進階應用統計第三次作業建議解答

December 9/2004 Due　　Fall 2004

1. 以下 100 筆資料乃自常態分配 $N(\mu,\sigma^2)$ 抽出的隨機樣本(已排序)：

12.8	42.7	51.2	62.5	73.9
14.8	42.8	51.7	62.7	74.2
21.5	43.5	51.9	62.9	74.8
22.2	44.2	52.4	63.1	75.2
23.8	44.2	53.0	63.8	75.5
30.0	45.1	53.3	63.9	75.8
30.5	45.1	53.6	63.9	76.1
32.2	45.7	56.1	65.5	76.9
32.7	47.2	57.4	65.9	77.2
33.7	47.4	57.5	67.0	78.4
34.1	48.0	57.5	67.1	79.6
34.9	48.3	58.4	67.2	80.1
35.6	49.1	59.1	67.3	80.9
36.1	49.1	59.5	68.6	82.6
36.9	49.2	59.7	69.0	83.7
37.5	49.4	60.1	70.8	83.7
40.8	49.6	60.2	71.2	84.0
41.3	49.7	60.8	72.2	84.3
42.0	50.8	61.4	73.2	84.7
42.5	50.9	61.8	73.8	85.7

(a) 在不借助於樣本平均數及樣本變異數，估計期望值 μ 、變異數 σ^2 。

(b) 由你/妳從(a)估計出的期望值及變異數，驗證這 100 筆資料是否服從常態分配。(註：不能使用圖表，建議使用期望值與標準差！)

(c) 假設這些資料在記錄時，意外地將 90 筆來自 $N(\mu_1,\sigma^2)$分配、10 筆來自 $N(\mu_2,\sigma^2)$分配混在一起，其中 $\mu_1 \neq \mu_2$ 。請說明如何判斷 μ_1,μ_2 兩者何者較大，同時大略估計 μ_1,μ_2 兩者的差異大小。

Note: 上述資料是由 90 筆來自 $N(60,15^2)$分配、10 筆來自 $N(30,15^2)$分配組成。

Variable	N	Mean	Median	TrMean	StDev	SE Mean
第一組	90	59.45	59.94	59.53	14.75	1.55
第二組	10	29.76	26.93	28.40	13.92	4.40
合併	100	56.49	57.56	57.01	17.13	1.71

Variable	Minimum	Maximum	Q1	Q3
第一組	30.55	85.79	48.26	72.52
第二組	12.89	57.54	19.90	38.56
合併	12.89	85.79	44.51	70.38

→(a) 平均數(期望值)可由中位數代替，標準差可由 $\frac{\text{全距}}{4}$ 或 $\frac{\text{全距}}{6}$ 近似。本題中的樣本平均數為 56.49 與中位數 57.56 很接近；樣本標準差為 17.13 與 $\frac{\text{全距}}{4}=18.23$ 較為接近，而與 $\frac{\text{全距}}{6}=12.15$ 相去較遠。

(b) 樣本平均數加減一倍標準差約可涵蓋 68%的觀察值，樣本平均數加減兩倍標準差約可涵蓋 95%的觀察值。依此想法檢查，我們發現：

$$(\hat{\mu}\pm\hat{\sigma}) = (45.41, 69.71) \Rightarrow \text{包含 48 個觀察值。}$$

$$(\hat{\mu}\pm 2\hat{\sigma}) = (33.26, 81.86) \Rightarrow \text{包含 84 個觀察值。}$$

與預期有一段相當大的差距。(若代入 18.23 為標準差則有較佳的結果，分別有 69、98 個點落在區間內。)

(c) 由中位數及平均數的差異，可推得是否有少數的「離群值」，本例中的中位數較大表示有少數的較小數值，也就是$\mu_1 > \mu_2$。另外，因為中位數較不受離群值影響，合併後的平均數滿足$\mu = 0.9\mu_1 + 0.1\mu_2$，合併後的中位數與平均數差值大約等於 $0.1(\mu_1 - \mu_2)$，$\mu_1 - \mu_2 \cong 10(57.56\text{-}56.49)=10.7$；較好的修正是將中位數也向上修正 5 個數，或是 59.9(59.7 與 60.1 的平均)，如此的修正將使$\mu_1 - \mu_2 \cong 10(59.9\text{-}56.49)=35.1$ 較接近實際的差值 30。

2. 問卷分析的首要步驟為確定抽出的樣本與母體有相似的特性，也就是說樣本能反映母體，這就是上課提到的「樣本代表性」。一般而言，樣本代表性無法直接確定，通常藉由比較母體與樣本的常見特性(如：性別、年齡等結構)，以卡方的適合度檢定比較母體與樣本是否相似。仿照上課曾以 TVBS 民調討論如何定義問題及設計問卷題目，各組需說明調查的母體範圍，接著再根據母體蒐集相關的常見特性，比較母體與樣本是否相似。

   (a) 台北市長滿意度調查

   (b) 國親合併後政黨名稱調查

   (c) 桃園縣長滿意度與缺水民調

   （註：本題需要至 TVBS 網站下載原始資料，再以 Excel 等軟體整理資料。）

→以台北市長滿意度調查為例，在此使用「行政區域」、「年齡」、「性別」三個問項，測試樣本代表性。（註：人口資料可至內政部統計處下載。）

(1)下表為 838 位填寫行政區域的受訪者（剔除 17 位）之統計：

	中正	萬華	大同	中山	松山	信義	大安	文山	士林	北投	內湖	南港
樣本	36	67	33	60	61	75	75	88	118	81	109	35
預期值	50.91	63.88	41.29	69.22	65.58	74.83	99.84	81.82	92.25	79.43	82.87	36.06
母體	159599	200266	129440	216999	205593	234590	313011	256506	289194	249029	259789	113122

註：母體為 2003 年年底數值。

根據上列數值計算出卡方檢定值為 29.8731，大於卡方分配($\chi^2_{0.95}(11) = 19.675$ 及 $\chi^2_{0.99}(11) = 24.725$)臨界值，因此在行政區域問項上樣本與母體明顯不同。

(2) 下表為 849 位填寫年齡的受訪者（剔除 6 位）之統計：

	16～19	20～29	30～39	40～49	50～59	60 以上
樣本	48	119	172	223	142	145
預期值	55.24	158.88	170.98	184.80	127.41	151.68
母體	137603	395784	425924	460354	317379	377843

註：母體為 2003 年年底數值，16 至 19 歲的數值為 15 至 19 歲的 80%所得。

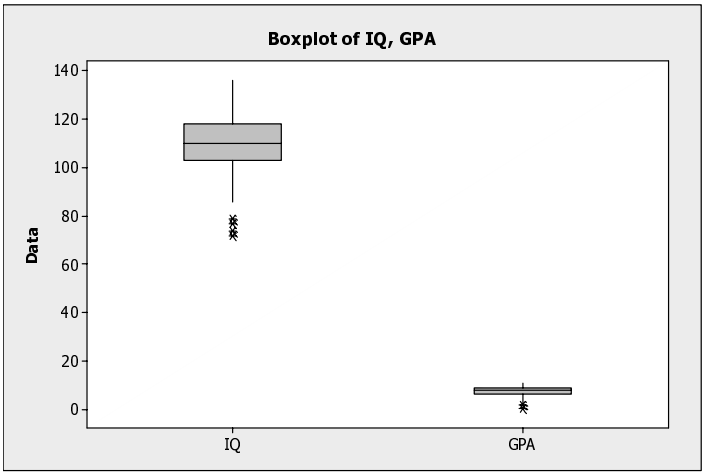
根據上列數值計算出卡方檢定值為 20.8261，大於卡方分配($\chi^2_{0.95}(5)=9.48773$ 及 $\chi^2_{0.99}(5)=15.0863$)臨界值，因此在年齡結構上樣本與母體明顯不同。

(3)在 854 位填寫性別的受訪者中（剔除 1 位），有 342 位男性、512 位女性。與母體的男女比例 48.414%及 51.586%比較，得出卡方檢定值 23.9393，相當於 p-value 為 0.0000，因此在性別比例上樣本與母體明顯不同。

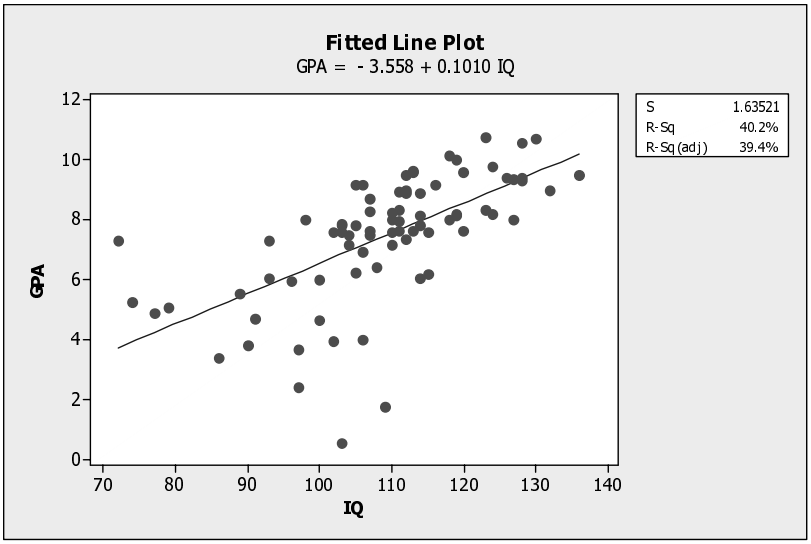
綜合以上三項，可知母體與樣本有不同的結構。

3. 離群值(Outlier)是較為不同的觀察值，然而何謂「不同」在不同問題或標準下不見得有統一的認定。請以美國中西部 78 位國中一年級學生的 GPA 與 IQ 資料，加上 IQ 比較高的人在校成績表現是否也比較好的考量，決定哪些觀察值可能是離群值，並詳細說明原因。

→(a) 如果由個別的IQ、GPA的boxplot來看，兩者各有4個及2個離群值。



(b) 由 IQ 及 GPA的散佈圖可看出這兩個變數間呈現線性正相關，相關係數為 0.634。以迴歸分析來看(檢視殘差值)， 有3個可能的離群值：(109,1.760), (103,0.530), (72,7.285)。



4. 延續第二題，請以調查資料進行統計分析，整理出：

(a) 各問題的調查結果

(b) （較有意義）問題間的交叉分析

根據你/妳的分析結果，訂出本調查的可能題目，並與 TVBS 的題目比較。

→TVBS 網站的原始報告未包含交叉分析，因此將著重於分析各問項中的關聯性，其中因為馬市長的支持者過去多為北市南區或是女性選民，以下就第二題的滿意度與居住地區、年齡、性別分別計算卡方檢定值，列表於下：

	卡方檢定值	p-value	d.f.	備　註
行政區域	19.793	0.596	22	樣本數 745
年齡	27.110	0.007	12	樣本數 745;年輕族群的滿意度較高
性別	5.294	0.151	3	樣本數 757

註：(1) 行政區域因樣本數不足，合併「不太滿意」、「很不滿意」為「不滿意」；

(2) 年齡因樣本數不足，合併「16~19」、「20~29」為「29 歲以下」。

以下為交叉檢定的建議操作程序：（在此僅以「是否支持馬市長選總統」、「年齡層」兩者為例）

Tabulated statistics: 是否支持，年齡層

Rows: 是否支持　　Columns: 年齡層

	1	2	3	4	5	6	72	99	All
1	30	69	96	104	65	64	0	1	429
2	14	31	40	62	28	26	0	0	201
5	0	0	0	0	0	0	1	0	1
95	4	19	36	57	49	55	0	4	224
All	48	119	172	223	142	145	1	5	855

明顯看出有資料輸入的問題，也有個數不滿5個的情形，需要先行處理。建議只針對「是否支持」＝1, 2, 95 與「年齡層」＝1, 2, 3, 4, 5, 6 作交叉分析，數據如下：

30	69	96	104	65	64
14	31	40	62	28	26
4	19	36	57	49	55

由第1行、第6行與第2列、第3列的數字可知這兩個問項不獨立，以下為檢定結果：

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	16-19	20-29	30-39	40-49	50-59	60+	Total
支持	30	69	96	104	65	64	428

	24.20	59.99	86.71	112.42	71.59	73.10	
	1.391	1.353	0.996	0.631	0.606	1.132	
不支持	14	31	40	62	28	26	201
	11.36	28.17	40.72	52.80	33.62	34.33	
	0.611	0.284	0.013	1.605	0.939	2.021	
拒答	4	19	36	57	49	55	220
	12.44	30.84	44.57	57.79	36.80	37.57	
	5.725	4.543	1.648	0.011	4.047	8.082	
Total	48	119	172	223	142	145	849

Chi-Sq = 35.637, DF = 10, P-Value = 0.000