Recap Handouts 1-4
Surya Tokdar

Stat Model

- Data $X \in S$, $\Theta$.
- Model $X \sim f, f \in \mathcal{F} =$ some family of pdfs on $S$.

**Parametric family**

- $\mathcal{F}$ can be identified with some $\Theta \subset \mathbb{R}^d$
- Example:
  - $\mathcal{F} = \{ \text{all bivariate normals } f(x_1, x_2) \}$, $\Theta = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$
  - $f \mapsto \left( \begin{array}{c}
  \mu_1 := \int x_1 f(x_1, x_2) dx_1 dx_2 \\
  \sigma_{11} := \int x_2 f(x_1, x_2) dx_1 dx_2 - \mu_1^2 \\
  \sigma_{22} := \int x_2^2 f(x_1, x_2) dx_1 dx_2 - \mu_2^2 \\
  \rho := \int \frac{x_1 x_2 f(x_1, x_2) dx_1 dx_2}{\sigma_{11} \sigma_{22}}
\end{array} \right)$
- Use $\theta \in \Theta$ as the index / parameter: $\mathcal{F} = \{ f(x|\theta) : \theta \in \Theta \}$.

**Non-parametric family**

- $\mathcal{F}$ can **not** be identified with any finite dimensional Euclidean set!
- Example:
  - $\mathcal{F} = \{ \text{all finite mixtures of bivariate normals} \}$, $\Theta = \{ \text{all } k \text{-component mixtures of BVN} \}$
- Another example:
  - $\mathcal{F} = \{ \text{all bivariate } f(x_1, x_2) \text{ with } \int x_1^2 f, \int x_2^2 f < \infty \}$
  - Includes all finite mixtures of bivariate normals

**Hypothesis Testing**

- Model $X \sim f, f \in \mathcal{F}$ (parametric or non-parametric)
- $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$ [partition generated from scientific context]
- Decide between $H_0 : f \in \mathcal{F}_0$ vs. $H_1 : f \in \mathcal{F}_1$

- For a parametric model $\mathcal{F} = \{ f(x|\theta) : \theta \in \Theta \}$
  - $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \iff \Theta = \Theta_0 \cup \Theta_1$
  - Decide between $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$

**Fisher-Bayes-Laplace: Likelihood**

- Once $X = x$ is observed,
  - Score every $f \in \mathcal{F}$ with $L_x(f) = f(x)$
  - Relative score: $\frac{L_x(f_1)}{L_x(f_2)} = \frac{f_1(x)}{f_2(x)}$
- Relative score = 2 means $f_1$ explains observed data $X = x$ twice as better as $f_2$.

- Log-likelihood $\ell_x(f) = \log L_x(f) = \log f(x)$
  - log-relative score: $\ell_x(f_1) - \ell_x(f_2)$

- For a parametric model, equivalent to work with $L_\theta(\theta), \ell_\theta(\theta)$. 
Laplace-Bayes: Bayesian Approach

- Once $X = x$ is observed,
  - Construct a "pdf" $\pi(f|x)$ on $\mathcal{F}$: $\pi(f|x) \propto L_0(f)$
  - Decision based on
    $$P(H_0|X = x) = \int_{F_0} \pi(f|x) df = 1 - P(H_1|X = x).$$
- Easiest to think of parametric model:
  - $\mathcal{F} = \{(f(x|\theta) : \theta \in \Theta), \Theta \subset \mathbb{R}^d$,
  - $\pi(\theta|x) \propto L_0(\theta)$, valid if $L_0(\theta) d\theta < \infty$.
- More generally, $\pi(f|x) \propto L_0(f)\pi(f)$, $\pi(f)$ is prior pdf
  - Can handle non-parametric $\mathcal{F}$

Fisher: ML

- Parametric model: $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$
  $$LR(X) = \frac{\max_{\theta_0 \in \Theta_0} L_X(\theta)}{\max_{\theta_0 \in \Theta_0} L_X(\theta)}$$
- Larger $LR(X) \iff$ stronger evidence against $\Theta_0$
- Significance testing once $X = x$ is observed:
  $$p-value = \max_{\theta_0 \in \Theta_0} P(LR(X) \geq LR(x|\theta))$$

Misleading Interpretation

- $p-value = 2\%$ means either $H_0$ is false or something rare (2 in 100 chance) has happened.
- Does not assign any relative odds to “either” and “or” parts
- Frequently misinterpreted as saying $P(H_0|X = x) = 2\%$.
  - That’s not how Fisher would interpret
  - To talk about $P(H_0|X = x)$ requires “pdf” on $\mathcal{F}$, i.e., Bayesian approach
  - May not numerically match a/any reasonable Bayesian answer

A Small Example

- Rare disease $D$, affects 1% of the population
- Clinical test, success rate = 99% if $D$ and 98% if $\overline{D}$
- Data $X \in \{\text{negative}, +\}$ test result on some individual with unknown disease status $\theta$
  - $H_0 : \theta = \overline{D}$ vs $H_1 : \theta = D$
  - $LR(+) = \frac{0.99}{0.98} = 49.5$, $LR(\text{−}) = \frac{0.98}{0.99} = 1$.
- Observation: $X = +$.
  - $p-value = 2\%$
  - Standard Bayes rule: $P(H_0|X = +) = \frac{0.99 	imes 0.01}{0.99 	imes 0.01 + 0.02 	imes 0.99} = \frac{2}{3}$

Neyman-Pearson: Frequentist Error Guarantees

- Data analysis = Applying a statistical rule to observed data
- Focus: “frequentist” performance and guarantees of the rule
- Test rules:
  - "reject $H_0$ if $T(X) \geq c’$ $[T(X) =$ test stat, $c’ =$ critical value]
  - "reject $H_0$ if $X \in R”$ $[R \subset S$, critical region]
- $\text{pow}(f) = P(\text{reject } H_0 | f) = \begin{cases} P(\text{type I error at } f), & f \in F_0 \\ 1 - P(\text{type II err at } f) & f \in F_1 \end{cases}$

N-P’s Fixed Level Testing Recipe

- Fix a small level $\alpha \in (0, 1)$
- Decide on a test statistic $T(X)$
- Choose $c$ so that the rule “reject $H_0$ if $T(X) \geq c$" satisfies
  $$\text{size} := \max_{f \in F_0} \text{pow}(f) \leq \alpha$$
- Once $X = x$ is observed, report decision based on how $T(X)$ compares to $c$
  - Must be careful with which hypothesis gets the $H_0$ label.
Comparing Test Rules

- \((T_1, c_1)\) is a better rule than \((T_2, c_2)\) at level \(\alpha\) if
  - either rule has size \(\leq \alpha\)
  - \(\text{pow}(f|T_1, c_1) \geq \text{pow}(f|T_2, c_2)\) for every \(f \in \mathcal{F}_1\).

Optimality of LR: N-P Lemma, MLR

- NP lemma establishes optimality of LR(X) for testing
  \[ H_0 : f = f_0 \text{ vs. } H_1 : f = f_1 \]
  within the simple model \(X \sim f, f \in \mathcal{F} = \{f_0, f_1\}\)
- Limited extension possible to one-sided tests for parametric models with a scalar parameter and possessing the MLR property.

Size Calculation for LR(X): Gaussian Linear Models

- Exact calculations possible for Gaussian linear model
  \[ Y_i = \beta^T x_i + \epsilon_i, \quad i \sim \mathcal{N}(0, 1) \]
  and for testing (for given \(a \in \mathbb{R}^p, b \in \mathbb{R}\))
  \[ H_0 : a^T \beta = b \text{ vs. } H_1 : a^T \beta \neq b \]
  \[ \text{LR}(X) \geq k \iff b \notin a^T \hat{\Theta}_{13} \pm \frac{c}{\sqrt{n}} \iff \begin{cases} \sqrt{n}^\frac{1}{2} a^T \hat{\Theta}_{13} - b \geq c \\ \sqrt{n}^\frac{1}{2} a^T \hat{\Theta}_{13} - b \leq -c \end{cases} \]
  - With a one-one correspondence between \(k\) and \(c\)
  - \(n_0 = 1/(a^T (Z^T Z)^{-1} a)\).
  - Size equals \(2(1 - \Phi_{n_0}(c))\)
  - Size \(\alpha\) test uses \(c = \Phi_{n_0}^{-1}(1 - \alpha/2)\)
- Similar calculations for one-sided hypotheses

LR for IID Non-Gaussian Parametric Models

- Observations \(X_1, X_2, \ldots\)
- Data at stage \(n\): \(X^{(n)} = (X_1, \ldots, X_n)\)
- Model: \(X_i \sim g(x_i|\theta), \theta \in \Theta\)
- \(\hat{\theta}_n = \hat{\theta}_{a2}(X^{(n)}), I_n = I_{22}(X^{(n)})\)

Size Calculation for LR(X): Asymptotic Approximation

- Under regularity assumptions on \(\{g(x|\theta) : \theta \in \Theta\}\)
  - \(\ell_{\text{ml}}(\theta)\) is quadratic around \(\theta_0\) with curvature \(I_n\)
  - For testing \(H_0 : a^T \theta = b \text{ vs. } H_1 : a^T \theta \neq b\)
  \[ \text{LR}(X) \geq k \iff b \notin a^T \hat{\Theta}_{13} \pm \frac{c}{\sqrt{n}^\frac{1}{2} a^T a} \]
  - When truth is \(\theta = \theta_0\) (for any \(\theta_0 \in \Theta\) or at least any \(\theta_0 \in \Theta_0\))
  \[ \sqrt{n}(\hat{\theta}_n - \theta_0) \sim N_d(0, I_n(\theta_0)^{-1}) \text{ and } \frac{1}{n} I_n \overset{P}{\rightarrow} I_r(\theta_0) \]
- Size-\(\alpha\) ML test given by “reject \(H_0\) if \(a^T \hat{\Theta}_{13} - b \geq \Phi^{-1}(1 - \alpha/2)\)”
  - Similar stuff for one-sided hypotheses
  - Could replace \(I_n\) in the test statistic with \(n I_r(\hat{\theta}_n)\) if a closed-form expression for \(I_r(\hat{\theta}_n)\) is available
General IID Data Models

- Observations $X_1, X_2, \ldots$
- Data at stage $n$: $X^{(n)} = (X_1, \ldots, X_n)$
- Model: $X_i \overset{iid}{\sim} g$, $g \in G$ potentially non-parametric
- Focus on $\theta = \theta(g)$, a $d$-dim functional
- To test $H_0: a^T \theta = b$ vs. $a^T \theta \neq b$
- $LR(X)$ may fail if $\dim(G) = \infty$.

A General Recipe of Estimators that are AN

- M-estimators: $\hat{\theta}_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^n m(X_i, \theta)$
- The “criterion function” $m(X_i, \theta)$ satisfies:
  \[ \theta \to E\{m(X_i, \theta) | \theta_0\} \] has unique max at $\theta = \theta_0$
- Z-estimators: $\hat{\theta}_n$ solves $\sum_{i=1}^n \psi(X_i, \theta) = 0$
- The “score function” $\psi(X_i, \theta)$ satisfies:
  \[ \theta \to E\{\psi(X_i, \theta) | \theta_0\} \] has unique zero at $\theta = \theta_0$
- Mostly equivalent with $\psi = \nabla m$, but there are stray cases where $\hat{\theta}_n$ is an M-est but not a Z-est or the opposite

Beyond the Standard Approach: Special Methods

- For certain models, asymptotically normal estimators could be built in other ways, e.g., by using ranks etc.
  - e.g., Mann-Whitney (or Wilcoxon) test for location shift
  - an attractive alternative to the standard t-test
- For certain testing problems, one could use other heuristics to get a test statistic, and could evaluate size via direct Monte Carlo simulation to approximate probability of rejection under the null.
  - e.g., testing CSR for point patterns with $K$-statistic
  - Power calculation is non-trivial
- Somewhat opportunistic constructions, but they all work out under N-P’s fixed level testing recipe.

Standard Approach Toward Test Construction

- Get a consistent estimator sequence $\hat{\theta}_n = \hat{\theta}(X^{(n)})$ such that
  when truth is $g = g_0$ (for any $g_0 \in G$ or at least $G_0$)
  \[ \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N_d(0, \Sigma(g_0)) \]
  where $\theta_0 := \theta(g_0)$ denotes the truth about $\theta$.
- Get hold of a consistent estimator $\hat{\Sigma}_n \overset{p}{\to} \Sigma(g_0)$
- Approximately size-$\alpha$ test (for large $n$):
  \[ \text{reject } H_0 \text{ if } |a^T \hat{\theta}_n - b| \sqrt{\frac{a^T \hat{\Sigma}_n a}{n}} \geq \Phi^{-1}(1 - \frac{\alpha}{2}) \]

Examples

- MLE for regular parametric models
- Least squares linear/non-linear regression with potentially non-Gaussian errors
- Quantile regression
- Method of moments
- Error-in-variable regression
- ...

Asymptotic Comparison of Tests

- Multiple ways to construct test rules.
- How to rank order based on relative performance?
- Exact calculations difficult
  - Asymptotics help
  - Will talk about Pitman’s Relative Efficiency of Tests.
The Pitman Relative Efficiency

- Look at \( \lim_{n \to \infty} \pi_n(\theta_n) \) over a sequence of increasingly difficult alternatives \( \theta_n \to \partial \Theta_0 \) [the bndry of \( \Theta_0 \)].

- **Defn.** Let \( \{ T_{n1} : n \geq 1 \} \) and \( \{ T_{n2} : n \geq 1 \} \) be two test statistics sequences. For any given \( \alpha, \gamma \in (0, 1), \) let \( n_1(\alpha, \gamma, \theta) \) denote the minimum \( n \) needed so that a level \( \alpha \) test based on \( T_{n1} \) has power at least \( \gamma \) at \( \theta \). Let \( n_2(\alpha, \gamma, \theta) \) denote the same for the other sequence. The Pitman relative efficiency of the first sequence against the second, at level \( \alpha \), power \( \gamma \) is defined as:

\[
\lim_{\theta_n \to \partial \Theta_0} \frac{n_2(\alpha, \gamma, \theta_n)}{n_1(\alpha, \gamma, \theta_n)}
\]

provided the limit exists.

The Setup: Sequence of Models

- Infinite sequence of statistical models \( X^{(o)} \sim p_n(\theta) \), \( \theta \in \Theta \).
- All sharing a common parameter \( \theta \).
- Example: IID data case
  - \( X_1, X_2, \ldots \) modeled as \( X_i \mid \theta \sim \mathcal{N}(\mu, \sigma^2) \), \( \theta \in \Theta \)
  - \( X^{(o)} = (X_1, \ldots, X_n) \)
  - \( p_n(s^{(o)}) = g(s_1|\theta) \times \cdots \times g(s_n|\theta) \)

Comparison: Basics

- Test sequence \( \{(T_n, c_n) : n \geq 1\} \) is asymptotically size \( \alpha \) if,

\[
\lim_{n \to \infty} \sup_{\theta \in \Theta_0} P(T_n \geq c_n|\theta) = \alpha.
\]

- Consider two sequences \( (T_{n1}, c_{n1}), (T_{n2}, c_{n2}) \), \( n = 1, 2, \ldots \)
  - Corresponding power function sequences:
    - \( \pi_{n1}(\theta) = P(T_{n1} \geq c_{n1}|\theta) \)
    - \( \pi_{n2}(\theta) = P(T_{n2} \geq c_{n2}|\theta) \)
  - Suppose both sequences are asymptotically size \( \alpha \).
  - Declare seq 1 asymptotically better than seq 2 if

\[
\lim_{n \to \infty} \pi_{n1}(\theta) \geq \lim_{n \to \infty} \pi_{n2}(\theta) \text{ for every } \theta \in \Theta_0.
\]

The Setup: Test Rule Sequences

- \( H_0 : \theta \in \Theta_0 \) vs. \( H_1 : \theta \in \Theta_1^\gamma \)
- Sequences of test procedures:
  - “reject \( H_0 \) if \( T_n \geq c_n \)”, \( n = 1, 2, \ldots \)
  - where \( T_n \) is based on data \( X^{(o)} \).
- Typically \( T_n \) given by the same recipe applied to all \( n \), e.g.,
  - ML, or
  - Z-estimation associated with a certain score function \( \psi \),
  - A certain type of rank-based tests, etc.

Not a Useful Comparison

- For almost for all reasonable asymptotically size \( \alpha \) test sequences, the limiting power equals 1 at any \( \theta \in \Theta_0^\gamma \).
- Example:
  - IID data model: \( X_i \sim \mathcal{N}(\mu_i, 1), i = 1, 2, \ldots \)
  - An asymptotically size 5% test for \( H_0 : \mu = 0 \) vs. \( H_1 : \mu \geq 0 \) is
    - "reject \( H_0 \) if \( \sqrt{n}X^{(o)} \geq 1.64 \)."
  - At any \( \mu > 0 \), \( \pi_\alpha(\mu) = 1 - \Phi(1.64 - \sqrt{n}\mu) \to 1 \) as \( n \to \infty \).

The Pitman Relative Efficiency

- Look at \( \lim_{n \to \infty} \pi_n(\theta_n) \) over a sequence of increasingly difficult alternatives \( \theta_n \to \partial \Theta_0 \) [the bndry of \( \Theta_0 \)].

- **Defn.** Let \( \{ T_{n1} : n \geq 1 \} \) and \( \{ T_{n2} : n \geq 1 \} \) be two test statistics sequences. For any given \( \alpha, \gamma \in (0, 1), \) let \( n_1(\alpha, \gamma, \theta) \) denote the minimum \( n \) needed so that a level \( \alpha \) test based on \( T_{n1} \) has power at least \( \gamma \) at \( \theta \). Let \( n_2(\alpha, \gamma, \theta) \) denote the same for the other sequence. The Pitman relative efficiency of the first sequence against the second, at level \( \alpha \), power \( \gamma \) is defined as:

\[
\lim_{\theta_n \to \partial \Theta_0} \frac{n_2(\alpha, \gamma, \theta_n)}{n_1(\alpha, \gamma, \theta_n)}
\]

provided the limit exists.

Doing the Math

- The limit may depend on \( \alpha, \gamma \) and the collapsing sequence \( \theta_n \)
- But it DOES NOT under some simplifying assumptions:
  - \( \Theta_0 = \{ \theta_0 \} \), a point null
  - \( \theta_n = \theta_0 + \nu h \)
    - \( h \) is a fixed unit vector
    - \( \nu \downarrow 0 \)
    - i.e., \( \theta_n \) approaches \( \theta_0 \) along a fixed ray from only one of the two directions.
  - Only “regular” test statistic sequences!
Regularity

- Test statistic sequence $T_n, n = 1, 2, \ldots$ is “regular” at $\theta_0$ if
  $\sqrt{n}(T_n - \mu(\theta_0 + h_n)) \xrightarrow{d} N(0, 1)$ under $p_{n,\theta_0 + h_n}$
  for some functions $\mu(\theta)$ and $\sigma(\theta)$ and any $h_n$.

- “Regularity” is stronger than requiring $\sqrt{n}(T_n - \mu(\theta)) \xrightarrow{d} N(0, 1)$
  at every $\theta \in \Theta$.

What Regularity Brings to the Table

- Sequence of increasingly difficult alternatives:
  $\theta_n = \theta_0 + h_n, n \geq 1$,
  for any fixed $h \neq 0$.

- If $(T_n, c_n, n \geq 1)$ is asymptotically size $\alpha$ and $T_n$ is regular,
  then,
  $\lim_{n \to \infty} \pi_n(\theta) = 1 - \Phi \left( \frac{c_n - h^T \mu(\theta)}{\sigma(\theta)} \right)$.

- Quantity $\mu(\theta_0)/\sigma(\theta_0)$ is the slope of $T_n$ at $\theta_0$.

Regularity and Pitman Relative Efficiency

Theorem
Consider statistical models $(p_{n,\theta} : \theta \geq 0), n = 1, 2, \ldots$ such that
$\|p_{n,\theta} - p_{n,0}\| \to 0$ as $\theta \to 0$ for every fixed $n$. Let $T_{n,1}$ and $T_{n,2}$ be
two sequences of test statistics that satisfy the regularity condition with $\mu_i(0) > 0$ and $\sigma_i(0) > 0, i = 1, 2$. Then the Pitman relative efficiency of the first sequence to the second equals
$$
\left( \frac{\mu_1(0)}{\sigma_1(0)} \right)^2
$$
for every sequence $\theta_n \downarrow 0$ independently of $\alpha$ and $\gamma$.

Slope Calculation

- Direct calculations for normal models.
- Advanced asymptotic calculations for other smooth models:
  - Le Cam’s third lemma.
  - Local asymptotic normality

Le Cam’s Third Lemma

Theorem (Le Cam’s third lemma)
For every $n = 1, 2, \ldots$, let $p_{n,\theta}, q_n$ be two pdfs defined on a common
space $X_n$ and let $T_n : X_n \to \mathbb{R}^k$ be a map such that,
$$
T_n(X(n)) \xrightarrow{d} N_{k+1} \left( \begin{pmatrix} \mu \\ \Sigma \end{pmatrix}, \tau \sigma^2 \right)
$$
under $X(n) \sim p_{n,\theta}$.

Then $T_n(X_n) \xrightarrow{d} N_k(\mu + \tau, \Sigma)$ under $X(n) \sim q_n$.

IID Models and LAN

- IID observations model
  - $X_1, X_2, \ldots \sim g(x|\theta), \theta \in \Theta$
  - $p_{n,\theta}(x^{(n)}) = g(x|\theta) \cdots g(x_n|\theta)$

- Cramér conditions at $\theta_0$ implies LAN
  $$
  \log \frac{p_{n,\theta_0 + h_n}}{p_{n,\theta_0}} \sim \frac{1}{2} h^T I_f(\theta_0) - \frac{1}{2} h^T I_f(\theta_0) h + \text{small}
  \xrightarrow{d} N \left( -\frac{1}{2} \sigma^2, \sigma^2 \right)
  $$
  under $p_{n,\theta_0}$

with $\sigma^2 = h^T I_f(\theta_0) h$.

- Invoke Le Cam’s third lemma:
  - Establish a bivariate CLT between $T_n$ and $h^T \frac{1}{\sqrt{n}} T_n(\theta_0)$
  - Calculate the limiting covariance vector $\tau$.
Example: T-Test vs. Mann-Whitney Test

- P’s relative efficiency of M-W vs. t-test: $12\sigma^2_n \left\{ \int g(w)^2 dw \right\}^2$.

<table>
<thead>
<tr>
<th>g(w)</th>
<th>Relative Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>$\pi^2/9 = 110%$</td>
</tr>
<tr>
<td>Normal</td>
<td>$3/\pi = 95%$</td>
</tr>
<tr>
<td>Uniform</td>
<td>100%</td>
</tr>
<tr>
<td>$t(3)$</td>
<td>124%</td>
</tr>
<tr>
<td>$t(5)$</td>
<td>190%</td>
</tr>
<tr>
<td>$\frac{1}{2}(1 - w^2)I(</td>
<td>w</td>
</tr>
</tbody>
</table>

- M-W never too worse for any parametric sub-model. Could be much superior for some.

Parametric Models: Optimality of ML tests

- Under LAN conditions there is non-trivial bound on $\lim_{n \to \infty} \pi_n(\theta_0 + h/\sqrt{n})$.
- MLE based test statistic sequence attains this bound.
- How to think about it:
  - For every given pdf $g$ on $\mathbb{R}$, the M-W rule for $H_0: \theta \leq 0$ vs $H_1: \theta > 0$ will be (asymptotically) dominated by the ML test based on the parametric sub-model $W_i \overset{iid}{\sim} g(\cdot - \theta), V_j \overset{iid}{\sim} g, \theta \in \mathbb{R}$.
  - But we saw, on the whole M-W is supremely competitive.