การแปลงข้อมูลสำหรับวิเคราะห์ทางสถิติด้านชีววิทยา

Transformation of Some Experimental Data for Biological Statistics

จิตติ จันทร์แสง	Chitti Chansang
กองกีฏวิทยาทางแพทย์	Divison of Medical Entomology
กรมวิทยาศาสตร์การแพทย์	Department of Medical Sciences

ตีพิมพ์ใน วารสารกรมวิทยาศาสตร์การแพทย์ ปีที่ 32 ฉบับที่ 1 มกราคม-มีนาคม 2533.

บทคัดย่อ

วิธีวิเคราะห์ทางสถิติด้านชีววิทยาที่ใช้ส่วนมาก เช่น t-test, การวิเคราะห์ความแปรปรวน, multiple comparision และ อื่นๆ ข้อมูลที่นำมาทดสอบต้องมีคุณลักษณะตามเงื่อนไขต่างๆ ที่กำหนด บางการทดลองข้อมูลไม่อยู่ในเงื่อนไขที่กำหนด วิธีที่สามารถดำเนินการได้ คือ (i) การวิเคราะห์แบบอื่น เช่น nonparametric methods หรือ (ii) ใช้การแปลงข้อมูลก่อนการวิเคราะห์ เพื่อให้ข้อมูลเป็นไปตามเงื่อนไขที่กำหนดโดยใช้รากที่สอง ลอกการิทึม หรือ อาร์คไซด์ การเลือกวิธีการแปลงข้อมูลวิธีใดขึ้นอยู่กับลักษณะของข้อมูลนั้น เมื่อการแปลงข้อมูลถูกนำมาใช้วิธีการวิเคราะห์ทางสถิติจะสามารถใช้กับข้อมูลที่แปลงค่านั้น แต่เมื่อจะเสนอข้อมูลเกี่ยวกับการประมาณค่าเฉลี่ย หรือ ช่วงความเชื่อมั่นต้องใช้ข้อมูลที่ไม่ได้แปลงค่า

Abstract

For many biological statistics techinques such as t-test, analysis of variance, multiple comparision. The characteristics of data must follow the statistic assumptions. In some experiments, the data do not meet the assumptions. It is necessary therefore (i) using another statistics such as nonparametric methods or (ii) to transform the data to another scale in order to be able to use those statistics techniques. Transformation can be done by using logarithmic, square root or arcsine depend on characteristics of data. When a transformation is applied, test of significance are performed on the transformed data, but estimates of means or confidence interval are usually given in the familiar untransformed scale.

Keywords

Data transformation, statistical analysis

บทความนี้มีจุดมุ่งหมายเพื่อเสริมสร้างและฟื้นฟูความรู้ทางด้านสถิติที่คาดว่าจะเป็นประโยชน์ช่วยทำให้การตรวจสอบนัยสำคัญของผลการทดลองบางเรื่อง ให้มีความถูกต้องเข้ากับเงื่อนไขที่กำหนดอย่างมีประสิทธิภาพ จากการวิเคราะห์ทางสถิติด้านชีววิทยา มักใช้ t-test หรือ z-test แล้วแต่กรณีโดยใช้กับการทดลองที่มี 2 สิ่งทดลอง (treatments) ส่วนการวิเคราะห์หาความแปรปรวน (Analysis of variance) ใช้ได้กับการทดลองที่มีมากกว่า 2 สิ่งทดลองและเป็นการวิเคราะห์ความผันแปรทั้งหมดของข้อมูลที่เกิดขึ้น ว่ามีสาเหตุมาจากอะไรแล้ววัดความผันแปรของข้อมูลที่เกิดขึ้นจากแต่ละสาเหตุออกมา การวิเคราะห์หาความแปรปรวนมักแบ่งออกเป็น 3 แผนการทดลองคือ แผนการทดลองแบบสุ่มตลอด (Completely Random Design; CRD) ใช้เมื่อตัวอย่างมีลักษณะสม่ำเสมอแผนการทดลองแบบสุ่มภายในบล็อก (Randomized Block Design; RBD) ใช้เมื่อทราบทิศทางความแปรปรวนของตัวอย่างนั้น และแผนการทดลองแบบละตินสแควร์ (Latin Square Design; LSD) ใช้เมื่อทราบทิศทางแปรปรวนของตัวอย่างนั้นเป็น 2 แหล่ง จากนั้นสามารถทำการทดสอบต่อไปว่าค่าเฉลี่ยของสิ่งทดลองคู่ใดบ้างที่มีความแตกต่างหรือไม่แตกต่างกันโดยการทดสอบที่เรียกว่า Multiple comparisons ซึ่งมีหลายวิธี แต่ที่ใช้บ่อยและพบมากได้แก่ The Least Significant Difference, Duncan's New Multiple Range Test, Tukey's W procedure เป็นต้น

การวิเคราะห์ที่กล่าวถึงมีข้อกำหนดคือ ทุกตัวอย่างที่ทำการศึกษาจะต้องเลือกมาโดยวิธีสุ่มจากประชากรที่มีการแจกแจงแบบปกติ (normal distribution) ความคลาดเคลื่อน (error) ของการทดลองจะต้องมีการกระจายแบบปกติอย่างอิสระ มีวาเรียนซ์ร่วมกัน (homogeneity of variances) อิทธิพลต่างๆ รวมกันแบบบวกสะสม งานวิจัยส่วนมากจะเข้ากฎเกณฑ์ข้างต้น แต่มีงานวิจัยบางส่วนไม่อยู่ในเกณฑ์ดังกล่าว ซึ่งมักเป็นงานวิจัยที่เกี่ยวข้องกับงานด้านกีฏวิทยาบางส่วน เช่น จำนวนแมลงที่ดักได้จากกับดัก เป็นต้น งานวิจัยซึ่งข้อมูลการทดลองได้มาจากตัวอย่างที่มีการกระจายแบบ Poisson มากกว่า Normal เช่น จำนวนบักเตรีที่นับจากจานเลี้ยงเชื้อ เป็นต้น งานวิจัยที่ข้อมูลการทดลองได้มาจากตัวอย่างที่มีการกระจายแบบ Binomial หรือข้อมูลที่ได้เป็นรูปของอัตราส่วนหรือร้อยละที่ไม่อยู่ในช่วง ร้อยละ 30-70 เช่นจำนวนร้อยละที่เกิดโรคของต้นพืชในแปลงทดลอง เป็นต้น เมื่อพบปัญหาเช่นนี้มี 2 วิธีการที่สามารถดำเนินการคือ ใช้การวิเคราะห์แบบอื่นเช่น nonparametric methods เป็นต้น หรือวิธีการแปลงข้อมูลก่อนการวิเคราะห์ข้อมูลซึ่งต่อไปนี้จะกล่าวถึงเฉพาะวิธีการแปลงข้อมูล

การแปลงข้อมูลก็คือวิธีการจัดการวัดต่างมาตรากันเท่านั้น เช่น 16 มากกว่า 9 ในขณะที่ค่ารากที่สองของ 16 ก็มากกว่ารากที่สองของ 9 เป็นต้น การวิเคราะห์ t-test หรือวิเคราะห์ความแปรปรวน จะใช้กับข้อมูลที่แปลงค่าแล้วสำหรับในกรณีที่ต้องการประมาณค่าเฉลี่ยต่างๆ หรือคำนวณค่า confidence interval จะทำกับข้อมูลเดิม การแปลงข้อมูลก่อนการวิเคราะห์ที่นิยมใช้ 3 วิธีคือ การแปลงโดยใช้รากที่สอง การแปลงโดยใช้ลอกการิทึม และการแปลงโดยใช้อาร์คไซด์ ปัญหาที่ตามมาก็คือจะเลือกวิธีการแปลงข้อมูลแบบใดถึงจะเหมาะสมกับข้อมูลการทดลองที่มีอยู่ซึ่งจะอธิบายในรายละเอียดต่อไป

การแปลงโดยใช้รากที่สอง (The square root transformation, $\sqrt{X}$)

ใช้กับข้อมูลที่ได้จากการนับ เช่น การนับจำนวนโคโลนีของบักเตรีบนจานเลี้ยงเชื้อ จำนวนเม็ดเลือดใน haemocytometer square จำนวนแมลงหรือพืชในพื้นที่ที่กำหนด เป็นต้น ข้อมูลที่ได้มักมีการกระจายแบบ Poisson มากกว่า Normal ซึ่งค่าเฉลี่ยกับค่าวาเรียนซ์ มักจะใกล้เคียงกันหรือเป็นสัดส่วนกัน (proportional) การแปลงโดยใช้รากที่สองนี้ช่วยทำให้ค่าเฉลี่ยและวาเรียนซ์เป็นอิสระกันและทำให้การกระจายแบบ Poissonเดิมเป็น Normal มากขึ้น การแปลงโดยใช้รากที่สองนี้ อาจใช้ $\sqrt{X}$ หรือใช้$\sqrt{X+0.5}$ ในกรณีที่มีค่าศูนย์ (0) ในข้อมูล ตัวอย่างการแปลงโดยใช้รากที่สอง ดังตารางที่ 1

ตารางที่ 1 จำนวนตัวเต็มวัยแมลง *Drosophila* sp. ***ที่เกิดจากหนอนที่เลี้ยงด้วยอาหาร 2 ชนิด โดยทดลองแบบ Single-pair cultures โดยอาหาร A มี DDT ผสมอยู่***

(1) Number of Flies emerging Y	(2) Square root of umber of flies $\sqrt{Y}$	(3) Medium A f	(4) Medium B f
0	0.00	1	-
1	1.00	5	-
2	1.41	6	-
3	1.73	-	-
4	2.00	3	-
5	2.24	-	-
6	2.45	-	-
7	2.65	-	2
8	2.83	-	1
9	3.00	-	2
10	3.16	-	3
11	3.32	-	1
12	3.46	-	1
13	3.61	-	1
14	3.74	-	1
15	3.87	-	1
16	4.00	-	2
		15	15
Untransformed variable			
$\overline{Y}$		1.933	11.133
S^2		1.495	9.410

Square root tranformation		
$\overline{\sqrt{Y}}$	1.299	3.307
$s^2_{\sqrt{Y}}$	0.2634	0.2099

Tests of equality of variances

Untransformed

$$F_s = \frac{S_2^2}{S_1^2} = \frac{9.410}{1.495} = 6.294^{**} \qquad F_{0.025[14.14]} = 2.98$$

Transformed

$$F_s = \frac{s^2_{\sqrt{Y_1}}}{s^2_{\sqrt{Y_2}}} = \frac{0.2634}{0.2099} = 1.255 \text{ ns}$$

Back-transformed (squared) means

	Medium A	Medium B
$\left(\overline{\sqrt{Y}}\right)^2$	1.687	10.937
95% confidence limits		
$L_1 = \overline{\sqrt{Y}} - t_{0.05}s_{\overline{\sqrt{Y}}}$	$1.297 - 2.145\sqrt{\frac{0.2634}{15}}$	$3.307 - 2.145\sqrt{\frac{0.2099}{15}}$
	= 1.015	= 3.053
$L_2 = \overline{\sqrt{Y}} - t_{0.05}s_{\overline{\sqrt{Y}}}$	1.583	3.561
Back-transformed (squared) confidence limits		
L_1^2	1.030	9.324
L_2^2	2.507	12.681

ที่มา : Sokal, R.R. and Rohif, F.J. 1987. Introduction to biostatistics

พบว่า ก่อนการแปลงข้อมูลมีความแตกต่างกันทางด้านสถิติของอาหารที่เลี้ยง แต่เมื่อแปลงด้วยรากที่สองกลับไม่มีความแตกต่างกันทางด้านสถิติ

การแปลงโดยใช้ลอกการิทึม (The logarithmic transformation, log (X))

ใช้กับข้อมูลที่ได้จากการนับเช่นเดียวกับข้อมูลที่ใช้การแปลงโดยใช้รากที่สอง แต่วิธีนี้จะเหมาะสมเมื่อค่าเฉลี่ยและวาเรียนซ์เป็นสัดส่วนกัน (proportional) ข้อมูลที่มีลักษณะการแจกแจงเป็นแบบเบ้ขวา (skewed to the right) การแปลงเป็นลอกการิทึมจะทำให้ทุกค่ามีวาเรียนซ์ใกล้เคียงกัน นอกจากนั้นทำให้อิทธิพลต่างๆ ที่มีผลแบบคูณในข้อมูลเดิมเปลี่ยนเป็นมีผลบวกในมาตราลอกการิทึม ในกรณีที่มีค่าศูนย์ (0) ในข้อมูลการแปลงแบบ log (x+1) ใช้ได้ดีกว่า log (x) ตัวอย่างการแปลงโดยใช้ลอกการิทึม ดังตารางที่ 2

ตารางที่ 2 การทดลองของ Williams ทดลองใช้กับดัก 2 ชนิดในการดักแมลงแต่ละคืน มีความแตกต่างกันหรือไม่ระหว่างกับดัก 2 ชนิดนี้

Trap	Nights						
	1	2	3	4	5	6	Total
Original data							
A	5	15	47	1,000	2	8	1,077
B	4	19	22	99	50	17	211
Logarithmic transformed data							
A	0.70	1.18	1.67	3.00	0.30	0.90	7.75
B	0.60	1.28	1.34	2.00	1.70	1.23	8.15

ที่มา : Williams, C.B. 1937. Appl.Biol. 24: 404-414, (อ้างตาม LECLERG, E.L., LEONARD, W.H., AND CLARD, A.G., 1996)

จากผลการทดลอง แนวโน้มจากทุกๆ คืน กับดัก B สามารถดักแมลงมากกว่ากับดัก A แต่มีเฉพาะคืนที่ 4 ที่กับดัก A มีแมลงเข้าถึง 1,000 ตัว ทำให้ค่าเพิ่มสูงขึ้น ก่อนการแปลงข้อมูลเมื่อดูค่าเฉลี่ยจะพบว่ากับดัก A ดักแมลงได้มากกว่ากับดัก B แต่เมื่อใช้การแปลงเป็นค่าลอกการิทึม แมลงที่เข้ากับดัก B กลับมากกว่ากับดัก A

การแปลงโดยใช้อาร์คไซด์ (The arcsine transformation, arcsin$\sqrt{X}$)

ใช้กับข้อมูลที่อยู่ในรูปร้อยละหรืออัตราส่วน มักมีลักษณะการแจกแจงแบบ Binomial เป็นเหตุการณ์ที่ค่าสังเกตมีเพียงเกิดหรือไม่เกิดเหตุการณ์นั้นเช่น อัตราส่วนหรือร้อยละของหนูตัวเมียในประชากรหนูที่ศึกษา เป็นต้น แต่ถ้าข้อมูลนั้นอยู่ในช่วง ร้อยละ 30-70 ก็ไม่จำเป็นต้องแปลงข้อมูล วิธีการแปลงใช้สูตร

$$\text{angle} = \arcsin\sqrt{\text{PERCENTAGE}}$$

ซึ่งสามารถหาค่าได้จากการเปิดตารางในหนังสือสถิติหรือใช้วิธีคำนวณ เช่น ข้อมูลเดิมร้อยละ 43.1 หาค่าองศาได้ดังนี้ หารากที่สองของ 0.431 = 0.6565 ใช้เครื่องคิดเลขหา $\sin^{-1}$ (0.6565) = 41.03 เป็นต้น ตัวอย่างการแปลงข้อมูลโดยใช้อาร์คไซด์ ดังตารางที่ 3

ตารางที่ 3 การทดลองใช้วิธีการป้องกัน 3 แบบ เพื่อป้องกันการทำลายของหนอนเจาะฝักข้าวโพด โดยวิธี A เป็นวิธีเปรียบเทียบ

จำนวนร้อยละของฝักข้าวโพดที่ถูกทำลาย

Treatments	BLOCK 1	2	3	4	5	6	MEAN	$(SD)^2$	
		Original data							
A	42.4	34.3	24.1	39.5	55.5	49.1	40.82	121.88	
B	33.3	33.3	5.0	26.3	30.2	28.6	26.11	114.38	
C	8.5	21.9	6.2	16.0	13.5	15.4	13.58	31.69	
D	16.6	19.3	16.6	2.1	11.1	11.1	12.80	38.32	
				Angle = arcsin $\sqrt{\text{PERCENTAGE}}$					%
A	40.6	35.8	29.4	38.9	48.2	44.5	39.6	43.56	40.6
B	35.2	35.2	12.9	30.9	33.3	32.3	29.9	72.69	24.9
C	17.0	27.9	14.4	23.6	21.6	23.1	21.3	23.65	13.2
D	24.0	26.1	24.0	8.3	19.5	19.5	20.2	41.25	11.9

ที่มา : Cochran W.G., 1940 Ann. Math. Statist., 11: 344-348 อ้างตาม Snedecor, G.W., Cochran, W.G., 1967

จากวาเรียนซ์ = $(SD)^2$ ก่อนการแปลงข้อมูล ค่าวาเรียนซ์ของแต่ละสิ่งทดลองมีการกระจายอยู่ระหว่าง 31.69-121.88 แต่เมื่อทำการแปลงข้อมูลค่าวาเรียนซ์ของแต่ละสิ่งทดลองอยู่ระหว่าง 23.62-72.59 กล่าวได้ว่าทำให้ค่าวาเรียนซ์มีเอกภาพหรือร่วมมากขึ้น (hemogeneity of variances) ตามข้อกำหนดของวิธีการที่ต้องการให้วาเรียนซ์มีเอกภาพหรือรวมกัน (homogeneity of variance)

การตรวจสอบนัยสำคัญของผลการทดลอง ซึ่งมีข้อกำหนดว่าความคลาดเคลื่อนของการทดลองต้องมีการกระจายแบบปกติอย่างอิสระ มีวาเรียนซ์รวมกัน อิทธิพลต่างๆ รวมกันแบบบวกสะสม จากการใช้วิธีการสุ่มประกันความเป็นอิสระกันระหว่างความคลาดเคลื่อน ในบางครั้งความคลาดเคลื่อนอาจมีวาเรียนซ์ไม่เท่ากัน อาจเนื่องจากการกระจายเป็นแบบไม่ปกติ ซึ่งอาจสังเกตจากบางสิ่งทดลองมีค่าเฉลี่ยสูงและมีวาเรียนซ์ของค่าเฉลี่ยสูงกว่าสิ่งทดลองอื่นๆ ในกรณีเช่นนี้จำเป็นต้องแปลงข้อมูลซึ่งมีจุดประสงค์เพื่อแปลงข้อมูลแล้วข้อมูลใหม่มีการกระจายเป็นแบบปกติหรือใกล้เคียง ค่าเฉลี่ยและวาเรียนซ์ของข้อมูลที่แปลงแล้วเป็นอิสระต่อกัน ซึ่งทำให้วาเรียนซ์มีลักษณะเอกภาพ สำหรับข้อมูลทั้งหมดวิธีการแปลงข้อมูลมีการใช้รากที่สองลอกการิทึมหรืออาร์คไซด์ซึ่งจะเหมาะกับข้อมูลต่างๆ ตามที่กล่าวมาแต่ถ้ามีปัญหาไม่ทราบจะใช้วิธีการแปลงข้อมูลแบบใด ก็อาจใช้วิธีเขียนกราฟการกระจายของข้อมูลที่มีอยู่โดยใช้สเกลต่างๆ คือ รากที่สองหรือลอกการิทึมหรืออาร์คไซด์ แล้วดูว่าสเกลใดที่ทำให้การกระจายข้อมูลเป็นแบบปกติมากที่สุดก็ใช้วิธีการนั้นแปลงข้อมูล แต่อย่างไรก็ตามการแปลงข้อมูลนี้เป็นเพียงวิธีการหนึ่งทางสถิติที่จะช่วยแก้ปัญหาตามข้อกำหนดดังกล่าว แต่ถ้าไม่สามารถแก้ปัญหาได้ก็ต้องใช้วิธีการวิเคราะห์แบบอื่นแทน

เอกสารอ้างอิง

1. จรัญ จันทลักขณา. 2523. สถิติวิธีวิเคราะห์และวางแผนงานวิจัย. สำนักพิมพ์ไทยวัฒนาพานิช จำกัด กรุงเทพฯ. หน้า 468.
2. สมบูรณ์ สุขพงษ์ และ เปรมใจ ตรีสรานุวัฒนา. 2524. หลักสถิติ 2 วิธีวิเคราะห์และการวางแผนงานทดลองเบื้องต้น. มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ. หน้า 179.
3. Leclerg, E.L., Leonard, W.H. and Clark, A.G. 1966. Field Plot Technique. 2D ED., Burgess Publishing Company, Minneapolis, Minnesota. p.373.
4. Snedecor, G.W., and Cochran, W.G. 1967. Statistical Methods. 6th ed., The Iowa State University Press, AMES, IOWA. p.593.
5. Bailey, N.J. 1959. Statistical Methods in Biolog. The English Universities Press Ltd. Warwick Lane, London. p.200.
6. Sokal, R.R., and Rohif, F.J. 1987. Introduction to biostatistics. 2nd.ed., W.H. Freeman and Company, New York. p.363.