Advanced Methods of Political Analysis

Problem Set 2 Answers

Sanghoon Park

2024-06-18

Load packages

```
library(ezpickr);library(knitr);library(patchwork);library(tidyverse)
```

Fundamentals of statistics and statistical tests

```
ggplot2::theme_set(theme_bw() + theme(legend.position = "bottom",
                                      legend.title = element_blank()))
data ← ezpickr::pick("W2Dataset.csv")
```

1. Create a table that gives the frequencies of observations across joint values of `v2elembcap_ord` and `v2elembaut_ord`.
   Note: you may need to set the factor levels for your variable first.

```
class(data$v2elembcap_ord);table(data$v2elembcap_ord)
class(data$v2elembaut_ord);table(data$v2elembaut_ord)

# According to the codebook, set the character values as factors
data ← data ▷
  mutate(
    v2elembcap_fct = factor(v2elembcap_ord,
                            levels = c("No", "Not really",
                             ↪  "Ambiguous", "Almost", "Yes"),
                            ordered = T),
    v2elembaut_fct = factor(v2elembaut_ord,
                            levels = c("No", "Somewhat",
                             ↪  "Ambiguous" , "Almost", "Yes"),
                            ordered = T))

# Frequencies of observations across joint values
```

```
freq.tab ← table(data$v2elembcap_fct, data$v2elembaut_fct)
freq.tab ▷
  kable(digits = 3,
        caption = "Frequencies of obervations across joint
        ↪  value")
```

2. Calculate the marginal probabilities for each value of `v2elembcap_ord` and `v2elembaut_ord`.
What is the probability that EMB capacity is greater than two ("Ambiguous")?
What is the probability that EMB autonomy is less than two ("Ambiguous")?

```
# Note: For convenience of reading, rounded up to deciaml(2)
# marginal probabilities of embcap
round(apply(prop.table(freq.tab), 1, sum), 2)

    # marginal probabilities of embaut
round(apply(prop.table(freq.tab), 2, sum), 2)
```

   - The probability that EMP capacity is greater than two ("Ambiguous") is about 0.4 ("Almost", "Yes"). Also, the probability that EMP autonomy is less than two ("Ambigous") is about 0.63 ("No", "Somewhat").

3. Plot the probability density function for `v2elembcap_ord`.

```
data ▷ drop_na(v2elembcap_fct) ▷
  ggplot(aes(as.numeric(v2elembcap_fct))) +
  geom_density(fill = futurevisions::futurevisions("mars")[2],
               color = futurevisions::futurevisions("mars")[2],
               alpha = 0.4) +
  scale_x_continuous(breaks = c(1:5),
                     labels = c(levels(data$v2elembcap_fct))) +
  labs(x = "EMB Capability") +
  theme(axis.title.x = element_text(vjust = -1))
```

4. Calculate the conditional probabilities for each value of `v2elembaut_ord` and the conditional expectation of `v2elembaut_ord` when `v2elembcap_ord`=0 ("No").
How strong should we expect EMB autonomy to be when EMB capacity is at zero?

```
table(data$v2elembcap_fct,
      data$v2elembaut_fct)
table(data$v2elembcap_fct,
      data$v2elembaut_fct)[1,]/sum(table(data$v2elembcap_fct,
                                         data$v2elembaut_fct)[1,])
```

```r
data <- data |> mutate(
  v2elembaut_nu = as.numeric(v2elembaut_fct) - 1,
  v2elembcap_nu = as.numeric(v2elembcap_fct) - 1
)
data |>
  summarize(
  mean = mean(as.numeric(v2elembaut_nu), na.rm = T)
)

data |> dplyr::filter(v2elembcap_nu==0) |>
  summarize(
    mean = mean(as.numeric(v2elembaut_nu), na.rm = T))
```

5. Using values from the frequency table, calculate the covariance of `v2elembcap_ord` and `v2elembaut_ord`. Show your work and verify the answer using the `cov()` function.
   Note: You may need to subset non-missing values first.

```r
# In freq.table, missings are omitted.
expect.cap <-
  sum((c(0:4) * margin.table(freq.tab, 1)) /
        sum(margin.table(freq.tab, 1)))
expect.aut <-
  sum((c(0:4) * margin.table(freq.tab, 2)) /
        sum(margin.table(freq.tab, 2)))

comb <- cbind(0:4) %*% rbind(0:4)
expect.capaut <- sum((freq.tab/sum(freq.tab)) * comb)

# Calculate!
cov.autocap <- expect.capaut - (expect.cap * expect.aut)

# Verify it. Compelete.obs omits the missings.
cov.verify <- cov(as.numeric(data$v2elembcap_fct),
                  as.numeric(data$v2elembaut_fct),
                  use = "complete.obs")

# Small differences come from the degree of freedom.
round(cov.verify, 3) == round(cov.autocap, 3)
```

6. Calculate the correlation coefficient for `v2elembcap_ord` and `v2elembaut_ord`. Show your work and verify the answer using the `cor()` function.
   How strong is the relationship between election management body capacity and election management body autonomy?

- $\text{Cor}(X, Y) = \text{Cov}(X, Y)/\sigma_x \times \sigma_y$

7. Treat the variable `v2elembcap` in 2019 as a complete population and calculate the probability of obtaining a value less than or equal to 1 using the z-score table provided here: https://www.math.arizona.edu/~rsims/ma464/standardnormaltable.pdf.
   What is the probability of obtaining a value greater than or equal to 1?

```
# It is a problem to check the sampling error.
summarized ← data ▷
  drop_na(v2elembcap, v2elembaut) ▷
  dplyr::filter(year == 2019) ▷
  summarize(
    sd = sd(v2elembcap),
    mean = mean(v2elembcap),
    zscore = (1 - mean) / sd
  )

# Pr(value ⩽ 1)
pnorm(summarized$zscore, lower.tail = T)

# Pr(value ⩾ 1)
pnorm(summarized$zscore, lower.tail = F)
```

8. Treat values of `v2elembcap` for countries in Latin America and the Caribbean as a sample taken from the population (from the previous question). Construct the 99% confidence interval for the population mean based on the sample.
   (Use the table of t distribution critical values provided here: https://www.stat.tamu.edu/~lzhou/stat302/T-Table.pdf)

```
data ▷
  dplyr::filter(e_regionpol_6C == "L. America and the
↪  Carribean") ▷
  dplyr::filter(year == 2019) ▷
  drop_na(v2elembcap) ▷
  summarize(
    mean = mean(v2elembcap),
    sd = sd(v2elembcap),
    n = n(),
    se = sd/sqrt(n-1),
    tstat = (mean - mean(data$v2elembcap, na.rm = T)) / se,
    ll = mean - (2.576 * se), # lower bound, z* for 99% = 2.576
    ul = mean + (2.576 * se), # upper bound
  ) ▷ kable(
```

```
    caption = "Descriptive statistics of sample")
```

- The results estimate the population mean using the sample of countries in Latin America and the Carribean with 99% confidence interval.
- Although I compute the 99% confidence interval based on the point estimate (sample mean), expecting the mean of population will fall between the boundaries, the population mean is likely to be different from the sample mean (t-statistics ≈ 6.42). We can consider the the difference between the sample mean and the population mean is much greater than the difference which can be driven by sampling (difference is less likely caused by chance).

9. Plot the probability density functions for `v2elembcap_ord` for observations in which there was and was not a civil war (`e_civil_war`).

```
data ▷ drop_na(e_civil_war) ▷ mutate(
  civilwar = ifelse(e_civil_war==1, "Civil war",
                    "Not civil war")
  ) ▷
  ggplot(aes(as.numeric(v2elembcap_fct),
             color = as.factor(civilwar),
             fill = as.factor(civilwar))) +
  geom_density(alpha = 0.3) +
  scale_fill_manual(values =
   ↪  futurevisions::futurevisions("earth")) +
  scale_color_manual(values =
   ↪  futurevisions::futurevisions("earth")) +
  labs(x = "v2elembcap_ord")
```

10. Perform a t-test for subsamples of observations in which there was and was not a civil war (*e_civil_war*). Is there a difference in election management body capacity between countries with and without civil wars? How do you know?

```
Warsub ←
  data ▷ drop_na(e_civil_war) ▷
  group_by(e_civil_war) ▷
  summarize(
    mean = mean(v2elembcap, na.rm = T),
    sd = sd(v2elembcap, na.rm = T),
    n = n()
  )

tofWar ← # Not war - War
Warsub ▷
```

```
  mutate(
    t = (mean[1]- mean[2]) /
      sqrt(sd[1]^2/n[1] + sd[2]^2/n[2]))
```

- There is a difference in election management body capacity between countries with and without civil wars.
- The t-statistics tell that the difference is more likely to driven by actual difference in election management body capacity between countries between with and without civil wars than by chance.
- About 18.89, t-statics imply that the difference in EMB capacity in countries between with and without civil war is eighteen times greater than the possible variation cased by sampling. Typically, scholars set the 1.96 of t-statistics as a 5% significance level. 18.89 is much greater than 1.96, meaning the differences is challenging to be rejected.