110學年度高等水文分析作業(一)

第一部分－統計假設檢定 (10 or 12分)

個別作業，請於10月14日(星期四) 3am以前將電子檔繳交至CEIBA網站
請1-3組每位同學準備簡報，10月14日8:10抽籤，決定哪兩位同學第一節課報告成果

一、 繳交說明：

(1) 作業繳交時間原則上為星期一的3am以前（本作業例外）。

(2) 遲交分數打折以週為單位，若未準時繳交，但遲交時間不滿一週者，分數打九折；遲交時間超過一週、未達兩週者，分數打八折；依此類推。

(3) 若有作業抽換情形，即以最後一次抽換上傳CEIBA的時間，當作是繳交時間。

(4) 如有部分題目準時繳交、部分遲交者，準時上傳CEIBA部分分數不打折，遲交部分（郵寄給老師＋助教）按照打折原則辦理。

二、 作業要點：

繳交各次作業的內容，都建議採用三段式報告方式呈現：

(1)「問題分析」或「問題理解」之文字說明；

(2)「分析方法」，包含算式(algorithm)和電腦程式(若不足一頁可放在分析，超過一頁建議放在附錄)；

(3)「結果分析」之圖、表與文字說明。

其他建議注意要點包括：

(4) 作業勿附大筆數據以增加份量，任何20筆以上的數字均請儘量用圖形表示，否則扣分。

(5) 圖形顯示時，請注意說明座標軸的意義及單位。

三、作業內容：

1. 2021年美國Boston市長選舉，9月14日第一輪投票選出兩位候選人，分別是吳弭(Michelle Wu)和Annissa Essaibi George；將於11月2日第二輪投票，由兩位初選領先者中選出市長[1]。第一輪投票率僅25%，其中，吳弭和Annissa的得票率分別為33.3%和22.4%，其餘44.3%的票投給了另外三位候選人。

   A. 假設第二輪投票吳弭和A.E. George的支持率分別為(33.3/0.557)%和(22.4/0.557)%（即 H_0：吳弭的支持率為 $p=0.598$，A.E. George的支持率為 $q=1-p=0.402$，只要達到法定的投票率後，無效票不計入分母），且決定民調隨機採樣人數為150人，在two-side Type I error $\alpha/2$ 都不大於0.025、α 不大於0.05的條件下，決定拒絕虛無假設的人數範圍。並在此條件下，計算吳弭和A.E. George的真實支持度分別為 $p=0.598\pm0.030$ 與 $q=0.402\pm0.030$ 的Type II error β。（2分）

   B. 若吳弭、A.E. George的真實支持率和假設的差異都不超過3個百分點，two-side Type I error $\alpha/2$ 都不大於0.025、Type II error的機率 β 也都不大於0.05的要求下，民調採樣人數應該是多少？以及各自拒絕虛無假設的範圍為何？（2分）

   C. 請再利用民意調查新論（該方法假設樣本數量大，乃使用常態分佈測試值近似法[2]）的近似法計算比較。（1分）

2. 分析中央氣象局臺北站1961到2009年，共49年的七月日最高溫紀錄，「氣候變遷」現象是否統計顯著[3]？

   A. 中央氣象局臺北站的溫度紀錄檔案為466920chkd.txt，請找出49年的七月「日最高溫」觀測值共49×31=1519筆；繪日最高溫樣本的histogram，利用卡方檢定、顯著水準 $\alpha=0.05$，判斷數據分布是否通過常態分布的虛無假設？（1分）

   B. 按照大小排序1961到2000年的1240筆「日最高溫」紀錄，找出觀測值最

---

[1] 吳弭和埃塞比・喬治贏得市長初選 成為波士頓選舉歷史上新時代，or Annissa Essaibi George Vs. Michelle Wu: Your Guide to Boston's Mayoral Election

[2] 講義 1-11 頁表示：若樣本數小於 20，則(如例題 1.1)利用二項(binomial)分佈計算。對於樣本數較大者，二項分佈逐漸趨近於常態分佈，可以採用常態分佈的測試值近似之。亦可參考維基百科：https://zh.wikipedia.org/wiki/二項式分布，or https://en.wikipedia.org/wiki/Binomial_distribution。

[3] 氣候與氣候變遷的定義：Climate is the pattern of weather that we observe geographically and over the seasons. And that's describe in terms of averaging, variation and probability. Thus, climate change is the change in average values, variation patterns or probabilities.

小的31筆和最大的31筆（31/1240＝0.025），採用序號31、32的平均值和1209、1210的平均值，當作接受或拒絕虛無假設的門檻值；計算2001年到2009年七月「日最高溫」紀錄小於低門檻值與大於高門檻值的百分比；請說明這兩個百分比，能不能判斷「七月日最高溫紀錄是否存在氣候變遷」現象？為什麼？（1分）

C. 請用蒙地卡羅法（根據多次重複試驗結果，估計某種事件出現機率的方法），產生10,000組、每組1519筆彼此相互獨立的標準常態分佈數據。依循B小題的方法，將每組的1519筆數據分為前1240筆和後279筆，用前1240筆排序計算$\alpha/2=0.025$的高、低門檻值，再統計後279筆中，小於低門檻值與大於高門檻值的百分比，共10,000組。分析10,000個小於低門檻值百分比的蒙地卡羅樣本，找出其中最小和最大的250個百分比值蒙地卡羅樣本，決定$\alpha/2=0.025$的百分比門檻值；同理，找出大於高門檻值的$\alpha/2=0.025$的百分比門檻值。（2分）

D. 判斷B小題得到的「小於低門檻值的百分比數據」與「大於高門檻值的百分比數據」，用C小題得到的結果來判斷，能不能說明「七月日最高溫紀錄是否存在氣候變遷」現象？為什麼？（1分）

E. 假設氣象局臺北站1961-2000年的1240筆七月「日最高溫」觀測值數據呈皮爾森第三型分布，利用動差法估計其三參數[4]，再利用卡方檢定、顯著水準$\alpha=0.05$，判斷其是否通過皮爾森第三型分布的虛無假設？其次，使用蒙地卡羅法產生10,000組、每組1519筆彼此相互獨立、參數值與1240筆相同的皮爾森第三型分布分佈數據。依循B小題方法，以每組的前1240筆，排序計算$\alpha/2=0.025$的高、低門檻值，再統計後279筆中，小於低門檻值與大於高門檻值的百分比，共10,000組。分析10,000個小於低門檻值百分比的蒙地卡羅樣本，找出其中最小和最大的250個百分比值蒙地卡羅樣本，決定$\alpha/2=0.025$的百分比門檻值；同理，找出大於高門檻值的$\alpha/2=0.025$的百分比門檻值。判斷B小題得到的「小於低門檻值的百分比數據」與「大於高門檻值的百分比數據」，用本小題得到的結果來判斷，能不能說明「七月日最高溫紀錄是否存在氣候變遷」現象？為什麼？（2分）

---

[4] 若偏態係數的絕對值$|c_s|$太接近 0，Matlab 或 R 無法計算$x_{40cs}=\gamma_{cs}+icdf('Gamma',0.975,\beta_{cs},\alpha_{cs})$，則 E 小題就不用做了，因為產生皮爾森第三型分佈數據的困難會太大，挑戰就會變成是「如何產生皮爾森第三型分佈數據」，而不再是設計問題的本意－統計測試了。

附註：

皮爾森第三型分佈的機率密度函數為：

$$f(x)=\frac{[|x-\gamma|/\alpha]^{\beta-1}\,\boldsymbol{exp}[-|x-\gamma|/\alpha]}{\alpha\cdot\Gamma(\beta)}$$ [5]

其中，α 是scale parameter，β 是shape parameter，γ 是position parameter。可以利用method of moment計算參數值 $\beta_{cs}=4/c_s^2$ ，$\alpha_{cs}=\sigma_x/\sqrt{\beta_{cs}}$ ，$\gamma_{cs}=\mu_x-\sigma_x\sqrt{\beta_{cs}}$ (if $c_s>0$) 或 $\gamma_{cs}=\mu_x+\sigma_x\sqrt{\beta_{cs}}$ (if $c_s<0$)。

若已知累積機率值 $F(x)$，計算皮爾森第三型分佈變數值 x 的方法，可以使用Matlab的正偏態係數Gamma分布[6]的累積機率函數值，再加減位置參數的方式計算。例如，重現期 $T=40$ 年、累積機率值 $F(x)=0.975$，對應的 x_{40} 數值的計算方法為：

若 $c_s>0$ ， $x_{40cs}=\gamma_{cs}+icdf('Gamma',0.975,\beta_{cs},\alpha_{cs})$ [7]

若 $c_s<0$ ， $x_{40cs}=\gamma_{cs}-icdf('Gamma',0.025,\beta_{cs},\alpha_{cs})$

若 $c_s=0$ ， $x_{40,cs=0}=icdf('Normal',0.975,\mu_x,\sigma_x)$ [8]

---

[5] 若偏態係數值在 $|c_s|<0.2$ 的範圍內，皮爾森第三型分佈機率密度函數的分子和分母都包含階乘，且兩者的數值都已經很大，分別計算的數值可能都會超過數值容量，無法再相除得到機率密度函數值，誤差可能也很大；解決技巧是使用 DO Loop，每次乘一個《分子階乘數值和分母階乘數值的比值》。

[6] 三參數皮爾森第三型分佈和兩參數 Gamma 分佈的主要差別是：當偏態係數大於 0，Gamma 分佈的下限值是 0，三參數皮爾森第三型分佈的下限值是 γ；當偏態係數小於 0，Gamma 分佈的上限值是 0，三參數皮爾森第三型分佈的上限值是 γ。故三參數皮爾森第三型分佈又稱為三參數 Gamma 分佈（比兩參數 Gamma 分佈多了位置的自由度，或多了一個位置參數），詳見水文學第 11 章講義最後兩頁的圖表。若有需要，也可參考 Matlab 的 Gamma Distribution 說明和維基百科的 Gamma Distribution 說明。

[7] 請參考 Matlab 的 `icdf` 網頁說明。

[8] 當偏態係數等於 0 時，皮爾森第三型分佈或 Gamma 分布退化成常態分佈，或說常態分佈是皮爾森第三型分佈在偏態係數等於 0 的特例。