單元 4: 機率分布

母體定義爲回答欲探索問題所需依據的整體 (全部) 度量值的集合. 機率分布爲模型化由母體承襲的變異性時, 所採用的理想數學模型 (或爲一理想數學模型, 用以模型化由母體承襲的變異性). 本章的目標爲

1. 討論隨機變數與機率分布的觀念

2. 討論環境統計中重要的機率分布

3. 展示統計軟體 S+ 及其環統模組中, 繪製機率分布圖形與計算機率分布的相關量的功能

一. 何謂隨機變數?

執行環境研究時, 由母體中取樣並且記錄此樣本的度量值, 這些實際的數值 (度量值) 爲一種觀念, 稱作隨機變數, 的實現 (值) (realizations). 或者說, 一隨機變數爲一實驗的下一個觀察值; 是由母體中取出的下一個實體樣本 (physical sample) 的度量值. 隨機變數的結

果通常隨著觀察 (或採樣) 而異, 起因於經常會有多個變異源對最後值的影響.

例. 擲銅板並觀察出現 "正面" 或 "反面". 母體爲擲銅板無窮多次的結果的集合 = {"正面"、"反面"}. 隨機變數爲擲下一個銅板後的出現結果.

二. 離散隨機變數對連續隨機變數

一隨機變數僅可能取有限個或可數無限個值時, 稱其爲離散 (discrete) 隨機變數

若一隨機變數可取無限 (不可數) 個值時, 則稱其爲連續 (continuous) 隨機變數, 如土壤樣本中四氯 (代) 苯的濃度理論上可爲無限多個可能值.

註. 實際上, 連續隨機變數的值經常受限於儀器的精準度以及數學的進位或刪除.

例. 實驗: 由 2 個不同的區域採集土壤樣本, 則母體爲所有可能的土壤樣本的度量值的集合; 隨機變數爲將觀察到的一土壤樣本中四氯代苯的濃度.

三. 何謂機率分布?

將實驗觀察值繪成相對頻率直方圖 (relative frequency histogram) (或密度直方圖, density histogram) (亦即, 落在每一個等級 class (bar) 內的觀察值在全部觀察值中所占的比率等於此等級的面積), 可顯示各種觀察值出現 (分布) 的相對頻率.

1. 針對一僅取整數值的離散隨機變數, 在等級的長度取爲 1 時, 一機率分布可視爲取大量樣本 (或無限多個樣本) 下的密度直方圖所呈現的結果.

2. 針對一連續隨機變數, 一機率分布可視爲取樣數愈來愈多, 且直方圖等級愈來愈窄下的密度直方圖所呈現的結果.

四. 機率密度函數 (PDF)

機率密度函數 (pdf) 是描述一隨機變數的相對頻率的數學公式; 有時此公式的圖形亦稱作機率密度函數.

離散隨機變數的機率密度函數又稱作機率密度質量 (probability mass function) 簡稱 (pmf), 因爲它顯示出隨機變數在每一個可能值的機率 "質量". 如, 擲公正銅板的機率密度質量爲

$$f(x) = Pr(X = x) = \begin{cases} 0.5 & \text{若 } x = 0 \\ 0.5 & \text{若 } x = 1 \end{cases}$$

lognormal 分布的機率密度函數爲

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}[\log(x) - \mu]^2\right\}$$

$x > 0$, 其中

$$\mu = \log\left(\frac{\theta}{\sqrt{\tau^2 + 1}}\right), \sigma = [\log(\tau^2 + 1)]^{1/2}$$

$$\theta = 0.6, \tau = 0.5 \text{ (某二個特定值)}$$

對於連續隨機變數, 隨機變數落在某區間內的機率等於機率密度函數在此區間上所圍出區域的面積, 如數學的表達方式:

$$Pr(0.75 \leq X \leq 1) = \int_{0.75}^{1} f(x)dx$$

1. 繪機率密度函數的圖形

S+ 與環統模組中共有 38 種可用的機率分布, 它們可作爲母體的模型. 大部份的機率分布都可由某理論數學模型獲得, 如, 二項分布針對二種結果, Poisson 分布針對 "稀有" 事件, Weibull 分布針對極端值, 常態分布針對多個隨機變數的和...等等. 每種分布都有對應的選單 (menu) 與函式 (function) (指令) 繪出其機率密度函數.

例. 繪二項分布與 lognormal 機率密度函數.

2. 計算機率密度函數的值

可使用 S+ 環統模組的選單與指令計算任一內建機率分布的機率密度函數的值.

例. 求二項分布的機率密度函數在 x 等於 0 與 1 的值以及 lognormal 的機率密度函數在 x 爲 0.5, 0.75 與 1 的值.

五. 累積分布函數 (CDF)

隨機變數 X 的累積分布函數

$$F(x) \stackrel{\text{def}}{=} Pr(X \le x) = \begin{cases} \int_{-\infty}^{x} f(t)dt & \text{若 } X \text{ 是連續} \\ \sum_{x_i \le x} f(x_i) & \text{若 } X \text{ 是離散} \end{cases}$$

其中 $-\infty < x < \infty$.

可用累積分布函數來求隨機變數落在某一特定區間內的機率, 如

$$\begin{aligned} Pr(0.75 \le X \le 1) &= \int_{0.75}^{1} f(x)dx \\ &= Pr(X \le 1) - Pr(X \le 0.75) + \\ &\quad Pr(X = 0.75) \\ &= F(1) - F(0.75) + Pr(X = 0.75) \\ &(= F(1) - F(0.75), \text{ 若 } X \text{ 是連續}) \end{aligned}$$

(因爲在連續時, $P(X = 0.75) =$ 在 0.75 與 0.75 間的累積分布函數所圍出區域的面積 $= 0$)

1. 繪累積分布函數的圖形

   離散隨機變數的累積分布函數的圖形爲一階梯函數 (step function), 在每一可能值 x_i 處有一跳躍 (jump) $f(x_i)$.

連續隨機變數的累積分布函數的圖形爲一平滑由 0 開始遞增至 1 的曲線.

例. 繪二項分布與 lognormal 的累積分布函數的圖形.

2. 求累積分布函數的值

可使用 S+ 與環統模組的選單與指令計算任一內建機率分布的累積分布函數的值.

例. 求二項分布的累積分布函數在 $x = 0, 0.5$ 與 1 的值以及 lognormal 累積分布函數在 $x = 0.5, 0.75$ 與 1 的值.

六. 量分位數與百分位數

簡言之, 一母體的第 p 個量分位數 (p^{th} quantile) 是這 (一) 數使得母體中小於或等於此數的比例爲 p; 第 p 個量分位數又稱作第 $100p$ 百分位數 ($100p^{\text{th}}$ percentile). 嚴格之, 令 X 爲某一特定分布的隨機變

數, X 的分布的第 p 個量分位數, 記作 x_p, 定義爲這(一)數滿足

$$Pr(X < x_P) \leq p \leq Pr(X \leq x_p) \qquad (1)$$

其中 $0 \leq p \leq 1$. 圖例說明如下:

1. 在單調轉換下, 量分位數是不變的 (Invariant under Monotonic Transformations)

   令 X 爲一隨機變數且 $Y = g(X)$, 其中 g 爲一單調函數, 則 X 的第 p 個量分位數 x_p 與 Y 的第 p 個量分位數之間的關係如下:

   $$y_p = g(x_p) \text{ 且 } x_p = g^{-1}(y_p)$$

   如, 若 X 的分布爲 lognormal, 則 $Y = \log(X)$ 的分布爲 normal. 因爲 log 爲一遞增函 (單調) 函數且其反函數爲 exp, 所以 $x_p = \exp(y_p)$. 因此, 若 Y 的 median (第 0.5 個量分位數) 爲 10, 則 X 的 median 就是 exp(10).

2. 求量分位數與百分位數

(a) 由累積密度函數的圖形可輕易地以視覺法求出一些重要的量分位數, 如第 50 百分位數與第 95 百分位數.

如, 由二項分布的累積密度函數圖形知,
x_0 = 任一 ≤ 0 之數
$x_{0.5}$ = 任一 ≥ 0 且 ≤ 1 之數
x_1 = 任一 ≥ 1 之數
(習慣上, 令 $x_{0.5} = \frac{1}{2}(0+1) = 0.5$)

由 lognormal 分布的累積密度函數, 可輕易地得知

$$x_{0.5} \approx 0.5 \text{ 且 } x_{0.95} \approx 1.1$$

(b) 使用 S+ 與環統模組的選單或指令求需要的量分位數.

例. 求二項分布的第 0, 25, 50, 75 與 100 百分位數 (與你想的值一樣嗎? S+ 的內定公式爲何?); 以及 lognormal 分布的第 50 與 95 百分位數.

七. 由機率分布生成隨機數

使用 S+ 與環統模組的選單或指令可生成所有內建機率分布的隨機數 (又稱仿隨機數, pseudo-random numbers). 當設定隨機數 (random number) 的種子 (seed) 爲某一特定值時, 一定會得到此特定種子所對應出一串 "隨機數" (此項似乎不隨機的特性方便於比較與驗証之用).

例. 生成 10 個與 100 個由二項分布形成的隨機數以及 100 個由 lognormal 分布形成的隨機數.

八. 機率分布的特質 (性)

每一個機率分布都有某些特質. 統計學家已定義了各種量以描述一機率分布的特性, 如, 平均值 (mean), 中位數 (median), mode, 變異數 (variance), 標準差 (standard deviation), 變異係數 (coefficient of variation), 偏斜性 (skew), 聳立性 (kurtosis). 在第 3 單元已討論過以上 8 種經由隨機樣本求得的樣本統計量 (sample statistics). 此處討論的乃是與母體相關連的 8 種量, 分述如下:

1. 平均值, 中位數與 Mode

均爲機率分布的中央傾向 (central tendency) 的度量.

中位數 = 第 50 百分位數, 不一定是唯一的.

mode = 產生機率密度函數最大值的量分位數, 不一定是唯一的 (因爲可能有多個量分位數會產生相同的最大值, 如二項分布的 mode 爲 0 與 1).

平均值 (又稱作期望值, expected value) 定義如下:

$$\begin{aligned} \mu &= E(X) \\ &= \begin{cases} \int_{-\infty}^{\infty} x f(x) dx & \text{若 } X \text{ 爲連續} \\ \sum_x x f(x) & \text{若 } X \text{ 爲離散} \end{cases} \end{aligned}$$

2. 變異數與標準差

度量分布在平均值附近的擴散情形.

$$\begin{aligned} \sigma^2 &= Var(X) \\ &= E[(X-\mu)^2] \\ &= \begin{cases} \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx & \text{若 } X \text{ 爲連續} \\ \sum_x (x-\mu)^2 f(x) & \text{若 } X \text{ 爲離散} \end{cases} \end{aligned}$$

(單位: (隨機變數的單位)^2)

$$\sigma = \text{sd}(X) = \sqrt{Var(X)}$$

(單位: 與隨機變數有相同的單位, 故 σ 爲常被採用的擴散性度量)

3. 變異係數

簡記爲 CV, 爲一無單位的度量, 描述分布相對於 "mean 的大小" 的擴散性, 定義如下:

$$\text{CV} \stackrel{\text{def}}{=} \sigma/\mu$$

4. 偏斜性與聳立性

過去, 一分布的偏斜性與聳立性常被用來作爲此分布與常態分布的比較.

(a) 偏斜性

定義一隨機變數 X 的第 r 個中央動差 (r^{th} central moment)

$$\mu_r = E[(X-\mu)^r]$$

註. 1^{st} central moment $\mu_1 = 0$, 2^{nd} central moment $\mu_2 = \sigma^2$.

一分布的偏離性係數 (coefficient of sekewness)

$$\text{Skew} \stackrel{\text{def}}{=} \mu_3/\sigma^3$$

註. 偏斜性度量隨機變數的值相對於平均值的分布情形. 若小的值成串 (群) 地靠近平均值且大的值遠散於平均值之上, 則偏斜性將爲正值, 且稱分布爲正偏斜 (positively skewed) 或 right skewed (右偏斜).

若小的值遠散於平均值之下且大的值傾向成串地靠近平均值, 則偏斜性將爲負, 且稱分布爲負偏斜 (negatively skewed) 或左偏斜 (left skewed).

(b) 聳立性

一分布的聳立性係數 (coefficient of kurtosis)

$$\text{kurtosis} \stackrel{\text{def}}{=} \mu_4/\sigma^4$$

註. 一些相關的說明:

- 聳立性度量一分布的尾部中的值所占的比例.
- 聳立性與分布的平均值和變異數是相互獨立的 (亦即, 同一種類分布的聳立性是相同的, 不會因平均值和變異數而改變)。
- 常態分布的偏斜性係數爲 3; 聳立性係數 $<$ 3 的分布稱作 platykurtic, 相較於常態分布, 有較短的尾部; 聳立性係數 $>$ 3 的分布稱作 leptokurtic, 相較於常態分布, 有較重 (大) 的尾部.
- 有些場合, 聳立性意謂著 ''超越聳立性係數'' (coefficient of excess kurtosis), 亦即, ''coefficient of kurtosis - 3''.

九. 環境統計中重要的分布

1. 常態分布 (Normal Distribution)

- 機率密度函數 pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

其中 $-\infty < x < \infty$, μ: 平均值, σ: 標準差. 常以 $N(\mu, \sigma)$ 表示之.

- 當 $\mu = 0$, $\sigma = 1$ 時, 稱作標準常態分布, 並以 $N(0, 1)$ 表示之.

- 不管 μ 與 σ 的值爲何, 機率密度函數的圖形始終爲一鐘型曲線 (bell-shaped curve).

- 獨立常態隨機變數的和或平均值依然爲常態分布: 令 $X_i \sim N(\mu_i, \sigma_i)$, $i = 1, \ldots, n$ 且相互獨立, 則和

$$Y \stackrel{\text{def}}{=} \sum_{i=1}^{n} X_i \text{ 爲常態分布}$$

且

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} E(X_i) \\ &= \sum_{i=1}^{n} \mu_i \end{aligned}$$

以及

$$\begin{aligned} Var(Y) &= Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} Var(X_i) \\ &\quad \text{(因爲 } X_i \text{ 相互獨立)} \\ &= \sum_{i=1}^{n} \sigma_i^2 \end{aligned}$$

又平均值

$$\bar{X} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i \text{ 爲常態分布}$$

且

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu_i \end{aligned}$$

以及

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 \end{aligned}$$

註 1. 設母體 $\sim N(\mu, \sigma)$, 並由其中取出 n 個樣本 (以 $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma)$ 表示之), 則

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

稱作樣本平均值 (sample mean), 爲一常態分布其 $E(\bar{X}) = \mu$, $Var(\bar{X}) = \sigma^2/n$ 且 $\text{sd}(\bar{X}) = \sigma/\sqrt{n}$.

註 2. 上述有關平均值, 變異數與標準差的公式不限於常態分布; 任何分布均成立, 只要隨機變數間是相互獨立即可.

- 中央極限定理: 常態分布經常爲資料提供一個好的模型, 因爲一個可証明的事實是, 無論隨機變數的根本分布或基本分布 (underlying distribution) 爲何, 多個獨立隨機變數的和或平均值都會近似於一常態分布. 這個事實稱作中央極限定理 (Central Limit Theorem, CLT).

  註. 近似的好壞與 "樣本大小" 和 "基本分布與常態分布的差異" 有關 (亦即, 樣本愈大, 差異愈小, 則近似的愈好).

- 機率與偏離平均值的程度 (Probabilities and Deviation from the Mean)

  以標準差 (standard deviation) 度量偏離平均值的程度.

  設 $X \sim N(\mu, \sigma)$, 則

  $$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx .68$$

  (亦即, 有 68% 的機會, 隨機變數的值落在 μ 的一個標準差內)

  $$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx .95$$

(亦即, 落在 μ 的 2 個標準差的範圍內的機會大約爲 95%)

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 100\%$$

(亦即, 差不多所有的值都會落在 μ 的 3 個標準差內)

這是一個很重要的性質, 因爲即使會有取相當正與負的值的機會 (因爲機率密度函數恆爲正), 但幾乎不會有值會落在 3 個標準差之外; 故有時可使用常態分布模型化只僅取正值的環境資料.

- 轉換回標準常態分布: Z-轉換

  若 $X \sim N(\mu, \sigma)$, 則

  $$Z \overset{\text{def}}{=} \frac{X - \mu}{\sigma} \sim N(0, 1)$$

  (稱此種轉換爲 Z-轉換) 亦即,

  $$X \overset{Z\text{-轉換}}{\Rightarrow} N(0, 1)$$

  應用:

  (1) 將樣本平均值轉換成標準常態分布, 亦即, 若

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

則

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

且

$$Z \overset{\text{def}}{=} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

註. 即使根本 (基本, underlying) 分布不是常態分布, 由中央極限定理可得

$$\bar{X} \approx N(\mu, \sigma/\sqrt{n})$$

故,

$$Z \overset{\text{def}}{=} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

(還是一個相當有用的結果)

(2) 常態隨機 X 的值 x, 可透過它的 z 轉換的值, 顯示它離開 μ 有多遠 (以標準差爲單位). 如,
$z = 0.5$ 表示比 μ 多 0.5 個標準差;
$z = -0.7$ 表示比 μ 少 0.7 個標準差;
z 落在 $[-2, 2]$ 之外的機會僅約 5%;
非常不可能 z 會落在 $[-3, 3]$ 之外.

2. 對數常態分布 (Lognormal Distribution)

一隨機變數取對數轉換後的分布若爲常態分布時, 則稱此隨機變數的分布爲雙參數對數常態分布分布, 亦即, 若

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

則稱 X 有對數常態分布, 並以

$$\Lambda(\mu, \sigma)$$

表示之, 其中 μ, σ 分別表示轉換後隨機變數的平均值與標準差. 如,

$$X \sim \Lambda(0, 1) \Leftrightarrow X = \exp(Z)$$

其中 $Z \sim N(0, 1)$

- 對數常態分布的機率密度函數

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{\log(x) - \mu}{\sigma}\right]^2\right\},$$

$x > 0$.

(Why? 因爲 $Y = \log(X) \sim N(\mu, \sigma)$, 所以由變數變換, X 的機率密度函數

$$f(x) = g(y) \left.\frac{dy}{dx}\right|_{y=\log x}$$

其中

$$
\begin{aligned}
g(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{\log(x)-\mu}{\sigma}\right]^2\right\}\frac{1}{x} \\
&= \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{\log(x)-\mu}{\sigma}\right]^2\right\}
\end{aligned}
$$

此處 $-\infty < y < \infty$, $x > 0$.

- 兩個主要特性: 下界爲 0 且右偏斜; 這些特性也是環境資料中非常普通的現象.

- 對數常態分布會很自然發生的理由: CLT 說明多個隨機變數的和 $\approx$ 常態分布, 亦即, 當 n 夠大時,

$$X_1 + \cdots + X_n \approx \text{ 常態分布}$$

此結果可引申出

$$\text{多個隨機變數的乘積} \approx \text{對數常態分布}$$

Why? 因爲根據 CLT,

$$
\begin{aligned}
&\log(X_1 \cdots X_n) \\
&\quad = \log(X_1) + \cdots + \log(X_n) \\
&\quad \approx \text{ 常態分布}
\end{aligned}
$$

故, 由定義

$$X_1 \cdot X_2 \cdots X_n \approx \text{ 對數常態分布}$$

而相加與相乘都是很自然就有的運算, 所以常態分布與對數常態分布會是很自然的現象.

- 僅因爲資料的值是以 0 爲下界且右偏斜, 不足以推論說對數常態分布是一個好的模型, 這是不幸的地方, 因爲還有許多其它的分布也是以 0 爲下界且右偏斜的 (如, gamma, generalized extreme value, Weibull, mixture of lognormal...等). 這些分布在中位數附近都很像, 卻在極端的尾部 (extreme tail) 有明顯的不同 (如, 第 90 與第 95 百分位數), 然而要辨識出一個分布的尾部特性常需要大量的觀察值, 所以模型化偏斜的環境資料是相當困難的.

- 另類參數化 (Alternative Parameterization): 平均值與變異係數

  除了以轉換後隨機變數的平均值與標準差特質化 (刻劃) 對數常態分布外, 有時也可用原隨機變數分布 (亦即, 對數常態分布) 的平均值與變

異係數 (coefficient of variation) 來特質化, 亦即, 令

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

則

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

其中 μ 與 σ 分別爲轉換後隨機變數 Y 的平均值與標準差. 令 $\theta = X$ 的平均值, 且 $\tau = X$ 的變異係數, 則 μ, σ, θ 與 τ 之間的關係如下:

1. $\theta = E(X) = \exp\left(\mu + \sigma^2/2\right)$

   Why? 直接求一階動差 (1^{st} moment)

$$E(X) = \int_0^\infty x \frac{1}{x\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2}\left[\frac{\log(x) - \mu}{\sigma}\right]^2\right\} dx$$

$$= \int_{-\infty}^\infty \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left[(y-\mu)^2 - 2\sigma^2 y\right]\right\} dy$$

   第二個等號成立是因爲變數變換

$$y = \log x,\ dy = \frac{1}{x} dx$$

然後，經由完全平方，上式等於

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left[y^2- 2(\mu+\sigma^2)y+(\mu+\sigma^2)^2+ \mu^2-(\mu+\sigma^2)^2\right]\right\} dy$$

繼續化簡，又等於

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left[y- (\mu+\sigma^2)\right]^2+\frac{1}{2\sigma^2}\left[\mu^2+ 2\mu\sigma^2+\sigma^4-\mu^2\right]\right\} dy \\
= &\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left[y- (\mu+\sigma^2)\right]^2\right\} \exp\left(\mu+\frac{\sigma^2}{2}\right) dy \\
= &\exp\left(\mu+\frac{\sigma^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left[y-(\mu+\sigma^2)\right]^2\right\} dy \\
= &\exp\left(\mu+\frac{\sigma^2}{2}\right)
\end{aligned}$$

(最後一個等號成立是因爲倒數第二個式子的被積函數爲 $N(\mu+\sigma^2)$ 的機率密度函數)

2. $\tau = \mathrm{CV}(X) = \sqrt{\exp(\sigma^2) - 1}$ (求 2^{nd} moment $E(X^2)$, 再根據變異係數的定義自行驗証.)

3. $\mu = E(Y) = \log\left(\frac{\theta}{\sqrt{\tau^2+1}}\right)$

4. $\sigma = \sqrt{Var(Y)} = \sqrt{\log(\tau^2 + 1)}$

   Why? 對

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

   的兩邊取 log 得

$$\log\theta = \mu + \frac{\sigma^2}{2} \qquad (2)$$

   又對

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

   的兩邊平方並化簡, 得

$$\tau^2 + 1 = \exp(\sigma^2)$$

   再對上式兩邊取 log 得

$$\sigma^2 = \log(\tau^2 + 1) \qquad (3)$$

代 (3) 入 (2), 得

$$\begin{aligned}\mu &= \log\theta - \frac{1}{2}\log(\tau^2+1)\\ &= \log\left(\frac{\theta}{\sqrt{\tau^2+1}}\right)\end{aligned}$$

將 (3) 開根號, 得

$$\sigma = \sqrt{\log(\tau^2+1)}$$

註 1. 雖然 $X = \exp(Y)$ 且 $E(Y) = \mu$ 但 $\theta = E(X) \neq \exp(\mu)$ (亦即, 透過單調轉換的平均值不是不變的 (invariant)) 事實上, 因爲百分位數是不變的且中位數等於第 50 百分位數, 故

$$\begin{aligned}X \text{ 的中位數} &= \exp(Y \text{ 的中位數})\\ &= \exp(E(Y))\\ &= \exp(\mu)\\ &< \exp\left(\mu + \frac{\sigma^2}{2}\right) = E(X)\end{aligned}$$

亦即, 對數常態的中位數 $<$ 平均值 (可理解, 因爲 X 是右偏斜的).

第二個等號成立是因爲 Y 的機率密度函數 $>$ 0 且對稱.

註 2. 有關 X 的變異係數:

$$\begin{aligned} \tau &= CV(X) \\ &= \frac{\sqrt{Var(X)}}{E(X)} \\ &= \sqrt{\exp(\sigma^2) - 1} \end{aligned}$$

不等於直觀上 (因爲 $X = \exp(Y)$ 而有) 的

$$\frac{\sqrt{\exp(Var(Y))}}{\exp(E(Y))} = \frac{\sqrt{\exp(\sigma^2)}}{\exp(\mu)}$$

而是只與 Y 的標準差有關.

- 對數常態分布的

$$\text{skew} = 3CV + CV^3$$

所以, CV 愈大, 愈偏斜.

- 三參數的對數常態分布: 多一個門檻參數 γ, 用以決定下界, 亦即, 若

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

則稱 X 的分布爲一三參數的對數常態分布, 並以

$$\Lambda(\mu, \sigma, \gamma)$$

表示之.

註 1. $X = e^Y + \gamma$ 可推導出

1. $X \geq \gamma$, 所以 γ 爲一下界.

2. 平均值與變異數分別爲

$$E(X) = E(e^Y) + \gamma$$

$$Var(X) = Var(e^Y)$$

因此, 位置向右移 γ 個單位, 且變異數不變.

註 2. $\gamma = 0$ 時, 三參數等於二參數.

註 3. 三參數的對數常態分布常用於水文學上有關降雨量, 流速與流量, 污染承載量...等的模型化.

3. 二項分布 (Binomial Distribution)

用以模型化在 n 個獨立試驗中某一特定事件發生的次數, 如在環境監測中, 用以模型化某污染物的觀察值中超出清潔標準值的比例或用以比較背景觀察值中與 compliance 觀察值中分別超出標準值的比例.

- 二項分布的機率密度函數

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

  $x = 0, 1, 2, \ldots, n$, 其中 $n =$ 試驗的次數, $p =$ 每次試驗 "成功" 的機率 (亦即, "成功" $\Leftrightarrow$ 特定事件發生).

- 以

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

  表示隨機變數 X 的分布爲二項分布.

- $E(X) = np$, $Var(X) = np(1-p)$

註. 當 n 固定, $np(1-p)$ 爲一拋物線, 圖形如下:

由此可知, 當 $p = 1/2$ 時, 有最大的變異性 (亦即, 當出現正, 反的機率一樣時, 最難猜測下一個試驗

的結果); 當 $p \downarrow 0$ 或 $p \uparrow 1$ 時, 變異性 $\downarrow 0$ (亦即, 愈確知正面或反面一定會發生時, 愈容易猜測結果).

4. 超幾何分布 (Hypergeometric Distribution)

二項分布 $\Leftrightarrow$ 由一有限母體中, 採放回取樣 (replacement sampling) 的結果, 如由一疊 52 張牌中, 任選一張記錄是否爲人頭後, 就放回. 重覆此抽取, 放回的動作 n 次, 並令 X 爲出現的總人頭次數, 則 $X \sim B(n, 12/52)$.

超幾何分布 $\Leftrightarrow$ 由一有限母體中, 採不放回取樣 (without replacement sampling) 的結果, 如以不放回的方式取 5 張牌. 令 $X =$ 此 5 張牌中出現人頭的總數, 則

$$P(X = x) = \binom{12}{x}\binom{40}{5-x}/\binom{52}{5},$$

$x = 0, 1, \ldots, 5$; 並稱 X 的分布爲超幾何分布, 並以

$$H(12, 40, 5)$$

表示之, 其中 $12 =$ 母體中 "成功" (人頭) 的總數, $40 =$ 母體中 "失敗" (非人頭) 的總數, $5 =$ 樣本大小.

- 超幾何分布的機率密度函數

$$f(x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}},$$

$x = 0, 1, 2, \ldots, k$, 其中 $m =$ 母體中的 "成功" 總數 (或某種特定個體 (如, 紅色) 的總數), $n =$ 母體中的 "失敗" 總數 (或非某種特定個體 (如, 非紅色) 的總數), $k =$ 樣本大小.

- 總母體的大小爲 $m + n$, 且以

$$X \sim H(m, n, k)$$

表示隨機變數 X的分布爲超幾何分布, 其機率密度函數如上述.

- 當樣本大小 k 相對小時 (如, $k/(m+n) < 0.1$), 不放回取樣的效應是輕微的, 故

$$H(m, n, k) \approx B(k, m/(m+n))$$

(因爲不放回 $\approx$ 放回的取 k 次, 且成功機率 $= m/(m+n)$)

例. 繪 $H(12, 40, 5)$ 與 $B(5, 12/52)$ 的機率密度函數圖形並比較之.

- 超幾何分布的平均值與變異數分別如下:

$$E(X) = k\left(\frac{m}{m+n}\right)$$

$$Var(X) = k\left(\frac{m}{m+n}\right)\cdot \left[1-\left(\frac{m}{m+n}\right)\right]\left(\frac{m+n-k}{m+n-1}\right)$$

(自行驗証; 參考機統課本)

註. 取 $k = n$ 且 $\frac{m}{m+n} = p$ 時,

$$E(X) = np$$

與二項分布 $B(n,p)$ 的期望值一樣, 且

$$Var(X) = np(1-p)\left(\frac{m}{m+n-1}\right)$$

與 $B(n,p)$ 的變異數差一個有限母體修正因子 (finite population correction factor).

5. Poisson 分布

當 $n \to \infty$, $p \to 0$ 且 $np = \lambda$ (常數) 時, $B(n,p)$ 的機率密度函數

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

$$\to e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0,1,2,\ldots$$

亦爲一種機率密度函數, 並以其中一位發現者的名字稱之 (1873).

適用於模型化一特定地區 (或一特定時段) 內 ''稀有'' 事件的發生次數, 如: 放射性粒子的釋放數, 在某一路口的經過車輛, 隊伍中的人數, 某區域內的植物數, 動物數或微生物體數等.

- 構成 Poisson 分布的條件:
  1. 在非重疊的時間或區域內, 發生的個數是相互獨立的.
  2. 在所有長度相同的時段或面積相同的區域內的平均發生數均爲同一常數.
  3. 當時段的長度或區域的面積 $\downarrow 0$ 時, 在一時段或區域內的平均發生數 $\downarrow 0$.
- 在環境檢測中, Poisson 分布常用於模型化:
  1. 在一個給定的時段之下, 違反污染標準值的次數.
  2. 32 種揮發性有機污染物的掃描中, 偵測出的複合物個數.

3. 化學物質濃度的分布 (單位: ppb).

- Poisson 分布的機率密度函數

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!},$$

$x = 0, 1, 2, \ldots$, 其中 $\lambda =$ 在某一給定時段或區域內的平均事件數.

- 以

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

表示隨機變數 X 的分布爲 Poisson.

- Poisson 分布的期望值與變異數分別如下:

$$E(X) = \lambda,\ Var(X) = \lambda$$

註. 期望值愈大, 變異數就愈大.

例. 繪 Poisson(1) 與 Poisson(3) 的機率密度函數圖形並比較之.

6. Gamma (Pearson Type Ⅲ) 分布

三參數的 gamma 分布, 又稱作 Pearson Type Ⅲ 分布, 與三參數的對數常態分布的應用情形一樣

，有時用於水文學上有關降雨量，流速與流量，污染承載量等的模型化.

- 三參數 gamma 分布的機率密度函數

$$f(x) = \frac{(x-\gamma)^{\alpha-1}\exp\left[-\frac{(x-\gamma)}{\beta}\right]}{\beta^{\alpha}\Gamma(\alpha)},$$

$\alpha > 0, \beta > 0, x > \gamma$, 其中 gamma function

$$\Gamma(x) = \begin{cases} \int_0^\infty t^{x-1}e^{-t}dt & \text{若 } x > 0 \\ \frac{\Gamma(x+1)}{x} & \text{若 } x < 0 \text{ 且} \\ & x \neq -1, -2, \ldots \end{cases}$$

註. α 決定機率密度函數的型態 (shape); β 決定機率密度函數的規模 (scale); γ 決定位置 (location), 又稱作門檻 (threshold) 參數.

- 以 $X \sim \text{gamma}(\alpha, \beta, \gamma)$ 表示隨機變數 X 的分布爲三參數 gamma 分布.

- 當位置參數 $\gamma = 0$ 時, 稱作二參數 gamma 分布, 其

$$E(X) = \alpha\beta$$

$$Var(X) = \alpha\beta^2$$

$$\text{skew}(X) = 2/\sqrt{\alpha}$$

且

$$\text{CV}(X) = 1/\sqrt{\alpha}$$

註. 偏斜性 (skew) 與變異係數 (CV) 完全由型態參考數 α 所決定, 且與 $\sqrt{\alpha}$ 成反比.

例. 繪二參數 gamma(2, 1) 與 gamma(3, 2) 的機率密度函數圖形並比較之.

7. 極值分布 (Extreme Value Distribution)

用以模型化某過程或一序列度量值中的極值 (最大或最小值), 如: 每日的最高溫, 每日最高空氣污染濃度, 日降雨量的年度最大值等.

事實上極值分布是 n 個獨立同分布 (連續) 隨機變數的最大 (或最小) 值的極限分布, 故稱之爲極值分布.

- 有三種不同家族的極值分布. 此處討論的爲 Type Ⅰ, 又稱作 Gumbd 極值分布, 以

$$\text{EV}(\eta, \theta)$$

表示之.

- 二參數的極值分布的機率密度函數

$$f(x) = \frac{1}{\theta} e^{-(x-\eta)/\theta} \exp\left[-e^{-(x-\eta)/\theta}\right],$$

$-\infty < x < \infty$, $-\infty < \eta < \infty$, $\theta > 0$, 其中 η 爲位置 (location) 參數, θ 爲規模 (scale) 參數.

註. 此分布爲最大值分布, 用來模型化最大值.

- 二參數的極值分布的 Mode, 其望值與變異數如下:

$$\text{Mode}(X) = \eta, \ E(X) = \eta + \epsilon\theta$$

$$Var(X) = \frac{\theta^2\pi^2}{6}$$

其中 ϵ 爲 Euler 常數 ≈ 0.5772157.

- skew(≈ 1.14) 與 kurtosis(≈ 5.4) 均爲常數.

例. 繪 EV$(10, 1)$ 的機率密度函數與累積分布函數的圖形並與其他不同參數的 EV 分布比較之.

8. 廣義的極值分布 (Generalized Extreme Value Distribution)

   GEV 除了有位置參數 η 與規模參數 θ 外, 再含型態 (shpae) 參數 κ (kappa).

   - 三種家族的極值分布均爲 GEV 的特例:

     1. $\kappa = 0$, GEV $\Rightarrow$ Type I EV (或 Gumbd) 分布.

     2. $\kappa > 0$, GEV $\Rightarrow$ Type II EV 分布.

     3. $\kappa < 0$, GEV $\Rightarrow$ Type III EV 分布.

   - 環統模組的輔助說明檔提供 GEV 的更多資訊 (自行參考).

9. 混合分布 (Mixture Distribution)

   當母體某部份來自某一分布且其餘部份來自另一分布時, 會產生一個混合分布. 混合分布常用於模型化含有一些或數個顯著 (大) 的離群值 (outliers) 的資料集合, 也即模型化大部份觀察值都來自於某一分布且其餘來自於平均值平移 (或標準差平移) 的分布的資料集合.

- 在環境統計中, 混合分布常用於模型化來自於含有剩餘 (residual) 污染的修復 (補救) (remediated) 地區的資料, 如 Cleanup 區域的 TcCB 資料.

- 由二分布形成的混合分布的機率密度函數爲

$$f(x) = (1-p)f_1(x) + pf_2(x)$$

其中 $f_1()$ 表示第一個分布的機率密度函數, $f_2()$ 表示第二個分布的機率密度函數, 且 $0 < p < 1$ 爲一混合比例 (mixing proportion).

- 混合分布的期望值與變異數如下:

$$E(X) = (1-p)\mu_1 + p\mu_2$$

$$Var(X) = (1-p)\sigma_1^2 + p\sigma_2^2 + p(1-p)(\mu_1 - \mu_2)^2$$

(與課本的不同, 請同學驗証何者正確), 其中 μ_1 與 μ_2 分別爲第一與第二分布的期望值, σ_1 與 σ_2 分別爲第一與第二分布的標準差且 p 爲混合比例.

- S+ 與環統模組可計算二常態分布的混合分布與二對數常態分布的混合分布的機率密度函數, 累積分布函數, 量分位數與隨機數.

例. 試繪 $N(5,1)$ 與 $N(10,2)$ 且 $p = 0.3$ 的常態分布的混合分布. (亦即, $N(5,1)$ 的 70% 與 $N(10,2)$ 的 30% 混合分布.)

10. 零修飾分布 (Zero-Modified Distribution)

一種給予 0 額外機率質量的分布, 爲混合分布的一種特例, 即母體中的一部份來自於原分布, 但其餘部份均設定爲 0 的分布.

- 在環統中, 零修飾對數常態分布 (又稱作 dalta 分布) 常用於內含無法偵側觀察值的資料集合的模型化 (這些無法偵測值 (nondetects) 均被設定爲 0).

- 零修飾分布的機率密度函數爲

$$h(x) = \begin{cases} p & \text{若 } x = 0 \\ (1-p)f(x) & \text{若 } x \neq 0 \end{cases}$$

其中 $f()$ 爲原分布的機率密度函數, p 表示母體中 0 值的比例.

- 零修飾分布的期望值與變異數如下:

$$E(X) = (1-p)\mu$$

$$Var(X) = (1-p)\sigma^2 + p(1-p)\mu^2$$

(與混合分布中的 $Var(X)$ 吻合), 其中 μ 與 σ 分別爲原分布的期望值與標準差.

- S+ 與環統模組可計算零修飾長態分布與零修飾對數常態分布的機率密度函數, 累積分布函數, 量分位數與隨機數.

例. 試繪期望值爲 3, 變異係數爲 0.5 且 0 的比例 $p = 20\%$ 的零修飾對數常態分布.