HW1

Kai-Chien Chen

2024-10-03

Table of contents

Question 2

Data Import

```
setwd("C:/Users/user/Desktop/  /4-1(113-1)/stat_consult/HW2")
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(readxl)
library(knitr)
# Read the dataset (adjust the path as necessary)
file_path <- "cleaned_dataset.xlsx"
data <- read_excel(file_path)

# View the first few rows of the data to understand its structure
kable(head(data),caption = "head of data")
```

Table 1: head of data

v1	v2	v3	v4_1	v4_2	v4_3	v4_4	v4_5	v4_6	v4_7	v4_8	v5	v6	v7	v8
中西區	遺漏值	西區警察	無反應	遺漏值	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	無反應	50	以上	女性
中西區	遺漏值	西區警察	無反應	(4號)	(7號)	(10號)	遺漏值	遺漏舊	遺漏舊	遺漏舊	遺漏舊	(7號)	50	初中	男性
中西區	遺漏值	西區警察	(4號)	遺漏值	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	(6號)	50	大學以上	女性
北區	大和里	遺漏值	(3號)	(10號)	遺漏值	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	(7號)	50	初中	女性
北區	成功里	遺漏值	無反應	遺漏值	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	或跳答	50	初中	女性
中西區	遺漏值	永凱里	無反應	遺漏值	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	遺漏舊	(10號)	50	高中	女性

v4 3

```
library(showtext)
```

```
Loading required package: sysfonts
```

```
Loading required package: showtextdb
```

```
# Enable showtext to use system fonts
showtext_auto()

# Filter data related to Candidate 3
candidate_3_data <- data %>%
  filter(grepl("3 ", v4_1) | grepl("3 ", v4_2) | grepl("3 ", v4_3))
```

```r
# Group by region and summarize the support count for Candidate 3
support_by_region <- candidate_3_data %>%
  group_by(v1) %>%
  summarize(count = n())
kable(support_by_region, caption = "Table of support by region")
```

Table 2: Table of support by region

v1	count
中西區	90
北區	319

```r
# Group by age and summarize the support count for Candidate 3
support_by_age <- candidate_3_data %>%
  group_by(v6) %>%
  summarize(count = n())
kable(support_by_age, caption = "Table of support by age")
```

Table 3: Table of support by age

v6	count
20-29歲	6
30-39歲	23
40-49歲	62
50-59歲	100
60歲及以上	216
不知道/拒答	2

```r
# Group by gender and summarize the support count for Candidate 3
support_by_gender <- candidate_3_data %>%
  group_by(v8) %>%
  summarize(count = n())
kable(support_by_gender, caption = "Table of support by gender")
```

Table 4: Table of support by gender

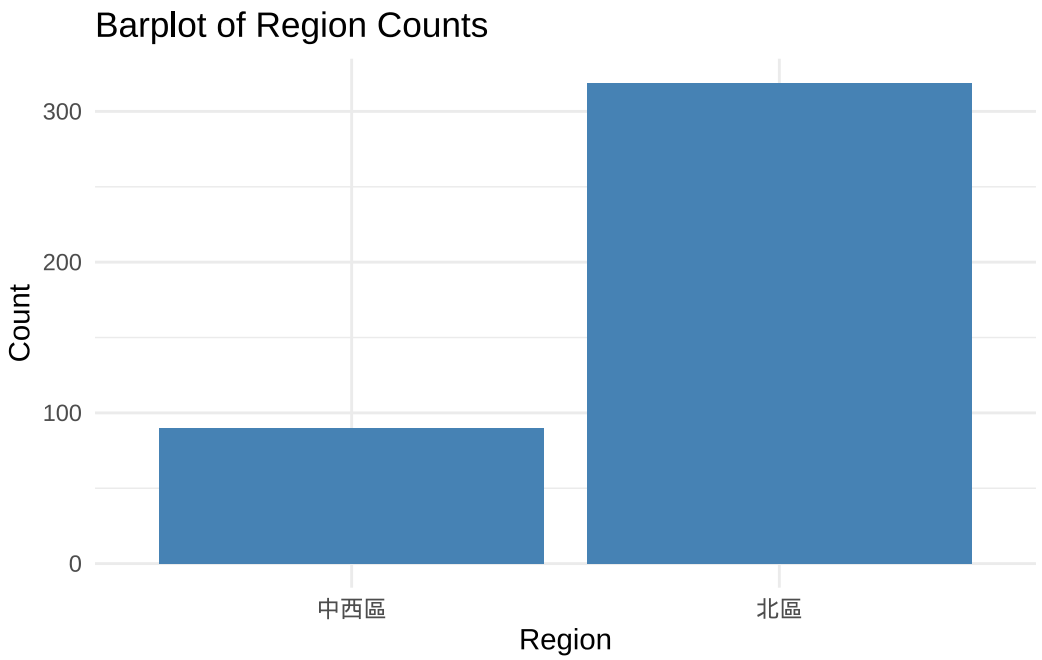
v8	count
女性	219
男性	190

```
# Group by education and summarize the support count for Candidate 3
support_by_education <- candidate_3_data %>%
  group_by(v7) %>%
  summarize(count = n())
kable(support_by_education, caption = "Table of support by education")
```

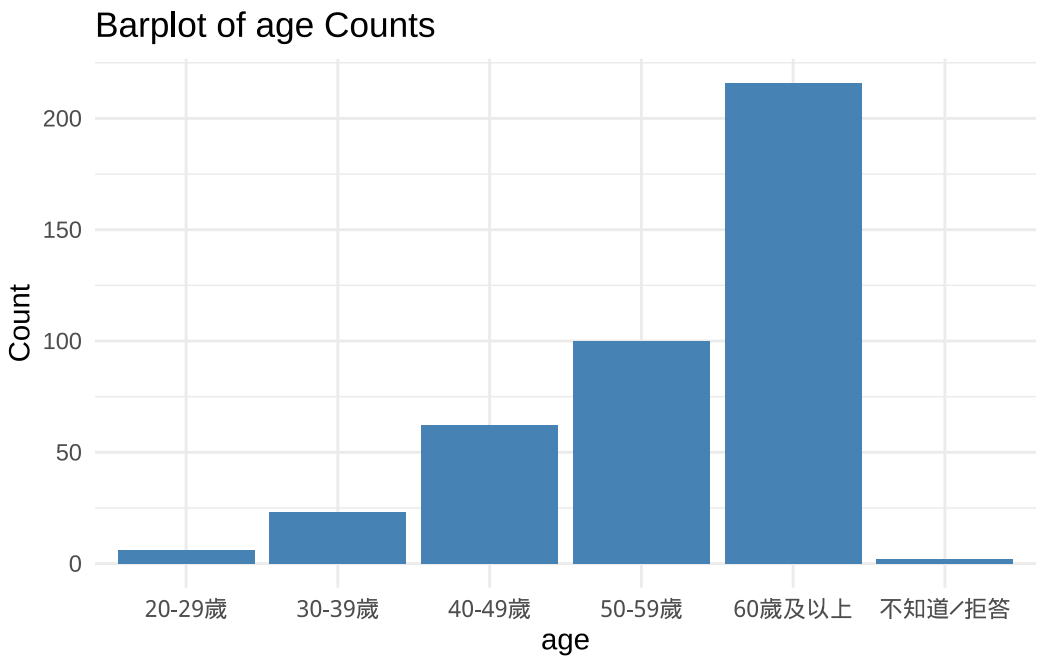
Table 5: Table of support by education

v7	count
初中、國中	40
大學以上	141
專科	53
小學或以下	51
拒答	2
高中、高職	122

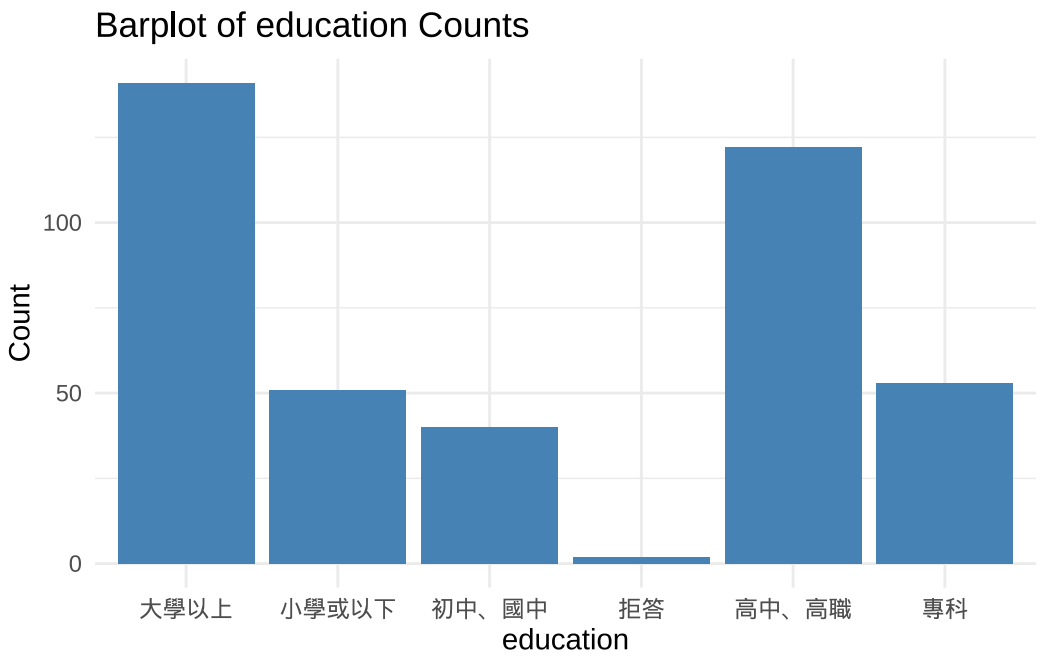
```
#plot
library(ggplot2)
ggplot(support_by_region, aes(x = v1, y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "Region", y = "Count", title = "Barplot of Region Counts") +
  theme_minimal()
```



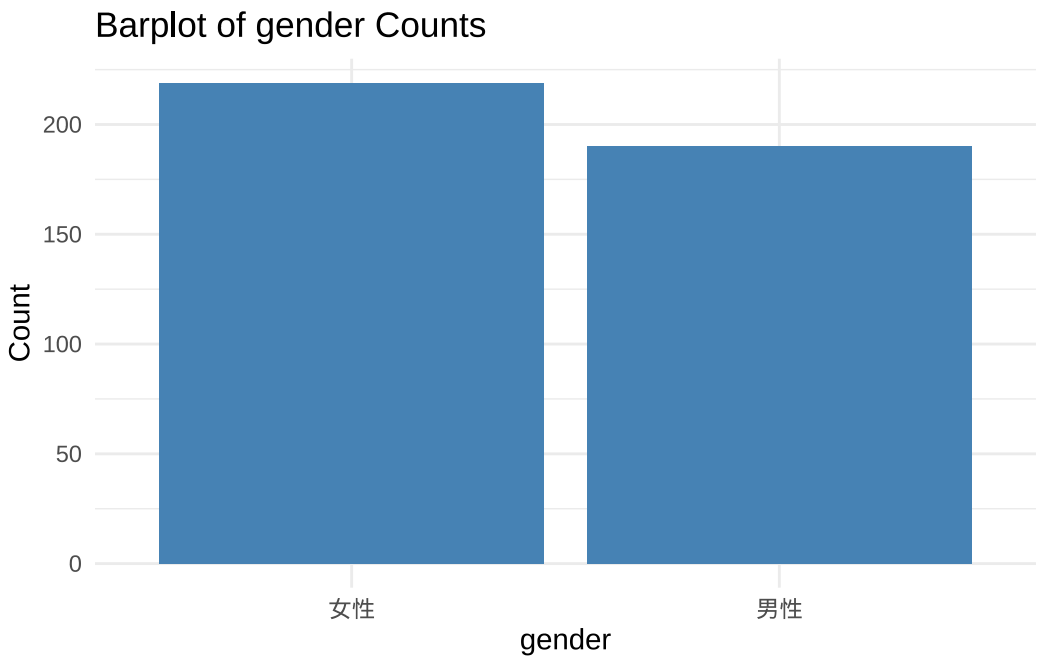

```
ggplot(support_by_age, aes(x = v6, y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "age", y = "Count", title = "Barplot of age Counts") +
  theme_minimal()
```




```
ggplot(support_by_education, aes(x = v7, y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "education", y = "Count", title = "Barplot of education Counts") +
  theme_minimal()
```




```
ggplot(support_by_gender, aes(x = v8, y = count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "gender", y = "Count", title = "Barplot of gender Counts") +
  theme_minimal()
```



從上方之表和圖可以發現知道3號候選人有一些地區上和年齡上的差異。

Chi-squre test

```
####################chi-square test###############
#
gender_region <- table(candidate_3_data$v8, candidate_3_data$v1)

#
kable(gender_region,caption ="     ")
```

Table 6: 性別與地區的交叉表

	中西區	北區
女性	51	168
男性	39	151

```
#      (      )
chi_g_r <- chisq.test(gender_region)

#
print(chi_g_r)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  gender_region
X-squared = 0.30542, df = 1, p-value = 0.5805
```

從上方的chi-square test中可以發現其p-value> 0.05，因此不拒絕虛無假設(地區與性別獨立)，這代表在知道3號候選人的人們中，候選人不需要因為地區之不同後改變針對性別相關議題的策略。

```
###
#
age_region <- table(candidate_3_data$v6, candidate_3_data$v1)


#
kable(age_region,caption = "      ")
```

Table 7: 年齡與地區的交叉表

	中西區	北區
20-29歲	2	4
30-39歲	2	21
40-49歲	8	54
50-59歲	17	83
60歲及以上	61	155
不知道/拒答	0	2

```
#     (     )
chi_a_r <- chisq.test(age_region)
```

```
Warning in chisq.test(age_region): Chi-squared approximation may be incorrect
```

```
#
kable(chi_a_r$expected,caption = "    ")
```

Table 8: 各分類之期望值

	中西區	北區
20-29歲	1.3202934	4.679707
30-39歲	5.0611247	17.938875
40-49歲	13.6430318	48.356968
50-59歲	22.0048900	77.995110
60歲及以上	47.5305623	168.469438
不知道/拒答	0.4400978	1.559902

從Table 8來看，有些地方期望值無大於5，因此重新分類，將20-39歲的定義為年輕人，40-59歲定義為中年人和60歲以上定義為老年人。

```
#
candidate_3_data_1 <- candidate_3_data %>%
  mutate(age_group = case_when(
    v6 %in% c("20-29 ", "30-39 ") ~ "  ",
    v6 %in% c("40-49 ", "50-59 ") ~ "  ",
    v6 == "60   " ~ "  ",
    TRUE ~ v6  #
  ))

#
age_region_combined <- table(candidate_3_data_1$age_group, candidate_3_data_1$v1)

#
kable(age_region_combined,caption = "     ")
```

Table 9: 年齡與地區的交叉表

	中西區	北區
不知道/拒答	0	2
中年人	25	137
年輕人	4	25
老年人	61	155

```
#
chi_a_r_combined <- chisq.test(age_region_combined)
```

```
Warning in chisq.test(age_region_combined): Chi-squared approximation may be
incorrect
```

```
#
print(chi_a_r_combined) #p-value<0.05  H0  3
```

```

    Pearson's Chi-squared test

data:  age_region_combined
X-squared = 10.675, df = 3, p-value = 0.01362
```

```
#
kable(chi_a_r_combined$expected,caption = "    ")
```

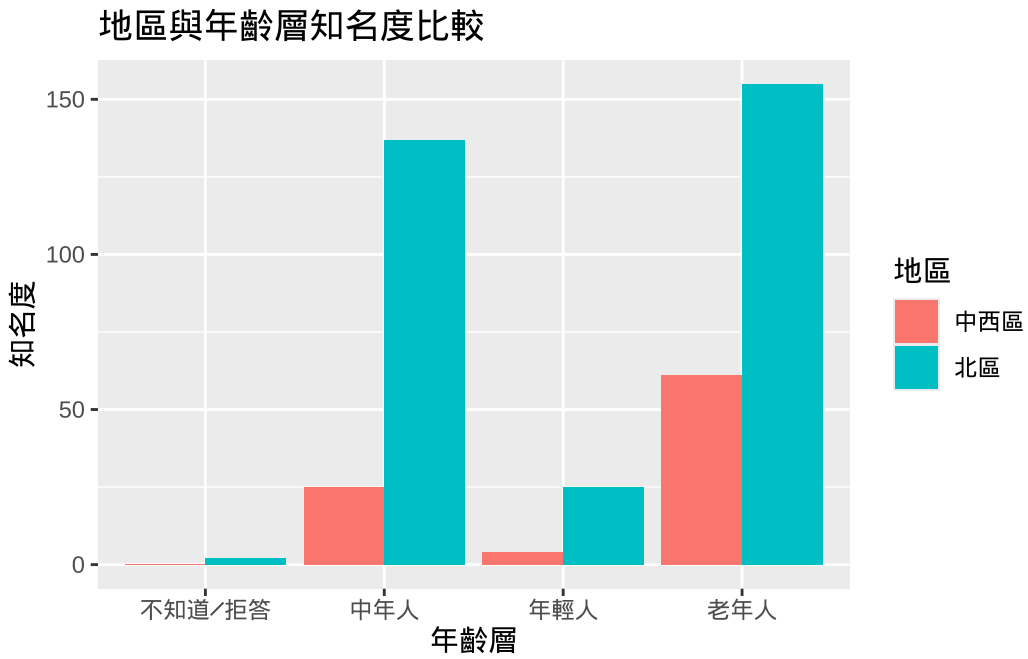
Table 10: 各分類之期望值

	中西區	北區
不知道/拒答	0.4400978	1.559902
中年人	35.6479218	126.352078
年輕人	6.3814181	22.618582
老年人	47.5305623	168.469438

```
library(ggplot2)

#
age_region_df <- as.data.frame(age_region_combined)
```

```r
#
ggplot(age_region_df, aes(Var1, Freq, fill = Var2)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "      ", x = "  ", y = "  ", fill = " ")
```



因為不知道/拒答的比例不高，將其省略，而從卡方檢定、交叉表和各分類之期望 值可以得到以下結論:

- **卡方檢定**:
  - 從結果可知其拒絕虛無假設，代表年齡層和地區是有相關的。
- **中西區**：
  - **老年人**：實際支持 61 人，明顯高於期望支持數 47.53。這說明 60 歲及以上的年齡層在中西區的支持較強，應著重針對這個群體進行宣傳。
  - **中年人**：實際支持 25 人，低於期望支持數 35.65。這個年齡層的支持度低於預期，可能需要加強宣傳。
  - **年輕人**：實際支持 4 人，略低於期望支持數 6.38。這個年齡層的支持者數量較少，但差異不大，應視資源投入情況決定是否加強針對這個群體的宣傳。
- **北區**：

- **老年人**：實際支持 155 人，略低於期望支持數 168.47。這個年齡層支持度仍然相當高，但有些低於期望，可能需要進一步鞏固這個群體的支持。
- **中年人**：實際支持 137 人，高於期望支持數 126.35，這表明這個年齡層對候選人的支持較高，可以維持此群體的宣傳。
- **年輕人歲**：實際支持 25 人，高於期望支持數 22.62，這表明這個年齡層的支持略高於期望，可以考慮繼續關注。

```
###
#
gender_age <- table(candidate_3_data$v8, candidate_3_data$v6)

#
kable(gender_age,caption = "     ")
```

Table 11: 性別與地區之交叉表

	20-29歲	30-39歲	40-49歲	50-59歲	60歲及以上	不知道/拒答
女性	4	7	32	54	120	2
男性	2	16	30	46	96	0

```
#     (     )
chi_g_a <- chisq.test(gender_age)
```

```
Warning in chisq.test(gender_age): Chi-squared approximation may be incorrect
```

```
#
kable(chi_g_a$expected,caption = "     ")
```

Table 12: 性別與地區之期望值

	20-29歲	30-39歲	40-49歲	50-59歲	60歲及以上	不知道/拒答
女性	3.212714	12.3154	33.19804	53.54523	115.6577	1.0709046
男性	2.787286	10.6846	28.80196	46.45477	100.3423	0.9290954

```
#
age_gender_combined <- table(candidate_3_data_1$age_group, candidate_3_data_1$v8)

#
kable(age_gender_combined,caption = "     ")
```

Table 13: 性別與地區之交叉表

	女性	男性
不知道/拒答	2	0
中年人	86	76
年輕人	11	18
老年人	120	96

```
#
chi_a_g_combined <- chisq.test(age_gender_combined)
```

```
Warning in chisq.test(age_gender_combined): Chi-squared approximation may be
incorrect
```

```
#
print(chi_a_g_combined)
```

```

    Pearson's Chi-squared test

data:  age_gender_combined
X-squared = 4.9422, df = 3, p-value = 0.1761
```

```
#
kable(chi_a_g_combined$expected,caption = "       ")
```

Table 14: 年齡與性別的交叉表之期望值

	女性	男性
不知道/拒答	1.070905	0.9290954
中年人	86.743276	75.2567237

	女性	男性
年輕人	15.528117	13.4718826
老年人	115.657702	100.3422983

做與地區和年齡層的卡方檢定的一樣動作，而從上方的表和檢定可知其檢定結果不拒絕虛無假設，也就是說在知道3號候選人的樣本中性別和年齡層為獨立的。