Probability distributions and inference

Computing a z-score

Jenine Harris
Brown School

Import the opioid distance data

```
# bring in the opioid policy data and check it out
dist.mat <- read.csv(file = "/Users/harrisj/Box/teaching/Teaching/Fall20

# check the data frame
summary(object = dist.mat)
```

```
##     STATEFP         COUNTYFP          YEAR       INDICATOR
##  Min.   : 1.00   Min.   :  1.0   Min.   :2017   Length:3214
##  1st Qu.:19.00   1st Qu.: 35.0   1st Qu.:2017   Class :character
##  Median :30.00   Median : 79.0   Median :2017   Mode  :character
##  Mean   :31.25   Mean   :101.9   Mean   :2017
##  3rd Qu.:46.00   3rd Qu.:133.0   3rd Qu.:2017
##  Max.   :72.00   Max.   :840.0   Max.   :2017
##      VALUE            STATE           STATEABBREVIATION     COUNTY
##  Min.   :  0.00   Length:3214        Length:3214        Length:3214
##  1st Qu.:  9.25   Class :character   Class :character   Class :character
##  Median : 18.17   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 24.04
##  3rd Qu.: 31.00
##  Max.   :414.86
```

Transform distance to MAT facility

```
# open the tidyverse
library(package = "tidyverse")

# transforming the variable
dist.mat.cleaned <- dist.mat %>%
  mutate(miles.cube.root = VALUE^(1/3)) %>%
  mutate(miles.log = log(VALUE)) %>%
  mutate(miles.inverse = 1/VALUE) %>%
  mutate(miles.sqrt = sqrt(VALUE))

# summary stats
summary(object = dist.mat.cleaned)
```

```
##     STATEFP          COUNTYFP          YEAR       INDICATOR
##  Min.   : 1.00   Min.   :  1.0   Min.   :2017   Length:3214
##  1st Qu.:19.00   1st Qu.: 35.0   1st Qu.:2017   Class :character
##  Median :30.00   Median : 79.0   Median :2017   Mode  :character
##  Mean   :31.25   Mean   :101.9   Mean   :2017
##  3rd Qu.:46.00   3rd Qu.:133.0   3rd Qu.:2017
##  Max.   :72.00   Max.   :840.0   Max.   :2017
##      VALUE           STATE           STATEABBREVIATION     COUNTY
##  Min.   :  0.00   Length:3214        Length:3214        Length:3214
##  1st Qu.:  9.25   Class :character   Class :character   Class :character
##  Median : 18.17   Mode  :character   Mode  :character   Mode  :character
```

Computing and interpreting z-scores to compare observations to groups

- Regardless of what the mean and standard deviation are, a normally distributed variable has approximately:
  - 68% of values within one standard deviation of the mean
  - 95% of values within two standard deviations of the mean
  - 99.7% of values within three standard deviations of the mean
- These characteristics of the normal distribution can be used to describe and compare how far individual observations are from a mean value.

Checking the mean and sd for the distance data

```
dist.mat.cleaned %>%
  summarize(mean.dist.cube = mean(miles.cube.root),
            sd.dist.cube = sd(miles.cube.root))
```

```
##   mean.dist.cube sd.dist.cube
## 1       2.662915    0.7923114
```

- In the data on distance to treatment facility with medication-assisted therapy, for example, about 68% of counties are between 2.66 - .79 and 2.66 + .79 cube root of miles from a facility.
- About 68% of counties have between 1.87 and 3.45 cube root miles to the nearest substance abuse facility with MAT.
- Transforming these values back into miles would be cubing them so, 6.539203 to 41.063625 miles.
- In addition, about 95% of counties would be between 1.259712 and 76.225024 miles to travel to the nearest substance abuse facility with MAT.
- Kiara explained that this information was used to create z-scores, which allow description and comparison of where an observation falls compared to the other observations for a normally distributed variable.

Defining the z-score

- The z-score is *the number of standard deviations an observation is away from the mean*.

$$z_i = \frac{x_i - m_x}{s_x}$$

- The x_i represents the value of variable x for a single observation
- m_x is the mean of the x variable
- s_x is the standard deviation of the x variable.
- So, z_i is the difference between the observation value and the mean value for a variable and is converted by the denominator into standard deviations. The final z-score for an observation is *the number of standard deviations it is from the mean*.

Calculating and interpreting z-scores

- Use the z-score formula to calculate z for a county with residents who have to travel 50 miles to the nearest facility.
- In the transformed miles variable, this would be the cube root of 50, or a value of 3.68.
- Substituted the value the equation to compute z.

$$z = \frac{3.68 - 2.66}{.79} = 1.29$$

- A score of z = 1.29 indicated that the transformed distance to a facility with MAT for this example county was 1.29 standard deviations above the mean transformed distance from a county to a facility with MAT.
- This county was further away from MAT than the mean distance for a county.

Compute and interpret a z-score

- A county with a 10-mile distance to a facility with MAT, which is a value of 2.15 in the transformed distance variable, was .65 standard deviations *below* the mean transformed distance (z = -.65):

$$z = \frac{2.15 - 2.66}{.79} = -.65$$

- z-scores are positive for counties with a distance from MAT that was higher than the mean and a negative value for a county with a distance that was lower than the mean.

- The z-score not only indicates how many standard deviations away from the mean an observation is, but whether the observed value is above or below the mean.