การใช้โปรแกรม **R** เพื่อการวิจัย

ดร.วนิดา พงษ์ศักดิ์ชาติ

บทที่ 1

โปรแกรม 

R เป็นโปรแกรมคอมพิวเตอร์โปรแกรมหนึ่งที่มีความสามารถสูงในการวิเคราะห์ข้อมูลเชิงสถิติ ซึ่งในปัจจุบัน **R** เป็นที่รู้จักของนักวิจัยในสาขาต่าง ๆ และถูกนำมาใช้อย่างแพร่หลาย สาเหตุที่ **R** ได้รับความนิยมมากในปัจจุบันก็เนื่องจาก **R** เป็นซอฟแวร์ประเภท Open source ที่ทุกคนสามารถนำมาใช้ได้โดยไม่มีค่าใช้จ่ายใด ๆ และผู้ใช้ไม่ต้องกังวลกับเรื่องของการละเมิดลิขสิทธิ์เหมือนกับโปรแกรมสำเร็จรูปทางสถิติอื่น ๆ

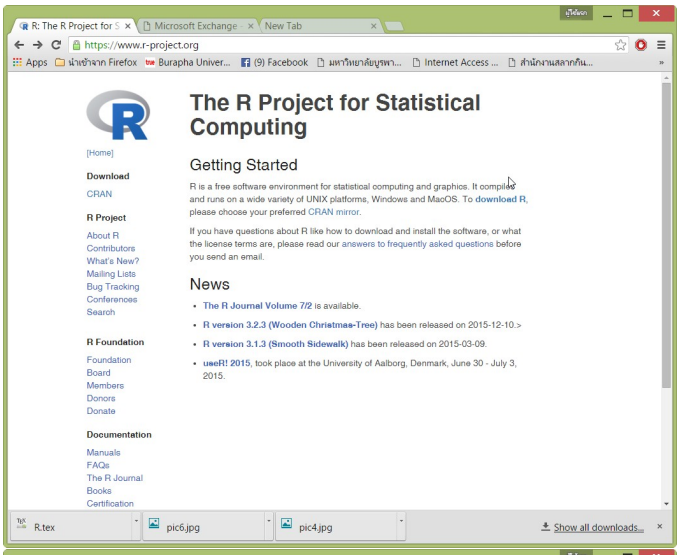
R เป็นโปรแกรมที่อยู่ภายใต้การดูแลของมูลนิธิที่ไม่แสวงหากำไรชื่อ **R** Foundation โดยมี Robert Gentleman และ Ross Ihaka จากภาควิชาสถิติ มหาวิทยาลัย Auckland เป็นผู้ที่เริ่มพัฒนา **R** ขึ้น และมีสมาชิกหลักจำนวนหนึ่งซึ่งดูแลและจัดการเกี่ยวกับ **R** ให้กับผู้ใช้ตั้งแต่ปี 1997 จนถึงปัจจุบัน (รายละเอียดสามารถดูได้จาก http://www.r-project.org) **R** เป็นส่วนหนึ่งในโครงการของ GNU การใช้ **R** สามารถใช้ได้ทั้งบนระบบปฏิบัติการ Unix Macintosh และ Windows ข้อมูลต่าง ๆ เกี่ยวกับ **R** และตัวโปรแกรมสามารถหาได้จาก http://www.r-project.org

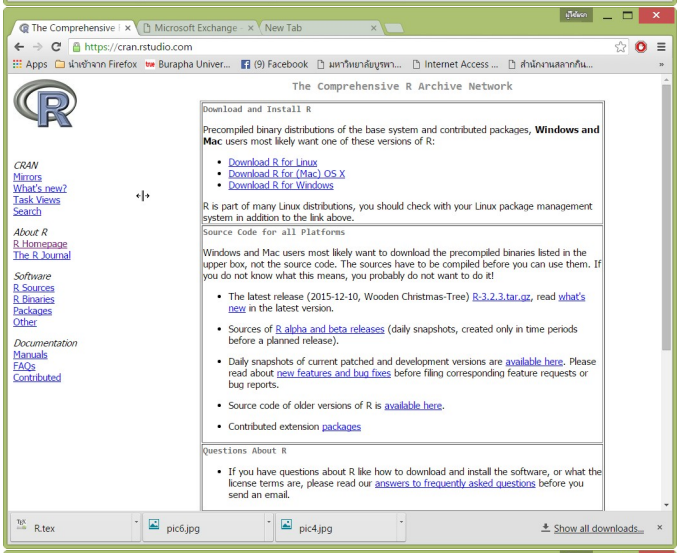
R เป็นโปรแกรมที่ดีมากสำหรับใช้ในการเรียนรู้ทางสถิติ เนื่องจากสามารถทำให้ผู้เรียนเกิดความเข้าใจในกระบวนการทางสถิติได้ดียิ่งขึ้น อีกทั้งยังเป็นโปรแกรมที่มีความยืดหยุ่นในการวิเคราะห์ทางสถิติ จึงทำให้ผู้ใช้สามารถขยายกระบวนการวิเคราะห์ออกไปได้ตามความต้องการ

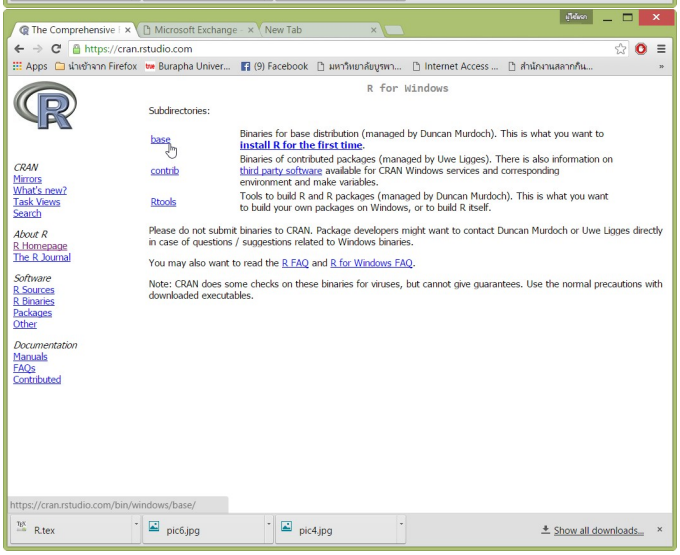
เนื่องจาก **R** เป็น open-source ทำให้ผู้ใช้สามารถหามาใช้ได้ง่าย นอกจาก **R** จะเหมาะกับนักศึกษาเพื่อใช้ในการเรียนรู้ทางสถิติแล้ว **R** ยังเหมาะกับนักวิจัยที่ต้องการใช้วิธีเชิงสถิติในการวิเคราะห์ข้อมูลอีกด้วย

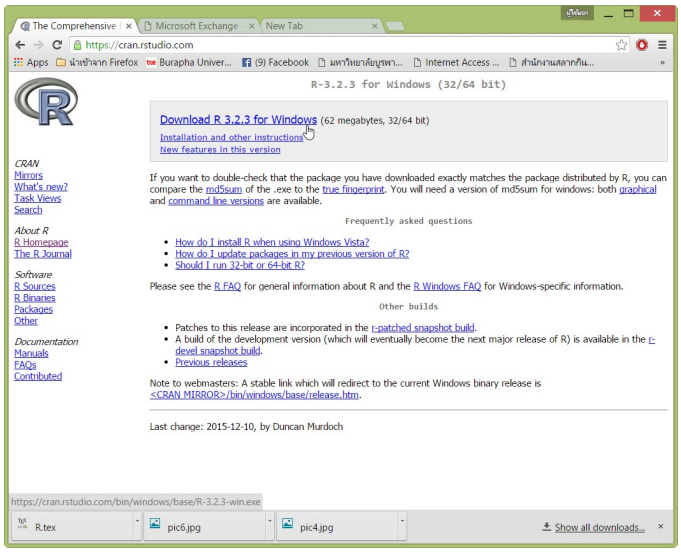
1.1 การดาวน์โหลดและติดตั้ง R

โปรแกรม **R** สามารถดาวน์โหลดได้จาก http://www.r-project.org/ เลือก Download CRAN (Comprehensive R Archive Network) จากนั้นเลือกดาวน์โหลดจากเวปไซต์ที่อยู่ใกล้กับเรามากที่สุด เพื่อความเร็วในการดาวน์โหลด เมื่อเลือก mirror แล้วให้เลือก Download R for Windows แล้วเลือก base เลือก Download R 3.2.3 for Windows แล้วบันทึกไฟล์นี้ไว้ในโฟล์เดอร์ที่ต้องการ

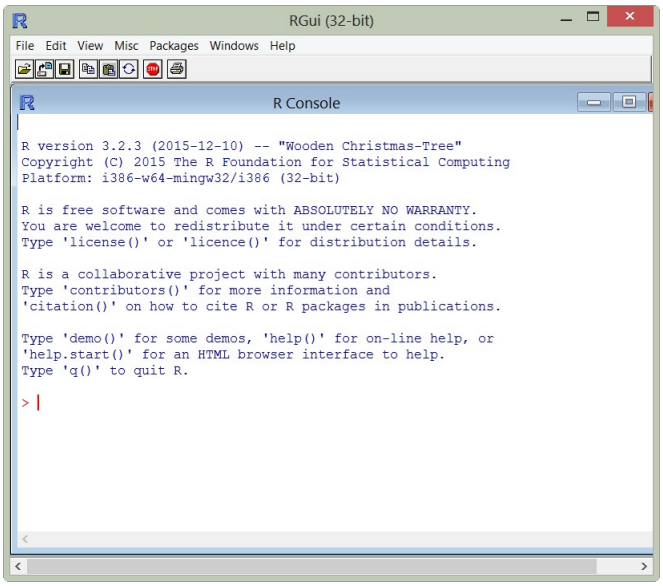

R: The R Project for S
Microsoft Exchange
New Tab
https://www.r-project.org
The R Project for Statistical Computing
Getting Started
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.
News
The R Journal Volume 7/2 is available.
R version 3.2.3 (Wooden Christmas-Tree) has been released on 2015-12-10.
R version 3.1.3 (Smooth Sidewalk) has been released on 2015-03-09.
useR! 2015, took place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
R.tex
pic6.jpg
pic4.jpg
Show all downloads...


The Comprehensive
https://cran.rstudio.com
The Comprehensive R Archive Network
Download and Install R
Precompiled binary distributions of the base system and contributed packages, Windows and Mac users most likely want one of these versions of R:
Download R for Linux
Download R for (Mac) OS X
Download R for Windows
R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.
Source Code for all Platforms
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!
The latest release (2015-12-10, Wooden Christmas-Tree) R-3.2.3.tar.gz, read what's new in the latest version.
Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
Source code of older versions of R is available here.
Contributed extension packages
Questions About R
If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
R.tex
pic6.jpg
pic4.jpg
Show all downloads...


The Comprehensive
https://cran.rstudio.com
R for Windows
Subdirectories:
base
Binaries for base distribution (managed by Duncan Murdoch). This is what you want to install R for the first time.
contrib
Binaries of contributed packages (managed by Uwe Ligges). There is also information on third party software available for CRAN Windows services and corresponding environment and make variables.
Rtools
Tools to build R and R packages (managed by Duncan Murdoch). This is what you want to build your own packages on Windows, or to build R itself.
Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.
You may also want to read the R FAQ and R for Windows FAQ.
Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.
CRAN
Mirrors
What's new?
Task Views
Search
About R
R Homepage
The R Journal
Software
R Sources
R Binaries
Packages
Other
Documentation
Manuals
FAQs
Contributed
https://cran.rstudio.com/bin/windows/base/
R.tex
pic6.jpg
pic4.jpg
Show all downloads...



เมื่อดาวน์โหลด **R** เรียบร้อยแล้วก็เริ่มทำการติดตั้ง **R** ซึ่งสำหรับระบบปฏิบัติการ Windows สามารถทำได้ง่ายมากโดยการเปิดไฟล์ R-3.2.3-win.exe ที่บันทึกไว้ แล้วคลิก Next ไปเรื่อย ๆ **R** จะถูกติดตั้งทันที เมื่อติดตั้ง **R** เรียบร้อยแล้ว เราสามารถเริ่มใช้งาน **R** ได้โดย double-click icon ของ **R** จะปรากฏหน้าต่างของโปรแกรม **R** ดังนี้



หน้าต่าง **R** จะแบ่งเป็นสองหน้าต่างย่อยคือ **R Console** และ **R Graphics** โดยหน้าต่าง **R** Console เป็นหน้าต่างสำหรับการเขียนคำสั่ง และแสดงผลลัพธ์ ส่วนหน้าต่าง **R** Graphics เป็นหน้าต่างสำหรับแสดงกราฟ ซึ่งจะปรากฏขึ้นเมื่อมีการสร้างกราฟเท่านั้น

Note ทุก ๆ 6 เดือน จะมีการออกรุ่นใหม่ของ **R** ผู้ใช้จึงควรอัพเกรดโปรแกรม **R** อยู่เสมอ

1.2 การใช้ R เบื้องต้น

1.2.1 การใช้ R เป็นเครื่องคิดเลข

การใช้ **R** แบบง่าย ๆ แบบหนึ่งคือการใช้เสมือนเป็นเครื่องคิดเลข **R** จะใช้เครื่องหมายทางคณิตศาสตร์ที่เราคุ้นเคยกันดีอยู่แล้ว เช่น +, -, * และ / และใช้ ^ เป็นเครื่องหมายของการยกกำลัง ตัวอย่างต่อไปนี้แสดงให้เห็นถึงการใช้ **R** เป็นเครื่องคิดเลข

```
> 2+2
[1] 4
> 2^2
[1] 4
> (1-2)*3
[1] -3
> -2*3
[1] -6
```

ผลที่ได้จากการคำนวณจะปรากฏในบรรทัดถัดมาพร้อมด้วย [1] (จะอธิบายถึงสัญลักษณ์นี้ต่อไปในเรื่องของ data vector)

Functions ใน **R** มีฟังก์ชันทางคณิตศาสตร์และสถิติที่สามารถนำมาใช้ได้มากมาย และมีวิธีใช้ที่คล้ายคลึงกับการใช้ฟังก์ชันทางคณิตศาสตร์ในเครื่องคิดเลข และโปรแกรมคอมพิวเตอร์อื่น ๆ เช่น EXCEL เป็นต้น เวลาใช้ฟังก์ชันเหล่านี้ ให้พิมพ์ชื่อของฟังก์ชันที่ต้องการใช้ ตามด้วยวงเล็บ, ( ) ซึ่งในวงเล็บนี้เราจะใส่ argument เข้าไป ตัวอย่างต่อไปนี้แสดงการใช้ฟังก์ชันใน **R**

```
> sqrt(2)
[1] 1.414214
> sin(pi)
[1] 1.224606e-16
> exp(1)
[1] 2.718282
> log(10)
[1] 2.302585
```

ฟังก์ชันหลายฟังก์ชันใน **R** มี argument ที่นอกเหนือจากที่ใช้ไปแล้ว ซึ่งทำให้เราสามารถเปลี่ยน defualt ที่กำหนดไว้ในฟังก์ชันได้ เช่น ฟังก์ชัน log() โดย default แล้วเป็น log ฐาน *e* แต่ถ้าเราต้องการใช้ log ฐาน 10 สามารถทำได้ดังนี้

```
> log(10,10)
[1] 1
> log(10,base=10)
[1] 1
```

สำหรับคำสั่งแรกจะใช้ได้เมื่อเราทราบอยู่แล้วว่า argument ที่สองเป็นการกำหนดฐานของ log ในคำสั่งที่สองเราใช้ชื่อของ argument ได้แก่ base = บอกให้ฟังก์ชันทราบว่าเราต้องการใช้ log ฐาน 10 แบบแรกนั้นมีประโยชน์คือประหยัดเวลาในการพิมพ์ แต่ในแบบที่สองนั้นง่ายต่อการจำและการอ่าน

1.2.2 Assignment

โดยปกติเรามักต้องการกำหนดชื่อให้กับค่าบางค่าเพื่อนำไปใช้ในคราวต่อไปได้โดยไม่ต้องพิมพ์ค่านั้นอีก เช่น

```
> x=2
> x+3
[1] 5
> e=exp(1)
> e^2
[1] 7.389056
```

การกำหนดชื่อตัวแปรสามารถใช้ได้ทั้งตัวอักษร ตัวเลข และ "." หรือ "_" แต่ไม่สามารถใช้เครื่องหมายของการคำนวณทางคณิตศาสตร์ เช่น +, -, *, / ได้ ตัวอย่างของการตั้งชื่อตัวแปร

```
> x=2
> n=25
> a.really.long.number = 123456789
> AReallySmallNumber = 0.000000001
```

การตั้งชื่อตัวแปรควรระวังในเรื่องของการใช้ตัวอักษรตัวใหญ่หรือตัวเล็ก เนื่องจาก **R** จะถือว่าตัวอักษรตัวใหญ่หรือตัวเล็กไม่ใช่ตัวแปรเดียวกัน

1.3 ชนิดและรูปแบบของข้อมูลใน R

ข้อมูลที่ใช้ใน **R** สามารถจำแนกได้เป็น 3 ชนิดใหญ่ ๆ คือ

Numeric เป็นข้อมูลที่มีค่าเป็นตัวเลข หรือในทางสถิติเรียกข้อมูลชนิดนี้ว่าข้อมูลเชิงปริมาณ (Quantitative data) เช่น อายุ รายได้ และคะแนนสอบ เป็นต้น ซึ่งข้อมูลชนิดนี้สามารถนำมาคำนวณหาค่าทางสถิติต่าง ๆ เช่น ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน เป็นต้น

Character เป็นข้อมูลที่ค่าของมันไม่ใช่ตัวเลข โดยปรกติมักเป็นข้อความ หรือตัวอักษร ในทางสถิติเรียกข้อมูลชนิดนี้ว่าข้อมูลเชิงคุณภาพ (Qualitative data) เช่น เพศ (ชาย, หญิง) ระดับการศึกษา (มัธยมศึกษา, ปริญญาตรี, ปริญญาโท) เป็นต้น ข้อมูลชนิดนี้ไม่สามารถนำมาคำนวณเพื่อหาค่าทางสถิติได้ แต่มีวิธีเชิงสถิติบางวิธีที่สามารถใช้ในการวิเคราะห์เพื่อหาคำตอบที่ผู้วิจัยต้องการจากข้อมูลชนิดนี้ได้ ซึ่งโดยทั่วไปจะใช้วิธีกลุ่มของสถิติไม่อิงพารามิเตอร์ (Nonparametric statistics)

Logical เป็นข้อมูลเชิงตรรกที่มีค่าอยู่สองค่าเท่านั้น คือ TRUE แทนค่าเป็นจริง หรือ FALSE แทนค่าที่เป็นเท็จ

ข้อมูลที่ใช้ใน **R** สามารถเก็บไว้ในรูปของ object ที่มีอยู่ด้วยกันหลายชนิด ได้แก่

Factor เป็นลักษณะการเก็บข้อมูลที่เป็นข้อมูลเชิงกลุ่ม (categorical data) หรือข้อมูลเชิงคุณภาพ ที่ไม่ได้มีค่าเป็นตัวเลขอย่างแท้จริง แต่ใช้ในการจำแนกข้อมูลออกเป็นกลุ่ม ๆ ตามลักษณะเฉพาะที่สนใจ เช่น เพศ (ชาย, หญิง) เป็นต้น

Vector ข้อมูลที่อยู่ในรูปของเวกเตอร์อาจเป็นตัวเลข หรือตัวอักษร ก็ได้ โดยสมาชิกในเวกเตอร์ต้องเป็นข้อมูลชนิดเดียวกัน ลักษณะการเก็บข้อมูลของเวกเตอร์ใน **R** จะเป็นเวกเตอร์แถว (row vector)

Matrix เป็นลักษณะของการเก็บข้อมูลในสองมิติ (2 dimensions) ข้อมูลที่อยู่ในรูปของเมทริกซ์อาจเป็นตัวเลข หรือตัวอักษรก็ได้ แต่สมาชิกทุกตัวในเมทริกซ์หนึ่งต้องเป็นข้อมูลชนิดเดียวกันเท่านั้น

Data frame เป็นลักษณะการเก็บข้อมูลในรูปตาราง ที่ประกอบด้วยเวกเตอร์หนึ่งหรือหลายเวกเตอร์มารวมกัน โดยแต่ละเวกเตอร์อาจเป็นข้อมูลชนิดเดียวกันหรือต่างชนิดกันก็ได้ แต่ความยาวของทุกเวกเตอร์ต้องเท่ากัน

List เป็นลักษณะการเก็บข้อมูลที่สามารถนำ object แต่ละแบบมารวมกันเก็บเป็น list ได้ จะเป็นข้อมูลชนิดเดียวกันหรือไม่ก็ได้ และไม่จำกัดในเรื่องของขนาดและความยาวของ objects ที่มารวมกัน ซึ่งโดยทั่วไปผลการวิเคราะห์ของ **R** มักอยู่ในรูปของ list

1.4 **R** Commander

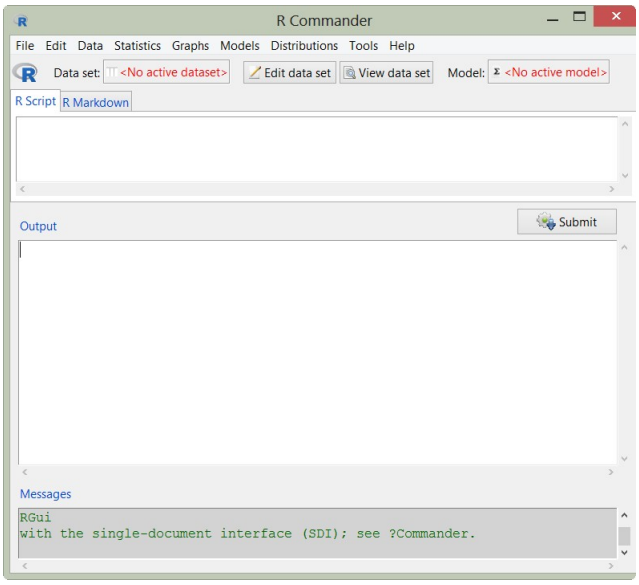
เนื่องจากการใช้ **R** ในรูปของ RGui ผู้ใช้มักประสบปัญหาในการใช้ฟังก์ชันต่าง ๆ เนื่องจากจำรูปแบบของคำสั่งไม่ได้ หรือกำหนดค่า argument ต่าง ๆ ไม่ถูกต้อง ทำให้ผู้เริ่มต้นใช้ **R** ในการวิเคราะห์ข้อมูลเชิงสถิติมีความรู้สึกว่า **R** เป็นโปรแกรมที่ยุ่งยากซับซ้อน และผู้ใช้จำเป็นต้องมีความรู้ในเรื่องของการโปรแกรมพอสมควร ซึ่ง Prof. John Fox ได้ทราบถึงปัญหานี้ จึงได้พัฒนาแพคเกจของ **R** ชื่อว่า [Rcmdr] ซึ่งช่วยให้การทำงานใน **R** ง่ายขึ้น โดยแพคเกจนี้จะสร้างอินเตอร์เฟส (interface) ที่มีลักษณะการทำงานแบบเมนู คล้ายคลึงกับโปรแกรมสำเร็จรูปทางสถิติอื่น ๆ เช่น SPSS หรือ MINITAB ดังนั้น ผู้ใช้จึงสามารถใช้ **R** ในการวิเคราะห์ข้อมูลได้ง่ายขึ้น อย่างไรก็ตาม หากผู้ใช้ต้องการทำงานที่มีความซับซ้อน หรืองานที่นอกเหนือจากที่ Rcmdr ทำได้ ก็จำเป็นต้องใช้ RGui เช่นเดิม

1.4.1 การติดตั้ง **R** Commander

1. เปิดโปรแกรม **R** ที่ติดตั้งไว้แล้วขึ้นมา ปรากฏหน้าต่าง RGui
2. เลือกเมนู **Packages** -> **Install package(s)...** จะปรากฏหน้าต่าง CRAN mirror
3. ให้เลือก mirror ที่อยู่ใกล้ที่สุด นั่นคือของประเทศไทย จากนั้นคลิก OK จะปรากฏหน้าต่าง Packages
4. เลือกแพคเกจชื่อ Rcmdr จากนั้นคลิก OK
5. โปรแกรมจะทำการติดตั้ง **R** commander ให้ โดยจะมีข้อความถามว่าต้องการติดตั้งแพคเกจที่ใช้ร่วมกับ **R** commander หรือไม่ ให้ตอบ YES จากนั้นการติดตั้งจะเสร็จสมบูรณ์

1.4.2 การใช้ **R** Commander

1. เปิดโปรแกรม **R**
2. ไปที่เมนู **Packages** -> **Load package** จะปรากฏหน้าต่าง **Select one** -> เลือก **Rcmdr** จะปรากฏหน้าต่าง **R Commander** ดังนี้



Note: เราสามารถกำหนดให้ **R** โหลด **R** Commander อัตโนมัติทุกครั้งที่มีการเปิดโปรแกรม **R** ได้โดยการเพิ่มคำสั่งต่อไปนี้ลงในไฟล์ชื่อ **Rprofile.site** ในไดเรกทอรี่ย่อยของ **R** ชื่อ etc

```
local({
old <- getOption("defaultPackages")
options(defaultPackages = c(old, "Rcmdr"))
})
```

หน้าต่างของ **R** Commander ประกอบด้วยหน้าต่างย่อยอีก 4 หน้าต่าง คือ

Script Window: ที่หน้าต่างนี้ผู้ใช้สามารถพิมพ์คำสั่ง **R** ได้ที่นี่ จากนั้นคลิก Submit แล้ว **R** จะทำการประมวลผลคำสั่งที่ได้รับ นอกจากนั้น ผู้ใช้สามารถเลือกการใช้คำสั่งต่าง ๆ ในการวิเคราะห์ข้อมูลจากเมนูที่อยู่ด้านบนของหน้าต่างได้ และสามารถบันทึกชุดคำสั่งที่ใช้ลงไฟล์เพื่อเรียกใช้ในคราวต่อไปได้ หรือสามารถเปิดไฟล์คำสั่ง (script file) ที่มีอยู่แล้วมาใช้ก็ได้

R Markdown: ใช้สำหรับการสร้างรายงานการทำงานของผู้ใช้

Output Window: เป็นพื้นที่สำหรับแสดงผลการทำงานตามคำสั่งจาก Script Window

Message Window: เป็นพื้นที่แสดงข้อความต่าง ๆ ที่เกิดจากการใช้งานคำสั่ง โดยข้อความที่เป็นสีแดงคือ Error message ข้อความสีเขียวคือ Warnings และข้อความสีน้ำเงินคือ ข้อมูลอื่นๆ

Note: กราฟจะปรากฏในอีกหน้าต่างหนึ่งแยกต่างหาก เป็น **R Graphics Device Window** โดยกราฟที่ปรากฏจะเป็นกราฟรูปล่าสุดที่ประมวลผล หากต้องการดูกราฟก่อนหน้าให้ใช้ page up และ page down keys

ที่ด้านบนของหน้าต่าง **R** Commander ประกอบด้วยเมนูต่างๆ ดังนี้

File ประกอบด้วยคำสั่งที่ใช้ในการเปิด หรือบันทึกไฟล์ประเภทต่าง (Open, Save) เช่น ไฟล์คำสั่ง และไฟล์ผลลัพท์ และยังมีคำสั่งที่ใช้ในการออกจากโปรแกรมด้วย (Exit)

Edit ประกอบด้วยคำสั่งตัด คัดลอก วาง (Cut, Copy, Paste, เป็นต้น) ที่ใช้ในการแก้ไขรายละเอียดของหน้าต่าง script หรือ output

Data ประกอบด้วยเมนูย่อย ที่มีคำสั่งในการสร้างข้อมูลใหม่ อ่านข้อมูลจากไฟล์ หรือจัดการข้อมูล

Statistics ประกอบด้วยเมนูย่อยที่มีคำสั่งเกี่ยวกับการวิเคราะห์เชิงสถิติเบื้องต้นต่าง ๆ

Graphs ประกอบด้วยคำสั่งในการสร้างกราฟทางสถิติ

Models ประกอบด้วยเมนูย่อยและคำสั่งในการหาค่าสรุปทางสถิติ การสร้างช่วงความเชื่อมั่น การทดสอบสมมติฐาน การตรวจสอบข้อสมมติเบื้องต้นของตัวแบบทางสถิติ และกราฟต่าง ๆ ที่ใช้ในการตรวจสอบข้อสมมติของตัวแบบทางสถิติ

Distributions ประกอบด้วยคำสั่งเกี่ยวกับการแจกแจงความน่าจะเป็นทางสถิติ

Tools ประกอบด้วยคำสั่งในการโหลดแพคเกจของ **R** และ Rcmdr plug-in มาใช้งาน

Help ให้ข้อมูลเกี่ยวกับ **R** Commander และความช่วยเหลือในการใช้ **R** Commander

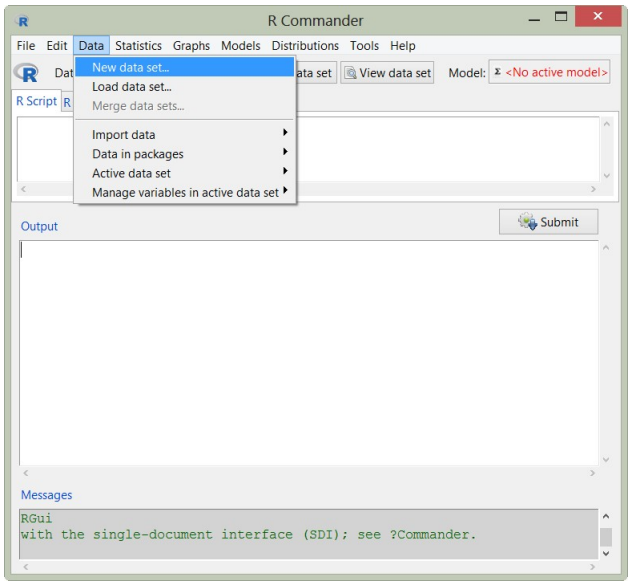
1.5 การนำข้อมูลมาใช้ด้วย **R** Commamder

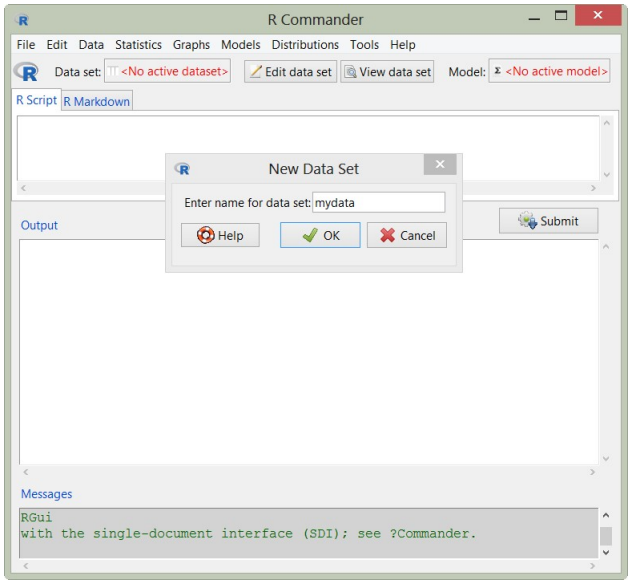
การนำข้อมูลเข้ามาใช้ใน **R** Commander สามารถทำได้หลายวิธีด้วยกัน เช่น การป้อนข้อมูลเข้าไปใหม่ การเรียกไฟล์ข้อมูลที่มีอยู่แล้วในรูปแบบต่าง ๆ เข้ามาใช้งาน

1.5.1 การป้อนข้อมูลใหม่

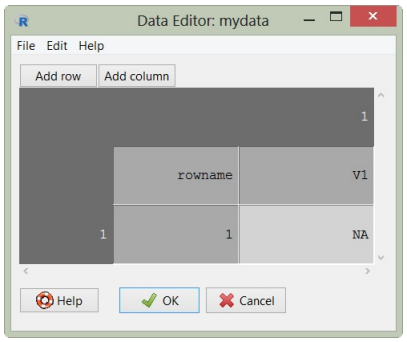
การป้อนข้อมูลใหม่โดยตรงเป็นวิธีที่เหมาะกับข้อมูลที่มีขนาดเล็ก โดยทำได้ดังนี้

1. เลือกเมนู **Data** -> **New data set** จะปรากฏหน้าต่าง **New Data Set**



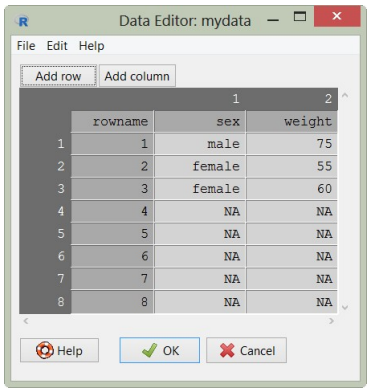


2. ใส่ชื่อของชุดข้อมูลลงไป -> **OK** จะปรากฏหน้าต่าง **Data Editor** ที่มีลักษณะเป็น spreadsheet



3. ที่หน้าต่าง **Data Editor** เราป้อนข้อมูลลงไปได้ตามต้องการ โดยการกำหนดชื่อตัวแปรให้คลิกที่ชื่อคอลัมน์ (V1) และสามารถพิมพ์ชื่อตัวแปรที่ต้องการลงไปแทนที่ได้เลย และสามารถเพิ่มคอลัมน์ได้ในกรณีที่ข้อมูลประกอบด้วยตัวแปรหลายตัว โดยคลิกที่ปุ่ม Add column

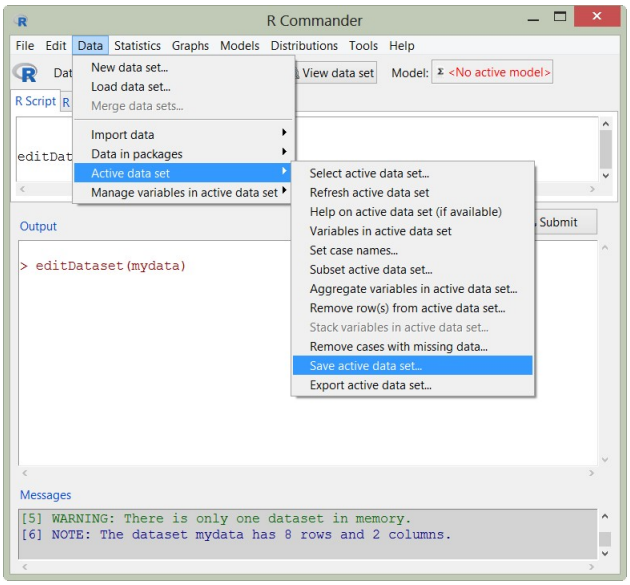
เมื่อกำหนดชื่อตัวแปรเรียบร้อยแล้ว ผู้ใช้สามารถป้อนข้อมูลในหน้าต่างนี้ได้เลย โดยป้อนข้อมูลลงไปตาม cell ในตารางข้อมูล



	rowname	sex	weight
1	1	male	75
2	2	female	55
3	3	female	60
4	4	NA	NA
5	5	NA	NA
6	6	NA	NA
7	7	NA	NA
8	8	NA	NA

4. เมื่อป้อนข้อมูลเสร็จ ชุดข้อมูลนี้จะอยู่ในรูปของ data frame และพร้อมใช้งาน จากนั้นให้ปิดหน้าต่างนี้ ชุดข้อมูลที่สร้างไว้นี้สามารถเรียกมาดูหรือแก้ไขได้โดยการคลิกเรียกที่ปุ่ม Edit data set

5. ข้อมูลชุดที่สร้างไว้นี้มีอยู่ในหน่วยความจำเท่านั้น ยังไม่ได้ถูกบันทึกเป็นไฟล์ การบันทึกชุดข้อมูลลงไฟล์ให้เลือกเมนู **Data** -> **Active data set** -> **Save active data set...**

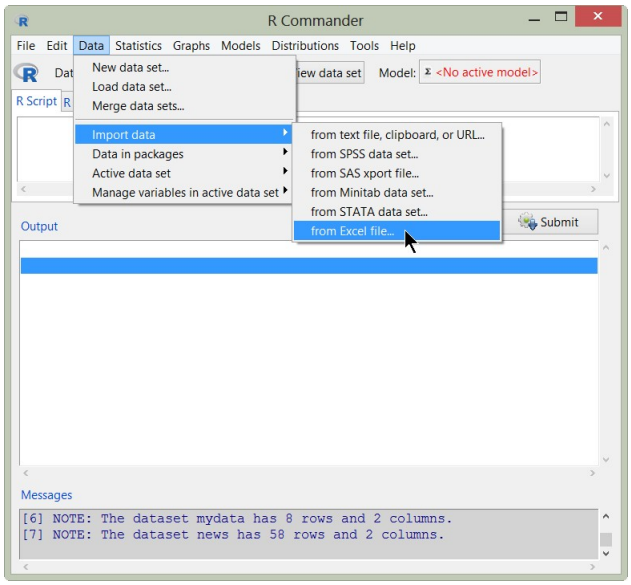


1.5.2 การนำข้อมูลเข้ามาใช้ใน R จาก Textfile, EXCEL, SPSS และ MINITAB

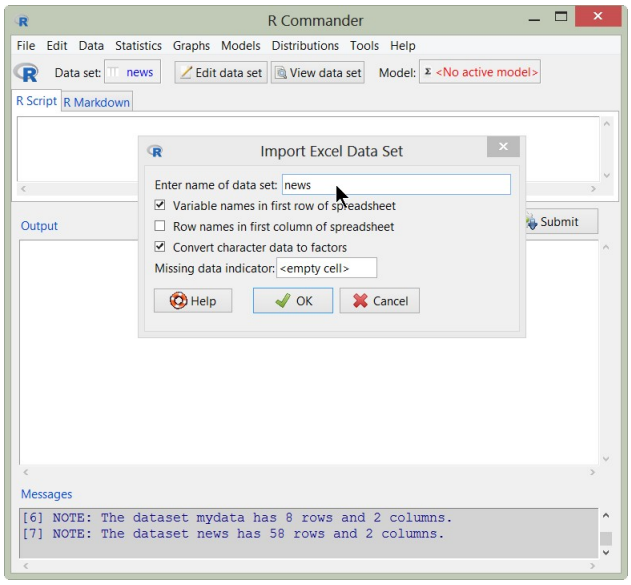
โดยส่วนใหญ่ผู้ใช้มักมีการบันทึกข้อมูลไว้ก่อนที่จะทำการวิเคราะห์เชิงสถิติแล้ว ซึ่งไฟล์เหล่านั้นอาจอยู่ในรูปแบบต่าง ๆ เช่น textfile SPSS EXCEL หรือ MINITAB การเรียกข้อมูลที่อยู่ในไฟล์รูปแบบเหล่านี้มาใช้งานใน **R** สามารถทำได้โดยง่ายตามชนิดของไฟล์ สำหรับไฟล์ข้อมูลที่เป็น textfile ข้อมูลต้องถูกจัดอยู่ในรูปแบบของ data frame โดยแต่ละคอลัมน์คือแต่ละตัวแปร เช่น เพศ อายุ รายได้ เป็นต้น แต่ละแถวคือแต่ละ case หรือข้อมูลของแต่ละคน และแถวบนสุดของไฟล์คือชื่อตัวแปร

ส่วนไฟล์ข้อมูลที่เป็น .xlsx ไฟล์ มีวิธีการ import file ดังนี้

1. **Data -> Import data -> from Excel file...**



ปรากฏหน้าต่าง



2. ใส่ชื่อของชุดข้อมูลสำหรับการเรียกใช้ใน **R** จากนั้นคลิก OK
3. เลือกโฟลเดอร์ และชื่อไฟล์ที่ต้องการเปิด และเปิดไฟล์ข้อมูล
4. ตรวจสอบว่าไฟล์ข้อมูลที่เปิดนั้นถูกต้อง คือครบทุกคอลัมน์ ทุกแถวหรือไม่
5. เรียกดูชุดข้อมูลโดยใช้ **View data set** ที่หน้าต่าง **R** Commander

ส่วนไฟล์ประเภทอื่น ๆ ก็ทำได้เช่นเดียวกัน

บทที่ 2

การใช้ R Commander ในการวิเคราะห์ข้อมูลเบื้องต้น

ในบทนี้จะศึกษาถึงการใช้ **R** Commander ในการวิเคราะห์ข้อมูลเบื้องต้น ซึ่งมีอยู่ด้วยกันหลายวิธี เช่น การแจกแจงความถี่ การคำนวณหาค่าสถิติต่าง ๆ เพื่ออธิบายเกี่ยวกับลักษณะของข้อมูล เช่น ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน เป็นต้น นอกจากนั้นยังมีการใช้กราฟรูปแบบต่าง ๆ ในการนำเสนอข้อมูลด้วย

2.1 ข้อมูลและชนิดของข้อมูล

ก่อนที่จะเริ่มทำการวิเคราะห์ข้อมูลในทางสถิติ เราต้องทำความเข้าใจเกี่ยวกับชนิดของข้อมูลก่อน เนื่องจากข้อมูลแต่ละชนิดจะมีวิธีการวิเคราะห์ที่แตกต่างกัน ซึ่งหากเราใช้วิธีเชิงสถิติที่ไม่สอดคล้องกับชนิดของข้อมูล จะทำให้ผลการวิเคราะห์ที่ได้ไม่ถูกต้อง ไม่น่าเชื่อถือ และไม่สามารถนำมาใช้ประโยชน์ได้ ในทางสถิติได้แบ่งชนิดของข้อมูลออกเป็นสองกลุ่มใหญ่ ๆ คือ

1. ข้อมูลเชิงคุณภาพหรือข้อมูลเชิงกลุ่ม (Qualitative variable) ค่าของข้อมูลชนิดนี้ไม่ใช่ตัวเลข แต่เป็นข้อความ เช่น เพศ (ชาย, หญิง) ระดับการศึกษา (ต่ำกว่าปริญญาตรี, ปริญญาตรี, ปริญญาโท, สูงกว่าปริญญาโท) เป็นต้น

2. ข้อมูลเชิงปริมาณ (Quantitative variable) ข้อมูลชนิดนี้จะมีค่าเป็นตัวเลขอย่างแท้จริง อาจเป็นค่าแบบต่อเนื่อง หรือไม่ต่อเนื่องก็ได้ เช่น อายุ รายได้ คะแนนสอบ จำนวนลูกค้า เป็นต้น

2.2 การเตรียมข้อมูลเพื่อการวิเคราะห์

เครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูลมีอยู่หลายชนิด เช่น แบบสอบถาม แบบสัมภาษณ์ และแบบสังเกต เป็นต้น ไม่ว่าผู้วิจัยจะใช้เครื่องมือใดในการเก็บรวบรวมข้อมูล เมื่อได้ข้อมูลมาแล้วจะต้องมีการจัดเตรียมก่อนเริ่มต้นวิเคราะห์ข้อมูล ในที่นี้จะสมมติว่าใช้แบบสอบถามเป็นเครื่องมือ ดังตัวอย่างต่อไปนี้

การสร้างรหัส และการกำหนดชื่อตัวแปร

ตัวอย่างแบบสอบถาม

สำหรับเจ้าหน้าที่
[ ] [ ] ID

ส่วนที่ 1 ข้อมูลส่วนบุคคล

1. เพศ [ ] 1. ชาย [ ] 2. หญิง [ ] SEX

2. อายุ _____ ปี [ ] AGE

3. สถานภาพ [ ] STATUS
[ ] 1. โสด [ ] 2. แต่งงาน [ ] 3. หม้าย/หย่า

จากตัวอย่าง ตัวแปร ID คือหมายเลขแบบสอบถาม มี 2 ช่อง หมายความว่าตัวเลขที่ใช้จะเป็นตัวเลขสองหลัก

ในข้อ 1. ตัวแปร SEX ใช้เลข 1 แทนเพศชาย และเลข 2 แทนเพศหญิง

ในข้อ 2. ตัวแปร AGE ใช้ตัวเลขอายุที่ผู้ตอบแบบสอบถามตอบ

ในข้อ 3. ตัวแปร STATUS ใช้เลข 1 แทนโสด 2 แทนแต่งงาน 3 แทนหม้ายหรือหย่า

ตัวอย่างการลงรหัสแบบสอบถาม

สำหรับเจ้าหน้าที่
[0] [1] ID

ส่วนที่ 1 ข้อมูลส่วนบุคคล

1. เพศ [/] 1. ชาย [ ] 2. หญิง [1] SEX

2. อายุ 20 ปี [20] AGE

3. สถานภาพ [1] STATUS
[/] 1. โสด [ ] 2. แต่งงาน [ ] 3. หม้าย/หย่า

แบบสอบถามที่ข้อคำถามผู้ตอบสามารถเลือกคำตอบหลายคำตอบ

โปรดเลือกรายการโทรทัศน์ที่ท่านชอบดู (เลือกได้มากกว่า 1 ข้อ)	
[ ] 1. ข่าว	[ ] TV1
[ ] 2. สารคดี	[ ] TV2
[ ] 3. ละคร	[ ] TV3
[ ] 4. เกมโชว์	[ ] TV4

จากตัวอย่างนี้จะเห็นได้ว่า ผู้ตอบสามารถเลือกตอบได้มากกว่า 1 ตัวเลือก ดังนั้นการลงรหัสจะให้เลข 1 แทนความหมายที่ผู้ตอบเลือกตอบตัวเลือกนั้น และ 0 แทนความหมายที่ผู้ตอบไม่เลือกตอบตัวเลือกนั้น ดังนี้

โปรดเลือกรายการโทรทัศน์ที่ท่านชอบดู (เลือกได้มากกว่า 1 ข้อ)	
[/] 1. ข่าว	[1] TV1
[/] 2. สารคดี	[1] TV2
[ ] 3. ละคร	[0] TV3
[ ] 4. เกมโชว์	[0] TV4

ข้อคำถามที่ให้ผู้ตอบเรียงลำดับข้อคำตอบ

ให้ท่านเรียงลำดับรายการโทรทัศน์ที่ท่านชอบมากที่สุดเป็นลำดับที่ 1 และรายการที่ท่านชอบรองลงมาเป็นลำดับที่ 2 3 และ 4	
[2] 1. ข่าว	[2] TV1
[1] 2. สารคดี	[1] TV2
[3] 3. ละคร	[3] TV3
[4] 4. เกมโชว์	[4] TV4

จากตัวอย่าง เป็นการลงรหัสโดยใช้อันดับที่เลือก เช่น ผู้ตอบเลือกสารคดีเป็นอันดับที่ 1 จึงใส่ 1 ในช่อง TV2 เลือกข่าวเป็นอันดับที่ 2 จึงใส่ 2 ในช่อง TV1 เลือกละครเป็นอันดับที่ 3 จึงใส่ 3 ในช่อง TV3 และเลือกเกมโชว์เป็นอันดับที่ 4 จึงใส่ 4 ในช่อง TV4

หลังจากที่กำหนดวิธีการลงรหัสแบบสอบถามเรียบร้อยแล้ว ซึ่งขั้นตอนนี้ควรทำก่อนการเก็บข้อมูล เมื่อเก็บข้อมูลเรียบร้อยแล้ว นำแบบสอบถามที่รวบรวมได้ทั้งหมดมาลงรหัส โดยสามารถทำได้โดยใช้โปรแกรมที่เหมาะกับการเก็บข้อมูล เช่น **R** Commander EXCEL SPSS หรือ Textfile ก็ได้

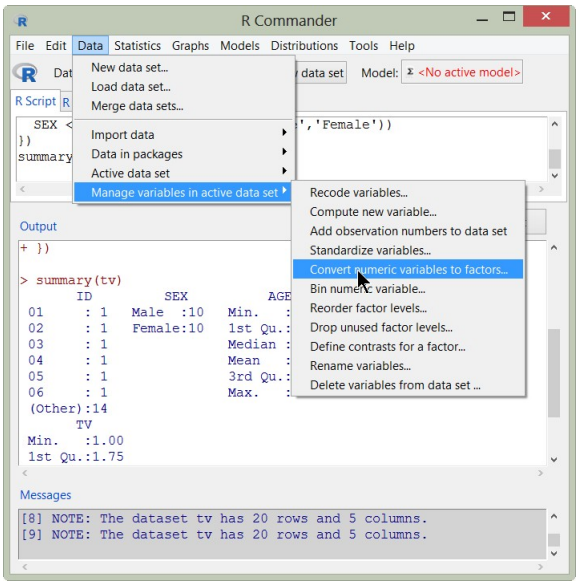
สมมติได้ข้อมูลดังนี้

ID	SEX	AGE	STATUS
01	1	28	1
02	2	35	2
03	1	29	1
04	1	32	3
05	2	34	2
06	1	28	2
07	2	25	1
08	1	32	1
09	2	33	1
10	2	38	2
11	2	39	3
12	2	34	3
13	2	32	1
14	1	26	1
15	1	27	2
16	2	36	2
17	1	32	2
18	1	33	1
19	1	29	1
20	2	25	3

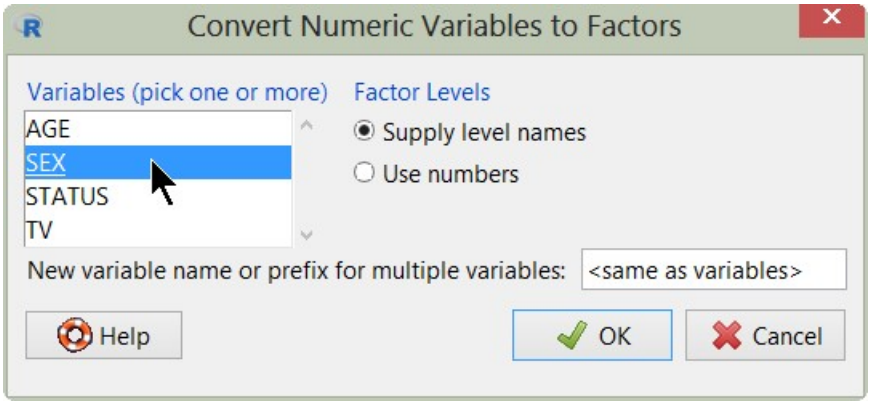
เมื่อนำข้อมูลชุดนี้มาวิเคราะห์ต้องมีการกำหนดชนิดของตัวแปรก่อน ซึ่งในข้อมูลชุดนี้ ตัวแปร SEX และ STATUS เป็นตัวแปรเชิงกลุ่ม ส่วนตัวแปร AGE เป็นตัวแปรเชิงปริมาณ เมื่อใช้ **R** Commander จึงต้องมีการกำหนดชนิดของตัวแปร SEX และ STATUS ให้เป็น Factor ดังนี้

1. เลือกเมนู **Data -> Manage variables in active data set -> Convert numeric variable to factors...**

   จะได้หน้าต่าง

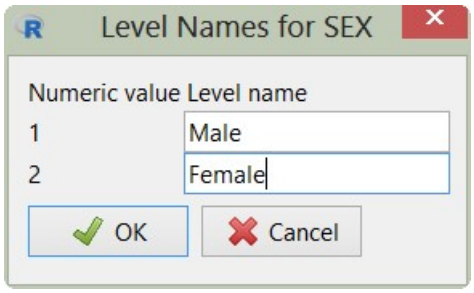


2. เลือกตัวแปรที่ต้องการเปลี่ยนให้เป็น factor เช่น SEX



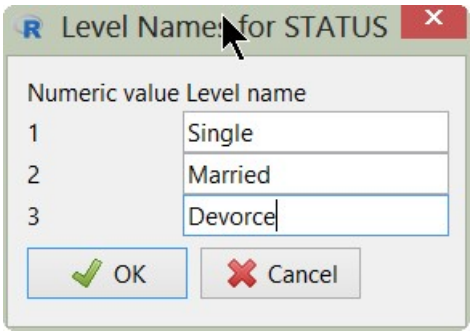
3. ผู้ใช้สามารถกำหนดให้ **R** เก็บค่าของตัวแปรที่เปลี่ยนแปลงแล้วในตัวแปรใหม่ก็ได้ หรือเก็บไว้ในชื่อเดิมก็ได้ ในตัวอย่างนี้จะเก็บไว้ในชื่อเดิม

4. หากคลิกเลือกที่ช่อง Supply level names แล้วคลิก **OK** จะได้หน้าต่างสำหรับการกำหนดชื่อของแต่ละระดับของตัวแปร ดังนี้



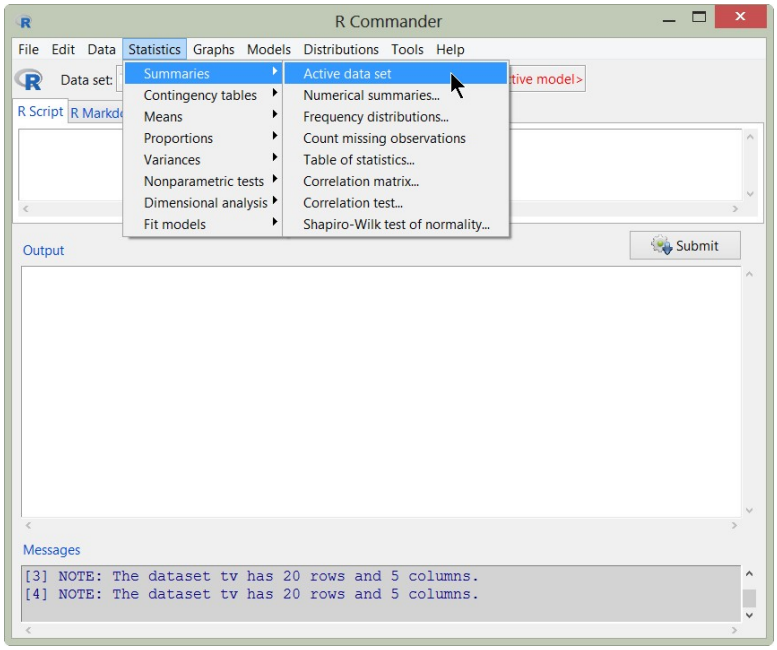
จากตัวอย่าง กำหนดให้ 1 เป็น Male และ 2 เป็น Female

สำหรับตัวแปร STATUS กำหนดดังนี้



ซึ่งจากตัวอย่าง กำหนดให้ 1 เป็น Single ให้ 2 เป็น Married และ 3 เป็น Divorce เราสามารถตรวจสอบชนิดของตัวแปรได้ด้วยการใช้คำสั่ง

Statistics -> Summaries -> Active data set



ได้ผลดังนี้

ID		SEX		AGE		STATUS		TV	
1	: 1	Male	:10	Min.	:25.00	Single	:9	Min.	:1.00
2	: 1	Female	:10	1st Qu.	:28.00	Married	:7	1st Qu.	:1.75
3	: 1			Median	:32.00	Divorce	:4	Median	:2.00
4	: 1			Mean	:31.35			Mean	:2.40
5	: 1			3rd Qu.	:34.00			3rd Qu.	:3.00
6	: 1			Max.	:39.00			Max.	:4.00
(Other)	:14								

จากผลที่ได้ จะเห็นได้ว่าตัวแปร ID เป็นตัวอักษร (character) ซึ่งไม่สามารถนำมาวิเคราะห์ได้ ตัวแปร SEX และ STATUS เป็นตัวแปรเชิงกลุ่มหรือ factor แบ่งเป็นระดับตามที่เรากำหนด และ AGE เป็นตัวแปรเชิงปริมาณมีค่าต่ำสุดเท่ากับ 25 ค่าสูงสุดเท่ากับ 39 และค่าเฉลี่ยเท่ากับ 31.35 เป็นต้น

ในขั้นตอนนี้ หากตรวจสอบแล้วพบว่าตัวแปรในชุดข้อมูลไม่ได้มีชนิดของตัวแปรตามที่ควรจะเป็น ก็สามารถทำการเปลี่ยนได้

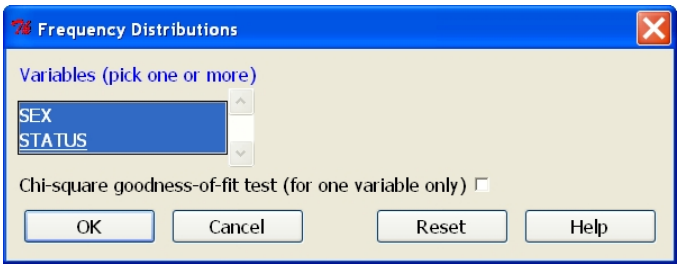
2.3 การแจกแจงความถี่

การแจกแจงความถี่เป็นวิธีเชิงสถิติที่ใช้กับตัวแปรเชิงกลุ่ม หรือตัวแปรเชิงคุณภาพ

2.3.1 การใช้ R Commander สำหรับการแจกแจงความถี่แบบทางเดียว

1. เลือกเมนู **Statistics -> Summaries -> Frequency distributions...**

   จะปรากฏหน้าต่าง Frequency Dsitributions ดังนี้



   ตัวแปรที่ปรากฏจะเป็นตัวแปรเชิงกลุ่มเท่านั้น

2. เลือกตัวแปรที่ต้องการแจกแจงความถี่ ในที่นี้จะเลือกตัวแปร SEX และ STATUS ได้ผลการวิเคราะห์ดังนี้

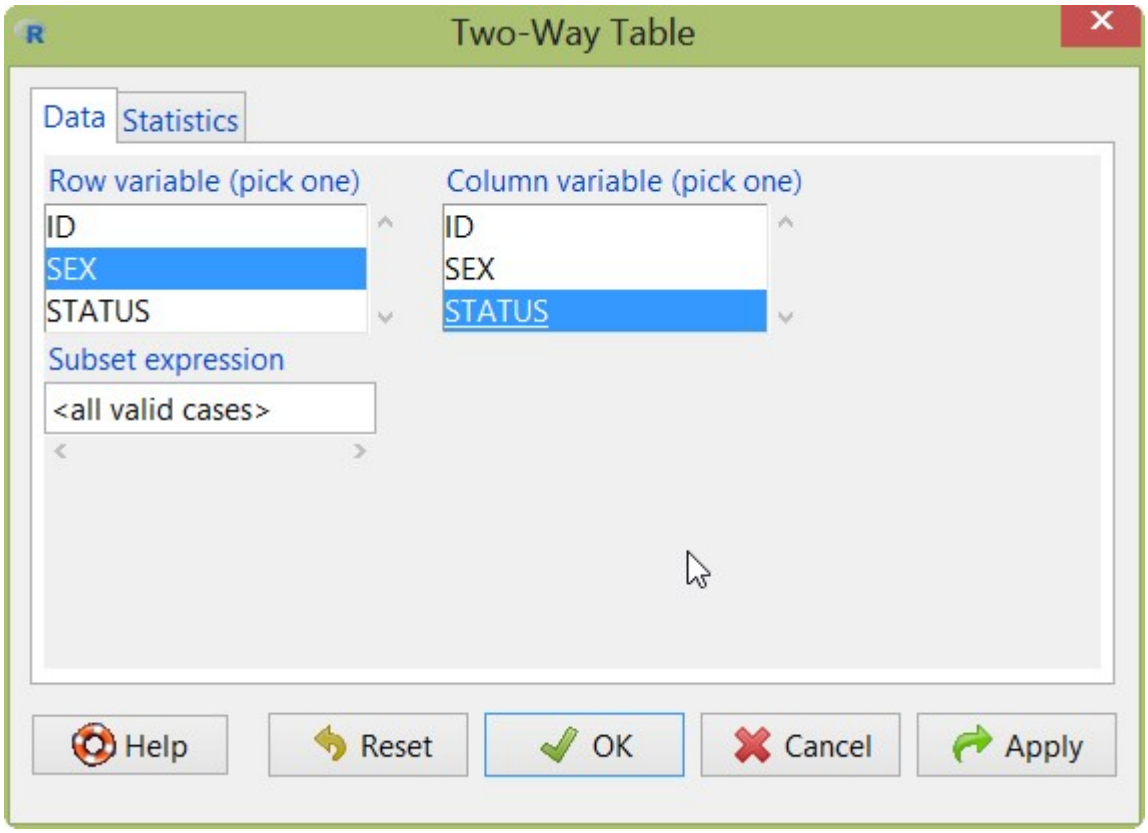
```
Male Female
  10     10

Male Female
  50     50

Single Married Divorce
   9       7       4

Single Married Divorce
  45      35      20
```

โดย output ในส่วนแรกจะแสดงความถี่ของแต่ละกลุ่ม และ output ส่วนที่สองจะแสดงร้อยละของแต่ละกลุ่ม

2.3.2 การใช้ R Commander สำหรับการแจกแจงความถี่แบบสองทาง

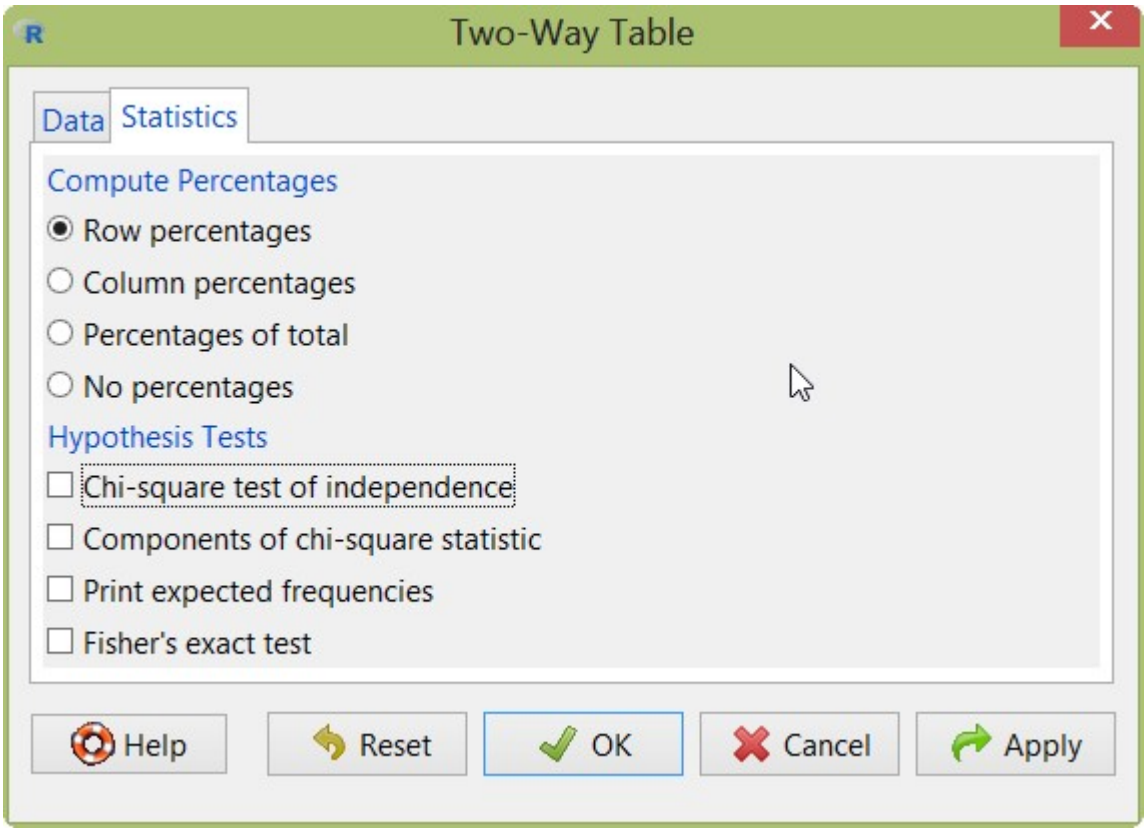
1. เลือกเมนู **Statistics -> Contingency tables -> Two-way tables...**

   จะปรากฏหน้าต่าง **Two-Way Table** ดังนี้



2. ที่ช่อง **Row variable** ให้เลือกตัวแปรที่ต้องการไว้ด้านแถวของตาราง จากตัวอย่างเลือก SEX ที่ช่อง **Column variable** ให้เลือกตัวแปรที่ต้องการไว้ด้านคอลัมน์ จากตัวอย่างเลือก STATUS

3. คลิกที่ปุ่ม Statistics ปรากฏหน้าต่าง



เลือกการแสดงร้อยละตามต้องการ ต้องการร้อยละทางด้านใด ให้คลิกที่ช่องที่ต้องการ

ร้อยละด้านแถว เลือก **Row percentages**

ร้อยละด้านคอลัมน์ เลือก **Column percenterages**

ร้อยละรวม เลือก **Percentages of total**

ไม่ต้องการร้อยละ เลือก **No percentages**

ในที่นี้เลือกร้อยละด้านแถว ได้ผลดังนี้

```
STATUS
SEX      Single Married Divorce
  Male        6       3       1
  Female      3       4       3

> rowPercents(.Table) # Row Percentages
         STATUS
SEX      Single Married Divorce Total Count
  Male       60      30      10   100    10
  Female     30      40      30   100    10
```

ตารางบนแสดงความถี่หรือจำนวนนับในแต่ละกลุ่ม ตารางล่างแสดงค่าร้อยละด้านแถว

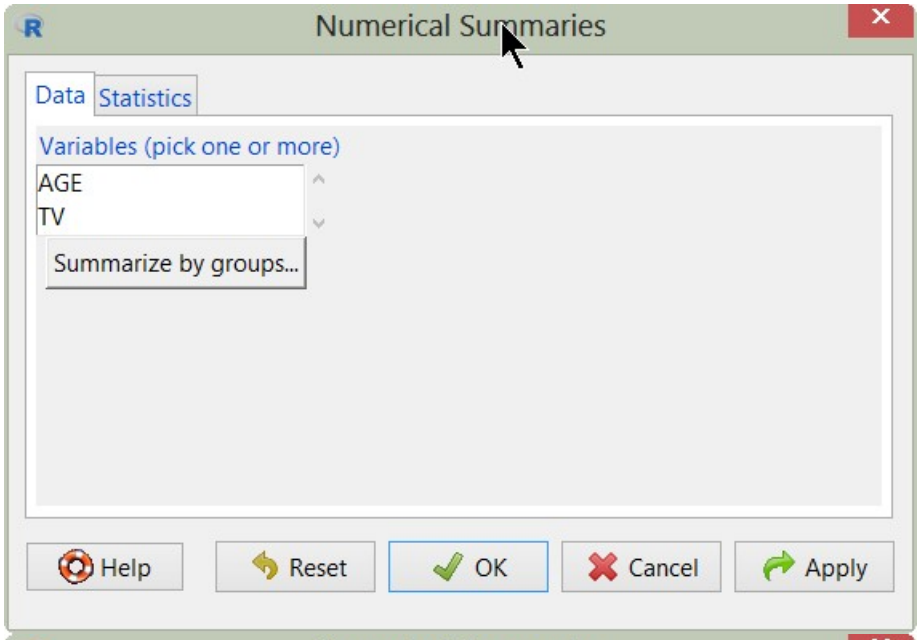
2.4 การคำนวณค่าสถิติเบื้องต้น

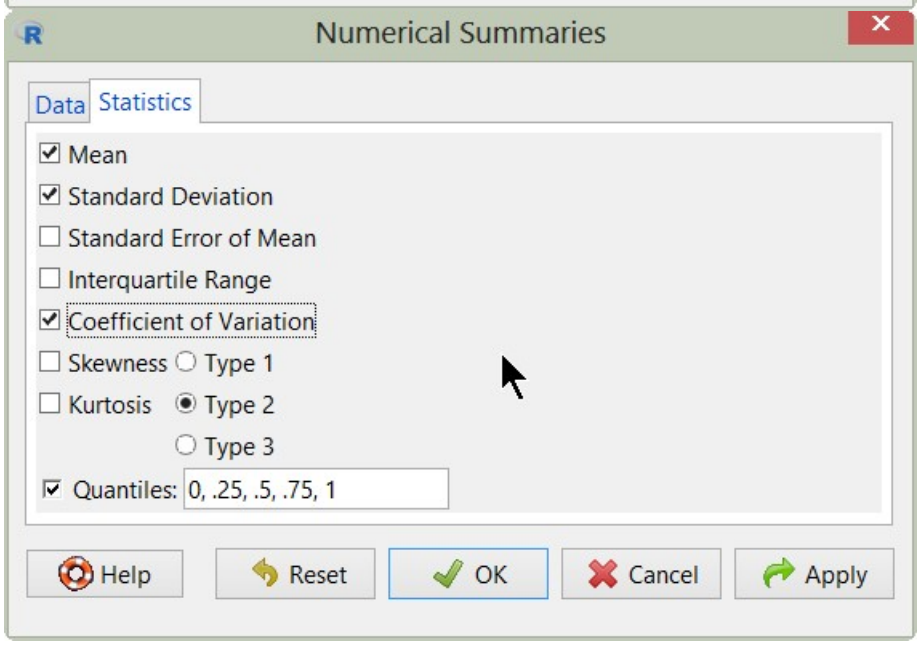
สำหรับข้อมูลเชิงปริมาณ เราสามารถอธิบายถึงลักษณะของข้อมูลเบื้องต้นได้ โดยใช้ค่าสถิติ เช่น ค่าเฉลี่ย ค่ามัธยฐาน และค่าเบี่ยงเบนมาตรฐาน เป็นต้น

การใช้ **R** Commander ในการหาค่าสถิติเหล่านี้ ทำได้ดังนี้

1. เลือกเมนู **Statistics -> Summaries -> Numerical Summaries...**

   จะได้หน้าต่างต่อไปนี้





2. เลือกตัวแปรที่ต้องการคำนวณค่าสถิติเบื้องต้น โดยคลิกที่ตัวแปรนั้น จะเห็นว่า **R** แสดงเฉพาะตัวแปรที่มีชนิดเป็น numeric เท่านั้น

3. เลือกค่าสถิติที่ต้องการหา

4. คลิก **OK** ได้ผลลัพธ์ดังนี้

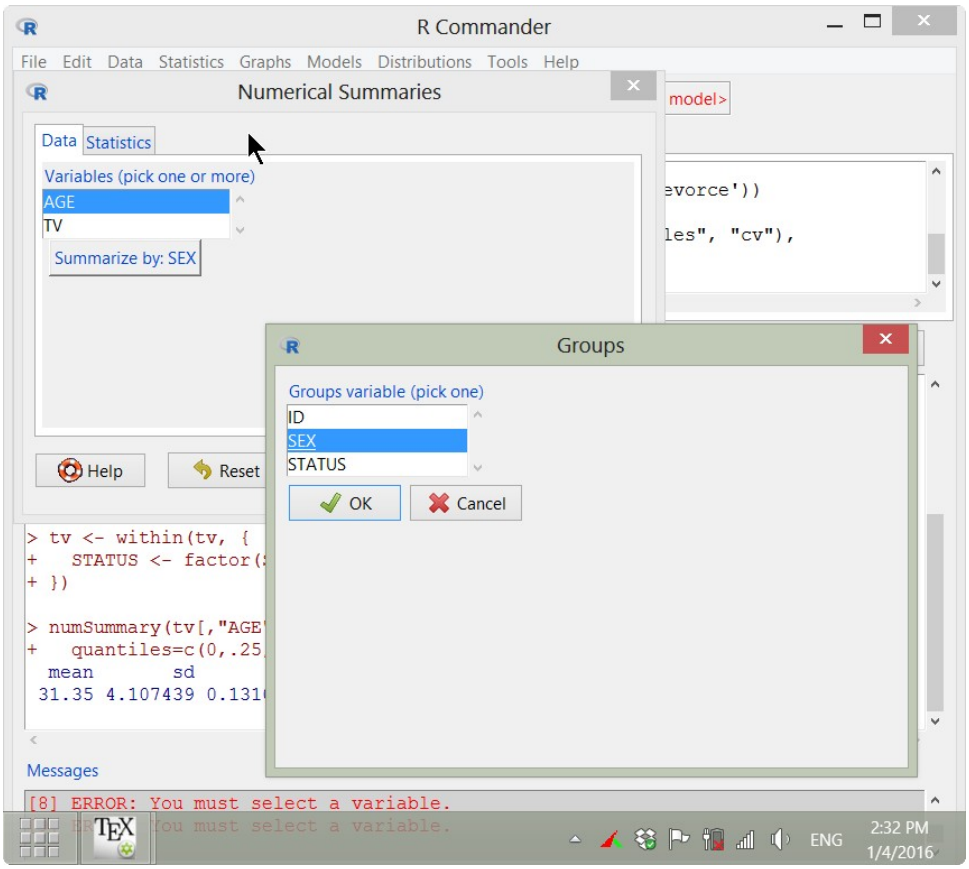
mean	sd	cv	0%	25%	50%	75%	100%	n
31.35	4.107439	0.1310188	25	28	32	34	39	20

นอกจากการคำนวณค่าสถิติเบื้องต้นตามตัวแปรแต่ละตัวแปรแล้ว เรายังสามารถใช้ คำนวณค่าสถิติเหล่านี้จำแนกตามกลุ่มย่อยของตัวแปรเชิงกลุ่มที่สนใจได้อีกด้วย เช่น เราต้องการหาอายุเฉลี่ยของผู้ตอบแบบสอบถามจำแนกตามเพศ สามารถทำได้ดังนี้

1. เลือกเมนู **Statistics -> Summaries -> Numerical Summaries...**

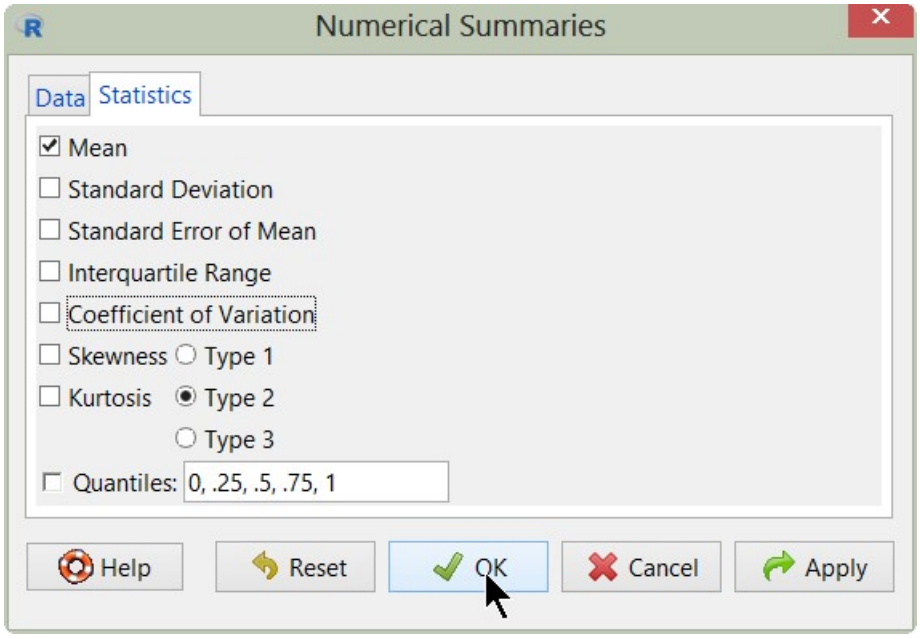
2. เลือกตัวแปรที่ต้องการ ในที่นี้คือ AGE

3. คลิก **Summarize by groups...** จะปรากฏหน้าต่าง



4. เลือกตัวแปรเชิงกลุ่มที่ต้องการ ในที่นี้คือ SEX จากนั้นคลิก **OK**

5. คลิกที่ปุ่ม Statistics เลือกค่าสถิติที่ต้องการ ในที่นี้เลือก Mean



6. คลิก **OK** ได้ผลลัพธ์ดังนี้

	mean	n
Male	29.6	10
Female	33.1	10

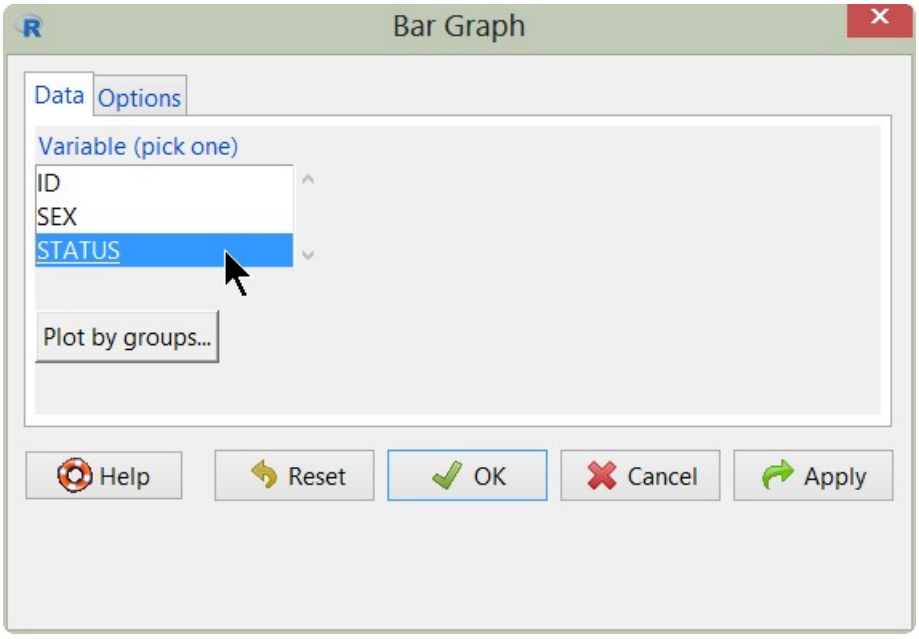
2.5 การสร้างกราฟเพื่อการนำเสนอข้อมูล

การนำเสนอข้อมูลด้วยกราฟเป็นวิธีการของสถิติเชิงพรรณณา ซึ่งกราฟที่ใช้มีอยู่ด้วยกันหลายรูปแบบ ได้แก่ กราฟแท่ง กราฟวงกลม และฮิสโตแกรม เป็นต้น

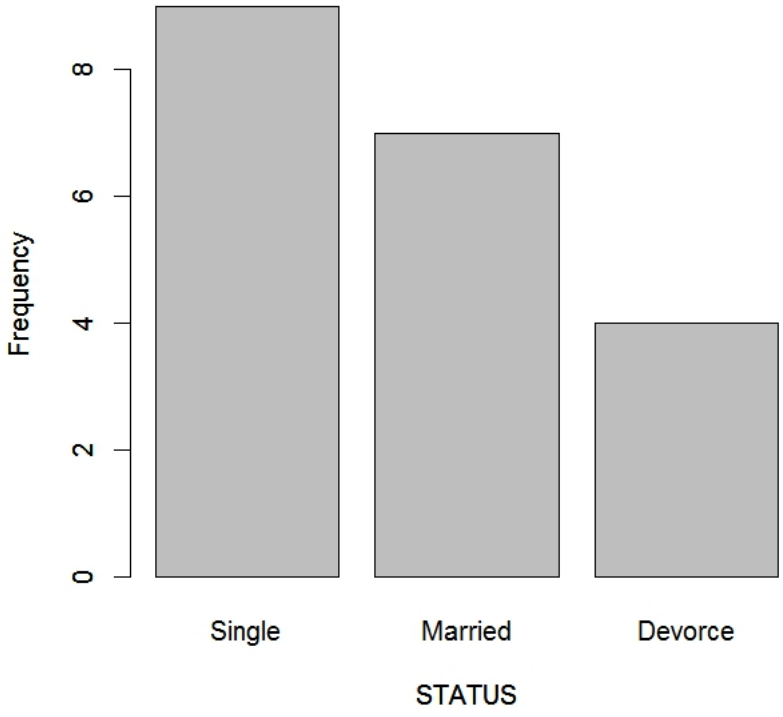
2.5.1 การสร้างกราฟสำหรับข้อมูลเชิงกลุ่ม

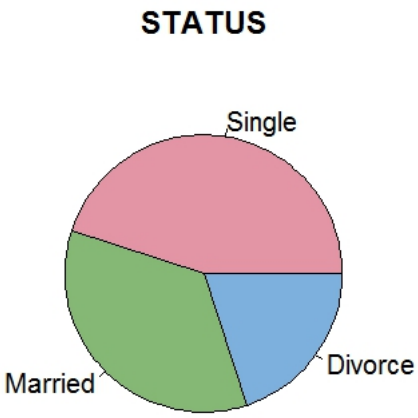
สำหรับข้อมูลเชิงกลุ่ม กราฟที่นิยมใช้ในการนำเสนอข้อมูล ได้แก่ กราฟแท่ง และกราฟวงกลม การใช้ **R** Commander สร้างกราฟเหล่านี้ ทำได้โดย

1. เลือกเมนู **Graphs -> Bar graph...** จะปรากฏหน้าต่าง



2. เลือกตัวแปรที่ต้องการสร้างกราฟ (ในที่นี้เลือกตัวแปร STATUS) จากนั้นคลิก **OK** จะได้กราฟดังรูป





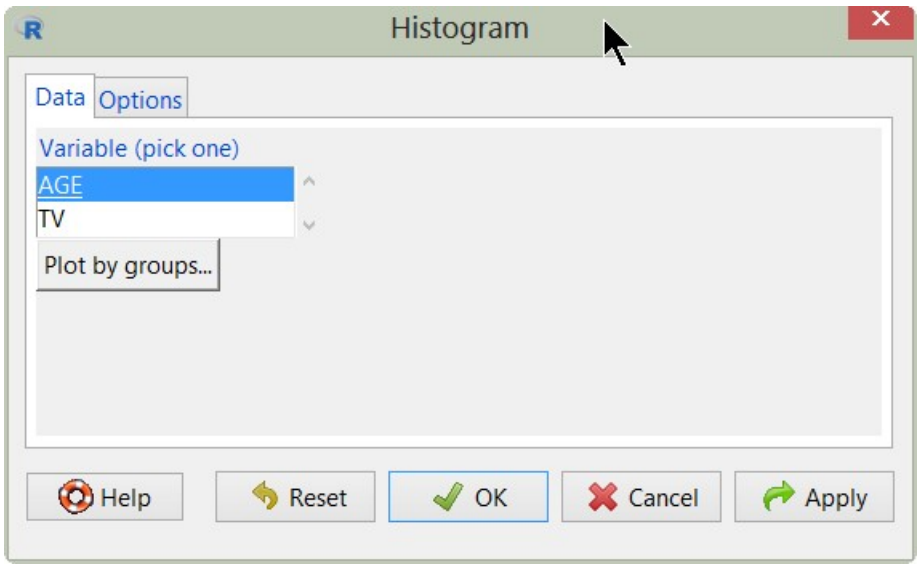
Note: การสร้างกราฟวงกลมทำได้ด้วยวิธีเดียวกัน เพียงแต่เปลี่ยนจากคำสั่ง **Bar graph** เป็น **Pie chart** เท่านั้น

2.5.2 กราฟสำหรับข้อมูลเชิงปริมาณ

สำหรับข้อมูลเชิงปริมาณ กราฟที่นิยมใช้ในการนำเสนอข้อมูลได้แก่ ฮิสโตแกรม stem and leaf display และ boxplot การใช้ **R** Commander ในการสร้างกราฟเหล่านี้ทำได้ ดังนี้

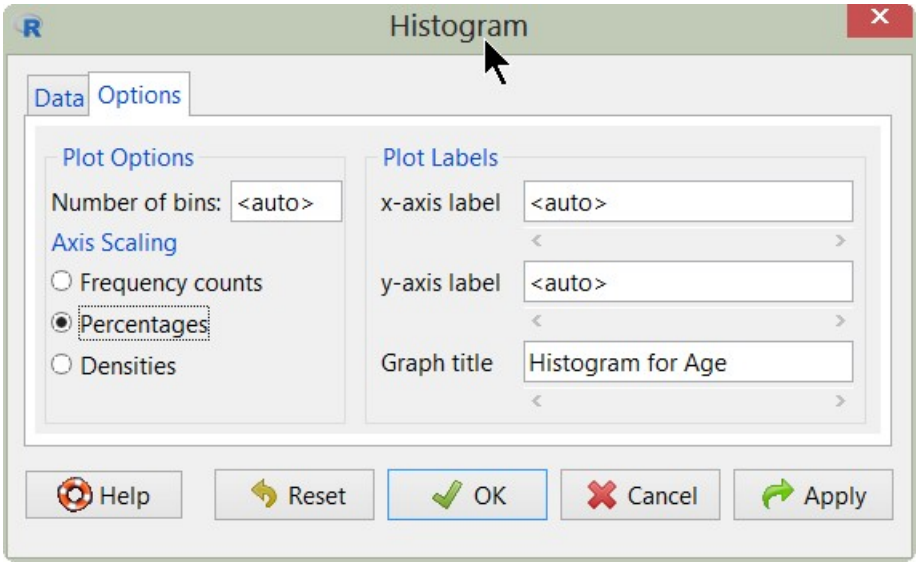
การสร้างฮิสโตแกรม

1. เลือกเมนู **Graphs -> Histogram** จะปรากฏหน้าต่าง **Histogram** ดังนี้



2. เลือกตัวแปรที่ต้องการสร้างกราฟ (ต้องเป็นตัวแปรเชิงปริมาณ) ในที่นี้เลือก AGE

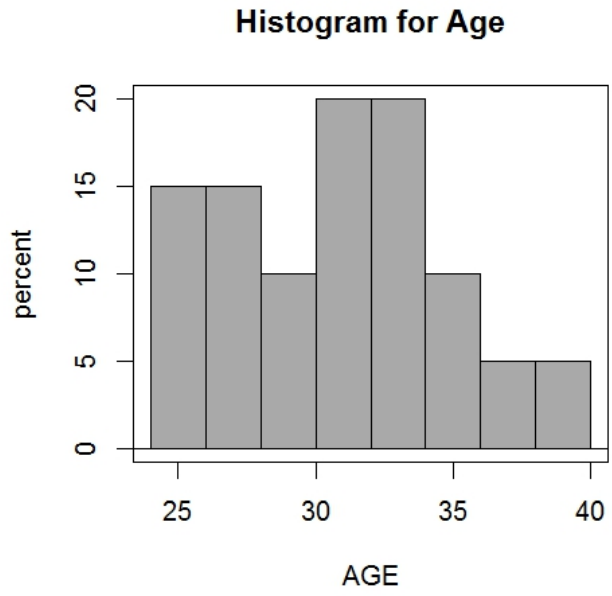
3. เลือกลักษณะการแสดงกราฟ ว่าจะให้แสดงในรูปความถี่ (Frequency counts) ร้อยละ (Percentage) หรือความน่าจะเป็น (Density) ให้คลิกที่ Options ได้หน้าต่างดังนี้



ในที่นี้เลือก **Percentages**

นอกจากนั้นยังสามารถใส่ชื่อกราฟได้ที่ Graph title ในรูปนี้ใส่คำว่า "Histigram for Age"

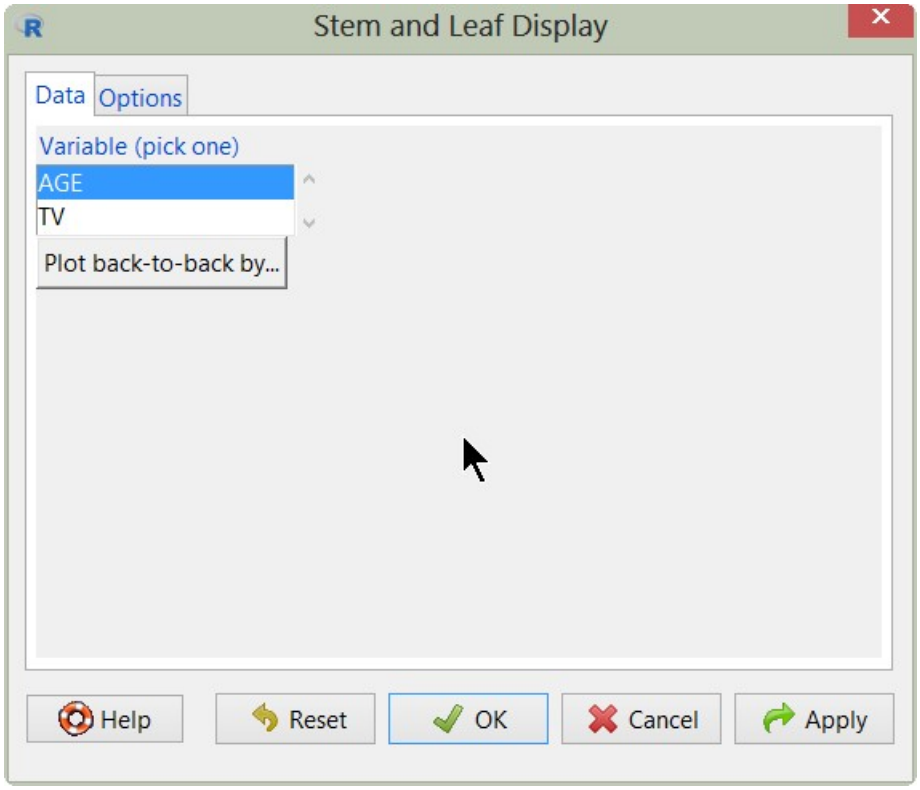
4. คลิก **OK** จะได้กราฟดังรูปข้างล่าง



การสร้างกราฟ Stem and Leaf

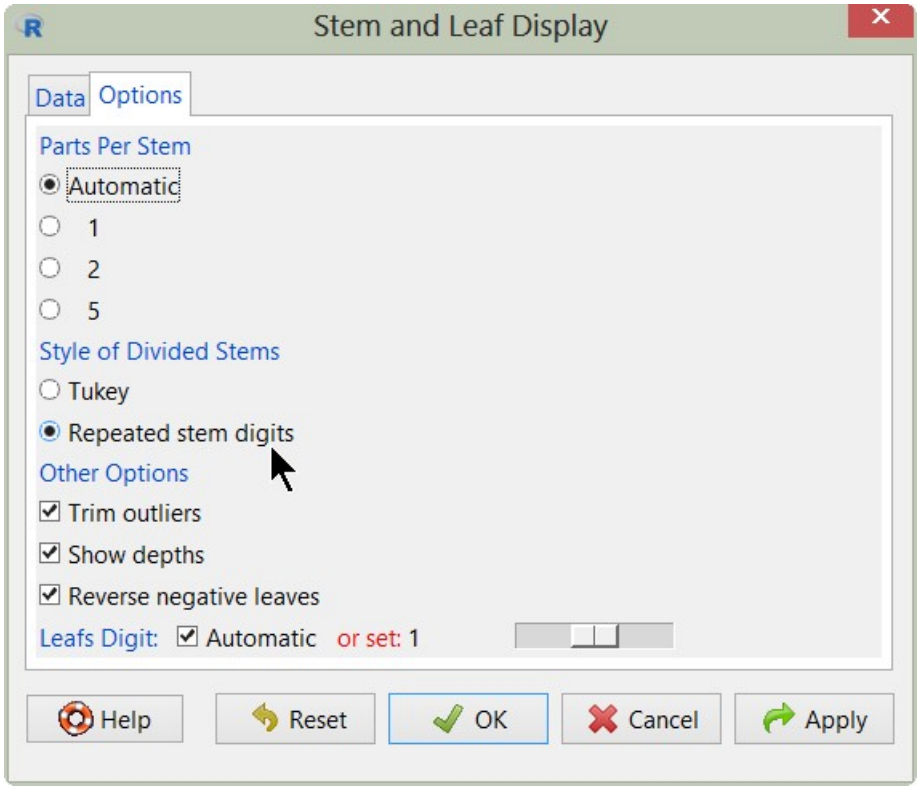
กราฟ Stem and Leaf เป็นกราฟที่มีลักษณะคล้ายกราฟแท่งในแนวนอน แต่ใช้ค่าจริงของข้อมูลมาแสดงในกราฟแทน การใช้ **R** Commander สร้างกราฟชนิดนี้ทำได้ดังนี้

1. เลือกเมนู **Graphs -> Stem and leaf display...** จะปรากฏหน้าต่าง **Stem and Leaf Display** ดังนี้



2. คลิกเลือกตัวแปรที่ต้องการสร้างกราฟ (ต้องเป็นตัวแปรเชิงปริมาณเท่านั้น) ในที่นี้เลือก AGE

3. คลิกที่ปุ่ม Options เพื่อเลือกตัวเลือกต่าง ๆ ตามความต้องการ



Leaf Digit ใช้กำหนดหน่วยของ leaf ที่สามารถปรับได้ครั้งละ 10 หน่วย

Parts Per Stem ใช้กำหนดจำนวน stem ดังนี้

- ถ้ากำหนดให้เป็น 1 จะมีจำนวน stem เป็น 0 1 2 3 4
- ถ้ากำหนดเป็น 2 จะมีจำนวน stem เป็น 00 11 22 33 44

Style of Divided Stem ใช้กำหนดวิธีการแบ่ง stem ในที่นี้ใช้ Repeated stem digits

Other Options ใช้เปลี่ยนวิธีการแสดงค่าต่าง ๆ ในผลลัพธ์

Trim outliers ใช้จัดการเกี่ยวกับค่าผิดปกติของตัวแปร โดยให้แยกค่าผิดปกติแสดงออกมาต่างหากไม่ต้องไปรวมไว้ใน stem and leaf

Show depth ให้แสดงจำนวนข้อมูลที่แสดงในแถวนั้น ๆ

Reverse negative leaves กรณีที่มีข้อมูลมีค่าเป็นลบ ให้เรียงลำดับข้อมูลตามหลักคณิตศาสตร์

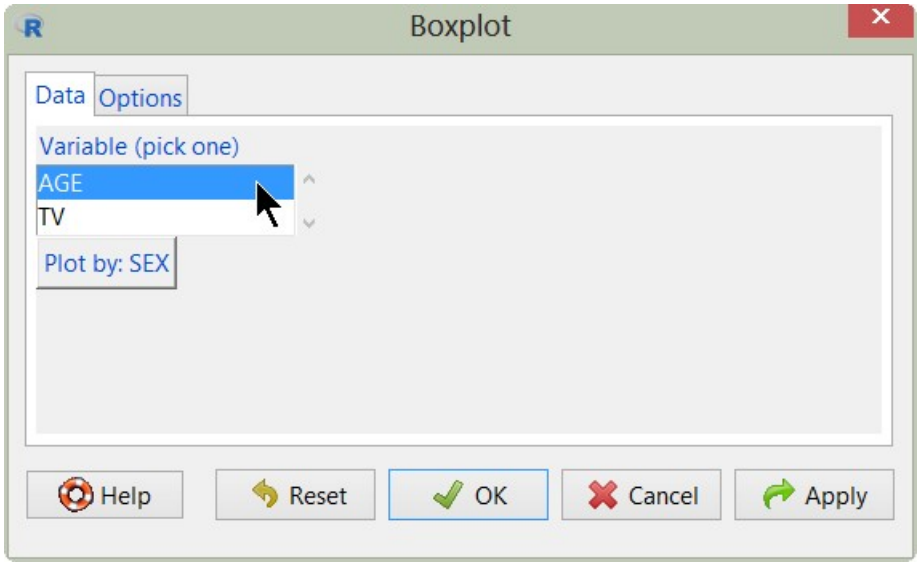
4. คลิก **OK** จะได้ผลลัพธ์ดังนี้

```
1 | 2: represents 12
 leaf unit: 1
           n: 20
  2    2 | 55
  4    2 | 67
  8    2 | 8899
       3 |
 (6)   3 | 222233
  6    3 | 445
  3    3 | 6
  2    3 | 89
```

จากกราฟที่ได้ จะสรุปได้ว่าอายุของผู้ตอบแบบสอบถามที่น้อยที่สุดคือ 25 ปี อายุมากที่สุดคือ 39 ปี และมีอายุ 32 ถึง 33 ปีมากที่สุด

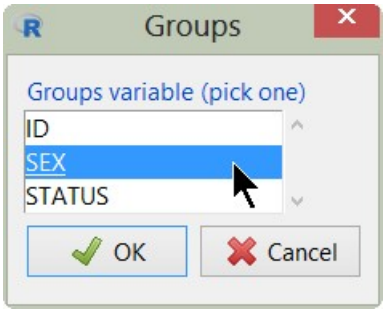
การสร้าง Boxplot

1. เลือกเมนู **Graphs -> Boxplot...** จะปรากฏหน้าต่าง ดังนี้



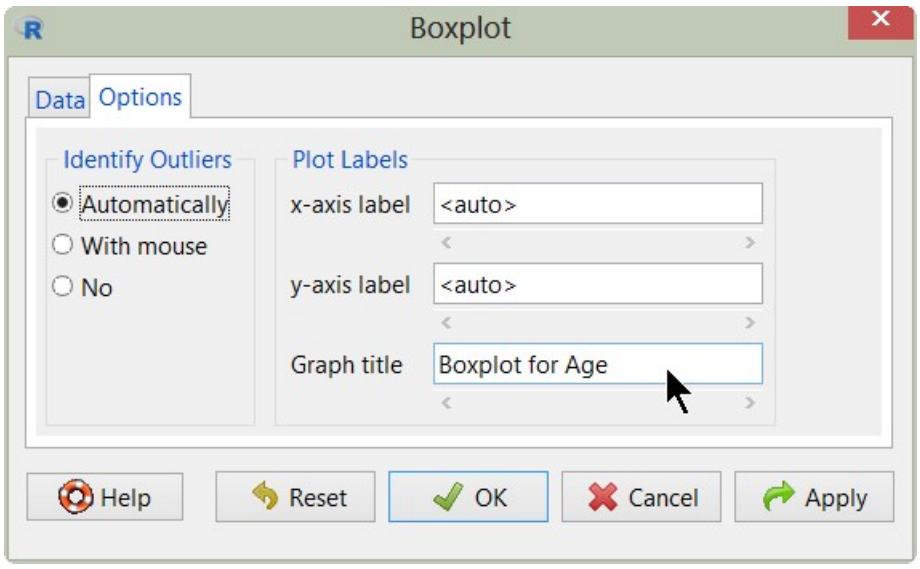
2. เลือกตัวแปรที่ต้องการสร้างกราฟ ในที่นี้เลือก AGE

3. หากต้องการสร้าง boxplot แยกตามกลุ่มของตัวแปรเชิงกลุ่มที่สนใจ เช่น เพศ ให้คลิกเลือก **Plot by groups...** จะปรากฏหน้าต่าง **Groups** ดังนี้

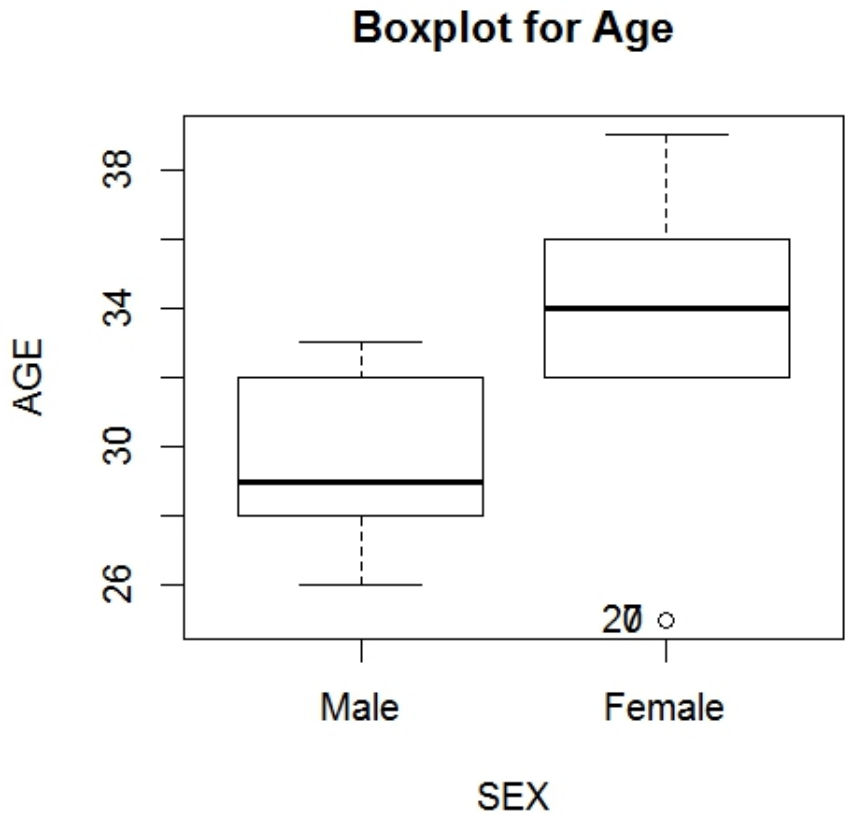


4. ในหน้าต่างจะปรากฏเฉพาะตัวแปรที่กำหนดไว้เป็น Factor เท่านั้น ให้คลิกเลือกตัวแปรที่ต้องการ ในที่นี้เลือก SEX จากนั้นคลิก **OK**

5. ที่ปุ่ม Options สามารถเลือกกำหนดการแสดงค่า Outlier ได้



6. คลิก OK



บทที่ 3

การทดสอบสมมุติฐานทางสถิติ

3.1 หลักการของการทดสอบสมมุติฐาน

การทดสอบสมมุติฐานทางสถิติเป็นวิธีการที่ผู้วิจัยสามารถนำมาใช้ในการตรวจสอบข้อสมมุติหรือข้อสงสัยเกี่ยวกับการวิจัยได้ ซึ่งข้อสมมุติหรือข้อสงสัยนี้เรียกว่า **สมมุติฐานทางการวิจัย (Research hypothesis)** ซึ่งการทดสอบสมมุติฐานนี้จะใช้ความรู้จากข้อมูลตัวอย่างที่เก็บรวบรวมมาได้ โดยเปลี่ยนสมมุติฐานทางการวิจัยให้เป็นสมมุติฐานทางสถิติ (Statistical hypothesis) ซึ่งขั้นตอนของการทดสอบสมมุติฐานสรุปได้ดังนี้

1. ตั้งสมมุติฐานทางสถิติ
2. กำหนดระดับนัยสำคัญของการทดสอบ
3. เลือกวิธีการทางสถิติที่เหมาะสม
4. หาขอบเขตของการตัดสินใจ
5. คำนวณค่าสถิติจากข้อมูลตัวอย่าง
6. ตัดสินใจว่าจะปฏิเสธหรือยอมรับสมมุติฐาน
7. สรุปผลการทดสอบที่ได้

3.1.1 การตั้งสมมุติฐานทางสถิติ

เป็นขั้นตอนแรกของการทดสอบสมมุติฐาน โดยสมมุติฐานทางสถิติมีอยู่ด้วยกัน 2 ชนิด ได้แก่

1. **สมมุติฐานหลัก (Null hypothesis)** ใช้สัญลักษณ์ H_0 โดยต้องกำหนดไว้ในความหมายว่า *เท่ากับ ไม่แตกต่าง ไม่น้อยกว่า ไม่มากกว่า ไม่มีความสัมพันธ์กัน* เป็นต้น ซึ่งอาจกำหนดในรูปสัญลักษณ์ทางสถิติ เช่น μ (ค่าเฉลี่ยของประชากร) p สัดส่วนของสิ่งที่สนใจในประชากร หรือกำหนดในรูปข้อความก็ได้ เช่น

H_0 : รายได้เฉลี่ยต่อเดือนของคนไทยเท่ากับ 5000 บาท หรือ $H_0 : \mu = 5000$

2. สมมุติฐานทางเลือก (Alternative hypothesis) ใช้สัญลักษณ์ H_a เป็นสมมุติฐานที่ต้องตั้งคู่และมีความหมายตรงข้างกับสมมุติฐานหลักเสมอ เช่น

H_a : รายได้เฉลี่ยต่อเดือนของคนไทยไม่เท่ากับ 5000 บาท หรือ $H_0 : \mu \neq 5000$

การตั้งสมมุติฐานทางสถิติจะต้องใช้สมมุติฐานทางการวิจัยเป็นแนวทาง ดังตัวอย่างต่อไปนี้

สมมุติฐานทางการวิจัย	สมมุติฐานทางสถิติ
1. คนไทยมีอายุขัยเฉลี่ย 73 ปี	$H_0 : \mu = 73$ $H_a : \mu \neq 73$
2. คนไทยอายุ 15 ปีขึ้นไป มากกว่าร้อยละ 30 มีภาวะอ้วน	$H_0 : p \leq 0.30$ $H_a : p > 0.30$
3. คนไทยเพศหญิงมีอายุยืนกว่าเพศชาย	$H_0 : \mu_{\text{หญิง}} \leq \mu_{\text{ชาย}}$ $H_a : \mu_{\text{หญิง}} > \mu_{\text{ชาย}}$
4. ผู้หญิงป่วยเป็นโรคมะเร็งปอดน้อยกว่าผู้ชาย	$H_0 : p_{\text{หญิง}} \geq p_{\text{ชาย}}$ $H_a : p_{\text{หญิง}} < p_{\text{ชาย}}$

ประเภทของสมมุติฐาน

การทดสอบสมมุติฐานในทางสถิติจะมีอยู่ด้วยกันสองประเภท ได้แก่

1. **การทดสอบแบบสองทาง (Two-tail test)** เป็นการทดสอบที่ให้ผลสรุปแบบกว้าง ๆ ในความหมายของ *การเท่ากับ* หรือ *ไม่เท่ากับ มีความสัมพันธ์* หรือ *ไม่มีความสัมพันธ์* เช่น ตัวอย่างที่ 1 ในตารางด้านบน

2. **การทดสอบแบบทางเดียว (One-tail test)** เป็นการทดสอบที่สามารถสรุปผลได้ด้านใดด้านหนึ่ง เช่น มากกว่า น้อยกว่า เช่น ตัวอย่างที่ 2 - 4 ในตารางด้านบน

3.1.2 การกำหนดระดับนัยสำคัญ

ระดับนัยสำคัญที่ใช้ในการทดสอบสมมุติฐาน ถือเป็นเกณฑ์ในการตัดสินใจว่าผู้วิจัยจะยอมรับ (Accept) หรือปฏิเสธ (Reject) สมมุติฐานหลักในทางสถิติ โดยระดับนัยสำคัญ (Level of significant) ใช้สัญลักษณ์ α หมายถึง *ความน่าจะเป็นหรือโอกาสที่ผู้วิจัยจะปฏิเสธ H_0 เมื่อ H_0 เป็นจริงหรือถูกต้อง* ซึ่งถือว่าเป็นโอกาสที่ผู้วิจัยจะตัดสินใจผิด การตัดสินใจผิดแบบนี้ เรียกว่า **ความคลาดเคลื่อนประเภทที่ 1 (Type I error)**

ในวิจัยโดยทั่วไป ผู้วิจัยจะเป็นผู้กำหนดระดับนัยสำคัญไว้ล่วงหน้าก่อนการทดสอบสมมุติฐานเสมอ โดยระดับนัยสำคัญที่นิยมใช้ คือ 0.01 0.05 และ 0.10 การเลือกระดับนัยสำคัญเท่าใดนั้นขึ้นอยู่กับความเหมาะสมและความเสียที่จะเกิดขึ้นหากมีความคลาดเคลื่อนของการตัดสินใจเกิดขึ้น หากผลเสียที่เกิดขึ้นไม่รุนแรงมากนัก เช่น ไม่ได้ทำให้เกิดการบาดเจ็บล้มตาย ก็อาจระดับนัยสำคัญ 0.05 หรือ 0.01 แต่หากผลเสียที่เกิดขึ้นอาจก่อ

ให้เกิดความเสียหายอย่างมาก ก็อาจต้องใช้ระดับนัยสำคัญต่ำ ๆ เช่น ในงานวิจัยทางการแพทย์อาจต้องใช้ระดับนัยสำคัญ 0.01 หรือ 0.001

3.1.3 การเลือกวิธีการทางสถิติหรือตัวสถิติทดสอบ

การเลือกวิธีการทางสถิติหรือตัวสถิติทดสอบเป็นขั้นตอนที่ต้องอาศัยความรู้ในทางสถิติของผู้วิจัยพอสมควร เพื่อให้วิธีการทางสถิติที่เลือกเป็นวิธีที่ถูกต้องเหมาะสม ซึ่งจะส่งผลให้ผลการทดสอบและข้อสรุปที่ได้ถูกต้องและเชื่อถือได้ ในการทดสอบเรื่องหนึ่ง มักมีวิธีการทางสถิติให้เลือกใช้ได้หลายชนิด แต่วิธีใดจะเป็นวิธีที่เหมาะสม ผู้วิจัยต้องพิจารณาถึงข้อสมมุติเบื้องต้น (Assumption) หรือข้อกำหนดของวิธีการนั้น ๆ ว่าสอดคล้องกับชนิดของข้อมูลและลักษณะของข้อมูลที่มีอยู่หรือไม่ รวมทั้งต้องคำนึงด้วยว่าตัวสถิติทดสอบนั้นสามารถใช้ในการตอบวัตถุประสงค์ของการวิจัยได้หรือไม่ และผู้วิจัยควรคำนึงไว้เสมอว่า ผู้วิจัยควรตัวสถิติทดสอบที่ง่ายที่สุดที่สามารถตอบวัตถุประสงค์ของการวิจัยได้ คำว่า "ง่าย" ในที่นี้หมายความว่า ง่ายต่อการใช้ ง่ายต่อการวิเคราะห์การคำนวณ และง่ายต่อการแปลความหมายของค่าสถิติที่ได้ เพราะหากผู้วิจัยเลือกวิธีการทางสถิติที่ซับซ้อนมากเกินไป โดยที่ไม่มีความรู้ทางสถิติในเรื่องนั้น ๆ อย่างเพียงพออาจทำให้เกิดผลเสียได้ นั่นคือ อาจมีการคำนวณและข้อมูลผิดพลาด และการแปลความหมายของค่าสถิติต่าง ๆ จะมีความซับซ้อนและยากมากขึ้น ซึ่งอาจทำให้ได้ข้อสรุปที่ผิดไปจากความเป็นจริงได้ อย่างไรก็ตาม หากในงานวิจัยนั้นจำเป็นต้องใช้วิธีการทางสถิติขั้นสูงซึ่งมีความซับซ้อน ผู้วิจัยควรจะศึกษาถึงหลักการและเงื่อนไขของวิธีเหล่านั้นให้เข้าใจเสียก่อนที่จะนำไปใช้ หรืออาจนำงานวิจัยนั้นไปปรึกษานักสถิติเพื่อขอคำแนะนำในการใช้วิธีการทางสถิติที่เหมาะสมได้

3.1.4 ขอบเขตของการตัดสินใจ

ขอบเขตของการตัดสินใจคือค่าทางสถิติที่ใช้สำหรับการตัดสินใจว่าจะยอมรับหรือปฏิเสธสมมุติฐานหลัก ซึ่งโดยทั่วไปมีอยู่ด้วยกัน 2 วิธี ได้แก่

- ใช้ตารางสถิติ
- ไม่ใช้ตารางสถิติ

ในที่นี้เราจะพิจารณาถึงเฉพาะวิธีที่ 2 เท่านั้น เนื่องจากวิธีแรกนั้นนักวิจัยต้องมีตารางสถิติสำหรับวิธีการทดสอบทางสถิติที่ใช้อยู่ด้วยทุกครั้ง ซึ่งไม่ค่อยสะดวกในการทำงานมากนัก อีกประการหนึ่งตารางสถิติสำหรับวิธีการทางสถิติวิธีหนึ่งอาจมีอยู่หลายตาราง และมีวิธีการอ่านค่าที่แตกต่างกัน ทำให้ผู้ใช้ที่ไม่มีความรู้ทางสถิติดีพออาจเกิดความสับสน และอ่านค่าผิดได้ ส่วนวิธีที่สองนั้นไม่จำเป็นต้องใช้ตารางสถิติแต่จะใช้ค่าสถิติค่าหนึ่งแทนค่าสถิตินี้เรียกว่าค่า *p*-value หมายถึงความน่าจะเป็นหรือโอกาสค่าสถิติทดสอบจะมีค่ามากกว่าค่าที่คำนวณได้จากข้อมูลตัวอย่าง หรือคือความน่าจะเป็นที่จะปฏิเสธสมมุติฐานหลักที่คำนวณได้จากข้อมูลตัวอย่าง เมื่อสมมุติฐานหลักเป็นจริง เนื่องจาก *p*-value เป็นความน่าจะเป็นจึงมีค่า 0 - 1 เท่านั้น จะสังเกตเห็นได้ว่าค่า *p*-value และระดับนัยสำคัญ α มีความหมายเหมือนกัน นั่นคือ ทั้งสองค่าต่างก็เป็นความน่าจะเป็นในการปฏิเสธสมมุติฐานหลักเมื่อสมมุติฐานหลักเป็นจริงทั้งคู่ แต่ที่มาของค่าสองค่านี้แตกต่างกัน โดยที่ระดับนัยสำคัญ α นั้นผู้วิจัยเป็นผู้กำหนดเอง เรียกว่า nominal value แต่ค่า *p*-value นั้นได้มาจากการคำนวณโดยใช้ข้อมูลตัวอย่างที่เก็บรวบรวมมา ดังนั้น ในการทดสอบสมมุติฐาน เราจึงสามารถใช้ค่า *p*-value เปรียบเทียบกับระดับนัยสำคัญ α ในการตัดสินใจได้ ซึ่งวิธีการใช้ค่า *p*-value ในการตัดสินใจที่จะปฏิเสธหรือยอมรับสมมุติฐานหลักทำได้ดังนี้

- ยอมรับสมมุติฐานหลัก H_0 เมื่อ *p*-value มากกว่าหรือเท่ากับระดับนัยสำคัญ α

- ปฏิเสธสมมุติฐานหลัก H_0 เมื่อ p-value น้อยกว่าระดับนัยสำคัญ α

3.1.5 การหาค่าสถิติทดสอบจากข้อมูลตัวอย่าง

ค่าสถิติทดสอบนี้คำนวณหาได้จากข้อมูลตัวอย่างที่เก็บรวบรวมมาได้ โดยมีสูตรหรือวิธีการคำนวณที่แตกต่างกันไปตามตัวสถิติทดสอบที่เลือกใช้ เมื่อได้ค่าสถิติทดสอบแล้วจึงจะนำค่าที่ได้ไปเปรียบเทียบกับขอบเขตของการตัดสินใจดังที่กล่าวมาแล้ว

3.1.6 การตัดสินใจที่จะปฏิเสธหรือยอมรับสมมุติฐาน

ในการตัดสินใจที่จะปฏิเสธหรือยอมรับสมมุติฐานหลักนั้น ผู้วิจัยต้องนำค่าสถิติทดสอบที่ได้ไปเปรียบเทียบกับเกณฑ์หรือขอบเขตของการตัดสินใจที่หาไว้ก่อนหน้านี้ แล้วจึงทำการตัดสินใจว่าจะปฏิเสธหรือยอมรับสมมุติฐานหลัก (การตัดสินในจะกระทำกับสมมุติฐานหลักเท่านั้น เนื่องจากสมมุติฐานหลักและสมมุติฐานทางเลือกมีความหมายตรงข้ามกัน ดังนั้น เมื่อปฏิเสธ H_0 ก็เท่ากับยอมรับ H_a นั่นเอง)

3.1.7 การสรุปผลการทดสอบ

การสรุปผลการทดสอบเป็นขั้นตอนสุดท้ายของการทดสอบสมมุติฐาน ผู้วิจัยต้องสรุปผลการทดสอบที่ได้เพื่อนำไปสู่ข้อสรุปของการวิจัย ซึ่งในการสรุปผลผู้วิจัยควรระบุเกณฑ์ที่ใช้ในการทดสอบ หรือระดับนัยสำคัญ α ด้วย ในการสรุปผลไม่จำเป็นต้องใช้สัญลักษณ์ทางสถิติ แต่ควรใช้ภาษาหรือคำอธิบายที่เข้าใจง่าย คนทั่วไปสามารถเข้าใจได้ โดยยึดสมมุติฐานทางการวิจัยเป็นหลักในการสรุปผล เช่น สมมุติฐานทางการวิจัยกำหนดไว้ว่า "คนไทยมีรายได้เฉลี่ยต่อเดือนน้อยกว่า 6000 บาท" ซึ่งมีสมมุติฐานทางสถิติคือ

$$H_0 \ : \ \mu \geq 6000$$
$$H_a \ : \ \mu < 6000$$

จากข้อมูลตัวอย่างที่เก็บรวบรวมมาได้ เมื่อทำการทดสอบสมมุติฐานพบว่าเกิดการปฏิเสธสมมุติฐานหลัก จึงสรุปได้ว่า "คนไทยมีรายได้เฉลี่ยน้อยกว่า 6000 บาทต่อเดือน ที่ระดับนัยสำคัญ 0.05 (หรือที่เกณฑ์ 5%)"

3.2 การทดสอบค่าเฉลี่ยของประชากรหนึ่งกลุ่ม (μ)

ในการทดสอบสมมุติฐานเกี่ยวกับค่าเฉลี่ยของประชากรหนึ่งกลุ่ม มีวิธีการทดสอบอยู่ 2 วิธี คือ t test และ z test โดยมีเงื่อนไขในการใช้ต่างกัน ดังนี้

- **t test** ใช้เมื่อตัวอย่างมีขนาดเล็ก ($n < 30$) และไม่ทราบค่าเบี่ยงเบนมาตรฐานของประชากร σ โดยประชากรของข้อมูลต้องมีการแจกแจงปรกติหรือใกล้เคียง ดังนั้น ก่อนที่จะใช้ t test ในการทดสอบสมมุติฐาน จะต้องมีการตรวจสอบก่อนเสมอว่า ข้อมูลตัวอย่างนั้นมาจากประชากรที่มีการแจกแจงปรกติหรือไม่ ซึ่งวิธีการตรวจสอบมีอยู่ด้วยกันหลายวิธี วิธีที่นิยมใช้มากได้แก่ การพิจารณาจากกราฟ เช่น ฮิสโตแกรม stem and leaf display หรือ boxplot

- **z test** ใช้เมื่อตัวอย่างมีขนาดใหญ่ ($n \geq 30$) หรือทราบค่าเบี่ยงเบนมาตรฐานของประชากร (σ)

Note: สำหรับ *z* test กรณีที่เราจะทราบค่าเบี่ยงเบนมาตรฐานของประชากรเป็นไปได้ยาก เนื่องจาก σ เป็นค่าพารามิเตอร์ที่ไม่ทราบค่า ส่วนในกรณีที่ตัวอย่างมีขนาดใหญ่ หากตัวอย่างมีขนาดใหญ่มาก ๆ เราอาจใช้ *t* test แทน *z* test ได้ เนื่องจากคุณสมบัติของการแจกแจงแบบ *t* นั่นคือ เมื่อ *n* มีขนาดใหญ่ขึ้น การแจกแจงแบบ *t* จะมีคุณสมบัติใกล้เคียงการแจกแจงปรกติมาตรฐาน ทำให้ในโปรแกรมสำเร็จรูปบางโปรแกรม เช่น SPSS จะไม่มี *z* test มีแต่ *t* test เท่านั้น

การใช้ R ในการทดสอบค่าเฉลี่ยของประชากร 1 กลุ่ม

จากข้อมูลเกี่ยวกับการสำรวจการดูรายการโทรทัศน์ในบทที่แล้ว หากเราต้องการทดสอบว่าอายุเฉลี่ยประชากรมากกว่า 30 ปี หรือไม่ ตั้งสมมุติฐานทางสถิติ คือ

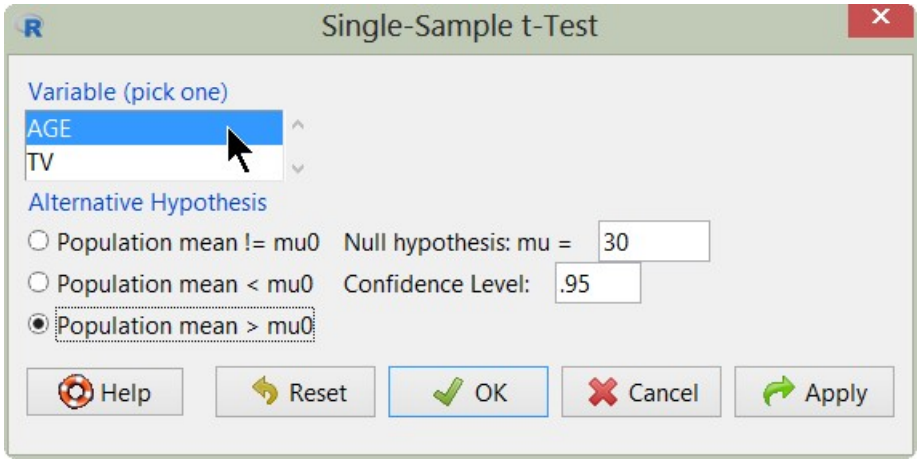
$$H_0 \ : \ \mu \leq 30$$
$$H_a \ : \ \mu > 30$$

ในการทดสอบสมมุติฐานสำหรับข้อมูลชุดนี้ เนื่องจากข้อมูลมีจำนวนน้อยกว่า 30 เมื่อจะใช้ *t* test ในการทดสอบต้องทำการตรวจสอบลักษณะการแจกแจงของข้อมูลก่อน เนื่องจากข้อมูลมีจำนวนน้อย กราฟที่เหมาะสมคือ stem and leaf display สำหรับข้อมูลชุดได้กราฟมีลักษณะดังนี้

```
1 | 2: represents 12
 leaf unit: 1
           n: 20
  2    2 | 55
  4    2 | 67
  8    2 | 8899
       3 |
 (6)   3 | 222233
  6    3 | 445
  3    3 | 6
  2    3 | 89
```

จากกราฟจะเห็นได้ว่าข้อมูลมีการแจกแจงคล้ายรูประฆังคว่ำ ซึ่งเป็นลักษณะของการแจกแจงปรกติ จึงสรุปได้ว่าข้อมูลชุดนี้มาจากประชากรที่มีการแจกแจงปรกติ ซึ่งตรงกับข้อสมมุติเบื้องต้นหรือข้อกำหนดของ *t* test ดังนั้น เราจะใช้ *t* test ทดสอบสมมุติฐาน ดังนี้

1. เลือกเมนู **Statistics -> Means -> Single-sample t-test...** จะได้หน้าต่าง



2. เลือกตัวแปรที่ต้องการทดสอบ ในที่นี้เลือก AGE

3. กำหนดค่าพารามิเตอร์ที่ต้องการทดสอบ จากตัวอย่างเราต้องการทดสอบว่าอายุเฉลี่ยมากกว่า 30 หรือไม่ จึงกำหนด **Null hypothesis: mu= 30**

4. เลือกรูปแบบของสมมุติฐานทางเลือกที่ต้องการทดสอบ โดยคลิกเลือกที่ **Alternative Hypothesis**

   Population mean != mu0 เป็นการทดสอบสมมุติฐานแบบสองทาง (Two tailed Test)

   Population mean < mu0 เป็นการทดสอบสมมุติฐานแบบทางเดียวทางซ้าย (Left tailed test)

   Population mean > mu0 เป็นการทดสอบสมมุติฐานแบบทางเดียวทางขวา (Right tailed test)

   ในที่นี้เราเลือกทดสอบสมมุติฐานแบบทางเดียวทางด้านมากกว่าตามสมมุติฐานที่ตั้งไว้ ดังนั้น คลิกเลือกที่ **Population mean > mu0**

5. คลิก **OK** ได้ผลลัพธ์ดังนี้

```
One Sample t-test

data:  tv$AGE
t = 1.4699, df = 19, p-value = 0.07898
alternative hypothesis: true mean is greater than 30
95 percent confidence interval:
 29.76188      Inf
sample estimates:
mean of x
    31.35
```

ผลการวิเคราะห์ที่ได้มีความหมายดังนี้

1. ค่า **t=1.4699** คือค่าสถิติทดสอบ t ที่คำนวณได้จากข้อมูลตัวอย่าง มีจำนวนองศาเสรีเท่ากับ 19 และมีค่า p-value=0.07898 หากเรากำหนดระดับนัยสำคัญ $\alpha = 0.05$ จะเห็นได้ว่า ค่า p-value นั้นมีค่ามากกว่าระดับนัยสำคัญที่กำหนด จึงยอมรับสมมุติฐานหลัก H_0 หมายความว่า ที่ระดับนัยสำคัญ 0.05 อายุเฉลี่ยของประชากรเป้าหมายนั้นไม่ได้มากกว่า 30 ปี

2. สำหรับค่า **95 percent confidence interval** (29.76188 Inf) คือช่วงความเชื่อมั่นทางด้านขวาของค่าเฉลี่ยของประชากร หมายความว่าประชากรเป้าหมายมีอายุเฉลี่ยตั้งแต่ 29.7 ปี

3. ส่วนของ **mean of x** หมายถึงค่าเฉลี่ยของข้อมูลตัวอย่าง ในที่นี้เท่ากับ 31.35 ปี หมายความคนที่ถูกสุ่มมาเป็นตัวอย่างมีอายุเฉลี่ยเท่ากับ 31.35 ปี

3.3 การทดสอบค่าเฉลี่ยของประชากร 2 กลุ่ม

การทดสอบความแตกต่างระหว่างค่าเฉลี่ยของประชากร 2 กลุ่ม เป็นการทดสอบเพื่อเปรียบเทียบว่าประชากรสองกลุ่มนั้นมีลักษณะแตกต่างกันอย่างไร เช่น คนไทยผู้ชายและผู้หญิงมีอายุขัยโดยเฉลี่ยแตกต่างกันหรือไม่ หรือผู้ชายไทยมีรายได้เฉลี่ยต่อปีมากกว่าผู้หญิงหรือไม่ โดยข้อมูลที่สามารถนำมาทดสอบค่าเฉลี่ยได้ต้องอยู่ในมาตราแบบช่วงขึ้นไป คือต้องเป็นข้อมูลเชิงปริมาณ วิธีการทางสถิติที่ใช้ ได้แก่ *t* test และ *z* test โดยแต่ละวิธีจะข้อกำหนดที่แตกต่างกัน อย่างไรก็ตาม ในการทดสอบค่าเฉลี่ยของประชากร 2 กลุ่มนี้ ต้องคำนึงถึงลักษณะของข้อมูลด้วย โดยจะจำแนกลักษณะของข้อมูลเป็น 2 ประเภท คือ

1. ตัวอย่างสองกลุ่มที่สุ่มมาเป็นอิสระกัน (Independent samples)

2. ตัวอย่างสองชุดที่สุ่มมาไม่เป็นอิสระกัน (Dependent samples)

การจำแนกลักษณะของกลุ่มตัวอย่าง 2 ประเภทนี้ มีความสำคัญมาก เพราะวิธีการวิเคราะห์ที่ใช้จะมีความแตกต่างกัน ดังนั้น ผู้วิจัยต้องจำแนกให้ได้ว่ากลุ่มตัวอย่างในงานวิจัยของตนเป็นประเภทไหน เช่น ต้องการเปรียบอายุขัยเฉลี่ยของผู้ชายและผู้หญิง ลักษณะเช่นนี้ถือเป็นกรณีที่ตัวอย่างสองกลุ่มเป็นอิสระกัน แต่ถ้าผู้วิจัยต้องการเปรียบเทียบรายได้เฉลี่ยของน้ำหนักเฉลี่ยก่อนและหลังเข้าโปรแกรมลดน้ำหนัก ลักษณะเช่นนี้ต้องใช้กรณีที่กลุ่มตัวอย่างสองกลุ่มเป็นอิสระกัน

3.3.1 การทดสอบความแตกต่างของค่าเฉลี่ย 2 กลุ่ม เมื่อกลุ่มตัวอย่างเป็นอิสระกัน

การทดสอบความแตกต่างของค่าเฉลี่ยของประชากร 2 กลุ่ม เมื่อกลุ่มตัวอย่างเป็นอิสระกัน ด้วย *t* test หรือ *z* test ข้อมูลตัวอย่างทั้งสองกลุ่มต้องเป็นข้อมูลเชิงปริมาณและมีวิธีการเลือกวิธีทดสอบ ดังนี้

z **test** ใช้เมื่อตัวอย่างทั้งสองกลุ่มมีขนาดใหญ่ (n_1 และ n_2 มากกว่า 30) หรือทราบค่าเบี่ยงเบนมาตรฐานของประชากรที่หนึ่งและสอง (ทราบค่า σ_1 และ σ_2)

t **test** ใช้เมื่อตัวอย่างมีขนาดเล็ก (n_1 หรือ n_2 น้อยกว่าหรือเท่ากับ 30) และไม่ทราบค่า σ_1 และ σ_2 เมื่อประชากรทั้งสองกลุ่มมีการแจกแจงปรกติหรือใกล้เคียง

เช่นเดียวกับการทดสอบสมมุติฐานเกี่ยวกับค่าเฉลี่ยของประชากรหนึ่งกลุ่ม ในกรณีของ *z* test การที่เราจะทราบค่าของ σ_1 และ σ_2 เป็นไปได้ยาก และเมื่อ n_1 และ n_2 มีขนาดใหญ่ การแจกแจงแบบ *t* จะใกล้เคียง

การแจกแจงปรกติมาตรฐาน ทำให้เราสามารถใช้ t test แทน z test ได้ ดังนั้น ในที่นี้จะกล่าวถึงเฉพาะ t test เท่านั้น

อย่างไรก็ตาม การทดสอบโดยใช้ t test ยังแบ่งเป็น 2 กรณี ดังนี้

1. $\sigma_1 = \sigma_2$ ในกรณีนี้ ค่าสถิติทดสอบคือ

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left\{ \frac{1}{n_1} + \frac{1}{n_2} \right\}}}$$

ซึ่งมีการแจกแจงแบบ t ด้วยองศาเสรีเท่ากับ $n_1 + n_2 - 2$

2. $\sigma_1 \neq \sigma_2$ ในกรณีนี้ ค่าสถิติทดสอบคือ

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

ซึ่งมีการแจกแจงแบบ t ด้วยองศาเสรีเท่ากับ

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

ดังนั้นในการตัดสินใจว่าจะใช้ t test กรณีใดจึงเหมาะสม เราต้องทำการทดสอบก่อนว่าค่าเบี่ยงเบนมาตรฐานของประชากร 2 กลุ่ม แตกต่างกันหรือไม่ ซึ่งการทดสอบทำได้โดยการใช้ F test

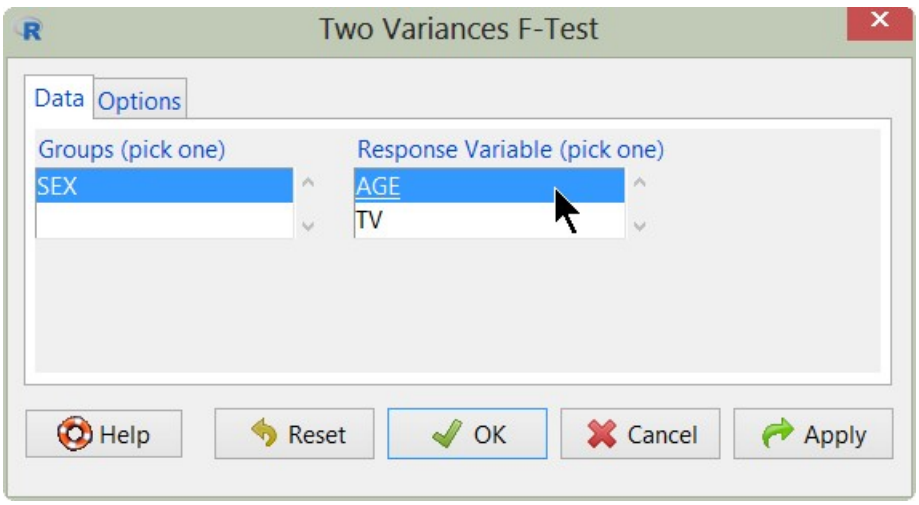
จากตัวอย่างการสำรวจข้อมูลการดูโทรทัศน์ หากเราต้องการทดสอบว่าอายุเฉลี่ยของผู้ชายและผู้หญิงที่เป็นกลุ่มเป้าหมายนั้นแตกต่างกัน มีสมมุติฐานทางสถิติคือ

$$H_0 \quad : \quad \mu_{\text{ชาย}} = \mu_{\text{หญิง}} \tag{3.1}$$

$$H_a \quad : \quad \mu_{\text{ชาย}} \neq \mu_{\text{หญิง}} \tag{3.2}$$

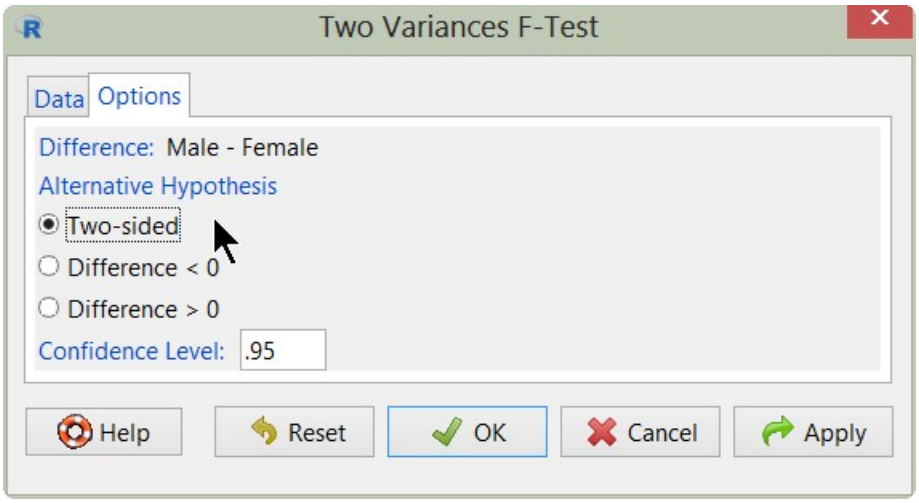
วิธีใช้ **R** Commander ทดสอบสมมุติฐานนี้ ทำได้โดยการทดสอบก่อนว่าค่าเบี่ยงเบนมาตรฐานของประชากร 2 กลุ่ม เท่ากันหรือไม่ ดังนี้

1. เลือกเมนู **Statistics -> Variances -> Two-variance F-test...**



2. ที่ **Groups (pick one)** เลือกตัวแปรแบ่งกลุ่ม ในที่นี้คือ SEX และที่ **Response variable (pick one)** เลือกตัวแปรที่ต้องการทดสอบ ในที่นี้คือ AGE

3. คลิกปุ่ม Options ได้หน้าต่าง



Alternative hypothesis คลิกเลือก **Two-sided**

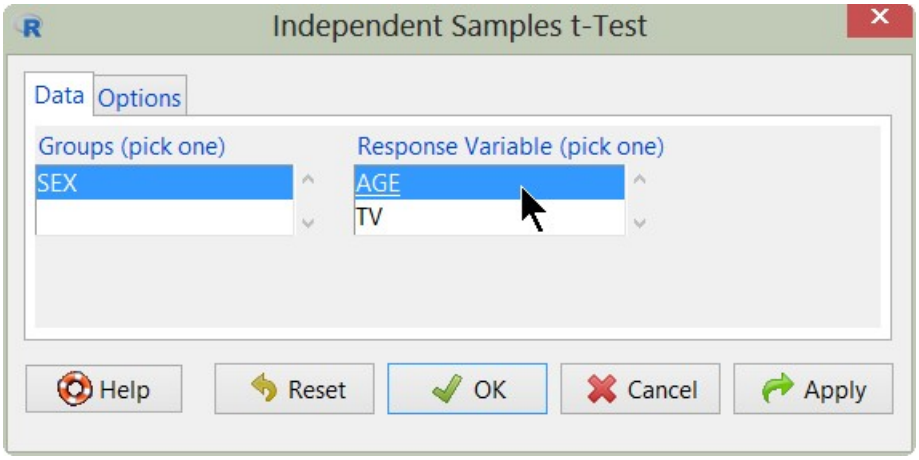
4. คลิก **OK** ได้ผลลัพธ์ดังนี้

```
F test to compare two variances

data:  AGE by SEX
F = 0.2655, num df = 9, denom df = 9, p-value = 0.0612
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.06594529 1.06888278
sample estimates:
ratio of variances
         0.2654954
```

การใช้ผลลัพธ์ที่ได้ทดสอบความแตกต่างของค่าเบี่ยงเบนมาตรฐานของประชากร 2 กลุ่ม ให้ดูที่ค่า **F = 0.2655** และค่า p-value = 0.0612 หากกำหนดระดับนัยสำคัญของการทดสอบเท่ากับ 0.05 จะเห็นได้ว่าค่า p-value มากกว่า 0.05 จึงสรุปได้ว่าค่าเบี่ยงเบนมาตรฐานของประชากร 2 กลุ่มนี้ ไม่แตกต่างกัน ที่ระดับนัยสำคัญ 0.05

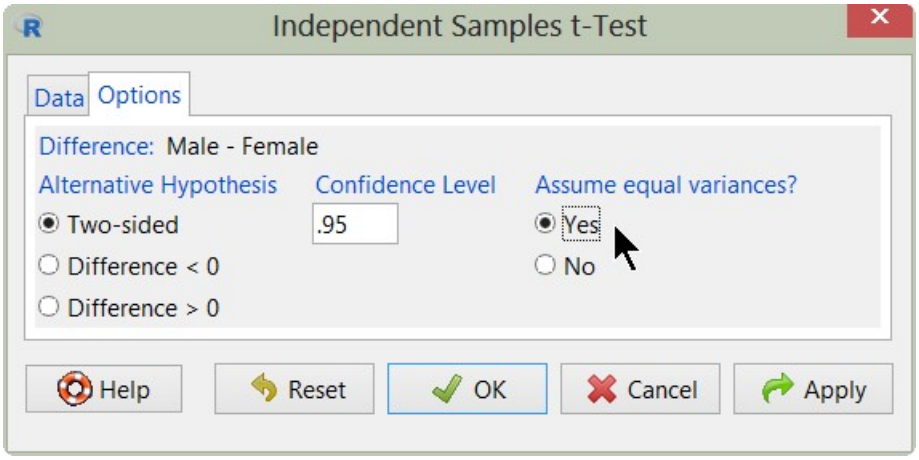
ดังนั้น ในการทดสอบความแตกต่างของค่าเฉลี่ยของประชากร 2 กลุ่ม จึงใช้ t test ในกรณีที่ $\sigma_1 \neq \sigma_2$ ดังนี้

1. เลือกเมนู **Statistics -> Means -> Independent-sample t-test...**



2. ที่ **Groups (pick one)** เลือกตัวแปรแบ่งกลุ่ม คือ SEX และที่ **Response Variable (pick one)** เลือกตัวแปรที่ต้องการทดสอบ คือ AGE

3. คลิกที่ปุ่ม Options ได้หน้าต่างดังนี้



4. ที่ **Alternative Hypothesis** เลือกสมมุติฐานทางเลือกที่ต้องการทดสอบ ในที่นี้เลือก **Two-sided** เนื่องจากสมมุติฐานที่กำหนดไว้เป็นสมมุติฐานสองทาง

5. ที่ **Assume equal variances?** เลือก **Yes** เนื่องจากในการทดสอบความแตกต่างของค่าเบี่ยงเบนมาตรฐาน พบว่าความเบี่ยงเบนมาตรฐานของประชากร 2 กลุ่ม ไม่แตกต่างกัน (แต่ถ้าหากใช้ *F* test ทดสอบแล้วพบว่า $\sigma_1 \neq \sigma_2$ ให้เลือก **No**)

6. คลิก **OK** ได้ผลลัพธ์ดังนี้

```
Two Sample t-test

data:  AGE by SEX
t = -2.062, df = 18, p-value = 0.05395
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.06607063  0.06607063
sample estimates:
  mean in group Male mean in group Female
                29.6                 33.1
```

ในการใช้ผลลัพธ์ที่ได้ทดสอบสมมุติฐานที่กำหนดไว้ ให้พิจารณาที่ค่าสถิติทดสอบ **t = -2.062** และ **p-value = 0.05395** หากเรากำหนดระดับนัยสำคัญของการทดสอบเท่ากับ 0.05 จะเห็นได้ว่า p-value มากกว่าระดับนัยสำคัญ 0.05 จึงยอมรับสมมุติฐานหลัก หมายความว่าอายุเฉลี่ยของผู้ชายและผู้หญิงในประชากรเป้าหมายนั้นไม่แตกต่างกัน ที่ระดับนัยสำคัญ 0.05

3.3.2 การทดสอบความแตกต่างระหว่างค่าเฉลี่ยประชากร 2 กลุ่ม เมื่อกลุ่มตัวอย่างไม่เป็นอิสระกัน

การทดสอบความแตกต่างระหว่างค่าเฉลี่ยของประชากร 2 กลุ่ม เมื่อกลุ่มตัวอย่างไม่เป็นอิสระกัน ซึ่งข้อมูลที่ได้จากกลุ่มตัวอย่างอาจได้จากการวัดค่าจากกลุ่มตัวอย่างเดียวกัน 2 ครั้ง เช่น การเปรียบเทียบน้ำหนักก่อนและหลังการเข้าโปรแกรมลดน้ำหนักของคนกลุ่มเดียวกัน หรือการเปรียบเทียบคะแนนก่อนเรียนกับหลังเรียนของนิสิตกลุ่มเดียวกัน เป็นต้น หรืออีกกรณีหนึ่งข้อมูลได้มาโดยการวัดค่าจากการจับกลุ่มตัวอย่างเป็นคู่ ๆ โดยใช้คุณลักษณะบางอย่าง เช่น การเปรียบเทียบระดับ IQ ของฝาแฝด หรือการเปรียบเทียบรายได้เริ่มต้นของผู้จบการศึกษาสาขาบริหารชายและหญิงที่มี GPA เท่ากัน

สำหรับการทดสอบสมมุติฐานเพื่อเปรียบเทียบค่าเฉลี่ยของประชากร 2 กลุ่ม เมื่อกลุ่มตัวอย่างไม่เป็นอิสระกัน ใช้วิธีการทดสอบ *t* ที่เรียกว่า **Dependent *t* test** หรือ **Match paired *t* test**

ตัวอย่างต่อไปนี้เป็นข้อมูลคะแนนสอบกลางภาคและปลายภาคของนิสิต 44 คน และเราต้องการทดสอบว่านิสิตมีพัฒนาการในการทำข้อสอบหรือไม่ นั่นคือ ต้องการทดสอบว่าคะแนนสอบปลายภาคมากกว่าคะแนนสอบปลายภาคหรือไม่

i	Midterm	Final
1	0.81	4.72
2	2.84	8.19
3	12.16	10.24
...	...	...
42	3.65	2.83
43	17.57	19.06
44	17.16	15.75

จากตารางข้อมูล จะเห็นได้ว่ามีตัวแปร 2 ตัวแปร คือ Midterm คือคะแนนสอบกลางภาค และ Final คือคะแนนสอบปลายภาค โดยไม่มีตัวแปรแบ่งกลุ่ม จะเห็นว่าลักษณะการเก็บข้อมูลจะไม่เหมือนกับการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของประชากร 2 กลุ่ม ที่เป็นอิสระกัน

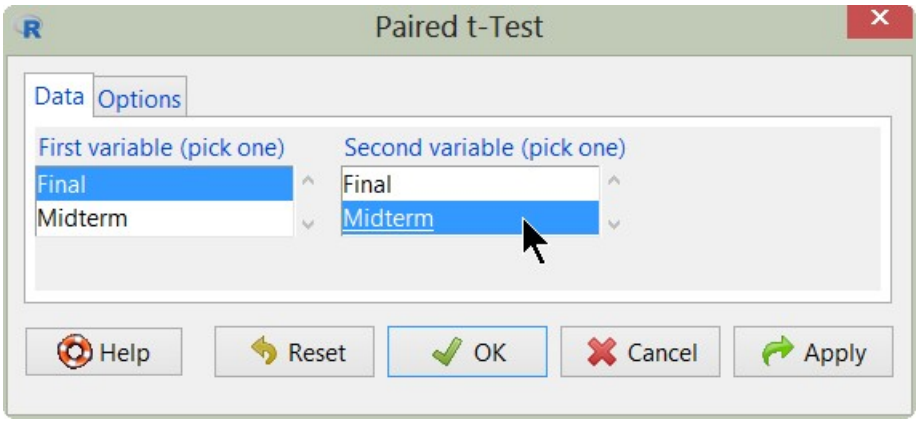
สำหรับตัวอย่างนี้ สมมุติฐานทางสถิติที่ทดสอบ คือ

$H_0 : \mu_D \leq 0$ (คะแนนสอบปลายภาคไม่มากกว่าคะแนนสอบกลางภาค)

$H_a : \mu_D > 0$ (คะแนนสอบปลายภาคมากกว่าคะแนนสอบกลางภาค)

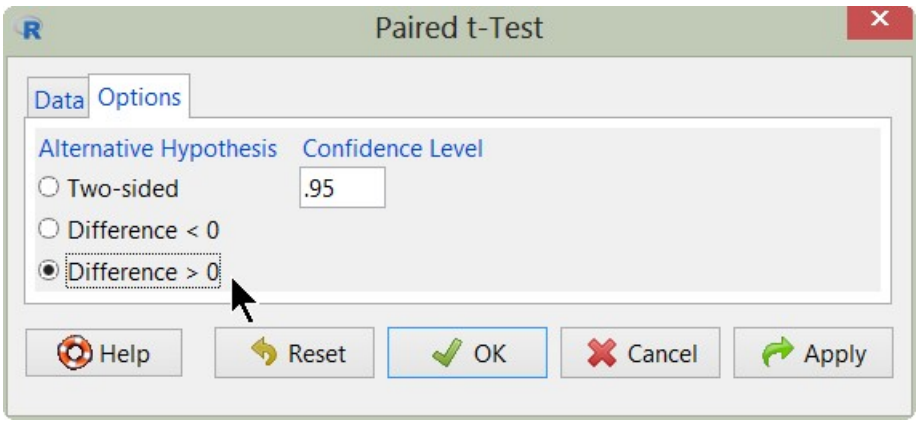
ใช้ **R** Commander ทำการทดสอบโดยใช้ Dependent t test ดังนี้

1. เลือกเมนู **Statistics -> Means -> Paired t-test...** จะได้หน้าต่าง



2. ที่ **First variable** เลือกตัวแปรตัวที่ 1 ในที่นี้เลือก Final และที่ **Second variable** เลือกตัวแปร ในที่นี้เลือก Midterm

3. คลิกปุ่ม Options ได้หน้าต่างดังนี้



4. ที่ **Alternative Hypothesis** เลือกชนิดของสมมุติฐานทางเลือกที่ต้องการทดสอบ ในที่นี้เลือก **Difference > 0** ตามสมมุติฐานทางสถิติที่ตั้งไว้)

5. คลิก **OK** ได้ผลลัพธ์ ดังนี้

```
Paired t-test

data:  score$Final and score$Midterm
t = 0.3707, df = 43, p-value = 0.3564
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.7966742        Inf
sample estimates:
mean of the differences
              0.2253381
```

จากผลลัพธ์ที่ได้ ได้ค่า **t = 0.3707** จำนวนองศาเสรี **df = 43** และ **p-value = 0.3564** หากเรากำหนดระดับนัยสำคัญของการทดสอบเท่ากับ 0.05 จะเห็นได้ว่าค่า p-value มากกว่าระดับนัยสำคัญ จึงยอมรับสมมุติฐานหลัก สรุปได้ว่า ที่ระดับนัยสำคัญ 0.05 คะแนนสอบปลายภาคเฉลี่ยของนิสิตไม่ได้มากกว่าคะแนนสอบกลางภาคเฉลี่ย หรืออีกนัยหนึ่ง นิสิตไม่มีพัฒนาการในการทำข้อสอบ เนื่องจากคะแนนสอบปลายภาคไม่ได้ดีกว่าคะแนนสอบกลางภาค