Journal of Propagations in Probability and Statistics
15(2), 97-103 Dec. 2015

統 計 薪 傳
第十五卷 第二期
民國104年12月97-103頁

用於分層隨機抽樣的一個自由度公式

陳宇宏 張光昭
展欣科技企業公司 輔仁大學

摘 要 在分層隨機抽樣的理論中，有一個用於區間估計而關於 t 分佈之自由度的公式。許多大專院校的學生與教師，包括統計科系的學生與教師，可能對於這個公式並不知曉，因爲這個公式只能從某些特定的抽樣學書籍中才找得到。在這一篇教學短文裡，作者們介紹這個公式並搭配一個計算例題，以餉廣大的統計薪傳讀者。作者希望本文的內容對於需要用到統計學之莘莘學子與教師們有一些可用之處。

關鍵字詞 分層隨機抽樣、區間估計、t 分佈、自由度、層大小、樣本大小、樣本變異數、樣本平均數、簡單隨機抽樣、有限母體、研究變數、母體大小。

□民國一百零四年五月收稿，一百零四年八月修訂、十月定稿。
□本文第一作者爲展欣科技企業有限公司負責人，電子郵址: techcom5054@hotmail.com；第二作者爲輔仁大學統計資訊學系專任教授；電子郵址: stat1016@mail.fju.edu.tw。

英文摘要/ English Abstract

A Formula of Degrees of Freedom in Stratified Random Sampling

Ardor Chen **Kuang-Chao Chang**
Techcom Information Corp. *Fu Jen Catholic University*

ABSTRACT In the theory of stratified random sampling, there is a formula of *degrees of freedom* (d.f.) for interval estimation based on *t* distributon. This formula may not be known to many college/university students and teachers, including those majoring in statistics, as it can be found only in certain books on sampling theory. In this short article, we introduce this formula of d.f. along with a computational example to the broad audience of JPPS. We hope the contents of this article can be useful to students and teachers who use statistics.

Keywords Stratified random sampling; Interval estimation; *t* distribution; Degrees of freedom; Stratum size; Sample size; Sample variance; Sample mean; Simple random sampling; Finite population; Study variable; Population size.

□Received May 2015, revised August 2015, in final form October 2015.
□Ardor Chen is the founder and CEO of Techcom Information Corp., Taipei, Taiwan, ROC; email: techcom5054@hotmail.com. Kuang-Chao Chang is a Professor in the Department of Statistics and Information Science at Fu Jen Catholic University, Hsinchuang, New Taipei City, Taiwan, ROC; email: stat1016@mail.fju.edu.tw.

 ISSN 1607-7083

1. 前言

在統計抽樣學的**分層隨機抽樣**(stratified random sampling)理論中，有一個用於**區間估計**(interval estimation)而較不爲人知的公式，用來計算小樣本區間估計所使用之 ***t* 分佈**(*t* distribution)的**自由度**(degrees of freedom; d.f.)。在一般較爲初階的抽樣學書籍裡，例如 Scheaffer *et al.* (2012)一書，對於分層隨機抽樣的區間估計這個議題主要只會談到一個屬於大樣本的公式，來計算母體平均數 μ 的100(1−α)% 信賴區間，如下:

$$\left(\hat{\mu}_{\mathrm{St}} - z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\mu}_{\mathrm{St}})},\ \ \hat{\mu}_{\mathrm{St}} + z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\mu}_{\mathrm{St}})}\right), \tag{1.1}$$

其中 $\hat{\mu}_{\mathrm{St}}$ 爲母體平均數 μ 的加權型式點估計量， $z_{\alpha/2}$ 爲標準常態分佈的第$[1-(\alpha/2)]$百分位置點。如果將公式(1.1)用於小樣本之情況，雖然也未嘗不可，但其精準度會降低。

在一些較爲高階的抽樣學書籍裡，例如 Cochran (1977)一書，會談到一個與公式(1.1)相似但是精準度較高的小樣本區間估計公式，如下:

$$\left(\hat{\mu}_{\mathrm{St}} - t_{\alpha/2,\,\nu}\sqrt{\widehat{Var}(\hat{\mu}_{\mathrm{St}})},\ \ \hat{\mu}_{\mathrm{St}} + t_{\alpha/2,\,\nu}\sqrt{\widehat{Var}(\hat{\mu}_{\mathrm{St}})}\right), \tag{1.2}$$

其中$t_{\alpha/2,\,\nu}$是自由度爲 ν 之 t 分佈的第$[1-(\alpha/2)]$百分位置點。這個t分佈的自由度 ν 是多少，就是這一篇教學短文的主題。在下一節，我們將會介紹一個關於這個自由度 ν 的計算公式並搭配一個計算例題。

2. 一個關於 *t* 分佈之自由度的計算公式

在前一節公式(1.2)之中的第$[1-(\alpha/2)]$百分位置點$t_{\alpha/2,\,\nu}$，其自由度ν的的來龍去脈可是相當地不簡單，它牽涉到 Satterwaite (1946)以及 Ames and Webster (1991)這兩篇學術論文。此外，在 Cochran (1977) 一書的 p. 96 以及 Govindarajulu (1999) 一書的 pp. 78-79，也有討論這個很特殊的自由度ν ，最後得到一個近似的計算公式如下:

$$\nu \approx \frac{\left(\sum_{h=1}^{L}\frac{N_h(N_h-n_h)}{n_h}S_h^2\right)^2}{\sum_{h=1}^{L}\frac{1}{n_h-1}\left(\frac{N_h(N_h-n_h)}{n_h}S_h^2\right)^2}, \tag{2.1}$$

式中諸多數學符號的意義分述於下:

L = 母體內的總層數；(因此，母體可以被完整地切割爲 L 個互不相交的層別)

N_h = 第 h 層之內的元素總數量 = 第 h 層的**層大小**(stratum size)， $h=1,2,\cdots,L$ ；

n_h = 從第 h 層抽樣的**樣本大小**(sample size)， $h=1,2,\cdots,L$ ；

S_h^2 = 從第 h 層抽得 n_h 個樣本數據 Y_{hj} ， $j=1,2,\cdots,n_h$ ， 所產生的**樣本變異數**(sample variance)

$$= \frac{1}{n_h-1}\sum_{j=1}^{n_h}(Y_{hj}-\bar{Y}_h)^2 \text{ ， } h=1,2,\cdots,L \text{ ， 其中}$$

$\bar{Y}_h$ = 從第 h 層抽得 n_h 個樣本數據 Y_{hj} ， $j = 1, 2, \cdots, n_h$ ， 所產生的**樣本平均數**(sample mean)

$= \dfrac{1}{n_h}\sum_{j=1}^{n_h} Y_{hj}$ ， $h = 1, 2, \cdots, L$ 。

我們以一個計算範例來說明如何使用公式(2.1)以及公式(1.2)， 並搭配**簡單隨機抽樣**(simple random sampling)之情況來相互比較。

計算範例 表 2.1之數據爲想像之中某一間大學某科系某班級的學生身高資料，以性別區分爲二個層別。

表 2.1 某大學某科系某班級的學生身高資料

男生(20 人)		女生(40 人)			
序號	身高 (公分)	序號	身高 (公分)	序號	身高 (公分)
1	181	21	162	41	168
2	168	22	160	42	157
3	172	23	155	43	153
4	165	24	159	44	164
5	169	25	156	45	151
6	173	26	164	46	163
7	178	27	166	47	155
8	166	28	160	48	165
9	170	29	158	49	162
10	162	30	165	50	152
11	184	31	152	51	162
12	168	32	159	52	164
13	175	33	150	53	163
14	179	34	166	54	164
15	163	35	154	55	165
16	164	36	151	56	159
17	176	37	156	57	171
18	172	38	148	58	152
19	171	39	152	59	156
20	174	40	160	60	154

若將這一個班級視爲一個**有限母體**(finite population)，每一位學生視爲一個元素，學生的身高視爲**研究變數**(study variable)，則母體之中 60 位學生的研究變數值依序爲 $y_1 = 181$, $y_2 = 168, \cdots,$ $y_{60} = 154$ 。所以，

N = **母體大小**(population size) = 母體之中的元素總個數 = 60，

$$\mu = \text{母體平均數} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{60}(181 + 168 + \cdots + 154) = \frac{9783}{60} = 163.05 \text{ 。}$$

我們暫且假裝不知道母體平均數 μ 的答案是多少，然後使用簡單隨機抽樣之方法從母體抽出 n = 8 個元素，來估計母體平均數 μ ，那麼也就是要從 60 位學生之中隨機地抽出 8 位學生，而這就需要介於 1 與 60 之間的 8 個亂數，來選出 8 位學生。因此，我們從某一本統計學書籍中節錄一小部份的亂數表，如以下表 2.1所示，以便稍後模擬抽樣之用。

表 2.1 從統計學書籍中節錄的一小部份亂數表

	column							
line	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
1	62956	95735	70988	86027	27648	65155	46301	27217
2	17143	50118	41681	87224	75674	43371	09846	83403
3	99285	01369	94610	71099	69207	01999	23931	34711

我們就從表 2.1的第一個橫列由左至右依序選取 8 個二位數字如後: 62、95、69、57、35、70、98、88。由於這 8 個亂數之中的 62、95、69、70、98、88 六個數字皆大於 60，所以必須將此六個數字減去 60，而得到 2、35、9、10、38、28。因此，原先選取的 8 個二位數字亂數就修正爲 2、35、9、57、35、10、38、28。但是，這 8 個修正之後的亂數之中的 35 出現了兩次，這就違反了簡單隨機抽樣是屬於“不置回抽樣”的基本原則。所以，我們將重複出現的亂數只保留出現一次，然後繼續抽取新的二位數字亂數(並加以修正，如果必要的話)，直到 8 個亂數完全沒有重複出現，才停止繼續抽取新的亂數。最後，我們得到 8 個完全沒有重複出現的修正後亂數如後: 2、35、9、57、10、38、28、60，這就是我們要隨機地抽出 8 位學生的序號，那麼抽得的 8 個身高值就是 168、154、170、156、162、148、160、154。雖然以上 8 個身高值也有重複出現的情形(154 出現兩次)，但是這可就沒有違反簡單隨機抽樣是屬於不置回抽樣的基本原則啦！因此，使用簡單隨機抽樣之方法來抽樣並推估母體平均數 μ ，其估計值爲樣本平均數，如下:

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{1}{8}(168+154+\cdots+154)=\frac{1272}{8}=159\ 。$$

接下來，如果我們採用分層隨機抽樣之方法從母體抽出 $n=8$ 個元素，來估計母體平均數 μ ，那麼爲了抽樣的簡便，並考慮到母體之中的男生人數較少而女生人數較多，不妨從男生這一層抽出$n_1=3$位學生，從女生層抽出$n_2=5$位學生，如此一來樣本之中的男女生比例就和母體之中的男女生比例頗爲接近，同時$n_1+n_2=3+5=8=n$。我們依然可以利用先前抽得的 8 個尚未修正之亂數: 62、95、69、57、35、70、98、88，來進行分層隨機抽樣，如下:

第一步　將 8 個亂數的前三個亂數乘以 0.2 倍之後再將小數點之後的數字四捨五入，變成爲介於 1 與 20 之間的三個整數: 12、19、14。然後，從男生這一層抽出序號爲這三個整數的三位學生，他們的身高值依序爲 168、171、179。

第二步　將 8 個亂數的後五個亂數乘以 0.4 倍之後再將小數點之後的數字四捨五入，然後再加上 20，變成爲介於 31 與 60 之間的五個整數: 43、34、48、59、55。隨後，從女生這一層抽出序號爲這五個整數的五位學生，她們的身高值依序爲 153、166、151、156、165。

有了以上二個步驟所抽得的樣本數據，則(1.1)式之中的加權型態估計量 $\hat{\mu}_{St}$ 所產生的估計值爲

$$\hat{\mu}_{St}=\sum_{h=1}^{2}\left(\frac{N_h}{N}\right)\bar{Y}_h=\left(\frac{20}{60}\right)\bar{Y}_1+\left(\frac{40}{60}\right)\bar{Y}_2=\frac{1}{3}\left(\frac{1}{3}(168+171+179)\right)+\frac{2}{3}\left(\frac{1}{5}(153+\cdots+165)\right)$$

$$=\frac{518}{9}+\frac{1582}{15}=\frac{7336}{45}\approx 163.02\ 。$$

由於母體平均數的標準答案是 $\mu=163.05$，因此就以 $\bar{Y}=159$ 與 $\hat{\mu}_{\text{St}}=163.02$ 這兩個點估計值來相互比較，很顯然 $\hat{\mu}_{\text{St}}=163.02$ 的估計誤差較小，也就是說，採用分層隨機抽樣較優於採用簡單隨機抽樣。不過，我們主要的目的是要進行區間估計，而在簡單隨機抽樣的情況下，μ 的 $100(1-\alpha)\%$ 小樣本近似信賴區間公式爲

$$\left(\bar{Y}-t_{\alpha/2,\,n-1}\sqrt{\left(\frac{N-n}{N}\right)\frac{S^2}{n}},\ \bar{Y}+t_{\alpha/2,\,n-1}\sqrt{\left(\frac{N-n}{N}\right)\frac{S^2}{n}}\right), \tag{2.2}$$

其中 $t_{\alpha/2,\,n-1}$ 是自由度爲 $(n-1)$ 之 t 分佈的第 $[1-(\alpha/2)]$ 百分位置點，S^2 爲依據簡單隨機抽樣之樣本數據所產生的樣本變異數。因此，

採用簡單隨機抽樣的樣本平均數 $\bar{Y}$ 所求得之 μ 的95%信賴區間

$$=\bar{Y}\mp t_{\alpha/2,\,n-1}\sqrt{\left(\frac{N-n}{N}\right)\frac{S^2}{n}}=159\mp t_{0.025,\,7}\sqrt{\left(\frac{60-8}{60}\right)\cdot\frac{56}{8}}$$

$$\approx 159\mp(2.365)(2.463)$$

$$\approx 159\mp 5.825=(153.175,\ 164.825)，$$

其中

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2\right)=\frac{1}{7}[(168^2+154^2+\cdots+154^2)-8\cdot(159)^2]$$

$$=\frac{1}{7}(202640-202248)=\frac{392}{7}=56\ 。$$

以上公式(2.2) 可以從 Cochran (1977) 一書的 p. 27 以及 Thompson (1992) 一書的 p. 27 查閱得到。接下來，我們利用公式(1.2)來計算在分層隨機抽樣之下母體平均數 μ 的95%信賴區間。首先，我們計算公式(2.1)的特殊自由度 ν，如下:

$$S_1^2=\frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j}-\bar{Y}_1)^2=\frac{1}{n_1-1}\left(\sum_{j=1}^{n_1}Y_{1j}^2-n_1(\bar{Y}_1)^2\right)$$

$$=(\frac{1}{3-1})\left((168)^2+(171)^2+(179)^2-3\cdot[\frac{1}{3}(168+171+179)]^2\right)$$

$$=\frac{1}{2}\left(89506-\frac{1}{3}(518)^2\right)=\frac{97}{3}=32.\overline{3}\ ；$$

$$S_2^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j}-\bar{Y}_2)^2=\frac{1}{n_2-1}\left(\sum_{j=1}^{n_2}Y_{2j}^2-n_2(\bar{Y}_2)^2\right)$$

$$=(\frac{1}{5-1})\left((153)^2+\cdots+(165)^2-5\cdot[\frac{1}{5}(153+\cdots+165)]^2\right)$$

$$=\frac{1}{4}\left(125327-\frac{1}{5}(791)^2\right)=47.7\ ；$$

$$\nu \approx \frac{\left(\sum_{h=1}^{L} \frac{N_h(N_h-n_h)}{n_h} S_h^2\right)^2}{\sum_{h=1}^{L} \frac{1}{n_h-1}\left(\frac{N_h(N_h-n_h)}{n_h} S_h^2\right)^2}$$

$$= \frac{\left(\frac{20(20-3)}{3}\cdot\frac{97}{3} + \frac{40(40-5)}{5}(47.7)\right)^2}{(\frac{1}{3-1})\left(\frac{20(20-3)}{3}\cdot\frac{97}{3}\right)^2 + (\frac{1}{5-1})\left(\frac{40(40-5)}{5}(47.7)\right)^2} \approx 5.646 \text{ 。}$$

因爲$\nu \approx 5.646$不是整數，所以我們將它四捨五入，得$\nu = 6$。接著，我們計算$Var(\hat{\mu}_{St})$之不偏估計量的估計值，如下:

$$\widehat{Var(\hat{\mu}_{St})} = \sum_{h=1}^{L} W_h^2 \left(\frac{N_h-n_h}{N_h}\right)\frac{S_h^2}{n_h} = (\frac{1}{3})^2(\frac{20-3}{20\cdot 3})\frac{97}{3} + (\frac{2}{3})^2(\frac{40-5}{40\cdot 5})(47.7) \approx 4.7279 \text{ 。}$$

因此，依據公式(1.2)，分層隨機抽樣的加權推估所求得之μ的95%信賴區間爲

$$\left(\hat{\mu}_{St} - t_{\alpha/2,\,\nu}\sqrt{\widehat{Var(\hat{\mu}_{St})}},\ \ \hat{\mu}_{St} + t_{\alpha/2,\,\nu}\sqrt{\widehat{Var(\hat{\mu}_{St})}}\right) = 163.02 \mp t_{0.025,\,6}\sqrt{\widehat{Var(\hat{\mu}_{St})}}$$

$$\approx 163.02 \mp (2.447)\sqrt{4.7279}$$

$$\approx 163.02 \mp 5.32 = (157.7,\ 168.34) \text{ ，}$$

其中$t_{0.025,\,6} = 2.447$可從一般統計學書籍中的 t 分佈百分位數表查閱得到。

以上採用分層隨機抽樣求得之信賴區間的一半寬度約爲5.32，小於簡單隨機抽樣求得之信賴區間的一半寬度5.825，因此相較之下，分層隨機抽樣優於簡單隨機抽樣。 □

3. 結語

在這一篇教學短文裡，作者們介紹了一個用於分層隨機抽樣的自由度公式，雖然這個公式未必適用於現實生活中的調查工作，不過它卻具有抽樣學理論上的意義。作者們希望本文的內容對於需要用到統計學之莘莘學子與教師們有一些可用之處。

參考文獻

Ames, M. H. and Webster, J. T. (1991). On estimating approximate degrees of freedom, *The American Statisticians*, 45, 45-50.

Cochran, W. G. (1977). Sampling Techniques, 3rd ed., John Wiley & Sons.

Govindarajulu, Z. (1999). Elememts of Sampling Theory and Methods, Prentice Hall.

Satterwaite, F. E. (1946). An approximate distribution of estimates of variance components, *Biometrics*, **2**, 110-114.

Scheaffer, R. L., Mendenhall, W., Ott, R. L., and Gerow, K. G. (2012). Survey Sampling, 7th ed., Brooks/Cole, Cengage Learning.

Thompson, S. K. (1992). Sampling, John Wiley & Sons.

(魏蘇珊文教事業機構發行，總公司: 中華民國臺灣新竹市建美路 2 巷 26 號。)

 ISSN 1607-7083