Basic Statistical Concepts

อ.รุ่งกานต์ บุญนาถกร

สถิติคืออะไร?

สถิติ คือ กระบวนการตัดสินใจจากข้อมูลที่มีการรวบรวมและวิเคราะห์อย่างเป็นระบบ

การเก็บรวบรวมข้อมูล

แบ่งออกเป็น 3 วิธี

1. โดยการสำรวจ (survey)
2. **โดยการทดลอง (experiment)**
3. โดยการสำรวจกึ่งทดลอง (Quasi-experiment)

การทดลอง

การเก็บรวบรวมข้อมูลจากการทดลอง ผู้ศึกษาจะต้องสร้างการทดลองขึ้นมาโดยมีวัตถุประสงค์ที่จะศึกษาอิทธิพลของปัจจัยบางปัจจัย โดยอาจเป็นการทดลองเพื่อศึกษาปัจจัยใดปัจจัยหนึ่ง เพียงปัจจัยเดียวหรือหลายปัจจัยไปพร้อมๆกัน โดยปัจจัยหรืออิทธิพลอื่นที่ไม่ต้องการศึกษาจะต้องถูกควบคุมให้อยู่ในสภาวะเดียวกัน ดังนั้นก่อนจะทำการทดลองจะต้องมีการ**ออกแบบหรือวางแผนให้เหมาะสม** เพื่อให้ได้ข้อมูลที่ถูกต้องเป็นไปตามเงื่อนไขของวิธีการทางสถิติที่ใช้ในการวิเคราะห์ และปราศจากความลำเอียง (bias)

แผนการทดลอง

แบ่งออกเป็น 2 ประเภทใหญ่ๆ

1. แผนการทดลองแบบระบบ (systematic design) เป็นแผนการทดลองที่จัดทรีทเมนต์ให้กับหน่วยทดลองเป็นระบบโดยไม่ได้ใช้การสุ่ม (ทรีทเมนต์บางทรีทเมนต์อาจได้เปรียบทรีทเมนต์อื่น) ถ้าวิเคราะห์ข้อมูลด้วยวิธีการทางสถิติทั่วไปอาจทำให้ได้ผลสรุปที่มีความคลาดเคลื่อนสูง
2. **แผนการทดลองแบบสุ่ม (randomized design) เป็นแผนการทดลองที่จัดทรีทเมนต์ให้หน่วยทดลองโดยคำนึงถึงกฎแห่งโอกาส ทุกทรีทเมนต์จะต้องได้รับโอกาสที่เท่าเทียมกัน**

แผนการทดลองแบบสุ่ม

แบ่งออกเป็น 2 ประเภทใหญ่ๆ

1. แผนการทดลองปัจจัยเดียว (single factor experiment)
   1) **แผนการทดลองแบบสุ่มสมบูรณ์ (CRD)**
   2) **แผนการทดลองแบบสุ่มในบลอคสมบูรณ์ (RCBD)**
   3) แผนการทดลองแบบลาตินสแควร์
2. แผนการทดลองหลายปัจจัย (single factor experiment)
   1) แผนการทดลองแบบสุ่มซ้อน (Nested design)
   2) **การทดลองแบบแฟคทอเรียล (Factorial experiment)**
      a) Factorial in CRD
      b) Factorial in RCBD

การวัด (Measurement)

- การวัด (Measurement) เป็นเครื่องมือพื้นฐานสำหรับงานวิจัย ออกแบบ และพัฒนา การวัดยังเป็นส่วนประกอบหลักในกระบวนการควบคุมการทำงานเพื่อตรวจสอบสมรรถนะที่ถูกต้อง
- การวัด คือ กระบวนการที่ทำการเปรียบเทียบปริมาณที่ไม่ทราบค่าของตัวแปรกับค่ามาตรฐานที่กำหนดไว้ ถ้าค่าวัดได้ใกล้เคียงกับค่าจริงมากเพียงใดแสดงว่ามีความแม่นยำหรือถูกต้องสูง
- การวัดทุกครั้งมีค่าความผิดพลาดเสมอ ในทางปฏิบัติไม่มีเครื่องมือวัดใดที่วัดค่าแล้วมีความถูกต้องเท่าค่าจริง ดังนั้น การใช้เครื่องมือวัดต้องเข้าใจหลักการและเทคนิคพื้นฐานของการวัดอย่างถูกต้อง และคำนึงถึงสาเหตุที่ทำให้ เกิดค่าความผิดพลาดขึ้นเพื่อหลีกเลี่ยงให้เหลือน้อยที่สุด

ความผิดพลาด (Error)

- ความผิดพลาด (error) คือความแตกต่างระหว่างค่าวัด (measured value) กับค่าจริง (true value)
- การวัดค่าตัวแปรในกระบวนการต่างๆนั้น เราไม่มีทางรู้ค่าจริงของตัวแปร ∴ ไม่มีทางรู้ค่าผิดพลาดที่แท้จริง เพราะถ้ารู้ความผิดพลาดเราจะสามารถแก้ไขวิธีเก็บข้อมูลเพื่อให้ได้ข้อมูลที่ถูกต้องได้ นั่นคือเหตุผลหลักที่ต้องมีการปรับเทียบเครื่องมือวัด ในการวัดค่าปริมาณในกระบวนการ

การจำแนกสาเหตุของความผิดพลาด

1. ความผิดพลาดโดยผู้วัด เช่น อ่านค่าไม่ถูกต้อง บันทึกค่าไม่ถูกต้อง คำนวณผลการวัดไม่ถูกต้อง
2. ความผิดพลาดของเครื่องมือวัด เช่น ขาดการบำรุงรักษา อายุการใช้งาน ไม่ได้สอบเทียบ ความละเอียดของเครื่องมือ การใช้เครื่องมือวัดไม่ถูกวิธี
3. ความผิดพลาดจากสิ่งแวดล้อม เช่น อุณหภูมิ ความชื้น การสั่นสะเทือน ความกดอากาศ แรงดันไฟฟ้า สนามแม่เหล็ก ฯลฯ ซึ่งอาจไปรบกวนการทำงานของเครื่องมือ ทำให้การอ่านค่าผิดพลาด

ความผิดพลาด (Error)

แบ่งเป็น 2 ชนิด

- ความผิดพลาดระบบ (systematic error) เกิดจากเครื่องมือ สภาพแวดล้อม หรือผู้วัดส่งผลต่อความถูกต้องของข้อมูล
- ความผิดพลาดสุ่ม (random error) เป็นความผิดพลาดที่เกิดขึ้นทุกครั้งของการวัด ซึ่งควบคุมไม่ได้ อาจเกิดจากผู้วัด หรือเครื่องมือที่ใช้วัด เป็นต้น

ในการวัดถ้าผลการวัดมีความผิดพลาดระบบน้อย นั่นคือมีค่าใกล้เคียงค่าจริงมากเรียกว่าผลการวัดนั้นมี **ความแม่น** (accuracy) สูง และถ้าผลการวัดมีความผิดพลาดแบบสุ่มน้อย นั่นคือมีการกระจายของการวัดน้อย เรียกผลการวัดนั้นมี**ความเที่ยงตรง** (precision) สูง

ความถูกต้อง (Accuracy)

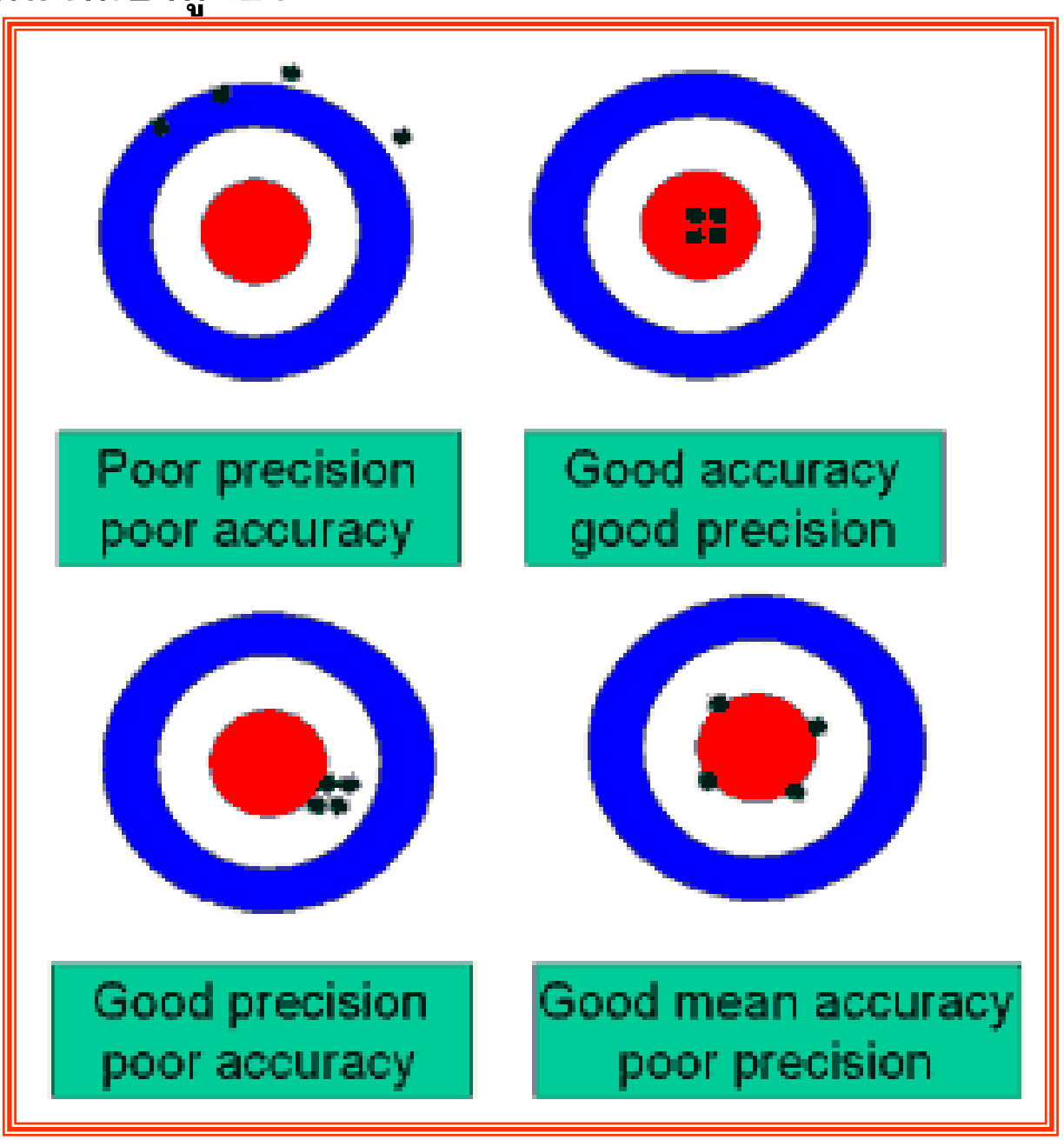
- ความแม่นยำหรือความถูกต้อง (accuracy) คือความสามารถของระบบที่แสดงผลลัพธ์ของการวัดได้อย่างถูกต้อง หรือ ความใกล้เคียงกันระหว่างค่าที่วัดได้กับค่าจริงของตัวแปรที่ถูกวัด

ความเที่ยง (Precision)

- ความเที่ยง (precision) หรือความสามารถในการวัดซ้ำได้ (repeatability) หมายถึง ความสามารถของระบบในการอ่านค่าเฉพาะได้ซ้ำๆ หรือ ความสามารถของเครื่องมือวัดที่วัดค่าแต่ละครั้งมีความแตกต่างของค่าที่วัดได้น้อยมาก เมื่อใช้เครื่องมือนั้นไปวัดปริมาณตัวแปรเดิม

ความเที่ยงตรงของการวัดนั้นสามารถหาได้จากการคำนวณหาความผิดพลาดแบบสุ่ม ซึ่งนับว่าเป็นส่วนสำคัญส่วนหนึ่งของการทดลอง ∴ ความผิดพลาดที่กล่าวถึงคือความผิดพลาดแบบสุ่ม

แนวคิดของความเที่ยง ความถูกต้อง ความผิดพลาดแบบสุ่ม และความผิดพลาดระบบในการวัดสามารถแสดงได้โดยการยิงเป้า ดังแสดงในรูป เป้าหมายคือวงกลมกลางเป้า จุดดำคือตำแหน่งที่เป้าถูกยิง



การใช้สถิติในการวิเคราะห์ข้อมูล

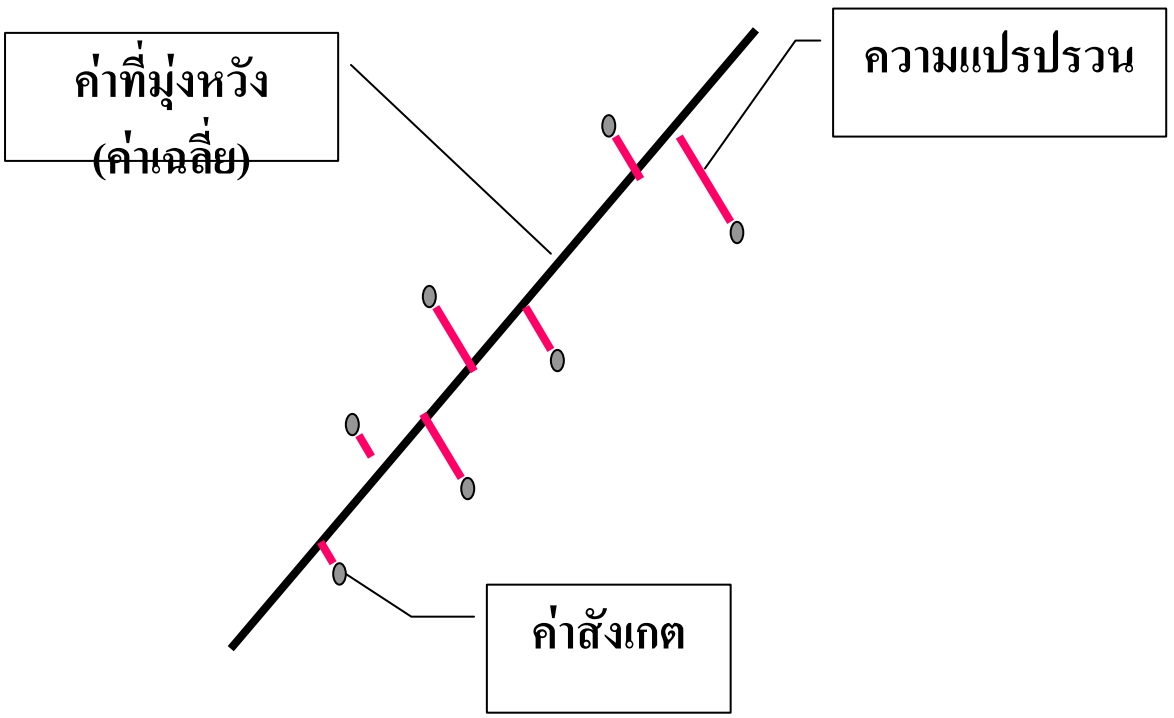
10	10	7
10	9	15
10	11	8
10	12	5
10	8	15
10	10	10

ค่าสถิติ

ค่าสถิติ คือ ตัวเลขที่แสดงลักษณะสำคัญของตัวอย่าง

- ค่าเฉลี่ย (mean) - เป็นค่าที่เป็นตัวแทนของข้อมูล ค่าเฉลี่ยเลขคณิตคำนวณจากการรวมค่าของตัวแปรทั้งหมดหารด้วยจำนวนตัวแปรที่นำมาคิด
- ความผิดพลาดจากการวัด (measurement error) - ค่าความแตกต่างของตัวแปรกับค่าเฉลี่ยหารด้วยค่าที่น้อยกว่าจำนวนตัวแปรทั้งหมดอยู่หนึ่ง
- ความเบี่ยงเบน หรือความแปรปรวน (varience) - ผลบวกกำลังสองของความแตกต่างระหว่างแต่ละค่ากับค่าเฉลี่ยหารด้วยค่าซึ่งน้อยกว่าจำนวนตัวแปรทั้งหมดอยู่หนึ่ง
- ค่าเบี่ยงเบนมาตรฐาน (standard deviation; S.D.) -ปริมาณที่ใช้บอกการกระจายของข้อมูลที่วัดได้รอบค่าเฉลี่ย

ความแปรปรวน



ค่าเฉลี่ยผลรวมกำลังสองของเส้นสีชมพูแต่ละเส้น

ปริมาณโซเดียมในน้ำองุ่น

น้ำองุ่น (mg/20 ml)	$x - \bar{x}$	$(x - \bar{x})^2$
1.43	-0.545	0.297
1.56	-0.415	0.172
2.54	+0.565	0.319
2.61	+0.635	0.403
1.34	-0.635	0.403
2.37	+0.395	0.156
1.975	0	0.35

$$S.D. = \sqrt{\frac{(x - \bar{x})^2}{n - 1}} = \sqrt{0.35} = 0.592 mg / 20ml$$

ความแปรปรวน

$$S^2 = \frac{\sum (x - \bar{x})}{n-1} \text{ or } \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{n-1}$$

$$SS = \sum (x - \bar{x})^2 \text{ or } \sum x^2 - \frac{(\sum x)^2}{N}$$

d.*f. (degree of freedom)* – องศาความเป็นอิสระ คือ โอกาสของตัวแปรที่จะมีค่าเปลี่ยนแปรไปอย่างอิสระ

การใช้สถิติในการตัดสินใจ

ปริมาณไขมัน

กลุ่มที่ 1		กลุ่มที่ 2
17.47%	<	20.03%
S.D. = 1.52		S.D. = 5.20

การทดสอบสมมติฐาน

สมมติฐาน คือคำตอบที่ผู้วิจัยคาดคะเนไว้ล่วงหน้าอย่างมีเหตุผล

สมมติฐานหลัก H_o – *สมมติฐานการทดลองตามวัตถุประสงค์การทดลอง เมื่อกำหนดสมมติฐานหลักแล้ว จะต้องหาหลักฐาน(ผลจากการทดลอง) มาใช้หักล้างสมมติฐานที่กำหนด*

สมมติฐานรอง H_A – *เป็นสมมติฐานที่ถูกกำหนดให้มีข้อความที่ขัดแย้งกับสมมติฐานหลัก*

ความผิดพลาดในการตัดสินใจ

การตัดสินใจทุกครั้งมีความเสี่ยง ถึงแม้จะมีการวิเคราะห์ด้วยสถิติ ตามหลักการวิเคราะห์ที่ถูกต้อง แต่ก็ไม่สามารถรับประกันได้ 100% ว่าการตัดสินใจนั้นถูกต้อง

ความผิดพลาดจากการตัดสินใจ สามารถแบ่งออกได้เป็น 2 ประเภท

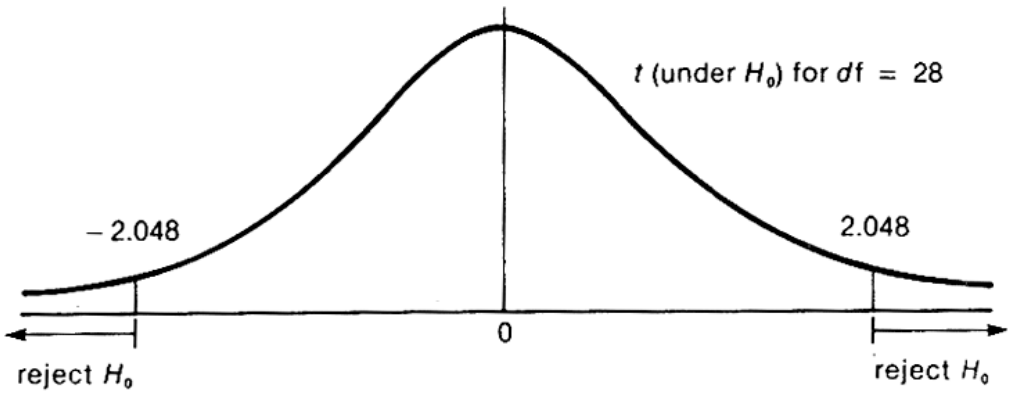
- ความผิดพลาดประเภทที่ 1 (type I error) เป็นความผิดที่เกิดจากการไม่ยอมรับสมมติฐานหลัก ทั้งๆที่สมมติฐานหลักเป็นจริง แทนด้วยสัญลักษณ์ α
- ความผิดพลาดประเภทที่ 2 (type 2 error) เป็นความผิดที่เกิดจากการยอมรับสมมติฐานหลัก ทั้งๆที่สมมติฐานหลักไม่เป็นจริง แทนด้วยสัญลักษณ์ β

สรุปขั้นตอนการทดสอบสมมติฐาน

1. กำหนดสมมติฐานทางสถิติ H_0, H_A
2. เลือกตัวทดสอบสถิติ (ค่า Z, ค่า t or ค่าF)
3. คำนวณหาค่าทดสอบสถิติ
4. กำหนดเกณฑ์การตัดสินใจ ที่จะยอมรับหรือปฏิเสธสมมติหลักตามระดับนัยสำคัญ (α)
5. สรุปผลการทดสอบ

เกณฑ์ในการพิจารณา

1. ปฏิเสธ H_0 ถ้าค่าทดสอบที่คำนวณได้ตกอยู่ในเขตวิกฤตหรือเขตปฏิเสธ



2. ปฏิเสธ H_0 ถ้าค่าความน่าจะเป็นของค่าทดสอบสถิติที่คำนวณได้ p-value มีค่าน้อยกว่าระดับนัยสำคัญที่กำหนด ($p<\alpha$)

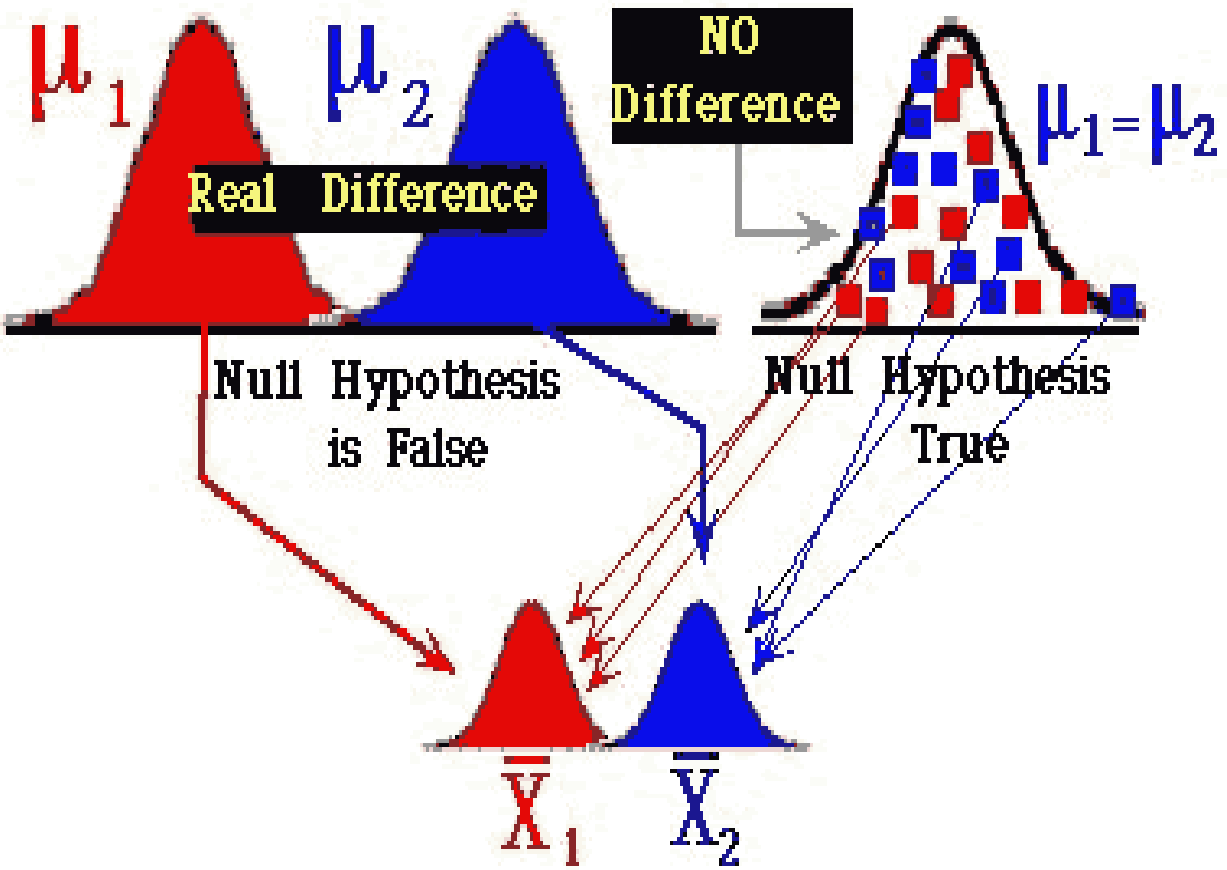
แบบฝึกหัด

จากการศึกษาวิธีผสมอาหาร 4 วิธี ได้ผลการทดลองค่า Tensile Strength (lb/in^2) ดังแสดงในตาราง

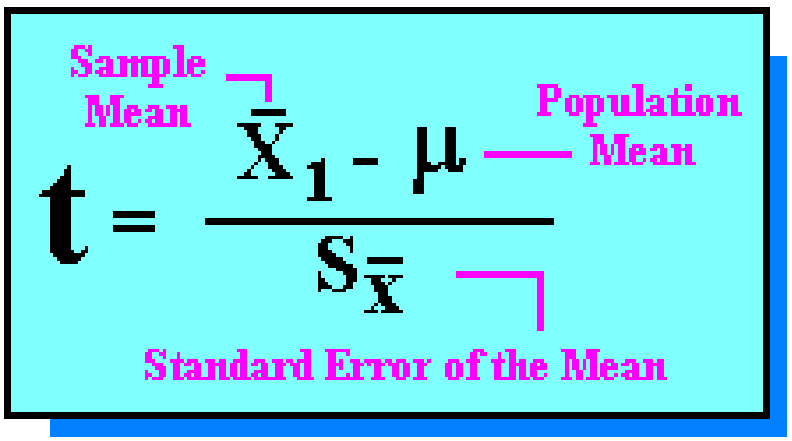
Mixing technique	Tensile Strength (lb/in^2)			
1	3129	3000	2865	2890
2	3200	3300	2975	3150
3	2800	2900	2985	3050
4	2600	2700	2600	2765

ให้นิสิตคำนวณค่า Mean ค่า S^2 และค่า S.D. ของวิธีผสมอาหารทั้ง 4 วิธี

T-test


μ_1
μ_2
Real Difference
NO Difference
$\mu_1 = \mu_2$
Null Hypothesis is False
Null Hypothesis True
$\bar{X}_1$
$\bar{X}_2$

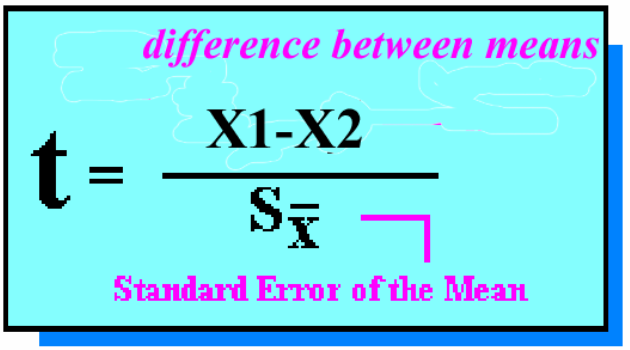
ONE-SAMPLE t TEST



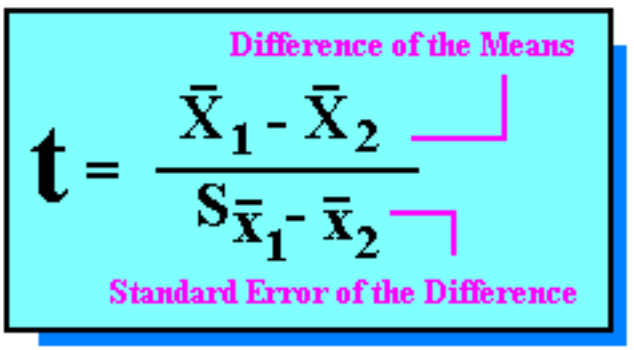
$$t = \frac{\bar{X} - \mu}{S/\sqrt{N}} = \frac{17 - 16}{3.13/\sqrt{10}} = 1.01$$

T-test

Related Samples



Independent Samples



Related Samples

No	Score		d	d^2
	white	red		
1	20	22	-2	4
2	18	19	-1	1
3	19	17	2	4
4	22	18	4	16
5	17	21	-4	16
6	20	23	-3	9
7	19	19	0	0
8	16	20	-4	16
9	21	22	-1	1
10	19	20	-1	1
sum	191	201	-10	68

การคำนวณ

$$S = \sqrt{\frac{\sum d^2 - \frac{\left(\sum d\right)^2}{N}}{n-1}}$$

$$S = \sqrt{\frac{68 - \frac{100}{10}}{9}} = 2.53$$

$$t = \frac{\overline{d}}{S/\sqrt{N}}$$

$$t = \frac{-1}{2.53/\sqrt{10}} = -1.25$$

RCBD; $t_{\text{คำนวณ}} = -1.25$ $t_{0.05} = 2.26$

Independent Samples

	A	B
	3 5 6 4 3 3 4	5 8 9 6 12 9 7
mean	4	8
SS	8	32

การคำนวณ

$$t = \frac{\overline{X} - \overline{Y}}{S_p \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

$$S_p = Pooled\ S\tan dard\ Deviation$$

$$S_p = \sqrt{\frac{SS_X + SS_Y}{N_1 + N_2 - 2}}$$

$$S_p = \sqrt{\frac{8+32}{7+7-2}}$$

$$\therefore t = \frac{4-8}{1.825 \times \sqrt{\frac{1}{7} + \frac{1}{7}}} = -4.10$$

CRD; $t_{คำนวณ} = -4.10$ $t_{0.05} = 2.17$

เปรียบเทียบขนมที่ทำจากแป้ง 2 ชนิด

ซ้ำ	แป้ง A	แป้ง B
1	10	20
2	11	23
3	18	46
4	22	44
5	13	20

CRD; $t_{\text{คำนวณ}} = -2.50$ $t_{0.05} = 2.306$ $t_{0.02} = 2.896$

RCBD; $t_{\text{คำนวณ}} = -3.996$ $t_{0.05} = 2.776$ $t_{0.02} = 3.747$

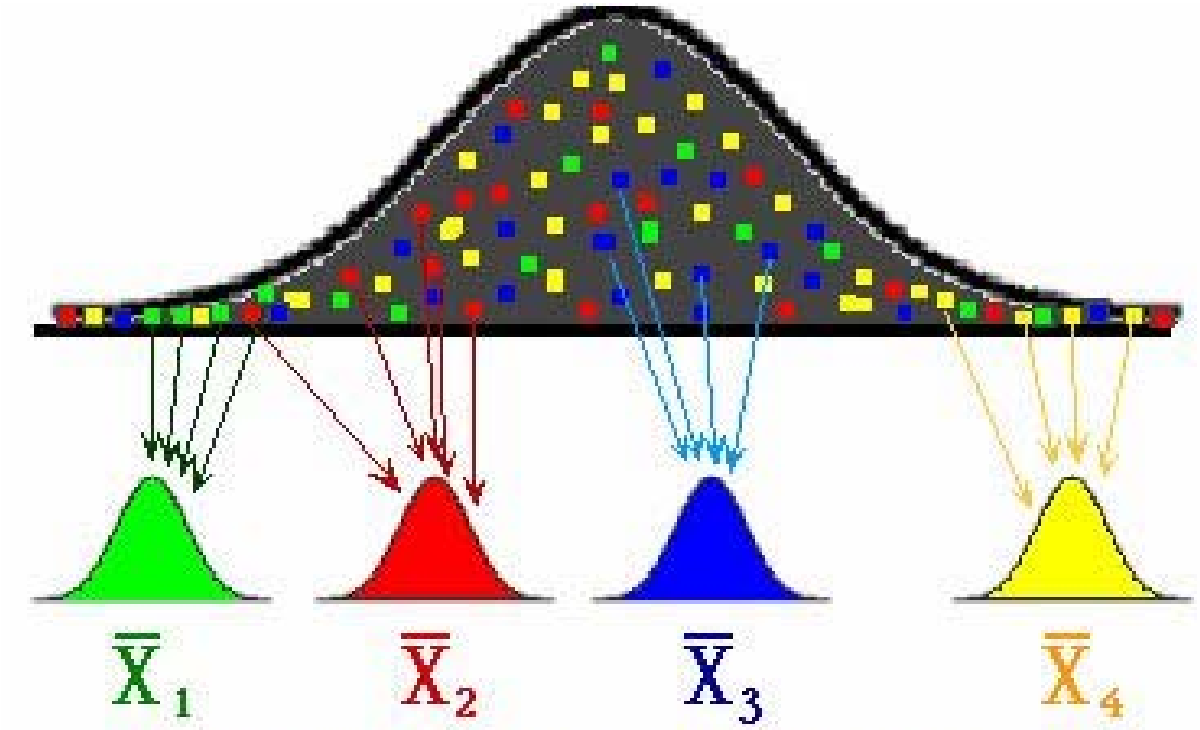
เปรียบเทียบขนมที่ทำจากแป้ง 2 ชนิด

ซ้ำ	แป้ง A	แป้ง B
1	5	10
2	11	23
3	18	46
4	11	22
5	13	20

CRD; $t_{\text{คำนวณ}} = -2.007$ $t_{0.05} = 2.306$

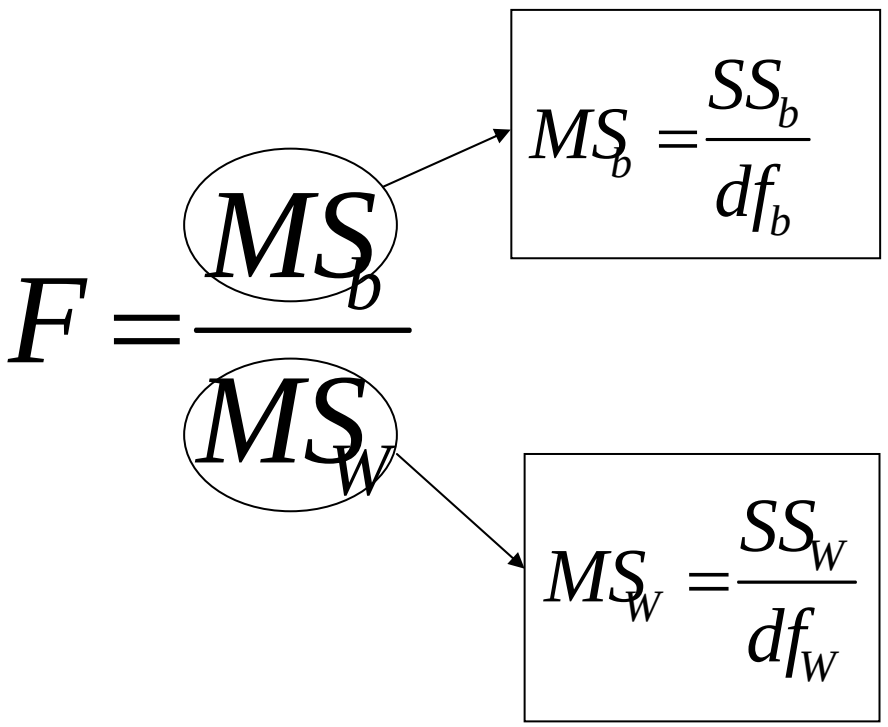
RCBD; $t_{\text{คำนวณ}} = -3.106$ $t_{0.05} = 2.776$

ANOVA



F-test

คือ การวัดความแปรปรวนระหว่างชุดข้อมูลกับความแปรปรวนภายในชุดข้อมูล ถ้าความแปรปรวนภายในชุดข้อมูลมีค่าน้อยกว่าความแปรปรวนระหว่างชุดข้อมูล แสดงว่าชุดข้อมูลมีความแตกต่างกัน



F-test

$$SS_T = SS_b + SS_{E \text{ or } W}$$

$$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}^2 - \frac{T^2}{N}$$

$$SS_b = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{N}$$

$$SS_W = SS_t - SS_b$$

ตารางวิเคราะห์ความแปรปรวน

ที่มาของความแปรปรวน	df	ความแปรปรวน(SS)	ค่าเฉลี่ยความแปรปรวน (MS)	อัตราส่วน*F*
ระหว่างชุดข้อมูล	c-1	SS_b	$MS_b = \frac{SS_b}{df_b}$	$F = \frac{MS_b}{MS_W}$
ภายในชุดข้อมูล	N-c	SS_W	$MS_W = \frac{SS_W}{df_W}$	
รวม	N-1			

ตารางวิเคราะห์ความแปรปรวน

ที่มาของความแปรปรวน	df	ความแปรปรวน(SS)	ค่าเฉลี่ยความแปรปรวน (MS)	อัตราส่วน*F*
ชุดข้อมูล	c-1	SS_b	$MS_b = \frac{SS_b}{df_b}$	$F = \frac{MS_b}{MS_W}$
block	r-1	SSr		
error	N-c-r	SS_W	$MS_W = \frac{SS_W}{df_W}$	
รวม	N-1			

เปรียบเทียบขนมที่ทำจากแป้ง 3 ชนิด

แป้ง A	แป้ง B	แป้ง C	
10	14	8	32
5	10	6	21
14	18	16	48
11	15	11	37
6	9	4	19
46	66	45	157

CRD; $F_{\text{คำนวณ}} = 1.704$ $F_{0.05} = 3.89$

RCBD; $F_{\text{คำนวณ}} = 28.542$ $F_{0.05} = 4.46$

คำนวณ

$$SST = (10^2+5^2+14^2+11^2+6^2+14^2+10^2+18^2 +15^2 +9^2+8^2+6^2+16^2+11^2+4^2)-157^2/15$$

$$= 1897-1643.26=253.74$$

$$SSB = (46^2/5+66^2/5+45^2/5)-157^2/15$$

$$= 1699.4-1643.26=56.14$$

$$SSE = 253.74-56.14 = 197.6$$

F-test

ที่มาของความแปรปรวน	df	ความแปรปรวน(SS)	ค่าเฉลี่ยความแปรปรวน (MS)	อัตราส่วน*F*
ระหว่างชุดข้อมูล	2	56.14	28.07	1.704
ภายในชุดข้อมูล	12	197.6	16.47	
รวม	14			

สรุปสูตร

- ความแปรปรวน

$$S^2 = \frac{\sum (x - \bar{x})^2}{n-1} \text{ or } \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{n-1}$$

- Sum of Square

$$SS = \sum (x - \bar{x})^2 \text{ or } \sum x^2 - \frac{(\sum x)^2}{N}$$

- ส่วนเบี่ยงเบนมาตรฐาน

$$\sqrt{S^2} = \sqrt{\frac{\sum (x - \bar{x})}{n-1}} \text{or } \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{n-1}}$$

- T-test
  - Relate sample t-test

$$t = \frac{\overline{d}}{S/\sqrt{N}}$$

$$S = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{N}}{n-1}}$$

สรุปสูตร

- T-test
  - Independent sample t-test

$$t = \frac{\overline{X} - \overline{Y}}{S/\sqrt{N}}$$

$$t = \frac{\left(\overline{X} - \overline{Y}\right) + \sqrt{\frac{N_1 N_2 \left(N_1 + N_2 - 2\right)}{N_1 + N_2}}}{\sqrt{\sum X^2 - \frac{\left(\sum X\right)^2}{N_1}} + \sqrt{\sum Y^2 - \frac{\left(\sum Y\right)^2}{N_2}}}$$

Or;

$$t = \frac{\overline{X} - \overline{Y}}{S_p \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

$$S_p = Pooled\ S\tan dard\ Deviation$$

$$S_p = \sqrt{\frac{SS_X + SS_Y}{N_1 + N_2 - 2}}$$

สรุปสูตร

- F-test
  - Completely randomized design; CRD

$$SS_T = SS_b + SS_{E \text{ or } W}$$

$$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}^2 - \frac{T^2}{N}$$

$$SS_b = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{N}$$

$$SS_W = SS_t - SS_b$$

สรุปสูตร

- F-test
  - Randomized Complete Block design; RCBD

$$SS_T = SS_b + SS_r + SS_{E \text{ or } W}$$

$$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}^2 - \frac{T^2}{N}$$

$$SS_b = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{N}$$

$$SS_r = \sum_{i=1}^{n}\left(\frac{R_i^2}{n_i}\right) - \frac{T^2}{N}$$

$$SS_W = SS_t - SS_b - SS_r$$