《統計學》

試題評析

今年度考題第一、二、四題都屬於觀念題，計算容易，中上程度考生應該都可以拿到分數，唯獨第三題需要推導。第一題在總複習講義P51《例題5》；第二題（一）在總複習講義P65《例題1》、（二）在總複習講義P65、（三）在總複習講義P62《例題2》、（四）在總複習講義P3；第三題在迴歸分析正規講義第一回P71《例題6》、第四題在總複習講義P30有談到偏性與效性。

一、某年全省民營加油站經省度量衡檢定所抽查檢驗油量計1795具，發現超過法定誤差範圍的有24具，而公營加油站抽驗的940具中，只有7具不合格。A先生認為24/1795>7/940故度量衡檢定所應對民營加油站加強檢驗工作。而B先生則認為應該進行統計檢定，並選擇 $H_0: P_{公} \le P_{民}$ vs. $H_1: P_{公} > P_{民}$ 和顯著水準 $\alpha = 0.05$，但是C先生卻認為虛無和對立假設應設 $H_0: P_{公} \ge P_{民}$ vs. $H_0: P_{公} < P_{民}$。（已知 $Z_{0.05} = 1.645$，$Z_{0.025} = 1.96$）

(一)請協助B先生進行統計檢定。（10分）

(二)請問B先生和C先生的虛無和對立假設之設定對統計檢定的影響。（5分）

(三)請問您支持那一位（A,B,C）的結論？請說明理由。（5分）

答：

$$\hat{P}_{民} = \frac{24}{1795} = 0.0134 \ , \ \hat{P}_{公} = \frac{7}{940} = 0.0074$$

(一)1. $\begin{cases} H_0: P_{公} \le P_{民} \\ H_1: P_{公} > P_{民} \end{cases}$

2. $C = \left\{ Z \,\middle|\, Z > Z_{0.05} = 1.645 \right\}$

3. $Z = \dfrac{\hat{P}_{公} - \hat{P}_{民}}{\sqrt{\dfrac{\hat{P}_{公}(1-\hat{P}_{公})}{n_1} + \dfrac{\hat{P}_{民}(1-\hat{P}_{民})}{n_2}}} = -1.54 \notin C$

$\therefore$ Not reject H_0，民營不合格率較高。

(二) $\hat{P}_{公} < \hat{P}_{民} \iff \hat{P}_{公} - \hat{P}_{民} < 0$

則檢定統計量 $Z = \dfrac{\hat{P}_{公} - \hat{P}_{民}}{\sqrt{\dfrac{\hat{P}_{公}(1-\hat{P}_{公})}{n_1} + \dfrac{\hat{P}_{民}(1-\hat{P}_{民})}{n_2}}} < 0$

1. B先生：

$$\begin{cases} H_0: P_{公} \le P_{民} \\ H_1: P_{公} > P_{民} \ \text{(右尾檢定)} \end{cases}$$

則檢定統計量 $Z \notin C$（統計值與拒絕域方向相反）

檢定結果必爲Not reject H_0

2.C先生：

(1) $\begin{cases} H_0 : P_{公} \geq P_{民} \\ H_1 : P_{公} < P_{民} \ \text{(左尾檢定)} \end{cases}$

(2) $C = \left\{ Z \mid Z < -Z_{0.05} = -1.645 \right\}$

(3) $Z = \dfrac{\hat{P}_{公} - \hat{P}_{民}}{\sqrt{\dfrac{\hat{P}_{公}(1-\hat{P}_{公})}{n_1} + \dfrac{\hat{P}_{民}(1-\hat{P}_{民})}{n_2}}} = -1.54 \notin C$

$\therefore$ Not reject H_0，公營不合格率較高。

(三)支持C

1. A先生：

點估計只抽一組樣本代入估計式後，用得到的估計值去估計參數 p ，估計誤差大

$\therefore$ 用 $\hat{p}$ 去度量檢驗並不適合

2. B先生：

$\because H_0$ 與 H_1 顯然放置不對，檢定結果一定是 Not reject H_0

$\therefore$檢定不具意義

3. C先生：

利用假設檢定，可靠度達 $(1-\alpha)\% = 95\%$

二、不動產公司欲比較3位不動產經紀人對不動產估價的一致性。這3位不動產經紀人皆評估該公司主管隨機選取的5棟建築物之價值（單位：萬元）。一位公司員工認為可用一因子變異數分析予以比較，而另一位員工則認為應該用二因子變異數分析予以比較，他們分別就上述資料求變異數分析表：

一因子

來源	SS	自由度	MS	F值	P-值
經紀人	(1)	(2)	(4)	(5)	0.95
誤差	14039700	(3)			
總和	14161950				

二因子

來源	SS	自由度	MS	F值	P-值
經紀人	(6)	(8)		(11)	0.046
建築物	13933666	(9)		(12)	0.0001
誤差	(7)	(10)			
總和	14161950				

(一)請完成此二個變異數分析表，寫出表中(1)至(12)的值。(10分)

(二)請問經紀人和建築物是隨機效果還是固定效果因子？請說明理由。(5分)

(三)您建議採用那一個變異數分析表進行統計檢定？請說明理由。(10分)

(四)何謂P-值？請據此，用選擇的表進行統計檢定和推論。(10分)

顯著水準 $\alpha = 0.05$

答：

(一)

(1) $14161950 - 14039700 = 122250$

(2) $3 - 1 = 2$ (3) $15 - 3 = 12$ (4) $\dfrac{122250}{2} = 61125$

(5) $\dfrac{61125}{14039700/12} = 0.0522$　　(6) 122250　　(7) 14039700 –13933666 = 106034

(8) 3 – 1 = 2　　(9) 5 – 1 = 4　　(10) $2 \times 4 = 8$

(11) $\dfrac{122250/2}{106034/8} = 4.6117$　　(12) $\dfrac{13933666/4}{106034/8} = 262.8151$

(二)執行ANOVA　R.B.D.檢定

模型假設 $\sum_{i=1}^{3} \alpha_i = 0$ ， $\sum_{j=1}^{5} \beta_j = 0$ (影響效應總和爲0)

Where α_i ：列因子主效果

β_j ：行因子主效果

∴ 固定效果因子

(三)

1.採用ANOVA　R.B.D (二因子變異數分析)

2.原因：

三位經紀人分別對相同的建築物估價（在相同的block之下做試驗）比較客觀，而且可以控制變因，降低實驗誤差（SSE），有效的檢定出主效果 α_i 是否有差異。

(四)

1. p – value：由樣本所得之統計量決定 reject H_0 時，所求最小顯著水準。

左尾 p – value =P (統計量 < 統計值 | H_0)

右尾 p – value = P (統計量 > 統計值 | H_0)

兩尾 p – value =2×min{P (統計量 < 統計值 | H_0),
P (統計量 > 統計值 | H_0)}

2. 二因子檢定：

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \\ H_1 : \text{不全爲}0 \end{cases}$$

p – value = 0.046 < α = 0.05

∴ reject H_0 ，三位經紀人估價不一致。

三、研究人員欲探討員工薪資(Y_{ij})與工作年資(X_i)的相關程度，而進行調查得下列資料型態：

工作年資	薪資	平均薪資
1	$Y_{11}Y_{12}\cdots Y_{1m_1}$	$\bar{Y}_{1.}$
2	$Y_{21}Y_{22}\cdots Y_{2m_2}$	$\bar{Y}_{2.}$
.	.	.
.	.	.
.	.	.
30	$Y_{301}Y_{302}\cdots Y_{30m_{30}}$	$\bar{Y}_{30.}$

研究人員欲配適 $Y_{ij}=\beta_0+\beta_1 X_i+\varepsilon_{ij}$, $j=1,2,...,m_i$ 。其研究助理為了簡化計算認為可藉由員工薪資平均後的模式 $\bar{Y}_{i.}=\alpha_0+\alpha_1 X_i+\varepsilon_{i.}$ 得 β_1 之估計值，其中 $\bar{Y}_{i.}=\frac{1}{m_i}\sum_{j=1}^{m_i}Y_{ij}$ 。請評估此種做法的可行性：

(一)寫出β_1和α_1的最小平方法的估計式。（10分）

(二)當m_i相同時$(m_i=m)$，探討β_1和α_1的最小平方法估計式之關係。（5分）

(三)當m_i不同時，探討β_1和α_1的最小平方法估計式之關係。（5分）

(四)那一條模式的R^2（判定係數）較高？（5分）

答：

(一) $n=30$

1. $$\hat{\beta}_1=\frac{S_{XY}}{S_{XX}}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(X_i-\bar{X})(Y_{ij}-\bar{Y})}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(X_i-\bar{X})^2} \quad , \quad \text{where } \bar{X}=\frac{\sum_{i=1}^{n}m_iX_i}{\sum_{i=1}^{n}m_i} \text{，} \bar{Y}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}Y_{ij}}{\sum_{i=1}^{n}m_i}$$

2. $$\hat{\alpha}_1=\frac{S_{XY}}{S_{XX}}=\frac{\sum_{i=1}^{n}(X_i-\bar{\bar{X}})(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})}{\sum_{i=1}^{n}(X_i-\bar{\bar{X}})^2} \quad , \quad \text{where } \bar{\bar{X}}=\frac{\sum_{i=1}^{n}X_i}{n} \text{，} \bar{Y}_{\bullet\bullet}=\frac{\sum_{i=1}^{n}\bar{Y}_{i\bullet}}{n}$$

(二)當 $m_i=m$ 時，則 $\bar{Y}_{\bullet\bullet}=\bar{Y}$ ，$\bar{X}=\frac{m\sum X_i}{nm}=\bar{\bar{X}}$

$$\therefore\ \hat{\beta}_1=\frac{S_{XY}}{S_{XX}}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(X_i-\bar{X})(Y_{ij}-\bar{Y})}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(X_i-\bar{X})^2}$$

$$=\frac{m\sum_{i=1}^{n}(X_i-\bar{\bar{X}})(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})}{m\sum_{i=1}^{n}(X_i-\bar{\bar{X}})^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{\bar{X}})(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})}{\sum_{i=1}^{n}(X_i-\bar{\bar{X}})^2}=\hat{\alpha}_1$$

(三) $\hat{\beta}_1=\dfrac{S_{XY}}{S_{XX}}=\dfrac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(X_i-\bar{X})(Y_{ij}-\bar{Y})}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(X_i-\bar{X})^2}$

$$=\frac{\sum\sum X_iY_{ij}-\dfrac{\left(\sum m_iX_i\right)\left(\sum\sum Y_{ij}\right)}{\sum m_i}}{\sum m_iX_i^2-\dfrac{\left(\sum m_iX_i\right)^2}{\sum m_i}}=\frac{\sum m_iX_i\bar{Y}_{i\bullet}-\dfrac{\left(\sum m_iX_i\right)\left(\sum m_i\bar{Y}_{i\bullet}\right)}{\sum m_i}}{\sum m_iX_i^2-\dfrac{\left(\sum m_iX_i\right)^2}{\sum m_i}}$$

$$\hat{\alpha}_1=\frac{S_{XY}}{S_{XX}}=\frac{\sum_{i=1}^{n}(X_i-\bar{\bar{X}})(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})}{\sum_{i=1}^{n}(X_i-\bar{\bar{X}})^2}$$

$$=\frac{\sum X_i\bar{Y}_{i\bullet}-\dfrac{\left(\sum X_i\right)\left(\sum \bar{Y}_{i\bullet}\right)}{n}}{\sum X_i^2-\dfrac{\left(\sum X_i\right)^2}{n}}=\frac{\sum X_i\bar{Y}_{i\bullet}-\dfrac{\left(\sum X_i\right)\left(\sum \bar{Y}_{i\bullet}\right)}{\sum 1}}{\sum X_i^2-\dfrac{\left(\sum X_i\right)^2}{\sum 1}}$$

$\begin{cases}\hat{\beta}_1\text{每一個}\sum\text{後皆爲}m_i\text{倍}\\ \hat{\alpha}_1\text{每一個}\sum\text{後皆爲1倍}\end{cases}$，其餘公式相同

表示 $\hat{\beta}_1$：i th個母體抽 m_i 個樣本

表示 $\hat{\alpha}_1$：每個母體只抽一個樣本

(四)model：$\bar{Y}_{i\bullet}=\alpha_0+\alpha_1x_i+\varepsilon_i$

i th個母體用 $\bar{Y}_{i\bullet}$ 取代 m_i 個樣本，會產生更大的抽樣誤差，即 $SSE\uparrow$，$SSR\downarrow$，因此 $R^2=\dfrac{SSR}{SSRO}\downarrow$

$\therefore$ model $Y_{ij}=\beta_0+\beta_1x_i+\varepsilon_{ij}$ R^2 較大

四、欲估計民眾對電費調整的支度(p)。一位研究人員分別由北、高兩地取樣，令 $\dfrac{X_1}{n_1}$ 和 $\dfrac{X_2}{n_2}$ 分別代表北、高兩地的樣本支持百分比，下列有兩種估計方式：

(1) $\dfrac{X_1+X_2}{n_1+n_2}$ 和(2) $\left(\dfrac{X_1}{n_1}+\dfrac{X_2}{n_2}\right)/2$

(一)那一種估計方式所得估計值較大？（5分）

(二)那一種估計方式具備不偏性？那一種估計方式的標準誤較大？（10分）

(三)您建議採用那一種估計方式？請說明理由。（5分）

答：

(一) $\dfrac{X_1+X_2}{n_1+n_2}-\left(\dfrac{X_1}{n_1}+\dfrac{X_2}{n_2}\right)\cdot\dfrac{1}{2}=\dfrac{2n_1n_2X_1+2n_1n_2X_2-n_1n_2X_1-n_1^2X_2-n_2^2X_1-n_1n_2X_2}{2n_1n_2(n_1+n_2)}$

$$=\frac{n_1n_2X_1+n_1n_2X_2-n_1^2X_2-n_2^2X_1}{2n_1n_2(n_1+n_2)}=\frac{n_2X_1(n_1-n_2)+n_1X_2(n_2-n_1)}{2n_1n_2(n_1+n_2)}$$

$$=\frac{(n_1-n_2)(n_2X_1-n_1X_2)}{2n_1n_2(n_1+n_2)}$$

∴ 當 $n_1=n_2$ 或 $n_2X_1=n_1X_2$ 時，兩估計式相等

當 $\begin{matrix} n_1>n_2 \\ n_1<n_2 \end{matrix}$ 且 $\begin{matrix} n_2X_1>n_1X_2 \\ n_2X_1<n_1X_2 \end{matrix}$ 時，$\dfrac{X_1+X_2}{n_1+n_2}$ 較大

當 $\begin{matrix} n_1<n_2 \\ n_1>n_2 \end{matrix}$ 且 $\begin{matrix} n_2X_1>n_1X_2 \\ n_2X_1<n_1X_2 \end{matrix}$ 時，$\left(\dfrac{X_1}{n_1}+\dfrac{X_2}{n_2}\right)\Big/2$ 較大

(二)1. $E\left(\dfrac{X_1+X_2}{n_1+n_2}\right)=\dfrac{n_1p+n_2p}{n_1+n_2}=p$

$$E\left(\left(\frac{X_1}{n_1}+\frac{X_2}{n_2}\right)\cdot\frac{1}{2}\right)=\left(\frac{n_1p}{n_1}+\frac{n_2p}{n_2}\right)\cdot\frac{1}{2}=p$$

∴皆為不偏估計式

2. $Var\left(\dfrac{X_1+X_2}{n_1+n_2}\right)=\dfrac{n_1p(1-p)+n_2p(1-p)}{\left(n_1+n_2\right)^2}=\dfrac{p(1-p)}{n_1+n_2}$

$$Var\left[\left(\frac{X_1}{n_1}+\frac{X_2}{n_2}\right)\cdot\frac{1}{2}\right]=\frac{1}{4}\left[\frac{p(1-p)}{n_1}+\frac{p(1-p)}{n_2}\right]=\frac{n_1+n_2}{4n_1n_2}p(1-p)$$

$$\frac{n_1+n_2}{4n_1n_2}-\frac{1}{n_1+n_2}=\frac{\left(n_1+n_2\right)^2-4n_1n_2}{4n_1n_2(n_1+n_2)}=\frac{\left(n_1-n_2\right)^2}{4n_1n_2(n_1+n_2)}\geq 0$$

當 $n_1\neq n_2$ 時，$\left(\dfrac{X_1}{n_1}+\dfrac{X_2}{n_2}\right)\Big/2$ 標準誤較大

(三)兩估計式均為不偏

但因 $\dfrac{X_1+X_2}{n_1+n_2}$ 標準誤較小

∴ 建議採用 $\dfrac{X_1+X_2}{n_1+n_2}$