Computing and reporting descriptive statistics

Measures of spread for the mode

Jenine Harris
Brown School

Spread to report with the mode

- The mode tends to be used more often with categorical variables compared to continuous variables.
- Measures of spread for categorical variables determine how observations are spread across categories, sometimes described as *diversity*.
- These measures are often called **indices of qualitative variation** and have a range of 0 to 1.
- Values are *high* when observations are spread out among categories and *low* when they are not.
- For example, if a data set had a marital status variable and there were 3 people in each marital status category, the data would be considered perfectly spread across groups and the index value would be 1. Likewise if everyone in the data set was in one category (e.g., unmarried), the index value would be 0 for no spread at all.

Computing spread in R

- The **B index** in the qualvar package is one good option.
- The B index uses a vector of frequencies for a categorical variable.
- Instead of making a table as a new object, add the `table()` code directly into the `B()` function.

```
# import brfss data
brfss.trans.2014 <- read.csv(file = "~/Box/teaching/Teaching/Fall2020/da
# open tidyverse for data management
library(package = "tidyverse")
# cleaning the TRNSGNDR variable
brfss.2014.cleaned <- brfss.trans.2014 %>%
  mutate(TRNSGNDR = recode_factor(.x = TRNSGNDR,
                                  `1` = 'Male to female',
                                  `2` = 'Female to male',
                                  `3` = 'Gender non-conforming',
                                  `4` = 'Not transgender',
                                  `7` = 'Not sure',
                                  `9` = 'Refused'))
# B index of transgender variable
qualvar::B(x = table(brfss.2014.cleaned$TRNSGNDR))
```

```
## [1] 0.0009940017
```

Interpreting the B index

- The resulting value of 0.001 indicates that observations in this data set are not well spread out among the six categories of `TRNSGNDR`.
- While it is true that there are people in all categories, the *Not transgender* category contains a much larger group than any of the other categories, so the small value of B reflects this lack of even spread of observations across categories of `TRNSGNDR`.

Interpreting the descriptive statistics

- The mean number of days of poor health per month for participants in the 2014 BRFSS was 4.22 (s = 8.78).
- The median number of days of poor health per month for participants in the 2014 BRFSS was 0 (IQR = 3).
- The most common response (mode) to the transgender question was *not transgender*. The responses were not spread out very evenly among the categories with over 150,000 in the *not transgender* category and just 116 gender non-conforming category (B = .001).

Check your understanding

Find the central tendency and spread for the age variable (`X_AGE80`). Examine the variable and codebook first to see if it needs to be cleaned.

Answer

```
# check variable
summary(brfss.2014.cleaned$X_AGE80)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   44.00   58.00   55.49   69.00   80.00
```

```
# examine distribution
brfss.2014.cleaned %>%
  ggplot(aes(x = X_AGE80)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Answer (part 2)

```
# median and IQR
brfss.2014.cleaned %>%
  drop_na(X_AGE80) %>%
  summarize(med.age = median(x = X_AGE80),
            iqr.age = IQR(x = X_AGE80))
```

```
##   med.age iqr.age
## 1      58      25
```

```
# median age of 58, IQR of 25
```