เอกสารประกอบการสอนวิชา
บธ.603 การวิจัยทางธุรกิจ

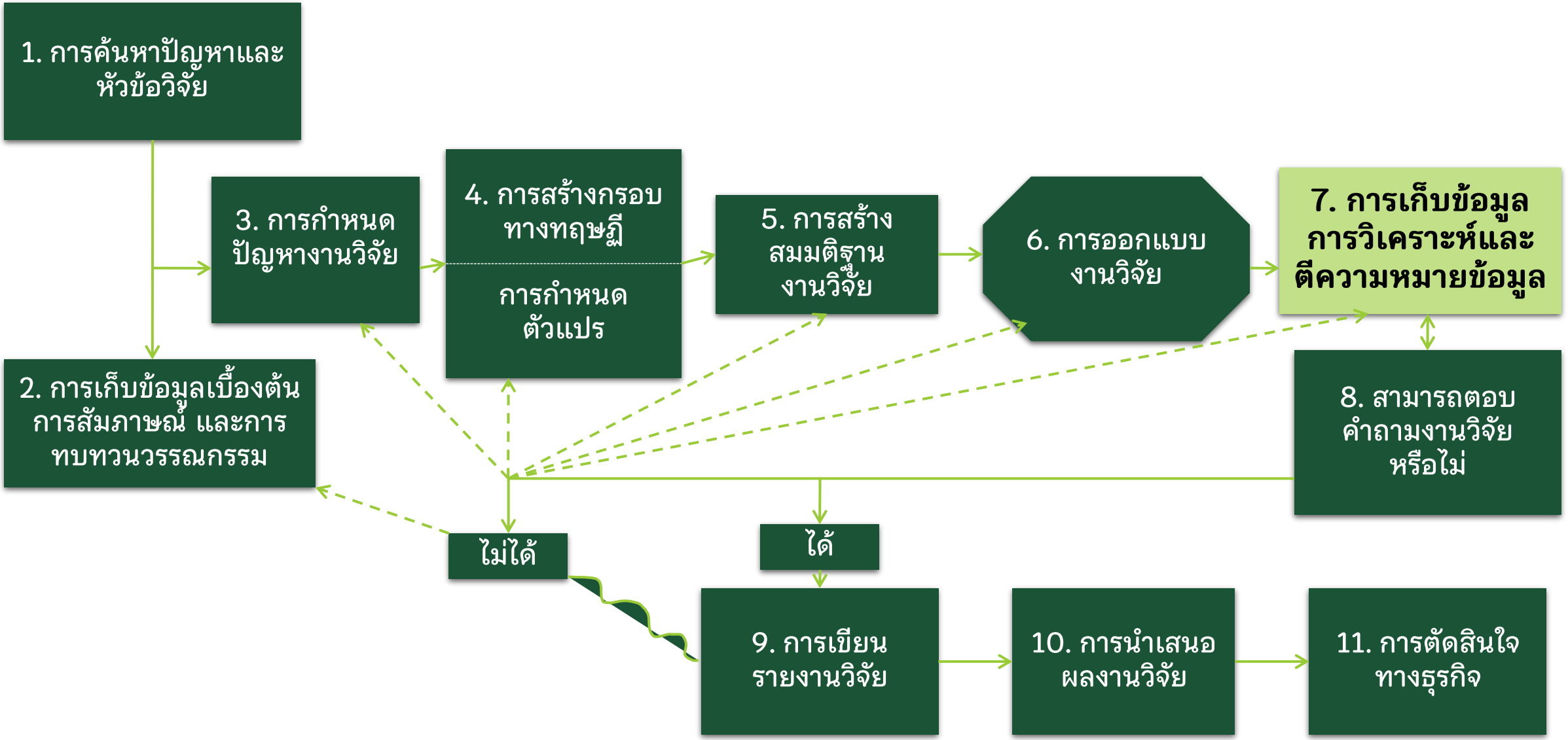
Data analysis and hypothesis testing

BA603 2026

Winai Nadee
Email: winain@staff.tu.ac.th
Website: https://mentor2code.com/

Suwanna Sayruamyat
Email: suwanna.s@ku.th
Facebook: Suwanna Sayruamyat
Page: EatEcon
Website: www.eatecon.com

กระบวนการในการทำงานวิจัย



ประเด็นสำคัญในการวิเคราะห์ข้อมูลและการแปลความหมายงานวิจัย

การเตรียมข้อมูลให้พร้อมในการวิเคราะห์

1. การตรวจสอบแก้ไขข้อมูล
2. การจัดการกับคำตอบที่ว่างเปล่า
3. การลงรหัสข้อมูล
4. การจำแนกประเภทข้อมูล
5. การสร้างไฟล์ข้อมูล
6. การโปรแกรมข้อมูล

จัดการข้อมูล

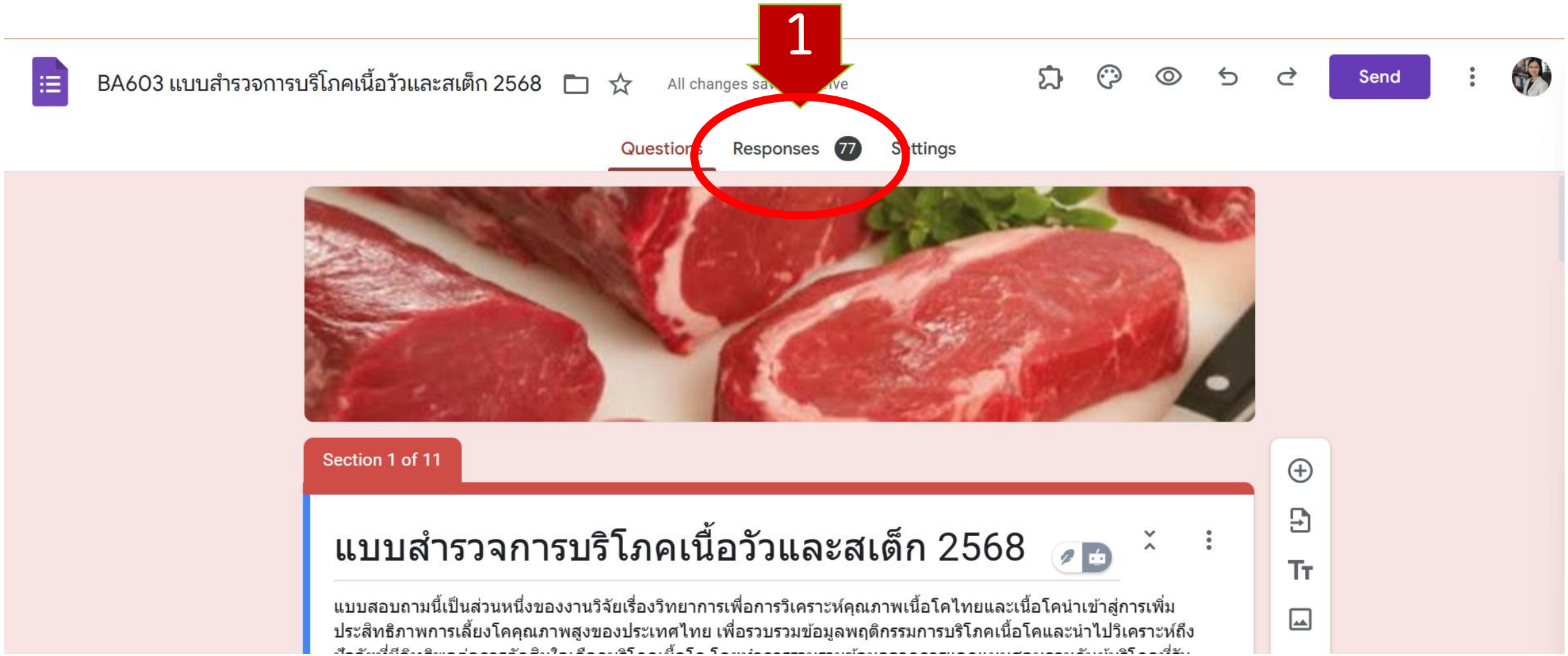
วิเคราะห์ข้อมูล

- **การรับรู้ข้อมูล**
  - การแจกแจงความถี่
  - ค่ากลาง /ค่าเฉลี่ย
  - ค่าเบี่ยงเบนมาตรฐาน
  - ค่าความสัมพันธ์ (Correlation)
- **การทดสอบความถูกต้องของข้อมูล**
  - ความเที่ยงตรง (Validity)
  - ความน่าเชื่อถือ (Reliability)
- **การทดสอบสมมติฐาน**

การใส่ข้อมูล (Entering Data)

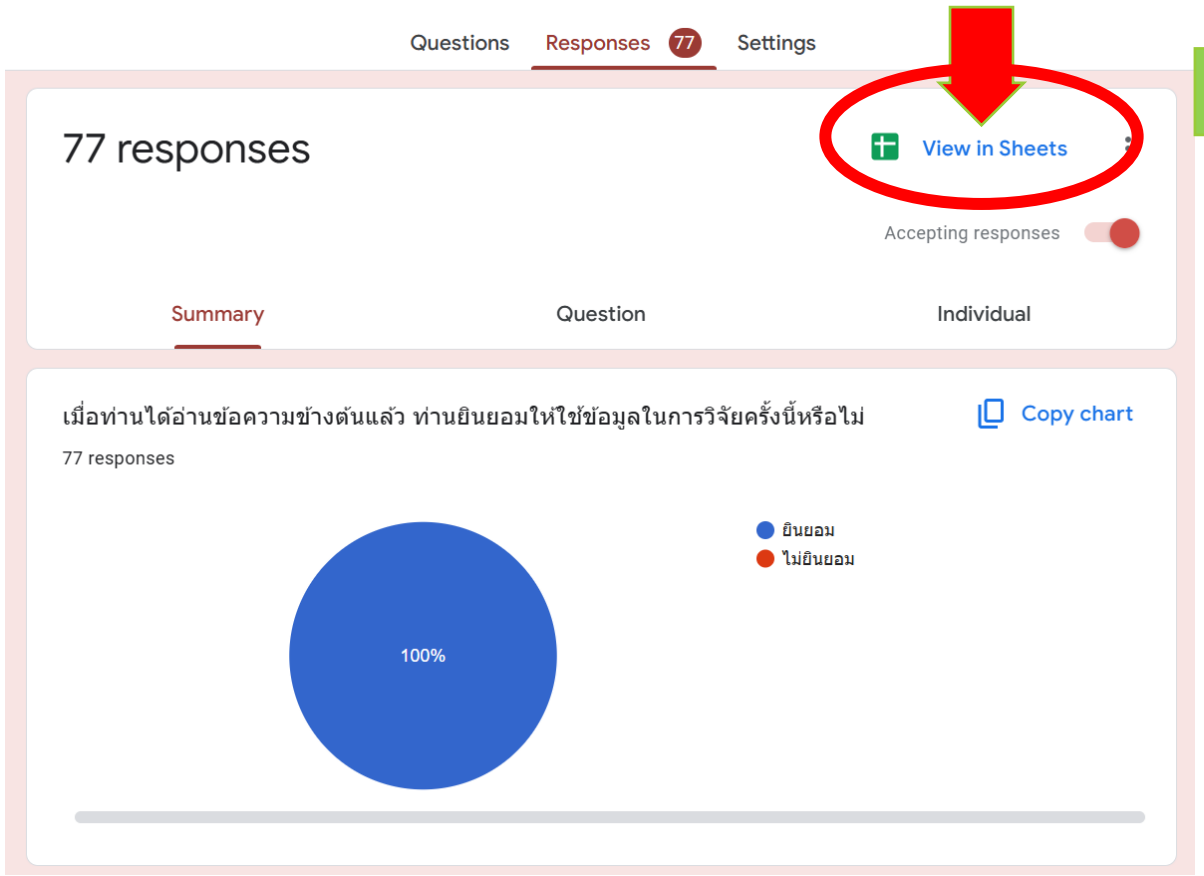
- บันทึกข้อมูลอัตโนมัติจากแบบสอบถามออนไลน์
- Export ข้อมูลจาก Google Form เข้ามาเป็น CSV format เพื่อเปิดด้วย Excel
- แก้ไขคำตอบใน Excel จากตัวอักษรให้เป็นตัวเลขตาม Coding Table
- Copy จาก Excel เข้า SPSS ได้โดยตรง
- กำหนดชื่อตัวแปร และนิยามตัวแปร ตามข้อกำหนดของ SPSS
- ข้อมูลที่ต้องจัดการเป็นพิเศษ
  - ข้อมูลที่เลือกตอบได้มากกว่าหนึ่งตัวเลือกต้องสร้างเป็น Dummy Variable
  - ข้อมูล Categorical หรือ Ordinal ที่ใช้เป็นตัวแปรอิสระและต้องการนำมาคำนวณหาความสัมพันธ์กับตัวแปรตาม ต้องสร้างเป็น Dummy Variable
  - การแปลง Scale ให้กลับด้าน ตามหัวข้อที่แล้ว

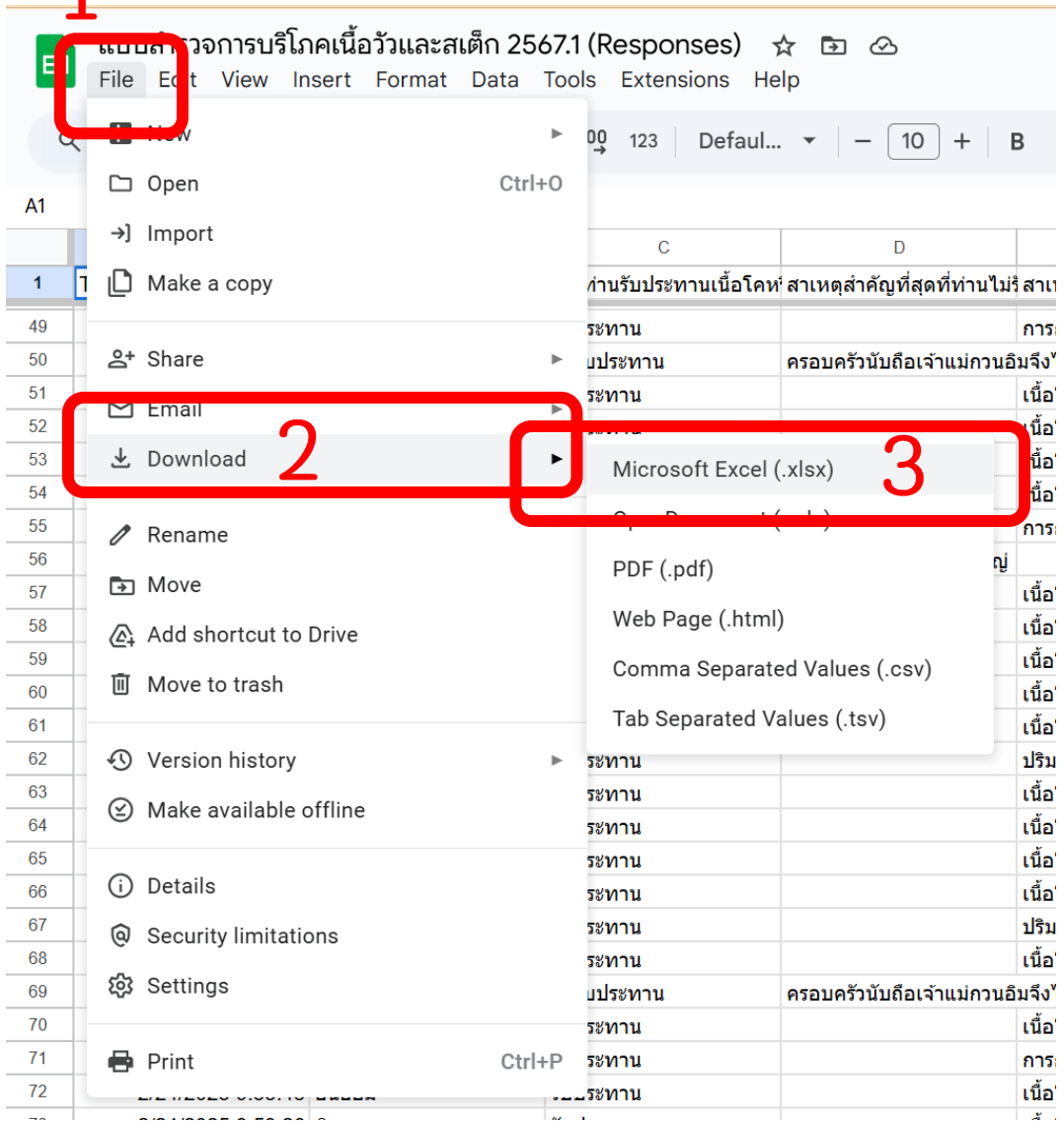
Click: Responses



Click: View in Sheets

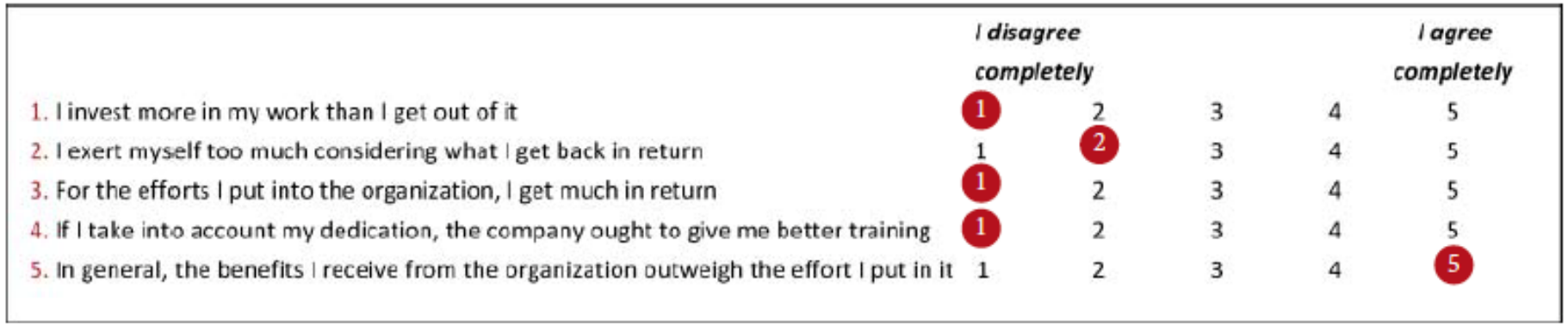
> File > Download > Microsoft Excel (.xlsx)





การตรวจสอบแก้ไขข้อมูล (Editing Data)

- ปัญหาของข้อมูลที่ได้จากการเก็บแบบสอบถาม แบ่งเป็น ข้อมูลที่ไม่สมเหตุสมผล (Illogical) เช่น ข้อมูลจากแบบสอบถาม เลือกไม่รับประทานเนื้อวัว แต่มีข้อมูลการรับประทานเนื้อวัว ผู้วิจัยควรแก้ไขข้อมูลให้สอดคล้องสมเหตุสมผล
- การตอบแบบสอบถามที่ขัดแย้งกันเองในคำตอบ (Inconsistent responses) ดังภาพด้านล่าง ข้อ 1-4 แสดงว่าผู้ตอบไม่พอใจในสภาวะการทำงานและค่าตอบแทน การที่ข้อ 5 ตอบ เห็นด้วยอย่างยิ่ง เป็นคำตอบที่ขัดแย้งกันเอง ผู้วิจัยควรติดต่อสอบถามผู้ตอบ หรือ แก้ไขให้อยู่ในกลุ่มเดียวกัน



	I disagree completely				*I agree completely*
1. I invest more in my work than I get out of it	**1**	2	3	4	5
2. I exert myself too much considering what I get back in return	1	**2**	3	4	5
3. For the efforts I put into the organization, I get much in return	**1**	2	3	4	5
4. If I take into account my dedication, the company ought to give me better training	**1**	2	3	4	5
5. In general, the benefits I receive from the organization outweigh the effort I put in it	1	2	3	4	**5**

การลงรหัสข้อมูล (Coding)

ส่วนที่ 1 พฤติกรรมการบริโภคเนื้อโค

2. B1 ท่านรับประทานเนื้อโคหรือไม่ *

Mark only one oval.

- ◯ รับประทาน *Skip to question 4*
- ◯ ไม่รับประทาน *Skip to question 3*

Variable code	Variable name	Coding
B1	Eating beef	0 = No
		1 = Yes

การลงรหัสข้อมูล (Coding)

สาเหตุสำคัญที่สุดที่ท่านไม่รับประทานเนื้อโค คือ *

Mark only one oval.

- เนื้อโคย่อยยากกว่าเนื้อสัตว์ประเภทอื่น
- ไม่ชอบกลิ่นของเนื้อโค
- ครอบครัวนับถือเจ้าแม่กวนอิมจึงไม่รับประทานเนื้อโค
- เนื้อโคเหนียว
- เนื้อโคมีราคาแพง
- หลักศาสนาที่นับถือห้ามกินเนื้อโค
- ไม่รับประทานเนื้อโคเนื่องจากปัญหาสุขภาพ
- ไม่รับประทานเนื้อโคเพราะสงสาร/ไม่ชอบการทารุณกรรมสัตว์
- ไม่รับประทานเนื้อสัตว์ใหญ่
- ไม่รับประทานเนื้อแดง (Red meat)
- เคยบนต่อสิ่งศักดิ์ไว้ว่าจะเลิกรับประทานเนื้อวัว

Variable code	Variable name	Coding
Reject	เหตุผลที่ไม่รับประทานเนื้อโค	1 = เนื้อโคย่อยยากกว่าเนื้อสัตว์ประเภทอื่น
		2 = ไม่ชอบกลิ่นเนื้อโค
		3 = ครอบครัวนับถือเจ้าแม่กวนอิม
		4 = …
		5 = …
		6 = …
		.
		.
		.
		11 = เคยบนต่อสิ่งศักดิ์สิทธิ์

Coding: คำถามที่ผู้ตอบเลือกได้มากกว่า 1 คำตอบ (Multiple Response)

F4 ประเภทของร้านอาหารที่ท่านไปรับประทานเมนูจากเนื้อโคเป็นประจำ

..... 1) ร้านส้มตำ/อาหารตามสั่ง ..... 2) ร้านก๋วยเตี๋ยว ..... 3) ร้านสเต็ก

..... 4) ร้านปิ้งย่าง ..... 5) ร้านสุกี้/ชาบู/จิ้มจุ่ม ..... 6) ร้านอาหารบุฟเฟต์

..... 7) ร้านอาหารญี่ปุ่น/เกาหลี/จีน ..... 8) อื่น ๆ โปรดระบุ .....

Categorical Variable

1 = ร้านส้มตำ / อาหารตามสั่ง
2 = ร้านก๋วยเตี๋ยว
3 = ร้านเสต็ก
:
:
8 = อื่น ๆ

Dichotomy Variable

ให้สำหรับการวิเคราะห์ทางเศรษฐมิติ กำหนดให้เป็นตัวแปรหุ่น (Dummy) ตามจำนวนทางเลือกที่มี โดย

Resta1	เลือกร้านส้มตำ = 1	ไม่เลือกร้านส้มตำ= 0
Resta2	เลือกร้านก๋วยเตี๋ยว = 1	ไม่เลือกร้านก๋วยเตี๋ยว = 0
Resta3	เลือกร้านสเต็ก = 1	ไม่เลือกร้านสเต็ก = 0
:		
:		
Resta4	เลือกอื่น ๆ = 1	ไม่เลือกอื่น ๆ = 0

Coding: คำถามที่ให้ใส่ลำดับที่ (Rank Question)

F8 เนื้อโคลักษณะใดที่ท่านชอบบริโภค ขอให้ท่านเรียงตามลำดับความชอบ
โดย 1= ชอบมากที่สุด 2= ชอบ และ 3= ชอบน้อยที่สุด
..... เนื้อโคขุนไทยแบบไม่มีไขมันแทรก หรือเนื้อนุ่ม
..... เนื้อโคขุนไทยที่มีไขมันแทรก
..... เนื้อโคนำเข้า

กรณี ให้จัดลำดับน้อยกว่าทางเลือกที่มีอยู่
เป็นการให้ผู้ตอบจัดลำดับน้อยกว่าทางเลือกที่มีอยู่ เช่น มีอยู่ 5 ทางเลือกแต่ให้จัดลำดับแค่ 3 ทางเลือก ซึ่งในกรณีนี้เราจะพบว่าผู้ตอบคำถามอาจจะตอบไม่ครบทั้ง 3 ลำดับก็ได้ ทั้งนี้เราสามารถกำหนดวิธีการกำหนดตัวแปรได้ 2 แบบคือ

1. กำหนดให้มีจำนวนตัวแปร = จำนวนทางเลือก
2. กำหนดให้มีจำนวนตัวแปร = จำนวนลำดับที่ให้จัด

แบบที่ 1

ชอบมากที่สุด	สร้าง Fav1 กำหนดให้	1	= เนื้อโคไทยไม่มีไขมันแทรก
		2	= เนื้อโคไทยมีไขมันแทรก
		3	= เนื้อโคนำเข้า
ชอบอันดับ 2	สร้าง Fav2 กำหนดให้	1	= เนื้อโคไทยไม่มีไขมันแทรก
		2	= เนื้อโคไทยมีไขมันแทรก
		3	= เนื้อโคนำเข้า
ชอบอันดับ 3	สร้าง Fav3 กำหนดให้	1	= เนื้อโคไทยไม่มีไขมันแทรก
		2	= เนื้อโคไทยมีไขมันแทรก
		3	= เนื้อโคนำเข้า

แบบที่ 2

LocBeef1	= 1	ถ้าเลือกเนื้อโคไทยไม่มีไขมันแทรก เป็นลำดับที่ 1
	= 2	ถ้าเลือกเนื้อโคไทยไม่มีไขมันแทรก เป็นลำดับที่ 2
	= 3	ถ้าเลือกเนื้อโคไทยไม่มีไขมันแทรก เป็นลำดับที่ 3
LocBeef2	= 1	ถ้าเลือกเนื้อโคไทยมีไขมันแทรก เป็นลำดับที่ 1
	= 2	ถ้าเลือกเนื้อโคไทยมีไขมันแทรก เป็นลำดับที่ 2
	= 3	ถ้าเลือกเนื้อโคไทยมีไขมันแทรก เป็นลำดับที่ 3
ImBeef	= 1	ถ้าเลือกเนื้อโคนำเข้า เป็นลำดับที่ 1
	2	ถ้าเลือกเนื้อโคนำเข้า เป็นลำดับที่ 2
	3	ถ้าเลือกเนื้อโคนำเข้า เป็นลำดับที่ 3

Coding: คำถามปลายปิด (Closed-end Question)

B10 หากเนื้อตัวอย่างถูกนำบรรจุใส่ในถาดดังรูป ท่านชอบรูปแบบบรรจุภัณฑ์ใด (WTP1)

..... รูปแบบ 1
ถาดธรรมดา

..... รูปแบบ 2
บรรจุถุงสุญญากาศ

..... รูปแบบ 3
Skin pack film

รหัสสำหรับการวิเคราะห์ทางสถิติอย่างง่าย เช่น ความถี่ ร้อยละ สามารถกำหนดให้

- เลือกรูปแบบที่ 1 = 1
- เลือกรูปแบบที่ 2 = 2
- เลือกรูปแบบที่ 3 = 3

รหัสสำหรับการวิเคราะห์ทางเศรษฐมิติ กำหนดให้เป็นตัวแปรหุ่น (Dummy) โดย

Pic1	เลือกรูปแบบที่ 1 = 1	ไม่เลือกรูปแบบที่ 1 = 0
Pic2	เลือกรูปแบบที่ 2 = 1	ไม่เลือกรูปแบบที่ 2 = 0
Pic3	เลือกรูปแบบที่ 3 = 1	ไม่เลือกรูปแบบที่ 3 = 0

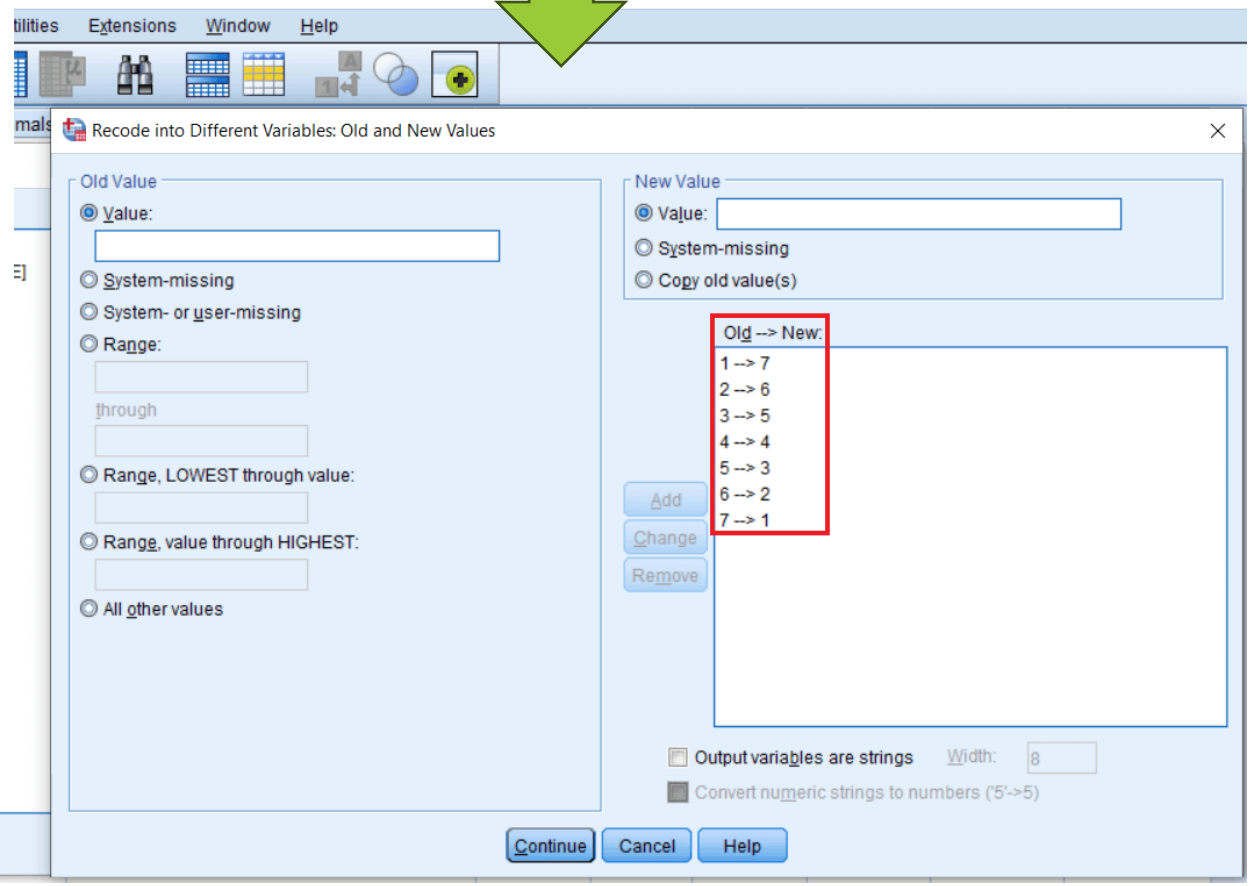
Coding: คำถามที่ให้แสดงระดับความมากน้อย (Scale Questions)

	1	2	3	4	5
รายการ	**ไม่เห็นด้วยอย่างยิ่ง**	**ไม่เห็นด้วย**	**ไม่แน่ใจ/เฉยๆ**	**เห็นด้วย**	**เห็นด้วยอย่างยิ่ง**
1. เนื้อสัตว์อร่อย					
2. เนื้อสัตว์เป็นแหล่งโปรตีนที่สำคัญต่อร่างกาย					
3. การรับประทานเนื้อสัตว์ส่งผลให้น้ำหนักเพิ่มขึ้น					
4. การรับประทานเนื้อสัตว์ส่งผลเสียต่อสุขภาพ					
5. เพื่อลดความเสี่ยงด้านสุขภาพ ท่านสามารถลดการบริโภคเนื้อสัตว์และผลิตภัณฑ์ได้					

แบบสอบถาม

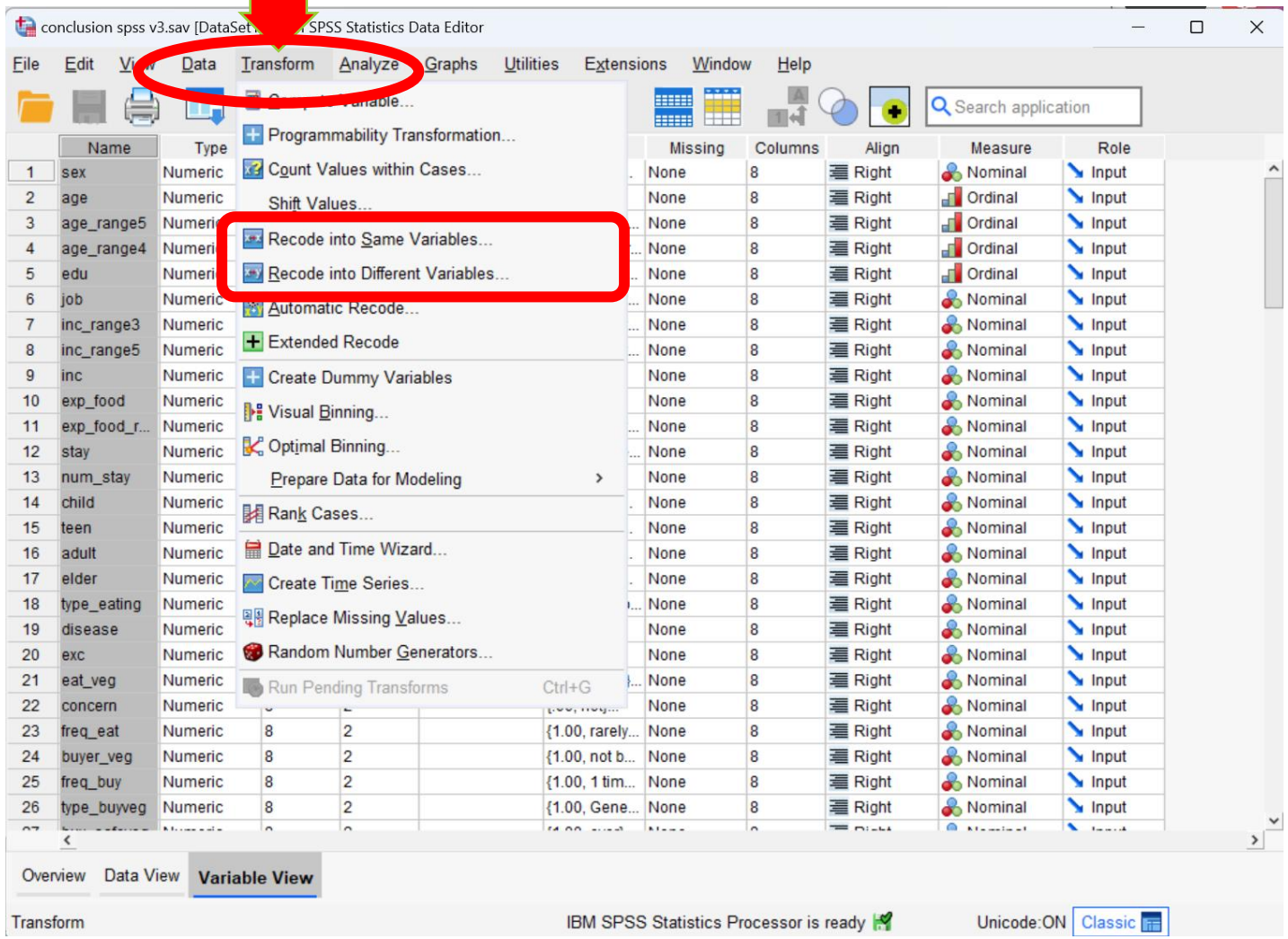
Personality Questionnaire								
1	I often feel sad*	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree
2	I like to take charge of situations and events	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree
3	I experience deep and varied emotions*	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree
4	I am very regular in my approach.	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree
5	I keep my promises	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree
6	I enjoy being center of attention	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree
7	I find it difficult to approach others*	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree
8	I love being in company.	1 Strongly disagree.	2	3	4	5	6	7 Strongly Agree

คำถาม 1, 3, และ 7 ต้องมีการ Recode กลับด้าน



การจำแนกประเภทข้อมูล (Categorization)

- คำตอบที่เป็นการวัดระดับความพึงพอใจ ระดับความชอบ ระดับการเห็นด้วย หรือ คำตอบที่มีลักษณะที่เป็น Scale หากมีการถามในเชิงลบ จะต้องมีการ Recode เพื่อให้สามารถนำมาวิเคราะห์รวมกับคำถามข้ออื่น ๆ ได้
- ใน SPSS มีคำสั่ง Transform / Recode เพื่อใช้ในการเปลี่ยนค่าคำตอบในแบบสอบถาม



คำถามที่ไม่ได้รับคำตอบ (Missing Data)

ในบางครั้งผู้ตอบจะไม่ตอบคำถามทุกข้อในแบบสอบถาม แต่อาจจะมีการละเว้นไม่ตอบคำถามในบางคำถาม โดยเฉพาะคำถามที่ต้องแสดงความรู้สึกในบางเรื่องหรือคำถามที่เกี่ยวกับฐานะ เป็นต้น ซึ่งการไม่ตอบสามารถสรุปได้ว่า มีสาเหตุ ที่สำคัญดังนี้

1) คำถามนั้นไม่ต้องตอบ (Not Applicable)

คำถามในแบบสอบถามบางคำถาม ผู้ตอบไม่ต้องตอบทุกข้อ เช่น ในกรณีเป็นคำถามต่อเนื่อง เช่น

ท่านรับประทานสเต็กเนื้อโคหรือไม่

( ) ไม่รับประทาน (ข้ามไปข้อที่ S5) ( ) รับประทาน

S3 ราคาสเต็กที่ท่านเคยรับประทานแพงที่สุด (เฉพาะตัวท่าน) ..... บาท (เฉพาะเมนูสเต็กเท่านั้น)

ถ้าผู้ตอบไม่เคยรับประทานสเต็ก ก็จะไม่ตอบคำถามข้อS3 จึงทำให้ไม่มีคำตอบในข้อนี้ ผู้วิจัยอาจจะกำหนดรหัสสำหรับผู้ที่ไม่ต้องตอบคำถามในข้อนี้ เป็น 8888 เป็นต้น

2) ไม่ตอบ (No Response)

ผู้ตอบบางคนอาจไม่ตอบคำถามบางคำถามทั้งที่ทราบคำตอบ หรือ อาจจะลืมตอบ เช่น ถ้าถามเรื่องรายได้ แต่ผู้ตอบไม่อยากจะตอบ ก็จะตั้งใจเว้นคำถามข้อนี้ไว้ ผู้วิจัยอาจจะกำหนดรหัสสำหรับผู้ที่ไม่ตอบเป็น 9999 เป็นต้น

การจัดการกับคำตอบว่างเปล่า (Omissions)

- ผู้ตอบไม่ตอบคำถาม อาจเกิดจากการไม่เข้าใจคำถาม หรือไม่ทราบคำตอบ ไม่เต็มใจที่จะตอบ หรือคำตอบอาจเหมือน ๆ ข้อก่อนหน้า เลยไม่อยากตอบ
- หากมีคำตอบว่างเปล่าเกิน 25% ผู้วิจัยไม่ควรนำแบบสอบถามชุดนั้นเข้ามาวิเคราะห์และประมวลผล
- หากคำตอบที่เว้นว่างมีเพียง 2-3 ข้อ จาก 40 ข้อ อาจใช้วิธีดังต่อไปนี้
  - ใช้ค่าตำแหน่งกลางของคำตอบข้อนั้นเป็นคำตอบลงไปแทนที่ เช่น 3 ในกรณีคำตอบมี 5 ตัวเลือก
  - ให้โปรแกรมวิเคราะห์ข้ามข้อมูลชุดนั้น (n ลดลง เฉพาะข้อที่มีเว้นว่าง)
  - ใช้ค่าเฉลี่ยของคำตอบข้อนั้น ของแบบสอบถามทุกชุด มาแทนที่
  - ใช้ค่าเฉลี่ยจากข้ออื่น ๆ ของคนนั้น มาแทนที่
  - นำค่าของคำตอบก่อนหน้าและหลังจากข้อที่คำตอบว่างเปล่า มาบวกกันแล้วหารสอง
- การที่มีแบบสอบถามจำนวนมาก ช่วยลดปัญหาของคำตอบว่างเปล่าได้

สร้างฐานข้อมูล : การจัดทำคู่มือลงรหัส

- โดยทั่วไปผู้วิจัยควรจะจัดทำคู่มือการกำหนดรหัส เพื่อให้ผู้พิมพ์ข้อมูลทุกคนเข้าใจได้ตรงกัน และในกรณีที่มีจำนวนคำถามในแบบสอบถามมาก ๆ ผู้พิมพ์หรือผู้ใส่รหัสอาจจะจำได้ไม่ครบ

	A	B	C	D	E
1	No.	Variable	Description	Code	Note
2	1	gen	เพศ	1	ชาย
3				2	หญิง
4	2	age	อายุ		ปี
5	3	edu	ระดับการศึกษา	1	ประถมศึกษา
6				2	มัธยมศึกษาตอนต้น
7				3	มัธยมศึกษาตอนปลาย
8				4	อนุปริญญา
9				5	ปริญญาตรี/เทียบเท่า
10				6	อื่นๆ
11	4	inc	รายได้		บาท/เดือน
12	5	occ	อาชีพ	1	ข้าราชการ/พนักงานรัฐ/รัฐวิสหกิจ
13				2	ธุรกิจส่วนตัว/อิสระ
14				3	พนักงานลูกจ้างเอกชน
15				4	ว่างงาน
16				5	อื่นๆ
17	6.1	hinf	ข้อมูลด้านสุขภาพ	1	มีโรคประจำตัว

Code +

ส่วนที่ 1 ข้อมูลทั่วไปของผู้ตอบแบบสอบถาม

1. เพศ

☐ ชาย ☐ หญิง

2. อายุ ..... ปี

3. ระดับการศึกษา

☐ ประถมศึกษา ☐ มัธยมศึกษาตอนต้น ☐ มัธยมศึกษาตอนปลาย

☐ อนุปริญญา ☐ ปริญญาตรี/เทียบเท่า ☐ อื่นๆ (โปรดระบุ) .....

4. รายได้ ..... บาทต่อเดือน

5. อาชีพ

☐ ข้าราชการ/พนักงานรัฐ/รัฐวิสาหกิจ ☐ ธุรกิจส่วนตัว/อิสระ

☐ พนักงานเอกชน/ลูกจ้าง ☐ ว่างงาน

☐ อื่นๆ (โปรดระบุ) .....

ฐานข้อมูล

	A	B	C	D	E
1	No.	Variable	Description	Code	Note
2	1	gen	เพศ	1	ชาย
3				2	หญิง
4	2	age	อายุ		ปี
5	3	edu	ระดับการศึกษา	1	ประถมศึกษา
6				2	มัธยมศึกษาตอนต้น
7				3	มัธยมศึกษาตอนปลาย
8				4	อนุปริญญา
9				5	ปริญญาตรี/เทียบเท่า
10				6	อื่นๆ
11	4	inc	รายได้		บาท/เดือน
12	5	occ	อาชีพ	1	ข้าราชการ/พนักงานรัฐ/รัฐวิสหกิจ
13				2	ธุรกิจส่วนตัว/อิสระ
14				3	พนักงานลูกจ้างเอกชน
15				4	ว่างงาน
16				5	อื่นๆ
17	6.1	hinf	ข้อมูลด้านสุขภาพ	1	มีโรคประจำตัว

Code +

	A	B	C	D	E	F	G	H	I
1	ID	gen	age	edu	inc	occ	hinf	risk_occ	veh
2	1	1	35	1	30000	2	1	5	1
3	2	2	25	3	15000	1	2	3	2
4	3	2	18	4	6000	4	1	5	2
5	4	1	40	3	40000	3	1	1	1
6	5	2	36	3	20000	2	2	2	1
7									
8									
9									

วัตถุประสงค์ขั้นพื้นฐานในการวิเคราะห์ข้อมูล

- ความรู้สึกได้ถึงความถูกต้องของข้อมูล เป็นการประเมินว่าสเกล (ชุดคำถาม) ใช้วัดสิ่งที่ต้องการรู้คำตอบได้ดีแค่ไหน การลงรหัสถูกต้องหรือไม่
- ทดสอบความถูกต้องของข้อมูลที่เก็บมาได้ด้วยวิธีทางสถิติ เช่น การหาความเชื่อมั่นจากสัมประสิทธิ์ Cronbach's Alpha การวิเคราะห์องค์ประกอบ (Factor Analysis) เป็นต้น
- การนำเสนอข้อมูลจะต้องคำนึงถึงประเภทของตัวแปร
  - การวัดค่ากลาง เช่น ค่ามัธยฐาน ค่าเฉลี่ย ค่าฐานนิยม
  - การวัดการกระจาย เช่น ค่าเบี่ยงเบนมาตรฐาน ค่าพิสัย
  - การนำเสนอข้อมูลของตัวแปร เช่น กราฟแท่ง กราฟวงกลม ฮิสโตแกรม
  - การนำเสนอความสัมพันธ์ เช่น กราฟแท่งแบบซ้อน กราฟแท่งแบบกลุ่ม ผังการกระจาย
- การทดสอบสมมติฐานตามวัตถุประสงค์การวิจัย โดยใช้เทคนิคทางสถิติที่เหมาะสมกับประเภทของตัวแปร

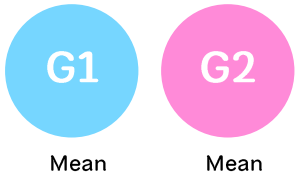
Types of Data Analysis

Descriptive

ความถี่ ร้อยละ
ค่าเฉลี่ย ส่วน
เบี่ยงเบนมาตรฐาน
ความแปรปรวน

Comparative

T-test



Independent t-test
Ho: mean G1 = mean G2
Ha: mean G1 != mean G2

One-way ANOVA

Post Hoc Comparisons
Ho: two groups mean equal
Ha: two groups mean not equal

One-way ANOVA
Ho: all group means equal
Ha: at least one group different

Associative

Correlation

Ho: no linear relationship bet Two variables
Ha: there is linear relationship bet two variables
* Correlation does not imply causation

Crosstabulation

Ho: no relationship bet Two variables
Ha: there is relationship bet two variables

Chi-square test

Predictive

Regression

Ho: all coeff. = 0
Ha: at least one coif. != 0

Structural Equation modelling (SEM)

สถิติเชิงพรรณนา (Descriptive statistics)

เป็นหลักการที่ใช้ในการเก็บรวบรวมข้อมูล นำเสนอข้อมูล และคำนวณค่าสถิติเบื้องต้น ซึ่งเป็นการอธิบายหรือบรรยายลักษณะของข้อมูลที่เก็บรวบรวม แต่จะไม่สามารถอ้างอิงถึงลักษณะประชากรได้ หรืออาจใช้สรุปลักษณะประชากรในกรณีที่ทำการเก็บรวบรวมข้อมูลทั้งหมดของประชากร ดังนั้น สถิติเชิงพรรณนาจึงเป็นการสรุปเฉพาะลักษณะที่สำคัญของข้อมูลของที่ศึกษาเท่านั้น

การนำเสนอข้อมูล

- การนำเสนอข้อมูลในรูปบทความ
- การนำเสนอข้อมูลในรูปตาราง หรือ ร้อยละ
- การนำเสนอข้อมูลในรูปกราฟ

การวัดค่ากลางของข้อมูล

- ค่าเฉลี่ย (Mean)
- มัธยฐาน (Median)
- ฐานนิยม (Mode)

การวัดการกระจายของข้อมูล

- **พิสัย (Range)**
- **พิสัยควอไทล์ (Interquartile Range)**
- **ค่าความแปรปรวน (Variance)**
- **ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)**
- **ค่าความคลาดเคลื่อนมาตรฐานของค่าเฉลี่ย (Standard Error of Mean)**
- **สัมประสิทธิ์ความแปรผัน (Coefficient of Variation)**
- **การวัดความเบ้ (Skewness)**
- **การวัดความโด่ง (Kurtosis)**

วิธีการทางสถิติที่ใช้ในการวิเคราะห์ข้อมูลตามประเภทของตัวแปร

การวิเคราะห์ประเภท	ประเภท	ตัวอย่างวิธีการทางสถิติ
การเปรียบเทียบ	เชิงปริมาณ หรือ เชิงปริมาณ + เชิงคุณภาพ เชิงปริมาณ + เชิงปริมาณ	การทดสอบเกี่ยวกับค่าเฉลี่ย Z-test, t-test, pair-test สหสัมพันธ์ (Correlation)
ความสัมพันธ์	เชิงคุณภาพ + เชิงคุณภาพ เชิงปริมาณ + เชิงปริมาณ เชิงปริมาณ + เชิงปริมาณ	ตาราง Crosstab (Chi-square) วิเคราะห์ความถดถอย (Regression analysis) สหสัมพันธ์ (Correlation)
เหตุผล	เชิงคุณภาพ + เชิงคุณภาพ เชิงปริมาณ + เชิงคุณภาพ	ตาราง Crosstab การวิเคราะห์ความแปรปรวน (ANOVA)

การวิเคราะห์ข้อมูลด้วยวิธีทางสถิติ

1. การทดสอบสมมติฐานค่าเฉลี่ยของตัวอย่างกลุ่มเดียว
2. การทดสอบสมมติฐานค่าเฉลี่ยของตัวอย่างสองกลุ่มที่เป็นอิสระจากกัน
3. การทดสอบสมมติฐานค่าเฉลี่ยของตัวอย่างสองกลุ่มที่ไม่เป็นอิสระจากกัน
4. การทดสอบสมมติฐานค่าเฉลี่ยของตัวอย่างมากกว่าสองกลุ่ม (One-way ANOVA)
5. การทดสอบสมมติฐานความสัมพันธ์ระหว่างตัวแปรเชิงคุณภาพ (Chi-Square)
6. การวิเคราะห์สหสัมพันธ์ (Correlation Analysis)
7. การวิเคราะห์การถดถอย (Regression Analysis)

การทดสอบสมมุติฐานเกี่ยวกับค่าเฉลี่ยของประชากรในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากร

- ถ้าไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากร (σ) เราจะใช้ค่าเบี่ยงเบนมาตรฐานของกลุ่มตัวอย่าง (s) แทน
- เราจะใช้ t-distribution เป็นค่าสถิติทดสอบ

TESTING A MEAN, σ UNKNOWN

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \qquad [10\text{–}2]$$

with $n - 1$ degrees of freedom, where:

$\overline{X}$ is the sample mean.
μ is the hypothesized population mean.
s is the sample standard deviation.
n is the number of observations in the sample.

1. การทดสอบสมมุติฐานเกี่ยวกับค่าเฉลี่ยของประชากรในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากร

- บริษัทประกันแห่งหนึ่งได้รายงานต้นทุนเฉลี่ยในการ claim ประกันต่อครั้งซึ่งมีค่าเท่ากับ $60 อย่างไรก็ตามบริษัทได้มีมาตรการลดต้นทุนใหม่ เพื่อทดสอบมาตรการใหม่นี้ บริษัทได้สุ่มเลือก claim มา 26 claim เพื่อหาต้นทุนของแต่ละ claim โดยได้ค่าตามตารางตามข้างล่างนี้

- ให้ใช้ระดับนัยสำคัญที่ 0.01 เราจะสามารถสรุปได้หรือไม่ว่าต้นทุนของ claim หลังจากมีมาตรการใหม่ต่ำกว่า $60 ต่อ claim

$45	$49	$62	$40	$43	$61
48	53	67	63	78	64
48	54	51	56	63	69
58	51	58	59	56	57
38	76				

ขั้นตอนที่ 1: ตั้ง null hypothesis และ alternate hypothesis

$$H_0: \mu \geq \$60$$
$$H_1: \mu < \$60$$

ขั้นตอนที่ 2: เลือกระดับนัยสำคัญ

$$\alpha = 0.01$$

ขั้นตอนที่ 3: เลือกค่าสถิติทดสอบ

เลือก t-distribution เนื่องจากไม่ทราบ σ

1.การทดสอบสมมุติฐานเกี่ยวกับค่าเฉลี่ยของประชากรในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากร

ขั้นตอนที่ 4: สร้างกฎการตัดสินใจ

ปฏิเสธ H_0 ถ้า $t < -t_{\alpha,n-1}$

$$t < -t_{\alpha,n-1}$$

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} < -t_{\alpha,n-1}$$

$$\frac{\$56.42-\$60}{\$10.04/\sqrt{26}} < -t_{.01,26-1}$$

-1.818 is not < -2.485

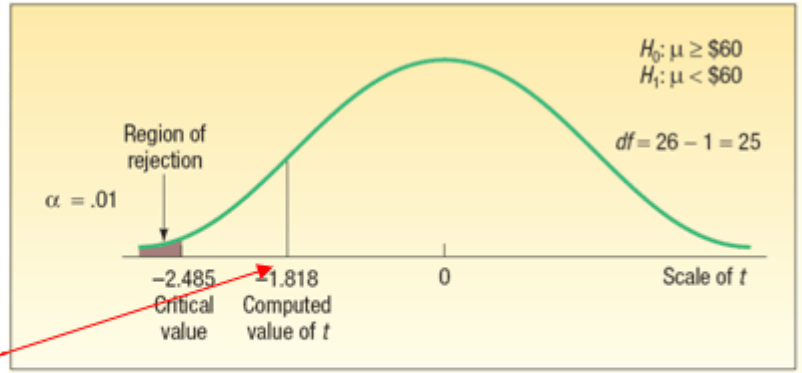


TABLE 10–1 A Portion of the *t* Distribution Table

	Confidence Intervals					
	80%	90%	95%	98%	99%	99.9%
df	Level of Significance for One-Tailed Test, α					
	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646

ขั้นตอนที่ 5: ตัดสินใจและแปลผล

เนื่องจากค่า t ที่คำนวณได้มีค่าเท่ากับ -1.818 ซึ่งอยู่นอกเหนือพื้นที่ในการปฏิเสธสมมุติฐาน เราจึงไม่ปฏิเสธ H_0 ที่ระดับนัยสำคัญที่ .01 เราจึงไม่สามารถสรุปได้ว่ามาตรการใหม่ช่วยลดต้นทุนให้ต่ำกว่า$60 ค่าความแตกต่าง $3.58 ($56.42 - $60) ระหว่างค่าเฉลี่ยของกลุ่มตัวอย่างกับค่าเฉลี่ยประชากรที่ทดสอบอาจเกิดขึ้นจากความผิดพลาดของการสุ่มตัวอย่างเท่านั้น

2.1 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรหรือกลุ่มตัวอย่างมีค่ามากกว่า 30

- กลุ่มตัวอย่างมาจากประชากรสองกลุ่มที่เป็นอิสระต่อกัน
- ใช้ z distribution ถ้าทราบค่าเบี่ยงเบนมาตรฐานของประชากรของทั้งสองกลุ่ม หรือถ้าขนาดของกลุ่มตัวอย่างของแต่ละกลุ่มทั้งสองกลุ่มมีค่ามากกว่า 30

Use if sample sizes $>$ 30 or if σ_1 and σ_2 are known

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Use if sample sizes $>$ 30 and if σ_1 and σ_2 are unknown

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2.1 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรหรือกลุ่มตัวอย่างมีค่ามากกว่า 30

- จากการทดสอบความเร็วในการใช้เทคโนโลยี U-Scan ในการคิดเงินลูกค้าในร้านสะดวกซื้อเมื่อเทียบกับวิธีการตามมาตรฐาน ได้ค่าตามตารางข้างล่าง ผู้จัดการร้านต้องการทราบว่าค่าเฉลี่ยของเวลาคิดเงินโดยใช้วิธีตามมาตรฐานมีค่ามากกว่าการใช้เทคโนโลยี U-Scan หรือไม่ (ให้ใช้ระดับนัยสำคัญที่.01)

ขั้นตอนที่ 1: ตั้ง null hypothesis และ alternate hypothesis

$H_0: \mu_s \leq \mu_u$

$H_1: \mu_s > \mu_u$

ขั้นตอนที่ 2: เลือกระดับนัยสำคัญ

$\alpha = 0.01$

ขั้นตอนที่ 3: เลือกค่าสถิติทดสอบ

เนื่องจากกลุ่มตัวอย่างทั้งสองกลุ่มมีค่ามากกว่า 30 และทราบค่าเบี่ยงเบนมาตรฐานของประชากร จึงใช้ z-distribution

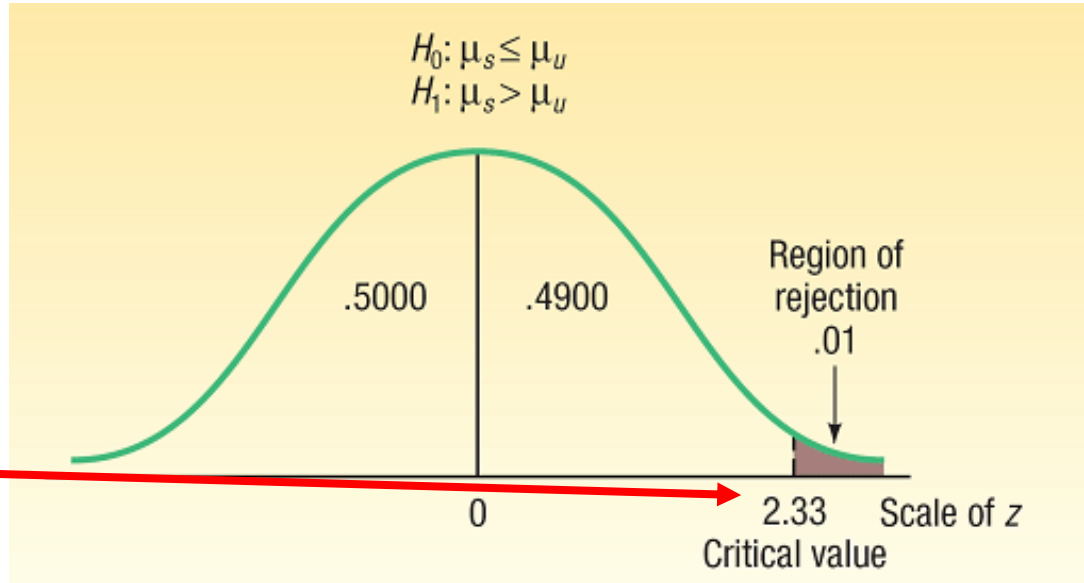
ขั้นตอนที่ 4: สร้างกฎการตัดสินใจ

จะปฏิเสธ H0 ถ้า $Z > Z_\alpha$

Z > 2.33

Z_α คือ ค่าวิกฤติ

Customer Type	Sample Mean	Population Standard Deviation	Sample Size
Standard	5.50 minutes	0.40 minutes	50
U-Scan	5.30 minutes	0.30 minutes	100



2.1 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรหรือกลุ่มตัวอย่างมีค่ามากกว่า 30

ขั้นตอนที่ 5: ตัดสินใจและแปลผล

$$z = \frac{\bar{X}_s - \bar{X}_u}{\sqrt{\frac{\sigma_s^2}{n_s} + \frac{\sigma_u^2}{n_u}}}$$

$$= \frac{5.5 - 5.3}{\sqrt{\frac{0.40^2}{50} + \frac{0.30^2}{100}}}$$

$$= \frac{0.2}{0.064} = 3.13$$

- ค่าที่คำนวณได้มีค่าเท่ากับ 3.13 ซึ่งมากกว่าค่าวิกฤติที่ 2.33 ดังนั้นเราจึงปฏิเสธ null hypothesis สรุปได้ว่ามีหลักฐานที่จะสรุปว่าเวลาเฉลี่ยในการใช้วิธีมาตรฐานนานกว่าเวลาเฉลี่ยในการใช้เทคโนโลยี U-Scan หรือการใช้ U-Scan จะให้ความเร็วสูงกว่าการใช้วิธีมาตรฐาน

2.2 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรและกลุ่มตัวอย่างมีค่าน้อยกว่า 30 (Pooled t-test)

เราจะใช้ t distribution เป็นสถิติทดสอบ ถ้าหนึ่งในกลุ่มตัวอย่างมีจำนวนไม่ถึง 30 โดยมีสมมุติฐานดังต่อไปนี้

1. ประชากรทั้งสองกลุ่มมีการกระจายตัวแบบปกติ
2. ประชากรทั้งสองกลุ่มมีค่าเบี่ยงเบนมาตรฐานเท่ากัน
3. กลุ่มตัวอย่างถูกสุ่มเลือกมาจากประชากรสองกลุ่มที่เป็นอิสระต่อกัน

- แบ่งออกเป็น 2 ขั้นตอนดังต่อไปนี้

1. หาค่าเบี่ยงเบนมาตรฐานรวม

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

2. แทนค่าเบี่ยงเบนมาตรฐานรวมลงในสูตรหาค่า t

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

2.2 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรและกลุ่มตัวอย่างมีค่าน้อยกว่า 30 (Pooled t-test)

- จากการสุ่มทดสอบวิธีในการผลิตสองวิธี (Welles กับ Atkins) ได้ค่าของเวลาที่ใช้ในการผลิตดังแสดงในตาราง โรงงานไม่ทราบค่าเบี่ยงเบนมาตรฐานจากการผลิต แต่คาดว่าการผลิตทั้งสองวิธีมีค่าเบี่ยงเบนมาตรฐานไม่แตกต่างกัน
- จากข้อมูลที่ให้สามารถสรุปได้หรือไม่ว่าค่าเฉลี่ยของเวลาที่ใช้ทั้งสองวิธีนี้มีความแตกต่างกันอย่างมีนัยสำคัญที่ 0.1

Welles (minutes)	Atkins (minutes)
2	3
4	7
9	5
3	8
2	4
	3

ขั้นตอนที่ 1: ตั้ง null hypothesis และ alternate hypothesis

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

ขั้นตอนที่ 2: เลือกระดับนัยสำคัญ

$\alpha = 0.1$ (Confidence Interval = 90%)

ขั้นตอนที่ 3: เลือกค่าสถิติทดสอบ

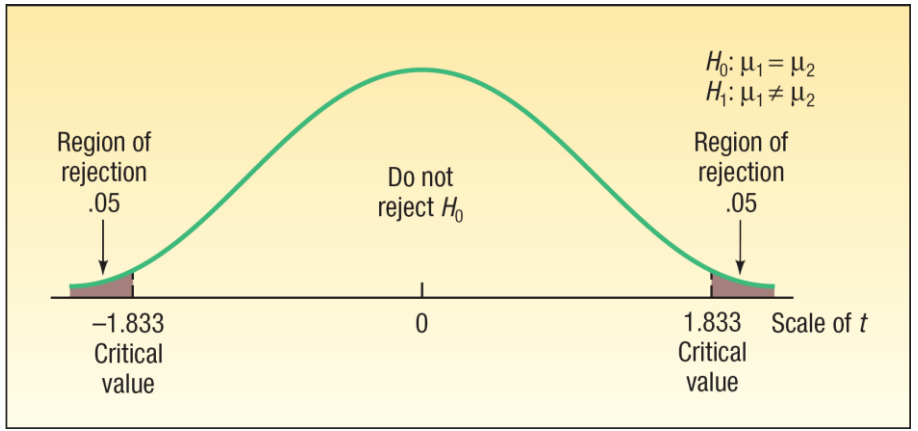
เนื่องจากไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากร เราจึงใช้ pooled t-test.

ขั้นตอนที่ 4: สร้างกฎการตัดสินใจ

จะปฏิเสธ H_0 เมื่อ $t > t_{\frac{\alpha}{2},n_1+n_2-2}$ or $t < -t_{\frac{\alpha}{2},n_1+n_2-2}$

$t > t_{.05,9}$ or $t < -t_{.05,9}$

$t > 1.833$ or $t < -1.833$



2.2 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรและกลุ่มตัวอย่างมีค่าน้อยกว่า 30 (Pooled t-test)

ขั้นตอนที่ 5: ตัดสินใจและแปลผล

Welles Method		Atkins Method	
X_1	$(X_1 - \bar{X}_1)^2$	X_2	$(X_2 - \bar{X}_2)^2$
2	$(2-4)^2 = 4$	3	$(3-5)^2 = 4$
4	$(4-4)^2 = 0$	7	$(7-5)^2 = 4$
9	$(9-4)^2 = 25$	5	$(5-5)^2 = 0$
3	$(3-4)^2 = 1$	8	$(8-5)^2 = 9$
2	$(2-4)^2 = 4$	4	$(4-5)^2 = 1$
20	34	30	22

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{20}{5} = 4 \qquad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{30}{6} = 5$$

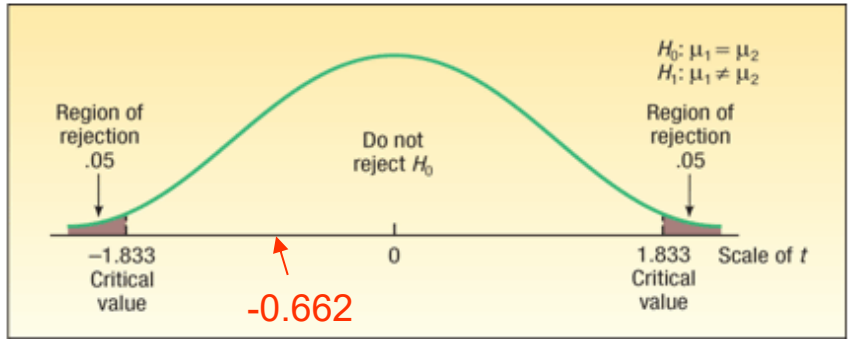
$$s_1 = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1}} = \sqrt{\frac{34}{5-1}} = 2.9155 \qquad s_2 = \sqrt{\frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1}} = \sqrt{\frac{22}{6-1}} = 2.0976$$

(b) Calculate the ***pooled*** sample standard deviation

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(5-1)(2.9155)^2 + (6-1)(2.0976)^2}{5 + 6 - 2} = 6.2222$$

(c) Determine the value of t

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{4.00 - 5.00}{\sqrt{6.2222\left(\frac{1}{5} + \frac{1}{6}\right)}} = -0.662$$



- จากค่าที่ได้ เราจะไม่สามารถปฏิเสธสมมุติฐานได้ เนื่องจากค่า t ที่คำนวณได้อยู่ระหว่างค่าวิกฤติที่ -1.833 และ 1.833.
- เราจึงสรุปว่าไม่มีหลักฐานเพียงพอที่จะสรุปว่ามีความแตกต่างระหว่างค่าเฉลี่ยของเวลาที่ใช้ทั้งสองวิธี

2.3 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรและกลุ่มตัวอย่างมีค่าน้อยกว่า 30 และค่าเบี่ยงเบนมาตรฐานของประชากรแตกต่างกัน

ในการทดสอบความสามารถในการดูดซับของกระดาษชำระที่เป็นยี่ห้อของร้านกับกระดาษชำระที่เป็นยี่ห้ออื่น โดยการสุ่มเลือกกระดาษชำระยี่ห้อของร้านจำนวน 9 แผ่นได้ผลดังต่อไปนี้ (หน่วยเป็นมิลลิลิตรของน้ำที่กระดาษชำระสามารถดูดซับได้)

8 8 3 1 9 7 5 5 12

และได้ทำการสุ่มเลือกกระดาษชำระยี่ห้ออื่นมา 12 แผ่น วัดการดูดซับเป็นมิลลิลิตรของน้ำ ได้ผลดังต่อไปนี้

12 11 10 6 8 9 9 10 11 9 8 10

ให้ใช้ระดับนัยสำคัญที่ .10 เราสามารถสรุปได้หรือไม่ว่าค่าเฉลี่ยของการดูดซับระหว่างกระดาษชำระยี่ห้อของร้านกับยี่ห้ออื่นมีความแตกต่างกัน (กำหนดให้ความแปรปรวนของประชากรแตกต่างกัน)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

2.3 การเปรียบเทียบค่าเฉลี่ยของประชากรสองกลุ่มในกรณีที่ไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรและกลุ่มตัวอย่างมีค่าน้อยกว่า 30 และค่าเบี่ยงเบนมาตรฐานของประชากรแตกต่างกัน

ขั้นตอนที่ 1: ตั้ง null hypothesis และ alternate hypothesis

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

ขั้นตอนที่ 2: เลือกระดับนัยสำคัญ

$\alpha = 0.10$

ขั้นตอนที่ 3: เลือกค่าสถิติทดสอบ

เนื่องจากไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากรและประชากรมีความแปรปรวนที่ต่างกัน เราจึงใช้ unequal variances t-test

ขั้นตอนที่ 4: สร้างกฎการตัดสินใจ

จะปฏิเสธ H_0 ถ้า $t > t_{\alpha/2,d.f.}$ or $t < -t_{\alpha/2,d.f.}$

$t > t_{.05,11}$ or $t < -t_{.05,11}$

$t > 1.796$ or $t < -1.796$

$$= \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{[(3.32^2/9) + (1.621^2/12)]^2}{\frac{(3.32^2/9)^2}{9 - 1} + \frac{(1.621^2/12)^2}{12 - 1}} = \frac{1.4}{.1875}$$

Descriptive Statistics: Store, Name

Variable	N	Mean	StDev
Store	9	6.44	3.32
Name	12	9.417	1.621

$$= \frac{\bar{X}_1 + \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{6.44 - 9.417}{\sqrt{\frac{3.32^2}{9} + \frac{1.621^2}{12}}} = -2.478$$

ขั้นตอนที่ 5: ตัดสินใจและแปลผล

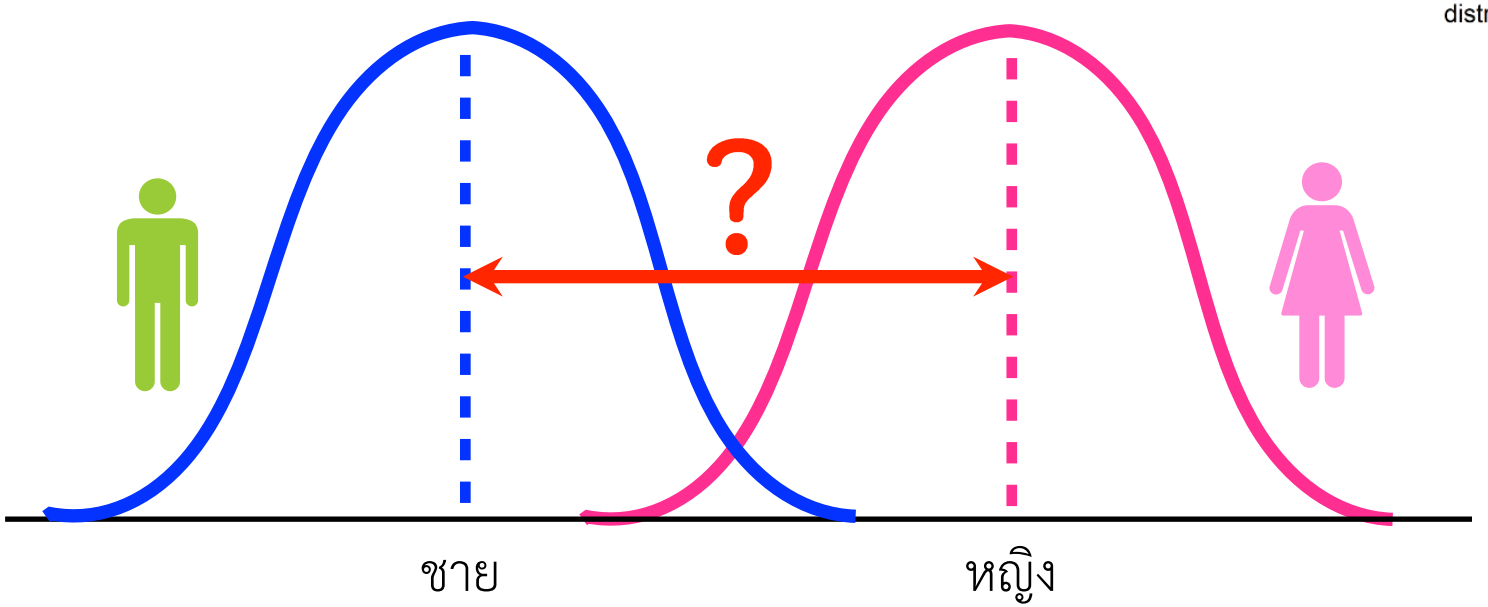
เนื่องจากค่า t ที่คำนวณได้มีค่าน้อยกว่าค่าวิกฤติดังนั้นเราจึงปฏิเสธ null hypothesis และสามารถสรุปได้ว่าค่าเฉลี่ยของการดูดซับน้ำของกระดาษชำระทั้งสองยี่ห้อมีความแตกต่างกัน

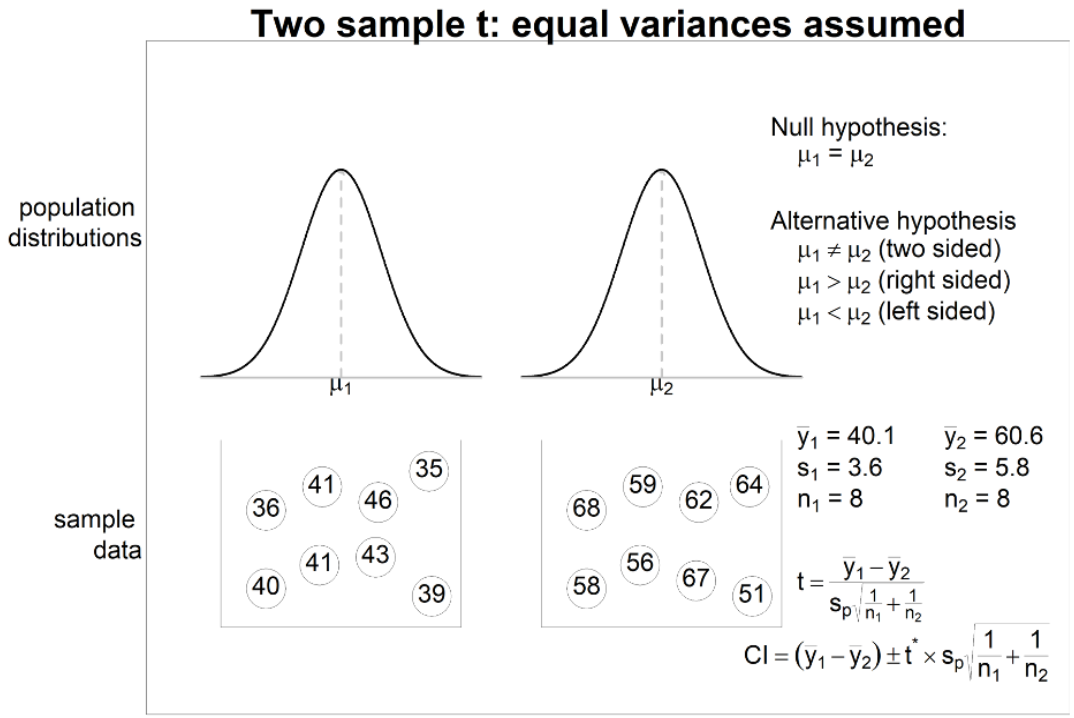
Independent Samples T-Test

2 groups are significantly different from each other on your continuous variable, The variable is **normally distributed** and have a similar spread between your 2 groups. Your **2 groups should be independent (not related to each other)** and you should have enough data (**more than 5 values in each group**).

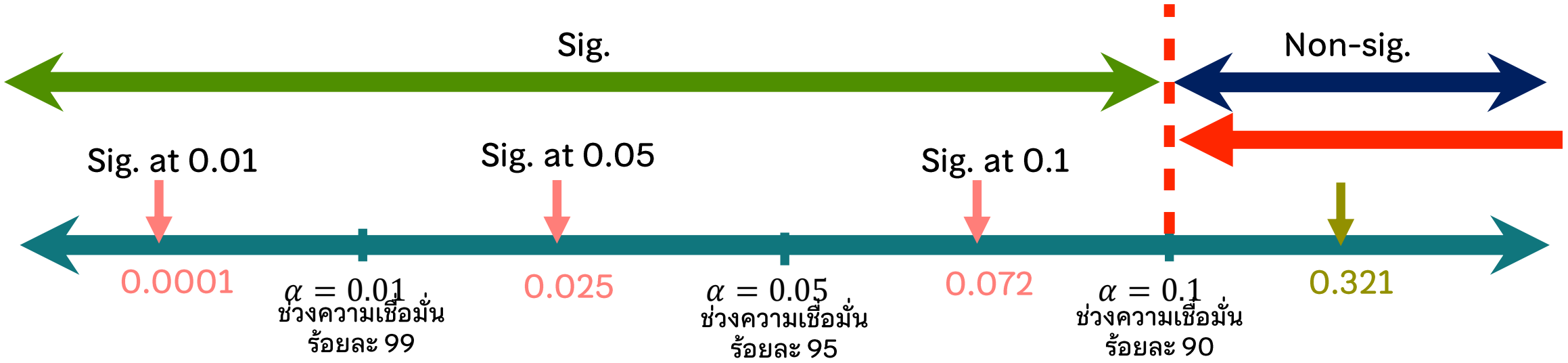
ตัวอย่าง

- รายได้ของเพศชายกับเพศหญิงต่างกันหรือไม่
- รายได้เฉลี่ยของเพศชายกับเพศหญิงเท่ากับ 57,000 บาท หรือไม่
- เพศชายมีความพึงพอใจมากกว่าเพศหญิงหรือไม่





P-value


Sig.
Non-sig.
Sig. at 0.01
Sig. at 0.05
Sig. at 0.1
0.0001
$\alpha = 0.01$
ช่วงความเชื่อมั่น
ร้อยละ 99
0.025
$\alpha = 0.05$
ช่วงความเชื่อมั่น
ร้อยละ 95
0.072
$\alpha = 0.1$
ช่วงความเชื่อมั่น
ร้อยละ 90
0.321

3. การเปรียบเทียบค่าเฉลี่ยสำหรับประชากรสองกลุ่มที่ไม่เป็นอิสระจากกัน

- **กลุ่มตัวอย่างที่ไม่เป็นอิสระจากกัน จะมีความเกี่ยวเนื่องกัน (Dependent samples) คือกลุ่มตัวอย่างที่ถูกจับคู่หรือมีความเกี่ยวข้องกันในทางใดทางหนึ่ง**
- **ตัวอย่าง**
  - การเปรียบเทียบราคารถรุ่นเดียวกันกับตัวแทนจำหน่าย 2 แห่ง
  - การเปรียบเทียบน้ำหนักก่อนทำการลดความอ้วนกับน้ำหนักหลังจากทำการลดความอ้วนแล้วสำหรับคน ๆ เดียวกัน
- **ในกรณีนี้ใช้ค่า t ดังคำนวณได้ตามสูตรดังต่อไปนี้**

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

โดยที่

$\bar{d}$ คือค่าเฉลี่ยของความแตกต่าง

sd คือค่าเบี่ยงเบนมาตรฐานของความแตกต่าง

n คือจำนวนคู่ที่ทดสอบ

3. การเปรียบเทียบค่าเฉลี่ยสำหรับประชากรสองกลุ่มที่ไม่เป็นอิสระจากกัน

- บริษัทได้ทำการเปรียบเทียบราคาบ้านที่ถูกประเมินโดยบริษัทประเมิน 2 แห่ง โดยได้สุ่มเลือกบ้านมาประมาณจำนวน 10 หลัง ได้ผลดังแสดงในตาราง (หน่วยเป็น $1,000) จากข้อมูลดังกล่าว หากใช้ระดับนัยสำคัญที่ 0.05 เราสามารถสรุปได้หรือไม่ว่ามีความแตกต่างระหว่างค่าเฉลี่ยของราคาประเมินของบริษัททั้งสองบริษัทนี้

Home	Schadek	Bowyer
1	235	228
2	210	205
3	231	219
4	242	240
5	205	198
6	230	223
7	231	227
8	210	215
9	225	222
10	249	245

3. การเปรียบเทียบค่าเฉลี่ยสำหรับประชากรสองกลุ่มที่ไม่เป็นอิสระจากกัน

ขั้นตอนที่ 1: ตั้ง null hypothesis และ alternate hypothesis

$$H_0: \mu_d = 0$$
$$H_1: \mu_1 \neq \mu_2$$

ขั้นตอนที่ 2: เลือกระดับนัยสำคัญ

$$\alpha = 0.05$$

ขั้นตอนที่ 3: เลือกค่าสถิติทดสอบ

ใช้ paired t-test

ขั้นตอนที่ 4: สร้างกฎการตัดสินใจ

จะปฏิเสธ H_0 ถ้า $t > t_{\alpha/2,n-1}$ or $t < -t_{\alpha/2,n-1}$

$t > t_{.025,9}$ or $t < -t_{.025,9}$

$t > 2.262$ or $t < -2.262$

TABLE 11–2 A Portion of the *t* Distribution from Appendix B.2

	Confidence Intervals					
	80%	90%	95%	98%	99%	99.9%
	Level of Significance for One-Tailed Test					
df	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of Significance for Two-Tailed Test					
	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587

3. การเปรียบเทียบค่าเฉลี่ยสำหรับประชากรสองกลุ่มที่ไม่เป็นอิสระจากกัน

ขั้นตอนที่ 5: ตัดสินใจและแปลผล

เนื่องจากค่า t ที่คำนวณได้มีค่ามากกว่าค่าวิกฤติ เราจึงปฏิเสธ null hypothesis และสามารถสรุปได้ว่ามีความแตกต่างระหว่างค่าเฉลี่ยของราคาประเมินของทั้งสองบริษัทนี้

Home	Schadek	Bowyer	Difference, d	$(d - \bar{d})$	$(d - \bar{d})^2$
1	235	228	7	2.4	5.76
2	210	205	5	0.4	0.16
3	231	219	12	7.4	54.76
4	242	240	2	−2.6	6.76
5	205	198	7	2.4	5.76
6	230	223	7	2.4	5.76
7	231	227	4	−0.6	0.36
8	210	215	−5	−9.6	92.16
9	225	222	3	−1.6	2.56
10	249	245	4	−0.6	0.36
			46	0	174.40

$$\bar{d} = \frac{\Sigma d}{n} = \frac{46}{10} = 4.60$$

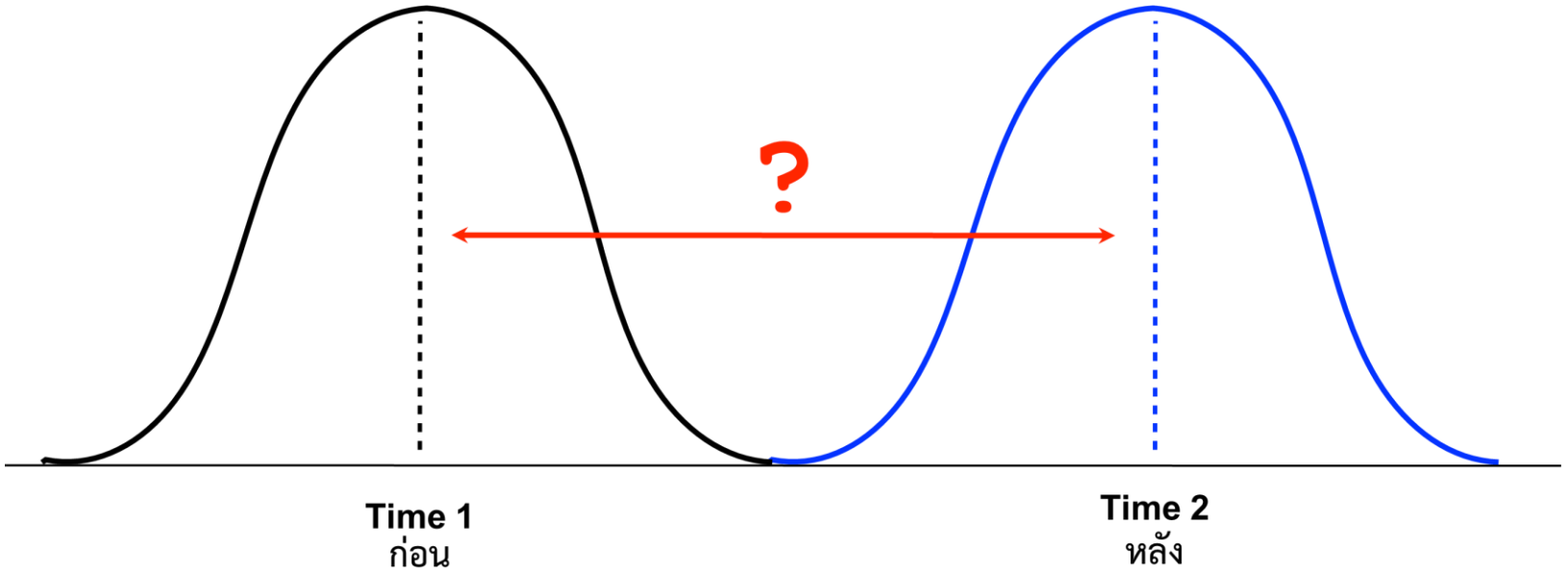
$$s_d = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n - 1}} = \sqrt{\frac{174.4}{10 - 1}} = 4.402$$

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{4.6}{4.402/\sqrt{10}} = 3.305$$

Paired Samples T-Test

The Paired Samples T-Test is also called the Paired Sample T-Test, Dependent Sample T-Test and the Paired T-Test.

Paired t-tests assess **paired observations**, which are often **two measurements on the same person** or item.



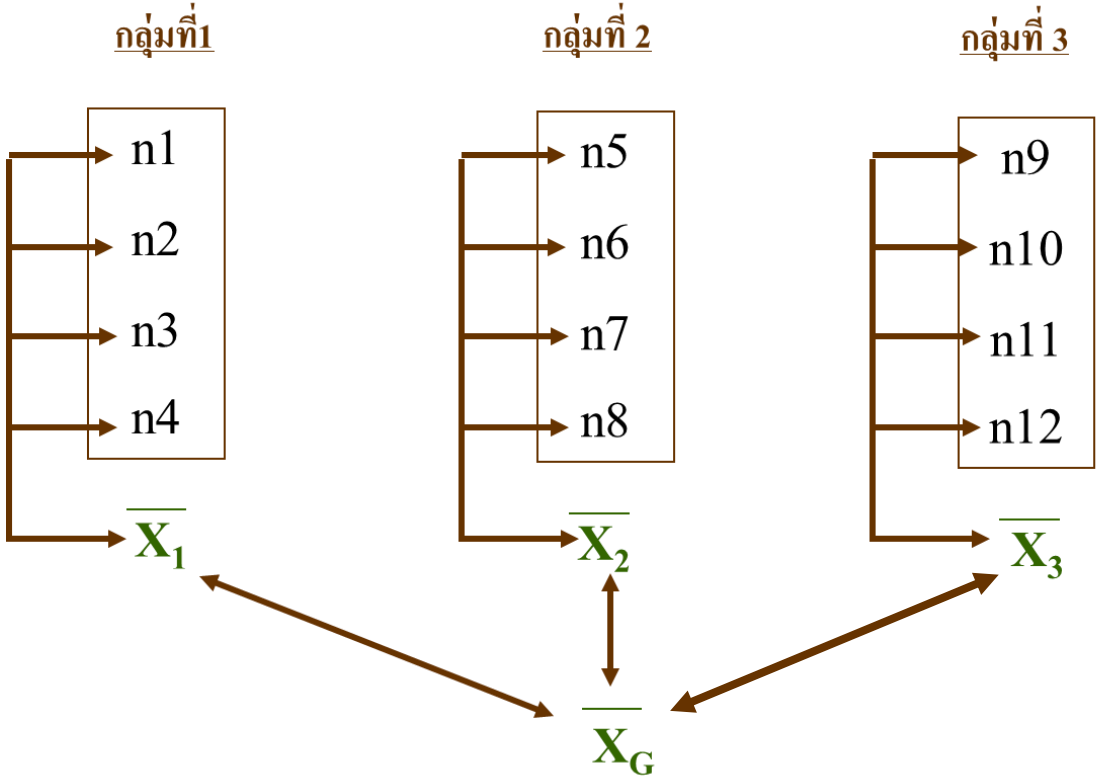
The assumptions for the Paired Samples T-Test include:

1. Continuous
2. Normally Distributed
3. Random Sample
4. Enough Data : more than 5 obs. each is OK. If your sample size is greater than 30 should use Paired Z-test.
5. Similar Spread Between Groups

Note: A Paired Samples T-Test can only be used to compare two groups (i.e. two observations from one group) on your variable of interest.

4.การทดสอบความแตกต่างระหว่างประชากรมากกว่าสองกลุ่มด้วยเทคนิค ANOVA (ANalysis Of VAriance)

- เป็นการทดสอบเพื่อที่จะให้ทราบว่า**ค่าเฉลี่ยจากกลุ่มตัวอย่างหลายกลุ่ม**ที่ได้นั้นมาจากประชากรกลุ่มเดียวกัน (มีค่าเฉลี่ยเท่ากัน) หรือมาจากประชากรคนละกลุ่มกัน
- เราจะ**เปรียบเทียบค่าเฉลี่ยของแต่ละกลุ่มนี้ผ่านทางค่าความแปรปรวน** โดยจะทำการประมาณความแปรปรวนของประชากรโดยใช้สองวิธีและหาอัตราส่วนของความแปรปรวนที่ได้จากสองวิธีดังกล่าว
- ถ้าพบว่าอัตราส่วนนั้นมีค่าประมาณ 1 ก็แสดงว่าการหาความแปรปรวนทั้งสองวิธีนั้นให้ค่าความแปรปรวนที่ใกล้เคียงกัน เราจึงสามารถสรุปได้ว่ากลุ่มตัวอย่างหลายกลุ่มนั้นมาจากประชากรเดียวกัน
- ถ้าค่าอัตราส่วนนั้นแตกต่างจาก 1 ก็แสดงว่าการหาความแปรปรวนโดยใช้ 2 วิธีนั้นให้ค่าที่แตกต่างกัน สรุปได้ว่ากลุ่มตัวอย่างมาจากประชากรที่มีค่าเฉลี่ยที่แตกต่างกัน
- เราจะใช้ **F distribution** ในการตัดสินว่าค่าที่มากกว่า 1 นั้นมีนัยสำคัญทางสถิติหรือไม่



เราต้องการที่จะทราบว่ากลุ่มที่ 1 2 และ 3 มาจากประชากรที่มีค่าเฉลี่ย m เท่ากันหรือไม่

One way ANOVA

An analysis of variance (ANOVA) is used to **compare the means of two or more independent samples** and to test whether the differences between the means are statistically significant. The one-way analysis of Variance (one-way ANOVA) can be thought of as **an extension of a t- test for independent samples.** It is used when there are two or more independent groups.

เปรียบเทียบระหว่างประชากร/หลายกลุ่ม

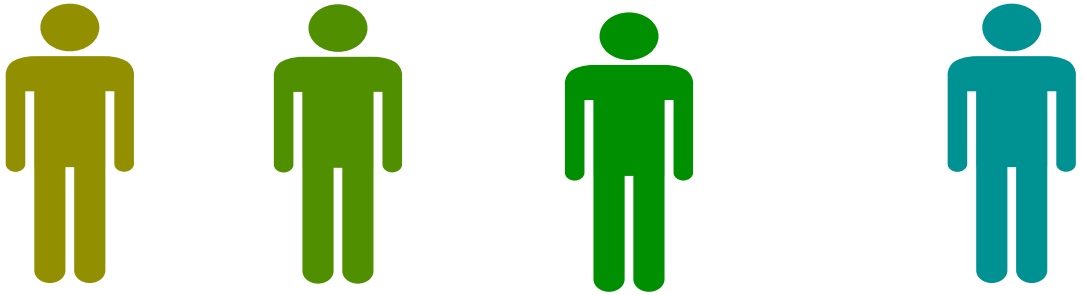
$H_0: \mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$

H_1: *Not all the means are equal*

ค่าใช้จ่ายเฉลี่ยในการบริโภคอาหารแต่ละอาชีพแตกต่างกันหรือไม่

สมมติฐาน

ค่าใช้จ่ายเฉลี่ยในการบริโภคอาหารแต่ละอาชีพแตกต่างกัน



ว่างงาน อาชีพอิสระ พนักงานเอกชน พนักงานภาครัฐ

4.การทดสอบ ANOVA

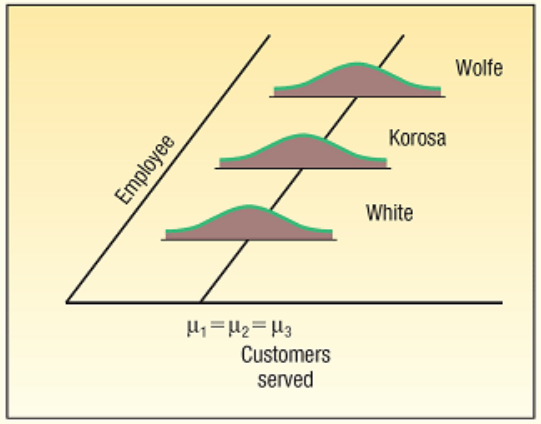
- ความแตกต่างรวม (Total Variation) = ความแตกต่างระหว่างกลุ่ม(Treatment Variation) + ความแตกต่างภายในกลุ่ม (Random Variation)
- ความแตกต่างรวม (Total Variation) = ผลรวมของผลต่างกำลังสองของข้อมูลแต่ละค่ากับค่าเฉลี่ยรวม
- ความแตกต่างระหว่างกลุ่ม (Treatment Variation) = ผลรวมของผลต่างกำลังสองของค่าเฉลี่ยของกลุ่มกับค่าเฉลี่ยรวม
- ความแตกต่างภายในกลุ่ม (Random Variation) = ผลรวมของผลต่างกำลังสองของข้อมูลแต่ละค่ากับค่าเฉลี่ยของกลุ่ม

ผู้จัดการร้านแห่งหนึ่งต้องการที่จะเปรียบเทียบอัตราการทำงานของพนักงาน 3 คน (วัดจากจำนวนลูกค้าที่ให้บริการได้ต่อวัน) เขาจึงทำการสุ่มนับจำนวนลูกค้าในระหว่างการทำงานของพนักงานแต่ละคน ได้ผลดังต่อไปนี้

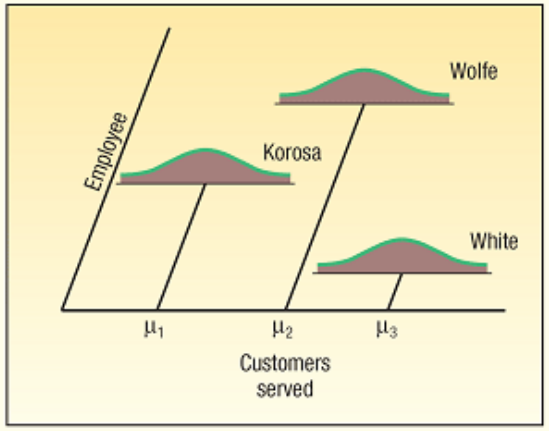
Wolfe	White	Korosa
55	66	47
54	76	51
59	67	46
56	71	48

จากผลที่แสดงดังกล่าว เราสามารถสรุปได้หรือไม่ว่าค่าเฉลี่ยของอัตราการทำงานของแต่ละคนมีความแตกต่างกัน

4.การทดสอบ ANOVA



Case Where Treatment Means Are the Same



Case Where Treatment Means Are Different

- จากตัวอย่าง ค่าเฉลี่ยรวม = (55 + 54 + … + 48)/12 = 58
- ความแตกต่างรวม (Total variation) = (55 – 58)2 + (54 – 58)2 + … + (48 – 58)2 = 1,082
- ค่าเฉลี่ยของกลุ่มที่ 1 = (55 + 54 + 59 + 56)/4 = 56 ค่าเฉลี่ยของกลุ่มที่ 2 และ 3 มีค่าเท่ากับ 70 และ 48 ตามลำดับ
- ความแตกต่างระหว่างกลุ่ม (treatment variation) = (56 – 58)2 + (56 – 58)2 + … + (48 – 58)2 = 4(56 – 58)2 + 4(70 – 58)2 + 4(48 – 58)2 = 992
- ความแตกต่างภายในกลุ่ม (error (random) variation) = (55 – 56)2 + (54 – 56)2 + … + (48 – 48)2 = 90

4.การทดสอบ ANOVA

$$F = \frac{\text{การประมาณค่าความแปรปรวนของประชากรจากความแตกต่างของค่าเฉลี่ยระหว่างกลุ่ม (between group หรือ treatment)}}{\text{การประมาณค่าความแปรปรวนของประชากรจากความแตกต่างของค่าเฉลี่ยภายในกลุ่ม (within group or random หรือ error)}}$$

Degrees of freedom ของค่าสถิติ F ใน ANOVA

ถ้ามีจำนวนประชากรเท่ากับ k กลุ่ม numerator degrees of freedom จะเท่ากับ k – 1

ถ้ามีจำนวนตัวอย่างทั้งหมดเท่ากับ n denominator degrees of freedom จะเท่ากับ n – k

4.การทดสอบ ANOVA

- จากตัวอย่าง การประมาณค่าของความแปรปรวนจากความแตกต่างระหว่างกลุ่ม (treatments หรือ between group) = 992/2 (เราหารด้วย 2 ซึ่งเท่ากับจำนวนของกลุ่มลบด้วย 1)
- การประมาณค่าของความแปรปรวนจากความแตกต่างภายในกลุ่ม (within group หรือ random) = 90/9
- F = (992/2) = 49.6
  (90/9)
- ค่าอัตราส่วนนี้แตกต่างจาก 1 มาก ดังนั้นเราอาจจะมีแนวโน้มที่จะสรุปว่ากลุ่มตัวอย่างทั้ง 3 กลุ่มนี้ไม่ได้ถูกเลือกมาจากประชากรที่มีค่าเฉลี่ยเท่ากัน (หรือไม่ได้ถูกเลือกมาจากประชากรเดียวกันนั่นเอง)

- Null hypothesis คือค่าเฉลี่ยของประชากรทุกกลุ่มมีค่าเท่ากันและ Alternate hypothesis คือมีค่าเฉลี่ยอย่างน้อยหนึ่งค่าที่มีค่าแตกต่างจากค่าอื่น
- ค่าทดสอบสถิติที่ใช้คือค่าการกระจายตัวแบบ F
- กฎการตัดสินใจคือเราจะปฏิเสธ null hypothesis ถ้าค่า F ที่คำนวณได้มีค่ามากกว่าค่า F วิกฤติ ซึ่งจะหาได้จากการเปิดตารางและพิจารณา numerator and denominator degrees of freedom
- การตั้งสมมุติฐาน

  H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

  H_1: มีค่าเฉลี่ยอย่างน้อยหนึ่งค่าที่แตกต่างจากค่าอื่น

  จะปฏิเสธ H_0 ถ้า $F > F_{\alpha,k-1,n-k}$

4.การทดสอบ ANOVA

- ถ้ามีจำนวนกลุ่มประชากรจำนวน k กลุ่ม จะมีค่า numerator degrees of freedom เท่ากับ k – 1.
- ถ้ามีจำนวนกลุ่มตัวอย่างเท่ากับ n ค่า denominator degrees of freedom จะเท่ากับ n – k.
- ค่าสถิติที่ใช้ทดสอบสามารถคำนวณได้โดย

สูตร

$$F = \frac{SST/(k-1)}{SSE/(n-k)}$$

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Treatments	SST	$k-1$	$SST/(k-1) = MST$	MST/MSE
Error	SSE	$n-k$	$SSE/(n-k) = MSE$	
Total	SS total	$n-1$		

$$\text{SS total} = \Sigma(X - \bar{X}_G)^2$$

where:
X is each sample observation.
$\bar{X}_G$ is the overall or grand mean.

$$\text{SSE} = \Sigma(X - \bar{X}_c)^2$$

where:
$\bar{X}_c$ is the sample mean for treatment c.

4.การทดสอบ ANOVA

- จากการเก็บข้อมูลความพึงพอใจลูกค้าของสายการบิน 4 สายการบิน ได้ค่าความพึงพอใจตามตาราง จากข้อมูลของกลุ่มตัวอย่างดังกล่าว เราสามารถสรุปได้หรือไม่ว่าค่าเฉลี่ยของความพึงพอใจของลูกค้ามีความแตกต่างกันในแต่ละสายการบิน โดยใช้ระดับนัยสำคัญที่ 0.01

Eastern	TWA	Allegheny	Ozark
94	75	70	68
90	68	73	70
85	77	76	72
80	83	78	65
	88	80	74
		68	65
		65	

ขั้นตอนที่ 1: ตั้ง null hypothesis และ alternate hypothesis

H_0: $\mu_E = \mu_A = \mu_T = \mu_O$

H_1: มีค่าเฉลี่ยอย่างน้อยหนึ่งค่าที่แตกต่างจากค่าอื่น

ปฏิเสธ H_0 ถ้า $F > F_{\alpha,k-1,n-k}$

ขั้นตอนที่ 2: เลือกระดับนัยสำคัญ

$\alpha = 0.01$

ขั้นตอนที่ 3: เลือกค่าสถิติทดสอบ

ใช้ F-distribution

ขั้นตอนที่ 4: สร้างกฎการตัดสินใจ

ปฏิเสธ H_0 ถ้า $F_{\alpha,k-1,n-k}$

$F_{.01,4-1,22-4}$

$F_{.01,3,18}$

$F > 5.09$

4.การทดสอบ ANOVA

- ขั้นตอนที่ 5: ตัดสินใจและแปลผล

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Treatments	SST	$k-1$	$SST/(k-1) = MST$	MST/MSE
Error	SSE	$n-k$	$SSE/(n-k) = MSE$	
Total	SS total	$n-1$		

$$SS\ total = \Sigma(X - \overline{X}_G)^2$$

where:
X is each sample observation.
$\overline{X}_G$ is the overall or grand mean.

$$SSE = \Sigma(X - \overline{X}_c)^2$$

where:
$\overline{X}_c$ is the sample mean for treatment c.

$$\overline{X}_G = \frac{1{,}664}{22} = 75.64$$

	Eastern	TWA	Allegheny	Ozark	Total
	94	75	70	68	
	90	68	73	70	
	85	77	76	72	
	80	83	78	65	
		88	80	74	
			68	65	
			65		
Column total	349	391	510	414	1,664
n	4	5	7	6	22
Mean	87.25	78.20	72.86	69.00	75.64

4.การทดสอบ ANOVA คำนวณค่า SS Total และ SSE

$(X - \bar{X}_G)^2$

Eastern	TWA	Allegheny	Ozark
18.36	−0.64	−5.64	−7.64
14.36	−7.64	−2.64	−5.64
9.36	1.36	0.36	−3.64
4.36	7.36	2.36	−10.64
	12.36	4.36	−1.64
		−7.64	−10.64
		−10.64	

$(X - \bar{X}_G)^2$

	Eastern	TWA	Allegheny	Ozark	Total
	337.09	0.41	31.81	58.37	
	206.21	58.37	6.97	31.81	
	87.61	1.85	0.13	13.25	
	19.0	54.17	5.57	113.21	
		152.77	19.01	2.69	
			58.37	113.21	
			113.21		
Total	649.91	267.57	235.07	332.54	1,485.09

SS total = $\Sigma(X - \bar{X}_G)^2$

$(X - \bar{X}_c)$

Eastern	TWA	Allegheny	Ozark
6.75	−3.2	−2.86	−1
2.75	−10.2	0.14	1
−2.25	−1.2	3.14	3
−7.25	4.8	5.14	−4
	9.8	7.14	5
		−4.86	−4
		−7.86	

$(X - \bar{X}_c)^2$

	Eastern	TWA	Allegheny	Ozark	Total
	45.5625	10.24	8.18	1	
	7.5625	104.04	0.02	1	
	5.0625	1.44	9.86	9	
	52.5625	23.04	26.42	16	
		96.04	50.98	25	
			23.62	16	
			61.78		
Total	110.7500	234.80	180.86	68	594.41

SSE = $\Sigma(X - \bar{X}_c)^2$

SST = SS total − SSE = 1,485.09 − 594.41 = 890.68.

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Treatments	SST	$k - 1$	SST/$(k - 1)$ = MST	MST/MSE
Error	SSE	$n - k$	SSE/$(n - k)$ = MSE	
Total	SS total	$n - 1$		

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Treatments	890.68	3	296.89	8.99
Error	594.41	18	33.02	
Total	1,485.09	21		

ค่า F ที่คำนวณได้มีค่าเท่ากับ 8.99 ซึ่งสูงกว่าค่า F วิกฤติที่ 5.09 ดังนั้น เราจึงปฏิเสธ Null Hypothesis และสรุปว่ามีค่าเฉลี่ยอย่างน้อย 1 ค่าที่มีความแตกต่างจากค่าอื่น

5. การทดสอบสมมติฐานความสัมพันธ์ระหว่างตัวแปรเชิงคุณภาพ (Chi Square)

- ตาราง contingency เป็นตารางที่ใช้ในการศึกษาความสัมพันธ์ของตัวแปรสองตัว เราจะใช้สมมุติฐานในการวิเคราะห์หาความสัมพันธ์โดยมีการคำนวณดังต่อไปนี้
  - ค่า degrees of freedom มีค่าเท่ากับ (จำนวนแถว-1)(จำนวนคอลัมน์-1)

EXPECTED FREQUENCY

$$f_e = \frac{(\text{Row total})(\text{Column total})}{\text{Grand total}}$$

Residence after Release from Prison	Adjustment to Civilian Life				
	Outstanding	Good	Fair	Unsatisfactory	Total
Hometown	27	35	33	25	120
Not hometown	13	15	27	25	80
Total	40	50	60	50	200

5. การทดสอบสมมติฐานความสัมพันธ์ระหว่างตัวแปรเชิงคุณภาพ (Chi Square)

- จากการศึกษาในเรื่องความสัมพันธ์ระหว่างการปรับตัวเข้าสู่สังคมกับผู้ที่เพิ่งพ้นโทษกับการกลับสู่บ้านเกิด ได้มีการสุ่มถามผู้ที่เพิ่งพ้นโทษจำนวน 200 คน ได้ข้อมูลตามตาราง จากข้อมูลของกลุ่มตัวอย่างดังกล่าว เราสามารถสรุปได้หรือไม่ว่ามีความสัมพันธ์กันระหว่างระดับการปรับตัวเข้าสู่สังคมกับการกลับสู่บ้านเกิด ให้ใช้ระดับนัยสำคัญที่ 0.01

Residence after Release from Prison	Adjustment to Civilian Life			
	Outstanding	Good	Fair	Unsatisfactory
Hometown	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 III	𝍸 𝍸 𝍸 𝍸 𝍸
Not hometown	𝍸 𝍸 III	𝍸 𝍸 𝍸	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II	𝍸 𝍸 𝍸 𝍸 𝍸

Residence after Release from Prison	Adjustment to Civilian Life				
	Outstanding	Good	Fair	Unsatisfactory	Total
Hometown	27	35	33	25	120
Not hometown	13	15	27	25	80
Total	40	50	60	50	200

ขั้นตอนที่ 1: ตั้ง null hypothesis และ alternate hypothesis

H_0: ไม่มีความสัมพันธ์กันระหว่างการปรับตัวเข้าสู่สังคมกับการกลับสู่บ้านเกิด

H_1: มีความสัมพันธ์กันระหว่างการปรับตัวเข้าสู่สังคมกับการกลับสู่บ้านเกิด

ขั้นตอนที่ 2: เลือกระดับนัยสำคัญ

$\alpha = 0.01$

ขั้นตอนที่ 3: เลือกค่าสถิติทดสอบ

ใช้ $\chi 2$

CHI-SQUARE TEST STATISTIC $\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right]$

5. การทดสอบสมมติฐานความสัมพันธ์ระหว่างตัวแปรเชิงคุณภาพ (Chi Square)

- ขั้นตอนที่ 4: สร้างกฎการตัดสินใจ

Reject H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{\alpha,(2-1)(4-1)}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.01,(1)(3)}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > \chi^2_{.01,3}$$

$$\sum\left[\frac{(f_o - f_e)^2}{f_e}\right] > 11.345$$

- การคำนวณหาความถี่ที่คาดหวัง (f_e)

EXPECTED FREQUENCY $f_e = \dfrac{\text{(Row total)(Column total)}}{\text{Grand total}}$

(120)(50) / 200

Residence after Release from Prison	Adjustment to Civilian Life: Outstanding f_o	Outstanding f_e	Good f_o	Good f_e	Fair f_o	Fair f_e	Unsatisfactory f_o	Unsatisfactory f_e	Total f_o	Total f_e
Hometown	27	24	35	30	33	36	25	30	120	120
Not hometown	13	16	15	20	27	24	25	20	80	80
Total	40	40	50	50	60	60	50	50	200	200

Must be equal

(80)(50) / 200

Must be equal

5. การทดสอบสมมติฐานความสัมพันธ์ระหว่างตัวแปรเชิงคุณภาพ (Chi Square)

การคำนวณหาค่าสถิติ Chi-square

Residence after Release from Prison	Adjustment to Civilian Life: Outstanding f_o	Outstanding f_e	Good f_o	Good f_e	Fair f_o	Fair f_e	Unsatisfactory f_o	Unsatisfactory f_e	Total f_o	Total f_e
Hometown	27	24	35	30	33	36	25	30	120	120
Not hometown	13	16	15	20	27	24	25	20	80	80
Total	40	40	50	50	60	60	50	50	200	200

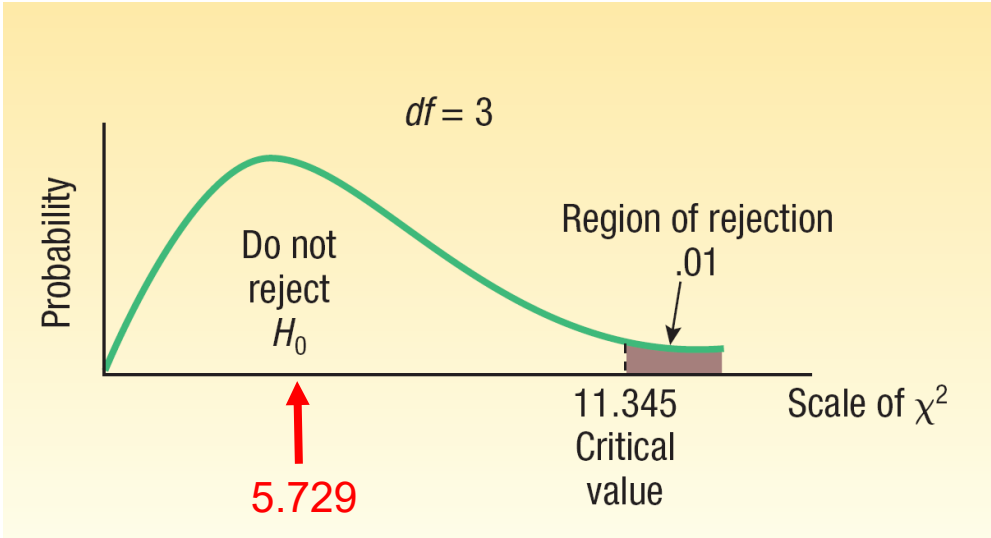
$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

Starting with the upper left cell:

$$\chi^2 = \frac{(27-24)^2}{24} + \frac{(35-30)^2}{30} + \frac{(33-36)^2}{36} + \frac{(25-30)^2}{30} + \frac{(13-16)^2}{16} + \frac{(15-20)^2}{20} + \frac{(27-24)^2}{24} + \frac{(25-20)^2}{20}$$

$$= 0.375 + 0.833 + 0.250 + 0.833 + 0.563 + 1.250 + 0.375 + 1.250$$

$$= 5.729$$



ค่า $\chi 2$ ที่คำนวณได้มีค่าเท่ากับ 5.729 ซึ่งอยู่ในพื้นที่ที่ไม่ปฏิเสธสมมุติฐาน เราจึงไม่ปฏิเสธ null hypothesis ในระดับนัยสำคัญที่ .01 และสรุปได้ว่าไม่มีความสัมพันธ์กันระหว่างการปรับตัวเข้าสู่สังคมกับการกลับสู่บ้านเกิดของผู้ที่เพิ่งพ้นโทษ

Chi-Square Test for Independence

The **Chi Square** statistic is commonly used for testing **relationships between categorical variables**. The null **hypothesis** of the **Chi-Square** test is that no relationship exists on the categorical variables in the population; they are independent.

H_0 : Variable A is independent of variable B
H_1 : Variable A is not independent of variable B

H_0 : There is no relationship between the variables.
H_1 : There is a relationship between the variables.

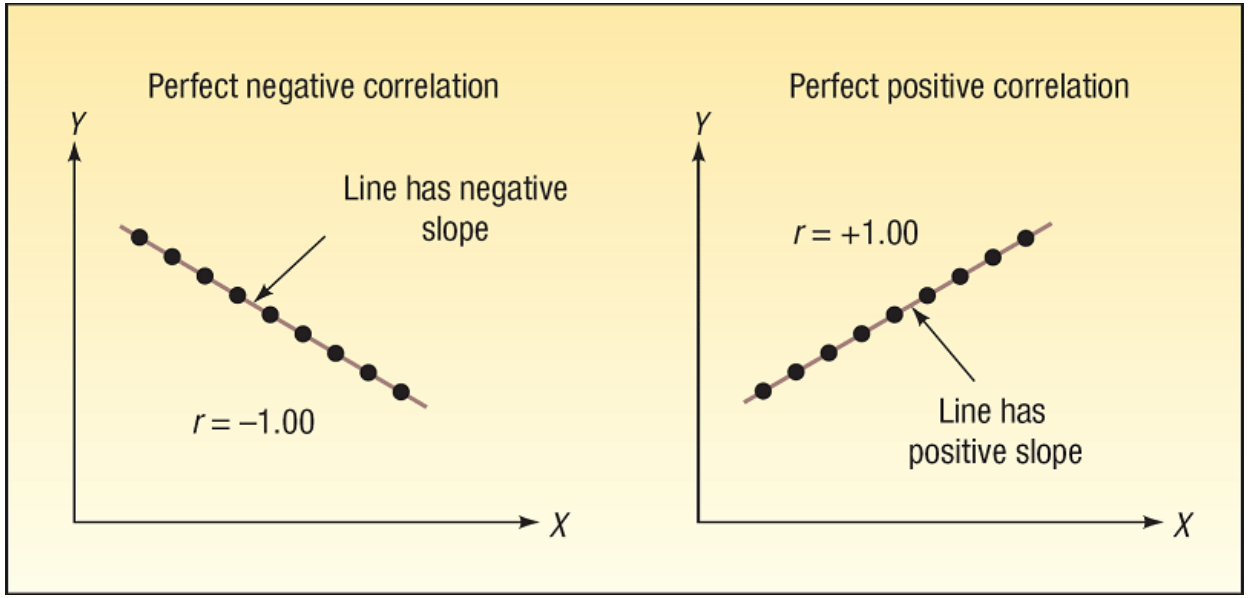
6.การวิเคราะห์สหสัมพันธ์ (Correlation, r)

- ในการวิเคราะห์สหสัมพันธ์เราจะพิจารณาค่าความสัมพันธ์ระหว่างตัวแปรที่เรียกว่าสัมประสิทธิ์สหสัมพันธ์ (Correlation: r)
- สัมประสิทธิ์สหสัมพันธ์ คือตัววัดระดับความสัมพันธ์ระหว่างตัวแปรสองตัวแปร โดยที่ตัวแปรทั้งสองตัวแปรนั้นจะต้องถูกวัดในระดับ interval หรือ ratio-scaled เท่านั้น
  - ค่าของสัมประสิทธิ์จะมีค่าอยู่ระหว่าง -1.00 ถึง1.00
  - ค่า -1.00 หรือ 1.00 แสดงถึงค่าความสัมพันธ์ที่สมบูรณ์ระหว่างตัวแปรทั้งสองตัวแปร
  - ค่าที่ใกล้ 0.0 แสดงว่ามีความสัมพันธ์กันน้อยมาก
  - ค่าติดลบแสดงให้เห็นถึงความสัมพันธ์ในทิศทางตรงกันข้าม ในขณะที่ค่าบวกแสดงให้เห็นว่ามีความสัมพันธ์ในทิศทางเดียวกัน

สูตรการคำนวณค่าสัมประสิทธิ์สหสัมพันธ์

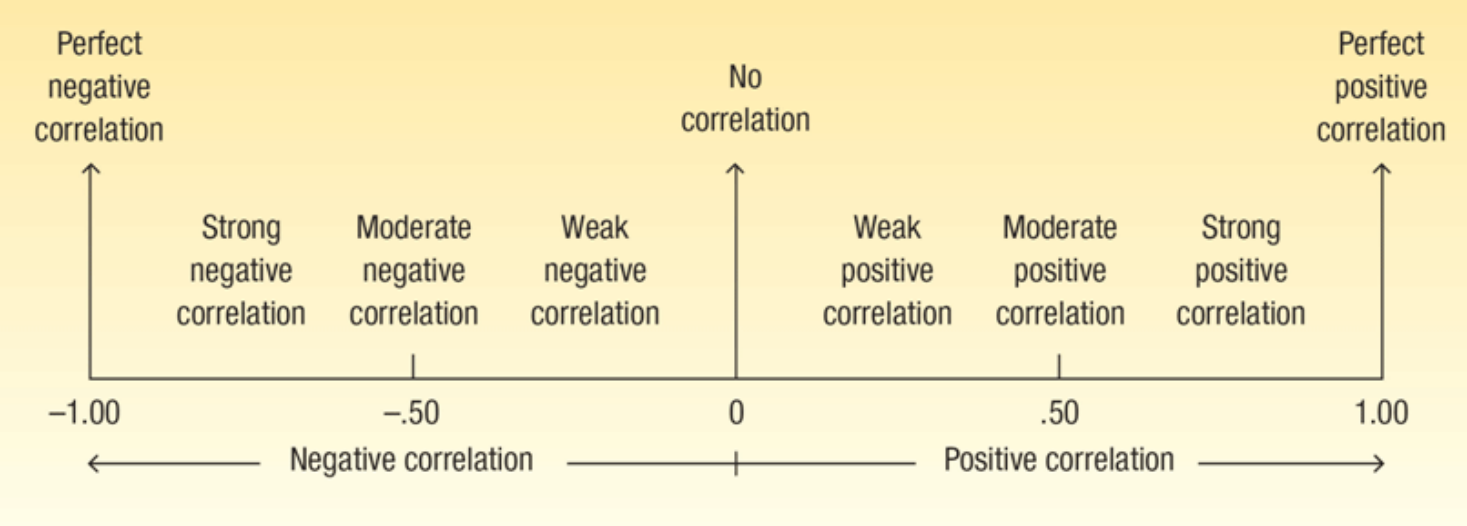
$$\text{CORRELATION COEFFICIENT} \quad r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n-1)s_x s_y}$$

ความสัมพันธ์กับแบบสมบูรณ์



6.การวิเคราะห์สหสัมพันธ์ (Correlation, r)

การตีความหมายของค่าสัมประสิทธิ์สหสัมพันธ์



ค่าสัมประสิทธิ์ตัวกำหนด (R^2)

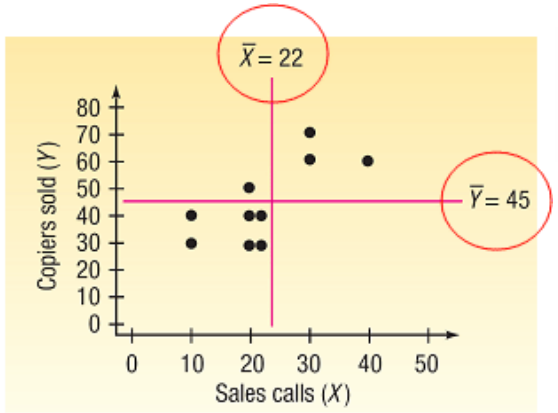
ค่าสัมประสิทธิ์ตัวกำหนดหรือที่เรียกว่า coefficient of determination (r2) เป็นค่าที่แสดงสัดส่วนของความผันแปรของตัวแปรตามที่สามารถถูกอธิบายได้โดยความผันแปรของตัวแปรอิสระ ซึ่งค่านี้คำนวณได้โดยการยกกำลังสองของค่าสัมประสิทธิ์สหสัมพันธ์

- จะมีค่าระหว่าง 0 ถึง 1
- ค่าที่ได้**ไม่สามารถใช้บอกทิศทางของความสัมพันธ์ได้**

6.การวิเคราะห์สหสัมพันธ์ (Correlation, r)

Using the formula:
CORRELATION COEFFICIENT

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y}$$



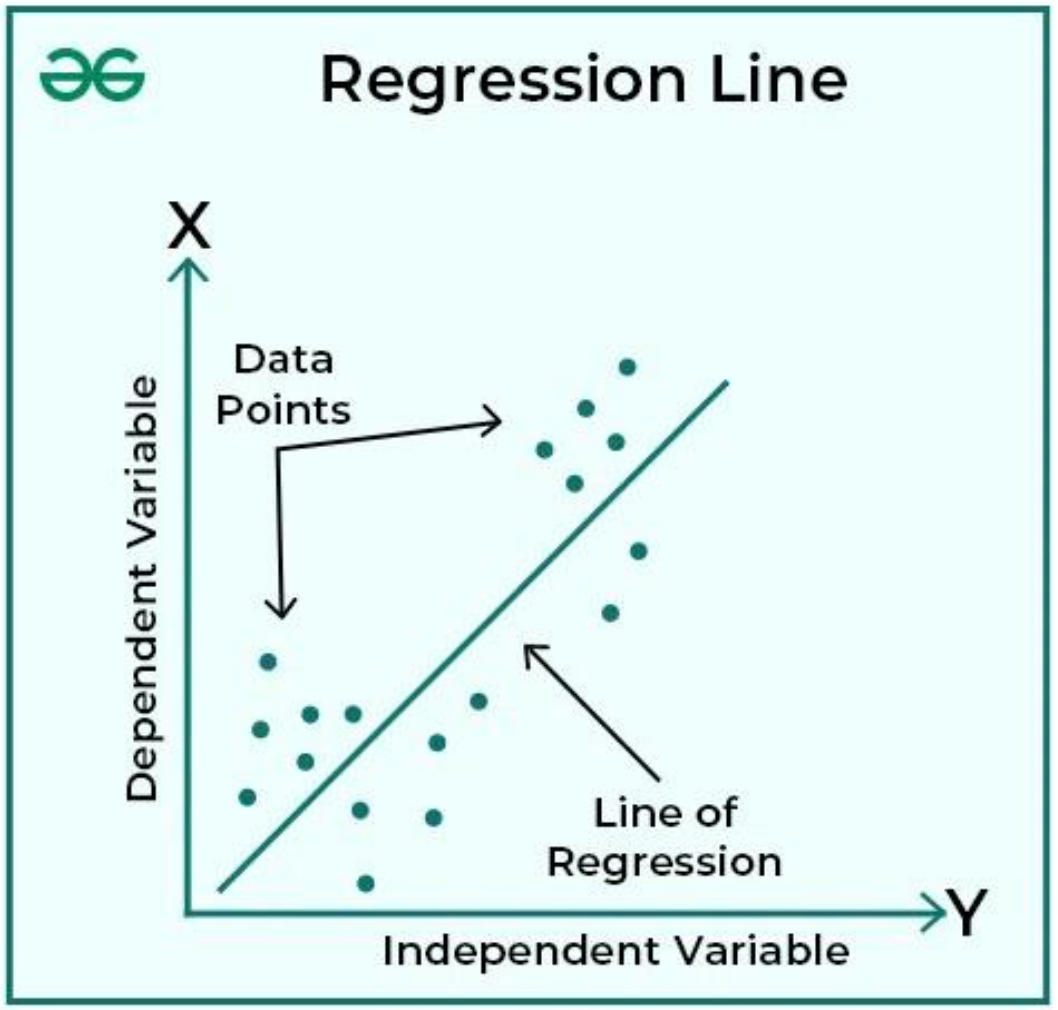
Sales Representative	Calls, Y	Sales, X	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	−2	−15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	−2	−5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	−12	−15	180
Carlos Ramirez	10	40	−12	−5	60
Rich Niles	20	40	−2	−5	10
Mike Kiel	20	50	−2	5	−10
Mark Reynolds	20	30	−2	−15	30
Soni Jones	30	70	8	25	200
					900

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} = \frac{900}{(10 - 1)(9.189)(14.337)} = 0.759$$

ค่าสัมประสิทธิ์สหสัมพันธ์ได้ค่าเท่ากับ 0.759 ซึ่งเป็นค่าบวกที่ใกล้กับ 1 ตีความหมายได้ว่าจำนวนครั้งในการไปพบลูกค้ามีความสัมพันธ์กันอย่างสูงกับจำนวนเครื่องถ่ายเอกสารที่ขายได้ อย่างไรก็ตามเราไม่สามารถสรุปได้ว่าจำนวนครั้งในการไปพบลูกค้าเป็นสาเหตุทำให้เกิดยอดขายที่เพิ่มขึ้น เราบอกได้เพียงแค่ว่ามีความสัมพันธ์กันเท่านั้น

7. การวิเคราะห์การถดถอย (Regression Analysis)

- การศึกษาความสัมพันธ์ของตัวแปรตั้งแต่ 2 ตัวขึ้นไป โดยมีวัตถุประสงค์เพื่อพยากรณ์ค่าตัวแปรหนึ่งจากตัวแปรอื่น ๆ ที่มีความสัมพันธ์กับตัวแปรที่ต้องการพยากรณ์ ดังนั้นจึงแบ่งกลุ่มตัวแปรได้เป็น 2 กลุ่มคือ
  1. **ตัวแปรอิสระ** (Independent Variable) คือ ตัวแปรที่ทราบค่าล่วงหน้า มักใช้สัญลักษณ์ X
  2. **ตัวแปรตาม** (Dependent Variable) คือ ตัวแปรที่เปลี่ยนแปลงไปตามการเปลี่ยนแปลงของตัวแปรอิสระ มักใช้สัญลักษณ์ Y
- การวิเคราะห์การถดถอยสามารถแบ่งได้เป็น 2 ประเภท คือ
  1. **การวิเคราะห์การถดถอยอย่างง่าย** (Simple Regression) เป็นการศึกษาความสัมพันธ์เฉพาะตัวแปรตาม 1 ตัว กับตัวแปรอิสระ 1 ตัว
  2. **การวิเคราะห์การถดถอยเชิงพหุหรือเชิงซ้อน** (Multiple Regression) เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตาม 1 ตัว และตัวแปรอิสระมากกว่า 1 ตัวขึ้นไป



https://www.geeksforgeeks.org/what-is-regression-line/

7. การวิเคราะห์การถดถอย

ข้อสมมติเบื้องต้นในการวิเคราะห์การถดถอย

1. ตัวอย่างที่นำมาวิเคราะห์ ต้องได้จากการสุ่ม นั่นคือความผิดพลาดหรือความคลาดเคลื่อนต้องเป็นอิสระต่อกัน (No Autocorrelation) ทดสอบจากค่า Durbin-Watson มีค่าใกล้ๆ 2 (อยู่ระหว่าง 1.5 - 2.5)

2. ε_i มีการแจกแจงแบบปกติ ค่าเฉลี่ย = 0 ทดสอบจากนำค่าความคลาดเคลื่อนมาพล็อตเป็นฮิสโตแกรม ให้พิจารณาว่ากราฟเป็นรูประฆังคว่ำโดยมีค่าความคลาดเคลื่อนมาตรฐานอยู่ระหว่าง -3 ถึง 3 และความแปรปรวนของความคลาดเคลื่อนคงที่ทุกค่าของ X (มีค่า = σ^2Homoscedasticity) โดยดูจากกราฟของค่าความคลาดเคลื่อนกับค่า X มีการกระจายรอบ ๆ ค่า 0 และไม่กว้างหรือแคบเข้ามาเมื่อค่า X สูงขึ้น

3. ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์แบบเส้นตรง ทดสอบจากค่า F ของค่า β

ตัวอย่าง

- การโฆษณากับมีความสัมพันธ์กับยอดขายของบริษัทหรือไม่
- ค่าไฟฟ้าในการทำความร้อนกับขนาดพื้นที่ของบ้านมีความสัมพันธ์กันหรือไม่
- การใช้น้ำมันกับขนาดของเครื่องยนต์ มีความสัมพันธ์กันหรือไม่
- ชั่วโมงการอ่านหนังสือกับคะแนนสอบของนักศึกษามีความสัมพันธ์กันหรือไม่

7. การวิเคราะห์การถดถอย

สมการเส้นตรง $Y_i = a + \beta X_i + \varepsilon_i$

โดยที่

Yi คือ ตัวแปรตาม (Dependent Variable) สำหรับค่าสังเกตตัวที่ i

Xi คือ ตัวแปรอิสระ (Independent Variable) สำหรับค่าสังเกตตัวที่ i

a คือ พารามิเตอร์ แทนจุดตัดแกน Y (ค่า Y เมื่อ X มีค่าเป็น 0)

ε_i คือ ความคลาดเคลื่อนอย่างสุ่ม สำหรับค่าสังเกตตัวที่ i

β คือ ความชัน (Slope) ของเส้นตรง แสดงการเปลี่ยนไปของ Y เมื่อ X เปลี่ยนไป 1 หน่วย บางครั้งเรียกว่า สัมประสิทธิ์การถดถอย (Regression Coefficient)

$\beta > 0$ X และ Y มีความสัมพันธ์ทางเดียวกัน

$\beta < 0$ X และ Y มีความสัมพันธ์ตรงกันข้าม

$\beta \rightarrow 0$ X และ Y มีความสัมพันธ์กันน้อยมาก

$\beta = 0$ X และ Y ไม่มีความสัมพันธ์กันเลย

7. การวิเคราะห์การถดถอย

ในการวิเคราะห์การถดถอยเราจะใช้ตัวแปรอิสระ (X) ในการทำนายตัวแปรตาม (Y)

- ความสัมพันธ์ระหว่างตัวแปรจะเป็นความสัมพันธ์เชิงเส้น
- ตัวแปรทั้งสองจะต้องถูกวัดอย่างน้อยในระดับ interval scale
- เราจะใช้หลักการของ least squares criterion ในการหาสมการ

GENERAL FORM OF LINEAR REGRESSION EQUATION $\hat{Y} = a + bX$

where

$\hat{Y}$ read Y hat, is the estimated value of the Y variable for a selected X value.
a is the Y-intercept. It is the estimated value of Y when $X = 0$. Another way to put it is: a is the estimated value of Y where the regression line crosses the Y-axis when X is zero.
b is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit (either increase or decrease) in the independent variable X.
X is any value of the independent variable that is selected.

REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual Y values and the predicted values of Y.

7. การวิเคราะห์การถดถอย

การคำนวณหาความชันของเส้น (b)

SLOPE OF THE REGRESSION LINE

$$b = r\frac{s_y}{s_x}$$

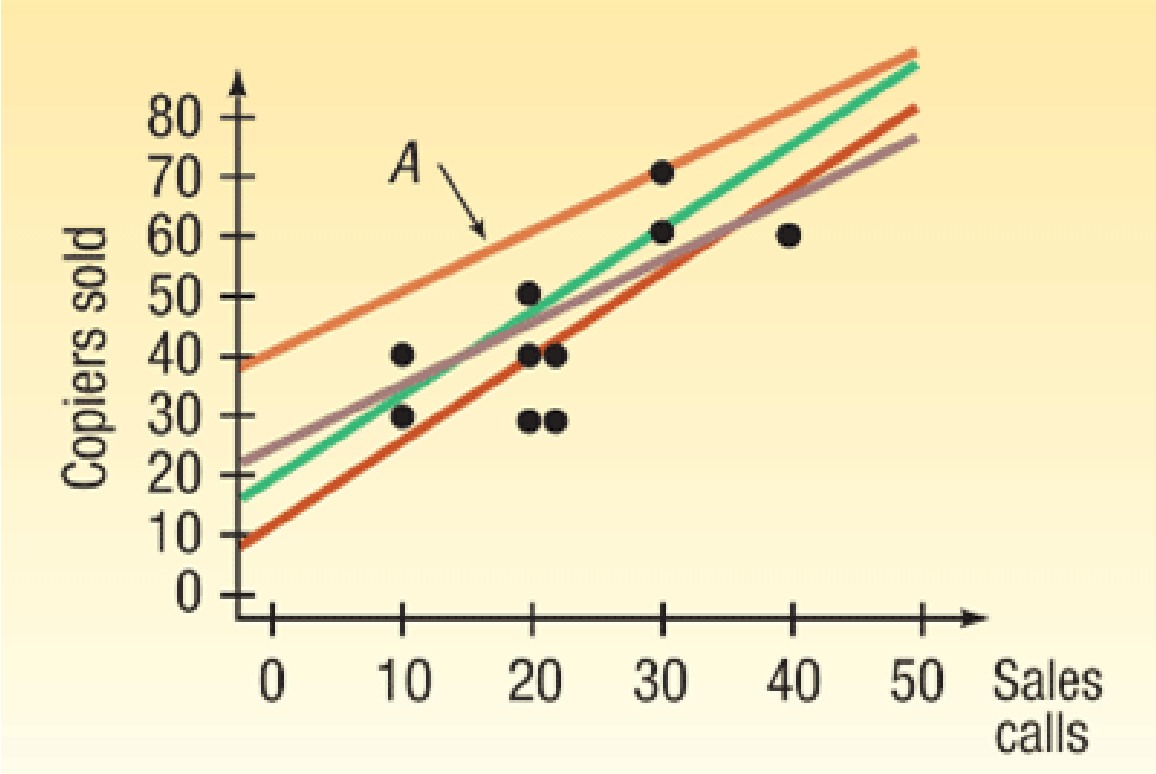
where

r is the correlation coefficient.

s_y is the standard deviation of Y (the dependent variable).

s_x is the standard deviation of X (the independent variable).

หลักการ Least Squares Regression



7. การวิเคราะห์การถดถอย

- ให้หาสมการถดถอย เพื่อใช้ทำนายยอดขายเครื่องถ่ายเอกสารของพนักงานที่ไปพบลูกค้าจำนวน 20 ครั้งต่อเดือน

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Step 1 – Find the slope (b) of the line

$$b = r\left(\frac{s_y}{s_x}\right) = .759\left(\frac{14.337}{9.189}\right) = 1.1842$$

Step 2 – Find the y-intercept (a)

$$a = \bar{Y} - b\bar{X} = 45 - 1.1842(22) = 18.9476$$

The regression equation is :

$$\hat{Y} = a + bX$$

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(20)$$

$$\hat{Y} = 42.6316$$

7. การวิเคราะห์การถดถอย

- ใช้สมการที่ได้ โดยการแทนค่า X (จำนวนครั้งในการไปพบลูกค้า) เพื่อใช้ในการประมาณค่า Y (จำนวนเครื่องถ่ายเอกสารที่ขายได้)

Sales Representative	Sales Calls (X)	Estimated Sales ($\hat{Y}$)	Sales Representative	Sales Calls (X)	Estimated Sales ($\hat{Y}$)
Tom Keller	20	42.6316	Carlos Ramirez	10	30.7896
Jeff Hall	40	66.3156	Rich Niles	20	42.6316
Brian Virost	20	42.6316	Mike Kiel	20	42.6316
Greg Fish	30	54.4736	Mark Reynolds	20	42.6316
Susan Welch	10	30.7896	Soni Jones	30	54.4736

Tom Keller

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(20)$$

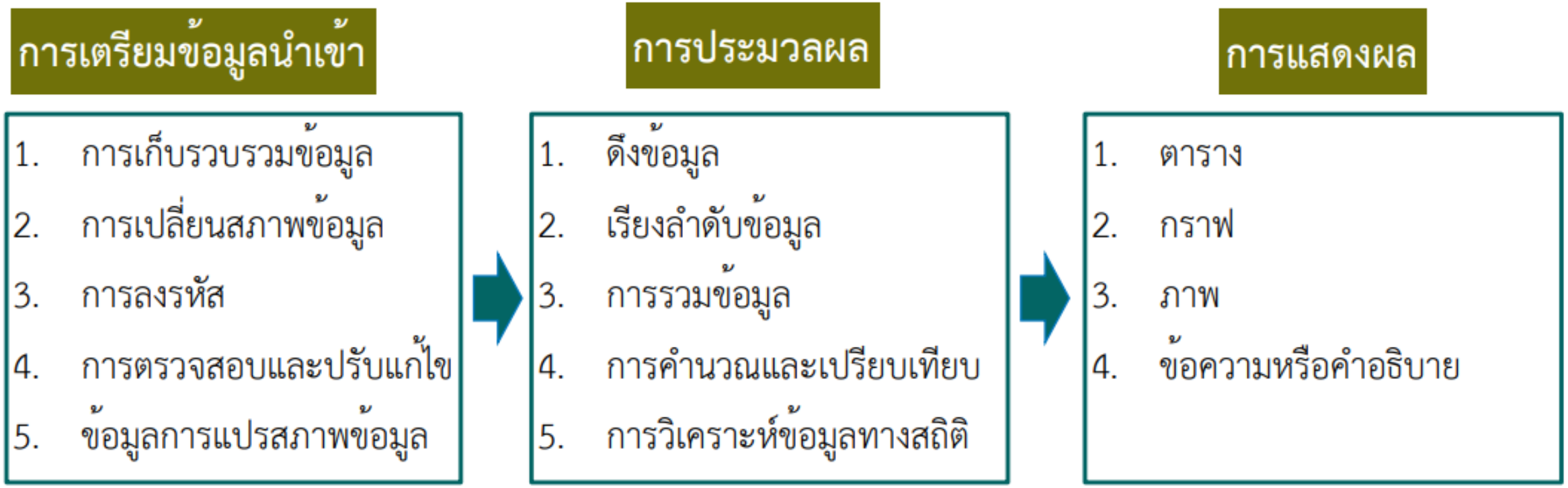
$$\hat{Y} = 42.6316$$

Soni Jones

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(30)$$

$$\hat{Y} = 54.4736$$


การเตรียมข้อมูลนำเข้า
1. การเก็บรวบรวมข้อมูล
2. การเปลี่ยนสภาพข้อมูล
3. การลงรหัส
4. การตรวจสอบและปรับแก้ไข
5. ข้อมูลการแปรสภาพข้อมูล
การประมวลผล
1. ดึงข้อมูล
2. เรียงลำดับข้อมูล
3. การรวมข้อมูล
4. การคำนวณและเปรียบเทียบ
5. การวิเคราะห์ข้อมูลทางสถิติ
การแสดงผล
1. ตาราง
2. กราฟ
3. ภาพ
4. ข้อความหรือคำอธิบาย

ประเภทของตัวแปร กับการนำเสนอข้อมูล

Scale	Examples	Measures of central tendency.... for a single variable	Measures of dispersion.... for a single variable	Visual summary.... for a single variable	Measures of relation.... between variable	Visual summary.... between variable
Nominal	Gender	Mode	-	Bar chart, pie chart	Cross-tab	Stacked bars, clustered bars
Ordinal	Satisfaction rating on a 5-point scale (1=not satisfied at all; 5 = extremely satisfied	Median	Semi-interquartile range	Bar chart, pie chart	Cross-tab	Stacked bars, clustered bars
Interval		Mean	Min, Max, SD, variance, coefficient of variation	Histogram, scatterplot, box-and-whisker plot	Correlation	Scatterplots
Ratio	Age, income, sales	Mean	Min, Max, SD, variance, coefficient of variation	Histogram, scatterplot, box-and-whisker plot	Correlation	Scatterplots

Semi-interquartile range [พิสัยกึ่งควอไทล์ = (Q3 - Q1) / 2]

coefficient of variation [CV= (SD/Mean)*100%]