# 經濟部所屬事業機構110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)。<br>2.可使用本甄試簡章規定之電子計算器。<br>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。<br>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。<br>5.本試題採雙面印刷，請注意正、背面試題。<br>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。<br>7.考試時間：90分鐘。 |

1. x和y的樣本標準差分別為10和15，x和y之間的共同變異為-120，x和y之間的相關係數為下列何者？
(A) -0.8 (B) -0.5 (C) 0.5 (D) 0.8

2. 假設樣本迴歸方程式為$\hat{Y} = 3 + 0.40x$。假設當x為10時，觀察到y為8，模型預測的殘差為何？模型是否低估或高估了y的值？
(A)殘差= -1，模型高估了y (B)殘差= 1，模型高估了y
(C)殘差= 1，模型低估了y (D)殘差= -1，模型低估了y

3. 下列哪一項的估計需要抽樣？
(A)台灣失業率 (B) 2021年台北市的總降雨量
(C)新竹攻城獅職業籃球隊2021年的命中率 (D)某大學新生的入學考試平均分數

4. 一項針對150名主管的調查，他們被詢問在面試中求職者最常犯的錯誤是什麼，並給出了六個不同的錯誤類型。下列哪一項最適合呈現上述訊息？
(A)長條圖 (B)直方圖 (C)莖葉圖 (D)列聯表

5. 假設A及B是獨立事件，若P(A) = 0.5及P(B) = 0.3，則$P(A \cup B)$之值為何？
(A) 0 (B) 0.15 (C) 0.2 (D) 0.8

6. 在設計假設檢定時，下列哪一項較適合？
(A)最大化$\alpha$並最小化$\beta$ (B)最小化$\alpha$並最大化$\beta$ (C)最大化$\alpha$和$\beta$ (D)最小化$\alpha$和$\beta$

7. 下列哪一項電腦儲存容量單位最大？
(A) ExaByte (EB) (B) GigaByte (GB) (C) KiloByte (KB) (D) PetaByte (PB)

8. 下列哪一項違反了迴歸分析的假設？
(A)誤差項呈正態分佈 (B)誤差項與解釋變量相關
(C)誤差項的均值為零 (D)誤差項有一個恆定的變異

9. 某統計學教授發現學生會在他的考試中作弊，他估計有95 %的機率能抓到班上的學生作弊，但有1 %的機率是他指控學生作弊但學生實際並沒有作弊，考慮虛無假設「學生沒有作弊」，發生型一錯誤的機率是多少？
(A) 1 % (B) 5 % (C) 95 % (D)資訊不夠，無法計算

10. 一位經濟學家對研究一個國家消費者的收入感興趣，已知其母體標準差為1,000美元，隨機抽樣50人得到的平均收入為15,000美元，在這樣的條件下，90 %信賴區間的寬度為多少？
註：$Z_{0.01} = 2.326$ , $Z_{0.05} = 1.645$ , $Z_{0.10} = 1.282$
(A) 232.60美元 (B) 364.30美元 (C) 465.23美元 (D) 728.60美元

11. 根據某銀行的信用卡資料，發現在過去約有7 %的客戶會違約無法繳清帳款，導致銀行產生呆帳。此外，銀行中不會違約的客戶，約有10 %會不小心忘了繳交一個月的卡費。若某客戶有一個月的卡費沒有繳交，那麼該客戶最終違約無法繳清帳款的機率為多少？
(A) 0.13 (B) 0.23 (C) 0.33 (D) 0.43

12. 一位分析師正在進行滿意度調查，從1,000名來自四家房屋仲介的新房購買者名單中抽樣，分析師從購屋者中抽取400名為樣本，他隨機抽取每家仲介的100名購屋者，請問這是一個簡單隨機樣本(simple random sampling)的例子嗎？
(A)是，因為樣本中的每個買家都是隨機抽樣的
(B)是，因為每家仲介的購買者在樣本中都有同等的代表
(C)否，因為400位買家樣本被選中的機會並不均等
(D)否，因為人口由四種不同房屋仲介的購買者組成

13. 您已經為$\mu$建立了95 %的信賴區間為$10 \le \mu \le 15$，若我們測試虛無假設H0 : μ等於16，與對立假設H1 : μ不等於16，在顯著水準$\alpha = 0.10$下，您將做出什麼決定？
(A)不拒絕H0　　(B)拒絕H0
(C)接受H0　　(D)無法從所提供的訊息判斷我們的決定

14. 關於卡方檢驗，下列敘述何者正確？
①卡方檢驗的結論說明所研究的變量是否相關，以及它們是如何相關
②要進行卡方檢驗，您必須有足夠大的樣本，此條件為每個觀測值至少是5
③要進行卡方檢驗，您必須有足夠大的樣本，此條件為每個預期值至少是5
(A) ①　(B) ②　(C) ③　(D) ①②③

15. 關於逐步迴歸(stepwise regression)，下列敘述何者正確？
(A)逐步迴歸用於從一大組變數中確定哪些變數可用於預測應變數的值
(B)逐步迴歸結果應直接當作預測Y應變數的最終模型
(C)在逐步迴歸中，犯一個或多個型一或型二錯誤的機率非常小
(D)當考慮包含一個或多個虛擬變量(dummy variable)時，不能使用逐步迴歸

16. 關於統計資料蒐集與分析，下列敘述何者有誤？
(A)用迴歸模型分析資料庫資料，可以確定獨立變數與應變數的因果關係
(B)運用實驗設計的好處是可以排除其他不感興趣因素的干擾
(C)使用問卷調查時，採用便利抽樣可以節省成本
(D)對於開放性的問題探索，可以使用半結構式訪談來蒐集資料

17. 關於中位數與平均數，下列敘述何者有誤？
(A)常態分配的平均數與中位數相等
(B)平均數容易受到極端值影響
(C)母體平均數高於中位數時，母體資料分配很可能是右偏的
(D)在貧富差距很大的國家中，人民薪資的平均數會比中位數來的小

18. 下列關於假設檢定的敘述，有幾個是正確的？
① p-value的定義是虛無假設成立，獲得檢定統計量及更極端數值之機率
②我們可以用建立平均值的信賴區間來檢驗，虛無假設H0：$\mu \ge 2$，對立假設HA：$\mu < 2$
③$\alpha$表示虛無假設為真的機率
④如果第一個假設檢定的p-value值為0.00029，而第二個假設檢定的p-value值為0.001，我們可以得出結論：第一個檢定有更多的證據可以拒絕虛無假設
(A) 0　(B) 1　(C) 2　(D) 3

19. 受訪者被問到「你現在的體重比3年前多還是少？」這個回應是什麼測量尺度？
(A)名目尺度(nominal) (B)順序尺度(ordinal) (C)區間尺度(interval) (D)比例尺度(ratio)

20. 某汽車廠的數據顯示，過去進貨的電子零件一萬個中，有5 %是劣質品，若從電子零件中隨機抽4個來測試，至少有一個是劣質品的機率為多少？
(A) 0.250　(B) 0.226　(C) 0.185　(D) 0.171

21. 當發生型一錯誤時，下列敘述何者正確？
(A)你拒絕一個正確的虛無假設　　(B)你不拒絕一個正確的虛無假設
(C)你拒絕一個錯誤的虛無假設　　(D)你不拒絕一個錯誤的虛無假設

22. 某公司的客服中心，平均每三分鐘會接到一通電話，請問六分鐘內都沒有接到電話的機率是多少？
(A) 0.135　(B) 0.105　(C) 0.065　(D) 0.035

23. 在標準化模式中，關於常態分配之敘述，下列何者正確？
(A)平均數為1，變異數為0　　(B)平均數為0，標準差為1
(C)機率面積等於0.5　　(D)不能用於近似離散的機率分配

24. 關於中央極限定理，下列敘述何者有誤？
(A)若母體為均勻分配，僅抽10個樣本，其樣本平均數分配即為常態分配
(B)有了中央極限定理，我們才能從樣本統計量去推估母體參數
(C)若母體為常態分配，僅抽10個樣本，其樣本平均數即為常態分配
(D)若母體為指數分配，僅抽5個樣本，其樣本平均數即為常態分配

25. 歷年台北市長各候選人的選舉得票率，皆與士林區天玉里開出之結果很接近，若在台北市長選舉前，對天玉里的居民隨機調查各候選人的支持度，此種作法較類似於下列何種抽樣方式？
(A)簡單隨機抽樣 (B)群集抽樣 (C)分層隨機抽樣 (D)便利抽樣

26. 巨量資料最後提出的第四個V是什麼？
(A)大量化Volume (B)多樣化Varity (C)快速化Velocity (D)真實性Veracity

27. 下列何者屬於類神經網路的機器學習方法？
(A) SVM (B) Gaussian Mixture Model
(C) BERT (D) Adaboost

28. 下列何者不是維度縮減(Dimensionality Reduction)的方法？
(A) Random Forest (B) t-Distributed Stochastic Neighbor Embedding
(C) Principal Component Analysis (D) Linear Discriminant Analysis

29. 關於巨量資料，下列敘述何者正確？
(A)巨量資料的分析目的是從大量資料中找出因果關係
(B)隨著巨量資料分析技術的進步，數據的真實性已不影響分析結果
(C)好的巨量資料運算服務，可以根據運算需求與時效性，動態的擴增運算資源
(D)如何妥善處理及保存大量的數據資料並非巨量資料分析涵蓋的範疇

30. 關於巨量資料思維，下列敘述何者正確？
(A)巨量資料中包含大量的雜訊，運用傳統統計方法較能降低雜訊對模型的影響
(B)企業在蒐集資料的過程中可能觸及個人隱私的疑慮
(C)巨量資料分析純粹是應用於商業競爭上的技術，無法廣泛被應用於日常生活中
(D)巨量資料分析建模的過程中不需要考慮資料是否與背景假設吻合

31. 機器學習(Machine Learning)是屬於下列哪一個領域的子集合？
(A)深度學習 (B)人工智慧 (C)資料處理 (D)網路系統

32. 通常處理有問題的資料時，不包括下列哪一項？
(A)移除 (B)插補 (C)加權 (D)替換

33. 與資料庫(Database)比較，資料倉儲具有下列哪一種特性？
(A)具有跨平台作業能力 (B)具有資料修改及新增功能
(C)資料結構高度正規化 (D)儲存短期詳細的作業資料

34. 關於平行運算(Parallel Computing)，下列何者不是分散式計算(Distributed Computing)的特性？
(A)共享稀有資源 (B)平衡負載 (C)共享記憶體 (D)允許計算錯誤

35. 請問以500 Mbps(Mega bit per second)的網路速度傳送1 Petabyte，不考量其他影響，大約需要多少時間才能傳送完畢？
(A)大約2年 (B)大約200天 (C)大約20天 (D)大約2小時

36. 下列何者可能是模型有配適不足(Under-fitting)的狀況？
(A)訓練誤差較小，測試誤差較小 (B)訓練誤差較小，測試誤差較大
(C)訓練誤差較大，測試誤差較小 (D)訓練誤差較大，測試誤差較大

37. 關於監督式學習(Supervised Learning)，下列敘述何者有誤？
(A)需要有標記(labeled)的數據做為訓練集
(B)決策樹(Decision tree)是一種監督式學習方式
(C) K-mean分群法(k-means clustering)是一種監督式學習方式
(D)監督式學習是機器學習方法中的一個子類別

38. 關於MAP Reduce Programming Model，下列敘述何者正確？
(A)是一種處理地理資訊的程式設計方法
(B)是一種資料壓縮的技術
(C)是一種類神經網路的程式設計架構
(D)是一種用平行化、在分散式系統上處理數據的程式設計架構

39. 下列何種方法能對大量特徵屬性的資料進行分析以萃取出重要訊息？
(A)主成分分析(Principal Component Analysis)
(B)關聯法則(Association Rule)
(C) K均值法(K-means)
(D) K近鄰法(K-nearest Neighbors)

40. 關於監督式學習(Supervised Learning)，下列敘述何者有誤？
(A)函數的輸出可以是預測一個分類標籤或是一個連續的值
(B)訓練資料是由輸入及預期輸出所組成
(C)訓練資料過少時，可利用拔靴法(Bootstrap)進行修正
(D)主成分分析是典型的監督式學習方法

41. 關於非監督式學習(Unsupervised Learning)，下列敘述何者正確？
(A)訓練資料是由輸入及預期輸出所組成
(B)可用於做分群與降維的工具
(C) K Nearest Neighbor演算法屬於非監督式學習方法
(D)決策樹(Decision Tree)是一種常用的非監督式學習方法

42. 關於深度學習，下列敘述何者有誤？
(A)與傳統機器學習最主要的區別在於數據依賴性，亦即資料量多寡會影響演算法效能
(B)企業過往採用傳統機器學習方法的各項應用，只要升級成以深度學習方法為基礎，都能在效率與效果上有卓越的提升
(C)卷積神經網路(Convolutional Neural Network)是一個可以被應用於圖形辨識的深度學習方法
(D)深度學習是機器學習的一種演進技術

43. 信用卡評分系統是應用機器學習的哪一個概念？
(A)社群網絡分析　(B)非監督式學習　(C)關聯法則　(D)監督式學習

44. 關於MapReduce，下列敘述何者正確？
(A)特別適合用於小型的計算任務
(B)運作原理是將待處理的大量資料劃分為多個數據塊，每個數據塊對應到一個計算任務來運行
(C)需要搭配運算效能強大的伺服器才能發揮效果
(D)使用在需要即時回傳結果的實時計算上極具效率

45. 下列哪一個方法能在建模過程中添加隨機噪訊，來解決模型變異過高的問題？
(A)拔靴集成法(Bootstrap Aggregating)
(B) XGBoost(eXtreme Gradient Boosting)
(C)支援向量機(Support Vector Machine)
(D)生成式對抗網路(Generative Adversarial Networks)

46. 下列何者不屬於關係資料庫管理系統(Relational Database Management System)？
(A) Spark SQL　(B) Oracle　(C) Hive　(D) HBase

47. 關於隨機森林(Random Forest)，下列敘述何者正確？
(A)在建立的過程是隨機採樣的，因此即使不剪枝也不容易出現過度配適(Over-fitting)的現象
(B)在資料取樣進行拔靴集成法(Bootstrap Aggregating)的過程中，每次會取出 n 筆資料，且不重複選取直到資料完全被使用
(C)無法在決定類別的同時，評估變數的重要性
(D)隨機森林中的決策樹個數增加，也不會增加運算效能的負擔

48. 下列何者非Hive支援之資料型別？
(A) Long　(B) Float　(C) Boolean　(D) Array

49. 關於Hadoop Distributed File System(HDFS)，下列敘述何者正確？
(A)預設資料塊大小為16 MB
(B)支援資料的隨機讀寫
(C)除了Client外，組成架構還包含Name Node、Secondary Name Node及Data Node
(D)如果Name Node當機，Secondary Name Node可做為備份來維持系統正常運作

50. Hadoop Distributed File System(HDFS)中的Name Node職責不包含下列何者？
(A)管理資料塊對映資訊　(B)管理HDFS的名稱空間
(C)處理客戶端讀寫請求　(D)儲存實際的資料塊