Seth Temple
sdtemple@uw.edu

Notes

1 Decision Theory

We introduce four criteria to measure the performance of decision rules: Bayes risk, minimaxity, Γ-minimaxity, and admissibility.

1.1 Definitions

- An action a from an action space $\mathcal{A}$ is the realization of a probability distribution conditional on data.
- A **decision rule** $D(\cdot|X = x)$ is a probability distribution conditional on data. We commonly use deterministic decision rules, i.e. the probability distribution is degenerate.
- A loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ measures the quality of an action $a \in \mathcal{A}$ when $\theta \in \Theta$.
- Risk is a function that measures the quality of a decision rule given $\theta \in \Theta$.

$$\mathcal{R}(D, \theta) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(a, \theta) D(da|x) dP_\theta(x).$$

- A (decision) rule is **inadmissible** if there is another decision rule that has risk less than or equal to it everywhere and risk strictly less than it somewhere. A decision rule is **admissible** if it is not inadmissible.
- A rule is **minimax** if its worst case risk is the infimum of the worst case risks of various rules. That is,

$$\sup_{\theta \in \Theta} \mathcal{R}(D^\star, \theta) = \inf_{D \in \mathcal{D}} \sup_{\theta \in \Theta} \mathcal{R}(D, \theta).$$

- The Bayes risk with respect to prior Π is the expectation of a rule's risk. That is,

$$r(D, \Pi) = \int_\Theta \mathcal{R}(D, \theta) d\Pi(\theta).$$

- A rule is a **Bayes rule** with respect to a prior Π if it achieves the smallest Bayes risk. That is,

$$r(D^\star, \Pi) = \inf_{D \in \mathcal{D}} r(D, \Pi).$$

- A rule is unique Bayes for a prior if any other Bayes rule equals it except on null sets for all P_θ. A rule is unique minimax if any other minimax rule equals it except on null sets for all P_θ.

- An estimator $D^\star$ is Γ-minimax w.r.t. a loss function if

$$\sup_{\Pi \in \Gamma} r(D^\star, \Pi) = \inf_T \sup_{\Pi \in \Gamma} r(T, \Pi).$$

  This definition is analogous to minimax but w.r.t. Bayes risk as opposed to risk.

- The **kernel** of a posterior distribution is the function depending on the parameter, not the proportionality constant. The kernel uniquely determines the posterior distribution.

- A conjugate prior is one such that the posterior distribution is in the same family as the prior distribution.

- A prior is **least favorable** if the Bayes risk for the Bayes rule and the prior achieves the supremum over priors of Bayes risks for paired Bayes rules and priors.

- A sequence of priors is least favorable if, for all priors Π,

$$r(D_\Pi, \Pi) \leq \liminf_{k \to \infty} r(D_{\Pi_k}, \Pi_k).$$

1.2 Results

- **Finding Bayes rules by minimizing the conditional expected loss.** Suppose that $\boldsymbol{\theta} \sim \Pi$, $X | \boldsymbol{\theta} = \theta \sim P_\theta$, and the loss is nonnegative. If,

  (i) there is a rule with finite Bayes risk, and

  (ii) there exists $D_\Pi \in \mathcal{D}$ for almost all x that minimizes the conditional expected loss,

  then D_Π is a Bayes rule.

- If the loss function is convex for fixed θ, decision rules are unrestricted, the action space is convex, and there is a Bayes rule, then there is a **deterministic Bayes rule**.

- **Constant risk theorem.** If Π satisfies

$$r(D_\Pi, \Pi) = \sup_{\theta \in \Theta} \mathcal{R}(D_\Pi, \theta),$$

  then (i) D_Π is minimax, (ii) unique Bayes D_Π w.r.t. Π implies unique minimax, and (iii) Π is a least favorable prior.

- **Constant risk theorem (ii).** For a sequence of priors $\{\Pi_k\}$, if $D \in \mathcal{D}$ satisfies

$$\sup_{\theta \in \Theta} \mathcal{R}(D, \theta) = \liminf_{k \to \infty} r(D_{\Pi_k}, \Pi_k),$$

  then (i) D is minimax, and (ii) $\{\Pi_k\}$ is a least favorable prior sequence.

- **Constant Bayes risk theorem.** If $\Pi^\star \in \Gamma$ satisfies
$$r(D_{\Pi_\star}, \Pi_\star) = \sup_{\Pi \in \Gamma} r(D_{\Pi^\star}, \Pi)$$
then (i) $D_{\Pi_\star}$ is Γ-minimax, (ii) unique Bayes $D_{\Pi^\star}$ implies unique minimax, and (iii) $\Pi^\star$ is a least favorable prior.

- If a minimax estimator for a smaller model has no worse worst-case risk scenario for a large model, then the estimator is minimax for the larger model as well. (See slide 36.)

- **Some admissible estimators.** (i) Any unique Bayes rule is admissible, and (ii) any unique minimax rule is admissible. Moreover, for squared error loss, finite $r(D_\Pi, \Pi)$, and $\int P_\theta(X \in A) d\Pi(\theta) = 0$ implying $P_\theta(X \in A) = 0$ for all θ, we have that D_Π is a unique Bayes rule.

- **Stein's lemma.** Let $Y \sim N(\mu, \sigma^2)$ and let $g : \mathbb{R} \to \mathbb{R}$ be such that $\mathbb{E}[|g'(Y)|] < \infty$. Then, $\mathbb{E}[g(Y)(Y - \mu)] = \sigma^2 \mathbb{E}[g'(Y)]$. (See slides 74-76 for multivariate generalization.)

1.3 Examples

- The posterior mean is the (deterministic) Bayes rule under L_2 loss.

- The posterior median is the (deterministic) Bayes rule under L_1 loss.

- The posterior mode is the (deterministic) Bayes rule under 0-1 loss.

- The sample mean is a minimax estimator for θ in an $X \equiv (X_1, \cdots, X_n)$ iid sample of $N(\theta, \sigma^2)$ random variables.

- The sample mean is an admissible estimator for θ in the case of univariate normals with mean θ. The sample mean is inadmissible for dimension three or higher.

- The James-Stein estimator beats the sample mean when estimating a mean vector of dimension 3 or more. The positive-part James-Stein estimator beats the James-Stein estimator, so the James-Stein estimator is inadmissible too. Intuitively, the James-Stein estimator shrinks the sample mean towards zero (or some other point).
$$T^{JS}(x) = \begin{cases} (1 - \frac{(d-2)\sigma^2}{n||\bar{x}_n||^2})\bar{x}_n & \bar{x}_n \neq (0, \cdots, 0) \\ 0 & \text{otherwise} \end{cases}.$$
In some cases, we can reduce the James-Stein estimator to lower dimensions by leveraging the fact that it is a spherically symmetric estimator, where a spherically symmetric estimator is of the form $T_\tau(x) = \tau(||x||)x$. This fact solicits some geometric intuition (see slides 63-72). The shrinkage property creates bias and may be inappropriate for estimating individual means. Finally, we can motivate these estimators from an empirical Bayes perspective.

2 Large Sample Theory

2.1 Definitions

- (convergence almost surely) $A_n \to_{a.s.} A$ if $P(\lim_{n\to\infty} ||A_n - A|| = 0) = 1$.
- (convergence in probability) $A_n \to_p A$ if, for all $\varepsilon > 0$, $P(||A_n - A|| > \varepsilon) \to 0$.
- $\mathbb{R}^d$-valued random variable A_n converges in distribution to A if, for all bounded, continuous functions $f : \mathbb{R}^d \to \mathbb{R}$,

  $$\mathbb{E}[f(A_n)] \to \mathbb{E}[f(A)].$$

  This convergence is sometimes referred to a weak convergence or convergence in law.
- Uniform integrability: $\{X_n\}$ is u.i. if $\sup_n \mathbb{E}[|X_n| \cdot 1\{|X_n| \geq a\}] \to 0$. This is a condition controlling tail probabilities. (Weak convergence and u.i. implies convergence of means.)
- **Order notations.**
  - $x_n = O(r_n)$ if $\limsup |x_n/r_n| < \infty$. Equivalently, there exists $M > 0$ such that $I\{|x_n| \leq M|r_n|\} \to 1$. In layman's terms, x_n is within some multiplicative constant of r_n.
  - $x_n = o(r_n)$ if $\limsup |x_n/r_n| = 0$. Identically, for all $M > 0$, $I\{|x_n| \leq M|r_n|\} \to 1$. In layman's terms, x_n changes slower than r_n.
  - $X_n = O_P(R_n)$ if, for all $\varepsilon > 0$, there exists $M > 0$ s.t.

    $$\liminf P(||X_n|| \leq M||R_n||) > 1 - \varepsilon.$$

  - $X_n = o_P(R_n)$ if, for all $M > 0$,

    $$P(||X_n|| \leq M||R_n||) \to 1.$$

  - Stochastic and determisitic notations are equivalent when $X_n \overset{a.s.}{=} x_n$ and $R_n \overset{a.s.}{=} r_n$.
  - $X_n = o_P(1)$ if and only if $X_n \to_p 0$.
  - $X_n = O_P(1)$ is also referred to as the random sequence being uniformly tight.
  - $o_P(1)$ is related to convergence in probability whereas $O_P(1)$ is related to weak convergence.
  - Read these useful properties left to right as an implication:

    (1) $X_n = o_P(R_n)$ if and only if $X_n = R_n Y_n$ for some $Y_n = o_P(1)$;
    (2) $X_n = O_P(R_n)$ if and only if $X_n = R_n Y_n$ for some $Y_n = O_P(1)$;

(3) $o_P(1) + o_P(1) = o_P(1)$;
(4) $o_P(1) + O_P(1) = O_P(1)$;
(5) $O_P(1)O_P(1) = O_P(1)$;
(6) $o_P(1)O_P(1) = o_P(1)$;
(7) $[1 + o_P(1)]^{-1} = O_P(1)$;
(8) $X_n = o_P(1)$ implies $X_n = O_P(1)$.

2.2 Results

- Almost sure convergence implies convergence in probability implies weak convergence.
- Let $\{A_n\}, A$, and B be defined on a common probability space. $A_n \to_{a.s.} A$ and $A_n \to_{a.s.} B$ implies $A = B$ almost surely. $A_n \to_p A$ and $A_n \to_p B$ implies $A = B$ almost surely. Similarly, if $X_n \to_d X$ and $X_n \to_d \tilde{X}$, then $X \stackrel{d}{=} \tilde{X}$. This juxtaposition highlights that the weak limit X is only unique up to its distribution, whereas the other convergence limits are unique.
- **Portmanteau theorem.** TFAE definitions for weak convergence $\Rightarrow$. Some of these interpretations of weak convergence are more useful than others in proving certain results. Search this list for the appropriate definition for any proof at hand.
  (i) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous f.
  (ii) $P(X_n \leq x) \to P(X \leq x)$ for all continuity points x of $P(X \leq \cdot)$.
  (iii) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, Lipschitz-continuous f.
  (iv) $\limsup \mathbb{E}[f(X_n)] \leq \mathbb{E}[f(A)]$ for every upper semicontinuous f bounded above.
  (v) $\liminf \mathbb{E}[f(X_n)] \geq \mathbb{E}[f(X)]$ for every lower semicontinuous f bounded below.
  (vi) $\limsup P(X_n \in F) \leq P(X \in F)$ for all closed sets F.
  (vii) $\liminf P(X_n \in O) \geq P(X \in O)$ for all open sets O.
  (viii) $P(X_n \in C) \to P(X \in C)$ for all continuity sets C, i.e. sets C s.t. $P(X \in \partial C) = 0$.
  (ix) $\mathbb{E}[\exp\{it^T X_n\}] \to \mathbb{E}[\exp\{it^T X\}]$ for all vectors t. (Lévy continuity)
  (x) $t^T X_n \Rightarrow t^T X$ for all vectors t. (Cramér-Wold)
- **Continuous mapping.** Let f be continuous at every point of C s.t. $P(X \in C) = 1$.
  (i) $X_n \Rightarrow X$ implies $g(X_n) \Rightarrow g(X)$;
  (ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$;
  (iii) $X_n \to_{a.s} X$ implies $g(X_n) \to_{a.s} g(X)$.

- **Slutsky-like lemmas.**
  (i) $X_n \Rightarrow X$ and $||X_n - Y_n|| \to_p 0$ implies $Y_n \Rightarrow X$;
  (ii) $X_n \Rightarrow X$ and $Y_n \to_p c$ for constant c implies $(X_n, Y_n) \Rightarrow (X, c)$.

- **Slutsky's lemma.** This result provides a way to combine random vectors and random variables in the asymptote. Be careful with random vectors versus random variables in between (i) versus (ii) and (iii).
  (i) $X_n \Rightarrow X$ and $Y_n \to_p c$ for multidimensional constant c implies $X_n + Y_n \Rightarrow X + c$;
  (ii) $X_n \Rightarrow X$ and $Y_n \to_p c$ for 1-dim constant c implies $X_n Y_n \Rightarrow cX$;
  (iii) $X_n \Rightarrow X$ and $Y_n \to_p c$ for nonzero 1-dim constant c implies $X_n + Y_n \Rightarrow X/c$.

- **Laws of large numbers**. Let $\mathbb{E}[|X|] < \infty$. Then,
  - (weak law of large numbers) $\bar{X}_n \to_p \mathbb{E}[X]$;
  - (strong law of large numbers) $\bar{X}_n \to_{a.s} \mathbb{E}[X]$.

- **Prokhorov theorem.** This theorem relates to how $o_P(1)$ is linked with converge in probability. Here we see a relationship between $O_P(1)$ and weak convergence. The result is not quite an if and only if result. (ii) is similar to the Bolzano-Weierstrass theorem from real analysis.
  (i) $X_n \Rightarrow X$ for some X implies that $X = O_P(1)$.
  (ii) $X_n = O_P(1)$ implies that there is a subsequence $\{X_{n_i}\}$ s.t. $X_{n_i} \Rightarrow X$ for some X.

- **Central limit theorems.**
  - (vanilla univariate CLT) iid sample and finite second moment implies
  $$\sqrt{n}(\bar{X}_n - \mu) \Rightarrow N(0, \sigma^2);$$
  - (vanilla multivariate CLT) iid sample and finite expected norm squared implies
  $$\sqrt{n}(\bar{X}_n - \mu) \Rightarrow N_d(0, \sigma^2 \cdot \mathrm{Id}_d).$$
  - (Lindeberg-Feller) The setup is triangular array $\{X_{ni}\}_{i=1}^n$ with independent rows, $\mathbb{E}[X_{ni}] = \mu_{ni}$, finite $\mathrm{Var}(X_{ni}) = \sigma_{ni}^2$, $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2 > 0$, and $Y_{ni} = (X_{ni} - \mu_{ni})/\sigma_n^2$. Then, the Lindeberg condition implies $\sum_{i=1}^n Y_{ni} \Rightarrow N(0, 1)$.
  - (Lindeberg) For all $\varepsilon > 0$, $\sum_{i=1}^n \mathbb{E}[Y_{ni}^2 \cdot I\{|Y_{ni}| \geq \varepsilon\}] \to 0$ as n gets large.
  - (Lyapunov) For some $\delta > 0$, $\sum_{i=1}^n \mathbb{E}[Y_{ni}^{2+\delta}] \to 0$ as n gets large. The Lyapunov condition implies the Lindeberg conditon.

- **Delta methods.**
  - (univariate) If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at ψ_0 and $r_n(\psi_n - \psi_0) \Rightarrow Z$, then
  $$r_n(f(\psi_n) - f(\psi_0)) \Rightarrow \langle Z, \nabla f(\psi_0) \rangle;$$
  - (multivariate) if $f : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable at ψ_0 and $r_n(\psi_n - \psi_0) \Rightarrow Z$, then
  $$r_n(f(\psi_n) - f(\psi_0)) \Rightarrow J_f Z,$$
  where J_f is the Jacobian with respect to function f.

2.3 Examples

- Examples
  - The vanilla univariate central limit theorem is a special case of the Lindeberg-Feller central limit theorem when we have iid samples.
  - See slides 40-44 applying Lindeberg-Feller for simple linear regression with a fixed design. This example is also discussed in Amy Willis's BIOST 533. Moreover, we can find another Lindeberg-Feller example on the BIOST 533 final exam.
  - Samples from standard multivariate normals have nice weak convergence results. See Homework 2. The crux is to decompose the random vector into polar coordinates and consider generic orthogonal transformations.
  - Estimation of relative risk using a delta method. See end of Chapter 2 slides.
- Counterexamples
  - Convergence in probability does not imply almost sure convergence. We consider a sequence of indicators that splits [0,1] into halves, then thirds, then fourths, and so on. As n gets large, this sequence of random variables converges in probability to zero. However, the indicator is triggered infinitely often, so the sequence does not converge almost surely to zero.
  - Convergence in distribution does not imply convergence in probability. We consider a sequence that converges weakly to a symmetric distribution.
  - Dependent sequences that marginally converge weakly may not jointly converge weakly. We consider sequences where the covariance between random variables alternate between -1 and 1. With independence, marginal weak convergences imply joint weak convergence.

3 M-, Z- Estimation

We introduce two paradigms for deriving estimators for a parameter θ. M for"maximu" in M-estimation involves maximizing a criterion function. Z for "zero" in Z-estimation involves finding roots of a criterion function.

3.1 Definitions

- Empirical process notation provides shorthand

$$Pf \equiv \int f(x)\, dP(x)$$

$$P_n f \equiv \frac{1}{n}\sum_i^n f(X_i)$$

- If $\phi_0 \equiv \phi(\theta_0) \in \arg\max_\phi P_0 m_\phi$, then $\phi_n \in \arg\max_\phi P_n m_\phi$ is an M-estimator. More generally, we consider M_0 and M_n that do not have to be $P_0 m_\phi$ and $P_n m_\phi$.
- If ϕ_0 is a solution to $P_0 z_\phi = \underline{0}$, then the solution ϕ_n to $P_n z_\phi = \underline{0}$ is a Z-estimator. More generally, we consider Z_0 and Z_n that do not have to be $P_0 m_\phi$ and $P_n m_\phi$.
- We call $\{m_\phi : \phi \in S \supset \mathrm{Im}(\Phi)\}$ a P_0-GC class if $\sup_\phi |(P_n - P_0)m_\phi| = o_P(1)$
- The bracketing number measures how complex the class of functions $\mathcal{F}$ is. See slides 21-26 for an introduction with application in the Glivenko-Cantelli theorem.
- The root density $\theta \mapsto \sqrt{p_\theta}$ is **differentiable in quadratic mean** at θ if there exists a **pseudoscore** function $\dot{\ell}_\theta$ such that

$$\sup_{||h||=1} \int \left(\frac{\sqrt{p_{\theta+\varepsilon h}} - \sqrt{p_\theta}}{\varepsilon} - \frac{h^T \dot{\ell}_\theta}{2}\sqrt{p_\theta} \right)^2 d\mu \xrightarrow{\varepsilon \to 0} 0$$

  This definition is akin to differentiability except an integral is thrown into the mix. It is exactly what we require to weaken the regularity condition of first and second order differentiability for asymptotic normality. We say a model is QMD if its root density is QMD.

- The squared **Hellinger distance** is involved in QMD arguments.

$$H^2(P_\epsilon, P_0) \equiv \int \left(\sqrt{p_\epsilon} - \sqrt{p_0} \right)^2 d\mu$$

- $\mathcal{L}^2(\mu)$ contains μ-measurable functions f such that $\int f^2 d\mu$ is finite. This space is equipped with inner product and norm, so the triangle and reverse triangle inequalities hold.

3.2 Results

- Under conditions for Radon-Nikodym derivatives, the Kullback-Leibler divergence serves as an m_ϕ justifying maximum likelihood estimators as M-estimators.

- Under certain conditions, the M-estimation problem (maximizing) can be expressed as a Z-estimation problem (root-finding). On the other hand, $M_\theta(\phi) = -||Z_\theta(\phi)||$ expresses a Z-estimation problem as an M-estimation problem.

- **Uniform consistency.** Suppose that

  (i) a near-maximizer for M_n is available: ϕ_n satisfies $M_n(\phi_n) \geq \sup_\phi M_n(\phi) - o_P(1)$;

  (ii) M_n is uniformly consistent: $\sup_\phi |M_n(\phi) - M_0(\phi)| \to_p 0$; and

  (iii) ϕ_0 is well-separated: $\forall \varepsilon > 0$, $M_0(\phi_0) > \sup_{||\phi - \phi_0|| > \varepsilon} M_0(\phi)$.

  Then, $\phi_n \to_p \phi_0$.

  Finding a maximum is a good way to satisfy (i). In Homework 4 Problem 1(a), we formulate a case where missing condition (iii) results in the conclusion not holding.

- **Consistency of Z-estimators in one dimension.** Let Im(Φ)$\subset \mathbb{R}$ and, for all ϕ, $Z_n(\phi) \to_p Z_0(\phi)$. One or the other or both must hold:

  (i) $\phi \mapsto Z_n(\phi)$ is continuous and has one root ϕ_n.

  (ii) $\phi \mapsto Z_n(\phi)$ is nondecreasing and there is ϕ_n such that $Z_n(\phi_n) = o_P(1)$.

  ϕ_0 such that, for all $\varepsilon > 0$, $Z_0(\phi_0 - \varepsilon) < 0 < Z_0(\phi_0 + \varepsilon)$ implies $\phi_n \to_p \phi_0$.

  This result is analogous to Homework 4 Problem 4(b).

- **Glivenko-Cantelli theorems.**

  (i) If $\mathcal{F}$ is a class of functions with finite bracketing number for all $\varepsilon > 0$, $\mathcal{F}$ is P_0-GC:

  $$||P_n - P_0||_{\mathcal{F}} \equiv \sup_f |(P_n - P_0)f| = o_P(1).$$

  (ii) Suppose $\mathcal{F} \equiv \{f_\phi : \phi \in K\}$ for $K \subset \mathbb{R}^d$ compact. If $\phi(x) \mapsto f_\phi(x)$ is continuous for all x and there is an envelope function F satisfying $P_0 F < \infty$ for which $\sup_\phi |f_\phi(x)| \leq F(x)$ for all x, then the bracketing number is finite for $\varepsilon > 0$.

- Under regularity conditions, we have **asymptotic normality** for Z- and M-estimators. Regularity conditions are assumptions required such that we achieve our desired result. See Chapter 3 slides 29-36 and van der Vaart Theorems 5.21 and 5.23. Below we write some of these regularity conditions for $Z-$estimation.

(i) ϕ is an open subset of $\mathbb{R}^d$;

(ii) $\mathbb{E}_0||z_{\phi_0}(X)||^2 < \infty$;

(iii) $\phi \mapsto P_{z_\phi}$ differentiable at zero ϕ_0 with nonsingular Jacobian matrix V_{ϕ_0};

(iv) There exists function G satisfying $P_0 G^2 < \infty$ so that, for all x and every ϕ and $\tilde{\phi}$ in some neighborhood of ϕ_0, $||z_\phi(x) - z_{\tilde{\phi}}(x)|| \leq ||\phi - \tilde{\phi}||G(x)$;

(v) There exists ϕ_n satisfying $P_n z_{\phi_n} = o_P(n^{-1/2})$ and $\phi_n = \phi_0 + o_P(1)$

- **Sufficient conditions for QMD.** For every θ in an open subset of $\mathbb{R}^d$, if

  (i) The root density is continuously differentiable for every x

  (ii) The information (matrix) is well-defined and continuous in θ

  then the root density is QMD and the pseudoscore is the score.

- QMD implies that the score is mean zero and that the information (matrix) exists.

- **Asymptotic normality of MLEs under QMD**. We list the regularities required.

  (i) The model is QMD at an inner point θ_0 in Θ

  (ii) There is a measurable function G with $P_0 G^2 < \infty$ such that for every θ_1 and θ_2 in a neighborhood of θ_0

  $$|\ell_{\theta_1}(x) - \ell_{\theta_2}(x)| \leq G(x)||\theta_1 - \theta_2||$$

  (iii) I_{θ_0} is nonsingular

  (iv) MLE is consistent

  Then,

  $$\sqrt{n}(\hat{\theta} - \theta_0) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum \dot{\ell}_{\theta_0}(X_i) + o_P(1) \Rightarrow N(0, I_{\theta_0}^{-1})$$

3.3 Examples

- Method of moments estimators are Z-estimators.
- The sample median ϕ_n satisfies $(1/n)\sum_i \text{sign}(X_i - \phi) = 0$ in one dimension.
- Location-scale families are QMD under assumptions. See Homework 4 Problem 2.
- See Homework 4 Problem 3 for some general QMD models.

4 Hypothesis Testing

First, we compare the Wald, score (Rao), and likelihood ratio tests that all converge in distribution to χ^2 under regularity conditions. Second, we consider local alternatives where in sampling from the alternative we maintain desirable properties for our estimators. See Chapter 4 slides for the hypothesis testing framework.

4.1 Definitions

- The (randomized) test function $\phi_n(X)$ indicates when we reject the null.
- The **power** function $\pi_n(\theta) = \mathbb{E}_\theta[\phi_n(X)]$ measures the probability we reject the null.
- The **size** of the test is $\sup_{\theta_0 \in \Theta_0} \pi_n(\theta_0)$.
- Q is absolutely continuous w.r.t. P means that $P(A) = 0$ implies $Q(A) = 0$.
- Q_n is **contiguous** w.r.t P_n means that $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$.
- A **local alternative** is some $\theta + h/\sqrt{n}$ where h describes some (small) perturbation from the null in an arbitrary direction.
- A **regular estimator** is an estimator whose sampling distribution is invariant to local perturbations of the data-generating distribution.

4.2 Results

- **Wald test.** This test rejects the null when the estimate $\hat{\psi}$ of ψ is far from zero. The test statistic $W_n \equiv n\hat{\psi}^T A_{\hat{\theta}} \hat{\psi} \Rightarrow \chi^2(m)$ under the null where m is the dimension of the space ψ lives in. This test can be easy to implement when considering many hypotheses because the we only find one MLE.
- **Likelihood ratio test.** This test rejects the null when the KL divergence is large. The test statistic $L_n \equiv 2nP_n[\ell_{\hat{\theta}} - \ell_{\hat{\theta}_0}] \Rightarrow \chi^2(m)$. This test better controls the type 1 error in small samples.
- **Score test.** This test rejects the null when the empirical mean of the score is far from zero. The test statistic $S_n \equiv Z_n(\hat{\theta}_0) I_{\hat{\theta}_0}^{-1} Z_n(\hat{\theta}_0) \Rightarrow \chi^2(m)$. This test is easy to implement because Θ_0 is often a lower-dimensional space.
- The pairwise differences between these three tests converge in probability to zero under the null hypothesis.

- **Le Cam's First Lemma**. This lemma helps us to show and characterize contiguity. TFAE:

  (i) Q_n is contiguous w.r.t. P_n

  (ii) $L_n \equiv \frac{dQ_n^a}{dP_n}(Z_n) \overset{P_n}{\Rightarrow} V$ along a subsequence implies that $\mathbb{E}[V] = 1$.

  (iii) $\frac{dP_n^a}{dQ_n}(Z_n) \overset{Q_n}{\Rightarrow} U$ along a subsequence implies that $P(U > 0) = 1$.

- **Le Cam's Third Lemma.** This lemma shows us how to use contiguity when studying alternatives. Suppose that Q_n is contiguous w.r.t. P_n and $(T_n, L_n) \overset{P_n}{\Rightarrow} (T, V)$. For all measurable $A \subset \mathbb{R}^d$, let $R(A) = \mathbb{E}[I_A(T)V]$. Then R is a probability measure and $T_n \overset{Q_n}{\Rightarrow} R$.

- **Asymptotic normality of log likelihood ratio.** Suppose $\log L_n \overset{P_n}{\Rightarrow} N(\mu, \sigma^2)$. Then Q_n is contiguous w.r.t. P_n if and only if $\mu = -\sigma^2/2$. This result emphasizes that it can sometimes be useful to consider the weak limit of the log likelihood ratio.

- See van der Vaart Theorem 7.2 and Chapter 4 slides 30-40. With this theorem and Le Cam's Third Lemma, we derive results for Wald tests under local alternatives and regular estimators.

- We achieve perfect asymptotic power at fixed alternatives. That is, if we collect enough data, we will always reject a false null hypothesis.

- **Wald statistic under local alternative.** We assume the regularity conditions for the asymptotic normality of the MLE. Then, the Wald statistic W_n has a noncentral $\chi^2(m)$ weak limit where the noncentrality parameter is $h_\psi^T A_\theta h_\psi$. Because χ^2 random variables are stochastically increasing in their noncentrality parameter, the Wald test achieves non-trivial power at local alternatives.

- We skipped the section on relative efficiency. See the final Chapter 4 slides.

5 Optimality

We study when the MLE is an optimal estimator.

5.1 Results

- **Pointwise asymptotic optimality.** Suppose there is an estimator θ_n for each θ such that $\sqrt{n}(\theta_n - \theta) \overset{\theta}{\Rightarrow} Q_\theta$. For any θ there exists another estimator $\tilde{\theta}_n$ such that

$$\sqrt{n}(\tilde{\theta}_n - \theta') \overset{\theta'}{\Rightarrow} \begin{cases} Q_{\theta'} & \theta' \neq \theta \\ \delta_0 & \text{otherwise} \end{cases}$$

  That is, we can always construct a dominating estimator sequence. The above is the Hodges' estimator. See graph in Chapter 5 slides and van der Vaart Chapter 8.

- **Almost everywhere convolution.** Assume a QMD model at every θ with nonsingular information I_θ. Suppose $\sqrt{n}(\theta_n - \theta) \overset{\theta}{\Rightarrow} Q_\theta$ for every θ. Then, **for almost every** θ, there exists M_θ s.t.

$$Q_\theta \equiv Z + \epsilon$$

  where $Z \sim N(0, I_\theta^{-1})$ and $\epsilon \sim M_\theta$.

- **Convolution for regular estimators.** Let θ_n be a regular estimator sequence. Under the same conditions as the a.e. convolution theorem, **for all** θ, there exists M_θ such that

$$Q_\theta \equiv Z + \epsilon$$

  where $Z \sim N(0, I_\theta^{-1})$ and $\epsilon \sim M_\theta$.

- **Anderson's lemma.** This theorem coupled with a convolution theorem shows that the MLE is asymptotically optimal, i.e. achieves a lower bound. Let $Z \sim N(0, \Sigma)$ and $\epsilon \sim M$ be independent. If the loss L is **quasiconvex** and **centrally symmetric**, then $\mathbb{E}[L(Z)] \leq \mathbb{E}[L(Z + \epsilon)]$.

6 Miscellaneous

- **Reverse triangle inequality.**

$$\Big| ||x||_2 - ||y||_2 \Big| \leq ||x-y||_2$$

- **Differentiability in higher dimensions.** Below are two representations of differentiability at ψ_0. The first representation is especially useful for proving delta methods. For all $\varepsilon > 0$,

$$\lim_{\varepsilon \to 0} \sup_{||h||=1} \frac{|f(\psi_0 + \varepsilon h) - f(\psi) - \varepsilon \langle h, \nabla f(\psi_0) \rangle|}{\varepsilon} \to 0;$$

$$\lim_{h \to 0} \frac{||f(\psi_0 + h) - f(\psi) - J_f(h)||}{||h||} \to 0.$$

  Read more about the second representation at the Wikipedia article for differentiable functions in higher dimensions. A sufficient condition for differentiability is that partial derivatives exist and the linear map J_f is the Jacobian matrix. In class, we claim this sufficient condition as that f is partially differentiable in a neighborhood around ψ_0 and the partial derivatives are continuous at ψ_0. Lastly, note that h in the two representations are different!

- Find a Lipschitz constant by finding the maximum of the norm of the gradient.

- **Taylor series.** We approximate differentiable functions by the following and the remainder vanishes in the asymptote.

$$f(x) = \frac{f^{(0)}(a)}{0!}(x-a)^0 + \frac{f^{(1)}(a)}{1!}(x-a)^1 + \frac{f^{(2)}(a)}{2!}(x-a)^2 + \text{remainder}$$

- **Bolzano-Weierstrass.** Every bounded sequence has a convergent subsequence.

- **Semi-continuity.** See graphs under Examples.
  - A function f is lower-semicontinuous at x_0 if for all $\varepsilon > 0$ there is a neighborhood around x_0 such that $f(x) \geq f(x_0) - \varepsilon$
  - A function f is upper-semicontinuous at x_0 if for all $\varepsilon > 0$ there is a neighborhood around x_0 such that $f(x) \leq f(x_0) + \varepsilon$.

- **Fisher-Cramér.** This result from MDP's STAT 513 says that the MLE is asymptotically consistant and normal if we have a regularity condition the depends of first and second derivatives of the log likelihood. Important estimators like medians may not fit into this framework. We weaken this assumption using QMD.

- We say that X is stochastically larger than Y if if $P(X \leq x) \leq P(Y \leq x)$.

- Use this equality $\inf(-x_n) = -\sup x_n$ to redefine the $X_n = O_P(R_n)$ notation.