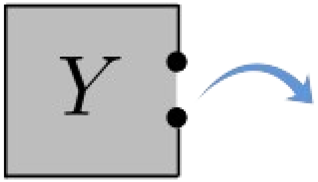
Statistics and Data Science for Engineers
E178 / ME276DS

Statistical inference:
Point estimation

Statistical inference

No inputs



$\{y_i\}_N = \mathcal{D}$.

Inference: Statement based on data.

Assumption: Sampling is iid.

- Y does not change btwn samples
- Samples are independent.

Given $\mathcal{D}$.

Three types of inferences

1) Point estimation : "My best guess for some parameter θ of P_X is $\hat{\theta}_N$"

2) Confidence intervals : "Parameter θ lies in the interval I with confidence γ".

3) Hypothesis tests :

null hypothesis

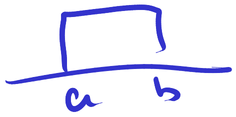
H_0 : ______

H_1 : ______

"H_0 is rejected in favor of H_1"

"H_0 is not rejected in favor of H_1"

Point estimation



estimate

$\boxed{\mathcal{U}(a,b)}$

Given $\mathcal{D} = \{y_i\}_N \overset{\text{iid}}{\sim} Y$, find a best estimate $\hat{\theta}_N$ of a property or parameter θ

data points.

$E[Y]$

$\mathrm{Var}[Y]$

a in $\mathcal{U}(a,b)$

Estimator $\qquad \hat{\theta}_N = g_N(y_1, \ldots, y_N)$

$$\{y_i\}_N \xrightarrow{g_N} \hat{\theta}_N \approx \theta$$

Is g_N a good estimator?

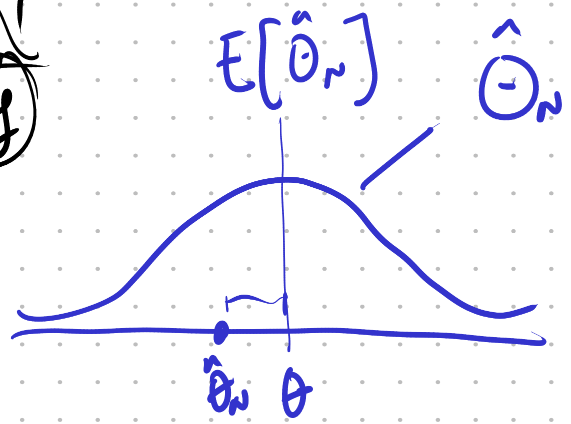
Means that "expect $\hat{\theta}_N \approx \theta$"

Make an assumption about Y

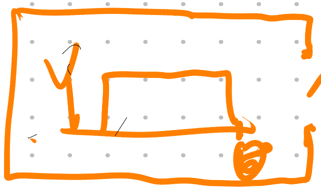
$\{Y_i\}_N \overset{iid}{\sim} Y$

"Estimator"

$\{Y_i\}_N \xrightarrow{g_N} \hat{\Theta}_N$



$\hat{\Theta}_N$



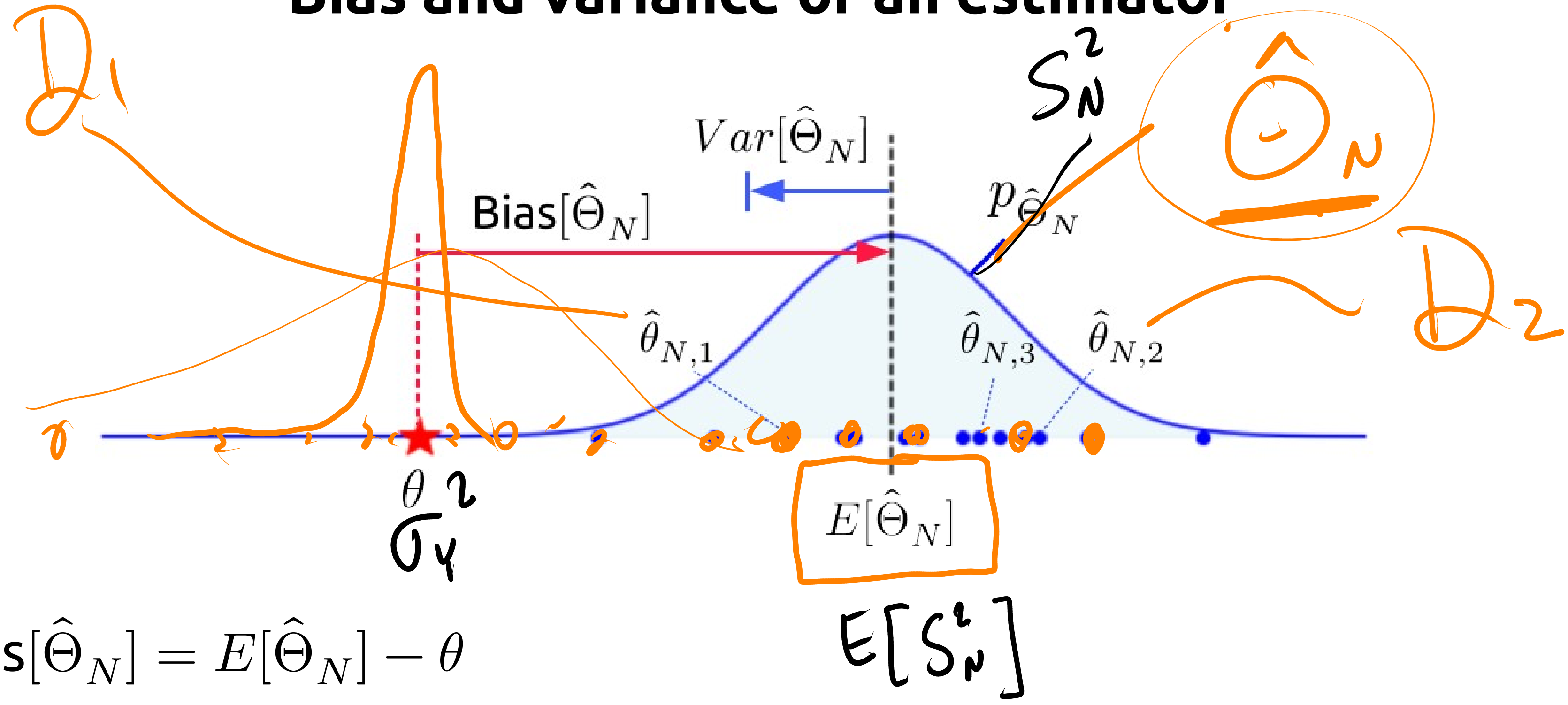
$\rightarrow \{y_1, \dots, y_{100}\} = D.$

θ... size of largest apple in the orchard

$g_N^1: \max(y_1, \dots, y_{100}) = \hat{\Theta}_N$

$g_N^2: \max(y_1, \dots, y_{100}) + 10\text{ cm.}$

Bias and variance of an estimator



- $\text{Bias}[\hat{\Theta}_N] = E[\hat{\Theta}_N] - \theta$

- $\text{Var}[\hat{\Theta}_N] = E\left[\left(\hat{\Theta}_N - E[\hat{\Theta}_N]\right)^2\right]$

Estimating the mean

$Y \sim \mu_Y$ μ_Y

The *sample mean*:

$\hat{\Theta}_N$

$$\hat{\mu}_N = g_N(y_1, \ldots, y_N) = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$\hat{\Theta}_N$

$$\bar{Y}_N = g_N(Y_1, \ldots, Y_N) = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$E[\bar{Y}_N] = E\left[\frac{1}{N}\sum_{i=1}^{N} Y_i\right] = \frac{1}{N}\sum E[Y_i]$$

$$= \frac{1}{N}\sum E[Y] = \frac{1}{N} N\, E[Y] = E[Y] = \mu_Y$$

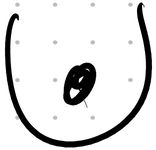
Estimating the mean

Proof

- Bias $\left[\bar{Y}_N\right] = 0$
- $Var\left[\bar{Y}_N\right] = \frac{\sigma_Y^2}{N}$

Prove: $Var\left[\bar{Y}_N\right] = Var\left[\frac{1}{N}\sum Y_i\right]$.

$$= \left(\frac{1}{N}\right)^2 \sum \underbrace{Var\left[Y_i\right]}_{\sigma_Y^2} = \frac{N\sigma_Y^2}{N^2} = \frac{\sigma_Y^2}{N}.$$

$N = 1$  $\bar{Y}_N = Y$ $\quad Var[\bar{Y}_N] = \sigma_Y^2$

$N = 10$  $Var[\bar{Y}_N] = \sigma_Y^2 / 10$

Var : σ_Y^2 / N

Std : $\sigma_Y / \sqrt{N}$ ✓

Estimating the variance

σ_Y^2

"Unbiased" sample variance:

$\hat{\theta}_N$ ... $\hat{\Theta}_N$

$$\hat{\sigma}_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{\mu}_N)^2 \qquad \boxed{S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2}$$

$$\Uparrow$$

$$\mathrm{Var}[Y] \approx \frac{1}{N}\sum\left[(y - \hat{\mu}_N)^2\right].$$

Estimating the variance

- Bias $[S_N^2] = 0$

Proof in the reader.

- $Var[S_N^2] =$ complicated.

Y is Gaussian.

$\Downarrow$

S_N^2 ... χ^2 distribution

Biased sample variance: $$\tilde{S}_N^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2$$

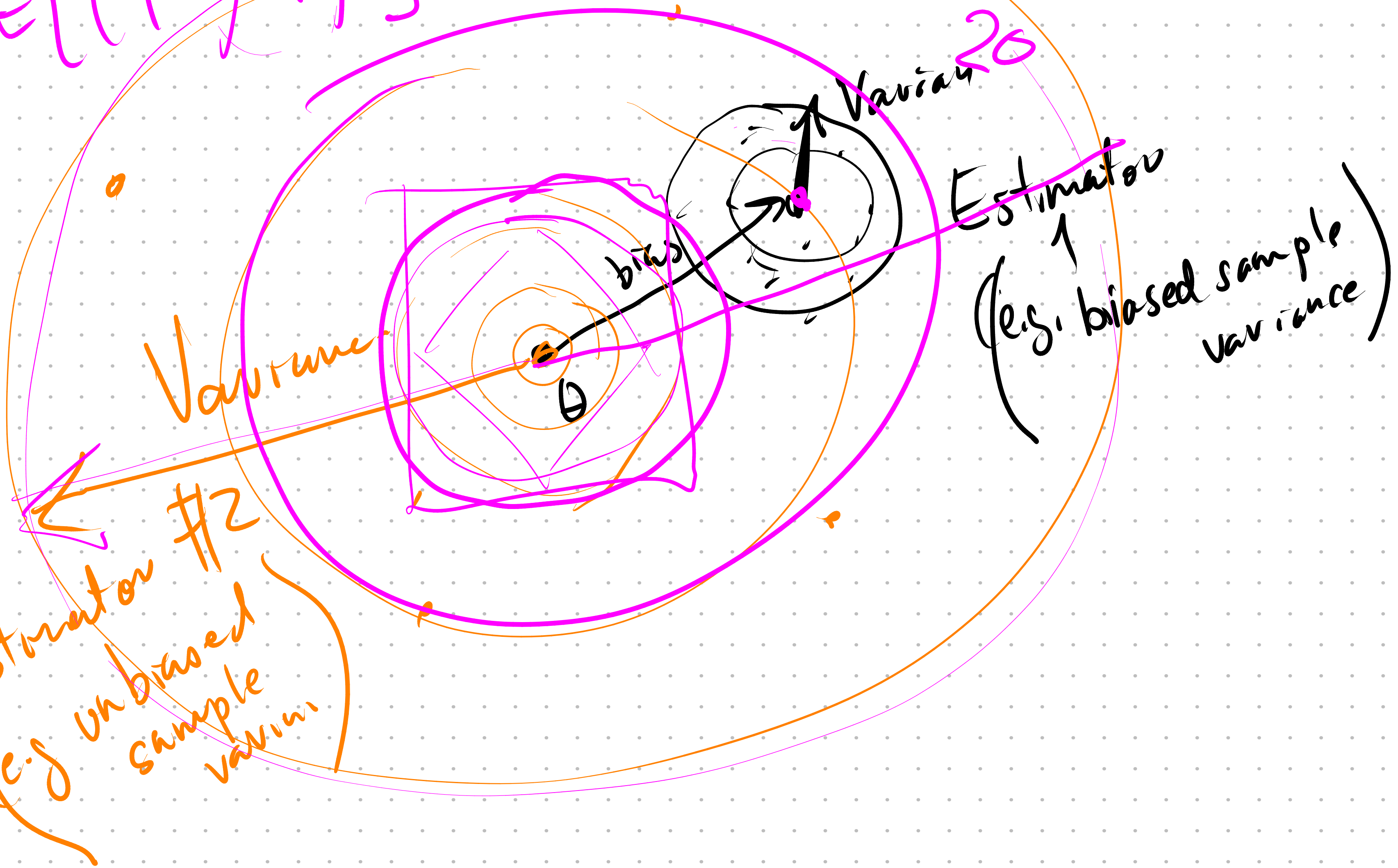
- Bias$[\tilde{S}_N^2] =$ $E[\tilde{S}_N^2] - \sigma_Y^2 = \frac{N-1}{N}\sigma_Y^2 - \sigma_Y^2 = -\frac{1}{N}\sigma_Y^2$

$\tilde{S}_N^2 = \frac{N-1}{N} S_N^2 \longrightarrow E[\tilde{S}_N^2] = \frac{N-1}{N} E[S_N^2] = \frac{N-1}{N} \cdot \sigma_Y^2$

- $Var[\tilde{S}_N^2] =$ Complicated.

  smaller than $Var[S_N^2]$.

$$\mathrm{Var} = E\left[(Y-\mu_Y)^2\right].$$

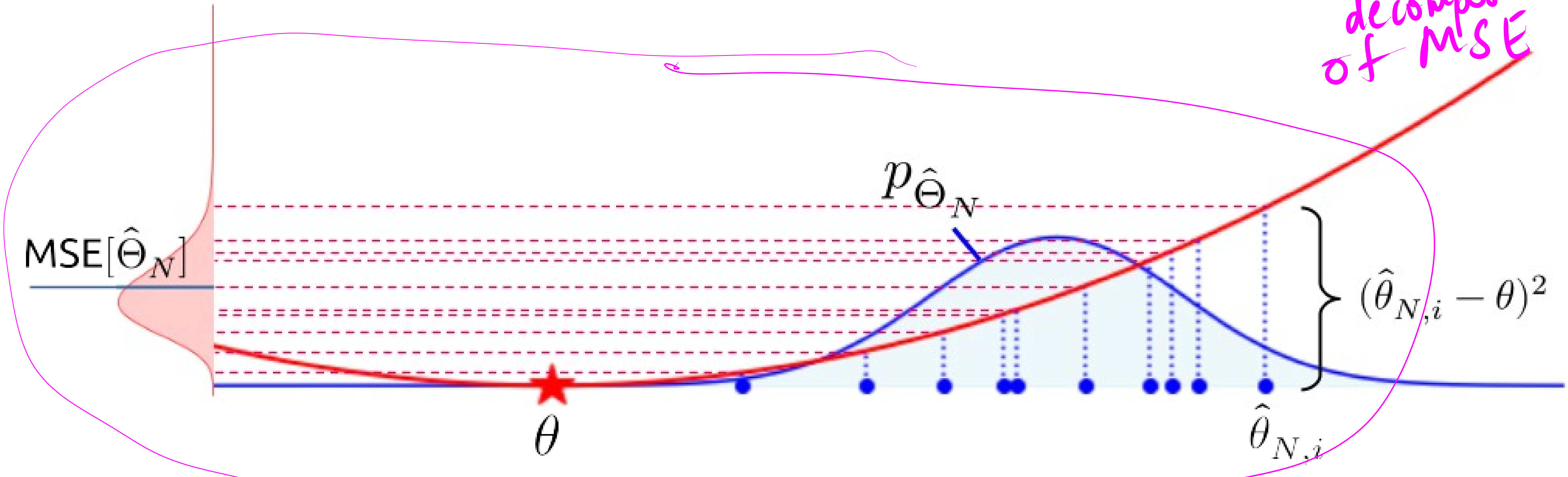


Mean squared error (MSE)

$Var[\hat{\Theta}_N] = E[(\hat{\Theta}_N - E[\hat{\Theta}_N])^2]$

$$\text{MSE}[\hat{\Theta}_N] = \boxed{E[(\hat{\Theta}_N - \theta)^2]}$$
$$= Var[\hat{\Theta}_N] + \left(\text{Bias}[\hat{\Theta}_N]\right)^2$$

prove this Bias/Variance decomposition of MSE



SAMPLE

$\text{MSE}[\bar{Y}_N] = \text{Var}[\bar{Y}_N] + (\text{Bias}[\bar{Y}_N])^2$

$\sigma_Y^2/N + 0 = \sigma_N^2/N.$

$\text{MSE}[S_N^2]$

$\text{MSE}[\tilde{S}_N^2]$

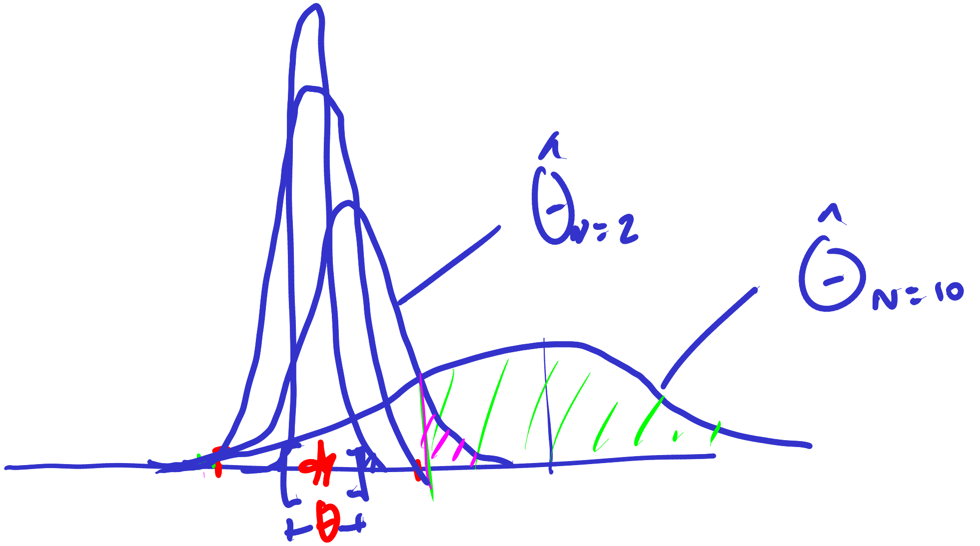
Asymptotic properties

- Asymptotic unbiasedness: $$\lim_{N\to\infty} \text{Bias}[\hat{\Theta}_N] = 0$$

- Consistency: $$\lim_{N\to\infty} P\left(|\hat{\Theta}_N - \theta| \geq \epsilon\right) = 0$$

$\varepsilon = 0.1$

$\varepsilon = 0.01$



$MSE = Var + Bias^2.$

As $N \to \infty$

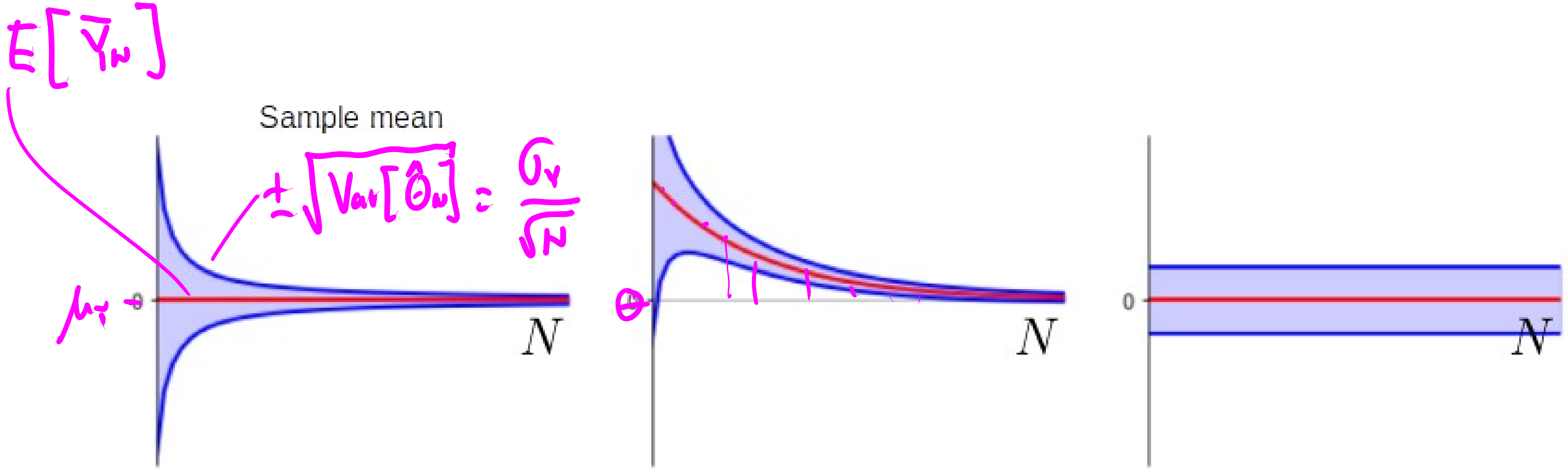
$$\begin{matrix} Var \to 0 \\ Bias \to 0 \end{matrix} \iff MSE \to 0.$$

$$\begin{matrix} \text{if } MSE \to 0 \\ \text{as } N \to \infty \end{matrix} \implies \text{consistent.}$$

MSE is "stronger" than consistent

$$\underline{\text{consistent}} \not\Longrightarrow \begin{matrix} MSE \to 0 \\ \text{as } N \to \infty. \end{matrix}$$

"usually" it does.



Unbiased,
Consistent

Biased ,
Asymptoticall unbiased,
Consistent

Unbiased,
Inconsistent

Maximum Likelihood Estimation (MLE)

$$\hat{\theta}_N = g_N(\mathcal{D})$$ … general point estimation

- MLE:

  1) Pick a parametrization.

  $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, $Y \sim \mathcal{B}(p)$, ...

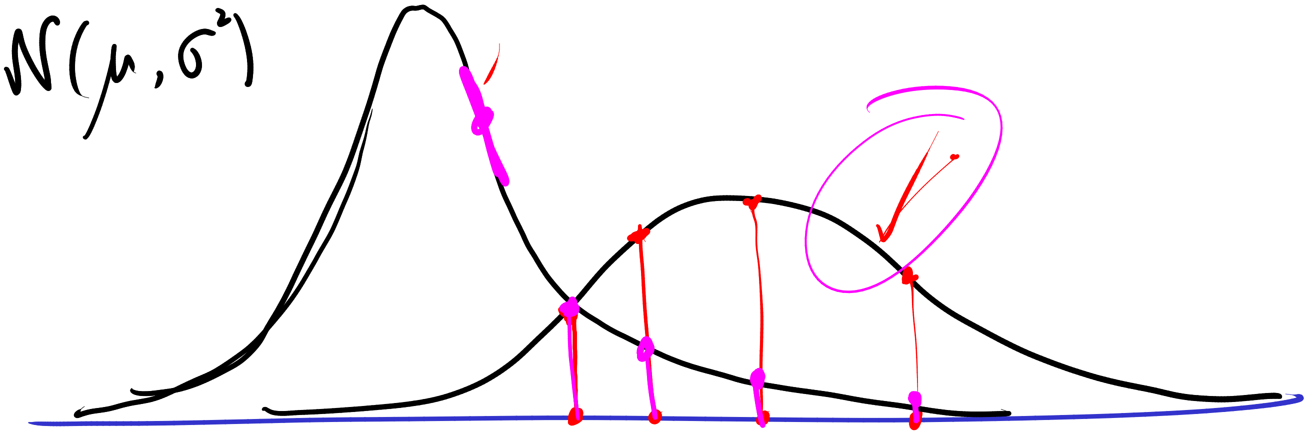
  $\underline{\theta} = (\mu_Y, \sigma_Y^2)$. $\underline{\theta} = p$.

  2) Solve: $$\underline{\hat{\theta}}_{\text{MLE}} = \underset{\underline{\theta}}{\text{argmax}}\ \underline{\mathcal{L}(\underline{\theta}; \mathcal{D})}$$

  Likelihood: Probability of obtaining the dataset $\mathcal{D}$ if parameters are $\underline{\theta}$

- Likelihood: $$\mathcal{L}(\underline{\theta}\,; \mathcal{D}) = \prod_{i=1}^{N} p_Y(y_i; \underline{\theta})$$

$Y \sim N(\mu, \sigma^2)$



Example

- 4 marbles in a bag, all either black or white
- pick 5 times with replacement

black white

$D = \{1, 0, 1, 0, 0\}$

Estimate the number of black marbles in the bag. θ

$\theta \ldots \{0, 1, 2, 3, 4\}$.

$p = \theta/4 = \{0, 1/4, 1/2, 3/4, 1\}$.

1. Pick a family: $Y \sim B(p)$

2. Solve: $\underset{\theta}{\text{maximize}} \prod_{i=1}^{5} p_Y(\theta; y_i)$

$$p_Y(\theta; y_i) = p^{y_i}(1-p)^{1-y_i} = \begin{cases} p = \theta/4 & y_i = 1 \\ 1-p = 1-\frac{\theta}{4} & y_i = 0. \end{cases}$$

$$\underset{\theta}{\text{maximize}} \prod_{i=1}^{5} p_Y(\theta; y_i) = \underset{\theta}{\text{maximize}} \overbrace{\left(\frac{\theta}{4}\right)^2\left(1-\frac{\theta}{4}\right)^3}^{J(\theta)}.$$

θ	0	1	2	3	4
$J(\theta)$	0	$27/4^5$	$32/4^5$	$9/4^5$	0

$$\frac{1}{4^2}\cdot\frac{3^3}{4^3} \qquad \frac{2^2\,2^3}{4^5} \qquad \frac{3^2\cdot 1^3}{4^5}$$

Log-likelihood

$$
\begin{aligned}
\underline{\hat{\theta}}_{\text{MLE}} &= \underset{\underline{\theta}}{\text{argmax}}\, \mathcal{L}(\underline{\theta}; \mathcal{D}) \\
&= \underset{\underline{\theta}}{\text{argmax}}\; \ln \mathcal{L}(\underline{\theta}; \mathcal{D}) \\
&= \underset{\underline{\theta}}{\text{argmax}}\; \ln \left(\prod_{i=1}^{N} p_Y(y_i; \underline{\theta}) \right) \\
&= \underset{\underline{\theta}}{\text{argmax}}\; \sum_{i=1}^{N} \ln p_Y(y_i; \underline{\theta})
\end{aligned}
$$

Example: Gaussian data

Assume: $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \quad , \quad \underline{\theta} = (\mu_Y, \sigma_Y^2)$

$$(\hat{\mu}_{\mathsf{MLE}}, \hat{\sigma}^2_{\mathsf{MLE}}) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \ln p_Y(y_i; \mu, \sigma^2)$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2} \right) \right)$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left(-\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right)$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmin}} \left(\frac{N}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right)$$

Example: Gaussian data

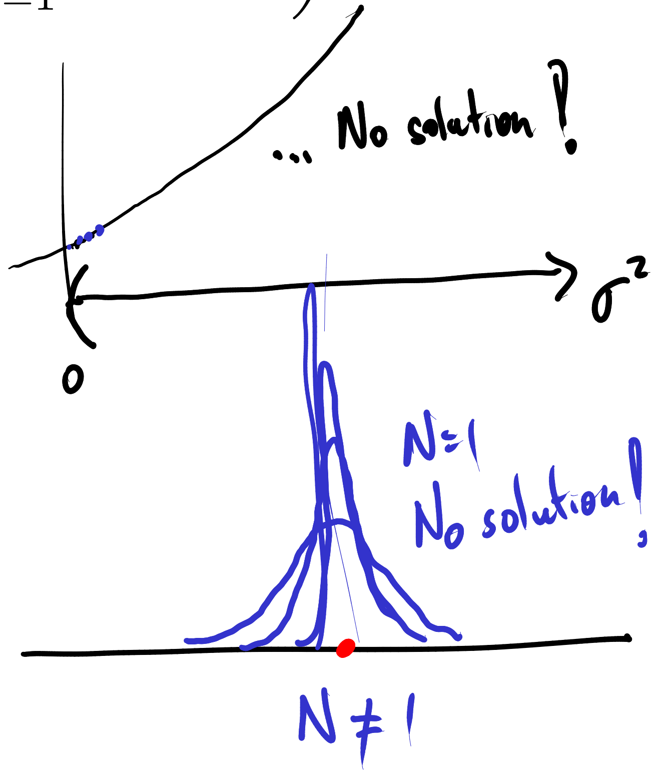
$$(\hat{\mu}_{\mathsf{MLE}}, \hat{\sigma}^2_{\mathsf{MLE}}) = \operatorname*{argmin}_{\mu,\sigma^2} \overbrace{\left(\frac{N}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right)}^{J(\mu,\sigma^2)}$$

.... convex ✓

subject to: $\sigma^2 > 0$.

$$J(\mu, \sigma^2) = \frac{N}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2$$

$$\nabla J(\mu, \sigma^2) = \left(\frac{\partial J}{\partial \mu}, \frac{\partial J}{\partial \sigma^2} \right) = 0$$



Example: Gaussian data

$$
\begin{aligned}
\frac{\partial J}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right) \\
&= \frac{1}{2\sigma^2}\sum_{i=1}^{N}\frac{\partial}{\partial \mu}(y_i-\mu)^2 \\
&= -\frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\mu) \\
&= \frac{N\mu}{\sigma^2}-\frac{1}{\sigma^2}\sum_{i=1}^{N}y_i = 0
\end{aligned}
$$

$$N\mu = \sum_{i=1}^{N} y_i$$

$$\therefore\ \hat{\mu}_{MLE} = \underline{\frac{1}{N}\sum_{i=1}^{N} y_i} = \hat{\mu}_N\,!$$

Example: Gaussian data

$$\frac{\partial J}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left(\frac{N}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right)$$

$$= \frac{N}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i-\mu)^2$$

$$= \frac{1}{\cancel{2\sigma^2}}\left(N - \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\mu)^2\right) \stackrel{!}{=} 0.$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_N)^2$$ ... biased sample variance!

Properties of MLE

N is finite.

- MLE has no finite-sample properties.
  - not necessarily unbiased
  - not necessarily minimum MSE.

- MLE has good asymptotic properties.
  - consistent
  - usually asymptotically unbiased