# การค้นหาความรู้โดยการวิเคราะห์ข้อมูลด้วย R และ Weka Knowledge Discovery by Data Analysis with R and Weka

## สุณี รักษาเกียรติศักดิ์

# การค้นหาความรู้

การค้นหาความรู้ในสิ่งที่เราอยากทราบหรือเราอาจจะเรียกว่ากระบวนการวิจัย สามารถดำเนินการ ได้ 2 แนวทางคือ forward และ backward

Forward เริ่มจากวัตถุประสงค์/คำถามการวิจัย หรือโจทย์หรือปัญหาการวิจัยนั่นเอง (Objective หรือ Problem definititions) จากนั้นก็จะเป็นวิธีการดำเนินการ (ซึ่งอาจจะต้องมีกระบวนการทบทวนวรรณกรรม – review literature หรือไม่ต้องมีก็ได้ เพราะโจทย์อาจจะเป็นปัญหาที่เกี่ยวกับการปฏิบัติงาน ซึ่งก็ดำเนินการ ้ได้เลย) ในวิธีการดำเนินงาน จะเกี่ยวข้องกับการออกแบบและ/หรือการพัฒนาเพื่อให้ได้คำตอบของโจทย์ที่ ในกระบวนการนี้อาจจะต้องมีการออกแบบเครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลเพื่อตอบโจทย์ ต้องการ เครื่องมืออย่างง่ายที่ใช้กันคือการออกแบบสอบถาม จากนั้นก็เป็นการกำหนดกลุ่มตัวอย่างที่จะตอบ แบบสอบถาม (sample) ซึ่งตามหลักการแล้วก็ต้องมีการนิยามประชากรที่เราสนใจจะศึกษา (population) ้ด้วย และระบุว่ากลุ่มตัวอย่างที่ได้มาและเป็นตัวแทน (representative) ของประชากรนั้น ได้มาด้วยวิธีการสุ่ม ้ตัวอย่าง (sampling) แบบใด เช่น การสุ่มอย่างง่าย (random sampling) ซึ่งคือการสุ่มที่ทุกหน่วยของ ประชากรมีโอกาสได้รับการสุ่มเข้ามาด้วยความน่าจะเป็นที่เท่ากัน หรือการสุ่มแบบแบ่งชั้น (stratify sampling) ซึ่งคือการสุ่มที่ให้ได้สัดส่วนของตัวแปรที่ต้องการแบ่งชั้นตามต้องการ เป็นต้น แต่การวิเคราะห์ ข้อมูลเพื่อตอบโจทย์ในการทำงานมักใช้กลุ่มตัวอย่างตามสะดวก (convenient sample) ตัวอย่างเช่น ในการ ้ปฏิบัติงานทุกอย่าง ต้องมีการประเมินผลการปฏิบัติงานของกิจกรรมใด กิจกรรมหนึ่ง ก็ใช้กลุ่มตัวอย่างของผู้ ที่เข้ามาร่วมกิจกรรม เป็นตัวแทนของประชากรที่ต้องการ

ในการวิเคราะห์ข้อมูลนั้นเราศึกษาจากกลุ่มตัวอย่าง (sample) ที่เรามีข้อมูล โดยคำนวณค่าสถิติ (statistics) ต่าง ๆ จากกลุ่มตัวอย่าง ที่เราเรียกว่าสถิติพรรณนา (descriptive statistics) ทำให้เราได้ความรู้ เกี่ยวกับลักษณะของกลุ่มตัวอย่างที่เราศึกษา แต่เราจะขยายผลของการศึกษาจากกลุ่มตัวอย่างนี้ไปอนุมาน (inference) เพื่อให้ได้ความรู้เกี่ยวกับประชากร (parameters) ได้โดยกระบวนการทดสอบสมมุติฐานทางสถิติ (statistical hypothesis testing) ซึ่งเป็นกระบวนการใช้สถิติอนุมาน/สถิติอ้างอิง (inferential statisitcs) ซึ่ง คำนวณจากสถิติพรรณนา (descriptive statistics) สำหรับการทดสอบคุณลักษณะของประชากร (parameters) ที่ต้องการ โดยใช้หลักการความน่าจะเป็น (probability distribution) ของการแจกแจง (distribution) ของสถิติอ้างอิงนั้น เพื่อสรุปความจริงเกี่ยวกับประชากร สถิติอนุมานมี 2 แบบ คือ สถิติพาราเมตริก (parametric) และสถิตินอนพาราเมตริก (non parametric) สถิติพาราเมตริกจะมีข้อตกลงเบื้องต้น (assumption) ข้อมูล/ตัวแปร (variable) ต้องมีการแจก แจงความน่าจะเป็นแบบปกติ (normal distribution) ซึ่งโดยธรรมชาติแล้ว ข้อมูลที่เกิดโดยธรรมชาติจะมีการ แจกแจงแบบปกติ ถ้าข้อมูลไม่มีการแจกแจงแบบปกติควรจะต้องใช้สถิตินอนพาราเมตริกแทน ในการ คำนวณของสถิตินอนพาราเมตริกจะไม่ใช้ข้อมูลดิบของตัวแปรมาคำนวณ แต่จะใช้อันดับ (rank) ของข้อมูล มาคำนวณ ซึ่งเป็นวิธีการที่หยาบกว่า ดังนั้นหากสามารถจะใช้วิธีการทางสถิติ (statistical methods) ที่ เรียกว่าพาราเมตริกได้ ก็ควรทำในเบื้องแรก หากไม่สามารถทำได้ เช่น ข้อมูลไม่มีการแจกแจงแบบปกติ หรือ ข้อตกลงเบื้องต้นของวิธีการทางสถิตินั้นไม่เป็นจริง ก็ค่อยไปใช้วิธีการแบบนอนพาราเมตริกแทน ข้อ ได้เปรียบของวิธีการพาราเมตริกลือให้กำลังการทดสอบ (power) สูงกว่าวิธีการนอนพาราเมตริก

### วิธีการทาง backward มี 2 กรณี คือ

กรณีที่ 1: เราอาจจะไม่ได้ระบุคำถามการวิจัยที่ชัดเจนนักหรือเพียงระบุวัตถุประสงค์กว้าง ๆ ไว้ แล้วเก็บ ข้อมูลต่าง ๆ ที่คาดว่าจะเกี่ยวข้องกับวัตถุประสงค์ของโจทย์ เช่น **วัตถุประสงค์ คือ ต้องการศึกษาการ ฝึกงานของนิสิตเอกวิทยาการคอมพิวเตอร์ (วิชา CP493 Internship)** ซึ่งเป็นวิชาบังคับของหลักสูตร ข้อมูลและเครื่องมือที่จะใช้ในการเก็บรวบรวมข้อมูลที่จะตอบโจทย์ยังไม่ชัดเจน (ซึ่งจะเป็นโจทย์ให้ผู้อบรมได้ ฝึกปฏิบัติต่อไป)

กรณีที่ 2: การค้นหาความรู้จากข้อมูลที่อยู่ในฐานข้อมูลการปฏิบัติงาน (Knowledge Discovery in Database หรือ KDD) การค้นหาความรู้ในกรณีนี้จะใช้วิธีการทางสถิติ (statistical methods) หรือวิธีการทางการทำ เหมืองข้อมูล (data mining) ด้วยวิธีการทางการเรียนรู้ของเครื่อง (machine learning) ก็ได้

การค้นหาความรู้โดยวิธีการเรียนรู้ของเครื่องจะตั้งอยู่บนสมมติฐานว่าจะแบ่งข้อมูลออกเป็น 2 ชุด ชุดหนึ่งคือชุดในการสร้างโมเดล (training set) ซึ่งจะเป็นข้อมูลส่วนใหญ่ และข้อมูลอีกชุดหนึ่งเรียกว่าชุด ทดสอบโมเดล (test set) ข้อมูลสองชุดนี้ต้องเป็นอิสระต่อกัน (independent) คือต้องไม่ซ้ำกัน เมื่อทำการ train แล้ว จะได้ผลลัพธ์เป็นความรู้หรือโมเดล ซึ่งอาจจะอยู่ในรูปสมการ หรือ กฏ เป็นต้น โมเดล ที่ได้จะมีความน่าเชื่อถือเพียงใดต้องมีการประเมินโมเดล โดยจะใช้ test set (จะกล่าวในรายละเอียดต่อไป)

## วัตถุประสงค์ของการอบรม

- 1. แสดงการใช้เครื่องมือออนไลน์ Google WORK (FORMS และ SHEETS) ในการเก็บรวบรวมข้อมูล
- แสดงการวิเคราะห์ข้อมูลอย่างง่ายโดย Google สำหรับข้อมูลที่จัดเก็บโดย FORMS และวิเคราะห์ ข้อจำกัดของการประมวลผลโดย Google จากโจทย์ "การฝึกงานของนิสิต"
- แสดงการนำข้อมูลจาก Google WORK มาวิเคราะห์ข้อมูลทางสถิติโดยใช้โปรแกรม R โดยมีการสร้างตัว แปรใหม่ วิเคราะห์สถิติพรรณา สถิติอนุมาน เช่น การเปรียบเทียบค่าเฉลี่ย 2 กลุ่มด้วย t-test การ

เปรียบเทียบค่าเฉลี่ยมากกว่า 2 กลุ่มด้วย One way ANOVA และการทดสอบความแตกต่างของความถึ่ ของกลุ่มด้วยไคร์สแคว์

- ใช้ข้อมูลของ Weka เป็นตัวอย่างในสร้างโมเดลโดยใช้การวิเคราะห์ Regression และ Logistic Regression โดยวิเคราะห์ด้วยโปรแกรม R
- 5. ใช้ Weka ในการทำ data mining โดยใช้ข้อมูลตัวอย่างของ Weka และวิเคราะห์โจทย์ประเภท Prediction/Regression, Classification, Association, Clustering โดยใช้ Explorer Interface
- เปรียบเทียบการวิเคราะห์โจทย์ประเภท Classification ด้วย data set หลายชุด และใช้หลาย algorithm โดยใช้ Experimenter Interface

# การใช้โปรแกรม R ในการวิเคราะห์ข้อมูลทางสถิติ

สุณี รักษาเกียรติศักดิ์

### การ Install โปรแกรม R และ Package RcmdR

- ไป download โปรแกรมที่ R Project <u>http://www.r-project.org/</u> เลือก **Download R** แล้วเลือก CRAN Mirrors เป็น **Thailand**
- 2. เลือก Download R for Windows แล้วเลือก base จะได้ดังรูป

Download R 3.4.3 for Windows (62 megabytes, 32/64 bit)

Installation and other instructions New features in this version

3. เลือก **Download R 3.4.3 for Windows** ให้ save Application ที่ download (R-3.4.3-win.exe)

4. ทำการ run โปรแกรม (install) เลือกโฟลเดอร์ตาม default

🔂 Setup - R for Windows 3.4.3 —	[		×
Select Destination Location Where should R for Windows 3.4.3 be installed?		q	R
Setup will install R for Windows 3.4.3 into the following folde	r.		
To continue, click Next. If you would like to select a different folder, o	lick Brov	vse.	
C:\Program Files\R\R-3.4.3	B <u>r</u> ov	vse	
At least 1.2 MB of free disk space is required.			
< <u>B</u> ack <u>N</u> ext >		Cancel	

5. Select Components

## ให้เอา 32-bit Files ออกสำหรับเครื่อง 64 bits

🕞 Setup - R for Windows 3.4.3	_		×
Select Components Which components should be installed?			R
Select the components you want to install; clear the components install. Click Next when you are ready to continue.	you do no	t want to	
Custom installation		$\sim$	
Core Files		83.7 MB	
32-bit Files		48.7 MB	
G4-bit Files		50.2 MB	
Message translations		7.3 MB	
Current selection requires at least 142.1 MB of disk space.			
< <u>B</u> ack <u>N</u> e	xt >	Cano	cel

6. เลือกตาม default และ Next จนระบบติดตั้งเสร็จ



5



- 7. เมื่อ Install เรียบร้อยแล้วจะปรากฏไอคอนของโปรแกรม R บน Desktop
- 8. ให้ double click โปรแกรมเพื่อใช้งาน

โดยต้อง Install package เพิ่มสำหรับเมนูการวิเคราะห์ข้อมูลทางสถิติ

โดยไปที่ Packages > Install package(s)



- 9. เลือก CRAN mirror
- 10. เลือก Packages เป็น Rcmdr

ระบบจะถาม



ให้ตอบ Yes



ระบบจะทำการ download packages ต่าง ๆ มาเก็บไว้

11. เลือก Packages > Load Packages จะขึ้นดังรูป

Ø	×
8	The following packages used by Rcmdr are missing: sem, rmarkdown, rgl, multcomp, Imtest, leaps, colorspace, aplpack Without these packages, some features will not be available. Install these packages?
	<u>Y</u> es <u>N</u> o

12. ให้กด Yes แล้วกด OK คือให้ Install จาก CRAN

7% Install Missing Packages	
Install Packages From: CRAN	
Local package directory (must include PACKAGES index file)	Specify package directory: Browse
OK Cancel	Help

13. ช่วงนี้จะใช้เวลาสักพัก รอจน download เสร็จ (ระบบจะ download ทุก packages ที่จำเป็น)

14. เมื่อเรียบร้อยจะขึ้นหน้าต่าง R Commander สามารถทดสอบโดยอ่านข้อมูลจาก Excel ได้ โดยไป ที่เมนู Data > Import data > from Excel



15. เมื่อเลิกใช้งานให้ปิด R Commander และ RGui

16. เมื่อต้องการใช้งานใหม่ เมื่อเปิด RGui ขึ้นมาต้อง Load Package Rcmdr มาใช้งานเสมอ <u>หมายเหตุ</u> เนื่องจากการติดตั้งต้องใช้ Internet ดังนั้นจึงได้เตรียมโปรแกรม R ที่ install เรียบร้อยแล้วใน Virtual Machine สำหรับภาคปฏิบัติ

# การใช้เครื่องมือออนไลน์ Google WORK (FORMS & SHEETS) ในการเก็บรวบรวมข้อมูล

### Google SHEETS

เครื่องมือในการเก็บรวบรวมข้อมูล ในแบบเดิมมักจะใช้แบบสอบถามที่เป็นกระดาษ เช่น ดังตัวอย่าง แบบรายงานการฝึกงานของนิสิต ชั้นปีที่ 3 .... ดังแสดงในไฟล์ *Report\_Internship* จากนั้นนำข้อมูลมา คีย์ข้อมูลลงใน Excel หรือ Goggle SHEETS ได้ดังแสดงในไฟล์ชื่อ *Report\_Internship* 

### Google FORMS

การเก็บรวบรวมข้อมูลออนไลน์ด้วย Google FORMS กำลังเป็นที่นิยมอย่างมากเนื่องด้วยการสร้าง ฟอร์ม (FORM) ทำได้ง่าย และสามารถจะเก็บข้อมูลแบบออนไลน์ได้ทันที ทำให้ประหยัดค่าใช้จ่าย และ Google FORM ยังประมวลผลข้อมูลอย่างง่ายให้ในระดับตัวแปรเดียว เช่น จำนวนและร้อยละ เป็นต้น หาก ผู้วิจัยต้องการวิเคราะห์ข้อมูลเพิ่มเติมด้วยโปรแกรมสำเร็จรูปทางสถิติอื่น ๆ เช่น โปรแกรม SPSS หรือ โปรแกรม R ก็สามารถ download ข้อมูลจาก Form\_responses มาใน Excel แล้วดำเนินการต่อตามต้องการ ได้

จากโจทย์ "การฝึกงานของนิสิต" ผู้วิจัยได้สร้างเครื่องมือในการเก็บข้อมูลเพิ่มเติมจากการฝึกงาน ของนิสิต โดยใช้ Google FORMS ดังแสดงในไฟล์ *Form* และข้อมูลจากการตอบแบบสอบถาม ดังแสดงใน ไฟล์ *Form\_Responses* โดย Google จะประมวลผลข้อมูลอย่างง่ายให้ดังแสดงในไฟล์ *Summary\_Responses* 

# การจัดเตรียมข้อมูลสำหรับการวิเคราะห์

ผู้วิจัยได้จัดเตรียมข้อมูลสำหรับการวิเคราะห์โดยนำตัวแปรที่สนใจจากไฟล์ *Report\_Internship* และจากไฟล์ *Form\_Responses* โดยทำการ join ด้วย ID ซึ่งก็คือหลักการทำ ETL นั้นเอง แต่เนื่องจาก ข้อมูลมีจำนวนน้อยและผู้วิจัยรู้จักข้อมูลดี จึงทำด้วยมือ โครงสร้างข้อมูลที่ได้ดังแสดงในตารางที่ 1 และ ข้อมูลดังแสดงในไฟล์ *DataAll* 

4		ና	ັ	e ع		4		99
ตารางท่	1	เคร	งสราง	ານອ	มลของ	"การฝา	กงานขอ	เงนสต"

ชื่อตัวแปร	คำอธิบายตัวแปร	ค่าของตัวแปร	ระดับการวัด
		(Domain)	(Measure)
ID	รหัสนิสิต	รหัส 4 หลักสุดท้าย	-
GENDER	เพศ	ชาย, หญิง	
GPA	เกรดเฉลี่ย	1.00-4.00	
ORG_SIZE	ขนาดขององค์กรที่ผึกงาน	ขนาดเล็ก (น้อยกว่า 15 คน)	
		ขนาดกลาง (15 - 50 คน)	
		ขนาดใหญ่ (มากกว่า 50 คน)	
PAY	ค่าเบี้ยเลี้ยง	ไม่ได้รับเบี้ยเลี้ยง	
		น้อยกว่า 100	
		100 - 200	
		201 - 300	
		301 - 400	
		มากกว่า 400	
DAYS	จำนวนวันที่ทำงาน	20 – 60	
HOURS	จำนวนชั่วโมงที่ทำงาน	130 - 500	
CP121,,	ความรู้ที่ใช้ในวิชาที่เรียนมา	1 – 5 (น้อย – มาก)	
CP444			
WSK1,, ความรู้และทักษะในการทำงาน		4=ดีมาก, 3=ดี, 2=ปานกลาง,	
WSK6		1=น้อย, 0=น้อยมาก	
USK1,, อัตลักษณ์นิสิตของมหาวิทยาลัย		4=ดีมาก, 3=ดี, 2=ปานกลาง,	
USK5	"มีทักษะการสื่อสาร"	1=น้อย, 0=น้อยมาก	
Letter	ใบรับรองการฝึกงาน	0=ไม่มี, 1=มี	
Stamp	ใบประทับตราบริษัท	0=ไม่มี, 1=มี	

### ระดับการวัดของข้อมูล (Level of Measurement)

ข้อมูลหรือตัวแปรหรือแอททริบิวต์ (Attribute) ที่จะวิเคราะห์เพื่อหาค่าสถิติของข้อมูลในเบื้องตันนั้น เราต้องทราบว่าข้อมูลของนั้นมีระดับการวัด (measure) เป็นอย่างไร

ระดับการวัดแบ่งเป็น 2 ระดับใหญ่ ๆ คือ

- ข้อมูลเชิงคุณลักษณะหรือข้อมูลเชิงคุณภาพ (Qualitative data) แบ่งออกเป็น 2 ระดับคือ
  - Nominal เป็นข้อมูลเชิงกลุ่มหรือนามบัญญัติ เช่น เพศ หรือ Gender มี 2 กลุ่ม/ค่า คือ
     Male และ Female หรือ ชาย และ หญิง เป็นตัน
  - Ordinal เป็นข้อมูลเชิงกลุ่มแบบมีอันดับ เช่น วุฒิการศึกษา หรือ EDU มี 3 กลุ่ม/ค่า คือ ต่ำกว่าปริญญาตรี ปริญญาตรี สูงกว่าปริญญาตรี เป็นต้น ซึ่งการศึกษามีอันดับจากต่ำ ไปสูง (มีปริมาณ หรือ magnitude แต่ไม่มี scale หรือหน่วยวัด)
- ข้อมูลเชิงปริมาณ (Quantitative data) คือข้อมูลที่มีค่าต่อเนื่อง (จำนวนจริง แต่ในทาง ปฏิบัติอาจจะบันทึกเป็นจำนวนเต็ม) มีหน่วยวัด (scale) แบ่งเป็น 2 ระดับคือ
  - Interval เป็นข้อมูลแบบช่วง มีหน่วยวัด เช่น คะแนน เป็นข้อมูลที่ไม่มีศูนย์สัมบูรณ์ มัก เป็นเครื่องมือวัดทางจิตวิทยา (psychological measurement) เครื่องมือสอบถามความ คิดเห็น ข้อสอบวัดต่าง ๆ หรือเครื่องมือวัด KPI ต่าง ๆ
  - Ratio เป็นข้อมูลแบบอัตราส่วน มีหน่วยวัด และมักเป็นเครื่องมือวัดทางกายภาพ (physical measurement) ที่เป็นสากล มีศูนย์สัมบูรณ์หรือศูนย์แท้ ทำให้สามารถ เปรียบเทียบค่าของข้อมูลแบบอัตราส่วนได้ เช่น อายุ (ปี) น้ำหนัก (กก.) ส่วนสูง (ซม.) เป็นตัน

<u>หมายเหตุ</u> ในการวิเคราะห์ข้อมูลทางสถิติ สถิติที่เลือกใช้จะขึ้นอยู่กับลักษณะของข้อมูล 3 ระดับเท่านั้น คือ Nominal, Ordinal, Scale (Interval/Ratio)

	nominal	ordinal	Scale (interval/ratio)
SPSS	nominal	ordinal	scale
Excel	text	text	number
R	character, factor	character, factor	numeric
Weka	nominal	nominal	numeric

Measure ในโปรแกรม SPSS, Excel, R, Weka เป็นดังนี้

Workshop 1: จงใส่ระดับการวัดของข้อมูล (Measure: nominal, ordinal, scale) ลงในตารางที่ 1

## สถิติและพารามิเตอร์

สถิติ (statistics) เป็นตัวเลขเป็นค่าที่คำนวณได้จากกลุ่มตัวอย่าง เป็นคุณลักษณะของกลุ่มตัวอย่าง (sample)

พารามิเตอร์ (parameter) เป็นคุณลักษณะของประชากร (population) ซึ่งเราไม่ทราบค่า สถิติ เป็น**ตัวประมาณค่า (estimator)** ของพารามิเตอร์

Workshop 2: จงเขียนสัญลักษณ์แทนสถิติและพารามิเตอร์ต่อไปนี้ และแสดงสูตรการคำนวณของค่าสถิติ (ตัวอย่างสัญลักษณ์: x̄, SD, r, μ, เป็นตัน)

	สถิติ (สัญลักษณ์และสูตรการคำนวณ)	พารามิเตอร์ (สัญลักษณ์)
ค่าเฉลี่ยหรือ Mean		
ค่าเบี่ยงเบนมาตรฐาน หรือ		
Standard deviation		
ค่าสัมประสิทธิ์สหสัมพันธ์		
หรือ Correlation		

## การหาค่าสถิติสำหรับตัวแปรเดียว

ตัวแปรเชิงคุณลักษณะ: ค่าความถี่ (จำนวนและร้อยละ) หรือ Frequency & percentage

	กราฟ: กราฟแท่ง (bar chart) หรือกราฟพาย (pie chart)
ตัวแปรเชิงปริมาณ:	ค่ากลาง และค่าการกระจาย หรือค่าเปอร์เซ็นไทล์ต่าง ๆ ได้แก่
	ค่าเฉลี่ย (Mean) และค่าเบี่ยงเบนมาตรฐาน (Standard deviation)
	ค่ามัธยฐาน (Median) และค่า Interquartile Range (Q3 – Q1)
	ค่าเปอร์เซ็นไทล์ (Percentile) ต่าง ๆ เช่น P10, P25=Q1, P50=Q2=Median, P75,
	P90 เป็นต้น
	กราฟ: ฮีสโตรแกรม (histogram) แสดงการแจกแจง (distribution) ของข้อมูล หรือ
	box plot (แสดงค่า Q1, Q2=Median, Q3)

Workshop 3: ใช้โปรแกรม R หาค่าสถิติเบื้องต้นของโจทย์ "การฝึกงานของนิสิต" ข้อมูลอยู่ในไฟล์ DataAll และสร้างตัวแปรใหม่

1. เปิดโปรแกรม R โดยดับเบิลคลิก 🚾 บน desktop มาทำงาน

2. เลือก Packages > Load package (รูปซ้าย) แล้วเลือก Rcmdr (รูปขวา)

RGui (32-bit)				
File Edit View Misc Pa	ckages Windows Help			
R Console	Load package Set CRAN mirror Select repositories		highr Hmisc KernSmooth knitr	
R version 3.1.2 Copyright (C) 201 Platform: i386-we	Install package(s) Update packages	computing	latticeExtra leaps Ime4 Imtest	
R is free software You are welcome to Type 'license()' or	and comes with ABSOLUTELY NO redistribute it under certain 'licence()' for distribution	WARRANTY. conditions. details.	markdown MASS Matrix matrixcalc methods	
R is a collaborativ Type 'contributors( 'citation()' on how	e project with many contribut )' for more information and to cite R or R packages in p	ors. ublications.	mgcv mime minqa multcomp	
<pre>Type 'demo()' for s 'help.start()' for Type 'q()' to quit &gt;  </pre>	ome demos, 'help()' for on-li an HTML browser interface to R.	ne help, or help.	nlme nloptr nnet parallel pbkrtest Rcmdr BrmdrMisc	

จะได้ R Commander ซึ่งเป็นเมนูที่ผูกฟังก์ชันการทำงานของโปรแกรมคำนวณค่าสถิติต่าง ๆ (เหมือน

เมนูของ SPSS) ทำให้ง่ายต่อการใช้งาน

R Commander			
File Edit Data Statisti	cs Graphs Models Distril	butions Tools Help	
💿 Data set: 🔲 < No	active dataset> 🛛 🖊 Edit da	ata set 💽 View data set	Model: E <no active="" model=""></no>
R Script R Markdown			
K Warkdown			
			_
∢ [			•
Output			Submit
			×
			· · · · · · · · · · · · · · · · · · ·
Massages			4
wessayes			
With the single-c	locument interface (S	SDI); see ?Command	er. E
			•
•			· · · ·

อ่านข้อมูล Excel ด้วยคำสั่ง Data > Import data > from Excel, ...

File Edit	Data Statistics Graphs Models Di	stributions Tools Help
R Script R	New data set Load data set Merge data sets	data set 🔯 View data set Model: 🗵 <
4	Import data Data in packages Active data set Manage variables in active data set	from text file, clipboard, or URL from SPSS data set from SAS xport file from Minitab data set from STATA data set
		from Excel, Access or dBase data set

# ตั้งชื่อ dataset เป็น **DataAll** แล้วกด OK



เลือกไฟล์ DataAll.xlsx จาก CD (เลือกไฟล์ type เป็น MS Excel 2007 file (.xlsx) แล้วเลือก Sheet

DataAll จะได้

R Commander		x
File Edit Data Statistics Graphs Models Distributions Tools Help		
Data set: DataAll Z Edit data set 🔯 View data set Model: 2 <1	No active mod	del>
R Script R Markdown		
<pre>library(RODBC, pos=14) DataAll &lt;- sqlQuery(channel = 1, select * from [DataAll\$]) &lt;</pre>	4	•
Output	Submit	
> library(RODBC, pos=14)		Â
> DataAll <- sqlQuery(channel = 1, select * from [DataAll\$	:1)	+
A messages	•	
with the single-document interface (SDI); see ?Commander. [3] NOTE: The dataset DataAll has 996 rows and 35 columns.		4 III +
•	+	

หน้าจอการทำงานของโปรแกรม R จะแบ่งเป็น 3 ส่วนคือ

- R Script เป็นคำสั่ง R จากเมนูที่เลือก
- Output แสดงผลลัพธ์การประมวลผล
- Messages แสดงข้อความสำหรับการทำงานนั้น (อ่านข้อมูลเข้า 996 rows 35 columns)

ถ้าเกิด error ข้อความก็จะเป็นสีแดง

R Markdown	<u>หมายเหตุ:</u> R Markdown เป็น feature ใหม่ใน
R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R. It combines the core syntax of markdown (an easy-to-write plain text format) with embedded R code chunks that are run so their output can be included in the final document. R Markdown — Dynamic Documents for R markdown, rstudio.com/	version นี้ (ยังไม่ได้ทดสอบการใช้งาน) ไม่ เกี่ยวกับการประมวลผลทางสถิติ
Nieu dete cet	

4. ดูข้อมูลโดยกดปุ่ม 🎑 View data set

ถ้าไม่มีข้อมูล (missing value) R จะแทนด้วยอะไร? .....

5. Clear window (ทำเป็นครั้งคราวตามต้องการ) โดยคลิกขวาที่ window ที่ต้องการ clear เลือก Clear

window

6. หาค่าสถิติเบื้องต้น (descriptive statistics) ของแต่ละตัวแปรเพื่อสรุปลักษณะของข้อมูล

*Qualitative variables:* GENDER, ORG\_SIZE, PAY, Letter, Stamp ด้วยการหาการแจงแจงความถี่ (จำนวนและร้อยละ) ด้วยคำสั่ง Frequency distributions และแสดงกราฟวงกลม ด้วยคำสั่ง Pie Chart

Quantitative variables: GPA, DAYS, HOURS, CP121, , CP444, WSK1, , WSK6, USK1, USK6 ด้วยการหาค่ากลางและค่าการกระจาย ด้วยคำสั่ง Numerical summaries

### สำหรับ Qualitative variables:

6.1 เลือกคำสั่ง Statistics > Summaries > Frequency distributions จะได้

R Frequency Distril	outions			×
Variables (pick one	or more)			
ORG_SIZE	Â			
PAY	-			
Chi-square go	odness-of-fit tes	t (for one varia	ble only)	
🔯 Help	🧄 Reset	🚽 ок	💥 Cancel	Apply

จะเห็นได้ว่ามีแค่ 3 ตัวแปร ที่เป็น factor (ใน R ตัวแปร character จะเป็น factor โดยอัตโนมัติ) ไม่ มีตัวแปร Letter และ Stamp เนื่องด้วยค่าของตัวแปรเป็นตัวเลข (จาก Excel) R เลยให้ชนิดเป็น numeric

6.2 แปลงตัวแปร Letter, Stamp ให้เป็น factor ด้วยคำสั่ง Data > Manages variables in active data set > Convert numeric variables to factors





ใส่ชื่อค่าของข้อมูล (Numeric value Level name) ดังภาพ แล้วกด OK

ลอง 🗟 View data set ดู จะได้

Q I	DataA	11												ē.	e,	
	SK_	Q5 W	sk_	Q6 .	USK_	Q1 U	sk_	Q2 US	5K_(	23 U	ISK_	Q4 1	USK_	Q5 L	ett	er Stamp Letter_F
1	4		4		4		3		4		4		4		4	0 0 ไม่มีใบรับรองฝึกงาน 🔺
2		3		3		3		3		4		3		1		1 มีใบรับรองฝึกงาน
3		4		2		3		2		2		3		2		1 1 มีใบรับรองฝึกงาน
4		4		4	1	4		4		4		4	ł	0		0 ไม่มีใบรับรองฝึกงาน
5		4		4		4		4		4		4		4		0 0 ไม่มีใบรับรองฝึกงาน
6		4		4		4		4		4		4		0		0 ไม่มีใบรับรองฝึกงาน
7	1	4	1		4	3	3	3			3		4		3	0 1 ไม่มีใบรับรองฝึกงาน
8		3		3		3		3		3		3		1		0 มีใบรับรองฝึกงาน
9	1	4	1		3	3	3	3			3		3		3	1 0 มีใบรับรองฝึกงาน
10		4		4		4		4		4		4		0		0 ไม่มีใบรับรองฝึกงาน
11	3		3		3		3		3		4		3		3	1 0 มิใบรับรองฝึกงาน
12		4		4		3		3		4		4		1		0 มีใบรับรองฝึกงาน
13	4		4		3		4		3		3		4		3	0 0 ไม่มีใบรับรองฝึกงาน
14		4		4		4		4		4		4		4		0 1 ไม่มีใบรับรองฝึกงาน
15		4		3		3		3		4		4		0		0 ไม่มีใบรับรองฝึกงาน
16		3		4		3		3		3		4		4		0 0 ไม่มีใบรับรองฝึกงาน
17		4		4		3		4		4		4		1		1 มิไบรับรองฝึกงาน
18		4		4		3		3		4		4		0		0 ไม่มีใบรับรองฝึกงาน
19	F.	NZ	7	1	A	NZ	7	NA		N	A	N	IA	N.	A	1 0 มิโบรีบรองฝึกงาน
20	3		4		4		3		3		4		4		4	0 0 เมม เบรบรองผ่องาน
21	3	. 1	1		4		3	4			4		4		4	0 0 เมม เบรบรองผกงาน
22		4		4		4		4		4		4		0		0 เมม เบรบรองผกงาน
23		2		2		3		2		3		3		0		0 เมม เบรบรองผกงาน
24		3		4	~	4		4		4		4		4		0 0 เมม เบรบรองผิกงาน
25		2			3	2		2		, 3			3	23		บ ⊥ เมม เบรบรองผกงาน ∩ ไม่อีในรับรองอื่อ หา
20		*		7		3		3		*		1		1		บ เมม เบวบวย∨ผก∨าน 1 มีในรับธองยืองอน
21		4		4		4		7		4		4		1		⊥ มเบวบวยงผกงาน 0 ไม่มีในต้นตวงยื่องวน
28		*		2		2		2		*		- *		~		∪ เมม เบรบรองศึกงาน 0. ไม่มีในธันธวงศึกงวน
30	1	3		2	4	~		3		3	4	3	4	· ·	4	0 0 ไม่บีในรับรองยืองวน
30	1 1		,		7	-	1				7		-		-	0 0 the transminister

ข้อมูลของแต่ละคอลัมน์จะไม่ตรงเนื่องจากการแสดงผลภาษาไทย

แปลงตัวแปร Stamp ให้เป็น factor โดยใช้ **Use Number** สำหรับ Factor Levels และแทนที่ตัวแปรเดิม (<same as variables>) เพื่อความรวดเร็ว หรือจะตั้งชื่อตัวแปรใหม่เป็น Stamp\_F <u>หมายเหต</u>ุ ถ้าเราต้องการใช้ตัวแปรเดิมที่เป็น numeric เราต้องสร้างตัวแปรใหม่ให้เป็น factor จะได้ มีตัวแปร 2 ตัว เลือกใช้ได้ตามความเหมาะสม

<u>หมายเหตุ</u> ใน R มี data type ได้หลายแบบ ได้แก่ scalars, vectors (numerical, character, logical), matrices, data frames, and lists. ซึ่ง data frame ก็คือชนิดข้อมูลที่เป็นตาราง (table หรือ relation) นั่นเอง คือแต่ละคอลัมน์จะเป็นตัวแปร ซึ่งจะมีชนิดเป็น numeric หรือ character (ซึ่งก็คือ string) ถ้า ข้อมูลเป็น character R จะถือเป็นnominal อัตโนมัติแล้วให้ type เป็น **factor** ส่วนตัวแปรที่เป็น numeric ถ้าจะให้เป็น factor ก็ต้องดำเนินการ ดังตัวอย่างข้างต้น 6.3 เลือกคำสั่ง Statistics > Summaries > Frequency distributions ใหม่ เลือกตัวแปรทั้งหมด (5 ตัว)



ν	1 2
จะ	ิด

R Scrip	Output
local({	counts:
.Table <- with(DataAll, table(Stamp_F))	Stamp_F
cat("\ncounts:\n")	0 1
print(.Table)	58 12
cat("\npercentages:\n")	
print(round(100*.Table/sum(.Table), 2))	percentages:
	Stamp_F
	0 1
	82.86 17.14
local({	counts:
.Table <- with(DataAll, table(GENDER))	GENDER
cat("\ncounts:\n")	ชาย หญิง
print(.Table)	40 32
cat("\npercentages:\n")	
print(round(100*.Table/sum(.Table), 2))	percentages:
})	GENDER
	ชาย หญิง
	55.56 44.44
<u>หมายเหตุ</u> หากข้อมูลมีมากกว่า 3 ค่า การแสดง v	alue label ที่เป็นภาษาไทยจะเรียงตาม
ตัวอักษร ่ถ้าข้อมูลเป็น Ordinal จะสับสน เพราะมั	ันจะไม่เรียงอันดับให้ หากต้องการให้เรียง
อันดับก็ควรใส่ 1_, 2_, 3_, ไว้ข้างหน้า	

6.4 แสดงกราฟพายด้วยคำสั่ง Graphs > Pie Chart ของตัวแปร Letter\_F ตาม default ได้



ให้ระวังและสังเกตว่าผลลัพธ์ของกราฟจะแสดงอยู่ใน RGUI ดังนั้นต้องยุบหน้าต่าง R Commnader ก่อน ลองเรียก Pie Chart อีกครั้ง

Variable (pick one)	- Plot Labels -	
GENDER	🚖 x-axis label	<auto></auto>
ORG_SIZE		< F
PAY	y-axis label	<auto></auto>
Stamp_F	Ŧ	<
	Graph title	<auto></auto>
		< >
Help	🥎 Reset	V OK Cancel Apply

### สำหรับ Quantitative variables:

6.5 เลือกคำสั่ง Statistics > Summaries > Numerical summaries แล้วเลือกกลุ่มวิชา CP…

ตาม default จะให้ค่า Statistics เป็น Mean, Standard Deviation, Quantiles: 0, .25, .5, .75, 1

R Numerical Summaries	R Numerical Summaries
Data Statistics	Data Statistics
Variables (pick one or more)	Mean
CP352	Standard Deviation
CP431	Standard Error of Mean
CP445	Interquartile Range
CP482	Coefficient of Variation
DAYS	Skewness 🔿 Type 1
Summarize by groups	🗌 Kurtosis 🔘 Type 2
	© Туре 3
	✓ Quantiles: 0, .25, .5, .75, 1
🚯 Help 🔦 Reset 🖌 OK 🗱 Cancel 🌈 Apply	🔞 Help 🦘 Reset 🖌 OK 🗱 Cancel 🎓 Apply

#### R Scrip:

numSummary(DataAll[,c("CP111", "CP121", "CP201", "CP212", "CP241", "CP251", "CP316", "CP322", "CP323", "CP342", "CP352", "CP431", "CP444", "CP445", "CP482")], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))

#### Output:

	mean	sd	IQR	0%	25%	50%	75%	100%	n
CP111	2.666667	1.3737991	3.00	1	1.00	3	4.00	5	72
CP121	2.041667	1.2830147	2.00	1	1.00	1	3.00	5	72
CP201	1.944444	0.9913554	1.25	1	1.00	2	2.25	5	72
CP212	2.888889	1.4394400	2.00	1	2.00	3	4.00	5	72
CP241	2.805556	1.2292089	2.00	1	2.00	3	4.00	5	72
CP251	3.750000	1.3188770	2.00	1	3.00	4	5.00	5	72
CP316	2.638889	1.3871194	3.00	1	1.00	3	4.00	5	72
CP322	2.055556	1.0991532	2.00	1	1.00	2	3.00	5	72
CP323	2.291667	1.1800871	2.00	1	1.00	2	3.00	5	72
CP342	3.388889	1.3692350	2.00	1	2.00	4	4.00	5	72
CP352	2.625000	1.2608682	2.00	1	2.00	3	4.00	5	72
CP431	2.777778	1.3862730	2.25	1	1.75	3	4.00	5	72
CP444	1.972222	1.1864520	2.00	1	1.00	1	3.00	5	72
CP445	2.083333	1.2189813	2.00	1	1.00	2	3.00	5	72
CP482	2.305556	1.1705168	2.00	1	1.00	2	3.00	5	72

7. สร้างตัวแปรตัวใหม่เพื่อวัด "ระดับความรู้ที่ใช้" ตั้งชื่อว่า **CP\_All** โดยเป็นค่าเฉลี่ยของ CP ทั้งหมด

เลือกคำสั่ง Data > Manage variables in active data set > **Compute new variable** 



ลอง View Data Set เพื่อดูค่าตัวแปรตัวใหม่ ต้องมีค่าอยู่ระหว่าง 1 – 5

8. ลอง plot กราฟ ของตัวแปรเชิงปริมาณ ด้วย histogram และ box plot

```
คำสั่ง: Graphs > Histogram (รูปซ้าย
```

```
หรือ Graphs > Boxplot (รูปขวา)
```



และหากจะสำเนากราฟก็คลิกขวาที่รูปแล้วเลือก Copy as bitmap

หาค่าสถิติเบื้องตันของ CP\_All คำสั่งที่ใช้ .....
 และให้ผลสรุป

22

Script & Output

ถ้าให้ความหมายของ Rating scale 5 ระดับ เป็น

1=น้อยที่สุด, 2=น้อย, 3=ปานกลาง, 4=มาก, 5=มากที่สุด ก็จะได้เกณฑ์การประเมิน 2 แบบ คือ แบบที่ 1: ใช้ค่ากลางในการให้ความหมาย จะได้เกณฑ์ดังรูปซ้าย ซึ่งช่วงของหัวท้ายจะไม่เท่าช่วงตรง กลาง

**แบบที่ 2:** ใช้แบบช่วงเท่าในการให้ความหมาย คือ แบ่ง max – min ออกเป็น 5 ช่วงเท่า ๆ กัน (ช่วงละ .8) จะได้เกณฑ์ดังรู**ปขวา** 

1.00 – 1.50 = น้อยที่สุด	1.00 – 1.80 = น้อยที่สุด
1.51 – 2.50 = น้อย	1.81 – 2.60 = น้อย
2.51 – 3.50 = ปานกลาง	2.61 – 3.40 = ปานกลาง
3.51 – 4.50 = มาก	3.41 – 4.20 = มาก
4.51 – 5.00 = มากที่สุด	4.21 – 5.00 = มากที่สุด

<u>แบบฝึกหัด</u> ลองสร้างตัวใหม่ตัวใหม่ เพื่อวัด "ด้านความรู้และทักษะการทำงาน" WSK และ "ด้านอัตลักษณ์นิสิตของมหาวิทยาลัย – มีทักษะการสื่อสาร" USK ว่าอยู่ในระดับใด

10. จัดเก็บผลลัพธ์การทำงานไว้ใช้ในโอกาสต่อไป

```
ใน R Commander: File > Save R Workspace as ดั้งชื่อไฟล์ว่า Training
```



## การทดสอบสมมุติฐานทางสถิติ

การทดสอบสมมุติฐานทางสถิติเป็นกระบวนสรุปความจริงเกี่ยวกับประชากร (population) โดยศึกษาจาก กลุ่มตัวอย่าง (sample) ที่เป็นตัวแทนของประชากร สถิติที่ใช้จะเรียกว่าสถิติอนุมาน (inferential statistics)

ขั้นตอนการทดสอบสมมุติฐานทางสถิติ (Hypothesis Testing Template)

1. สมมุติฐานที่ทดสอบ H₀ และ H₁ (คุณลักษณะของประชากร ต้องเป็นพารามิเตอร์เสมอ) H<sub>0</sub>: .....

H<sub>4</sub>: .....

- กำหนดระดับนัยสำคัญของการทดสอบ α = ...... (ปกติจะเป็น .05 หรือ .01 หรือ .10)
- เลือกสถิติที่เหมาะสมกับสมมุติฐานที่ทดสอบในข้อ 1 สถิตินี้คือสถิติอนุมาน/สถิติอ้างอิง และตัวสถิตินี้มีการแจกแจงความน่าจะเป็นแบบใด?

Z =(สูตร)	มีการแจกแจงแบบ Z (ปกติมาตรฐาน)
-----------	--------------------------------

- t
   = .....(สูตร).....
   มีการแจกแจงแบบ t ที่ df = ......

   X<sup>2</sup>
   = .....(สูตร)....
   มีการแจกแจงแบบ X<sup>2</sup> ที่ df = ......

   F
   = ......(สูตร)....
   มีการแจกแจงแบบ F ที่ df<sub>1</sub> = .... และ df<sub>2</sub> = .....
- 4. กฏการตัดสินใจ ปฏิเสธ H₀ ถ้า p-value < α
- 5. คำนวณค่าสถิติจากกลุ่มตัวอย่าง (โปรแกรมสำเร็จรูปจะให้ p-value ด้วย)

เช่น t = ..... p-value = .....

6. ตัดสินใจ ตีความและให้ความหมาย

ตัดสินใจ: ปฏิเสธ H<sub>o</sub> / ไม่ปฏิเสธ H<sub>o</sub>

์ตีความและให้ความหมาย .....

## อธิบายเพิ่มเติม

2. ระดับนัยสำคัญของการทดสอบ หรือ Type I error คือ ความน่าจะเป็นที่ปฏิเสธ H<sub>0</sub> แต่ H<sub>0</sub> เป็นจริง

	ความจริงเกี่ยวกับสมมุติฐานที่ทดสอบ (Actual/Model)							
การตัดสินใจ	H <sub>0</sub> เป็นจริง	H <sub>0</sub> เป็นเท็จ						
(Predicted/Test)	(Positive)	(Negative)						
<i>ไม่ปฏิเสธ <mark>H<sub>o</sub></mark> /</i> H <sub>o</sub> เป็นจริง	ตัดสินใจถูก	ตัดสินใจผิด (Type II error = $eta$ )						
(Positive)	(True Positive: TP)	(False Negative: FN)						
	Confidence Level $(1 - \alpha)$							
<i>ปฏิเสธ H<sub>o</sub> /</i> H <sub>o</sub> เป็นเท็จ	ตัดสินใจผิด (Type I error = α)	ตัดสินใจถูก						
(Negative)	(False Positive: FP)	True Negative: TN)						
		Power of the test $(1 - \beta)$						



3. การแจกแจงความน่าจะเป็นของสถิติทดสอบ และค่าวิกฤตที่ระดับ lpha = .05

4. กฎการตัดสินใจ

<u>แบบเดิม</u>: หาค่าวิกฤติ โดยเปิดตารางการแจกแจงของข้อ 3 เช่น สำหรับการแจกแจงแบบ Z ค่าวิกฤตคือ -1.96 และ 1.96 บริเวณวิกฤตคือ Z < -1.96 และ Z > 1.96 ปฏิเสธ H<sub>0</sub> ถ้า ค่าสถิติในข้อ 3 ตกอยู่ในบริเวณวิกฤติ (บริเวณปฏิเสธ) ซึ่งมีความน่าจะเป็น = **Q** <u>แบบใหม่</u>: ใช้ค่าพี (p-value) (คือความน่าจะเป็นจากค่าสถิติจนถึงหางของการแจกแจง) เช่น ถ้า Z = 2.00 p-value คือ P(Z >= 2.00) + P(Z <= 2.00) = 0.02275013 + 0.02275013 = 0.0455

 ถ้าการตัดสินใจเป็น ปฏิเสธ H<sub>0</sub> เรียกว่า การทดสอบมีนัยสำคัญ (the test is significant) แสดงว่า H<sub>0</sub> (ความรู้เดิม) ไม่จริง H<sub>1</sub> จริง (H<sub>1</sub> คือความรู้ใหม่ที่นักวิจัยคันพบ) ดังนั้นเป้าหมายของการวิจัยคือ การปฏิเสธ H<sub>0</sub> ซึ่งก็คือต้องการให้การทดสอบมีนัยสำคัญ ซึ่งต่างจากการทดสอบข้อตกลงเบื้องตัน (assumption) เช่น การทดสอบ Normal distribution หรือการ ทดสอบการเท่ากันของ Variance คือ<u>ไม่</u>ต้องการปฏิเสธ H<sub>0</sub> Workshop 4: การแจกแจงความน่าจะเป็นแบบต่าง ๆ การหาค่าวิกฤตที่ระดับนัยสำคัญต่าง ๆ และการหา

ค่าพี (p-value)

1. Plot การแจกแจงความน่าจะเป็นของการแจกแจงปกติมาตรฐาน N(0,1)

Distribution > Continuous distributions > Normal distribution > Plot normal distribution



2. หาค่าวิกฤต (ค่าสถิติ) ที่ระดับนัยสำคัญต่าง ๆ (α)

Distributtion > Continuous distributions > Normal distribution > Normal quantiles

R Normal Quantiles	
Probabilities	.05
Mean	0
Standard deviation	1
Lower tail	
O Upper tail	
🔞 Help	♦ Reset ✓ OK Cancel Apply

Probabilities คือค่า lpha (one tail)

ผลลัพธ์ที่ได้คือค่าวิกฤต (ค่าสถิติ)

#### Script & Output

> qnorm(c(0.05), mean=0, sd=1, lower.tail=TRUE)
[1] -1.644854

หากต้องการค่าวิกฤตของ lpha 2 tails ต้องหาร 2 คือต้องใส่ Probabilites = .025 จะได้ -1.959964

หรือ -1.96

ลองดำเนินการด้วยค่า lpha อื่น ๆ

<u>สรุป</u> P(Z < -1.96 and Z > 1.96) = 0.05

P(Z < -1.64 and Z > 1.64) = 0.10

P(Z < -2.58 and Z > 2.58) = 0.01

3. หาค่าพี (p-value) จากค่าสถิติที่คำนวณได้

Distributtion > Continuous distributions > Normal distribution > Normal probabilities

R Normal Probabilitie	s X
Variable value(s)	3.00
Mean Standard deviation	1
<ul> <li>Lower tail</li> <li>Upper tail</li> </ul>	
( Help	♦ Reset

ผลลัพธ์ที่ได้คือค่าพี (p-value) ของ**ค่าส**ถิติที่ใส่ใน Variable value(s)

#### Script & Output

```
> pnorm(c(3.0), mean=0, sd=1, lower.tail=FALSE)
[1] 0.001349898
```

ดังนั้น p-value ของ 3.00 one tail คือ 0.001349898 ถ้าต้องการ 2 tails ต้องคูณ 2 จะได้ 0.002699796

สามารถลองเล่นกับ distribution อื่น ๆ เช่น t distribution, Chi-squared distribution, F distribution

### การแจกแจงแบบปกติ

โดยธรรมชาติแล้วข้อมูลเชิงปริมาณส่วนใหญ่จะมีการแจกแจงแบบปกติ (Normal distribution) เช่น X1, X2, X3, .... เป็นค่าของข้อมูลของกลุ่มตัวอย่างที่มาสุ่มมาจากประชากรที่มีการแจกแจงปกติ เขียน สัญลักษณ์ X<sub>i</sub> ~ N (μ, σ<sup>2</sup>) และ ค่าเฉลี่ยของ X คือ Xี ซึ่งจะมีการแจกแจงแบบ N (μ, σ<sup>2</sup>/N)

การแจกแจงแบบปกติเป็นข้อตกลงเบื้องต้นของสถิติ**พาราเมตริก** (parametric statistics) หาก ข้อมูลไม่มีการแจกแจงแบบปกติก็ต้องไปใช้สถิติแบบ**นอนพาราเมตริก** (nonparametric statistics)

การทำ transformation ของข้อมูลอาจจะมีความจำเป็นในบางกรณีในกระบวนการวิเคราะห์ข้อมูล ด้วย data mining ที่สำคัญ 2 อย่างคือ การทำให้เป็นปกติมาตรฐาน (standardization) และการ normalization เพื่อให้ค่าของตัวแปรอยู่ในมาตรวัดเดียวกัน

การทำให้เป็นปกติมาตรฐานคือการลบด้วยค่าเฉลี่ยและหารด้วยค่าเบี่ยงเบนมาตรฐาน ค่าปกติ มาตรฐานจะมีค่าเฉลี่ยเป็น 0 และค่าเบี่ยงเบนมาตรฐานเป็น 1 ซึ่งหลักการนี้เป็นหลักการที่สำคัญในสถิติ อนุมาน

การ normalization คือการที่ทำให้ข้อมูลมีค่าอยู่ระหว่าง 0 และ 1 ทำได้โดยลบข้อมูลด้วย min และ หารด้วย max – min

```
Workshop 5 จงทดสอบว่า CP_All มีการแจกแจงปกติหรือไม่
```

```
Statistics > Summaries > Shapio-Wilk test for normality เลือก Variable > CP_All
```

#### Script & Output

### การทดสอบสมมุติฐานทางสถิติ

- 1. สมมุติฐานที่ทดสอบ H<sub>o</sub> และ H<sub>1</sub>
  - H₀: CP\_All มีการแจกแจงแบบปกติ
  - H₁: CP\_All ไม่มีการแจกแจงแบบปกติ
- กำหนดระดับนัยสำคัญของการทดสอบ **α** = 0.05
- 3. สถิติทดสอบ

Shapiro-Wilk test for normality

- 4. กฏการตัดสินใจ ปฏิเสธ H<sub>0</sub> ถ้า p-value <  $\alpha$
- 6ำนวณค่าสถิติจากกลุ่มตัวอย่าง
   ค่าสถิติ W = 0.9796 p-value = 0.292
- 6. ตัดสินใจ ตีความและให้ความหมาย ไม่ปฏิเสธ H<sub>0</sub>
   CP\_All มีการแจกแจงปกติ

# ทดสอบค่าเฉลี่ยของประชากร 1 กลุ่ม

```
การหา 95% ช่วงความเชื่อมั้นของ μ = X̄ ± t<sub>0.025</sub>*SD/sqrt(N)
```

Workshop 6: จงหา 95% ช่วงความเชื่อมั่นของค่าเฉลี่ยของ CP\_All (1 sample t-test)

คำสั่ง: Statistics > Mean > **Single-Sample t-test** ใส่ค่า Null hypothesis mu = 2.549074

R Single-Sample t-Test		×
Variable (pick one)		
CP431		
CP444		
CP445		
CP482		
CP_AII		
DAYS -		
Alternative Hypothesis		
Population mean != mu0	Null hypothesis: mu = 2.549074	
Population mean < mu0	Confidence Level: .95	
Population mean > mu0		
🔞 Help 🦘 Re	eset 🗸 OK 🎗 Cance	l 🔶 Apply

#### Script & Output

95% ช่วงความเชื่อมั่นของค่าเฉลี่ยของ CP\_All

จงทดสอบให้เห็นจริง โดยการแทนค่าในสูตร 2.549074 ± 1.993943\*(0.8238611/sqrt(72))

= 2.549074 ± 0.193598

## ทดสอบความแตกต่างของค่าเฉลี่ยของประชากร 2 กลุ่ม

Workshop 7: นิสิตหญิงและนิสิตชายประเมินการใช้ความรู้ที่เรียนในชั้นเรียนกับการทำงาน (CP\_All) แตกต่างกันหรือไม่

<u>หมายเหตุ 1.</u> ตัวแปรที่เราต้องการจะศึกษา/ต้องการที่จะทดสอบ (CP\_All) เรียกว่าตัวแปรตาม (Dependent variable) ตัวแปรกลุ่ม (GENDER) เรียกว่าตัวแปรตัน (Independent variable) หรือตัวแปรกลุ่ม (Group variable)

<u>หมายเหตุ 2.</u> ใน Data warehouse Measure attribute ก็คือ Dependent variable และ Dimension attribute ก็คือ Independent variable หรือ Group variable นั่นเอง

1. หาค่าสถิติพรรณนา Mean, SD, N ของ CP\_ALL ของกลุ่ม ชาย และ หญิง

คำสั่ง: Statistics > Summaries > Numerical Summaries (ซ้าย) คลิก Summarize by groups (ขวา)



#### Script & Output

```
> numSummary(DataAll[,"CP_All"], groups=DataAll$GENDER, statistics=c("mean",
+ "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
mean sd IQR 0% 25% 50% 75% 100% data:n
ซาย 2.660000 0.9011294 1.133333 1 2.20 2.533333 3.333333 4.6 40
หญิง 2.410417 0.7051494 0.750000 1 2.05 2.366667 2.800000 4.0 32
```

ให้สังเกตว่า ผลลัพธ์ข้างต้น ก็คือ เมเชอร์ 8 ตัว (mean, sd, IQR, 0%, 25%, 50%, 75%, 100%, data:n) ซึ่ง data:n aggregation function คือ count นั่นเอง ส่วนเมเชอร์ตัวอื่น ๆ ก็คือ CM (Calculated Member)

2. ทดสอบค่าเฉลี่ย 2 ค่า

<u>คำถาม แบบที่ 1</u>: Mean ของ CP\_All ของชาย (2.660000) และของหญิง (2.410417) แตกต่างกันอย่างมี นัยสำคัญหรือไม่?

<u>คำถาม แบบที่ 2</u>: Mean ของ CP\_All ของชาย และของหญิง แตกต่างกันหรือไม่?

ในแบบที่ 1 Mean คือ Sample Mean ( $\overline{X}_{
m yrg}$  และ  $\overline{X}_{
m kn ilde{u}}$ )

ในแบบที่ 2 Mean คือ Population Mean (µ<sub>ชาย</sub> และ µ<sub>หญิง</sub>)

้ค้นหาคำตอบด้วยการทดสอบ Independent samples t-test

# คำสั่ง: Statistics > Means > Independent samples t-test

R Independent Samples t-Test	R Independent Samples t-Test
Data Options	Data Options
Groups (pick one)     Response Variable (pick one)       GENDER     CP445       Letter_F     CP482       Stamp_F     CP_All       DAVS     GPA	Difference: ชาม - หญิง       Alternative Hypothesis     Confidence Level     Assume equal variances?       (a) Two-sided     .95     Yes       Difference < 0     (a) No       Difference > 0
HOURS -	Help Seset OK Cancel Apply

ซึ่ง Independent samples t-test มี 2 สูตร คือ สูตร Variances เท่า และ ไม่เท่า

ดังนั้นจึงต้องทดสอบ Variances ก่อน

3. ทดสอบ Variance 2 ค่า

### คำสั่ง: Statistics > Variances > Lavene's test

Statistics Graphs Models Distributions Tools He	มีสถิติให้เลือก 3 ตัว คือ
Summaries	1.Two-variances F-test
Means	2.Bartlett's test
Proportions	3.Levene's test
Variances  Two-variances F-test	   1 และ 2 มีสตรคำนวณปรากกในตำราสถิติ
Nonparametric tests  Bartlett's test	
Dimensional analysis      Levene's test	พื้นฐานทั่วไป
Fit models groups=DataAll\$GENDE	3. SPSS ใช้ตัวนี้
🕼 Levene's Test	×
Factors (pick one or more) Response Variable (pick one) GENDER CP445	
Letter_F CP482	
PAY DAYS	
Stamp_F	
Center	
median	
🔘 mean	
🚯 Help 🦘 Reset 🗸 OK 🗶 Cancel 🥐 Apply	Y

#### Script & Output

```
> leveneTest(CP_All ~ GENDER, data=DataAll, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
Df F value Pr(>F)
group 1 2.3079 0.1332
70
```

ทดสอบสมมุติฐานในใจ: p=value = 0.1332 แสดงว่าไม่ปฏิเสธ H₀, H₀ คือ Variances เท่า

ดังนั้น สรุปได้ว่า Variance เท่า

4. ทดสอบค่าเฉลี่ย 2 ค่า โดยใช้ Independent samples t-test สูตร Variance เท่า

คำสั่ง: Statistics > Means > Independent samples t-test (ใน Options: Assume equal variances? = Yes)

R Independent Samples t-Test	R Independent Samples t-Test
Data Options	Data Options
Groups (pick one)       GENDER     CP445       Letter_F     CP482       Stamp_F     CP41       DAYS     GPA       HOURS     +	Difference: ସୀଧ - Xญିଏ Alternative Hypothesis Confidence Level Assume equal variances? I Two-sided 95 I Yes Difference < 0 No Difference > 0
🔇 Help 🦘 Reset 🖌 OK 🗱 Cancel 🌈 Apply	🔞 Help 🦘 Reset 🖌 OK 🗱 Cancel 🌈 Apply

#### Script & Output

```
> t.test(CP_All~GENDER, alternative='two.sided', conf.level=.95,
+ var.equal=TRUE, data=DataAll)
Two Sample t-test
data: CP_All by GENDER
t = 1.2831, df = 70, p-value = 0.2037
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.1383594 0.6375260
sample estimates:
mean in group %78 mean in group MQN
2.660000 2.410417
```

### การทดสอบสมมุติฐานทางสถิติ

1. สมมุติฐานที่ทดสอบ H₀ และ H₁

H₀: μ<sub>ชาย</sub> = μ<sub>หญิง</sub> หรือ μ<sub>ชาย</sub> - μ<sub>หญิง</sub> = 0 เมื่อ μ คือค่าเฉลี่ยของ CP\_All

 $H_1$ :  $\mu_{ extsf{vre}} 
eq \mu_{ extsf{kn}}$ ง หรือ  $\mu_{ extsf{vre}}$  -  $\mu_{ extsf{kn}} 
eq 0$ 

- กำหนดระดับนัยสำคัญของการทดสอบ α = 0.05
- 3. สถิติทดสอบ

Independent samples t-test

$$t = \frac{(\bar{x}1 - \bar{x}2)}{\sqrt{\frac{(n1 - 1)s1^2 + (n2 - 1)s2^2}{n1 + n2 - 2}}} \quad \vec{u} \text{for sum on use t } \vec{n} \text{ df} = 40 + 32 - 2 = 70 \text{ is } \vec{u} \text{ o } \text{H}_0 \text{ is } \vec{u} \text{ o } \vec{s} \text{ o }$$

- 4. กฏการตัดสินใจ ปฏิเสธ H $_{0}$ ถ้า p-value < lpha
- 6ำนวณค่าสถิติจากกลุ่มตัวอย่าง
   ค่าสถิติ t = 1.2831 p-value = 0.2037
- 6. ตัดสินใจ ตีความและให้ความหมาย

ไม่ปฏิเสธ H₀

<mark>คำตอบแบบที่ 1</mark>: Mean ของ CP\_All ของชาย และของหญิง แตกต่างกันอย่าง<u>ไม่</u>มีนัยสำคัญ <mark>คำตอบแบบที่ 2</mark>: Mean ของ CP\_All ของชาย และของหญิง <u>ไม่</u>แตกต่างกัน

<u>หมายเหตุ</u>Independent samples t-test\_เป็นสถิติพาราเมตริก ซึ่งมีข้อตกลงเบื้องต้นว่าข้อมูลของ ประชากร 2 กลุ่มต้องมีการแจกแจงปกติ หากข้อตกลงเบื้องต้นนี้ไม่เป็นจริง ต้องไปใช้สถิตินอนพาราเมตริก ที่คู่กัน ด้วยคำสั่ง Statistics > Nonparametric tests > **Two-sample Wilcoxon test** การดำเนินการและการแปลผลใช้หลักการเหมือนกัน

## ทดสอบความแตกต่างของค่าเฉลี่ยของประชากร 3 กลุ่ม

Workshop 8: นิสิต 3 กลุ่ม ที่มี GPA 3 ระดับคือ ต่ำ (< 2.50) ปานกลาง (2.50 – 2.99) สูง (>= 3.00) ประเมิน การใช้ความรู้ที่เรียนในชั้นเรียนกับการทำงาน (CP\_All) แตกต่างกันหรือไม่ (One Way ANOVA)

 สร้างตัวแปรตัวใหม่ (transformation) ชื่อ GPA\_GP โดยแปลงค่า GPA ที่มีค่าต่อเนื่องให้มีค่าเป็น 3 กลุ่ม (discretization) ตามโจทย์

คำสั่ง: Data > Manages variables in active data set > **Recode variables** 

Recode Variables	a Canada a	1.000	×			
Variables to recode (pick one or more)						
CP482						
CP_AII						
DAYS	=					
GENDER						
GPA						
HOURS	Ŧ					
New variable name or pr	refix for multiple recodes:	GPA_GP				
📝 Make (each) new va	riable a factor					
Enter recode directives						
1.00:2.49 = "1 ต่	า"	*				
2.50:2.99 = "2 ก	ลาง"					
3.00:4.00 = "3]	v"					
· · · · · · · · · · · · · · · · · · ·						
		-				
<						
		•				
۲ الم	Reset V OK	Cancel	Apply			
<	🖌 Reset 🛛 🗸 OK	Cancel	Apply			

<u>หมายเหตุ</u> หากไม่ทราบว่าจะ Enter recode directives อย่างไร ให้กดปุ่ม Help

#### Script

```
> DataAll <- within(DataAll, {
+ GPA_GP <- Recode(GPA,
+ '1.00:2.49 = "1_min"; 2.50:3.00 = "2_nanv"; 3.00:4.00 = "3_av"; ; ;',
+ as.factor.result=TRUE)
+ })</pre>
```

#### Output (View data set)

R	DataA						
	SK1	USK2	USK3	USK4 U	ISK5 Le	tter	Stamp Letter_F Stamp_F CP_All GPA_GP
1	4	3	4	4	4	4	0 0 ไม่มีใบรับรองฝึกงาน 0 2.200000 3_สูง 🔺
2	3	3	4	3	1	1	มีใบรับรองฝึกงาน 1 3.333333 1_ต่า
3	3	3 2	2 2	3	2	1	L 1 มีใบรับรองฝึกงาน 1.1.000000 1_ตำ
4	4	4	4	4	0	0	) ไม่มีใบรับรองฝึกงาน 0 2.333333 3_สูง
5	4	4	4	4	4		0 ใม่มิใบรบรองฝึกงาน 03.33333333_สูง
6	4	4	4	4	0	0	เมม เบรบรองผลงาน 0 2.066667 2_กลาง
7	4	3	3	3	4 3		0 1 ไมม เบรบรองผ่องาน 1 3.866667 2_อลาง
8	3	3	3	3	1	0	มเบรบรองผกงาน 02.533333 1_ตา
9	3	3	3	3	3 3	0	1 0 มเบรบรองผกงาน 02.466667 1_ตา ไม่มีในสัมธุรณใจงาน 0.2.522222 1 ต่ำ
	7	-	-	-	2	2	เมม เบรบรองผกง 1 0 มีในสับสวน 0 2.555555 ⊥_พ1
	1	ູ	4	4	1	٠ ١	บใบรับรองป้องวน 0.1.333333.1 ต่ว
	Ľ,	4	3	3	4	ູ້	0 0 ไม่มีในรับรองยืองาน 0 1 400000 1 ต่ำ
14	<b>1</b>	4	4	4	4	Č	1 ไม่มีใบรับรองฝึกงาน 1 2.533333 2 กลาง
15	3	3	4	4	0	0	ไม่มีใบรับรองฝึกงาน 0 2.466667 2 กลาง
16	3	3	3	4	4	0	0 ไม่มีใบรับรองฝึกงาน 0 1.400000 2 กลาง
17	3	4	4	4	1	1	มีใบรับรองฝึกงาน 1 2.200000 3 สูง
18	3	3	4	4	0	0	ไม่มีใบรับรองฝึกงาน 0 3.933333 2 กลาง
19	AI (	NA	NA	NA N	IA NA		1 0 มีใบรับรองฝึกงาน 04.0000002_กลาง
20	4	3	3	4	4	4	0 0 ไม่มีใบรับรองฝึกงาน 0 1.666667 3_สูง
21	4	3	4	4	4 4		0 0 ไม่มีใบรับรองฝึกงาน 0 1.866667 2_กลาง
22	4	4	4	4	0	0	ไม่มีใบรับรองฝึกงาน 0 2.200000 2_กลาง
23	3	2	3	3	0	0	ไม่มีใบรับรองฝึกงาน 0 3.200000 2_กลาง
24	4	4	4	4	4	0	0 ใมมิใบรับรองฝึกงาน 0 2.133333 2_กลาง
25	3	3	2	3	3 3		0 1 ไมม เบรบรองฝกงาน 1 2.333333 3_สูง
20	3	3	4	4	0	0	เม่ม เปรียรองผู้กงาน 0 1.133333 2_กลาง
	4	4	4	4	1	1	มเบรบรองผกงาน ⊥ 3.266667 3_ลูง ไม่มีในสัมธวงปีองวน 0.1 666667 1 ตัว
20	2	3	2	3	0	~ ~ `	เมม เบรบรองศึกง ∩น 0 ±.000007 ±.พ.เ ไม่มีในรับรองศึกงาน 0 1 000000 1 ต่ำ
30	4	4	4	4	4 4	0	0 0 ไม่บี้ในรับรองยืองวน 0.2.266667 3 สง –
	4	*		•			

#### 2. ตรวจสอบผลการ Discritization

คำสั่ง: Statistics > Summaries > Frequency distributions

```
> local({
    .Table <- with(DataAll, table(GPA_GP))
+
+
   cat("\ncounts:\n")
  print(.Table)
+
+
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
+
+ })
counts:
GPA GP
1_ต่ำ 2_กลาง 3_สูง
17 33 22
percentages:
GPA GP
1_ต่ำ 2_กลาง 3_สูง
23.61 45.83 30.56
```

หาค่าสถิติพรรณนา Mean, SD, N ของ CP\_ALL ของ 3 กลุ่ม (GPA\_GP)
 คำสั่ง: Statistics > Summaries > Numerical Summaries Variables: เลือก CP\_All,
 Summarize by groups: เลือก GPA\_GP, Statistics เลือก Mean, Standard Deviation

#### Script & Output

4. One Way ANOVA มีข้อตกลงเบื้องตันว่า Variance ของ 3 กลุ่มต้องเท่ากัน ถ้าไม่เท่าต้องไปใช้

Nonparametric ที่คู่กัน

ผลการทดสอบ Variance เท่า ดังนั้นจึงใช้ One Way ANOVA ได้ (ลองทดสอบ)

คำสั่ง: Statistics > Means > **One-way ANOVA** (เลือก **Pairwise comparisons of means**)

R One-Way Analysis of	Variance	a la constante en	×			
Enter name for model:	AnovaModel.3	]				
Groups (pick one)	Response Variable	(pick one)				
GENDER	▲ CP444	*				
GPA GP	CP445					
Letter_F	CP482	=				
ORG_SIZE	CP_AII					
PAY	DAYS					
Stamp_F		T				
Pairwise comparisons of means						
🔞 Help 🦘 Reset 🗸 OK 🎇 Cancel 🦽 Apply						

#### Script & Output

```
95% family-wise confidence level
   2_กลาง - 1_ต่ำ
    3_สุง-1_ต่ำ =
   3_สุง - 2_กลาง -
         -0.5
                 0.0
                         0.5
                                 10
                  Linear Function
> local({
  .Pairs <- glht(AnovaModel.1, linfct = mcp(GPA_GP = "Tukey"))
+
+
  print(summary(.Pairs)) # pairwise tests
  print(confint(.Pairs)) # confidence intervals
+
+
  print(cld(.Pairs)) # compact letter display
  old.oma <- par(oma=c(0,5,0,0))</pre>
+
  plot(confint(.Pairs))
+
   par(old.oma)
+
+ })
         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = CP_All ~ GPA_GP, data = DataAll)
Linear Hypotheses:
                     Estimate Std. Error t value Pr(>|t|)
2 กลาง – 1 ต่ำ == 0 0.30838 0.24617 1.253
                                                   0.425
3 สูง – 1 ต่ำ == 0 0.32050 0.26627 1.204
                                                    0.454
3 สูง – 2 กลาง == 0 0.01212
                              0.22696 0.053
                                                   0.998
(Adjusted p values reported -- single-step method)
```
```
Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = CP_All ~ GPA_GP, data = DataAll)

Quantile = 2.3927

95% family-wise confidence level

Linear Hypotheses:

Estimate lwr upr

2_nanv - 1_min == 0 0.30838 -0.28064 0.89740

3_qv - 1_min == 0 0.32050 -0.31663 0.95762

3_qv - 2_nanv == 0 0.01212 -0.53093 0.55517

1_min 2_nanv 3_qv

"a" "a" "a"
```

## การทดสอบสมมติฐานทางสถิติ

- สมมุติฐานที่ทดสอบ H₀ และ H₁
   H₀: µ₁ = µ₂ = µ₃ เมื่อ µᵢ คือค่าเฉลี่ยของ CP\_All ของ GPA\_GP แต่ละกลุ่ม (ต่ำ, กลาง, สูง)
   H₁: มี µᵢ ≠ µⱼ อย่างน้อย 1 คೖ (มีกี่คೖ่?)
- กำหนดระดับนัยสำคัญของการทดสอบ **α** = 0.05
- 3. สถิติทดสอบ

One way ANOVA (ANalysis Of Variances)

F =  $\frac{MS_b}{MS_w} = \frac{SS_b/(k-1)}{SS_w/(N-k)}$ ; มีการแจกแจงแบบ F ที่ df = 3-1=2 และ df = 72-3 =69 (~F<sub>2,69</sub>) เมื่อ H<sub>0</sub> เป็นจริง

- 4. กฏการตัดสินใจ ปฏิเสธ H<sub>0</sub> ถ้า p-value <  $\alpha$
- 5. คำนวณค่าสถิติจากกลุ่มตัวอย่าง

ค่าสถิติ F = 0.967 p-value = 0.385

6. ตัดสินใจ ตีความและให้ความหมาย

## ไม่ปฏิเสธ H₀

้ค่าเฉลี่ยของ CP\_All ของนิสิต 3 กลุ่ม (GPA\_GP) ไม่แตกต่างกัน

ซึ่งสอดคล้องกัยการเปรียบเทียบรายคู่ (Miltiple Cpmparisons of Means) ด้วยวิธี Tukey ไม่มีคู่ใดที่แตกต่างกัน ดังนั้นค่าเฉลี่ยของทุกกลุ่มจะอยู่ใน set เดียวกัน (สัญลักษณ์ "a")

<u>หมายเหต</u>ุ One-way ANOVA เป็นสถิติพาราเมตริก ซึ่งมีข้อตกลงเบื้องต้นว่าข้อมูลของประชากรทุกกลุ่ม ต้องมีการแจกแจงปกติ และความแปรปรวนของประชากรทุกกลุ่มต้องเท่ากัน หากข้อตกลงเบื้องต้นนี้ไม่ เป็นจริง ต้องไปใช้สถิตินอนพาราเมตริกที่คู่กัน ด้วยคำสั่ง Statistics > Nonparametric tests > Kruskal-Wallis test การดำเนินการและการแปลผลใช้หลักการเหมือนกัน

## ทดสอบความเป็นอิสระของตัวแปรเชิงคุณลักษณะ

Workshop 9: การออกใบรับรองการฝึกงาน (Letter) เป็นอิสระหรือสัมพันธ์กับขนาดขององค์กรที่ฝึกงาน (ORG\_SIZE) (Chi-square test for independent)

1. หาค่าสถิติพรรณนา จะได้เป็นตารางแจกแจงความถี่ 2 ทาง (Cross tab หรือ Table)

คำสั่ง: Statistics > Contingency tables > Two-way tables

Two-Way Table	Two-Way Table
Data Statistics	Data Statistics
Row variable (pick one) Column variable (pick one)	Compute Percentages
GENDER A GENDER A	Row percentages
GPA_GP GPA_GP	Column percentages
Letter_F ORG_SIZE ORG_SIZE	Percentages of total
PAY PAY	No percentages
Stamp_F v Stamp_F v	Hypothesis Tests
Subset expression	Chi-square test of independence
<all cases="" valid=""></all>	Components of chi-square statistic
	Print expected frequencies
	Fisher's evact test
🔞 Help 🔸 Reset 🖌 OK 🗱 Cancel Apply	🔞 Help 🧄 Reset 🗸 OK 🗱 Cancel 🌈 Apply

### Script & Output

> local({			
+ .Table <- xtabs(~ORG SI	IZE+Letter F, data=DataAll)		
+ cat("\nFrequency table;	:\n")		
+ print(.Table)			
+ cat("\nRow percentages;	:\n")		
+ print (rowPercents (.Tab)	Le))		
+ .Test <- chisg.test(.Ta	able, correct=FALSE)		
+ print(.Test)			
+ cat("\nExpected counts;	:\n")		
+ print(.Test\$expected)			
+ })			
Frequency table:			
	Letter_F		
ORG_SIZE	ไม่มีใบรับรองฝึกงาน มีใบรับรองฝึกงาน		
ขนาดกลาง (15 – 50 คน)	10	1	
ขนาดเล็ก (น้อยกว่า 15 คน)	7	1	
ขนาดใหญ่ (มากกว่า 50 คน)	29	23	
Row percentages:			
	Letter_F		
ORG_SIZE	ไม่มีใบรับรองฝึกงาน มีใบรับรองฝึกงาน	Total Count	
ขนาดกลาง (15 – 50 คน)	90.9	9.1 100	11
ขนาดเล็ก (น้อยกว่า 15 คน)	87.5	12.5 100	8
ขนาดไหญ่ (มากกว่า 50 คน)	55.8	44.2 100	52

```
Pearson's Chi-squared test
data: .Table
X-squared = 6.9529, df = 2, p-value = 0.03092
Expected counts:
                           Letter F
ORG SIZE
                            ไม่มีใบรับรองฝึกงาน มีใบรับรองฝึกงาน
                                      7.126761 3.873239
 ขนาดกลาง (15 – 50 คน)
  ขนาดเล็ก (น้อยกว่า 15 คน)
                                  5.183099
                                                    2.816901
  ขนาดใหญ่ (มากกว่า 50 คน)
                                    33.690141
                                                     18.309859
Messages
[5] WARNING:
2 expected frequencies are less than 5
```

<u>หมายเหตุ</u> ให้สังเกต ใน WARNING: สำหรับการทดสอบไค-สแคว์ ค่าความถี่ควรต้องมีจำนวน > 5

39

## การทดสอบสมมติฐานทางสถิติ

1. สมมุติฐานที่ทดสอบ H<sub>0</sub> และ H<sub>1</sub>

H<sub>0</sub>: Letter และ ORG\_SIZE เป็นอิสระต่อกัน (คือไม่มีความสัมพันธ์กัน) H<sub>1</sub>: Letter และ ORG\_SIZE ไม่เป็นอิสระต่อกัน (คือมีความสัมพันธ์กัน)

- กำหนดระดับนัยสำคัญของการทดสอบ α = 0.05
- 3. สถิติทดสอบ: Chi-square test for independence

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}; E_i = np_i$$
 มีการแจกแจงแบบ  $\chi^2$  ที่ df = 2 เมื่อ H<sub>o</sub> เป็นจริง

- 4. กฏการตัดสินใจ ปฏิเสธ H<sub>0</sub> ถ้า p-value < lpha
- 6ำนวณค่าสถิติจากกลุ่มตัวอย่าง
   ค่าสถิติ X-squared = 6.9529 p-value = 0.03092
- 6. ตัดสินใจ ตีความและให้ความหมาย

ปฏิเ**สธ H**₀

การออกใบรับรองการฝึกงาน (Letter) ไม่เป็นอิสระหรือมีสัมพันธ์กับขนาดขององค์กรที่ฝึกงาน (ORG\_SIZE)

2. Save active data set ไว้ใช้งานต่อไป

คำสั่ง: Data > Active data set > **Save active data set** 

## ทดสอบความสัมพันธ์ของตัวแปรเชิงปริมาณ

ความรู้และทักษะในการปฏิบัติงาน (WSK), อัตลักษณ์นิสิต "ทักษะการสื่อสาร" (USK), GPA มี ความสัมพันธ์กันหรือไม่ (Correlation)

Pearson correlation สำหรับข้อมูลที่มีการแจกแจงปกติ

Spearman rank correlation สำหรับข้อมูลที่ไม่มีการแจกแจงปกติ

## Workshop 10:

- 1. Load data set Training2 มาใช้งาน (มีตัวแปร WSK และ USK แล้ว)
- ทดสอบว่า WSK, USK, GPA มีการแจกแจงปกติหรือไม่ ผลการทดสอบ: WSK, USK <u>ไม่</u>มีการแจกแจงปกติ, GPA มีการแจกแจงปกติ
- 3. หาความสัมพันธ์

## คำสั่ง: Statistics > Summaries > Correlation test

Correlation Test	
Variables (pick two)	
GPA 🔺	
HOURS	
ID	
Letter	
Stamp	
USK -	
Type of Correlation	Alternative Hypothesis
Pearson product-moment	Two-sided
Spearman rank-order	Correlation < 0
🔘 Kendall's tau	Correlation > 0
🔞 Help 🦘 Res	et 🗸 OK 🎇 Cancel 🌈 Apply

## Script & Output

```
> with(DataAll, cor.test(GPA, USK, alternative="two.sided", method="spearman"))
        Spearman's rank correlation rho
data: GPA and USK
S = 51671.69, p-value = 0.4295
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
0.0959376
```

```
> with(DataAll, cor.test(GPA, WSK, alternative="two.sided", method="spearman"))
        Spearman's rank correlation rho
data: GPA and WSK
S = 42469.79, p-value = 0.1219
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.189415
> with(DataAll, cor.test(USK, WSK, alternative="two.sided", method="spearman"))
         Spearman's rank correlation rho
data: USK and WSK
S = 11932.41, p-value = 1.246e-14
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7722563
สรุป: ตัวแปรที่สัมพันธ์กันมีเพียงคู่เดียวคือ WSK และ USK
```

สัมประสิทธ์สหสัมพันธ์ (Correlation coefficient) = 0.7722563

เมื่อทดสอบสมมุติฐานทางสถิติจะสรุปได้ว่า ความสัมพันธ์ไม่เป็นศูนย์

4. ทดสอบว่าค่าเฉลี่ย WSK กับ USK แตกต่างกันหรือไม่ (Paired t-test)

```
คำสั่ง: Statistics > Means > Paired t-test
```

R Paired t-Test			1	×
Data Options				
First variable (p	ick one) Sec	ond variable (pic	k one)	
Stamp	🔺 USK	5	*	
USK	WS	<		
USK1	WSI	a		
USK2	≡ WS	(2	=	
USK3	WS	З	-	
USK4	+ WS	(4	-	
(A) Help	👆 Reset	и ок	Cancel	Apply
- Contraction	, Meser			( · Appy
·				

### Script & Output

สรุป ทักษะไม่แตกต่างกัน

<u>หมายเหตุ</u> Paired t-test\_เป็นสถิติพาราเมตริก ซึ่งมีข้อตกลงเบื้องดันว่าผลต่างของข้อมูล (paired) ของ ประชากร ต้องมีการแจกแจงปกติ หากข้อตกลงเบื้องต้นนี้ไม่เป็นจริง ต้องไปใช้สถิตินอนพาราเมตริกที่คู่กัน ด้วยคำสั่ง Statistics > Nonparametric tests > Paired-samples Wilcoxon test การดำเนินการและการแปลผลใช้หลักการเหมือนกัน

## การพยากรณ์หรือการวิเคราะห์การถดถอย (Regression)

การวิเคราะห์การถดถอยเป็นการพยากรณ์ตัวแปรที่สนใจ (ตัวแปรตาม/dependent variable 1 ตัว) หรือเป็นการอธิบายความแปรปรวนของตัวแปรที่สนใจ ด้วยตัวพยากรณ์หรือตัวอธิบาย (predictor/explainatory/independent variables หลายตัว) โดยที่ตัวแปรตันและตัวแปรตามต้องมี ความสัมพันธ์กันแบบเชิงเส้นอย่างมีนัยสำคัญ ตัวแปรทั้งหมดต้องมีระดับการวัดเป็น scale หากเป็นตัวแปร nominal ต้องมี 2 ค่าเท่านั้น หากมีมากกว่า 2 ค่าต้องมีการทำเป็น dummy variables วิธีการนี้เรียกว่า Multuple Linear Regression และมีข้อตกลงเบื้องตันว่าตัวแปรตามที่ระดับต่าง ๆ ของตัวแปรตันต้องมีการ แจกแจงแบบปกติ และมีความแปรปรวนเท่ากัน หากตรวจสอบข้อตกลงเบื้องต้นนี้ไม่เป็นจริงอาจจะต้องทำ การ Transformation ข้อมูล

หากตัวแปรตามเป็น nominal จะต้องใช้วิธี Logistic Regression

Workshop 11: หา Linear Regression Model ของ CPU data set (จาก Weka data set)

- 1. อ่านไฟล์ cpu.xls เข้ามาด้วยคำสั่ง .....
- View data set จะได้

R	PU				s.,			<u> </u>	โจท	เย้ สร้างโมเดลหรือสมการพยากรณ์
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	class		٩١٩٩	สิทธิกาพของการประบาลแลของ CPU
1	125	256	6000	256	16	128	198	*	1190	
2	29	8000	32000	32	8	32	269		ດ້າະ	เดกเล้กษณะของดอมพิวเตอร์
3	29	8000	32000	32	8	32	220	=	¥1 0 L	
4	29	8000	32000	32	8	32	172		Der	dondont variable da alaga
5	29	8000	16000	32	8	16	132		Deb	
6	26	8000	32000	64	8	32	318		(9)5	*สิทธิ์ภาพ)
7	23	16000	32000	64	16	32	367		(п а	∞ (N K I L N I )
8	23	16000	52000	64	16	32	489		_	
10	23	32000	64000	129	30	52	1144		Pre	dictor variables: MYC1, MMIN,
11	400	1000	3000	120	1	2	38		N 4 N /	
12	400	512	3500	4	1	6	40		IVIIV	IAX, CACH, CHMIN, CHMAX ที่งหมด 6
13	60	2000	8000	65	1	8	92		ຕັ້ວ	
14	50	4000	16000	65	1	8	138		٩IJ	
15	350	64	64	0	1	4	10		d v	<u>چ</u>
16	200	512	16000	0	4	32	35		มขอ	อมูลทงหมด 209 รายการ
17	167	524	2000	8	4	15	19			-
18	143	512	5000	0	7	32	28		ตัวเ	เปรทั้งหมดเป็น scale
19	143	1000	2000	0	5	16	31			
20	110	5000	5000	142	8	64	120		คำเ	าาม· ตัวแปรทกตัวมีการแจกแจงแบบปกติ
21	143	1500	6300	0	5	32	30			
22	143	3100	6200	0	5	20	33		หรือ	าไม่?
23	143	2300	6200	0	6	64	61			
24	110	3100	6200	0	6	64	76		ല്	prodictor แต่ละตัวปีความสัมพับธ์เชินสับ
25	320	128	6000	0	1	12	23		AI 9	
26	320	512	2000	4	1	3	69		กับเ	ข้าแปร dass หรือไม่?
27	320	256	3000	0	1	6	33			VI 4 16 1 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
20	320	200	5000	- 4	1	5	21			
30	320	256	5000	4	1	6	27	-		

3. ทดสอบการแจกแจงปกติ สมมติว่าทุกตัวมีการแจกแจงแบบปกติ

4. ทดสอบว่าตัวแปร class มีความสัมพันธ์กับตัว predictor แต่ละตัวแบบเชิงเส้น

## คำสั่ง: Graphs > **Scratterplot**

R Scatterplot	Zamme 3	-	-		x
Data Options					
x-variable (pick one) CACH CHMAX CHMIN E class MMAX MMIN T	y-variable (pick one) CHMAX CHMIN class MMAX MMIN MYCT	E			
Plot by groups Subset expression <all cases="" valid=""></all>					
🔞 Help 🧄 Re	set	🚽 ок	X Cano	:el 🔶 A	pply

### Script & Output

```
> showData(CPU, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80,
+ maxheight=30)
> scatterplot(class~CACH, reg.line=lm, smooth=FALSE, spread=FALSE,
+ id.method='mahal', id.n = 2, boxplots='xy', span=0.5, data=CPU)
1 200
1 200
```



5. สมมติตัวแปร predictor ทุกตัวมีความสัมพันธ์แบบเชิงเส้นกับ class

หาความสัมพันธ์ระหว่าง predictors แต่ละตัวกับ class

Predictor	r	t, df	p-value
CASH	0.6626414	12.7297, 207	< 2.2e-16
CHMAX	0.6052093	10.9381, 207	< 2.2e-16
CHMIN	0.6089033	11.044, 207	< 2.2e-16
MMAX	0.8630041	24.5775, 207	< 2.2e-16
MMIN	0.7949313	18.8513, 207	< 2.2e-16
МҮСТ	-0.3070994	-4.6427, 207	= 6.102e-06

<u>หมายเหตุ</u>  $t = rac{r * \sqrt{N-2}}{1-r^2}$ 

โดยปกติแล้วตัว predictors ก็จะมีความสัมพันธ์กันเอง

ลองหาความสัมพันธ์ของตัวแปรทั้งหมดด้วย

คำสั่ง: Statistics > Summaries > Correlation matrix เลือก Pairwise p-value



Script & Output

```
> rcorr.adjust(CPU[,c("CACH","CHMAX","CHMIN","class","MMAX","M
   type="pearson", use="complete")
 Pearson correlations:
          CACH CHMAX CHMIN class
                                             MMAX MMIN
                                                                MYCT
CACH 1.0000 0.4878 0.5822 0.6626 0.5380 0.5347 -0.3210
CHMAX 0.4878 1.0000 0.5483 0.6052 0.5272 0.2669 -0.2505
CHMIN 0.5822 0.5483 1.0000 0.6089 0.5605 0.5172 -0.3011
class 0.6626 0.6052 0.6089 1.0000 0.8630 0.7949 -0.3071
MMAX 0.5380 0.5272 0.5605 0.8630 1.0000 0.7582 -0.3786
MMIN 0.5347 0.2669 0.5172 0.7949 0.7582 1.0000 -0.3356
MYCT -0.3210 -0.2505 -0.3011 -0.3071 -0.3786 -0.3356 1.0000
 Number of observations: 209
 Pairwise two-sided p-values:
      CACH CHMAX CHMIN class MMAX MMIN MYCT
CACH <.0001 <.0001 <.0001 <.0001 <.0001 <.0001 <.0001 CHMAX <.0001 <.0001 <.0001 <.0001 <.0001 0.0003
CHMIN <.0001 <.0001 <.0001 <.0001 <.0001 <.0001

        class <.0001 <.0001 <.0001</td>
        <.0001 <.0001 <.0001</td>

        MMAX <.0001 <.0001 <.0001 <.0001</td>
        <.0001 <.0001</td>

MMIN <.0001 <.0001 <.0001 <.0001 <.0001
                                                       <.0001
MYCT <.0001 0.0003 <.0001 <.0001 <.0001 <.0001
```

6. หา Regression Model (Explanatory variables ทุกตัว ยกเว้น class ซึ่งเป็น Response variable)

คำสั่ง: Statistics > Fits models > Linear Regression

R Linear Regression	
Enter name for model: RegMo	odel.2
Response variable (pick one)	Explanatory variables (pick one or more)
CACH CHMAX CHMIN class MMAX MMIN	CHMAX CHMIN class MMAX MMIN MYCT
Subset expression <all cases="" valid="">  Help</all>	set 🧹 OK 🎇 Cancel 🥟 Apply

#### Script & Output

```
> RegModel.2 <- lm(class~CACH+CHMAX+CHMIN+MMAX+MMIN+MYCT, data=CPU)
> summary(RegModel.2)
```

```
Call:
lm(formula = class ~ CACH + CHMAX + CHMIN + MMAX + MMIN + MYCT,
   data = CPU)
Residuals:
   Min 1Q Median 3Q Max
-195.82 -25.17 5.40 26.52 385.75
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.589e+01 8.045e+00 -6.948 5.00e-11 ***
           6.414e-01 1.396e-01 4.596 7.59e-06 ***
CACH
CHMAX
           1.482e+00 2.200e-01 6.737 1.65e-10 ***
CHMIN
           -2.704e-01 8.557e-01 -0.316 0.7524
MMAX
           5.571e-03 6.418e-04 8.681 1.32e-15 ***
MMTN
           1.529e-02 1.827e-03 8.371 9.42e-15 ***
MYCT
           4.885e-02 1.752e-02 2.789 0.0058 **
____
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 59.99 on 202 degrees of freedom
Multiple R-squared: 0.8649, Adjusted R-squared: 0.8609
F-statistic: 215.5 on 6 and 202 DF, p-value: < 2.2e-16
```

คำสัง: Models > Stepwise model selection

R Stepwise Model Selection	×
Direction	Criterion
ø backward/forward	BIC
forward/backward	◎ AIC
🔘 backward	
◎ forward	
🔞 Help 🥠 Reset 🗸 O	K Cancel Apply

#### Script & Output

```
> stepwise (RegModel.2, direction='backward/forward', criterion='BIC')
Direction: backward/forward
Criterion: BIC
Start: AIC=1741.63
class ~ CACH + CHMAX + CHMIN + MMAX + MMIN + MYCT
      Df Sum of Sq RSS AIC
- CHMIN 1 359 727279 1736.4
                   726920 1741.6
<none>
            27985 754905 1744.2
- MYCT
      1
- CACH 1
             76009 802929 1757.1
- CHMAX 1 163347 890267 1778.7
- MMIN 1 252179 979099 1798.5
- MMAX 1 271177 998097 1802.5
```

```
Step: AIC=1736.39
class ~ CACH + CHMAX + MMAX + MMIN + MYCT
                      RSS AIC
       Df Sum of Sq
        727279 1736.4
<none>
             28343 755623 1739.0
- MYCT 1
+ CHMIN 1
                359 726920 1741.6
              78715 805995 1752.5
- CACH 1
- CHMAX 1 177114 904393 1776.6
- MMIN 1 258252 985531 1794.6
- MMAX 1 270856 998135 1797.2
Call:
lm(formula = class ~ CACH + CHMAX + MMAX + MMIN + MYCT, data = CPU)
Coefficients:
Coefficient
(Intercept) 0.629824
                                             MMAX MMIN MYCT
0055562 0.015180 0.049113
                  CACH CHMAX MMAX
.629824 1.459877 0.005562
```

Regression Model ที่ได้คือ

class = -56.075 + 0.630\*CASH + 1.460\*CHMAX + 0.006\*MMAX + 0.015\*MMIN + 0.049\*MYCT Model นี้มีความแม่นยำเท่าใด? Multuple R-squared = 0.8649? ลองทดสอบ Model ในส่วนอื่น

48

	สรุปคำสั่ง R (version 3.1.2)
การจัดการข้อมูล	
การอ่านข้อมูลจากไฟล์ Excel	Data > Import data > from Excel
การสร้างตัวแปรตัวใหม่	Data > Manage variables in active data set > Compute new variable
การเปลี่ยนแปลงค่าตัวแปร	Data > Manages variables in active data set > <b>Recode variables</b>
การแปลงตัวแปร numeric ให้เป็น factor	Data > Manages variables in active data set > <b>Convert</b>
และการให้ชื่อค่าของตัวแปร (value label)	numeric variables to factors
การจัดเก็บไฟล์ข้อมูล R	Data > Active data set > Save active data set
การเปิดไฟล์ข้อมูล R	Data > Load data set
สถิติสำหรับตัวแปรเดียว	
ตัวแปรเชิงคุณลักษณะ	
การแจกแจงความถื่	Statistics > Summaries > Frequency distributions
กราฟแท่งและกราฟพาย	Graphs > Bar graph / Pie Chart
ตัวแปรเชิงปริมาณ	
ค่าสถิติเบื้องตัน (N, Mean และ SD)	Statistics > Summaries > Numerical summaries
กราฟฮีสโตแกรม	Graphs > Histogram
การแจกแจงความน่าจะเป็นแบบต่าง ๆ เช่น	Distributtion > Continuous distributions > Normal distribution >
การแจกแจงปกติ (Normal distribution)	Plot normal distribution
การหาค่าวิกฤต เมื่อกำหนดระดับนัยสำคัญ	Normal quantiles
การหาค่าพี (p-value) ของค่าสถิติ	Normal probabilities
การทดสอบการแจกแจงปกติ	Statistics > Summaries > Shapio-Wilk test for normality
การทดสอบค่าเฉลี่ย	Statistics > Mean > Single-Sample t-test
สถิติสำหรับสองตัวแปร	
การเปรียบเทียบค่าความแปรปรวน 2 ค่า	Statistics > Variances > Two-variances F-test
การเปรียบเทียบค่าความแปรปรวน > 2 ค่า	Statistics > Variances > Bartlett's test / Levene's test
การเปรียบเทียบค่าเฉลี่ย 2 ค่า (Independent	Statistics > Means > Independent samples t-tes
Samples)	Statistics > Nonparametric tests > Two-sample Wilcoxon test
การเปรียบเทียบค่าเฉลี่ย 2 ค่า (Paired Samples)	Statistics > Means > Paired t-test
	Statistics > Nonparametric tests > Paired-samples Wilcoxon test
การเปรียบเทียบค่าเฉลี่ย 2 ค่าหรือมากกว่า	Statistics > Means > One-way ANOVA
(Independent Samples)	Statistics > Nonparametric tests > Kruskal-Wallis test
การแจกแจงความถี่สองทางและการ	Statistics > Contingency tables > Two-way tables
ทดสอบความเป็นอิสระของตัวแปร 2 ตัว	
การหากราฟความสัมพันธ์ของ 2 ตัวแปร	Graphs > Scratterplot
การหาความสัมพันธ์ของ 2 ตัวแปร	Statistics > Summaries > Correlation test / Correlation matrix
การวิเคราะห์การถดถอย	Statistics > Fits models > Linear Regression
	Models > Stepwise model selection

# การทำเหมืองข้อมูลด้วย Weka (Data Mining with Weka)

สุณี รักษาเกียรติศักดิ์

# การติดตั้ง Weka

- 1. Download โปรแกรมที่ <u>http://www.cs.waikato.ac.nz/ml/weka/downloading.html</u>
- ดัลเบิลคลิกเพื่อติดตั้ง

ในที่นี้จะใช้ version 3.7.11 (Book version 3.6)

ติดตั้งตาม default และ Next

讶 Weka 3.7.11 Setup	
	Welcome to the Weka 3.7.11 Setup Wizard
	This wizard will guide you through the installation of Weka 3.7.11.
	It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.
	Click Next to continue.
	Next > Cancel

จนถึง

讶 Weka 3.7.11 Setup		
Weka	Installing Please wait while Weka 3.7.11 is b	eing installed.
Extract: UnivariateEqualFr	equencyHistogramEstimator.html 1	00%
Extract: KernelEstimator. Extract: MahalanobisEstii Extract: MultivariateEstin Extract: MultivariateGaus Extract: NDConditionalEs Extract: NNConditionalEs Extract: NormalEstimator Extract: PoissonEstimato Extract: UnivariateDensit	html 100% nator.html 100% ator.html 100% sianEstimator.html 100% timator.html 100% timator.html 100% .html 100% yEstimator.html 100%	100%
Nullsoft Install System v08-Ma	r-2013.cvs	Next > Cancel

## 3. Install JRE

Java Setup - Welcome
Java" ORACLE
Welcome to Java
Java provides safe and secure access to the world of amazing Java content. From business solutions to helpful utilities and entertainment, Java makes your internet experience come to life.
Note: No personal information is gathered as part of our install process. Click here for more information on what we do collect.
Click Install to accept the license agreement and install Java now.
Change destination folder

กดปุ่ม Install

# 4. ติดตั้งเรียบร้อย

💮 Weka 3.7.11 Setup						
Weka "	stallation Complete Setup was completed successfully.					
Completed						
Output folder: C: \Program Files\Weka-3-7         Execute: RunJREInstaller.bat         Delete file: C: \Program Files\Weka-3-7\RunJREInstaller.bat         Created uninstaller: C: \Program Files\Weka-3-7\runinstall.exe         Output folder: C: \Users\Admin \AppData \Roaming\Microsoft\Windows\Start Menu\Pr         Create shortcut: C: \Users\Admin \AppData \Roaming\Microsoft\Windows\Start Menu\         Completed						
Nullsoft Install System v08-Mar-2	013,cvs					
Weka 3.7.11 Setup	Completing the Weka 3.7.11 Setup Wizard Weka 3.7.11 has been installed on your computer. Click Finish to close this wizard.					
	< <u>B</u> ack <b>Einish</b> Cancel					

## การเรียกใช้ Weka

1. All Programs > Weka 3.7



หรือ สำเนา short cut Weka 3.7 มาไว้บน desktop เพื่อสะดวกในการเรียกใช้



คลิก Do not show this message again แล้วคลิก OK

### Weka GUI Chooser



## มี Interface การทำงาน 4 แบบ

#### 1. Explorer

-		
Weka Explorer		
Preprocess Classify Cluster Associate Select attributes Visualize		
Open file Open URL Open DB Gener	Undo Edit	Save
Filter		
Choose None		Apply
Ourrent relation	Selected attribute	
Relation: None Attributes: None	Name: None	Type: None
Instances: None Sum of weights: None	Missing: None Distinct: None	Unique: None
Attributes		
		Mauritan All
Remove		
Chat and the second sec		
Status Welcome to the Weka Explorer		
recome to the recta Explorer		

## 2. Experimenter

🗿 Weka Experiment Environment				- • ×
Setup Run Analyse				
Experiment Configuration Mode:		Simple	<u>A</u> dvance	d
Open	<u><u>s</u></u>	ave	]	lew
Results Destination       ARFF file <ul> <li>Filename:</li> <li>Filename:</li> </ul>				Browse
Experiment Type		Iteration Control		
Cross-validation		Number of repetitions:		
Number of folds:		O Data sets first		
Classification     Classification		<ul> <li>Algorithms first</li> </ul>		
Datasets		Algorithms		
Add new Edit selected	Delete selected	Add new	Edit selected	Delete selected
Use relative paths	Down	Load options	Save options	Up
	No	tes		
	140			

## 3. KnowledgeFlow

🕝 Weka KnowledgeFlow Environment				
Data mining processes				
		Q Q 🖬 🖬 🖷	* • • • • >	) 🗈 🛤 🎟 🏟 😡
Design				
, onucui v				
DataSources				<u>^</u>
DataSinks				
Elusterers				E
Associations				
Tools				
Elow				
				-
	m			•
Status Log	2			
Component	Parameters	Time Status		
[Knowledgef	low]	- Welcom	e to the Weka Knowledge Flow	
				I

### 4. Simple CLI

```
SimpleCLI
                                                                    Welcome to the WEKA SimpleCLI
                                                                                ٠
Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.
> help
                                                                                Ξ
Command must be one of:
          java <classname> <args> [ > file]
          break
          kill
          capabilities <classname> <args>
          cls
          history
          exit
          help <command>
```

## Weka\_Workshops

สุณี รักษาเกียรติศักดิ์

## **WS#1: Numeric Prediction**

1 %

Data set: cpu (Weka\_data จะมีไฟล์ format เป็น .arff)

2 % As used by Kilpatrick, D. & Cameron-Jones, M. (1998). Numeric prediction 3 % using instance-based learning with encoding length selection. In Progress 4 % in <u>Connectionist</u>-Based Information Systems. Singapore: Springer-<u>Verlag</u>. 8 5 6 % Deleted "vendor" attribute to make data consistent with with what we 7 % used in the data mining book. 8 % 9 @relation 'cpu' 10 @attribute MYCT numeric 11 @attribute MMIN numeric 12 @attribute MMAX numeric 13 @attribute CACH numeric 14 @attribute CHMIN numeric 15 @attribute CHMAX numeric 16 @attribute class numeric 17 @data 18 125,256,6000,256,16,128,198 19 29,8000,32000,32,8,32,269 20 29,8000,32000,32,8,32,220 29,8000,32000,32,8,32,172 21 22 29,8000,16000,32,8,16,132 23 26 8000 32000 64 8 32 318

Look in:	🔒 Weka dat	a				) (1)			
Look jii									
(Pa)	g breast-car	ncer	segment-test		Invo Invo	ke options	dialog		
	Contact-le	enses	soybean						
Recent Items	Cou.with	vendor	unbalanced		Note:				
	Credit-q	rendor	vote		Some fil	e formats	offer addition	al	
	diabetes		🤕 weather.nominal		options when in	which can voking the	be customized options dialog	d   a.	
Desktop	🥥 glass		weather.numeric			i onang and		<u>,</u>	
	🥥 ionosphei	re							
	iris.2D								
Mu Desumente	labor								
My Documents	ReutersCo	orn-test							
	ReutersCo	orn-train							
	ReutersGr	ain-test							
Computer	ReutersGr	ain-train							
	Segment-	challenge							
	File name:	cpu.arff					Open	ן ר	
Network	Files of type:		Cl (* CD				Connel		
	riles or <u>cype</u> .	Arff data	files (*.arff)			•	Cancer		
Weka Explore Preprocess Class Open file Filter	sify Cluster /	Associate   URL	Select attributes Visualize Open DB Ge	nerate	Undo	Ed	it		Save.
Weka Explore Preprocess Class Open file Filter Choose Noo	sify Cluster / Open	Associate   URL	Select attributes Visualize	nerate	Undo	Ed	it		Save.
Weka Explore Preprocess Class Open file Filter Choose Noi Current relation	sify Cluster / Open	Associate	Select attributes Visualize	nerate	Undo	Ed	it		Save.
Weka Explore Preprocess Class Open file Filter Choose Noi Current relation Relation: cpu Instances: 209	er sify Cluster / / Open ne	Associate   URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209	Selected attribut Name: MYCT Missing: 0 (0%)	Undo	Ed	it	rpe: N jue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Non Current relation Relation: cpu Instances: 209 Attributes	er sify   Cluster   / Open ne	Associate	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 205	Selected attribut Name: MYCT Missing: 0 (0%) Statistic	Undo	Ed nct: 60 Val	it Ty Uniq ue	pe: N jue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Non Current relation Relation: cpu Instances: 209 Attributes	er sify Cluster / / Open ne	Associate   URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum	Undo	Ed nct: 60 Val	it Ty Uniq ue	rpe: N jue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Non Current relation Relation: cpu Instances: 209 Attributes All	r sify Cluster / / Open ne None	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Maximum	Undo	Ed	it Ty Uniq ue	pe: Nue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Nor Current relation Relation: cpu Instances: 209 Attributes All No. N	r sify Cluster / / Open ne None Name	Associate	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Maximum Mean StdDev	Undo	Ed	it Ty Uniq ue 3.823 263	ppe: Nuue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Nor Current relation Relation: cpu Instances: 209 Attributes All No. N 1 M	er sify Cluster / / Open ne None Vame YCT	Associate	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Maximum Mean StdDev	Undo	Ed	it Ty Uniq ue 3.823 263	rpe: N uue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Nor Current relation Relation: cpu Instances: 209 Attributes All No. N 1 M 2 MI	r sify Cluster / / Open ne None Vame YCT MIN	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Maximum Mean StdDev	Undo	Ed	it Ty Uniq ue 3.823 263	rpe: N uue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Noi Current relation Relation: cpu Instances: 209 Attributes All No. N 1 M 2 M 3 M 4 C	r sify Cluster / / Open ne None Vame YCT MIN MAX ACH	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Maximum Mean StdDev	Undo	Ed	it Ty Uniq ue 3.823 3.263	ppe: N uue: 1	Save.
Weka Explore Preprocess Class  Open file  Filter Choose Noi Relation: cpu Instances: 209 Attributes All No. N 1 M 2 M 1 M 2 M 1 C 1 M 2 M 1 C 5 C 5 C 5 C 5 C 5 C 1 5 C 1 M 1 M 1 C 1 M 1 M 1 C 1 M 1 M 1 C 1 M 1 M 1 C 1 M 1 M 1 M 1 M 1 M 1 M 1 M 1 M 1 M 1 M	r sify Cluster / / Open ne None None Vame YCT MIN MAX ACH HMIN	Associate   URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	nerate Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Maximum Mean StdDev Class: class (Num)	Undo	Ed	it Ty Uniq ue 00 3.823 3.263	ppe: N ue: 1	Save.
Weka Explore Preprocess Class  Copen file  Filter Choose Noi Relation: cpu Instances: 209 Attributes All No. N 1 M 2 M 1 M 2 M 1 C 5 C 6 C 6 C	er sify Cluster / / sify Cluster / / Open ne None None Vame YCT MIN MAX ACH HMIN HMAX	Associate   URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: class (Num)	Undo	Ed	it Ty Uniq ue 00 3.823 0.263	ppe: N ue: 1	Save.
Weka Explore Preprocess Class  Copen file  Filter Choose Noi Relation: cpu Instances: 209 Attributes All No. N 1 M 2 M 1 M 2 M 1 4 C 5 C 6 C 7 da	er sify Cluster / / Sify Cluster / / Open ne None None Vame YCT MIN MAX ACH HMIN HMAX ass	Associate   URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: dass (Num)	Undo	Ed	it Ty Uniq ue 00 3.823 0.263	ppe: N ue: 1	Save.
Weka Explore Preprocess Class  Copen file  Filter Choose Noi Current relation Relation: cpu Instances: 209 Attributes All No. N 1 M 2 MI 3 MI 4 C 5 CC 6 C 7 da	er sify Cluster / / Open ne None Vame Vame VAme VAT MIN MAX ACH HMIN HMAX ass	Associate   URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: class (Num) 137	Undo	Ed	it Ty Uniq ue 00 3.823 3.263	rpe: N ue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Non Current relation Relation: cpu Instances: 209 Attributes All No. N 1 M 2 M 3 M 4 C/ 5 C 6 C 7 Cda	er sify Cluster / / Sify Cluster / / Open ne None None None None None Name YCT MIN MIN HMIN HMIN HMIN HMAX ass	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	nerate Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: dass (Num) 137	Undo	Ed	it Ty Uniq ue 00 3.823 0.263	rpe: N uue: 1	Save.
Weka Explore Preprocess Class Open file Filter Choose Non Current relation Relation: cpu Instances: 209 Attributes All No. N 1 M 2 M 3 M 4 C/ 5 C 6 C 7 Cda	er sify Cluster / / Sify Cluster / / Open ne None None None None None None None None None None None None None None	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	nerate Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: class (Num) 137	Undo	Ed	it Ty Uniq ue 00 3.823 0.263	rpe: N uue: 1	Save.
Weka Explore Preprocess Class  Open file  Filter Choose Nor Current relation Relation: cpu Instances: 209 Attributes All No. N  1 M  2 M  3 M  4 C/ 5 Cr  6 Cr  7 Cda	er sify Cluster / / Sify Cluster / / Open ne None	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	nerate Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: dass (Num) 137 24	Undo	Ed	it Ty Uniq ue 00 3.823 0.263	rpe: Nue: 1	Save.
Weka Explore Preprocess Class  Open file  Filter Choose Nor Current relation Relation: cpu Instances: 209 Attributes All No. N  Current relation No. N  Current relation Relation: cpu Instances: 209 Relat	er sify Cluster / / Sify Cluster / / Open ne None	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 205 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: class (Num) 137	Undo	Ed	it Ty Uniq ue 3.823 0.263	rpe: N ue: 1	Save.
Weka Explore Preprocess Class  Open file  Filter Choose Nor Current relation Relation: cpu Instances: 209 Attributes All No. N     No. N	er sify Cluster / / Open ne None None Vame YCT MIN MAX ACH HMIN HMAX ass	Associate URL	Select attributes Visualize Open DB Ge Attributes: 7 Sum of weights: 209 Invert Pattern	Selected attribut Name: MYCT Missing: 0 (0%) Statistic Minimum Mean StdDev Class: class (Num) 137	Undo	Ed	it Ty Uniq ue 00 3.823 0.263	pe: N ue: 1	Save.

 แอททริบิวต์ตัวสุดท้ายจะเป็นตัวแปร class หากต้องการให้ตัวแปรอื่นเป็นตัวแปร class ให้ Edit แล้ว คลิกขวาตัวแปรที่ต้องการ เลือก Attribute as class ตัวแปรนั้นก็จะมาอยู่เป็นตัวสุดท้าย และแสดงเป็น

## ตัวหนา

v v	iewer		-	-		-		x	
Pelation: cou									
Relation	on: cpu	D. 141471	o	4.0401	E. CLIMITAL		-		
INO.	1: MYCI Numeric	2: MMIIN	3: MIMAX	4: CACH	5: CHMIN	6: CHMAX /	class		
122	1500.0	769.0	1000.0	Numeric	Nu.	Get mean			
123	1500.0	760.0	2000.0	0.0					
124	1300.0	760.0	2000.0	0.0		Set all values t	to		
125	490.0	/68.0	2000.0	22.0		C			
208	480.0	512.0	8000.0	32.0		Set missing va	alues to.		
209	480.0	1000.0	4000.0	0.0		Replace value	s with		
11	400.0	1000.0	3000.0	0.0					
12	400.0	512.0	3500.0	4.0		Rename attrib	ute		
13	60.0	2000.0	8000.0	65.0		Association and all			
14	50.0	4000.0	16000.0	65.0		Attribute as c	ass		
15	350.0	64.0	64.0	0.0		Delete attribut	te		
25	320.0	128.0	6000.0	0.0		Delete attribut	tor		
26	320.0	512.0	2000.0	4.0		Delete attribu			
27	320.0	256.0	6000.0	0.0		Sort data (asc	ending)		
28	320.0	256.0	3000.0	4.0					
29	320.0	512.0	5000.0	4.0		Optimal colur	mn widt	h (currei	nt)
30	320.0	256.0	5000.0	4.0		Ontimal colu	nn widt	h (all)	
37	50.0	500.0	2000.0	8.0			Loro		
38	50.0	1000.0	4000.0	8.0	1.0	5.0	29.0		
39	50.0	2000.0	8000.0	8.0	1.0	5.0	71.0		
47	810.0	512.0	512.0	8.0	1.0	) 1.0	18.0		
48	810.0	1000.0	5000.0	0.0	1.0	0 1.0	20.0		
49	320.0	512.0	8000.0	4.0	1.0	5.0	40.0		
50	200.0	512.0	8000.0	8.0	1.0	8.0	62.0		
51	700.0	384.0	8000.0	0.0	1.0	) 1.0	24.0	-	
	Undo OK Cancel								

## <u>หมายเหต</u>ุ จะใช้ว่า แอททริบิวต์ หรือ ตัวแปร ก็ได้

้ถ้าต้องการได้ไฟล์ลง Excel ก็ ctrl+A (Select All) แล้วไป paste ลงใน Excel แล้วเพิ่มบรรทัดแรกเป็น ชื่อตัวแปร

Weka จะสรุปค่า descriptive statistics ของแต่ละตัวแปรให้ พร้อมแสดงกราฟ (histogram) หรือคลิก
 Visualize All เพื่อด histogram ของตัวแปรทกตัวพร้อมกัน



4. Visualize scatter plot (ตัวแปร 2 ตัว) โดยคลิกที่แท็บ Visualize



5. Fit linear regressiom model

Classify > Choose Classifier = functions > Linear Regression

Test options = Use training ser (default)

Weka Explorer	Weka Explorer
Preprocess Classify Cluster Associate	Preprocess Classify Cluster Associate Select attri
Classifier	Classifier
weka	Choose LinearRegression -5 0 -R 1.0E-8
a bayes	Test options
GaussianProcesses	Use training set
- LinearRegression	Supplied test set Set
Logistic     MultilaverPerceptron	○ Cross-validation Folds 10
♦ SGD	Percentage split % 66
• SGDText	More options
SimpleLinearRegression	
_( ♦ SMO	(Num) class
SMOreg	
VotedPerceptron	Start Stop
F meta	Result list (right-dick for options)

## กด Start จะได้ผลลัพธ์

🔮 Weka Explorer			
Preprocess Classify Cluster Associate S	elect attributes Visualize		
Classifier			
Choose LinearRegression -5 0 -R 1	.0E-8		
Test options	Classifier output		
<ul> <li>Use training set</li> </ul>	Linear Regression Model		A
Supplied test set Set Cross-validation Eolds 10	class =		
	0.0491 * MYCT +		
Percentage spire % 00	0.0152 * MMIN +		
More options	0.0056 * MMAX +		
	0.6298 * CACH +		
(Num) class 🔹 👻	1.4599 * CHMAX +		
	-56.075		
Start Stop	Time taken to build model: 0.03 seconds		
Result list (right-click for options)			
18:15:55 - functions.LinearRegression	=== Evaluation on training set ===		
	Time taken to test model on training da	ta: 0.02 seconds	E
	=== Summary ===		
	Correlation coefficient	0.93	
	Mean absolute error	37.9748	
	Root mean squared error	58.9899	
	Relative absolute error	39.592 %	
	Root relative squared error	36.7663 %	<b>T</b>
Status			
ок			Log 💉 x 0

- 6. การแปลผล
  - 6.1 Model: Class/PRP = -56.075 + 0.0491\*MYCT + 0.0152\*MMIN + 0.0056 \*MMAX
    - + 0.6298\*CACH + 1.4599\*CHMAX

ตัวแปรใดที่ไม่เข้าในโมเดล?

```
Linear Regression Model

class =

0.0491 * MYCT +

0.0152 * MMIN +

0.0056 * MMAX +

0.6298 * CACH +

1.4599 * CHMAX +

-56.075
```

6.2 Evaluation: Correlation coefficient R=0.93 or R<sup>2</sup>=0.8649

ตัวแปร predictor 5 ตัว คือ MYCT, MMIN, MMAX, CACH, CHMAX สามารถอธิบายความ แปรปรวนของตัวแปร Class/PRP ได้ประมาณ 81%

Correlation coefficient 0.93

6.3 Deployment: เครื่องคอมพิวเตอร์เครื่องใหม่ 1 เครื่อง มีคุณสมบัติดังนี้ MYCT=200, MMIN=1000, MMAX=2000, CASH=0, CHMAX=64 จะมีประสิทธิภาพ (PRP) = 73.58

คลิกขวาที่ Result แล้ว Save model

 ให้ทดสอบทั้ง 3 test options จะได้ว่า Weka ใช้ full training set ในการสร้างโมเดล ซึ่งผลลัพธ์ของ โมเดลจะเหมือนกัน ซึ่งสอดคล้องกับ Stepwise Regression ใน SPSS และใน R แต่ผลลัพธ์ของ Evaluation จาก Correlation coefficient ของทั้ง 3 test mode คือ 1) Use training set, 2) Crossvalidation (10 folds), 3) Percentage split (66%) จะไม่เท่ากัน

### WS#2: Classification by Logistic Regression

#### Data set: diabetes

1. Preprocess > Open file เลือก diabetes

eprocess Classify Cluster Associate Select attributes Visualize		
Open file Open URL Open DB Gene	rate Undo	Edit Save
Choose None		Apply
urrent relation Relation: pima_diabetes Attributes: 9 Instances: 768 Sum of weights: 768	Selected attribute Name: preg Missing: 0 (0%) Distinct:	Type: Numeric 17 Unique: 2 (0%)
ttributes	Statistic	Value
All None Invert Pattern	Minimum	0
All None Divert Pattern	Maximum	17
la Neura	Mean	3.845
3 pres 4 skin 5 nsu	(lace: dace (Nom)	▼ Vicualize ∆
6 mass	Class, class (Non)	Visualize A
/	246	
9 das		
Remove	103 75 50 45	<sup>24</sup> 11 19 2 1 1
	0	8.5
tatus		

2. Edit เพื่อดู type และ ค่าของแต่ละตัวแปร

-										
🖆 V	iewer									×
Relati	on: pima_d	liabetes								
No.	1: preg	2: plas	3: pres	4: skin	5: insu	6: mass	7: pedi	8: age	9: class	1
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive	
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative	
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_positive	
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_negative	
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_positive	
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_negative	
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive	
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative	
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive	
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive	
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative	
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive	
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative	
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive	
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive	
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive	
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive	
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive	
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_negative	
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_positive	
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested_negative	
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested_negative	
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested_positive	
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested_positive	-
										Connel
								Und	OK	Cancel

ตัวแปร class (dependent/target) มี type เป็น nominal มี 2 ค่า คือ tested negative/tested positive ตัวแปร classifier (independent/predictor) มี 8 ตัว มี type เป็น numeric ทั้งหมด  ดู descriptive statistics ของแต่ละตัวแปรที่ Weka สรุปให้ พร้อมแสดงกราฟ (histogram) หรือคลิก Visualize All เพื่อดู histogram ของตัวแปรทุกตัวพร้อมกัน

😡 Weka Explorer	
Preprocess Classify Cluster Associate Select attributes Visualize	
Open file Open URL Open DB Gene	rate Undo Edit Save
Filter	
Choose None	Apply
Current relation	Selected attribute
Relation: pima_diabetes Attributes: 9	Name: preg Type: Numeric
Instances: 768 Sum of weights: 768	Missing: 0 (0%) Distinct: 17 Unique: 2 (0%)
Attributes	Statistic Value
All None Invert Pattern	Minimum 0
	Maximum 17
No. Name	StdDev 3.37
1 🔲 preg	
2 plas	
4 skin	
5 insu	Class: class (Nom)
6 mass	
/ pedi	246
9 dass	
	125
	103
	50 45 66
Remove	24 11 19 2 1 1
	0 8.5 17
Status	
OK	Log 💞 x 0

ข้อมูลนี้มีทั้งหมด 768 ตัว (Instances/Cases/N)

ตัวแปร 1 ถึง 8 เป็นตัวแปร classifier มี type เป็น numeric ดังนั้น descriptive statistics จึงเป็น Minimum, Maximum, Mean, Standard Deviation (StdDev)

ตัวแปร class จะเป็นตัวแปรสุดท้ายเสมอ มี type เป็น nominal

ลองคลิก Visualize All จะเห็นกราฟความสัมพันธ์ของตัวแปร class กับตัวแปร classifier แต่ละตัว

- 4. Visualize scatter plot (ตัวแปร 2 ตัว) โดยคลิกที่แท็บ Visualize
- 5. Fit logistic regressiom model

เลือกแท็บ Classify > Choose Classifier = functions >Logistic (Regression) Default Test option = Cross validation Fold 10

Preprocess       Classify       Cluster       Associate       Select         Classifier       Classifiers       Classifier       Classifier         Image: Select       Image: Select       Supplement of the select       Select         Image: Select       Image: Select       Supplement of the select       Select         Image: Select       Supplement of the select       Select       Select         Image: Select       Select       Select       Select         Image: Select	🜍 Weka Explorer	🜍 Weka Explorer
	Preprocess Classify Cluster Associate Select Classifier weka classifiers weka classifiers bayes functions GaussianProcesses LinearRegression Logistic MultilayerPerceptron SGD SGDText SimpleLogistic SimpleLogistic SMO SMOreg VotedPerceptron lazy meta misc rules	Preprocess       Classify       Cluster       Associate         Classifier       Choose       Logistic -R 1.0E-8 -M -1         Test options       Use training set       Supplied test set       Set         O Supplied test set       Set       Set         O Cross-validation       Folds       10         Percentage split       %       66         More options       More options         (Nom) class       Start       Stop         Result list (right-click for options)       Stop

## กด Start จะได้ผลล*ั*พธ์

## 6. การแปลผล

Classifier output Logistic Regression with ridge		Odds Ratios Class Variable tested_negative	
Variable	Class tested_negative	preg plas	0.8841 0.9654
preg plas	-0.1232 -0.0352	pres skin insu	1.0134 0.9994 1.0012
pres skin	0.0133	mass pedi	0.9142
insu mass	0.0012	age	0.9852
pedi age Intercept	-0.9452 -0.0149 8.4047		

### 6.2 Evaluation:

Correctly Classified Instances	593	77.2135 %	
Incorrectly Classified Instances	175	22.7865 %	

## 6.3 **Deployment:** แทนค่าลงในสมการข้อ 6.1

คลิกขวาที่ Result แล้ว Save model

7. ให้ทดสอบทั้ง 2 test options (Use training set, Cross-validation Folds 10 จะได้ว่า Weka ใช้ full training set ในการสร้างโมเดล ซึ่งผลลัพธ์ของโมเดลจะเหมือนกัน แต่ accuracy จะไม่เหมือนกัน ในวิธีการ Cross-validation Folds 10 Weka จะแบ่งข้อมูลทั้งหมดออกเป็น 10 ส่วน และรัน 10 ครั้ง โดยครั้งที่ 1 ใช้ ส่วนที่ 1 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set ครั้งที่ 2 ใช้ ส่วนที่ 2 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set จนถึงครั้งที่ 10 ใช้ ส่วนที่ 10 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set จนถึงครั้งที่ 10 ใช้ ส่วนที่ 10 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set จนถึงครั้งที่ 10 ใช้ ส่วนที่ 10 เป็น test set ส่วนที่เหลือทั้งหมดอีก 9 ส่วนเป็น training set โดย evaluation accuracy จะเป็นค่าเฉลี่ยของทั้ง 10 ครั้ง จากนั้นจะรันโมเดลอีกครั้งโดยใช้ full training set ทั้งหมดในการรันโมเดล

### WS#3: Classification by ML algorithm

#### Data set: diabetes

- 1. Preprocess > Open file เลือก diabetes
- 2. Fit classification model Classifier = rules > ZeroR (baseline accuracy)

#### การแปลผล

#### **Model:** Class = tested\_negative

```
=== Classifier model (full training set) ===
ZeroR predicts class value: tested negative
```

#### Evaluation:

```
Correctly Classified Instances 500 65.1042 %
Incorrectly Classified Instances 268 34.8958 %
=== Confusion Matrix ===
a b <-- classified as
500 0 | a = tested_negative
268 0 | b = tested_positive
```

Deployment: tested negative

3. Fit classification model Classifier = rules > OneR

#### การแปลผล

#### Model: ไม่ make sence

(587/768 instances correct)

#### Evaluation: (cross-validation)

Correctly Classified Instances	549	71.4844 %
Incorrectly Classified Instances	219	28.5156 %

```
=== Confusion Matrix ===
a b <-- classified as
433 67 | a = tested_negative
152 116 | b = tested_positive</pre>
```

**Deployment:** ใช้กฎ (apply rule)

4. Fit classification model Classifier = trees > J48

<u>หมายเหตุ</u>ตาม default parameter จะได้จะได้ 20 กฏ ซึ่งบางกฏ ก็เป็นจริงสำหรับข้อมูลจำนวนน้อย ดังนั้นควรจะปรับ parameter "minNumObj" เช่นถ้าต้องการให้กฏเป็นจริงสำหรับข้อมูล 10% ก็ปรับ minNumObj = 77 โดยคลิกที่ **J48** แก้ minNumObj

68

	Choose J48 -C 0.25 -M 77				
1	🔮 weka.gui.GenericObjectEditor				
l	weka.classifiers.trees.J48 About				
I	Class for generating a prun	ed or unpruned C4.			
	binarySplits	False E			
l	collapseTree	True			
	confidenceFactor	0.25			
	debug	False			
l	doNotCheckCapabilities	False			
l	doNotMakeSplitPointActualValue	False			
l	minNumObj	77			
l	numFolds	3			
	•				
	Open Save	OK Cancel			

## การแปลผล

Model: คลิกขวาที่ Result เลือก Visualize tree



Evaluation: (cross-validation)

=== Stratified cross-validation === === Summary ===		
Correctly Classified Instances	570	74.2188 %
Incorrectly Classified Instances	198	25.7813 %
=== Confusion Matrix ===		
a b < classified as		
445 55   a = tested_negative		
143 125   b = tested_positive		

**Deployment:** คลิกขวาที่ Result แล้ว Save model

 ให้เปรียบเทียบผลการทดลองด้วย Classifier ตัวอื่น ๆ เช่น rules > PART, rules > JRIP (จะดำเนินการใน Experimenter Interface)

## WS#4: Association Problem Type

## Data set: supermarket

1. Preprocess > Open file เลือก supermarket

## Edit เพื่อดูการจัดเก็บข้อมูล

4	🖄 Viewer						
Re	Relation: supermarket						
at	20: canned fruit	21: canned vegetables	22: breakfast food	23: cigs-tobacco pkts			
	Nominal	Nominal	Nominal	Nominal			
		t			^		
	t	t		C	_		
	t						
		t	t				
		t	t				
	t	t	t				
		t					
	-	-					
	t	t	-				
	t	t	t				
	t	t	-	-			
			t	t			
			-				
			t				
	t	t	t				
	t						
				-			
			t	t			
				t			
		t	t				
					-		
1	× •						
	Undo OK Cancel						

แถวหนึ่ง ๆ คือตะกร้าสินค้าหนึ่ง ๆ t (=true) มีสินค้าชนิดนั้น

2. Fit Association model โดยเลือกแท็บ Associate

Preprocess Classify Cluster Associate	Preprocess Classify Cluster Associate Select attributes Visuali:
weka	Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.2 -S -1.
Apriori	veka.gui.GenericObjectEditor
FilteredAssociator	weka.associations.Apriori
เลือก Associator algorithm เป็น	Class implementing an Apriori-type algorithm.
Apriori	
คลิกที่ Apriori (รูปทางขวา)	car False E
ปรับ parameter ที่สำคัญ 3 ตัวคือ	classIndex -1
-lowerBoundMinSupport = 0.2	delta 0.05
-minMetric (min Confidence) = 0.9	lowerBoundMinSupport 0.2
-numRules = 10	metricType Confidence
ผลลัพธ์: ไม่พบกฏใด ๆ เลย	minMetric 0.9
ลองปรับ	numPulan 10
-minMetric (min Confidence) = 0.85	
	Open Save OK Cancel

## การแปลผล

### Model:

Best rules found:	
1. biscuits=t frozen foods=t fruit=t vegetables=t 1039 ==> bread and cake=t 929 <conf:(0.89)> lift:(1.1)</conf:(0.89)>	24) lev:(0.
2. fruit=t vegetables=t total=high 1050 ==> bread and cake=t 938 <conf:(0.89)> lift:(1.24) lev:(0.04)</conf:(0.89)>	[182] conv:
3. fruit=t total=high 1243 ==> bread and cake=t 1104 <conf:(0.89)> lift:(1.23) lev:(0.05) [209] conv:(</conf:(0.89)>	2.49)
4. biscuits=t total=high 1228 ==> bread and cake=t 1082 <conf:(0.88)> lift:(1.22) lev:(0.04) [198] con</conf:(0.88)>	7:(2.34)
5. milk-cream=t total=high 1217 ==> bread and cake=t 1071 <conf:(0.88)> lift:(1.22) lev:(0.04) [195] c</conf:(0.88)>	onv:(2.32)
6. biscuits=t margarine=t vegetables=t 1054 ==> bread and cake=t 925 <conf:(0.88)> lift:(1.22) lev:(0.</conf:(0.88)>	)4) [166] c
7. frozen foods=t total=high 1273 ==> bread and cake=t 1117 <conf:(0.88)> lift:(1.22) lev:(0.04) [200]</conf:(0.88)>	conv: (2.25
8. biscuits=t margarine=t fruit=t 1073 ==> bread and cake=t 938 <conf:(0.87)> lift:(1.21) lev:(0.04) [</conf:(0.87)>	65] conv:
9. party snack foods=t total=high 1120 ==> bread and cake=t 979 <conf:(0.87)> lift:(1.21) lev:(0.04) [</conf:(0.87)>	72] conv:
10. vegetables=t total=high 1270 ==> bread and cake=t 1110 <conf:(0.87)> lift:(1.21) lev:(0.04) [195] c</conf:(0.87)>	onv:(2.21)

Evaluation: Min Support และ Confidence

Deployment: เอา กฏ ไปใช้

## WS#5: Clustering

#### Data set: iris

- Preprocess > Open file เลือก iris
   ในที่นี้ iris มีตัวแปร class แต่เราจะไม่ใช้ในการ clustering แต่จะใช้ในการเปรียบเทียบกับค่าจากโมเดล
- 2. Fit Clustering model โดยเลือกแท็บ Cluster


process Classify Cluster Associate Select	attributes Visualize	คลก SimpleKMeans
Choose SimpleKMeans -init 0 -max-candid	dates 100 -periodic-pruning 10000 -min-density	
weka.gui.GenericObjectEditor	×	חבת numCluster = 3
weka.clusterers.SimpleKMeans		
About	<b>^</b>	
Cluster data using the k means algorith	Im. Mor Capabi	
canopyMaxNumCanopiesToHoldInMemory 1	100	
canopyMinimumCanopyDensity 2	2.0	
canopyPeriodicPruningRate 1	10000	
canopyT1 -	1.25	
canopyT2 -	10 E	
debug [	Elles	
debug [		
displaystopevs [	aise	
distanceFunction	Choose EuclideanDistance -R f	
doNotCheckCapabilities F	alse	
dontReplaceMissingValues F	alse	
fastDistanceCalc F	alse	
initializationMethod R	Random	
maxIterations 5	500	
numClusters 3	8	
<		

## การแปลผล

#### Model:

Final cluster	centroids:			
		Cluster#		
Attribute	Full Data	0	1	2
	(150)	(61)	(50)	(39)
sepallength	5.8433	5.8885	5.006	6.8462
sepalwidth	3.054	2.7377	3.418	3.0821
petallength	3.7587	4.3967	1.464	5.7026
petalwidth	1.1987	1.418	0.244	2.0795

#### **Evaluation:**

Within cluster sum of squared errors: 6.998114004826762

Deployment: คลิกขวาที่ Result แล้ว Save model

# 74

#### Witten: "Data mining is an experimental science"

## WS#6 จงเปรียบเทียบการวิเคราะห์โจทย์ Classification โดยใช้หลาย algoithms

#### **Experiment 1**

#### Data set: iris

Algoritms: trees > J48, rules > OneR, rules > ZeroR

1. Weka GUI Chooser เลือก Experimenter

Weka Experiment Environment				- <b>•</b> ×
Setup Run Analyse				
Experiment Configuration Mode:		<u>Simple</u>	<u>A</u> dvanc	ed
Open	<u></u>	ave		New
Results Destination				
ARFF file v Filename:				Browse
Experiment Type		Iteration Control		
Cross-validation	Ŧ	Number of repetitions:		
Number of folds:		Oata sets first		
Classification     Classification	ı	<ul> <li>Algorithms first</li> </ul>		
Datasets		Algorithms		
Add new Edit selected	Delete selected	Add new	Edit selected	Delete selected
Use relative paths				
Up	Down	Load options	Save options	Up Down
	No	tes		

### 2. คลิก **New**

Weka Experiment Environment	Part of the local division of the local divi		
Setup Run Analyse			
Experiment Configuration Mode:	Simple	<u>A</u> dvanced	
	Save	New	
Results Destination           ARFF file <ul> <li>Filename:</li> </ul>			Browse
Experiment Type	Iteration Control		
Cross-validation	<ul> <li>Number of repetition</li> </ul>	ns: 10	
Number of folds: 10	<ul> <li>Data sets first</li> </ul>		
Classification     Classification	Algorithms first	:	
Datasets  Add new Edit selected Delete selecte Use relative paths	Algorithms d Add new	Edit selected Delete se	lected
Up Down	Load options	. Save options Up	Down
	Notes		
			-

3. ที่ Datasets คลิก Add new แล้วเปิดเลือก iris

ที่ Algorithms คลิก Add new แล้วเลือก Choose: rules > ZeroR, rules > OneR, trees > J48 (Add new 3 ครั้ง ตาม default)

G Weka Experiment Environment	
Setup Run Analyse	
Experiment Configuration Mode:	
Open	<u>a</u> ve <u>N</u> ew
Results Destination	
ARFF file - Filename:	Browse
Experiment Type	Iteration Control
Cross-validation -	Number of repetitions: 10
Number of folds: 10	Oata sets first
Classification     Classification	Algorithms first
Datasets	Algorithms
Add ne Edit s Delete	Add Edit sel Delete s
Use relati	149 CO 35 M 3
ColDrooram EilopiWoka 2 71dataliria arff	OneR -B 6
C. Program nes (weka-5-7 juata justam	ZeroR
Up Down	Load op Save op
N	otes
	,

4. ไปที่แท็บ Run คลิก Start

🕝 Weka Experiment Environment		🔮 Weka Experiment Environment
Setup Run Analyse		Setup Run Analyse
<u>S</u> tart S	Stop	Start Stop
Log		Log
		18:48:02: Started
		18:48:02: Finished
		18:48:02: There were 0 errors
Status		Status
Not running		Not running

5. ไปที่แท็บ Analyze คลิก Experiment

🕝 Weka Experiment E	nvironment	
Setup Run Analyse		
Source		
Got 300 results		<u>File</u> <u>D</u> atabase <u>Experiment</u>
Configure test		Test output
Testing <u>w</u> ith	Paired T-Tester (corrected)	Available resultsets
Select <u>r</u> ows and cols	Rows Cols Swap	<pre>(1) trees.J48 '-C 0.25 -M 2' -217733168393644444 (2) rules.OneR '-B 6' -3459427003147861443 (3) rules.ZeroR '' 48055541465867954</pre>
Comparison field	Percent_correct	
Significance	0.05	
Sorting (asc.) by	<default></default>	
Test <u>b</u> ase	Select	
Displayed Columns	Select	
Show std. devi <u>a</u> tions		
Output Format	Select	
Perform <u>t</u> est Result list	Save output	

#### 6. คลิก Perform test

🕝 Weka Experiment E	nvironment		x
Setup Run Analyse			
Source			
Got 300 results		<u>Fi</u> le <u>D</u> atabase <u>E</u> xperimen	nt
Configure test		Test output	
Testing <u>w</u> ith	Paired T-Tester (corrected)	Tester: weka.experiment.PairedCorrectedTTester Analysing: Percent correct	Â
Select rows and cols	Rows Cols Swap	Datasets: 1 Resultsets: 3	
Comparison field	Percent_correct 🔹	Confidence: 0.05 (two tailed) Sorted by: -	
Significance	0.05	Date: 26/12/2557, 19:01 u.	
Sorting (asc.) by	<default></default>	Dataset (1) trees.J4   (2) rules (3) rules	
Test <u>b</u> ase	Select	iris (100) 94.73   92.53 33.33 *	•
Displayed Columns	Select	(v/ /*)   (0/1/0) (0/0/1)	
Show std. deviations		Trans.	
<u>O</u> utput Format	Select	<pre>key: (1) trees.J48 '-C 0.25 -M 2' -217733168393644444 (2) rules.OneR '-B 6' -3459427003147861443</pre>	
Perform <u>t</u> est	Save output	(3) rules.ZeroR '' 48055541465867954	•

7. การแปลผลจาก Witten

Dataset	<li>(1) trees.J4</li>	(2) rules (3) rules
iris	(100) 94.73	92.53 33.33 *
	(100) 0000	
	(V/ /*)	(0/1/0) (0/0/1)
Key:		
(2) rules.OneR		
(3) rules.ZeroR		

- v significantly better
- \* significantly worse

- ZeroR (33.3%) is significantly worse than J48 (94.7%)
- Cannot be sure that OneR (92.5%) is significantly worse than J48
- ✤ ... at the 5% level of statistical significance
- ↔ J48 seems better than ZeroR: pretty sure (5% level) that this is not due to chance

# Experiment 2:

78

#### Data set: iris, breast-cancer, glass, ionosphere, segment-challenge

Algoritms: trees > J48, rules > OneR, rules > ZeroR

- 1. ไปที่แท็บ Set up คลิก New
- ที่ Datasets คลิก Add new แล้วเปิดเลือก iris, breast-cancer, glass, ionosphere, segmentchallenge (Add new 5 ครั้ง)

ที่ Algorithms คลิก Add new แล้วเลือก Choose: rules > ZeroR, rules > OneR, trees > J48 (Add new 3 ครั้ง ตาม default)

Weka Experiment Environment		
Setup Run Analyse		
Experiment Configuration Mode:	Simple	Advanced
Open	Save	New
Results Destination		
ARFF file   Filename:		Browse
Experiment Type	Iteration Control	
Cross-validation	<ul> <li>Number of repetition</li> </ul>	ns: 10
Number of folds: 10	O Data sets first	·,
Classification	Algorithms first	
Datasets	Algorithms	
Add new Edit sele Del	Add n	Edit select Delete sele
Use relative	148 -C 0.25 -M 2	
C:\Program Files\Weka-3-7\data\iris.arff	OneR -B 6	
C:\Program Files\Weka-3-7\data\breast-cancer	arff	
C: \Program Files \Weka-3-7\data \glass.arff C: \Program Files \Weka-3-7\data \joposphere ar	ff I	
C:\Program Files\Weka-3-7\data\segment-chall	enge.arff	
Up Down	Load opti	Save opti
	Notes	

- 3. ไปที่แท็บ Run คลิก Start
- 4. ไปที่แท็บ Analyze คลิก Experiment

### 5. คลิก Perform test

🕝 Weka Experiment Er	nvironment		a, Million	-		- 🗆 🗙
Setup Run Analyse						
Source						
Got 1500 results			File.	Datab	base	Experiment
Configure test		Test output				
Testing with	Paired T-Tester (corrected)	Tester: weka.experi Analysing: Percent_cor	iment.Pai: rrect	redCorrecte	edTTester	
Select rows and cols	Rows Cols Swap	Datasets: 5 Resultsets: 3				
Comparison field	Percent_correct	Confidence: 0.05 (two t Sorted by: -	ailed)			
Significance	0.05	Date: 26/12/2557,	19:26 u	•		
Sorting (asc.) by	<default></default>	Dataset	(1) t:	rees.J4	(2) rules	(3) rules
Test base	Select	iris	(100)	94.73	92.53	33.33 *
Displayed Columns	Select	breast-cancer Glass	(100)	74.28   67.63	66.91 * 57.40 *	70.30 35.51 *
Show std. deviations		ionosphere segment	(100) (100)	89.74   95.71	82.28 * 64.35 *	64.10 * 15.73 *
Output Format	Select			(v/ /*)	(0/1/4)	(0/1/4)
Perform test         Save output           Result list         (1) trees.J48 '-C 0.25 -M 2' -217733168393644444           19:00:39 - Available resultsets         (2) rules.OneR '-B 6' -3459427003147861443           19:01:51 - Percent_correct - trees.J48 'C 0.25 -M 2' -21773316839           19:26:05 - Available resultsets           19:26:05 - Percent_correct - trees.J48 'C 0.25 -M 2' -21773316839           (2) rules.ZeroR '' 48055541465867954						

6. เปลี่ยน **Test base** Select เป็น rules.OneR

🕝 Weka Experiment E	nvironment			
Setup Run Analyse				
Source				
Got 1500 results			Eile	Experiment
Configure test		Test output		
Testing <u>w</u> ith	Paired T-Tester (corrected)	Tester: we Analysing: Pe	ka.experiment.PairedCorrectedTTester rcent_correct	<u>^</u>
Select <u>r</u> ows and cols	Rows Cols Swap	Datasets: 5 Resultsets: 3	-	
Comparison field	Percent_correct	Sorted by: -	05 (two tailed)	
Significance	0.05	Date: 26	/12/2557, 19:50 u.	
Sorting (asc.) by	<default></default>	Dataset	(2) rules.On   (1) trees (3	3) rules =
Test <u>b</u> ase	Select	iris	(100) 92.53   94.73	33.33 *
Displayed Columns	Select	breast-cancer Glass	(100) 66.91   74.28 v (100) 57.40   67.63 v	70.30 35.51 *
Show std. deviations		ionosphere segment	(100) 82.28   89.74 v (100) 64.35   95.71 v	64.10 * 15.73 *
<u>O</u> utput Format	Select		(v/ /*)   (4/1/0)	(0/1/4)
Perform <u>t</u> est	Save output	Key:		
Result list		<pre>(1) trees.J48 (2) rules.OneF</pre>	'-C 0.25 -M 2' -217733168393644444 '-B 6' -3459427003147861443	_
19:30:44 - Percent_con	III FUIL - ULES - ULES - MI Z - 21/73316	<		•
		P		