國立交通大學

電機資訊學院 電信學程

碩士論文

題目：稀少性的輸入資訊下所造成的分佈不匹配問題在語者確認上之可靠度分析

Reliability Analysis Focusing on Sparse Input Data Caused Distribution Mismatch Problems for Speaker Verification

學生姓名：羅文輝

指導教授：陳信宏　博士

中　華　民　國　九　十　五　年　七　月

題目：稀少性的輸入資訊下所造成的分佈不匹配問題在語者確認上之可靠度分析

Reliability Analysis Focusing on Sparse Input Data Caused Distribution Mismatch Problems for Speaker Verification

研 究 生：羅文輝　　Student：Wen-Hui Lo

指導教授：陳信宏　　Advisor：Dr. Sin-Horng Chen

國 立 交 通 大 學

電機資訊學院 電信學程

碩 士 論 文

A Thesis

Submitted to Degree Program of Electrical Engineering and Computer Science

College of Electrical Engineering and Computer Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Communication Engineering

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年七月

稀少性的輸入資訊下所造成的分佈不匹配問題在語者確認上之可靠度分析

研究生：羅文輝　　指導教授：陳信宏 教授

國立交通大學電機資訊學院 電信學程（研究所）碩士班

摘　　要

在語音辨識的領域上，往往需要使用少量的資料來對模型進行校估藉以使得模型更爲強健(robust)。在語者確認的問題上，時常也需要面對資料量很少的情形之下從事語者模型的訓練或測試的問題。

本研究首先提出稀少性資料(sparse data)的輸入情況下，語者確認(speaker verification) 的問題在混合高斯 GMM (Gaussian mixture model)模型上的度量分數分佈情形會產生和原先假設之間有落差的現象。本研究稱此種現象爲「分佈不匹配(distribution mismatch)的問題」。針對此分佈不匹配的問題，本研究首先提出使用截尾分佈機率密度函數(truncated probability distribution function)的概念來近似。最後以此爲基礎，使用次序統計(order statistic)量的概念，推導得出一個以圖(graph)爲基礎的聯合分佈機率模型；可以同時以機率的形式描述完整機率密度函數和截尾分佈機率密度函數。

本研究建立一個以輸入資料，資料之最小值，資料之分佈範圍大小，資料分佈範圍下的累積機率（覆蓋率）及資料長度五個隨機變數的聯合分佈機率密度函數。配合 Gaussian quadrature 積分的取樣概念，得出最少取樣點下最精準的估計公式。最終的目的是希望以較優勢的資訊量補償在傳統的統計推估上，因爲資料量稀少所造成的估計標準誤增加的問題。

最後，本研究以語者語句所獲得之相對於 UBM(universal background model)模型規一化平均分數對 EER(equal error rate)進行假設檢定(hypothesis test)；由實驗的結果得知，假設檢定可以有效的減少語者確認時，因爲抽樣誤差所造成的誤判。

本研究的另外的主要成果在於確立稀少性的輸入資訊下，如果要出現原先我們所假設的分佈狀況的可能性將是一個機率的隨機行爲；本研究所得出的結論：「當輸入的樣本數量小於 20 的時候，輸入樣本的覆蓋範圍和原來的假設 PDF 之間會互相匹配一致」的假設必須使用機率事件來描述才能完全掌握整體隨機變數的特性，而本研究完成了這個機率事件的描述公式。

Reliability Analysis Focusing on Sparse Input Data Caused Distribution Mismatch Problems for Speaker Verification

student：Wen-Hui Lo　　Advisors： Dr. Sin-Horng Chen

Degree Program of Electrical Engineering Computer Science

National Chiao Tung University

ABSTRACT

It is a frequent facing problem for sparse data input to make a robust model testing with speech recognition. This phenomenon also encountered in the field of speaker verification with small data enrollment to do training or testing.

A new approach to sparse data input caused problems named “distribution mismatch(DM)” was addressed. The core of DM which was on account of the coverage of the probability distribution function(PDF) of the input data which are applied to GMM(Gaussian mixture model) score calculation is not full mapping to the original PDF assumption. There maybe be some differences between the original assumption PDF to the new one generated by sparse data input and we suggested to using the truncated probability distribution function for modeling this situation.

The most important addition to be made to what we have said about DM is that we have derived a new joint PDF based on graph theory with order statistic and the new formula would act as the truncated PDF or the original PDF measured by this joint PDF.

We have succeeded establishing the joint PDF which is compose of five random variables, including the input data, the minimum order of input

data, the range of input data, the coverage of input data and the sample size of input data to estimate with Gaussian quadrature integration.

In the end of experiment, we take a hypothesis test to the equal error rate(EER) of the average score per frame of per sentence announced by the speaker normalized to the universal background model(UBM) and the same score announced by imposter normalize to the UBM model.

There are good evidences to show that hypothesis test could decrease the error probability for speaker verification. The other finding finished by this study is that we discover a special fact caused by sparse data input.

We usually regard the input random variable submitted to a certain probability distribution function but it is probabilistic to agree with this assumption when the input sample size is less than 20. Finally, we have derived the joint probability distribution function about it.

致　　謝

感謝口試委員王教授小川和王教授逸如悉心與耐心的指正。也感謝指導教授陳教授信宏的敦敦教誨。我到語音實驗室的時光已經四年，這四年的時間算是無中生有，慢慢累積起對一種完全陌生領域的了解。感謝眾家學弟在資料前處理的協助。感謝葉人鳳和曹志欣兩位先進帶領我入門。也感謝俊良和阿樹把 HTK 研習心得傳授給我。更感謝那位正在從軍報國的 lubo，主動把它的實驗結果都留存給我一份。因爲這些人，使我能夠踩在巨人的肩膀上縮短學習的歷程。

寫完了這本碩士論文之後，對古人的思緒與堅持彷彿有了更深一層的體認。陸遊-「山窮水盡疑無路，柳暗花明又一村」，蘇軾-「不識廬山真面目，只緣身在此山中」。事物的外表總是虛幻充滿未知，有時候， 如果靜下心來細細品味，其實虛幻也可以慢慢建立起認知的基礎。

感謝父親和母親，能夠支持他們的孩子進修第二個碩士學位。如果我能夠有任何一絲的成就，都要感謝父母親日常生活的照顧，還有我那身處在遠方的妻子小蓉，亦感謝她的包容與體諒。

羅文輝　　July 31, 2006

目錄

表目錄

圖目錄

1. 緒論

1.1. 研究緣起

語音辨識(speech recognition)是目前在自動化辨識領域之中最爲成熟的一門技術，並且已經進入商品化的階段。在目前的研究項目之中，稀少性資料(sparse data)的訓練方法是一個熱門的研究領域。我們總是希望以很少的資料得出很好的辨識結果。但是資料量一但減少，很多統計假設都會產生估計偏差，如何減少這些偏差就成爲稀少性資料輸入的熱門研究對象。

1.2. 研究動機

在進行語音辨識的起始階段，通常都要進行人機之間的模式校估階段，藉此來使得辨識的精確度更爲提高。通常在這個階段，我們希望利用最少的資料得到最迅速又正確的結果。這個動作通常我們稱之爲語者調適 (speaker adaptation)。語者調適的前提是我們有足夠的語料訓練基本的辨識模型，然後利用少量的資料得出其他不存在語者的訓練模型。如果資料量真的很稀少，到了某一個臨界值(threshold)以下時，是否我們原來對於資料的假設依然正確呢？

1.3. 研究方法

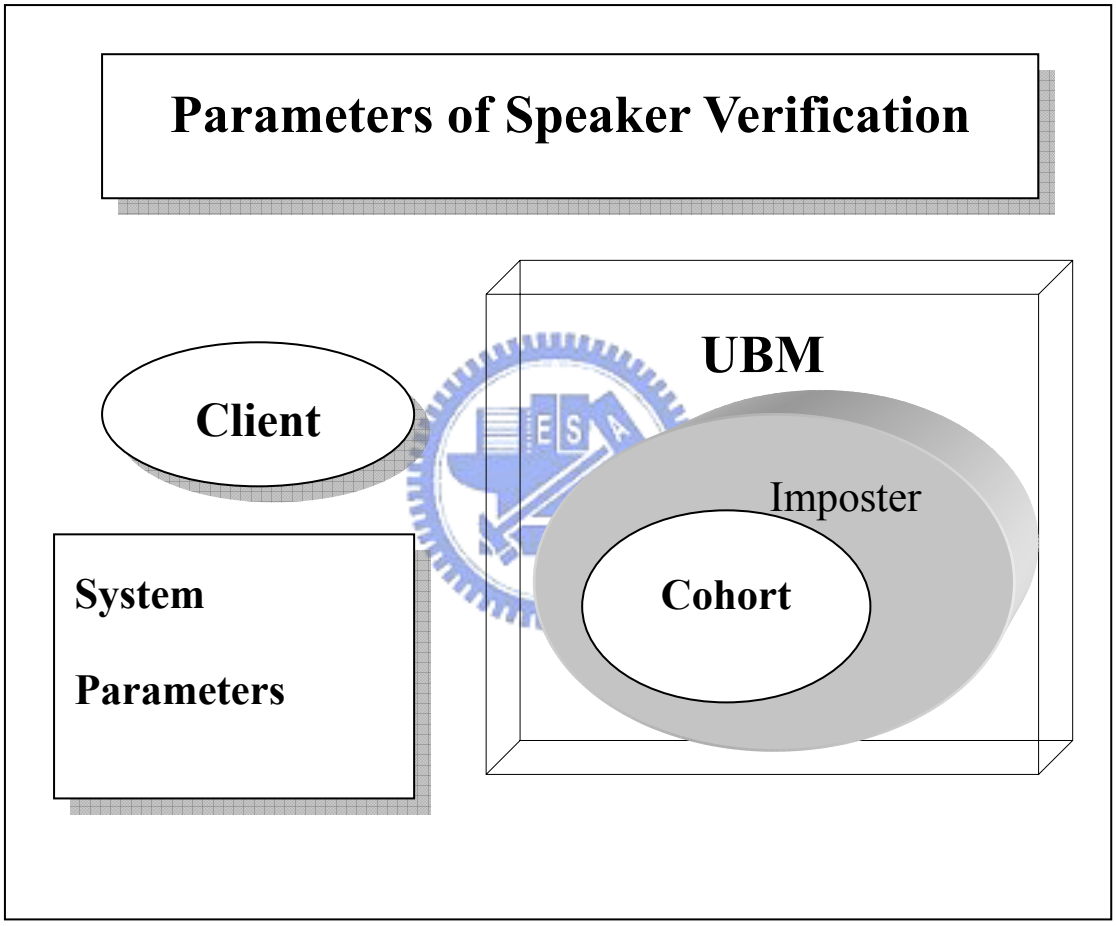
本研究之主要目的在於分析稀少性小樣本的抽樣特性。一般而言，當我們的抽樣數量變的稀少時，往往會和我們原來的假設有所偏離。本研究稱這種現象爲 Distribution Mismatch(DM)，當 DM 現象開始發生時，如果我們能夠使用數學工具將這種效應清楚地描述出來，則模型與資料之間的一致性將會更好，對於促進辨識結果會更有幫助。

DM 可能發生的另外一個情形，就是遮蔽性的輸入資料。假設我們在實驗室的錄音行爲，有時候爲了去除背景的低頻雜音，我們會選擇開始錄音的前一分鐘左右的訊號爲背景成分，然後再將錄音訊號減去背景訊號得出真正乾淨的人體發音訊號成分。如果我們使用遮蔽資料的分析方式，直接將低頻部份截除是否仍然可以保持相同辨識率？另外可能產生的疑問？究竟可以截取多少的資料予以忽略呢？

另外可能附帶產生的疑問？如果我們本身擁有完整的資料，卻故意將資料予以設限(censored data)，只取其中一部分的資料來進行實驗觀察，在這種設限資料(censored data)的推估下，所可能產稱的最佳利益是爲何種？又如何予以應用呢？

1.4. 語者確認文獻回顧

■ 傳統的語者確認方法(Conventional Speaker Verification)



圖表 1 語者確認上之參數蒐集分類

Client: 當事人，語者確認時被假設的對象

UBM: universal background model or world model，對照模型；所有非當事人的語料所構成的聯合模型

Imposter: 僞裝者，在 UBM 的集合中，僞裝成 client 時容易通過受測的個體所成的集合

Cohort: 同隊集合，假定語料庫中所有的 client model 都已經完成。此時想要檢測

其中的某一個 client 時，以該 client 爲基準，以統計方法度量尋找 imposter 中的所有模型參數值，以參數值最接近者，分項組合出最近似受測 client 的模型，作爲對照模型。或者也可以選擇在 UBM 集合中，所得到的測試分數前 Top N 所對應的 imposters 作爲選擇組成 cohort set 的標準。

System parameters:常用的系統參數
錯誤率：主要包含兩個部份，錯誤發報(false alarm,FA)和失誤(false rejection,FR)。
門檻值(threshold)：用來判定輸入與料是否隸屬於假設中的當事人(client)的標準值。常用的兩種門檻值：
EER(equal error rate):使得 false alarm 等於 false rejection 的門檻值。
HER(half error rate)：新的門檻值選取是根據前次的錯誤率一半的效用來決定。HER=1/2(FA+FR)。
Detection Cost Function (DCF)➔偵測成本函數，用來反映整個語者確認系統的好壞程度。

$$DCF = C_{FR} \cdot P_{client} \cdot P(FR \mid client) + C_{FA} \cdot P_{imposter} \cdot P(FA \mid imposter) \quad (1.1)$$

常用的典型值：

$P_{client} = 0.01$ and $P_{imposter} = 0.99$

$C_{FR} = 10$ following the NIST recommendation

$C_{FA} = 1$

根據語料內容，一般將語音轉化成 MFCC(Mel scale frequency cestral coefficient) 特徵參數，經過向量量化(vector quantization)以及分群(clustering)演算。最後再使用混合高斯模型 GMM (Gaussian mixture model)，建立語者(client)的模型以及背景(UBM, universal background model)。以下假設有 M 位 speaker 欲進行 speaker verification。則實際上的執行步驟如下：

使用 EM(Expectation maximization)演算法，對每一位 client 求算 GMM 參數➔所以每一位 client 的 PDF(probability distribution function)都可以使用混合高斯進行展開。

$$\mathbf{P}(x_j \mid \hat{s}) \Rightarrow \mathbf{P}_{\hat{s}}(x_j) = \sum_k c_k^{(\hat{s})} \cdot N(x_j; u_k^{(\hat{s})}, \Sigma_k^{(\hat{s})}) \quad (1.2)$$

$P_{\hat{s}}(x_j)$: *PDF* of client，每位當事人的機率密度函數

$c_k^{(\hat{s})}$: mixture weight，在client上，每個混合高斯成分的權重值

$u_k^{(\hat{s})}$: *mean* vector，client上的混合高斯成份平均向量

$\Sigma_k^{(\hat{s})}$: covariance matrix，client上的混合高斯成份協方差矩陣

x_j: *the* j-th input frame of feature vector，第j幅輸入MFCC向量

k:mixture index，混合高斯成分計數編號

針對當事人以外的所有受測試語者訓練背景模型 universal background model(UBM)參數。通常最簡單的方式是將這些非當事人的 client model 進行平均。

$$P(x_j \mid \Omega) \Rightarrow \mathrm{P}_{\Omega}(x_j) = \frac{1}{M-1}\sum_i \log\left(\sum_{i,m} c_{i,m}^{(\Omega_i)} \cdot N(x_j; u_{i,m}^{(\Omega_i)}, \Sigma_{i,m}^{(\Omega_i)})\right) \tag{1.3}$$

$\mathrm{P}_{\Omega}(x_j)$:背景模型之機率密度函數

$\mathrm{c}_{\mathrm{i,m}}^{(\Omega_i)}$: 第i位偽裝者的第m個混合高斯成分權重值

i=1,2…M-1,偽裝者集合中的語者編號

$u_{i,m}^{(\Omega_i)}$:第i位偽裝者的第m個混合高斯成分平均向量

$\Sigma_{i,m}^{(\Omega_i)}$: 第i位偽裝者的第m個混合高斯成分協方差矩陣

x_j:第j幅輸入MFCC向量

計算概度比值並且進行判斷

$$llr(x_j) \Rightarrow \log\left(\frac{\mathrm{P}_{\hat{s}}(x_j)}{\mathrm{P}_{\Omega}(x_j)}\right) = \log\left(\sum_k c_k^{(\hat{s})} \cdot N(x_j; u_k^{(\hat{s})}, \Sigma_k^{(\hat{s})})\right)$$
$$-\frac{1}{M-1}\sum_i \log\left(\sum_{i,m} c_{i,m}^{(\Omega_i)} \cdot N(x_j; u_{i,m}^{(\Omega_i)}, \Sigma_{i,m}^{(\Omega_i)})\right) \underset{\Omega}{\overset{\hat{s}}{>}} \log \Lambda \tag{1.4}$$

$$\text{相似度比}\left(likelihood\ \text{ratio}\right) \Rightarrow LR(x) = \frac{p(x \mid \hat{s})}{p(x \mid \Omega)}$$

$\hat{s}$:當事人(client)模型

Ω:偽裝者模型

x:輸入MFCC特徵向量 (1.5)

■ 決策準則

$$if\ LR(x)\begin{cases} > \Lambda, decide\ \mathrm{x} \in \hat{s} \\ < \Lambda, decide\ \mathrm{x} \in \Omega \end{cases} \tag{1.6}$$

Λ：自行選定的門檻值

以上(1.2)到(1.6)就是典型的語者確認過程。

■ 相似度分數標準化(Likelihood Score Normalization)

上一節所描述的是單一 frame 作爲輸入的過程，但實際上這樣的決策風險過高，並不實用。所以實用上通常是取多個 frame 的輸入分數進行平均。

首先將 GMM 的輸出分數取對數 log(.)運算。GMM 的輸出取 log(.)運算之後將會形成高斯分佈

$$Define \Rightarrow\ \mathrm{llr}(\mathrm{x}_j) = \log(LR(x_j)) \tag{1.7}$$

多重輸入 n frame 下之決策分數爲：

$$\text{Sequence Decision} \Rightarrow llr(X) = \frac{1}{n}\sum_j \log(\mathrm{P}_{\hat{s}}(x_j)) - \frac{1}{n}\sum_j \log(\mathrm{P}_{\Omega}(x_j))$$

$$X = [x_1, x_2 \cdots x_j \cdots x_n], x_j \text{ is feature vector(MFCC)}$$

(1.8)

$$\begin{cases} if\ (llr(X)) > \log\Lambda, \text{ decide client} \\ if\ (llr(X)) < \log\Lambda, \text{ decide imposter} \end{cases} \tag{1.9}$$

式子(1.9)是一般在進行speaker verification的判別式子，但是一般因爲語者確認在應用上的區別，我們會將(1.8)進行所謂的score normalization藉以獲得更好的

辨識結果。一般語者確認常用的score normalization分別有T norm和Z norm兩種R. Auckenthaler et. all.[1], C. Barras and J.-L. Gauvain[2]：

■ 針對僞裝者模型之分數標準化（Score Normalization of Imposters of UBM or Cohort Set）

最基本的分數計算可以使用 UBM 標準化分數，統計輸入的測試語料在當事人和僞裝者兩個不同集合的平均分數差距。另外 cohort 的分數計算方式也相當近似，只是 cohort set 通常會選取和當事人（client）比較近似的僞裝者(imposter)，其餘剩下的 imposters 會被捨去，所以 cohort set 通常是 UBM set 之部分集合。

$$S_{imposter}(X,\hat{s}) = \frac{\gamma}{n}\sum_{j}\log(p_{\hat{s}}(x_j \mid \hat{s})) - \frac{\gamma}{n}\sum_{j}\log(p_{\Omega}(x_j \mid \Omega)) \quad (1.10)$$

γ : compensating coefficient for independent assumption：**補償係數**

$S_{imposter}(X,\hat{s})$：決策分數

γ 稱爲補償係數，用來補償因爲假設 client 和 imposter 兩個集合是完全獨立的兩個集合所產生的誤差。

T norm(Test norm):只針對 client 部分的分數進行歸一化的處理。

如果將每次的輸入序列訊號 X 在 client model 上的平均分數 ($\rightarrow \frac{\gamma}{n}\sum_{j}\log(p_{\hat{s}}(x_j \mid \hat{s}))$) 視爲是一個隨機變數 $\hat{s}_c(X)$

$$S_{T\_norm}(X,\hat{s}) = \frac{\hat{s}_c(X) - u_{\hat{s}_c}(X)}{\sigma_{\hat{s}_c}(X)} \quad (1.11)$$

$u_{\hat{s}_c}(X)$：X 對所有的 imposter 成員 model 所得出的分數平均值

$\sigma_{\hat{s}_c}(X)$：X 對所有的 imposter 成員 model 所得出的分數標準差

這種T norm的好處是它可以離線單獨先進行運算出平均數和標準差。另外一種T norm的寫法是針對輸入的測試訊號進行score normalization 由Mariethoz, J.and Bengio, S. [3]所發表，假設 $\boldsymbol{X}$ 是某一句測試語料，如果對所有個別的僞裝者模型都進行分數計算，然後取得分數歸一化的結果，稱爲T norm。

$$llr_i(x_j) \Rightarrow \log(\sum_k c_k^{(\hat{s})} \cdot N(x_j; u_k^{(\hat{s})}, \Sigma_k^{(\hat{s})})) - \log(\sum_{i,m} c_{i,m}^{(\Omega_i)} \cdot N(x_j; u_{i,m}^{(\Omega_i)}, \Sigma_{i,m}^{(\Omega_i)}))$$

(1.12)

總共有 M-1 位 imposter 模型，

$$llr_i(X) \triangleq\Rightarrow$$

$$\frac{1}{n}\sum_{j=1}^{n}\left\{\log(\sum_k c_k^{(\hat{s})} \cdot N(x_j; u_k^{(\hat{s})}, \Sigma_k^{(\hat{s})})) - \log(\sum_{i,m} c_{i,m}^{(\Omega_i)} \cdot N(x_j; u_{i,m}^{(\Omega_i)}, \Sigma_{i,m}^{(\Omega_i)}))\right\}$$

n :length of X (1.13)

$$S_{T\_norm(i)}(X,\hat{s}) = \frac{llr_i(X,\hat{s}) - u_{T(X,\hat{s})}}{\sigma_{T(X,\hat{s})}} \quad (1.14)$$

$$u_T(X,\hat{s}) = \frac{1}{M-1}\sum_{i=1}^{M-1} llr_i(X)$$

$$\sigma_T(X,\hat{s}) = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M-1}(llr_i(X) - u_T(X,\hat{s}))^2}$$

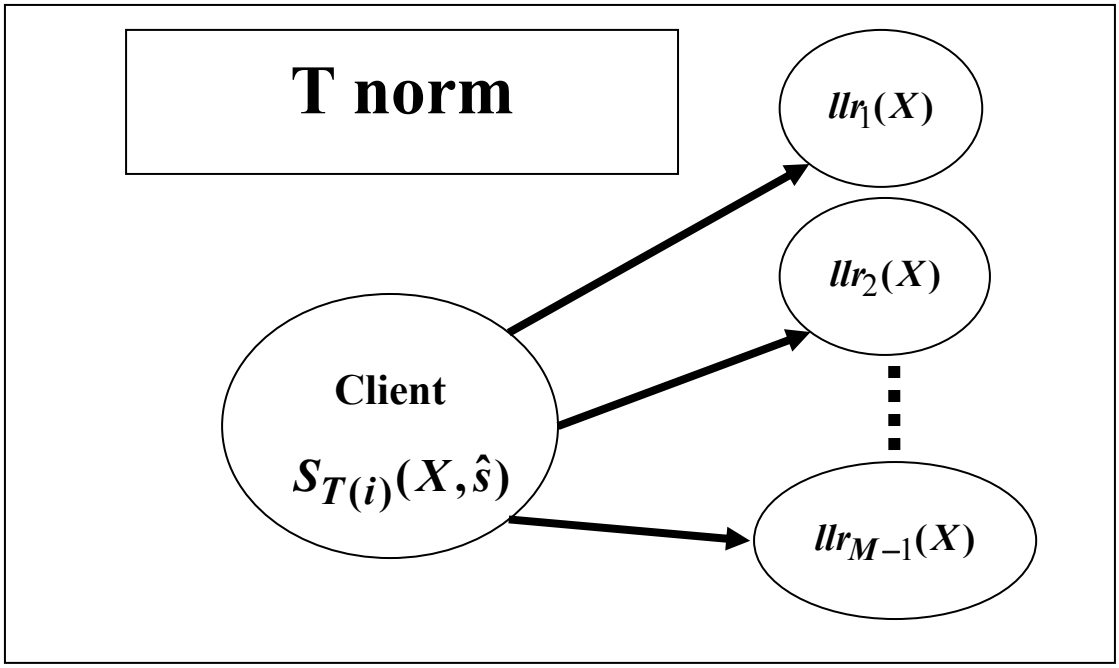
i: index of imposter

ŝ:client

$X = [x_0, x_1, \cdots x_j \cdots x_n]$

j : *index* of frame

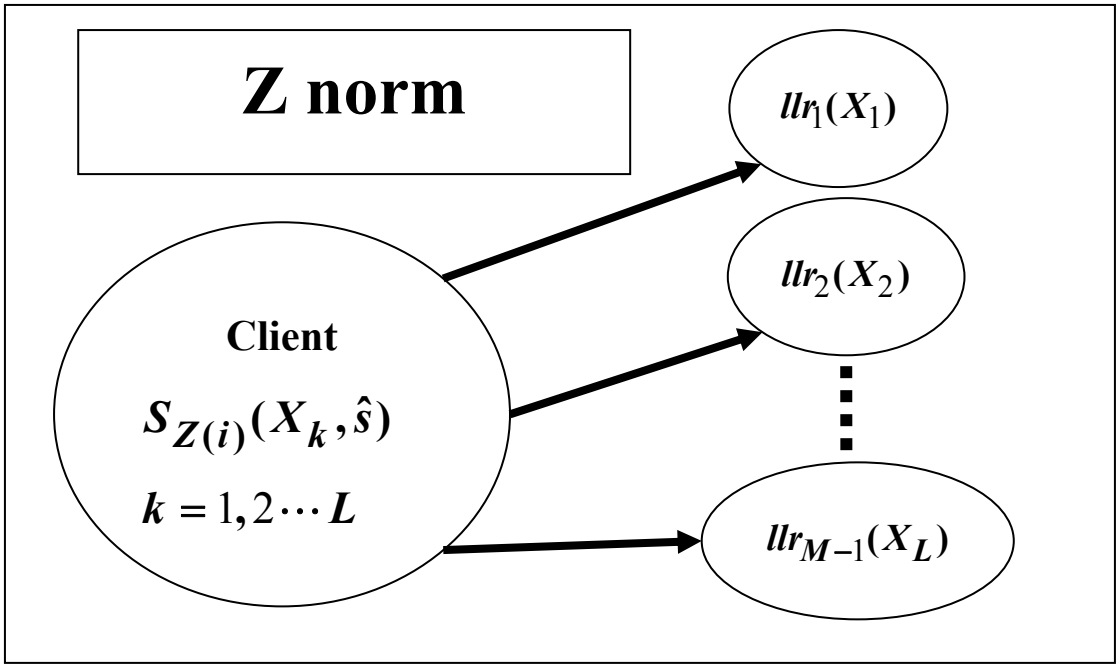
M-1: number of imposter



圖表 2 T norm 之構成概念

T norm 的精神是測試語料對於 client 和 imposter 兩個模型所可能產生的分數進行歸一化的處理。

Z norm(Zero norm): 如果總共有 L 句已知來自於 imposter 的測試語料。$[X_1, X_2, \cdots X_k \cdots, X_L], k = 1 \cdots L$ 針對僞裝者(imposter)集合所得出的決策分數 $S(X_k, \hat{s})$ 可以視爲常態分佈，然後進行歸一化處理即爲 Z norm。



圖表 3 Z norm 之架構

$$S_{Z(i)}(X_k,\hat{s})=\frac{llr_i(X_k,\hat{s})-u_{S_Z(X,\hat{s})}}{\sigma_{S_Z(X,\hat{s})}} \tag{1.15}$$

$X_k\in\Omega_i, k=1,2,\cdots L, i=1,2,\cdots M-1$, relation of L and M is independent

$$u_{S_Z(X,\hat{s})}=\frac{1}{L}\sum_{k=1}^{L} llr_i(X_k,\hat{s})$$

$$\sigma_{S_Z(X,\hat{s})}=\sqrt{\frac{1}{L}\sum_{k=1}^{L}\left\{llr_i(X_k,\hat{s})-u_{S_Z(X,\hat{s})}\right\}^2}$$

(1.15)代表將(1.9)序列輸入訊號 X 的 $llr(X)$ 分數進行歸一化(normalization)。然後再使用歸一化之後的分數來進行語者確認上的判定工作。

2. 可靠度相關文獻回顧

在語者確認的問題上，以資料量多寡爲依據進行可靠度分析評估的文獻目前尚付之闕如；在語音的研究問題上，大部分的可靠度分析問題都是以雜訊或是環境的影響作爲研究分析對象。

2.1. 以雜訊爲影響基礎之可靠度分析

一般現存之文獻在評量可靠度的問題時多以雜訊作爲影響變因或者整合環境和辨認器之間的交互行爲探討。如Ganchev. et al.[4]，針對時變環境下的雜訊提出MMSE(Minimum mean square error)的criteria加強語音訊號MFCC(Mel-Frequency Cepstral Coefficients)的方法，得出近似乾淨(clean)的語音訊號。Richiardi et all.[5]針對較惡劣的環境下語者確認的問題，提出一套信賴度及可靠度分析的方法，Jonas Richiardi比較大的突破點，是其所建立的評估方法可以適應當GMM的輸出分數不再假設爲常態分佈時依然有效。Chaudhari. et. al. [6]針對影像及語音雙重資料輸入的情況下，互動式的影音雙重語者辨認器。Chaudhari. 根據所接收的訊號，和原來的訊號成分互相比較之後，建立評分等級，作爲處理訊號時的可靠度依據。

M. Arcienega, A. Drygajlo[7] 進行融合pitch 與spectral envelope features兩項特徵參數的語者確認問題。研究中將受到雜訊影響的資料予以捨去，然後使用Bayesian network的圖論模型輔助求解GMM參數，結果顯示可以有效的提升語者確認模型在雜訊影響下的可靠度。Leung et al. [8] 等人則是研究語者自身的特性對模型可靠度的影響，語者本身受到先天性語言學習與發音的影響，即使未來面對熟析的語言發音時，仍然會受到先天性語言學習所遺留的發音習慣所影響，因此提出以連接音 (articulatory)爲基礎，建構一套條件式發音模型(conditional pronunciation modeling)的語者確認機制，結果也能提升語者確認問題在不同語言之間適用的可靠度。

可靠度分析的問題實際和語音上的(confidence measure)問題相似，當我們的語者確認模型要適用到新的環境或者是語者結構產生新的改變，測試語料產生大幅度的變動時，這時候會想要評估原來所訓練的語者確認模型可以正確工作的可靠度有多少？如果可靠度不足時，可以重新訓練語者確認模型或者進行模型調適

(adaptation)，藉此獲得更好的工作模型。

E. Mengusoglu [9]將一般confidence measure使用在word 或phoneme上的事後(posterior)與事前(prior)機率比較的概念。轉移到語者確認的問題上進行研究。有別於以往的作法，E. Mengusoglu 將GMM的參數規劃爲兩個state。分別是word或syllable等級的silence和speech以及phoneme等級的voice和unvoice各兩個state。以normalization score（如T norm或Z norm）的相關係數(correlation coefficient)進行inverse Fisher轉換，然後度量系統模型的可靠度，決定是否要對模型進行調適。

E. Mengusoglu 在語音上比較大的突破，在於將原本應用在 text-dependent 的 confidence measure 工具發展到 text-independent 的語者確認問題上進行應用。

2.2. 使用統計觀點來看待語者確認中之分數標準化過程

本研究的主要觀點在於評估稀少資料輸入情況下對於語者確認模型強健程度的影響。研究的假設前提認爲訓練模型具有強健性。但是當測試環境和訓練環境產生不匹配的問題時，這時候使用該模型的可靠度剩下多少？

本研究首先著眼於 GMM 輸出值取對數之後的分佈行爲。如果我們將 GMM 的輸出值取對數之後的結果視爲一個隨機變數，一般在處理上都視爲常態分佈。

一般在處理這類的問題時常會使用 T norm 或者是 Z norm 這兩種 score normalization 技術，由於前一章已經對 T norm 或者是 Z norm 進行過介紹。此處便不再贅述。有鑒於 T norm 或者是 Z norm 基本上還是在處理大樣本的問題，使用上並無法符合本研究小樣本的要求。所以本研究在此處自行提出小樣本的 score normalization 方法。對 GMM 取對數運算的輸出分數進行判別的工作，實際上可以視爲是檢定兩個常態分佈隨機變數的平均值差異現象，我們可以直接將 GMM 的輸出判別過程整理如下：

$$p_{\hat{s}}(x_j)=\sum_k c_k^{(\hat{s})}\mathrm{N}(x_j;u_k^{(\hat{s})},\Sigma_k^{(\hat{s})})$$
$$p_{\Omega_i}(x_j)=\sum_{i,m} c_{i,m}^{(\Omega_i)}\mathrm{N}(x_j;u_{i,m}^{(\Omega_i)},\Sigma_{i,m}^{(\Omega_i)}) \quad (2.1)$$

按：$p_{\Omega}(x_j)$模型也可以直接將所有的 *speaker* 語料輸入，使用 *GMM* 訓練得出。

$$llr(x_j)=\log(\sum_k c_k^{(\hat{s})}\mathrm{N}(x_j;u_k^{(\hat{s})},\Sigma_k^{(\hat{s})}))$$
$$-\frac{1}{M-1}\sum_i \log(\sum_{i,m} c_{i,m}^{(\Omega_i)}\mathrm{N}(x_j;u_{i,m}^{(\Omega_i)},\Sigma_{i,m}^{(\Omega_i)})) \quad (2.2)$$

$X=[x_1,x_2,\cdots x_n]=\{x_j\}, j=1\cdots n$:

$$llr(X)=\frac{1}{n}\sum_{j=1}^{n}\log(LR(x_j))$$
$$\Rightarrow \frac{1}{n}\left\{\sum_{j=1}^{n}\{\log(\sum_k c_k^{(\hat{s})}\mathrm{N}(x_j;u_k^{(\hat{s})},\Sigma_k^{(\hat{s})}))-\frac{1}{M-1}\sum_i \log(\sum_{i(m)} c_{i(m)}^{(\Omega_i)}\mathrm{N}(x_j;u_{i(m)}^{(\Omega_i)},\Sigma_{i(m)}^{(\Omega_i)}))\}\right\}$$
$\Rightarrow$ Approximate student t distribution
(2.3)

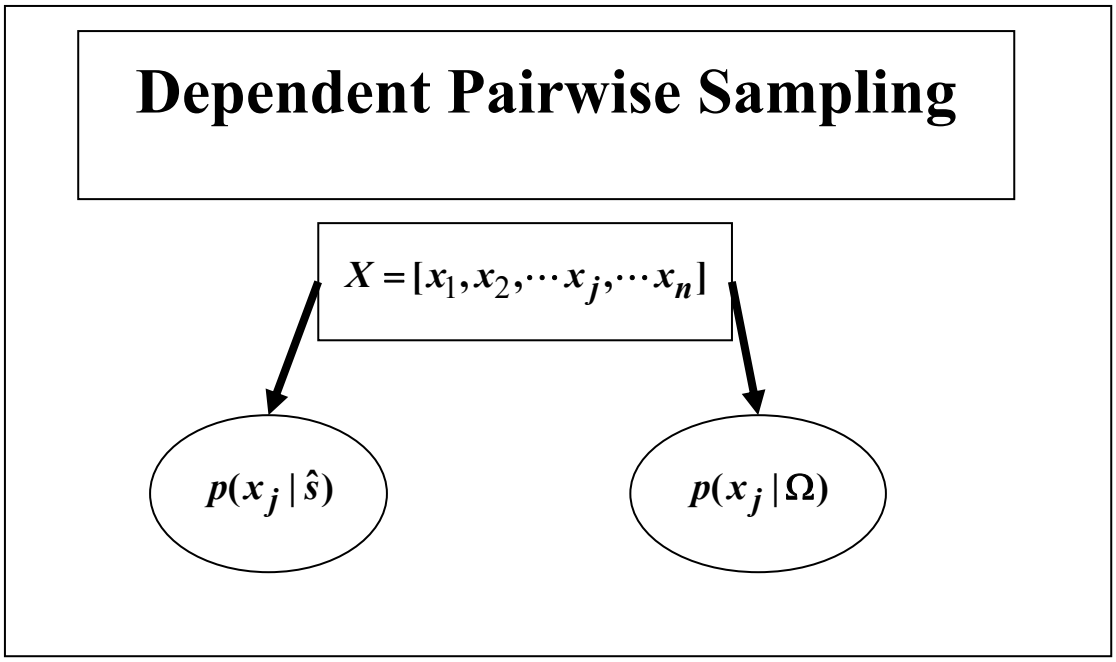
依據統計學的理論，對兩個相關樣本集合的抽樣分佈平均值度量，可以使用 student t 分佈進行等效轉換。所以(2.3)將會近似於自由度爲 n-1 的 t 分佈。

Student-t Distribution of sample size n

$$f_n(t)=\frac{\Gamma(\frac{1}{2}n)}{\sqrt{(n-1)\pi}\,\Gamma(\frac{1}{2}(n-1))(1+\frac{t^2}{(n-1)})^{\frac{n}{2}}} \quad (2.4)$$

Γ : ***Gamma*** function

n:sample size of random variable of score normalization



圖表 4 相似度比值視爲相關樣本之抽樣行爲

語者確認在當事人和僞裝者兩個集合上所得的分數是建構在同一串輸入序列 X 上，所以這兩個集合上的分數比較，可視爲相關樣本下的兩個樣本的平均值差異度量，顏月珠，「商用統計學」[10]。

$$\overline{s(X)}=\frac{1}{n}\sum_{j=1}^{n}\log\left(\sum_{k}c_k^{(s)}\mathrm{N}(x_j;u_k^{(s)},\Sigma_k^{(\hat{s})})\right)$$
$$\overline{\Omega(X)}=\frac{1}{n(M-1)}\sum_{i=1}^{M-1}\sum_{j=1}^{n}\log\left(\sum_{i,m}c_{i,m}^{(\Omega_i)}\mathrm{N}(x_j;u_{i,m}^{(\Omega_i)},\Sigma_{i,m}^{(\Omega_i)})\right) \quad (2.5)$$

在(2.5)的前假設下，決策分數 $llr(X)$ 可以視爲兩個常態分佈的平均值進行相減所得出的隨機變數。

$$\overline{llr(X)}=\overline{s(X)}-\overline{\Omega(X)} \quad (2.6)$$

$X=[x_1,x_2,\cdots x_n]=\{x_j\}, j=1\cdots n$: input testing sequence of feature vector

■ Hard Decision

傳統的語者確認方法會將所得到的分數和事先我們自己所選定好的門檻值

來進行比較，如果分數大於門檻值(threshold)，則輸出結果確認是假設中的當事人(client)，反之，則判定爲僞裝者(imposter)。至目前爲止，門檻值(threshold)的選取方式多半只有兩種➔EER 或 HER，使用 deterministic 的選取方式來決策。

$$\text{Sequence Decision} \Rightarrow \boldsymbol{llr(X)} = \frac{1}{\boldsymbol{n}}\sum_{\boldsymbol{j}} \mathbf{log}(\mathbf{P}_{\hat{\boldsymbol{s}}}(\boldsymbol{x_j})) - \frac{1}{\boldsymbol{n}}\sum_{\boldsymbol{j}} \mathbf{log}(\mathbf{P}_{\boldsymbol{\Omega}}(\boldsymbol{x_j})) \tag{2.7}$$

$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x_j} \cdots \boldsymbol{x_n}], \boldsymbol{x_j}$ is single frame of feature vector,j=1,2⋯n

$$\begin{cases} \boldsymbol{if}\ (\boldsymbol{llr(X)}) > \mathbf{log}\,\boldsymbol{\Lambda}, \text{ decide client} \\ \boldsymbol{if}\ (\boldsymbol{llr(X)}) < \mathbf{log}\,\boldsymbol{\Lambda}, \text{ decide imposter} \end{cases} \tag{2.8}$$

$\boldsymbol{\Lambda}$: ***threshold***

■ Soft Decision

因爲我們所研究的目標是稀少資料的小樣本情形，所以此處針對小樣本的情況引入新的 criteria。給予兩個常態分佈的母體，如果抽樣的順序是成對抽取(pairwise sampling)的模式，同時在兩個母群體中對相同的隨機變數 x_j 進行 log likelihood 之計算)，此時可以視爲相關的小樣本處理。

$$\boldsymbol{llr(X_k)} = \overline{\boldsymbol{s(X_k)}} - \overline{\boldsymbol{\Omega(X_k)}} \tag{2.9}$$

k : k-th input frame sequence

$$Let \quad D_j = S(x_j) - \boldsymbol{\Omega}(x_j), j = 1 \cdots n$$

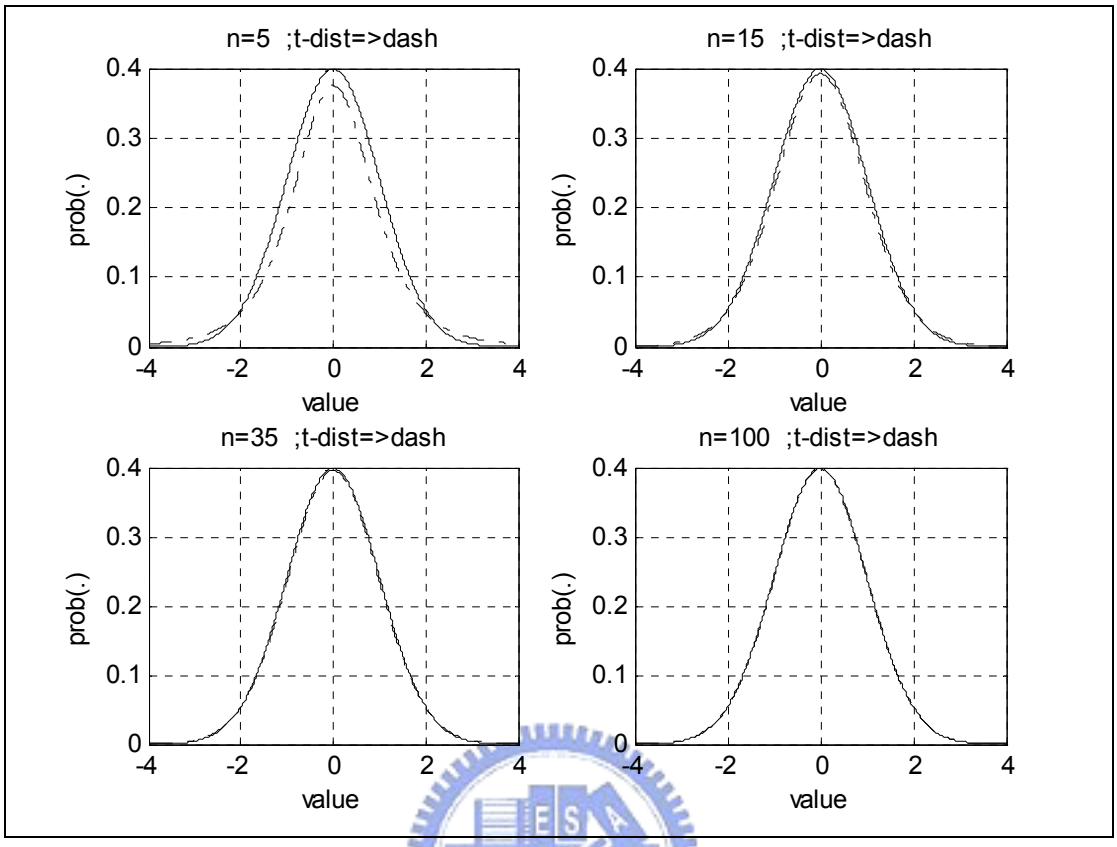
$$\Rightarrow sample \text{ mean} : \overline{D} = \frac{\sum_{j=1}^{n} D_j}{n} = \overline{S(x)} - \overline{\boldsymbol{\Omega}(x)} \tag{2.10}$$

$$sample \text{ variance} \Rightarrow s^2(D) = \frac{\sum_{j=1}^{n} (D_j - \overline{D})^2}{n-1} \tag{2.11}$$

應用樣本的平均數和變異數，可以得出近似的 t 分佈。

$$t = \frac{\overline{D} - u(D)}{s(D)/\sqrt{n}} \tag{2.12}$$

T 分佈的參數僅和自由度（degree of freedom）有關，df=n-1。



圖表 5 difference between normal and student-t distribution

- **使用假設檢定的方式計算出 false alarm 和 false rejection**

在統計理論上，兩個相關樣本的平均值差可以使用(2.12)的 t distribution 來近似。對(2.12)的 t distribution 定義式進行比對，可以寫出原來的 likelihood ratio test 轉換到 t distribution 上的 Hypothesis test 的對應寫法。

- **Condition➔輸入序列長度爲 n**

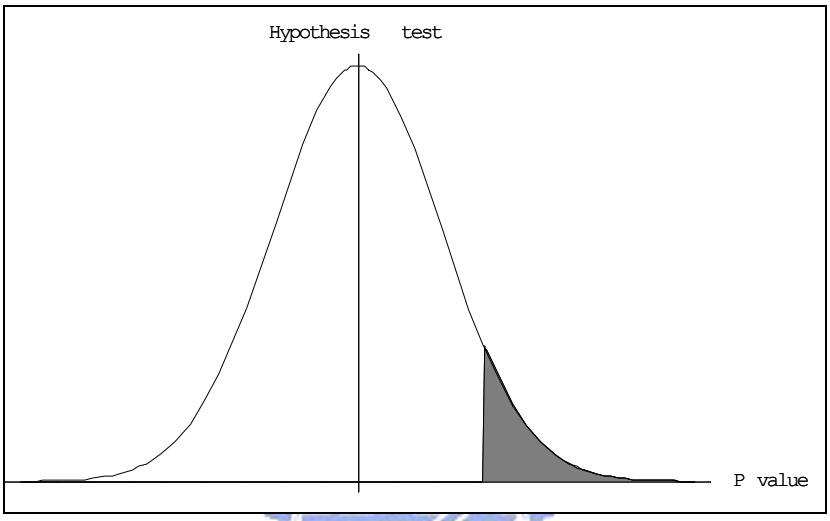
- **檢定是否爲 client? Right-tailed test**

$$\begin{aligned} &\mathrm{H}_0 : \mu(D) \le \log(\Lambda) \\ &\mathrm{H}_1 : \mu(D) > \log(\Lambda) \\ &if \quad t_0 = \frac{decision(x) - \log(\Lambda)}{s(D)/\sqrt{n}} > t_{(1-\alpha,n-1)} \Rightarrow reject\ \mathrm{H}_0 \end{aligned} \tag{2.13}$$

- **檢定是否爲 imposter? Left-tailed test**

$$\mathbf{H}_0 : \mu(D) \geq \log(\Lambda)$$
$$\mathbf{H}_1 : \mu(D) < \log(\Lambda) \qquad (2.14)$$
$$if \quad t_0 = \frac{decision(x) - \log(\Lambda)}{s(D)/\sqrt{n}} < -t_{(1-\alpha,n-1)} \Rightarrow reject\ \mathbf{H}_0$$

$\alpha : risk$ (level of significance)



圖表 6 Hypothesis test (P value test)

如果 training 的結果已經到達 optimization，Hypothesis test 的 p value 將會到達 minimum。P value 值越小表示相對風險(risk)就會越小，也就是說，這時候的 GMM 參數訓練的結果，其相對於其它的值是風險最小的。

本研究之所以稱第二種假設檢定的方法爲 soft decision，主要的觀點在於 level of significance➔α， 是我們可以自由挑選的一個參數。一般統計上在進行假設檢定時α的典型值是 0.1 或 0.05。

小結：T norm 和 Z norm 實際上是上述 t student distribution 的變形處理，然而在語者確認上的應用卻可以增強模型的強健程度(robustness)。本研究在此處想要建立稀少資料量(sparse data)之下的語者確認(speaker verification)模型穩定度(reliability)分析。

本研究欲進行的研究方向：

一、因爲取樣數量之減少所造成的分佈不匹配(distribution mismatch)問題之下的語者確認實驗結果。

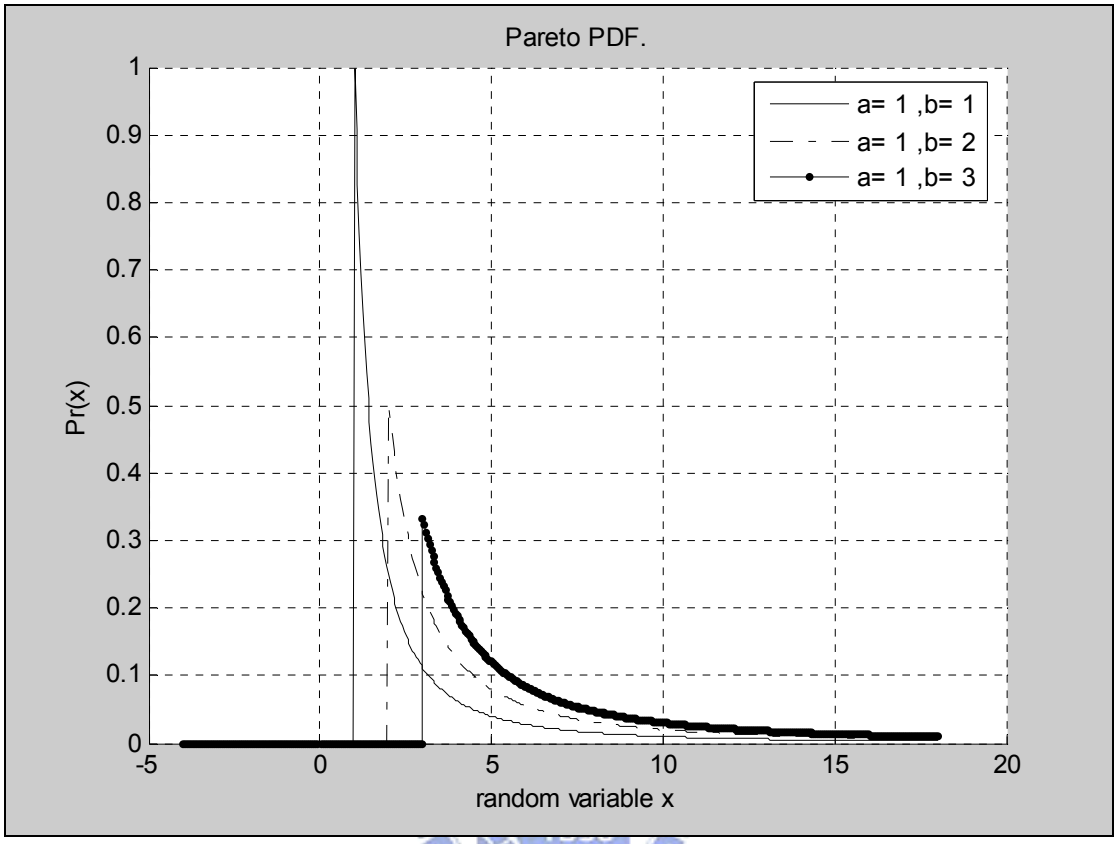
將原來的機率密度函數進行遮罩處理的研究結果在以往的研究結果中曾經有相當不錯的成果。最主要的研究方向是和時間有關的應用。例如工業工程上面用來進行生產線之產品良率估計實驗設計。在大量生產的產品線上，有時候因爲產品的數量太多或者是產品的價錢昂貴，不可能讓研究者等到所有的產品都損壞之後才去推估全部產品的平均生命週期(lifetime)，這時候就必須假設產品的生命週期是服從某種分佈，然後研究者將這個分佈的左端或右端進行遮罩(mask)處理，但是估計時仍然以整個未受遮罩處理的樣本空間爲對象做分析。應用這種概念可以使用少量的時間快速得到整體的實驗結果。工業上在此方面的應用稱爲產品生命週期分析(lifetime analysis)。

2.3. 工業產品之壽命分析(Lifetime Analysis)

在大量生產的生產線上，如果想要對所有的產品進行壽命測試得出結果。通常是費時而且不合乎成本原則的考慮，因爲我們無法預知所有的產品其測試時間需要多久。在以成本爲前提的考量之下，通常我們將只會蒐集部分產品進行測試，其所節省下來的時間就是我們所獲得的效益。 最典型的問題可以想像一般我們購買電腦時的燒機測試，通常我們只進行了數十個小時的開機測試，如果電腦仍然能夠正確無誤的工作時，可以推論這台電腦系統在未來的時間裡頭將可以穩定的工作。

工業上一般將產品的特性施以測試，將所有產品的性能是否到達所要求的品質測試過程視爲是一種可靠度(reliability)的測試。在規劃這種可靠度的實驗時，總是希望能夠以最少的時間得到最佳的實驗結果。而以設限資料(censored data)的方式來蒐集實驗數據取代做完所有實驗即可達成此種目的。

H.I. Hamdy[11]使用Pareto distribution來描述產品的生命週期。下列是Hamdy所舉出的實例，主要的用途是觀測 15 種新興行業從起始到結束的營運時間。



圖表 7 Pareto 的機率密度函數圖形

$$p(x) = \frac{a \cdot b^a}{x^{a+1}}, x \geq b \tag{2.15}$$

$$D(x) = 1 - \left(\frac{b}{x}\right)^a, x \geq b \tag{2.16}$$

P(x): PDF of Pareto distribution
D(x): CDF of Pareto distribution

假設公司營運的時間長度符合 Pareto distribution，假定 $x_{i:n}$,i=1..k 代表結束營運的 k 家新興行業。n 爲全部起始投入的營運企業總數。

假設這些新興企業的存活時間（以年爲單位）可以使用 Pareto distribution 來描述。則剩下的(n-k)家新興企業的可能存活時間便是我們想要快速知道的結果。Hamdy 以樣本數 n=15 個新興行業運作時間爲實驗資料，每個新興行業的運作時

間至少爲一年。根據(2.15)➔ $p(x)=\frac{a\cdot b^{a}}{x^{a+1}}, x\geq b$ (2.17)

則可以知道 b=1，a 在 Pareto distribution 中稱爲 shape parameter.

Hamdy 只觀察前 10 個新興行業停止運作的時間，求出 shape parameter a=4.23。 緊接著，使用次序統計量求出$\mathbf{E}\{n_{10:15}\}$=1.28(年)，表示如果每次只觀察到第 10 種新興行業停業，就停止實驗觀察的總共使用時間平均值。另外再求出$\mathbf{E}\{n_{10:10}\}$=2.11(年)。Hamdy 做出結論，如果採用設限資料(censored data)(例如本例中共有 15 種新興行業需要調查，但是觀察實驗只進行到第 10 個就停止)進行分析，可能所使用的時間比相同樣本大小的完整資料觀察實驗還要少。在本例中

$$\mathbf{E}\{n_{10:15}\}=1.28<\mathbf{E}\{n_{10:10}\}=2.11$$

總共節省了大約 39%的寶貴時間。

2.4. 醫學上之臨床統計應用(Survival Analysis)

另外一個應用方向是醫學統計，醫學統計上針對不同病人不同時間之內所獲得之臨床資料往往無法進行相互之間的比較。例如後天免疫性症候群(AIDS)的患者，既無法斷定其何時開始感染病菌，也很難推估出這名患者未來還可以存活的時間有多久；這時候就必須使用從前病人所留下的部份臨床資料來進行統計資料的全面性推估。醫學統計上的這種特殊統計分析稱爲倖存分析(survival analysis)。

Lagakos et al.[12]曾經對 258 名成人和 37 名兒童進行AIDS的病症觀察研究。研究觀察期間自 1978 年的 4 月 1 日至 1986 年的 6 月 30 日。在這期間凡是經過確認已經藉由血液途徑感染病毒，並且在觀察中止時間前產生AIDS病徵者皆列入觀察對象。其餘已經確認遭到病毒感染，但是於觀察時間結束前尚未產生AIDS病徵者已以剔除。期觀察紀錄形式如下（僅爲部分資料）：

表格 1 以觀察期間進行 AIDS 之研究，單位（年）

Infection time (感染期)[1]	Adult induction time (誘導期)[2]	Children induction time
0.00	5	
0.25	6.75	
0.75	5, 5, 7.25	
1.00	4.25, 5.75, 6.25, 6.5	5.5
1.25	4, 4.25, 4.75, 5.75	
1.50	2.75, 3.75, 5, 5.5, 6.5	2.25
1.75	2.75, 3, 5.25, 5.25	
2.00	2.25, 3, 4, 4.5, 4.75.5, 5.25, 5.25, 5.5, 5.5, 6	
2.25	3, 5.5	3
2.50	2.25, 2.25, 2.25, 2.25, 2.5, 2.75, 3, 3.25, 3.25, 4, 4, 4	

醫學上常使用這種部分時間的觀察資料來進行母體特徵的推估。以本例而言，在觀察時間區間之內沒有 AIDS 病徵者的樣本將被剔除。這樣子才能節省實驗的研究時間。在母體行爲推論上，雖然有部分的抽樣樣本被刪除，但推估整體母體的行爲時，還是必須包含這些已經被剔除的樣本範圍。

[1] 傳染期(infection time)：病原侵入至宿主體內的時間

[2]誘導期(induction time)：病原侵入至宿主產生臨床症狀或徵候的期間.

3. 截尾分佈之介紹

本研究中的語者確認模型與時間尚無關係，固本研究以資料量的出現與否?作爲遮罩資料的處理前提。本研究將遮罩資料視爲是原來的 PDF 函數進行部份截取(truncation)的結果。本研究將這種原 PDF 經過截取之後新形成的 PDF 函數稱爲 truncated probability distribution function (TPDF). TPDF 的分類上大概有三種形式，以下以常態分佈(normal distribution)爲例進行介紹：

常態分佈函數

$$f(x_i;u,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-(x_i-u)^2}{\sigma^2}),-\infty<x_i<\infty \qquad (3.1)$$

x_i : random variable of index i

u:mean

σ:standard deviation

一、單一左端點截取常態分佈(Singly left truncated normal distribution at $x_i=x_L$)

$$f_{Tr}(x_i;u,\sigma,x_L)=\frac{f(x_i;u,\sigma)}{1-F(x_L)}UnitStep(x_i-x_L) \qquad (3.2)$$

$$F(x_L)=\int_{-\infty}^{x_L}f(x_i;u,\sigma)dx_i$$

$UnitStep(.)$: unitstep function

二、單一右端點截取常態分佈(Singly right truncated normal distribution at $x_i=x_R$)

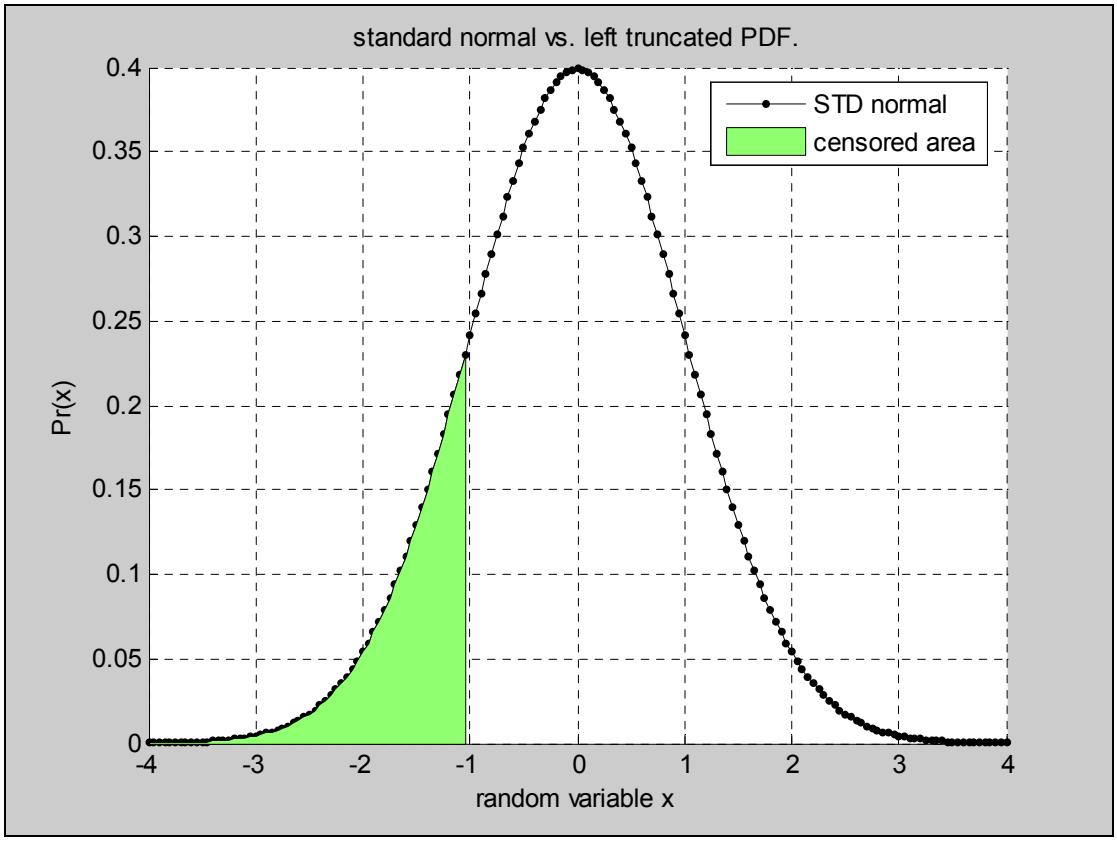
$$f_{Tr}(x_i;u,\sigma,x_R)=\frac{f(x_i;u,\sigma)}{F(x_R)}(1-UnitStep(x_i-x_R)) \qquad (3.3)$$

三、雙截尾常態分佈(Doubly truncated normal distribution at $x = x_L, x = x_R$)

$$f_{Tr}(x_i;u,\sigma,x_L,x_R) = \frac{f(x_i;u,\sigma)}{F(x_R)-F(x_L)} \times (UnitStep(x_i - x_L) - UnitStep(x_i - x_R)) \quad (3.4)$$

和截尾分佈相關聯的應用方式稱爲資料設限(data censoring)，例如因爲實驗的因素得到許多飽和裝況下的訊號成分，這時候可以考慮將高百分位數(percentile)的訊號成分移除之後進行樣本參數推估的方法。

(一)、 左端設限資料(left censored data)的機率密度函數



圖表 8 設限資料(censored data)機率密度函數和原機率密度函數之間的關係

單一左端點設限常態分佈(Singly left censored normal distribution at $x_i = x_L$)

這種狀況下樣本有兩種情形。一是x_i落在$(-\infty, x_L]$的區間內，但資料必須使用設限點(censoring point，在本例之中x_L=-1)來取代。另一個情況是

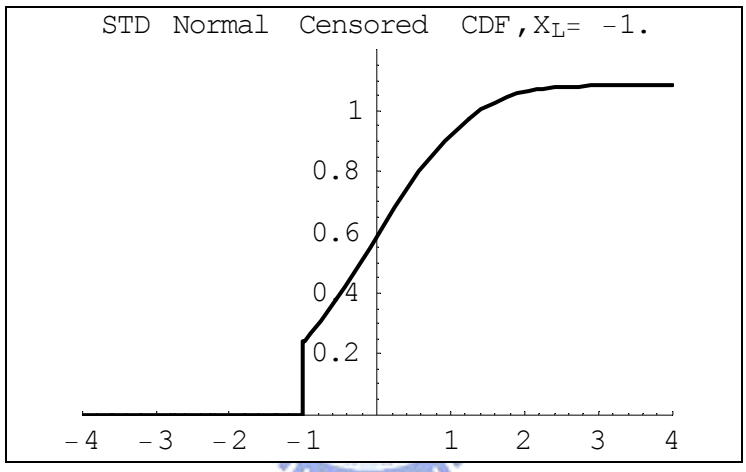
$x_i > x_L$，這時候將是正常的常態分佈情形。以下我們將這種情形使用數學的方程式描寫出來。

$$f_{Ce}(x_i;u,\sigma,x_L) = (F(x_L)UnitStep(x-x_L))^{1-g(x;x_L)} f(x_i;u,\sigma)^{g(x;x_L)} \quad (3.5)$$

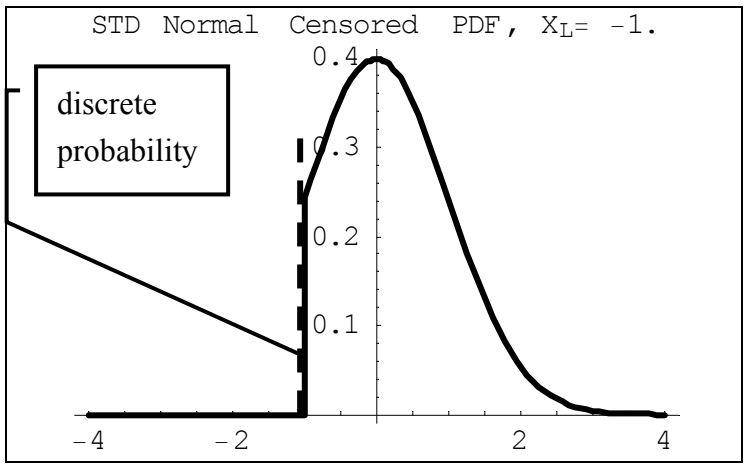
$$UnitStep(x-x_L) = \begin{cases} 1, if\ x \geq x_L \\ 0, \text{ otherwise} \end{cases}$$

$$g(x;x_L) = Boolean(x > x_L)$$

$$F(x_L) = \int_{-\infty}^{x_L} f(x_i;u,\sigma)dx_i$$



圖表 9 standard normal and its censored CDF



圖表 10 standard normal and its censored PDF

式子(3.5)是機率密度函數的表示式，因爲在左端點的位置有一個離散的機率密度函數起始值，所以累積機率密度函數應爲

$$F_{Ce}(x_i;u,\sigma,x_L) = \int_{-\infty}^{\infty} f_{Ce}(x_i;u,\sigma,x_L) + f_{Ce}(x_L;u,\sigma,x_L) \quad (3.6)$$

是一個連續機率加上另外一個離散的點機率所形成的累積機率密度函數。結果如圖表 9 所示。

(二)、 單一右端點設限常態分佈(Singly right censored normal distribution at $x_i = x_R$)

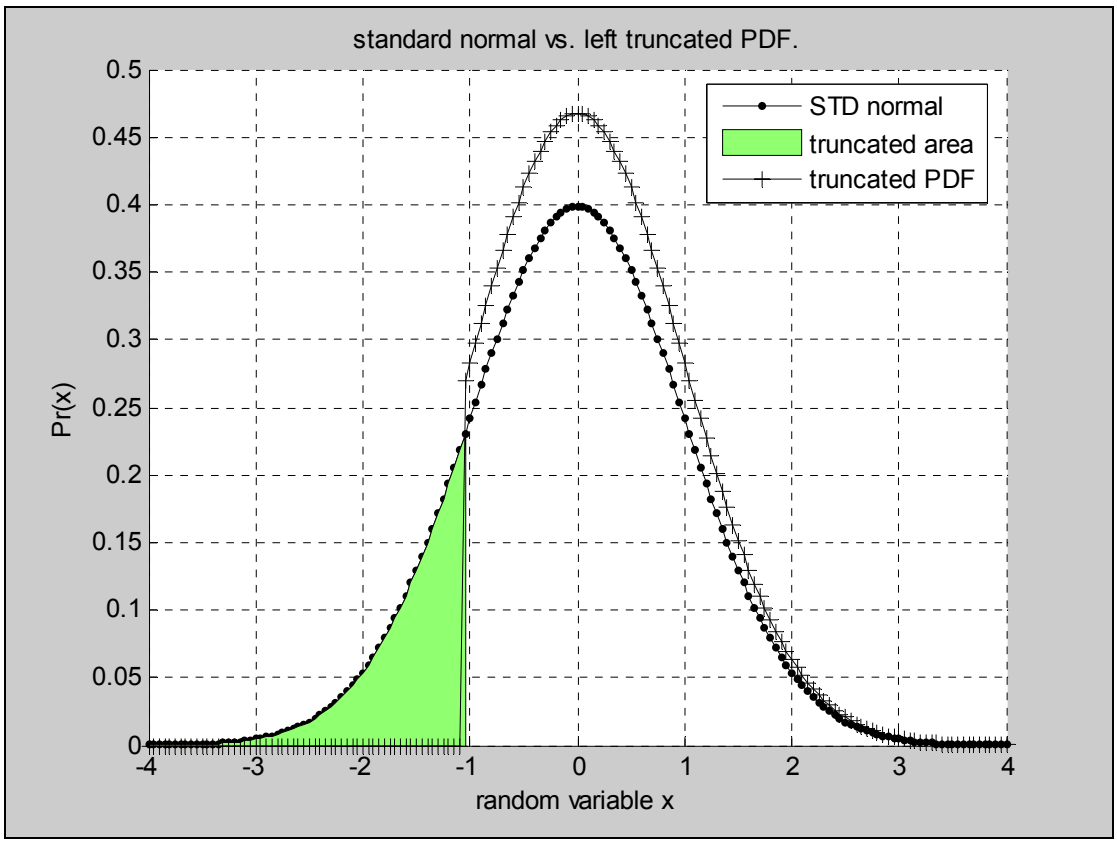
參考式子(3.5)，可以得出單一右端點設限常態分佈的機率密度函數

$$f_{Ce}(x_i;u,\sigma,x_R) = (1 - F(x_R))^j f(x_i;u,\sigma)^{1-j} \tag{3.7}$$

$$F(x_R) = \int_{-\infty}^{x_R} f(x_i;u,\sigma)dx_i$$

4. 截尾分佈之推導

4.1. 左截尾常態分佈之最大概度估計(Maximum Likelihood Estimators for Left Truncated Normal Distribution)



圖表 11 左截尾長態分佈之圖示

變數定義

N:全部樣本數量

n:有效的樣本數量

$\bar{x}$:有效樣本平均值

u:全體樣本平均值

x_L:左截斷點

$f(.)$: probability distribution function of normal distribution

$F(.)$: cumulative probability distribution function of normal distribution

$\phi(.)$: probability distribution function of standard normal distribution

$\Phi(.)$: cumulative probability distribution function of standard normal distribution

一個單一左截尾常態分佈可表示如下：

$$f(x_i;u,\sigma,x_L)=\frac{f(x_i;u,\sigma)}{1-F(x_L)}UnitStep(x_i-x_L) \tag{4.1}$$

$$f(x_i;u,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-(x_i-u)^2}{\sigma^2}\right),-\infty<x_i<\infty$$

$$F(x_L)=\int_{-\infty}^{x_L}f(x;u,\sigma)dx$$

$UnitStep(.)$: unitstep function

如果我們是由全部 N 個樣本，其中 c 個樣本被截取，n=N-c 個樣本抽樣所組成的概度函數則可表示成：

$$L(x;u,\sigma,x_L)=\prod_{i=1}^{n}\frac{f(x_i;u,\sigma)}{\left(1-F(x_L)\right)^n}UnitStep(x_i-x_L) \tag{4.2}$$

$$x=\{x_i\},i=1,2,\cdots n$$

對兩邊取對數，得出 log likelihood

$$\log\{L(x;u,\sigma,x_L)\}=-n\log(1-F(x_L))+\sum_{i=1}^{n}\log(f(x_i;u,\sigma)) \tag{4.3}$$

$$+\sum_{i=1}^{n}UnitStep(x_i-x_L)$$

$$\Rightarrow \log\{L(x;u,\sigma,x_L)\} = -n\log(1-F(x_L)) - \frac{n}{2}\log(2\pi\sigma^2)$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-u)^2 + \sum_{i=1}^{n} UnitStep(x_i - x_L) \tag{4.4}$$

將(4.4)式對 u 進行微分得出：

$$\frac{\partial L(x;u,\sigma,x_L)}{\partial u} = 0 \Rightarrow \frac{-n\cdot e^{\frac{-(x_L-u)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}\sigma}}{1-F(x_L)} + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-u) = 0 \tag{4.5}$$

令 $\tau = \frac{x_L - u}{\sigma}$ ， $\phi(\tau) = f(\tau;0,1)$
$\Phi(\tau) = F(\tau;0,1)$

則(4.5)可進行化簡爲

$$\frac{\partial L(x;u,\sigma,\tau)}{\partial u} = 0 \Rightarrow \frac{-n\phi(\tau)}{\sigma(1-\Phi(\tau))} + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-u) = 0 \tag{4.6}$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-u) = \frac{n\phi(\tau)}{\sigma(1-\Phi(\tau))}$$
$$Let \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \Rightarrow (\bar{x}-u) = \frac{\sigma\phi(\tau)}{(1-\Phi(\tau))} \tag{4.7}$$

同理，將 log likelihood 對 standard deviation σ 微分且令其等於零。

$$\Rightarrow \log\{L(x;u,\sigma,x_L)\} = -n\log(1-\Phi(\tau)) - \frac{n}{2}\log(2\pi\sigma^2)$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-u)^2 + \sum_{i=1}^{n} UnitStep(x_i - x_L) \tag{4.8}$$

$$\Rightarrow \frac{\partial}{\partial\sigma}\log\{L(x;u,\sigma,\tau)\} = \frac{-n\cdot(-\phi(\tau))(\frac{-\tau}{\sigma})}{1-\Phi(\tau)} - \frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-u)^2 \tag{4.9}$$

$$\Rightarrow \frac{\partial}{\partial \sigma}\log\{L(x;u,\sigma,\tau)\} = \frac{-n\cdot(\tau\cdot\phi(\tau))}{\sigma(1-\Phi(\tau))} - \frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-u)^2 = 0$$

$$\Rightarrow \frac{\tau\cdot\phi(\tau)}{\sigma(1-\Phi(\tau))} + \frac{1}{\sigma} = \frac{1}{n\sigma^3}\sum_{i=1}^{n}(x_i-u)^2$$

(4.10)

(4.10)繼續進行化簡

$$\frac{\tau\cdot\phi(\tau)}{(1-\Phi(\tau))} + 1 = \frac{1}{n\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-u)^2 \qquad (4.11)$$

其中(4.11)的右端可以進行展開化簡：

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

$$\Rightarrow \frac{1}{n\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2+(\bar{x}-u)^2+2(x_i-\bar{x})(\bar{x}-u) \qquad (4.12)$$

$$\Rightarrow \frac{1}{\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+(\bar{x}-u)^2+\frac{2}{n}\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-u)\right\}$$

$$\Rightarrow \frac{1}{\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+(\bar{x}-u)^2+\frac{2}{n}\sum_{i=1}^{n}x_i\bar{x}-x_iu-\bar{x}^2+\bar{x}u\right\}$$

$$\Rightarrow \frac{1}{\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+(\bar{x}-u)^2+2(\bar{x}^2-\bar{x}u-\bar{x}^2+\bar{x}u)\right\}$$

$$\Rightarrow \frac{1}{\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+(\bar{x}-u)^2\right\}$$

$$\Rightarrow \frac{1}{\sigma^2}\left\{S^2+(\bar{x}-u)^2\right\}$$

(4.13)

S^2 :sample variance

回到(4.11)式，得出$\frac{\partial L(x,u,\sigma,\tau)}{\partial\sigma}$的最簡化形式，

$$\frac{\tau\cdot\phi(\tau)}{(1-\Phi(\tau))}+1=\frac{1}{\sigma^2}\left\{S^2+(\bar{x}-u)^2\right\} \tag{4.14}$$

結合$\frac{\partial L(x,u,\sigma,\tau)}{\partial u}=0$的式子，$\Rightarrow(\bar{x}-u)=\frac{\sigma\phi(\tau)}{(1-\Phi(\tau))}$ (4.15)

(4.14)，(4.15)總共得出兩個式子，我們的主要目的是想要求出u,σ兩個未知數。所以聯立這兩個方程式是有機會求解出u,σ的值。不過主要的問題出在有三個未知數，$\tau=\frac{x_L-u}{\sigma}$在這個聯立方成組之中也是未知數。所以在此處，我們必須進行某種假設條件才可以求解u,σ。

- 再將(4.15)的結果代入(4.14)中，(4.14)可以在化簡成爲

$$\Rightarrow S^2=\sigma^2\left(-\left\{\frac{\phi(\tau)}{(1-\Phi(\tau))}\right\}^2+\frac{\tau\cdot\phi(\tau)}{(1-\Phi(\tau))}+1\right) \tag{4.16}$$

(4.15)也可以先進行替代化簡，

$$\begin{aligned}&\Rightarrow(\bar{x}-u)=\frac{\sigma\phi(\tau)}{(1-\Phi(\tau))}\\&\Rightarrow(\bar{x}-x_L+x_L-u)=\frac{\sigma\phi(\tau)}{(1-\Phi(\tau))}\end{aligned} \tag{4.17}$$

$$\begin{aligned}&\Rightarrow(\bar{x}-x_L)=\frac{\sigma\phi(\tau)}{(1-\Phi(\tau))}-\sigma\left(\frac{x_L-u}{\sigma}\right)\\&\Rightarrow(\bar{x}-x_L)=\sigma\left(\frac{\phi(\tau)}{(1-\Phi(\tau))}-\tau\right)\end{aligned} \tag{4.18}$$

- 現在如果假設標準差σ爲已知，(4.16)，(4.18)再共同成爲新的聯立方程組，標準差在合併的過程之中可以約分。如此可以得出一個變係數的方程式，使用最佳數值解來近似。

$$\Rightarrow\alpha(\tau)\triangleq\frac{S^2}{(\bar{x}-x_L)^2}=\frac{\left(-\left\{\frac{\phi(\tau)}{(1-\Phi(\tau))}\right\}^2+\frac{\tau\cdot\phi(\tau)}{(1-\Phi(\tau))}+1\right)}{\left(\frac{\phi(\tau)}{(1-\Phi(\tau))}-\tau\right)^2} \tag{4.19}$$

- (4.19)式的左端 $\frac{S^2}{(\bar{x}-x_L)^2}$ 都是可計算的已知數，右端的式子裡頭只剩下一個變數 τ ，於是求解這個變係數的方程式可以得出 τ 的最佳值。
- 解出 τ 之後，再代回(4.18)式，可以得出標準差 σ
- 最後利用定義式 $\tau=\frac{x_L-u}{\sigma}$ ，就可以估計出來，在完全樣本的考量之下，母體的最佳平均估計值 u 。

4.2. 右截尾常態分佈之最大概度估計(Maximum Likelihood Estimators for Right Truncated Normal Distribution)

變數定義

N: 全部樣本數量

n: 有效的樣本數量

$\bar{x}$: 有效樣本平均值

u: 全體樣本平均值

x_R : 右截斷點

$x=\{x_i\}, i=1,2\cdots n$

S^2 : 有效樣本之變異數

$f(.)$: 常態分佈機率密度函數

$F(.)$: 常態分佈累積機率函數

$\phi(.)$: 標準常態分佈機率密度函數

$\Phi(.)$: 標準常態分佈累積機率函數

一個單一右截尾常態分佈可表示如下：

$$f(x_i;u,\sigma,x_R)=\frac{f(x_i;u,\sigma)}{F(x_R)}(1-UnitStep(x_i-x_R)) \qquad (4.20)$$

$$f(x_i;u,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-(x_i-u)^2}{\sigma^2}),-\infty<x_i<\infty$$

$$F(x_R)=\int_{-\infty}^{x_R} f(x;u,\sigma)dx$$

$UnitStep(.)$: unitstep function

$$L(.)\Rightarrow L(x;u,\sigma,x_R)=\prod_{i=1}^{n}\frac{f(x_i;u,\sigma)}{F(x_R)^n}(1-UnitStep(x_i-x_R)) \tag{4.21}$$

對兩邊取對數，得出 log likelihood

$$\log\{L(x;u,\sigma,x_R)\}=-n\log(F(x_R))+\sum_{i=1}^{n}\log(f(x_i;u,\sigma))$$
$$+\sum_{i=1}^{n}(1-UnitStep(x_i-x_R)) \tag{4.22}$$

將(4.22)式對 u 進行微分得出：

$$\frac{\partial L(x;u,\sigma,\tau)}{\partial u}=0\Rightarrow\frac{-n\phi(\tau)}{\sigma(\Phi(\tau))}+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-u)=0 \tag{4.23}$$

$$\Rightarrow\frac{1}{n}\sum_{i=1}^{n}(x_i-u)=\frac{\sigma\phi(\tau)}{\Phi(\tau)}$$
$$Let\quad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\Rightarrow(\bar{x}-u)=\frac{\sigma\phi(\tau)}{\Phi(\tau)} \tag{4.24}$$

$$\Rightarrow(\bar{x}-u)=\frac{\sigma\phi(\tau)}{\Phi(\tau)} \tag{4.25}$$

同理，將 log likelihood 對 standard deviation σ 微分且令其等於零。

$$\Rightarrow \frac{\partial}{\partial \sigma}\log\{L(x;u,\sigma,\tau)\} = \frac{n(\tau\cdot\phi(\tau))}{\sigma\Phi(\tau)} - \frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-u)^2 = 0$$

$$\Rightarrow -\frac{n\cdot\tau\cdot\phi(\tau)}{\sigma\Phi(\tau)} + \frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-u)^2$$

(4.26)

$$-\frac{\tau\cdot\phi(\tau)}{\Phi(\tau)} + 1 = \frac{1}{n\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-u)^2 \tag{4.27}$$

回到(4.26)式，得出$\frac{\partial L(x,u,\sigma,\tau)}{\partial \sigma}$的最簡化形式，

$$-\frac{\tau\cdot\phi(\tau)}{\Phi(\tau)} + 1 = \frac{1}{\sigma^2}\left\{S^2 + (\bar{x}-u)^2\right\} \tag{4.28}$$

結合$\frac{\partial L(x,u,\sigma,\tau)}{\partial u} = 0$的式子，$\Rightarrow (\bar{x}-u) = \frac{\sigma\phi(\tau)}{\Phi(\tau)}$ (4.29)

$$\Rightarrow S^2 = \sigma^2\left(-\left\{\frac{\tau\cdot\phi(\tau)}{\Phi(\tau)}\right\}^2 - \frac{\tau\cdot\phi(\tau)}{\Phi(\tau)} + 1\right) \tag{4.30}$$

$$\Rightarrow (\bar{x}-u) = \frac{\sigma\phi(\tau)}{\Phi(\tau)}$$

$$\Rightarrow (\bar{x}-x_R+x_R-u) = \frac{\sigma\phi(\tau)}{\Phi(\tau)} \tag{4.31}$$

$$\Rightarrow (\bar{x}-x_R) = \frac{\sigma\phi(\tau)}{\Phi(\tau)} - \sigma\left(\frac{x_R-u}{\sigma}\right)$$

$$\Rightarrow (\bar{x}-x_R) = \sigma\left(\frac{\phi(\tau)}{\Phi(\tau)} - \tau\right) \tag{4.32}$$

$$\Rightarrow \alpha(\xi,\tau) \triangleq \frac{S^2}{(\bar{x}-x_R)^2} = \frac{\left(-\left\{\frac{\tau\cdot\phi(\tau)}{\Phi(\tau)}\right\}^2 + \frac{\tau\cdot\phi(\tau)}{\Phi(\tau)} + 1\right)}{\left(\frac{\phi(\tau)}{\Phi(\tau)} - \tau\right)^2} \quad (4.33)$$

- (4.33)式的左端 $\frac{S^2}{(\bar{x}-x_R)^2}$ 都是可計算的已知數，右端的式子裡頭只剩下一個變數 τ ，於是求解這個變係數的方程式可以得出 τ 的最佳值。
- 解出 τ 之後，再代回(4.32)式，可以得出標準差 σ
- 最後利用定義式 $\tau = \frac{x_L - u}{\sigma}$ ，就可以估計出來，在完全樣本的考量之下，母體的最佳平均估計值。

4.3. 雙截尾常態分佈之最大概度估計(Maximum Likelihood Estimation for Doubly Truncated Normal Distribution)

變數定義

x_L:左截斷點

n:有效的樣本數量

$\bar{x}$:有效樣本平均值

u:全體樣本平均值

x_R:右截斷點

$x = \{x_i\}, i = 1,2,\cdots n$

σ:全體標準差

S^2:有效樣本之變異數

$f(.)$:常態分佈機率密度函數

$F(.)$:常態分佈累積機率函數

$\phi(.)$:標準常態分佈機率密度函數

$\Phi(.)$:標準常態分佈累積機率函數

■ 機率密度函數

$$f_{Tr}(x;u,\sigma,x_L,x_R)=\frac{f(x;u,\sigma)}{(F(x_R)-F(x_L))}\times \\ (UnitStep(x_i-x_L)-UnitStep(x_i-x_R)) \tag{4.34}$$

■ 最大概度函數

$$Log\{L(f_{Tr}(x;u,\sigma,x_L,x_R))\}=-n\log(\Phi(\tau_R)-\Phi(\tau_L))-\frac{n}{2}\log(2\pi\sigma^2) \\ -\sum_{i=1}^{n}\frac{(x_i-u)^2}{2\sigma^2}+\log\{\sum_{i=1}^{n}(UnitStep(x_i-x_L)-UnitStep(x_i-x_R))\} \tag{4.35}$$

對於(4.35)分別對u,σ微分，並且令其等於零。

$$\frac{\partial}{\partial u}\{log(f_{Tr}(x;u,\sigma,x_L,x_R))\}=0 \\ \Rightarrow\frac{n(\phi(\tau_L)-\phi(\tau_R))}{\sigma(\Phi(\tau_R)-\Phi(\tau_L))}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-u) \tag{4.36}$$

令$\Theta_L=\frac{\phi(\tau_L)}{\Phi(\tau_R)-\Phi(\tau_L)},\Theta_R=\frac{\phi(\tau_R)}{\Phi(\tau_R)-\Phi(\tau_L)}$

則(4.36)可以整理出比較簡單的樣子。

$$\bar{x}-u=\sigma(\Theta_L-\Theta_R) \tag{4.37}$$

對σ微分

$$\frac{\partial}{\partial\sigma}\{\log(f(x;u,\sigma,x_L,x_R))\}=0 \\ \Rightarrow\frac{-n(\tau_L\phi(\tau_L)-\tau_R\phi(\tau_R))}{\sigma(\Phi(\tau_R)-\Phi(\tau_L))}-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-u)^2=0 \tag{4.38}$$

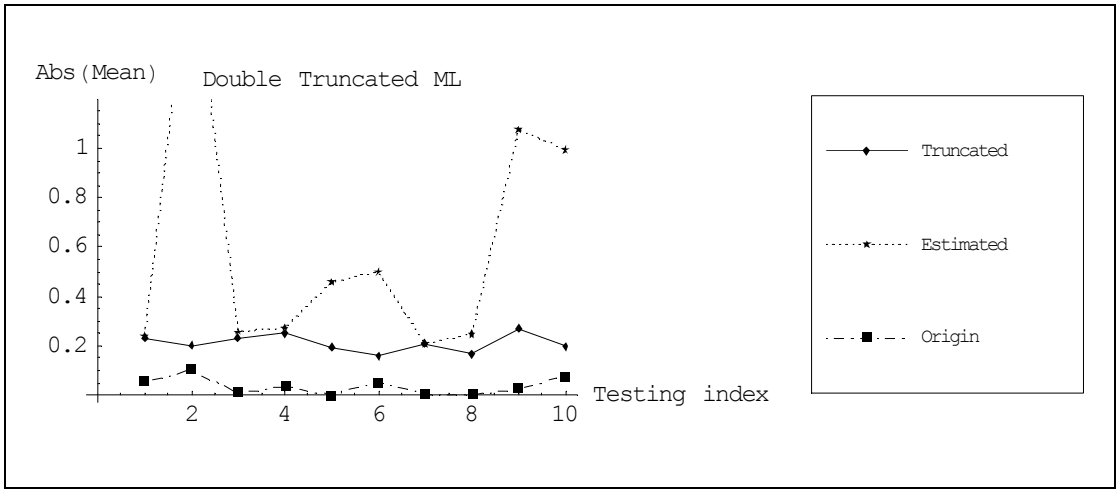
$$\sigma^2\left\{\frac{(\tau_L\phi(\tau_L)-\tau_R\phi(\tau_R))}{(\Phi(\tau_R)-\Phi(\tau_L))}+1\right\}=\frac{1}{n}\sum_{i=1}^{n}(x_i-u)^2$$

$$\sigma^2\left\{\tau_L\Theta_L-\tau_R\Theta_R+1\right\}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{1}{n}\sum_{i=1}^{n}(\bar{x}-u)^2 \qquad (4.39)$$

$$=S^2+(\bar{x}-u)^2$$

(4.39)式子中，令$\Theta_L=\frac{\phi(\tau_L)}{\Phi(\tau_R)-\Phi(\tau_L)},\Theta_R=\frac{\phi(\tau_R)}{\Phi(\tau_R)-\Phi(\tau_L)}$可以得出上述的結果。

這裡的$(\bar{x}-u)^2$可以利用(4.37)所推導的結果進行疊代之後得出最後的結果。



圖表 12 雙截尾常態分佈之最大概度母體平均數推估一

Truncated: 經過截尾刪除後的樣本資料

Estimated:使用 (4.39)的最大概度所推估出來的估計值

Origin:原始未經過截尾刪除處理的樣本資料

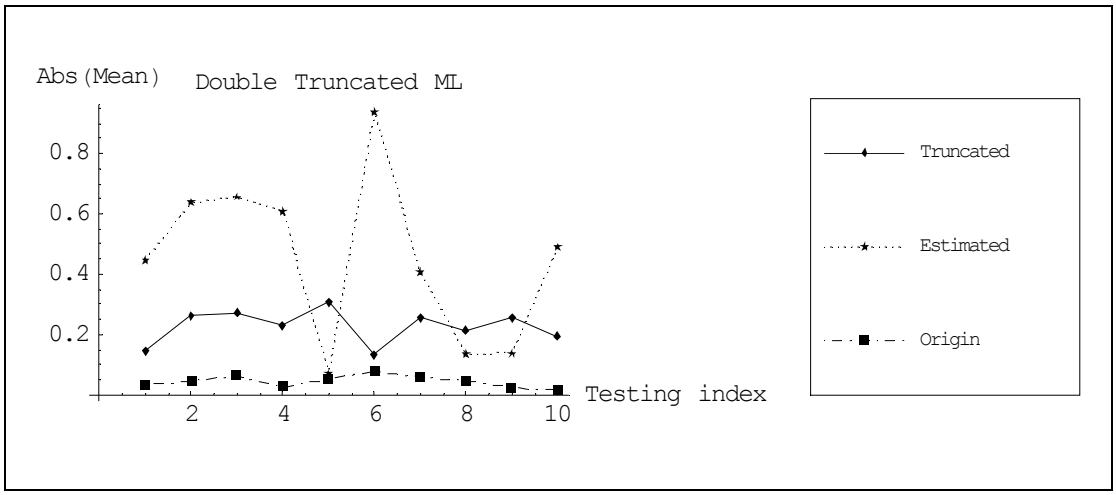
左截斷點：-1.4

右截斷點：0.8

每次使用未受截尾處理的全部樣本數：25， 每次計算一個估計值。

將每 15 次的推估結果平均之後作爲一次測試(Testing)的觀察值，總共進行 10 次測試。

圖表 12 是推估由 standard normal distribution N(0,1)所產生的樣本，經過最大概度函數估計之後的結果。



圖表 13 雙截尾常態分佈之最大概度母體平均數推估二

小結：雙截尾分佈的結果並不理想，單截尾分佈的結果較好。

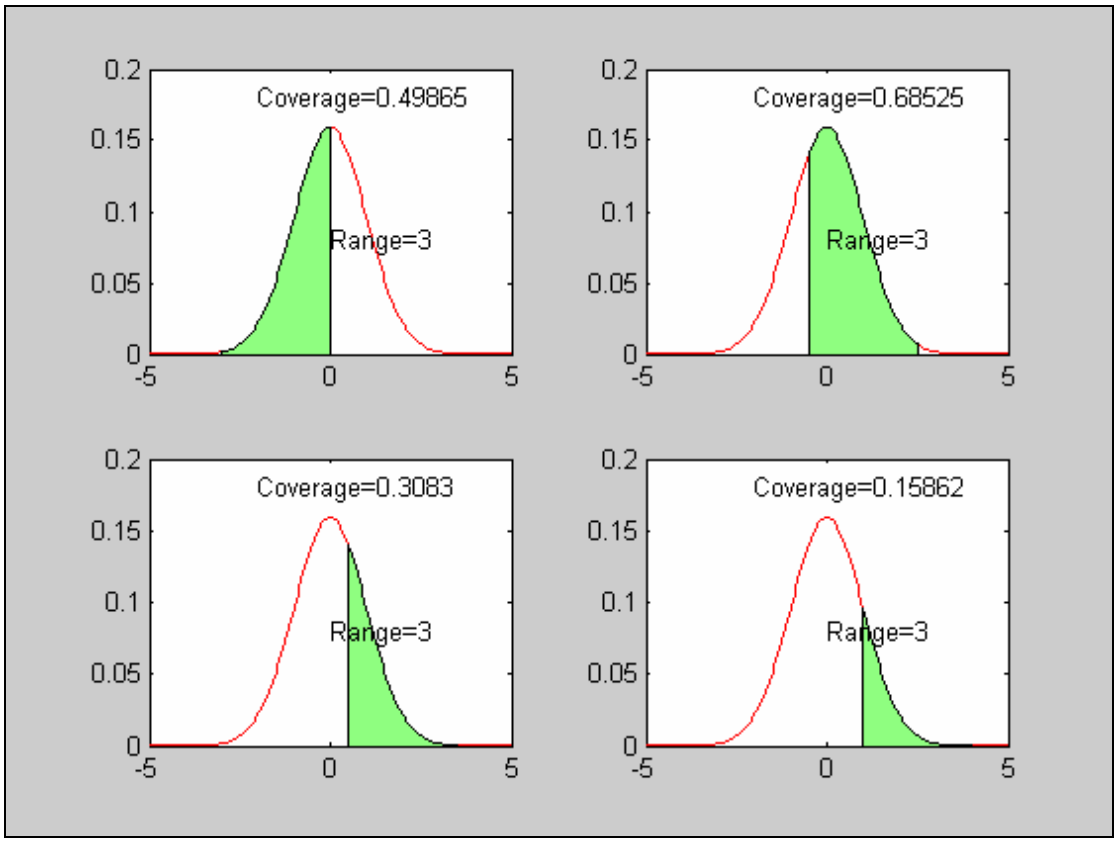
5. 模式建立

5.1. 模型定義

全距(range)： 在固定的抽樣數量之下，所得出的抽樣結果經過排序之後由最小值到最大值之間的區間範圍。

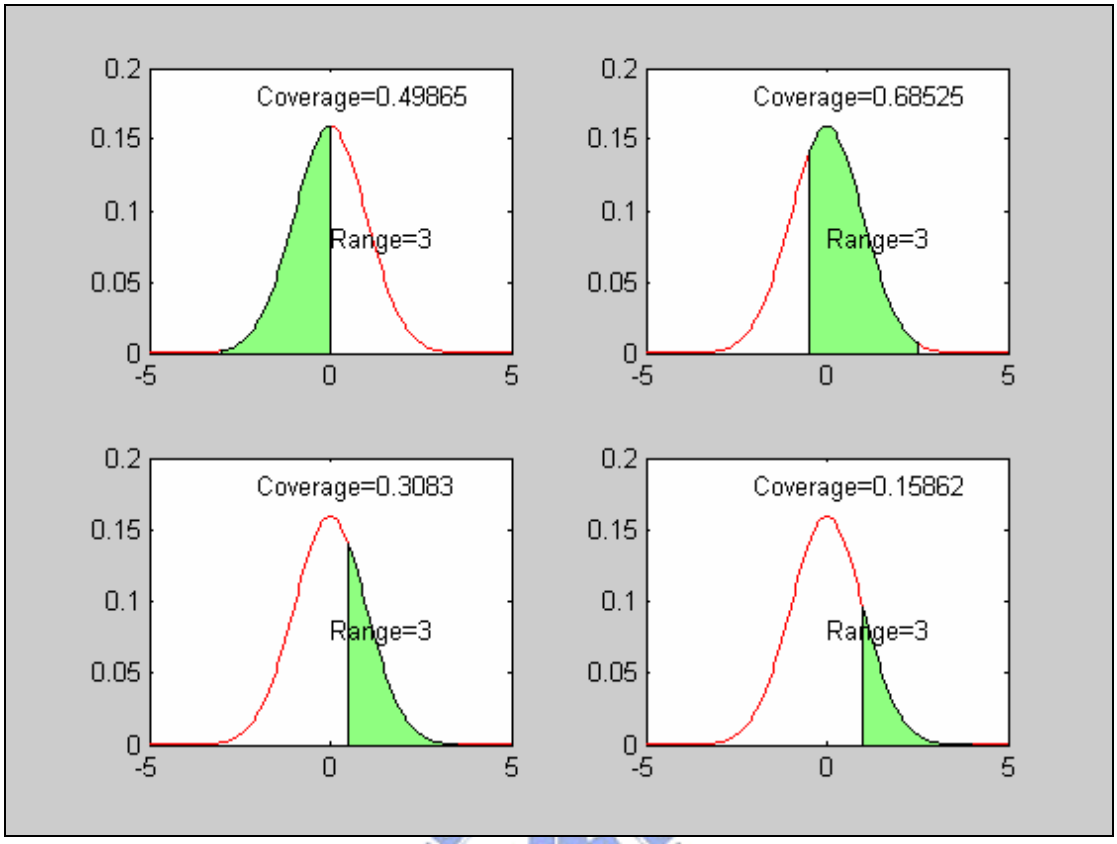
覆蓋率(coverage)：最小值到最大值之間的累積機率(cumulative probability)

5.2. 覆蓋率之實例解釋

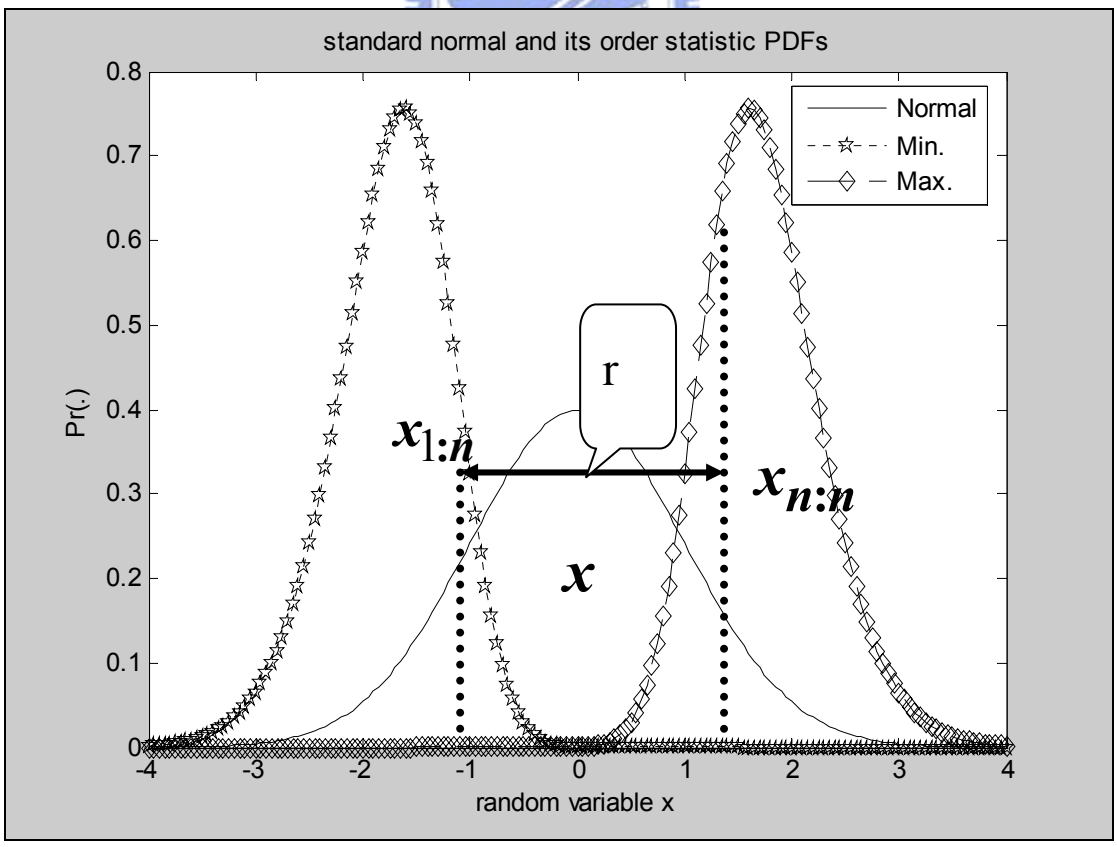


圖表 14 相同全距之下所得到之不同覆蓋率結果 一

如果對於一個母體取樣不足得時候，我們可能會在相同的全距(range)之下，看到不同的覆蓋率(coverage)。



圖表 15 相同全距之下所得到之不同覆蓋率結果二



圖表 16 Standard normal distribution and its minimum order and maximum order distribution

圖表 16 是本研究將一個基本的機率密度函數 PDF 和它的相關次序統計量的隨機變數，以及全距(range)，覆蓋率(coverage)的相互之間關係圖示說明。這個說明圖將一直沿用到本論文的結尾。本研究的最終目的就是將這些變數寫成聯合機率密度函數，用來描述稀少資料的隨機行爲。

變數定義：

$\boldsymbol{x}$: 常態分佈下的隨機變數

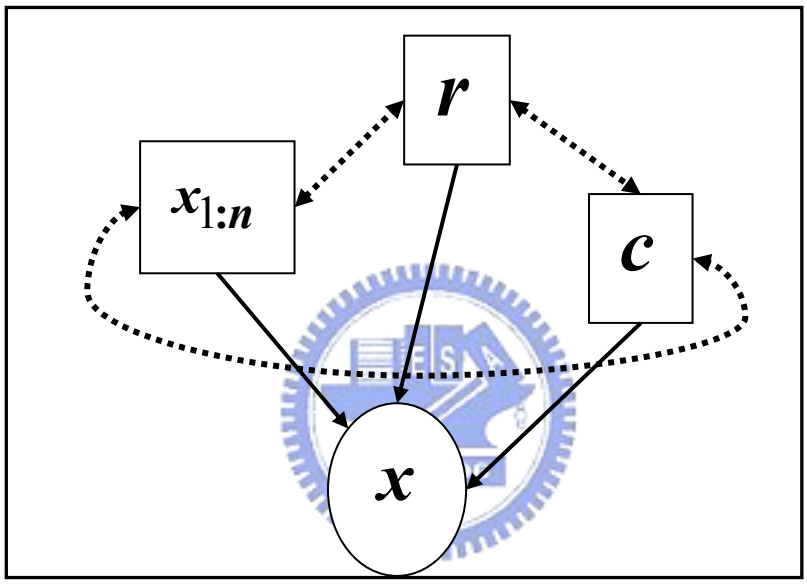
$\boldsymbol{x}_{1:n}$ 最小次序隨機變數(minimum order random variable): 從 $\boldsymbol{x}$ 所服從的常態分佈中，隨機抽取 n 個樣本。經過排序後，最小次序的隨機變數。

$\boldsymbol{x}_{n:n}$ 最大次序隨機變數(maximum order random variable):從 $\boldsymbol{x}$ 所服從的常態分佈中，隨機抽取 n 個樣本。經過排序後，最大次序的隨機變數。

$\boldsymbol{r}$ 全距(range)，$\boldsymbol{r} = \boldsymbol{x}_{n:n} - \boldsymbol{x}_{1:n}$：任意對於隨機變數 $\boldsymbol{x}$ 取 n 個樣本，此 n 個樣本最大值和最小值的差距。

$\boldsymbol{c}$ 覆蓋率(coverage), $\boldsymbol{c} = \int_{\boldsymbol{x}_{1:n}}^{\boldsymbol{x}_{n:n}} \boldsymbol{f}(\boldsymbol{x})d\boldsymbol{x}$，$\boldsymbol{f}(\boldsymbol{x})$ 爲 $\boldsymbol{x}$ 的機率密度函數(probability distribution function)，在本研究中爲常態分佈。

5.3. 聯合機率分佈函數(Joint Probability Distribution Function) $p(x, x_{1:n}, r, c|n)$ 之模型假設與推導



圖表 17 圖模式(graph model)下的變數關係

如果我們任意對於隨機變數 x 取了 n 個樣本，這些隨機變數的可能分佈範圍可以透過最小次序 $x_{1:n}$，全距, $r = x_{n:n} - x_{1:n}$ 和覆蓋率 $c = \int_{x_{1:n}}^{x_{n:n}} f(x)dx$ 來作更精確的描述。另外 $x_{1:n}, r, c$ 除了會影響 random variable x 之外，彼此之間也會產生交互影響。所以可以考慮使用聯合機率的方式來描述這四個變數之間的關係，另外一個相關的變數就是樣本大小 n，它直接影響其它四個變數。

實線箭頭部份表示參數 $x_{1:n}, r, c$ 對於隨機變數的影響，虛線雙箭頭表示參數 $x_{1:n}, r, c$ 彼此之間的交互影響效應。

■ 模式目的

➔建立一個評估樣本大小，樣本覆蓋率，樣本覆蓋區間大小，樣本覆蓋區間端點，樣本自變數四個變數之聯合機率密度函數

$$p(x, x_{1:n}, r, c \mid n) \tag{5.1}$$

對於任意的隨機變數 x 從常態分佈的機率密度函數中隨機抽取 n 個樣本時，我們可以假設它是個雙截尾常態分佈(doubly truncated normal distribution)。

雙截尾常態分佈

(Doubly truncated normal distribution at $x = x_{1:n}, x_{n:n} = x_{1:n} + r$)

$$f(x;u,\sigma,x_{1:n},x_{1:n}+r) = \frac{f(x;u,\sigma)}{F(x_{1:n}+r) - F(x_{1:n})} \times \\ (UnitStep(x - x_{1:n}) - UnitStep(x - x_{1:n} - r)) \tag{5.2}$$

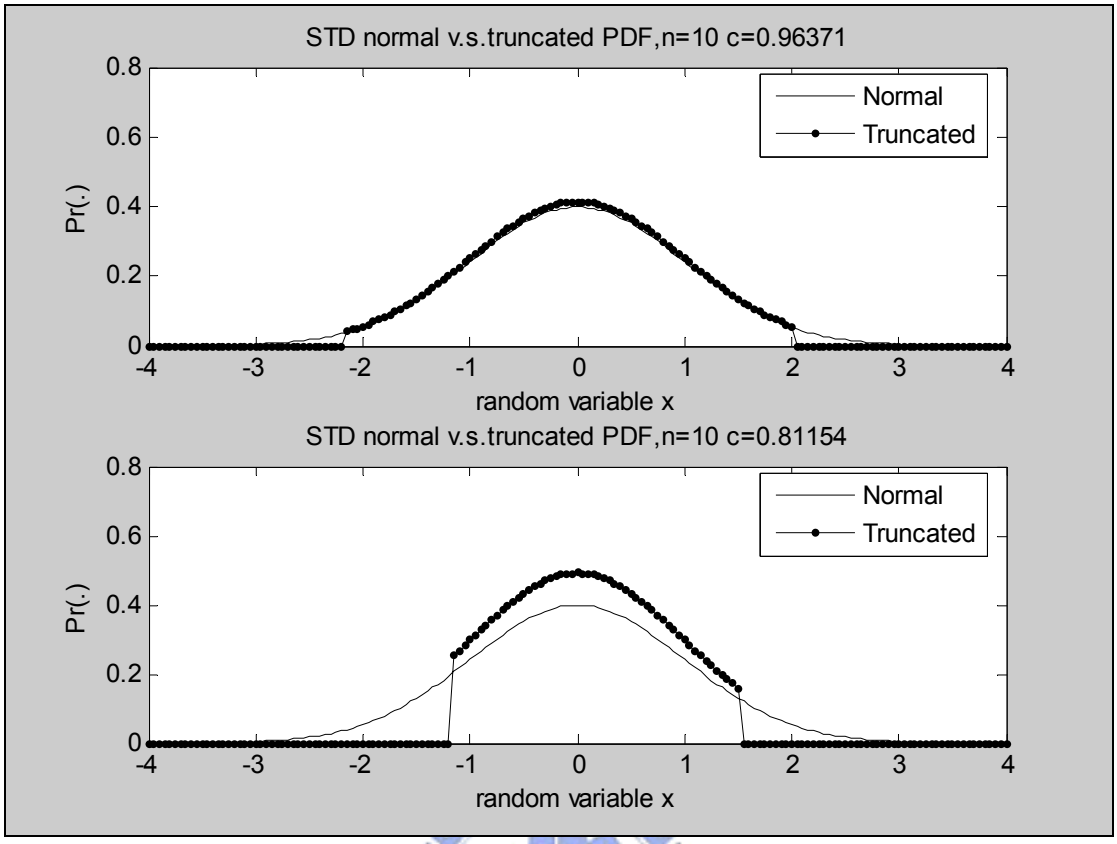
緊接著，令

$$p(x, x_{1:n}, r, c; u, \sigma \mid n) = p(x;u,\sigma \mid x_{1:n}, r, c, n) \times p(x_{1:n} \mid r, n) \times \\ p(r \mid c, n) \times p(c \mid n) \tag{5.3}$$

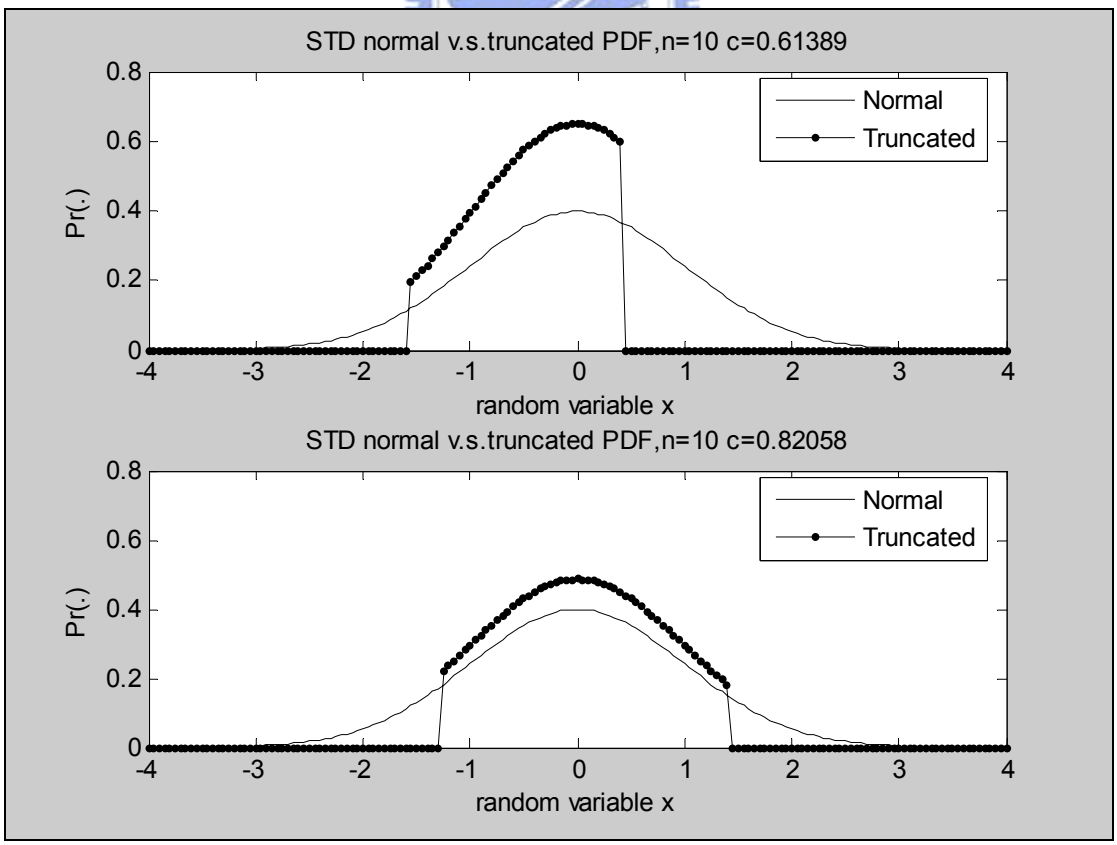
其中，$p(x;u,\sigma \mid x_{1:n}, r, c, n)$ 即為(5.2)。以下只要繼續求出 $p(x_{1:n} \mid r, n)$，$p(r \mid c, n)$ 和 $p(c \mid n)$ 即可完成。

➔使用截尾機率密度函數取代原來的全域機率密度函數，建立一個以覆蓋率為基礎的度量系統並且評估在此環境之下的可靠度。

如圖表 18，圖表 19，黑色粗線表示截尾分佈，細線代表原來的標準常態分佈。所謂的可靠度分析，就是我們想要使用能夠掌握的黑色粗線變化來取代細線，然後去測定黑色粗線受到覆蓋率減少的影響偏離原來的細線程度，和語者確認辨識率之間的相互關係。



圖表 18 標準常態分佈和其截尾分佈因爲覆蓋率所產生之落差(一）



圖表 19 標準常態分佈和其截尾分佈因爲覆蓋率所產生之落差(二)

次序統計量 (Order Statistic)提供了一個可以估計全距(range)和覆蓋率(coverage)的演算方法。

■ 次序統計量

假設我們自一個常態分佈中任意抽取 n 個相互獨立的隨機樣本。

$x_1, x_2, x_3, \ldots\ldots, x_n$ 。令 $x_{1:n}, x_{2:n}, x_{3:n}, \ldots\ldots, x_{n:n}$ 是這些隨機變數經過排序之後由小到大的樣本。這些經過排序之後的隨機變數，可以使用次序統計量(order statistic)來描述每一個變數的分佈情形。

定理一：來自同一個母體的 n 個經過排序之後的樣本。則排序位置由小算到大第 k 個位置的累積機率函數爲：

$$F_{k:n}(x_{k:n}) = p(x_{k:n} \le x) \quad \text{at least k order statistics } x_{k:n} \le x \tag{5.4}$$

$$p(k \text{ or even more than k number } x_{i:n} \in (-\infty, x]) \tag{5.5}$$

$$= \sum_{j=k}^{n} C_j^n F(x)^j \left(1 - F(x)\right)^{n-j} \tag{5.6}$$

$x_{k:n}$：第 k 個次序統計量的隨機變數

$F_{k:n}(x_{k:n})$：第 k 個次序統計量的累積機率函數

$$C_j^n = \frac{n!}{j!(n-j)!}$$

第 k 個次序統計量的機率密度函數爲

$$f_{k:n}(x) = \frac{dF_{k:n}(x)}{dx} \Rightarrow \frac{n!}{(k-1)!(n-k)!} f(x) \cdot F(x)^{k-1} \cdot \left(1 - F(x)\right)^{n-k} \tag{5.7}$$

$$f_{k:n}(x)=\frac{dF_{k:n}(x)}{dx}$$

$$=\sum_{j=k}^{n} C_j^n \cdot j \cdot F(x)^{j-1} f(x)\left(1-F(x)\right)^{n-j}+\sum_{j=k}^{n} C_j^n \cdot F(x)^j (n-j)(1-F(z))^{n-j-1}\left(-f(x)\right)$$

$$\Rightarrow \sum_{j=k}^{n} \frac{n!}{j!(n-j)!}\cdot j \cdot f(x)\cdot F(x)^{j-1}\left(1-F(x)\right)^{n-j}$$

$$-\sum_{j=k}^{n} \frac{n!}{j!(n-j)!}\cdot (n-j)\cdot f(x)\cdot F(x)^{j}\left(1-F(x)\right)^{n-j-1}$$

$$\Rightarrow \sum_{j=k}^{n} \frac{n!}{(j-1)!\cdot(n-j)!} f(x)\cdot F(x)^{j-1}\left(1-F(x)\right)^{n-j}$$

$$-\sum_{m=k+1}^{n} \frac{n!}{(m-1)!(n-m)!} f(x)\cdot F(x)^{m-1}\left(1-F(x)\right)^{n-m}$$

$$=\frac{n!}{(k-1)!\left(n-k\right)!} f(x)\cdot F(x)^{k-1}\cdot\left(1-F(x)\right)^{n-k}$$

(5.8)

根據(5.8)式我們可以得出最小次序 $x_{1:n}$ 和最大次序 $x_{n:n}$ 的機率密度函數：

$$f_{1:n}(x_{1:n})=n\cdot f(x_{1:n})\cdot(1-F(x_{1:n}))^{n-1}$$
$$f_{n:n}(x_{n:n})=n\cdot f(x_{n:n})\cdot(F(x_{n:n}))^{n-1} \qquad (5.9)$$

有了最小和最大次序分佈的機率密度函數之後，可以定義出全距(range)

r(range):全距隨機變數

$$\textit{Define}\ \mathrm{r}=x_{n:n}-x_{1:n}$$

$$P(x_{1:n}>x, x_{n:n}\le y)$$

$$=P(x<x_{i:n}\le y,\forall_i) \quad x_{1:n}, x_{2:n},\cdots,x_{n:n} \text{ is i.i.d}$$

$$=(F(y)-F(x))^n$$

$$P(x_{n:n}\le y)=P(x_{i:n}\le y,\forall_i)=F(y)^n$$

$$\Rightarrow F_{1,n:n}(x,y)=P(x_{1:n}\le x, x_{n:n}\le y)$$

$$=P(x_{1:n}\le x, x_{n:n}\le y)=P(x_{n:n}\le y)-P(x_{1:n}>x, x_{n:n}\le y)$$

$$\Rightarrow F(y)^n-(F(y)-F(x))^n$$

如果每次的輸入序列，經過排序之後， $x_{n:n}$ 就是最大值， $x_{1:n}$ 就是最小值。所

以也可以直接將 y 用 $x_{n:n}$ 代替， x 用 $x_{1:n}$ 代替。此處將 $F_{1,n:n}(x,y)$ 記爲 $F_{1,n:n}(x_{1:n},x_{n:n})$ 便於統一符號名稱。

$$F_{1,n:n}(x_{1:n},x_{n:n}) = F(x_{n:n})^n - \left(F(x_{n:n}) - F(x_{1:n})\right)^n \qquad (5.10)$$

$$\Rightarrow f_{1,n:n}(x_{1:n},x_{n:n}) = \frac{\partial^2}{\partial x_{1:n}\partial x_{n:n}} F_{1,n:n}(x_{1:n},x_{n:n})$$

$$\Rightarrow n(n-1)\cdot f(x_{1:n})\cdot f(x_{n:n})(F(x_{n:n}) - F(x_{1:n}))^{n-2}$$

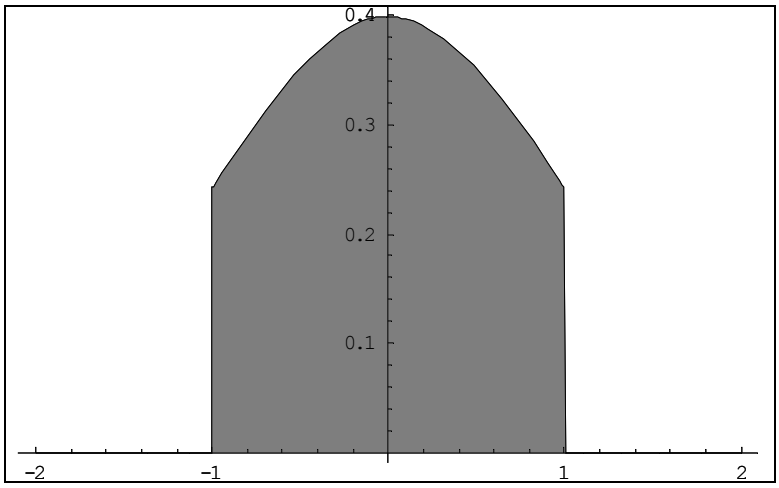
進行變數轉換， $r = x_{n:n} - x_{1:n}$

$$f_R(r) = \int_{-\infty}^{\infty} f_{x_1,n:n}(x_{1:n},r)dx_{1:n}$$
$$\int_{-\infty}^{\infty} n(n-1)\cdot f(x_{1:n})\cdot f(r+x_{1:n})(F(r+x_{1:n}) - F(x_{1:n}))^{n-2}dx_{1:n}, r>0 \qquad (5.11)$$

n :樣本大小
$f(x)$:機率密度函數
$F(x)$: 累積機率函數
r : 全距

(5.11)是全距 range 的公式，它和取樣的大小以及隨機變數的機率密度函數有關係。

如果在進行語者確認的過程之中，對於全體母體(population)的涵蓋率有興趣時，可以引入另外一個物理量 –覆蓋率(coverage)，coverage 所代表的意義是指 range 範圍之內的累積機率值。我們可以假想 coverage 是對應到某一位 speaker 的取樣覆蓋率，覆蓋率不夠充份時，可能很容易會和其他的僞裝者之間的特性產生混淆；如果覆蓋率無法符合預先界定的水準時，可以考慮捨棄該次所估計的統計參數或者使用補強的措施來增加統計參數的強健(robust)程度。

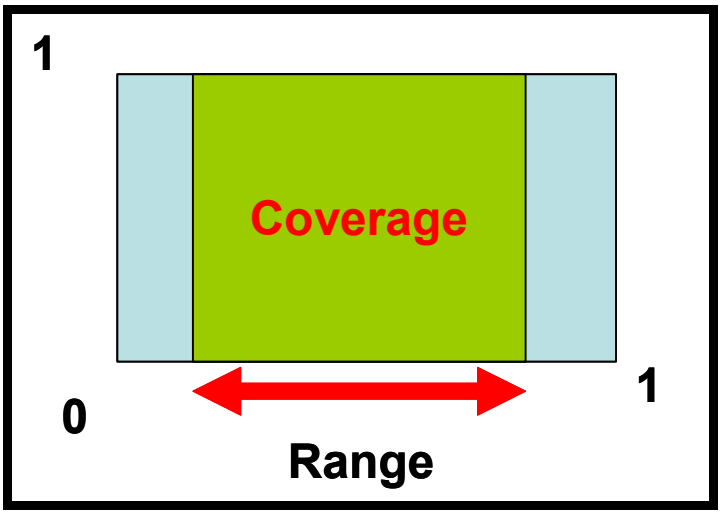


圖表 20 覆蓋率的物理意義

5.4. 覆蓋率之機率密度函數➔ $p(c \mid n)$ 之計算

$$p(x, x_{1:n}, r, c; u, \sigma \mid n) = p(x; u, \sigma \mid x_{1:n}, r, c, n) \times p(x_{1:n} \mid r, n) \times p(r \mid c, n) \times p(c \mid n) \quad (5.12)$$

如果我們進行變數轉換，令 $u_{x_{i:n}} = \int_{-\infty}^{x_{i:n}} f(x)dx, 1 \le i \le n$ ….. (5.13)。於是，我們將會得出一個新的 uniform distribution, $u_{x_{i:n}} \sim U[0,1]$，在這個分佈裡頭，將會出現一個特性，那就是 range 的值會等於 coverage；然而，任何的 PDF 經過上式(5.13)的轉換之後都會形成 $u_{x_{i:n}} \sim U[0,1]$ 的 PDF，也就是說；我們將可以獲得一個 distribution free 的通用公式來計算 coverage 的機率。



$$Define \triangleq \hat{r} = u_{x_{n:n}} - u_{x_{1:n}} \tag{5.14}$$

$$\begin{aligned} \hat{r} &= u_{x_{n:n}} - u_{x_{1:n}} \\ &\Rightarrow \int_{-\infty}^{x_{n:n}} f(x)dx - \int_{-\infty}^{x_{1:n}} f(x)dx \\ &= \int_{x_{1:n}}^{x_{n:n}} f(x)dx = c(r) \end{aligned} \tag{5.15}$$

$\hat{r}$:均等分佈 [0,1]的隨機變數因為抽樣所形成的全距隨機變數

$x_{i:n}$:次序統計量隨機變數

f(x): 隨機變數 x 的機率密度函數

c(r): 全距 r 所相對應的覆蓋率

原來，新的 range $\hat{r}$ 會等於原來的 coverage $\Rightarrow c(r)$

所以, $p_r(c(r)) = p_r(\hat{r})$,

從次序統計的理論中可以得出，任何分佈下，其抽樣大小爲 n 的樣本全距分佈爲下式:

$$\begin{aligned} &f_R(r) = \int_{-\infty}^{\infty} f_{x_{1:n},x_{n:n}}(x_{1:n}, x_{1:n} + r)dx_{1:n} \\ &\int_{-\infty}^{\infty} n(n-1) \cdot f(x_{1:n}) \cdot f(r + x_{1:n})(F(r + x_{1:n}) - F(x_{1:n}))^{n-2} dx_{1:n}, r > 0 \end{aligned}$$

(5.16)

n :樣本大小

$f(x)$:機率密度函數

$F(x)$: 累積機率函數

r: 全距

覆蓋率本身是全距的函數 c(r)，但是爲了方便起見，在往後的文章中以 c 作爲符號代表。(5.16) 實際上和樣本大小有關係，所以應該把它記爲：

$$p(r \mid n) = \int_{-\infty}^{\infty} n(n-1) f(x_{1:n}) f(x_{1:n} + r)\{F(x_{1:n} + r) - F(x_{1:n})\}^{n-2} dx_{1:n}$$

(5.17)

5.5. 使用均等分佈 U[0,1]下的全距分佈公式 $\hat{r}$ 作爲覆蓋率的機率密度函數

由上一節的分析得出結論，任何機率分佈下，因爲抽樣大小所產生的全距，其所對應的覆蓋率隨機變數；實際上對應到變數轉換之後的均等分佈[0,1]的全距分佈。在此，我們可以簡單的將均等分佈下的全距公式求出。根據次序統計量的推導結果，全距公式適用所有的分佈，所以均等分佈的全距公式如下：

$$p(\hat{r}) = \int_0^{1-\hat{r}} n(n-1) \cdot 1 \cdot 1 \left(u_{x_{1:n}} + \hat{r} - u_{x_{1:n}} \right)^{n-2} du_{x_{1:n}}$$

(5.18)

$$\Rightarrow n(n-1)\hat{r}^{n-2}(1-\hat{r}), \hat{r} > 0 \qquad (5.19)$$

均等分佈下的全距累積機率函數爲

$$F_R(\hat{r}) = \int p(\hat{r}) d\hat{r} = \hat{r}(n + (1-n)\hat{r}), \hat{r} > 0 \qquad (5.20)$$

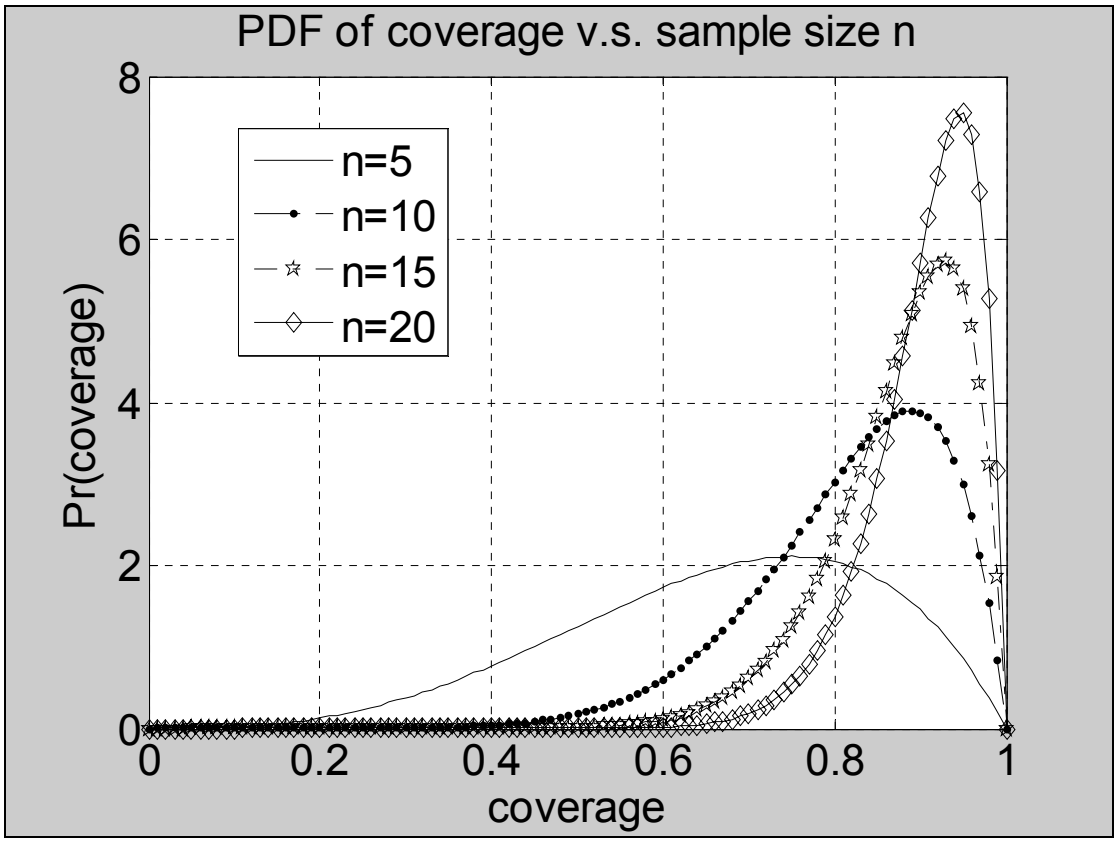
所以，原來的隨機變數的覆蓋率機率密度函數爲:

$$\boxed{\Rightarrow p(c) = n(n-1)\left(c\right)^{n-2}(1-c), c > 0} \qquad (5.21)$$

原來的隨機變數的覆蓋率累積機率函數爲:

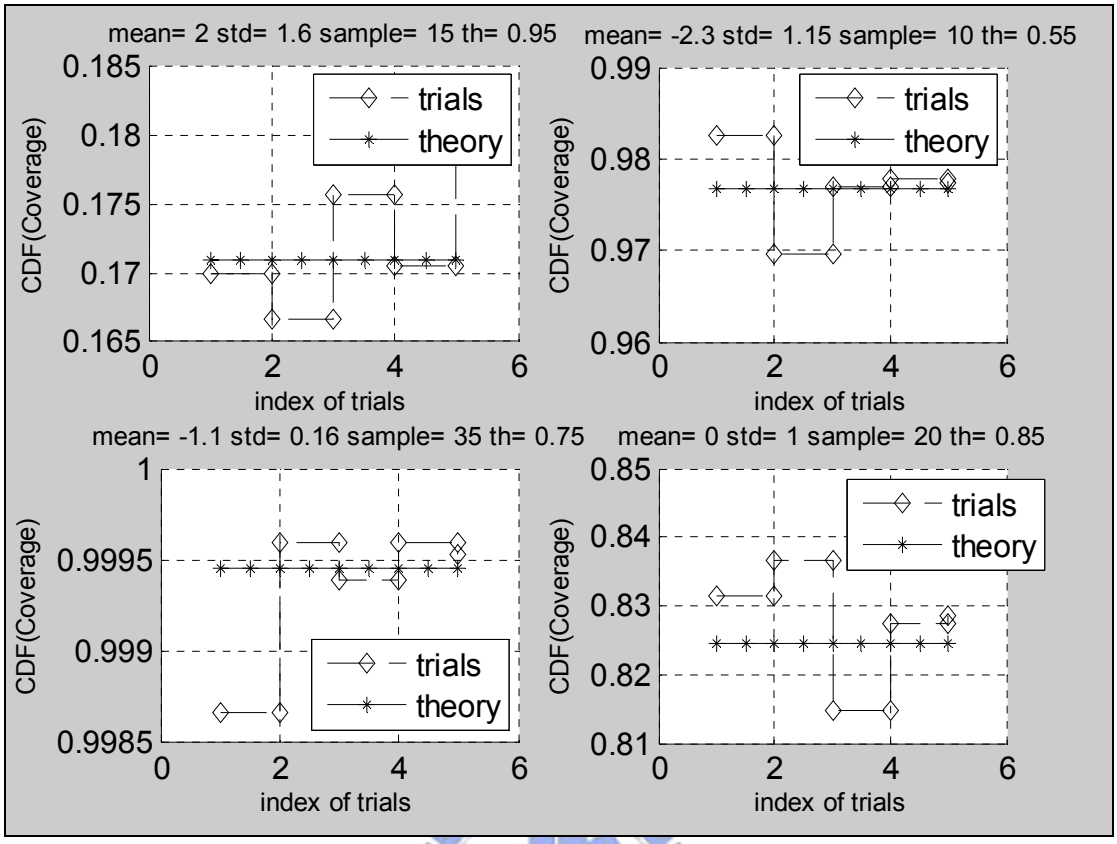
$$F(c) = \int p_r(c)dc = c^{n-1}(n + (1-n)c), c > 0 \qquad (5.22)$$

在這一節中是使用變數轉換的觀點來解決求算覆蓋率的機率密度函數。**所有的機率分佈都可以藉由累積機率的轉換式(5.13)將自身的變數轉換到均等分佈 U[0,1]之上。在 U[0,1]的分佈下，全距的值會等於覆蓋率的值；所以只要知道全距的值就等於知道了覆蓋率的值。恰巧的是，覆蓋率並沒有因爲變數轉換而改變（這是定積分的性質）**。所以在這裡經由變數轉換的技巧，得出了覆蓋率的機率密度函數型態。



圖表 21 不同的樣本數量下的覆蓋率(coverage) 機率分布

圖表 21 是將推導公式(5.21)直接作圖畫出，由結果的圖中可以發現到—當取樣的樣本數爲 20，覆蓋率幾乎都集中在 0.9 附近。也就間接說明，如果研究的只是一維的資料，覆蓋率不足的現象只會發生在樣本數小於 20 的情況。



圖表 22 覆蓋率公式之測試結果

圖表 22 是公式(5.22)的累積機率測試結果。四個測試樣本分別針對不同的平均值(mean)，標準差(std)，樣本數量(sample)，以及門檻值(th)進行測試。試行(trial)的次數大小分成五個種類，在圖中分別以橫軸的數字 1~5 代表。縱軸代表這些累積機率值的平均

$$\frac{1-F(th)}{N} \Rightarrow \frac{1-(th(n+(1-n)th))}{N} \tag{5.23}$$

N ：試行次數

n ：樣本大小

$F(.)$ ：覆蓋率之累積機率函數

Trial 所分成的五種不同數量所衍生的類別分類依序爲 $[30,50,100,200,300]$ 五種不同的試行次數，分別對應到橫軸座標的 $[1,2,3,4,5]$ 。

由結果可以看出來，本研究所推導的理論值和實際值非常吻合。

5.6. 條件機率 $p(r \mid c,n)$ 之推導

$$p(x,x_{1:n},r,c;u,\sigma \mid n) = p(x;u,\sigma \mid x_{1:n},r,c,n) \times p(x_{1:n} \mid r,n) \times p(r \mid c,n) \times p(c \mid n) \tag{5.24}$$

r：全距

c：覆蓋率

n：樣本大小

從次序統計量中，可以得出 $p(r \mid n)$，

$$p(r \mid n) = \int_{-\infty}^{\infty} n(n-1) f(x_{1:n}) f(x_{1:n}+r)\{F(x_{1:n}+r) - F(x_{1:n})\}^{n-2} dx_{1:n} \quad (5.25)$$

$$f(x;u,\sigma) \triangleq f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-u)^2}{2\sigma^2}}, PDF \text{ of normal distribution}$$

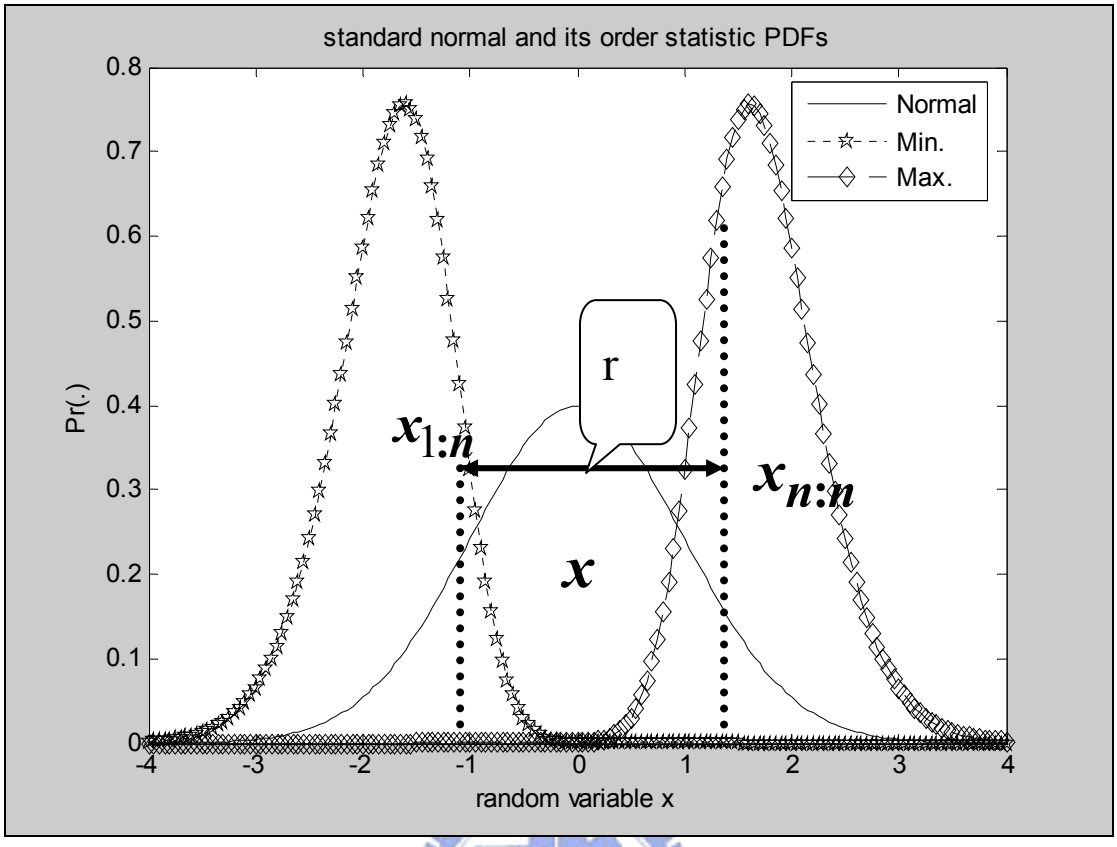
$$F(x;u,\sigma) \triangleq F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(t-u)^2}{2\sigma^2}} dt, CDF \text{ of normal distribution}$$

$x_{1:n}$：最小次序統計量，任取 n 個隨機標準常態分佈變數，經過排序之後的最小值。

r 全距(range):任取 n 個隨機標準常態分佈變數，最大值隨機變數 $x_{n:n}$ 和最小值隨機變數 $x_{1:n}$ 的差值。

c 覆蓋率(coverage)：➔全距範圍內的 PDF 積分值大小

n: 樣本數量大小



圖表 23 Standard normal distribution and its minimum order and maximum order distribution

現在要寫出 $p(r \mid c,n)$，它表示覆蓋率和樣本大小固定時所看到的全距 分佈情形。 仔細看(5.25)式，裡頭的 $F(x_{1:n}+r)-F(x_{1:n})$ 就是全距變數，只要讓他固定成某個觀察值，則 $p(r \mid c,n)$ 就成功了。

■ 條件機率計算剖析

首先，先觀察一般條件機率的變化情形。如圖表 24，二維之機率密度函數

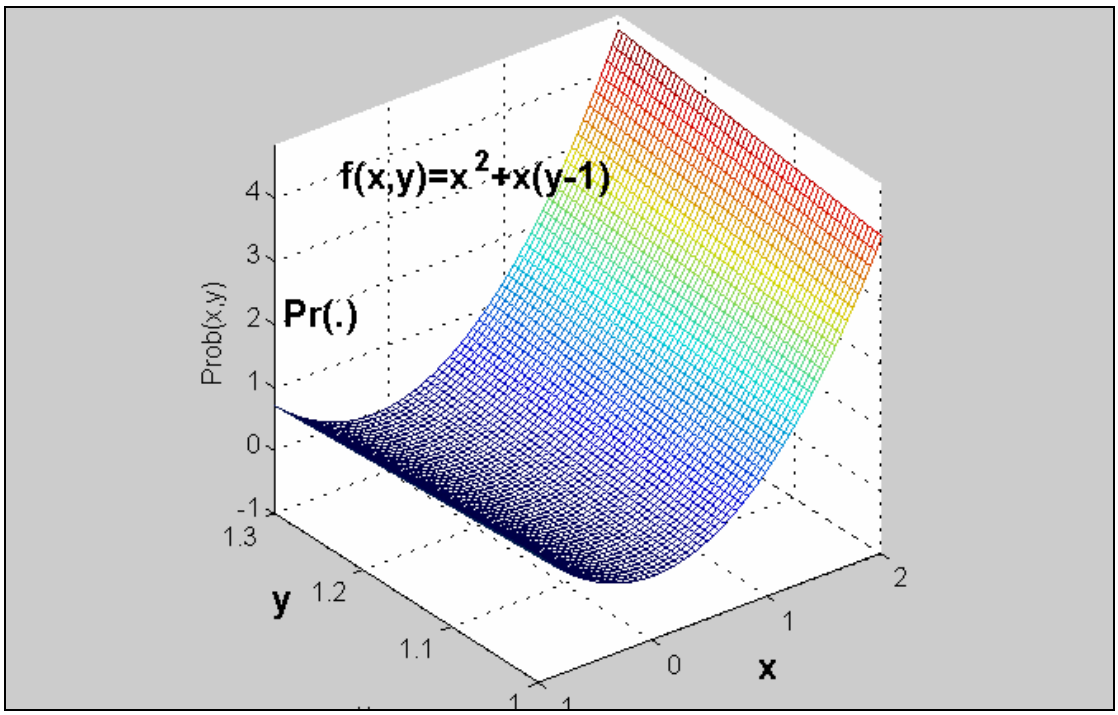
$$f(x,y)=x^2+x(y-1),\begin{cases}-1\le x\le 2\\ 1\le y\le 1.3\end{cases} \tag{5.26}$$

如果我們想要求取條件機率 $p(x \mid y)$，一般的作法是計算 $\dfrac{f(x,y)}{\int_{-1}^{2} f(x,y)dx}$，其結果等於

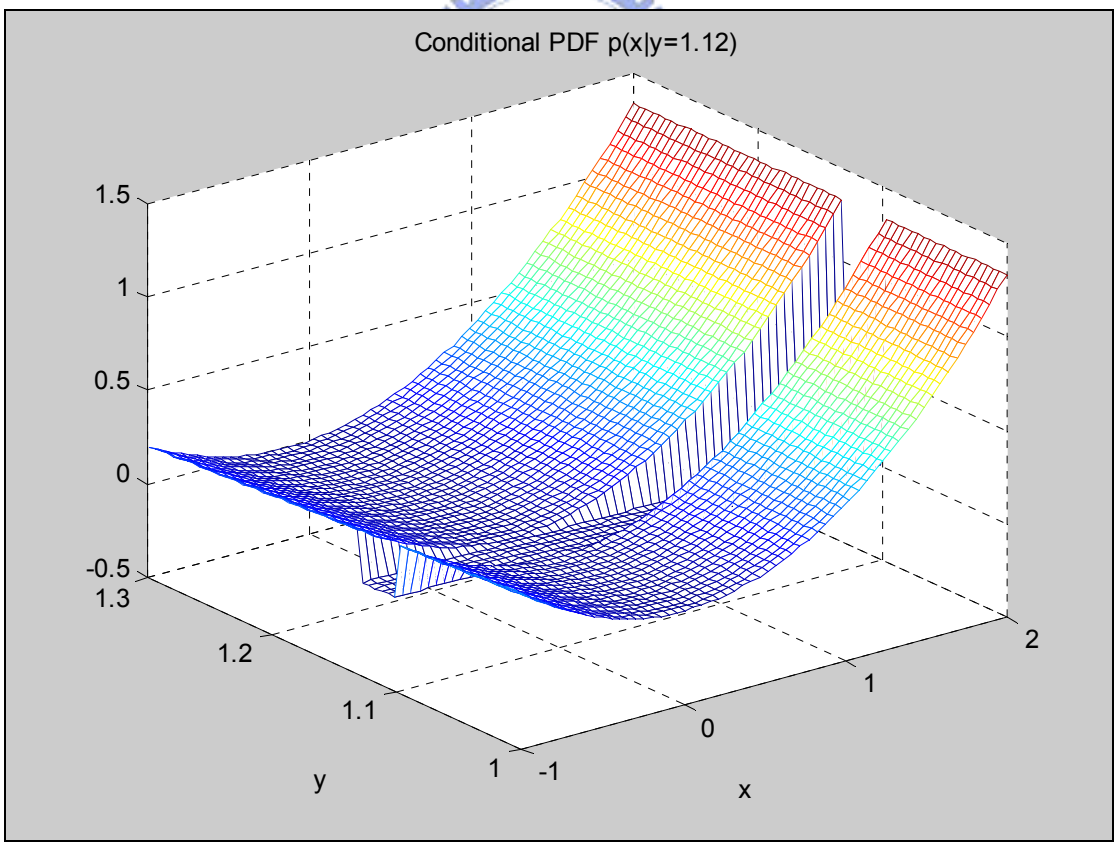
$$\frac{\frac{2}{3}x(x+y-1)}{y+1} \tag{5.27}$$

(5.27)條件機率 $p(x \mid y)$ 在 $f(x,y)$ domain 上的圖形如圖表 25 所示。圖表

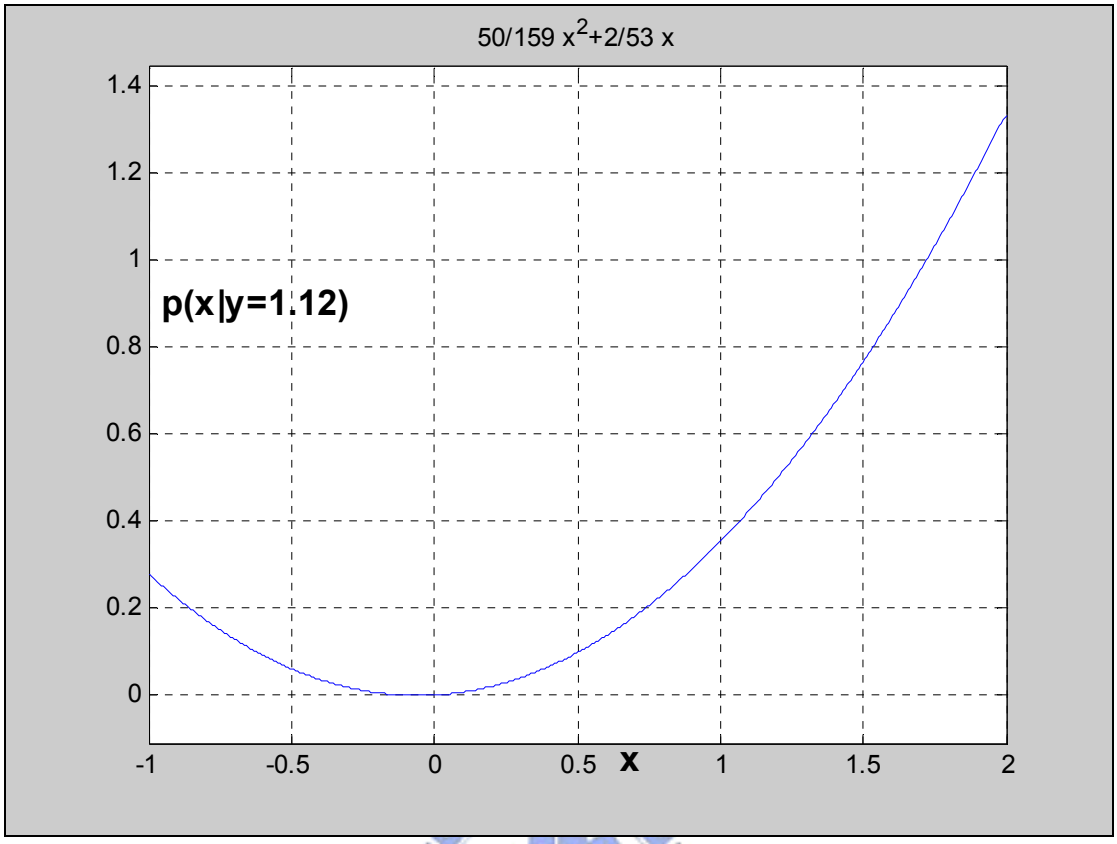
25 挖空的部份爲圖表 26 $p(x \mid y = 1.12)$之圖形。



圖表 24　x^2+x(y-1)之機率密度函數



圖表 25 p(x|y)之所有條件機率 p(x|Y=y)圖形



圖表 26 特殊情形 p(x|y=1.12)之 PDF 圖形

將圖表 26 $p(x \mid y=1.12)$ 之條件機率密度函數進行 x 之定義域積分

$$\int_{-1}^{2}\left(\frac{50}{159}x^2+\frac{2}{53}x\right)dx=1 \qquad (5.28)$$

由(5.28)可以知道，所謂的條件機率就是指條件值被設定爲常數時，其餘的式子在原來變數的定義域之內可以滿足機率密度函數的定義。

由以上的知識，(5.28)式中的條件機率我們可以使用另外一個角度來觀察。

一、令機率密度函數 $f(x,y)=x^2+x(y-1),\begin{cases}-1\le x\le 2\\ 1\le y\le 1.3\end{cases}$ 式中的 y=1.12。

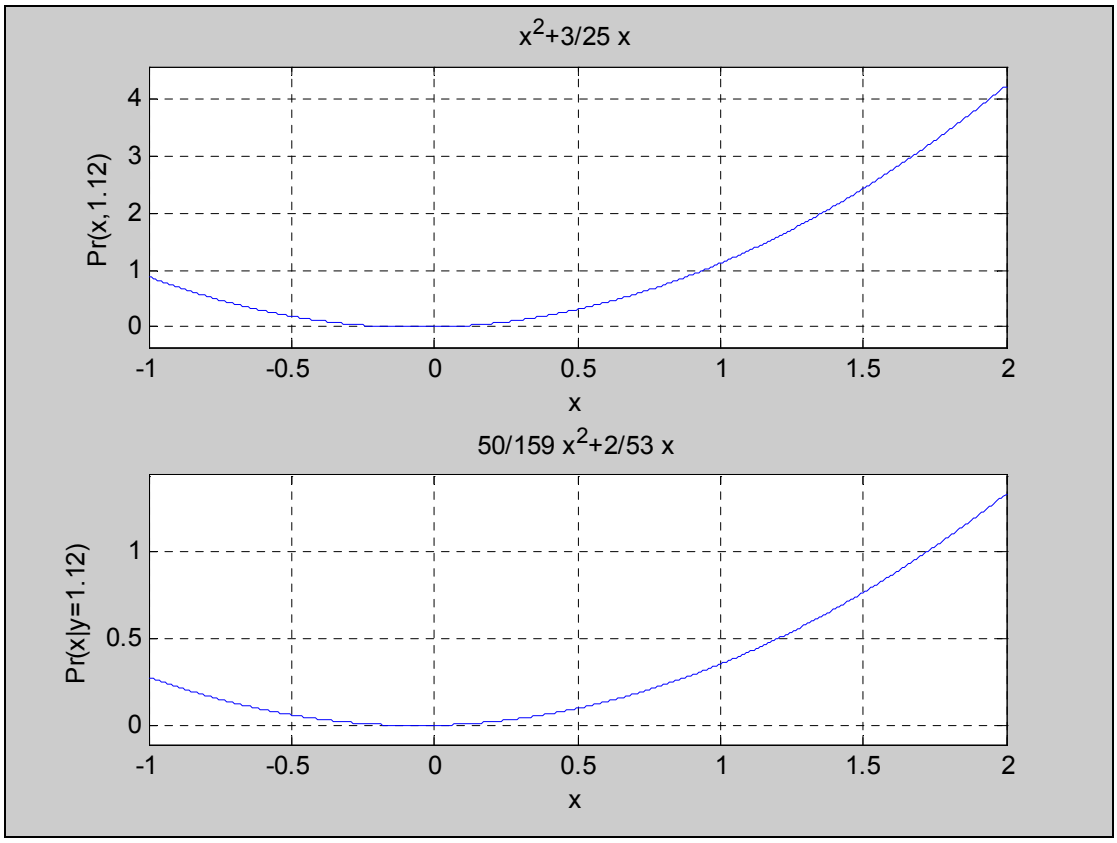
二、求算 $p(x,y=1.12)$ 之聯合分佈中，x domain 下的機率分佈情形。其結果如圖表 27 的上方圖形。

三、以 $\dfrac{p(x,y=1.12)}{\int_{-1}^{2}p(x,y=1.12)dx}$ 得出 $p(x \mid y=1.12)=\dfrac{50}{159}x^2+\dfrac{2}{53}x$，這個結果和(5.28)相同，同圖表 27 之下方圖形。

由以上的結果顯示出來，如果我們想要求算一個條件機率時，可以使用一般的求法得出通式，然後再進行選擇從哪一個條件相關值進行 PDF 觀測（這個動作好像是圖表 25 進行切片的行爲，切片的結果就是圖表 26）。每一個切片都是 PDF 可以滿足機率空間的行爲。

另外一種作法，就是如圖表 27，直接在聯合分佈中選擇 y=1.12，然後統計 y=1.12 時，在聯合分佈 $p(x, y=1.12)$ 中的 x 值所產生的聯合機率值。(圖表 27 的上圖)。最後，使用 $p(x, y=1.12)$ 除上 x 值的聯合機率總積分值(圖表 27 上，$x^2+\frac{3}{25}x$ 的曲線下方總面積)。這樣子的做法也可以得出條件機率式 $p(x \mid y=1.12)$。

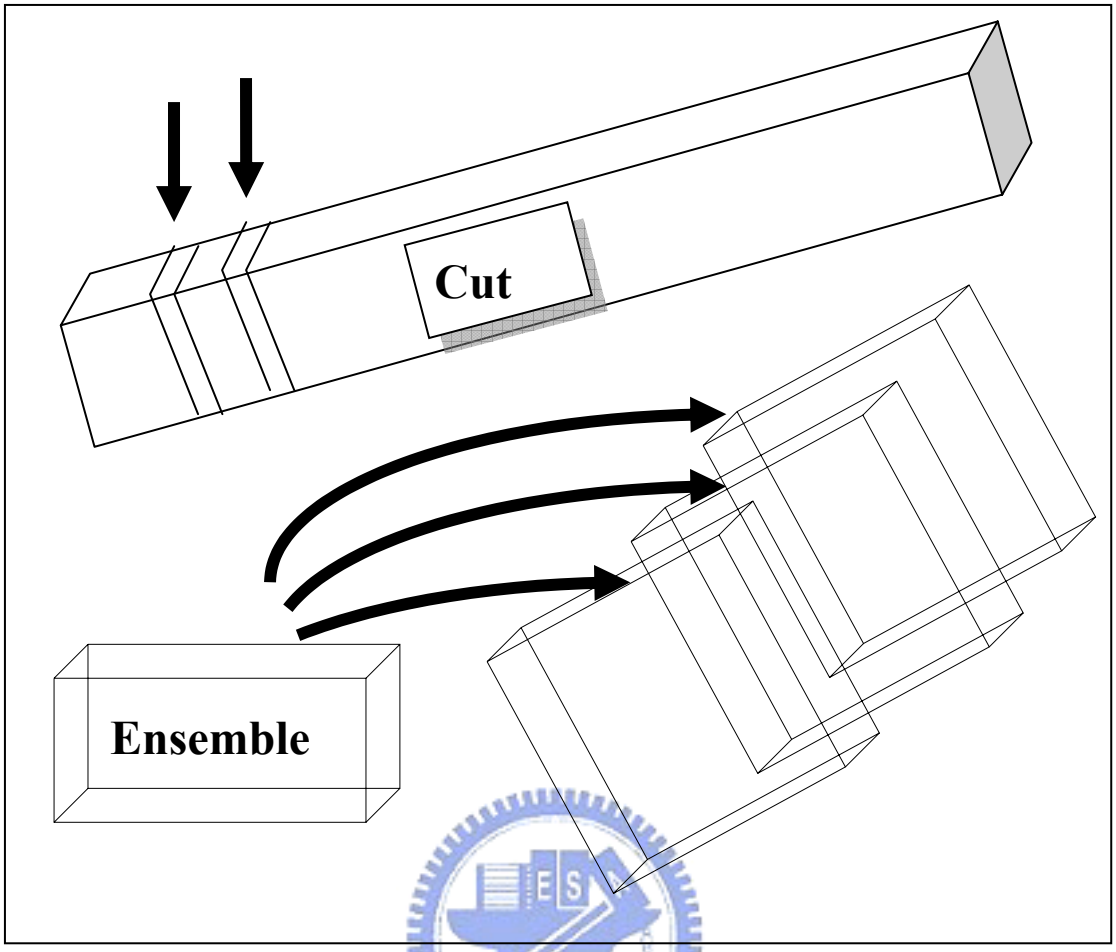
上述的兩種方法從條件機率的定義式 $p(x \mid y)=\frac{p(x,y)}{\int_D p(x,y)dx}$ 來看是相同並無二致，但實際上在操作上有所區別。端看從何處下手較簡易來執行。



圖表 27 先固定觀察值之條件機率密度函數求算

討論：

- 當條件機率 $p(x \mid y)=\frac{f(x,y)}{\int_D f(x,y)dx}$ 在計算上有窒礙難行之處，無法一次得出整個通式，而恰巧只對其中相關變數 y 的部分值或固定的幾個值有興趣時，可以考慮使用第二種方法來計算條件機率。



圖表 28 組合式條件機率之想法

- 如圖表 28，我們將以往的條件機率計算方式稱爲切片(cut)，現在本研究要使用的方法命名爲組合法(ensemble)。當條件機率的通式(general form)容易求出時，使用切片法可以很容易完成整個隨機變數的樣本活動空間度量。當通式無法求出時，可以考慮改用組合法，將切片逐一求出，最後組合回來整個想要觀察的機率空間樣本行爲。

- 在此將組合法(ensemble)的整個執行程序作成一個列表，以備下一節說明時對照使用。

表格 2 組合法(ensemble)求算條件機率之步驟

執行動作次序	執行動作內容
A 指定相關變數	令相關變數爲一常數 y=constant
B 計算聯合分佈下的 x 邊際效用	計算 y=constant, 在 x domain 下之 probability
C 執行切片之條件機率計算 $p(x \mid y) = \frac{f(x, y = cons\tan t)}{\int_D f(x, y = cons\tan t)dx}$	畫出方案 b 的圖形，並且將此圖形對圖形曲線下方的總面積進行歸一化處理即爲所求。如圖表 27

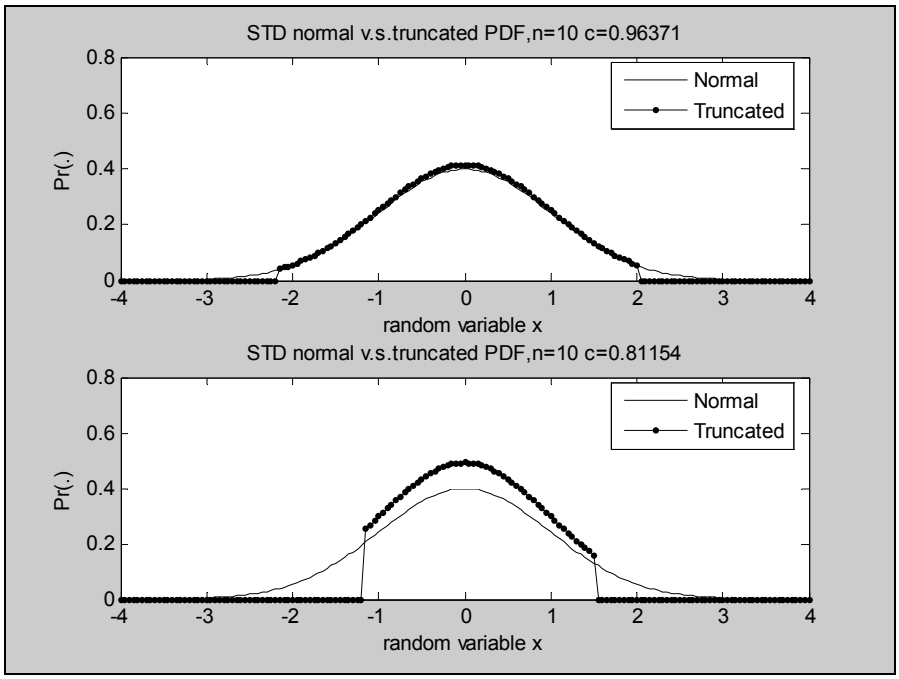
5.7. 使用聯合機率的角度來思考全距(range)公式

對於上一小節之全距公式

$$p(r \mid n) = \int_{-\infty}^{\infty} n(n-1) f(x_{1:n}) f(x_{1:n} + r)\{F(x_{1:n} + r) - F(x_{1:n})\}^{n-2} dx_{1:n} \tag{5.29}$$

令 $c = F(x_{1:n} + r) - F(x_{1:n})$，則 range 公式可以視爲全距 r，和覆蓋率 c 的聯合分佈

$$p(r, c \mid n) = \int_{-\infty}^{\infty} n(n-1) f(x_{1:n}) f(x_{1:n} + r) c^{n-2} dx_{1:n} \tag{5.30}$$

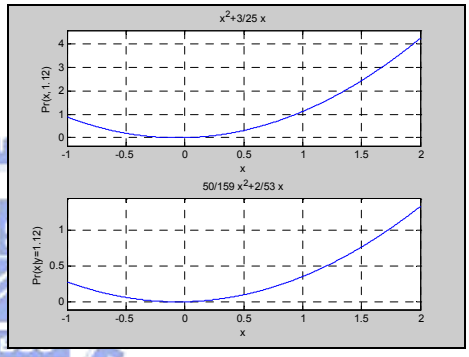


圖表 29 覆蓋率對於推估樣本的影響

現在的條件變數是覆蓋率 c 和樣本大小 n，如圖表 29 所示。覆蓋率如果太少的話，我們進行估計時所使用的黑色實線會偏離原來的分佈。所以我們有興趣的範圍只是覆蓋率的部份，在這個前提之下就很適合使用組合法(ensemble)進行條件機率之求算。

■ 步驟 A

表格 3 組合法(ensemble)求算條件機率之步驟

執行動作次序	執行動作內容
A 指定相關變數	令相關變數爲一常數 y=constant
B 計算聯合分佈下的 x 邊際效用	計算 y=constant, 在 x domain 下之 probability
C 執行切片之條件機率計算 $p(x \mid y) = \frac{f(x, y = cons\tan t)}{\int_D f(x, y = cons\tan t)dx}$	畫出方案 b 的圖形，並且將此圖形對圖形曲線下方的總面積進行歸一化處理即爲所求。如圖表 27 

現在要執行上一節所提出求取條件機率觀點的 A 步驟。先令條件變數 c 爲一常數 Cc。

令 $F(x_{1:n} + r) - F(x_{1:n}) - Cc = 0$

組合法(ensemble)之條件機率之求法

$$p(r \mid c = Cc, n) = \frac{p(r, c = Cc \mid n)}{\int_{dr} p(r, c = Cc \mid n)} \tag{5.31}$$

本研究提出的想法，使用 ***dirac delta*** 函數來固定覆蓋率的值。則 $p(r, c = Cc \mid n)$ 可以寫成：

$$p(r, c = Cc \mid n) =$$
$$\int_{-\infty}^{\infty} n(n-1) f(x_{1:n}) f(x_{1:n} + r) \times$$
$$\{\left(F(x_{1:n} + r) - F(x_{1:n})\right)^{n-2} \times \delta(F(x_{1:n} + r) - F(x_{1:n}) - Cc)\} dx_{1:n}$$
(5.32)

(5.32)式是處於積分環境下的公式，考量這種情形，可以考慮使用 dirac delta(.)函數來進行展開。它具有以下的性質：

$$\int_{-\infty}^{\infty} f(x)\delta\left(x - a\right) = f(a) \qquad (5.33)$$

(5.33)的成立前提是單根(single root)的情況。其效能類似於對函數 $f(x)$ 進行取樣，$\delta(.)$ 於是可以視爲是取樣函數。

表格 4 標準常態分佈下，p(r|c=0.95,n=15)之最小 x 左端點求解

全距(range)	Second root of $x_{1:n}$	First root of $x_{1:n}$
3.9200	-1.9742	-1.9458
3.9700	-2.1462	-1.8238
4.0200	-2.2386	-1.7814
4.0700	-2.3163	-1.7537
4.1200	-2.3865	-1.7335
4.1700	-2.4519	-1.7181
4.2200	-2.5142	-1.7058
4.2700	-2.5742	-1.6958
4.3200	-2.6323	-1.6877
4.3700	-2.6891	-1.6809
4.4200	-2.7447	-1.6753
4.4700	-2.7995	-1.6705

- 多根的情形：

表格 4 是標準常態分佈下，覆蓋率固定要求 95%，抽樣大小數量爲 15。求解 $F(x_{1:n} + r) - F(x_{1:n}) - 0.95=0$，輸入 r，求解根 $x_{1:n}$ 的結果。由表

中的結果可以發現；當覆蓋率固定時，相同大小的全距會出現在不同的區間上。也就是說，$F(x_{1:n}+r)-F(x_{1:n})-Cc=0$的解將不只一個。於是我們必須討論 dirac delta 函數在大於一個根的情況下表示方法。

令$\delta(g(x))$爲一新的取樣函數，在本研究中等於 $g(x)=F(x_{1:n}+r)-F(x_{1:n})-Cc$。如果令$g(x)$=0 的根不只一個，則(5.33)的結果將必須重新考慮。

此處$F(x_{1:n}+r)-F(x_{1:n})-Cc=0$可能會有多個根，此時(5.32)的結果必須參考多根的 dirac delta 函數寫法。

$$令 \mathrm{H}(x_{1:n})=n(n-1)f(x_{1:n})f(x_{1:n}+r)\times\left(F(x_{1:n}+r)-F(x_{1:n})\right)^{n-2}$$

$$\Rightarrow\int_{-\infty}^{\infty}\mathrm{H}(x_{1:n})\cdot\delta\left(g(x_{1:n})\right)dx_{1:n}=\sum_{j=1}^{k}\int_{\eta_j-\varepsilon}^{\eta_j+\varepsilon}\mathrm{H}(x_{1:n})\cdot\delta\left(g(x_{1:n})\right)dx_{1:n} \tag{5.34}$$

選擇在η_j進行泰勒函數展開，假設$\varepsilon>0,\varepsilon\to 0$，我們可以考慮在所有的根$\eta_j$附近將積分區間予以摺疊，於是整個實數軸上的積分效應可以拆開成很多段相加：

$$\sum_{j=1}^{k}\int_{\eta_j-\varepsilon}^{\eta_j+\varepsilon}\mathrm{H}(x_{1:n})\cdot\delta\left(g(\eta_j)+g'(\eta_j)(x_{1:n}-\eta_j)+O(\frac{g''(\eta_j)}{2!})\right)dx_{1:n}$$

$$\simeq\sum_{j=1}^{k}\int_{\eta_j-\varepsilon}^{\eta_j+\varepsilon}\mathrm{H}(x_{1:n})\cdot\delta\left(g'(\eta_j)(x_{1:n}-\eta_j)\right)dx_{1:n} \tag{5.35}$$

使用輔助公式

$$\delta(a\,x_{1:n})=\frac{\delta(x_{1:n})}{|a|} \tag{5.36}$$

則(5.35)可以繼續化簡爲：

$$\simeq \sum_{j=1}^{k} \frac{1}{\left|g'(\eta_j)\right|} \int_{\eta_j-\varepsilon}^{\eta_j+\varepsilon} \mathrm{H}(x_{1:n}) \cdot \delta(x_{1:n} - \eta_j) dx_{1:n} \tag{5.37}$$

$$\Rightarrow \sum_{j=1}^{k} \frac{\mathrm{H}(\eta_j)}{\left|g'(\eta_j)\right|} \tag{5.38}$$

由(5.34)和(5.38)比對之後，可以得出

$$\delta(g(x_{1:n})) = \sum_{j=1}^{k} \frac{\delta(x_{1:n} - \eta_j)}{\left|g'(\eta_j)\right|} \tag{5.39}$$

最後，根據$g(x_{1:n})$根的數量，可以得出以下的歸納做爲整理。

$$\int_{-\infty}^{\infty} \mathrm{H}(x_{1:n}) \delta(g(x_{1:n})) dx_{1:n} \Rightarrow g(\eta_j) = 0 \begin{cases} j = 1, f(\eta_j) \\ j \geq 2, \sum_j \dfrac{f(\eta_j)}{\left| \dfrac{\partial}{\partial x_{1:n}} g(x_{1:n}) \Big|_{x_{1:n}=\eta_j} \right|} \end{cases}$$

$$,\eta_j \text{ 不得為重根，且必須為實數，另外} (\frac{\partial}{\partial x_{1:n}} g(x_{1:n}) \Big|_{x_{1:n}=\eta_j}) \neq 0 \tag{5.40}$$

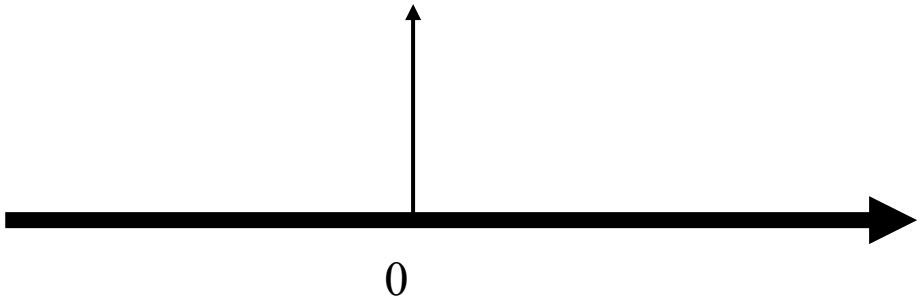
■ 解釋$\delta(g(x_{1:n}))$的物理意義

由前一小節的分析可以了解到，當$\delta(g(x_{1:n}))$只有一個根的時候，它的功用類似於取樣。但是當$\delta(g(x_{1:n}))$有超過一個以上的根時，由公式的結果並無法直接有所領悟。由於 dirac delta 函數，是一種抽象的定義，在實體的自然界之中，並無實際的事物可以與之對應，所以只能夠透過間接觀察的方式來進行了解。

Dirac delta 函數的定義就是單位步級函數(UnitStep(.))的微分

$$\delta(x_{1:n}) = \frac{\partial}{\partial x_{1:n}} UnitStep(x_{1:n}) \tag{5.41}$$

它的對應圖形是在$x_{1:n} = 0$的地方有一個無限大能量的脈衝。



圖表 30 dirac delta 之定義

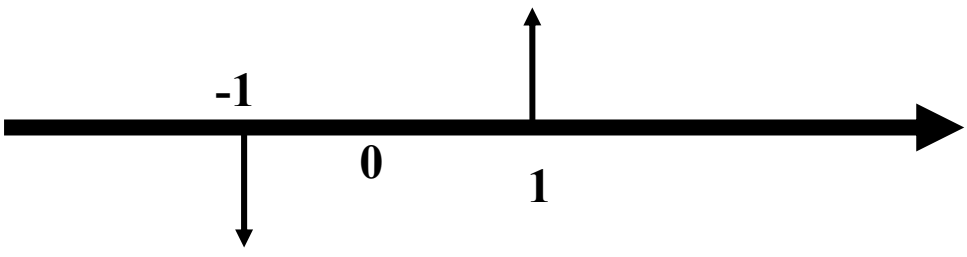
圖表 32 是單位步級函數 UnitStep(.)以及相關的合成函數輸出結果。圖表 32-1 是原來$\boldsymbol{UnitStep}(x_{1:n})$的圖形；圖表 32-2 是$\boldsymbol{UnitStep}(x_{1:n}^2-1)$，圖表 32-3 是$\boldsymbol{UnitStep}((x_{1:n}^2-1)(x_{1:n}+\frac{1}{2}))$。

以圖表 32-2 爲例，總共有兩個根。分別在 x=1,x=-1，可以視爲在 x=1 及 x=-1 將有兩個不會同時存在的$\boldsymbol{UnitStep}(x_{1:n}-1)$和$\boldsymbol{UnitStep}(x_{1:n}+1)$

$$\boldsymbol{UnitStep}(x_{1:n}+1)=\begin{cases}1,\boldsymbol{if}\ x_{1:n}\geq -1\\0,\boldsymbol{if}\ x_{1:n}<-1\end{cases} \tag{5.42}$$

$$\boldsymbol{UnitStep}(x_{1:n}-1)=\begin{cases}1,\boldsymbol{if}\ x_{1:n}\geq 1\\0,\boldsymbol{if}\ x_{1:n}<1\end{cases} \tag{5.43}$$

圖表 32-2 就是同時能夠滿足(5.42)，(5.43)的邏輯輸出結果。現在如果對圖表 32-2 的圖形進行對 x 的微分，會在 x=-1,x=1 的兩個位置形成兩個脈衝。



圖表 31 UnitStep(x^2-1)之微分輸出結果

$$\frac{\partial}{\partial x}\boldsymbol{UnitStep}(x^2-1)=\delta(x^2-1).(2x) \tag{5.44}$$

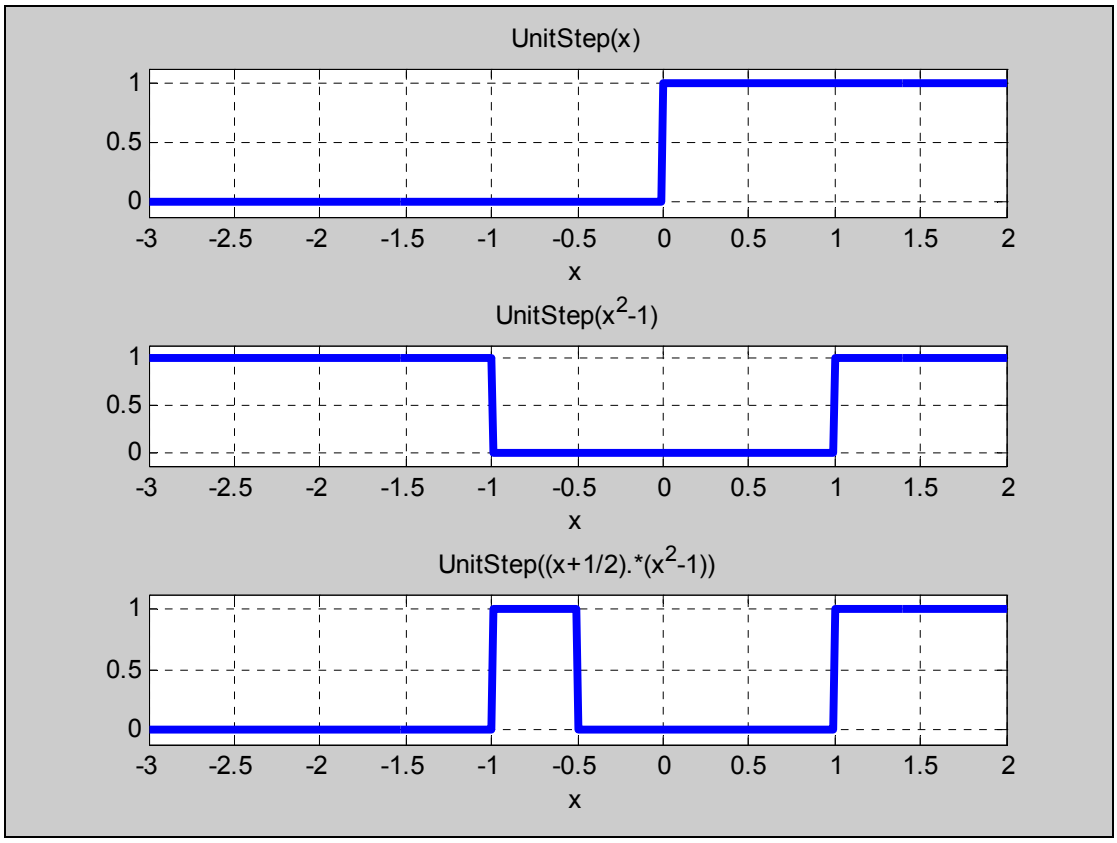
又圖表 31 之輸出結果爲

$$-\delta(x+1)+\delta(x-1) \tag{5.45}$$

(5.44)的結果要等於(5.45)，

$$\delta(x^2-1)=\frac{\delta(x+1)}{|2x|}+\frac{\delta(x-1)}{|2x|}$$
$$\Rightarrow \delta(g(x))=\sum_{j=1}^{k}\frac{\delta(x-\eta_j)}{|g'(\eta_j)|}, g(\eta_j)=0, \textit{for } \text{j=1,2}\cdots\text{k} \tag{5.46}$$

(5.46)是將(5.44)，(5.45)視爲相等所得出的結果。至此，我們可以了解到，$\delta(g(x))$會將所有的根相互之間的影響統一考量，然後輸出。這一點，可以由圖表 32-2，圖表 32-3 的結果觀察得到。因爲$\delta(g(x))$正是從圖表 32-2 的結果微分所推導得出。



圖表 32 複合式合成函數在 UnitStep(.)上之結果

根據上述的歸納原理，可以知道如果$F(x_{1:n}+r)-F(x_{1:n})-Cc=0$有 k 個根滿足(5.40) ，$\eta_j, j=1..k$ 。則(5.32)的結果成爲:

$$\sum_{j=1}^{k}\{\frac{\mathrm{H}(\eta_j)}{|f_C{}'(\eta_j)|}\}, f_C(\eta_j)=0 \ \textit{and}\ f_C{}'(\eta_j)\neq 0 \text{ and } \eta_j\in\mathbb{R} \tag{5.47}$$

$$\mathrm{H}\left(\eta_j\right)=n(n-1)f(\eta_j)f(\eta_j+r)\left(F(\eta_j+r)-F(\eta_j)\right)^{n-2}$$

$$f_C(x_{1:n}):F(x_{1:n}+r)-F(x_{1:n})-Cc$$
$$\forall\eta_j, f_{C(\eta_j)}=0, j=1,2\cdots k$$

$$f_C'(\eta_j)=\frac{\partial}{\partial x}(F(x_{1:n}+r)-F(x_{1:n})-Cc)=f(x_{1:n}+r)-f(x_{1:n})$$
$$Thus,\Rightarrow f_C'(x_{1:n})\Big|_{x_{1:n}=\eta_j}=f(\eta_j+r)-f(\eta_j)$$

展開(5.47)式，就可以比較清楚看到全貌，了解爲什麼覆蓋率會成爲固定值。

$$p(r,c=Cc\mid n).$$

$$\sum_{j=1}^{k}\{\frac{\mathrm{H}(\eta_j)}{\left|f_C'(\eta_j)\right|}\}$$
$$\Rightarrow\sum_{j=1}^{k}\{\frac{n(n-1)f(\eta_j)f(\eta_j+r)\{F(\eta_j+r)-F(\eta_j)\}^{n-2}}{\left|f(\eta_j+r)-f(\eta_j)\right|}\}\quad(5.48)$$

此處引入我們所給予的限制式$\Rightarrow F(\eta_j+r)-F(\eta_j)=Cc$

$$\Rightarrow p(r,c=Cc\mid n)=\sum_{j=1}^{k}\{\frac{n(n-1)f(\eta_j)f(\eta_j+r)\left(Cc\right)^{n-2}}{\left|f(\eta_j+r)-f(\eta_j)\right|}\}\quad(5.49)$$
$$f(\eta_j+r)-f(\eta_j)\neq 0 \text{ and } \eta_j\in\mathbb{R}$$

在這個小節裡頭，使用了 dirac delta(.)函數的特性，成功的除去了原來 range 公式的積分符號。

(5.49)在處理上必須利用$x_{1:n}$的 domain 做限制捨去增根，然後再進行展開。另外一個需要注意的是$\left(-4\sigma\le x_{1:n}\right)\wedge\left(x_{1:n}+r\le 4\sigma\right)$也必須考慮加入（因爲常態分佈在$\pm4\sigma$以外的機率可以視爲零），作爲$x_{1:n}$的限制式，全距 r

必須使得 $x_{1:n}$ 落在限制條件內。

根據上述的探討，在此寫出步驟 A 的結果。

$$\text{限制式} \Rightarrow F(\eta_j + r) - F(\eta_j) = Cc$$

$$\Rightarrow p(r, c = Cc \mid n) = \sum_{j=1}^{k} \{ \frac{n(n-1) f(\eta_j) f(\eta_j + r) (Cc)^{n-2}}{\left| f(\eta_j + r) - f(\eta_j) \right|} \} \quad (5.50)$$

$f(\eta_j + r) - f(\eta_j) \neq 0$ and $\eta_j \in \mathbb{R}$ *and* η_j 不得為重根

$\{\eta_{j, j=1,2\cdots n} \mid g(\eta_j)=0 \wedge lb \leq \eta_j \leq ub \wedge 0 \leq r \leq \left|(ub - lb)\right| \wedge lb \leq \eta_j + r \leq ub\}$

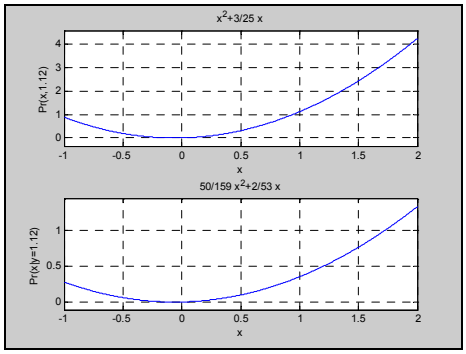
lb $\Rightarrow (-4\sigma)$:常態分佈的下界

ub $\Rightarrow (4\sigma)$:常態分佈的上界

σ:標準差

執行上一節的步驟 B

表格 5 組合法(ensemble)求算條件機率之步驟

執行動作次序	執行動作內容
A 指定相關變數	令相關變數爲一常數 y=constant
B 計算聯合分佈下的 x 邊際效用	計算 y=constant, 在 x domain 下之 probability
C 執行切片之條件機率計算 $p(x \mid y) = \frac{f(x, y = const ant)}{\int_D f(x, y = const ant) dx}$	畫出方案 b 的圖形，並且將此圖形對圖形曲線下方的總面積進行歸一化處理即爲所求。如圖表 27 

步驟 B 要開始計算全距(range) r 在覆蓋率爲固定常數時的聯合機率分佈

$p(r, c = Cc \mid n)$，其計算方式就是將(5.50)予以實現即可。首先計算

$$F(x_{1:n} + r) - F(x_{1:n}) - Cc = 0 \quad (5.51)$$

$F(x_{1:n}) = \frac{1}{2} + \frac{1}{2} erf(\frac{x_{1:n}}{\sqrt{2}\sigma})$，如果直接將(5.51)式給予電腦疊代求解，電腦將只會給出一個最佳解。這樣子的結果並不符合原本的期待，所以此處必須將$F(x_{1:n})$使用厄米特多項式(Hermite Polynomials chaos)進行展開。

$$\text{Hermite}(i, x_{1:n}) = (-1)^i \exp(x_{1:n}^2) \frac{\partial^i}{\partial x_{1:n}^i} \exp(-x_{1:n}^2) \quad (5.52)$$

取前面的 15 階多項式來近似

$$F(x_{1:n}) = \sum_{i=0}^{15} t_i \text{Hermite}[i, x_{1:n}] \quad (5.53)$$

t_i：每一階多項式的迴歸係數

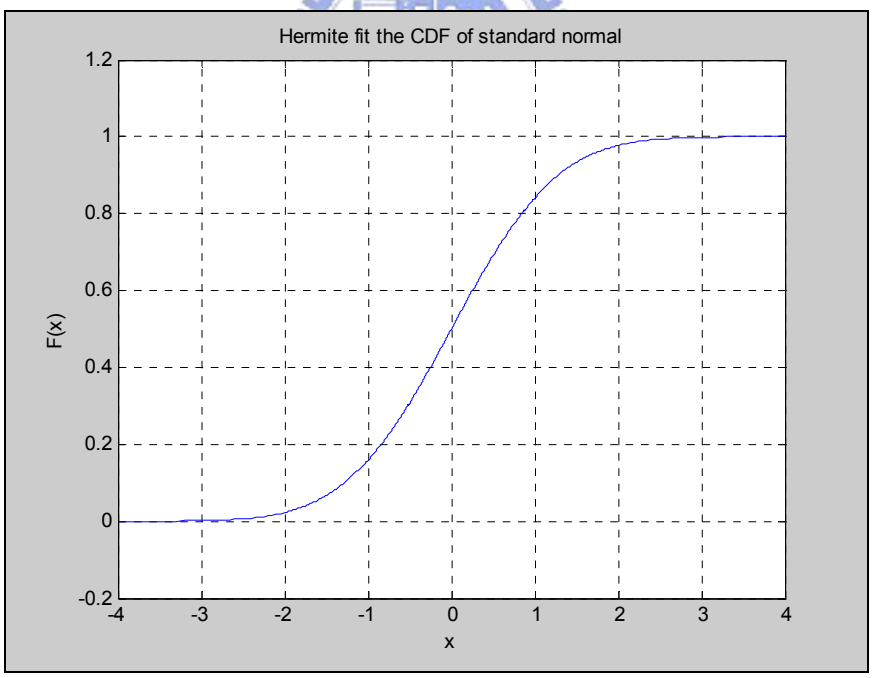
下一段是標準常態分佈的執行結果，以及所有的迴歸係數

Linear model:

result(x) = t0 + t1*2*x + t2*(-2+4*x^2) + t3*(-12*x+8*x^3) +
t4*(12-48*x^2+16*x^4) + t5*(120*x-160*x^3+32*x^5) +
t6*(-120+720*x^2-480*x^4+64*x^6) + t7*(16*x*(-105+210*x^2-84*x^4+8*x^6)) +
t8*(1680-13440*x^2+13440*x^4-3584*x^6+256*x^8) +
t9*(32*x*(945-2520*x^2+1512*x^4-288*x^6+16*x^8)) +
t10*(-30240+302400*x^2-403200*x^4+161280*x^6-23040*x^8+1024*x^10) +
t11*(64*x*(-10395+34650*x^2-27720*x^4+7920*x^6-880*x^8+32*x^10)) +
t12*(665280-7983360*x^2+13305600*x^4-7096320*x^6+1520640*x^8-135168*
x^10+4096*x^12) +
t13*(128*x*(135135-540540*x^2+540540*x^4-205920*x^6+34320*x^8-2496*x^10
+64*x^12)) +
t14*(-17297280+242161920*x^2-484323840*x^4+322882560*x^6-92252160*x^8+
12300288*x^10-745472*x^12+16384*x^14) +
t15*(256*x*(-2027025+9459450*x^2-11351340*x^4+5405400*x^6-1201200*x^8+1
31040*x^10-6720*x^12+128*x^14))

t_i effitients (with 95% tonfidente bounds):

t0 = 0.5 (0.5, 0.5)
t1 = 0.1629 (0.1629, 0.1629)
t2 = 2.351e-017 (-3.047e-007, 3.047e-007)
t3 = -0.004524 (-0.004524, -0.004524)
t4 = 2.242e-018 (-3.687e-008, 3.687e-008)
t5 = 0.0001131 (0.0001131, 0.0001131)
t6 = -2.389e-019 (-2.84e-009, 2.84e-009)
t7 = -2.242e-006 (-2.242e-006, -2.241e-006)
t8 = 6.527e-021 (-1.538e-010, 1.538e-010)
t9 = 3.607e-008 (3.604e-008, 3.611e-008)
t10 = -7.42e-024 (-6.425e-012, 6.425e-012)
t11 = -4.883e-010 (-4.898e-010, -4.869e-010)
t12 = -4.218e-024 (-1.318e-013, 1.318e-013)
t13 = 4.919e-012 (4.897e-012, 4.941e-012)
t14 = 2.263e-026 (-4.225e-015, 4.225e-015)
t15 = -5.687e-014 (-5.792e-014, -5.583e-014)



圖表 33 常態分佈之累積機率函數使用 Hermite Polynomials 展開

其次判斷 $F(x_{1:n}+r)-F(x_{1:n})-Cc=0$ 解出來的根是否合乎限制式？

$f(\eta_j + r) - f(\eta_j) \neq 0$ and $\eta_j \in \mathbb{R}$ *and* η_j 不得為重根

$\{\eta_{j,j=1,2\cdots n} \mid g(\eta_j)=0 \wedge lb \leq \eta_j \leq ub \wedge 0 \leq r \leq |(ub - lb)| \wedge lb \leq \eta_j + r \leq ub\}$

lb $\Rightarrow$ (-4σ):常態分佈的下界

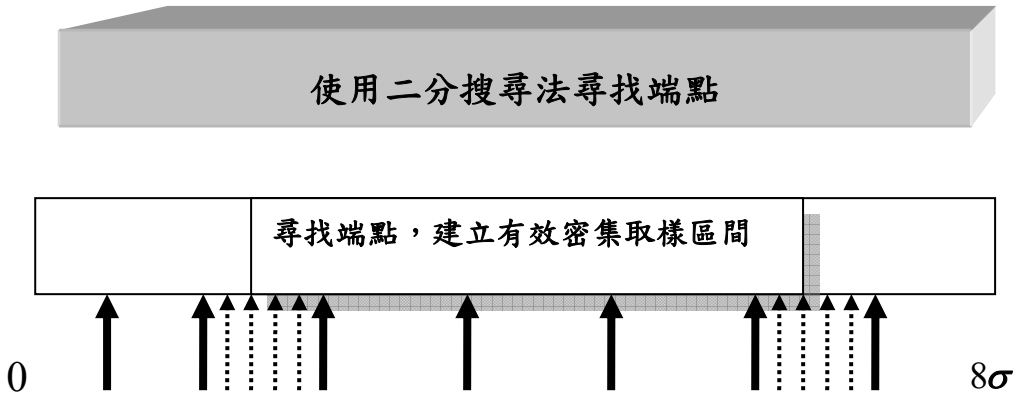
ub $\Rightarrow$ (4σ):常態分佈的上界

σ:標準差

在這個階段，先將 $x_{1:n}$ 的範圍予以界定，然後將 r 的抽樣值視爲輸入（r 的抽樣範圍是 $[0,8\sigma]$）；如此一來，$F(x_{1:n} + r) - F(x_{1:n}) - Cc = 0$ 就只剩下一個變數，於是可以開始求根。

■ 建立端點，減少電腦運算時間

因爲每取一個 r 值輸入，就要求解一次 15 階的多項式根。所以此階段的演算將非常耗費電腦執行的時間。所以此階段要準備進行一項資料的搜尋動作，找尋 range r 的 domain 中開始對應到聯合機率值不等於零的兩個端點。在兩個端點內的有效區間執行 range r 的取樣輸入即可。如此才能夠節省下大量的演算時間。因爲 range r 的分佈區間從 0 到 8σ，我們不知道 r 的 domain 從何時開始有聯合機率值開始對應；所以此處剛開始使用大範圍間距取樣，當發現有聯合機率開始對應不爲零時，使用更小的取樣間距，以二分搜尋法找尋下一個不爲零的端點，重覆此一步驟，直到搜尋間距小於事先界定的範圍以內才停止。



■ 留下合乎限制式的根

以下表格 6 是以 n=15，覆蓋率 c=0.95 所計算得出的結果。二分搜尋法在 3.92 和 5.62 找到最佳端點。在最佳端點的有效區間之內，取樣間距是 0.05

由表中可以發現，Hermite polynomials 所解出的根有兩個，直接疊代法只能求出一個根。Hermite polynomial 所求出的第一個根和疊代法所得出的結果幾乎是一致的。

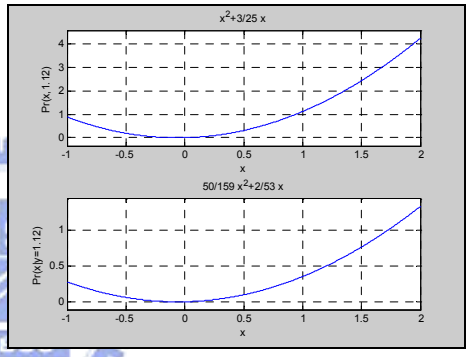
表格 6 使用 Hermite polynomials 求解與直接疊代法之求解比較

覆蓋率 r	Hermite polynomial chaos		Iteration
	Second $x_{1:n}$	First $x_{1:n}$	$x_{1:n}$
3.9200	-1.9742	-1.9458	-1.9539
3.9700	-2.1462	-1.8238	-1.8244
4.0200	-2.2386	-1.7814	-1.7817
4.0700	-2.3163	-1.7537	-1.7539
4.1200	-2.3865	-1.7335	-1.7336
4.1700	-2.4519	-1.7181	-1.7180
4.2200	-2.5142	-1.7058	-1.7056
4.2700	-2.5742	-1.6958	-1.6956
4.3200	-2.6323	-1.6877	-1.6874
4.3700	-2.6891	-1.6809	-1.6806
4.4200	-2.7447	-1.6753	-1.6749
4.4700	-2.7995	-1.6705	-1.6702
4.5200	-2.8535	-1.6665	-1.6662

4.5700	-2.9069	-1.6631	-1.6628
4.6200	-2.9599	-1.6601	-1.6600
4.6700	-3.0124	-1.6576	-1.6575
4.7200	-3.0645	-1.6555	-1.6555
4.7700	-3.1163	-1.6537	-1.6538
4.8200	-3.1678	-1.6522	-1.6523
4.8700	-3.2191	-1.6509	-1.6511
4.9200	-3.2702	-1.6498	-1.6501
4.9700	-3.3210	-1.6490	-1.6492
5.0200	-3.3717	-1.6483	-1.6485
5.0700	-3.4222	-1.6478	-1.6479
5.1200	-3.4726	-1.6474	-1.6474
5.1700	-3.5229	-1.6471	-1.6469
5.2200	-3.5731	-1.6469	-1.6466
5.2700	-3.6233	-1.6467	-1.6463
5.3200	-3.6735	-1.6465	-1.6460
5.3700	-3.7237	-1.6463	-1.6458
5.4200	-3.7741	-1.6459	-1.6456
5.4700	-3.8245	-1.6455	-1.6455
5.5200	-3.8750	-1.6450	-1.6454
5.5700	-3.9252	-1.6448	-1.6453
5.6200	-3.9750	-1.6450	-1.6452

■ 執行上一節的步驟 C

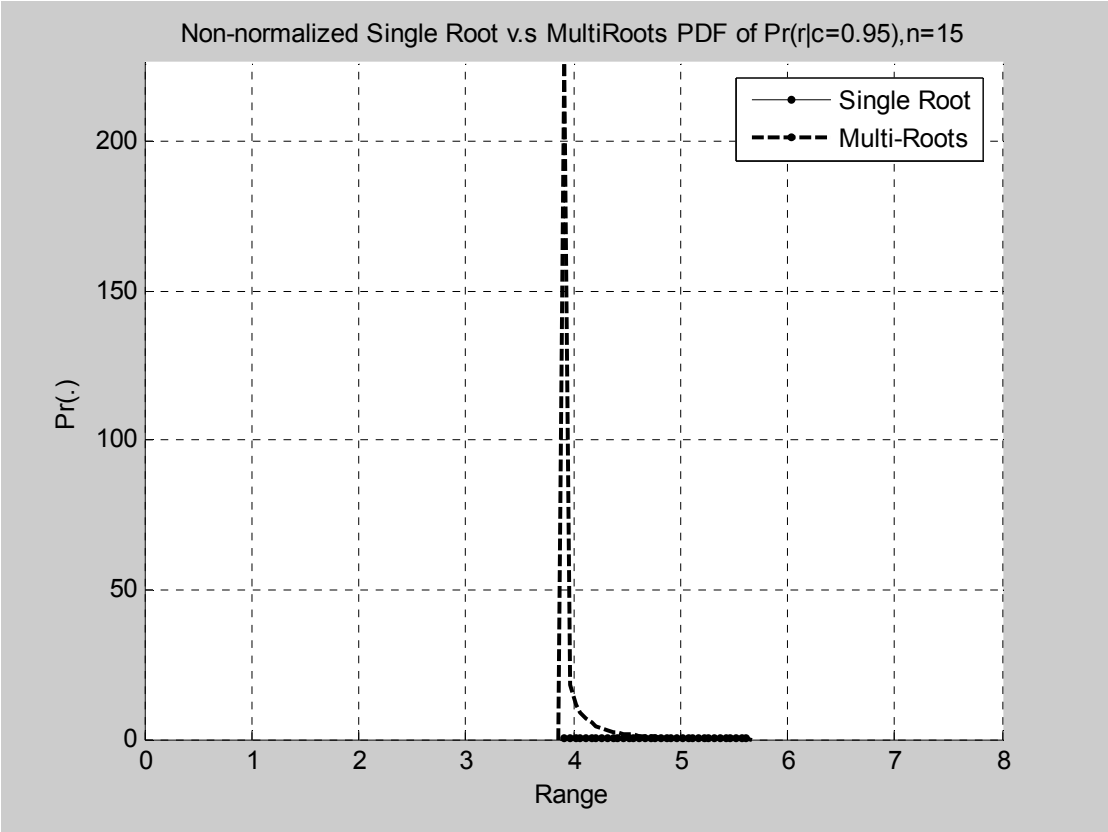
表格 7 組合法(ensemble)求算條件機率之步驟

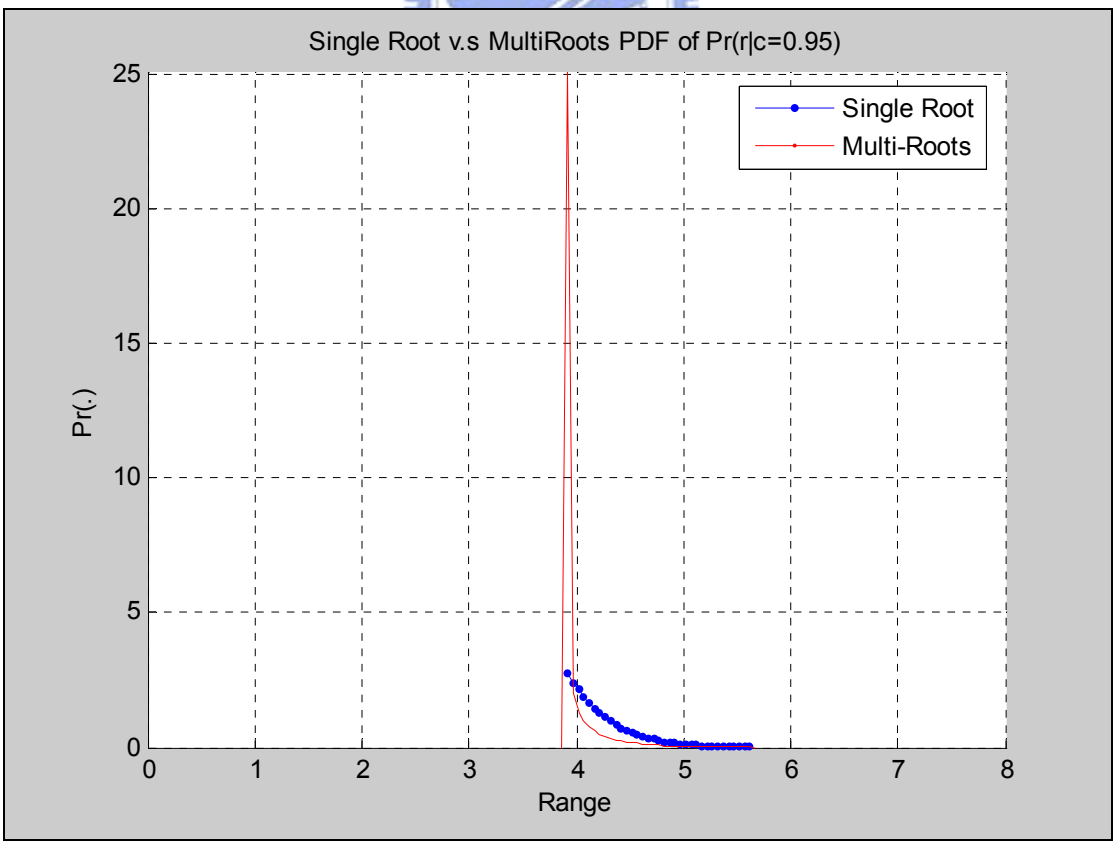
執行動作次序	執行動作內容
A 指定相關變數	令相關變數爲一常數 y=constant
B 計算聯合分佈下的 x 邊際效用	計算 y=constant, 在 x domain 下之 probability
C 執行切片之條件機率計算 $p(x \mid y) = \frac{f(x, y = cons\tan t)}{\int_D f(x, y = cons\tan t)dx}$	畫出方案 b 的圖形，並且將此圖形對圖形曲線下方的總面積進行歸一化處理即爲所求。如圖表 27 

現在準備畫出圖表 27-1 圖形。所需使用之公式

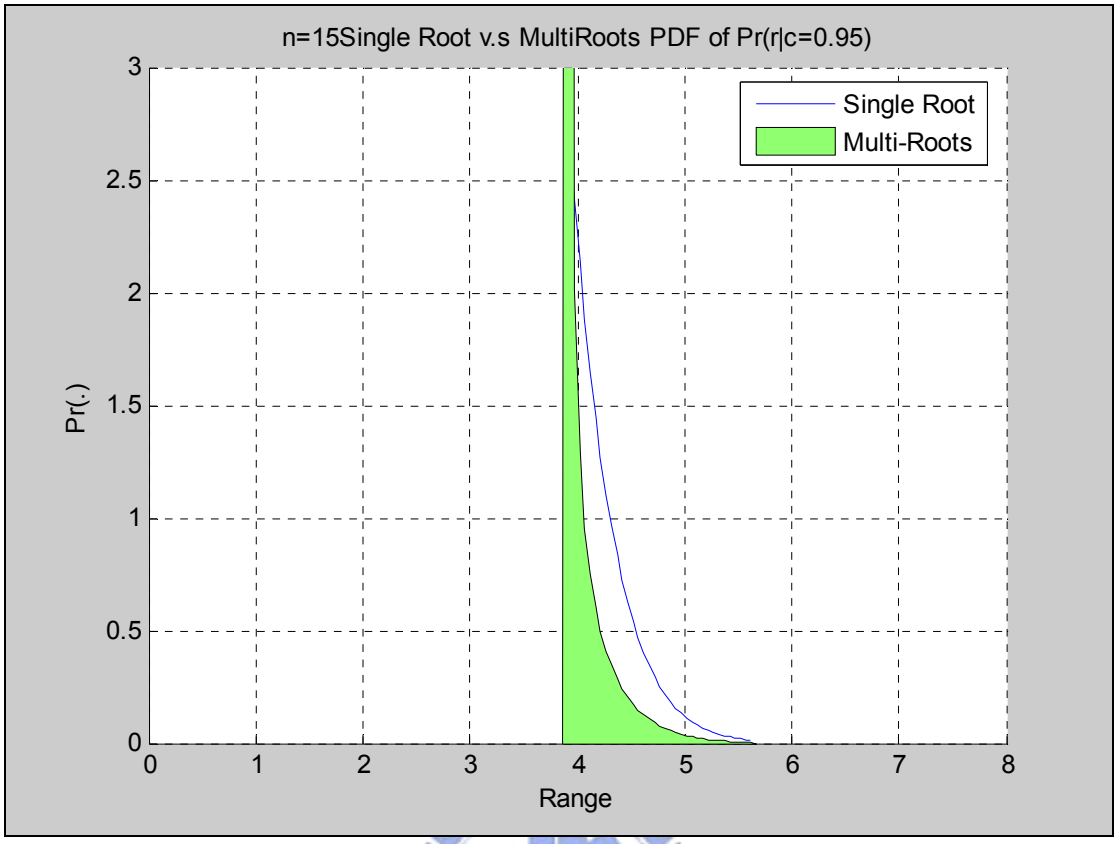
$$\Rightarrow p(r, c = Cc \mid n) = \sum_{j=1}^{k} \{ \frac{n(n-1) f(\eta_j) f(\eta_j + r) (Cc)^{n-2}}{\left| f(\eta_j + r) - f(\eta_j) \right|} \} \quad (5.54)$$

將前一節整理完所留下有效的 r 值逐一代入(5.54)式中就會得出圖表 27-1。然後再對曲線下的面積進行歸一化相除就會得出圖表 27-2 的條件機率。也就是 $p(r \mid c = Cc, n)$。

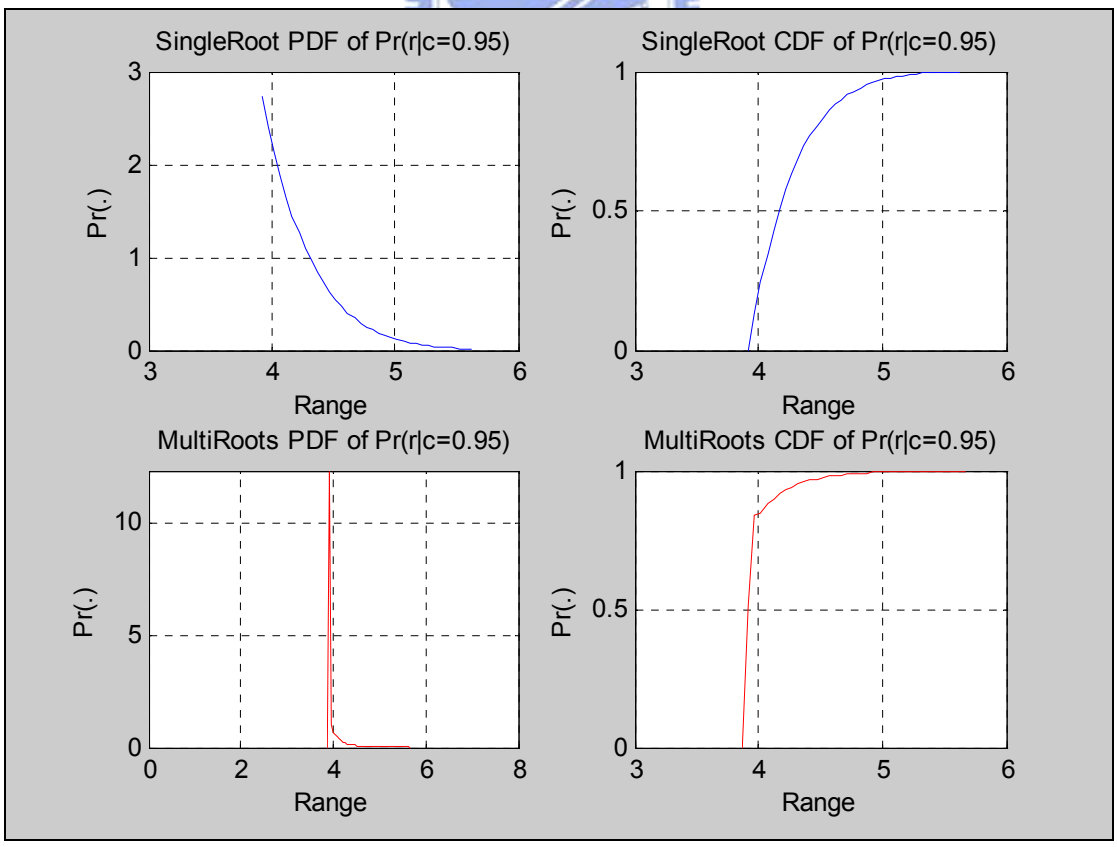




圖表 34 p(r|c,n)之 PDF,n=15



圖表 35 p(r|c,n)之 PDF 單根與多根之比較



圖表 36 p(r|c,n)單根與多根之比較

5.8. 條件機率 $p(x_{1:n}\,|\,r,n)$

$$p(x,x_{1:n},r,c;u,\sigma\,|\,n)=p(x;u,\sigma\,|\,x_{1:n},r,c,n)\times p(x_{1:n}\,|\,r,n)\times p(r\,|\,c,n)\times p(c\,|\,n) \tag{5.55}$$

利用條件機率的定義式子就可以求出 $p(x_{1:n}\,|\,r,n)$ ：

$$p(x_{1:n}\,|\,r,n)=\frac{p(x_{1:n},r\,|\,n)}{p(r\,|\,n)}$$

$$\Rightarrow \frac{n(n-1)f(x_{1:n})f(x_{1:n}+r)(F(x_{1:n}+r)-F(x_{1:n}))^{n-2}}{\int_{dx_{1:n}} n(n-1)f(x_{1:n})f(x_{1:n}+r)(F(x_{1:n}+r)-F(x_{1:n}))^{n-2}} \tag{5.56}$$

圖形如圖表 37，因爲形狀連續而且足構平滑，可以考慮使用 gaussian quadrature 來近似展開。Gaussian quadrature 有好幾種定義和計算方式，在此考慮 Gauss-Hermite formula 進行定積分的近似展開：

對於任意形狀連續且曲線足夠平滑的函數 $h(x)$

$$\int_a^b h(x)dx=\int_a^b \left\{e^{x^2}h(x)\right\}e^{-x^2}dx \tag{5.57}$$

$$\Rightarrow \sum_{i=1}^{m}\left\{e^{\gamma_i^2}h(\gamma_i)\right\}w_m(\gamma_i)+R_m(x)$$

γ_i :第 m 階的 Hermite polynomials 令其等於零時所有的根

$w_m(\gamma_i)$:weighting coefficient $w_m(\gamma_i)=\dfrac{2^{m-1}m!\sqrt{\pi}}{m^2\left[\text{Hermite}_{m-1}(\gamma_i)\right]^2}$

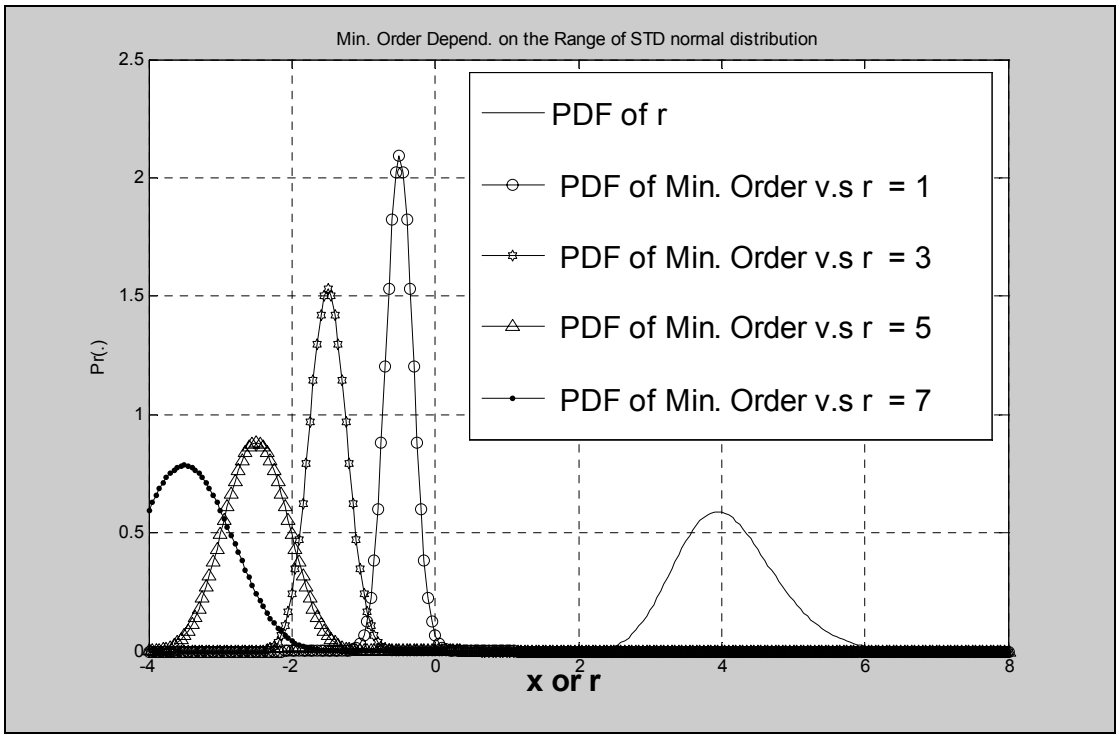
$R_m(x)$:誤差項 $R_m(x)=\dfrac{m!\sqrt{\pi}}{2^m(2m)!}\dfrac{\partial^{2m}}{\partial x^{2m}}h(x)$

(5.58)

$$p(x_{1:n} \mid r,n) = \frac{p(x_{1:n}, r \mid n)}{p(r \mid n)}$$

$$\Rightarrow \frac{n(n-1) f(x_{1:n}) f(x_{1:n}+r)(F(x_{1:n}+r) - F(x_{1:n}))^{n-2}}{\sum_{i=1}^{m} \left\{ e^{\gamma_i^2} h(\gamma_i) \right\} w_m(\gamma_i)} \quad (5.59)$$

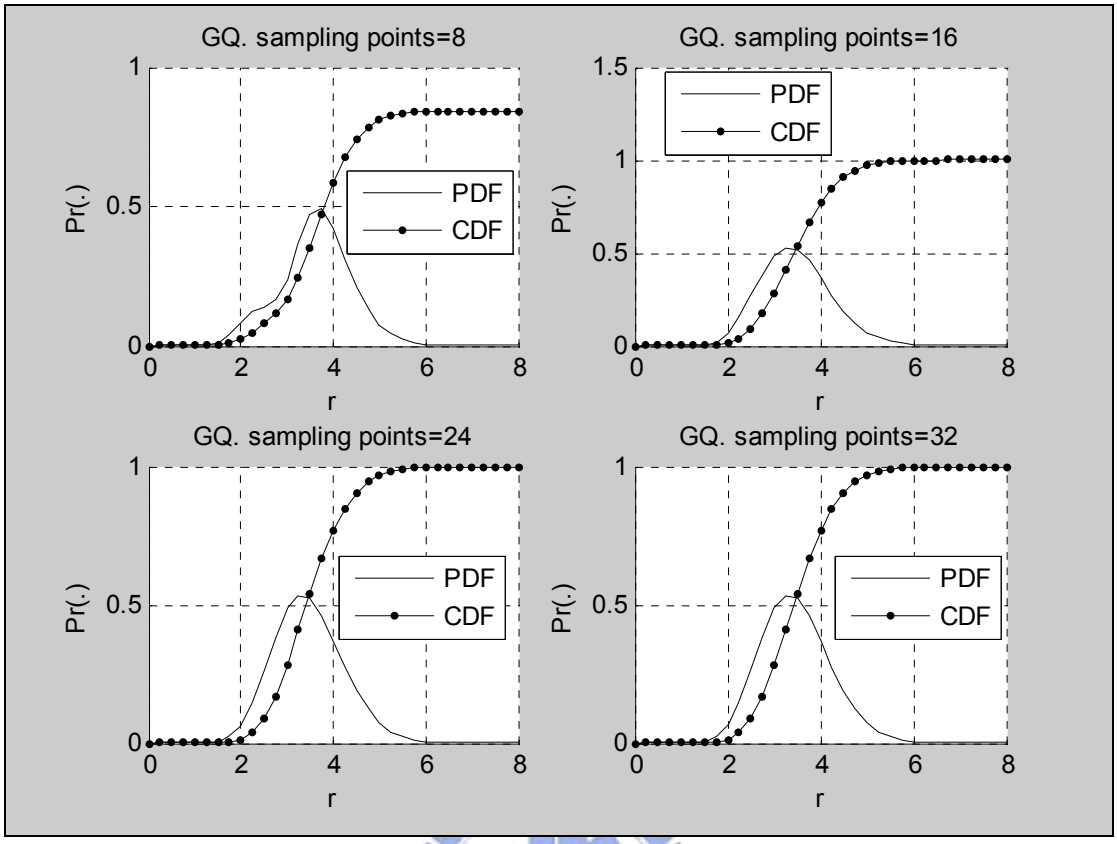
圖表 37 是 $p(x_{1:n} \mid r,n)$ 在 n=15，不同的 r 值圖形，最右方是原來 range 的機率分佈。



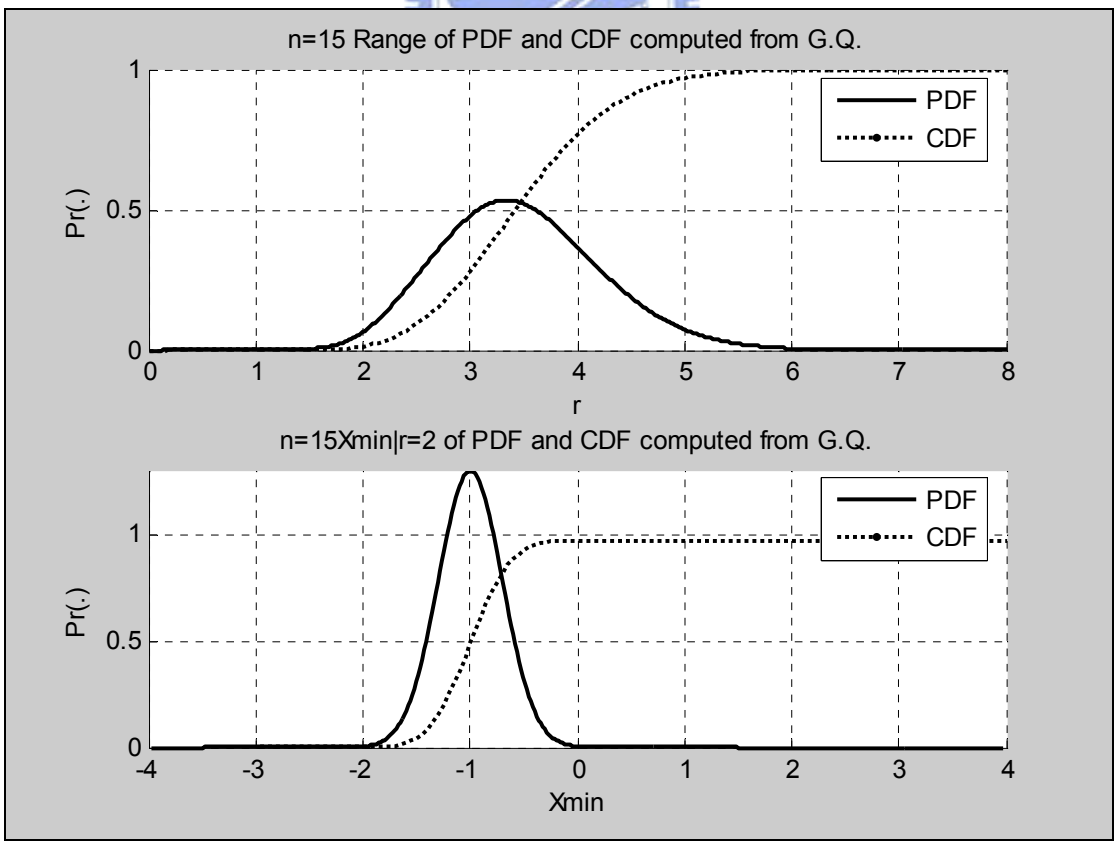
圖表 37 Xmin dependent on r and n

圖表 38 是測試不同的取樣數量之下，使用 gauss quadrature 近似 ranage 的機率密度函數及累積機率函數。

圖表 39 則是 n=15，使用 gauss quadrature 近似 $p(r), p(x_{1:n} \mid r,n)$ 兩個機率密度函數的結果。從以上的結果可以看出來，只要取樣的數量足夠，使用 gauss quadrature 來近似機率密度函數所產生的誤差幾乎可以忽略不計。



圖表 38 Gaussian quadrature 積分取樣數量測試一



圖表 39 Gaussian quadrature 積分取樣測試二

最後的結果

$$p(x, x_{1:n}, r, c = Cc \mid n) =$$

$$\{f(x)\left[\frac{UnitStep(x - x_{1:n}) - UnitStep(x - x_{1:n} - r)}{c}\right] \times$$

$$\left[\frac{f(x_{1:n}) f(x_{1:n} + r)\left\{F(x_{1:n} + r) - F(x_{1:n})\right\}^{n-2}}{\sum_{i=1}^{m}\left\{\left[e^{\gamma_i^2} f(\gamma_i) f(\gamma_i + r)\left\{F(\gamma_i + r) - F(\gamma_i)\right\}^{n-2}\right] w_m(\gamma_i)\right\}}\right] \times$$

$$\left(\sum_{j=1}^{k}\left\{\frac{f(\eta_j) f(\eta_j + r)\left\{F(\eta_j + r) - F(\eta_j)\right\}^{n-2}}{\left|f(\eta_j + r) - f(\eta_j)\right|}\right\}\right) \times$$

$$\frac{1}{Z(r, c, n)} n(n-1)(c^{n-2} - c^{n-1}))\} \tag{5.60}$$

x：常態分佈隨機變數

n：樣本大小

$x_{1:n}$：從常態分佈之中隨機抽取 n 個變數，經過排序在最小位置的隨機變數

$f(x)$：x之機率密度函數

$F(x)$：x之累積機率函數

$r = x_{n:n} - x_{1:n}$，全距隨機變數

$c = F(x_{n:n}) - F(x_{1:n})$，覆蓋率，全距範圍下之累積機率值，亦爲一隨機變數

Cc：覆蓋率c的某個固定常數值

γ_i:第 m 階的 Hermite polynomials 令其等於零時所有的根

$w_m(\gamma_i)$:weighting coefficient $w_m(\gamma_i) = \dfrac{2^{m-1} m! \sqrt{\pi}}{m^2 \left[\text{Hermite}_{m-1}(\gamma_i)\right]^2}$

η_j：$F(\eta_j + r) - F(\eta_j) - Cc = 0$的根，$\eta_j$須滿足以下限制式（假設有 k 個滿足）

$f(\eta_j + r) - f(\eta_j) \neq 0$ and $\eta_j \in \mathbb{R}$ *and* η_j 不得為重根

$\{\eta_{j,j=1,2\cdots k} \mid F(\eta_j + r) - F(\eta_j) - Cc = 0 \wedge lb \leq \eta_j \leq ub$

$\wedge 0 \leq r \leq |(ub - lb)| \wedge lb \leq \eta_j + r \leq ub\}$

lb $\Rightarrow (-4\sigma)$:常態分佈的實際合理下界

ub $\Rightarrow (4\sigma)$:常態分佈的實際合理上界

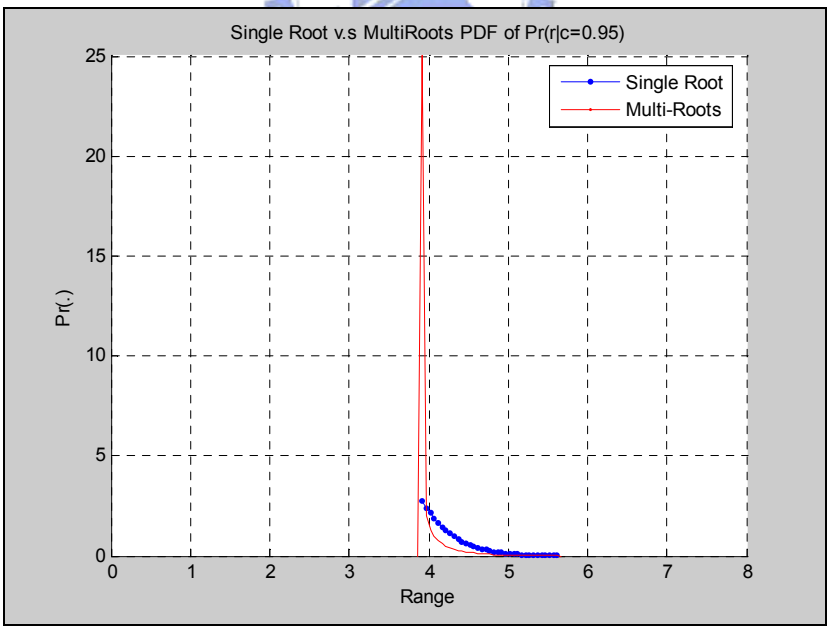
σ:標準差

$$Z(r,c,n) = \int_{dr} \left[\sum_{j=1}^{k} \{ \frac{f(\eta_j) f(\eta_j + r) \left\{ F(\eta_j + r) - F(\eta_j) \right\}^{n-2}}{\left| f(\eta_j + r) - f(\eta_j) \right|} \} \right]$$

$Z(r,c,n)$就是計算步驟C中，聯合機率 $p(r, c = Cc \mid n)$曲線下的面積。

5.9. 組合切片，進行區間估計

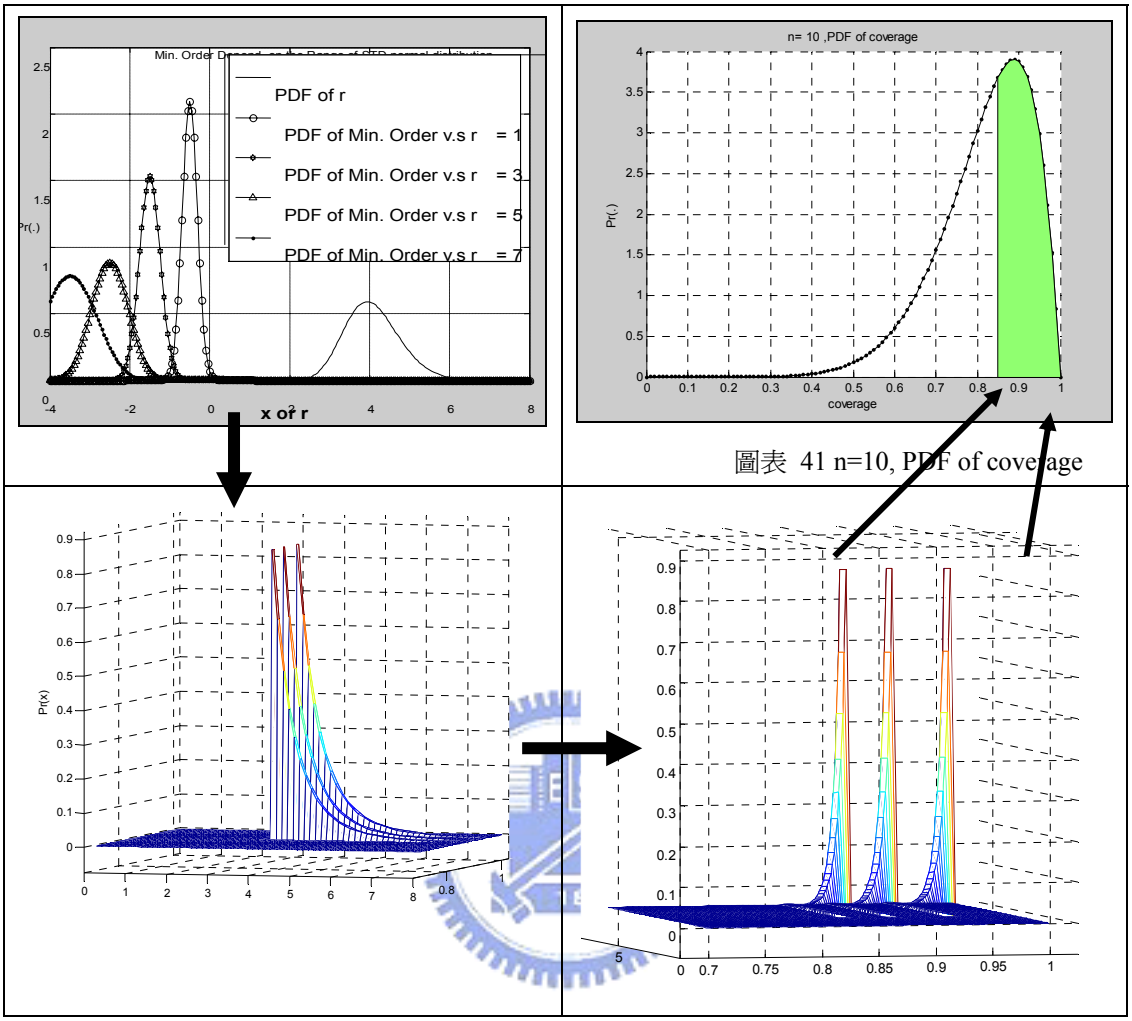
(5.60)只是一個固定覆蓋率(coverage)下的點估計(point estimation)效應。如果我們想要針對不同的覆蓋率進行區間估計(interval estimation)時，就必須計算出不同的點估計效應，然後進行組合推估。



圖表 40 p(r|c,n)之 PDF,n=15

每一個切片的剖面圖都是類似圖表 40，所以應該將有興趣估計的覆蓋率區間進行覆蓋率取樣，以獲得最佳的估計結果。表格 8 就是整個從點估計到區間估計的執行過程。

表格 8 切片組合之影響路徑



圖表 41 n=10, PDF of coverage

在表格 8 中的圖表 41 中，覆蓋率大於 0.85~1 之間的範圍是我們對輸入語者資料的覆蓋率有興趣的範圍。其下方的雙箭頭區間表示相對應的範圍區間正在使用切片進行組合。下方左邊的圖型是圖表 34 的正視圖。

由本節的分析可以知道，覆蓋率的取樣多寡將會影響整個聯合機率的估計結果。因爲覆蓋率的分佈是一個連續函數，所以比較好的處理方式，應該是將有興趣的區間分散成很多個和切片對照的小區間來計算機率值。

5.10. 再一次使用 gaussian quadrature

在 5.8 中，本研究曾經使用 hermite-gauss 的數值積分演算法來近似全距(range)的分佈，主要的用意是希望避開積分符號。這裡考慮在度使用 gaussian quadrature；會有這樣的想法，主要是因爲現在所面對的課題正是切片數量與精確度之間的關係。究竟要算出多少片的切片來進行組合，才能精準的近似到我們所能夠容忍的誤差範圍以內？

■ Gauss-Legendre Integration

考慮一個任意區間上的定積分$\int_a^b h(x)dx$，Gauss-Legendre 積分的主要應用是將這個積分轉到$[-1,1]$的區間上，使用 Legendre 正交多項式進行展開。

$$Legendre\ polynomial \Rightarrow P_v(x) = \frac{1}{2^v v!}\frac{\partial^v}{\partial x^v}(x^2-1)^v, v = 0,1,2\cdots \tag{5.61}$$

(5.61)是一個 v 階的 Legendre polynomial 公式。在區間$[-1,1]$上會形成一個 complete orthogonal set。其轉換進行式如下：

$$\int_a^b h(x)dx = \int_{-1}^1 h(\frac{b-a}{2}\xi + \frac{b+a}{2})\frac{(b-a)}{2}d\xi$$
$$\Rightarrow \frac{b-a}{2}\int_{-1}^1 g(\xi)d\xi = \frac{b-a}{2}\sum_{\tau=1}^{v} w_v(\xi_\tau)g(\xi_\tau) + R_v(\xi) \tag{5.62}$$

$$\Rightarrow \frac{b-a}{2}\sum_{\tau=1}^{v} w_v(\xi_\tau)h(\frac{b-a}{2}\xi_\tau + \frac{b+a}{2}) + R_v(\xi) \tag{5.63}$$

$$x = \frac{b-a}{2}\xi + \frac{b+a}{2}, -1 < \xi < 1$$

ξ_τ : the τth root of $P_v(x)$

w_v : weight coefficient $\Rightarrow \dfrac{(b-a)}{(1-\xi_\tau^2)(P_v'(\xi_\tau))^2}$

$$g(\xi) = h(\frac{b-a}{2}\xi + \frac{b+a}{2})$$

$R_v(\xi)$: error term $\Rightarrow \dfrac{2^{(2v+1)}(v!)^4}{(2v+1)((2v)!)^3}g^{(2v)}(\xi)$

以下，本研究用實際的例子比較 Gauss-Legendre integration 和一般積分近似法之間的效率。以下是實際的標準常態分佈計算$[-4,4]$之間的累積機率，使用 Gauss-Legendre 積分，取樣點數和積分近似結果。

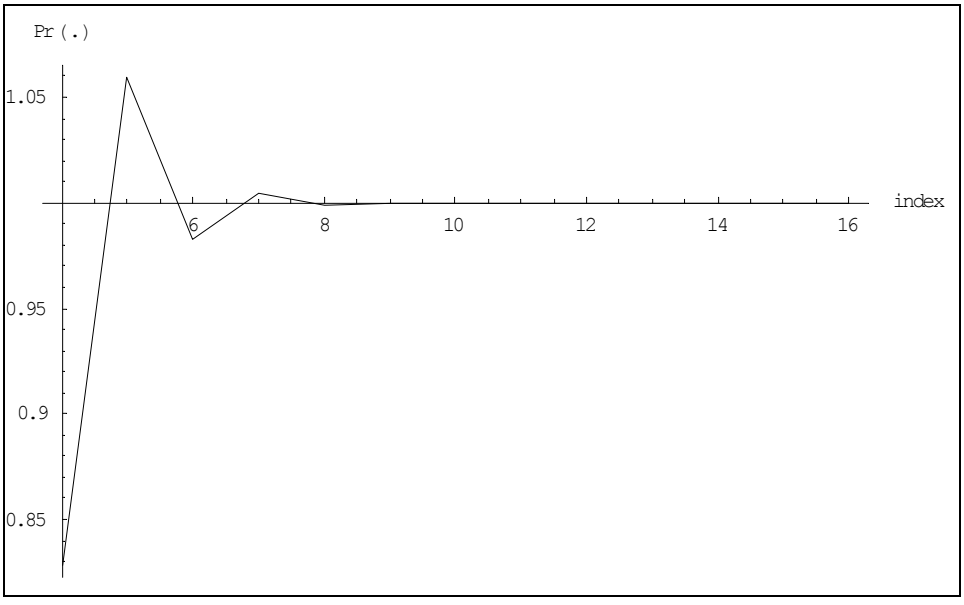
表格 9 Gauss-Legendre 積分結果分析

{取樣點數，近似結果}	取樣點數=13， {取樣位置 x_τ ，權重值} $(w_v(\xi_\tau))$ }
{{4, 0.8285135869455158}, {5, 1.059056043914255}, {6, 0.9823506486292622}, {7, 1.0045374967517489}, {8, 0.9988645254752475}, {9, 1.0001616381338159}, {10, 0.9998937524987197}, {11, 0.99994415164309}, {12, 0.9999354505731359}, {13, 0.9999368377840447}, {14, 0.9999366324202007}, {15, 0.9999366607873392}, {16, 0.999936657115609}}	$\begin{pmatrix} -3.93673 & 0.161936 \\ -3.67039 & 0.368486 \\ -3.20631 & 0.555494 \\ -2.5694 & 0.712584 \\ -1.79397 & 0.831264 \\ -0.921833 & 0.905133 \\ 0. & 0.930206 \\ 0.921833 & 0.905133 \\ 1.79397 & 0.831264 \\ 2.5694 & 0.712584 \\ 3.20631 & 0.555494 \\ 3.67039 & 0.368486 \\ 3.93673 & 0.161936 \end{pmatrix}$

$$x_\tau = \frac{b-a}{2}\xi_\tau + \frac{b+a}{2}, -1 < \xi < 1$$

ξ_τ : the τth root of $P_v(x)$

$w_v(\xi_\tau)$: weight coefficient $\Rightarrow \dfrac{(b-a)}{(1-\xi_\tau^2)(P_v'(\xi_\tau))^2}$



圖表 42 Gauss-Legendre 取樣積分的結果

由上述的結果分析可以了解，使用 Gauss-Legendre 進行取樣數值積分，只要大約取 12 個樣本點就可以得出近似完美的結果（積分結果趨近於 1）。

應用 Gauss-Legendre 積分的概念，我們可以成功的使用組合切片的概念來拼湊出原來的機率密度函數。例如，我們如果想要估計在固定的樣本數量 n=15 的情況之下，覆蓋率大於 0.85 以上的成本函數，其整體操作程序如下：

表示式

$$\int_{0.85}^{1}\iiint p(x,x_{1:n},r,c\,|\,n)\cdot Cost(x,x_{1:n},r,c,n)\,dxdx_{1:n}drdc \quad (5.64)$$

■ 首先計算出切片的位置

考慮使用 16 點的 Gauss-Legendre 積分，覆蓋率(c)的機率密度函數爲：

$$\Rightarrow p(c)=n(n-1)\left(c\right)^{n-2}(1-c),c>0 \quad (5.65)$$

則 16 個切片的近似結果：

```
{{1, 0.8574609342373792},
 {2, 0.6910463328813958},
 {3, 0.6814941099154919},
 {4, 0.6814142104225789},
 {5, 0.6814140245862208},
 {6, 0.6814140244542831},
 {7, 0.6814140244542619},
 {8, 0.6814140244542567},
 {9, 0.681414024454257},
 {10, 0.681414024454255},
 {11, 0.6814140244542067},
 {12, 0.681414024454209},
 {13, 0.6814140244543205},
 {14, 0.6814140244539989},
 {15, 0.6814140244548879},
 {16, 0.6814140244553407}}
```

由上述的結果可以臆測，其實只要切四片即可；所以此時可以立即修正爲只切四片：

位置和權重值分別爲

表格 10 覆蓋率>0.85，n=15 時的最佳切片位置

Cc_t	位置	權重值
t=1	0.860415	0.0260891
t=2	0.899501	0.0489109
t=3	0.950499	0.0489109
t=4	0.989585	0.0260891

(5.64)的估計式子現在可以表示成爲：

$$\int_{0.85}^{1}\iiint p(x,x_{1:n},r,c\,|\,n)\cdot Cost(x,x_{1:n},r,c,n)\,dxdx_{1:n}drdc$$

$$\Rightarrow\sum_{t=1}^{4}\iiint p(x\,|\,x_{1:n},r,c,n)p(x_{1:n}\,|\,r,n)p(r\,|\,c=Cc_t,n)p(c=Cc_t)w_4(\mathrm{Cc_t})$$

$$Cost(x,x_{1:n},r,c,n)dxdx_{1:n}dr \tag{5.66}$$

接下來的結果在前面的章節提過：

$$p(x,x_{1:n},r,c=Cc\,|\,n)=$$

$$\{f(x)\left[\frac{UnitStep(x-x_{1:n})-UnitStep(x-x_{1:n}-r)}{c}\right]\times$$

$$\left[\frac{f(x_{1:n})f(x_{1:n}+r)\{F(x_{1:n}+r)-F(x_{1:n})\}^{n-2}}{\sum_{i=1}^{m}\left\{\left[e^{\gamma_i^2}f(\gamma_i)f(\gamma_i+r)\{F(\gamma_i+r)-F(\gamma_i)\}^{n-2}\right]w_m(\gamma_i)\right\}}\right]\times$$

$$\left(\sum_{j=1}^{k}\{\frac{f(\eta_j)f(\eta_j+r)\{F(\eta_j+r)-F(\eta_j)\}^{n-2}}{|f(\eta_j+r)-f(\eta_j)|}\}\right)\times$$

$$\frac{1}{Z(r,c,n)}n(n-1)(c^{n-2}-c^{n-1}))\} \tag{5.67}$$

所以(5.64)式的結果爲：

$$\int_{0.85}^{1}\iiint p(x,x_{1:n},r,c\,|\,n)\cdot Cost(x,x_{1:n},r,c,n)\,dxdx_{1:n}drdc$$

$$\Rightarrow\sum_{t=1}^{4}\iiint p(x\,|\,x_{1:n},r,c,n)p(x_{1:n}\,|\,r,n)p(r\,|\,c=Cc_t,n)p(c=Cc_t)w_{P4}(\mathrm{Cc_t})$$
$$Cost(x,x_{1:n},r,c,n)dxdx_{1:n}dr$$

(5.68)

$$\Rightarrow\sum_{t=1}^{4}\iiint f(x)\left[\frac{UnitStep(x-x_{1:n})-UnitStep(x-x_{1:n}-r)}{Cc_t}\right]\times$$
$$\left[\frac{f(x_{1:n})f(x_{1:n}+r)\left\{F(x_{1:n}+r)-F(x_{1:n})\right\}^{n-2}}{\sum_{i=1}^{m}\left\{\left[e^{\gamma_i^2}f(\gamma_i)f(\gamma_i+r)\left\{F(\gamma_i+r)-F(\gamma_i)\right\}^{n-2}\right]w_{Hm}(\gamma_i)\right\}}\right]\times$$
$$\left(\sum_{j=1}^{k}\left\{\frac{n(n-1)f(\eta_j)f(\eta_j+r)\left\{F(\eta_j+r)-F(\eta_j)\right\}^{n-2}}{\left|f(\eta_j+r)-f(\eta_j)\right|}\right\}\right)\times$$
$$\frac{1}{Z(r,Cc_t,n)}n(n-1)(Cc_t^{\,n-2}-Cc_t^{\,n-1})\times$$
$$w_{P4}(\mathrm{Cc_t})\,Cost(x,x_{1:n},r,c,n)dxdx_{1:n}dr$$

(5.69)

x：常態分佈隨機變數

n：樣本大小

$x_{1:n}$：從常態分佈之中隨機抽取 n 個變數，經過排序在最小位置的隨機變數

$f(x)$：x之機率密度函數

$F(x)$：x之累積機率函數

$r=x_{n:n}-x_{1:n}$，全距隨機變數

$c=F(x_{n:n})-F(x_{1:n})$，覆蓋率，全距範圍下之累積機率值，亦爲一隨機變數

Cc_t：覆蓋率c的某個固定常數值

γ_i:第 m 階的 Hermite polynomials 令其等於零時所有的根

$w_{Hm}(\gamma_i)$:**weighting coefficient for the m-th order of i-th root of Hermite**

polynomial $\Rightarrow w_{Hm}(\gamma_i)=\dfrac{2^{m-1}m!\sqrt{\pi}}{m^2\left[\text{Hermite}_{m-1}(\gamma_i)\right]^2}$

η_j：$F(\eta_j+r)-F(\eta_j)-Cc=0$的根，$\eta_j$須滿足以下限制式（假設有 k 個滿足）

$f(\eta_j+r)-f(\eta_j)\neq 0$ and $\eta_j\in\mathbb{R}$ *and* η_j 不得為重根

$\{\eta_{j,j=1,2\cdots k}\mid F(\eta_j+r)-F(\eta_j)-Cc=0\wedge lb\le\eta_j\le ub$

$\wedge 0\le r\le\left|(ub-lb)\right|\wedge lb\le\eta_j+r\le ub\}$

lb $\Rightarrow$ (-4σ):常態分佈的實際合理下界

ub $\Rightarrow$ (4σ):常態分佈的實際合理上界

σ:標準差

$$Z(r,Cc_t,n)=\int_{dr}\left[\sum_{j=1}^{k}\{\frac{n(n-1)f(\eta_j)f(\eta_j+r)\left\{F(\eta_j+r)-F(\eta_j)\right\}^{n-2}}{\left|f(\eta_j+r)-f(\eta_j)\right|}\}\right]$$

$Z(r,Cc_t,n)$就是計算步驟C中，聯合機率$p(r,c=Cc_t\mid n)$曲線下的面積。

$$x_\tau=\frac{b-a}{2}\xi_\tau+\frac{b+a}{2},-1<\xi_\tau<1$$

ξ_τ : the τth root of $P_v(x)$

$w_{Lv}(\xi_\tau)$: weight coefficient of the v-th order of τ-th root of

Legendre polynomial $\Rightarrow\dfrac{(b-a)}{(1-\xi_\tau^2)(P_v'(\xi_\tau))^2}$

$[a,b]$：我們想要估計的覆蓋率區間範圍

$$P_v(x)(Legendre\ polynomial)\Rightarrow\frac{1}{2^v v!}\frac{\partial^v}{\partial x^v}(x^2-1)^v,v=0,1,2\cdots$$

6. 實驗設計

首先將 TCC300 語料分成四個部份，分別是當事人(client)，背景模型(world model,UBM or Cohort)，還有僞裝者(imposter)以及部分的語料作爲系統參數訓練使用。

Client set:用來計算每個語者的個別 GMM 模型

World set: 使用來計算 UBM (Universal background model)或同隊(Cohort)模型

Imposter set: 作爲僞裝者，用來測試 client model

Development set：使用作爲系統其他參數推估

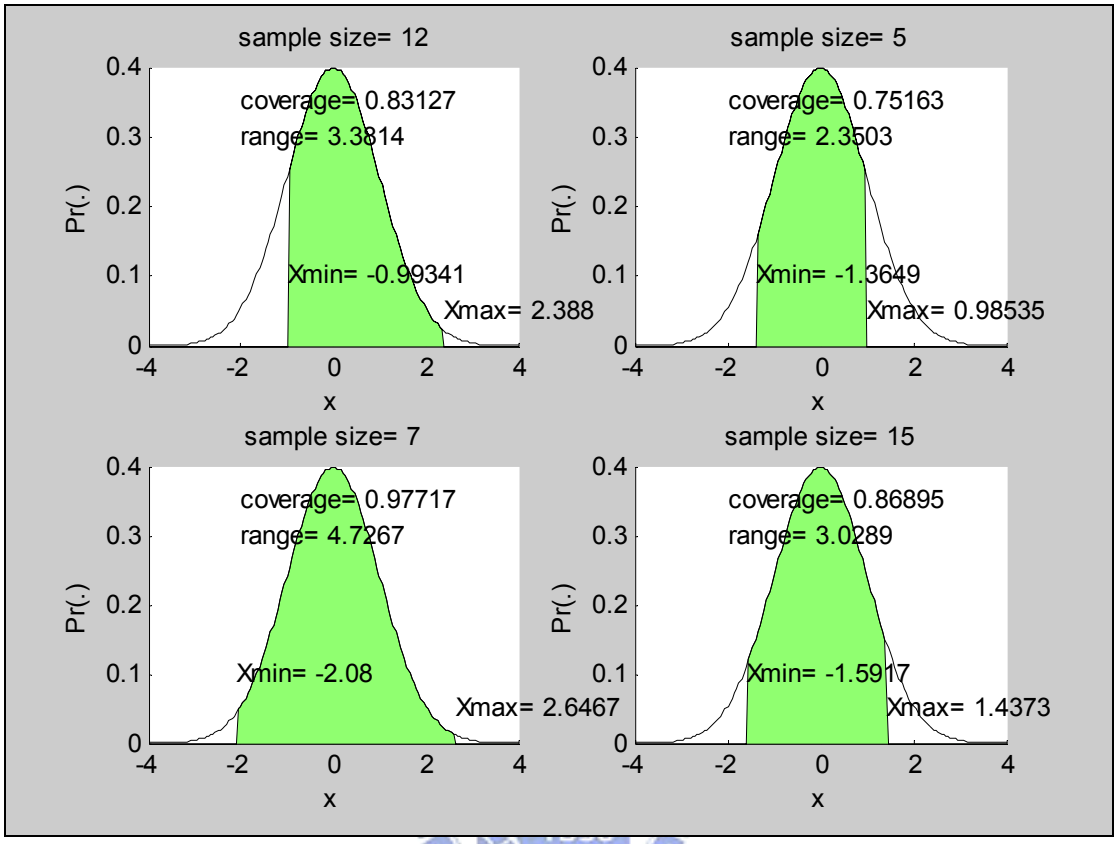
測試主題一

基本的實驗參數結果比較測試

測試主題二：

使用統計上的假設檢定方式，以 soft decision 的形式來進行可靠度分析，並且和以往的 hard decision 結果進行比較。

6.1. 稀少資料的隨機分佈現象



圖表 43 模型之參數變化

由上圖的結果知道，當樣本數量稀少時，實際上數量比較少的樣本組其覆蓋率不一定就會比較少；只是樣本數量比較少的組別其覆蓋率大於樣本比較大的組別的機率相對而言會比較小，這個現像就是本研究所推導描述的聯合機率的內容。

測試主題三：

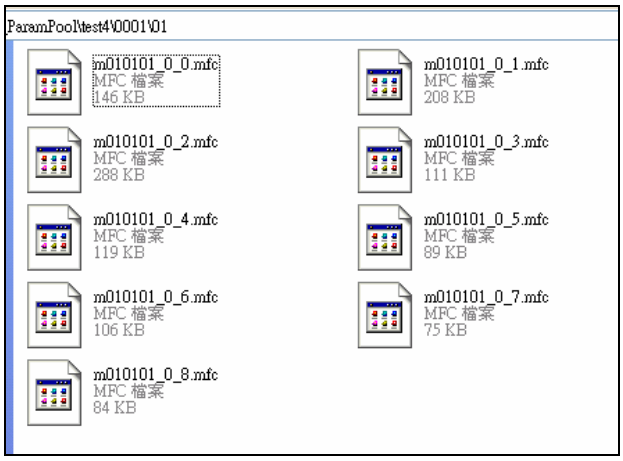
如果已知受檢驗的隨機變數是一個常態分佈，在同時擁有變數集合本身 $\boldsymbol{x}$，集合最小值 $\boldsymbol{x}_{1:n}$ or $\boldsymbol{x}_{\min}$，集合分佈區間大小 $\boldsymbol{range}$，以及集合分佈區間所佔之總體覆蓋率 $\mathbf{coverage}$ 和樣本大小 $\boldsymbol{n}$ 的聯合機率分佈時，是否可可以有效地提升辨識率？

6.2. 實驗環境設定

我們想進行語者確認之分析，使用交通大學電信工程學系語音實驗室 TCC300 語料庫進行分析。
語者確認之型式：text independent

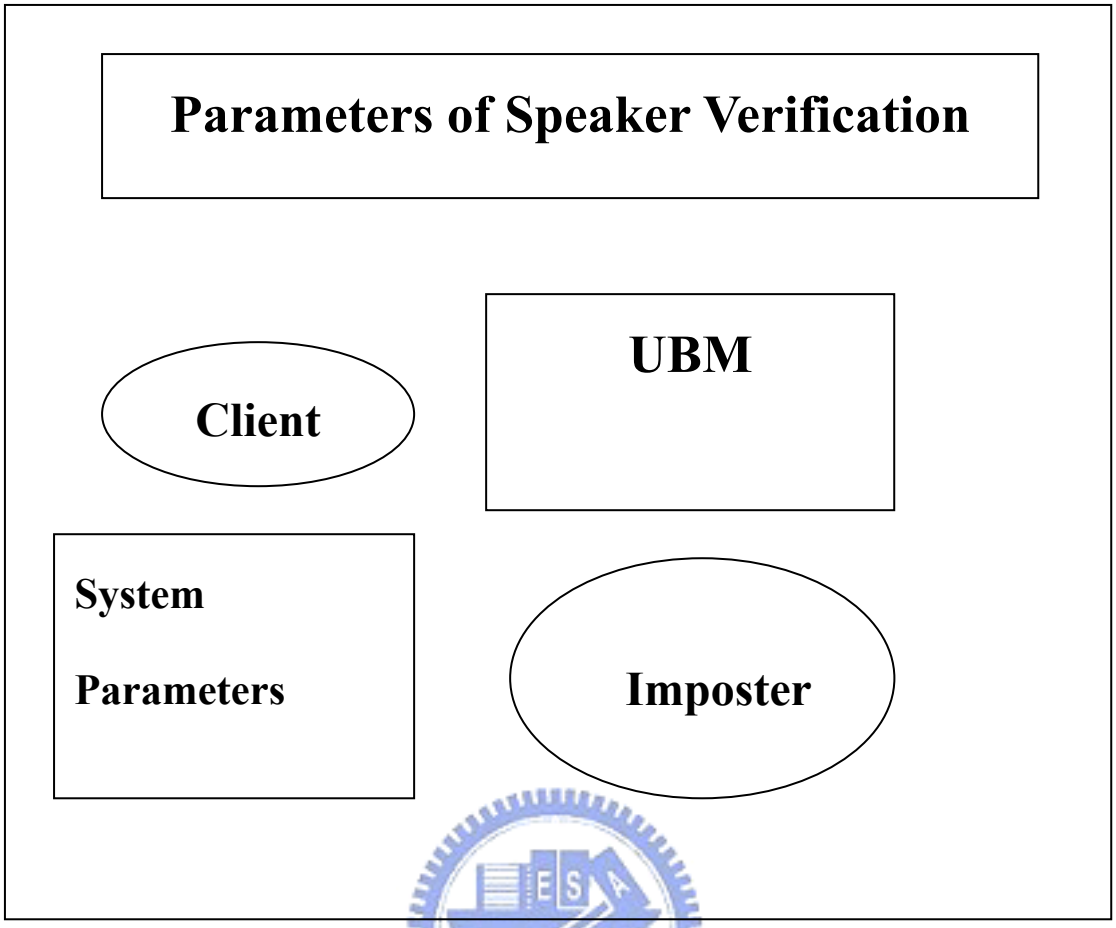
使用分析工具：HTK(Hidden Markov Model Toolkit)工具程式及其相關之MASV shareware(**M**unich **A**utomatic **S**peaker **V**erification)進行小規模程式改寫。

實驗假設情況:我們想使用 TCC300 語料庫進行語者確認之研究,以句子(sentence)作爲分析個體。但是 TCC300 每位語者之語料發音多在 20 句之內，符合本研究稀少性輸入資料之假設。



圖表 44 以句子爲觀察單位，則每位語者的資料量是稀少的

首先，修該第一章之假設，將所有的語料以語者作爲單位，分成四個互相獨立的集合。



圖表 45 語者確認上之參數蒐集分類

各個方塊之參數使用如下：

一、UBM(Universal background model or World model)集合內的語者，所有的語者資料訓練一個 GMM 模型。

二、Client 集合內的語者，每位語者使用自己的語料訓練屬於自己的 GMM 模型。

三、System parameters:用來求取系統參數使用，HER(half error rate), ERR(equal error rate)。

四、Imposter 集合內的語者，不需要訓練任何模型。

五、假設所有的語者識別只使用一個系統臨界值(threshold)來進行判定。

符號：

假設 client 語者之中，第 c_i 位 speaker 需要被確認，則第 c_i 個 GMM client 模型。

記爲 $\hat{s}_{c_i}$ ， x_{c_i} 表示來自第 c_i 位 client 語者的語料。

有 k 位 imposter 預備進行僞裝測試，x_{I_k} 表示第 k 位 imposter 的語料。

UBM 模型記爲Ω

假設 GMM 模型的輸出取對數運算之後會服從高斯分佈，這個高斯分佈隨機變數我們稱它爲分數(score)。定義下列二種判別測試：

一自我判別測試(self testing)：

$$\bar{J}(x_{c_{i,j}}) = \frac{1}{d_{i,j}^{(c|\hat{s}_{c_i})}} \sum_{\lambda=1}^{j_n} \log(p(x_{c_{i,j}}(\lambda) \mid \hat{s}_{c_i})) - \frac{1}{d_{i,j}^{(c|\Omega)}} \sum_{\lambda=1}^{j_n} \log(p(x_{c_{i,j}}(\lambda) \mid \Omega)) - \log \Lambda \quad (6.1)$$

$\bar{J}(x_{c_{i,j}})$：自我判別測試函數，$\begin{cases} \bar{J}(x_{c_{i,j}}) > 0, x_{c_{i,j}} \in client \\ \bar{J}(x_{c_{i,j}}) < 0, x_{c_{i,j}} \in imposter \end{cases}$

$x_{c_{i,j}}$：屬於 client 中的第 i 位 speaker，第 j 句語料發音。

$x_{c_{i,j}}(\lambda)$：client 集合中，第 i 位 speaker 所說的第 j 句話，其中的第λ個 phone 或 syllable(視分析單元大小而定)，$\lambda = 1,2 \cdots j_n$

j_n：$x_{c_{i,j}}$所包含的 phone 總數量

$\hat{s}_{c_i}$：client 集合中，第 i 位 speaker 的 GMM 模型

$\log(p(x_{c_{i,j}}(\lambda) \mid \hat{s}_{c_i}))$：client 集合中，第 i 位 speaker 所說的第 j 句話，其中的第λ個 phone 對應到自己的模型$\hat{s}_{c_i}$所獲得的 score。

$d_{i,j}^{(c|\hat{s}_i)}$：來自於 client 集合中的第 i 位語者所說的第 j 句話，對應到自己的 $\hat{s}_{c_i}$ GMM 模型之下進行 force alignment，該句話中的所有 phones 所單獨對應的 duration 總和。

$\log(p(x_{c_{i,j}}(\lambda)|\Omega))$：client 集合中，第 i 位 speaker 所說的第 j 句話，其中的第 λ 個 phone 對應到 UBM 模型 Ω 所獲得的 score。

$d_{i,j}^{(c|\Omega)}$：來自於 client 集合中的第 i 位語者所說的第 j 句話，對應到 UBM Ω 的 GMM 模型下進行 force alignment，該句話中的所有 phones 所單獨對應的 duration 總和。

Λ：threshold

表格 11 是 client 集合中，編號 0001 的語者對自己的 16mixtures GMM 模型進行 force alignment 的結果。分割單元使用的是 syllable，因爲執行的語者確認形式爲 text independent，所以所有的 syllable 都 tie 成一個名稱爲 gmmstate 的 GMM 模型。

➔終止時間減去起始時間就是該 syllable 所對應的 duration。

由實驗的結果可以觀察得出，第一個 syllable 與其它的 syllable 分數互相差距過大，故每句話的第一個 syllable 分數都予以忽略不計入平均計算。

表格 11Client 對自己的模型 force alignment 的結果

起始時間	終止時間	GMM 名稱	分數(Score)
0	91099997	gmmstate	-50202.347656
91100000	91200000	gmmstate	-54.636047
91200000	91300000	gmmstate	-53.327332
91300000	91400000	gmmstate	-56.178677
91400000	91500000	gmmstate	-52.698994
91500000	91600000	gmmstate	-54.373470
91600000	91700000	gmmstate	-55.059395
91700000	91800000	gmmstate	-56.624462
91800000	91900000	gmmstate	-52.812164
91900000	92000000	gmmstate	-56.019474
92000000	92100000	gmmstate	-54.247986
92100000	92200000	gmmstate	-58.101490

92200000	92300000	gmmstate	-54.202251
92300000	92400000	gmmstate	-56.527683
92400000	92500000	gmmstate	-58.027462
92500000	92600000	gmmstate	-57.978786
92600000	92700000	gmmstate	-57.719040

.相同於上表格的內容，client 對於 UBM 模型也會得出如表格 11 的 force alignment 結果。

將所有的資料代入公式(6.1) 中，計算出$\bar{J}(x_{c_{i,j}})$，它的意義爲該句話在平均每個 frame 之下所獲得的的相對分數（因爲式子中減去 UBM 模型和$\log\Lambda$的分數，所以稱此時的分數爲相對分數）。

如果先暫時假設 threshold=0，則此時所計算出來的相對分數由小到大依序爲：

表格 12 自我判別分數統計

自我判別分數	相對應語句編號
-0.5132	#0
-0.5074	#2
-0.3142	#3
-0.2562	#7
-0.1588	#6
-0.0988	#5
-0.0704	#1
0.1915	#4
0.3059	#8

表格 12 的內容如果判讀出現錯誤時，此時 false rejection 就會發生。

定義二：

僞裝者判別測試(Imposter testing)

$$\bar{J}(x_{I_{k,j}}) = \frac{1}{d_{k,j}^{(I|\hat{s}_{c_i})}} \sum_{\lambda=1}^{j_\tau} \log(p(x_{I_{k,j}}(\lambda)\,|\,\hat{s}_{c_i})) - \frac{1}{d_{k,j}^{(I|\Omega)}} \sum_{\lambda=1}^{j_\tau} \log(p(x_{I_{k,j}}(\lambda)\,|\,\Omega)) - \log\Lambda \tag{6.2}$$

$x_{I_{k,j}}$：來自於僞裝者集合(imposter set)中，第 k 位 speaker，所說的第 j 句話。

其餘的變數意義同(6.1)的解釋。

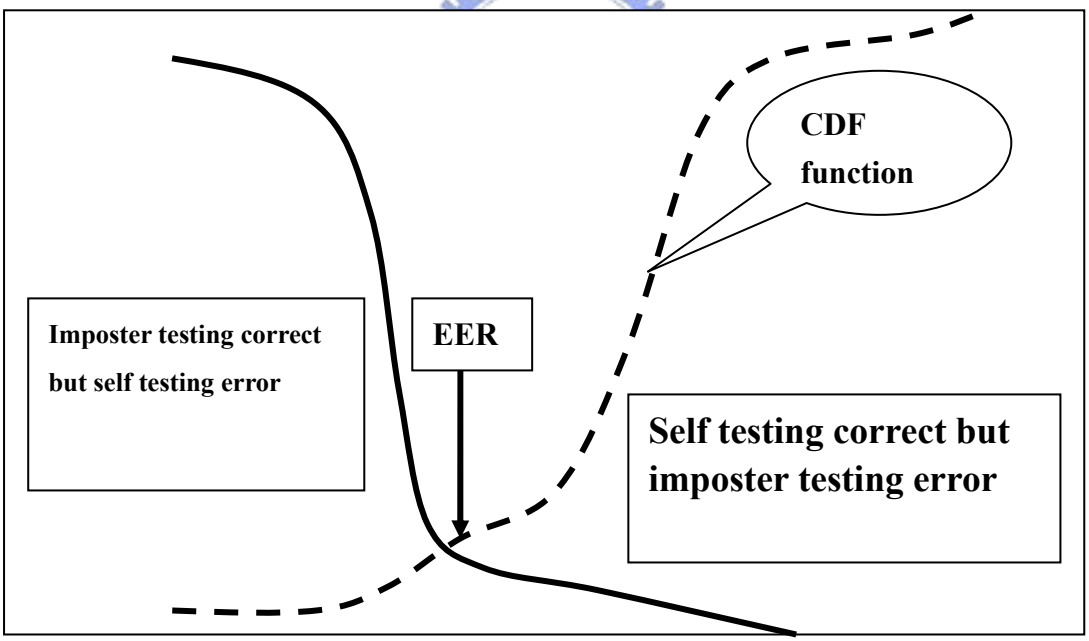
定義二之僞裝者判別測試是相對於定義一來描述的，如同定義一的結果，在僞裝者判別測試上，我們也可以制定出相對分數。

僞裝者判別測試和自我判別測試最大的不同點在於資料量之多寡，因爲此時我們想要一次推估出整個僞裝者(imposter)集合所可能產生的 false alarm rate。所以僞裝者在語者確認的處理上通常都是以集合型態出現，不會面臨稀少資料量的問題。　本例中僞裝者集合的句子總數量爲 145，僞裝者判讀如果出現錯誤時，此時所犯的錯誤爲 false alarm。

6.3. 將自我判讀及僞裝者測試所得之相對分數視爲隨機分佈處理

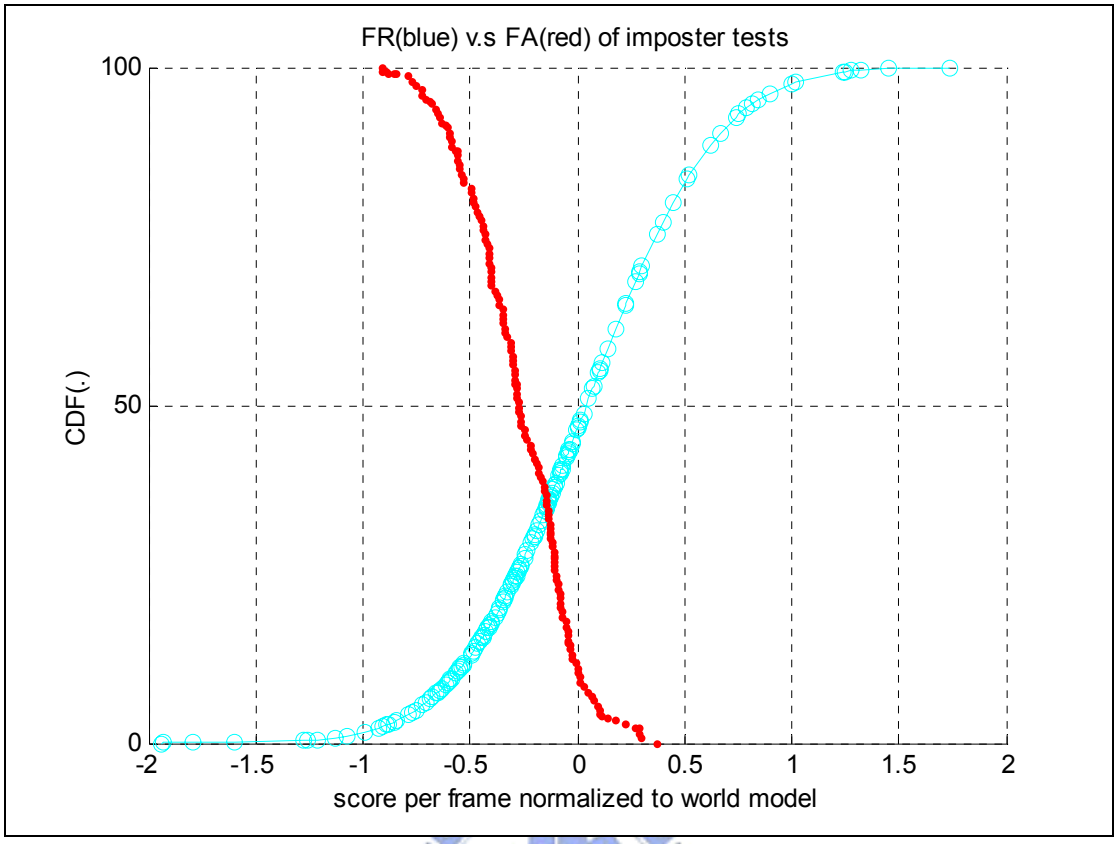
我們可以將僞裝者測試和自我判別測試視爲兩個分佈(distribution)處理。在第二章文獻回顧的部份，本研究曾經提到這種平均值（整句話中平均每個 frame 所獲得的 score）的差所形成的新隨機變數可以視爲常態分佈來處理。

語者確認上如果使用 EER(equal error rate)來作爲臨界值(threshold)選定的條件時，分佈處理可以使用累積機率(cumulative probability)來表示比較明確。



圖表 46 使用 CDF 來描述與處理語者確認之臨界值選取

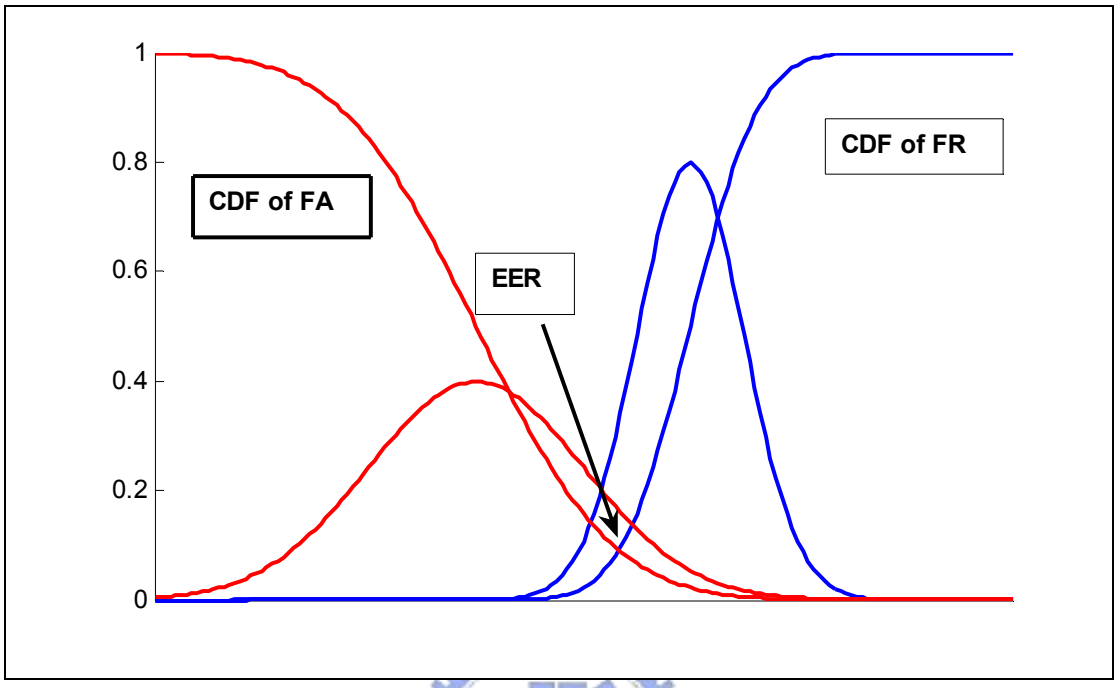
圖表 46 將兩種測試使用累積機率函數(CDF)來進行描述時，EER 的選取點就剛好是兩個 CDF 函數的交點。



圖表 47 使用 10 句話的語料進行 EER 策定之結果

圖表 47 是使用十句語料進行 EER 預測的結果。其結果並不好。

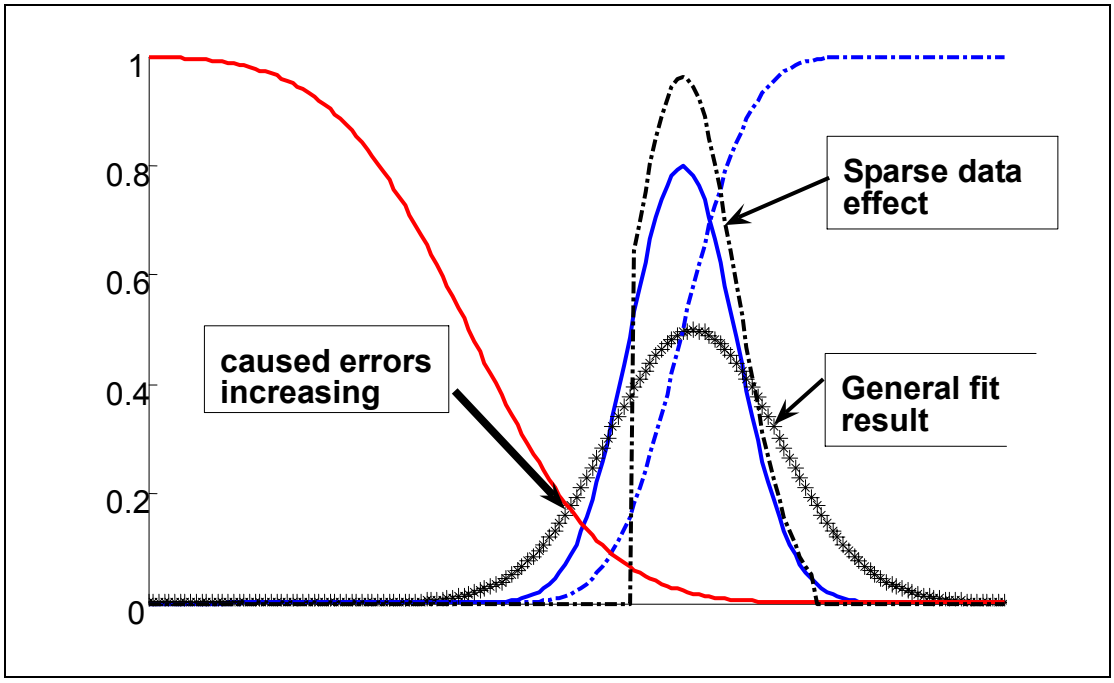
6.4. 問題的分析



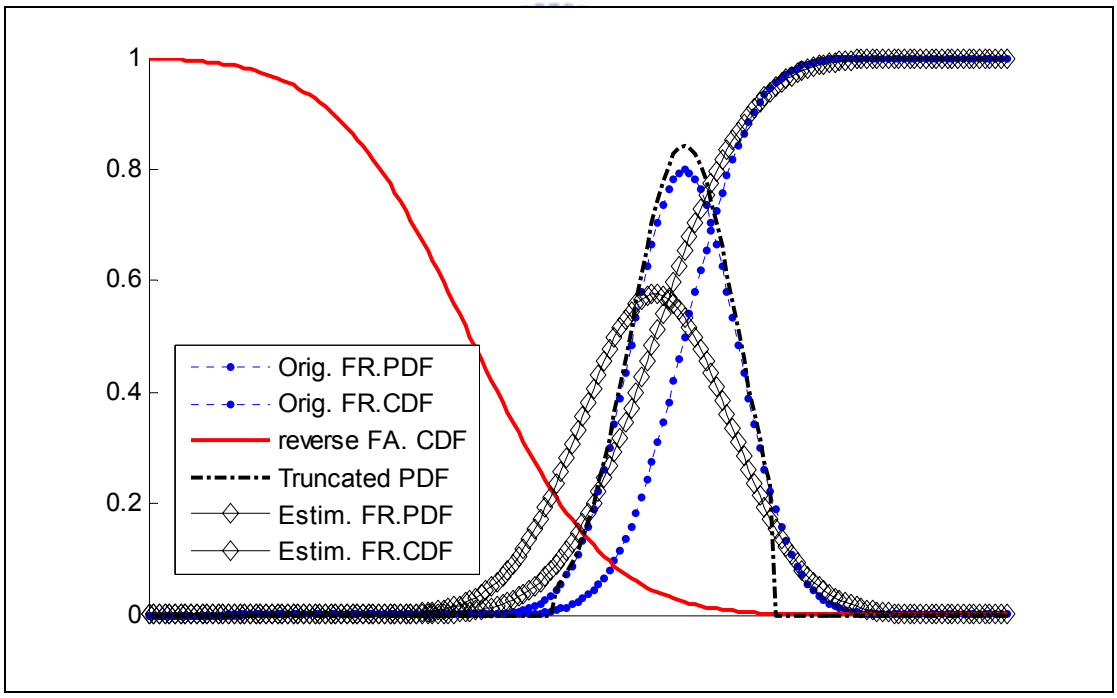
圖表 48 問題分析示意圖

圖表 48 右方藍色線條表示自我測試，左方紅色線表示僞裝者測試。累積機率函數的交點就是 EER 的最佳參考值。因爲藍色線條表示爲 client 進行 self testing 的結果，所以他的變異性比較小，使用比較陡峭的常態分佈來表示。紅色線條表示整體 imposter testing 的分佈，其變異性較大，所以使用比較平緩的常態分佈來描述。

根據本研究的結論，當輸入資料量稀少時，會引發分佈不匹配（distribution mismatch）的問題。如下圖之黑色虛線所示，這時候我們很可能會在資訊不齊全的情況下認爲底下的黑色分佈是原來的分佈。當然這時候所產生的誤認情況，有可能還會增加錯誤率的產生。

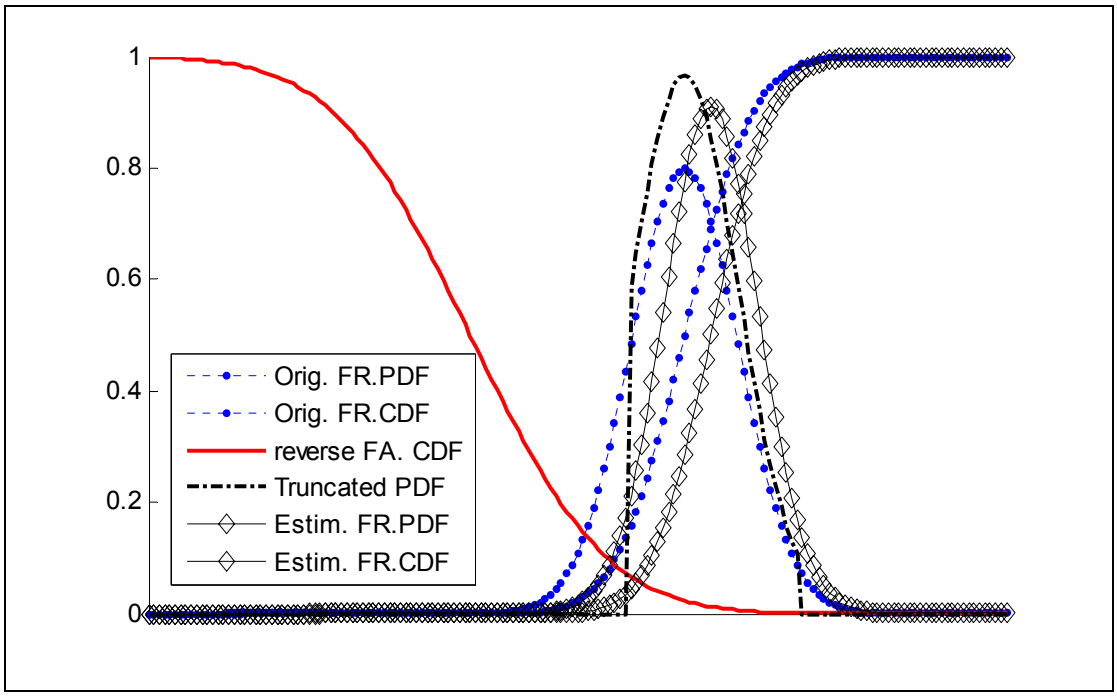


圖表 49 稀少性輸入資料所引發的誤差增加問題



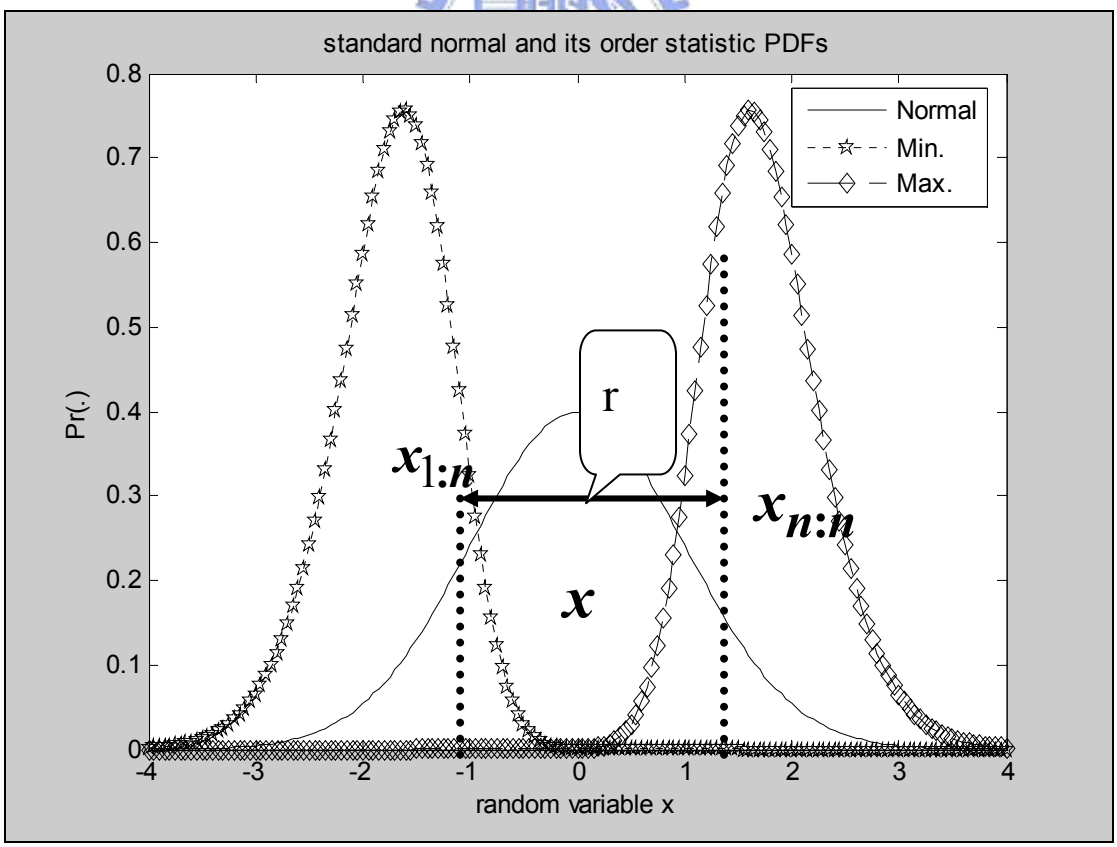
圖表 50 EER 往左方移動，會導致 false alarm 增加

圖表 50 中，推估的機率密度函數和真正原來的機率密度函數比較起來往左方漂移，結果是 EER 也會往左方移動，決策時會導致 false alarm 增加。



圖表 51 EER 往右方移動，會導致決策時的 false rejection 增加

下圖是一個標準常態分佈和它的最大和最小次序統計量分佈。現在我們將圖表 51 中的 speaker false rejection. PDF 使用下圖中的元素來描述。



圖表 52 標準常態分佈下的最大和最小次序統計量分佈

依據模式所得出的公式，我們可以定義如下：

$\boldsymbol{x}$：自我判別測試(self testing) 項目，可能會導致 false rejection 的輸出相對分數，即 $\boldsymbol{x} \Rightarrow \overline{\boldsymbol{J}}(\boldsymbol{x}_{c_{i,j}})$

$$\overline{\boldsymbol{J}}(\boldsymbol{x}_{c_{i,j}}) = \frac{1}{\boldsymbol{d}_{i,j}^{(c|\hat{s}_i)}} \sum_{\lambda=1}^{j_n} \log(\boldsymbol{p}(\boldsymbol{x}_{c_{i,j}}(\lambda) \mid \hat{\boldsymbol{s}}_{c_i})) - \frac{1}{\boldsymbol{d}_{i,j}^{(c|\Omega)}} \sum_{\lambda=1}^{j_n} \log(\boldsymbol{p}(\boldsymbol{x}_{c_{i,j}}(\lambda) \mid \Omega)) - \log \Lambda \tag{6.3}$$

處理程序上，先假設 threshold $\Lambda = 1$。隨機變數 $\boldsymbol{x}$ 是由兩個平均數 $\frac{1}{\boldsymbol{d}_{i,j}^{(c|\hat{s}_i)}} \sum_{\lambda=1}^{j_n} \log(\boldsymbol{p}(\boldsymbol{x}_{c_{i,j}}(\lambda) \mid \hat{\boldsymbol{s}}_i))$ 減 $\frac{1}{\boldsymbol{d}_{i,j}^{(c|\Omega)}} \sum_{\lambda=1}^{j_n} \log(\boldsymbol{p}(\boldsymbol{x}_{c_{i,j}}(\lambda) \mid \Omega))$ 相減所得出，可以視爲常態分配處理，單位是（score/per frame）。

$\boldsymbol{x}_{c_{i,j}}$：屬於 client 中的第 i 位 speaker，第 j 句語料發音。

底下的數學式子，爲了簡化符號內容，$\boldsymbol{x}_{c_{i,j}}$ 都記作 $\boldsymbol{x}$。

$\boldsymbol{n}$：樣本大小，也就是每一位 speaker 自我判別（self test）的輸出分數總數量，這個數量值也會等於每一位 speaker 在語料庫中所說的句子數量總數。

$\boldsymbol{x}_{1:n}$：最小次序統計量，每一位 speaker 經過自我判別所輸出的 n 個隨機標準常態分佈變數，經過排序之後的最小值。

$\boldsymbol{r}$ 全距(range): 每一位 speaker 經過自我判別所輸出的 n 個隨機標準常態分佈變數，最大值隨機變數 $\boldsymbol{x}_{n:n}$ 和最小值隨機變數 $\boldsymbol{x}_{1:n}$ 的差值。

$\boldsymbol{c}$ 覆蓋率(coverage)：➔全距範圍內的 PDF 積分值大小

$\boldsymbol{x}$：自我判別的輸出分數，假設爲常態分佈，$\boldsymbol{x} \sim \boldsymbol{N}(\boldsymbol{u}, \boldsymbol{\sigma}^2)$

$$f(x;u,\sigma) \triangleq f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-u)^2}{2\sigma^2}}, PDF \text{ of normal distribution}$$

$$F(x;u,\sigma) \triangleq F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(t-u)^2}{2\sigma^2}} dt, CDF \text{ of normal distribution}$$

本研究已經建立完成 $x, x_{1:n}, r, c \mid n; u, \sigma$ 的聯合分布型態如下，

$$p(x, x_{1:n}, r, c; u, \sigma \mid n) = p(x; u, \sigma \mid x_{1:n}, r, c, n) \times p(x_{1:n} \mid r, n) \times p(r \mid c, n) \times p(c \mid n) \tag{6.4}$$

$$p(x, x_{1:n}, r, c = Cc \mid n) =$$
$$\{f(x)\left[\frac{UnitStep(x - x_{1:n}) - UnitStep(x - x_{1:n} - r)}{c}\right] \times$$
$$\left[\frac{f(x_{1:n}) f(x_{1:n} + r)\{F(x_{1:n} + r) - F(x_{1:n})\}^{n-2}}{\sum_{i=1}^{m}\left\{\left[e^{\gamma_i^2} f(\gamma_i) f(\gamma_i + r)\{F(\gamma_i + r) - F(\gamma_i)\}^{n-2}\right] w_m(\gamma_i)\right\}}\right] \times$$
$$\left(\sum_{j=1}^{k}\left\{\frac{f(\eta_j) f(\eta_j + r)\left\{F(\eta_j + r) - F(\eta_j)\right\}^{n-2}}{\left|f(\eta_j + r) - f(\eta_j)\right|}\right\}\right) \times$$
$$\frac{1}{Z(r,c,n)} n(n-1)(c^{n-2} - c^{n-1}))\} \tag{6.5}$$

x：常態分佈隨機變數

n：樣本大小

$x_{1:n}$：從常態分佈之中隨機抽取 n 個變數，經過排序在最小位置的隨機變數

$f(x)$：x 之機率密度函數

$F(x)$：x 之累積機率函數

$r = x_{n:n} - x_{1:n}$，全距隨機變數

$c = F(x_{n:n}) - F(x_{1:n})$，覆蓋率，全距範圍下之累積機率值，亦爲一隨機變數

Cc：覆蓋率 c 的某個固定常數值

γ_i:第 m 階的 Hermite polynomials 令其等於零時所有的根

$w_m(\gamma_i)$:weighting coefficient $w_m(\gamma_i) = \dfrac{2^{m-1} m! \sqrt{\pi}}{m^2 \left[\text{Hermite}_{m-1}(\gamma_i)\right]^2}$

η_j：$F(\eta_j + r) - F(\eta_j) - Cc = 0$的根，$\eta_j$須滿足以下限制式（假設有 k 個滿足）

$f(\eta_j + r) - f(\eta_j) \neq 0$ and $\eta_j \in \mathbb{R}$ *and* η_j 不得為重根

$\{\eta_{j,j=1,2\cdots k} \mid F(\eta_j + r) - F(\eta_j) - Cc = 0 \wedge lb \leq \eta_j \leq ub$

$\wedge 0 \leq r \leq |(ub - lb)| \wedge lb \leq \eta_j + r \leq ub\}$

lb $\Rightarrow$ (-4σ):常態分佈的實際合理下界

ub $\Rightarrow$ (4σ):常態分佈的實際合理上界

σ:標準差

$$Z(r,c,n) = \int_{dr} \left[\sum_{j=1}^{k} \{ \frac{f(\eta_j) f(\eta_j + r) \left\{ F(\eta_j + r) - F(\eta_j) \right\}^{n-2}}{\left| f(\eta_j + r) - f(\eta_j) \right|} \} \right]$$

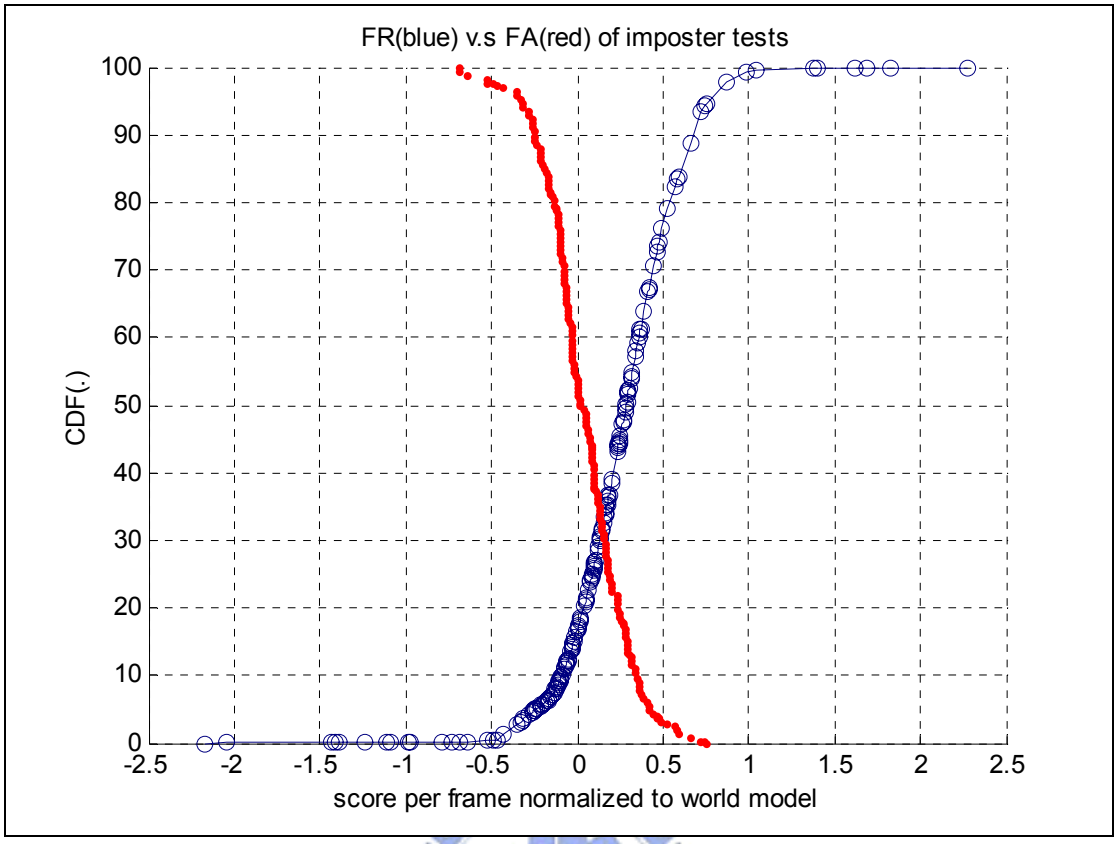
6.5. 實驗 Case 1：基本組態實驗性能測試

輸入狀況組態：
輸入特徵向量：39 維 MFCC (MFCC_E_D_A)
Mixture component: 64
基本 speaker 分成四群，imposter 15 人，client 20 人，world set 20 人(for UBM)，development set 8 人，總共 63 人。

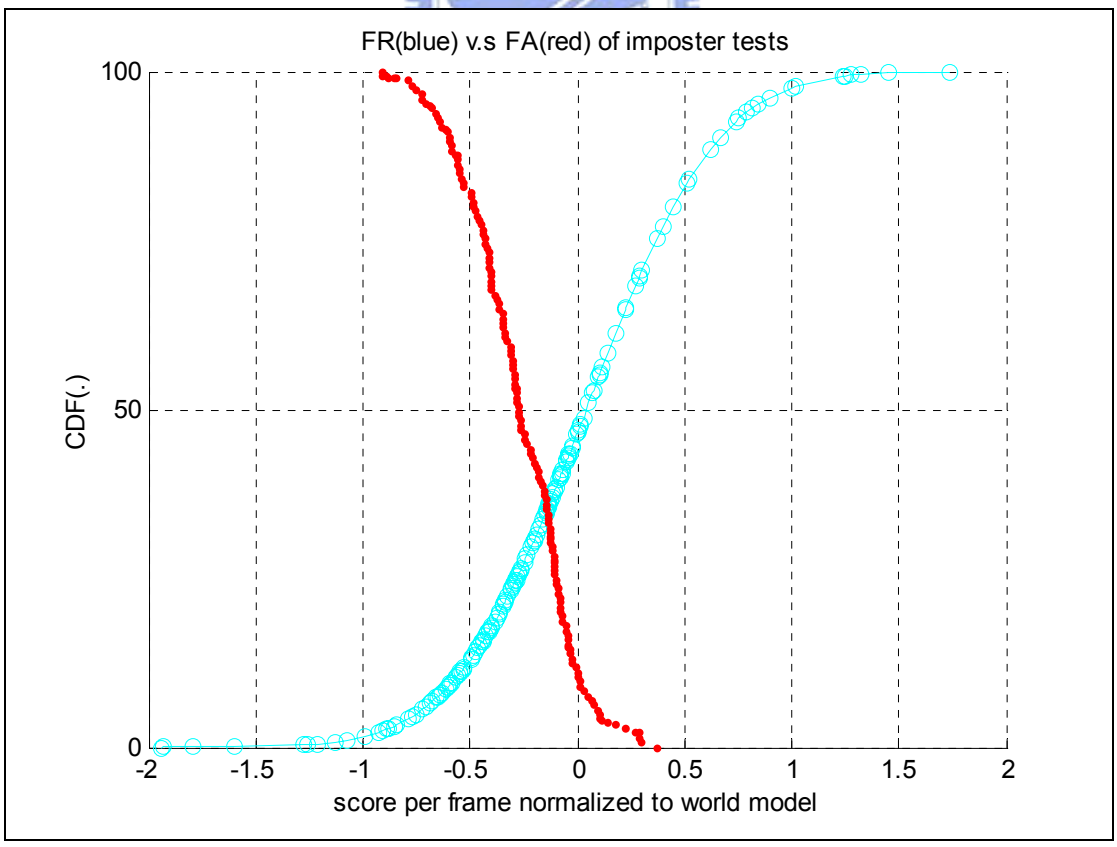
HTK 操作方面，以 flat start 先進行粗估，然後再進行微調。所有的 client model 總共進行兩次微調，UBM model 則進行三次微調。

Corpus 之處理：包含所有的語音元素，且視爲同一個元素。此因只有一個 GMM 模型，所以所有的語音元素都會被相等對待視之。此處爲了簡化處理，所以使用比較大的 syllable 作爲基本分析單元。

實驗的結果如下圖：



圖表 53 64mixtures,speaker verification



圖表 54 16mixtures, speaker verification

表格 13 基本實驗之結果

項目	EER mean	FA mean	FR mean	Threshold
64 mixtures	30.68	31.64	30.6	0.002168
16 mixtures	34.51	35.17	34.5	-0.001723

這個結果並未如 MASV 所描述一般，可以達到接近 80%的確認率。但是本研究的先決條件是在稀少性的輸入資料量，所以本研究的結果尚且差強人意。

由上圖的實驗結果可以知道，mixture 數目增加，對語者確認上的工作是有幫助的。原來是 16 個 mixture，結果確認率 66%，增加 mixture 之後，確認率幾乎接近 70%。

下列表格是僞裝者對受測者當事人的結果統計，由統計的結果，可以作爲重新對 speaker 分類進行測試時的考量依據。

僞裝者，共有二十人（imposter set）

speakerlist{test_set}=['0002','0007','0009','0010','0013','0017','0023','0025','0027','0030','0037','0047','0048','0049','0050','0051','0053','0055','0058','0059'];共有二十人

當事人，共有十五人(client set)

speakerlist{training_set}=['0005','0006','0012','0014','0018','0020','0021','0026','0032','0035','0039','0040','0043','0054','0057'];共有十五人

由統計表中可以約略發現，編號'0032','0035','0039','0040','0043'被僞裝成功的次數明顯比較低。這些 speaker 的特性可以在新一輪的 training set 做爲分類的依據。

下表是 imposter 僞裝成功統計

表格 14 僞裝者對受測者之成功次數統計分佈

	0002	0007	0009	0010	0013	0017	0023	0025	0027
0005	6	3	3	4	5	2	5	4	5
0006	5	5	3	6	3	4	5	4	7
0012	8	4	3	7	4	6	5	5	6
0014	7	4	4	7	5	7	7	5	6
0018	7	7	5	6	8	6	6	4	5
0020	7	7	5	5	8	4	5	3	6
0021	8	5	3	6	5	5	5	3	4
0026	1	5	2	3	2	4	2	1	3
0032	1	1	2	2	4	2	2	1	2
0035	2	0	0	2	1	3	1	1	2

0039	1	0	0	2	3	3	0	2	2
0040	1	0	2	1	4	3	0	1	5
0043	1	1	1	1	4	3	0	1	1
0054	3	5	4	2	2	5	3	2	2
0057	4	7	5	3	4	6	6	2	5

	0030	0037	0047	0048	0049	0050	0051	0053	0055	0058	0059
0005	3	2	2	2	5	4	3	8	4	6	2
0006	4	4	3	4	3	5	4	6	6	5	4
0012	5	4	1	4	5	4	3	6	7	5	6
0014	7	4	4	4	6	6	4	5	7	5	7
0018	6	6	4	7	6	5	4	2	8	4	8
0020	7	6	6	6	9	6	5	4	6	4	8
0021	5	6	1	4	4	5	2	3	4	3	5
0026	1	0	1	1	1	2	1	1	4	1	4
0032	3	7	2	4	3	3	4	2	3	2	3
0035	2	3	0	3	3	2	4	1	4	1	3
0039	1	4	1	3	1	2	2	1	3	2	3
0040	2	6	1	3	2	1	3	1	4	2	4
0043	1	4	1	3	3	2	3	0	1	3	2
0054	3	3	3	4	4	4	3	1	3	4	5
0057	5	4	4	4	5	3	4	4	5	4	4

6.6. 實驗 Case 2➔將稀少性樣本視爲 truncated probability distribution 處理

在 case 1，本研究使用 normal distribution 處理稀少性輸入樣本。現在 case 2 使用 truncated normal distribution 來處理稀少性的輸入樣本問題。輸入 client set 中的某一位 speaker 全部 n 句語料。計算出 empirical mean 和 empirical variance。

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x \tag{6.6}$$

$$S^2 = \frac{1}{n}\sum_{j=1}^{n}(x-\bar{x})^2 \tag{6.7}$$

$x \Leftarrow x_{c_{i,j}}$ ：屬於 client 中的第 i 位 speaker，第 j 句語料發音經過 self testing 計算得出相對於 UBM model 的相對平均分數。單位(score/per frame)

將 empirical mean 和 empirical variance 代入(6.5)作爲起始疊代參數，利用第四章所使用的 truncated probability distribution maximum likelihood estimation 估計出新的 mean 和 variance。然後使用新的 mean 和 variance 作爲 self testing 的分數分佈，最後進行決策。

$$\phi(x) = f(x;0,1)$$
$$\Phi(x) = F(x;0,1)$$

由第四章所推導之關係，

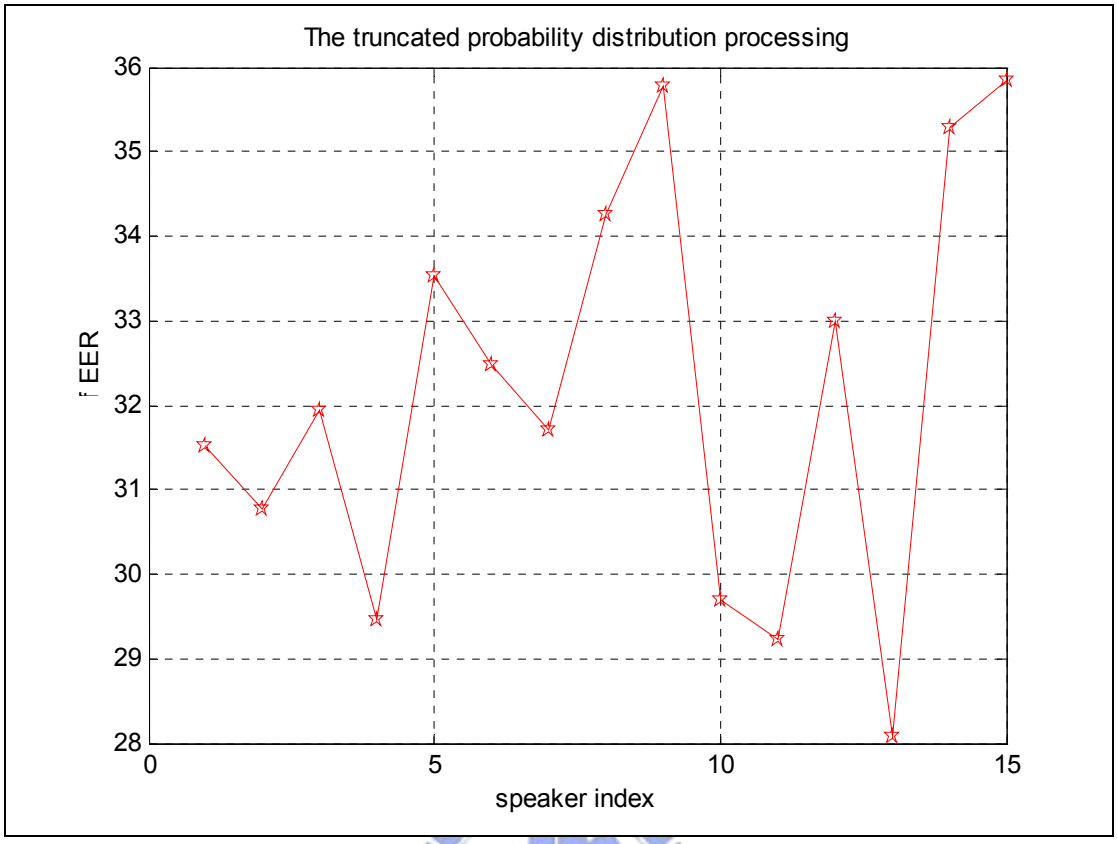
令 $\Theta_L = \frac{\phi(x_{1:n})}{\Phi(x_{n:n})-\Phi(x_{1:n})}, \Theta_R = \frac{\phi(x_{n:n})}{\Phi(x_{n:n})-\Phi(x_{1:n})}$，則

$$\bar{x} - u = \sigma(\Theta_L - \Theta_R) \tag{6.8}$$

$$\sigma^2\left\{x_{1:n}\Theta_L - x_{n:n}\Theta_R + 1\right\} = S^2 + (\bar{x}-u)^2 \tag{6.9}$$

(6.8)和(6.9)有兩個式子，剛好可以解開兩個未知數 u,σ 。

結果：



圖表 55 truncated probability distribution function ML test

總共有十五位當事人(client)，所以結果分成十五組陳列。每一個點代表二十位 imposter 的 EER 總體估計值。由結果可以發現，truncated probability distribution maximum likelihood estimator 的表現欠佳，只有在分佈情形很明顯不對稱的時候，truncated estimator 表現會比 empirical mean（Case 1）來的好。

6.7. 使用 Hypothesis Test 輔助判別

- 檢定已知的 **imposter** 是否爲 **client? right-tailed test**

$$\mathrm{H}_0 : \mu(D) \le \log \Lambda$$
$$\mathrm{H}_1 : \mu(D) > \log \Lambda$$
$$if \quad z_0 = \frac{\bar{J}(x_{I_{k,j}})}{s(D)/\sqrt{n}} > z_{(1-\alpha)}$$
$$\Rightarrow reject\ \mathrm{H}_0$$

s(D):整體imposter的標準差

n:imposter 整個集合的句子總數

H_0：虛無假設

H_1：對立假設

$$\mu(D): \frac{1}{d_{k,j}^{(I|\hat{s}_{c_i})}} \sum_{\lambda=1}^{j_\tau} \log(p(x_{I_{k,j}}(\lambda)|\hat{s}_{c_i})) - \frac{1}{d_{k,j}^{(I|\Omega)}} \sum_{\lambda=1}^{j_\tau} \log(p(x_{I_{k,j}}(\lambda)|\Omega))$$

整個 imposter 集合的句子總數量是大樣本，所以採 z 分數檢定

■ 檢定已知的 **client** 是否爲 **imposter? left-tailed test**

$$\mathrm{H}_0 : \mu(D) \geq \log\Lambda$$
$$\mathrm{H}_1 : \mu(D) < \log\Lambda$$
$$if \quad t_0 = \frac{\bar{J}(x_{c_{i,j}})}{s(D)/\sqrt{n}} < -t_{(1-\alpha,n-1)}$$
$$\Rightarrow reject\ \mathrm{H}_0$$

s(D):client語者c_i的標準差

n:語者c_i的句子數量

$$\mu(D): \frac{1}{d_{i,j}^{(c|\hat{s}_i)}} \sum_{\lambda=1}^{j_n} \log(p(x_{c_{i,j}}(\lambda)|\hat{s}_{c_i})) - \frac{1}{d_{i,j}^{(c|\Omega)}} \sum_{\lambda=1}^{j_n} \log(p(x_{c_{i,j}}(\lambda)|\Omega))$$

Client set 中的任何一位 speaker c_i 的句子總數量在本實驗之中設定最多不會超過 10 句話，屬於小樣本，所以採用 t 分數檢定。

使用 Hypothesis Test 輔助之結果

表格 15 使用假設檢定輔助判別的結果分析

	16 Mixtures				64 Mixtures			
原始	FR tests	FR hits	FA tests	FA hits	FR tests	FR hits	FA tests	FA hits
	465	255	2490	888	465	251	2490	774
alpha=0.05	465	25	2490	127	465	21	2490	121
alpha=0.1	465	45	2490	247	465	38	2490	223

由表格 15 可以知道，speaker verification 的 GMM 模型由 16 mixtures 上升

至 64 mixtures 之後，所減少的錯誤率部分主要來自於 false alarm 的減少。表格中的兩個顯著水準(level of significance)值 alpha=0.05 和 alpha=0.1 都是單尾檢定值。

$$z_{1-0.05}=1.645,\quad z_{1-0.1}=1.285,$$

$$-t_{1-0.05,10-1}=-1.833,\quad -t_{1-0.1,10-1}=-1.383$$

由假設檢定的結果可以知道，speaker verification 如果使用假設檢定作爲輔助判別時，錯誤率可以有效地大幅降低。但是不論 alpha 等於 0.05 或 0.1，16 mixtures 和 64 mixtures 相對於假設檢定的輸出結果機乎相同。

6.8. 實驗 Case 3

若每一位 speaker 有 n 句語料。採用 leave one out 的模式，每次使用 n-1 句語料計算 $x_{1:n-1}, r, c$ 和 u, σ，則我們將會擁有 n-1 組資料。將此 n-1 組資料代入公式中，

$$p(x, x_{1:n}, r, c; u, \sigma \mid n) = p(x; u, \sigma \mid x_{1:n}, r, c, n) \times p(x_{1:n} \mid r, n) \times p(r \mid c, n) \times p(c \mid n)$$

取 likelihood 值最大的一組進行語者確認使用？

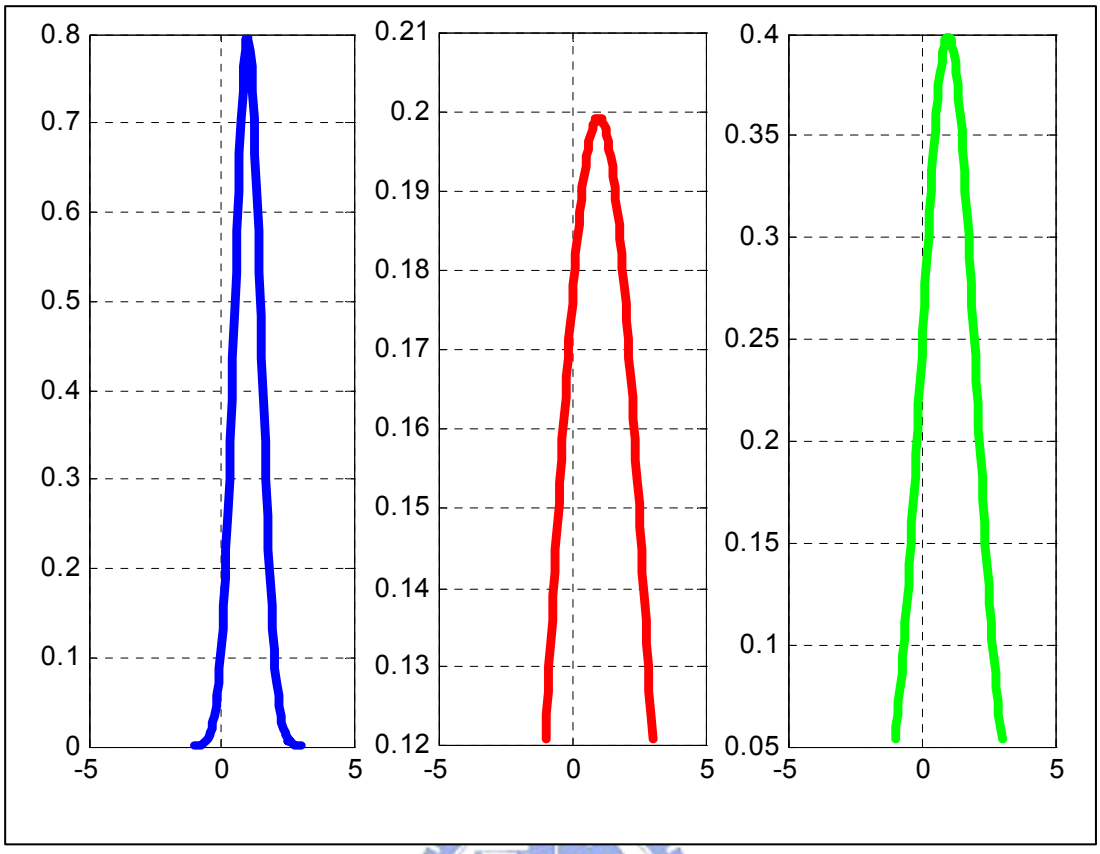
計算方式：以權重方式相加：

當 n-1 個 sample 輸入確立之後，$p(x^{\omega}_{1:n-1} \mid r, n-1), p(r \mid c, n-1), p(c \mid n-1)$，三者將成爲已知，此時便可以視爲權重值，然後將此 n-1 組的權重值相加作爲最終的判別分數。

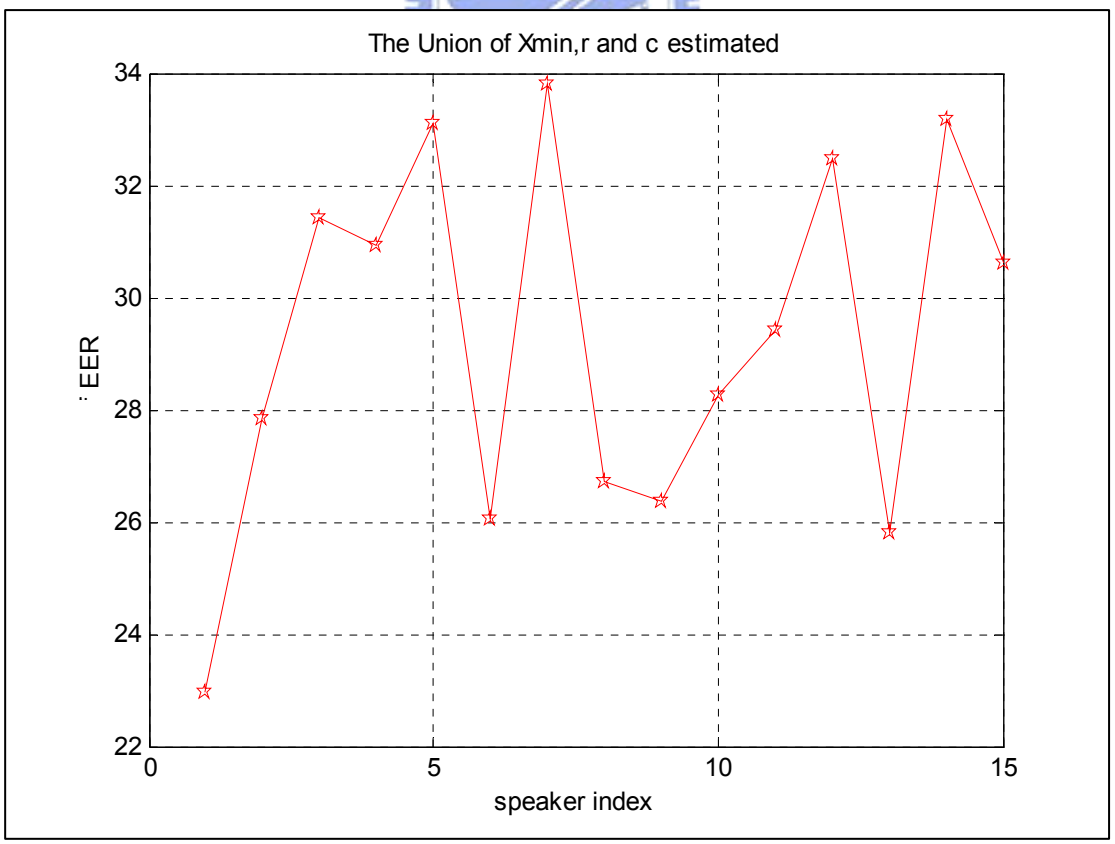
$$p(x, x_{1:n}, r, c; u, \sigma \mid n) = \sum_{\omega=1}^{n-1} p(x; u, \sigma \mid x^{\omega}_{1:n-1}, r, c, n-1) \times p(x^{\omega}_{1:n-1} \mid r, n-1) \times p(r \mid c, n-1) \times p(c \mid n-1) \tag{6.10}$$

ω ：組別，共有 n-1 組

此處採用 leave one out 的概念在於希望將潛在的高斯成份通通都萃取出來進行權重值相加，然後以分數較高者作爲 client 當事人的 self testing normal distribution 參數。



圖表 56 將潛在的高斯成份進行權重值相加



圖表 57 leave one out 的結果

表格 16 綜合比較不同的實驗操作結果

項目	EER mean	FA mean	FR mean	Threshold
64 mixtures	30.68	31.64	30.6	0.002168
16 mixtures	34.51	35.17	34.5	-0.001723
64 mixtures	***29.26***	***30.72***	***29.2***	***-0.002374***

表格 16 中灰色的部份就是 leave one out 的結果。由結果可以發現，leave one out 和同樣是 64 mixtures 的 GMM baseline 實驗結果並無明顯差異；也就是說，本實驗方法並沒有達到實質降低錯誤的效用。

7. 結論與未來展望

一、稀少資料量的可靠度分析在語音問題的分析尚未被提出討論，本研究進行了最初步的探討。

二、本研究在截尾分佈（truncated distribution)和原假設分佈之間已經築起一座橋樑，未來應該還有努力空間。

三、在語者確認的系統實做時，未來可以考慮加入假設檢定，減少過多的錯誤。

四、Hypothesis test 的結果顯示雖然 mixture 數量和辨識率有關，但若加上 Hypothesis test 時，mixture 數量和辨識結果關係不大。

五、語者確認的問題上，64 個高斯單元是比較合適的系統設定。

六、在語音結構方面，更細小或多元的分類也可一併考慮使用，比如除去呼吸聲，short pause 等。

8. 參考文獻

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, ”Score Normalization for Text-Independent Speaker Verification Systems”, Digital Signal Precessing, Vol. 10, pp.42-54, 2000

[2] C. Barras and J.-L. Gauvain, "Feature and Score Normalization for Speaker Verification of Cellular Data," in Proceedings of ICASSP, May 2003.

[3] Mariethoz, J.and Bengio, S.,”A Unified Framework for Score Normalization Techniques Applied to Text-Independent Speaker Verification”, IEEE Signal Processing Letters, Vol. 12, No. 7, pp. 532-535, July 2005

[4] T. Ganchev, I. Potamitis, N. Fakotakis, “Noise-Source Modeling for Robust Speaker Verification in Adverse Environments”, Competitive Environment, Renewable Energy, Distributed Generation ISAP 2003, Lemnos, Greece, August 31st - September 3rd, 2003. paper ISAP03/137

[5] Jonas Richiardi, Plamen Prodanov, Andrzej Drygajlo, “Speaker Verification with Confidence and Reliability Measures”, IEEE International Conference on Acoustics, Speech and Signal processing, 2006.

[6] U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos, and C. Neti, “Audio-Visual Speaker Recognition Using Time-Varying Stream Reliability Prediction”, Proc. International Conference on Acoust, Speech and Signal Processing, Vol. V, pp. 712-715, Hong Kong, Apr. 2003.

[7] Mijail Arcienega and Andrzej Drygajlo, “A Bayesian Network Approach for Combining Pitch and Reliable Spectral Envelope Features for Robust Speaker Verification”, lecture note in Computer Science, Springer, 2003.

[8] K.Y. Leung, M.W. Mak, M.H. Siu, and S.Y. Kung, “Adaptive Articulatory Feature-Based Conditional Pronunciation Modeling for Speaker Verification”, Speech Communications, Vol. 48, Issue 1, pp. 71-84, Jan. 2006.

[9] Erhan Mengusoglu, “Confidence Measure Based Model Adaptation for Speaker Verification”, Proc. 2nd IASTED International Conference on Communications, Internet and Information Technology, Scottsdale, AZ, USA, 17-19, November 2003.

[10] 顏月珠, ”商用統計學”，三民書局，民 79

[11] Hamdy, H. I., “Bayesian Predication Bounds for the Pareto Lifetime Model”, Commun, Statust.-Theory Method, Vol. 16, Issue 6, pp.1761-1772, 1987.

[12] Lagakos, S. W., Barraj, L. M., and Degruttola, V., “Nonparametric Analysis of Truncated Survival Data, with Application to AIDS”, Biometrika 75, pp.515-523, 1988.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, ”Score Normalization for Text-Independent Speaker Verification Systems”, Digital Signal Precessing, Vol. 10, pp.42-54, 2000.

[14] C. Barras and J.-L. Gauvain, "Feature and Score Normalization for Speaker Verification of Cellular Data," in Proceedings of ICASSP, May 2003.

15.COHEN, A. CLIFFORD, “Truncated and Censored Samples: Theory and Applications”, Marcel Dekker ,1991

16.Helmet Schneider , “Truncated and Censored Samples from Normal Populations”, ISBN 0-8247-7591-0. Marcel Dekker, 1986.

17.H.A.DAVID, “Order Statistics, 2nd Edition”, Iowa State University, 1981.

18.N. Balakrishnan and A. Clifford Cohen, “Order Statistics and Inference: Estimation Methods”, Academic Press, Inc., 1991.

19.Barry C. Arnold, N. Balakrishnan and H. N. Nagaraja, “A First Course in Order Statistics”, John Wiley & Sons, Inc., 1992.

20.John P. Klein and Melvin L. Moeschberger, “Survival Analysis Techniques for Censored and Truncated data, 2nd Edition”, Springer, 2003.

21.Vincent Wan, “Speaker Verification using Support Vector Machines”, Ph.D. thesis, University of Sheffield, U.K., June 2003.

22.Chun-Nan Hsu and Hau-Chung Yu and Bo-Hou Yang. "Speaker Verification without Background Speaker Models," In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2003, Hong Kong, China, 2003.

23.C. T. Liao and H. K. Iyer, “A Tolerance Interval for the Normal Distribution with Several Variance Components”, Statistica Sinica, Vol.14, pp,217-229, 2004

24.J HH Barrett and KJ Myers, “The Dirac Delta and other Generalized Functions,” in Foundations of Image Science, John Wiley & Sons, Inc., New Jersey, pp. 63-94, 2004.

25.Arfken, G. “Appendix 2: Gaussian Quadrature.”, Mathematical Methods for Physicists, 3rd Edition. Orlando, FL: Academic Press, pp. 968-974, 1985.

26.Taischi, Chi., “A Segment-Based Speaker Verification System Using SUMMIT. Massachusetts”, Massachusetts Institute of Technology, Department of Electrical Engineering and. Computer Science, Master thesis, 1997