Simulation-based inference review & Sampling distributions

Dr. Maria Tackett

11.12.19

STA 199

Click for PDF of slides

Announcements

- Team Feedback #3 **due Thu at 11:59p**
- Writing Exercise #3 draft **due Thu, at 11:59p**
- HW 04 **due Thu, Nov 14 at 11:59p**
- Project data analysis due December 3

Thinking about inference

Let's walk through the thought process for conducting statistical inference:

Getting Started

Step 1: Start by asking questions

- **Question 1:** What is the parameter you're interested in understanding?
- **Question 2:** What is the sample statistic associated with this parameter?
- **Question 3:** What is the objective - estimation or testing a claim?

Confidence intervals (estimation)

Step 2: Use the sample data to generate a bootstrap distribution

Step 3: Use the bootstrapped distribution to calculate the upper and lower bounds for the confidence interval

Step 4: Interpret the interval in the context of the data

Testing a claim (hypothesis tests)

Step 2: State the null and alternative hypotheses

Step 3: Use the parameter(s) specified in the null hypothesis to generate the null distribution

Step 4: Use the null distribution, observed sample statistic, and alternative hypothesis to calculate the p-value

Step 5: Compare the p-value to the significance level α to make a conclusion (reject or fail to reject H_0)

Step 6: State the conclusion in the context of the data

Inference for a single numeric variable

Hypothesis testing for a single numeric variable

Let's go back to the price to rent a one-bedroom apartment in Manhattan.

```r
library(tidyverse)
manhattan <- read_csv("data/manhattan.csv")
```

```r
manhattan %>% slice(1:10)
```

```
## # A tibble: 10 x 1
##     rent
##    <dbl>
##  1  3850
##  2  3800
##  3  2350
##  4  3200
##  5  2150
##  6  3267
##  7  2495
##  8  2349
##  9  3950
## 10  1795
```

```r
manhattan %>% slice(11:20)
```

```
## # A tibble: 10 x 1
##     rent
##    <dbl>
##  1  2145
##  2  2300
##  3  1775
##  4  2000
##  5  2175
##  6  2350
##  7  2550
##  8  4195
##  9  1470
## 10  2350
```

Rent in Manhattan

```
manhattan %>% summarise(mean=mean(rent))
```

```
## # A tibble: 1 x 1
##     mean
##    <dbl>
## 1 2626.
```

According to the site Rent Jungle, the average price to rent an apartment in LA is around $2400. **Is the average rent for a one-bedroom in Manhattan significantly different from $2400?**

Step 1: Start by asking questions:

- What is the parameter you're interested in understanding?
- What is the sample statistic associated with this parameter?
- What is the objective - estimation or testing a claim?

Rent in Manhattan

Step 2: State the null and alternative hypotheses

$$H_0 : \mu = 2400$$
$$H_a : \mu \neq 2400$$

Step 3: Use the parameter(s) specified in the null hypothesis to generate the null distribution

- In practice, we'll use the **`generate`** function in the infer package to generate the null distribution
- Let's talk about what is going on when we use the **`generate`** function

Simulation process

We will use bootstrapping to generate the null distribution, ie. a sampling distribution of sample means under the assumption H_0 is true.

1. Take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample.

2. Calculate the mean of the bootstrap sample.

3. Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrapped means.

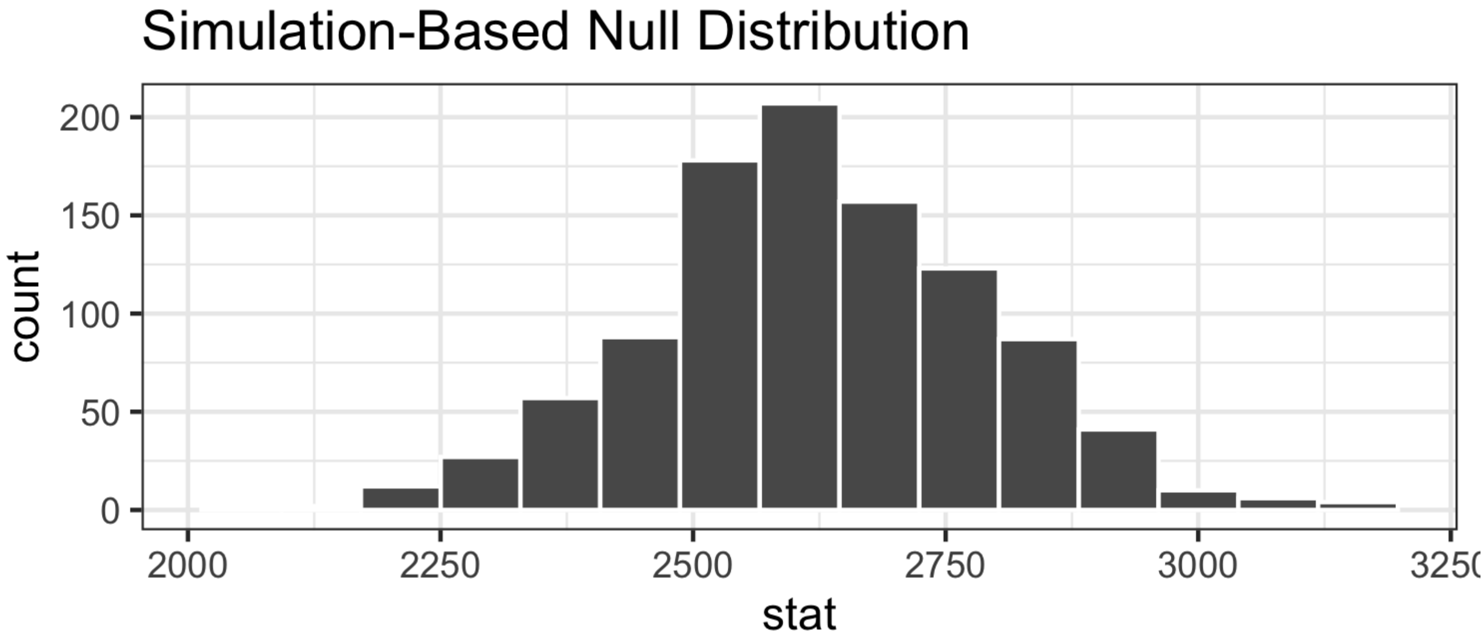
4. Shift the bootstrap distribution to be centered at the null value by subtracting/adding the difference between the center of the bootstrap distribution and the null value to each bootstrap mean.

Simulation: Take Bootstrap Sample

Take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample

```
set.seed(111219)
rent_bootstrap <- manhattan %>%
  specify(response = "rent") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")
```

Simulation: Take Bootstrap Sample



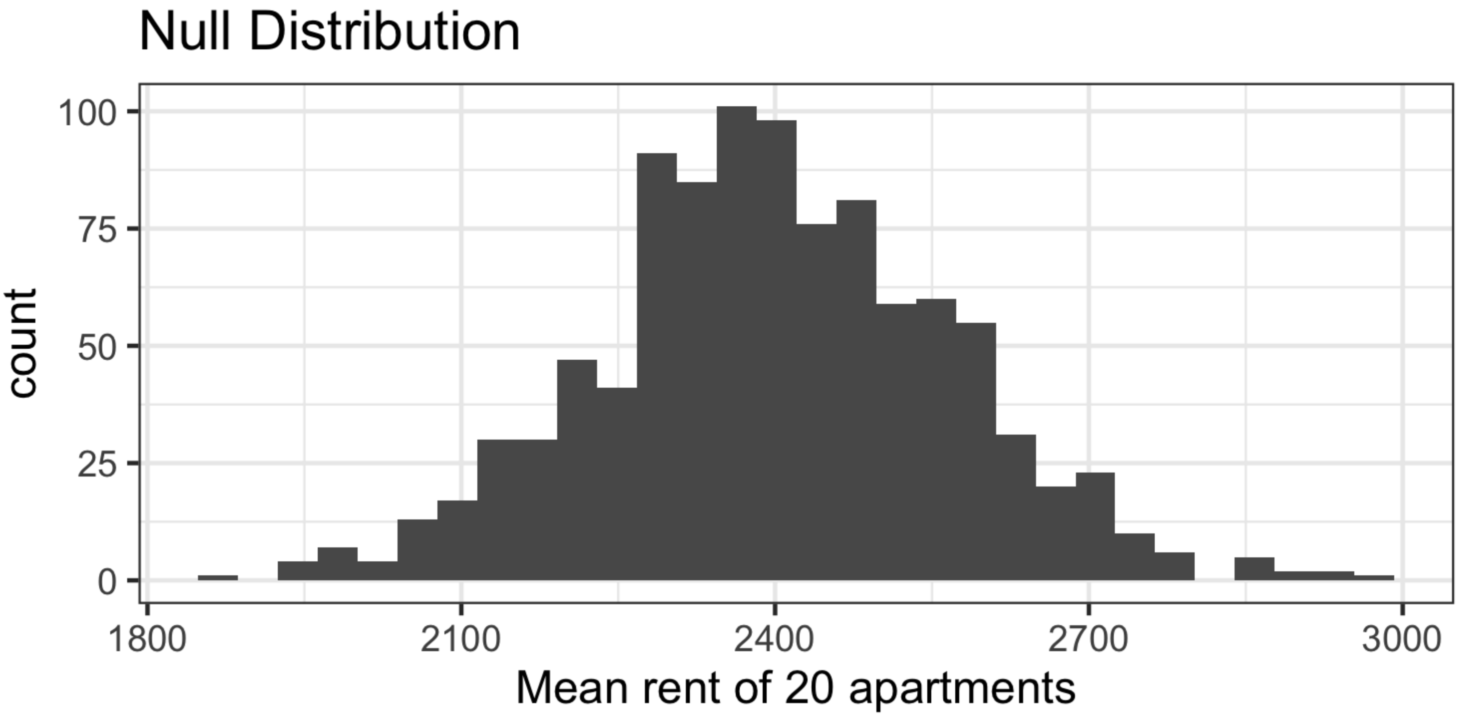
Where is the center of the distribution? What should it be under the null hypothesis?

Simulation: Shift Distribution

Shift the bootstrap distribution to be centered at the null value by subtracting/adding the difference between the center of the bootstrap distribution and the null value to each bootstrap mean.

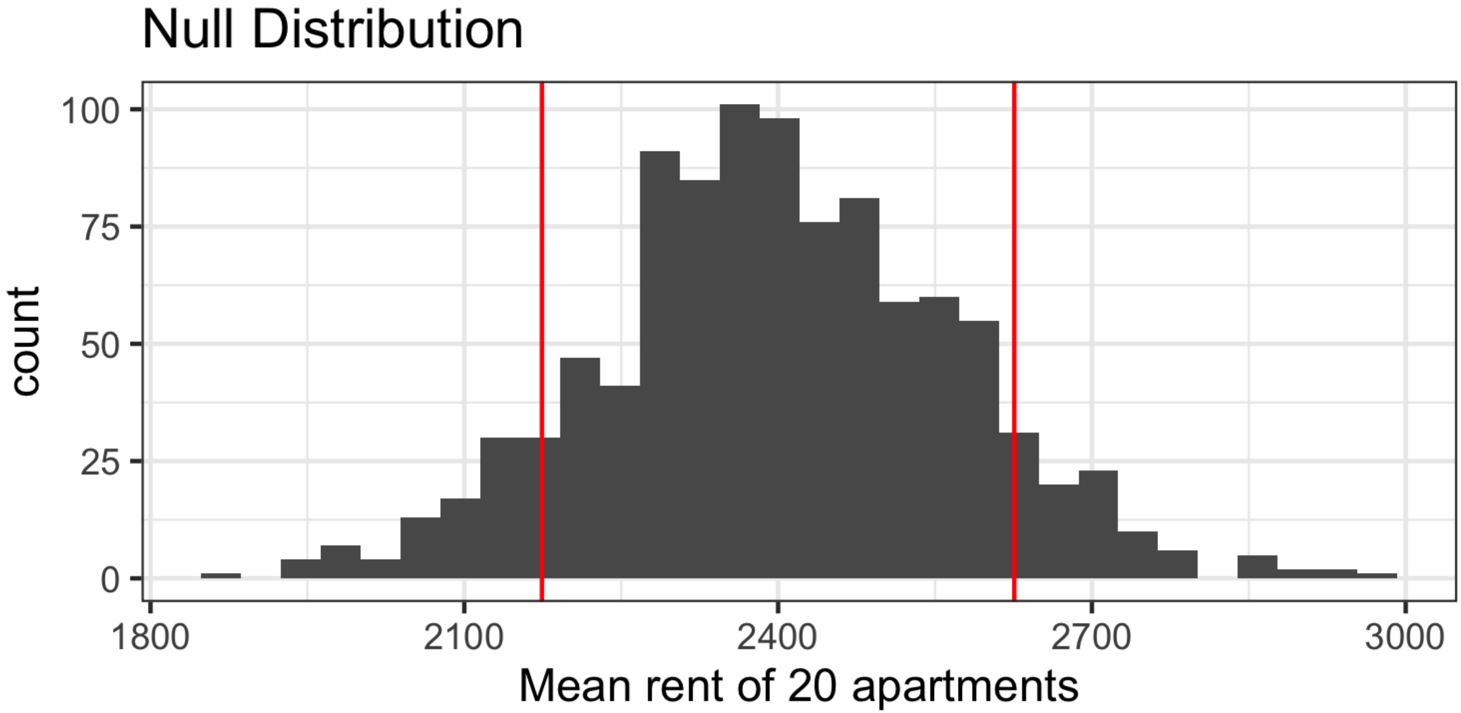
```r
rent_boot_mean <- rent_bootstrap %>%
  summarise(mean = mean(stat)) %>% pull()


null_dist <- rent_bootstrap %>%
  mutate(null_dist_stat = stat - (rent_boot_mean - 2400))
```

Simulation: Shift Distribution



Rent in Manhattan

Step 4: Use the null distribution, observed sample statistic, and alternative hypothesis to calculate the p-value.



Rent in Manhattan

Step 4: Use the null distribution, observed sample statistic, and alternative hypothesis to calculate the p-value.

```
p_val <- null_dist %>%
  filter(null_dist_stat >= 2625.8) %>%
  summarise(pval = 2 * n() / nrow(null_dist)) %>% pull()
p_val
```

```
## [1] 0.172
```

Rent in Manhattan

The p-value is 0.172. Use a significance level of $\alpha = 0.05$ to complete steps 5 and 6.

Step 5: Compare the p-value to the significance level α to make a conclusion (reject or fail to reject H_0).

Step 6: State the conclusion in the context of the data.

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

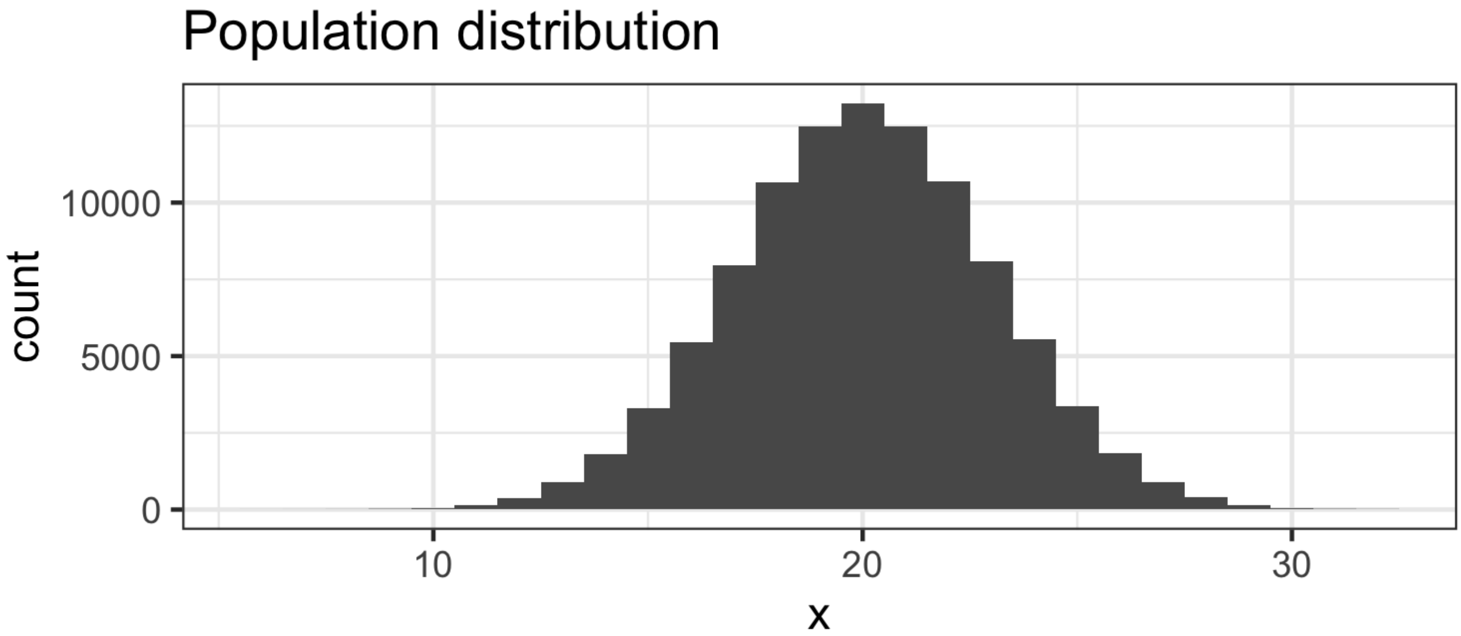
What is the difference between the sampling distribution and bootstrap distribution?

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

```r
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

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

```r
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
```

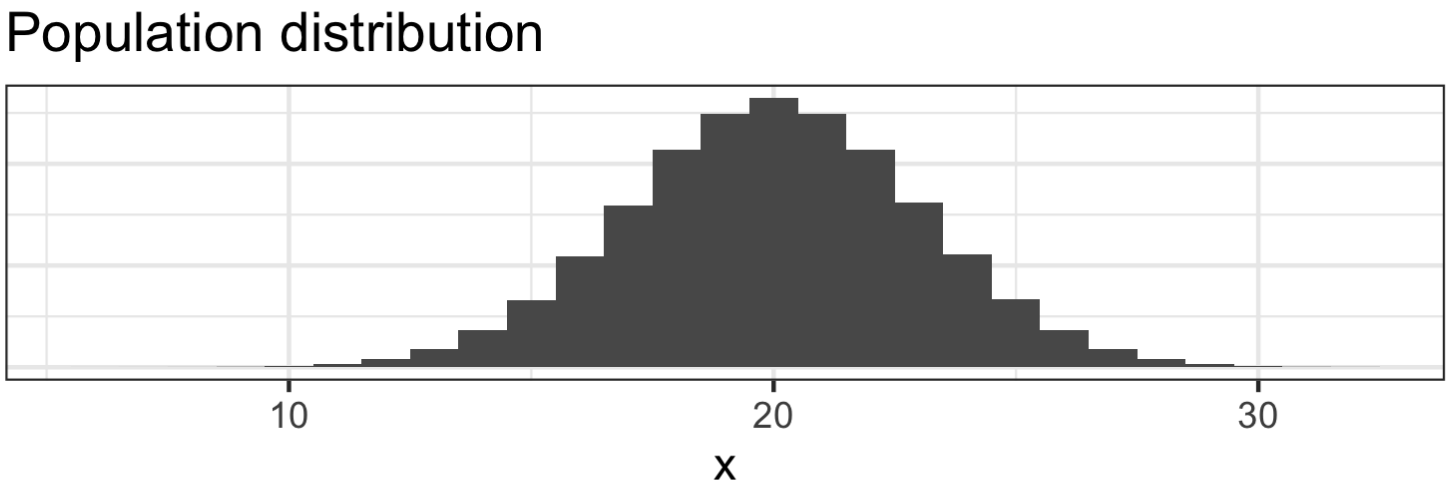
keep repeating...

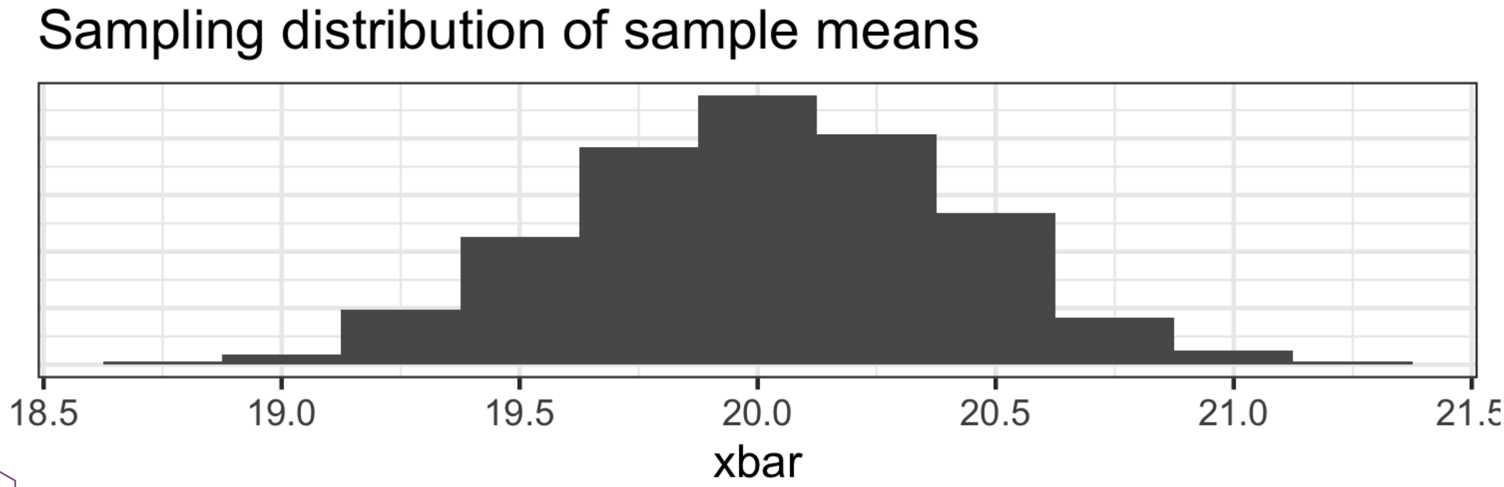
Sampling distribution

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

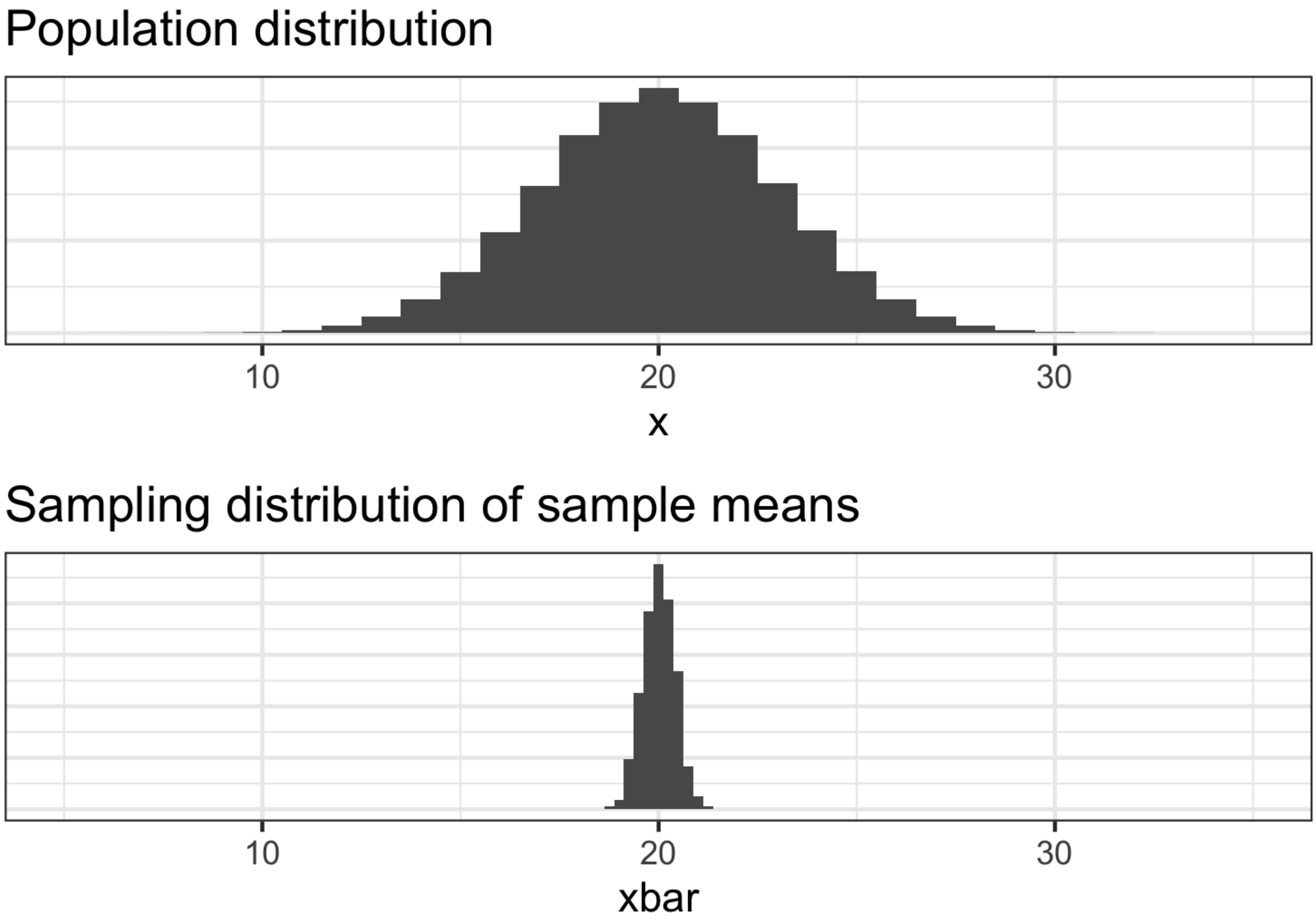
```
## # A tibble: 1,000 x 2
##    replicate  xbar
##        <int> <dbl>
##  1         1  19.4
##  2         2  20.9
##  3         3  20.4
##  4         4  19.5
##  5         5  19.9
##  6         6  19.6
##  7         7  19.8
##  8         8  20.4
##  9         9  20.4
## 10        10  19.4
## # … with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

Population vs. sampling





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))
```

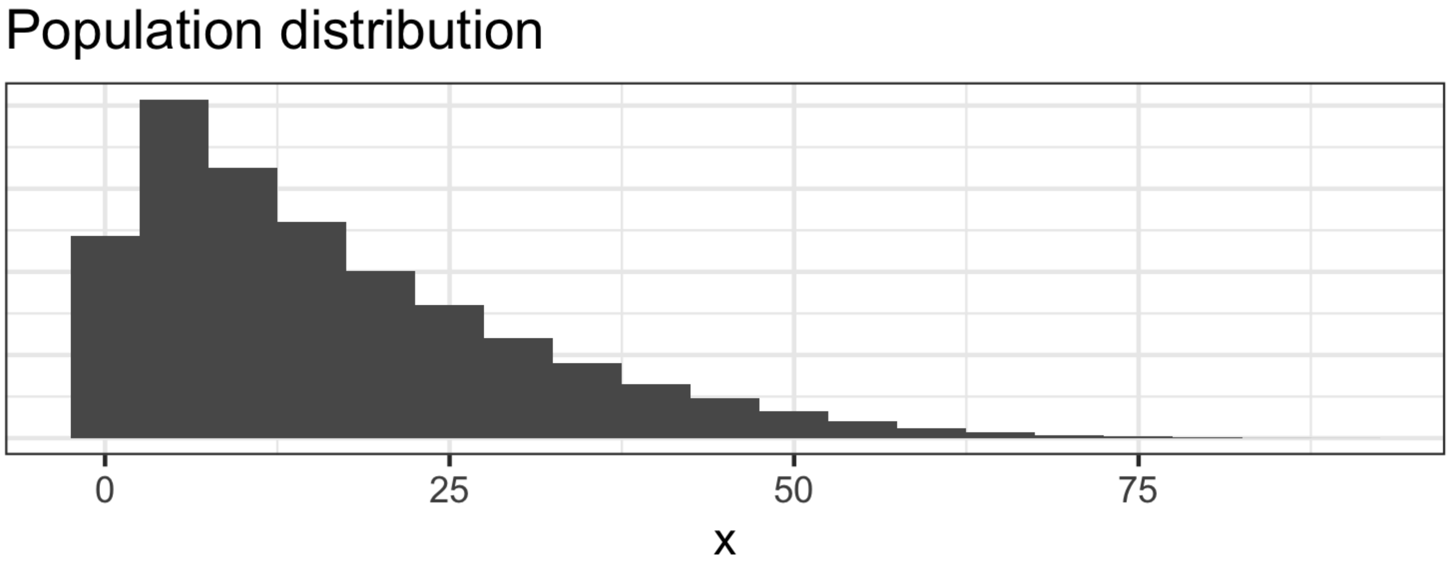
```
## # A tibble: 1 x 2
##      mu sigma
##   <dbl> <dbl>
## 1  20.0  3.00
```

```
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

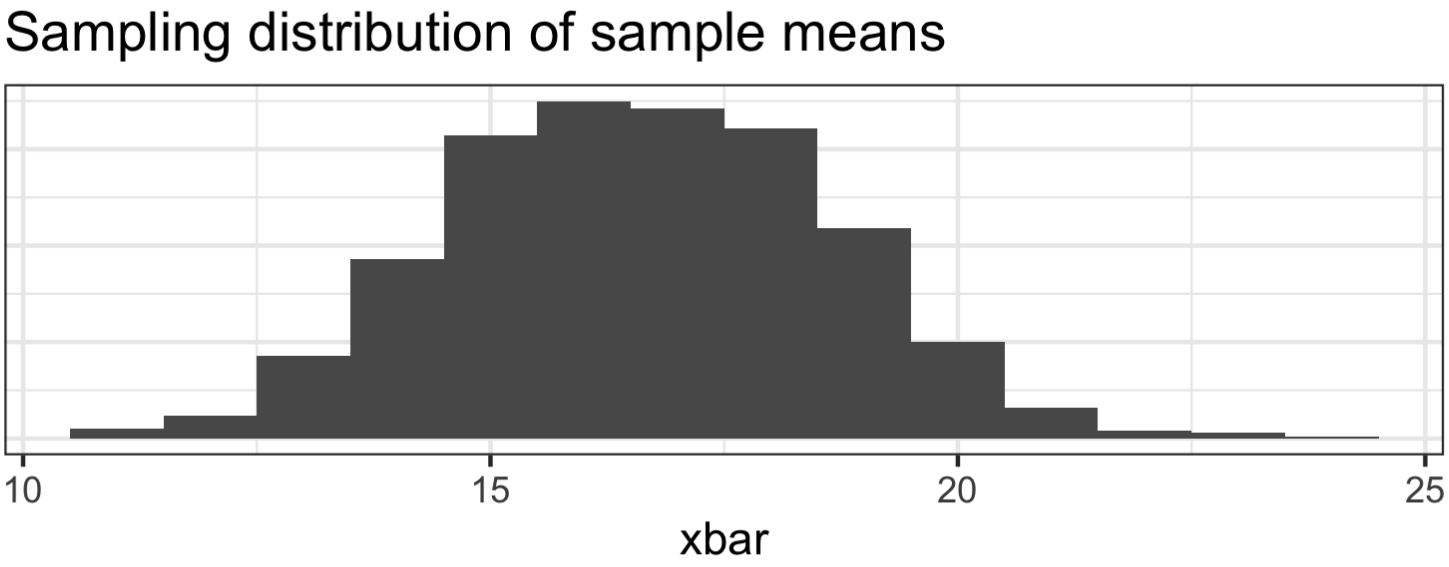
```
rs_pop <- tibble(x = rbeta(100000, 1, 5) * 100)
```



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

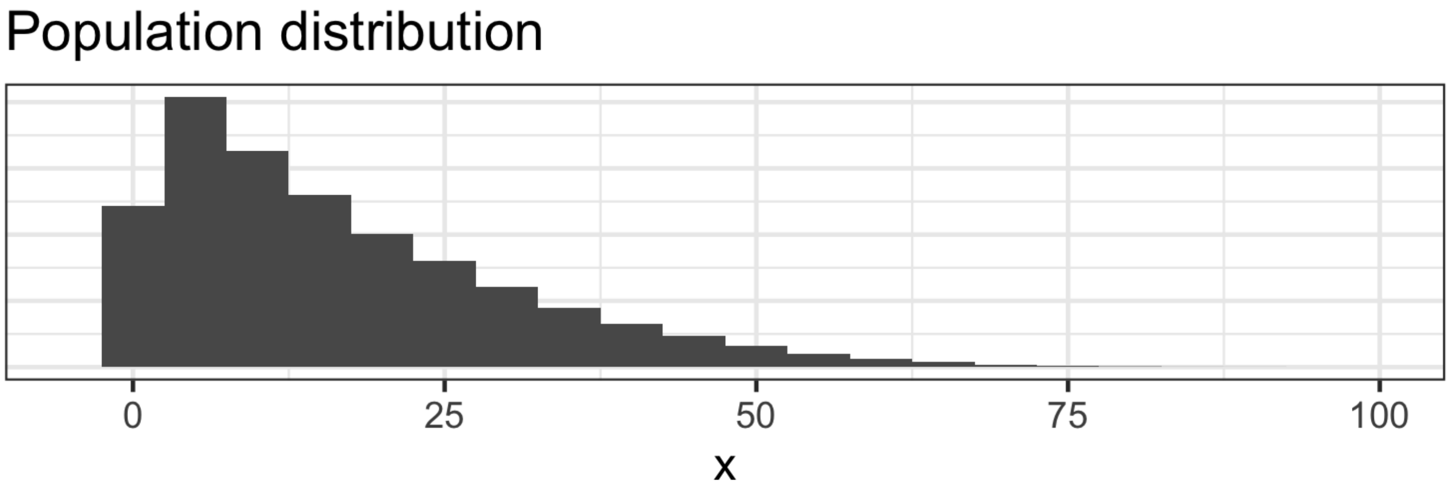
Sampling distribution

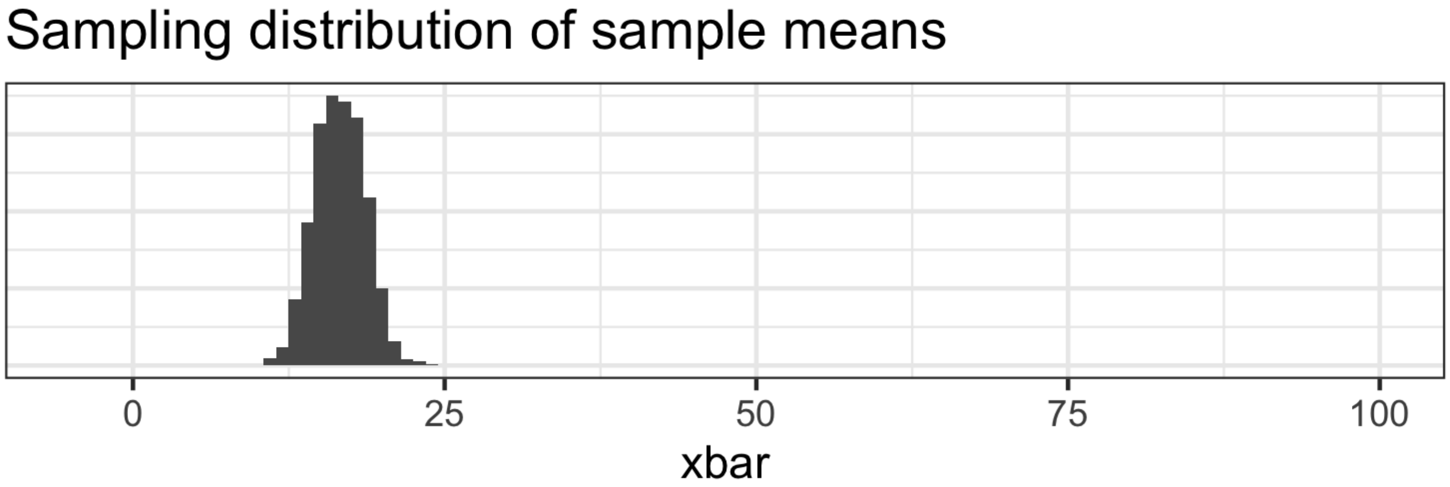
```
sampling <- rs_pop %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 1000) %>%
  group_by(replicate) %>%
  summarise(xbar = mean(x))
```



```
## # 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?


Population distribution


Sampling distribution of sample means

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?