【統計、經建行政、工業行政、農業行政、交通技術】

《統計學》

試題評析

今年高考統計學考題內容涵蓋機率分配、期望值及變異數的計算、迴歸係數的點估計及區間估計、應變數平均值的信賴區間估計、應變數值的預測區間估計、一因子變異數分析及 Tukey 多重比較法。考題略難，且部份考題計算需時較長，例如迴歸部份資料較多，且需以分組資料的計算方式來估算迴歸係數；而 Tukey 多重比較法是所有多重比較法中較少考的，對這部份不熟稔的學生也不易拿分。一般考生考 50 分以上不是問題，程度好的考生應可拿 70 分以上。

一、若某一種大樂透（lottery）係顧客從號碼 1 至 49 中挑選 6 個號碼，但不考慮 6 個號碼之前後次序，且每次開獎係由號碼球形質大小重量相同之二部開獎機中抽選一部開獎機執行開獎，先開 6 個中獎號碼（winning numbers），再開特別號（special number）。若開獎後，與 6 個中獎號碼完全相同者，則中獎者皆可獲$2,500,000；若與特別號及 5 個中獎號碼完全相同者，則中獎者皆可獲$1,200,000；若與 5 個中獎號碼完全相同者，則中獎者皆可獲$100,000；若與 4 個中獎號碼完全相同者，則中獎者皆可獲$1,000；若與 3 個中獎號碼完全相同者，則中獎者皆可獲$200。則：

（一）請定義隨機變數（random variable）以表示樂透玩家各種中獎事件。並列出相對應之衍生樣本空間（induced sample space）。（5 分）

（二）請推導（Derive）該隨機變數之機率函數（probability function），並試求其機率各為何？（15 分）

（三）試問樂透玩家中獎金額之期望數（expected number）與變異數（variance）？（10 分）

答：

(一)令 $X=\begin{cases}1,\text{表示樂透玩家中頭獎}\\2,\text{表示樂透玩家中二獎}\\3,\text{表示樂透玩家中三獎}\\4,\text{表示樂透玩家中四獎}\\5,\text{表示樂透玩家中五獎}\\6,\text{表示樂透玩家沒中獎}\end{cases}$，而 $Y=\begin{cases}2500000, X=1\\1200000, X=2\\100000, X=3\\1000, X=4\\200, X=5\\0, X=6\end{cases}$ 表樂透玩家可能獲得的獎金，

則可能中獎情形的樣本空間爲 $\Omega=\{x \mid x=1,2,3,4,5,6\}$

(二)

x	1	2	3	4	5	6
$f_X(x)$	$\frac{1}{\binom{49}{6}}$	$\frac{\binom{6}{5}\binom{1}{1}}{\binom{49}{6}}$	$\frac{\binom{6}{5}\binom{42}{1}}{\binom{49}{6}}$	$\frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}}$	$\frac{\binom{6}{3}\binom{43}{3}}{\binom{49}{6}}$	$1-f_X(1)-f_X(2)-f_X(3)-f_X(4)-f_X(5)$

(三)由(二)可得 Y 之分配如下：

y	2500000	1200000	100000	1000	200	0
$f_Y(y)$	$\frac{1}{\binom{49}{6}}$	$\frac{\binom{6}{5}\binom{1}{1}}{\binom{49}{6}}$	$\frac{\binom{6}{5}\binom{42}{1}}{\binom{49}{6}}$	$\frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}}$	$\frac{\binom{6}{3}\binom{43}{3}}{\binom{49}{6}}$	$1-f_Y(1)-f_Y(2)-f_Y(3)-f_Y(4)-f_Y(5)$

$$\therefore E(Y)=\sum y\cdot f_Y(y)=6.9944$$

$$E(Y^2)=\sum y^2 f_Y(y)=1246685$$

$$Var(Y)=E(Y^2)-[E(Y)]^2=1246685-(6.9944)^2=1246636$$

二、茲為研究某地區夫妻就業所得之關係，乃於該區隨機抽出 100 對夫妻之月薪（單位：千元）作為樣本之次數分配表如下；惟研究者若以妻之就業所得 x 為自變數（independent variable），而夫之就業所得 y 為因變數(dependent variable)之線性迴歸模式（linear regression model）為 $y=\alpha+\beta x+\varepsilon$ ，以研究夫妻就業所得之關係。

x \ y	195-200	200-205	205-210	210-215	215-220	220-225	225-230	230-235	235-240
190-195								1	2
185-190						1	2	3	1
180-185				2	3	5	3	2	
175-180			1	2	6	5	2		
170-175			1	6	8	4	1	2	
165-170		1	3	4	5	1	1		
160-165		2	4	4	2	1			
155-160	1	2	3	1					
150-155	1	1							

(一)簡要說明線性迴歸模式之必要假設為何？（5分）

以下各小題滿足線性迴歸模式之必要假設。

(二)試求 α 與 β 之估計值，並列出線性迴歸方程式（linear regression equation）。（10分）

(三)試列出 α 之99%信賴區間估計式（estimator），並求 α 之99%信賴區間估計值（estimate）。（5分）

(四)試列出 β 之99%信賴區間估計式，並求 β 之99%信賴區間估計值。（5分）

(五)若妻之就業所得為180，試求夫就業所得平均數之95%信賴區間估計值。（5分）

(六)若妻之就業所得為180，試求夫就業所得之95%信賴區間估計值。（5分）

答：(一) $Y_i=\alpha+\beta X_i+\varepsilon_i, i=1,2,\ldots,n$

1. $E(\varepsilon_i)=0$
2. $Var(\varepsilon_i)=\sigma^2$
3. $Cov(\varepsilon_i,\varepsilon_j)=0$ ，$\forall i\neq j$
4. $\varepsilon_i \overset{i.i.d.}{\sim}$ Normal（若欲找 β_0 、 β_1 之 MLE 或對 β_0 、 β_1 作區間估計、檢定等統計推論時需此假設）。

(二)1. α,β 的估計式為

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}=\frac{SSXY}{SSX}$$

$$=\frac{\sum_{i=1}^{n}(X_i)(Y_i)-n(\overline{X})(\overline{Y})}{\sum_{i=1}^{n}(X_i)^2-n(\overline{X})^2}=\frac{3769788-100(\frac{17295}{100})(\frac{21760}{100})}{2999875-100(\frac{17295}{100})^2}=0.7347$$

$$\hat{\alpha}=\overline{Y}-\hat{\beta}\,\overline{X}=\frac{21760}{100}-0.734714(\frac{17295}{100})=90.5313$$

2. $\hat{y}=\hat{\alpha}+\hat{\beta}x=90.5313+0.7347x$

(三)若欲求 α 的 99%C.I.，則（一）之假設中的 1.、2.、3.、4.皆需滿足。

其 99%信賴區間估計式為：

$$(\hat{\alpha}-t_{0.005}(n-2)S(\hat{\alpha}),\hat{\alpha}+t_{0.005}(n-2)S(\hat{\alpha}))$$

$$=(90.5313-2.576\cdot(11.0149),90.5313+2.576\cdot(11.0149))=(62.1570,118.9055)$$

其中 $S(\hat{\alpha})=\sqrt{[\frac{1}{n}+\frac{\overline{X}^2}{SSX}]MSE}=\sqrt{[\frac{1}{100}+\frac{(\frac{17295}{100})^2}{2999875-100(\frac{17295}{100})^2}](35.2055)}=11.0149$ 為 $\hat{\alpha}$ 的估計標準誤，而

$$\text{MSE}=\frac{SSE}{n-2}=\frac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}\,X_i)^2}{n-2}=\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2-\hat{\beta}^2\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-2}$$

$$=\frac{4743125-100(\frac{21760}{100})^2-(0.7347)^2\left[2999875-100(\frac{17295}{100})^2\right]}{100-2}=35.2055$$

(四)若欲求 β 的 99%C.I.，則（一）之假設中的 1.、2.、3.、4.皆需滿足。

其 99%信賴區間估計式為：

$$(\hat{\beta}-t_{0.005}(n-2)S(\hat{\beta}),\hat{\beta}+t_{0.005}(n-2)S(\hat{\beta}))$$

$$=(0.7347-2.576\cdot0.0636,0.7347+2.576\cdot0.0636)=(0.5709,0.8985)$$

其中 $S(\hat{\beta})=\sqrt{\frac{MSE}{SSX}}=\sqrt{\frac{35.2055}{2999875-100(\frac{17295}{100})^2}}=0.0636$ 為 $\hat{\beta}$ 的估計標準誤。

(五)1. X_h=180 時 $E(Y_h)$ 的點估計為

$$E(\hat{Y_h})=\hat{y}=90.5313+0.7347x=90.5313+0.7347(180)=222.7797$$

2. $E(Y_h)$ 的 100$(1-\alpha)$%區間估計為：

$$E(\hat{Y}_h) \pm t_{\frac{\alpha}{2}}(n-2)S(E(\hat{Y}_h))$$

$$= (222.7797 - 1.96 \cdot 0.7437, 222.7797 + 1.96 \cdot 0.7437) = (221.3221, 224.2374)$$

其中 $S(E(\hat{Y}_h)) = \sqrt{\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SSX}\right] MSE} = 0.7437$

(六)1. X_h=180 時 $E(Y_h)$ 的點估計爲

$$E(\hat{Y}_h) = \hat{y} = 90.5313 + 0.7347x = 90.5313 + 0.7347(180) = 222.7797$$

2. Y_h 的 100$(1-\alpha)$%預測區間估計爲：

$$E(\hat{Y}_h) \pm t_{\frac{\alpha}{2}}(n-2)S(Y_h - E(\hat{Y}_h))$$ ，

$$= (222.7797 - 1.96 \cdot 5.9798, 222.7797 + 1.96 \cdot 5.9798) = (211.0592, 234.5002)$$

其中

$$S(Y_h - E(\hat{Y}_h)) = \sqrt{\left[1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SSX}\right] MSE} = \sqrt{\left[1 + \frac{1}{100} + \frac{(180 - \overline{X})^2}{SSX}\right] (35.2055)} = 5.9798$$

三、茲為瞭解台灣地區五星級觀光大飯店營業所得之實情，乃於該區隨機抽出 4 家五星級觀光大飯店同時調查 6 週營業所得（單位：千元新台幣）為樣本如下表。惟若本題滿足單因子變異數分析設計之必要假設。

週別 / 觀光大飯店	1	2	3	4	5	6
1	1430	2200	1140	880	1670	990
2	980	1400	1200	1300	1300	550
3	1780	2890	1500	1470	2400	1600
4	2300	2680	2000	1900	2540	1900

(一)請建構單因子變異數分析（one-way ANOVA）之數學模式與單因子變異數分析表（one-way ANOVA Table）。(10 分)

(二)若顯著水準為 5%，試以危險域法(critical region approach)與 p 值法(p-value approach)檢定 4 家五星級觀光大飯店之平均營業所得是否不同？(7 分)

(三)若顯著水準為 5%，試以 Tukey-Kramer procedure 查明 4 家五星級觀光大飯店之平均營業所得不同者為何？(10 分)

(四)請簡要說明單因子變異數分析設計之必要假設為何？試討論本題是否滿足單因子變異數分析設計之必要假設？請簡要說明理由為何？但若本題不滿足單因子變異數分析設計之必要假設，請簡要說明合適之設計與理由為何？(8 分)

答：(一)1.one-way ANOVA 的數學模式爲

$$Y_{ij} = \mu_i + \varepsilon_{ij} = (\mu + \alpha_i) + \varepsilon_{ij} \ , \ i = 1, \cdots, k \ , \ j = 1, \cdots, n_i \ , \ n = \sum_{i=1}^{k} n_i$$

且 $\varepsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2), \forall i, j$（$\therefore Y_{ij} \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2), \forall i, j$）

其中，Y_{ij}：第 i 個因子水準的第 j 個觀察值。

μ_i：爲第 i 個因子水準的平均值（爲未知常數參數）。

$\mu = \frac{1}{k}\sum_{i=1}^{k}\mu_i$ 爲母體總平均。

$\alpha_i = \mu_i - \mu$ 稱爲第 i 個因子水準的主效果（$\sum_{i=1}^{k}\alpha_i = 0$）。

ε_{ij}：爲隨機誤差項（error term）。

$Y_{\bullet\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$ ，$\bar{Y}_{\bullet\bullet} = \frac{Y_{\bullet\bullet}}{n} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{n}$ 。

2.模型假設檢定

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ versus $H_1 : \mu_i$ 不全相等

$\Leftrightarrow H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ versus $H_1 : \alpha_i$ 不全爲 0

3.變異數分析表（ANOVA Table）：

變異來源	自由度(df)	平方和	均方和	F 值	p-value
處理	k-1	SSTR(SSB)	$MSTR = \frac{SSTR}{k-1}$	$F = \frac{MSTR}{MSE}$	$P(F > \frac{MSTR}{MSE})$
誤差	n- k	SSE(SSW)	$MSE = \frac{SSE}{n-k}$		
總和	n -1	SSTO(SST)			

4.決策：

(1)Reject $H_0 \Leftrightarrow F = \frac{MSTR}{MSE} > F_\alpha(k-1, n-k)$ 或

(2)Reject $H_0 \Leftrightarrow$ p-value= $P(F > \frac{MSTR}{MSE}) < \alpha$

其中 $SSTO = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{n}$

$SSTR = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{n}$

$SSE = SSTO - SSTR$

(二)

飯店	1	2	3	4	
	1430	980	1780	2300	
	2200	1400	2890	2680	
	1140	1200	1500	2000	
	880	1300	1470	1900	
	1670	1300	2400	2540	
	990	550	1600	1900	
$Y_{i\bullet}$	8310	6730	11640	13320	40000
$\bar{Y}_{i\bullet}$	1385	1121.667	1940	2220	1666.667

① $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

② $H_1 : \mu_i$ 不全相等

③ α = 0.05

④C={ $F \mid F > F_{0.05}(3,20) = 3.0984$ }，其中 $P(F > F_\alpha) = \alpha$

⑤ $F = \dfrac{MSTR}{MSE} = \dfrac{SSTR/(k-1)}{SSE/(n-k)} = \dfrac{4543500/(4-1)}{3956033/(24-4)} = 7.6567 \in C$

其中 $SSTO = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \dfrac{Y_{\bullet\bullet}^2}{n}$

$= \left[(1430)^2 + \cdots + (1900)^2\right] - \dfrac{(40000)^2}{24} = 8499533$

$SSTR = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\dfrac{Y_{i\bullet}^2}{n_i} - \dfrac{Y_{\bullet\bullet}^2}{n}$

$= \dfrac{1}{6}\left[(8310)^2 + \cdots (13320)^2\right] - \dfrac{(40000)^2}{24} = 4543500$

$SSE = SSTO - SSTR$ = 8499533-4543500 = 3956033

而其 p-value= $P(F > 7.6567) \cong 0.0013 < 0.05$

⑥結論：拒絕 H_0，亦即在顯著水準 α = 0.05 下，我們的資料顯示四家五星級觀光大飯店之平均營業所得有顯著的差異。

ANOVA TABLE						
變源	SS	自由度	MS	F	P-值	臨界值
組間	4543500	3	1514500	7.656659	0.00134	3.098393
組內	3956033	20	197801.7			
總和	8499533	23				

(三)1.Tukey-Kramer 多重比較法（Tukey-Kramer multiple comparison）：

若 $n_1 = n_2 = \cdots = n_k = n_*$，則

$$(\bar{Y}_{i\bullet} - \bar{Y}_{j\bullet}) \pm \frac{1}{\sqrt{2}} q_\alpha(k, n-k) \times \sqrt{MSE \times \left(\frac{1}{n_*} + \frac{1}{n_*}\right)}$$

稱爲 $100(1-\alpha)\%$ Tukey 聯立信賴區間

2.各95%的Tukey聯立信賴區間如下：

$$\left[L_{12}, U_{12}\right] = (\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}) \pm \frac{1}{\sqrt{2}} q_\alpha(k, n-k) \times \sqrt{MSE \times \left(\frac{1}{n_*} + \frac{1}{n_*}\right)}$$

$$= (1385 - 1121.667) \pm \frac{1}{\sqrt{2}} \times 3.9583 \times \sqrt{197801 \times \left(\frac{1}{6} + \frac{1}{6}\right)} = (-455.3751, 982.0417)$$

同理可得

$\left[L_{13}, U_{13}\right] = \left[-1273.7084, 163.7084\right]$

$\left[L_{14}, U_{14}\right] = \left[-1553.7084, -116.2916\right]$

$[L_{23},U_{23}]=[-1537.0417,-99.6249]$

$[L_{24},U_{24}]=[-1817.0417,-379.6249]$

$[L_{34},U_{34}]=[-998.7084,438.7084]$

因爲$[L_{12},U_{12}]$、$[L_{34},U_{34}]$包含0，而$[L_{14},U_{14}]$、$[L_{23},U_{23}]$、$[L_{24},U_{24}]$、$[L_{13},U_{13}]$不包含0，所以我們可得如下結論：

$\mu_1=\mu_2\neq\mu_3=\mu_4$

亦即在顯著水準0.05下，由Tukey-Kramer Procedure我們知道第1家和第2家觀光大飯店的平均營業所得差不多，第3家和第4家觀光大飯店的平均營業所得差不多，而第1、2家與第3、4家的平均營業所得有顯著的差異。

(四)1. $\varepsilon_{ij}\overset{i.i.d.}{\sim}N(0,\sigma^2),\forall i,j$（$\therefore Y_{ij}\overset{i.i.d.}{\sim}N(\mu_i,\sigma^2),\forall i,j$）

亦即各組處理誤差的平均數爲 0，變異數爲齊一，且其值服從常態分配，而組內各實驗值之間無關。就此題而言，上述檢定需假設四家五星級觀光大飯店之營業所得所對應的母體分配皆爲常態分配，且變異數必須相同。

2.欲檢定各組母體是否變異數齊一，可採 Bartlett test

① $H_0:\sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2=\sigma^2$

② $H_1:\sigma_i^2$ 不全相等

③ α = 0.05

④C={ $\chi^2\mid\chi^2>\chi^2_{0.05}(3)=7.815$ }其中$P(\chi^2>\chi^2_{\alpha})=\alpha$

⑤

飯店	營業額						$Y_{i\bullet}$	S_i^2	$\overline{Y}_{i\bullet}$
1	1430	2200	1140	880	1670	990	8310	243710	1385
2	980	1400	1200	1300	1300	550	6730	98816.67	1121.667
3	1780	2890	1500	1470	2400	1600	11640	333960	1940
4	2300	2680	2000	1900	2540	1900	13320	114720	2220

$$MSE=\frac{SSE}{n-k}=\frac{5(243710)+5(98816.67)+5(333960)+5(114720)}{24-4}=197801.7$$

$$C=1+\frac{1}{3(k-1)}[(\sum_{i=1}^{k}\frac{1}{n_i-1})-\frac{1}{n-k}]$$

$$=1+\frac{1}{3(4-1)}[4(\frac{1}{5})-\frac{1}{20}]\cong 1.0833$$

$$B=\frac{2.3026}{C}[(n-k)\log_{10}(MSE)-\sum_{i=1}^{k}(n_i-1)\log_{10}S_i^2]$$

$$=\frac{2.3026}{1.0833}[(20)(5.2962)-5(5.3869+4.9948+5.5237+5.0596)]\cong 2.3368\notin C$$

⑥結論：無法拒絕H_0，亦即在顯著水準 α = 0.05 下，我們的資料顯示各母體變異數無顯著的差異。

3.若此題不滿足 one-way ANOVA 的各組母體需滿足常態分配的假設，則可改採無母數統計方法中的 Krustal-Wallis 檢定。