您好：

為了評估您與此職缺的相符程度，請嘗試完成以下題目，並於email中提及的截止日期前繳交下述兩份檔案。

**【繳交檔案】**

* **程式檔 (.R 或 .ipynb)**：按題目要求使用 R 或 Python 所撰寫完成的程式碼
* **資料檔 (.csv)**：按題目要求所產出的資料檔案
* **文件檔 (.docx)**：附上您對各題的解釋、說明或回覆

**【題目說明】**

* 若您認為有不明確的專有名詞，請自行定義。交卷前請確認繳交的程式碼的可執行性及結果的再現性

**【題目】**

近年來，台灣流行著一種內分泌疾病，盛行率約為30%人口。醫師依靠症狀評估作為臨床診斷，仍有一定誤診率，且每位醫師不同。若要100%確診，則需要進行精密的實驗室檢查。而您，作為一位傑出的算命師，聲稱僅需觀看民眾的照片即可診斷其是否罹患此疾病，且診斷準確率高達90%。然而，您的診斷能力遭受質疑與批評，為了證明自己，您計劃進行一項驗證研究以驗證您的診斷能力。請透過以下步驟展開您的驗證研究：

**<資料準備>**

1. 首先，您邀請了500位民眾參與研究，收集民眾基本資訊後，您建立了一個包含500列的資料集。請依照下表的定義建立四個欄位，並填入每個欄位的值。

|  |  |
| --- | --- |
| **欄位名稱** | **欄位定義** |
| **ID** | * 請為這500筆資料建立流水號，並以「Test\_」作為每一筆資料的編號開頭 * 每一筆資料代表一位民眾 |
| **Gender** | * 類別變項，包含「Female」和「Male」兩類 * 用隨機分配設定500筆資料的性別(Gender)，需符合「Female和Male各150筆」條件 |
| **Age** | * 連續變項，介於「0至100歲」之間的整數 * 用隨機分配設定500筆資料的年齡(Age)，需符合以下條件   + 男性：平均值為59、SD為10；女性：平均值為62、SD為12 |
| **Diagnosis** | * 類別變項，包含「Positive」和「Negative」兩類 * 用隨機分配設定500筆資料的最終診斷(Diagnosis)，需符合   + 盛行率約為30%人口 |

1. 您邀請了三位經驗豐富的專家醫師各自獨立對500位民眾進行評估與診斷，希望將其診斷結果作為比較基準。請於資料集中新增四個欄位，分別代表三位專科醫師(A、B、C)對這500位民眾進行臨床症狀所得到的臨床診斷結果，以及您對疾病的診斷。

|  |  |
| --- | --- |
| **欄位名稱** | **欄位定義** |
| **Doctor\_A** | * 類別變項，包含「Positive」和「Negative」兩類，為醫師A對疾病的診斷 * 用隨機分配設定500筆資料的醫師A之臨床診斷(Doctor\_A)，需符合75-80%準確率 |
| **Doctor\_B** | * 類別變項，包含「Positive」和「Negative」兩類，為醫師B對疾病的診斷 * 用隨機分配設定500筆資料的醫師B之臨床診斷(Doctor\_B)，需符合80-85%準確率 |
| **Doctor\_C** | * 類別變項，包含「Positive」和「Negative」兩類，為醫師C對疾病的診斷 * 用隨機分配設定500筆資料的醫師C之臨床診斷(Doctor\_C)，需符合85-90%準確率 |
| **FT\_Output** | * 類別變項，包含「Positive」和「Negative」兩類，為您對疾病的診斷 * 用隨機分配設定500筆資料的您的診斷，需符合90-92%準確率 |

**<統計分析>**

1. 請根據上述資料集，繪製不同性別(Gender)的年齡(Age)直方圖。
2. 您挑選了某種合適的統計方法，進行了醫師間臨床診斷一致性的比較分析。請提供您使用的統計方法、情境設定與假設、實際分析步驟以及分析結果的詳細解釋。
3. 您選用了合適的統計方法，進行了驗證分析，驗證您的診斷能力，是否如同您聲稱的「高達90%」。請提供您使用的統計方法、情境設定與假設、實際分析步驟以及分析結果的詳細解釋。
4. 您選用了合適的統計方法，進行了驗證分析，比較您的診斷能力與醫師相比如何？請提供您使用的統計方法、情境設定與假設、實際分析步驟以及分析結果的詳細解釋。
5. 最後，關於這個研究，您有什麼想要補充的嗎？

謝謝您，敬業地同時扮演了算命師及數據分析師的角色，我們期待收到您的回覆。