**Homework 1-1**

生工所 李世耀 F08622011

問題分析

以2021波士頓市長選舉第一輪投票結果，假設進入第二輪投票的兩位候選人支持率：

虛無假設H0：吳弭支持度*p* = 0.598，A.E. George支持度*q* = 0.402。

替代假設H­1：吳弭支持度*p* = 0.598±0.03，A.E. George支持度*q* = 0.402±0.03。

**第A小題**

**以民調隨機抽樣150人進行檢定，在雙尾檢定型一錯誤發生機率的條件下，計算拒絕H0的的人數範圍，以及真實支持度為*p* = 0.598±0.03發生型二錯誤的機率。**

民調抽樣150人可視為進行150次成功機率為*p* (即吳弭支持度)的獨立伯努力試驗，

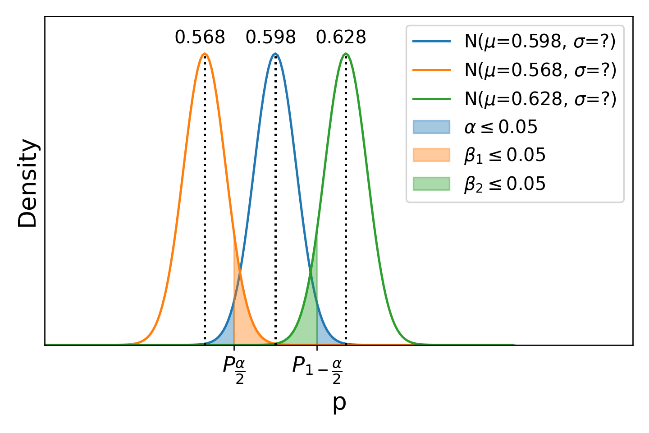
令*S*為150次試驗成功次數(即支持吳弭人數)，其服從二項式分布，當抽樣人數*n*足夠大時，依中央極限定理，*S*的分布近似於平均值為*np*、標準差為*np*(1－*p*)的常態分布，則

吳弭支持度估計值為：

**第B小題**

**在真實支持率和假設的差異在±3%，且符合雙尾檢定型一錯誤發生機率、型二錯誤發生機率要求下，民調抽樣人數應為多少人，以及其拒絕虛無假設的人數範圍。**

由第A小題的分析可知，當抽樣人數增加時，則支持度估計值服從的常態分布標準差變小，替代假設*p* = 0.598±0.03對應的分布與虛無假設接受域重疊部分減少(下圖橘色與綠色區域的面積)，即型二錯誤發生機率(下圖中與)下降。



**圖 1**　型一錯誤與型二錯誤示意圖

**第C小題**

**利用民意調查新論的近似法計算比較。**

《民意調查新論》中提到當樣本數量大時，抽樣誤差為，若估計方式是採百分比，則樣本百分比的標準差會隨百分比變動但不超過，因此樣本百分比抽樣誤差可表示為：

分析方法

**第A小題**

虛無假設吳弭支持度*p* = 0.598，抽樣人數*n* = 150人，支持度估計值分布近似於：

其對應百分比與百分比的數值為：

拒絕虛無假設H0的人數範圍即可由抽樣人數乘上與獲得。或使用非參數化檢定，假設150次獨立伯努力試驗成功()的機率為*p*，統計值*T*為樣本中成功的次數(即支持吳弭的人數)，統計值*T*不通過檢定的上下界為：

而型二錯誤發生機率即計算當真實支持率為*p* = 0.598±0.03時，支持吳弭的人數*s*落在接受域範圍的機率，即：

**第B小題**

如圖 1，分別計算符合雙尾檢定型一錯誤發生機率、型二錯誤發生機率對應的閾值：

*p* = 0.598，Type I error

下界：

上界：

*p* = 0.568，Type II error ：

*p* = 0.628，Type II error ：

分別解、獲得對應抽樣人數，拒絕H0的人數範圍則利用A小題方法。

**第C小題**

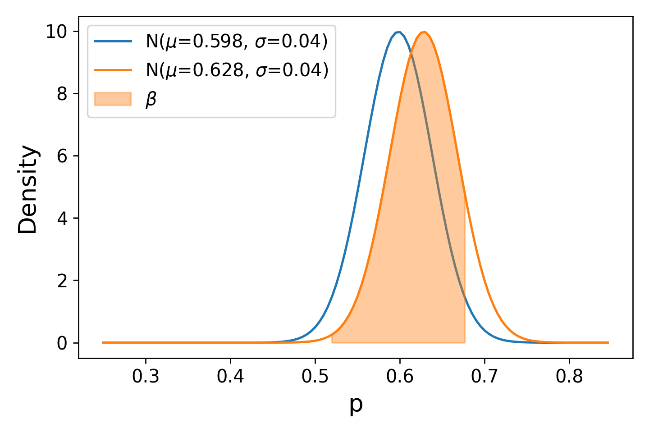
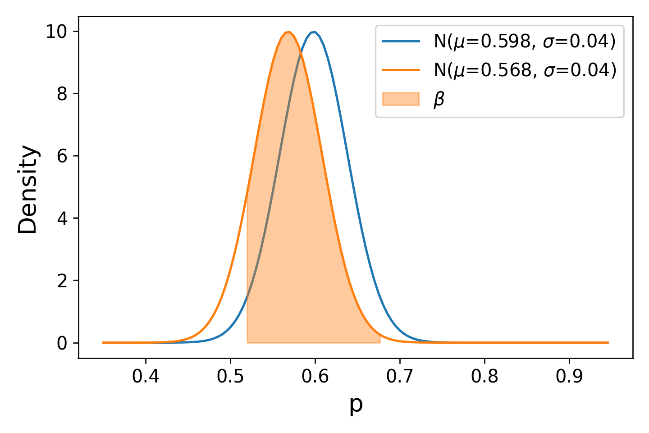
抽樣誤差設定為正負3個百分點，利用民調《民意調查新論》方法計算抽樣人數：

結果分析

**第A小題**

因此，拒絕H0的人數範圍為0 ~77 (< 78)或102~150 (> 101)，利用非參數檢定方法可以獲得相同的結果：

而型二錯誤發生機率分別為：



**圖 2**　真實支持度*p* = 0.568與*p* = 0.628第二型錯誤發生機率

**第B小題**

拒絕H0的人數範圍分別為0 ~ 2039或2153 ~ 3505(*p* = 0.568)與0 ~ 1993或2106 ~ 3427(*p* = 0.628)。

**第C小題**

以民調《民意調查新論》方法計算抽樣人數，應抽樣人數為，此時拒絕H0的人數範圍為0 ~ 632或698 ~ 1112，而型二錯誤發生機率為：

問題分析

以氣象局臺北站1961~2009年7月日最高溫資料分析「氣候變遷」現象是否統計顯著。

虛無假設H0：49年共1519筆資料來自**相同**分布，即「氣候變遷」現象**無**統計顯著。

替代假設H­1：49年共1519筆資料來自**不同**分布，即「氣候變遷」現象**有**統計顯著。

**第A小題**

**以卡方檢定，顯著水準α= 0.05，判斷1519筆7月日最高溫資料是否通過常態分布的虛無假設。**

卡方適合度檢定(Goodness of fit test)虛無假設為樣本來自於假設的機率分布，其檢定統計值為：

其中*k*為一次隨機試驗可能結果的互斥類別(事件)個數，*ni*和*ei*分別為在*n*次試驗中第*i*類內觀察到的次數與假設分布的期望次數，當*n*趨於無限大時，統計值T分布趨向自由度為*k*－1的卡方分布 (若假設分布的參數需由樣本估計，則自由度為*k*－*m*－1，*m*為參數數量)。當統計檢定值，則拒絕虛無假設。

**第B小題**

**排序並找出1961~2000年7月日最高溫低門檻值與高門檻值(前後2.5%)，計算2001~2009年7月日最高溫小於低門檻值與大於高門檻值的百分比，是否能判斷7月日最高溫紀錄存在氣候變遷。**

若前30年與後9年資料來自相同分布(虛無假設)，則兩者落在低門檻值以下和高門檻值以上的資料數量百分比會接近2.5%。若存在氣候變遷(以溫度升高說明)，則後9年資料小於低門檻的百分比會下降(< 2.5%)而大於高門檻的百分比會上升(> 2.5%)，即後9年資料往高溫偏移，前30年與後9年資料來自不同分布。兩個百分比只能說明可能存在氣候變遷，然而是否具有統計顯著則需要進一步討論。

**第C小題**

**利用蒙地卡羅法模擬10,000組、每組1519筆彼此互相獨立的標準常態分布資料，依循第B小題計算10,000筆小於低門檻值與大於高門檻值的百分比。**

為檢定第B小題獲得的百分比是否顯著偏離2.5%，假設資料符合常態分布，利用蒙地卡羅法產生10,000組、每組1519筆來自相同分布(虛無假設)的資料，將每組資料分為前1240筆和後279筆，找出前1240筆低門檻值與高門檻值並計算後279筆小於低門檻值和大於高門檻值的百分比，兩個百分比具有隨機性為隨機變數，透過上述的方法可以產生兩個百分比的10,000筆樣本值，兩個百分比樣本平均值應會接近2.5%。

**第D小題**

**利用第B小題與第C小題的結果判斷，能不能說明「7月日最高溫紀錄是否存在氣候變遷」。**

第C小題產生的10,000筆百分比樣本可以分別找出「小於低門檻值的百分比」對應α/ 2= 0.025的百分比門檻值(*LBL,M*, *UBL,M*)與「大於高門檻值的百分比」對應α/ 2= 0.025的百分比門檻值(*LBU,M*, *UBU,M*)，和第B小題得到的百分比(*LBO*, *UBO*)進行比較：

虛無假設H0：1519筆資料來自相同分布(P)且呈常態分布(Q)，即氣候變遷現象無統計顯著。

1. 若且，則不拒絕虛無假設，即說明「氣候變遷」現象無統計顯著。
2. 若或或或，則拒絕虛無假設，但因第A小題拒絕常態分布的虛無假設，P∩Q→R, ~R→~P∪~Q，故也無法說明氣候變遷現象是否統計顯著筆資料來自不同分布(~P)。

分析方法

**第A小題**

進行卡方檢定前需將資料分類為互斥事件，分類方式分可為等間距方法或等機率方法，為避免部分類別中樣本數過少，因此採取等機率分類方法(Romanowski et al., 1978)[[1]](#footnote-1)。類別個數*k*則由下式決定(D’Agostino, 1986)[[2]](#footnote-2)，詳細程式碼與類別內樣本數請參見附錄。

因為常態分布的參數未知，卡方檢定自由度為*k*－*m*－1，m＝2 (平均值、標準差)。

**第B小題**

依題意利用numpy.quantile()找到1961~2000年1240筆資料的低門檻值與高門檻值，並計算2001~2009年279筆資料小於低門檻值或大於高門檻值的百分比。

**第C小題**

利用scipy.stat.norm.rvs()產生樣本大小為(10000, 1519)彼此互相獨立的標準常態分布資料，依題意計算10,000筆「小於低門檻值的百分比」與「大於高門檻值的百分比」的樣本值，且找到兩個百分比各自的低門檻值與高門檻值(前後2.5%，一樣利用numpy.quantile())。

**第D小題**

利用第B小題和第C小題結果依問題分析說明進行比較。

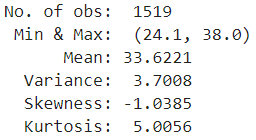
**第E小題**

將常態分布虛無假設改為皮爾森第三型分布假設，重複第A小題至第D小題。

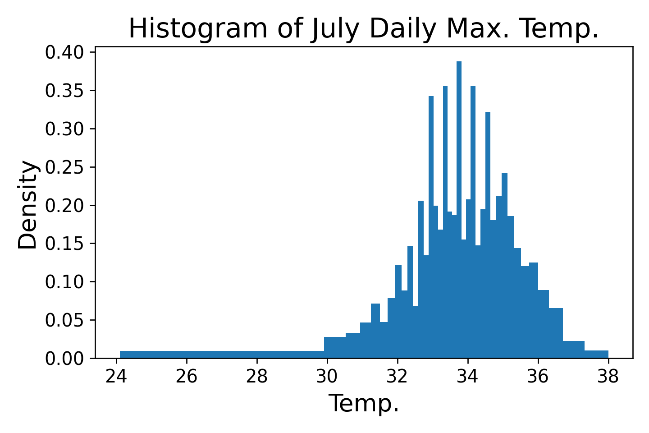
結果分析

**第A小題**

原始資料經前處理取出歷年7月日最高溫資料，資料基本統計值如下**圖 3**。資料依等機率方法分為37組繪製直方圖如圖 4。



**圖 3**七月日最高溫資料(1519筆)統計值



**圖 4**七月日最高溫資料直方圖

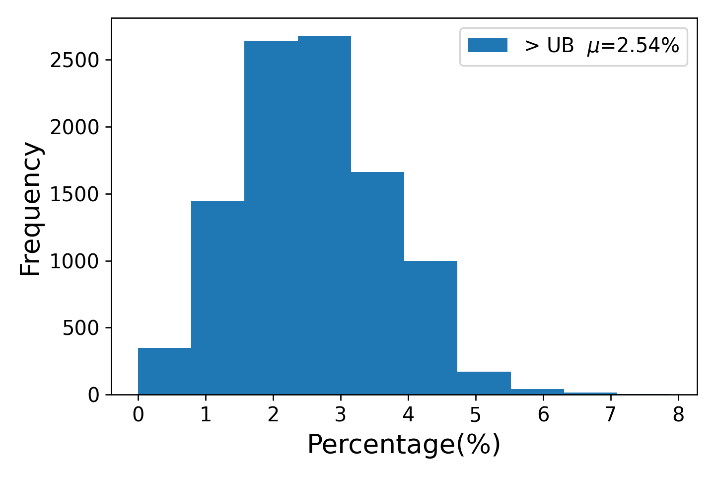
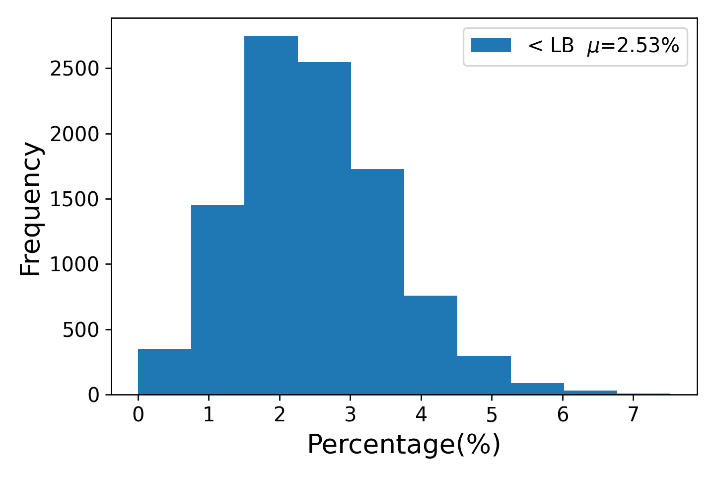
卡方檢定測試統計值為290.079，而自由度*k*－*m*－1 = 34的卡方檢定門檻值為48.602，因，拒絕常態分布的虛無假設。

**第B小題**

1961~2000年7月日最高溫低門檻值與高門檻值分別為**28.4℃**與**36.6℃**，2001~2009年279筆資料小於低門檻值與大於高門檻值的百分比分別為**0.717%**、**3.584%**，由這兩個百分比(< 2.5%與> 2.5%)乍看之下7月日最高溫可能存在氣候變遷，但是否有統計顯著並無法說明，需透過第C小題更進一步討論。

**第C小題**

10,000筆「小於低門檻值的百分比」與「大於高門檻值的百分比」的直方圖如圖 5，兩者平均值都接近2.5%，高低門檻值皆為[0.717%, 4.659%]。



**圖 5**　「<低門檻值百分比」與「>高門檻值百分比」樣本直方圖

**第D小題**

由第B小題與第C小題結果：

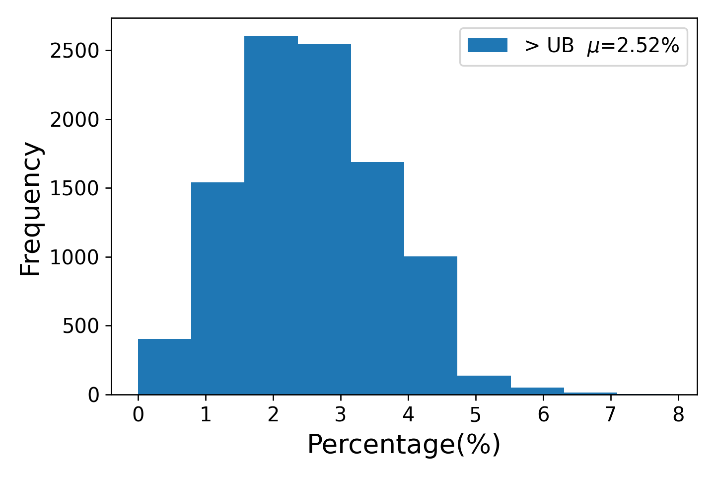
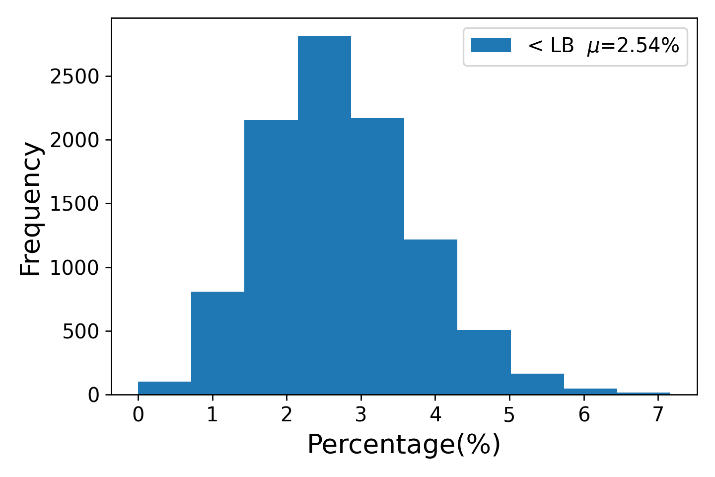
且

因此不拒絕虛無假設，即說明「氣候變遷」現象無統計顯著。

**第E小題**

假設1961~2000年7月日最高溫1240筆資料呈皮爾森第三型分布進行卡方檢定，一樣使用等機率分類方法將1240筆資料分為37組(自由度為*k*－*m*－1 = 33)，測試統計值為246.674，卡方檢定的門檻值為47.4，因，拒絕皮爾森第三型分布的虛無假設。

利用scipy.stat.pearson.rvs()產生樣本大小為(10000, 1519)彼此互相獨立且參數值與1240筆相同的皮爾森第三型分布分佈資料，計算10,000筆「小於低門檻值的百分比」與「大於高門檻值的百分比」的直方圖如圖 6，兩者平均值都接近2.5%，高低門檻值與第C小題相同皆為[0.717%, 4.659%]，因此與第D小題結論相同。



**圖 6**　「<低門檻值百分比」與「>高門檻值百分比」樣本直方圖 (Pearson Type III)

1. Romanowski, M., McConnell, R. K., &Halpenny, J. F. (1978). The equiprobable interval method of sample classification and the validity of the χ2 test. Metrologia, 14(4), 185–187. [↑](#footnote-ref-1)
2. D’Agostino, R. B. (1986). Goodness-of-fit-techniques (Vol. 68). CRC press. [↑](#footnote-ref-2)