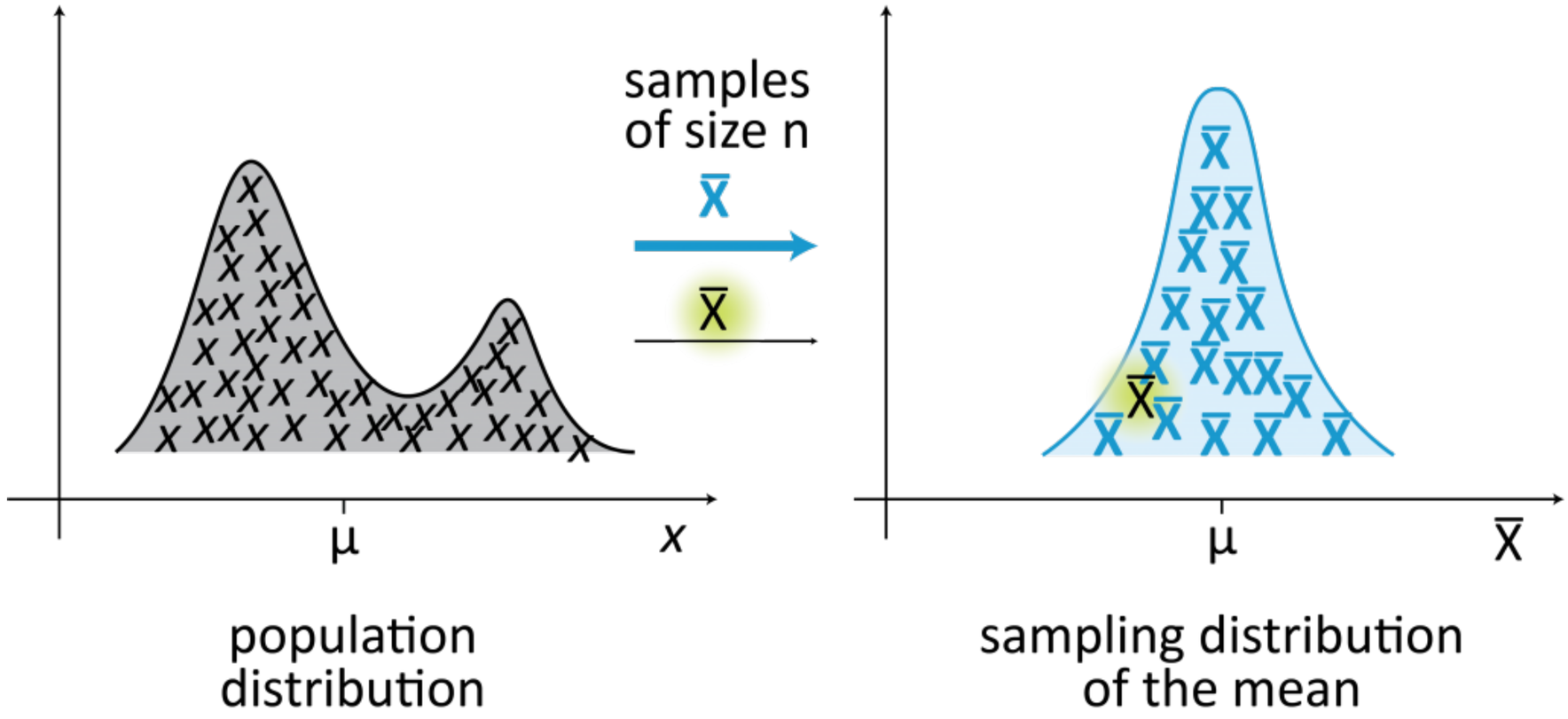
Foundational Statistics

Sampling and Parmeter Estimation



Why do we need parameter estimates?

1. Compare features of a system to values that are important to us practically or scientifically

2. Compare parameters between different populations (supported by hypothesis testing)

3. Need for theoretical work that seeks to understand what parameter values are realistic

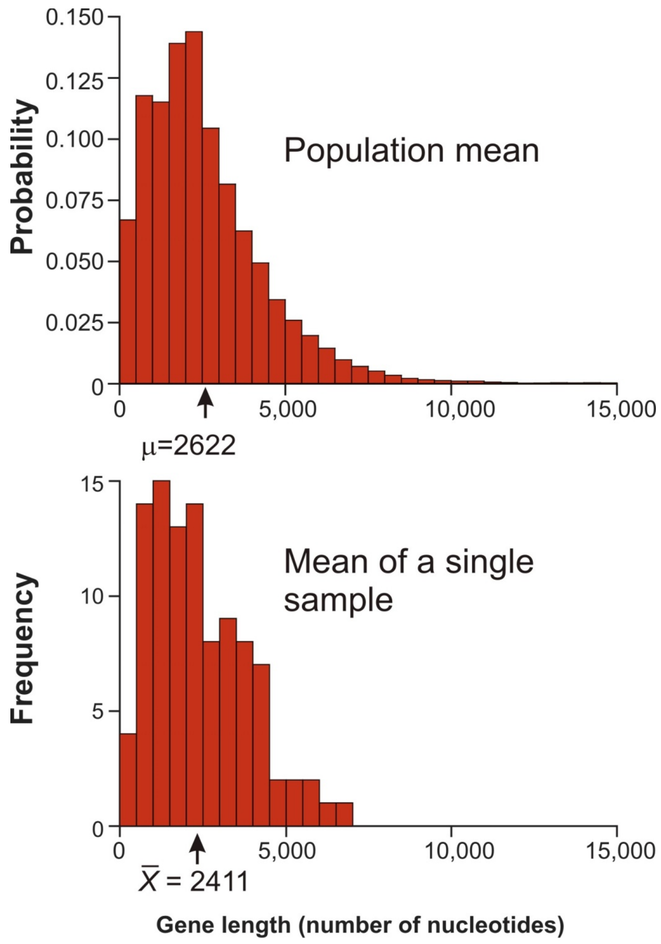
Estimation

- The process of inferring a population parameter from sample data
- The value of a sample estimate is almost never the same as the population parameter because of random sampling error
- Sampling distribution of an estimate
  - -all values we might have obtained from our sample
  - -probabilities of occurrence
- Standard error of an estimate
  - -standard deviation of a sampling distribution
  - -measures the reliability of a parameter estimate
  - -All reported estimates should include it, or a version of it.
- Estimates are more useful than your p-values for future research, because the standard error accounts for random sampling bias!

Estimating the sample mean from a single sample

We want to know the mean of the variable in the population

We have a sample mean (random sample of 100 genes)

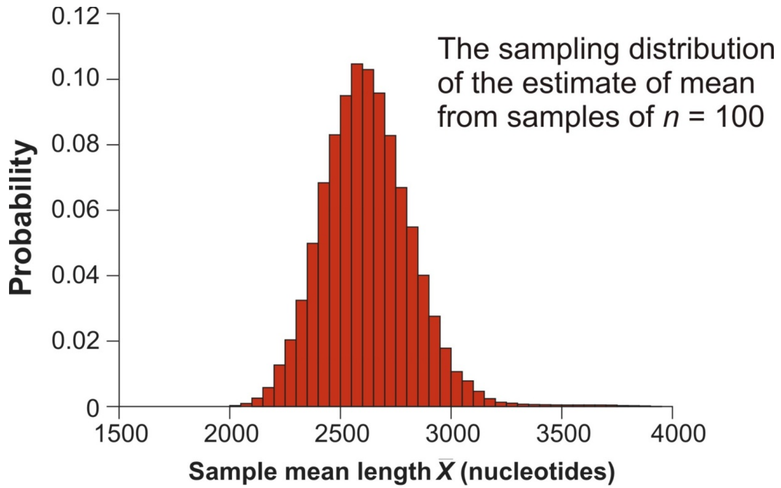


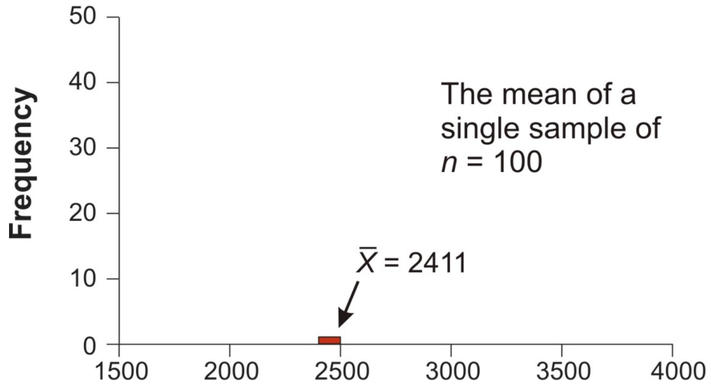
Estimating the sample mean from a single sample

Don’t have the true mean

Ideally we want the sampling distribution, which is all possible sample-based estimates and their probabilities

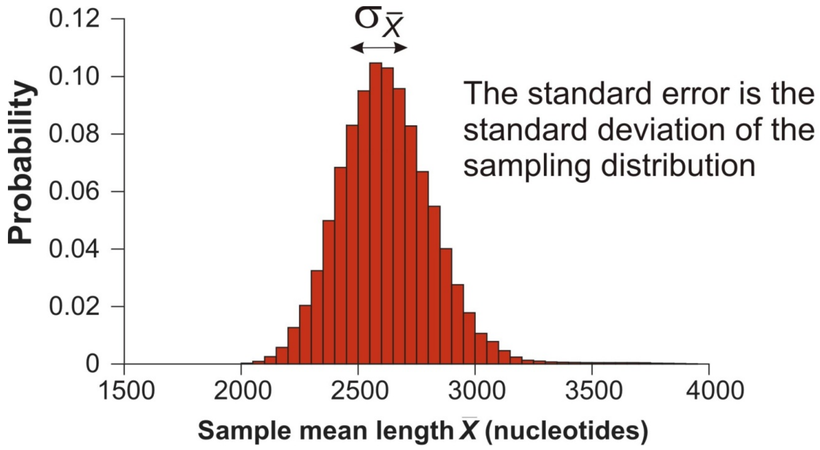
But we have just **one** sample mean



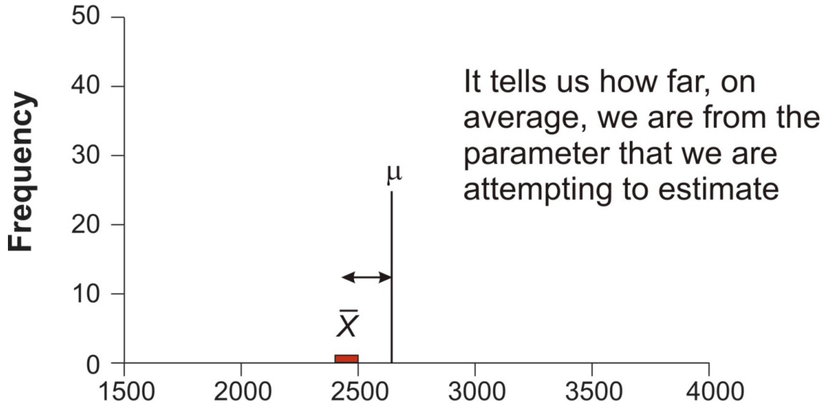


Estimating the sample mean from a single sample

Mostly want the standard deviation of the sampling distribution (aka the standard error)



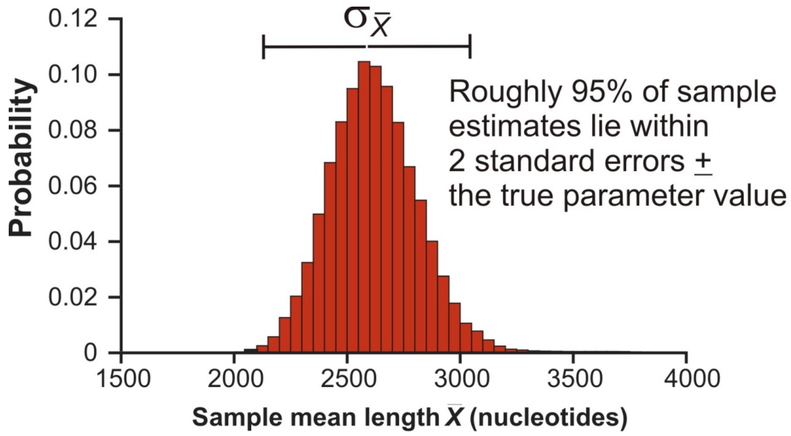
The standard error (SE) measures variation of the sample estimates around the population parameter

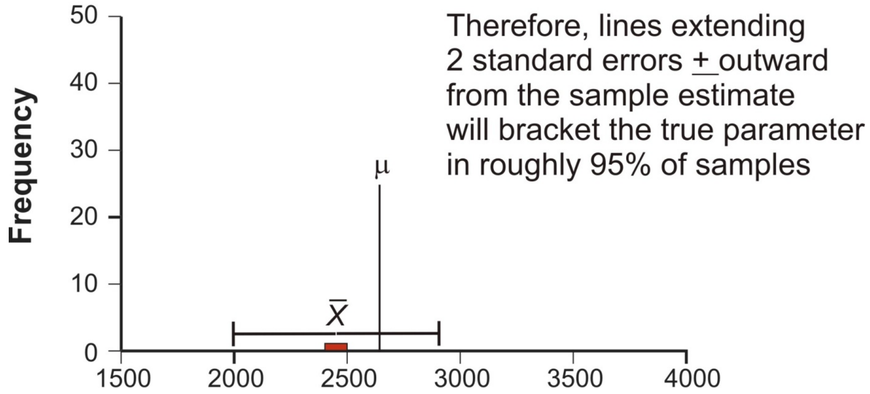


Estimating the sample mean from a single sample

About 95% of estimates fall within 2 SE's of population parameter

2 x SE approximates 95% confidence interval.

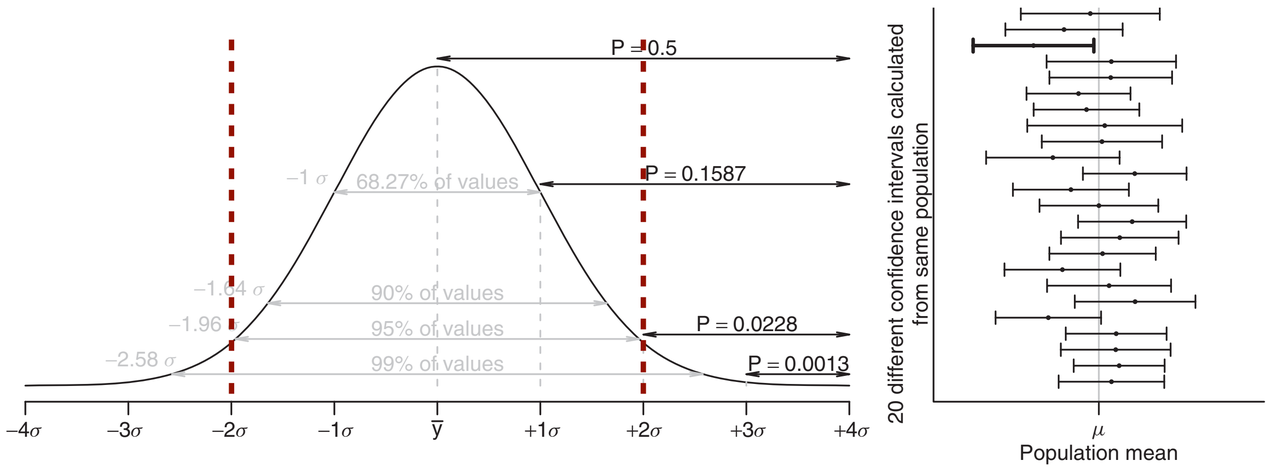




Estimating the sample mean from a single sample

Central limit theorem - any set of averaged values drawn from an identical population will converge towards the normal distribution

Sample distribution



95% Confidence Interval based on +/- 2 standard errors

Estimating the sample mean from a single sample - standard error of the mean (SEM)

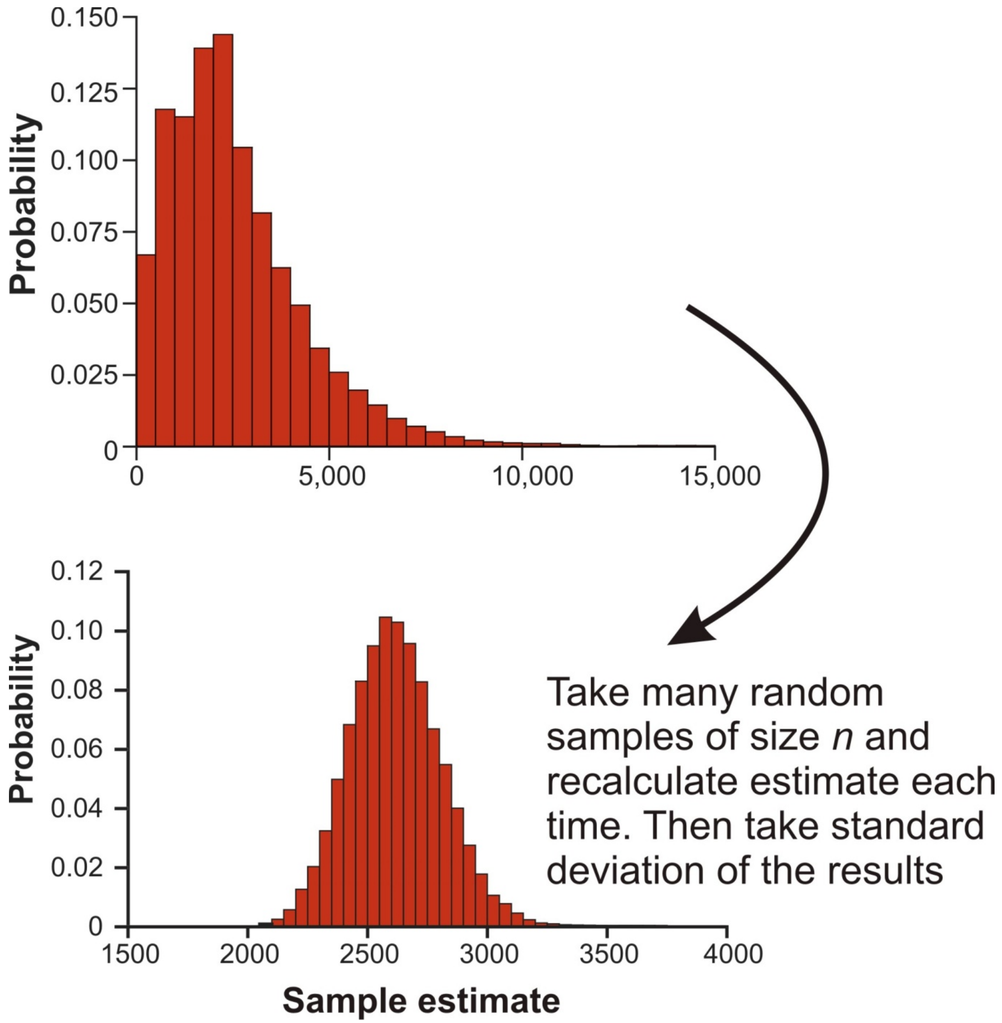
The SEM can be estimated from a single sample, thanks to the central limit theorem and the following equation:

$$\sigma_{\bar{x}} \approx s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$s_{\bar{x}}$ *is the estimated standard error. It is usually referred to just as the "standard error of the mean".*

Most other kinds of estimates do not have this amazing property.
What to do?
One answer: generate your own, approximate sampling distribution for the estimate using the '
Method invented by Efron (1979).

Estimating population parameters by “bootstrap” resampling

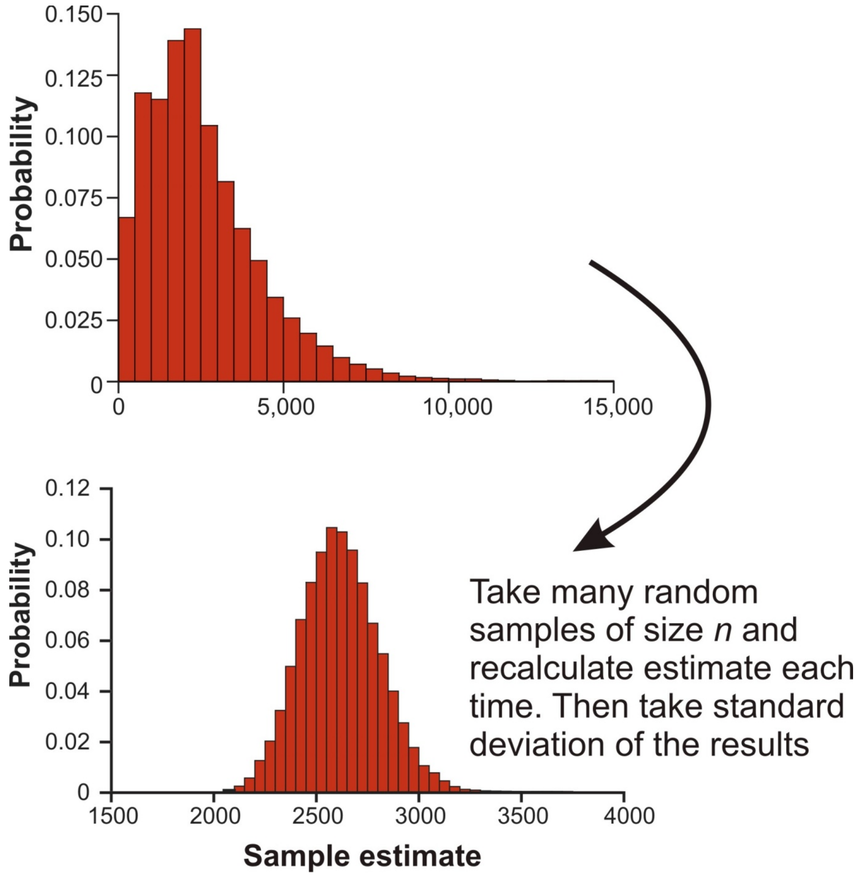


Estimating population parameters by “bootstrap” resampling

Ideal: Sample many times from the same population.

Calculate SE as the standard deviation of the resulting sampling distribution

But: Only have one sample, hence one estimate, so this is usually impractical



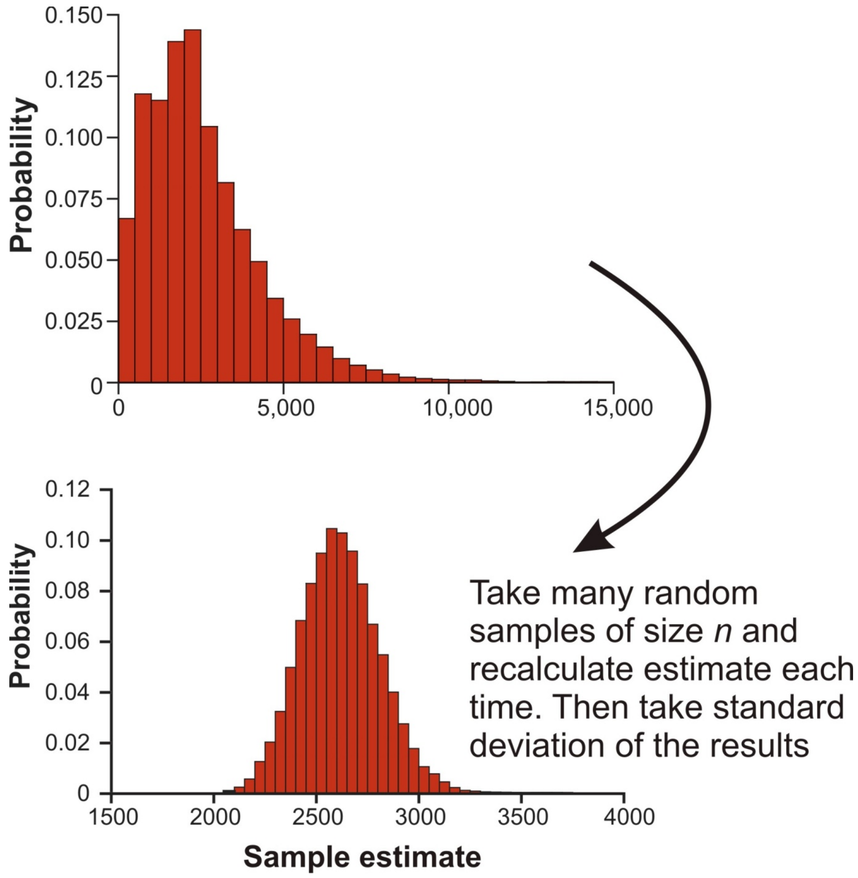
Estimating population parameters by “bootstrap” resampling

The bootstrap sampling distribution: next best thing

Sample many times from the single sample instead.

As if it were the population! Sampling is “with replacement” .

The SD deviation is **bootstrap standard error**



The bootstrap algorithm

- Use the computer to take a random sample of individuals from the original data, with replacement.
- Calculate the estimate using the measurements in the bootstrap sample (step 1). The first **bootstrap replicate estimate**.
- Repeat steps 1 and 2 a large number of times (1000 times is reasonable).
- Calculate the sample standard deviation of all the bootstrap replicate estimates obtained in step 3.
- The resulting quantity is called the **bootstrap standard error**.

Why the bootstrap is useful

- Can be applied to almost any sample statistic (including means, proportions, correlations, regression)
- Works when there is no ready formula for a standard error (e.g., median, trimmed mean, correlation, eigenvalue, etc.)
- Is nonparametric, so doesn't require normally-distributed data
- Works for estimates based on complicated sampling procedures or calculations (for example, it is used in phylogeny estimation)