

圖解醫學統計學

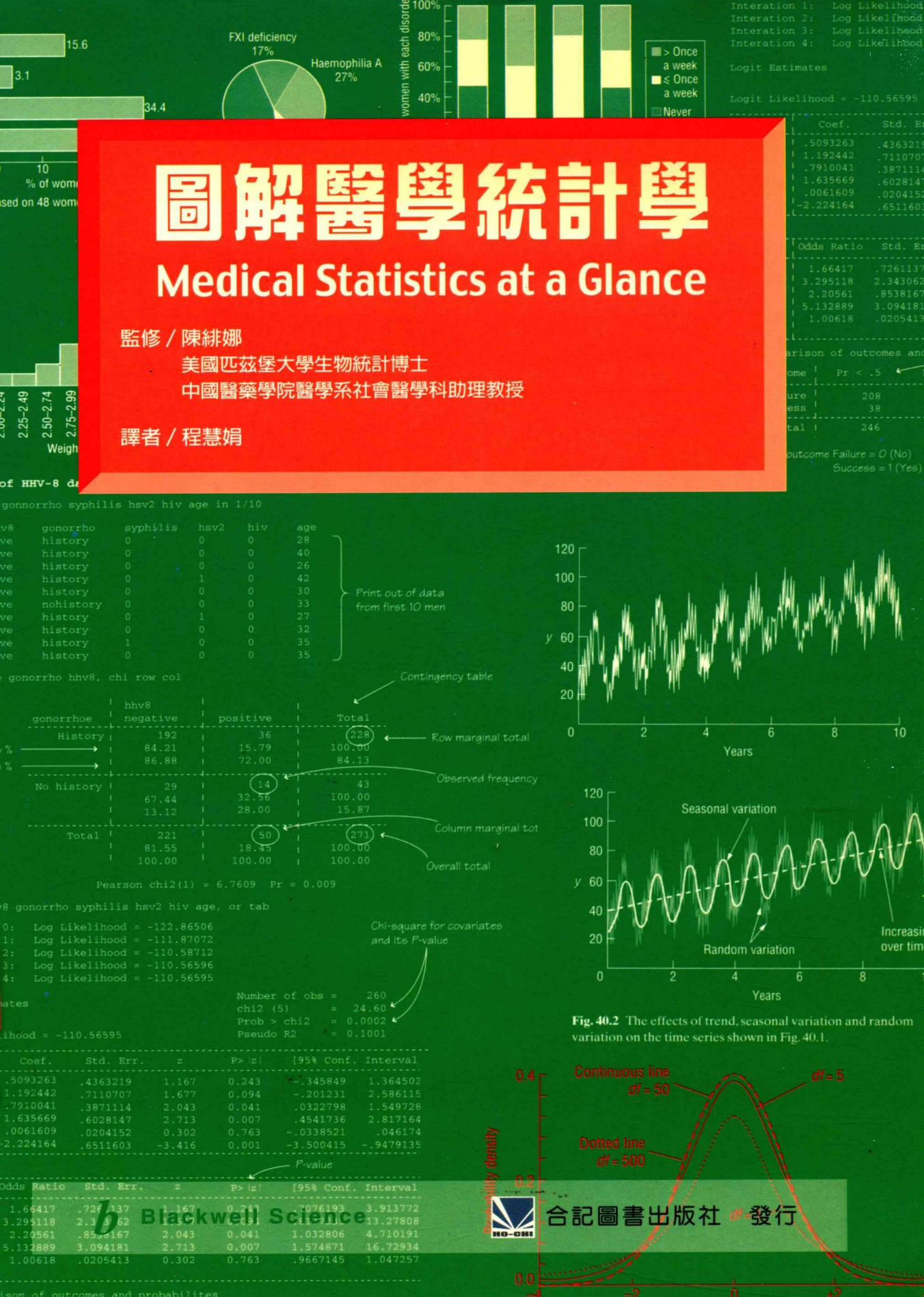
Medical Statistics at a Glance

監修 / 陳緋娜

美國匹茲堡大學生物統計博士

中國醫藥學院醫學系社會醫學科助理教授

譯者 / 程慧娟





圖解醫學統計學

Medical Statistics at a Glance

監修 / 陳緋娜

美國匹茲堡大學生物統計博士

中國醫藥學院醫學系社會醫學科助理教授

譯者 / 程慧娟

analysis of HIV-8 data

list hiv8 gonorrhoea syphilis hsv2 hiv age in 1/10

hiv8	gonorrhoea	syphilis	hsv2	hiv	age
negative	history	0	0	0	28
negative	history	0	0	0	40
negative	history	0	0	0	26
negative	history	0	1	0	42
positive	history	0	0	0	30
negative	nohistory	0	0	0	33
negative	history	0	1	0	27
positive	history	0	0	0	32
negative	history	1	0	0	35
positive	history	0	0	0	35

Print out of data from first 10 men

tabulate gonorrhoea hiv8, chi row col

	gonorrhoea	hiv8 negative	positive	Total
History	192	36	228	
Row %	84.21	15.79	100.00	
Column %	86.88	72.00	84.13	
No history	29	14	43	
	67.44	32.56	100.00	
	13.12	28.00	15.87	
Total	221	50	271	
	81.55	18.45	100.00	
	100.00	100.00	100.00	

Pearson chi2(1) = 6.7609 Pr = 0.009

logit hiv8 gonorrhoea syphilis hsv2 hiv age, or tab

Iteration 0: Log Likelihood = -122.86506
Iteration 1: Log Likelihood = -111.87072
Iteration 2: Log Likelihood = -110.58712
Iteration 3: Log Likelihood = -110.56596
Iteration 4: Log Likelihood = -110.56595

Estimates

Log Likelihood = -110.56595

hiv8	Coef.	Std. Err.	z	P
gonorrhoea	1.5093263	.4363219	1.167	0.244
syphilis	1.192442	.7110707	1.677	0.095
hsv2	.7910041	.3871114	2.043	0.041
hiv	1.635669	.6028147	2.713	0.007
age	.0061609	.0204152	0.302	0.763
Constant	-2.224164	.6511603	-3.418	0.001

P-value

hiv8	Odds Ratio	Std. Err.	z	P	95% Conf. Interval
gonorrhoea	1.84117	.7110707	1.167	0.244	0.748493 4.5913772
syphilis	3.295118	1.070707	1.677	0.095	1.127608 11.10130
hsv2	2.20561	.581147	2.043	0.041	1.032806 4.710130
hiv	5.132889	.6028147	2.713	0.007	1.571371 16.72314
age	1.00618	.0205413	0.302	0.763	.9667145 1.047257

Comparison of outcomes and probabilities

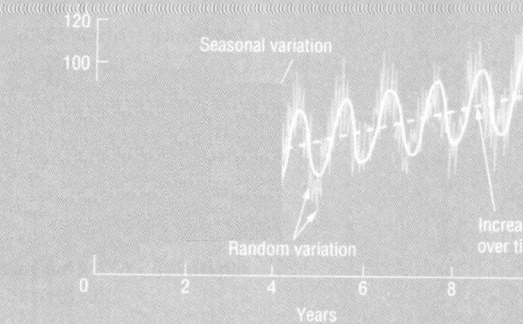
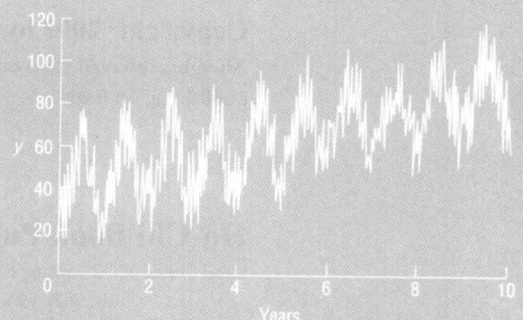


Fig. 40.2 The effects of trend, seasonal variation and random variation on the time series shown in Fig. 40.1



合記圖書出版社 發行

圖解醫學統計學 / Aviva Petrie, Caroline Sabin

原著：程慧娟編譯 -- 初版 -- 臺北市：合記
，2003【民 92】

面： 公分

譯自：Medical statistics at a glance

ISBN 957-666-982-0 (平裝)

1. 醫學 - 統計方法

410.28

92009776

書 名 圖解醫學統計學

編 譯 程慧娟

執行編輯 黎琬琦

發行人 吳富章

發行所 合記圖書出版社

登記證 局版臺業字第 0698 號

社 址 台北市內湖區(114)安康路 322-2 號

電 話 (02)27940168

傳 真 (02)27924702

網 址 <http://www.hochi.com.tw/>

總經銷 合記書局

北 醫 店 臺北市信義區(110)吳興街 249 號

電 話 (02)27239404

臺 大 店 臺北市中正區(100)羅斯福路四段 12 巷 7 號

電 話 (02)23651544 (02)23671444

榮 總 店 臺北市北投區(112)石牌路二段 120 號

電 話 (02)28265375

臺 中 店 臺中市北區(404)育德路 24 號

電 話 (04)22030795 (04)22032317

高 雄 店 高雄市三民區(807)北平一街 1 號

電 話 (07)3226177

花 蓮 店 花蓮市(970)中山路 632 號

電 話 (03)8463459

郵政劃撥 帳號 19197512 戶名 合記書局有限公司

西元 2005 年 2 月 10 日 初版二刷



目錄

1 資料的形式	8	25 類別資料：超過兩個類別以上 ...	64
2 輸入資料	10	26 相關性	67
3 檢查錯誤與極端值	12	27 線性迴歸理論	70
4 用圖表展示資料	14	28 執行線性迴歸分析	72
5 描述資料 1：一般平均值	16	29 多元線性迴歸	75
6 描述資料 2：分散程度	18	30 多項式及羅吉斯回歸分析	78
7 理論分佈 1：常態分布	20	31 統計模型	80
8 理論分佈 2：其他的分布	22	32 檢視預設	82
9 轉換	24	33 計算樣本數	84
10 取樣與取樣分布	26	34 發表成果	87
11 信賴區間	28	35 診斷方法	90
12 研究設計 1	30	36 評估一致性	93
13 研究設計 2	32	37 實證醫學	96
14 臨床實驗	34	38 全面性回顧和後分析	98
15 世代研究	37	39 重複性測量的方法	101
16 病例對照研究	40	40 時間數列	104
17 假設檢定	42	41 存活分析	106
18 檢定假說時發生的錯誤	44	42 貝氏方法	109
19 數值資料：單一組別	46		
20 數值資料：兩相關的組別	49	附錄	
21 數值資料：兩個不相關的組別 ...	52	A 統計表	112
22 數值資料：兩組以上的資料	55	B 計算樣本數的阿特慢表	119
23 類別資料：單一比例	58	C 一般電腦報表輸出格式	120
24 類別資料：兩個比例值	61	D 術語的字彙	127



圖解醫學統計學

Medical Statistics at a Glance

監修 / 陳緋娜

美國匹茲堡大學生物統計博士

中國醫藥學院醫學系社會醫學科助理教授

譯者 / 程慧娟

analysis of HIV-8 data
list hhv8 gonorrhoea syphilis hsv2 hiv age in 1/10

hhv8	gonorrhoea	syphilis	hsv2	hiv	age
negative	history	0	0	0	28
negative	history	0	0	0	40
negative	history	0	0	0	26
negative	history	0	1	0	42
positive	history	0	0	0	30
negative	nohistory	0	0	0	33
negative	history	0	1	0	27
positive	history	0	0	0	32
negative	history	1	0	0	35
positive	history	0	0	0	35

Print out of data
from first 10 men

tabulate gonorrhoea hhv8, chi row col

	hhv8 negative	positive	Total
gonorrhoea History	192 84.21	36 15.79	228 100.00
Column %	86.88	72.00	84.13
No history	29 67.44	14 32.56	43 100.00
Row %	13.12	28.00	15.87
Total	221 81.55	50 18.45	271 100.00
Column %	100.00	100.00	100.00

Pearson chi2(1) = 6.7609 Pr = 0.009

Contingency table

Row marginal total

Observed frequency

Column marginal tot

Overall total

logit hhv8 gonorrhoea syphilis hsv2 hiv age, or tab

Iteration 0: Log Likelihood = -122.86506
Iteration 1: Log Likelihood = -111.87072
Iteration 2: Log Likelihood = -110.58712
Iteration 3: Log Likelihood = -110.56596
Iteration 4: Log Likelihood = -110.56595

Chi-square for covariates
and its P-value

Number of obs = 260
chi2 (5) = 24.60
Prob > chi2 = 0.0002
Pseudo R2 = 0.1001

Estimates

Log Likelihood = -110.56595

hhv8	Coef.	Std. Err.	z	P
gonorrhoea	1.5093263	.4363219	1.167	0
syphilis	1.192442	.7110707	1.677	0
hsv2	.7910041	.3871114	2.043	0
hiv	1.635669	.6028147	2.713	0
age	.0061609	.0204152	0.302	0.763
Constant	-2.224164	.6511603	-3.418	0.001

P-value

hhv8	Odds Ratio	Std. Err.	z	P	95% Conf. Interval
gonorrhoea	1.84417	.7110707	1.677	0	1.192442 3.192442
syphilis	3.295118	.7110707	1.677	0	1.192442 9.13772
hsv2	2.20561	.3871114	2.043	0	1.032806 4.710793
hiv	5.132889	.6028147	2.713	0	1.573371 16.72914
age	1.00618	.0205413	0.302	0.763	.9667145 1.047257

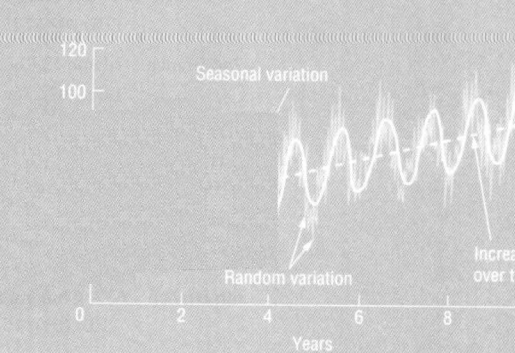
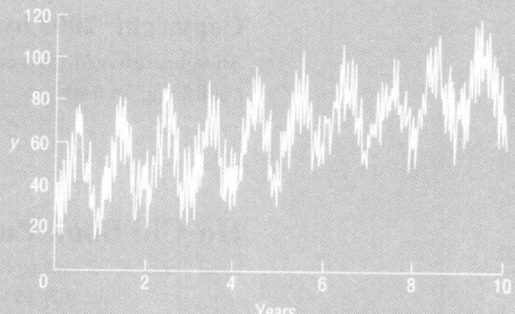
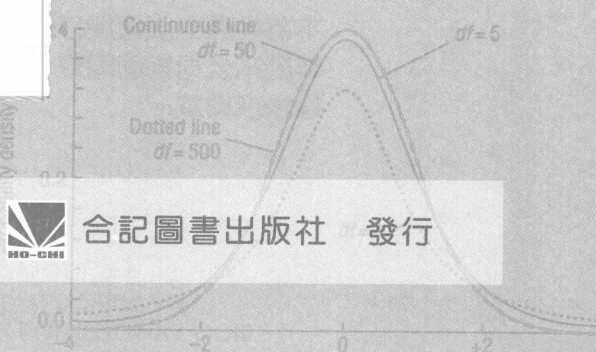


Fig. 40.2 The effects of trend, seasonal variation and random variation on the time series shown in Fig. 40.1



Blackwell Science



合記圖書出版社 發行

MEDICAL STATISTICS AT A GLANCE

By AVIVA PETRIE & CAROLINE SABIN

ISBN 0-632-05075-6

Copyright © By Blackwell Science Ltd, Oxford

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

Copyright©2003 by Ho-Chi Book Publishing Co.

All rights reserved. This edition is published by arrangement with Blackwell publishing, Oxford.

Ho-Chi Book Publishing Co.

Head Office 322-2 Ankang Road, NeiHu Dist., Taipei Taiwan 114

TEL: (02)2794-0168 FAX:(02)2792-4702

1st Branch 249 Wu-Shing Street, Taipei 110, Taiwan, ROC

TEL: (02)2723-9404 FAX:(02)2723-0997

2nd Branch 7 Lane 12, Roosevelt Rd, Sec 4, Taipei 100, Taiwan.

TEL: (02)2365-1544 FAX:(02)2367-1266

3rd Branch 120 Shih-Pai Road, Sec 2, Taipei 112, Taiwan.

TEL: (02)2826-5375 FAX:(02)2823-9604

4th Branch 24 Yu-Der Road, Taichung(404), Taiwan

TEL: (04)2203-0795 FAX: (04)2202-5093

5th Branch 1 Pei-Peng 1st Street, Kaoshiung 800, Taiwan.

TEL: (07)322-6177 FAX:(07)323-5118

6th Branch 632, ChungShan Road, Hualien 970, Taiwan.

TEL: (03)846-3459

本書經原出版者授權翻譯、出版、發行；版權所有。
非經本公司書面同意，請勿以任何形式作翻印、攝影、
拷錄或轉載。



醫學統計學概要一書是針對醫學生，醫學相關研究者，生醫類別的研究生，以及製藥從業人員所寫。這些人在從業過程中可能或多或少會握有一些定量的資料(可能是從別人那裡或是從自己的資料)必須進行嚴謹的評估與詮釋，或是他們得準備統計學的考試。應付這些需要，必須能了解統計概念與方法學。我們不只想燃起各位讀者對於統計學的興趣，同時也考慮到實際的情形。我們的目的是提供學生與研究者以及那些臨床醫生在閱讀醫學文獻時碰到統計的一些概念時，能夠有一本容易閱讀，詳實，高度相關的書，且能夠作實際的運用。

醫學統計學概要可以作為統計學授課的輔助讀物且可以作為參考資料指引。另外，讀者可以試試網站 www.medstatsaag.com 上的習題來評估她／她們的自學的進步程度。這個網站也有一套完整的參考資料(其中有些直接可以連上 MEDLINE)來補充文中所引用的參考資料並提供例子相關的一些有用的背景資料。對於那些想要更加充實了解醫學統計某些部份，我們在此推薦下列書：

- Altman, D.G. (1991) *Practical Statistics for Medical Research*. Chapman and Hall, London
- Armitage, P., Berry, G. (1994) *Statistical methods in Medical Research*, 3rd edn. Blackwell Scientific Publications, Oxford
- Pocock, S.J. (1983) *Clinical Trials: A practical Approach*. Wiley, Chichester.

與其它的概要系列書籍(AT A GLANCE系列)類似，我們試著引導讀者來研讀一系列的主題，每個主題約2到3頁，每個主題討論醫學統計學的不同部份。從我們的教學經驗知道且也考慮了讀者在研讀醫學統計學時會碰到的問題。也就因為這樣此書有關理論的內容只容納能幫助了解步驟的部份，盡量不要容納太多超過實際執行層面的內容。

醫學統計學是範圍廣泛的一門學科，涵蓋很多

的主題。我們提供一個基本的介紹來了解醫學統計學的背後的概念並指引大家了解最常使用的統計學步驟。流行病學與醫學統計學有密切的關係。因此流行病學中主要的議題只要與研究設計與詮釋相關者我們才進行討論。另外還包含一些比較偶爾才會用到的一些主題，且對於醫學研究的許多區域很重要的部份，比如說實證醫學，全面性回顧和整合分析，時間數列，存活分析與貝氏方法。我們解釋了這些主題背後的原則，這樣讀者能夠了解並詮釋文獻中所呈現的研究成果。更詳細的討論可以在網站上所列的參考資料找到。

整個內文引用了很多的麻醉資料來幫助讀者將各個步驟連接在一起。術語字彙的部份(附錄D)提供常用術語的一些基本的解釋。附錄A有一些基本的統計圖表。如果不使用電腦計算，由Routledge出版的Neave, H.R. (1981)所編纂的Elementary Statistical Tables與Ciba-Geigy出版的第八版第二冊的Geigy Scientific Tables提供了讀者更完全的結果。

我們知道對於非統計背景的人最大的困難之一是選擇適當的統計技術。因此我們製作了兩個流程圖可以幫助我們來決定什麼情形之下使用什麼方法並很容易在書中找到使用那種方法。在封面的內側有明顯的圖示。

每一個主題都有一個例子伴隨著來解釋一種統計技術。例子所使用的資料是從我們或是同僚有參與的研究而來。在某些例子，我們的資料是來自已經出版的論文。可能的話在不同的主題我們盡量選用同樣的一套資料來反映資料分析的實際情形，也就是通常不會只使用同一種技術或是方法來分析。雖然我們認為公式還有方法的邏輯應該確實地加以解釋來幫助了解，複雜的計算細節還是應該盡量避免，因為大部份的讀者都會使用電腦或是只有簡單的部份才會自行計算，所以我們不在這裡執行任何複雜的演算。

我們認為讀者必須能夠解讀電腦套裝軟體的輸出結果。因此只要可能我們會把電腦的輸出結果列出。有人在解讀某些例子的結果可能會發生困難，所以我們在 APPENDIX C 附上資料分析後完整的電腦輸出結果。有很多的統計軟體可以提供使用。爲了讓讀者了解各個軟體的輸出結果有所不同，我們不只用一特定的電腦套裝軟體，而使用了三種多數人所熟知的電腦軟體：SAA, SPSS，與 STATA。

我們要謝謝那些提供例子資料的人。特別感謝 RICHARD MORRIS, FIONA LAMPE 以及 SHAK HAJAT，他們讀了整本書，以及 ABUL BASAR 他讀了一大部份，他們都提供了很多寶貴的意見。當然還是我們對內文與例子負絕對的文責。

我們還要謝謝那些一起工作的夥伴：MIKE, GERALD, NINA, ANDREW, KAREN，與 DIANE。他們的耐心與了解，尤其是在完成前的幾個月讓我們能夠專心地完成這件工作。





譯者序

統計學常常是一門讓人望而生畏的基礎學科。醫學統計學則比較多應用的層面，正確地應用統計方法比去鑽研理論要實際的多。希望翻譯本書能夠成功地幫助讀者，從實例與解說之中一窺醫學統計學的堂奧。

譯者謹在此感謝諸位好友與同僚的協助：佳音，慧貞，與令治，在譯者受到催稿的壓力時不吝嗇地提供各式支持，吃吃喝喝工作之餘的海濱散步

與聊天將會是人生美好的回憶之一；另外要感謝大經表哥大方地提供重要的傳道解惑與技術協助；與津如的無數討論，雖然無關乎統計學與國計民生，開闊了譯者的視野也是工作之餘重要的精神支撐養份。一切的一切，銘記在心。

程慧娟

2003年5月於花蓮



目錄

1 資料的形式	8	25 類別資料：超過兩個類別以上 ...	64
2 輸入資料	10	26 相關性	67
3 檢查錯誤與極端值	12	27 線性迴歸理論	70
4 用圖表展示資料	14	28 執行線性迴歸分析	72
5 描述資料 1：一般平均值	16	29 多元線性迴歸	75
6 描述資料 2：分散程度	18	30 多項式及羅吉斯回歸分析	78
7 理論分佈 1：常態分佈	20	31 統計模型	80
8 理論分佈 2：其他的分佈	22	32 檢視預設	82
9 轉換	24	33 計算樣本數	84
10 取樣與取樣分佈	26	34 發表成果	87
11 信賴區間	28	35 診斷方法	90
12 研究設計 1	30	36 評估一致性	93
13 研究設計 2	32	37 實證醫學	96
14 臨床實驗	34	38 全面性回顧和後分析	98
15 世代研究	37	39 重複性測量的方法	101
16 病例對照研究	40	40 時間數列	104
17 假設檢定	42	41 存活分析	106
18 檢定假說時發生的錯誤	44	42 貝氏方法	109
19 數值資料：單一組別	46		
20 數值資料：兩相關的組別	49	附錄	
21 數值資料：兩個不相關的組別 ...	52	A 統計表	112
22 數值資料：兩組以上的資料	55	B 計算樣本數的阿特慢表	119
23 類別資料：單一比例	58	C 一般電腦報表輸出格式	120
24 類別資料：兩個比例值	61	D 術語的字彙	127

資料與統計 (Data and statistics)

大部份研究目的在於收集資料 (data) 以獲得研究相關資訊。所謂的資料是我們針對一或多個變數觀察 (observations) 所得到的觀察紀錄值；所謂變數 (variable) 指的是相同性質資料的項目名稱，例如，在搜集病患的臨床基本資料與人口學資料時，我們可以找性別，年齡，身高的資料。

通常我們從樣本 (sample) 來蒐集資料，前提是樣本要能代表研究群體 (population)，才能從樣本得到有意義的資訊。統計學 (statistics) 涵蓋了如何收集資料，概括資料，分析資料並從之獲得結論。統計方法能幫我們達成目的。

資料有各種不同形式。資料中的變數種類必須先弄清楚，才能決定使用何種統計方法。變數資料有兩類：類別變數 (categorical) 或是數值變數 (numerical) 圖 1.1。

項目 (定性) 資料

[Categorical (qualitative) data]

當個體只能被歸在諸多類別中的一類時稱之。

- 類別資料 (Nominal data) — 各個類別無法排出順序，只是命名個別不同，像血型 (A, B, AB, 及 O 型)，婚姻別 (已婚 / 鰥寡 / 單身) 屬於這一

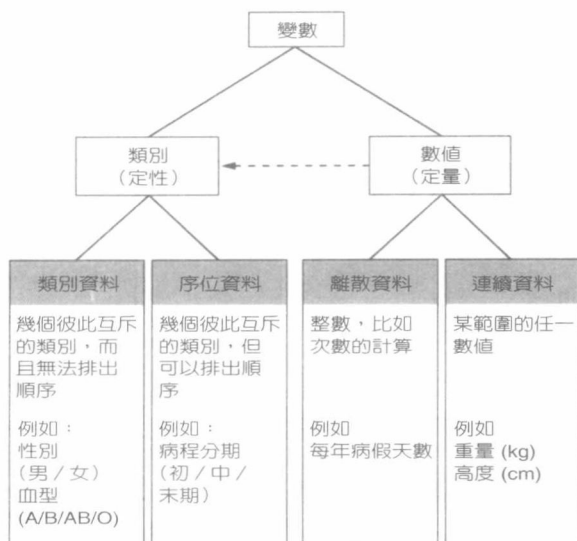


圖 1.1 不同類型的變數

類。‘已婚’與‘單身’沒有孰優孰劣之分！

- 序位資料 (Ordinal data) — 各個項目可以排出順序，例如疾病分期系統 (disease staging systems，重度，中度，輕微，零期之分 (advanced，moderate，mild，none) 及疼痛程度 (有‘重度’，‘中度’，‘輕微’，和‘無’的區分)。

當項目資料只能分為兩個類別時稱之為二項式資料 (binary or dichotomous)，例如“是 / 否”，“死 / 活”或“有病 / 無病”屬之。

數值 (定量) 資料

[Numerical (quantitative) data]

當變數為數值形式時我們可將數值資料再細分為兩類。

- 離散資料 (Discrete data) — 當資料呈整數形式時稱之，例如事件發生數，每年就醫次數，過去五年發病次數屬之。
- 連續資料 (Continuous data) — 如果資料可以是任何數值時稱之含小數點。例如身高，體重 (除非測量時受到限制則不算在內)。

區分資料形式

(Distinguishing between data types)

我們視資料的不同選用不同的統計方法。要分辨 (定性) 資料或數值 (定量) 資料一般沒有問題，但某些情形則可能不易決定。例如某變數有很多排好順序的類別 (例如有 7 個類別的疼痛指數) 可能不容易和不連續數值資料作區別；這不見得對分析結果影響很大，但有時要分別不連續性與連續性數值資料可能更難。像年齡，常常我們當作不連續性資料來處理，可是它其實是連續資料：一般我們用“最近一次生日時的歲數” (age at last birthday) 來表示所謂的‘年齡’而不用“確實年齡” (age)，一個女人自稱 30 歲其實可能是剛過 30 歲，也可能已經將近 31 歲。

在記錄數值資料時不要一開始將它變成項目資料來記錄，這樣往往會失去重要資訊 (例如只記錄病人所屬年齡別，而不記錄確實年齡)；只要我們將資料收集好，將來要把數值資料轉換成項目資料並不麻煩。

衍生資料 (Derived data)

在醫藥文獻可以看到各式資料，如：

- 百分比 (Percentages)－比如病人接受治療後病情改善程度，像病人的肺功能（病人第一秒的強力呼氣量，FEV1）在接受某藥治療後增加達24%。我們感興趣、想知道的是改善程度的百分比，而不是指增加的絕對值。
- 比例值 (Ratios) 或商數 (Quotients)－偶而會碰到兩變數以比例值 (ratios) 的形式呈現。例如身體質量係數 (body mass index, BMI)，計算方法乃用重量除以高度的平方，用來評估體重過輕或過重。
- 比率 (Rates)－疾病發生的頻率，即疾病發生的次數除以時間間隔，常用於流行病學研究（見第12講）
- 評分 (Scores)－對於不能測量的變數，有時我們憑主觀打分數。例如對於生活品質進行一系列評

估後可以得到一個生活品質總得分。

以上的變數都是連續變數，所以大多數的分析方法都可以用。當一個變數是用多個數值來算出時（例如百分比是由分子除以分母），一定要將各數值紀錄下來。例如說，經治療後狀況改善了10%，視治療前的測量值是多少會有不同的臨床意義。

普查資料 (Censored data)

在以下的情形，我們會碰到普查 (censored) 資料。

- 某機器只偵測高於某特定數值的實驗數據，當數值低於某特定量時不會被偵測出來；以測量病毒含量為例，即使取樣標本含有病毒，但當含量太低，低於偵測極限時，沒有辦法紀錄任何觀察值。
- 另外會碰到在試驗追蹤期病人中途退出臨床試驗的情形。第41講將對此作較詳細的討論。

進行研究需要將資料輸入電腦。電腦對於收集資料及分析的速度及正確性有卓越的貢獻，使得檢查錯誤，用圖形來概括資料以及創造新變數都變得很方便。在輸入資料前值得花點時間先計劃一下，這樣將來會省下很多的時間。

資料輸入的格式 (Formats for data entry)

資料輸入與儲存有不同的方式。大部份的統計軟體可以讓我們直接輸入資料，然而這樣作有一個限制就是你再來沒辦法將資料移到另一個軟體。簡單的變通方法是將資料存在試算表裡或是資料庫軟體裡；只是，這些軟體的統計功能多半不強，常常還是需要用特定的統計軟體來輔助分析。

一個比較有彈性的方法是把資料轉換成為 ASCII 碼或是文字 (text) 檔。一旦存成 ASCII 碼，大多數的軟體都可以讀取這種資料。ASCII 碼的格式是成排的文字格式，可以從電腦螢幕來查看。通常檔案的每個變數與變數之間會用一個限制元 (delimiter) 分開來，它可能是一個空格 (space) 或是一個逗點 (comma)。這個叫做自由格式 (free format)。

用 ASCII 碼格式輸入資料的最簡單方法是用文字編輯軟體直接輸入資料，或者已輸入試算表的資料也可以用 ASCII 的格式儲存。不管用那一種方法，一般都習慣同一個個體的資料擺在一排，一個變數的資料擺在一欄。變數有時會多到得換行。

輸入資料前的計劃 (Planning data entry)

收集資料需要用問卷來作紀錄。小心計劃好的話，事後輸入資料時你可以省下很多時間。一般問卷會有空格來紀錄資料用一通常資料數字的每一位數，空格獨立分開以便輸入。

類別資料 (Categorical data)

一些統計軟體在處理非數值變數時會有問題，因此你得在輸入電腦前替各個特定類別編一個數字碼，比如說，“不痛”“輕微疼痛”“中度疼痛”“嚴

重疼痛”的數字碼分別為 1，2，3，4，在收集資料時可以直接填在問卷上，像二元資料的‘是’‘否’，可以指定 1 與 0 數字碼。

- 單一碼的變數 (Single-coded variables) — 某問題只有一個答案，比如說“病人過世了沒？”這種問題的答案不可能同時回答“是否”。
- 多重碼的變數 (Multi-coded variables) — 問題可以同時存在有一個以上的答案，比如說，“病人出現什麼症狀？”，症狀可以不只一種。有兩種方法可以來處理這種資料，視情況而定：

- ①可能的症狀就是那些，病人可以出現它們其中的大部份症狀。這時我們可以創造一些不同的二元變數看看個別的症狀出現與否。比如說，“病人有沒有咳嗽？”“病人的喉嚨痛不痛？”
- ②可能的症狀有很多很多不過病人會出現的症狀只有其中的一些。這時我們可以創造不同的類別變數，再來的每一個變數讓你可以知道病人的症狀為何，比如說，“病人出現的第一個症狀是什麼？”“第二個症狀是什麼”，這樣你也許得先決定病人可能有的症狀最多大概數目有多少。

數值變數 (Numerical data)

數值變數的準確度應該跟測量時一樣，變數的單位必須一致，比如說，紀錄重量的單位是公斤就一直是公斤，是磅就一直是磅作單位，不要換來換去。

一個病人有多份問卷時 (Multiple forms per patient)

有時同一個病人不同時間收集了一次以上的資料，最好問卷上有針對個體的特別的辨認方法（例如序號，serial number），可以將某個個體的所有資料彙整在一起。

資料與時間上的一些問題 (Problems with dates and times)

輸入時間和日期應一致，比如說日 / 月 / 年或是月 / 日 / 年，不要換來換去。先弄清楚統計軟體

讀取的格式是那一種。

資料遺失時如何編碼 (Coding missing values)

在輸入資料前先考慮遺失資料要如何處理，大部份的情況下你會用某些符號來代替這些遺失的資料，每種統計軟體有不同的做法，有些使用特殊的符號比如句號或是星形符號來表示，其他有些軟體

你得自己去指定一個碼給它，一般常常給一個值 9，999 或是 -99，值的選定應該是以那個變數來說是不可能出現的數來代表，比如當輸入有四個類別的類別變數資料時（數字碼為 1，2，3，4）你可以選 9 來代表遺失的資料；如果變數是兒童的年齡的話那你得另作選擇。遺失資料在第 3 講有更詳細的討論。

例題

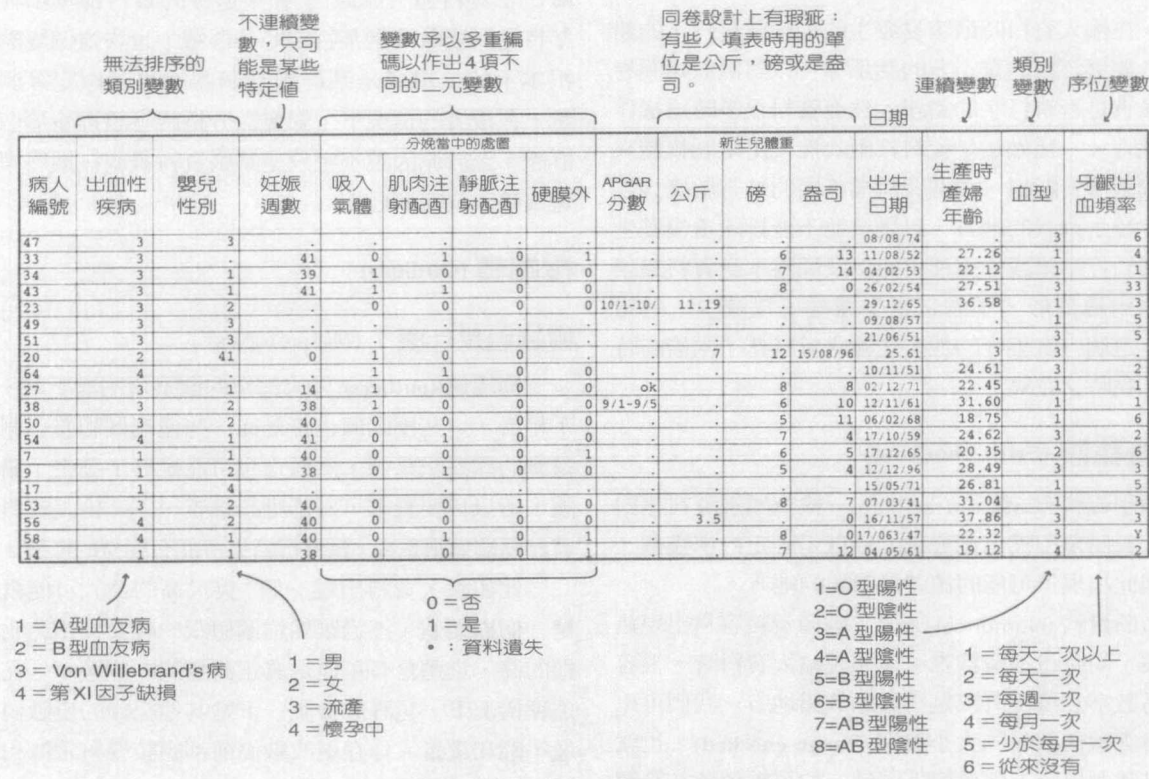


圖 2.1 64 個罹患遺傳性出血疾病的女病人的部分資料。

在倫敦的某血友病中心收集了 64 個女病人的資料，以研究遺傳性出血疾病對於懷孕與分娩的影響。收集的資料包括了她們的病情以及她們第一次懷孕的情形。圖 2.1 的資料表有一部分資料還沒檢查錯誤。圖 2.1 下方註明我們對於

類別變數的編碼辦法，每一排代表了一個個體的資料，如果收集資料正當懷孕期間，則媽媽的年齡以到達預產期時的年齡來紀錄。與活產的相關資料列在第 34 講。

Data kindly provided by Dr R.A. Kadir, University Department of Obstetrics and Gynaecology, and Professor C.A. Lee, Haemophilia Centre and Haemostasis Unit, Royal Free Hospital, London.

任一個研究蒐集的資料都有可能發生錯誤，也許是發生在一開始測量時、收集資料時、轉錄資料時，或是將資料輸入電腦時發生。要完全把錯誤排除不太容易。不過我們可以在資料輸入後小心地檢查以減少打字時或是轉錄時所發生的錯誤。光用眼睛掃視一下也許就可以發現很明顯的錯誤。本講談我們提供一些檢查錯誤的方法。

打字的錯誤 (Typing errors)

在輸入資料時最容易發生的錯誤就是打字的錯誤。如果資料的量不多的話那你可以對照原始問卷來檢查是否有打字的錯誤。然而資料很多時這樣作很耗時。一種做法是資料打兩次然後用電腦軟體來比較兩個資料檔。兩個資料檔不同的地方就應該是打字發生錯誤的地方，只是這並不能排除重複發生同樣的打字錯誤的這種可能以及問卷本身資料錯誤的這兩種情形，不過它至少還是可以減少一些錯誤。這個方法不好的地方是輸入資料花了兩倍的時間，耗時又花錢。

檢查錯誤 (Error checking)

- 類別資料 (Categorical data) — 檢查類別資料相對上較簡單因為每個變數的值就是固定的那幾個，因此出現不可能的值就是錯誤的地方。
- 數值資料 (Numerical data) — 數值資料容易出現錯誤，同時也很難檢查。比如說輸入資料時，很容易數字互換位置或是小數點放錯地方。我們可以檢查數值資料的大小範圍 (range checked)，也就是我們界定上限值與下限值，如果數值落在範圍以外則把它標示出來，然後調查原因。
- 日期 (Dates) — 日期正確與否常常很難檢查，雖然也許你會有概念知道日期資料的大概範圍為何。你可以檢查確定它們是否對或不對，比如說，二月30日，或是月超過12一定不對。也可以用邏輯的原則來檢查錯誤，比如說，病人的出生資料應該和年齡相符合，或病人的出生日期應該在開始研究日期之前。另外追蹤的資料上不可

能出現已經身故的病人！

所有的檢查必須先確定真的有錯誤才能更改，不可以只因為出現不尋常的數值就擅自更改數值。

遺失資料的處理 (Handling missing data)

遺失資料也是會發生的情況。如果有很多的資料都遺失了，分析結果根本就不可靠。我們應該了解一下資料遺失的原因-如果遺失的資料都集中在某特定的變數或是某特定的子群體，也許這個變數根本不能使用或是某群體的個體都沒有作紀錄測量。在後者的情況下，對變數分析時應該排除那個群體。至於原因說不定只是某張資料紙放在抽屜裡還沒輸入資料而已！

極端值 (Outliers)

極端值是什麼？(What are outliers?)

極端值 (outliers) 與大部份的觀察值明顯不同，不相合。這些極端值也許是從一些極端個體真正測量到的觀察紀錄值，然而它也可能是打字發生了錯誤，所以只要覺得可疑都應該檢查一下。檢查資料有無極端值很重要，對分析結果可能有相當的影響。

比如說，資料出現一個7英尺高的女人可能就是一個極端值，不過雖然這個值跟一般女人相比明顯的高，它還是有可能是真正測量到的觀察值。在這種情況下，也許先檢查一下她其他紀錄的變數值像年齡和重量。你在更改數值前都應該要有證據和理由才能作更動。

檢查極端值 (Checking for outliers)

如果資料不多而且極端值明顯地高或是明顯地低於其他，最簡單的方法就是印出來然後目視檢查。檢查數值範圍也是一個方法，另外我們可以把資料製圖（第4講）—用長條圖 (histograms) 或是散狀圖 (scatter plots) 可以明顯地將極端值辨認出來。

很重要的一點是不可以因為數值高於或小於期望值就把個體從分析剔除。可是把極端值包括進來做統計分析時又會影響結果。一個簡單的做法是先把它們包括進去作一次分析，然後予以剔除再作一

次分析。如果結果一樣表示極端值的影響不大。可是如果結果變化很大的話，可能要選擇適當而不會被極端值影響的研究方法來作分析。方法包括使用轉換（第9講）或是無母數檢定（第17講）。

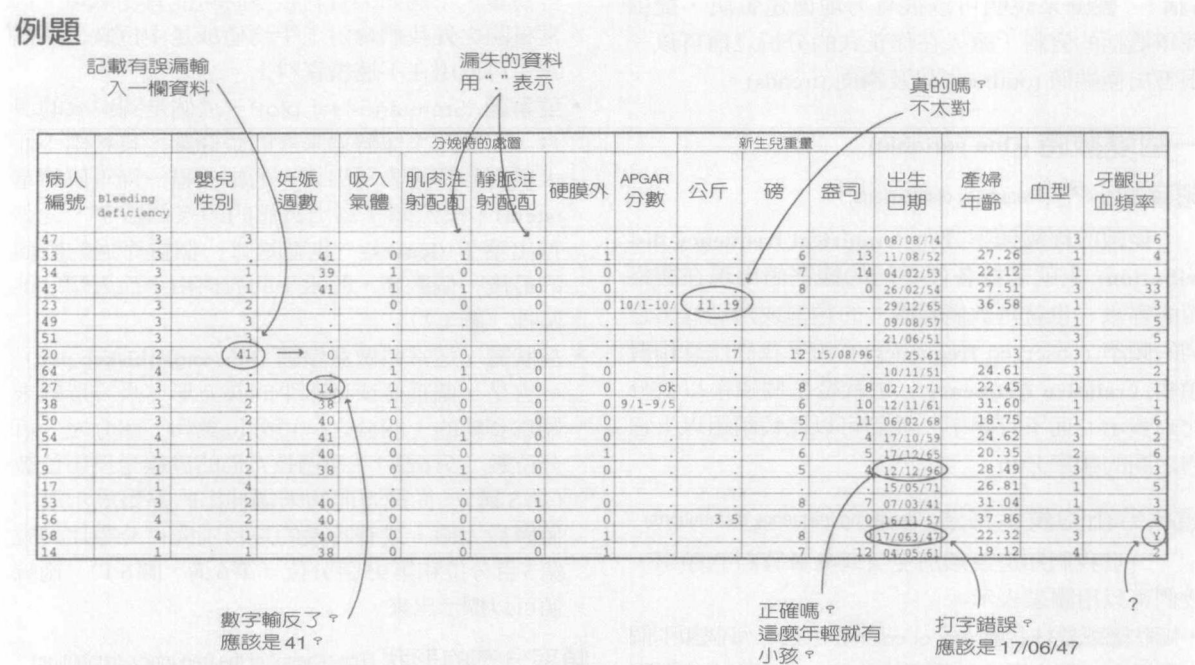


圖3.1 檢查資料記載上有無錯誤

在如第2講所述輸入資料以後，現在要來檢查資料檔的錯誤。這裡標示出來的矛盾處是資料建檔時發生的錯誤。比如說，編號20的病人，寶寶性別欄“41”並不正確，因為性別欄的資料不見了而剩下的資料輪到不正確的欄位裡，其他的值（比如懷孕週數與重量不太尋常）

可能錯誤，但在決定更改以前，應該檢查筆記，因為它們可能是真正的極端值。編號27的病人這個例子，懷孕週數41週，不可能重量達11.19公斤。由於我們沒辦法找到正確的值我們把它以數值遺失來表示。