STA409

Answer to Assignment 1

12112627
李乐平

1. Solution.

For the 2 given functions, it is easy to discover that $\mu_1=\mu_2=0$.

Now we first find the variance of these 2 functions.

$$\sigma_1^2 = E(X_1-\mu_1)^2 = E(X_1^2) = \int_{-\infty}^{+\infty} x^2 f_1(x)\mathrm{d}x$$
$$= 2(\int_0^{0.9399} 0.3334x^2\mathrm{d}x + \int_{0.9399}^{2.3242} 0.2945\mathrm{d}x + 0) = 1.00$$
$$\sigma_2^2 = E(X_2-\mu_2)^2 = E(X_2^2) = \int_{-\infty}^{+\infty} x^2 f_2(x)\mathrm{d}x$$
$$= 2(\int_0^{2.4495} 0.4082x^2 - 0.1667x^2 \,|\, x \,|\mathrm{d}x) = 1.00$$

By the definition of population kurtosis $\frac{E(X-\mu)^4}{\sigma^4}$, we can respectively compute the kurtosis of 2 functions as follows.

Kurtosis of f_1:

$$\frac{E(X_1-\mu_1)^4}{\sigma_1^4} = \frac{E(X_1^4)}{1^2} = \int_{-\infty}^{+\infty} x^4 f_1(x)\mathrm{d}x$$
$$= 2(\int_0^{0.9399} 0.3334x^4\mathrm{d}x + \int_{0.9399}^{2.3242} 0.2945x^2\mathrm{d}x) = 2.40$$

Kurtosis of f_2:

$$\frac{E(X_2-\mu_2)^4}{\sigma_2^4} = E(X_2^4) = \int_{-\infty}^{+\infty} x^4 f_2(x)\mathrm{d}x$$
$$= 2(\int_0^{2.4495} 0.4082x^4 - 0.1667x^4 \,|\, x \,|\mathrm{d}x) = 2.40$$

We can discover that though the shapes of 2 distributions are different, but their variances and kurtosis are the same. Actually, kurtosis is a measure of whether data are heavy-tailed or light-tailed relative to a normal distribution, that is to say, it measures the probability of outliers present in a distribution.

2. Solution.

Kendall's τ is computed as

$$\begin{aligned}\tau &= (\#\text{concordant pairs} - \#\text{discordant pairs}) / (\#\text{pairs}) \\ &= 1 - 2\,(\#\text{discordant pairs} / \#\text{pairs}) \\ &= 1 - 2 * (5 / 45) \\ &= 7 / 9\end{aligned}$$

Though the rankings of 2 TAs given are totally different, the Kendall's τ holds a relative high value, showing a high consistency exists between 2 TAs.

3. Solution.

The equation expressed by the assumption of MCAR is given by

$$\Pr(M=1 \,|\, X,Y) = \Pr(M=1)$$

This equation shows that the probability of Y missing is not dependant to the value of X and Y.

The equation expressed by the assumption of MAR is given by

$$\Pr(M=1|X,Y)=\Pr(M=1|X)$$

This equation shows that the probability of Y missing is not dependant to the value of Y when X is given, but may be dependant to the value of X.

4. The output is as follows.

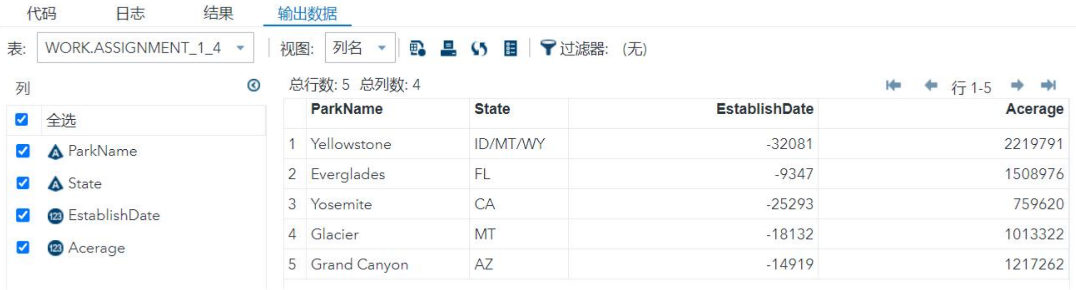


表: WORK.ASSIGNMENT_1_4

视图: 列名

列

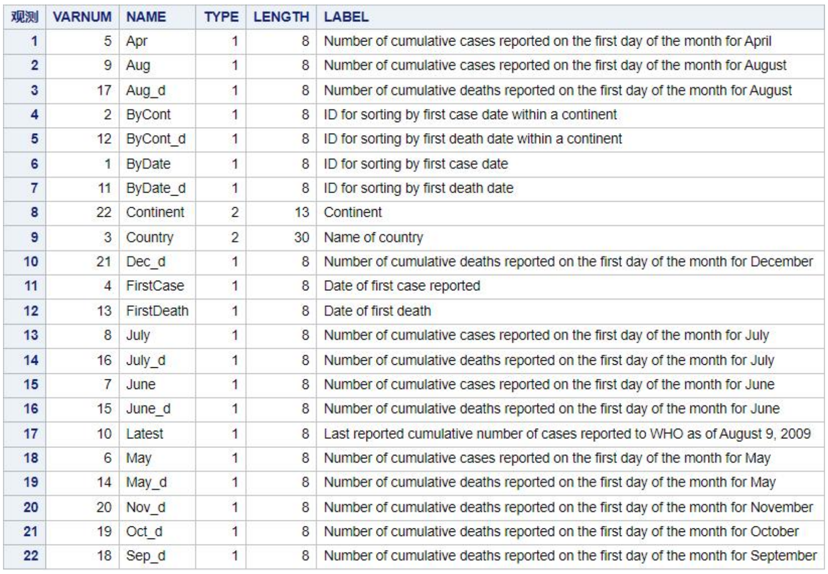
☑ 全选
☑ ParkName
☑ State
☑ EstablishDate
☑ Acerage

总行数: 5 总列数: 4

	ParkName	State	EstablishDate	Acerage
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

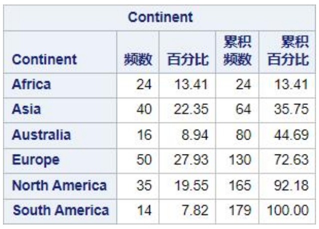
Figure 4.1: Formatted Table of "NationalPark.txt"

5. (1) The output is as follows.



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

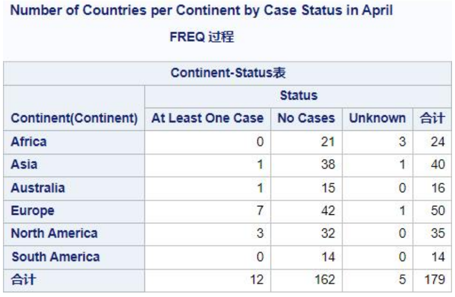
Figure 5.(1).1: Information of Attributes in Table "sff.sas7bdat"



Continent				
Continent	频数	百分比	累积 频数	累积 百分比
Africa	24	13.41	24	13.41
Asia	40	22.35	64	35.75
Australia	16	8.94	80	44.69
Europe	50	27.93	130	72.63
North America	35	19.55	165	92.18
South America	14	7.82	179	100.00

Figure 5.(1).2: Number of Countries per Continent

5. (2) The output is as follows.



Number of Countries per Continent by Case Status in April

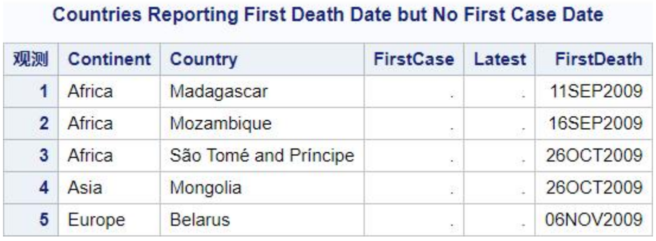
FREQ 过程

Continent-Status表

	Status			
Continent(Continent)	At Least One Case	No Cases	Unknown	合计
Africa	0	21	3	24
Asia	1	38	1	40
Australia	1	15	0	16
Europe	7	42	1	50
North America	3	32	0	35
South America	0	14	0	14
合计	12	162	5	179

Figure 5.(2).1: Number of Countries per Continent by Case Status in April

5. (3) The output is as follows.

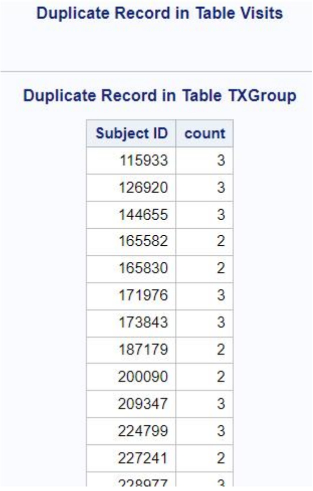


Countries Reporting First Death Date but No First Case Date

观测	Continent	Country	FirstCase	Latest	FirstDeath
1	Africa	Madagascar	.	.	11SEP2009
2	Africa	Mozambique	.	.	16SEP2009
3	Africa	São Tomé and Príncipe	.	.	26OCT2009
4	Asia	Mongolia	.	.	26OCT2009
5	Europe	Belarus	.	.	06NOV2009

Figure 5.(3).1: Countries Reporting First Death Date but No First Case Date

6. (1) The output is as follows. As the result shows, there is dulicate records in table "txgroup" while there is no duplicate record in table "visits".



Duplicate Record in Table Visits

Duplicate Record in Table TXGroup

Subject ID	count
115933	3
126920	3
144655	3
165582	2
165830	2
171976	3
173843	3
187179	2
200090	2
209347	3
224799	3
227241	2
228977	3

Figure 6.(1).1: Duplicate Records in Tables "visits.sas7bdat" and "txgroup.sas7bdat"

6. (2) The output is as follows.

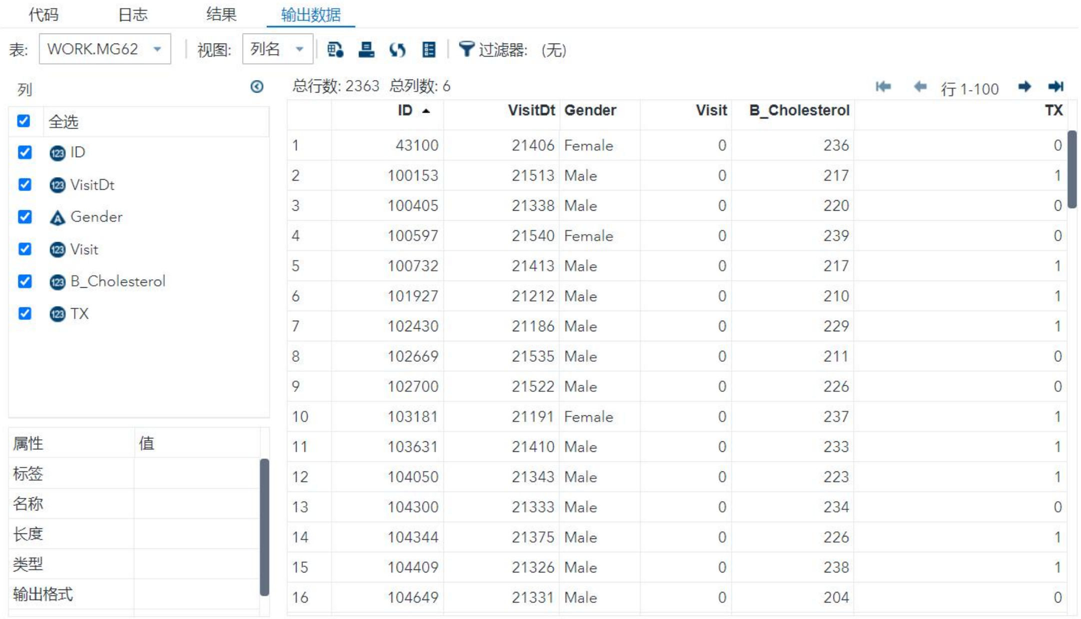


表: WORK.MG62

视图: 列名

过滤器: (无)

总行数: 2363 总列数: 6

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

属性	值
标签	
名称	
长度	
类型	
输出格式	

	ID	VisitDt	Gender	Visit	B_Cholesterol	TX
1	43100	21406	Female	0	236	0
2	100153	21513	Male	0	217	1
3	100405	21338	Male	0	220	0
4	100597	21540	Female	0	239	0
5	100732	21413	Male	0	217	1
6	101927	21212	Male	0	210	1
7	102430	21186	Male	0	229	1
8	102669	21535	Male	0	211	0
9	102700	21522	Male	0	226	0
10	103181	21191	Female	0	237	1
11	103631	21410	Male	0	233	1
12	104050	21343	Male	0	223	1
13	104300	21333	Male	0	234	0
14	104344	21375	Male	0	226	1
15	104409	21326	Male	0	238	1
16	104649	21331	Male	0	204	0

Figure 6.(2).1: Table Merged by Tables "visits" and "txgroup"

6. (3) The output is as follows.

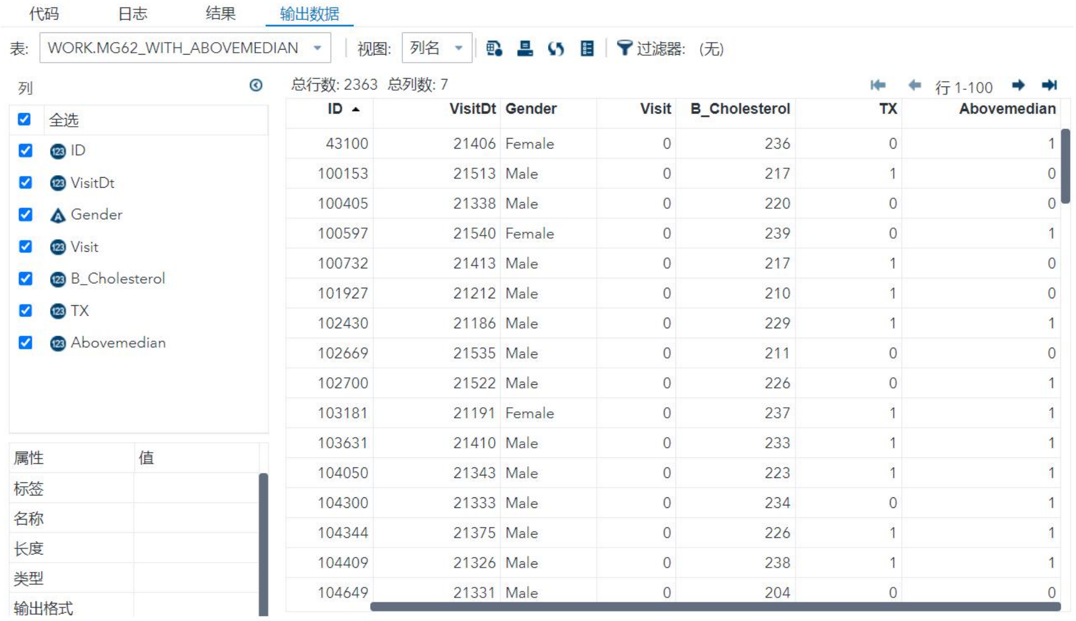


表: WORK.MG62_WITH_ABOVEMEDIAN

视图: 列名

过滤器: (无)

总行数: 2363 总列数: 7

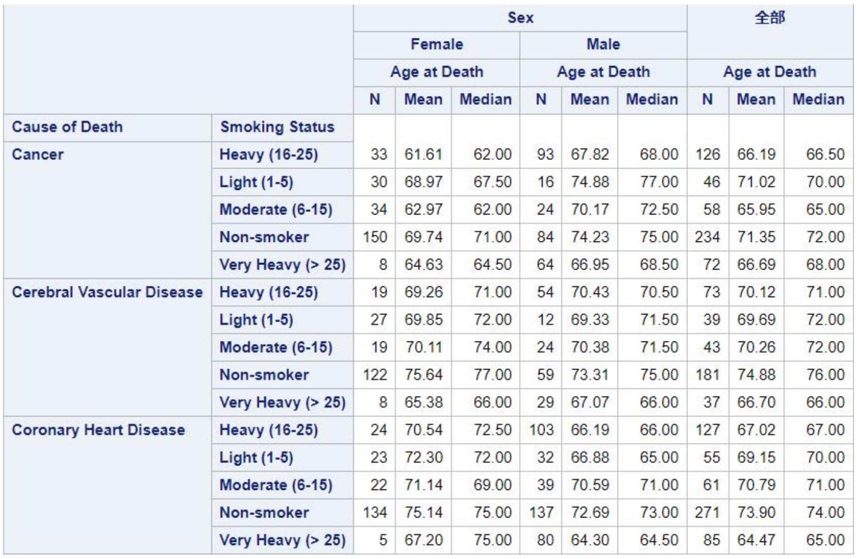
☑ 全选
☑ ID
☑ VisitDt
☑ Gender
☑ Visit
☑ B_Cholesterol
☑ TX
☑ Abovemedian

属性	值
标签	
名称	
长度	
类型	
输出格式	

ID	VisitDt	Gender	Visit	B_Cholesterol	TX	Abovemedian
43100	21406	Female	0	236	0	1
100153	21513	Male	0	217	1	0
100405	21338	Male	0	220	0	0
100597	21540	Female	0	239	0	1
100732	21413	Male	0	217	1	0
101927	21212	Male	0	210	1	0
102430	21186	Male	0	229	1	1
102669	21535	Male	0	211	0	0
102700	21522	Male	0	226	0	1
103181	21191	Female	0	237	1	1
103631	21410	Male	0	233	1	1
104050	21343	Male	0	223	1	1
104300	21333	Male	0	234	0	1
104344	21375	Male	0	226	1	1
104409	21326	Male	0	238	1	1
104649	21331	Male	0	204	0	0

Figure 6.(3).1: Merged Table with "Abovemedian" Attribute

7. (1) The output is as follows.



		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

Figure 7.(1).1: Tabulation of Several Attributes in Table "Heart"

7. (2) The output is as follows.

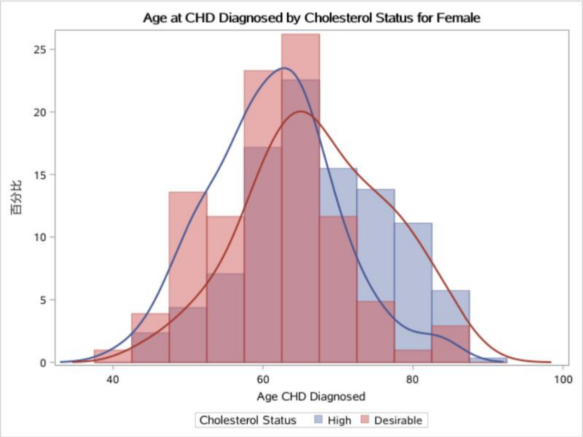
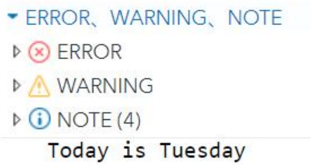


Figure 7.(2).1: Age at CHD Diagnosed by Cholesterol Status for Female

7. (3) The output is as follows. The macro function will first print the weekday, then draw the plot corresponding to the weekday.



```
Today is Tuesday
```

Figure 7.(3).1: The Weekday the Snapshot been Generated

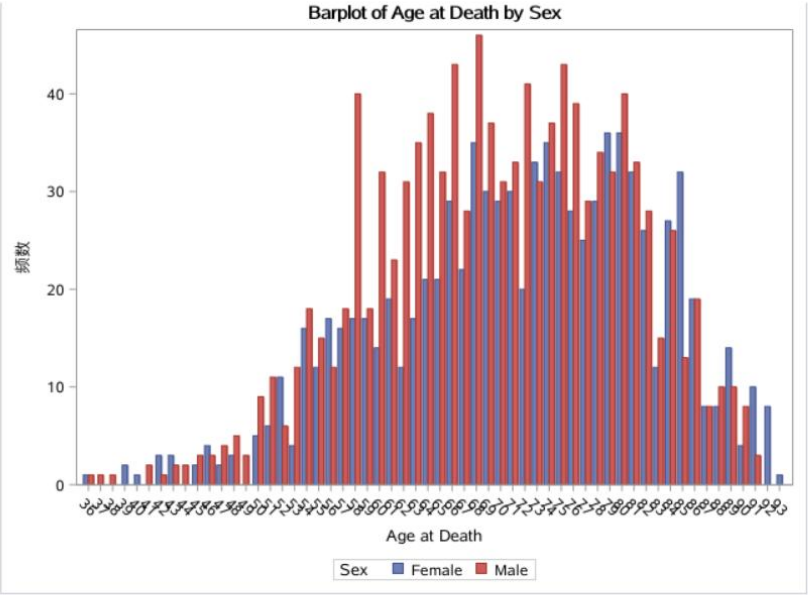


Figure 7.(3).2: The Barplot Generated at Tuesday

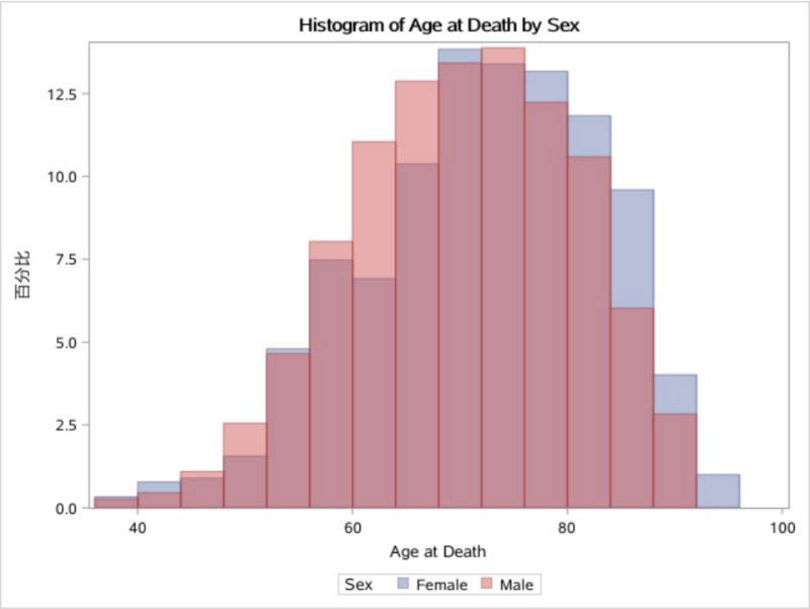


Figure 7.(3).3: The Histogram Generated at Wedensday