Bootstrapping and Confidence Intervals

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

Bootstrap

In the lack of active imagination, parallel universes and the likes, we need an alternative way of producing fake data set that resemble the parallel universes.

Bootstrapping is the practice of sampling from the observed data (X,Y) in estimating statistical properties.

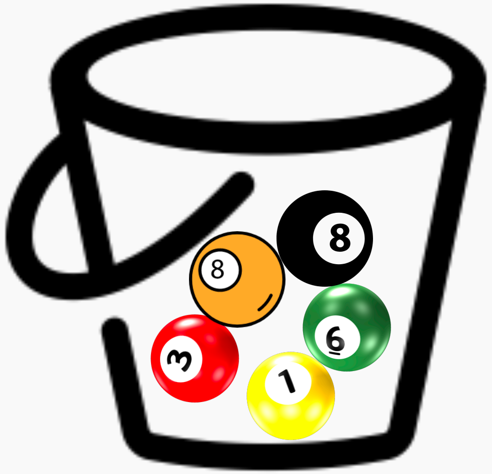
Bootstrap

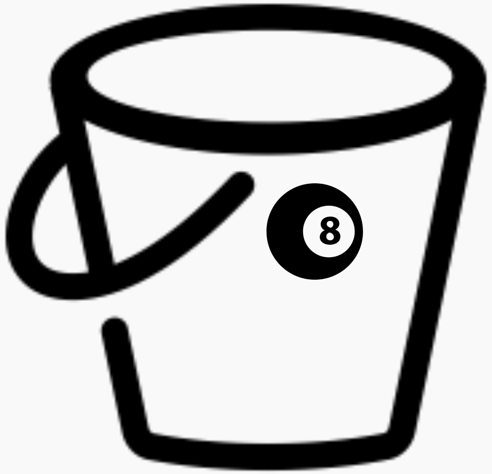
Imagine we have 5 billiard balls in a bucket.

Bootstrap

We first pick randomly a ball and replicate it. This is called **sampling with replacement.** We move the replicated ball to another bucket.

Bootstrap

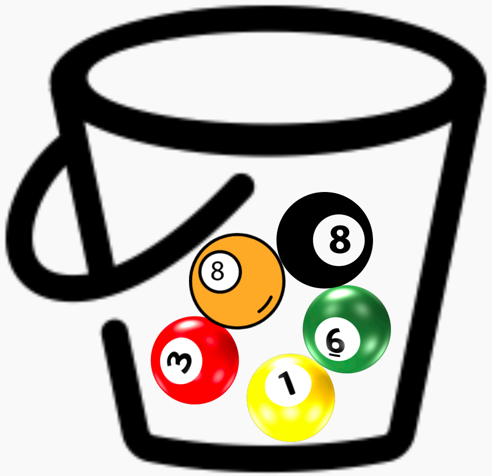


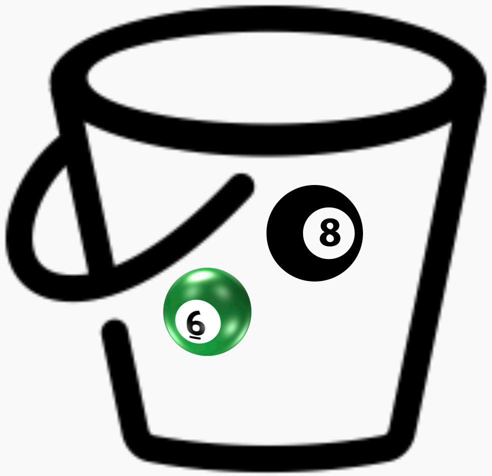


Bootstrap

We then randomly pick another ball and again we replicate it. As before, we move the replicated ball to the other bucket.

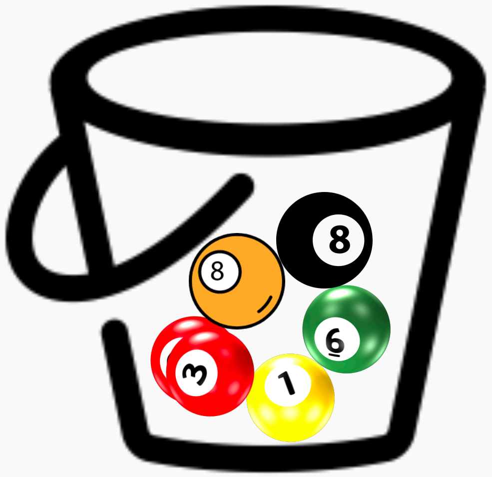
Bootstrap

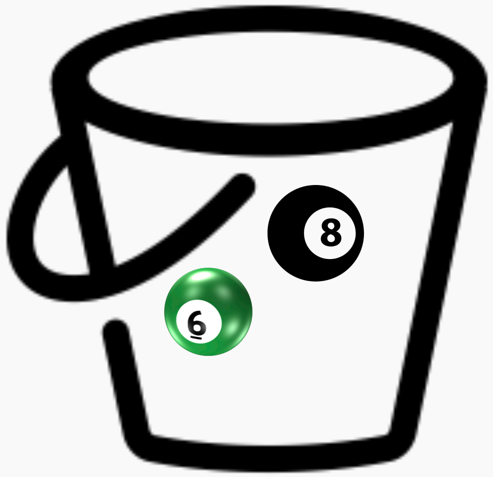




Bootstrap

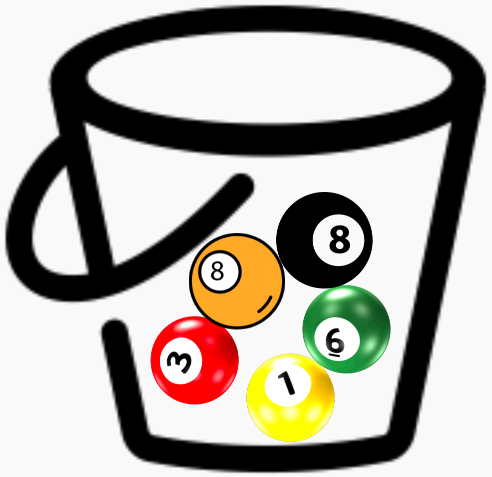
We repeat this process.

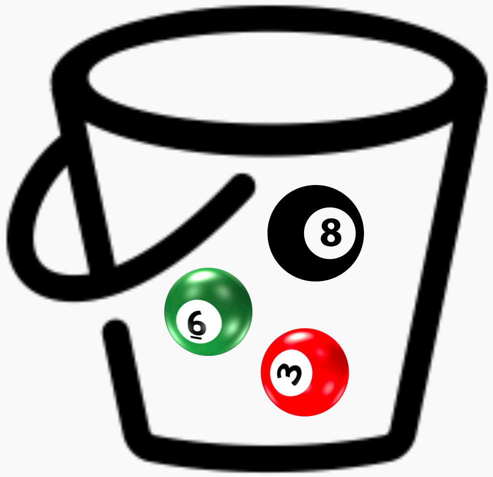




Bootstrap

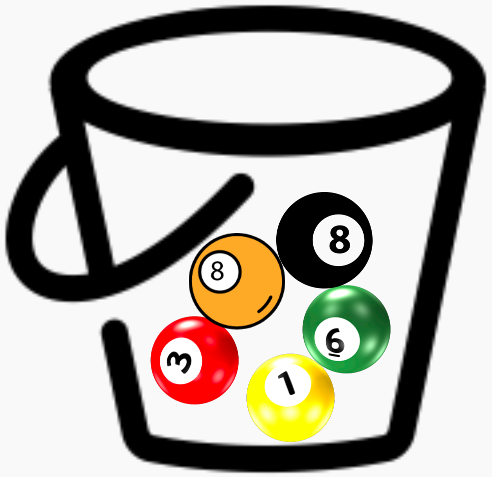
Again

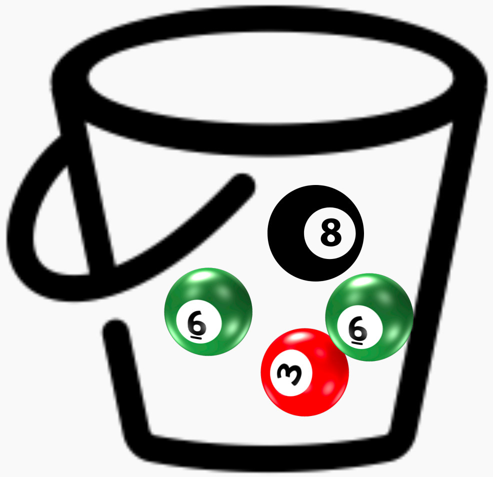




Bootstrap

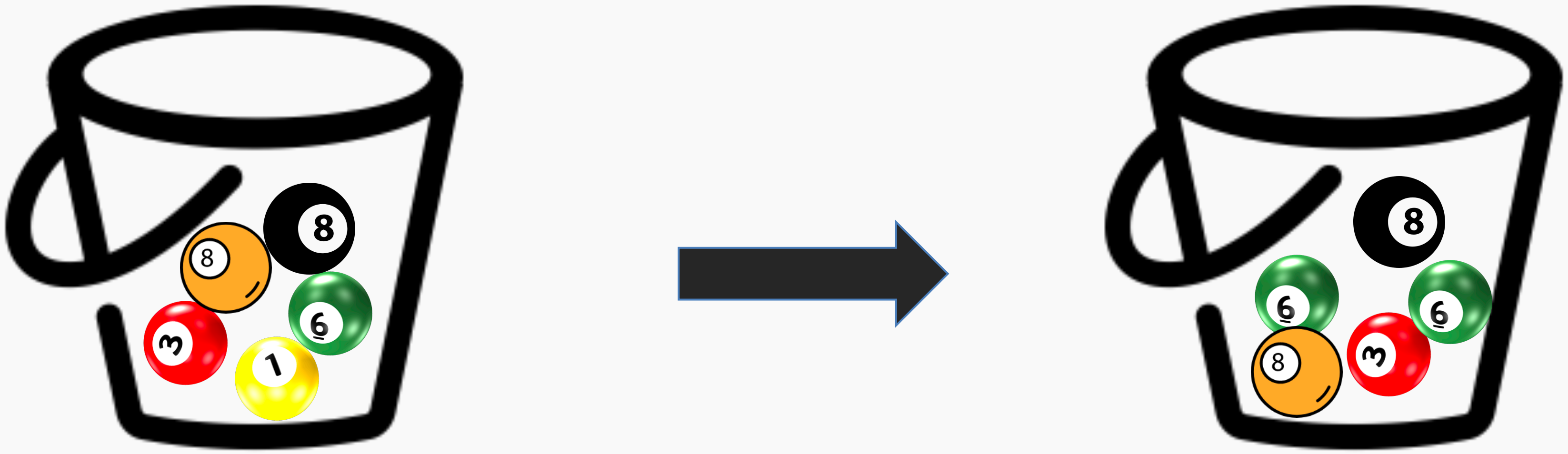
And again





Bootstrap

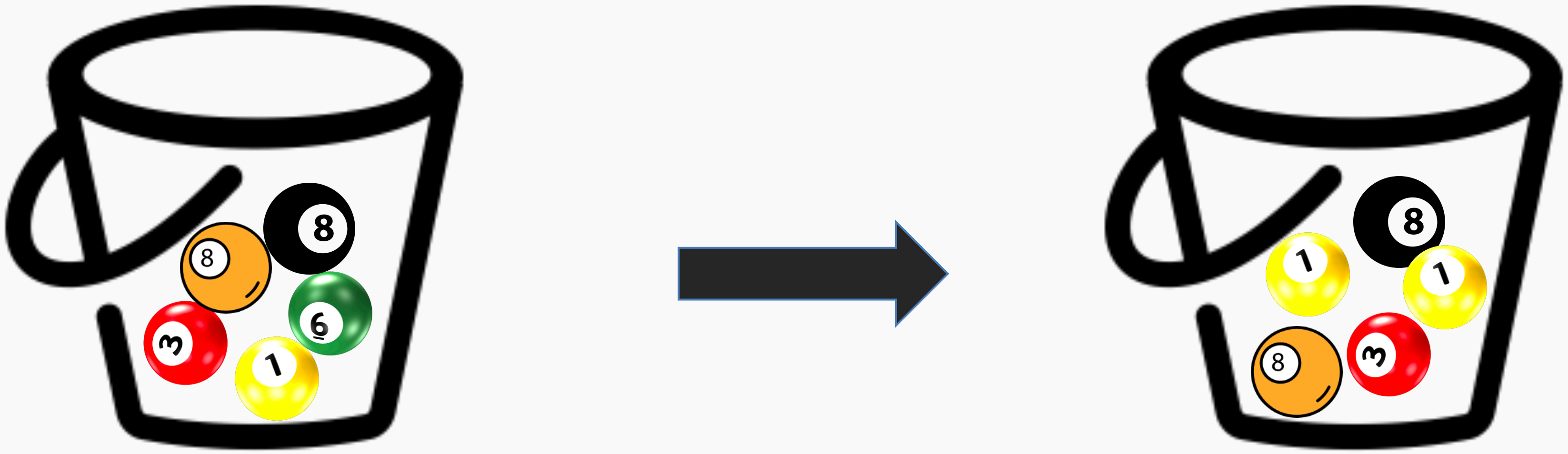
We continue until the “other” bucket has **the same number of balls** as the original one.



This new bucket represents a new parallel universe

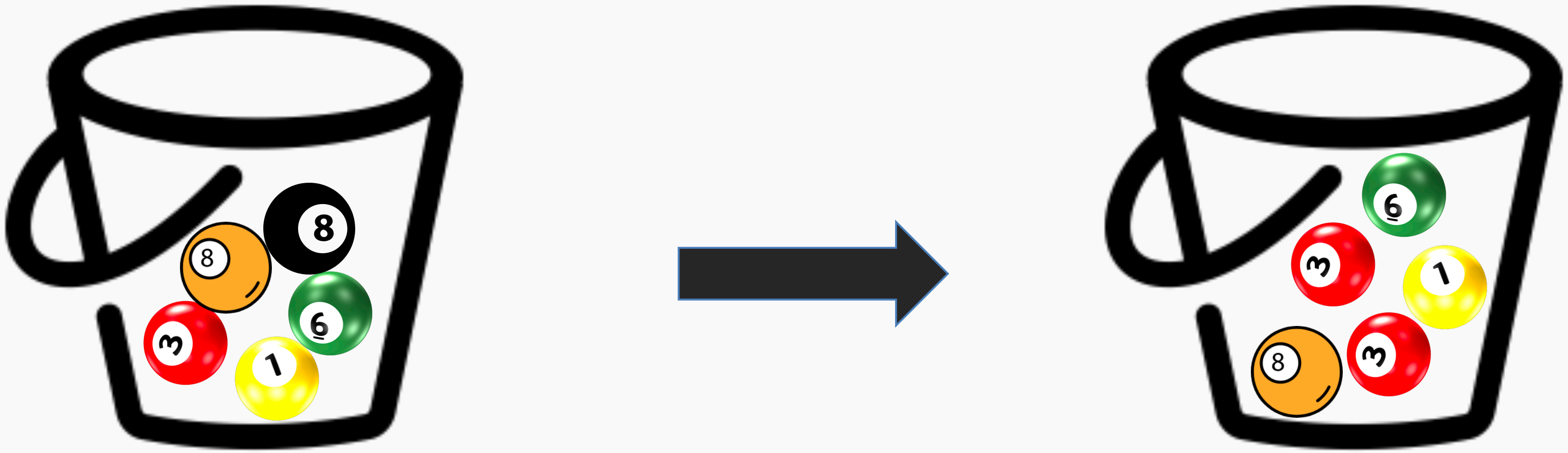
Bootstrap

We repeat the same process and acquire another sample.

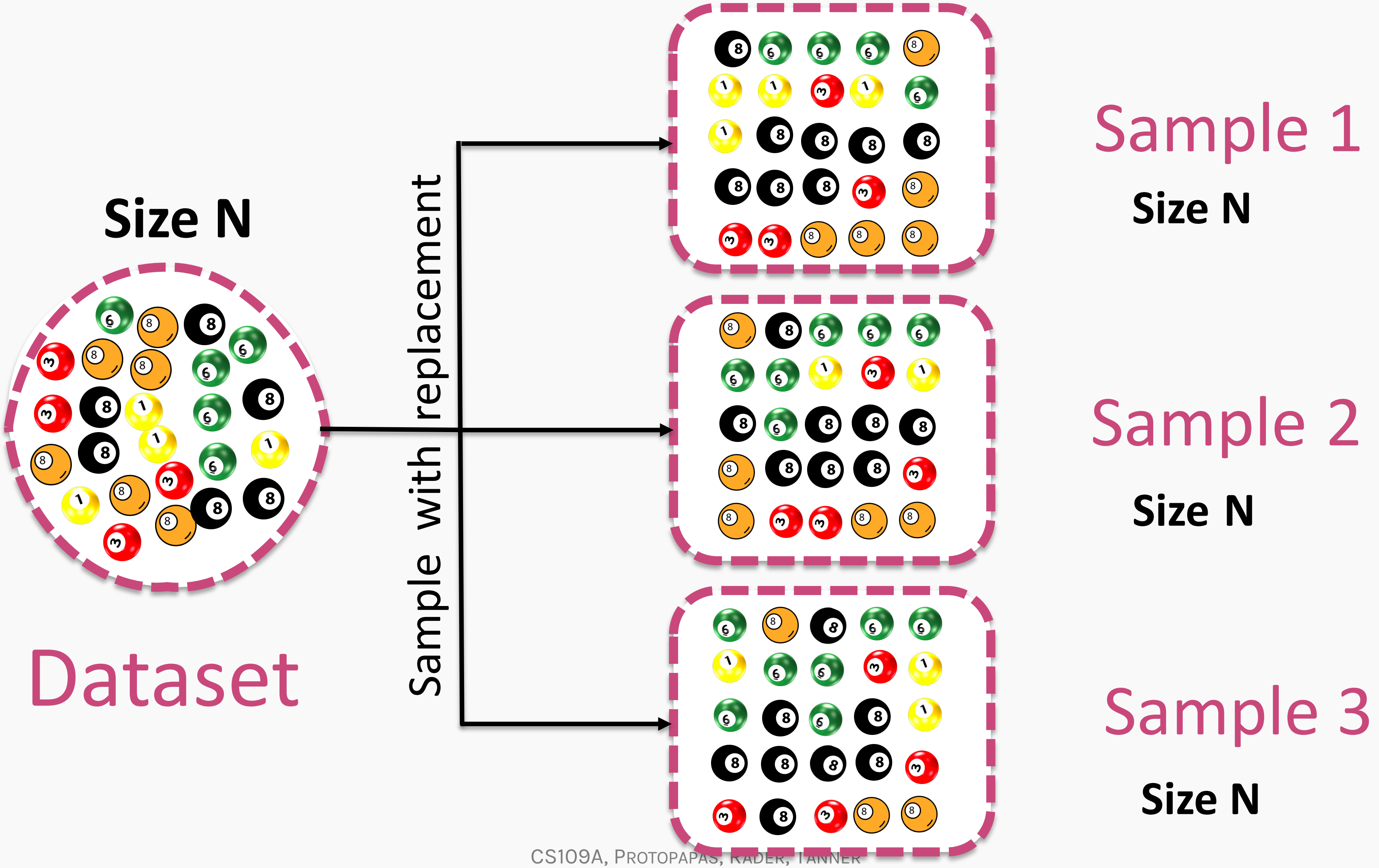


Bootstrap

We repeat the same process and acquire another sample.



Bootstrap


Size N
Dataset
Sample with replacement
Sample 1
Size N
Sample 2
Size N
Sample 3
Size N

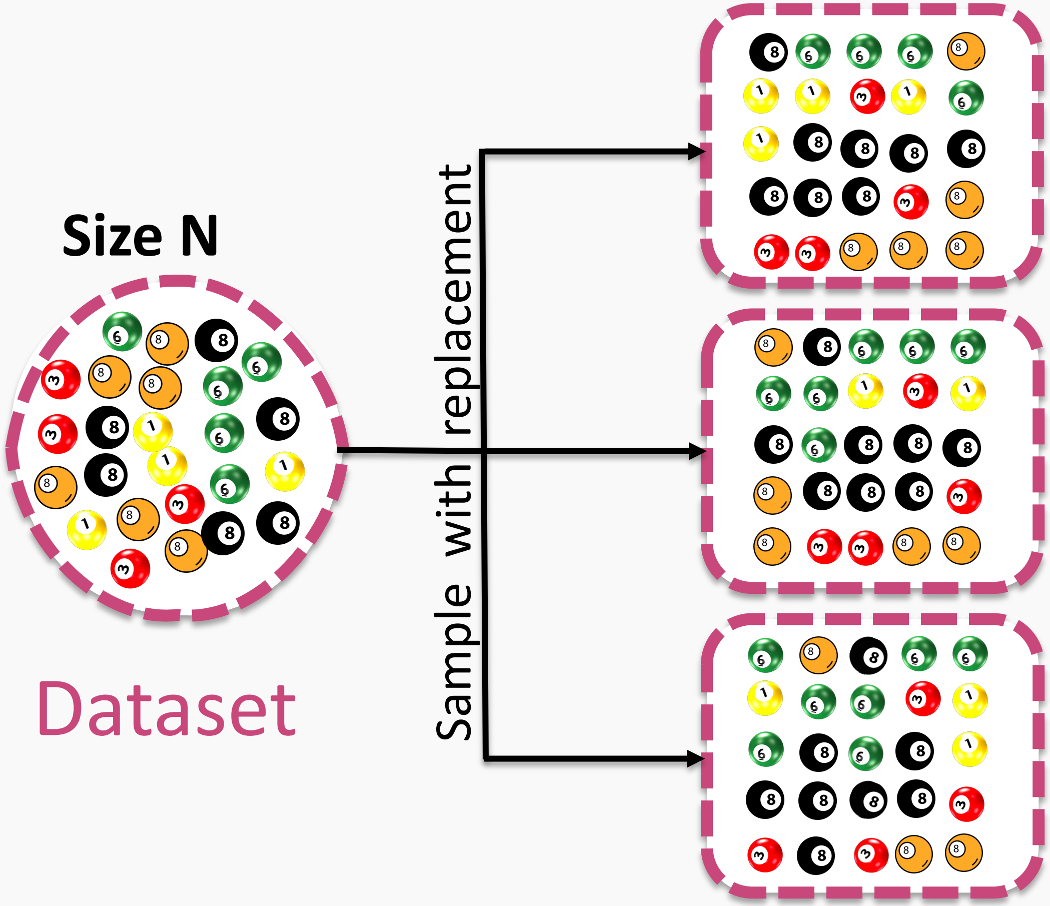
Bootstrapping for Estimating Sampling Error

Definition

Bootstrapping is the practice of estimating properties of an estimator by measuring those properties by, for example, sampling from the observed data.

For example, we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ multiple times by randomly sampling from our data set. We then use the variance of our multiple estimates to approximate the true variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

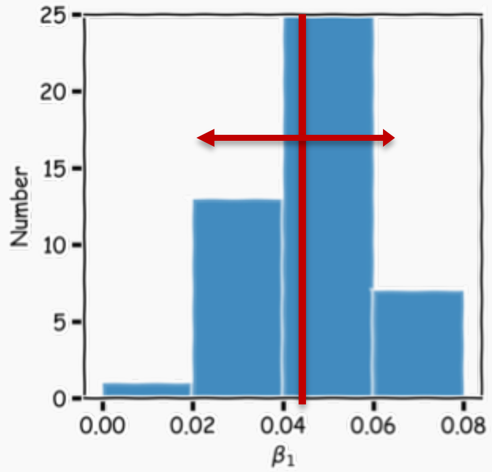
Bootstrap



train → Model 1: $\hat{y} = \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x$

train → Model 2: $\hat{y} = \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x$

train → Model s: $\hat{y} = \hat{\beta}_0^{(s)} + \hat{\beta}_1^{(s)} x$



$$\mu_{\hat{\beta}} = \frac{1}{s} \sum_{i=1}^{s} \hat{\beta}^{(i)}$$

$$\sigma_{\hat{\beta}} = \sqrt{\frac{1}{s} \sum_{i=1}^{s} \left(\hat{\beta}^{(i)} - \bar{\beta}\right)^2}$$

Confidence intervals for the predictors estimates: **Standard Errors**

We can empirically estimate the standard deviations $\sigma_{\hat{\beta}}$ which are called the **standard errors,** $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ through bootstrapping.

Alternatively:

If we know the variance σ_ϵ^2 of the noise ϵ, we can compute $SE(\hat{\beta}_0), SE(\hat{\beta}_1)$ analytically using the formulae below (no need to bootstrap):

$$SE\left(\hat{\beta}_0\right) = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$

$$SE\left(\hat{\beta}_1\right) = \frac{\sigma_\epsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Where n is the number of observations

$\bar{x}$ is the mean value of the predictor.

Standard Errors

More data: $n \uparrow$ and $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Larger coverage: $var(x)$ or $\sum_i (x_i - \bar{x})^2 \uparrow \Longrightarrow SE \downarrow$

Better data: $\sigma_\epsilon^2 \downarrow \Rightarrow SE \downarrow$

$$SE\left(\hat{\beta}_0\right) = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$

$$SE\left(\hat{\beta}_1\right) = \frac{\sigma(\epsilon)}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Better model: $(\hat{f} - y_i) \downarrow \Longrightarrow \sigma_\epsilon \downarrow \Longrightarrow SE \downarrow$

$$\sigma(\epsilon) = \sqrt{\sum \frac{\left(\hat{f}(x) - y_i\right)^2}{n-2}}$$

Question: What happens to the $\widehat{\beta_0}$, $\widehat{\beta_1}$ under these scenarios?

Standard Errors

In practice, we do not know the value of σ_ϵ since we do not know the exact distribution of the noise ϵ.

However, if we make the following assumptions,

- the errors $\epsilon_i = y_i - \hat{y}_i$ and $\epsilon_j = y_j - \hat{y}_j$ are uncorrelated, for $i \neq j$,
- each ϵ_i has a mean 0 and variance σ_ϵ^2,

then, we can empirically estimate σ^2, from the data and our regression line:

$$\sigma_\epsilon = \sqrt{\frac{n \cdot MSE}{n-2}} = \sqrt{\sum \frac{\left(\hat{f}(x) - y_i\right)^2}{n-2}}$$

Remember: $y_i = f(x_i) + \epsilon_i \Longrightarrow \epsilon_i = y_i - f(x_i)$

Standard Errors

The following results are for the coefficients for TV advertising:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0061
Bootstrap	0.0061

The coefficients for TV advertising but restricting the coverage of x are:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.0068
Bootstrap	0.0068

SE increase

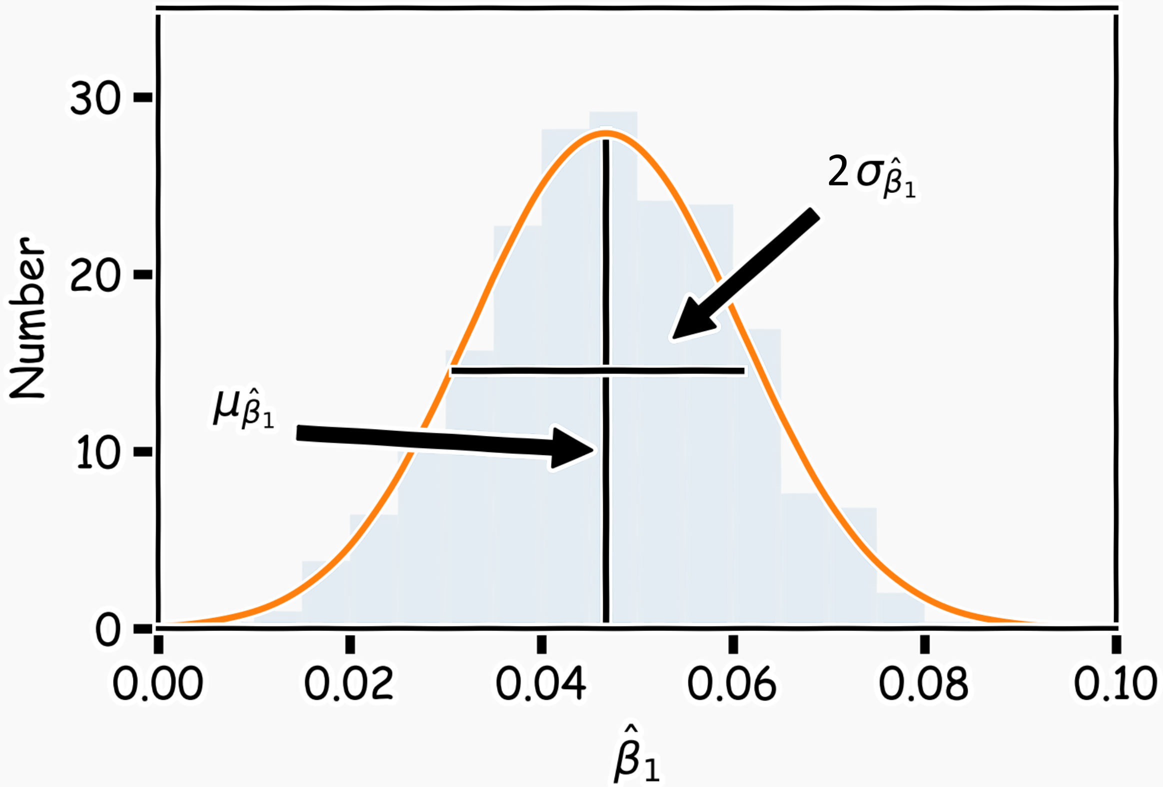
The coefficients for TV advertising but with added **extra** noise:

Method	$SE(\hat{\beta}_1)$
Analytic Formula	0.028
Bootstrap	0.023

SE increase

Confidence intervals for the predictors estimates (cont)

We can now estimate the mean and standard deviation of the estimates of $\hat{\beta}_0, \hat{\beta}_1$.

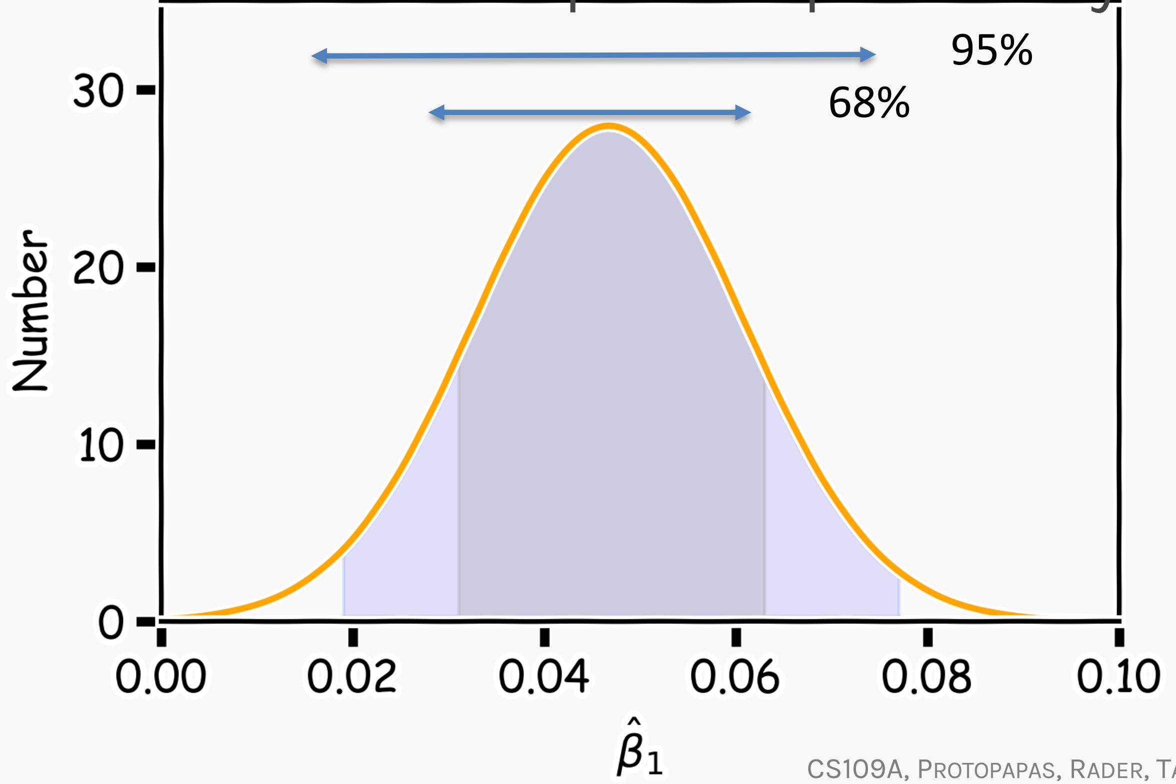


21

Confidence intervals for the predictors estimates (cont)

The standard errors give us a sense of our uncertainty over our estimates.

Typically we express this uncertainty as a 95% confidence interval, which is the range of values such that the **true** value of β_1 is contained in this interval with *95%* percent probability.



$$\boldsymbol{CI_{\hat{\beta}} = (\hat{\beta} - 2\sigma_{\hat{\beta}}, \hat{\beta} + 2\sigma_{\hat{\beta}})}$$