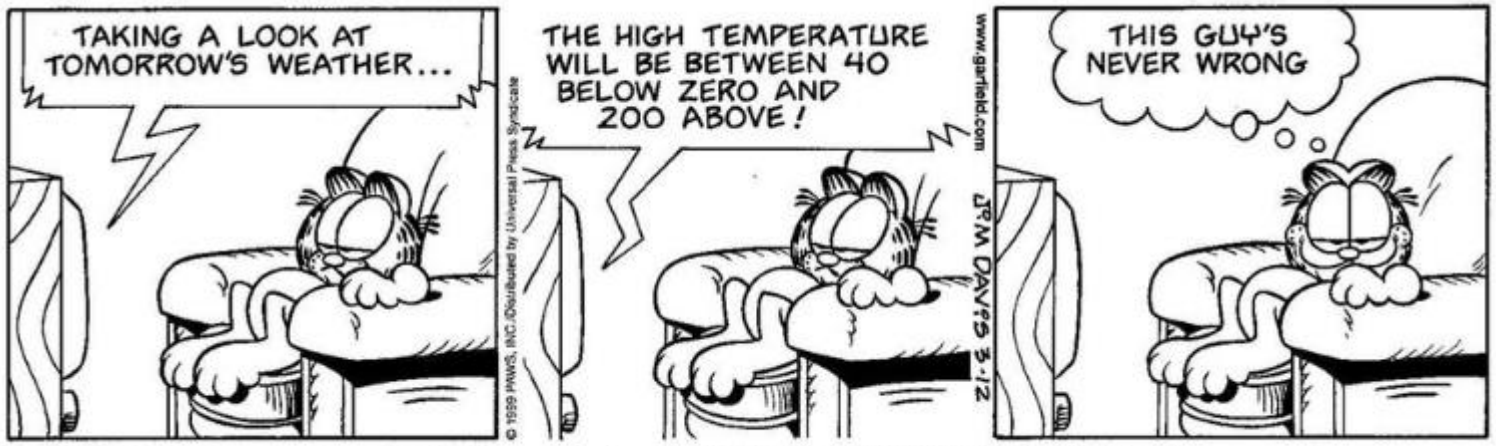
Bootstrap and confidence intervals

Introduction to data analysis: Lecture 7

Ori Plonsky

Spring 2023



RE: Percentiles

- The *p*th percentile is the value below which *p*% of the observations lie
  - The median is the 50th percentile
- Q1 = the first quartile = the 25th percentile
- Q3 = the third quartile = the 75th percentile
- IQR = Inter-quartile range = Q3-Q1

Percentiles

- The *p*th percentile is the value below which *p*% of the observations lie
  - The median is the 50th percentile
- Q1 = the first quartile = the 25th percentile
- Q3 = the third quartile = the 75th percentile
- IQR = Inter-quartile range = Q3-Q1

- For a percentile that does not exactly correspond to an observed element, there are many interpolation methods.
- We'll take the first element that is greater.
  - 100, 92, 97, 83, 67
    - Median? Q1? Q3?
- We'll use numpy percentile (with "method='higher'")

  Or pandas quantile (with "interpolation='higher'")

Estimation

- How big is an unknown parameter?
- If you have a census (that is, the whole population):
  - Just calculate the parameter and you're done
- If you don't have a census:
  - Take a random sample from the population
  - Use a statistic as an estimate of the parameter

Variability of an estimate

- One sample ➜ One estimate
- But the random sample could have come out differently
- And so the estimate could have been (very?) different

Variability of an estimate

- One sample ➜ One estimate
- But the random sample could have come out differently
- And so the estimate could have been (very?) different
- Main question:

How different could the estimate have been?

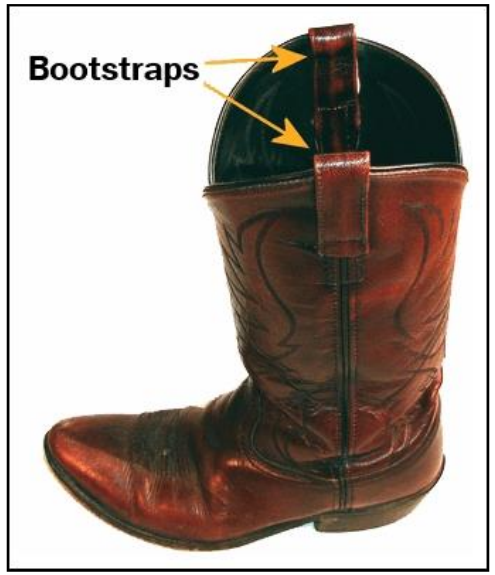
- The variability of the estimate tells us something about how accurate the estimate is:
  estimate = parameter + error

Another sample?

- One sample ➜ One estimate
- To get many values of the estimate, we need many random samples
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?

Another sample?

- One sample ➜ One estimate
- To get many values of the estimate, we need many random samples
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?



The Bootstrap

- A technique for **simulating** repeated random sampling
- All that we have is the original sample
  - … which is large and random
  - Therefore, it probably resembles the population
    - According to?
  - That is, the empirical distribution of the sample looks like the probability distribution of the population (hopefully)

The Bootstrap

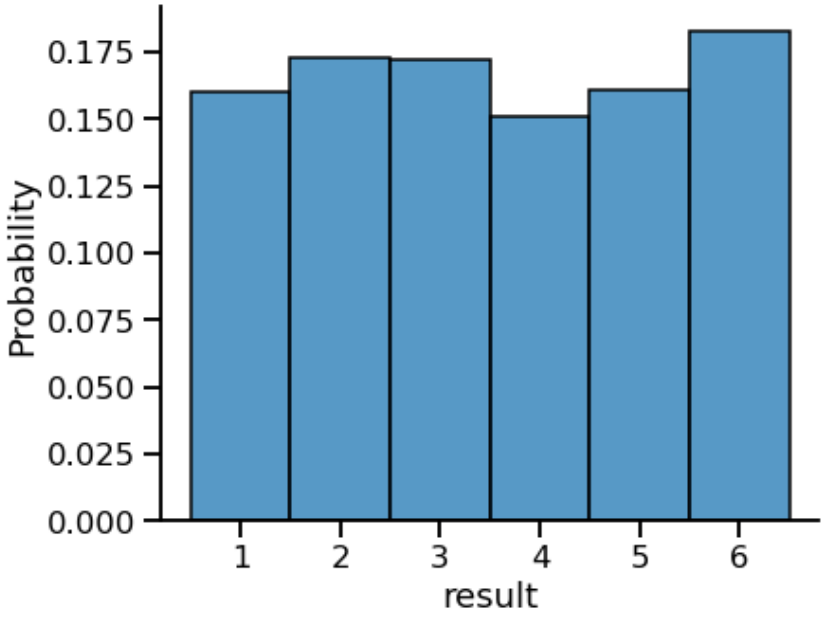
- A technique for **simulating** repeated random sampling
- All that we have is the original sample
  - … which is large and random
  - Therefore, it probably resembles the population
    - According to?
  - That is, the empirical distribution of the sample looks like the probability distribution of the population (hopefully)
- So we sample at random from the original sample!
  - Behave as if they are the same

Intuition

- A couple of lectures ago, we simulated rolls of a die

```
# empirical distribution of roll of a die
def empirical_hist_die(num_rolls):
    die_df = pd.DataFrame({'result': np.random.choice(die, num_rolls)})
    sns.displot(die_df, x='result', bins=die_bins, stat='probability', height=5, aspect=1.3)
    return die_df

die_simulation_results = empirical_hist_die(1000)
```



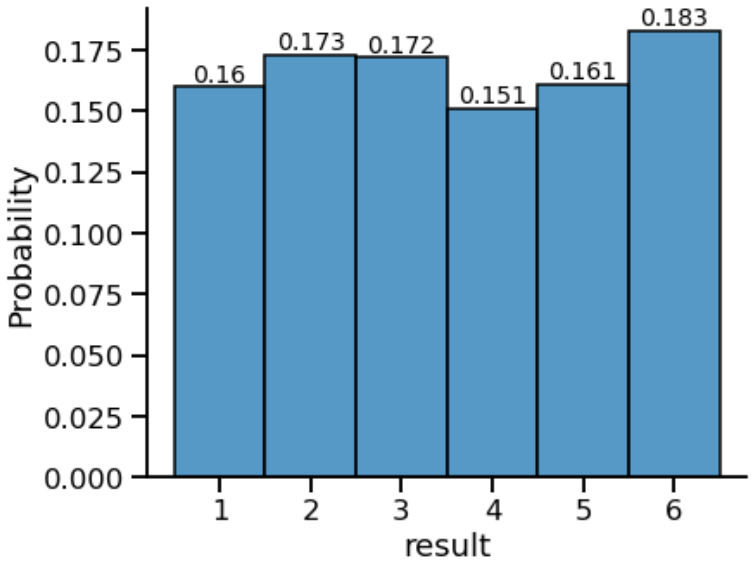
Intuition

- The empirical distribution of the sample of 1000 die rolls we got is not uniform, but it is not very far from uniform

```
die_probs = []
num_rolls = die_simulation_results.shape[0]
for i in range(1,7):
    die_probs.append(np.count_nonzero(die_simulation_results == i)/num_rolls)

die_probs
```

```
[0.16, 0.173, 0.172, 0.151, 0.161, 0.183]
```



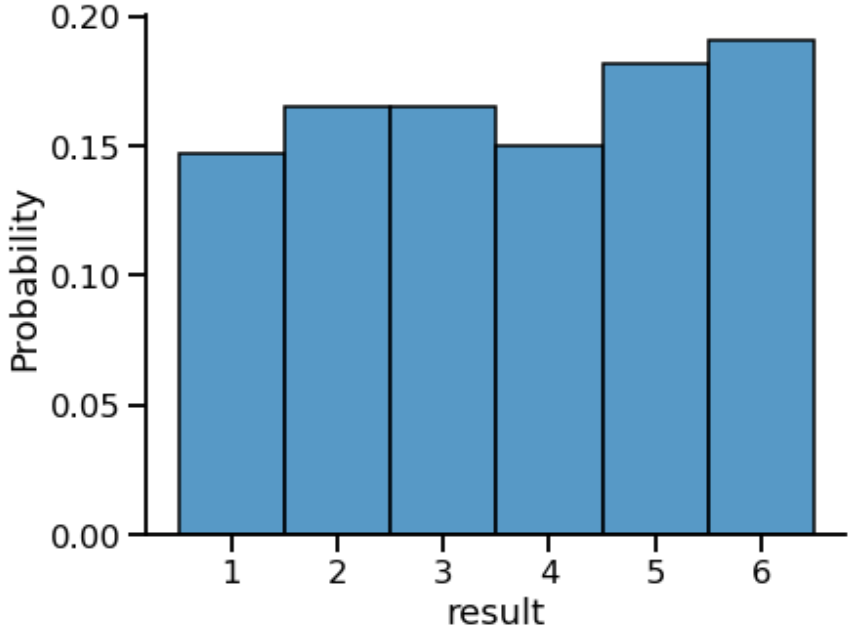
- We can use this empirical distribution (which is close to the true distribution) to generate additional simulated samples

Intuition

- We can use this empirical distribution (which is close to the true distribution) to generate additional simulated samples

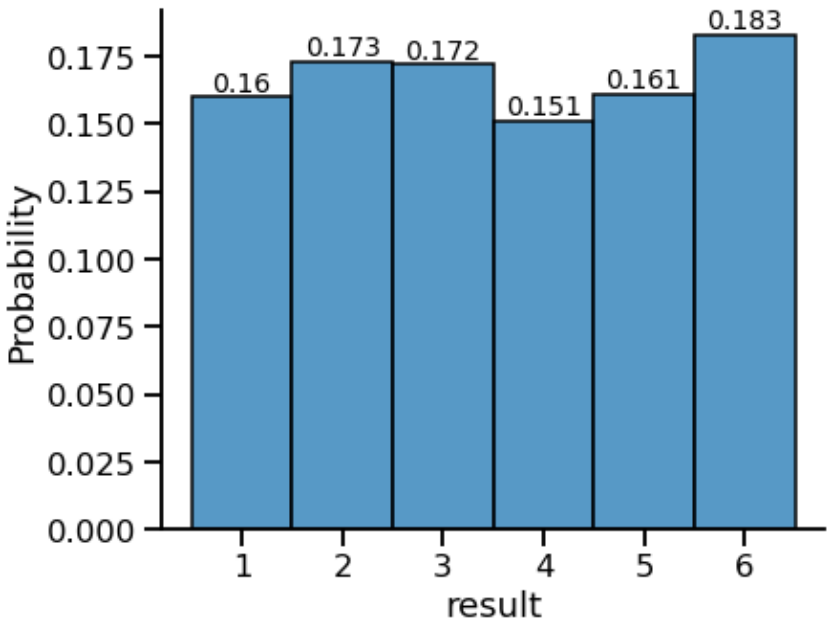
```
# Bootstrap distribution of roll of a die
def empirical_hist_die(num_rolls):
    die_df = pd.DataFrame({'result': np.random.choice(die, p=die_probs, size=num_rolls)})
    sns.displot(die_df, x='result', bins=die_bins, stat='probability', height=5, aspect=1.3)
    return die_df

bootstrap_die_simulation_results = empirical_hist_die(1000)
```

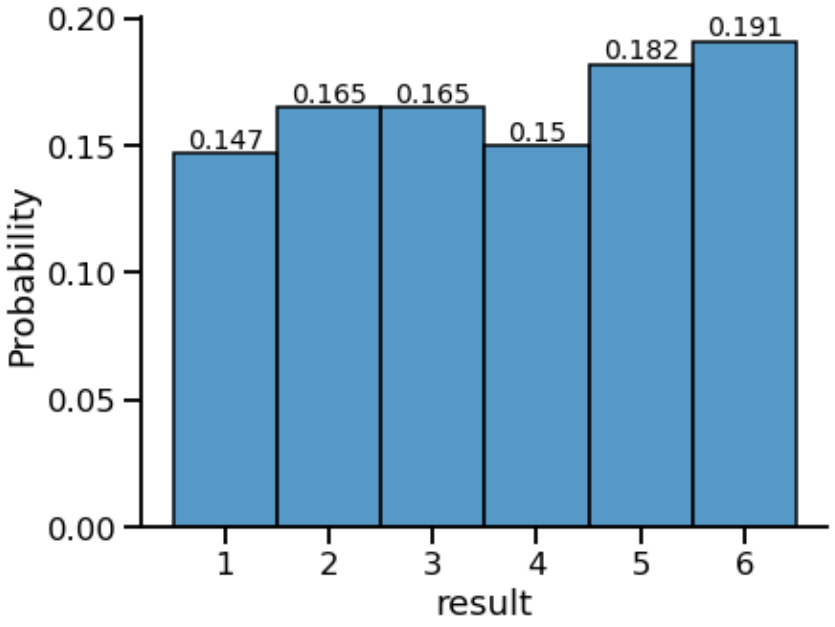


Intuition

The distribution of the *bootstrapped* sample we got is similar to the distribution of the sample we got when using the true distribution of the die



1000 simulations of a die roll using its true probability distribution



1000 simulations of a die roll using the empirical distribution of our initial sample

Key to resampling

From the original sample

- Draw at random
- With replacement
- As many value as the original sample contained

Key to resampling

From the original sample

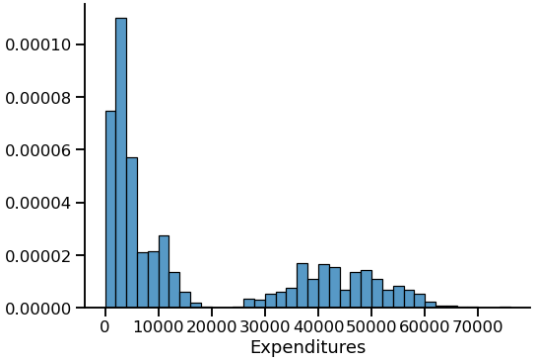
- Draw at random
- With replacement
- As many value as the original sample contained

- The size of the new sample has to be the same as the original so that the two estimates we get are comparable
  - The variability of the estimate is linked to the sample size

[notebook]

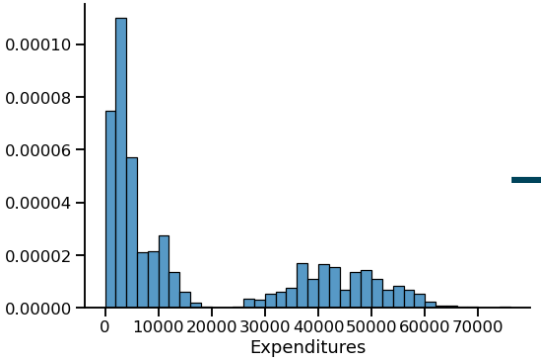
Why the bootstrap works

population

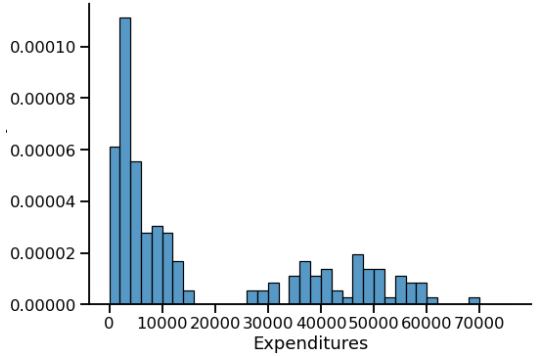


Why the bootstrap works

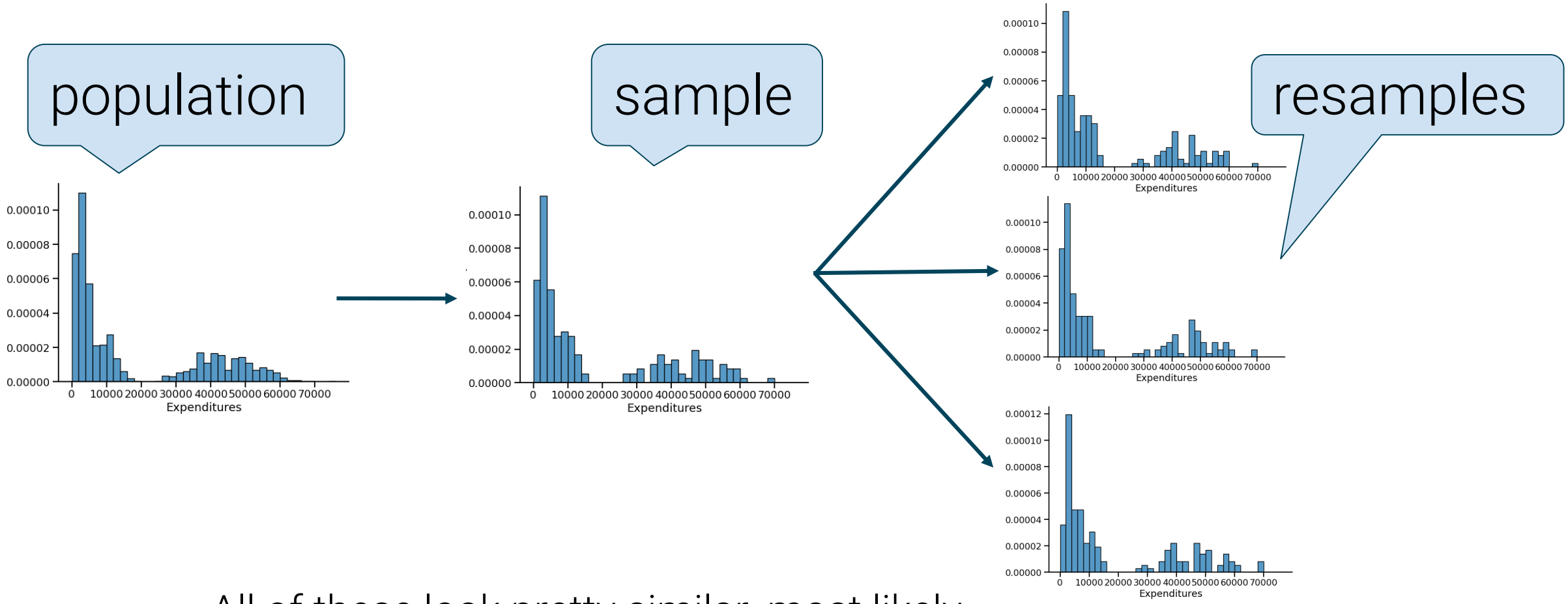
population



sample



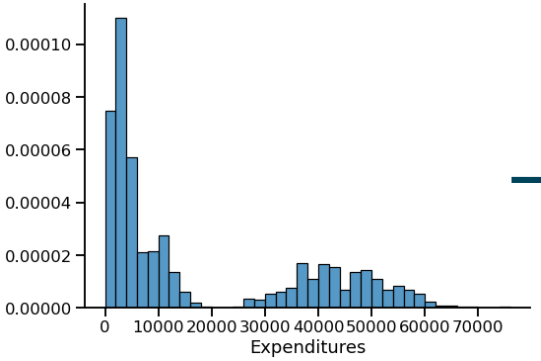
Why the bootstrap works



All of these look pretty similar, most likely.

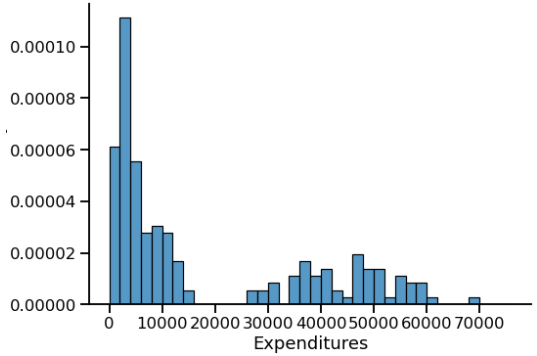
Why we need the bootstrap

population



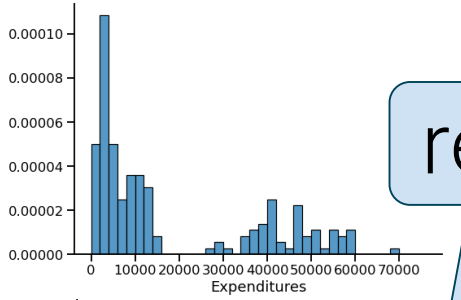
What we wish we could get

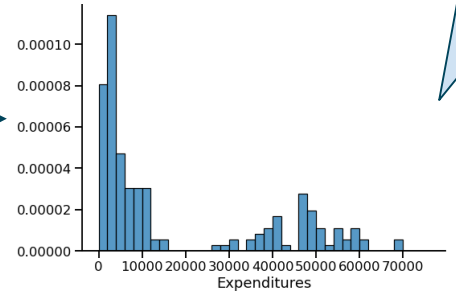
sample

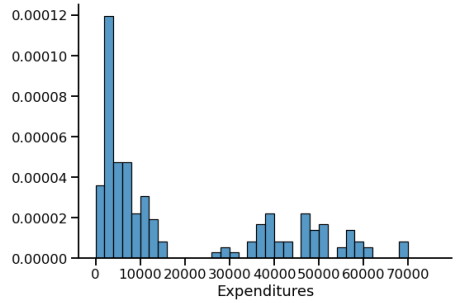


What we really get

resamples







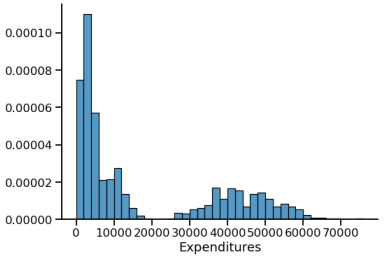
Estimating parameters using bootstrap

- We want to know how variable our estimate is, but we only have one sample!

Estimating parameters using bootstrap

- We want to know how variable our estimate is, but we only have one sample!

The truth
(how the world
behaves)



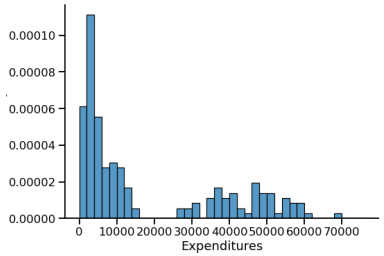
Estimating parameters using bootstrap

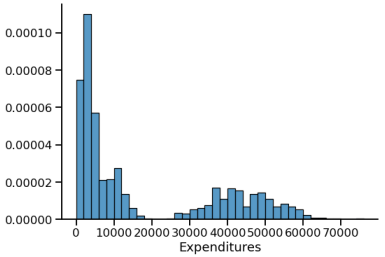
- We want to know how variable our estimate is, but we only have one sample!

The truth
(how the world
behaves)

Observations

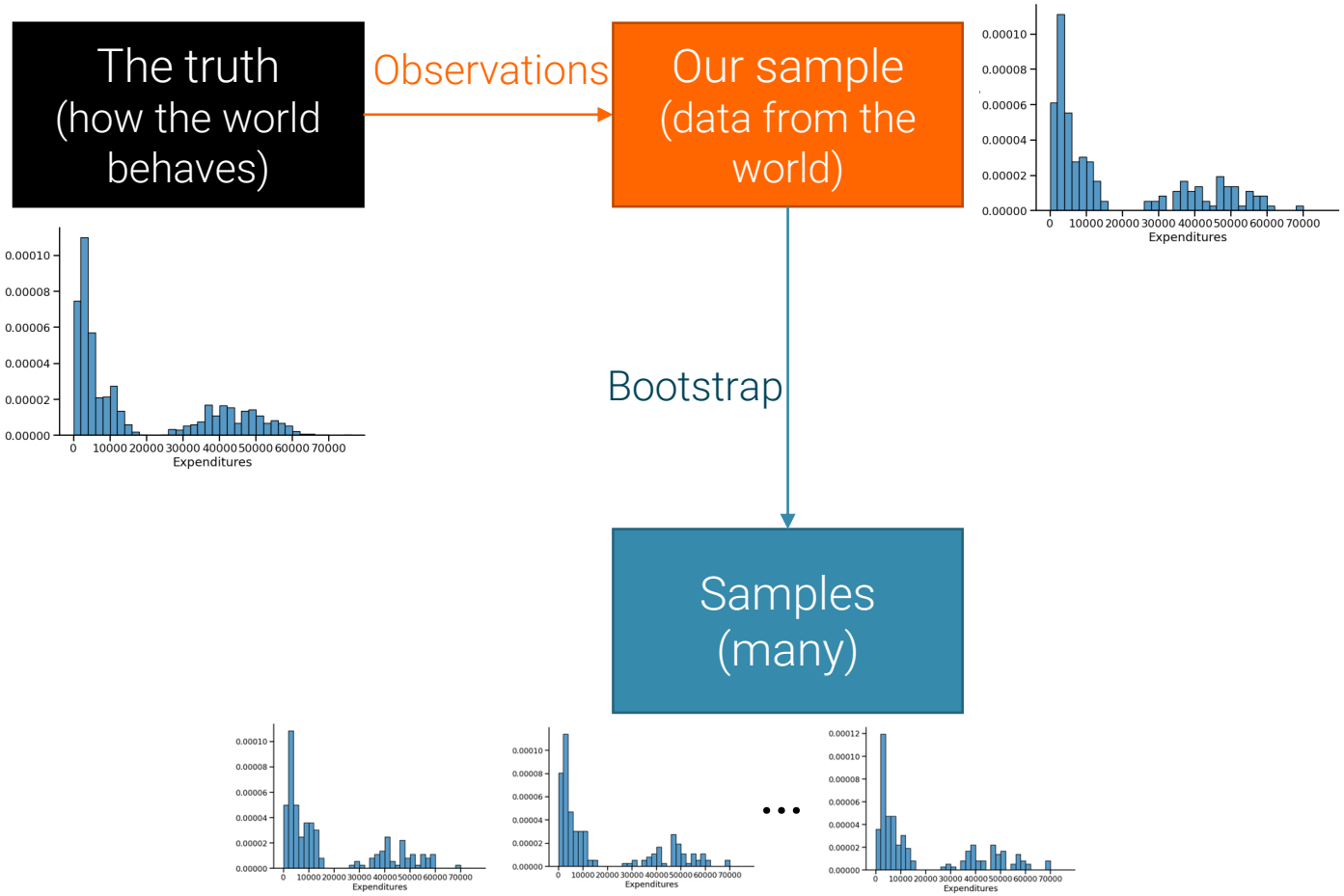
Our sample
(data from the
world)





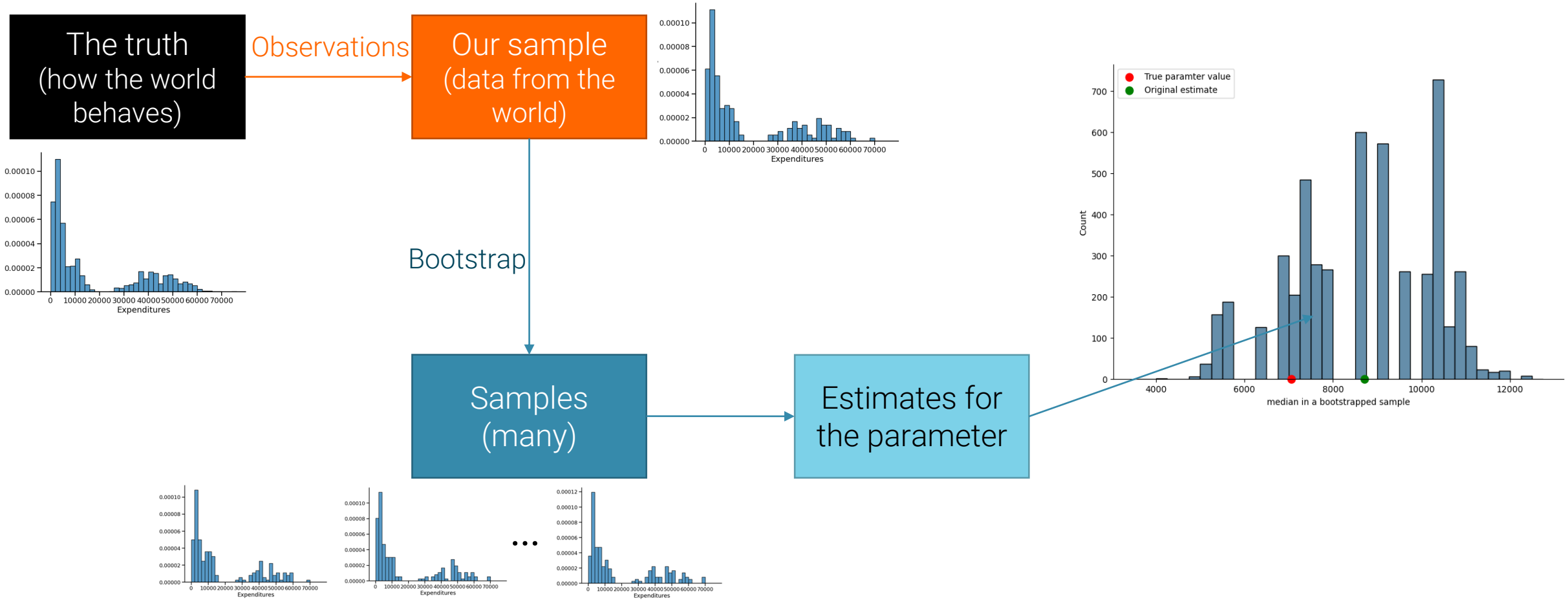
Estimating parameters using bootstrap

- We want to know how variable our estimate is, but we only have one sample!



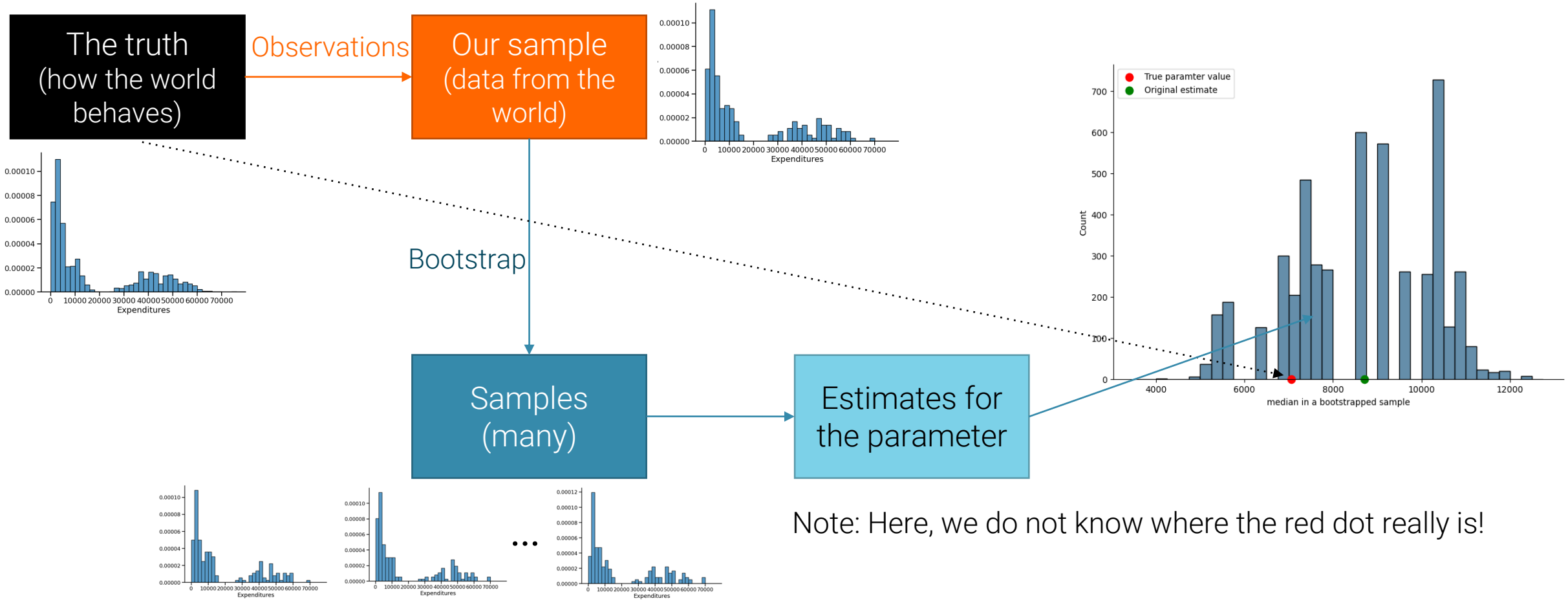
Estimating parameters using bootstrap

- We want to know how variable our estimate is, but we only have one sample!



Estimating parameters using bootstrap

- We want to know how variable our estimate is, but we only have one sample!



Note: Here, we do not know where the red dot really is!

Confidence Interval (רווח סמך)

- So, we got many possible estimates for the parameter
  - One for each bootstrapped sample
  - The distribution we get tells us how our estimate may change due to variability in data
- Some of these are more likely than others
  - Values that are in the tail of the distribution of estimates are less likely
- It is very unlikely that the one estimate we got from the original sample equals the true parameter exactly, but it **is likely** that true parameter lies within a range where we see most of the estimates we got

Confidence Interval (רווח סמך)

- So, we got many possible estimates for the parameter
  - One for each bootstrapped sample
  - The distribution we get tells us how our estimate may change due to variability in data
- Some of these are more likely than others
  - Values that are in the tail of the distribution of estimates are less likely
- It is very unlikely that the one estimate we got from the original sample equals the true parameter exactly, but it **is likely** that true parameter lies within a range where we see most of the estimates we got
- Confidence interval = a range of values within which we expect the true parameter lies
  - In a certain percentage of the cases
  - Based on random sampling (and the distribution of bootstrapped estimates)

Confidence level

- Often, we look for “95% confidence intervals”
- 95% is called the **confidence level**
  - Could be any percent between 0 and 100
  - Higher level means wider intervals

Confidence level

- Often, we look for “95% confidence intervals”
- 95% is called the **confidence level**
  - Could be any percent between 0 and 100
  - Higher level means wider intervals
- The confidence is in the process that generated the interval:
  - It generates a “good” interval about 95% of the time.

Confidence level

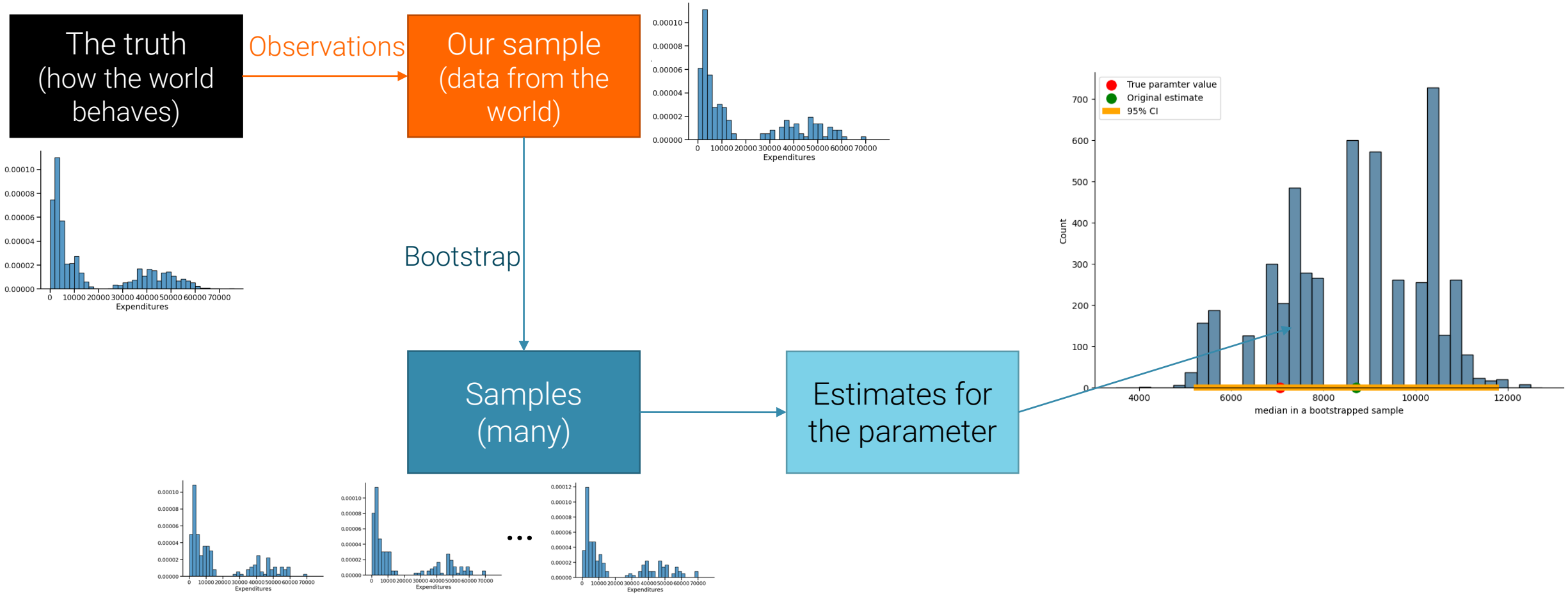
- Often, we look for “95% confidence intervals”
- 95% is called the **confidence level**
  - Could be any percent between 0 and 100
  - Higher level means wider intervals
- The confidence is in the process that generated the interval:
  - It generates a “good” interval about 95% of the time.

- *Default seaborn error bars (e.g. in bar plots) show 95% confidence interval based on bootstrap with 1000 replications*

(notebook)

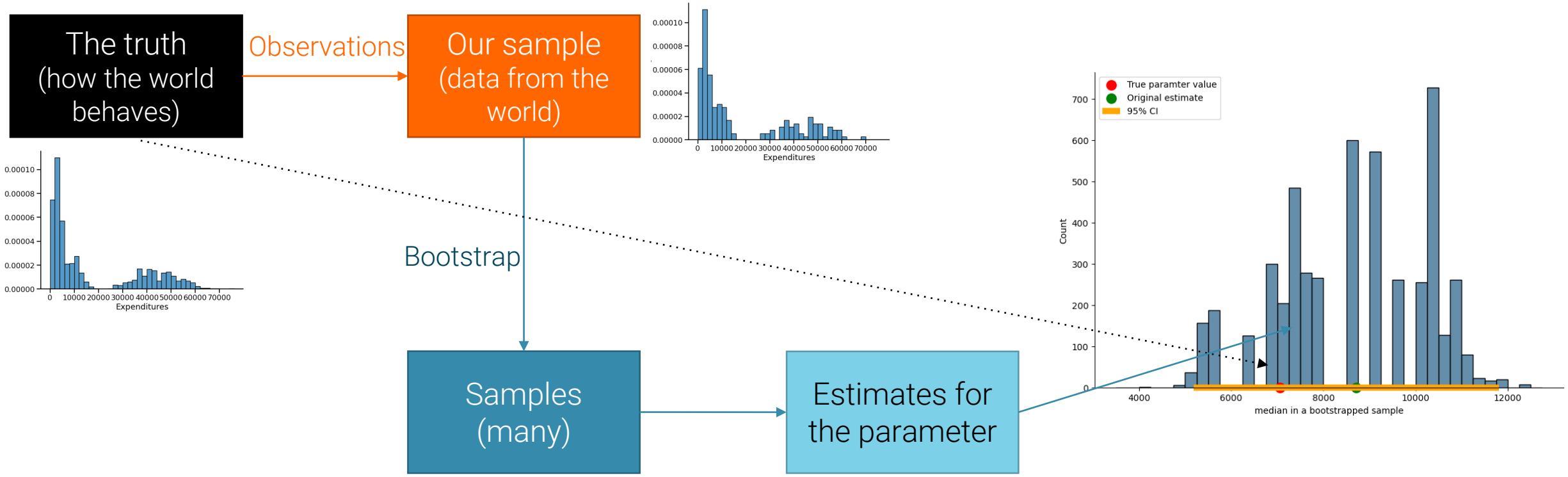
Estimating parameters using bootstrap

- We want to know how variable our estimate is, but we only have one sample!



Estimating parameters using bootstrap

- We want to know how variable our estimate is, but we only have one sample!



Note: Here, we will not know where the red dot really is! But if our sample is large and random, the 95% confidence interval (orange line) would include it with high likelihood

True or False?

“An approximate 95% confidence interval for the average gross income of movies in the population is (48,966,237, 53,282,655) dollars.”

True or False:

True or False?

“An approximate 95% confidence interval for the average gross income of movies in the population is (48,966,237, 53,282,655) dollars.”

True or False:

1. About 95% of the movies in the population had gross income between $48,966,237 and $53,282,655.

True or False?

“An approximate 95% confidence interval for the average gross income of movies in the population is (48,966,237, 53,282,655) dollars.”

True or False:

1. About 95% of the movies in the population had gross income between $48,966,237 and $53,282,655.

2. There is a 0.95 probability that the average gross income of movies in the population is in the range (48,966,237, 53,282,655).

True or False?

"An approximate 95% confidence interval for the average gross income of movies in the population is (48,966,237, 53,282,655) dollars."

True or False:

1. About 95% of the movies in the population had gross income between $48,966,237 and $53,282,655.

2. There is a 0.95 probability that the average gross income of movies in the population is in the range (48,966,237, 53,282,655).

3. If we create 1000 confidence intervals using the bootstrap, we expect around 950 of them to include the true value of the average gross income of movies in the population

Illustrating CI

When *Not* to use the Bootstrap

- If you're trying to estimate very high or very low percentiles, like min or max
- If you're trying to estimate any parameter that's greatly affected by rare elements of the population
- If observations are dependent on one another (time series, spatial data)
- If the original sample is too small to reliably represent the probability distribution in the population
- In general, if your sample is biased in some way (not random), then bootstrap will not help you

Using CI for hypothesis testing

- Our approximate 95% confidence interval for the average gross income of movies in the population is (48,966,237, 53,282,655) dollars
- This estimation process is good in 95% of the time (assuming our sample is large and random)

Using CI for hypothesis testing

- Our approximate 95% confidence interval for the average gross income of movies in the population is (48,966,237, 53,282,655) dollars
- This estimation process is good in 95% of the time (assuming our sample is large and random)
- Suppose someone wants to test the following hypotheses:

  H_0 (null hypothesis): The mean gross income of movies is $55M

  H_1 (alternative hypothesis): The mean gross income of movies is **not** $55M

Using CI for hypothesis testing

- Our approximate 95% confidence interval for the average gross income of movies in the population is (48,966,237, 53,282,655) dollars
- This estimation process is good in 95% of the time (assuming our sample is large and random)
- Suppose someone wants to test the following hypotheses:

  H_0 (null hypothesis): The mean gross income of movies is $55M

  H_1 (alternative hypothesis): The mean gross income of movies is **not** $55M
- If our significance level is 0.05, we will reject the null hypothesis!
  - We are 95% confident that (48,966,237, 53,282,655) contains the true mean gross income of movies.
  - If the true mean income were 55M, it is highly unlikely we would get a sample that leads us to a CI of (48,966,237, 53,282,655)
  - In particular, we are (at least) 95% confident that $55M is not the true mean gross income of movies → We reject H_0 at the 5% level

CI for testing population mean

- Null hypothesis: Population average = x
- Alternative hypothesis: Population average ≠ x
- Cutoff for P-value: p%
- Method:
  - Construct a (100-p)% confidence interval for the population average
  - If x is not in the interval, reject the null
  - If x is in the interval, can't reject the null

CI for testing difference in population means

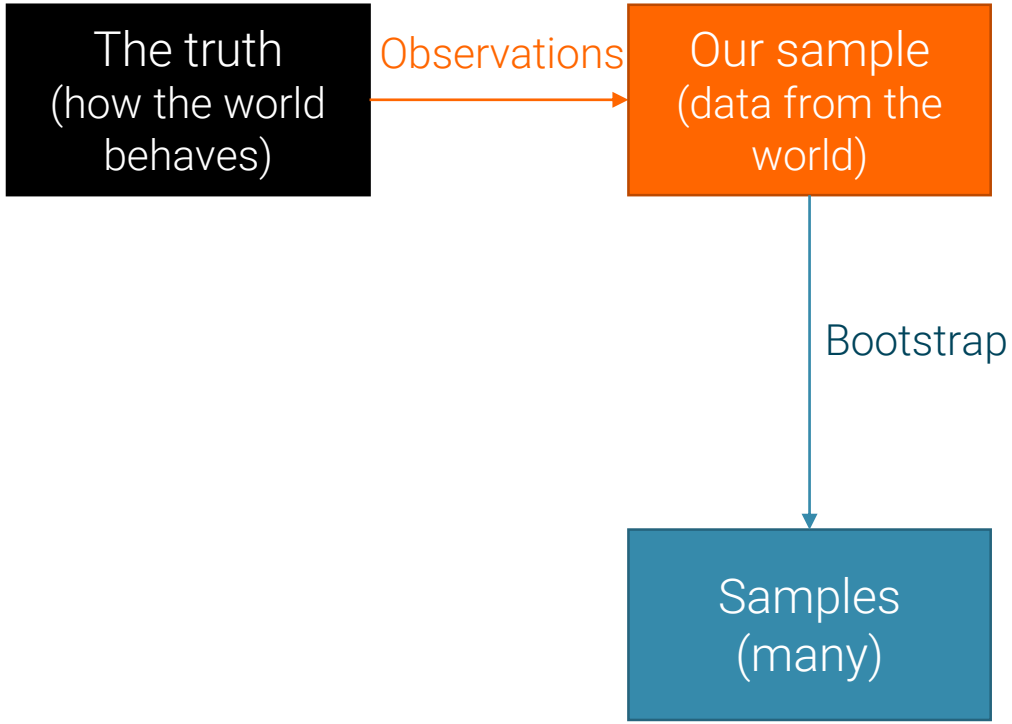
- Null hypothesis: Difference between population averages = *0*
- Alternative hypothesis: Difference between population averages ≠ *0*
- Cutoff for P-value: *p*%
- Method:
  - Construct a (100-*p*)% confidence interval for the difference between population averages
  - If 0 is not in the interval, reject the null
  - If 0 is in the interval, can't reject the null

Hypothesis testing using bootstrap

We want to know something about a parameter, but we only have a sample!

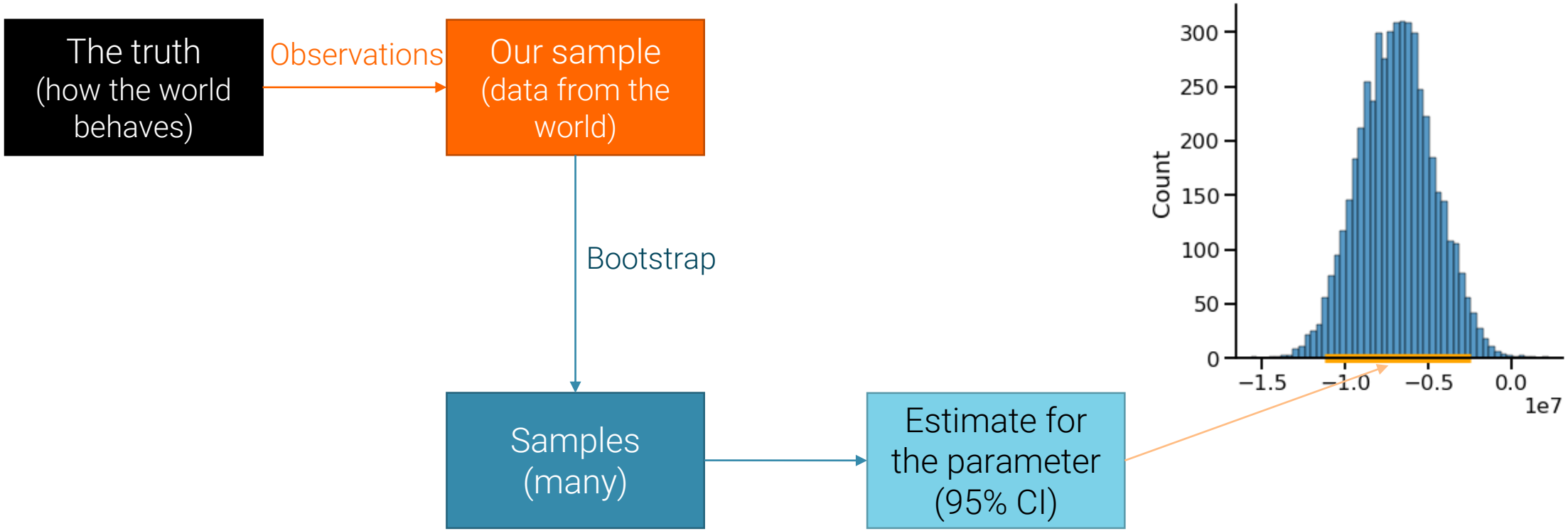
Hypothesis testing using bootstrap

We want to know something about a parameter, but we only have a sample!



Hypothesis testing using bootstrap

We want to know something about a parameter, but we only have a sample!



Hypothesis testing using bootstrap

We want to know something about a parameter, but we only have a sample!