CS 215
Data Analysis and Interpretation

Estimation

Suyash P. Awate

Sample

- **Definition:**
  If random variables $X_1, \ldots, X_N$, are **i.i.d.**,
  then they constitute a random **sample** of size N from the common distribution
- N = "sample size"
- One set of observed data is one instance/realization of the sample
  - i.e., $\{x_1, \ldots, x_N\}$
- The common distribution from which data was "drawn" is usually unknown

Statistic

- **Definition:**
  Let $X_1, \dots, X_N$ denote a sample associated with random variable X (i.e., all of $X_1, \dots, X_N$ have the same distribution as X).
  Let $T(X_1, \dots, X_N)$ be a **function of the sample.**
  Then, random variable T is called the **statistic.**
- For the drawn sample $\{x_1, \dots, x_N\}$,
  the value $t := T(x_1, \dots, x_N)$ is an instance of the statistic

Model

- **Statistical model**
  - Typically, a probabilistic description of real-world phenomena
  - Description involves a distribution that may involve some parameters
    - e.g., $P(X; \theta)$
  - Describes/represents a data-**generation** process
  - Designed by people
    - Unlike data that is observed/measured/acquired
    - Nature doesn't generate models

Estimation

- **Estimation theory**
  - A branch of statistics that deals with estimating the values of parameters (underlying a statistical model) based on measured/empirical data
  - While data generation starts with parameters and leads to data, estimation starts with data and leads to parameters
- **Estimation problem**
  - Given: Data
  - Assumption: Data was generated from a parametric family of distributions (i.e., a family of models)
  - Goal: To infer the distribution parameters (i.e., the distribution/model instance from the family of distributions/models) that the data was generated from

Estimator, Estimate

- **Estimator**
  - A deterministic (not stochastic) rule/formula/algorithm for calculating/computing an estimate of a given quantity (e.g., a parameter value) based on observed data
  - An estimator is also a statistic
- **Estimate**
  - A value resulting from applying the estimator to data

Estimator Mean, Variance, Bias

- Let $X_1, \ldots, X_N$ be a sample on a random variable X with PDF/PMF $P(X; \theta)$
- Let $T(X_1, \ldots, X_N)$ be a statistic
- **Mean of the estimator (definition)**: Expected value of T, i.e., $E[T]$
- **Variance of the estimator (definition)**: $Var(T) := E[(T - E[T])^2]$
- **Bias of the estimator (definition)**: $Bias(T) := E[T] - \theta$
- **Mean squared error (MSE) of the estimator (definition)**
  - Expected value of the squared error $MSE(T) := E[(T - \theta)^2]$
- **Unbiased estimator (definition)**: T is unbiased if $Bias(T) = 0$
- **Consistent estimator (definition)**
  - Estimator is $T_N = T(X_1, \ldots, X_N)$ is consistent if $\forall \epsilon > 0, \lim_{N \to \infty} P(|T_N - \theta| \geq \epsilon) = 0$
  - Thus, T_N is said to "converge in probability" to θ

Estimator MSE, Bias, Variance

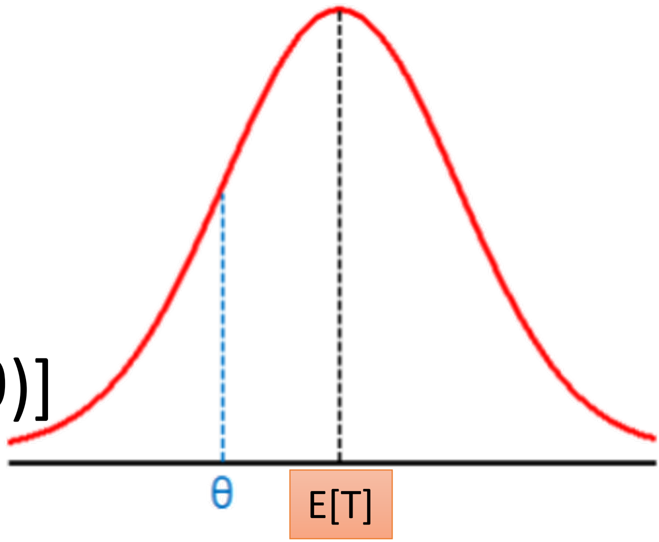
- $\mathrm{MSE}(T) := E[(T - \theta)^2]$

$= E[(T - E[T] + E[T] - \theta)^2]$

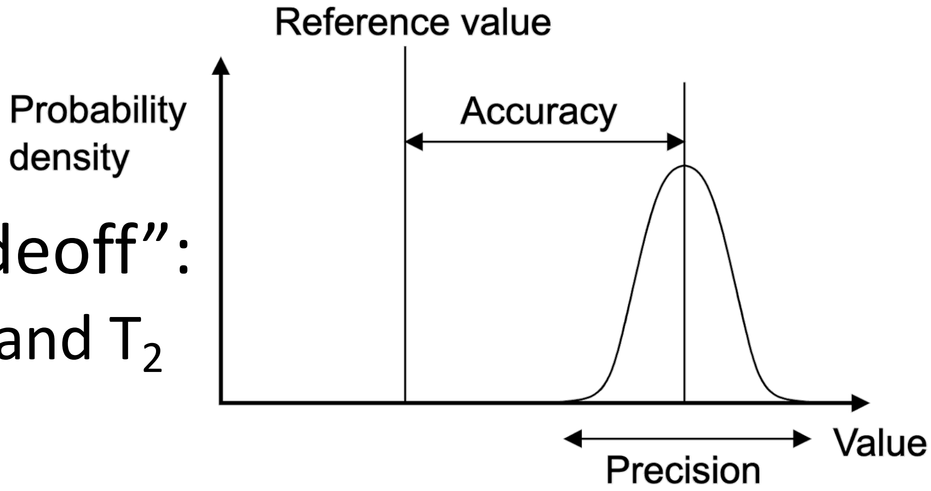
$= E[(T - E[T])^2] + E[(E[T] - \theta)^2] + E[2(T - E[T])(E[T] - \theta)]$

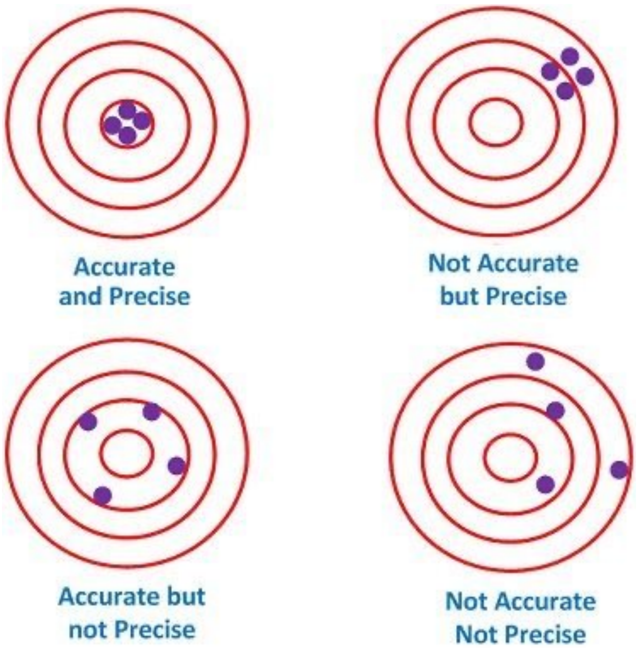
$= \mathrm{Var}(T) + (\mathrm{Bias}(T))^2 + 0$

: Variance + Bias^2



- Bias-variance decomposition/“tradeoff”:
  - If two estimators T_1 and T_2 have same MSE, then if one estimator (say, T_1) has a smaller bias magnitude, it (i.e., T_1) also has a larger variance





Likelihood Function

- Let $X_1, \ldots, X_N$ be a sample on a random variable X with PDF/PMF $P(X; \theta)$
- **Definition:** Likelihood function $L(\theta; X_1, \ldots, X_N) := \prod_{i=1}^{N} P(X_i; \theta)$
- We want to use the likelihood function to estimate θ from the sample
- Sometimes, analysis relies on $\log(L(\theta; X_1, \ldots, X_N))$,
  leveraging that log(.) is strictly monotonically increasing
- Some assumptions (#)
  1. Different values of θ correspond to different CDFs associated with $P(X; \theta)$
     - i.e., parameter θ identifies a unique distribution
  2. All PMFs/PDFs have common support for all parameters θ
     - i.e., support of X cannot depend on θ
  - Under these assumptions, the likelihood function has a nice property
    (as discussed next)

Likelihood Function

- **Theorem:** Let θ_{true} be the parameter value that led to sample $X_1, \ldots, X_N$. Assume $E_{P(X;\theta_{\text{true}})}[P(X;\theta)/P(X;\theta_{\text{true}})]$ exists (e.g., it is finite). Then,
$$\lim_{N\to\infty} P\left(L(\theta_{\text{true}}; X_1, \cdots, X_N) > L(\theta; X_1, \cdots, X_N); \theta_{\text{true}}\right) = 1, \forall\theta \neq \theta_{\text{true}}$$
- Proof:
  - Event $L(\theta_{\text{true}}; X_1, \cdots, X_N) > L(\theta; X_1, \cdots, X_N) \equiv \frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{P(X_i;\theta)}{P(X_i;\theta_{\text{true}})}\right] < 0$
  - We want to show that, as $N\rightarrow\infty$, this event has prob. 1 (i.e., inequality is true)
  - Because of the law of large numbers:
$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{P(X_i;\theta)}{P(X_i;\theta_{\text{true}})}\right] \to E_{P(X;\theta_{\text{true}})}\left[\log \frac{P(X;\theta)}{P(X;\theta_{\text{true}})}\right]$$
  - Because log(.) is a (strictly) concave function, Jensen's inequality makes the above expectation $< \log\left(E_{P(X;\theta_{\text{true}})}\left[\frac{P(X;\theta)}{P(X;\theta_{\text{true}})}\right]\right) = \log(1) = 0$

Law of large numbers:
For all $\varepsilon > 0$, as $n\rightarrow\infty$, $P(|\overline{X} - \mu| \geq \varepsilon) \rightarrow 0$

Maximum Likelihood (ML) Estimation

- **Definition**:
  An estimator $T = T(X_1, \ldots, X_N)$ is a "maximum likelihood (ML) estimator" if $T := \arg\max_\theta L(\theta; X_1, \cdots, X_N)$
  - "$\arg\max_\theta g(\theta)$": the argument (i.e., θ) that maximizes the function g(.)
  - "$\max_\theta g(\theta)$": the maximum possible value of the function g(.) across all θ
- Properties of ML estimation
  - Sometimes, ML estimator may not exist, or it may not be unique
  - When assumptions (#) hold, and max of likelihood function exists & is unique, then ML estimator is a consistent estimator
    - When sample size is finite, it loses guarantee to find true parameter value
      - When sample size is finite, this behavior holds for most methods,
        unless very strong assumptions (usually not holding in practice) are made on the data
  - In practice, a large enough sample size take ML estimate T sufficiently close to θ_{true} so that the ML estimate T is still useful

MLE for Bernoulli

- Let θ := probability of success
  - θ must lie within [0,1]
- Likelihood function L(θ) := $\prod_{i=1}^{N} \theta^{X_i} (1-\theta)^{(1-X_i)}$
- ML estimate for θ is what ?
  - At maximum of L(θ):
    - First derivative must be zero
      - This gives one equation in one unknown θ
    - Second derivative must be negative
  - ML estimate is sample mean, i.e., $\sum_{i=1}^{N} X_i \,/ N$

MLE for Binomial

- Let θ := probability of success
  - θ must lie within [0,1]
- Let M := number of Bernoulli tries for each Binomial random variable
- Let { X_i : i = 1, …, N} model repeated draws from Binomial, where X_i models number of successes in i-th draw from Binomial
- ML estimate for θ is sample mean $\sum_{i=1}^{N} X_i \,/(NM)$
- Interpretation:
  - N independent Binomials draws,
    where each Binomial has M independent Bernoulli draws,
    is equivalent to NM independent Bernoulli draws
  - Total number of successes in NM Bernoulli trials is $\sum_{i=1}^{N} X_i$

MLE for Poisson

- Parameter is average rate of arrivals/hits λ
- ML estimate is sample mean
- Note that λ is both mean and variance of the Poisson random variable

MLE for Gaussian

- Parameters are mean μ and standard deviation σ
- Likelihood function L(μ,σ) is a function of 2 variables
- Maximizing likelihood function L(μ,σ) is equivalent to maximizing log-likelihood function log(L(μ,σ))
  - Because log(.) function is a (strictly) monotonically increasing
- Need to solve for 2 equations in 2 unknowns
- ML estimate for μ is sample mean
- ML estimate for σ^2 is sample variance

MLE for Uniform Distribution

- Parameters are lower limit ‘a’ and upper limit ‘b’ ($a < b$)
- Let sample instance be $\{ x_1 , \ldots, x_N \}$, **sorted** in increasing order, & $x_1 < x_N$
- What are ML estimates ?
  - First, $a \le x_1$, else likelihood function is zero
  - Also, $x_N \le b$, else likelihood function is zero
  - Likelihood function $L(a,b) := (1/(b-a))^N$
  - Log-likelihood function $\log(L(a,b)) = -N.\log(b-a)$
    - Partial derivative w.r.t. ‘a’ is $N/(b-a) > 0$
    - Partial derivative w.r.t. ‘b’ is $(-N/(b-a)) < 0$
  - $L(a,b)$ is maximum when $a = x_1$ and $b = x_N$

On Preparation for Events (Exams) in Life

- From the Iron Man
  - “I don’t really prepare for anything like an event.”
  - “The goal is to be at a certain level of fitness.”
  - “I should be able to run a full marathon whenever I want.”
  - “That is the constant level of fitness that I aspire to.”
  - “I keep my fitness level as a goal, not an event as a goal.”
  - “There is no such thing as a good shortcut.”
  - “If you want to be healthy,
    and you want to be fit,
    and you want to be happy,
    you have to work hard.”
  - https://youtu.be/x_96xVfdzu0?t=303