# Bootstrap for Statistical Uncertainty

Ian Lundberg
Soc 114

Winter 2025

# Learning goals for today

At the end of class, you will be able to:

1. assess statistical uncertainty (sample-to-sample variability) by a computational procedure

# A motivating problem

- Sample of 10 Dodger players
- Mean salary = \$3.8 million

How much do you trust this as an estimate of the population mean salary?

```
# A tibble: 3 × 2
  `Salary Among Sampled Dodgers`    Value
  <chr>                             <dbl>
1 sample_mean                    3829119.
2 sample_standard_deviation      6357851.
3 sample_size                         10 
```

# Estimator: Sample mean

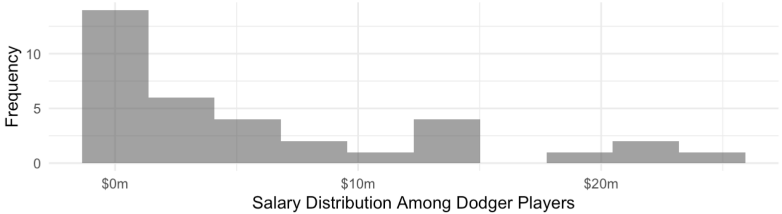
$$\hat{\mu} = \frac{1}{n} \sum_i Y_i$$

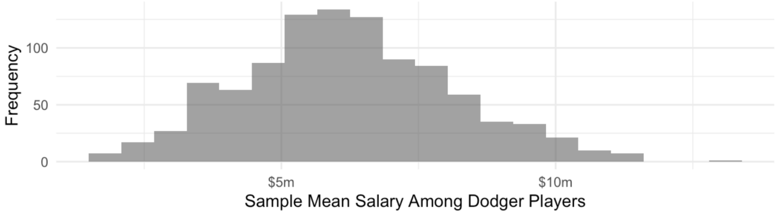
How statistically uncertain is $\hat{\mu}$?

# Standard error of the sample mean

$$\mathsf{SD}(\hat{\mu}) = \sqrt{\mathsf{V}(\hat{\mu})} = \frac{\mathsf{SD}(Y)}{\sqrt{n}}$$

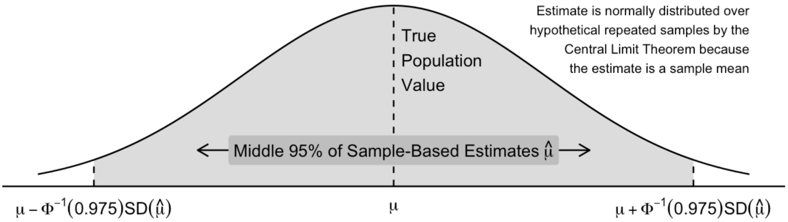
# A standard error captures sample-to-sample variability of the sample mean (second plot)



# Confidence interval

$$\hat{\mu} \to \text{Normal}\left(\text{Mean} = \text{E}(Y), \quad \text{SD} = \frac{\text{SD}(Y)}{\sqrt{n}}\right)$$



# Confidence interval

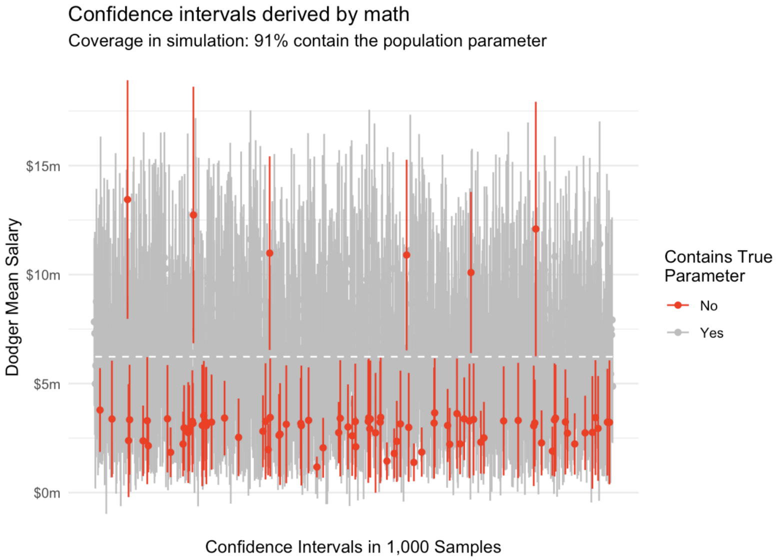
A 95% confidence interval is a range $(\hat{\mu}_{\text{Lower}}, \hat{\mu}_{\text{Upper}})$ such that

$$P(\hat{\mu}_{\text{Lower}} < \mu < \hat{\mu}_{\text{Upper}}) = .95$$

You may know this formula:

$$\hat{\mu} \pm 1.96 \times \widehat{\text{SD}}(\hat{\mu})$$

where 1.96 comes from the properties of the normal distribution.



Confidence intervals derived by math
Coverage in simulation: 91% contain the population parameter
Dodger Mean Salary
$15m
$10m
$5m
$0m
Confidence Intervals in 1,000 Samples
Contains True Parameter
No
Yes

## Replacing math with computation: The bootstrap

# How our estimate comes to be

$$F \to \texttt{data} \to s(\texttt{data})$$

# How our estimate comes to be

$$F \rightarrow \texttt{data} \rightarrow s(\texttt{data})$$

1. The world produces data

# How our estimate comes to be

$$F \rightarrow \texttt{data} \rightarrow s(\texttt{data})$$

1. The world produces data
2. Our estimator function $s()$ converts data to an estimate

```r
estimator <- function(data) {
  data |>
    summarize(estimate = mean(salary)) |>
    pull(estimate)
}
```

# The bootstrap idea

$$F \to \texttt{data} \to s(\texttt{data})$$

# The bootstrap idea

$$F \rightarrow \texttt{data} \rightarrow s(\texttt{data})$$

$$\hat{F} \rightarrow \texttt{data}^* \rightarrow s(\texttt{data}^*)$$

# The bootstrap idea

$$F \to \texttt{data} \to s(\texttt{data})$$

$$\hat{F} \to \texttt{data}^* \to s(\texttt{data}^*)$$

- $F$ is the true distribution of data in the population
- $\hat{F}$ is a plug-in estimator: our empirical data distribution

# The bootstrap idea

1. Generate `data`$^*$ by sampling with replacement from `data`
2. Apply the estimator function
3. Repeat (1–2) many times. Get a distribution.

# Original sample

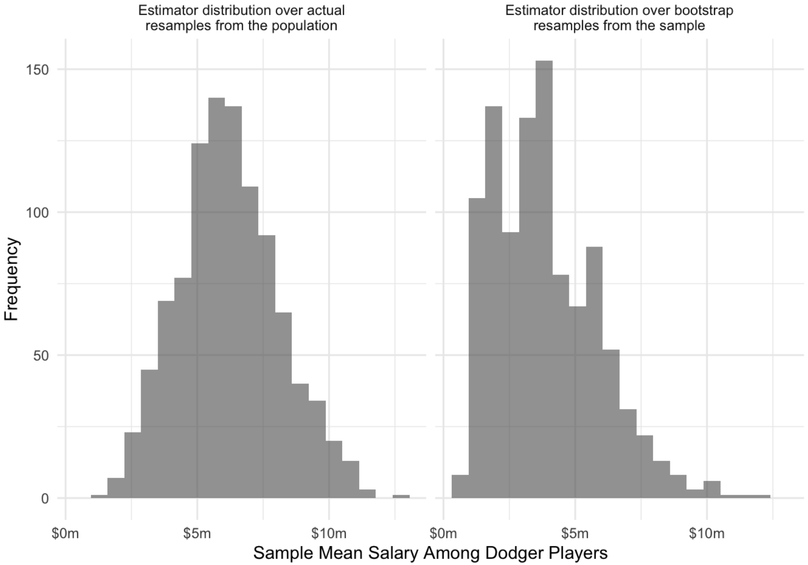
```
# A tibble: 10 × 3
   player           team           salary
   <chr>            <chr>           <dbl>
 1 Barnes, Austin   L.A. Dodgers  3500000
 2 Reyes, Alex*     L.A. Dodgers  1100000
 3 Betts, Mookie    L.A. Dodgers 21158692
 4 Vargas, Miguel   L.A. Dodgers   722500
 5 May, Dustin      L.A. Dodgers  1675000
 6 Bickford, Phil   L.A. Dodgers   740000
 7 Jackson, Andre   L.A. Dodgers   722500
 8 Thompson, Trayce L.A. Dodgers  1450000
 9 Pepiot, Ryan*    L.A. Dodgers   722500
10 Peralta, David   L.A. Dodgers  6500000
```

# Bootstrap sample

```
sample |>
  slice_sample(prop = 1, replace = TRUE)
```

```
# A tibble: 10 × 3
   player         team           salary
   <chr>          <chr>           <dbl>
 1 Betts, Mookie  L.A. Dodgers 21158692
 2 Peralta, David L.A. Dodgers  6500000
 3 Barnes, Austin L.A. Dodgers  3500000
 4 Pepiot, Ryan*  L.A. Dodgers   722500
 5 Jackson, Andre L.A. Dodgers   722500
 6 May, Dustin    L.A. Dodgers  1675000
 7 Reyes, Alex*   L.A. Dodgers  1100000
 8 May, Dustin    L.A. Dodgers  1675000
 9 Vargas, Miguel L.A. Dodgers   722500
10 Peralta, David L.A. Dodgers  6500000
```

# Bootstrap: Many sample estimates



# Bootstrap standard errors

# Bootstrap standard errors

**Goal:** Standard deviation across hypothetical sample estimates

# Bootstrap standard errors

**Goal:** Standard deviation across hypothetical sample estimates
**Estimator:** Standard deviation across bootstrap estimates

$$\widehat{\mathrm{SD}}(s) = \frac{1}{B-1} \sum_{r=1}^{B} \left( s(\mathtt{data}_r^*) - s(\mathtt{data}_\bullet^*) \right)^2$$

# Bootstrap confidence intervals

Two (of many) approaches

- normal approximation
- percentile method

# Bootstrap confidence intervals

## Normal approximation

Point estimate + Bootstrap Standard Error + Normal Approximation

# Bootstrap confidence intervals

## Normal approximation

Point estimate + Bootstrap Standard Error + Normal Approximation

$$s(\mathbf{data}) \pm \Phi^{-1}(.975)\mathrm{SD}\big(s(\mathrm{data}^*)\big)$$

```
estimator(sample) + c(-1,1) * qnorm(.975) * sd(bootstrap_estimates)
```

```
[1]  -22353.11 7680591.51
```

# Bootstrap confidence intervals

## Percentile method

Point estimate + Bootstrap Distribution + Percentiles

# Bootstrap confidence intervals

Percentile method

Point estimate + Bootstrap Distribution + Percentiles

```
quantile(bootstrap_estimates, probs = c(.025, .975))
```

```
   2.5%   97.5%
1103406 8216408
```

(requires a larger number of bootstrap samples)

# Bootstrap discussion: Causal outcome model

Suppose a researcher carries out the following procedure.

1. Sample $n$ units from the population
2. Learn an algorithm $\hat{f} : \{A, \vec{X}\} \rightarrow Y$ to minimize squared error
3. Predict the average causal effect

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}(A = 1, \vec{X} = \vec{x}_i) - \hat{f}(A = 0, \vec{X} = \vec{x}_i) \right)$$

How would you make a bootstrap confidence interval for $\hat{\tau}$?

# Bootstrap discussion: Causal outcome model

Suppose a researcher carries out the following procedure.

1. Sample $n$ units from the population
2. Learn an algorithm $\hat{f} : \{A, \vec{X}\} \rightarrow Y$ to minimize squared error
3. Predict the average causal effect

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(A = 1, \vec{X} = \vec{x}_i) - \hat{f}(A = 0, \vec{X} = \vec{x}_i)\right)$$

How would you make a bootstrap confidence interval for $\hat{\tau}$?

# Bootstrap discussion: Causal outcome model

For each replicate $r = 1, \ldots,10{,}000$,

1. Draw bootstrap sample $\texttt{data}_r^*$
2. Estimate $\hat{\tau}_r^*$

Produces many estimates $\hat{\tau}_1^*, \ldots, \hat{\tau}_{10{,}000}^*$
Report the 2.5 and 97.5 percentiles of those

# Complex samples

- stratified
- clustered
- beyond

# Simple random sample

Sample 150 players at random.
(standard bootstrap applies)

# Stratified sample

Sample 10 players on each of 30 teams

- Why doesn't the simple bootstrap mimic this sampling variability well?

# Stratified sample

Sample 10 players on each of 30 teams

- Why doesn't the simple bootstrap mimic this sampling variability well?

Solution: Stratified bootstrap

- Take resamples within groups
- Preserve distribution across groups

# Clustered sample

Sample 10 teams. Record data on all players in sampled teams.

- Why doesn't the simple bootstrap mimic this sampling variability well?

# Clustered sample

Sample 10 teams. Record data on all players in sampled teams.

- Why doesn't the simple bootstrap mimic this sampling variability well?

Solution: Cluster bootstrap

- Bootstrap the groups

# Complex survey sample

- Often stratified and clustered, in multiple stages
- Strata and clusters are often restricted geographic identifiers

# Complex survey sample: Replicate weights

| | name | weight | employed | repwt1 | repwt2 | repwt3 |
|---|---|---|---|---|---|---|
| 1 | Luis | 4 | 1 | 3 | 5 | 3 |
| 2 | William | 1 | 0 | 1 | 2 | 2 |
| 3 | Susan | 1 | 0 | 3 | 1 | 1 |
| 4 | Ayesha | 4 | 1 | 5 | 3 | 4 |

- Point estimate $\hat{\tau}$
- Replicate estimates $\hat{\tau}^1, \hat{\tau}^2, \ldots$

# Complex survey sample: Replicate weights

Re-aggregate as directed by survey documentation.
Current Population Survey (example with documentation)

# Complex survey sample: Replicate weights

Re-aggregate as directed by survey documentation.
Current Population Survey (example with documentation)

$$\text{StandardError}(\hat{\tau}) = \sqrt{\frac{4}{160}\sum_{r=1}^{160}(\hat{\tau}_r^* - \hat{\tau})^2}$$

# Words of Warning

The bootstrap makes inference easy, but there are catches.

- biased estimator
- estimator is something like $\max(\vec{y})$

# Words of Warning

The bootstrap makes inference easy, but there are catches.

- biased estimator
  - not centered correctly $\rightarrow$ undercoverage
- estimator is something like $\max(\vec{y})$

# Words of Warning

The bootstrap makes inference easy, but there are catches.

- biased estimator
  - not centered correctly $\rightarrow$ undercoverage
- estimator is something like $\max(\vec{y})$
  - $\max(\vec{y}^*)$ never above $\max(\vec{y})$

# Words of Warning

The bootstrap makes inference easy, but there are catches.

- biased estimator
  - not centered correctly $\rightarrow$ undercoverage
- estimator is something like $\max(\vec{y})$
  - $\max(\vec{y}^*)$ never above $\max(\vec{y})$
  - depends heavily on a particular point

# Learning goals for today

At the end of class, you will be able to:

1. assess statistical uncertainty (sample-to-sample variability) by a computational procedure