SAS — Assignment 1

牛圣杰 11910901

Problem 1:

1°. For $f_1(x)$, We have kurtosis $= \frac{E[(X-\mu_1)^4]}{\sigma_1^4}$,

$$\Rightarrow \mu = E_{X_1}(X) = \int_{-\infty}^{+\infty} x f_1(x) dx = \int_{-0.9399}^{0.9399} 0.3334 x dx + \int_{0.9399}^{2.3243} \frac{0.2945}{x} dx + \int_{-2.3243}^{-0.9399} \frac{0.2945}{x} dx$$

$$= 0 + \ln(0.2945x)\Big|_{0.9399}^{2.3243} + \ln(-0.2945x)\Big|_{-2.3243}^{0.9399} = 0.$$

$$\Rightarrow E[(X-\mu)^4] = E(X^4) = \int_{-\infty}^{+\infty} X^4 f_1(x) dx = \int_{-0.9399}^{0.9399} 0.3334 X^4 dx + \int_{0.9399}^{2.3243} 0.2945 X^2 dx + \int_{-2.3243}^{-0.9399} 0.2945 X^2 dx$$

$$= \frac{0.3334}{5} X^5 \Big|_{-0.9399}^{0.9399} + \frac{0.2945}{3} X^3 \Big|_{0.9399}^{2.3243} + \frac{0.2945}{3} X^3 \Big|_{-2.3243}^{-0.9399}$$

$$= \frac{0.3334}{5}[2 \times 0.9399^5] + \frac{0.2945}{3}[2.3243^3 - 0.9399^3 - 0.9399^3 + 2.3243^3]$$

$$= 0.09782 + 2.30229 \approx 2.40$$

$$\Rightarrow \sigma_1^2 = Var(X) = E[X-\mu_1]^2 = E(X^2) = \int_{-0.9399}^{0.9399} 0.3334 X^2 dX + \int_{0.9399}^{2.3243} 0.2945 dX + \int_{-2.3243}^{-0.9399} 0.2945 dX$$

$$= \frac{0.3334}{3}[2 \times 0.9399^3] + 0.2945[2.3243 - 0.9399 - 0.9399 + 2.3243]$$

$$= 0.18455 + 0.81541 \approx 1.00$$

$\Rightarrow$ The population kurtosis of $f_1(x)$ is $\frac{E[(X-\mu)^4]}{\sigma_1^2} = 2.40 < 3$

$\Rightarrow$ this distribution produces less outliers than the normal distribution.

1°. For $f_2(x)$, We have kurtosis $= \frac{E[(X-\mu_1)^4]}{\sigma_1^4}$,

$$\Rightarrow \mu_2 = E_{X_2}(X) = \int_{-\infty}^{+\infty} x f_2(x) dx = \int_{-2.4495}^{2.4495} 0.4082x - 0.1667|x|x \, dx = \int_{-2.4495}^{0} 0.4082x + 0.1667x^2 dx + \int_{0}^{2.4495} 0.4082x - 0.1667x^2 dx$$

$$= \left(\frac{0.4082}{2}x^2 + \frac{0.1667}{3}x^3\right)\Big|_{-2.4495}^{0} + \left(\frac{0.4082}{2}x^2 - \frac{0.1667}{3}x^3\right)\Big|_{0}^{2.4495}$$

$$= -\frac{0.4082}{2}\cdot(2.4495)^2 + \frac{0.1667}{3}\cdot(2.4495)^3 + \frac{0.4082}{2}(2.4495)^2 - \frac{0.1667}{3}(2.4495)^3 = 0$$

$$\Rightarrow E[(X_2-\mu_2)^4] = E(X_2^4) = \int_{0}^{2.4495} 0.4082x^4 - 0.1667x^5 dx + \int_{-2.4495}^{0} (0.4082x^4 + 0.1667x^5) dx$$

$$= \left(\frac{0.4082}{5}x^5 - \frac{0.1667}{6}x^6\right)\Big|_{0}^{2.4495} + \left(\frac{0.4082}{5}x^5 + \frac{0.1667}{6}x^6\right)\Big|_{-2.4495}^{0}$$

$$= \frac{0.4082}{5}\cdot(2.4495)^5 - \frac{0.1667}{6}(2.4495)^6 + \frac{0.4082}{5}\cdot(2.4495)^5 - \frac{0.1667}{6}(2.4495)^6$$

$$\approx 2.40$$

$$\Rightarrow \sigma_2^2 = Var(X_2) = E(X_2-\mu_2)^2 = \int_{0}^{2.4495} 0.4082x^2 - 0.1667x^3 dx + \int_{-2.4495}^{0} 0.4082x^2 + 0.1667x^3 dx$$

$$= 2\left[\frac{0.4082}{3}\cdot(2.4495)^3 - \frac{0.1667}{4}\cdot(2.4495)^4\right] \approx 1.00$$

⇒ The population kurtosis of $f_1(x)$ is $\frac{E[(X-\mu_2^4]}{\sigma_2^2} = 2.40 < 3$

⇒ this distribution produces less outliers than the normal distribution.

Note that for $f_1(x)$, $f_2(x)$, these two distribution have the same mean, variance, skewness and kurtosis, but the distribution are not same ⇒ We can't judge the distribution only by mean, variance, skewness and kurtosis.

Problem 2:

Note that there are totally $\frac{n(n-1)}{2} = 45$ pairwise matching methods

And note that there is 5 discordant pairs and 40 concordant pairs.

⇒ Thus, the Kendall's Tau is $\tau = \frac{40-5}{45} = \frac{7}{9}$

⇒ Thus, we have that the rank of projects given by 2 TAs has a monotonic positively relationships

Problem 3:

MCAR: $Pr(M=1 \mid \underset{\sim}{X}, Y) = Pr(M=1)$

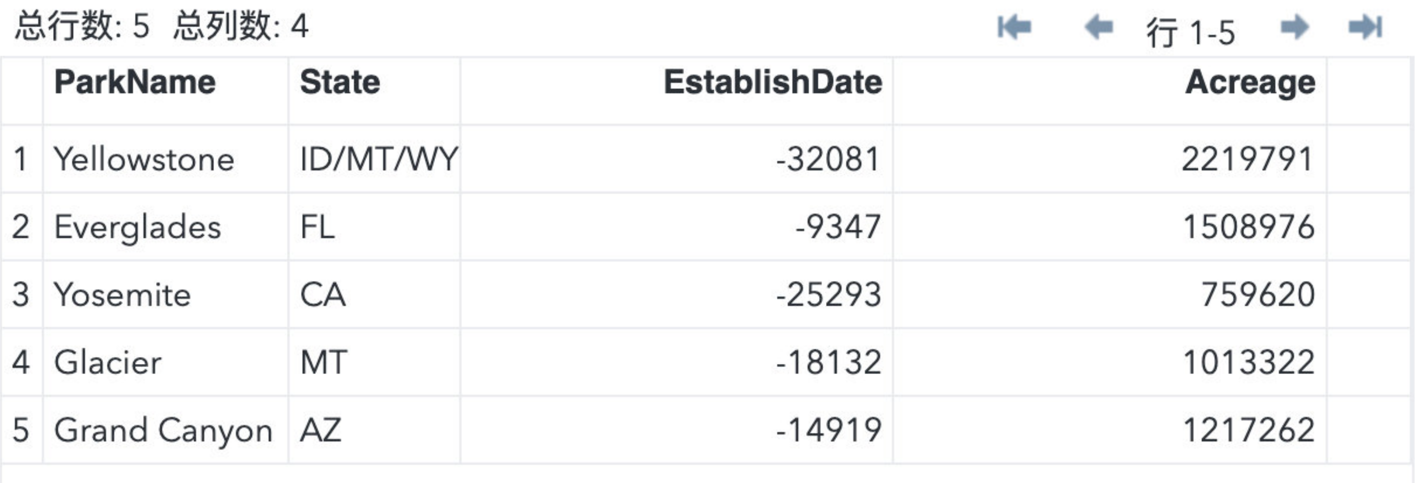
⇒ that is, the probability that Y is missing depends neither on the observed variables X nor on the possibly missing values of Z itself.

MAR: $Pr(M=1 \mid \underset{\sim}{X}, Y) = Pr(M=1 \mid \underset{\sim}{X})$

⇒ missingness on Y may depend on $\underset{\sim}{X}$, but it does not depend on Y itself (after adjusting for $\underset{\sim}{X}$)

Problem 4

```
DATA Problem4;
INFILE "/home/u60917423/Assignments/Assignment1/NationalPark.txt";
INPUT @1 ParkName $ 12.
  @ 15 State $ 8.
  +2 EstablishDate mmddyy10.
  +2 Acreage comma9.;
RUN;
```



总行数: 5 总列数: 4

行 1-5

	ParkName	State	EstablishDate	Acreage
1	Yellowstone	ID/MT/WY	-32081	2219791
2	Everglades	FL	-9347	1508976
3	Yosemite	CA	-25293	759620
4	Glacier	MT	-18132	1013322
5	Grand Canyon	AZ	-14919	1217262

Problem 5

(a)

```
LIBNAME ass "/home/u60917423/my_shared_file_links/u44964922/Assignments";
PROC CONTENTS DATA = ass.sff POSITION;
RUN;
DATA problem5;
  SET ASS.SFF;
PROC FREQ DATA=PROBLEM5;
  TABLES Continent/NOPERCENT NOCUM;
RUN;
```

按创建时间排序的变量

#	变量	类型	长度	标签
1	ByDate	数值	8	ID for sorting by first case date
2	ByCont	数值	8	ID for sorting by first case date within a continent
3	Country	字符	30	Name of country
4	FirstCase	数值	8	Date of first case reported
5	Apr	数值	8	Number of cumulative cases reported on the first day of the month for April
6	May	数值	8	Number of cumulative cases reported on the first day of the month for May
7	June	数值	8	Number of cumulative cases reported on the first day of the month for June
8	July	数值	8	Number of cumulative cases reported on the first day of the month for July
9	Aug	数值	8	Number of cumulative cases reported on the first day of the month for August
10	Latest	数值	8	Last reported cumulative number of cases reported to WHO as of August 9, 2009
11	ByDate_d	数值	8	ID for sorting by first death date
12	ByCont_d	数值	8	ID for sorting by first death date within a continent
13	FirstDeath	数值	8	Date of first death
14	May_d	数值	8	Number of cumulative deaths reported on the first day of the month for May
15	June_d	数值	8	Number of cumulative deaths reported on the first day of the month for June
16	July_d	数值	8	Number of cumulative deaths reported on the first day of the month for July
17	Aug_d	数值	8	Number of cumulative deaths reported on the first day of the month for August
18	Sep_d	数值	8	Number of cumulative deaths reported on the first day of the month for September
19	Oct_d	数值	8	Number of cumulative deaths reported on the first day of the month for October
20	Nov_d	数值	8	Number of cumulative deaths reported on the first day of the month for November
21	Dec_d	数值	8	Number of cumulative deaths reported on the first day of the month for December
22	Continent	字符	13	Continent

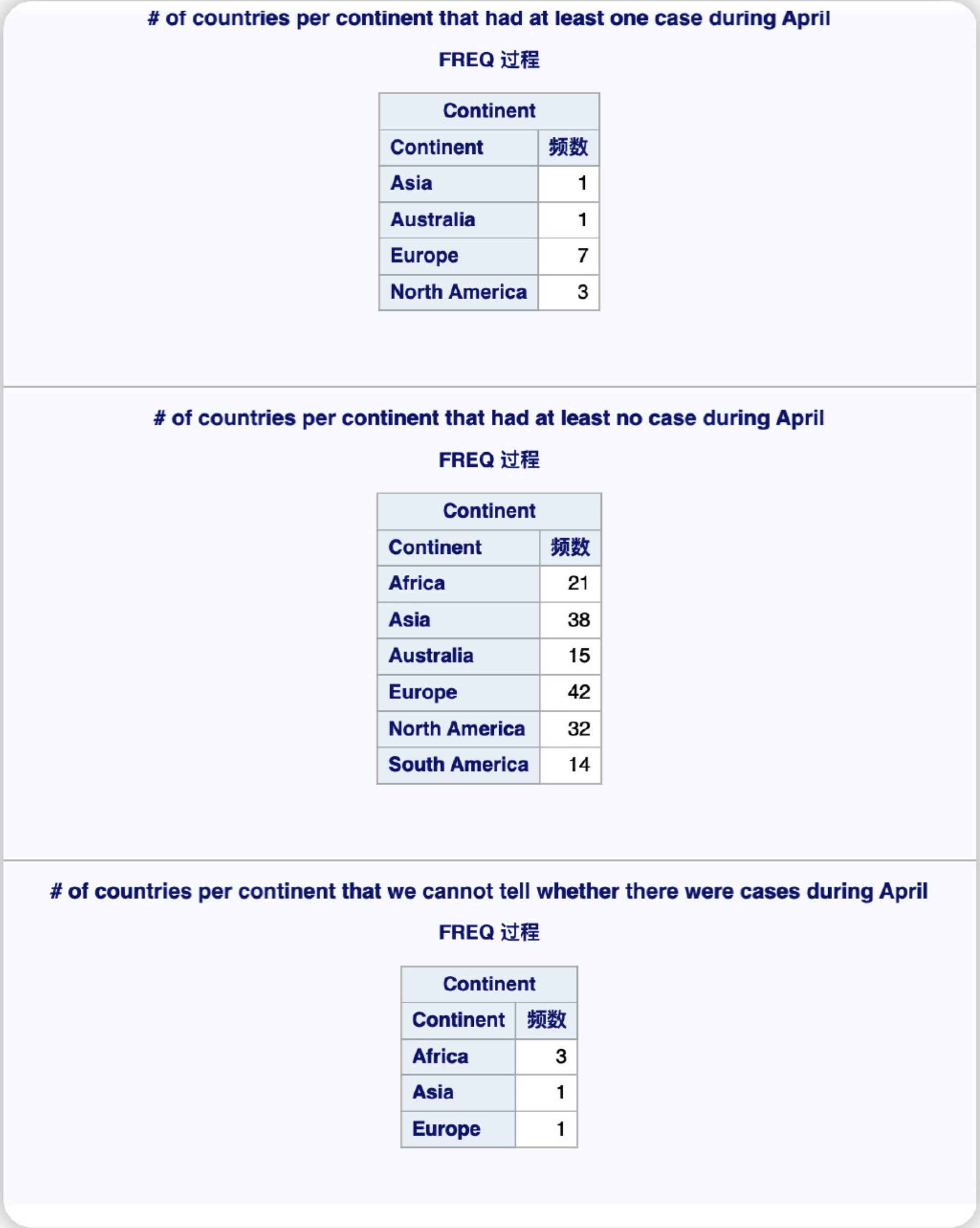
Continent

Continent	频数
Africa	24
Asia	40
Australia	16
Europe	50
North America	35
South America	14

(b)

```
* Note that the number for 05-01-2009 is 18018 and the number for 04-01-2009 is 17988;
PROC SQL;
  create table data1 AS
  SELECT FirstCase, Continent
  from PROBLEM5
  where 17988 <= FirstCase <18018;
```

```
quit;

PROC FREQ DATA=data1;
  TABLES Continent/NOPERCENT NOCUM;
  TITLE "# of countries per continent that had at least one case during April"
RUN;

PROC SQL;
  create table data2 AS
  SELECT FirstCase, Continent
  from PROBLEM5
  where FirstCase >= 18018;
quit;

PROC FREQ DATA=data2;
  TABLES Continent/NOPERCENT NOCUM;
  TITLE "# of countries per continent that had at least no case during April"
RUN;

PROC SQL;
  create table data3 AS
  SELECT FirstCase, Continent
  from PROBLEM5
  where FirstCase <17988;
quit;

PROC FREQ DATA=data3;
  TABLES Continent/NOPERCENT NOCUM;
  TITLE "# of countries per continent that we cannot tell whether there were cases during April."
RUN;
```



# of countries per continent that had at least one case during April

FREQ 过程

Continent	
Continent	频数
Asia	1
Australia	1
Europe	7
North America	3

# of countries per continent that had at least no case during April

FREQ 过程

Continent	
Continent	频数
Africa	21
Asia	38
Australia	15
Europe	42
North America	32
South America	14

# of countries per continent that we cannot tell whether there were cases during April

FREQ 过程

Continent	
Continent	频数
Africa	3
Asia	1
Europe	1

(c)

```
PROC SQL;
  SELECT CONTINENT, COUNTRY, FIRSTCASE, LATEST, FIRSTDEATH format=DDMMYY10.
  FROM ASS.SFF
  WHERE FIRSTCASE=. AND FIRSTDEATH <> .
ORDER BY CONTINENT;
QUIT;
```

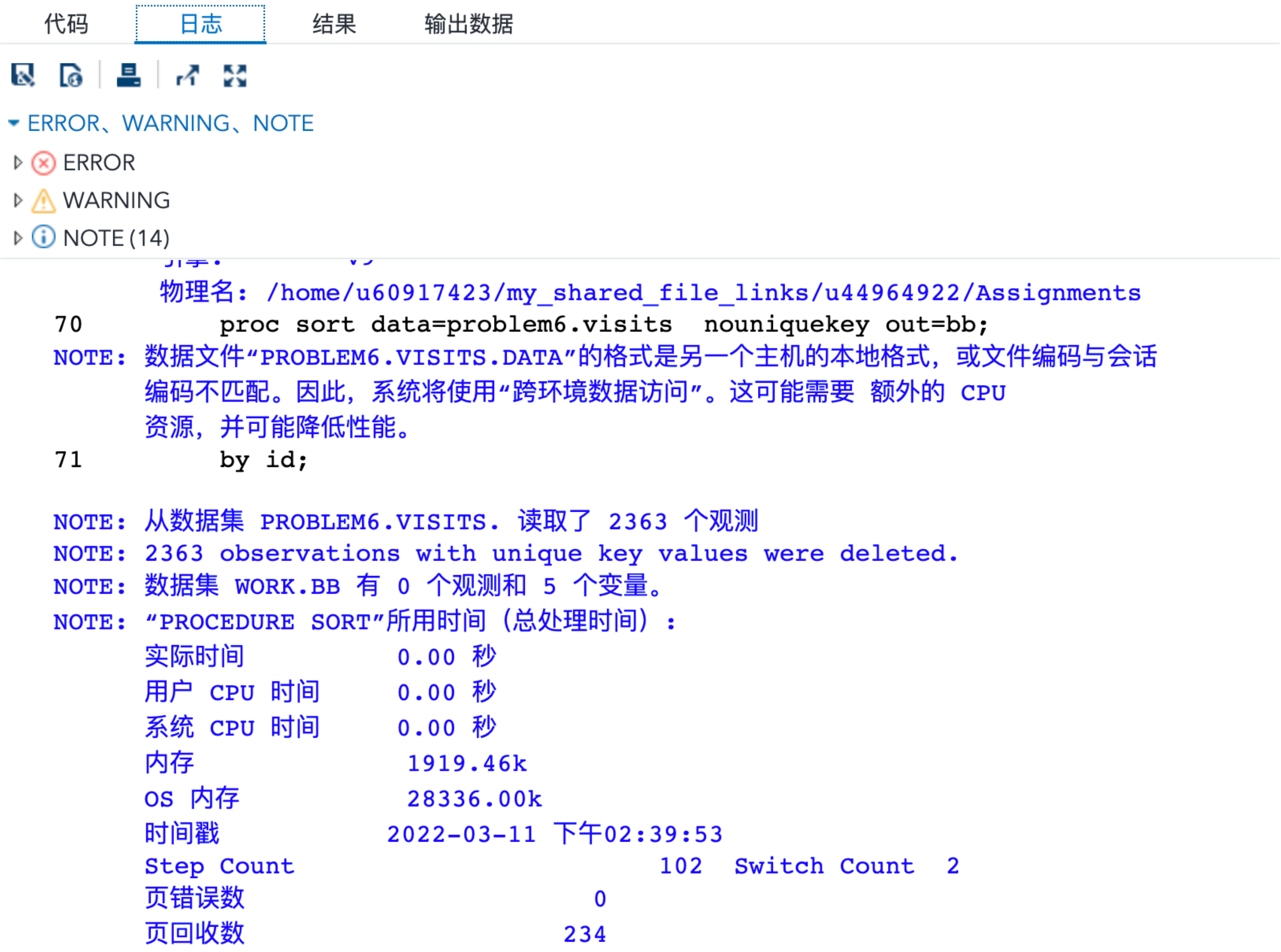
Continent	Name of country	Date of first case reported	Last reported cumulative number of cases reported to WHO as of August 9, 2009	Date of first death
Africa	São Tomé and Príncipe	.	.	26/10/2009
Africa	Madagascar	.	.	11/09/2009
Africa	Mozambique	.	.	16/09/2009
Asia	Mongolia	.	.	26/10/2009
Europe	Belarus	.	.	06/11/2009

Problem 6

(a)

According to the note, we have that there is 0 obs and 5 variables in dataset WORK.BB, which means there is no multiple records for 'visit' data.

```
LIBNAME problem6 "~/my_shared_file_links/u44964922/Assignments";
proc sort data=problem6.visits  nouniquekey out=bb;
by id;
proc print;
run;
proc sort data=problem6.txgroup  nouniquekey out=bb;
by id;
run;
```



```
         物理名: /home/u60917423/my_shared_file_links/u44964922/Assignments
 70          proc sort data=problem6.visits  nouniquekey out=bb;
 NOTE: 数据文件“PROBLEM6.VISITS.DATA”的格式是另一个主机的本地格式，或文件编码与会话
       编码不匹配。因此，系统将使用“跨环境数据访问”。这可能需要 额外的 CPU
       资源，并可能降低性能。
 71          by id;

 NOTE: 从数据集 PROBLEM6.VISITS. 读取了 2363 个观测
 NOTE: 2363 observations with unique key values were deleted.
 NOTE: 数据集 WORK.BB 有 0 个观测和 5 个变量。
 NOTE: “PROCEDURE SORT”所用时间（总处理时间）:
       实际时间            0.00 秒
       用户 CPU 时间       0.00 秒
       系统 CPU 时间       0.00 秒
       内存                 1919.46k
       OS 内存              28336.00k
       时间戳              2022-03-11 下午02:39:53
       Step Count                        102  Switch Count  2
       页错误数                               0
       页回收数                             234
```

(b)

```
PROC SQL;
  create table Visits AS
  select ID, visitdt, gender, visit, B_cholesterol
  from problem6.visits;
quit;
proc sort data=Visits;
  by id;
run;
PROC SQL;
  create table TX AS
  select ID, TX
  from problem6.txgroup;
quit;
proc sort data=TX  NODUPKEY;
  by id;
run;

DATA temp;
  MERGE Visits TX;
  BY ID;
PROC SQL;
  create table combine AS
  select ID, TX, visit, visitdt, B_cholesterol
  from temp
  where TX = 1;
RUN;
```

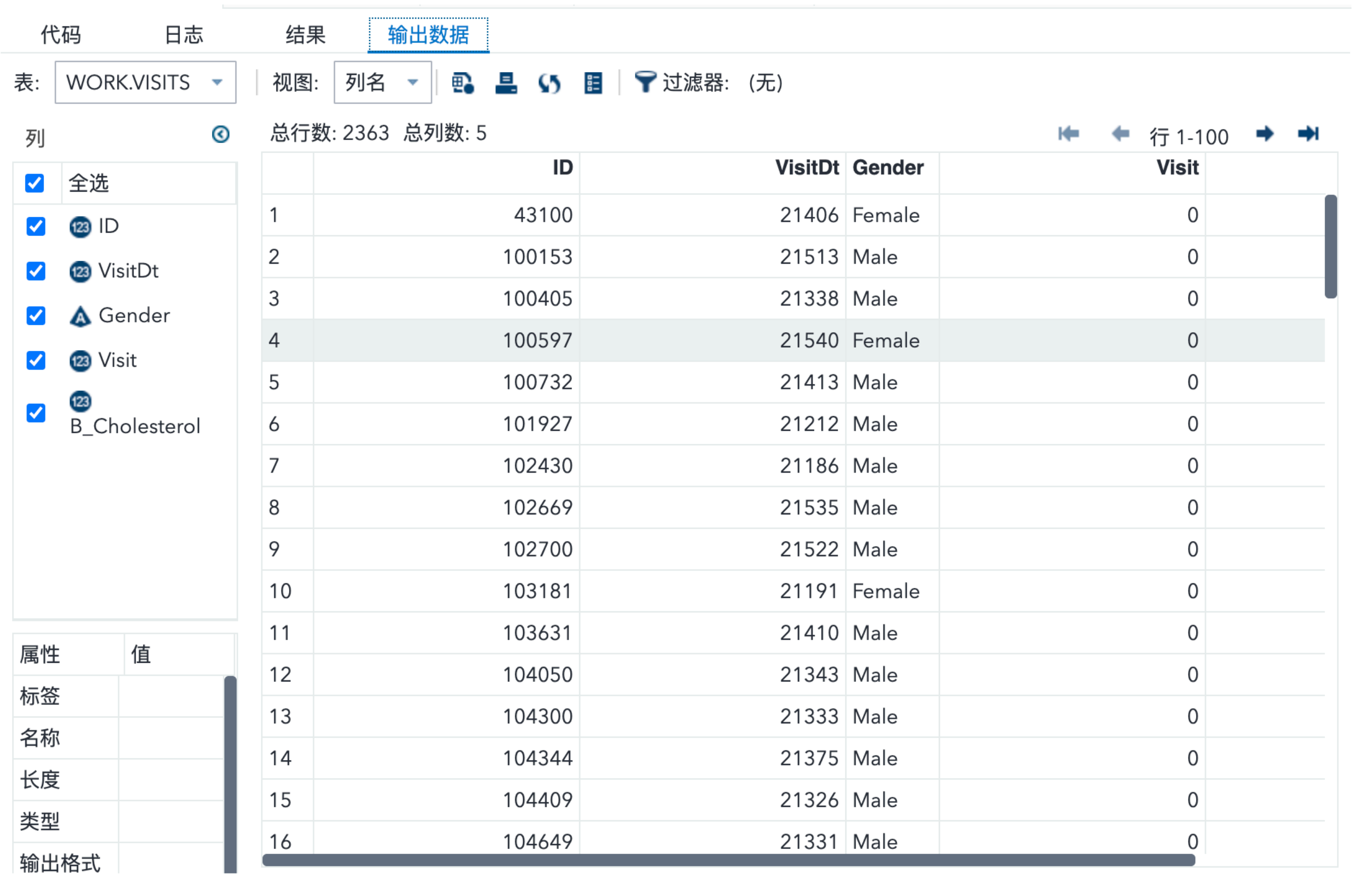


表: WORK.VISITS　视图: 列名　过滤器: (无)

总行数: 2363　总列数: 5　行 1-100

☑ 全选
☑ ID
☑ VisitDt
☑ Gender
☑ Visit
☑ B_Cholesterol

	ID	VisitDt	Gender	Visit
1	43100	21406	Female	0
2	100153	21513	Male	0
3	100405	21338	Male	0
4	100597	21540	Female	0
5	100732	21413	Male	0
6	101927	21212	Male	0
7	102430	21186	Male	0
8	102669	21535	Male	0
9	102700	21522	Male	0
10	103181	21191	Female	0
11	103631	21410	Male	0
12	104050	21343	Male	0
13	104300	21333	Male	0
14	104344	21375	Male	0
15	104409	21326	Male	0
16	104649	21331	Male	0

(c)

```
proc sql;
create table median as
select median(B_cholesterol) as Median from combine ;
quit;

PROC SQL;
  CREATE TABLE Dummy AS
  SELECT ID, TX, visitdt,visit,B_Cholesterol,Median
  FROM combine median;
QUIT;
Data Dummy;
  SET Dummy;
  Length Abovemedian $ 2.;
  IF B_Cholesterol>Median Then Abovemedian=1;
  ELSE Abovemedian=0;
RUN;
PROC PRINT data=Dummy;
RUN;
```

Problem 7

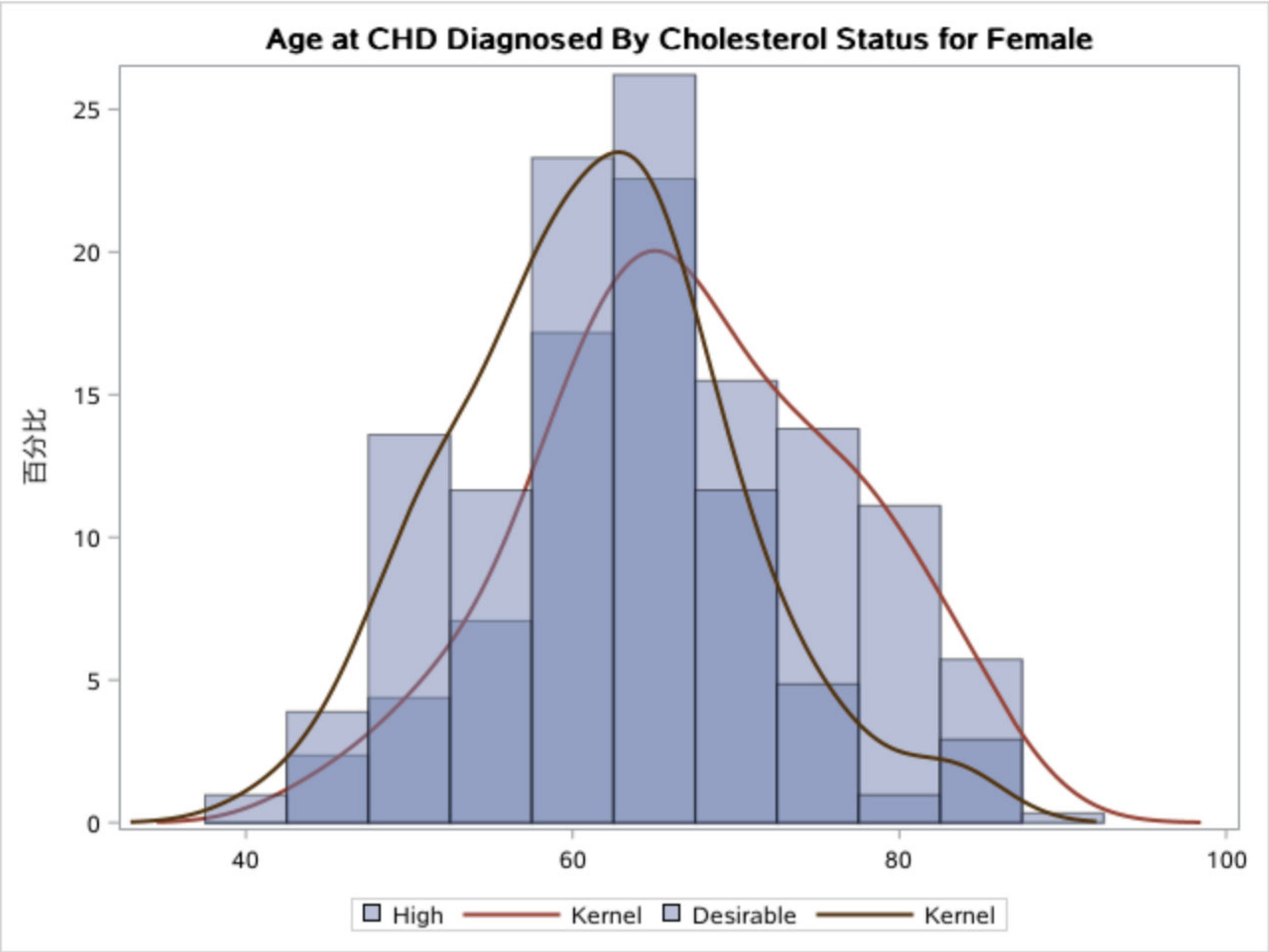
(a)

```
LIBNAME ass "/home/u60917423/my_shared_file_links/u44964922/Assignments";
PROC CONTENTS DATA = ass.sff POSITION;
RUN;

DATA problem5;
  SET ASS.SFF;
PROC FREQ DATA=PROBLEM5;
  TABLES Continent/NOPERCENT NOCUM;
RUN;
```

		Sex						全部		
		Female			Male					
		Age at Death			Age at Death			Age at Death		
		N	Mean	Median	N	Mean	Median	N	Mean	Median
Cause of Death	Smoking Status									
Cancer	Heavy (16-25)	33	61.61	62.00	93	67.82	68.00	126	66.19	66.50
	Light (1-5)	30	68.97	67.50	16	74.88	77.00	46	71.02	70.00
	Moderate (6-15)	34	62.97	62.00	24	70.17	72.50	58	65.95	65.00
	Non-smoker	150	69.74	71.00	84	74.23	75.00	234	71.35	72.00
	Very Heavy (> 25)	8	64.63	64.50	64	66.95	68.50	72	66.69	68.00
Cerebral Vascular Disease	Heavy (16-25)	19	69.26	71.00	54	70.43	70.50	73	70.12	71.00
	Light (1-5)	27	69.85	72.00	12	69.33	71.50	39	69.69	72.00
	Moderate (6-15)	19	70.11	74.00	24	70.38	71.50	43	70.26	72.00
	Non-smoker	122	75.64	77.00	59	73.31	75.00	181	74.88	76.00
	Very Heavy (> 25)	8	65.38	66.00	29	67.07	66.00	37	66.70	66.00
Coronary Heart Disease	Heavy (16-25)	24	70.54	72.50	103	66.19	66.00	127	67.02	67.00
	Light (1-5)	23	72.30	72.00	32	66.88	65.00	55	69.15	70.00
	Moderate (6-15)	22	71.14	69.00	39	70.59	71.00	61	70.79	71.00
	Non-smoker	134	75.14	75.00	137	72.69	73.00	271	73.90	74.00
	Very Heavy (> 25)	5	67.20	75.00	80	64.30	64.50	85	64.47	65.00

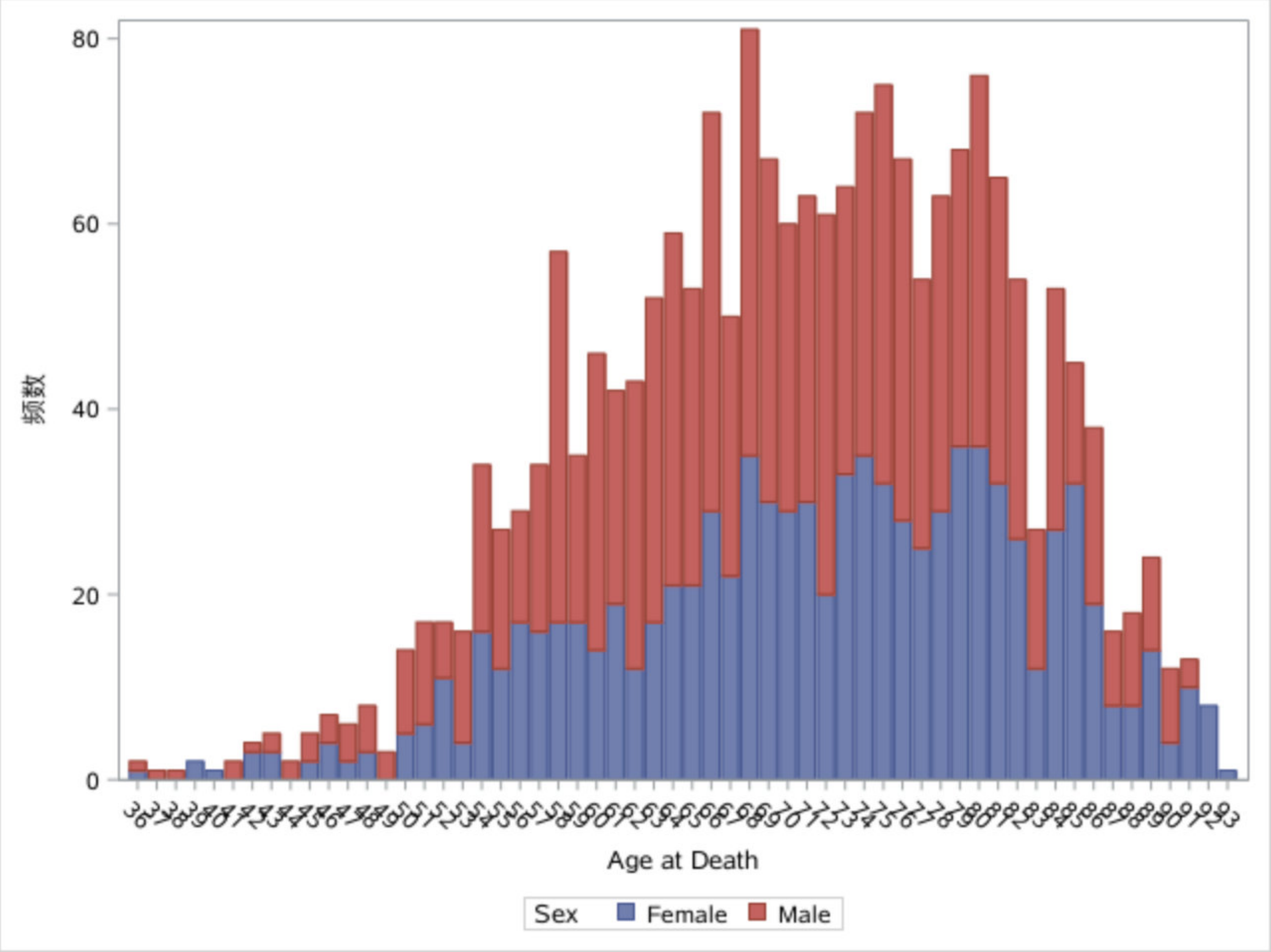
(b)

```
DATA heart1;
  set SASHELP.HEART;
  IF Sex ~= 'Female' Then Delete;
  IF Chol_Status ~= 'High' Then Delete;
run;
DATA heart2;
  set SASHELP.HEART;
  IF Sex ~= 'Female' Then Delete;
  IF Chol_Status ~= 'Desirable' Then Delete;
run;
DATA h1;
  set heart1;
  CHDhigh = AgeCHDdiag;
run;
DATA h2;
  set heart2;
  CHDdesir = AgeCHDdiag;
run;
DATA heart3;
  merge h1 h2;
run;
title 'Age at CHD Diagnosed By Cholesterol Status for Female';
PROC SGPLOT DATA=heart3;
  histogram CHDhigh/fillattrs=graphdata1 name='s' legendlabel='High'
  transparency=0.5 binwidth=5;
  density CHDhigh/type=kernel legendlabel='Kernel' lineattrs=(pattern=solid);

  histogram CHDdesir/fillattrs=graphdata1 name='d' legendlabel='Desirable'
  transparency=0.5 binwidth=5;
  density CHDdesir/type=kernel legendlabel='Kernel' lineattrs=(pattern=solid);
  xaxis display=(nolabel);
Run;
```



(c)

```
* create a macro named %DrawPlot;
%MACRO DrawPlot(cate=);
  %IF &SYSDAY = Monday %THEN %DO;
    PROC SGPLOT DATA = SASHELP.Heart;
      histogram AgeAtDeath/group=&cate;
    RUN;
  %END;
  %ELSE %IF &SYSDAY = Wednesday %THEN %DO;
    PROC SGPLOT DATA = SASHELP.Heart;
      histogram AgeAtDeath/group=&cate;
    RUN;
  %END;
  %ELSE %IF &SYSDAY = Tuesday %THEN %DO;
    PROC SGPLOT DATA = SASHELP.Heart;
      VBAR AgeAtDeath/group=&cate;
    RUN;
  %END;
  %ELSE %IF &SYSDAY = Thursday %THEN %DO;
    PROC SGPLOT DATA = SASHELP.Heart;
      VBAR AgeAtDeath/group=&cate;
    RUN;
  %END;
%MEND;

* invoke the macro;
%DrawPlot(cate=Sex)
* This figure was drawn on thursday
```


80
60
40
20
0
频数
Age at Death
Sex
Female
Male