Bootstrapping and Confidence Intervals

Pavlos Protopapas, Ignacio Becker

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 $\hat{f}$

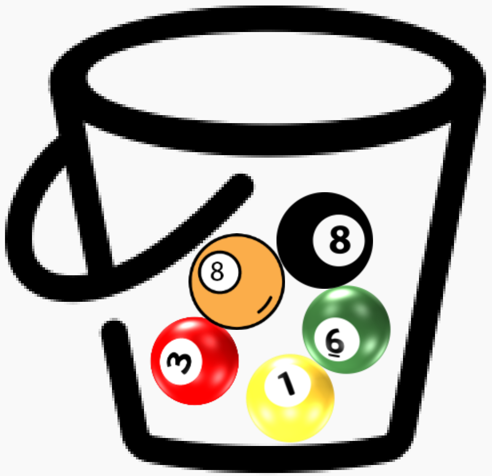
Lack of Active Imagination

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.

Bootstrap

Imagine we have 5 billiard balls in a bucket.



Bootstrap

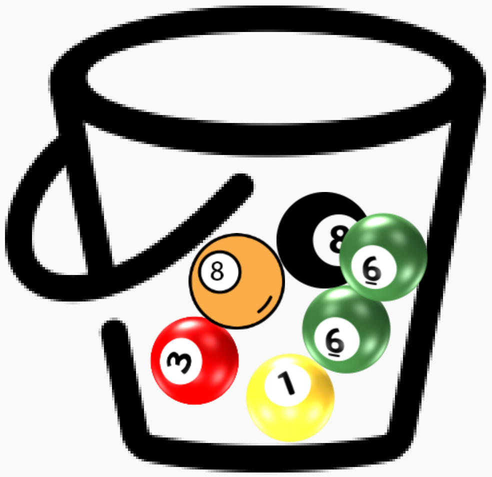
We first pick randomly a ball and replicate it. This is called **sampling with replacement.**

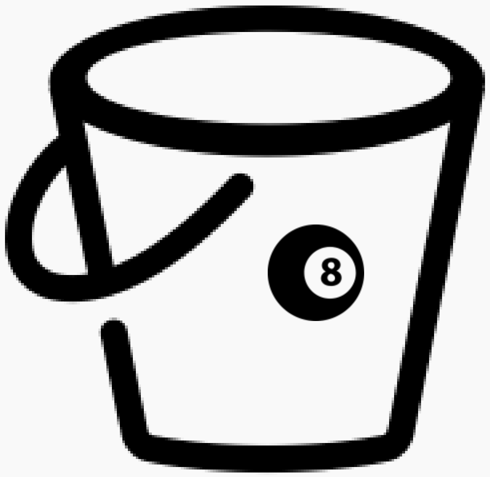
Bootstrap

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

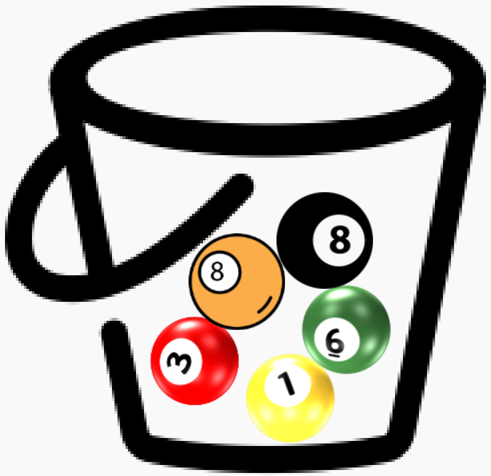
We repeat this process.

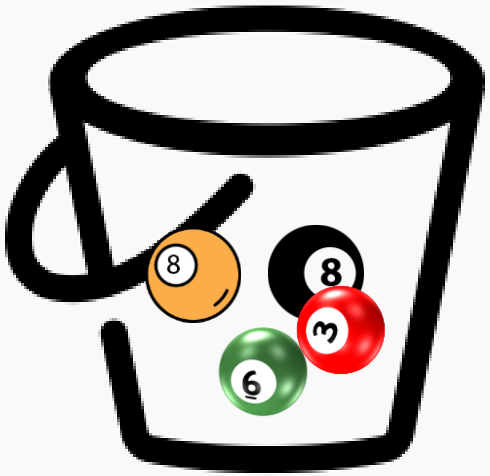
Bootstrap

We repeat this process.

Bootstrap

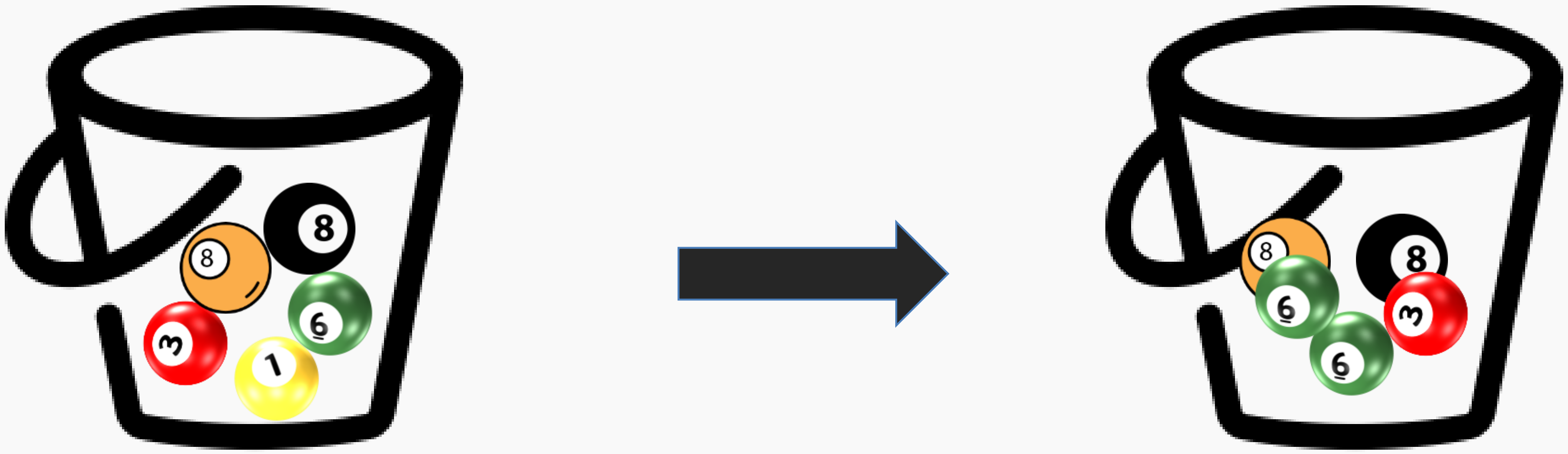
Again





Bootstrap

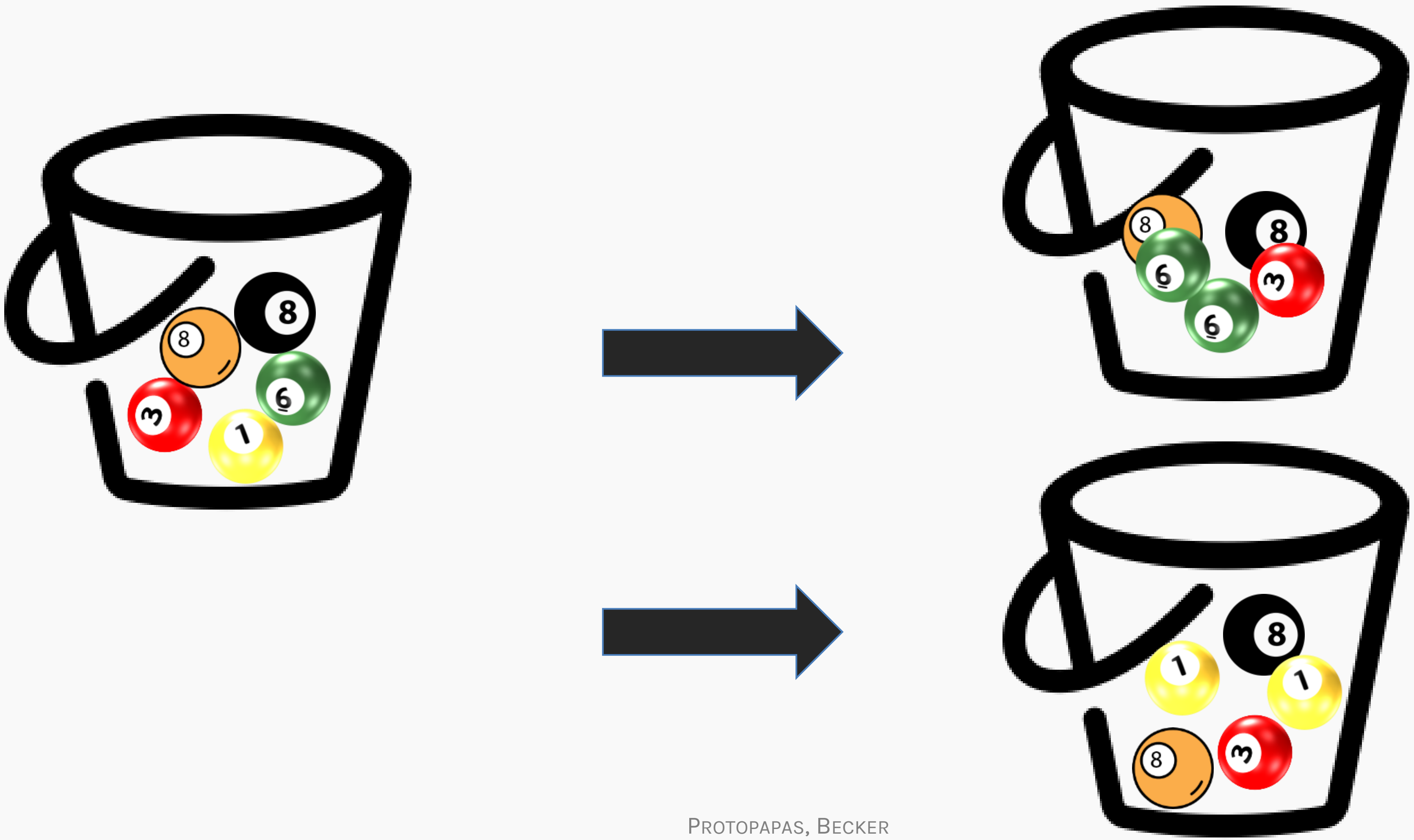
We continue until the "other" bucket has **the same number of balls** as the original one.



This new bucket represents a new parallel universe

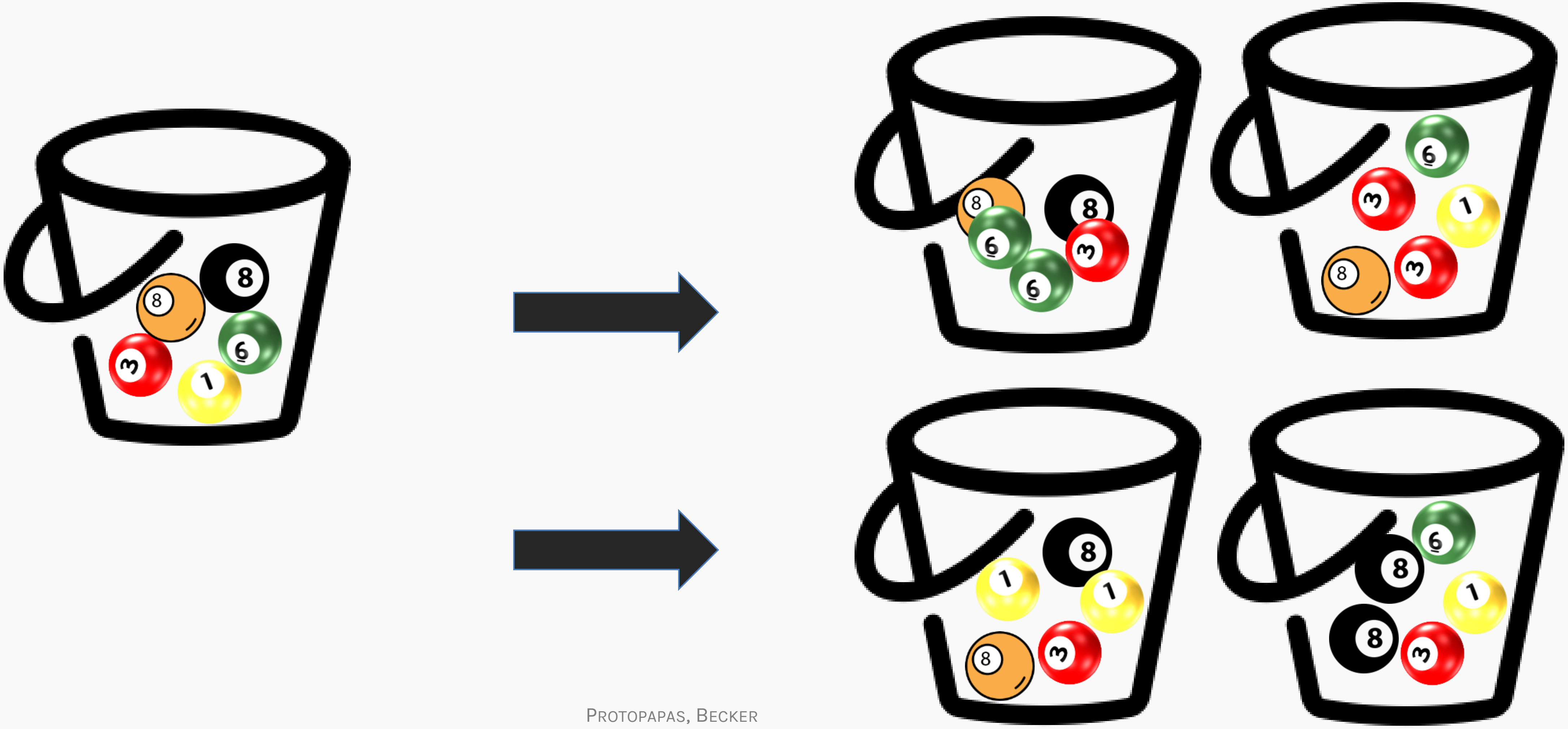
Bootstrap

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



Bootstrap

We repeat the same process and acquire many bootstrapped observations.

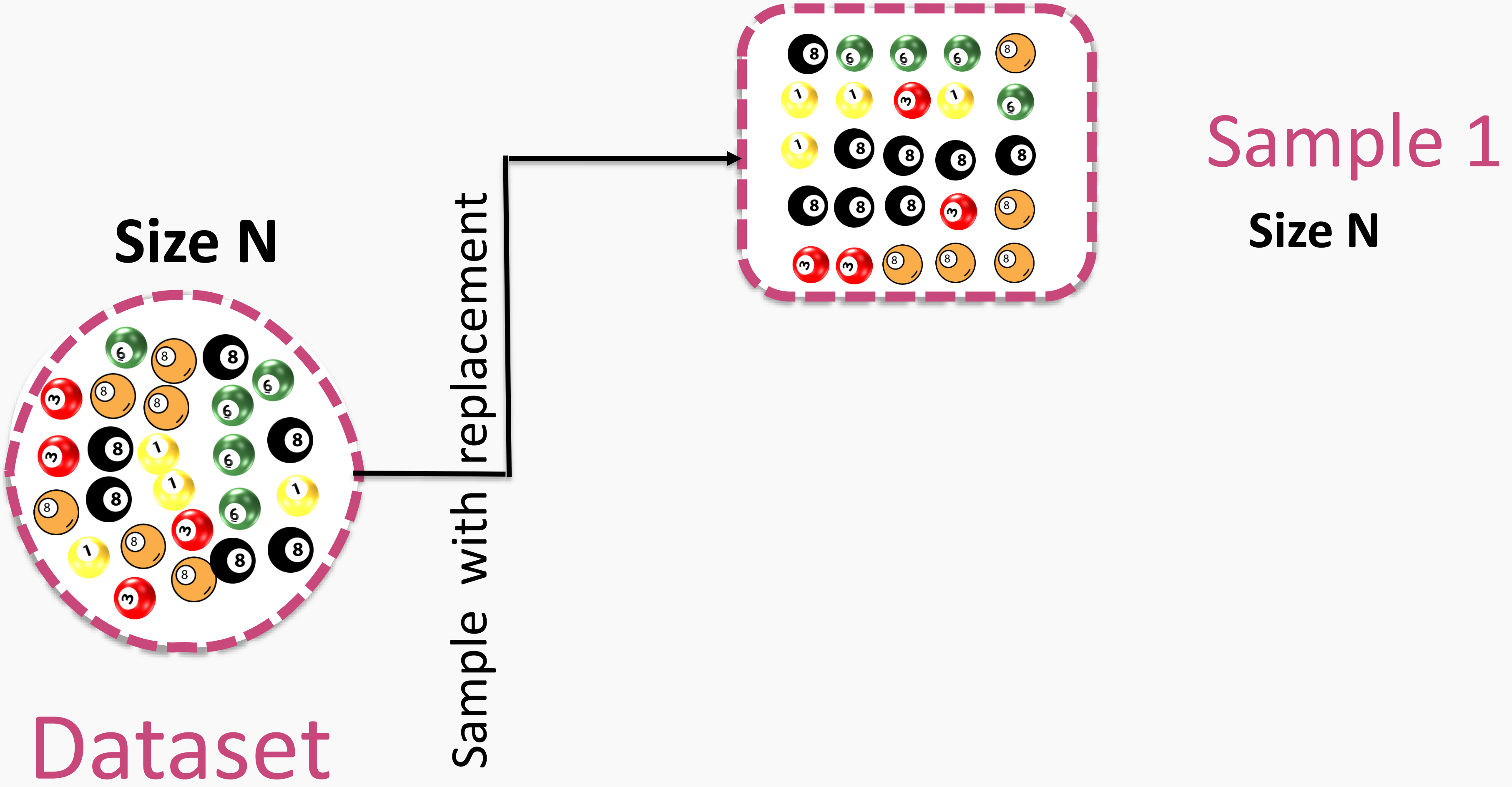


Bootstrap

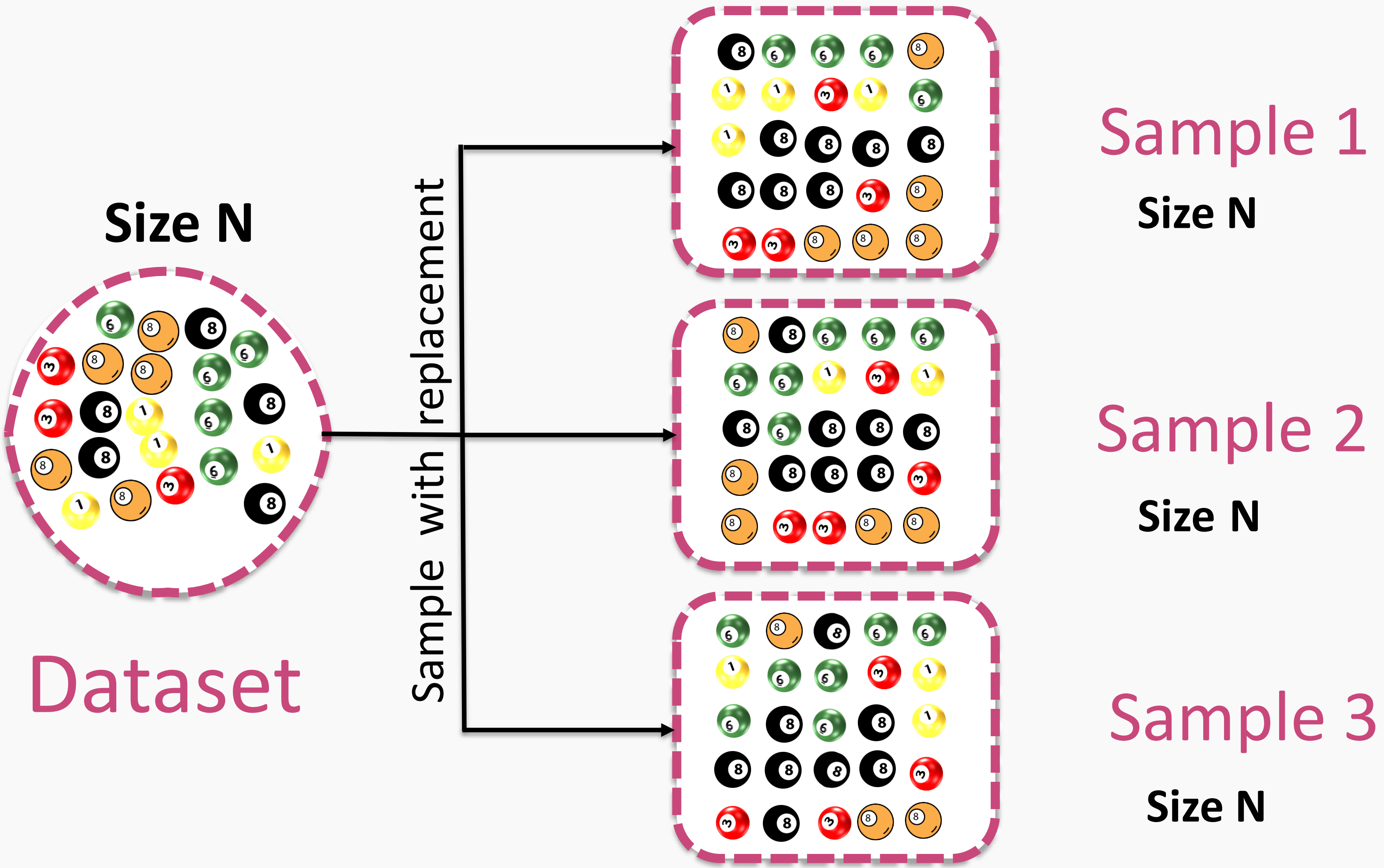
Size N

Dataset

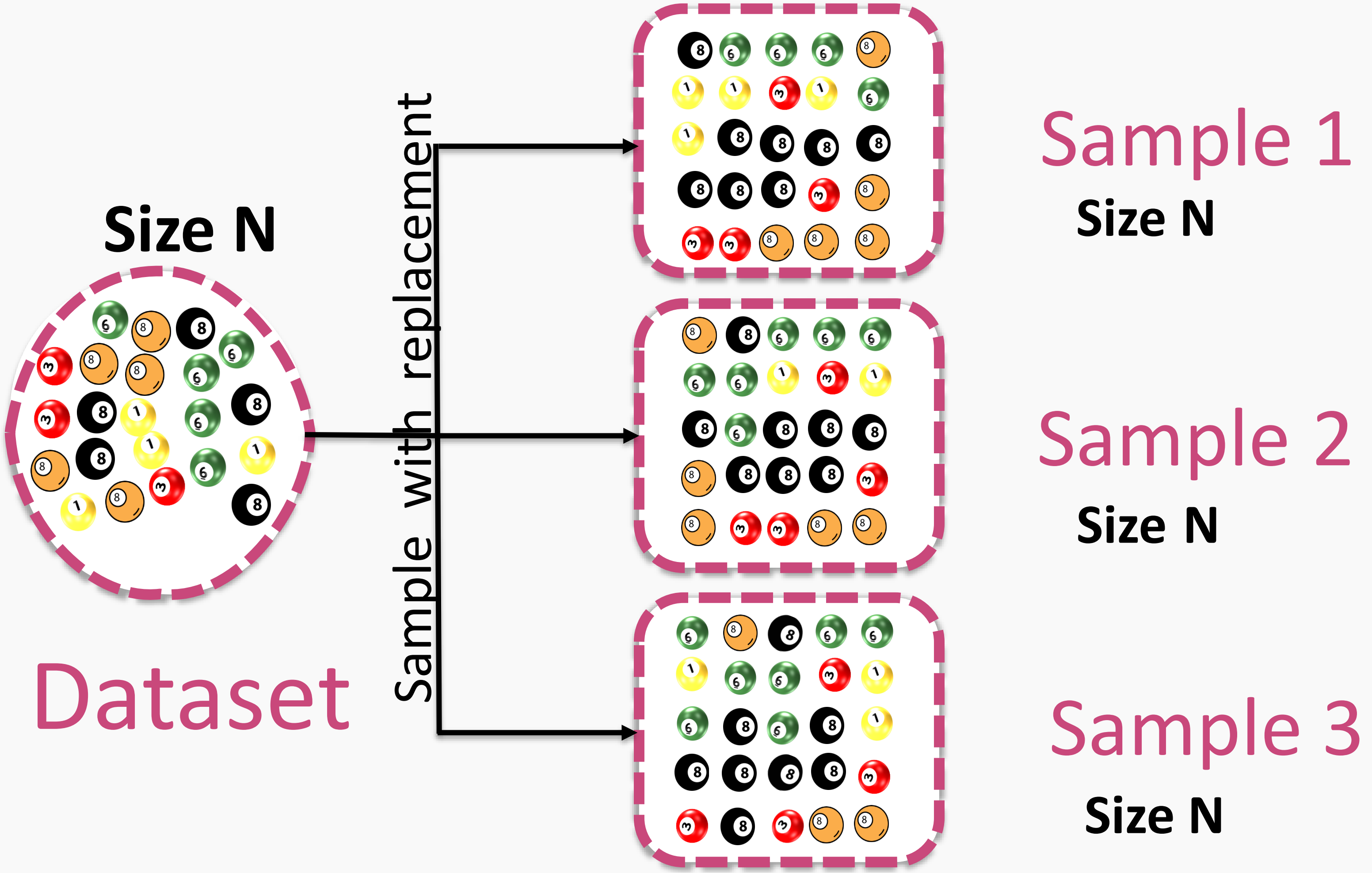
Bootstrap



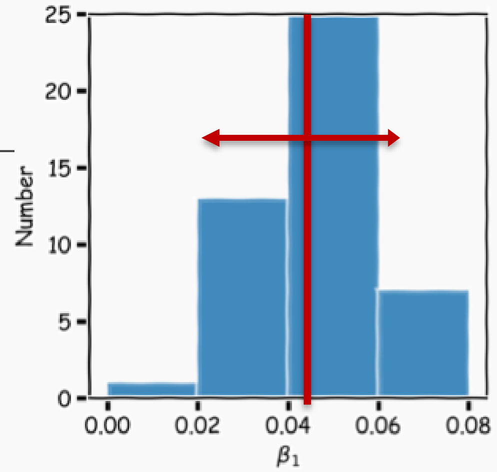
Bootstrap

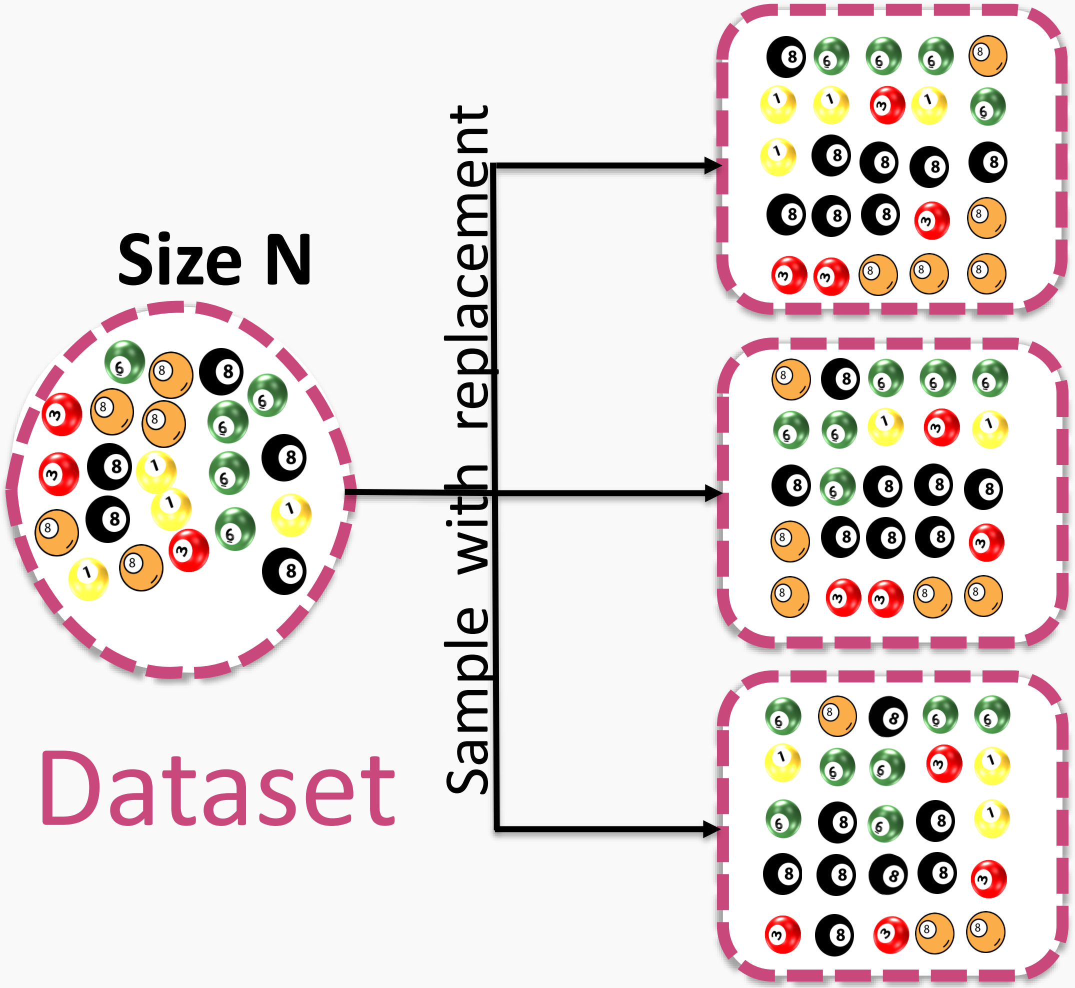

Size N
Dataset
Sample with replacement
Sample 1
Size N
Sample 2
Size N
Sample 3
Size N

Bootstrap


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

Bootstrap





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

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

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

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

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

Bootstrapping for Estimating Sampling Error

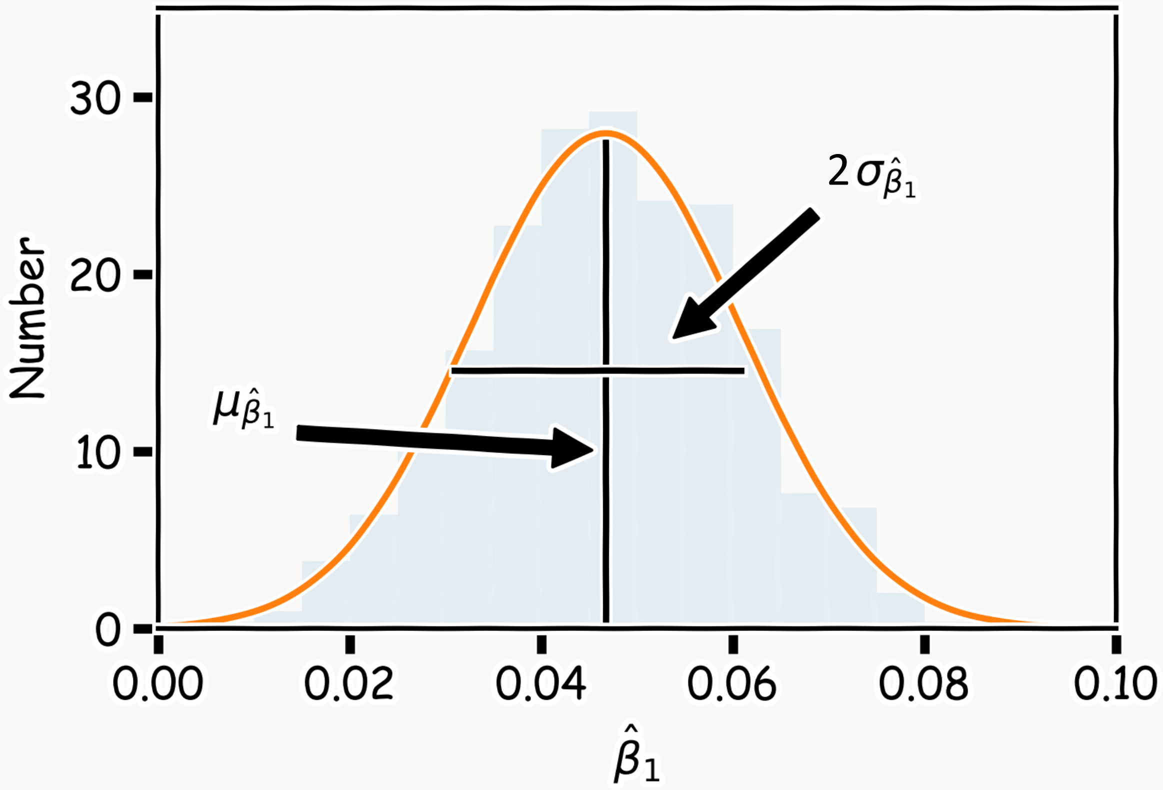
Definition

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 $\hat{\beta}_1$.

Confidence intervals for the predictors estimates (cont)

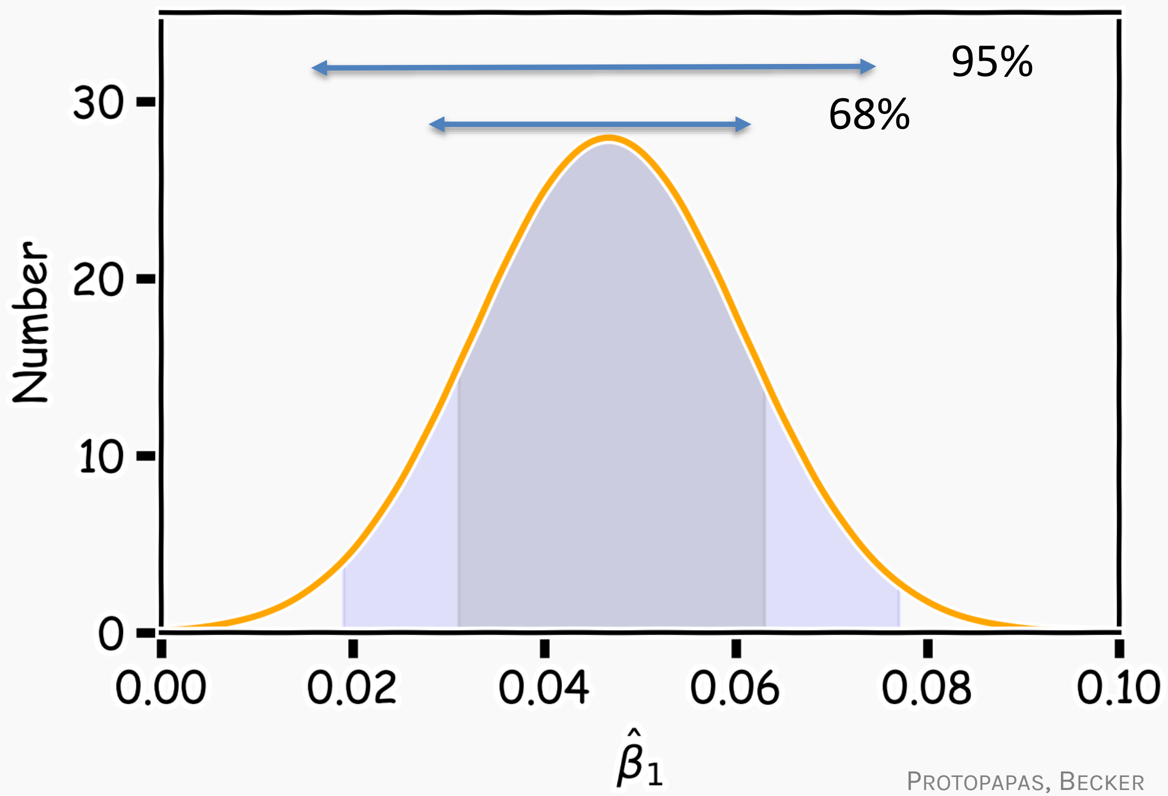
We can now estimate the mean and standard deviation of the estimates of $\hat{\beta}_0, \hat{\beta}_1$.



Confidence intervals for the predictors estimates (cont)

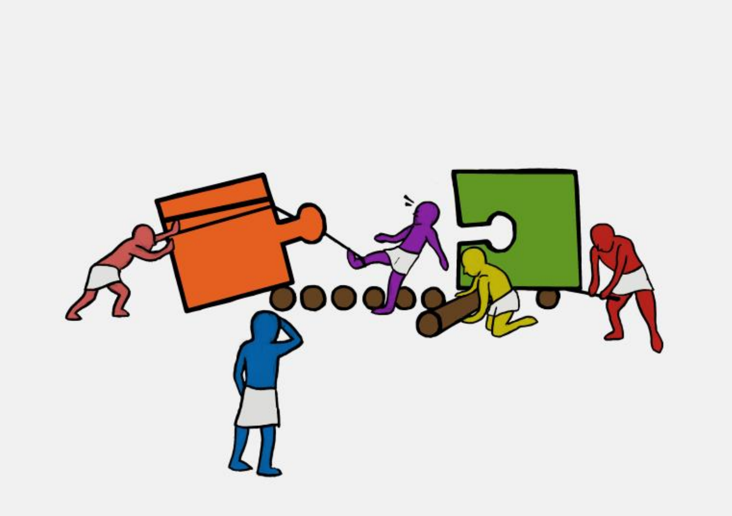
The standard errors give us a sense of our uncertainty over our estimates.

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.



If we assume normality, then:

$$CI_{\hat{\beta}}(95\%) = (\hat{\beta} - 2\sigma_{\hat{\beta}}, \hat{\beta} + 2\sigma_{\hat{\beta}})$$



Description

🧑‍🏫 Exercise: Beta Values for Data using Bootstrapping

Solve the previous exercise by building your own bootstrap function.

Instructions

- Define a function `bootstrap` that takes a dataframe as the input. Use NumPy's `random.randint()` function to generate random integers in the range of the length of the dataset. These integers will be used as the indices to access the rows of the dataset.
- Similar to the previous exercise, compute the β_0 and β_1 values for each instance of the dataframe.
- Plot the β_0, β_1 histograms.

Hints

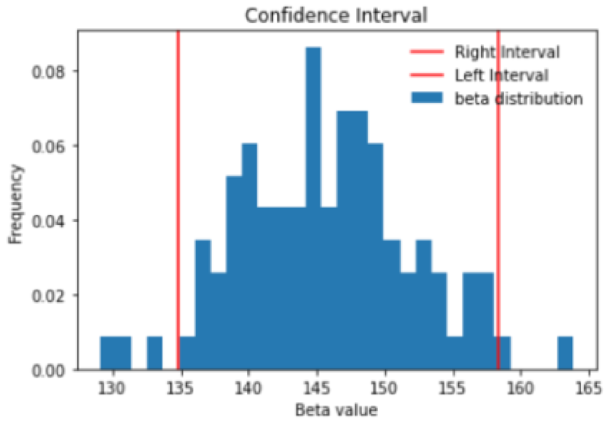
- To compute the beta values use the following equations:
  - $\beta_0 = \bar{y} - (b_1 * \bar{x})$
  - $\beta_1 = \frac{\sum(x-\bar{x})*(y-\bar{y})}{\sum(x-\bar{x})^2}$

where $\bar{x}$ is the mean of x and $\bar{y}$ is the mean of y

Description

🧑‍🏫 Exercise: Confidence Intervals for Beta value

The goal of this exercise is to create a plot like the one given below for β_0 and β_1.



Instructions:

- Follow the steps from the previous exercise to get the lists of beta values.
- Sort the list of beta values in ascending order (from low to high).
- To compute the 95% confidence interval, find the 2.5 percentile and the 97.5 percentile using `np.percentile()` .
- Use the helper code `plot_simulation()` to visualise the β values along with its confidence interval

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