld (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distribution," Journal of the American Statistical Association, 84, 1065-1071.

SAS Institute Inc. (1990), *SAS Guide to TABULATE Processing, Second Edition*, Cary, NC: SAS Institute Inc., 208.

SAS Institute Inc. (1997), *SAS Macro Language: Reference*, Cary, NC: SAS Institute Inc., 96-99, 175-176.

Veronesi U. et al. (1981), "Comparing radical mastectomy with quadrantectomy, axillary dissection and radiotherapy in patients with small cancers of the breast," New England Journal of Medicine, 305, 6-11.

Modified code is not supported by the authors or SAS Institute.

Figure 1. Calculate Event Rates for Early Breast Cancer Patients inVeronesi et al.'s (1981) study

Breast Cancer Sample	Number Interval	Number Censored	Number Failed	Person-Years	Mortality Rate (%)
Time Interval					
[0,5)	10	0	1	50	2.00
[5,10)	9	0	1	44	2.29
[10,15)	8	4	2	25	8.08
15+	2	2	0	6	0
Overall	29	6	4	125	3.21

Figure 2: Calculate Virus Positivity Rates for Each Treatment Group for Blood Samples at Weeks 4, 8, 12 and All Combined Blood Samples

Week 4		Number Interval	Number Censored	Number Failed	Culture-Days	Event Rate (%)
Treatment Group	Time Interval					
Placebo	[0,5)	12	0	3	56	5.36
	[5,10)	9	0	8	32	25.00
	[10,15)	1	0	1	5	20.00
Low Dose	[0,5)	11	0	1	54	1.85
	[5,10)	10	0	3	41	7.32
	[10,15)	7	0	1	35	2.86
	[15,20)	6	0	2	24	8.33
	[20,25)	4	0	0	20	0
	[25,30)	4	3	0	13	0
	30+	1	0	1	1	100.00
High Dose	[0,5)	13	0	1	63	1.59
	[5,10)	12	0	4	51	7.84
	[10,15)	8	0	3	33	9.09
	[15,20)	5	1	1	20	5.00
	[20,25)	3	0	2	7	28.57
	[25,30)	1	1	0	2	0

Figure 2 (continued): Calculate Virus Positivity Rates for Each Treatment Group for Blood Samples at Weeks 4, 8, 12 and All Combined Blood Samples

Week 8		Number Interval	Number Censored	Number Failed	Culture-Days	Event Rate (%)
Treatment Group	Time Interval					
Placebo	[0,5)	9	0	2	45	4.44
	[5,10)	7	0	4	23	17.39
	[10,15)	3	0	1	12	8.33
	[15,20)	2	1	1	9	11.11
Low Dose	[0,5)	11	0	3	51	5.88
	[5,10)	8	0	1	37	2.70
	[10,15)	7	0	1	32	3.13
	[15,20)	6	2	2	21	9.52
	[20,25)	2	2	0	2	0
High Dose	[0,5)	12	0	2	57	3.51
	[5,10)	10	0	5	40	12.50
	[10,15)	5	0	3	18	16.67
	[15,20)	2	1	1	7	14.29

Figure 2 (continued): Calculate Virus Positivity Rates for Each Treatment Group for Blood Samples at Weeks 4, 8, 12 and All Combined Blood Samples

Week 12		Number Interval	Number Censored	Number Failed	Culture-Days	Event Rate (%)
Treatment Group	Time Interval					
Placebo	[0,5)	11	0	1	53	1.89
	[5,10)	10	0	5	35	14.29
	[10,15)	5	0	1	22	4.55
	[15,20)	4	2	1	14	7.14
	[20,25)	1	1	0	1	0
Low Dose	[0,5)	10	0	1	50	2.00
	[5,10)	9	0	2	39	5.13
	[10,15)	7	0	0	35	0
	[15,20)	7	1	2	26	7.69
	[20,25)	4	3	1	7	14.29
High Dose	[0,5)	13	0	1	63	1.59
	[5,10)	12	1	5	45	11.11
	[10,15)	6	0	1	30	3.33
	[15,20)	5	1	1	22	4.55
	[20,25)	3	3	0	7	0

Figure 2 (continued): Calculate Virus Positivity Rates for Each Treatment Group for Blood Samples at Weeks 4, 8, 12 and All Combined Blood Samples

All Blood Samples		Number Interval	Number Censored	Number Failed	Culture-Days	Event Rate (%)
Treatment Group	Time Interval					
Placebo	[0,5)	32	0	6	154	3.90
	[5,10)	26	0	17	90	18.89
	[10,15)	9	0	3	39	7.69
	[15,20)	6	3	2	23	8.70
	[20,25)	1	1	0	1	0
Low Dose	[0,5)	32	0	5	155	3.23
	[5,10)	27	0	6	117	5.13
	[10,15)	21	0	2	102	1.96
	[15,20)	19	3	6	71	8.45
	[20,25)	10	5	1	29	3.45
	[25,30)	4	3	0	13	0
	30+	1	0	1	1	100.00
High Dose	[0,5)	38	0	4	183	2.19
	[5,10)	34	1	14	136	10.29
	[10,15)	19	0	7	81	8.64
	[15,20)	12	3	3	49	6.12
	[20,25)	6	3	2	14	14.29
	[25,30)	1	1	0	2	0

Figure 3: Descriptive Summary Statistics of Time to Virus Positivity Detection by Treatment, Event Status and Blood Sample Collection Week

Sample Week 4		Failure Times (Quartiles)								Total		Failure Times (Days)				
		3 to 7		8 to 10		11 to 19		20 to 31								
		N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	MEAN	STD	MIN	MAX
Treatment / Placebo	Event Status															
	Event	5	41.7	6	50.0	1	8.3	0	0	12	100.0	12	7.8	3.3	3.0	15.0
	Total	5	41.7	6	50.0	1	8.3	0	0	12	100.0	12	7.8	3.3	3.0	15.0
Low Dose	Event Status															
	Event	3	37.5	1	12.5	3	37.5	1	12.5	8	100.0	8	13.1	8.9	4.0	31.0
	Censored	0	0	0	0	0	0	3	100.0	3	100.0	3	27.7	0.6	27.0	28.0
	Total	3	27.3	1	9.1	3	27.3	4	36.4	11	100.0	11	17.1	10.1	4.0	31.0
High Dose	Event Status															
	Event	2	18.2	3	27.3	4	36.4	2	18.2	11	100.0	11	11.8	5.8	3.0	21.0
	Censored	0	0	0	0	1	50.0	1	50.0	2	100.0	2	23.0	5.7	19.0	27.0
	Total	2	15.4	3	23.1	5	38.5	3	23.1	13	100.0	13	13.5	7.0	3.0	27.0

Figure 3 (continued): Descriptive Summary Statistics of Time to Virus Positivity Detection by Treatment, Event Status and Blood Sample Collection Week

Sample Week 8		Failure Times (Quartiles)								Total		Failure Times (Days)				
		3 to 7		8 to 10		11 to 19		20 to 31								
		N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	MEAN	STD	MIN	MAX
Treatment	Event Status															
Placebo	Event	5	62.5	1	12.5	2	25.0	0	0	8	100.0	8	8.6	4.7	5.0	19.0
	Censored	0	0	0	0	0	0	1	100.0	1	100.0	1	20.0	0	20.0	20.0
	Total	5	55.6	1	11.1	2	22.2	1	11.1	9	100.0	9	9.9	5.8	5.0	20.0
Low Dose	Event Status															
	Event	4	57.1	0	0	3	42.9	0	0	7	100.0	7	9.0	5.9	3.0	17.0
	Censored	0	0	0	0	2	50.0	2	50.0	4	100.0	4	20.0	1.2	19.0	21.0
	Total	4	36.4	0	0	5	45.5	2	18.2	11	100.0	11	13.0	7.2	3.0	21.0
High Dose	Event Status															
	Event	4	36.4	3	27.3	4	36.4	0	0	11	100.0	11	9.5	4.7	3.0	19.0
	Censored	0	0	0	0	1	100.0	0	0	1	100.0	1	18.0	0	18.0	18.0
	Total	4	33.3	3	25.0	5	41.7	0	0	12	100.0	12	10.2	5.1	3.0	19.0

Figure 3 (continued): Descriptive Summary Statistics of Time to Virus Positivity Detection by Treatment, Event Status and Blood Sample Collection Week

Sample Week 12		Failure Times (Quartiles)								Total		Failure Times (Days)				
		3 to 7		8 to 10		11 to 19		20 to 31								
		N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	MEAN	STD	MIN	MAX
Treatment	Event Status															
Placebo	Event	5	62.5	1	12.5	2	25.0	0	0	8	100.0	8	8.5	4.7	3.0	18.0
	Censored	0	0	0	0	2	66.7	1	33.3	3	100.0	3	19.0	2.0	17.0	21.0
	Total	5	45.5	1	9.1	4	36.4	1	9.1	11	100.0	11	11.4	6.4	3.0	21.0
Low Dose	Event Status															
	Event	3	50.0	0	0	2	33.3	1	16.7	6	100.0	6	12.2	6.8	5.0	22.0
	Censored	0	0	0	0	1	25.0	3	75.0	4	100.0	4	21.0	1.6	19.0	23.0
	Total	3	30.0	0	0	3	30.0	4	40.0	10	100.0	10	15.7	6.9	5.0	23.0
High Dose	Event Status															
	Event	4	50.0	2	25.0	1	12.5	1	12.5	8	100.0	8	9.4	5.5	3.0	20.0
	Censored	0	0	1	20.0	1	20.0	3	60.0	5	100.0	5	18.4	6.5	8.0	25.0
	Total	4	30.8	3	23.1	2	15.4	4	30.8	13	100.0	13	12.8	7.3	3.0	25.0

Figure 3 (continued): Descriptive Summary Statistics of Time to Virus Positivity Detection by Treatment, Event Status and Blood Sample Collection Week

Total		Failure Times (Quartiles)								Total		Failure Times (Days)				
		3 to 7		8 to 10		11 to 19		20 to 31								
		N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	MEAN	STD	MIN	MAX
Treatment	Event Status															
Placebo	Event	15	53.6	8	28.6	5	17.9	0	0	28	100.0	28	8.2	4.0	3.0	19.0
	Censored	0	0	0	0	2	50.0	2	50.0	4	100.0	4	19.3	1.7	17.0	21.0
	Total	15	46.9	8	25.0	7	21.9	2	6.3	32	100.0	32	9.6	5.3	3.0	21.0
Low Dose	Event Status															
	Event	10	47.6	1	4.8	8	38.1	2	9.5	21	100.0	21	11.5	7.3	3.0	31.0
	Censored	0	0	0	0	3	27.3	8	72.7	11	100.0	11	22.5	3.6	19.0	28.0
	Total	10	31.3	1	3.1	11	34.4	10	31.3	32	100.0	32	15.3	8.2	3.0	31.0
High Dose	Event Status															
	Event	10	33.3	8	26.7	9	30.0	3	10.0	30	100.0	30	10.3	5.3	3.0	21.0
	Censored	0	0	1	12.5	3	37.5	4	50.0	8	100.0	8	19.5	5.8	8.0	27.0
	Total	10	26.3	9	23.7	12	31.6	7	18.4	38	100.0	38	12.2	6.5	3.0	27.0

Figure 4: Descriptive Summary Statistics of Time to Detection of Virus Positivity by Treatment and Blood Sample Collection Week

		Failure Times (Quartiles)								Total		Failure Times (Days)				
		3 to 7		8 to 10		11 to 19		20 to 31								
		N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	MEAN	STD	MIN	MAX
Treatment	Sample															
Placebo	Week 4	5	41.7	6	50.0	1	8.3	0	0	12	100.0	12	7.8	3.3	3.0	15.0
	Week 8	5	55.6	1	11.1	2	22.2	1	11.1	9	100.0	9	9.9	5.8	5.0	20.0
	Week 12	5	45.5	1	9.1	4	36.4	1	9.1	11	100.0	11	11.4	6.4	3.0	21.0
	Total	15	46.9	8	25.0	7	21.9	2	6.3	32	100.0	32	9.6	5.3	3.0	21.0
Low Dose	Sample															
	Week 4	3	27.3	1	9.1	3	27.3	4	36.4	11	100.0	11	17.1	10.1	4.0	31.0
	Week 8	4	36.4	0	0	5	45.5	2	18.2	11	100.0	11	13.0	7.2	3.0	21.0
	Week 12	3	30.0	0	0	3	30.0	4	40.0	10	100.0	10	15.7	6.9	5.0	23.0
	Total	10	31.3	1	3.1	11	34.4	10	31.3	32	100.0	32	15.3	8.2	3.0	31.0
High Dose	Sample															
	Week 4	2	15.4	3	23.1	5	38.5	3	23.1	13	100.0	13	13.5	7.0	3.0	27.0
	Week 8	4	33.3	3	25.0	5	41.7	0	0	12	100.0	12	10.2	5.1	3.0	19.0
	Week 12	4	30.8	3	23.1	2	15.4	4	30.8	13	100.0	13	12.8	7.3	3.0	25.0
	Total	10	26.3	9	23.7	12	31.6	7	18.4	38	100.0	38	12.2	6.5	3.0	27.0

Figure 4 (continued): Descriptive Summary Statistics of Time to Detection of Virus Positivity by Treatment and Blood Sample Collection Week

		Failure Times (Quartiles)								Total		Failure Times (Days)				
		3 to 7		8 to 10		11 to 19		20 to 31								
		N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	PCTN	N	MEAN	STD	MIN	MAX
All Group	Sample															
	Week 4	10	27.8	10	27.8	9	25.0	7	19.4	36	100.0	36	12.7	8.0	3.0	31.0
	Week 8	13	40.6	4	12.5	12	37.5	3	9.4	32	100.0	32	11.1	6.1	3.0	21.0
	Week 12	12	35.3	4	11.8	9	26.5	9	26.5	34	100.0	34	13.2	6.9	3.0	25.0
	Total	35	34.3	18	17.6	30	29.4	19	18.6	102	100.0	102	12.4	7.1	3.0	31.0