# Parameters and confidence intervals

FOUNDATIONS OF INFERENCE

**Jo Hardin**
Instructor

# Research questions

| Hypothesis test | Confidence interval |
|---|---|
| Under which diet plan will participants lose more weight on average? | How much should participants expect to lose on average? |
| Which of two car manufacturers are users more likely to recommend to their friends? | What percent of users are likely to recommend Subaru to their friends? |
| Are education level and average income linearly related? | For each additional year of education, what is the predicted average income? |

# Parameter

- A parameter is a numerical value from the population
- Examples (continued):
  - The true average amount all dieters will lose on a particular program
  - The proportion of individuals in a population who recommend Subaru cars
  - The average income of all individuals in the population with a particular education level

# Confidence interval

- Range of numbers that (hopefully) captures the true parameter
- "95% confident that between 12% and 34% of the entire population recommends Subarus"

# Let's practice!

FOUNDATIONS OF INFERENCE

# Bootstrapping

FOUNDATIONS OF INFERENCE

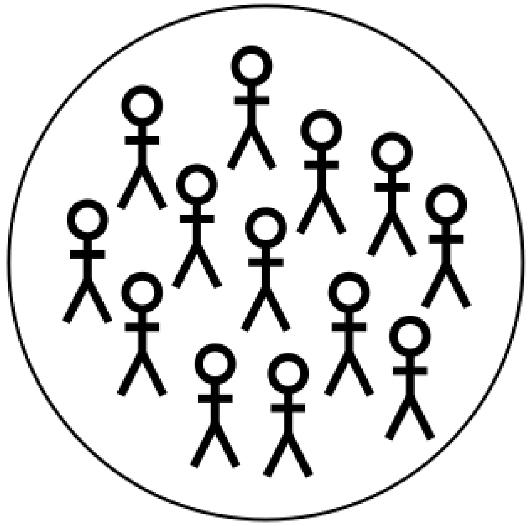
**Jo Hardin**
Instructor

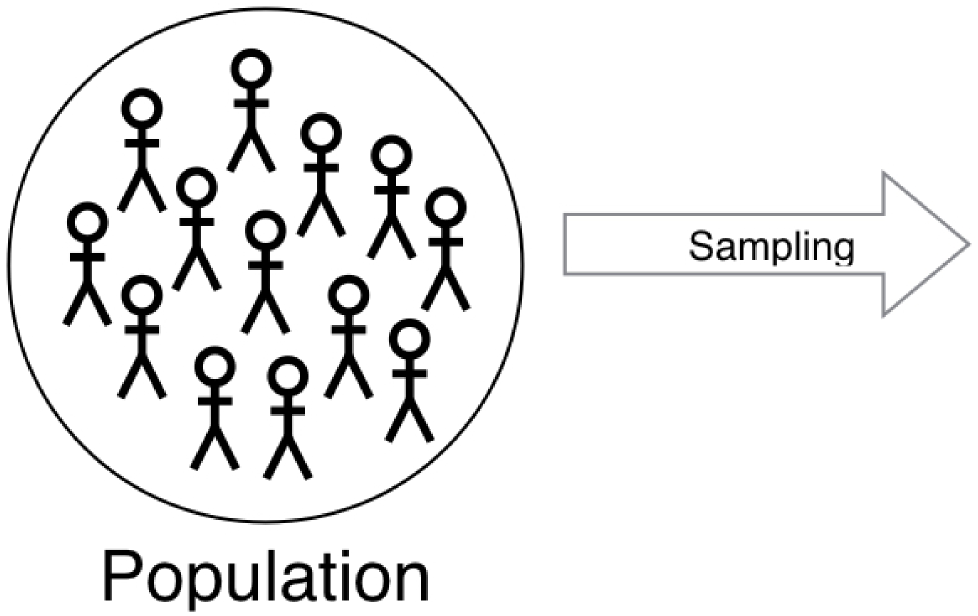
# Hypothesis testing

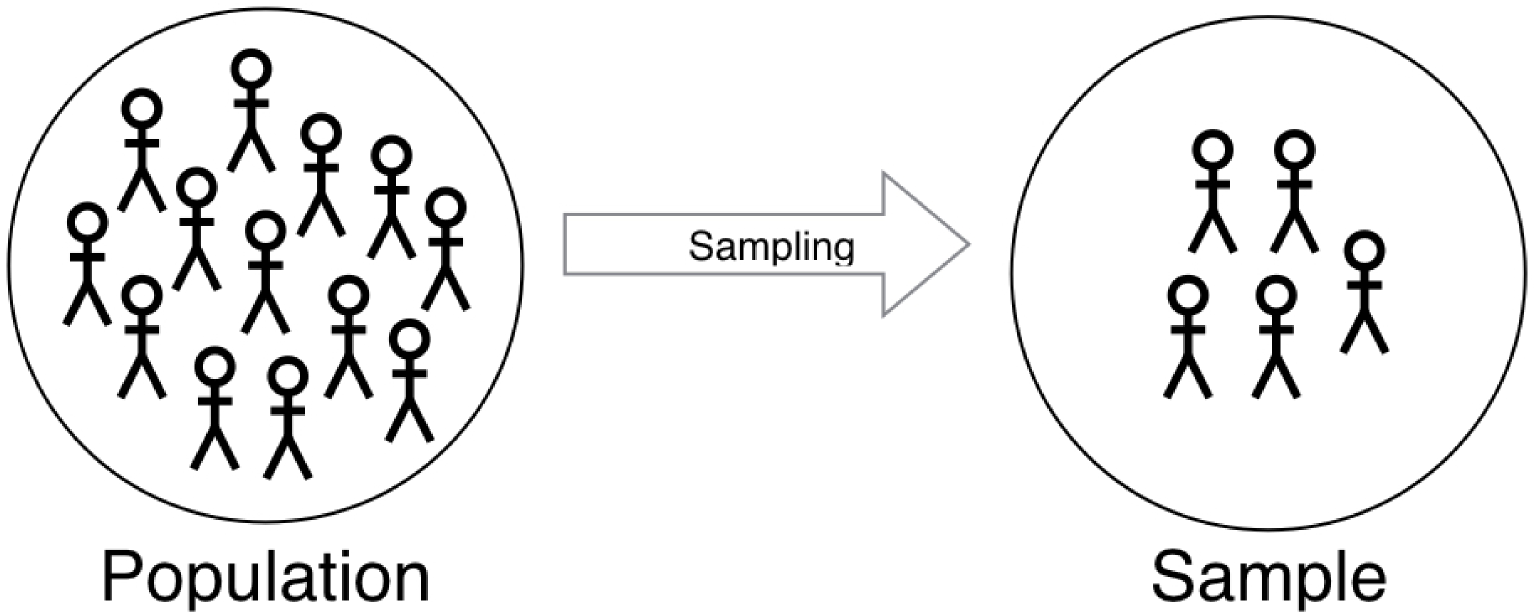
- How do samples from the null population vary?
- **Statistic**, proportion of successes in *sample* $\rightarrow \hat{p}$
- **Parameter**, proportion of successes in *population* $\rightarrow p$

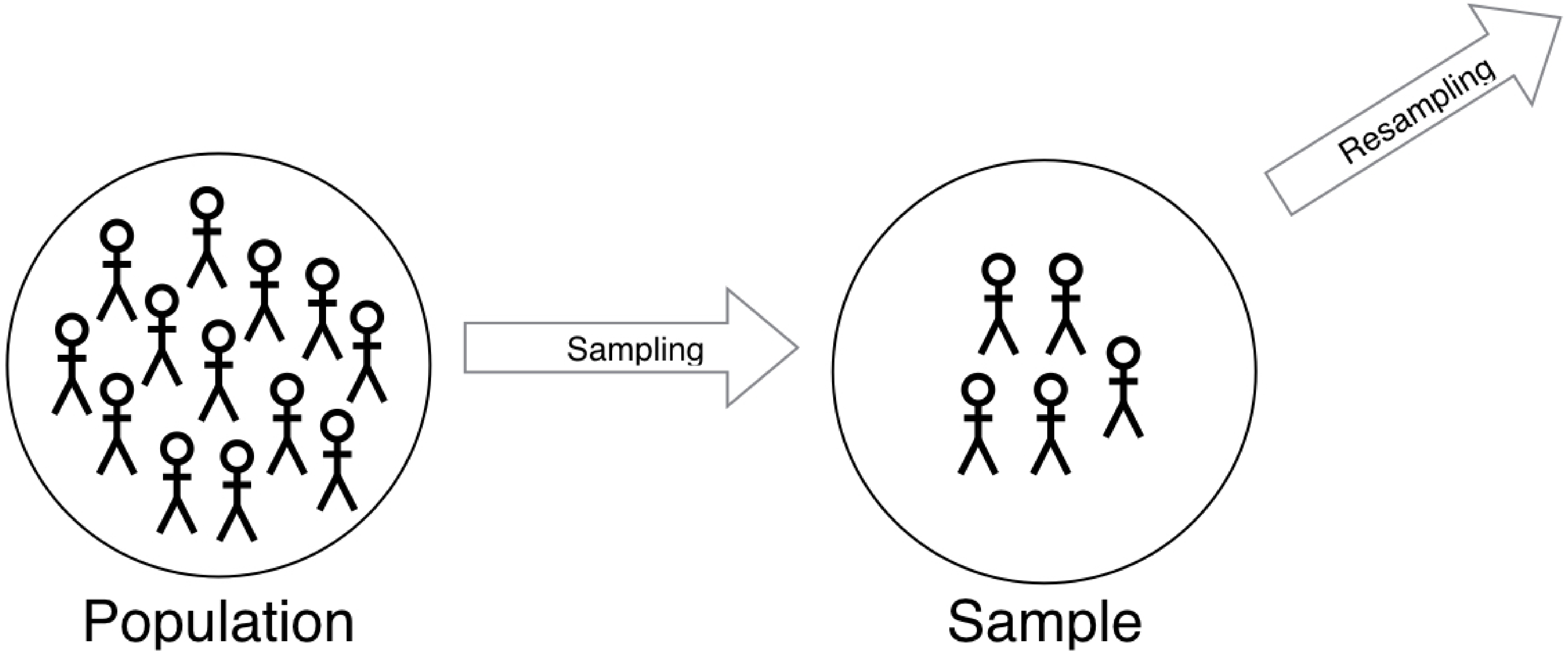
# Confidence intervals

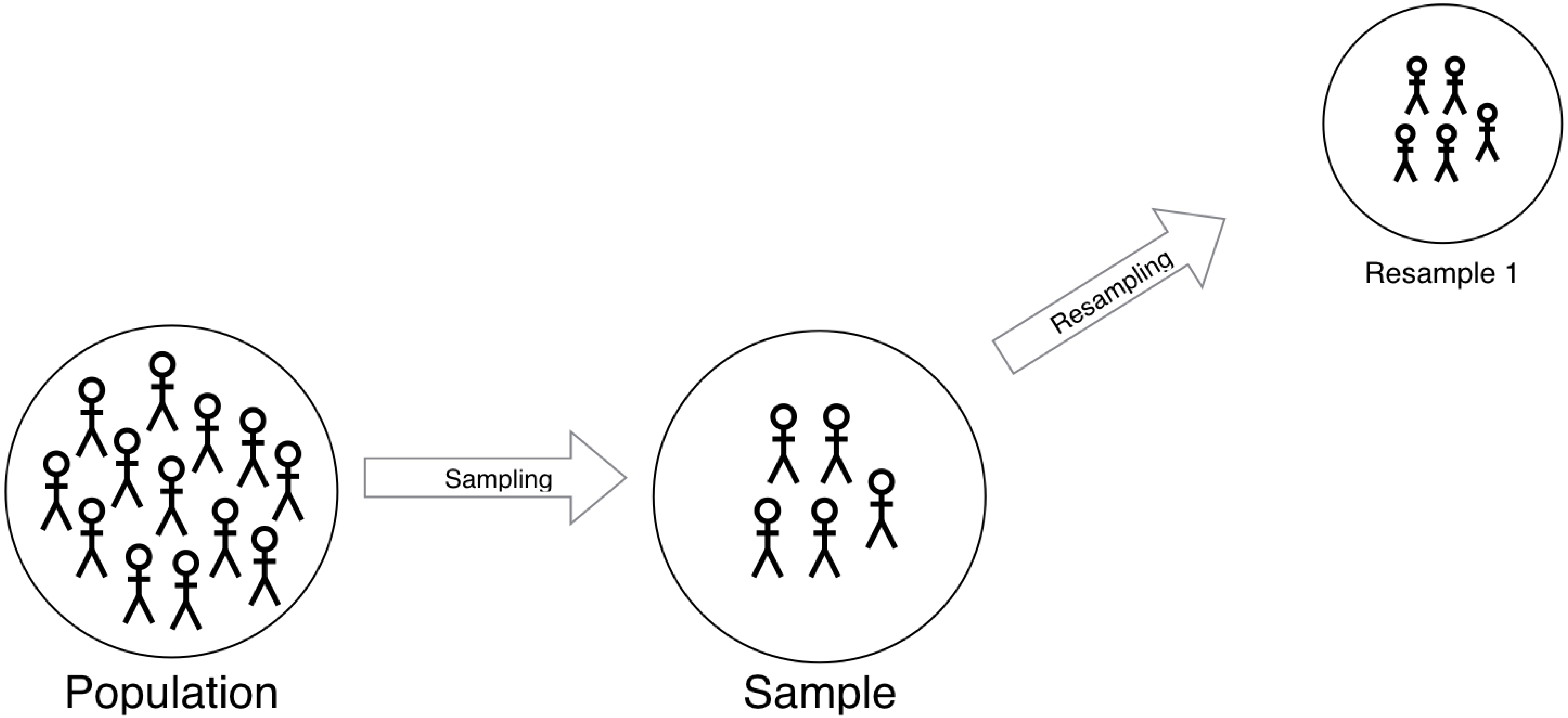
- No null population, unlike in hypothesis testing
- How do $p$ and $\hat{p}$ vary?

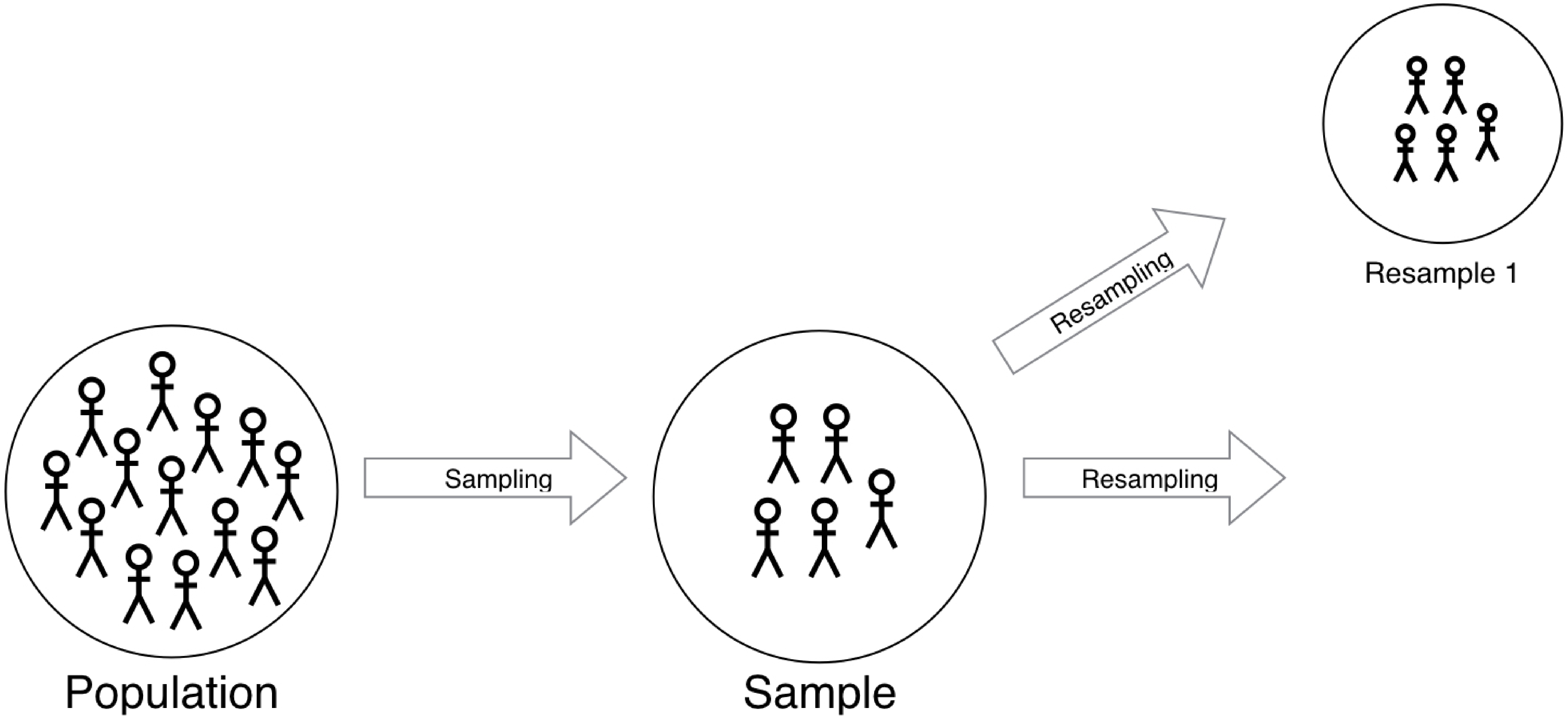

Population

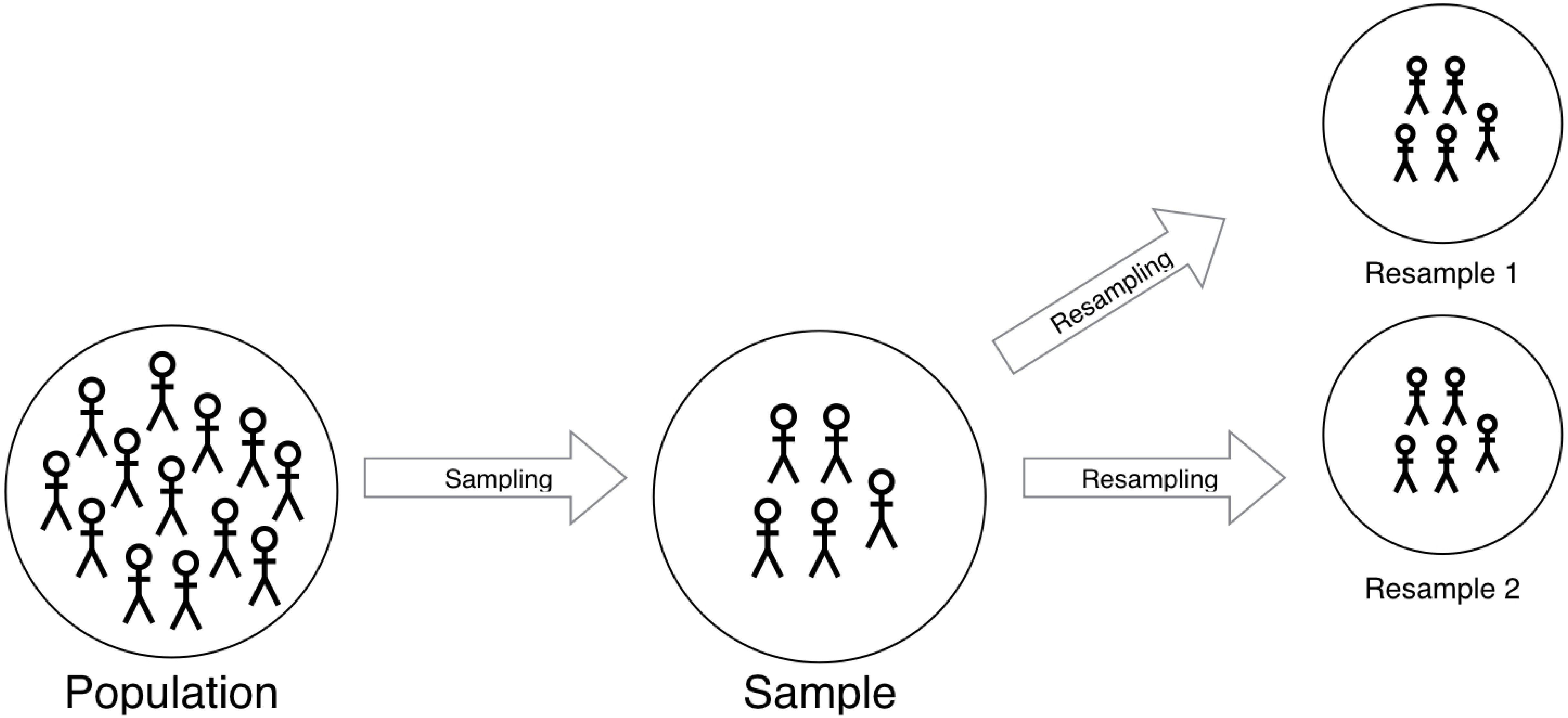

Sampling
Population

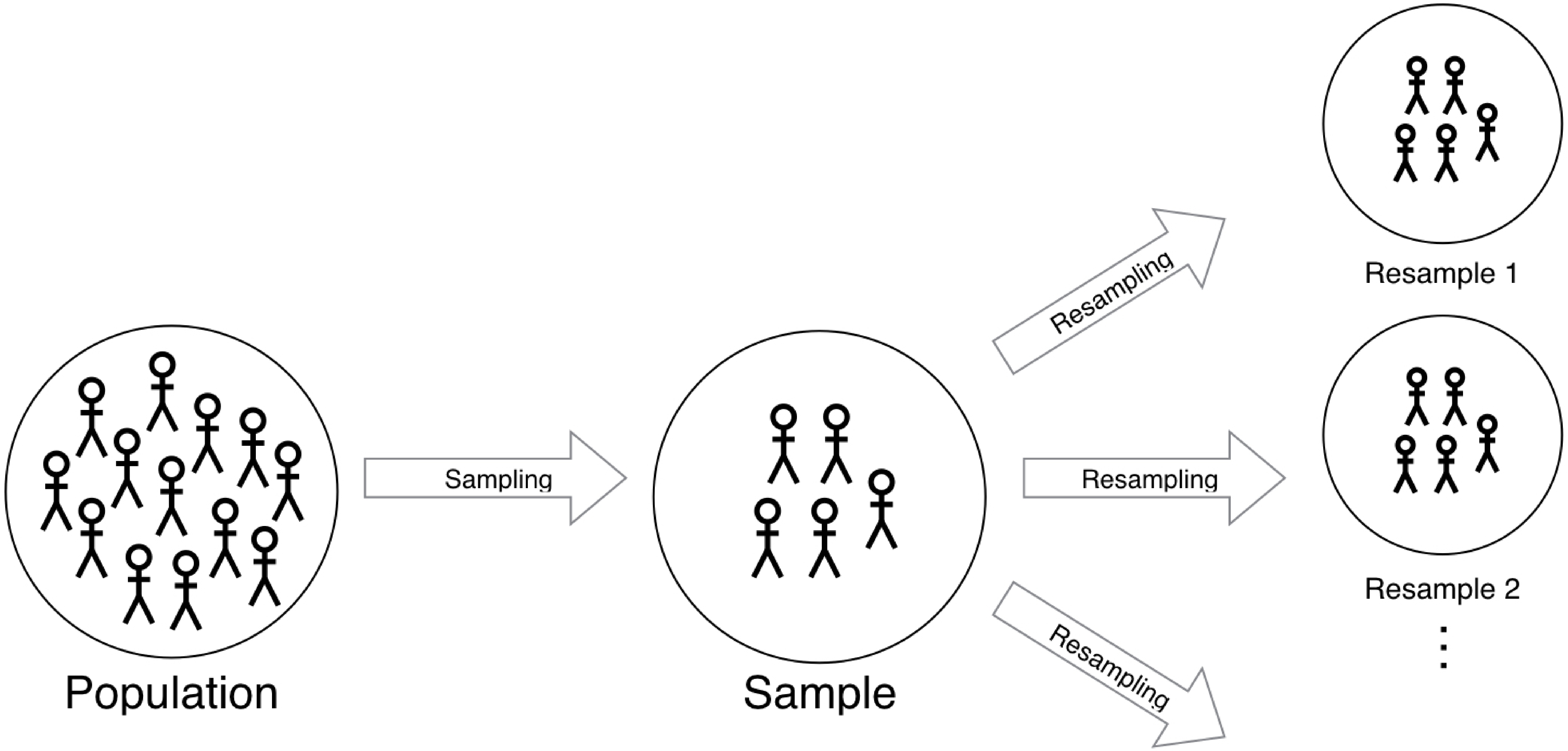

Sampling
Population
Sample

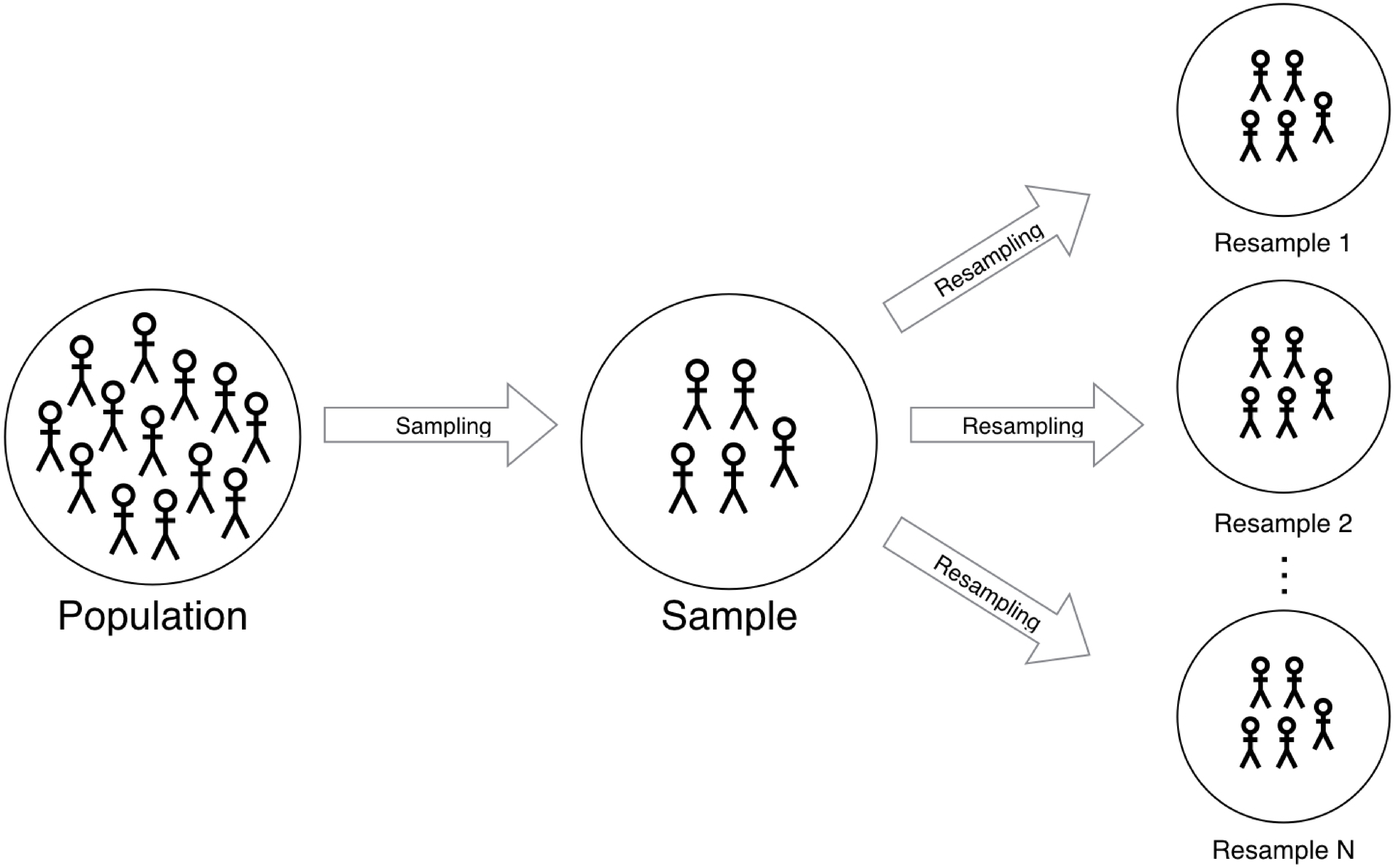

Resampling
Sampling
Population
Sample


Resample 1
Resampling
Sampling
Population
Sample


Resample 1
Resampling
Sampling
Resampling
Population
Sample


Resample 1
Resampling
Sampling
Resampling
Resample 2
Population
Sample


Sampling
Resampling
Resampling
Resampling
Resample 1
Resample 2
Population
Sample


Sampling
Resampling
Resampling
Resampling
Population
Sample
Resample 1
Resample 2
Resample N

# Polling

```
# Original data
Source: local data frame [30 x 3]
```

```
      flip_num  flip
         <int>  <chr>
1            1      H
2            2      H
3            3      H
4            4      T
5            5      H
6            6      H
# ... with 24 more rows
```

Original data

| Candidate X | Total voters | Proportion X |
|---|---|---|
| 17 | 30 | 0.5667 |

# Polling

```
# First resample
Source: local data frame [30 x 3]
```

```
   replicate flip_num  flip
       <dbl>    <int> <chr>
1          1        7     H
2          1       17     T
3          1       13     H
4          1       14     H
5          1       24     H
6          1       28     T
# ... with 24 more rows
```

First resample

| Candidate X | Total voters | Proportion X |
|---|---|---|
| 17 | 30 | 0.5667 |
| 14 | 30 | 0.4667 |

# Polling

```
# Second resample
Source: local data frame [30 x 3]
```

```
   replicate flip_num  flip
       <dbl>    <int> <chr>
1          2       21     H
2          2       19     T
3          2       25     H
4          2       24     T
5          2       21     H
6          2       28     T
7          2       13     H
8          2       23     H
9          2       24     T
10         2       24     T
# ... with 20 more rows
```

Second resample

| Candidate X | Total voters | Proportion X |
|---|---|---|
| 17 | 30 | 0.5667 |
| 14 | 30 | 0.4667 |
| 18 | 30 | 0.6 |

# Polling

```
# Third resample
Source: local data frame [30 x 3]
```

```
   replicate flip_num  flip
       <dbl>    <int> <chr>
1          3        6     H
2          3       19     H
3          3        1     H
4          3       24     T
5          3       11     H
6          3       28     T
7          3       16     H
8          3       13     H
9          3       21     T
10         3       29     H
# ... with 20 more rows
```

Third resample

| Candidate X | Total voters | Proportion X |
|---|---|---|
| 17 | 30 | 0.5667 |
| 14 | 30 | 0.4667 |
| 18 | 30 | 0.6 |
| 12 | 30 | 0.4 |

# Standard error

- Obtained standard error of 0.09 by resampling many times
- Describes how the statistic varies around parameter
- Bootstrap provides an approximation of the standard error

# Variability of p-hat from the population

```
# Compute p-hat for each poll
ex1_props <- recommend %>%
    group_by(poll) %>%
    summarize(prop_yes =
                mean(vote == "yes"))
```

```
# Variability of p-hat
ex1_props %>%
    summarize(sd(prop_yes))
```

```
# A tibble: 1 × 1
  `sd(prop_yes)`
           <dbl>
1     0.08523512
```

# Variability of p-hat from the sample (bootstrapping)

```
# Select one poll from which to resample
one_poll <- all_polls %>%
    filter(poll ==1) %>%
    select(vote)

# Compute p-hat for each resampled poll
ex2_props <- one_poll %>%
    specify(response = vote,
            success = "yes") %>%
    generate(reps = 1000,
            type = "bootstrap")
```

```
# Variability of p-hat
ex2_props %>%
    summarize(sd(stat))
```

```
# A tibble: 1 × 1
  `sd(stat)`
       <dbl>
1   0.08691885
```

# Let's practice!

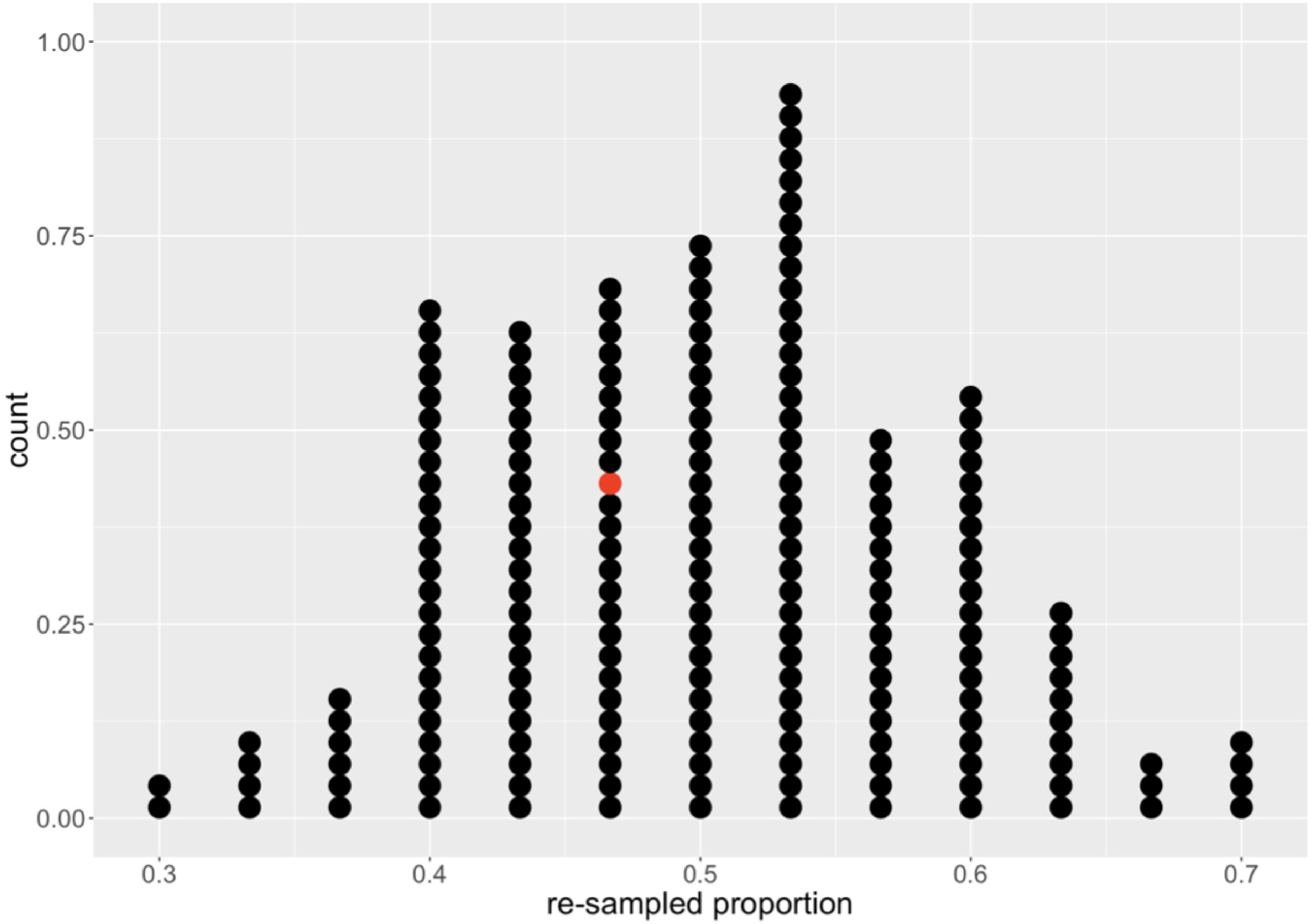
FOUNDATIONS OF INFERENCE

# Variability in p-hat

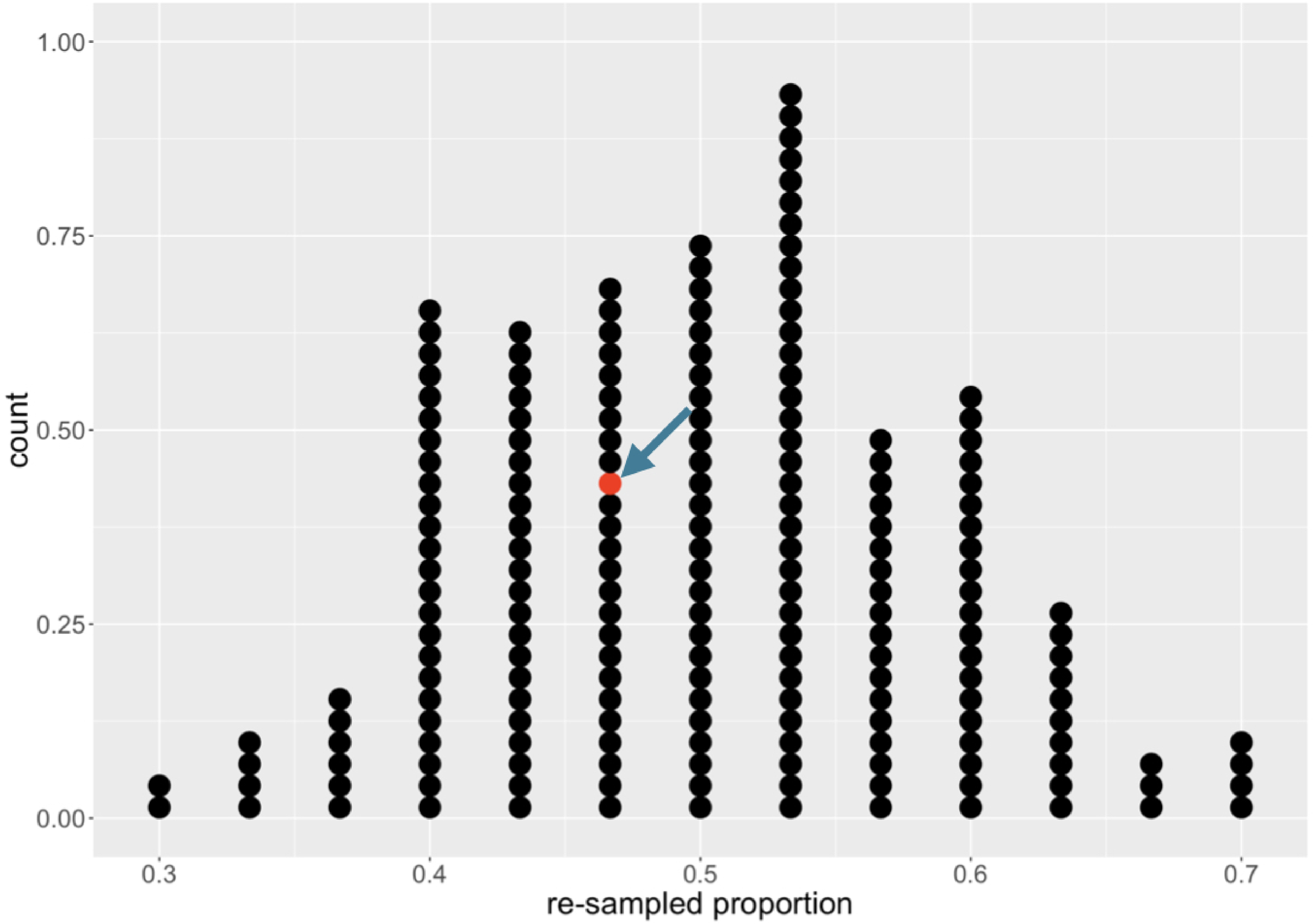
FOUNDATIONS OF INFERENCE

Jo Hardin
Instructor

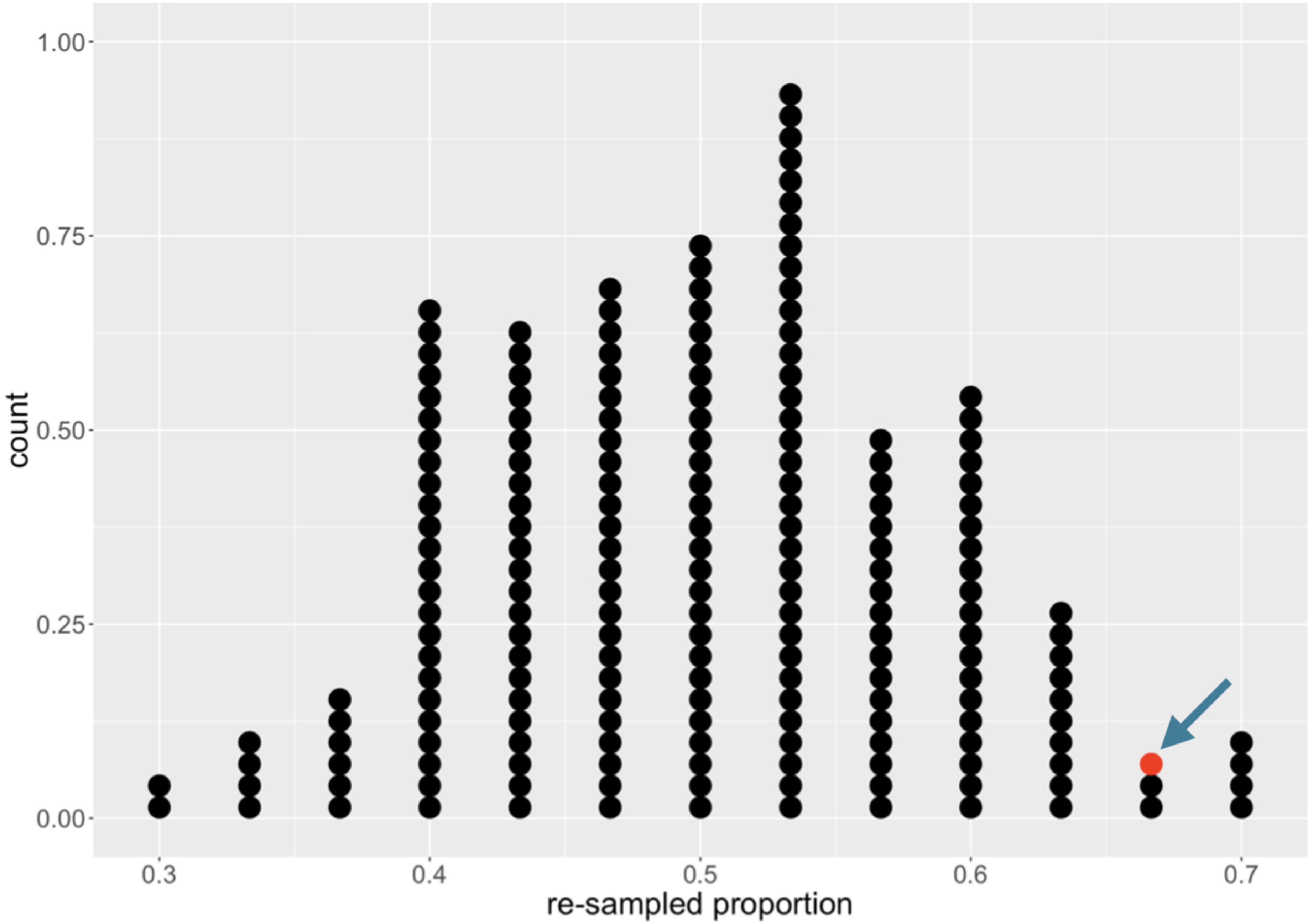
# How far are the data from the parameter?



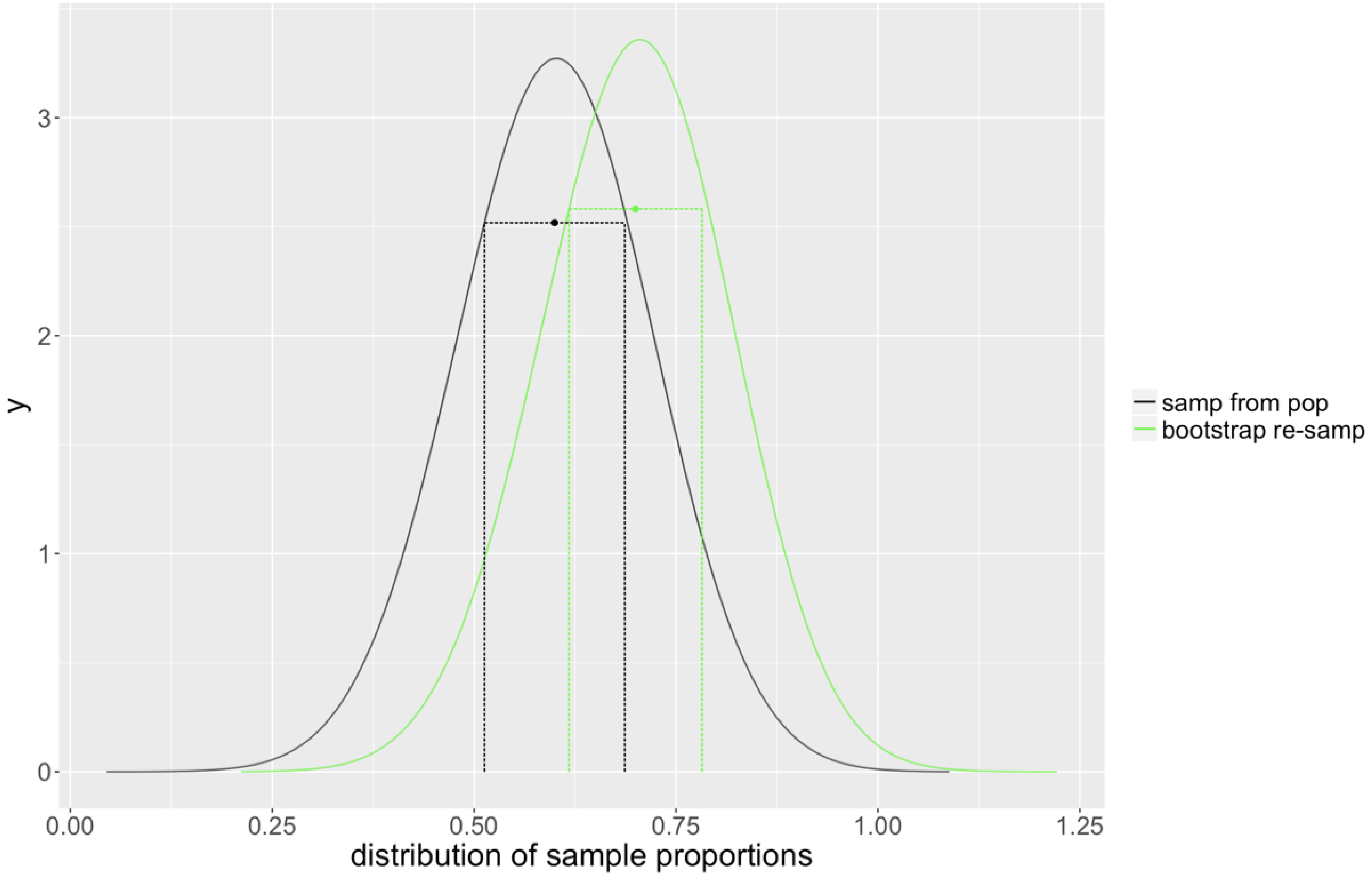
# How far are the data from the parameter?



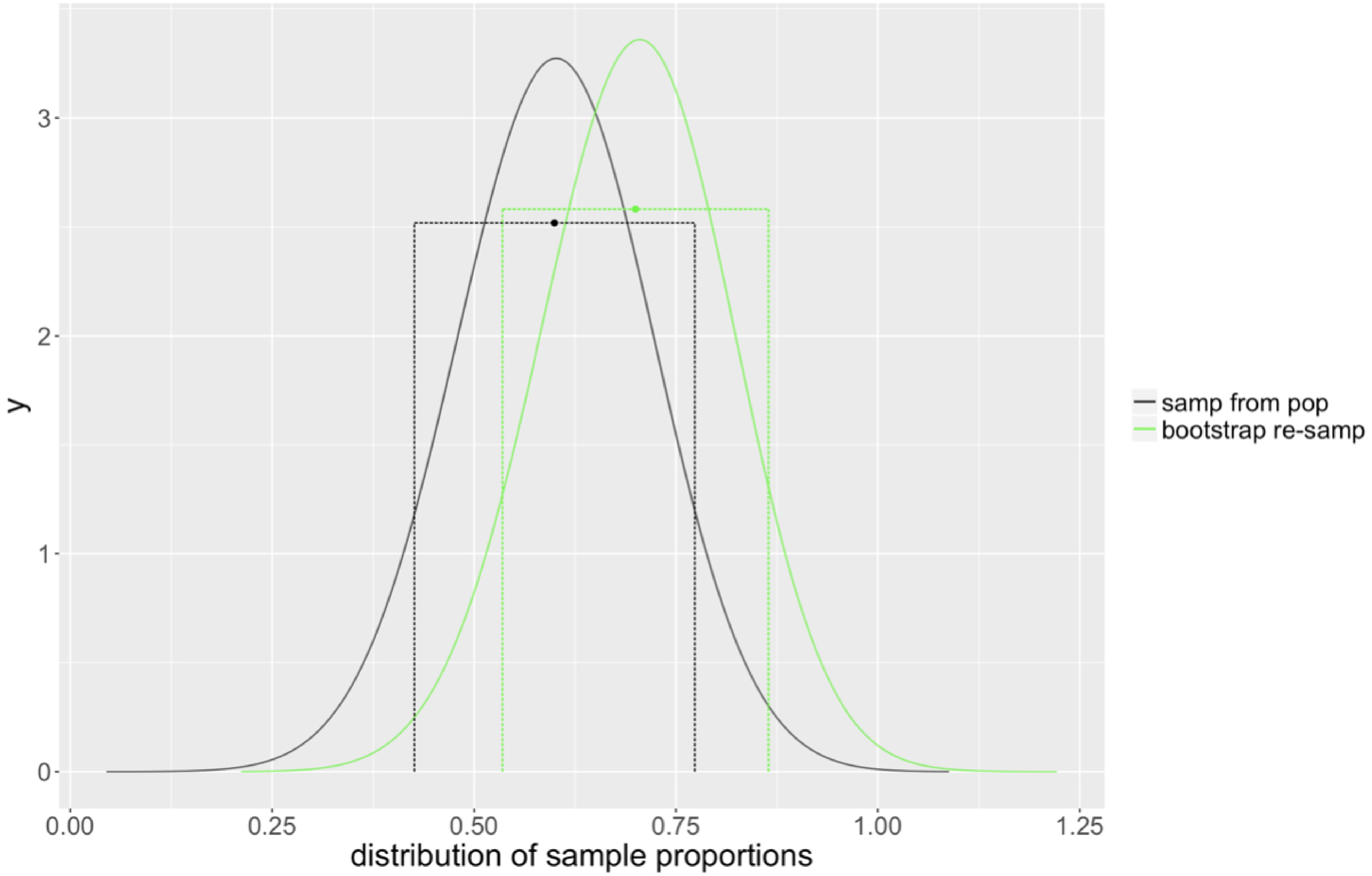
# How far are the data from the parameter?



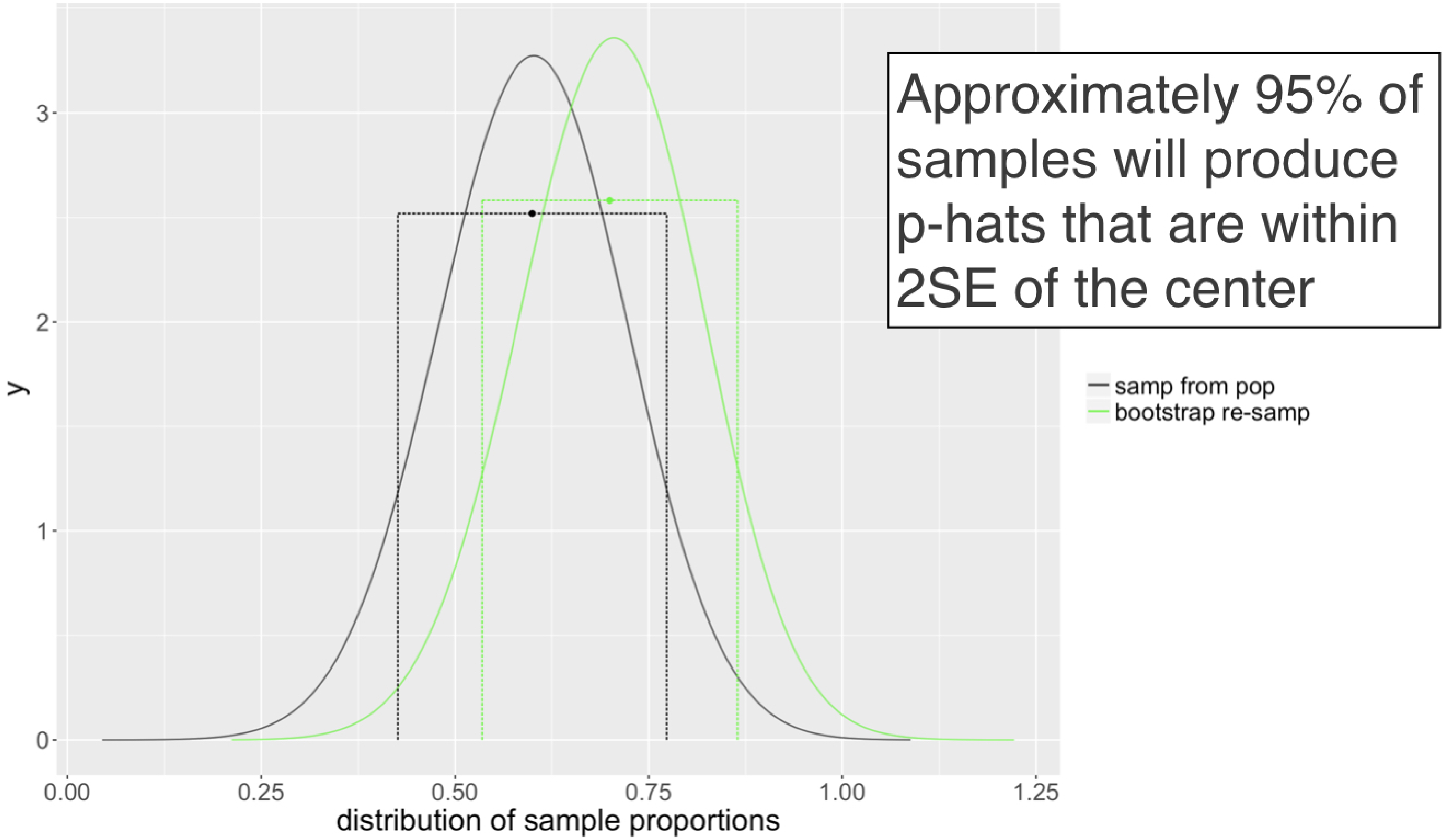
# Standard error of p-hat



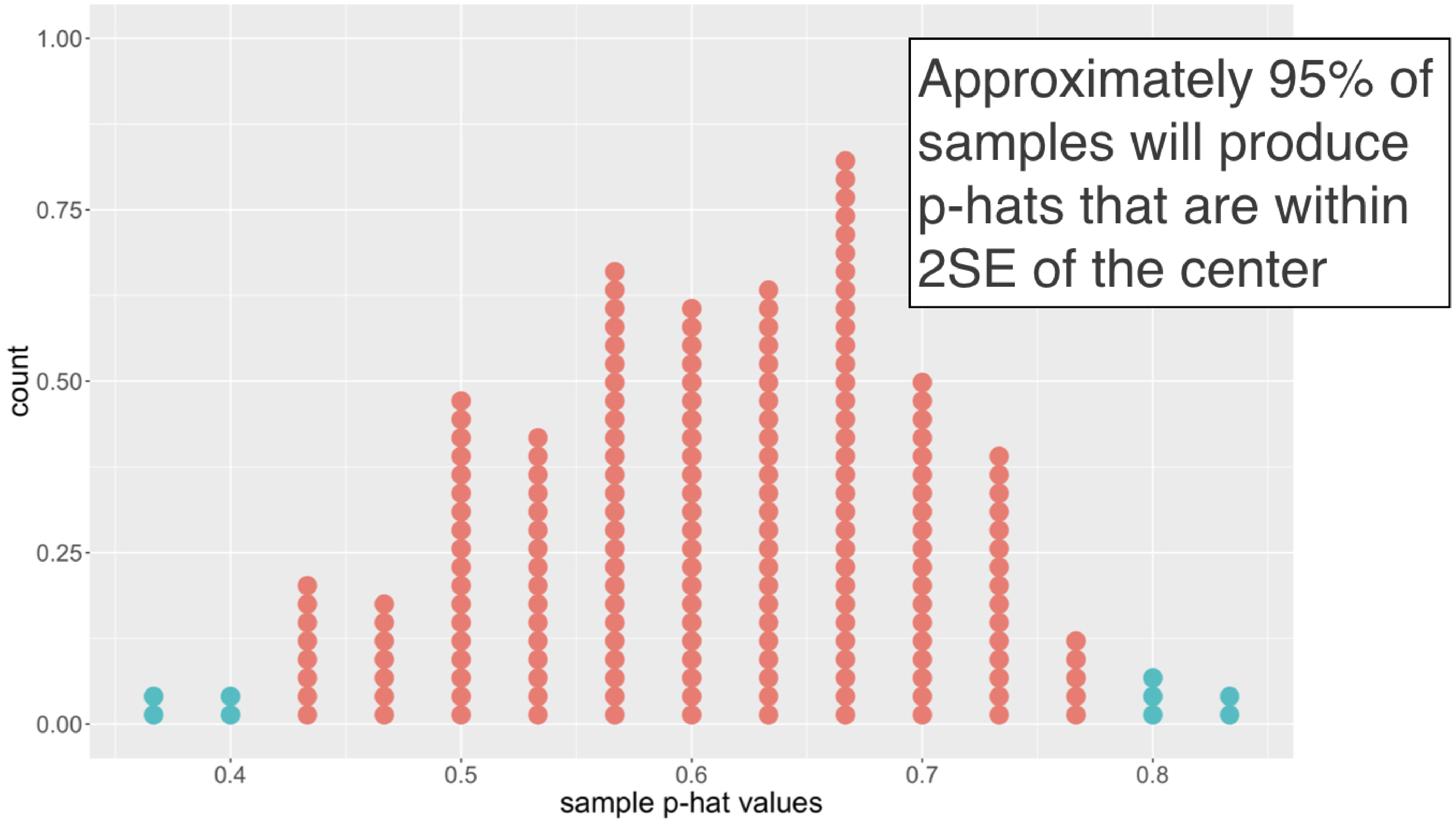
# Empirical rule



# Empirical rule



# Empirical rule



# Let's practice!

FOUNDATIONS OF INFERENCE

# Interpreting CIs and technical conditions

FOUNDATIONS OF INFERENCE

**Jo Hardin**
Instructor

# Creating CIs

```
# Compare confidence intervals
one_poll_boot %>% summarize(
    lower = p_hat - 2 *
            sd(prop_yes_boot),
    upper = p_hat + 2 *
            sd(prop_yes_boot))
```

```
# A tibble: 1 × 2
     lower    upper
     <dbl>    <dbl>
1 0.536148 0.863852
```

```
# Find 2.5% and 97.5% of p-hat vals
one_poll_boot %>% summarize(
    q025_prop = quantile(prop_yes_boot,
                         p = .025),
    q975_prop = quantile(prop_yes_boot,
                         p = .975))
```

```
# A tibble: 1 × 2
  q025_prop q975_prop
      <dbl>     <dbl>
1 0.5333333 0.8333333
```

# Motivating CIs

- Goal is to find the parameter when all we know is the statistic
- Never know whether the sample you collected actually contains the true parameter

# Interpreting the CIs

- Bootstrap t-CI: (0.536, 0.864)
- Percentile interval: (0.533, 0.833)

***We are 95% confident that the true proportion of people planning to vote for candidate X is between 0.536 and 0.864 (or 0.533 and 0.833)***

# Technical conditions

- Sampling distribution of the statistic is reasonably symmetric and bell-shaped
- Sample size is reasonably large
- Variability of resampled proportions

# Let's practice!

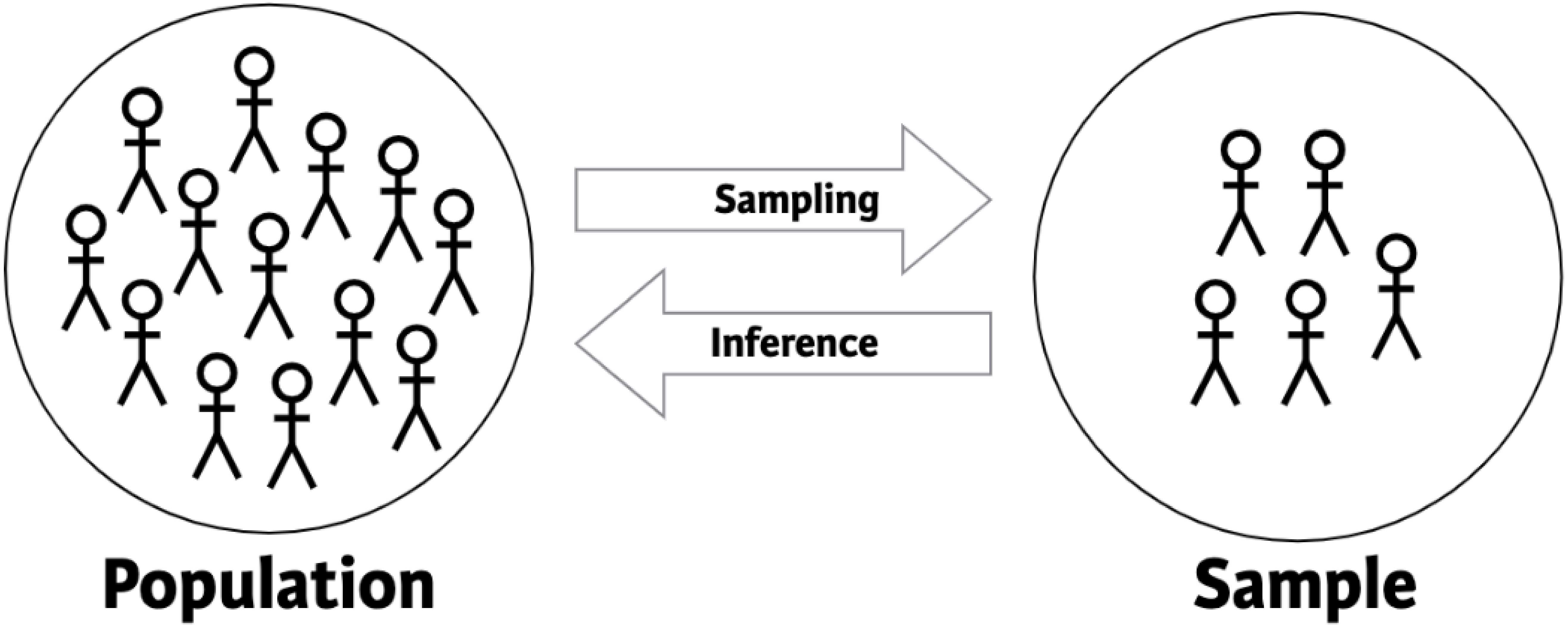
FOUNDATIONS OF INFERENCE

# Summary of statistical inference

FOUNDATIONS OF INFERENCE

**Jo Hardin**
Instructor

# Inference



# Testing

- $H_0$: There is no gender discrimination in hiring
- $H_A$: Men are more likely to be promoted than women

Test

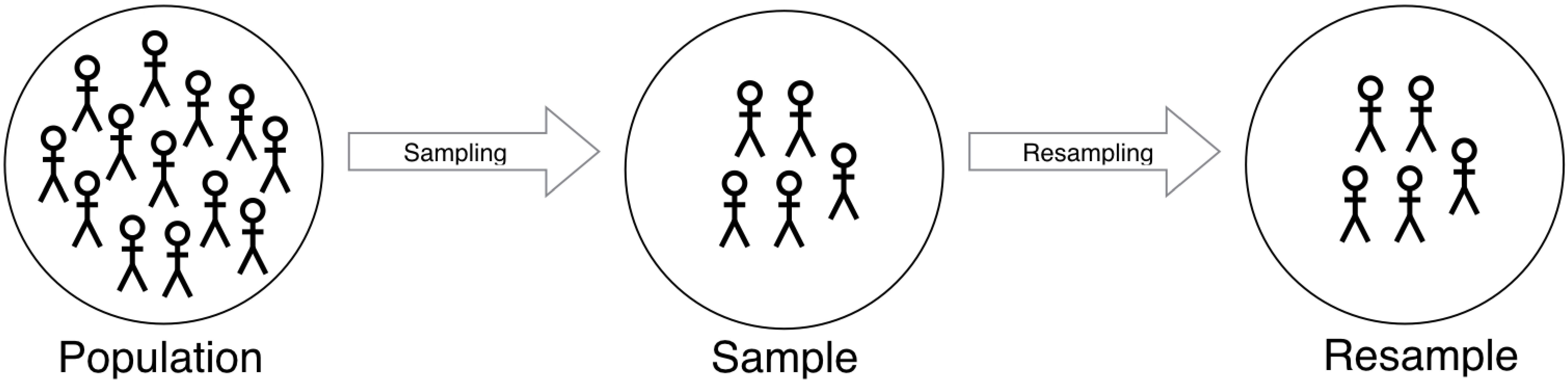
Truth

| | Do not reject $H_0$ | Reject $H_0$ in favor of $H_A$ |
|---|---|---|
| $H_0$ true | ✓ | Type I error |
| $H_A$ true | Type II error | ✓ |

# Estimation

**What proportion of the voters will select candidate X?**

# Bootstrapping



# Congratulations!

FOUNDATIONS OF INFERENCE