

經濟部所屬事業機構 113 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，答錯不倒扣；畫記多於 1 個選項或未作答者，該題不予計分。
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. 袋子中有 2 個紅球、3 個黑球、5 個白球，每次從袋中抽出 1 球後放回，共抽 3 次，請問 3 球中有 2 個是紅球之機率為何？
(A) 0.024 (B) 0.032 (C) 0.064 (D) 0.096
- [B] 2. 大學生打工之比例為 0.3，若隨機抽取 3 位大學生，其中至少 2 位有打工之機率為何？
(A) 0.116 (B) 0.216 (C) 0.316 (D) 0.416
- [D] 3. 欲比較兩公司員工薪資之離散程度，可採用下列何者統計量？
(A) 全距 (B) 平均數 (C) 變異數 (D) 變異係數
- [B] 4. 假設 A 、 B 為樣本空間 S 之兩事件，下列何者與其他敘述之意義不同？
(A) A 、 B 為互斥事件 (B) $P(A \cap B) = P(A)P(B)$
(C) $P(A \cap B) = 0$ (D) $P(A \cup B) = P(A) + P(B)$
- [B] 5. 已知 $E(X + 4) = 10$ 且 $E[(X + 4)^2] = 116$ ，試求 $Var(X)$ 為何？
(A) 4 (B) 16 (C) 100 (D) 116
- [D] 6. 均方根誤差(RMSE)是藉以衡量下列何者？
(A) 樣本量大小 (B) 指數平滑度 (C) 移動平均週期 (D) 預測的準確性
- [C] 7. 下列何種圖表最適合用來顯示資料隨著時間變化之趨勢？
(A) 散點圖 (B) 長條圖 (C) 折線圖 (D) 圓餅圖
- [A] 8. 颱風正接近台灣，某地方首長須決定次日是否停班停課，故設立 2 個假設為 H_0 ：颱風會登陸、 H_1 ：颱風不會登陸。以「該放假而不放假」之情形，係犯何種型態錯誤？
(A) 型 I 錯誤 (B) 型 II 錯誤 (C) 型 III 錯誤 (D) 型 IV 錯誤
- [B] 9. 若右尾檢定的顯著水準(α 值)愈小，下列何者正確？
(A) p 值(p-value)愈大 (B) 臨界值(Critical Value)愈大
(C) 樣本平均數愈大 (D) 母體平均數愈大
- [A] 10. 下列何者屬於古典迴歸分析之基本假設？ $\textcircled{甲}$ ：誤差項服從常態、 $\textcircled{乙}$ ：誤差項彼此不相關、 $\textcircled{丙}$ ：反應變數 Y 服從常態分配、 $\textcircled{丁}$ ：解釋變數間不相關。
(A) $\textcircled{甲}$ 、 $\textcircled{乙}$ (B) $\textcircled{甲}$ 、 $\textcircled{丙}$ (C) $\textcircled{甲}$ 、 $\textcircled{丁}$ (D) $\textcircled{乙}$ 、 $\textcircled{丁}$
- [C] 11. 當 A 、 B 的共變異數(Covariance) $Cov[A, B] = 5$ 時，試求 $Cov[2A, B + 1]$ 為何？
(A) 5 (B) 6 (C) 10 (D) 11
- [A] 12. 若有 4 家供應商提供原料，欲檢定此 4 家原料平均數是否相等，可用下列何種檢定？
(A) F 檢定 (B) t 檢定 (C) Z 檢定 (D) 卡方檢定

- [D] 13. 下列何者受離群值(outliers)的影響最小?
 (A)全距 (B)標準差 (C)變異係數 (D)四分位數
- [C] 14. 有關迴歸模式(Regression Models)的最小平方估計法(Least Square Estimation), 下列敘述何者正確?
 (A)所求得之迴歸係數, 使得依變數之估計值與0的誤差平方和最小
 (B)所求得之迴歸係數, 使得依變數(Y)與其平均數之誤差平方和最小
 (C)所求得之迴歸係數, 使得依變數之估計值與依變數之觀察值的誤差平方和最小
 (D)所求得之迴歸係數, 使得依變數之估計值與依變數之平均數的誤差平方和最小
- [B] 15. 某資料分配的偏態係數(Coefficient of Skewness) = -3, 請問該資料分配的平均數、中位數與眾數的順序關係為何?
 (A)平均數=中位數=眾數 (B)平均數<中位數<眾數
 (C)眾數<平均數<中位數 (D)平均數<眾數<中位數
- [D] 16. 有關信賴區間之敘述, 下列何者有誤?
 (A)在變數固定下增加樣本數, 區間長度變短
 (B)在樣本數固定下增加信賴係數, 區間長度變長
 (C)信賴係數是指欲推估參數會落在信賴區間的機率
 (D)在樣本數固定下增加信賴係數, 區間估計的精確度提升
- [A] 17. 根據過去資料顯示, 某公司員工離職人數為常態分配, 其A、B兩區分公司員工離職人數標準差相同, σ 均為2人/年, 若A、B兩區各隨機抽取8家分公司, 其平均員工離職人數分別為12、10人, 欲以 $\alpha = 0.05$ 檢定兩區分公司平均員工離職人數是否相同, 下列敘述何者有誤?
 (A) $Z = 1$ (B)臨界值為 ± 1.96
 (C) H_0 : 兩區分公司平均員工離職人數相同 (D)結論為拒絕 H_0
- [D] 18. 假設 X_1, \dots, X_n 係一組來自母體平均數為 μ 、母體變異數為 σ^2 的隨機樣本, 下列敘述何者有誤?
 (A) $\bar{X} = (\sum_{i=1}^n X_i)/n$ 是 μ 的一致估計量
 (B) $\bar{X} = (\sum_{i=1}^n X_i)/n$ 是 μ 的不偏估計量
 (C) $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$ 是 σ^2 的不偏估計量
 (D) $S = \sqrt{S^2}$ 是 σ 的不偏估計量
- [B] 19. 有關 t 分配與常態分配的峰態係數, 下列敘述何者正確?
 (A) t 分配的峰態係數比較小 (B) t 分配的峰態係數比較大
 (C) 兩分配的峰態係數一樣大 (D) 兩分配的峰態係數無法比較
- [C] 20. 統計學家證實, 要提高抽樣的準確度, 最好的方式為下列何者?
 (A)使用速度最快的電腦硬體 (B)使用最精準的分析軟體
 (C)做到隨機抽樣 (D)增加樣本數
- [D] 21. 假設某隨機變數 X 服從平均數為10的卜瓦松分配(Poisson Distribution), 若自其中隨機抽取100個觀察值構成樣本平均數, 請問此統計量服從的抽樣分配為何?
 (A)平均數為10的卜瓦松分配 (B)平均數與變異數皆為10的常態分配
 (C)平均數為10、變異數為1的常態分配 (D)平均數為10、變異數為0.1的常態分配
- [C] 22. 若樣本資料(4, 6, 6, 8, 10, 14)等6個數值來自於相同的對稱分配, 根據此樣本, 母體中位數最佳不偏估計元的數值為何?
 (A) 6 (B) 7 (C) 8 (D)最佳不偏估計元不存在
- [A] 23. 檢定母體平均數時, 若母體為常態分布且小樣本, 下列敘述何者正確?
 (A)母體變異數已知時用 Z 檢定 (B)母體偏離常態時用 t 檢定
 (C)小樣本時用 t 檢定 (D)母體為常態分布用 F 檢定

- [A] 24. 若於一複迴歸分析中使用100個觀察值得到下列迴歸估計式，其總平方和SST = 1075，誤差平方和SSE = 275，試求迴歸均方MSR為何？
 $\hat{Y} = -1.5 + 7.5X_1 - 6.4X_2 + 3.8X_3 + 0.9X_4$
 (A) 200 (B) 275 (C) 345 (D) 800
- [C] 25. 某常態分布之母體，其變異數為 σ^2 ，欲檢定虛無假設 $H_0: \sigma^2 = \sigma_0^2$ 的真偽(σ_0^2 為一定數)，使用的統計檢定量與下列何種機率分布有直接關聯性？
 (A) F 分布 (B) t 分布 (C) 卡方分布 (D) 指數分布
- [C] 26. 有關統計與機器學習的差異，下列敘述何者正確？
 (A) 統計是機器學習的一個子領域，兩者沒有明顯差異
 (B) 統計主要用於數據分析，機器學習主要用於模式識別和預測
 (C) 統計強調推論和參數估計，機器學習更側重模式識別和模型訓練
 (D) 統計使用傳統方法進行數據分析，機器學習使用深度學習方法進行模式識別
- [B] 27. 下列何者非屬資料前處理(Data Preprocessing)的一環？
 (A) 特徵轉換 (B) 資料分群 (C) 遺失值填補 (D) 異常值檢測與排除
- [C] 28. 何謂結構化資料(Structured Data)？
 (A) 由感測器產生的資料 (B) 社群媒體上的使用者互動資料
 (C) 具固定格式，如表格資料 (D) 無固定格式，如文字、影像
- [A] 29. 下列何者非屬低品質的資料？
 (A) 重複值 (B) 離群值 (C) 錯誤 (D) 雜訊
- [A、D] 30. 下列何者非屬分群的應用範圍？
 (A) 垃圾郵件過濾 (B) 異常檢測 (C) 顧客細分 (D) 降維技術
- [C] 31. 下列何者非屬離群值的處理方式？
 (A) 直接刪除 (B) 群集分析
 (C) 使用屬性絕對值 (D) 用其他數值替換，將資料範圍正規化
- [B] 32. Spark的DAG(Directed Acyclic Graph)在資料處理之作用，下列何者正確？
 (A) 定義資料內部的儲存結構 (B) 定義資料處理的邏輯流程
 (C) 定義資料判斷的決策規則 (D) 定義資料外部的呈現形式
- [C] 33. 下列何者非屬群集分析(Clustering Analysis)的4個主要階段？
 (A) 資料準備與特徵選取 (B) 相似度計算
 (C) 非線性分類 (D) 分群演算法
- [D] 34. 有關巨量資料的長尾效應(Long Tail Effect)，下列敘述何者正確？
 (A) 資料量越大，價值就越大 (B) 少數資料佔據大部分的價值
 (C) 資料的價值會隨著時間推移而衰減 (D) 大量資料中隱藏著小眾但有價值的資訊
- [A] 35. 在資料分群中，用以確定最佳聚類數量的技術為何？
 (A) Elbow Method (B) Gradient Descent
 (C) Principal Component Analysis (D) Random Forest
- [D] 36. 何謂遷移學習(Transfer Learning)？
 (A) 在不同的環境中部署模型的過程
 (B) 在訓練模型中將資料轉移到不同存儲位置的過程
 (C) 將模型從一個硬體平台轉移到另一個硬體平台的過程
 (D) 在不同的機器學習任務之間轉移模型權重和知識的過程

- [B] 37. Apache Spark 在巨量資料環境中的主要用途為何？
 (A)資料清理 (B)即時與批次處理 (C)機器學習模型訓練 (D)資料視覺化
- [D] 38. 何謂交叉驗證(Cross-Validation)？
 (A)將資料集隨機分為訓練集和測試集
 (B)將資料集按照特徵分為訓練集和測試集
 (C)將資料集按照時間順序分為訓練集和測試集
 (D)將資料集多次隨機分為訓練集和測試集，取平均結果
- [A] 39. 有關自然語言處理(NLP)之步驟，下列何者正確？
 (A)斷詞→詞性標記→相依剖析 (B)斷詞→相依剖析→詞性標記
 (C)相依剖析→詞性標記→斷詞 (D)相依剖析→斷詞→詞性標記
- [B] 40. 在機器學習演算法中，下列何者最能避免過度配適(Overfitting)？
 (A)決策樹 (B)隨機森林 (C)羅吉斯迴歸 (D) K-means演算法
- [B] 41. 在Hadoop中，HDFS的資料冗餘機制為何？
 (A)資料壓縮 (B)資料複製 (C)資料分片 (D)資料加密
- [A、B] 42. 假設使用半導體晶圓資料，良率資料之反應值以二元類別表示，請問上述情境最適合使用何種分析技巧？
 (A)羅吉斯迴歸 (B)決策樹 (C)階層式聚類分析 (D)主成分分析
- [D] 43. 在資料進行機器學習的過程中，正規化(Regularization)係指下列何者？
 (A)一種降低資料維度的技術
 (B)一種平衡類別分佈的方法
 (C)一種改善模型可解釋性的方式
 (D)一種透過在損失函數中添加懲罰防止過度配適的技術
- [B] 44. 下列何種機器學習技術適合偵測大型資料集中的異常值(Anomalies)？
 (A) Apriori演算法 (B)隔離森林(Isolation Forest)
 (C)線性迴歸(Linear Regression) (D)神經網路(Neural Networks)
- [D] 45. 下列何種神經網路架構最適合分析序列資料，如時間序列或自然語言？
 (A)生成對抗網路(GAN) (B)卷積神經網路(CNN)
 (C)前饋神經網路(FNN) (D)循環神經網路(RNN)
- [C] 46. 下列何者為使用機器學習時須注意之事項？
 (A)確保模型訓練時間足夠長 (B)模型的複雜度越高越好
 (C)需要有過去資料且資料充足 (D)模型的性能與模型的可解釋性無關
- [D] 47. 有關支持向量機(Support Vector Machine)中的「核技巧」(Kernel Trick)，其作用為下列何者？
 (A)處理高維資料 (B)減少支持向量的數量
 (C)提高支持向量機的計算速度 (D)將非線性可分的資料轉換為線性可分的資料
- [C] 48. 在 ETL 過程中，Transform 主要功能是將抽取之資料進行轉換，下列何者非屬其主要任務？
 (A)資料格式轉換 (B)資料清洗 (C)資料備份 (D)資料聚合
- [A] 49. 生成對抗網路(GAN)較常應用在下列何種技術上？
 (A)影像生成 (B)語音識別 (C)文字分類 (D)物體檢測
- [B] 50. 下列何種技術可實現即時分析？
 (A) OLAP (B)串流處理 (C)批次處理 (D)資料倉儲