經濟部所屬事業機構107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。

[C] 1. 一位正在競選的政治人物， 對30,000名註冊選民進行一項民意抽調。接受抽調訪談的200位註冊選民中，48%的選民表明將票投給他。下列敘述何者有誤？
(A)感興趣的母體為30,000名註冊選民 (B)樣本為接受訪談的200位註冊選民
(C) 48%是樣本參數 (D)參數是支持率

[D] 2. 當資料有正偏態時，其平均數、中位數與眾數的大小順序為何？
(A)平均數<中位數<眾數 (B)中位數<平均數<眾數
(C)平均數<眾數<中位數 (D)眾數<中位數<平均數

[C] 3. 當母體的觀察個數總是比樣本的觀察個數來得多，對於樣本統計量的敘述，下列何者正確？
(A)絕不能大於母體參數 (B)絕不能等於母體參數
(C)可能大於、小於或等於母體參數 (D)絕不能小於母體參數

[D] 4. 若A及B是獨立事件，有 $P(A)=0.65$ 及 $P(A \cap B) = 0.26$, 則 $P(A \cup B)$ 之值為何？
(A) 0.35 (B) 0.4 (C) 0.48 (D) 0.79

[D] 5. 某一個實驗包含3個步驟，第一個步驟有3種可能的結果、第二個步驟有5種可能的結果、及第三個步驟有4種可能的結果。則可能實驗結果的總數有多少種？
(A) 12 (B) 15 (C) 20 (D) 60

[D] 6. 目前具有2年經驗電腦程式設計師的時薪已達$240。由於對該職業的需求遽增，人們相信該職業的時薪應該有所增加。為了確定時薪是否已經提高，其對應的假設為下列何者？
(A) $H_0: \mu > 240$ $H_a: \mu \leq 240$ (B) $H_0: \mu = 240$ $H_a: \mu > 280$
(C) $H_0: \mu > 240$ $H_a: \mu \leq 270$ (D) $H_0: \mu \leq 240$ $H_a: \mu > 240$

[C] 7. 隨機變數X的機率分配函數為 $f(X) = X/15$ for X = 2, 3, 4 or 6。求 $3X+4$ 的期望值為何？
(A) 13 (B) 15 (C) 17 (D) 20

[B] 8. 一家汽車噴漆公司根據歷史數據發現，每部車噴漆所需時間在45至90分鐘之間呈現均勻分配。請問1部車噴漆所需時間不超過1小時的機率為何？
(A) 0.255 (B) 0.333 (C) 0.49 (D) 0.665

[B] 9. 關於適合度檢定(goodness of fit test)，下列敘述何者正確？
(A)永遠都是左尾檢定(lower-tail test) (B)永遠都是右尾檢定(upper-tail test)
(C)永遠都是雙尾檢定(two-tail test) (D)無法檢定

[C] 10. 收集應變數(Y)與自變數(X)的資料並進行簡單線性迴歸分析，分析的部分訊息為：
$\Sigma X = 90$, $\Sigma Y = 170$, $SSE = 505.98$, $n = 10$,
$\Sigma (Y - \overline{Y})(X - \overline{X}) = 466$, $\Sigma (X - \overline{X})^2 = 234$, $\Sigma (Y - \overline{Y})^2 = 1434$
下列何者有誤？
(A) SSR (sum of squares for regression)=928.02 (B)相關係數為 0.8045
(C)迴歸線之斜率估計值為 2.89 (D)判定係數為 0.6472

[C] 11. 進行銷售量(Y, 單位：千元)與銷售單價(X, 單位：元)的迴歸分析，產生結果：$\hat{Y}=60-8X$。下列敘述何者正確？
(A)單價每增加 1 元，銷售量減少 60 元 (B)單價每增加 1 元，銷售量減少 52 元
(C)單價每增加 1 元，銷售量減少 8000 元 (D)單價每增加 1 元，銷售量減少 52000 元

[C] 12. 設A、B、C為樣本空間S之三事件，且A、B、C為獨立事件，已知P(A)=0.4，P(B)=0.4，P(C)=0.2，求 $P((A \cup B) \cap C)$ 之值為何？
(A) 0.032 (B) 0.072 (C) 0.128 (D) 0.288

[B] 13. 簡單線性回歸分析中，已知SSE=500，SSR=300，請問判定係數 R^2 為何？
(A) 0.167 (B) 0.375 (C) 0.600 (D) 0.625

[A] 14. 已知組裝某機器的零件，所需時間具有平均數為14分鐘的指數分配。求組裝該零件所需時間不超過7分鐘的機率為何？
(A) $1-e^{-0.5}$ (B) $1-e^{-2}$ (C) $2-e^{-2}$ (D) $2-e^{-1}$

[B] 15. 某民調針對某候選人的支持度做調查，以電話隨機抽樣20歲以上民眾於1000份有效樣本中，顯示此候選人在95 %的信心水準下的信賴區間為 (0.33, 0.39)，請問若將信心水準改成99 %，此信賴區間的間距會有下列何種變化？
(A)變小 (B)變大 (C)不變 (D)無法判斷

[C] 16. 下列何種機率分配，其期望值等於變異數？
(A)指數分配 (B)常態分配 (C)卜瓦松分配 (D)二項分配

[A] 17. 有4組數字G1= (7,9,9,7,5)、G2=(7,6,5,6,7)、G3=(6,6,6,6,6)、G4=(3,4,5,4,3)，請問哪一組資料的標準差最大？
(A) G1 (B) G2 (C) G3 (D) G4

[C] 18. 假如 ρ_{XY} 表示隨機變數X和Y的相關係數，則下列何項正確？
(A) $\rho_{XY}=0$ 表示 X 和 Y 獨立 (B) ρ_{XY} 可以看出 X 和 Y 有非線性的相關
(C) $-1 \le \rho_{XY} \le 1$ (D) $\rho_{XY}=1$ 稱為完全負相關

[B] 19. 在下列的敘述中，何種條件下表示有愈多的證據拒絕虛無假設？
(A)有愈小的顯著水準 (B)有愈小的 p 值
(C)有愈小的臨界值(critical value) (D)有愈小檢定力 (power)

[C] 20. 已知修統計學課程的學生中有 40 % 會參加統計讀書會。根據以往的資料，參加統計讀書會的學生中有 65 % 會拿到成績A，而沒有參加統計讀書會的學生中有 10 % 會拿到成績A。假如在已知某位學生拿到成績A的情況下，求此位學生有參加統計讀書會的機率為何？
(A) 0.5642 (B) 0.75 (C) 0.8125 (D) 0.9215

[A] 21. 下列何者不是量測資料分散程度的統計量？
(A)眾數 (B)標準差 (C)全距 (D)變異數

[C] 22. 隨機抽取49包二砂糖，樣本平均數為60公斤，樣本變異數為12.25 公斤^2。已知 $t_{48,0.025}=2.0$ ，$t_{48,0.05}=1.7$ ，求母體的平均數之 95 % 信賴區間為下列何者 ？
(A) (58,62) (B) (58.3,61.7) (C) (59,61) (D) (59.15,60.85)

[B] 23. 一部門共有10位成員，其月薪(單位: 萬元)分別為：5,22,6,8,5,6,7,5,12,4。請問中位數為何？
(A) 5 (B) 6 (C) 6.5 (D) 8

[C] 24. 二項分配和超幾何分配之間，主要的差別在於超幾何分配具有下列何種特質？
(A)成功的機率必須小於 0.5 (B)成功的機率必須大於 0.5
(C)每次試驗彼此並不獨立 (D)隨機變數是連續的

[D] 25. 以ANOVA過程分析來自4個母體的資料，分別由每個母體抽出包括30個觀察值的樣本。此時檢定所需的 F 臨界值(critical value)，其分子與分母的自由度分別為下列何者？
(A) 3 與 30 (B) 4 與 30 (C) 4 與 120 (D) 3 與 116

[C] 26. 下列何者不是維度縮減(dimensionality reduction)之方法？
(A) Random projection (B) Principal component analysis (PCA)
(C) Clustering algorithms (D) Classical multidimensional scaling (cMDS)

[A] 27. 有關Hadoop的軟體疊層架構中之元素，下列何者有誤？
(A) Big Table (B) Hadoop MapReduce
(C) HDFS (D) HBase

[D] 28. 下列選項何者不是Big Data之應用技術？
(A) Google 用以指引 Web(index Web) 之技術
(B) Facebook 用以建立社交圖(build social graph)之技術
(C) Netflix 用以推薦電影(recommend movies)之技術
(D)比特幣(Bitcoin)用以預防盜竊及保證匿名之技術

[A] 29. Gartner Group於2012年定義巨量資料所具備3V的特性，下列敘述何者有誤？
(A) 差異(Variation) (B)多樣化(Variety) (C) 超大容量(Volume) (D)高流速(Velocity)

[B] 30. 學者Endsley(1995)針對決策過程所提出的處境察覺(Situation Awareness)模型中，決策人員察覺所處環境的3個狀態：①規劃(projection) ②知覺(perception) ③理解(comprehension)，請問此3個狀態正確的步驟順序為下列何者？
(A) ②①③ (B) ②③① (C) ③①② (D) ③②①

[A] 31. 下列何者是巨量資料領域的資料倉儲系統？
(A) HIVE (B) RDBMS (C) HDFS (D) Spark

[B] 32. GB、PB、TB、EB為4種電腦容量的單位，若依容量由大至小的排序，下列何者正確？
(A) PB＞TB＞EB＞GB (B) EB＞PB＞TB＞GB (C) PB＞EB＞GB＞TB (D) TB＞EB＞GB＞PB

[A] 33. 有關K-means集群(clustering)演算法，下列敘述何者有誤？
(A)不論不相似度測度(dissimilarity measure)為何，均適合採用 K-means 演算法
(B)同一筆資料，用 K-means 演算法分群兩次，可能得到不同之分群結果
(C) K-means 演算法的目標是使各個群組內部的均方誤差總和達到最小
(D) K-means 演算法的目標是使各個群組間之均方誤差總和達到最大

[D] 34. 巨量資料分析資料時，下列何者不是最常用的資料檔案格式來源？
(A) CSV (B) XML (C) JSON (D) TIF

[D] 35. 下列何者不是屬於NoSQL類型的資料儲存？
(A) MongoDB (B) CouchDB (C) Redis (D) MySQL

[A] 36. 某位數據分析師試圖自海量數據中提取潛在且有價值之資訊，此作法稱為下列何者？
(A) 資料探勘 (B)資料加密 (C) 資料維護 (D)資料查詢

[D] 37. 在巨量資料時代中，互聯網上所流動的網路行為資料可被用來從事許多極具價值之商業課題分析，試問下列哪一個工具無法用來捕捉網路流量？
(A) Google Analytics (B)百度統計 (C) Google趨勢 (D)微軟Power BI

[D] 38. 下列哪一選項不屬於「巨量資料」領域中所稱的資料型態特性？
(A)結構化資料 (B)非結構化資料 (C)半結構化資料 (D)去識別化資料

[A] 39. 試問apriori關連法則演算法中，哪兩項門檻值異動最為顯著影響資料探勘法則之數量？
(A)支持度、信賴度 (B)廣泛度、強弱度 (C)精密度、準確度 (D)清晰度、複雜度

[A] 40. 下列資料何者為結構化資料(Structured Data)？
(A)客戶交易資料表 (B)照片分享資料 (C)影音上傳資料 (D)社群討論文章資料

[B] 41. 關於ETL三個步驟的正確英文全名，下列何者正確？
(A) Extraction, Transport, Loading (B) Extraction, Transform, Loading
(C) Export, Transform, Loading (D) Extraction, Transform, Lifting

[A] 42. 在眾多巨量資料儲存作為中，若將同一份資料以副本方式分別存放在5個不同的場域，此舉主要是希望落實下列哪一個選項？
(A)提高資料異地備援能力與系統容錯性 (B)提高資料存取速度與可存取性
(C)提高資料儲存空間與可辨別性 (D)提高資料安全防護能力與不可否認性

[D] 43. 關於巨量資料領域常使用到的集群分析演算法具體作為，下列選項何者有誤？
(A) partitioning method (B) hierarchical method
(C) density-based method (D) bellman-ford method

[B] 44. 以巨觀的觀點，請將以下各個階段：①資料分析或知識挖掘 ②資料組織 ③資料視覺化或報告 ④資料收集與準備，按照巨量資料管道(pipelines)的正確順序排序，下列何者正確？
(A) ③④②① (B) ④②①③ (C) ④①②③ (D) ②④③①

[C] 45. 假設您取得了一份含有500位顧客資料的表單，其中的資料欄位包括顧客編號、生日、居住地、交易額，試問此表單最為滿足R語言中的哪一種資料結構？
(A)向量 vector (B)矩陣 matrix (C)資料框架 data.frame (D)串列 list

[D] 46. MapReduce是由Google所提出的一個巨量資料運算架構，試問下列哪一個選項正確表達該架構的資料輸入至輸出之順序？
(A)對應(Map)→歸納(Reduce)→排序(Sort)→合併(Merge)
(B)排序(Sort)→歸納(Reduce)→合併(Merge)→對應(Map)
(C)合併(Merge)→對應(Map)→歸納(Reduce)→排序(Sort)
(D)對應(Map)→排序(Sort)→合併(Merge)→歸納(Reduce)

[B] 47. 關於資料標準化，下列敘述何者有誤？
(A)標準化可用來消除變數之間的尺度差異問題 (B)資料標準化作業發生在資料模型建立後
(C)標準化可將各變數的資料範圍予以調整 (D)標準化可用來降低變數之間的變異程度

[A] 48. 請問下列何者等於1個EB(Exabyte)？
(A) 1024 PB(Petabyte) (B) 1024 TB(Terabyte) (C) 1024 GB(Gigabyte) (D) 1024 ZB(Zettabyte)

[C] 49. 關於Python語言的特性，下列敘述何者有誤？
(A)支援多種作業系統 (B)具備資料分析與視覺化繪圖能力
(C)屬於一種常見的編譯式程式語言 (D)可免費使用

[A] 50. 下列有關深度學習(Deep Leaning)之敘述，何者有誤？
(A)深度學習神經網路就是有很多中間層(hidden layers)的反向傳播神經網路(Back-Propagation Net)。
(B)強大的計算能力與高品質的大數據是促成深度學習成功的重要因素。
(C)卷積神經網路(Convolutional Neural Network, CNN)是機器視覺領域最有效的深度學習演算法。
(D)AlphaGo主要是使用深度學習並結合蒙地卡羅樹搜尋(Monte Carlo tree search，MCTS)演算法完成的。