Statistical Theory: revision handout

Kolyan Ray

Disclaimer: this is a **brief overview** of the **main ideas** of the course. It is **not** an exhaustive list of everything that is examinable, nor is anything covered guaranteed to appear in the exam. Results are also written less precisely: refer to the notes for more rigorous formulations.

1. Principles of point estimation

Definition. *A family of distributions is a k-parameter exponential family if its pmf/pdf takes the form:*

$$f_\theta(x) = \exp\left\{\sum_{i=1}^{k} c_i(\theta)T_i(x) - d(\theta) + S(x)\right\},$$

and the support of f_θ does not depend on θ.

Definition. *A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of X given $T(X)$ does not depend on θ.*

Theorem 1.1 (Factorization criterion)**.** *$T = T(X)$ is sufficient for θ if and only if*

$$f_\theta(x) = g(T(x), \theta)h(x)$$

for some functions g and h.

Example. *If X has distribution belonging a k-parameter exponential family, then by the factorization criterion, $(T_1(X), \cdots, T_k(X))$ is sufficient for θ.*

Definition. *A sufficient statistic $T(X)$ is minimal if it is a function of every other sufficient statistic.*

Sufficient and minimal sufficient statistics are not unique: any bijective function of a minimal sufficient statistic is also minimal.

Theorem 1.2. *If $T = T(X)$ satisfies*

$$\frac{f_\theta(x)}{f_\theta(x')} \text{ does not depend on } \theta \iff T(x) = T(x').$$

Then T is minimal sufficient for θ.

We can rewrite the MSE using the *bias-variance decomposition*:

$$\mathrm{MSE}_\theta(\hat\theta) = E_\theta(\hat\theta - \theta)^2 = \mathrm{var}_\theta(\hat\theta) + b_\theta(\hat\theta)^2.$$

bijective functions preserves sufficient statistics
and minimal sufficient statistics

Theorem 1.3 (Rao-Blackwell theorem)**.** *Let $T = T(X)$ be sufficient for θ and $\tilde\theta(X)$ be an estimator for θ. Let $\hat\theta(X) = E[\tilde\theta(X)|T(X)]$. Then for all $\theta \in \Theta$,*

$$b_\theta(\hat\theta) = b_\theta(\tilde\theta) \qquad \text{and} \qquad \mathrm{var}_\theta(\hat\theta) \le \mathrm{var}_\theta(\tilde\theta).$$

with equality if and only if $\tilde\theta$ is a function of T. [Note: $MSE_\theta(\hat\theta) \le MSE_\theta(\tilde\theta)$].

Lemma 1.1. *Let T_1 and T_2 be two sufficient statistics for θ and $\tilde\theta(X)$ be an estimator for θ. Let $\hat\theta_i(X) = E[\tilde\theta(X)|T_i(X)]$, $i = 1, 2$. If $T_2 = h(T_1)$, then for all $\theta \in \Theta$,*

$$\mathrm{var}_\theta(\hat\theta_2) \le \mathrm{var}_\theta(\hat\theta_1).$$

Remark.
- *$\hat\theta(X)$ does not depend on θ by sufficiency.*
- *Best variance reduction arises from conditioning on a minimal sufficient statistic.*
- *Variance of $\hat\theta(X) = E[\tilde\theta(X)|T(X)]$ does depend on baselines estimator $\tilde\theta$.*

2. Likelihood-based estimation

The likelihood function $L : \Theta \to \mathbb{R}$,

$$L(\theta) = L_n(\theta) = f_{n,\theta}(x)$$

is considered as a *function of* θ for a *fixed* $X = x$. The log-likelihood is $l(\theta) = \log L(\theta)$. A (not necessarily unique) maximum likelihood estimator (MLE) is defined as *any* element $\hat{\theta} \in \Theta$ for which

$$L_n(\hat{\theta}) = \max_{\theta \in \Theta} L_n(\theta).$$

Remark. *Some strategies for finding MLEs when dimension $p = 1$:*

- *Solve $l'(\hat{\theta}) = 0$ (stationary point) and show $l''(\theta) < 0$ for all θ (global maximum).*
- *Solve $l'(\hat{\theta}) = 0$ (stationary point) and show $l'(\theta) \geq 0 \iff \theta \leq \hat{\theta}$.*
- *Maximize $l(\theta)$ by direct arguments (often useful when the support of f_θ depends on θ).*

Theorem 2.1 (Invariance of MLE)**.** *If $\hat{\theta}_{ML}$ is an MLE for θ and $g(\theta)$ is any (measurable) function, then $g(\hat{\theta}_{ML})$ is an MLE for $g(\theta)$.*

Lemma 2.1. *Suppose $E_\theta|\log f_\theta(X)| < \infty$ for all $\theta \in \Theta$. If $X \sim f_{\theta_0}$ for some true $\theta_0 \in \Theta$, then for any $\theta \in \Theta$,*

$$E_{\theta_0}[l(\theta)] \leq E_{\theta_0}[l(\theta_0)],$$

i.e. $\theta \mapsto E_{\theta_0}[l(\theta)]$ is maximized at θ_0.

Definition. *For $\Theta \subseteq \mathbb{R}^p$ and $\theta \mapsto l_n(\theta)$ differentiable, the score function is defined as*

$$S_n(\theta) = \nabla_\theta l_n(\theta) = \left(\frac{\partial}{\partial \theta_1} l_n(\theta), \ldots, \frac{\partial}{\partial \theta_p} l_n(\theta) \right)^T.$$

When $p = 1$, this is just $S_n(\theta) = l_n'(\theta)$, which is often used to solve $l_n'(\theta) = 0$.

Lemma 2.2. *Consider a model $\{f_\theta : \theta \in \Theta\}$ that is regular enough that differentiation (in θ) and integration (in x) can be exchanged. Then for all $\theta \in int(\Theta)$,*

$$E_\theta[\nabla_\theta \log f_\theta(X)] = 0.$$

Remark. *When the support of f_θ depends on θ, it is not generally possible to interchange differentiation and integration, e.g. for $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U[0, \theta]$. Be careful when assuming regularity conditions!*

In particular, this implies $E_{\theta_0}[\nabla_\theta \log f_{\theta_0}(X)] = 0$ at the true parameter θ_0.

Definition. *For a parameter space $\Theta \subseteq \mathbb{R}^p$, we define for all $\theta \in int(\Theta)$ the Fisher information matrix as*

$$I_{ij}(\theta) = E_\theta \left[\frac{\partial}{\partial \theta_i} \log f_\theta(X) \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right] \quad \left(= E_\theta \left[l'(\theta; X)^2 \right] \right), \qquad 1 \leq i, j \leq p.$$

Lemma 2.3. *Under the same regularity assumptions as Lemma 2.2, for all $\theta \in int(\Theta)$,*

$$I_{ij}(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_\theta(X) \right] \quad (= -E_\theta[l''(\theta)]), \qquad 1 \leq i, j \leq p.$$

Proposition 2.1. *If $X = (X_1, \ldots, X_n)$ with X_i i.i.d. random variables, then*

$$I_X(\theta) = n I_{X_1}(\theta).$$

Theorem 2.2 (Cramer-Rao lower bound)**.** *Consider a model $\{f_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ (i.e. $p = 1$) under the same regularity assumptions as Lemma 2.2. Let $\hat{\theta} = \hat{\theta}(X)$ be an unbiased estimator of θ based on an observation X from this model. Then for all $\theta \in int(\Theta)$,*

$$\text{var}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] \geq \frac{1}{I_X(\theta)} \qquad \left(=^{iid} \frac{1}{n I_{X_1}(\theta)} \right).$$

for estimators of g(θ), the lower bound is (g'(θ) + b'(θ))² / I(θ) where b(θ) is bias

Proposition 2.2 (Attaining the CR bound)**.** *Assume regularity conditions and $p = 1$. An unbiased statistic $\hat{\theta}(X)$ attains the Cramer-Rao lower bound if and only if X belongs to the exponential family*

$$f_\theta(x) = \exp\left(A(\theta)\hat{\theta}(x) + B(\theta) + S(x)\right).$$

- The CR bound is not always attained.
- If an unbiased estimator attains the CR bound for all $\theta \in \Theta$, it is the UMVUE.

by this theorem, an estimator of θ when X is an exponential family

unbiased statistics attains CR lower bound iff it is a linear transformation of T_i(X)

3. Asymptotic theory for MLEs

Definition. *Consider $X_1, \ldots, X_n \sim^{iid} P_{\theta_0}$. A sequence of estimators $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is consistent if $\hat{\theta}_n \to^P \theta_0$ as $n \to \infty$ for all $\theta_0 \in \Theta$.*

Strategies for consistency:

- If $\hat{\theta}_n = \bar{X}_n$ use WLLN.
- If $\hat{\theta}_n = h(\bar{X})$ use WLLN and continuous mapping theorem.
- Use Markov's inequality

$$P_{\theta_0}\left(|\hat{\theta}_n - \theta_0| > \epsilon\right) = P_{\theta_0}\left((\hat{\theta}_n - \theta_0)^2 > \epsilon^2\right) \leq \frac{E_{\theta_0}[(\hat{\theta}_n - \theta_0)^2]}{\epsilon^2} = \frac{\mathrm{MSE}_{\theta_0}(\hat{\theta}_n)}{\epsilon^2}.$$

- Use general asymptotic theory for MLEs.

Assumption 3.1 (Model regularity)**.** *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_\theta$, where f_θ is a pmf/pdf such that:*

1. *The parameter space Θ is an open subset of $\mathbb{R}$ (i.e. no boundary points)*
2. *$\theta \mapsto l_{X_1}(\theta)$ is twice continuously differentiable in θ for all $x \in \mathcal{X}$.*
3. *$E_\theta[l''_{X_1}(\theta)] < \infty$ for all $\theta \in \Theta$.*
4. *We can exchange integration/summation in x with two-times differentiation in θ (support of f_θ should not depend on θ):*

this means differentiation can be exchanged with expectation for both continuous and discrete r.v.

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathcal{X}} f_\theta(x)dx = \int_{\mathcal{X}} \frac{\mathrm{d}}{\mathrm{d}\theta} f_\theta(x)dx, \qquad \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\int_{\mathcal{X}} f_\theta(x)dx = \int_{\mathcal{X}} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} f_\theta(x)dx.$$

Theorem 3.1 (Consistency and asymptotic normality of the MLE)**.** *Let $\{f_\theta : \theta \in \Theta\}$ satisfy Assumption 3.1 and suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0}$ for some true $\theta_0 \in \Theta$. Then the MLE $\hat{\theta}$ satisfies, as $n \to \infty$,*

$$\hat{\theta}_{ML} \to^P \theta_0,$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N\left(0, \frac{1}{\mathcal{I}_{X_1}(\theta_0)}\right).$$

Remark.
- *For $\theta \in \mathbb{R}^p$, $p \geq 1$, one has*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N_p(0, \mathcal{I}_{X_1}^{-1}(\theta_0)).$$

- *If $\hat{\theta}_n = h(\bar{X}_n)$, often easier to use CLT and continuous mapping theorem (consistency) or the delta method (asymptotic normality), sometimes combined with Slutsky's theorem.*

Theorem 3.2 (Delta method $p = 1$)**.** *Let $g : \Theta \to \mathbb{R}$ be continuously differentiable at θ_0 with derivative $g'(\theta_0) \neq 0$. Let (Y_n) be random variables such that $\sqrt{n}(Y_n - \theta_0) \to^d Z$. Then*

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \to^d g'(\theta_0)^T Z.$$

Corollary 3.1. *If $Z \sim N(0, \sigma^2)$, then*

$$\sqrt{n}(g(Y_n) - g(\theta_0)) \to^d N(0, g'(\theta_0)^2\sigma^2).$$

Remark. *If the MLE $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \to^d N(0, 1/\mathcal{I}_{X_1}(\theta_0)$ as $n \to \infty$, then*

$$\sqrt{n}(g(\hat{\theta}_{ML}) - g(\theta_0)) \to^d N(0, g'(\theta_0)^2\mathcal{I}^{-1}(\theta_0)).$$

4. Bayesian inference

- Let $X \sim f_\theta$, $\theta \in \Theta$.
- Treat $\theta \sim \pi$ as a random variable with prior π.
- Observe X ($=^{iid} X_1, \ldots, X_n$).
- Compute the *posterior* distribution $\pi(\theta|X)$ (i.e. conditional distribution) by Bayes' theorem:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f_X(x)} = \frac{f_\theta(x)\pi(\theta)}{\int_\Theta f_{\theta'}(x)\pi(\theta')\,\mathrm{d}\theta'} \propto f(x|\theta)\pi(\theta) = L(\theta)\pi(\theta).$$

The constant of proportionality is chosen to make the posterior integrate/sum to one. Commonly used point estimators are the posterior mean, median and mode (depends on the *loss function*).

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(θ) and assume a Beta(α, β) prior distribution for θ:*

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad 0 < \theta < 1,$$

where $\alpha, \beta > 0$ are known. Then the posterior distribution of θ given observations $X_1 = x_1, \ldots, X_n = x_n$ satisfies (keeping track of only the θ terms)

$$\pi(\theta|x) \propto f_\theta(x)\pi(\theta) \propto \left(\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}\right)\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}.$$

We recognize this as the density (as a function of θ) of a Beta($\sum x_i + \alpha, n - \sum x_i + \beta$) distribution. Thus from the formula for the Beta distribution, we can read off the normalizing constant:

$$\pi(\theta|x) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\sum_{i=1}^n x_i + \alpha)\Gamma(n - \sum_{i=1}^n x_i + \beta)}\theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}.$$

Therefore, the posterior distribution is also a Beta distribution, but with updated parameters.

We can analyze Bayesian methods under the frequentist assumption that $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_{\theta_0}$ for some true θ_0.

1. Compute the posterior as usual.
2. *Then* see how it (or its estimators) behave under this assumption.

Example. *The posterior mean equals (properties of Beta distributions)*

$$E[\theta|x] = \frac{\sum x_i + \alpha}{n+\alpha+\beta} = \frac{n}{n+\alpha+\beta}\bar{X}_n + \frac{\alpha}{n+\alpha+\beta}.$$

If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Bernoulli(θ_0), by WLLN, $\bar{X}_n \to^p E_{\theta_0}X_1 = \theta_0$. Since $\frac{n}{n+\alpha+\beta} \to^p 1$ and $\frac{\alpha}{n+\alpha+\beta} \to^p 0$ (deterministic convergence implies convergence in probability), we have by Slutsky's theorem that $E[\theta|x] \to^d 1 \times \theta_0 + 0 = \theta_0$ as $n \to \infty$, i.e. consistency.

Can work out bias or variance of $E[\theta|x]$, or can do something similar to work out asymptotic normality of $E[\theta|x]$ (replace WLLN by CLT).

The frequentist variance of $E[\theta|x]$ is **not** the same as the posterior variance.

Remark. *When the posterior is in the same family of distributions as the prior, the prior is called conjugate. This allows you to compute the integral $f_X(x) = \int_\Theta f_{\theta'}(x)\pi(\theta')\,\mathrm{d}\theta'$ in the denominator of Bayes' formula by 'recognizing' the form of the distribution.*

Definition. *A non-negative prior function π with $\int_\Theta \pi(\theta)d\theta = \infty$ is called an improper prior.*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$. Assign to θ the prior $\pi(\theta) \propto 1$.*

Definition. *The prior $\pi(\theta) \propto \sqrt{det(I(\theta))}$ is called the Jeffreys prior. For $p = 1$, this is $\pi(\theta) \propto I(\theta)^{1/2}$.*

- This prior might not be proper.
- Since the Fisher information is additive over independent observations (Proposition 2.1), the Jeffreys prior for n i.i.d. observations is the same as for a single observation.

Lemma 4.1. *If θ has Jeffreys prior and $\varphi = h(\theta)$ is a smooth reparametrization, then φ also has Jeffreys prior.*

5. Optimality in Estimation

5.1. Decision Theory

An action space $\mathcal{A}$ is a set of actions and a decision rule is a map from the observation space $\mathcal{X}$ to $\mathcal{A}$, i.e.

$$\delta : \mathcal{X} \to \mathcal{A}.$$

A loss function $L : \mathcal{A} \times \Theta \to [0, \infty)$ is a non-negative function that determines the cost of a particular action for a given parameter θ.

Definition. *For a loss function L, a decision rule δ and an observation $X \sim f_\theta$, the risk function is*

$$R(\delta, \theta) = E_\theta[L(\delta(X), \theta)] = \int_{\mathcal{X}} L(\delta(x), \theta) f_\theta(x) dx.$$

Example.

- *Estimation: $\mathcal{A} = \Theta$ and the decision $\delta(X) = \hat{\theta}(X)$ is an estimator. Two commonly used loss functions are the squared error loss and absolute error:*

  $$L(a, \theta) = (a - \theta)^2 \quad or \quad L(a, \theta) = |a - \theta|.$$

  The risk function is $R(\delta, \theta) = E_\theta[(\delta(X) - \theta)^2] = MSE_\theta(\delta)$ (mean-squared error) or $R(\delta, \theta) = E_\theta[|\delta(X) - \theta|]$ (expected absolute error).
- *Hypothesis testing: $\mathcal{A} = \{0, 1\}$ and the decision $\delta(X)$ is a test. Writing $a \in \{0, 1\}$ for the chosen hypothesis and $\theta \in \{0, 1\}$ for the true hypothesis, use 0-1 loss:*

  $$L(a, \theta) = \mathbb{1}_{\{a \neq \theta\}}.$$

  The risk function $R(\delta, \theta) = E_\theta[\mathbb{1}_{\{\delta(X) \neq \theta\}}] = P_\theta(\delta(X) \neq \theta)$ describes the probability of making a (type I/II) error.

The risk function of a decision rule is a function of θ, and different decision rules can each perform better on different parts of the parameter space Θ. We cannot normally minimize $R(\delta, \theta)$ *uniformly* in $\theta \in \Theta$.

Definition. *For a loss function L and parameter space Θ, a decision rule δ is inadmissible if there exists a decision rule $\delta^*(X)$ such that $R(\delta^*, \theta) \leq R(\delta, \theta)$ for all $\theta \in \Theta$, and the inequality is strict for some $\theta \in \Theta$. If no such δ^* exists, then δ is admissible.*

Definition. *Given a prior $\pi(\theta)$ on Θ and a loss function L, the π-Bayes risk for the decision rule δ is*

$$R_\pi(\delta) = E_{\theta \sim \pi}[R(\delta, \theta)] = \int_\Theta R(\delta, \theta)\pi(\theta) d\theta = \int_\Theta \int_{\mathcal{X}} L(\delta(x), \theta) f_\theta(x) \pi(\theta) d\theta.$$

A π-Bayes decision rule δ_π is any decision rule that minimizes $R_\pi(\delta)$.

Example. *Consider $X \sim Binomial(n, \theta)$, a prior $\theta \sim U[0, 1]$, a decision rule (estimator) $\hat{\theta}_1(X) = X/n$ and squared error loss.*

1. *Compute risk function $R(\delta, \theta) = MSE_\theta(\hat{\theta}) = \mathrm{var}_\theta(\hat{\theta}) = \theta(1 - \theta)/n$.*

2. *Take expectation over prior:*

$$R_\pi(X/n) = E_{\theta\sim\pi}[R(\delta,\theta)] = \frac{1}{n}\int_0^1 \theta(1-\theta)d\theta = \frac{1}{6n}.$$

Definition. *The* <u>posterior risk</u> *is defined as the average loss under the posterior distribution for an observation $X \in \mathcal{X}$:*

$$R_\pi(\delta(x)) = E_\pi[L(\delta(x),\theta)|x] = \int_\Theta L(\delta(x),\theta)\pi(\theta|x)d\theta.$$

Proposition 5.1. *An estimator δ that minimizes the π-posterior risk also minimizes the π-Bayes risk.*

The converse is true under mild conditions. In particular, this tells us that if the minimizer of the posterior risk is *unique*, then so is the minimizer of the π-Bayes risk.

Proposition 5.2. *Suppose δ_π minimizes the Bayes risk $R_\pi(\delta)$ and $R_\pi(\delta_\pi) < \infty$. Then $\delta_\pi(x)$ minimizes the posterior risk $R_\pi(\delta(x))$ (with probability one under the prior predictive distribution $f_\pi(x) = \int f_\theta(x)\pi(\theta)d\theta$).*

To find the minimizer of the posterior risk

1. Compute the posterior distribution as a function of x (for arbitrary x).
2. Compute the posterior risk $E_\pi[L(\delta(x),\theta)|x]$ by taking the expectation over $\theta \sim \pi(\theta|x)$ for fixed x as a function of $\delta(x)$. Pick $\delta(x)$ that minimizes this.
3. The minimizer is $\delta(x)$ as a function of x.

Example. *For estimation with squared error loss $L(a,\theta) = (a-\theta)^2$, the minimizing decision rule is the posterior mean $\delta(x) = \int_\Theta \theta\pi(\theta|x)d\theta = E_\pi[\theta|x]$. For absolute error loss $L(a,\theta) = |a-\theta|$, it is the posterior median.*

To find the minimizer of the *Bayes risk* either

- Find the minimizer of the posterior risk.
- Evaluate the Bayes risk and directly minimize this.

Proposition 5.3. *If a Bayes estimator $\hat{\theta}_{Bayes}$ is unique, then it is admissible.*

Strategies for proving admissibility:

- Show estimator is the unique Bayes estimator for *some* prior.
- Try to prove this from scratch (e.g. by contradiction).

Definition. *The* <u>minimax risk</u> *is defined as the infimum ('min') over all decision rules δ of the maximal ('max') risk over the whole parameter space Θ:*

$$\inf_\delta \sup_{\theta\in\Theta} R(\delta,\theta).$$

A decision rule that attains the minimax risk if called <u>minimax</u>.

Lemma 5.1 (Bayes and minimax risks)**.** *For any decision rule δ and prior π for θ,*

$$R_\pi(\delta) \le \sup_{\theta\in\Theta} R(\delta,\theta).$$

Proposition 5.4. *Let π be a prior on Θ such that*

$$R_\pi(\delta_\pi) = \sup_{\theta\in\Theta} R(\delta_\pi,\theta),$$

where δ_π is a (unique) π-Bayes rule. Then δ_π is (unique) minimax.

Hence if the maximal risk of a Bayes rules equals the Bayes risk, the corresponding Bayes rule is minimax.

Corollary 5.1. *If a (unique) Bayes rule δ_π has constant risk in θ, then it is (unique) minimax.*

Lemma 5.2. *If δ is admissible and has constant risk, then it is minimax.*

Strategies for finding a minimax estimator:

- Find a Bayes rule with constant risk.
- Find a Bayes rule whose Bayes risk equals its maximal risk over the parameter space.
- Find an admissible estimator with constant risk.

5.2. Minimum variance unbiased estimators

Definition. *Consider estimation of $g(\theta)$ based on data $X \sim P_\theta$, $\theta \in \Theta$. An unbiased estimator $\hat{g}(X)$ of $g(\theta)$ is a uniformly minimum variance unbiased estimator (UMVUE) if*

$$\text{var}_\theta(\hat{g}) \leq \text{var}_\theta(\tilde{g}) \qquad \forall \theta \in \Theta,$$

for any other unbiased estimator $\tilde{g}(X)$ of $g(\theta)$.

Definition. *Let $X \sim P_\theta$, $\theta \in \Theta$. A statistic $T = T(X)$ is complete for θ if, for any (measurable) function g,*

$$\text{if } E_\theta[g(T)] = 0 \text{ for all } \theta \in \Theta, \text{ then } P_\theta(g(T) = 0) = 1 \text{ for all } \theta \in \Theta.$$

complete statistics are preserved under bijective functions

Proposition 5.5. *Suppose $X = (X_1, \cdots, X_n)$ have joint distribution belonging to a k-parameter exponential family of distributions:*

$$f_\theta(x) = \exp\left\{\sum_{i=1}^{k} c_i(\theta)T_i(x) - d(\theta) + S(x)\right\}.$$

If the exponential family has full rank (roughly speaking, if $c_1(\theta), \ldots, c_k(\theta)$ are linearly independent and $T_1(x), \ldots, T_k(x)$ are also linearly independent), then $T = (T_1(X), \cdots, T_k(X))$ is complete for θ.

Theorem 5.1. *If a sufficient statistic T is complete, then it is minimal.*

The converse of the last theorem is not true.

Definition. *A statistic is an ancillary statistic if its distribution does not depend on the parameter θ.*

Example. *If $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N(\theta, 1)$, then $T(X) = \max_i X_i - \min_i X_i$ is ancillary for θ.*

Theorem 5.2 (Basu's Theorem)**.** *If T is a complete sufficient statistic for θ, then any ancillary statistic V is independent of T.*

Using a complete and (minimal) sufficient statistic, we can find the best unbiased estimator (**UMVUE**).

Theorem 5.3 (Lehmann-Scheffe Theorem)**.** *Let T be a sufficient and complete statistic for θ, and $\tilde{g}$ be an unbiased estimator of $g(\theta)$. If $\hat{g}(T(X)) = E[\tilde{g}(X)|T(X)]$, then $\hat{g}$ is the unique uniformly minimum variance unbiased estimator (UMVUE) of $g(\theta)$.*

Strategies for finding the UMVUE:

- If a complete sufficient statistic T exists:
  - Take an unbiased estimator $\tilde{g}$ of $g(\theta)$ and construct $\hat{g} = E[\tilde{g}|T]$. This is the UMVUE.
  - Find a function $h = h(T)$ of T that is unbiased: $E_\theta[h(T)] = g(\theta) \;\; \forall \theta \in \Theta$. Then $h(T)$ is the UMVUE.
- Find an estimator that achieves the CR lower bound for every $\theta \in \Theta$ (doesn't need completeness).

Remark. *The UMVUE need* ***not*** *attain the CR lower bound. In fact the UMVUE need not even exist.*

Example. *Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim}$ Poisson(θ), $\theta > 0$. It is easy to show that $\bar{X}_n$ is unbiased for θ and it is a function of the complete and sufficient statistic $T = \sum_{i=1}^n X_i$. Therefore, $\bar{X}_n$ is the UMVUE of θ.*

6. Hypothesis testing and confidence intervals

Suppose we observe $X \sim P_\theta$ and we want to test the hypotheses

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_1.$$

Definition. *A test is a binary function $\phi : \mathcal{X} \to \{0,1\}$ from the sample space. If $\phi(X) = \mathbb{1}_R(X)$ is an indicator function, then R is called the critical region or rejection region.*

When performing a test, we may make two types of errors.

Type I error: reject H_0 when H_0 is true.
Type II error: reject H_1 when H_1 is true.

Remark. *The null hypothesis and alternative hypothesis are **not** considered equally. By default, we assume the null hypothesis is true and we need a lot of evidence to reject it.*

Definition. *The power function $\pi_\phi : \Theta \to [0,1]$ of a test $\phi = \mathbb{1}_R$ with rejection region R is*

$$\pi_\phi(\theta) = P_\theta(X \in R_\phi) = E_\theta[\phi(X)] = P_\theta(\text{reject } H_0).$$

A good test should have π_ϕ small for $\theta \in \Theta_0$ and large for $\theta \in \Theta_1$.

Definition. *The size of a test ϕ is*

$$\alpha = \sup_{\theta \in \Theta_0} \pi_\phi(\theta).$$

Definition. *A test ϕ is a level α test if*

$$\sup_{\theta \in \Theta_0} \pi_\phi(\theta) \leq \alpha.$$

Definition. *A test ϕ is uniformly most powerful (UMP) at level α for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ if:*

(i) $\sup_{\theta \in \Theta_0} \pi_\phi(\theta) \leq \alpha$ *(level α test),*
(ii) for any other test level α test ϕ^*, we have $\pi_{\phi^*}(\theta) \leq \pi_\phi(\theta)$ for all $\theta \in \Theta_1$.

UMP tests do not necessarily exist.

Simple hypotheses. Consider *simple* hypotheses:

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta = \theta_1,$$

where θ_0 and θ_1 are known. The *likelihood ratio* of the two simple hypotheses H_0 and H_1 given data x is

$$\Lambda(x) = \Lambda(x; H_0, H_1) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}.$$

A *likelihood ratio test* (LRT) is one where the critical/rejection region takes the form

$$R = \{x : \Lambda(x; H_0, H_1) > k\}.$$

Lemma 6.1 (Neyman-Pearson lemma)**.** *Suppose $X \sim f_\theta(x)$ and consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Then among all tests of size α, the test with the largest power is the likelihood ratio test of size α:*

$$\phi(x) = \mathbb{1}\{x : \Lambda(x; H_0, H_1) > k\} = \begin{cases} 1 & \text{if } f_{\theta_1}(x) > k f_{\theta_0}(x), \\ 0 & \text{if } f_{\theta_1}(x) \leq k f_{\theta_0}(x), \end{cases}$$

where $k > 0$ is such that $E_{\theta_0}[\phi(X)] = P_{\theta_0}(f_{\theta_1}(X) > k f_{\theta_0}(X)) = \alpha$.

Remark. *We assume there exists a k such that $E_{\theta_0}[\phi(X)] = \alpha$ exactly. Otherwise, we might have $E_{\theta_0}[\phi(X)] < \alpha$, which can be dealt with using a randomized test.*

Strategy for UMP test for simple hypotheses:

note: if you find $\wedge(x)$ is monotone increasing/decreasing in x, can use CDF of x to obtain value of k

1. Compute the likelihood ratio statistic $\Lambda(x) = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}$.
2. Rearrange the inequality $\{\Lambda(x) > k\}$ in terms of a nice statistic $T(X)$ (e.g. $T(X) = \bar{X}_n$).
3. Work out the distribution of T under H_0.
4. Find k such that $P_{\theta_0}(\Lambda(x) > k) = \alpha$ by rewriting the probability in terms of T and using the H_0-distribution of T.
5. (This will have largest power over H_1, i.e. $\pi_\phi(\theta_1)$ is maximized over all level α tests)

One-sided hypotheses. Suppose $X \sim f_\theta$, where $\theta \in \Theta \subseteq \mathbb{R}$, and consider the one-sided hypotheses

$$H_0 : \theta \leq \theta_0, \qquad H_1 : \theta > \theta_0.$$

Definition. *A family of distributions $\{f_\theta(x) : \theta \in \Theta\}$ has monotone likelihood ratio if there exists a function $T(x)$ such that for any $\theta_2 > \theta_1$, the ratio $\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)}$ is a non-decreasing function of $T(x)$.*

Karlin-Rubin theorem also works when ratio is a decreasing function of T(x)

Theorem 6.1 (Karlin-Rubin theorem)**.** *Suppose $X \sim f_\theta(x)$ and consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. If $\{f_\theta(x) : \theta \in \Theta\}$ has monotone likelihood ratio in a statistic $T(x)$, then the UMP test at level α is*

$$\phi(x) = \mathbb{1}\{x : T(x) \geq k\} = \begin{cases} 1 & \text{if } T(x) \geq k, \\ 0 & \text{if } T(x) < k, \end{cases}$$

for k such that $P_{\theta_0}(T(X) \geq k) = \alpha$.

Strategy for UMP test one-sided hypotheses:

- Check whether family of distribution has monotone likelihood ratio.
- If so, compute test threshold k based on significance level α as above.

Remark. *For testing $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, the UMP test at level α is similarly*

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) \leq k, \\ 0 & \text{if } T(x) > k, \end{cases}$$

where $\alpha = P_{\theta_0}(T(X) \leq k)$.

Definition. *Let $X \sim f_\theta(x)$, where $\theta \in \Theta$. The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is defined as*

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} = \frac{L(\hat{\theta}_{ML})}{L(\hat{\theta}_0)},$$

where $\hat{\theta}_0$ and $\hat{\theta}_{ML}$ are the MLEs for θ under the models Θ_0 and Θ, respectively.

A likelihood ratio test (LRT) at level α rejects H_0 if $\Lambda(x) \geq k$, where $k \geq 1$ is chosen so that

$$\sup_{\theta \in \Theta_0} P_\theta(\Lambda(X) \geq k) = \alpha.$$

Theorem 6.2 (Wilks' theorem)**.** *Let $\{f_\theta : \theta \in \Theta\}$ be a statistical model satisfying Assumption 3.1, except $\Theta \subseteq \mathbb{R}^p$ for possibly $p \geq 1$. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f_\theta$ and consider the testing problem $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Then under H_0, as $n \to \infty$,*

$$2 \log \Lambda(X) \to^d \chi^2_p.$$

Strategy for computing (asymptotic) likelihood ratio test:

1. Compute the MLEs $\hat{\theta}_0$ and $\hat{\theta}_{ML}$ under H_0 and the full model, respectively ($\hat{\theta}_0 = \theta_0$ if $H_0 : \theta = \theta_0$).
2. Evaluate the likelihood ratio test statistic $\Lambda(x)$.
3. Reject H_0 if $2 \log \Lambda(x) > k_\alpha$, where k_α satisfies $P(\chi^2_p > k_\alpha) = \alpha$.

4. Try to rearrange $\{\Lambda(x) > k_\alpha\}$ in terms of a 'nice' statistic $T(X)$.

Definition. *Let $X \sim f_\theta$. For $0 < \alpha < 1$, a set $C = C(X)$ is called a $100(1-\alpha)\%$ confidence set (or interval if $p = 1$) for θ if*

$$P_\theta(\theta \in C(X)) = 1 - \alpha$$

(or $\geq 1 - \alpha$) for all $\theta \in \Theta$. The probability $1 - \alpha$ is called the coverage.

If we calculate $C(x)$ for a large number of samples $X = x$, then approximately $100(1-\alpha)\%$ of them will cover (contain) the true value of θ.

Definition. *A random variable $Q(X, \theta)$ is a pivotal quantity if its distribution does not depend on the parameter θ.*

Strategy 1 to construct a confidence interval:

1. Find a pivotal quantity $Q(X, \theta)$ such that the P_θ-distribution of $Q(X, \theta)$ does not depend on θ (for all $\theta \in \Theta$).
2. Write down a probability statement of the form $P_\theta(a \leq Q(X, \theta) \leq b) = 1 - \alpha$ [a, b will not depend on θ since Q is pivotal]
3. Rearrange the inequalities inside $P_\theta(\dots)$ to find an interval for θ.

Example. *Suppose $X_1, \dots, X_n \overset{\text{i.i.d}}{\sim} N(\mu, 1)$ and we want to construct a 95% confidence interval for μ. We know $\bar{X}_n \sim N(\mu, 1/n)$, so $Q(X, \mu) = \sqrt{n}(\bar{X}_n - \mu) \sim N(0, 1)$ is pivotal quantity for μ. For $Z \sim N(0, 1)$, we can find a, b such that*

$$P(a \leq Z \leq b) = P_\mu(a \leq \sqrt{n}(\bar{X}_n - \mu) \leq b) = 1 - \alpha$$

for all $\mu \in \mathbb{R}$. Rearranging the inequalities in terms of μ gives a confidence interval.

Remark. *One can construct asymptotic confidence intervals using asymptotically pivotal quantities. Suppose $Q_n(X, \theta) \to^d Z$, where the distribution of Z does not depend on θ. Then we instead use*

$$P_\theta(a \leq Q_n(X, \theta) \leq b) \approx P_\theta(a \leq Z \leq b) = 1 - \alpha$$

and rearrange the inequalities for θ.

Definition. *The acceptance region A of a test is the complement of the critical/rejection region R.*

Theorem 6.3. *For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. Then the set*

$$C(X) = \{\theta : X \in A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence set for θ. Conversely, let $C(X)$ be a $100(1-\alpha)\%$ confidence set for θ. Then

$$A(\theta_0) = \{X : \theta_0 \in C(X)\}$$

is the acceptance region for a level α test of $H_0 : \theta = \theta_0$.

Strategy 2 to construct a confidence interval:

1. Consider tests $(\phi_{\theta_0} : \theta_0 \in \Theta)$, each with null hypothesis $H_0 : \theta = \theta_0$.
2. Work out the acceptance/non-rejection region for every $\theta_0 \in \Theta$:

$$A(\theta_0) = \{x : \phi_{\theta_0}(x) = 0\}.$$

3. Rearrange the condition $\{\phi_{\theta_0}(x) = 0\}$ in terms of θ_0 (depends on the form of ϕ_{θ_0}).
4. Define $C(X) = \{\theta : X \in A(\theta)\} = \{\theta : \phi_\theta(X) = 0\}$ and substitute in the rearrangement in the last set.
5. (Can see if this simplifies into an interval or half-interval, etc.)