電腦與統計公式的選擇

盧 欽 銘

電子計算機（electronic computer），俗稱電腦（electronic brain），目前已被廣泛使用於統計學界。同時也由於電腦的運用，統計學領域中與起了不少革新的措施，其中以採用「運算公式（computational formula）來替代「定義公式」（difinitional formula）一項最爲明顯。的確，應用運算公式的電腦作業，可以節省不少電腦時間。然而，電腦終究是人爲的機器，它祇可以在其有限的運作單位之準確範圍內，遵照人們的指示進行快速而正確的工作。這也就是說：電腦運作單位有一定的準確範圍（即有效數字之位數是固定的），要是利用電腦作業之資料的數字位數過大，就會產生「削截誤差」（truncation error），而減低了統計結果之準確性（precision）。因此，增加電腦運作單位之準確範圍，當有助於統計資料的電腦作業。唯在電腦運作單位的準確範圍能作合理的調整之前，選擇適當的統計公式，或採用“倍式準確性的浮點情況（double-precision floating point mode）（縮稱倍準情況）”以代替“單式準確性的浮點情況（single-precision floating point mode）（縮稱單準情況）”，是獲致比較準確的統計結果之變通辦法。

電腦運作單位的準確性，具範圍常有一定的限度，爲增加其準確性的範圍，乃有倍準情況的設置。在這種運作情況中，電腦準確性之運作單位的準確性範圍可以增加一倍。然而倍準情況下，電腦的作業甚爲複雜，而且又需用更多的電腦時間；因此，只有革新電腦運作單位準確性的範圍，才是處理數字位數較大的統計資料的最佳辦法。

本文之目的在以逐漸增大位數的資料之電腦作業，說明電腦運作單位的準確性的範圍，會對於統計結果之準確性，有頗爲顯著的影響；同時指出：如果採用適當之統計公式，當可以減少電腦作業中的削截誤差。

方 法

1. 研究工具：本文中所有數字資料，都是利用美國北科羅拉多大學電腦中心 IBM 360 型之電子計算機來處理的。該項機器性能頗優，在單準情況下其有效數字爲 7 位，倍準情況則爲 16 位。

2. 數字資料：本文共有八組數字資料，每組包括九十六數目，第 1 組的數目係從 1 至 96的範圍中選擇出來的（但並非 1 至 96 的 96 個自然數），其平均數爲 48·50，變異數爲 875·00。第 2 至第 8 各組之資料，都是由第 1 組資料加上常數的「直線轉換」而構成的。就以第 2 組資料爲例：它係將第 1 組資料各加上 9 而形成的，因係加上常數的直線轉換，第 2 組資料的平均數也比第 1 組的平均數增加 9，而爲 57·50；而變異數未受影響，仍爲 875·00。同樣地其餘第 3 至第 8 各組之資料，也都是循第 2 組資料相似的步驟，由第 1 組者加上常數的直線轉換而構成的。此八組資料之數字範圍、平均數和變異數都詳列於表一。

3. 處理步驟：

表一 各組數字資料之範圍、平均數和變異數

組別	數字範圍	平均數	變異數
1	1. — 96.	48.50	875.00
2	10. — 105.	57.50	875.00
3	100. — 195.	147.50	875.00
4	1000. — 1095.	1047.50	875.00
5	10000. — 10095.	10047.50	875.00
6	100000. — 100095.	100047.50	875.00
7	1000000. — 1000095.	1000047.50	875.00
8	10000000. — 10000095.	10000047.50	875.00

電腦作業係分別在「單準」和「倍準」兩種情況下，進行下列統計數量的運算：

(1) 利用公式Ⅰ、Ⅱ，求各組資料之平均數：計算平均數之定義公式和運算公式沒有區別，公式Ⅰ是定義公式，也就是運算公式，公式Ⅱ是簡捷法計算平均數的公式。

(2) 利用公式Ⅲ、Ⅳ，求各組資料之變異數：公式Ⅲ爲定義公式，公式Ⅳ爲運算公式。

(3) 利用公式Ⅴ、Ⅳ，求八組資料間之相關係數：前者爲定義公式，後者爲運算公式。

公式Ⅰ：

$$\bar{X} = \frac{\Sigma X}{N}$$

公式Ⅱ：

$$\bar{X} = X' + \frac{\Sigma (X - X')}{N}$$

公式Ⅲ：

$$S^2 = \frac{\Sigma (X - \bar{X})^2}{N - 1}$$

公式Ⅳ：

$$S^2 = \frac{N\Sigma X^2 - (\Sigma X)^2}{N (N-1)}$$

公式Ⅴ：

$$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2}\sqrt{\Sigma (Y - \bar{Y})^2}}$$

公式Ⅵ：

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

上列各式中：

X、Y 爲變項 X、Y 中之數值

$\bar{X}$、$\bar{Y}$ 爲變項 X、Y 中數值之平均數

S^2 爲變項 X 中數值之變異數

X' 爲變項 X 之假定平均數

r 爲變項 X、Y 間之相關係數

N 爲總次數

結　果

1. 各組數字之平均數

在單準、倍準兩種情況下，利用公式Ⅰ和公式Ⅱ所求得之各組資料之平均數如表二。比較表一和表二中各組之平均數，不難看出：祇有在倍準情況下，利用公式Ⅱ所求得之各組平均數，才與各組實際上之平均數（見表一）是完全相符合的。也就是說：儘管在倍準情況下，機器之運作範圍雖已經加倍，然而在利用公式Ⅰ計算平均數時，也不能完全免除削截誤差（truncation error）的產生。

除了在倍準情況下，利用公式Ⅱ所求得之各組平均數完全沒有削截誤差外，在其餘各種情況下所求得之各組之平均數，與各組實際上之平均數都不完全相符合；同時此等削截誤差也不是均等的，資料的數目字範圍小的諸組中，其平均數尚與實際上之平均數相等。資料的數目字範圍大的各組中，則有削截誤差存在，而且數字資料範圍越大時，其削截誤差也就愈大。在那些誤差中，小的雖祇有 .02，大的卻達到 7.50。

另外尚有一事實是值得注意的：就是無論是單準或倍準情況下，利用公式Ⅱ（簡捷法公式）所得各組平均數，都要比利用公式Ⅰ所得者來得準確些。

表二　由電腦處理所得之各組資料之平均數

組別	單準情況		倍準情況	
	公式Ⅰ	公式Ⅱ	公式Ⅰ	公式Ⅱ
1	48.50	48.50	48.50	48.50
2	57.50	57.50	57.50	57.50
3	147.50	147.50	147.50	147.50
4	1047.50	1047.50	1047.50	1047.50
5	10047.50	10047.50	10047.50	10047.50
6	100047.50	100047.50	100047.30	100047.50
7	1000047.56	1000047.50	1000047.30	1000047.50
8	10000040.00	10000047.00	10000045.53	10000047.50

2. 各組數字的變異數

在單準、倍準情況下，利用公式Ⅲ（定義公式）和公式Ⅳ（運算公式）所求得各組資料之變異數均見表3。由於第2至第8各組之數字資料，均係由第1組者加上常數的直線轉換所構成的，因此各組之變異數都應該相等（即都應等於875·00），但表3所列舉各組之變異數中，在單準或倍準情況下，利用公式Ⅲ所得者，各有六個與實際上之變異數相符合，其餘二個則與實際上的數值不盡一致，而且是當數字資料範圍越大時，其誤差也越顯著。

再看利用公式Ⅳ（運算公式）所求得之各組資料的變異數中，在兩種情況之下，各有五個數值與實際上的變異數不一致，其削截誤差至爲明顯，甚至在單式精確性的浮點情況下，其所得變異數尚發現有負值的情形。由此可見，由運算公式計算變異數，比由定義公式計算變異數時，可能會造成更大的誤差。

表三　由電腦處理所得各組資料的變異數

組別	單準情況		倍準情況	
	公式Ⅲ	公式Ⅳ	公式Ⅲ	公式Ⅳ
1	875.00	875.00	875.00	875.00
2	875.00	875.00	875.00	875.00
3	875.00	875.00	875.00	875.00
4	875·00	874.44	875.00	875.23
5	875.00	776.08	875.00	895.93
6	875.00	−27594.10	875.00	2949.77
7	878.95	2119227.00	875.04	208175.15
8	934.21	33907632.00	879.08	20729117.95

表四　單準情況下的各組間之相關係數

組別	1	2	3	4	5	6	7	8
1		1.000	1.000	.996	.756	.198	.007	.001
2	1.000		1.000	.996	.756	.198	.007	.001
3	1.000	1.000		.996	.753	.196	.008	.000
4	1.000	1.000	1.000		.750	.171	.009	.000
5	1.000	1.000	1.000	1.000		−.089	−.015	−.317
6	1.000	1.000	1.000	1.000	1.000		−.129	−.535
7	.998	.998	.998	.998	.998	.998		.242
8	.968	.968	.968	.968	.968	.968	.982	

附註：1. 在斜線下半部之相關係數是利用公式Ⅴ求得的

2. 在斜線上半部之相關係數是利用公式Ⅵ求得的

3. 各組數字間之相關係數

由於各組數字資料間具有直線轉換的關係，各組間之相關係數應該是完全正相關 (即$r=1.000$)，然而在單準情況下，利用運算公式所求得之二十八個相關係數中，祇有三個相關係數等於1.000，其餘二十五個都不等於1.000，甚至於有五個相關係數是負值的（見表四）。由此可見因電腦運作單位的範圍限制，所造成的誤差確甚顯著。但是利用公式Ⅴ（定義公式）計算出來的二十八個相關係數，則有半數以上的相關係數等於1.000，其餘不等於1.000的，其誤差也比較小些（見表四）。

至於在倍準情況下，無論利用公式Ⅴ（運算公式）或公式Ⅵ（定義公式）所求得之相關係數（見表五），都甚爲準確，其間部分相關係數與實際上應得之相關係數有百分之一或二的誤差，祇能說是因圓整作用所產生的誤差（rounded error）了。

表五 倍準情況下的各組間之相關係數

組別	1	2	3	4	5	6	7	8
1		1.000	1.000	1.000	1.000	1.000	1.000	.998
2	1.000		1.000	1.000	1.000	1.000	1.000	.998
3	1.000	1.000		1.000	1.000	1.000	1.000	.998
4	1.000	1.000	1.000		1.000	1.000	1.000	.998
5	1.000	1.000	1.000	1.000		1.000	1.000	.998
6	1.000	1.000	1.000	1.000	1.000		1.000	.998
7	1.000	1.000	1.000	1.000	1.000	1.000		.998
8	.999	.999	.999	.999	.999	.999	.999	

附註：1. 在斜線下半部之相關係數是利用公式Ⅴ求得的

2. 在斜線上半部之相關係數是利用公式Ⅵ求得的

綜上所述，我們不難看出：當統計資料的數目不大時，無論是單準或倍準情況下，利用定義公式和利用運算公式的電腦作業，所求得之統計數量都是一致的，不會有若何誤差；但是當統計資料的數目字增大時，仍用運算公式去進行電腦處理，則會受到電腦運作單位的範圍之限制，而造成削截誤差，對於統計數量之準確性，可能產生頗爲明顯的影響。此時若能以倍準情況代替單準情況，並採用定義公式取代運算公式，則上述誤差會因此大爲減少。

通常，統計數量的誤差，除了極少數的特殊現象（如變異數爲負數，相關係數的絕對值大於1.000）易於覺察者外，其他的誤差多半不會引人注意，而每予忽略。因而在使用電腦進行統計分析時，應視統計資料數目字的大小，選用適用的公式，以及單準或倍準情況，以免因爲機器功能的限制，而得到不準確的統計數量，從而作了不準確的結論或推斷。

摘　要

1. 統計資料的數目字不大時，利用運算公式和利用定義公式的電腦作業，會獲得一致的統計量數。

2. 統計資料的數目字增大時，採取定義公式的電腦作業所得之統計數量，較採用運算公式者準確。

3. 統計資料的數字很大時，宜改用倍式準確性的浮點情況，而避免使用單式精確性的浮點情況之電腦作業。

4. 革新電腦運作單位之準確性的範圍，以便應用運算公式進行統計資料的電腦作業，乃爲當務之急。

參考資料

朱寶珍、盧欽銘：電子計算機簡介，中國測驗學會，測驗年刊第十三輯，民國五十五年，55～57。

蕭慕岳：電子計算機原理，中國電機工程學會，民國五十三年。

Barrett, J. P.: Elementary computer programs for statistical analysis, California, Belmont: Duxbury press, 1971.

Ford, D. H.: Basis fortran IV programming) Illinois, Homewood: Richard D. Irwin, 1971.

Lehman, R. S. etal,: Digital computing, New-York: John Wrley and Sons, 1968.

Veldman, D. J.: Fortran programming for the behavioral sciences, New-York: Holt, Rinehart and winston, 1967.

TRUNCATION ERROR AND THE SELECTION OF STATISTICAL FORMULA IN COMPUTER OPERATION

CHING-MING LU

ABSTRACT

This paper used eight groups of numerical data, which were composited by some different length of digits, to demonstrate the existance of truncation error in the computer operation.

The IBM 360 computer was used as the instrument. It provided 7 significant digits in the single-precision floating point mode, and 16 in double-precision flotaing point mode.

The findings show when the length of digits was long, using definition formula in both single and double precision floating point mode would get more precise outputs than using the computation formula; and there was no truncation error, if the length of digits was short.