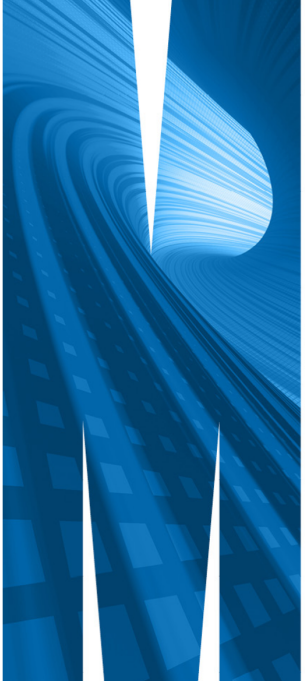
MONASH University

Bootstrap



Statistical Thinking (ETC2420 / ETC5242)

Week 5, Semester 2, 2025

Outline

Outline

Learning goals for Week 5

- Understand the basic idea of bootstrapping for assessing variability of statistics.
- Understand the percentile bootstrap technique.
- Calculate bootstrap confidence intervals for parameters.

(1-α)% confidence intervals (CIs) in general

- Common general format for CIs:

$$\text{estimate} \pm \text{quantile} \times \text{se(estimate)}$$

- If we know these quantities, we can calculate a CI
- In Week 3 (Estimation) we relied on the Central Limit Theorem (CLT) or on knowing the sampling distribution
- Today we learn about an alternative
- Useful when we can't use the CLT, e.g. using the median as our estimator.

[Reminder] “Confidence” vs probability

- What does the “(1-α)%” refer to?
- Probability of the CI (the interval **estimator**) covering the true value of the parameter under hypothetical repeated sampling.
- Only relevant **before** the data are observed.
- Once the data are observed, the CI (the interval **estimate**) is fixed (not random).
- The parameter is always fixed.
- The observed interval estimate gives a *plausible* set of values for the parameter, in light of the observed data.
- Higher confidence level $\Rightarrow$ greater plausibility.

Confidence intervals via a “bootstrap” approach

- **Bootstrap** techniques provide alternative approaches to constructing a confidence interval.
- Replace mathematical derivations by computational techniques.
- (Replace ‘brain work’ with brute-force…)
- There are several bootstrap approaches. Today we cover only the **percentile bootstrap**.

What is a “bootstrap”?

What is a “bootstrap”?

- “To pull oneself up by one’s bootstraps”
- **Bootstrapping** is a self-starting process that is supposed to proceed without external input.

Outline

Percentile bootstrap: preview of key ideas

- If we knew the population distribution, we could calculate the sampling distribution of our estimator.
- Use the sample to approximate the population.
- Use the **empirical cdf** to approximate the **population cdf**.
- Simulate approximate hypothetical datasets (*bootstrap samples*), by **re-sampling from the observed values**.
- This re-sampling is done **with replacement**, creating simulated datasets with the **same sample size**.
- Do this a large number of times, to approximate the sampling distribution of our estimator.

Bootstrap samples

- If we have a large enough sample, we can say that our sample represents a ‘pseudo-population’.
- We cannot take repeated samples from the population (it is unknown).
- But we can take repeated samples from our ‘pseudo-population’!
- Sampling with replacement from the sample (‘pseudo-population’) approximates sampling from the population. This is called a **bootstrap sample**.

Using the bootstrap samples

- From each bootstrap sample, we calculate our estimator.
- A large number of bootstrap samples gives us a large number of simulated values of our estimator.
- This set of values gives us an approximation to the sampling distribution of the estimator.
- Can use this to assess the variability of the estimate and calculate a confidence interval.

Percentile bootstrap CI for single population mean, μ

An approximate 95% CI can be obtained in two steps:

1. Generate B bootstrap samples and calculate $\hat{\mu}$ for each one.
   - Denote these as $\{\hat{\mu}^{[1]}, \hat{\mu}^{[2]}, \ldots, \hat{\mu}^{[B]}\}$
   - B should be a large number, e.g. $B = 1000$
2. Calculate the 2.5% and 97.5% sample percentiles of the B values of $\hat{\mu}$.
   - This gives an approximate 95% CI for μ.

Notes:

- The bootstrap samples provide an approximation of the sampling distribution of $\hat{\mu}$, by giving the following approximate (empirical) cdf:

$$\hat{\Pr}(\hat{\mu} \leqslant c) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}[\hat{\mu}^{[b]} \leqslant c].$$

- We want to select an interval from this cdf that has (empirical) probability of 95%.
  - The 2.5% and 97.5% sample percentiles provide such an interval.
- This gives an approximate 95% CI.

How to generate bootstrap samples

How to calculate $\hat{\mu}^{[b]}$?

- Suppose we have a sample of size n
- For each b in $\{1, 2, \ldots, B\}$
  - Resample n values from the sample, with replacement
  - Label these values as $\{x_1^{[b]}, x_2^{[b]}, \ldots, x_n^{[b]}\}$
  - Compute $\hat{\mu}^{[b]}$ from the bootstrap sample.

For example, if your estimator is the sample mean, $\hat{\mu} = \bar{X}$, then for bootstrap sample b you would calculate:

$$\hat{\mu}^{[b]} = \bar{x}^{[b]} = \frac{1}{n} \sum_{i=1}^{n} x_i^{[b]}$$

How to generate bootstrap samples in R

Can use the following functions in R:

- `sample(..., replace = TRUE)`
- `slice_sample(..., replace = TRUE)`

Resampling in R with `sample()` – permutation

```
x <- 1:10
x
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```
mean(x)
```

```
[1] 5.5
```

```
x1 <- sample(x, replace = FALSE)
x1
```

```
 [1]  5  8  7  9  4  2 10  6  3  1
```

```
mean(x1)
```

```
[1] 5.5
```

Resampling in R with `sample()`

```
x <- 1:10
x
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```
mean(x)
```

```
[1] 5.5
```

```
x1 <- sample(x, replace = TRUE)
x1
```

```
 [1]  5  8  7  7  2 10  8 10 10  5
```

```
mean(x1)
```

```
[1] 7.2
```

Resampling in R with `slice_sample()`

```
df <- tibble(a = 1:10, b = letters[1:10])
glimpse(df)
```

```
Rows: 10
Columns: 2
$ a <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
$ b <chr> "a", "b", "c", "d", "e", "f", "g", "h", "i", "j"
```

```
df2 <- slice_sample(df, n = nrow(df), replace = TRUE)
glimpse(df2)
```

```
Rows: 10
Columns: 2
$ a <int> 5, 8, 7, 7, 2, 10, 8, 10, 10, 5
$ b <chr> "e", "h", "g", "g", "b", "j", "h", "j", "j", "e"
```

Resampling in R with `slice_sample()`

```
mean(df$a)
```

```
[1] 5.5
```

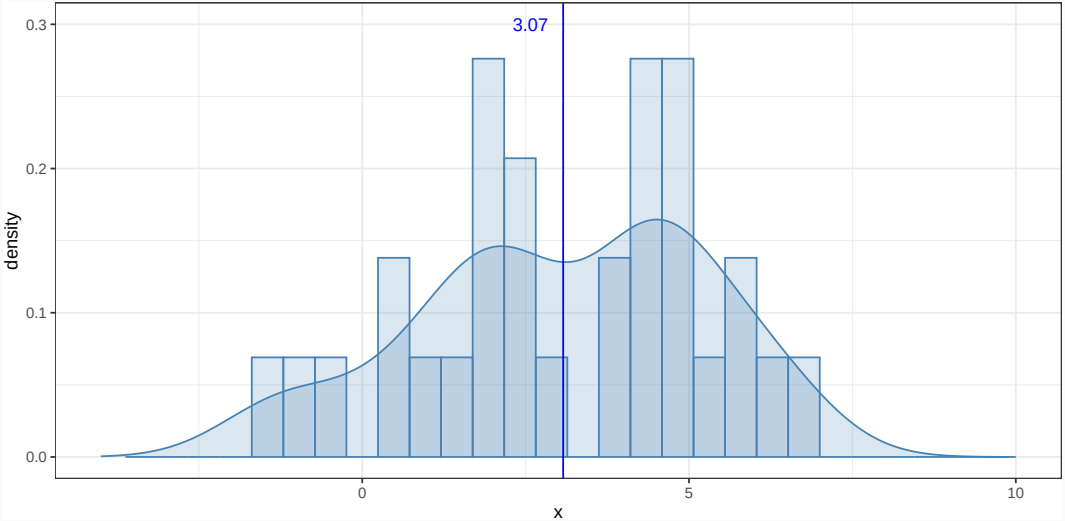
```
mean(df2$a)
```

```
[1] 7.2
```

Sample of size $n = 30$ from $\mathrm{N}(\mu = 3, 4)$

- Want to estimate μ using the sample mean, $\hat{\mu} = \bar{X}$.
- This is a theoretical example where we are drawing from a known population.
- We actually know the population mean (but we pretend that we don't).
- This is just to illustrate how a bootstrap works, rather than being a realistic example of where the bootstrap is necessary.

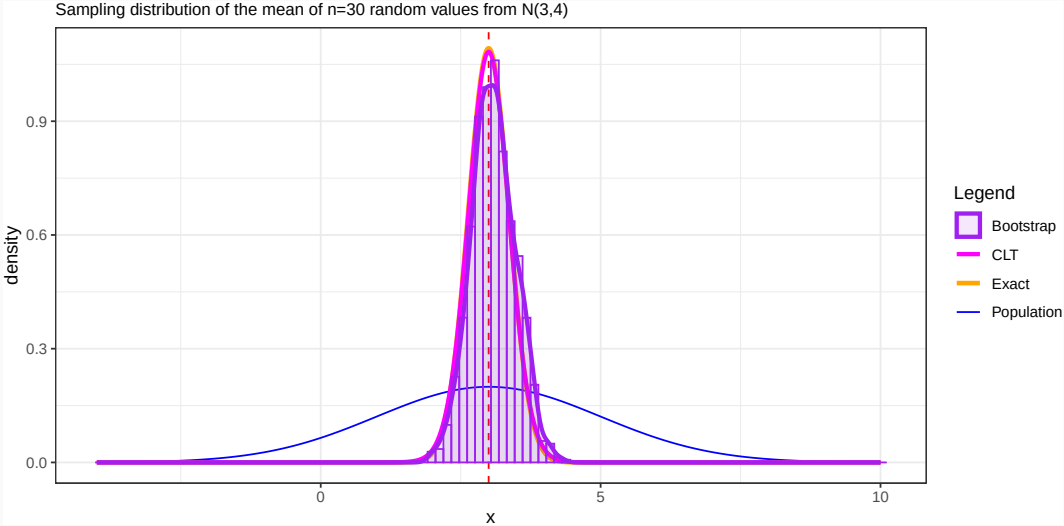
Our sample of size 30 looks like:



Now bootstrap

- Starting with our sample of size 30
- From it we sample **with replacement** a new sample of size 30
- Calculate our sample statistic (the sample mean) and store it
- Repeat many times; here we use $B = 1000$, but could have used many more
- Plot the distribution of the 1000 simulated means (this is an approximate sampling distribution of $\bar{x}$)
- Find the middle 95% of the simulated means
- This defines our CI.

What does it look like?



95% confidence intervals

- Take 2.5% off from each tail of the bootstrap empirical distribution
- Just sort the $\{\bar{x}_{obs}^{[b]}\}$ values and find
  - The 2.5% percentile $\Rightarrow L_{\text{Boot}}$
  - The 97.5% percentile $\Rightarrow U_{\text{Boot}}$
- Then $[L_{\text{Boot}},\ U_{\text{Boot}}]$ is an approximate 95% confidence interval for μ
- For $\mathrm{N}(3, 4)$ example: 95% CIs (approximate and exact)
  - $[L_{\text{Boot}},\ U_{\text{Boot}}] = [2.34,\ 3.83]$
  - $[L_{\text{CLT}},\ U_{\text{CLT}}] = [2.25,\ 3.90]$
  - $[L_{\text{Exact}},\ U_{\text{Exact}}] = [2.28,\ 3.72]$
- Exact available since $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(3, 4)$, therefore $\bar{X} \sim \mathrm{N}\left(3, \frac{4}{30}\right)$.

Outline

Bootstrap for paired samples

- Like with the CLT, we can apply the bootstrap to paired data

$$(X_{1,i}, X_{2,i}), \quad \text{for } i = 1, 2, \ldots, n.$$

- First calculate the sample of paired differences:

$$DD_n = \{Diff_i = X_{1,i} - X_{2,i}, \text{ for } i = 1, 2, \ldots, n\}$$

- Then apply the **percentile bootstrap** method (for a single population) to the DD_n sample.

Bootstrap for paired samples

- For each b in $\{1, 2, \ldots, B\}$:
  - Resample n values from the DD_n set, with replacement
  - Compute the average $\bar{Diff}^{[b]}$
- Use the empirical sample of $\{\bar{Diff}^{[b]}, \text{ for } b = 1, 2, \ldots, B\}$ to obtain a confidence interval for $\mu_{Diff} = \mu_1 - \mu_2$.
- (We will do this in the tutorial)

Bootstrap for the difference in two independent samples

- For unpaired data $D1_{n_1} = \{X_{1,i}, \text{ for } i = 1, 2, ..., n_1\}$ and $D2_{n_2} = \{X_{2,j}, \text{ for } j = 1, 2, ..., n_2\}$, we can use the bootstrap to build the relevant confidence interval.
- For each b,
  - resample with replacement n_1 observations from $D1_{n_1}$ to produce $\bar{x}^{[b]}_{1,obs}$,
  - resample with replacement n_2 observations from $D2n_2$ to produce $\bar{x}^{[b]}_{2,obs}$, and
  - calculate $(\bar{x}^{[b]}_{1,obs} - \bar{x}^{[b]}_{2,obs})$
- And compute an approximate 95% confidence interval using the lower 2.5% and 97.5% percentiles of $\{(\bar{x}^{[b]}_{1,obs} - \bar{x}^{[b]}_{2,obs}), \text{ for } b = 1, 2, \ldots, B\}$.

CLT and bootstrap CIs are similar...

- Both the CLT and bootstrap approaches **do not** require knowledge of the true underlying population distribution.
- Both the CLT and bootstrap approaches are straightforward to implement.

What's the advantage of the bootstrap approach?

- The CLT only works for the sampling distribution of $\bar{X}$ (sample mean)
- The bootstrap can be applied to (almost) **any point estimator**
- For example:
  - single population median
  - trimmed/robust mean
  - parameters of assumed models
  - to assess single population asymmetry: e.g. mean $-$ median
  - differences in the medians of two independent samples
  - etc.
- The CLT CI is symmetric by construction, but many bootstrap approaches can incorporate **skewness** in the data.

Example: Gamma distribution

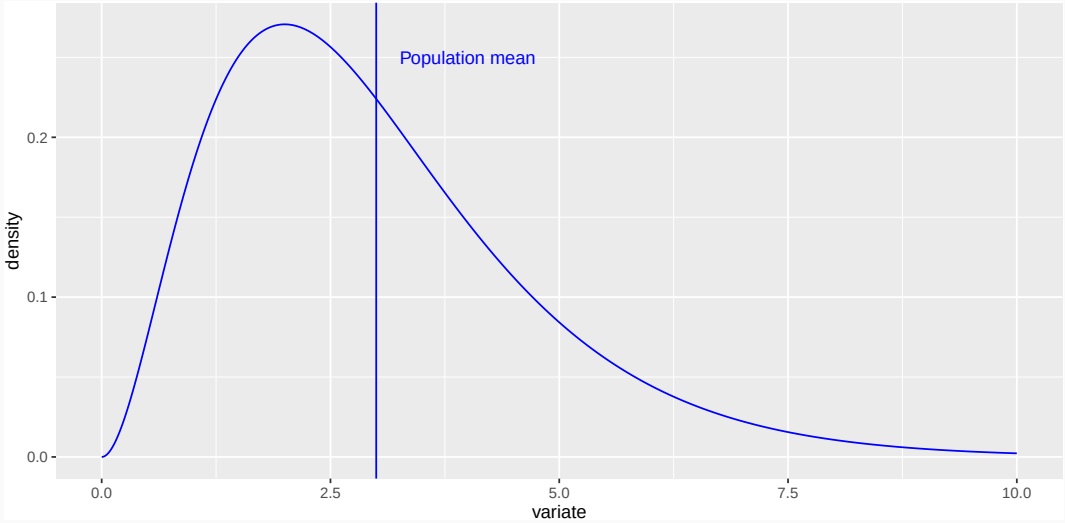
- Population: $\mathrm{Gamma}(\text{shape} = 3, \text{rate} = 1)$
- Parameter of interest (assumed unknown): population mean,

$$\mu = \frac{\text{shape}}{\text{rate}} = 3$$

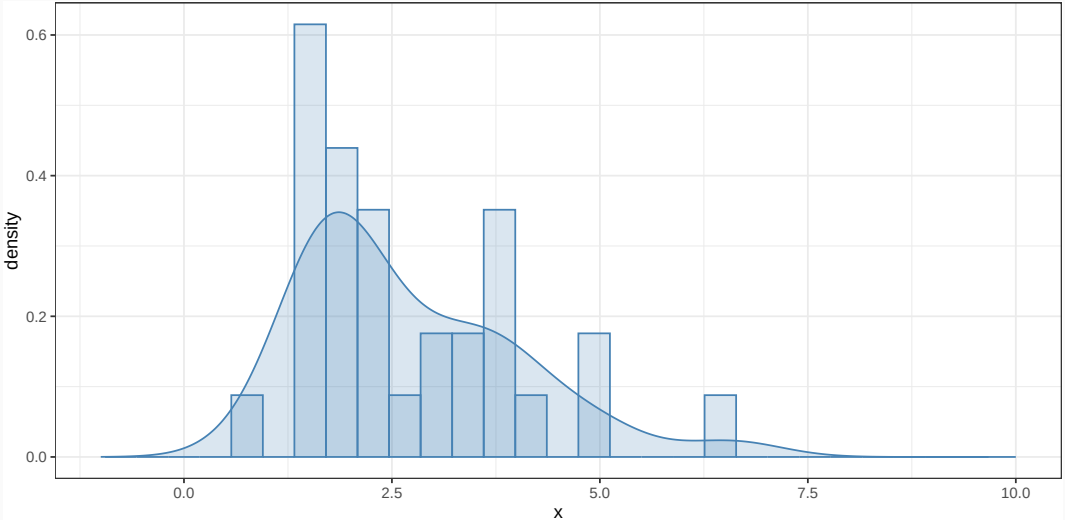
- Estimator: sample mean, $\bar{X}$

(Background info: the $\mathrm{Gamma}(\text{shape} = a, \text{rate} = b)$ distribution takes values only on the positive real line and has mean equal to $\mu = a/b$.)

Density of $\mathrm{Gamma}(\textbf{shape} = 3, \textbf{rate} = 1)$



Sample (n = 30)



The bootstrap CI

- The population distribution, and our sample, are positively skewed.
- The CLT tells us that the sampling distribution of the mean of the $\text{Gamma}(\text{shape} = 3, \text{rate} = 1)$ distribution will be approximately normal.
- We would expect that the bootstrap and CLT intervals will be similar, however the bootstrap may incorporate some of the skewness in our data.

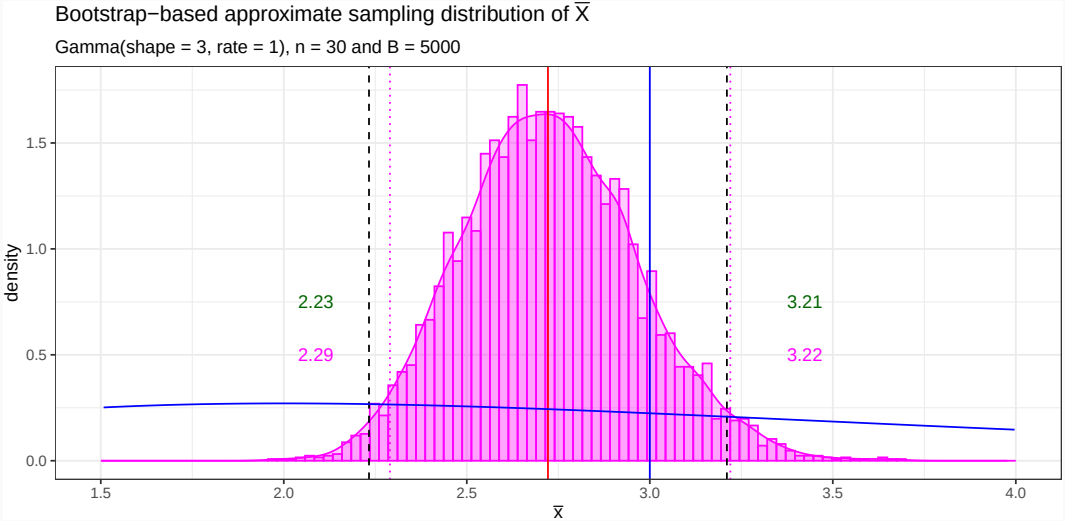
- The bootstrap CI is:

```
    2.5%    97.5%
2.287659 3.223428
```

- The CLT interval is:

```
    2.5%    97.5%
2.232840 3.210583
```

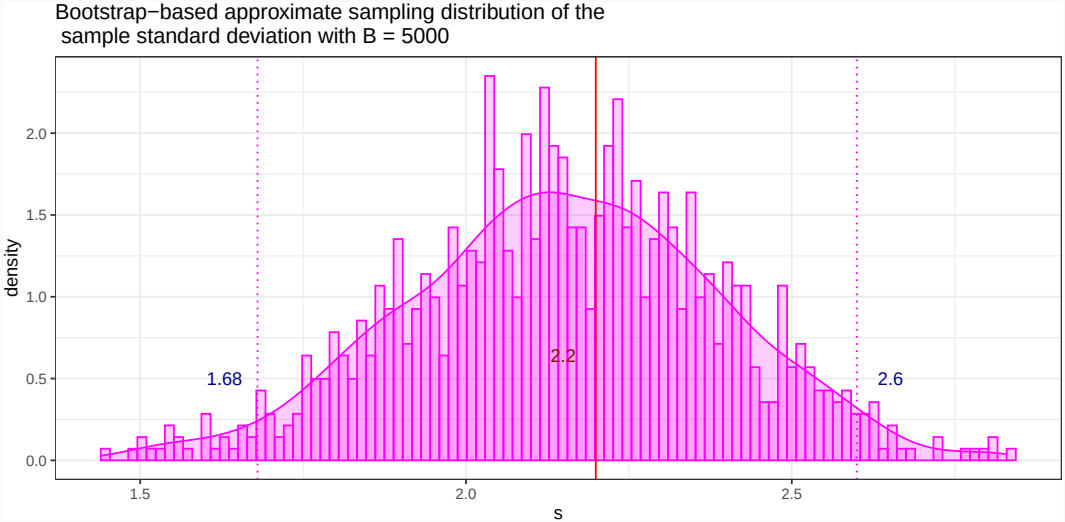
Visualisation



Bootstrap for the standard deviation

- The flexibility of the bootstrap technique is that we can calculate CIs for estimates for which we do not know the standard error or sampling distribution.
- We can't (based on what we have learnt thus far) use the CLT to obtain a CI for the sample standard deviation, s.
- The steps are the same as for the mean, except we use $\hat{\sigma} = s$ as our estimator.
- Return to the $\mathrm{N}(3, 4)$ example...

Plot of the approximate sampling distribution



The 95% bootstrap interval for s

- The bootstrap 95% confidence interval for σ is:

```
    2.5%    97.5%
1.679627 2.601597
```

- (For reference: the **true** standard deviation is $\sigma = 2$)
- Note: the sampling distribution looks normal, but the interval is not symmetric around the point estimate (of the standard deviation).

Outline

More advanced bootstrap methods

- There are many variations of the bootstrap approach
- We have only covered the **percentile bootstrap** today
- Key aspect common to all bootstrap methods: **resampling from the sample**
- What differs: which quantity is being resampled

Different bootstrap approaches

Some examples:

- **Nonparameteric bootstrap** (what we have learnt): replace cdf by empirical cdf
- **Smoothed bootstrap**: replace cdf by a smoothed empirical cdf (e.g., a "kernel density estimate")
- **Parameteric bootstrap**: draw samples from a fitted distribution rather than the empirical cdf.

The bootstrap is fallible

- When can the percentile bootstrap fail?
- If the empirical distribution is a poor approximation to the true distribution.
- For example:
  - Estimating extreme values (tail quantities, boundary points)
  - Small sample sizes
  - When there is substantial bias and skewness
- Diagnostics:
  - Visualise your bootstrap distribution
  - Check it is relatively smooth and symmetric.

Some fundamental principles

- Bootstrap main principle:

  Approximate the population by the sample.

- The ‘plug-in principle’:

  If something is unknown, substitute (‘plug in’) an estimator for it.

Hypothesis testing using bootstrap

- It is possible to use a bootstrap approach to carry out a hypothesis test
- But the process is more involved and easy to get wrong
- **We will not cover bootstrap hypothesis tests in this unit.**