

COMPUTING HIGHWAY CRASH RATES USING PROC SUMMARY

Elizabeth G. Hamilton

University of North Carolina Highway Safety Research Center

ABSTRACT

This paper describes the technique of linking accidents to the roadway and using PROC SUMMARY to compute the variables needed in the crash rate formula. Crash rate is the statistic most often used to compare roadways. The effectiveness of various countermeasures or roadway improvements is measured as a change in the crash rate. This is a better measure than the raw count of crashes because it incorporates the amount of roadway and the amount of traffic at the site. This method can be easily adapted to computing rates within specific roadway types, such as: comparing various lane widths, comparing two lane to multi-lane, comparing major traffic arteries to local feeder routes, etc.

These data files are part of a system called HSIS (Highway Safety Information System) which was developed by the Highway Safety Research Center under contract from the Federal Highway Administration (FHWA). HSIS is the primary database used by FHWA for its roadway-related safety analysis. It incorporates data from a select group of states - Illinois, Maine, Michigan, Minnesota, and Utah.

BACKGROUND

Typically, roadway and crash information is stored in separate datasets. For the HSIS system, the roadway information is stored in a SAS dataset where every observation represents a "homogeneous" roadway section. Within a homogeneous section, the characteristics of the roadway are identical (i.e. the same type of surface, the same number of lanes, the same speed limit, the same width of shoulders, etc.). The lengths of the individual sections are different because they are based on actual field measurements. Data for the roadway can be linked to other HSIS databases by matching the location variables (typically ROUTE and MILEPOST). Each roadway section has a beginning milepost, an ending milepost, and a section length variable.

The crash information within HSIS is stored in a set of three SAS datasets: the accident portion, the vehicle portion, and the occupant portion. These three parts can

be linked by matching on the unique crash number and vehicle position number. The crash information can be linked to the other HSIS databases by matching the location variables found on the accident portion.

One frequent type of analysis using the HSIS system involves the calculation of crash rates. The researcher/engineer will often be interested in computing a rate for a specific group of crash types and/or for a specific type of location. After examining the results, a follow-up study will often look more closely at the "interesting" segments.

This type of analysis involves handling HSIS data in two types of orientations. Many of the issues in linking and processing the data are similar regardless of the HSIS state chosen.

Each separate database in HSIS has an inherent type of orientation. For example, the roadway database has one record for each homogeneous section. This is often referred to as SEGMENT ORIENTATION. The crash database is in ACCIDENT ORIENTATION because there is a unique record for each crash. The vehicle database is in VEHICLE ORIENTATION because there is a unique record for each separate vehicle involved in a crash.

There are several common units used to express crash rates. These include: annual crashes per million vehicle miles (for sections) and annual crashes per million entering vehicles (for intersections).

The formula to compute the first crash rate is:

$$R = \frac{(\text{count of crashes in "n" years}) * 1,000,000}{\text{Sum}(\text{AADT} * \text{SECT_LNG}) * (365.25 \text{ days per year}) * n}$$

where

- R is the crash rate per million vehicle miles
- AADT is the Annual Average Daily Traffic count of vehicles
- SECT_LNG is the length of a section in miles
- n is the number of years of crash data included in the study

The formula to compute the second crash rate is:

$$r = \frac{(\text{count of crashes in "n" years}) * 1,000,000}{\text{Sum (entering AADTs)} * (365.25 \text{ days per year}) * n}$$

where

- r is the crash rate per million entering vehicles
- AADT is the Annual Average Daily Traffic count of vehicles
- n is the number of years of crash data included in the study

TECHNIQUE

In general, the steps necessary to compute crash rates are the following:

- 1) Before starting, you MUST have both a beginning and an ending milepost variable on each segment. This implies that you already have, or can easily compute, the length of the section in miles. For most HSIS states, this is already true and you can immediately proceed to step 2.
- 2) Select roadway segments which satisfy your selection criteria. For example, to compute statewide rates select all segments. To compute rural rates on two lane sections eliminate all non-rural segments and eliminate all sections with more than two lanes.
- 3) If necessary, sort the segments from step 2 in location order. There should NEVER be any overlap of mileposts. It is OK to have gaps between sections. It is OK to have adjacent sections (where the ending milepost of one section matches the beginning milepost of the next section).
- 4) (Optional) Subset the accident database to only the crashes which satisfy your criteria for crashes. For example, one could select only night crashes or only wet road crashes.

This step becomes more complicated when your selection depends on information only found in either the vehicle or occupant database. For example, if you want to compute a crash rate for tractor trailer trucks or elderly drivers, you must create a temporary file containing the accident crash number and all the necessary variables for

possible crashes. Next, the temporary case number file must be sorted and any duplicates dropped. Finally, these cases must be matched to the master accident database to select only those cases of interest.

- 5) Modify the generalized SAS matching program which creates a SEGMENT ORIENTED file including tallies for various accident types. As input use the SEGMENT ORIENTED dataset from step 3 and the ACCIDENT ORIENTED dataset from step 4.

For computing crash rates, one only needs to count the accidents of interest on each segment and output a segment oriented file which includes these tallies. The resulting output should contain EXACTLY the same number of observations as the SEGMENT ORIENTED input file. It is possible that many sections will have no accidents, meaning that we looked for matching crashes on this section but none occurred.

A second dataset is also created containing the crash number and selected variables from both the roadway and the crash dataset for every crash which occurred on one of these sections and contributed to the tallies. This dataset will typically contain FEWER observations than the ACCIDENT ORIENTED input file.

- 6) Using the SEGMENT ORIENTED dataset from step 5 and the SAS SUMMARY procedure, compute accident rates. Typically you must first multiply the length of the section in miles by the average annual daily traffic count to compute the average daily vehicle miles on each section. Next compute the total count of accidents and the total vehicle miles within categories of interest using PROC SUMMARY. Finally, compute the crash rates and print the results. The typical units are crashes per million vehicle miles per year.

After the crash rates are examined, the researcher/engineer may ask for additional information about the crashes. Further analysis can be done using selected crashes from the ACCIDENT ORIENTED dataset from step 5. If the follow-up request uses the same variables as the rate calculation, then this dataset may be useful directly. If additional variables are needed, then you must decide whether to extract the missing information and attach it to the study file or whether to

backup and repeat the above steps including extra information about the crashes. The best choice will depend on many factors.

CONCLUSIONS

In conclusion, PROC SUMMARY is an efficient way to compute total crash counts and total vehicle miles driven. These values can then be substituted in the formula to compute highway crash rates. By using the appropriate roadway variables as SUMMARY CLASS variables, then several specific crash rates can be computed in the same step. This technique is applicable to older versions of SAS (for example SAS 6.04 for DOS). The newer versions of SAS provide newer tools (like SQL) for linking information from several databases.

Author Contact:

Elizabeth G. Hamilton
 UNC Highway Safety Research Center
 134 1/2 E. Franklin St, CB#3430
 Chapel Hill, NC 27599-3430
 Internet: UHSREGH@unc.oit.unc.edu

SAS is a registered trademark of SAS Institute, Inc., Cary, NC.

EXAMPLE SAS PROGRAM

```

/*          Computation of HIGHWAY CRASH RATES          */
/*
/* This program demonstrates the technique for computing crash
/* rates. It includes an example of matching crashes and section.
/*
/* This example uses the UTAH crash file and roadway file.
/*
/* ASSUMPTIONS:
/*   1 - The UTAH data files are unzipped and residing on D:\
/*   2 - The UTAH standard format library is stored in the
/*       LIBRARY located at D:\CDW\FORMATS\UTAH
/*
/* SAS STEPS USED:
/*   DATA STEP - to LINK the roadway and crash files and
/*                output two files:
/*                1-an accident oriented file called LINKACC
/*                2-a segment oriented file with counters
/*                called LINKSEG
/*   PROC PRINT - to inspect the intermediate data files
/*   PROC FORMAT - to establish ranges for continuous variables
/*                and also to define a RURAL/URBAN based on the
/*                categorical variable for functional class.
/*   DATA STEP - to eliminate those sections without medians
/*   PROC SUMMARY- to subtotal accident counters and veh-miles
/*                within various categories of segment types.
/*                These values will be used in computing the
/*                various crash rates in the next DATA Step.
/*   DATA STEP - Using the values from the SUMMARY above,
/*                compute various crash rates.
/*   PROC PRINT - to produce the default report of the results
/*   PROC TABULATE-to produce a more sophisticated report

```

```

LIBNAME LIBRARY 'D:\CDW\FORMATS\UTAH';
LIBNAME DD2 'D:\';

/* The DATA STEP code to link roadway and crash data and create two */
/* SAS datasets: */
/* 1- LINKSEG - roadway segments with added crash counters */
/* 2- LINKACC - selected variables from the matching crashes */
/* has been eliminated for brevity */

PROC PRINT DATA=LINKSEG (OBS=25);
  VAR RTNUM BEGMP ENDMP FUNC MEDTYP MEDWID AADT SECT_LNG
      TOT ABCK TOT FAT TOTCNT TOT_ALL ;
  TITLE1 'UTAH ROADWAY SEGMENTS AFTER MATCHING AND COUNTING CRASHES';
  TITLE2 '-- SEGMENT ORIENTATION --';

/* Group values of continuous CLASS variables using PROC FORMAT */

PROC FORMAT;

  VALUE MEDWIDF 0 = '00' /* Median Width */
              1-12 = '1 - 12'
              13-24 = '13-24'
              25-50 = '25-50'
              51-99 = '51+';

  VALUE AADTF 0 = 'UNKNOWN' /* Average Annual Daily Traffic */
             1-20000 = '1 - 20,000'
             20001 - 40000 = '20,001-40,000'
             40001 -999999 = '40,001+' ;

  VALUE $FUNCFJ '01','02','06','07','09','08' = 'RURAL'
               '11','12','14','16','17','19' = 'URBAN';

/* Compute VEH-MILES by MULTIPLYING EACH SECTION LENGTH BY TRAFFIC */

DATA DD2.TEMP;
  FORMAT VEHMI 12.3;
  SET LINKSEG;
  IF MEDWID = 0 THEN DELETE;
  IF MEDTYP = 'X' THEN DELETE;
  VEHMI=SECT_LNG * AADT;
  LABEL FUNC = 'FUNCTIONAL CLASS';
  FORMAT AADT AADTF. MEDWID MEDWIDF. MEDTYP $MEDTYPF.
         FUNC $FUNCFJ.;
RUN;

```

```
PROC SUMMARY DATA=DD2.TEMP NWAY;
    CLASS FUNC MEDTYP MEDWID AADT;
    VAR    TOT_ABCK TOT_FAT TOTCNT TOT_ALL VEHMI SECT_LNG;
    OUTPUT OUT=SUMM2 SUM=;
RUN;
```

```
PROC PRINT DATA=SUMM2(OBS=20);
TITLE 'UTAH - TOTAL ACCS AND TOTAL VEHICLE-MILES';
RUN;
```

```
DATA RATES2;
    SET SUMM2;
    INJ_RATE  = (TOTCNT    * 1000000) / (VEHMI * 365.25 * 2);
    FAT_RATE  = (TOT_FAT   * 1000000) / (VEHMI * 365.25 * 2);
    T_INJRTE  = (TOT_ABCK  * 1000000) / (VEHMI * 365.25 * 2);
    TOT_RATE  = (TOT_ALL   * 1000000) / (VEHMI * 365.25 * 2);
    DROP _TYPE_; RENAME _FREQ_ = NUM_SEGS; RENAME SECT_LNG=T_MILES;
RUN;
```

```
PROC PRINT DATA=RATES2(OBS=10);
    FORMAT AADT AADTFF.
           MEDWID MEDWIDF.
           MEDTYP $MEDTYPF.
           FUNC  $FUNCFJ.;
TITLE 'UTAH - COMPUTED ACCIDENT RATES';
RUN;
```

```
PROC TABULATE DATA=RATES2;
    CLASS FUNC MEDTYP MEDWID AADT;
    VAR    TOT_RATE TOT_ALL T_MILES;
    TABLES (FUNC * MEDTYP='MEDIAN TYPES'*
    (MEDWID='MEDIAN WIDTH' ALL='SUB-TOTALS') ALL='FINAL TOTALS'),
    AADT='AADT' *(TOT_RATE='ACCIDENT'*SUM='RATE'*F=12.6
    T_MILES='MILEAGE'*SUM='TOTAL'*F=9.3
    TOT_ALL='TOTAL'*SUM='ACCS'*F=5.0)
    /RTS=37 MISSTEXT='0';
    TITLE 'ACCIDENT CHARACTERISTICS OF ROADWAY SEGMENTS';
    TITLE2 'UTAH DATA 1988-89';
RUN;
```