國立雲林科技大學工業工程與管理所
Graduate school of Industrial Engineering & Management,
National Yunlin University of Science & Technology

系統可靠度實驗室 System Reliability Lab.
http://campusweb.yuntech.edu.tw/~qre/index.htm

Statistical Quality Control for DNA Microarray Data: A Model of Type I Error

出處： Quality Engineering, 20:426–434, 2008
作者：Justin R. Chimka, Kevin J. Oden
報告學生：陳昀名
指導老師：童超塵 教授

- BACKGROUND
- INDIVIDUALS CONTROL CHART FOR ‘‘VALUE’’
- Multivariate Process Monitoring
- Common Observations Out of Control
- Models of Type I Error
- DISCUSSION AND CONCLUSIONS

BACKGROUND

- DNA microarray(稱DNA芯片)是一種複合分子生物學技術

- 依不同需求有不同品質要求(Steinmetz and Davis, 2004)
- 無特定標準評估芯片 品質(Ji and Davis, 2006)
- 因此，靠著複製芯片 去發現問題，既費時又耗成本

MicroArray Quality Control (MAQC Consortium, 2006)

- 致力於RNA(核糖核酸)樣本分析

External RNA Controls Consortium (ERCC, 2005)

- 開發管制以及分析系統

- 大部份的quality-control (QC)並無實際驗證或是無具體方法進行(Allison et al., 2006)

- 數據取得： National Center for Biotechnology Information(NCBI；http://www.ncbi.nlm.nih.gov)

- GSM123472
- GSM123473

Appendix: GSM123472 Sample Data Set

ID_REF	Spot mean intensity (Cyanine3)	Spot mean intensity (Cyanine5)	Background median intensity (Cyanine3)	Background median intensity (Cyanine5)	Spot normalized intensity (Cyanine3)	Spot normalized intensity (Cyanine5)	Spot SNR (Cyanine3)	Spot SNR (Cyanine5)	Spot ratio (Cyanine5)	Spot Log2 ratio (Cyanine5)	VALUE
1	11962.66	8654.26	257.5	265.5	11705.16	8388.76	17.4889	20.8672	0.8075145	−0.3084399	0.091164
2	7292.07	4421.14	441.5	250.5	6850.57	4170.64	7.6602	14.0176	0.6859708	−0.5437809	−0.043003
3	7936.64	5846.45	60	262	7876.64	5584.45	8.09997	21.03185	0.798858	−0.323989	0.176277
4	5820.16	4492.93	31	244	5789.16	4248.93	7.8408	8.97231	0.8269792	−O.2740771	0.230175
5	8915.94	7470.35	74	362	8841.94	7108.35	9.41442	13.521	0.9058391	−0.1426733	0.350781
6	7365.39	4854.05	52	235	7313 39	4619.05	7.07285	8.07475	0.7116463	−0.4907678	0.01473
7	8252.26	5358.76	52	126	8200.26	5232.76	12.35951	10.38593	0.7190068	−0.4759226	0.024803
8	4308.32	2782.942	69	162	4239.32	2620.942	6.24781	7.07601	0.6966126	−0.5215716	0.017512
9	5974.58	3705.33	72	268	5902.58	3437.33	8.09862	8.46053	0.6561583	−0.6078841	−0.099386
10	11347.14	5941.62	112	151	11235.14	5790.62	8.96749	17.39688	0.5807333	−0.7840523	−0.406066
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
53847	18940.79	15679.45	63	285	18877.79	15394.45	39.6795	41.4917	0.9188465	−0.1221042	0.005015
53848	12887.1	8371.96	0	360	12887.1	8011.96	13.29441	22.59044	0.7005093	−0.5135239	−0.308673
53849	17601.06	14357.47	309.5	336.5	17291.56	14020.97	13.94017	27.43378	0.9136375	−0.1303063	0.004653
53850	399.263	424.362	0	271	399.263	153.3625	−0.1810586	0.005187672	0.4328033	−1.208217	−0.287016
53851	15728.66	11724.22	0	274	15728.66	11450.22	21.90927	31.54038	0.820261	−0.2858451	−0.131451
53852	19357.14	16925.59	0	154	19357.14	16771.59	20.93652	65.425	0.9762544	−0.03467098	0.085646
53853	18152.3	16430.38	9	222	18143.3	16208.38	29.18001	42.4161	1.006592	0.009478454	0.142458
53854	23212.09	18503.41	29.5	68.5	23182 59	18434.91	32.3453	61.9608	0.8960021	−0.158426	−0.053479
53855	15775.18	13296.25	46.5	159.5	15728.68	13136.75	22.87246	32.0691	0.9410776	−0.08761434	0.068802
53856	209.5125	378.95	82.5	245	127.0125	133.95	−0.3394013	0.04018176	1.1883	0.2488997	1.335643

- “Value”
- Value is the normalized log (pre value)=(F1-B1)/(F2-B2)

- F1=arithmetic mean from channel 1
- F2=arithmetic mean from channel 2
- B1=arithmetic median from channel 1
- B2=arithmetic median from channel 2

(Khojasteh et al ; 2005)

- We analyze the Value with control charts for individual measurements, and the 10 original variables with Hotelling T^2 control charts, subject to traditional and more conservative decision rules for rejecting the hypothesis of control.

- Type I errors.

INDIVIDUALS CONTROL CHART FOR "VALUE"

- X random variable，define the moving range as

$$MR_i = |X_i - X_{i-1}| \qquad (1)$$

- The sample average is x-bar,

$$UCL = \bar{\bar{x}} + 3\frac{\overline{mr}}{d_2} = \bar{\bar{x}} + 3\frac{\overline{mr}}{1.128}$$

$$CL = \bar{\bar{x}} \qquad (2)$$

$$LCL = \bar{\bar{x}} - 3\frac{\overline{mr}}{d_2} = \bar{\bar{x}} - 3\frac{\overline{mr}}{1.128}$$

- The average of moving ranges is mr-bar

$$UCL = D_4\overline{mr} = 3.267\overline{mr}$$

$$CL = \overline{mr} \qquad (3)$$

$$LCL = D_3\overline{mr} = 0$$

- 其中 $$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad (4)$$

$$\overline{mr} = \frac{1}{m-1}\sum_{i=1}^{m-1} mr_i \qquad (5)$$

- m=microarray sample
- n=2可查表得到d2, D3, and D4

- LCL為0

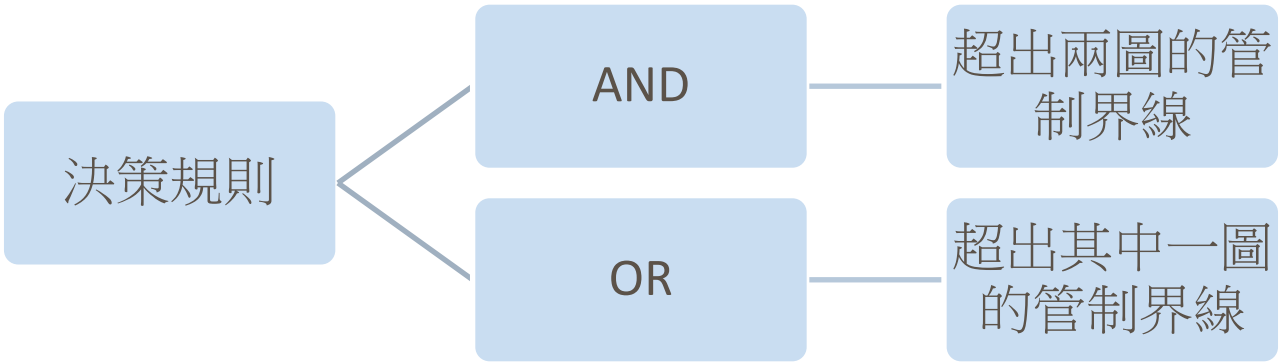

TABLE 1 Example Computation

		Individuals control chart			MR control chart				
Value	Value MR	UCL	LCL	OUT	UCL	LCL	OUT	AND	OR
0.212728	0.18481	0.720475	−0.7456	0	0.900444	0	0	0	0
−1.04172	1.254445	0.720475	−0.7456	1	0.900444	0	1	1	1
−1.50273	0.461009	0.720475	−0.7456	1	0.900444	0	0	0	1

- 總和各個OR以及AND
- OR是可能包含到AND的，因此OR point都會大於AND
- m=53856

TABLE 2 Individuals Out of Control

Data set	Number of "OR" points	Number of "AND" points
GSM123472	1879	587
GSM123473	1640	943
GSM123474	1956	606
GSM123475	1753	580
GSM123476	1704	516
GSM123478	1921	678
GSM123479	1556	438
GSM123481	1170	259
GSM123482	1423	361

Multivariate Process Monitoring

- 通常DNA microarray資料分析都透過T^2管制圖進行分析
- 多變量監控分為兩階段
- 階段I管制界線如右公式

$$UCL = \frac{(m-1)^2}{m}\beta_{\frac{\alpha}{2};\frac{p}{2};\frac{Q-p-2}{2}} \quad (6)$$

$$LCL = 0 \quad (7)$$

$$Q = \frac{2(m-1)^2}{3m-4} \quad (8)$$

- 其中

m=number of observations

P=number of characteristics/variables

β=beta distribution

- Hotelling T^2

$$T^2 = (x - \bar{x})' S^{-1} (x - \bar{x}) \qquad (9)$$

$$S = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} v_i v_i \qquad (10)$$

其中

$$v_i = x_{i+1} - x_i \qquad (11)$$

x=individual data point

x-bar=sample mean vector

m=number of observations

- H2+也是會包含H1的

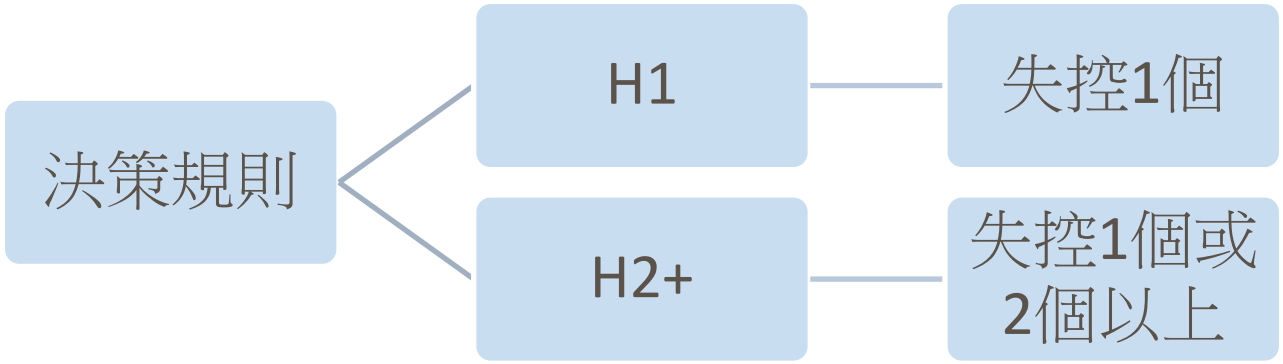

TABLE 3 Hotelling Observations Out of Control

Accession	Number of "H1" points	Number of "H2+" points
GSM123472	994	180
GSM123473	1924	796
GSM123474	888	418
GSM123475	3273	480
GSM123476	1429	250
GSM123478	1354	213
GSM123479	888	522
GSM123481	1565	355
GSM123482	2067	308

Common Observations Out of Control

- 將兩者作組合後如表4
- y=共同失控(稱為response)

TABLE 4 What is common among decision rules?

Data set	Decision rule 1	Decision rule 2	*y*
GSM123472	OR	OR	1879
	OR	H1	340
	OR	H2+	77
	AND	AND	587
	AND	H1	255
	AND	H2+	76
	H1	H1	994
	H2+	H2+	180
GSM123474	OR	OR	1956
	OR	H1	123
	OR	H2+	37
	AND	AND	606
	AND	H1	73
	AND	H2+	25
	H1	H1	888
	H2+	H2+	418

- 利用迴歸模型進行估計以及比較

$$y = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k + \epsilon \qquad (12)$$

$$\hat{y} = b_0 + b_1 z_1 + \cdots + b_k z_k \qquad (13)$$

其中

- y=response
- zk=kth predictor
- βk=kth population regression coefficient
- ε=error term
- bk=estimate of kth population regression coefficient
- ŷ=fitted response

- 再利用R^2調整此模型

$$R^2_{\text{adj}} = 1 - \left(\frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}\right)\left(\frac{n-1}{n-p-1}\right) \qquad (14)$$

- yi=ith observed response value
- ŷi=ith fitted response
- y-bar=mean response
- n=number of observations
- p= number of terms in the model

Models of Type I Error

- 在此假設為常態，並確立Z1~Z6

TABLE 5 Variable Meta Model Setup

		Indicator variables					
		Decision rule 1			Decision rule 2		
Decision rule 1	Decision rule 2	z_1	z_2	z_3	z_4	z_5	z_6
OR	OR	0	0	0	0	0	0
OR	H1	0	0	0	0	1	0
OR	H2+	0	0	0	0	0	1
AND	AND	1	0	0	1	0	0
AND	H1	1	0	0	0	1	0
AND	H2+	1	0	0	0	0	1
H1	H1	0	1	0	0	1	0
H2+	H2+	0	0	1	0	0	1

- 尋找最佳線性迴歸
  1. y
  2. $\sqrt{y}$
  3. ln(y)
  4. $\frac{1}{y}$.
- 調整R^2如表6

TABLE 6 Adjusted R^2 Values

Response	Adj-R^2
y	80.0%
$\sqrt{y}$	86.7%
$\ln(y)$	83.8%
$\frac{1}{y}$	40.8%

- 可利用回歸方程式求得如表7之預測值
- H1-H1標準差最大(紅框)，這可能是因為GSM123475的H1-H1過大關係

TABLE 7 What is Observed, and What is Expected

Decision rule		Indicator variables						Expected value		Actual values	
1	2	z_1	z_2	z_3	z_4	z_5	z_6	$y^{1/2}$	y	*Mean*	*Standard deviation*
OR	OR	0	0	0	0	0	0	40.70	1656.49	1666.89	240.68
OR	H1	0	0	0	0	1	0	17.90	320.41	353.33	93.28
OR	H2+	0	0	0	0	0	1	12.10	146.41	138.22	56.13
AND	AND	1	0	0	1	0	0	22.31	497.74	508.00	124.73
AND	H1	1	0	0	0	1	0	15.11	228.31	217.56	67.45
AND	H2+	1	0	0	0	0	1	9.31	86.68	103.22	37.59
H1	H1	0	1	0	0	1	0	39.00	1521.00	1598.00	714.46
H2+	H2+	0	0	1	0	0	1	19.29	372.10	391.33	180.77

- Figure 1 is an illustration of the expected values

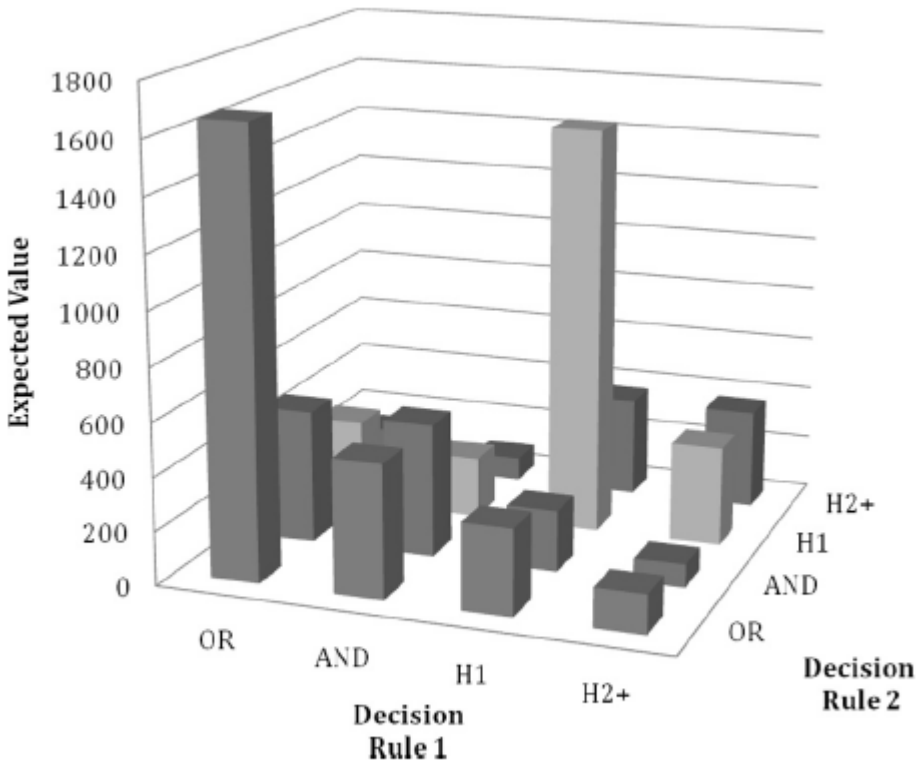

FIGURE 1 Expected values between decision rules.

- 在此也列出相關係數以及P值(相關係數)
- 從Z1可知P=0.057

TABLE 8 Details of the chosen Meta model

$y^{1/2}$	Coefficient	Standard error	t	P
z_1	−2.790637	1.440577	−1.94	0.057
z_2	21.14904	1.905704	11.10	0.000
z_3	7.193926	1.905704	3.77	0.000
z_4	−15.5794	2.495152	−6.24	0.000
z_5	−22.77121	1.905704	−11.95	0.000
z_6	−28.61551	1.905704	−15.02	0.000
Constant	40.71364	1.440577	28.26	0.000

- 最後分析反應(response)以及決策規則(decision rules)

- 在此利用K-means：“assigns each item to the cluster having the nearest mean”(Johnson and Wichern, 2007)分析

DISCUSSION AND CONCLUSIONS

- 此圖為各個規則之型1錯誤，OR與H1明顯都高於AND以及H2+，以及H2+是最接近0.0027

TABLE 9 Average Type I Error Rates across Data Sets

	Type 1 error rate
OR	0.0309514
AND	0.0102497
H1	0.0296717
H2+	0.0072663

- Hotelling T^2 control charts and different decision rules for chart use might be better predicted with Poisson or negative binomial regression.
- 之後可用CUSUM或EWMA等不同方法進行

THE END

K-means

- For example, when we partition the items into two clusters, one of the clusters includes all of the observations associated with OR/OR.
- The remaining items in that same cluster are 2/3 of the observations associated with H1/H1. When we partition the items into three clusters, one of the clusters includes 2/3 of the observations associated with AND/AND.
- Another one of the three clusters includes 8/9 of the observations associated with OR/OR, and the remaining items in that same cluster are 2/3 of the observations associated with H1/H1.

K-means基本定義

- 把每個data point 做數值化,轉換成 n-tuples (X1,X2,X3,..,Xk),然後用 Ecleadian distance方式去計算距離,就是各分量相減平方和開根號, 出各個cluster 的中心,去做 clustering的動作,每加進一筆就從新計算該cluster的mean(中心)

- Department of Industrial Engineering, University of Arkansas
- Fayetteville , Arkansas.