Week 4

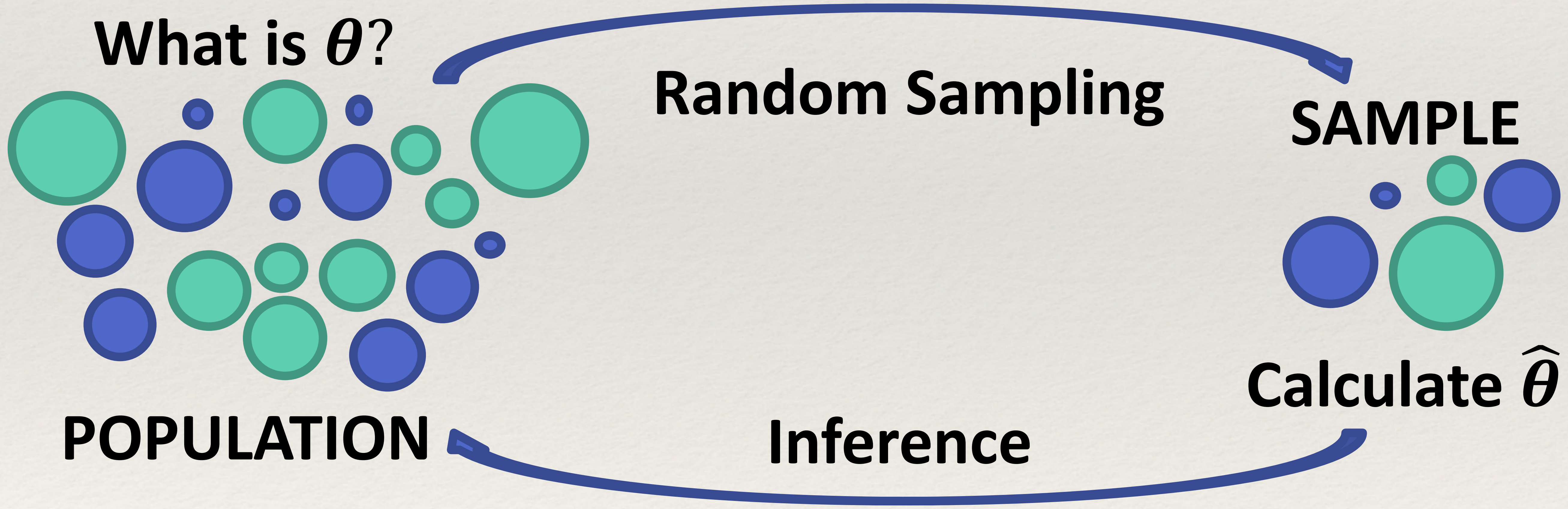
Sampling Variability & The Central Limit Theorem

ST 314
Introduction to Statistics for Engineers

Oregon State University

Inferential Statistics

Recall that **inferential statistics** use information from a sample to estimate or test characteristics from a population of interest.



How good is the sampled statistic $\hat{\theta}$ at estimating the population parameter θ?

Things to consider:

- Method of data collection
- Sampling variability
- Sample size

Point Estimates

As part of a quality control process for computer chips, an engineer a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(a) What population is under consideration in the data set?

All chips manufactured at this factory during the week of production.

(b) What is the parameter being estimated?

p = proportion of defective chips from the pop.

(c) Based on the sample what is the **point estimate** for the parameter?

$$\hat{p} = \frac{27}{212} = 0.127$$

Sampling Variability

Suppose the study previously described was repeated by two other engineers in the same week. The following table gives the results from each of the three studies.

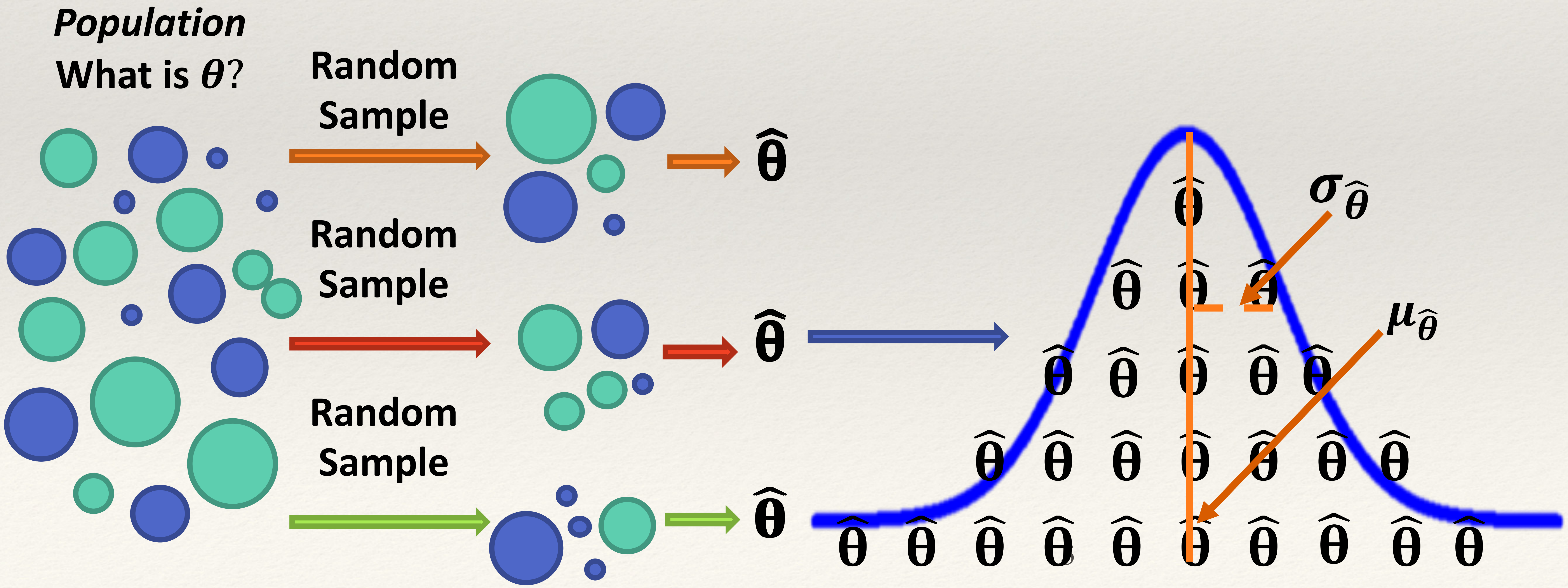
Study	Number of Chips Sampled	Number of Defective Chips	$\hat{p}$
1	212	27	0.127
2	212	19	0.090
3	212	23	0.108

Compare the point estimates from the three studies. What do you notice? Which point estimate is the "best"?

All point estimates are different

Sampling Distributions

The probability distribution of a statistic, $\hat{\theta}$, is the **Sampling Distribution.** The sampling distribution defines the variability of $\hat{\theta}$ and quantifies the chance occurrence of specific values.



Statistics are random variables! If N is the number of units in the population and n is the sample size, there are $\binom{N}{n}$ possible sample combinations.

Unbiased Estimators

- Common statistics, such as $\hat{p}, \bar{x},$ and s^2 are **unbiased.**
- The expected values of these estimators are equal to their parameters:

$$E(\hat{\theta}) = \theta$$

$E(\bar{x}) = \mu \leftarrow$ pop. mean

$E(\hat{p}) = p \leftarrow$ pop. proportion

$E(s^2) = \sigma^2 \leftarrow$ pop. variance

sample variance

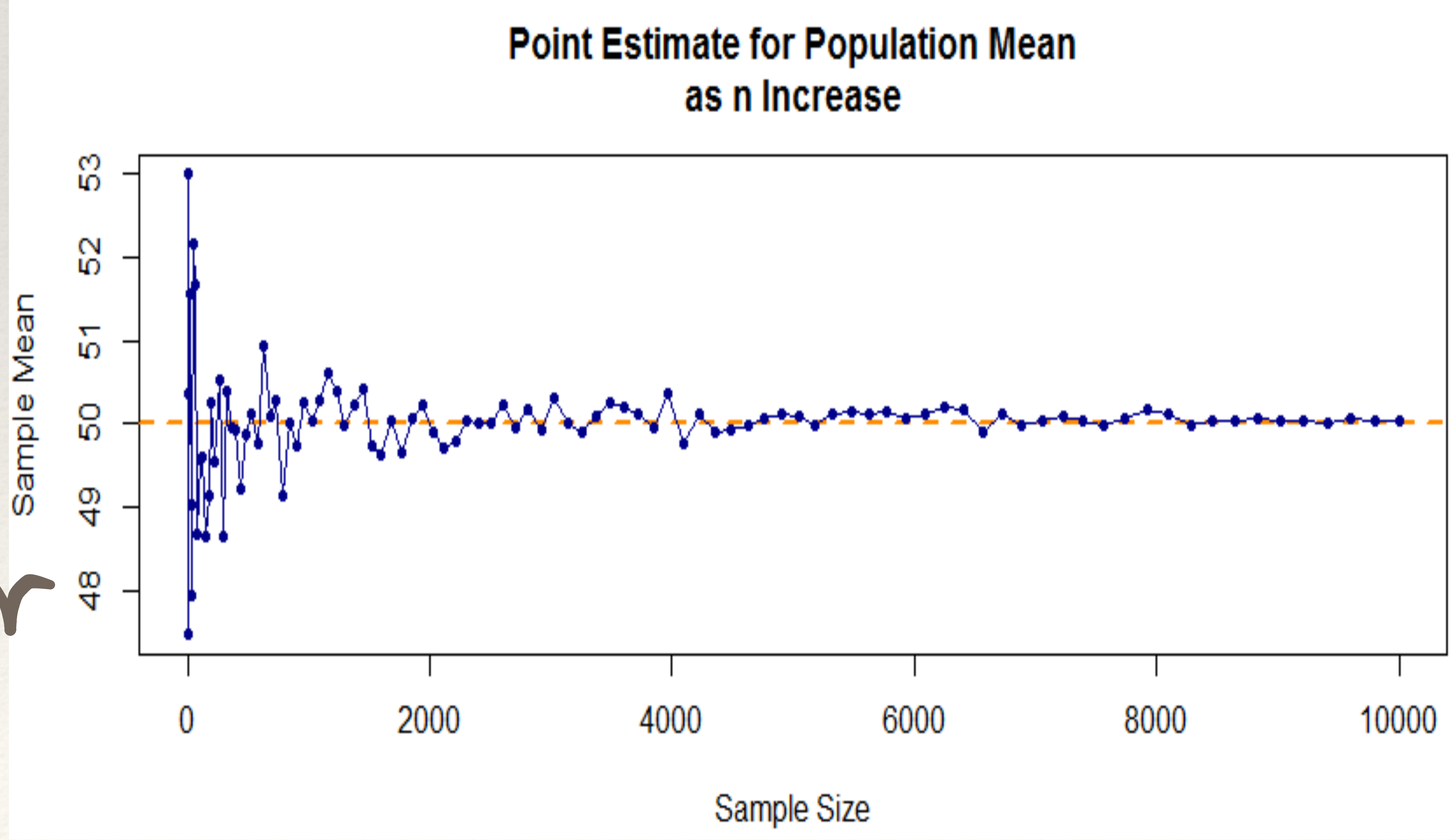
Law of Large Numbers

The **Law of Large Numbers** states that as n increases, the statistic will approach the true population parameter.

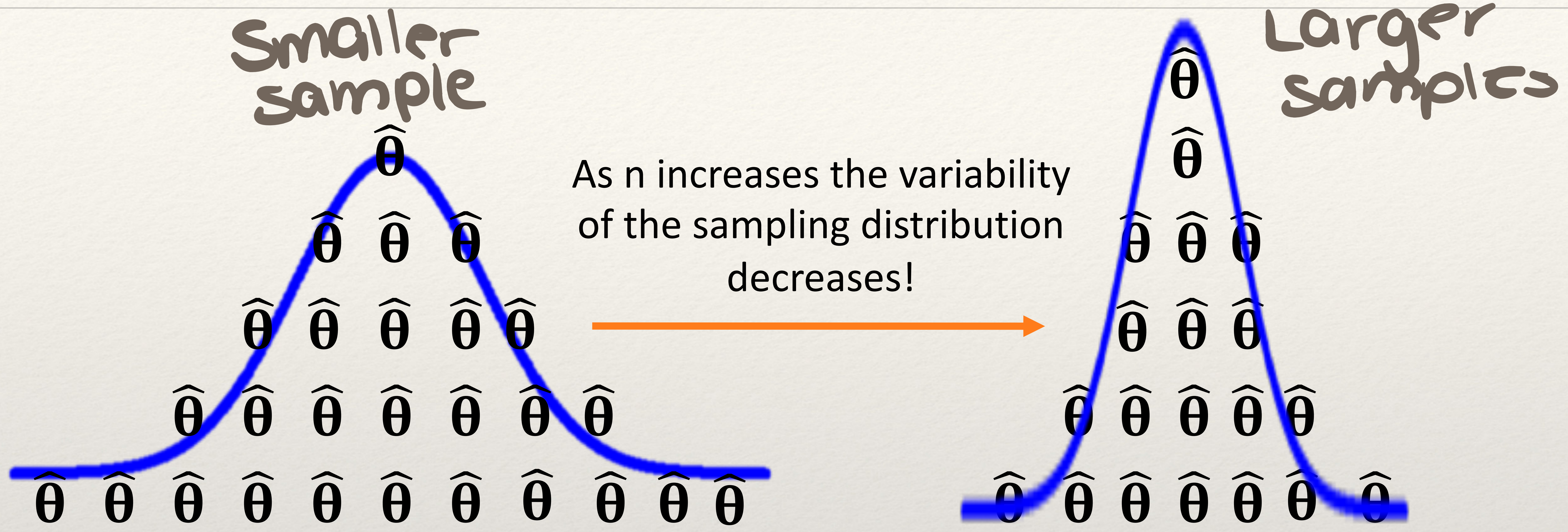
As $n \to \underset{\text{pop. size}}{N}$, $\hat{\theta} \to \theta$

More formally, $\hat{\theta}$ converges in probability to θ. This implies that $\hat{\theta}$ is a consistent estimator. $\bar{x}$, $\hat{p}$, and s^2 are all consistent

Larger sample size doesn't guarantee a closer estimate.



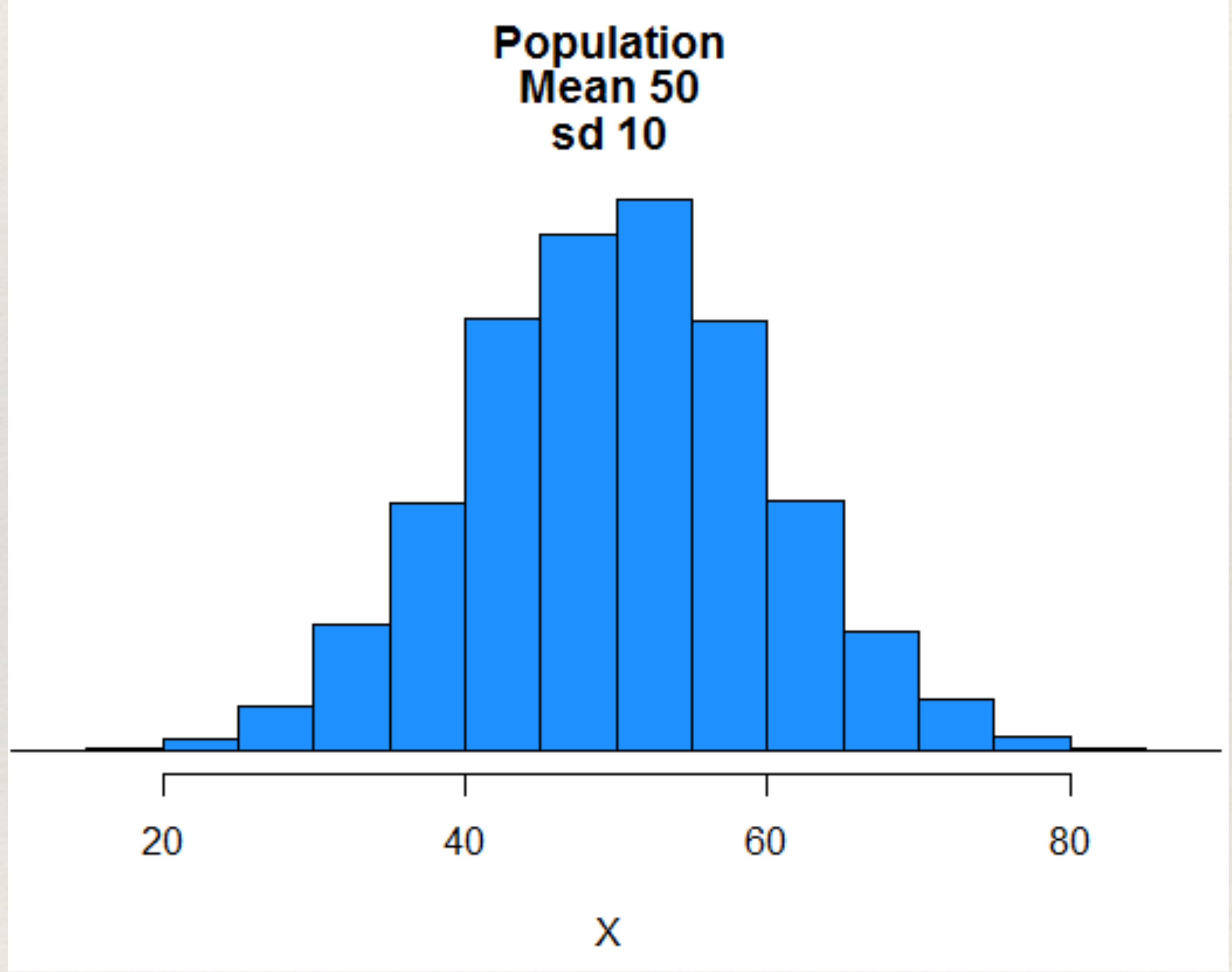
Sample Size & Sampling Variability

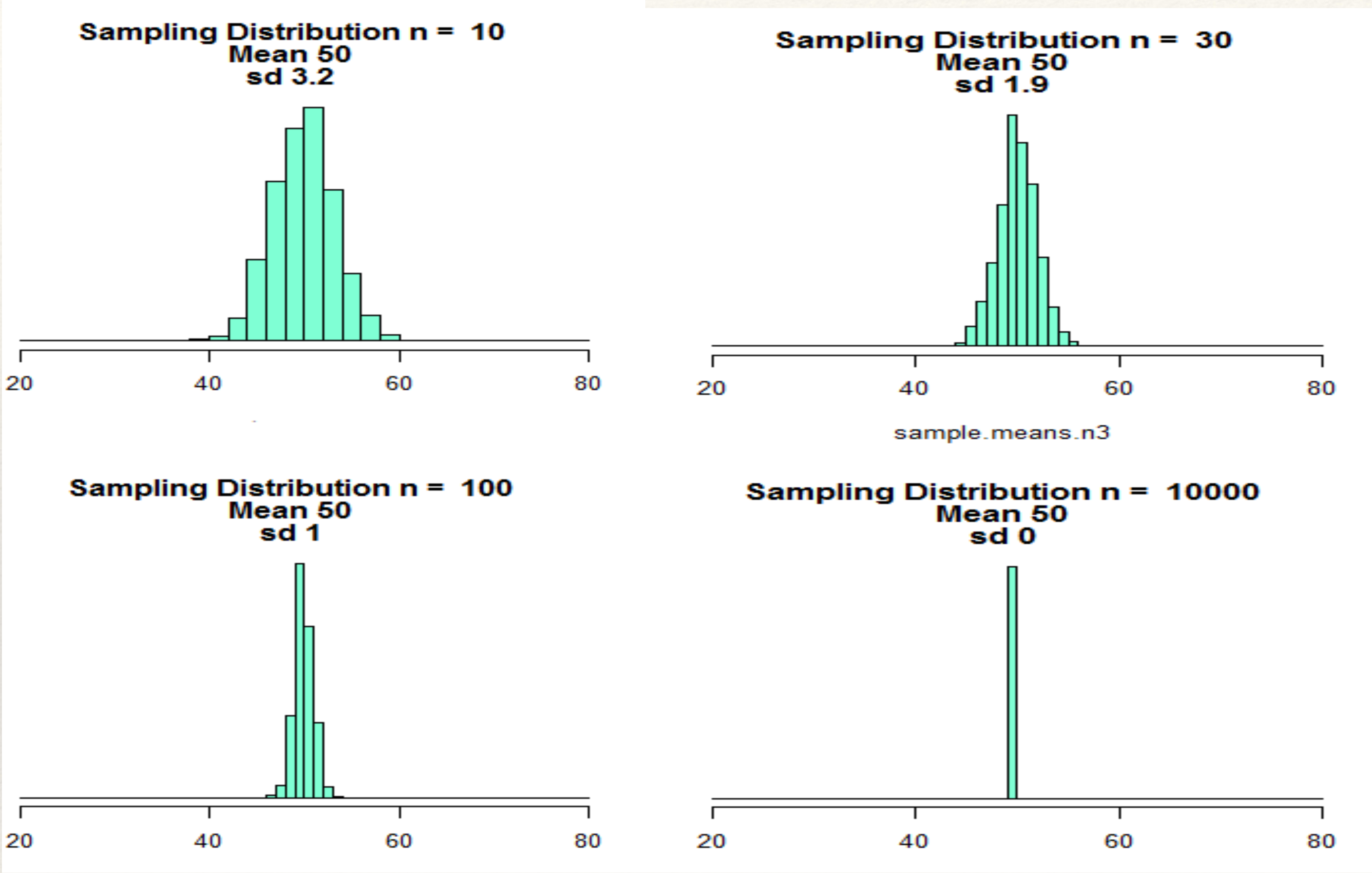


The variability of the sampling distribution of $\hat{\theta}$ is referred to as the Standard error, denoted by $SE_{\hat{\theta}}$ or $\sigma_{\hat{\theta}}$. The standard error describes the typical error or uncertainty of the statistic. It is the standard deviation of the statistic.

Sample Size & Sampling Variability

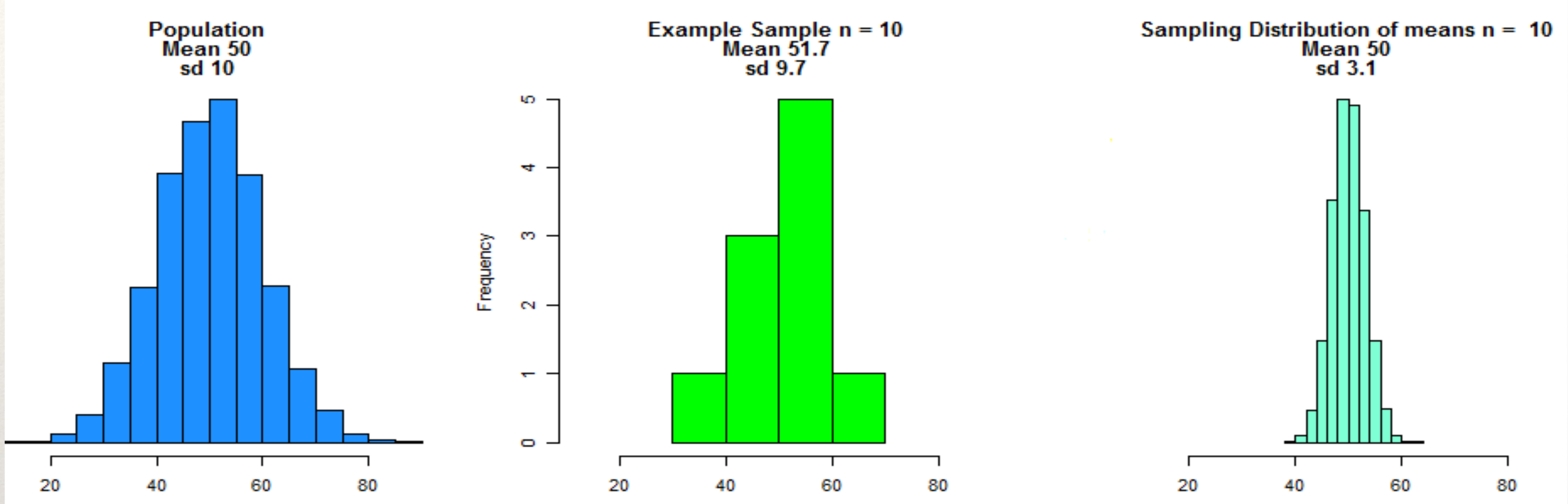
A random variable X is simulated from a normal distribution with population parameters μ_x and σ_x.





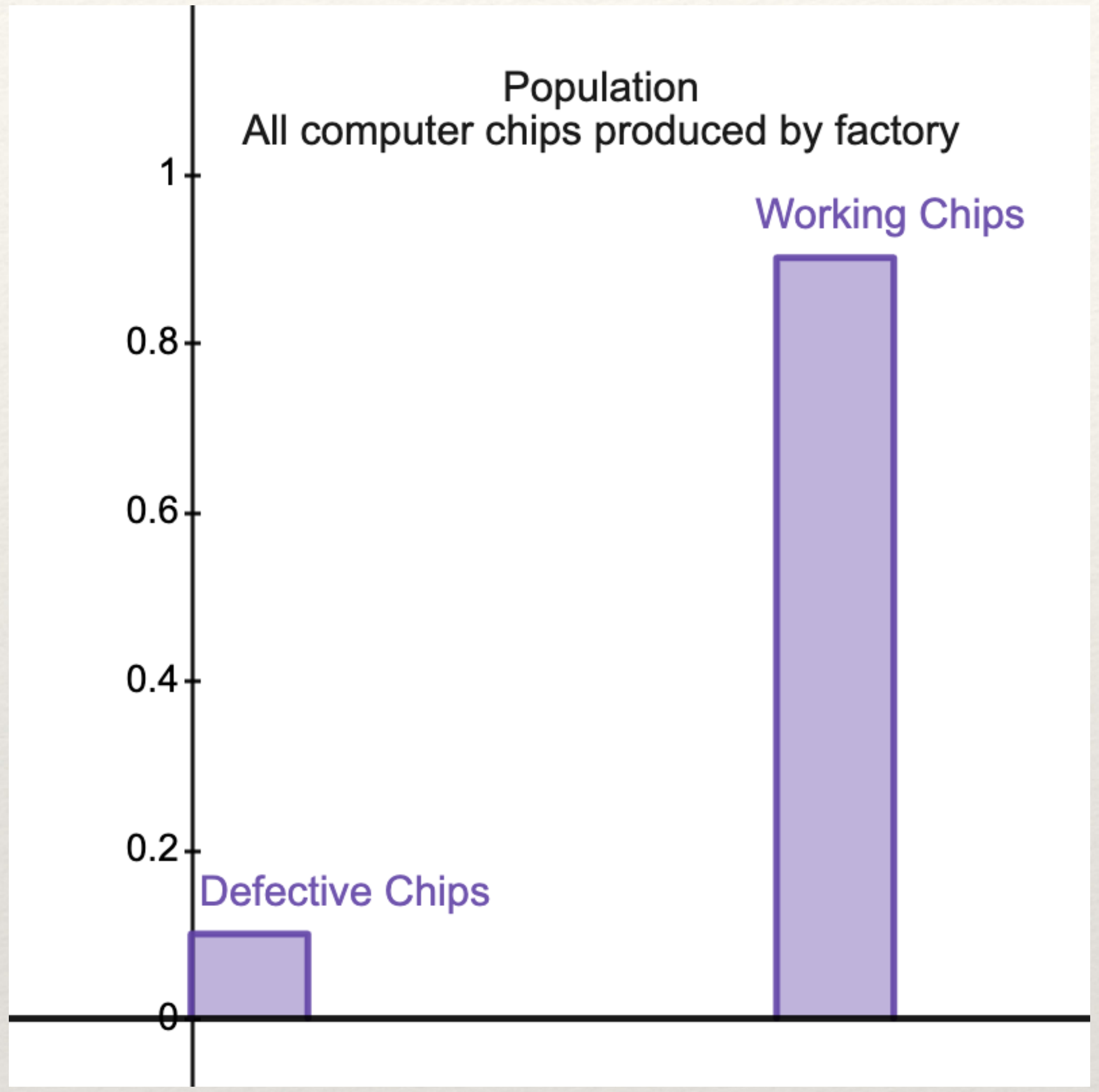
As n increases, the variability of the sampling distribution decreases.

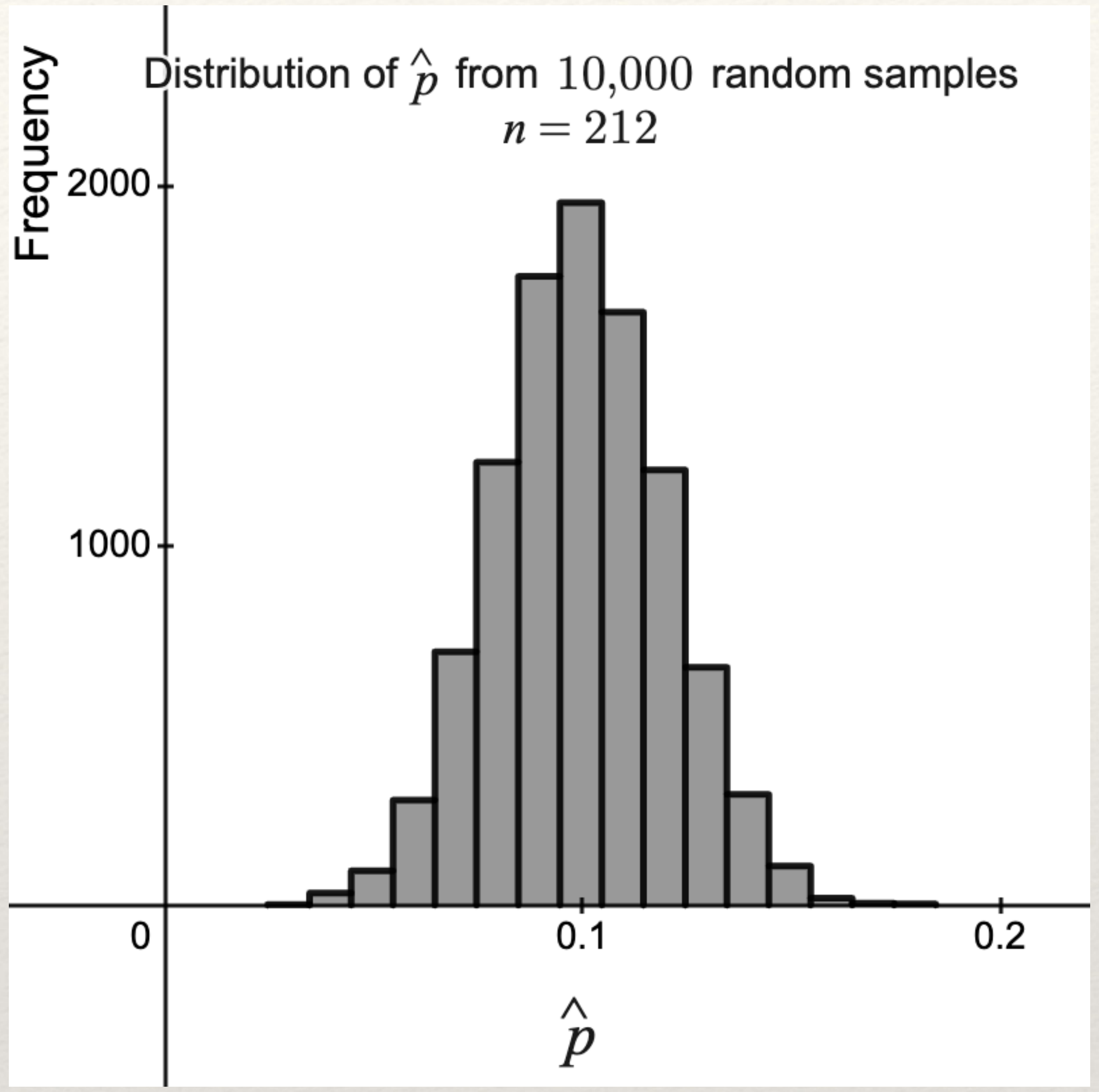
Distributions in inference



Population Distribution	Sampled Distribution	Sampling Distribution
Distribution of the entire collection of observations	Distribution of n obs. obtained from a single sample	Distribution of a sampled statistic from repeated samples of size n from population

Sampling Distribution of the Sample Proportion

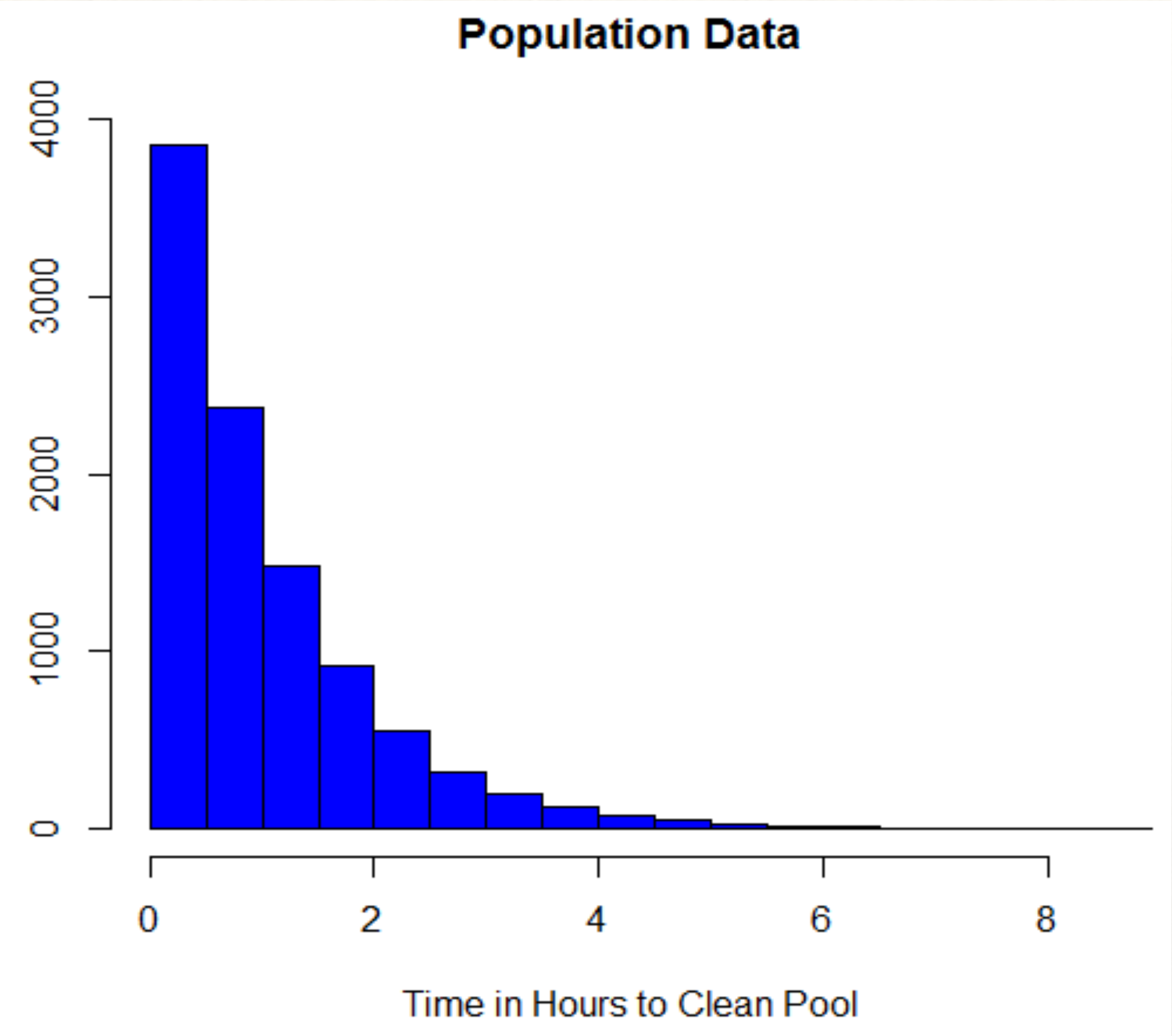


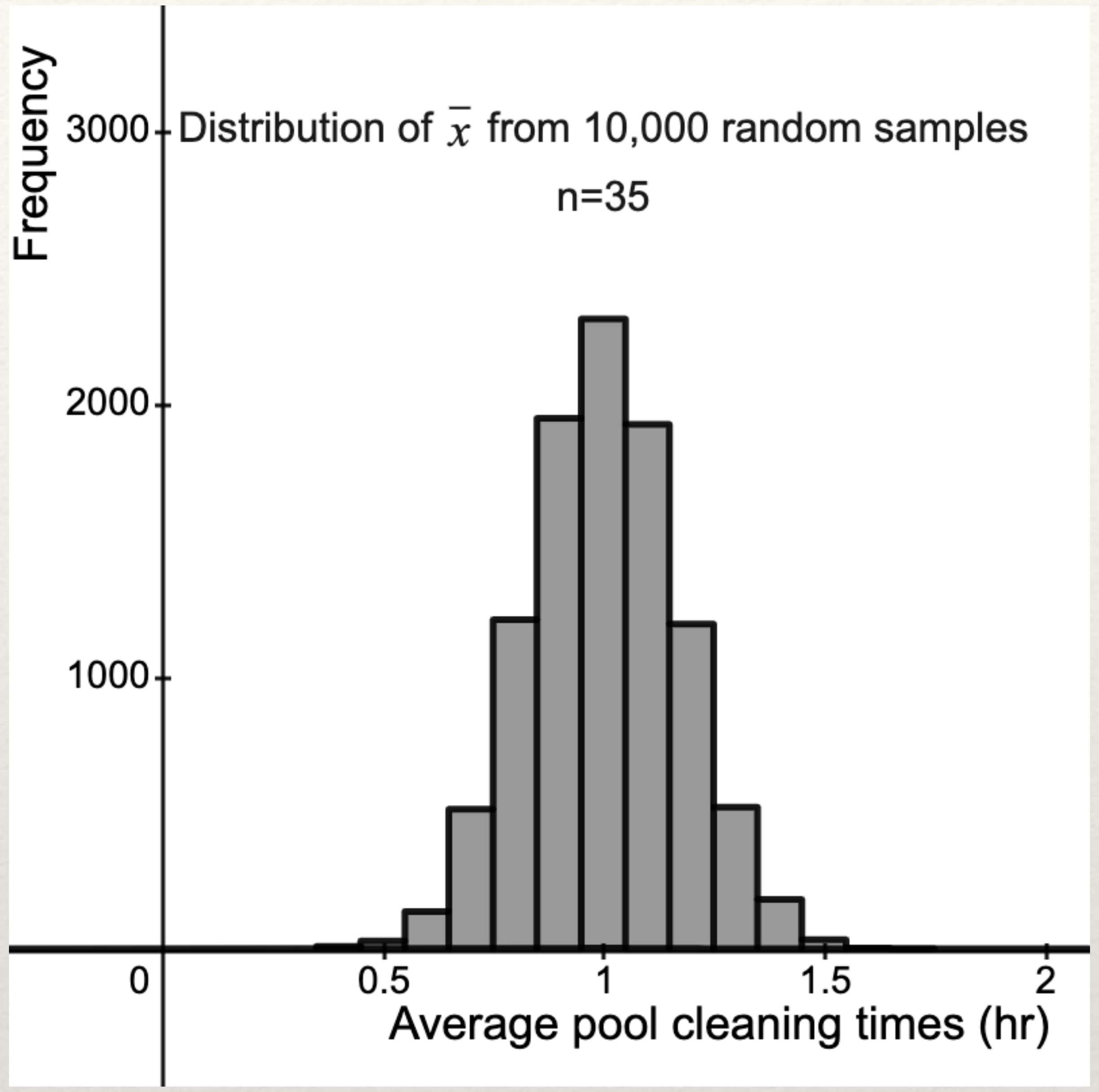


If n is sufficiently large, then the **Central Limit Theorem** states the sampling distribution of the statistic $\hat{p}$ is:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Sampling Distribution of the Sample Mean





If n is sufficiently large, then the **Central Limit Theorem** states the sampling distribution of the statistic $\bar{x}$ is:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

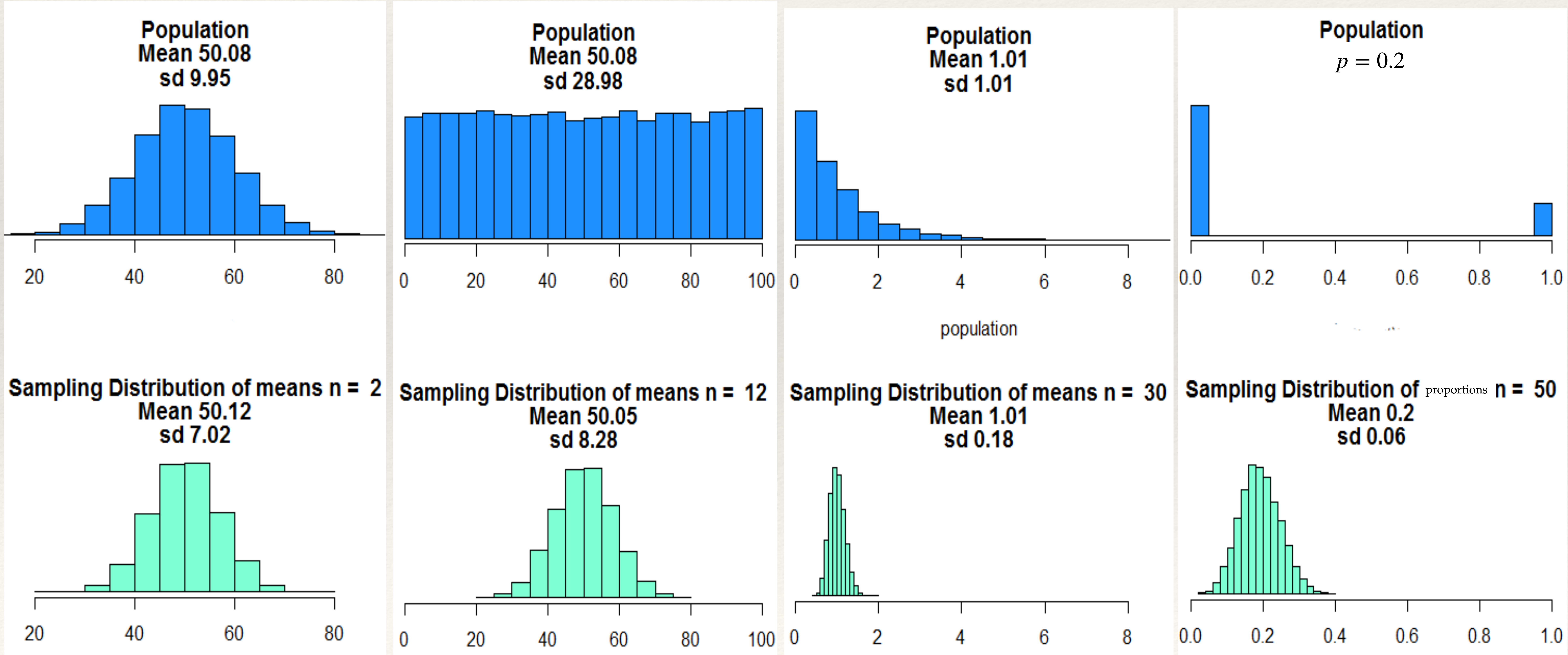
population mean

pop. standard dev.

Sufficiently Large Samples

$n\hat{p} \geq 10$
$n(1-\hat{p}) \geq 10$

Normal Data	Symmetric Data	Skewed Data	Bernoulli Data
$n \geq 1$	$n \geq 12$	$n \geq 30$	$n \geq \frac{10}{p}$ and $\frac{10}{1-p}$



CAUTION To apply CLT, sample size must be sufficiently large

Whether n is **sufficiently large enough** is determined by the population shape, excluding binary data, $n \geq 30$ is the rule of thumb.

Sampling Distributions

Sampling distributions are never observed, but we keep them in mind!