淺論資料品質檢核與假設檢定

朱昌綸／金融聯合徵信中心資訊部

統計學有一個很有趣的理論，稱爲「虛無假設」（Null Hypothesis），應用於假設檢定或統計顯著性檢定，簡言之，就是當認爲某件事情是對的，卻找不到方法證實時，就先假設這件事情是錯的，再找方法證實它是錯的，當最後仍無法證實這件事情是錯的，這件事情便是對的；雖然無法證明是錯的事情，並不代表一定就是對的事情，但是這種換個角度思考事情的方法，總比假設條件先憑空臆測，後又拿捏失準可靠多了。本文僅就資料品質檢核中，對於「虛無假設」的檢定，綜合引述說明。

假設檢定的四種判斷結果

所謂「假設檢定」，乃是依據有限的樣本數量，來推測母體的實際分佈，其正確判斷的型態有二類（參表一）。

1. specificity（特異性）

〔接受假設丨假設為真〕，A ／（A ＋ B），也就是說，當假設它是真實的，但經過研究亦判定它是正確的而接受它，這種判斷也稱為 True Positive。

2. sensitivity（敏感度）

〔拒絕假設丨假設為偽〕，D ／（C ＋ D），也就是說，當假設它是錯誤的，但經過研究亦判定它是錯誤的而拒絕它，這種判斷也稱為 True Negative。

然而，假設檢定之錯誤判斷的風險亦有二類（參表一）。

1. Type 1 error

〔拒絕假設丨假設為真〕，B ／（A ＋ B），也就是說，當假設它是真實的，但經過研究卻判定它是錯誤的而拒絕它，這種誤判也稱為 α Error 或 False Positive。

2. Type 2 error

〔接受假設丨假設為偽〕，C ／（C ＋ D），也就是說，當假設它是錯誤的，但經過研究卻判定它是正確的而接受它，這種誤判也稱為 β Error 或 False Negative。

表一

		研究判斷	
		接受假設	拒絕假設
真實情況	假設真實	正確判斷 特異性 A	錯誤判斷 type 1 error B
	假設錯誤	錯誤判斷 type 2 error C	正確判斷 敏感度 D

資料品質檢核之假設檢定

金融機構報送至聯徵中心的資料，必須經過資料品質的檢核；舉例而言，探討某一筆報送資料有否正確？有人認為：資料是正確的，但也有人認為：資料是錯誤的；依照虛無假設：資料是正確的。

1. 如果真實情況係假設為真，即資料是正確的；但檢核的結果，卻判定該資料不符檢核邏輯，即拒絕虛無假設，那麼，這個錯誤的檢核可能會將正確的資料視為錯誤，剔退於信用資料庫外，這種誤判，就是type 1 error。
2. 如果真實情況係假設為偽，即資料是錯誤的；但檢核的結果，卻判定該資料符合檢核邏輯，即接受虛無假設，那麼，這個錯誤的檢核可能會將錯誤的資料視為正確，建置於信用資料庫內，這種誤判，就是type 2 error。

聯徵中心將資料檢核作業區分為二階段，第一階段檢核作業稱為「資料實體檢核」，即資料建入資料庫前先行的檢核作業，為了避免將正確的資料視為錯誤而遭剔退，資料實體檢核的邏輯設計必須儘量降低type 1 error發生；第二階段檢核作業稱為「資料邏輯檢核」，即資料建入資料庫後再行的檢核作業，為了避免將錯誤的資料視為正確而遭誤植，資料邏輯檢核的邏輯設計必須儘量降低type 2 error發生。

高敏感度議題之假設檢定

所謂「敏感度高」，係指普遍性認為兩造之間存有關聯，個案的發生絕非偶然，即真實情況係假設為偽，拒絕虛無假設；舉例而言，探討授信戶「呆帳」與「學歷」有無關聯？大多數人認為：學歷低者因知識及專業能力不足，較易發生無法還款之情事，但也有少數人認為：這純粹是巧合，被認列為呆帳的授信戶恰巧學歷較低；依照虛無假設：這只是巧合的現象，沒有相互關聯。

1. 如果真實情況係假設為真，即學歷低者發生呆帳只是巧合的現象，沒有相互關聯；但檢核的結果，卻判定這現象並非只是巧合的現象，即拒絕虛無假設，那麼，這個錯誤的檢核可能會無中生有、創造出一堆歧視性的理論學說，這種誤判，就是type 1 error。
2. 如果真實情況係假設為偽，即學歷低者發生呆帳不單是巧合的現象；但檢核的結果，卻判定這現象只是巧合的現象，即接受虛無假設，那麼，這個錯誤的檢核可能會造成真理被忽視、真實的社會現象沒被發現，這種誤判，就是type 2 error。

上述案例是屬於社會行為的探討，type 1 error相較於type 2 error所造成的後果，是比較不符合科學精神的，因此，必須儘量降低type 1 error。即對於社會行為的探討，要儘量相信這只是巧合的現象而已，換句話說，儘量接受虛無假設，寧願錯失發現真理的機會，也不要創造出無中生有的理論；申言之，資料建入資料庫前先行檢核時（實體檢核），為了避免將正確的資料視為錯誤，應儘量接受虛無假設，寧願將品質可疑之資料暫時建置，也不要創造出自以為是的檢核邏輯，將資料逕自剔退。

高特異性議題之假設檢定

所謂「特異性高」，係指普遍性認為兩造

之間毫無關聯，個案的發生純屬偶然，即真實情況係假設為真，接受虛無假設；舉例而言，探討信用卡戶「上月為強制停卡」與「本月為一般停卡」有無關聯？大多數人認為：這純粹是巧合，信用卡機構恰巧因一般性作業疏失，旋即更正資料罷了，並無與卡戶交換利益之情事，但也有少數人認為：信用卡機構利用信用資訊做為稽催工具，待卡戶欠款繳清後，再將其不良紀錄漂白；依照虛無假設：這只是巧合的現象，沒有相互關聯。

1. 如果真實情況係假設為真，即上月為異常而本月為正常只是巧合的現象，沒有相互關聯；但檢核的結果，卻判定這現象並非只是巧合的現象，即拒絕虛無假設，那麼，這個錯誤的檢核可能會造成某些信用卡機構的作業疏失行為，卻被當成惡意企圖而受冤枉，這種誤判，就是type 1 error。
2. 如果真實情況係假設為偽，即上月為異常而本月為正常不單是巧合的現象；但檢核的結果，卻判定這現象只是巧合的現象，即接受虛無假設，那麼，這個錯誤的檢核可能會造成某些信用卡機構惡意利用信用資訊做為稽催工具，卻得以免被除追究責任，進而導致聯徵中心信用資料庫不受金融機構信任，這種誤判，就是type 2 error。

上述案例是屬於公共紀律的維持，type 2 error相較於type 1 error所造成的後果，是難以承擔的，因此，必須儘量降低type 2 error。即對於公共紀律的維持，要儘量拒絕相信這只是巧合的現象而已，換句話說，儘量拒絕接受虛無假設，寧願冤枉無辜，也不要縱放惡人；申言之，資料建入資料庫後再行檢核時（邏輯檢核），為了避免將錯誤的資料視為正確，應儘量拒絕接受虛無假設，寧願將品質可疑之資料要求報送機構提出說明，也不要視此現象習以為常而閒置不管。

結語

當處理資料檢核時，對於敏感度較高的議題，因檢定時容易產生type 1 error，如社會行為的探討，故資料檢核需要降低type 1 error，反之，對於特異性較高的議題，因檢定時容易產生type 2 error，如公共紀律的維持，故資料檢核需要降低type 2 error。所以，在資料檢核的設計上，就必須視不同的檢核需要而調整。有鑑於此，聯徵中心評估資料品質時，必須謹慎思考各項資料欄位之屬性，由於資料欄位核心屬性不同而各自有其不同的檢核門檻及容忍誤差範圍，因此，資料檢核的邏輯設計及執行強度並沒有通用的標準。

聯徵中心於執行資料建入資料庫前先行的檢核作業時，應儘量避免將金融機構所報送正確的資料視為錯誤，逕自剔退於信用資料庫外，此外，聯徵中心於執行資料建入資料庫後再行的檢核作業時，應儘量避免將金融機構所報送錯誤的資料視為正確，放任建置於信用資料庫內；誠如前言，聯徵中心將持續秉持對資料報送品質持妥適的假設，研究如何設計適切的資料檢核邏輯，再藉由資料品質假設的檢定驗證，學習瞭解資料真正的內涵與品質。