MATH357 - Statistics

Summary

Winter 2025, Prof. Abbas Khalili.
Notes by Louis Meunier

1 Probability Prerequisites

Definition 1: $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$

Theorem 1 (Properties of Normal Distributions): Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

(i) $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$;

(ii) $\overline{X}_n$ and S_n^2 are independent;

(iii) $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$;

(iv) If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{(\nu)}$, $\frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$. In particular,

$$\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \sim t(n-1).$$

Similarly, if $Y_j \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, $j = 1, ..., m$ another independent normal sample, then

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu - \tilde{\mu})}{S_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(m+n-2), \qquad S^2_{\text{pooled}} := \frac{(n-1)S_n^2 + (m-1)S_m^2}{m+n-2}.$$

(v) If $U \sim \chi^2_{(m)}$, $V \sim \chi^2_{(n)}$ are independent rv's, then $\frac{U/m}{V/n} \sim F(m,n)$.

Theorem 2 (Order Statistics): If $X_1, ..., X_n$ iid rv's with CDF F, the CDF's of the min, max order statistics are respectively

$$F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n, \qquad F_{X_{(n)}}(x) = [F(x)]^n,$$

and generally, for $1 \leq j \leq n$,

$$F_{X_{(j)}}(x) = \sum_{k=j}^n \binom{n}{k} F^k(x)[1 - F(x)]^{n-k}.$$

Theorem 3 (Convergence Theorems):

(i) (Slutsky's) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a$, then $X_n + Y_n \xrightarrow{d} X + a$, $X_n Y_n \xrightarrow{d} aX$ and, if $a \neq 0$, $X_n / Y_n \xrightarrow{d} X/a$.

(ii) (Continuous Mapping Theorem) If $X_n \xrightarrow{P,d} X$ and g continuous on a set C where $P(X \in C) = 1$, then $g(X_n) \xrightarrow{P,d} g(X)$.

(iii) (WLLN) If X_i iid rv's with mean μ and finite second moment, $\overline{X}_n \xrightarrow{P} \mu$.

(iv) (First-Order Delta Method) If $\sqrt{n}(X_n - \mu) \xrightarrow{d} V$ and g a function such that g' exist and is nonzero at $x = \mu$, then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} g'(\mu) \cdot V.$$

(v) (Second-Order Delta Method) If $\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, and g a function with $g'(\mu) = 0$ but $g''(\mu) \neq 0$, then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} \mathcal{N}\left(0, g'(\mu)^2 \sigma^2\right).$$

Theorem 4 (Empirical CDF Properties): Let $X_1, ..., X_n$ be iid with cdf F. The ECDF is the rv defined by, for $x \in \mathbb{R}$, $F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$. The following hold:

(i) $nF_n(x) \sim \text{Bin}(n, F(x))$; in particular,

$$\mathbb{E}[F_n(x)] = F(x), \qquad \text{Var}(F_n(x)) = \frac{1}{n} F(x)(1 - F(x))$$

(ii) $\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d} \mathcal{N}(0, 1)$

(iii) $F_n(x) \xrightarrow{P} F(x)$

2 Parametric Inference

Definition 2 (Qualities of Estimators):

(i) The *bias* of an estimator $\hat{\theta}$ of θ is defined $\text{Bias}\left(\hat{\theta}\right) = \mathbb{E}_\theta\left[\hat{\theta}\right] - \theta$. $\hat{\theta}$ is *unbiased* if it has zero bias.

(ii) The *mean-squared error* (MSE) is defined $\text{MSE}\left(\hat{\theta}\right) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$.

(iii) We say $\hat{\theta}$ *unbiased* if $\hat{\theta} \xrightarrow{P} \theta$.

Theorem 5 (Cramer-Rau Lower Bound): For a parametric family $\{p(\cdot, \theta) : \theta \in \Theta\}$, if $T(\boldsymbol{X})$ an unbiased estimator of a function of a parameter $\tau(\theta)$, with finite variance, then

$$\text{Var}(T(\boldsymbol{X})) \geq \frac{[\tau'(\theta)]^2}{I(\theta)},$$

for every $\theta \in \Theta$ in the, where $I(\theta) := \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\boldsymbol{X})\right)^2\right]$ the *Fisher information* of the parametric family and assuming the denominator is finite, and moreover:

(i) $\{p_\theta : \theta \in \Theta\}$ has common support independent of θ

(ii) for any $\boldsymbol{x}$ and $\theta \in \Theta$, $\frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\boldsymbol{x}) < \infty$

(iii) for any statistic $h(\boldsymbol{X})$ with finite first absolute moment, differentiation under the integral holds ie $\frac{\mathrm{d}}{\mathrm{d}\theta} \int h(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int h(\boldsymbol{x}) \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$

Moreover, equality occurs iff there exists a function $a(\theta)$ such that $a(\theta)\{T(\boldsymbol{x}) - \tau(\theta)\} = \frac{\mathrm{d}}{\mathrm{d}\theta} \log p(\boldsymbol{x}; \theta)$.

Remark 1: If p_θ twice differentiable in θ and $\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\boldsymbol{X})\right]$ differentiable "under the integral sign", then $I(\theta) = -\mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} p_\theta(\boldsymbol{X})\right]$.

If working with iid rv's, then the denominator becomes $n I_1(\theta)$ where $I_1(\theta)$ the Fisher information of a single rv.

Theorem 6 (Neyman-Fisher Factorization): A statistic $T(\boldsymbol{X})$, $\boldsymbol{X} \sim p_\theta(\cdot)$ is called *sufficient* for θ if the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X}) = t$ is independent of θ. $T(\boldsymbol{X})$ is sufficient iff there are functions $h(\cdot), g(\cdot; \theta)$ such that $p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) g(T(\boldsymbol{x}), \theta)$.

Theorem 7: Any one-to-one function of a sufficient statistic is still sufficient.

Theorem 8 (Minimal Sufficiency): A sufficient statistic is minimal if it is a function of every other sufficient statistic. For a parametrized pdf $p_\theta(\cdot)$, suppose $T(\boldsymbol{x}) = T(\boldsymbol{y}) \Leftrightarrow \frac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{y})}$ does not depend on θ. Then, $T(\boldsymbol{X})$ is minimally sufficient.

Definition 3 (Completeness): An estimator $\hat{\theta}$ is called *complete* if $\mathbb{E}\left[g\left(\hat{\theta}\right)\right] = 0$ for every θ implies $g = 0$ (a.s.).

Theorem 9 (Rao-Blackwell): Let $U(\boldsymbol{X})$ be unbiased for $\tau(\theta)$ and $T(\boldsymbol{X})$ sufficient, and define $\delta(t) := \mathbb{E}_\theta[U(\boldsymbol{X}) \mid T(\boldsymbol{X}) = t]$. Then $\delta(\boldsymbol{X})$ is unbiased for $\tau(\theta)$, and has smaller variance then $U(\boldsymbol{X})$.

Theorem 10 (Lehmann-Scheffé): Let $T(\boldsymbol{X})$ be complete and sufficient and $U(\boldsymbol{X}) = h(T(\boldsymbol{X}))$ unbiased with finite second moment, then $U(\boldsymbol{X})$ is the UMVUE for $\tau(\theta)$.

Remark 2: Combine these two theorems to systematically construct UMVUEs starting from an (arbitrary) unbiased estimator and a complete and sufficient statistic.

Theorem 11 (Existence of a UMVUE): An estimator $U(\boldsymbol{X})$ of $\tau(\theta) = \mathbb{E}[U(\boldsymbol{X})]$ is the best unbiased estimator iff $\mathrm{Cov}(\delta(\boldsymbol{X}), U(\boldsymbol{X})) = 0$ for every estimator $\delta(\boldsymbol{X})$ such that $\mathbb{E}[\delta(\boldsymbol{X})] = 0$.

3 Systematic Parameter Estimation

Definition 4 (Method of Moments): The *method of moments* estimator(s) for rv's $X_1, \dots, X_n \overset{\text{iid}}{\sim} f_\theta$ is given by solving the system

$$\frac{1}{n} \sum_{i=1}^{n} X_i^j = \mu_j(\theta) := \mathbb{E}\left[X_i^j\right],$$

for j as high as we need for the system of equations to have solutions.

Definition 5 (Minimum Likelihood Estimation (MLE)): An estimator $\hat{\theta}_n$ is said to be an MLE of a parametric family if it maximizes the likelihood (resp. log likelihood) function (for any post-experimental data $\boldsymbol{x}$)

$$\begin{aligned} L_n &: \Theta \to [0, \infty) \\ L_n(\theta) &= p_\theta(\boldsymbol{x}) \end{aligned}, \qquad \left(\begin{aligned} \ell_n &: \Theta \to (-\infty, \infty) \\ \ell_n(\theta) &= \log L_n(\theta) \end{aligned} \right).$$

If differentiable, one can solve for the (at least a candidate) MLE by solving the likelihood equations $\partial_\theta L_n = 0$ or equivalently $\partial_\theta \ell_n = 0$.

Remark 3: Since log monotonic increasing, the likelihood/log-likelihood functions are equivalent and thus one should use which ever one is more convenient (lots of parametric families have exponentials, so using log is helpful).

Theorem 12 (Properties of MLEs): We assume "the regularity conditions".

(i) (Invariance) If $\hat{\theta}$ the MLE of θ and $\tau(\theta)$ a function of θ, then $\tau\left(\hat{\theta}\right)$ the MLE of $\tau(\theta)$.
(ii) $\hat{\theta}$ is consistent.
(iii) $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, [I_1^{-1}(\theta_0)])$ where θ_0 the "true value".
(iv) (1st Bartlett Identity) $\mathbb{E}_\theta\left[\frac{\partial \log f(X)}{\partial \theta}\right] = 0$.

Definition 6 (Bayesian Estimation): Let $\boldsymbol{X} \sim p_\theta$ where θ also random, with pdf/pmf $\pi(\theta)$, called the *prior distribution* of θ. The *posterior distribution* is defined as $\pi(\theta|\boldsymbol{x})$, which by Baye's is proportional to $p_\theta(\boldsymbol{x})\pi(\theta)$. A *loss function* $L(\delta(\boldsymbol{X}), \theta)$ is a function assigning a "penalty" to an estimator $\delta(\boldsymbol{X})$, for instance the L^2-loss given by $(\delta(\boldsymbol{X}) - \theta)^2$. *Baye's risk* given a loss function L is defined

$$R(\delta) := \mathbb{E}_\pi\left[\mathbb{E}_{\boldsymbol{X}|\theta}[L(\delta(\boldsymbol{X}), \theta)]\right].$$

Then, *Baye's estimator* is simply $\hat{\delta}(\boldsymbol{X}) := \operatorname{argmin}_\delta R(\delta)$.

Theorem 13: For L the L^2-loss function, the Baye's estimator is

$$\hat{\delta}(\boldsymbol{X}) = \mathbb{E}_{\theta|\boldsymbol{X}=x}[\theta|\boldsymbol{X}].$$

Remark 4: So, given p_θ and $\pi(\theta)$, the typical steps to finding $\hat{\delta}(\boldsymbol{X})$ are:

(i) compute $p_\theta(\boldsymbol{x})\pi(\theta)$, and deduce the distribution of $(\theta|\boldsymbol{X})$; if deducing is not possible, one will have to compute the full proportionality constant i.e.

$$\pi(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{\int_\Theta p(\boldsymbol{x}|\theta)\pi(\theta)\,\mathrm{d}\theta}.$$

(ii) hopefully the distribution found in (i) has a well-known mean, which is then equal to the Baye's estimator $\hat{\delta}(\boldsymbol{X})$ by the previous theorem; else, one in general would have to solve $\mathbb{E}_{\theta|\boldsymbol{X}}[\theta|\boldsymbol{X}]$.

4 Confidence Intervals and Hypothesis Testing

Definition 7 (Pivotal Quantity): A random function $Q = Q(\boldsymbol{X}; \theta)$ is called a *pivotal quantity* (PQ) for a distribution if its distribution is independent of θ.

Remark 5: Given a *confidence level* α, we wish to find $L(\boldsymbol{X}), U(\boldsymbol{X})$ such that $P(L \leq \theta \leq U) = 1 - \alpha$. Supposing we have a PQ Q, first find constants c_1, c_2 (which will by virtue be independent of θ) such that

$$P(c_1 \leq Q(\boldsymbol{X}; \theta) \leq c_2) = 1 - \alpha.$$

Invert/solve then Q for $\boldsymbol{X}$ to find $L(\boldsymbol{X}), U(\boldsymbol{X})$ as functions c_1, c_2.

Remark 6: The general technique to find PQs is to start with a minimal sufficient statistic, and transform its distribution to be independent of θ and moreover to be one for which we have easy access to its quantiles (typically chi-squared, since many statistics involve exponentials so its often possible to rescale such into chi-squareds).

Remark 7: If not possible to find (or just difficult) to find an exact confidence interval, one can just appeal to CLT and compute an approximate CI using normal-distribution theory.

Theorem 14 (Neyman-Pearson Lemma): Let

$$\phi(\boldsymbol{X}) := \begin{cases} 1 \text{ if } p(\boldsymbol{X}; \theta_1) > k \cdot p(\boldsymbol{X}; \theta_0) \\ 0 \text{ if if } p(\boldsymbol{X}; \theta_1) < k \cdot p(\boldsymbol{X}; \theta_0) \end{cases},$$

and either if equal, where k is such that $P_{\mathcal{H}_0}(\text{rejecting } \mathcal{H}_0) = \alpha$. Then, ϕ is the UMP test in the class of all tests at significance level α.

Remark 8: If simple-simple, *always* use this lemma!

Definition 8 (Likelihood Ratio Statistic): The *likelihood ratio statistic* (LR) is the quantity

$$\lambda_n(\boldsymbol{X}) := \frac{L_n\left(\hat{\theta}_{\mathrm{MLE}, \mathcal{H}_0}\right)}{L_n\left(\hat{\theta}_{\mathrm{MLE}}\right)}.$$

A test based on LR is

$$\phi(\boldsymbol{X}) = \begin{cases} 1 \text{ if } \lambda_n(\boldsymbol{X}) < C \\ 0 \text{ else} \end{cases}, \qquad C \text{ s.t. } P(\lambda_n(\boldsymbol{X}) < C) = \alpha.$$

Remark 9: This test should be used when the hypotheses are not simple-simple.

Theorem 15: Under the regularity conditions, $-2\log(\lambda(\boldsymbol{X})) \overset{d}{\approx} \chi^2_d$, where $d := \dim(\Theta) - \dim(\Theta_0)$.

Remark 10: Sometimes its hard to manipulate/solve the necessary condition $P(\lambda_n(\boldsymbol{X}) < C) = \alpha$ explicitly for what C should be. This theorem says that you can take $C = \exp\left(-\frac{\chi^2_{d,\alpha}}{2}\right)$ to find an approximate test.

5 Some MLEs and Such To Remember

Distribution	**Sufficient Statistic**	**UMVUE**	**MLE**
Exponential, $f(x, \theta) = h(x)c(\theta)\exp(\omega(\theta)T_1(x))$	$\sum_{i=1}^n T_1(X_i)$	$\frac{1}{n}\sum_{i=1}^n T_1(X_i)$	
Poisson(λ)	$f\left(\sum_{i=1}^n X_i\right)$	$\overline{X}_n$	$\overline{X}_n$

$\mathcal{U}(0, \theta)$	$X_{(n)}$	$\frac{n+1}{n} X_{(n)}$	$X_{(n)}$
$\mathcal{N}(\mu, \sigma^2)$ μ, σ^2 unknown	$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$	$\left(\overline{X}_n, S_n^2\right)$	$\left(\overline{X}_n, \frac{n-1}{n} S_n^2\right)$
$\text{Ber}(\theta)$	$\sum_{i=1}^n X_i$	$\overline{X}_n$	$\overline{X}_n$
$f(x; \theta) = e^{-(x-\theta)}, x \geq \theta$	$X_{(1)}$	$X_{(1)} - \frac{1}{n}$	$X_{(1)}$
$\theta e^{-\theta x}$	$\sum_{i=1}^n X_i$	$(n-1)/\sum_i^n X_i$	$1/\overline{X}_n$

Remark 11: Recall that any one-to-one function of a (minimal) sufficient statistic is still a (minimal) sufficient statistic.