行政院國家科學委員會專題研究計畫 成果報告

「至多選取 k 項」等類形的複選題分析
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 94-2118-M-004-007-
執 行 期 間 ：94 年 08 月 01 日至 95 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人 ：江振東

計畫參與人員：此計畫無參與人員：無

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

中 華 民 國 95 年 11 月 11 日

行政院國家科學委員會專題研究計畫成果報告

「至多選取 *k* 項」等類型的複選題分析

Analyzing "Choose at Most *k* Items" Type of Questions

計畫編號：NSC 94-2118-M-004-007

執行期限：94 年 8 月 1 日至 95 年 7 月 31 日

主持人：江振東 國立政治大學統計系

一、計畫中文摘要

複選式的問項型式在問卷調查中經常被採用，其中填答者可以勾選的項數多半不受限制，然而至多選取 *k* 項等選取總數受限的情形，也屢見不鮮。只是針對這類問題的統計資料分析，似乎都僅侷限在敘述性統計的呈現層次，更深入的統計分析，並不多見。此外，文獻中似乎有沒有見過有針對此一議題作探討的論述。因此在此一計畫中，我們就這種選取總數受限的複選式問題，提出幾種具體可行的統計分析方法。

關鍵詞：複選題、至多選取 *k* 項

Abstract

Although "Check All That Apply" questions are most frequently spotted in a questionnaire, "Choose at Most *k* Items" type of questions can also be seen from time to time. However, statistical analyses on this type of questions never seem to go beyond the level of summary statistics. One of the reasons may be due to the fact that statistical methods that can be applied to analyze the data collected this way are not well documented. In this study, we propose several statistical methods that can be used for this purpose.

Keywords: "Check All That Apply" questions, "Choose at most *k* Items" questions

二、計畫緣由與目的

這個計畫基本上可以視為「複選題的分析—CMH 統計量的一個應用」這一篇文章的一個延續。在進行統計諮詢時，我們常可見到複選題式的問題，出現在問卷裡。比方說：

1.請就下列筆記型電腦品牌，勾選您覺得最值得信賴的廠家(可複選)

□宏碁 □華塑 □技嘉 □精英 □聯強 □HP □IBM □Sony □DELL

2.請就下列科目，分別依照您的喜好程度分別作勾選：

	非常喜歡	喜歡	普通	不喜歡	非常不喜歡
國文	□	□	□	□	□
英文	□	□	□	□	□
數學	□	□	□	□	□
自然	□	□	□	□	□
社會	□	□	□	□	□

3.請就下列四種品牌鮮奶，分別勾選您考慮購買的最主要原因：

	價格	品質	品牌知名度
光泉	□	□	□
義美	□	□	□
統一	□	□	□
味全	□	□	□

這三種類型的問題，我們泛稱為廣義的複選題，其中第一種類型就是我們一般所常見的典型的複選題。雖然這三種類型的複選式問題在我們生活周遭經常被用來作為調查的工具，但是針對這一類問題的統計分析，基本上如果不是採用忽略資料本身相關性的方法來作處理，多半就只有敘述性統計量的呈現而已。前者由於忽略相關性，分析的結果可能不盡然可信；而後者則無法進一步做統計推論。此外，由於文獻中以複選題分析的統計方法作為探討主軸的論述，也並不多見，因此這是否意謂著複選題的統計分析並不容易進行，因此我們只能退而求其次，採用較為簡略的方式來呈現結果。其實並非如此。雖然相關的文獻確實並不多見，但是適合用來處理分析複選題的統計方法，確不在少數。Agresti and Liu(1999, 2001)採用以 model-based 的方式，針對典型複選題的統計分析來作探討；而前述所提及的「複選題的分析—CMH 統計量的一個應用」我們則是利用 CMH 統計量，提出一種非 model-based 的方式來分析廣義複選題的問題。由於典型複選題基本上可以視為廣義複選題的一種特例，因此 CMH 統計量也適用於典型複選題的分析。前述這三篇文章所提出的典型複選題的分析方式，其實採用的都是既有的統計方法，只是這些方法不曾被應用來作為複選式問題的統計分析工具罷了。因此針對前面三類複選式的問題的統計分析，實際上是可行的。

不過除了前述三種問題形式外，我們也常見到如下的問項型式：

4.請就下列筆記型電腦品牌，勾選您覺得最值得信賴的廠家(最多勾選三項)：

□宏碁 □華塑 □技嘉 □精英 □聯強 □HP □IBM □Sony □DELL

5.請就下列筆記型電腦，勾選三家您覺得最值得信賴的廠家？

□宏碁 □華塑 □技嘉 □精英 □聯強 □HP □IBM □Sony □DELL

6.請就下列筆記型電腦，分別以 1,2,3,標示出您所認為最值得信賴的前三名廠家？

□宏碁 □華塑 □技嘉 □精英 □聯強 □HP □IBM □Sony □DELL

這三種類型的複選題與典型複選題的最大差異，在於後者的勾選數目可以少到一項也不勾選，多則可以每個選項都做勾選，而前者則是限制填答者可以勾選的總數。其中的第 4 類型，至多僅能勾選三項；第 5 類型則是勾選數目限定為 3 個；第 6 類型則是在勾選的 3 項中還得依喜好程度排列順序。這三種類型的問題型式雖然不如前三者那麼廣泛被採用，但是也不時可以見著。然而相關的統計分析方法的探討，在文獻上似乎不曾見過，因此這也是本研究計畫所要探討的主題。

三、計畫結果與討論

就一般的複選題而言，我們可以令 $(y_{i1}, y_{i2}, \ldots, y_{iq})$ 來表示第 i 位受訪者就 q 個問題的回答結果，其中 $i = 1, 2, \ldots, n$，而 $y_{ij} \in \{0,1\}$。由於每一個問題的回應情況，可以有兩種不同的選擇，因此 $(y_{i1}, y_{i2}, \ldots, y_{iq})$ 總共可以有 2^q 種不同的回應組合。這一點其實就是我們所要探討的主題，與一般的複選題的主要不同點。就第 4 類型的問題，也就是至多選取 k 項的問題來說，由於 $0 \le \sum_j y_{ij} \le k,\ k < q$，因此不同的回應組合，總計只有 $\binom{q}{0} + \binom{q}{1} + \cdots + \binom{q}{k} = \sum_{j=1}^{k} \binom{q}{j}$ 種；至於第 5 類型的問題，則只有 $\binom{q}{k}$ 種回應組合。以 $n = 4,\ k = 2$ 為例，

No	y_1	y_2	y_3	y_4
1	0	0	0	0
2	0	0	0	1
3	0	0	1	0
4	0	1	0	0
5	1	0	0	0
6	0	0	1	1
7	0	1	0	1
8	1	0	0	1
9	0	1	1	0
10	1	0	1	0
11	1	1	0	0
12	0	1	1	1
13	1	0	1	1
14	1	1	0	1
15	1	1	1	0
16	1	1	1	1

一般的複選題總計有 16 種回應組合，而第 4 類型與第 5 類型的問題，則各只有 11 種與 6 種回應可能。如果我們將一般複選題的 16 種回應組合，視為一個 2×2×2×2 的列聯表的話，那麼第 4 類型與第 5 類型的問題則可以分別視為是一種 incomplete tables。因此前述三種情況實際上可以分別視為來自 multinomial$(n;\pi_1,\ldots,\pi_{16})$，multinomial$(n;\pi_1,\ldots,\pi_{11})$，或 multinomial$(n;\pi_6,\ldots,\pi_{11})$ 的數據，如此一來，任何可以用來處理一般複選題的分析方法，只要經過適度調整，理論上都可以用來處理第 4 類型或第 5 類型的問題。

無論哪一種資料型態，我們想要檢定的是 $H_0: P(Y_1=1)=\cdots=P(Y_q=1)$。由於第 5 類型可以視為第 4 類型問題的特例，因此我們將僅就第 4 類型的問題分析做說明。

(一) Model-Based 方式

(1) Cochran's Q

就一般的複選題而言，我們可以透過 Cochran's Q 統計量來作分析。由於資料結構相同，唯一的不同點只在於回應的可能組合數較少而已，因此 Cochran's Q 自然也適用於此。由於 Cochran's Q 近似於卡分配，雖然一般的複選題與第 4 類型問題，我們所使用的是相同的統計量來作分析，我們也觀察到Q統計量使用於第4類型問題的分析時，收斂速度似乎有較快的傾向。

(2)至於小樣本時，我們可以使用 permutation test 來進行分析，步驟如下：首先，使用原始資料來計算出 $S_{obs}=\sum_{j=1}^{q}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$。接下來，就每一位受訪者的資料，進行「 permutation」的過程。總計共有 $\prod_{i=1}^{n}\binom{q}{y_{i\cdot}}$ 種情況。就前述的每一種情況，分別計算出 $S_c=\sum_{j=1}^{q}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2,\ c=1,\ldots,\prod\binom{q}{y_{i\cdot}}$。檢定的 p-值，我們可以定義為 $(S_c\geq S_{obs})$ 的個數 $\Big/\prod_{i=1}^{n}\binom{q}{y_{i\cdot}}$。

(二) Nonmodel-Based 方式

考慮模型如下：

$\ell(P(Y_j=1))=\alpha+\beta_j,\ j=1,\ldots,q$,其中連結函數(link function) ℓ 可以是 identity function，亦即 $\ell(x)=x$，或者是 logit function，亦即 $\ell(x)=\log\frac{x}{1-x}$。唯一的麻煩之處在於 $P(Y_j=1),\ j=1,\ldots,q$ 的定義方式。我們仍以 $n=4,\ k=2$ 為例來作說明。由於

$$P(Y_1=1)=\pi_{1000}+\pi_{1001}+\pi_{1010}+\pi_{1100}$$
$$P(Y_2=1)=\pi_{0100}+\pi_{0101}+\pi_{0110}+\pi_{1100}$$
$$P(Y_3=1)=\pi_{0010}+\pi_{0011}+\pi_{0110}+\pi_{1010}$$
$$P(Y_4=1)=\pi_{0001}+\pi_{0011}+\pi_{0101}+\pi_{1001}$$

就模型 $P(Y_j=1)=\alpha+\beta_j,\ j=1,\ldots,q$ 而

言，我們可以表示為 $\mathbf{A\pi = X\beta}$ 的形式；就就模型 $\text{logit}(P(Y_j = 1)) = \alpha + \beta_j,\ j = 1,...,q$ 而言，則可以表示為 $\mathbf{C}\log\mathbf{A\pi = X\beta}$ 的形式，其中 **C**、**A**、及 **X** 為矩陣，而 $\boldsymbol{\pi} = (\pi_{0000}, \pi_{0001}, \pi_{0010}, \pi_{0100}, \pi_{1000}, \pi_{0011}, \pi_{0101}, \pi_{1001}, \pi_{0110}, \pi_{1010}, \pi_{1100})'$。由於這些數據可以視為來自於 $\text{multinomial}(n; \pi_1, \ldots, \pi_{11})$，因此我們可以透過 constrainted mle 的方式，來求出參數估計值，並進行檢定

至於第 6 類的問題，我們可以使用 Friedman's Test 來作處理，只是由於我們僅能就最喜好的幾個選項做勾選並排序，未被選取到的項目可以視為 tie 的情況，因此我們的主要工作就是要對 tie 存在的情況做處理。由於 Friedman's Test 原本就可以處理 tie 存在的問題，因此用於第 6 類問題的處理上，並不會衍生任何麻煩。此外就小樣本的情況，我們依舊可以使用 permutation test 來進行分析。

四、計畫成果自評

文獻中有關「複選題分析」這類標題的探討，似乎並不多見，然而可以用來處理類似問題的方式倒是不少。「至多選取 k 項」等類型的複選題，與一般複選題的最大差異僅在於回應的可能組合數較少，因而形成 incomplete tables 的資料結構。因此我們所需要克服的唯一關卡就是機率模型的調整，除此之外一般複選題的分析方式都可以沿用。因此雖然在此計畫中，就統計方法而言，我們並沒有新的創見，然而結合 incomplete tables 的想法，與一般複選題的分析方式，我們成功的提出了適用於「至多選取 k 項」等類型的複選題的分析方式，相信這應該可以提供一般大眾在處理類似問題時，一種別於敘述性統計呈現的選擇。

五、參考文獻

1. 江振東 (2005)。「複選題的分析—CMH 統計量的一個應用」。智慧科技與應用統計學報，第三卷，第二期。
2. Agresti, A. (2002). *Categorical Data Analysis*. New York:Wiley.
3. Agresti, A. and Liu, I. (2001). "Strateqies for Modeling a Categorical Variables Allowing Multiple Category Choices" *Sociological Methods & Research*, 29:403-434.
4. Agresti, A. and Liu, I. (1999). "Marginal Modeling of a Categorical Variable Allowing Arbitrarily Many Category Choices" *Biometrics*, 55:936-943.
5. Berry, K.J., and Mielke, P.W. (2003). "Permutation Analysis of Data with Multiple Binary Category Choices" *Psychological Reports*, 92:91-98
6. Bilder, C.R., Loughin, T.M., and Nettleton, D. (2000). "Multiple Marginal Independence Testing for Pick Any/c Variables" *Communications in Statistics-Simulation and Computation*, 29(4):1285-1316.
7. Cochran, W.G. (1950). "The Comparison of Percentages in Matched Samples" *Biometrics*, 37:256-266.
8. Patil, K.D. (1975). "Cochran's Q Test: Exact Distribution" *Journal of the American Statistical Association*, 70:186-189.
9. Tate, M.W., and Brown, S.M. (1970). "Note on the Cochran's Q Test" *Journal of the American Statistical Association*, 65:155-160.