Inference for Modeling ([Reading: 18])

Learning goals:

- Review the procedure for constructing bootstrap confidence intervals.
- Learn new applications of the bootstrap for inference on models.

UC Berkeley Data 100 Summer 2019
Sam Lau

(Slides adapted from John DeNero)

Announcements

- Project 2 out
  - Due next Tuesday, Aug 5.
- Small group tutoring: tinyurl.com/d100-tutor-week6

Statistical Inference

Remember this slide?

Statistical inference estimates attributes of the **population given a sample**:

Population

$$p^* \approx 0.51$$

We know that $\hat{p}$ is an unbiased estimator of the parameter: p^*

Sample

$$X_1, X_2, \ldots, X_{100}$$

$$\hat{p} = \frac{1}{n}(X_1 + \cdots + X_{100})$$

$$= 0.51$$

Problem: Estimate of p* will almost always be wrong! Why?

Confidence Intervals

- Although our estimates won't be exactly right, they will (hopefully) get close.
- **Confidence intervals** (CIs) quantify how close we think population parameter is to an estimate.
- Intuition: If our estimator has high variance, we are less certain that our parameter will be close to estimate.
  - IOW: High estimator variance => bigger CI

Estimator Variance

- The variance of a random variable X is:

$$Var(X) = E\left((X - E(X))^2\right)$$

IOW: How much does X vary around its long-run average?

- Estimators are RVs (constructed using other RVs).
- E.g. sample mean μ̂ estimator for population mean μ*.

$$Var(\hat{\mu}) = E\left((\hat{\mu} - E(\hat{\mu}))^2\right)$$

But average sample mean doesn't change.

If sample were different

Sample mean would be different

What does it mean?

$$Var(\hat{\mu}) = E\left((\hat{\mu} - E(\hat{\mu}))^2\right)$$

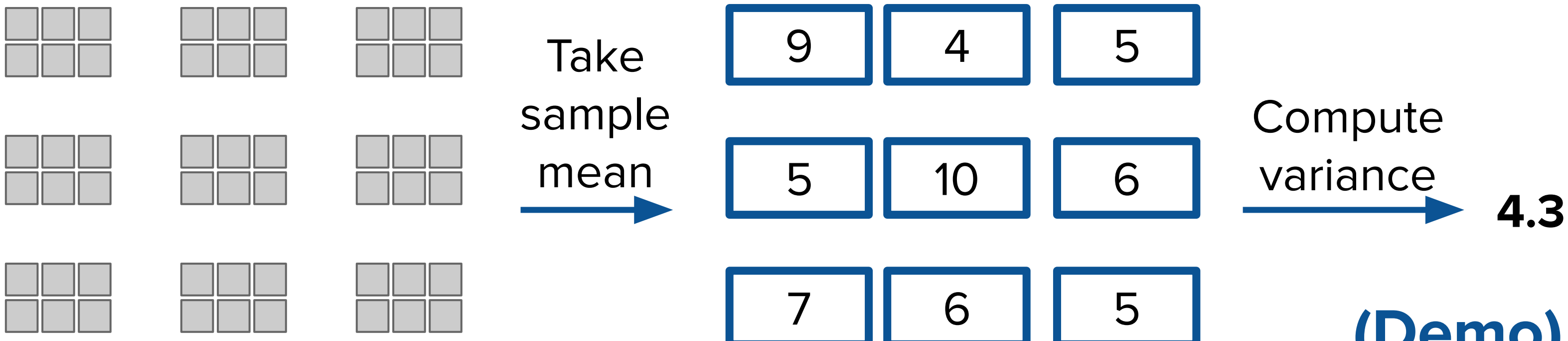
- Suppose μ̂ is sample mean for sample of size n.
- Imagine repeated taking samples from population.
  - For each sample, compute μ̂, the sample mean.
- E(μ̂) is long-run average of sample means.
  - μ̂ is unbiased, so expect to get close to μ*
- Var(μ̂) is how much the sample means vary.
- Can't find Var(μ̂)! So we have to estimate it from sample.

Same reasoning applies for any estimator θ, not just the sample mean. I use the sample mean here because it's familiar.

Estimating the Estimator Variance

$$Var(\hat{\mu}) = E\left((\hat{\mu} - E(\hat{\mu}))^2\right)$$

- Impractical approach that would work:
  - Draw m samples of size n.
  - Compute m sample means.
  - $Var(\hat{\mu}) \approx$ Variance of these sample means.



(Demo)

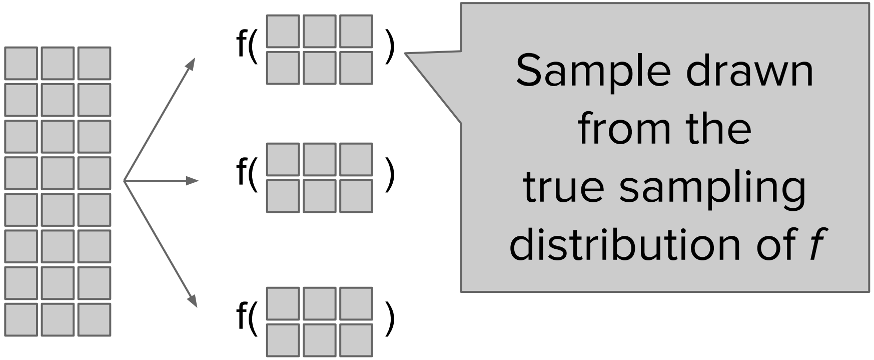
Bootstrap Resampling

Bootstrap Resampling

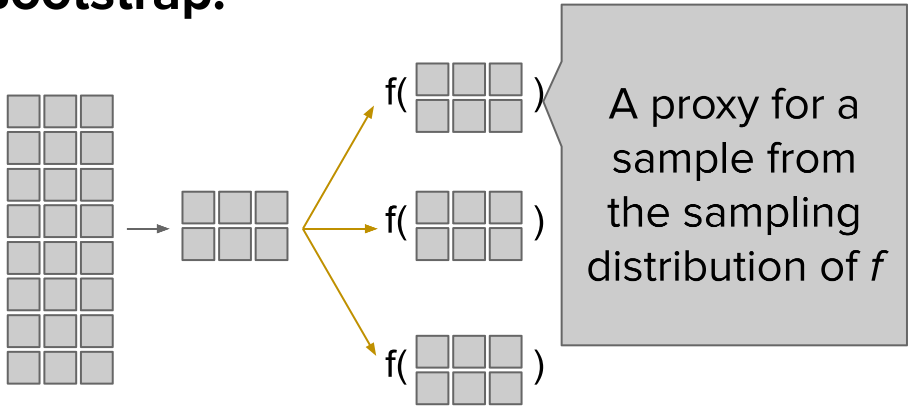
- **Bootstrap resampling** is a technique for estimating parameters of the sampling distribution of an estimator.
- Intuition: If sample looks like population, we can pretend that the sample is the population.
    - We resample **n** points from our sample **with replacement** to simulate a sample from population.
    - Usually only works if sample drawn randomly!

Bootstrap Resampling

Impractical:



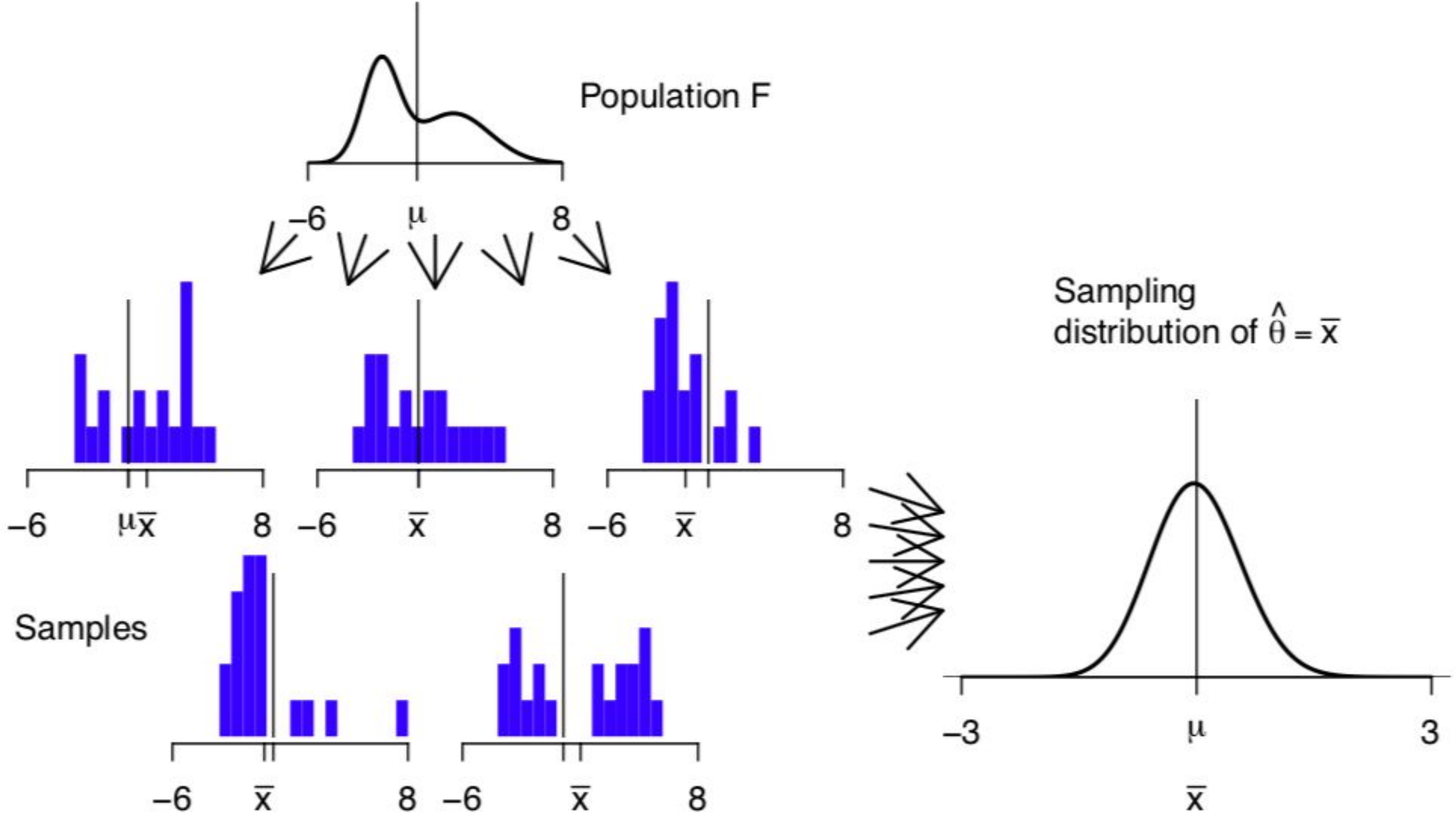
Bootstrap:



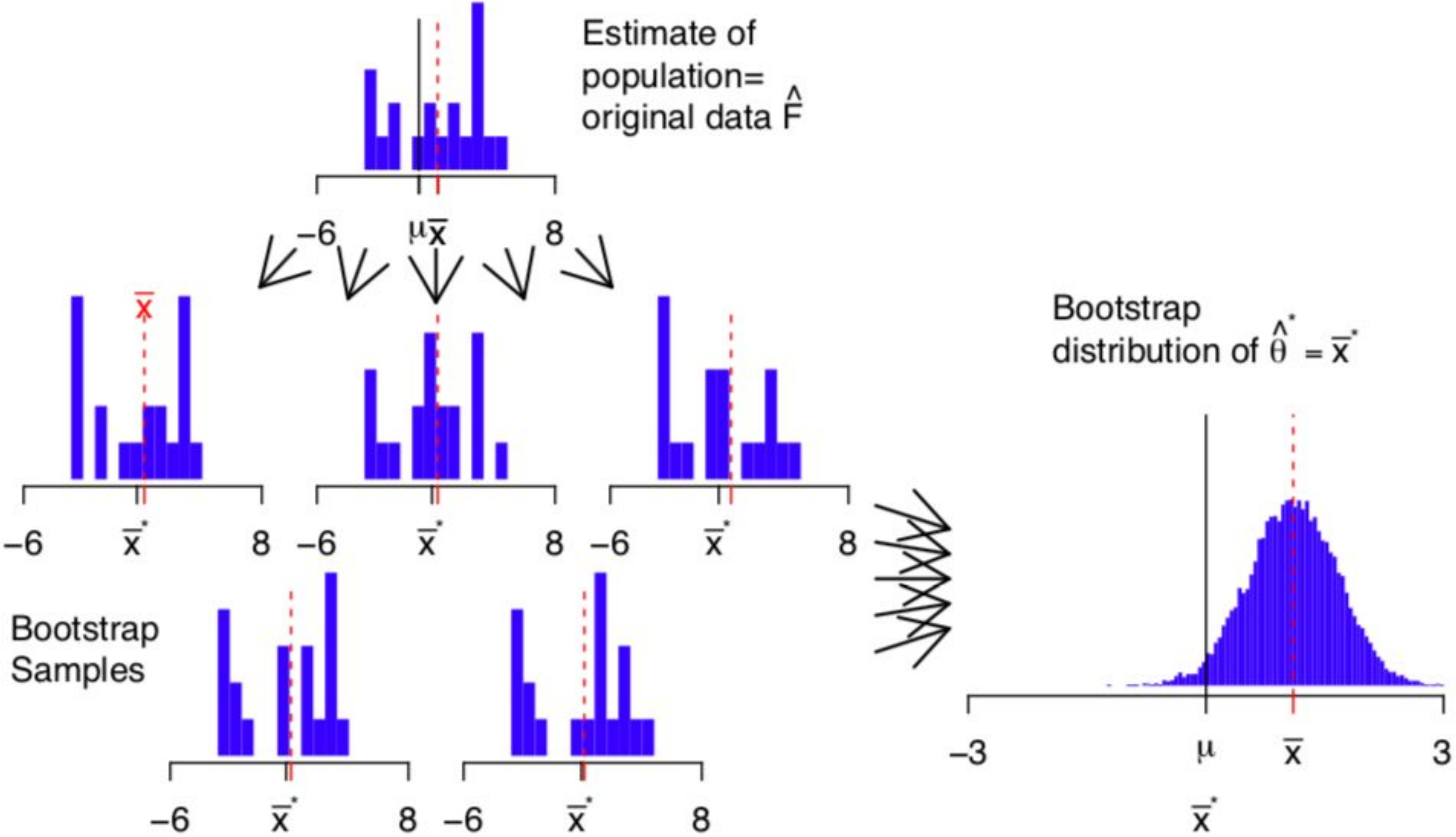
(Demo)

Bootstrap Discussion

- All RVs have a distribution. If the RV is an estimator computed using a sample, its distribution is called the **sampling distribution**.
- The **bootstrap sampling distribution** is the distribution estimated by computing estimate across resamples.
  - The center and spread are both wrong (but usu close).
  - Bootstrap distribution centered at sample's center (not population's center)
  - More usefully, the variance of bootstrap distribution often close to estimator variance.



Population F
−6
μ
8
Samples
−6
μ
$\bar{x}$
8
−6
$\bar{x}$
8
−6
$\bar{x}$
8
−6
$\bar{x}$
8
−6
$\bar{x}$
8
Sampling
distribution of $\hat{\theta} = \bar{x}$
−3
μ
3
$\bar{x}$


Estimate of
population=
original data $\hat{F}$
−6
μ $\bar{x}$
8
Bootstrap
distribution of $\hat{\theta}^* = \bar{x}^*$
$\bar{x}$
−6
$\bar{x}^*$
8
Bootstrap
Samples
−3
μ
$\bar{x}$
3
$\bar{x}^*$

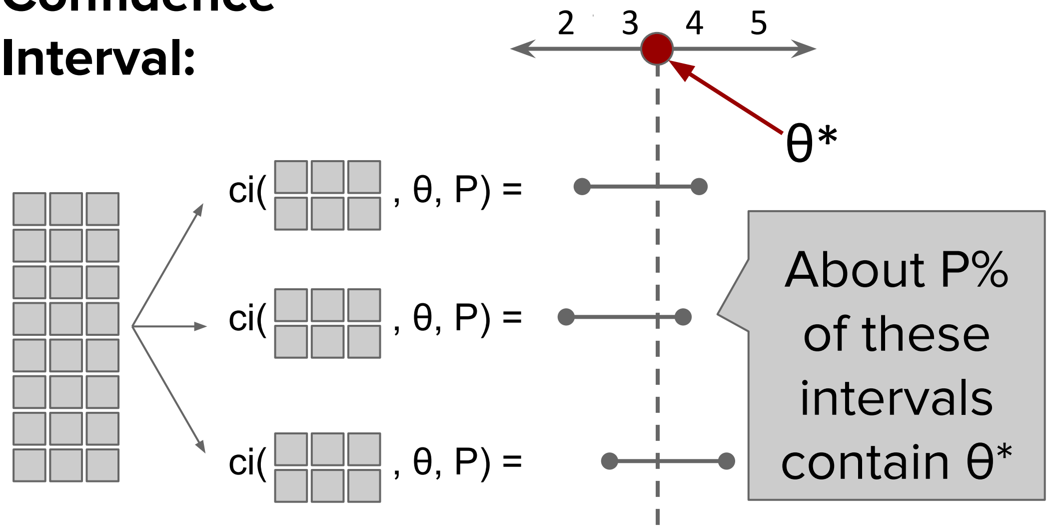
Bootstrap Confidence Intervals

Confidence Intervals (Without Bootstrap)

- Intuition: Estimate an interval where we think population parameter is based on center and variance of estimator.
- What does a P% confidence interval mean?
- Imagine the following procedure:
  - Take a sample from population.
  - Find sampling distribution of estimator.
  - Compute confidence interval with distribution.
- If we repeat this procedure, population parameter will be in our interval P% of the time in the long run.

Confidence Intervals (Without Bootstrap)

Confidence Interval:

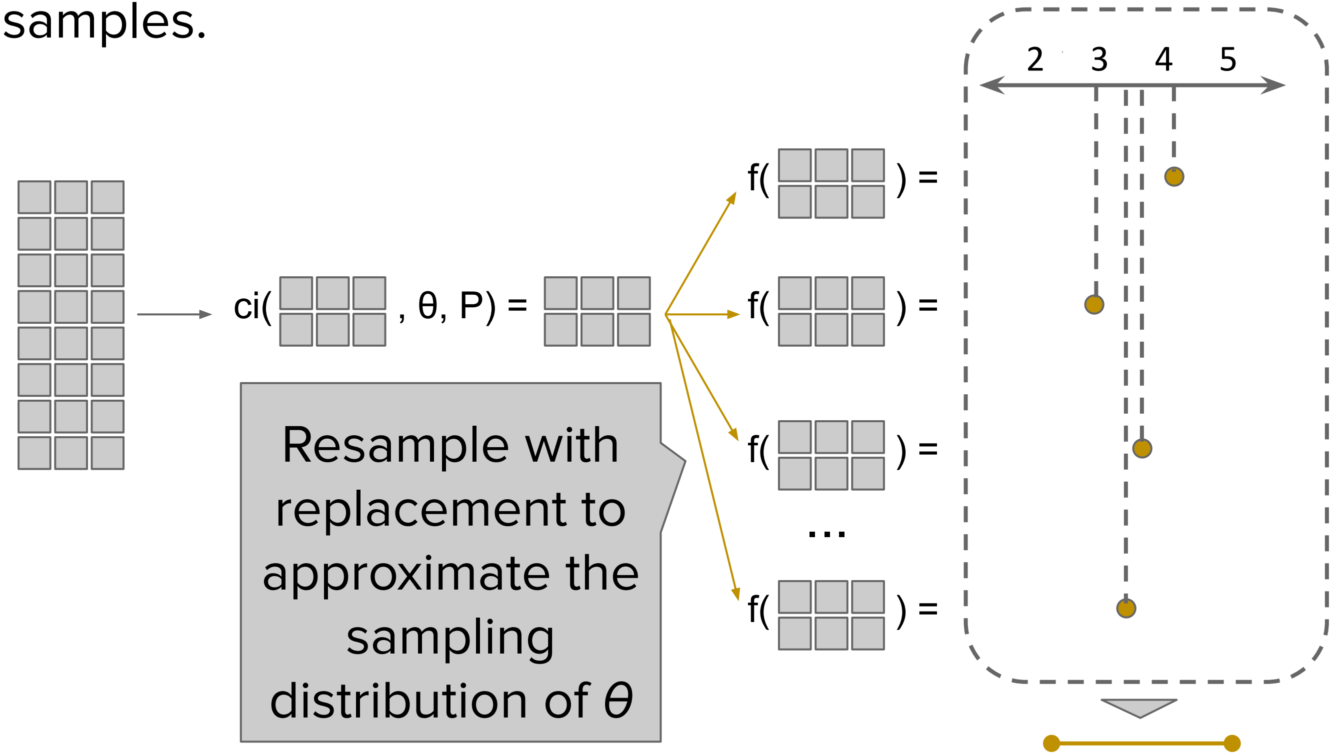


How to compute `ci(`s`,`θ`,`P`)`?

- Approximate the sampling distribution of θ using the sample s.
- Choose the middle P% of samples from this approximate distribution.

Bootstrap Confidence Intervals

An estimator *ci* for a P% confidence interval for θ is a function that takes a sample and returns an interval. This interval will (ideally) contain θ* for P% of samples.



Resample with replacement to approximate the sampling distribution of θ

Choose an interval containing the middle P% of these values.

Aka: the percentile bootstrap CI

(Demo)

Break!

Fill out Attendance:

http://bit.ly/at-d100

Applications For Modeling

Estimating Population Accuracy

- In modeling: Training set = (X_{train}, y_{train}); Test set = (X_{test}, y_{test}).
- Both are random samples from the same population.
- We're interested in accuracy of model on population.
- Can't find this, so we estimate it with test set accuracy

$\texttt{model.fit}(X_{train}, y_{train})$

$y_{guess} = \texttt{model.predict}(X_{test})$

$\texttt{accuracy}(y_{guess}, y_{test})$

Estimator for population accuracy

Same reasoning works for estimating error of model on population

Therefore, resample the test set only.

(Demo)

Comparing Two Models

- Training set = (X_{train}, y_{train}); Test set = (X_{test}, y_{test}).
- Fit two models. Which has higher population accuracy?

```
model₁.fit(X_train, y_train)
model₂.fit(X_train, y_train)
y₁ = model₁.predict(X_test)
y₂ = model₂.predict(X_test)
acc(y₁, y_test) - acc(y₂, y_test)
```

Estimator for difference in accuracy

If CI for diff contains 0, models not significantly better for population.

Therefore, resample the test set and compute accuracy differences.

(Demo)

Estimating Linear Regression Parameters

- Training set = (X_{train}, y_{train}); Test set = (X_{test}, y_{test}).
- What associations are useful for linear model?

```
model = LinearRegression().fit(X_train, y_train)
```

$\boxed{\theta_i = \texttt{model.coef[i]}}$

Estimator for slope associated with feature i

If CI for slope contains 0, the feature not significantly important for model.

Therefore, resample the training set.

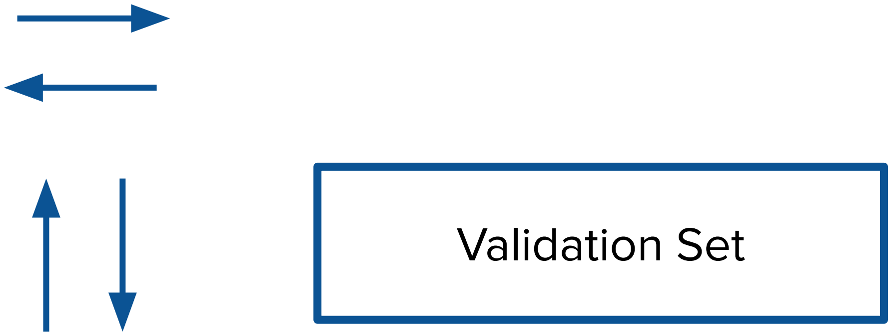
(Demo)

Bootstrap Warnings

- Although bootstrap is very useful, be careful:
- **Small samples** and **skewed sampling distributions** cause inaccurate estimates.
  - Usually means that the procedure to construct a 95% CI will not capture parameter 95% of the time.
- Many extensions to the bootstrap to work around these.
  - E.g. studentized bootstrap for small samples (details in the textbook)
  - Covered in more advanced classes on estimation.

Summary

- We quantify uncertainty about estimation via confidence intervals.
- The bootstrap allows us to estimate the variance of the sampling distribution via resampling.
- Bootstrap confidence intervals can be used to estimate many types of population parameters, including the error of a model on the population.



(Demo)