Problem Set 1

Nicholas Rangel

2025-01-28

My Responses

1. For this problem set, I chose Angola. There are 1200 respondents in the survey and the interviews were conducted between February and March of 2022.

```
## commands used:
library(haven)
data <- read_sav("/Users/nicholasrangel/Data Analysis Class/Week 3/AngolaAfrobarometer.sav")
summary(data$RESPNO)
```

```
##    Length     Class      Mode
##      1200 character character
```

```
summary(data$DATEINTR)
```

```
##          Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-02-09" "2022-02-12" "2022-02-19" "2022-02-18" "2022-02-23" "2022-03-08"
```

2. The median age of the respondents is 30, while the mean is 34.29. There is approximately an equal amount of male and female respondents, as the mean is 1.503 (where 1 indicates a male and 2 indicates a female). Regarding language spoken, 970 of the 1200 respondents spoke Portuguese, 71 respondents spoke Umbundu, 67 spoke Chokwe, and a combined 92 respondents spoke 8 other languages. The number of adults in the respondent's household had a mean of 2.74.

```
## commands used:
summary(data$Q1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   23.00   30.00   34.29   40.00  998.00
```

```
summary(data$THISINT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.503   2.000   2.000
```

```
table(data$Q2)
```

```
##
##    3 1750 1751 1752 1753 1754 1755 1756 1757 1758 9995
##  970   71   28    5   67   19   19    5   14    1    1
```

```
summary(data$ADULT_CT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    2.00    2.74    3.00   13.00
```

```
## Bonus: Table showing just the means of Q1, THISINT, ADULT_CT, and median of Q2
mean_table <- tibble(
  Variable = c("Q1", "THISINT", "Q2", "ADULT_CT"),
  Mean = c(
    mean(data$Q1, na.rm = TRUE),
    mean(data$THISINT, na.rm = TRUE),
    median(data$Q2, na.rm = TRUE),
    mean(data$ADULT_CT, na.rm = TRUE)
  )
)
```

3. Q78A in the dataset refers to the following question: Do you think that the economic and political influence of each of the following countries on Angola is mostly positive, mostly negative, or haven't you heard enough to say: China?

The values have the following meanings: 1=Very negative, 2=Somewhat negative, 3=Neither positive nor negative, 4=Somewhat positive, 5=Very positive, 8=Refused, 9=Don't know, -1=Missing.

Creating a frequency table with all of the aforementioned values would not accurately represent people's opinions, so I filtered out respondents who refused to answer (8), those who did not have an answer (9), and missing responses (-1), keeping only the 1-5 responses.

Findings show that on average, respondents see Chinese economic and political influence as somewhat positive, as the median is 4, and the mean is 3.6. This eliminated 334 respondents who did not have an answer (9) and 31 respondents who refused to answer (8). It is important to note that the "don't know" category (9) received the most votes, indicating that there is a large amount of people who did not feel confident enough in providing a solid answer.

```
## commands used:
summary(data$Q78A)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   5.000   5.192   9.000   9.000
```

```
library(dplyr)
filteredQ78a <- data %>%
  filter(Q78A != 8, Q78A != 9, Q78A != -1) ## excludes responses (8), (9), (-1)

summary(filteredQ78a$Q78A) ## shows Min, Mean, Max for filtered responses
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   3.564   5.000   5.000
```

```
table(filteredQ78a$Q78A) ## shows exact values each category received as votes
```

```
##
##   1   2   3   4   5
## 135  84  88 231 297
```

```
table(data$Q78A) ## shows exact values of the excluded responses
```

```
##
##   1   2   3   4   5   8   9
## 135  84  88 231 297  31 334
```

```
filteredQ78a %>%
  count(Q78A) %>%
  mutate(percentage = n / sum(n) * 100) ## relative frequency table for Q78A responses excluding answer
```

```
## # A tibble: 5 x 3
##   Q78A                               n percentage
##   <dbl+lbl>                      <int>      <dbl>
## 1 1 [Muito negativa]               135       16.2
## 2 2 [De algum modo negativa]        84       10.1
## 3 3 [Nem positiva nem negativa]     88       10.5
## 4 4 [De algum modo positiva]       231       27.7
## 5 5 [Muito positiva]               297       35.6
```

```
data %>%
  count(Q78A) %>%
  mutate(percentage = n / sum(n) * 100) ## relative frequency table for Q78A responses including the pr
```

```
## # A tibble: 7 x 3
##   Q78A                               n percentage
##   <dbl+lbl>                      <int>      <dbl>
## 1 1 [Muito negativa]               135      11.2
## 2 2 [De algum modo negativa]        84       7
## 3 3 [Nem positiva nem negativa]     88       7.33
## 4 4 [De algum modo positiva]       231      19.2
## 5 5 [Muito positiva]               297      24.8
## 6 8 [Recusou]                       31       2.58
## 7 9 [Não sabe]                     334      27.8
```

4. Q78B in the dataset refers to the following question: Do you think that the economic and political influence of each of the following countries on Angola is mostly positive, mostly negative, or haven't you heard enough to say: United States?

The values have the following meanings: 1=Very negative, 2=Somewhat negative, 3=Neither positive nor negative, 4=Somewhat positive, 5=Very positive, 8=Refused, 9=Don't know, -1=Missing.

Creating a frequency table with all of the aforementioned values would not accurately represent people's opinions, so I filtered out respondents who refused to answer (8), those who did not have an answer (9), and missing responses (-1), keeping only the 1-5 responses.

Findings show that on average, respondents see American economic and political influence as somewhat positive, as the median is 4, and the mean is 3.8. This eliminated 407 respondents who did not have an answer (9) and 34 respondents who refused to answer (8). It is important to note that the "don't know" category (9) received the most votes, indicating that there is a large amount of people who did not feel confident enough in providing a solid answer.

```
filteredQ78b <- data %>%
  filter(Q78B != 8, Q78B != 9, Q78B != -1) ## excludes responses (8), (9), (-1)

summary(filteredQ78b$Q78B) ## shows Min, Mean, Max for filtered responses
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.773   5.000   5.000
```

```
table(filteredQ78b$Q78B) ## shows exact values each category received as votes
```

```
##
##   1   2   3   4   5
##  78  59  98 246 278
```

```
table(data$Q78B) ## shows exact values of the excluded responses
```

```
##
##   1   2   3   4   5   8   9
##  78  59  98 246 278  34 407
```

```
filteredQ78b %>%
  count(Q78B) %>%
  mutate(percentage = n / sum(n) * 100) ## relative frequency table for Q78B responses excluding answer
```

```
## # A tibble: 5 x 3
##   Q78B                                n percentage
##   <dbl+lbl>                       <int>      <dbl>
## 1 1 [Muito negativa]                 78      10.3
## 2 2 [De algum modo negativa]         59       7.77
## 3 3 [Nem positiva nem negativa]      98      12.9
## 4 4 [De algum modo positiva]        246      32.4
## 5 5 [Muito positiva]                278      36.6
```

```
data %>%
  count(Q78B) %>%
  mutate(percentage = n / sum(n) * 100) ## relative frequency table for Q78B responses including the pr
```

```
## # A tibble: 7 x 3
##   Q78B                                n percentage
##   <dbl+lbl>                       <int>      <dbl>
## 1 1 [Muito negativa]                 78       6.5
## 2 2 [De algum modo negativa]       59       4.92
## 3 3 [Nem positiva nem negativa]    98       8.17
## 4 4 [De algum modo positiva]      246      20.5
## 5 5 [Muito positiva]              278      23.2
## 6 8 [Recusou]                      34       2.83
## 7 9 [Não sabe]                    407      33.9
```

5. After using the cleaning up the data to explude dk/na and refusals, I conducted a paired t-test to evaluate the difference between the two perceptions (US and China). The results show that the t-stat is -4.21, meaning that there is generally a strong difference between the two. The p-value being small (-0.00087) further emphasizes this as being statistically significant.

```
data1 =
  data %>%
  mutate(
    across(
      Q78A:Q78B,
      ~ if_else(.x %in% 1:5, .x, NA))) ## this cleans the data of dk/na and refusals

t.test(data1$Q78A, data1$Q78B, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data1$Q78A and data1$Q78B
## t = -4.21, df = 736, p-value = 2.87e-05
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.2964462 -0.1078958
## sample estimates:
## mean difference
##       -0.202171
```