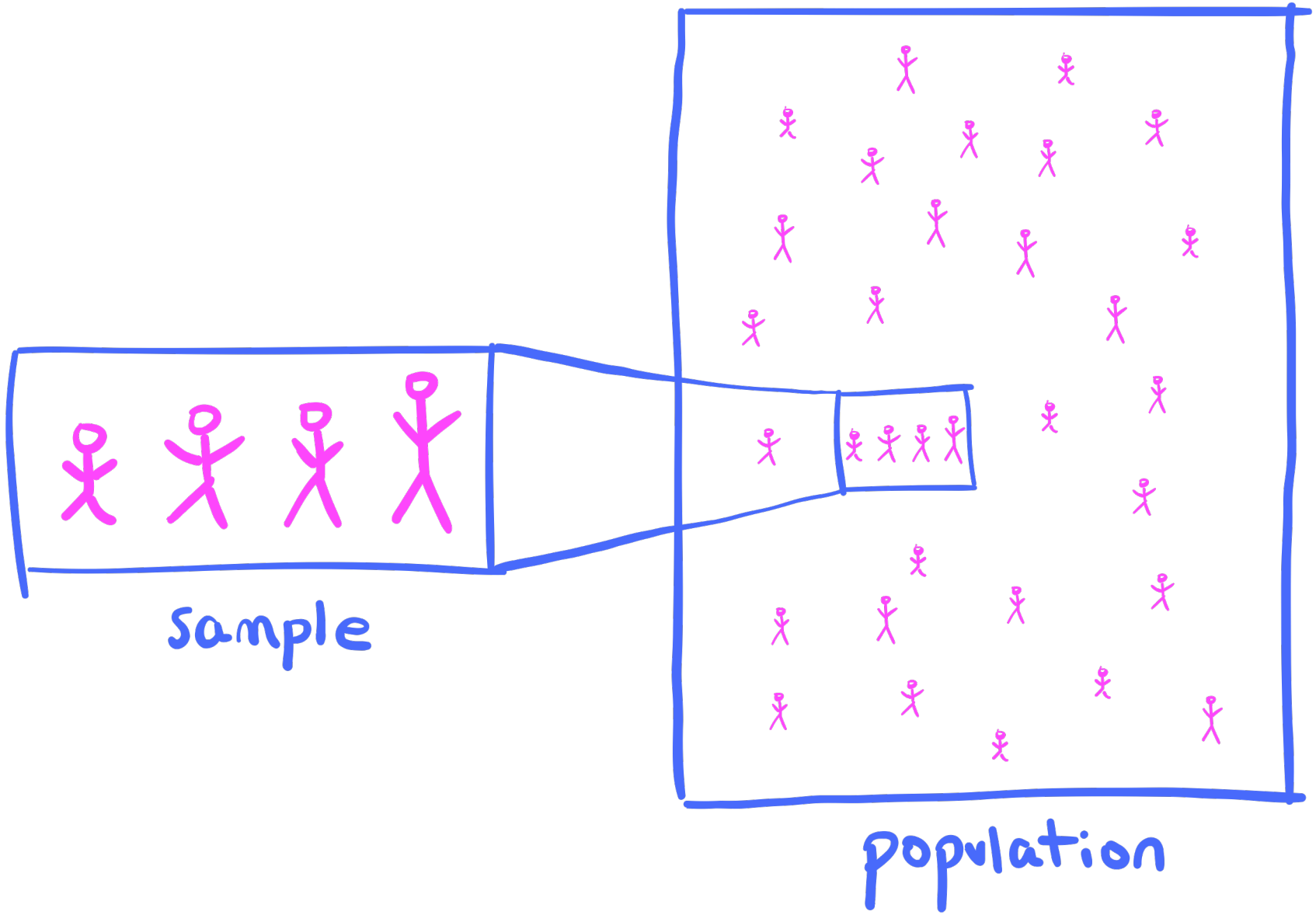
1.1. The goal of inferential statistics

> Neither perfect precision nor absolute certainty exist, in all cases, in the real world.


Sample
Population

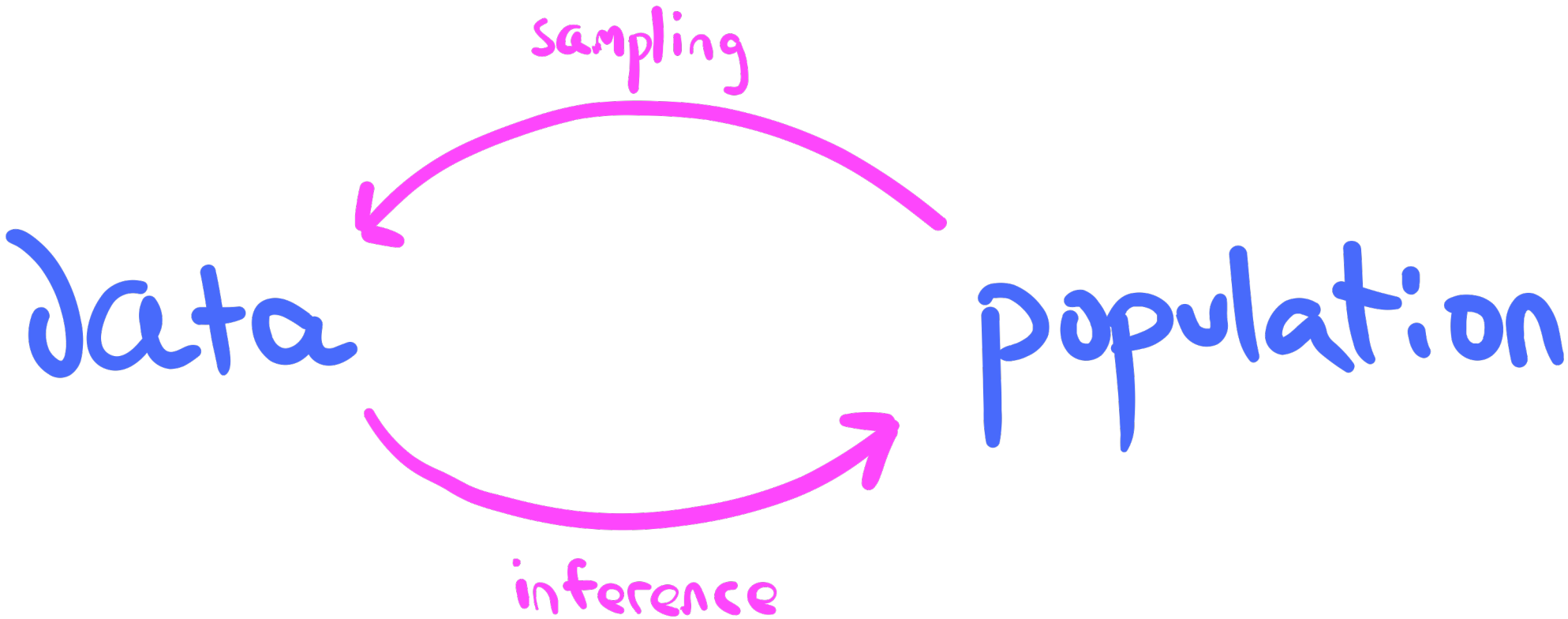
1. We estimate the median GPA of *all* incoming freshmen at a certain university based on a sample of incoming freshmen.
2. We estimate the likelihood that a sports team with a given record wins a championship based on the winning records of past champions.
3. We estimate the likelihood that a tumor of a certain size is malignant or benign based on the diagnoses of past tumors and their sizes.
4. We predict the next word in a new sentence fragment based on a sample of similar, complete sentences. (This is, in essence, the goal of so-called *large language models* in artificial intelligence.)
5. We estimate the true speed of light based on a sample of measurements obtained in a laboratory.

Definition 1.1

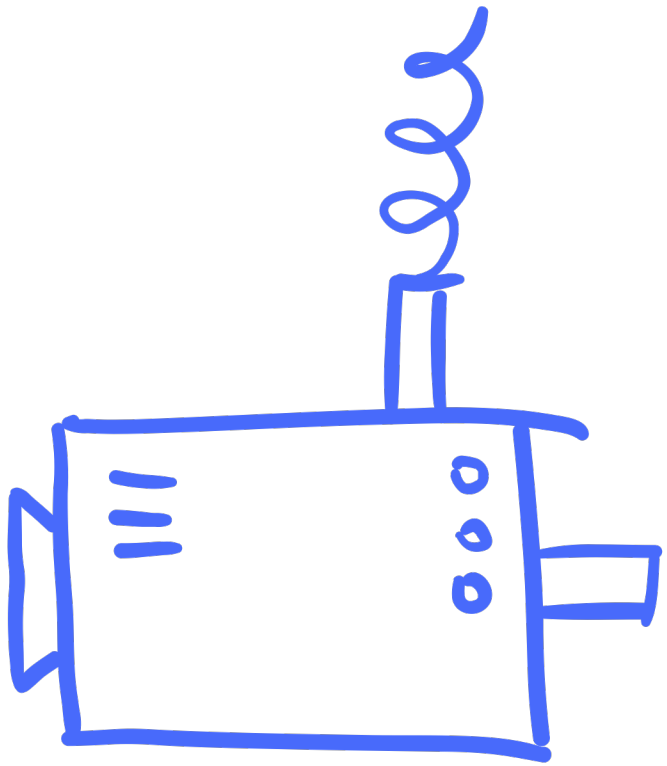
The *population* in a given scenario is the collection of all items or objects under consideration.

1. The median GPA of the population. (A *single* number.)
2. A probability for each win-loss record. (A *collection* of numbers.)
3. A probability for each tumor size. (A *collection* of numbers.)
4. A probability for each word. (A *collection* of numbers.)
5. The value of the single “true” speed of light. (A *single* number.)

> The **goal of inferential statistics** is to estimate population parameters based on samples and datasets.


sampling
data
population
inference

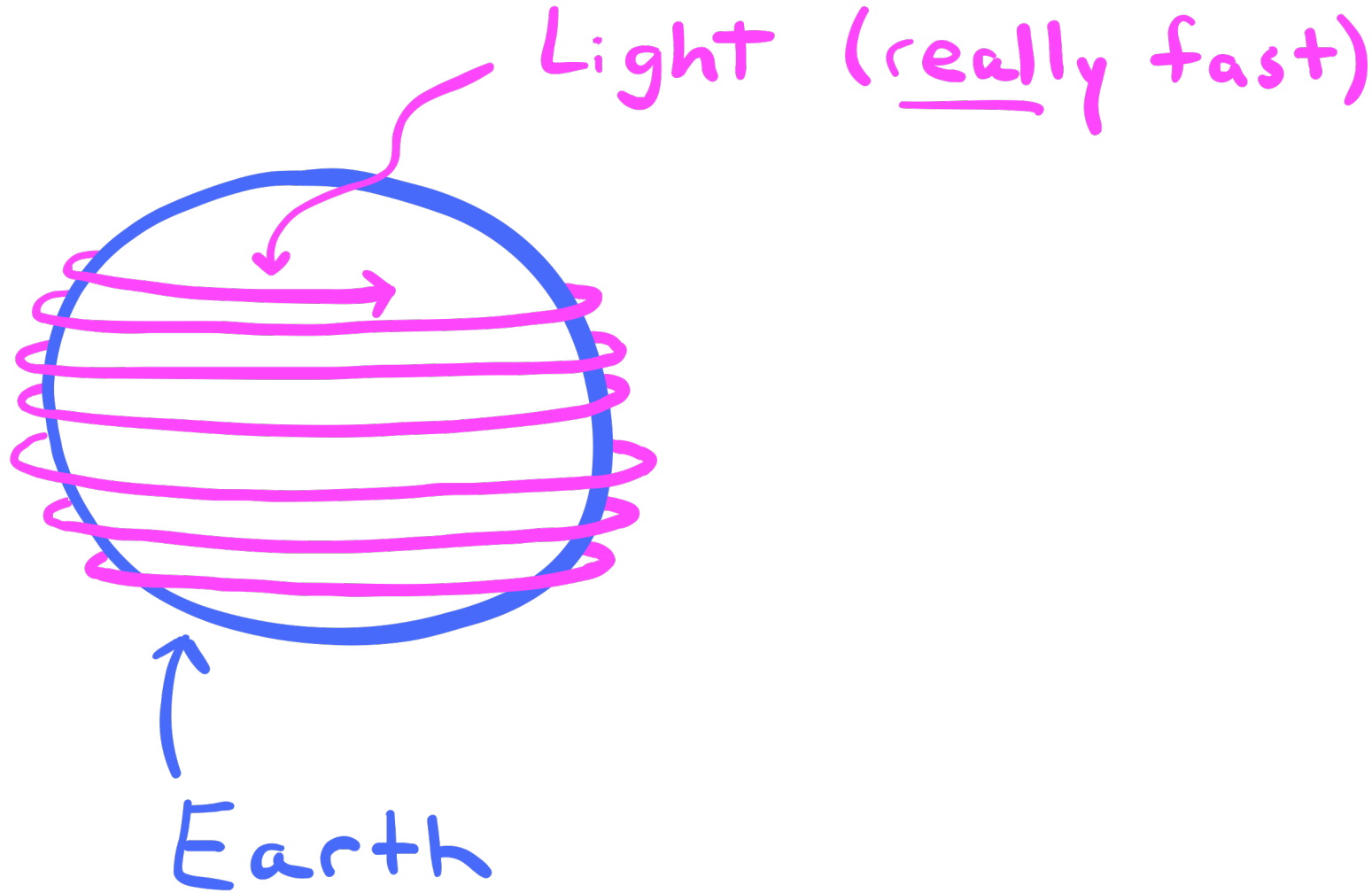
Inference from data



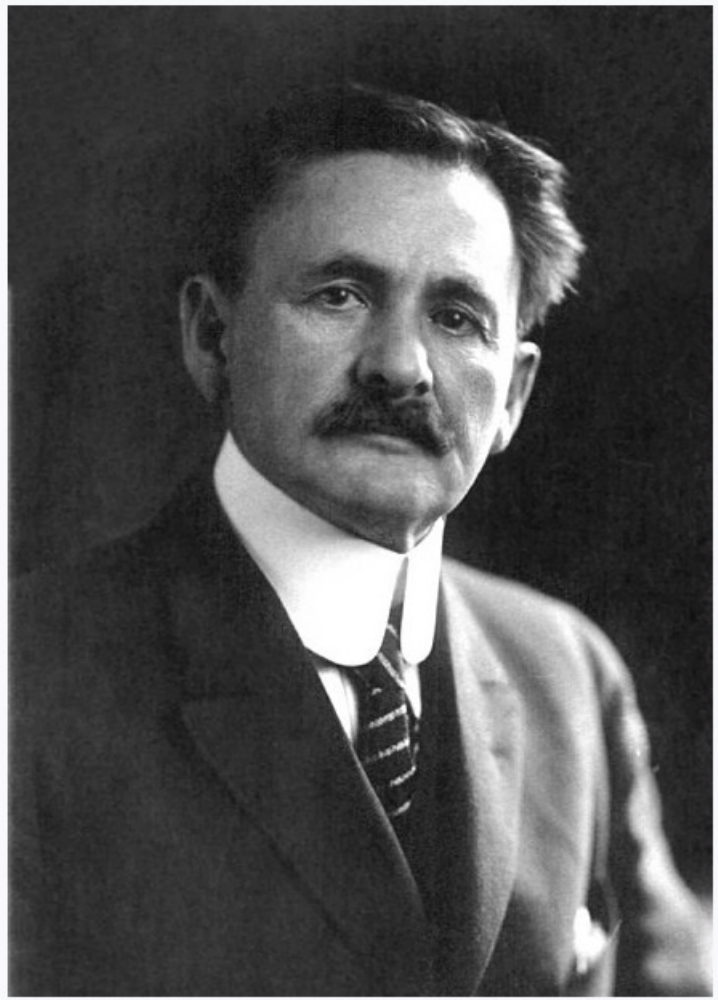
data → estimator → estimate of population parameter

1.2. Case study: the speed of light and statistical estimators

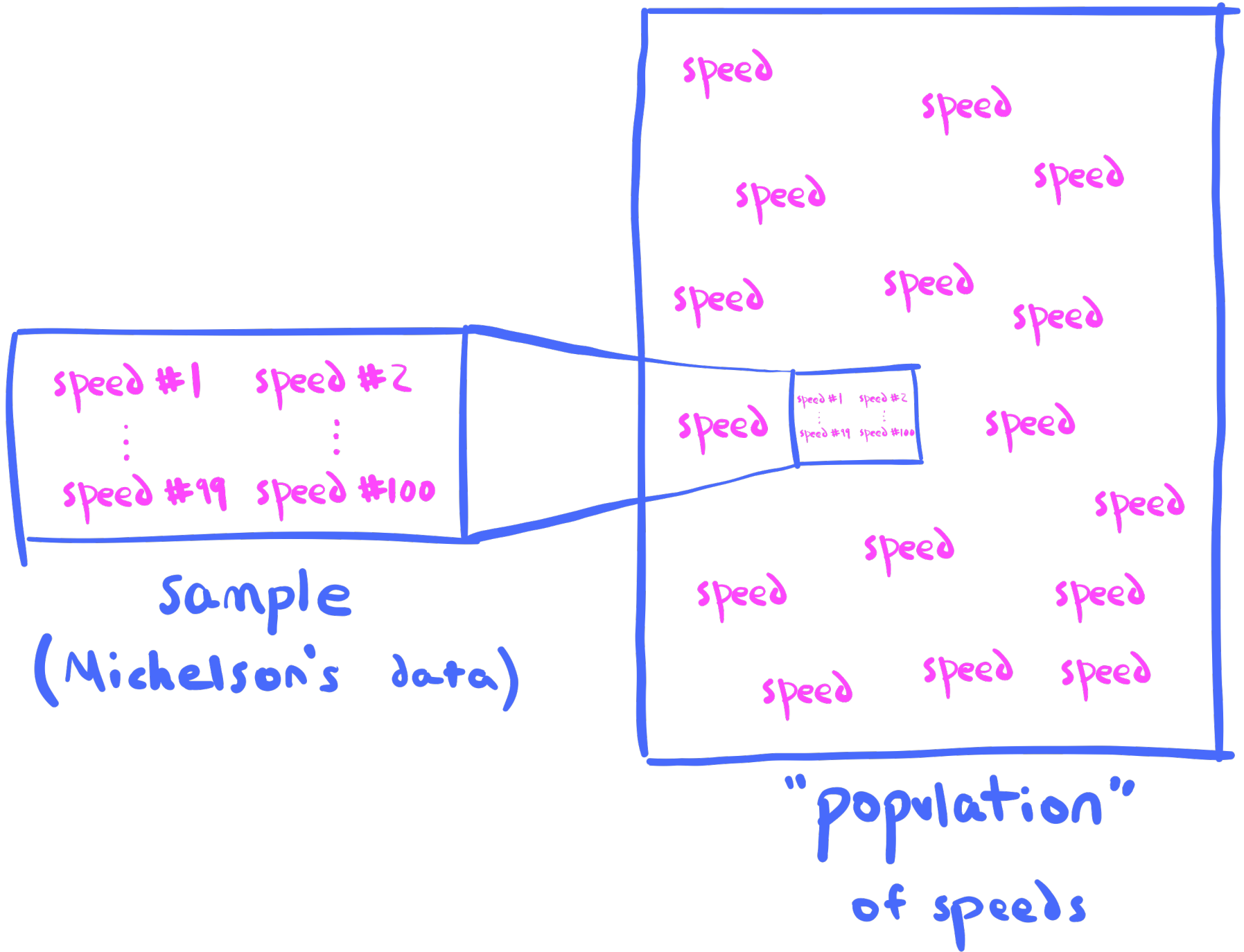
> Probability theory gives us the tools to *quantify* our confidence in our estimators.

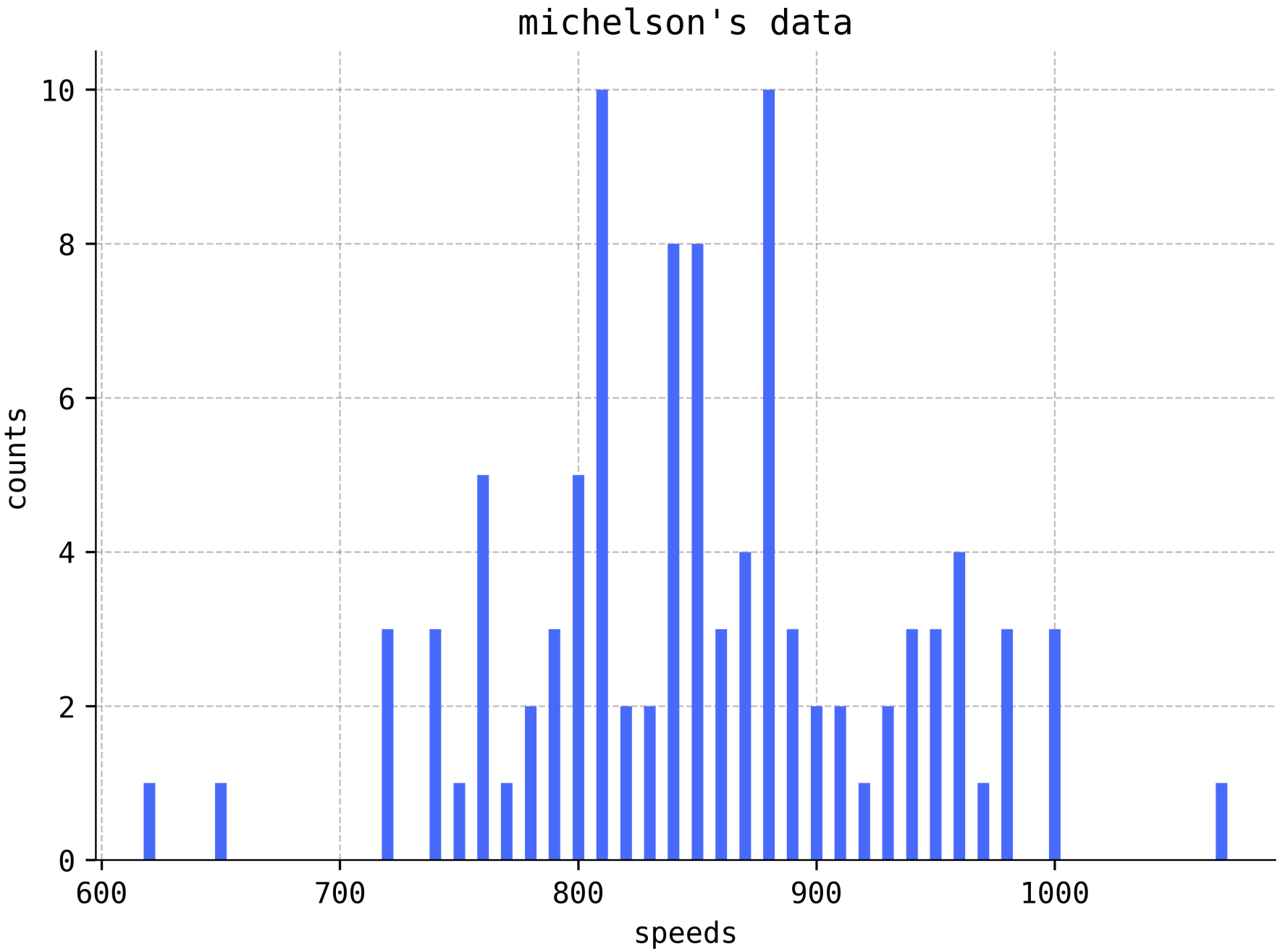

Light (really fast)
Earth

Albert A. Michelson

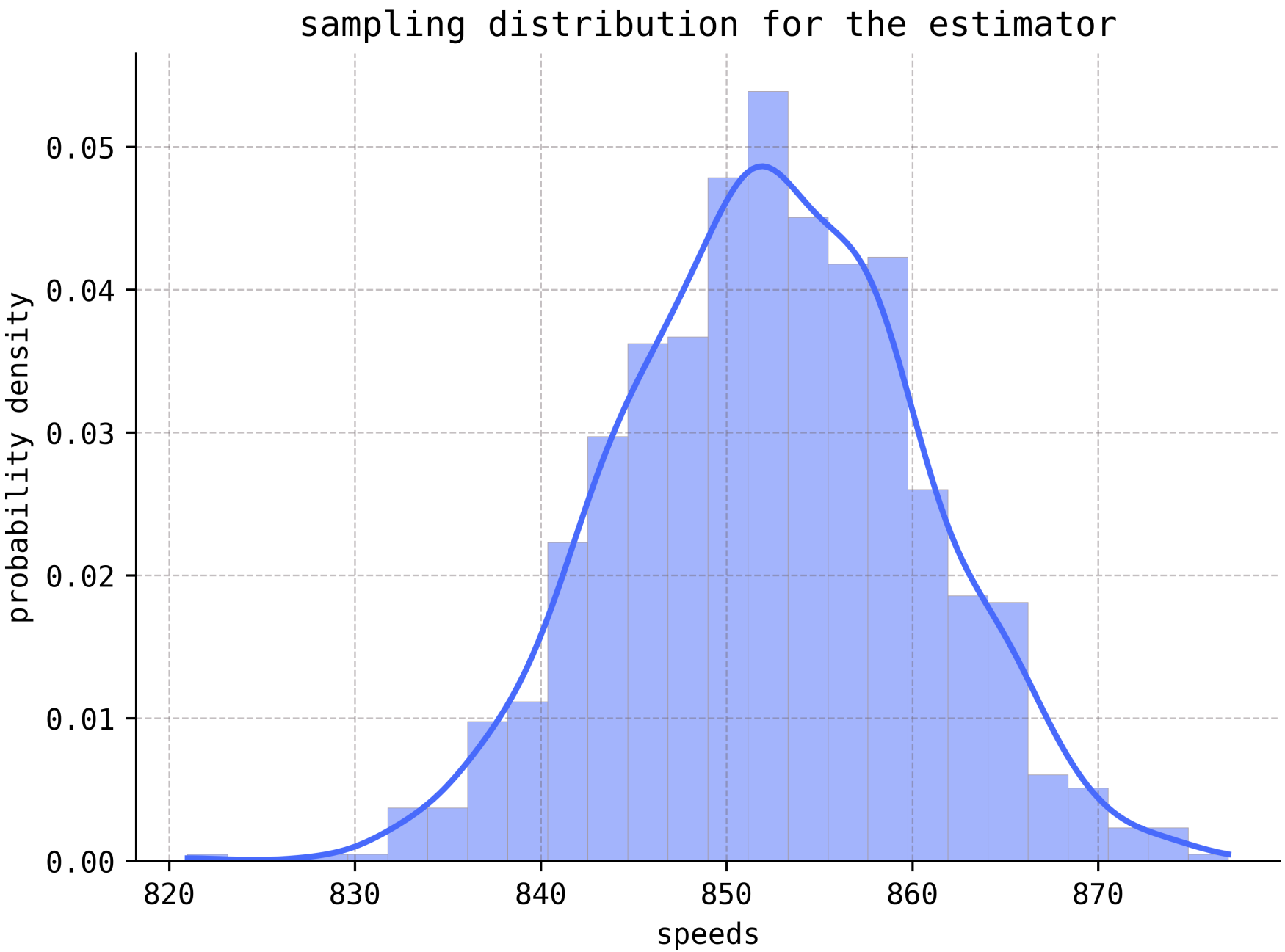


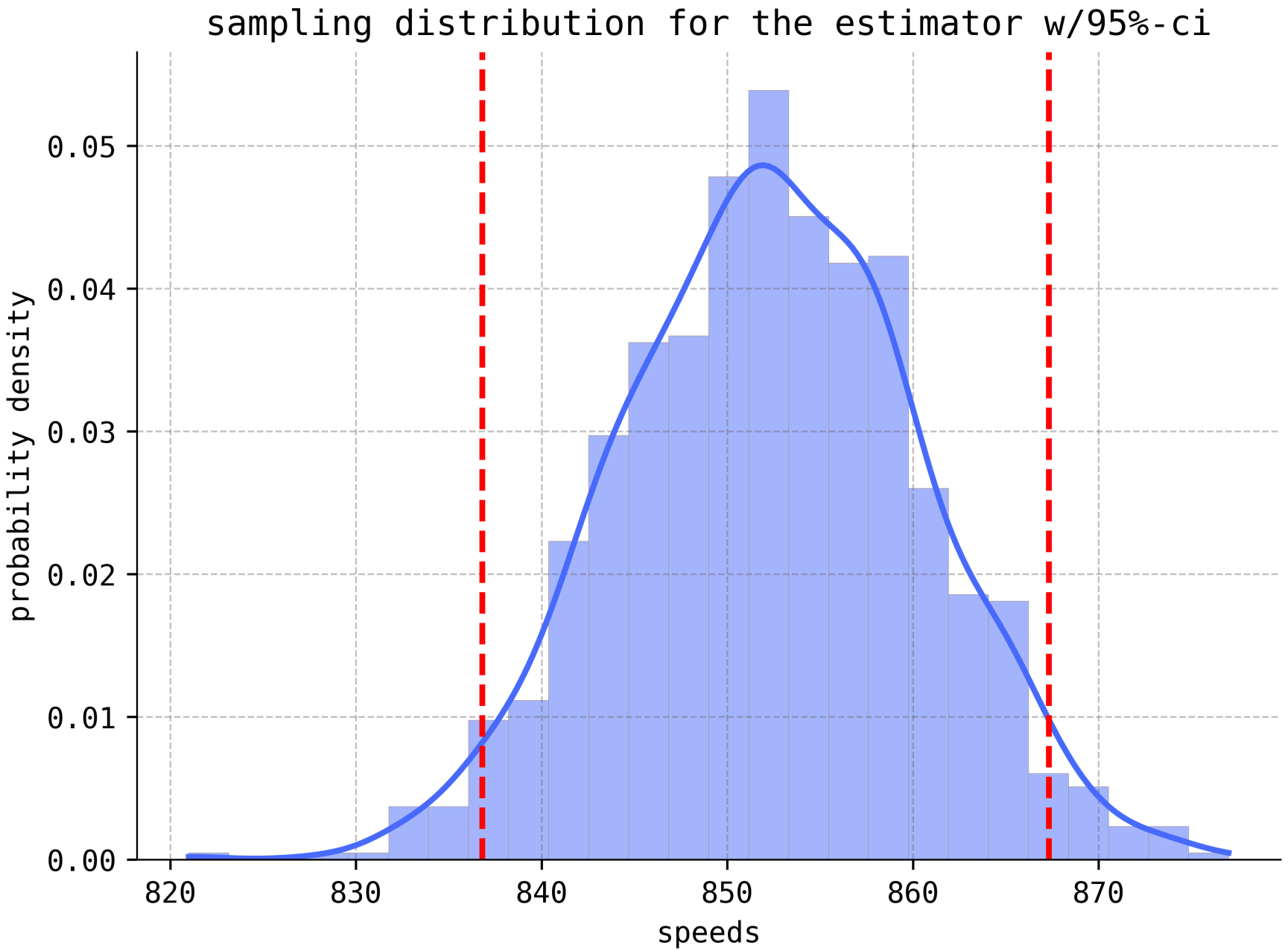
	speeds
0	850
1	1000
2	960
3	830
4	880
...	...
95	840
96	760
97	780
98	850
99	870


speed #1
speed #2
speed #99
speed #100
Sample
(Michelson's data)
speed
"population"
of speeds


michelson's data
counts
speeds
0
2
4
6
8
10
600
700
800
900
1000

> Since the dataset has variability, so too does our estimator.


sampling distribution for the estimator
probability density
0.00
0.01
0.02
0.03
0.04
0.05
820
830
840
850
860
870
speeds


sampling distribution for the estimator w/95%-ci
probability density
0.00
0.01
0.02
0.03
0.04
0.05
820
830
840
850
860
870
speeds

> The variability in our estimator is relatively small-ish, which is evidence for a high-ish level of confidence in our estimator.

1.3. Summary

- First, one must construct an estimator for the target population parameter.

- Then, one conducts a statistical and probabilistic analysis to describe the variability (i.e., precision), the bias (i.e., accuracy) and other properties of the estimator. These metrics serve as measures for our “confidence” in an estimator’s outputs and allow us to compare one estimator to another.