Statistical Decision Theory

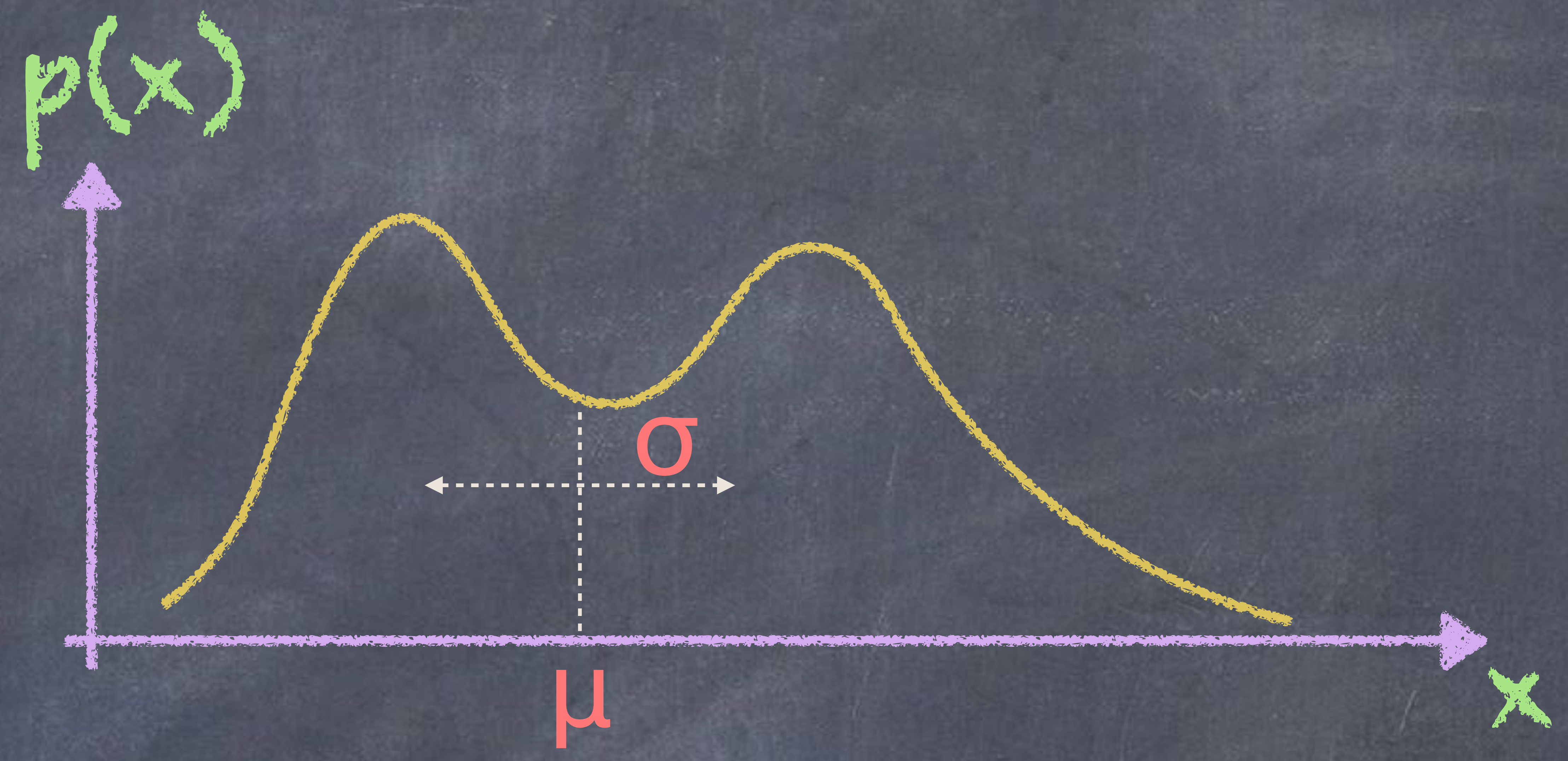
Sadeep Jayasumana

Some Terms

- Population: A set of similar items of interest. A large body of data, existing or conceptual
  - E.g. all undergraduate students in Sri Lanka, all possible throws of a dice
- Sample: A subset of the population chosen to represent the population
- Measurements in the sample are employed to make an inference about the characteristics of the population.
- Parameter <-> Population, Statistic <-> Sample

Sampling Distribution

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

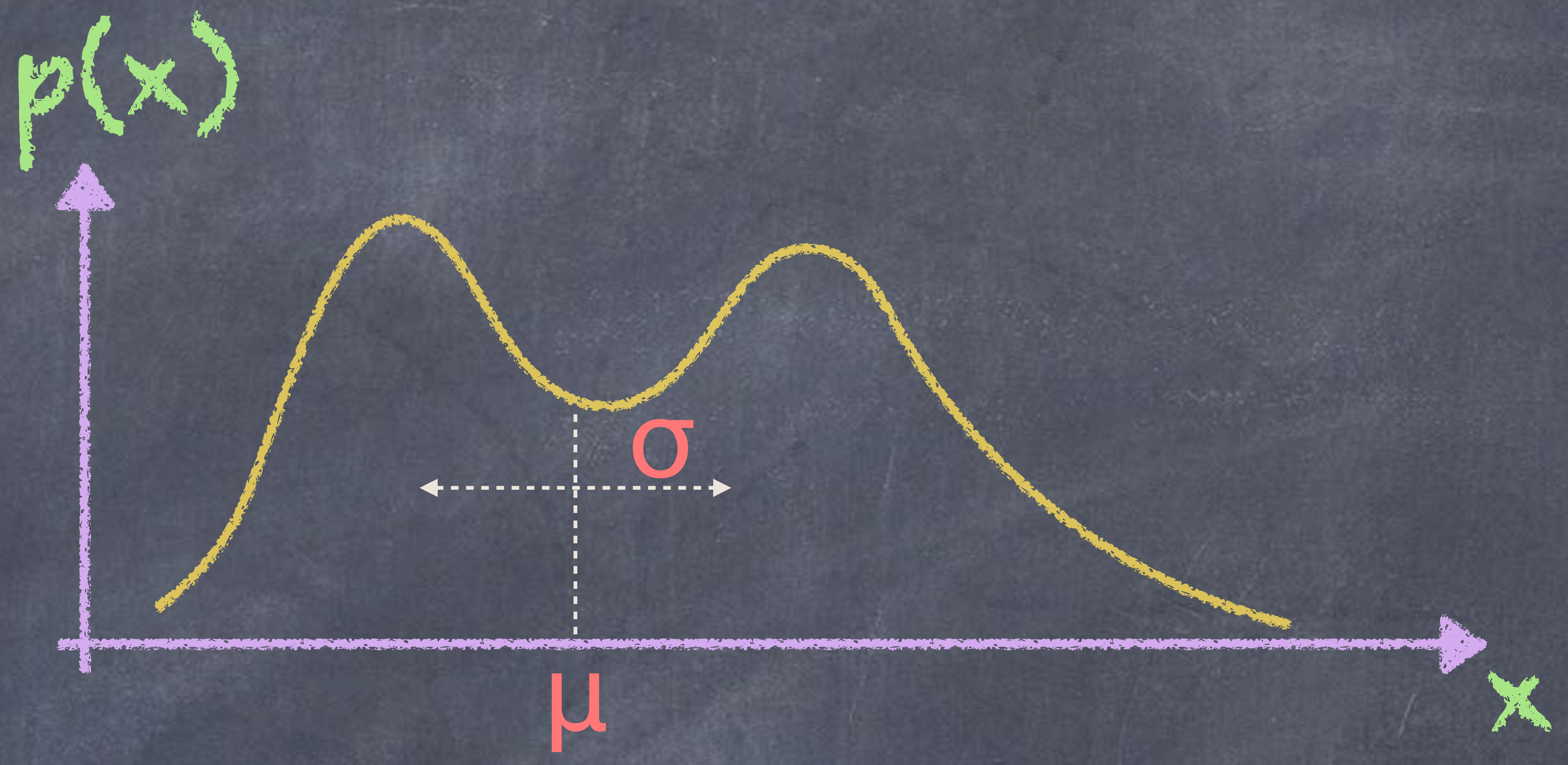


Sampling Distribution

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

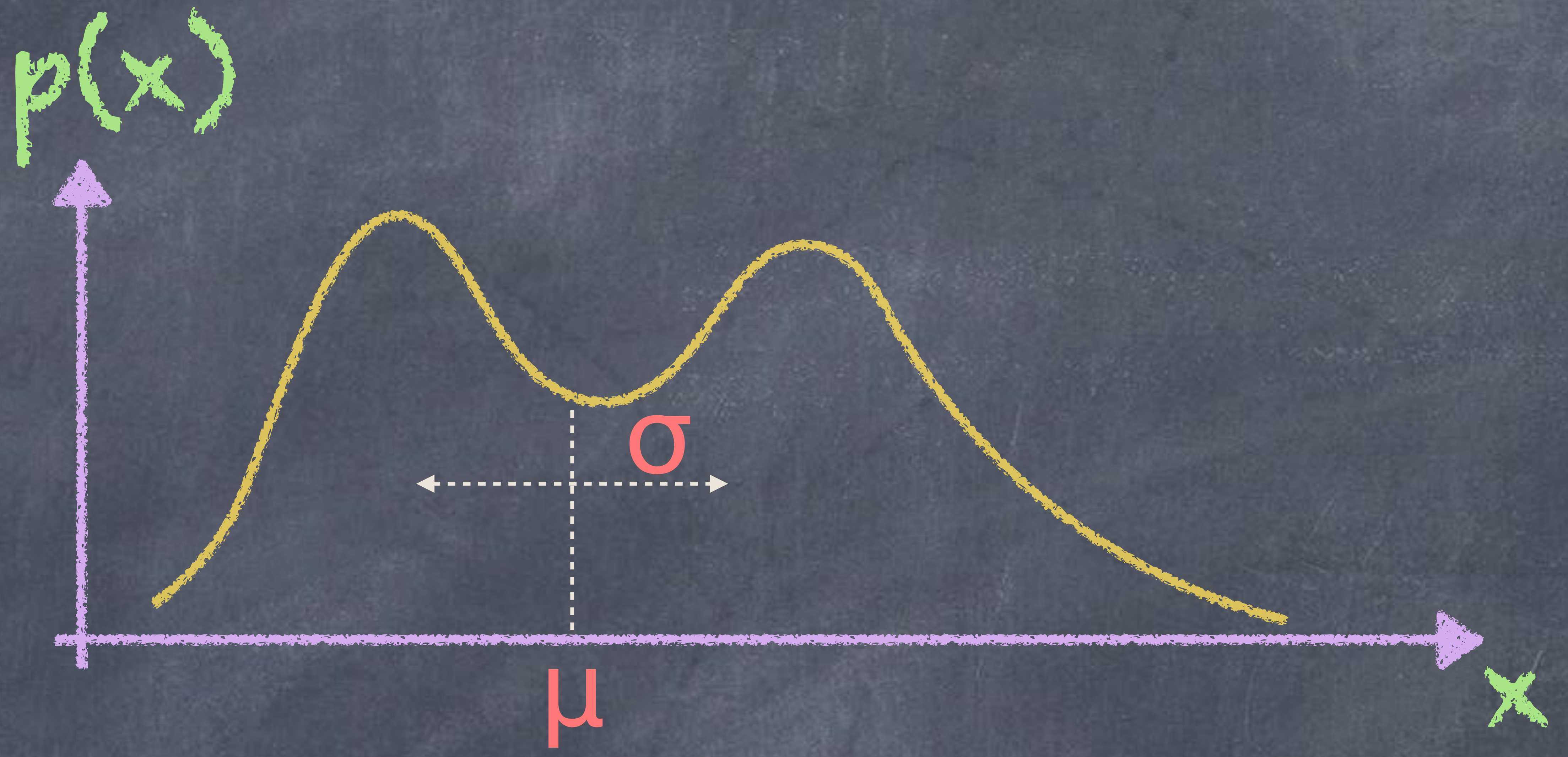
$$E[\bar{X}] = ?$$

$$V[\bar{X}] = ?$$



Sampling Distribution

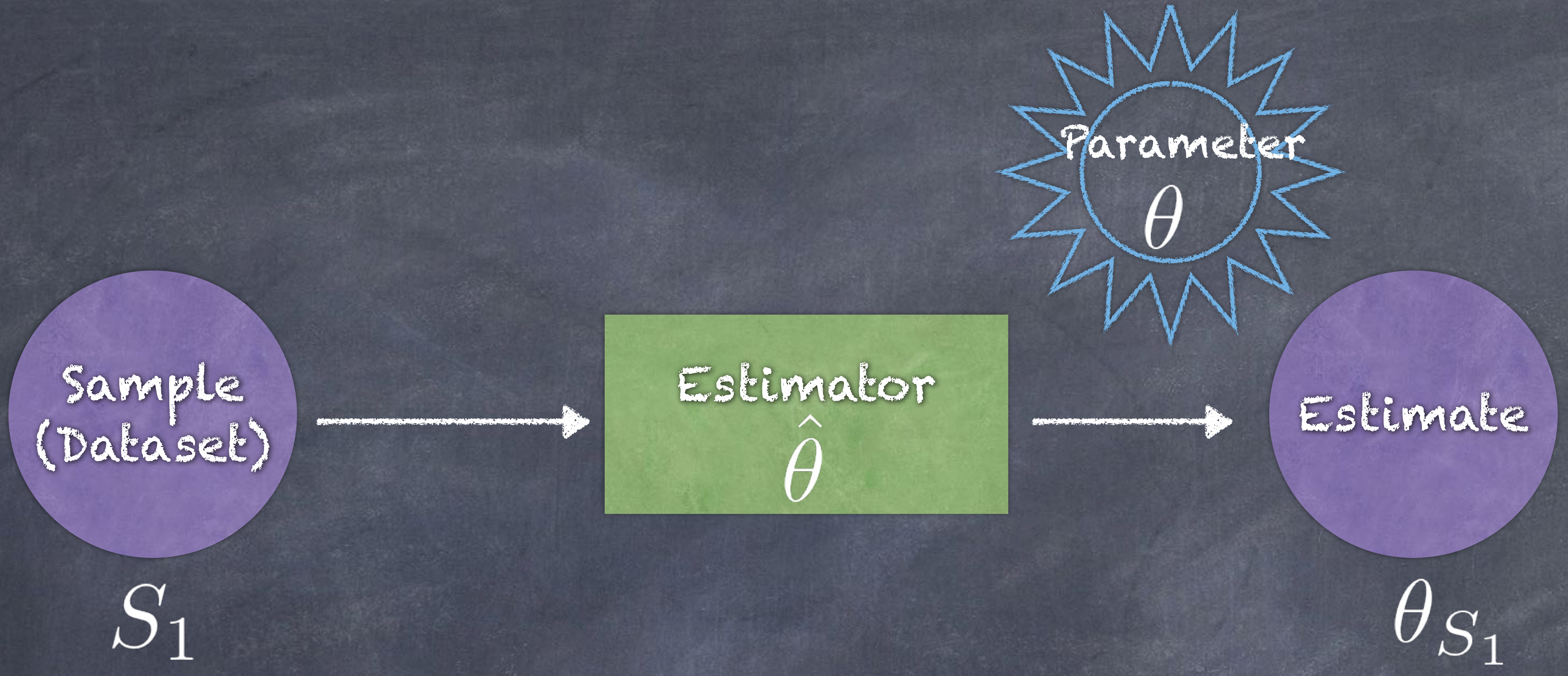
Population's distribution:



Sampling distribution:

Estimators

- Estimator - a rule (formula) that tells us how to calculate an estimate given the measurements contained in a sample.
- E.g. sample mean is an (point) estimator for the population mean



Example: Sample mean

{165, 163, 175, 170, 160} $\rightarrow$ $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\rightarrow$ 166.6

μ

Properties of Point Estimators

$\hat{\theta}$ Estimator

θ Parameter (the true value)

$E[\hat{\theta}]$ Average value of the estimator (across different random samples)

$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$ Bias of the estimator. How far off are we on average?

$V(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$ Variance of the estimator (across different random samples)

Mean Squared Error

- Average of the square of the distance between the estimator and its target parameter

$$\mathrm{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

- Prove that:

$$\mathrm{MSE}(\hat{\theta}) = [\mathrm{Bias}(\hat{\theta})]^2 + V(\hat{\theta})$$

Bias and Variance

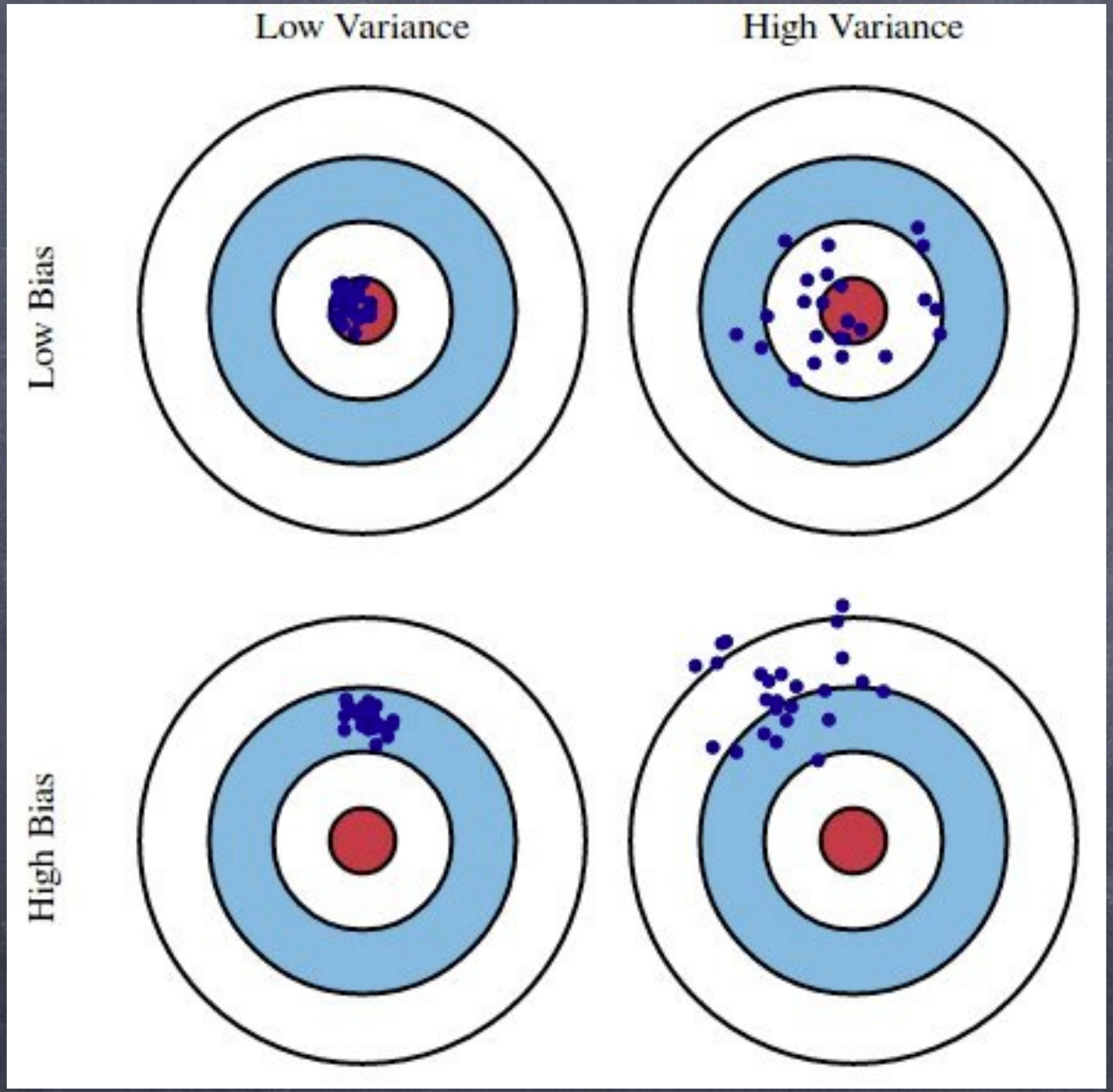


Image from: https://www.kdnuggets.com/2016/08/bias-variance-tradeoff-overview.html

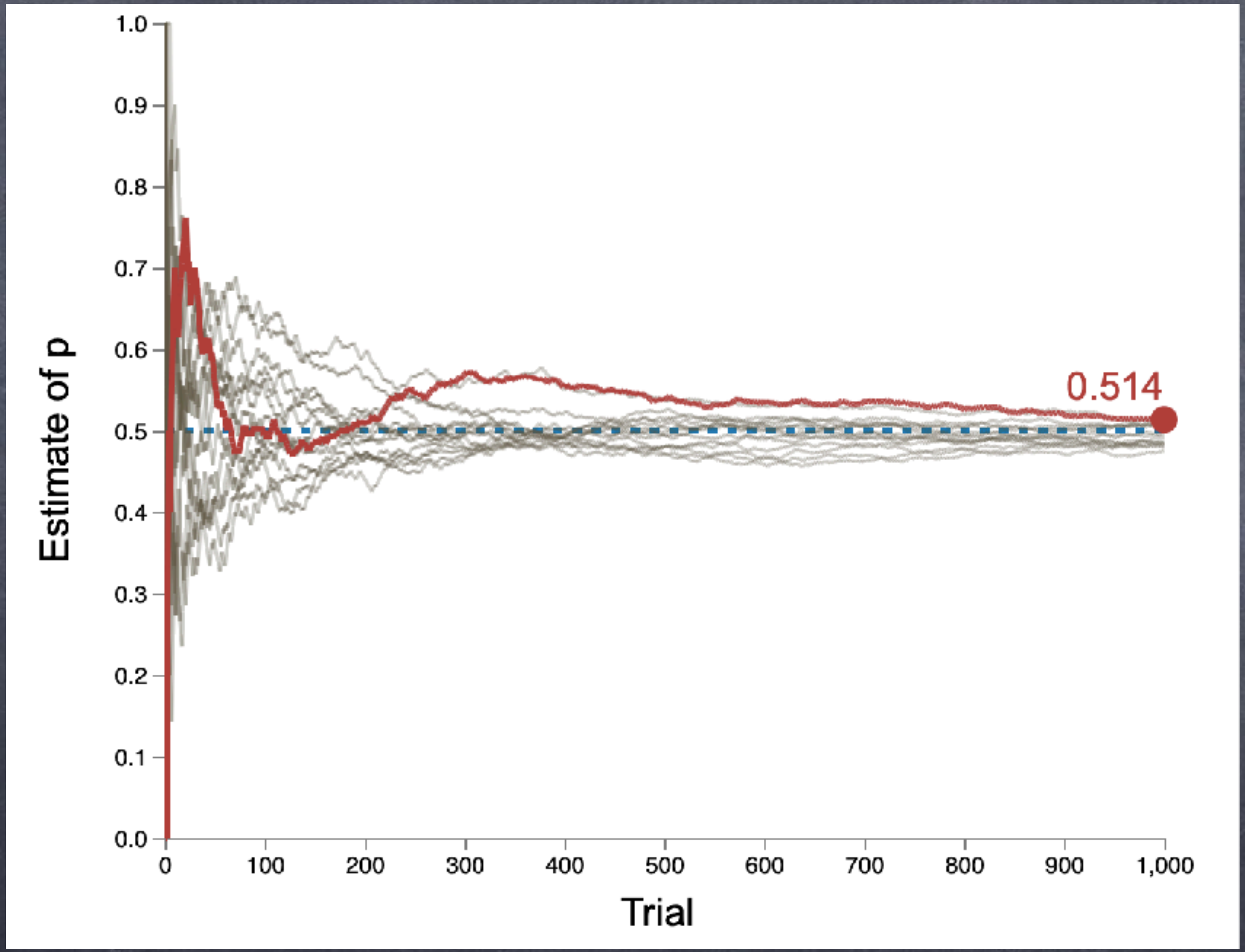
Desirable Properties

- Unbiased estimator - We're hitting the target on average
- Low-variance estimator - Results don't change much from sample to sample
- Consistent estimator - We're hitting the target asymptotically. (estimator converges in probability to the true value)

An estimator $\hat{\theta}_n$ is said to be consistent if, for any positive ϵ > 0 ,

$$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| \leq \epsilon) = 1.$$

Desirable Properties



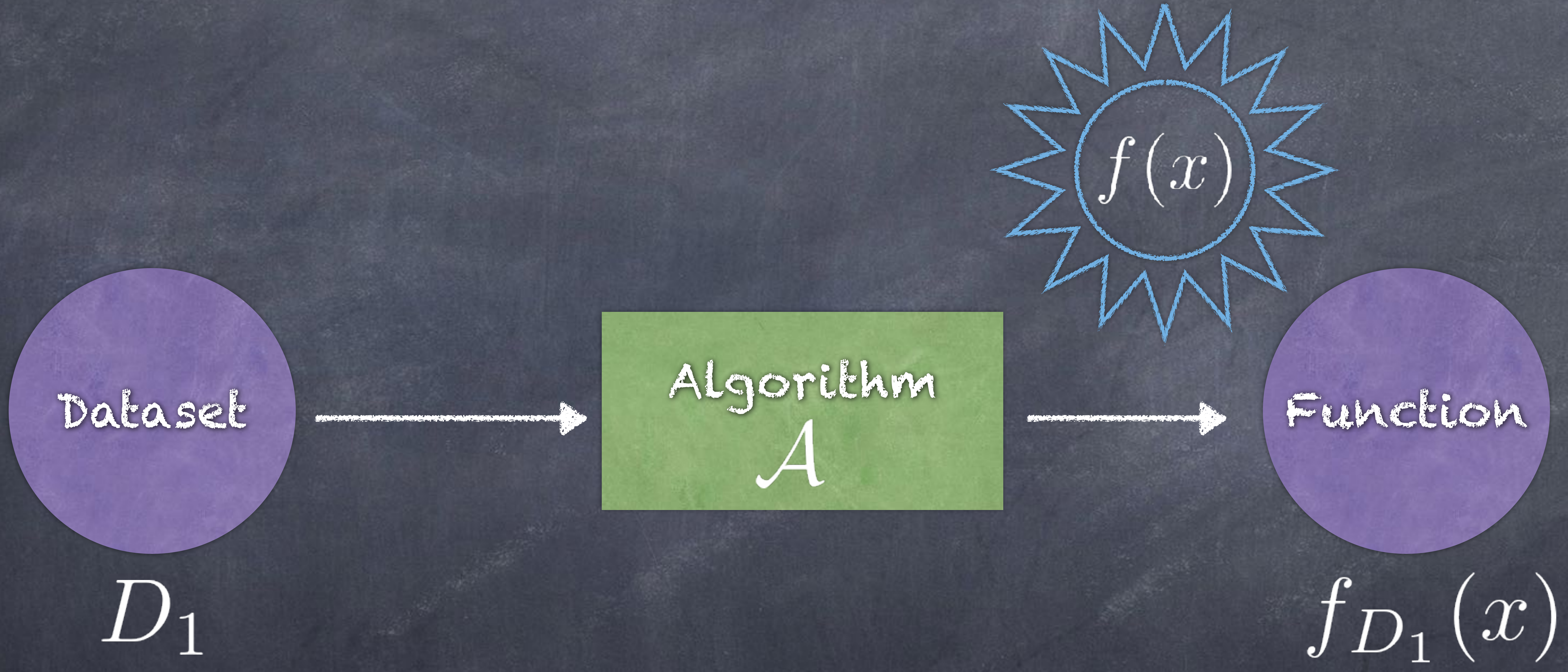
https://college.cengage.com/nextbook/statistics/wackerly_966371/student/html/09_point_estimation_long.html

Exercise

- Show that the sample mean is an unbiased estimator for the population mean.
- Let $S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Show that, $E[S'^2] = \frac{(n-1)}{n}\sigma^2$,

  where σ^2 is the population variance. Derive an unbiased estimator for the population variance.

Analyzing an ML Algorithm



Analyzing an ML Algorithm

$f_D(x)$ Hypothesis function learned with the dataset D

$\bar{f}(x)$ Average/expected hypothesis function, obtained by averaging $f_D(x)$ across all D.

$f(x)$ Bayes-optimal function. The same as E(Y|X=x). The average/expected y value for the given x value.

Bias-Variance Decomposition

$$E[(f_D(X) - Y)^2] = E[(\bar{f}(X) - f(X))^2] + E[(f_D(X) - \bar{f}(X))^2] + E[(f(X) - Y)^2]$$

Bias-Variance Decomposition

$$E[(f_D(X) - Y)^2] = \underbrace{E[(\bar{f} - f)^2]}_{\text{Bias}^2} + \underbrace{E[(f_D - \bar{f})^2]}_{\text{Variance}} + \underbrace{E[(f(X) - Y)^2]}_{\text{Noise}^2}$$

The Bayes (Optimal) Classifier

$$g^*(x) = \text{argmax}_k \Pr(Y = k|X = x)$$

- Theoretically, the best classifier we can have
- But usually we cant get it since we dont know the real P(Y|X)
- Bayes (error) rate = Error rate of the Bayes classifier

The Bayes (Optimal) Classifier

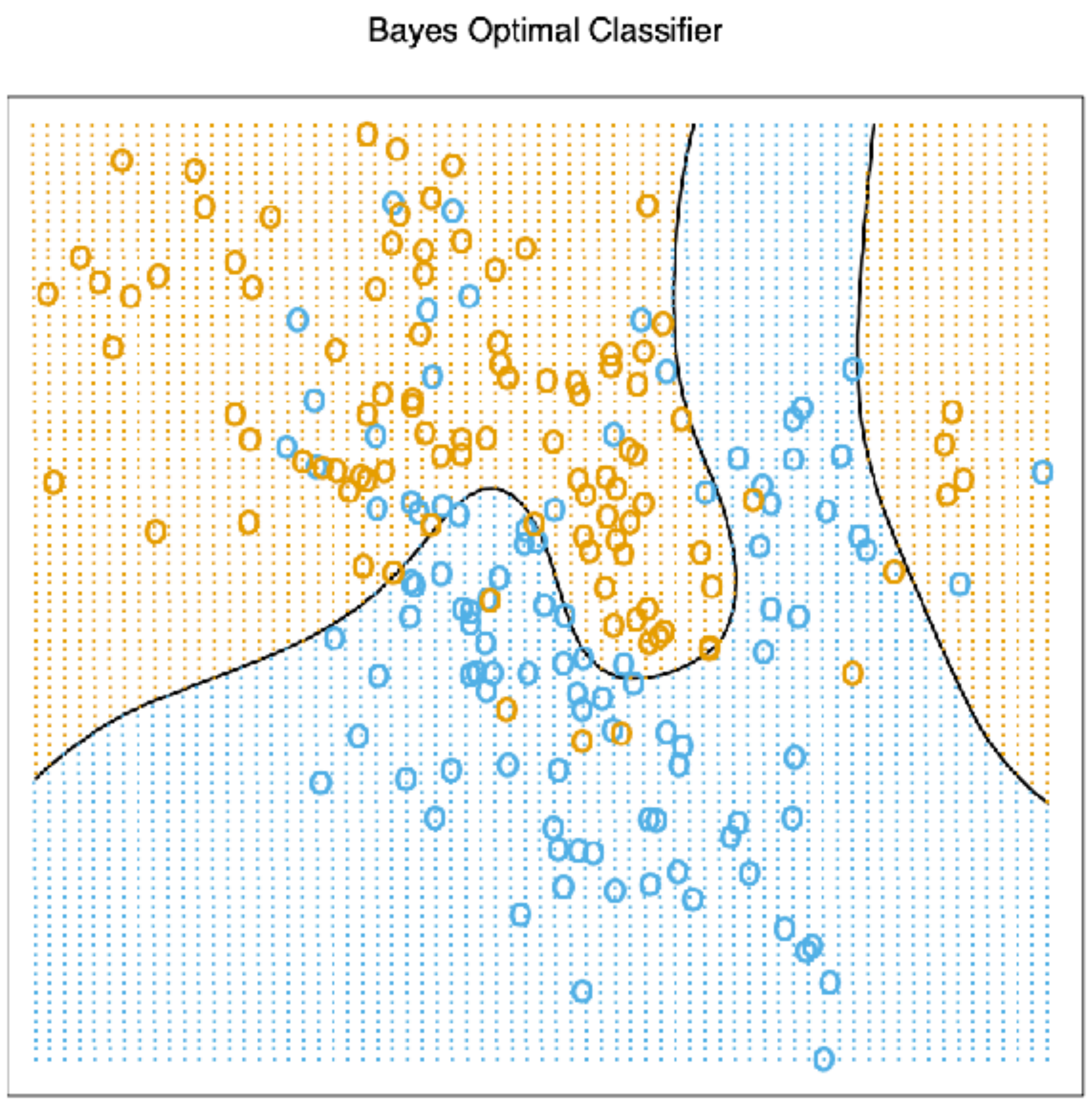


FIGURE 2.5. *The optimal Bayes decision boundary for the simulation example of Figures 2.1, 2.2 and 2.3. Since the generating density is known for each class, this boundary can be calculated exactly (Exercise 2.2).*

Source: Elements of Statistical Learning

Empirical Risk Minimization (ERM)

- Supervised learning - find a hypothesis $\tilde{f}$ from a class of functions $\mathcal{F}$ such that the risk is minimized:

$$\tilde{f} = \text{argmin}_{f \in \mathcal{F}} R(f); \qquad R(f) = E[L(Y, f(X))]$$

- But we can't know the true risk, so we minimize an empirical estimate of it:

$$\tilde{f} = \text{argmin}_{f \in \mathcal{F}} \hat{R}(f); \qquad \hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$

Thank you!