

# Maximum Likelihood Estimation

Berlin Chen

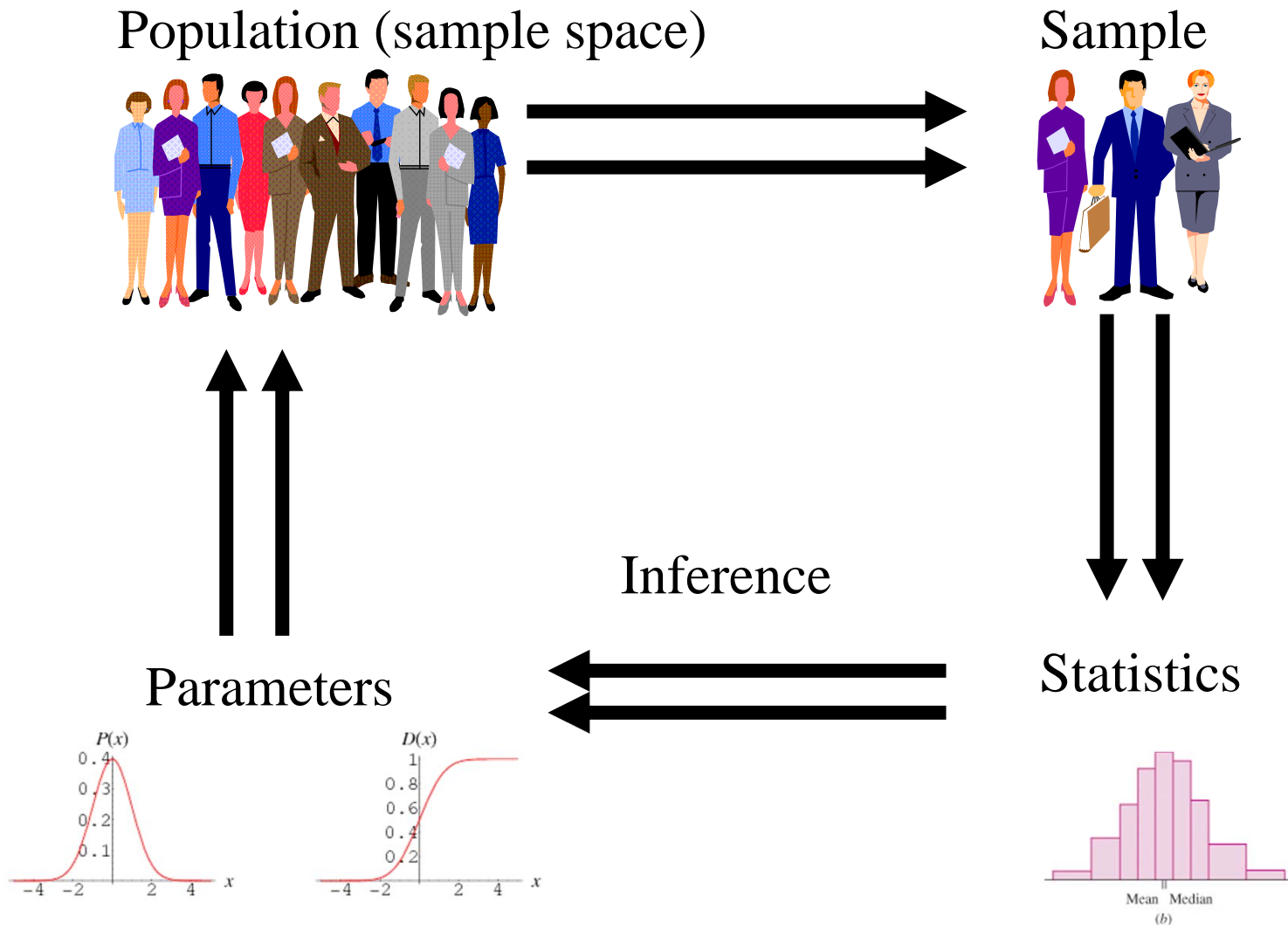
Department of Computer Science & Information Engineering  
National Taiwan Normal University

## References:

1. Ethem Alpaydin, *Introduction to Machine Learning*, Chapter 4, MIT Press, 2004

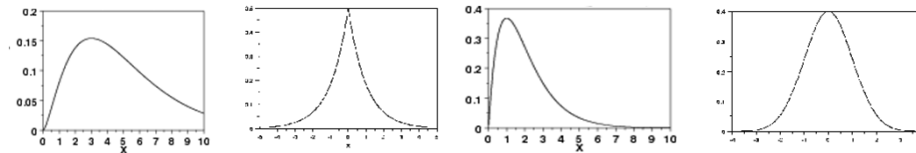
# Sample Statistics and Population Parameters

- A Schematic Depiction



# Introduction

- **Statistic**
  - Any value (or function) that is calculated from a given sample
  - Statistical inference: make a decision using the information provided by a sample (or a set of examples/instances)
- **Parametric methods**
  - Assume that examples are drawn from some distribution that obeys a known model  $p(x)$




- Advantage: the model is well defined up to a small number of parameters
  - E.g., mean and variance are **sufficient statistics** for the Gaussian distribution
- Model parameters are typically estimated by either maximum likelihood estimation or Bayesian (MAP) estimation

# Maximum Likelihood Estimation (MLE) (1/2)

- Assume the instances  $\mathbf{x} = \{x^1, x^2, \dots, x^t, \dots, x^N\}$  are independent and identically distributed (*iid*), and drawn from some known probability distribution  $X$ 
  - $X^t \sim p(x^t | \theta)$
  - $\theta$  : model parameters (assumed to be fixed but unknown here)
- MLE attempts to find  $\theta$  that make  $\mathbf{x}$  the most likely to be drawn
  - Namely, maximize the likelihood of the instances

$$l(\theta | \mathbf{x}) = p(\mathbf{x} | \theta) = p(x^1, \dots, x^N | \theta) = \prod_{t=1}^N p(x^t | \theta)$$

$x^1, \dots, x^N$  are *iid*



## MLE (2/2)

- Because logarithm will not change the value of  $\theta$  when it take its maximum (monotonically increasing/decreasing)
  - Finding  $\theta$  that maximizes the likelihood of the instances is equivalent to finding  $\theta$  that maximizes the log likelihood of the samples

$$L(\theta|\mathbf{x}) = \log l(\theta|\mathbf{x}) = \sum_{t=1}^N \log p(x^t|\theta)$$

$$a \geq b \\ \Rightarrow \log a \geq \log b$$

- As we shall see, logarithmic operation can further simplify the computation when estimating the parameters of those distributions that have exponents

# MLE: Bernoulli Distribution (1/3)

- Bernoulli Distribution

- A random variable  $X$  takes either the value  $x=0$  (with probability  $r$ ) or the value  $x=1$  (with probability  $1-r$ )
  - Can be thought of as  $X$  is generated from two distinct states
- The associated probability distribution

$$P(x) = r^x(1-r)^{1-x}, x \in \{0, 1\}$$

- The log likelihood for a set of *iid* instances  $\mathbf{x}$  drawn from Bernoulli distribution

$$\mathbf{x} = \{x^1, x^2, \dots, x^t, \dots, x^N\}$$

$$\begin{aligned} L(r|X) &= \log \prod_{t=1}^N r^{x^t} (1-r)^{(1-x^t)} \\ \theta & \nearrow \\ &= \left( \sum_{t=1}^N x^t \right) \log r + \left( N - \sum_{t=1}^N x^t \right) \log (1-r) \end{aligned}$$

## MLE: Bernoulli Distribution (2/3)

- MLE of the distribution parameter  $r$

$$\hat{r} = \frac{\sum_{t=1}^N x^t}{N}$$

- The estimate for  $r$  is the ratio of the number of occurrences of the event ( $x^t = 1$ ) to the number of experiments

- The expected value for  $X$

$$E[X] = \sum_{x \in \{0,1\}} x \cdot P(x) = 0 \cdot (1-r) + 1 \cdot r = r$$

- The variance value for  $X$

$$\text{var}(X) = E[X^2] - (E[X])^2 = r - r^2 = r(1-r)$$

# MLE: Bernoulli Distribution (3/3)

- Appendix A

$$\frac{dL(r|X)}{dr} = \frac{\partial \left[ \binom{N}{\sum_{t=1}^N x^t} \log r + \left( N - \sum_{t=1}^N x^t \right) \log (1 - r) \right]}{dr} = 0$$

$$\Rightarrow \frac{\binom{N}{\sum_{t=1}^N x^t}}{r} - \frac{\binom{N}{N - \sum_{t=1}^N x^t}}{1 - r} = 0$$

$$\frac{d \log y}{dy} = \frac{1}{y}$$

$$\Rightarrow \hat{r} = \frac{\sum_{t=1}^N x^t}{N}$$

The maximum likelihood estimate of the mean is the sample average



# MLE: Multinomial Distribution (1/4)

- Multinomial Distribution

- A generalization of Bernoulli distribution
- The value of a random variable  $X$  can be one of  $K$  mutually exclusive and exhaustive states  $x \in \{s_1, s_2, \dots, s_K\}$  with probabilities  $r_1, r_2, \dots, r_K$ , respectively
- The associated probability distribution

$$p(x) = \prod_{i=1}^K r_i^{s_i}, \quad \sum_{i=1}^K r_i = 1$$
$$s_i = \begin{cases} 1 & \text{if } X \text{ choose state } s_i \\ 0 & \text{otherwise} \end{cases}$$

- The log likelihood for a set of *iid* instances  $\mathbf{x}$  drawn from a multinomial distribution  $X$

$$L(\mathbf{r}|\mathbf{x}) = \log \prod_{t=1}^N \prod_{i=1}^K r_i^{s_i^t} \quad \mathbf{x} = \{x^1, x^2, \dots, x^t, \dots, x^N\}$$

## MLE: Multinomial Distribution (2/4)

- MLE of the distribution parameter  $r_i$

$$\hat{r}_i = \frac{\sum_{t=1}^N s_i^t}{N}$$

- The estimate for  $r_i$  is the ratio of the number of experiments with outcome of state  $i$  ( $s_i^t = 1$ ) to the number of experiments

# MLE: Multinomial Distribution (3/4)

- Appendix B

$$L(\mathbf{r}|\mathbf{x}) = \log \prod_{t=1}^N \prod_{i=1}^K r_i^{s_i^t} = \sum_{t=1}^N \sum_{i=1}^K \log r_i^{s_i^t}, \quad \text{with constraint : } \sum_{i=1}^K r_i = 1$$

$$\frac{\partial \bar{L}(\mathbf{r}|\mathbf{x})}{\partial r_i} = \frac{\partial \left[ \sum_{t=1}^N \sum_{i=1}^K s_i^t \cdot \log r_i + \lambda \left( \sum_{i=1}^K r_i - 1 \right) \right]}{\partial r_i} = 0$$

Lagrange Multiplier

$$\Rightarrow \sum_{t=1}^N s_i^t \cdot \frac{1}{r_i} + \lambda = 0$$

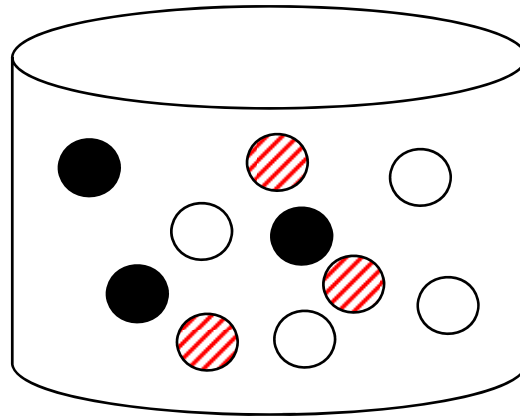
$$\Rightarrow r_i = -\frac{1}{\lambda} \sum_{t=1}^N s_i^t$$

$$\Rightarrow \sum_{i=1}^K r_i = 1 = -\frac{1}{\lambda} \sum_{t=1}^N \left( \sum_{i=1}^K s_i^t \right) = 1$$

$$\Rightarrow \lambda = -N$$

$$\Rightarrow \hat{r}_i = \frac{\sum_{t=1}^N s_i^t}{N}$$

# MLE: Multinomial Distribution (4/4)



$$P(B)=3/10$$

$$P(W)=4/10$$

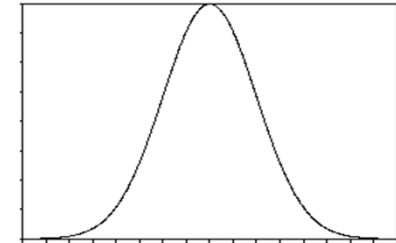
$$P(R)=3/10$$

# MLE: Gaussian Distribution (1/3)

- Also called Normal Distribution

- Characterized with mean  $\mu$  and variance  $\sigma^2$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$



- Recall that mean and variance are sufficient statistics for Gaussian

- The log likelihood for a set of *iid* instances drawn from Gaussian distribution  $X$

$$L(\mu, \sigma | \mathbf{x}) = \log \prod_{t=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x^t - \mu)^2}{2\sigma^2}\right)} \quad \mathbf{x} = \{x^1, x^2, \dots, x^t, \dots, x^N\}$$

$$= -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{\sum_{t=1}^N (x^t - \mu)^2}{2\sigma^2}$$

## MLE: Gaussian Distribution (2/3)

- MLE of the distribution parameters  $\mu$  and  $\sigma^2$

$$m = \hat{\mu} = \frac{\sum_{t=1}^N x^t}{N} \quad \text{sample average}$$

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{t=1}^N (x^t - m)^2}{N} \quad \text{sample variance}$$

- Remind that  $\mu$  and  $\sigma^2$  are still fixed but unknown

# MLE: Gaussian Distribution (3/3)

- Appendix C

$$L(\mu, \sigma | \mathbf{x}) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 - \frac{\sum_{t=1}^N (x^t - \mu)^2}{2\sigma^2}$$

$$\frac{\partial L(\mu, \sigma | \mathbf{x})}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2} \sum_{t=1}^N (x^t - \mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum_{t=1}^N x^t}{N}$$

$$\frac{\partial L(\mu, \sigma | \mathbf{x})}{\partial \sigma^2} = 0 \Rightarrow -N + \frac{1}{\sigma^2} \sum_{t=1}^N (x^t - \hat{\mu})^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_{t=1}^N (x^t - \hat{\mu})^2}{N}$$

# Evaluating an Estimator : Bias and Variance (1/6)

- The mean square error of the estimator  $d$  can be further decomposed into two parts respectively composed of bias and variance

$$\begin{aligned}
 r(d, \theta) &= E[(d - \theta)^2] \\
 &= E[(d - E[d] + E[d] - \theta)^2] \\
 &= E[(d - E[d])^2 + (E[d] - \theta)^2 + 2(d - E[d])(E[d] - \theta)] \\
 &= E[(d - E[d])^2] + E[\underbrace{(E[d] - \theta)^2}_{\text{constant}}] + 2E[\underbrace{(d - E[d])(E[d] - \theta)}_{\text{constant}}] \\
 &= E[(d - E[d])^2] + (E[d] - \theta)^2 + 2E[\cancel{(d - E[d])}(E[d] - \theta)}] \\
 &= \underbrace{E[(d - E[d])^2]}_{\text{variance}} + \underbrace{(E[d] - \theta)^2}_{\text{bias}^2}
 \end{aligned}$$



## Evaluating an Estimator : Bias and Variance (2/6)

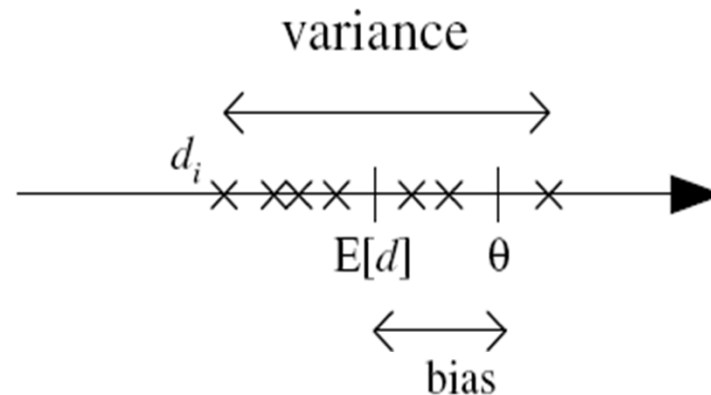


Figure 4.1:  $\theta$  is the parameter to be estimated.  $d_i$  are several estimates (denoted by 'x') over different samples. Bias is the difference between the expected value of  $d$  and  $\theta$ . Variance is how much  $d_i$  are scattered around the expected value. We would like both to be small.

# Evaluating an Estimator : Bias and Variance (3/6)

- Example 1: **sample average and sample variance**

- Assume samples  $\mathbf{x} = \{x^1, x^2, \dots, x^t, \dots, x^N\}$  are independent and identically distributed (*iid*), and drawn from some known probability distribution  $X$  with mean  $\mu$  and variance  $\sigma^2$

- Mean  $\mu = E[X] = \sum_x x \cdot p(x)$

- Variance  $\sigma^2 = E[(X - \mu)^2] = E[X^2] - (E[X])^2$

- Sample average (mean) for the observed samples  $m = \frac{1}{N} \sum_{t=1}^N x^t$

- Sample variance for the observed samples  $s^2 = \frac{1}{N} \sum_{t=1}^N (x^t - m)^2$

or  $s^2 = \frac{1}{N-1} \sum_{t=1}^N (x^t - m)^2$  ?

# Evaluating an Estimator : Bias and Variance (4/6)

- Example 1 (count.)
  - Sample average  $m$  is an **unbiased** estimator of the mean  $\mu$

$$E[m] = E\left[\frac{1}{N} \sum_{t=1}^N X^t\right] = \frac{1}{N} \sum_{t=1}^N E[X] = \frac{N \cdot \mu}{N} = \mu$$

$$\therefore E[m] - \mu = 0$$

- $m$  is also a **consistent** estimator:  $\text{Var}(m) \rightarrow 0$  as  $N \rightarrow \infty$

$$\text{Var}(m) = \text{Var}\left(\frac{1}{N} \sum_{t=1}^N X^t\right) = \frac{1}{N^2} \sum_{t=1}^N \text{Var}(X) = \frac{N \cdot \sigma^2}{N^2} = \frac{\sigma^2}{N} \xrightarrow{N=\infty} 0$$

$$\text{Var}(aX + b) = a^2 \cdot \text{Var}(X)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

# Evaluating an Estimator : Bias and Variance (5/6)

- Example 1 (count.)

- Sample variance  $s^2$  is an **asymptotically unbiased** estimator of the variance  $\sigma^2$

$$s^2 = \frac{1}{N} \sum_{t=1}^N (x^t - m)^2$$

$$\sum_{t=1}^N x^t = N \cdot m$$

$$\begin{aligned} E[s^2] &= E \left[ \frac{1}{N} \sum_{t=1}^N (X^t - m)^2 \right] \\ &= E \left[ \frac{1}{N} \sum_{t=1}^N (X - m)^2 \right] \quad (X^t \text{'s are i.i.d.}) \\ &= E \left[ \frac{1}{N} \sum_{t=1}^N (X^2 - \underline{2X \cdot m} + m^2) \right] \\ &= E \left[ \frac{N \cdot X^2 - 2N \cdot m^2 + Nm^2}{N} \right] \\ &= E \left[ \frac{N \cdot X^2 - N \cdot m^2}{N} \right] = \frac{N \cdot E[X^2] - N \cdot E[m^2]}{N} \end{aligned}$$

# Evaluating an Estimator : Bias and Variance (6/6)

- Example 1 (count.)

- Sample variance  $s^2$  is an **asymptotically unbiased** estimator of the variance  $\sigma^2$

$$\text{Var}(m) = \frac{\sigma^2}{N} = E[m^2] - (E[m])^2$$

$$\Rightarrow E[m^2] = \frac{\sigma^2}{N} + (E[m])^2 = \frac{\sigma^2}{N} + \mu^2$$

$$E[s^2] = \frac{N \cdot E[X^2] - N \cdot E[m^2]}{N}$$

$$= \frac{N(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right)}{N}$$

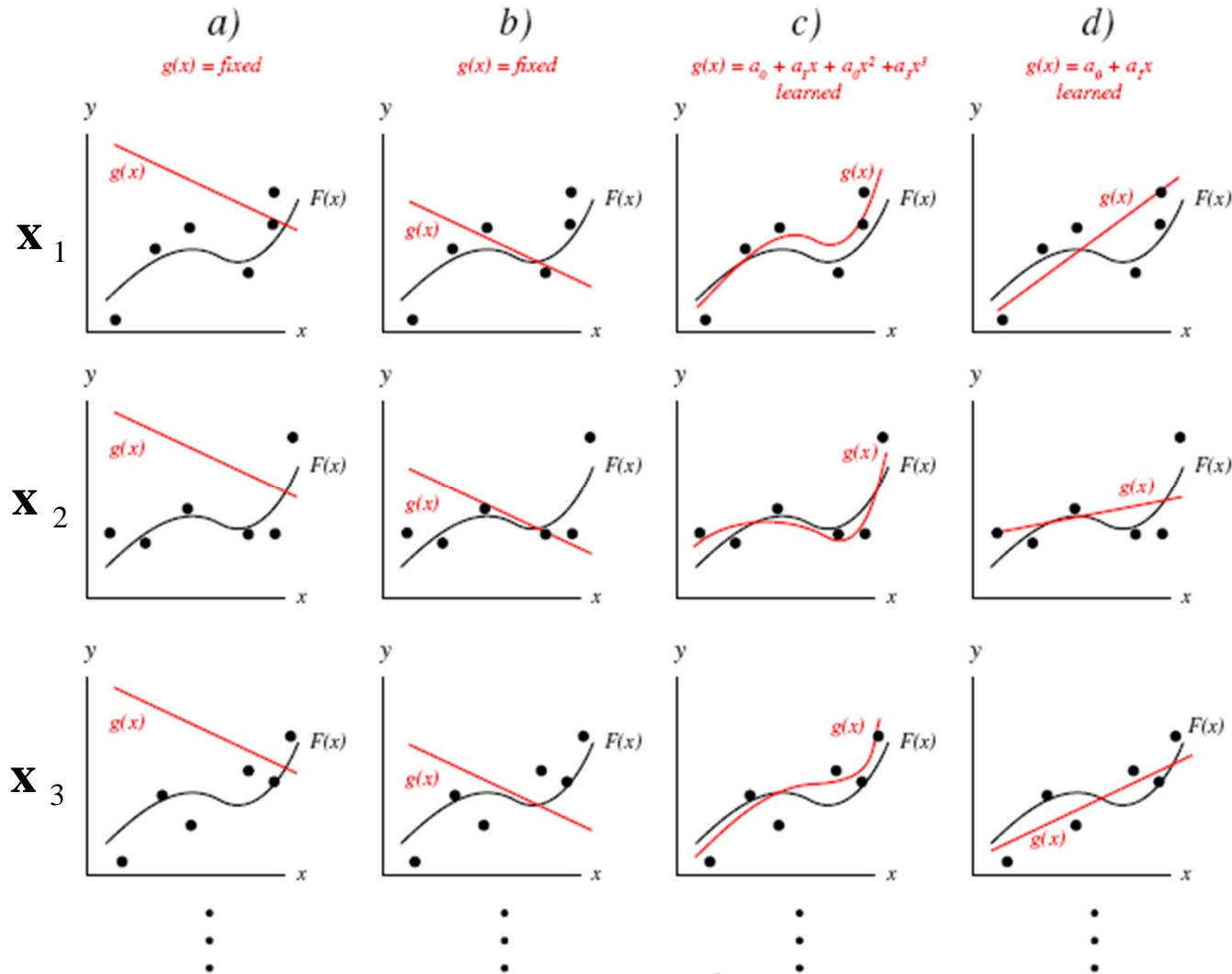
$$= \frac{(N-1)\sigma^2}{N} \xrightarrow{N \rightarrow \infty} \sigma^2$$

$$\text{Var}(X) = \sigma^2 = E[X^2] - (E[X])^2$$

$$\Rightarrow E[X^2] = \sigma^2 + (E[X])^2 = \sigma^2 + \mu^2$$

The size of the observed sample set

# Bias and Variance: Example 2



different samples for an unknown population

$$X \rightarrow (x, y)$$

$$y = F(x)$$

$$y' = F(x) + \varepsilon$$

error of measurement

*Simple is Elegant ?*