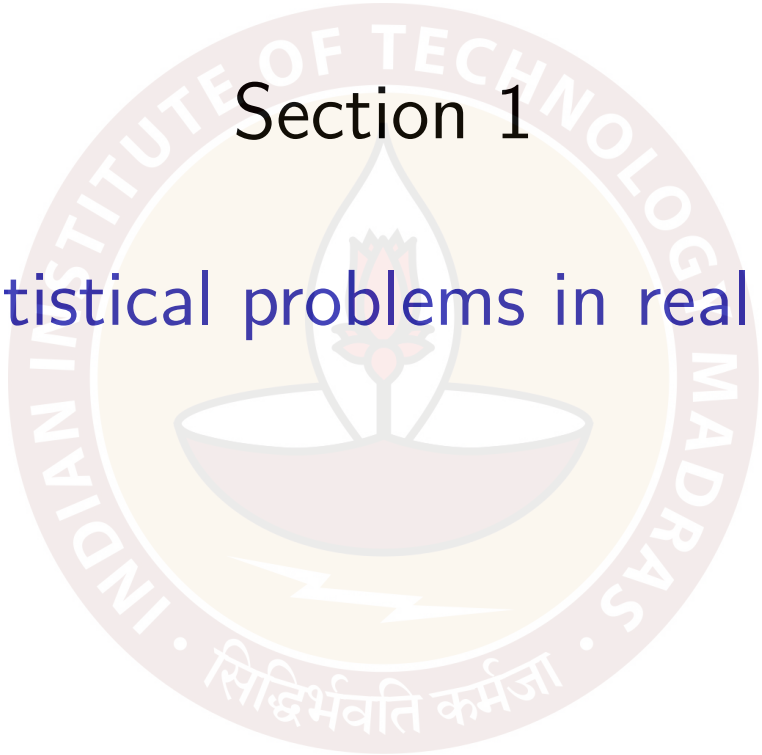
Parameter estimation

Andrew Thangaraj

IIT Madras

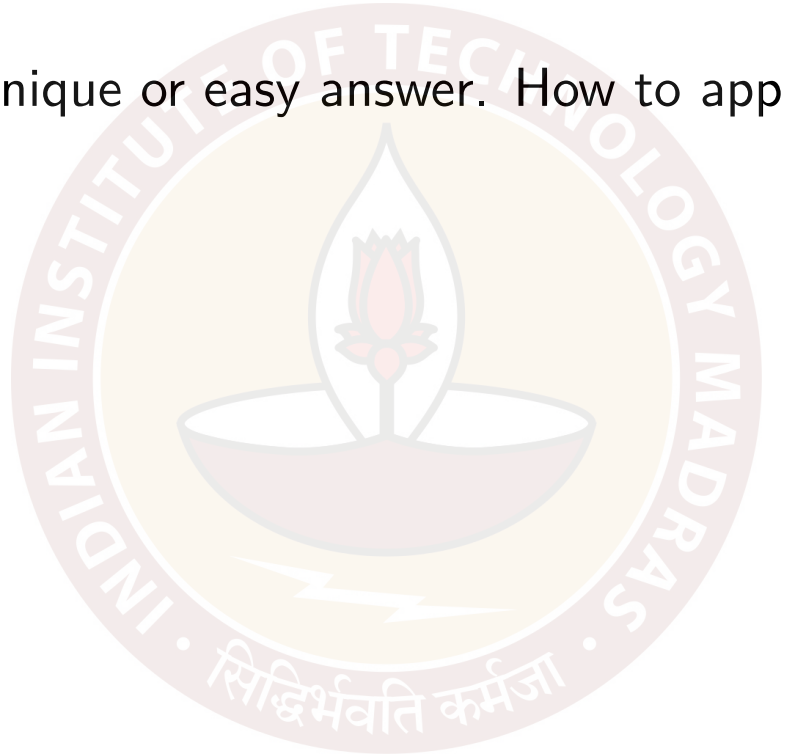


Section 1

Statistical problems in real life

Example 1: Who is the best captain in the IPL?

- Problem and planning
  - What are the qualities of a good captain? How to quantify it?
  - Typically, no unique or easy answer. How to approach?



Example 1: Who is the best captain in the IPL?

- Problem and planning
  - What are the qualities of a good captain? How to quantify it?
  - Typically, no unique or easy answer. How to approach?

Example 1: Who is the best captain in the IPL?

- Problem and planning
  - What are the qualities of a good captain? How to quantify it?
  - Typically, no unique or easy answer. How to approach?
- Data
  - What data is available? Scoresheets. How to collect it and consolidate?
  - Should some sampling of experts or fans be done?

Example 1: Who is the best captain in the IPL?

- Problem and planning
  - What are the qualities of a good captain? How to quantify it?
  - Typically, no unique or easy answer. How to approach?
- Data
  - What data is available? Scoresheets. How to collect it and consolidate?
  - Should some sampling of experts or fans be done?
- Analysis
  - Study data: descriptive stats, histograms, scatter plots
  - Find patterns and fit models or form hypotheses
    - ★ Use statistical procedures to find unknown parameters or test hypothesis
  - Derive metrics that measure captaincy in the IPL

Example 1: Who is the best captain in the IPL?

- Problem and planning
  - What are the qualities of a good captain? How to quantify it?
  - Typically, no unique or easy answer. How to approach?
- Data
  - What data is available? Scoresheets. How to collect it and consolidate?
  - Should some sampling of experts or fans be done?
- Analysis
  - Study data: descriptive stats, histograms, scatter plots
  - Find patterns and fit models or form hypotheses
    - ★ Use statistical procedures to find unknown parameters or test hypothesis
  - Derive metrics that measure captaincy in the IPL
- Conclusion and communication
  - Develop visualizations for communicating results

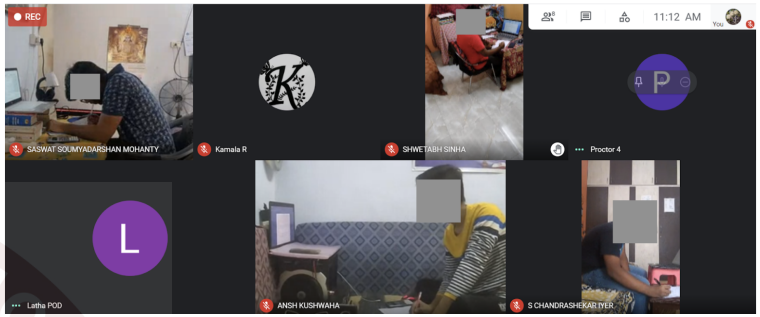
Example 2: How many tigers are there in India?

- National Tiger Conservation Authority (NTCA)
  - Statutory body for strengthening tiger conservation [ntca.gov.in]
- Tiger census
  - Sampling over multiple phases/methods
    - ★ Survey by field forest staff
    - ★ Landscape characterization using satellite and other data
    - ★ Intensive camera traps
- Statistical methods
  - Find relationships between tiger population and various factors
  - Find a joint distribution likelihood model
  - Estimate number of tigers not camera-trapped

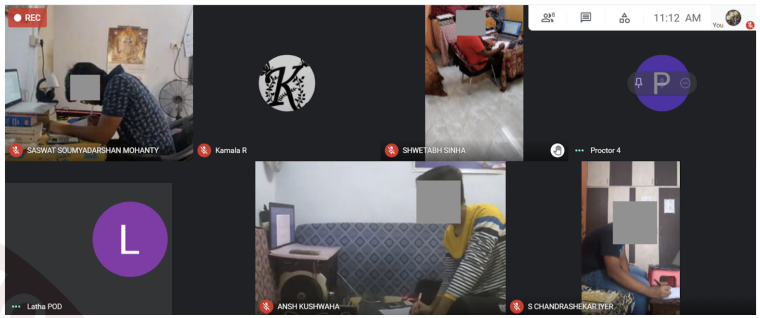
2018: 2461 tigers camera-trapped, 2967 total estimated tigers

Example 3: Was a remote-proctored exam successful?

- Problem and planning
  - How to assess success of exam?
  - Honor code, possible collaboration

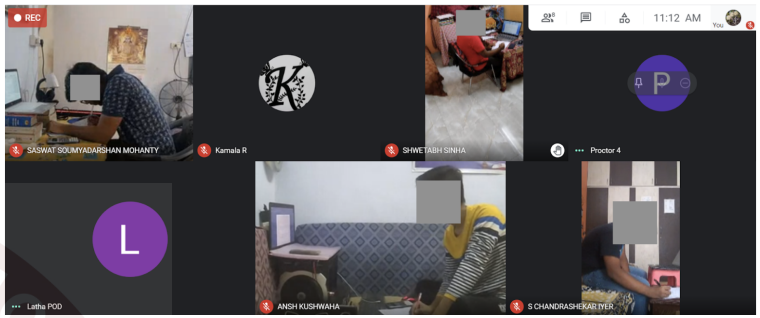


Example 3: Was a remote-proctored exam successful?



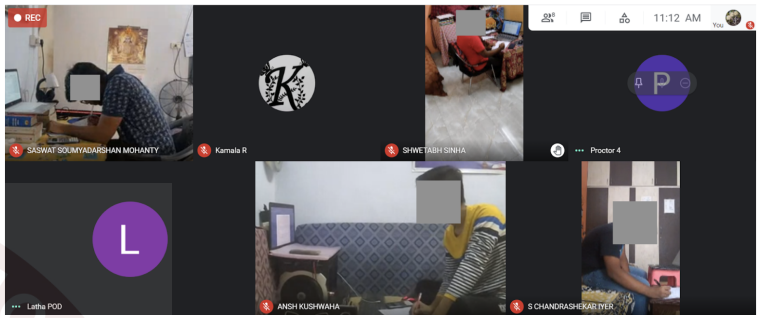
- Problem and planning
  - How to assess success of exam?
  - Honor code, possible collaboration
- Data
  - Scores in online exam
  - Scores in previous in-person exams

Example 3: Was a remote-proctored exam successful?



- Problem and planning
  - How to assess success of exam?
  - Honor code, possible collaboration
- Data
  - Scores in online exam
  - Scores in previous in-person exams
- Analysis
  - Test the hypothesis that "honor code" was violated
  - Estimate number of violations
  - Detect violators or groups of collaborators

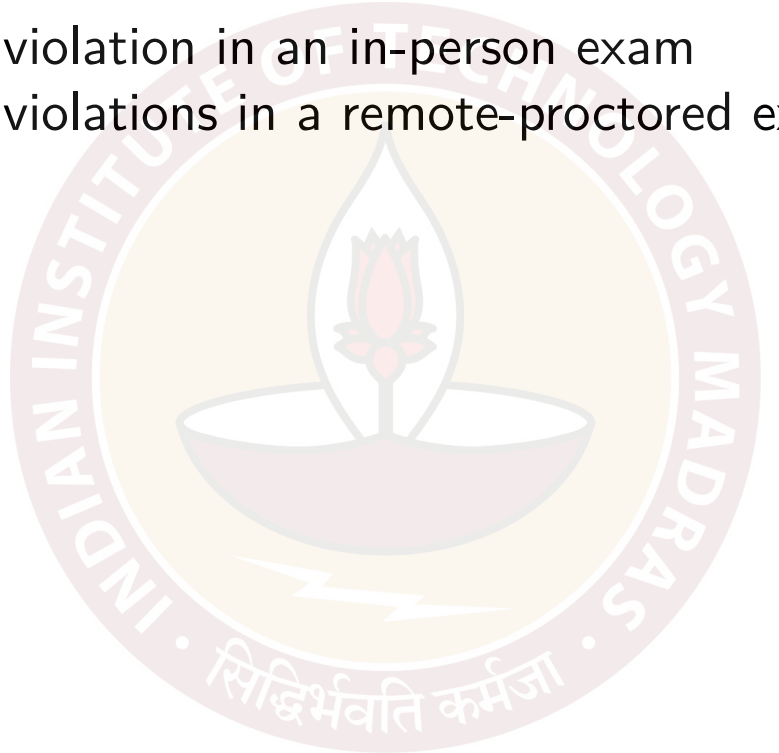
Example 3: Was a remote-proctored exam successful?



- Problem and planning
  - How to assess success of exam?
  - Honor code, possible collaboration
- Data
  - Scores in online exam
  - Scores in previous in-person exams
- Analysis
  - Test the hypothesis that "honor code" was violated
  - Estimate number of violations
  - Detect violators or groups of collaborators
- Conclusion and communication
  - To university authorities
  - To students

The importance of communication

- Consider 1500 students in a course
  - 1 honor code violation in an in-person exam
  - 2 honor code violations in a remote-proctored exam



The importance of communication

- Consider 1500 students in a course
  - 1 honor code violation in an in-person exam
  - 2 honor code violations in a remote-proctored exam
- Communication 1: **100% increase in honor code violations in remote-proctored exams**
- Communication 2: **Honor code violations within 0.15% under remote proctoring**

The importance of communication

- Consider 1500 students in a course
  - 1 honor code violation in an in-person exam
  - 2 honor code violations in a remote-proctored exam
- Communication 1: **100% increase in honor code violations in remote-proctored exams**
- Communication 2: **Honor code violations within 0.15% under remote proctoring**
- You will see such communication in the press and in social media
- Truthful representation of what the data has conveyed is often difficult to find

Summary

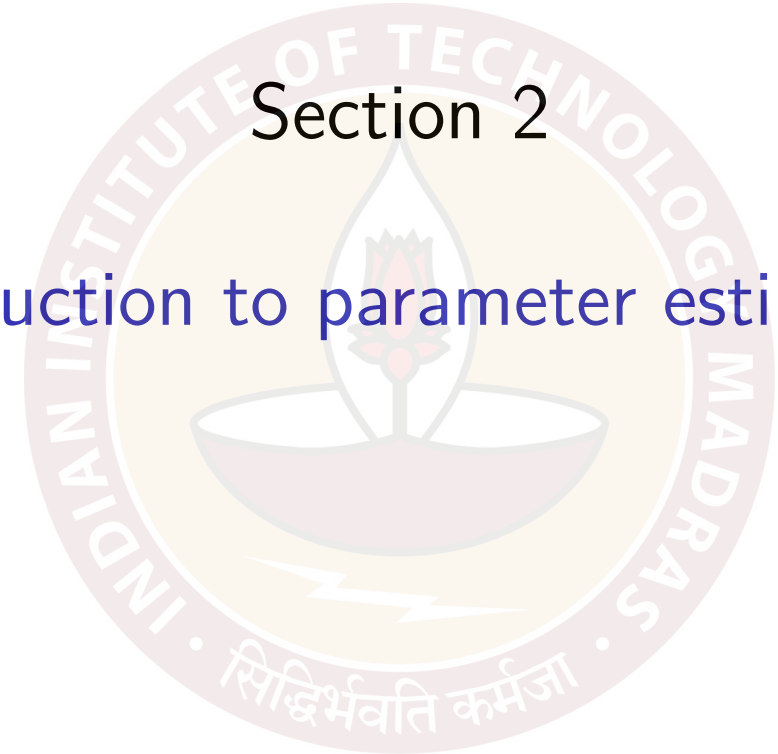
- Statistical problems in real life
  - Usually very complex and involve several competing factors that are difficult to measure
  - Many such questions have very serious consequences
    - ★ Medical problems, societal problems, policy and development issues
  - Majority of time is usually spent on planning, acquiring data and in formulating a clear communication plan

Summary

- Statistical problems in real life
  - Usually very complex and involve several competing factors that are difficult to measure
  - Many such questions have very serious consequences
    - ★ Medical problems, societal problems, policy and development issues
  - Majority of time is usually spent on planning, acquiring data and in formulating a clear communication plan
- In this course
  - We will focus on the "Analysis" part
  - Analysis: involves well-defined statistical procedures assuming an iid sample model for available data
  - Estimation of unknown parameters in
    - ★ probabilistic models formulated for data
    - ★ relationship models between factors
  - Testing of hypothesis using data

Books

- Our textbook for the course
- Mathematical Statistics and Data Analysis by John A. Rice, CENGAGE learning
  - Good reference with all theory and equations
  - Has lots of data and practical examples
- The Art of Statistics (Learning from Data) by David Spiegelhalter, Pelican
  - A popular book about the entire statistical approach to problem solving and understanding phenomenon
  - No equations
  - Emphasis on all aspects of the underlying problem



Section 2

Introduction to parameter estimation

Illustrative example 1: Bernoulli(p) trials

- Setting
  - n Bernoulli(p) trials with p unknown
  - One set of samples ($n = 10$): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - Can you guess p?

Illustrative example 1: Bernoulli(p) trials

- Setting
  - n Bernoulli(p) trials with p unknown
  - One set of samples ($n = 10$): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - Can you guess p?
- Above is an example of a simple "parameter estimation" problem
  - Result of Bernoulli trial is a random variable: $X \sim \{0, 1\}$
    - ★ Distribution of X: $\text{Prob}(X = 0) = 1 - p$, $\text{Prob}(X = 1) = p$
    - ★ p: a parameter of the distribution
  - We observe a certain number of iid samples from the distribution
    - ★ Using the observed samples, we are required to estimate a parameter

Illustrative example 2: Emissions of alpha particles

- Number of particles N emitted in a 10 sec interval
  - Modelled as Poisson: $P(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}$
  - λ is a parameter: average number of particles emitted

Illustrative example 2: Emissions of alpha particles

- Number of particles N emitted in a 10 sec interval
  - Modelled as Poisson: $P(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}$
  - λ is a parameter: average number of particles emitted
- Samples from one observation

n	Observed	Poisson fit
0-2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	132.7
8	164	166.9
9	161	155.6

n	Observed	Poisson fit
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1

Illustrative example 2: Emissions of alpha particles

- Number of particles N emitted in a 10 sec interval
  - Modelled as Poisson: $P(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}$
  - λ is a parameter: average number of particles emitted
- Samples from one observation

n	Observed	Poisson fit
0-2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	132.7
8	164	166.9
9	161	155.6

n	Observed	Poisson fit
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1

- Parameter esimation: What is λ?

Illustrative example 3: Noise in electronic circuits

- Voltage or current measured in circuits will show random fluctuations
  - The same voltage measured at different times will give slightly different values
- Popular model for such voltages or currents in circuits

$$X \sim \text{Normal}(\mu, \sigma^2)$$

- Two parameters
  - μ: a parameter (average voltage or current)
  - σ^2: another parameter (variance)

Illustrative example 3: Noise in electronic circuits

- Voltage or current measured in circuits will show random fluctuations
  - The same voltage measured at different times will give slightly different values
- Popular model for such voltages or currents in circuits

$$X \sim \text{Normal}(\mu, \sigma^2)$$

- Two parameters
  - μ: a parameter (average voltage or current)
  - σ^2: another parameter (variance)
- 10 measurements: 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
- Parameter estimation: What is μ? What is σ?

Parameter estimation

- iid samples

$$X_1, \ldots, X_n \sim \text{iid } X$$

- X has a distribution described by some parameters θ_1, θ_2, $\ldots$
  - Parameters take real values, $\theta_i \in \mathbb{R}$
- Parameter estimation: What is θ_1? What is θ_2? $\cdots$

Parameter estimation

- iid samples

$$X_1, \ldots, X_n \sim \text{iid } X$$

- X has a distribution described by some parameters θ_1, θ_2, ...
  - Parameters take real values, $\theta_i \in \mathbb{R}$
- Parameter estimation: What is θ_1? What is θ_2? $\cdots$
- Estimator for a parameter θ
  - Function of the samples: $\hat{\theta}(X_1, \ldots, X_n)$
  - Notation: $\hat{\theta}$ is an estimator for parameter θ

Parameter estimation

- iid samples
$$X_1, \ldots, X_n \sim \text{iid } X$$
- X has a distribution described by some parameters θ_1, θ_2, ...
  - Parameters take real values, $\theta_i \in \mathbb{R}$
- Parameter estimation: What is θ_1? What is θ_2? $\cdots$
- Estimator for a parameter θ
  - Function of the samples: $\hat{\theta}(X_1, \ldots, X_n)$
  - Notation: $\hat{\theta}$ is an estimator for parameter θ
- Parameter vs estimator
  - θ: constant parameter, not a random variable
  - $\hat{\theta}$: function of n random variables; therefore, it is a random variable
    - ★ $\hat{\theta}$ will have a distribution, PMF or PDF

Example: Bernoulli(p) trials

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Parameter: p
  - We had denoted this as θ earlier

Example: Bernoulli(p) trials

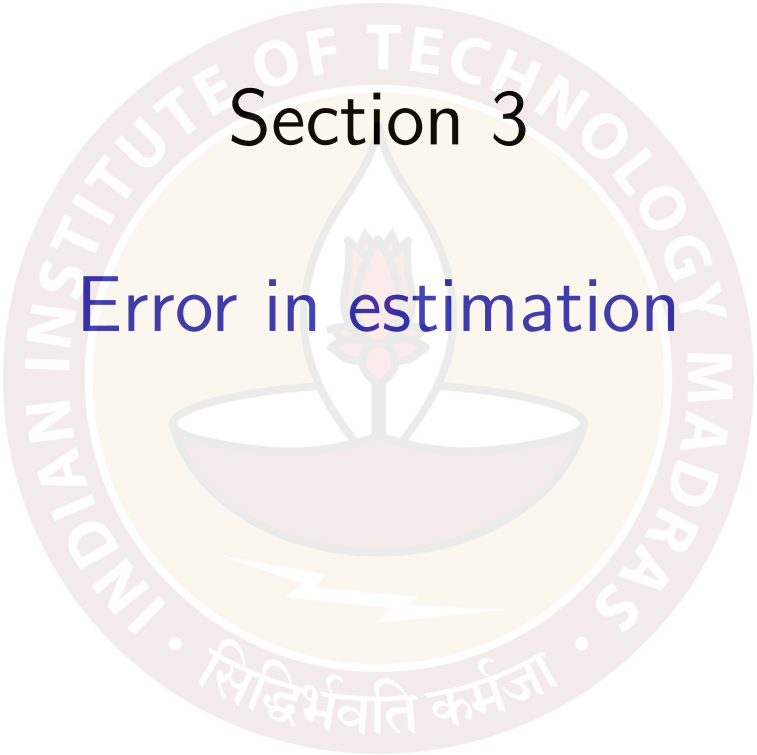
$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Parameter: p
  - We had denoted this as θ earlier
- Estimator 1: $\hat{p}_1 = 1/2$
- Estimator 2: $\hat{p}_2 = (X_1 + X_2)/2$
- Estimator 3: $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$

Example: Bernoulli(p) trials

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Parameter: p
  - We had denoted this as θ earlier
- Estimator 1: $\hat{p}_1 = 1/2$
- Estimator 2: $\hat{p}_2 = (X_1 + X_2)/2$
- Estimator 3: $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$
- An infinite number of estimators are possible
  - How to characterize *good* estimators?
  - How to design estimators?

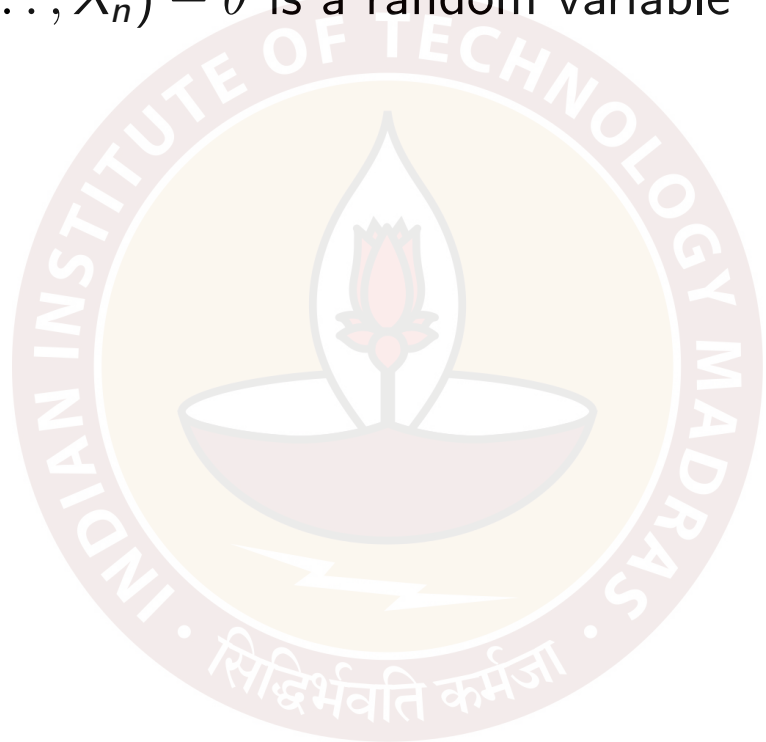


Section 3

Error in estimation

Estimation error

- θ: parameter, $\hat{\theta}(X_1, \ldots, X_n)$: estimator
  - Error: $\hat{\theta}(X_1, \ldots, X_n) - \theta$ is a random variable



Estimation error

- θ: parameter, $\hat{\theta}(X_1, \ldots, X_n)$: estimator
  - Error: $\hat{\theta}(X_1, \ldots, X_n) - \theta$ is a random variable
- We expect the estimator random variable $\hat{\theta}(X_1, \ldots, X_n)$ to take values around the actual value of the parameter θ. So, the random variable 'Error' should take values close to 0.
  - How to express this mathematically? $P(|\text{Error}| > \delta)$ should be small

Estimation error

- θ: parameter, $\hat{\theta}(X_1, \ldots, X_n)$: estimator
  - Error: $\hat{\theta}(X_1, \ldots, X_n) - \theta$ is a random variable
- We expect the estimator random variable $\hat{\theta}(X_1, \ldots, X_n)$ to take values around the actual value of the parameter θ. So, the random variable 'Error' should take values close to 0.
  - How to express this mathematically? $P(|\text{Error}| > \delta)$ should be small
- Parameter will be in a certain range, and estimator error should be low over the entire range
  - How to quantify 'low'?
  - Example: In Bernoulli(p) trials, $p \in [0, 1]$, and the same estimator has to give low error for all values of p
    - ★ What is 'low'? |Error| should be small compared to p.
    - ★ 10% or lower error: $|\text{Error}| \leq p/10$

Example: Bernoulli(p) trials

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$, $\hat{p}_2 = (X_1 + X_2)/2$, $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$

Example: Bernoulli(p) trials

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$, $\hat{p}_2 = (X_1 + X_2)/2$, $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$
- Variation in samples $=>$ variation in estimation
  - 10 samples of Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
    - ★ $\hat{p}_1 = 0.5$, $\hat{p}_2 = 0.5$, $\hat{p}_3 = 0.5$
  - 10 samples in another round: 1, 0, 0, 1, 0, 1, 0, 1, 0, 0
    - ★ $\hat{p}_1 = 0.5$, $\hat{p}_2 = 0.5$, $\hat{p}_3 = 0.4$
  - 10 samples in another round: 1, 1, 0, 0, 0, 1, 0, 1, 0, 1
    - ★ $\hat{p}_1 = 0.5$, $\hat{p}_2 = 1$, $\hat{p}_3 = 0.5$

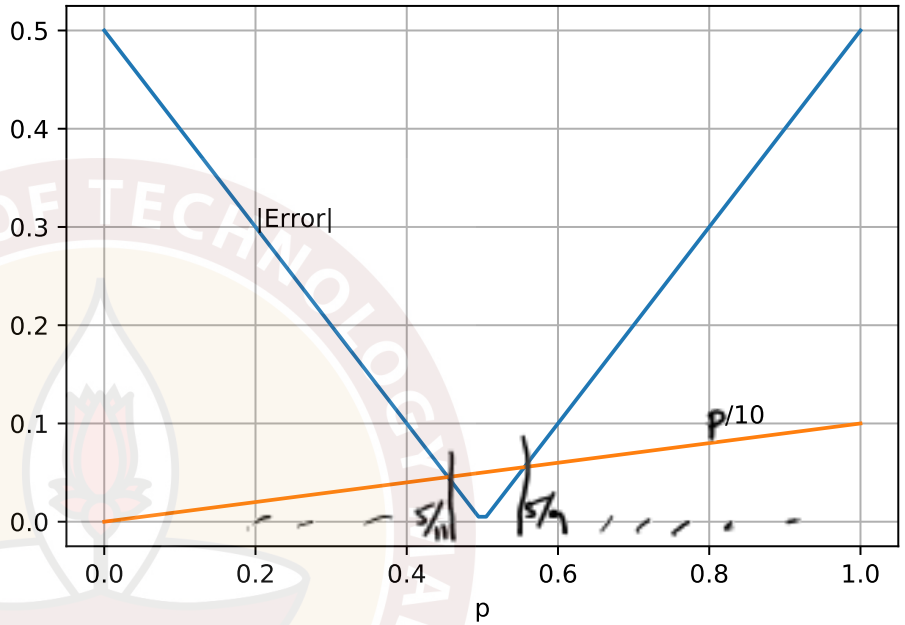
Example: Bernoulli(p) trials

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$, $\hat{p}_2 = (X_1 + X_2)/2$, $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$
- Variation in samples $=>$ variation in estimation
  - 10 samples of Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
    - $\hat{p}_1 = 0.5$, $\hat{p}_2 = 0.5$, $\hat{p}_3 = 0.5$
  - 10 samples in another round: 1, 0, 0, 1, 0, 1, 0, 1, 0, 0
    - $\hat{p}_1 = 0.5$, $\hat{p}_2 = 0.5$, $\hat{p}_3 = 0.4$
  - 10 samples in another round: 1, 1, 0, 0, 0, 1, 0, 1, 0, 1
    - $\hat{p}_1 = 0.5$, $\hat{p}_2 = 1$, $\hat{p}_3 = 0.5$
- $\hat{p}_1$: does not work for all values of p
- $\hat{p}_2$: varies a lot with variation in samples
- $\hat{p}_3$: seems to be promising

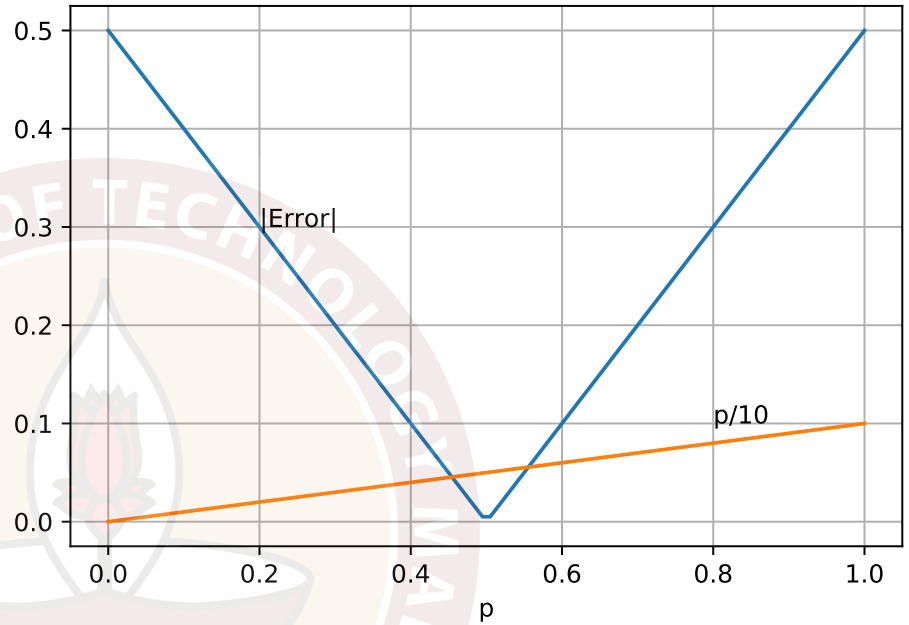
Example: Bernoulli(p) trials (contd)

- $\hat{p}_1 = 1/2$
  - Error: $1/2 - p$
- $P(|\text{Error}| > p/10) = 1$ if $p < 5/11$ or $p > 5/9$



Example: Bernoulli(p) trials (contd)

- $\hat{p}_1 = 1/2$
  - Error: $1/2 - p$
- $P(|\text{Error}| > p/10) = 1$ if $p < 5/11$ or $p > 5/9$



- $\hat{p}_2 = (X_1 + X_2)/2$
  - Error: $\frac{X_1+X_2}{2} - p$
- $P(|\text{Error}| > p/10) = 1$ if $p < 5/11$ or $5/9 < p < 10/11$

x_1	x_2	$e = \frac{x_1+x_2}{2} - p$	$\Pr(\text{Error} = e)$
0	0	$-p$	$(1-p)^2$
0	1	$1/2 - p$	$p(1-p)$
1	0	$1/2 - p$	$p(1-p)$
1	1	$1 - p$	p^2

Example: Bernoulli(p) trials (contd)

- $\hat{p}_3 = (X_1 + \cdots + X_n)/n$
  - Error: $(X_1 + \cdots + X_n)/n - p$

Example: Bernoulli(p) trials (contd)

- $\hat{p}_3 = (X_1 + \cdots + X_n)/n$
  - Error: $(X_1 + \cdots + X_n)/n - p$ is a r.v.
- Recall: Chebyshev bound

$$P(|\text{Error} - \overbrace{E[\text{Error}]}^{0}| > \delta) \leq \frac{\overbrace{\text{Var}(\text{Error})}^{\frac{p(1-p)}{n}}}{\delta^2}$$

Example: Bernoulli(p) trials (contd)

- $\hat{p}_3 = (X_1 + \cdots + X_n)/n$
  - Error: $(X_1 + \cdots + X_n)/n - p$
- Recall: Chebyshev bound

$$P(|\text{Error} - E[\text{Error}]| > \delta) \leq \frac{\text{Var(Error)}}{\delta^2}$$

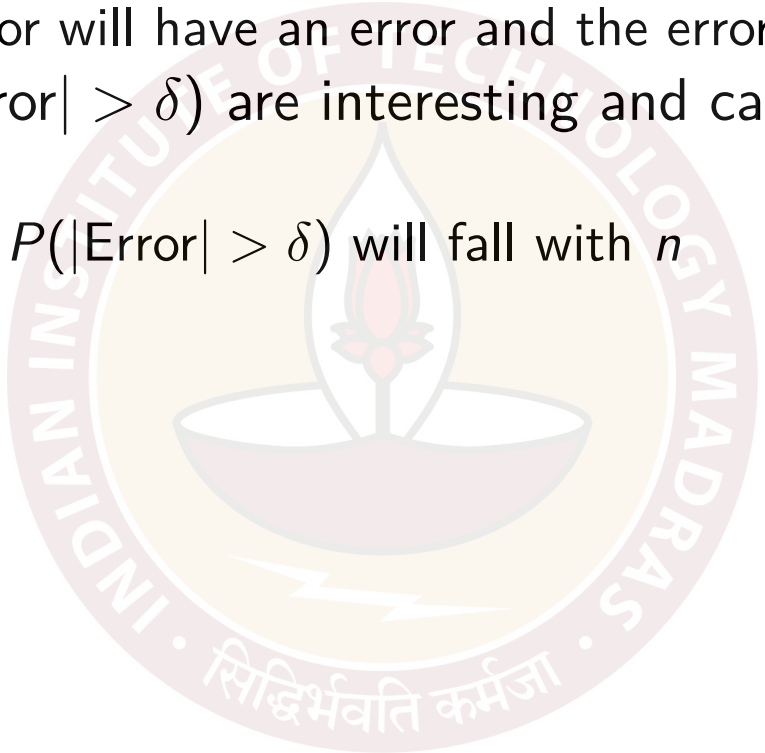
- Using above bound,

$$P(|\text{Error}| > p/10) \leq \frac{p(1-p)/n}{p^2/100} \leq \frac{100(1-p)}{np}$$

- For any fixed p, the above probability tends to 0 as $n \to \infty$
  - Chebyshev bound results in fall of $1/n$
  - Use Chernoff bound or concentration to get exponential fall with n

Observations

- Various estimators are usually possible
  - Every estimator will have an error and the error will have a distribution
- Bounds on $P(|\text{Error}| > \delta)$ are interesting and capture useful properties of the estimator
  - Good design: $P(|\text{Error}| > \delta)$ will fall with n



Observations

- Various estimators are usually possible
  - Every estimator will have an error and the error will have a distribution
- Bounds on $P(|\text{Error}| > \delta)$ are interesting and capture useful properties of the estimator
  - Good design: $P(|\text{Error}| > \delta)$ will fall with n
- Chebyshev bound is a useful tool

$$P(|\text{Error} - E[\text{Error}]| > \delta) \leq \frac{\text{Var}(\text{Error})}{\delta^2}$$

Observations

- Various estimators are usually possible
  - Every estimator will have an error and the error will have a distribution
- Bounds on $P(|\text{Error}| > \delta)$ are interesting and capture useful properties of the estimator
  - Good design: $P(|\text{Error}| > \delta)$ will fall with n
- Chebyshev bound is a useful tool

$$P(|\text{Error} - E[\text{Error}]| > \delta) \leq \frac{\text{Var(Error)}}{\delta^2}$$

- Good design principles
  - $E[\text{Error}]$ should be close to or equal to 0
  - Var(Error)$\rightarrow$ 0 with n

Section 4

Bias, Variance and Risk of an Estimator

Recap: Mean and variance

- Random variable X taking values in a set $\mathcal{X}$
  - Assume discrete with a PMF f_X
- Mean or expected value of X
  - $E[X] = \sum_{x \in \mathcal{X}} x f_X(x)$
  - Average value
  - Denoted μ
- Second moment: $E[X^2] = \sum_{x \in \mathcal{X}} x^2 f_X(x)$
- Variance: $\text{Var}(X) = E[(X - \mu)^2]$
  - $\text{Var}(X) = E[X^2] - \mu^2$ or $E[X^2] = \text{Var}(X) + \mu^2$
  - Spread of the distribution
- Low variance: random variable takes values around μ

Continuous with PDF $f_X(x)$

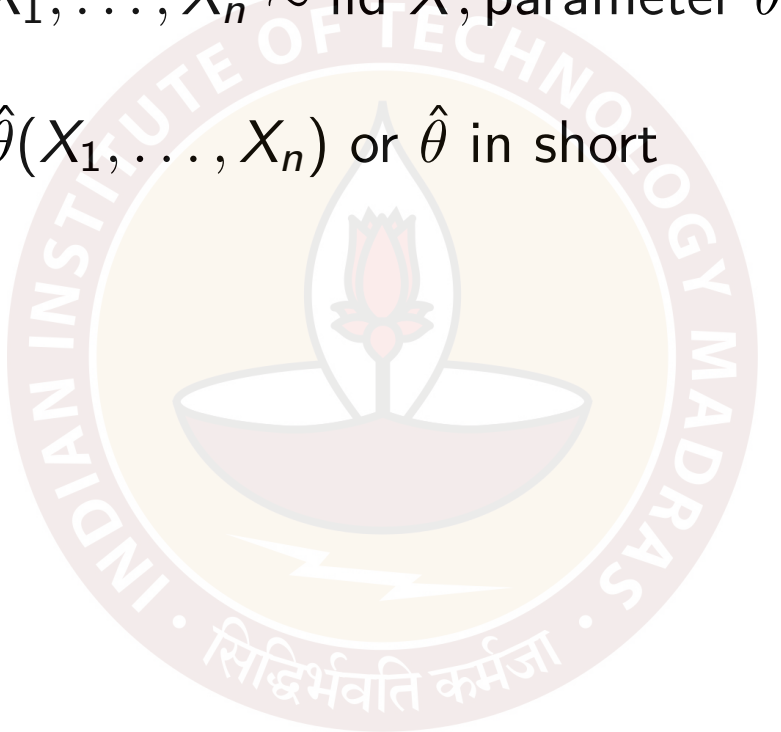
$E[X] = \int x f_X(x)\, dx$

$E[X^2] = \int x^2 f_X(x)\, dx$

Bias

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short



Bias

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Definition (Bias)

The bias of the estimator $\hat{\theta}$ for a parameter θ, denoted $\text{Bias}(\hat{\theta}, \theta)$ is defined as

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

Bias

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Definition (Bias)

The bias of the estimator $\hat{\theta}$ for a parameter θ, denoted $\text{Bias}(\hat{\theta}, \theta)$ is defined as

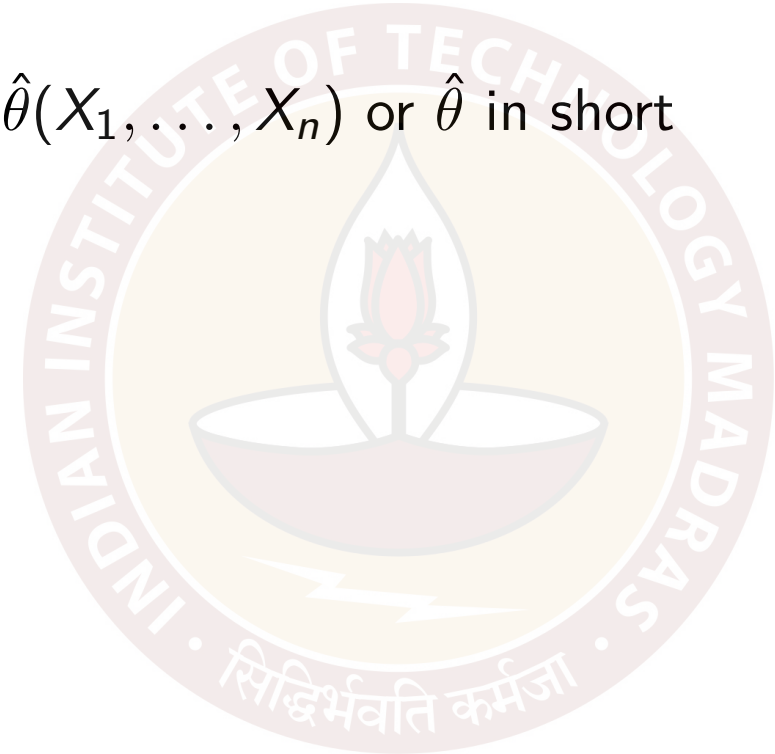
$$\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta.$$

- Since Error $= \hat{\theta} - \theta$, bias is the expected value of Error
- An estimator with bias equal to 0 is said to be an *unbiased* estimator

Risk (squared error)

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short



Risk (squared error)

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Definition (Risk)

The (squared-error) risk of the estimator $\hat{\theta}$ for a parameter θ, denoted $\text{Risk}(\hat{\theta}, \theta)$, is defined as

$$\text{Risk}(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2].$$

Risk (squared error)

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Definition (Risk)

The (squared-error) risk of the estimator $\hat{\theta}$ for a parameter θ, denoted $\text{Risk}(\hat{\theta}, \theta)$, is defined as

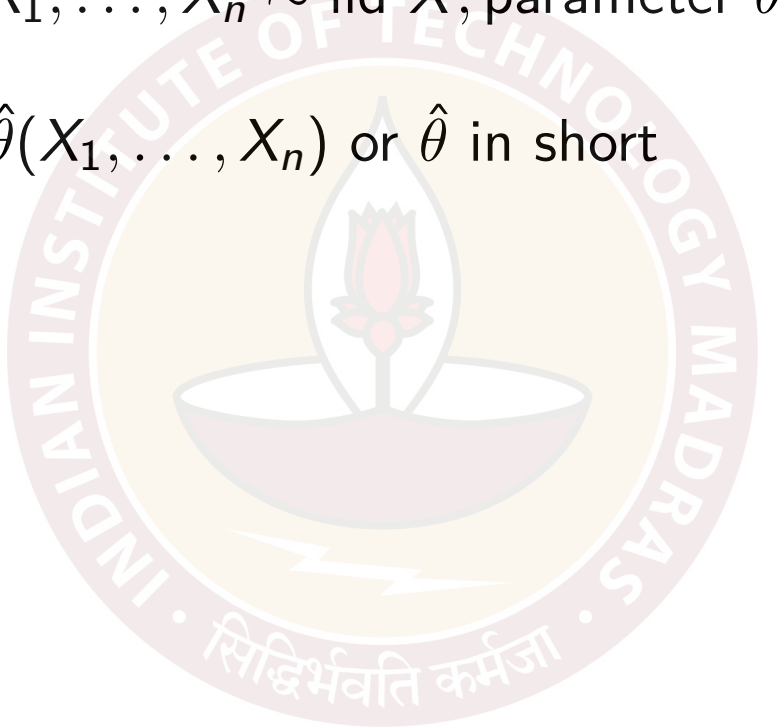
$$\text{Risk}(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2].$$

- Since Error $= \hat{\theta} - \theta$, risk is the expected value of "squared error" and is also called mean squared error (MSE) often
- Squared-error risk is the second moment of Error

Variance

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short



Variance

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short
- Variance of estimator

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

Variance

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short
- Variance of estimator

$$\text{Var}(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

- Variance of error: $\text{Error} = \hat{\theta} - \theta$
  - Error is a "translated" version of the estimator $\hat{\theta}$
  - Remember: θ is a constant

$$\text{Var}(\text{Error}) = \text{Var}(\hat{\theta})$$

Bias-variance tradeoff

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Theorem (Bias-variance tradeoff)

The risk of the estimator satisfies the following relationship:

$$Risk(\hat{\theta}, \theta) = Bias(\hat{\theta}, \theta)^2 + Var(\hat{\theta})$$

Bias-variance tradeoff

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Theorem (Bias-variance tradeoff)

The risk of the estimator satisfies the following relationship:

$$Risk(\hat{\theta}, \theta) = Bias(\hat{\theta}, \theta)^2 + Var(\hat{\theta})$$

- Expanded form

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

Bias-variance tradeoff

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter } \theta$$

- Estimator for θ: $\hat{\theta}(X_1, \ldots, X_n)$ or $\hat{\theta}$ in short

Theorem (Bias-variance tradeoff)

The risk of the estimator satisfies the following relationship:

$$Risk(\hat{\theta}, \theta) = Bias(\hat{\theta}, \theta)^2 + Var(\hat{\theta})$$

- Expanded form

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

- Proof
  - Risk = E[Error^2] = Mean[Error]^2 + Var[Error]

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$
  - Bias $= 1/2 - p$, Variance $= 0$, Risk $= (1/2 - p)^2$

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$
  - Bias $= 1/2 - p$, Variance $= 0$, Risk $= (1/2 - p)^2$
- $\hat{p}_2 = (X_1 + X_2)/2$
  - Bias $= 0$
  - Variance $= \frac{1}{4}(\text{Var}(X_1) + \text{Var}(X_2)) = p(1-p)/2$
  - Risk $= p(1-p)/2$

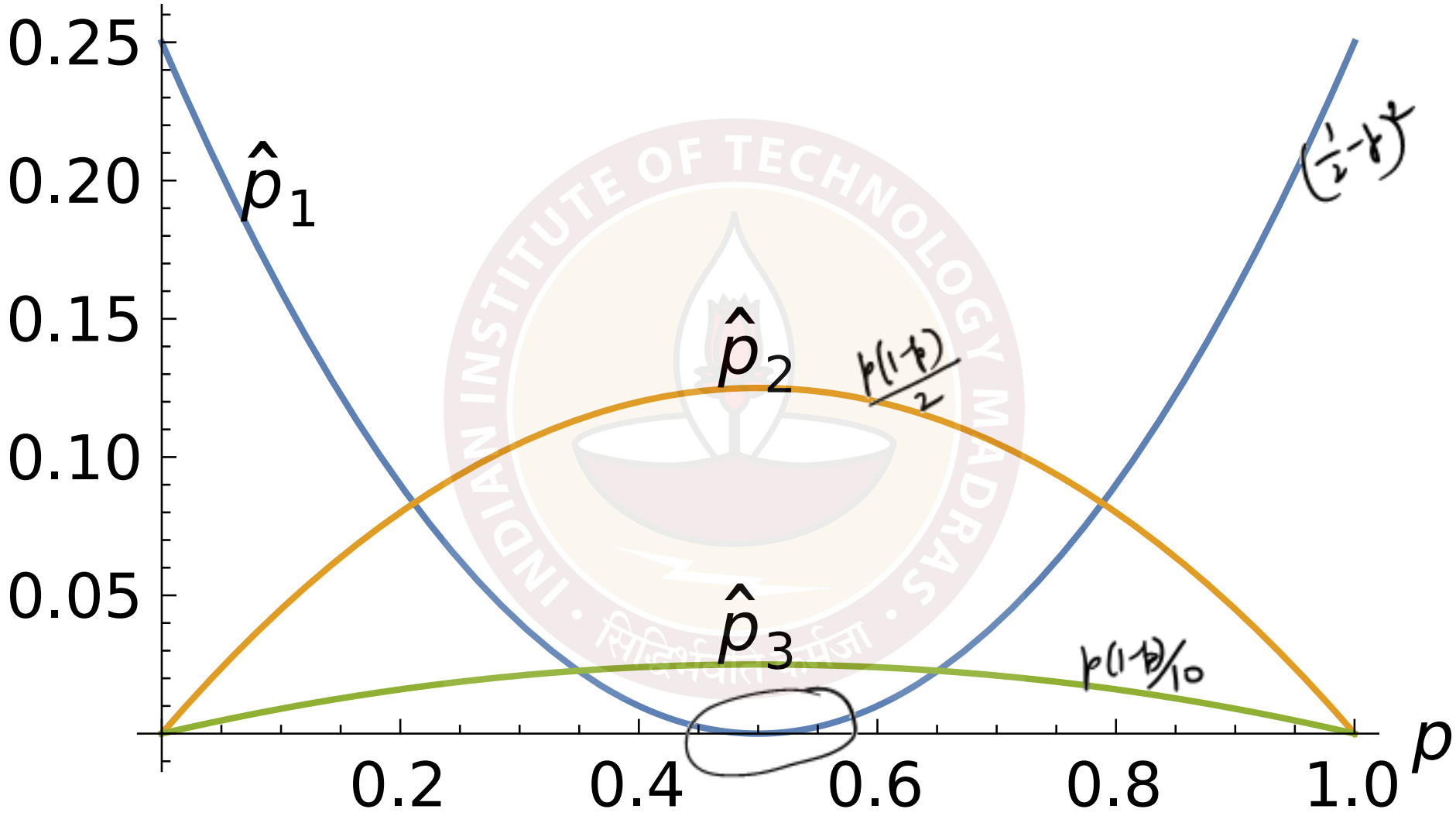
$$E[\hat{p}_2] = E\left[\frac{X_1+X_2}{2}\right] = \frac{1}{2}E[X_1] + \frac{1}{2}E[X_2] = p$$

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $\hat{p}_1 = 1/2$
  - Bias $= 1/2 - p$, Variance $= 0$, Risk $= (1/2 - p)^2$
- $\hat{p}_2 = (X_1 + X_2)/2$
  - Bias $= 0$
  - Variance $= \frac{1}{4}(\text{Var}(X_1) + \text{Var}(X_2)) = p(1-p)/2$
  - Risk $= p(1-p)/2$
- $\hat{p}_3 = (X_1 + X_2 + \cdots + X_n)/n$ $\quad E[\hat{p}_3] = p$
  - Bias $= 0$
  - Variance $= \frac{1}{n^2}(\text{Var}(X_1) + \cdots + \text{Var}(X_n)) = p(1-p)/n$ ($p(1-p)$, $p(1-p)$)
  - Risk $= p(1-p)/n$ $\to$ falling with n ($\leq 1/4$)

Plot of Risk versus p, $n = 10$



Problem: Computing bias, variance, risk

Let $X_1, \ldots, X_n \sim$ iid Bernoulli(p). Consider the estimator

$$\hat{p} = \frac{X_1 + \cdots + X_n + \sqrt{n}/2}{n + \sqrt{n}}.$$

Find the bias, variance and risk of $\hat{p}$.

$$E[\hat{p}] = \frac{np + \sqrt{n}/2}{n+\sqrt{n}} \qquad \text{Bias} = \frac{np+\sqrt{n}/2}{n+\sqrt{n}} - p = \frac{\sqrt{n}/2 - p\sqrt{n}}{n+\sqrt{n}} = \frac{\sqrt{n}(1/2 - p)}{n+\sqrt{n}}$$

$$\text{Var}(\hat{p}) = \frac{1}{(n+\sqrt{n})^2}\left(np(1-p)\right)$$

squared

$$\text{Risk} = \frac{n(1/2-p)^2}{(n+\sqrt{n})^2} + \frac{np(1-p)}{(n+\sqrt{n})^2} = \frac{n}{(n+\sqrt{n})^2}\left(\frac{1}{4} - p + p^2 + p - p^2\right)$$

$$= \frac{n}{4(n+\sqrt{n})^2}$$

Section 5

Estimator design approach: Method of moments

Moments and parameters

$$X \sim f_X(x), \text{parameters } \theta_1, \theta_2, \ldots$$

- Moments $E[X]$, $E[X^2]$, etc can be expressed as functions of the parameters

Moments and parameters

$$X \sim f_X(x), \text{parameters } \theta_1, \theta_2, \ldots$$

- Moments $E[X]$, $E[X^2]$, etc can be expressed as functions of the parameters
- Bernoulli(p)
  - $E[X] = p$
- Poisson(λ)
  - $E[X] = \lambda$
- Exponential(λ)
  - $E[X] = 1/\lambda$

Moments and parameters

$$X \sim f_X(x), \text{parameters } \theta_1, \theta_2, \ldots$$

- Moments $E[X]$, $E[X^2]$, etc can be expressed as functions of the parameters
- Bernoulli(p)
  - $E[X] = p$
- Poisson(λ)
  - $E[X] = \lambda$
- Exponential(λ)
  - $E[X] = 1/\lambda$
- Normal(μ, σ^2)
  - $E[X] = \mu$, $E[X^2] = \mu^2 + \sigma^2$
- Gamma(α, β) $f_X(x) \propto x^{\alpha-1} e^{-\beta x}$
  - $E[X] = \alpha/\beta$,
    $E[X^2] = \alpha^2/\beta^2 + \alpha/\beta^2$

$X \sim \text{Binomial}(N, p)$

$E[X] = Np$ $\quad E[X^2] = (Np)^2 + Np(1-p)$

Moments of samples

$$X_1, \ldots, X_n \sim \text{iid } X$$

- Sample moments

$$M_k(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Moments of samples

$$X_1, \ldots, X_n \sim \text{iid } X$$

- Sample moments

$$M_k(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

- One sampling instance: $x_1, \ldots, x_n$
  - 1st sample moment: $m_1 = \frac{1}{n}(x_1 + \cdots + x_n)$
  - k-th sample moment: $m_k = \frac{1}{n}(x_1^k + \cdots + x_n^k)$

Moments of samples

$$X_1, \ldots, X_n \sim \text{iid } X$$

- Sample moments

$$M_k(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

- One sampling instance: $x_1, \ldots, x_n$
  - 1st sample moment: $m_1 = \frac{1}{n}(x_1 + \cdots + x_n)$
  - k-th sample moment: $m_k = \frac{1}{n}(x_1^k + \cdots + x_n^k)$
- M_k is a random variable, and m_k is the value taken by it in one sampling instance
  - If sampling is repeated, the random variable M_k will take different values
  - We expect that M_k will take values around $E[X^k]$
    - ★ Justified by WLLN, CLT, concentration

Method of moments

- Procedure
  - Equate sample moments to expression for moments in terms of unknown parameters
  - Solve for the unknown parameters

Method of moments

- Procedure
  - Equate sample moments to expression for moments in terms of unknown parameters
  - Solve for the unknown parameters
- One parameter θ usually needs one moment
  - Sample moment: m_1
  - Distribution moment: $E[X] = f(\theta)$
  - Solve for θ from $f(\theta) = m_1$ in terms of m_1
  - $\hat{\theta}$: replace m_1 by M_1 in above solution

Method of moments

- Procedure
  - Equate sample moments to expression for moments in terms of unknown parameters
  - Solve for the unknown parameters
- One parameter θ usually needs one moment
  - Sample moment: m_1
  - Distribution moment: $E[X] = f(\theta)$
  - Solve for θ from $f(\theta) = m_1$ in terms of m_1
  - $\hat{\theta}$: replace m_1 by M_1 in above solution
- Two parameters θ_1, θ_2 usually needs two moments
  - Sample moments: m_1, m_2
  - Distribution moments: $E[X] = f(\theta_1, \theta_2)$, $E[X^2] = g(\theta_1, \theta_2)$
  - Solve for θ_1, θ_2 from $f(\theta_1, \theta_2) = m_1$, $g(\theta_1, \theta_2) = m_2$ in terms of m_1, m_2
  - $\hat{\theta}_1, \hat{\theta}_2$: replace m_1 by M_1 and m_2 by M_2 in above solution

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- $E[X] = p$
- Method of moments equation

$$p = m_1$$

- Estimator

$$\hat{p} = M_1 = \frac{1}{n}(X_1 + \cdots + X_n)$$

Example: Poisson

$$X_1, \ldots, X_n \sim \text{iid Poisson}(\lambda)$$

- $E[X] = \lambda$
- Method of moments equation

$$\lambda = m_1$$

- Estimator

$$\hat{\lambda} = M_1 = \frac{1}{n}(X_1 + \cdots + X_n)$$

Example: Exponential

$$X_1, \ldots, X_n \sim \text{iid Exp}(\lambda)$$

- $E[X] = 1/\lambda$
- Method of moments equation: $1/\lambda = m_1$
- Solution: $\lambda = 1/m_1$
- Estimator

$$\hat{\lambda} = 1/M_1 = \frac{n}{X_1 + \cdots + X_n}$$

Example: Normal

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- $E[X] = \mu$, $E[X^2] = \mu^2 + \sigma^2$
- Method of moments equation: $\mu = m_1$, $\mu^2 + \sigma^2 = m_2$
- Solution: $\mu = m_1$, $\sigma = \sqrt{m_2 - m_1^2}$
- Estimator for μ

$$\hat{\mu} = M_1 = \frac{X_1 + \cdots + X_n}{n}$$

- Estimator for σ

$$\hat{\sigma} = \sqrt{\frac{X_1^2 + \cdots + X_n^2}{n} - \frac{(X_1 + \cdots + X_n)^2}{n^2}}$$

Problem: Gamma

$$X_1, \ldots, X_n \sim \text{iid Gamma}(\alpha, \beta)$$

- $E[X] = \alpha/\beta$, $E[X^2] = \alpha^2/\beta^2 + \alpha/\beta^2$

$\alpha/\beta = m_1$

$\alpha = \beta m_1$

Solve for α and β

$\frac{\alpha^2}{\beta^2} + \frac{\alpha}{\beta^2} = m_2$

$\frac{\beta^2 m_1^2}{\beta^2} + \frac{\beta m_1}{\beta^2} = m_2$

$\frac{m_1}{\beta} = m_2 - m_1^2$

$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2}$

(or) $\beta = \frac{m_1}{m_2 - m_1^2}$

$\hat{\beta} = \frac{m_1}{m_2 - m_1^2}$

$\Rightarrow \alpha = \frac{m_1^2}{m_2 - m_1^2}$

Problem: Binomial(N, p)

$$X_1, \ldots, X_n \sim \text{iid Binomial}(N, p)$$

- $E[X] = Np$, $E[X^2] = N^2p^2 + Np(1-p)$

$Np = m_1 \Rightarrow N = \frac{m_1}{p}$

$N^2p^2 + Np(1-p) = m_2$

$\frac{m_1^2}{p^2} \cdot p^2 + \frac{m_1}{p} \cdot p(1-p) = m_2$

$m_1(1-p) = m_2 - m_1^2$

$1 - p = \frac{m_2 - m_1^2}{m_1}$

$p = 1 - \frac{m_2 - m_1^2}{m_1} = \frac{m_1^2 + (m_1 - m_2)}{m_1}$

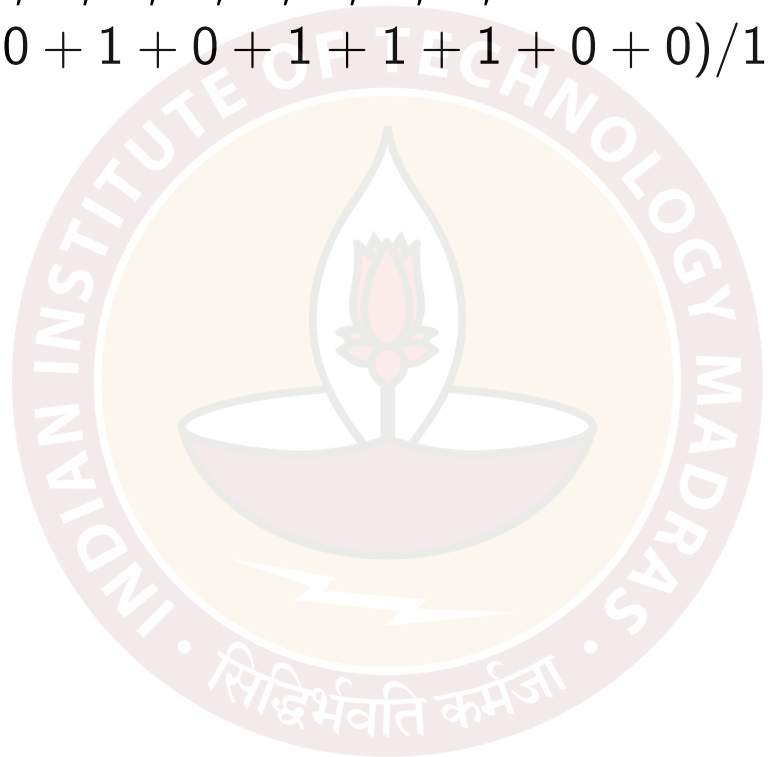
$N = \frac{m_1^2}{m_1^2 + (m_1 - m_2)}$

$\hat{p} = \frac{M_1^2 + (M_1 - M_2)}{M_1}$

$\hat{N} = \frac{M_1^2}{M_1^2 + (M_1 - M_2)}$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$



Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$
- Alpha particles emission in 10sec: Poisson(λ)
  - No of particles emitted per second $= 0.8392$
  - $\hat{\lambda}$ = Average number of particles emitted in 10 seconds $= 8.392$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = \underbrace{(1+0+0+1+0+1+1+1+0+0)/10}_{m_1} = 5/10$
- Alpha particles emission in 10sec: Poisson(λ)
  - No of particles emitted per second $= 0.8392$
  - $\hat{\lambda}$ = Average number of particles emitted in 10 seconds $= 8.392$
- Normal(μ, σ^2): 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
  - $\hat{\mu} = m_1 = (1.07 + 0.91 + \cdots + 1.08)/10 = 1.024$
  - $\hat{\sigma} = \sqrt{m_2 - m_1^2} = \sqrt{1.05482 - 1.024^2} = 0.079$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$
- Alpha particles emission in 10sec: Poisson(λ)
  - No of particles emitted per second $= 0.8392$
  - $\hat{\lambda}$ = Average number of particles emitted in 10 seconds $= 8.392$
- Normal(μ, σ^2): 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
  - $\hat{\mu} = m_1 = (1.07 + 0.91 + \cdots + 1.08)/10 = 1.024$
  - $\hat{\sigma} = \sqrt{m_2 - m_1^2} = \sqrt{1.05482 - 1.024^2} = 0.079$
- Binomial(N, p): 8, 7, 6, 11, 8, 5, 3, 7, 6, 9
  - $\hat{N} = 19$
  - $\hat{p} = 0.371$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$
- Alpha particles emission in 10sec: Poisson(λ)
  - No of particles emitted per second $= 0.8392$
  - $\hat{\lambda}$ = Average number of particles emitted in 10 seconds $= 8.392$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$
- Alpha particles emission in 10sec: Poisson(λ)
  - No of particles emitted per second $= 0.8392$
  - $\hat{\lambda}$ = Average number of particles emitted in 10 seconds $= 8.392$
- Normal(μ, σ^2): 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
  - $\hat{\mu} = m_1 = (1.07 + 0.91 + \cdots + 1.08)/10 = 1.024$
  - $\hat{\sigma} = \sqrt{m_2 - m_1^2} = \sqrt{1.05482 - 1.024^2} = 0.079$

Method of moments estimation

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $\hat{p} = (1+0+0+1+0+1+1+1+0+0)/10 = 5/10$
- Alpha particles emission in 10sec: Poisson(λ)
  - No of particles emitted per second $= 0.8392$
  - $\hat{\lambda}$ = Average number of particles emitted in 10 seconds $= 8.392$
- Normal(μ, σ^2): 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
  - $\hat{\mu} = m_1 = (1.07 + 0.91 + \cdots + 1.08)/10 = 1.024$
  - $\hat{\sigma} = \sqrt{m_2 - m_1^2} = \sqrt{1.05482 - 1.024^2} = 0.079$
- Binomial(N, p): 8, 7, 6, 11, 8, 5, 3, 7, 6, 9
  - $\hat{N} = 19$
  - $\hat{p} = 0.371$

Section 6

Estimator design approach: Maximum likelihood

Likelihood of *iid* samples

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- $f_X(x)$: depends on $\theta_1, \theta_2, \ldots$
  - To bring this out, we will write $f_X(x)$ as $f_X(x; \theta_1, \theta_2, \ldots)$
  - Example: Normal(μ, σ^2)
    - ★ $f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

Likelihood of *iid* samples

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- $f_X(x)$: depends on $\theta_1, \theta_2, \ldots$
  - To bring this out, we will write $f_X(x)$ as $f_X(x; \theta_1, \theta_2, \ldots)$
  - Example: Normal(μ, σ^2)
    - ★ $f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$
- Likelihood of a sampling $x_1, x_2, \ldots, x_n$, denoted $L(x_1, \ldots, x_n)$

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

Likelihood of *iid* samples

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- $f_X(x)$: depends on $\theta_1, \theta_2, \ldots$
  - To bring this out, we will write $f_X(x)$ as $f_X(x; \theta_1, \theta_2, \ldots)$
  - Example: Normal(μ, σ^2)
    - ★ $f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$
- Likelihood of a sampling $x_1, x_2, \ldots, x_n$, denoted $L(x_1, \ldots, x_n)$

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $L = p(1-p)(1-p)p(1-p)ppp(1-p)(1-p) = p^5(1-p)^5$

#1s in the sample

#0s in the sample

$n - \#1s$

$f_X(x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$

Likelihood of *iid* samples

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- $f_X(x)$: depends on $\theta_1, \theta_2, \ldots$
  - To bring this out, we will write $f_X(x)$ as $f_X(x; \theta_1, \theta_2, \ldots)$
  - Example: Normal(μ, σ^2)
    - ★ $f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$
- Likelihood of a sampling $x_1, x_2, \ldots, x_n$, denoted $L(x_1, \ldots, x_n)$

$$L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

- Bernoulli(p): 1, 0, 0, 1, 0, 1, 1, 1, 0, 0
  - $L = p(1-p)(1-p)p(1-p)ppp(1-p)(1-p) = p^5(1-p)^5$
- Normal(μ, σ^2): 1.07,0.91,0.88,1.07,1.15,1.02,0.99,0.99,1.08,1.08
  - $L = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1.07-\mu)^2/2\sigma^2} \cdots \frac{1}{\sqrt{2\pi}\sigma} e^{-(1.08-\mu)^2/2\sigma^2}$
  - Simplified: $\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{10} e^{-((1.07-\mu)^2+\cdots+(1.08-\mu)^2)/2\sigma^2}$

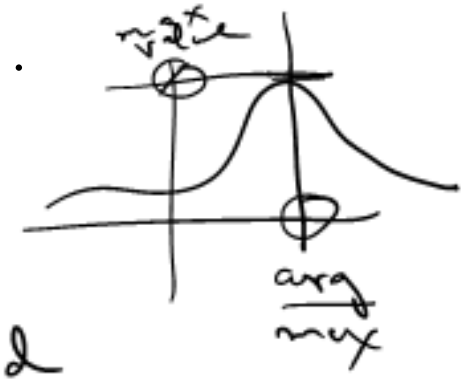
Maximum likelihood (ML) estimator

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- Likelihood $L(x_1, \ldots, x_n)$ is a function of parameters

Maximum likelihood (ML) estimator

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$



- Likelihood $L(x_1, \ldots, x_n)$ is a function of parameters
- Maximum likelihood (ML) estimation
  - Sampling: $x_1, \ldots, x_n$

$$\theta_1^*, \theta_2^*, \ldots = \arg\max_{\theta_1, \theta_2} \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

likelihood

- Find parameters that maximize likelihood for a given set of samples

Maximum likelihood (ML) estimator

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameters: } \theta_1, \theta_2, \ldots$$

- Likelihood $L(x_1, \ldots, x_n)$ is a function of parameters
- Maximum likelihood (ML) estimation
  - Sampling: $x_1, \ldots, x_n$

$$\theta_1^*, \theta_2^*, \ldots = \arg\max_{\theta_1, \theta_2} \prod_{i=1}^{n} f_X(x_i; \theta_1, \theta_2, \ldots)$$

- Find parameters that maximize likelihood for a given set of samples
- When the maximization problem has a closed-form solution, the estimator can be expressed in terms of the samples.
  - This will need a lot of algebraic manipulation.
- In many cases, the maximization problem will need a numerical routine.
  - Several standard modules are available for optimization.

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Samples: $x_1, x_2, \ldots, x_n$
  - $x_i = 0$ or 1

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Samples: $x_1, x_2, \ldots, x_n$
  - $x_i = 0$ or 1
- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$
  - $f_X(x_i) = p$ if $x_i = 1$, or $f_X(x_i) = 1 - p$ if $x_i = 0$
  - Let w denote the number of 1s in the sample

$$L(x_1, \ldots, x_n) = p^w (1-p)^{n-w}$$

Maximize a function: "Differentiate and equate to zero"

→ Maximizing L $\iff$ Maximizing $\log L$

$h(p) = \log L = w \log p + (n-w) \log(1-p)$

$\frac{dh(p)}{dp} = w \frac{1}{p} + (n-w) \frac{1}{1-p}(-1) = 0$

$w(1-p) = (n-w)p$

$p = w/n$

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Samples: $x_1, x_2, \ldots, x_n$
  - $x_i = 0$ or 1
- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$
  - $f_X(x_i) = p$ if $x_i = 1$, or $f_X(x_i) = 1 - p$ if $x_i = 0$
  - Let w denote the number of 1s in the sample

$$L(x_1, \ldots, x_n) = p^w (1-p)^{n-w}$$

- ML estimation: $p^* = \arg\max_p p^w (1-p)^{n-w}$
  - How to find the p that maximizes the above expression?
  - Differentiate w.r.t. p and equate to 0 and solve for p
  - $p^* = w/n = \frac{x_1 + \cdots + x_n}{n}$

$w = \underbrace{x_1 + x_2 + \cdots + x_n}_{\#\text{1s in the sample}}$

Example: Bernoulli(p)

$$X_1, \ldots, X_n \sim \text{iid Bernoulli}(p)$$

- Samples: $x_1, x_2, \ldots, x_n$
  - $x_i = 0$ or 1
- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$
  - $f_X(x_i) = p$ if $x_i = 1$, or $f_X(x_i) = 1 - p$ if $x_i = 0$
  - Let w denote the number of 1s in the sample

$$L(x_1, \ldots, x_n) = p^w (1-p)^{n-w}$$

- ML estimation: $p^* = \arg\max_p p^w (1-p)^{n-w}$
  - How to find the p that maximizes the above expression?
  - Differentiate w.r.t. p and equate to 0 and solve for p
  - $p^* = w/n = \frac{x_1 + \cdots + x_n}{n}$

replace x_i by X_i

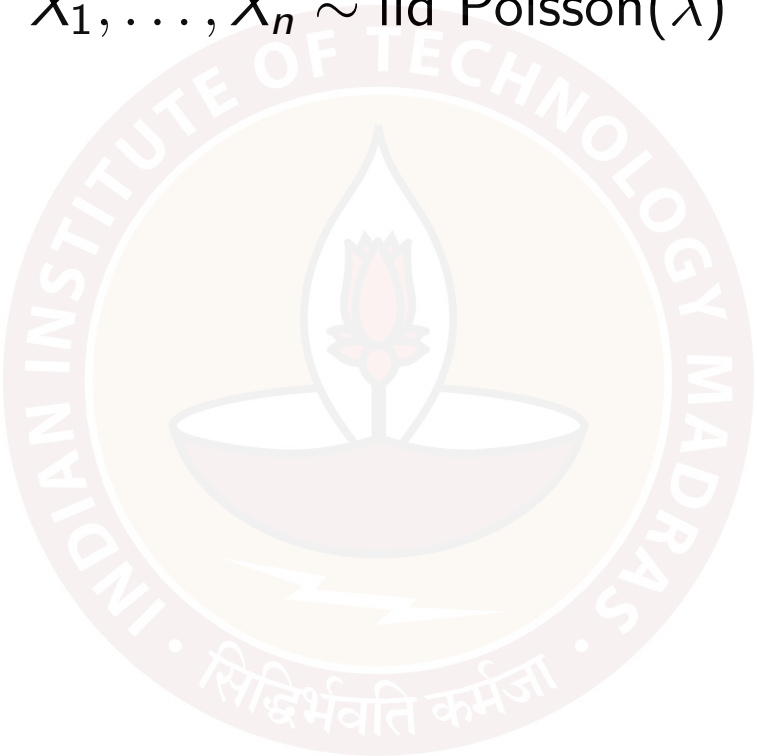
$$\hat{p}_{ML} = \frac{X_1 + \cdots + X_n}{n}$$

Same as MME (method of moments estimator)

Working

Example: Poisson(λ)

$$X_1, \ldots, X_n \sim \text{iid Poisson}(\lambda)$$

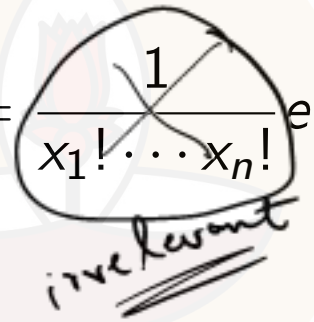


Example: Poisson(λ)

$$X_1, \ldots, X_n \sim \text{iid Poisson}(\lambda)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$

$$L(x_1, \ldots, x_n) = \frac{1}{x_1! \cdots x_n!} e^{-n\lambda} \lambda^{x_1 + \cdots + x_n}$$



Example: Poisson(λ)

$$X_1, \ldots, X_n \sim \text{iid Poisson}(\lambda)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$

$$L(x_1, \ldots, x_n) = \frac{1}{x_1! \cdots x_n!} e^{-n\lambda} \lambda^{x_1 + \cdots + x_n}$$

- ML estimation: $\lambda^* = \arg\max_{\lambda} [(x_1 + \cdots + x_n) \log \lambda - n\lambda]$
  - $\lambda^* = \frac{x_1 + \cdots + x_n}{n}$

Example: Poisson(λ)

$$X_1, \ldots, X_n \sim \text{iid Poisson}(\lambda)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$

$$L(x_1, \ldots, x_n) = \frac{1}{x_1! \cdots x_n!} e^{-n\lambda} \lambda^{x_1 + \cdots + x_n}$$

- ML estimation: $\lambda^* = \arg\max_{\lambda} [(x_1 + \cdots + x_n) \log \lambda - n\lambda]$
  - $\lambda^* = \frac{x_1 + \cdots + x_n}{n}$

$$\hat{\lambda}_{ML} = \frac{X_1 + \cdots + X_n}{n}$$

sample mean

same as MME

Example: Normal(μ, σ^2)

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

Example: Normal(μ, σ^2)

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2}$

$$L(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Example: Normal(μ, σ^2)

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

$$L(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\log(\cdot) \rightarrow -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - n\log\sigma$$

- ML estimation: $\mu^*, \sigma^* = \arg\min_{\mu,\sigma}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + n\log\sigma\right]$

$$\updownarrow$$

$$\arg\max_{\mu,\sigma} -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - n\log\sigma$$

Example: Normal(μ, σ^2)

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- Likelihood: $L(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

$$L(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$x_i^2 - 2\mu x_i + \mu^2$

- ML estimation: $\mu^*, \sigma^* = \arg\min_{\mu,\sigma} \left[\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 + n\log\sigma\right]$

Sample mean same as MME $\longrightarrow$ $\hat{\mu}_{ML} = \frac{X_1 + \cdots + X_n}{n}$ ← diff. w.r.t. μ (treat σ as constant)

$$\hat{\sigma}^2_{ML} = \frac{(X_1 - \hat{\mu}_{ML})^2 + \cdots + (X_n - \hat{\mu}_{ML})^2}{n}$$ ← diff. w.r.t. σ (treat μ as constant)

Sample variance

Observations

- Maximum likelihood is a very popular method for deriving estimators
- Theoretically and intuitively appealing: maximize the probability or likelihood of the observed samples
- Deriving the actual estimator needs some careful calculus
- Numerous questions
  - How do ML estimators look? They seem similar to MME, so far.
  - How does MME compare with ML? How to compare estimators?

Section 7

Finding MME and ML estimators

Problem: Exp(λ)

$$X_1, \ldots, X_n \sim \text{iid Exp}(\lambda) \qquad \lambda e^{-\lambda x}$$

$$\hat{\lambda}_{mme} = \frac{n}{x_1 + \cdots + x_n} \quad \text{(seen before)}$$

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(x_1 + \cdots + x_n)}$$

$$\lambda^* = \arg\max_{\lambda} \left[n \log \lambda - \lambda (x_1 + \cdots + x_n) \right]$$

$$\frac{n}{\lambda} - (x_1 + \cdots + x_n) = 0$$

$$= \frac{n}{x_1 + \cdots + x_n}$$

$$\hat{\lambda}_{ML} = \frac{n}{x_1 + \cdots + x_n}$$

Problem: $\{1, 2, 3\}$ w.p. p_1, p_2, p_3

parameters

$$X_1, \ldots, X_n \sim \text{iid } \{1, 2, 3\} \quad \begin{matrix} p_1 & p_2 & p_3 \end{matrix}$$

$p_3 = 1 - p_1 - p_2$

p_1, p_2: unknown parameters

$m_1 = \bar{X} = \frac{1}{n}\sum X_i$

$m_2 = \frac{1}{n}\sum X_i^2$

MME

$m_1 = p_1 + 2p_2 + 3p_3$

$m_2 = p_1 + 4p_2 + 9p_3$

Put $p_3 = 1 - p_1 - p_2$

$m_1 = p_1 + 2p_2 + 3 - 3p_1 - 3p_2$

$= 3 - 2p_1 - p_2$

$2p_1 + p_2 = 3 - m_1$ — (1)

$m_2 = p_1 + 4p_2 + 9 - 9p_1 - 9p_2$

$= 9 - 8p_1 - 5p_2$

$8p_1 + 5p_2 = 9 - m_2$ — (2)

5×(1) − (2): $2p_1 = 6 - 5m_1 + m_2$

$p_1 = 3 - \frac{5m_1}{2} + \frac{m_2}{2}$

$p_2 = 3 - m_1 - 6 + 5m_1 - m_2$

$= 4m_1 - m_2 - 3$

$\hat{p}_{1,MME} = 3 - \frac{5}{2}m_1 + \frac{m_2}{2}$

$\hat{p}_{2,MME} = 4m_1 - m_2 - 3$

ML

$L = p_1^{\omega_1} p_2^{\omega_2} (1 - p_1 - p_2)^{n - \omega_1 - \omega_2}$

ω_1: #1s, ω_2 = #2s

$p_1^*, p_2^* = \arg\max_{p_1, p_2} \left[\omega_1 \log p_1 + \omega_2 \log p_2 + (n - \omega_1 - \omega_2)\log(1 - p_1 - p_2)\right]$

$p_1^* = \frac{\omega_1}{n}, \; p_2^* = \frac{\omega_2}{n}$

$\hat{p}_{1,ML} = \frac{\#1\text{s in sample}}{n}$

$\hat{p}_{2,ML} = \frac{\#2\text{s in sample}}{n}$

Problem: Uniform$[0, \theta]$

$X_1, \ldots, X_n \sim$ iid Uniform$[0, \theta]$

$$f_X(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & \text{otherwise} \end{cases}$$

MME

$m_1 = \frac{\theta}{2}$

$\theta = 2m_1$

$\hat{\theta}_{MME} = 2m_1 = 2\bar{X}$

$= \frac{2(X_1 + \cdots + X_n)}{n}$

ML

$$L(x_1, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n} \\ 0 \end{cases}$$

if $0 \le x_1, \ldots, x_n \le \theta$

otherwise.

to maximize L, we have to avoid 0.

$\theta \ge \max(x_1, \ldots, x_n)$

$\arg\max_{\theta} \frac{1}{\theta^n}$

⇓

Pick the least possible θ.

$\hat{\theta}_{ML} = \max(X_1, \ldots, X_n)$

$n = 3$: $\frac{1}{2}, \frac{1}{2}, 8$ — bigger

$\hat{\theta}_{MME} = \frac{\frac{1}{2} + \frac{1}{2} + 8}{3} \cdot 2 = 6$

$\hat{\theta}_{ML} = 8$

Problem: Uniform$\{1, 2, \ldots, N\}$

$$X_1, \ldots, X_n \sim \text{iid Uniform}\{1, 2, \ldots, N\}$$

MME

$$m_1 = \frac{1}{N} \cdot 1 + \frac{1}{N} \cdot 2 + \cdots + \frac{1}{N} \cdot N$$

$$= \frac{1}{N}(1 + 2 + \cdots + N)$$

$$= \frac{1}{N} \cdot \frac{N(N+1)}{2}$$

$$= \frac{N+1}{2}$$

$$N = 2m_1 - 1$$

$$\hat{N}_{MME} = 2\bar{X} - 1$$

ML

$$\hat{N}_{ML} = \max(X_1, \ldots, X_n)$$

Problem: Gamma(α, β)

$$X_1, \ldots, X_n \sim \text{iid Gamma}(\alpha, \beta), f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

MME (seen before)

ML

$$L = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i} = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} (x_1 \cdots x_n)^{\alpha-1} e^{-\beta(x_1+\cdots+x_n)}$$

$$\alpha^*, \beta^* = \arg\max_{\alpha,\beta} \; n\alpha \log\beta - n\log\Gamma(\alpha) + (\alpha-1)\log(x_1 \cdots x_n) - \beta \sum_{i=1}^{n} x_i$$

diff. w.r.t. β: $n\alpha \cdot \frac{1}{\beta} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \alpha = \beta \frac{\sum_{i=1}^{n} x_i}{n}$

diff. w.r.t. α: $n\log\beta - \frac{n}{\Gamma(\alpha)}\Gamma'(\alpha) + \log(x_1 \cdots x_n) = 0$

Solve numerically.

$$\frac{\sum_{i=1}^{n} \log x_i}{n} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log\beta$$

Problem: Binomial(N, p)

$$X_1, \ldots, X_n \sim \text{iid Binomial}(N, p)$$

MME (seen before)

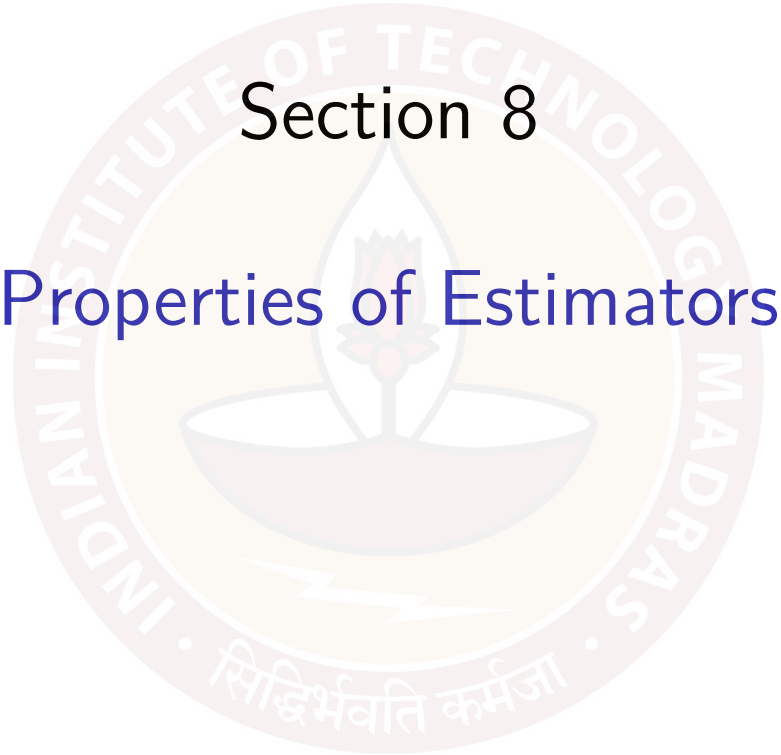
ML

$$L = \prod_{i=1}^{n} \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i} = \binom{N}{x_1} \cdots \binom{N}{x_n} p^{x_1+\cdots+x_n} (1-p)^{Nn-(x_1+\cdots+x_n)}$$

$$\log L = \log \binom{N}{x_1} + \cdots + \log \binom{N}{x_n} + (x_1+\cdots+x_n) \log p + (Nn - (x_1+\cdots+x_n)) \log (1-p) = 0$$

Diff. w.r.t. p: $$Np = \frac{x_1+\cdots+x_n}{n}$$

Diff w.r.t. N: very complicated



Section 8

Properties of Estimators

Consistency of estimators

$$X_1, \ldots, X_n \sim \text{iid } f_X(x;\theta)$$

- Estimator: $\hat{\theta}$, Error $= \hat{\theta} - \theta$

Consistency of estimators

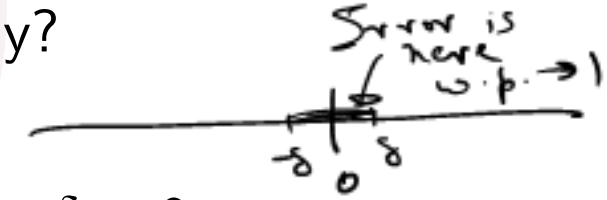
$$X_1, \ldots, X_n \sim \text{iid } f_X(x;\theta)$$

- Estimator: $\hat{\theta}$, Error $= \hat{\theta} - \theta$
- 'Error' is a random variable
- As n increases, we expect 'Error' to take values that are close to zero. How to capture this requirement mathematically?

Consistency of estimators

$$X_1, \ldots, X_n \sim \text{iid } f_X(x;\theta)$$

- Estimator: $\hat{\theta}$, Error $= \hat{\theta} - \theta$
- 'Error' is a random variable
- As n increases, we expect 'Error' to take values that are close to zero. How to capture this requirement mathematically?



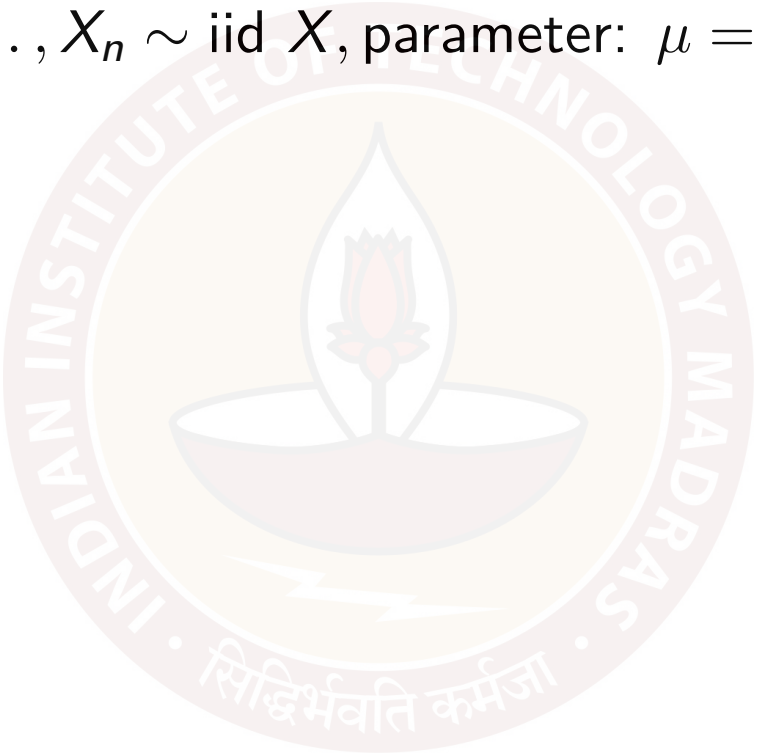
number: depends on 'n'

$$P(|\text{Error}| > \delta) \to 0 \text{ as } n \to \infty \text{ for any } \delta > 0$$

- If an estimator satisfies the above requirement, it is said to be *consistent*
- Technically, the above requirement is called *convergence in probability*

Examples: bias and consistency

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter: } \mu = E[X]$$



Examples: bias and consistency

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter: } \mu = E[X]$$

- Estimator 1: $\hat{\mu}_1 = M_1 = (X_1 + \cdots + X_n)/n$
  - This estimator is unbiased, $E[\hat{\mu}] = \mu$
  - This estimator is consistent (Proof: WLLN)

$P(|M_1 - M| > \delta) \leq$ quantity that $\to 0$ as $n \to \infty$

WLLN

Sample moments $\longrightarrow$ Distribution moments as $n \to \infty$ (in prob.)

Examples: bias and consistency

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter: } \mu = E[X]$$

- Estimator 1: $\hat{\mu}_1 = M_1 = (X_1 + \cdots + X_n)/n$
  - This estimator is unbiased, $E[\hat{\mu}] = \mu$
  - This estimator is consistent (Proof: WLLN)
- Estimator 2: $\hat{\mu}_2 = M_1 = (X_1 + \cdots + X_n)/(n-1)$
  - This estimator is biased, $E[\hat{\mu}] = n\mu/(n-1) \neq \mu$
  - This estimator is consistent (Proof: $\hat{\mu}_2 = \hat{\mu}_1(1 - 1/n)$

Examples: bias and consistency

$$X_1, \ldots, X_n \sim \text{iid } X, \text{parameter: } \mu = E[X]$$

- Estimator 1: $\hat{\mu}_1 = M_1 = (X_1 + \cdots + X_n)/n$
  - This estimator is unbiased, $E[\hat{\mu}] = \mu$
  - This estimator is consistent (Proof: WLLN)
- Estimator 2: $\hat{\mu}_2 = M_1 = (X_1 + \cdots + X_n)/(n-1)$
  - This estimator is biased, $E[\hat{\mu}] = n\mu/(n-1) \neq \mu$
  - This estimator is consistent (Proof: $\hat{\mu}_2 = \hat{\mu}_1(1 - 1/n)$
- Estimator 3: $\hat{\mu}_3 = X_1$
  - This estimator is unbiased, $E[\hat{\mu}] = \mu$
  - This estimator is inconsistent (for most non-trivial distributions)

Estimator designs, properties and comparisons

- Method of moments estimators
  - If parameter is mean or variance, they will be unbiased
  - For most other cases, they may be biased
  - Usually, consistent
    - ★ Sample moments converge to distribution moments
    - ★ If MME estimate is a continuous function of moments, then the estimate converges too.

Estimator designs, properties and comparisons

- Method of moments estimators
  - If parameter is mean or variance, they will be unbiased
  - For most other cases, they may be biased
  - Usually, consistent
    - ★ Sample moments converge to distribution moments
    - ★ If MME estimate is a continuous function of moments, then the estimate converges too.
- Maximum likelihood estimators
  - Consistent
  - Bias vanishes in a limiting sense with growing n
  - Several interesting properties
    - ★ Functional invariance: ML estimate of $g(\theta)$ is $g(\hat{\theta}_{ML})$ for smooth functions g

Estimator designs, properties and comparisons

- Method of moments estimators
  - If parameter is mean or variance, they will be unbiased
  - For most other cases, they may be biased
  - Usually, consistent
    - ★ Sample moments converge to distribution moments
    - ★ If MME estimate is a continuous function of moments, then the estimate converges too.
- Maximum likelihood estimators
  - Consistent
  - Bias vanishes in a limiting sense with growing n
  - Several interesting properties
    - ★ Functional invariance: ML estimate of $g(\theta)$ is $g(\hat{\theta}_{ML})$ for smooth functions g
- How to compare estimators? Squared-error risk or Mean Squared Error (MSE) is one option

MSE or Risk of estimators

- Finding the risk of an estimator theoretically usually involves some calculations of expectations
- Example: $X_1, \ldots, X_n \sim$ iid Uniform$[0, \theta]$
  - $\hat{\theta}_{MME} = 2M_1$
    - ★ Bias $= 0$, Risk $=$ Variance
    - ★ Risk$(\hat{\theta}_{MME}) = \frac{\theta^2}{3n}$
  - $\hat{\theta}_{ML} = \max(X_1, \ldots, X_n)$
    - ★ $f_{\hat{\theta}}(t) = \frac{nt^{n-1}}{\theta^n}$, $E[\hat{\theta}] = \frac{n\theta}{n+1}$, $E[\hat{\theta}^2] = \frac{n\theta^2}{n+2}$, $\text{Var}(\hat{\theta}) = \frac{n\theta^2}{(n+2)(n+1)^2}$
    - ★ Bias $= -\theta/(n+1)$, Risk $= \frac{2\theta^2}{(n+1)(n+2)} \leq \frac{2\theta^2}{n^2}$
  - ML is a factor of $1/n$ better than MME!

MSE or Risk of estimators

- Finding the risk of an estimator theoretically usually involves some calculations of expectations
- Example: $X_1, \ldots, X_n \sim$ iid Uniform$[0, \theta]$
  - $\hat{\theta}_{MME} = 2M_1$
    - ★ Bias $= 0$, Risk $=$ Variance
    - ★ Risk$(\hat{\theta}_{MME}) = \dfrac{\theta^2}{3n}$
  - $\hat{\theta}_{ML} = \max(X_1, \ldots, X_n)$
    - ★ $f_{\hat{\theta}}(t) = \frac{nt^{n-1}}{\theta^n}$, $E[\hat{\theta}] = \frac{n\theta}{n+1}$, $E[\hat{\theta}^2] = \frac{n\theta^2}{n+2}$, Var$(\hat{\theta}) = \frac{n\theta^2}{(n+2)(n+1)^2}$
    - ★ Bias $= -\theta/(n+1)$, Risk $= \dfrac{2\theta^2}{(n+1)(n+2)} \leq \dfrac{2\theta^2}{n^2}$
  - ML is a factor of $1/n$ better than MME!
- A good alternative to theoretical computations is Monte Carlo simulations, which will work in most cases
  - Colab exercise: build a simulation for above and show ML estimator's risk is $1/n$ better than MME

Section 9

Confidence intervals

Example: Surveys and results

- "82% Indians willing to take Covid-19 vaccine," from a *Gallup Survey*
  - **3045** people were called and asked if they were willing to take a vaccine between **Nov 24, 2020 and Jan 8, 2021**
  - Languages spoken: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Odia, Punjabi, Assamese
  - 95% confidence level with a margin of error of 3%

Example: Surveys and results

- "82% Indians willing to take Covid-19 vaccine," from a *Gallup Survey*
  - **3045** people were called and asked if they were willing to take a vaccine between **Nov 24, 2020 and Jan 8, 2021**
  - Languages spoken: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Odia, Punjabi, Assamese
  - 95% confidence level with a margin of error of 3%
- "40% Indians willing to take vaccine," from a *LocalCircles survey*
  - **9628** votes on **Jan 25, 2021** through the LocalCircles app
  - Demographics: people from 299 districts, 48% tier 1, 27% tier 2, 25% rural

Example: Surveys and results

- "82% Indians willing to take Covid-19 vaccine," from a *Gallup Survey*
  - **3045** people were called and asked if they were willing to take a vaccine between **Nov 24, 2020 and Jan 8, 2021**
  - Languages spoken: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Odia, Punjabi, Assamese
  - 95% confidence level with a margin of error of 3%
- "40% Indians willing to take vaccine," from a *LocalCircles survey*
  - **9628** votes on **Jan 25, 2021** through the LocalCircles app
  - Demographics: people from 299 districts, 48% tier 1, 27% tier 2, 25% rural

This lecture: What are confidence level and margin of error?

Estimation of sample mean and confidence interval

$$X_1, \ldots, X_n \sim \text{iid } X, \mu = E[X]$$

- Estimator: $\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$

Estimation of sample mean and confidence interval

$$X_1, \ldots, X_n \sim \text{iid } X, \mu = E[X]$$

- Estimator: $\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$
- Suppose $\Pr(|\hat{\mu} - \mu| < 0.03) = 0.95$.
  - Probability that μ lies in the interval $[\hat{\mu} - 0.03, \hat{\mu} + 0.03]$ is 0.95
  - $[\hat{\mu} - 0.03, \hat{\mu} + 0.03]$: called *95%-confidence interval*
  - $\hat{\mu}$ in one sampling instance: estimate with margin of error 3% at confidence level 95%

Estimation of sample mean and confidence interval

$$X_1, \ldots, X_n \sim \text{iid } X, \mu = E[X]$$

- Estimator: $\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$
- Suppose $\Pr(|\hat{\mu} - \mu| < 0.03) = 0.95$.
  - Probability that μ lies in the interval $[\hat{\mu} - 0.03, \hat{\mu} + 0.03]$ is 0.95
  - $[\hat{\mu} - 0.03, \hat{\mu} + 0.03]$: called *95%-confidence interval*
  - $\hat{\mu}$ in one sampling instance: estimate with margin of error 3% at confidence level 95%
- **Confidence interval** (in general)
  - Suppose $\Pr(|\hat{\mu} - \mu| < \alpha) = \beta$, where α is a small fraction and β is a large fraction
  - $\hat{\mu}$ in one sampling instance: estimate with **margin of error** $(100\alpha)\%$ at confidence level $(100\beta)\%$

How to find α, β for which $\Pr(|X - \mu| < \alpha) = \beta$?

- Suppose X is continuous and has CDF F_X
  - $P(X \leq x) = F_X(x)$

$$P(|X - \mu| < \alpha) = P(\mu - \alpha < X < \mu + \alpha) = F_X(\mu + \alpha) - F_X(\mu - \alpha)$$

$\to 0$, $\alpha = 0$
↑ increases with α

- Given β, find α such that $F_X(\mu + \alpha) - F_X(\mu - \alpha) = \beta$

How to find α, β for which $\Pr(|X - \mu| < \alpha) = \beta$?

- Suppose X is continuous and has CDF F_X
  - $P(X \leq x) = F_X(x)$

$$P(|X - \mu| < \alpha) = P(\mu - \alpha < X < \mu + \alpha) = F_X(\mu + \alpha) - F_X(\mu - \alpha)$$

- Given β, find α such that $F_X(\mu + \alpha) - F_X(\mu - \alpha) = \beta$
- Suppse X is symmetric about the mean,
  i.e. $P(X < \mu - \alpha) = P(X > \mu + \alpha)$

How to find α, β for which $\Pr(|X - \mu| < \alpha) = \beta$?

- Suppose X is continuous and has CDF F_X
  - $P(X \leq x) = F_X(x)$

    $$P(|X - \mu| < \alpha) = P(\mu - \alpha < X < \mu + \alpha) = F_X(\mu + \alpha) - F_X(\mu - \alpha)$$

- Given β, find α such that $F_X(\mu + \alpha) - F_X(\mu - \alpha) = \beta$
- Suppse X is symmetric about the mean,
  i.e. $P(X < \mu - \alpha) = P(X > \mu + \alpha)$
  - $F_X(\mu + \alpha) = 1 - P(X > \mu + \alpha) = 1 - P(X < \mu - \alpha) = 1 - F_X(\mu - \alpha)$

How to find α, β for which $\Pr(|X - \mu| < \alpha) = \beta$?

- Suppose X is continuous and has CDF F_X
  - $P(X \leq x) = F_X(x)$

    $$P(|X - \mu| < \alpha) = P(\mu - \alpha < X < \mu + \alpha) = F_X(\mu + \alpha) - F_X(\mu - \alpha)$$

- Given β, find α such that $F_X(\mu + \alpha) - F_X(\mu - \alpha) = \beta$
- Suppse X is symmetric about the mean,
  i.e. $P(X < \mu - \alpha) = P(X > \mu + \alpha)$
  - $F_X(\mu + \alpha) = 1 - P(X > \mu + \alpha) = 1 - P(X < \mu - \alpha) = 1 - F_X(\mu - \alpha)$
  - $F_X(\mu + \alpha) - F_X(\mu - \alpha) = 1 - 2F_X(\mu - \alpha)$

How to find α, β for which $\Pr(|X - \mu| < \alpha) = \beta$?

- Suppose X is continuous and has CDF F_X
  - $P(X \leq x) = F_X(x)$

    $$P(|X - \mu| < \alpha) = P(\mu - \alpha < X < \mu + \alpha) = F_X(\mu + \alpha) - F_X(\mu - \alpha)$$

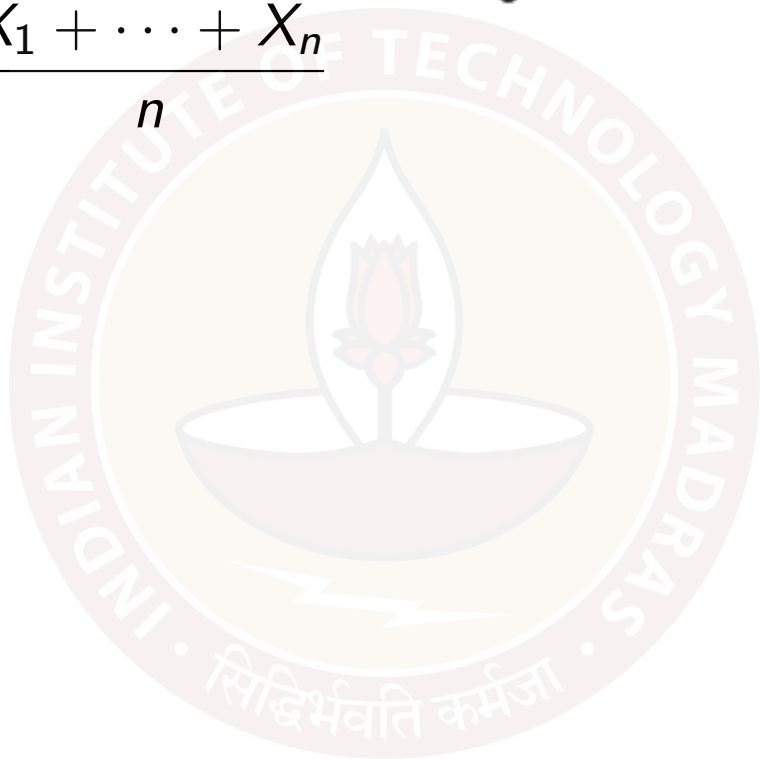
- Given β, find α such that $F_X(\mu + \alpha) - F_X(\mu - \alpha) = \beta$
- Suppse X is symmetric about the mean,
  i.e. $P(X < \mu - \alpha) = P(X > \mu + \alpha)$
  - $F_X(\mu + \alpha) = 1 - P(X > \mu + \alpha) = 1 - P(X < \mu - \alpha) = 1 - F_X(\mu - \alpha)$
  - $F_X(\mu + \alpha) - F_X(\mu - \alpha) = 1 - 2F_X(\mu - \alpha)$
  - Given β, find α s.t. $1 - 2F_X(\mu - \alpha) = \beta$ or $F_X(\mu - \alpha) = (1 - \beta)/2$

Normal samples with known variance

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2), \sigma^2 \text{ known}$$

unknown

- Estimator: $\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$



Normal samples with known variance

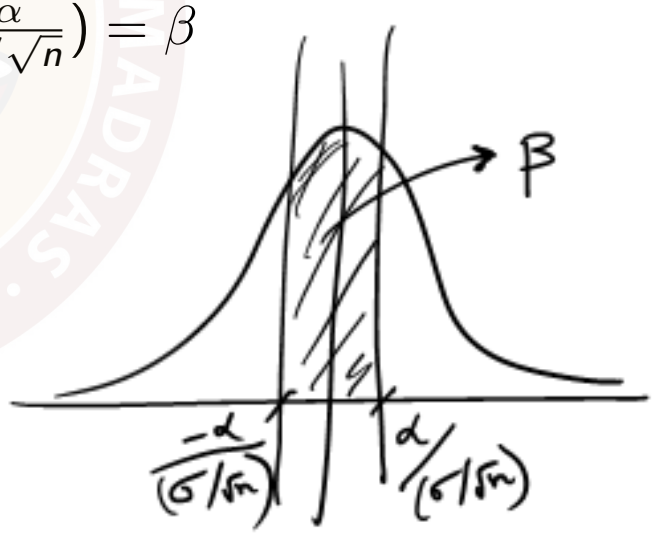
$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2), \sigma^2 \text{ known}$$

- Estimator: $\hat{\mu} = \dfrac{X_1 + \cdots + X_n}{n}$
- $\hat{\mu} \sim \text{Normal}(\mu, \sigma^2/n)$, $Z = \dfrac{\hat{\mu} - \mu}{(\sigma/\sqrt{n})} \sim \text{Normal}(0,1)$
  - $P(|\hat{\mu} - \mu| < \alpha) = \beta \leftrightarrow P(\left|\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}\right| < \frac{\alpha}{\sigma/\sqrt{n}}) = \beta$
  - $\leftrightarrow P(|\text{Normal}(0,1)| < \frac{\alpha}{\sigma/\sqrt{n}}) = \beta.$

Standard ← its CDF is well-known

β	$\frac{\alpha}{\sigma/\sqrt{n}}$
0.68	0.99
0.90	1.64
0.95	1.96

Using CDF of Standard normal



Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and $\sigma = 3$. Find a 95% confidence interval for μ.

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and $\sigma = 3$. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\sigma = 3$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and $\sigma = 3$. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\sigma = 3$
- $\beta = 0.95$, and, using the CDF of Normal(0,1), we get $\frac{\alpha}{\sigma/\sqrt{n}} = 1.96$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and $\sigma = 3$. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\sigma = 3$
- $\beta = 0.95$, and, using the CDF of Normal(0,1), we get $\frac{\alpha}{\sigma/\sqrt{n}} = 1.96$
- $\alpha = 1.96 \times 3/\sqrt{16} = 1.47$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and $\sigma = 3$. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\sigma = 3$
- $\beta = 0.95$, and, using the CDF of Normal(0,1), we get $\frac{\alpha}{\sigma/\sqrt{n}} = 1.96$
- $\alpha = 1.96 \times 3/\sqrt{16} = 1.47$
  - $P(|\hat{\mu} - \mu| < 1.47) = 0.95$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and $\sigma = 3$. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\sigma = 3$
- $\beta = 0.95$, and, using the CDF of Normal(0,1), we get $\frac{\alpha}{\sigma/\sqrt{n}} = 1.96$
- $\alpha = 1.96 \times 3/\sqrt{16} = 1.47$
  - $P(|\hat{\mu} - \mu| < 1.47) = 0.95$
- 95% confidence interval: $[10.06 - 1.47, 10.06 + 1.47] = [8.59, 11.53]$

Normal samples and t-distribution

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
  - $\overline{X} \sim \text{Normal}(\mu, \sigma^2/n)$, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
  - $\overline{X}$ and S are independent

Normal samples and t-distribution

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
  - $\overline{X} \sim \text{Normal}(\mu, \sigma^2/n)$, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
  - $\overline{X}$ and S are independent
- $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$

Normal samples and t-distribution

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2)$$

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
  - $\overline{X} \sim \text{Normal}(0, \sigma^2/n)$, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$
  - $\overline{X}$ and S are independent
- $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$

- $\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim$ t-distribution with $n-1$ degrees of freedom, denoted t_{n-1}
  - PDF of t_n: proportional to $(1 + x^2/n)^{-(n+1)/2}$
  - Assume that CDF of t_n is known in calculations. Computer packages can provide CDF.

Normal samples with unknown variance

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2), \sigma^2 \text{ unknown}$$

- Sample instance: $x_1, \ldots, x_n$
- Estimated mean and variance: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$

Normal samples with unknown variance

$$X_1, \ldots, X_n \sim \text{iid Normal}(\mu, \sigma^2), \sigma^2 \text{ unknown}$$

- Sample instance: $x_1, \ldots, x_n$
- Estimated mean and variance: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

$\hat{\mu} = (x_1 + \cdots + x_n)/n$

- $\hat{\mu} \sim \text{Normal}(\mu, \sigma^2/n)$, $Z = \frac{\hat{\mu} - \mu}{(S/\sqrt{n})} \sim t_{n-1}$
  - $P(|\hat{\mu} - \mu| < \alpha) = \beta$ approx. $P(|\frac{\hat{\mu}-\mu}{S/\sqrt{n}}| < \frac{\alpha}{\hat{\sigma}/\sqrt{n}}) = \beta$
  - $\leftrightarrow P(|t_{n-1}| < \frac{\alpha}{\hat{\sigma}/\sqrt{n}}) = \beta$.

Solve using CDF of t_{n-1}

Given β, find $\frac{\alpha}{\hat{\sigma}/\sqrt{n}}$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and σ unknown. Find a 95% confidence interval for μ.

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and σ unknown. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\hat{\sigma} = 3.297$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and σ unknown. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\hat{\sigma} = 3.297$
- $\beta = 0.95$ and, using CDF of t_{15}, we get $\frac{\alpha}{\hat{\sigma}/\sqrt{n}} = 2.13$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and σ unknown. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\hat{\sigma} = 3.297$
- $\beta = 0.95$ and, using CDF of t_{15}, we get $\frac{\alpha}{\hat{\sigma}/\sqrt{n}} = 2.13$
- $\alpha = 2.13 \times 3.297/\sqrt{16} = 1.76$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and σ unknown. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\hat{\sigma} = 3.297$
- $\beta = 0.95$ and, using CDF of t_{15}, we get $\frac{\alpha}{\hat{\sigma}/\sqrt{n}} = 2.13$
- $\alpha = 2.13 \times 3.297/\sqrt{16} = 1.76$
  - $P(|\hat{\mu} - \mu| < 1.76) \approx 0.95$

Problem: Normal samples

Suppose 8.39, 12.78, 6.44, 9.36, 9.48, 14.39, 15.67, 12.6 , 9.37, 6.7, 7.02, 6.49, 6.77, 8.85, 16.05, 10.65 are *iid* normal samples with μ unknown and σ unknown. Find a 95% confidence interval for μ.

- $n = 16$, $\hat{\mu} = 10.06$, $\hat{\sigma} = 3.297$
- $\beta = 0.95$ and, using CDF of t_{15}, we get $\frac{\alpha}{\hat{\sigma}/\sqrt{n}} = 2.13$
- $\alpha = 2.13 \times 3.297/\sqrt{16} = 1.76$
  - $P(|\hat{\mu} - \mu| < 1.76) \approx 0.95$
- 95% confidence interval: $[10.06 - 1.76, 10.06 + 1.76] = [8.30, 11.82]$

What if samples are not normal?

- Use CLT to argue that sample mean will have a normal distribution
  - Use the same procedure as for normal, if reasonable

What if samples are not normal?

- Use CLT to argue that sample mean will have a normal distribution
  - Use the same procedure as for normal, if reasonable
- If the specific distribution is known, an expression or bound for the sample variance may be possible
- Bernoulli(p) samples
  - This is common in most sampling surveys
    - ★ Response is either yes or no
  - Sample variance: $\frac{p(1-p)}{n} \leq \frac{0.25}{n}$
    - ★ $\frac{\hat{\sigma}}{\sqrt{n}} \approx \sqrt{\frac{0.25}{n}}$ is commonly used
    - ★ 95% confidence interval: $[\hat{\mu} - 1.96\sqrt{\frac{0.25}{n}}, \hat{\mu} + 1.96\sqrt{\frac{0.25}{n}}]$

Revisit survey example

- "82% Indians willing to take Covid-19 vaccine," from a *Gallup Survey*
  - **3045** people were called and asked if they were willing to take a vaccine between **Nov 24, 2020 and Jan 8, 2021**
  - Languages spoken: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu, Odia, Punjabi, Assamese
  - 95% confidence level with a margin of error of 3%
- "40% Indians willing to take vaccine," from a *LocalCircles survey*
  - **9628** votes on **Jan 25, 2021** through the LocalCircles app
  - Demographics: people from 299 districts, 48% tier 1, 27% tier 2, 25% rural
  - 95% confidence interval will have a lower margin of error

Most probably, the two surveys are sampling different types of populations!