指考排名計算

在大學聯考的時代，大眾一般依照各科系的最低錄取總分作為科系排名的依據。然而在指考的框架下，每個科系以選擇採計科目與各科加權的比重，因此公布的最低錄取總分無法直接比較。要解決這個問題，一個簡單的做法是把錄取總分除以權重總和，得到科目的加權平均分數。每個科系都這樣計算，就可以用這個加權平均分數排名。舉例而言，2018 年交大電機的權重為國文 x1.00，英文 x1.50，數甲 x2.00，物理 x2.00，化學 x1.00。錄取分數為 640.75，因此加權平均分數為 640.75 / 7.5 = 85.43。

這個做法主要的問題是各科系的平均分數可能經由不同權重調整，因此相互比較的可能有失公平。為了解決這個問題，我們發展了另一套調整權重的方式，可以讓我們更公平的比較各科系的最低錄取分數。這個調整方式可以在給定某科系考科權重與總分的前提下，估計某科系指定考科權重為 1 時的錄取總分。

這是一個條件期望值的估計問題。在假設各科目分數的分配為多變數常態的前提下，我們可以導出條件期望值的明解 (Closed-form Solution)。

具體而言，我們可以不失一般性的假設某年指考共有十考科。考生某甲的分數為$X = [x_1, x_2, \dots, x_{10}]^T$。令某科系$d$的加權向量為$a_d$，那考生某甲的總分為$a_d^T X$。我們想要知道的是，如果這個考生的加權向量不是$a_d$，而是另一個加權向量$a_0$，那這個新的總分$a_0^T X$應該是多少?

至於加權向量a_0的設值，我們將所有科系依據採計科目分為第一類組、第二類組與第三類組，將第一類組科系的國文、英文、數乙、歷史、地理、公民，第二類組科系的國文、英文、數甲、物理、化學，以及第三類組科系的國文、英文、數甲、物理、化學、生物科目權重設為 1，其餘設為 0。

我們假設這十個考科的分數服從多變數常態分配，均數向量為$\mu = \left[\mu_1, \mu_2, \dots, \mu_{10}\right]^T$，共變異數矩陣為$\Sigma$。那我們關心的就是條件機率分布$\mathrm{P}(\mathrm{a}_0^{\mathrm{T}} X \middle| a_d^T X)$的均數。

由於常態隨機變數的線性組合一樣也服從常態分布，因此我們知道$\mathrm{a}_0^{\mathrm{T}} X$與$a_d^T X$服從雙變數常態分布，均數為

$$[\mathrm{a}_0^{\mathrm{T}} \mu, a_d^T \mu]$$

變異數矩陣為

$$\begin{bmatrix} a_0^T \Sigma a_0 & a_0^T \Sigma a_d \\ a_d^T \Sigma a_0 & a_d^T \Sigma a_d \end{bmatrix}$$

由$\mathrm{P}(\mathrm{a}_0^{\mathrm{T}}X \big| a_d^T X) = \frac{p(a_0^T \mathrm{X}, a_d^T \mathrm{X})}{p(a_d^T \mathrm{X})}$，經過推導與化簡之後可以得到條件期望值的明解:

$$\mathrm{E}[\mathrm{a}_0^{\mathrm{T}}X | a_d^T X] = \mathrm{a}_0^{\mathrm{T}}\mu + \frac{a_0^T \Sigma a_d}{a_d^T \Sigma a_d}(a_d^T X - a_d^T \mu) \quad \text{....} \quad \text{公式(*)}$$

熟悉迴歸模型的人看到這個明解應該會很親切。

這裡雖然已經有公式(*)，但卻無法直接應用。原因是各科分數的共變異數矩陣Σ未知。雖然大考中心有公佈詳細的各科分數的累積機率分配的資訊，卻沒有公佈科目分數的共變異數矩陣。因此如果要能使用這個公式，必須要先由公開資訊估計共變異數矩陣。

估計共變異數矩陣看似不可能的任務。然而，大學考試入學分發委員會有公佈每個考科組合總分的累積機率分配，例如數乙、歷史、地理的未加權總分人數累計表，以及數甲、物理、化學的未加權的總分人數累計表等。以2019 年為例，扣除有採計音樂、體育、美術的組合之外，共有 58 個考科組合可以做為後續分析使用。我們的方法可以用這些考科組合資訊反推需要的共變異數矩陣。

這個做法的原理是利用隨機變數和的動差關係。為了方便說明，考慮某考科組合$\mathrm{Y} = \mathrm{X}_1 + X_2 + X_3$，則

$$Var(Y) = \mathrm{Var}(\mathrm{X}_1) + Var(X_2) + Var(X_3) + 2(\sigma_{1,2} + \sigma_{1,3} + \sigma_{2,3})$$

其中Var(Y)可以由未加權的各組合總分人數累計表算出，$\mathrm{Var}(\mathrm{X}_1)$、$\mathrm{Var}(\mathrm{X}_2)$ 、$\mathrm{Var}(\mathrm{X}_3)$可由各科的人數累計表算出。如果我們將上式整理一下，可以得到

$$\frac{Var(Y) - \mathrm{Var}(\mathrm{X}_1) - Var(X_2) - Var(X_3)}{2} = \sigma_{1,2} + \sigma_{1,3} + \sigma_{2,3}$$

其中左邊是已知，右邊是未知的數值。將所有能找到的考科組合累計次數表依照上述的方法處理，就可以用來估計未知的相關係數。

這個做法有幾個重要的細節。第一，有些相關係數實際上沒有資料可以估計，如物理與歷史。我們將這些相關係數直接設為 0。第二，扣除直接設為 0 的相關係數之後，共有 29 個相關係數需要估計，而考科組合的數量一般大於需要估計的相關係數總數。以 2019 年為例，我們共有 58 個考科組合，但只需估計 29 個未知數。我們的解決方法也很直觀，就是找一組相關係數讓配適誤差最小。這件事可以很方便的使用現成的迴歸函數來執行。第三，有一些考科與組合因為缺考的關係，左尾有異常高的頻率，我們會先將這些離群值去除。

以這個方法求出的 2019 年指考科目的相關係數矩陣如下:

相關係數矩陣	國文	英文	數甲	數乙	歷史	地理	公民	物理	化學	生物
國文	1	0.569	0.410	0.589	0.860	0.778	0.921	0.393	0.479	0.563
英文		1	0.640	0.619	0.637	0.654	0.644	0.619	0.666	0.810
數甲			1	0	0	0	0	0.761	0.760	0.901
數乙				1	0.547	0.575	0.556	0	0	0
歷史					1	0.830	0.869	0	0	0
地理						1	0.827	0	0	0
公民							1	0	0	0
物理								1	0.854	0.951
化學									1	0.973
生物										1

由於考科的相關係數沒有公開資料，我們並沒有辦法直接驗證這個做法的正確性，只能以直觀分析結果。所有考科相關係數最高的是生物與化學，相關係數高達 0.973，接下來是生物與物理，有 0.951。生物與數甲的相關係數也高達 0.901。相關係數最低的組合是物理與國文，只有 0.393。第二低的是國文與數甲，為 0.410。化學與國文也是低檔(0.479)。估計出來的結果與直觀還算是相符合。

求得相關係數矩陣後就能反推共變異數矩陣並帶入公式(*)，得到條件期望值的明解，也就是某科系考科指定權重為 1 時的估計錄取總分。每個系所都得到經過調整的錄取總分後，即可根據計算排名。