

經濟部所屬事業機構 109 年新進職員甄試試題

類別：統計資訊

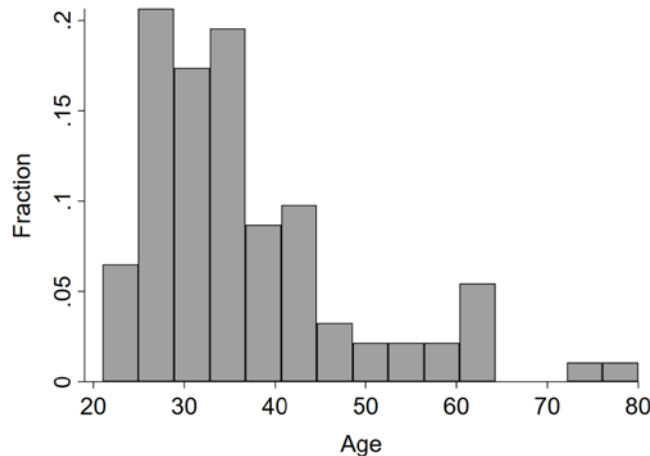
節次：第二節

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. 下圖顯示自 1929 年至 2019 年，奧斯卡金像獎最佳女主角得主的(得獎時)年齡分布。根據此圖，得主年齡的平均數與中位數最接近何者？



- (A) 平均數：36；中位數：33
(B) 平均數：33；中位數：36
(C) 平均數：36；中位數：28
(D) 平均數：33；中位數：45
2. 某教授蒐集 80 位勞工的資料，將他們的時薪(單位：新臺幣)做為應變數，性別做為自變數，進行迴歸分析。得到結果如下：

	估計值	標準誤
性別	11.8	3.2
常數	160.9	10.5

其中男性勞工的性別值為 1，女性的性別值為 0。若重新定義性別變數，讓男性的性別值為 0，女性的性別值為 1。使用同一樣本估計迴歸模型，得到的常數值和性別係數值，將分別是多少？

- (A) 172.7, -11.8 (B) 172.7, 11.8 (C) 160.9, -11.8 (D) 160.9, 11.8
3. 根據一項研究，國道三號中和至土城間的車行時速，大致符合平均 90 公里、標準差 5 公里的常態分配(normal distribution)。該路段設有一台測速照相機，凡超過速限 100 公里視為超速。假設每輛車的速度彼此獨立，請問 3 台車行經該測速照相機，皆無超速的機率最接近下列何者？
- (A) 99 % (B) 93 % (C) 89 % (D) 85 %

4. 配適一條簡單迴歸模式： $Y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_i$ ，

其中， ε_i 服從平均數為0，標準差為1的常態分配， $i = 1, 2, \dots, 150$ ， $\bar{X} = 150$ ， $\bar{Y} = 85$ ，

$\sum_1^{150} (X_i - \bar{X})^2 = 20$ ， $\sum_1^{150} (Y_i - \bar{Y})^2 = 2000$ ， $\sum_1^{150} (X_i - \bar{X})(Y_i - \bar{Y}) = -180$ 。請問 ANOVA 表內，MSR 為多少？

Source	自由度	SS	MS	F-value	p-value
Model	a	SSR	MSR	F	<0.0001
Error	b	SSE	MSE		
Total	c	2000			

- (A) 16.2 (B) 180 (C) 810 (D) 1620

5. 變異係數的定義為下列何者？

- (A) 標準差/平均數 (B) 平均數/標準差 (C) 變異數/平均數 (D) 平均數/變異數

6. 1位小兒科醫生，想瞭解2019年台北市新生兒的體重。他從台北市12個行政區隨機抽選3個，然後蒐集該年度3個行政區所有新生兒的體重資料。請問他採用的是何種抽樣方法？

- (A) 簡單隨機抽樣(simple random sampling) (B) 分層抽樣(stratified sampling)
(C) 群集抽樣(cluster sampling) (D) 多階段抽樣(multistage sampling)

7. 某航空公司托運行李收費的政策如下：每位乘客的第1件行李20美金，第2件行李50美金，第3件以上不予載運，且不得協助他人托運行李。假設50%的乘客沒有托運行李，40%的乘客托運1件行李，10%的乘客托運2件行李。1班載有200名乘客的飛機，預期可為航空公司帶來多少行李費收入？

- (A) 2,000美金 (B) 2,600美金 (C) 3,000美金 (D) 5,000美金

8. 有關型一錯誤(Type I Error)的敘述，下列何者最為正確？

- (A) 其機率為1-型二錯誤機率
(B) 其機率永遠設為5%
(C) 是拒絕真的虛無假設時，所犯的錯誤
(D) 是對立假設為真時，不拒絕虛無假設所犯的錯誤

9. 在多元迴歸模型中，若要進行聯合檢定(joint hypothesis test)，應使用下列何種檢定法？

- (A) t 檢定 (B) Z 檢定 (C) 卡方檢定 (D) F 檢定

10. 指數分配是伽瑪分配的一個特例，當伽瑪分配中的何項參數固定時，伽瑪分配將退化成指數分配？

- (A) $\alpha = 1$ (B) $\alpha = 2$ (C) $\beta = 1$ (D) $\beta = 2$

11. 小美回到宿舍拿起雜誌翻閱，突然間上個月才更換的燈泡燒毀了！包裝盒上明明寫著可以照亮3,000個小時的燈泡，總共才使用30個小時就燒毀，已知該燈泡的壽命是服從指數分配，請問1,000個使用此品牌燈泡的消費者當中，比小美更倒楣的人約有多少？

- (A) 1 (B) 10 (C) 100 (D) 條件不足無法計算

12. 下列有關變異數分析的敘述，何者有誤？

- (A) 每一組資料都必須服從常態分配 (B) 每一組資料的母體變異數都必須相同
(C) 用以檢定平均數 (D) 用以檢定變異數

13. 假設隨機變數 X 與 Y 的聯合機率分配為 $f(x, y) = (x + y)/30$, $x = 0, 1, 2, 3$, $y = 0, 1, 2$, 則 $P(x > y)$ 為下列何者?
 (A) 0 (B) 1/2 (C) 2/3 (D) 3/5
14. 公司舉辦健行活動, 某員工去程平均時速為6公里, 回程平均時速為3公里, 則該員工的總平均時速為下列何者?
 (A) 3.5公里 (B) 4公里 (C) 4.5公里 (D) 5公里
15. 就同一組資料進行假設檢定時, 下列敘述何者有誤?
 (A) 右尾檢定和左尾檢定所計算出來的檢定統計量相同
 (B) 單尾檢定和雙尾檢定所計算出來的檢定統計量相同
 (C) 右尾檢定和左尾檢定所計算出來的p值相同
 (D) 右尾檢定和左尾檢定所計算出來的p值和為1
16. 某公司販賣的10公克果醬包, 根據過去的資料顯示其重量的標準差為0.2公克, 今任取1包該公司販賣的10公克果醬包, 其重量介於9.6公克到10.4公克之間的機率至少為下列何者?
 (A) 1/2 (B) 2/3 (C) 3/4 (D) 4/5
17. 1磅精心調配的綜合咖啡豆當中包含了非洲、美洲、亞洲等3地生產的咖啡豆, 假設 X 與 Y 分別代表這1磅的綜合咖啡豆之中非洲豆和美洲豆的重量, 已知 X 與 Y 的聯合機率密度函數為 $f(x, y) = 24xy$, $0 < x < 1$, $0 < y < 1$, $x + y < 1$ 。若非洲豆的重量為0.75磅, 試問美洲豆重量小於0.1磅的機率為下列何者?
 (A) 1/25 (B) 2/25 (C) 3/25 (D) 4/25
18. 從台灣全省抽樣1,000家公司, 調查其去年的業績, 發現結果如下: 業績成長的有150家, 業績衰退的有550家, 業績不變的有300家, 而其中服務業所佔的比例分別為45%, 30%, 50%。若從中選取1家公司, 其為服務業的機率為下列何者?
 (A) 0.3825 (B) 0.4016 (C) 0.4167 (D) 0.4207
19. 令 (X_1, X_2, X_3) 為由常態母體 $N(\mu, \sigma^2)$ 抽出的一組隨機樣本, T_1, T_2, T_3, T_4 均為 μ 的估計量, $T_1 = (3X_1 + 3X_2 + 4X_3)/10$, $T_2 = (X_1 + X_2 + X_3)/3$, $T_3 = (X_1 + 2X_2 + 3X_3)/6$, $T_4 = (2X_1 + 3X_2 + 4X_3)/9$, 請問下列何者為 μ 的不偏估計量中變異數最小者?
 (A) T_1 (B) T_2 (C) T_3 (D) T_4
20. 棒球教練想要透過假說檢定確認某選手的打擊率是否超過3成, 乃蒐集過去50次的打擊紀錄做為樣本, 得到的打擊率為0.33。假設該選手的每次打擊都是獨立事件, 請問在設定顯著水準為5%的條件下, p值(p-value)係指下列何種機率值?
 (A) 0.05 (B) $\Pr(Z > 1.645)$
 (C) $\Pr(\text{樣本打擊率} > 0.3 \mid \text{真實打擊率} = 0.3)$ (D) $\Pr(\text{樣本打擊率} > 0.33 \mid \text{真實打擊率} = 0.3)$
21. 已知某股票的報酬率服從期望值為 μ , 標準差為 σ 的對數常態分配, 則該股票報酬率的期望值為下列何者?
 (A) μ (B) e^μ (C) $e^{\frac{\mu + \sigma^2}{2}}$ (D) $e^{\mu + \frac{\sigma^2}{2}}$
22. 為了解台灣人民的網路使用情形, 隨機抽取600位年滿15歲以上的國民調查, 其中有360位每天都使用網路, 據此估計台灣15歲以上的國民每天使用網路的比率為0.6, 則在信賴係數(信心水準)為95%時, 估計誤差之最大值為下列何者?
 ($Z_{0.05} = 1.645$, $Z_{0.025} = 1.96$)
 (A) 0.0200 (B) 0.0337 (C) 0.0392 (D) 0.0475
23. 已知 \bar{X} 服從常態分配 $N(\mu, \sigma)$, 設 μ 的95%信賴區間為 (L_1, U_1) , μ 的90%信賴區間為 (L_2, U_2) , 下列敘述何者正確?
 (A) $L_1 < L_2 < U_2 < U_1$ (B) $L_2 < L_1 < U_2 < U_1$ (C) $L_1 < L_2 < U_1 < U_2$ (D) $L_2 < L_1 < U_1 < U_2$

24. 為了解房屋售價(X)與面積(Y)之間的關係，隨機選取12戶已成交房屋，所得資料為 $\sum_{i=1}^{12} x_i = 3177$ 、 $\sum_{i=1}^{12} y_i = 272$ 、 $\sum_{i=1}^{12} x_i^2 = 869111$ 、 $\sum_{i=1}^{12} y_i^2 = 6464$ 、 $\sum_{i=1}^{12} x_i y_i = 74113$ ，X與Y的相關係數為下列何者？
 (A) 0.68 (B) 0.73 (C) 0.82 (D) 0.89
25. 二個互斥事件A、B，機率分別是0.5、0.6，則 $\Pr \{ A^c \cup B^c \}$ 的值為何？(註： A^c 、 B^c 分別表示A、B的餘集合)
 (A) 0.7 (B) 0.8 (C) 0.9 (D) 1.0
26. 下列何者不是Apache Hadoop之特色？
 (A)使用MapReduce程式框架
 (B)支援Java語言
 (C) Apache Mahout是一種用來支援Apache Hadoop分散式工作程序管理的程式庫
 (D)使用Hadoop Distributed File System
27. 對於HDFS Shell指令，下列敘述何者有誤？
 (A) Hadoop fs -ls用來列出HDFS檔案與目錄列表
 (B) Hadoop fs -rm input/masters刪除HDFS上的檔案
 (C) Hadoop fs -cat input/slaves壓縮HDFS內之檔案
 (D) Hadoop fs -put conf input將本地端檔案上傳至HDFS
28. 對於k-均值(k-means)聚類演算法的敘述，下列何者有誤？
 (A) k-均值中update的程序，將更新聚類中心
 (B) k-均值中assign的程序，將比較各資料點之間的距離，並將各資料點以隨機方式分配至其中一個聚類
 (C) k-均值中的k值表示資料將分成幾類，需事先給定
 (D)每一次執行的k-均值演算法，其結果可能會不一樣
29. 對於NoSQL資料庫說明，下列何者有誤？
 (A)可採用Key-DM資料架構來建立資料庫 (B)使用記憶體方式建立分散資料庫
 (C) MongoDB是一種NoSQL的資料庫 (D)各種NoSQL資料庫所支援的語言可能不同
30. 對監督式學習(supervised learning)的說明，下列何者有誤？
 (A)監督式學習需要使用標記過類別的資料(labeled data)進行訓練(training)
 (B)監督式學習可以跟非監督式學習整合，進行資料分析
 (C)目前所有的監督式學習的方法，皆無法對線性不可分(nonlinear)的資料進行分類
 (D)將資料的所有特徵(feature)放入監督式學習，不一定會找到最好的分類方式
31. 以機器學習對於巨量資料進行分析後，通常會使用混淆矩陣(confusion matrix)，對於所產生的分類器進行評估，其中將分析結果分為true positive (TP)、true negative (TN)、false positive (FP)及false negative (FN)，下列敘述何者有誤？
 (A)正確率Accuracy = $(TP + TN) / (TP + TN + FP + FN)$
 (B)召回率Recall = $TP / (TP + FN)$
 (C)精確率Precision = $TP / (TP + FP)$
 (D) F1 Score = $TP / (TP + FP + FN)$
32. 對於大量資料分析的技術，下列敘述何者有誤？
 (A) PageRank是用來對於數值資料進行資料壓縮的演算法
 (B)支持向量機(support vector machine)的核函式(kernel function)選擇會影響分析結果
 (C) k-最近鄰居(k-nearest neighbors)是用來對於資料分類的監督式演算法
 (D) k-中心點(k-medoids)是用來對於資料分類的非監督式演算法

33. 使用MapReduce框架來設計一個字數統計(word count)程式，其程式所進行的常用標準程序應為下列何者？
- (A) Input → Splitting → Mapping → Shuffling → Reducing → Final Result
 (B) Input → Mapping → Splitting → Shuffling → Reducing → Final Result
 (C) Input → Mapping → Splitting → Reducing → Shuffling → Final Result
 (D) Input → Reducing → Splitting → Shuffling → Mapping → Final Result
34. CAP定理可用來分析NoSQL資料庫的特性，下列對於NoSQL資料庫及CAP定理之敘述，何者有誤？
- (A) CAP定理的「C」代表的是一致性(consistency)性質
 (B) CAP定理的「A」代表的是原子性(atomicity)性質
 (C) CAP定理的「P」代表的是分區容錯(partition tolerance)性質
 (D) NoSQL資料庫輸出內容，可以搭配非監督式演算法進行資料探勘分析
35. 下列工作何者適合在一般資料庫進行，但不適合在NoSQL環境？
- (A)維持保證多方同時交易一致性的管理機制(concurrency control)
 (B)複雜度高的加總計算
 (C)有時效性的趨勢分析
 (D)綱要(schema)尚未穩定的彈性分散式資料儲存與擷取
36. 下列何種計算方法原則是先綜觀全局，再分層深化處理的廣度優先策略？
- (A) Apriori原則找所有frequent patterns (B) FP-growth計算association rules
 (C) Agglomerative hierarchical clustering (D) MapReduce
37. 對關聯規則(association rule) $X \rightarrow Y$ 的理解，下列何者較為正確？
- (A) X的值決定Y的值 (B) X是因，Y是果
 (C) X之後的下一階段是Y (D) X出現時，也容易見到Y
38. 文字探勘(text mining)常見的TF-IDF處理，IDF是以甚麼為單位的值？(Document--D代表文件，Term--T代表字詞，Weight--W代表加權比重)
- (A) IDF (D, W) (B) IDF (T) (C) IDF (T, D) (D) IDF (W)
39. 下列何者為公有區塊鏈(block-chain)的特性？
- (A)由鏈外仲裁者驗證資料 (B)由認證金融組織負責Bitcoin(比特幣)運作
 (C)已上鏈資料無法更改 (D)不支援智能合約的數位服務
40. 有關巨量資料的多類(variety)特性，下列敘述何者正確？
- (A)一般感知器(sensor)所回傳的資料為無結構性資料(unstructured data)
 (B)監視器所錄下的視訊(video)為半結構性資料(semi-structured data)
 (C)線上訂房的網頁資料為半結構性資料(semi-structured data)
 (D)書籍文本文字(text)為結構性資料(structured data)
41. 關於Hadoop分散式檔案系統HDFS的檔案文件儲存，下列敘述何者有誤？
- (A)檔案內容將被切割為區塊(chunk)儲存
 (B)檔案區塊大小不一，視檔案內容而定
 (C)檔案區塊大小通常為64 MB以上
 (D)每一檔案區塊至少將會複製二份存放

42. 資料倉儲設計會希望是主題導向(subject-oriented)，下列敘述何者正確？
(A)主題不應被期待在倉儲系統運作後自然浮現
(B)分析維度的準備與主題制定是分別獨立的设计工作
(C)資料倉儲設計不易，最好盡量納入多元主題
(D)主題需要經常性的檢討並重新訂定
43. 資料立方(data cube)是由資料倉儲綱要所建立的多維度數值統計資訊，若決策者希望獲得某單一維度的部分條件之統計量來分析資料時，可以用下列何種 OLAP 的運算來達成？
(A) roll up (B) drill down (C) slice (D) dice
44. 對Hadoop Distributed File System (HDFS)的敘述，下列何者有誤？
(A)提供容錯功能 (B)至少包含一台data node
(C)至少包含一台name node (D)至少包含一台analytics node
45. 有關MapReduce程式的執行，下列敘述何者正確？
(A)工作追蹤器(job tracker)主要是回報資料節點中 Map 或 Reduce 任務的執行情況
(B)主節點(master node)若發生故障，只有主節點上的任務(task)會失敗
(C)資料節點(data node)若發生故障，該節點的任務(task)將會重新指定給其他資料節點
(D)為獲取最大的平行計算效益，Map任務和Reduce任務不會安排至在同一資料節點上執行
46. 以資料分析為目的構建資料倉儲(Data Warehouse)時，其資料特性將不包括下列何者？
(A)主題導向性(subject-oriented) (B)資料異動性(volatile)
(C)多重整合性(integrated) (D)時間變動性(time variant)
47. 當在具有數值屬性(numerical attribute)的資料集中探勘關聯式規則(association rule)時，必須預先對屬性資料完成何種處理？
(A)補值處理(missing value imputation) (B)離散化(discretization)
(C)比例轉換(scaling) (D)正規化(normalization)
48. 巨量資料分析前進行屬性特徵選擇(Feature-Selection)時，下列何種方法不適合用來做為選擇的標準依據？
(A)資訊增益(Information Gain) (B)均方根誤差(Root Mean Squared Error)
(C)卡方係數(Chi-Squared coefficient) (D)相關係數(Pearson's correlation coefficient)
49. 深度神經網路(deep neural networks)的神經元中通常輸出時會經過激發函數(activation function)的轉換，下列針對常用激發函數的敘述何者有誤？
(A) ReLU可以避免過度擬合(overfit)的問題
(B) Sigmoid會有梯度消失(vanishing gradient)的問題
(C) ReLU會發生死亡神經元(dead neural)的問題
(D) Sigmoid會有梯度爆炸問題(vanishing gradient)的問題
50. 集成式分類方法是將弱分類器(weak classifiers)集合起來用以增強分類的準確率與穩定度。請問下列何者不是集成式分類方法？
(A) AdaBoost (B) Gradient Boosted Trees
(C) Random Forest (D) K-Nearest Neighbor