ORF524 Midterm Notes

Sunay Joshi

July 3, 2026

Three tricks that will take you far: (1) Add and subtract zero, (2) Plug-in principle, (3) Markov inequality.

The angel in statistics: the normal distribution.

The devil in statistics: the uniform distribution. (Its support depends on the parameter!)

The overarching tension: efficiency (concentration) vs robustness (what if parametrized family is wrong?)

1 Basics and Likelihood theory

Plug-in principle: to estimate a parameter $\beta = \psi(P_\theta)$, use $\hat{\beta} = \psi(\hat{P}_\theta)$, where $\hat{P}_\theta$ is the empirical distribution.

Res: $E[Y|X] = \text{argmin } E[(Y - g(X))^2]$.

(I.e. $E[Y|X]$ minimizes L^2 distance to Y among all X-mb square-integrable rvs. Of course, because it is the L^2-projection of Y onto this space. As a corollary, $Y - E[Y|X]$ is orthogonal to all X-mb square-integrable rvs.)

$\hat{\theta}_{ML} = \text{argsup } \ell_n(\theta)$. (Example of an M-estimator.)

Res (information ineq): $E_{\theta_0}[\ell(\theta)] \leqslant E_{\theta_0}[\ell(\theta_0)]$.

Score $\ell_{\theta,n}(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta)$.

Res (score identity): $E_\theta[\ell_\theta(\theta)] = 0$ (FOC).

Hessian $\ell_{\theta\theta,n}(\theta) = \frac{\partial}{\partial\theta\partial\theta'}\ell_n(\theta)$.

Res (info matrix): $-E_\theta[\ell_{\theta\theta}(\theta)] = E_\theta[\ell_\theta(\theta)\ell_\theta(\theta)'] = V_\theta[\ell_\theta(\theta)] =: I(\theta)$. (Second equality uses expectation of score equals zero. SOC: hessian is negative definite.)

CRLB: if T is an estimator, we have the following lower bound on its variance under P_θ:

$$V_\theta[T(X)] \geqslant (\frac{\partial}{\partial\theta}E_\theta[T(X)])' \cdot I(\theta)^{-1} \cdot (\frac{\partial}{\partial\theta}E_\theta[T(X)]) \tag{1}$$

This tells us the maximum "concentration" we can obtain.

Ex: if T is unbiased, then $V_\theta[T(X)] \geqslant I(\theta)^{-1}$, uniformly in T ("the manner in which you manipulate the data").

Location scale family: given a density f, consider the family $\{f_X(x) = \frac{1}{\sigma}f(\frac{x-\mu}{\sigma}) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$.

($f(x) \mapsto f(x-\mu)$ shifts the density to the right by μ. $f(x) \mapsto \frac{1}{\sigma}f(\frac{x}{\sigma})$ stretches the density by σ horizontally and shrinks it by σ vertically to preserve the integral.)

Exponential family in k-dimensions: if $x \in \mathbb{R}^n$ is the sample, $f(x;\theta) = h(x)c(\theta)\exp(w(\theta)'T(x))$, where $T(x) \in \mathbb{R}^k$. Alternatively written as $f(x;\theta) = h(x)\exp(w(\theta)'T(x) - A(\theta))$, where A is the log partition function with nice properties.

Sufficiency: $T(X)$ is sufficient for θ if the distribution of $X|T(X)$ is independent of θ. "You've distilled all there is to learn about θ into T!"

(I.e. the density $f(x|T(X)=t) = \frac{f(X=x)}{f(T(X)=t)}$ should be independent of θ.)

Res (factorization thm): $T(X)$ is sufficient for θ iff the density of the sample can be written as $f_X(x;\theta) = g(T(x),\theta)h(x)$.

(One direction: if the density can be written like that, then given $T(X)=t$, the conditional density equals $f_X(x|T(X)=t,\theta) = \frac{g(t,\theta)h(x)}{\int g(t,\theta)h(y)dy} = \frac{1}{\int h(y)dy}h(x)$, which is independent of θ.)

Important cor: for the k-dimensional exponential family, $T(X) \in \mathbb{R}^k$ is sufficient. (Look at it!)

Ex: the normal distribution has density $\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\mu^2}{2\sigma^2})\exp(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2)$. If (μ,σ^2) are both unknown, then we have a 2-dimensional exponential family with natural sufficient statistic $T(X) = (\sum X_i, \sum X_i^2)$.

Ex: given a sample, the uniform distribution has density $\frac{1}{\theta^n} \cdot \mathbf{1}_{X_{(1)}>0} \cdot \mathbf{1}_{X_{(n)}<\theta}$. By the factorization thm, $X_{(n)}$ is a sufficient statistic.

A sufficient statistic $T(X)$ is minimal sufficient **if $T(X)$ is a measurable function of every sufficient statistic $S(X)$.**

A complete sufficient statistic is a sufficient statistic such that if $E_\theta[g(T(X))] = 0$ for all θ, then $P_\theta(g(T(X)) = 0) = 1$ for all θ.

(A condition for uniqueness. If you modify T via g such that you obtain an unbiased estimator of zero, then g must be a.s. zero.)

Res (Lehmann-Scheffe): unbiased estimators based on complete sufficient statistics are unique.

(Easy: suppose you had two such estimators $\psi_1(T(X))$ and $\psi_2(T(X))$. Since these are both

unbiased, we may take the difference to find $E_\theta[(\psi_1 - \psi_2)(T(X))] = 0$. By completeness, we find $\psi_1 = \psi_2$ a.s.)

When is a sufficient statistic complete? Master theorem:

Res (exponential family): the natural sufficient statistic $T(X) \in \mathbb{R}^k$ is complete if the set $\{w(\theta) : \theta \in \Theta\}$ contains an open set in $\mathbb{R}^k$.

(This doesn't apply to uniform, Laplace, etc. This doesn't apply to the curved exponential family $\{N(\theta, \theta^2) : \theta \in \mathbb{R}\}$, because here $f(x;\theta) = \frac{1}{\sqrt{2\pi}\theta} \exp(-\frac{1}{2}) \exp(\frac{1}{\theta}x - \frac{1}{2\theta^2}x^2)$, so that $w(\theta) = (\frac{1}{\theta}, -\frac{1}{2\theta^2})$ is a parabola that contains no open set in $\mathbb{R}^2$.)

2 Point estimation

In framework of statistical decision theory, we seek an estimator that minimizes risk (MSE).

Bias-variance decomposition: $MSE_\theta(X, a) = V_\theta[X] + (Bias_\theta(X, a))^2$.

Impose a constraint of unbiasedness to rule out silly estimators. We write $\mathcal{T}_u$ for the class of unbiased estimators of θ. (However in real life, this isn't always a natural assumption!)

An estimator is UMVU if it is an unbiased estimator of θ and is minimum variance in this class uniformly in θ.

Res (Rao-Blackwell): given an estimator $\hat{\theta}_n \in \mathcal{T}_u$ of θ, and given a sufficient statistic $T(X)$ for θ, we have:

- $\tilde{\theta}_n := E_\theta[\hat{\theta}_n | T(X)] \in \mathcal{T}_u$
- $V_\theta[\tilde{\theta}_n] \leqslant V_\theta[\hat{\theta}_n]$ for all θ

(In other words, given an unbiased estimator of θ, we can project onto $T(X)$-mb rvs to obtain a new unbiased estimator with lower variance! Only estimators based on sufficient statistics can be UMVU.)

(Proof is easy: since $T(X)$ is sufficient, $\tilde{\theta}_n$ does not depend on θ (?). By tower property, $\tilde{\theta}_n$ is unbiased. By basic properties of projection, the L^2 norm of $\tilde{\theta}_n$ is at most the norm of $\hat{\theta}_n$. And the norm squared is the variance by unbiasedness.)

(A formal way of showing that the variance decreases: the law of total variance. $V[Y] = E[V[Y|X]] + V[E[Y|X]]$.)

With Rao-Blackwell and Lehmann-Scheffe, the key to find the UMVU is to find a complete sufficient statistic $T(X)$ and then find an unbiased estimator based on $T(X)$.

Ex: recall that for $Unif(0, \theta)$, $X_{(n)}$ is complete sufficient for θ. Its expectation however is $\frac{n}{n+1}\theta$. So $\hat{\theta}_{UMVU} = \frac{n+1}{n}X_{(n)}$.)

You might expect the variance of the UMVU to match the CRLB. Unfortunately the CRLB was derived assuming regularity conditions. It fails for instance for the devil, $Unif(0,\theta)$! Add...)

Aside from UMVU estimation, there are two ad-hoc methods for estimation.

Method of moments: given g, $\hat{\theta}_{MM}$ is chosen such that $E_\theta[g(X)] = \frac{1}{n}\sum_{i=1}^n g(X_i)$.

Ex: suppose we match the first moment for $Unif(0,\theta)$. Then $\frac{\theta}{2} = \bar{X}$, so the MME is $2\bar{X}$. (Note this is not sufficient! We can reduce its variance by conditioning on $X_{(n)}$, and this gives the UMVU.)

3 Hypothesis testing

We ask a less ambitious question: is the true parameter in the null Θ_0 or the alternative Θ_1?

Our tests will be of the form "reject the null" iff $T(X) > c$, where T is some statistic.

Neyman approach: as with the theory of optimal point estimation, we have constraints and objectives. We maximize power subject to a bound on Type I error. (This is not the only way.)

Subject to $\sup_{\theta\in\Theta_0} P_\theta(\text{Type I}) \leqslant \alpha$ (level α), we maximize the power $\beta(\theta) = P_\theta(\text{reject } H_0)$ for $\theta \in \Theta_1$. We seek a UMP test among the class of all level α tests.

Plotting the power $\beta(\theta)$ is helpful. It should be small and $\leqslant \alpha$ for $\theta \in \Theta_0$, and as close to 1 for $\theta \in \Theta_1$. Size of the test is the maximum value in the null region.

Res (Neyman-Pearson lem): consider "simple vs simple" testing problem $H_0 : \theta = \theta_0$ cs $H_1 : \theta = \theta_1$. Then for all α, there exists a (randomized) test of *size* α of the form

$$T_*(X) = \mathbf{1}_{p_1(x)>cp_0(x)} + \gamma \cdot \mathbf{1}_{p_1(x)=cp_0(x)} \tag{2}$$

and such a test is UMP of level α.

More informally, the optimal test statistic is the likelihood ratio $T(X) = \frac{f_X(x;\theta_1)}{f_X(x;\theta_0)}$, and we choose c s.t. $P_{\theta_0}(T(X) > c) = \alpha$. (The complicated form above handles when the distribution of the likelihood ratio is not continuous, for which we need *randomized* tests!)

We use a size α test because if we don't, then intuitively we are "wasting" power. We could be rejecting more and increasing our power while staying within the level α constraint.

If there exists a sufficient statistic $S(X)$, then by the factorization thm, the likelihood ratio will be a function of $S(X)$.

We can also handle "composite vs composite" tests. A family of densities satisfies the monotone likelihood ratio (MLR) in the statistic $T(x)$ if $\frac{f(x;\theta_1)}{f(x;\theta_0)}$ is an increasing function of $T(x)$.

Res (Karlin-Rubin): consider the testing problem $H_0 : \theta \leqslant \theta_0$ vs $H_1 : \theta > \theta_0$. If $T(X)$ is sufficient for θ with density satisfying the MLR property, then the likelihood ratio test given above is UMP at level α. (?)

4 Confidence intervals

A hypothesis test defines a critical region of samples X for which we reject the null, i.e. for which Θ_0 is inconsistent with the data (at level α).

Confidence intervals go the other way. Given a sample X, we define a confidence set of parameter values θ that are consistent with X at level α.

An $(1-\alpha)\%$ confidence interval $I(X)$ is a *random interval* such that the coverage probability is at least $1-\alpha$: $P_\theta(\theta \in I(X)) \geqslant 1-\alpha$.

(Here since we are using the law P_θ, the "true" parameter is θ.)

Interpretation: if many (random) intervals are constructed, then $(1-\alpha)\%$ of them will contain the true θ.

Duality... add

5 Asymptotics basics

When I write "the a-tail of Y" I mean the quantity $P(|Y| \geqslant a)$.

Convergence in probability: for any ϵ, the ϵ-tails of $X_n - X_\infty$ go to zero.

Bounded in probability: for any ϵ, there exists M such that the M-tails are at most ϵ past a certain point.

Convergence in distribution: the cdfs converge. (If $X_n \xrightarrow{d} c$, then we also have convergence in probability.)

Continuous mapping thm: continuous functions preserve convergence in probability.

What if we want to combine different types of convergence?

Slutsky's thm: if $X_n \xrightarrow{d} X_\infty$ and $Y_n \xrightarrow{p} c$, then $X_n Y_n \xrightarrow{d} X_\infty c$, etc.

Stochastic orders: o_p, O_p.

$X_n = o_p(Y_n)$ if $X_n/Y_n \xrightarrow{p} 0$. $X_n = O_p(Y_n)$ if X_n/Y_n is bounded in probability.

If $r_n \to 0$ and $X_n = O_p(1)$, then $r_n X_n = o_p(1)$. Add properties...

LLN: sample mean converges to EX_1 under various assumptions. (Second moment method.)

CLT: sample mean is asymptotically normal upon centering and scaling.

Delta method: if $r_n(X_n - \theta) \xrightarrow{d} X_\infty$, then $r_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot X_\infty$.

(Proof: Taylor expand the LHS and bound insignificant terms... add)

$\hat{\theta}_n$ is consistent for θ if $\hat{\theta}_n \xrightarrow{p} \theta$. (It is r_n-consistent if $r_n(\hat{\theta}_n - \theta) = O_p(1)$.)

$\hat{\theta}_{1,n}$, $\hat{\theta}_{2,n}$ are asymptotically equivalent if $\sqrt{n}(\hat{\theta}_{1,n} - \hat{\theta}_{2,n}) = o_p(1)$.

Asymptotic bias: if $r_n(\hat{\theta}_n - \theta) \xrightarrow{d} Y$, then $AsyBias(\theta) = E[Y]$.

Asymptotic variance: if $r_n(\hat{\theta}_n - \theta) \xrightarrow{d} Y$, then $AsyVar(\theta) = \frac{V[Y]}{r_n^2}$.

Asymptotic MSE: if $r_n(\hat{\theta}_n - \theta) \xrightarrow{d} Y$, then $AsyMSE(\theta) = \frac{E[Y^2]}{r_n^2}$.

Asymptotic relative efficiency: given $\hat{\theta}_{1,n}$ and $\hat{\theta}_{2,n}$, we define $e_{12}(\theta) = \frac{AsyMSE_1(\theta)}{AsyMSE_2(\theta)}$.

Asymptotic inference: for hypothesis testing, take limsup's; for confidence intervals, take liminf's.

(Reflection: asymptotic normality (say $\sqrt{n}(X_n - \theta) \xrightarrow{d} N(0,1)$) gives you more than consistency. By blowing up by $\sqrt{n}$, asymptotic normality tells you about the fluctuations at scale $\frac{1}{\sqrt{n}}$. Of course, it doesn't say anything about finer fluctuations, say at scale $\frac{1}{n}$.)

6 M and Z estimation

M-estimation provides a framework that generalizes MLE.

Given a sample X and a loss function ρ, we choose the parameter b such that

$$\hat{\beta}_n = \text{argmin } \frac{1}{n} \sum_{i=1}^{n} \rho(X_i; b) \tag{3}$$

(The argument is the plug-in estimator of $E_{\theta_0}[\rho(X_i; b)]$.) We would like $\sqrt{n}$-consistency and asymptotic normality of such a $\hat{\beta}$.

(Ex: if $\rho(x; b) = -\log f(X_i; b)$ we recover MLE. Also, $\rho(x; b) = (x - b)^2$ gives $E_{\theta_0}[X_1]$. Also, MME... add)

Consistency: does $\hat{\beta}_n \xrightarrow{p} \beta_0$? This is asking if the minima of a sequence of random objective functions converges to the minimum of the limiting $E_{\theta_0}[\rho(X_1; b)]$ objective function. If we could swap *plim* and argmin, then this would hold.

Res (consistency): let $\hat{Q}_n(b)$ be the empirical objective function, and let $Q(b)$ be the objective function. Suppose that the empirical objective functions converge uniformly in probability to $Q(b)$: $\sup_{b \in B} |Q_n(b) - Q(b)| \xrightarrow{p} 0$. Then $\hat{\beta}_n \xrightarrow{p} \beta_0$.

(Pf: "uniform LLN"... uniform consistency... add...)

The following is the master theorem for asymptotic normality.

Res (asymptotic normality of M-estimators): under numerous technical assumptions including uniform consistency and sufficient differentiability, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, H_0^{-1}\Sigma_0 H_0^{-1}) \tag{4}$$

Here $\Sigma_0 = V[\frac{\partial}{\partial\beta}\rho(X_i;\beta_0)]$ and $H_0 = E[\frac{\partial^2}{\partial\beta\partial\beta'}\rho(X_i;\beta_0)]$. We call $V_0 := H_0^{-1}\Sigma_0 H_0^{-1}$ the "sandwich matrix".

(Pf: Taylor expansion... In the end, we find a sample average lurking: $\sqrt{n}(\hat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i;\beta_0) + o_p(1)$, where $\psi(x;\beta) = -H_0^{-1}\frac{\partial}{\partial\beta}\rho(x;\beta)$ is our "influence function". By uniform consistency, the RHS consists of mean zero rvs, so CLT implies that the average tends to $N(0,V)$, where $V = V[-H_0^{-1}\frac{\partial\rho}{\partial\beta}]$. But this is exactly $H_0^{-1}\Sigma_0 H_0^{-1}$.)

(In hindsight, with only the CLT at our disposal, how else could we obtain AN without reducing to an asymptotic linear representation?)

Z-estimation considers the FOC of an M-estimation problem. Specifically, assuming differentiability, the Z-estimator is the solution to

$$\hat{\beta}_n = \text{argzero } \frac{1}{n}\sum_{i=1}^n \psi(X_i; b) \tag{5}$$

Here we write $\psi := \frac{\partial}{\partial b}\rho$. (The argument is the plug-in estimator of $\frac{\partial}{\partial b}E_{\theta_0}[\rho(X_1;b)]$.)

We can now deploy the theory of M-estimators to familiar estimators like MLE and MME.

Ex: MLE uses $\rho(X_i;\theta) = -\log f(X_i;\theta)$. Note that $H_0 = -I_0$ and $\Sigma_0 = I_0$, where I_0 is the information matrix. So $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically $N(0, I_0^{-1})$. And we recognize I_0^{-1} as the CLRB! Therefore the MLE is asymptotically efficient; it attains minimum variance as $n \to \infty$.

Add MME...

We now consider asymptotic inference. A common issue is that the asymptotic variance is unknown and must be estimated via a standard error. Consider the testing problem $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$.

Wald test ("parameter estimate"): since $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, V_0)$, we may left-multiply by $V_0^{-1/2}$ and take $\|\cdot\|_2^2$ to find $T_{Wald} = n\|V_0^{-1/2}(\hat{\beta}_n - \beta_0)\|^2 \xrightarrow{d} \chi^2_{dim(B)}$. Since the RHS is pivotal, we can do testing with the LHS as the statistic.

LM (Lagrange multiplier) ("FOC"): consider the FOC of M-estimation. Then by CLT $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\beta}\rho(X_i;\beta_0) \xrightarrow{d} N(0,\Sigma_0)$. Left-multiplying by $\Sigma_0^{-1/2}$ and taking $\|\cdot\|_2^2$, we find $T_{LM} = n\|\Sigma_0^{-1/2}\cdot\frac{1}{n}\sum\frac{\partial}{\partial\beta}\rho(X_i;\beta_0)\|^2 \xrightarrow{d} \chi^2_{dim(B)}$.

LR ("objective function"): this is generalized to the "distance difference test". We use the statistic $T_{DD} = n(\frac{1}{n}\sum_{i=1}^n \rho(X_i;\hat{\beta}_n) - \frac{1}{n}\sum_{i=1}^n \rho(X_i;\beta_0)) \xrightarrow{d} \chi^2_{dim(B)}$.

Why do we have this convergence? add...

Nuisance parameters? add...

We've seen LR before; but what do Wald and LM mean more generally?... add

Add specifics for MLE in terms of general composite testing, careful with argmin versus argmax...

Add precise definitions of each test statistic according to slides (including weird denominator of Wald)...

7 Linear regression

Add...

8 How to calculate

If you transform a rv X to obtain $Y := \phi(X)$, the density transforms as

$f_Y(y) = |\det \phi'|^{-1} \cdot f_X(\phi^{-1}(y))$

$\int f_Y(y)dy = \int |\det \phi'|^{-1} f_X(x) \cdot |\det \phi'|dx = \int f_X(x)dx$

A sanity check: suppose we scale the random variable X by k. Then $f_Y(y) = \frac{1}{k} f_X(\frac{y}{k})$.

Calculating with multivariate gaussians... independence iff covariance zero... projection...

Multivariate derivatives in likelihood computations, multivariate chain rule...

Common distributions...

9 Practice problems

F22 midterm:

Q1:

(1) $m(S(X))$ sufficient implies $S(X)$ sufficient.

First thoughts: we know that $f(x|m(S(X)) = t, \theta)$ is independent of θ...

Actually, factorization thm: $f(x, \theta) = h(x)g(m(S(x)), \theta)$

The second is certainly a function of $S(x), \theta$! Formally, let $\tilde{g}(S(x), \theta) = g(m(S(x)), \theta)$.

Then $f(x, \theta) = h(x)\tilde{g}(S(x), \theta)$. Hence $S(X)$ is sufficient.

(If we tried the converse: $f(x, \theta) = h(x)g(S(x), \theta)$. If m is not injective, then $g(S(x), \theta)$ is not necessarily a function of $m(S(x))$. Think about writing x as a function of x^2!)

(2) $S(X)$ sufficient and m^{-1} exists and is measurable implies that $m(S(X))$ is sufficient.

Obvious from the above: $S(x) = m^{-1}(m(S(x))$, which is a measurable function of $m(S(x))$.

(3) Add...

(4) If $S(X)$ complete, then $m(S(X))$ complete.

Go back to the definition: if g is a measurable function such that $E[g(m(S(X))] = 0$, then writing $h = g \circ m$, we have $E[h(S(X))] = 0$. By completeness of $S(X)$, this implies $h = 0$ a.s.

(5) (From precept) Suppose S is complete sufficient and bounded. Suppose T is minimal sufficient. Then there exists measurable ψ such that $T = \psi(S)$. Then there exists measurable g such that $g(S) = S - E[S|T]$. (Integrable because bounded. The conditional expectation is a measurable function of S, hence of T.) But $E[g(S)] = 0$ by tower property. Since S is complete, $g = 0$ a.s. Hence $E[S|T] = S$ a.s. The LHS is a measurable function of T, so that $S = \phi(T)$. Since T is minimal sufficient, S is also minimal sufficient.

Q2:

(1) For $x \leqslant 0$, clearly $F(x) = 0$. For $x \geqslant \theta^2$, clearly $F(x) = 1$. For x in between,

$$F(x) = P(X_i \leqslant x) \tag{6}$$

$$= \int_0^x \frac{2x}{\theta^4} dx \tag{7}$$

$$= \frac{x^2}{\theta^4} \tag{8}$$

(2) (a) Let's compute $E[\sqrt{X_i}]$; it seems relevant.

$$E[\sqrt{X_i}] = \int_0^{\theta^2} \sqrt{x} \cdot \frac{2x}{\theta^4} dx \tag{9}$$

$$= \frac{2x^{5/2}}{5/2} \frac{1}{\theta^4}\Big|_{\theta^2} \tag{10}$$

$$= \frac{4}{5}\theta \tag{11}$$

Therefore $E[\hat{\theta}] = \frac{4}{5}\theta$, and $\hat{\theta}_U = \frac{5}{4}\hat{\theta}$ is unbiased.

Method of moments? Match $\hat{E}[\sqrt{X_i}]$ and $E[\sqrt{X_i}]$.

(b)

$$L(\theta) = \prod_i \frac{2X_i}{\theta^4} \cdot \mathbf{1}_{X_{(n)} \leqslant \theta^2} \tag{12}$$

$$\ell(\theta) = -4n \log \theta + \sum_i \log(2X_i) + \log \mathbf{1}_{X_{(n)} \leqslant \theta^2} \tag{13}$$

If $X_{(n)} > \theta^2$, the log-likelihood is $-\infty$. We won't find a maximum there!

To the right of $X_{(n)}$, the likelihood is decreasing, since it's $-4n \log \theta + \text{const}$. Therefore $\hat{\theta}_{ML} = \sqrt{X_{(n)}}$. Boris is right; we need not fear non-differentiable likelihood functions!

How do you show complete sufficient? Sufficiency follows from the factorization theorem; θ only appears with $X_{(n)}$ in the density $L(\theta) \equiv f(X; \theta)$. Completeness? Not sure, since it's not an exponential family... hard, add

(c) We have a complete sufficient statistic $\hat{\theta}_{ML}$, so we need only manipulate it to obtain an unbiased one to recover the UMVU. Let's compute its expectation.

$$P(\sqrt{X_{(n)}} \leqslant x) = \prod_i P(X_i \leqslant x^2) \tag{14}$$

$$= \prod_i \frac{x^4}{\theta^4} \tag{15}$$

$$= (x/\theta)^{4n} \tag{16}$$

Differentiate to find the density: $f(x; \theta) = \frac{4nx^{4n-1}}{\theta^{4n}}$. Ah; I keep forgetting the integrand besides the measure in expectations!

$$E[\sqrt{X_{(n)}}] = \int_0^{\theta} x \cdot \frac{4nx^{4n-1}}{\theta^{4n}} dx \tag{17}$$

$$= \frac{4n}{4n+1}\theta \tag{18}$$

Therefore $\hat{\theta}_{UMVU} = \frac{4n+1}{4n}\hat{\theta}_{ML}$.

(d) Calculating $V[\hat{\theta}_U]$: easy by independence!

$$V[\hat{\theta}_U] = \frac{25}{16}\frac{1}{n}V[\sqrt{X_i}] \tag{19}$$

What's that?

$$V[\sqrt{X_i}] = E[X_i] - E[\sqrt{X_i}]^2 \tag{20}$$

$$= \int_0^{\theta^2} x \cdot \frac{2x}{\theta^4}dx - (\frac{4}{5}\theta)^2 \tag{21}$$

$$= \frac{2\theta^2}{3} - \frac{16\theta^2}{25} \tag{22}$$

$$= \frac{2}{75}\theta^2 \tag{23}$$

So $V[\hat{\theta}_U] = \frac{1}{24n}\theta^2$. (**Don't forget what you're solving for!**)

Calculating $V[\hat{\theta}_{UMVU}]$:

Clearly $V[\hat{\theta}_{UMVU}] = (\frac{4n+1}{4n})^2 V[\hat{\theta}_{ML}]$. Need $V[\hat{\theta}_{ML}] = E[X_{(n)}] - E[\hat{\theta}_{ML}]^2$. Know the second term is $(\frac{4n}{4n+1})^2\theta^2$. What about the first term?

$P(X_{(n)} \leqslant x) = x^{2n}/\theta^{4n}$. So $f(x;\theta) = 2nx^{2n-1}/\theta^{4n}$. Hence $E[X_{(n)}] = \int_0^{\theta^2} 2nx^{2n}/\theta^{4n}dx = \frac{2n}{2n+1}\theta^2$. Therefore the above is $((\frac{4n+1}{4n})^2\frac{2n}{2n+1} - 1)\theta^2$. Simplify: $\frac{1}{16n^2+8n}\theta^2$.

Therefore the UMVU is strictly less for $n \geqslant 2$.

(3) Ah, a one-sided test to the left. We use a test statistic of $\hat{\theta}_{ML}/\theta_0$.

(a) We want $\sup_{\theta \leqslant \theta_0} P_\theta[\hat{\theta}_{ML} > c\theta_0] = \alpha$, where $\alpha = 0.05$. (Equality, since size!)

Luckily, the expression is the one-minus-cdf: $1 - F(c\theta_0) = 1 - (\frac{c\theta_0}{\theta})^{4n}$. This is clearly maximized at $\theta = \theta_0$, at which we have power $1 - c^{4n}$. Setting this equal to α, we need $c_\alpha = (1-\alpha)^{1/4n}$.

(b) We need to determine an expression for the power function. Intuitively, it will be close to zero for $\theta < \theta_0$, α at $\theta = \theta_0$, and close to 1 for $\theta > \theta_0$. From our calculations above, $\beta(\theta) = \max\{0, 1 - (\frac{c\theta_0}{\theta})^{4n}\}$. (The max is because for $c\theta_0 > \theta$, the one-minus-cdf is zero.)

(c) Dual one-sided CI?

Choose all θ_0 such that we don't reject H_0, i.e. such that $\hat{\theta}_{ML} \leqslant c\theta_0$, i.e. $\theta_0 \in [\hat{\theta}_{ML}/c, \infty)$.

F22 final:

Q2: (Optimal conditional Z-estimation) Oh boy, what have I gotten myself into...

(1) Condition on x_i. Want to show: if h is a function such that $E[g(x_i)h(x_i)] = 0$ for all $g \in L^2$, then $h(x_i) = 0$ a.s. (If $h(x_i)$ is nonzero with positive probability, then we could localize g such that the expectation is nonzero... add)

(The point of this problem is to show that θ_0 is the optimizer of the infinite-sample Z-estimation argument – hence Z-estimation has a hope of yielding a consistent estimator!)

(2) How to prove consistency of Z-estimator? Using uniform continuity, interchange argzero and plim! Then the result follows from (1) and LLN.

(I guess we need to show that uniform continuity implies uniform convergence of the objective function in the θ argument...)

(3) Straight from the master theorem?

Think of $\psi(X_i;\theta) = g(x_i)m(z_i,\theta)$.

Then $\tilde{\psi} = (-E[\frac{\partial\psi}{\partial\theta}])^{-1}\psi$ (which we unfortunately call regular ψ).

So $\tilde{\psi}(X_i;\theta_0) = (-E[g(x_i)\frac{\partial m}{\partial\theta^T}(z_i,\theta_0)])^{-1}\psi(X_i;\theta_0)$ should go in the given equation.

(We **transpose** because the resulting derivative should be a matrix! **Beware of dimensions!**)

(4) Easy to rederive, now that we have (3). The CLT kicks in on the RHS, and with Slutsky, we find that the RHS tends in distribution to $N(0,V_0)$, where V_0 satisfies Huber's sandwich formula! This is:

$$V_0 = \Omega^{-1}\Sigma\Omega^{-1} \tag{24}$$

where $\Omega = E[g(x_i)\frac{\partial m}{\partial\theta^T}(z_i,\theta_0)]$ and $\Sigma = V_{\theta_0}[\psi(X_i;\theta_0)] = V[g(x_i)m(z_i,\theta)]$. (The covariance matrix of AZ is simply $AV[Z]A^T$, which is what we used above.)

(5) Propose an estimator $\hat{V}_n$ of $V_0(\theta_0)$. If I don't know something, I should put a hat on it! We use $\hat{V}_n = \hat{\Omega}_n^{-1}\hat{\Sigma}_n\hat{\Omega}_n^{-1}$, where

$$\hat{\Omega}_n = \frac{1}{n}\sum_i g(x_i)\frac{\partial m}{\partial\theta^T}(z_i,\hat{\theta}_n) \tag{25}$$

$$\hat{\Sigma}_n = \frac{1}{n}\sum_i (g(x_i)m(z_i,\hat{\theta}_n))(g(x_i)m(z_i,\hat{\theta}_n))^T \tag{26}$$

(Notice we have plugged in the Z-estimate $\hat{\theta}_n$, as we do not know θ_0!)

Consistency follows by expanding $\partial m/\partial\theta^T$ or m around θ_0 to first order, and invoking LLN or asymptotic normality to control the various terms... add

(6) At last, time for asymptotic inference.

Delta method:

$$\sqrt{n}(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} h'(\theta_0)\cdot N(0,V_0) = N(0,h'(\theta_0)V_0h(\theta_0)^T) \tag{27}$$

So

$$\sqrt{n}\cdot(h'(\theta_0)V_0h'(\theta_0)^T)^{-1/2}\cdot(h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} N(0,I) \tag{28}$$

By consistency, we may swap out the denominator with estimators. As Boris said, we do this to **make the statistic a quadratic form in** θ_0:

$$\sqrt{n} \cdot (h'(\hat{\theta}_n)\hat{V}_n h'(\hat{\theta}_n)^T)^{-1/2} \cdot (h(\hat{\theta}_n) - h(\theta_0)) \xrightarrow{d} N(0, I) \tag{29}$$

Now we take the norm squared:

$$T_W(X) = \|\sqrt{n} \cdot (h'(\hat{\theta}_n)\hat{V}_n h'(\hat{\theta}_n)^T)^{-1/2} \cdot (h(\hat{\theta}_n) - h(\theta_0))\|^2 \tag{30}$$

Since this goes to χ^2_1 in distribution, our two-sided test rejects if it lands outside $[x_{1-\frac{\alpha}{2}}, x_{\frac{\alpha}{2}}]$.

(7) Two-sided CI for $\beta_0 = c^T\theta_0$? Again, Delta method.

$$\sqrt{n}(c^T\hat{\theta}_n - c^T\theta_0) \xrightarrow{d} N(0, c^T V_0 c) \tag{31}$$

We may use a Wald test statistic and dualize to obtain a two-sided CI as follows.

$$T_W(X) = (\sqrt{n} \cdot (c^T\hat{V}_n c)^{-1/2} \cdot (c^T\hat{\theta}_n - c^T\theta_0)^2)^2 \tag{32}$$

Our two-sided CI consists of all θ_0 for which $T_W(X) \in [x_{1-\frac{\alpha}{2}}, x_{\frac{\alpha}{2}}]$. This can be solved explicitly; it's "centered" around $c^T\hat{\theta}_n$ with widths $(c^T\hat{V}_n c)^{1/2}\sqrt{\frac{x_\beta}{n}}$... add, robust?, check signs?

(8) Add... expand?

(9) Should be easy with (8) in place; (8) tells us that $V(g) - V(g^*)$ is PSD, hence $V(g) \succcurlyeq V(g^*)$ in our partial order.

Estimator of θ_0? $\hat{\theta}_n(\hat{g}^*_n)$? add...

(10) Normal linear model... compute above, add

F21 midterm:

Scan a given question to get a sense of the direction.

Q1: (Probability theory)

Simple Chernoff-like inequalities via sub-gaussian property... add...

Q2:

(1) Obvious; how to prove formally? add...

Attempt: if F is the law of V, then $p_A = \int_A dF(v)$. Since F does not depend on θ, this doesn't depend on θ.

Note that the conditional distribution $X|T$ doesn't depend on θ, so the distribution of $V|T$ doesn't depend on θ, and the argument above applies.

(2) The first is just tower property. How to show second?

Ah, **think about what information hasn't been used.** We need to use the fact that T is *complete* sufficient. Rewrite to find $E_\theta[g(T)] = 0$, where $g(T) = q_A(T) - p_A$ is a measurable function of T. By completeness, we need $g(T) = 0$ a.s., hence $p_A = q_A(T)$ a.s.

(3) Show that T, V are independent. (This is somewhat intuitive; the distribution of T captures the dependence on θ, and the distribution of V doesn't depend on θ at all, so they should be independent!)

Note $p_A = q_A(T) = P(V \in A|T)$ a.s. How to get $T \in B$ here?

$P(T \in B|V \in A) \stackrel{?}{=} P(T \in B)$

$q_A(t) = P(V \in A|T = t) = \frac{P(T=t|V\in A)P(V\in A)}{P(T=t)}$

$P(T = t|V \in A) = q_A(t)\frac{P(T=t)}{P(V\in A)} = p_A\frac{P(T=t)}{P(V\in A)} = P(T = t)$

$P(T \in B|V \in A) = P(T \in B)$

Check logic...

(4)

(a) Need to find distribution of $\bar{X}(w)$. Mean $\frac{\mu}{n}\sum_i w_i$. Variance $\frac{\sigma^2}{n^2}\sum_i w_i^2$.

Ancillary for μ iff $\sum_i w_i = 0$; then the distribution doesn't depend on μ.

(b) Clearly we want to use the previous parts. Suppose $\sum_i w_i = 0$, so that $\bar{X}(w)$ is ancillary. We know from the factorization theorem that $\bar{X}$ is complete sufficient. Then by (3), $\bar{X}$ and $\bar{X}(w)$ are independent.

Q3:

Consider the cdf $F(x;\theta) = (a + bx^{-1/\theta})\mathbf{1}_{x>2}$.

(1)

(a) $F(\infty) = 1$ implies $a = 1$.

We need F to be right-continuous at $x = 2$. At $x = 2$, $F = 0$. The right-limit is $a + b2^{-1/\theta}$. Hence $b2^{-1/\theta} = -a$, so $b = -2^{1/\theta}$.

(b) The density is the derivative in x:

$f(x;\theta) = b(-\frac{1}{\theta})x^{-\frac{1}{\theta}-1} = \frac{2^{1/\theta}}{\theta}x^{-\frac{\theta+1}{\theta}}\mathbf{1}_{x>2}$.

We can write this in exponential family form: $f(x;\theta) = \mathbf{1}_{x>2} \cdot \frac{2^{1/\theta}}{\theta} \cdot \exp((-\frac{\theta+1}{\theta}) \cdot \log x)$.

(Hence the complete sufficient statistic is $\sum_i \log X_i$.)

Don't forget indicator! Exponential family refers to density, not cdf!

(2)

(a)

$$E[X_i] = \int_2^\infty x \cdot \frac{2^{1/\theta}}{\theta} x^{-\frac{\theta+1}{\theta}} dx \tag{33}$$

$$= \frac{2^{1/\theta}}{\theta} \int_2^\infty x^{-\frac{1}{\theta}} dx \tag{34}$$

$$= \frac{2^{1/\theta}}{\theta} \frac{2^{\frac{\theta-1}{\theta}}}{\frac{1-\theta}{\theta}} \tag{35}$$

$$= \frac{2}{1-\theta} \tag{36}$$

If $\frac{2}{1-\hat{\theta}_{MM}} = \bar{X}$, then $\hat{\theta}_{MM} = \frac{\bar{X}-2}{\bar{X}}$.

Direction of bias... add

(b)

$$P(\log(X_i/2) \leqslant x) = P(X_i \leqslant 2e^x) \tag{37}$$

$$= F(2e^x) \tag{38}$$

$$= (1 - 2^{1/\theta}(2e^x)^{-1/\theta}) \tag{39}$$

$$= 1 - e^{-x/\theta} \tag{40}$$

So $\log(X_i/2) \sim \text{Expon}(\theta)$. Thus $E[\hat{\theta}_{ML}] = \theta$.

(c) We saw above that $T(X) = \sum_i \log X_i$ is a complete sufficient. So $\hat{\theta}_{ML}$ is based on $T(X)$, not $\hat{\theta}_{MM}$.

(d) Let's compute CRLB. Recall this is $I(\theta)^{-1}$, where the Fisher information is given by $I(\theta) = V[\ell_\theta(\theta)]$ (based on sample size n). We must compute!

$L(\theta) = \mathbf{1}_{X_{(1)}>2} \cdot (\frac{2^{1/\theta}}{\theta})^n \cdot \exp((-\frac{\theta+1}{\theta}) \cdot \sum_i \log X_i)$

$\ell(\theta) = \log \mathbf{1}_{X_{(1)}>2} + n \log \frac{2^{1/\theta}}{\theta} - \frac{\theta+1}{\theta} \cdot \sum_i \log X_i$

$= ... + \frac{n}{\theta} \log 2 - n \log \theta - \frac{\theta+1}{\theta} \cdot \sum_i \log X_i$

$\ell_\theta(\theta) = -\frac{n}{\theta^2} \log 2 - \frac{n}{\theta} + \frac{1}{\theta^2} \cdot \sum_i \log X_i$

$= -\frac{n}{\theta} + \frac{1}{\theta^2} \cdot \sum_i \log \frac{X_i}{2}$

(Matches with $\hat{\theta}_{ML}$; but onwards.)

The desired variance is $\frac{n}{\theta^4} \sum V[\log \frac{X_i}{2}]$. And we showed $\log \frac{X_i}{2} \sim \text{Expon}(\theta)$! If only I remembered the moments of the exponential, dang... certainly the mean is θ. I think the variance is θ^2. This implies that $I(\theta) = \frac{n}{\theta^2}$, and $CRLB(\theta) = \frac{\theta^2}{n}$.

An UMVU estimator, by Rao-Blackwell, must be based on a sufficient statistic. This immediately eliminates $\hat{\theta}_{MM}$. We know $\hat{\theta}_{ML}$ is based on a complete sufficient statistic, so since it is unbiased, it is UMVU.

Let's check if it attains CRLB. By the formula for $\hat{\theta}_{ML}$, we have $V[\hat{\theta}_{ML}] = \frac{1}{n}V[\text{Expon}(\theta)] = \frac{\theta^2}{n}$. Thus CRLB is attained!

(3) Time to do inference for this family. One-sided testing to the right.

(a) Find UMP test among level α tests. **Karlin-Rubin** (not Neyman-Pearson!) to the rescue. This will imply that the likelihood ratio test of size α is UMP. But first, we must check that the parametric family satisfies the **MLR property** in a given sufficient statistic.

Sufficient statistic for this 1d exponential family: $T(X) = \sum_i \log \frac{X_i}{2}$. (Why this? It's related to the MLE, and we know its distribution! It's Gamma(n, θ).)

Fix $0 < \theta_1 < \theta_0$. The log of the likelihood ratio is

$$\ell(\theta_1) - \ell(\theta_0) = \text{const} + (\frac{1}{\theta_0} - \frac{1}{\theta_1})T(X) \tag{41}$$

The quantity in parentheses is negative, so we have a **decreasing** LR in the sufficient statistic $T(X)$? (Checked **backwards**?... add)

We can imagine writing down the LR statistic. **We can pick any $\theta_1 \in \Theta_1$ to design our statistic.** Take $\theta_1 < \theta_0$. Inverting the above, we see that we reject the null for **small** values of $T(X)$. This is equivalent to $\hat{\theta}_{ML}/\theta_0 < c$ for some c. Since $\hat{\theta}_{ML}/\theta_0 \sim$ Gamma$(n, 1/n)$, we have a **pivotal** distribution. We reject iff $\hat{\theta}_{ML}/\theta_0 < c_\alpha$, the α-quantile of Gamma$(n, 1/n)$. This is size α, so it is UMP among level α tests.

(We are using **corrected** notation for quantile; how much mass lies below.)

(b) Exact $(1-\alpha)$-CI for θ? The set of all θ_0 for which H_0 is not rejected above. I.e., the set of all θ_0 for which $\hat{\theta}_{ML}/\theta_0 \geqslant c_\alpha$, i.e. $\theta_0 \leqslant \hat{\theta}_{ML}/c_\alpha$. The interval is $(0, \hat{\theta}_{ML}/c_\alpha]$.

(c) Power function.

$\beta_n(\theta) = P_\theta(\hat{\theta}_{ML}/\theta_0 < c_\alpha)$

Now $\hat{\theta}_{ML} \sim$ Gamma$(n, \theta/n)$, so that $\hat{\theta}_{ML}/\theta \sim$ Gamma$(n, 1/n)$.

Multiplying both sides by $\frac{\theta_0}{\theta}$, we find $P(\text{Gamma}(n, 1/n) < c\theta_0/\theta)$, i.e. $F(c\theta_0/\theta)$, where F is the cdf of Gamma$(n, 1/n)$.

Intuitively, for $\theta < \theta_0$, we want close to 1, for $\theta > \theta_0$, we want close to 0, and for $\theta = \theta_0$, we want α. Indeed, $F(c\theta_0/\theta)$ is decreasing in θ.

(d) Limiting power function.

We want $\lim_n F_n(x)$, where $\lambda_n = 1/n$ and $x = \frac{c\theta_0}{\theta}$.

Aha; x **is not a constant!** It is a sequence depending on n: x_n. The quantile changes.

$P(Y_n \leqslant x_n) = P(\sqrt{n}(Y_n - 1) \leqslant \sqrt{n}(x_n - 1)) \to \Phi(\lim_n \sqrt{n}(x_n - 1))$

If $x_n = c_n$, then we find $\sqrt{n}(c_n - 1) \to \alpha$.

Now we set $x_n = \frac{c_n \theta_0}{\theta}$.

If $\theta = \theta_0$, we get limiting power α. If $\theta > \theta_0$, then there is a negative contribution which goes to $-\infty$, hence power is zero. If $\theta < \theta_0$, then there is a positive contribution which goes to ∞, hence power is one.

F21 final:

Q1: U and V statistics, and asymptotics.

(1) $E[\bar{L}_n] = E[L(X_i)] = 0$ clearly.

For $i \neq j$, $E[\bar{Q}_n] = E[Q(X_i, X_j)] = E[X_i - \mu]E[X_j - \mu] = 0$.

Now we need to handle $E[\bar{L}_n \bar{Q}_n]$... all terms of the form $E[L(X_i)Q(X_j, X_k)]$ for i, j, k distinct die. The only remaining ones are $E[L(X_i)Q(X_i, X_j)]$. Aha, but independence prevails: this equals $E[2\mu(X_i - \mu)^2(X_j - \mu)] = 0$.

It follows that $\bar{L}_n, \bar{Q}_n$ are uncorrelated, so that $V[\bar{U}_n] = V[\bar{L}_n] + V[\bar{Q}_n]$.

(The beauty of the Hoeffding decomposition is that it decomposes a U-statistic into two uncorrelated statistics, a linear centered sample average and a quadratic centered sample average.)

(2) $V[\bar{L}_n] = \frac{1}{n} V[L(X_i)] = \frac{1}{n} 4\mu^2 V[X_i] = O(n^{-1})$.

$V[\bar{Q}_n] = \frac{1}{\binom{n}{2}} V[Q(X_i, X_j)] = \frac{1}{\binom{n}{2}} V[(X_i - \mu)(X_j - \mu)]$.

Note in general, $V[XY] = E[X^2Y^2] - E[XY]^2 = E[X^2]E[Y^2] - E[X]^2E[Y]^2$. For iid mean zero, $V[XY] = E[X^2]^2 = V[X]^2$.

So $V[\bar{Q}_n] = \frac{1}{\binom{n}{2}} V[X_i]^2 = O(n^{-2})$.

Now it's time to handle stochastic orders. $n^{1/2}\bar{L}_n$ has variance $O(1)$. We claim it is bounded in probability. By Chebyshev, $P(n^{1/2}\bar{L}_n > m) \leqslant O(1)/m^2$. Given ϵ, there exists m s.t. the RHS is at most ϵ, uniformly in n. Thus $n^{1/2}\bar{L}_n$ is bounded in probability.

An identical argument shows that $n\bar{Q}_n$ is bounded in probability, so that $\bar{Q}_n = O_p(n^{-1})$.

(3) Let us do this nicely.

$\bar{V} = \frac{1}{n^2} \sum X_i^2 + \frac{2}{n^2} \binom{n}{2} \bar{U}$

$= \frac{1}{n^2} \sum X_i^2 + \bar{U} - \frac{1}{n} \bar{U}$

So $\bar{V} - \bar{U} = \frac{1}{n}(\frac{1}{n} \sum X_i^2 - \bar{U})$

By LLN, the first term in parentheses tends to a constant in probability. The term $\bar{U}$ is $O_p(1)$

by the previous part. Since there is a factor of $1/n$ in front of each, the whole sum is $O_p(n^{-1})$.

(I am repeatedly using the fact that if $X_n \xrightarrow{p} c$, then $X_n = O_p(1)$. Does this make sense? Yes; given η and ϵ, there exists N s.t. for all $n \geqslant N$, the tails satisfy $P(|X_n - c| > \eta) \leqslant \epsilon$. Hence $P(|X_n| > c + \eta) < \epsilon$ for all $n \geqslant N$, and so $X_n = O_p(1)$.)

(4) If $\mu \neq 0$, then we want to show that $\bar{U}_n$ is asymptotically normal. The minus μ^2 is suggestive of the Hoeffding decomposition.

$\sqrt{n}(\bar{U} - \mu^2) = \sqrt{n}\bar{L} + \sqrt{n}\bar{Q}$

By CLT the first tends to $N(0, 4\mu^2)$ (we used $\mu \neq 0$). The second tends to 0 in probability by (2). By Slutsky, the sum tends to $N(0, 4\mu^2)$.

(5) Now $\sqrt{n}\bar{L}$ can't be handled immediately with LLN.

Go via V statistic (intuitively why?):

$\frac{n-1}{n}\bar{U} = \bar{V} - \frac{1}{n^2}\sum X_i^2$

$n\bar{U} = \frac{n^2}{n-1}\bar{V} - \frac{1}{n-1}\sum X_i^2$

$= \frac{n^2}{n-1}(\frac{1}{n}\sum X_i)^2 - \frac{n}{n-1}\frac{1}{n}\sum X_i^2$

$= \frac{n}{n-1}(\frac{1}{\sqrt{n}}\sum X_i)^2 - \frac{n}{n-1}\frac{1}{n}\sum X_i^2$

The second tends to 1 in probability by LLN. By CLT and Slutsky, the first tends to χ_1^2. Therefore by Slutsky once more, the entire expression tends to $\chi_1^2 - 1$.

Q2: (Uniform inference at a boundary)

$\hat{\theta} = \max(\bar{X}, 0)$.

(1) $L(\theta) = \prod_i \frac{1}{\sqrt{2\pi}} \exp(-\frac{(X_i-\theta)^2}{2})$

$\ell(\theta) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}\sum(X_i - \theta)^2$

Strictly concave. Recall $\theta \in [0, \infty)$.

$\ell_\theta(\theta) = -\sum(\theta - X_i) = n(\bar{X} - \theta)$

If $\bar{X} \geqslant 0$, then global max at $\bar{X}$, and this is our MLE.

If $\bar{X} < 0$, then ℓ is decreasing on $[\bar{X}, \infty)$, hence on $[0, \infty)$ it is maximized at the MLE $\theta = 0$.

(2) Note $\bar{X}_n \xrightarrow{p} \theta$ by LLN. Consider the continuous function $g(t) = \max(t, 0)$. Then applying g to both sides of the convergence and using continuous mapping theorem yields consistency of $\hat{\theta}_n$.

(3) This is CLT for $\bar{X}_n$ plus the Delta method applied with g, since g is differentiable at $t > 0$ with slope 1.

(4) Ah, g is not differentiable at 0! The Delta method doesn't apply.

$P(\sqrt{n}\hat{\theta}_n \leqslant t) = P(\max(0, \sqrt{n}\bar{X}_n) \leqslant t)$

$t < 0$: equals zero.

$t \geqslant 0$: equals $P(\sqrt{n}\bar{X}_n \leqslant t)$, which tends to $\Phi(t)$.

So limiting cdf is $\mathbf{1}_{t\geqslant 0}\Phi(t)$, i.e. what is denoted $\max(N(0,1), 0)$. (We push all the mass left of zero to zero.)

(5) Add... could this have been useful in pset?

(6) We are given a particular testing procedure (supposedly Wald).

(a) Show that the test is pointwise size α asymptotically.

Let's start with the power under the null: $P_{\theta_0}(|T_n| > z_{\alpha/2})$

We know T_n is asymptotically normal under the null for $\theta_0 > 0$, so that $P_{\theta_0}(|T_n| > z_{\alpha/2})$ tends to α. The inner limsup is a lim equal to α, and the outer sup does nothing. We have shown that for each value of θ_0, the asymptotic power is α.

(b) Now we consider $\sup_{\theta_0>0} P_{\theta_0}(|T_n(\theta_0)| > z_{\alpha/2})$. Hmm... add...

Q3: (Large sample estimation and inference)

Ahh, the median!

(1) General definition of median: any number m for which $P(X \leqslant m) \geqslant 1/2$ and $P(X \geqslant m) \geqslant 1/2$. (Think of a cdf with a flat portion equal to 1/2; then you get an interval of medians.)

Add population median is unique; by evenness, $F(0) = 1/2$. Since $f(0) > 0$, $F(x) < 1/2$ for $x < 0$ and $F(x) > 1/2$ for $x > 0$. Hence $m = 0$ is unique.

$\hat{\theta}_{1/2}$ is unique and consistent for θ: consistency follows from $\hat{\theta}_n - \theta = O_p(n^{-1/2})$, where we used asymptotic normality. For uniqueness, note that $P(Y \leqslant m) = \sum_{i=1}^n \mathbf{1}_{X_i \leqslant m}$ and $P(Y \geqslant m) = \sum_{i=1}^n \mathbf{1}_{X_i \geqslant m}$.

Suppose $m = X_{(k)}$ for some k. Then $P(Y \leqslant m) = \frac{k}{n}$ and $P(Y \geqslant m) = \frac{n-k+1}{n}$. This implies $k = \frac{n+1}{2}$.

Suppose $m \in (X_{(k)}, X_{(k+1)})$ for some k. We show this is impossible. In this case $P(Y \leqslant m) = \frac{k}{n}$ and $P(Y \geqslant m) = 1 - \frac{k}{n}$. If these sum to 1, they must both be 1/2. Then $k = n/2$, which is not an integer, contradiction. Hence $m = X_{((n+1)/2)}$ is the unique sample median.

Is $E[\hat{\theta}_{(1/2)}]$ unbiased for θ? Yes, by symmetry: $X \sim 2\theta - X$. So $\hat{\theta}_{1/2}$ and $2\theta - \hat{\theta}_{1/2}$ have the same distribution. Take the expectation. (?)

(2) Find log-likelihood, score, hessian, CRLB.

$$L(\theta) = \prod_i f(X_i - \theta) \tag{42}$$

$$\ell(\theta) = \sum_i \log f(X_i - \theta) \tag{43}$$

$$\ell_\theta(\theta) = -\sum_i \frac{f'(X_i - \theta)}{f(X_i - \theta)} \tag{44}$$

$$\ell_{\theta\theta}(\theta) = \sum_i \frac{f''(X_i - \theta)f(X_i - \theta) - f'(X_i - \theta)^2}{f(X_i - \theta)^2} \tag{45}$$

$I(\theta) = nE[(\frac{f'(X_i-\theta)}{f(X_i-\theta)})^2] = n\int \frac{f'(x)^2}{f(x)}dx$

The CRLB is $CRLB(\theta) = I(\theta)^{-1}$. It does not depend on θ.

Include $1/n$ in these functions? Careful with the definition of I; size of the sample matters... Add...

(3) Did this in homework; we can take any consistent estimator and use a one-step update to upgrade it to an asymptotically *efficient* estimator.

(4) Also did in homework, via the asymptotic linear representation of $\hat{\theta}_{ML}$:

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{1}{\sqrt{n}} \sum_i \tilde{\psi}(X_i; \theta) + o_p(1) \tag{46}$$

where $\tilde{\psi} = (-E[\frac{\partial\psi}{\partial\theta}])^{-1}\psi$ is our modified influence function. Take the difference with the identity in (c) and manipulate.

(5) Let's restrict to shifts of the Cauchy density $f(x) = \frac{1}{\pi(1+x^2)}$.

(a) Show that Cauchy$(\theta, 1)$ is stable. Let us go via characteristic functions, since it is provided in the table:

$$E[\exp(it\bar{X})] = E[\exp(i\frac{t}{n}X_1)]^n \tag{47}$$

$$= \exp(i\theta\frac{t}{n} - |\frac{t}{n}|)^n \tag{48}$$

$$= \exp(i\theta t - |t|) \tag{49}$$

$$= E[\exp(itX_1)] \tag{50}$$

hence $\bar{X} \sim$ Cauchy$(\theta, 1)$.

$\bar{X}$ is consistent for θ by LLN.

(b) To compute ℓ_θ, we need to compute $f'(x)$.

$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$

$f'(x) = \frac{1}{\pi}\frac{-2x}{(1+x^2)^2}$

Thus $\frac{f'(x)}{f(x)} = \frac{-2x}{1+x^2}$.

The score equation is $\sum \frac{(X_i-\theta)}{1+(X_i-\theta)^2} = 0$. Why does this have at least one root? This is continuous, so let's look at extreme values of θ. For very negative θ, the function is positive. For very positive θ, the function is negative. Therefore by IVT there exists a root θ^*.

Assume FTSOC that $\theta^* > \max X_i$. Then the function is negative. Similarly, if $\theta^* < \min X_i$, then the function is positive. So we must have θ^* in between these quantities.

(c) By asymptotic efficiency, we have $\sqrt{n}(\hat{\theta}_{step} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$

$\sqrt{n}(\hat{\theta}_{1/2} - \theta) \xrightarrow{d} N(0, \frac{1}{4f(0)^2}) = N(0, \frac{\pi^2}{4})$

We must compute $I(\theta)$ (for a single observation) for the Cauchy family. Note that

$$I(\theta) = \int \frac{f'(x)^2}{f(x)} dx \tag{51}$$

$$= \int \frac{1}{\pi} \frac{4x^2}{(1+x^2)^3} dx \tag{52}$$

$$= \frac{1}{\pi} \int \frac{4}{(1+x^2)^2} dx - \frac{4}{\pi} \int \frac{1}{(1+x^2)^3} dx \tag{53}$$

Aha, they gave us a bivariate integrand so we can differentiate under the integral sign! The derivative is $\frac{\pi}{2u^{3/2}} = \frac{\pi}{2}$. The second derivative divided by 2 is $\frac{3\pi}{8}$. So the above information equals $2 - 3/2 = 1/2$ and the asymptotic variance is 2.

Therefore the relative efficiency is $2/(\pi^2/4) = 8/\pi^2 \approx 0.81$.

(d) One-sided asymptotic α level test based on $\hat{\theta}_{step}$. Show consistency (i.e. on the alternative, power goes to 1).

Use the Wald statistic (i.e. "based on an estimator") $T_n = \frac{1}{\sqrt{2}} \cdot \sqrt{n}(\hat{\theta}_{step} - \theta_0)$ which is asymptotically $N(0,1)$. This is in (z_α, ∞) with probability tending to α. So reject if $T_n > c$, where $c = z_\alpha$.

Power function on the alternative space: $\beta_n(\theta) = P_\theta(T_n > c)$

$= P_\theta(\frac{1}{\sqrt{2}}\sqrt{n}(\hat{\theta}_{step} - \theta) + \frac{1}{\sqrt{2}}\sqrt{n}(\theta - \theta_0) > c)$

$\to 1 - \Phi(c - \frac{1}{\sqrt{2}}\sqrt{n}(\theta - \theta_0))$

For $\theta > \theta_0$, this tends to $1 - \Phi(-\infty) = 1$, proving consistency.

F20 midterm:

Q1: Probability theory.

Done. Don't forget Cauchy-Schwarz!

Add...