SAPIENZA
UNIVERSITÀ DI ROMA

Sampling distributions and the bootstrap(1)

Serena Rosignoli

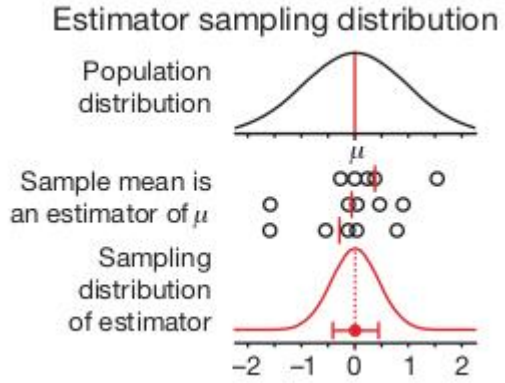
Data Analysis 2019/2020

Master degree in Genetics and Molecular Biology

*<< Samples are our windows to the population, and their statistics are used to **estimate** those of the population >> (2)*

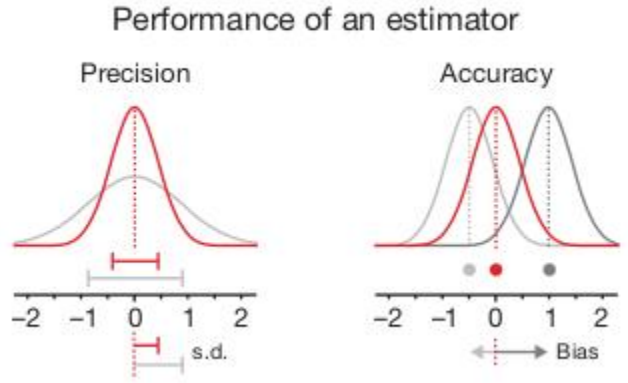
Central limit theorem (CLT): the sampling distribution of estimator become increasingly close to a normal distribution as the sample size increases.

$$\mu_{\bar{X}} = \mu$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$



- **Precision** can be measured as the standard error of an estimator and it tells us how much we can expect estimates to vary between experiments.

- To assess **accuracy** we need to measure bias, thus the expected difference between the estimate and the true value.



The **BOOTSTRAP** is a computational method used to resample by independently sampling with replacement from an existing sample data. (3-4)

- Introduced by Bradley Efron(5), with the aim of evaluating the accuracy of an estimator in the field of statistic inference.

Sample $X_1^{*(1)}, \cdots, X_n^{*(1)}$

↓

Random observation from the dataset

↓

Repeat *n* times $X_1^{*(2)}, \cdots, X_n^{*(2)}$

↓

Compute the statistic

Repeat **B** times $X_1^{*(B)}, \cdots, X_n^{*(B)}$

```
1  sample = [1.5, 7.8, 3.5, 6.0, 7.0, 7.5, 0.7]
```

```
1  [random.choice(sample) for n in range(6)]
```

Repeat **m** times

```
[7.5, 7.5, 0.7, 3.5, 3.5, 7.0]
```

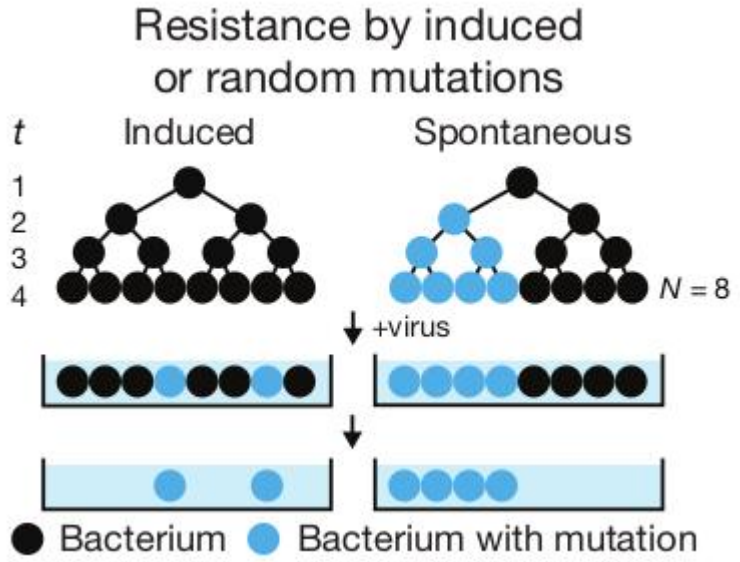
~~*random.sample()*~~

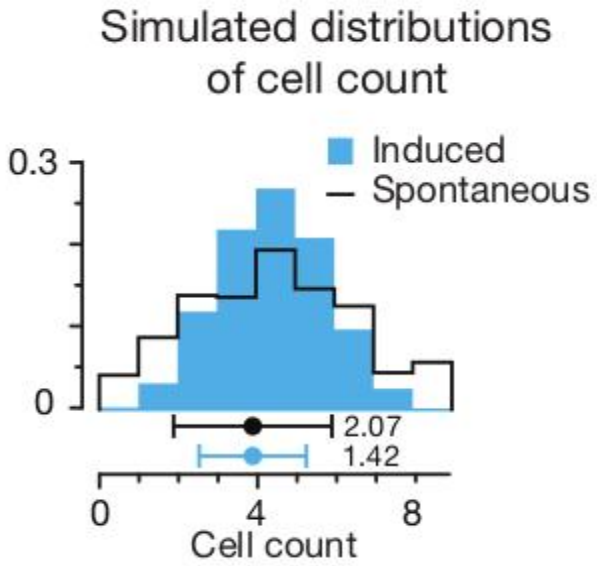
or

resample() function from **scikit** library

```
1  sklearn.utils.resample(data, replace=True,
                          n_samples=4, random_state=1)
```

Luria-Delbruck experiment
(1943) (6)





- Poisson
- Increase in Variance: not Poisson

Variance-to-Mean Ratio (VMR)

H0: VMR = 1
H1: VMR >> 1

HOW BOOTSTRAP CAN BE USED TO ESTIMATE THE UNCERTAINTY AND BIAS OF THE VMR USING MODEST SAMPLE SIZES.

25 cultures: as a sample, the count of cells in each culture is used.

- **Sample's mean** = 5.48
- **Variance** = 55.3
- **VMR** = 10.1

How much is the uncertainty of VMR estimator?

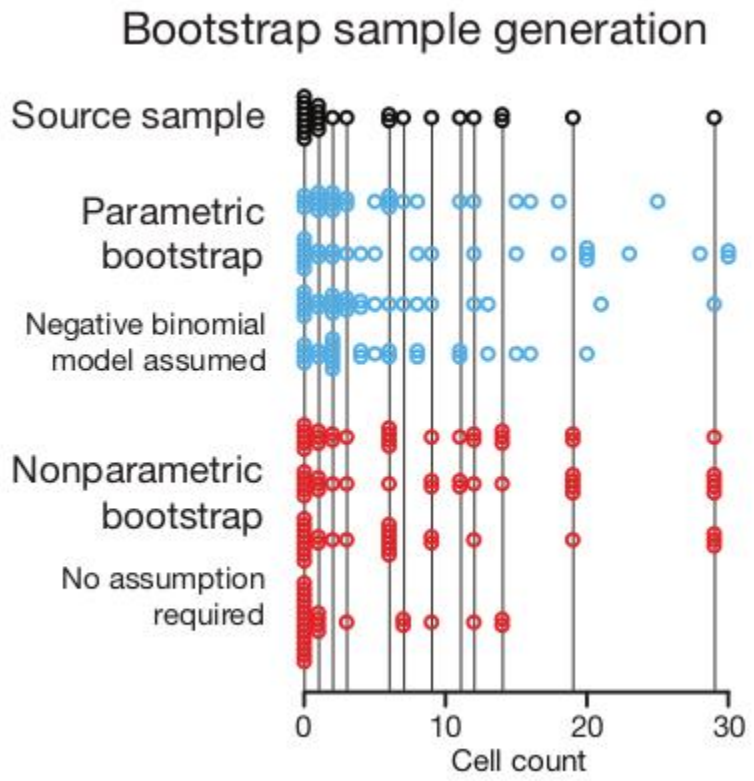
We don't have access to the sampling distribution

Plate more cultures?

Simulate more samples with the bootstrap

PARAMETRIC	NON PARAMETRIC
An underlying parametric distribution for the source sample is assumed.	No distribution is assumed

THE SAMPLING DISTRIBUTION OF COMPLEX QUANTITIES SUCH AS THE **VARIANCE-TO-MEAN RATIO (VMR)** CAN BE GENERATED FROM OBSERVED DATA USING THE BOOTSTRAP

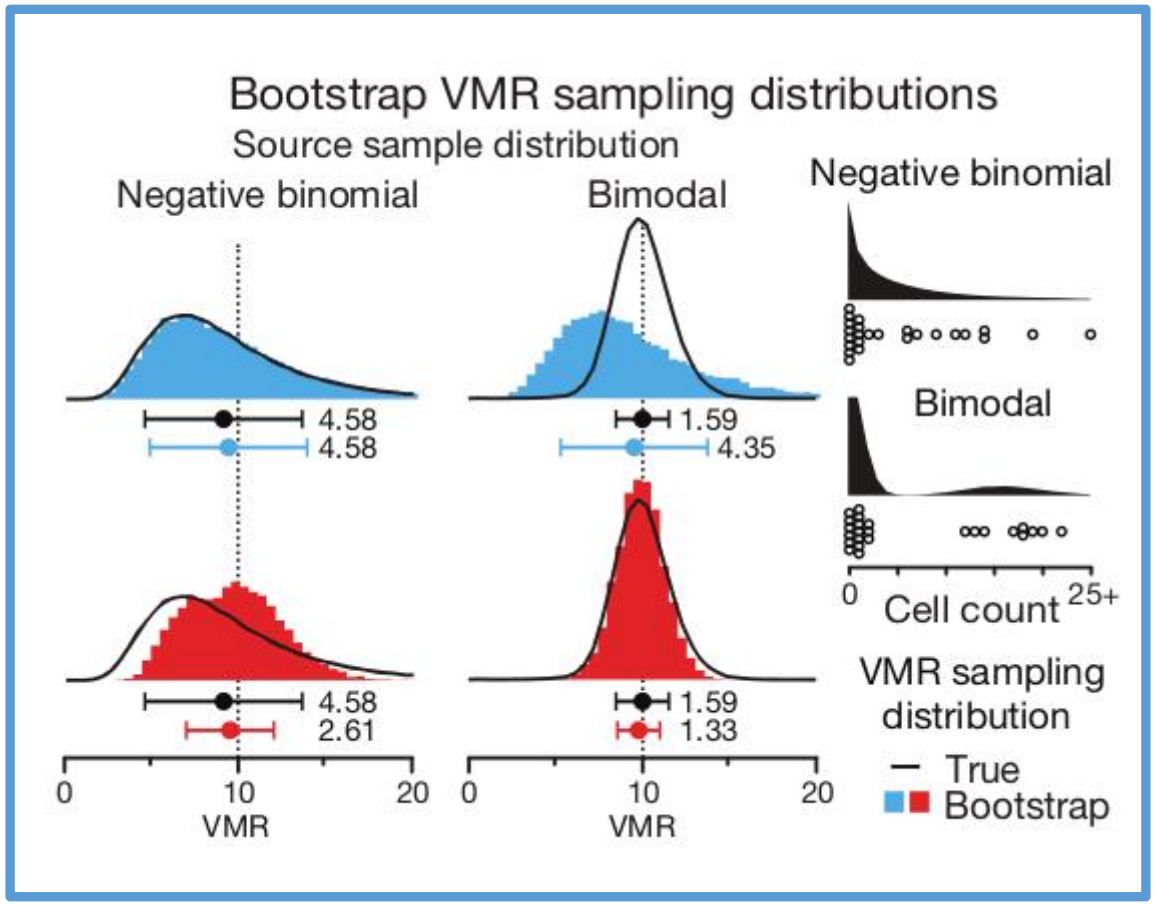


4* samples simulated using a parametric bootstrap that assumes a negative binomial model

4* samples simulated using a non-parametric bootstrap

*(4 of 10.000 samples with size n=25)

The s.d. measured suggest that the re-sampling method can be used as a measure of precision, given that we assumed the proper source distribution.



The non parametric bootstrap may understimate the sampling distribution s.d.

It is generally safer to use the non parametric bootstrap when we are uncertain of the source distribution.

No significant bias were present in the simulations.

PROS

- Simplicity
- Generality

CONS

- Not perfect with small samples
- Not all the type of parameters are well estimated (e.g. minimum of a function)
- Rare extreme values may be understimated

APPLICATIONS

- Generation of phylogenetic trees
- Machine learning algorithms performance measurment
- p-value adjustment
- CI estimation

CONTROVERSY

The so-called “parametric bootstrap” is not just a parametric simulation?

REFERENCES

1) Kulesa, A., Krzywinski, M., Blainey, P. et al. *Sampling distributions and the bootstrap*. Nat Methods 12, 477-478 (2015). https://doi.org/10.1038/nmeth.3414
2) Krzywinski, M. & Altman, N. Nat. Methods 10, 809-810 (2013) - *Importance of being Uncertain*
3) https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/
4) https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60
5) Efron, B.; Tibshirani, R. *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy.* Statist. Sci. 1 (1986), no. 1, 54--75. doi:10.1214/ss/1177013815. https://projecteuclid.org/euclid.ss/1177013815
6) Luria, S.E. & Delbrück, M. Genetics 28, 491-511 (1943)