9: Resampling Methods (The Bootstrap)

```
$ echo "Data Science Institute"
```

Activity (15 minutes)

Watch this video up tp Bootstrapping part: https://www.youtube.com/watch?v=uGsf3spCM3Y

The Bootstrap

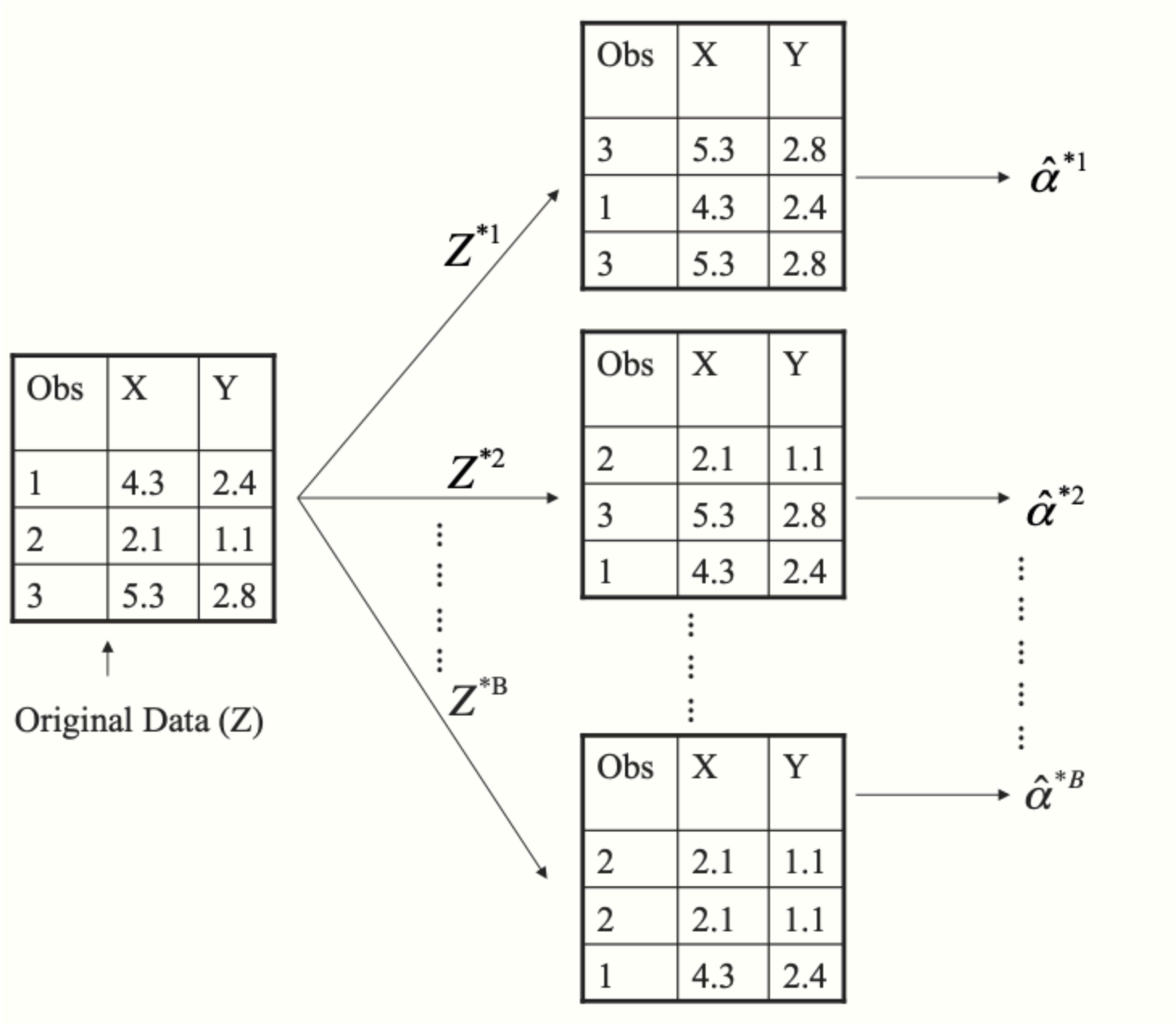
Suppose we wish to find the average age of the population of Toronto μ and we have a sample of size n. We can find the mean of the sample μ_s but this does not give any indication for how this compares to the true population mean μ.

The Bootstrap

♦ ***The bootstrap can be used to quantify the uncertainty of an estimate*** ♦ in the following way:

- Randomly sample n observations from the original sample to acquire a new sample of the same size (repeat observations are allowed).
- Compute the desired statistic (i.e. average age) of this new sample.
- Repeat steps 1-2 many times.
- Compute the standard error (SE) of the estimates.

This method is able to give us an estimate of the variability associated with our sample mean μ_s.


Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$

Breakout Room

When should you use Bootstrapping over Cross Validation methods?

Exercises: The Bootstrap

Open the The Bootstrap Jupyter Notebook file.

- Go over the "The Bootstrap" section together as a class.

`Portfolio` data in the `ISLP` package

Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y respectively, where X and Y are random quantities. We will invest a fraction α of our money in X, and will invest the remaining $1 - \alpha$ in Y . Since there is variability associated with the returns on these two assets, we wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $\mathrm{Var}(\alpha X + (1 - \alpha)Y)$. One can show that the value that minimizes the risk is given by:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

where $\sigma_Y^2 = \mathrm{Var}(Y), \sigma_X^2 = \mathrm{Var}(X), \sigma_{XY} = \mathrm{Cov}(X, Y)$

`Portfolio` data in the `ISLP` package

We can compute estimates for these quantities, $\hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\sigma}_{XY}$ using a data set that contains past measurements for X and Y. We can then estimate the value of α that minimizes the variance of our investment using:

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

Our goal in this exercise is to estimate the sampling variance of the parameter α given in the formula above.

References

Chapter 5 of the ISLP book:

James, Gareth, et al. "Resampling Methods." An Introduction to Statistical Learning: with Applications in Python, Springer, 2023.