**Homework 9**

1. Refer to DOGS1 data. The investigator may want to ensure that the dogs allocated to each treatment group were of similar compositions with respect to gender and hair coat. Use PROC FREQ to conduct Fisher’s exact test to see if the concentration of the drug received was statistically independent of the gender of the dog. Likewise, see if the length of the coat and the drug treatment were statistically independent with Fisher’s exact test. Write your interpretation of the results of these tests.

**CODE**

**data** dogs1;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\data\dogs1.txt' firstobs = **2**;

input Dog $ Conc **16** Sex $ **17** Age **31**-**32** Haircoat $ @**32** Weight **43**-**48**;

**run**;

**proc** **sort** data=dogs1;

by dog;

**proc** **freq** data=dogs1;

title 'Concentration vs Gender';

tables conc\*sex/exact;

**run**;

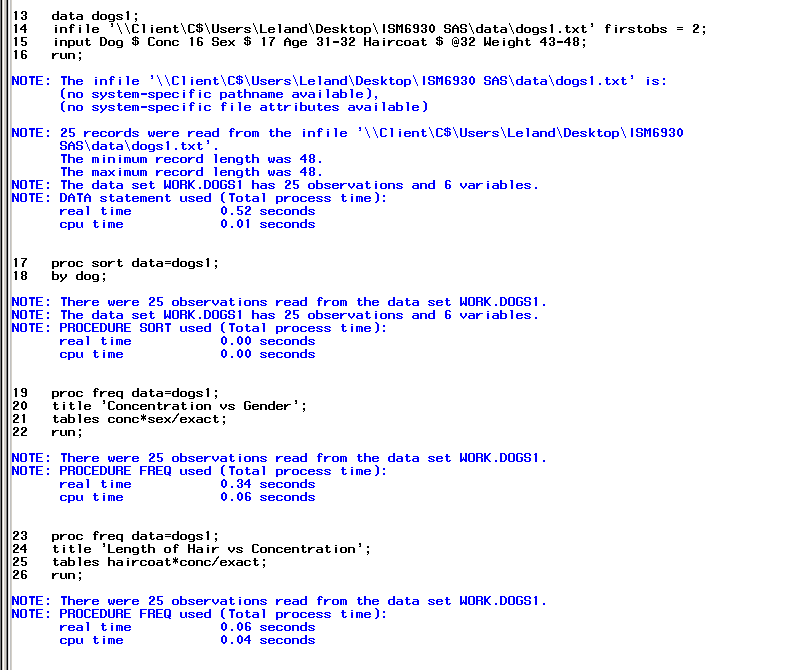
**proc** **freq** data=dogs1;

title 'Length of Hair vs Concentration';

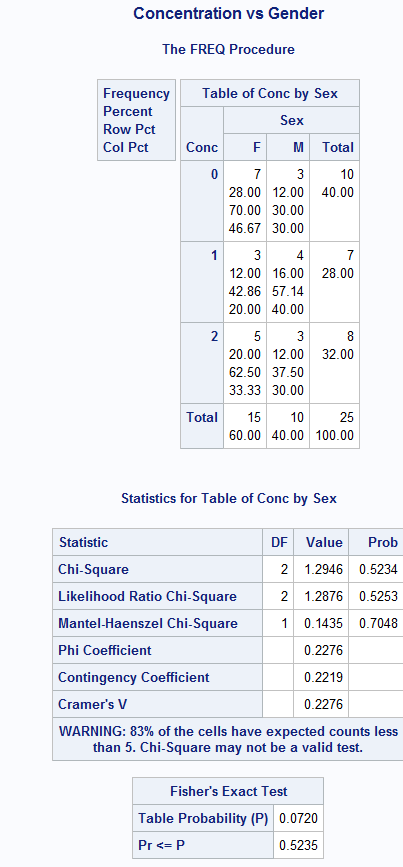
tables haircoat\*conc/exact;

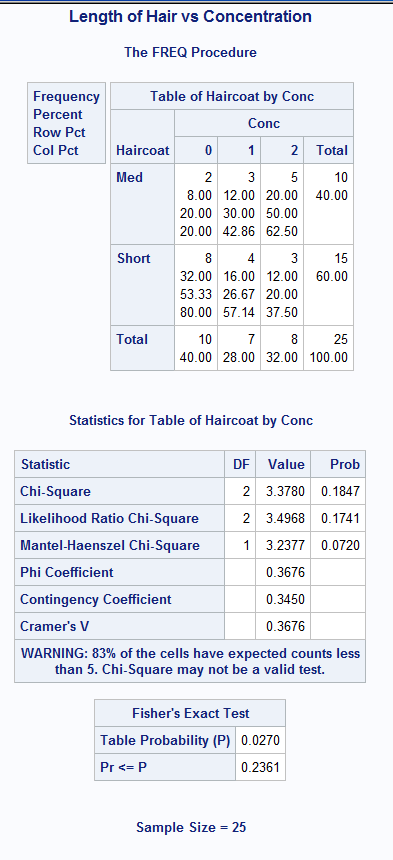
**run**;

**LOG**



**OUTPUT**





1. Refer to the MANATEES data. Your task is to see if the proportion of manatees killed by human-related causes has remained about the same through time or if this proportion has changed significantly from year to year. Create a dataset with 23 observations and three variables: year, cause indicator (human or non-human), and total number of deaths. Define deaths attributed to humans as the sum of deaths from all watercraft (WATERC), canal locks (LOCK), and other human-related deaths (RELAT). Non-human deaths include perinatal (PERIN), other natural (NATUR), and undetermined causes of death (UNDET). Then apply a chi-square test to the table of year versus cause of death (human or non-human). Make sure that this table correctly indicates total number of manatees in each cell.

**CODE**

**data** manatees;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\data\manatees.txt' dlm='09'x firstobs = **2**;

input Year Waterc Lock Relat Perin Natur Undet;

label waterc = 'Watercraft'

Relat = 'Human Related'

Perin= 'Flood Gates'

Lock = 'Canal Locks';

**run**;

**data** man\_new;

set manatees;

Cause = 'Human Related';

HumanRelatedDeaths = waterc + lock + relat;

drop waterc lock perin natur undet relat;

**run**;

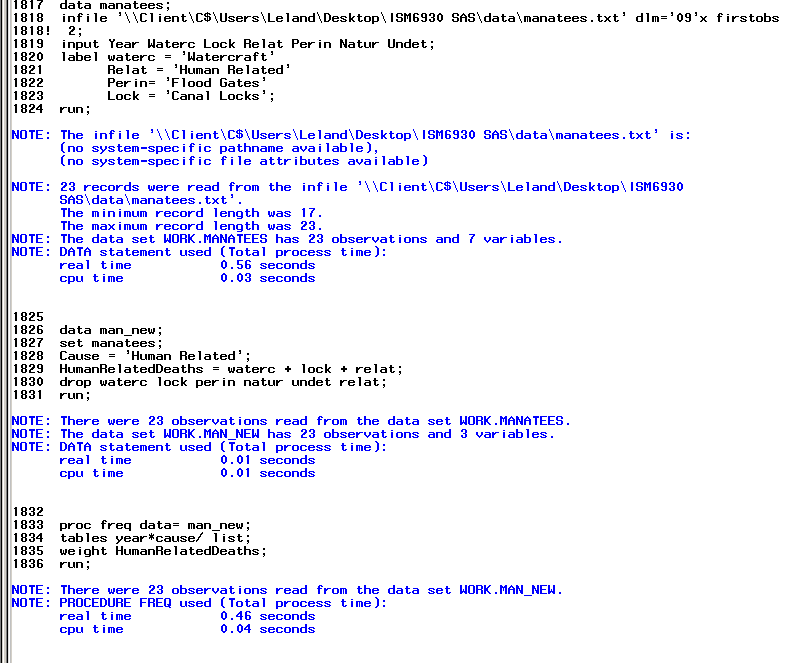
**proc** **freq** data= man\_new;

tables year\*cause/chisq;

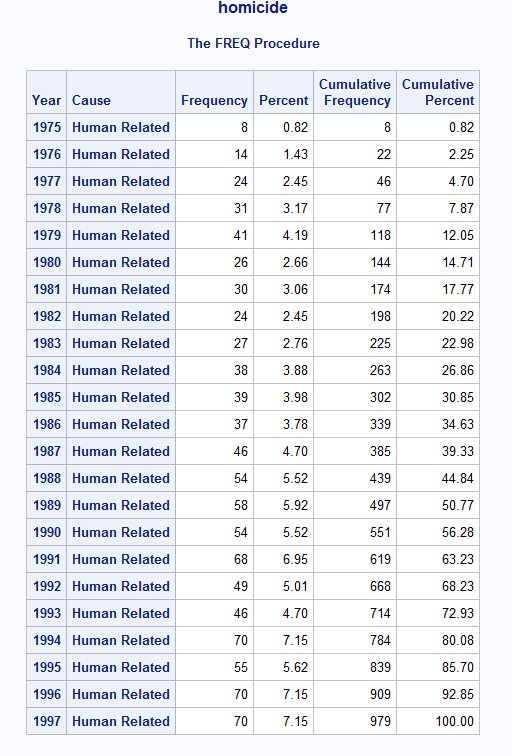
weight HumanRelatedDeaths;

**run**;

**LOG**



**OUTPUT**



1. Refer to the LIMES dataset. Fruits and vegetables are sometimes classified into groups by size for sale on the market. Classify the diameters of the lines as “small”, “medium”, and “large”, using respective cutoff points of less then 5 cm, 5 to 6 cm, and over 6 cm. Likewise, classify the juice volumes as “low”, “medium”, and “high”, using cutoff points of less than 20 ml, 20 to 40 ml, and more than 40 ml. Delete any observations which have missing diameters or missing juice volumes from the dataset. Prepare a contingency (weight) table showing the size classification versus the juiciness classification. Make sure that the rows and columns in the table are labeled with the words shown above and that they appear in the proper order from the lowest category to the highest category. Does the table indicate that larger limes tend to have more juice?

* WEIGHT statement allows you to compute the chi-square from the contingency table when you don’t want to use raw data.

How do you know what the contingency(weight) statement should be ?

Why would I not want to use raw data?

**CODE**

**data** limes;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 3\Homework\limes.txt'

dlm=',' firstobs=**2** dsd;

input Date :MMDDYY10. FruitDia FruitLength FruitWt FruitVol JuiceVol JuiceWt PeelWt ;

format date MMDDYY8.;

label Fruitdia = "Fruit Diameter"

FruitLength= "Fruit Length"

FruitWt= "Fruit Weight"

FruitVol= "Fruit Volume"

JuiceVol= "Juice Volume"

JuiceWt= "Juice Weight"

PeelWt= "Peel Weight";

**run**;

**data** newlimes;

set limes;

\*why are letters of the words being cut off;

\*will adjust where medium is first to fix;

if juicevol=**.** then delete;

else if juicevol>=**20** and juicevol<=**40** then Volume = "Medium";

else if juicevol < **20** then Volume = "Small";

else if juicevol > **40** then volume = "Large";

if fruitdia=**.** then delete;

else if fruitdia >=**5** and fruitdia <=**6** then Diameter = "Medium";

else if fruitdia < **5** then Diameter = "Small";

else if fruitdia > **6** then Diameter = "Large";

drop fruitdia juicevol;

label volume= "Juiciness"

Diameter= "Size";

**run**;

**proc** **sort** data= newlimes;

by descending diameter;

**run**;

**proc** **print** data = newlimes label;

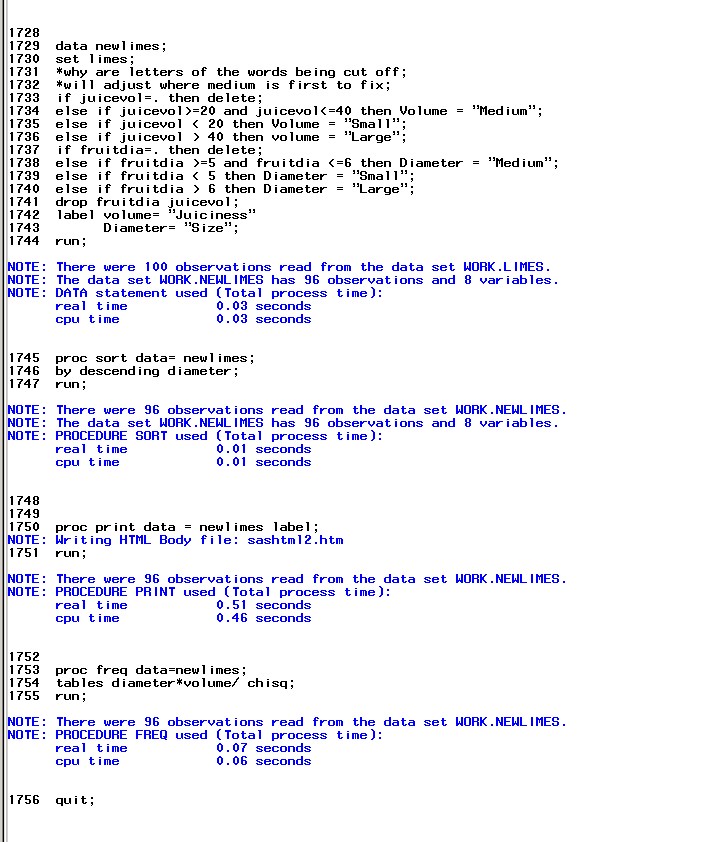
**run**;

**proc** **freq** data=newlimes;

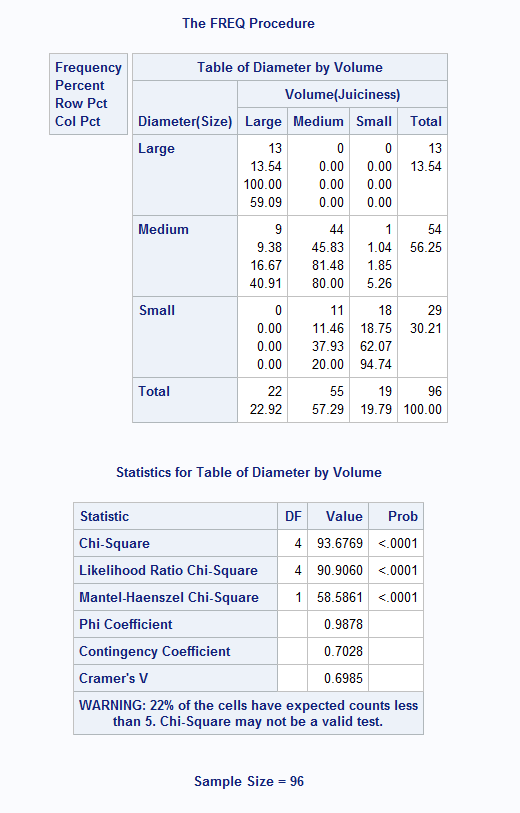
tables diameter\*volume/ chisq;

**run**;

**quit**;

**LOG**

**OUTPUT**



1. Removed
2. The *Tampa Tribune* recently published this list of manufacturers and models for the 15 cars which are most likely to be stolen. In descending order of popularity among thieves, these were:

Manufacturer Model

Honda Accord

Toyota Camry

Oldsmobile Cutlass

Honda Civic

Ford Mustang

Chevrolet C/K

Nissan Maxima

Jeep Grand Cherokee

Ford F150

Jeep Cherokee

Cadillac Deville

Ford Taurus

Chevrolet Caprice

Plymouth Voyager

Suppose that you work for an insurance company which provides coverage for used car dealers in Tampa. You want to find which dealerships have the highest risk of having cars stolen from their lots. Create one SAS dataset from the USEDCARS data and another dataset from the information given above. Use programming statements with those two datasets to classify each car as “high-risk” or “low-risk”, where a car is at high risk if it appears in the list above. Then, prepare a table showing the car dealerships versus the risk categories.

**CODE**

**data** usedcars;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 4\Homework\usedcars.txt'

firstobs=**2** obs=**51** dsd missover;

input Year **1**-**2**

@**9** Manufacturer :$10.

@**24** Model $11.

@**38** Miles comma7.0

@**49** Price dollar10.0

@**61** Dealer $23. ;

label Miles = 'Milage'

Price = 'Cost';

**run**;

**proc** **sort** data=usedcars;

by price asd;

**run**;

**proc** **print** data=usedcars label;

format milage comma7. price dollar10.;

title 'usedcars';

**run**;

**data** tribune;

Input Manufacturer $ Model $;

datalines;

Honda Accord

Toyota Camry

Oldsmobile Cutlass

Honda Civic

Ford Mustang

Chevrolet C/K

Nissan Maxima

Jeep Grand Cherokee

Ford F150

Jeep Cherokee

Cadillac Deville

Ford Taurus

Chevrolet Caprice

Plymouth Voyager

;

**data** tampa;

set tribune;

Risk="high risk";

**proc** **sort** data =usedcars;

by manufacturer model;

**run**;

**proc** **sort** data =tampa;

by manufacturer model;

**run**;

**data** carmerge;

merge usedcars tampa;

by manufacturer model;

if year=**.** then delete;

if Risk^= "high risk" then Risk= "low risk";

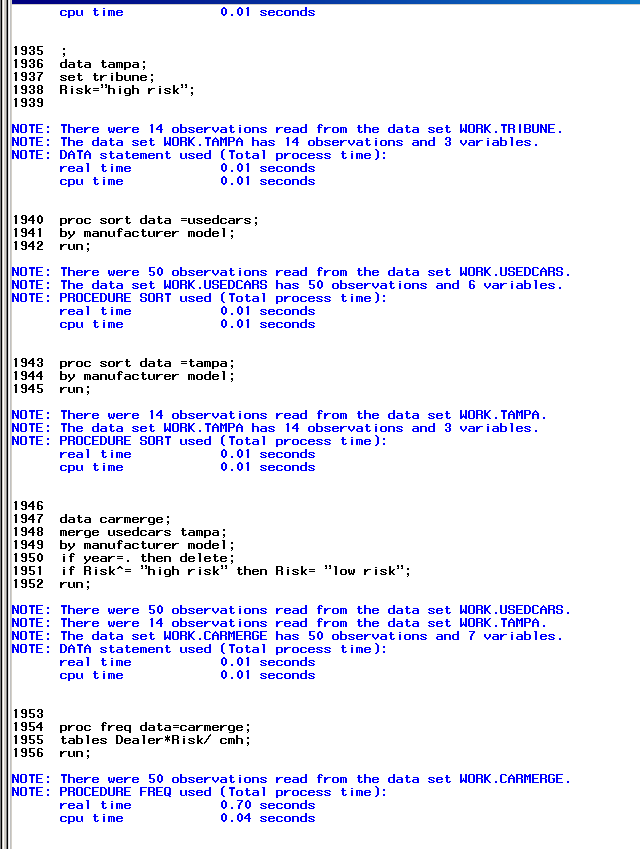
**run**;

**proc** **freq** data=carmerge;

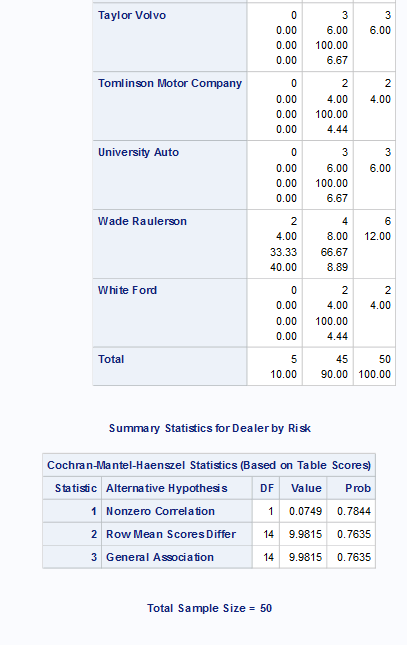
tables Dealer\*Risk/ cmh;

**run**;

**LOG**



**OUTPUT**

1. Refer to the following data, obtained from Michael Radelet, in the book Categorical Data Analysis by Alan Agresti. The data describe the circumstances of 326 homicide cases in Florida from 1976-1977.

Defendant’s race Victim’s race Death penalty Count

Black White Yes 11

Black White No 52

Black Black Yes 6

Black Black No 97

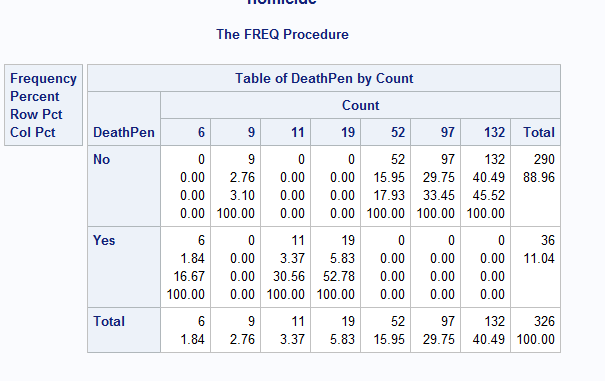
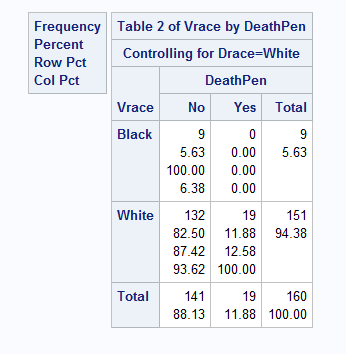
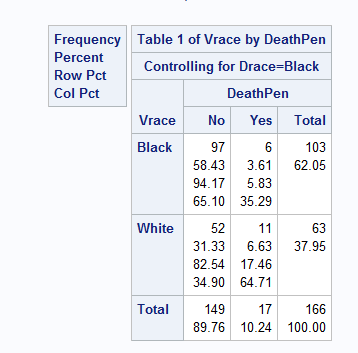
White White Yes 19

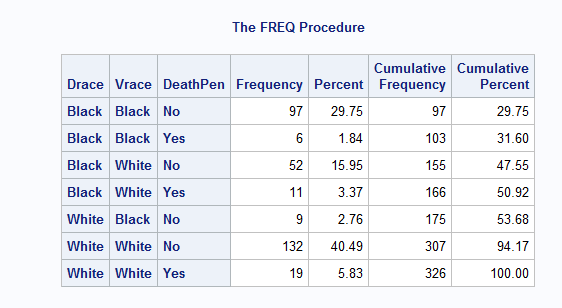
White White No 132

White Black Yes 0

White Black No 9

Use PROC FREQ to create appropriate tables to answer the following questions. Write down your answers to these questions, and mark (circle or highlight) those same numbers on your SAS output. (In other words, make sure that your tables explicitly show the requested percentages.)

1. In what percentage of cases was the death penalty verdict given? 11.04%  
   
2. When the defendant was white and the victim was black, in what percentage of cases was the death penalty verdict given? 0%  
   
3. When the defendant was black and the victim was white, in what percentage of cases was the death penalty verdict given? 6.63%  
   
4. When the races of the victim and the defendant were the same, in what percentage of cases was the death penalty given? 1.84 +5.83 = 7.67



**CODE**

**Data** homicide;

input Drace $ Vrace $ DeathPen $ Count;

datalines;

Black White Yes 11

Black White No 52

Black Black Yes 6

Black Black No 97

White White Yes 19

White White No 132

White Black Yes 0

White Black No 9

;

\*percentage of yes vs no;

**proc** **freq** data=homicide;

title 'homicide';

tables deathpen\*Count/ ;

weight count;

**run**;

\*White defendant vs black victim;

\*black defendant vs white victim;

**proc** **freq** data=homicide;

title 'homicide';

tables drace\*vrace\*deathpen;

weight count;

**run**;

\*WvW BvB;

**proc** **freq** data=homicide;

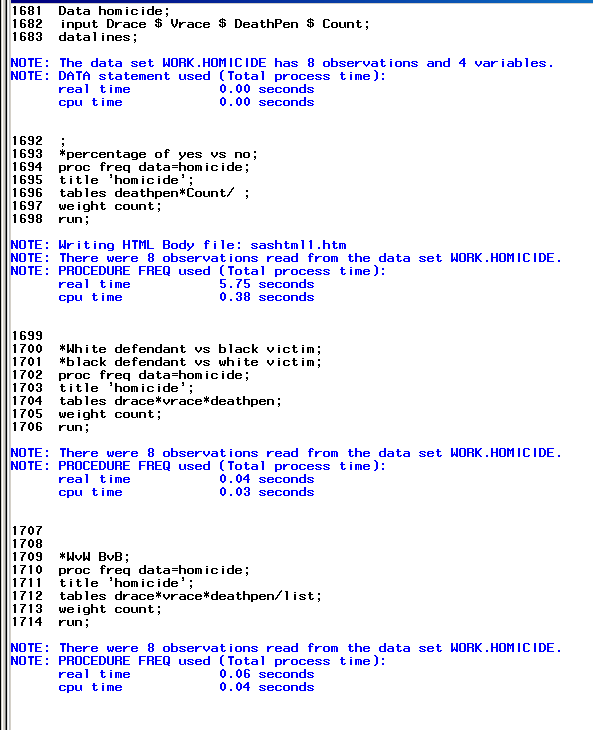
title 'homicide';

tables drace\*vrace\*deathpen/list;

weight count;

**run**;

**LOG**



1. Refer to HOCKEY data. Did Ohio State have a home-ice advantage? Use PROC FREQ to perform Fisher’s exact test for the 2x2 table of the outcome of the game (“won” or “lost or tied”) versus location (Columbus, Ohio or elsewhere). Don’t forget to change the score of the final game to Boston College 5, Ohio State 2.

**CODE**

**data** hockey;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 4\Homework\hockey.csv' dlm=',' dsd missover firstobs=**2**;

INPUT Game\_Date : MMDDYY10.

Team : $27.

City : $17.

State : $13.

OSU

OPP;

FORMAT Game\_Date mmddyy10.;

if team = 'Boston College' then opp=**5**;

if team = 'Boston College' then osu=**2**;

**RUN**;

**data** record;

set hockey;

if osu<opp then outcome = "lost";

else if osu>opp then outcome = "won";

else if osu=opp then outcome = "tied";

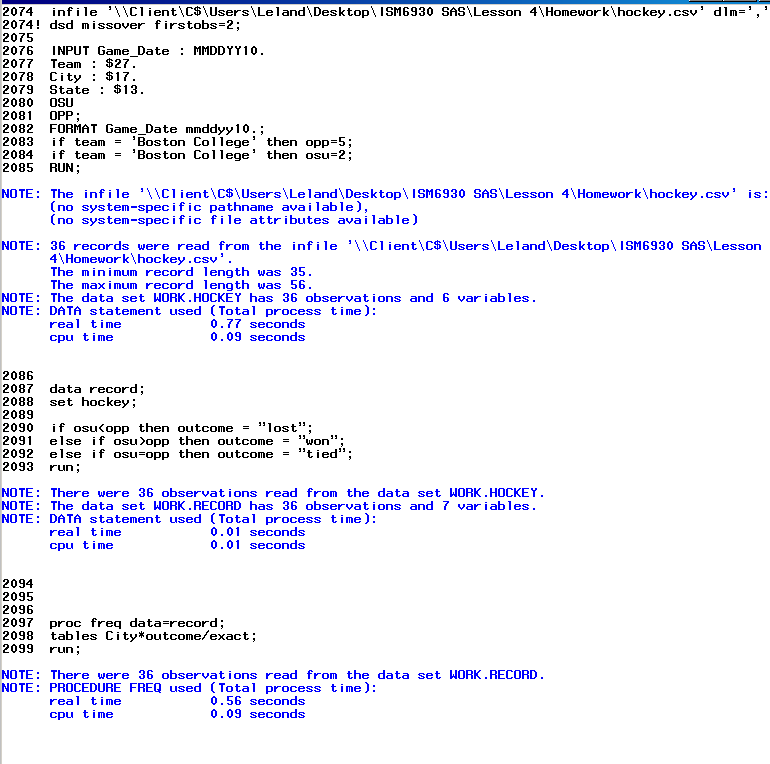
**run**;

**proc** **freq** data=record;

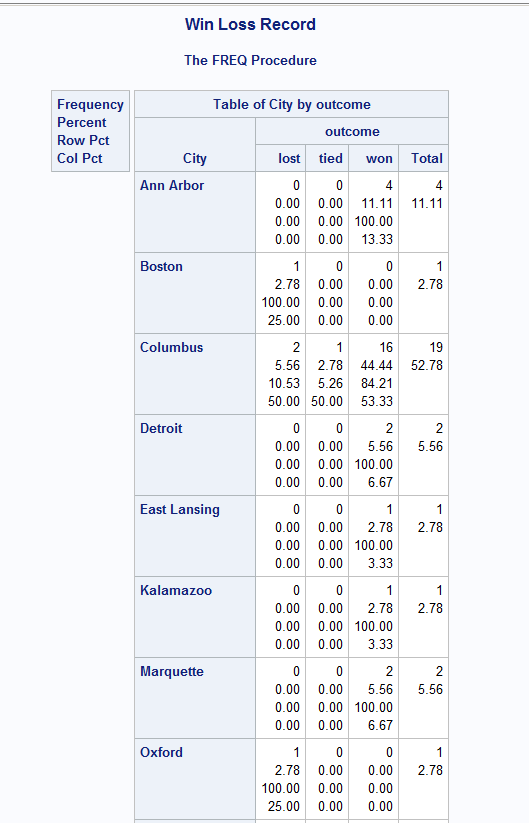
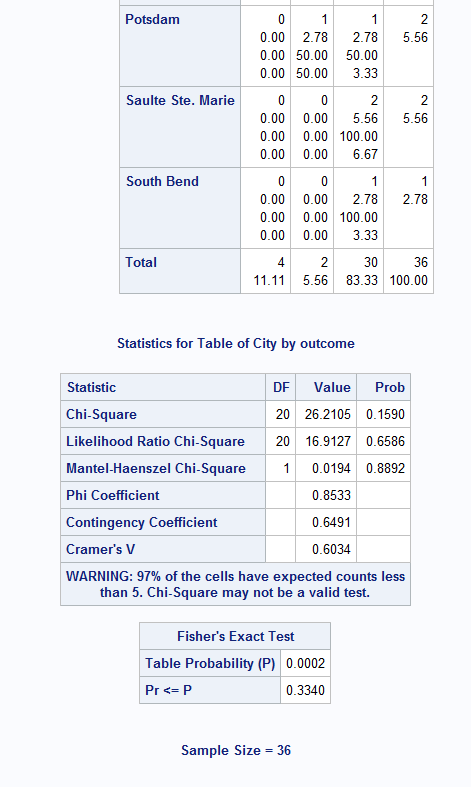
title 'Win Loss Record';

tables City\*outcome/exact;

**run**;

**LOG** 

**OUTPUT**

1. Refer to the HANKS data. Your task is to use a SAS program to count the number of movies in which Tom Hanks appeared in each of the years from 1984-1998, then make a scatterplot with the number of movies made in each year on the vertical axis versus the year on the horizontal axis. In some years, such as 1997, he did not appear in any movies. On the scatterplot, indicate those years by plotting a point at zero. To do this you could create another dataset with all of the years from 1984-1998 by using a Do loop, then MERGE that dataset with the dataset containing the movie counts. Then, you will need to replace missing values for movie counts with zeroes.

Does this relate to R. Questions with Multiple Answers in the Slides?

I think I need to proc transpose the years where the years become variables to get a count of how many movies happen in a year.

**CODE**

**data** hanks;

infile '\\Client\C$\Users\Leland\Desktop\ISM6930 SAS\Lesson 4\Homework\hanks.txt'

dlm='09'x missover firstobs=**2**;

input Title $ **1**-**25** Year **26**-**29** Length **34**-**36** MPAA $ **42**-**46** Action **50**-**51** Drama **58**-**59** Humor **66**-**67** Sex **74** Violence **82**-**83** Suspense **90** Offbeat **98** ;

**run**;

**proc** **format**;

value $rating 'G' = 'Suitable'

'PG' = 'Questionable'

'PG-13' = 'Questionable'

'R' = 'Not Suitable'

;

**run**;

**proc** **sort** data=hanks;

by year;

**run**;

**proc** **freq** data=hanks;

tables year/out=hankscount(drop=percent);

**run**;

**data** years;

do year= **1984** to **1998**;

output;

end;

**proc** **sort** data=years;

by year;

**run**;

**proc** **sort** data=hankscount;

by year;

**run**;

**data** hanksmerge;

merge years hankscount;

by year;

if count=**.** then count = **0**;

**run**;

**proc** **print** data=hanksmerge;

title 'hankscount';

**run**;

**proc** **gplot** data=hanksmerge;

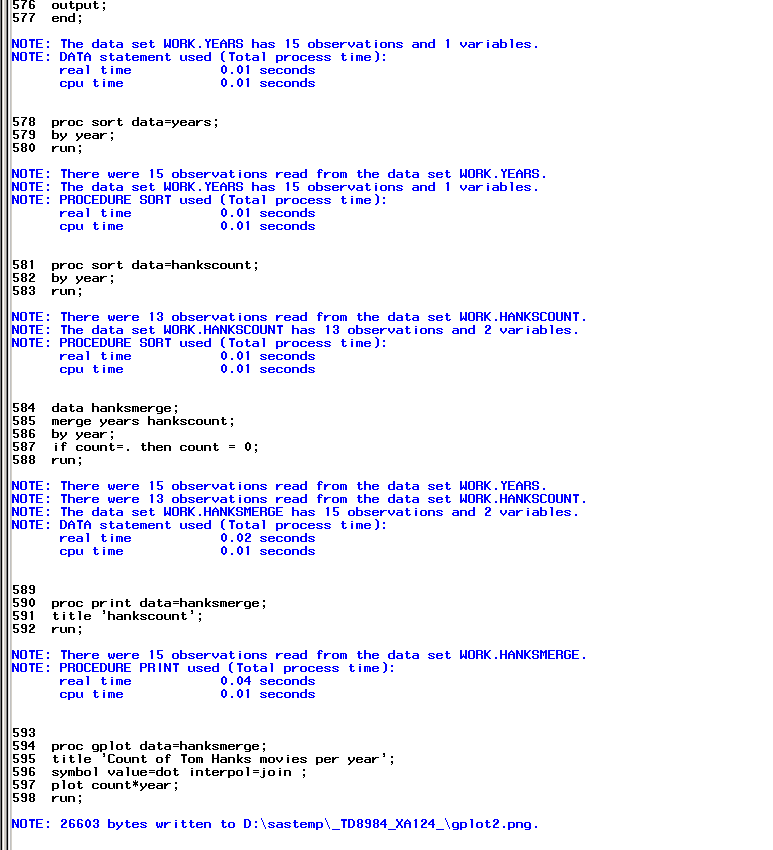
title 'Count of Tom Hanks movies per year';

symbol value=dot interpol=join ;

plot count\*year;

**run**;

**LOG**



**OUTPUT**

