Sampling distributions & Central Limit Theorem

Dr. Maria Tackett

11.14.19

Click for PDF of slides

Announcements

- Writing Exercise #3 draft **due TODAY 11:59p**
- HW 04 **due TODAY at 11:59p**
- Team Feedback #3 **due Sunday at 11:59p**

Sampling Distributions & Central Limit Theorem

Sample Statistics and Sampling Distributions

Notation

- Means:
  - **Population**: mean = μ, standard deviation = σ
  - **Sample**: mean = $\bar{x}$, standard deviation = s

- Proportions:
  - **Population**: p
  - **Sample**: $\hat{p}$

- **Standard error**: SE

Variability of sample statistics

- Each sample from the population yields a slightly different sample statistic (sample mean, sample proportion, etc.)
- The variability of these sample statistics is measured by the standard error
- Previously we quantified this value via simulation
- Today we talk about the theory underlying **sampling distributions**

Sampling distribution

- Sampling distribution is the distribution of sample statistics of random samples of size n taken with replacement from a population
- In practice it is impossible to construct sampling distributions since it would require having access to the entire population
- Today for demonstration purposes we will assume we have access to the population data, and construct sampling distributions, and examine their shapes, centers, and spreads

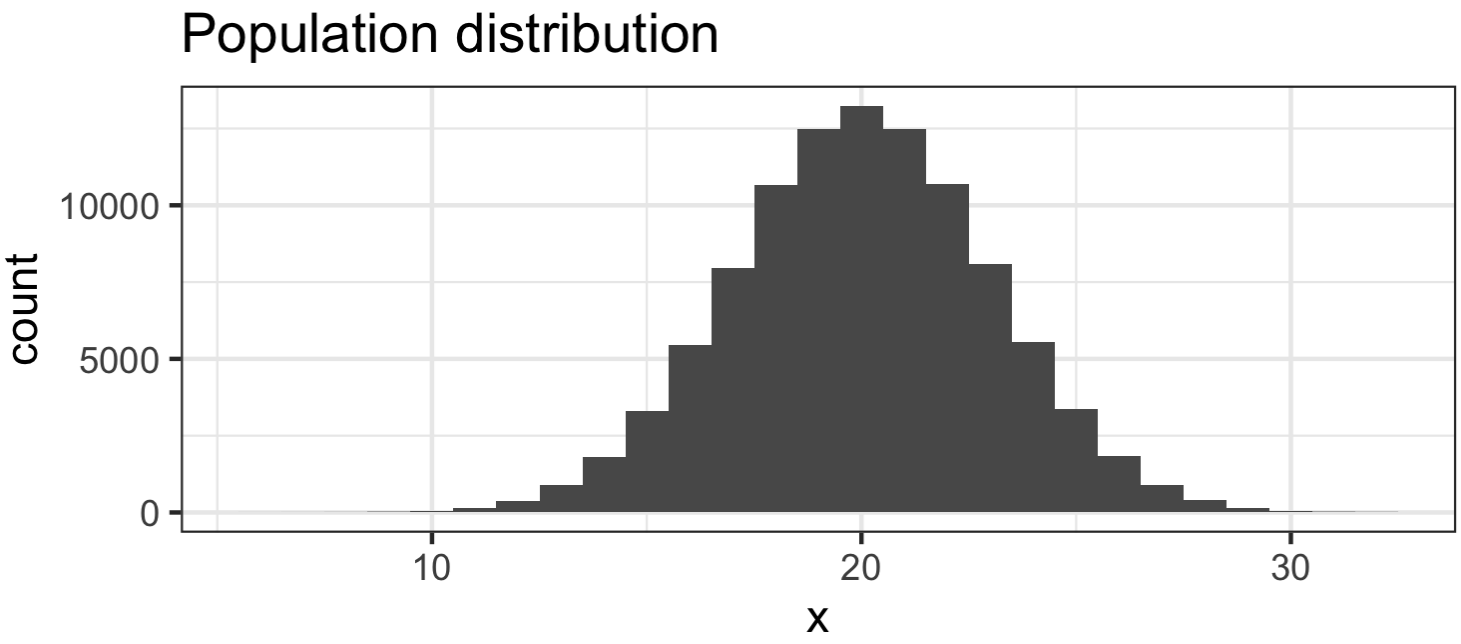
What is the difference between the sampling distribution and bootstrap distribution?

The sampling distribution

We have a population that is normally distributed with mean 20 and standard deviation 3. Suppose we take samples of size 50 from this distribution, and plot their sample means. What shape, center, and spread will this distribution have?

The population

```
set.seed(111219)
norm_pop <- tibble(x = rnorm(n = 100000, mean = 20, sd = 3))
ggplot(data = norm_pop, aes(x = x)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Population distribution")
```



Sampling from the population - 1

```
samp_1 <- norm_pop %>%
  sample_n(size = 50, replace = TRUE)
```

```
samp_1 %>%
  summarise(x_bar = mean(x))
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  20.9
```

Sampling from the population - 2

```
samp_2 <- norm_pop %>%
  sample_n(size = 50, replace = TRUE)
```

```
samp_2 %>%
  summarise(x_bar = mean(x))
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  19.9
```

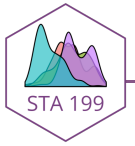
Sampling from the population - 3

```
samp_3 <- norm_pop %>%
  sample_n(size = 50, replace = TRUE)
```

```
samp_3 %>%
  summarise(x_bar = mean(x))
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  19.0
```

keep repeating...



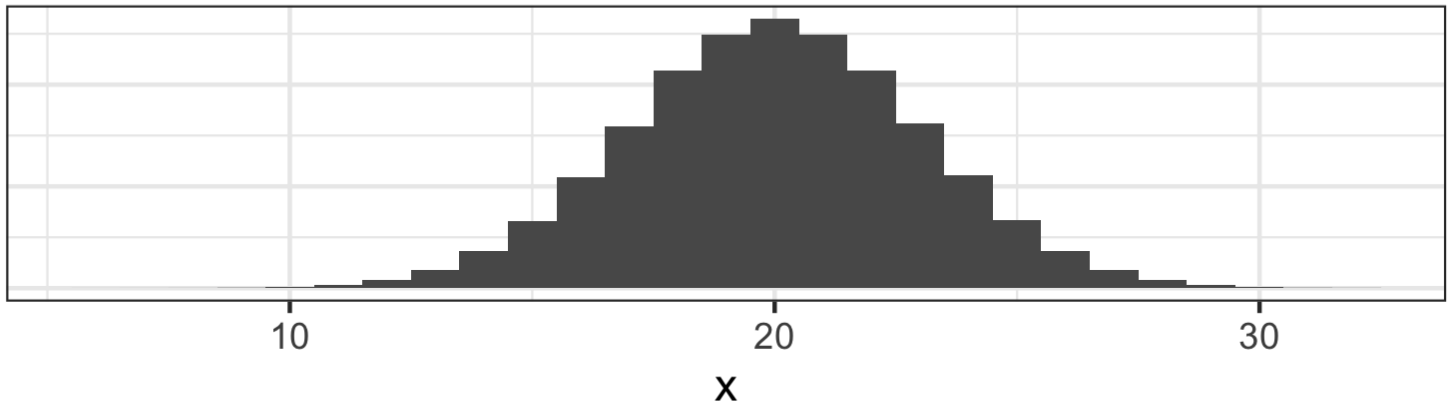
Sampling distribution

```
sampling <- norm_pop %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 1000) %>%
  group_by(replicate) %>%
  summarise(xbar = mean(x))
sampling
```

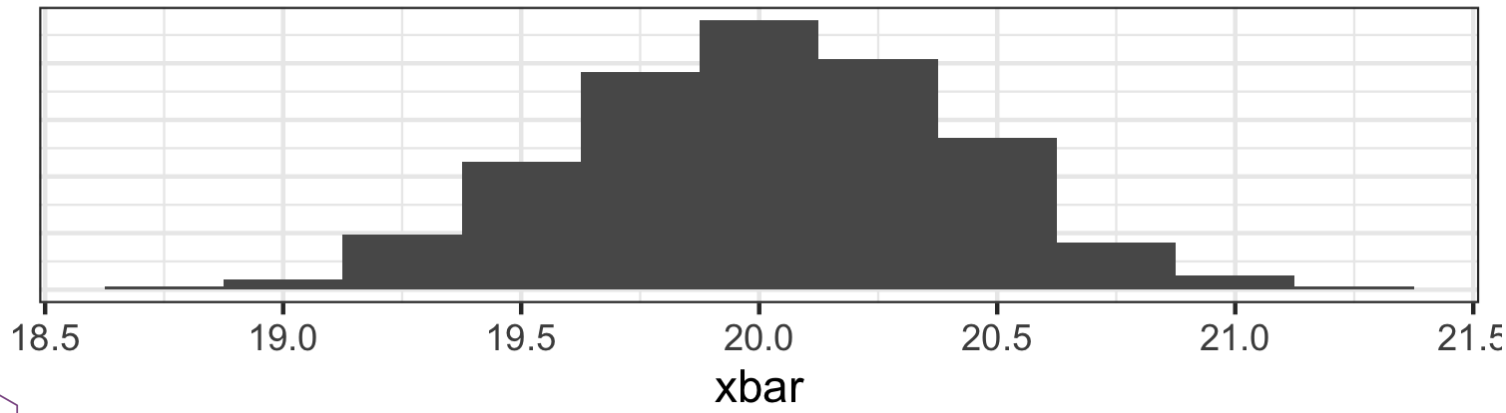
```
## # A tibble: 1,000 x 2
##    replicate  xbar
##        <int> <dbl>
##  1         1  19.4
##  2         2  20.9
##  3         3  20.4
##  4         4  19.5
##  5         5  19.9
##  6         6  19.6
##  7         7  19.8
##  8         8  20.4
##  9         9  20.4
## 10        10  19.4
## # … with 990 more rows
```

Population vs. sampling

Population distribution

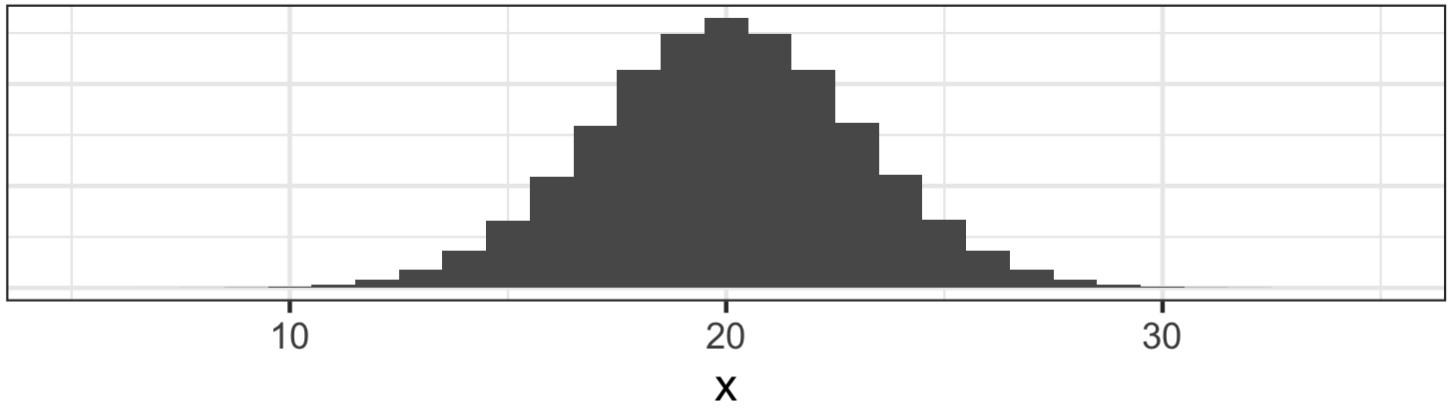


Sampling distribution of sample means

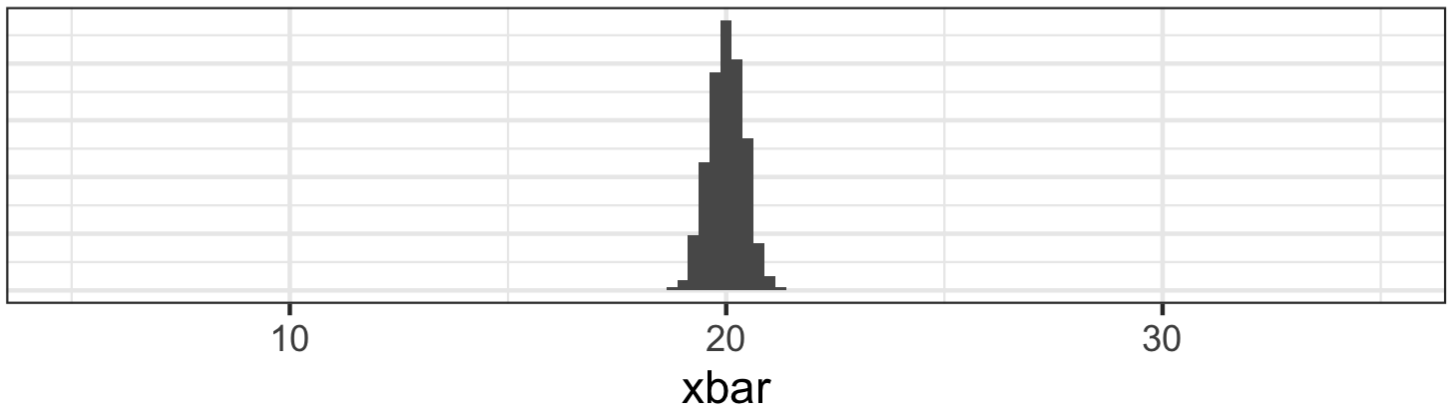


How do the shapes, centers, and spreads of these distributions compare?

Population distribution



Sampling distribution of sample means



How do the centers and spreads of these distributions compare?

```
norm_pop %>%
  summarise(mu = mean(x), sigma = sd(x))
```

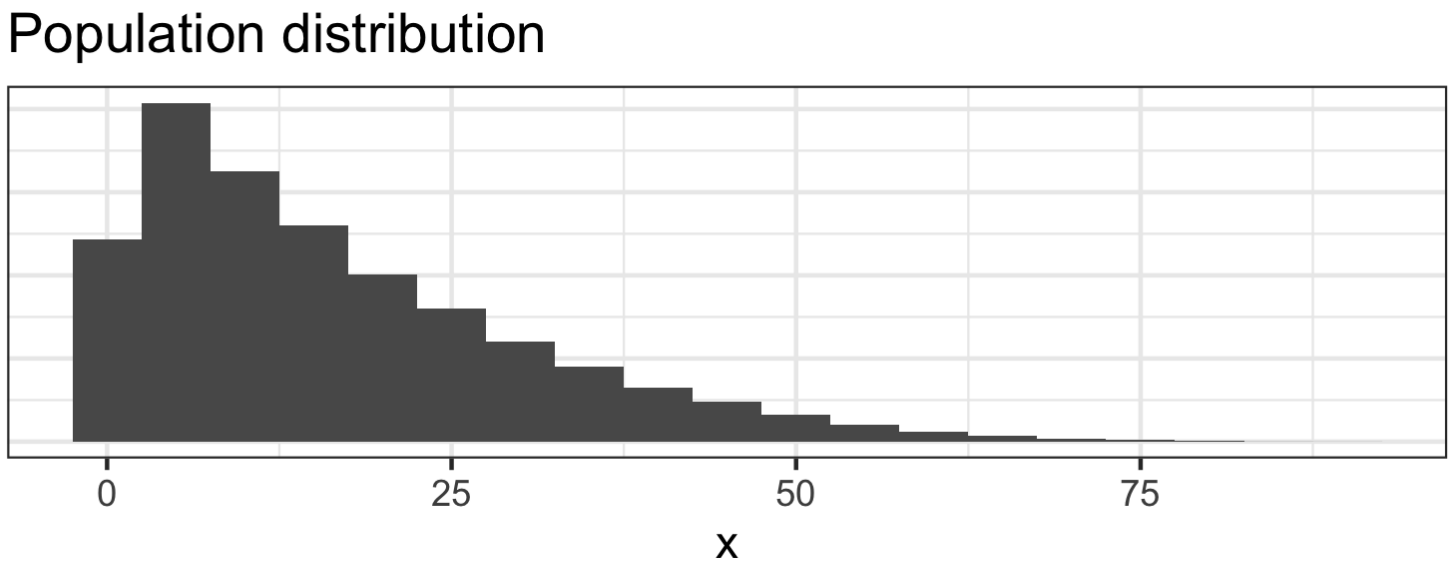
```
## # A tibble: 1 x 2
##      mu sigma
##   <dbl> <dbl>
## 1  20.0  3.00
```

```
sampling %>%
  summarise(mean = mean(xbar), se = sd(xbar))
```

```
## # A tibble: 1 x 2
##    mean    se
##   <dbl> <dbl>
## 1  20.0 0.402
```

Simulating another sampling distribution

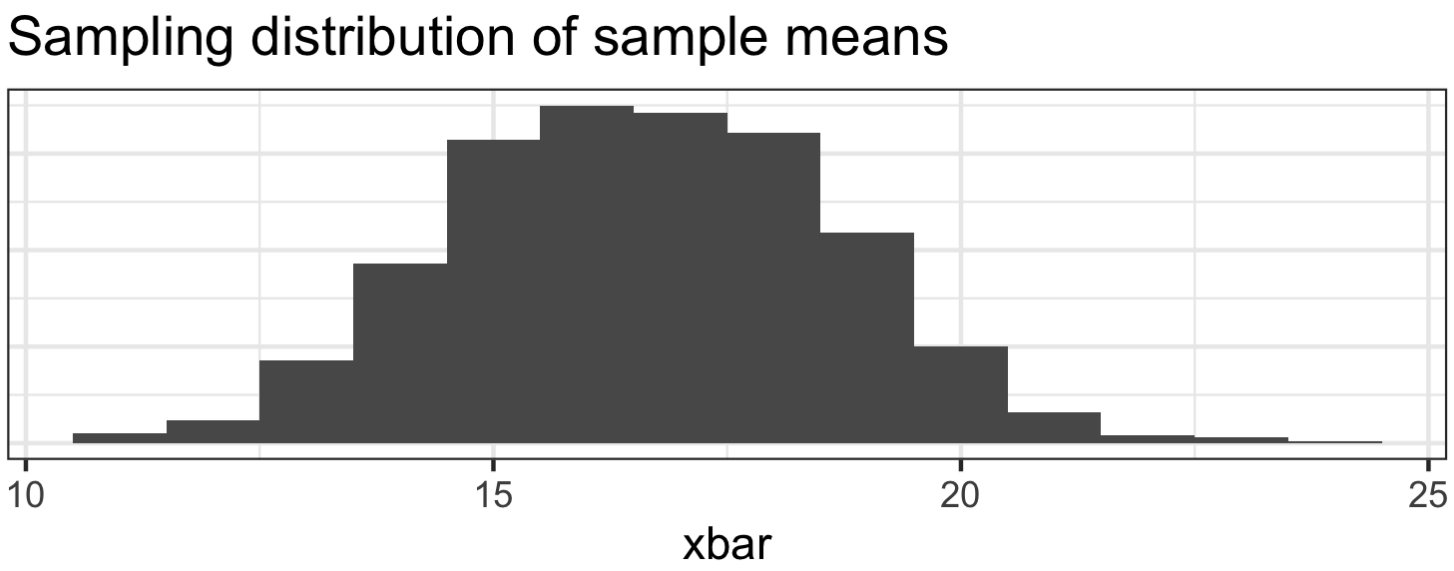
```
rs_pop <- tibble(x = rbeta(100000, 1, 5) * 100)
```




```
## # A tibble: 1 x 2
##      mu sigma
##   <dbl> <dbl>
## 1  16.6  14.1
```

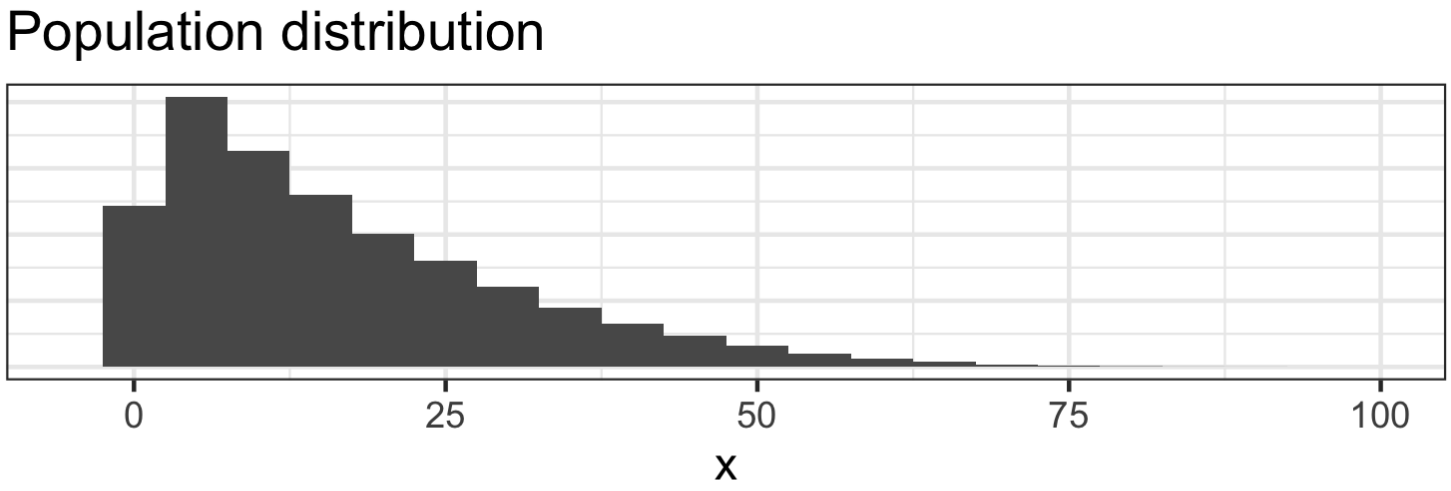
Sampling distribution

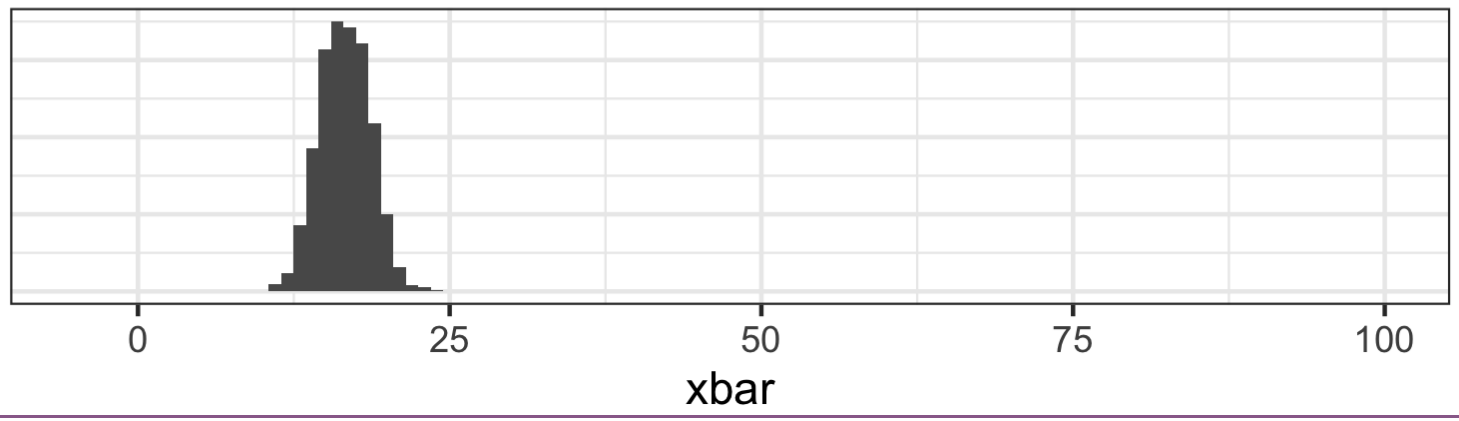
```
sampling <- rs_pop %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 1000) %>%
  group_by(replicate) %>%
  summarise(xbar = mean(x))
```




```
## # A tibble: 1 x 2
##    mean    se
##   <dbl> <dbl>
## 1  16.6  2.02
```

How do the shapes, centers, and spreads of these distributions compare?





Recap

- Regardless of the shape of the population distribution, the sampling distribution follows a normal distribution.
- The center of the sampling distribution is at the center of the population distribution.
- The sampling distribution is less variable than the population distribution.

What was the one (very unrealistic) assumption we had in simulating these sampling distributions?

Central Limit Theorem

In practice...

We can't directly know what the sampling distributions looks like, because we only draw a single sample.

- The whole point of statistical inference is to deal with this issue: observe only one sample, try to make inference about the entire population
- We have already seen that there are simulation based methods that help us estimate the sampling distribution
- Additionally, there are theoretical results (Central Limit Theorem) that tell us what the sampling distribution should look like (for certain sample statistics)

Central Limit Theorem

If certain conditions are met (more on this in a bit), the sampling distribution of the sample statistic will be

- nearly normal
- mean equal to the unknown population parameter
- standard error proportional to the inverse of the square root of the sample size.

Central Limit Theorem

One Sample:

- Single mean: $\bar{x} \sim N\left(mean = \mu, sd = \frac{\sigma}{\sqrt{n}}\right)$
- Single proportion: $\hat{p} \sim N\left(mean = p, sd = \sqrt{\frac{p(1-p)}{n}}\right)$

Two Sample:

- Difference between two means:
  $(\bar{x}_1 - \bar{x}_2) \sim N\left(mean = (\mu_1 - \mu_2), sd = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
- Difference between two proportions:
  $(\hat{p}_1 - \hat{p}_2) \sim N\left(mean = (p_1 - p_2), sd = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$

Conditions for inference

- **Independence:** The sampled observations must be independent. This is difficult to check, but the following are useful guidelines:
  - the sample must be random
  - if sampling without replacement, sample size must be less than 10% of the population size
- **Sample size / distribution:**
  - numerical data: The more skewed the sample (and hence the population) distribution, the larger samples we need. Usually n > 30 is considered a large enough sample for population distributions that are not extremely skewed.
  - categorical data: At least 10 successes and 10 failures.
- If comparing two populations, the groups must be independent of each other, and all conditions should be checked for both groups.

Standard Error

The **standard error** is the *standard deviation* of the *sampling distribution*, calculated using sample statistics (since we don't know the population parameters like σ or p).

- Single mean: $SE = \frac{s}{\sqrt{n}}$
- Difference between two means: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Single proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Difference between two proportions: $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

How are standard error and sample size associated? What does that say about how the spread of the sampling distribution changes as n increases?

What is the normal distribution?

Normal distribution

- is unimodal and symmetric
- follows the 68-95-99.7 rule