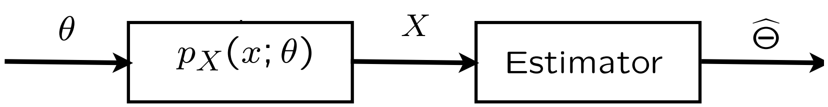
LECTURE 20: An introduction to classical statistics

- 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

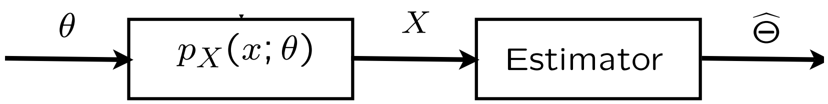
- Inference using the Bayes rule:
  unknown Θ and observation X are both random variables
  - Find $p_{\Theta|X}$
- Classical statistics: unknown constant θ



  - also for vectors X and θ: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
  - $p_X(x;\theta)$ 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** $\widehat{\Theta}$, to "keep estimation **error** $\widehat{\Theta} - \theta$ small"

Estimating a mean

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

$\widehat{\Theta}_n = \text{sample mean} = M_n = \dfrac{X_1 + \cdots + X_n}{n}$ $\qquad$ $\widehat{\Theta}_n$: estimator (a random variable)

Properties and terminology:

- $\mathbf{E}[\widehat{\Theta}_n] = \theta$ (unbiased)

- WLLN: $\widehat{\Theta}_n \to \theta$ (consistency)

- mean squared error (MSE): $\mathbf{E}\Big[(\widehat{\Theta}_n - \theta)^2\Big]$

On the mean squared error of an estimator

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

$$\mathbf{E}[(\widehat{\Theta} - \theta)^2] \;=\; \mathrm{var}(\widehat{\Theta} - \theta) + \left(\mathbf{E}[\widehat{\Theta} - \theta]\right)^2 \;=\; \mathrm{var}(\widehat{\Theta}) + (\mathrm{bias})^2$$

- $\sqrt{\mathrm{var}(\widehat{\Theta})}$ is called the standard error

Confidence intervals (CIs)

- The value of an estimator $\widehat{\Theta}$ may not be informative enough

- An $1-\alpha$ **confidence interval** is an interval $[\widehat{\Theta}^-, \widehat{\Theta}^+]$,

  s.t. $\quad \mathbf{P}(\widehat{\Theta}^- \leq \theta \leq \widehat{\Theta}^+) \geq 1-\alpha, \quad$ for all θ

  - often $\alpha = 0.05$, or 0.025, or 0.01

  - interpretation is subtle

CI for the estimation of the mean

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

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

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

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

Confidence intervals for the mean when σ is unknown

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

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

- **Option 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: $\hat{\sigma} = \sqrt{\widehat{\Theta}_n(1 - \widehat{\Theta}_n)}$

Confidence intervals for the mean when σ is unknown

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

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

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

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

(but do not know θ)

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

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

Other natural estimators

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

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

- $v_X = \mathsf{var}(X) = \mathbf{E}\Big[(X - \theta_X)^2\Big]$ $\qquad$ $\widehat{v}_X = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \big(X_i - \widehat{\Theta}_X)^2$

- $\mathsf{cov}(X,Y) = \mathbf{E}\Big[(X - \theta_X)(Y - \theta_Y)\Big]$ $\qquad$ $\widehat{\mathsf{cov}}(X,Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \big(X_i - \widehat{\Theta}_X\big)\big(Y_i - \widehat{\Theta}_Y\big)$

- $\rho = \dfrac{\mathsf{cov}(X,Y)}{\sqrt{v_x} \cdot \sqrt{v_Y}}$ $\qquad$ $\widehat{\rho} = \dfrac{\widehat{\mathsf{cov}}(X,Y)}{\sqrt{\widehat{v}_x} \cdot \sqrt{\widehat{v}_Y}}$

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

Maximum Likelihood (ML) estimation

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

- 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: $\qquad p_{\Theta|X}(\theta\,|\,x) = \dfrac{p_{X|\Theta}(x\,|\,\theta)\,p_{\Theta}(\theta)}{p_X(x)}$

  – 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: $\widehat{\Theta}_n \to \theta$
  - asumptotically normal: $\dfrac{\widehat{\Theta}_n - \theta}{\sigma(\widehat{\Theta}_n)} \longrightarrow N(0,1)$ (CDF convergence)
- analytical and simulation methods for calculating $\hat{\sigma} \approx \sigma(\widehat{\Theta}_n)$
  - hence confidence intervals $\mathbf{P}\left(\widehat{\Theta}_n - 1.96\,\hat{\sigma} \le \theta \le \widehat{\Theta}_n + 1.96\,\hat{\sigma}\right) \approx 0.95$
  - asymptotically "efficient" ("best")

ML estimation example: parameter of binomial

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

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

$$\hat{\theta}_{\mathrm{ML}} = \frac{k}{n} \qquad \widehat{\Theta}_{\mathrm{ML}} = \frac{K}{n}$$

- same as MAP estimator with uniform prior on θ

ML estimation example — normal mean and variance

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

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

  – minimize w.r.t. μ: $\qquad \widehat{\mu} = \frac{x_1 + \cdots + x_n}{n}$

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