# 大數據和統計學之間的關係，你怎麼看？

2017-04-28 由 西線學院 發表于科技


普遍的定義認為，統計學是關於數據的科學，研究如何收集數據，併科學地推斷總體特徵。大數據和統計學還是存在一定區別的，其一是數據分析時不再進行抽樣，而是採用 population（n=all）；其二是分析方法，側重所有變量之間的相關性，而不再根據背景學科理論篩選變量，進行假設檢驗。

現在社會上有一種流行的說法，認為在大數據時代，「樣本=全體」，人們得到的不是抽樣數據而是全數據，因而只需要簡單地數一數就可以下結論了，複雜的統計學方法可以不再需要了。

普查和抽樣調查是傳統的兩大數據收集方法。普查不需要統計學方法進行推斷估計，因為通過普查，已經取得了所有個體數據和總體的實際分布，這也是為什麼人類開始懂得計數就開始進行普查。抽樣調查是利用抽樣理論解決如何科學設計樣本，取得樣本個體數據，併科學地推斷總體分布及特徵。無論是普查還是抽樣調查，其核心問題之一是要取得準確的「個體數據」。但在大數據時代，一切皆可量化，一切皆可記錄，如何利用更全面、更及時、更經濟的網絡電子化數據，以及通過對這些數據使用新的分

析及挖掘技術，產生新的見解和認識，是我們面臨的重大機遇。

大數據的應用可以說是在減少人類處理數據時帶入的主觀假設的影響，而完全依靠數據間的相關性來闡述。而由於消除人為因素帶入的誤差，已經分析人員作出假設的限制（如果教育背景和保險購買額是相關的，而分析人員沒想到，那這個結論就不會被分析出來，這在實際案例中是很容易發生的，大數據的核心也就在於它能更充分的發掘數據的全部真實含義。

在大數據時代，數據分析的很多根本性問題和小數據時代並沒有本質區別。當然，大數據的特點，確實對數據分析提出了全新挑戰。例如，許多傳統統計方法應用到大數據上，巨大計算量和存儲量往往使其難以承受；對結構複雜、來源多樣的數據，如何建立有效的統計學模型也需要新的探索和嘗試。對於新時代的數據科學而言，這些挑戰也同時意味著巨大的機遇，有可能會產生新的思想、方法和技術。

西線學院培訓機構提供良好的教學環境，良好的師資以及行業資源，使得西線學院教學永遠都是跟隨行業進步的步伐。說了這麼多，其實就是想讓你更加了解大數據。如此優秀的資源和別人望眼欲穿的實習機會，再不行動就要被後來居上的技術人員拍死在沙灘上了。

原文網址：https://kknews.cc/tech/nm86283.html