經濟部所屬事業機構112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。

1. 假設X的標準差為5，試求$Var(2X-5)$為何？
(A) 100 (B) 50 (C) 10 (D) 20

2. 當資料分析者蒐集20筆資料來配適簡單線性迴歸模型時，計算出$SS_{Reg}=84$，MSE$=2$，試求判定係數(R^2)為何？
(A) 0.700 (B) 0.977 (C) 0.677 (D) 0.800

3. 已知$P(Y)=0.6$，$P(X|Y)=0.8$，$P(X|Y^c)=0.4$，試求$P(Y|X^c)$為何？
(A) 0.750 (B) 0.250 (C) 0.667 (D) 0.333

4. 投擲一公平的硬幣直到出現正面就結束此試驗，請問在偶數次投擲到正面之機率為何？
(A) 0.500 (B) 0.333 (C) 0.250 (D) 0.667

5. 請問$P(A\cap B)$等於下列何者？
(A) $P(A)P(B)$ (B) $P(B)/P(A)$ (C) $P(\mathrm{A})/P(B)$ (D) $P(A|B)P(B)$

6. 卡方分配是伽瑪(Gamma)分配的一個特例，當伽瑪分配中的參數以形狀母數(α)和比例母數(β)表示時，下列何組伽瑪分配之參數等同卡方分配自由度為10？
(A) $\alpha=5$，$\beta=1$ (B) $\alpha=20$，$\beta=1$ (C) $\alpha=5$，$\beta=2$ (D) $\alpha=20$，$\beta=2$

7. 假設資料$x=0,1,2,3$，$y=1,2,6,8$，研究者想要比較2個預測模型$\dot{y}=2x$和$\ddot{y}=x^2$，若依據最小平方法來評估2個預測模型優劣，下列敘述何者正確？
(A) $\dot{y}$模型優於$\ddot{y}$ (B) $\ddot{y}$模型優於$\dot{y}$
(C) $\dot{y}$模型之殘差平方和為7 (D) $\ddot{y}$模型之殘差平方和為6

8. 假設母體來自參數為λ之卜瓦松(Poisson)分配，抽取n個隨機樣本估計得λ之最大概似估計量為$\bar{X}$，請問母體標準差之最大概似估計量為何？
(A) $\sqrt{\bar{X}}$ (B) $\sqrt{n\bar{X}}$ (C) $\bar{X}$ (D) $\bar{X}^2$

9. 假設$X|N\sim\mathrm{Gamma}(N,\beta)$，且$N\sim\mathrm{Poi}(\lambda)$，請問$E(X)$為何？
(A) $N\beta$ (B) $N\beta\lambda$ (C) $\beta\lambda$ (D) $N\lambda$

10. 下列何種情形為型Ⅱ誤差？
(A)拒絕一個正確的虛無假設 (B)無法拒絕一個正確的虛無假設
(C)拒絕一個錯誤的虛無假設 (D)無法拒絕一個錯誤的虛無假設

11. 在進行統計檢定時，若檢定值的p-value小於顯著水準α，則表示下列何種結果？
(A)無法拒絕虛無假設 (B)無法拒絕對立假設
(C)拒絕虛無假設 (D)拒絕對立假設

12. 某樣本(n=100)的標準誤為30。若要將標準誤降為15，下列樣本處理方式，何者正確？
(A)樣本數增至400 (B)樣本數增至200 (C)樣本數減至50 (D)樣本數減至25

13. 統計檢定的檢定力是正確判定下列何種情形之能力？
(A)拒絕一個正確的虛無假設 (B)無法拒絕一個正確的虛無假設
(C)拒絕一個錯誤的虛無假設 (D)無法拒絕一個錯誤的虛無假設

14. 某一研究宣稱每10位醫生中，有9位會推薦病人服用止痛藥以緩解頭痛。為驗證此敘述(對立假說為推薦病人服用止痛藥的醫生比率<0.9)，實際取樣100位醫生，其中有83位推薦病人服用止痛藥。請問其檢定統計量為何？
(A) -4.12 (B) -2.33 (C) -1.86 (D) -0.07

15. 在複迴歸模型中，下列關於Adj-R^2之敘述，何者正確？
(A) Adj-R^2的值恆大於R^2的值 (B) Adj-R^2的值恆大於1
(C) Adj-R^2的值恆為正值 (D) Adj-R^2的值可能是負值

16. 在複迴歸模型中，下列哪種情形需增加模型中解釋變數的交叉交乘項？(如：X_1X_2)
(A)模型中解釋變數過多 (B)應變數與解釋變數間有曲度關係
(C)解釋變數X_1與X_2的參數估計值皆不顯著 (D) X_1和應變數間的關係受到X_2的影響

17. 假設變數X與Y之相關係數(Coefficient of Correlation)為0.375，且檢定$H_0: \rho = 0, H_1: \rho > 0$之p-value為0.256。試求檢定$H_0: \rho = 0, H_1: \rho \neq 0$之p-value為何?
(A) $1 - (0.256/2)$ (B) $1 - 0.256$ (C) $0.256/2$ (D) 0.256×2

18. 圖書館館長希望知道每日借出的圖書數量，館員提供的數據為在95%信賴區間下，每日借出的圖書數量在740本至920本之間。假設標準差已知為150，請問館員分析數據的樣本數為何？($Z_{0.025}$=1.96，$Z_{0.05}$=1.645)
(A) 125 (B) 13 (C) 11 (D) 4

19. 假設某大學男學生參加各種運動的比例如下：

羽球	桌球	網球	羽球及桌球	羽球及網球	桌球及網球	三者皆有
30%	20%	20%	5%	10%	5%	2%

若隨意抽取1名男學生，試求其至少參加1種運動的機率為何？
(A)0.7 (B) 0.5 (C) 0.48 (D) 0.52

20. 在一個盒中放4個球，編號分別為1、2、3、4，自此盒中隨機抽出2球(採抽出不放回法)，令X_1、X_2為其編號數，且$(X_1+X_2)/2 = \bar{X}$，請問$\bar{X}$之期望值與變異數分別為何？
(A) 2.5, 0.42 (B) 2.5, 6.45 (C) 2.5, 2.5 (D) 2.92, 2.92

21. 若X服從常態分配，期望值與標準差皆為60，令$Y = 600 + 5X + 0.3X^2$，請問E(Y)為何？
(A) 3060 (B) 2340 (C) 2440 (D) 2600

22. 若某飛機每個月發生意外事件之次數為0.2次，試求一年中發生2次意外事件之機率為何？
(A) 0.27 (B) 0.24 (C) 0.25 (D) 0.26

23. 根據經驗法則，如果數據呈現「鐘形」常態分配，約有多少百分比的觀察值落在算術平均數上下1個標準差的範圍內？
(A) 68.26 (B) 75.00 (C) 88.89 (D) 93.75

24. 在一完全對稱的分配中，下列敘述何者有誤？
(A) Q1至Q2的距離等於Q2至Q3的距離 (B)最小值至Q1的距離等於Q3至最大值的距離
(C)最小值至Q2的距離等於Q2至最大值的距離 (D) Q1至Q3的距離為最小值至最大值距離的一半

25. 下列哪項非屬指數平滑法(Exponential Smoothing Method)之特性？
(A)可以平滑資料中的季節特徵(Seasonal Components)
(B)可以平滑資料中的循環特徵(Cyclical Components)
(C)可以一次產出超過1期以上的預測值
(D)可以產生1期預測值

26. 在機器學習中，深度強化學習(Deep Reinforcement Learning)主要面臨的挑戰為下列何者？
(A)過擬合(Overfitting)　(B)計算效率
(C)缺乏大規模資料　(D)探索與利用的平衡

27. 在機器學習中，遞迴類神經網路(RNN)主要用於處理下列何種類型之資料？
(A)非結構化資料　(B)圖像資料　(C)時序資料　(D)結構化資料

28. 在巨量資料分析中，下列何者為主成分分析(PCA)之用途？
(A)資料壓縮　(B)資料可視化　(C)資料備份　(D)特徵選擇與降維

29. 在巨量資料分析中，推薦系統的主要目標為何？
(A)評估資料相關性　(B)預測未來趨勢　(C)辨識異常行為　(D)提供個性化建議

30. 維度災難(Curse of Dimensionality)指下列何者？
(A)資料集的數據維度過多　(B)資料集的數據維度過少，導致計算困難
(C)資料集的數據分佈不平均　(D)資料集中的數據缺乏多樣性

31. 考慮巨量資料的特性，下列何者最能描述「Velocity」之特性？
(A)資料的快速產生、傳輸、即時處理　(B)資料多樣性
(C)資料的大規模與容量　(D)資料的真實性與準確性

32. 當面對結構化和非結構化的巨量資料時，下列哪項存儲技術最適合存放非結構化資料？
(A)資料倉儲　(B)關聯式資料庫　(C) NoSQL資料庫　(D) OLAP立方體

33. 在資料工程項目中為確保資料完整性，下列哪種技術是必要的？
(A)資料加密(Encryption)　(B)資料驗證(Validation)
(C)資料壓縮(Compression)　(D)資料隨機化(Randomization)

34. 針對巨量資料處理，台灣的一家公司決定使用MapReduce進行數據處理，下列何者最能說明MapReduce之操作方式？
(A)首先彙整資料，然後再分配給不同的節點進行處理
(B)在單一伺服器上進行所有計算，然後將結果分配給客戶端
(C)先將資料映射到各節點，再從各節點彙整結果
(D)資料首先被分配給節點，每個節點獨立運算，不需要合併結果

35. 在數據科學領域中，當數據分佈不均衡時，下列哪個評估指標最適合用於分類器性能評估？
(A)精確度(Accuracy)　(B) F1分數(F1-Score)
(C)正確率(Precision)　(D)召回率(Recall)

36. 下列有關過擬合(Overfitting)之處理策略，何者有誤？
(A)利用數據增強如圖片旋轉，增加數據多樣性
(B)不斷增加訓練次數，直到訓練誤差趨近於零
(C)使用Dropout技巧隨機丟棄神經元
(D)使用早停法(Early Stopping)

37. 下列何者屬於資料探勘研究的範疇？
(A)專家系統的推論　(B) SQL查詢處理
(C)敘述統計處理　(D)分群處理

38. 使用卷積神經網路(CNN)進行圖像辨識時，下列哪一層負責提取圖像的基本特徵？
(A)全連接層(Fully Connected Layer) (B)池化層(Pooling Layer)
(C)歸一化層(Normalization Layer) (D)卷積層(Convolution Layer)

39. 當一家醫院希望分析病人的醫療紀錄以預測其再入院的風險時，最適合使用下列哪一種機器學習任務？
(A)迴歸(Regression) (B)分類(Classification)
(C)聚類(Clustering) (D)降維(Dimensionality Reduction)

40. 在機器學習中，哪一種演算法是基於樹狀結構，且適用於分類與預測問題？
(A)支持向量機(Support Vector Machine) (B)隨機森林(Random Forest)
(C) K均值聚類法(K-Means Clustering) (D)主成分分析(Principal Component Analysis)

41. 在機器學習中，特徵工程(Feature Engineering)指下列何者？
(A)創造新的數據特徵以提高模型的特性 (B)調整模型的超參數以改善性能
(C)清理數據以去除離群值 (D)壓縮數據以節省儲存空間

42. 在類神經網路中，反向傳播(Backpropagation)是用來執行下列哪個關鍵任務？
(A)將資料從輸入層傳遞至輸出層 (B)調整神經元的啟動函數(Activation Function)
(C)計算損失函數的梯度，以更新權重 (D)創造新的神經元層

43. 假設銷售點有3類項目集{a，b，c}；資料庫共有3筆交易：{b}，{b、c}，{a、b、c}。若關聯規則要求最小支持度為2/3，最小信心度為0.7。下列哪條關聯規則符合該要求？
(A) b→a (B) b→c (C) c→b (D) ab→c

44. 下列哪項目標不易透過資料探勘客戶交易資料達成？
(A)顧客的薪資預測 (B)顧客會一起購買的產品項
(C)猜測那些產品會先後購買 (D)猜測顧客信用卡被盜刷與否

45. 下列何者非屬OLAP操作？
(A)下鑽(Drill down) (B)資料修改(Update) (C)切片(Slicing) (D)切丁(Dicing)

46. 下列何者非屬資料倉儲用的資料模型？
(A)物件導向(OOP) (B)星型(Star) (C)雪花(Snowflake) (D)星雲(Constellation)

47. 下列哪項方法較合適使用在進行數據分析時，可有效評估機器學習模型的性能以確保其在實際應用中具有良好的泛化能力？
(A)使用全部數據進行訓練 (B)增加特徵數量
(C)使用交叉驗證 (D)增加神經網路的深度

48. 在分群演算法中，下列哪種度量不會用來衡量節點不純度(Node Impurity)？
(A)變異數 (B) TF/IDF
(C)基尼係數(Gini index) (D)分類誤差

49. 下列何者非屬基於內容推薦(Content-Based Recommendation)之優點？
(A)沒有稀疏(Sparsity)問題 (B)可建立使用者輪廓檔案
(C)能夠提供推薦的原因 (D)對於新品與冷門商品也會推薦

50. 有關分散式系統之優勢，下列何者有誤？
(A)彈性擴展(Scalability) (B)高容錯性(Fault-tolerance)
(C)資料集中性(Centralization) (D)資源最大化(Maximized Resource Utilization)