

經濟部所屬事業機構 105 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 6 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，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. 下列對Poisson分配的敘述，何者正確？
I: 所有的Poisson分配為正偏態(positive skewness)
II: Poisson分配的期望值和標準差相等
III: Poisson分配為離散(discrete)機率分配
(A) I、II、和III (B) I和II (C) I和III (D) II和III
2. 一簡單隨機樣本，其 \bar{X} 抽樣分配的特徵(平均數、變異數和分配形狀)，容易受到下列何者之影響？
I: 樣本大小 II: 母體的變異數 III: 母體的平均數
(A) I (B) I和II (C) I和III (D) I、II、和III
3. 下列敘述何者正確？
I: 若 $X \sim N(\mu, \sigma^2)$ ，令 $Y = X - \mu$ ，則 $Y \sim N(0, \sigma^2)$
II: 若 $X \sim t(n)$ (自由度為n的t分配)，則 $X^2 \sim F(n, 1)$
III: 若 $Z \sim N(0, 1)$ ，則 $Z^2 \sim \chi^2(1)$ (自由度為1的chi-square分配)
(A) I和II (B) I和III (C) II和III (D) III
4. 令 S_1, S_2 分別為母體平均數 μ 的二個估計量(estimator)，又知道 $\mu=10$ 。對一已知的樣本數， S_1 抽樣分配其平均數為10，變異數為12；而相同的樣本數， S_2 抽樣分配其平均數為10，變異數為10。下列敘述何者正確？
(A) S_1 是不偏(unbiased)，但 S_2 有一偏差(bias)是2
(B) S_1 和 S_2 都是有偏差的
(C) S_1 的效率較 S_2 低
(D) S_1 和 S_2 有相同的效率
5. 某一分析家利用 $n = 500$ 個家庭的隨機樣本，估計家庭平均月收入的90 %信賴區間為 $60000 \leq \mu \leq 80000$ 。若分析家想以99 %信賴係數取代，則信賴區間會？
(A) 變窄且會有一較大的錯誤風險 (B) 變寬且會有一較大的錯誤風險
(C) 變窄且會有一較小的錯誤風險 (D) 變寬且會有一較小的錯誤風險
6. 下列哪些機率分配為離散(discrete)分配？
I: 超幾何 II: 指數 III: 二項 IV: 幾何
(A) I、III、和IV (B) I、II、和III (C) I和IV (D) I和III

7. 某一自助餐餐廳的主菜，分成四大類：豬肉、雞肉、海鮮和其他。隨機抽取700位顧客，所點之主菜結果如下：

主菜： 豬肉 雞肉 海鮮 其他
 人數： 370 172 115 43

想檢定個別的母體比例，依序主菜四大類分別為0.5、0.2、0.2、和0.1，則此檢定的檢定統計量為何？

- (A) 23.34 (B) 24.71 (C) 29.94 (D) 36.72

8. 在研究市場報酬率(X)和甲股票報酬率(Y)的一簡單線性迴歸中，有下列的結果：

$n = 5, \sum_i X_i = 0, \sum_i Y_i = 15, \sum_i X_i^2 = 20, \sum_i Y_i^2 = 55, \sum_i X_i Y_i = 5, \sum_i (X_i - \bar{X})^2 = 20$
 $\sum_i (Y_i - \bar{Y})^2 = 10$ ，則估計的迴歸函數為？

- (A) $3 + 0.25X$ (B) $-3 + 0.25X$ (C) $0.25 + 3X$ (D) $0.25 - 3X$

9. 某一影印機每100頁中印壞1張，若某僱員要影印500頁的報告，則在影印過程中沒有印壞的機率為何？

- (A) e^{-1} (B) $5e^{-1}$ (C) e^{-5} (D) $5e^{-5}$

10. X, Y為二個隨機變數，已知 $Var(2X - Y) = 32, Var(Y) = 4, Cov(X, Y) = -3$ ，則X和Y的相關係數為何？

- (A) -0.75 (B) -0.70 (C) -0.53 (D) -0.50

11. 在隨機完全區集設計(Randomized complete block design)有4個處理分佈在6個區集中，下列為其變異數分析表的部分結果，請問處理和區集的自由度分別為多少？

變異來源	平方和	自由度	均方	F值
處理				5
區集				
誤差	360			
總和	1440			

- (A) 4 ; 6 (B) 3 ; 6 (C) 4 ; 5 (D) 3 ; 5

12. 使用自然的地理位置或是其他界線將母體區分為許多區塊，並在每一個區塊中進行簡單隨機抽樣來組成樣本，此種抽樣方法稱為？

- (A) 區塊抽樣 (B) 分層抽樣 (C) 簡單抽樣 (D) 群集抽樣

13. 在假設檢定時，如果虛無假設為真，不拒絕虛無假設的機率是0.95，如果虛無假設為假，拒絕虛無假設的機率是0.9，則下列敘述何者有誤？

- (A) 型 I 錯誤機率為 0.05 (B) 型 II 錯誤機率為 0.1
 (C) 兩種檢定錯誤的機率總合為 1.00 (D) 統計檢定力(power)為 0.9

14. 有一個隨機變數X，其機率分配如下：

$$f(x) = \begin{cases} \frac{ax^3}{x!}, & x = 1, 2, 3 \\ 0.5, & \text{all other values} \end{cases}, \text{ 則常數 } a \text{ 為?}$$

- (A) 2/57 (B) 2/43 (C) 1/19 (D) 3/28

15. 下列哪一種統計圖可以決定第90百分位數的大約位置？

- (A) 直方圖 (B) 箱型圖 (C) 莖葉圖 (D) 肩型圖

16. 下表列示了一組隨機樣本的資料，其相關係數為何？

X	27	31	16	8	11
Y	41	47	24.5	12.5	17

- (A) -0.1 (B) 0.1 (C) 0.5 (D) 1

17. 從一組母體中選取68個觀察值為樣本，樣本平均數是1.72、樣本標準差是0.64。另從第二組母體中選取33個觀察值為樣本，樣本平均數是0.82、樣本標準差是0.48。請進行下列的假設檢定，

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

並使用顯著水準0.05，下列敘述何者有誤？

- (A)此為雙尾檢定 (B)自由度為99的T檢定
(C)檢定統計量為 3.15 (D)拒絕 H_0
18. 關於卡方檢定 (Chi-square Test) 之適用，下列何種有誤？
(A)國籍和宗教有無關聯 (B)骰子是否公正
(C)性別和是否吃素有無差異 (D)性別在智商上有無差別
19. 關於二因子變異數分析中檢定的敘述，下列何者正確？
(A)先檢定主效應顯著性，若不顯著，再檢定交互作用效應顯著性
(B)先檢定交互作用效應顯著性，若不顯著，再檢定主效應顯著性
(C)只需要檢定主效應顯著性，不必檢定交互作用效應顯著性
(D)先檢定交互作用效應顯著性，若顯著，再檢定主效應顯著性
20. 已知 $P(A) = 0.35$ ， $P(B|A) = 0.4$ ，則 $P(A \cap B^c)$ 為多少?(其中 B^c 為事件B的餘事件)
(A) 0.14 (B) 0.21 (C) 0.35 (D) 0.6
21. 關於複迴歸分析，下列敘述何者正確？
(A)自變數越多， R^2 越大 (B)自變數越多，adj- R^2 越大
(C)自變數越少， R^2 越大 (D) adj- R^2 一定大於 R^2
22. 下列何者為母體參數？
(A)行政院主計總處調查顯示，上個月台灣地區失業率為4.91 %
(B)有十萬名網友參加的網路民調的結果
(C)內政部每年公布的台灣地區新生兒出生數
(D)衛福部抽查鴨蛋、鵝肉殘留抗生素的不合格率
23. 等式 $SS(\text{Total}) = SST(\text{treatment}) + SSB(\text{block}) + SSE$ 適用於哪一種實驗設計模型？
(A)完全隨機設計 (B)隨機區集設計 (C)二因子設計 (D)不完全區集設計
24. 關於Tukey多重比較的用途，下列何者正確？
(A)檢定常態性 (B)檢定成對平均數的差異
(C)檢定一群平均數的差異 (D)檢定變異數的差異
25. 要決定資料是否來自於特定多項分配(Multinomial distribution)時，我們要選用哪一種檢定？
(A)適合度檢定 (B)列聯表獨立性檢定
(C)比較多個母體的卡方檢定 (D)檢定常態性的卡方檢定
26. 費林分類法(Flynn's Taxonomy)是一種計算機架構的分類方式，根據指令和資料的相對關係，可以區分為四種，如以單一處理器來執行單一程式段落及單一資料的模式，下列何者正確？
(A) MIMD (B) SIMD (C) SISD (D) MISD
27. 在CUDA(Compute Unified Device Architecture)中，memory的分配上是相當重要的問題，GPU上具有shared memory、global memory、constant memory，下列敘述何者正確？
(A)存取速度: global memory > shared memory > constant memory
(B)容量大小: global memory > constant memory > shared memory
(C)要在 shared memory 中宣告變數需在前面加上 device
(D)使用 shared memory 配置的變數，可以被grid中的所有thread存取

28. 在CUDA(Compute Unified Device Architecture)程式編寫中，如果我們要給每一個thread唯一的ID，起始值為0，且為整數，例如:0、1、2、3...循序下去，可以透過下列何者取得？
 (A) $\text{blockIdx.x} * \text{blockDim.x} + \text{threadIdx.x}$ (B) $\text{blockIdx.x} * \text{blockDim.x} - \text{threadIdx.x}$
 (C) $\text{blockIdx.x} * \text{threadIdx.x} + \text{blockDim.x}$ (D) $\text{blockIdx.x} / \text{blockDim.x} + \text{threadIdx.x}$
29. 關於資料庫中使用的索引結構，下列何者有誤？
 (A) B+-Tree (B) FLA-Tree (C) QuadTree (D) R*-Tree
30. 關於支持向量機(Support Vector Machine, SVM)，下列敘述何者有誤？
 (A) Kernel Function不只一種
 (B)使用Kernel Trick來增加效率
 (C)可對資料進行非線性分類
 (D)將高維度的資料降至低維度以提升準確率
31. 隨著巨量資料成為許多公司的競爭優勢，所有產業的架構也將重新調整，關於巨量資料對企業的影響，下列敘述何者有誤？
 (A)各公司能得到的利益不會是平等的
 (B)得利最多的會是大型企業和中型企業，小型企業將遭到嚴重衝擊
 (C)持有許多大型資料來源，能輕鬆取用資料才是優勢
 (D)競爭優勢的最核心重點，不是擺在硬體設施本身的規模
32. 分析巨量社群網路資料時，下列敘述何者有誤？
 (A)分析使用者的朋友數對了解社群網路是有幫助的
 (B)分析使用者打卡(check in)地點有助於了解使用者空間上的資訊
 (C)在分析時，使用者皆被視為只有一個社群網路帳號
 (D)社群網路上的使用者會隨著時間而增減
33. 關於機器學習演算法，下列敘述何者有誤？
 (A) AdaBoost透過調整訓練資料(training data)被抽樣到的機率以提升效能
 (B) K-Nearest Neighbor可用於監督式學習
 (C)決策樹(decision tree)可以不是二元樹(binary tree)
 (D)過度解讀(overfit)可提升機器學習演算法之準確率，故大部分機器學習演算法皆嘗試過度解讀訓練資料
34. 關於巨量資料的特色，下列敘述何者有誤？
 (A)巨量資料特點之一是產生速度很快 (B)線上社群網站是巨量資料來源之一
 (C)巨量資料一定要搭配雲端運算才能處理 (D)使用者的網路瀏覽紀錄亦可視為巨量資料
35. 關於衡量機器學習中分類器(Classifier)效能的指標，下列何者有誤？
 (A) Walter-Kimplin Curve (B) Precision
 (C) Recall (D) ROC Curve
36. 關於從社群網站抓取社群網路拓撲(Social Network Topology)資料，下列敘述何者有誤？
 (A)深度優先搜尋(Depth-First-Search)是常用的方法之一
 (B)廣度優先搜尋(Breadth-First-Search)是常用的方法之一
 (C)取得之社群網路拓撲可視為整個線上社群網路的抽樣(sample)
 (D)為了避免偏差(bias)，可抓取多份不同的社群網路拓撲進行分析
37. 關於巨量資料分析常使用的 NoSQL 資料庫，下列敘述何者有誤？
 (A) NoSQL 指的是 Not only SQL，通常不使用 SQL 做為查詢語言
 (B) NoSQL 與傳統資料庫相同，都必須隨時確保資料一致
 (C) NoSQL 沒有資料庫 schema 的欄位架構，因此可以隨著資料變動有彈性的進行欄位調整
 (D) NoSQL 通常不支援 Join 操作

38. 對於兩事件 A, B 的關聯式規則(Association Rule)，下列敘述何者有誤？
 (A) $A \Rightarrow B$ 的支持度(Support)定義為 A, B 兩事件共同出現的機率
 (B) $A \Rightarrow B$ 的信心度(Confidence)定義為事件 A 發生的情形下，事件 B 也同時發生的條件機率
 (C) 項目 A, B 順序的調換對支持度不會有影響
 (D) 項目 A, B 順序的調換對信心度不會有影響
39. 關於分群演算法(Clustering)，下列敘述何者正確？
 (A) 分群屬於非監督式學習 (B) k-means 演算法進行分群前，可先不決定 k 值
 (C) 分群演算法通常計算複雜度較分類要高 (D) 分群的效果與資料數量、群集數量都無關
40. 關於 MapReduce model，下列敘述何者有誤？
 (A) MapReduce model 是 Google 所提出，用於大規模資料的平行運算
 (B) 在 MapReduce 分散式計算 model 中，只有 Map 及 Reduce 兩種運算
 (C) Map 和 Reduce 的概念是從 functional programming 而來
 (D) Hadoop 為目前較為知名的 open source MapReduce project
41. 關於 HDFS 架構，下列敘述何者有誤？
 (A) Hadoop JobTracker 必須也是 HDFS 的 Namenode
 (B) Hadoop JobTracker 負責分配工作，而 TaskTracker 負責執行工作
 (C) Namenode 只能有一個，而 Datanode 通常有很多個
 (D) Namenode 主要負責儲存檔案系統的索引，而 Datanode 負責儲存檔案的 data blocks
42. 下列何種方法比較不適合進行平行化？
 (A) K-Means clustering (B) Logistic Regression
 (C) Newton's method for finding roots (D) PageRank
43. 根據巨量資料分析的資料屬性，下列何者與其他屬性差異最大？
 (A) 社群網路分析 (B) 股市趨勢分析 (C) 路況車流分析 (D) 環境感測資料分析
44. 我們會定義資料為巨量、大或者海量，最主要的原因，是目前的資訊科技環境，提供了以下何種重要特性，使其稱為 Big Data？
 (A) Variety (B) Veracity (C) Velocity (D) Volume
45. Spark 的 Word Count 程式片段範例如下所示，下列敘述何者有誤？
- ```
text_file = sc.textFile("hdfs://...")
counts = text_file.flatMap(lambda line: line.split(" ")) \
 .map(lambda word: (word, 1)) \
 .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://...")
```
- (A) counts 從頭到尾就是單一個 RDD  
 (B) flatMap() 是依照分隔符號 (" ") 將文字檔分成一個一個的字  
 (C) map() 將每個字計算一次  
 (D) reduceByKey() 將相同字的值相加
46. 在巨量資料的想法中，強調的是去尋找事件或事物的相互關係(Correlation)，非著重在傳統因果關係(Causality)的尋找，主要是因為當今環境變遷迅速，資料可以被即時收集且運算，因此，相互關係的尋找，更優於確立因果關係。請問，以下何種模型，可以用來協助尋找相互關係？  
 (A) 時間序列分析 (B) 類神經網路分析 (C) 關聯規則分析 (D) 決策樹分析

47. 於巨量資料當中進行關聯規則的探勘(Association Rule Mining)，採用Apriori 的方法，將會於每階段產生大量的組合，使得計算顯得沒有效率。但實際上，此方法因下列何種特性，使其不用去進行每一個候選組合的測試，而可以大量減少其演算時間？  
(A)大數法則            (B)反單調性            (C) 80/20準則            (D)維爾特定律
48. 在巨量資料的技術堆疊中，哪一層將透過MapReduce，慣用程式碼的額外處理與建構中介資料結構，諸如：統計模型或資料立方體...等，所產生的結構，做為額外分析或傳統查詢工具查詢之用，使巨量資料做好接受進一步分析的準備？  
(A)應用程式碼            (B)資料            (C)商業觀點            (D)儲存
49. 雲端運算技術是巨量資料處理的基礎，關於雲端運算技術，下列敘述何者正確？  
(A)雲端運算強調基於實體化等技術，在分散式硬體環境上提供共用資源服務  
(B)雲端運算以伺服器集群為中心，運算和資料儲存都由網路中心的雲端完成  
(C)雲端運算的伺服器集群仍存在不穩定性，這使得基於雲端運算實現應用的範圍受限  
(D)雲端運算伺服器之間是分散式結構，流量具天然的對稱特點，符合現階段的網路頻寬特點
50. 下列哪一項目不是Google 於2003~2006年前後，發表奠定巨量資料技術理論基石的技術？  
(A) BigTable            (B) GFS            (C) Hadoop            (D) MapReduce