東　海　大　學

工業工程與經營資訊研究所

碩士論文

題目：強森分配擬合方法於

輸入資料分析中之探討

研 究 生：林保佑

指導教授：姚銘忠 博士

蔡禎騰 博士

彭　泉 博士

林水順 博士

中華民國九十一年六月

強森分配擬合方法於輸入資料分析中之探討

學生：林保佑　　指導教授：姚銘忠 博士

蔡禎騰 博士

彭　泉 博士

林水順 博士

東海大學工業工程與經營資訊研究所

摘要

在系統模擬的研究中，決策者若是在輸入資料分析時發生錯誤，可能造成運用錯誤的機率分配來產生隨機變數，明顯地將造成模擬的明確性受到質疑，甚至嚴重地影響系統模擬結果的正確性。在實際的應用中，傳統的輸入資料分析必須以已知的機率分配來擬合樣本資料點。在此常遭遇一個窘境，即在嘗試所有已知的機率分配後，仍無法在所要求的信賴度下通過配合度檢驗。在此狀況下，模擬模式建構者可能被迫採取運用經驗分配。但在實際運用經驗分配時，又常遭遇下列的問題：因為大部分的連續型機率分配都有峰度偏右的傾向，如果樣本數不大時，則樣本由機率密度函數右尾部分所產生的希望則非常渺茫，而且經驗分配機率密度函數無法產生比最大樣本點大的隨機變數，這些問題會造成系統模擬結果嚴重的誤差。

為解決上述的問題，在最近的十幾年內，已經有許多的學者嘗試在系統模擬的輸入分析中運用強森分配。以往研究學者嘗試以下列四種擬合方法來擬合強森分配與樣本資料點：動差擬合法、百分比擬合法、最小平方法、最小 L_p-norm 法。在文獻中，學者已運用第 3 級及第 4 級動差之合理範圍建立(b_1, b_2)平面，供資料分析者運用強森分配轉換系統時依資料型態選擇適用分配族(distribution family)之參考。本研究的重點在於判斷在平面上不同的分佈位置，運用何種擬合方法其效果最好。本研究運用隨機方式，產生均勻分佈於(b_1, b_2)平面 4,000

組以上，每組 25 個樣本點的實驗資料進行擬合，再以 Kolmogorov-Smirnov Test 進行適合度檢驗。本研究發現動差擬合法及百分比擬合法兩個方法的擬合效果較差，而最小平方法及最小 L_p-norm 法運用較佳的實驗資料點，則以假設檢定與 Logistic Regression 斷定出其在(b_1, b_2)平面上個別較佳的區域。本研究的主要貢獻在於協助系統模擬之資料分析者，能較快且準確的選擇適合的擬合方法，決定最適的強森分配，作為系統模擬輸入資料分析時方便有用的工具。

On Choosing the Best Fitting Approach When Applying Johnson Distribution in Input Data Analysis

Student: Pao-Yu Lin

Advisor: Dr. Ming-Jong Yao
Dr. Jen-Teng Tsai
Dr. Chyuan Perng
Dr. Shui-Shun Lin

Institute of Industrial Engineering and Enterprise Information

Tunghai University

Abstract

Input data analysis (IDA) plays an extremely important role in system simulation. Inappropriate IDA may lead an analyst to employ incorrect probability distributions (*p.d.*'s) to generate random variables, and consequently, it would jeopardize validity of the results from system simulation.

Using conventional IDA, an analyst may have to try all the known *p.d.*'s to fit the sample data points. However, it often leads to an embarrassing situation – an analyst may have no probability distribution which could meet the confidence level required in the goodness-of-fit test. Therefore, one may be enforced to use empirical distributions for data fitting. However, an analyst may have the following problems: (1) most of the continuous *p.d.*'s are skewed-to-the-right, if the size of the sample data points is not large, the probability of generating random variables at the right-tail of the *p.d.* is slim; (2) empirical distribution is unable to generate any random variable that is larger than the largest sample point. These problems may cause serious consequence in the results of system simulation.

In the past two decades, researchers have been addressing their efforts in applying Johnson distribution to improve the conventional IDA. Four fitting approaches have been derived to fit the sample data points

with certain Johnson distribution; they are: (1) moment matching, (2) percentile matching, (3) least square, and (4) least L_p-norm approaches. In the literature, researchers proposed to use the value of (β_1, β_2) (from the skewness and the kurtosis of the sample data) to choose an appropriate distribution family when one fits the sample data using Johnson distribution.

The focus of this study is to help the analyst to determine the most suitable fitting approach using the information of β_1 and β_2. For this purpose, we generate more than 4000 sets of randomly data with 25 sample points in each data set. Then, we analyze each data set using the four aforementioned fitting approaches and evaluate the goodness-of-fit for these approaches by Kolmogorov-Smirnov test. We assert that both the moment matching and the percentile matching approaches are inferior to the other two. Importantly, using a hypothesis testing method and logistic regression, we indicate the specific regions on the (β_1, β_2) plane where the least-square approach surpasses the least L_p-norm approach, and vice versa. Our conclusion assists an analyst to efficiently choose the best fitting approaches when the analyst employs Johnson distribution to find the most appropriate distribution in Input Data Analysis.

誌謝

隨著時序更迭，樹也如正有意無意地提醒別離的到來。兩年的時間過得很快，研究所的生活即將步入尾聲，在東海求學的六個年頭，有許多令人難忘的回憶，而這些回憶不僅將伴隨我踏進下一個旅程，也豐富了我的人生。

兩年來的研究所生活，在課業、智識及生活上感謝多位師長從旁的指導與協助，讓我得以順利完成學業。首先感謝指導老師姚銘忠博士在論文及專案計畫上的殷勤指導，蔡禎騰博士除論文指導外，在日常生活及人生哲學上亦給予莫大的啟益，而彭泉博士與林水順博士在論文的指導上也不遺餘力，使我受益良多。在此衷心感謝老師們對論文辛勤的指導及對學生的關愛。

另外，要感謝口試委員鄭順林博士、盧希鵬博士及陳同孝博士在論文口試時提供寶貴的意見與指正，使得論文更趨完善。還要感謝淑惠學姊與創鈞學長對我的關心及論文的指正。

另一方面，研究室諸位同學及學弟們對我的鼓勵與支持讓我倍感溫馨。謝謝同學錦宗、維駿、永祥、志賢、博仁、英欽、宗輝、俊武及湘翎，以及學弟政憲、穎威、志宏、崇民、子麟、勝翔、宣彥、茂洲在這些日子裡帶給我許多美好的回憶。此外要特別感謝系上的助教素卿、俊良、雅惠及宏華對我生活及論文上種種協助與照顧，使我能夠順利的完成論文。

最後，要感謝我的雙親與家人，以及雅娟、君華、筆勝、士恩、偉正、哲旭、裕聰給我的支持和關懷，讓我在求學生涯中能順利完成我的目標及理想。在此謹將此小小的成果與關心我的人一起分享。

林保佑　謹誌於

東海大學工業工程與經營資訊研究所

中華民國九十一年七月

目錄

頁次

中文摘要 i

英文摘要 iii

誌謝 v

目錄 vi

表目錄 ix

圖目錄 x

第一章　緒論 1

1.1 **研究背景**.....1

1.2 **研究動機與目的** 2

1.3 **研究方法與步驟** 4

1.4 **研究範圍與限制** 5

1.5 **研究架構**.....6

第二章 文獻探討 8

2.1 **輸入資料分析**.....8

2.2 **輸入資料分析流程介紹**.....10

2.2.1 **假設樣本所屬分配族**.....10

2.2.2 **參數估計**.....12

2.2.3 **以卡方**(Chi-square test)**進行適合度檢定**.....13

2.3 **輸入資料分析問題點**.....14

2.4 **強森分配轉換系統**.....15

2.5 **資料分析技術的應用**.....16
2.5.1 **資料挖掘的內涵**.....17
2.5.2 **迴歸技術於資料挖掘的應用**.....17
2.5.3 **邏輯迴歸**(logistic regression)**介紹**.....18
2.6 **本章結論**.....19
第三章　強森分配數學模式.....**20**
3.1 **強森分配數學模式**.....20
3.2 **強森分配擬合方法**.....22
3.2.1 **動差擬合法**(Moment Matching).....22
3.2.2 **百分比擬合法（**Percentile Matching**）**.....24
3.2.3 **最小平方法（**Least Squares**）**.....25
3.2.4 **最小** L_p-norm **法（**Minimum L_p-norm Estimation**）**.....26
3.3 **遺傳演算法**(Genetic algorithm)**產生樣本資料點**.....27
3.3.1 **遺傳演算法的介紹**.....28
3.3.2 **樣本產生程序**.....30
3.3.3 **樣本產生方式**.....31
3.4 **樣本數據**.....33
第四章　樣本資料分析.....**35**
4.1 **以強森分配進行擬合**.....35
4.2 **資料整理**.....37
4.3 **擬合結果分析**.....40

4.3.1 **擬合結果整理與判別**.....40

4.3.2 **統計檢定及邏輯迴歸判別**.....40

第五章　樣本資料分類 42

5.1 **邏輯迴歸模型**.....42

5.2 **邏輯迴歸假設與檢定**.....43

5.3 **迴歸模型適合度** 43

5.4 **邏輯迴歸分析**.....44

5.4.1 **分析程序**.....44

5.4.2 **資料分析**.....45

5.5 **邏輯迴歸結果於二維**$(\hat{a}_1,\hat{a}_2)$**平面之分析**.....46

第六章　結論與建議 48

6.1 **結論** 48

6.2 **建議** 50

參考文獻 51

表目錄

表 2-1 **模擬應用中之隨機資源..... 9**

表 2-2 **統計量摘要.....11**

表 2-3 **樣本百分位數結構摘要..... 12**

表 3-1 Data1、Data2 **資料表..... 33**

圖目錄

圖 1-1 $\hat{a}_1$、β_2與強森分配族間之關係圖 4
圖 1-2 研究架構圖 7
圖 2-1 輸入資料分析之步驟 10
圖 2-2 箱形圖 12
圖 3-1 $\hat{a}_1$與β_2之關係圖 23
圖 3-2 GA 搜尋程式流程圖 30
圖 3-3 以 GA 產生樣本之流程圖 32
圖 3-3 Data1 之直方圖 33
圖 3-4 Data1 之箱形圖 34
圖 3-5 Data2 之直方圖 34
圖 3-6 Data2 之箱形圖 34
圖 4-1 擬合步驟 36
圖 4-2 樣本資料之β_1、β_2分圖 37
圖 4-3 動差擬合法 38
圖 4-4 百分比擬合法 38
圖 4-5 最小平方擬合法 39
圖 4-6 最小 L_p-norm 法 39
圖 4-7 二維平面區隔圖 41
圖 5-1 邏輯迴歸方程式於最小 L_p-norm 法 47
圖 5-2 邏輯迴歸方程式於最小平方法 47

第一章　緒論

1.1 研究背景

在系統模擬(system simulation)的研究中，決策者須運用輸入資料分析(input data analysis)的方法，針對模擬模式中的隨機變數斷定其機率分配。然後運用該機率分配，在模擬模式中產生該部分所需之隨機變數，以利進行系統模擬。若是在輸入資料分析時發生錯誤，則可能造成運用錯誤的機率分配來產生隨機變數，明顯地將造成模擬模式的明確性(validity)受到質疑，甚至嚴重地影響系統模擬結果的正確性，因此輸入資料分析在系統模擬的研究中扮演相當重要的角色。

傳統在進行輸入資料分析必須在機率分配已知的情況下來擬合樣本資料點，以進行資料分析的動作。但決策者常會遇到一困境，即在嘗試所有機率分配後仍無法找到一合適的分配，在此情形下決策者只好被迫採用經驗分配來處理，但如 Kelton and Law[24]所評論，在實際運用經驗分配時，常遭遇一問題：大部分的連續型(continuous)機率分配都有偏度偏右(skewed to the right)的傾向，如此以經驗分配做為樣本資料點擬合的依據，其所得的結果往往與實際值有所出入，如此便無法對資料做正確的分析與處理。

決策者在進行資料分析時除了可能面臨上述困境外，亦會遭遇樣本數多寡的問題，一般實際應用上常假設其服從常態分配，但若樣本數不大時，此常態假設所得結果會有明顯的偏差，若利用經驗分配來解決此問題，則會因偏度偏右導致樣本由機率密度函數右尾部分所產生的希望渺茫，如果無法產生如此的隨機變數，則決策者將無法觀察到模擬系統會有如此的現象，而對整個系統的特性有所誤判。

資料分析在實際應用方面非常廣泛，在統計檢定及品質管制的範疇中，對於所得樣本資料亦必須在已知或某特定的分配下方能進行分析。如在統計製程管制(statistical process control)中，常必須面臨到非常態的資料型態，然而大多數處理統計製程管制問題之方法皆須在其

產品資料型態為常態下方能進行。又如在分配未知的報童問題，在決策點 t 時估計未來一銷售時點 T 之需求量，若所依據之分配並非適合此報童問題模式，則所預測之需求量，可能會發生缺貨或存貨過多等問題，因而導致成本的損失。

綜上所述，資料分析其應用的範圍廣泛且深入，在不同的樣本資料特徵(pattern)下，決策者如何應用資料分析方法，處理不同的作業要求，並能快速且有效率的獲得分析競爭激烈的環境中更形重要。

1.2 研究動機與目的

系統模擬之輸入資料分析運用的分析技術，皆以樣本資料點為分析的基礎，尋找適合的機率分配來描述母體。多數的系統模擬書籍[2,7,17,30,31,32]建議資料分析者，運用下列之流程，針對將進行研究的隨機變數，進行輸入資料分析：

1. 收集所欲研究隨機變數的樣本資料點。
2. 以長方圖等方法對樣本資料點描圖，並對於隨機變數的機率分配(通常為常見已知的機率分配)進行假設。
3. 以樣本資料點估計所假設之機率分配其各個參數。
4. 以 Chi-Square test 或 Kolmogorov-Smirnov test (簡稱 K-S test)進行配合度檢驗(Goodness-of-fit test)。
5. 如果所假設的機率分配無法在所要求的信賴度下，為配合度檢驗所接受，則再假設另一機率分配，並回到第 2 步驟繼續嘗試，以期通過配合度檢驗。

在實際的應用中，如此的流程常會遭遇一個窘境：在第 2 步驟中，模擬模式建構者可能嘗試過所有已知的機率分配，但仍無法在所要求的信賴度下通過配合度檢驗。在此狀況下，模擬模式建構者可能

被迫採取運用經驗分配。但經驗分配因大部分的連續型(continuous)機率分配都有偏度偏右的傾向，所產生的機率分配之隨機變數無法完整且正確描述樣本資料。

為解決上述經驗分配所可能造成的問題，在最近的十幾年內，已經有許多的學者嘗試在系統模擬的輸入分析中運用強森系統(Johnson System)，以 Wilson 為首的研究群，已經在此主題上有相當不錯的研究成果[12,37,38,39,40]。以往研究學者嘗試以下列四種擬合方法來擬合強森分配與樣本資料點：動差擬合法（Moment Matching)、百分比擬合法（Percentile Matching)、最小平方法（Least Squares)、最小 L_p-norm 法。

蔡瑞隆[53]指出在強森分配與其它外形多樣之分配的比較中，強森分配總能正確的描述樣本之母體，且能有效呈現母體特性。而在強森分配擬合方法的比較中，林保佑[44]指出在樣本數不大(<30)，且母體分配未知的情況下，強森分配其擬合的效果優於統計常用的分配，且在相同的樣本條件限制下，四種強森的擬合方法中以最小 L_p-norm 法最佳。

由上所述在輸入資料分析的範疇中，應用強森分配做為樣本資料點擬合的依據上有相當程度的助益。而在以強森分配轉換系統(Johnson translation system)擬合樣本資料，根據 Johnson 建議應依資料型態(data pattern)，尤其是其 3 級及 4 級動差之相關資訊，對應其適當的分配族(distribution family)。在文獻中，學者已運用 3 級及 4 級動差之合理範圍建立二維之(β_1, β_2)平面(圖 1-1)，供資料分析者運用強森分配轉換系統時選擇適用分配族之參考。

根據此二維之(β_1, β_2)平面，可了解在不同的動態及偏態下，樣本資料其所屬的強森分配族類。而本研究即根據此平面作為研究基礎，以強森分配為原始資料分析之工具，擬分析與判斷不同的資料型態，以強森分配擬合方法進行擬合，其擬合結果的優劣在此二維平面上的分佈狀況，歸納出在不同的樣本特徵時，何種擬合方法的效果最

好，藉以幫助決策者在面臨不同樣本資料時能正確且有效對資料進行分析，而能針對不同資料結構做一較佳的決策。

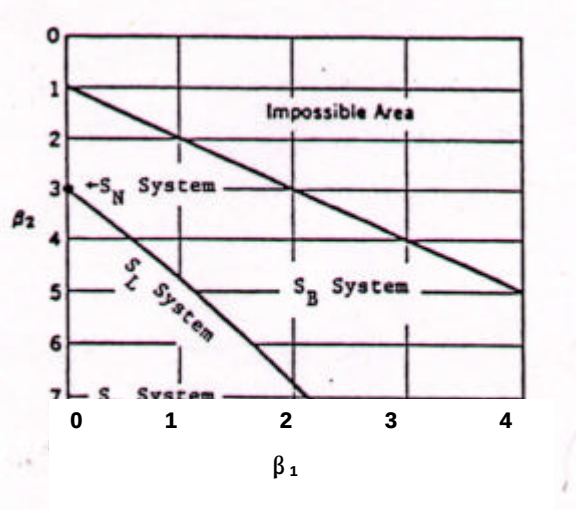


圖 1-1 β_1、β_2與強森分配族間之關係圖

1.3 研究方法與步驟

過去在進行資料分析、銷售預測、產品抽樣與品質檢驗時，已有許多的學者嘗試在處理資料時，以樣本的特性(如平均數、變異數)或樣本的順序統計量作為依據，搜尋強森分配的四個參數，欲獲得擬合效果最佳的強森分配結果，故有了以下四種擬合的方法：

1. **動差擬合法**(Moment Matching)
2. **百分比擬合法**(Percentile Matching)
3. **最小平方法**(Least Squares)
4. **最小** L_p-norm **法**(Minimum L_p Norm Estimation)

根據前述目的，本研究運用隨機方式，產生均勻分佈於二維(β_1, β_2)平面的 4000 個樣本點，透過對樣本資料處理所得資料的偏態及峰態，運用強森分配之四種擬合法則擬合樣本資料，再以 Kolmogorov-Smirnov Test(簡稱 K-S 檢定)進行適合度檢驗，判斷在平面上不同的分佈位置，何種擬合方法的效果最好。藉此提供系統模擬

之資料分析者能較快且準確的選擇適合的擬合方法，減少時間的浪費及錯誤的發生。

本研究之研究步驟可分下列五個步聚：

1. 蒐集輸入資料分析及強森分配等相關文獻，整理歸納其中之內涵。
2. 探討輸入資分析在系統模擬中所面臨的問題點，及強森四種擬合方法間之特性和優劣比較，並瞭解強森分配在輸入資料分析上的應用。
3. 利用遺傳演算法(Genetic Algorithm)隨機產生所欲分析之樣本資料，利用強森之四種擬合方法針對所產生之資料點進行擬合，所得結果再以 Kolmogorov-Smirnov Test(簡稱 K-S 檢定)進行適合度檢驗，將所得結果分佈於(β_1, β_2)之二度平面。
4. 利用分群的手法，對分佈於(β_1, β_2)平面上之點進行分群，擬找出在不同的區域上，最佳之擬合方法為何，並歸納其結果。
5. 提出結論與未來研究方向之建議。

1.4 研究範圍與限制

本研究利用遺傳演算法，以 Visual Basic 程式產生所需之樣本資料，以強森分配之四種擬合方法對樣本資料進行擬合，透過統計分析技術進行擬合結果的分析。在樣本資料的型態上有以下限制：

1. 資料以隨機方式產生。
2. 每組樣本共 25 個數值。
3. 數值範圍介於 0 及 2500 之間。
4. 每組樣本均為分配未知且偏度偏右之資料型態。

5. 樣本資料之 β_1, β_2 均勻分佈於二維平面上，不在此平面範圍上之樣本點，本研究不進行探討。

在擬合方法上，僅以強森分配所提之四種擬合法則進行擬合，它種擬合法則不納入本研究範圍。本研究之擬合法則為：

1. 動差擬合法
2. 百分比擬合法
3. 最小平方法
4. 最小 L_p-norm 法

1.5 研究架構

本研究之架構依章節排列，共分為五個章節，如圖 1-2 所示。第一章節簡述本研究的研究背景、研究動機與目的、研究方法、研究步驟與研究範圍與限制。第二章則是針對輸入資料分析、強森分配、遺傳演算法及 k-s 檢定法等課題進行相關文獻的探討。第三章針對強森分配模型及資料點的產生、擬合與分析進行探討，並包含若干樣本數據以茲說明。第四章為進行資料點的擬合及各擬合法則結果的整理。第五章為以統計檢定及邏輯迴歸對樣本資料進行平面結果的整理與探討。第六章為本研究歸納之結論與建議。

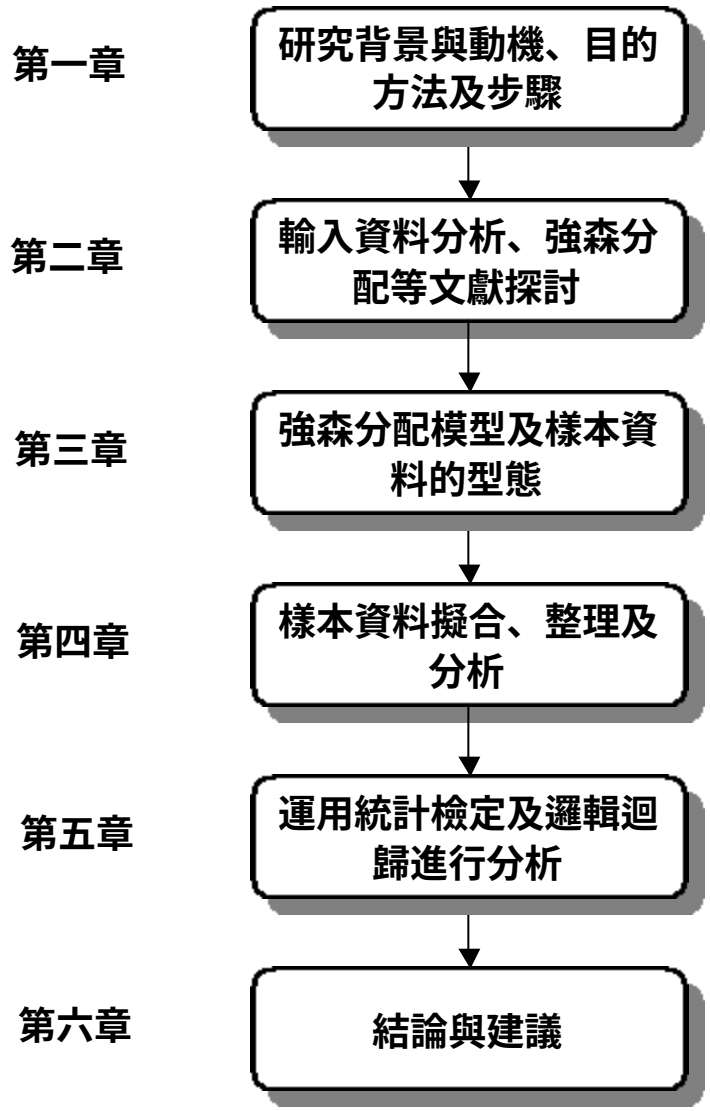

第一章
研究背景與動機、目的
方法及步驟
第二章
輸入資料分析、強森分
配等文獻探討
第三章
強森分配模型及樣本資
料的型態
第四章
樣本資料擬合、整理及
分析
第五章
運用統計檢定及邏輯迴
歸進行分析
第六章
結論與建議

圖 1-2 研究架構圖

第二章 文獻探討

在系統模擬(system simulation)的研究中，決策者須運用輸入資料分析(input data analysis)的方法，針對模擬模式中的隨機變數斷定其機率分配。然後運用該機率分配，在模擬模式中產生該部分所需之隨機變數，以利進行系統模擬。若是在輸入資料分析時發生錯誤，則可能造成運用錯誤的機率分配來產生隨機變數，明顯地將造成模擬模式的明確性(validity)受到質疑，甚至嚴重地影響系統模擬結果的正確性，故輸入資料分析在系統模擬的研究中扮演相當重要的角色。

2.1 輸入資料分析

使用隨機輸入的方式來執行模擬的動作時，需明確定義其資料的機率分配，如單一服務窗口之等候線系統其抵達時間或是存貨管理系統需求量大小的預測等，其模擬模型之輸入隨機變數皆需遵循特定的分配，並從這些分配中產生模擬所需的隨機變數值。

所有實際世界系統幾乎都包含一或多個隨機的資料型態，而在模擬應用的系統及資源如表 2-1 所示。在每個不同系統型態中，其隨機資料導循著特定的分配，模擬模式建構者可依其既定的分配型態進行系統模擬，藉此推斷出此系統運作模型及特性。如一等候系統其資料符合一指數分配等。

Kelton and Law 在其著作中提到輸入資料分析中，對一系統若能正確且完整的蒐集其資料則可使用以下的方法來明確定義其所屬分配，方法如下：

1. 資料值本身即可直接使用於模擬中。例如資料值是服務時間，則在模擬過程中須要一服務所需時間的資訊，則某一組資料之初始值即可直接使用。此即所謂回溯模擬(trace-driven simulation) 。
2. 資料值本身被用來定義為經驗分配的函數，其涵義為若資料代表服務的時間，則當模擬過程中須要一服務所需時間的資訊時，可自此

經驗分配中進行抽樣。

3. 在統計推論的標準技術中常利用統計理論分配型態來擬合(fit)資料，如指數分配(Exponential distribution)、波以耳分配(Poisson distribution)等，擬合結果經由假設檢定來決定其擬合的適合度。若此理論分配擬合結果的參數值對服務時間資料為一良好的模型，若模擬過程中須要此服務時間的資料時，可自此理論分配中進行抽樣。

表 2-1 模擬應用中之隨機資源

系統型態	隨機資源
製造	加工時間、機器故障時間、機器維修時間等
防禦系統	飛機或飛彈的抵達時間和負載、武器偏差距離等
溝通協調	資料傳遞時間、資料型態、資料大小
運輸	上下貨物的時間，顧客抵達車站時間等

資料來源：Law, A.M. and W.D. Kelton, *Simulation Modeling & Analysis*, 7th ED., New York,1990,p.326.

輸入資料分析透過對欲模擬的系統蒐集所需資料，藉上述所提的方法對樣本資料尋其合適的分配模型，藉此分配模型展開一連串的模擬的動作，因之資料的蒐集及分配的選擇乃是輸入資料分析在系統模擬進行程序中的開端，其詳細的分析步驟及方法在許多系統模擬書籍中皆有論述，運用這些步驟，針對將進行研究的隨機變數，進行輸入資料分析。其步驟如圖 2-1 所示，各步驟介紹於下章節詳述。

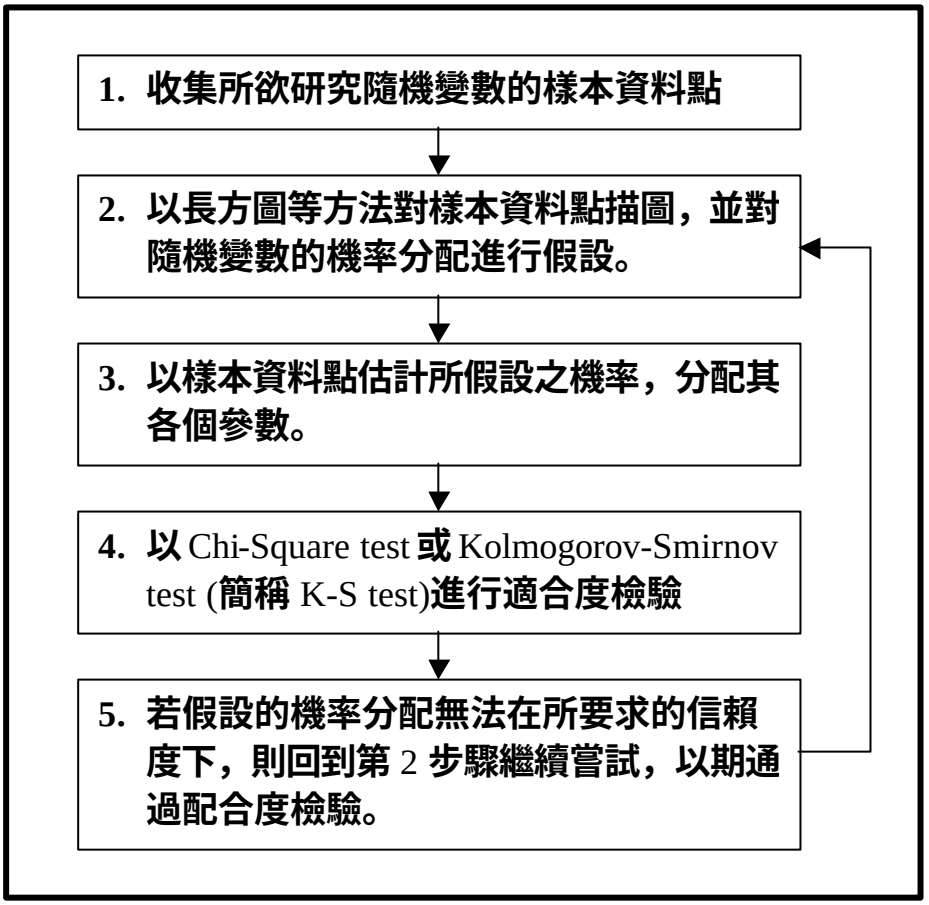


圖 2-1 輸入資料分析之步驟

2.2 輸入資料分析流程介紹

在對所欲分析的系統正確且完整的蒐集其隨機產生的樣本資料後，便可著手輸入資料分析對資料分析所執行的步驟，以下便是各步驟的詳細介紹。

2.2.1 假設樣本所屬分配族

對樣本點選擇其所屬分配的階段中最主要的課題為決定所屬分配族類，亦即考慮樣本資料是否有相類似的圖形型態，而不去考慮分配族類是否有明確的參數值。如：指數分配、常態分配或波以耳分配等。本小節描述一般用來假設分配族類常用的技術。

1. **統計量**

某些分配的不同點可藉由其統計量的參數值以茲分辨，表 2-2 **為幾個常用的統計量，透過自獨立且同型分配之資料中進行函數估計的動作。**

表 2-2 統計量摘要

統計量	**樣本估計值**
最大值、最小值	$X_{(1)}, X_{(n)}$
平均數 μ	$\bar{X}_{(n)}$
中位數 $X_{0.5}$	$\hat{x}_{0.5}(n) = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \left[X_{(n/2)} + X((n/2)+1)\right]/2 & \text{if } n \text{ is even} \end{cases}$
變異數 σ^2	$S^2(n)$
變異係數, $cv = \frac{\sqrt{s^2}}{m}$	$\hat{cv}(n) = \frac{\sqrt{S^2(n)}}{\bar{X}(p)}$
偏態值, $u = \frac{E\left[(X-m)^3\right]}{(s^2)^{3/2}}$	$\hat{u}(n) = \frac{\sum_{i=1}^{n}\left[X_i - \bar{X}(n)\right]^3 / n}{\left[S^2(n)\right]^{3/2}}$

資料來源：Law, A.M. and W.D. Kelton, *Simulation Modeling & Analysis*, 7th ED., New York,1990,p.326.

2. **直方圖**

對連續的資料型態來說，直方圖乃藉由輪廓繪製的方式來估計資料點 $X_1, X_2, \ldots X_n$ **的分配其所屬之機率密度函數。此以描繪資料的型態來估計機率密度的方式，對找尋資料所屬分配來說是個不錯的好方法，且對資料點來說亦是一個好的模型。至於直方圖製作的方法本研究不再詳述。**

3. **分位數說明文摘**(Quantile Summaries)**與箱形圖**(Box Plots)

分位數說明文摘乃是對樣本作概要的敘述，可用來決定樣本機率密度函數的圖形是屬對稱、右偏或左偏，且對連續或離散的資料集合

皆適用。而對樣本資料 $X_1, X_2, .. X_n$ 的分位數概要如表 2-3 所示。

表 2-3 樣本百分位數結構摘要

分位數	項數	樣本值	中點
中位數	$i = (n+1)/2$	$X_{(i)}$	$X_{(i)}$
四分位數	$j = (\lfloor i \rfloor + 1)/2$	$X_{(j)}$ $X_{(n-j+1)}$	$[X_{(j)} + X_{(n-j+1)}]/2$
十分位數	$k = (\lfloor j \rfloor + 1)/2$	$X_{(k)}$ $X_{(n-k+1)}$	$[X_{(k)} + X_{(n-k+1)}]/2$
端點	1	$X_{(1)}$ $X_{(n)}$	$[X_{(1)} + X_{(n)}]/2$

資料來源：Law, A.M. and W.D. Kelton, *Simulation Modeling & Analysis*, 7th ED., New York,1990,p.326.

箱形圖為分位數概要的圖形展現。一般有 50%的觀測值會落於 $[X_{0.25}, X_{0.75}]$間之水平邊界內，如圖 2-2 所示。

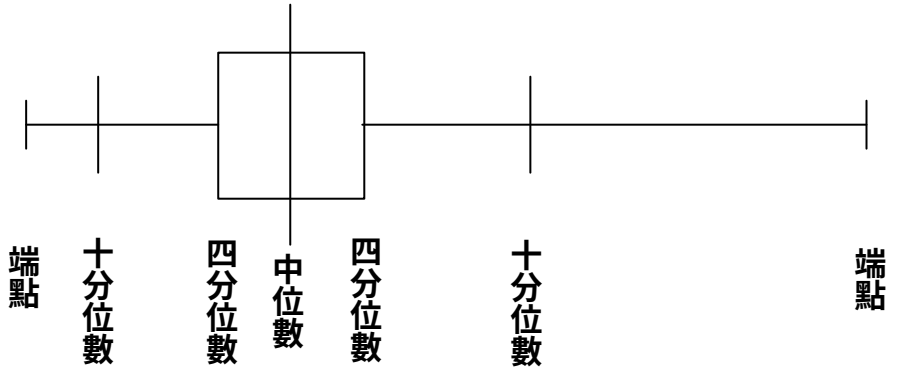


圖 2-2 箱形圖

2.2.2 參數估計

經假設所選定的一或多個分配族後，在此步驟之主要工作為具體指出其參數值，以便在模擬中能使用完整且明確的分配。若樣本 $X_1, X_2, .. X_n$ 有同型且各自獨立之分配(IID Independent Identical Distribution)，除了可從此種樣本中推論可能符合的分配外，亦可於這些資料中估計其參數。上述這種透過資料直接來求得未知參數值的

方法稱為估計參數值。

一般所謂估計量(estimator)即為樣本資料的數值函數，在一給定的分配中要獲得一特定參數的估計量有多種方法可採用，且亦有許多的方法可用來評估此估計量品質的好壞。一般大多採用最大概氏估計式(maximum-likelihood estimators , MLEs)，因為最大概氏估計式有以下的優點：

1. MLEs 有幾個有用的工具作為找尋估計值時的選擇。如最小平方估計量、不偏估計量與動差法等。
2. 使用 MLEs 的結果可利用卡方(chi-square)進行適合度檢定。
3. MLEs 有一較明顯的直覺式的裁定。

而關於最大概氏估計式有兩點需要注意：

1. 取自一分配之有限個樣本所估計之參數，係屬於估計參數，而非真正參數。
2. 若分配之選擇不佳，一最大概氏估計不會產生一良好之結果，例如，若對於成指數分配之隨機數字，估計為常態分配之參數，則這些參數可能描述對於該項數據之最近似常態分配。然而，所估計之分配對抽取的樣品分配來說不是一個好的分配估計。

2.2.3 以卡方(Chi-square test)進行適合度檢定

最大概氏估計技巧，會產生最佳可能的參數集合。然而，使用最大概氏方法，並不保證估計分配會是該樣本的一個可接受之近似分配。於是需要有一適合度檢定以決定此近似分配的最近似結果。

傳統估計與檢定大都要求母體分配已知，或為常態分配，若母體分配不合乎基本假設時，則傳統方法將會失效。適合度檢定即在假設母體之分配形式下，進行推估及檢定的工作，適合度檢定就是用來檢驗對母體分配所建立的假設是否存在的統計方法，而卡方檢定及 k-s

即是在進行檢定時所利用的工具。

卡方檢定係作為驗證隨機母體之均勻性的一種方法，該項卡方檢定之應用為其作為一最適度度檢定更通用函數之一特例。當卡方被應用於一有估計參數之分配時，自由度之計算會改變。若 S 個參數被估計，則自由度為 K-S-1。對於卡方統計量之一定值，減低自由度使接受試驗更為嚴格。

2.3 輸入資料分析問題點

於 2.2 小節所述輸入資料分析的步驟於實際的應用中，常會遭遇一個窘境即在第 2 步驟中，模擬模式建構者可能嘗試過所有已知的機率分配，但仍無法在所要求的信賴度下通過適合度檢驗。在此狀況下，模擬模式建構者可能被迫採取運用經驗分配 (empirical distribution)。如果將樣本點 $X_{(i)}$ 由小至大排列，即 $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$，則經驗分配可寫成如下：

$$F(x)=\begin{cases} 0, & \text{if } x < X_{(1)} \\ \dfrac{i\text{-}1}{n-1}+\dfrac{x-X_{(i)}}{(n-1)(X_{(i+1)}-X_{(i)})}, & \text{if } X_{(i)} \le x < X_{(i+1)}, \text{for } i=1,\wedge,n \\ 1, & X_{(n)} \le x \end{cases}$$

Kelton and Law 所評論，在實際運用經驗分配時，常遭遇一些問題：大部分的連續型(continuous)機率分配都有偏度偏右(skewed to the right)的傾向；如果樣本數不大時，則樣本由機率密度函數右尾部分所產生的希望則非常渺茫。而且上述的經驗分配機率密度函數無法產生比 $X_{(n)}$ 大的隨機變數，這會造成系統模擬中一些嚴重的問題，例如：一個特長的服務時間，極可能在等候線系統中造成嚴重的雍塞。如果無法產生如此的隨機變數，則決策者將無法觀察到系統會有如此的現象，而對系統的特性有所誤判。

誠如 Kelton and Law 所評論，進行輸入資料分析的過程中，運用經驗分配所可能遭遇偏度偏右的問題外，樣本數大小亦是可能問題點

之一。傳統在資料分析、統計檢定或品質管制的分析過程中，分析者常假設樣本服從常態分配，藉常態分配的特性，如分配形狀為左右對稱若鐘形之曲線，配合標準差的觀念轉化為標準常態分配、任何點與平均數 X 間在常態曲線下之面積是一定且已知的等，對所蒐集整理的樣本資料進行分析。但在此假設過程中，當樣本數不大(小於 30)時，便無法以大數法則使其服從常態分配，如此與經驗分配一般，無法完整且正確的描述母體，導致分析結果產生誤差。

2.4 強森分配轉換系統

Johnson[20]提出強森常態轉換系統對一連續單變量母體建模，利用樣本之平均數、變異數、峰態(kurtosis)、偏態(skewness)以樣本之動差進行強森常態轉換系統中參數之擬合，並將非常態之資料轉換為標準常態之資料，常態轉換式如所示，強森轉換系統及其擬合方法於本研究中第三章將有一詳細討。

$$Z = g + d.\ f\left(\frac{X - e}{l}\right)$$

其中

γ及 δ為型態參數(shape parameters),

λ為尺度參數(scale parameter),

ε為位置參數(location parameter)

其後 Hahn and Shapiro[14]針於強森分配中各分配參數界限做一詳細討論。Johnson and Lowe[19]針對樣本之峰態及偏態做一探討，求得峰態及偏態之上下界。

Slifker and Shapiro[33]利用百分比擬合法求強森分配中四個參數值。Mage(1980)[26]探討 S_B 分配族運用百分比擬合法，針對四個百分比值進行討論。Bowman and Shenton[4]針對百分比擬合法中四個百分比值之間具有某些倍數關係時，其對參數做一探討並使用導出之公式求出參數，另外也針對在母體不偏情形下(偏態為零)時求得其四個參

數。Bowman and Shenton[5]亦研究在 S_B 和 S_U 分配族中，百分比擬合法的四個樣本百分比值在任何倍數的情況下均可運用插補方法求出強森的四個參數。

Swain, Venkatraman and Wilson[37]提出最小平方法擬合樣本資料，並求出強森轉換系統中四個參數。DeBrota, et al[12]提出使用視覺方法來進行資料的擬合。Chou, et al(1994)[10]提出一些常用的分配如:Uniform、Beta、Exponential、Chi-square、Gamma、F、Normal、t、Logistic 都可依(1)樣本範圍(2)峰態及偏態(3)Quartile ratio 之一進行分類成強森分配中四個分配族。

在強森轉換系統應用方面，Spedding and Rawlings[34]運用強森轉換系統於統計製程管制上，將非常態資料轉換為常態資料後與管制界線的使用做一探討。Chou, Polansky and Mason[9]提出一程序將非常態資料轉為常態後應用於品質管制上。Simpson 以 Chou, Polansky and Mason 等所提之程序，透過一強森轉換式，將非常態之資料型態轉成為常態。Hubele and Lawrence[18]利用資料轉換來進行 C_{pk} 指標的估計，提出藉由資料的轉換有助於估計一區間之 C_{pk} 值和未合格區零件之數目。Hsin-Hung Wu[41]提出使用強森分配族群的方式來計算 Clement 的方法以改善以偏態及峰態值及使用順序統計量來決定 99.865、50、0.135 等三個百分位數之值所造成的不穩定現象。

2.5 資料分析技術的應用

大量資料中潛藏著許多有用的特徵或知識。若能妥善的進行資料分析與探勘，挖掘出這些隱藏的特徵，可以提供主管，做為決策的重要依據[50]。本研究透過強森分配擬合法則對樣本資料進行擬合，針對其擬合結果進行資料分析，擬藉此找出資料中有用的訊息，提供決策人員參考。上述經由萃取找出隱藏於資料中有用資訊或知識的過程，乃是資料挖掘(data mining)的一個部份。本研究應用邏輯迴歸進行資料分析，以下針對資料挖掘的內涵及方法進行介紹，並探討迴歸

技術在其中所扮演的角色。

2.5.1 資料挖掘的內涵

Data Mining 是指找尋隱藏在資料中的訊息，如趨勢(Trend)、特徵(Pattern)及相關性(Relationship)的過程，也就是從資料中發掘資訊或知識(有人稱為 Knowledge Discovery in Databases, KDD)，或稱為「資料樣型分析」(Data Pattern Analysis)或「功能相依分析」(Functional Dependency Analysis)等等。資料採礦所要處理的問題，就是在龐大的資料庫中尋找出有價值的隱藏事件，並且加以分析。而其主要的貢獻在於，它能從資料庫中獲取有意義的資訊及對資料歸納出有結構的模式，以作為企業在進行決策時之參考依據。常見資料挖掘型態有下列八項[45]：特徵描述(characterization)、區辨法則(discriminate rules)、類別法則(classification rule)、關聯法則(association rule)、群集規則(clustering rule)、預測(prediction)、偏差(deviation mining)、連續性樣式(sequential pattern)等等。

2.5.2 迴歸技術於資料挖掘的應用

上述八種資料挖掘型態中，區別法則主要用在區分兩個類別之間的不同特色或性質，被測者稱為目的類別，其他類別稱為對照類別。此區分類別與統計之區別分析(discriminant analysis)其目的皆為對兩個不同資料類別做一個區分。區別分析乃在於計算一組「預測變項」的線性組合，對依變項加以分類，並檢查其再分組的正確率，與自變項間的線性關係[42]。區別分析與多變項變異數分析及多元迴歸分析有密切關係，預測變項的線性組合類似多元迴歸方程式右邊乘積和，區別分析中它是變項與區別函數係數的乘積總和。

統計中的群體規則，為以群集分析方法，依各事物的特性分別出來，主要以相對位置遠近作為分群依據。而群集分析也是一種多變量分析程式，其目的在於將資料分成幾個相異性最大的群組，而群組內

的相似程度最高。由於群集分析時，使用之分析方法不同，結果便有所不同，不同研究者對同一觀察值進行群集分析時，所決定群集數也未必一致，因此在研究應用上，常與區別分析一起使用。

預測它能判斷或預測某些遺失資料可能的值或是在一群物件中確定屬性值的分佈，基於選擇類似的物件利用某些統計分析來發現一群有相關或有興趣的屬性及預測值的分佈。如線性迴歸、多元迴歸等分析方法。

2.5.3 邏輯迴歸(logistic regression)介紹

迴歸技術應用於資料挖掘的範圍廣泛，透過所求線性方程式即區別函數來區分資料中不同的特徵或屬性，規範出在不同資料特徵間之分界和其所適之區域。若資料分類方法屬二分數據的分析，如二分依變項(或稱反應變項)值是上榜、落榜的結果，或疾病經過治療後治癒、復發的兩種可能。陳世雄[8]指出當一研究有二分的依變數時最好使用邏輯迴歸來分析資料。

邏輯迴歸分析的目的是為了找出這個依變項值與一組連續變項(或稱自變項)之間的線性關係。這個線性關係的表示可用依變項的對數奇數比單位(logit)、常態數單位(normit)或雙對數單位(log-log)等三類線性函數表示法。Fisher 和 Yates 第一次以 logit 處理有關二元結果的資料分析，Cox[11]於其著作中討論 logistic 迴歸模式的應用與影響，柳克婷[46]指出若二變數中有一為數值變值或為有自然大小順序的類別變數，另一為類別變數，且二者間有線性關係時，應考慮二變數間的 logisatic 迴歸模式自可得到更多的資訊以供參考。

本研究分析在二維平面上資料的分佈情況，應用邏輯迴歸模式，判斷不同擬合法則間之線性關係，藉以規範不同資料間之適用區域，以提供決策時的參考，邏輯迴歸的模型將於以後章節進行討論。

2.6 本章結論

綜合以上文獻探討，可以歸納出輸入資料分析對系統模擬正確與否佔有關鍵性的角色，其中尤以分析的第二步驟，母體機率分配假設的部份影響最大。而藉由問題點的探討了解到假設母體分配的困難處及歸納出藉由強森分配轉換系統可針對此間題進行解決。

目前有關強森分配擬合法於輸入資料分析的應用上，並未有學者針對二維之(β_1, β_2)平面，判斷在不同的樣本特徵下，以何種擬合法進行擬合其效果優劣的比較。因此，本研究將針對此議題，透過遺傳演算法產生樣本數據，經過資料的擬合，分析及整理，提出一判斷在不同樣本特徵下適合使用的擬合方法，提供分析師、企業決策者在進行資料分析及決策時的參考。

第三章　強森分配數學模式

在約五十年前，Johnson **即提出強森分配轉換系統**(Johnson translation system of distributions)**，其貢獻是能有效地對連續單變量母體建模，將不具常態性質的資料**(non-normal data)**轉為成具有常態性質的資料點，相關的研究在多數書籍中可找到更詳細、完整的討論**[23,3,36,21]**。雖然統計上有** Beta **分配、**Gamma **分配、**Erlang **分配、**Lognormal **分配、**Log-logistic **分配、**Weibull **分配、**Pearson **分配等都具有型態多樣性的特性。但強森分配轉換系統不僅能做資料擬合之分配，更能將非常態資料轉換成常態分配資料，這一點在實務應用上極具價值，如在統計製程管制及品質管制上的應用**

本節將詳細討論強森分配轉換系統的數學模式，及擬合強森分配的四種方法。

3.1 強森分配數學模式

1.**常態轉換**（Normalizing Translation)**。**

(1)**變數定義（**Nomenclature**）**

a. X **為一個連續的隨機變數。**

b. X **的累積分配函數**(CDF)**為：**

$$F(x) = \Pr\{X \le x\} \quad -\infty < x < \infty$$

c. X **的機率分配函數**(PDF)**為：**

$$P(x) = F'(x) \qquad -\infty < x < \infty$$

d. X **的相關動差有：**

$$m \equiv E(X) \ \ and \ \ m_c \equiv E\left[(X-m)^c\right],\ c = 2,3,4$$

e. X **的偏態係數**(Skewness)**和峰態係數**(kurtosis)**分別為：**

$$\sqrt{b_1} \equiv \frac{m_3}{m_2^{3/2}} \quad and \quad b_2 \equiv \frac{m_4}{m_2^2}$$

(2)**常態轉換的通式**（General Form of the Normalizing Translation）

$$Z = g + d \cdot f\left(\frac{X - e}{l}\right)$$

其中

γ及 δ為型態參數(shape parameters),

λ為尺度參數(scale parameter),

ε為位置參數(location parameter),

依據不同的分配族，再由下列四種函數選擇其一套入：

$$f(y) = \begin{cases} \ln(y) & for\ the\ S_L\ (\log normal)\ family \\ \ln\left[y + \sqrt{y^2 + 1}\right] & for\ the\ S_U\ (unbounded)\ family \\ \ln\left[y/(1-y)\right] & for\ the\ S_B\ (bounded)\ family \\ y & for\ the\ S_N\ (normal)\ family \end{cases}$$

(3)**設定條件**（Conventions）

a. **其中我們取** $\delta > 0$ **及** $\lambda > 0$。

b. **對於** S_N (normal)**族，我們取** $\lambda \equiv 1$ **及** $\varepsilon \equiv 0$。

c. **對於** S_L (lognormal)**族，我們取** $\lambda \equiv 1$。

2.**產生強森隨機變數的方式**（Generating Johnson Variates）

在選定適當的分配族，完成常態轉換之後，模擬模式建構者可依據不同的分配族及給定之標準常態變數 Z **值, 由下列四種反函數代入以求取** X **的值，以產生隨機變數：**

$$X = e + l \cdot f^{-1}\left(\frac{Z - g}{d}\right),$$

其中

$$f^{-1}(z) = \begin{cases} e^z, & for\ the\ S_L(\log normal)\ family, \\ \frac{1}{2}\left(e^z - e^{-z}\right), & for\ the\ S_U\ (unbounded)\ family, \\ 1/\left(1 + e^{-z}\right), & for\ the\ S_B\ (bounded)\ family, \\ z, & for\ the\ S_N\ (normal)\ family. \end{cases}$$

3.**強森分配密度函數的特性**（Properties of Johnson Densities）

四種不同的分配族的強森分配，其機率密度函數可以下式表示：

$$p(x)=\frac{\delta}{\lambda\sqrt{2\pi}}f'\left(\frac{x-\varepsilon}{\lambda}\right)\exp\left\{-\frac{1}{2}\left[\gamma+\delta\cdot f\left(\frac{x-\varepsilon}{\lambda}\right)\right]^2\right\}$$ ， **對於** $x\in H$

其中

$$f'(y)=\begin{cases}\dfrac{1}{y} & for\ the\ S_L(\log normal)\ family,\\ 1\Big/\sqrt{y^2+1} & for\ the\ S_U(unbonded)\ family,\\ 1/[y(1-y)] & for\ S_B(bounded)\ family,\\ 1 & for\ the\ S_N(normal)\ family,\end{cases}$$

H **在不同的分配形式時，其支援集合**(support set)**為：**

$$H=\begin{cases}[\varepsilon,+\infty) & for\ the\ S_L(\log normal)\ family,\\ [-\infty,+\infty] & for\ the\ S_U(unbounded)\ family,\\ [\varepsilon,\varepsilon+\lambda] & for\ the\ S_B(bounded)\ family,\\ (-\infty,+\infty) & for\ the\ S_N(normal)\ family,\end{cases}$$

3.2 強森分配擬合方法

由文獻資料中，我們歸納出以強森分配擬合(Johnson distribution fitting)**樣本資料點可分為下列四種方法：**

1.**動差擬合法**(Moment Matching)

2.**百分比擬合法**(Percentile Matching)

3.**最小平方法**(Least Squares)

4.**最小** L_p-norm **法**(Minimum L_p Norm Estimation)

因為這四種擬合方法為本研究計劃的研究核心，我們在下列針對每一種擬合方法及其精神，逐一進行概述。

3.2.1 動差擬合法(Moment Matching)

樣本資料的前四個動差常被使用於描述機率分配機率函數的型態[6,35]**。動差擬合法的理論背景是：針對一組樣本的偏態**(Skewness)

及峰態(Kurtosis)，僅有一個強森分配其偏態及峰態恰能擬合其值。故動差擬合法將以樣本資料的偏態及峰態的值，搜尋欲擬合的強森分配。動差擬合法的相關細節，可參考如下的論文：Hill, Hill and Holder [16]，Venkatraman and Wilson[39]及 Debrota et al[12]。而動差擬合法其運用的過程，可簡述如下：

(1) 目標 CDF 函數 $F(\cdot)$之隨機樣本$\{x_j : j = 1,\ldots,n\}$

(2) 求出樣本資料的前兩個動差(moment) $\overline{X}$ 及 s^2，並以下列公式計算偏態估計值 $\hat{a}_1 = \sqrt{b_1}$ 及峰態估計值 $\hat{a}_2 = b_2$ ：

$$\hat{a}_1 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_j - \overline{X}}{s}\right)^3 , \quad \hat{a}_2 = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_j - \overline{X}}{s}\right)^4$$

(3) 依(β_1, β_2)及圖 3-1 判斷出所適合的強森分配族類(family)。

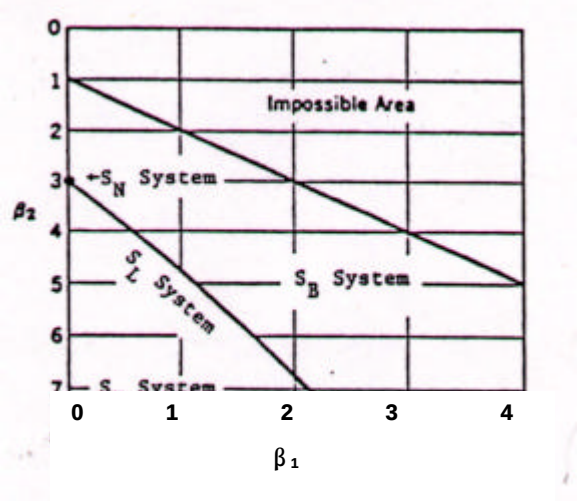


圖 3-1 β_1與β_2之關係圖

(4) 以非線性規劃的演算法擬合強森分配。

(a) 依步驟(3)所判斷出的強森分配族類，導出適當的偏態(Skewness)及峰態(Kurtosis)的函數式：

$$x = x + l.\ f^{-1}\left(\frac{Z-g}{d}\right), \text{ 令 } W = f^{-1}\left(\frac{Z-g}{d}\right)$$

$$a_1 = \sqrt{b_1} \equiv \frac{m_3}{m_2^{3/2}} = E\left[W^3\right]$$

$$= \int_{-\infty}^{\infty}\left[f^{-1}\left(\frac{Z-g}{d}\right)\right]^3 \frac{1}{\sqrt{2p}} e^{-Z^2/2}\, dZ \quad (skewness)$$

$$\alpha_2 = \beta_2 \equiv \frac{m_4}{m_2^2} = E[W^4] \quad (kurtosis)$$

(b) **以下列二個非線性聯立方程式，找出強森分配的兩個參數 γ 和 δ 的估計值** $(\hat{\gamma}, \hat{\delta})$**：**

$$\begin{cases} \alpha_1(W; \hat{\gamma}, \hat{\delta}) = \hat{\alpha}_1 \quad \wedge \ (1) \\ \alpha_2(W; \hat{\gamma}, \hat{\delta}) = \hat{\alpha}_2 \quad \wedge \ (2) \end{cases}$$

運用如下的方法求解：

(I) Venkatraman and Wilson[39]**採用** Levenberg-Marquardt **的最佳化演算法進行求解；請參見** Marquardt[27]**的論文或** Avriel[1]**的書。**

(II) **利用數學軟體求解，如** LINGO[25]**或** MATLAB (**參考** Hanselman and Littlefield[15]**的書**)**。**

(c) **以變數轉換的方法，求出強森分配的另外兩個參數** ξ **和** λ **的估計值** $(\hat{\xi}, \hat{\lambda})$**：**

$$X = \xi + \lambda \cdot W \quad ,$$

$$m_x = E[X] = \xi + \lambda E[W] = \xi + \lambda m_w$$

$$\sigma_x^2 = Var[X] = \lambda^2 \cdot Var[X] = \lambda^2 \sigma_W^2$$

將樣本代入，得

$$\begin{cases} \bar{x} = \hat{\xi} + \hat{\lambda}\,\bar{W} \\ S_x = \hat{\lambda}\, S_W \end{cases} \quad \Rightarrow \quad \begin{cases} \hat{\lambda} = \dfrac{S_x}{S_W} \\ \hat{\xi} = \dfrac{\bar{x} - \hat{\lambda}}{\bar{W}} \end{cases}$$

3.2.2 百分比擬合法（Percentile Matching）

樣本資料的百分位數(Percentiles)**也常被使用於描述機率分配機率函數的型態如** Kahneman, Slovic and Tversky[22] **及** Doubilet et al [13]**的論文所提。百分比擬合法相關的細節，可參考**[26,33,10,9,29]**等的論文。而百分比擬合法其運用的過程，可簡述如下：**

(1) **選擇** k **個百分比**$\{a_j : j = 1,..,k\}$**在**(0,1)**範圍內。**

(2) **要估計強森分配的** k **個參數，須在強森分配所要估計的目標參數中選擇** k **個百分比。**

$$\left\{\hat{x}_{a_j} : j = 1,...,k\right\}$$

(3) **根據樣本所求出的**b_1, b_2**的值選擇適當的分配族類，以適當的** $f(\cdot)$**代入下列的標準轉換式。**

$$Z = g + d \cdot f\left(\frac{X - x}{l}\right)$$

(4) **令樣本之**$\hat{x}_{a_j}$ **等於強森分配常態轉換系統中**Z_{a_j} **所對應的值。**

$$\hat{x}_{a_j} = x + l \cdot f^{-1}\left(\frac{z_{a_j} - g}{d}\right), \quad j = 1,..., k$$

解上式聯立的 k **個非線性方程式，求得強森分配的四個參數**$\{g, d, l, x\}$**。**

3.2.3 最小平方法（Least Squares）

假設 $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$**為自** $F_X(\cdot)$**所取出之隨機樣本，則最小平方法的理論基礎為：被轉換過的隨機變數** $R_j = F_X(X_{(j)})$**為** n **個由均勻分配中隨機抽樣而得的隨機樣本中，第** j **小的樣本值**(that is, $F_X(X_{(j)})$ has the distribution of the j^{th} uniform order statistic; **本理論請參考** Kendall and Stuart [23]**一書)。若是**

$F_X(x) \cong F_X(x) = \Phi\left\{g + d \cdot f\left[\frac{x - x}{l}\right]\right\}$, $-\infty < x < \infty$,

其中$\Phi(\cdot)$**常態分配的累積機率分配，則** $R_j = \Phi\left\{g + d \cdot f\left[\frac{x - x}{l}\right]\right\}$ (for j = 1,..,n)。

假設$m_{Ri} = r_j$**，則最小平方法在於找尋一強森分配（由其四個參數所決定之），使** R **與** ρ **在** n **維的歐氏空間**(Euclidean space)**中其距離為最短（及誤差的最小平方和為最小）。最小平方法相關的細節，可參考**[37,12]**兩篇論文。最小平方法其運用的過程，可簡述如下：**

(1)**有一順序統計量，將樣本值順序排列，**

$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$

(2)**第** i **個** uniformized **順序統計量如下：**

$$U_{(i)} = \Phi\left\{g + d \cdot f\left[\frac{x_{(i)} - x}{l}\right]\right\} \quad , i = 1,\ldots,n$$

(3)**第** i **個順序統計量**

$$\text{mean} = r_i = \frac{i}{n+1}$$

$$\text{variance} = \frac{r_i(1-r_i)}{n+2}$$

(4)**令** $e_i \equiv U_{(i)} - r_i$ **則**

$$E[e_i] = 0 \quad and \quad Var[e_i] = \frac{r_i(1-r_i)}{(n+2)}$$

(5) **根據樣本選擇適當的** f(·)**，並指派** weights{w_i}**給**{e_i}

(6) **解下列非線性規劃問題：**

$$\underset{g,d,l,x}{\text{minimize}} \quad \sum_{i=1}^{n} w_i \, x_i^2$$

subject to $d > 0$

$$l \begin{cases} > 0 & for\ S_U, \\ > x_{(n)} - x & for\ S_B, \\ = 1 & for\ S_L\ and\ S_N \end{cases}$$

$$x \begin{cases} < x_{(1)} & for\ S_L and\ S_B \\ = 0 & for\ S_N \end{cases}$$

(7) **當** $w_i = 1$ **時是求出** ordinary least square(OLS)**估計值，**

當 $w_i = \frac{1}{Var(e_i)} \quad for\ i = 1,\ldots,n$ **時** diagonally-weighted least squares **估計值。**

3.2.4 最小 L_p-norm 法（Minimum L_p-norm Estimation）

最小 L_p-norm **法與最小平方法的涵義十分接近，都意在找尋位於** n **維的歐氏空間**(Euclidean space)**中其距離為最短的強森分配。但是於**

最小 L_p-norm **法中距離以** L_p-norm **定義；即**

$$\left\| F_n - \hat{F} \right\|_p \equiv \left[\int_{-\infty}^{\infty} \left| F_n(x) - \hat{F}(x) \right|^p d\hat{F}(x) \right]^{1/p}$$ **。**

(1)**經驗分配函數**

$$F_n(x) \equiv \frac{number\ of\ x_j \le x}{n}, \quad -\infty < x < \infty$$

(2)**擬合的分配函數**

$$\hat{F}(x) = \Phi\left\{ g + d \cdot f\left[\frac{x - x}{l} \right] \right\}, -\infty < x < \infty$$

(3) *Minimize* $L_p - norm$**，求得強森分配的四個參數：**

$$\underset{g,d,l,x}{\text{minimize}} \quad \left\| F_n - \hat{F}_n \right\|_p$$

$$Subject\ to \quad d > 0$$

$$l \begin{cases} > 0 & for\ S_U, \\ > x_{(n)} - x & for\ S_B, \\ = 1 & for\ S_L\ and\ S_N \end{cases}$$

$$x \begin{cases} < x_{(1)} & for\ S_L\ and\ S_B \\ = 0 & for\ S_N \end{cases}$$

本研究在 p **的選擇上，以** p **為** Infinity **時之特例作為本研究** p **之次數選擇時之依據。**

3.3 遺傳演算法(Genetic algorithm)產生樣本資料點

遺傳演算法係由 John Holland **於** 1975 **年首度發表。經過了多年的發展，遺傳演算被証明為一有效的最佳化的搜尋方法。在最近幾年，許多學者投入這個領域繼續對演化式計算做更深一層的探索。綜觀人工智慧在最近幾年的發展，相信模糊理論、類神經網路與遺傳演算法將是重要的研究方向。**

3.3.1 遺傳演算法的介紹

GA 主要是以達爾文的「進化論」為基礎，模擬生物界依「適者生存，不適者淘汰」的生存演化法則；而 GA 的主要目的在於建立一個保有自然特性的「人工遺傳系統」，以模擬和解釋生物自然進化的過程。

GA 主要是以操作染色體來進行演化過程，在反覆演化的過程中，可看成是在問題的可行區域中做系統化的多維空間搜尋，而其特點為多點搜尋、只需適應值資訊、轉移規則是隨機性而非決定性的。另外 GA 的搜尋方式是屬於平行式而非循序的，這點和 Tabu、模擬退火法等的搜尋方式有很大的不同。本研究使用簡單遺傳演算法(SGA)。

GA 的搜尋技術是以隨機搜尋為架構，但是 GA 絕非僅是一種單純的隨機搜尋方法，因為 GA 保存了演化過程中所提供的資訊，所以能展現出比單純的隨機搜尋方式更好的求解能力。GA 的優點在於它是一種穩健且有效的搜尋技術，而且相較於其它演算法，GA 有較小的機率會陷入局部最佳解中；而 GA 的缺點則在於計算時間長，但是此缺點也由於電腦技術的進步而漸漸的克服了。如今 GA 被廣泛的應用在各領域；如參數設計、機器人、排程問題、分類系統、控制系統工程等等，且都有不錯的成果。而 GA 之操作流程如圖 3-2，以下介紹 GA 相關之名詞及操作因子。

1.**母代**(Population)

所謂的母代就是指每一世代中共同生存在環境中的個體。其主要目的是為了提供表現不同的個體。由於個體間的差異性，天擇的結果與基因交換的機制，才能產生更好的下一代，如此週而復始，演化便持續的進行。

2.**染色體**(chromosome)

在生物界中，遺傳性狀的最基本單位被稱之為基因(gene)。而相關連

的基因則以一連串的方式存在於染色體中，因此，藉由染色體，便可決定該個體之遺傳特徵。所有個體皆由一組編碼在染色體中的數據來表示。儲存在染色體中的資訊將影響該個體在模擬環境中的表現。

3.**適應值**(fitness)

依爾文進化論的觀點，對環境適存存機率，更有機會繁衍下一代。在自然界中，生物的演化是天擇的結果。而天擇的依據即為該個體的 fitness。在遺傳演算法中，同樣需要 fitness 的機制。Fitness 的定義在遺傳演算法中扮演的角色，將引導演化的方向。由於 Fitness 使得演化將不是漫無目的，而是朝著所要求的方向前進。

4.**交配**(crossover)

交配的目的在於希望能夠製造出同時兼具親代優點的新個體。然而子代亦可能同時遺傳親代的缺點，交配不一定保証能造出更好的子代，但是透過天擇的結果，較差的子代自然會被淘汰。遺傳演算法中，仍然保留了算然界中 crossover 的機制，以成混合二的。其交配的方法有單點交配、兩點交配及均等交配三種，在實際的應用時，究竟該採用何種交配方法，將依不同問題而定。

5.**突變**(mutation)

純粹靠交配及重複產生(reproduction)此二操作子並不能夠使得演化造出一個具有新特性的個體。自然界中，生物藉著突變(mutation)來造出新的物種。在遺傳演算法中，仍然藉著自然界中的這個運算，來增加 population 的變化，使演化能盡可能的朝多個新方向進行，而不侷限在少數的個體上。

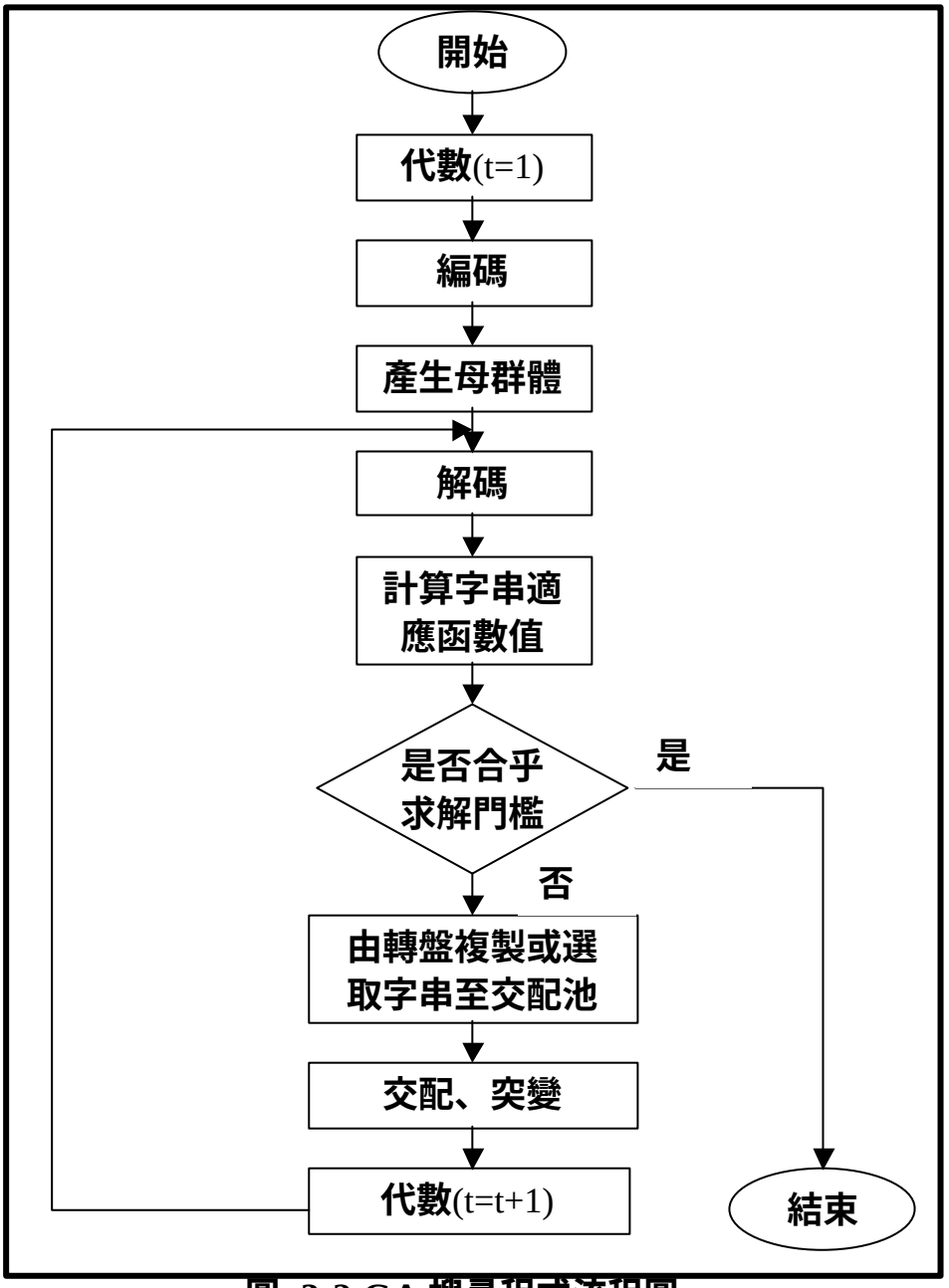


圖 3-2 GA 搜尋程式流程圖

3.3.2 樣本產生程序

本研究在樣本產生過程中使用遺傳演算法，有別以往之數值方法，運用遺傳演算法之特性，不需起始點的情形下即可得到全域最佳解之近似值。本研究在產生樣本點過程共分為四項前置作業，分別敘述如下：

1. **編碼**

本研究為數字的搜尋，故採一般的二進位編碼方式。

2. **定義函數**

樣本之偏態與峰態值需在要求的範圍內，因此為搜尋最小值。

3. **複製**

本研究所使用之複製方法採常用的輪盤法(roulette wheel method)**或稱之為比例選擇法**(proportional selection method)**。盤中每個槽的大小都根據每個個體適應度佔總適應度的百分比來設定的，也就是適應度愈高者所佔據的盤面比例愈大。**

4. **交配率**(crossover rate)**及突變率**(mutation rate)

本研究中 GA **的參數設定非主要的研究對內容，透過不同交配率及突變率以找出最適合之樣本點。本研所採用之交配法則為雙點交配：首先在母代染色體上隨機選擇兩切點，然後將兩交配點間的字元互換。**

3.3.3 樣本產生方式

了解上述 GA **之前置作業，本研究以** Microsoft Visual Basic **軟體撰寫** GA **程式。藉由此演算程式產生本研究所需之樣本數據。以** GA **產生所求之樣本的流程如下**(**圖** 3-3)**：**

1. **利用所撰寫之** GA **程式，隨機產生** 25 **個樣本數據** X_1**、**X_2**、、、**X_{25}**。根據此** 25 **個樣本資料，藉由動差的計算找出其偏態係數平方值** $\hat{a}_1$ **與峰態係數** $\hat{a}_2$ **之值。**
2. **設定欲產生之樣本數據的偏態係數平方值** $\hat{a}^*_1$ **與峰態係數值** $\hat{a}^*_2$**。計算**$(\hat{a}^*_1-\hat{a}_1)^2+(\hat{a}^*_2-\hat{a}_2)^2$**，若此二平方值之和於所要求的誤差範圍內，則此樣本即為所求。若其誤差範圍過大則進行** GA **求解過程。**
3. **以** GA **演算的程序，依次進行複製、交配和突變，在設定的演化次數下，求得最小的誤差值。若無法在誤差範圍下求得所要求的**

樣本數據，則反覆以 GA 進行求解，以求獲得誤差範圍內之樣本數據。

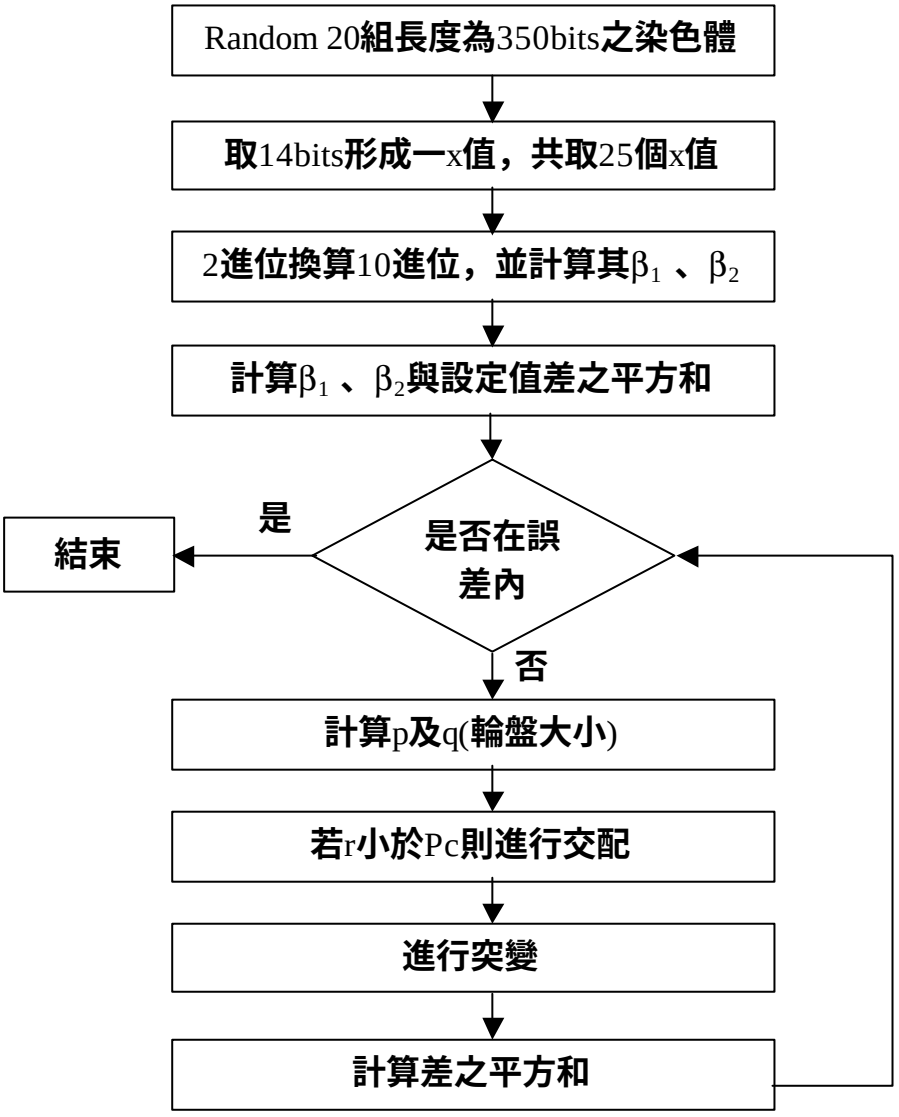


圖 3-3 以 GA 產生樣本之流程圖

3.4 樣本數據

本研究配合遺傳演算法，利用電腦隨機產生 4000 組數據，每一組 25 個樣本數，每組數據皆為分配未知之樣本，每組樣本值介於 0~2500 之間(不為 0)。以下列出二組數據資料以供參考如表 3-1。

表 3-1 Data1、Data2 資料表

Data1	582,802,90,624,465,378,397,406,612,321,873,407,972 1081,501,171, 424,1817,2044,551,252,823,251,995,403			
	平均數	**標準差**	**偏態**	**峰態**
	649.7	457.2	1.643	5.433
Data2	549,831,359,308,64,2030,643,696,702,1070,1068, 1103, 906,909,63,571,221,1014,936,919,579,585,673,1086,1523			
	平均數	**標準差**	**偏態**	**峰態**
	776.3	427.6	0.7245	4.211

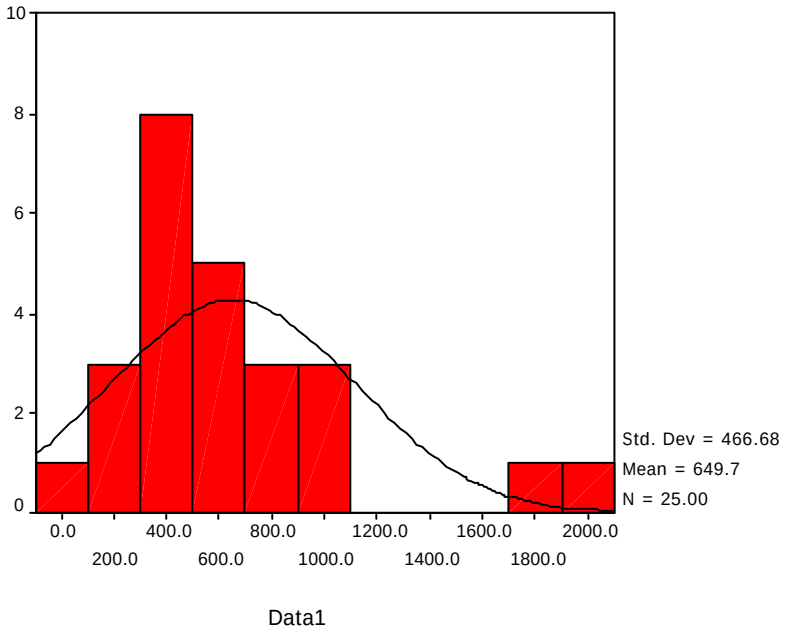
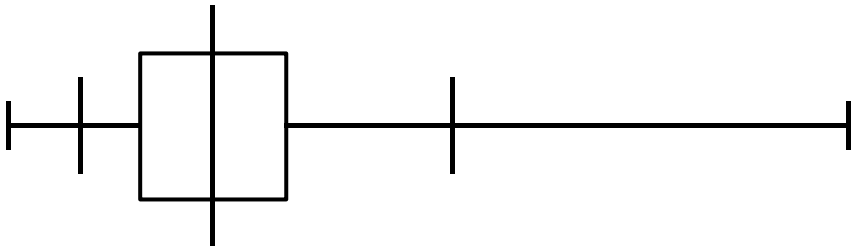


圖 3-3 Data1 之直方圖

圖 3-3 為 Data1 之直方圖，圖 3-4 為 Data1 之箱形圖



90 251 397 501 873 1081 2044

圖 3-4 Data1 之箱形圖

圖 3-5 為 Data2 之直方圖，圖 3-6 為 Data2 之箱形圖

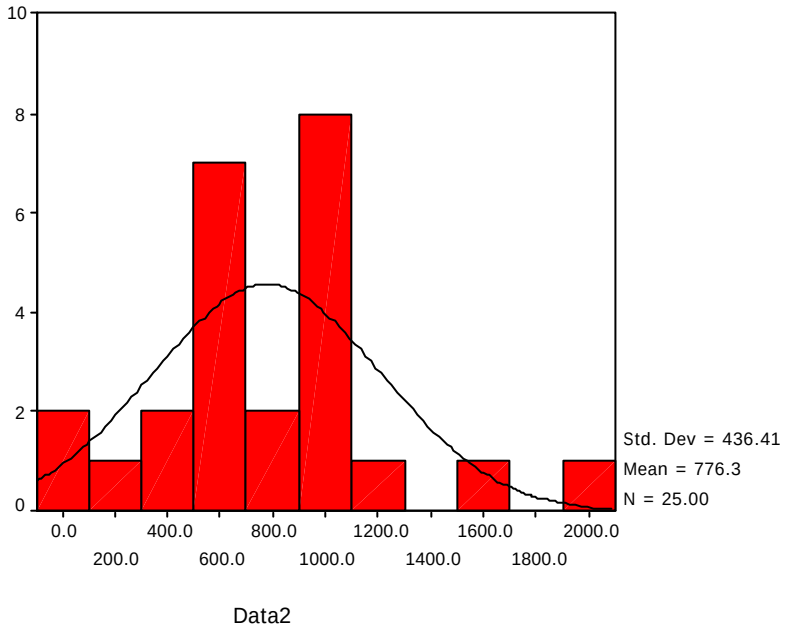
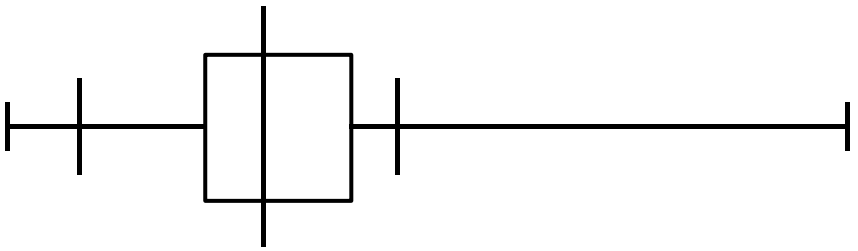


圖 3-5 Data2 之直方圖



63 221 571 702 1068 1103 2030

圖 3-6 Data2 之箱形圖

第四章　樣本資料分析

本章旨在介紹樣本資料分析的內容及程序，並透過整理歸納其結果。藉由第三章所介紹之遺傳演算法產生欲分析之 4000 組樣本資料點，以強森分配之擬合方法對樣本點進行擬合的動作，擬合結果以 K-S 值進行檢定，並在二度之(β_1, β_2)平面表現出不同之區域其適合之擬合方法，並判斷其分布之情況。

4.1 以強森分配進行擬合

以強森分配之四種擬合方法進行資料的擬合，其擬合的步驟如下：

1. 產生樣本點

本研究利用遺傳演算法隨機產生 4000 組樣本資料點，每組 25 個樣本數，每組數據為分配未知之樣本，每組樣本值介於 0~2500 之間(不為 0)。

2. 決定其所屬分配族

根據偏態(Skewness)及峰態(Kurtosis)之二維平面，可判斷出在不同的偏態及峰態值下可界定出其所屬之分配族類(如圖 3-1)，本研究只針對 S_L(lognormal)、S_U(unbounded)、S_B(bounded)、S_N(normal)等四個分配族類加以探討。

3. 以擬合方法進行擬合

決定其所屬之分配族類，選擇強森分配之擬合方法分別對樣本資料進行擬合。在此程序中本研究對四種擬合方法皆進行擬合動作，藉以歸納出在不同的分配族下(亦即在不同的二維平面)，何種擬合方法其擬合結果較佳。

4. 估計強森之四個參數

四種擬合方法各以不同的演算法則求出各組樣本其強森分配估

計之四個參數值(γ δ λ ε)，根據此四個估計參數進行適合度檢定。

5. 統計檢定強森分配之適合度

由步驟四求得之四個參數值, 以統計 K-S 檢定方法進行適合度檢定。判斷的法則為四種擬合方法中 K-S 值最小者即為研究所求之最佳擬合方法。本研究 K-S 值之門檻值選擇的標準為 k-s 值在 0.1 以下為選擇依據，因為在 0.1 的門檻值下，為檢定結果有 90%的信賴水準，故以 0.1 為本研究所設之門檻值。擬合步驟如圖 4-1。

圖 4-1 擬合步驟

4.2 資料整理

根據遺傳演算法產生之 4000 組樣本資料，撰寫 Matlab 程式使樣本之 β_1 及 β_2 分佈於二維平面，其分佈之情況如圖 4-2。圖中標示 S_L(lognormal)、S_U(unbounded)、S_B(bounded)、S_N(normal)等四個分配族類所在之(β_1, β_2)二維平面。本研究產生之樣本資料，務必使其均勻分佈於(β_1, β_2)二維平面，如此方可獲得正確的分析結果。

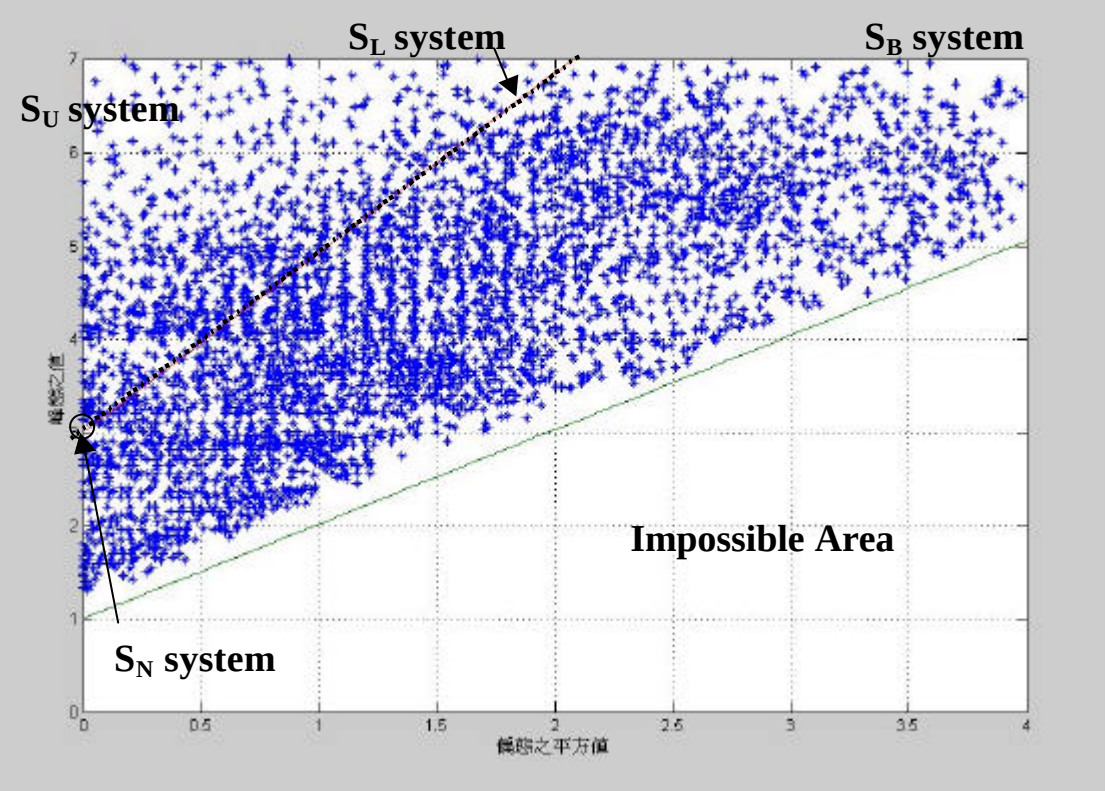


圖 4-2 樣本資料之 β_1、β_2 分圖

以步驟 5 之法則為基準，判斷四種擬合方法中 k-s 值最小者即為最佳之擬合方法。分別將四種擬合方法其所屬之最佳擬合效果的樣本點整理如圖 4-3(動差擬合法)、圖 4-4(百分比擬合法)、圖 4-5(最小平方法)及圖 4-6(最小 L_p-norm 法)。

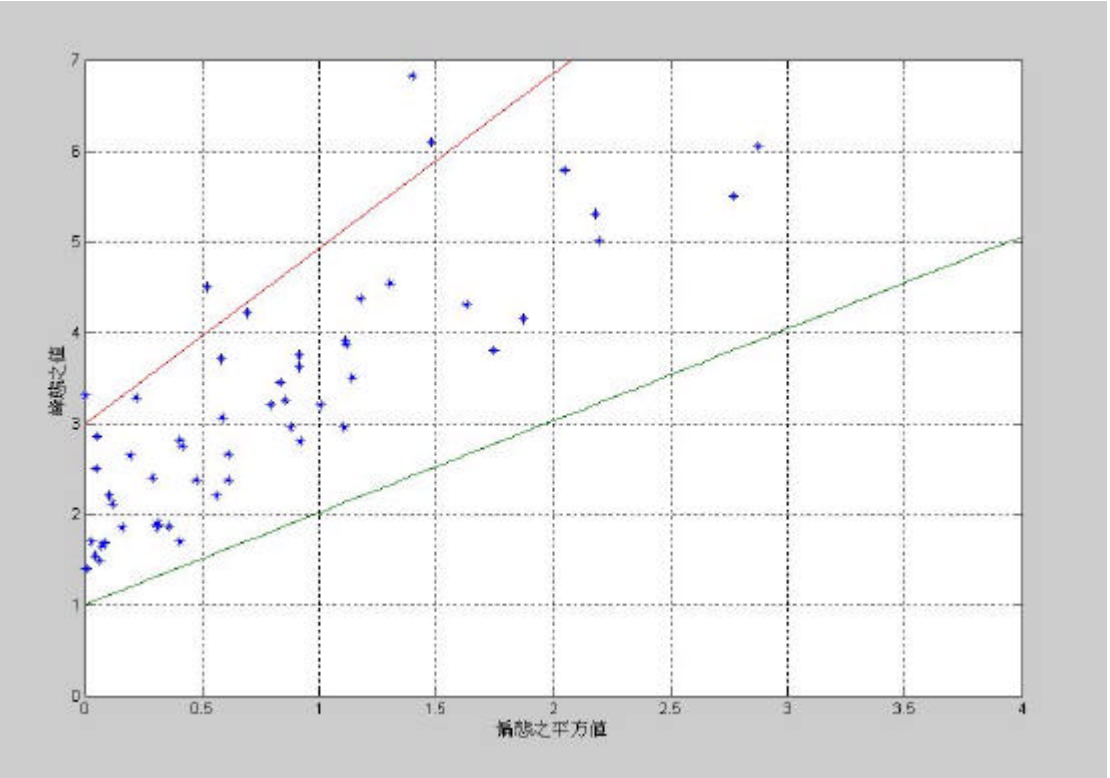


圖 4-3 動差擬合法

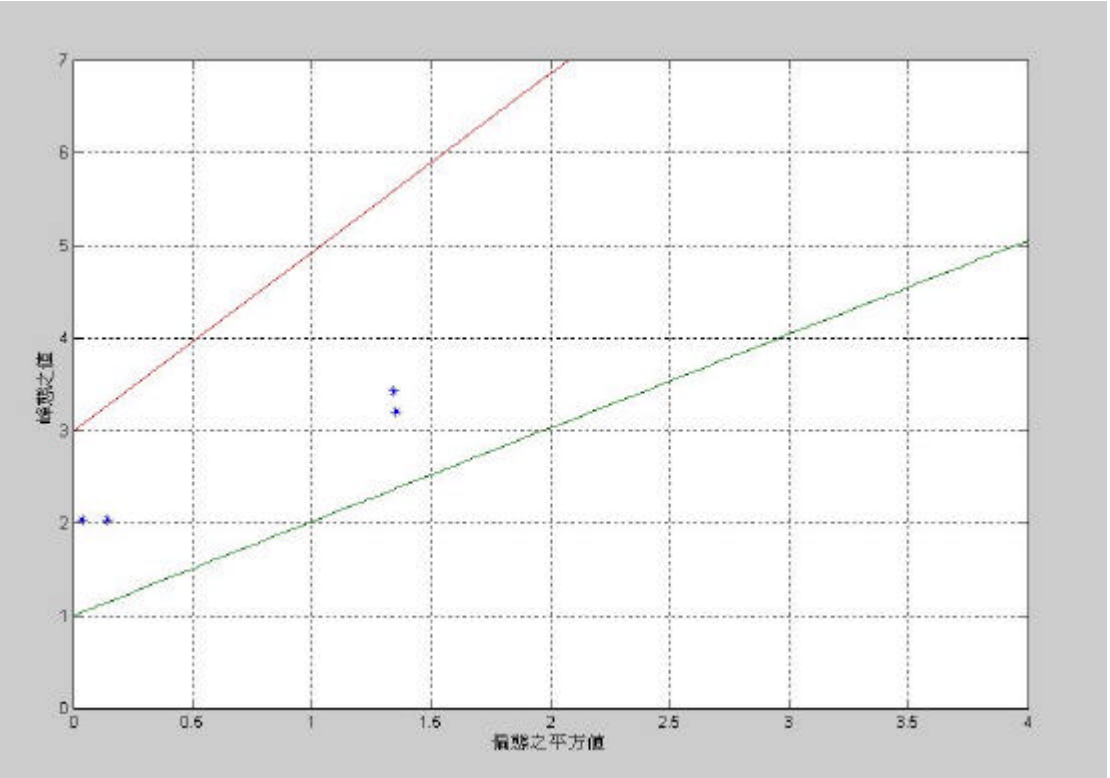


圖 4-4 百分比擬合法

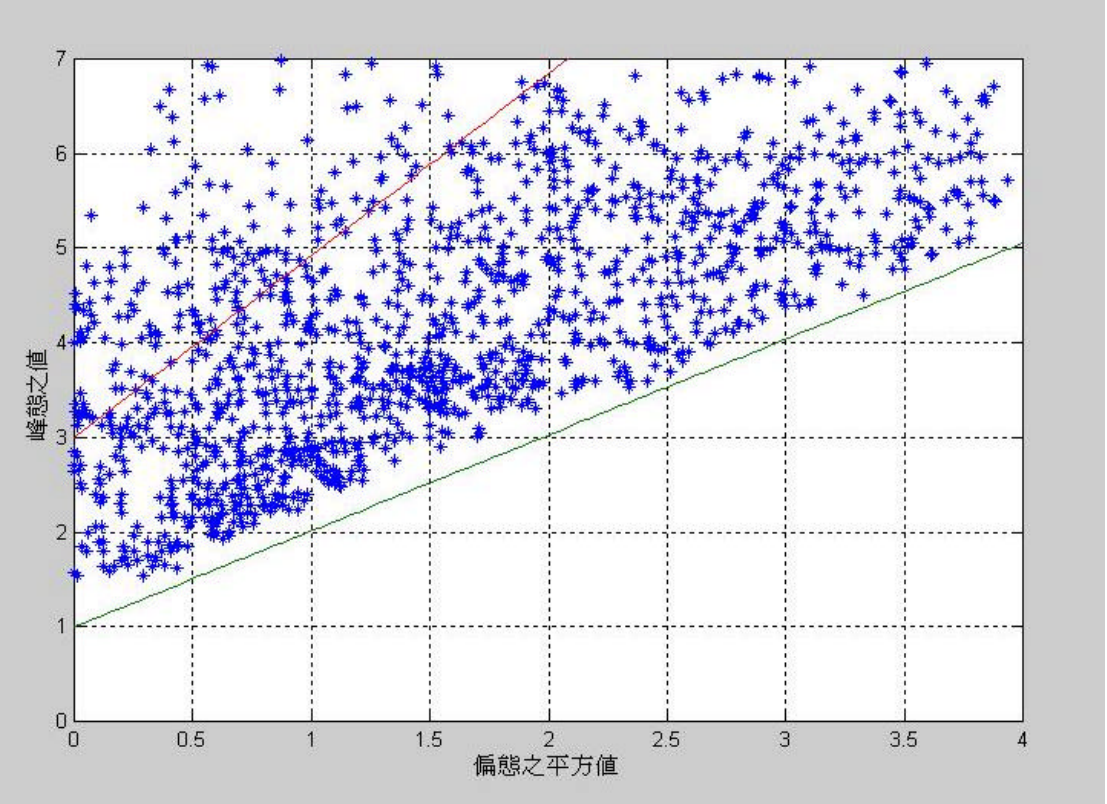


圖 4-5 最小平方擬合法

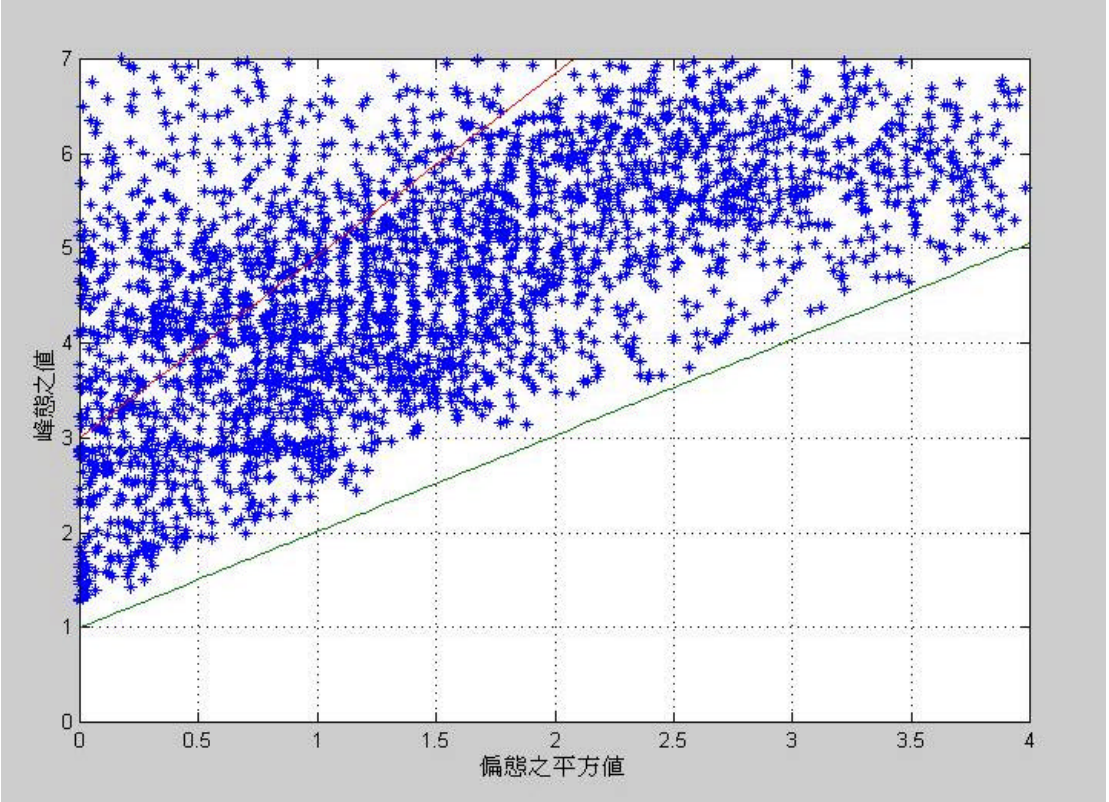


圖 4-6 最小 L_p-norm 法

4.3 擬合結果分析

於 4.2 節中已將四種擬合方法之擬合結果及各別之資料分佈情況做了介紹，本節利用統計檢定與資料分類的技術，試圖針對樣本資料分佈情況進行分析，擬整理出各擬合方法其適合的資料特徵，提供分析者一參考的依據。

4.3.1 擬合結果整理與判別

1. 摒除動差擬合法與百分比擬合法。

根據圖 4-3 及圖 4-4 可發現，屬此二擬合法效果最佳之樣本點在所有樣本集合中僅佔 1.425%(57/4000)，因其所佔比率極小，所以此二擬合法可予忽略。亦即決策者進行資料分析時，可將此二擬合法則摒除，如此除可加快分析的時間外，亦可減少錯誤的產生，藉此幫助決策者能正確且有效率的獲得結果。

2. 最小平方法、最小 L_p-norm 法之判別。

於圖 4-5 及圖 4-6 可觀察到，二方法於二維平面上分佈情況分別在某特定區域其擬合效果較佳。因此本研究以分配族所在區域的不同, 做為區分資料的依據, 因資料均散佈於 S_U(unbounded), S_B(bounded) 二區域上所以將二維平面區分為二個區域(圖 4-7)，分別利用合適的分析方法進行分析。

4.3.2 統計檢定及邏輯迴歸判別

1. 區域一中二擬合方法之分佈情況較分散，明顯不同處僅在於樣本點數的多寡，無法透過資料分類手法找出群集的現象。因此，本研究採用統計檢定的方法，求出此區塊中何種擬合方法效果較明顯，統計之假設及檢定結果如下：

假設此區塊之數據中, 至少有 80%自最小 L_p-norm 擬合法中取得，於此 714 組數據中有 154 組非來自於最小 L_p-norm 擬合法，在顯

著水準(α=0.05)之下驗証上述假設是否成立。

H_0:P≧0.8　　H_1:P＜0.8

$$Z^0 = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.784 - 0.8}{0.01496} = -1.049$$

$Z^0 > Z_{(\alpha)}$=-1.645　故 Non-Reject H_0 最小 L_p-norm 法其擬合效果優於最小平方法，亦即資料之 β_1 及 β_2 分佈於 S_U(unbounded)之區塊中，則決策者可採用最小 L_p-norm 法進行資料的分析處理，如此可有較佳的決策品質。

2. 區域二上二擬合法分佈狀況明顯在特定的區域上，其各自的擬合效果較佳，在此利用統計之邏輯迴歸模式，擬找出一線性方程式做為區隔依據，決策者可依線性方程式做為其判斷的準則。邏輯迴歸的應用與分析結果在下章有詳細的介紹。

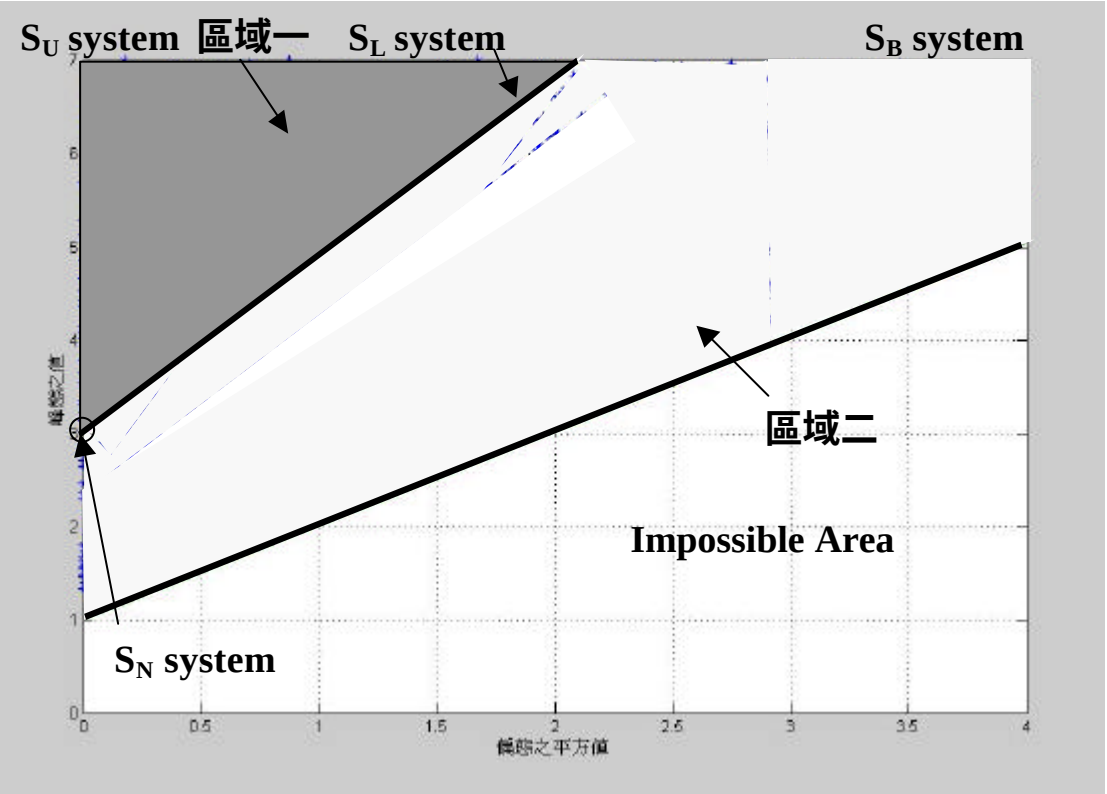


圖 4-7 二維平面區隔圖

第五章　樣本資料分類

本研究資料分析第二部分為利用邏輯迴歸程序進行分析，其資料分類方法屬二分數據的分析，利用最大可能率估計法找出一個迴歸模型的參數估計值、或實驗數據以及類別數據中的底線率(Threshold Response Rate)。本研究為最小平方法、最小 L_p-norm 法兩種情況，不論其定義如何，透過邏輯迴歸分析的目的是為了找出此兩種方法之間的線性關係。藉由求得之線性關係規範出不同的樣本特徵於此兩種方法之適用狀況。

5.1 邏輯迴歸模型

若反應變項可以是一個二分的變項或次序變項則模型的量化單位可以是 logit、normit、log-log 等三種，以下介紹邏輯迴歸模型的種類：

1. 二分反應變項的模型

若反應變項可以是一個二分的(如：1=正向的結果，2=負項的結果)，則任何一觀察體在此變項上得 1 的機率，p=prob(Y=1丨X)，可用對數奇數比的單位來表示，如：

$$\text{logit}(p)=\log[p/(1-P)]=\alpha+\beta' X$$

在此，X 代表一組自變項，α是模型中的截距，β是一組與 X 對應的迴歸係數。這種迴歸模型與一般的線性模型無異，都代表依變項 Y 的均數，即 prob(Y=1)與一組連續變項間的函數對應關係。Nelder 與 Wedderburn[28]將此種函數對應的模型稱作附會函數(Link Function)

2. 次序變項的模型

若變項的數值有大、小或高低之分，如 1=高中畢業，2=大(專)學畢業，3=研究所畢業或以上的學歷..，則用 1,..,k,k+1 的整數來代表這些組別。由於組別數可能不止 2，因此上述的函數表示法改寫成：

$g[Prob(Y \leqq i \mid X)] = \alpha_i + \beta' X \qquad 1 \leqq i \leqq k$

所以，(k+1)**個組只需** k **個截距再加上一組**(k **個**)**與斜率有關的參數即可解釋次序變項上反應分佈的情況。**

5.2 邏輯迴歸假設與檢定

應用四個統計假設做為評估邏輯迴歸模型適合度(model-fit)**及統計推論的依據**[8]**：**

1. **獨立觀測值：觀測值假設為獨立。**
2. **精確的衡量：獨立變數假設皆已精確衡量。**
3. **樣本數很大。**
4. **所有有關變項需包含於分析中，無關之變項不包含於分析中。**

邏輯迴歸模型之虛無假設為依變項之機率(Y=1)**與變項** X **之關係符合線性函數模型，檢定此虛無假設的方法為皮爾森卡方檢定**(Pearson chi-square test)**。而模型整體的有效度以對數可能率來表示，其值等於**[-2 log likelihood]**。不論模型的形式是簡單的**(**只含一個或數個截距**)**或是複雜的**(k **個截距與** k **個斜率參數**)**，這個對數可能率的檢定都是針對模型中所有參數的聯合有效度而設計。因此，每個參數個別對模型的影響力則必須看其他的統計量，如** Wald **氏的** X^2 **檢定。**

5.3 迴歸模型適合度

Logistic **程序根據三個指標來鑑定迴歸模型的優劣：**

(1)-2 log likelihood

這個指標的定義如下：

$$-2\log L = -2\sum_j w_j \log(\hat{p}_j)$$

在此，w_j **是第** j **個觀察體的加權值，**$\hat{p}_j$ **就是所定義的** $Prob(Y_j=i \mid X_j)$ **之最大值。**

(2)**赤池資訊量指標**(Akaike Information Criterion，**又作** AIC)

AIC=-2LogL+2(k+s)

在此，L **的定義如**(1)**所示，**k **代表反應變項之組別數減** 1，s **代表自變項的個數**。

(3)**蕭氏指標**(Schwartz Criterion,**又作** SC)

SC=-2LogL+(k+s)log(N)

在此，L、K、S **的定義如**(1)(2)**所示，**N **則代表樣本數的大小**。

上述的三個指標都是對數可能率的函數。其中，(1)**的統計量可用** $\div^2$ **的抽樣分配來鑑定其值是。其餘兩個統量則是對數可能率的衍生值，分別對自變項數目或樣本數作矯正。**AIC **與** SC **最主要的功能是比較各個模型的優劣，愈是優秀的模型，其所對應的** AIC **與** SC **值都(相對地)愈小**。

5.4 邏輯迴歸分析

綜上所述，本研究資料型態在(β_1, β_2**)二維平面之** S_U(unbounded)**區域，屬二分數據之分類型態，其二分的依變項值為最小平方法、最小** L_p-norm **法兩種結果，因此符合二分反應變項模型。透過資料篩選、軟體程式執行、結果輸出及整理等程序進行資料的分析研究。**

5.4.1 分析程序

資料篩選過程中需摒除劣質數據(outliers)**，一般藉由長方圖、箱形圖的描繪了解資料的分佈狀況，從而摒除可能的劣值數據。本研究在資料的產生過程中，即利用程式產生區域內所要的資料點，規避劣質數據的發生。**

在邏輯迴歸分析中，軟體程式執行之基本流程包括：自變數與依變數的輸入、選擇執行法則(如順向選擇法、反向選擇法等，本研究為順向選擇法)、選擇輸出形式等。本研究採用 SAS **套裝軟體為分析**

工具。

根據程式輸出結果，估計邏輯迴歸之參數值，並檢驗各估計值是否在檢定的範圍內，從而歸納其結果。

5.4.2 資料分析

樣本資料中之 Y 為依變數，若 Y=1 則表事件發生結果為最小 L_p-norm 法；若 Y=0 則表最小 L_p-norm 法事件未發生。兩自變數 β_1、β_2 分別表偏態平方值及峰態值。根據統計程序分析可得結果如下 logit 線性函數結果如下：

logistic regression

The LOGISTIC Procedure

A.Model Information

Data Set	WORK.NEWS
Response Variable	c
Number of Response Levels	2
Number of Observations	3132
Link Function	Logit
Optimization Technique	Fisher's scoring

B.Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	161.0661	2	<.0001
Score	156.2089	2	<.0001
Wald	149.9817	2	<.0001

C.Analysis of Maximum Likelihood Estimates Standard

Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	0.9744	0.1450	45.1878	<.0001
a	1	0.6454	0.0657	96.5579	<.0001
b	1	-0.6181	0.0505	149.9199	<.0001

D.Association of Predicted Probabilities and Observed Responses

Percent Concordant	65.8	Somers' D	0.280
Percent Discordant	33.7	Gamma	0.282
Percent Tied	0.5	Tau-a	0.129
Pairs	2248052	c	0.640

根據報表可判讀，資料刪除後所得數據為 3132 組，而根據報表上標為“ Analysis of Maximum Likelihood Estimates Standard ”的部份。根據這一部份參數的估計值，可得：

$$\text{Logit}(p)=0.9744 + 0.6454*\beta_1 - 0.6181*\beta_2$$

根據報表上標為“ Association of Predicted Probabilities and Observed Responses ”的部份，可判讀此參數之估 65.8%機率相契合。

本研究欲求之線性方程式乃二分之機率值相等的情況下，以 p=0.5 代入，得

$$\beta_2=1.5764+1.0442*\beta_1$$

5.5 邏輯迴歸結果於二維(β_1,β_2)平面之分析

將邏輯迴歸分析所得之線性方程式表示於二維(β_1,β_2)之 S_B 區域上，圖 5-1 為最小 L_p-norm 法圖，圖 5-2 為最小平方法。

自圖 5.1 及圖 5.2 中可判斷，在 S_B 區域內藉由邏輯迴歸求得之方程式 $\beta_2=1.5764+1.0442*\beta_1$ 劃分為兩個區域(區域一、區域二)。圖 5-1 之區塊一內的樣本點數(1862 個樣本點，佔 68.75%)明顯優於圖 5-2 之區塊一內的樣本點數(886 個樣本點，佔 31.25%)，而在圖 5-2 之區塊二內的樣本點數(228 個樣本點，佔 60%)明顯優於圖 5-1 之區塊二內的樣本點數(156 個樣本點，佔 40%)，且在模型檢定程序中對此邏輯迴歸所求之方程式亦有 70%的信賴水準。

綜上所述，在 S_B 區域中，可藉由 $\beta_2=1.5764+1.0442*\beta_1$ 方程式做為判斷的依據。若決策者所面臨資料點之 β_1、β_2 落於區塊一，建議可採最小平方法做為擬合方法；若資料點之 β_1、β_2 落於區塊二則可採最小 L_p-norm 法進行分析，如此可減少決策耗費的時間，亦可減少錯誤的發生。

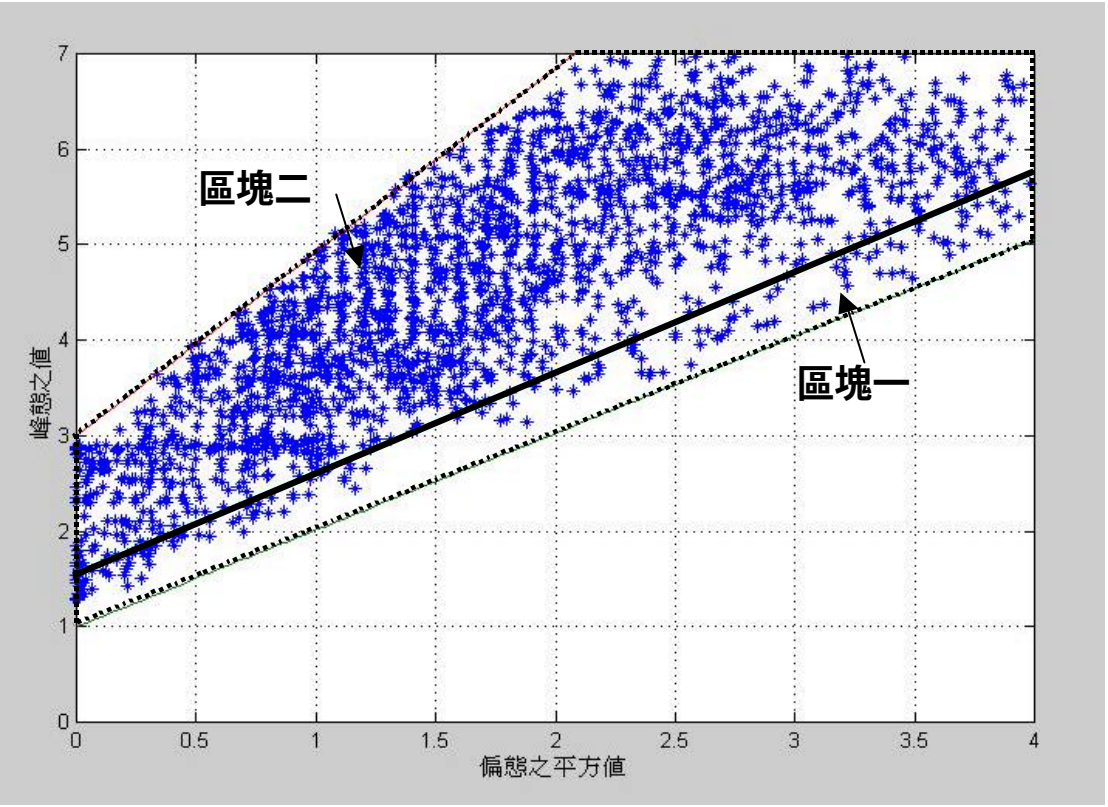


圖 5-1 邏輯迴歸方程式於最小 L_p-norm 法

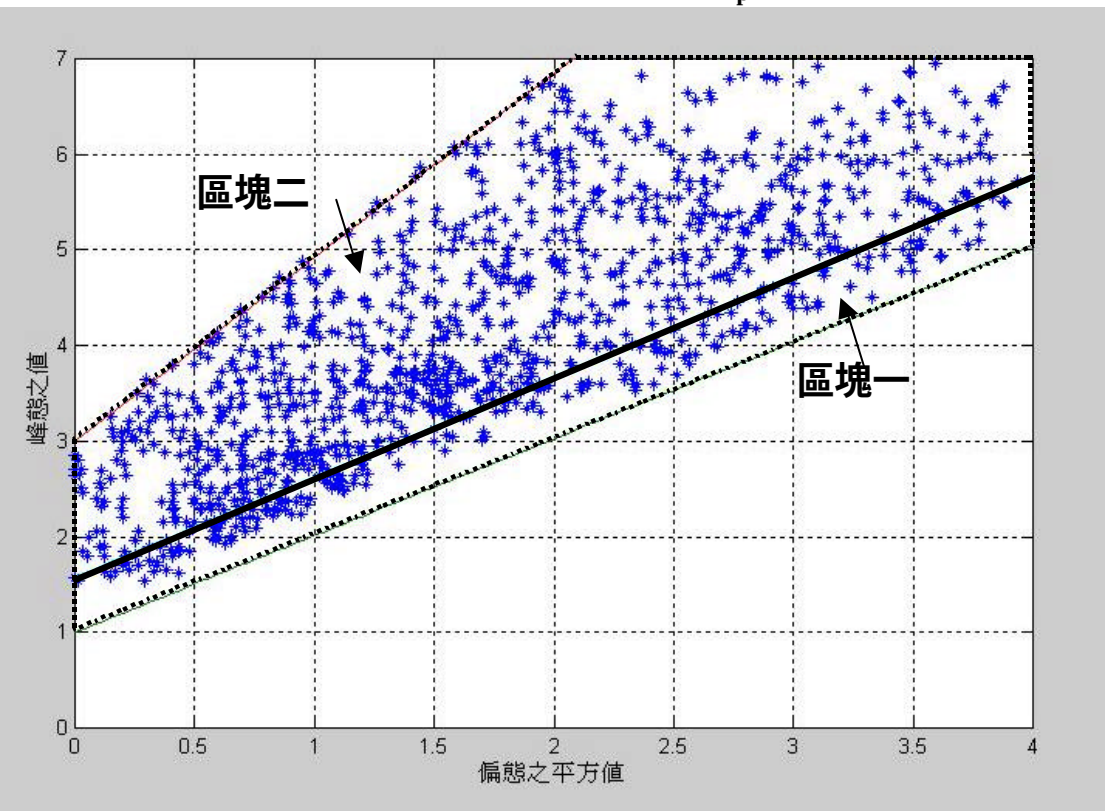


圖 5-2 邏輯迴歸方程式於最小平方法

第六章　結論與建議

6.1 結論

面對多樣且具不同特徵的資料型態，藉由輸入資料分析方法分析與整理資料本身隱含的資訊，以提供決策人員參考。此利用資料，從資料中發掘有用的資訊以輔助決策制訂之用的過程，乃是資料挖掘(data mining)的一個部份。使用者根據本身需求，對所蒐集與整理的資料，加以處理、轉換、發掘至評估的一連串過程，期能找出真實世界運行時隱含其內的運作現象，以輔助解決問題之用。而輸入資料分析乃是此過程中所使用的技巧與工具之一，藉此顯示資料所隱藏的含義。

傳統在進行輸入資料分析時必須在機率分配已知的情況下來擬合樣本資料點，以利進行資料分析的動作。但決策者常會遇到一困境，即在嘗試所有機率分配後仍無法找到一合適的分配。因此，本研究的目的為利用樣本 3 級及 4 級動差的合理範圍所建構之二維(β_1, β_2)平面，判斷強森分配擬合法則在平面上的擬合效果，整理不同法則所適之最佳區域，做為找尋樣本所屬分配時的參考。

本研究藉由國內外相關文獻探討，針對輸入資料分析時所面臨問題點及強森分配擬合法則的特性，歸納整理在進行輸入資料分析時以強森四種擬合法則在二維(β_1, β_2)平面上的擬合效果，藉此改善分析時時間的耗費及避免錯誤的發生。透過遺傳演算法產生研究所欲分析之樣本資料點，且使此資料點之 β_1、β_2 值均勻分佈於二維平面上，利用強森擬合程序對每組樣本各別進行四種擬合法則的擬合，並以 k-s 值進行適合度檢定，根據檢定結果，以統計假設檢定及邏輯迴歸等分析技術整理各別擬合法則適用情況，並判斷分析結果的正確性。

經由上述資料分析的過程，在樣本數不大(<30)，資料屬於右偏且母體分配未知的情況下，四種擬合法則於二維之(β_1, β_2)平面適用情況，整理結果如下：

1. **動差擬合法與百分比擬合法其所屬最佳之擬合效果不顯著，僅佔** 1.425%(57/4000)**，在進行資料分析時，可予以摒除，避免時間的耗費及錯誤的發生。**
2. **最小平方法及最小** L_p-norm **法以圖** 4-7 **為依據，圖中以** S_L **為分界線將二維之(**β_1, β_2**)平面劃分為二個區域，區域一(**S_U**)及區域二(**S_B**)。區域一內之資料分佈較分散且不規則，透過統計檢定得結果為，至少有** 80%**信心研判最小** L_p-norm **法最佳。**
3. **區域二中，在資料分佈上有明顯規則，利用邏輯迴歸程序進行分析所得結果為，以方程式** $\hat{a}_2=1.5764+1.0442*\hat{a}_1$ **區分為二部份，方程式以上為最小** L_p-norm **法擬合效果較佳；方程式以上為最小平方法較佳。**

最後就學術意義而言，本研究具有下列貢獻：

1. **透過文獻資料的蒐集與整理，歸納輸入資料分析所面臨的困難點及可行的解決方法。**
2. **藉由資料挖掘的方式，以強森分配擬合法則輔以遺傳演算法，分析資料在不同特徵下與各四種擬合法間的關係。**
3. **由統計分析與邏輯迴歸的方法，整理出樣本在不同**β_1**、**β_2**下其適合的擬合方法，提供資料分析或相關決策人員一個參考的依據。**

6.2 建議

本研究主要針對輸入資料分析提出在二維(β_1, β_2)平面上，不同擬合方法的適用情況。然而強森分配系統在一般的統計及品質管制上亦扮演極具重要的角色，因此對有意從事相關研究者，本研究提出以下幾點建議作為未來研究方向的參考。

1. 強森分配轉換系統具有將非常態性質的資料(non-normal data)轉為常態性質資料的性質，此性質亦可與(β_1, β_2)之二維平面相結合，藉由資料點的產生、分析與轉換，可進一步判斷在不同區域上常態轉換效果的優劣。
2. 在統計及製程管制上對資料點的整理與分析亦需將非常態性質的資料轉化為常態，未來可針對此部份配合(β_1, β_2)之二維平面做進一步的研究，並可與傳統之統計及製程管制的手法相互比較，歸納出其差異性及適合的樣本特徵，作為統計及製程管制時的參考。
3. 本研究可應用於未知分配報童模型存貨問題上，藉由(β_1, β_2)之二維平面擬合結果所提供的資訊，探討其在資訊價值(EVAI)方面的應用。

參考文獻

1. Avriel, M. (1976), *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, New Jersey.
2. Banks, J. and J.S. Carson,II (1984), *Discrete-Event Simulation*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
3. Bawman, K.O. and L.R. Shenton (1983), "Johnson's systems of distributions," *Encyclopedia of Statistical Sciences* 4, 303-314.
4. Bawman, K.O. and L.R. Shenton (1988), "Solution to Johnson's SB and SU," *Communications in Statistics, Part B- Simulation and Computation*, Vol.17(2), 343-348.
5. Bawman, K.O. and L.R. Shenton (1989), "SB and SU Distribution Fitted by Percentles：A General Criterion," *Communications in Statistics, Part B- Simulation and Computation*, Vol.18(1), 1-13.
6. Beach, L. R. and R.G. Swenson (1966), "Intuitive estimation of means," *Psychonological Science* 5, 161-162.
7. Bratley, P.B, B. L. Fox, and L. E. Schrage (1987), *A Guide to Simulation.* 2nd. Ed. Springer-Verlag, New York.
8. Chen, S.S. and C. Darryl (1998), "Principle and Method of Using Logistic Regression Analysis," *Journal of Dahan Junior College of Engineering and Business*, Vol.12.
9. Chou, Y.M., A.M. Polansky, and R.L. Mason (1998), "Transforming non-normal data to normality in statistical process control." *Journal of Quality Technology* 31, 133-141.
10. Chou, Y.M., S. Turner, S. Henson, D. Meyer, and K.S. Chen (1994), "On using percentiles to fit data by a Johnson distribution," *Communications on Statistics – Simulation and Computation* 23, 341-354.

11. Cox, D.R.(1970), *The Analysis of Binary Data*, London: Chapman and Hall.
12. DeBrota, C., S. D. Roberts, R. S. Dittus, and J. R. Wilson (1988), "Visual interactive fitting of probability distributions," *Simulation* 52, 199-205.
13. Doubilet, P., C.B. Begg, M. C. Weinstein, P. Braun, and B.J. McNeil (1985), "Probabilistic sensitivity analysis using Monte Carlo simulation, a practical approach," *Medical Decision Making* 5,157-177.
14. Hahn, G. J. and S. S. Shapiro (1967), *Statistic Models in Engineering*, New York：John Wiley and Sons, Inc.
15. Hanselman, D. and B. Littlefield (1995), *Mastering MATLAB: A Comprehensive Tutoring and Reference*. Prentice-Hall, New York, NY.
16. Hill, I.D., R. Hill, and R.L. Holder (1976), "Fitting Johnson curves by moments," *Applied Statistics*, 25, 180-189.
17. Hoover, S.V. and R.F. Perry (1989), *Simulation: a Problem-Solving Approach*, Addison-Welsle Publishing, New York, NY.
18. Hubele, N.F. and F.P. Lawrence (1995), "C_{pk} Index Estimation Using Data Transformation," *The 17th International Conference on Computers and Industrial Engineering*,Vol.29,No.1-4,55-58.
19. Johnson, M. E. and V.W. Lowe (1979), "Bounds on the Sample Skewness and Kurtosis," *Technometrics* , Vol.21(5), 377-378.
20. Johnson, N.L. (1949), "Systems of frequency curves generated by methods of translation," *Biometrika* 36, 149-176.
21. Johnson, N.L., S. Kotz, and N. Balakrishnan (1994), *Continuous Univariate Distributions*, Vol. 1. John Wiley & Sons, New York, NY.
22. Kahneman, D., P. Slovic, and A. Tversky (1982), *Judgement under*

uncertainty: Heuristics and biases,Cambridge University Press.

23. Kendall, M.G. and A. Stuart (1979), *The Advanced Theory of Statistics*, Vol. 2, 4th ed. Macmillian, New York.
24. Law, A.M. and W.D. Kelton (1990), *Simulation Modeling and Analysis* ,McGraw-Hill Book Company, New York.
25. LINDO, *LINGO: Optimization Modeling Language*,LINDO System, Inc., Chicago, IL.
26. Mage, D.T. (1980), " An explicit solution for S_B parameters using four percentile points, " *Technometrics* 22, 247-251.
27. Marquardt, D.W. (1963), " An algorithm for least-square estimation of nonlinear parameters, " *Journal of the SIAM* 11, 431-441.
28. Nelder, J.A. and R.W.M. Wedderburn (1972), " Generalized linear models, " *Journal of the Royal Statistical Society*, Series A,135 370-384.
29. Polansky, A.M., Y.M. Chou, and R.L. Mason (1999), " An algorithm for fitting Johnson transformations to non-normal data, " *Journal of Quality Technology* 31, 345-350.
30. Roberts, S.D. (1983), *Simulation Modeling and Analysis with INSIGHT*, Regenstrief Institute for Health Care, Indianapolis, Indiana.
31. Schmeiser, B.W. and S.J. Deutsch (1977), " A versatile family of probability distributions suitable for simulation, " *AIIE Transactions* 9, 176-182.
32. Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.
33. Slifker, J.F. and S.S. Shapiro (1980), " The Johnson system: selection and parameter estimation, " *Technometrics* 22, 239-246.
34. Spedding, T.A. and P.L. Rawlings (1994), " Non-normality in

Statistical Process Control Measurements,” *International Journal of Quality & Reliability Management*, Vol.11(6), 27-37.

35. Spencer, J. (1963), “A further study of estimating averages,” *Ergonomics* 6,255-265.
36. Stuart, A. and J. K. Ord (1987), *Kendall’s Advanced Theory of Statistics*, Vol. 1, 5th ed. Oxford University Press, New York, NY.
37. Swain, J.J., S. Venkatraman, and J.R. Wilson (1988), “Least squares estimation of distribution functions in Johnson’s translation system,” *Journal of Statistical Computation and Simulation* 29, 271-297.
38. Venkatraman, S. (1988), *Modeling multivariate populations with translation systems*,Unpublished Ph.D. dissertation, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.
39. Venkatraman, S. and J.R. Wilson (1987), Modeling univariate populations with Johnson’s translation system-Description of the FITTR1 software. Research Memorandum, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.
40. Wilson, J.R., D.K. Vaughan, E. Naylor, and R.G. Voss (1982), “Analysis of Space Shuttle ground operations,” *Simulation* 38, 187-203.
41. Wu, H.H.(2000), “Using the Johnson System in Clements-based Process Capability Indices for Non-Normal Processes,” *The 5th Annual International Conference on Industrial Engineering.*Hsinchu, Taiwan.
42. **吳明隆，**SPSS **統計應用實務，第二版，松崗圖書，台北，**2000 **年** 8 **月。**
43. **李隆安、葉昭瑛，「論邏輯迴歸模式的估計」，國立政治大學學報，第六十七期，頁** 493-511**，民國** 82 **年** 10 **月。**

44. 林保佑，「運用強森分配於系統模擬及輸入資料分析之探討」，中國工業工程學會，2001 年 12 月。
45. 林寶香，「智慧型代理人於電子商務之整合與應用」，東海大學工業工程研究所碩士論文，台中，民國 89 年。
46. 柳克婷，「類別變數資料分析方法之研究--ODDS 比與 LOGISTIC 迴歸模式」，中國工商學報，第十七期，頁 295-308，民國 84 年。
47. 胡坤德、姚銘忠，「應用強森分配在系統模擬輸入分析時演算法之改善」，國科會計劃提案，民國 90 年。
48. 唐明月，應用統計學，第三版，中興管理顧問，台北，民國 71 年。
49. 張子傑、徐銘傑，應用統計學講義，第六版，鼎茂圖書，台北，民國 88 年 2 月。
50. 許中川，「圖形化線上資料分析」，工業工程學刊，　十　八　　，四期頁 37-35，民國 90 年 7 月。
51. 彭昭英，SAS 與統計分析，第五版，儒林圖書，台北，民國 82 年。
52. 劉睦雄，系統模擬，中央圖書出版社，台北，民國 81 年。
53. 蔡瑞隆，「供應鏈中存貨系統資訊價值之探討」，東海大學工業工程研究所碩士論文，台中，民國 90 年。

簡歷

姓　　名：林保佑

出 生 地：台灣省彰化縣

出生日期：民國六十六年十一月二十三日

學　　歷：民國八十九年東海大學工業工程與經營資訊學系畢業

聯 絡 處：彰化縣埔心鄉東門村員鹿路二段 335 **號**