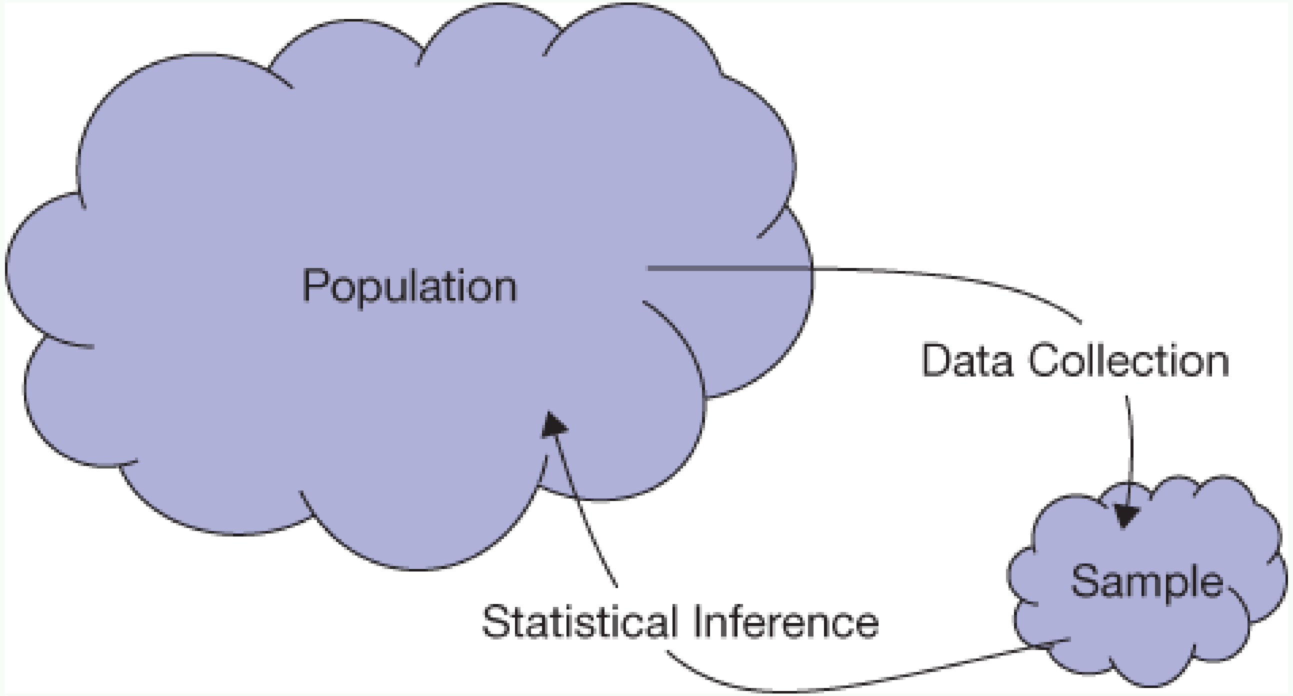
Sampling Distribution and Bootstrap

Stat 120

April 09 2023

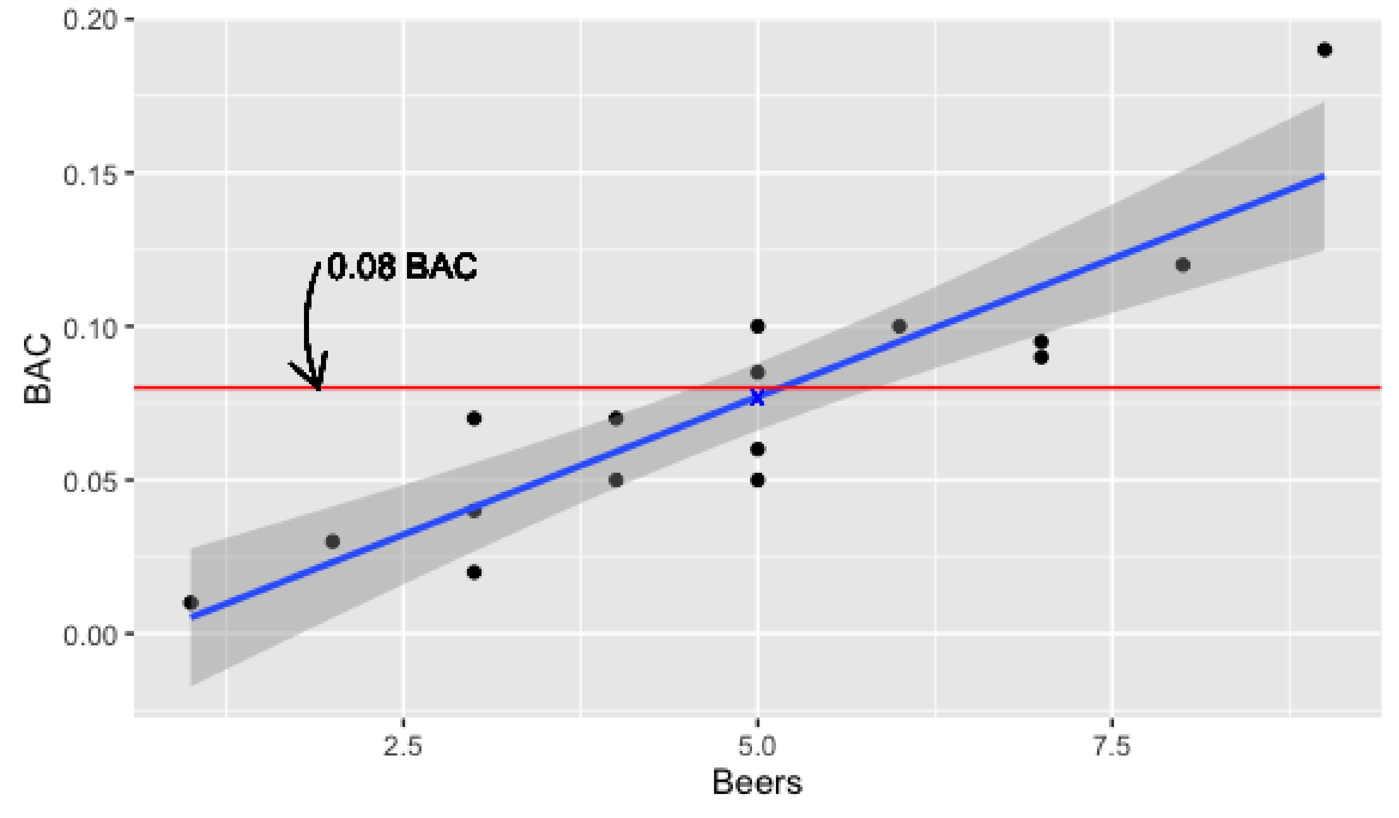
Statistical Inference

Statistical inference ***is the process of drawing conclusions about the entire population based on information in a sample.***



Statistical Inference

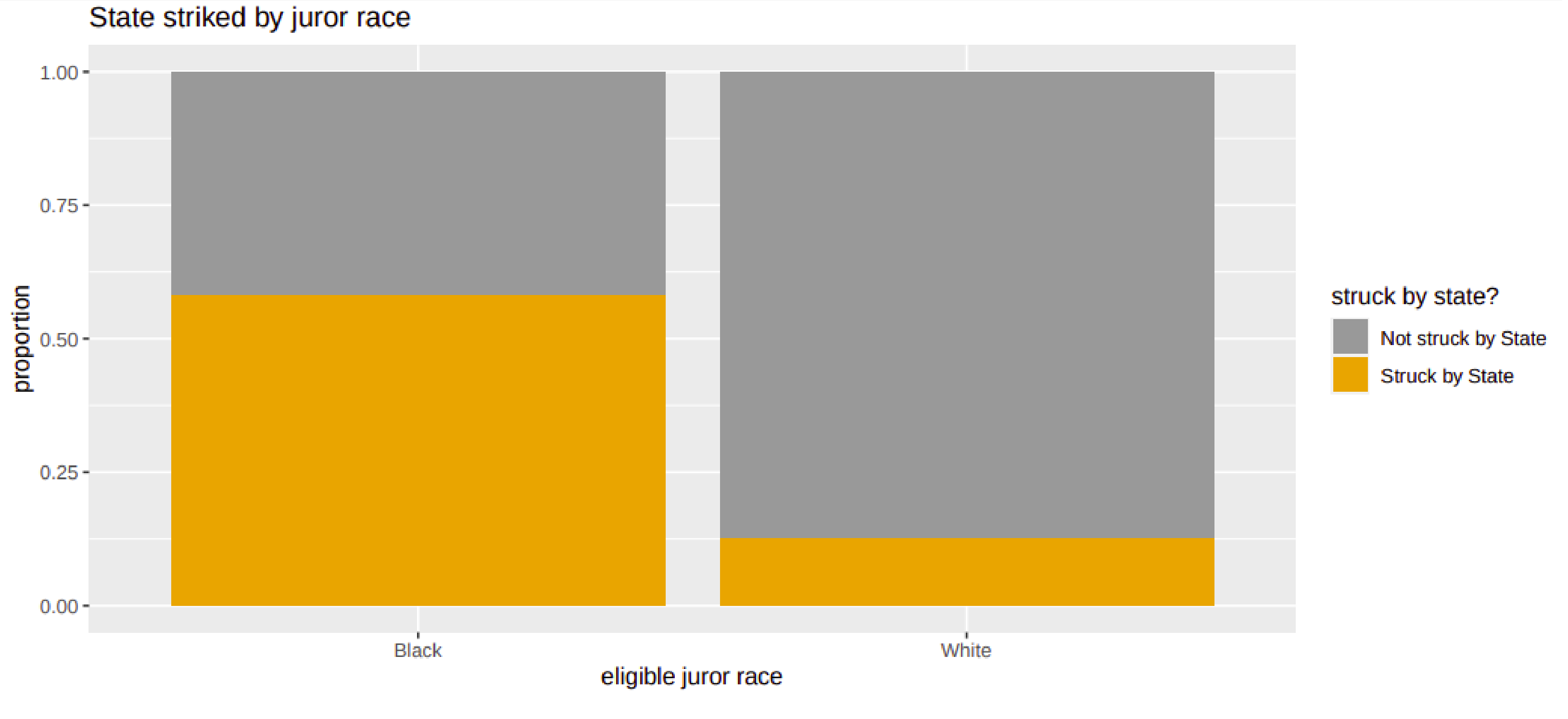
Motivating Example 1



Regression line of Bood alcohol content (BAC) Vs. number of beers

Can you drink 5 beers and stay under the 0.08 limit?

Motivating Example 2



Striking rates by race

Do the observed differences in strike rates between black and white eligible jurors indicate a potential bias, or are the differences just due to chance?

Statistic and Parameter

- A **parameter** is a number that describes some aspect of a population.
- A **statistic** is a number that is computed from data in a sample.

	Parameter	Statistic
Mean	μ	$\bar{x}$
Proportion	p	$\hat{p}$
Std. Dev.	σ	s
Correlation	ρ	r
Slope	β	b

Parameter Vs. Statistic

State whether the quantity described is a **parameter** or a **statistic**, and give the correct notation.

a. Average household income for all houses in the US, using data from the US census

b. The proportion of all residents in a county who voted in the last presidential election.

c. The difference in proportion who have ever smoked cigarettes, between a sample of 500 people who are 60 years old and a sample of 200 people who are 25 years old.

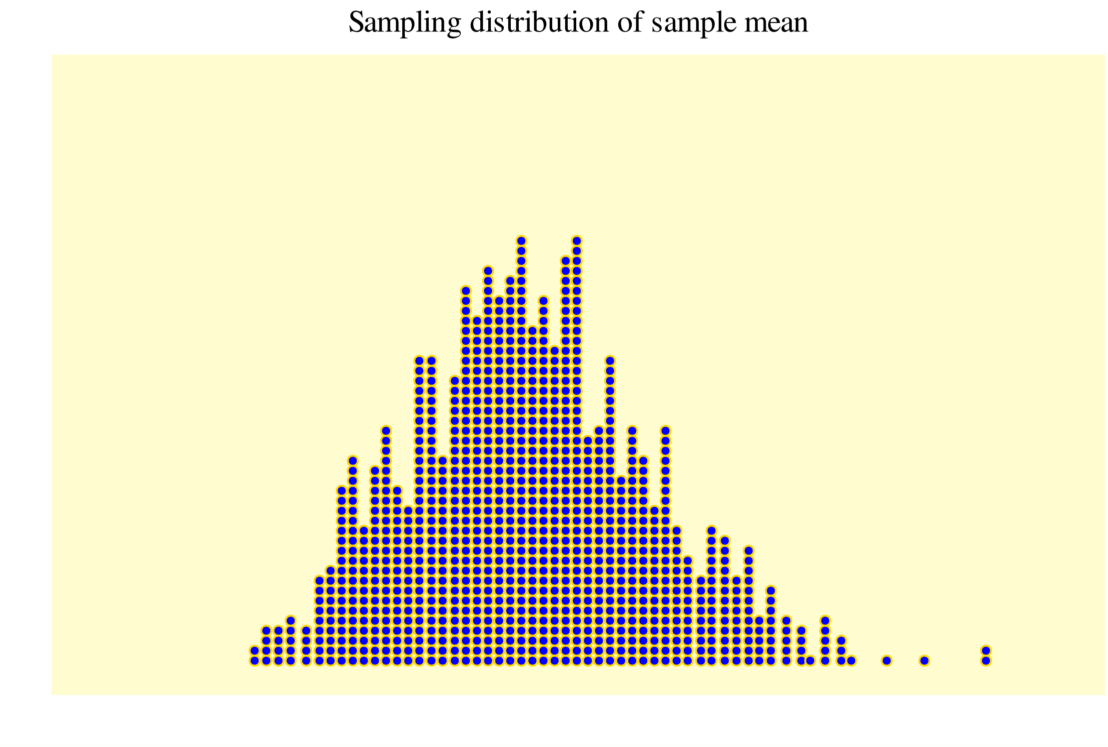
Point Estimate (PE)

- ***Point estimate*** *is a single value constructed from the sample data*
- ***Sample statistic*** *can serve as a point estimate for an unknown parameter*

Sampling Distribution

*A **sampling distribution** is the distribution of sample statistics computed for different samples of the same size from the same population.*

- *Sample statistics varies from sample to sample*
- *Sampling distribution gives us an idea of the variation*



Center and Shape

Center: *If samples are randomly selected, the sampling distribution will be centered around the population parameter.*

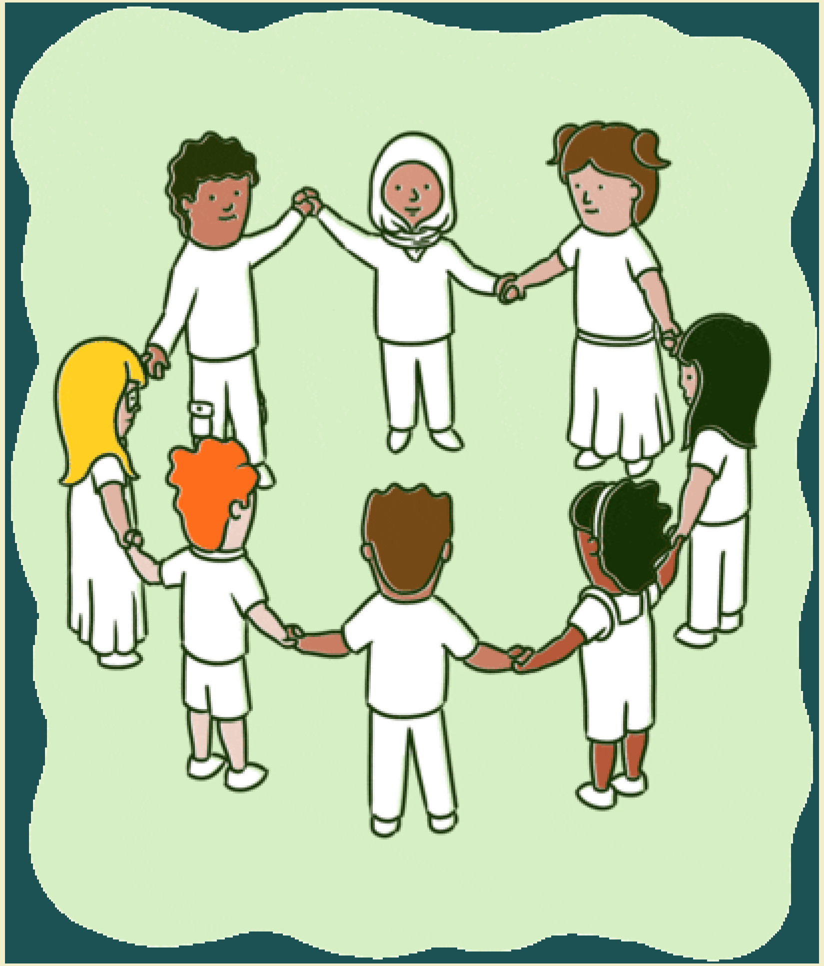
Shape: *For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.*

Standard Error

Uncertainty in point estimates measured by the standard error (SE)

- *The standard error of a statistic is the standard deviation of the sampling distribution*
- *The standard error measures how much the statistic varies from sample to sample*

SHORT DEMO



A Short demo on Sampling distribution

Recall: Gettysburg Address

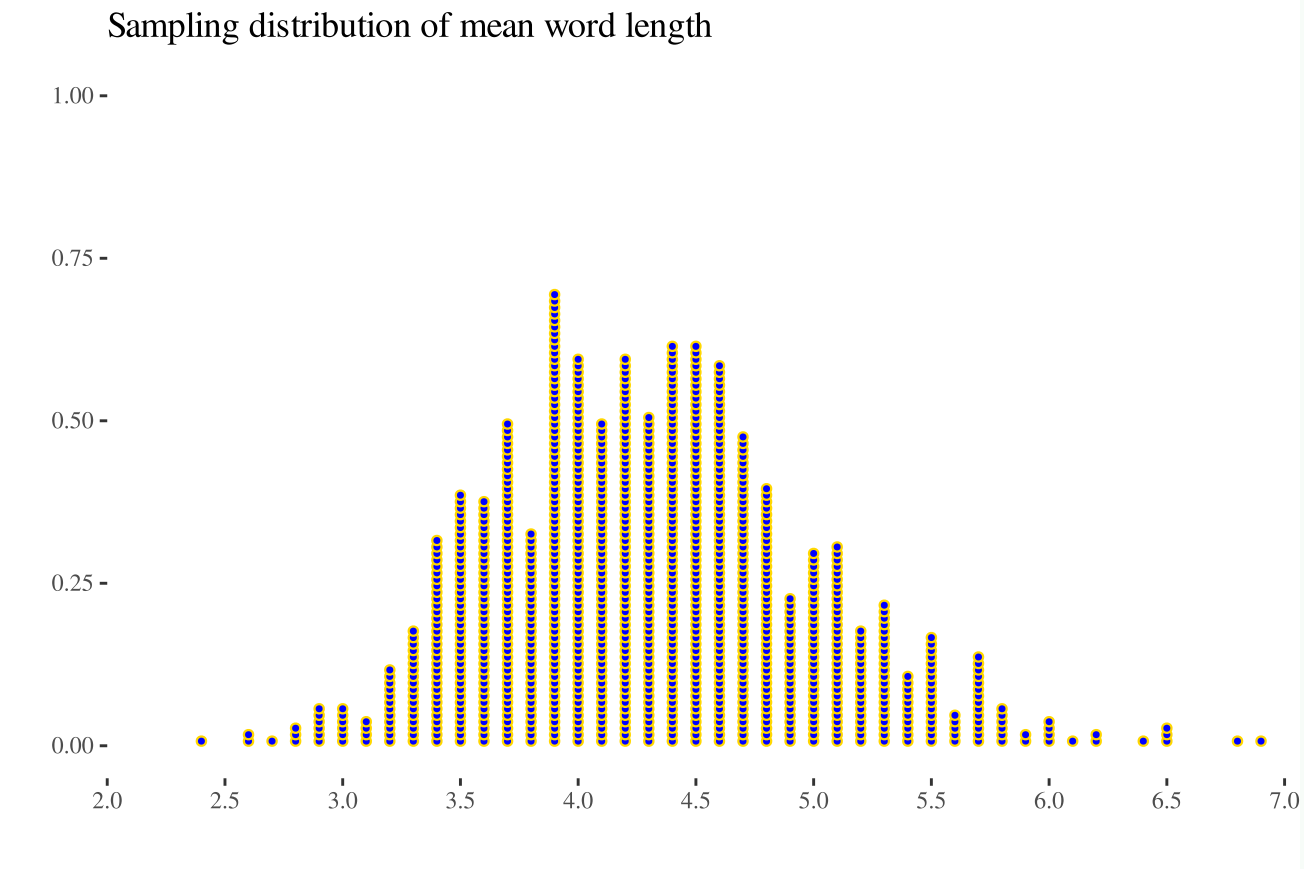
The standard error for the average word size in a random sample of 10 words is closest to

a. 0.5

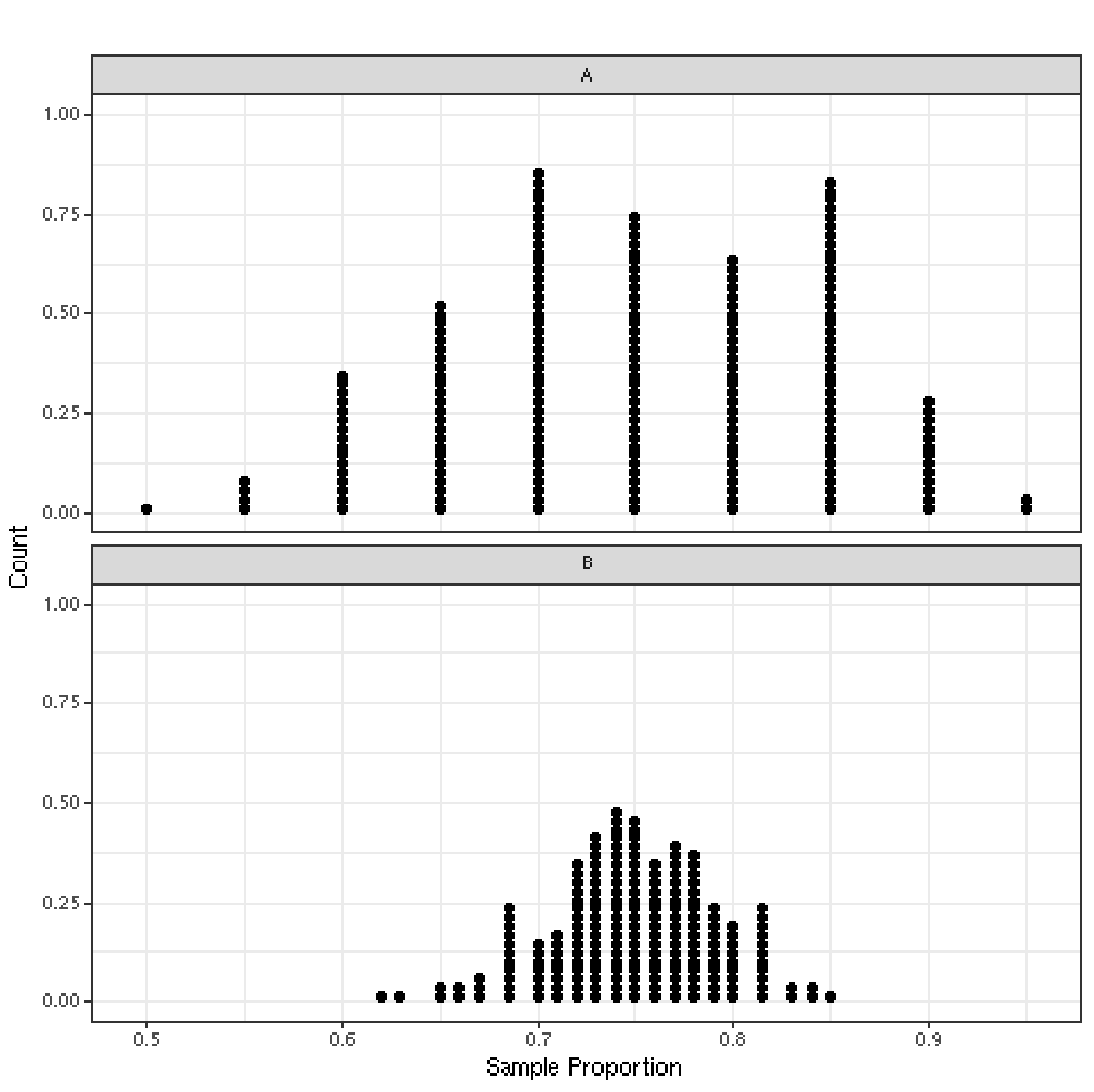
b. 0.7

c. 1.0

d. 1.5



Sample Size Matters!



- As the sample size increases, the variability (SE) of the sample statistics tends to decrease.
- Smaller SE means the sample statistics tend to be closer to the true population parameter value!

Other Factors

Q. What else affects the standard error of a statistic?

A. The variability of the population!

- **Quantitative variable:** the larger the population standard deviation, the larger the standard error of a statistic (like a mean)
- **Categorical variable:** the closer the population proportion is to 0.5, the larger the standard error of the sample proportion

Sample Size vs. Simulation size

Do not confuse sample size and simulation size !!

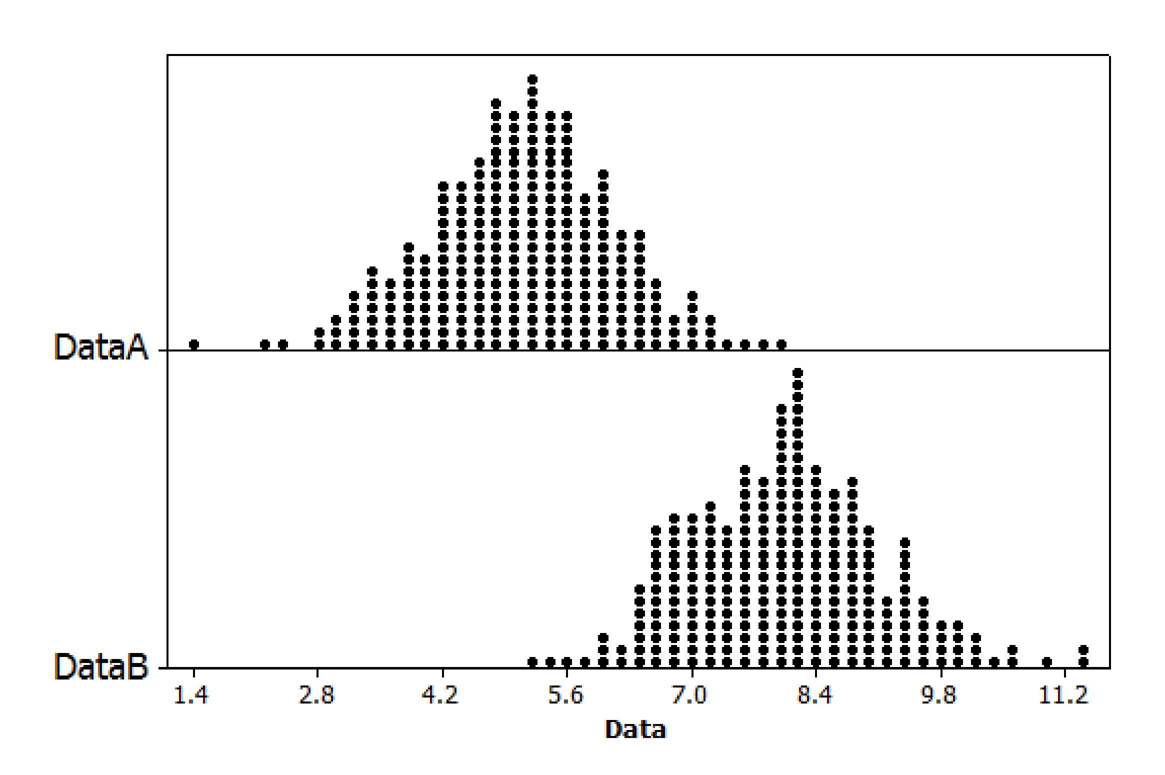
Sample size (n) *= how many individuals are in the sample used to compute our stat?*

Simulation size (N) *= how many random samples did we take from the population to simulate the sampling distribution of our stat?*

- The SE of your stat gets smaller as n get bigger.
- Once you've simulated a couple 100 samples, the shape/center/spread of the sampling distribution should remain about the same as you increase the simulation size.

Random Vs. Non-random

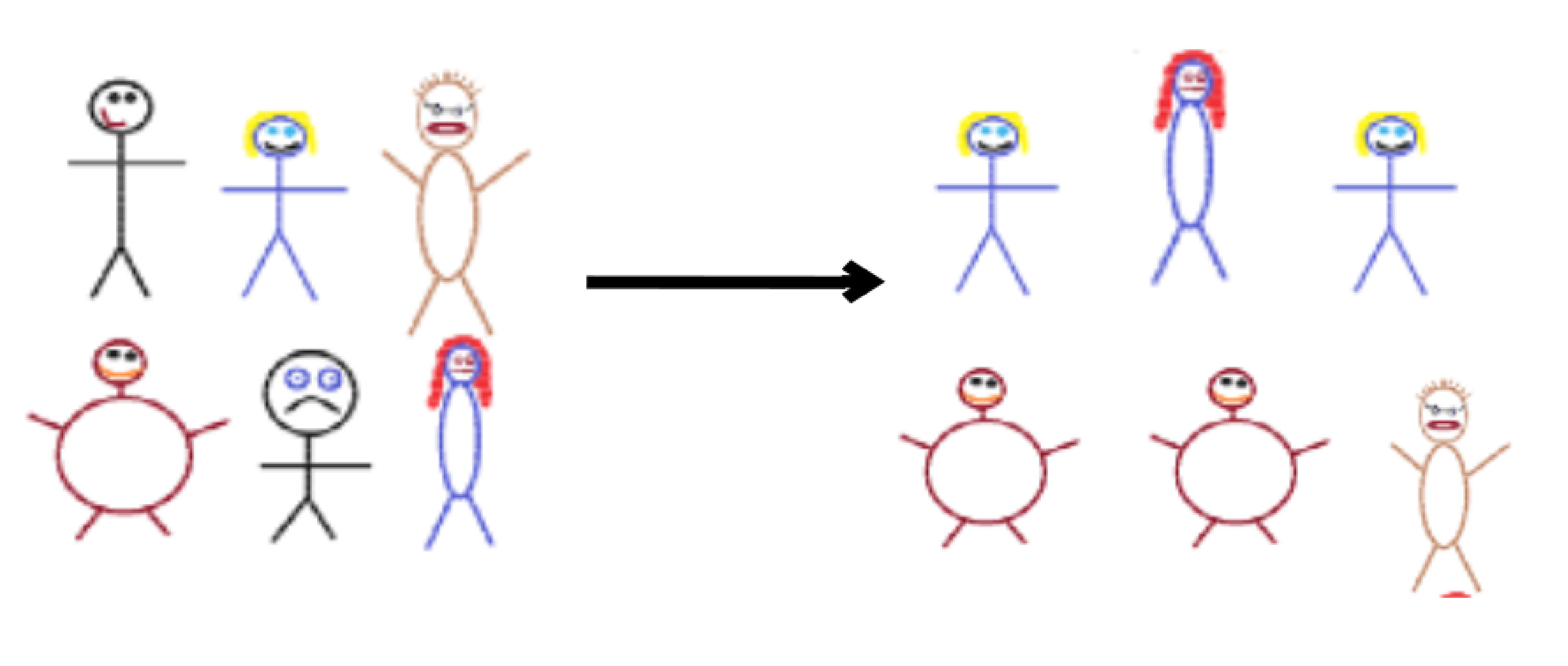
Samples of size 5 are taken from a large population with population mean 8, and the sampling distributions for the sample means are shown. Dataset A (top) and Dataset B (bottom) were collected using different sampling methods. Which dataset (A or B) used random sampling?



Random Vs. non-random data distribution

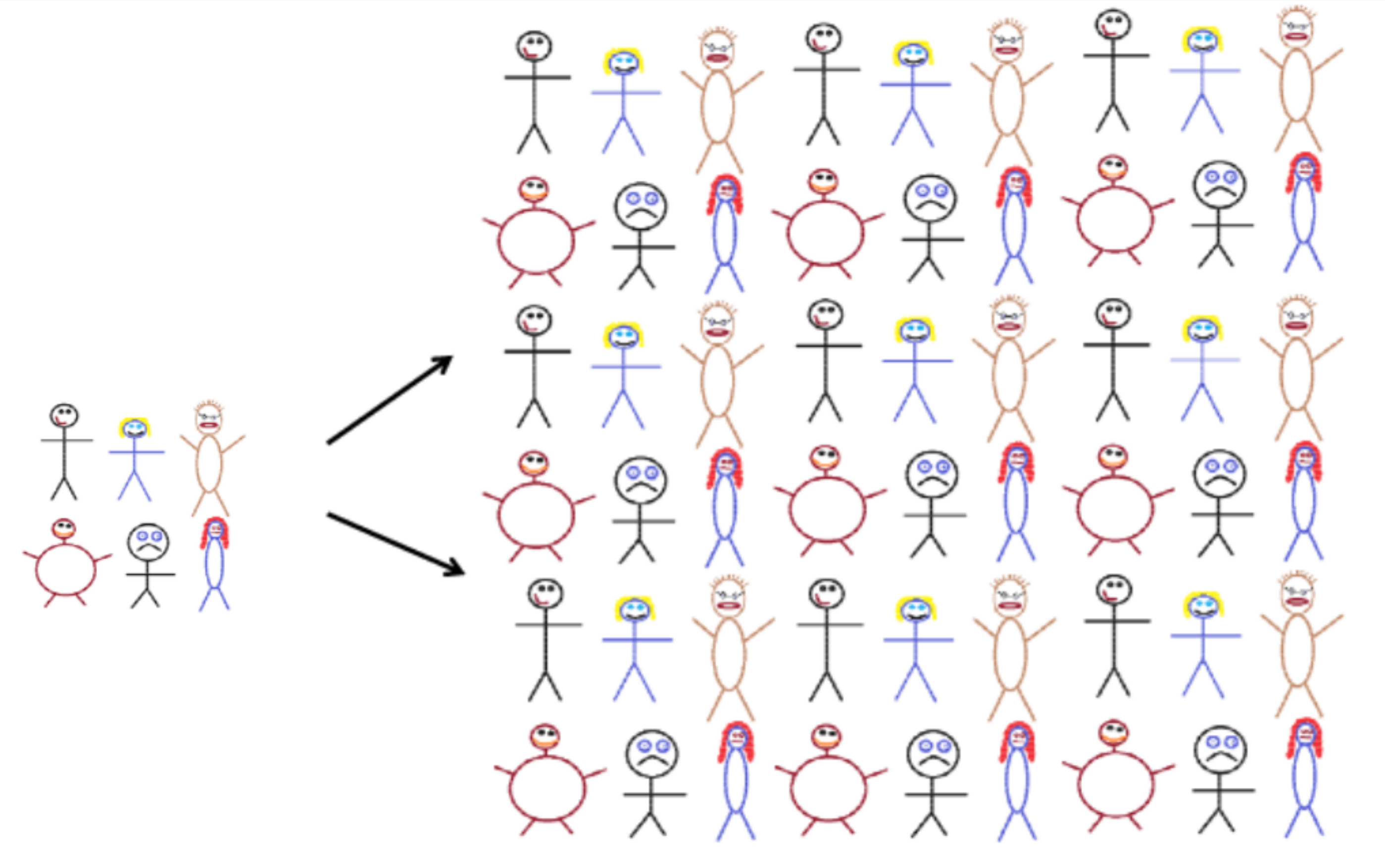
Bootstrap

Bootstrap: Sample with replacement from the original sample, using the same sample size.



Original sample (left) to bootstrap sample (right)

Bootstrap



Original sample (left) to population (right)

Creating a bootstrap sample is the same as using the data simulate a "population" that contains an infinite number of copies of the data.

Bootstrap Sampling in R

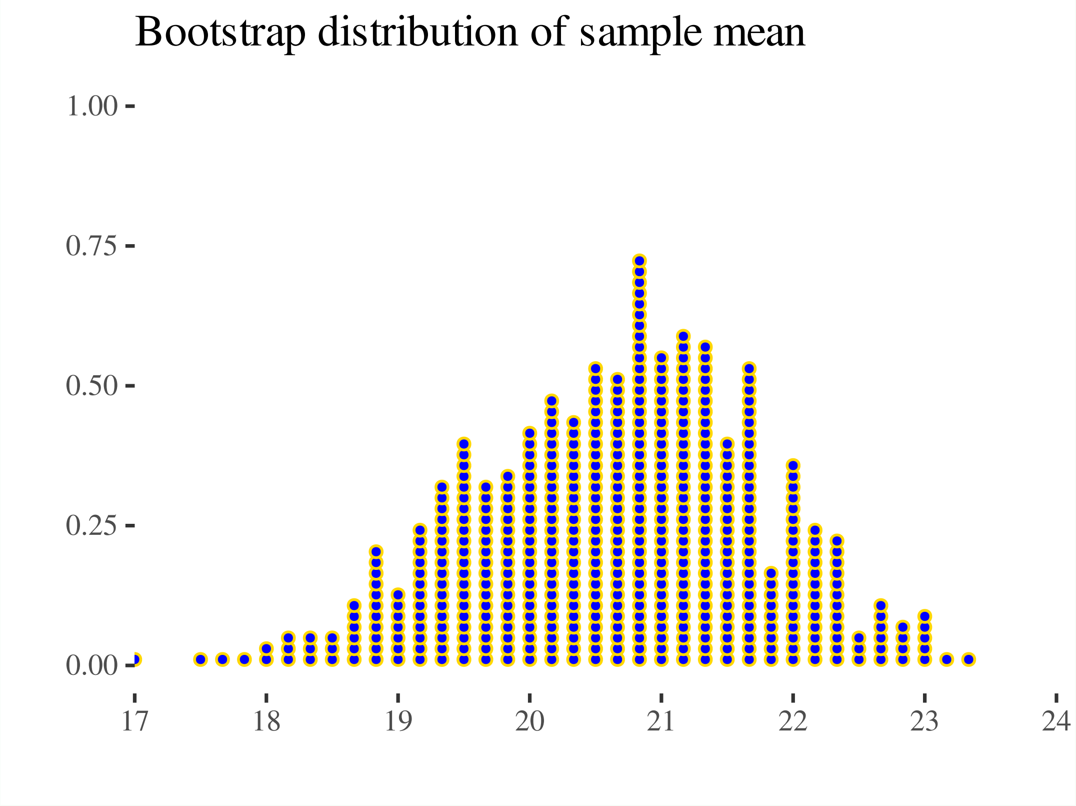
- ***resample a set of observations with replacement***
- ***same data points can appear multiple times***

	Data	Statistic
Original sample	$x_1, x_2, \ldots, x_n$	$\bar{x}_n$
Resample	$x_1^*, x_2^*, \ldots, x_n^*$	$\bar{x}_n^*$

```r
# R-code
boot <- sample(x, size, replace = TRUE)
```

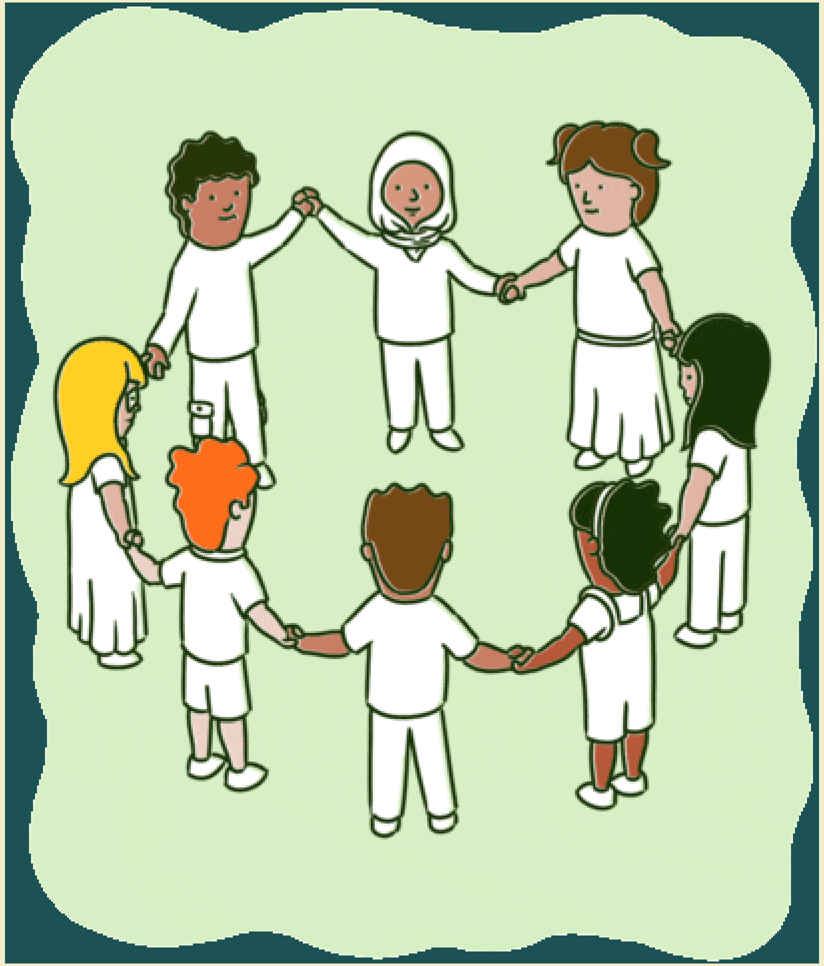
Bootstrap Distrubution

```
library(purrr)
set.seed(143)
X <- c(20,24,19,23,22,16)

# Prototype simulation with N = 500
bootstrapped_means <- tibble(
  iteration = 1:500,
  mean = map_dbl(iteration,
               ~mean(sample(X, replace = TRUE)))
)

ggplot(bootstrapped_means, aes(x = mean)) +
  geom_dotplot(dotsize = 0.7,
             stackratio = 0.9,
             binwidth = .13,
             color = "gold",
             fill = "blue") +
  ggtitle("") + xlab("") + ylab("") +
  scale_x_continuous(limits = c(17, 24),
                   expand = c(0, 0),
                   breaks = seq(17, 24, 1)) +
  labs(title = "Bootstrap distribution of sample mear
```



YOUR TURN 1

10:00



Please go through the remainder of the class activity file and try to answer the questions