Computing and reporting descriptive statistics

Mean, median, and mode

Jenine Harris
Brown School

Why frequency distributions do not work for numeric variables

- Numeric variables are measured on a continuum, and can be truly continuous or just close to continuous.
- These types of variables are not well described using frequency distributions.
- For example, a frequency table of the age variable (`X_AGE80`) looks like this:

```
# import brfss data
brfss.trans.2014 <- read.csv(file = "data/transgender_hc_ch2.csv")

# table with frequencies from the age variable
table(brfss.trans.2014$X_AGE80)
```

```
##
##    18    19    20    21    22    23    24    25    26    27    28    29    3
##  3447  3209  3147  3470  3470  3632  3825  3982  3723  3943  4191  4054  471
##    31    32    33    34    35    36    37    38    39    40    41    42    4
##  4169  4988  4888  4925  5373  5033  5109  5152  4891  5897  4672  6029  621
##    44    45    46    47    48    49    50    51    52    53    54    55    5
##  6091  6463  6252  6963  6994  7019  8925  7571  9060  9015  9268  9876  954
##    57    58    59    60    61    62    63    64    65    66    67    68    6
## 10346 10052 10293 11651 10310 11842 10955 10683 11513 10704 11583  9129  809
##    70    71    72    73    74    75    76    77    78    79    80
```

Defining and calculating central tendency

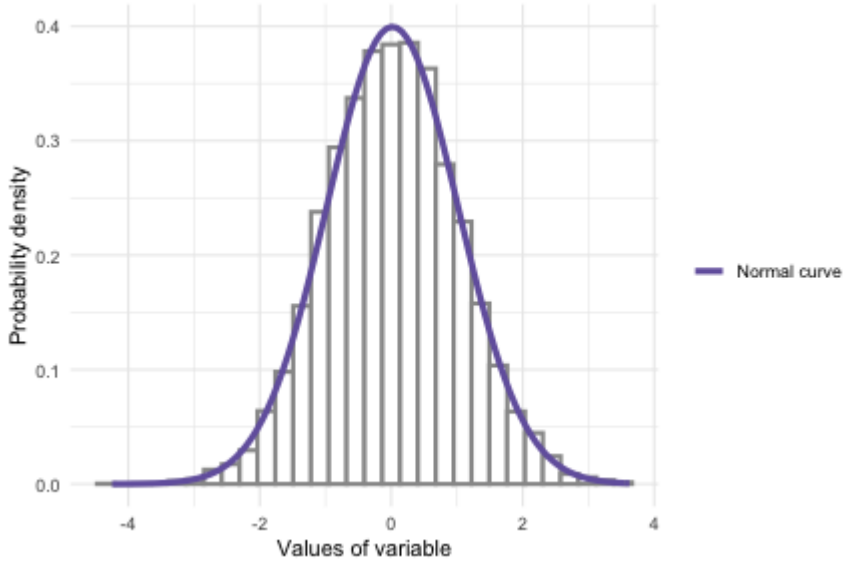
- Instead of frequencies and percentages, report central tendency and spread for continuous variables
- Central tendency measures:
  - The **mean** is the sum of the values divided by the number of values
  - The **median** is the middle value (or the mean of the two middle values if there is an even number of observations)
  - The **mode** is the most common value or values

Using the mean

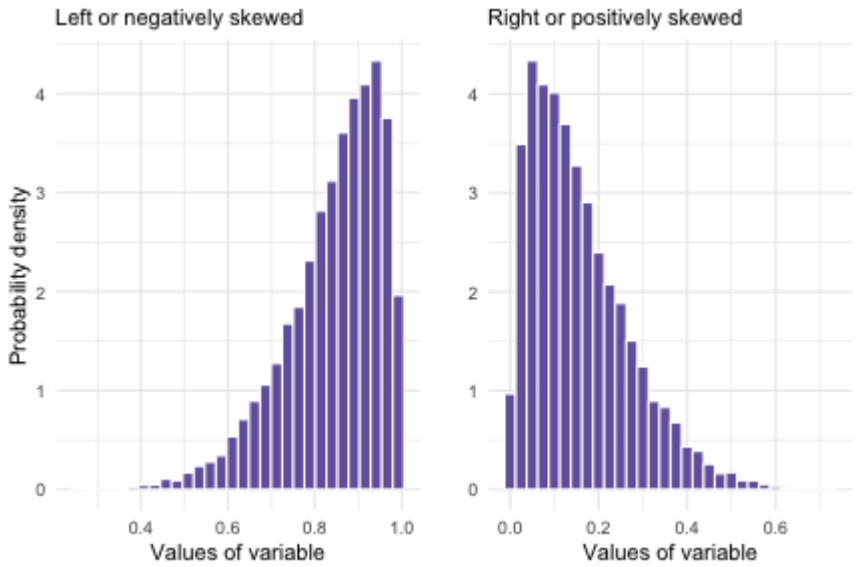
- The most well-understood and widely used measure of central tendency is the mean, which is calculated:

$$m_x = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean is useful for normally distributed data



The median is appropriate for skewed data



How skewed is too skewed?

- **Skewness** is a measure of the extent to which a distribution is skewed.

$$skewness_x = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - m_x}{s_x} \right)^3$$

Computing skewness in R

```
# skewness of a variable
semTools::skew(object = variable)
```

- The output will include four values, use the z to assess skewness
  - If the sample size is small (n < 50), z values outside the -2 to 2 range are a problem.
  - If the sample size is between 50 and 300, z values outside the -3.29 to 3.29 range are a problem.
  - For large samples (n > 300), using a visual is recommended over the statistics, but generally z values outside the range of -7 to 7 can be considered problematic.

R code for the measures of central tendency

- The three measures can be computed with R:

```
# mean, median, and mode
mean(x = data$variable)
median(x = data$variable)
names(x = sort(x = table(data$variable), decreasing = TRUE))[1]
```