HW2

Jiang, Tai-Jia

2024-10-05

Table of contents

將檔案讀取並檢查。由summary()可以看出，此資料多為類別型變數。

```
#
library(readxl)
```

```
Warning: package 'readxl' was built under R version 4.3.3
```

```
data <- read_excel("C:/Users/zach1/Downloads/cleaned_dataset.xlsx")
head(data)
```

```
# A tibble: 6 x 15
  v1    v2    v3    v4_1  v4_2  v4_3  v4_4  v4_5  v4_6  v4_7  v4_8  v5    v6
  <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
1 ~     ~     ~     ~     ~     ~     ~     ~     ~     ~     ~     60    ~
2 ~     ~     (1 )  (4 )  (7 )  (10~  ~     ~     ~     ~     (7 )  60    ~
3 ~     ~     (4 )  ~     ~     ~     ~     ~     ~     ~     (6 )  60    ~
4 ~     ~     (3 )  (10~  ~     ~     ~     ~     ~     ~     (7 )  60    ~
5 ~     ~     ~     ~     ~     ~     ~     ~     ~     ~     ~     60    ~
6 ~     ~     ~     ~     ~     ~     ~     ~     ~     ~     (10~  60    ~
# i 2 more variables: v7 <chr>, v8 <chr>
```

```
summary(data)
```

```
       v1                  v2                  v3                  v4_1
 Length:1671         Length:1671         Length:1671         Length:1671
 Class :character    Class :character    Class :character    Class :character
 Mode  :character    Mode  :character    Mode  :character    Mode  :character
      v4_2                v4_3                v4_4                v4_5
 Length:1671         Length:1671         Length:1671         Length:1671
 Class :character    Class :character    Class :character    Class :character
 Mode  :character    Mode  :character    Mode  :character    Mode  :character
      v4_6                v4_7                v4_8                 v5
 Length:1671         Length:1671         Length:1671         Length:1671
 Class :character    Class :character    Class :character    Class :character
 Mode  :character    Mode  :character    Mode  :character    Mode  :character
       v6                  v7                  v8
 Length:1671         Length:1671         Length:1671
 Class :character    Class :character    Class :character
 Mode  :character    Mode  :character    Mode  :character
```

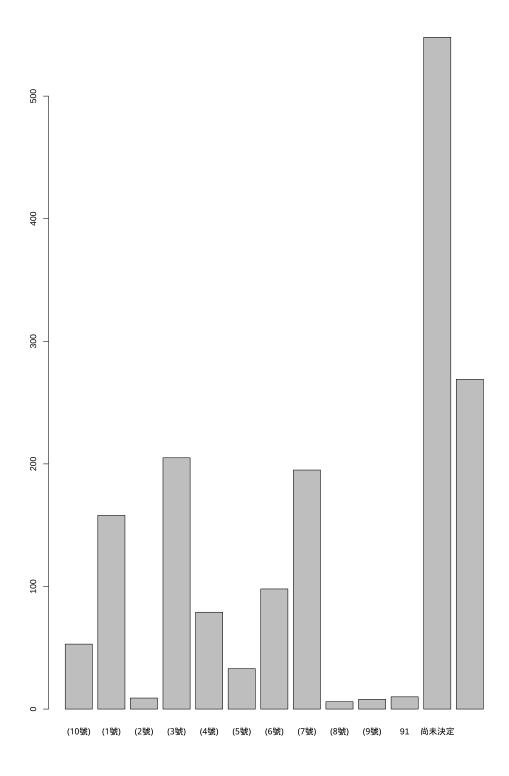將問題5(若明天就要投票，請問您會把票投給誰?)做為候選人的支持度，並分別列出由小到大的表格、樣本比例表格與次數的直條圖。

```r
sort(table(data$v5),decreasing = T)
```

```
        (3 )    (7 )    (1 )    (6 )    (4 )   (10 )
    548     269     205     195     158      98      79      53
 (5 )      91    (2 )    (9 )    (8 )
     33      10       9       8       6
```
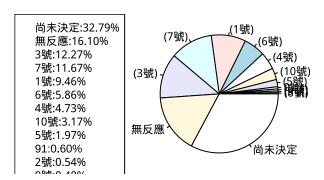
```r
round(sort(table(data$v5),decreasing = T)/length(data$v1),digits = 4)
```

```
        (3 )    (7 )    (1 )    (6 )    (4 )   (10 )
 0.3279  0.1610  0.1227  0.1167  0.0946  0.0586  0.0473  0.0317
 (5 )      91    (2 )    (9 )    (8 )
 0.0197  0.0060  0.0054  0.0048  0.0036
```

```r
barplot(table(data$v5))
```

(10號) (1號) (2號) (3號) (4號) (5號) (6號) (7號) (8號) (9號) 91 尚未決定

若將樣本比例以圓餅圖表示(如下)，則可以發現去除"無反應"與"尚未決定"的選項，以3號候選人所佔比例最大(12.27%)，其次為7號候選人(11.67%)。

```
pie(sort(table(data$v5)),cex=0.7)
legend("topleft",legend=c("  :32.79%","  :16.10%","3 :12.27%",
                          "7 :11.67%","1 :9.46%","6 :5.86%","4 :4.73%",
                          "10 :3.17%","5 :1.97%","91:0.60%","2 :0.54%",
                          "9 :0.48%","8 :0.36%"),cex = 0.7)
```

接著，假設這10位候選人與"無反應"、"尚未決定"和"91"服從Multinomial distribution，如下Eq(1)所示，則可以進行區間估計，如Eq(2)所示。

$$\mathsf{X} \sim Multinom(n, \mathsf{p}), \text{where } \mathsf{p} \text{ equals to } [p_1, p_2, ..., p_{12}]. \ (1)$$

$$\hat{se}(X_i) = \sqrt{\frac{\hat{p}_i(\hat{p}_i - 1)}{n}}, C.I.(p_i) = \hat{p}_i \pm Z_{\frac{\alpha}{2}} (2)$$
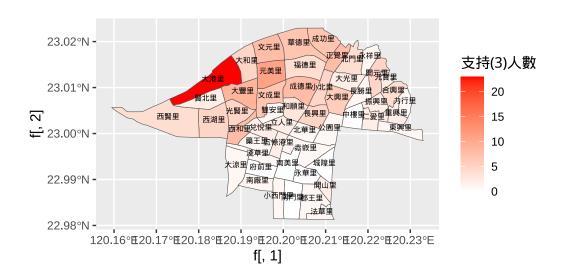
計算 p_i 的 confidence interval

4

```r
n <- length(data$v1)
p <- sort(table(data$v5),decreasing = T)/n
se <- ((p*(1-p))/n)^0.5
lower <- p-1.96*se
upper <- p+1.96*se
round(se,3)
```

```
        (3 )    (7 )    (1 )    (6 )    (4 )    (10 )
 0.011   0.009   0.008   0.008   0.007   0.006   0.005   0.004
 (5 )      91    (2 )    (9 )    (8 )
 0.003   0.002   0.002   0.002   0.001
```

```r
CI <- data.frame(matrix(ncol=1))
for (i in 1:13){
    CI[i] <- paste("[",round(lower[i],digits = 3),",",round(upper[i],3),"]")
}
colnames(CI) <- c("  "," ","3 ","7 ","1 ","6 ","4 ","10 ","5 ","91","2 ","9 ","8 ")

CI
```

```
                              3                   7
1 [ 0.305 , 0.35 ] [ 0.143 , 0.179 ] [ 0.107 , 0.138 ] [ 0.101 , 0.132 ]
              1               6               4               10
1 [ 0.081 , 0.109 ] [ 0.047 , 0.07 ] [ 0.037 , 0.057 ] [ 0.023 , 0.04 ]
              5               91              2               9
1 [ 0.013 , 0.026 ] [ 0.002 , 0.01 ] [ 0.002 , 0.009 ] [ 0.001 , 0.008 ]
              8
1 [ 0.001 , 0.006 ]
```

使用地圖進行競選策略規劃。

```r
library(sf)
```

```
Warning: package 'sf' was built under R version 4.3.3

Linking to GEOS 3.11.2, GDAL 3.8.2, PROJ 9.3.1; sf_use_s2() is TRUE
```

```
#
map <- st_read("C:/Users/zach1/Desktop/ /  / ( )(TWD97  )1130807/VILLAGE_NLSC_1130807.shp")
```

Reading layer `VILLAGE_NLSC_1130807' from data source
  `C:\Users\zach1\Desktop\ \ \ ( )(TWD97)1130807\VILLAGE_NLSC_1130807.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 7953 features and 11 fields
Geometry type: MULTIPOLYGON
Dimension: XY
Bounding box: xmin: 114.3593 ymin: 10.37135 xmax: 124.5612 ymax: 26.38528
Geodetic CRS: GCS_TWD97[2020]

```
map <- map[map$COUNTYNAME=="  ",]
map <- map[map$TOWNNAME=="  "|map$TOWNNAME=="  ",]
```

```
#         dataframe
#
candidate_3_data <- data[data$v5=="(3 )",]
num <- data.frame(table(c(candidate_3_data$v2,candidate_3_data$v3)))
num <- num[c(-33,-48,-10),]
colnames(num) <- c("VILLNAME", "Freq")
map <- merge(map,num,by="VILLNAME",all=T)
map[is.na(map)] <- 0
```

```
#
library(ggplot2)
summary(map$Freq)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
 0.000   1.000   2.000   3.509   5.000  23.000
```

```
map_N <- map[map$TOWNNAME=="  ",]
map_MW <- map[map$TOWNNAME=="  ",]
```

由兩區地圖可以看出，三號候選人在西北方有較高的支持度，尤其是在大港里。在東南部分支持者較少。

```
#
coord <- as.data.frame(st_centroid(map))
```

```
Warning: st_centroid assumes attributes are constant over geometries
```

```r
f <- st_coordinates(st_centroid(map))
```

```
Warning: st_centroid assumes attributes are constant over geometries
```

```r
ggplot(data = map) +
  geom_sf(aes(fill = Freq)) +
  scale_fill_gradient(low="white",high="red",name = " (3) ",limits=c(0,23))+
  geom_text(data = coord, aes(x = f[,1], y = f[,2], label = coord$VILLNAME),
           size = 2, color = "black", vjust =0)
```


從單一北區地圖可以看到北區有許多人支持三號候選人。
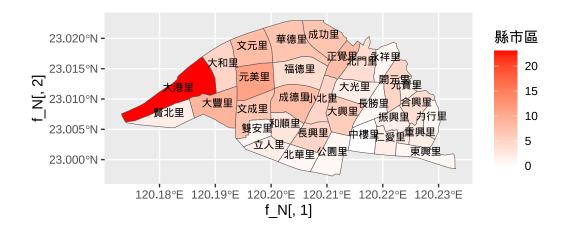
```r
#
coord_N <- as.data.frame(st_centroid(map_N))
```

```
Warning: st_centroid assumes attributes are constant over geometries
```

```
f_N <- st_coordinates(st_centroid(map_N))
```

Warning: st_centroid assumes attributes are constant over geometries

```
ggplot(data = map_N) +
    geom_sf(aes(fill = Freq)) +
    scale_fill_gradient(low="white",high="red",name = "  ",limits=c(0,23))+
    geom_text(data = coord_N, aes(x = f_N[,1], y = f_N[,2], label = coord_N$VILLNAME),
        size = 2.75, color = "black", vjust =0)
```


從中西區地圖來看，在中西區支持三號候選人的人較少，如果想要增加影響力可以考慮往中西區進行宣傳。

```
#
coord_MW <- as.data.frame(st_centroid(map_MW))
```

Warning: st_centroid assumes attributes are constant over geometries

```
f_MW <- st_coordinates(st_centroid(map_MW))
```

```r
ggplot(data = map_MW) +
  geom_sf(aes(fill = Freq)) +
  scale_fill_gradient(low="white",high="red",name = "  ",limits=c(0,23))+
  geom_text(data = coord_MW, aes(x = f_MW[,1], y = f_MW[,2], label = coord_MW$VILLNAME),
            size = 2.75, color = "black", vjust =0)
```