Probability distributions and inference

Population means from sample means

Jenine Harris
Brown School

Samples and populations

- The characteristics of the normal curve are exceptionally useful in better understanding the characteristics of a population when it is impossible to measure the entire population.
- For example, there was no real way to measure the height or weight or income of every single person in the US.
- Instead, researchers often use a **representative sample** from the population they are interested in and use properties of the normal distribution to understand what is likely happening in the whole population.
- A **representative sample** is a sample taken carefully so that it does a good job of representing the characteristics of the population.
- For example, if a sample of US citizens were taken and the sample was 75% female, this would not be **representative** of the distribution of sex in the population.
- There are many strategies for sampling that help to ensure a representative sample.

Using the normal distribution and samples to understand populations

- Use the distance to a treatment facility variable to understand how to use a sample mean to estimate a population mean.
- There is no need to transform the variable because the theory works for continuous variables whether they are normally distributed or not.
- Import the data and rename the distance variable to be `distance` since that is easier to remember.

```
# distance to substance abuse facility with medication assisted treatmen
dist.mat <- read.csv("/Users/harrisj/Box/teaching/Teaching/Fall2020/data
# rename variable
library(package = "tidyverse")
dist.mat.cleaned <- dist.mat %>%
  rename('distance' = VALUE)
```

Summarize the 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
```

Examine the distance data

- `summarize()` the distance variable as a reminder of the mean and standard deviation.

```
# get mean and sd from cleaned data
dist.mat.cleaned %>%
  summarize(mean.dist = mean(x = distance, na.rm = TRUE),
            sd.dist = sd(x = distance, na.rm = TRUE),
            n = n())
```

```
##   mean.dist  sd.dist    n
## 1  24.04277 22.66486 3214
```

- The mean distance to the nearest substance abuse facility that provides at least one type of MAT is 24.04 miles away, with a standard deviation of 22.66 miles.
- There were 3,214 counties in the data set representing all of the counties in the US.
- This is all the US counties, so it is not a sample of counties, it is the population of counties.

Examining a sample from a population

How close can you get to the true mean distance to a facility with MAT (m = 24.04) if you only had enough time and money to collect data on distances from 500 counties rather than all 3,214 counties?

- Use `set.seed()` for reproducibility, `sample_n()` to sample, and `summarize()` to examine the results

```
# set a starting value for sampling
set.seed(seed = 1945)

# sample 500 counties at random
counties.500 <- dist.mat.cleaned %>%
  sample_n(size = 500, replace = TRUE)

# compute the mean death rate in the sample
counties.500 %>%
  summarize(mean.s1 = mean(x = distance, na.rm = TRUE))
```

```
##    mean.s1
## 1 24.40444
```

Interpreting the results

- The result is 24.40 miles, which is close to the population mean of 24.04 miles, but not exactly the same.
- Try it again with a different seed value to choose a different set of 500 counties.

```
# set a different starting value for sampling
set.seed(seed = 48)

# sample 500 counties at random
counties.500.2 <- dist.mat.cleaned %>%
  sample_n(size = 500, replace = TRUE)

# compute the mean death rate in the sample
counties.500.2 %>%
  summarize(mean.s2 = mean(x = distance, na.rm = TRUE))
```

```
##    mean.s2
## 1 23.49652
```

- This time the mean distance is 23.50 miles, which is slightly lower than the first sample mean of 24.40 miles and still close to the population value of 24.04 miles.

Examining a sample of samples from a population

- What would happen if you took 20 samples of counties where each sample had 500 counties in it.

```
# get 20 samples
# each sample has 500 counties
# put samples in a data frame with each sample having
# a unique id called sample_num
set.seed(seed = 111)
samples.20 <- bind_rows(replicate(n = 20, dist.mat.cleaned %>%
                                  sample_n(size = 500, replace = TRUE),
                                  simplify = FALSE), .id = "sample_num")

# find the mean for each sample
sample.20.means <- samples.20 %>%
  group_by(sample_num) %>%
  summarize(mean.distance = mean(x = distance, na.rm = TRUE))
sample.20.means
```

Review the 20 means

```
## # A tibble: 20 x 2
##    sample_num mean.distance
##    <chr>              <dbl>
##  1 1                   24.2
##  2 10                  22.0
##  3 11                  23.9
##  4 12                  23.8
##  5 13                  23.1
##  6 14                  23.0
##  7 15                  22.6
##  8 16                  24.4
##  9 17                  24.4
## 10 18                  24.0
## 11 19                  23.7
## 12 2                   24.2
## 13 20                  23.1
## 14 3                   23.9
## 15 4                   24.4
## 16 5                   24.7
## 17 6                   22.8
## 18 7                   24.2
## 19 8                   23.9
## 20 9                   24.2
```

What is the mean of the 20 sample means

- With the mean distance to substance abuse facilities for each of 20 samples, including 500 counties, try taking the mean of the sample means.

```
# find the mean of the 20 sample means
sample.20.means %>%
  summarize(mean.20.means = mean(mean.distance))
```

```
## # A tibble: 1 x 1
##   mean.20.means
##           <dbl>
## 1          23.7
```

What is the mean of 100 sample means

- Change the `n = 20` to `n = 100` for 100 samples and find the mean of the 100 sample means.

```
# get 100 samples
set.seed(seed = 143)
samples.100 <- bind_rows(replicate(n = 100, dist.mat.cleaned %>%
                                   sample_n(size = 500, replace = TRUE),
                                   simplify = FALSE), .id = "sample_num")

# find the mean for each sample
sample.100.means <- samples.100 %>%
  group_by(sample_num) %>%
  summarize(mean.distance = mean(x = distance, na.rm = TRUE))

# find the mean of the 100 sample means
sample.100.means %>%
  summarize(mean.100.means = mean(mean.distance))
```

```
## # A tibble: 1 x 1
##   mean.100.means
##            <dbl>
## 1           24.0
```

Larger samples and more samples

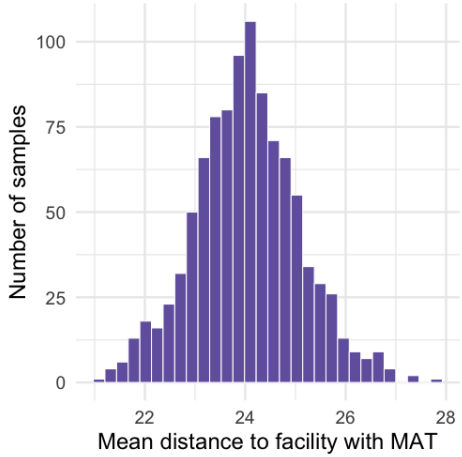
- Get closer to the population mean if they use more samples and larger sample sizes.
- Try taking 1,000 samples and getting the mean.

```
# get 1000 samples
set.seed(seed = 159)
samples.1000 <- bind_rows(replicate(n = 1000, dist.mat.cleaned %>%
                                    sample_n(size = 500, replace = TRUE),
                                    simplify = FALSE), .id = "sample_num")

# find the mean for each sample
sample.1000.means <- samples.1000 %>%
  group_by(sample_num) %>%
  summarize(mean.distance = mean(x = distance, na.rm = TRUE))

# find the mean of the sample means
sample.1000.means %>%
  summarize(mean.1000.means = mean(mean.distance))
```

```
## # A tibble: 1 x 1
##   mean.1000.means
##             <dbl>
## 1            24.0
```

Graphing the means from 1000 samples

```
# histogram of the 1000 means
sample.1000.means %>%
  ggplot(aes(x = mean.distance)) +
  geom_histogram(fill = "#7463AC", color = "white") +
  labs(x = "Mean distance to facility with MAT",
       y = "Number of samples") +
  theme_minimal(base_size = 24)
```



Interpreting the mean of sample means

- The mean of the sample means of 24.036 is very close to the population mean of 24.04 and the graph of the sample means looks a lot like a normal distribution.
- Taking a lot of large samples and graphing their means results in a **sampling distribution** that looks like a normal distribution, and, more importantly, *the mean of the sample means is nearly the same as the population mean*.
- A **sampling distribution** is the distribution of summary statistics from repeated samples taken from a population.

The Central Limit Theorem

- This phenomenon is called the **Central Limit Theorem** and that it holds true for continuous variables that are normally distributed and those that not normally distributed.
- The Central Limit Theorem is one of the most important ideas for *inferential statistics*, or statistical approaches that infer population characteristics based on sample data.
- Another characteristic of the **Central Limit Theorem** is that the standard deviation of the sample means can be estimated using the population standard deviation (σ) and the size of the samples used to make the sampling distribution (n).

$$s_{samp.dist} = \frac{\sigma}{\sqrt{n}}$$

Computing the sampling dist sd from the population

- Because the `var()` function uses the Bessel correction of n-1 in the denominator to account for the limited variation in a sample compared to in a population, it is a little tricky to compute the population variance.
- To reverse the Bessel correction, use `var()*((n-1)/n)` to get the population variance of σ^2 and then use `sqrt()` from there to get population standard deviation, σ.
- This could then be used to get the estimate of the sampling distribution standard deviation.
- Leslie thought that sounded like a good plan and wrote some code.

```
# compute estimated standard dev of sampling dist
dist.mat.cleaned %>%
  drop_na(distance) %>%
  summarize(n = n(),
            pop.var = var(x = distance)*((n-1)/n),
            pop.s = sqrt(x = pop.var),
            s.samp.dist.est = pop.s/sqrt(x = 500))
```

```
##      n pop.var    pop.s s.samp.dist.est
## 1 3214 513.536 22.66133        1.013446
```

Computing the sampling distribution sd directly

- Compute the standard deviation of the sampling distribution directly since the 1000 sample means are saved in a data frame.

```
# computing the samp dist standard devation
# directly from the 1000 sample means
sd(x = sample.1000.means$mean.distance, na.rm = T)
```

```
## [1] 1.04966
```

- Noticed that the results are similar (1.01 and 1.05), but not identical.

The standard error

- Since the *mean of the sample means* in the sampling distribution is very close to the *population mean*, the standard deviation of the sampling distribution shows how much we expect sample means to vary from the population mean.
- Specifically, given that the distribution of sample means is relatively normally distributed, 68% of sample means will be within one standard deviation of the mean of the sampling distribution and 95% of sample means will be within two standard deviations of the sampling distribution mean.
- Since the sampling distribution mean is a good estimate of the population mean, it follows that **most of the sample means are within one or two standard deviations of the population mean**.
- It is unusual to have the entire population for computing the population standard deviation, and it is also unusual to have a large number of samples from one population, so a close approximation to this value is called the *standard error of the mean*.

Calculating the standard error

- The standard error is computed by dividing the standard deviation *of a sample* by the square root of the sample size.

$$se = \frac{s}{\sqrt{n}}$$

- Where s is a sample standard deviation and n is the sample size.

Calculating the standard error for a sample

- Compute the standard error for the mean of distance in the first sample of 500 counties:

```
# mean, sd, se for first sample of 500 counties
counties.500 %>%
  summarize(mean.dist = mean(x = distance),
            sd.dist = sd(x = distance),
            se.dist = sd(x = distance)/sqrt(x = length(x = distance)))
```

```
##   mean.dist  sd.dist  se.dist
## 1  24.40444 23.79142 1.063985
```

Calculating the standard error for another sample

- Compute the mean, standard deviation, and standard error for the second sample of 500 counties:

```
# mean, sd, se for second sample
counties.500.2 %>%
  summarize(mean.dist = mean(x = distance),
            sd.dist = sd(x = distance),
            se.dist = sd(x = distance)/sqrt(x = length(x = distance)))
```

```
##   mean.dist  sd.dist   se.dist
## 1  23.49652 20.08756 0.8983431
```

Interpreting the standard errors

- Both of the standard error (se) values are close to the sampling distribution standard deviation of 1.05, but are not exactly the same.
- The first sample standard error of 1.06 was a little above and the second sample standard error of .90 was a little below.

Summarizing the ideas so far

- The standard deviation of the sampling distribution is 1.05.
- The standard error from the first sample is 1.06.
- The standard error from the second sample is 0.9.

Using sample information to understand populations

- Most of the time researchers have a single sample and so the only feasible way to determine the standard deviation of the sampling distribution is by computing the standard error of the single sample.
- This value tends to be a good estimate of the standard deviation of sample means.
    - about 68% of sample means are within one standard deviation of the sampling distribution mean (i.e., mean-of-sample-means)
    - about 95% of sample means are within two standard deviations of the sampling distribution mean (i.e., mean-of-sample-means)
    - the mean of a sampling distribution tends to be close to the population mean
    - the standard error is a good estimate of the sampling distribution standard deviation
- **The mean of any given sample (given good data collection practices) is likely to be within two standard erors of the population mean for that variable.**
- This is one of the foundational ideas of **inferential statistics**.

Standard deviation vs. standard error

- There is a difference between a sample standard deviation and a sample standard error.
- The standard deviation is a measure of the **variability in the sample** while the standard error is an estimate of **how closely the sample represents the population.**
- If there were no way to measure every county to get the population mean for the distance to the nearest substance abuse facility in a county, use the first sample of 500 counties to estimate that there is a good chance that the population mean distance is between 24.40 - 1.06 and 24.40 + 1.06 or 23.34 and 25.46.
- There is an even higher probability that the population mean distance is between $24.40 - 1.06 \cdot 2$ and $24.40 + 1.06 \cdot 2$ or 22.28 and 26.52.
- While it is unusual to have the population mean to compare sample results to, this time we have it and can see that the population mean of 24.04 miles to the nearest substance abuse facility is represented well by the sample mean of 24.40 and its standard error of 1.06.