Survey Data Analysis with R (2)

Chia-hung Tsai

Aug. 27, 2018

Goals

- Cross-table
- Descriptive statistics

Preparation

- Install R from here
- Install RStudio IDE
- Install `survey`, `foreign`, `car` packages

Dataset

We will run a survey data to show how R works, so you should download **TEDS2016_indQ.sav** file to your working directory

Data

You can import dataset from RStudio's **File** and **Import Dataset**. The codes will looks like the following:

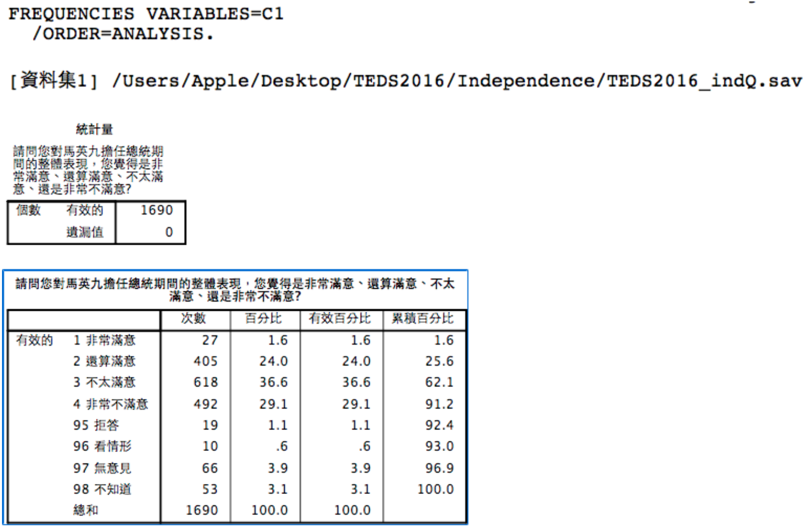
```r
library(haven)
TEDS2016_indQ <- read_sav("TEDS2016_indQ.sav")
```

Or you can use library **foreign** to read the data:

```r
library(foreign)
df <- read.spss("TEDS2016_indQ.sav", to.data.frame=T,
                use.value.labels = F)
```

Cross-table

Let's analyze the relationship between gender, age, and satisfaction with President Ma. Ma's performance is:



```
FREQUENCIES VARIABLES=C1
  /ORDER=ANALYSIS.
```

```
[資料集1] /Users/Apple/Desktop/TEDS2016/Independence/TEDS2016_indQ.sav
```

統計量

請問您對馬英九擔任總統期間的整體表現，您覺得是非常滿意、還算滿意、不太滿意、還是非常不滿意?

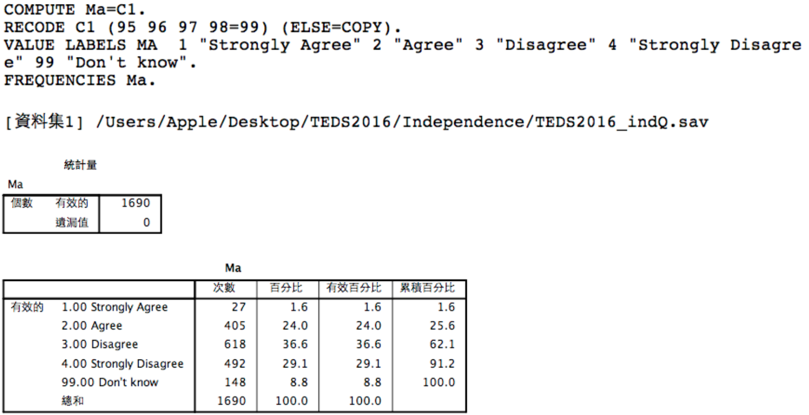
個數	有效的	1690
	遺漏值	0

請問您對馬英九擔任總統期間的整體表現，您覺得是非常滿意、還算滿意、不太滿意、還是非常不滿意?

		次數	百分比	有效百分比	累積百分比
有效的	1 非常滿意	27	1.6	1.6	1.6
	2 還算滿意	405	24.0	24.0	25.6
	3 不太滿意	618	36.6	36.6	62.1
	4 非常不滿意	492	29.1	29.1	91.2
	95 拒答	19	1.1	1.1	92.4
	96 看情形	10	.6	.6	93.0
	97 無意見	66	3.9	3.9	96.9
	98 不知道	53	3.1	3.1	100.0
	總和	1690	100.0	100.0	

Recode (SPSS)

Before further analysis, we can create a “Don’t Know” category to contain every ambiguous response.



```
COMPUTE Ma=C1.
RECODE C1 (95 96 97 98=99) (ELSE=COPY).
VALUE LABELS MA  1 "Strongly Agree" 2 "Agree" 3 "Disagree" 4 "Strongly Disagre
e" 99 "Don't know".
FREQUENCIES Ma.
```

```
[資料集1] /Users/Apple/Desktop/TEDS2016/Independence/TEDS2016_indQ.sav
```

統計量

Ma

個數	有效的	1690
	遺漏值	0

Ma

		次數	百分比	有效百分比	累積百分比
有效的	1.00 Strongly Agree	27	1.6	1.6	1.6
	2.00 Agree	405	24.0	24.0	25.6
	3.00 Disagree	618	36.6	36.6	62.1
	4.00 Strongly Disagree	492	29.1	29.1	91.2
	99.00 Don't know	148	8.8	8.8	100.0
	總和	1690	100.0	100.0	

Recode (R)

```r
library(foreign)
df <- read.spss("TEDS2016_indQ.sav", to.data.frame=T,
                use.value.labels = T)
class(df$C1)
```

```
## [1] "factor"
```

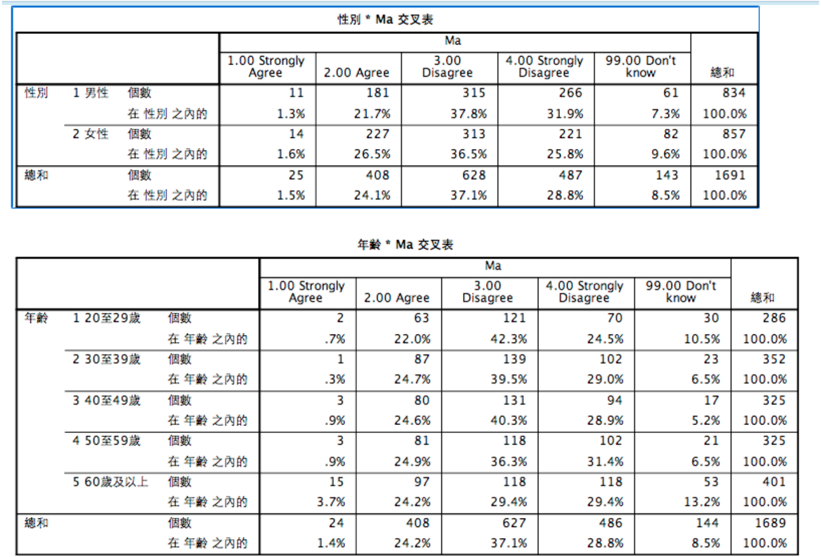
```r
df$Ma <- as.numeric(df$C1)
```

```r
library(car)
df$Ma <- recode(df$Ma, "5:8=99")
table(df$Ma)
```

```
##
##   1   2   3   4  99
##  27 405 618 492 148
```

Cross-table by SPSS

```
WEIGHT BY w.
CROSSTABS
  /TABLES=Sex Age BY Ma
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```



性別 * Ma 交叉表

			Ma					
			1.00 Strongly Agree	2.00 Agree	3.00 Disagree	4.00 Strongly Disagree	99.00 Don't know	總和
性別	1 男性	個數	11	181	315	266	61	834
		在 性別 之內的	1.3%	21.7%	37.8%	31.9%	7.3%	100.0%
	2 女性	個數	14	227	313	221	82	857
		在 性別 之內的	1.6%	26.5%	36.5%	25.8%	9.6%	100.0%
總和		個數	25	408	628	487	143	1691
		在 性別 之內的	1.5%	24.1%	37.1%	28.8%	8.5%	100.0%

年齡 * Ma 交叉表

			Ma					
			1.00 Strongly Agree	2.00 Agree	3.00 Disagree	4.00 Strongly Disagree	99.00 Don't know	總和
年齡	1 20至29歲	個數	2	63	121	70	30	286
		在 年齡 之內的	.7%	22.0%	42.3%	24.5%	10.5%	100.0%
	2 30至39歲	個數	1	87	139	102	23	352
		在 年齡 之內的	.3%	24.7%	39.5%	29.0%	6.5%	100.0%
	3 40至49歲	個數	3	80	131	94	17	325
		在 年齡 之內的	.9%	24.6%	40.3%	28.9%	5.2%	100.0%
	4 50至59歲	個數	3	81	118	102	21	325
		在 年齡 之內的	.9%	24.9%	36.3%	31.4%	6.5%	100.0%
	5 60歲及以上	個數	15	97	118	118	53	401
		在 年齡 之內的	3.7%	24.2%	29.4%	29.4%	13.2%	100.0%
總和		個數	24	408	627	486	144	1689
		在 年齡 之內的	1.4%	24.2%	37.1%	28.8%	8.5%	100.0%

Figure 1:

Cross-table by R

```
library(survey)
# attributes
df$Gender <- as.numeric (df$Sex)
df$Age5 <- as.numeric (df$Age)
#weigting
dfw <- svydesign(ids = ~1, data = df, weights = df$w)
#cross-table
svytable(~Gender+Ma, design=dfw)
```

```
##       Ma
## Gender     1     2     3     4    99
##      1  10.6 181.4 314.5 265.7  61.0
##      2  13.6 226.7 313.3 221.3  82.0
```

```
100*prop.table(svytable(~Age5+Ma, design=dfw),1)
```

```
##     Ma
## Age5      1      2      3      4     99
```

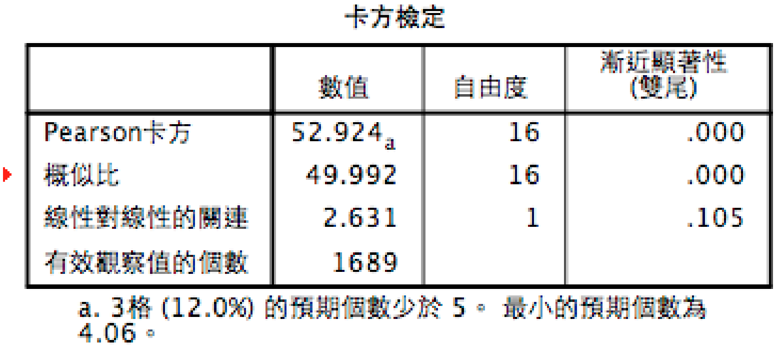
Chi-squared

- We use chi-squared value to see if the two variables are indepedent. If chi-squared value is large, the probability of observing such value is very small. Therefore, we can reject the null hypothesis that the two variables are independent. In other words, these two variables are associated. When variable A changes, variable B will also change.
- We can calculate the chi-squared value of survey data with R

```
svychisq(~Age+Ma, design=dfw,
               statistic="Chisq")
```

```
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  svychisq(~Age + Ma, design = dfw, statistic = "Ch
## X-squared = 50, df = 20, p-value = 6e-05
```

```
CROSSTABS
  /TABLES=Age BY Ma
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```



卡方檢定

	數值	自由度	漸近顯著性 (雙尾)
Pearson卡方	52.924[a]	16	.000
概似比	49.992	16	.000
線性對線性的關連	2.631	1	.105
有效觀察值的個數	1689		

a. 3格 (12.0%) 的預期個數少於 5。 最小的預期個數為 4.06。