**Homework 1-2**

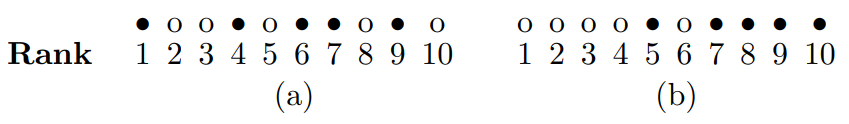
生工所 李世耀 F08622011



問題分析

**以序號總和測試分析臺北站1961~1980年、1990~2009年的7月日最高溫，α = 0.05顯著水準下，判斷兩組7月日最高溫紀錄分布是否相同。**

序號總和測試(Wilcoxon rank-sum test or Mann–Whitney U test)為非參數化檢定，主要測試兩組樣本分布是否相同，虛無假設為兩組樣本分布相同，因測試統計值只使用序號而非樣本數值，故為非參數化檢定。將兩組樣本混合並排序，若兩組樣本分布相同，我們預期兩組樣本在排序資料中為隨機分布，如圖 1 (a)；若其中一組小於或大於另一組，分布則可能呈現如圖 1 (b)。



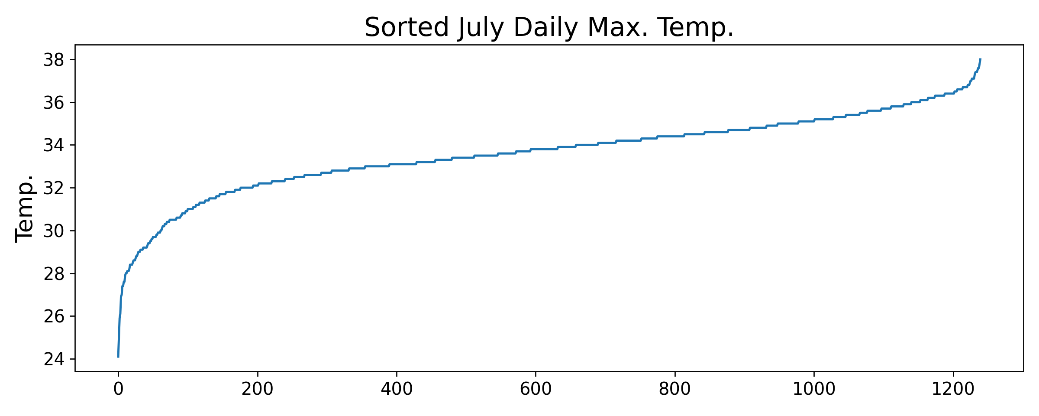
**圖 1**樣本來自相同分布與不同分布示意圖 (from: Chris Wild)[[1]](#footnote-1)

計算樣本個數較小的一組所有樣本的序號總和*W*，當兩組樣本數足夠大( > 10)時，*W*的分布可以使用常態分布近似，將*W*標準化後可得到標準常態分布的測試統計值*zrs*。若| *zrs* | > *z*1-α/2，拒絕虛無假設，兩組樣本分布不相同。

分析方法

給予1961~1980年、1990~2009年兩組7月日最高溫樣本不同標籤後，將兩組樣本混合並排序，因兩組樣本個數皆為620筆，因此可任選一組計算樣本序號總和*W*，利用下式計算測試統計值*zrs*，詳細程式請參見附錄。

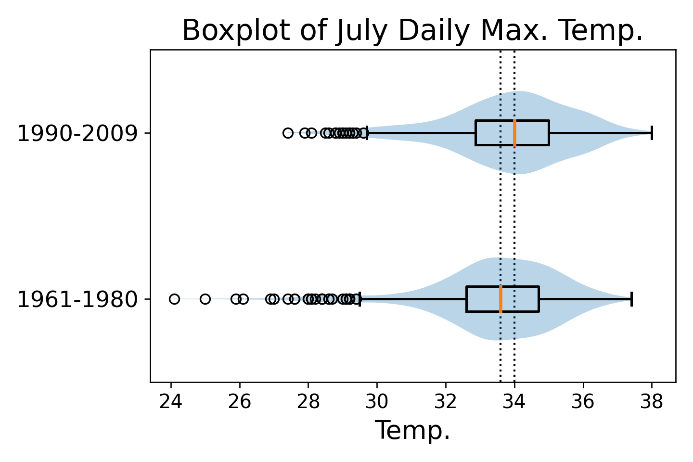
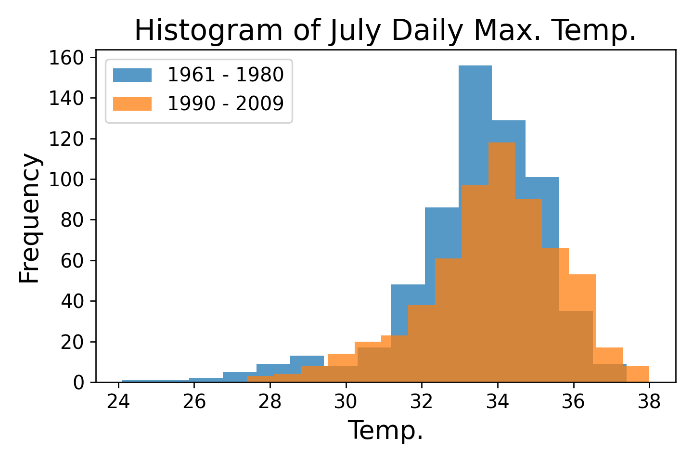
由圖 2可知數值相等(水平線段)的組數(1134組)並非少數，則上式標準偏差*σ*需改由下式估計。



**圖 2**　兩組7月日最高溫混合排序資料呈現(共1240筆)

結果分析

測試統計值***zrs* = 3.567 > *z*0.975 = 1.96**，故**拒絕虛無假設，兩組7月日最高溫樣本分布不相同**，兩組樣本直方圖與盒鬚圖呈現於圖 3，1990~2009年7月日最高溫略高於1961~1980年7月日最高溫。(使用scipy.stats.ranksums()與scipy.stats.mannwhitneyu()進行檢定，結果相同。)

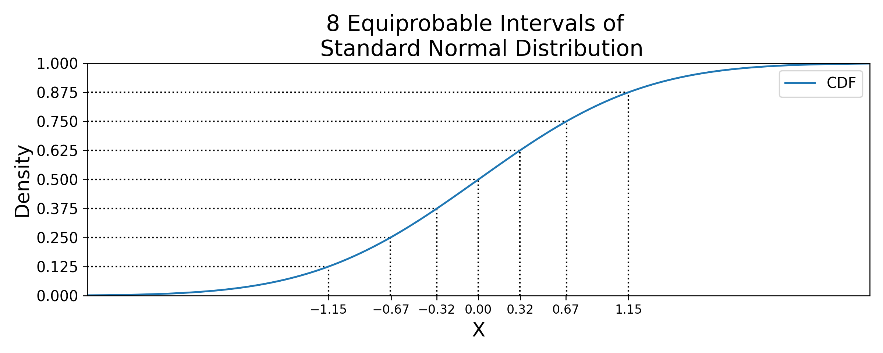


**圖 3**　1961~1980年、1990~2009年 7月日最高溫直方圖與盒鬚圖

問題分析

**樣本資料呈極端值第一型分布，但虛無假設為「隨機樣本呈常態分布」的「假設錯誤」類別問題中，顯著水準α= 0.05下，利用蒙地卡羅法，分析在不同樣本數的條件下，卡方檢定發生第二型錯誤的機率及其隨樣本數變化的曲線。**

常態分布為統計分析常用的假設分布，而樣本資料是否服從常態分布可以透過卡分檢定測試，但卡方檢定對於樣本數相當敏感，若樣本數太少，在「假設錯誤」的狀況下，卡方檢定則無法有效拒絕虛無假設(第二型錯誤發生機率β高)。利用蒙地卡羅法，產生不同樣本數量的資料各10,000組，分別進行卡方檢定，可以找出使卡方檢定第二型錯誤發生機率β< 0.05所需的最小樣本數。常態分布8個等機率區間(每個區間累積機率為0.125)如圖 4所示。

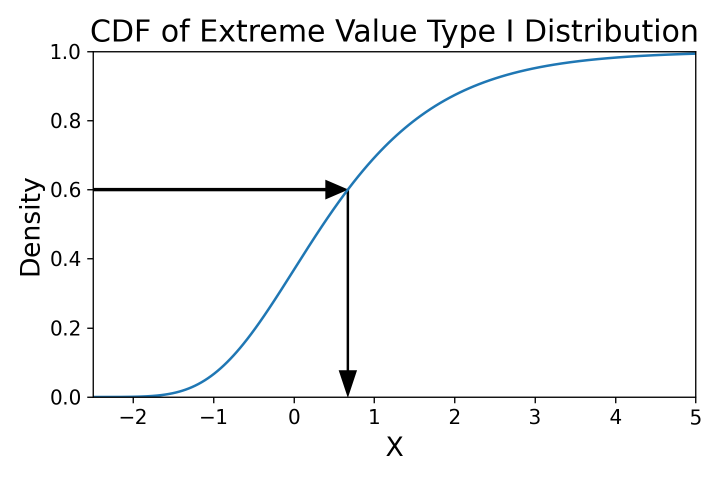


**圖 4**　標準常態分布等機率區間

分析方法

極端值第一型分布機率密度函數和累積分布函數如下式，

極端值第一型分布樣本可以透過產生0 ~ 1的均勻分布樣本(即累積機率)與利用極端值第一型分布累積分布函數之反函數轉換獲得，如下式與圖 5所示。

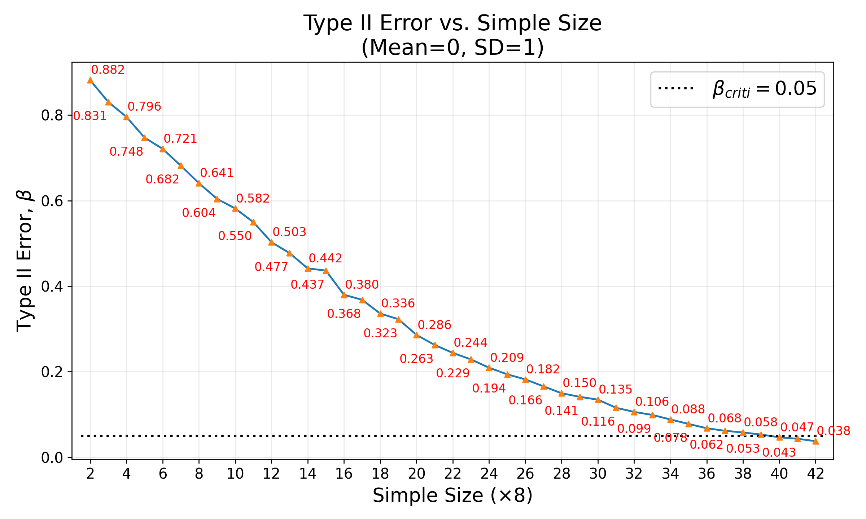


**圖 5**　極端值第一型分布樣本產生示意圖

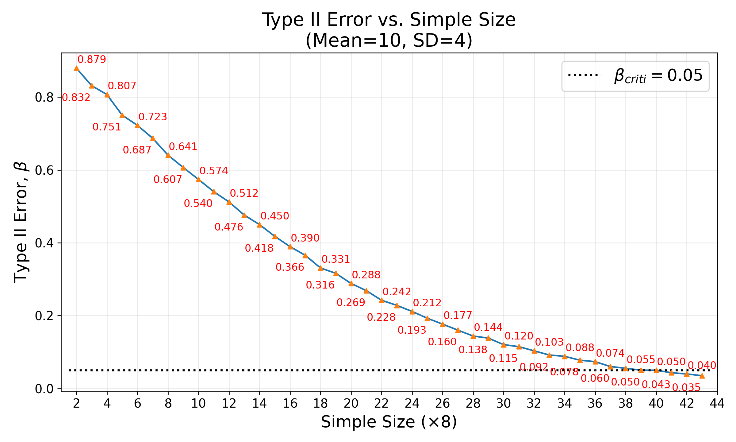
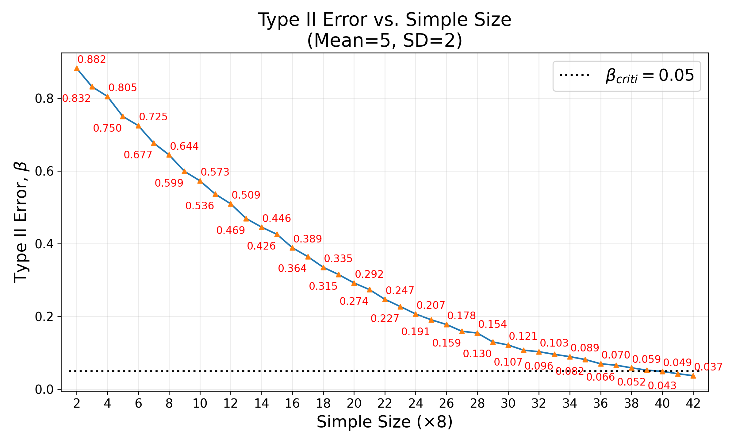
假設樣本平均值為0、標準差為1，利用scipy.stats.uniform.rvs()產生size = (10000, 8×*n*) (*n* = 2,3,4,…)的均勻分布樣本，再透過scipy.stats.gumbel\_r.ppf()(即累積分布函數之反函數)將其轉換為(10000, 8×*n*)的極端值第一型分布樣本，對不同樣本數(8×*n*)的樣本進行10,000次卡方檢定，並計算第二型錯誤發生機率，詳細程式請參見附錄。

結果分析

不同樣本數(8×*n*)樣本10,000次卡方檢定，第二型錯誤發生機率隨樣本數變化曲線如下圖 6，可以發現在樣本增加初期，第二型錯誤發生機率下降較快，而要使β< 0.05連續3次，*n*需要增加至42或43，不同參數假設(圖 7)下獲得的結果一致。



**圖 6**　第二型錯誤發生機率β隨樣本數變化曲線



**圖 7**　不同參數假設第二型錯誤發生機率β隨樣本數變化曲線

問題分析

五個溫度測站O(0,0)、A(20,10)、B(25, 50)、C(-80、30)、D(-20, -60)，日均溫均為常態分佈的隨機變數，期望值均為30度、標準偏差均為3度，相關係數為距離的函數*ρ*(*d*) = *exp*(−*d* / 30)。

**第A小題**

**某日O站和D站日均溫分別為33.3度和29.7度，請「最佳化估計」A、B、C三站的溫度，計算三個估計值不確定性的變異數，及、、是否與、、相同。**

多元常態分布可以下列矩陣形式表示，

給定部分隨機變數觀測值時，可以將***Z***分拆為兩部分，以下列矩陣形式表示，

其中，條件機率的平均值為給定***X***時***Y***的最小方均誤差估計量(minimum-mean-square- error estimator)，即給定***X***時***Y***的最佳估計，而為***X***時***Y***的共變異數矩陣，其對角線上的值即為估計值不確定性的變異數。

相關係數計算公式為下式，給定部分隨機變數觀測值時，剩餘隨機變數共變異數矩陣()也會改變，因此相關係數也可能會隨之改變。

**第B小題**

**若某日O站缺測，擬用 ，*i* = A、B、C、D補遺估計，決定四測站的權重係數值*wi*。**

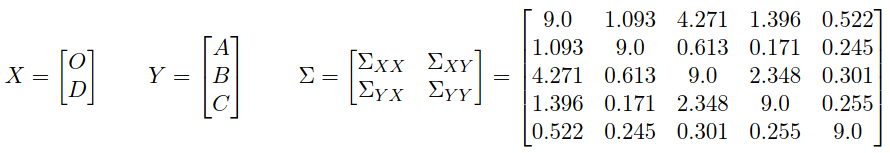
已知，i = A, B, C, D為給定四測站觀測值時O站的最小方均誤差估計量，可以透過將轉換為，求得A、B、C、D四測站的權重係數值*wi*，

可以發現上式的估計式比原始估計式多出一項，是因為上式並不是不偏估計(Unbiased estimator)，若加入(unbiased condition)條件，則與原始估計式相同，但需引入拉格朗日乘數(Lagrange multiplier)求解權重係數*wi*。

分析方法

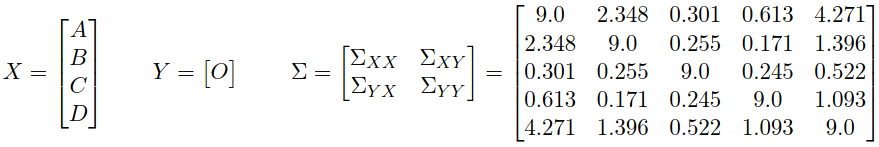
**第A小題**

O、A、B、C、D五站日均溫平均值皆為30度、標準偏差為3度，透過計算共變異數矩陣如下式，



給定O、D兩站日均溫，利用下列三式計算A、B、C三站條件機率平均值、共變異數矩震和相關係數矩陣。

**第B小題**

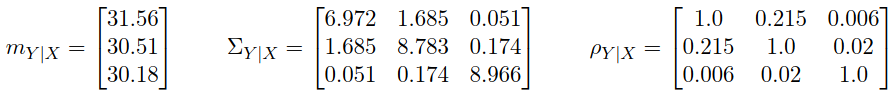


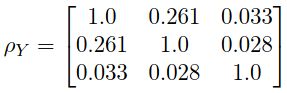
O站缺測，給定A、B、C、D四站觀測值估計O站日均溫，利用下式計算各站權重係數(僅使用最小方均誤差條件)。

結果分析

**第A小題**

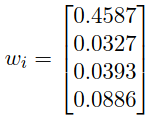
A、B、C三站最佳估計、估計不確定性變異數(共變異數矩陣對角線值)、相關係數等計算結果如下式，其中與並不相同。





**第B小題**

A、B、C、D各站權重係數(僅使用最小方均誤差條件)計算結果如下式，權重和不等於1。



1. <https://www.stat.auckland.ac.nz/~wild/ChanceEnc/index.shtml> [↑](#footnote-ref-1)