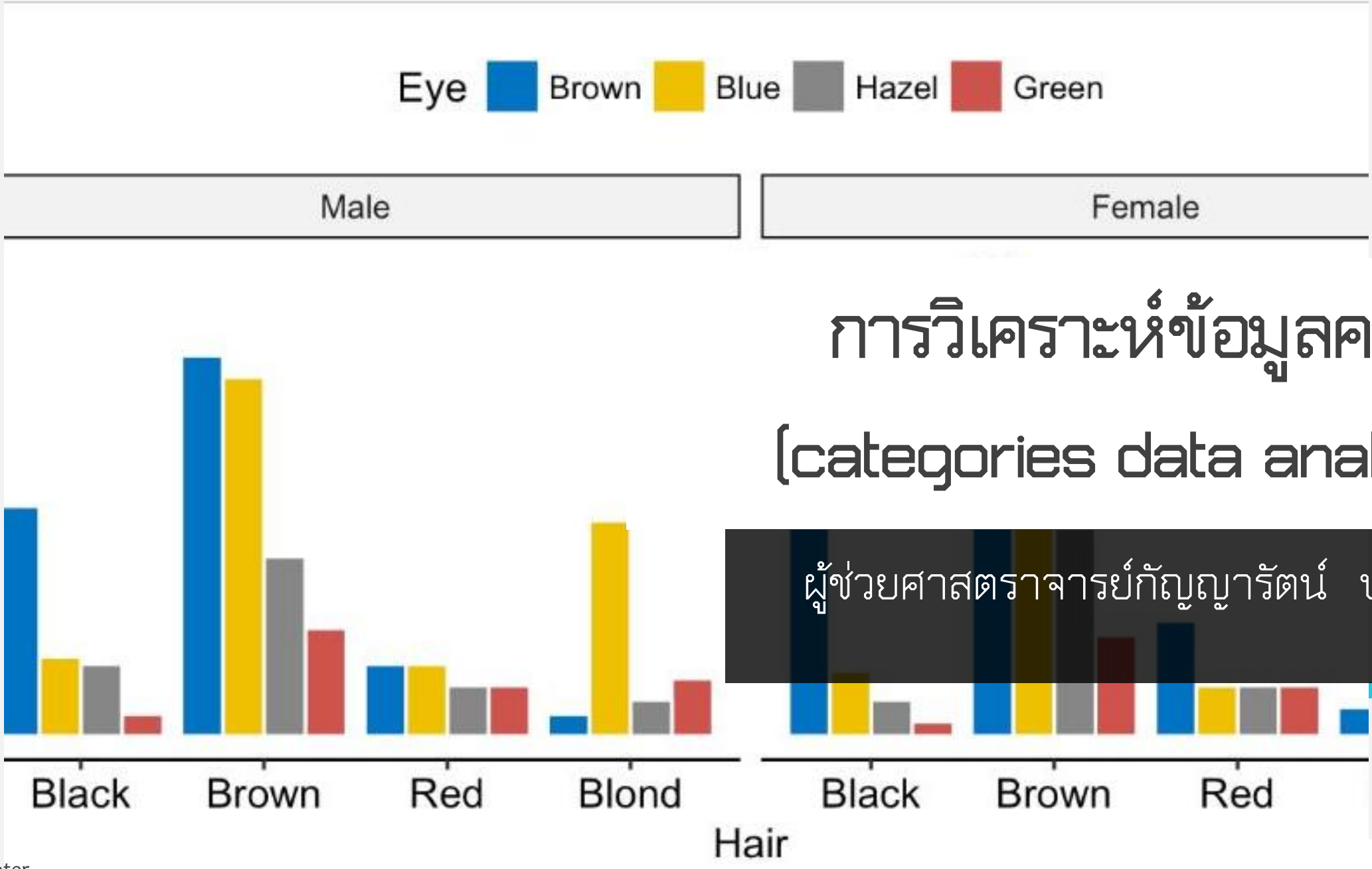

การวิเคราะห์ข้อมูลความถี่ (categories data analysis)

ผู้ช่วยศาสตราจารย์กัญญารัตน์ บุษบรรณ

หลักการ

ศึกษาเปรียบเทียบจำนวนความถี่ที่สังเกตได้จากตัวอย่างเรียกว่า Observed Frequency แทนด้วย O_i กับจำนวนความถี่ที่คาดว่าจะเกิดขึ้นหรือเป็นไปได้ตามทฤษฎีเรียกว่า Expected Frequency แทนด้วย E_i

เหตุการณ์	1	2	3 ...	k
ความถี่ที่สังเกตได้	O_1	O_2	O_3 ...	O_k
ความถี่คาดหวัง	E_1	E_2	E_3 ...	E_k

Jump to first page

χ^2 - Test

การทดสอบความเหมาะสม (Goodness of Fit Test)

- การทดสอบอัตราส่วน
- การทดสอบการแจกแจงของประชากร

การทดสอบความเป็นอิสระ (Test of Independence)

การทดสอบอัตราส่วน : ข้อมูลแจกแจงทางเดียว

สถิติที่ใช้ทดสอบ: $\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - E_i\right)^2}{E_i}$

χ^2 มีการแจกแจงประมาณด้วยการแจกแจงไคสแควร์มีองศาแห่งความอิสระเท่ากับ k-1

โดยที่ k คือจำนวนเหตุการณ์ที่สนใจ

การทดสอบอัตราส่วน : ข้อมูลแจกแจงทางเดียว

สมมติฐานทดสอบ

$$H_0 : A_1 : A_2 : A_3 : \ldots : A_k = C_1 : C_2 : C_3 : \ldots : C_k$$

$$H_1 : A_1 : A_2 : A_3 : \ldots : A_k \neq C_1 : C_2 : C_3 : \ldots : C_k$$

ตัวสถิติทดสอบ

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

ขอบเขตวิกฤต $\chi^2_{cal} \geq \chi^2_{\alpha}$ ที่จำนวนองศาความเป็นอิสระเท่ากับ k-1

การสรุปผลการทดสอบ ถ้า χ^2_{cal} ตกอยู่ในบริเวณวิกฤต สรุปได้ว่าข้อมูลที่รวบรวมได้มาจากประชากรซึ่งการเกิดขึ้นของเหตุการณ์ต่างๆไม่เป็นไปตามอัตราส่วนของสมมติฐาน H_0

ตัวอย่างที่ 1 จากข้อมูลในตารางต้องการทดสอบว่าการลางานของบุคลากรในองค์กรในแต่ละวันเป็นอัตราส่วนเท่ากับ 2:1:2:1:2 หรือไม่ โดยใช้ α =0.01

วัน	จันทร์	อังคาร	พุธ	พฤหัสบดี	ศุกร์
จำนวนพนักงาน	124	74	104	98	120

สมมติฐานทดสอบ

H_0 : $A_จ : A_อ : A_พ : A_{พฤ} : A_ศ = 2 : 1 : 2 : 1 : 2$

H_1 : $A_จ : A_อ : A_พ : A_{พฤ} : A_ศ \neq 2 : 1 : 2 : 1 : 2$

วัน	จันทร์	อังคาร	พุธ	พฤหัสบดี	ศุกร์	รวม
O_i	124	74	104	98	120	520
ความน่าจะเป็นภายใต้ H_0(P)	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{2}{8}$	1
E_i	130	65	130	65	130	520
$\frac{(O-E)^2}{E}$	0.28	1.25	5.2	16.75	0.77	24.25

$$\chi^2 = \frac{(124-130)^2}{130} + \frac{(74-65)^2}{65} + \frac{(104-130)^2}{130} + \frac{(98-65)^2}{65} + \frac{(120-130)^2}{130} = 24.25$$

$\chi^2_{0.01,4} = 13.2767$

เนื่องจาก 24.25 >13.2767 ดังนั้น ปฏิเสธ H_0

นั่นคือ การลางานของบุคลากรในองค์กรในแต่ละวันเป็นอัตราส่วนเท่ากับ 2:1:2:1:2

ตัวอย่างที่ 2 ฝ่ายวิจัยการตลาดของบริษัทต้องการประเมินความชอบของลูกค้าเกี่ยวกับสีของเครื่องสุขภัณฑ์ 3 สี คือ ขาว น้ำเงิน ครีม โดยสุ่มตัวอย่างลูกค้าที่เข้ามาเลือกซื้อผลิตภัณฑ์ 150 ได้ข้อมูลดังนี้

	ขาว	น้ำเงิน	ครีม	รวม
จำนวนลูกค้า	63	56	31	150

จากข้อมูลดังกล่าว จงทดสอบว่าความชอบของลูกค้าต่อสีของสุขภัณฑ์แตกต่างกันหรือไม่ ที่ระดับนัยสำคัญ 0.05

สมมติฐานทดสอบ

$H_0 : A_ข : A_น : A_ค = 1 : 1 : 1$

$H_1 : A_ข : A_น : A_ค \neq 1 : 1 : 1$

	ขาว	น้ำเงิน	ครีม	รวม
O	63	56	31	150
P	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1
E	50	50	50	150
$\frac{(O-E)^2}{E}$	3.38	0.72	7.22	11.32

$\chi^2_{0.05,3} = 7.8147$

$$\chi^2_{cal} = \frac{(63-50)^2}{50} + \frac{(56-50)^2}{50} + \frac{(31-50)^2}{50} = 11.32$$

เนื่องจาก 11.32 > **7.8147** ดังนั้น ปฏิเสธ H_0

นั่นคือ ความชอบของลูกค้าต่อสีของสุขภัณฑ์แตกต่างกัน

ข้อจำกัดของการใช้ไคสแควร์

1. ถ้าความถี่คาดหวังของเหตุการณ์ใดน้อยกว่า 5 จะปรับแก้โดยรวมความถี่ของเหตุการณ์นั้นเข้ากับความถี่ของเหตุการณ์ที่อยู่ใกล้เคียง หรือรวมกับเหตุการณ์คล้ายกันลักษณะใกล้เคียง
2. การหาองศาความเป็นอิสระสำหรับข้อมูลที่มี k เหตุการณ์ องศาความเป็นอิสระเท่ากับ k-r-1เมื่อ r คือ จำนวนพารามิเตอร์ที่ต้องประมาณจากข้อมูลเพื่อใช้ในการหาค่าความถี่คาดหวัง
3. กรณีองศาความเป็นอิสระเท่ากับ 1 การทดสอบไคสแควร์จะใช้ได้ไม่ค่อยดีจึงใช้ตัวสถิติของ Frank Yate ซึ่งปรับแก้ตัวสถิติไคสแควร์ เพื่อให้การใช้ไคสแควร์ดีขึ้น

สถิติไคสแควร์ของ Frank Yate

$$\text{corrected}\chi^2 = \sum_{i=1}^{k} \frac{\left(\left|O_i - E_i\right| - \frac{1}{2}\right)^2}{E_i}$$

การทดสอบการแจกแจงของประชากร

ใช้สำหรับทดสอบข้อสงสัยเกี่ยวกับการแจกแจงของตัวแปรที่ต้องการศึกษา

ขั้นตอนการทดสอบ

ขั้นที่ 1 **ตั้งสมมติฐาน** H_0: ตัวแปรมีการแจกแจงตามข้อสงสัย

H_1: ตัวแปรไม่ได้มีการแจกแจงตามข้อสงสัย

ขั้นที่ 2 **คำนวณค่าสถิติ** χ^2_{cal}

ขั้นที่ 3 **พิจารณาบริเวณวิกฤต**ภายใต้ระดับนัยสำคัญ α

$\chi^2_{cal} \geq \chi^2_{\alpha,k-r-1}$ เมื่อ r เป็นจำนวนพารามิเตอร์ที่ต้องประมาณ k เป็นจำนวนเหตุการณ์ที่ปรับค่าคาดหวังแล้ว

ขั้นที่ 4 **สรุปผล** ถ้า χ^2_{cal} ตกอยู่ในบริเวณวิกฤต ไม่สามารถสรุปได้ว่าตัวแปรมีการแจกแจงตามข้อสงสัย

ตัวอย่าง3 ถ้าผู้บริหารไปรษณีย์เชื่อว่าเปอร์เซนต์ที่จดหมายที่ส่งทางไปรษณีย์จะเสียหาย เช่นเลอะเทอะ ฉีกขาดเป็น 15% ทางไปรษณีย์สุ่มคนมา 310 คน จะส่งจดหมายให้คนละ 2 ฉบับ ปรากฏว่ามี 260 คนที่ได้รับจดหมาย 2 ฉบับในสภาพที่ดี มี 40 คนได้รับจดหมายดี 1 ฉบับและเสียหาย 1ฉบับ และมี 10 คนที่ได้รับจดหมายที่เสียหายทั้งสองฉบับ จงทดสอบว่าจำนวนจดหมายที่เสียหายมีการแจกแจงแบบทวินามที่มีสัดส่วนจดหมายเสียหาย เท่ากับ 0.15 ด้วยระดับนัยสำคัญ 0.10

ให้ X = จำนวนจดหมายที่เสียหายของแต่ละคน

เหตุการณ์	O	P	E
X=0	260	0.7225	223.975
X=1	40	0.2550	79.05
X=2	10	0.0225	6.975

สมมติฐานทดสอบ

H_0 : X มีการแจกแจงแบบทวินามที่ n =2 p=0.15

H_1 : X ไม่ได้มีการแจกแจงทวินามที่ p=0.15

ภายใต้ H_0 จะได้ว่า $P_i = P(X=x) = \binom{2}{x}(0.15)^x(1-0.15)^{2-X}; X=0,1,2$

$$P_0 = P(X=x) = \binom{2}{0}(0.15)^0(1-0.15)^{2-0} = 0.7225$$

$$P_1 = P(X=x) = \binom{2}{1}(0.15)^1(1-0.15)^{2-1} = 0.2550$$

$$P_2 = P(X=x) = \binom{2}{2}(0.15)^2(1-0.15)^{2-2} = 0.0225$$

$\chi^2_{cal} = 26.39$

$\chi^2_{0.10,2} = 4.6052$

เนื่องจาก 26.39 > 4.0652 ดังนั้น ปฏิเสธ H_0

สรุป จำนวนจดหมายที่เสียหายไม่ได้มีการแจกแจงทวินามที่ P= 0.15 ที่นัยสำคัญ 0.10

ตัวอย่าง 4 องค์การโทรศัพท์ได้บันทึกจำนวนครั้งของการต่อโทรศัพท์หมายเลขผิดเข้ามาทีโรงพยาบาลแห่งหนึ่ง ในช่วงเวลา 9.00 – 12.00 น. เป็นเวลา 100 วัน ผลการบันทึกดังนี้

จำนวนครั้งที่ต่อหมายเลขผิด	0	1	2	3	4
จำนวนวัน	32	24	24	12	8

อยากทราบว่าจำนวนครั้งที่ต่อหมายเลขผิด น่าจะมีการแจกแจงแบบใด และจงทดสอบการแจกแจงนั้น ที่ระดับนัยสำคัญ 0.01

พิจารณาการเก็บข้อมูลเป็นการเกิดเหตุการณ์ที่สนใจในช่วงเวลาหนึ่ง ถ้าให้ X แทนจำนวนครั้งที่ต่อหมายเลขผิด X = 0,1,2,3,... จะพบว่า X มีคุณสมบัติของการแจกแจงปัวซอง จึงจะทำการทดสอบการแจกแจงปัวซอง

การคำนวณความถี่คาดหวัง ต้องใช้ความน่าจะเป็นของการแจกแจงปัวซอง ซึ่งต้องใช้ค่าพารามิเตอร์ λ ในที่นี้ไม่ได้กำหนดให้ดังนั้นต้องประมาณด้วยค่า เฉลี่ย $\bar{X}$

สมมติฐานทดสอบ

H_0: จำนวนครั้งที่ต่อหมายเลขผิดมีการแจกแจงปัวซอง

H_1: จำนวนครั้งที่ต่อหมายเลขผิดไม่มีการแจกแจงปัวซอง

คำนวณจำนวนครั้งเฉลี่ยที่ต่อหมายเลขผิดในช่วงเวลาดังกล่าว

$$\bar{X}=\sum\frac{fx}{N}=\frac{140}{100}=1.4;\lambda\approx 1.4$$

$$\chi^2_{.01,5-1-1}=11.34 \qquad \chi^2=6.71$$

สรุป จำนวนครั้งที่ต่อหมายเลขผิดมีการแจกแจงปัวซอง ที่ระดับนัยสำคัญ 0.01

x:จำนวนครั้งที่ต่อหมายเลขผิด	จำนวนวัน O:		P(X= x)		E=nP(X= x)	
0	32		0.2466		24.66	
1	24		0.3452		34.52	
2	24		0.2417		24.17	
3	12		0.1128		11.28	
4	8	8	0.0395		3.95	5.37
ตั้งแต่ 5 ตัวขึ้นไป	0		0.0142		1.42	
รวม	100		1		100	

ตัวอย่างที่ 5 ข้อมูลที่กำหนดให้ต่อไปนี้เป็นความยาวของทารกแรกเกิด จำนวน 125 คน สรุปได้หรือไม่ว่าความยาวของทารกแรกเกิดมีการแจกแจงปกติ ที่ระดับนัยสำคัญ .05

ความยาว(ซ.ม.)	45-46.9	47-48.9	49-50.9	51-52.9	53-54.9
จำนวน(คน)	28	32	35	20	10

วิธีทำ 1. H_0 : ความยาวทารกแรกเกิดแจกแจงแบบปกติ

H_1 :ความยาวทารกแรกเกิดไม่แจกแจงแบบปกติ

2. เนื่องจากเราไม่ทราบค่าพารามิเตอร์ μ,σ

ดังนั้นเราจะประมาณ μ ด้วย $\overline{x}$ และประมาณ σ ด้วย s

จากข้อมูลจะได้ $\overline{X}$ =49.18 และ S = 2.45

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 12.91$$

ค่าวิกฤต $\chi^2 .05,6-2-1=7.8$ สรุป ไม่สามารถสรุปได้ว่า ความยาวทารกแรกเกิดแจกแจงแบบปกติ