數學傳播　31卷3期, pp. 79-84

假設檢定的特點與一個特例
——兼談為什麼要用尾部檢定

徐澤泉 · 唐志華

1. 引言

近年來, 爲了適應資訊時代發展的需要, 國際上大多數國家都把概率統計知識納入到中學的數學教學。

例如, 中國大陸在新的高中數學課程標準裏, 已經“把最基本的資料處理、統計知識” 等作爲一種新的數學基礎和基本技能之一。

在高中數學必修3中, 統計約16課時。除了一些最基本的概念外, 還介紹了從樣本資料中提取基本的數量特徵 (如平均數、方差), 會用這些基本的數量特徵估計總體, 並能體會所得樣本資料和這些數量特徵的隨機性, 會用資料分析進行合理決策等等。

選修1-2中, 統計案例約14課時。已涉及到獨立性檢定, 並通過典型案例, 如“質量檢驗”、“新藥檢驗”等, 向學生介紹了實際推斷原理和假設檢定的基本思想和方法等。

選修2-3中, 統計與概率約22課時。在選修1-2統計案例的基礎上, 又加深了對這些基本思想和方法的應用, 典型案例也進一步增加。還特別指出了統計思想與方法的特點是, “統計推論可能會犯錯誤”和“估計結果的隨機性”等。

選修4-9 的風險與決策, 只有內容要求、說明與建議, 沒有具體課時安排。要求中學生能夠掌握一些簡單的統計決策方面的知識和方法, 形成初步的決策意識。

此外, 在選修系列3-1的數學史選講中還專門有一節“近代統計學的緣起”。

這一系列統計知識的安排, 似乎已經把現代統計的基本思想和方法簡明扼要地全都融入到了中學數學課程之中。它是進一步提高中學生的數學修養, 打好數學基礎, 提高數學的現代應用意識的的一種新的安排。作者認爲, 要使學生能夠初步掌握這些最基本也是最簡單的統計知識與方法, 就要使他們在掌握這些知識與方法的過程中, 同時瞭解與領會推論統計學的基本思想和特點。數理統計學是不確定性數學隨機數學的一個分支學科, 有它獨特的思考問題的方法和

推理過程。儘管中學數學教學只要求學生瞭解幾種統計方法的基本思想及其初步應用, 而對其理論基礎不作要求; 但是, 只有當學生瞭解與體驗到這些統計思想與方法的獨特的思維過程後, 才能有效地改變他們只會單純記憶和機械套用公式的習慣。這裏的關鍵在教師。師生們普遍認爲在統計知識中, 最難於理解和掌握的要算是假設檢定的基本思想與方法了。因此, 作爲教者來講, 如對其中的一些理論基礎不甚了了, 就很難勝任這一教學任務。本文打算就這一問題通過一個實例作一些扼要的討論。

2. 統計推論與假設檢定的基本思想與概念

數理統計方法不同於一般的資料統計, 它更側重於應用隨機現象本身的規律性來考慮資料的收集、整理與分析, 從而找出隨機變數的分佈規律或它的數量特徵。它所關心的乃是某些規定的總體或集合, 是以掌握事物總體的數量特徵爲目標的。這就需要用統計方法對總體進行推論。而假設檢定又是統計推論的基本問題之一。

概率論中有一條基本原理, 那就是"小概率事件實際上的不可能性原理"。它是假設檢定用它進行統計推論的一個理論基礎。所謂小概率事件, 一般說來是不可能出現的。正是由於人們相信和承認"小概率事件的不可能性原理", 才能大膽地進行工作、學習, 遊玩與休息。不然, 誰還敢出門, 誰還敢乘車? 當然話又說回來, 就是不出門也不能保證絕對安全。

假設檢定一般都以第二種可能性作爲出發點, 設立原假設 (也稱虛無假設), 它的對立面就是對立假設 (或備擇假設也叫研究假設等)。坦率地說, 原假設就是我們要達到它的對立假設, 而假定它的出現是由於隨機因素或偶然因素造成的, 準備通過統計試驗中出現的小概率事件來否定它, 從而說明我們所作的努力是有成效的。假設檢定的基本思想是簡單的。它是使用帶有概率性質的反證法, 間接運用"小概率事件實際上的不可能性"原理。

那麼, 何爲小概率事件呢? 當然, 這僅是相對而論。這要看事件本身的重要性程度而定。比如, 在一千支青黴素藥品中有一支是廢品, 這就不能算作是小概率事件, 但一千顆普通紐扣中有一粒是次品的話, 那無關緊要。一般地, 人們經常把概率小於等於0.01、0.05等的事件作爲小概率事件。

這些小概率事件都對應著分佈曲線下面積的比例問題, 而正態分佈又是概率論中最重要的一種分佈形態, 它在實際應用和理論研究中佔有頭等重要的地位。一般說來, 若影響某一數量指標的隨機因素很多, 而每一個因素所起的作用不太大, 則這個指標服從正態分佈。正態曲線的密度函數爲

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

其中, μ、σ 就是我們所熟知的平均數和標準差, σ^2 爲方差。如果參數 μ 和 σ 確定了, 那麼該密度函數從而它的分佈形態也就確定了, 故正態分佈一般可記作 $N(\mu, \sigma^2)$, 本文所涉及的假設檢

定都是關於正態分佈下的參數假設檢定, 且限定爲單參數的非此即彼的簡單假設檢定。圓括號中的 $\frac{X-\mu}{\sigma}$ 稱之爲標準分數, 如果用字母 z 代替它, 方程式便化爲:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

它是應用中及其方便的標準正態曲線, 標準正態分佈就是以平均數爲0和標準差爲1的分佈, 即 $N(0,1)$ 分佈, 它的示意圖如下:

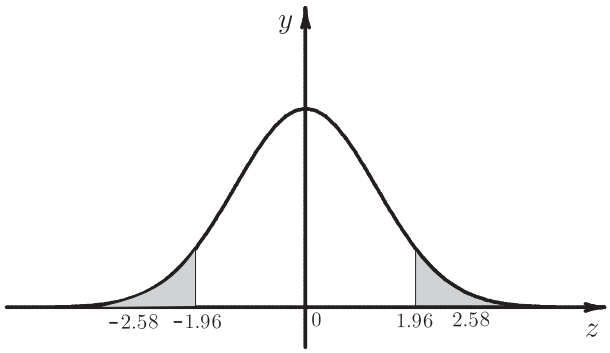

圖1

可以證明, 它與 z 軸所圍成的面積即其概率積分之值 $\mathcal{P} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = 1$。當 $|z| \geq 1.96$ 時, 其中雙尾部分的面積 $p = 0.05$, 我們稱之爲 p 值小於等於0.05; $|z| > 2.58$ 時, p 值等於 0.01。如果從實得的樣本統計量計算所得的 p 值落入該區間內, 我們就說小概率事件發生了。習慣上, 人們把0.05和0.01等 p 值用字母 α 來表示, 並把它稱爲檢定的顯著性水平。當 $\alpha = 0.05$ 時, 我們說差異顯著; 當 $\alpha = 0.01$ 時, 我們說差異非常顯著或極其顯著。餘類推。與之對應的數 $z = 1.96$ 和 2.58 等數值叫做 α 的臨界值。在進行檢定之前就應當選擇好顯著性水平, 即 α 的值。檢定結果所要求的顯著性水平越高, α 值就越小, 這時需要超過其臨界值的實得統計量 z 值就越大。

假設檢定的基本步驟如下:

(1) 提出假設;

(2) 選擇並計算檢定統計量;

(3) 確定檢定形式;

(4) 統計決斷。

下面我們用一個實例來說明之, 並引入人們習慣上所使用的數學運算式與符號。自上個世紀80年代以來, 江蘇省無錫市和南京市的許多學校都普遍開展了“現代班級集體建設的理論與實踐” (注, 所謂現代班級集體建設就是指, 要使人類社會自200多年以來的班級授課制能跟上

現代社會與科學技術的發展) 教育科學實驗。其中有一個項目是關於現代班級集體建設中的“結構要素”的研究與探索。所謂結構要素主要是指班級的教學管理和生活管理中的目標管理、人際關係、輿論建設、組織形式 (如小組教學個別教學) 等等。

實例: 某市現代班級集體建設結構要素科學實驗中, 測得該市某年在班級集體結構要素之一的“輿論建設”上總體發展水平的平均值 $\mu_0 = 28.98$ (滿分爲40分), 標準差爲 $\sigma_0 = 1.72$。在同一次抽樣調查中, 測得該市第016號班級集體在輿論建設上的平均值是 $\overline{x} = 33.63$, 假定方差不變。試問該班在班級集體輿論建設上的發展水平與市平均水平有無顯著性差異?

把它化爲統計推論問題就是: 假如在方差 σ_0^2 不變的情況下, 問我們樣本平均數 $\overline{x}$ 是否來自於同一正態總體 $N(28.98, 1.72^2)$, 還是不同的正態總體 $N(\mu_1, 1.72^2)$。這時需要我們進行差異的顯著性檢驗。步驟如下:

(1) 提出假設 如前所述, 我們可認爲樣本圍繞總體成正態分佈。

$$H_0 : \overline{x} \sim N(\mu_0, \sigma_0^2) \quad \mu_0 = 28.98 \quad \text{(表示該班與市總體平均水平無差異)}$$

$$H_1 : \overline{x} \sim N(\mu_1, \sigma_0^2) \quad \mu_1 \neq 28.98 \quad \text{(表示該班與市平均水平有顯著差異)}$$

(2) 選擇並計算檢定統計量

因爲是正態分佈, 故選擇 z 檢定統計量

$$z = \frac{|\overline{x} - \mu_0|}{\sigma} = \frac{33.63 - 28.98}{1.72} = 2.7$$

(3) 確定檢定形式

順便指出, z 檢定有單尾檢定與雙尾檢定兩種形式。如果暫沒有足夠資料說明該班在輿論水平上所取得的分值顯著高於市平均水平的話, 一般說來, 用雙尾檢定; 如以往的樣本中可以估計會優於市平均的話, 也可使用單尾檢定。這裏我們用雙尾檢定。

(4) 統計決斷

選擇顯著性水平 $\alpha = 0.01$, 查得相應的臨界值

$$|Z|_{p(2).01} = 2.58$$

其中足標 $p(2).01$ 表示對 z 的雙尾檢定, 顯著性水平 $\alpha = 0.01$。而實得的 $z = 2.70$, 且

$$z = 2.70 > |Z|_{p(2).01} = 2.58$$

故 p 值小於 $\alpha = 0.01$。實際上, 這裏的 p 值就是隨機樣本落入正態曲線兩尾部之概率。在 $\alpha = 0.01$ 的水平上就相當於我們把概率小於等於0.01的事件作爲小概率事件。那麼, 這一事件在100次試驗中平均最多出現一次, 而現在竟然在一次抽樣中就出現了 (小概率事件)。這

一事實是由於不合理的原假設 (虛無假設) 造成的。所以, 我們有99%的信心和把握來否定原假設; 反之, 若實際計算的 值小於其臨界值2.58, 則不能否定原假設。

(5) 結論: 該市第016號班集體在輿論建設上的發展水平高於市平均且差異非常顯著。

當然, 這僅僅是一項輿論指標在反映班集體發展水平上的一個參考。

這裏我們使用的是雙尾檢定, 當然也可使用單尾檢定的方法。但需要說明的是, 在同樣的信心水平 $1-\alpha$ 下, 雙尾檢定的顯著性要求高於單尾檢定的顯著性要求。比如在某項檢定中, 要把0.05的概率值置於曲線的兩尾, 則每尾各占0.025, 對應的臨界值 $z \approx \pm 1.96$; 而使用單尾檢定的話只要把0.05的 p 值置於曲線的一尾, 可查獲其相應的臨界值 $z \approx \pm 1.64$。換言之, 採用雙尾檢定時統計量的臨界值的絕對值要大於採用單尾檢定的值, 故實際計算得來的 z 值越過它的可能性就小, 就比較不容易否定原假設, 相對說來就不易獲得顯著性差異的結論。

3. 兩類錯誤與檢驗的優化

由上可知, 當 $|z| \geq 1.96$ 時, 正態曲線下雙尾部分的面積 $p = 0.05$, $|z| \geq 2.58$ 時, p 值等於 0.01 這樣, 在顯著性水平 α 被選定的前提下, 無論是雙尾檢定還是單尾檢定, 其臨界值就把標準正態曲線下的面積一分爲二 (如圖1)。我們把區域 (或集合) $W = \{z : |z| \geq 1.96\}$、$\{z : |z| \geq 2.58\}$ 叫做拒絕域 (或否定域); 而把它們的補集 $\overline{W} = \{z : -1.96 \leq z \leq 1.96\}$、$\{z : -2.58 \leq z \leq 2.58\}$, 稱爲接受域。

然而, 由於假設檢定是根據一次抽樣所得樣本統計量的值而作出的判斷, 這樣判斷的結果有可能發生錯誤。本來成立的原假設由於實得的 z 值落入否定域 W 而錯誤地被拒絕; 也可能是本來並不成立的原假設由於實得的 z 值落入接受域 $\overline{W}$ 而錯誤地被接受。通常人們把原假設爲眞而被拒絕的錯誤稱爲第一類型錯誤 (或簡稱棄眞錯誤), 而當原假設不眞時反倒被接受的錯誤稱之爲第二類型錯誤 (或簡稱取僞錯誤)。

因爲抽樣的隨機性, 雖說犯上述錯誤在所難免, 但卻具有確定的概率。從上可以看到, 犯棄眞錯誤的概率恰好就等於顯著性水平 α。然而要減少犯第一類錯誤的概率 α, 就會引起接受域的擴大, 愈容易接受原假設而隨之犯第二類型錯誤的概率也愈大。當然, 它們的值與抽樣的大小和檢定的參數有關。從下文的圖2可以看到, 如果以 μ_0 爲原假設、μ_2 或 μ_1 爲它的對立假設, 那麼 μ_2 所犯第二類錯誤的概率較小 (參見圖 2-1)。因此, 在不增加樣本容量和參數不變的情況下, 要同時減小這兩類錯誤的概率幾乎是不可能的。國際統計學家的先軀們, 如美國的 J. 奈曼和英國的 E.S 皮爾遜於20世紀三、四十年代建立了一套優化檢定的理論, 該理論提出先控制住犯第一類型錯誤的概率 α, 使 α 不大於某個值。在此基礎上選取一個檢驗法 ϕ, 使得犯第二類錯誤的概率即取僞的概率盡可能地小。對於每一個這樣的 α 的值, 假設檢驗的拒絕域 W 也隨之而確定。按這一準則建立的檢定叫做 α-水平顯著性檢定。若設棄眞概率爲 α, 取僞概率爲

β, 則顯然 $\alpha = \phi_W(z)$, 其中 $z \sim N(\mu, \sigma^2)$; $\beta = \phi_{\overline{W}}(z)$, 其中 $z \sim N(\mu_1, \sigma^2)$, $\mu_1 \neq \mu$。特別需要指出的是, 這裏 $\beta \neq 1 - \alpha$ 只有當 $\mu_1 = \mu$ 時才有 $\beta = 1 - \alpha$ 成立。但我們有下列關係式:

$\phi_W(z) = 1 - \phi_{\overline{W}}(z) = 1 - \beta$, 當 $z \sim N(\mu_1, \sigma^2)$ 時。

我們把由否定域所確定的概率函數 $\phi_W(z)$ 稱之爲 α-水平檢驗的功效函數, 當 $z \sim N(\mu, \sigma^2)$ 時, 它就是犯第一類錯誤之概率, 即棄眞概率; 當 $z \sim N(\mu_1, \sigma^2)$ 時, 它恰好是去僞之概率。此時一個好的檢驗法 ϕ_W 就是當 $\phi_W(z) \leq \alpha$ 時, 能使取僞概率 $\beta = \phi_{\overline{W}}(z)$ 最小。或者說功效函數 $\phi_W(z) = 1 - \beta$ 最大。

如果有兩個檢驗法 $\phi^i_W(z)$, $i = 1, 2$, 在滿足 $\phi^i_W(z) \leq \alpha$, $z \sim (\mu, \sigma^2)$ 的條件下, 若 $\phi^1_W(z) > \phi^2_W(z)$, $z \sim N(\mu_1, \sigma^2)$ 成立, 則我們就說 ϕ^1_W 比 ϕ^2_W 有效。

由上可知, 我們就知道假設檢定中爲什麼要用正態曲線下的雙尾檢定或單尾檢驗了。下圖2-1中兩個燕尾否定域所示面積在理論上完全可以和概率密度曲線下其他任何特定區域的面積(如圖2-2中所示之條形域) 相等, 這時意味著所犯棄眞錯誤的概率相等。但它們所犯第二類錯誤的概率即取僞錯誤的概率卻明顯不等 (左邊小於右邊, 讀者可從圖 2 中仔細辨認), 從而說明採用尾部檢定 (這裏是雙尾檢定) 法要比採用非尾部檢定法 (如條型的情形) 有效。

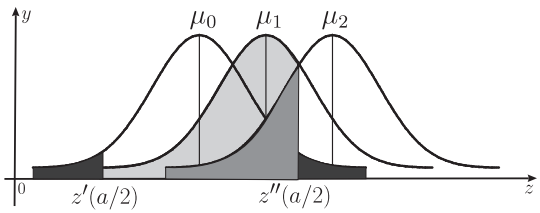

圖2-1. 拒絕域爲雙尾時第二類錯誤概率

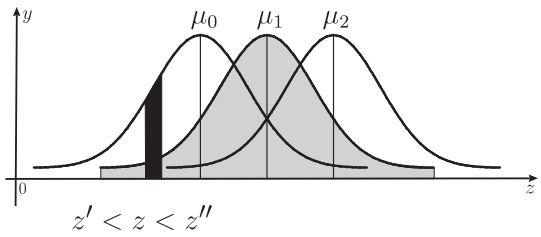

圖 2-2. 拒絕域爲條型時第二類錯誤概率

—本文作者徐瀝泉任職於江蘇省無錫市教育研究中心、唐志華系江蘇省南京幼兒高等師範學校校長—