相似度41%夠大嗎？

黃文璋
國立高雄大學應用數學系

1 前言

“美女上錯身”(Drop Dead Diva), 是美國一部有關律師事務所的電視影集。在某一集裡, 有人委請律師, 對抄襲他的書出版者提出控告。上了法庭, 被告律師答辯, 依反抄襲軟體, 兩書相似度僅41%, 以此佐證他的委託人並未抄襲, 或者說就算有抄襲, 也不嚴重。法官支持被告, 同意41%不夠高。原告及其律師當然都很不服氣, 明明是抄襲, 怎麼被視為像是如有雷同, 純屬巧合呢?

再看一則新聞的標題“女子整容酷似楊冪難分辨 相似度高達95%”。楊冪(1986-)是位有相當名氣的大陸演員。有人心儀其容貌, 經醫師妙手整容, 化腐朽為神奇, 與楊冪的相似度達到95%, 差不多是一個模樣了。比起95%, 41%的確太低了, 難怪被認為未達抄襲程度。

究竟41%是否算低, 95%是否算高, 其實並不能由表面上的數字大小來判定。如果影印機, 印出來與原件相似度為95%, 這種影印機品質並不算好。相機照出來, 與本人相似度, 即使高達99%, 也會有人不滿意, 覺得差了一點。但補習班考前猜題, 若命中率能有41%, 就很驚人。所以相似度高或低, 應由發生的難易程度, 即依發生機率的大小而定。

我們不知反抄襲軟體如何判定相似度, 其中的機率模型為何也毫無概

念。但因有某統計學家關切此議題, 遂擬以幾個能建立機率模型的例子, 來略討論相似度的高低。

2 $0,1$數列

假設有一長達100個的$0,1$數列, 由你依序猜。如果你沒有特異功能, 只是隨機地猜, 即每次猜對的機率為$1/2$。則全都對的機率為$(1/2)^{100}$, 且平均可對50個。我們以對的個數除以100, 當做吻合度。則吻合度50%以上, 並不稀奇。由於對的個數, 有參數100及$1/2$的二項分佈, 即$\mathcal{B}(100,1/2)$, 期望值是50, 標準差是5, 以常態分佈來近似, 將得對的個數比期望值至少多1個標準差, 即吻合度55%以上, 機率約為0.1587; 對的個數比期望值至少多2個標準差, 即吻合度60%以上, 機率約為0.0228; 至於吻合度65%(比期望值多3個標準差)以上, 機率約為0.0013。

二項分佈以常態分佈來近似, 若採連續性的更正(continuity correction), 將較精確些, 但差異並非太大。為了簡便, 此處就不採用了。對$0,1$數列, 可看出即使吻合度60%以上, 都已相當不容易了, 因機率才比百分之2稍大些。吻合度若有65%以上, 更是神乎其技, 因機率僅約千分之1.3。至於若吻合度達到95%以上(超過9個標準差, 機率約為$1.1422\cdot 10^{-19}$), 恐怕將沒人會相信你真的是隨機地猜。

對參數n,p的二項分佈$\mathcal{B}(n,p)$, 期望值為np, 標準差為$\sqrt{np(1-p)}$。現假設$0,1$數列長達10,000個, 你依序猜。如果是隨機猜, 則對的個數有$\mathcal{B}(10{,}000,1/2)$分佈, 期望值是5,000, 標準差是50。諸位看, 數列長度增為100倍, 期望值亦隨之增為100倍, 但標準差僅增為10倍。如此一來, 吻合度比期望值分別至少多1個標準差, 2個標準差, 及3個標準差, 即吻合度達50.5%, 51%, 及51.5%以上, 其機率分別約為0.1587, 0.0228, 及0.0013。在此情況下, 吻合度要有55%以上, 幾乎是萬萬不可能(超過10個標準差, 機率約為$7.6946\cdot 10^{-24}$)。

$0,1$數列長度若是100, 吻合度55%以上不太難, 因機率約0.1587, 即有超過7分之1的機會。但數列長度若增至10,000, 吻合度便極雖達

到55%。一般若數列長度增為k^2倍, 標準差將僅增為k倍。可以這麼想, 若數列長度只有1, 則要全對的機率為1/2, 輕而易舉; 若數列長度為2, 則全對的機率降為1/4, 餘此類推, 愈長愈難全對。所以即使皆為0, 1數列, 對同樣的吻合度, 還要看數列的長短, 才能判定難易程度。數列愈長便愈難。

3 英文字母數列

現將0, 1數列, 改為26個英文字母數列, 且不去考慮大小寫的問題。假設隨機猜, 則當數列長度為n, 對的個數有$\mathcal{B}(n, 1/26)$分佈, 期望值$= n/26$, 標準差$= \sqrt{n(1/26)(25/26)} = 5\sqrt{n}/26$。若$n = 10{,}000$, 則期望值約為384.615, 標準差約為19.231。因此吻合度比期望值分別至少多1個標準差, 2個標準差, 及3個標準差, 即吻合度4.038%, 4.231%, 4.423%以上, 其機率分別約為0.1587, 0.0228, 及0.0013。可看出此時吻合度能有4.4%就很稀奇了, 因機率不過千分之1.3左右。至於想要吻合度達41%以上, 那眞是天方夜譚。

4 結語

由前述0, 1數列, 及英文字母數列的兩個例子顯示, 在不同情況下, 吻合度的難易, 並無法一起評比。至於書籍如何斷定其相似度, 我們雖不甚了了, 但應非僅簡單地看單字的吻合程度。前述例子顯示, 至少與書的長短必然有關。無論如何, 就是不能僅由表面上41%, 比統計裡信賴區間動不動就是90%, 95%低很多, 就覺得41%的相似度極低, 判定無抄襲之虞。