經濟部所屬事業機構 108 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 4 頁(A3 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。

1. 令 X 為間斷隨機變數，其 E(X) = 5，Var(X) = 2，試求 $E(X^2 + 3X + 3)$為何？
(A) 43 (B) 20 (C) 45 (D) 17

2. 下列何種計算機率方法假設事件(events)發生之機率都相等？
(A)古典(classical) (B)實證(empirical) (C)主觀(subjective) (D)互斥(mutually exclusive)

3. 設隨機變數 Z~ N(0 , 1) 標準常態分布。試問 P(Z < 5) 最接近下列哪個數值？
(A) 0.95 (B) 0.99 (C) 0.50 (D) 0.10

4. 若 t_n 代表自由度為 n 之 t 分布，下列何者最接近標準常態分布？
(A) t_{100} (B) t_{36} (C) t_{25} (D) t_1

5. 設 A 和 B 是兩個獨立之事件，則條件機率 P(A | B) 等於下列何者？
(A) P(B) (B) $P(A \cap B)$ (C) P(A) (D) $P(A \cup B)$

6. 何種統計圖表會呈現四分位距(inter quartile range)？
(A)點圖(dot plot) (B)散布圖(scatter diagram)
(C)箱型圖(box plot) (D)列聯表(contingency table)

7. 設 X_1, X_2, X_3, X_4 為 4 個獨立之隨機變數且都來自於常態分布 N(8 , 16)，已知 $\overline{X} = \sum_1^4 X_i/4$，試問下列何者為 $\overline{X}$ 之標準誤($\sqrt{Var(\overline{X})}$)？
(A) 16 (B) 8 (C) 2 (D) 1

8. 班上學生人數共 20 人，第一次統計考試中，學生唸書時間及成績之判定係數(coefficient of determination)為 80 %。迴歸方程式之變異數 σ^2 估計式的標準誤(standard error of estimate)為 10。以上資訊可編製變異數分析表(ANOVA)表，試問總變異(total sum of square)為何？
(A) 7,200 (B) 9,000 (C) 8,000 (D) 5,400

9. 下列何種抽樣方法可達到群內變異大、群間變異小之結果？
(A)分層抽樣(stratified random sampling) (B)系統抽樣(systematic sampling)
(C)部落抽樣(cluster sampling) (D)簡單隨機抽樣(simple random sampling)

10. 設有一組資料{11,15,13,15,9,8,4,5,5,15}，其最後一個數字由 15 改為 14 ，試問下列何者不變？
(A)平均數 (B)變異數 (C)變異係數 (D)中位數

11. 若隨機變數 X 服從於均勻分布 U(0 , 2)，則X的變異數Var(X)為何？
(A) 18 (B) 1/3 (C) 4 (D) 1/12

 【請翻頁繼續作答】

12. 設 X_1 和 X_2 為獨立同態之 2 個柏努利分布(Bernoulli distribution)，且其值為 1 之機率為 0.4，即 $P(X=1)=0.4=1-P(X=0)$。則樣本平均值 $(X_1+X_2)/2$ 介於 0.25 和 0.75 之間的機率為何？
(A) 0.16 (B) 0.32 (C) 0.36 (D) 0.48

13. 關於敘述統計之陳述，下列何者正確？
(A)一個右偏分布其偏斜度(skewness)大於 0
(B)一個右偏分布通常其中位數會大於平均值
(C)一個對稱的分布，其峰度(kurtosis)必等於 3
(D)一個分布，若知道其前4個動差(Moment)值，則此分布就可決定

14. 對於標準常態分布 Z，設 Z_α 表示 $P(Z > Z_\alpha)=\alpha$ 之百分位點，$0<\alpha<1$。下列何者正確？
(A) $Z_{0.5}=0.5$ (B) $Z_{0.5}=1.96$ (C) $Z_{0.975}=1.96$ (D) $Z_\alpha=-Z_{1-\alpha}$

15. 設 $X_1, X_2, \ldots, X_n$ 表一組獨立且來自於常態分布 $N(\mu, 1)$ 之隨機樣本。下列何者不是 μ 之不偏估計(unbiased estimate)？
(A) $\bar{X}$ (樣本平均值) (B) X_1 (C) $(X_1+X_2)/2$ (D) $X_{(1)}$ (最小順序統計量)

16. 某樣本資料為 26, 21, 24, 9, 17, 23, 18, 22, 20，下列何者正確？
(A)四分位距為 8 (B)全距為 16 (C)變異係數為 25 % (D)此資料有異常值

17. 設樣本空間 $S=\{E_1, E_2, E_3, E_4, E_5\}$，其中 $E_1, E_2, \ldots, E_5$ 為樣本點(sample point)。各樣本點機率為 $P(E_1)=0.3$, $P(E_2)=0.3$, $P(E_3)=0.1$, $P(E_4)=0.15$。令 $A=\{E_1, E_4, E_5\}$ 和 $B=\{E_3, E_4\}$，下列何者正確？
(A) $P(E_5)=0.1$ (B) $P(A \cap B^c)=0.4$ (C) $P(B \mid A)=0.25$ (D) A 和 B 不獨立

18. 設事件 A_1 和 A_2 之驗前機率為 $P(A_1)=0.4$ 和 $P(A_2)=0.6$，已知 A_1 和 A_2 互斥，$P(B \mid A_1)=0.2$ 和 $P(B \mid A_2)=0.1$，下列何者正確？
(A) $P(A_1 \mid B)=3/7$ (B) $P(A_1 \cap B)=0.06$ (C) $P(B)=0.14$ (D) $P(A_2 \cup B)=0.46$

19. 關於顯著水準之敘述，下列何者正確？
I：是 1 減信賴水準； II：是 P 值； III：是最大可容許型一誤差發生之機率
(A) I和II (B) I和III (C) II和III (D) I、II和III

20. 盒子中有 8 顆球，其中 4 顆是白球，其餘是黑球。以取後不放回方式隨機取 2 顆球，令 X 為取到白球之個數。下列何者正確？
(A) $E(X)=0.5$ (B) $Var(X)=3/7$ (C) $P(X=1)=3/7$ (D) $P(X \le 1)=5/7$

21. 隨機選取 n 個樣本欲計算母體比例之95 % 信賴區間，若希望誤差界限在 0.05 以內，則需要幾個樣本數？
(A) 196 (B) 271 (C) 385 (D) 1,068

22. 兩個隨機變數 X 和 Y 之線性關係為 $Y=0.5X+\epsilon$ ，其中隨機誤差 ϵ 服從常態分布 N(0 , 1) 且與 X 獨立，X 之期望值與變異數各為 $E(X)=0$, $Var(X)=1$。則 X 與 Y 之皮爾森相關係數(Pearson correlation coefficient)為下列何者？
(A) $1/\sqrt{2}$ (B) $1/\sqrt{5}$ (C) 3/4 (D) 1/4

23. 考慮下列線性迴歸模型 $Y=\beta X+\epsilon$ 。若我們有 n 對 (X_i, Y_i) 觀察值，且 β 之最小平方估計為 $\hat{\beta}$ ，下列何者正確？
(A) $\hat{\beta}=\frac{\sum_1^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sum_1^n (X_i-\bar{X})^2}$ (B) $\hat{\beta}=\frac{\sum_1^n X_i Y_i}{\sum_1^n X_i^2}$
(C) ϵ_i 必須服從常態分布 (D) X_i 必須服從常態分布

24. 對於 3 個解釋變數之迴歸模型 $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$, $i = 1,\ldots,n$，$\epsilon \sim N(0, \sigma^2)$，若 $n = 30$。下列何者正確？
(A)想要檢定 β_1 和 β_2 是否同時為 0 (H_0：$\beta_1 = \beta_2 = 0$)，可使用 partial F 檢定
(B)想要檢定 β_1 和 β_2 是否同時為 0 ，則對立假設應為 H_1： $\beta_1 \neq 0$且 $\beta_2 \neq 0$
(C)想要檢定 β_3 是否為 0 (H_0：$\beta_3 = 0$)，H_0 為真時檢定統計量服從於 t_1
(D)檢定H_0：$\beta_1 = \beta_2 = \beta_3 = 0$，須使用自由度為 3 之 t 分布

25. 資料進行變異數分析(analysis of variance)時，不需下列何種假設？
(A)資料呈常態分配　(B)各組母體變異數相等
(C)各組資料間獨立　(D)各組母體平均數相等

26. 關於雲端運算之敘述，下列何者有誤?
(A)公有雲不一定免費，但可降低硬體投資和機房管理等成本
(B)使用雲端服務之付費方式採用「Pay-As-You-Go」
(C) AWS EC2 屬於雲端服務中之 SaaS
(D) OAuth 2.0 協議的授權碼授予模式，需要服務端之認證伺服器許可

27. 何者非屬監督式學習之演算法？
(A)決策樹　(B)隨機森林　(C)支持向量機　(D)關聯規則

28. 針對巨量資料之特性，下列何者有誤？
(A)巨量資料之巨量性質(volume)意指存放數據量超過 PB
(B)巨量資料之即時性質(velocity)意指數據擷取時間不到 1 秒
(C)巨量資料等同於巨大價值(value)
(D)根據巨量資料分析需求之改變趨勢，視覺化(visualization)在分析中日趨重要

29. 針對 Apache Spark，下列何者有誤？
(A) in-memory 之計算框架　(B)不允許用戶將資料載入至叢集記憶體內儲存
(C)多次記憶體重覆運算　(D)非常適合用於機器學習演算法

30. 影響資料分析技術之重要資料集特質，下列何者有誤？
(A)維度(dimensionality)　(B)稀疏性(sparsity)
(C)連續性(continuity)　(D)分辨度(resolution)

31. 巨量資料之定義為何？
(A)巨量資料中有 70 % 都為結構化資料
(B)巨量資料除資料量龐大外，其資料特性具變化速度快及多樣性
(C)儲存的資料內容不包含影片或電子郵件
(D)巨量資料強調資料數量能為企業帶來商業機會

32. 若欲將大量資料進行分群，下列何種方法不適合？
(A)決策樹法　(B) K-means法　(C)階層式方法　(D) SOM方法

33. 關於巨量資料之特性，下列何者正確？
(A)堅持原始資料都要做到標準化與精確　(B)陷入資料獨裁之問題
(C)利用「隨機取樣」處理所有的資料　(D)看重資料之間的因果關係

34. 下列何者非屬巨量資料分析工具？
(A) Spark　(B) Python　(C) Spigot　(D) Julia

35. 關於由小到大的電腦容量(單位)，下列何者正確？
(A) YB＜GB＜TB＜PB　(B) GB＜TB＜PB＜EB　(C) TB＜PB＜YB＜EB　(D) GB＜TB＜ZB＜EB

36. 巨量資料分析所蒐集之資料來源，下列何者與其他來源差異最大？
(A)養殖水產保險　(B)網路溫度計　(C)豆腐指數　(D) 753感冒指數

【請翻頁繼續作答】

37. 關於工業 4.0 製造模式轉變，下列何者正確？
(A)將原本 B2C 之製造模式轉變為 C2B (B)將原本 B2C 之製造模式轉變為 B2B
(C)將原本 B2B 之製造模式轉變為 C2B (D)將原本 B2B 之製造模式轉變為 C2C

38. 針對巨量資料分析進行資料探勘(data mining)，下列何者有誤？
(A)找尋趨勢 (B)找尋特徵 (C)找尋相關性 (D)無法發掘出各種假設

39. 下列何者非屬邏輯迴歸(logistic regression)之特性？
(A)離散選擇法模型之一 (B)屬於多重變量分析範疇
(C)需要常態分配的假設 (D)羅吉斯迴歸用到的對數函數是Sigmoid函數

40. 強化學習(reinforcement learning)系統中不包括下列何者？
(A)智能體(agent) (B)獎賞(reward) (C)回應(response) (D)環境(environment)

41. 關於遷移學習(transfer learning)特性，下列何者有誤？
(A)遷移學習之重點是不必儲存已解決一個問題之模型
(B)遷移學習被應用於認知科學
(C)可使用遷移學習重新利用既有神經網絡
(D)遷移網絡之應用包括語句分類，篩選垃圾郵件與簡訊以及發現癌症種類

42. 配置 Hadoop 時，JAVA_HOME 包含在下列何者配置檔案中？
(A) hadoop-default.xml (B) hadoop-site.xml
(C) configuration.xsl (D) hadoop-env.sh

43. Java 語言之 Buffered Reader 類別是將資料讀入下列何者當緩衝區？
(A)陣列 (B)資料庫 (C)檔案 (D)變數

44. 巨量資料分析之資料存在著資料量大、非結構化、高度異質性等特性，下列何種資料庫工具最不適宜運用在此類型工作？
(A) Mongodb (B) Redis (C) Sybase (D) Hbase

45. 關於關聯式資料庫資料表(table)之敘述，下列何者正確？
(A)是一維資料組成之集合 (B)資料表由橫列和直行所組成
(C)一般不會設定主索引鍵 (D)資料表每一列表示屬性

46. 關於遞歸神經網路(RNN)之基本概念，下列何者有誤？
(A)反向傳播的權重更新不會造成梯度爆炸
(B)長短期記憶模組共 4 層，比 RNN 多 3 個 S 型函數層
(C)長短期記憶模組，能改善 RNN 在長期記憶之不足
(D)可用來處理時間序列資料

47. 關於機器學習之敘述，下列何者正確？
(A)主成分分析法(PCA)是用於資料之降維
(B)用大量人力對訓練資料集來標籤特徵，是強化學習(RL)之特色
(C) Q-Learning 之 γ 數值趨向於 1 ，表示 agent 只在乎目前可獲得之獎勵
(D)監督式學習之演算法有邏輯迴歸和 K-means 等

48. 有關卷積神經網路之基本概念，下列何者有誤？
(A)運作流程：輸入的圖片→特徵擷取→分類辨識
(B)池化層會使用到 ReLU 之激勵函數
(C)得到之特徵圖比原圖要小，被稱為valid padding
(D)可應用於人臉識別、語音識別等

49. 關於巨量資料中之資料庫，下列何者有誤？
(A) HBase 技術提供非關聯式資料庫(NoSQL)之儲存環境
(B) HBase 技術支援隨機存取功能
(C)無法直接透過 SQL 來查詢 Hadoop 儲存之資料
(D) Apache Hive 就是把 SQL 編譯成 Map Reduce，從而讀取和操作 Hadoop 上之資料

50. 下列何者非屬資料操作語言(data manipulation language)？
(A) INSERT (B) UPDATE (C) DELETE (D) CREATE