SAS® GLOBAL FORUM 2020 #SASGF

Paper 4699-2020

Basic Concepts for Research Study Reporting

Julie Plano, Keli Sorrentino, Yale University

ABSTRACT

Data reports can be produced using SAS to convey the progress and potential needs of an ongoing study to research staff and investigators. It is critical this information is accurate and timely. Tailoring these reports to your team's needs allows for maximum communication. It is imperative to begin by knowing your data, the structure, variable formats, and meanings. What you report is highly dependent on the audience; this paper reviews basic concepts of report production. We cover the most requested types of information for current field studies: cumulative counts over time, frequencies, percentages, missing values, descriptive statistics, and plots. Presentation can be just as important as content. Creating visually appealing output can be done with a few extra lines of code. Including titles, formatting and system options as well as utilizing the SAS output delivery system (ODS) can improve appearance and clarity.

INTRODUCTION

Data reports are a requirement of any project. The progress of a research study needs to be reported regularly to the principle investigator, project coordinator, and the funding agency. There also may be a Data Safety Monitoring Board with different report needs. Building reports that tell the story of what has been done, what is scheduled to happen, and where the project is in relation to goals is critical. The first step to building quality reports is to know your data. What variables do you have? What variables need to be created? Use SAS to explore then determine what information will be used to create tables and graphs to best present data. Include some thought as to how the story will flow so the information is conveyed clearly, then use the software's many options to enhance your report.

Data used for examples in this paper are from the Framingham Heart Study, *Sashelp.heart*. These data are available to users and located in the Sashelp library. For demonstration purposes the data were adapted to include three additional variables (ID, Outcome and Out_Date) and 253 new observations (&HRT dataset). The new observations represent individuals who were not successfully enrolled and have missing heart data. Full code and a sample report are provided in the Appendix.

KNOW YOUR DATA

STRUCTURE

Become familiar with the data. There will be enrollment statistics from the system used to track subjects as they move through the study from inquiry to completion as well as data from questionnaires. There may also be data regarding samples collected. The programmer needs to know what is available, how the data is formatted, and how to link information together.

Importing the raw data to SAS and running a PROC CONTENTS should always be the first step. Saving a copy of the results (.docx or .pdf) is important to refer to names and formats. This can be time-saving if datasets will be linked for the report and can be critical later if the data collection processes are still developing. If the report depends on data from different sources, then import individual datasets and check the contents of each. Compare like variables to determine if the types and lengths match.

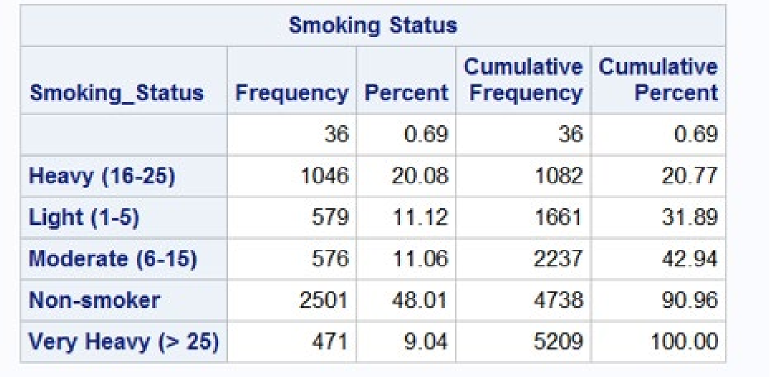
CREATING VARIABLES FOR ANALYSES

An input program can be used to assign labels, change type, keep or drop variables, build or format dates, rename and create new variables as needed. At this point, determine if you have missing data or values that are inconsistent/out of range. Example code for several of the topics discussed below is provided at the end of the section.

Handling missing values

Using options offered in select procedures can help to determine what variables include problem values and how to clean or report them. In the FREQUENCY procedure, using the options of LIST and MISSING will include these values as part of the list. Once data is displayed in the output window the user can subset or stratify the dataset for further investigation or cleaning. Errors in data entry and patterns in coding can be identified at this step, before a report is finalized and shared. Use PROC FREQ to produce Table 1:

```
   proc freq data = sashelp.heart ; tables Smoking_Status/list missing; run;
```



Smoking Status

Smoking_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	36	0.69	36	0.69
Heavy (16-25)	1046	20.08	1082	20.77
Light (1-5)	579	11.12	1661	31.89
Moderate (6-15)	576	11.06	2237	42.94
Non-smoker	2501	48.01	4738	90.96
Very Heavy (> 25)	471	9.04	5209	100.00

Table 1 - Checking for Missing Values

Table 1 shows there are 36 missing values for smoking status in the set. It should be confirmed if the values are truly missing or if there is a data entry error that can be corrected. For confirmed missing values, in lieu of leaving the table with a blank row the values should be coded in the program with an appropriate label such as ‘Not Reported’.

Inconsistent values

Examine the accuracy of similar variables or those which can be cross-checked. If one variable is dependent on the other or should line up exactly with another. For example, in sashelp.heart, the variable STATUS is coded as ‘Dead’ or ‘Alive’. There is also a variable for age at death. If we cross-check these two the expectation is that all observations with a status of ‘Dead’ have data for age at death and those with a status of ‘Alive’ are blank for age at death. In this dataset the expectation is confirmed (data not shown) by running:

```
   proc freq data = sashelp.heart; tables status*ageatdeath/list missing; run;
```

Create new variables/Change variable type

Variables can be coded in many ways, all of which can ultimately have the same meaning. However, depending on what you are going to do with the variable (frequencies, graph or modeling) it may be useful to have it in different formats. In the code shown at the end of this section, the character variable SEX was used to create three numeric variables – Male, Female and Gender. In the sashelp.heart dataset, CHOL_STATUS, WEIGHT_STATUS and SMOKING_STATUS are examples of categorical variables that were created from continuous variables in the set (CHOLESTEROL, HEIGHT, WEIGHT, and SMOKING).

Dates

Determine how the dates are formatted in your raw data. Dates should be changed to the format that makes them the most useful for the tasks required by the report. It may also be helpful to create a variable for the week, month or year that a specific event took place. Calculating the time between two dates may also be useful.

The input program code below provides examples discussed in this section:

```
DATA &RPT; /*SEE DOCUMENTATION SECTION FOR NAMING REFERENCE*/
SET &HRT;
format out_date mmddyy8.;  /*FORMAT OUT_DATE TO 01/01/2020*/

year = year(out_date);   /*CREATE A VARIABLE OF ONLY YEAR*/

if outcome = 1 then do;
if smoking_status = '' then do smoking_status = 'Not Reported';end; end;
   /*SUBSET THIS CHANGE TO APPLY ONLY TO ENROLLED INDIVIDUALS*/

if outcome = . then outcome = 0; /*COMBINED WITH PROC FORMAT TO LABEL
                                   MISSINGS 'PENDING'*/

Male = .; /*THE REMAINING LINES CREATE THE NUMERICAL VARIABLES MALE,
            FEMALE AND GENDER*/
Female = .;
Gender =. ;

if sex = 'Male' then do Male =1; end;
if sex = 'Female' then do Female =1; end;

if sex = 'Male' then do Gender =1; end;
if sex = 'Female' then do Gender =2; end;
run;
```

DOCUMENTATION

A critical aspect of tracking and reporting research is setting up protocols for storing and documenting how data is handled throughout the project. Reports should be produced on a regular schedule. There may be a weekly report which provides a look at the current activity of the project and a monthly report that shows cumulative information. Best practice includes storing dated copies of all datasets used to develop a report. Each time a report is made, copy, date, and store the base data prior to moving forward with running any programs. Part of the input program should be to save a dated copy of the SAS dataset with the new variables needed:

```
LIBNAME GF  'W:/Global Forum';

%LET HRT = GF.HEART_GF_011020;/*HRT IS THE DATASET ADAPTED FROM SASHELP.HEART*/

%LET RPT = GF.REPORT_GF_011020;/*RPT IS THE REPORT READY DATASET*/
```

Our preference for report output is .rtf or .pdf (a full list can be found in the SAS ODS documentation). A dated copy of one or both file types is important for historical purposes. As a research project progresses, outcomes and other coding used for reporting may change. The ability to go back in time to review reports and the exact datasets they were created from is crucial. Changes to report programming should be commented in the program.

A procedures manual of report processes should be kept up to date. If the primary programmer is absent it is vital that another programmer can step in to duplicate a report when needed. Keeping a log of any changes to variables in the raw data is important. Every time a change is made to a value it should be recorded in a change log. This can be vital in piecing together how the numbers may have shifted over time.

INCLUDED IN THE REPORT

When creating a report utilize the study team to know what to include. Listen to your colleagues and the questions they are asking. For example, the all-important (and constantly asked!) questions of “How many?” and “What percent?”.

SAS procedures offer a multitude of ways to accomplish a single task. Often a programmer writing a report will try out several procedures before settling on what will accurately and clearly convey the information. The following examples refer to the exploration phase and suggest options for presenting data in a report to an audience.

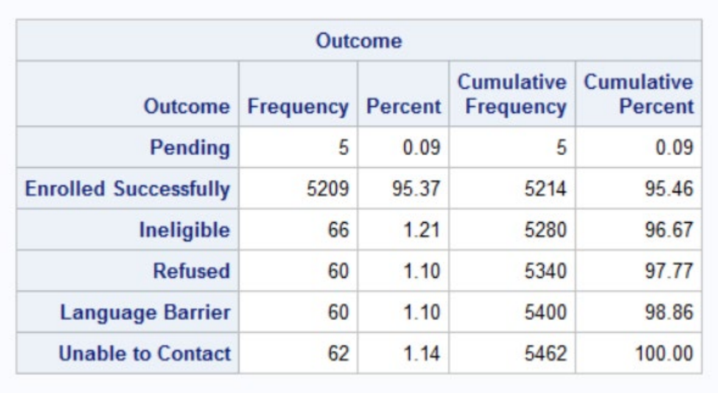
CUMULATIVE COUNTS

Explore

There are several procedures (FREQ, SQL, TABULATE, REPORT and MEANS) that can sum and list counts over time. Using the dataset &HRT, we will show some enrollment statistics for the study.

The FREQUENCY procedure (PROC FREQ) can easily tackle the questions of variable distribution across a population and present the information in a couple of ways to keep everyone satisfied. The counts can be of one variable, multiple variables or cross-tabulated. Table 2 shows the enrollment outcomes for all the potential study participants. To make the table presentation-ready PROC FORMAT is used to change numerical values into easy to understand text (See full code in Appendix):

```
proc freq data = &HRT; tables outcome;
format outcome OUT.; run;
```

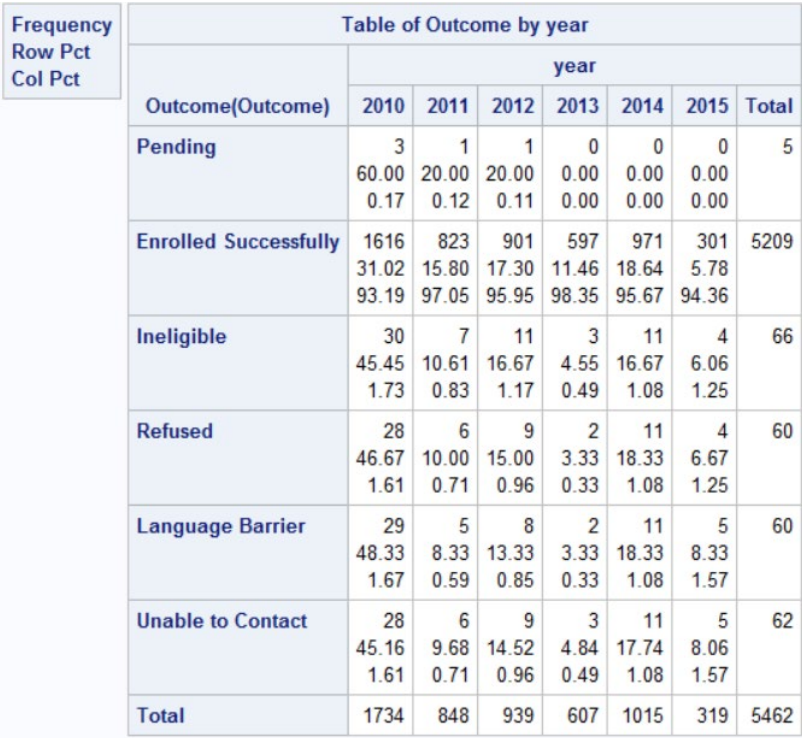


Outcome				
Outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Pending	5	0.09	5	0.09
Enrolled Successfully	5209	95.37	5214	95.46
Ineligible	66	1.21	5280	96.67
Refused	60	1.10	5340	97.77
Language Barrier	60	1.10	5400	98.86
Unable to Contact	62	1.14	5462	100.00

Table 2. Enrollment Outcomes for All Potential Study Participants

Showing enrollment rates over time can be important as well. PROC FREQ can be used to cross-tabulate enrollment by year:

```
proc freq data = &RPT;
tables outcome*year /nopercent;
format outcome OUT.;
 run;
```



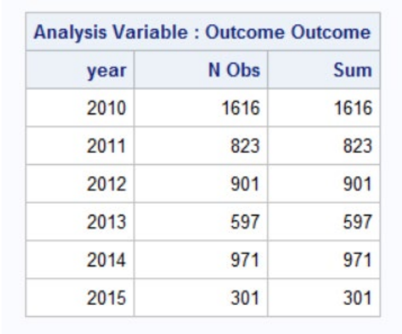
Frequency
Row Pct
Col Pct

Table of Outcome by year							
	year						
Outcome(Outcome)	2010	2011	2012	2013	2014	2015	Total
Pending	3 60.00 0.17	1 20.00 0.12	1 20.00 0.11	0 0.00 0.00	0 0.00 0.00	0 0.00 0.00	5
Enrolled Successfully	1616 31.02 93.19	823 15.80 97.05	901 17.30 95.95	597 11.46 98.35	971 18.64 95.67	301 5.78 94.36	5209
Ineligible	30 45.45 1.73	7 10.61 0.83	11 16.67 1.17	3 4.55 0.49	11 16.67 1.08	4 6.06 1.25	66
Refused	28 46.67 1.61	6 10.00 0.71	9 15.00 0.96	2 3.33 0.33	11 18.33 1.08	4 6.67 1.25	60
Language Barrier	29 48.33 1.67	5 8.33 0.59	8 13.33 0.85	2 3.33 0.33	11 18.33 1.08	5 8.33 1.57	60
Unable to Contact	28 45.16 1.61	6 9.68 0.71	9 14.52 0.96	3 4.84 0.49	11 17.74 1.08	5 8.06 1.57	62
Total	1734	848	939	607	1015	319	5462

Table 3 - Enrollment Outcomes by Year

Table 3 is cumbersome however. Limiting the data to only successfully enrolled cases will help. PROC MEANS is commonly used to calculate descriptive statistics. The NWAY options will place only the observations with the highest _type_ value. MAXDEC =0 will round to the whole number. CLASS specifies the classification variable. VAR indicates which variable should be analyzed. The output statement will create a dataset which will be helpful later for including this data in a report. Using AUTONAME will automatically name the created variable(s) and store them in the output dataset:

```
proc means data = rpt2 sum nway maxdec=0;
class year;
var outcome;
output out = enrolled sum = /autoname;
run;
```



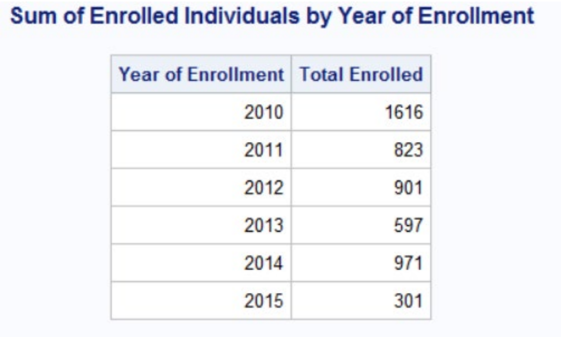
Analysis Variable : Outcome Outcome		
year	N Obs	Sum
2010	1616	1616
2011	823	823
2012	901	901
2013	597	597
2014	971	971
2015	301	301

Table 4 - Total Enrolled Individuals by Year

Presentation

Table 2 is good just the way it is. Sometimes simple is best and there is no need to change the output. Table 4 does show the information we are looking for, but the labels and the seeming duplication of data will likely be confusing to the audience. By adding a few lines of code, it will be report worthy. Include a descriptive title, remove the observation count and label variables with full text words for clarity. Table 5 is simple and to the point:

```
Title 'Sum of Enrolled Individuals by Year of Enrollment';
proc print data = enrolled noobs label;
var year outcome_sum;
label outcome_sum = 'Total Enrolled';
label year = 'Year of Enrollment';
run;
```



Sum of Enrolled Individuals by Year of Enrollment

Year of Enrollment	Total Enrolled
2010	1616
2011	823
2012	901
2013	597
2014	971
2015	301

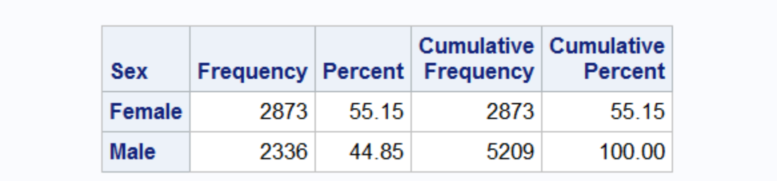
Table 5 - Simple Table for Enrolled Individuals by Year

DESCRIPTIVE STATISTICS

Explore

PROC FREQ returns a table or list of frequency counts and percentages. Let’s look at the data &RPT and the distribution of sex and smoking status. First, determine the frequency of gender in the enrolled population:

```
proc freq data = &RPT; tables sex/list missing; run;
```



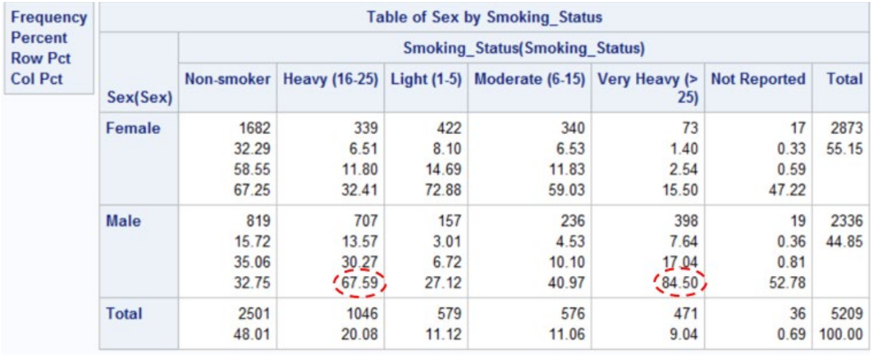
Sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	2873	55.15	2873	55.15
Male	2336	44.85	5209	100.00

Table 6 - Gender Distribution in Enrolled Population

The distribution of males and females in the population is 45% and 55%, respectively, and there is no missing data on sex (Table 6).

Investigating the relationship between two variables by placing an asterisk (*) between them can give us a look at their association. The code includes the option ORDER = FREQ which will print the values by descending count of the total. Results are shown in Table 7:

```
proc freq data = &RPT order=freq; tables Smoking_Status*sex; run;
```



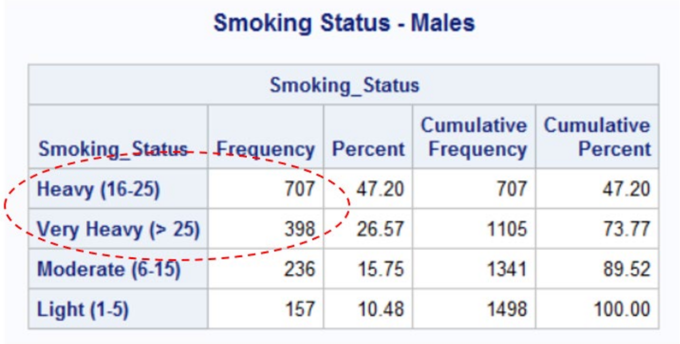
Frequency Percent Row Pct Col Pct	Table of Sex by Smoking_Status						
	Smoking_Status(Smoking_Status)						
Sex(Sex)	Non-smoker	Heavy (16-25)	Light (1-5)	Moderate (6-15)	Very Heavy (> 25)	Not Reported	Total
Female	1682 32.29 58.55 67.25	339 6.51 11.80 32.41	422 8.10 14.69 72.88	340 6.53 11.83 59.03	73 1.40 2.54 15.50	17 0.33 0.59 47.22	2873 55.15
Male	819 15.72 35.06 32.75	707 13.57 30.27 67.59	157 3.01 6.72 27.12	236 4.53 10.10 40.97	398 7.64 17.04 84.50	19 0.36 0.81 52.78	2336 44.85
Total	2501 48.01	1046 20.08	579 11.12	576 11.06	471 9.04	36 0.69	5209 100.00

Table 7 - Smoking Status by Gender

Presentation

Table 7 indicates a high percentage of males in the categories of heavy to very heavy smoking status. This information is somewhat buried in the table though. We could present it more clearly by removing the 'non-smoker' and 'not reported' groups and dividing the tables by gender. Use ODS to output a .rtf file and include options NOPROCTITLE to suppress the procedure title and style to enhance the appearance (full code in Appendix):

```
proc freq data = &RPT order = freq; tables smoking_status ;
where Smoking_Status not in ('Not Reported', 'Non-smoker') and gender=1; run;
```



Smoking Status - Males

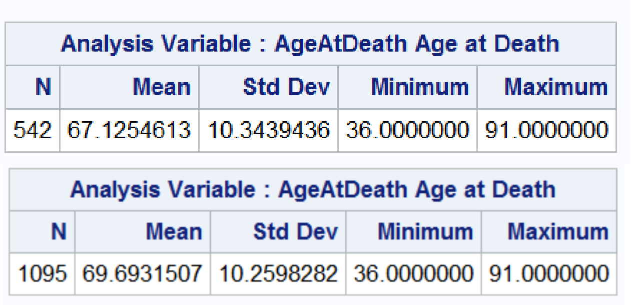
Smoking_Status				
Smoking_Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Heavy (16-25)	707	47.20	707	47.20
Very Heavy (> 25)	398	26.57	1105	73.77
Moderate (6-15)	236	15.75	1341	89.52
Light (1-5)	157	10.48	1498	100.00

Table 8 - Smoking Status of Enrolled Males

Explore

To drill down further and investigate the group of males with heavy to very heavy smoking status, age at death is of interest. This frequency output listing is miserably long as it is inclusive of ages 36 to 91 by single years. A better way is to use the MEANS procedure.

When the variable of interest is continuous PROC MEANS or PROC UNIVARIATE can be used to look at the distribution. Note the MEANS procedure excludes class variables with missing values.



Analysis Variable : AgeAtDeath Age at Death				
N	Mean	Std Dev	Minimum	Maximum
542	67.1254613	10.3439436	36.0000000	91.0000000

Analysis Variable : AgeAtDeath Age at Death				
N	Mean	Std Dev	Minimum	Maximum
1095	69.6931507	10.2598282	36.0000000	91.0000000

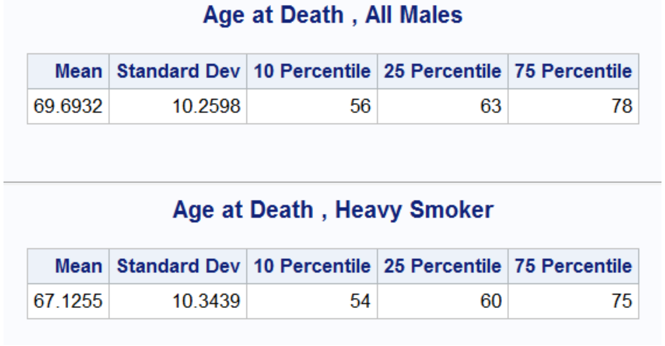
Table 9 - Subset of Heavy to Very Heavy Smokers (top) & All Males in Study (bottom)

This PROC MEANS reports a mean age at death of 67 for the heavy to very heavy smoker group while the mean age is almost 70 for all males in the study (Table 9).

Presentation

Adding an output statement in your PROC MEANS will produce a dataset including basic statistics for the variable in the var statement, the dataset is named in the OUT=option. Additional options for inclusion are median and percentiles. When printing the output dataset for presentation include the MAXDEC=0 in the PROC MEANS for the output dataset variables to be rounded to a whole number (Table 10):

```
proc print data =subsetmeansSMK noobs label; var AgeatDeath_Mean
AgeatDeath_StdDev AgeAtDeath_P10 AgeatDeath_P25 AgeatDeath_P75;
label AgeatDeath_Mean= 'Mean';
label AgeatDeath_StdDev = 'Standard Dev';
label AgeAtDeath_P10 = '10 Percentile';
label AgeatDeath_P25 = '25 Percentile';
label AgeatDeath_P75 = '75 Percentile';
run;
```



Age at Death , All Males

Mean	Standard Dev	10 Percentile	25 Percentile	75 Percentile
69.6932	10.2598	56	63	78

Age at Death , Heavy Smoker

Mean	Standard Dev	10 Percentile	25 Percentile	75 Percentile
67.1255	10.3439	54	60	75

Table 10 - Improved Presentation of PROC MEANS

PLOTS

Explore

The next question: what caused their death? To visually communicate the answer, we create a chart using PROC SGPLOT. With simple syntax, graphs are created and show counts by the selected category. In Figure 1 we can see for each gender which causes of death occurred most often:

```
proc sgplot data=&RPT;
vbar deathcause;
by Gender; run;
```

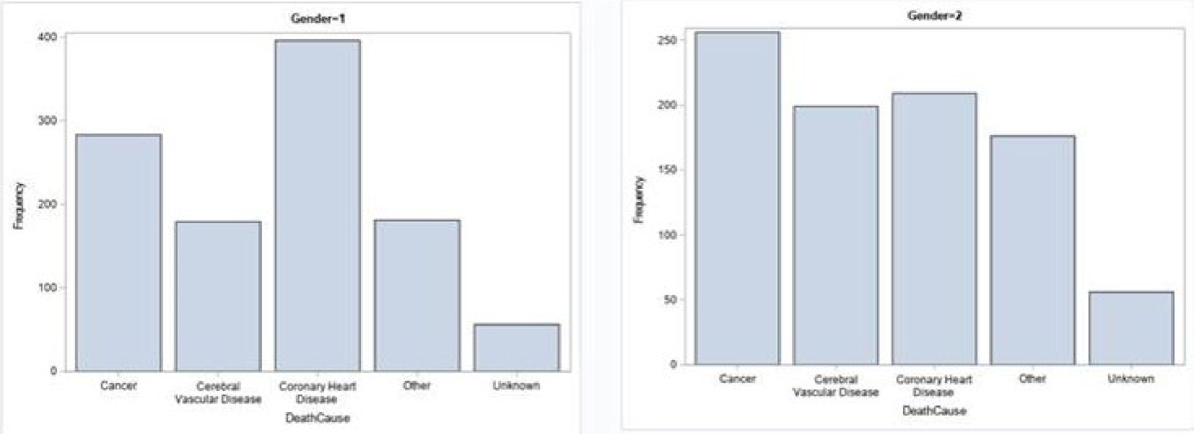


Figure 1 - PROC SGPLOT by gender

Presentation

To produce graphics ready for presentation, overlay the two bar charts and add some options to enhance the visualization. To do this, MALE and FEMALE need to be independent variables (see input program on p.3). Include two VBAR statements, vertical bars representing deathcause, response variables by gender. NOSTATLABEL removes the statistic from the legend (i.e. “sum”). The STYLESTTRS statement allows the user to customize graph elements. DATACOLORS specifies the color name for each group. A great reference for color names can be found here: https://support.sas.com/content/dam/SAS/support/en/books/pro-template-made-easy-a-guide-for-sas-users/62007_Appendix.pdf . The XAXIS AND YAXIS specify the display of each axis (tick marks, labels, font size and color):

```
Title ‘Cause of Death by Sex’;
proc sgplot data = &RPT;
styleattrs DATACOLORS=(BigB STPK);
vbar deathcause / response=Male  nostatlabel transparency=0.25 ;
vbar deathcause/ response=Female nostatlabel barwidth=0.5 nooutline;
xaxis display= (nolabel) valueattrs=(size=6 color=Black);
yaxis grid display=(noline noticks nolabel) valueattrs=(size=8 color=Black);
run;
```

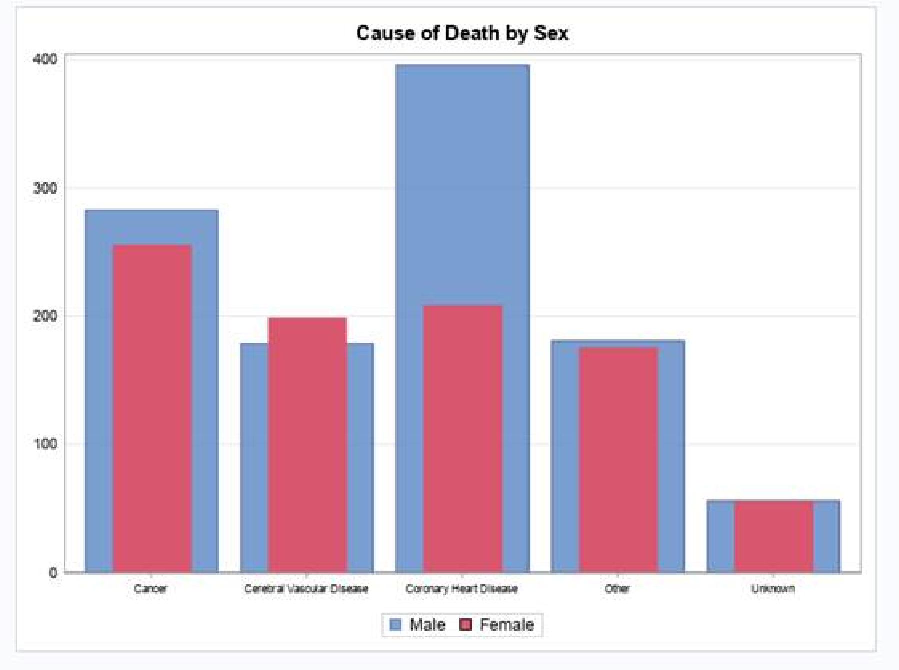


Figure 2 - Improved PROC SGPLOT

ORGANIZATION OF TABLES AND GRAPHS

The organization of tables and graphs can be planned prior to writing a report program. It will often change as the data is processed and output is viewed. Tell the story from start to finish. With research data this may start by describing inquiries to the study including how many, type of contact and/or location within the recruitment area. From there, continue to drill through the layers of a subject’s progress through each study phase. At the point of enrollment, additional statistics such as demographics can be useful. Add descriptive tables to show study focus outcomes clearly.

CONCLUSION

It is important for programmers working in research to deliver reports that are concise and easy to follow. Producing these reports does not require fancy programming. Base procedures like PROC FREQ, PROC PRINT and PROC MEANS can provide most of the tables necessary to tell the story of the study data. Keep tables uncomplicated so the reader can easily absorb what is being portrayed. Often two simple tables are better than one complex one. Plots and charts can be helpful when a graphical representation will better convey the information. Use of additional output options in .rtf or .pdf include adding text, adjusting layout, changing fonts and adding colors will improve the final product. The options are endless, and many papers exist showing how to make a report even more pleasing.

ACKNOWLEDGMENTS

We would like to acknowledge our colleagues and friends at the Center for Perinatal, Pediatric & Environmental Epidemiology.

CONTACT INFORMATION

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

Julie Plano
Yale School of Public Health
Julie.colburn@yale.edu

Keli Sorrentino
Yale School of Public Health
Keli.sorrentino@yale.edu

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

APPENDIX

```
LIBNAME GF  'W:/Global Forum';

%LET HRT = GF.HEART_GF_011020;  /*HRT IS THE DATASET ADAPTED FROM SASHELP.HEART*/
%LET RPT = GF.REPORT_GF_011020;  /*RPT IS THE REPORT READY DATASET*/

proc contents data = &hrt; run;


/*INPUT PROGRAMMING*/
/*DEVELOP DATASET FOR RUNNING REPORT*/

DATA &RPT; /*SEE DOCUMENTATION SECTION FOR NAMING REFERENCE*/
SET &HRT;
format out_date mmddyy8.;  /*FORMAT OUT_DATE TO 01/01/2020*/

year = year(out_date);   /*CREATE A VARIABLE OF ONLY YEAR*/

if outcome = 1 then do;
if smoking_status = '' then do smoking_status = 'Not Reported';end; end; /*SUBSET
THIS CHANGE TO APPLY ONLY TO ENROLLED INDIVIDUALS*/

if outcome = . then outcome = 0; /*COMBINED WITH PROC FORMAT TO LABEL MISSINGS
'PENDING'*/

Male = .; /*THE REMAINING LINES CREATE THE NUMERICAL VARIABLES MALE, FEMALE AND
GENDER*/
Female = .;
Gender =. ;

if sex = 'Male' then do Male =1; end;
if sex = 'Female' then do Female =1; end;

if sex = 'Male' then do Gender =1; end;
if sex = 'Female' then do Gender =2; end;
run;

proc contents data = &rpt; run;

/*FORMAT NUMERIC VALUES SO THEY WILL HAVE CLEAR LABELS IN TABLES FOR EASE OF READING*/

proc format ;
value $Smk
'Very Heavy (> 25)' =1
'Heavy (16-25)'= 2
'Moderate (6-15)'= 3
'Light (1-5)' = 4 ; run;

proc format;
value out
      1 = 'Enrolled Successfully'
      2 = 'Ineligible'
      3 = 'Refused'
      4 = 'Language Barrier'
      5 = 'Unable to Contact'
      0 = 'Pending';
run;
```

```
ods  rtf file = "W:\Global Forum\GF_FinalReport_012120.rtf"
      style=styles.mimictitle startpage = no;
ods graphics on /noborder; ods noproctitle; ods escapechar = '^';

TITLE 'Research Data Report 1/21/20';

ods text = ' ';  /*INSERTS A BLANK LINE. THIS IS AN EASY WAY TO CONTROL SPACING.*/
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Enrollment Outcomes - All
Potential Study Participants'; /*TITLE OF THE OUTPUT WILL BE CENTERED WITH A
SPECIFIED FONT SIZE*/

/****FIRST LOOK AT SOME ENROLLMENT STATISTICS FOR THE ENTIRE POPULATION*****/
proc freq data = &RPT; tables outcome;
format outcome OUT.;
run;


/**REDUCE DATA DOWN TO ONLY CASES WITH A SUCCESSFUL ENROLLMENT IN THE STUDY**/
data rpt2;
set &rpt;
if outcome ne 1 then delete;
run;


proc means data = rpt2 sum nway maxdec=0 noprint;
class year;
var outcome;
output out = enrolled sum = /autoname;
run;


ods text = ' ';
ods text = ' ';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Sum of Enrolled Individuals
by Year of Enrollment';
proc print data = enrolled noobs label;
var year outcome_sum;
label outcome_sum = 'Total Enrolled';
label year = 'Year of Enrollment';
run;


ods text = ' ';
ods text = ' ';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Gender of Enrolled
Individuals';
proc freq data = RPT2 ; tables sex/list missing; run;

ods startpage = now;

Title 'Smoking Status';
ods text = '^S={width = 100% just = c fontsize = 12pt}Smoking Status of Enrolled
Individuals';

proc freq data = RPT2; tables smoking_status/ list missing; run;
```

```
ods text = '^S={width = 100% just = c fontsize = 12pt}Smoking Status by Gender';
proc freq data = RPT2 order=freq; tables Smoking_Status*sex/missprint; run;

ods startpage = now;

/*****REPORT TABLES FOR MALES ENROLLED IN STUDY***/

Title 'Smoking Status - Males';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Smoking Status for
Enrolled Males';
ods text = '^S={width = 100% just = c fontsize = 9pt}(Excludes Non-Smokers
and Unreported Smoking Status Cases)';
proc freq data = RPT2 order = freq; tables smoking_status ;
where Smoking_Status  not in ('Not Reported', 'Non-smoker') and gender=1 ;
run;

proc means data = RPT2 maxdec=0 noprint; Where sex = 'Male'; var ageatdeath;
output out=subsetmeansmale
        mean=
        median=
        std=
        min=
        max=
        p10=
        p25=
        p75=
        p90=
        / autoname; run;
proc means data = RPT2 maxdec=0 noprint; Where sex = 'Male'
and Smoking_Status in ('Heavy (16-25)','Very Heavy (> 25)') ; var ageatdeath;
output out=subsetmeansSMKmale
        mean=
        median=
        std=
        min=
        max=
        p10=
        p25=
        p75=
        p90=
        /autoname; run;

ods text = ' ';
ods text = ' ';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Mean Age of Death –
All Enrolled Males';
proc print data =subsetmeansmale noobs label; var AgeatDeath_Mean
AgeatDeath_StdDev AgeAtDeath_P10 AgeatDeath_P25 AgeatDeath_P75;
label AgeatDeath_Mean= 'Mean';
label AgeatDeath_StdDev = 'Standard Dev';
label AgeAtDeath_P10 = '10 Percentile';
label AgeatDeath_P25 = '25 Percentile';
label AgeatDeath_P75 = '75 Percentile';
run;
```

```
ods text = ' ';
ods text = ' ';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Mean Age of Death –
Enrolled Males who Reported Heavy to Very Heavy Smoking';
proc print data =subsetmeansSMKmale noobs label; var AgeatDeath_Mean
AgeatDeath_StdDev AgeAtDeath_P10 AgeatDeath_P25 AgeatDeath_P75;
label AgeatDeath_Mean= 'Mean';
label AgeatDeath_StdDev = 'Standard Dev';
label AgeAtDeath_P10 = '10 Percentile';
label AgeatDeath_P25 = '25 Percentile';
label AgeatDeath_P75 = '75 Percentile';
run;

 ods startpage = now;

/*****REPORT TABLES FOR FEMALES ENROLLED IN STUDY***/

Title 'Smoking Status - Females';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Smoking Status for Enrolled
Females';
ods text = '^S={width = 100% just = c fontsize = 9pt}(Excludes Non-Smokers and
Unreported Smoking Status Cases)';
proc freq data = RPT2 order = freq; tables smoking_status ;
where Smoking_Status  not in ('Not Reported', 'Non-smoker') and gender=2 ; run;


proc means data = RPT2 maxdec=0 noprint; Where sex = 'Female'; var ageatdeath;
output out=subsetmeansfemale
        mean=
        median=
        std=
        min=
        max=
        p10=
        p25=
        p75=
        p90=
        / autoname; run;
proc means data = RPT2 maxdec=0 noprint; Where sex = 'Female'
and Smoking_Status in ('Heavy (16-25)','Very Heavy (> 25)') ;
var ageatdeath;
output out=subsetmeansSMKfemale
        mean=
        median=
        std=
        min=
        max=
        p10=
        p25=
        p75=
        p90=
        /autoname; run;
```

```
ods text = ' ';
ods text = ' ';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Mean Age of Death –
All Enrolled Females';
proc print data =subsetmeansfemale noobs label; var AgeatDeath_Mean
AgeatDeath_StdDev AgeAtDeath_P10 AgeatDeath_P25 AgeatDeath_P75;
label AgeatDeath_Mean= 'Mean';
label AgeatDeath_StdDev = 'Standard Dev';
label AgeAtDeath_P10 = '10 Percentile';
label AgeatDeath_P25 = '25 Percentile';
label AgeatDeath_P75 = '75 Percentile';
run;

ods text = ' ';
ods text = ' ';
ods text = ' ';
ods text = '^S={width = 100% just = c fontsize = 12pt}Mean Age of Death –
Enrolled Females who Reported Heavy to Very Heavy Smoking';
proc print data =subsetmeansSMKfemale noobs label; var AgeatDeath_Mean
AgeatDeath_StdDev AgeAtDeath_P10 AgeatDeath_P25 AgeatDeath_P75;
label AgeatDeath_Mean= 'Mean';
label AgeatDeath_StdDev = 'Standard Dev';
label AgeAtDeath_P10 = '10 Percentile';
label AgeatDeath_P25 = '25 Percentile';
label AgeatDeath_P75 = '75 Percentile';
run;

ods startpage = now;

/*****PLOT FOR BOTH GENDERS****/
proc sort data = &RPT; by gender ;run;

Title ' ';
ods text = ' ';
ods text = ' ';
ods text = ' ';
Title 'Cause of Death by Sex';
proc sgplot data = &RPT;
styleattrs DATACOLORS=(BigB STPK);  /* Specify bar colors */
vbar deathcause / response=Male nostatlabel transparency=0.25 ;     /* Male Bars */
vbar deathcause/ response = Female nostatlabel barwidth=0.5 nooutline  ;  /*Female
Bars*/
  xaxis display= (nolabel)  valueattrs=(size=6 color=Black); /* Xaxis - nolabel */
/*to be able to read all categories change font size*/
  yaxis grid display=(noline noticks nolabel) valueattrs=(size=8 color=Black);  /*
Yaxis*/
  run;


ods rtf close;
```

Research Data Report 1/21/20

Enrollment Outcomes - All Potential Study Participants

Outcome				
Outcome	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Pending	5	0.09	5	0.09
Enrolled Successfully	5209	95.37	5214	95.46
Ineligible	66	1.21	5280	96.67
Refused	60	1.10	5340	97.77
Language Barrier	60	1.10	5400	98.86
Unable to Contact	62	1.14	5462	100.00

Sum of Enrolled Individuals by Year of Enrollment

Year of Enrollment	Total Enrolled
2010	1616
2011	823
2012	901
2013	597
2014	971
2015	301

Gender of Enrolled Individuals

Sex				
Sex	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Female	2873	55.15	2873	55.15
Male	2336	44.85	5209	100.00

Smoking Status

Smoking Status of Enrolled Individuals

Smoking_Status				
Smoking_Status	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Heavy (16-25)	1046	20.08	1046	20.08
Light (1-5)	579	11.12	1625	31.20
Moderate (6-15)	576	11.06	2201	42.25
Non-smoker	2501	48.01	4702	90.27
Not Reported	36	0.69	4738	90.96
Very Heavy (> 25)	471	9.04	5209	100.00

Smoking Status by Gender

Table of Smoking_Status by Sex			
Smoking_Status(Smoking_Status)	**Sex(Sex)**		
Frequency **Percent** **Row Pct** **Col Pct**	**Female**	**Male**	**Total**
Non-smoker	1682 32.29 67.25 58.55	819 15.72 32.75 35.06	2501 48.01
Heavy (16-25)	339 6.51 32.41 11.80	707 13.57 67.59 30.27	1046 20.08
Light (1-5)	422 8.10 72.88 14.69	157 3.01 27.12 6.72	579 11.12
Moderate (6-15)	340 6.53 59.03 11.83	236 4.53 40.97 10.10	576 11.06
Very Heavy (> 25)	73 1.40 15.50 2.54	398 7.64 84.50 17.04	471 9.04
Not Reported	17 0.33 47.22 0.59	19 0.36 52.78 0.81	36 0.69
Total	2873 55.15	2336 44.85	5209 100.00

Smoking Status - Males

Smoking Status for Enrolled Males
(Excludes Non-Smokers and Unreported Smoking Status Cases)

Smoking_Status				
Smoking_Status	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Heavy (16-25)	707	47.20	707	47.20
Very Heavy (> 25)	398	26.57	1105	73.77
Moderate (6-15)	236	15.75	1341	89.52
Light (1-5)	157	10.48	1498	100.00

Mean Age of Death - All Enrolled Males

Mean	Standard Dev	10 Percentile	25 Percentile	75 Percentile
69.6932	10.2598	56	63	78

Mean Age of Death - Enrolled Males who Reported Heavy to Very Heavy Smoking

Mean	Standard Dev	10 Percentile	25 Percentile	75 Percentile
67.1255	10.3439	54	60	75

Smoking Status - Females

Smoking Status for Enrolled Females

(Excludes Non-Smokers and Unreported Smoking Status Cases)

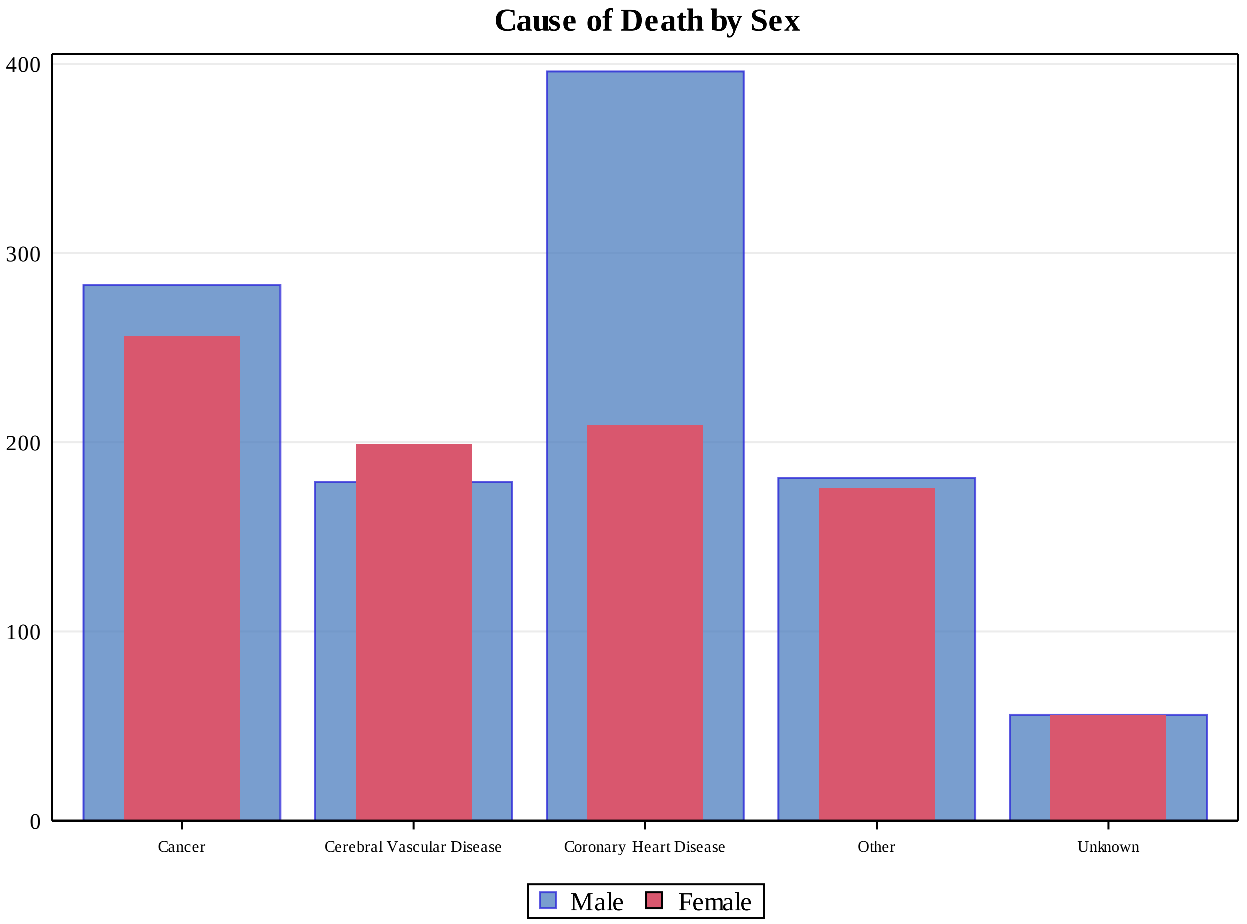
Smoking_Status				
Smoking_Status	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Light (1-5)	422	35.95	422	35.95
Moderate (6-15)	340	28.96	762	64.91
Heavy (16-25)	339	28.88	1101	93.78
Very Heavy (> 25)	73	6.22	1174	100.00

Mean Age of Death - All Enrolled Females

Mean	Standard Dev	10 Percentile	25 Percentile	75 Percentile
71.5670	10.8313	56	64	80

Mean Age of Death - Enrolled Females who Reported Heavy to Very Heavy Smoking

Mean	Standard Dev	10 Percentile	25 Percentile	75 Percentile
67.0652	11.5854	51	58	75


Cause of Death by Sex
400
300
200
100
0
Cancer
Cerebral Vascular Disease
Coronary Heart Disease
Other
Unknown
Male
Female