300 Quals Guide

July 5, 2025

1 Misc facts

Definition 1 (Convexity)
For all $0 \leqslant t \leqslant 1$ and $x_1, x_2 \in X$:

$$f(tx_1 + (1-t)x_2) \leqslant tf(x_1) + (1-t)f(x_2).$$

Alternative, check the second derivative $\geqslant 0$

Theorem 2 (Jensen's Inequality)
If ϕ convex, then $\mathbf{E}\phi(X) \geqslant \phi(EX)$. Eg $EX^2 \geqslant (EX)^2$. Inequality is strict if ϕ is strictly convex and X is not degenerate (constant). Also conditional version holds.

2 Exponential Families

Definition 3 (Exponential Family)

$$p_\theta(x) = \exp(\sum_{i=1}^{s} T_i(X)\eta_i(\theta) - A(\theta))h(x)$$

$$p_\eta(x) = \exp(\sum_{i=1}^{s} T_i(X)\eta_i - \tilde{A}(\eta))h(x)$$

Fact 4 (Expectation and covariance of sufficient stats in exponential fam)

$$\mathbf{E}[T(X)] = \nabla A(\eta)$$

$$\mathbf{Cov}(T_i(X), T_j(X)) = \partial_{\eta_i}\partial_{\eta_j} A(\eta)$$

2.1 All the Expo family examples

Include curved..

3 Sufficiency

"Throwing away everything else besides this statistic entails no loss of info in estimating θ"

Definition 5 (Sufficiency)
If for all t, the distribution of $X|T = t$ does not depend on θ.

If you collect data $X_1, \ldots, X_n$, then throw away data except for $T = t$, then can construct $\tilde{X}_1, \ldots \tilde{X}_n$ with same distribution as $\underline{X}$ just by knowing T.

Make a table of good examples including Unif$(0, \theta)$

Theorem 6 (Factorization Theorem)
T sufficient for θ (or for the model) $\iff$ we can write $p_\theta(\underline{x}) = g_\theta(t(\underline{x}))h(\underline{x})$. Ie, can factor the density into a part that depends on statistic and θ, and another part that depends on data but not θ.

Example 7 (Expo(θ))
Can write $p_\theta(\underline{x}) = \theta^n \exp(-\theta \sum x_i)$ so $T(X) = \sum_{i=1}^n X_i$ is sufficient.

Example 8 (Exponential families)
Look at the form – T is sufficient.

Definition 9 (Minimal sufficiency)
For all T' sufficient, T is a function of T'.

Theorem 10 (Rao-Blackwell)
If T sufficient, L is any convex loss, and δ estimates $g(\theta)$, define:

$$\eta(T) = \mathbf{E}[\delta(X)|T],$$

then using Tower property and Jensen's:

$$R(\theta, \delta(X) = \mathbf{E}[g(\theta), \delta(X)] \geqslant \mathbf{E}[L(\theta, \eta(T)) = R(\theta, \eta(T))$$

Definition 11 (Ancillary Statistic)
Distribution of $A(X)$ does not depend on θ.

Definition 12 (First Order Ancillary)
If $\mathbf{E}_\theta A(X)$ does not depend on θ. Weaker than Ancillary.

4 Completeness

Definition 13 (Completeness)
$T(X)$ complete if no non-constant function of T is even 1st order ancillary, or equivalently:

$$\mathbf{E}f(T) = 0 \implies f(T) = 0 \text{ a.s.}$$

Example 14 ($X_i \sim \text{Unif}(0, \theta)$)
See coaching notes– break h into positive and negative parts to show $h = 0$ as.

Theorem 15 (Full rank exponential family)
Ie if $(\eta_1, \dots, \eta_k) : \eta \in \Omega$ contains a k–dimensional rectangle, then T is not only sufficient, but **also complete**.

Theorem 16 (Basu's Theorem)
If T is CSS and A Ancillary, then $T \perp\!\!\!\perp A$.

A common strategy with Basu's theorem is to consider a submodel– show independence in the submodel, then based on arbitrariness of submodel, show true for full model.

Example 17 (Show sample mean and sample variance in $N(\mu, \sigma^2)$ are independent)
(Assume both unknown). then in the submodel where $\sigma = \sigma_0$ known, then $\overline{X}$ is CSS. Can show that $\sum(X_i - \overline{X})^2$ is Ancillary. So $\overline{X} \perp\!\!\!\perp \sum(X_i - \overline{X})^2$. This is true for all μ, and for $\sigma = \sigma_0$ fixed. But σ_0 arbtirary, so true for all μ, σ.

Add a table here including 2 param Unif

1. $\text{Unif}(0, \theta) : T = X_{(n)}$
2. $\text{Unif}(\theta_1, \theta_2)$: $T = (X_{(1)}, X_{(n)})$

5 UMVU

Note that we can't take existence of an unbiased estimator for granted. If $x \sim Bin(n, \theta)$. Take like $g(\theta) = \frac{1-\theta}{\theta}$ or a polynomial of degree $> n$– no unbiased estimators. Strategy to show no unbiased estimator is just write out the expectation of an estimator.

Definition 18 (UMVU)
δ^* is unbiased and for all other δ, $R(\delta^*, g(\theta)) \leqslant R(\delta, g(\theta))$ for all θ.

Note that if we have an unbiased estimator, we can Rao-Blackwellize with a sufficient statistic and still have unbiased estimator.

Theorem 19 (Lehman-Scheffe)
If unbiased estimator is function of CSS, it is UMVU. (Since there exists at most one unbiased estimator that's function of T by completeness).

Alternate statement: if there exists any unbiased estimator and a CSS T, then there is a unique unbiased estimator that's a function of T, which is UMVU (UMRU for any convex loss). Unique UMRU if strict convex loss, since this makes Jensen strict.

Recipe 20 (UMVU with CSS)
Steps:

1. Find a CSS
2. Find an unbiased estimator function of CSS. (Alternatively, find a dumb unbiased estimator and RB)
3. $\implies$ UMVU

Note that UMVU can be inadmissible - see James-Stein or the Poisson UMVU for $g(\theta)$. $= \exp(-3\lambda)$ example in Lecture 5.

Theorem 21 (Orthogonality Condition)
Let $\hat{\theta}$ be an unbiased estimator with $E_P\hat{\theta}^2 < \infty$ if $U(X)$ is an unbiased estimator of 0 for all P, then:

$$\hat{\theta} \text{ UMVU} \iff E_P[\hat{\theta}(X)U(X)] = 0 \text{ for all unbiased estimators of 0 and for all P}$$

An application is showing that the addition of UMVUs is UMVU. Also, product of UMVUs is UMVU for its expectation.
Also reasonable to try to show not UMVU. Come back to Ex 2.3 in Notes

Theorem 22 (Cramer Rao LB)
For any $\hat{\theta}$ unbiased for θ:

$$\mathbf{Var}_\theta \hat{\theta} > I_\theta^{-1}$$

when $\{P_\theta\}$ is QMD with non singular I_θ.

5.1 UMVU Examples

1. Basic Expo Fam examples – eg $\overline{X}$ in Bernouli or Normal with known variance
2. Empirical CDF in $\mathcal{N}(\theta, 1)$.
   - CSS is $\overline{X}$, $\delta = \mathbf{1}[X_1 < u]$ unbiased.
   - Idea to RB then add and subtract $\overline{X}$ since $X_1 - \overline{X}$ is ancillary, then apply Basu
   - UMVU is $\Phi(\frac{u-\overline{X}}{\sqrt{(n-1)/n}})$

Non parametric examples

1. $X_i \sim F \in \mathcal{F} = \{$ all distributions with density wrt Lebesgue and finite variance $\}$. $g(\theta) = E_F X_i$.
   - Note that $\overline{X}$ is unbiased in the big family and is UMVU in the normal **subfamily**
   - Order statistics CSS (always sufficient - complete by subfamily arg and bijection with sums of powers): $(X_{(1)}, \dots, X_{(n)}) \iff (\sum X_i, \dots, \sum X_i^n)$ bijection.
   - So $\overline{X}$ is UMVU
2. X_i iid symmetric about θ, $EX_i = \theta$. Finite variance.
   - Two subclasses: normal family, Unif(θ_1, θ_2) family - have different UMVUs and both are unbiased in the original class

Recipe 23 (Subfamily UMVU Argument)
If UMVU in subfamily and unbiased in big family, must be UMVU in big family **if** an UMVU exists in the big family. Because the UMVU **uniquely** minimizes variance in the subfamily.

Proof. Take the big family. Take a function such that $E_\theta f(X) = 0$ for all θ. Then true for subfamily. So $P_\theta(f(X) = 0)$ for all θ in the subfamily. Same null sets as big family, means that also $P_\theta(f(X) = 0)$ for the big family. ☐

Fact 24 (Completeness in subfamily)
If $\mathcal{F}_0 \subset \mathcal{F}$ and they have the same null sets, then completeness in $\mathcal{F}_0$ implies completeness in $\mathcal{F}$.

Recipe 25 (Non-existence of Non-parametric UMVU)
Find two different subclasses with different unique UMVU that are also unbiased in the big class – no UMVU in big class.

If we can't use Lehman-Scheffe, try one of the following ideas:

1. Cramer-Rao lower bound
2. Orthogonality condition
3. Subfamily arguments

5.2 Non-convex loss functions

If loss is bounded, there is no UMRU estimator. This is unbiased:

$$\delta_\pi = \begin{cases} g(\theta_0) \text{ wp } 1-\pi \\ \frac{1}{\pi}[\delta_0(X) - g(\theta_0)] + g(\theta_0) \end{cases}$$

and its risk is πM, so it could be arbitrarily small.

6 MRE

Big picture– there's an easy recipe for MRE for square error or absolute error if we pick any old δ_0 equivariant function of CSS.

Definition 26 (Location models)
Density satisfies

$$f_{\theta+h}(x+h) = f_\theta(x)$$

Think: normal at its mean has same density as a shifted normal at shifted mean.
A **location invariant loss** is $\ell(a+h, \theta+h) = \ell(a, \theta)$ for all a, θ, h.

Want location equivariant estimators, ie $\hat{\theta}(x+h) = \hat{(\theta)} + h$ – if our observations all shift by some amount, our estimate of the location should also shift by that amount.

Fact 27 (Bias, variance, risk of equivariant estimators)
Do not depend on θ. Ie,

$$\mathbf{E}_\theta \hat{\theta} = \theta + b \text{ for all } \theta$$

for risk, since it's constant, we can hope to find the best among all equivariant estimators.

Definition 28 (MRE)
Satisfies equivariance condition:

$$\hat{\theta}(x + c, u) = \hat{\theta}(x, u) + c$$

and minimum risk condition amongst all equivariant estimators.

Definition 29 (Characterization of location **invariant** estimators)
Location invariant means $U(X_1 + c, \dots, X_n + c) = U(X)$. Characterize by: U location invariant $\iff U = V(y_1, \dots, y_{n-1})$ where $y_i = X_i - X_n$.

Fact 30 (Characterization of location **equivariant** estimators)
Let δ_0 be **any** location equivariant estimator, eg $\delta_0 = \overline{X}$.

$$\delta \text{ is location equivariant } \iff \delta(X_1, \dots, X_n) = \delta_0(X_1, \dots, X_n) + U(X_1, \dots, X_n)$$

where U is location **invariant**. Then from invariant characterization:

$$\iff \delta(X_1, \dots, X_n) = \delta_0(X_1, \dots, X_n) + V(Y_1, \dots, Y_{n-1})$$

Recipe 31 (Finding the MRE)
Let X_i observations iid from location model and let $Y = (X_1 - X_n, \dots, X_{n-1} - X_n)$. Let $\hat{\theta}_0$ be any location equivariant estimator of θ_0 with finite risk. If the following is well-defined:

$$v(y) = \arg\min_v E_0[\ell(\hat{\theta}_0(X) - v, 0) \mid Y = y]$$

Then there exists an MRE $\hat{\theta}^*(X) = \hat{\theta}_0(X) - v(Y)$.

Want to minimize

$$\mathbf{E}_\theta[\rho(\delta_0(X) - V(Y) - \theta)] = \mathbf{E}_0 \rho(\delta_0(X) - V(Y))$$

Apply Tower property and minimize the inner expectation. For square error yields:

$$V = \mathbf{E}_0[\delta_0(X)|Y)]$$

So $\delta^* = \delta_0 - E_0[\delta_0|Y]$ Notice that the Y_i here are ancillary so if we pick δ_0 function of CSS, easy.
If we can make δ_0 a function of CSS, then we can apply Basu

Recipe 32 (MRE for Square Error Loss)
Choose δ_0 equivariant function of CSS. Then $\delta^* = \delta_0 - E_0\delta_0$ is MRE.

Example 33 ($X_i \sim N(\theta, 1)$)
Let $\delta_0 = \overline{X}$. Then $E_0\delta_0 = 0$. So $\overline{X}$ is MRE.

Note for absolute error, just do $\text{median}_{\theta=0}[\delta(X)|Y]$ instead of expectation. Ie find m such that $P(X \leqslant m) \geqslant 1/2$ and $P(X \geqslant m) \geqslant 1/2$.

Theorem 34 (Existence of MRE)
If loss is convex and not monotone, then MRE exists by the previous theorem. If strictly convex, unique.

Fact 35 (MRE is Unbiased (under squared error loss))
Since the bias does not depend on θ, just subtract off whatever bias.

Fact 36 (UMVU is MRE if UMVU is location equivariant)
In a location model, the UMVU is location equivariant. Since MRE is the best amongst equivariant estimators, and any competing equivariant estimators are unbiased.. Is this just with square error loss?

Theorem 37 (Anderson's Lemma)
Review this $Z \sim N(0, \Sigma)$, and ℓ is bowl shaped, then $\mathbf{E}\ell(Z) \leqslant \mathbf{E}\ell(Z + U)$ for any $U \perp\!\!\!\perp Z$.

Definition 38 (Pitman Estimator)

Theorem 39 (Mini-convolution theorem)

7 James-Stein Estimator

Setup: $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known.

Theorem 40 (SURE - Stein Unbiased Risk Estimator)
Letting $\hat{\mu}(x) = x + g(x)$ for $g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ almost differentiable, and assume that $\mathbf{E}[\sum_{i=1}^p |\partial_i g_i(X)|] < \infty$. Then:

$$\mathbf{E}_\mu[\|\hat{\mu}(X) - \mu\|^2] = p\sigma^2 + \mathbf{E}\left[\|g(X)\|^2 + 2\sigma^2 \sum_{i=1}^p \partial_i g_i(X)\right].$$

Proved using integration by Parts – see Lec 17 300c.

Fact 41 (UMVU is not admissible in $\mathcal{N}(\mu, 1)$ model)
Because James Stein renders X inadmissible.

J-S estimator is given by:

$$\hat{\mu}^{JS}(X) = \left(1 - \frac{\sigma^2(p-2)}{\|X\|_2^2}\right) X,$$

biased towards the origin. Prove that it has better risk by SURE estimator.

8 Bayes Estimators

Average risk over some choice of prior $\Lambda(\theta)$ on parameter space.

Want to minimize:

$$\int R(\theta, \delta) d\Lambda(\theta) = \mathbf{E}_{(X,\Theta)} L(\Theta, \delta(X)).$$

So our P_θ is $X|\Theta = \theta$. By tower property, we can minimize this by minimizing the following for almost all x:

$$\arg\min_{\delta} \mathbf{E}[L(\Theta, \delta(X) \mid X = x]$$

ie this is an integral with respect to posterior distribution. For square error: $\delta = \mathbf{E}[g(\Theta) \mid X]$, abs error gives $\delta = \text{median}\, g(\Theta)|X$. See the quals notes for a list. For square error, often (?) gives a convex combination of UMVU and the prior mean.

Theorem 42 (When Bayes is Unique?)
???

Fact 43 (Unique Bayes is Admissible)
Idea is that if $R(\hat{\theta}', \theta) \leqslant R(\hat{\theta}, \theta)$ for all θ, it would then be Bayes.
provided the prior isn't super weird (eg continuous dist with an atom)

Fact 44 (Constant risk Bayes is minimax)
If not, some other estimator would render Bayes inadmissible, which would make that estimator the Bayes estimator.

Fact 45 (Bayes is not UMVU if $r_\Lambda < \infty$)
Under square error loss, Bayes estimators are biased, unless $r_\Lambda = 0$.

An example of an unbiased Bayes estimator would be if $X \sim Bin(n, \theta)$ and Λ puts mass on $\{0, 1\}$ only. Then there exists an estimator of 0 risk– X/n.

9 Minimax

Want δ to minimize $\sup_{\theta\in\Omega} R(\theta, \delta)$.

Note that for any prior, the bayes risk is upper bounded by the minimax (frequentist) risk.
If δ' is our guess:

$$\inf_\delta \int R(\delta, \theta) d\Pi(\theta) \leqslant \inf_\delta \sup_\theta R(\delta, \theta) \leqslant \sup_\theta R(\delta', \theta).$$

If we can find a prior such that δ' is Bayes, then $\inf_\delta \int R(\delta, \theta) d\Pi(\theta) = \int R(\delta', \theta) d\Pi(\theta)$. If this also equals the sup risk RHS, then δ' is minimax.

Next best thing is a sequence Π_n such that $r_{\Lambda_n} \longrightarrow \sup_\theta R(\delta', \theta)$. Then we squeeze the above inequality by taking a limit in $n-$ so that

$$\inf_\delta \sup_\theta R(\delta, \theta) = \sup_\theta R(\delta', \theta)$$

.

Recipe 46 (Minimax - Constant risk)
A Bayes or admissible estimator is constant risk, then it is minimax.

So: **find a prior with constant frequentist risk.**

1. Hopefully conjugate prior
2. Find Bayes estimator
3. Write the frequentist risk $R(\theta, \delta_\Lambda) = \mathbf{E}_X L(\theta, \delta_{Lambda})$
4. Find parameters of the prior Λ such that $R(\theta, \delta_\Lambda)$ is constant
5. Conclude this estimator is minimax

If unique Bayes, unique minimax. Prior Λ is *least favorable.*

The above is most useful in case when parameter space Θ is compact. Is the supremum achieved by some Λ, making the frequenist constant, so that the inequality at beginning of section is equality.

Example 47 (Multinomial minimax for p)
See Theory Coaching. Use conjugate prior Dirichlet. One idea is that a least favorable prior must be symmetric in $\alpha_1, \ldots, \alpha_n$.

Analogous to in class Binomial minimax for p using Beta prior.

Definition 48 (Minimax - Least favorable prior)
Λ least favorable if $r_\Lambda \geqslant r_{\Lambda'}$ for any other Λ'.

A sequence of priors Λ_m is least favorable if for any Λ', $r_{\Lambda'} \leqslant \lim_m r_{\Lambda_m}$.

Recipe 49 (Minimax - Least favorable sequence of priors)
If $\sup R(\theta, \delta) = r$ and have a sequence such that $r_{\Lambda_m} \longrightarrow r$ then Λ_m is least favorable and δ is minimax.

1. Guess some δ by looking at limit of Bayes estimator
2. Find the sup risk of our guess δ, ie $\sup_\theta R(\delta, \theta)$.
3. Find a sequence of priors s.t. $r_{\Lambda_m} \longrightarrow r$
4. Conclude δ minimax

Example 50 (Normal location model minimax)
If $X_i \overset{iid}{\sim} N(\theta, \sigma^2)$ with σ known, then we can show that $\overline{X}$ is minimax by the above approach. Let $\Lambda_m = N(0, b_m^2)$ with $b_m \longrightarrow \infty$ eg $b_m = m$. Note that $\sup R(\theta, \overline{X}_n) = \sigma^2/n$– it's constant risk so a good candidate.

$$r_{\Lambda_m} = \mathbf{E}[\mathbf{Var}(\Theta|X)] \longrightarrow r$$

Conclude $\overline{X}$ is minimax.

Connection that just because an estimator is minimax doesn't mean it's admissible. $\overline{X}$ is UMVU, minimax, MRE.. but not admissible.

Recipe 51 (Minimax in subfamily argument)
δ minimax in Λ_0 then it is also minimax in Λ **if**

$$\sup_{\theta \in \Lambda_0} R(\theta, \delta) = \sup_{\theta \in \Omega} R(\theta, \delta)$$

Eg, $\overline{X}$ in $\Lambda_0 = \{\sigma^2 = b\}$ and $\Lambda = \{\sigma^2 \leqslant b\}$.

Can extend to non parametric big family– $\overline{X}$ is minimax if $X_i \overset{iid}{\sim} F$ since don't increase sup risk in the larger family.

10 Testing

Types of errors:

1. Type 1: Reject the null when the null is true (restricted to size α size of test)
2. Type 2: Retain the null when the null is false (given by 1-power)

10.1 Simple-simple

Theorem 52 (Neyman-Pearson (NP) Lemma)
For simple versus simple, a MP test always exists and is given by

$$\phi(x) = \begin{cases} 1 & p_1(x) > kp_0(x) \\ \gamma & p_1(x) = kp_0(x) \\ 0 & p_0(x) < kp_0(x) \end{cases}$$

where constants determined by level constraint. This is just the likelihood ratio test. First find a k, then find the γ.

Also note that MP test must have level exactly α unless there exists a test with power 1 with size $< \alpha$.

Note that non-uniqueness comes from the = condition– how to randomize.

Fact 53 (NP characterizes MP tests)
The NP is actually an if and only if statement. If MP test, then satisfies the above along with level constraint. MP test must be of this form.

10.2 Simple-Composite

A UMP test only exists if the NP test is the same for **all** alternatives! If not, have different MP tests, no UMP test! If you want to show that a UMP does not exist, show that the MP test depends on the alternative– to show exist, does not depend. Eg, testing normal $\theta = 0$ versus $\theta \neq 0$, we get different MP tests depending on which side our alternative is on.

Occurs most often in a one-dimensional parametric family with monotone-likelihood ratio.

Definition 54 (Monotone Likelihood Ratio (MLR))
If exists $T : \mathcal{X} \longrightarrow \mathbb{R}$ statistic that likelihood ratio $p_{\theta'}/p_{\theta_0}$ is **non-decreasing** in T *whenever* $\theta_0 < \theta'$.

Ie if we have a singleton null and for alternative $\mathcal{H}_1 : \theta' > \theta_0$, then we check whether there's a statistic such that likelihood ratio is increasing.

Theorem 55 (UMP in MLR)
Testing $\theta = \theta_0$ versus $\theta > \theta_0$, if there's a MLR, then the UMP test given by

$$\phi(x) = \mathbf{1}[T(X) > C] + \gamma\mathbf{1}[T(X) = C].$$

Fact 56 (MLR in Exponential Families)
Exponential families satisfy MLR! Also satisfied in non-central chi-square.

10.3 Composite-Simple

An important insight is that:

$$\{\text{tests valid for full composite null}\} \subset \{\text{tests valid for a specific point in the null}\}.$$

So if you can find a best test for a specific point, and it is also valid for the full null, then it is the best for the composite null.

If we have a composite null, to find UMP test we put a prior on the null parameter space and reduce to a simple-simple. If the MP test in this simple-simple problem is also **valid** for the original problem null, then it is the UMP for composite-simple testing problem. Ie:

$$\tilde{P}_0(\cdot) = \int_{H_0} P(\cdot)d\Pi(P) \quad \text{weighted average of null measures}$$

Definition 57 (Least favorable prior)
The prior with the lowest powered MP test, ie $\beta_\Lambda \leqslant \beta_{\Lambda'}$ for all Λ' prior on Θ_0.

Recipe 58 (Composite-Simple)
To find UMP test:

1. Posit a least favorable prior
2. Compute the simple-simple MP test ϕ_Π wrt the simple null from the LFP
3. Show that ϕ_Π is *valid* for the original problem, ie $\mathbf{E}_{\theta_0}\phi_\Pi \leqslant \alpha$ for all $\theta_0 \in \mathcal{H}_0$.
4. Conclude that ϕ_Π is UMP in the original problem

Unique UMP if ϕ_Π is unique in the simple-simple case.

For least favorable prior, try to make it hard to reject. Ie, mass on the boundary.

10.4 Composite-Composite

UMP rarely exists in composite-composite. To check, always first fix an alternative, find MP test for composite-simple case. Then show that it does or doesn't depend on the alternative chosen. Here are the times when it does:

1. One-side MLR testing: $\mathcal{H}_0 : \theta \leqslant \theta_0, \mathcal{H}_1 : \theta > \theta_0$
2. Bioequivalence: Test null that $\theta \notin (\theta_1, \theta_2)$. If we have a 1-param expo family, then $\phi(x) = \mathbf{1}[T \in (C_1, C_2)]$ is UMP
3. One sided testing of Gaussian mean, ie $\mathcal{H}_0 : v^T\mu = \delta$ versus $\mathcal{H}_1 : v^T\mu > \delta$ for fixed v with $v^T\Sigma v > 0$, then reject for large $v^T X$.

10.5 Examples

Example 59 ($X \sim N(\mu_x, \sigma^2, Y \sim N(\mu_y, \sigma^2)$ with σ known)
Want to test $\mu_y \leqslant \mu_x$ versus not.

1. Fix an alternative θ_x, θ_y.
2. Put a point mass prior on average of the $\mu_x = \mu_y$ line ie $(\theta_x + \theta_y)/2$ for both
3. Get NP test: $\frac{Y-x}{\sqrt{2}\sigma} > z_{1-\alpha}$
4. Need to check that level $\alpha for original problem.$ So fix a null μ_x, μ_y such that $\mu_y \leqslant \mu_x$.
5. Conclude that MP for testing $\mathcal{H}_0$ versus (θ_x, θ_y)
6. Since doesn't depend on the specific alternative chosen, it's UMP

Example 60 (Non-parametric - hypothesis about the cdf $\mathcal{H}_0 : F_{X_i}(u) \geqslant p_0$)
Idea is to parametrize P by conditional distributions:

1. $p = P(X_i \leqslant u)$
2. P^- Distribution of $X_i|X_i \leqslant u$
3. P^+ Distribution of $X_i|X_i > u$

Least favorable prior will be have the same conditional distributions with a different p such that in the null space. (Point mass). Becomes isomorphic to testing $Bin(n, p)$ $p \geqslant p_0$.

Weird: we have a MP but not UMP for testing $H_0 : \sigma \geqslant \sigma_0$ where both μ, σ unknown– place point mass LFP. By contrast, testing $\sigma \leqslant \sigma_0$ we can't use a point mass- need a normal prior. Leads to UMP test. $\phi = \mathbf{1}[\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2} > c_{n-1,1-\alpha}]$.

10.6 P-values

Definition 61 (p-value)
Given some nested rejection regions for valid test, then

$$\hat{p}(X) = \inf\{\alpha : X \in R_\alpha\}.$$

Fact 62 (P-values are subuniform)
P values are subuniform. If we have exact level α for any α, then $\hat{p}$ is exactly uniform.

11 Stochastic Convergence

Helpful tools

1. Don't forget Portmanteau
2. CLT, SLLN
3. CMT
4. Slutsky

Theorem 63 (CMT)
If g is continuous on a set of probability 1

$$X_n \longrightarrow^* X \implies g(X_n) \longrightarrow^* g(X)$$

for conv in dist, prob, or as

Theorem 64 (Slutsky)
A few parts

1. $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$
2. If $\|X_n - Y_n\| \xrightarrow{p} 0$ then if $X_n \xrightarrow{d} X$ we have also $Y_n \xrightarrow{d} X$.
3. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ then
$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

Corrolary via Slutsky and CMT: $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ gives us

1. $Y_n X_n \xrightarrow{d} cX$
2. $Y_n + X_n \xrightarrow{d} c + X$
3. $Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$

Definition 65 (Uniform Tightness)
A collection $\{X_\alpha\}_{\alpha \in A}$ is uniformly tight if for all $\epsilon > 0$ there exists an $M < \infty$ such that

$$P(\|X_\alpha\| > M) \leqslant \epsilon \quad \text{for all } \alpha \in A$$

Example 66 (Markov, Tightness)
If all X_α have the same ℓ-th moment, then just use Markov to prove tightness.

In general if a collection has increasing means or something, it won't be tight, eg $X_n \sim N(n, 1)$.

Theorem 67 (Prohorov)
Uniformly tight $\implies$ for all sequences there is a a subsequence that converges in distribution to a random variable.

Conversely, Convergence in distribution $\implies$ unifromly tight.

11.1 Big-O, Little-o

Non-stochastic versions,

$$f(x) = O(g(x)) \iff \limsup_{\varepsilon \longrightarrow 0} \frac{f(\epsilon)}{g(\varepsilon)} < \infty$$

.

$$f(x) = o(g(x)) \iff \lim_{\varepsilon \longrightarrow 0} \frac{f(\varepsilon)}{g(\varepsilon)} = 0$$

Definition 68 (Little-o_p)

$$X_n = o_p(R_n) \iff \exists Y_n \text{ such that} X_n = R_n Y_n \text{ with } Y_n \xrightarrow{p} 0$$

Definition 69 (Big O_p)

$$X_n = O_p(R_n) \iff X_n = R_n Y_n \text{ where } Y_n \text{ uniformly tight}$$

Theorem 70 (Delta Method)
Let $r_n \longrightarrow \infty$ and f differentiable at θ. If $r_n(T_n - \theta) \xrightarrow{d} A$ then:

1. $$r_n(f(T_n) - f(\theta)) \xrightarrow{d} f'_\theta A$$

2. $$r_n(f(t_n) - f(\theta)) - f'_\theta(r_n(T_n - \theta)) \xrightarrow{p} 0$$

Proof. Main idea:
$$f(\theta + h) - f(\theta) = f'_\theta h + o(\|h\|) \text{ as } h \longrightarrow 0$$

Take $h = T_n - \theta$ □

Theorem 71 (Higher order delta method)

12 MLE

Definition 72 (M-estimators)

$$\hat{\theta}_n = \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i)$$

where m_θ is some known function, eg ℓ_θ– maximize log likelihood

Definition 73 (Z-estimator)

$$\hat{\theta}_n = \{\theta : n^{-1} \sum_{i=1}^{n} \Psi_\theta(X_i) = 0\}$$

eg $\nabla \ell_\theta = 0$.

12.1 Consistency of MLE

Consistency of $P_n \ell_\theta \xrightarrow{p} P\ell_\theta$ (WLLN) is **not** enough for consistency of MLE $\hat{\theta}_n \xrightarrow{p} \theta^*$. For arbitrary sample size, $P_n \ell_\theta$ not necessarily maximized at (or near) θ^* if we converge too non-uniformly. See picture. For consistency of MLE, need:

1. Uniform convergence (in probability)
2. Well-separation

Definition 74 (Uniform convergence)
$M_n(\theta)$ converges uniformly to $M(\theta)$ if

$$\sup_\theta |M_n(\theta - M(\theta)| \xrightarrow{p} 0$$

Definition 75 (Well-separation)
definition........
eg, strong convexity.

Some more primitive conditions. Identifiability and a finite sample space $|\Theta| < \infty$ are enough for consistency of MLE.

Definition 76 (Identifiability)
Identifiable if for $\theta \neq \theta'$, $P_\theta \neq P'_\theta$ ie KL divergence is strictly positive.

Example 77 (Example of non-identifiability)
In my notes

12.2 Asymptotic normality of the MLE

Under a regularity condition ("smooth, nice"), the MLE is normal.

Definition 78 (Smooth/Nice at θ)
See notes..

1. Hessian of log likelihood is Lipschitz near θ^* - see notes
2. Bounded gradient $P_\theta^* \|\nabla \ell_\theta^*\|^2 < \infty$

Theorem 79 (Asymptotic normality of MLE)
If

1. $\{P_\theta\}_{\theta \in \Theta}$ is smooth/nice at θ^*
2. Θ open subset of $\mathbb{R}^d$,
3. Hessian has finite mean (or alt that exchange order of differentiation wrt θ and expectation).
4. $\hat{\theta}_n$ the MLE is consistent

Then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma_\theta^*)$. **If we can also exchange order of differentiation and expectation** , $\Sigma_\theta^* = I_\theta^{-1}$.

13 Fisher Information

Definition 80 (Fisher Information)
Outer product of the score

$$I_\theta = \mathbf{E}_{P_\theta}[\nabla_\theta \ell_\theta (\nabla_\theta \ell_\theta)^T]$$

If we can exchange order of differentiation and expectation then:

$$I_\theta = \mathbf{Cov}(\nabla \ell_\theta) = -\mathbf{E}[\nabla^2 \ell_\theta]$$

Lots of information, small variance. Look at hessian to look at curvature– if it's really peaked, we have lots of information.

Theorem 81 (Cramer-Rao)

Definition 82 (Asymptotic efficiency)

Distribution	**Parameter(s)**	**Fisher information** $I(\theta)$
Bernoulli $\mathrm{Bern}(p)$	$p \in (0,1)$	$\frac{1}{p(1-p)}$
Binomial $\mathrm{Bin}(m,p)$ (fixed m)	$p \in (0,1)$	$\frac{m}{p(1-p)}$
Poisson $\mathrm{Pois}(\lambda)$	$\lambda > 0$	$\frac{1}{\lambda}$
Exponential $\mathrm{Exp}(\lambda)$	$\lambda > 0$	$\frac{1}{\lambda^2}$
Gamma $\mathrm{Gamma}(\alpha,\theta)$ (fixed α)	$\theta > 0$	$\frac{\alpha}{\theta^2}$
Normal $\mathcal{N}(\mu,\sigma^2)$ (known σ^2)	$\mu \in \mathbb{R}$	$\frac{1}{\sigma^2}$
Normal $\mathcal{N}(\mu,\sigma^2)$ (known μ)	$\sigma^2 > 0$	$\frac{1}{2\sigma^4}$
Normal $\mathcal{N}(\mu,\sigma^2)$ (both unknown)		$\begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$

Table 1: Perobservation Fisher information for common parametric families. Multiply by n for a sample of size n.