經濟部所屬事業機構111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5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，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. 在統計學概念及方法中，下列哪些會受極端值影響？
①中位數②平均數③迴歸直線的係數④全距⑤眾數⑥四分位數距(IQR)
(A) ①②③④ (B) ②③④ (C) ①②⑤ (D) ④⑥

2. 已知箱中置有4顆藍色球、6顆紅色球及n顆白色球。參加者自箱中抽出1球，確定顏色後放回。若抽得藍色球者可得1,000元，抽得紅色球者可得2,000元，抽得白色球者得0元。在抽出任1球之機率相等的條件下，若主辦單位希望參加者所得金額的期望值為400元，則n等於下列何者？
(A) 10 (B) 20 (C) 30 (D) 60

3. 某校有1,000位學生，英文段考成績呈常態分布，平均成績70分，標準差10分。請問此次段考成績超過90分的學生人數最靠近下列哪個數值？
註: $Z_{0.05} = 1.645$, $Z_{0.025} = 1.96$, $Z_{0.01} = 2.326$
(A) 1 (B) 5 (C) 25 (D) 50

4. 欲求常態母體平均數μ的信賴區間，抽取1組樣本數為30的隨機樣本。假設母體標準差未知，若信心水準不變，但樣本數增加為原本的9倍，而樣本標準差增加為原本的3倍，則新的信賴區間長度，約是原來信賴區間長度的多少倍？
(A) 0.5 (B) 1 (C) 1.5 (D) 3

5. 欲檢定虛無假設$H_0: \mu = 25$與對立假設$H_A: \mu < 25$，自標準差未知的常態分布中選取1組大小為n的隨機樣本，下列哪個結果最有可能支持H_A？
(A) $\overline{X} = 20, s = 3$ (B) $\overline{X} = 20, s = 7$ (C) $\overline{X} = 22, s = 3$ (D) $\overline{X} = 22, s = 7$

6. 在正斜分配資料中，下列何者正確？
(A)中位數等於平均值 (B)中位數小於平均值
(C)中位數大於平均值 (D)平均值大於或小於中位數

7. 有關P-值(P-value)之敘述，下列何者正確？
(A) P-值為型一錯誤的機率 (B) P-值是虛無假設為真的機率
(C) P-值等於0.05 (D) P-值越小，越可能會拒絕H_0

8. 若已知X和Y兩變數之線性迴歸方程式為$\hat{y} = 1.05 + 0.80x$，且$\sum_{i=1}^{20}(x_i - \bar{x})^2 = 144$，$\sum_{i=1}^{20}(y_i - \bar{y})^2 = 256$，則X和Y的相關係數為何？
(A) 0.2 (B) 0.4 (C) 0.6 (D) 0.8

9. 某市調公司欲探討3家飲料廠商(A、B、C)之市場佔有率是否相等。今自市場者中隨機抽取120位消費者，發現其中是A、B、C廠商的消費者分別為50、40、30人。根據此一資料，在 $\alpha = 0.05$的情況下，下列敘述何者正確？
(A)應採用變異數分析(ANOVA)
(B)此檢定的檢定統計量之值為5
(C)此檢定的臨界值為 $-\chi^2(0.025, df = 2) = -7.38$及 $\chi^2(0.025, df = 2) = 7.38$
(D)此檢定的P-值大於0.05,則在$\alpha = 0.05$的情況下，沒有足夠的證據去支持3家飲料廠商市場佔有率相同的說法

10. 欲檢定100個隨機樣本是否來自均勻分配，假設隨機變數的可能範圍為已知，並據此範圍以等間距的方式分成5組。假設第i組的觀察個數為Oi，使用卡方為檢定統計量，該檢定統計量的自由度為何？
(A) 4 (B) 5 (C) 20 (D) 99

11. 假設隨機抽取1組樣本數等於100的樣本，發現樣本平均數是10、樣本標準差是10，根據中央極限定理(Central Limit Theorem)，上述樣本平均數的抽樣分配，最接近下列哪一個機率分配？
(A)平均數等於0，變異數等於1的常態分配
(B)平均數等於0，變異數等於10的常態分配
(C)平均數等於10，變異數等於1的常態分配
(D)平均數等於10，變異數等於10的常態分配

12. 假設已知某個城市有60 %的大學生有打工經驗，也發現該城市有70 %的大學生有意願再繼續就讀研究所。若使用簡單隨機抽樣，由該城市的大學生中抽出1位同學，下列何者有誤？
(A)被抽出同學有打工經驗且有意願再繼續就讀研究所的機率至多為0.6
(B)被抽出同學有打工經驗且有意願再繼續就讀研究所的機率至少為0.3
(C)被抽出同學有意願再繼續就讀研究所的機率為0.63
(D)被抽出同學有意願再繼續就讀研究所的機率為0.7

13. 假設隨機變數x的平均數為15，標準差為4，若$y = 2x$，$z = x + 15$，以下敘述何者正確？
(A) z之變異數等於x之變異數 (B) x與z之相關係數小於1
(C) y之標準差最小 (D) y之平均數大於z之平均數

14. 若因子變異數分析如下表，則甲、乙、丙之值，下列何者有誤？

	平方和	自由度	平均平方和	F
組間	20	2	乙	丙
組內	30	甲	5	
總和	50	8		

(A)甲= 6 (B)甲= 4 (C)乙= 10 (D)丙= 2

15. 某知名人士想競選民意代表，但不清楚自己的民意基礎為若干，於是以抽樣方式了解他的支持度p。若估計誤差不超過3 %，請問需要抽多少樣本(請以95 %信心水準計算)？
(A) 654 (B) 1,068 (C) 1,243 (D) 4,269

16. 某候選人希望能提前自行宣布勝選，因此進行出口民調及統計檢定。隨機訪問400位該區已投票的民眾，其中228位投給該候選人。已知得票率需大於50 %才勝選，則該檢定的標準化檢定統計量之值為下列何者？
(A) 1.4 (B) 1.5 (C) 2.8 (D) 3

17. 在1所大學抽取100位學生隨機樣本之平均GPA為3.1，若用上述結果來估計此大學所有學生之平均GPA，則是在作下列哪一項統計？
(A)敘述統計 (B)母體統計 (C)樣本統計 (D)推論統計

18. 下列資料為500件發生在星期六晚上之意外事件分析，其中牽涉到酒駕與1輛汽車之意外事件比例為何？

牽涉汽車數				
酒駕否？	1	2	3	總數
是	60	110	30	200
否	40	215	45	300
總數	100	325	75	500

(A) 0.08 (B) 0.12 (C) 0.2 (D) 0.4

19. 某所高中三年級學生的數學成績近似於具有平均數300分與標準差35分的常態分配。藉由簡單隨機抽樣，選取4位高中三年級學生，他們的平均成績高於335分的機率大約為何？
(A) 0.15 (B) 0.10 (C) 0.05 (D) 0.025

20. 在最近針對100名小學六年級學生的研究中，得知他們每週平均看電視的時間是20小時，標準差為5小時，請問全體六年級學生每週平均看電視時間的95 %信賴區間，最接近下列何者？
(A) (20小時, 25小時) (B) (19小時, 21小時)
(C) (17小時, 23小時) (D) (15小時, 25小時)

21. 從蒐集來的100個樣本資料值計算得到$\sum X_i = 300$及$\sum {X_i}^2 = 925$，則其變異係數為何？
(A) 0.1675 (B) 0.2525 (C) 0.5025 (D) 5.9701

22. 有3法則可決定結果發生機率，包含古典法則、相對次數法及主觀認定法。請問下列3種情形，分別依據哪一種法則認定？
①美國人將贏明年法國公開賽
②公平骰子之任一點出現機率為1/6
③根據過去經驗，假定某種教科書每月之平均銷售為5,000本是合理的。
(A) ①相對次數法②古典法則③主觀認定法
(B) ①相對次數法②主觀認定法③古典法則
(C) ①古典法則②相對次數法③主觀認定法
(D) ①主觀認定法②古典法則③相對次數法

23. 有關F分配百分位數之說法，下列何者正確？
(A) $F_{0.10,10,20} = 1/F_{0.90,10,20}$ (B) $F_{0.10,10,20} = 1/F_{0.10,20,10}$
(C) $F_{0.90,10,20} = 1/F_{0.90,20,10}$ (D) $F_{0.90,10,20} = 1/F_{0.10,20,10}$

24. 在線性迴歸模式中，相較於y特定值之區間估計，y平均值之區間估計，下列何者正確？
(A)較窄 (B)較寬 (C)一樣 (D)無法判斷

25. 1個社區大學有150台PC，1天任1台要修之機率為0.025。為求剛好25台PC要修之機率，應使用下列何種機率分配近似求原機率？
①二項機率分配②波松機率分配③常態機率分配④指數機率分配
(A) ③近似① (B) ③近似④ (C) ②近似④ (D) ②近似①

26. 下列何者非屬關聯式資料庫(Structured Query Language, SQL)的資料模型？
(A)表格 (B)列 (C)欄 (D)超連結

27. 下列何者非屬使用巨量資料處理技術之主要目的？
(A)開發人工智慧導向的應用系統
(B)以即時或低延遲的方式擷取、處理和分析未繫結的資料串流
(C)轉換非結構化資料以進行分析和報告
(D)儲存和處理的資料數量對於傳統資料庫來說太大時

28. Spark巨量資料處理技術所使用的主要資料模型是彈性分散式資料集(Resilient Distributed Dataset, RDD)，下列何者不是RDD的核心成分？
(A) Partition Set (B) Relational Scheme (C) Dependencies Set (D) Partition Scheme

29. 下列關聯式資料庫一致性屬性之概念，何者正確？
(A)要求從意外的系統故障或停電狀況還原成上個已知狀態的能力
(B)要求並行的交易必須分開執行
(C)要求進行交易時資料就必須符合資料庫結構描述
(D)要求交易完整執行或完全不執行

30. 下列何者非屬MapReduce的主要計算動作？
(A) Combine (B) Map (C) Split (D) Visualize

31. 下列何者非屬盲信號分離技術？
(A)關聯規則分析 (B)主成分分析 (C)非負矩陣分解 (D)奇異值分解

32. 巨量資料由於數量大與特徵多元的特性，若適當透過探索式資料分析的技巧，能幫助我們了解資料的多種隱含現象，但不包含下列哪一個項目？
(A)分析各變數間的關聯性，找出重要的變數
(B)瞭解資料，獲取資料的資訊、結構和特點
(C)解析資料或事件的因果關係
(D)檢查有無離群值或異常值，看資料是否有誤

33. 卷積神經網路中將輸入的資料劃分為若干個矩形區域，對每個子區域輸出最大值，此機制為哪個網路層的運作目的？
(A)線性整流層 (B)記憶層 (C)完全連接層 (D)池化層

34. 特徵選擇主要目的是從資料集中選出最重要或最相關的特徵，來給機器學習建立模型，下列何者非屬特徵選擇的方法？
(A) Filter Methods (B) Allocate Methods (C) Wrapper Methods (D) Embedded Methods

35. Apache Hadoop系統是由下列何者領導進行開發案？
(A) Doug Cutting (B) Elon Musk (C) Jeff Bezos (D) Mark Elliot Zuckerberg

36. 如果有3位候選人競選台北市長，最後3人得票數很接近時，依照Information Theory的理論，此事件之Entropy值應該接近下列何者？
(A) 0.000 (B) 0.333 (C) 0.666 (D) 1.000

37. 針對數學模型：$Y = a_1X_1 + a_2X_2 + a_3X_3$，如果Y是類別型變數(Categorical Variable)，則該模型適用於下列何種資料分析方法？
(A) Data Clustering (B) Data Classification (C) Association Rules (D) Regression Analysis

38. 關於MapReduce程式設計與運算，下列敘述何者有誤？
(A) Map()方法是將資料區塊(Data Blocks)映射成鍵值對(key, value)
(B) Reduce()方法是將具有相同 key 的鍵值對，針對值的部分進行加總
(C) 為方便 Reduce()加總運算，會先對 Map()的結果依據值的大小排序
(D) Reduce加總運算後的結果就是輸出(Output)

39. 雲端運算上，NameNode非常重要，當NameNode壞掉時，所有資料儲存地址都會消失，導致資料會全部遺失。下列預防NameNode壞掉之方式，何者有誤？
(A)利用Network File System (NFS)將NameNode儲存資料做備份
(B)直接建立第2個NameNode當作備份
(C)每個DataNode節點都可以設置1個NameNode，避免遺失
(D)將NameNode執行於高階電腦上

【請另頁繼續作答】

40. 一般會要求Spark的使用者將RDD(Resilient Distributed Dataset)變數宣告為唯讀性質的val，原因是下列何者？
(A)確保資料在平行計算時的一致性　(B)確保資料在平行計算時的安全性
(C)確保資料在分散儲存時的完整性　(D)確保資料在分散儲存時的獨立性

41. 進行資料分群時，下列敘述何者表示好的分群結果？
(A)群內成員相似度高，群間成員相似度低
(B)群內成員相似度高，群間成員相似度高
(C)群內成員相似度低，群間成員相似度低
(D)群內成員相似度低，群間成員相似度高

42. 有關資料前處理方法-獨熱編碼(One-Hot-Encoding, OHE)，下列敘述何者有誤？
(A)若某個資料屬性有 N 種類別，則需要 N 個維度的二位元向量來表示
(B)若編碼後的向量，其維度 i 之值為 1，則代表資料屬於類別 i
(C)若(A)中的 N 值很大，則編碼後的向量中，1 出現的次數可能會不只 1 次
(D)編碼後的向量將會是稀疏向量(Sparse Vector)

43. Gmail屬於下列何者服務？
(A) Infrastructure as a Service (IaaS)　(B) Platform as a Service (PaaS)
(C) Software as a Service (SaaS)　(D) Function as a Service (FaaS)

44. 對於1年份刷卡資料，欄位有時間、店名、物品、刷卡金額及付款方式，若要使用MapReduce得到每家店此年銷售總額，請問應該使用下列何種Key-Value Pair？
(A)物品-刷卡金額　(B)店名-付款方式　(C)店名-刷卡金額　(D)時間-店名

45. 對於為什麼HDFS的檔案塊(Blocks)預設成128MB為最佳狀態之敘述，下列何者正確？
(A)如設成128KB，則1個cluster會有太多檔案塊，造成NameNode很難儲存每個檔案塊之位置
(B)因為我們只能夠在1個節點上，存1個檔案的其中1個檔案塊
(C)因為若把檔案塊設成128KB，對於多個Mappers要同時處理來說太小了
(D)因為這樣每8個檔案塊剛好就是1GB，剛好適合伺服器核心數都是4或8的倍數

46. 人工智慧(AI)、深度學習(DL)與機器學習(ML)之關係為下列何者？
(A) DL ⊂ ML ⊂ AI　(B) DL ⊂ AI ⊂ ML　(C) DL ⊂ ML = AI　(D) AI ⊂ ML ⊂ DL

47. 有關使用深度學習(Deep Learning)於序列資料中，下列敘述何者有誤？
(A)可以使用卷積神經網路(Convolutional Neural Networks)
(B)可以使用循環神經網路(Recurrent Neural Network)
(C)循環神經網路(Recurrent Neural Network)使用平行計算效率佳
(D)循環神經網路(Recurrent Neural Network)之優化較卷積神經網路(Convolutional Neural Networks)複雜

48. 視覺化是從巨量資料中得到洞見之重要工具，若我們想對於不同職業之平均收入繪圖，使用下列何圖較合適？
(A)直方圖　(B)折線圖　(C)河流圖　(D)散佈圖

49. 強化學習(Reinforcement Learning)容易應用在遊戲，而較難應用在真實世界的主要原因，下列何者正確？
(A)產生序列過長，不易訓練　(B)獎勵(Reward)不明確
(C)真實世界動作(Action)為連續動作　(D)對於真實世界無法完整模擬

50. 當雲端叢集有1個DataNode故障時，系統將進行何種操作？
(A)資料會遺失
(B) Hadoop會自動複製裡面所儲存之檔案塊於其他DataNode
(C) Hadoop會自動寄信給管理員警告此狀況
(D) Hadoop會繼續執行，只是有些資料塊會只有1份備份