單元 5: 估計分布參數與量分位數

根據繪圖工具以及產生資料的機制的認識, 可針對採樣計劃所得的資料予以某一特別機率分布的模型化. 一當決定欲使用的機率分布後, 通常必需估計與此分布相關的參數. 本章的目標爲

1. 探討估計不同機率分布的參數與量分位數的方法以及這些估計量的信賴區間的製造方法.
2. 環統模組執行上述方法的示範.

一. 估計分布參數的方法

三個最常用於估計機率分布參數的方法爲: 動差法估計量 (method of moment estimator) (MME), 最大概似估計量 (maximum likelihood estimator) (MLE), 以及最小變異不偏估計量 (minimum variance unbiased estimator)(MVUE).

近來，L-動差估計量 (L-moment estimator) 也被引進，並常用於水文文獻中.

有些時候，這些方法的估計量的公式會有略微差異，但經常它們產生相同的公式.

本節略述 MMEs, MLEs 與 MVUEs 背後的基本概念如下. 至於 L-moment estimator 的細節內容，請參看環統模組的輔助說明檔.

1. 估計量也是隨機變數

- 以 X 表示一隨機變數，x 表示經由樣本而得的此隨機變數的一個觀察值.

- 令 $x_1, x_2, \ldots, x_n$ 表示某一特別機率分布的 n 個觀察值 (經由抽樣而得)，則所有的估計量 (estimator) 均是使用由樣本而得的觀察值對母體參數予以估計. 以上概念的形式 (公式) 可化爲：令 $\widehat{\theta}$ 表某母體參數 θ 的估計量，則

$$\widehat{\theta} = h(x_1, x_2, \ldots, x_n)$$

此處 $h()$ 表示某函數. 如, 估計母體平均值的樣本平均值 (sample mean) 的 $h()$ 爲

$$h(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- 由於樣本是變動的, 估計量爲隨機變數的函數, 故估計量本身也是一隨機變數並有其對應的機率分布.

2. 動差法估計量 (MME)

基本概念爲, 以資料求得的樣本動差 (sample moments) 估計對應的機率分布的動差.

- X 的分布的 r 階動差 (r^{th} moment)

  $$\mu_r' \stackrel{\text{def}}{=} E(X^r)$$

  註. 一階動差 $\mu_1' = E(X) = \mu$ (常用的符號, 俗稱 X 的期望值 (mean)).

- r 階樣本動差 (r^{th} sample moment)

  $$\widehat{\mu}_r' \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i^r$$

作爲 r 階動差 μ'_r 的估計量.

註. 期望值的 mme (method of moments estimator)

$$\widehat{\mu}_{mme} = \widehat{\mu}'_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

等於俗稱的樣本期望值 (sample mean).

- X 的 r 階中央動差 (r^{th} central moment)

$$\mu_r \stackrel{\text{def}}{=} E[(X-\mu)^r]$$

註. X 的二階的中央動差
$\mu_2 = E[(X-\mu)^2] = \sigma^2$ (常用的符號, 俗稱 X 的變異數 (variance)).

- r 階樣本中央動差 (r^{th} sample central moment)

$$\widehat{\mu}_r \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r$$

爲 r 階中央動差 μ_r 的估計量.

註 1. 變異數的 mme

$$\hat{\sigma}^2_{mme} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_m^2$$

等於俗稱的有偏樣本變異數 (biased sample variance).

註 2. 標準差 sd 的 mme

$$\hat{\sigma}_{mme} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = s_m$$

等於俗稱的有偏樣本標準差 (biased sample sd), 此處有偏是指變異數的估計量爲有偏的.

例 1. 常態分布的 MME

常態分布由期望值 (μ) 及變異數 (σ^2) (或 μ 及標準差 (σ)) 所刻劃. 這些參數 (刻劃出分布特質的量) 的 MME 如下:

$$\hat{\mu}_{mme} = \bar{x},\ \hat{\sigma}^2_{mme} = s_m^2,\ \hat{\sigma}_{mme} = s_m$$

例 2. 對數常態分布的 MME

有二個方式刻劃出對數常態分布的特質:

1. 由期望值 (θ) 與標準差 (η) (或期望值 (θ) 與 CV (τ))

2. 由對數轉換後的期望值 (μ) 與標準差 (σ)

對應的 MME 如下:

1. $\widehat{\theta}_{mme} = \bar{x}, \quad \widehat{\eta}^2_{mme} = s^2_m$ 或

   $\widehat{\eta}_{mme} = s_m, \quad \widehat{\tau}_{mme} = s_m/\bar{x}$

2. $\widehat{\mu}_{mme} = \bar{y}$, $\widehat{\sigma}^2_{mme} = s^2_{y,m}$, $\widehat{\sigma}_{mme} = s_{y,m}$

   其中

$$y_i = \log(x_i), i = 1, 2, \ldots, n$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s^2_{y,m} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

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

例 3. 二項分布的 MME

由二參數 n 表示試驗的次數與 p 表示一次試驗成功的機率, 共同刻劃出二項分布的特質. 由於 n 事先設定, 僅需估計 p 如下. 若

$$X \sim B(n,p)$$

則

$$E(X) = \mu = np$$

故

$$p = \frac{\mu}{n}$$

因爲 X 表示在 n 次試驗中成功的次數且一個單一觀察值爲 x, 所以, 以此單一觀察值而得的

$$\widehat{\mu}_{mme} = x$$

因此, p 的 MME

$$\widehat{p}_{mme} = \frac{\widehat{\mu}_{mme}}{n} = \frac{x}{n}$$

即 "成功" 機率的 MME = (成功的次數)/(試驗的次數).

另類觀點:

$$X \sim B(n,p) = \sum_{i=1}^{n} X_i$$

其中

$$X_i \overset{\text{iid}}{\sim} B(1,p)$$

所以,

$$\widehat{p}_{mme} = \frac{x}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

3. 最大概似估計量 (MLE)

基本概念爲參數值應爲使得此特殊結果發生的機率爲最大. 也就是說, 不同的參數值對應出此特殊結果發生的機率, 而目前觀察到了此特殊結果, 這意謂著此特殊結果發生的機率爲最大, 故參數值應爲產生此特殊結果的最大發生機率的那一個參數值 (或此特殊結果的最大發生機率所對應的那一個參數值).

如, 擲銅板 100 次, 出現 80 次正面, 20次反面. 問出現正面的機率較可能爲 0.5 或 0.8?

若 $p = 0.5$, 則

$$\begin{aligned} P(\text{此特殊結果}) &= \binom{100}{80}(0.5)^{80}(0.5)^{20} \\ &\approx 0.422 \times 10^{-9} \end{aligned}$$

(幾乎不可能出現)

若 $p = 0.8$, 則

$$P(\text{此特殊結果}) = \binom{100}{80}(0.8)^{80}(0.2)^{20} \approx 0.099$$

(此機率 $\gg p = 0.5$ 時的機率)

而目前此特殊結果發生了, 故參數應爲使得發生機率爲最大的那一個參數. 因此, 猜測出現正面的機率爲 0.8 較合理. 另外, 也可繪 $B(100, 0.5)$ 與 $B(100, 0.8)$ 的機率密度函數圖形, 並由圖形得知 $B(100, 0.8)$ 的分布最可能產生 "100 次投擲中會出現 80 次正面" 的結果.

- 概似函數 (The Likelihood Function) 以機率密度函數 (pdf) 求機率時, 假設分布參數已知 (亦

即, 有明確的分布), 再觀察結果並求此結果發生的機率.

然而 MLE 方法剛好顚倒上述假設而爲: 觀察結果已知, 但分布參數未知, 此時 pdf 轉換成 likelihood function.

如, $X \sim B(n,p)$ 的 pdf

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

(讀作給定 p 值, $X = x$ 的機率).

若擲 n 次, 出現正面 x 次, 則 likelihood function

$$f(p|x) = \binom{n}{x} p^x (1-p)^{n-x}$$

(讀作給定 x 值, 參數爲 p 時, $X = x$ 的機率).

問. 如何在未知的 $0 \le p \le 1$ 的範圍內, 求 $f(p|x)$ 的最大值?

因爲 x 固定且 log 函數爲遞增函數, 所以,

$$\text{maximize}_{\,0 \le p \le 1}\ f(p|x)$$

相當於

$$\text{maximize}_{\,0 \le p \le 1}\ x \log p + (n-x)\log(1-p)$$

根據微積分求極值的典型方法，需解

$$p: \frac{x}{p} - \frac{n-x}{1-p} = 0$$

由此可得

$$x(1-p) - p(n-x) = 0$$

化簡後，

$$x - np = 0$$

因此，

$$p = \frac{x}{n}$$

(驗證前僅能視 $p = x/n$ 爲一臨界點 (critical point)

接著需要驗證在 $p = x/n$ 確實有最大值:

$$p < \frac{x}{n}, \text{ 一階導函數 } > 0$$

且

$$p > \frac{x}{n}, \text{ 一階導函數 } < 0$$

所以，在 $p = x/n$，$f(p|x)$ 有最大值．故

$$\widehat{p}_{mle} = \frac{x}{n} = \widehat{p}_{mme}$$

推廣: 設 n 個觀察值

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{pdf } f()$$

則概似函數

$$f(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(\theta|x_i)$$

(因爲獨立性), 其中 θ 爲此分布的參數向量 (如, 期望值與標準差).

註. 因爲 log 函數爲一遞增函數, 所以最大化

$$f(\theta|x_1, x_2, \ldots, x_n)$$

相當於最大化

$$\begin{aligned}
&\log[f(\theta|x_1, x_2, \ldots, x_n)] \\
\overset{\text{表成}}{=}\; & l(\theta|x_1, x_2, \ldots, x_n) \\
=\; & \log\left[\prod_{i=1}^{n} f(\theta|x_i)\right] \\
=\; & \sum_{i=1}^{n} \log[f(\theta|x_i)]
\end{aligned}$$

(稱此爲對數-概似函數, log-likelihood), 而且通常最大化

$$l(\theta|x_1, x_2, \ldots, x_n)$$

比較容易 (因爲求和的導函數比求乘積的導函數容易).

例 1. 常態分布的 MLE

設

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma)$$

則

$$\begin{aligned} & l(\mu, \sigma | x_1, x_2, \cdots, x_n) \\ = & \sum_{i=1}^{n} \log[f(\mu, \sigma | x_i)] \\ = & \sum_{i=1}^{n} \log \left\{ \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \right\} \\ = & \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right] \\ = & -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

接著, 最大化

$$l(\mu, \sigma | x_1, x_2, \ldots, x_n)$$

相當於解

$$\frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]=0$$

以及

$$\frac{\partial}{\partial \mu}\left[-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]=0$$

實際微分後，亦相當於解

$$\begin{cases} -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2=0 \\ 0-\frac{1}{2\sigma^2}2(-1)\sum_{i=1}^{n}(x_i-\mu)=0 \end{cases}$$

化簡後，相當於解

$$-n\sigma^2+\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \quad (1)$$

$$\left(\sum_{i=1}^{n}x_i\right)-n\mu = 0 \quad (2)$$

首先由 (2) 式，得

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

將 $\bar{x}$ 代入 (1) 式，得

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

再由二階導數檢定法 (請自行驗證), 得

$$\mu = \bar{x} \text{ 且 } \sigma^2 = s_m^2$$

時,

$$l(\mu, \sigma | x_1, x_2, \ldots, x_n)$$

有最大值. 故

$$\widehat{\mu}_{mle} = \bar{x} \text{ 且 } \widehat{\sigma}^2_{mle} = s_m^2$$

註 1 分布參數的單調函數的 MLE = 分布參數的 MLE 代入此單調函數的值, 即 $h(\theta)$ 的 MLE 等於

$$h(\widehat{\theta}_{mle})$$

此處, h 爲一單調函數 (因爲單調函數是遞增的, 所以單調轉換後的參數與原參數會在相同的值使得概似函數爲最大)

因爲開方爲單調的, 故標準差的 MLE

$$\widehat{\sigma}_{mle} = \sqrt{\widehat{\sigma}^2_{mle}} = \sqrt{s_m^2} = s_m$$

註 2. 常態分布的 MLEs = MMEs.

例 2. 對數常態分布的 MLE

先計算對數轉換後的期望值 (μ) 與變異數 (σ^2) 的 MLE (因爲其爲 $N(\mu, \sigma)$, 所以可由例 1 得之), 再透過與原尺度的參數 (期望值 (θ), CV (τ) 與標準差 (η)) 的關係

$$\theta = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$\tau = \sqrt{exp(\sigma^2) - 1}$$

$$\eta = \theta\tau$$

可得 (因爲透過單調函數的轉換):

$$\begin{aligned} \hat{\theta}_{mle} &= \exp\left[\hat{\mu}_{mle} + \frac{1}{2}\hat{\sigma}^2_{mle}\right] \\ &= \exp\left[\bar{y} + \frac{1}{2}s^2_{y,m}\right] \end{aligned}$$

其中,

$$y_i = \log(x_i), i = 1, 2, \ldots, n$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s^2_{y,m} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

以及

$$\hat{\tau}_{mle} = \sqrt{\exp(\hat{\sigma}^2_{mle}) - 1} = \sqrt{\exp(s^2_{y,m}) - 1}$$

$$\hat{\eta}_{mle} = \hat{\theta}_{mle}\hat{\tau}_{mle}$$

註. 對數常態分布的期望值 (θ) 與標準差 (η) 的 MLEs $\neq$ MMEs (與常態分布相異之處).

4. 最小變異不偏估計量 (MVUE)

- 設 $\hat{\theta}$ 爲母體參數 θ 的估計量, 則 $\hat{\theta}$ 的偏離度 (bias) (或稱作系統誤差, systematic error)

$$\text{Bias}(\hat{\theta}) \stackrel{\text{def}}{=} E(\hat{\theta} - \theta)$$

(亦即, 估計量的平均值與它所要估計的參數之間的差). 註. $E(\hat{\theta}) = \theta$, 則稱 $\hat{\theta}$ 爲不偏 (unbiased).

$E(\hat{\theta}) < \theta$, 則稱 $\hat{\theta}$ 爲負偏差 (negatively biased).

$E(\hat{\theta}) > \theta$, 則稱 $\hat{\theta}$ 爲正偏差 (positively biased).

無論是正或負的偏差, 統稱爲有偏的 (biased).

- 期望值與變異數的不偏估計量

樣本期望值

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

爲期望值的不偏估計量 (unbiased estimator) (因爲 $E(\bar{x}) = E(X) = \mu$).

樣本變異數

$$s^2 \overset{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

爲變異數的不偏估計量 (因爲 $E(s^2) = Var(X) = \sigma^2$) (自行驗証).

註 1. 樣本標準差

$$s \overset{\text{def}}{=} \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

不是母體標準差 σ 的不偏估計量.

註 2 因為

$$
\begin{aligned}
s_m^2 &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= (\hat{\sigma}^2_{mme}，\text{變異數的動差法估計量}) \\
&= \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{n-1}{n}s^2
\end{aligned}
$$

所以,

$$s_m^2 < s^2$$

- MVUE 的定義

母體參數 θ 的所有不偏估計量當中, 有最小變異數的那一個, 稱作 θ 的最小變異不偏估計量 (MVUE).

一般數理統計的書中都詳述求 MVUE 的技巧 (請自行參看).

例 1. 常態分布的 MVUE

$$\widehat{\mu}_{mvue} = \bar{x} \text{ (樣本期望值)}$$

$$\hat{\sigma}^2_{mvue} = s^2 \text{ (樣本變異數)}$$

例 2. 對數常態分布的 MVUE

1941 年由 Finney 求得對數常態 X 的期望值 (θ) 與變異數 (η^2) 的 MVUE, 分述如下:

$$\hat{\theta}_{mvue} = e^{\bar{y}} g_{n-1}\left(\frac{s_y^2}{2}\right)$$

$$\hat{\eta}^2_{mvue} = e^{2\bar{y}}\left\{g_{n-1}(2s_y^2) - g_{n-1}\left[\frac{(n-2)s_y^2}{n-1}\right]\right\}$$

此處

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ (y_i = \log(x_i), i = 1, 2, \ldots, n)$$

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

(註. $s_y^2 = \hat{\sigma}^2_{mvue}$ (對數轉換後的隨機變數 $\sim N(\mu, \sigma)$ 的變異數 σ^2 的不偏估計量))

$$g_m(z) = \sum_{i=0}^{\infty}\frac{m^i(m+2i)}{m(m+2)(m+2i)}\left(\frac{m}{m+1}\right)^i\left(\frac{z^i}{i!}\right)$$

例 3. 二項分布的 MVUE

$$\widehat{p}_{mvue} = \frac{x}{n} \ (= \widehat{p}_{mme} = \widehat{p}_{mle})$$

二. 使用環境統計模組估計分布參數

幾乎所有在第四章討論過的機率分布的參數都可根據樣本以環統模組估計之. 僅示範數例如下:

例 1. 估計常態分布的參數

資料架構 (data frame) `new.epa.94b.tccb.df` 中 Reference area 的 TcCB 資料是來自於對數常態分布 (透過模型化確定的結果).

試估計對數轉換後的資料 (爲一常態分布) 的期望值與標準差 (分別近似於 −0.62 與 0.47 log(ppb)), 並繪出此資料的密度直方圖以及附上建基於估計值而得的配適常態分布 (fitted normal distribution).

例 2. 估計對數常態分布的參數

假設 Reference area 的 TcCB 資料爲來自於對數常態分布, 今以不同的估計量求得的期望直 (θ) 與 CV (τ) 的估計如下表:

參數	估計量	値
θ	MME	0.60
θ	MLE	0.60
θ	QMLE	0.60
θ	MVUE	0.60
τ	MME	0.47
τ	MLE	0.49
τ	QMLE	0.49
τ	MVUE	0.49

註 1. 刪到百分位時, 所有期望値的估計値都一樣; CV 的估計值, 除了 MME 的以外, 均一樣.

註 2. QMLE 爲準最大概似估計量 (Quasi-Maximum Likelihood Estimator), 與 MLE 的式子一樣, 除了σ^2 的估計量由

$$s_{y,m}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

換成

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

註 3. 期望值的 QMLE 與 MLE 對於典型的環境資料會產生嚴重的偏差 (biased), 故建議用 MVUE.

註 4. CV 的估計值 = 標準差的估計值/期望值的估計值.

另外展示此資料的密度直方圖以及基建於 MVUE 估計值的配適對數常態分布.

例 3. 估計二項分布的參數

資料架構 `epa.92c.benzene1.df` 中, 內含 6 個背景井 (background wells), 各 6 個月 (每月採集一次) 的地下水樣本中苯 (benzene) 的濃度, 共有 36 筆資料. 無法檢測值 (nondetect values) 以 "$<$ 2" 記錄之, 共有33筆. 試以此組資料估計在 6 井中會觀察到一個無法檢測值的機率.

透過二項分布的任一 p 的估計量可得

$$\widehat{p} = 33/36 \approx 0.92$$

(因爲 $\widehat{p}_{mme} = \widehat{p}_{mle} = \widehat{p}_{mvue} = x/n$)

三. 不同估計量的比較

所有估計量均爲隨機變數的函數, 故其本身亦爲一隨機變數, 其值會在欲估計的參數附近跳動, 變動的程度隨著估計量而異. 如:

1. $N(10, 2)$ 的樣本期望值 $\bar{x}$ (分別取樣本大小 $n = 4$ 與 $n = 30$) 的分布圖形如下:

   圖形顯示:

   $n = 30$ 的 $\bar{x}$: 較集中在期望值附近.

   $n = 4$ 的 $\bar{x}$: 變動性較大.

2. $N(0, 2)$ 的兩種 σ^2 的估計量

$$\widehat{\sigma}^2_{mme} = \widehat{\sigma}^2_{mle} = s^2_m$$

與

$$\hat{\sigma}^2_{mvue} = s^2$$

在取樣大小 $n = 4$ 時的分布圖形如下:

眞正的 $\sigma^2 = 4$, 故 MVUE 比較不偏.

3. 取樣大小 $n = 30$ 時, $N(0, 2)$ 的 σ^2 的 2 種估計量

$$\hat{\sigma}^2_{mme} = \hat{\sigma}^2_{mle} = s^2_m$$

與

$$\hat{\sigma}^2_{mvue} = s^2$$

的分布圖形如下:

當樣本大小 $n \uparrow$ 時, 兩者差異不大. (MVUE 還是比較不偏)

如何判斷估計量的好壞 (或選擇估計量)? 有如下 2 種方法:

1. 誤差平方期望值 (Mean Squre Error, MSE)

一種常用來比較估計量的判斷準則, 定義如下: 若 $\hat{\theta}$ 爲 θ 的估計量, 則 $\hat{\theta}$ 的 MSE 爲

$$\begin{aligned}\text{MSE}(\hat{\theta}) &\stackrel{\text{def}}{=} E[(\hat{\theta}-\theta)^2] \\ &= Var(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2\end{aligned}$$

(亦即, $\hat{\theta}$ 與欲估計參數 θ 之差的平方的平均值.)

Why?

$$\begin{aligned}&E[(\hat{\theta}-\theta)^2] \\ &= E[(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta)^2] \\ &= E[(\hat{\theta}-E(\hat{\theta}))^2+2(\hat{\theta}-E(\hat{\theta}))\cdot \\ &\quad (E(\hat{\theta})-\theta)+(E(\hat{\theta})-\theta)^2] \\ &= E[(\hat{\theta}-E(\hat{\theta}))^2]+2(E(\hat{\theta})-\theta)\cdot \\ &\quad E(\hat{\theta}-E(\hat{\theta}))+[E(\hat{\theta})-\theta]^2 \\ &= Var(\hat{\theta})+2[E(\hat{\theta})-\theta][E(\hat{\theta})-E(\hat{\theta})]+ \\ &\quad [\text{Bias}(\hat{\theta})]^2 \\ &= Var(\hat{\theta})+[\text{Bias}(\hat{\theta})]^2\end{aligned}$$

註. 若 $\hat{\theta}$ 爲不偏, 則 MSE $(\hat{\theta}) = Var(\hat{\theta})$ (因爲此時 $\text{Bias}(\hat{\theta}) = 0$).

例 1. 常態分布的樣本期望值與樣本中位數的比較

設

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)$$

(亦即，由 $N(\mu,\sigma)$ 抽樣取得的大小爲 n 的隨機樣本)．因爲常態分布的母體期望值 $=$ 母體中位數，故可用樣本期望值

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

或樣本中位數

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

來估計 μ．問．何者較優?

分別考慮各自的 MSE: (i) $\bar{x}$: μ 的不偏估計量．所以，

$$\text{MSE}(\bar{x}) = Var(\bar{x}) = \frac{\sigma^2}{n}$$

(ii) 樣本中位數: μ 的不偏估計量 (針對常態分布而言)．故

$$\text{MSE}(\text{樣本中位數}) = Var(\text{樣本中位數}) \approx \frac{\pi}{2}\frac{\sigma^2}{n}$$

(參考指定的文獻) 因此,

$$\text{MSE}(\bar{x})/\text{MES}(\text{樣本中位數}) \approx \frac{2}{\pi} = 63.7\%$$

稱樣本中位數的效率僅是樣本期望值的 63.7% (針對常態分布而言).

例 2. 常態分布的變異數的估計量的比較

設

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)$$

有兩種估計 σ^2 的估計量:

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

與

$$\text{MME/MLE } s_m^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

問. 何者較優?

可由各自的 MSE 比較之:

(i) MVUE s^2: σ^2 的不偏估計量. 所以,

$$\text{MSE}(s^2) = Var(s^2) = 2\sigma^4/(n-1)$$

Why?

$$(n-1)s^2/\sigma^2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

(當 $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)$ 時, 成立. 參考數理統計的書) 所以,

$$\begin{aligned} Var(s^2) &= Var\left(\frac{\sigma^2}{n-1}\chi^2(n-1)\right) \\ &= \frac{\sigma^4}{(n-1)^2}Var(\chi^2(n-1)) \\ &= \frac{\sigma^4}{(n-1)^2}2(n-1) \\ &= 2\sigma^4/(n-1) \end{aligned}$$

(因爲 $Var(\chi^2(r)) = 2r$, $E(\chi^2(r)) = r$)

(ii) MME/MLE s_m^2: σ^2 的偏差估計量. 所以,

$$\text{MSE}(s_m^2) = Var(s_m^2) + [\text{Bias}(s_m^2)]^2$$

接著需要分別計算出 $\text{Bias}(s_m^2)$ 與 $Var(s_m^2)$. 首先,

$$E(s_m^2) = \frac{n-1}{n}\sigma^2$$

Why? 因爲

$$ns_m^2/\sigma^2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

得

$$
\begin{aligned}
E(s_m^2) &= E\left(\frac{\sigma^2}{n}\chi^2(n-1)\right) \\
&= \frac{\sigma^2}{n}E(\chi^2(n-1)) \\
&= \frac{\sigma^2}{n}(n-1)
\end{aligned}
$$

所以,

$$
\begin{aligned}
\text{Bias}(s_m^2) &= E(s_m^2) - \sigma^2 \\
&= \frac{n-1}{n}\sigma^2 - \sigma^2 \\
&= -\frac{\sigma^2}{n}
\end{aligned}
$$

又

$$
\begin{aligned}
Var(s_m^2) &= Var\left(\frac{\sigma^2}{n}\chi^2(n-1)\right) \\
&= \frac{\sigma^4}{n^2}2(n-1) \\
&= \frac{2(n-1)}{n^2}\sigma^4
\end{aligned}
$$

因此,

$$\begin{aligned}\mathrm{MSE}(s_m^2) &= \frac{2(n-1)}{n^2}\sigma^4 + \left(-\frac{\sigma^2}{n}\right)^2 \\ &= \frac{2n-2+1}{n^2}\sigma^4 \\ &= \frac{2n-1}{n^2}\sigma^4\end{aligned}$$

MSE 的比值如下:

$$\begin{aligned}\frac{\mathrm{MSE}(s^2)}{\mathrm{MSE}(s_m^2)} &= \frac{2\sigma^4/(n-1)}{(2n-1)\sigma^4/n^2} \\ &= \frac{2n^2}{(2n-1)(n-1)} \\ &= \frac{2n^2}{2n^2-3n+1}\end{aligned}$$

數個不同的樣本大小 n 所對應的比值如下表:

Sample size	MSE的比值
2	2.67
3	1.80
4	1.52
5	1.39
10	1.17
25	1.06
50	1.03
100	1.02

結論: 比值 >1, 但 $\downarrow 1$ 當 $n\uparrow$. 所以由 MSE 的角度而言, s_m^2 (有偏估計量) 是比 s^2 (不偏的, 常用的估計量) 有些微的效率.

2. 以信賴區間 (Confidence Intervals) 比較估計量

在實際上, 決策者不單需要知道估計值, 同時也需要知道估計值 "擺動" 的程度. 信賴區間可量化點估計的 "擺動" 程度, 所以透過由估計量得到的信賴區間的表現情況, 來比較估計量, 是一件合理的事情. 一種度量信賴區間表現的可行方法是考慮信賴區間寬度 (width) 的期望值或中位數.

註. 以 MSE 比較估計量和以信賴區間的效能比較估計量, 二者間並不必然存在一對一的對應關係 (亦即, MSE 顯示有效率不一定保証信賴區間的效能高; 反之亦然.) 這是因爲由估計量得到的信賴區間不單與此估計量在眞正參數附近的彈跳情形有關, 也與此估計量的變異數的估計量表現有關. 如, 例 1 中之

$$Var(\bar{x}) = \sigma^2/n$$

其中 σ 未知, 所以需要估計 σ^2 而得 $Var(\bar{x})$ 的估計量, 此估計量的表現會影響信賴區間的寬度 (width).

四. 準確度 (Accuracy), 偏差度 (Bias), 誤差平方均值 (Mean Square Error), 精密度 (Precision), 隨機誤差 (Random Error), 系統誤差 (Systematic Error) 與變異度 (Variability)

了解各名稱的明確定義並避免誤用是相當重要的, 分述如下:

1. Bias 與 Systematic Error: 平均而言, 此乃度量估計量 ($\widehat{\theta}$) 與眞正的母體參數 (θ) 間靠近的程度, 亦即,

$$\begin{aligned}\text{Bias}(\widehat{\theta}) &= \text{Systematic Error}(\widehat{\theta}) \\ &= E(\widehat{\theta}-\theta) \\ &= E(\widehat{\theta})-\theta\end{aligned}$$

若 Bias 或 Systematic Error

$$= \begin{cases} (+) & \text{稱估計量爲正偏} \\ & \text{(positively biased)} \\ 0 & \text{稱估計量爲不偏} \\ & \text{(unbiased)} \\ (-) & \text{稱估計量爲負偏} \\ & \text{(negatively biased)} \end{cases}$$

估計量有正或負偏時, 統稱爲偏差的 (biased), 並且此估計量擁有系統誤差 (systematic error).

2. Precision, Random Error 與 Variability: 以 Variance 度量估計量 ($\widehat{\theta}$) 在估計量平均值 ($E(\widehat{\theta})$) 附近彈跳的程度, 亦即,

$$Var(\widehat{\theta}) = E[(\widehat{\theta} - E(\widehat{\theta}))^2]$$

若一估計量有小的變異度 (variability), 稱其爲精密的 (precise); 有大的變異度, 稱其爲不精密 (imprecise).

3. Mean Square Error 與 Root Mean Square Error: MSE 度量估計量 ($\widehat{\theta}$) 在欲估計參數 (θ) 附近彈跳的程度, 亦即,

$$\text{MSE}(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2]$$

又

$$
\begin{aligned}
&\text{MSE}(\widehat{\theta}) = \\
&\quad \widehat{\theta} \text{ 的 variability (random error)} + \\
&\qquad [\widehat{\theta} \text{ 的 bias (systematic error)}]^2
\end{aligned}
$$

另外,

$$\text{RMSE}(\widehat{\theta}) \stackrel{\text{def}}{=} \sqrt{\text{MSE}(\widehat{\theta})}$$

保有原單位.

4. Accuracy: 有兩種混淆的定義:

(a) Accuracy = Bias (= Systematic Error) (= $E(\widehat{\theta}) - \theta$)

註. 稱估計量有高準確度 (high accuracy), 若其爲不偏的或僅有小的偏差; 有低準確度 (low accuracy), 若其有大的偏差.

(b) Accuracy = MSE (同時度量 bias 與 precision (或 systematic 與 random error))

註. 稱估計量有高準確度, 僅當有 "小的或無偏差" 以及高精密度 (high precision).

註. 由於 accuracy 有兩種不同的定義, 使用時易造成混淆, 故使用 bias 與 precision 以及相關的含意 (bias $\Leftrightarrow$ systematic error; precision $\Leftrightarrow$ random error ($\Rightarrow$variability)) 較合宜.

5. 例子 (不同偏差度 (bias) 與精密度 (precision) 的估計量) 一模擬實驗, 由 $N(10,1)$ 分布中隨機取樣出 100 個觀察值並以下列 4 種估計量估計母體的期望值 μ:

$$\hat{\mu}_1 = \bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i \text{ (unbiased)}$$

$$\hat{\mu}_2 = \frac{x_{(1)} + x_{(100)}}{2} \text{ (unbiased)}$$

$$\hat{\mu}_3 = \frac{1}{50}\sum_{i=51}^{100} x_{(i)} \text{ (biased)}$$

$$\hat{\mu}_4 = \frac{x_{(51)} + x_{(100)}}{2} \text{ (biased)}$$

執行模擬 500 次後, 4 種估計量的分布如下圖, 分別顯示出各自的偏差度 (bias) 與精密度 (precision):

五. 母體參數的母數信賴區間

提供決策者除了數字供其參考外; 更應說明這些數字的偏離度 (bias) 與精密度 (precision) (亦即, 這些數字 "擺動" 的程度). 其中一個常用的方法是提供母體參考數 (θ) 的估計量 ($\widehat{\theta}$) 之外, 並附上此參數的信賴區間, 其定義如下: **定義.** 母體參數 θ 的信賴區間是實數線上的一區間, 由樣本中的 n 個觀察值製造而得, 使得此區間有預先指定的 $(1-\alpha)100\%$ 的機率包含母體參數 θ, 其中 α 爲介於 0 與 1 之間的某比率 (通常小於 0.5). $(1-\alpha)100\%$ 稱作信賴區間的信賴係數 (confidence cofficient) 或信賴水平 (confidence level).

註 1. 信賴區間的假設條件：

- 樣本是某種隨機樣本.

- 觀察值彼此間是獨立的. (十一章與十二章會探討不符此假設的狀況: 時間序列資料與空間資料.)

  或

- 觀察值出自於某一特定的機率分布 (如, 常態, 對數常態, 二項, 卜松分布等). (根據此最後一個假設所得的信賴區間, 稱作母數信賴區間 (parametric confidence interval); 否則, 稱作無母數信賴區間 (nonparametric confidence interval).)

註 2. 信賴區間是一隨機區間 (random interval), 亦即, 區間的上 (或) 下界乃是根據樣本統計量 (sample statistics) 所求得的隨機變數. 在取特定樣本之前, 信賴區間將包含 θ 的機率為 $(1-\alpha)100\%$. 一當選取了特定樣本且根據此樣本求出了特定的信賴區間後, 此一特定區間會確定地包含 θ 或不包含 θ. 通常不知道信賴區間實際上是否包含 θ, 因為 θ 是未知的, 而這也是要去求信賴區間的理由!

註 3. 若重覆模擬實驗 N 次, 且每次實驗都取一樣本並根據此樣本製造一個 $(1-\alpha)100\%$ 的信賴區間, 則

實際包含 θ 的信賴區間的個數 $\sim B(N, 1-\alpha)$. (因爲有 N 個信賴區間, $P(\theta \in$ 信賴區間$) = 1-\alpha$, 且信賴區間是相互獨立的 (因爲重覆實驗).)

例. 由 $N(5,1)$ 選取大小爲 10 的樣本並製造 80% 的信賴區間. 重覆 100 次上述的模擬實驗, 並記錄包含 $\theta = 5$ 的信賴區間數. (如, 課本所述會有 80 個左右的區間包含 $\theta = 5$; 課本的結果是 84 個.)

1. 信賴水平的選取

   一般而言, 信賴區間的寬度 (width) 與樣本大小, 承襲於母體的變異程度以及所選取的信賴水平有關. 下列三種方法可提高信賴區間的效能 (亦即, 使信賴區間變小):

   (a) 增加樣本大小.

   (b) 降低信賴水平.

   (c) 改良取樣設計及處理協定以降低資料的變異度.

   無選取信賴水平的固定法則, 這是一個主觀的決策, 但通常選用 90% 或 95%.

2. 常態分布的期望值的信賴區間

設

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)$$

則樣本期望值 (sample mean)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \sim N(\mu, \sigma/\sqrt{n})$$

由 z-轉換,

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

設 z_p 爲 $N(0,1)$ 的第 p 個量分位數 (p^{th} quartile) ($p < 0.5$), 則

$$\begin{aligned} & P\left[z_p \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{1-p}\right] \\ = \ & P\left[z_p \leq N(0,1) \leq z_{1-p}\right] \\ = \ & P\left[N(0,1) \leq z_{1-p}] - P[N(0,1) \leq z_p\right] \\ = \ & (1-p) - p = 1 - 2p \end{aligned} \tag{3}$$

取 $p = \alpha/2$ 以及因爲常態分布對稱於期望值 μ ($\Rightarrow z_p = -z_{1-p}$), 式 (3) 相當於

$$P\left[-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

經過整理後, 亦相當於

$$P\left[\bar{x}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\leq\mu\leq\bar{x}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha \tag{4}$$

所以, 當標準差 σ 已知時, μ 的 $(1-\alpha)100\%$ 信賴區間:

$$\left[\bar{x}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

註 1. 通常 σ 未知, 所以

$$\left[\bar{x}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

並不實用, 可行的方法爲以樣本標準差 (sample sd) s (可由資料求得之量) 估計 σ 並將 z-統計量修改成 t-統計量

$$t=\frac{\bar{x}-\mu_{\bar{x}}}{\hat{\sigma}_{\bar{x}}}=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$$

(自由度爲 $(n-1)$ 的學生 (student's) 分布)

(註: t-分布對稱於 0 且自由度 $\uparrow\Rightarrow t$-分布 $\rightarrow N(0,1)$. 所以, 樣本大小夠大時, t-統計量 $\approx N(0,1)$.)

令 $t_{\nu,p}$ 爲自由度爲 ν 之 t-分布的第 p 個量分位數, 則類似於 (4) 式的推導 (亦即, 將 σ 與 z 轉換成 s 與 t), 可得

$$P\left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right] = 1-\alpha$$

因此, μ 的 $(1-\alpha)100\%$ 信賴區間:

$$\left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

(稱作雙邊信賴區間 (two-sided confidence interval))

註 2. 同理, 可得 μ 的 $(1-\alpha)100\%$ 單邊上信賴區間 (one-sided upper confedence interval):

$$\left[-\infty,\ \bar{x} + t_{n-1,1-\alpha}\frac{s}{\sqrt{n}}\right]$$

此乃因爲

$$\begin{aligned} 1-\alpha &= P\left[t_{n-1,\alpha} \le \frac{\bar{x}-\mu}{s/\sqrt{n}}\right] \\ &= P\left[\frac{\mu-\bar{x}}{s/\sqrt{n}} \le -t_{n-1,\alpha}\right] \\ &= P\left[\frac{\mu-\bar{x}}{s/\sqrt{n}} \le t_{n-1,1-\alpha}\right] \end{aligned}$$

(最後的等式成立是因爲 t 分布對稱於 0)

以及 μ 的 $(1-\alpha)100\%$ 單邊下信賴區間 (one-sided lower confidence interval):

$$\left[\bar{x} - t_{n-1,1-\alpha}\frac{s}{\sqrt{n}}, \infty\right]$$

理由如下:

$$\begin{aligned} 1-\alpha &= P\left[\frac{\bar{x}-\mu}{s/\sqrt{n}} \le t_{n-1,1-\alpha}\right] \\ &= P\left[-t_{n-1,1-\alpha} \le \frac{\mu-\bar{x}}{s/\sqrt{n}}\right] \end{aligned}$$

例. 對數轉換後 TcCB 濃度的期望值的信賴區間 : 針對資料架構 `new.epa.94b.tccb.df` 的資料以 MVUE 估計量求期望值與標準差的點估計, 得

-0.62 與 0.47 log(ppb). 另外, 可求得期望值的 95% 雙邊 CI: [-0.76, -0.48].

3. 常態分布的變異數的信賴區間

變異數的估計可達到下列目的, 故有其必要性:

- 量化母體期望值的點估計的不確定程度 (亦即, $Var(\bar{x}) = \sigma^2/n$, 其中 σ^2: 未知, 需要估計).
- 計算各種區間 (如, 預測 (prediction), 容忍 (tolerence) 區間) 及假設檢定所需的樣本大小時, 需要用到母體的變異數. 但此變異數是未知的, 一個好的構想是採用此變異數的某種上界, 而其中的一種可能上界就是變異數的上信賴極限 (upper confidence limit).

設

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)$$

且

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

則

$$\begin{aligned}\chi^2 &\stackrel{\text{def}}{=} \frac{(n-1)}{\sigma^2}s^2 \\ &= \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma^2}\right)^2 \sim \chi^2(n-1)\end{aligned}$$

(自由度爲 $(n-1)$ 的卡方分布)

令 $\chi^2_{\nu,p}$ 爲自由度爲 ν 的卡方分布的第 p 個量分位數, 則

$$\begin{aligned}&1-\alpha \\ &= P\left[\chi^2_{n-1,\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,1-\alpha/2}\right] \\ &= P\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right]\end{aligned}$$

所以, σ^2 的雙邊 $(1-\alpha)100\%$ 信賴區間:

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right]$$

同理, 得 σ^2 的單邊上 $(1-\alpha)100\%$ 信賴區間:

$$\left[0,\ \frac{(n-1)s^2}{\chi^2_{n-1,\alpha}}\right]$$

此乃因爲

$$
\begin{aligned}
1-\alpha &= P\left[\chi^2_{n-1,\alpha} \le \frac{(n-1)s^2}{\sigma^2}\right] \\
&= P\left[\sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,\alpha}}\right]
\end{aligned}
$$

以及 σ^2 的單邊下 $(1-\alpha)100\%$ 信賴區間:

$$
\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}},\ \infty\right]
$$

理由如下:

$$
\begin{aligned}
1-\alpha &= P\left[\frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,1-\alpha}\right] \\
&= P\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}} \le \sigma^2\right]
\end{aligned}
$$

註. 由於中央極限定理, 即使根本 (基) 機率分布 (underlying probability distribution) 不是常態分布, 期望值 μ 的信賴區間的公式:

$$
\left[\bar{x} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right]
$$

$$
\left[-\infty,\ \bar{x} + t_{n-1,1-\alpha}\frac{s}{\sqrt{n}}\right]
$$

以及

$$\left[\bar{x} - t_{n-1,1-\alpha}\frac{s}{\sqrt{n}},\ \infty\right]$$

依然適用, 只要樣本大小 "夠大". 但上述的 σ^2 的信賴區間對於 "根本機率分布偏離常態性" 是非常地敏感. 故不適用於其它分布的根本母體, 如根本母體是偏斜 (skewed) 時, 信賴區間將會太寬 (沒有實際的用途).

例. 對數轉換後 TcCB 濃度的變異數的信賴區間 : σ^2 的點估計為 0.22 以及 95% 的上 CI: [0, 0.32].

4. 母數信賴區間的一般公式

- 常態分布的期望值的信賴區間公式都是精確, 無誤差 (exact) 的, 因為常態分布的樣本期望值依然是常態分布以及常態分布和 student's t-分布之間的關係. 其他也有精確信賴區間公式的分布, 如, 二項, 指數和卜松分布.

- 通常一參數的信賴區間乃是基於 "常態近似" 而求得的, 也即, 假設估計量的分布近似於常態

分布, 即使估計量是取自於非常態分布的觀察值的函數 . (爲何可做此假設? 事實上, 在實務情況中均成立的某些條件下, 任一 MLE 的的分布均是漸近常態 (asymptatically normal) 分布.) 因此, θ 的近似 $(1-\alpha)100\%$ 信賴區間:

$$\left[\widehat{\theta}-z_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}},\ \widehat{\theta}+z_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\right] \tag{5}$$

其中 $\widehat{\theta}$ 爲 θ 的估計量, $\widehat{\sigma}_{\widehat{\theta}}$ 爲 $\widehat{\theta}$ 的標準差 $(\sigma_{\widehat{\theta}})$ 的估計量, 通常稱作估計量 $\widehat{\theta}$ 的標準誤差 (standard error).

(5) 式成立, 乃是因爲當 n 夠大時,

$$\frac{\widehat{\theta}-\theta}{\widehat{\sigma}_{\widehat{\theta}}}\approx N(0,1)$$

以及

$$P\left[-z_{1-\alpha/2}\le\frac{\widehat{\theta}-\theta}{\widehat{\sigma}_{\widehat{\theta}}}\le z_{1-\alpha/2}\right]\approx 1-\alpha$$

註. 一個較保守的信賴區間公式爲以 student's t-分布的量分位數 $(t_{\nu,p})$ 取代常態分布的量分位數 (z_p):

$$\left[\widehat{\theta}-t_{\nu,1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}},\ \widehat{\theta}+t_{\nu,1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\right] \tag{6}$$

其中 ν 爲與 θ 的估計相關連的自由度.

經常 (6) 式的信賴區間比 (5) 式的信賴區間寬; 但寬度的差異會隨著樣本大小的增大而逐漸消失.

5. 對數常態分布的期望值的信賴區間

有兩種方法可刻劃出對數常態分布的特質, 分別如下:

(a) 以對數轉換後的隨機變數 ($\sim N(\mu, \sigma)$) 的期望值 μ 與標準差 σ. (此時, 只需將資料取對數後, 以處理常態分布的方式用樣本期望值估計 μ 以及相關的公式而求得 μ 的信賴區間.)

(b) 以原隨機變數的期望值 θ 與 CV τ. (一個直觀的想法 (也是一些書上的建議): 計算對數轉換後資料的期望值的信賴區間, 再將信賴極限 (confidence limits) (亦即,信賴區間的端點) 指數化; 事實上, 此方法得到的應是中位數而不是期望值的信賴區間.) 有關計算對數常態分布的期望值 θ 的估計值以及信賴區間的一些方法和討論它們利弊的例子, 分別說明如後.

- 藍德氏精確法 (Land's Exact Method)

Land (1971, 1975) 求得計算 θ 的單邊 (上或下) 一致最精確不偏信賴區間 (uniformly most accurate unbiased confidence interval). 雙邊的信賴區間可結合最佳下信賴極限與最佳上信賴極限而得之, 此種求得的過程僅是漸近最佳的 (asymptotically optimal), 但對大多數的目標而言應是可接受的. 公式如下:

θ 的單邊上 $(1-\alpha)100\%$ 信賴區間:

$$\left\{0,\ \exp\left[\left(\bar{y}+\frac{s_y^2}{2}\right)+s_y\frac{H_{1-\alpha}}{\sqrt{n-1}}\right]\right\}$$

θ 的單邊下 $(1-\alpha)100\%$ 信賴區間:

$$\left\{\exp\left[\left(\bar{y}+\frac{s_y^2}{2}\right)+s_y\frac{H_{\alpha}}{\sqrt{n-1}}\right],\ \infty\right\}$$

θ 的雙邊 $(1-\alpha)100\%$ 信賴區間:

$$\left\{\exp\left[\left(\bar{y}+\frac{s_y^2}{2}\right)+s_y\frac{H_{\alpha/2}}{\sqrt{n-1}}\right],\right.$$
$$\left.\exp\left[\left(\bar{y}+\frac{s_y^2}{2}\right)+s_y\frac{H_{1-\alpha/2}}{\sqrt{n-1}}\right]\right\}$$

其中

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

$$y_i = \log(x_i),\ i = 1, 2, \ldots, n$$

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

H_p 內含複雜的數學, 可由 H-統計量的表格查得對應的近似值.

- 其它方法

  由 Parkin et al., Cox, El-Shaarawv 以及 Gilbert 所建議的方法均詳述於環統模組的輔助資料夾 **`elnorm.alt`** 內.

  Parkin et al. 透過蒙地卡羅模擬法探討各種製造對數常態的期望值的信賴區間方法的效能與 Land's Exact Optimal 方法比較後的結論如下:

(a) 小樣本 ($n \le 20$) 時, 無一近似方法 (對立於 Land's 方法的其他方法) 是非常的好.

(b) $n > 20$ 時, Parkin et al. 所建議的方法提供的涵蓋面 (coverage) 與假設涵蓋面 (assumed coverage) 相當合理地接近.

(c) $n > 60$ 時, Cox 的方法相當的好, 並且針對高偏斜母體時, 略優於 Parkin et al. 的方法.

例 1. Reference Area TcCB 濃度的期望值的信賴區間

經由環統模組的計算, 期望值的點估計為 0.66 ppb, 各種方法的 95% 雙邊信賴區間如下表:

信賴區間	方法
[0.52, 0.70]	Land
[0.50, 0.74]	Parkin et al.
[0.52, 0.68]	Cox
[0.51, 0.68]	常態近似

樣本大小 $n = 47$, 各種方法差不多都相同.

例 2. 鉻 (chromium) 濃度的期望值的上信賴極限

向量 `epa.92d.chromium.vec` 內含 15 個鉻濃度 (mg/kg) 的觀察值, 假設均取自於對數常態分布. 各種方法的 95% 上信賴區間如下表:

信賴區間	方法
[0, 497]	Land
[0, 230]	Parkin et al.
[0, 366]	Cox
[0, 261]	常態近似

相當大的差異: Land 的方法求得的上信賴極限差不多是 Parkin et al. 與常態近似方法的兩倍.

(注意: 資料中有一個觀察值爲 1300 $\gg$ 其它的觀察值.)

例 3. 小樣本時, 期望值的上信賴極限

由期望值 $\theta = 10$, CV $\tau = 2$ 的對數常態分布生成 10 個觀察值並比較各種方法所求得的 95% 的期望值的上信賴區間. 結果如下 (取種子 seed =23):

信賴區間	方法
[0, 182]	Land
[0, 96]	Parkin et al.
[0, 68]	Cox
[0, 28]	常態近似

討論:

– 資料中最大的觀察值爲95.8133610.

– Land 的方法中最令人不安的特性: 上信賴極限差不多是最大觀察值的 2 倍; 事實上, 在小樣本時, 由 Land 的方法所求得的上信賴極限會比最大觀察值多一個或以上個倍數.

– 雖然 Land 的方法是精確的, 並有最佳的性質, 卻對 "資料取自於對數常態分布" 的假

設非常地敏感, 與對數常態分布的偏差 (如, 外觀類似於對數常態分布的混合分布) 會導致相當大的上信賴極限.

– 模擬研究顯示的結果: 當樣本大小 ≤ 30 時, Land 的方法會製造出 "無法接受的高" (unacceptably high) 的上信賴極限, 特別是當 CV > 1 時更爲明顯, 建議採用無母數方法, 如環靴法 (bootstrap, 再取樣法).

6. 偏斜分布的期望值的單邊信賴區間

Chen (1995) 發展出一種修正的 t-統計量來執行偏斜分布的期望值的單邊假設檢定. 基於假設檢定與信賴區間的關係 (於第七章交待), 可採用 Chen 的方法製造偏斜分布的期望值的單邊信賴區間如下:

(a) 正偏斜分布的期望值的單邊下信賴區間, 如圖:

(b) 負偏斜分布的期望值的單邊上信賴區間, 如圖:

註. 由於環境資料通常爲正偏斜且經常需要雙邊的信賴極限或單邊的上信賴極限, 在實務上 Chen 的

製造信賴區間的方法不太被採用. 然而 Chen 的假設檢定可被發展成解決某一特定環統問題的方法 (詳見第七章). 例. 鉻濃度的期望值的下信賴極限

以向量 `epa.92d.chromium.vec`, 根據各種方法製造的期望值的 95% 下信賴區間如下表:

信賴區間	方法
$[95, \infty]$	Land
$[110, \infty]$	Parkin et al.
$[82, \infty]$	Cox
$[58, \infty]$	常態近似
$[97, \infty]$	Chen

Chen 的方法製造出的下信賴極限與 Land 的相當接近 (一致).

7. 二項分布成功率的信賴區間

設 $X \sim B(n, p)$. 成功機率 p 的估計量 $\widehat{p} = x/n$ (亦即, 成功的次數/試驗的總數).

兩種製造成功機率的信賴區間的方法: 精確 (exact) 方法與常態近似法, 分述如下:

- 精確信賴區間

p 的精確 $(1-\alpha)100\%$ 信賴區間:

$$[\mathrm{LCL}_{\alpha/2},\ \mathrm{UCL}_{\alpha/2}]$$

其中兩個信賴極限 $\mathrm{LCL}_{\alpha/2}$, $\mathrm{UCL}_{\alpha/2}$滿足:

$$P(X \geq x | p = \mathrm{LCL}_{\alpha/2}) = \alpha/2$$

亦即, $\mathrm{LCL}_{\alpha/2}$ 是最小的 p 使得

$$P(X \geq \text{ 實際觀察值 } x) \geq \alpha/2$$

(事實上, 等於 $\alpha/2$), 因爲 p 愈大, 上述機率愈大於 $\alpha/2$.

而且

$$P(X \leq x | p = \mathrm{UCL}_{\alpha/2}) = \alpha/2$$

亦即, $\mathrm{UCL}_{\alpha/2}$ 是最大的 p 使得

$$P(X \leq \text{ 實際觀察值 } x) \leq \alpha/2$$

(事實上, 等於 $\alpha/2$), 因爲 p 愈小, 上述機率愈小於 $\alpha/2$.

另外, 可証明 (看參考文獻):

$$\mathrm{LCL}_{\alpha/2} = \frac{x}{x + (n - x + 1)F_{\nu_1,\nu_2,1-\alpha/2}} \tag{7}$$

$$\begin{aligned}&\text{UCL}_{\alpha/2}\\&=\frac{(x+1)F_{\nu_2+2,\nu_1-2,1-\alpha/2}}{n-x+(x+1)F_{\nu_2+2,\nu_1-2,1-\alpha/2}}\end{aligned} \tag{8}$$

其中

$$\nu_1 = 2(n-x+1),\ \nu_2 = 2x$$

且

$$F_{\nu_1,\nu_2,r}$$

爲自由度爲 ν_1 與 ν_2 的 F 分布的第 r 個量分位數.

註 1. p 的單邊 $(1-\alpha)100\%$ 下信賴區間:

$$[\text{LCL}_\alpha,\ 1]$$

其中

$$\text{LCL}_\alpha = \frac{x}{x+(n-x+1)F_{\nu_1,\nu_2,1-\alpha}}$$

且 $\nu_1 = 2(n-x+1), \nu_2 = 2x$.

另外, p 的單邊 $(1-\alpha)100\%$ 上信賴區間:

$$[0,\ \text{UCL}_\alpha]$$

其中

$$\text{UCL}_\alpha = \frac{(x+1)F_{\nu_2+2,\nu_1-2,1-\alpha}}{n-x+(x+1)F_{\nu_2+2,\nu_1-2,1-\alpha}}$$

且 $\nu_1 = 2(n-x+1)$, $\nu_2 = 2x$.

註 2. 當 $x = 0$ (亦即, 所有的觀察值均為失敗), p 的上 $(1-\alpha)100\%$ 信賴極限

$$\text{UCL}_\alpha = 1 - \alpha^{1/n}$$

Why? 因為此時一個自由度會為 0, 故無法用 (7) 式與 (8) 式, 但可直接求之: 找最大的 p 使得

$$P(X \le 0) \le \alpha$$

此乃相當於

$$P(X = 0) = \alpha \Leftrightarrow (1-p)^n = \alpha$$

故,

$$1 - p = \alpha^{1/n}$$

因此,

$$\text{UCL}_\alpha = 1 - \alpha^{1/n}$$

同理, 當 $x = n$ (亦即, 所有的觀察值均成功), p 的下 $(1 - \alpha)100\%$ 信賴極限

$$\mathrm{LCL}_{\alpha} = \alpha^{1/n}$$

Why? 找最小的 p 使得

$$P(X \geq n) \geq \alpha$$

亦相當於

$$P(X = n) = \alpha \Leftrightarrow p^n = \alpha$$

由此得

$$p = \alpha^{1/n}$$

因此,

$$\mathrm{LCL}_{\alpha} = \alpha^{1/n}$$

- 常態近似的信賴區間

  根據 $\widehat{p}$ 的常態近似, p 的近似 $(1 - \alpha)100\%$ 信賴區間:

$$\left[\widehat{p} - z_{1-\alpha/2}\widehat{\sigma}_{\widehat{p}},\ \widehat{p} + z_{1-\alpha/2}\widehat{\sigma}_{\widehat{p}}\right]$$

其中

$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\widehat{p}(1 - \widehat{p})/n}$$

此乃因爲

$$\widehat{p}=\frac{x}{n},\ \sigma_{\widehat{p}}=\sqrt{\frac{np(1-p)}{n^2}}=\sqrt{\frac{p(1-p)}{n}}$$

註 1. 透過連續性的修正 (continuity correction) 可使得二項分布的曲線更接近常態分布的曲線, 對應的 $(1-\alpha)$100% 近似信賴區間:

$$\left[\widehat{p}-z_{1-\alpha/2}\widehat{\sigma}_{\widehat{p}}-\frac{1}{2n},\ \widehat{p}+z_{1-\alpha/2}\widehat{\sigma}_{\widehat{p}}+\frac{1}{2n}\right]$$

註 2. 小樣本時, 常態近似不理想, 特別當 np 或 $n(1-p)<5$ 時或當 p 太小 ($p<0.2$) 或 p 太大 ($p>0.8$) 時, 更爲明顯. 既然有精確的方法可用, 實在沒必要使用常態近似.

例. 無法偵測值 (nondetect) 的機率的信賴區間

在資料架構 `epa.92c.benezenel.df` 中, 無法偵測的觀察值記錄成 "<2", 在全部 36 個觀察值中共有 33 個, 約爲 92%. 以精確方法與常態近似法求得的觀察到無法偵側值的機率的 95% 信賴區間如下表:

信賴區間	方法
[0.78, 0.98]	精確法
[0.83, 1]	常態近似(無連續性修正)
[0.81, 1]	常態近似(連續性修正)

兩個常態近似的方法都產生不可能的上界 1 (因爲若 $p = 1$, 則不可能觀察到任何偵測值).

8. 卜松分布的期望值的信賴區間

卜松分布可用於模型化 "給定時段中違反污染標準的次數" 以及模型化 "多數觀察值爲不可偵測, 僅少數爲可偵測的化學物濃度的分布". 設

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$$

則 λ 的估計量

$$\hat{\lambda}_{(mme/mle/mvue)} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

共有 3 種求 λ 的信賴區間的方法, 分述如下:

- 精確的信賴區間

令 $Y = \sum_{i=1}^{n} X_i$, 則

$$Y \sim \text{Poisson}(n\lambda)$$

(因爲獨立卜松隨機變數的和依然爲卜松分布且參數爲對應參數的和). 因爲 Y 是事件發生的

總數, 透過 Y 的分布以及類似於二項分布成功機率的精確信賴區間的求法, λ 的精確 $(1-\alpha)100\%$ 信賴區間:

$$\left[\mathsf{LCL}_{\alpha/2},\ \mathsf{UCL}_{\alpha/2}\right]$$

其中兩個信賴極限 $\mathsf{LCL}_{\alpha/2}$ 與 $\mathsf{UCL}_{\alpha/2}$ 滿足

$$P\left(Y \geq y|\lambda = \mathsf{LCL}_{\alpha/2}\right) = \alpha/2$$

(亦即, $\mathsf{LCL}_{\alpha/2}$ 是最小的 λ 使得

$$P(Y \geq \text{實際觀察值 } y) \geq \alpha/2$$

(事實上等於 $\alpha/2$), 因為 λ 愈大, Y 愈大, 得上述機率會大於 $\alpha/2$), 且

$$P\left(Y \leq y|\lambda = \mathsf{UCL}_{\alpha/2}\right) = \alpha/2$$

(亦即, $\mathsf{UCL}_{\alpha/2}$ 是最大的 λ 使得

$$P(Y \leq \text{實際觀察值 } y) \leq \alpha/2$$

(事實上等於 $\alpha/2$), 因為 λ 愈小, Y 愈小, 得上述機率會小於 $\alpha/2$).

註 1. λ 的單邊下 $(1-\alpha)100\%$ 信賴區間:

$$[\mathsf{LCL}_{\alpha},\ \infty]$$

其中 LCL_α 滿足

$$P(Y \geq y|\lambda = \mathsf{LCL}_\alpha) = \alpha$$

另外, λ 的單邊上 $(1-\alpha)100\%$ 信賴區間:

$$[0,\ \mathsf{UCL}_\alpha]$$

其中 UCL_α 滿足

$$P(Y \leq y|\lambda = \mathsf{UCL}_\alpha) = \alpha$$

註 2. 當 $y = 0$ (也就是說, 沒有觀察到任何事件) 時,

$$\mathsf{UCL}_\alpha = -\frac{\log \alpha}{n}$$

Why? 找最大的 λ 使得

$$P(Y \leq 0) = \alpha \Leftrightarrow P(Y = 0) = \alpha$$

由此導出

$$e^{-\lambda n} = \alpha$$

(因爲, $Y \sim \text{Poisson}(\lambda n)$.) 因此,

$$\lambda = -\frac{\log \alpha}{n}$$

故,

$$\mathrm{UCL}_\alpha = -\frac{\log\alpha}{n}$$

- Pearson-Hartley 近似的信賴區間

  λ 的 Pearson-Hartley 近似的雙邊 $(1-\alpha)100\%$ 信賴區間:

$$\left[\frac{\chi^2_{2y,\alpha/2}}{2n},\ \frac{\chi^2_{2y+2,1-\alpha/2}}{2n}\right]$$

  其中 $\chi_{\nu,p}$ 爲自由度爲 ν 的卡方分布的第 p 個量分位數.

  λ 的 Pearson-Hartley 近似下單邊 $(1-\alpha)100\%$ 信賴區間:

$$\left[\frac{\chi^2_{2y,\alpha}}{2n},\ \infty\right]$$

  λ 的 Pearson-Hartley 近似上單邊 $(1-\alpha)100\%$ 信賴區間:

$$\left[0,\ \frac{\chi^2_{2y+2,1-\alpha}}{2n}\right]$$

- 常態近似的信賴區間

根據 $\hat{\lambda}$ 的常態近似, λ 的近似 $(1-\alpha)100\%$ 信賴區間:

$$\left[\hat{\lambda} - z_{1-\alpha/2}\hat{\sigma}_{\hat{\lambda}},\ \hat{\lambda} + z_{1-\alpha/2}\hat{\sigma}_{\hat{\lambda}}\right]$$

其中

$$\hat{\sigma}_{\hat{\lambda}} = \sqrt{\frac{\hat{\lambda}}{n}}$$

因爲 $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\sigma_{\hat{\lambda}} = \sqrt{\frac{Var(x_i)}{n}} = \sqrt{\frac{\lambda}{n}}$.

註. 常態近似不如 Pearson-Hartley 近似. 既然有精確方法, 故無實際的需要去使用兩種近似法.

例. 苯平均濃度的信賴區間

因爲資料中, 有大多數的無法偵測值 (36 個中有 32 個), USEPA 假設此組資料可以用卜松過程模型化. 四個偵測值分別代表在相當大量的水分子中, 苯的分子數量. 32 個無法偵測值也可表示成苯的分子數, 分別爲 0 或 1 (因爲它們被記錄成 "$<$ 2" ppb). USEPA 建議設定所有的無法偵測值均爲 1 個苯分子. 根據此設定後的資料, λ 的點估計爲 1.94 ppb, 由精確法與近似法求得的苯的平均濃度的上單邊 95% 信賴區間如下表:

信賴區間	方法
[0, 2.37]	精確法
[0, 2.37]	Pearson-Hartley 近似法
[0, 2.33]	常態近似法

註 1. 將無法偵測値設定成 0 或 2, 可作爲敏感性分析: 爲 0 時, 精確的上信賴極限爲 1.35 ppb; 爲 2 時, 精確的上信賴極限爲 3.37 ppb.

註 2. 以卜松分布模型化此組資料的合宜性並不顯然, 也不常被接受. 處理無法偵測値的方法是相當主觀的. 無法偵測値設定成 1 的資料的卜松 Q-Q 圖反應出卜松模型並不適當. 更愼重起見, 根據環統選單 Q-Q plot Gestalt 所產生的多個典型卜松 Q-Q 圖, 也相當明顯地指出卜松模型並不適宜.

六. 基於環靴法 (Bootstrapping) 的無母數信賴區間

製造母數信賴區間需要下列三種假設:

1. 樣本是某種隨機樣本.

2. 觀察值間是彼此互相獨立的.

3. 觀察值來自於某一特定的機率分布 (如, 常態, 對數常態, 二項, 卜松等).

概念:

假設某一特定的母體分布

$$\Downarrow$$

參數 (θ) 的估計量 ($\widehat{\theta}$) 的分布

$$\Downarrow$$

θ 的信賴區間

如, $N(\mu, \sigma) \Rightarrow \widehat{\mu} = \bar{x} \sim N(\mu, \sigma/\sqrt{n}) \Rightarrow \mu$ 的信賴區間.

以此推廣出製造無母數信賴區間的概念:

以經驗累積分布函數 (empirical cdf) 估計母體 cdf

$$\Downarrow$$

以 empirical cdf 估計 $\widehat{\theta}$ 的分布:

經由從 empirical cdf 隨機取樣數回,

以及每回都基於所取的隨機樣本求出點估計而得
(亦即, 由多個點估計導出 $\widehat{\theta}$ 的估計分布)

$$\Downarrow$$

θ 的信賴區間

Efron 稱此概念爲 ”bootstrap” (環靴法) (自力救濟法), 取自於一諺語 ”pulling yoursslf up by your own bootstraps” (靠自己力量而成功) 的含意.

環靴法的基本步驟如下:

1. 根據資料估計參數.

2. 以放回取樣的方式由資料中取樣 B 回, 再根據每回取出的樣本 (稱作環靴法樣本, bootstrap sample) 估計參數.

3. 用步驟 1 得到的參數估計值以及步驟 2 得到的估計量的環靴法分布來製造參數的信賴區間.

註 1. 以上 3 步驟完全沒有倚賴特定參數分布的假設.

註 2. 於步驟 2 使用經驗機率估計量 (empirical probability estimator)

$$\widehat{p}_i = \frac{\#\left[x_j \leq x_{(i)}\right]}{n}$$

(亦即, empirical cdf 的描點位置 (plotting position))

1. 製造環靴法信賴區間的方法

   討論 4 種製造環靴法信賴區間的方法:

   (a) 標準常態法 (Standard Normal Method)

   (b) 經驗百分位數法 (Empirical Percentile Method)

   (c) 偏差修正百分位數法 (Bias-Corrected Percentile Method)

   (d) 偏差修正暨調節百分位數法 (Bias-Corrected and Adjusted Percentile Method)

   其中 S+ 是根據第 2 與第 4 種方法計算信賴區間.

符號定義:

θ = 欲估計的參數

$\widehat{\theta}$ = 基於最初樣本而得的 θ 的估計

$\widehat{\theta}_b$ = 根據第 b 個環靴法樣本而得的 θ 的估計

$\widehat{\theta}^*$ = 一隨機變數, 其分布爲根據給定樣本 (original sample) 而得的 $\widehat{\theta}$ 的環靴法分布

- 標準常態法 (Standard Normal Method)

  設 $\widehat{\theta}$ 近似於平均値爲 θ 的常態分布, 則根據常態近似, θ 的雙邊 $(1-\alpha)100\%$ 信賴區間：

  $$\left[\widehat{\theta} - z_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}},\ \widehat{\theta} + z_{1-\alpha/2}\widehat{\sigma}_{\widehat{\theta}}\right]$$

  接著需要計算 $\widehat{\theta}$ 的標準差的估計值 $\widehat{\sigma}_{\widehat{\theta}}$, 求法如下:

(a) 生成 B 個環靴法樣本.

(b) 根據 (a) 中的每個環靴法樣本, 計算 θ 的估計值 $\widehat{\theta}_b$, 共得 B 個 θ 的估計值 $\widehat{\theta}_b, b = 1, \ldots B$.

(c) 根據 (b) 中的 B 個 θ 的估計值 θ_b, 計算 $\widehat{\theta}$ 的樣本標準差, 以此作爲 $\widehat{\sigma}_{\widehat{\theta}}$.

因此, $\widehat{\theta}$ 的標準差的估計值

$$\widehat{\sigma}_{\widehat{\theta}} = \sqrt{\frac{\sum_{b=1}^{B}\left(\widehat{\theta}_b - \bar{\widehat{\theta}}\right)^2}{B-1}}$$

其中

$$\bar{\widehat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b$$

註. 通常不超過 $B = 200$ 個環靴法就可得到 $\widehat{\theta}$ 的好的標準差的估計值 $\widehat{\sigma}_{\widehat{\theta}}$.

- 經驗百分位數法 (Empirical Percentile Method)

  假設 $\widehat{\theta}$ 的某一單調轉換 (monotonic transformation), 如 $f(\widehat{\theta})$ 的分布是近似於期望值爲 $f(\theta)$, 標準差爲 1 的常態分布 (即, $f(\widehat{\theta}) \approx N(f(\theta), 1)$). (註. $\widehat{\theta}$ 的分布不必要近似於期望值爲 θ 的常態分布.)

  則 $f(\theta)$ 的雙邊 $(1-\alpha)100\%$ 信賴區間:

$$\left[f(\widehat{\theta}) - z_{1-\alpha/2},\ f(\widehat{\theta}) + z_{1-\alpha/2}\right]$$

  有待解答的兩個問題:

  (1) θ 的信賴區間爲何?

(2) f 的型式爲何?

處理過程如下:

$$f(\widehat{\theta}) \approx N(f(\theta), 1) \text{ (假設條件)}$$

$$\Downarrow$$

$f(\widehat{\theta})$ 的環靴法分布 $\approx N(f(\widehat{\theta}), 1)$

(假設 f 的型式已知, 透過 $\widehat{\theta}$ 的環靴法分布, 經由 f 轉換可得)

$$\Downarrow$$

$f(\widehat{\theta}) + z_{1-\alpha/2} \approx f(\widehat{\theta})$ 的環靴法分布的 $(1-\alpha/2)$ 量分位數. (因爲若 $X \sim N(\mu, \sigma)$, 則 $x_p = \mu + \sigma z_p$

(因爲 $P(X \leq x_p) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x_p-\mu}{\sigma}\right)$

$= P(N(0,1) \leq z_p) = p)$,

以及 $f(\widehat{\theta})$ 的環靴法分布 $\approx N(f(\widehat{\theta}), 1))$

另外, 也可導出

$f(\widehat{\theta}) - z_{1-\alpha/2} = f(\widehat{\theta}) + z_{\alpha/2}$ (因爲 $-z_{1-\alpha/2} = z_{\alpha/2}$) $\approx$ $f(\widehat{\theta})$ 的環靴法分布的 $\alpha/2$ 量方位數.

又 f 是單調且百分位數在單調轉換下是不變的 (invariant), 所以 $f(\theta)$ 落在區間

$$[f(\widehat{\theta}) \text{ 的環靴法分布的 } \alpha/2 \text{ 量分位數},$$

$$f(\widehat{\theta}) \text{ 的環靴法分布的 } 1-\alpha/2 \text{ 量分位數}]$$

內也就相當於 θ 落在區間

$$[\widehat{\theta} \text{ 的環靴法分布的 } \alpha/2 \text{ 量分位數},$$

$$\widehat{\theta} \text{ 的環靴法分布的 } 1-\alpha/2 \text{ 量分位數}]$$

內. 故 θ 的近似雙邊 $(1-\alpha)100\%$ 信賴區間爲

$$[\widehat{\theta}^*_{\alpha/2},\ \widehat{\theta}^*_{1-\alpha/2}]$$

其中 $\widehat{\theta}^*_p$ 爲 $\widehat{\theta}$ 的環靴法分布的第 p 個量分位數. (亦即, 只需 $\widehat{\theta}$ 的環靴法分布以及對應的量分位數 (或稱經驗百分位數, empirical percentiles) 即可, 完全不需要知道 f 的型式.) 如, θ 的雙邊 95% 信賴區間:

$$[\widehat{\theta} \text{ 的環靴法分布的第 } 2.5 \text{ 個百分位數觀察值},$$

$$\widehat{\theta} \text{ 的環靴法分布的第 } 97.5 \text{ 個量分位數觀察值}]$$

註. 製造 90% 信賴區間時, 至少需用 $B=1000$ 個環靴法 (再取樣, resamples); 95% 信賴區間時, 至少需 $B=2000$ 個環靴

法. 這些要求亦適用於下面討論的另二種方法: "偏差修正百分位數法 (bias-corrected percentile method)" 與 "偏差修正暨調節百分位數法 (bias-corrected and adjusted percentile method)".

- 偏差修正百分位數法 (Bias-Corrected Percentile Method)

  若 $f(\widehat{\theta})$ 是 $f(\theta)$ 的偏差估計量時, 經驗百分位數法 (empirical percentile method) 的效果不好, 此時 $\widehat{\theta}$ 的環靴法分布的中位數不夠靠近 $\widehat{\theta}$. 需要將經驗百分位數法修正成偏差修正百分位數法以克服困難, 做法如下:

  更一般的假設: $f(\widehat{\theta})$ 的分布是近似於期望值爲 $f(\theta) - z_0$, 標準差爲 1 的常態分布 (亦即, $f(\widehat{\theta}) \approx N(f(\theta), 1)$, 其中偏差 z_0 爲一常數). 則根據常態近似, $f(\theta)$ 的 $(1-\alpha)100\%$ 信賴區間如下:

  $$[f(\widehat{\theta}) + z_0 - z_{1-\alpha/2},\ f(\widehat{\theta}) + z_0 + z_{1-\alpha/2}]$$

  此乃因爲

  $$P(z_{\alpha/2} \le f(\widehat{\theta}) - f(\theta) + z_0 \le z_{1-\alpha/2}) \approx 1 - \alpha$$

亦相當於

$$P(-z_{1-\alpha/2} \le f(\theta) - f(\widehat{\theta}) - z_0 \le z_{1-\alpha/2}) \approx 1-\alpha$$

也就是說

$$P(f(\widehat{\theta}) + z_0 - z_{1-\alpha/2} \le f(\theta) \le f(\widehat{\theta}) + z_0 + z_{1-\alpha/2}) \approx 1-\alpha$$

但我們的目標是求 θ 的信賴區間並且另一個待解決的問題是 f 的型式與偏差 z_0 的值爲何? 處理過程如下:

$$f(\widehat{\theta}) \approx N(f(\theta) - z_0, 1) \text{ (假設條件)}$$

$$\Downarrow$$

$f(\widehat{\theta})$ 的環靴法分布 $\approx N(f(\widehat{\theta}) - z_0, 1)$

(假設 f 的型式已知, 透過 $\widehat{\theta}$ 的環靴法分布經由 f 轉換可得)

$$\Downarrow$$

$$\begin{aligned} f(\widehat{\theta}) + z_0 + z_{1-\alpha/2} &= (f(\widehat{\theta}) - z_0) + 1 \cdot (2z_0 + z_{1-\alpha/2}) \\ &= (f(\widehat{\theta}) - z_0) + 1 \cdot z_{p_U} \end{aligned}$$

$\approx f(\widehat{\theta})$ 的環靴法分布的第 p_U 個量分位數

其中 $p_U \overset{\text{def}}{=} \Phi(2z_0 + z_{1-\alpha/2})$,

以及

$$
\begin{aligned}
& f(\widehat{\theta}) + z_0 - z_{1-\alpha/2} \\
& \quad = (f(\widehat{\theta}) - z_0) + 1(2z_0 - z_{1-\alpha/2}) \\
& \quad = (f(\widehat{\theta}) - z_0) + 1(2z_0 + z_{\alpha/2}) \\
& \quad = (f(\widehat{\theta}) - z_0) + 1z_{p_L} \\
& \quad \approx f(\widehat{\theta}) \text{ 的環靴法分布的第 } p_L \text{ 個量分位數}
\end{aligned}
$$

其中 $p_L \overset{\text{def}}{=} \Phi(2z_0 + z_{\alpha/2})$.

又 f 是單調且百分位數在單調轉換下是不變的 (invariant), 所以, $f(\theta)$ 落在區間

[$f(\widehat{\theta})$ 的環靴法的第 p_L個量分位數,

$f(\widehat{\theta})$ 的環靴法的第 p_U 個量分位數]

內, 就相當於 θ 落在如下的區間內:

[$\widehat{\theta}$ 的環靴法分布的第 p_L 個量分位數,

$\widehat{\theta}$ 的環靴法分布的第 p_U 個量分位數]

所以, θ 的近似雙邊 $(1-\alpha)100\%$ 信賴區間如下:

$$[\widehat{\theta}^*_{p_L},\ \widehat{\theta}^*_{p_U}]$$

其中 $\widehat{\theta}_p^*$ 爲 $\widehat{\theta}$ 的環靴法分布的第 p 個量分位數,

$$p_L = \Phi(2z_0 + z_{\alpha/2})$$

且

$$p_U = \Phi(2z_0 + z_{1-\alpha/2})$$

至此, 信賴區間與 f 的型式無關 (亦即, 不需要知道 f 的型式), 但仍然需要知道偏差 z_0 的值.

如何估計 z_0? 先看看 z_0 在原分布的眞正值爲何?

由假設 $f(\widehat{\theta}) \approx N(f(\theta) - z_0, 1)$, 得

$$f(\widehat{\theta}) - f(\theta) + z_0 \approx N(0, 1)$$

以及

$$\begin{aligned} \Phi(z_0) &= P(N(0,1) \le z_0) \\ &\approx P(f(\widehat{\theta}) - f(\theta) + z_0 \le z_0) \\ &= P(f(\widehat{\theta}) \le f(\theta)) \\ &= P(\widehat{\theta} \le \theta) \text{ (因爲 } f \text{ 爲單調)} \end{aligned}$$

所以, 令

$$p = P(\widehat{\theta} \le \theta) \tag{9}$$

則

$$\Phi(z_0) \approx p$$

因此,

$$z_0 \approx \Phi^{-1}(p) \tag{10}$$

(註. 若 θ 爲 $\widehat{\theta}$ 的分布的中位數, 則 $p = 1/2$, 故 $z_0 \approx \Phi^{-1}(\frac{1}{2}) = 0$, 回到經驗百分位數法的情況.) 接著, 根據 (9) 式, 以 θ 的環靴法估計值 $\widehat{\theta}_b \leq \widehat{\theta}$ 的比率估計 p, 得

$$\widehat{p} = \frac{1}{B}\sum_{b=1}^{B} I(\widehat{\theta}_b \leq \widehat{\theta})$$

其中指標函數 (indicator function)

$$I(x) = \begin{cases} 1, \text{ 若 } x \text{ 爲眞} \\ 0, \text{ 若 } x \text{ 爲假} \end{cases}$$

(此乃因爲由大數法則, (9) 式的含意就爲 $\widehat{\theta} \leq \theta$ 的比率, 又爲無母數估計, 故以環靴法估計值估計之.) 因此, 可根據 (10) 式估計 z_0, 得

$$\widehat{z}_0 = \Phi^{-1}(\widehat{p})$$

最後, 很自然的 θ 的近似雙邊 $(1-\alpha)100\%$ 信賴區間爲

$$[\widehat{\theta}^*_{\widehat{p}_L},\ \widehat{\theta}^*_{\widehat{p}_U}]$$

其中

$$\widehat{p}_L = \Phi(2\widehat{z}_0 + z_{\alpha/2})$$

$$\widehat{p}_U = \Phi(2\widehat{z}_0 + z_{1-\alpha/2})$$

- 偏差修正暨調節百分位數法 (Bias-Corrected and Adjusted Percentile Method)

  爲偏差修正百分位數法 (bias-corrected percentile method) 的修正, 其假設更爲一般化: $\widehat{\theta}$ 的單調轉換 $f(\widehat{\theta})$ 的分布是近似於期望值爲 $f(\theta) - z_0[1 + af(\theta)]$, 標準差爲 $1 + af(\theta)$ 的常態分布, 亦即,

  $$f(\theta) \approx N(f(\theta) - z_0[1 + af(\theta)], 1 + af(\theta))$$

  其中偏差 z_0 與加速量 (acceleration) a 均爲常數.

  (註. 加速量常數 a 使得 $f(\widehat{\theta})$ 的標準差爲 $f(\theta)$ 的線性函數; 當 $a = 0$ 時, 回到偏差修正百分位數法.)

  此時, θ 的近似雙邊 $(1 - \alpha)100\%$ 信賴區間如下:

  $$[\widehat{\theta}^*_{\widehat{p}_L},\ \widehat{\theta}^*_{\widehat{p}_U}]$$

  其中

$$
\begin{aligned}
\widehat{p}_L &= \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + z_{\alpha/2})}\right) \\
\widehat{p}_U &= \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{1-\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + z_{1-\alpha/2})}\right) \\
\widehat{z}_0 &= \Phi^{-1}(\widehat{p}),\ \widehat{p} = \frac{1}{B}\sum_{b=1}^{B} I(\widehat{\theta}_b \leq \widehat{\theta}\,) \\
\widehat{a} &= \frac{\sum_{i=1}^{n}(\widehat{\theta} - \widehat{\theta}_{-i})^3}{6\left[\sum_{i=1}^{n}(\widehat{\theta} - \widehat{\theta}_{-i})^2\right]^{3/2}},
\end{aligned}
$$

$\widehat{\theta}_{-i} =$ 去掉第 i 個觀察值後,
θ 的估計值

2. 應採用何種環靴法?

四種製造環靴法信賴區間的方法是以假設條件的限制性多寡而順序地討論. S+ 採用 "經驗百分位數法" 以及 "偏差修正暨調節百分位數法" (BCa method) 來計算信賴區間. BCa 法比 "經驗百分位數法" 需要更多的計算, 但也針對廣泛的母體分布都提供更接近假設的涵蓋面 (coverage).

3. 何時環靴法的運作結果是好的?

環靴法的運作好壞, 取決於三件事情:

- 母體的特性 (外型, shape)
- 樣本大小
- 欲估計的參數

註 1. 前兩項因素 (母體特性與樣本大小) 會影響經驗累積分布函數 (empirical cdf) 是否爲母體分布函數 (population cdf) 的好估計, 繼而影響環靴法運作的好壞. 如:

(a) 相當對稱的母體分布: 中等的樣本大小即可 (如, $n \geq 20$); 若製造中央位置, 如期望值與中位數的信賴區間時, 甚至更小的樣本就夠了.

(b) 高度偏斜的母體分布: 需要更大的樣本大小.

註 2. 製造眞實期望值爲 1 的指數分布的期望值的 95% 信賴區間時, 採用 BCa 方法 (在 $B = 1000$ 個環靴法 (再取樣) 下), 當樣本大小 $n = 20$ 時, 就會有相當好的結果.

註 3. 欲估計的參數對運作好壞的影響極大: 取 $n = 20$ 時, 指數分布的標準差的信賴區間所提供的涵蓋面是非常的小.

註 4. 在小樣本下, 不應假設環靴法的運作會有好的結果.

4. 期望值的無母數信賴區間

根據前面提過的資料製造對應的無母數信賴區間並與母數信賴區間比較之:

(1) 無母數信賴區間

資料	信賴區間	型態
參考區域的 TcCB	[0.52, 0.69]	雙邊
鉻	[0, 418]	上單邊
隨機樣本	[0, 41]	上單邊

註. 隨機樣本是由期望值爲 10, CV 爲 2 的對數常態分布所取樣出的.

(2) Reference area TcCB 的參數信賴區間

信賴區間	方法
[0.52，0.70]	Land
[0.50, 0.74]	Parkin et al.
[0.52, 0.69]	Cox
[0.51, 0.68]	常態近似

註. 與 Cox 方法的結果一樣; 與其他的結果也差不多相同.

(3) 鉻的參數信賴區間 (單邊上)

信賴區間	方法
[0, 497]	Land
[0, 230]	Parkin et al.
[0, 366]	Cox
[0, 261]	常態近似

註. 上信賴極限 418 介於 Cox 方法與 Land 方法的結果間.

(4) 隨機樣本的參數信賴區間 (單邊上)

信賴區間	方法
[0, 182]	Land
[0, 96]	Parkin et al.
[0, 68]	Cox
[0, 28]	常態近似

註. 上信賴極限 41 小於常態近似法以外的三種方法的結果.

另外, 可參考估計量的環靴法分布的值方圖以及附加密度函數圖. 也可繪估計量的環靴法分布的常態 Q-Q 圖.

註. BCa 方法並未假設估計量的分布 (因此估計量的環靴法分布) 爲常態近似; 而是假設估計量的某種轉換爲近似常態分布.

5. 有關正偏斜資料的期望值的上信賴極限涵蓋面的模擬研究

探討 3 個正偏斜母體分布, 當取小樣本到中樣本的情況下 (n =10, 15, 20, 25, 30):

(a) $\Lambda(10, 1)$, 期望值 $=$ 10, CV $=$ 1: 次正偏斜

(b) $\Lambda(10, 2)$: 最正偏斜

(c) mixed-$\Lambda(5, 1; 30, 0.5; 30\%)$, 期望值 $= 0.7 \times 5 + 0.3 \times 30 = 3.5 + 9 = 12.5$: 最小正偏斜

採用 $B = 2000$ 個環靴法的 BCa 方法下, 經過 1000 回的模擬, 得期望值的 95% 單邊上信賴極限的涵蓋率 (coverage) 如下表:

分布	樣本大小	涵蓋率 (%)
$\Lambda(10, 1)$	10	84
$\Lambda(10, 1)$	15	85
$\Lambda(10, 1)$	20	87
$\Lambda(10, 1)$	25	89
$\Lambda(10, 1)$	30	90
$\Lambda(10, 2)$	10	77
$\Lambda(10, 2)$	15	80
$\Lambda(10, 2)$	20	83
$\Lambda(10, 2)$	25	82
$\Lambda(10, 2)$	30	85
混合-$\Lambda(5, 1, 30, 0.5; 30\%)$	10	90
混合-$\Lambda(5, 1, 30, 0.5; 30\%)$	15	92
混合-$\Lambda(5, 1, 30, 0.5; 30\%)$	20	92
混合-$\Lambda(5, 1, 30, 0.5; 30\%)$	25	92
混合-$\Lambda(5, 1, 30, 0.5; 30\%)$	30	93

- 針對 $\Lambda(10, 1)$ 分布, $n = 30$ 時, 涵蓋率才合宜.

- 針對 $\Lambda(10, 2)$ 分布, $n = 30$ 時, 涵蓋率還遠小於假設的 95%, 只有 85% (因爲此分布最正偏斜).

- 斜對混合-對數常態分布, 即使當 $n = 10$ 時, 涵蓋率就不錯了; $n = 15$ 時, 就相當接近假設

的 95% (因爲此分布是最小的正偏斜).

- 針對極右偏斜分布的期望值的上信賴極限, 環靴法的運作效果不是很好, 除非樣本大小要相當地大. 這是因爲極高的值還是會有一個小量的發生機率, 而對相當敏感的期望值有影響; 再加上極大的值不太會出現在小樣本中, 導致經驗累積分布函數不是母體累積分布函數的好估計, 故以環靴法方法所導出的期望值的上信賴極限的涵蓋率會遠低於假設的信賴水平.

6. 二期望值或二中位數之間的差的無母數信賴區間

   環靴法不限於製造單一母體的參數的信賴區間, 反而是一相當廣義的方法, 適用於更複雜的問題, 諸如兩期望值或兩中位數的比較, 線性迴歸與時間序列分析. 此處示範說明如何應用環靴法於製造兩期望值的差以及兩中位數的差的信賴區間, 如下述:

   (a) 清理區期望值 − 參考區期望值: 點估計 $= 3.9 - 0.6 = 3.3$ ppb; 雙邊 95% 信賴區間 $= [0.75, 13]$ (根據 BCa 環靴法)

   (b) 清理區中位數 − 參考區中位數: 點估計 $= 0.43 - 0.54 = -0.11$ ppb; 雙邊 95% 信賴區間 $= [-0.25, 0.6]$ (根據 BCa 環靴法)

註. 兩期望值的差的信賴區間有極大的寬度, 反應出眞實的差大略是介於 1 與 13 ppb 之間以及淸理區 TcCB 的資料是大的偏斜. 兩中位數的差的信賴區間提供一較小的寬度, 顯示眞實的差是相對的小.

七. 分布的量分位數的估計與信賴區間

量分位數或百分位數常用於環境標準值與法規上. 舉二例說明如下:

例 1. 爲了決定是否符合法規, 可要求估計 "背景水平" 分布的極端百分位數 (extreme percentile, 如, 第 95 個百分位數), 然後再將監測井 (compliance wells) 或修復區 (remediated areas) 的觀察值與此百分位數 (或此百分位數的上信賴極限) 比較之. USEPA 稱此種比較爲 "熱度量比較" (Hot-Measurement Comparison).

例 2. 地下水檢測時, 現場也許以某一化學物的固定合法極限 (compliance limit) 進行合法檢測 (compliance monitoring). 固定合法極限可能爲一

"最大濃度極限" (MCL, maximum concentration limit) 或 "另類濃度極限" (ACL, alternative concentration limit). 事實上, 大多數的 MCL 代表著平均水平; 但有時 MCL 或 ACL 代表一個僅有極少數時候被超越的極限 (亦即, 極大多數時候都在此極限下, 只有非常小的次數超過此極限). 此時, 需要將地下水化學濃度分布的第 99 個百分位數與 MCL 比較之. 若此估計的第 99 個百分位數 (或此百分位數的下信賴極限) 大於 MCL, 則判定不合法.

註. 即使用第 99 個百分位數似乎是一件保守或安全的事, 仍需謹記估計量是隨機變數且極端百分位數的估計更是非常, 非常的隨機. 舉例:

1. 針對常態分布 (對數常態分布亦然), 第 50 個百分位數 (亦即, 中位數) 的估計比其他百分位數的估計有較高的精密度 (precision), 且精密度迅速地下降, 當估計較大或較小的百分位數時.

2. 即使取大樣本時, 極端百分位數的估計也是非常的不精密 (imprecise).

因此, 極端百分位數的估計及其信賴區間的精密度的量化就顯得格外的重要.

註. 量分位數的信賴區間的製造是根據第六章的容忍區間 (tolerence intervals) 的概念而來.

1. 第 p 個量分位數的公式

   第 p 個量分位數的定義 (複習): 一數 (這數) 使得母體中小於或等於此數的比率爲 p, 也就是說, 隨機變數 X 的 第 p 個量分位數 x_p 爲使得累積分布函數 (cdf)

   $$P(X \leq x_p) = F(x_p) = p$$

   的那個數, 若 X 爲連續時; 或最小的數, 若 X 爲離散時. 因此, 第 p 個量分位數的一般公式爲

   $$x_p = F^{-1}(p),\ 若\ X\ 爲連續$$

   或

   $$F^{-1}(p)\ 的最小值,\ 若\ X\ 爲離散$$

2. 量分位數的參數估計

步驟如下:

(a) 估計假設分布的參數.

(b) 根據 (a) 中的分布以及公式 $x_p = F^{-1}(p)$ 求得所需的量分位數的估計.

註. 必需相當地確定 (a) 中的假設分布對資料而言是一個好的模型.

3. 常態分布的估計以及信賴區間

設

$$X \sim N(\mu, \sigma)$$

則 X 的第 p 個量分位數

$$x_p = \mu + \sigma z_p$$

其中 z_p 爲 $N(0,1)$ 的第 p 個量分位數. 此乃因爲

$$\begin{aligned} P(X \le \mu + \sigma z_p) &= P\left(\frac{X-\mu}{\sigma} \le z_p\right) \\ &= P(N(0,1) \le z_p) \\ &= p \end{aligned}$$

所以, 一個相當自然的 x_p 的估計

$$\hat{x}_p = \bar{x} + sz_p$$

(亦即, 將參數以估計值取代之), 稱其爲 x_p 的 "準最大概似估計量" (quasi-mle) (因爲若採用 σ 的最大概似估計量 s_m 時, $\bar{x} + s_m z_p$ 就爲 x_p 的最大概似估計量).

根據常態分布容忍區間的定理, x_p 的單邊上 $(1-\alpha)100\%$ 信賴區間如下:

$$\left[-\infty,\ \bar{x} + t_{n-1, z_p\sqrt{n}, 1-\alpha} \frac{s}{\sqrt{n}}\right]$$

其中 $t_{\nu,\delta,p}$ 爲自由度爲 ν, 非中央參數 (non-centrality) 爲 δ 的第 p 個量分位數.

同理, x_p 的單邊下 $(1-\alpha)100\%$ 信賴區間爲:

$$\left[\bar{x} + t_{n-1, z_p\sqrt{n}, \alpha} \frac{s}{\sqrt{n}},\ \infty\right]$$

以及 x_p 的雙邊 $(1-\alpha)100\%$ 信賴區間爲:

$$\left[\bar{x} + t_{n-1, z_p\sqrt{n}, \alpha/2} \frac{s}{\sqrt{n}},\right.$$
$$\left.\bar{x} + t_{n-1, z_p\sqrt{n}, 1-\alpha/2} \frac{s}{\sqrt{n}}\right]$$

例. 涕滅威 (Aldicarb, 農藥) 的第 95 個百分位數的下信賴極限與 ACL 的比較

三個地下水監測井中, 個別 4 個月 (每個月一次) 的涕滅威濃度採樣資料存放爲資料架構 `epa.89b.aldicarb2.df`. 假設合格標準爲:

超過 50 ppm ACL 的觀察值數不應大於 5%

所以, 判定不合格的保守作法爲:

比較涕滅威濃度的第 95 個百分位數的下信賴極限是否大於 ACL; 若是, 則眞正大於 ACL 的觀察值數會相當確定地大於 5%, 故判定爲不合格 (亦即, 相當地確定眞正的 95% 會大於 ACL).

註. USEPA 計算第 95 個百分位數的上單邊信賴極限是不正確的做法, 因爲上信賴極限 $>$ ACL, 無法確定地保證眞正的第 95 個百分位數會大於 ACL.

透過環統模組計算的結果如下:

統計量	井 1	井 2	井 3
期望值	23.1	24.7	24.5
標準差	4.9	2.3	2.1
第 95 百分位數	31.2	28.4	28.0
第 95 百分位數的 LCL	25.3	25.7	25.5

三個井的第 95 百分位數的 LCL 均小於 50 ppm ACL, 故無一井是不合格的.

註. 爲了解第 95 百分位數估計的變異量程度, 可求其 99% 的雙邊信賴區間如下:

井 1 爲 [25, 80]
井 2 爲 [25, 51]
井 3 爲 [25, 49]

有不小的變異程度; 若以 UCL 與 ACL 比較, 則井 1 與井 2 可被判定爲不合格, 但此方法適乎不合宜. 有此現象的主要問題是因爲非常小的樣本, $n = 4$.

4. 對數常態分布的量分位數的估計與信賴區間

由於量分位數在單調轉換下是不變的 (invariant), 故對數常態分布的量分位數的估計與信賴區間的公式可透過常態分布的公式得之, 如下述: 設

$$X \sim \Lambda(\theta, \tau)$$

則

$$Y = \log(X) \sim N(\mu, \sigma)$$

根據常態分布的百分位數的公式以及單調轉換下的不變的特性，X 的第 p 個量分位數

$$x_p = \exp(y_p) = \exp(\mu + \sigma z_p)$$

所以，x_p 的準-最大概似估計量 (quasi-mle)

$$\hat{x}_p = \exp(\bar{y} + s_y z_p)$$

其中

$$y_i = \log(x_i), i = 1, 2, \ldots, n$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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

另外，x_p 的信賴區間分別如下.

單邊上 $(1-\alpha)100\%$ 信賴區間:

$$\left[-\infty,\ \exp\left(\bar{y} + t_{n-1, z_p\sqrt{n}, 1-\alpha}\frac{s_y}{\sqrt{n}}\right)\right]$$

單邊下 $(1-\alpha)100\%$ 信賴區間:

$$\left[\exp\left(\bar{y} + t_{n-1, z_p\sqrt{n}, \alpha}\frac{s_y}{\sqrt{n}}\right),\ \infty\right]$$

雙邊 $(1-\alpha)100\%$ 信賴區間:

$$\left[\exp\left(\bar{y}+t_{n-1,z_p\sqrt{n},\alpha/2}\frac{s_y}{\sqrt{n}}\right),\right.$$
$$\left.\exp\left(\bar{y}+t_{n-1,z_p\sqrt{n},1-\alpha/2}\frac{s_y}{\sqrt{n}}\right)\right]$$

其中 $\bar{y}$ 與 s_y 如上述.

例. Chrysene 的第 95 百分位數的下信賴極限與 ACL 的比較

5 個地下水監測井中, 個別 4 個月 (每月一次) 的 chrysene 濃度採樣資料存放爲資料架構 `epa.92c.chrysene.df`. 假設資料來自於對數常態分布 (亦即, 以對數常態分布模型化). 假設合格標準爲:

超過 80 ppb ACL 的觀察值數不應大於 5%

所以, 判定不合格的作法爲:

比較 chrysene 濃度的第 95 百分位數的下信賴極限是否大於 ACL; 若是, 則眞正大於 ACL 的觀察值數會相當確定地大於 5%, 故判定爲不合格.

註. USEPA 採用計算第 95 個百分位數的上信賴極限的不妥作法.

透過環統模組計算的結果如下表:

統計量	井 1	井 2	井 3
期望值	19.9	9.8	46.3
CV	0.7	0.5	0.4
第 95 百分位數	51.4	19.0	78.5
第 95 百分位數的 LCL	22.6	11.1	51.6

統計量	井 4	井 5
期望值	24.7	29.0
CV	0.2	0.5
第 95 百分位數	36.5	58.5
第 95 百分位數的 LCL	26.9	32.9

5 個井的第 95 個百分位數的 LCL 均小於 80 ppb ACL, 故無一井是不合格的.

註. 計算 99% 的雙邊信賴區間以便對第 95 百分位數估計的變異程度有所了解:

井 1 爲 [25, 178]
井 2 爲 [11, 63]
井 3 爲 [52, 235]
井 4 爲 [27, 95]

井 5 爲 [34, 185]

有很大的變異程度; 若以 UCL 與 ACL 比較, 則井 1, 井 3, 井 4, 井 5 均可被判爲不合格, 似乎不適宜. 造成此現象的主要原因爲極小的樣本數, $n = 4$.

5. ト松分布的量分位數的估計與信賴區間

先估計分布的參數, 再透過此分布的累積分布函數與公式

$$x_p = F^{-1}(p)$$

計算 x_p 的估計, 如下述: 設

$$X \sim \text{Poisson}(\lambda)$$

令

$$x_{p|\lambda}$$

爲 X 的第 p 個量分位數. 則第 p 個量分位數的估計

$$\widehat{x}_{p|\lambda} = x_{p|\lambda=\widehat{\lambda}}$$

其中

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

(爲一 mle). 因此, $\hat{x}_{p|\lambda}$ 爲 $x_{p|\lambda}$ 的 mle. 又根據卜松分布的容忍區間的定理, $x_{p|\lambda}$ 的信賴區間分別如下.

單邊上 $(1-\alpha)100\%$ 信賴區間:

$$[0,\ x_{p|\lambda=\text{UCL}}]$$

其中 UCL 爲最大的 λ 使得

$$P(X \leq x|\lambda = \text{UCL}) = \alpha$$

單邊下 $(1-\alpha)100\%$ 信賴區間:

$$[x_{p|\lambda=\text{LCL}},\ \infty]$$

其中 LCL 爲最小的 λ 使得

$$P(X \geq x|\lambda = \text{LCL}) = \alpha$$

雙邊 $(1-\alpha)100\%$ 信賴區間:

$$[x_{p|\lambda=\text{LCL}},\ x_{p|\lambda=\text{UCL}}]$$

其中 LCL 爲最小的 λ 使得

$$P(X \geq x|\lambda = \text{LCL}) = \alpha/2$$

且 UCL 爲最大的 λ 使得

$$P(X \leq x|\lambda = \text{UCL}) = \alpha/2$$

例. 苯濃度的第 90 百分位數的上信賴極限 (UCL)

將無法偵測觀察值 "$<$ 2" 設定爲 1, 並以卜松分布模型化此組資料成: 大量水分子中, 苯的分子數量. USEPA 建議計算 6 個背景井中苯濃度的第 95 百分位數的上信賴極限, 並以此 UCL 作爲門檻值 (亦即, 下游監測井中, 苯濃度超過此上極限, 則暗示地下水中可能有污染). 所以這是一種 "熱-度量比較" (Hot-Measurement Comparison). 透過環統模組的計算, 得

$$\widehat{\lambda} = 1.94,\ \widehat{x}_{.95|\lambda} = 4,\ x_{.95|\lambda = \mathrm{UCL}} = 5$$

6. 量分位數的無母數估計

概念如下:

(a) 以經驗累積分布 (empirical cdf) 估計母體的累積分布 (cdf).

(b) 以線性內差法 (linear intergolation) 估計量分位數.

S+ 中, 以無母數估計量分位數的明確方式如下:

(a) 以描點位置 (plotting position)

$$\hat{p}_i = \frac{i-1}{n-1}$$

製造經驗累積分布.

(b) 以線性內差法計算 x_p 的估計

$$\hat{x}_p = x_{(i+1)} - \left[x_{(i+1)} - x_{(i)}\right]\left(\frac{\hat{p}_{i+1} - p}{\hat{p}_{i+1} - \hat{p}_i}\right)$$

圖示如下:

註 1. 選用此種描點位置時, 以最小值 $x_{(1)}$ 估計第 0 個百分位數 (亦即, $\hat{x}_0 = x_{(1)}$) 且以最大值 $x_{(n)}$ 估計第 100 百分位數 (亦即, $\hat{x}_1 = x_{(n)}$).

Why? 因爲

$$\hat{p}_1 = \frac{1-1}{n-1} = 0 \text{ 如圖:}$$

所以,

$$\begin{aligned}
\hat{x}_0 &= x_{(2)} - \left[x_{(2)} - x_{(1)}\right]\left(\frac{\hat{p}_2 - 0}{\hat{p}_2 - \hat{p}_1}\right) \\
&= x_{(2)} - \left[x_{(2)} - x_{(1)}\right]\left(\frac{\hat{p}_2 - 0}{\hat{p}_2 - 0}\right) \\
&= x_{(2)} - \left[x_{(2)} - x_{(1)}\right] = x_{(1)}
\end{aligned}$$

又因爲

$$\widehat{p}_n = \frac{n-1}{n-1} = 1 \text{ 如圖:}$$

所以,

$$\begin{aligned}
\widehat{x}_1 &= x_{(n)} - \left[x_{(n)} - x_{(n-1)}\right]\left(\frac{\widehat{p}_n - 1}{\widehat{p}_n - \widehat{p}_{n-1}}\right) \\
&= x_{(n)} - \left[x_{(n)} - x_{(n-1)}\right]\left(\frac{1-1}{\widehat{p}_n - \widehat{p}_{n-1}}\right) \\
&= x_{(n)} - 0 = x_{(n)}
\end{aligned}$$

註 2. 對立於以參數方式估計量分位數, 使用無母數方式估計極端的量分位數時, 需要大量的觀察值以便得到好的精密度.

註 3. 以無母數方式估計量分位數的一大優勢爲: 容易處理缺失 (censored) 值, 因爲只需將他們排序即可.

7. 量分位數的無母數信賴區間

令

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} X$$

且

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$$

為對應的 n 個有序統計量. 若 x_p 為 X 的第 p 個量分位數, 則對於 $1 \le i \le n$ (參考圖示),

$$\begin{aligned}
P\left[X_{(i)} > x_p\right] &= P\left[X_1, X_2, \ldots, X_n \text{ 中至多有 } (i-1) \text{ 個是 } \le x_p\right] \\
&= \sum_{k=0}^{i-1} P\left[X_1, X_2, \ldots, X_n \text{ 中剛好有 } k \text{ 個是 } \le x_p\right] \\
&= \sum_{k=0}^{i-1} \binom{n}{k} \left[P(X \le x_p)\right]^k \left[P(X > x_p)\right]^{n-k} \quad \text{(因為獨立同分布)} \\
&= \sum_{k=0}^{i-1} \binom{n}{k} p^k (1-p)^{n-k} \\
&= F_{n,p}(i-1) \qquad (11)
\end{aligned}$$

其中 $F_{n,p}(y) = \text{Binomial}(n,p)$ 的 cdf 在 y 的值.

接著以上式製造量分位數的精確的無母數信賴區間如下:

- 第 p 個量分位數的雙邊無母數的信賴區間:

$$[x_{(r)},\ x_{(s)}]$$

$1 \le r \le n-1,\ 2 \le s \le n$ 且 $r < s$, 其信賴水平至少爲

$$F_{n,p}(s-1) - F_{n,p}(r-1)$$

Why? 因爲

$$\begin{aligned} P\left[X_{(r)} \le x_p \le X_{(s)}\right] &= P\left[x_p \le X_{(s)}\right] - P\left[x_p < X_{(r)}\right] \\ &\ge P\left[x_p < X_{(s)}\right] - P\left[x_p < X_{(r)}\right] \\ &= F_{n,p}(s-1) - F_{n,p}(r-1) \end{aligned}$$

(最後一個等號成立乃根據 (11) 式)

- 第 p 個量分位數的單邊無母數下信賴區間:

$$[x_{(r)},\ \infty]$$

且信賴水平爲

$$1 - F_{n,p}(r-1)$$

Why? 因爲

$$\begin{aligned} P\left[X_{(r)} \le x_p\right] &= 1 - P\left[x_p < X_{(r)}\right] \\ &= 1 - F_{n,p}(r-1) \end{aligned}$$

(由 (11) 式)

- 第 p 個量分位數的單邊無母數上信賴區間:

$$[-\infty,\ x_{(s)}]$$

且信賴水平至少爲

$$F_{n,p}(s-1)$$

Why? 因爲

$$\begin{aligned} P\left[x_p \le X_{(s)}\right] &\ge P\left[X_{(s)} > x_p\right] \\ &= F_{n,p}(s-1) \end{aligned}$$

(由 (11) 式)

註 因爲百分位數的無母數信賴區間是根據資料的秩值 (ranks) 而得, 故信賴水平很自然地會是離散的, 也即, 針對一特別的百分位數, 只能得到某些特定的信賴水平且這些信賴水平會受限於樣本大小. 如, 計算第 95 百分位數的上信賴區間時, 選取最大值爲上信賴極限 (UCL) (亦即, $s=n$), 得不同的樣本大小 (n) 所對應的信賴水平如下表:

樣本大小 (n)	信賴水平 (%)
5	23
10	40
15	54
20	64
25	72
50	92
75	98
100	99

- 信賴水平 $= F_{n,0.95}(n-1) \times 100\%$.
- 當樣本大小 $n > 50$ 才能得到超過 95% 的信賴水平.

例. 銅濃度的第 95 百分位數的無母數上信賴極限

對 5 個地下水井做銅濃度的監視: 3 個背景井, 2 個監測井 (compliance well), 每月由各井採樣, 共 8 個月, 但遺失 2 個監測井前 4 個月的樣本. 共得 32 個觀察值存放於資料架構 `epa.92c.copper2.df` 內 (24 個來自背景井, 8 個來自監測井); 其中背景井的 24 個觀察值中, 有 15 個標示爲 "<5" 的無法偵測值. USEPA 採用的 "熱-度量比較" (Hot-Measurement

Comparson) 如下: 以背景井求出的銅濃度極端 (如, 第 95) 百分位數的無母數 UCL 作爲判定監測井的門檻值 (亦即, 若監測井的濃度觀察值大於此門檻值, 則暗示有污染並判斷爲不合格). 計算與討論結果爲:

- 背景中有 $15/24 = 62.5\%$ 的無法偵測值 (" < 5" ppb) (超過一半), 故中位數的估計值 < 5.
- 第 95 百分位數的估計值爲 7.925.
- 以背景井的最大值 9.2 ppb 作爲第 95 百分位數的上信賴極限, 所得到的信賴區間僅有 71% 的信賴水平.
- 若以 9.2 ppb 作爲門檻值, 則暗示井 4 可能有污染而井 5 是合格的.
- 以環統模組計算 95% 的第 95 百分位數的 UCL, 並以此爲門檻值, 則結果爲何?

8. 根據環靴法求量分位數的無母數信賴區間

以環靴法求得的百分位數的無母數信賴區間, 本質上無異於根據秩值方法求得的結果.

八. 一個有關信賴區間警誡性的註解

- 信賴區間是一種統計推論 (statistical inference), 根據由母體得到的樣本針對母體特性做推論.

- 所有討論過的用以估計母數及製造對應信賴區間的方法均假設有一取自母體的具代表性樣本以及觀察值間的相互獨立性.

- 常發生的一種情況是, 所採集的樣本乃出自於一個比欲推論的母體還要小的母體, 並不適宜作爲眞正欲討論的母體的代表性樣本, 如, 欲推論某一特定背景地下水井中銅濃度的第 95 百分位數, 而此濃度分布極有可能受許多時間因子的影響, 以致於在月與年之間有某種特定型式的變異性, 故僅由一年中的某時段取樣, 則無法成爲具有月與年變異性母體的代表性樣本.

- 將所有造成環境資料變異性的因素均考慮進去並處理的抽樣設計是少有並不可行的. 故粗略而言, 大多數的信賴區間均太窄.