# Survey Data Analysis with R (1)

Chia-hung Tsai

Aug. 13, 2018

# Goals

- Using R to analyze survey data in SPSS format
- Using weight
- Frequency tables
- 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:

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

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

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

# Use survey package

Before we weight the data, let's look at the frequencies of sex and eduation.

```
prop.table(table(df$Sex))
```

```
##
##     1     2
## 0.514 0.486
```

```
prop.table(table(df$Age))
```

```
##
##     1     2     3     4     5
## 0.156 0.167 0.188 0.199 0.290
```

# Comparing SPSS and R

We can compare the results with SPSS output.

性別

| | | 次數 | 百分比 | 有效百分比 | 累積百分比 |
|---|---|---|---|---|---|
| 有效的 | 男性 | 868 | 51.4 | 51.4 | 51.4 |
| | 女性 | 822 | 48.6 | 48.6 | 100.0 |
| | 總和 | 1690 | 100.0 | 100.0 | |

年齡

| | | 次數 | 百分比 | 有效百分比 | 累積百分比 |
|---|---|---|---|---|---|
| 有效的 | 20至29歲 | 264 | 15.6 | 15.6 | 15.6 |
| | 30至39歲 | 282 | 16.7 | 16.7 | 32.3 |
| | 40至49歲 | 317 | 18.8 | 18.8 | 51.1 |
| | 50至59歲 | 337 | 19.9 | 19.9 | 71.0 |
| | 60歲及以上 | 490 | 29.0 | 29.0 | 100.0 |
| | 總和 | 1690 | 100.0 | 100.0 | |

# Using weight

We weight the data as follows.

```
library(survey)
dfw <- svydesign(ids = ~1, data = df, weights = df$w)
```

We can check the frequencies

```
prop.table(svytable(~Sex, design=dfw))
```

```
## Sex
##     1     2
## 0.493 0.507
```

```
prop.table(svytable(~Age, design=dfw))
```

```
## Age
##     1     2     3     4     5
## 0.170 0.209 0.192 0.192 0.237
```

# Comparing SPSS and R

We can compare the results with SPSS output.

性別

| | | 次數 | 百分比 | 有效百分比 | 累積百分比 |
|---|---|---|---|---|---|
| 有效的 | 男性 | 833 | 49.3 | 49.3 | 49.3 |
| | 女性 | 857 | 50.7 | 50.7 | 100.0 |
| | 總和 | 1690 | 100.0 | 100.0 | |

年齡

| | | 次數 | 百分比 | 有效百分比 | 累積百分比 |
|---|---|---|---|---|---|
| 有效的 | 20至29歲 | 287 | 17.0 | 17.0 | 17.0 |
| | 30至39歲 | 353 | 20.9 | 20.9 | 37.8 |
| | 40至49歲 | 325 | 19.2 | 19.2 | 57.1 |
| | 50至59歲 | 324 | 19.2 | 19.2 | 76.3 |
| | 60歲及以上 | 401 | 23.7 | 23.7 | 100.0 |
| | 總和 | 1690 | 100.0 | 100.0 | |

# Data cleaning

## Recode "Don't Know"

```
prop.table(svytable(~Edu, design=dfw))
```

```
## Edu
##        1        2        3        4        5        9
## 0.14757 0.12761 0.27894 0.12265 0.31741 0.00582
```

■ After recoding, be sure to weight your data. Otherwise you cannnot find the new weighted variable

```
library(car)
df$Edu.5<-recode(df$Edu, "9=NA")
dfw <- svydesign(ids = ~1, data = df, weights = df$w)
```

```r
prop.table(svytable(~Edu.5, design=dfw))
```

```
## Edu.5
##     1     2     3     4     5
## 0.148 0.128 0.281 0.123 0.319
```

## Reducing the number of categories

```
df$Edu.3<-recode(df$Edu, "1:2=1; 3=2; 4:5=3; 9=NA")
dfw <- svydesign(ids = ~1, data = df, weights = df$w)
prop.table(svytable(~Edu.3, design=dfw))
```

```
## Edu.3
##     1     2     3
## 0.277 0.281 0.443
```