Bootstrapping and Confidence Intervals

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai and Chris Gumb

Outline

Part A and B: Assessing the Accuracy of the Coefficient Estimates

Bootstrapping and confidence intervals

Part C: Evaluating Significance of Predictors

Does the outcome depend on the predictors?

Hypothesis testing

Part D: How well do we know $\hat{f}$

The confidence intervals of $\hat{f}$

Lack of Active Imagination

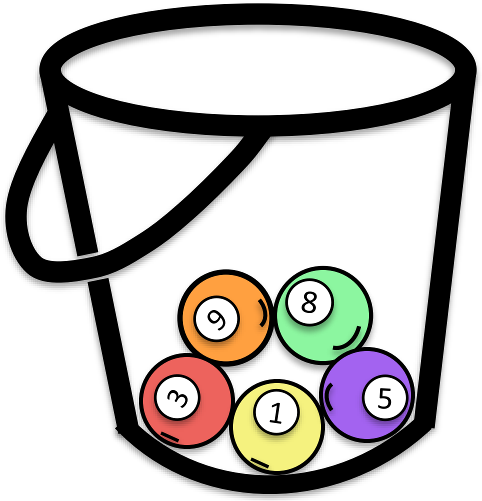
In the lack of active imagination, parallel universes and the likes, we need an alternative way of producing fake data set that resemble the parallel universes.

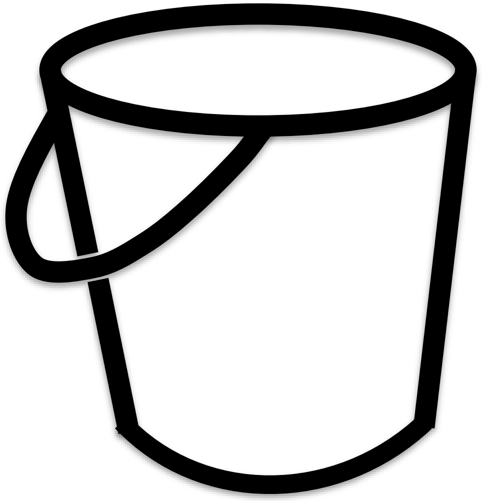
Bootstrapping is the practice of sampling from the observed data (X,Y) in estimating statistical properties.

*Note: this is not to create synthetic data to add to the actual observed data. It is to mimic the alternative universes mentioned previously.

Bootstrap

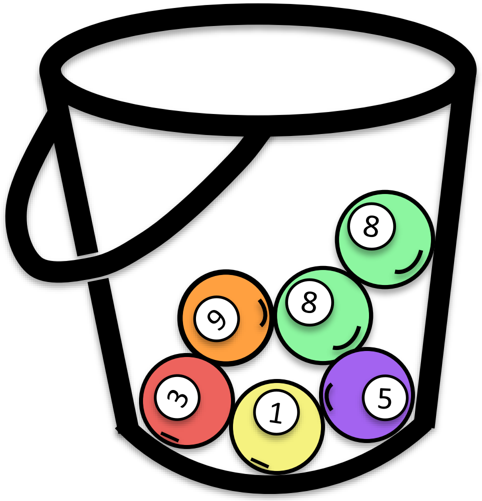
Imagine we have 5 billiard balls in a bucket.

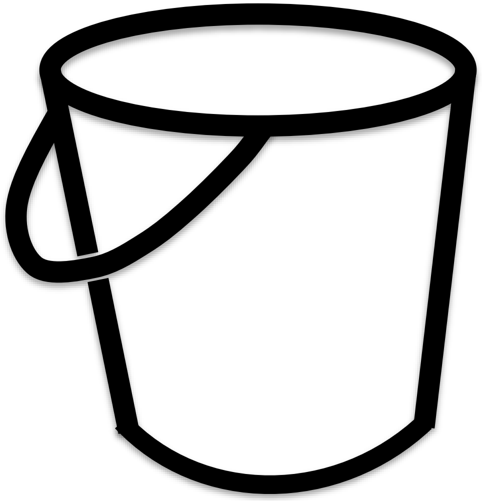




Bootstrap

We first pick randomly a ball and replicate it.





This is called **sampling with replacement.**

Bootstrap

We move the replicated ball to another bucket.

Bootstrap

We then randomly pick another ball and again we replicate it.

Bootstrap

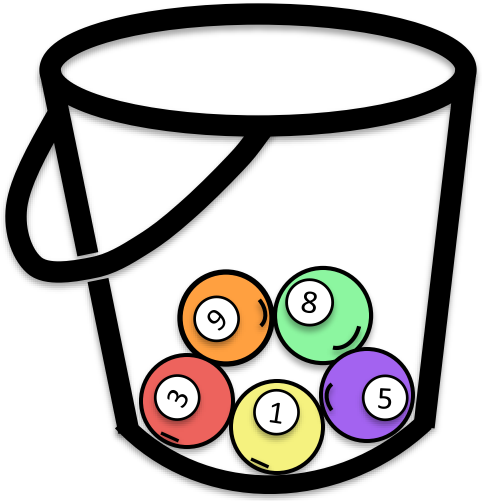
As before, we move the replicated ball to the other bucket.

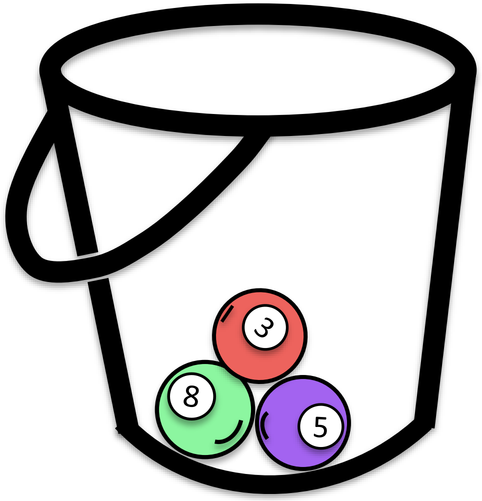
Bootstrap

We repeat this process.

Bootstrap

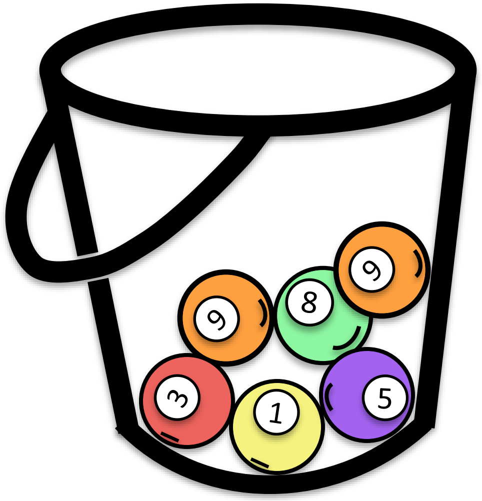
We repeat this process.

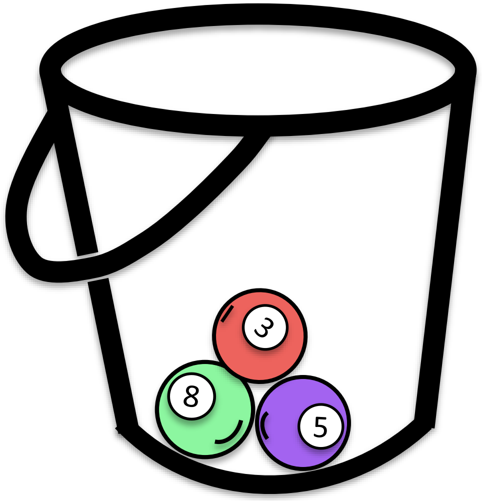




Bootstrap

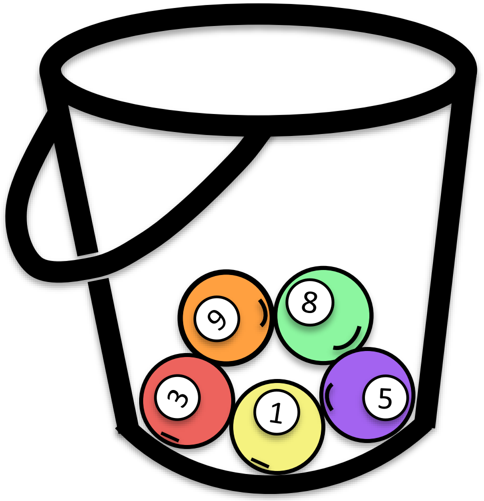
We repeat this process.

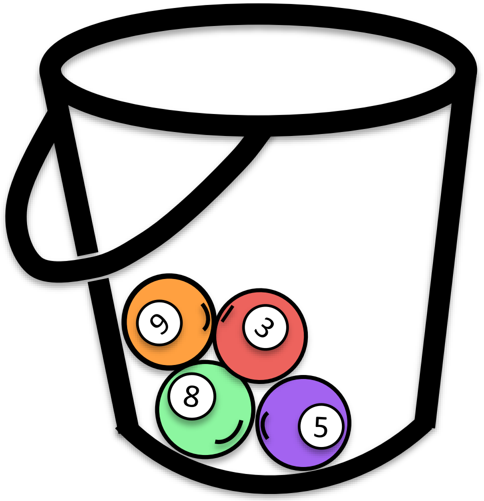




Bootstrap

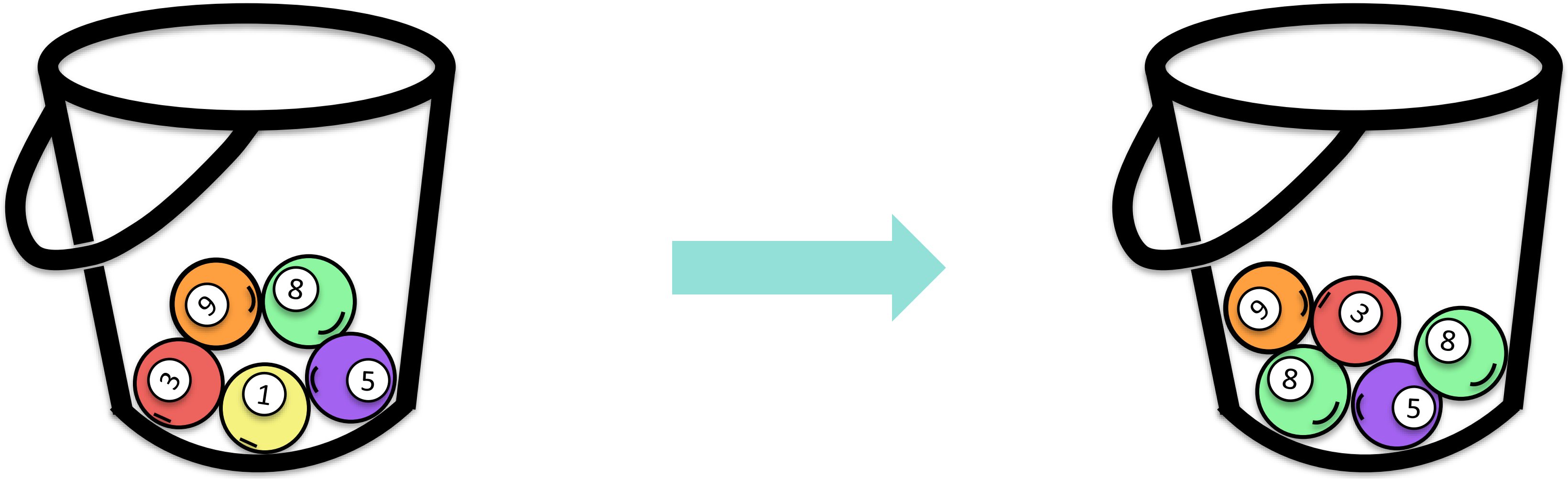
We repeat this process.





Bootstrap

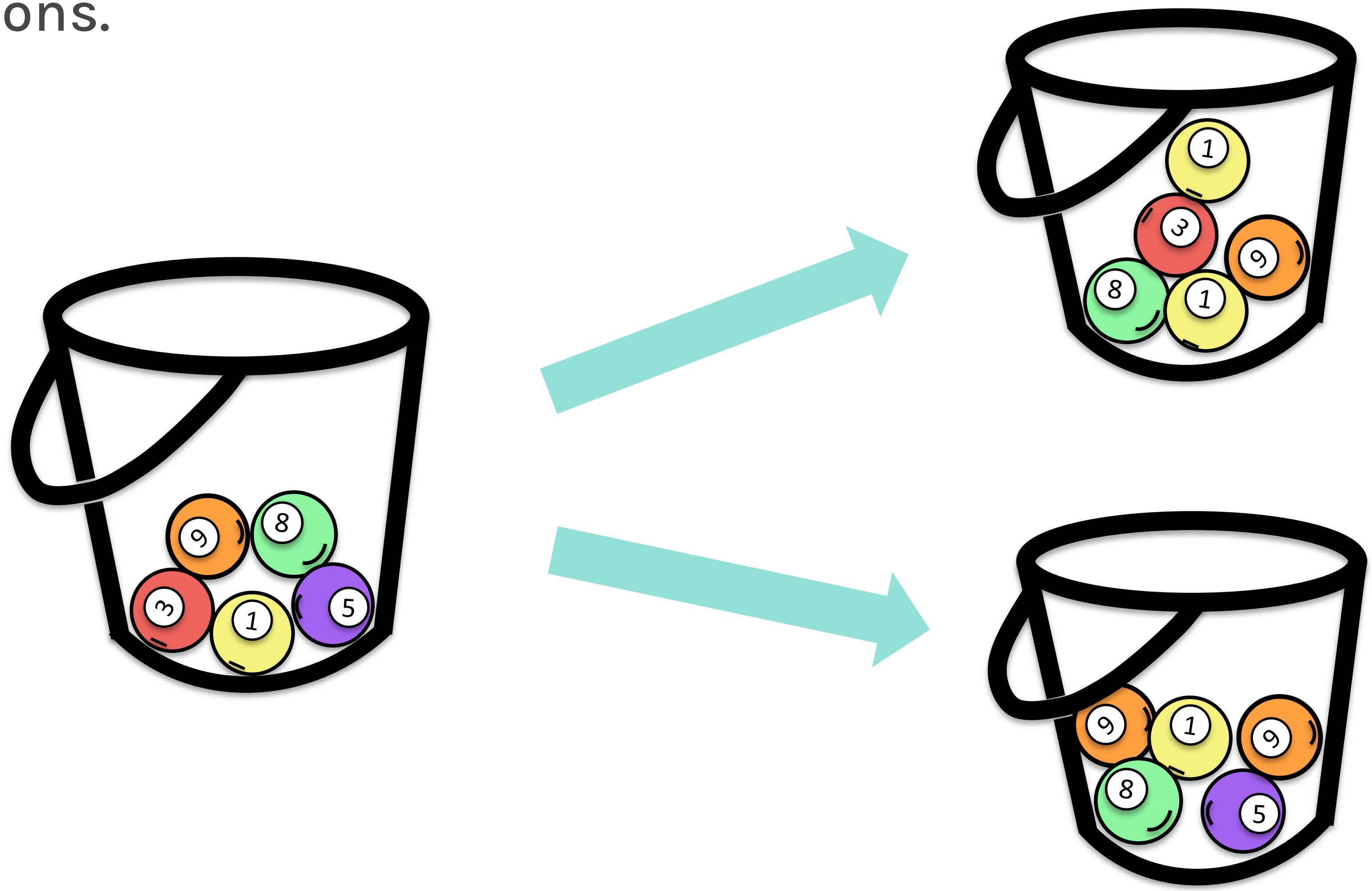
We continue until the “other” bucket has **the same number of balls** as the original one.



This new bucket represents a new parallel universe

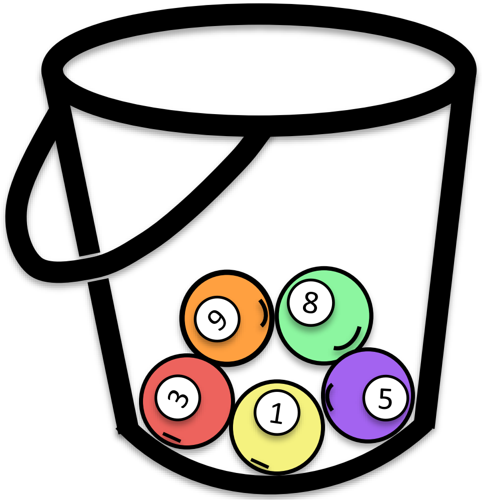
Bootstrap

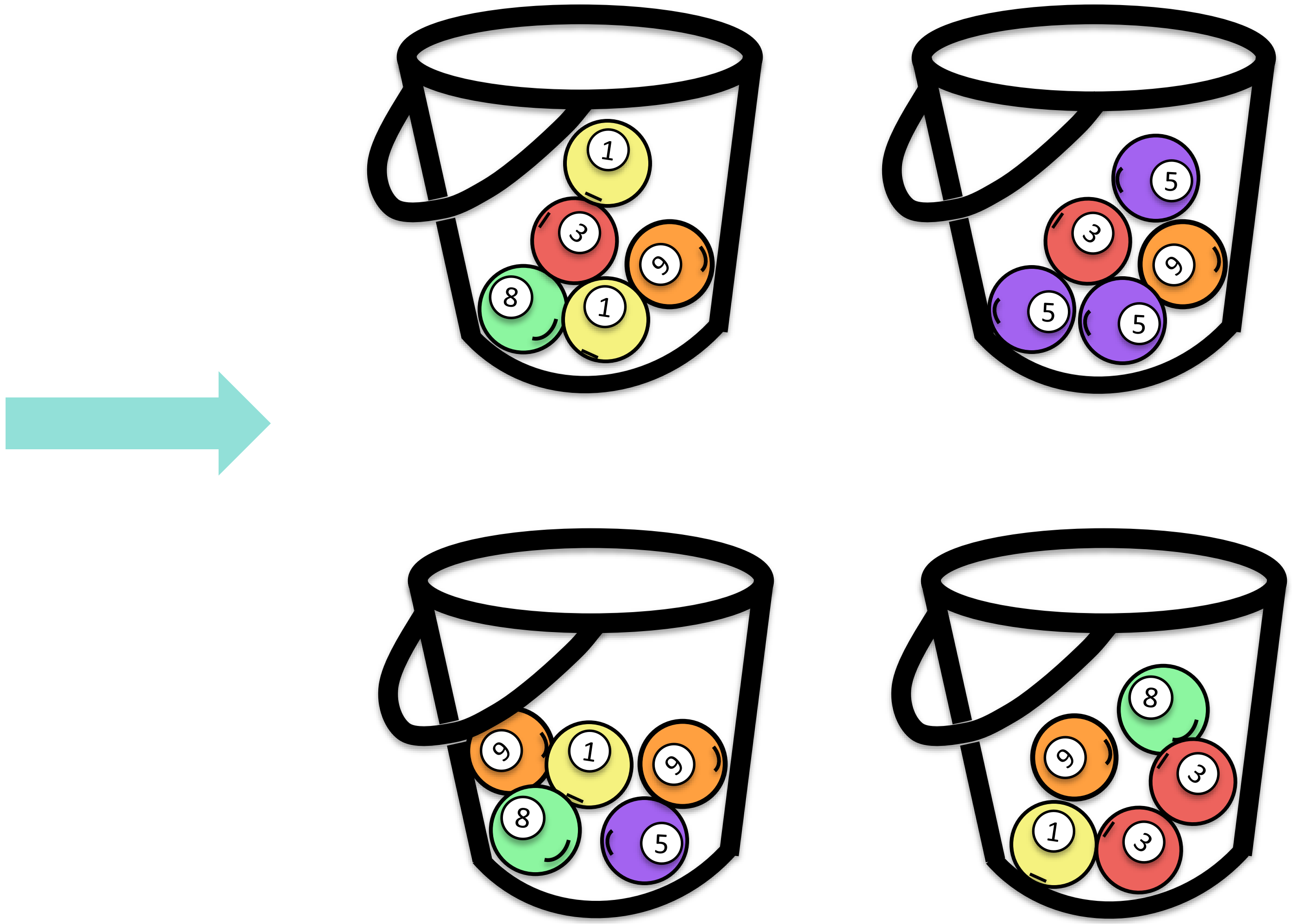
We repeat the same process and acquire another set of bootstrapped observations.



Bootstrap

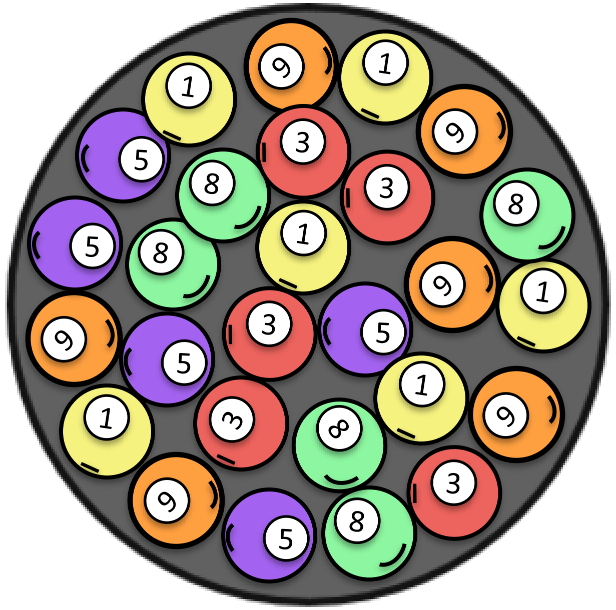
We repeat the same process and acquire another set of bootstrapped observations.





DATASET

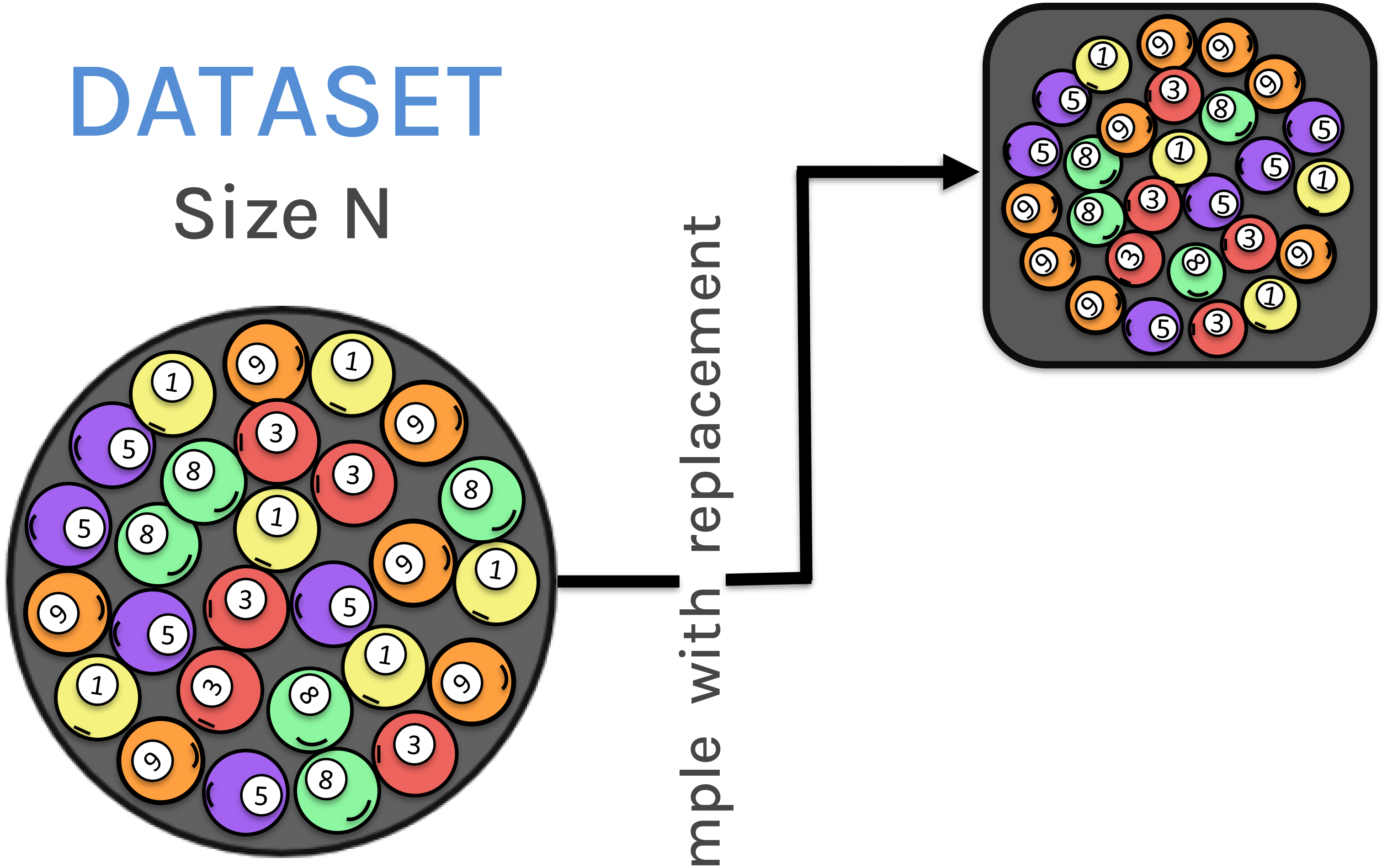
Size N



Bootstrap

DATASET

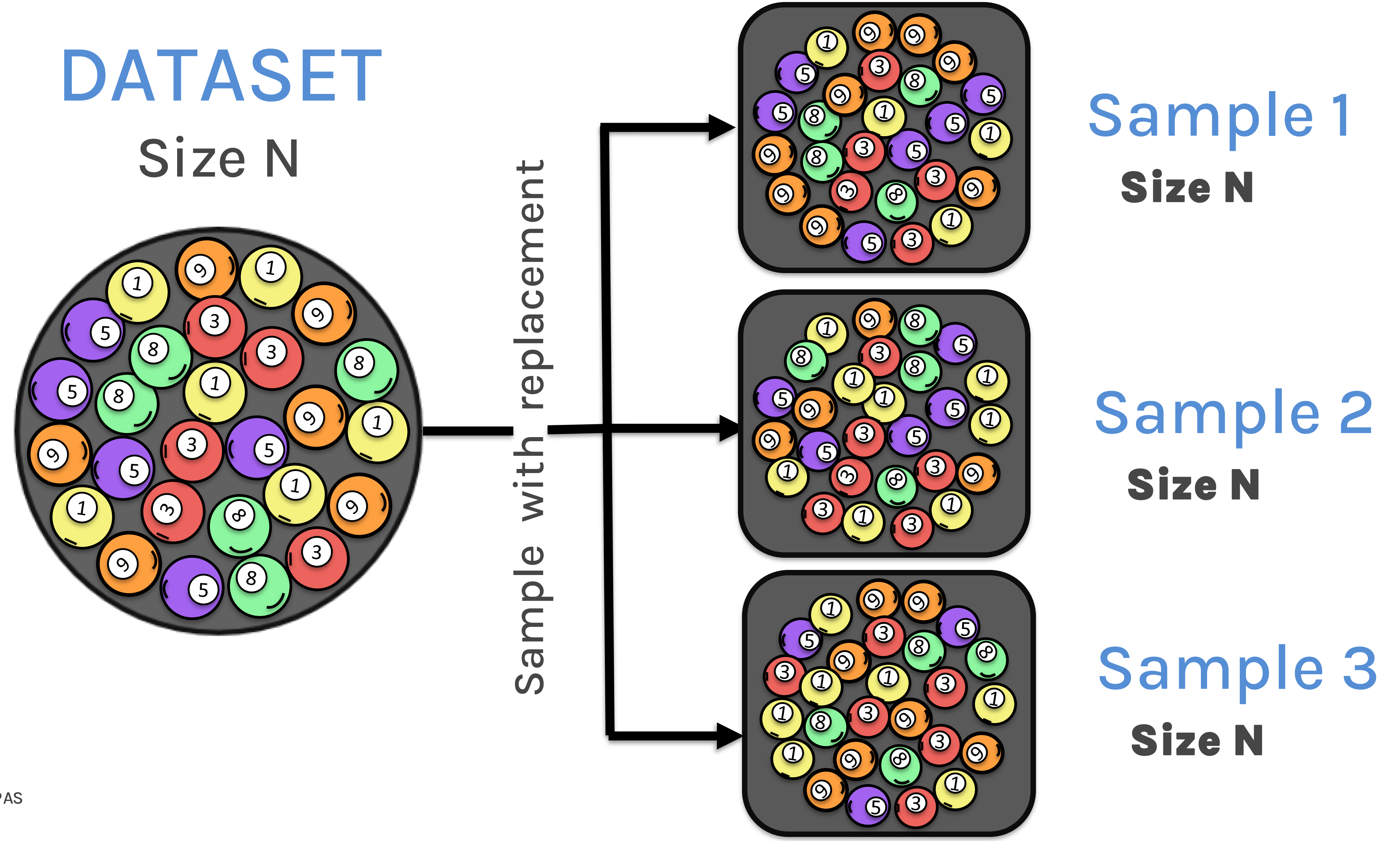
Size N



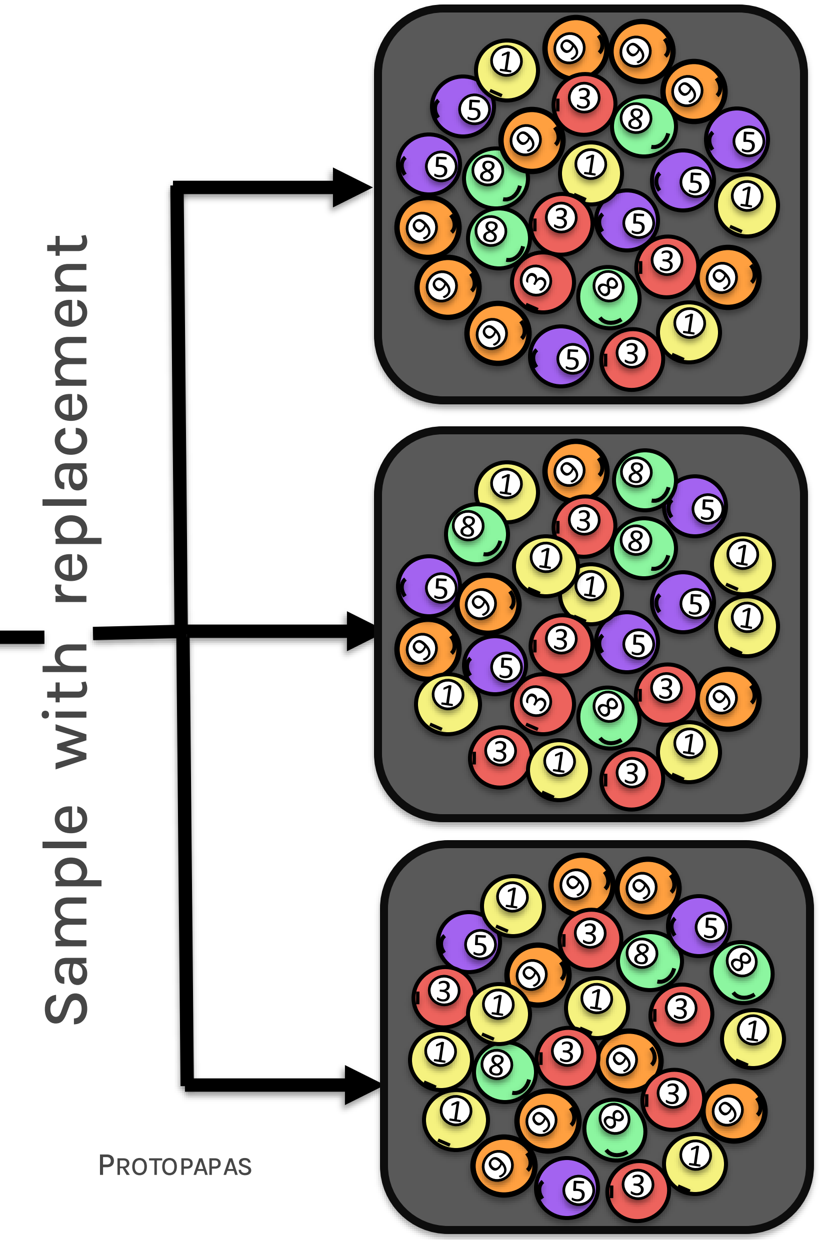
Sample 1

Size N

Bootstrap


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

Bootstrap



Sample 1

Size N

$\xrightarrow{\text{Train}}$ Model 1: $\hat{y} = \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x$

Sample 2

Size N

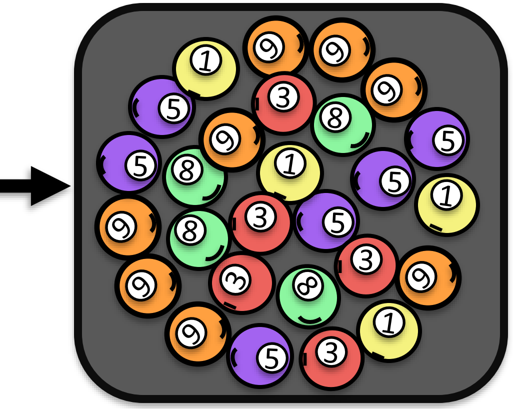
$\xrightarrow{\text{Train}}$ Model 2: $\hat{y} = \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x$

Sample 3

Size N

$\xrightarrow{\text{Train}}$ Model s: $\hat{y} = \hat{\beta}_0^{(s)} + \hat{\beta}_1^{(s)} x$

Bootstrap

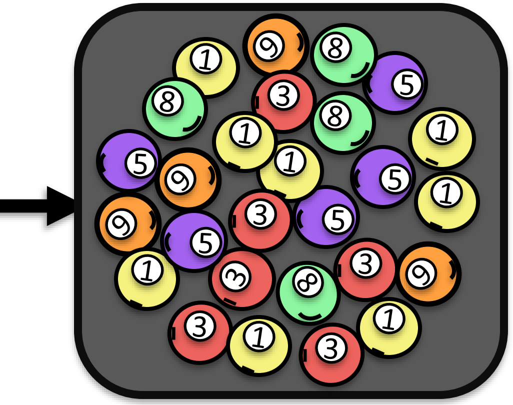


Sample 1

Size N

Train →

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

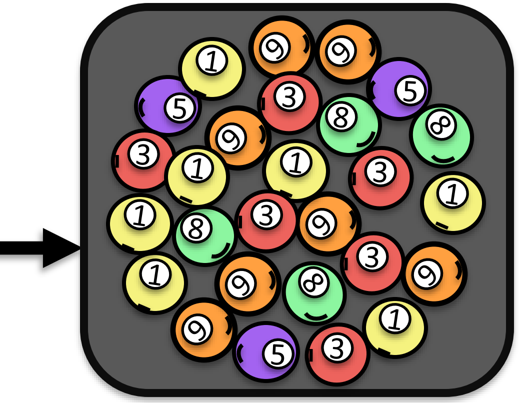


Sample 2

Size N

Train →

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



Sample 3

Size N

Train →

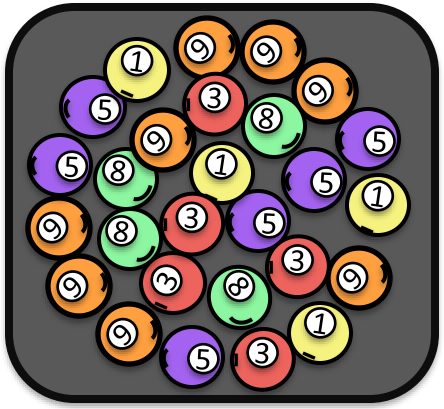
Model s: $\hat{y} = \hat{\beta}_0^{(s)} + \hat{\beta}_1^{(s)} x$

Combine models

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

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

In summary, for each “Parallel Universe”…



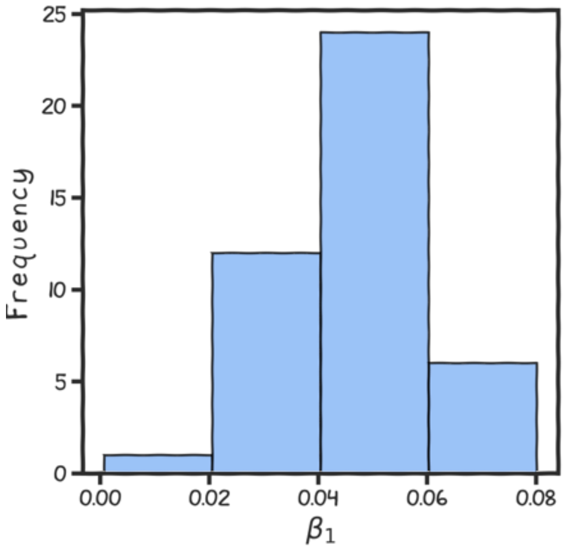
Train

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

Combine
all models

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

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



s models

Bootstrapping for Estimating Sampling Error

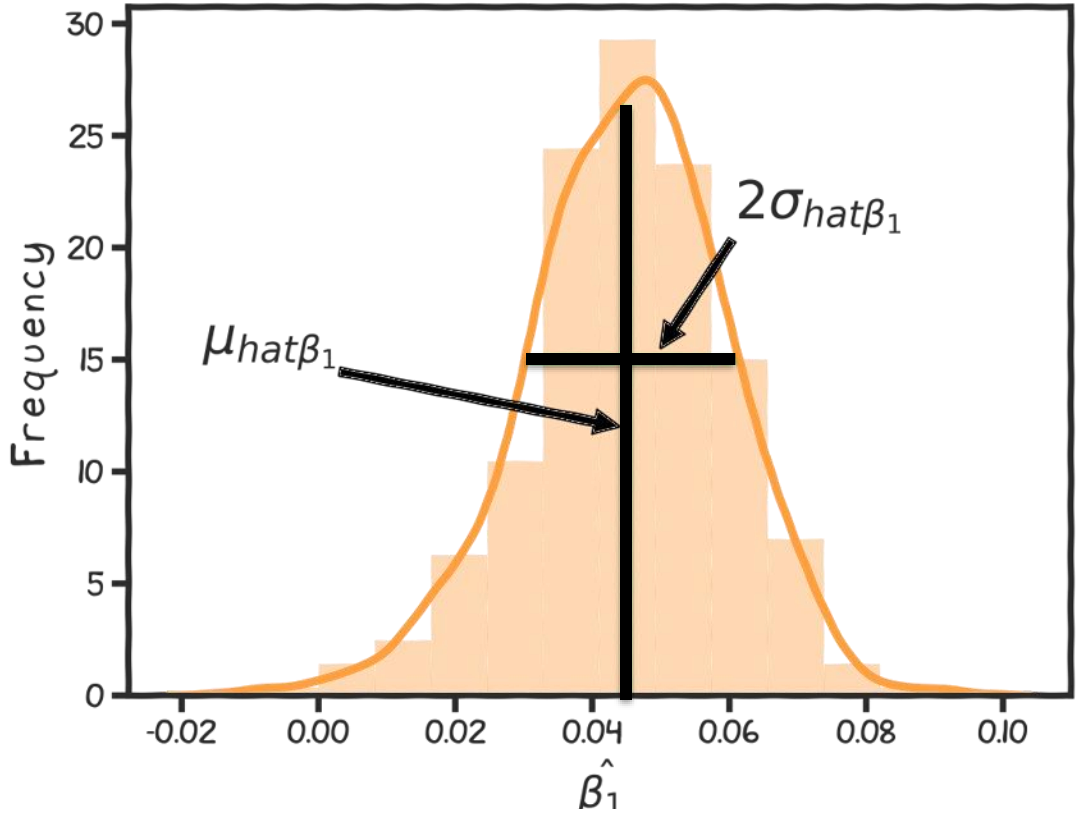
Definition

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

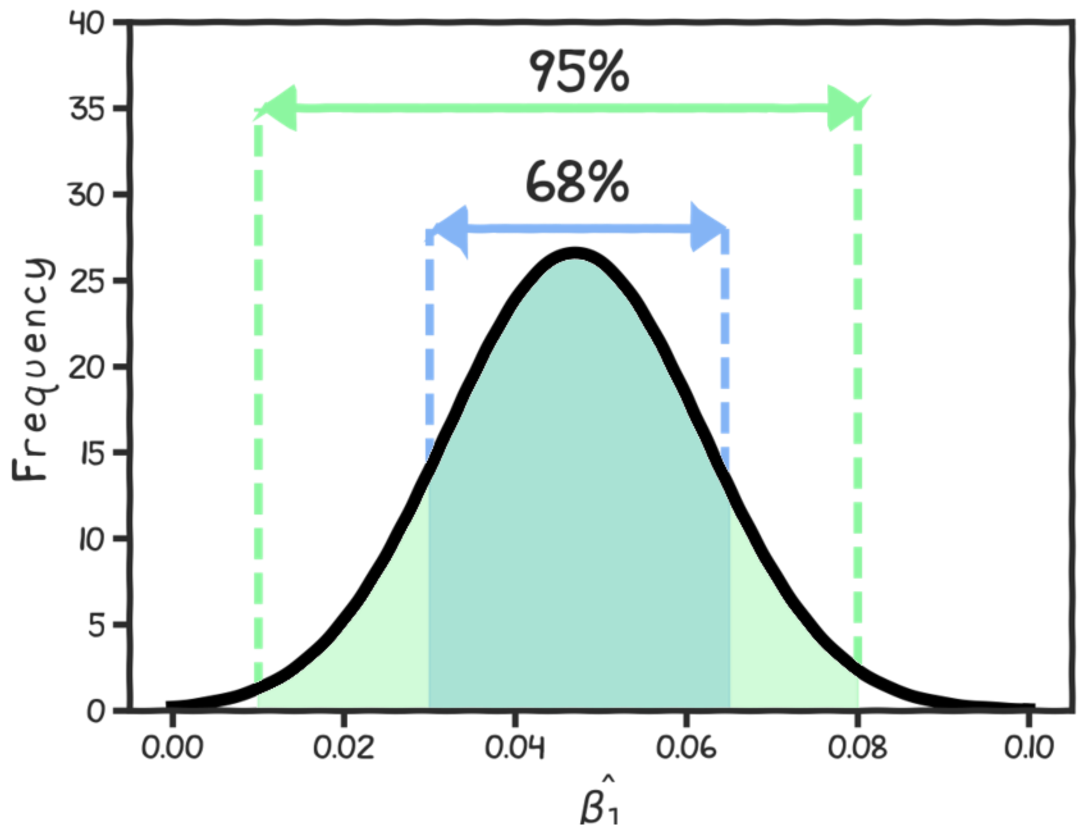
For example, we can compute $\hat{\beta}_0$ and $\hat{\beta}_1$ multiple times by randomly sampling from our data set. We then use the variance of our multiple estimates to approximate the true variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Confidence intervals for the predictors estimates (cont)

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



Confidence intervals for the predictor 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.

Standard Errors based on probability theory

Alternatively: If we assume normality, then:

And 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):

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

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

$$\boldsymbol{CI_{\widehat{\beta}}(95\%)} = [\widehat{\boldsymbol{\beta}} - \mathbf{2}SE(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\beta}} + \mathbf{2}SE(\widehat{\boldsymbol{\beta}})]$$

Where n is the number of observations.

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

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