單元 3: 凝視 (檢視) 資料

本單元目標: 探討如何使用摘要統計量 (Summary Statistics) 與繪圖 (graphs) 對環境資料的描述與觀察.

一. 摘要統計量

摘要統計量 (又稱敍述統計量 (descriptive statistics)) 爲一些數字, 用以摘要出一組觀察值中的訊息, 也稱作樣本統計量 (sample statistics), 因爲它是由樣本所計算出的統計量; 無法描述出整個母體.

二種摘要 (或敍述) 統計量的分類法:

- 根據所度量的值. 如, 位置 (location) (顯示中央趨勢 (central tendency)), 擴散度 (spread) (顯示變異性 (variability)), 偏斜性 (skew) (顯示單向長尾 (long-tail in one direction)), kurtosis (顯示聳立性 (peakedness))等.

- 根據不尋常極端值顯現時的反應. 如, 敏感型 (sensitive) 或強韌型 (robust).

參考表 3.1, 根據此二種分類法有關摘要統計量的簡述.

1. 公式: 下述公式中以 $x_1, x_2, \cdots x_n$ 表示 n 個觀察值; 以 $x_{(1)}, x_{(2)}, \cdots x_{(n)}$ 表示排序後, 由小到大的觀察值.

1. 位置的度量 (Measures of Location) (中央趨勢 (Central Tendency)):

   (a) Mean (又稱 average):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

   - 對極端值是敏感的(由公式顯而易見)

   (b) Trimmed Mean:

$$\bar{x}_{trimmed} = \frac{1}{n - 2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-[\alpha n]} x_{(i)}$$

   其中 $0 < \alpha < 0.5$ (稱作修剪率(trimming fraction)), $[y]$ = 小於或等於 y 的最大整數, (亦即, $\bar{x}_{trimmed}$ = 修剪掉前後各 $[\alpha n]$ 個觀察值後的平均值).

- 因爲刻意地修剪掉極端值, 不如 mean 那麼 sensitive.

(c) Median:

簡言之, 就是 50% trimmed mean.

$$\text{median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{若 } n\text{: 奇數} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{若 } n\text{: 偶數} \end{cases}$$

- 各有一半的觀察值位於 median 的上, 下.
- 對極端值是非常強韌的 (因爲不論極端值是如何地偏, 不影響中間的觀察值).

(d) Geometric Mean:

通常用來描述正值的資料; 觀察值經對數轉換後, 求 mean, 再取指數.

$$\bar{x}_g = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right]$$

- 用來估計 lognormal 分布的眞正 median. (如, Reference area 的 TcCB 資料中, median 與 geometric mean 均爲 .54ppb.)

- $\bar{x}_g \leq \bar{x}$, "$=$" 成立僅當所有觀察值 x_i 均相等 (因爲幾何平均數 $\leq$ 算術平均數, 自行驗證).

- 如同 median, geometric mean 也是強韌的.

2. 擴散性的度量 (Measures of Spread) (變異性 (variability))

   簡言之, 擴散性小 $\Rightarrow$ 樣本中位數或樣本平均值會是相當具 "代表性" 的觀察; 擴散性大 $\Rightarrow$ 會有些比樣本平均值或樣本中位數過小或過大之觀察值. 是一種說明 "描述分布特質" 好的程度的工具.

   (a) Range:

   最大值與最小值之間的差，快速地提供觀察值差異程度的認知.

   $$\text{range} = x_{(n)} - x_{(1)}$$

   - 敏感的 (對極端值而言).

   (b) Interquartile Range (又稱 IQR):

   令 x_p 資料的第 $p100$ 百分位數 ($p100^{\text{th}}$ percentile), 亦即, (簡言之) 爲一數使得約

$p100\%$ 的觀察值小於此數且約
$(1-p)100\%$ 的觀察值大於此數. (如, median =第 50 百分位數 $= x_{.5}$)

$25^{\text{th}}, 50^{\text{th}}$ 與 75^{th} percentile 又分別稱作 1st, 2nd 與 3rd 四分位數 (quartile).

Interquartile range 就是前後兩個四分位數間 (亦即, 75^{th} 百分位數與 25^{th} 百分位數) 的差距, 度量資料中間 50% 的範圍, 即

$$\text{IQR} = x_{.75} - x_{.25}$$

- 對極端值是強韌的.

(c) Variance:

觀察值與 mean 之間差距的平方的平均值.

$$S^2_{mm} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(稱作 動差法估計量, method of moment estimator)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(常用的不偏估計量 (unbiased estimator))

- 對端值是相當敏感的, 程度更甚於平均值 (因爲涉及觀察值與平均值的距離平方). (如, TcCB 資料中, Reference area 的 variance = .08ppb^2, 而 Cleanup area 的 variance =400ppb^2.)

(d) Standard Deviation:

variance 的平方根.

$$S_{mm} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(對應於動差法估計量)

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(常用的標準差, 對應於不偏估計量)

- 比 variance 更受歡迎的擴散性度量值 (因爲與原資料有相同的度量單位).

- 對極端值是相當敏感的.

(e) Geometric Standard Deviation:

通常用來描述正值的資料; 觀察值經對數轉換後, 求標準差, 再取指數.

$$S_g = e^{S_y}$$

其中

$$S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$y_i = \log(x_i),\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- 異於 $\bar{x}_g$ (geometric sample mean $\approx$ lognormal 分佈的 median), 不能用來估計 lognormal 分布的任一母體參數.

(f) Median Absolute Deviation (簡寫成 MAD):

觀察值與 median 之間差距的 median.

$$\begin{aligned} &\text{MAD} \\ &\quad = \ \text{median}(|x_1 - m|, \cdots, |x_n - m|) \end{aligned}$$

其中 $m = \text{median}(x_1, x_2, \ldots, x_n)$.

- 對極端值是強韌的 (異於 variance 與 standard deviation 之處).

(g) Coefficient of Variance (變異係數) (簡記成 CV):

標準差對平均值的比值.

$$\text{CV} = S/\bar{x}$$

- 無單位的度量, 描述分布相對於平均值的擴散性 (亦即, 每單位平均值的偏離度).
- 通常用來描述正, 右偏斜分布 (如, lognormal) 的特質.
- 如同 mean 與標準差, 對極端值是敏感的.

3. 偏離對稱 (或鐘形) 分布的度量 (Measures of Deviation from a Symmetric or Bell-Shaped Distribution):

度量資料與對稱 (或鐘形) 直方圖 (histogram) 的偏離程度. 對稱 (或鐘形) 直方圖是用來判斷資料可用常態 (normal, Gauss) 分布模型化的一個好的指標, 許多統計假設檢定均假設資料是取樣於常態的母體分布, 故此度量有其必要性. 在簡易地以電腦軟體繪圖與做配適度檢定 (goodness-of-fit test) 前, 常以二統計量:

skew 與 kurtosis, 做爲此種度量的報告 (雖然目前不廣爲採用).

(a) Skew (又稱偏斜係數, coefficient of skewness):

''觀察值與平均值差距的三次方的平均值'' 對 ''標準差三次方'' 的比值.

$$\text{Skew}_{mm} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 / S_{mm}^3$$

(對應於動差法估計量)

$$\text{Skew} = \frac{\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3}$$

(對應於不偏估計量)

- 無單位的值, 度量觀察值在 mean 附近分佈的情形.

  若直方圖

  – 相當對稱 $\Leftrightarrow$ skew 爲 0 或靠近 0.

  – 有數個大的值位於 mean 的右邊 (但不在 mean 的左邊) $\Rightarrow$ skew $>$ 0 且稱作右

偏斜 (right skewed) 或正偏斜 (positively skewed).

– 有數個小的值位於 mean 的左邊 (但不在 mean 的右邊) $\Rightarrow$ skew <0 且稱作左偏斜 (left skewed) 或負偏斜 (negatively skewed).

- 環境資料通常爲右偏斜 (正偏斜) (因爲環境資料常涉及度量化學物的濃度, 而濃度不會低於 0).

(b) Kurtosis (又稱作聳立係數, coefficient of kurtosis):

''觀察值與平均值的差距的四次方的平均值'' 對 ''標準差四次方'' 的比值.

$$\text{Kurtosis}_{mm} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 / S_{mm}^4$$

(對應於動差法估計量)

- 無單位的值, 度量 ''資料的直方圖'' 相對於 ''理想鐘形的直方圖'' 的聳立程度.

  理想鐘形的直方圖是根據常態 (高斯) 分布而得, 其 kurtosis 爲 3.

若資料的直方圖相對於理想直方圖

- 有許多觀察值位於尾部 (例如, 厚長的尾巴), 則 kurtosis $>$ 3.
- 有短的尾巴且大部份的觀察值均緊密地群聚在 mean 附近, 則 kurtosis $<$ 3.

2. 例子: 透過 S-PLUS 處理內建資料, 以了解各種摘要統計量.

1. 以資料架構 (data frame) `epa.94b.tccb.df` 展示 TcCB 變量 (variable) 的摘要統計量. (Cleanup 中的 "ND" 表示無法偵測, nondetect.)

2. 製造對數轉換後的 TcCB (命名爲 `log.TcCB`), 並加入原先的資料架構, 再命名成 `new.epa.94b.tccb.df`. 以此展示變量 `log.TcCB` 的摘要統計量. (註: 不計算 $\bar{x}_g$ 與 $\bar{S}_g$ (因爲資料爲負值).)

二. 單變數的圖形

1. 點狀圖 (Dot Plots)

以每一度量個別的標籤展示量化的資料; 也可以根據類別變數 (categarical variable), 展示各類觀察值所占的百分比. 資料由小到大排序後的圖形更能加強認知.

例. 美國東北部 41 個城市有關臭氧 (ozone) 的訊息分別在向量 (vectors): `ozone.city`, `ozone.median`, `ozone.quartile` 以及字串 (list): `ozone.xy` (內含經度 (longitude) 與緯度 (latitude)) 內.

- 製造一新的資料架構 (命名爲 `ozone.NE`) 內含上述訊息.
- 繪出緯度大於北緯 42 度的 13 個城市的 dot plot.
- 以指令 `ozone.NE <- cbind(ozone.NE, City.Ordered = reorder.factor (ozone.NE$City, ozone.NE$Median))` 加

入一個新的變量 (variable): `City.Ordered`, 並以此繪出有序的 (ordered) dot plot.

註. `attach(ozone.NE)` 後, `reorder.factor` 內就可直接用 `City` 與 `Median`, 之後再 `detach()`.

2. 長條圖 (Bar Charts)

如同 dot plot, 以每一度量個別的標籤展示出量化的資料; 也可根據類別變數展示各類觀察值所占的比例. bars 的長度會有一種附加上去的視覺效果; 當基線 (baseline) 不爲 0 時, bars 的長度並不代表實際的長度, 僅顯現觀察值相對大小的關係.

例. 繪出緯度 ≥ 42 度的城市的 bar chart. 指令: `barchart(City.Ordered ~ Median, data=ozone.NE, Subset=Latitude >= 42)`

3. 圓形圖 (Pie Charts)

用以展示及比較百分比. 圖形認知的實驗證明 pie charts 在傳遞認息上不如 dot plot 與 bar chart 來的可信.

例. 做爲實驗室間比較的銀濃度的資料中, 有 56 個觀察值, 其中 34 個以 "<DL" 的形式呈現出, 此處 DL 表示偵測極限 (detect limit). (以 "<DL" 形式呈現的觀察值又稱作缺失觀察值 (censored obervations) (此資料中, 共有 12 個不同的偵測極限 (detection limit))

被記成 "<DL" 的觀察值意謂著針對此特定的實驗室, 實體樣本, 及儀器設定, 化學物的濃度僅能被量化爲小於 DL 的一值, 亦即, 任一介於 0 與 DL 間的一數.

此資料存放於資料架構 helsel.cohn.88.silver.df 中, 內含變量 Ag.orig (爲一文字向量 (character vector), 是最原始的編碼值), Ag (爲一數字向量 (numeric vector), 是銀的濃度, 將 censored 值編碼成 censoring level (即, DL)) 以及 Censored (爲一邏輯向量 (logic vector), 指出一觀察值是否爲 censored).

處理程序如下:

- 加一個變量 Censoring.Level 以指出 censoring level (若此觀察值是 censored), 或 "Not Censored" (若此觀察值是 "not

censored"), 之後命名此資料架構爲 `new.helsel.cohn.88.silver.df`.

- 製造一命名爲 `silver.table` 的資料架構.

  註: 以指令方式執行 `reorder.factor`.

- 繪出 `silver.table` 的 pie chart.

- 繪出 `silver.table` 的 dot plot.

(自行參考課本的選單程序以及必要的指令, 實際操作以得到上述的結果.)

4. 細條圖 (Strip Plots)

   又稱作一維散播圖 (one-dimensional scatter plot), 是一種將每一觀察值展現的圖示, 可用來觀看一個資料集合的分布, 或比較二個或以上個資料集合的分布.

   例. 展示資料架構 `new.epa.94b.tccb.df` 中 Reference 與 Cleanup area 的對數轉換後之 TcCB 資料. 很明顯地, 可看出在清除區域中少數

的幾個離群值 (outliers), 它們可能是在清理 (修補、改善) (remediation) 過程中遺漏掉的熱點 (hot spots).

5. 直方圖 (Histograms)

異於 strip plot, 不展示每一個觀察值, 而是透過將觀察值區分在不同的區間 (又稱作 class 或 bin) 內, 並計算在每一區間內的數量的方式, 來摘要出資料的分布. y-軸可爲下列 4 種尺度之一:

- 數量 (number or counts): 長條 (bar) 的高度 = 在區間內的觀察值的個數.
- 百分比 (percent of total): 長條的高度 = 在區間內的觀察值占所有觀察值的百分比.
- 比率 (fraction of total): 以介於 0 與 1 之間的比率, 呈現出區間內的觀察值占總體的比率.
- 密度 (density): 長條的高度 $\times$ 區間的寬度 = 區間內的觀察值占總體的比率.

histogram 的外形取決於區間的寬度:

- 區間的寬度愈來愈小時, 會導致外形愈來愈殘破, 而趨向一 strip plot.

- 區間的寬度愈來愈大時, 會提供愈來愈少的訊息, 而趨向於只有一個區間, 含蓋所有觀察值的圖形.

- 選取適當的區間以免失去過多的準確性, 但卻沒有一個固定的規則可依循.

例. 繪 `new.epa.94b.tccb.df` 中對數轉換後的 TcCB 的 histogram (取區間寬度爲 0.3 或總區間數等於 25).

6. 密度圖 (Density Plots)

根據樣本觀察值, 求得的一個有關樣本背後的眞實機率分布 (true underlying probability distribution) 的估計; 故 x-軸上二點之間曲線下的面積會等於一觀察值落於此二點間的機率的估計.

例. 繪 Reference 區域的對數轉換後的 TcCB 的 density plot.

注意: strip plot 中在 0 附近有一間隔 (gap), 此現象在 density plot 中也呈現出; 卻在 histogram 中被隱藏掉 (無法顯示此一訊息).

7. 盒型圖 (Boxplots)

又稱作 box-and-whisker plot, 一種摘要一組觀察值的簡單圖形展式, 內含

- 一盒子 (box), 由 25^{th} 與 75^{th} 百分位數所定義的.
- 在盒子上以一線或一點標示 median (或 50^{th} 百分位數).
- 由 25^{th} 百分位數開始往下劃一條虛線 (鬍鬚, whisker) 直至在一個 步 (step) 內的最小觀察值. (註: 一步定義成 interquartile range 的 1.5 倍) (interquartile range = 75^{th} 與 25^{th} 百分位數間的距離) (稱作下鄰近值, lower adjacent value).
- 由 75^{th} 百分位數開始往上劃一條虛線 (whisker) 直至一步內的最大觀察值 (稱作上鄰近值, upper adjacent value).

- 盒子兩端一步以外的觀察值, 以符號 (如, $*$) 或直線呈現之 (稱作外面值, outside values).

一 boxplot 可以很快地呈現出下列特徵:

- 資料的中心 (median)
- 資料的擴散性 (variability)
- 資料的偏斜性 (透過觀看半個盒子的相對長度以及 whiskers)
- 任何不尋常 (outside) 的值 (註: outside values 不必然是離群值或壞的觀察值; 事實上, 由於環境資料多爲右偏斜, 在 boxplot 中會經常出現許多的外面值.)

例. 繪 `new.epa.94b.tccb.df` 中經過對數轉換後的 TcCB 並比較二區域間的差異.

8. 量分位數圖或經驗累積分布函數圖 (Quantile Plots or Empirical CDF Plots)

母體: 第 p 個量分位數 (p^{th} quantile) = 一數使得母體中 $\leq$ 此數的比率爲 p (相當於第 $100p^{\text{th}}$ 百分位數)

累積分布函數圖 (cdf plot, cumulative distribution function plot): 橫軸 (x-axis) 爲 quantiles, 縱軸 (y-axis) 爲 quantiles 對應出的 $\leq$ 其值的比率 (或百分比) 的圖形. (y-axis 通常標示爲累積機率 (cumulative probability) 或累積頻率 (cumulative frequency).)

樣本: 無法獲得每一觀察值所對應的 quantile (因爲不知道眞正母體的 quantiles), 所以只能用樣本資料估計觀察值所對應的 quantiles.

將排序後的觀察值 (由小到大) 作爲 x 座標, 對應的估計累積機率 (estimated cumulative probability) 作爲 y 座標, 而得的圖形, 稱作量分位數圖 (quantile plot) (又稱作經驗累積分布函數圖, empirical cdf plot).

estimated cumulative probability 又稱作 plotting position (描點位置), 其求法爲下列三式之一:

(1) $\widehat{p}_i = \#[x_j \leq x_{(i)}]/n$

(2) $\widehat{p}_i = i/n$

(3) $\widehat{p}_i = (i-a)/(n-2a+1)$

其中 $\widehat{p}_i$ = 第 i 個有序統計量 $x_{(i)}$ 所對應的估計累積機率, $i = 1, 2, \ldots, n$, 且 a 爲 0 與 1 間的一常數.

註 1. (1) 式的估計量說明小於或等於第 i 個有序觀察值的數在母體中所占的比率是由樣本中 $\leq$ 此觀察值 (第 i 個有序觀察值) 的觀察值在樣本中所占的比率來估計. ((1) 式的 estimator 有時又稱作 empirical probability estimator)

註 2. 在資料中沒有相同的觀察值 (no ties) 時, (2) 式等於 (1) 式.

註 3. (1) 式與 (2) 式的缺點爲它們暗示最大觀察值是母體的最大可能值 (100^{th} 百分位數).

註 4. 母體的分布若假設爲連續時, 常以 (3) 和一特定的 a 值計算估計累積機率. 參考表 3.7 有關特別母體對應的特定 a 值 (根據統計原理而得).

註 5. 一般 x 值的估計累積機率的求法:

- 離散分布: 針對 $x_{(i)} \le x < x_{(i+1)}$,

$$\widehat{p}_{(x)} = \widehat{p}_i$$

(⇒ quantile plot 維持水平 (flat), 直到碰到一個有序觀察值時, 才有一跳躍 (jump). 所以爲一 step function.)

- 連續分布: 針對 $x_{(i)} \le x < x_{(i+1)}$,

$$\widehat{p}_{(x)} = (1-r)\widehat{p}_i + r\widehat{p}_{i+1}$$

其中 $r = \frac{x - x_{(i)}}{x_{(i+1)} - x_{(i)}}$ (⇒ 根據線性內插法 (linear interpolation) 求之, 故以直線連接二相鄰觀察值所對應的點)

例. 以資料架構 `new.epa.94b.tccb.df` 中的 TcCB 繪出

- Reference area 的 TcCB 的 quantile plot (以點呈現)

- Reference area 的 TcCB 的 quantile plot (以線呈現) (以線性內插法求得)

註. 由上二圖形可約略看出 median $\approx$ 0.5, 第一四分位數 $\approx$ 0.4, 第三四分位數 $\approx$ 0.75; 0.8 附近突然平下來, 指出右偏斜.

- Reference area 的 TcCB 的 empirical cdf 與一特定的理論分布 (如, lognormal, 其參數由資料估計) 的 cdf 相比較的圖形

- Reference area 與 Cleanup area 的 TcCB 的 empirical cdf 的比較圖

9. 機率圖或量分位數-量分位數圖 (Probability Plots or Quantile-Quantile (Q-Q) Plots)

   爲一圖形展示, 說明一組資料與一特定機率分布或另一組資料的比較. 其本概念如下:

   - 若二母體分布完全相同, 則它們有相同的 quantiles. 所以, 第一個分布的 quantiles 對第 2 個分布的 quantiles 的圖 (亦即, x 座標爲第一個分布的 quartile, y 座標爲第二個分布的 quartile), 會落在 0-1 線上 (亦即, 截距爲 0, 斜率爲 1 的直線 $y = x$).

   - 若二母體有相同的外形與擴散性但有不同的位置 (locations), 則 quantiles 的圖會落在直

線 $y = a + x$ 上 (平行於 0-1 線), 其中 a 爲位置之間的差異 (因爲第一個母體的 quantiles $+$ a $=$ 第二個母體的 quantiles).

- 若二母體的位置差異爲 a 且又相差一個乘積常數 b, 則 quantiles 的圖會落在直線 $y = a + bx$ 上 (因爲 (第一個母體的 quantiles) $\times$ b $+$ a $=$ 第二個母體的 quantiles).

因此, 母體分布間的不同差異會導致各種偏離直線的種類.

註 1. 一組資料與一理論機率分布的比較

繪 empirical quantiles (亦即, 有序統計量或有序觀察值, ordered statistics) 爲 x 座標, 對應的理論機率的 quantiles 爲 y 座標的圖形 (稱作 Q-Q plot), 其中理論機率的 quantiles 乃根據所對應的有序觀察值的估計累積機率 (estimated cumulative probability) 求得的. 如, 將 Reference area 的 TcCB 與標準常態分布相比較, 計算它們的 quantiles 的步驟如下表:

Order Statistics	Plotting Position	Normal Quantiles
0.22	$\widehat{p}_1 = 0.013$	-2.2
0.23	$\widehat{p}_2 = 0.034$	-1.8
0.26	$\widehat{p}_3 = 0.056$	-1.6
...	...	...
1.14	$\widehat{p}_{45} = 0.944$	1.6
1.20	$\widehat{p}_{46} = 0.966$	1.8
1.33	$\widehat{p}_{47} = 0.987$	2.2

註. 上表中最後一行的 Normal Quantiles 是根據下列式子而得的:

第一列的 Normal Quantile 來自於
$P(Z \leq -2.2) = 0.013$

第二列的 Normal Quantile 來自於
$P(Z \leq -1.8) = 0.034$

⋮

最後一列的 Normal Quantile 來自於
$P(Z \leq 2.2) = 0.987$

例. Reference 區域的 TcCB 與 $N(0,1)$ 的比較: 呈現出一 $\bigcup$ 的外形, 表示右偏斜.

Reference 區域的 log(TcCB) 與 $N(0,1)$ 的比較: 落在一條直線上, 暗示 lognormal 分布可能是一好的模型.

註. normal Q-Q plot (亦即, 縱軸爲 $N(0,1)$ 的 quantiles, 橫軸爲樣本有序量 (empirical quantiles) 的 (Q-Q) plot) 的特性: 若樣本取自於常態分布的母體, 則 fitted line 的截距爲母體的 mean 的估計值, 且斜率爲母體的標準差的估計值. 因此, 觀察此 Normal Q-Q plot 可得截距爲 -0.6 ($\approx$ mean) 與斜率約 0.5 ($\approx$ 標準差)

註 2. 二組資料的比較

判斷此二組資料是否取自於同一母體. 縱座標爲第一組資料的 empirical quantiles, 橫座標爲對應的第二組的 empirical quantiles. 若二組資料的大小不一樣時, 選用較小資料組的 empirical quantiles, 較大資料組的對應的 quantiles 需透過較小資料組的 plotting positions (亦即, empirical cdf) 及線性內插法求得. 如, 令 $y_{(1)}, y_{(2)}, \cdots, y_{(m)}$ 爲第一資料組的 m 個有序觀察值且 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 爲第二組資料的 n 個有序觀察值. 若均取自連續分布且 $m < n$, 則選

用 $y_{(1)}, y_{(2)}, \ldots, y_{(m)}$ 爲縱座標, 對應的橫座標求法如下：

(a) 計算第一組中 $y_{(1)}, y_{(2)}, \ldots, y_{(m)}$ 的 m 個 plotting positions: $\widehat{p}_1, \widehat{p}_2, \cdots, \widehat{p}_m$.

(b) 計算第二組中 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ 的 n 個 plottin positions (亦即, empirical cdf); $\widehat{p}_1^*, \widehat{p}_2^*, \ldots, \widehat{p}_n^*$.

(c) 針對 $1 \le i \le m$, 若 $\widehat{p}_j^* \le \widehat{p}_i \le \widehat{p}_{j+1}^*$, 則令與 $y_{(i)}$ 對應的 quartile

$$x_{(i)}^* = (1-r)x_{(j)} + rx_{(j+1)}$$

其中

$$r = \frac{\widehat{p}_i - \widehat{p}_j^*}{\widehat{p}_{j+1}^* - \widehat{p}_j^*}$$

最後, 描點 $(y_{(i)}, x_{(i)}^*)$ (或 $(x_{(i)}^*, y_{(i)})$), $1 \le i \le m$, 得 Q-Q plot.

例. 比較 Reference area 與 Cleanup area 的 log(TcCB) 資料.

因爲 Reference area 中有 47 個觀察值 < Cleanup中有 77 個觀察值, 故選用 Reference area 中的 47 個有序觀察值爲 x 座標, 對應的 quartile (作爲 y 座標) 如下表求得:

參考區的 Quantiles	描點 位置	清理區的 Quantiles
-1.15	0.013	-2.41
-1.47	0.034	-2.40
-1.35	0.056	-2.12
...	...	...
0.13	0.944	1.77
0.18	0.966	2.88
0.29	0.987	4.66

註. 上表中第一行的參考區域的量分位數的求法如下:

$$\log(.22) = -1.51$$

$$\log(.23) = -1.47$$

$$\vdots$$

$$\log(1.2) = 0.18$$

$$\log(1.3) = 0.29$$

第三行的清理區域的量分位數的求導過程如下:

$$P(\text{清理區域的觀察值} \leq -2.41) \approx 0.013$$

$$P(\text{清理區域的觀察值} \leq -2.40) \approx 0.034$$

$$\vdots$$

$$P(\text{清理區域的觀察值} \leq 2.88) \approx 0.966$$

$$P(\text{清理區域的觀察值} \leq 4.66) \approx 0.987$$

描點後, 得一 Q-Q 圖, 顯示描點並未落在 0-1 直線上, 而傾向於落在二條斜率均相當大於 1 的不同直線上. 此種現象反應出其中一個樣本 (此處是 Cleanup 區域的資料) 可能取自於 "混合" 分布: 一些觀察值來自於與 Reference 區域分布有相似的外形與尺度的分布, 而另一些觀察值來自於右移及擴散更廣的分布 (相對於 Reference area 分布而言).

註 3. 偏離線性的評估 (Assessing Departures from Linearity)

基於下列理由, 一件重要的事就是, 對各類資料, 分布, 樣本大小的典型 (typical) Q-Q 圖的樣子應有一個好的認知:

- Q-Q 圖上的點僅是樣本而非母體的表現 (因此, 即使 underlying distribution 完全相同或僅差一個加法或乘法常數, Q-Q 圖上的點不會剛好落在一條直線上)
- 極端值點比靠近中心的點會造成更多的變異
- 由小樣本得到 Q-Q 圖比由大樣本得到的 Q-Q 圖更易變動

對於任意特定的分布與樣本大小, 可透過主選單點選 **EnvironmentalStats ▶ EDA ▶ Q-Q plot ▶ Q-Q plot GestaIt** 製造典型的 Q-Q 圖.

偏離線性的一般型式與對應的原因如下表:

Patteror (型式)	原因 (Cause)
$\cup$ 型	y 軸的分布是右偏斜 (相對於 x 軸的分布)
倒 $\cup$ 型 ($\Leftrightarrow \cap$)	x 軸的分布右偏斜 (相對於 y 軸的分布)
S型	x 軸的分布有較大 (重) 的尾部 (相較於 y 軸的分布)
左邊下彎, 中間直線, 右邊上彎 (亦即, 反 S 型)	y 軸的分布有較大 (重) 的尾部 (相較於 x 軸的分布)
二分開的直線	y 軸的分布是二不同分布的混合
大多數的點落在一直線上, 但一個或數個遠在直線的上或下	一個或多個離群值

註 4. Tukey Mean-Difference Q-Q Plots

一種評估偏離線性程度更好的方法 (相較於前面於 Q-Q plot 中加入一條配適迴歸直線 (fitted regression line), 簡稱 m-d plot, 此圖的 y 軸爲兩個量分位數的差 (difference, 來自於不同的兩組), x 軸爲兩個量分位數的平均值 (mean). 根據 m-d plot, 會有下列數種的評估結論:

(a) 若兩組量分位數來自於相同的母體分布, 則 m-d plot 中的點會大致落在水平線 $y = 0$ 上面.

(b) 若兩組量分位數來自於僅有位置 (location) 平移差異的兩個母體分布, 則 m-d plot 中的點會大致落在 $y = 0$ 之上或下的平行線上.

(c) 若兩組量分位數來自於有乘積常數差異的兩個母體分布, 則 m-d plot 中的點會大致落在一個有角度的直線上.

例. 試繪下述的三圖形.

1. Reference area 的 TcCB 對立於 $N(0, 1)$ 的 m-d plot

2. Reference area 的 TcCB 對立於 lognormal 分布的 m-d plot

3. Reference area 的 log(TcCB) 對立於 Cleanup area 的 log(TcCB)

10. Box-Cox 資料轉換 (Data Transformations) 與 Q-Q 圖

資料轉換有其必要性, 因爲

(a) 一些標準的參數假設檢定均會假定:

- 觀察值取自於常態分布.
- 若有多個母體同時涉入時, 它們要有相同的變異數.

(b) 標準線性迴歸模型中, 除了有上述的假設外, 還要求反應變數 (response variable) 與預測子變數 (predictor variable 或 variable) 間有線性關係的假設.

特別是環境資料都不符合上述的假設, 因爲原始資料多爲右偏斜以及 (或) 不是眞正常態分布的外形, 然而有時卻可透過轉換, 將原始資料轉變爲來自於常態 (或近似常態) 分布母體, 並且也可能導出變異一致性和反應與預測變數間的線性關係.

轉換方法:

(a) 理論上的考量: 如計數 (count) 資料多來自 Poisson 分布, 將觀察值開平方後, 會轉換成更有鐘形的外貌; 環境領域中, 化學濃度常來自於 lognormal 分布或正偏斜分布, 將觀察值取對數後, 會得到常態分布的資料.

(b) 對於生成資料的過程的認知以及圖形工具, 如: Q-Q 圖與直方圖.

(c) Box-Cox Data Transformation: 一正式 (公式化) 的方法, 定義如下: 給定一來自於正值分布的隨機變數 X, 則 Box-Cox family of power transformations

$$Y \stackrel{\text{def}}{=} \begin{cases} (X^\lambda - 1)/\lambda, & \text{若 } \lambda \neq 0 \\ \log(X), & \text{若 } \lambda = 0 \end{cases}$$

來自於常態分布, 其中 λ 爲 power of transformation (轉換冪次).

註 1. 轉換對 λ 而言是連續的 (用微積分的內容證證看); 保序的 (preserving order), 亦即, 若 $X_1 < X_2$ 則 $Y_1 < Y_2$ (想想看).

註 2. Box 與 Cox 建議透過最大化概似函數 (likelihood function) 的方式選取適常的 λ 值.

註 3. 若 $\lambda \neq 0$ 時, 可採用較簡單的轉換 $Y = X^\lambda$, 因爲與原先的轉換只有一個倍數和

原點平移的差異, 故有相同本質上的特性 (Why? 想想常態分布).

註 4. λ 值所對應的效應如下述:

- $\lambda = 1 \Leftrightarrow$ 無轉換
- $\lambda < 1$ 會縮小 X 中的大值, 故有效於右偏斜資料的轉換 (如圖:)
- $\lambda > 1$ 會放大 X 中的大值, 故有效於左偏斜資料的轉換 (如圖:)

幾個常用的 λ 值如下:

- 0 (對數轉換, log transformation)
- 0.5 (方根轉換, square-root transformation)
- -1 (倒數轉換, reciprocal transformation)
- -0.5 (倒方根轉換, reciprocal root transformation)

註 5. 不可直接將轉換後尺度 (transformed scale) 的結果反轉回原尺度 (original scale) 的結果; 在轉換後尺度中估計的量分位數 (如, means), 變異數和信心極限 (confidence limits) 被反轉回原尺度時通常會導致 biased 與 inconsistent 的估計, 如由對數轉換後的資料所得的 mean 的信賴極限, 經過指數化的反轉後, 所得的結果不是原尺度的 mean 的信賴區間, 而是 median 的信賴區間. 然而, 量分位數與順序墊基 (ranked-based) 的過程, 對於單調轉換 (monotonic transformation) 是不變的 (invariant) (亦即, 轉換後的某量估計, 反轉換後還是那量的估計) (Why? 想想看)

註 6. 可透過環境統計模組, 以下面三種判斷標準選擇最佳的 Box-Cox 轉換, 或計算 λ 的某範圍內所對應的特定判斷標準值:

- Probability Plot Correlation Coefficient (PPCC): 介於 -1 與 1 之間的數; 資料若來自於常態分布, 則 PPCC 會接近 1.

- Shapiro-Wilk Goodness-of-Fit Test Statistics (W): 與 PPCC 非常相關, 其值被限定介於 0 與 1 之間; 資料若來自於常態分布, 則 W 會靠近 1.

- Log-Likelihood Function: 愈大的值表示與常態分布有愈好的配適 (fit).

例. 繪 Reference area 的 TcCB 的 PPCC (對應於不同的 λ), 可明顯地看出 $\lambda = 0$ 時有最大值, 此時對應到 log 轉換 (我們一開始就採用的轉換, 此時得到一個確認). 除此, 還繪出 9 個 λ 值分別的 normal Q-Q plot 與 m-d plot, 共有 19 頁圖.

三. 雙或多變數的圖形

觀看 2 個或多個變數之間的關係:

1. 散播圖與局部加權迴歸平滑化線 (Scatter plots and Loess Smooths)

(a) 散播圖可被用於觀看 2 變數, 如 x 與 y 之間的關係, 這些變數都某種程度自然地以有序對的方式呈現. 通常我們有興趣於判斷兩個變數間是否有某種被資料中的噪音 (noise) 所隱藏的確定 (deterministic) 關係, (如, linear 或 curvilinear)

(b) loess curve: 一種非常方便的平滑化子 (smoother), 在不需要給予資料某種特定的模型 (如, 線性模型) 下, 可找出噪音中的訊號 ("signal" in the "noise"), 名稱來自於 "locally weighted regression", 製造過程如下:

i. 對每一 x 值, 製造一包含此值的視窗 (window 或 local neighborhood, bin).

ii. 僅針對在視窗內的觀察值, 以加權迴歸 (weighted regression) 方式 (亦即, 愈靠近 x 值的觀察值, 得到愈大的加權) 求出最配適的直線 (或 2 次曲線).

iii. 配適出直線後, 根據目前的 x 值預測 y 值. 接著通常使用一 robust regression 程序, 進行重覆配適過程 (iterative fitting

process), 給予較大餘數 (residual = 觀察值 - 配適值) 的觀察值較小的加權後, 再一次計算配適.

例. 探討 2 個資料架構: `air.df` 與 `environmental.df` 內含臭氧 (ozone) (`environmental.df` 的單位爲 ppb; `air.df` 的單位爲 $\text{ppb}^{1/3}$), 太陽輻射能 (solar radiation), 溫度 (temperature), 風速 (wind speed) 的 111 個由 1973 年 5 月至 9 月測得的觀察值.

(a) 繪 ozone 與 temperature 的 scatterplot 可得知臭氧與溫度爲正相關 (positively correlated) (亦即, 溫度 $\uparrow \Leftrightarrow$ 臭氧濃度 $\uparrow$) 以及溫度增高時, 臭氧的變異性會加大.

(b) 繪臭氧$^{1/3}$ 與溫度的 scatterplot (用 `air.df`) 可得知三次方根轉換後的臭氧與溫度較具線性關係, 同時也保有相同的臭氧變異性.

(c) 透過 smoother, loess curve, 可評估臭氧的三次方根與溫度是否眞有線性關係. 結果呈現出非線性, 而是些微地 curvilinear. 也許可用

2 個不同的直線模型化此組資料的母體; 一條直線描述溫度在華氏 75 度以下的關係, 另一條直線描述溫度超過 75 度時的關係.

2. 三維的繪圖 (Three-Dimensional Plots)

展示三個變數之間的關係, 有 6 種不同的繪製方式: 三維散播圖 (three-dimensional scatterplot) (又稱雲狀圖, cloud plot), 含文字散播圖 (scatterplot with text), 泡泡圖 (bubble plot), 等高線圖 (contour plot), 映象圖 (image plot) (又稱作水平層次圖, level plot) 以及曲面圖 (surface plot) (又稱絲網架構圖或透視圖), 分述如下:

(a) 三維散播圖

將點 (x, y, z) 以三度空間圖形的方式呈現出.

例. 繪 `environmental.df` 的臭氧對應於溫度與風速的三維散播圖. 由圖得知, 臭氧與溫度的關係隨著風速而改變; 風速愈大, 臭氧必然隨著溫度增高而變大. 可透過旋轉以不同角度的透視 (perspective) 觀看.

(b) 含文字散播圖

就是一個 (二維的) 散播圖, 其中以第三變數的值作爲描點符號, 是一種簡單, 自然的工具, 用以傳遞 z 如何隨著 x 與 y 而變化的訊息.

例. 以臭氧值作爲描點符號的溫度對風速的含文字散播圖中, 可得知臭氧的範圍爲 1 ppb 到 168 ppb; 一般而言, 低風速與高溫下, 有最大的臭氧值, 雖然在中等與強勁風速下有一些高的臭氧值; 在 "弱風及低溫" 和 "強風及高溫" 下都沒有觀察值.

(c) 泡泡圖

就是一散播圖, 其描點符號爲圓泡, 大小與第三變數的值成比率 (亦即, 值愈大, 相對的圓泡愈大).

例. 以臭氧值對應的圓泡作爲描點符號所繪出的散播圖可顯示出含文字散播圖所呈現的相同特性. 不需以視覺比較臭氧的值, 而直接以圓泡的大小比較所對應的臭氧值.

(d) 等高線圖

以等高線的方式展示 z 變數在 x-y 平面上的變化情形.

例. 繪描述臭氧如何隨著溫度與風速而變化的等高線圖.

(e) 映象圖 (或水平層次圖): 以彩色或灰色尺度網格展示 z 變數如何在 x-y 平面上變化.

例. 將水平層次圖與等高線圖合併成一填充等高線圖 (filled contour plot), 以描述臭氧隨著風速和溫度的變化情形.

(f) 曲面圖 (或絲網架構圖, 透視圖)

在三維圖形上, 運用多重內差平滑曲線 (multiple interpolated smooth curve) 顯示 z 值的變化. 在平滑曲線之間填充色彩或灰色尺度的圖形, 稱作填充曲面圖.

例. 以填充曲面圖展示臭氧值的變化.

註 1. 類似於以一溫度的函數配適給臭氧值一個平滑曲線 (smooth curve), 等高線圖, 映象圖和曲

面圖都以一溫度和風速的函數配適給臭氧值一個平滑曲面 (smooth surface).

註 2. 當資料不是均勻擴散在一方形的區域時, 會得到危險的訊息, 如等高線圖, 映象圖和曲面圖均顯示臭氧值快速的增加, 當 "風速與溫度增加" 以及當 "風速遞減到 0, 溫度在 75 度左右" 時, 但是實際情況卻是, 在含文字散播圖與泡泡圖中, 在上述兩種情形時, 根本就沒有觀察值. 所以我們無從得知在那二種區域內臭氧的行爲. 危險的成因是, 在那二處沒有觀察值的地方依然以內差法 (外差法) 估計臭氧值並以填充曲面呈現出.

註 3. 在等高線圖, 映象圖和曲面圖中所呈現出的僅是配適曲面 (fitted surface) 而沒有原始資料, 所以我們沒有任何觀察到的臭氧值的變異性. 相異於此種現象的是含文字散播圖, 從其中可同時看到 "原始資料內臭氧與溫度的關係" 以及局部加權迴歸配適 (loess fit).

3. 散播圖矩陣 (Scatterplot Matrix) 與拂拭化 (Brushing)

散播圖矩陣以共用座標軸的方式展現所有可能的一對一對的散播圖. 同時也可以互動的方式拂拭化 (brush) 散播圖矩陣 (亦即, 在某一散播圖內標記 (highlight) 某些點, 這些點會同時在其他的散播圖內也被標記出).

例. 繪 `environmental.df` 的散播圖矩陣後, 會呈現出下列結果:

(a) 臭氧與輻射能及溫度均分別成正相關 (positively correlated); 但與風速成負相關 (negatively correlated).

(b) 溫度與風速成負相關.

(c) ''輻射線與溫度'' 和 ''輻射線與風速'' 都無明顯的關係.

(d) ''臭氧對輻射能'' 的圖形呈現出倒 V 的外形, 顯示在低輻射線時, 不產生高的臭氧值, 這是因爲在產生臭氧的光化學反應中需要某種數量的太陽輻射能. 另外, 當高的輻射能時, 高的臭氧值也不發生, 成因爲風速傾向於中等至強勁或溫度傾向於中溫至高溫.

4. 多格窗條件化圖 (Multi-Panel Conditioning Plot, 簡稱 Coplot) 或格窗圖 (Trellis Plot)

用於觀察在給定一個或多個變數的值的條件下, ''一個變數的行爲'' 或 ''兩個或多個變數間的關係''.

例. 根據不同的條件試繪對應的格窗圖.

(1) 將風速根據觀察值的個數均分成 4 類 (亦即, 每一類中的觀察值個數大致相等) 的條件下, 繪出臭氧對溫度的多格窗條件化圖, 同時加上對應的局部加權迴歸曲線 (loess curves). 圖形呈現出下列結果:

(i) 高臭氧值大致發生在弱風的條件下.

(ii) 一般而言, 臭氧值隨著溫度升高而增大; 但是它們之間 (臭氧與溫度) 的關係卻因風速而異 (如, 在弱風下, 臭氧隨著溫度快速增加; 但在強風時，增加速度趨緩)。

(2) 根據風速的範圍均分成 3 類的條件下, 繪出臭氧對溫度的多格窗條件化圖, 並加上對應的 loess curves. 此圖呈現出與上圖大致相同的結果.