**การทดสอบความเป็นอิสระต่อกันของสองประชากร (Chi-square test)**

การทดสอบความเป็นอิสระต่อกันของสองตัวแปร (Test for independence, Contingency table test) บางครั้งเรามีข้อมูลลักษณะเป็นกลุ่ม (Category data) อยู่สองกลุ่ม เช่น เครื่องจักรกับชิ้นงานที่ผลิตได้ เพศของนักเรียนกับคณะที่เลือกอันดับหนึ่งในการสอบเข้าเรียนต่อในมหาวิทยาลัย เป็นต้น เราต้องการจะพิสูจน์ว่าตัวแปรที่หนึ่งเป็นเหตุหรือมีผลต่อตัวแปรที่สองหรือไม่ เครื่องมือที่ใช้เพื่อทำการทดสอบข้อมูลลักษณะนี้ เราเรียกว่า Contingency Table หรือบางครั้งก็เรียก Chi-square test การเก็บข้อมูล จะมีรูปแบบการบันทึกข้อมูลที่เป็นกฎเกณฑ์ ตามตารางดังต่อไปนี้

| | | Column Factor | | | | $Y_j$ | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | ..... | c | Totals |
| Row Factor $X_i$ | 1 | $O_1$ | $O_1$ | $O_1$ | ..... | $O_1$ | X |
| | 2 | $O_2$ | $O_2$ | $O_2$ | ..... | $O_2$ | X |
| | 3 | $O_3$ | $O_3$ | $O_3$ | ..... | $O_3$ | X |
| | ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| | ..... | ..... | ..... | ..... | ..... | ..... | ..... |
| | r | $O_r$ | $O_r$ | $O_r$ | ..... | $O_r$ | X |
| Totals | | Y | Y | Y | | Y | $F_{total}$ |

**การตั้งสมมติฐาน**

จุดประสงค์ของการทดสอบคือ เราต้องการทราบความสัมพันธ์ ระหว่างตัวแปรสองตัวแปร รูปแบบการตั้งสมมติฐานจะต้องเป็นตามรูปแบบต่อไปนี้

Ho: Factor 1 ไม่ขึ้นอยู่กับ Factor 2

Ha: Factor 1 ขึ้นอยู่กับ Factor 2

กำหนดค่า Alpha หรือกำหนดระดับนัยสำคัญ โดยปกติเราให้ a = 0.05 หมายถึงระดับในยังสำคัญที่ 95% ในกรณีนี้ เราใช้ c2 เป็น Test statistic

ขั้นตอนการพิสูจน์ มีตัวอย่างต่อไปนี้

**ตัวอย่างที่ 1** ในโรงงานผลิตสินค้าแห่งหนึ่ง วิศวกรต้องการทราบว่าจำนวนชิ้นงานที่พบข้อบกพร่องจากการผลิตขึ้นอยู่กับเครื่องจักรในสายการผลิตจำนวน 3 เครื่องนั้นหรือไม่ จึงได้สุ่มตรวจงานที่พบข้อบกพร่องโดยแยกรายละเอียดลักษณะของข้อบกพร่อง และแยกแต่ละเครื่องจักร บันทึกข้อมูลดังในตาราง

| | MC#1 | MC#2 | MC#3 |
|---|---|---|---|
| Scratch | 128 | 9 | 19 |
| Base crack | 44 | 87 | 40 |
| Arm bent | 18 | 21 | 53 |

คำถาม ลักษณะของข้องบกพร่องขึ้นอยู่กับแต่ละเครื่องจักรหรือไม่ เช่น ที่เครื่องจักร #1 พบว่าลักษณะข้อบกพร่อง Scratch มากกว่าลักษณะอื่นๆ ขณะที่เครื่อง #2 กลับแทบไม่มีอาการแบบนี้เลย กรณีที่เราได้ข้อมูลเป็นตามตาราง เราจะสรุปว่าอย่างไร

**ขั้นตอนการพิสูจน์**

**ตั้งสมมติฐาน**

Ho: ลักษณะของข้อบกพร่องของชิ้นงาน ไม่ได้ขึ้นอยู่กับ เครื่องจักร

Ha: ลักษณะของข้อบกพร่องของชิ้นงาน ขึ้นอยู่กับ เครื่องจักร

กำหนด a=0.05

หาผลรวมของข้อมูล ทั้งในแนว Column และ Row และ Total จากข้อมูลในตารางจะได้ดังนี้

| | MC#1 | MC#2 | MC#3 | |
|---|---|---|---|---|
| Scratch | 128 | 9 | 19 | 156 |
| Base crack | 44 | 87 | 40 | 171 |
| Arm bent | 18 | 21 | 53 | 92 |
| | 190 | 117 | 112 | 419 |

คำนวณหาค่าที่ควรจะเป็น Expected value (E) ของแต่ละช่อง (Cell) จากสมการ

$$Expected = \frac{X_{total} * Y_{total}}{Total}$$

เมื่อ Xtotal คือผลรวมในแนว Column ที่ Cell นั้นอยู่

Ytotal คือผลรวมในแนว Row ที่ Cell นั้นอยู่

Total คือผลรวมทั้งหมด

| | MC#1 | MC#2 | MC#3 |
|---|---|---|---|
| Scratch | E=190*156/419 | E=117*156/419 | E=112*156/419 |
| Base crack | E=190*171/419 | E=117*171/419 | E=112*171/419 |
| Arm bent | E=190*92/419 | E=117*92/419 | E=112*92/419 |

เมื่อคำนวณแล้วจะได้ค่า Expected ของแต่ละช่องดังต่อไปนี้

| | MC#1 | MC#2 | MC#3 |
|---|---|---|---|
| Scratch | E=70.74 | E=43.56 | E=41.70 |
| Base crack | E=77.54 | E=47.75 | E=45.71 |
| Arm bent | E=41.72 | E=25.69 | E=24.59 |

ค่า Expected value นี้เป็นค่าตามทฤษฎี โดยเปรียบเทียบว่า ถ้าผลรวมของข้อมูลในแนว Column และ Row ได้มาอย่างนี้ ในแต่ละ Cell นั้นควรจะได้จำนวนตัวอย่างมากเท่าใด (ตามกฎความน่าจะเป็น)

คำนวณหาค่า Chi-square ของแต่ Cell ตามสมการ

$$\chi^2 = \frac{(Actual - Expected)^2}{Expected}$$

ค่า c2 นี้ คือค่าความคลาดเคลื่อน ระหว่างค่า จริงกับ Expected value ในรูปกำลังสอง ซึ่งได้ค่าดังนี้

| | MC#1 | MC#2 | MC#3 |
|---|---|---|---|
| Scratch | c2 = 46.348 | c2 = 27.419 | c2 = 12.357 |
| Base crack | c2 = 14.507 | c2 = 32.263 | c2 = 0.713 |
| Arm bent | c2 = 13.486 | c2 = 0.856 | c2 = 32.823 |

6. คำนวณหาค่า Calculated Chi-square จาก

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

= 46.348+14.507+13.486+27.419+32.263+0.856+12.357+0.713+32.823

= 180.770

หาค่า Degree of freedom จาก

df = (C-1)(R-1)

เมื่อ C คือจำนวน Column

R คือจำนวน Row

ดังนั้น df = (3-1)(3-1) = 2*2 = 4

8. หาค่า $\chi^2_{critical}$ จาก Chi-square table ที่ df = 4 , a = 0.05

จากตารางได้ $\chi^2_{critical}$ = 9.488

9. สรุปผลการทดสอบสมมติฐาน

ถ้า $\chi^2_{cal} > \chi^2_{critical}$ เราถือว่า ปฏิเสธ *Ho*

จากสมมติฐาน

Ho: Factor 1 ไม่ขึ้นอยู่กับ Factor 2

Ha: Factor 1 ขึ้นอยู่กับ Factor 2

เมื่อเราปฏิเสธ Ho ในกรณีตัวอย่างนี้จึงสรุปได้ว่า เครื่องจักรแต่ละเครื่อง ทำให้เกิดลักษณะอาการข้องบกพร่องของชิ้นงาน แตกต่างกัน หมายความว่า หากจะลดอาการ Scratch ก็ต้องแก้ที่เครื่อง 1 อาจจะไม่จำเป็นต้องแก้ เครื่อง 2 หรือ 3 ก็ได้

กรณีตัวอย่างนี้ สมมติว่า เรายอมรับ Ho แปลว่า ลักษณะของข้อบกพร่อง ไม่ได้เกี่ยวกับ เครื่องจักรเลย นั่นคือ Scratch, Base crack และ Arm bent จะมีจำนวนเรียงลำดับอย่างนี้ทุกเครื่อง นั่นอาจเป็นเพราะ เหตุอื่นที่ไม่ใช่เครื่องจักรก็ได้ ที่ทำให้เกิดข้อบกพร่องในชิ้นงาน

#Assignment 9 Chi-Square Test หากพิสูจน์โดยใช้ โปรแกรม R จะได้ดังนี้

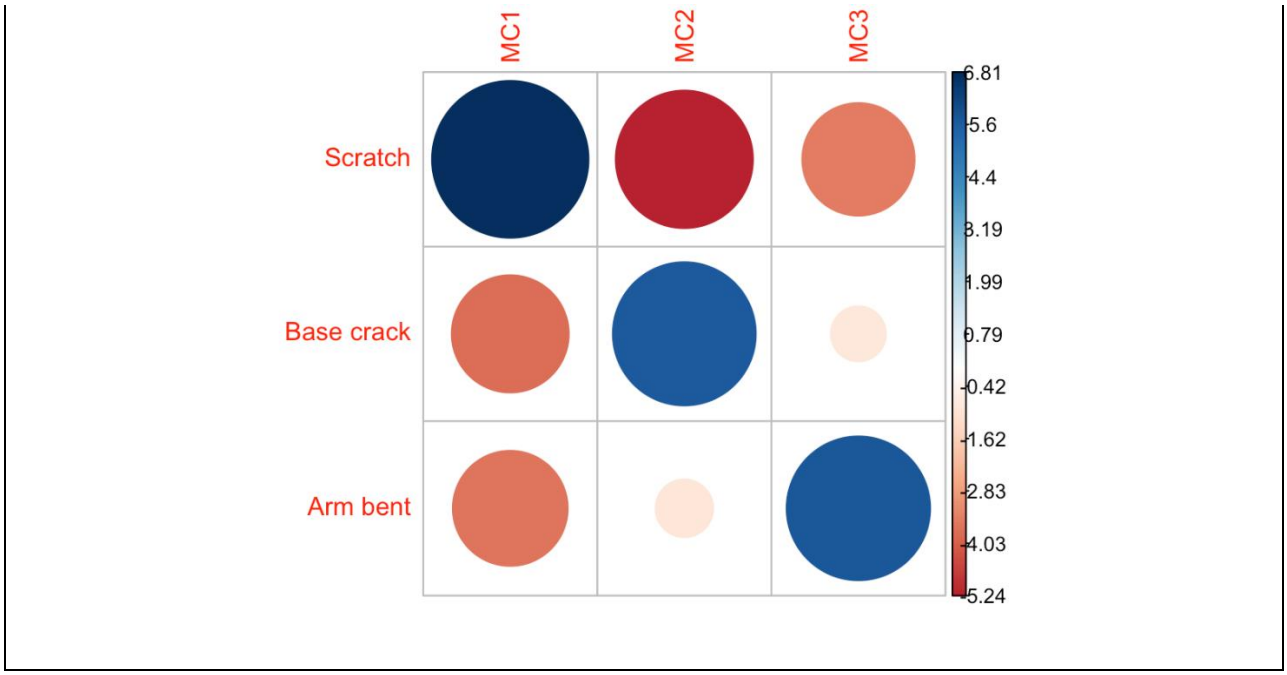
#Chi-Square Test: MC#1, MC#2 and MC#3

```
library(readr)
file_path <- "chi-square_ex_txt.txt"
chi_square_ex <- read.delim(file_path, row.names = 1)

chisq <- chisq.test(chi_square_ex)
chisq
```

```
        Pearson's Chi-squared test

data:  chi_square_ex
X-squared = 180.77, df = 4, p-value < 2.2e-16
```

#ค่า P-Value น้อยกว่า a แปลว่าเรา ปฏิเสธ Ho

```
# Observed counts
chisq$observed

# Expected counts
round(chisq$expected,2)

library(corrplot)
corrplot(chisq$residuals, is.cor = FALSE)
```




ตัวอย่างที่ 2 ผลสำรวจตัวอย่างของนักเรียนชั้นมัธยมศึกษาปีที่ 6 ที่กำลังจะสมัครสอบเข้าเรียนต่อในมหาวิทยาลัยเกี่ยวกับสาขาที่เลือกเป็นอันดับ 1 ได้ผลดังตารางต่อไปนี้

| | Male | Female |
|---|---|---|
| Engineering | 321 | 89 |
| Information Technology | 156 | 251 |
| Pure science | 89 | 174 |
| Law | 74 | 59 |
| Mass communication | 15 | 84 |
| Social Science | 21 | 47 |

คำถามคือ " เพศเป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1 ของนักเรียนหรือไม่ " ให้ทำการทดสอบสมมติฐาน โดยให้ระดับความมั่นใจที่ 95%

ขั้นตอนการพิสูจน์

**ตั้งสมมติฐาน**

สมมติฐานของผู้ที่ทำการวิจัยในข้อนี้คือ “เพศไม่ได้เป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1 ” เพราะตามหลักวิชาสถิติ สมมติฐานจะต้องถือว่า ปัจจัยหนึ่งไม่ได้ได้มีผลต่ออีกปัจจัยหนึ่ง จนกว่าจะมีหลักฐานมายืนยันว่ามีผล เมื่อเขียนสมมติฐานตามหลักวิชาสถิติจะได้ดังนี้

Ho: เพศไม่ได้เป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1

Ha: เพศเป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1

หาผลรวมในแนว Row และ Column ได้ดังนี้

| | Male | Female | |
|---|---|---|---|
| Engineering | 321 | 89 | 410 |
| Information Technology | 156 | 251 | 407 |
| Pure science | 89 | 174 | 263 |
| Law | 74 | 59 | 133 |
| Mass communication | 15 | 84 | 99 |
| Social Science | 21 | 47 | 68 |
| | 676 | 704 | 1380 |

3. คำนวณหาค่า Expected value ของแต่ละช่อง (Cell) ได้ดังนี้

| | Male | Female |
|---|---|---|
| Engineering | 200.84 | 209.16 |
| Information Technology | 199.37 | 207.63 |
| Pure science | 128.83 | 134.17 |
| Law | 65.15 | 67.85 |
| Mass communication | 48.5 | 50.5 |
| Social Science | 33.31 | 34.69 |

4. คำนวณหาค่า c2 ของแต่ละ Cell ได้ค่าดังต่อไปนี้

| | Male | Female |
|---|---|---|
| Engineering | 71.89 | 69.03 |

| | | |
|---|---|---|
| Information Technology | 9.43 | 9.06 |
| Pure science | 12.32 | 11.83 |
| Law | 1.2 | 1.15 |
| Mass communication | 23.14 | 22.22 |
| Social Science | 4.55 | 4.37 |

คำนวณหาค่า Chi-square

$\chi^2_{cal}$ = 71.89+9.43+12.32+1.20+23.14+4.55+69.03+9.06+11.83+1.15+22.22+4.37

= 240.179

6. หาค่า *df*

*df = (C-1)(R-1) = (2-1)(6-1) = 5*

10. หาค่า $\chi^2_{critical}$ จาก *Chi-square table* ที่ *df* = 5 , a = 0.05

จากตารางจะได้ $\chi^2_{critical}$ = 11.07

7. สรุปผลการทดสอบสมมติฐาน

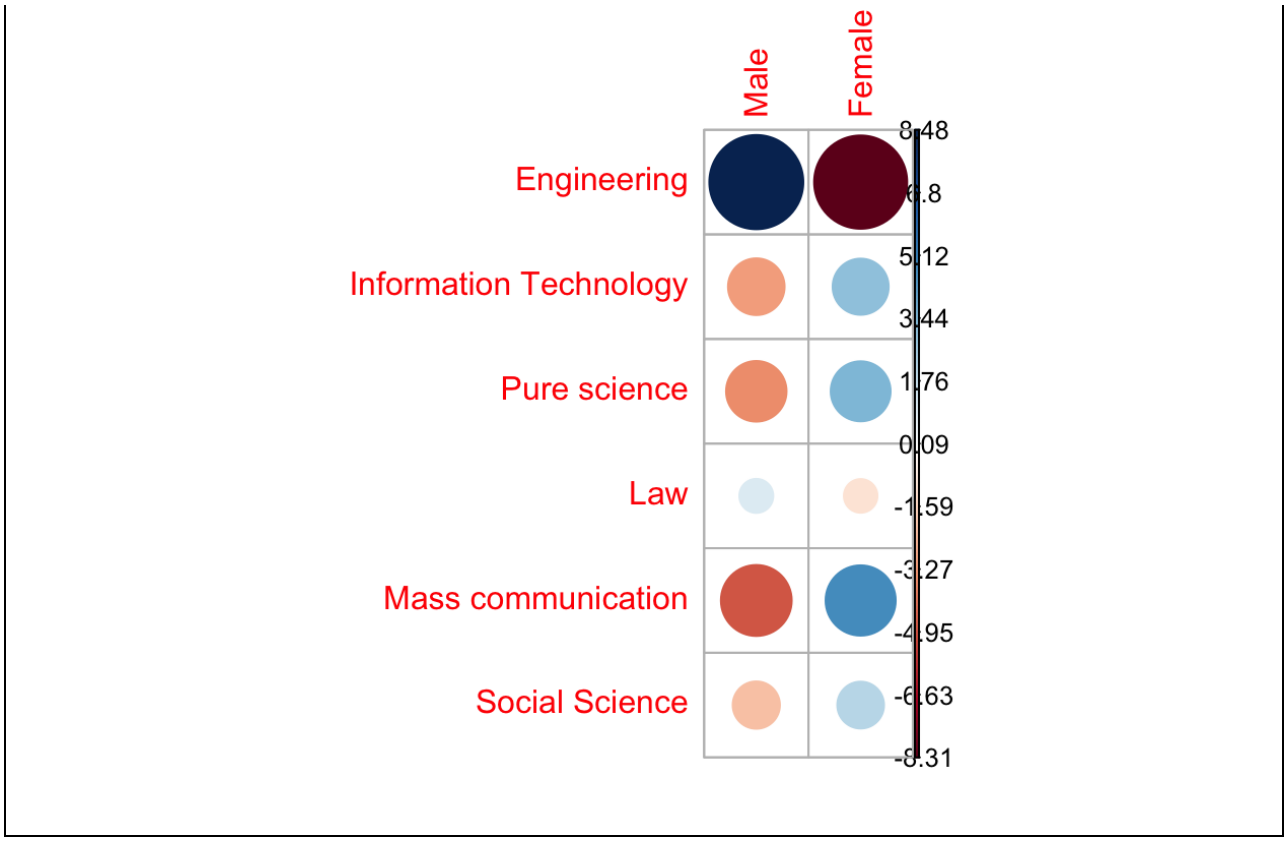
เนื่องจาก $\chi^2_{cal} > \chi^2_{critical}$ เราถือว่า ปฏิเสธ *Ho* *ที่ว่า* เพศไม่ได้เป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1

Assignment 9 Chi-Square Test เมื่อใช้ โปรแกรม R ในการวิเคราะห์ข้อมูล จะได้ดังนี้

```
file_path <- "chi-square_ex02.txt"
chi_square_ex02 <- read.delim(file_path, row.names = 1)

chisq <- chisq.test(chi_square_ex02)
chisq
```

```
        Pearson's Chi-squared test

data:  chi_square_ex02
X-squared = 240.18, df = 5, p-value < 2.2e-16
```

```
#เมื่อ P-Value น้อยกว่า a เราจึงปฏิเสธ Ho

# Observed counts
chisq$observed

# Expected counts
round(chisq$expected,2)

library(corrplot)
corrplot(chisq$residuals, is.cor = FALSE)
```




จากสมมติฐาน

Ho: เพศไม่ได้เป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1

Ha: เพศเป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1

เมื่อเราต้องปฏิเสธ Ho: เราสรุปว่า เพศเป็นปัจจัยที่มีผลต่อการเลือกคณะอันดับ 1 ในการสอบเข้าเรียนในมหาวิทวิทยาลัย ของนักเรียนมัธยม หมายความว่า นักเรียนชายก็จะเลือกคณะที่จะสอบเข้าอันดับ 1 แตกต่างจากนักเรียนหญิง เป็นส่วนใหญ่

ข้อกำหนด ในการใช้ Contingency table ในการพิสูจน์สมมติฐานที่พึงระวังอย่างมาก

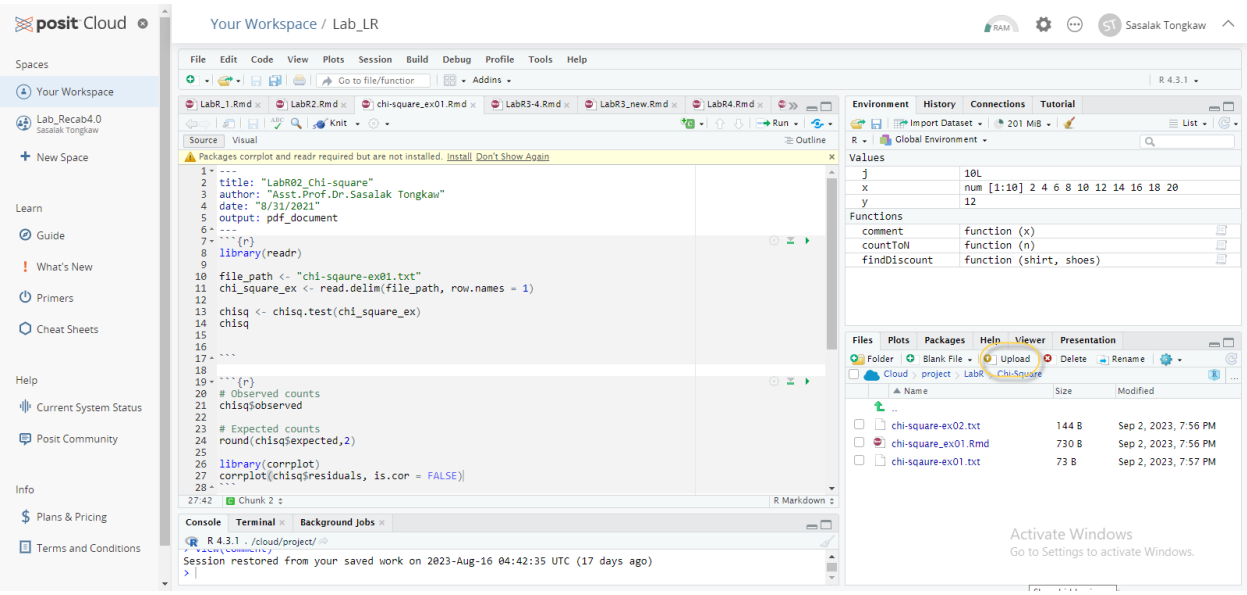
1. ไม่มีข้อกำหนดเรื่องขนาดของสิ่งตัวอย่างตายตัว เพียงแต่ในระหว่างเก็บข้อมูลหากข้อมูลที่เก็บมาได้ใน Cell ใด Cell หนึ่ง มีค่าน้อยกว่า 5 นั่นเป็นข้อบ่งบอกว่า เรายังเก็บข้อมูลไม่เพียงพอ จำเป็นต้องใช้ขนาดสิ่งตัวอย่างเพิ่มขึ้น ถ้าหากเรายังฝืนใช้ข้อมูลดังกล่าววิเคราะห์ ผลสรุปก็จะนำไปสู่การสรุปที่ผิดพลาดอยู่ดี เป็นเรื่องที่ไม่ควรทำอย่างมาก

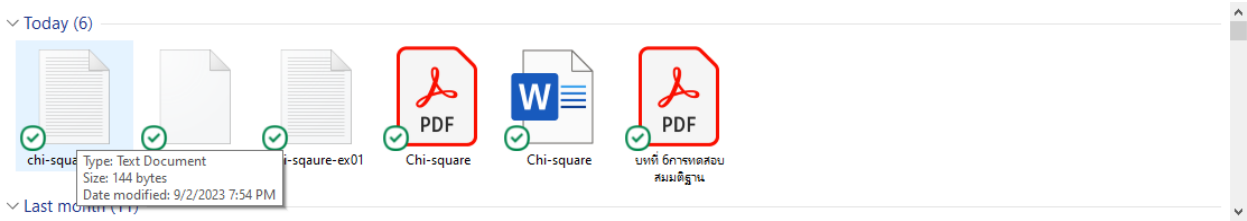
2. ในการทดสอบสมมติฐานนั้น สิ่งสำคัญอยู่ที่การเก็บและบันทึกข้อมูล จะต้องเก็บข้อมูลอย่างสุ่ม ขอบเขตของสิ่งตัวอย่างนั้นขึ้นอยู่กับขอบเขตการทำการวิจัย เช่น ถ้าผู้ทำการวิจัยตามตัวอย่างที่ 2 ต้องการศึกษาเฉพาะโรงเรียนในเขตกรุงเทพฯ ก็ให้ทำการเก็บข้อมูลโรงเรียนในกรุงเทพฯ เท่านั้น ในขณะเก็บข้อมูล จะต้องสุ่มสอบถามตัวอย่าง โดยที่ผู้สอบถามไม่ได้รู้ว่าตัวอย่างเลือกคณะอะไรมาก่อนเสมอ

3. Contingency Table ให้ข้อสรุปเพียงแค่ว่า Factor หนึ่งขึ้นอยู่กับอีก Factor หรือไม่ หรือ Factor หนึ่งมีผลหรือมีอิทธิพลต่ออีก Factor หรือไม่ เท่านั้น

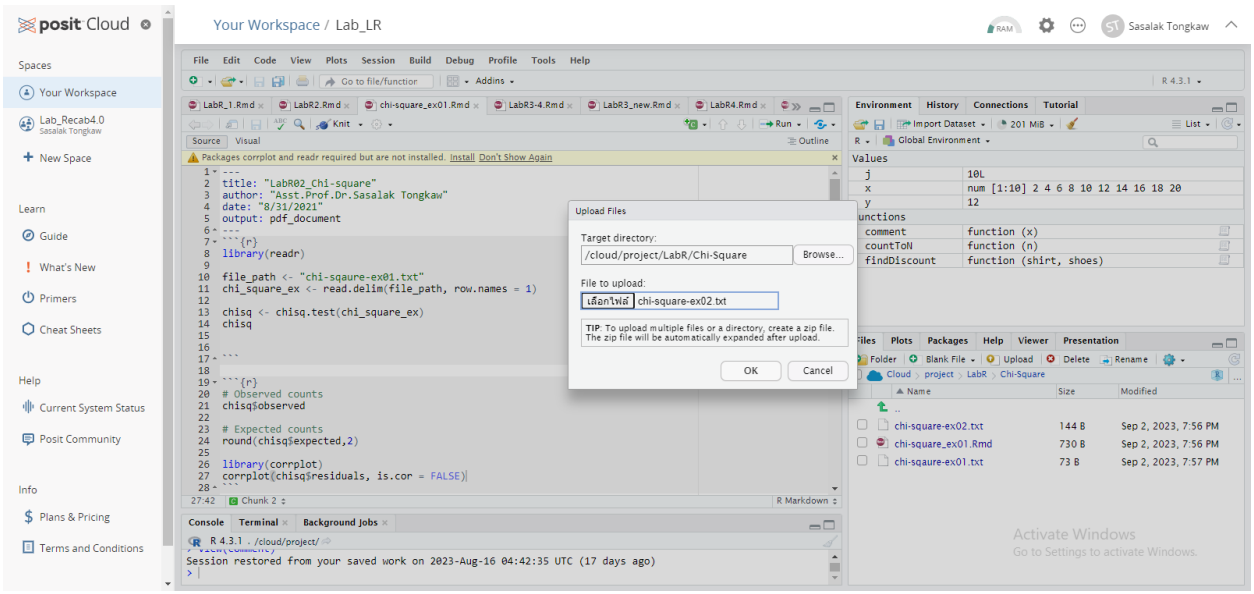
## แบบฝึกหัด

1. เปิดโปรแกรม Rstudio Cloud โดยให้คลิกที่ File ในส่วนขวาล่าง จะมีคำว่า upload ให้อัพโหลดไฟล์ทั้ง 3 ไฟล์ขึ้นไปไว้บน cloud ดังภาพ ไฟล์ทั้ง 3 ไฟล์ อาจารย์ให้ไว้ใน Assignment 9 Chi-Square Test

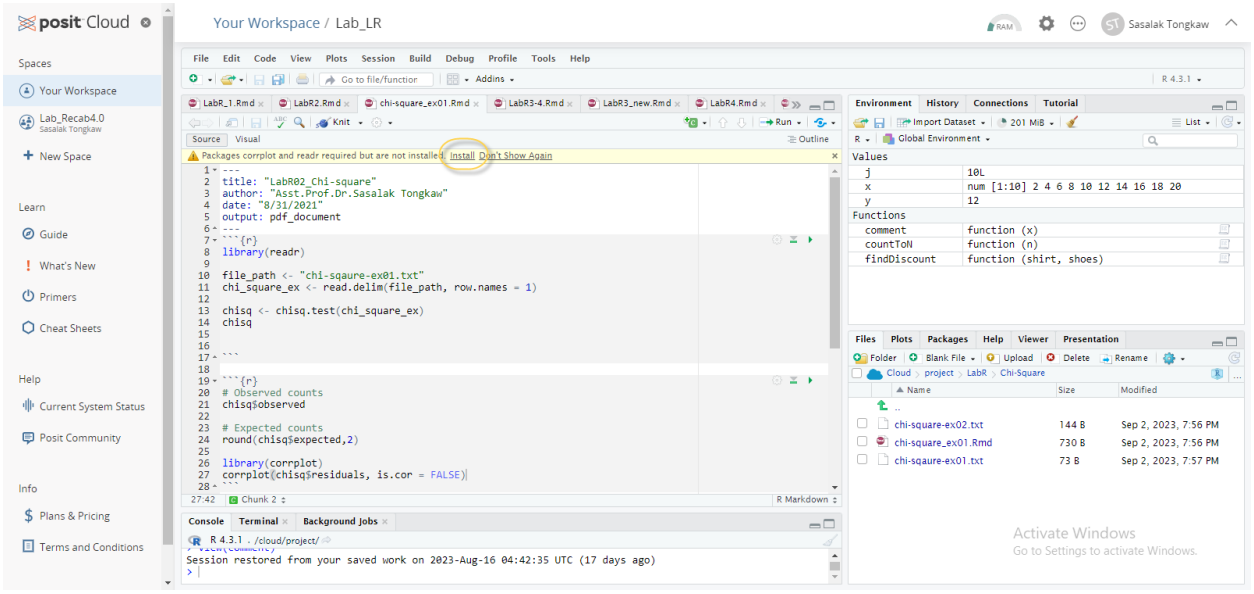



2. เลือกไฟล์ทีละไฟล์ แล้วคลิก OK ทำซ้ำ ทั้ง 3 ไฟล์




3. คลิกที่ คำว่า install ที่แถบสีเหลือง เพื่อ install package corrplot และ readr




4. ทดลองรันข้อมูล แล้ว Knit เป็น .PDF เปลี่ยนชื่อไฟล์เป็น LabR02.rmd.knit (ดูตัวอย่างของอาจารย์) แล้วอัพโหลดใน Assignment 9