EE4211: Data Science for the Internet of Things

Parameter Estimation from Data

Biplab Sikdar

Agenda

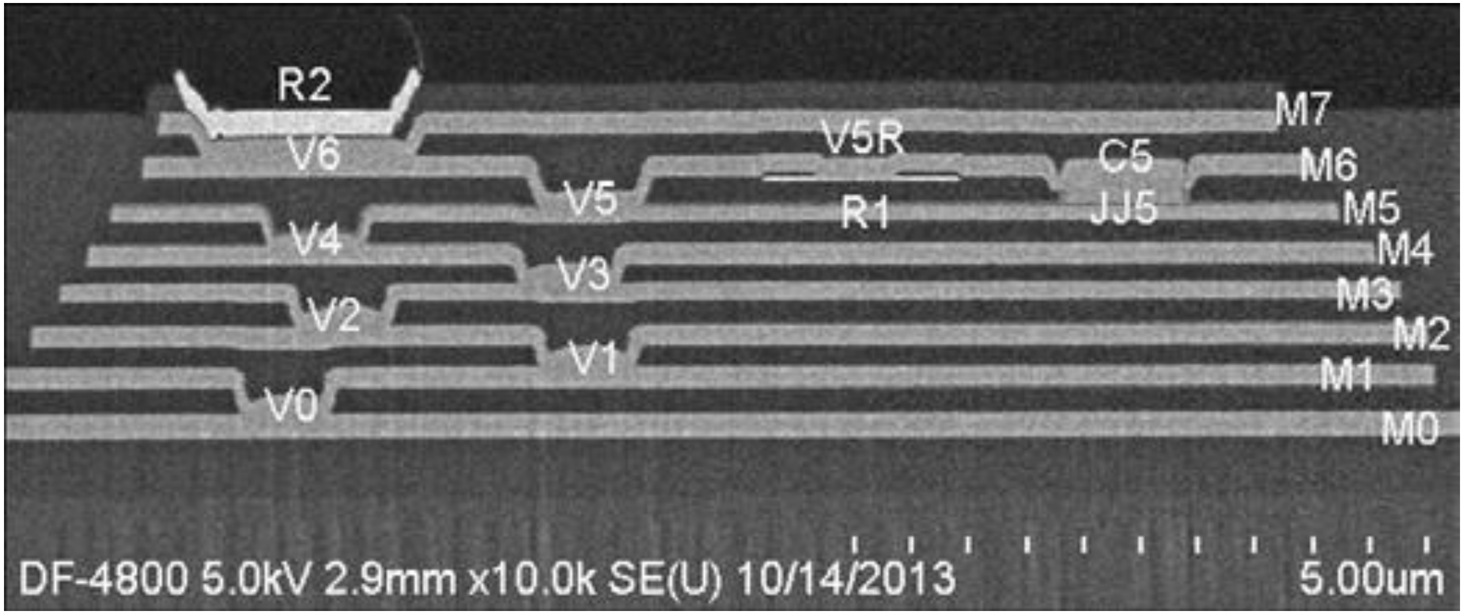
- Examples
- Moment estimation
- Maximum likelihood estimation
- Bayesian parameter estimation

Example 1: Superconducting VLSI

- Current VLSI technologies: CMOS (complimentary metal oxide semiconductor)
- Superconducting digital electronics: applications in high performance-computing due to a potential for much higher clock rates and lower energy dissipation
- Problem: Current superconducting digital circuits about 5 orders of magnitude lower integration scale than the typical CMOS circuits.
  - Largest demonstrated superconducting digital circuits have only about 10^5 switching elements whereas CMOS circuits routinely have over 10^{10} transistors.

Example 1: Superconducting VLSI

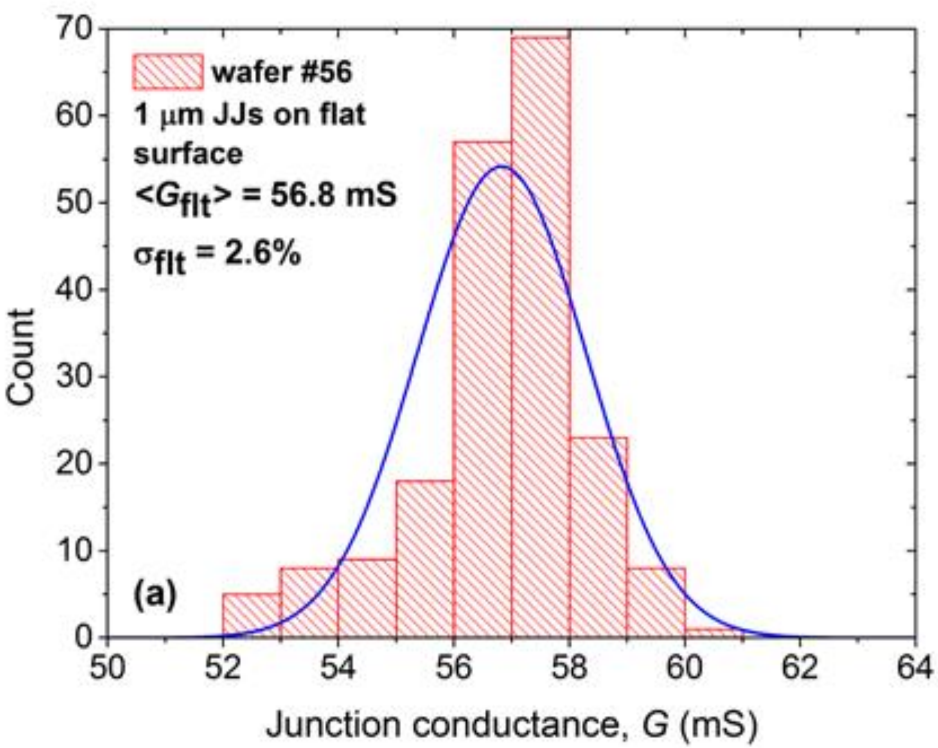
- Two processes developed at MIT Lincoln Labs: 8 and 9 superconducting layers

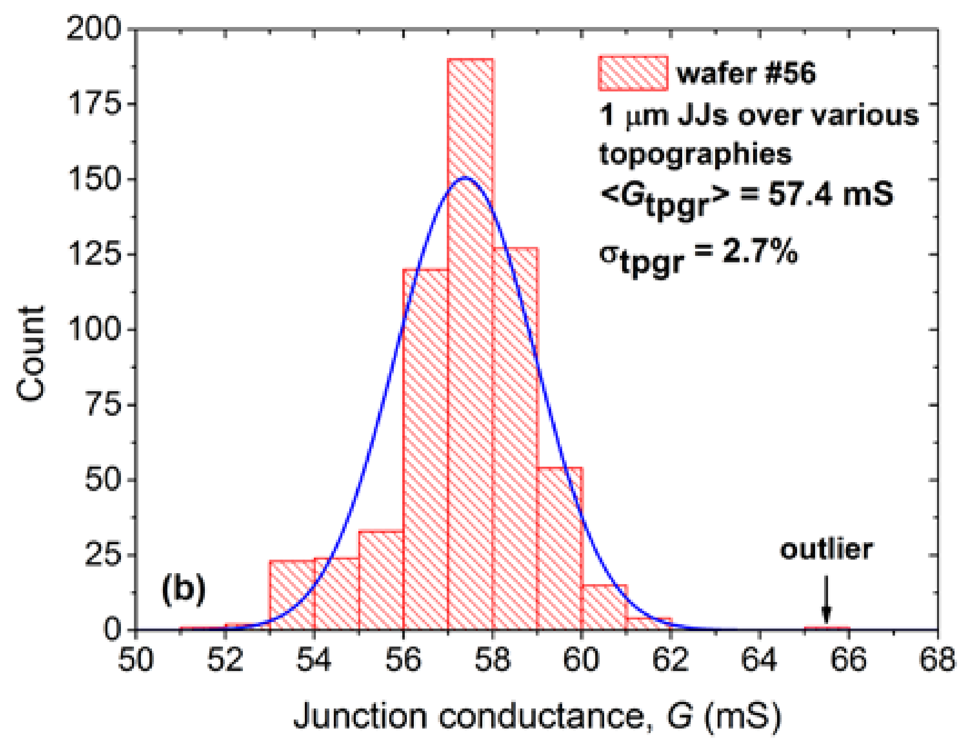


Scanning Electron Microscope image of a wafer cross section

Example 1: Superconducting VLSI

- Junction conductances at top and bottom layers

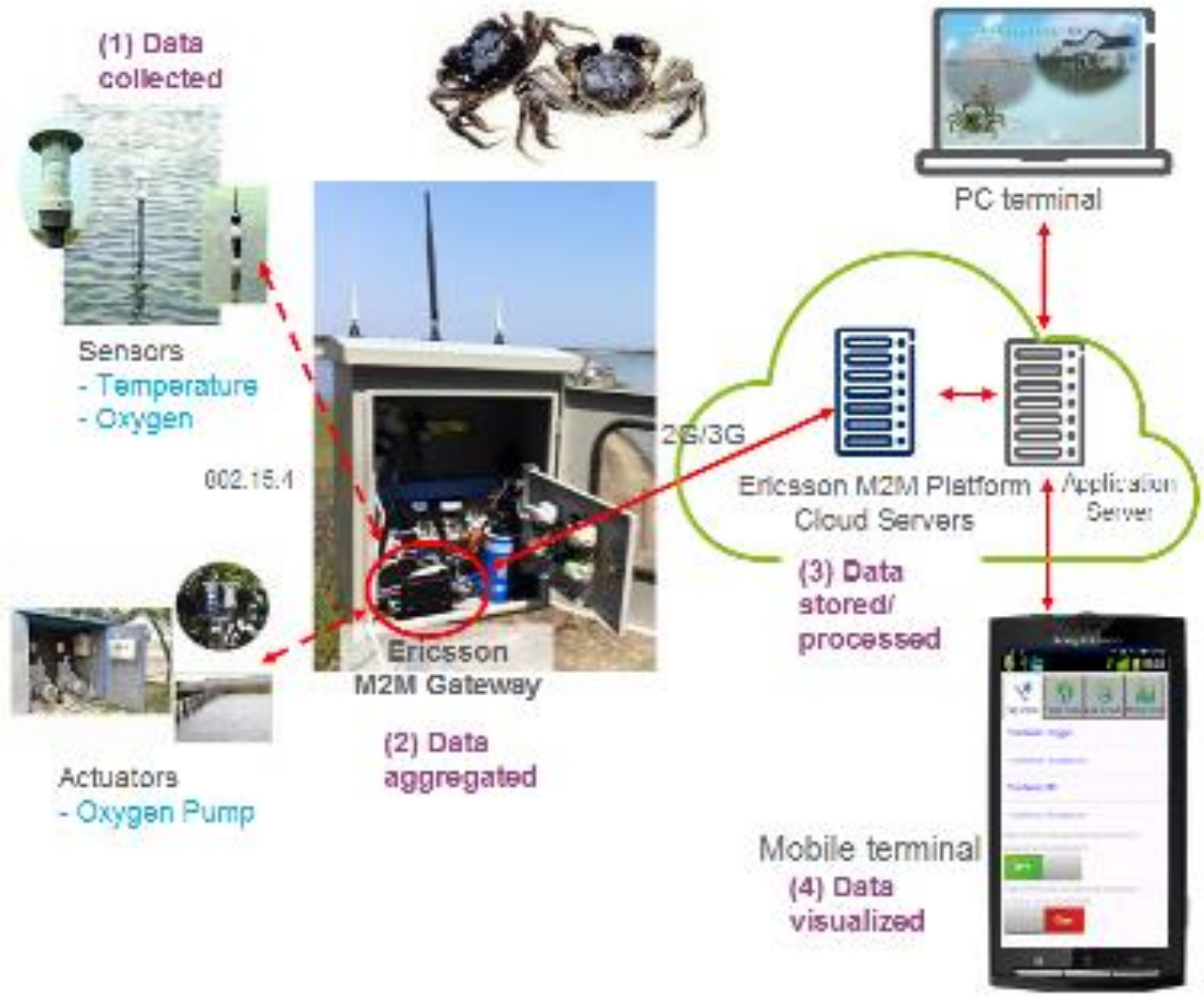




National University of Singapore

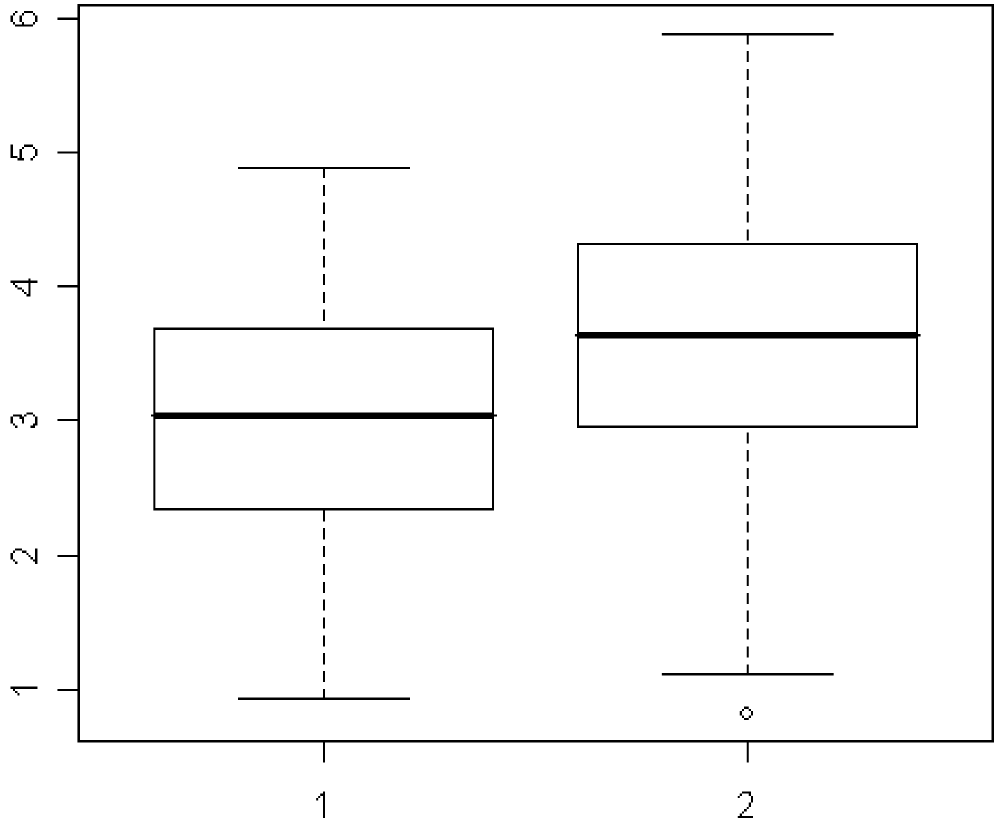
Example 2: Fish Farms

Example 2: Fish Farms



Example 2: Fish Farms

- Fish population fed with normal and high protein diet
- Estimate the weight gain due to change in diet



Parameter/Point Estimation

- Suppose we know we have data with values $x_1, x_2, \cdots, x_n$ drawn from some (e.g. exponential) distribution
- The exponential distribution $exp(\lambda)$ is not a single distribution but rather a one-parameter family of distributions
- Each value of λ defines a different distribution in the family, with pdf $f_X(x) = \lambda e^{-\lambda x}, x \geq 0$
- The question remains: which exponential distribution?
- We are interested in finding a point estimate to the parameter λ

Parameter Estimation

- Questions to ask:
  - How to estimate model parameters from data?
  - What are the factors to consider when choosing between estimators?
  - Is there an optimal way of estimating parameters from data?
  - How to compare different parameter values?

Parameter Estimation

- Most questions in statistics can be formulated in terms of making statements about underlying parameters
- Objective: devise a framework for estimating those parameters and making statements about our certainty in these estimates
- Three different approaches to making such statements
  - Moment estimators
  - Maximum likelihood estimators
  - Bayesian estimators

Moment Estimation

- Moment estimation techniques: parameter values are found that match sample moments (mean, variance, etc.) to those expected

Population (parameter)

$$X \sim N(\mu, \sigma^2)$$
$$E[X] = \mu$$
$$\text{var}(X) = E[(X-\mu)^2] = \sigma^2$$

Samples (statistic)

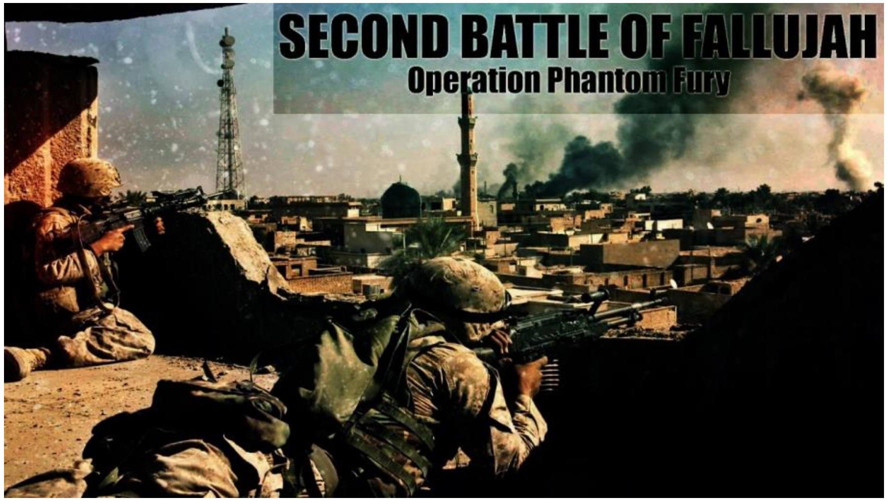
$$\frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}$$
$$\frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \hat{\sigma}^2$$

Moment Estimation

- Step 1: Start with the underlying distribution of the data
- Step 2: Obtain the expression for the first moment (mean) in terms of the parameters
- Step 3: Obtain expressions for higher order moments if distribution has more than one parameter
- Step 4: Compute the sample based moments
- Step 5: Substitute sample moments in the analytic expressions for the moments and solve to obtain the parameters

Example

- We measure the levels of white phosphorus in the air
- Used in munitions, chemical weapons

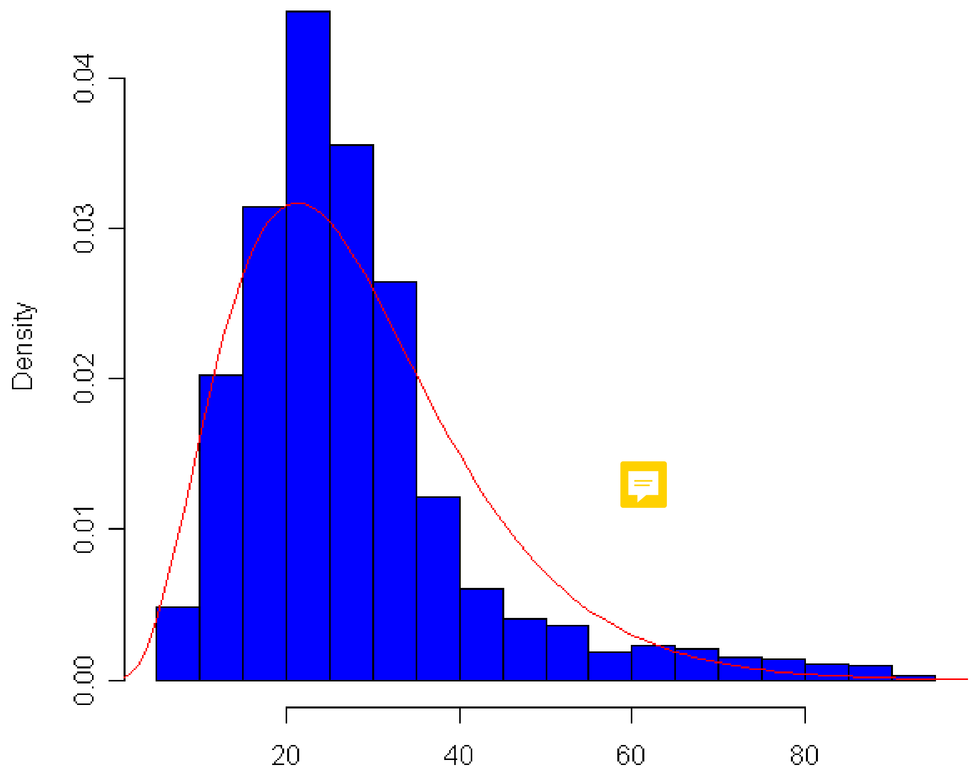


Example

- Phosphorus levels have a gamma distribution
- Gamma distribution:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

  - shape parameter: α
  - scale parameter: β
  - Distribution mean: α/β
  - Variance: α/β^2



$$\hat{\beta} = \frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0.14$$

$$\hat{\alpha} = \beta\bar{X} = 4.03$$

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

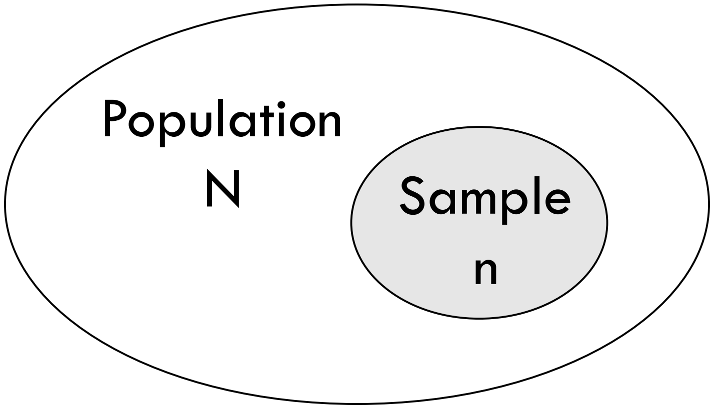
Bias

- Although the moment method looks sensible, it can lead to biased estimators
- Bias is measured by the difference between the expected estimate and the truth

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

- In the previous example, estimates of both parameters are upwardly biased
- Bias is not the only thing to worry about
- We also need to worry about the variance of an estimator

Bias

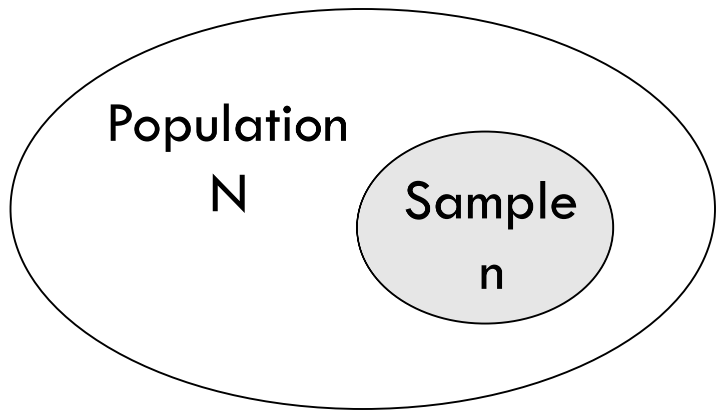


- Parameter: population mean, $\mu = E[X] = \frac{1}{N}\sum_{i=1}^{N} X_i$
- Statistic: sample mean, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

Bias

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

Bias



- Parameter: population mean, $\sigma^2 = E[(X - \mu)^2]$
- Statistic: sample mean, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$
- $\mathrm{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

Bias

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i^2 - 2X_i\hat{\mu} + \hat{\mu}^2\right)\right] \qquad (E[aX] = aE[X])$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}X_i^2 - 2\hat{\mu}\sum_{i=1}^{n}X_i + n\hat{\mu}^2\right]$$

$$= \frac{1}{n}E\left[\sum_{i=1}^{n}X_i^2 - 2\hat{\mu}n\hat{\mu} + n\hat{\mu}^2\right] \qquad \left(\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}X_i\right)$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}E\left[X_i^2\right] - nE[\hat{\mu}^2]\right] \qquad (E[X+Y] = E[X] + E[Y])$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n}(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \qquad \left(E[\hat{\mu}^2] = \frac{\sigma^2}{n} + \mu^2\right)$$

$$= \frac{(n-1)}{n}\sigma^2$$

NUS National University of Singapore

Mean Square Error of an Estimator

$$
\begin{aligned}
MSE[\hat{\theta}] &= E\left[(\hat{\theta}-\theta)^2\right] \\
&= E\left[(\hat{\theta}-E[\hat{\theta}]+E[\hat{\theta}]-\theta)^2\right] \\
&= E\left[(\hat{\theta}-E[\hat{\theta}])^2\right]+E\left[(E[\hat{\theta}]-\theta)^2\right] \\
&\quad +2E[(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)] \\
&= E\left[(\hat{\theta}-E[\hat{\theta}])^2\right]+(E[\hat{\theta}]-\theta)^2 \\
&\quad +2(E[\hat{\theta}]-\theta)E[(\hat{\theta}-E[\hat{\theta}])] && E[\hat{\theta}]-\theta\text{: constant} \\
&= E\left[(\hat{\theta}-E[\hat{\theta}])^2\right]+(E[\hat{\theta}]-\theta)^2 \\
&\quad +2(E[\hat{\theta}]-\theta)(E[\hat{\theta}]-E[\hat{\theta}]) && E[\hat{\theta}]\text{: constant} \\
&= VAR[\hat{\theta}]+(\mathrm{bias}[\hat{\theta}])^2
\end{aligned}
$$

NUS National University of Singapore

Efficiency

- An estimator $\hat{\theta}$ is said to be efficient if its mean square error is the smallest among all competitors
- Relative efficiency:

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE[\hat{\theta}_1]}{MSE[\hat{\theta}_2]}$$

NUS National University of Singapore

Problems with Moment Estimation

- They are usually not the "best estimators" available: do not achieve the minimum MSE
- Sometimes the estimates may be meaningless:
  - Uniform distribution: $U(0, \theta)$
  - Observed data: 3, 5, 6,18
- Expected value: $E[X] = \theta/2$
- Method of moments estimate of θ:

$$\hat{\theta} = 2E[X] = 2\frac{3+5+6+18}{4} = 16$$

- This is not acceptable, because we have a sample of 18

Maximum Likelihood Estimator

- We have data with values $x_1, x_2, \cdots, x_n$ drawn from some distribution with parameter θ
- We would like to obtain an estimate of θ: $\hat{\theta}$
- One of the approaches to estimating $\hat{\theta}$ is to find what value of parameter θ makes the current observation $x_1, x_2, \cdots, x_n$ most likely
  - For any model the maximum information about model parameters is obtained by considering the likelihood function

Maximum Likelihood Estimator

- Maximum likelihood estimate: joint-distribution of the observed data is given by

$$f_x(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f_x(x_i; \theta) \triangleq L(\theta)$$

- Maximizing the likelihood function:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta) = \arg\left\{\frac{dL(\theta)}{d\theta} = 0\right\}$$

- It is often easier to work with the natural log of the likelihood function: log likelihood

NUS National University of Singapore

Example

- Suppose that the time to failure of a photo-lithography equipment is modeled by an exponential distribution with (unknown) parameter λ
- Data: 2, 3, 1, 3, 4 years
- What is the MLE for λ?
- Exponential: $f_X(x) = \lambda e^{-\lambda x}, x \geq 0$
- The likelihood function:

$$L(\lambda) = \prod_{i=1}^{n} f_x(x_i)$$
$$= \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}$$
$$= \lambda^5 e^{-13\lambda}$$

Example

- Log likelihood:

$$\ln\big(L(\lambda)\big) = 5\ln\lambda - 13\lambda$$

- Maximizing the likelihood:

$$\frac{d\ln\big(L(\lambda)\big)}{d\lambda} = \frac{5}{\lambda} - 13 = 0$$

- Maximum likelihood estimate:

$$\hat{\lambda}_{MLE} = \frac{13}{5}$$

Why use Method of Moments?

- Consider a family of Gamma distribution with parameters $\theta_1 = \alpha$ and $\theta_2 = \beta$, with $\theta_1, \theta_2 > 0$:

$$f_X(x) = \frac{1}{\Gamma(\theta_1)\theta_2^{\theta_1}} x^{\theta_1 - 1} e^{-\frac{x}{\theta_2}}, \qquad x > 0$$

- For data with values $x_1, x_2, \cdots, x_n$ from this distribution:

$$L(\theta_1, \theta_2) = \prod_{i=1}^{n} f_x(x_i; \theta_1, \theta_2)$$
$$= \left[\frac{1}{\Gamma(\theta_1)\theta_2^{\theta_1}}\right]^n (x_1 x_2 \cdots x_n)^{\theta_1 - 1} e^{-\sum_{i=1}^{n} \frac{x_i}{\theta_2}}$$

- Gamma function makes it hard to find MLE in a closed form

Bayesian Estimation

- The main difference with respect to MLE is that in the Bayesian case θ is a random variable
- This notion is encapsulated in the use of a subjective prior for the parameters
- Basic idea:
  - Observed data: $x_1, x_2, \cdots, x_n$
  - Probability distribution for data given parameters: $f_X(x;\theta)$
  - Prior distribution for parameter: $f_\Theta(\theta)$
  - Goal: compute posterior probability

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{f_X(x)} \sim f_{X|\Theta}(x|\theta) f_\Theta(\theta)$$

Biplab Sikdar

Bayesian Estimation

- We have conditional distribution for parameter θ: $f_{\Theta|X}(\theta|x)$
- What if we are asked to make a point estimate?
- Option: θ that maximizes $f_{\Theta|X}(\theta|x)$: $\arg\max_{\theta} f_{\Theta|X}(\theta|x)$
- Option: Depending on the cost of error
  - Suppose the cost is $\left(\theta - \hat{\theta}\right)^2$: $\hat{\theta} = E[\theta|x]$
  - Because for a random variable Y, the expected value of the squared error, $E[(Y-b)^2]$, is minimized at $b = E[Y]$
  - Suppose the cost is $\left|\theta - \hat{\theta}\right|$: $\hat{\theta} = \mathrm{Median}[\theta|x]$
  - Expected value of $E[|Y-b|]$ is minimized at $b = \mathrm{Median}[Y]$

Example

- We toss a coin m times and observe n heads
- If I toss the coin again, what is the probability of a heads?
- Model for data: generated by a sequence of independent draws from a Bernoulli distribution, parameterized by θ, which is the probability of flipping a heads.
- MLE estimator for θ:

$$L(\theta) = \theta^n (1-\theta)^{m-n}$$
$$\mathrm{l}(\theta) = n \ln \theta + (m-n) \ln(1-\theta)$$

- MLE estimator for θ:

$$\frac{d l(\theta)}{d\theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{MLE} = \frac{n}{m}$$

National University of Singapore

Example

- Bayesian:

$$f_{\Theta|X}(\theta|x) \sim f_{X|\Theta}(x|\theta) f_\Theta(\theta)$$

- What should be our prior belief for θ, $f_\Theta(\theta)$?
- Ideally, we would like our posterior distribution to be from the same family as the prior distribution: conjugate distribution

$$f_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Example

- Bayesian (MAP) estimator:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)$$

$$= \arg\max_{\theta}\left(\ln f_{X|\Theta}(x|\theta) + \ln f_{\Theta}(\theta)\right)$$

$$= \arg\max_{\theta}(n \ln \theta + (m-n)\ln(1-\theta) + (\alpha - 1)\ln \theta$$

Example

$$\hat{\theta}_{MLE} = \frac{n}{m} \qquad \hat{\theta}_{MAP} = \frac{n + \alpha - 1}{n + \beta - 1 + \alpha - 1}$$

- The MAP estimate is equivalent to the ML estimate with $\alpha - 1$ additional heads and $\beta - 1$ additional Tails
- Example: if $\alpha = 7$ and $\beta = 3$ it is as if we had begun the experiment with 6 heads and 2 tails on the record
- Good idea if we initially believed probability of heads was 6/8
- Useful in reducing variance of the estimate for small samples.
- Example: data contains only one coin flip, heads. Then $\hat{\theta}_{MLE}$ =1. However, if we believe the coin is probably fair, then we can assign $\alpha = \beta = 3$ (or any $\alpha = \beta$), and we get $\hat{\theta}_{MAP} = 3/5$

National University of Singapore

Interval Estimation

- In most cases the chance that the point estimate we obtain for a parameter is actually the correct one is zero
- Generalize the idea of point estimation to interval estimation: rather than estimating a single value of a parameter we estimate a region of parameter space
- We make the inference that the parameter of interest lies within the defined region
- The coverage of an interval estimator is the fraction of times the parameter actually lies within the interval
- The idea of interval estimation is intimately linked to the notion of confidence intervals

Acknowledgements

- A number of the slides in this lecture are based on material from various sources:
  - Gil McVean
  - Jeremy Orloff
  - Jonathan Bloom
  - Sebastian Thrun
  - Alex Teichman