統計數據的檢驗

主要資料來源：《統計與真理》，C. R. Rao 著(美)，科學出版社，2004。

我們常常需要分析他人收集的數據，來獲得數據隱含的意義和用於解釋結果。進行分析前，了解數據來源是否正確當然十分重要，在進行數據正確性檢驗時，可採用以下的檢驗項目：

- 數據是如何收集、記錄的？
- 數據中含有測量誤差和記錄誤差嗎？
- 有關測量值的概念和定義明確嗎？
- 觀察值之間存在任何區別嗎？
- 數據的真實性如何？是原始資料還是經過調整、編纂或修飾過的？有無刪除過任何觀察值？
- 數據中有無存在異常值？對這些異常值是如何處理的？
- 母群體是什麼？
- 樣本中有多少部分(或全部)是沒有回應(答)的？
- 數據是來自單一母群體還是混合母群體？
- 抽出樣本中其單位的識別和分類等相關因素都有記錄嗎？
- 對所要調查的課題或觀察數據的性質，是否存在先驗訊息？

以上項目，有些通過直接與收集數據的調查人員交談即可獲知，其他的部份，可通過散佈圖、直方圖、坐標圖以及某些描述統計量的計算…等方法來檢測。

有時，一些研究結果的數據過於符合某些理論或出現特別的規律，反而是有問題的，這些數據有可能是編撰出來的。在任何大規模調查中，測量和記錄上出現誤差是不可避免的，可是如果這些值不是與其他值有顯著的不同時，要檢測出它們是很困難的。在調查中如果出現一個異常值時，帶有審查的程序就會向調查者發出警示，並容許調查者重複測量或確認被測量的個體是否屬於被研究的母群體。簡單的圖示，如直方圖和分佈圖也能夠幫助我們檢測。

政府常對統計數字非常熱心，並把數據累計作出漂亮的圖形。但永遠不要忘記：這些圖形是基於每一個來自基層的統計員，他們是可以偽造數據的。

接受一個新的理論，必須依賴對觀測數據的驗證，一個科學家有時也會被引誘去編造一些實驗數據來擬合理論，從而建立他自己的學術地位。雖然一個錯誤的理論遲早會被人發現，但在被發現之前已經造成了社會的危害。不要以

為知名科學家的數據就一定不會作假，如果一個實驗者十分了解統計學者，使用什麼樣的檢驗方式來檢測偽造的數據，那麼這個實驗者就可以知道如何去配合，來偽造數據，使些數據看起來無可懷疑且經得起檢測。

當一個科學家確信要建立他自己的理論時，便存在一種誘惑，使得他去尋找「事實」或歪曲事實以便擬合他自己的理論。就算當代的著名科學家孟德爾、牛頓、托勒密、伽利略、道爾頓、密立根……等人也曾被發現實驗數據上的問題。而且，這種利用偽造數據來建立錯誤假設結果的例子，到現在還是屢見不鮮。

電子計算機先驅英國科學家 C. Babbage，把某些科學家在處理數據和使用數據時採取的欺騙手法分為以下幾類：

- 修飾數據

  修剪那些與平均值有極大差異的觀察值，貼補那些看起來與平均值相比似乎太小的。

- 加工數據

  為了使普通的觀測值看起來正確而採用各種技巧。其中之一就是進行多次重複觀察，從中只選取那些一致的或非常接近一致的觀測值。

- 偽造數據

  並未做實際觀測，記錄的只是無中生有憑空捏造出來的數據。

那麼應該如何處理那些看起來的極端值呢？遺憾的是，除了上述的修飾或調整以外，至今仍沒有滿意的解決方法。當我們懷疑存在異常值時，應採取的科學方法是考慮下列的幾種可能：

- 再確認異常值是不是測量或記錄時發生顯著錯誤的結果。

- 確認與異常值有關的個體是否不屬於所研究的母群體，或者與樣本中其他部分有本質上的區別。

- 所研究母群體測量值的分佈其實是偏態的，出現較大或較小的值並非罕見。

處理懷疑為異常觀測值的第一步，就是驗證母群體有關的部分，也許可以找到合適的處置方法，偶爾，當再次觀察異常值時，會導致新的發現！當然，回到觀測原點的方法並不總是可行的，那就要依賴以下純統計學的檢驗方法去確定：

- 是否從研究對象的母群體中剔除異常觀測值，把剩下的部分作為有效樣本。

- 是否從研究對象的母群體中剔除異常觀測值，同時在統計分析的意義下做出相應的修正。

- 是否接受那些看起來似乎是異常觀測值是研究母群體中的正常現象，再利用合適的模型進行統計分析。

總之，目前還沒有適當的統計方法來處理上述問題，但統計學者正在從穩健推斷、檢出異常值和檢出有影響觀察值等各個方向進行這方面的工作，也許結合交叉數據檢驗，可以在推斷數據分析時提供一個統一的理論。