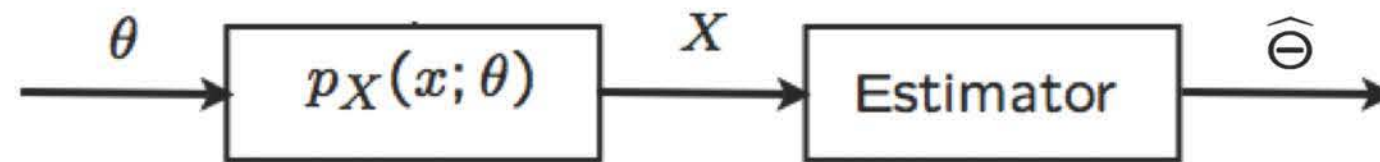


LECTURE 20: An introduction to classical statistics

- Unknown constant θ (not a r.v.)
- if $\theta = \mathbf{E}[X]$: estimate using the sample mean $(X_1 + \cdots + X_n)/n$
 - terminology and properties
- Confidence intervals (CIs)
 - CIs using the CLT
 - CIs when the variance is unknown
- Other uses of sample means
- Maximum Likelihood estimation

Classical statistics

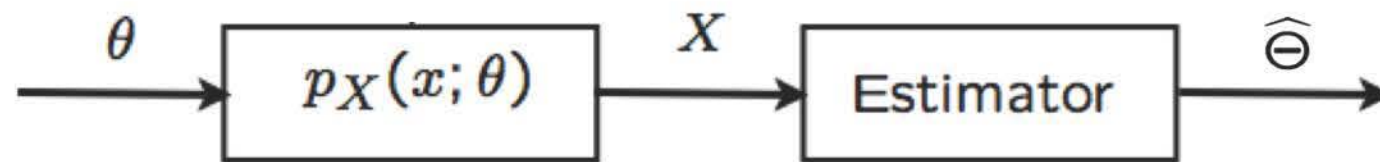
- Inference using the Bayes rule:
unknown Θ and observation X are both random variables
 - Find $p_{\Theta|X}$
- Classical statistics: unknown constant θ



- also for vectors X and θ : $p_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$
- $p_X(x; \theta)$ are NOT conditional probabilities; θ is NOT random
- mathematically: many models, one for each possible value of θ

Problem types in classical statistics

- Classical statistics: unknown constant θ



- Hypothesis testing: $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$
- Composite hypotheses: $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$
- Estimation: design an **estimator** $\hat{\Theta}$, to “keep estimation **error** $\hat{\Theta} - \theta$ small”

Estimating a mean

- X_1, \dots, X_n : i.i.d., mean θ , variance σ^2

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \dots + X_n}{n}$$

$\hat{\Theta}_n$: estimator (a random variable)

Properties and terminology:

- $E[\hat{\Theta}_n] = \theta$ (unbiased)
- WLLN: $\hat{\Theta}_n \rightarrow \theta$ (consistency)
- mean squared error (MSE): $E[(\hat{\Theta}_n - \theta)^2]$

On the mean squared error of an estimator

- For any estimator, using $\mathbf{E}[Z^2] = \text{var}(Z) + (\mathbf{E}[Z])^2$:

$$\mathbf{E}[(\hat{\Theta} - \theta)^2] = \text{var}(\hat{\Theta} - \theta) + (\mathbf{E}[\hat{\Theta} - \theta])^2 = \text{var}(\hat{\Theta}) + (\text{bias})^2$$

- $\sqrt{\text{var}(\hat{\Theta})}$ is called the **standard error**

Confidence intervals (CIs)

- The value of an estimator $\hat{\Theta}$ may not be informative enough
- An $1 - \alpha$ **confidence interval** is an interval $[\hat{\Theta}^-, \hat{\Theta}^+]$,
s.t. $P(\hat{\Theta}^- \leq \theta \leq \hat{\Theta}^+) \geq 1 - \alpha$, for all θ
 - often $\alpha = 0.05$, or 0.025 , or 0.01
 - interpretation is subtle

CI for the estimation of the mean

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

normal tables: $\Phi(1.96) = 0.975 = 1 - 0.025$

$$\mathbf{P}\left(\frac{|\hat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95 \quad (\text{CLT})$$

$$\mathbf{P}\left(\hat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

Confidence intervals for the mean when σ is unknown

$$\hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

$$\mathbf{P}\left(\hat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 1:** use **upper bound** on σ
 - if X_i Bernoulli: $\sigma \leq 1/2$
- **Option 2:** use **ad hoc estimate** of σ
 - if X_i Bernoulli: $\hat{\sigma} = \sqrt{\hat{\Theta}_n(1 - \hat{\Theta}_n)}$

Confidence intervals for the mean when σ is unknown

$$\mathbf{P}\left(\widehat{\Theta}_n - \frac{1.96 \sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96 \sigma}{\sqrt{n}}\right) \approx 0.95$$

- **Option 3:** Use **sample mean estimate** of the variance

- Two approximations involved here:
 - CLT: approximately normal
 - using estimate of σ
- correction for second approximation (t -tables) used when n is small

Start from $\sigma^2 = \mathbf{E}[(X_i - \theta)^2]$

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

(but do not know θ)

$$\frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\Theta}_n)^2 \rightarrow \sigma^2$$

Other natural estimators

- $\theta_X = \mathbf{E}[X]$ $\widehat{\Theta}_X = \frac{1}{n} \sum_{i=1}^n X_i$

- $\theta = \mathbf{E}[g(X)]$ $\widehat{\Theta} = \frac{1}{n} \sum_{i=1}^n g(X_i)$

- $v_X = \text{var}(X) = \mathbf{E}[(X - \theta_X)^2]$

$$\widehat{v}_X = \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\Theta}_X)^2$$

- $\text{cov}(X, Y) = \mathbf{E}[(X - \theta_X)(Y - \theta_Y)]$

$$\widehat{\text{cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \widehat{\Theta}_X)(Y_i - \widehat{\Theta}_Y)$$

- $\rho = \frac{\text{cov}(X, Y)}{\sqrt{v_X} \cdot \sqrt{v_Y}}$

$$\widehat{\rho} = \frac{\widehat{\text{cov}}(X, Y)}{\sqrt{\widehat{v}_X} \cdot \sqrt{\widehat{v}_Y}}$$

- next steps: find the distribution of $\widehat{\Theta}$, MSE, confidence intervals,...

Maximum Likelihood (ML) estimation

- $\theta = \mathbf{E}[g(X)]$ $\widehat{\Theta} = \frac{1}{n} \sum_{i=1}^n g(X_i)$

- Pick θ that “makes data most likely”

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} p_X(x; \theta)$$

– also applies when x , θ are vectors or x is continuous

- compare to Bayesian posterior: $p_{\Theta|X}(\theta | x) = \frac{p_{X|\Theta}(x | \theta) p_{\Theta}(\theta)}{p_X(x)}$

– interpretation is very different

Comments on ML

- maximize $p_X(x; \theta)$
- maximization is usually done numerically
- if have n i.i.d. data drawn from model $p_X(x; \theta)$, then, under mild assumptions:
 - consistent: $\hat{\Theta}_n \rightarrow \theta$
 - asymptotically normal: $\frac{\hat{\Theta}_n - \theta}{\sigma(\hat{\Theta}_n)} \rightarrow N(0, 1)$ (CDF convergence)
- analytical and simulation methods for calculating $\hat{\sigma} \approx \sigma(\hat{\Theta}_n)$
 - hence confidence intervals $\mathbf{P}\left(\hat{\Theta}_n - 1.96 \hat{\sigma} \leq \theta \leq \hat{\Theta}_n + 1.96 \hat{\sigma}\right) \approx 0.95$
 - asymptotically “efficient” (“best”)

ML estimation example: parameter of binomial

- K : binomial with parameters n (known), and θ (unknown)

$$p_K(k; \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$$

$$\hat{\theta}_{\text{ML}} = \frac{k}{n} \quad \hat{\Theta}_{\text{ML}} = \frac{K}{n}$$

- same as MAP estimator with uniform prior on θ

ML estimation example — normal mean and variance

- X_1, \dots, X_n : i.i.d., $N(\mu, v)$ $f_X(x; \mu, v) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi v}} \exp \left\{ -\frac{(x_i - \mu)^2}{2v} \right\}$

minimize $\frac{n}{2} \log v + \sum_{i=1}^n \frac{(x_i - \mu)^2}{2v}$

– minimize w.r.t. μ : $\hat{\mu} = \frac{x_1 + \dots + x_n}{n}$

– minimize w.r.t. v : $\hat{v} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Introduction to Probability
John Tsitsiklis and Patrick Jaillet

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.