108-2 流行病學 實習課 作業三 更新

梁嫚芳
B07801003
公衛二

繳交日期: 2020.06.05 提醒：

- 第一頁上方註明姓名、學號、系級
- 請以 PDF 檔案繳交
- 除了回答問題，結果還需包含文字敘述和 SAS 程式碼

題組：第一題至第三題

作業檔案 data1 為一個 Cohort study 研究資料，此檔案包含受試者的性別(sex)、年齡(age)、收縮壓(sbp)、舒張壓(dbp)、BMI(bmi)、中風(stroke)、追蹤時間(followup) 等資料。除了追蹤時間與中風情形是追蹤研究結束後得知，其餘資料都是在受試者剛開始參與追蹤研究時由測量或填寫問卷得知。請利用此檔案進行第一題分析。

請利用此檔案進行以下分析：

1. 請問在此 Cohort study 資料中有多少高血壓的盛行個案？(高血壓標準：收縮壓≧140mmHg 或舒張壓≧90mmHg)（13%）

FREQ 程序

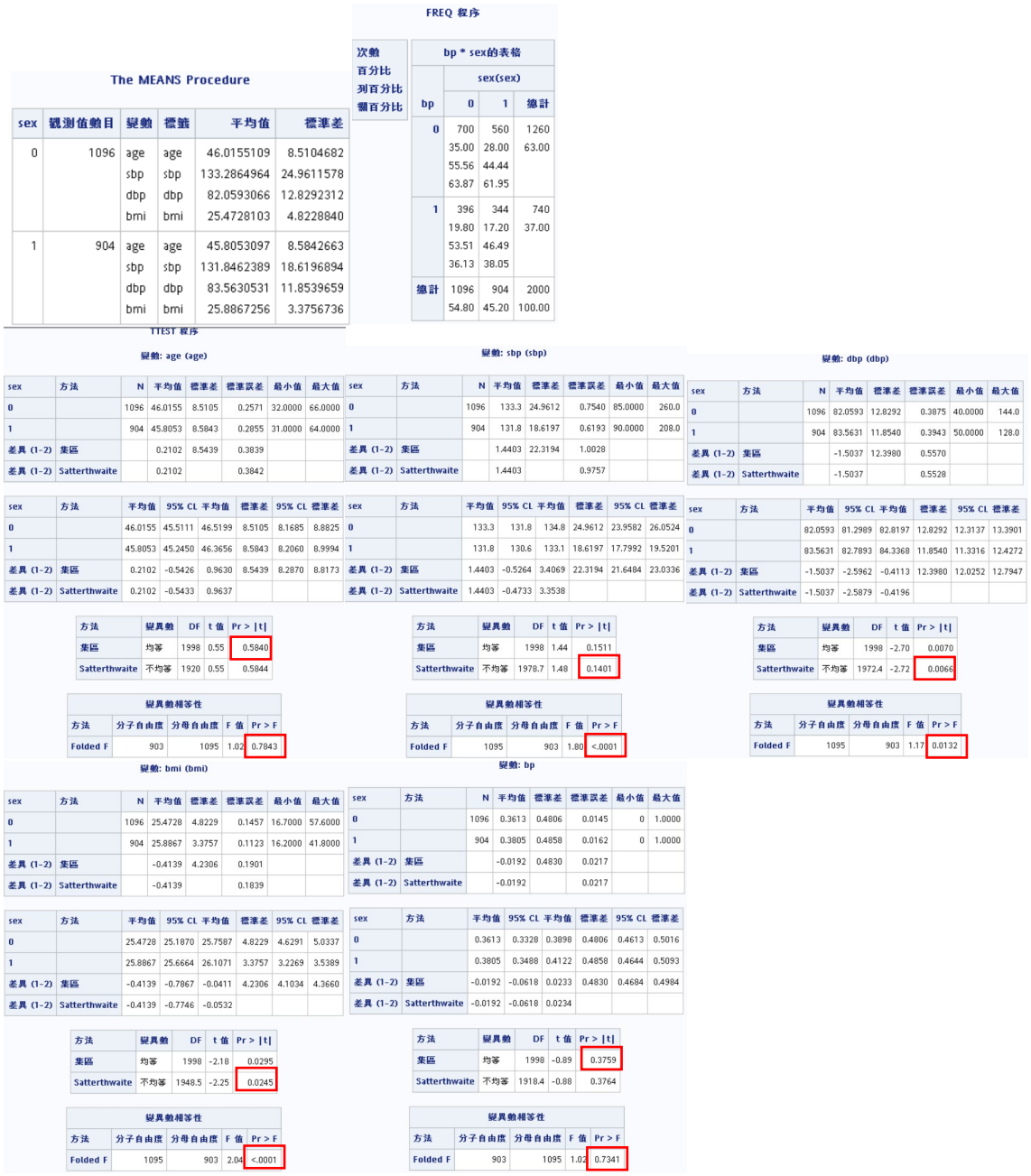
bp	次數	百分比	累積次數	累積百分比
0	1260	63.00	1260	63.00
1	740	37.00	2000	100.00

高血壓的盛行個案有 740 個，盛行率為 37%

```
/*1*/
/*將血壓分組*/
data data1_1 ;
set data1 ;
        if sbp >= 140 or dbp >= 90 then bp  = 1 ;
        else bp = 0 ;
run ;
/*呈現表格*/
PROC FREQ DATA=data1_1;
        TABLE bp ;
RUN;
```

2. 下表欲呈現該 Cohort study 的受試者男性與女性基本人口學資料，請協助完成。請依資料型態分別用平均值、標準差、樣本數及百分比呈現，並檢定不同性別該變項是否有顯著差異。(分析變項：年齡、收縮壓、舒張壓、BMI、高血壓)（27%）

	男性 (n=904, 45.2%)	女性 (n= 1096, 54.8%)	p-value
年齡 (mean, s.d)	45.80, 8.58	46.02, 8.51	0.5840
收縮壓 (mean, s.d)	131.85, 18.62	133.29, 24.96	0.1401
舒張壓 (mean, s.d)	83.56, 11.85	82.06, 12.83	0.0066
BMI (mean, s.d)	25.89, 3.38	25.47, 4.82	0.0245
罹患高血壓 (n, %)	344, 46.49	396, 53.51	0.3759



The MEANS Procedure

sex	觀測值數目	變數	標籤	平均值	標準差
0	1096	age	age	46.0155109	8.5104682
		sbp	sbp	133.2864964	24.9611578
		dbp	dbp	82.0593066	12.8292312
		bmi	bmi	25.4728103	4.8228840
1	904	age	age	45.8053097	8.5842663
		sbp	sbp	131.8462389	18.6196894
		dbp	dbp	83.5630531	11.8539659
		bmi	bmi	25.8867256	3.3756736

FREQ 程序

次數 百分比 列百分比 欄百分比	bp * sex的表格		
	sex(sex)		
bp	0	1	總計
0	700 35.00 55.56 63.87	560 28.00 44.44 61.95	1260 63.00
1	396 19.80 53.51 36.13	344 17.20 46.49 38.05	740 37.00
總計	1096 54.80	904 45.20	2000 100.00

TTEST 程序

變數: age (age)

sex	方法	N	平均值	標準差	標準誤差	最小值	最大值
0		1096	46.0155	8.5105	0.2571	32.0000	66.0000
1		904	45.8053	8.5843	0.2855	31.0000	64.0000
差異 (1-2)	集區		0.2102	8.5439	0.3839		
差異 (1-2)	Satterthwaite		0.2102		0.3842		

sex	方法	平均值	95% CL	平均值	標準差	95% CL	標準差
0		46.0155	45.5111	46.5199	8.5105	8.1685	8.8825
1		45.8053	45.2450	46.3656	8.5843	8.2060	8.9994
差異 (1-2)	集區	0.2102	-0.5426	0.9630	8.5439	8.2870	8.8173
差異 (1-2)	Satterthwaite	0.2102	-0.5433	0.9637			

方法	變異數	DF	t 值	Pr > |t|
集區	均等	1998	0.55	0.5840
Satterthwaite	不均等	1920	0.55	0.5844

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	903	1095	1.02	0.7843

變數: sbp (sbp)

sex	方法	N	平均值	標準差	標準誤差	最小值	最大值
0		1096	133.3	24.9612	0.7540	85.0000	260.0
1		904	131.8	18.6197	0.6193	90.0000	208.0
差異 (1-2)	集區		1.4403	22.3194	1.0028		
差異 (1-2)	Satterthwaite		1.4403		0.9757		

sex	方法	平均值	95% CL	平均值	標準差	95% CL	標準差
0		133.3	131.8	134.8	24.9612	23.9582	26.0524
1		131.8	130.6	133.1	18.6197	17.7992	19.5201
差異 (1-2)	集區	1.4403	-0.5264	3.4069	22.3194	21.6484	23.0336
差異 (1-2)	Satterthwaite	1.4403	-0.4733	3.3538			

方法	變異數	DF	t 值	Pr > |t|
集區	均等	1998	1.44	0.1511
Satterthwaite	不均等	1978.7	1.48	0.1401

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	1095	903	1.80	<.0001

變數: dbp (dbp)

sex	方法	N	平均值	標準差	標準誤差	最小值	最大值
0		1096	82.0593	12.8292	0.3875	40.0000	144.0
1		904	83.5631	11.8540	0.3943	50.0000	128.0
差異 (1-2)	集區		-1.5037	12.3980	0.5570		
差異 (1-2)	Satterthwaite		-1.5037		0.5528		

sex	方法	平均值	95% CL	平均值	標準差	95% CL	標準差
0		82.0593	81.2989	82.8197	12.8292	12.3137	13.3901
1		83.5631	82.7893	84.3368	11.8540	11.3316	12.4272
差異 (1-2)	集區	-1.5037	-2.5962	-0.4113	12.3980	12.0252	12.7947
差異 (1-2)	Satterthwaite	-1.5037	-2.5879	-0.4196			

方法	變異數	DF	t 值	Pr > |t|
集區	均等	1998	-2.70	0.0070
Satterthwaite	不均等	1972.4	-2.72	0.0066

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	1095	903	1.17	0.0132

變數: bmi (bmi)

sex	方法	N	平均值	標準差	標準誤差	最小值	最大值
0		1096	25.4728	4.8229	0.1457	16.7000	57.6000
1		904	25.8867	3.3757	0.1123	16.2000	41.8000
差異 (1-2)	集區		-0.4139	4.2306	0.1901		
差異 (1-2)	Satterthwaite		-0.4139		0.1839		

sex	方法	平均值	95% CL	平均值	標準差	95% CL	標準差
0		25.4728	25.1870	25.7587	4.8229	4.6291	5.0337
1		25.8867	25.6664	26.1071	3.3757	3.2269	3.5389
差異 (1-2)	集區	-0.4139	-0.7867	-0.0411	4.2306	4.1034	4.3660
差異 (1-2)	Satterthwaite	-0.4139	-0.7746	-0.0532			

方法	變異數	DF	t 值	Pr > |t|
集區	均等	1998	-2.18	0.0295
Satterthwaite	不均等	1948.5	-2.25	0.0245

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	1095	903	2.04	<.0001

變數: bp

sex	方法	N	平均值	標準差	標準誤差	最小值	最大值
0		1096	0.3613	0.4806	0.0145	0	1.0000
1		904	0.3805	0.4858	0.0162	0	1.0000
差異 (1-2)	集區		-0.0192	0.4830	0.0217		
差異 (1-2)	Satterthwaite		-0.0192		0.0217		

sex	方法	平均值	95% CL	平均值	標準差	95% CL	標準差
0		0.3613	0.3328	0.3898	0.4806	0.4613	0.5016
1		0.3805	0.3488	0.4122	0.4858	0.4644	0.5093
差異 (1-2)	集區	-0.0192	-0.0618	0.0233	0.4830	0.4684	0.4984
差異 (1-2)	Satterthwaite	-0.0192	-0.0618	0.0234			

方法	變異數	DF	t 值	Pr > |t|
集區	均等	1998	-0.89	0.3759
Satterthwaite	不均等	1918.4	-0.88	0.3764

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	903	1095	1.02	0.7341

H0: 男性的變項平均值=女性的變項平均值

變項分別代入：年齡、收縮壓、舒張壓、BMI、高血壓

結論：根據 t-test，變項舒張壓、bmi 的 p-value 分別為 0.0066、0.0245，皆小於 α (0.05)，拒絕虛無假說，因此舒張壓、bmi 在不同性別間具有統計顯著的差異。其餘變項年齡、收縮壓、罹患高血壓之 p-value 皆大於 α (0.05) ，不拒絕虛無假說，因此在不同性別間無統計顯著的差異。

```
/*2*/
/*呈現表格*/
PROC MEANS DATA=data1_1 mean std;
class sex;
VAR age sbp dbp bmi;
RUN;

PROC FREQ DATA=data1_1;
     TABLE bp*sex ;
RUN;

/*檢定*/
```

```
PROC TTEST DATA=data1_1;
     CLASS sex;
     VAR age sbp dbp bmi bp;
RUN;
```

3. 以存活分析中的 Kaplan-Meier method 估計性別不同其罹患中風 (stroke) 的狀況，並利用 Log-rank test 檢定不同性別的存活曲線是否有不同？需呈現存活曲線(survival curve)圖、列出虛無假設、檢定結果與針對圖表結果加以闡釋。(中風：stroke=1，無中風：stroke=0) （25%）

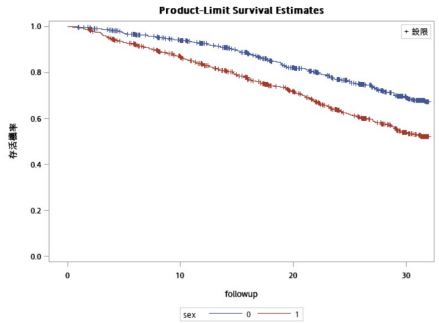
分層相等性的檢定

檢定	卡方	DF	Pr > 卡方
對數排名	48.2727	1	<.0001
Wilcoxon	52.6916	1	<.0001
-2Log(LR)	43.8461	1	<.0001

H0: $S_0(t)=S_1(t)$　　$S_0(t)$: sex=0

H1: $S_0(t)\neq S_1(t)$　　$S_1(t)$: sex=1

結論：根據 Log-rank test 檢定不同性別的存活曲線，p-value<0.0001<α(0.05) 拒絕 H0，$S_0(t)\neq S_1(t)$，不同性別間的存活曲線達統計顯著的差異。



女性的存活曲線皆高於男性的的存活曲線，即曲線下面積大於男性，女性中風的風險較男性低，性別不同有效影響中風情形。

```
/*3*/
/*Kaplan-Meier analysis and Log-Rank test*/
ods graphics on;
proc lifetest data=data1_1 method=pl plots=(survival);
/*method:Kaplan-Meier estimates*/
     time followup*stroke(0);
     strata sex;
run;
ods graphics off;
```

4. 利用 data2 進行分析。假設總膽固醇為變異數相等的常態分佈。BMI 依照過輕(<18.5)、正常(18.5-24)、過重(≥24)分為三組。將膽固醇視為反應變數，BMI 分組視為解釋變數，擬合回歸模型。請問膽固醇是否可以被 BMI 分組預測？請解釋結果，並說明判定係數 (R^2)。 (35%)

應變數: chol chol

來源	DF	平方和	均方	F 值	Pr > F
模型	2	103247.717	51623.858	39.80	<.0001
誤差	1997	2590354.361	1297.123		
已校正的總計	1999	2693602.078			

R 平方	變異係數	根 MSE	chol 平均值
0.038331	20.17330	36.01559	178.5310

來源	DF	類型 I SS	均方	F 值	Pr > F
bmi_g	2	103247.7169	51623.8584	39.80	<.0001

來源	DF	類型 III SS	均方	F 值	Pr > F
bmi_g	2	103247.7169	51623.8584	39.80	<.0001

參數	估計值		標準誤差	t 值	Pr > |t|
截距	185.5086705	B	1.22456686	151.49	<.0001
bmi_g 0	-27.1442637	B	3.53442084	-7.68	<.0001
bmi_g 1	-10.5725840	B	1.66583444	-6.35	<.0001
bmi_g 2	0.0000000	B	.	.	.

R-square = 0.038331，此迴歸模型中 bmi 分組可解釋膽固醇總變異的 3.8331%。

$Y=\beta_0+\beta_{1(1)}X_{1(1)}+\beta_{1(2)}X_{1(2)}+\epsilon$
Y: 膽固醇
ϵ: model 的誤差

BMI	X1(1)	X1(2)
過輕	1	0
正常	0	1
過重	0	0

β_0: model 的截距
$\beta_{1(1)}$: bmi 過輕相較於過重對膽固醇的影響。當固定其他變項後，bmi 過輕的膽固醇平均較過重減少 27.14mg/dL
$\beta_{1(2)}$: bmi 正常相較於過重對膽固醇的影響。當固定其他變項後，bmi 正常的膽固醇平均較過重減少 10.57mg/dL

$Y=185.50-27.14*X_{1(1)}-10.57*X_{1(2)}+\epsilon$

H0 : $\beta_i=0$; i=1(1) , 1(2)
H1 : $\beta_i\neq0$; i=1(1) , 1(2)

結論：$\beta_{1(1)}$與 $\beta_{1(2)}$的 p-value<0.0001<α(0.05)，皆拒絕 H0，$\beta_{1(1)}$與 $\beta_{1(2)}$與 0 達統計顯著的差異，因此膽固醇可以被 BMI 分組預測。

```
/*4*/
libname hw3 'C:\Users\user\OneDrive - 國立台灣大學\108-2\流病\期末考\SAS HW\SAS
Homework 3_for student' ;

/*bmi類別化*/
data data2_1; set hw3.hw3_data2;
bmi = weight / (height/100)**2 ;
      if bmi < 18.5 then bmi_g = 0 ;
      else if bmi < 24 then bmi_g = 1 ;
      else if bmi >= 24 then bmi_g = 2 ;
      else  bmi_g= . ;
run ;

/*linear regression*/
proc glm data=data2_1;
      class bmi_g; /*categorical variable*/model
      chol=bmi_g/solution; /*solution: to illustrate regression coefficients*/
run;
```

5. 假設我們對於社區居民糖尿病與抽菸狀況的關係有興趣。請將血糖值進行分組，血糖值≥126 mg/dL 視為糖尿病。目前已知與糖尿病相關的因子分別有：年齡、性別、體重和運動習慣。請針對以下題目進行解釋與說明結論。

(1) 請問糖尿病與抽菸之間的關係為何？（校正已知的干擾因子）(15%)

最大概度估計值分析

參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	-8.0155	0.8020	99.8956	<.0001
smoke	1	1	-13.4074	573.3	0.0005	0.9813
smoke	2	1	1.5004	0.7208	4.3325	0.0374
smoke	3	1	0.4008	0.2828	2.0089	0.1564
age		1	0.0544	0.00732	55.1234	<.0001
SEX	1	1	-0.3699	0.2835	1.7029	0.1919
weight		1	0.0349	0.00960	13.1779	0.0003
exercise	1	1	0.0413	0.2276	0.0329	0.8560

勝算比估計值和 Wald 信賴區間

效果	單位	估計值	95% 信賴界限	
smoke 1 與 0	1.0000	<0.001	<0.001	>999.999
smoke 2 與 0	1.0000	4.483	1.092	18.416
smoke 3 與 0	1.0000	1.493	0.858	2.599
age	1.0000	1.056	1.041	1.071
SEX 1 與 0	1.0000	0.691	0.396	1.204
weight	1.0000	1.035	1.016	1.055
exercise 1 與 0	1.0000	1.042	0.667	1.628

1. After adjusting for age, sex, weight, exercise habits,
   the OR of smoke=1 vs 0 is <0.001, p-value=0.9813, it's not significantly different from 0 at the 0.05 level;
   the OR of smoke=2 vs 0 is 4.483, p-value=0.0374, it's significantly different from 0 at the 0.05 level;
   the OR of smoke=3 vs 0 is 1.493, p-value= 0.1564, it's not significantly different from 0 at the 0.05 level.

2. After adjusting for age, sex, weight, exercise habits,

   the risk of smoke=1(僅嘗試吸過幾次而已) having diabetes was <0.001-fold higher than smoke=0(沒有吸過), and it is significantly different from 0 at the 0.05 level;

   the risk of smoke=2(有吸過，從以前到現在沒有吸超過 5 包（ 100 支）菸) having diabetes was 4.483-fold higher than smoke=0(沒有吸過), and it is significantly different from 0 at the 0.05 level;

   the risk of smoke=3(有吸過，從以前到現在有吸超過 5 包（ 100 支）菸) having diabetes was <1.493-fold higher than smoke=0(沒有吸過), and it is not significantly different from 0 at the 0.05 level.

```
/*5*/
/*糖尿病event*/
data hw3.hw3_data2_2; set hw3.hw3_data2;
     if glu >= 126 then diabetes = 1 ;
     else diabetes = 0 ;
run;
/*fit logistic model*/
proc logistic data=hw3.hw3_data2_2;
     class smoke(ref='0') sex(ref='0') exercise(ref='0')/ param=ref;
     model diabetes(event='1')=smoke age sex weight exercise/ risklimits;
run;
```

(2) 請問哪些干擾因子可能會影響糖尿病與抽菸之間的關係 (15%)

Age, weigh 之係數的 p-value 分別為<0.0001, 0.0003 ，皆<α (0.05)，拒絕 H0 ，與 0 達統計顯著的差異，因此會對 model 造成影響，而影響糖尿病與抽菸間的關係。