計算出恰當的樣本數目

◎楊沛昇 編譯

我們帶著興趣研讀「Fallacies of Statistical Significance」[1]專欄一文

我們很高興地注意到對顯著性檢定與資料分析析的常識性方法是被支持的，我們同意「基本上，實務顯著性是依據現實世界中的考量…」，我們也支持對繪製資料與信賴區間的建議。

如你尚未研讀此一比較新爐與舊爐所生產產品之玻璃拉伸強度中位數的專欄文章，考慮使用於執行顯著性檢定之玻璃拉伸強度資料的上述兩種情況，對於每一種情況，有表格說明玻璃拉伸強度中位數於五種樣本數目與三種檢定顯著水準間所發現在統計上有顯著差異的結果機率，這兩種情況是：

情況1：顯著性檢定指出當沒有實務顯著性時的統計上顯著差異，例如，你可能總結出舊有爐生產出玻璃拉伸強度中位數為35百萬帕斯卡(MPa)的玻璃，在統計上優於生產出玻璃拉伸強度中位數為34MPa之玻璃的新爐，但是35MPa與34MPa之間3%的差異是否重要？

在專欄中，在以100、200、500、1,000與1,500五種樣本數目來計算新舊爐間玻璃拉伸強度中位數中形成統計上顯著差異的機率。

情況2：顯著性檢定未指出當有實務顯著性時的統計上顯著差異，例如，你可能總結出舊有爐生產出玻璃拉伸強度中位數為35MPa的玻璃並無顯著優於生產出玻璃拉伸強度中位數為31MPa之玻璃的新爐，當無統計上顯著差異時，35MPa與31MPa間的11%差異就不重要嗎？

以10、20、50、75與100五種樣本數目來計算新舊爐間玻璃拉伸強度中位數中形成統計上顯著差異的機率。

在那篇專欄中，對以盲目依循假設結果之測試為基礎的實際行動提出質疑，並提出採用常識性方法的建議。

該專欄以下列評論情況2為「實際上是說明樣本至樣本間的變異太大，且/或取自新爐的樣本數目太少而無法讓你得到確切的結論，因而需要來自新爐的額外樣本以獲得更確定的結果」[2]，作為結束。

由研讀該專欄，我們學習到樣本數目可以過小，但樣本數目可以過大嗎？在看完該專欄後，在心中浮現了一個問題「對於顯著性檢定是否存在有最佳樣本數目？」。

最佳樣本數目

一般來說，對這個問題的回答是「有」，但是在討論如何推測出一個統計上的最佳樣本數目之前，必須要承認一些現實世界中的限制。

實際上來說，你可以利用任一的樣本數目[3]，例如，若檢驗是破壞性或受限於成本考量，像是讓飛機墜毀以取得數據，則實際的樣本數目將會不同於數據需要量測零件尺寸的檢驗，或者你受限於過程本身，例如一個爐只能放四個零件且循環時間是一個禮拜。

在這些情形中，你選擇你能合理提供的最大樣本數目並做到最好，記住一點，

越大的樣本數目的結果是越窄的信賴區間，以做出更精準的決定。

在本文中，我們假設檢驗樣本數目是不受限的，如前所述。然後，在計算出統計上最佳樣本數目前，必須先回答三個問題：

1.應使用何種假設檢定(顯著性檢定)？這有許多的選擇，諮詢統計專家是有幫助的。表1.[4]中列出了部分的假設檢驗，在本文使用的假設檢驗是Z檢驗(Z test)，而做出下列的假設：

++我們想要執行檢驗比較方法

++已知為常態分布

++已知標準差

請注意，這與上述專欄中之中位數值對照與對數常態分布不同。

表 1 部分假設檢定清單(顯著性檢定)

假設檢定(顯著性檢定)的部分列表		檢定統計		
		1個樣本	2個樣本	2個樣本以上
檢定比較平均	標準差已知	Z(normal)	Z	F
	標準差未知	t(stundent t)	t	F
	問題舉例	新舊爐產出玻璃的拉伸強度是否不同?	吸煙與未吸煙孕婦女生產嬰兒的體重是否不同?	五台弧焊機的平均焊接斷裂強度是否有統計上的不同?
檢定比較變異		X2(Chi Squared)	F	Bartlett's
	問題舉例	新舊爐產出玻璃的拉伸強度間是否有變異性?	吸煙與未吸煙孕婦女生產嬰兒的體重間是否有變異性?	五台弧焊機的平均焊接斷裂強度間是否有變異性?
分布未知或非變數(例如：是或不是的數據)				
檢定比較中位數/平均		Sign test	Sign test	Kruskal-Wallis
	問題舉例	和上述檢定相同除了分布為	和上述檢定相同除了分布為	和上述檢定相同除了分布為
檢定比較變異		N/A	Siegel-Tukey	N/A
	問題舉例		和上述檢定相同除了分布為	

2.所比較數值的變化是有意義的嗎？按照玻璃爐的例子，假如舊爐的中位玻璃拉伸強度是35MPa，是中位數中的什麼差異可以造成實際差異的？例如，若新爐生產的玻璃其中位玻璃拉伸強度是34.99MPa，你會認為新舊爐間並無實際差異。因此，需要求得最佳最佳樣本數目，實際差異必須被明定。中位數或平均值的什麼差異是有意義的?接下來會以在文章開頭所說明新舊兩爐的情況為基礎，來考慮一些實際比較的例子。

3.當在解釋顯著性檢定的結果時，什麼是合理的誤差機率？以上述專欄文章中的情況1與情況2中，有兩種誤差機率：

+首先，就是當沒有顯著差異而你結論出存在有顯著差異，依循玻璃拉伸強度的情況，得到結論的例子就是推斷新爐生產出玻璃拉伸強度較低的玻璃，但實際上並非如此，樣本若有偏差因而不能代表真實結果則有可能發生。(在統計學

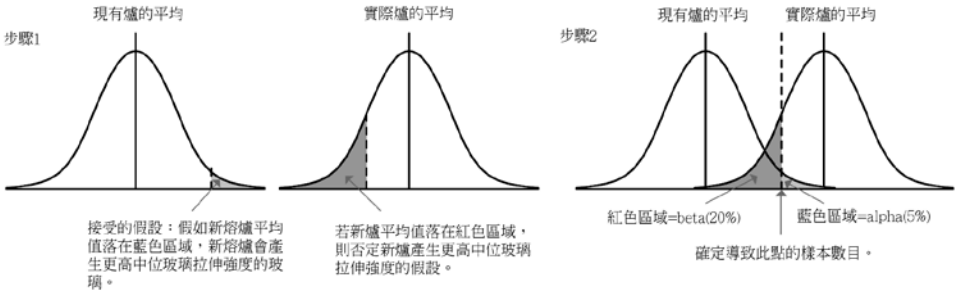
中，這是當為真時被否定之虛無假設的概率）

+其次，就是事實上存在有一個顯著差異但你推斷出沒有顯著差異的概率，依循玻璃拉伸強度的情況，得到結論的例子就是推斷新爐未生產出玻璃拉伸強度較低的玻璃，但實際上卻是有。如同前述，樣本若有偏差因而不能代表真實結果則有可能發生。（在統計學中，這是當為偽時被接受之虛無假設的概率）

在統計學中，第一型誤差機率稱為α(alpha)誤差，其中100(1-α)%分別等於單邊或雙邊假設檢定的信賴界線或信賴水準，第二型誤差機率稱為 β (beta)誤差，其中1- β 等於檢定的統計檢定力，在製造業中，第一型誤差機率的值通常為0.05，而第二型誤差機率的值通常為0.2(統計檢定力是80%)[5]。

給定 α 與 β 的值，接著必須解出兩個方程式以決定樣本數目，假設兩者均為常態分布，則過程如圖1所示，圖中展示的假設檢定例子是可用於決定新爐平均值是否高於舊爐平均值的單尾檢定。

圖 1 alpha 與 beta 的關聯



總體來看

我們現在說明在Excel表格中三個依循上述專欄文章中考慮的例子，如何可以簡單地計算出最佳樣本數目，但簡化了下列：

++玻璃拉伸強度之分布為常態分布

++新舊兩爐的拉伸強度分布均為已知

++新舊兩爐的標準差相同

++我們想知道新爐生產玻璃的拉伸強度平均值是否高於舊爐的，而非比較中位數。

鑑於上述假定，適當的假設檢定是單尾Z檢定。

在上述專欄中，舊爐生產玻璃的拉伸強度是中位數為35MPa、形狀參數為0.25的對數常態分布，由「工程與科學之概率與統計」[6]一書，可輕易決定出具有期望值36.1MPa與標準差9.2MPa之可比較常態分布的結果，利用這些數值來建構簡化情形中的常態拉伸強度分布。

但在計算樣本數目前滿必須對每一各例子回答三個樣本數目問題。

例子1：在此例中，對上述專欄中對比情況2的三個問題的答案是。

1.使用那種假設檢定？Z檢定。

2.所比較數值中變化的何者是實務顯著性？在專欄文章的情況2中，36.1MPa減去32.0MPa所得到的4.1MPa。35MPa對數常態分布中位數與31MPa的比較，對於這些可比較常態分布，分別將可比較常態分布中位數轉換為期望值36.1MPa與32.0MPa。

3.在解釋顯著性檢驗結果時，誤差機率是否合理？製造產業慣例為 α =0.05(信賴界/信賴水準=95%)與 β =0.2(統計檢定力=80%)

如圖2，可在EXCEL表格中依單尾Z檢定計算出樣本數目，為計算樣本數目，輸入所需訊息於欄位B14、B15與B16，將會產生灰色欄位中的數值，欄位B25即為最佳樣本數目，在此例中最佳樣本數目為32。

圖2

	A	B	C	D	E	F
1						
2	假設：					
3	－近乎常態分布					
4	－已知標準差					
5						
6	檢驗：					
7	－單側檢定(是一個平均值大於或小於另一個平均值?)					
8						
9	給定條件					
10	alpha/信賴區間(confidence interval)	0.05	95%			
11	beta/統計檢定力(power)	0.2	80%			
12						
13	輸入					
14	標準差	9.2				
15	現有平均值	36.1				
16	實際平均值	32		B欄方程式		
17	相差	4.1	⇐	=ABS(+B15-B16)		
18						
19						
20	Z(1-alpha)	1.645	⇐	=NORM.S.INV(1-B10)		
21	Z(1-beta)	0.842	⇐	=NORM.S.INV(1-B11)		
22						
23	樣品數目					
24	n(caculated)	31.1	⇐	=(+B14*((B20+B21)/(B17)))^2		
25	n(rounded up)	32	⇐	=ROUNDUP(+B24,0)		
26						

例子2：在此例中，對上述專欄中對比情況1的三個問題的答案是。

1. 使用那種假設檢定? Z檢定。

2.所比較數值中變化的何者是實務顯著性?在專欄文章的情況1中，36.1MPa減去35.1MPa所得到的1.0MPa。對數常態分布的中位數由35MPa至34MPa，分別將可比較常態分布中位數轉換為期望值36.1MPa與35.1MPa。

3.在解釋顯著性檢驗結果時，誤差機率是否合理？製造產業慣例為 α =0.05(信賴界/信賴水準=95%)與 β =0.2(統計檢定力=80%)

如圖3，可在EXCEL表格中決正最佳樣本數目為524。當兩個平均值的差異變小時，決定差異是否具統計性顯著時所需的最佳樣本數目則變大。類似地，假如標準差變小，則樣本數目則變小，如同下述例子3。

圖3

	A	B	C	D	E	F
1						
2	假設：					
3	－近乎常態分布					
4	－已知標準差					
5						
6	檢驗：					
7	－單側檢定(是一個平均值大於或小於另一個平均值?)					
8						
9	給定條件					
10	alpha/信賴區間(confidence interval)	0.05	95%			
11	beta/統計檢定力(power)	0.2	80%			
12						
13	輸入					
14	標準差	9.2				
15	現有平均值	36.1				
16	實際平均值	35.1		B欄方程式		
17	相差	1	⇐	=ABS(+B15-B16)		
18						
19						
20	Z(1-alpha)	1.645	⇐	=NORM.S.INV(1-B10)		
21	Z(1-beta)	0.842	⇐	=NORM.S.INV(1-B11)		
22						
23	樣品數目					
24	n(caculated)	523.3	⇐	=(+B14*((B20+B21)/(B17)))^2		
25	n(rounded up)	524	⇐	=ROUNDUP(+B24,0)		
26						

例子3：在此最後一個例子，使用與例子2相同的數據，除了標準差設為1.0外，如圖4中，以EXCEL表格決定出的最佳樣本數目為7。

圖4

	A	B	C	D	E	F
1						
2	假設：					
3	－近乎常態分布					
4	－已知標準差					
5						
6	檢驗：					
7	－單側檢定(是一個平均值大於或小於另一個平均值?)					
8						
9	給定條件					
10	alpha/信賴區間(confidence interval)	0.05	95%			
11	beta/統計檢定力(power)	0.2	80%			
12						
13	輸入					
14	標準差	1				
15	現有平均值	36.1				
16	實際平均值	35.1		B欄方程式		
17	相差	1	⇐	=ABS(+B15-B16)		
18						
19						
20	Z(1-alpha)	1.645	⇐	=NORM.S.INV(1-B10)		
21	Z(1-beta)	0.842	⇐	=NORM.S.INV(1-B11)		
22						
23	樣品數目					
24	n(caculated)	6.2	⇐	=(+B14*((B20+B21)/(B17)))^2		
25	n(rounded up)	7	⇐	=ROUNDUP(+B24,0)		
26						

最佳化方法

當對假設檢定為必須且必須計算出統計上地最佳樣本數目時，有三個問題必須處理(假設無成本、時間或過程限制)

1.使用何種假設檢定？

2.所比較數值中的那些變化為實務顯著性？

3.當解釋顯著性檢驗的結果時，合理的誤差機率為何？

提出用以回答上述問題的邏輯，接著，利用Excel表格計算出三個案例中的樣本數目，本文中三個案例中用以回答三個問題的答案是以專欄文章「Fallacies of Statistical Significance」中提供的數據為基礎。

假設為一個常態分布，在此說明的方法可用於計算假設檢定的最佳樣本數目，此外，假如分布情形為連續且近乎常態，但標準差為未知，此方法亦可用以計算一個合理近似的假設檢定的最佳樣本數目。

參考文獻和說明：

1.Necip Doganaksoy, Gerald J. Hahn and William Q. Meeker, "Statistics Spotlight: Fallacies of Statistical Significance," Quality Progress, November 2017, pp. 56-62.

2. Ibid.

3.The assumption is made that the data sample is representative of the population.

4.For a more complete list of tests of hypothesis, see Gopal K. Kanji, 100 Statistical Tests, third edition, Sage Publications, 2006.

5.Different choices for α or β may be preferable in different scenarios, and it can be helpful to explore a range of values to see how the results change. This may be a particularly useful exercise when confidence intervals are large.

6.Anthony Hayter, Probability and Statistics for Engineers and Scientists, third edition, Brooks/Cole, 1996, pp. 249-251.

7.To calculate the sample size for other tests of hypothesis, consult a statistician or, if statistically knowledgeable, make use of a commercial software program such as Minitab.

編輯者備註：

Christopher N. Bertoni died before this article was published.

作者：

Christopher N. Bertoni was an industrial engineer and a retired quality professional. He earned master's degrees in industrial and systems engineering at the University of Florida in Gainesville, and in business administration from Rensselaer Polytechnic Institute in Troy, NY. He was an ASQ member.

Bridget Bertoni is a policy associate at Acumen LLC in Burlingame, CA, and earned a doctorate in physics from the University of Washington in Seattle.

資料來源：Quality Progress December 2018, pp. 58-64