Problem Set 1

Costa Stasinopoulos

2025-01-23

R Markdown

1. This data set is for South Africa. There are 1580 observations for that same number of respondents. The interviews took place from 26 November 2022 to 17 December 2022. The data set contains 390 variables overall.
2. The table below contains basic demographic information on the respondents in the survey.Q1 refers to respondent age. From this variable, we can see that the average age of respondents is 39.93 years old, the median age is 38, and the minimum and maximum are 18 and 89, respectively. Q100 identifies the gender of the respondent, with 1 signifying man and 2 signifying woman. The minimum and maximum values of this variable are not very helpful, but the mean of 1.497 indicates that just over 50% of the respondents are men (because there are more men the mean is weighted slightly more toward 1 than 2).Q2 stands for the "language spoken at home" variable, while Q84A is ethnic community, cultural group or tribe. These are both important demographic variables, but they are categorical variables with different numbers assigned to different languages and ethnic groups, meaning that the summary statistics provided in the table to not provide useful information.

```
#Q2 is what is the primary language you speak at home
#Q84A is Ethnic community, cultural group or tribe
#Gender of respondent is Q100, 1=Man, 2=Woman
data %>%
  select(RESPNO, Q1, Q2, Q84A, Q100) %>%
  summary() %>%
  kable(caption = "Summary of Selected Variables")
```

Table 1: Summary of Selected Variables

RESPNO	Q1	Q2	Q84A	Q100
Length:1580	Min. :18.00	Min. : 1.0	Min. : 700.0	Min. :1.000
Class :character	1st Qu.:27.00	1st Qu.: 700.0	1st Qu.: 703.0	1st Qu.:1.000
Mode :character	Median :38.00	Median : 703.0	Median : 706.0	Median :1.000
NA	Mean :39.93	Mean : 574.7	Mean : 783.1	Mean :1.497
NA	3rd Qu.:50.00	3rd Qu.: 706.0	3rd Qu.: 710.0	3rd Qu.:2.000
NA	Max. :89.00	Max. :9998.0	Max. :9998.0	Max. :2.000

1.The "real" values for the variable Q78A are 1-5, with one reflecting an extremely negative view of Chinese influence, 5 reflecting an extremely positive view, and 2-3 being more moderate views on the spectrum. We see that responses have a local peak for 1, decline for 2-3, and are highest (and similar in frequency) for 4-5, indicating a high frequency of support for Chinese influence. Answers of 8 or 9 indicated refusals to answer or an answer of "I don't know", respectively.

```
#question 3
#prop.table takes a table and turns it into frequencies (percentage of total for each response type)
china_support <- prop.table(table(data$Q78A))

# Convert the table to a data frame and rename columns, this is necessary because it seems like there i
china_support_df <- as.data.frame(china_support) %>%
  setNames(c("Outlook on Chinese Influence (Q78A)", "Relative Frequency"))

#output table (which is actually a data frame now)
kable(china_support_df)
```

Outlook on Chinese Influence (Q78A)	Relative Frequency
1	0.1164557
2	0.0791139
3	0.0987342
4	0.1639241
5	0.1835443
8	0.0094937
9	0.3487342

1.The variable Q78B is the same as the variable in the previous question but in regard to US influence. The relative frequency table shows that relatively fewer respondents held strongly or moderately negative views of US influence and more had neutral to somewhat positive views of this influence. However, a substantially lower proportion of respondents were strongly postive about US influence vs Chinese influence (.127 vs .184). A significantly larger proportino of respondents answered "I don't know" in regard to US influence than in regard to Chinese influence (.421 vs .349.)

```
#question 4
#prop.table takes a table and turns it into frequencies (percentage of total for each response type)
usa_support <- prop.table(table(data$Q78B))

# Convert the table to a data frame and rename columns, this is necessary because it seems like there i
usa_support_df <- as.data.frame(usa_support) %>%
  setNames(c("Outlook on Chinese Influence (Q78A)", "Relative Frequency"))

#output table (which is actually a data frame now)
kable(usa_support_df)
```

Outlook on Chinese Influence (Q78A)	Relative Frequency
1	0.0702532
2	0.0620253
3	0.1291139
4	0.1791139
5	0.1272152
8	0.0107595
9	0.4215190

1. We can see in the t-test results that the mean of the first variable (Q78A) is about .45 lower than that of the second variable, Q74B. Additionally, the p-value is extremely low - effectively zero - meaning

that this two-tailed t-test is highly significant and it is very likely that the means of the two variables are different (and that the Chinese influence variable mean is lower). -.573 to -.337 provides the 95 percent confidence interval range for the actual difference in average mean value between Q74A and Q74B. This t-test shows that, with the non-value answers removed, it is very likely that on average responses regarding Chinese influence are more negative than those regarding US influence (most likely by around .455 points on a 5-point scale).

```
#question 5
data %>%
  mutate(
    across(Q78A:Q78B, # vars to work on
           ~ if_else(.x %in% 1:5, .x, NA) # function applied to vars
    )
  )
```

```
## # A tibble: 1,580 x 390
##    RESPNO  URBRUR    REGION        EA_SVC_A  EA_SVC_B EA_SVC_C EA_SVC_D EA_SVC_E
##    <chr>   <dbl+lbl> <dbl+lbl>     <dbl+lbl> <dbl+lb> <dbl+lb> <dbl+lb> <dbl+lb>
##  1 SAF0001 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  0 [No]   0 [No]
##  2 SAF0002 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  0 [No]   0 [No]
##  3 SAF0003 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  1 [Yes]  0 [No]
##  4 SAF0004 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  0 [No]   0 [No]
##  5 SAF0005 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  1 [Yes]  0 [No]
##  6 SAF0006 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  1 [Yes]  0 [No]
##  7 SAF0007 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  1 [Yes]  0 [No]
##  8 SAF0008 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  0 [No]   0 [No]
##  9 SAF0009 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  1 [Yes]  0 [No]
## 10 SAF0010 1 [Urban] 702 [Gauteng] 1 [Yes]   1 [Yes]  1 [Yes]  1 [Yes]  0 [No]
## # i 1,570 more rows
## # i 382 more variables: EA_FAC_A <dbl+lbl>, EA_FAC_B <dbl+lbl>,
## #   EA_FAC_C <dbl+lbl>, EA_FAC_D <dbl+lbl>, EA_FAC_E <dbl+lbl>,
## #   EA_FAC_F <dbl+lbl>, EA_FAC_F2 <dbl+lbl>, EA_FAC_G <dbl+lbl>,
## #   EA_SEC_A <dbl+lbl>, EA_SEC_B <dbl+lbl>, EA_SEC_C <dbl+lbl>,
## #   EA_SEC_D <dbl+lbl>, EA_SEC_E <dbl+lbl>, EA_ROAD_A <dbl+lbl>,
## #   EA_ROAD_B <dbl+lbl>, EA_ROAD_C <dbl+lbl>, NOCALL_1 <dbl+lbl>, ...
```

```
t.test(data$Q78A, data$Q78B, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data$Q78A and data$Q78B
## t = -7.5774, df = 1579, p-value = 5.972e-14
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.5728591 -0.3372675
## sample estimates:
## mean difference
##      -0.4550633
```