Probability distributions and inference

Characteristics of a normal distribution

Jenine Harris
Brown School

foundations in public health
biostatistics

Characteristics and uses of the normal distribution of a continuous variable

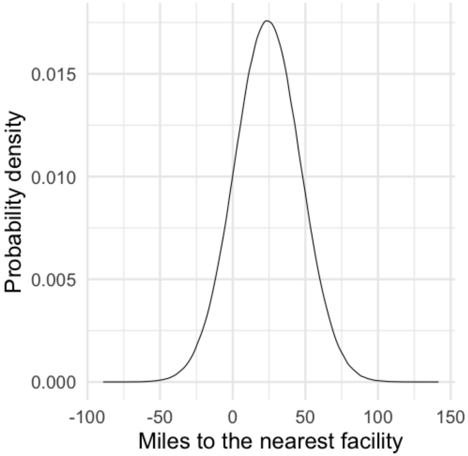
- Many of the variables of interest in social science are not binary so the binomial distribution and its related functions would not be all that useful.

- The probability distribution for a continuous variable is the normal distribution.

- Just as the shape of the binomial distribution is determined by n and p, the shape of the normal distribution for a variable in a sample is determined by μ and σ, the population mean and standard deviation, which are estimated by the sample mean and standard deviation, m and s.

Probability density function

- The normal distribution is used to find the likelihood of a certain value or range of values for a continuous variable.
- Like the probabilities from the binomial distribution are shown visually in a probability mass function graph, the normal distribution has a **probability density function** graph.
- The mean distance to the nearest substance abuse facility providing medication assisted treatment for all the counties in the amFAR data set was 24.04 miles with a standard deviation of 22.66 miles.
- Using the mean and standard deviation of 24.04 and 22.66, the probability density function graph for a variable with a mean of 24.04 and a standard deviation of 22.66 would look like:



Interpreting the probability density function graph

- The graph extends into negative numbers, which does not make sense for a measure of distance.
- There is no way to run or drive -2 miles.
- This variable might be *skewed* to the right rather than normally distributed, given the large standard deviation relative to its mean.
- This is a good opportunity to **transform** the variable to continue to discuss the normal distribution.
- For variables that are right skewed, a few transformations that could work to make the variable more normally distributed are: square root, cube root, reciprocal, and log.

Importing the data

- Import the distance data from amFAR and review it before data transformation.

```
# distance to substance abuse facility with medication assisted treatmen
dist.mat <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/data

# review the data
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
```

Data codebook

The variables in the data frame:

- STATEFP: Unique Federal Information Processing Standards (FIPS) code representing each state
- COUNTYFP: Unique FIPS code representing each county
- YEAR: Year data were collected
- INDICATOR: Label for value variable
- VALUE: Distance in miles to nearest substance abuse facility with medication assisted treatment (MAT)
- STATE: Name of state
- STATEABBREVIATION: Abbreviation for state
- COUNTY: Name of county

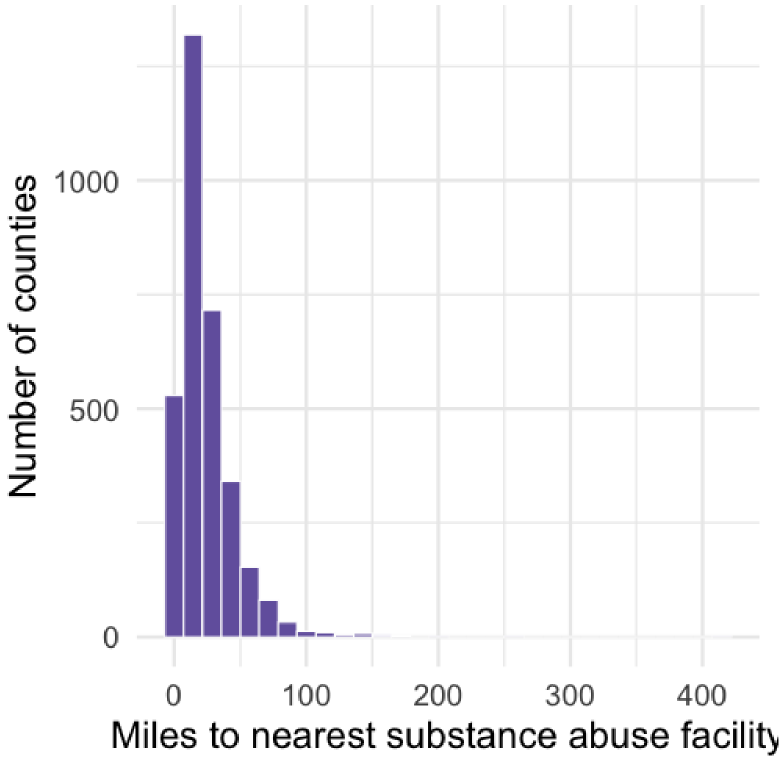
Examining the data

- The data are **county-level** data.
- The distances included in the data frame were the distance from the middle of the county to the nearest treatment facility with Medication-Assisted Therapy (MAT) for substance abuse.
- Examine the distance variable, `VALUE`, to confirm whether skew is a problem.

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

# graph the distance variable
dist.mat %>%
  ggplot(aes(x = VALUE)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal() +
  labs(x = "Miles to nearest substance abuse facility", y = "Number of c
```

Checking the plot



Check some transformations

- The distance variable was skewed for sure!
- Try the four transformations to see which is more useful for making the distance variable more normally distributed.

```
# 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))
```

Code to graph the transformations

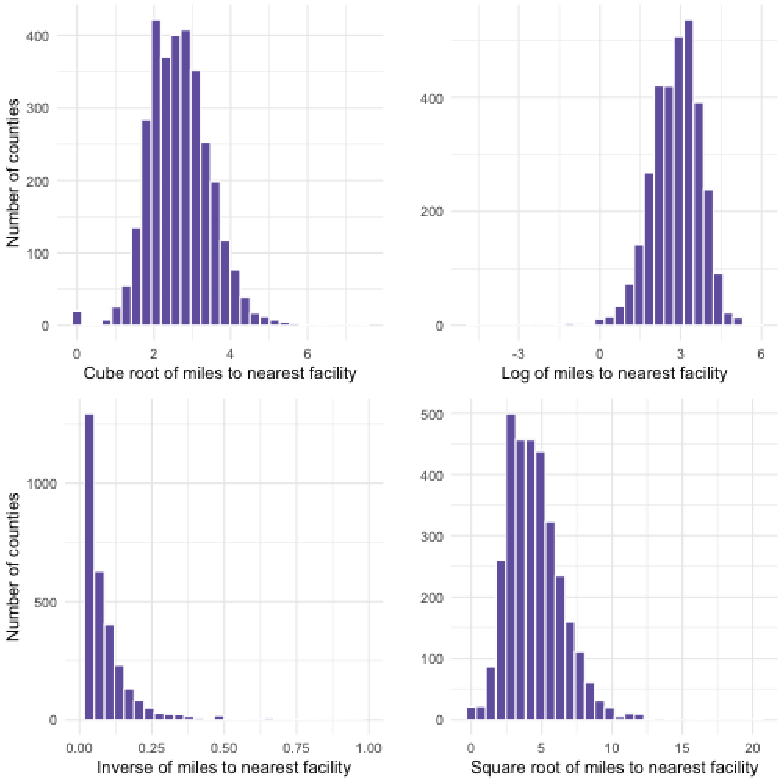
```
# graph the transformations
cuberoot <- dist.mat.cleaned %>%
  ggplot(aes(x = miles.cube.root)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 24) +
  labs(x = "Cube root of miles to nearest facility", y = "Number of coun

logged <- dist.mat.cleaned %>%
  ggplot(aes(x = miles.log)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 24) +
  labs(x = "Log of miles to nearest facility", y = "")

inversed <- dist.mat.cleaned %>%
  ggplot(aes(x = miles.inverse)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 24) + xlim(0,1) +
  labs(x = "Inverse of miles to nearest facility", y = "Number of counti

squareroot <- dist.mat.cleaned %>%
  ggplot(aes(x = miles.sqrt)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  theme_minimal(base_size = 24) +
  labs(x = "Square root of miles to nearest facility", y = "")
```

Examine the graphs



Choosing & using the transformed variable

- The cube root is the best of the four transformations for making the distribution appear normal, or *normalizing* the variable.
- The inverse did not work at all and made the variable appear even more skewed than it originally was.
- The log and square root both were fine, but the cube root was closest to normal.
- Find the mean and standard deviation of the cube root of distance:

```
# mean and standard deviation for cube root of miles
dist.mat.cleaned %>%
  summarize(mean.tran.dist = mean(miles.cube.root),
            sd.tran.dist = sd(miles.cube.root))
```

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

Graph the probability distribution for transformed variable

- Plot the probability distribution with these new summary statistics:

Interpreting the probability distribution

- The area under the curve in the figure represents 100% of observations.
- Using this **probability density function** graph to determine probabilities is a little different from using the **probability mass function** graph from the binomial distribution in the previous examples.
- With continuous variables the probability of any one specific value is going to be extremely low, often near zero.
- Instead, probabilities are usually computed for a *range of values*.
- For example, the shading under the curve represents US counties with the cube root of miles to a treatment facility being 4 or more, which is 4 cubed or 64 miles or more to the nearest substance abuse treatment facility with MAT.

Finding the area under the curve

- The `pnorm()` function is useful for finding the actual probability value for the shaded area under the curve.

- In this case, `pnorm()` could be used to determine the proportion of counties that are 4 or more cube root of miles to the nearest facility with MAT.

- The `pnorm()` command takes three arguments: q is the value of interest, the mean (m), and the standard deviation (s).

```
# shaded area under normal curve > 4
# when curve has mean of 2.66 and sd of .79
pnorm(q = 4, mean = 2.66, sd = .79)
```

```
## [1] 0.9550762
```

- The area shaded under the curve did not look like 95.5% of the area under the curve.

Using pnorm() to find area on the right side of curve

- The `pnorm()` function finds the area under the curve starting on the left up to, but not including, the q value entered, in this case 4.
- To get the area from 4 to under the right side tail of the distribution, add the `lower.tail = FALSE` option:

```
# shaded area under normal curve
# when curve has mean of 2.66 and sd of .79
pnorm(q = 4, mean = 2.66, sd = .79, lower.tail = FALSE)
```

```
## [1] 0.04492377
```

- It looked like 4.49% of observations are in the shaded part of this distribution and therefore have a value for the distance variable of 4 or greater.
- Reversing the transformation, this indicates that residents of 4.49% of counties have to travel or 64 miles or more to get to the nearest substance abuse facility providing medication-assisted treatment.

This seems really far to travel to get treatment, especially for people struggling with an opioid addiction or trying to help their family members and friends.