MS Theory Exam Topics 2022

Convergence Theory

Def: Let $F_1, ..., F_n$ be the corresponding CDFs of $Z_1, ..., Z_n$. For an RV Z with CDF F, we say that Z_n **converges in distribution** to Z iff $lim_{n\to\infty} F_n(x) = F(x)$ for every x.
[*Note*: We can show this by showing $lim_{n\to\infty} M_{X_n}(t) = M_X(t)$]

Def: We say that a sequence of RV, Z_n, **converges in probability** to an RV, Z, iff $lim_{n\to\infty} P(|Z_n - Z| > \epsilon) = 0$

Def: We say that a sequence of RV, Z_n, **converges almost surely** to an RV, Z, iff $P(lim_{n\to\infty} Z_n = Z) = 1$

Continuous Mapping Theorem: For a continuous function g,

$X_n \to^d X \Rightarrow g(X_n) \to^d g(X)$ and $X_n \to^p X \Rightarrow g(X_n) \to^p g(X)$

Slutsky's Theorem: Let $X_n \to^d X$, $Y_n \to^p c$.

Then (1) $X_n + Y_n \to^d X + c$, (2) $X_n Y_n \to^d cX$, and (3) $X_n / Y_n \to^d X_n / c$

Markov's Inequality: Let X be a nonnegative RV.

Then for any $\epsilon > 0$, $P(X \geq \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$

Chebyshev's Inequality: Let X be a RV with finite variance.

Then for any $\epsilon > 0$, $P(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}$

Weak LLN: If $X_1, ..., X_n$ are distributed iid with finite mean and variance, then $\bar{X} \to^p \mathbb{E}[X_1]$

Central Limit Theorem: If $X_1, ..., X_n$ are distributed iid with finite mean and variance, then $\sqrt{n}(\frac{\bar{X} - \mathbb{E}[X_1]}{Var(X_1)}) \to^d N(0, 1)$

Hoeffding's Inequality: Let $X_1, ..., X_n$ be iid RVs such that $0 \leq X_1 \leq 1$ and let $\bar{X}$ be the sample average. Then for any $\epsilon > 0$, $P(|\bar{X} - \mathbb{E}(\bar{X})| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$

Jensen's Inequality: If X is a RV and f is a convex function, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.

Moment Generating Functions

Def: The MGF of a RV X is $M_X(t) = \mathbb{E}(e^{tX})$. Moreover, the j^{th} moment of RV X,

$$\mathbb{E}[X^j] = M_X^{(j)}(0) = \frac{d^j M_X(t)}{dt^j}|_{t=0}$$

Note: (1) $M_{aX+b}(t) = e^{bt}M_X(at)$ and (2) $M_{X+Y}(t) = M_X(t)M_Y(t)$

Regression & Classification

- For a simple linear regression, the OLS $\hat{\beta}$ estimates are defined as

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \text{ and } \hat{\beta}_1 = \frac{\sum_i^n (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2} = \frac{(sample)Cov(x,y)}{(sample)Var(x)}$$

$$\text{where } Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2}{n\sum_i (x_i-\bar{x})^2},\ Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i-\bar{x})^2}, \text{ and } Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2\bar{x}}{\sum_i (x_i-\bar{x})^2}$$

- In general $\hat{\beta} = (X^TX)^{-1}X^TY$ and $Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$
- To measure how well $g(X)$ predicts Y we use

$$MSE(g) = \mathbb{E}[(Y - g(X))^2] = \mathbb{E}[g(X) - Y]^2 + Var(g(X) - Y)$$

- We can find the 'best' classifier $c(X)$ for Y (where 'best' is defined by some loss function, $L(c(X), Y)$) by finding the c which minimizes $R(c) = \mathbb{E}[L(c(X), Y)]$.

Other Estimators

Method of Moments estimator: $\hat{m}_j(\theta) = \frac{1}{n}\sum_i X_i^j$ estimates the j^{th} moment of X (i.e. $\mathbb{E}[X^j]$)

Bayesian estimators

- Posterior mean, $\hat{\theta}_\pi = \mathbb{E}(\theta|X_1, ..., X_n) = \int \theta \cdot \pi(\theta|X_1, ..., X_n)d\theta$
- Maximum a posteriori, $\hat{\theta}_{MAP} = argmax_\theta \pi(\theta|X_1, ..., X_n)$

where $\pi(\theta|X_1, ..., X_n)$ is the posterior distribution of θ

Empirical Risk Minimization: $\hat{\theta} = argmin_\beta \frac{1}{n}\sum_i^n L(Y_i, f_\beta(X_i))$ for some loss function $L(a, b)$

Note: Maximum likelihood/least squares estimation is the special case of ERM where

$$L(Y_i, f_\beta(X_i)) = (Y_i - X_i^T\beta)^2$$

Sufficient Statistics

Def (SS): (1) $T(X)$ is SS for $\mathcal{P}$ if $T(X)$ contains all relevant information that X provides about unknown $\mathcal{P}_\theta$; (2) $T(X)$ is SS for $\mathcal{P}$ if $X|T(X)$ does not depend on θ

Fisher-Neyman Factorization Theorem: $T(X)$ is SS wrt $\mathcal{B}$ iff the pdf/pmf $f_\theta(x)$ can be factorized as

$$f_\theta(x) = g_\theta(T(x))h(x)$$

Helpful Lemmas for SS

- *Lemma 11.1*: If $T(X)$ is SS wrt the class of pdfs, $\mathcal{B}$, and $\mathcal{B}_1 \subset \mathcal{B}$, then $T(X)$ is also SS wrt $\mathcal{B}_1$.
- *Lemma 11.2*: If $T(X)$ is SS for $(X, \mathcal{B})$ and $S(T(X))$ is SS for $(X, \mathcal{Q})$, then $S(T(X))$ is also SS for $(X, \mathcal{B})$

Def (Minimal SS): $T^*(X)$ is a minimal SS for $\mathcal{B}$ if, for any SS, $T(X)$, there exists h s.t $T^*(X) = h(T(X))$

Lehmann-Scheffe Theorem: Suppose $X \sim \{f_\theta(X), \theta \in \Omega\}$. Then $T^*(X)$ is minimal SS if it satisfies the following sufficient condition:

$$\text{For any } x, y, \in X,\ T^*(x) = T^*(y) \iff \frac{f_\theta(y)}{f_\theta(x)} \text{ is } \theta\text{-free}$$

Minimal SS for Special Cases

- *Prop 11.48*: Let X have pdf $f_\theta(x) = [a(\theta)]^n exp\{\theta_1 \sum_i T_1(x_i) + ... + \theta_k \sum_i T_k(x_i)\} \Pi_i^n h(x_i)$. Then $(\sum_i T_1(x_i), ..., \sum_i T_k(x_i))$ is minimal SS iff $\Omega = (\theta_1, ..., \theta_k)$ has $dim(k)$.
- *Prop 11.47*: Let X be distributed iid with pdf $[B(\theta)]^{-1}\mathbb{I}_{[\theta,a)}(x)b(x)$. Then $X_{(1)}$ is minimal SS for θ.
- *Prop 11.52*: Let X be distributed iid with pdf $[B(\theta_1, \theta_2)]^{-1}\mathbb{I}_{[\theta_1,\theta_2]}(x)b(x)$. Then $(X_{(1)}, X_{(n)})$ is minimal SS for $\boldsymbol{\theta}$.

Def (Ancillary Statistic): A statistic $V = V(X)$ is an ancillary statistic wrt a distribution family $\mathcal{B}$ if the distribution of V is θ-free.

Note: For the location/scale/location-scale family, any statistic which is location/scale/location-scale invariant is an ancillary statistic.

Def (Complete SS): A statistic $T(X)$ is complete wrt $\mathcal{B}$ if, for any function g,

$$\mathbb{E}_\theta[g(T(X))] \text{ is } \theta\text{-free} \Rightarrow g(T) \text{ is a constant function}$$

which is equivalent to:

$$\mathbb{E}_\theta[g(T(X))] = 0 \Rightarrow g(T) = 0$$

Helpful Theorems for Complete SS

- *Basu's Theorem*: If T is complete and sufficient, T is independent of any ancillary statistic V.
- *Theorem 12.1*: If T is complete, then no non-constant function of T is ancillary.
- *Theorem 12.2*: If T is a complete SS, it is also minimal.

Tools for Showing Complete SS

- *Prop 12.1*: Suppose $T = [T_1...T_k]^T$ has pdf $f_\theta(t_1, ..., t_k) = a(\theta)exp\{\sum_j^k \theta_j t_j\}h(t)$. Then if $(\theta_1, ..., \theta_k) = \Omega$ contains a k-dimensional rectangle, T is complete.
- *Prop 12.3*: Suppose $X_1, ..., X_n$ is an iid sample from the truncation pdf, $f_\theta(x) = [B(\theta)]^{-1}\mathbb{I}_{(a,\theta]}(x)b(x)$. Then $T = X_{(n)}$ is complete.

Tools for Showing an SS is *not* Complete

- Find ancillary statistic which is not independent of T (Basu's Theorem)
- Show that T is not minimal (Theorem 12.2)
- Find $g(T)$ which violates definition of complete SS

Def (UMVUE): An unbiased estimator $\hat{\tau}$ of $\tau(\theta)$ is the UMVUE if it has the smallest variance among all unbiased estimators of $\tau(\theta)$

RBLS Theorem: Assume (1) there is an unbiased estimator $\tilde{\tau}(X)$ of $\tau(\theta)$ and (2) there is a complete SS, $T = T(X)$ for θ. Then $\hat{\tau}(T) = \mathbb{E}[\tilde{\tau}(X)|T]$ is the unique UMVUE for $\tau(\theta)$.

Note: Aside from using RBLS Theorem directly, we can also find the UMVUE for $\tau(\theta)$ via the "UMVUE Supermarket": find $\phi(T)$ which is an unbiased estimator of $\tau(\theta)$. This is the UMVUE.

Information Inequality & MLE

Def (FIN): The Fisher Information Number (FIN) of a regular distribution family $\mathcal{B}$ is

$$I_x(\theta) = \mathbb{E}_\theta[(\frac{dlog\mathcal{L}(\theta)}{d\theta})^2] = -\mathbb{E}_\theta[\frac{d^2log\mathcal{L}(\theta)}{d\theta^2}] = Var_\theta(\frac{dlog\mathcal{L}(\theta)}{d\theta})$$

Cramer-Rao Lower Bound: Given statistical family $(X, \mathcal{B})$ and any estimator $T(X)$ then

$$Var_\theta(T(X)) \geq \{\frac{d}{d\theta}\mathbb{E}_\theta[T(X)]\}^2/I_x(\theta)$$

Note: Equality holds iff $f_\theta(x) = e^{A(\theta)}e^{B(\theta)T(x)}e^{C(x)}$

Def (FIM): The Fisher Information Matrix of a regular multivariate distribution family $\mathcal{B}$ is

$$I_x(\theta) = \mathbb{E}_\theta[\{\nabla_\theta log f_\theta\}\{\nabla_\theta log f_\theta\}^T]$$

where $\nabla_\theta log f_\theta = [\frac{\partial log f_\theta(x)}{\partial\theta_1}, ..., \frac{\partial log f_\theta(x)}{\partial\theta_k}]^T \in \mathbb{R}^k$

Note: $[I_x(\theta)]_{ij} = -\mathbb{E}_\theta[\frac{\partial^2 log f_\theta(x)}{\partial\theta_i\partial\theta_j}]$

Cramer-Rao Lower Bound (Multivariate):

$$Var_\theta(T(X)) \geq \{\nabla_\theta\mathbb{E}[T(X)]\}^T I_x(\theta)^{-1}\{\nabla_\theta\mathbb{E}[T(X)]\}$$

Def (MLE): The MLE is defined as $\hat{\theta} = argmax_\theta f_\theta(x)$

Fisher-Cramer Theorem: $\hat{\theta}$ is consistent and asymptotically attaining CR-LB

$$\iff \sqrt{n}(\hat{\theta} - \theta) \to^d N(0, I_{x_i}(\theta_0)^{-1})$$

Remark: By invariance of the MLE, delta method, and continuous mapping theorem,

$$\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \to^d N(0, [\tau'(\theta)]^2 I_{x_i}(\theta_0)^{-1})$$

Hypothesis Testing

Def (Neyman-Pearson Criterion):

1. *Power function*: The power function is the probability of rejecting the null hypothesis using test ϕ, given θ is the true parameter
$$\Pi_\phi(\theta) := \mathbb{E}_\theta[\phi(X)]$$
2. *Size*: The size of test ϕ is the worst potential Type I error rate of all $\theta \in \Omega_0$
$$sup_{\theta\in\Omega_0}\{\Pi_\phi(\theta)\} = sup_{\theta\in\Omega_0}\{\mathbb{E}[\phi(X)]\}$$
3. *Level*: A test ϕ has level α if its size is less than or equal to α
4. *Uniformly most powerful (UMP)*: A test is UMP level α if it is the test with smallest Type II error/highest power among all level α tests
$$\Pi_\phi(\theta) = sup_{\phi', level\alpha}\{\Pi_{\phi'}(\theta)\} \text{ for all } \theta \in \Omega_1$$

Two-point Test ($H_0 : \theta = \theta_0; H_1 : \theta = \theta_1$)

- *Neyman-Pearson Theorem*: The most powerful level α test for the two-point hypothesis is
$$\phi(x) = \begin{cases} 1, & \lambda(x) = \dfrac{f_1(x)}{f_0(x)} > c \\ 0, & \lambda(x) < c \\ \delta(x) & \lambda(x) \end{cases}$$
where c and $\delta(x)$ are chosen s.t. $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

One-sided Test ($H_0 : \theta = \theta_0; H_1 : \theta > \theta_0$ or $H_0 : \theta \leq \theta_0; H_1 : \theta > \theta_0$)

- *UMP Existence Theorem*: If $f_\theta(\cdot)$ is MLR in T, then the UMP level α test for a one-sided hypothesis is
$$\phi(x) = \begin{cases} 1, & t > c_\alpha \\ 0, & t < c_\alpha \\ \delta_\alpha & t = c_\alpha \end{cases}$$
where c_α and δ_α are chosen s.t. $\mathbb{E}_{\theta_0}[\phi(T)] = \alpha$.

  Note: $f_\theta(\cdot)$ is MLR in some $T = T(x)$ if $\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = g(T(x))$ increases with T for all $\theta_0 < \theta_1 \in \Omega$.

Two-sided Test ($H_0 : \theta \in \Omega_0; H_1 : \theta \in \Omega_1 = \Omega/\Omega_0$) *Note*: Be sure to plug in the MLE estimates to calculate $\lambda(x)$

- If Ω_0 and Ω_1 are uniformly or pointwise separated, the recommended test is
$$\phi(x) = \begin{cases} 1, & \lambda(x) = \dfrac{f_{\hat{\theta}_0}(x)}{f_{\hat{\theta}_1}(x)} < 1 \\ 0, & \lambda(x) \geq 1 \end{cases}$$
- If $\Omega_0, \Omega \in \mathbb{R}^p$, $dim(\Omega) = k - r$ and $dim(\Omega_0) = k - r - s$ then the recommended level α test is
$$\phi(x) = \begin{cases} 1, & -2log\lambda(x) < \chi^2_{s,1-\alpha} \\ 0, & -2log\lambda(x) > \chi^2_{s,1-\alpha} \end{cases}$$
because, by Wilk's Theorem, $-2log\lambda(x) \to^d \chi^2_s$ under the null hypothesis.