

实验概述：

【实验目的及要求】

实验目的：通过本实验项目，使学生掌握通过SAS过程步完成数据集的初步整理和统计初步分析，能够输出常见的描述统计和统计图表。在数据分析过程中掌握基本作图、格式加工等辅助功能。

一、下述实验数据是关于各个国家的人口数据集，从dataset for experiment 3.txt文件获得这些数据集，可以考虑用一个 infile语句完成，并用下面的输入语句来读取数据。

```
input country $20. birthrat deathrat inf_mort life_exp popurban perc_gnp  
lev_tech civillib;
```

(1) 创建一个数据集取名为world，创建如下的分类变量

Variable	Groupings	Category Variable
Infant mortality	< 24 = 1 (low) 24 – 73 = 2 (moderate) ≥ 74 = 3 (high)	infgrp
Level of technology	< 24 = 1 (low) ≥ 24 = 2 (high)	techgrp
Degree of civil liberties	1, 2 = 1 (low degree of denial) 3, 4, 5 = 2 (moderate degree of denial) 6, 7 = 3 (high degree of denial)	civilgrp

用libname语句创建自己的逻辑库，并将生成的数据集保存在该库中。

(2)计算birthrate、deathrate和popurban的33.3%和66.7%分位数，并将这些统计量保存在一个临时SAS数据集中，命名为stats。然后打印该数据集。

(3)

利用(2)中的分位数，分别对world数据集中的变量birthrate、deathrate和popurban进行分组，并相应创建三个分类变量birthgrp、deathgrp、popgrp，将数据集命名为world2并保存在同一逻辑库中。

(4)写一个新的SAS程序，利用proc

univariate过程步计算三个变量life_exp、perc_gnp、popurban的描述性统计量、极值、分位数和箱线图；选一个合适的t-

test检验这三个变量的总体参数分别为大于70年、低于3000\$、超过70%，用程序给出的p值进行检验并给出结论，同时计算这三个参数的95%的置信区间

; 做直方图分析三个变量的分布形态，并对每个变量做正态性检验，给出你的结论。

(5)根据分类变量birthgrp和popgrp分成的9个组合类别分别计算birthrate、deathrate、inf_mort三个变量的样本均值、标准差、最大与最小值，并将其输出到一个新的数据集stats1。同时添加一个proc
format过程来定义这些分类变量的输出格式，打印出数据集stats1。

(6)写一个新的SAS程序用(3)中数据world2在一个proc freq中完成下列要求：

a. 做出 2×2 的频数表格，

techgrp*infgrp、techgrp*civilgrp、popgrp*infgrp，计算卡方统计值、cell χ^2 、
cell期望值，但是不要输出行、列和单元格的占比百分数。

b. 用卡方检验统计量来检验a中几对变量的独立性

c.

考虑用每个国家的popgrp来预测infgrp的值，选用哪个值来度量二者之间的相关性程度合适？说明理由。

(7)写一个新的SAS程序利用(3)中的数据集world2，利用proc
gchart来完成下列的散点图

a. 计算根据变量infgrp作为y轴上分类，techgrp变量作为次要分类而分成组的
每个样本的mean life expectanc的水平条形图

b. 选择合适的midpoints做关于per capita

GNP的垂直条形图，用频数做为统计量并用变量civilgrp进行分组

二、数据集jisu为消化内科病人临床试验的数据，研究某种激素水平对于胃癌发生的作用。含有性别、萎缩程度、肠化程度以及激素水平等指标。

其中萎缩分为轻度（1）、中度（2）和重度（3）三级；

肠化分为无（1）、轻度（2）、中度（3）和重度（4）四级；

性别为男（1），女（2）；

年龄分组为：青年组、中年组和老人组。

(1) 创建你的逻辑库，将数据集jisu放入你的库中。

(2) 使用univariate过程步对激素水平进行分析，画出该指标的直方图和正态的
QQ图，并作正态性检验，试问该指标是否服从正态分布。

(3) 对激素水平进行对数化保存到新指标lnjisu中，将生成的新数据集保存到jis
su1中，验证lnjisu是否服从正态分布。数据对数变换是Box-Cox
变换的一种特殊情况，查询资料回答什么是Box-Cox
变换？数据的对数化处理是否存在缺点？

(4) 求出Injisu的33.33%和66.67%分位数，并以求出的两个分位数作为分界，将Injisu分成三类保存到分类变量jisugrp中，生成的新数据集保存到jisu2中。对jisugrp的三类分别用高、中和低标签的值进行格式化。

(5) 分别做年龄分组、肠化程度、萎缩程度与激素水平分组的列联表分析，检验他们之间的关联性，给出你的结论。

三、某项政策颁布后，通过调查得到了下面所示的列联表，其中显示了不同收入和性别的人对这一项政策的观点。

收入\观点	反对	支持	合计
低	9	92	101
中	31	53	84
高	40	25	65
合计	80	170	250

给定显著性水平为0.05，检验的假设为：

H0：观点和收入不显著相关

H1：观点和收入显著相关

利用已学知识和所给数据，给出分析过程以及所得的结论。

四、数据集sale是一份航空公司月度客运量的数据，其中包含两个变量date和sale。试完成以下工作。

(1) 选用合适的工具和变量进行描述统计分析，了解该数据集中月度数据的缺失情况。

(2) 试用数据步将月度客运量数据转换为年度数据；

(3) 利用Freq和means过程步将月度数据转化为年度数据；

(4) 用gplot绘制年度客运量数据的折线图，用gchart绘制数据集中每年包含月份数的条形图，用sgplot将这两种图形叠加在一张图上。

(5) 如果由你负责数据搜集和整理工作，请用sas编程把数据缺失的月份找出来，列出一份清单，以便向企业索要数据。

【实验原理】

一、通过means和univariate过程步完成描述统计分析，并选取适当的选项和参数完成作图和检验步骤。

二、通过univariate过程步完成数据的正态性检验包括直方图、QQ图和系列正态性检验法。

三、通过FREQ过程步完成列联表数据的汇总和关联性检验。

四、通过gplot、gchart和sgplot过程步完成SAS作图。

【实验环境】（使用的软硬件）

硬件：IBM PC 或其兼容机

软件：Microsoft Windows, Microsoft Word 2003 或更高版本, SAS 8.x 或更高版本.

注：上述内容同学都不要改动，请附上相应代码（可以截图）和步骤描述，具体结果可以整理好以图片或其他格式，原则为简洁整齐。

实验内容：

第一题

(1) 建一个数据集取名为world，创建如下的分类变量

```
LIBNAME MY_DATA "/home/u61827597/Sufe_SAS/Assignments";
DATA MY_DATA.raw_data;
  INFILE "/home/u61827597/Sufe_SAS/Assignments/hw3/dataset
for experiment 3.txt";
  INPUT country $20. birthrat deathrat inf_mort life_exp
popurban perc_gnp lev_tech civillib;
RUN;
DATA MY_DATA.world;
  SET MY_DATA.raw_data;
  IF inf_mort < 24 THEN infgrp = 1;
  ELSE IF inf_mort < 73 THEN infgrp = 2;
  ELSE IF inf_mort >= 73 THEN infgrp = 3;
  IF lev_tech < 24 THEN techgrp = 1;
  ELSE IF lev_tech >= 24 THEN techgrp = 2;
  IF civillib IN (1 , 2) THEN civilgrp = 1;
  ELSE IF civillib IN (3 , 4, 5) THEN civilgrp = 2;
  ELSE IF civillib IN (6 ,7) THEN civilgrp = 3;
  *DROP inf_mort civillib lev_tech;
RUN;
```

观测	country	birthrat	deathrat	inf_mort	life_exp	popurban	perc_gnp	lev_tech	civillib	infrgr	techgrp	civilgrp
1	ALGERIA	45	12	109	60	52	2400	17	6	3	1	3
2	ARGENTINA	24	8	35	70	83	2030	23	3	2	1	2
3	AUSTRALIA	16	7	10	75	86	9210	71	1	1	2	1
4	AUSTRIA	12	12	12	73	56	9210	50	1	1	2	1
5	BOLIVIA	42	16	124	51	46	510	10	3	3	1	2
6	BRAZIL	31	8	71	63	68	1890	15	3	2	1	2
7	BULGARIA	14	11	17	72	65	3900	44	7	1	2	3
8	CANADA	15	7	9	75	76	12000	75	1	1	2	1
9	CHILE	24	6	24	70	83	1870	22	5	2	1	2
10	COLOMBIA	28	7	53	64	67	1410	15	3	2	1	2
11	CZECHOSLOVAKIA	15	12	16	71	74	5800	72	6	1	2	3
12	DENMARK	10	11	8	74	83	11490	71	1	1	2	1
13	EGYPT	37	10	80	57	44	700	13	5	3	1	2
14	FINLAND	14	9	6	74	60	10440	57	2	1	2	1
15	FRANCE	14	10	9	75	73	11390	62	2	1	2	1
16	GHANA	47	15	107	52	40	320	10	5	3	1	2
17	GREECE	14	9	15	74	70	3970	23	2	1	1	1
18	HUNGARY	12	14	19	70	54	2150	49	5	1	2	2
19	ITALY	11	10	12	74	72	6350	41	2	1	2	1
20	INDIA	34	13	118	53	23	260	6	3	3	1	2
21	IRAQ	46	13	72	59	68	3400	12	7	2	1	3
22	IRELAND	19	9	11	73	56	4810	48	1	1	2	1
23	ISRAEL	24	6	14	74	87	5360	33	2	1	2	1
24	IVORY COAST	46	18	122	47	42	720	9	5	3	1	2
25	JAPAN	13	6	6	77	76	10100	53	1	1	2	1
26	KENYA	54	13	80	53	16	340	11	5	3	1	2
27	MADAGASCAR	45	17	67	50	22	290	12	6	2	1	3
28	MALAWI	52	20	165	45	12	210	12	7	3	1	3
29	MALAYSIA	29	7	29	67	32	1870	14	4	2	1	2
30	MOROCCO	41	12	99	58	42	750	12	5	3	1	2
31	NETHERLANDS	12	8	8	76	88	9910	68	1	1	2	1
32	NEW ZEALAND	16	8	13	74	83	7410	66	1	1	2	1
33	NIGERIA	48	17	105	50	28	760	8	3	3	1	2
34	NORWAY	12	10	8	76	71	13820	63	1	1	2	1
35	PAKISTAN	43	15	120	50	19	390	8	5	3	1	2
36	PERU	35	10	99	59	65	1040	12	3	3	1	2
37	PHILIPPINES	32	7	50	64	37	760	15	5	2	1	2
38	POLAND	20	10	19	71	59	4200	53	5	1	2	2
39	PORTUGAL	14	9	20	71	30	2190	22	2	1	1	1
40	ROMANIA	15	10	28	71	49	2200	33	6	2	2	3
41	SENEGAL	50	19	141	43	42	440	11	4	3	1	2
42	SOUTH AFRICA	35	14	92	54	56	2450	33	6	3	2	3
43	SPAIN	13	7	10	74	91	4800	28	2	1	2	1
44	SRI LANKA	27	6	34	68	22	330	9	4	2	1	2
45	SWEDEN	11	11	7	76	83	12400	81	1	1	2	1
46	SWITZERLAND	11	9	8	76	58	16390	57	1	1	2	1
47	SYRIA	47	7	57	64	47	1680	16	7	2	1	3
48	THAILAND	25	6	51	63	17	810	12	4	2	1	2
49	TOGO	45	17	113	49	20	280	15	6	3	1	3
50	TUNISIA	33	10	85	61	52	1290	15	5	3	1	2
51	TURKEY	35	10	110	63	45	1230	14	5	3	1	2
52	USSR	20	10	32	69	64	6350	59	7	2	2	3
53	UNITED KINGDOM	13	12	10	73	76	9050	61	1	1	2	1
54	UNITED STATES	16	9	11	75	74	14090	100	1	1	2	1
55	URUGUAY	18	9	32	69	84	2490	20	4	2	1	2
56	VENEZUELA	33	6	39	69	76	4100	25	2	2	2	1
57	WEST GERMANY	10	11	10	74	94	11420	66	2	1	2	1
58	YUGOSLAVIA	17	10	32	70	46	2570	23	5	2	1	2
59	ZAIRE	45	16	106	50	34	160	10	7	3	1	3
60	ZAMBIA	48	15	101	51	43	580	12	6	3	1	3

(2)计算birthrate、deathrate和popurban的33.3%和66.7%分位数，并将这些统计量保存在一个临时SAS数据集中，命名为stats。然后打印该数据集。

PROC UNIVARIATE DATA = my_data.world NOPRINT;

VAR birthrat deathrat popurban;

```

OUTPUT OUT = stats
PCTLPTS = 33.3 66.7
PCTLPRE = birthrat deathrat popurban;
RUN;
PROC PRINT DATA=stats;
RUN;

```

观测	birthrat33_3	deathrat33_3	popurban33_3	birthrat66_7	deathrat66_7	popurban66_7
1	15	9	45	35	12	71

(3) 利用(2)中的分位数, 分别对 world 数据集中的变量 birthrate、deathrate 和 popurban 进行分组, 并相应创建三个分类变量 birthgrp、deathgrp、popgrp, 将数据集命名为 world2 并保存在同一逻辑库中。

```

DATA MY_DATA.world2;
SET MY_DATA.raw_data;
IF birthrat < 15 THEN birthgrp = 1;
ELSE IF birthrat < 35 THEN birthgrp=2;
ELSE IF birthrat >= 35 THEN birthgrp=3;
IF deathrat < 9 THEN deathgrp = 1;
ELSE IF deathrat < 12 THEN deathgrp=2;
ELSE IF deathrat >= 12 THEN deathgrp=3;
IF popurban < 45 THEN popgrp = 1;
ELSE IF popurban < 71 THEN popgrp = 2;
ELSE IF popurban >= 71 THEN popgrp = 3;
RUN;
PROC PRINT DATA=MY_DATA.world2;
RUN;

```

观测	country	birthrat	deathrat	inf_mort	life_exp	popurban	perc_gnp	lev_tech	civilib	birthgrp	deathgrp	popgrp
1	ALGERIA	45	12	109	60	52	2400	17	6	3	3	2
2	ARGENTINA	24	8	35	70	83	2030	23	3	2	1	3
3	AUSTRALIA	16	7	10	75	86	9210	71	1	2	1	3
4	AUSTRIA	12	12	12	73	56	9210	50	1	1	3	2
5	BOLIVIA	42	16	124	51	46	510	10	3	3	3	2
6	BRAZIL	31	8	71	63	68	1890	15	3	2	1	2
7	BULGARIA	14	11	17	72	65	3900	44	7	1	2	2
8	CANADA	15	7	9	75	76	12000	75	1	2	1	3
9	CHILE	24	6	24	70	83	1870	22	5	2	1	3
10	COLOMBIA	28	7	53	64	67	1410	15	3	2	1	2
11	CZECHOSLOVAKIA	15	12	16	71	74	5800	72	6	2	3	3
12	DENMARK	10	11	8	74	83	11490	71	1	1	2	3
13	EGYPT	37	10	80	57	44	700	13	5	3	2	1
14	FINLAND	14	9	6	74	60	10440	57	2	1	2	2
15	FRANCE	14	10	9	75	73	11390	62	2	1	2	3
16	GHANA	47	15	107	52	40	320	10	5	3	3	1
17	GREECE	14	9	15	74	70	3970	23	2	1	2	2
18	HUNGARY	12	14	19	70	54	2150	49	5	1	3	2
19	ITALY	11	10	12	74	72	6350	41	2	1	2	3
20	INDIA	34	13	118	53	23	260	6	3	2	3	1
21	IRAQ	46	13	72	59	68	3400	12	7	3	3	2
22	IRELAND	19	9	11	73	56	4810	48	1	2	2	2
23	ISRAEL	24	6	14	74	87	5360	33	2	2	1	3
24	IVORY COAST	46	18	122	47	42	720	9	5	3	3	1
25	JAPAN	13	6	6	77	76	10100	53	1	1	1	3
26	KENYA	54	13	80	53	16	340	11	5	3	3	1
27	MADAGASCAR	45	17	67	50	22	290	12	6	3	3	1
28	MALAWI	52	20	165	45	12	210	12	7	3	3	1
29	MALAYSIA	29	7	29	67	32	1870	14	4	2	1	1
30	MOROCCO	41	12	99	58	42	750	12	5	3	3	1
31	NETHERLANDS	12	8	8	76	88	9910	68	1	1	1	3
32	NEW ZEALAND	16	8	13	74	83	7410	66	1	2	1	3
33	NIGERIA	48	17	105	50	28	760	8	3	3	3	1
34	NORWAY	12	10	8	76	71	13820	63	1	1	2	3
35	PAKISTAN	43	15	120	50	19	390	8	5	3	3	1
36	PERU	35	10	99	59	65	1040	12	3	3	2	2
37	PHILIPPINES	32	7	50	64	37	760	15	5	2	1	1
38	POLAND	20	10	19	71	59	4200	53	5	2	2	2
39	PORTUGAL	14	9	20	71	30	2190	22	2	1	2	1
40	ROMANIA	15	10	28	71	49	2200	33	6	2	2	2
41	SENEGAL	50	19	141	43	42	440	11	4	3	3	1
42	SOUTH AFRICA	35	14	92	54	56	2450	33	6	3	3	2
43	SPAIN	13	7	10	74	91	4800	28	2	1	1	3
44	SRI LANKA	27	6	34	68	22	330	9	4	2	1	1
45	SWEDEN	11	11	7	76	83	12400	81	1	1	2	3
46	SWITZERLAND	11	9	8	76	58	16390	57	1	1	2	2
47	SYRIA	47	7	57	64	47	1680	16	7	3	1	2
48	THAILAND	25	6	51	63	17	810	12	4	2	1	1
49	TOGO	45	17	113	49	20	280	15	6	3	3	1
50	TUNISIA	33	10	85	61	52	1290	15	5	2	2	2
51	TURKEY	35	10	110	63	45	1230	14	5	3	2	2
52	USSR	20	10	32	69	64	6350	59	7	2	2	2
53	UNITED KINGDOM	13	12	10	73	76	9050	61	1	1	3	3
54	UNITED STATES	16	9	11	75	74	14090	100	1	2	2	3
55	URUGUAY	18	9	32	69	84	2490	20	4	2	2	3
56	VENEZUELA	33	6	39	69	76	4100	25	2	2	1	3
57	WEST GERMANY	10	11	10	74	94	11420	66	2	1	2	3
58	YUGOSLAVIA	17	10	32	70	46	2570	23	5	2	2	2
59	ZAIRE	45	16	106	50	34	160	10	7	3	3	1
60	ZAMBIA	48	15	101	51	43	580	12	6	3	3	1

(4) 写一个新的 SAS 程序，利用 proc univariate 过程步计算三个变量 life_exp、perc_gnp、popurban 的描述性统计量、极值、分位数和箱线图；选一个合适的 t-test 检验这三个变量的总体参数分别为大于 70 年、低于 3000\$、超过 70%，用程序给出的 p 值进行检验并给出结论，同时计算这三个参数的 95%

的置信区间；做直方图分析三个变量的分布形态，并对每个变量做正态性检验，给出你的结论。

```

PROC PRINT DATA=MY_DATA.raw_data;
RUN;

PROC UNIVARIATE DATA=MY_DATA.raw_data PLOT NORMAL;
  VAR life_exp perc_gnp popurban;
  HISTOGRAM life_exp perc_gnp popurban;
  PROBPLOT life_exp perc_gnp popurban;
RUN;

PROC TTEST DATA=MY_DATA.raw_data H0=70 SIDE=U;
  VAR life_exp;
RUN;

PROC TTEST DATA=MY_DATA.raw_data H0=3000 SIDE=L;
  VAR perc_gnp;
RUN;

PROC TTEST DATA=MY_DATA.raw_data H0=70 SIDE=U;
  VAR popurban;
RUN;

```

UNIVARIATE 过程 变量: life_exp	
矩	
数目	60
均值	65.05
标准差	9.9121993
偏度	-0.658332
未校平方和	259687
变异系数	15.237816
权重总和	60
观测总和	3903
方差	98.2516949
峰度	-0.9408622
校正平方和	5796.85
标准误差均值	1.27965943

基本统计测度	
位置	变异性
均值	65.05000
中位数	69.00000
众数	74.00000
四分位间距	16.50000
标准差	9.91220
方差	98.25169
极差	34.00000

位置检验: Mu0=0			
检验	统计量	p 值	
Student t	t	50.83384	Pr > t <.0001
符号	M	30	Pr >= M <.0001
符号秩	S	915	Pr >= S <.0001

分位数 (定义 5)	
水平	分位数
100% 最大值	77.0
99%	77.0
95%	76.0
90%	75.0
75% Q3	74.0
50% 中位数	69.0
25% Q1	57.5
10%	50.0
5%	48.0
1%	43.0
0% 最小值	43.0

极值观测			
最小值	最大值		
值	观测	值	观测
43	41	76	31
45	28	76	34
47	24	76	45
49	49	76	46
50	59	77	25

UNIVARIATE 过程 变量: perc_gnp			
矩			
数目	60	权重总和	60
均值	4345.66667	观测总和	260740
标准差	4442.41359	方差	19735038.5
偏度	1.04589858	峰度	-0.1099338
未校平方和	2297456400	校正平方和	1164367273
变异系数	102.226285	标准误差均值	573.513129

基本统计测度			
位置		变异性	
均值	4345.667	标准差	4442
中位数	2300.000	方差	19735039
众数	760.000	极差	16230
		四分位间距	6125

注意: 显示的众数是 4 个众数中最小的众数, 其计数为 2。

位置检验: Mu0=0			
检验		统计量	p 值
Student t	t	7.577275	Pr > t
符号	M	30	Pr >= MI
符号秩	S	915	Pr >= ISI

UNIVARIATE 过程 变量: popurban			
矩			
数目	60	权重总和	60
均值	56.35	观测总和	3381
标准差	22.3613051	方差	500.027966
偏度	-0.2660574	峰度	-0.9754078
未校平方和	220021	校正平方和	29501.65
变异系数	39.682884	标准误差均值	2.88683208

基本统计测度			
位置		变异性	
均值	56.35000	标准差	22.36131
中位数	57.00000	方差	500.02797
众数	83.00000	极差	82.00000
		四分位间距	33.00000

位置检验: Mu0=0			
检验		统计量	p 值
Student t	t	19.51967	Pr > t
符号	M	30	Pr >= MI
符号秩	S	915	Pr >= ISI

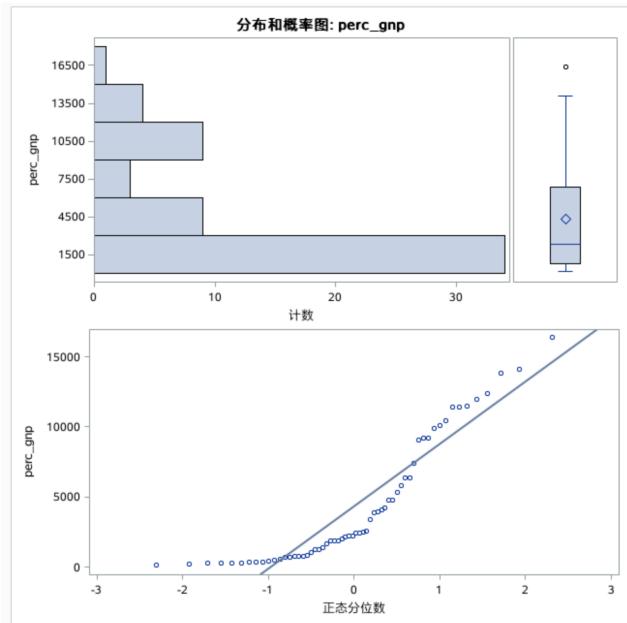
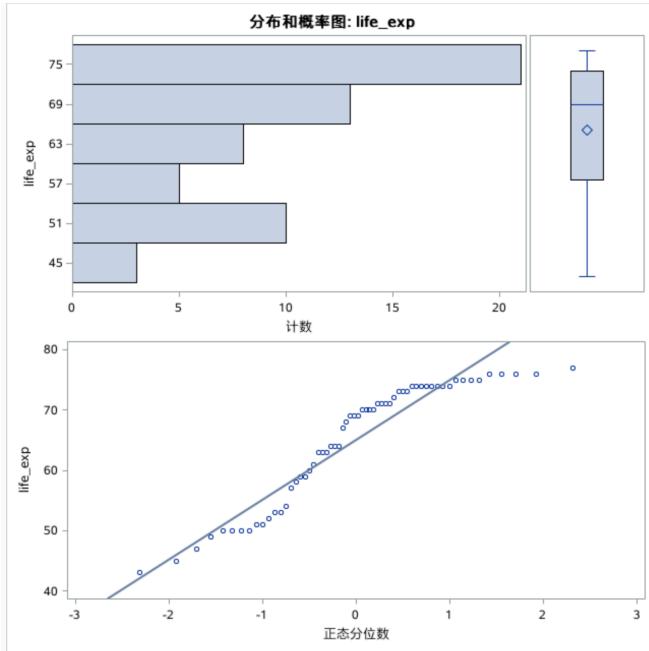
分位数 (定义 5)			
水平		分位数	
100% 最大值		16390	
99%		16390	
95%		13110	
90%		11455	
75% Q3		6880	
50% 中位数		2300	
25% Q1		755	
10%		325	
5%		270	
1%		160	
0% 最小值		160	

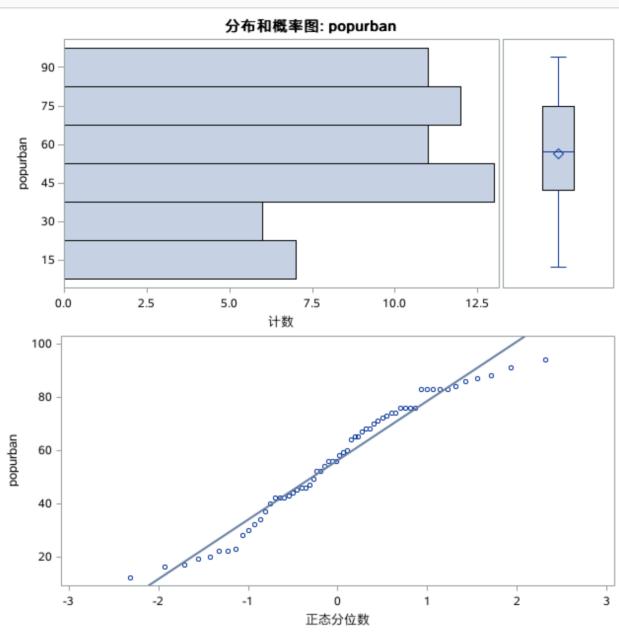
极值观测			
最小值		最大值	
值	观测	值	观测
160	59	12000	8
210	28	12400	45
260	20	13820	34
280	49	14090	54
290	27	16390	46

分位数 (定义 5)			
水平		分位数	
100% 最大值		94.0	
99%		94.0	
95%		87.5	
90%		83.5	
75% Q3		75.0	
50% 中位数		57.0	
25% Q1		42.0	
10%		22.0	
5%		18.0	
1%		12.0	
0% 最小值		12.0	

极值观测			
最小值		最大值	
值	观测	值	观测
12	28	86	3
16	26	87	23
17	48	88	31
19	35	91	43
20	49	94	57

<p>TTEST 过程</p> <p>变量: perc_gnp</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>数目</th> <th>均值</th> <th>标准差</th> <th>标准误差</th> <th>最小值</th> <th>最大值</th> </tr> </thead> <tbody> <tr> <td>60</td> <td>4345.7</td> <td>4442.4</td> <td>573.5</td> <td>160.0</td> <td>16390.0</td> </tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>均值</th> <th>95% 置信均值</th> <th>标准差</th> <th>95% 置信限标准差</th> </tr> </thead> <tbody> <tr> <td>4345.7</td> <td>-lnfty</td> <td>5304.1</td> <td>4442.4</td> </tr> <tr> <td></td> <td></td> <td></td> <td>3765.5</td> </tr> <tr> <td></td> <td></td> <td></td> <td>5418.2</td> </tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>自由度</th> <th>t 值</th> <th>Pr < t</th> </tr> </thead> <tbody> <tr> <td>59</td> <td>2.35</td> <td>0.9888</td> </tr> </tbody> </table>	数目	均值	标准差	标准误差	最小值	最大值	60	4345.7	4442.4	573.5	160.0	16390.0	均值	95% 置信均值	标准差	95% 置信限标准差	4345.7	-lnfty	5304.1	4442.4				3765.5				5418.2	自由度	t 值	Pr < t	59	2.35	0.9888	<p>由于 $p = 0.9888$, 因此接受原假设, 即不拒绝总体参数低于 3000 的原假设。</p> <p>95%置信区间 (-∞, 5304.1)</p>
数目	均值	标准差	标准误差	最小值	最大值																														
60	4345.7	4442.4	573.5	160.0	16390.0																														
均值	95% 置信均值	标准差	95% 置信限标准差																																
4345.7	-lnfty	5304.1	4442.4																																
			3765.5																																
			5418.2																																
自由度	t 值	Pr < t																																	
59	2.35	0.9888																																	
<p>TTEST 过程</p> <p>变量: popurban</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>数目</th> <th>均值</th> <th>标准差</th> <th>标准误差</th> <th>最小值</th> <th>最大值</th> </tr> </thead> <tbody> <tr> <td>60</td> <td>56.3500</td> <td>22.3613</td> <td>2.8868</td> <td>12.0000</td> <td>94.0000</td> </tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>均值</th> <th>95% 置信均值</th> <th>标准差</th> <th>95% 置信限标准差</th> </tr> </thead> <tbody> <tr> <td>56.3500</td> <td>51.5258</td> <td>lnfty</td> <td>22.3613</td> </tr> <tr> <td></td> <td></td> <td></td> <td>18.9542</td> </tr> <tr> <td></td> <td></td> <td></td> <td>27.2732</td> </tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>自由度</th> <th>t 值</th> <th>Pr > t</th> </tr> </thead> <tbody> <tr> <td>59</td> <td>-4.73</td> <td>1.0000</td> </tr> </tbody> </table>	数目	均值	标准差	标准误差	最小值	最大值	60	56.3500	22.3613	2.8868	12.0000	94.0000	均值	95% 置信均值	标准差	95% 置信限标准差	56.3500	51.5258	lnfty	22.3613				18.9542				27.2732	自由度	t 值	Pr > t	59	-4.73	1.0000	<p>由于 $p = 1$, 因此接受原假设, 即不拒绝总体参数大于 70%年的原假设。</p> <p>95%置信区间 (51.5258, +∞)</p>
数目	均值	标准差	标准误差	最小值	最大值																														
60	56.3500	22.3613	2.8868	12.0000	94.0000																														
均值	95% 置信均值	标准差	95% 置信限标准差																																
56.3500	51.5258	lnfty	22.3613																																
			18.9542																																
			27.2732																																
自由度	t 值	Pr > t																																	
59	-4.73	1.0000																																	





popurban 不服从正态分布

(5)根据分类变量 birthgrp 和 popgrp 分成的 9 个组合类别分别计算 birthrate、deathrate、inf_mort 三个变量的样本均值、标准差、最大与最小值，并将其输出到一个新的数据集 stats1。同时添加一个 proc format 过程来定义这些分类变量的输出格式，打印出数据集 stats1。

```

PROC MEANS DATA=MY_DATA.world2 MEAN STDDEV MAX MIN;
  VAR birthrat deathrat inf_mort;
  CLASS birthgrp popgrp;
  OUTPUT OUT=stats1_raw;
RUN;

DATA stats1;
  SET stats1_raw;

  IF nmiss(of _numeric_)>0 THEN
    DELETE;
  DROP _TYPE_;
  RENAME _FREQ_=freq _STAT_=stat_type;
RUN;

PROC FORMAT;
  VALUE birthfmt 1='low' 2='moderate' 3='high';
  VALUE popfmt 1='low' 2='moderate' 3='high';
RUN;

PROC PRINT DATA=stats1;
  FORMAT birthgrp birthfmt.
        popgrp popfmt.;
RUN;

```

MEANS PROCEDURE

birthgrp	popgrp	观测数	变量	均值	标准差	最大值	最小值
1	1	1	birthrat	14.0000000	.	14.0000000	14.0000000
			deathrat	9.0000000	.	9.0000000	9.0000000
			inf_mort	20.0000000	.	20.0000000	20.0000000
	2	6	birthrat	12.8333333	1.3291601	14.0000000	11.0000000
			deathrat	10.6666667	2.0655911	14.0000000	9.0000000
			inf_mort	12.8333333	5.1153364	19.0000000	6.0000000
	3	10	birthrat	11.9000000	1.3703203	14.0000000	10.0000000
			deathrat	9.6000000	1.9550504	12.0000000	6.0000000
			inf_mort	8.8000000	1.7511901	12.0000000	6.0000000
2	1	5	birthrat	29.4000000	3.6469165	34.0000000	25.0000000
			deathrat	7.8000000	2.9495762	13.0000000	6.0000000
			inf_mort	56.4000000	35.7673035	118.0000000	29.0000000
	2	8	birthrat	22.8750000	6.7915389	33.0000000	15.0000000
			deathrat	9.2500000	1.1649647	10.0000000	7.0000000
			inf_mort	41.3750000	25.8950492	85.0000000	11.0000000
	3	10	birthrat	20.1000000	5.9525905	33.0000000	15.0000000
			deathrat	7.8000000	1.8737959	12.0000000	6.0000000
			inf_mort	20.3000000	11.2945415	39.0000000	9.0000000
3	1	13	birthrat	46.2307692	4.4935852	54.0000000	37.0000000
			deathrat	15.6923077	2.8102377	20.0000000	10.0000000
			inf_mort	108.1538462	25.9738527	165.0000000	67.0000000
	2	7	birthrat	40.7142857	5.5592052	47.0000000	35.0000000
			deathrat	11.7142857	2.9840848	16.0000000	7.0000000
			inf_mort	94.7142857	23.3074687	124.0000000	57.0000000

观测	birthgrp	popgrp	freq	stat_type	birthrat	deathrat	inf_mort
1	low	low	1	N	1.0000	1.0000	1.000
2	low	low	1	MIN	14.0000	9.0000	20.000
3	low	low	1	MAX	14.0000	9.0000	20.000
4	low	low	1	MEAN	14.0000	9.0000	20.000
5	low	moderate	6	N	6.0000	6.0000	6.000
6	low	moderate	6	MIN	11.0000	9.0000	6.000
7	low	moderate	6	MAX	14.0000	14.0000	19.000
8	low	moderate	6	MEAN	12.8333	10.6667	12.833
9	low	moderate	6	STD	1.3292	2.0656	5.115
10	low	high	10	N	10.0000	10.0000	10.000
11	low	high	10	MIN	10.0000	6.0000	6.000
12	low	high	10	MAX	14.0000	12.0000	12.000
13	low	high	10	MEAN	11.9000	9.6000	8.800
14	low	high	10	STD	1.3703	1.9551	1.751
15	moderate	low	5	N	5.0000	5.0000	5.000
16	moderate	low	5	MIN	25.0000	6.0000	29.000
17	moderate	low	5	MAX	34.0000	13.0000	118.000
18	moderate	low	5	MEAN	29.4000	7.8000	56.400
19	moderate	low	5	STD	3.6469	2.9496	35.767
20	moderate	moderate	8	N	8.0000	8.0000	8.000
21	moderate	moderate	8	MIN	15.0000	7.0000	11.000
22	moderate	moderate	8	MAX	33.0000	10.0000	85.000
23	moderate	moderate	8	MEAN	22.8750	9.2500	41.375
24	moderate	moderate	8	STD	6.7915	1.1650	25.895
25	moderate	high	10	N	10.0000	10.0000	10.000
26	moderate	high	10	MIN	15.0000	6.0000	9.000
27	moderate	high	10	MAX	33.0000	12.0000	39.000
28	moderate	high	10	MEAN	20.1000	7.8000	20.300
29	moderate	high	10	STD	5.9526	1.8738	11.295
30	high	low	13	N	13.0000	13.0000	13.000
31	high	low	13	MIN	37.0000	10.0000	67.000
32	high	low	13	MAX	54.0000	20.0000	165.000
33	high	low	13	MEAN	46.2308	15.6923	108.154
34	high	low	13	STD	4.4936	2.8102	25.974
35	high	moderate	7	N	7.0000	7.0000	7.000
36	high	moderate	7	MIN	35.0000	7.0000	57.000
37	high	moderate	7	MAX	47.0000	16.0000	124.000
38	high	moderate	7	MEAN	40.7143	11.7143	94.714
39	high	moderate	7	STD	5.5592	2.9841	23.307

(6)写一个新的 SAS 程序用(3)中数据 world2 在一个 proc freq 中完成下列要求：

- a. 做出 2*2 的频数表格， techgrp*infgrp、 techgrp*civilgrp、 popgrp*infgrp，计算卡方统计值、 cell χ^2 、 cell 期望值，但是不要输出行、列和单元格的占比百分数。
- b. 用卡方检验统计量来检验 a 中几对变量的独立性
- c. 考虑用每个国家的 popgrp 来预测 infgrp 的值，选用哪个值来度量二者之间的相关性程度合适？说明理由。

```
DATA MY_DATA.world;
  SET MY_DATA.raw_data;
  IF inf_mort < 24    THEN infgrp = 1;
  ELSE IF inf_mort < 73 THEN infgrp = 2;
```

```

ELSE IF inf_mort >= 73 THEN infgrp = 3;
IF lev_tech < 24 THEN techgrp = 1;
ELSE IF lev_tech >= 24 THEN techgrp = 2;
IF civillib IN (1, 2) THEN civilgrp = 1;
ELSE IF civillib IN (3, 4, 5) THEN civilgrp = 2;
ELSE IF civillib IN (6, 7) THEN civilgrp = 3;
IF popurban < 45 THEN popgrp = 1;
ELSE IF popurban < 71 THEN popgrp = 2;
ELSE IF popurban >= 71 THEN popgrp = 3;
*DROP inf_mort civillib lev_tech;
RUN;
PROC PRINT DATA = MY_DATA.world;
RUN;
PROC FREQ DATA = MY_DATA.world;
  TABLES techgrp*infgrp;
RUN;
PROC FREQ DATA = MY_DATA.world;
  TABLES techgrp*infgrp / CHISQ CELLCHI2 EXPECTED NOCOL
NOROW NOPERCENT;
  TABLES techgrp*civilgrp / CHISQ CELLCHI2 EXPECTED NOCOL
NOROW NOPERCENT;
  TABLES popgrp*infgrp / CHISQ CELLCHI2 EXPECTED NOCOL
NOROW NOPERCENT MEASURES;
RUN;

```

频数 期望 单元格卡方		techgrp-infgrp表			
techgrp		infgrp			合计
		1	2	3	
1		2 13.75 10.041	13 8.8 2.0045	18 10.45 5.4548	33
2		23 11.25 12.272	3 7.2 2.45	1 8.55 6.667	27
合计		25	16	19	60

表“infgrp-techgrp”的统计量

统计量	自由度	值	概率
卡方	2	38.8894	<.0001
似然比卡方检验	2	45.3604	<.0001
Mantel-Haenszel 卡方	1	34.0997	<.0001
Phi 系数		0.8051	
列联系数		0.6271	
Cramer V		0.8051	

样本大小 = 60

由 χ^2 检验可知
 p 值很小，两者
 之间不独立

频数
期望
单元格卡方

techgrp-civilgrp表					
techgrp	civilgrp				
	1	2	3	合计	
1	2 12.1 8.4306	23 13.75 6.2227	8 7.15 0.101	33	
2	20 9.9 10.304	2 11.25 7.6056	5 5.85 0.1235	27	
合计	22	25	13	60	

表“civilgrp-techgrp”的统计量

由 χ^2 检验可知, p 值很小, 两者之间不独立

统计量	自由度	值	概率
卡方	2	32.7875	<.0001
似然比卡方检验	2	37.9110	<.0001
Mantel-Haenszel 卡方	1	14.1569	0.0002
Phi 系数		0.7392	
列联系数		0.5944	
Cramer V		0.7392	

样本大小 = 60

频数
期望
单元格卡方

popgrp-infgrp表					
popgrp	infgrp				
	1	2	3	合计	
1	1 7.9167 6.043	5 5.0667 0.0009	13 6.0167 8.1053	19	
2	8 8.75 0.0643	7 5.6 0.35	6 6.65 0.0635	21	
3	16 8.3333 7.0533	4 5.3333 0.3333	0 6.3333 6.3333	20	
合计	25	16	19	60	

表“infgrp-popgrp”的统计量

由 χ^2 检验可知, p 值很小, 两者之间不独立

统计量	自由度	值	概率
卡方	4	28.3470	<.0001
似然比卡方检验	4	34.7893	<.0001
Mantel-Haenszel 卡方	1	27.1451	<.0001
Phi 系数		0.6873	
列联系数		0.5664	
Cramer V		0.4860	

样本大小 = 60

统计量	值	ASE
Gamma	-0.8217	0.0657
Kendall's Tau-b	-0.6123	0.0681
Stuart's Tau-c	-0.6067	0.0672
Somers' D CIR	-0.6072	0.0662
Somers' D RIC	-0.6175	0.0707
Pearson 相关	-0.6783	0.0684
Spearman 相关	-0.6782	0.0690
Lambda 非对称 CIR	0.3429	0.0867
Lambda 非对称 RIC	0.3846	0.1319
Lambda 对称	0.3649	0.0995
不确定系数 CIR	0.2681	0.0600
不确定系数 RIC	0.2641	0.0575
不确定系数对称	0.2661	0.0587

样本大小 = 60

- (7) 写一个新的SAS程序利用(3)中的数据集world2，利用proc gchart来完成下列的散点图
- 计算根据变量infgrp作为y轴上分类，techgrp变量作为次要分类而分成组的每个样本的mean life expectanc的水平条形图
 - 选择合适的 midpoints 做关于 per capita GNP 的垂直条形图，用频数做为统计量并用变量 civilgrp 进行分组

```

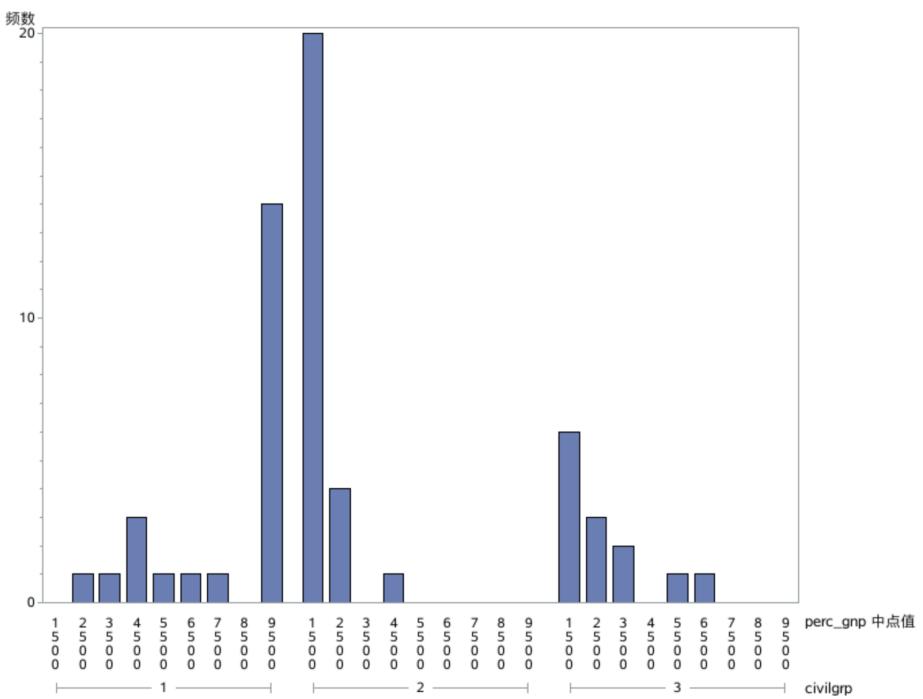
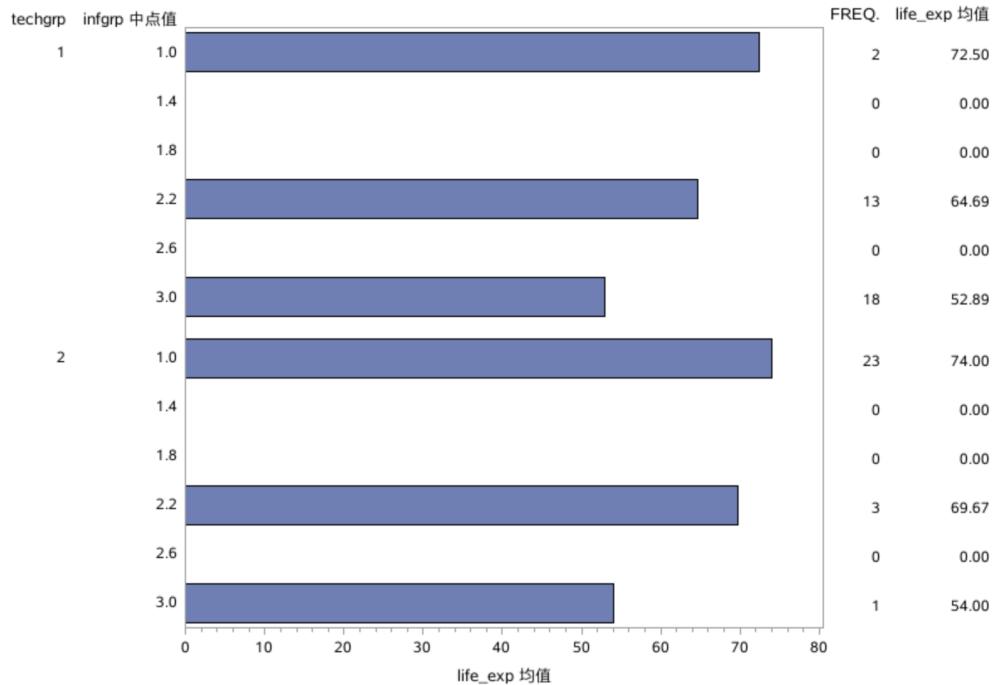
DATA MY_DATA.world;
  SET MY_DATA.raw_data;
  IF inf_mort < 24 THEN infgrp = 1;
  ELSE IF inf_mort < 73 THEN infgrp = 2;
  ELSE IF inf_mort >= 73 THEN infgrp = 3;
  IF lev_tech < 24 THEN techgrp = 1;
  ELSE IF lev_tech >= 24 THEN techgrp = 2;
  IF civillib IN (1, 2) THEN civilgrp = 1;
  ELSE IF civillib IN (3, 4, 5) THEN civilgrp = 2;
  ELSE IF civillib IN (6, 7) THEN civilgrp = 3;
  IF popurban < 45 THEN popgrp = 1;
  ELSE IF popurban < 71 THEN popgrp = 2;
  ELSE IF popurban >= 71 THEN popgrp = 3;
  *DROP inf_mort civillib lev_tech;
RUN;
PROC PRINT DATA=MY_DATA.world;
RUN;
PROC GCHART DATA = MY_DATA.world;
  HBAR infgrp /
    GROUP = techgrp
    TYPE = MEAN

```

```

SUMVAR = life_exp;
RUN;
PROC GCHART DATA = MY_DATA.world;
VBAR perc_gnp /
GROUP = civilgrp
TYPE = FREQ
MIDPOINTS = 1500 to 10000 by 1000;
RUN;

```



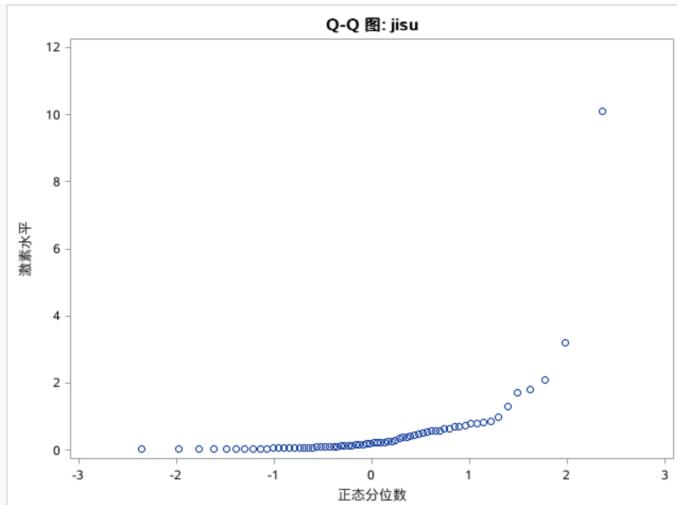
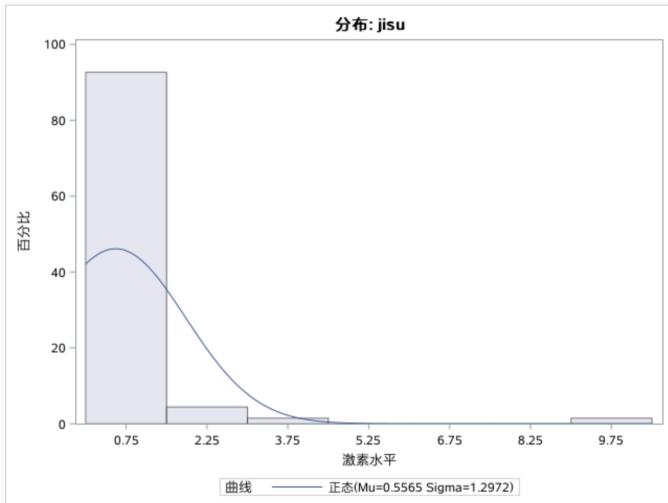
第二题

(1) 创建你的逻辑库，将数据集 `jisu` 放入你的库中。

```
LIBNAME hw "/home/u61827597/Sufe_SAS/Assignments/hw3/";
```

(2) 使用 univariate 过程步对激素水平进行分析，画出该指标的直方图和正态的 QQ 图，并作正态性检验，试问该指标是否服从正态分布。

```
proc univariate data=hw.jisu;
  var jisu;
  histogram jisu/normal;
  qqplot jisu/normal;
run;
```



UNIVARIATE 过程 拟合“正态”分布 - jisu (激素水平)				
“正态”分布的参数				
参数	符号	估计		
均值	Mu	0.556471		
标准差	Sigma	1.297226		

“正态”分布的拟合优度检验				
检验	统计量		p 值	
Kolmogorov-Smirnov	D	0.3481109	Pr > D	<0.010
Cramer-von Mises	W-Sq	2.4777333	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	12.7079193	Pr > A-Sq	<0.005

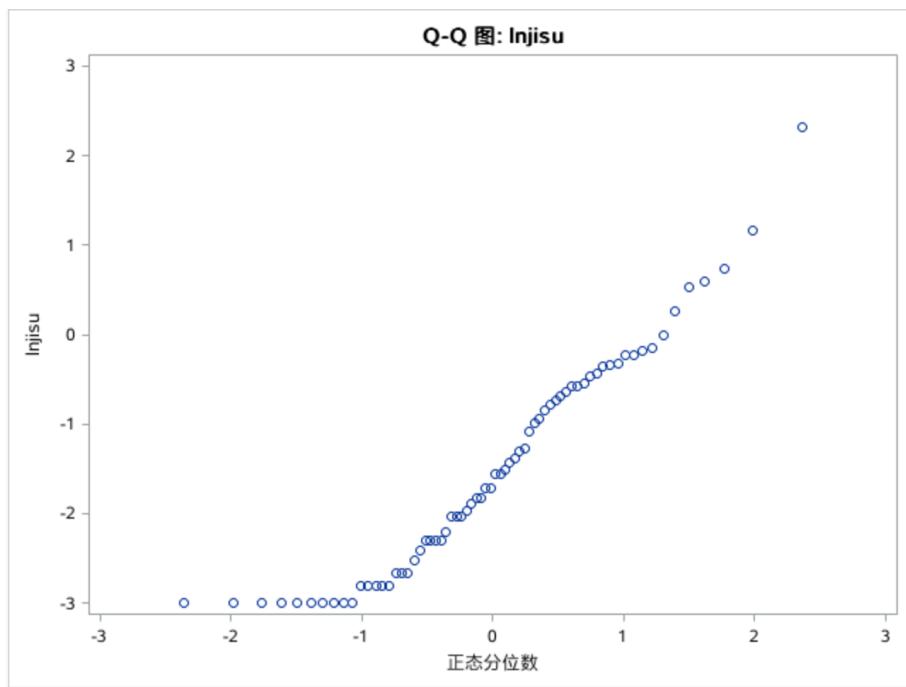
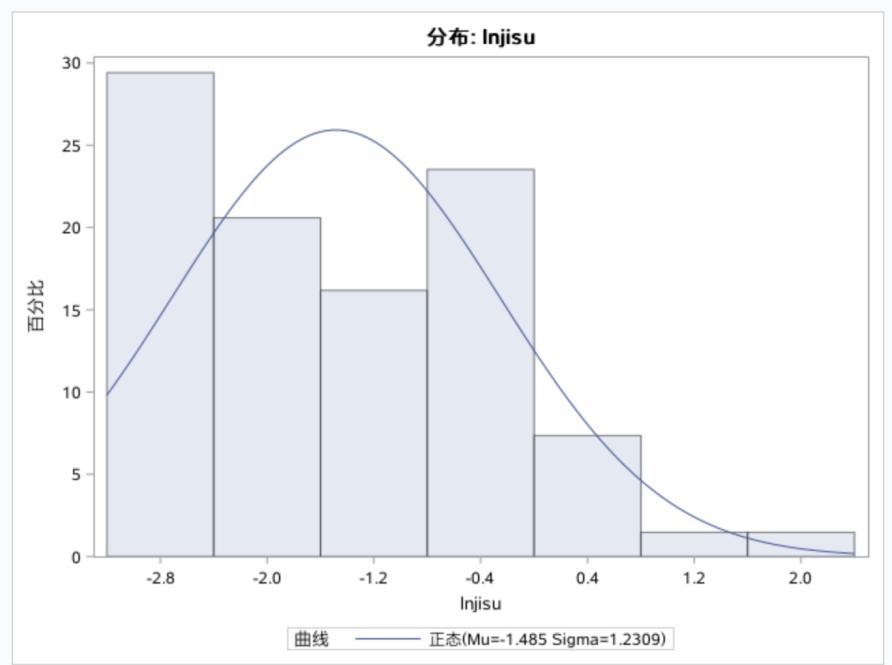
“正态”分布的分位数				
百分比	分位数			
	观测	估计		
1.0	0.05000	-2.46133		
5.0	0.05000	-1.57728		
10.0	0.05000	-1.10599		
25.0	0.07000	-0.31850		
50.0	0.19500	0.55647		
75.0	0.57000	1.43144		
90.0	1.00000	2.21893		
95.0	1.80000	2.69022		
99.0	10.10000	3.57427		

正态分布检验不通过，拒绝认为其服从正态分布

(3) 对激素水平进行对数化保存到新指标 `lnjis` 中，将生成的新数据集保存到 `jisu1` 中，验证 `lnjis` 是否服从正态分布。数据对数变换是 Box-Cox 变换的一种特殊情况，查询资料回答什么是 Box-Cox 变换？数据的对数化处理是否存在缺点？

```
data hw.jisu1;
  set hw.jisu;
  lnjis=log(jisu);
  drop jisu;
run;

proc univariate data=hw.jisu1;
  var lnjis;
  histogram lnjis/normal;
  qqplot lnjis/normal;
run;
```



UNIVARIATE 过程 “Injisu”的拟合正态分布				
“正态”分布的参数				
参数	符号	估计		
均值	Mu	-1.48468		
标准差	Sigma	1.230858		
“正态”分布的拟合优度检验				
检验	统计量		p 值	
Kolmogorov-Smirnov	D	0.10979095	Pr > D	0.041
Cramer-von Mises	W-Sq	0.16207107	Pr > W-Sq	0.017
Anderson-Darling	A-Sq	1.16052575	Pr > A-Sq	<0.005
“正态”分布的分位数				
百分比	分位数			
	观测	估计		
1.0	-2.99573	-4.34808		
5.0	-2.99573	-3.50926		
10.0	-2.99573	-3.06209		
25.0	-2.65926	-2.31488		
50.0	-1.63772	-1.48468		
75.0	-0.56227	-0.65448		
90.0	0.00000	0.09273		
95.0	0.58779	0.53990		
99.0	2.31254	1.37872		

正态分布检验不通过，拒绝认为其服从正态分布

Box-cox:

对观测得到的试验数据集 $(x'_j, y_j), i = 1, \dots, n$, 若经过回归诊断后得知, 它们不满足 Gauss-Markov 条件, 我们就要对数据采取“治疗”措施, 实践证明, 数据变换是处理有问题数据的一种好方法。Box-Cox 变换的主要特点是引入一个参数, 通过数据本身估计该参数, 从而确定应采取的数据变换形式, 实践证明, Box-Cox 变换对许多实际数据都是行之有效的。

Box-Cox 变换是对回归因变量的如下变换:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}$$

这里 λ 是一个待定变换参数。Box-Cox 变换时一族变换, 它包括了许多常见的变换, 诸如对数变换 ($\lambda = 0$), 倒数变换 ($\lambda = -1$) 和平方根变换 ($\lambda = 1/2$) 等等。

对因变量的 n 个观测值 y_1, \dots, y_n , 应用上述变换, 得到变换后的向量

$$y^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})'$$

$$y^{(\lambda)} = X\beta + e, e \sim N(0, \sigma^2 I)$$

即通过因变量的变换, 使得变换过的向量 $y^{(\lambda)}$ 与回归自变量之间具有线性

相依关系，误差也服从正态分布，误差各分量等方差且相互独立。因此，Box-Cox 变换时通过参数 λ 的选择，达到对原来数据的“综合治理”，使其满足一个正态线性回归模型的所有假设条件。

现在我们把 Box-Cox 变换的具体步骤归纳如下：

- (1) 对给定的 λ 值，计算出 $z_i^{(\lambda)}$ ；
- (2) 计算残差平方和 $SS_e(\lambda, z^{(\lambda)})$ ；
- (3) 对一系列的 λ 值，重复上述步骤，得到相应的残差平方和 $SS_e(\lambda, z^{(\lambda)})$ 的一串值，以 λ 为横轴，做出相应的曲线。用直观的方法，找出使 $SS_e(\lambda, z^{(\lambda)})$ 达到最小值点的 $\hat{\lambda}$
- (4) 求出 $\hat{\beta}(\hat{\lambda})$

对数变换有时会增大数据分布的偏度（左偏数据的偏度更大），或使得线性性更差；有时也会降低模型的可解释性。

(4) 求出 lnjisu 的 33.33% 和 66.67% 分位数，并以求出的两个分位数作为分界，将 lnjisu 分成三类保存到分类变量 jisugrp 中，生成的新数据集保存到 jisu2 中。对 jisugrp 的三类分别用高、中和低标签的值进行格式化。

```
proc univariate data=hw.jisu1;
  var lnjisu;
  histogram lnjisu/normal;
  qqplot lnjisu/normal;
run;

proc univariate data=hw.jisu1;
  var lnjisu;
  output out=hw.jsqtl pctlpts=33.3 66.7 pctlpre=lnjisu;
run;

proc print data=hw.jsqtl;
run;

data hw.jisu2;
  set hw.jisu1;
  format jisugrp $ 6.;
  if lnjisu<-2.30259 then
    jisugrp="low";
  else if lnjisu>=-2.3059 and lnjisu < -0.77653 then
    jisugrp="medium";
  else if lnjisu>=-0.77653 then
```

```

jisugrp="high";
run;

proc print data=hw.jisu2;
run;

```

观测	id	group	gender	age	weisuo	changhua	age_cls	Injisu	jisugrp
1	001	1	1	42	2	1	1	-1.71480	medium
2	002	1	1	38	1	2	1	-2.52573	low
3	003	1	1	45	2	3	1	-2.99573	low
4	004	1	1	45	2	2	1	-0.84397	medium
5	005	1	1	43	2	3	1	-1.83258	medium
6	006	1	2	44	3	2	1	-2.99573	low
7	007	1	1	45	1	1	1	-2.65926	low
8	008	1	1	45	3	3	1	-2.81341	low
9	009	1	1	27	3	1	1	-2.99573	low
10	010	1	1	36	1	4	1	-2.81341	low
11	011	1	2	39	1	2	1	-0.46204	high
12	012	1	2	45	1	2	1	-0.63488	high
13	013	1	2	44	2	3	1	-0.15082	high
14	014	1	2	42	2	2	1	-0.54473	high
15	015	1	1	42	2	2	1	-2.30259	medium
16	026	1	2	53	2	1	2	0.74194	high
17	027	1	1	50	2	1	2	-2.99573	low
18	028	1	1	57	3	3	2	-2.20727	medium
19	029	1	2	54	1	1	2	-2.30259	medium
20	030	1	1	47	1	3	2	-2.81341	low
21	031	1	2	53	3	3	2	-2.81341	low
22	032	1	2	48	1	3	2	-1.07881	medium
23	033	1	1	47	1	3	2	-2.99573	low
24	034	1	1	53	2	2	2	-1.96611	medium
25	035	1	2	50	3	3	2	-2.99573	low
26	036	1	1	46	2	3	2	-2.65926	low
27	037	1	1	51	2	2	2	0.00000	high
28	049	1	2	69	1	1	3	0.26236	high
29	050	1	1	69	2	3	3	-1.38629	medium
30	051	1	1	60	2	2	3	-2.99573	low
31	052	1	1	70	1	1	3	-0.73397	high
32	053	1	1	62	1	1	3	-2.99573	low
33	054	1	1	63	1	2	3	-2.04022	medium
34	055	1	2	61	2	2	3	-2.81341	low
35	056	1	1	60	3	4	3	-2.99573	low

(5) 分别做年龄分组、肠化程度、萎缩程度与激素水平分组的列联表分析，检验他们之间的关联性，给出你的结论。

```

proc freq data=hw.jisu2;
tables age*jisugrp / exact;
tables changhua*jisugrp / exact;
tables weisuo*jisugrp / exact;
*exact pchi;
run;

```

频数
百分比
行百分比
列百分比

age(年龄)	age-jisugrp表				
	high	low	medium	合计	
17	1 1.47 100.00 4.35	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 1.47
27	0 0.00 0.00 0.00	1 1.47 100.00 5.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 1.47
28	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 1.47 100.00 4.00	1 1.47 4.00	1 1.47
31	1 1.47 33.33 4.35	0 0.00 0.00 0.00	2 2.94 66.67 8.00	3 4.41	
32	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 1.47 100.00 4.00	1 1.47	
33	1 1.47 50.00 4.35	0 0.00 0.00 0.00	1 1.47 50.00 4.00	2 2.94	
36	0 0.00 0.00 0.00	1 1.47 100.00 5.00	0 0.00 0.00 0.00	1 1.47	
38	0 0.00 0.00 0.00	1 1.47 100.00 5.00	0 0.00 0.00 0.00	1 1.47	
39	1 1.47 50.00 4.35	0 0.00 0.00 0.00	1 1.47 50.00 4.00	2 2.94	
40	1 1.47 100.00 4.35	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 1.47	

表“jisugrp-age”的统计量

统计量	自由度	值	概率
卡方	66	62.3693	0.6040
似然比卡方检验	66	74.7027	0.2165
Mantel-Haenszel 卡方	1	0.8397	0.3595
Phi 系数		0.9577	
列联系数		0.6917	
Cramer V		0.6772	

WARNING: 100% 的单元格包含小于 5 的期望计数。卡方不是有效的检验。

Fisher 精确检验	
表概率 (P)	<.0001
Pr <= P	0.8063

样本大小 = 68

频数
百分比
行百分比
列百分比

		changhua-jisugrp表			
		jisugrp			
changhua(肠化)		high	low	medium	合计
1	3	4	3	10	
	8.11	10.81	8.11		27.03
	30.00	40.00	30.00		
	37.50	22.22	27.27		
2	4	4	4	12	
	10.81	10.81	10.81		32.43
	33.33	33.33	33.33		
3	1	8	4	13	
	2.70	21.62	10.81		35.14
	7.69	61.54	30.77		
	12.50	44.44	36.36		
4	0	2	0	2	
	0.00	5.41	0.00		5.41
	0.00	100.00	0.00		
	0.00	11.11	0.00		
合计		8	18	11	37
		21.62	48.65	29.73	100.00
频数缺失 = 31					

表“jisugrp-changhua”的统计量

统计量	自由度	值	概率
卡方	6	5.5973	0.4698
似然比卡方检验	6	6.6578	0.3537
Mantel-Haenszel 卡方	1	0.3827	0.5362
Phi 系数		0.3889	
列联系数		0.3625	
Cramer V		0.2750	
WARNING: 83% 的单元格包含小于 5 的期望计数。卡方不是有效的检验。			

Fisher 精确检验

表概率 (P)	0.0007
Pr <= P	0.5651

样本大小 = 37

频数缺失 = 31

频数
百分比
行百分比
列百分比

		weisuo-jisugrp表			
		jisugrp			
weisuo(萎缩)		high	low	medium	合计
1	4	6	4	14	
	10.81	16.22	10.81		37.84
	28.57	42.86	28.57		
	50.00	33.33	36.36		
2	4	6	6	16	
	10.81	16.22	16.22		43.24
	25.00	37.50	37.50		
3	0	6	1	7	
	0.00	16.22	2.70		18.92
	0.00	85.71	14.29		
	0.00	33.33	9.09		
合计		8	18	11	37
		21.62	48.65	29.73	100.00
频数缺失 = 31					

表“jisugrp-weisuo”的统计量

统计量	自由度	值	概率
卡方	4	5.2857	0.2592
似然比卡方检验	4	6.5461	0.1619
Mantel-Haenszel 卡方	1	0.2397	0.6244
Phi 系数		0.3780	
列联系数		0.3536	
Cramer V		0.2673	
WARNING: 78% 的单元格包含小于 5 的期望计数。卡方不是有效的检验。			

Fisher 精确检验

表概率 (P)	0.0019
Pr <= P	0.3266

样本大小 = 37

频数缺失 = 31

WARNING: 46% 数据缺失。

Fisher 检验结果可知其两两独立

第三题

```
DATA data;
  input income :$7. attitude:$7. numcell;
  cards;
  low against 9
  low support 92
  medium against 31
  medium support 52
  high against 40
  high support 25
  ;
run;

proc freq data=data;
  tables income * attitude / chisq expected;
  weight numcell;
run;
```

FREQ 过程				
		income-attitude表		
		attitude		
income		against	support	合计
high	40	25	65	
	20.884	44.116		
	16.06	10.04		26.10
	61.54	38.46		
	50.00	14.79		
low	9	92	101	
	32.45	68.55		
	3.61	36.95		40.56
	8.91	91.09		
	11.25	54.44		
medium	31	52	83	
	26.667	56.333		
	12.45	20.88		33.33
	37.35	62.65		
	38.75	30.77		
合计	80	169	249	
	32.13	67.87	100.00	

表"attitude-income"的统计量

统计量	自由度	值	概率
卡方	2	51.7877	<0.0001
似然比卡方检验	2	55.6591	<0.0001
Mantel-Haenszel 卡方	1	6.8043	0.0091
Phi 系数		0.4561	
列联系数		0.4149	
Cramer V		0.4561	

样本大小 = 249

由检验结果可知 p 值较小，拒绝原假设，认为其具有相关性影响

第四题

(1)选用合适的工具和变量进行描述统计分析，了解该数据集中月度数据的缺失情况。

```
libname hw_data "/home/u61827597/Sufe_SAS/Assignments/hw3";
```

```

proc print data=hw_data.sale;
run;

data sale_data;
  set hw_data.sale;
  year=year(DATE);
  month=month(DATE);
run;

proc means data=sale_data;
  class year;
run;

```

MEANS PROCEDURE								
year	观测数	变量	标签	数目	均值	标准差	最小值	最大值
1949	10	DATE sale month	international airline travel (thousands)	10	-3835.10	113.6470658	-4017.00	-3683.00
				10	126.7000000	14.9447129	104.0000000	148.0000000
				10	7.0000000	3.7416574	1.0000000	12.0000000
1950	10	DATE sale month	international airline travel (thousands)	10	-3476.30	118.0358232	-3652.00	-3318.00
				10	139.2000000	20.7888859	114.0000000	170.0000000
				10	6.8000000	3.8815804	1.0000000	12.0000000
1951	10	DATE sale month	international airline travel (thousands)	10	-3120.50	115.7710864	-3287.00	-2953.00
				10	169.5000000	19.6991258	145.0000000	199.0000000
				10	6.5000000	3.8078866	1.0000000	12.0000000
1952	11	DATE sale month	international airline travel (thousands)	11	-2769.91	101.1280916	-2922.00	-2617.00
				11	197.2727273	24.0669521	171.0000000	242.0000000
				11	6.0000000	3.3166248	1.0000000	11.0000000
1953	10	DATE sale month	international airline travel (thousands)	10	-2392.50	106.3424761	-2556.00	-2222.00
				10	228.4000000	27.1874154	196.0000000	272.0000000
				10	6.4000000	3.5023801	1.0000000	12.0000000
1954	11	DATE sale month	international airline travel (thousands)	11	-2014.73	109.4185459	-2191.00	-1857.00
				11	239.2727273	36.6062588	188.0000000	302.0000000
				11	6.8181818	3.6005050	1.0000000	12.0000000
1955	10	DATE sale month	international airline travel (thousands)	10	-1644.40	111.2806262	-1826.00	-1492.00
				10	286.0000000	42.0740089	237.0000000	364.0000000
				10	7.0000000	3.6514837	1.0000000	12.0000000
1956	11	DATE sale month	international airline travel (thousands)	11	-1300.64	112.4831300	-1461.00	-1126.00
				11	325.8181818	49.4142048	271.0000000	413.0000000
				11	6.2727273	3.6902821	1.0000000	12.0000000
1957	9	DATE sale month	international airline travel (thousands)	9	-926.8888889	97.6530138	-1064.00	-791.0000000
				9	372.0000000	61.2107017	301.0000000	467.0000000
				9	6.5555556	3.2058973	2.0000000	11.0000000
1958	11	DATE sale month	international airline travel (thousands)	11	-578.7272727	100.8584066	-730.0000000	-426.0000000
				11	385.0000000	66.1014372	310.0000000	505.0000000
				11	6.0000000	3.3166248	1.0000000	11.0000000
1959	11	DATE sale month	international airline travel (thousands)	11	-188.7272727	109.4185459	-365.0000000	-31.0000000
				11	430.3636364	72.8660040	342.0000000	559.0000000
				11	6.8181818	3.6005050	1.0000000	12.0000000
1960	10	DATE sale month	international airline travel (thousands)	10	155.1000000	111.6805464	0	335.0000000
				10	478.9000000	78.5372523	391.0000000	622.0000000
				10	6.1000000	3.6851512	1.0000000	12.0000000

(2) 试用数据步将月度客运量数据转换为年度数据;

```

data annual;
  set sale_data;
  drop DATE;
  by year;

  if first.year=1 then
    sum=0;
  sum+sale;
  annual_sale=(last.year)*sum;

```

```

if annual_sale=0 then
do;
  delete;
end;
keep year annual_sale;
run;
proc print data=annual;
run;

```

观测	year	annual_sale
1	1949	1267
2	1950	1392
3	1951	1695
4	1952	2170
5	1953	2284
6	1954	2632
7	1955	2860
8	1956	3584
9	1957	3348
10	1958	4235
11	1959	4734
12	1960	4789

(3) 利用Freq和means过程步将月度数据转化为年度数据;

```

proc means sum data=sale_data;
var sale;
class year;
output out=annual_data sum (sale)=sum_sale;
run;

```

MEANS PROCEDURE		
分析变量: sale international airline travel (thousands)		
year	观测数	总和
1949	10	1267.00
1950	10	1392.00
1951	10	1695.00
1952	11	2170.00
1953	10	2284.00
1954	11	2632.00
1955	10	2860.00
1956	11	3584.00
1957	9	3348.00
1958	11	4235.00
1959	11	4734.00
1960	10	4789.00

(4) 用gplot绘制年度客运量数据的折线图, 用gchart绘制数据集中每年包含月份数的条形图, 用sgplot将这两种图形叠加在一张图上。

```
proc gplot data=annual;
  plot annual_sale*year;
  symbol h=2 v=circle i=join;
run;

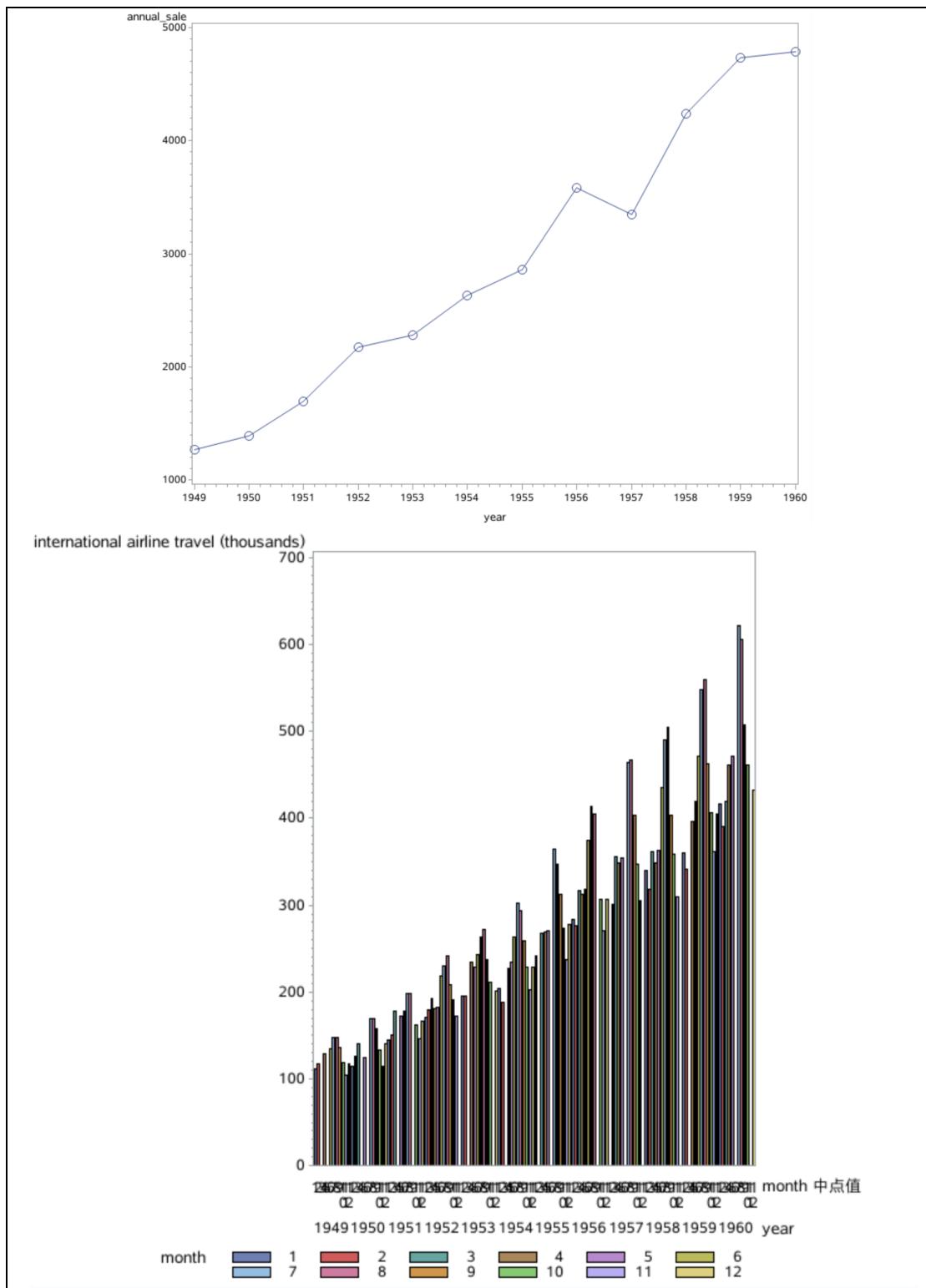
proc gchart data=sale_data;
  vbar month / type=sum sumvar=sale midpoints=1 to 12 by 1
group=year
  subgroup=month;
run;

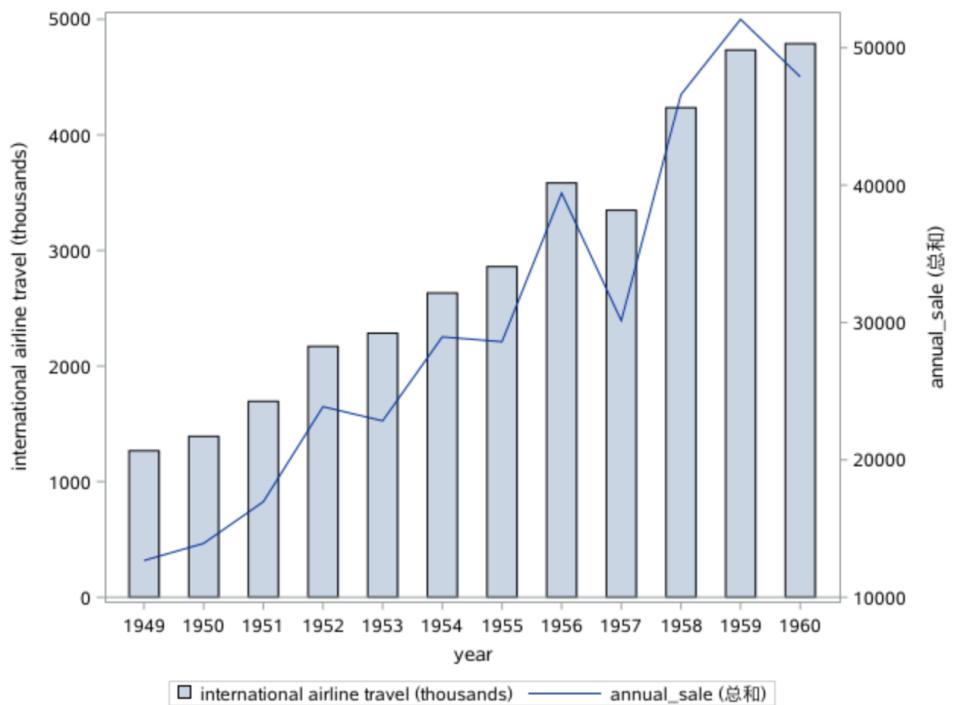
proc gchart data=sale_data;
  vbar month / type=sum sumvar=sale midpoints=1 to 12 by 1
group=year
  subgroup=month;
run;

data all;
  merge annual sale_data;
  by year;
run;

proc print data=all;
run;

proc sgplot data=all;
  vbar year / stat=sum response=sale
  groupdisplay=stack barwidth=0.5;
  vline year / response=annual_sale y2axis;
run;
```





(5) 如果由你负责数据搜集和整理工作，请用sas编程把数据缺失的月份找出来，列出一份清单，以便向企业索要数据。

```

data temp_report;
  set sort_data;
  format report missing;
  by year month;
  first=first.year;

  if first.year=1 then
    t_month=0;
  report=month-t_month-1;

  if (report >0) then
    do;
      missing=month-1;
      t_month+1;
    end;

  if last.year*month ne 12 and last.year ne 0 then
    missing=12;
  t_month+1;
run;

proc print data=temp_report;
run;

```

```

data report;
  set temp_report;

  if missing=. then
    delete;
  keep year missing;

proc print data=report;
run;

```

观测	DATE	sale	year	month	report	missing	first	t_month
1	JAN49	112	1949	1	0	.	1	1
2	FEB49	118	1949	2	0	.	0	2
3	APR49	129	1949	4	1	3	0	4
4	JUN49	135	1949	6	1	5	0	6
5	JUL49	148	1949	7	0	.	0	7
6	AUG49	148	1949	8	0	.	0	8
7	SEP49	136	1949	9	0	.	0	9
8	OCT49	119	1949	10	0	.	0	10
9	NOV49	104	1949	11	0	.	0	11
10	DEC49	118	1949	12	0	.	0	12
11	JAN50	115	1950	1	0	.	1	1
12	FEB50	126	1950	2	0	.	0	2
13	MAR50	141	1950	3	0	.	0	3
14	MAY50	125	1950	5	1	4	0	5
15	JUL50	170	1950	7	1	6	0	7
16	AUG50	170	1950	8	0	.	0	8
17	SEP50	158	1950	9	0	.	0	9
18	OCT50	133	1950	10	0	.	0	10
19	NOV50	114	1950	11	0	.	0	11
20	DEC50	140	1950	12	0	.	0	12
21	JAN51	145	1951	1	0	.	1	1
22	FEB51	150	1951	2	0	.	0	2
23	MAR51	178	1951	3	0	.	0	3
24	MAY51	172	1951	5	1	4	0	5
25	JUN51	178	1951	6	0	.	0	6
26	JUL51	199	1951	7	0	.	0	7
27	AUG51	199	1951	8	0	.	0	8
28	OCT51	162	1951	10	1	9	0	10
29	NOV51	146	1951	11	0	.	0	11
30	DEC51	166	1951	12	0	.	0	12

观测	year	missing
1	1949	3
2	1949	5
3	1950	4
4	1950	6
5	1951	4
6	1951	9
7	1952	12
8	1953	3
9	1953	11
10	1954	3
11	1955	2
12	1955	6
13	1956	9
14	1957	1
15	1957	6
16	1957	12
17	1958	12
18	1959	3
19	1960	6
20	1960	11

【小结】

本次的实验主要就针对一些统计问题进行了操作练习，包括 t 检验、卡方检验及其衍生检验、统计绘图，最后还进行了 SAS 编程内容练习。在本次操作中对于常规的 proc 步与 data 步也进行了进一步的熟悉。希望能够对编程的语句更加熟悉，提高编程效率。

指导教师评语及成绩：

成绩：

指导教师签名：

批阅日期：

附件：

实验报告说明

1. **实验项目名称：**要用最简练的语言反映实验的内容。
2. **实验类型：**一般需说明是验证型实验还是设计型实验、综合型实验或其他实验。
3. **实验目的与要求：**目的要明确，要抓住重点。
4. **实验原理：**简要说明本实验项目所涉及的理论知识。
5. **实验环境：**实验用的软硬件环境（配置）。
6. **实验方案设计（思路、步骤和方法等）：**这是实验报告极其重要的内容，概括整个实验过程。

对于验证型实验，要写明依据何种原理、何仲操作方法进行实验，并写明需要经过哪几个步骤。

对于设计型和综合型实验，在上述内容基础上还应该画出流程图、设计思路和设计方法，再配以相应的文字说明。

7. **实验过程（实验中涉及的记录、数据、分析）：**写明具体上述实验方案的具体实施，包括实验过程中的记录、数据和相应的分析。
8. **结论（结果）：**即根据实验过程中所见到的现象和测得的数据，做出结论。
9. **小结：**对本次实验的心得体会、思考和建议。
10. **指导教师评语及成绩：**指导教师依据学生的实际报告内容，用简练语言给出本次实验报告的评价和价值。