統計期末報告

資工二秦紫頤 406410035 資工四曹維廷 404410058

資工四林楷博 404410077 資工四陳柏諺 404410081

1. 主題目: 台灣近三年空氣污染分析

次題目→

(1) 近三年全台各區域數據整理與統計圖像化

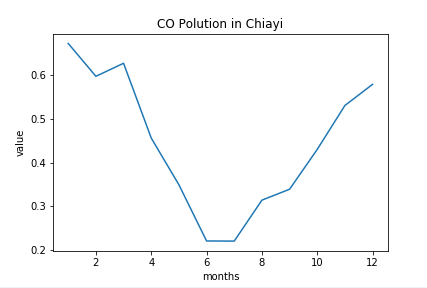
(2) 20年前(1998)與去年(2018) PM2.5汙染值比較

(3) 春季＋冬季的空氣污染是否有明顯比夏季＋秋季還嚴重

(4) 白天與夜晚的空污程度比較

1. 近三年全台各區域數據整理與統計圖像化

將各區域的數據中的空白值(NaN值)轉變成0以便處理，接著將數據中各污染值分別以24小時、8小時、每小時取最大值再做成日平均，再將日平均處理成月平均，如下圖：



1. 20年前與去年(2018) PM2.5汙染值比較
2. 資料蒐集: 若測站數不夠的區域符合Normal Distribution的話，會利用2.所得到的各測站的平均數，進行北中南東四個區域2018年和1998年PM2.5的變異數和T test，來驗證虛無假設。
3. 利用統計的資料分析:

(I) Normal Distribution的確立

先確定區域中，1998年和2018年中，測站數<30個區域，有符合Normal Distribution(目前須檢測的為中部(15個測站)、南部(25個測站)和東部(5個測站))。

→ (a.)先利用Pearson Coefficient看結果是否介於1和-1之間。

(b.)再利用Q1-1.5IQR和Q3+1.5IQR看是否有離群值存在。

(c.)若(a.)(b.)皆符合，即確定區域監測站數不影響結果。

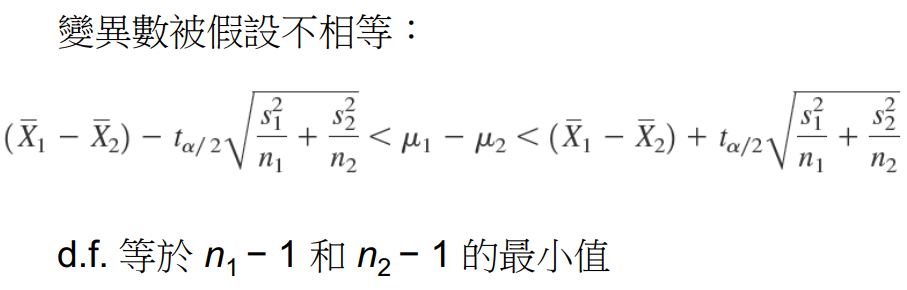
所需input: 1. 各個Sample(測站)的平均值。

2. 利用1.的資料，求出Sample的Standard Deviation s和Sample

Mean 。

3. 1.之中，所有數之間的Median。

(II) 利用9-2的T test進行統計分析



s1: 1998年的Sample之Standard Deviation, s2: 2018年的Sample之Standard Deviation, 1:1998年的Sample Mean, 2:2018年的Sample Mean, n1: 1998年該區域的測站數, n2: 2018年該區域的測站數。

虛無假設 H0: 北部地區1998年PM2.5汙染值μ1 = 2018年PM2.5汙染值μ2

對立假設 H1: 北部地區1998年PM2.5汙染值μ1 < 2018年PM2.5汙染值μ2

(同理，中部地區、南部地區及東部地區亦同) 。

若是最後區間中有包含0。則我們沒有足夠的證據認為2018年該區域的PM2.5汙染值與1998年該區域的PM2.5汙染值有顯著的差異。

反之，若是最後區間沒有包含0。我們則有足夠的證據支持2018年該區域的PM2.5汙染值與1998年該區域的PM2.5汙染值有顯著的差異。

1. 預期結果: 由四組分析去印證出全台灣各區在近20年的空氣汙染是否有越來越嚴重的趨勢。
2. 白天與夜晚的空污程度比較
3. 資料蒐集: 先將資料區分成白天及夜晚，設定6:00-18:00為白天，18:00-6:00為晚上，將所有測站的白天與夜晚的空污數值取平均，也會計算白天和夜晚AQI不健康以上的天數分別有幾天。
4. 利用統計資料分析:

我們會利用假設檢定的方式，檢定白天的PM2.5平均是否與夜晚的PM2.5平均值相等，還會檢定白天AQI大於不健康的比例是否與夜晚相等。

(I)PM2.5平均值:

利用t-test進行分析，公式如3.所示。

虛無假設 H0: 台灣白天PM2.5汙染值μ1 =台灣夜晚PM2.5汙染值μ2

對立假設 H1: 台灣白天PM2.5汙染值μ1 ≠ 台灣夜晚PM2.5汙染值μ2

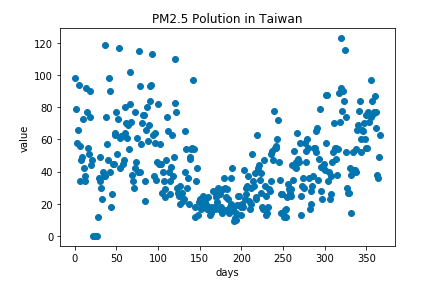
(II)AQI比例:

利用t-test進行分析，公式如上一項所示。

虛無假設 H0: 台灣白天AQI不健康以上的比例μ1=台灣夜晚AQI不健康以上的比例

對立假設 H1: 台灣白天AQI不健康以上的比例 ≠ 台灣夜晚AQI不健康以上的比例μ2

1. 預期結果: 預期結果為白天晚上空污程度應該差不多，若空污程度有差異，則會去尋找可能造成差異的原因，例如白天晚上的用電差異等等。
2. 春季＋冬季的空氣污染是否有明顯比夏季＋秋季還嚴重
3. 資料蒐集: 若測站數不夠的區域符合Normal Distribution的話，會利用2.所得到的各測站的平均數，進行北中南東四個區域，夏季＋秋季和春季＋冬季PM2.5的變異數和T test來驗證虛無假設。
4. 利用統計資料分析:



從統計圖中(上圖)可以隱約看出春季＋冬季的空污嚴重程度似乎好像遠高於夏季＋秋季，因此想實際上去驗證這是否屬實。

夏季＋秋季的PM2.5汙染值平均=μ1

春季＋冬季的PM2.5汙染值平均＝μ2

以下假設台灣4個地區（北、中、南、東）分別做探討

虛無假設H0: μ1 = μ2

對立假設H1: μ1 < μ2

實際上我們只有X̄1和X̄2，所以同3.必須先檢查是否符合normal distribution、利用t-test去預估μ1-μ2所位在的區間範圍。如果μ1-μ2所預估的區間有包含0那代表春季＋冬季和夏季＋秋季的空氣污染沒有顯著差異。但如果μ1-μ2的區間範圍界是落在小於0的數字，那此時我們的對立假設（春季＋冬季的空氣污染有明顯比夏季＋秋季還嚴重）就成立了。

1. 預期結果: 我們的預期結果是不管是哪個地區春季＋冬季的空氣品質會明顯比夏季＋秋季還差。
2. 目前進度
3. 近三年資料已依照測站在各縣市的所在地分類後，完成整理 (詳圖(a)和圖(b))

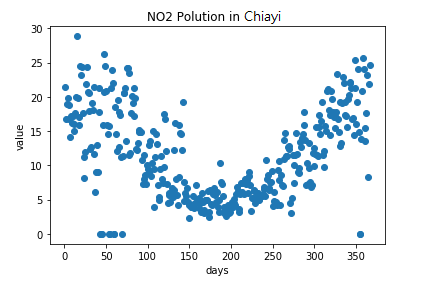


圖(a)



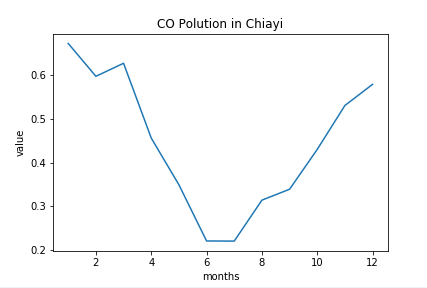
圖(b)

1. 資料程式碼整理並以天為單位，產生各縣市不同汙染物質的點圖分布 (詳圖(c))。



圖(c) 2018 Chiayi NO2的濃度隨著天數分布

1. 各個測站各月的平均數 (詳圖(d))



圖(d) 2018 Chiayi測站 CO的濃度各月的平均數