Maximum Likelihood Estimation

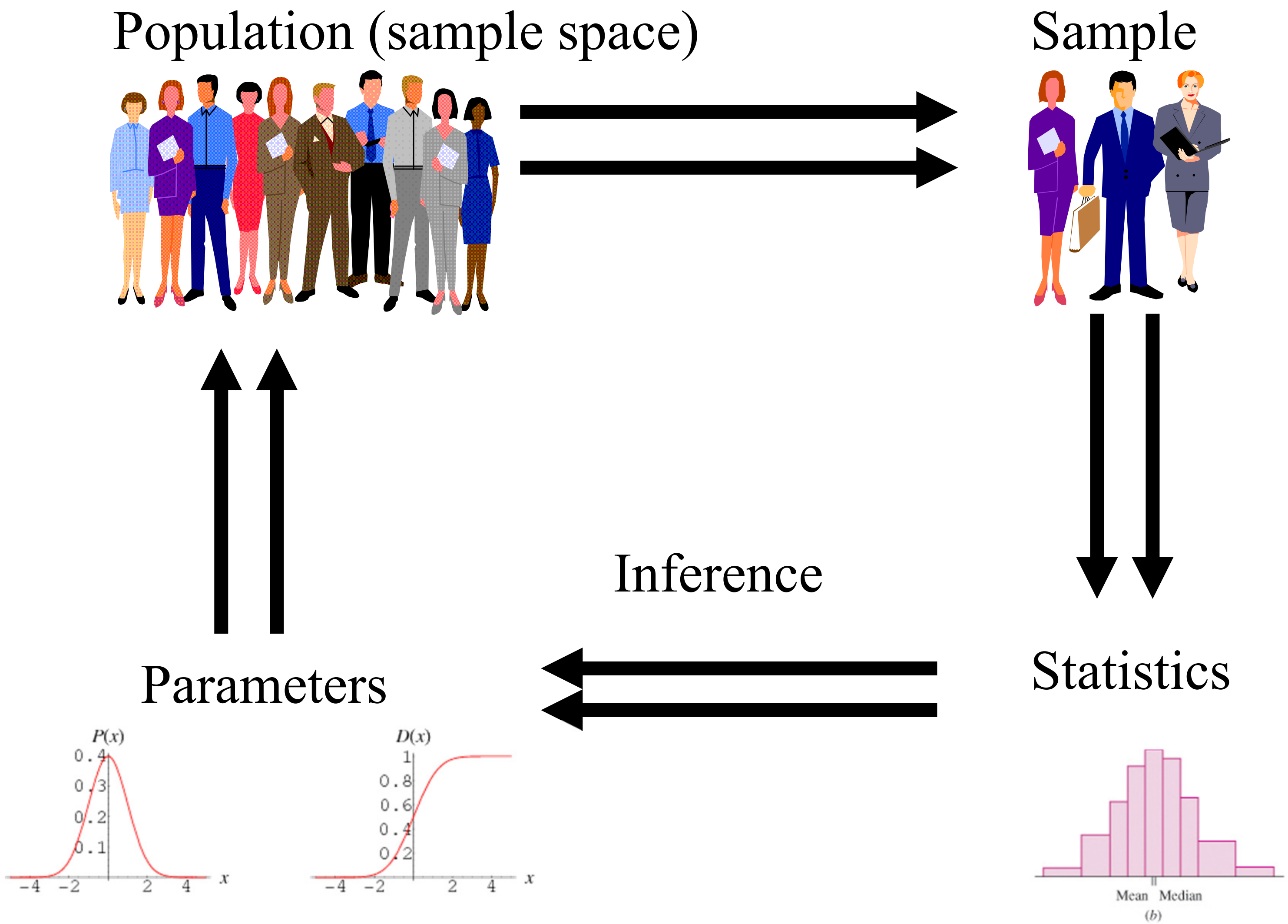
Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

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

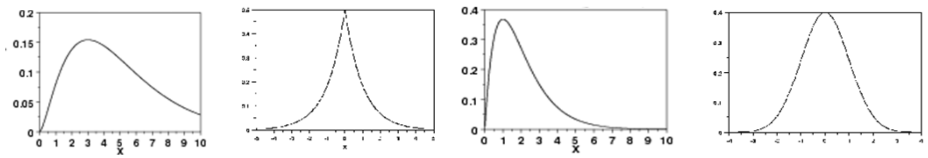
Sample Statistics and Population Parameters

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

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

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

MLE (2/2)

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

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

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

  - 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 $x=1$ (with probability r) or the value $x=1$ (with probability $1-r$)
    - Can be thought of as X is generated form two distinct states
  - The associated probability distribution

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

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

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

θ → L

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

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

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

MLE: Bernoulli Distribution (3/3)

- Appendix A

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

$$\Rightarrow \frac{\left(\sum_{t=1}^{N} x^t\right)}{r} - \frac{\left(N - \sum_{t=1}^{N} x^t\right)}{1-r} = 0$$

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

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

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 $x \in \{s_1, s_2, \cdots, s_K\}$ with probabilities $r_1, r_2, \cdots, r_K$, respectively
  - The associated probability distribution

$$p(x) = \prod_{i=1}^{K} r_i^{s_i}, \qquad \sum_{i=1}^{K} r_i = 1$$

$$s_i = \begin{cases} 1 & \text{if } X \text{ choose state } s_i \\ 0 & \text{otherwise} \end{cases}$$

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

$$L(\mathbf{r} \mid \mathbf{x}) = \log \prod_{t=1}^{N} \prod_{i=1}^{K} r_i^{s_i^t} \qquad \mathbf{x} = \{x^1, x^2, \ldots, x^t, \ldots, x^N\}$$

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

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

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

Lagrange Multiplier

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

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

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

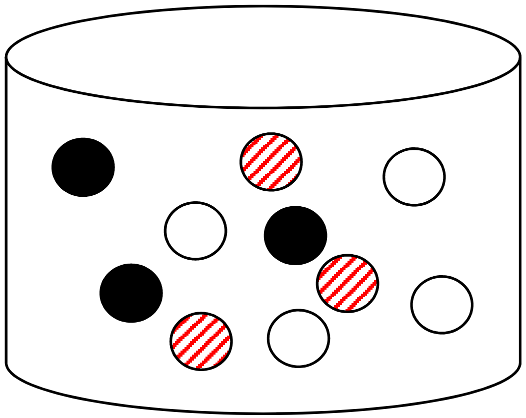
=1

$$\Rightarrow \lambda = -N$$

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

Lagrange Multiplier: http://www.slimy.com/~steuard/teaching/tutorials/Lagrange.html

MLE: Multinomial Distribution (4/4)



P(B)=3/10

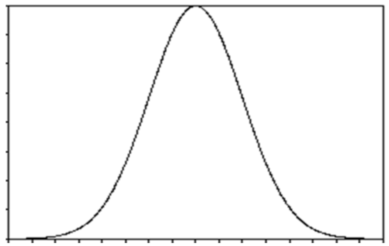
P(W)=4/10

P(R)=3/10

MLE: Gaussian Distribution (1/3)

- Also called Normal Distribution
  - Characterized with mean μ and variance σ^2

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



  - Recall that mean and variance are sufficient statistics for Gaussian
- The log likelihood for a set of *iid* instances drawn from Gaussian distribution X

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

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

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

MLE: Gaussian Distribution (2/3)

- MLE of the distribution parameters μ and σ^2

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

sample average

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

sample variance

- Remind that μ and σ^2 are still fixed but unknown

MLE: Gaussian Distribution (3/3)

- Appendix C

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

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

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

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

$$
\begin{aligned}
r(d,\theta) &= E\left[(d-\theta)^2\right] \\
&= E\left[(d-E[d]+E[d]-\theta)^2\right] \\
&= E\left[(d-E[d])^2+(E[d]-\theta)^2+2(d-E[d])(E[d]-\theta)\right] \\
&= E\left[(d-E[d])^2\right]+E\left[\underbrace{(E[d]-\theta)^2}_{\text{constant}}\right]+2E\left[(d-E[d])\underbrace{(E[d]-\theta)}_{\text{constant}}\right] \\
&= E\left[(d-E[d])^2\right]+(E[d]-\theta)^2+2\underbrace{E\left[(d-E[d])\right]}_{0}(E[d]-\theta) \\
&= \underbrace{E\left[(d-E[d])^2\right]}_{\textbf{variance}}+\underbrace{(E[d]-\theta)^2}_{\textbf{bias}^2}
\end{aligned}
$$

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

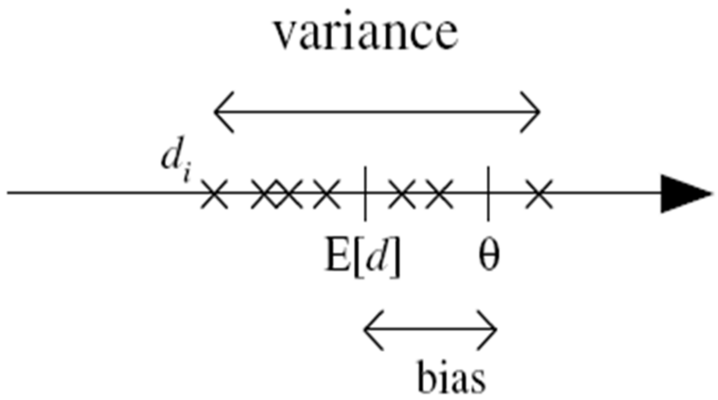


Figure 4.1: θ is the parameter to be estimated. d_i are several estimates (denoted by '$\times$') over different samples. Bias is the difference between the expected value of d and θ. Variance is how much d_i are scattered around the expected value. We would like both to be small.

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

- Example 1: **sample average and sample variance**
  - Assume samples $\mathbf{x} = \{x^1, x^2, \ldots, x^t, \ldots, x^N\}$ are independent and identically distributed (*iid*), and drawn from some known probability distribution X with mean μ and variance σ^2
    - Mean $\mu = E[X] = \sum_x x \cdot p(x)$
    - Variance $\sigma^2 = E\left[(X-\mu)^2\right] = E\left[X^2\right] - (E[X])^2$
    - Sample average (mean) for the observed samples $m = \frac{1}{N}\sum_{t=1}^{N} x^t$
    - Sample variance for the observed samples $s^2 = \frac{1}{N}\sum_{t=1}^{N}\left(x^t - m\right)^2$

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

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

- Example 1 (count.)
  - Sample average m is an unbiased estimator of the mean μ

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

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

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

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

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

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

- Example 1 (count.)
  - Sample variance s^2 is an asymptotically unbiased estimator of the variance σ^2

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

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

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

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

- Example 1 (count.)
  - Sample variance s^2 is an asymptotically unbiased estimator of the variance σ^2

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

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

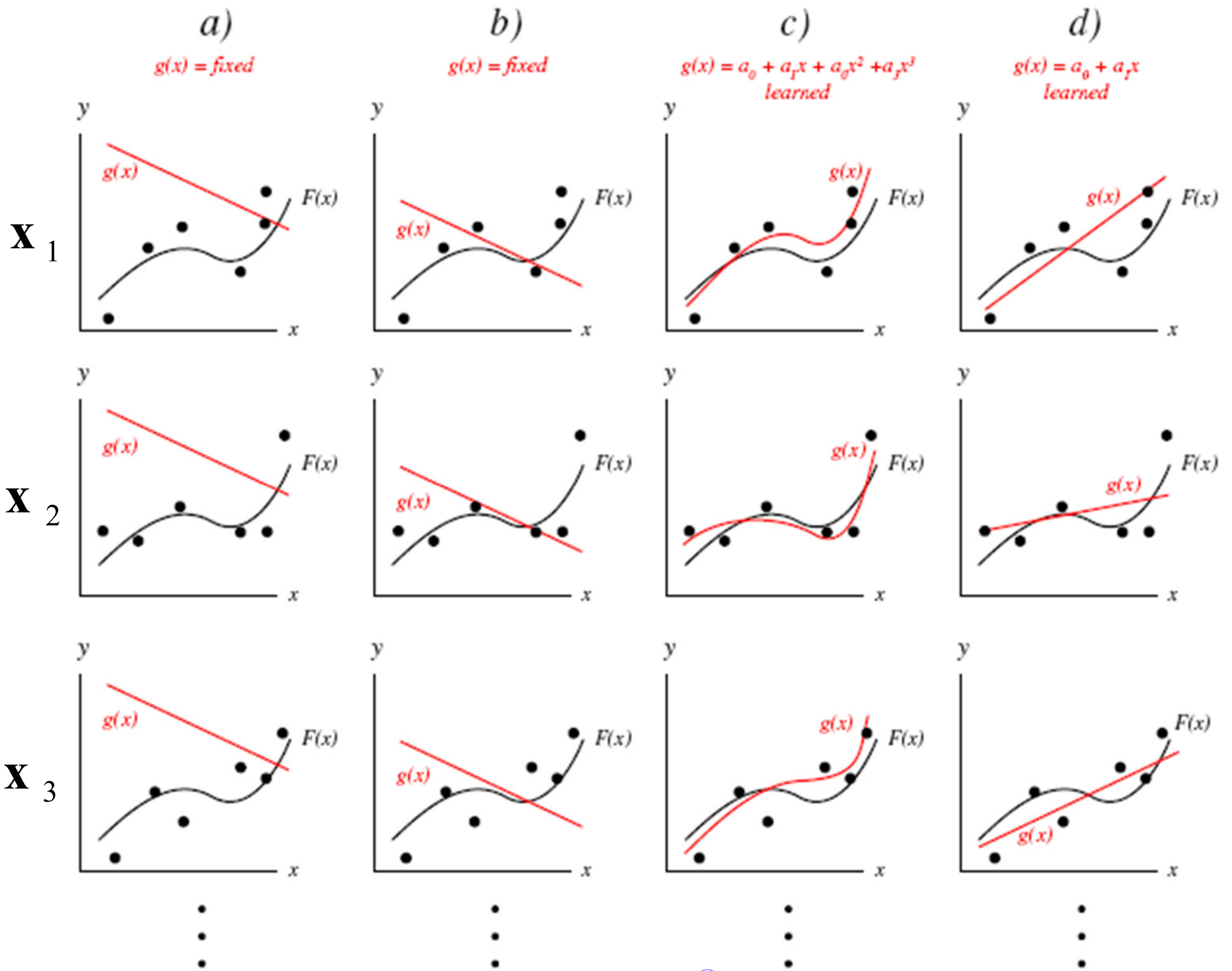
$$= \frac{N(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right)}{N}$$

$$\mathrm{Var}(X) = \sigma^2 = E[X^2] - (E[X])^2$$
$$\Rightarrow E[X^2] = \sigma^2 + (E[X])^2 = \sigma^2 + \mu^2$$

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

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 ?