經濟部所屬事業機構106年新進職員甄試試題

類別：統計資訊　　節次：第二節

科目：1.統計學 2.巨量資料概論

注意事項	1.本試題共6頁(含A3紙1張、A4紙1張)。 2.可使用本甄試簡章規定之電子計算器。 3.本試題為單選題共50題，每題2分，共100分，須用2B鉛筆在答案卡畫記作答，於本試題或其他紙張作答者不予計分。 4.請就各題選項中選出最適當者為答案，各題答對得該題所配分數，答錯或畫記多於1個選項者，倒扣該題所配分數3分之1，倒扣至本科之實得分數為零為止；未作答者，不給分亦不扣分。 5.本試題採雙面印刷，請注意正、背面試題。 6.考試結束前離場者，試題須隨答案卡繳回，俟本節考試結束後，始得至原試場或適當處所索取。 7.考試時間：90分鐘。

[D] 1. 隨機變數X服從一個參數為n和p的二項分配，下列敘述何者有誤？
(A)n大且p不要太小時，X近似常態分配　(B) n大且p不要太小時，X近似Poisson分配
(C)$p > 0.5$時X的分配為左偏　(D) $n-X$同樣是服從一個參數為n和p的二項分配

[C] 2. 我們想驗證台灣年輕人每星期花在精靈寶可夢(Pokémon Go)線上遊戲的平均時間是否為20小時，即在給定顯著水準α下檢定H_0: μ=20，令β為給定μ≠20的情況下犯了型二誤差之機率，下列敘述何者正確？
(A) $\alpha+\beta=1$　(B)若真正的$\mu=21$，則$\alpha>\beta$
(C)若α變大則β變小　(D)若μ變小則β變小

[A] 3. 以下ANOVA表為單一變數簡單迴歸分析的結果，令Y為反應變數，X為解釋變數，反應變數Y有多少百分比的變異量可以由變數X來解釋？

Source	df	SS	MS	F
Regression	1	300	300	4.50
Error	9	600	66.67	
Total	10	900		

(A) 33.3 %　(B) 50 %　(C) 66.7 %　(D) 100 %

[C] 4. 由最小平方法得到的迴歸方程式為$Y = 3.5 - 6X$，下列敘述何者有誤？
(A) X=0時Y的點估計為3.5　(B) X每增加1單位則Y減少6單位
(C) X和Y必為高度線性關係　(D) X和Y的樣本相關係數為負

[D] 5. 在具有多個自變數的複迴歸分析中，若同時考量模型的複雜度和解釋能力，以下那一個變量最適合用來當作模型選取的依據？
(A)每個迴歸係數對應的t值　(B)變異數影響因子(Variance Inflation Factor)
(C)判定係數(R^2)　(D)調整後之判定係數(adjusted R^2)

[D] 6. 假設台北市的全體居民平均每1,000之中有1人為原住民。如果我們採用置回抽樣(sampling with replacement)隨機查訪8,000位台北市居民，再令X代表這8,000人之中原住民的人數，則X的期望值$E(X)$是多少？
(A) $E(X)=2$　(B) $E(X)=4$　(C) $E(X)=6$　(D) $E(X)=8$

[B] 7. 如果某一母體具有常態分布，其變異數為σ^2，而我們想要檢定虛無假設H_0：$\sigma^2=\sigma_0^2$的真偽(σ_0^2為一定數)，那麼所使用的統計檢定量，和以下哪一種機率分布有直接的關聯性？
(A) t分布　(B)卡方分布　(C) F分布　(D)指數分布

[D] 8. 假設兩個隨機變數 X 與 Y 的聯合機率分布可表列如下:令$Cov(X,Y)$ 為 X 與 Y 的共變異數，下列敘述何者正確？

$f_{X,Y}(x,y)$		X			$f_Y(y)$
		-1	0	1	
Y	-1	1/9	2/9	0	1/3
	0	0	1/9	2/9	1/3
	1	2/9	0	1/9	1/3
$f_X(x)$		1/3	1/3	1/3	機率總和=1

(A) $Cov(X,Y)\neq 0$且X與Y為相互獨立　(B) $Cov(X,Y)\neq 0$且X與Y非相互獨立
(C) $Cov(X,Y)=0$且X與Y為相互獨立　(D) $Cov(X,Y)=0$且X與Y非相互獨立

[D] 9. 假設$\hat{\theta}$為母體之中某參數θ的一個點估計量(point estimator)，則$\hat{\theta}$之均方誤差($\mathrm{MSE}(\hat{\theta})$, mean-squarederror)，是下列哪一個數學式？
(A) $\mathrm{MSE}(\hat{\theta})=\left[E(\hat{\theta})-\theta\right]^2$　(B) $\mathrm{MSE}(\hat{\theta})=E(\hat{\theta}^2)-\theta^2$
(C) $\mathrm{MSE}(\hat{\theta})=\left[E(\hat{\theta})\right]^2-\theta^2$　(D) $\mathrm{MSE}(\hat{\theta})=E\left[(\hat{\theta}-\theta)^2\right]$

[A] 10. 某汽車公司為瞭解顧客群對於不同車型之喜好程度是否與性別有關聯性，隨機查訪150人作樣本，得到以下的列聯表(contingency table)樣本數據，若設定顯著水準(level of significance)為$\alpha=0.05$，且已知卡方檢定的臨界值(critical value)為5.991，則其檢定的p值(p-value)會落於以下哪一個區間？

性別/車型	中型車	小型車	迷你車	合計
男性	30	25	25	80
女性	15	20	35	70
合計	45	45	60	150

(A)p值<0.05　(B) 0.05<p值<0.10　(C) 0.10<p值<0.15　(D) 0.15<p值

[B] 11. 已知P(A) = 1/5、P(B) = 2/5，且事件A和事件B獨立，則P(A ∪ B)=？
(A) 3/25　(B) 13/25　(C) 3/5　(D)條件不足，無法決定

[C] 12. 若事件A和事件B為相依事件，下列敘述何者正確？
(A) P(A ∪ B)> 1　(B) P(A | B) = P(A)　(C) P(A) + P(B)> 1　(D) P(A ∩ B) < 0

[B] 13. 從一副橋牌中，以不放回方式依序抽取出8張牌，想計算這抽出的8張牌中有3張是10點的機率，應該使用哪種分配？
(A)Poisson分配　(B)超幾何分配　(C)多項式分配　(D)二項分配

[A] 14. 對常態分配而言，下列敘述何者正確？
(A) $P(2<X\leq 8)=P(2\leq X<8)$　(B) $P(X\leq 8)=P(X\leq 8.5)$
(C) $P(X<5)\neq P(X\leq 5)$　(D) $P(X>8)=P(X\geq 9)$

[B] 15. 已知樣本$n=324$和母體標準差$\sigma=9$，則在95 %信心水準下，最大的估計誤差是？
($Z_{0.05}$=1.645；$Z_{0.025}$=1.96)
(A) 0.8225 (B) 0.98 (C) 1.645 (D) 1.96

[D] 16. 某老師想了解學生寫作業的平均時間是多少，已知母體標準差σ=30分鐘，則在95 %信心水準和估計誤差至多10分鐘下，需要多大的樣本？($Z_{0.05}$=1.645；$Z_{0.025}$=1.96)
(A) 13 (B) 18 (C) 25 (D) 35

[B] 17. 下列何種情形將拒絕虛無假設(H_0)？
(A) α=0.01和p值=0.03 (B) α=0.05和p值=0.03 (C) α=0.05和p值=0.15 (D) α=0.10和p值=0.15

[B] 18. 型I誤差(Type I Error)的意思為下列何者？
(A)當虛無假設(H_0)為真時，接受H_0 (B)當虛無假設(H_0)為真時，接受對立假設(H_1)
(C)當對立假設(H_1)為真時，接受H_1 (D)當對立假設為(H_1)真時，接受虛無假設(H_0)

[A] 19. 計算列聯表(contingency table)的卡方值(Chi-Square value)，當這個值很小時，說明這兩個變數間可能的關係是下列何者？
(A)兩變數間獨立 (B)兩變數間相依
(C)兩變數間存在很強的關連 (D)各個觀察值遠大於其各個期望值

[C] 20. 令隨機變數Z服從標準常態分布、隨機變數T服從自由度(degrees of freedom)為9的t分布，而$f_Z(z)$、$f_T(t)$分別為Z、T的機率密度函數(probability density function，pdf)，下列何者正確？
(A) $f_Z(0)=f_T(0)$ (B) $f_Z(0)<f_T(0)$ (C) $f_Z(0)>f_T(0)$ (D) $f_Z(3)>f_T(3)$

[D] 21. 完全隨機集區設計(Randomized Block Design)下，總變異(SSTO)，處理變異(SSTR)，集區變異(SSBL)和誤差變異(SSE)間的關係為下列何者？
(A) SSTO = SSTR + SSBL (B) SSTO = SSTR + SSE
(C) SSTO = SSBL + SSE (D) SSTO = SSTR + SSBL + SSE

[D] 22. 一因子變異數分析(One-Way ANOVA)中，有3個處理，每個處理有5個觀察值，則誤差項的自由度為多少？
(A) 2 (B) 4 (C) 10 (D) 12

[A] 23. 若簡單迴歸方程式中相關係數等於0.8，則其斜率(Slope)值是下列何者？
(A)正值 (B)負值 (C)可正值也可能負值 (D)條件不足，無法決定

[C] 24. 某君每天喝一杯咖啡，且其只喝黑咖啡跟拿鐵。如果他今天喝黑咖啡，則明天會喝黑咖啡的機率是 0.3，喝拿鐵的機率是 0.7；如果他今天喝拿鐵，則明天會喝黑咖啡或喝拿鐵的機率各為 0.5。假設今天是第 1 天，某君喝的是黑咖啡，如果到了第 4 天某君是喝拿鐵，則第 2 天及第 3 天某君喝不同咖啡的機率介於下列何者？
(A) 0.39 與 0.49 之間 (B) 0.5 與 0.579 之間 (C) 0.58 與 0.62 之間 (D) 0.63 與 0.64 之間

[B] 25. 若成功的機率為50 %，試求50次獨立實驗下得到20次成功的機率為下列何者？
(A)無法使用常態分配表計算 (B)可用常態近似法求介於19.5到20.5間之機率
(C)等於0 (D)僅能用二項公式來求近似解

[一律給分] 26. 下列哪項不是Hadoop之特性？
(A)開放原始碼 (Open Source) (B)即時性(Real Time)
(C)以Java為其程式基礎 (D)分散式計算 (Distributed Computing)

[C] 27. 下列哪項不是 Apache Spark 2.0 所支援的程式語言？
(A) Scala (B) R (C) C++ (D) Python

[B] 28. 在類神經網路(Neural Network)分類器中，一個 sigmoid node 的輸出數值是下列何者？
(A)沒有限制，可以是任何實數　(B)介於0與1之間
(C)沒有限制，可以是任何整數　(D)介於-1與1之間

[C] 29. 有關類神經網路(Neural Network)之敘述，下列何者有誤？
(A)使用成千上萬個神經元(Neuron)作為處理器
(B)其學習過程是一個不斷的嘗試錯誤(Trial and Error)
(C)用於影響辨識時，須制訂一系列的規則來識別樣式 (Patterns)
(D)相較全連通網路(Fully Connected Networks)，卷積神經網路 (Convolutional Neural Networks)通常擁有較少須學習的參數

[D] 30. 有關應用卷積神經網路(Convolutional Neural Networks，簡稱CNNs)於影像分析之敘述，下列何者正確？
①較前面的處理層 (Layer) 過濾器通常可包含邊緣偵測器 (Edge Detector)
②和擁有相同層數以及每層內有相同神經元個數的全連通網路 (Fully-connected Networks)比起來，CNNs用有更多參數
③一個CNN可用來訓練非監督式學習模型，但一般的類神經網路則不行
④池化層(Pooling Layers) 可降低影像的空間解析度
(A)①②　(B)②③　(C)③④　(D)①④

[C] 31. 有關分析巨量社群網路資料之敘述，下列何者有誤？
(A)連結預測(Link Prediction) 之技術可用來推薦朋友
(B)使用者的朋友個數之分配通常遵從冪次法則(Power-law distribution)
(C)在計算資源有限時，我們通常對社群網路中的節點進行簡單隨機抽樣，以抽取出具有結構代表性的社群網路子圖(subgraph)
(D)社群網路中任兩位使用者的距離可能隨時間而增減

[A] 32. 有關頻繁樣式探勘(Frequent Pattern Mining) 之敘述，下列何者有誤？
(A)若一項目集合之所有子集合皆為頻繁，則該項目集合必定也為頻繁
(B)所有最大(maximal) 頻繁項目集合都是封閉的(closed)
(C)對於一關聯式規則(Association Rule)，若將一規則左手邊之項目搬到規則右手邊，不會降低該規則之信心度(Confidence)
(D)對於一關聯式規則，將規則左右兩邊之項目集合對調，不會影響支持度(Support)

[A] 33. 巨量資料庫管理最常用的操作包含Selection、Projection、Union與Intersection。對於此4種資料庫操作，若想實作在Hadoop的工作任務中，哪2種操作需要同時使用一個Mapper與一個Reducer？
(A) Selection與Projection　(B) Selection與Union
(C) Projection與Intersection　(D) Union與Intersection

[D] 34. 在資料分析中，經常遇到一分類或預測模型產生高或低的偏差(Bias) 與變異性(Variance)。以下有4個關於模型偏差與變異數之敘述，下列何者正確？
①過適(Overfitting)的模型通常有較高的偏差
②過適(Overfitting)的模型通常有較低的偏差
③乏適 (Underfitting)的模型通常有較高的變異性
④乏適 (Underfitting)的模型通常有較低的變異性
(A) ①③　(B) ①④　(C) ②③　(D) ②④

[C] 35. 下列何種資料格式經常被定義為半結構化資料(Semi-structured Data)？
(A)關聯式資料庫中的表格(Table)資料　(B)影音(Video)檔案
(C) JSON 檔案　(D) Email內文文字

[A] 36. 在巨量資料分析專案中，下列何者比較無助於提升分類模型準確度(Accuracy)？
(A)使用交叉驗證(Cross-Validation)
(B)增加或減少模型參數(Parameters)以提升或降低模型複雜度
(C)進行更進一步的特徵工程(Feature Engineering)
(D)取得與使用有更多變數/特徵(Variables/Features)的訓練資料

[B] 37. 下列的統計學習(Statistical Learning)方法經常被用於分析巨量資料，請問何者應用的領域問題與其它不同？
(A)隨機森林(Random Forests)　(B)主成分分析(Principal Component Analysis)
(C)彈性網絡(Elastic Net)　(D)分類迴歸樹(Classification and Regression Tree)

[C] 38. 有關隨機森林(Random Forests)之敘述，下列何者正確？
(A)採用 Boosting 方法
(B)使用多個淺層決策樹 (Decision Stump) 構建整個森林
(C)是一種重抽樣(Re-sampling)技術的應用
(D)模型訓練的運算較難平行化

[B] 39. 在統計學習(Statistical Learning)方法論中，有關模型的過適(Overfitting)之敘述，下列何者有誤？
(A)此種模型的錯誤(Error)常來自於高的方差(High Variance)
(B)此種模型的參數較少或複雜度過低
(C)此種模型過度學習訓練資料集(Training Dataset)
(D)採用Bagging技術有可能避免過度擬合

[D] 40. 有關近年來巨量資料處理與分析的發展，下列何者有誤？
(A)異質運算(Heterogeneous Computing)的興起
(B)記憶體內運算(In-memory Computing)的普及
(C)資料湖(Data Lake) 的採用
(D)關聯式資料庫系統(Relational Database Systems)的沒落

[A 或C] 41. 有關巨量資料分析的平台環境建置之敘述，下列何者正確？
(A)傳統的關聯式資料庫管理系統無法處理巨量資料
(B) NoSQL資料庫系統已逐漸取代傳統的關聯式資料庫系統
(C) Apache Hadoop可用來建構資料湖(Data Lake)
(D) Apache Spark不適合用來處理串流(Streaming) 資料

[C] 42. 有關 MapReduce程式模型(Programming Model)之敘述，下列何者有誤？
(A)比較不適合用來處理疊代式(Iterative) 的資料處理分析過程
(B)概念來自於函數式程式語言(Functional Programming Languages)
(C)不適合用來處理資料量相對較小，但運算量大的工作
(D) Combiner 的使用可以減少資料在各節點(Node)之間的傳輸

[D] 43. 有關統計學習(Statistical Learning)中的非監督式學習(Unsupervised Learning)方法，下列敘述何者有誤？
(A)可以做為維度縮減 (Dimensionality Reduction) 的工具
(B)此類方法也可用於特徵工程 (Feature Engineering)
(C)是一種無目標變數 (Target Variable) 的模型學習方法
(D) Receiver Operating Characteristic (ROC) 曲線常用來評估此類模型優劣

[A 或D] 44. 下列統計學習(Statistical Learning)的方法或過程，何者不易受到變數尺度(Scale)不同的影響？
(A)決策樹學習(Decision Tree Induction)　(B) K-近鄰算法 (k-nearest Neighbors Algorithm)
(C) K-均值聚類算法 (K-means Clustering）　(D)隨機梯度下降 (Stochastic Gradient Descent)

[B] 45. 如何降低模型過適(Overfitting)問題？
(A)降低資料量
(B)在目標函數中加入參數正則化(Regularization) 項
(C)對資料作正規化(Normalization)前處理
(D)使用測試資料建模

[A] 46. 資料前處理在資料探勘是一個非常重要的步驟，下列哪一個步驟不屬於資料前處理？
(A)模型解釋　(B)處理遺失值　(C)移除離群值　(D)資料數值正規化

[C] 47. 在Hadoop中，下列何者為有效的資料分析流程？
(A) Input → Reducer → Mapper → Combiner → Output
(B) Input → Mapper → Reducer → Combiner → Output
(C) Input → Mapper → Combiner → Reducer → Output
(D) Input → Reducer → Combiner → Mapper → Output

[A] 48. 資料視覺化是資料處理中一項重要工作，通常資料視覺化需要使用到下列哪種方法？
(A)資料維度降維　(B)建立分類模型　(C)建立迴歸模型　(D)調整模型參數

[B 或 D] 49. 下列何者並非Hadoop與Spark共同擁有的特色？
(A)皆為叢集計算環境
(B)皆擁有自己的檔案系統
(C)皆使用開放原始碼應用程式介面 (API) 來連結其他資料分析工具
(D)皆為資料處理平台

[A] 50. 下列哪項不是主成分分析(Principal Component Analysis)之特性？
(A)屬於非線性降維方法　(B)第一主成分跟第二主成分會正交
(C)通常可以提升模型訓練速度　(D)屬於非監督式方法