LECTURE 20: An introduction to classical statistics

- Unknown constant $\theta$ (not a r.v.)
- if $\theta = \mathbb{E}[X]$: 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 $\Theta$ and observation $X$ are both random variables
  - Find $p_{\Theta|X}$
- Classical statistics: unknown constant $\theta$
  - also for vectors $X$ and $\theta$: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
  - $p_X(x;\theta)$ are NOT conditional probabilities; $\theta$ is NOT random
  - mathematically: many models, one for each possible value of $\theta$
Problem types in classical statistics

- Classical statistics: unknown constant $\theta$

  \[ \theta \xrightarrow{p_X(x; \theta)} X \xrightarrow{\text{Estimator}} \hat{\theta} \]

- 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 $\hat{\theta}$, to “keep estimation error $\hat{\theta} - \theta$ small”

  *Art!*
Estimating a mean

- \( X_1, \ldots, X_n \): i.i.d., mean \( \theta \), variance \( \sigma^2 \)

\[ \Theta_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n} \]

\( \Theta_n \): estimator (a random variable)

Properties and terminology:

- \( \mathbb{E}[\Theta_n] = \theta \) (unbiased)
- WLLN: \( \Theta_n \to \theta \) (consistency)
- mean squared error (MSE): \( \mathbb{E}[(\Theta_n - \theta)^2] = \text{var}(\Theta_n) = \frac{\sigma^2}{n} \).
On the mean squared error of an estimator

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

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

\[
\hat{\Theta} = 0: \quad \text{MSE} = 0 + \theta^2
\]

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

- The value of an estimator $\hat{\theta}$ may not be informative enough
  - 95%
- 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 $\theta$
  - often $\alpha = 0.05$, or 0.025, or 0.01
  - interpretation is subtle

$P(0.3 < \theta < 0.52) \geq 0.95$
CI for the estimation of the mean

\[ \hat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \ldots + X_n}{n} \]

- 95% normal tables: \( \Phi(1.96) = 0.975 = 1 - 0.025 \)
- 90% normal tables: \( \Phi(1.645) = 0.95 \)

\[ P\left( \left| \hat{\Theta}_n - \theta \right| \leq 1.96 \right) \approx 0.95 \text{ (CLT)} \]

\[ P\left( \hat{\Theta}_n - \frac{1.96 \sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{1.96 \sigma}{\sqrt{n}} \right) \approx 0.95 \]
Confidence intervals for the mean when $\sigma$ is unknown

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

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

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

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

$\sigma = \sqrt{\Theta(1 - \Theta)}$
Confidence intervals for the mean when $\sigma$ is unknown

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

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

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

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

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

(but do not know $\theta$)

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{\theta}_n)^2 \to \sigma^2$$
Other natural estimators

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

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

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

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

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

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

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

- $\widehat{\rho} = \frac{\text{cov}(X, Y)}{\sqrt{\nu_X} \cdot \sqrt{\nu_Y}}$

- next steps: find the distribution of $\widehat{\Theta}$, MSE, confidence intervals, ...
Maximum Likelihood (ML) estimation

- Pick $\theta$ that "makes data most likely"

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

- also applies when $x, \theta$ 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 $P\left(\hat{\Theta}_n - 1.96\hat{\sigma} \leq \theta \leq \hat{\Theta}_n + 1.96\hat{\sigma}\right) \approx 0.95$
  
  - asymptotically “efficient” (“best”)
ML estimation example: parameter of binomial

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

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

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

\[
0 + \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \Rightarrow k \theta = n \theta - k \theta
\]

\[
\hat{\theta}_{\text{ML}} = \frac{k}{n} \quad \hat{\theta}_{\text{ML}} = \frac{K}{n}
\]

- same as MAP estimator with uniform prior on \( \theta \)
ML estimation example — normal mean and variance

- \(X_1, \ldots, X_n\): i.i.d., \(N(\mu, \nu)\)

\[
f_X(x; \mu, \nu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \nu}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\nu} \right\}
\]

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

- minimize w.r.t. \(\mu\):
  \[
  \hat{\mu} = \frac{x_1 + \cdots + x_n}{n}
  \]

  \[
  \frac{1}{\nu} \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \sum x_i = n \mu
  \]

- minimize w.r.t. \(\nu\):
  \[
  \hat{\nu} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2.
  \]

  \[
  \frac{m}{n} \cdot \frac{1}{\nu} \rightarrow \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{x_i \nu} = 0
  \]