LECTURE 20: An introduction to classical statistics

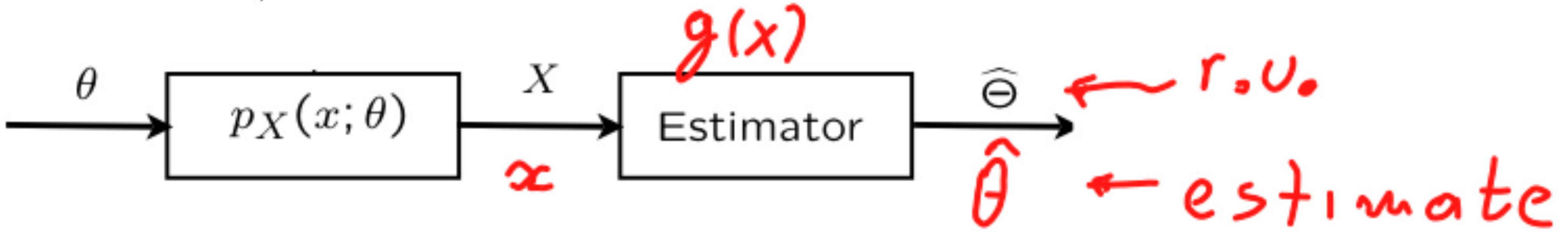
- Unknown constant θ (not a r.v.)
- if $\theta = \mathbf{E}[X]$: estimate using the sample mean $(X_1 + \cdots + X_n)/n$
  - terminology and properties
- Confidence intervals (CIs)
  - CIs using the CLT
  - CIs when the variance is unknown
- Other uses of sample means
- Maximum Likelihood estimation

Classical statistics

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

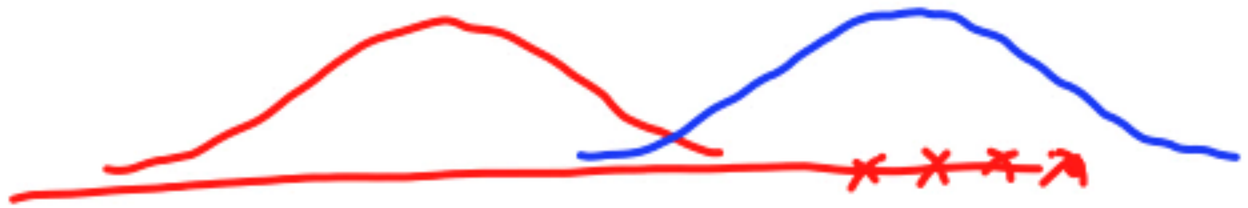
$P_\Theta \quad P_{X|\Theta}$

- Classical statistics: unknown constant θ



$p_{X|\cancel{\Theta}}(x|\theta)$

  - 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

$$\xrightarrow{\theta} \boxed{p_X(x;\theta)} \xrightarrow{X} \boxed{\text{Estimator}} \xrightarrow{\widehat{\Theta}}$$

- 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"

Art!

Estimating a mean

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

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

Properties and terminology:

$\hat{\Theta} = g(x)$

$E[\hat{\Theta}] = \sum_x g(x)\, p_X(x;\theta)$

- $\mathbf{E}[\widehat{\Theta}_n] = \theta$ (unbiased)
  for all θ 这个等式成立

- WLLN: $\widehat{\Theta}_n \xrightarrow{i.p.} \theta$ (consistency)
  for all θ

- mean squared error (MSE): $\mathbf{E}\left[(\widehat{\Theta}_n - \theta)^2\right] = \mathrm{var}(\hat{\Theta}_n) = \dfrac{\sigma^2}{n}$.

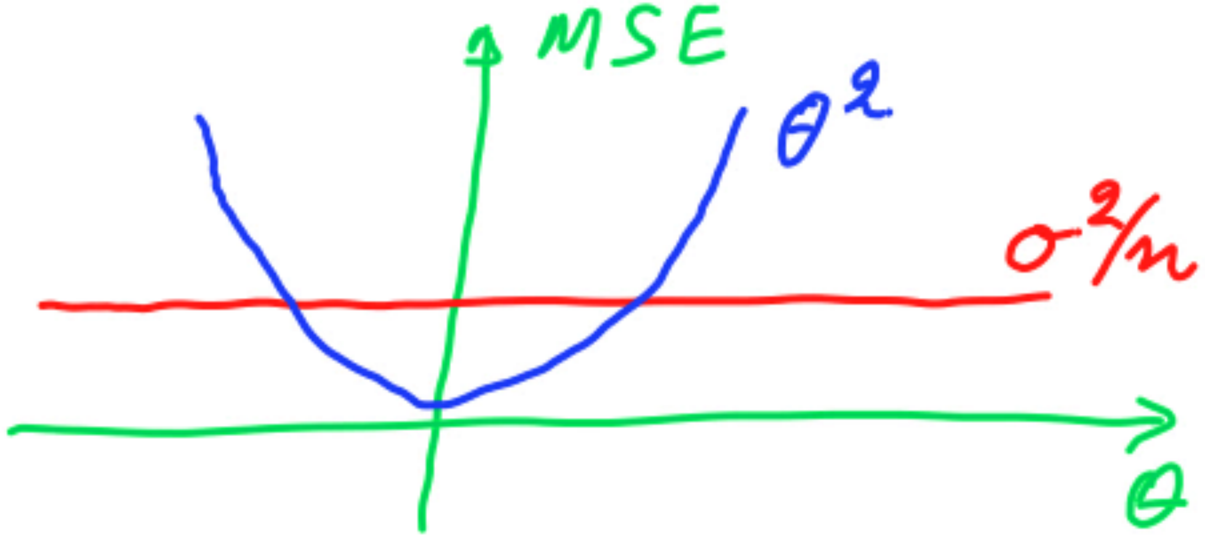
On the mean squared error of an estimator

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

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

$\hat{\Theta}_n = M_n$: $\quad MSE = \sigma^2/n + 0$

$\hat{\Theta} = 0$: $\quad MSE = 0 + \theta^2$



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

Confidence intervals (CIs)

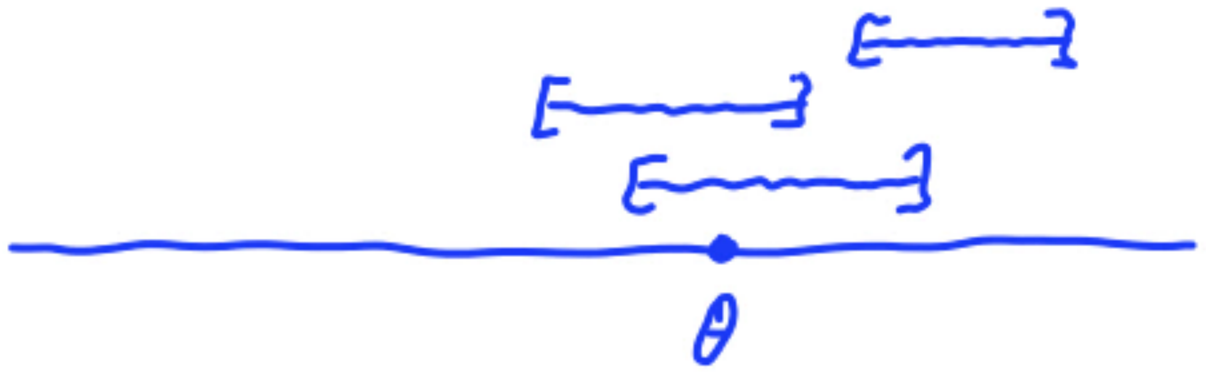
- The value of an estimator $\widehat{\Theta}$ may not be informative enough

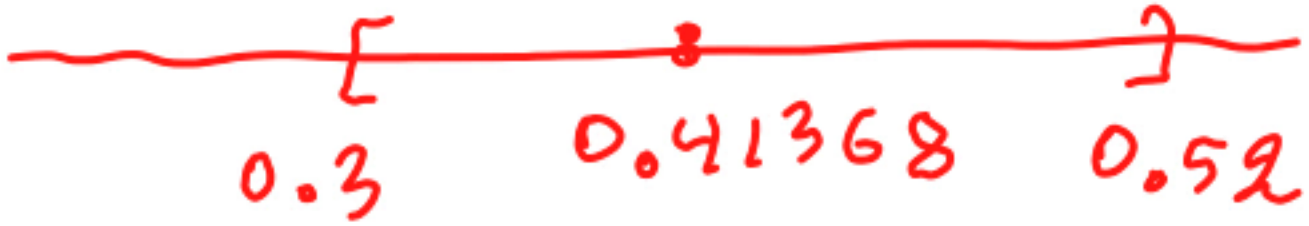
95%

- 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





$\cancel{P(0.3 < \theta < 0.52) \geq .95}$

not in one specific sample

CI for the estimation of the mean

iid θ σ^2

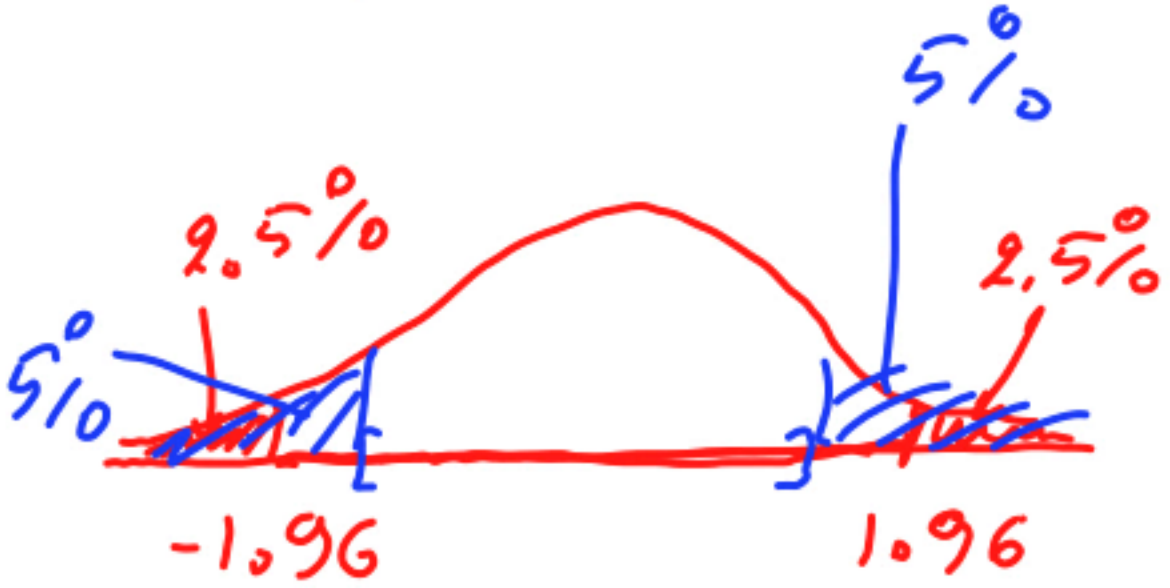
$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

95%

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

90%

$\Phi(1.645) = 0.95$



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

std of the sample mean(theta_n)

$$Z_n = \frac{(S_n - n\mu)/n}{\sqrt{n}\,\sigma/n}$$

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

Confidence intervals for the mean when σ is unknown

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

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

- **Option 1:** use upper bound on σ
  - if X_i Bernoulli: $\sigma \le 1/2$

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

$\sigma = \sqrt{\theta(1-\theta)}$

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)

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

- 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

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

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

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

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

(X_i, Y_i)

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

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

Maximum Likelihood (ML) estimation

- Pick θ that "makes data most likely"

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

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

constant

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

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

  - asymptotically "efficient" ("best")

ML estimation example: parameter of binomial

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

k

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

$$\log\left[\binom{n}{k}\right] + k\log\theta + (n-k)\log(1-\theta)$$

$$0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \Rightarrow k - \cancel{k\theta} = n\theta - \cancel{k\theta}$$

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

- 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: $\qquad \widehat{\mu} = \frac{x_1 + \cdots + x_n}{n}$

  $$\frac{1}{v}\sum_{i=1}^{n}(x_i - \mu) = 0 \Rightarrow \sum x_i = n\mu$$

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

  $$\frac{n}{\cancel{2}} \cdot \frac{1}{\cancel{v}} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\cancel{2}v^{\cancel{2}}} = 0$$