經濟部所屬事業機構104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，前 25 題每題各 1.5 分、其餘 25 題每題 2.5 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. 抽樣分配是指下列何者的分配？
(A)母數 (B)統計量 (C)母數與統計量 (D)非母數也非統計量

2. 某一公司有4個零件供應商(S1、S2、S3、S4)，其零件在S1、S2、S3、S4之供應比例分別是30%、20%、10%、40%，根據該公司之統計S1、S2、S3、S4所供應的零件之不良率分別是2%、4%、3%、2%，如果該公司隨機抽出一個零件為不良品，則該零件來自S2供應商的機率為下列何者？
(A) 0.008 (B) 0.2 (C) 0.24 (D) 0.32

3. 某甲欲研析全臺灣地區加油站之服務量，而執行一抽樣調查以推估總服務量，為顧及各地區之特性，宜使用下列何種抽樣設計？
(A)集群抽樣 (B)分層抽樣
(C)簡單隨機抽樣取出不放回 (D)簡單隨機抽樣取出放回

4. 在簡單線性迴歸模型，斜率的估計值代表下列何種意涵？
(A)在獨立變數為0時，依變數的平均估計值
(B)在獨立變數為0時，依變數的估計值
(C)當獨立變數變動一單位時，依變數的平均變動估計值
(D)觀察值的預測值

5. 下列何者正確？
(A)互斥事件彼此獨立 (B)獨立事件彼此互斥
(C)互斥事件不會同時發生 (D)獨立事件彼此互相影響

6. 若每一個可能的樣本被抽到的機會相等，此抽樣方法為下列何者？
(A)簡單隨機抽樣(Simple random sampling) (B)分層隨機抽樣(Stratified random sampling)
(C)部落抽樣(Cluster sampling) (D)系統抽樣(Systematic sampling)

7. 若A與B事件互相獨立，P(A) = 0.38且P(B) = 0.55，則 P(B｜A) 為下列何者？
(A) 0 (B) 0.17 (C) 0.38 (D) 0.55

8. 某甲欲研析新設加油站之顧客平均等待時間，作為安排服務人員人數之依據，請問下列哪一種分配最適合用來描述兩位顧客到達該加油站的時間間隔？
(A)常態分配(Normal distribution) (B)卜瓦松分配(Poisson distribution)
(C)指數分配(Exponential distribution) (D)二項分配(Binomial distribution)

9. X，Y為二服從標準常態分配之隨機變數且兩者獨立，則下列何者有誤？
(A) E(X/Y) = 1　(B) E(X+Y) = 0　(C) E(X2+Y2) = 2　(D) E(X-Y) = 0

10. 樣本統計量的期望值等於所欲估計的母數時，則此樣本統計量具有下列何種特性？
(A)最小變方　(B)隨機性　(C)不偏性　(D)一致性

11. 統計學家證實，要提高抽樣的準確度，最好的方式為下列何者？
(A)增加樣本數　(B)做到隨機抽樣
(C)使用最精準的分析軟體　(D)使用速度最快的電腦硬體

12. 下列何者不是用於資料的相關性分析(Correlation Analysis)？
(A)卡方檢定　(B)相關係數　(C)共變異數　(D)四分位數

13. 分析資料、建構模型來預測顧客的貸款申請是「安全的」或「有風險的」，是下列何者？
(A)關聯規則探勘　(B)分類　(C)迴歸　(D)群集分析

14. 下列何者為巨量資料最適當的定義？
(A)巨大資料量的資料集
(B)資料量大於1 TB的資料集
(C)資料量超出傳統資料庫的抓取、儲存、管理和分析能力的資料集
(D)資料量超出人類的抓取、儲存、管理和分析能力的資料集

15. 巨量資料會使分析資訊的方式產生三大改變，不包括下列何者？
(A)能夠取得、分析的資料量大為增加　(B)不會堅持一切都要做到精準
(C)放下長久以來對於因果關係的堅持　(D)不需找出資料之間的相關性

16. 關於巨量資料分析的概念，下列何者正確？
(A)巨量資料分析主要是針對量大的數據分析，因此對於資料來源多樣性和資料產生快慢等因素將不列入考慮
(B)巨量資料分析的資料格式僅限於結構化資料，因此非結構化與半結構化資料不納入其分析的範疇
(C)巨量資料分析的建模是透過由下而上(Bottom-Up)的數學歸納推理方式來解讀資料的規則，這和傳統資料庫系統以由上而下(Top-Down)的建模方式有所不同
(D)巨量資料分析方法可以利用現有資料庫管理系統技術來完成

17. 關於分類的訓練資料集與測試資料集的敘述，下列何者有誤？
(A)訓練資料是從要分析的資料庫中隨機取樣
(B)訓練資料必須已經知道其類別
(C)測試資料集不應該包含訓練資料集中的資料
(D)測試資料可以不知道其類別

18. 請問機器學習(Machine Learning)上所使用的深化學習(Deep Learning)和下列哪一個演算法有直接相關連？
(A)類神經網路　(B)迴歸分析　(C)貝氏網路　(D)決策樹

19. 巨量資料分析是一連串分階段流程的處理步驟(Pipeline)，針對此巨量資料分析的流程，下列何者正確？
(A) ETL(Extract Transform Load)的處理是屬於此分析流程中最後階段
(B)現階段我們已經有一套資訊系統可以完整涵蓋所有巨量資料分析的階段流程
(C)統計學習的建模與分析因為需大量人力介入，因此是獨立出來的步驟，所以它不屬於巨量資料分析流程的一環
(D)巨量資料分析流程必須要透過人為的介入與客製化的操作，因此目前整個巨量資料分析流程尚無法完全自動化完成

20. 一般巨量資料處理的單位為PB級；1 PB的資料為1 GB資料的幾倍大？
(A) 1,000 (B) 1,000,000 (C) 1,000,000,000 (D) 1,000,000,000,000

21. 下列哪一項資訊技術和巨量資料最不相關？
(A)資料探勘 (B)物件導向軟體開發 (C) Hadoop (D) NoSQL

22. 對於決策樹(Decision Tree)機器學習演算法，下列何者有誤？
(A)決策樹最末端葉面點(Leaves)是標示資料分類別的結果
(B)決策樹中間的點是提供資料分類時特徵值的判斷
(C)決策樹的分類需要將所有訓練資料集的資料正確分類
(D)隨機樹叢(Random Forest)演算法是整合多個小決策樹來進行資料分類

23. 關於Hadoop的敘述，下列何者有誤？
(A) Hadoop的做法，是將資料打散成小塊，分散到不同的電腦中
(B)由於資料量夠大，Hadoop不會儲存資料的備份
(C) Hadoop預設，由於資料量十分龐大，所以不可能在處理之前就清理乾淨、排序整齊
(D)與過去的關聯式資料庫相比，Hadoop輸出的結果比較不準確

24. 如果使用者在MapReduce中打算使用外部執行檔來定義其工作，需要利用下列何者？
(A) Virtual Machine (B) Streaming (C) Pipeline (D) Filter

25. 下列何者是知識發現(Knowledge Discovery)的正確程序？
(A)資料探勘、資料準備、樣式評估、知識呈現
(B)資料準備、資料探勘、樣式評估、知識呈現
(C)資料準備、樣式評估、資料探勘、知識呈現
(D)資料準備、資料探勘、知識呈現、樣式評估

26. 下表為臺北市及新北市居民對於是否應適度調漲電價來促進節約能源意識之比例

	臺北市	新北市
樣本數	$n_1 = 400$	$n_2 = 600$
贊成人數	300	300
反對人數	100	300

假定臺北市及新北市居民之贊成比例分別為P_1及P_2，
則檢定H_0：$P_1=P_2$之檢定統計量為$\frac{\hat{P}_1-\hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})[\frac{1}{n_1}+\frac{1}{n_2}]}}$，請問本檢定統計量中分母$\widehat{P}$之值為何？
(A) 0.25 (B) 0.5 (C) 0.6 (D) 0.75

27. 保險公司請求某家諮詢公司，幫忙確認非常高機率的假保險理賠事件。已知某工業的假理賠要求比例為3%。該諮詢公司決定從該工業隨機抽樣100家公司確認其是否申請保險理賠。他們認為從這100家公司的申請理賠數目會得到保險公司想要的信息。此諮詢公司最有可能是利用下列哪種機率分配來分析此假理賠問題？
(A)卜瓦松分配 (B)二項分配 (C)超幾何分配 (D)以上皆非

28. 某間航空公司預計從甲、乙、丙、丁這4種訂票系統中擇一，並希望該訂票系統讓乘客遇到較少障礙。因此該航空公司設計一實驗設計來收集資料，其中每個訂票系統隨機選擇5週接受訂票，共計20週，收集訂票者碰到訂票障礙的個數如下。
甲：(12, 14, 9, 11, 16) 乙：(2, 4, 7, 3, 1) 丙：(10, 9, 6, 10, 12) 丁：(7, 6, 6, 15, 12)
請問該做何種數據分析？
(A)檢定母體變異數間的差異 (B)以t檢定作平均數差異分析
(C)以Z檢定作平均數差異分析 (D)利用一維變異數分析，採用F檢定

29. 對兩個獨立母群體以T檢定檢定其母體均數是否相等，下列何者正確？
(A)兩樣本的變異數相等　(B)母體分配近似常態分配
(C)兩個樣本數相等　(D)以上皆是

30. 若 Z 是一個標準常態隨機變數，則 P(-1.5<Z<0)將會比 P(1.5<Z<3.0)
(A)小　(B)相等　(C)大　(D)以上皆非

31. 一個完全隨機設計，下列何者正確？
(A)含一個因子、一個集區和許多觀測值　(B)只含一個因子，而此因子含有數個試驗群組
(C)含一個因子和一個集區　(D)含數個因子，而因子含有數個試驗群組

32. 下表是一個不完整的變異數分析表(ANOVA table)，請問檢定統計量 F 值為何？

變異來源(SV)	平方和(SS)	自由度(df)	均方(MS)	F 值
組間	*	4	*	*
組內	60	*	*	
總和	140	19		

(A) 0.2　(B) 4.0　(C) 5.0　(D) 6.0

33. 欲建立母體均數的區間估計值，假設使用36個觀察值時，其母體均數的區間估計值為 19.76 ± 1.32，則當樣本大小*n*以144 取代36時，其母體均數的區間估計值應為下列何者？
(A) 19.76 ± 2.64　(B) 19.76 ± 0.66　(C) 9.88 ± 2.64　(D) 4.94 ± 1.32

34. 某甲收集100筆資料，其平均值為50，變異數為100，中位數為60。請問下列何者正確？
(A)至少有75筆資料值介於30與70之間　(B)大約有95筆資料介於30與70之間
(C)這100筆資料分佈為右偏　(D)這100筆資料分佈為對稱

35. 卡方分配(chi-square distribution)可應用在下列哪項？
(A)推論單一母體的變異數　(B)檢定配適度
(C)檢定兩個變數的獨立性　(D)以上皆是

36. 透過迴歸分析演算法可以進行資料關連性分析。現考量運用三種行銷廣告通路：電視、廣播、報紙的預算金額分配額度大小，找出它們對於產品銷售值(sales)的影響。請問下列敘述何者有誤？
(A)電視、廣播、報紙三種特徵值為獨立變數，而sales是我們要預測的結果，因此為相依變數
(B)對於獨立變數如電視、廣播、報紙，我們允許使用連續量化變數，同時也允許使用類別式變數(categorical variable)兩種資料型態，但是我們對於相依變數sales，僅允許使用類別式變數型態，因此一般性的迴歸分析演算法可以用來進行資料分類
(C)三種變數是否要納入迴歸分析方程式可以透過T-statistic的檢驗，並考量其F-statistic值越大、p-value的值愈小的變數者，其較具有影響力
(D)迴歸分析方程式有時需要考量獨立變數所表示特徵值是否有交互作用，而判斷交互作用的影響力也可運用T-statistic的檢驗，並透過高F-statistic值和低p-value值來判斷

37. 透過統計學習方法論，我們希望找出一個決策樹資料分類器以避免資料分類時的過度解讀(overfitting)，請問對於資料過度解讀的現象，下列何者有誤？
(A)決策樹分類器對於訓練資料集過度解讀，可以提高對於訓練資料集分類結果的準確率
(B)過度解讀對於測試資料的分類判斷結果其準確率很低
(C)我們可以用簡單決策樹分類器模型，來避免產生訓練資料集過度解讀的現象
(D)當每一次納入新的特徵值來成長決策樹結構時，當新增特徵值對於訓練資料集無法產生有效分類就該停止，因此用事後修剪決策樹的方法將無法避免資料過度解讀

38. 對於巨量資料分析Spark平台，下列何者有誤？
(A) Spark有提供結構化資料格式的巨量資料分析功能
(B) Spark能夠透過多種電腦語言，如Scala, Python, R, Java來呼叫系統引擎
(C) Hadoop原有的MapReduce巨量資料分析的計算原理無法在Spark上來進行
(D) Spark是選擇Apache開放性系統發展模式，因此我們可以看到Spark系統的原始程式碼

39. 一般推薦系統(Recommender System)時常會採用下列哪一個方法作為核心技術，來估計產品與使用者間的可能關係？
(A) Matrix Factorization (B) Hashing
(C) Linear Discriminative Analysis (LDA) (D) Part-of-Speech (POS) Tagging

40. 下列何種統計學習的演算法是用來進行資料的分群(Clustering)，但不能用來進行資料分類(Classification)?
(A) K-Means (B) Bayes Nets (C) Logistic Regression (D) Support Vector Machine

41. 巨量資料中的資料類別出現的頻率，時常會形成所謂的長尾現象，一般可利用以下哪種統計工具來描述此種形式的資料分佈？
(A) Zipf (B) Gaussian (C) Dirichlet (D) Uniform

42. 以分析使用者打卡(check in)的應用為例，如果打卡的位置為變數X。使用者A習慣在一個位置打卡(如公司)，其打卡位置的分佈統計模型為$P_A(X)$；使用者B時常在都市內到處打卡，其分佈統計模型為$P_B(X)$。請問何者的entropy，也就是H(.)較高？
(A)使用者A (B)使用者B (C)兩者一樣 (D)無法比較

43. 在視覺化工具中為了在二維畫面中檢視資料點之間的關係(相似度或距離)，例如社群網路、地圖中重要都市等，一般都使用下列哪種形式的演算法？
(A) Inverted-based Indexing (B) Multidimensional Scaling (MDS)
(C) Approximate Nearest Neighbor (ANN) (D) Ranking Methods

44. 關於Bayes Nets與Naïve Bayes兩種統計學習演算法，下列何者有誤？
(A) Naïve Bayes的演算法假設特徵值相互之間的關連性是獨立
(B) Naïve Bayes學習演算可以算是Bayes Nets的一個特例，因為Bayes Nets並未假設特徵值相互之間的關連性是獨立
(C) Bayes Nets的計算是用條件機率來表示特徵值之間的關連性，並利用這些條件機率的相乘積結果來表示分類別機率的大小，以判斷最後分類別
(D)因為Naïve Bayes演算法的侷限性，它僅能作為資料分群，而不能作為資料分類使用

45. Hadoop一般對於疊代式(Iterative)程序執行起來較沒效率，主要原因為下列何者？
(A) Iteration不易平行化 (B)跨Iteration間的狀態不容易維持
(C) Hadoop不支持Iteration (D) Hadoop不支援C語言

46. 對於巨量資料分析所需要的統計(或機器)學習技術，下列何者有誤？
(A)統計(或機器)學習主要涵蓋有監督式、半監督式、非監督式等學習型態
(B)監督式學習主要提供資料分類(Classification)，而非監督式學習只是提供資料的分群(Clustering)
(C)隨機樹叢(Random Forest)演算法是整合多個決策樹(Decision Trees)分類的演算法
(D)迴歸分析和KMeans兩種演算法常被用來進行巨量資料的分類

47. 巨量資料分析的R程式語言具有多種不同資料結構型態(data types)的表達格式，下列何者有誤？
(A) R的資料結構型態表達格式主要包括：vector, matrix, array, data frame, list
(B) Matrix可以表示二維空間的資料結構，在matrix中所有元素必須是屬於相同資料型態
(C) Lists可以表示最複雜的資料結構型態，舉凡vectors, arrays, data frames甚至於lists本身的資料結構都可以被包含進來
(D) Data Frame的資料可以用data.frame()函式呼叫來產生，在data frame資料結構中屬於不同列資料其型態必須相同

48. 關於巨量資料分析系統Hadoop平台，下列何者有誤？
(A)是一個開放式系統軟體，具有可靠度、可擴充性與分散式處理的功能
(B) HDFS主要是提供分散式檔案系統功能的軟體模組
(C) MapReduce主要是在叢聚式電腦系統上，進行資料分析時提供有效的資源管理
(D) HBase是一個具有可擴充性與結構化資料庫特性的軟體模組

49. 關於巨量資料分析系統Hadoop平台，下列何者正確?
(A)可以提供Iterative演算法的有效計算
(B)可以透過MapReduce運算，提供各種進階式的結構化資料分析
(C)容錯功能是透過資料複製多份，並每一次的執行都寫入到硬碟的方式來完成
(D) Hadoop和MapReduce的整合性運作，可以有效完成圖論結構資料的巨量資料分析

50. 在Spark大數據分析平台上執行下列的Python程式碼：

```
file=spark.textFile("hdfs://…")  //opens a file
counts=file.flatMap(lambda line: line.split(" "))  //iterate over the lines, split each line by space into words
.map(lambda word: (word, 1))  //for each word, create the tuple (word, 1)
.reduceByKey(lambda a, b: a+b)  //go over the tuples "by key" (first element)
and sum  the second elements
counts.saveAsTextFile("hdfs://…")
```

當輸入短文： a spark can start a fire that burns the entire prairie
請問最後存入到 HDFS 檔案的(key, value)一共有多少對？
(A) 8 (B) 9 (C) 10 (D) 11