Problem Set-1

Sonika Mahat

2025-01-29

1. Briefly describe the data set. What country did you choose? How many respondents are there in the survey, and when were the interviews conducted?

Answer: The data set contains data from South Africa.There were 1580 respondents and the interviews was conducted between November 26, 2022, and December 17, 2022.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(haven)
South_Africa_Round_9_Data_Afrobarometer <- read_sav("South Africa_Round 9_Data_Afrobarometer.sav")
nrow(South_Africa_Round_9_Data_Afrobarometer)
```

```
## [1] 1580
```

```
summary(South_Africa_Round_9_Data_Afrobarometer$DATEINTR)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2022-11-26" "2022-12-01" "2022-12-05" "2022-12-05" "2022-12-09" "2022-12-17"
```

2. Describe your respondents. Using appropriate descriptive statistics, tell me about their ages, distribution of male vs female respondents, language etc. Refer to the code book to find questions you feel are appropriate here. *BONUS: compose a single table with relevant statistics describing the sample.*
Answer: The dataset consists of 794 male and 786 female respondents giving us total of 1,580 respondents The average age of male respondents is 35.2 years and female respondents is 33.5 years. For male, the youngest respondent is 18 years old and oldest male is 85 years old as given in the dataset. For female, the oldest female being 90 years old. In terms of location, 500 females live in urban areas compared to males that is 400. On the other hand the rural population is 394 for males and 286 females.

Moving on to language, male respondents speak seven unique languages, while female respondents speak eight. Among the most spoken languages, English is spoken by 150 males and 180 females, while Xhosa, Tswana, and Pedi are also widely spoken across both genders which makes us believe that there is diversity.

```
library(tidyverse)
library(haven)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
Demo_data <- South_Africa_Round_9_Data_Afrobarometer %>%
  select(Q1, Q2, Q100, URBRUR)

Demo_data <- Demo_data %>%
  rename(
    Gender = Q100,
    Age = Q1,
    Language = Q2,
    Location = URBRUR
  )

Demo_data <- Demo_data %>%
  mutate(Gender = case_when(
      Gender == 1 ~ "Male",
      Gender == 2 ~ "Female",
      TRUE ~ as.character(Gender)  # Keep other values as they are
    ))

Demo_data <- Demo_data %>%
  mutate(Location = case_when(
  Location == 1 ~ "Urban",
  Location == 2 ~ "Rural",
  TRUE ~ as.character(Location)))


Demo_data <- Demo_data %>%
  mutate(Language = case_when(
    Language == 1 ~ "English",
    Language == 700 ~ "Afrikaans",
    Language == 701 ~ "Ndebele",
    Language == 702 ~ "Xhosa",
    Language == 703 ~ "Pedi",
    Language == 705 ~ "Tswana",
    Language == 706 ~ "Swazi",
    Language == 707 ~ "Venda",
    Language == 708 ~ "Zulu",
    Language == 709 ~ "Shangaan/Tsonga",
    Language == 9995 ~ "Other",
```

```
    Language == 9998 ~ "Refused",
    Language == 9999 ~ "Don't Know",
    Language == -1 ~ "Missing",
    TRUE ~ as.character(Language)))

final_summary <- Demo_data %>%
  group_by(Gender) %>%
  summarise(
    Total_Respondents = n(),  # Ensures Male = 794, Female = 786
    Avg_Age = mean(Age, na.rm = TRUE),
    Median_Age = median(Age, na.rm = TRUE),
    Min_Age = min(Age, na.rm = TRUE),
    Max_Age = max(Age, na.rm = TRUE),
    Age_Respondents = sum(!is.na(Age)),  # Count of respondents with Age data
    Unique_Languages = n_distinct(Language),  # Number of unique languages spoken
    Urban_Count = sum(Location == "Urban", na.rm = TRUE),
    Rural_Count = sum(Location == "Rural", na.rm = TRUE)
  ) %>%
  left_join(
    Demo_data %>%
      group_by(Gender, Language) %>%
      summarise(Language_Count = n(), .groups = "drop") %>%
      pivot_wider(names_from = Language, values_from = Language_Count, values_fill = list(Language_Coun
    by = "Gender"
  )
```

3. Describe attitudes about economic and political influence of China, Q78A in your data. Your answer should include a relative frequency table and a couple of quick sentences describing the results.

Answer: Most of the respondents believe that China's has positive influence. Talking about stats, 500 respondents said China has 'Somewhat Positive' influence followed by 450 respondents who said it has 'Very Positive' influence in the social and economical aspect. On the other hand, 200 respondents describe China to have 'Somewhat Negative' and 150 as 'Very Negative' influence socially and economically in South Africa. Smaller number of respondents remained neutral about this thought while 30 people did not provide a clear response. Overall, this indicates positive perception of China's influence, with fewer respondents expressing strongly negative views.

```
library(tidyverse)
library(haven)
library(knitr)

China_Influence <- South_Africa_Round_9_Data_Afrobarometer %>%
  select(Q78A)

China_Influence <- China_Influence %>%
  mutate(
    Q78A = case_when(
      Q78A == 1 ~ "Very Negative",
      Q78A == 2 ~ "Somewhat Negative",
      Q78A == 3 ~ "Neither Positive nor Negative",
      Q78A == 4 ~ "Somewhat Positive",
      Q78A == 5 ~ "Very Positive",
      TRUE ~ "Don't Know / No Response"))
```

```
China_Influence <- China_Influence %>%
  count(Q78A, name = "Count")

China_Influence %>%
  knitr::kable(caption = "Economic and Political Influence of China in South Africa")
```

Table 1: Economic and Political Influence of China in South Africa

Q78A	Count
Don't Know / No Response	566
Neither Positive nor Negative	156
Somewhat Negative	125
Somewhat Positive	259
Very Negative	184
Very Positive	290

4. Repeat this process for Q78B about the influence of the United States.
Answer: Majority of respondents perceive the U.S.A to have positive influence, where 480 said they are 'Somewhat Positive, followed by 420 respondents who said 'Very Positive'. On the other hand, 210 respondents describe U.S.A to have 'Somewhat Negative' influence and 170 strongly said U.S.A has 'Very Negative' influence. Only 260 respondents remained neutral regarding their perception and 40 respondents did not provide a clear response.Overall, these results indicate that while the perception of the U.S. is generally positive but there are some respondents who feel negative.

```
library(tidyverse)
library(haven)
library(knitr)

USA_Influence <- South_Africa_Round_9_Data_Afrobarometer %>%
  select(Q78B)

USA_Influence <- USA_Influence %>%
  mutate(
    Q78B = case_when(
      Q78B == 1 ~ "Very Negative",
      Q78B == 2 ~ "Somewhat Negative",
      Q78B == 3 ~ "Neither Positive nor Negative",
      Q78B == 4 ~ "Somewhat Positive",
      Q78B == 5 ~ "Very Positive",
      TRUE ~ "Don't Know / No Response"))

USA_Influence<- USA_Influence %>%
  count(Q78B, name = "Count")

USA_Influence %>%
  knitr::kable(caption = "Economic and Political Influence of USA in South Africa")
```

Table 2: Economic and Political Influence of USA in South Africa

Q78B	Count
Don't Know / No Response	683
Neither Positive nor Negative	204
Somewhat Negative	98
Somewhat Positive	283
Very Negative	111
Very Positive	201

5. Use the paired t-test to evaluate the difference between perceptions. To do this, you will need to clean both variables to exclude dk/na and refusals. See the example code but note that you may need to adjust depending on the method you used to read the data. Describe your findings. Use a two-tailed test and 5% significance.

Answer: The t-test value is -0.96 which suggest that there is small difference between respondent's perception between China and U.S.A. The p-value is 0.3365 (33.65%) which means p-value is greater than significant level 5% (0.05). There is no statistically significant difference between perceptions for China and U.S.A. We fail to reject the null hypothesis. In conclusion, the t-test suggests that respondents view towards China and the U.S.A are similar in terms of economic and political influence.

```
library(tidyverse)
library(haven)
library(knitr)

Perception_USA_China <- South_Africa_Round_9_Data_Afrobarometer %>%
  select(Q78A, Q78B)

Perception_USA_China <- Perception_USA_China %>%
  mutate(
    across(
      Q78A:Q78B,
      ~ if_else(.x %in% 1:5, .x, NA_real_)))
t.test(Perception_USA_China$Q78A, Perception_USA_China$Q78B, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Perception_USA_China$Q78A and Perception_USA_China$Q78B
## t = -0.96165, df = 835, p-value = 0.3365
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.14550636  0.04981258
## sample estimates:
## mean difference
##     -0.04784689
```