MATH 408 Midterm 2 Review

Qilin Ye

November 4, 2021

Contents

Disclaimer: I typed these within one evening, some of which directly copied from HW. Consequently the following content are extremely prone to typos. If something doesn't make sense, it's probably because I made a mistake. But good luck reviewing.

4 Estimation of Parameters

4.1 Sufficient Statistics

Definition: (4.9) Sufficient Statistics

Suppose $X = (X_1, \dots, X_n)$ and X_i has distribution $f_\theta \in \{f_\theta : \theta \in \Theta\}$. A statistic Y is **sufficient** for θ if for all y, the conditional distribution of $X = (X_1, \dots, X_n)$ given $Y = y$ does not depend on θ, i.e., the following quantity does not depend on θ:

$$\mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n) \mid Y = y).$$

Theorem: (4.12) Factorization Theorem

Y is sufficient for θ if and only if $f_\theta(x) = g_\theta(y)h(x)$ for some function $g_\theta(y)$ depending only on y and $h(x)$ depending only on x.

Example: HW3 P4. Let $\theta \in \mathbb{R}$ be an unknown parameter. Consider the density

$$f_\theta(x) := \begin{cases} \exp(-(x-\theta)) & x \geqslant \theta \\ 0 & x < \theta. \end{cases}$$

Suppose $X_1, \dots, X_n$ is a random sample of size n such that each X_i has density f_θ. Show that $X_{(1)} = \min\limits_{1 \leqslant i \leqslant n} X_i$ is a sufficient statistic for θ.

Proof. We first write $f_\theta(x) = \exp(-(x-\theta))\chi_{[\theta,\infty)}(x_i)$. Since X_i's are i.i.d., for $\vec{x} := (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$f_\theta(\vec{x}) = \prod_{i=1}^n f_\theta(x_i) = \exp(n\theta - \sum_{i=1}^n x_i) \prod_{i=1}^n \chi_{[\theta,\infty)}(x_i).$$

Note that $f_\theta(\vec{x}) \neq 0$ if and only if $x_i \geqslant \theta$ for all i, i.e., $X_{(1)} \geqslant \theta$. That is,

$$f_\theta(\tilde{x}) = \exp(n\theta - \sum_{i=1}^n x_i)\chi_{[\theta,\infty)}(X_{(1)}).$$

Therefore, $f_\theta(\tilde{x})$ admits a factorization

$$f_\theta(\tilde{x}) = \underbrace{\exp(n\theta)\chi_{[\theta,\infty)}(X_{(1)})}_{g_\theta(X_{(1)})} \cdot \underbrace{\exp(-\sum_{i=1}^n x_i)}_{h(x)},$$

which by the factorization theorem shows $X_{(1)}$ is sufficient. □

4.2 Evaluating Estimators

4.2.1 Conditional Expectation

Definition: (4.14) Conditional Expectation

For x, we define $g(y) := \mathbb{E}(X \mid Y = y)$. From this we obtain a random variable $g(Y)$, called the **conditional expectation** of X given Y, written $\mathbb{E}(X \mid Y)$.

Example: (4.14(i)). Let X, Y be random variables such that (X, Y) is uniformly distributed on the triangle

$$\{(x, y) \in \mathbb{R}^2 : x \geqslant 0, y \geqslant 0, x + y \leqslant 1\}.$$

Show that $\mathbb{E}(X \mid Y) = (1 - Y)/2$.

Proof. For any given $y \in [0, 1]$, $X \mid Y = y$ is uniformly distributed on $[0, 1 - y]$ so the expectation is $(1 - y)/2$. Then by definition $\mathbb{E}(X \mid Y) = (1 - Y)/2$. □

4.2.2 Variance of Estimators; UMVU

Definition: (4.15) UMVU

Let $X_1, \ldots, X_n$ be i.i.d. from a distribution in $\{f_\theta \in \theta \in \Theta\}$. Suppose we want to estimate $g(\theta)$. We say Y is the **uniformly minimum variance unbiased** estimator for $g(\theta)$ if:

(1) Y is unbiased (obviously), i.e., $\mathbb{E}_\theta(Y) = g(\theta)$, and

(2) if Z is also an estimator for $g(\theta)$ then $\mathrm{var}_\theta(Y) \leqslant \mathrm{var}_\theta(Z)$ (*hence the word "minimum"*) for all $\theta \in \Theta$ (*hence the word "uniformly"*).

Theorem: (4.17) Rao-Blackwell

Suppose $X_1, \ldots, X_n$ are i.i.d. with $f_\theta \in \{f_\theta : \theta \in \Theta\}$ and suppose Z is sufficient for θ. Let Y be unbiased for θ with $\mathrm{var}_\theta(Y) < \infty$ and define $W := \mathbb{E}_\theta(Y \mid Z)$. Then:

(1) *the total expectation theroem says W is also unbiased* (see HW3 P2), and

(2) $\mathrm{var}_\theta(W) \leqslant \mathrm{var}_\theta(Y)$, with = if and only if $W = Y$.

Example: (4.23). Let $X_1, \ldots, X_n$ be i.i.d. with unknown $\mu \in \mathbb{R}$ and known $\sigma^2 < \infty$. Suppose that $Y := \sum_{i=1}^{n} X_i$ is sufficient for μ. Clearly X_1 itself is an unbiased estimator as $\mathbb{E}X_1 = \mu$. Rao-Blackwell says

$$\mathbb{E}(X_1 \mid \sum_{i=1}^{n} X_i)$$

is unbiased with smaller variance. Since $X_1, \ldots, X_n$ are i.i.d., by HW3 P2

$$nW = \sum_{j=1}^{n} \mathbb{E}(X_j \mid \sum_{X=1}^{n} X_i) = \mathbb{E}(\sum_{j=1}^{n} X_j \mid \sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} X_i$$

so W upgrades X_1 to the sample mean, whose variance shrinks from σ^2 to σ^2/n.

4.3 Efficiency of an Estimator

4.3.1 Fisher Information

Definition: (4.24) Fisher Information

Let $\{f_\theta : \theta \in \Theta \subset \mathbb{R}\}$ be a family of distributions. Define the **Fisher information** of the family to be

$$I_X(\theta) := \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X)\right)^2\right].$$

Equivalent definitions include

$$I_X(\theta) = \mathrm{var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X)\right) = -\mathbb{E}_\theta\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f_\theta(X)\right).$$

Example: (4.25). Let $\sigma > 0$ and let $f_\theta(x)$ be the PDF of a Gaussian with mean θ and variance σ^2. We have

$$\log f_\theta(x) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x-\theta)^2}{2\sigma^2}$$

so

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X) = \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{-(X-\theta)^2}{2\sigma^2},$$

and so

$$I(\theta) = \mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta} \frac{-(X-\theta)^2}{2\sigma^2}\right)^2 = \mathbb{E}_\theta\left(\frac{X-\theta}{\sigma^2}\right)^2 = \frac{1}{\sigma^4}\mathrm{var}(X-\theta) = \frac{1}{\sigma^2}.$$

Alternatively, using the negative expectation of second derivative, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f_\theta(X) = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left(-\frac{(X-\theta)^2}{2\sigma^2}\right) = -\frac{1}{\sigma^2}$$

so $I(\theta) = -\mathbb{E}(-1/\sigma^2) = 1/\sigma^2$ again.

Proposition: (4.20)

If X, Y are independent then $I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$. In particular if $X_1, \ldots, X_n$ are i.i.d. then

$$I_{(X_1,\ldots,X_n)}(\theta) = nI_{X_1}(\theta).$$

4.3.2 Cramér-Rao and its relation to UMVU

Theorem: (4.28) Cramér-Rao / Information Inequality

Let $X : \Omega \to \mathbb{R}^n$ be a random variable with distribution from $\{f_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$. Let $Y := t(X)$ be a statistic. For $\theta \in \Theta$, define $g(\theta) := \mathbb{E}_\theta Y$. Then

$$\mathrm{var}_\theta(Y) \geqslant \frac{|g'(\theta)|^2}{I_X(\theta)} \qquad \text{for all } \theta \in \Theta.$$

In particular if Y is *unbiased* then $g(\theta) = \theta$ and $g'(\theta) = 1$, so

$$\mathrm{var}_\theta(Y) \geqslant \frac{1}{I_X(\theta)} \qquad \text{for all } \theta \in \Theta.$$

In both cases, "=" happens only when $\dfrac{\mathrm{d}/\mathrm{d}\theta(\log f_\theta(X))}{Y - \mathbb{E}_\theta Y} \in \mathbb{R}$ for some $\theta \in \Theta$.

This theorem provides a lower bound on the variance of unbiased estimators of θ — in general, we cannot get estimators with arbitrarily small variance.

Problem: HW4 P3

Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli distribution with parameter $\theta \in (0,1)$ unknown. In class we showed that $\sum_{i=1}^n X_i$ is sufficient for θ and that

$$\mathbb{E}_\theta(X_1 \mid \sum_{i=1}^n X_i) = \frac{1}{n}\sum_{i=1}^n X_i.$$

That is, Rao-Blackwell suggests that the sample mean has small variance among unbiased estimators for θ.

(1) Compute the Fisher information $I_{X_1}(\theta)$.

(2) Compute the Fisher information $I_{(X_1,\ldots,X_n)}(\theta)$.

(3) Show that $\mathrm{var}(n^{-1}\sum_{i=1}^n X_i) = \theta(1-\theta)/n$.

(4) Does the sample mean $n^{-1}\sum_{i=1}^n X_i$ achieve equality in Cramér-Rao? If so, then $n^{-1}\sum_{i=1}^n X_i$ is UMVU.

Solution. (1) For $x \in \{0,1\}$, the PMF of X is given by $\mathbb{P}(X = x) = \theta^x(1-\theta)^{1-x}$. Since

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\log(\theta^x(1-\theta)^{1-x}) &= \frac{1}{\theta^x(1-\theta)^{1-x}}(x\theta^{x-1}(1-\theta)^{1-x} - \theta^x(1-x)(1-\theta)^{-x}) \\ &= \frac{x\theta^{x-1}(1-\theta)^{1-x}}{\theta^x(1-\theta)^{1-x}} - \frac{\theta^x(1-x)(1-\theta)^{-x}}{\theta^x(1-\theta)^{1-x}} \\ &= \frac{x}{\theta} - \frac{1-x}{1-\theta},\end{aligned}$$

we see that

$$\begin{aligned}
I_{X_1}(\theta) &= \mathbb{E}\left[\frac{X}{\theta} - \frac{1-X}{1-\theta}\right]^2 \\
&= \mathbb{E}\left[\frac{X^2}{\theta^2}\right] - 2\mathbb{E}\left[\frac{X(1-X)}{\theta(1-\theta)}\right] + \mathbb{E}\left[\frac{(1-X)^2}{(1-\theta)^2}\right] \\
&= \mathbb{E}\left[\frac{X^2}{\theta^2}\right] - 2\mathbb{E}\left[\frac{X-X^2}{\theta(1-\theta)}\right] + \mathbb{E}\left[\frac{X^2-2X+1}{(1-\theta)^2}\right] \\
&\overset{\Delta}{=} \frac{\theta}{\theta^2} - 2\cdot\frac{\theta-\theta}{\theta(1-\theta)} + \frac{\theta-2\theta+1}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.
\end{aligned}$$

(The Δ step is because $\mathbb{E}X = \theta$ and $\mathbb{E}X^2 = \theta$, both of which follow from simple calculation.)

Alternatively, we can compute the second derivative (this is in general easier):

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(\theta^x(1-x)^{1-x}) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{x}{\theta} - \frac{1-x}{1-\theta}\right] \\
&= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}
\end{aligned}$$

whose expected value is

$$-\frac{1}{\theta^2}\mathbb{E}X - \frac{\mathbb{E}(1-X)}{(1-\theta)^2} = -\frac{1}{\theta} - \frac{1}{1-\theta} = -\frac{1}{\theta(1-\theta)}.$$

Taking the negative of above, we also obtain $I_{X_1}(\theta) = 1/(\theta(1-\theta))$.

(2) Since $X_1, \ldots, X_n$ are i.i.d. (in particular independent),

$$I_{(X_1,\ldots,X_n)}(\theta) = \sum_{i=1}^n I_{X_i}(\theta) = \frac{n}{\theta(1-\theta)}.$$

(3) This also follows from definition: since $X_1, \ldots, X_n$ are i.i.d. (in particular independent),

$$\mathrm{var}(n^{-1}\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}(X_i) = \frac{\theta(1-\theta)}{n}.$$

(4) Yes.

4.4 Maximum Likelihood Estimator (MLE)

Definition: (4.32) Maximum Likelihood Estimator (MLE)

The **maximum likelihood estimator** Y is the estimator maximizing the likelihood function, given that it exists (it might not a priori). That is, the MLE of θ is the θ estimating

$$\ell(\theta) := \prod_{i=1}^n f_\theta(x_i).$$

Sometimes it is more convenient to maximize $\log \ell(\theta)$ *instead.*

Example: (4.34). MLE need not exist and even if it does, it need not to be unique.

For example, let $X_1, .., X_n$ be i.i.d. from f_θ with $f_\theta(x_i) = \chi_{[\theta,\theta+1)}(x_i)$, i.e., X is uniform on $[\theta, \theta+1]$. Then

the joint PDF is $\prod_{i=1}^{n} \chi_{x_i \in [\theta, \theta+1]}$.
Suppose for example that $x_1 = \ldots = x_n = 0$. Then $\ell(\theta) = \chi_{0 \in [\theta, \theta+1]} = \chi_{\theta \in [-1,0]}$, so any $\theta \in [-1, 0]$ is a MLE in this case. Uncountably many!

Example: HW4 P6. Let $X_1, \ldots, X_n$ be a random sample of size n from a Poisson distribution with unknown parameter $\lambda > 0$ (so $\mathbb{P}(X_1 = k) = e^{-\lambda}\lambda^k / k!$ for $k \in \mathbb{N}$).

(1) Find an MLE for λ.

(2) Find an MLE for $e^{-\lambda}$.

(3) How do your results compare to the previous homework, where we found two different estimators for $e^{-\lambda}$ (one from the method of moments, and the other by applying the Rao-Blackwell Theorem)?

Solution. (1) The likelihood function $\ell(\lambda)$ is given by

$$\ell(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{(x_i)!}.$$

Taking log gives

$$\log \ell(\lambda) = -n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log(x_i!).$$

Hence

$$(\log \ell(\lambda))' = -n + \lambda^{-1} \sum_{i=1}^{n} x_i.$$

Setting this quantity to 0, we see that the critical point is the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$. Verify that this is a maximum:

$$(\log \ell(\overline{x}))'' = -(\overline{x})^{-2} \sum_{i=1}^{n} x_i = -n^2 / \sum_{i=1}^{n} x_i < 0.$$

Hence $\frac{1}{n}\sum_{i=1}^{n} X_i$ is an MLE estimator for λ.

(2) By functional equivariance of MLE (proposition 4.45), $\exp\left(-\frac{1}{n}\sum_{i=1}^{n} X_i\right)$ is an MLE for $e^{-\lambda}$.

(3) It does agree with the estimator obtained from MoM (see HW4 P1).

4.5 Convex Functions

Definition: Convex Functions

Let $\varphi : \mathbb{R} \to \mathbb{R}$. We say the function is **convex** if for all $x, y \in \mathbb{R}$ and $\lambda \in [0,1]$,

$$\varphi(\lambda x + (1-\lambda)y) \leqslant \lambda\varphi(x) + (1-\lambda)\varphi(y).$$

We say the function is *strictly convex* if $\leqslant$ can be replaced by $<$ above.
A generalization of higher-dimensional convex functions exist. Simply replace $x, y \in \mathbb{R}$ by $x, y \in \mathbb{R}^n$.

Example: (Quiz 4 P2). Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that if $x \in \mathbb{R}^n$ is a local minimum of f then it's a global minimum. In addition, if f is strictly convex, show that there exists at most one global minimum of f.

Proof. Let x be a local minimum. That is, there exists $r > 0$ such that x^* is the minimum on the r-neighborhood of x, i.e.,

$$f(k) \geqslant f(x) \qquad \text{for all } k \text{ with } \|k - y\| < r. \tag{1}$$

Suppose for contradiction that x is not a global minimum, so for some $y \in \mathbb{R}^n$ we have $f(y) < f(x)$. Consider the line segment connecting x and y, and consider $\epsilon > 0$ so small that $\epsilon y + (1-\epsilon)x$ is within r from x, i.e.,

$$\|x - (\epsilon y + (1-\epsilon)x)\| < r.$$

Call this point z. Then by assumption (1) $f(z) \geqslant f(x)$. However, by convexity,

$$f(z) \leqslant \epsilon f(y) + (1-\epsilon)f(x) < \epsilon f(x) + (1-\epsilon)f(x) = f(x),$$

contradiction. Hence x must be a global minimum.

In the case of a strictly convex function, if $x \neq y$ and x, y are both global minima, then by strict convexity, $f((x+y)/2) < f(x)/2 + f(y)/2 = f(x) = f(y)$, contradicting the assumption that x, y are minima. Hence global minimum is unique. □

Example: Quiz 4 P3. Let $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ be strictly convex and define $g : \mathbb{R}^n \to \mathbb{R}$ by

$$g(x_1, \ldots, x_n) := \sum_{i=1}^{n} f(x_i).$$

Show that g is strictly convex.

Proof. Let $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and $\lambda \in (0,1)$ be given. Then

$$\begin{aligned}
g(\lambda x + (1-\lambda)y) &= g(\lambda x_1 + (1-\lambda)y_1, \ldots, \lambda x_n + (1-\lambda)y_n) \\
&= \sum_{i=1}^{n} f(\lambda x_i + (1-\lambda)y_i) \\
&< \sum_{i=1}^{n} [\lambda f(x_i) + (1-\lambda)f(y_i)] && \text{(by strict convexity of } f) \\
&= \sum_{i=1}^{n} \lambda f(x_i) + \sum_{i=1}^{n} (1-\lambda)f(y_i) = \lambda g(x) + (1-\lambda)g(y).
\end{aligned}$$

□

Example: Quiz 4 P5. Let $f : \mathbb{R}^n \to \mathbb{R}$. Suppose for any fixed $1 \leqslant i \leqslant n$ and any fixed $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, the function

$$x_i \mapsto f(x_1, \ldots, x_n)$$

is strictly convex. Prove that f has at most one global minimum.

Proof. Suppose there exist two distinct global minima $x, y \in \mathbb{R}^n$. It follows that $x_i \neq y_i$ for some i. But then for the i^{th} component, if we fix $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, the map

$$x \mapsto f(x_1, \ldots, x_n)$$

is a strictly convex function with two distinct global minima. This is impossible by problem 2(b). Hence f has at most one global minimum. $\square$

5 Hypothesis Testing

5.0 Preliminaries

Definition: (5.1) Null Hypothesis & Alternative Hypothesis

Let $\{f_\theta : \theta \in \Theta\}$ be a family of distributions. Let $\Theta_0 \subset \Theta$. A **null hypothesis** H_0 is an event of form

$$\{\theta \in \Theta_0\}.$$

Define $\Theta_1 := \Theta_0^c$ so $\Theta = \Theta_0 \sqcup \Theta_1$ (disjoint union). The **alternative hypothesis** H_1 is the event $\{\theta \in \Theta_1\}$.

Definition: (5.4) Rejection Region, Power Function, & Significance Level

Let H_0 be a null hypothesis. A **hypothesis test** of H_0 vs. H_1 is specified by a subset $C \subset \mathbb{R}^n$. The set C is called the **critical region** or the **rejection region**.

(1) If $X \notin C$, we accept H_0.

(2) If $X \in C$, we reject H_0 and assert that H_1 is true.

The complement $C^c \subset \mathbb{R}^n$ is called the **acceptance region**. The performance of the test is quantified by the **power function** $\beta : \Theta \to [0,1]$ by

$$\beta(\theta) := \mathbb{P}_\theta(X \in C) = \mathbb{P}_\theta(\text{reject } H_0) = 1 - \mathbb{P}_\theta(X \notin C).$$

The **significance level** α is defined as

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta).$$

Remark. Ideally, we want to find a "perfect" test in the sense that $\beta(\theta) = 0$ for all $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for all $\theta \in \Theta_1$, i.e., if the null hypothesis is true then we accept it with probability 1 and if it is false, we accept it with probability 0. However, this might not always happen.

Definition: (5.5) Type II Error: *false negative*

A **type II error** for a hypothesis test occurs when $X \notin C$ with positive probability but H_0 is actually false. That is, $\beta(\theta) < 1$ for some $\theta \in \Theta_1$. In other words, H_0 is accepted to be true whereas it is actually false. The quantity $1 - \beta(\theta)$ is the probability of occurrence of a type II error for $\theta \in \Theta_1$.

Definition: (5.6) Type I Error: *false positive*

A **type I error** for a hypothesis test occurs when $X \in C$ with positive probability but H_1 is actually false. That is, $\beta(\theta) > 0$ for some $\theta \in \Theta_0$. In other words, H_0 is rejected whereas it is actually true. The value of $\beta(\theta)$ is the probability of occurrence of a type I error for $\theta \in \Theta_0$.

Definition: (5.8) Uniformly Most Powerful Test (UMP)

Let $\Theta_0 \subset \Theta$ and denote $\Theta_1 := \Theta_0^c$. Let H_0 be the hypothesis $\theta \in \Theta_0$ and H_1 be $\{\theta \in \Theta_1\}$. Let $\mathcal{T}$ be a family of hypothesis tests. A hypothesis test in $\mathcal{T}$ with power function $\beta(\theta)$ is called the **uniformly most powerful class $\mathcal{T}$ test** if $\beta(\theta) \geqslant \beta'(\theta)$ for all $\theta \in \Theta_1$ for every $\beta'(\theta)$ corresponding to a hypothesis test in $\mathcal{T}$. *In other words, the UMP test's power function is the largest on all of Θ_1.*

5.1 Neyman-Perason Testing

Lemma: (5.9) Neyman-Pearson

Let $\Theta = \{\theta_0, \theta_1\}$, $\Theta_0 := \{\theta_0\}$, and $\Theta_1 := \{\theta_1\}$. Let H_0 be the hypothesis $\{\theta = \theta_0\}$ and H_1 be $\{\theta = \theta_1\}$. Let $\{f_{\theta_0}, f_{\theta_1}\}$ be two multivariable PDFs or PMFs. Fix $k \geqslant 0$. Define the **likelihood ratio test** with rejection region C in the following way:

$$C := \{x \in \mathbb{R}^n : f_{\theta_1}(x) > k f_{\theta_0}(x)\}. \tag{1}$$

As usual, define

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = \mathbb{P}_{\theta_0}(X \in C). \tag{2}$$

Let $\mathcal{T}$ be the class of hypothesis tests with significance levels $\leqslant \alpha$. Then:

$$\text{The test is UMP level } \alpha \iff \text{ it is the likelihood ratio test with significance level } \alpha.$$

To put formally:

- *(Sufficiency) Any hypothesis test satisfying (1) is a UMP class $\mathcal{T}$ test.*
- *(Necessity) If there exists a hypothesis test satisfying (1) and (2) with $k > 0$, then any UMP class $\mathcal{T}$ test has significance level equal to α, and any UMP class $\mathcal{T}$ test satisfies (1), except possibly on a null set D with $\mathbb{P}_{\theta_1}(X \in D) = 0$.*

Example. Let X be a Poisson distribution with parameter λ, i.e., the distribution is given by $e^{-\lambda}\lambda^x/x!$. Find a UMP level α test for $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda = \lambda_1$, where λ_1, λ_0 are given and $\lambda_1 > \lambda_0$.

Solution. By Neyman-Perason, it suffices to find the likelihood ratio test with ratio k_α whose significance level is precisely α. The rejection region for k_α is characterized by

$$\text{Reject } H_0 \iff \frac{f_{\lambda_1}(x)}{f_{\lambda_0}(x)} \geqslant k_\alpha \iff \frac{e^{-\lambda_1}\lambda_1^x/x!}{e^{-\lambda_2}\lambda_2^x/x!} \geqslant k_\alpha$$
$$\iff e^{-\lambda_1+\lambda_0}\left(\frac{\lambda_1}{\lambda_0}\right)^x \geqslant k_\alpha.$$

The significance level of this test is

$$\begin{aligned}\sup_{\lambda\in\{\lambda_0\}} \mathbb{P}(\text{reject } H_0) &= \mathbb{P}_{\lambda_0}\left(e^{-\lambda_1+\lambda_0}(\lambda_1/\lambda_0)^x \geqslant k_\alpha\right) \\ &= \mathbb{P}_{\lambda_0}(\lambda_1+\lambda_0+x\log(x_1/x_0) \geqslant \log k_\alpha) \\ &= \mathbb{P}_{\lambda_0}\left(x \geqslant \frac{\log k_\alpha + \lambda_1 - \lambda_0}{\log(\lambda_1/\lambda_0)}\right) =: \mathbb{P}_\lambda(x \geqslant c_\alpha) \\ &= \sum_{x=\lceil c_\alpha\rceil}^{\infty} \mathbb{P}_{\lambda_0}(X=x) = \sum_{x=\lceil c_\alpha\rceil}^{\infty} \frac{e^{-\lambda_0}\lambda_0^x}{x!}.\end{aligned}$$

Whatever c_α (and thus k_α) makes the series equal α will give us the explicit rejection region, as long as it exists. (In the case of a Poisson, not every α can be a significance level of a likelihood ratio test.) The UMP is given by

$$\text{Reject } H_0 \iff f_{\lambda_1}(X)/f_{\lambda_0}(X) \geqslant k_\alpha \iff X \geqslant c_\alpha.$$

5.2 Confidence Intervals

Definition: (5.14) Confidence Interval, Confidence Region

Let $X : \Omega \to \mathbb{R}^n$ be a random variable with distribution $f_\theta \in \{f_\theta : \theta \in \Theta\}$. Let $g : \Theta \to \mathbb{R}$. Let $u, v : \mathbb{R}^n \to \mathbb{R}$ be such that $u(x) \leqslant g(x)$ for all $x \in \mathbb{R}^n$. A **$100(1-\alpha)$% confidence interval** for a parameter $g(\theta)$ is a random variable of form $[u(X), v(X)]$ satisfying

$$\mathbb{P}_\theta(g(\theta)) \in [u(X), v(X)] \geqslant 1-\alpha \qquad \text{for all } \theta \in \Theta.$$

More generally, in higher dimensions, if $c : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ (power set), then a ***$100(1-\alpha)$% confidence region*** *for $g(\theta)$ is a random set $c(X)$ satisfying*

$$\mathbb{P}_\theta(g(\theta) \in c(X)) \geqslant 1-\alpha \qquad \textit{for all } \theta \in \Theta.$$

Example. Let X_i be i.i.d. with distribution $\mathcal{N}(\mu, \sigma^2)$ with both parameters unknown. Find a $100(1-\alpha)$% confidence interval for μ.

Solution. If S is the sample standard deviation, then $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ has a student's t-distribution with $n-1$ degrees of freedom. Hence, we want to find $a < b$ such that

$$1-\alpha \leqslant \mathbb{P}\left(a \leqslant \frac{\overline{X}-\mu}{S/\sqrt{n}} \leqslant b\right).$$

One way is to find a, b such that

$$\mathbb{P}(t_{n-1} \leqslant \alpha) = \frac{\alpha}{2} \qquad \text{and } \mathbb{P}(t_{n-1} \geqslant b) = \frac{\alpha}{2}.$$

by symmetry of t-distribution, these are given by $a = -t_{n-1}(\alpha/2)$ and $b = t_{n-1}(\alpha/2)$. Algebra gives

$$\begin{aligned} 1-\alpha &\leqslant \mathbb{P}\left(-t_{n-1}(\alpha/2) \leqslant \frac{\overline{X}-\mu}{S/\sqrt{n}} \leqslant t_{n-1}(\alpha/2)\right) \\ &= \mathbb{P}\left(-\frac{t_{n-1}(\alpha/2)S}{\sqrt{n}} \leqslant \overline{X}-\mu \leqslant \frac{t_{n-1}(\alpha/2)S}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\mu \in \left[\overline{X} - \frac{t_{n-1}(\alpha/2)S}{\sqrt{n}}, \overline{X} + \frac{t_{n-1}(\alpha/2)S}{\sqrt{n}}\right]\right). \end{aligned}$$

The last interval is the $100(1-\alpha)\%$ confidence interval we are looking for.

Example: (MT1 P5). Consider a population of $30,000$ people, where half of them are given a vaccine for a disease. Suppose all $30,000$ people are exposed to a virus causing the disease. We observe that 90 of the unvaccinated people catch the disease, while 5 of the vaccinated people catch the disease. Consider the following statement:

If we have a population of $30,000$ *people exposed to the virus, with half of them vaccinated and half not. Then the number of infections of vaccinated people, divided by the number of infections of unvaccinated people, is less than* $15/100$.

Is the statement true with $\geqslant 90\%$ certainty?

Solution. Let U, V denote the number of infected people who are unvaccinated and vaccinated, respectively. Let $S := U/V$.

Then U is approximately Binomial$(15000, 90/15000)$ and V is approximately Binomial$(15000, 5/15000)$. By CLT, U and V are approximately

$$U \sim \mathcal{N}(90, \sqrt{90}) \qquad \text{and } V \sim \mathcal{N}(5, \sqrt{5}).$$

Since $\mathbb{P}(V \leqslant 5+\sqrt{5}) = \mathbb{P}(U \geqslant 90 - 2\sqrt{90}) \approx 0.9545$, we have

$$\mathbb{P}\left(V/U \leqslant \frac{5+2\sqrt{5}}{90-2\sqrt{90}}\right) \approx 0.9545^2 > 0.9.$$

Since $\dfrac{5+2\sqrt{5}}{90-2\sqrt{90}} < \dfrac{15}{100}$, the statement is true with $\geqslant 90\%$ certainty.

5.3 p-Value

Definition: (5.17) p-value

Let $X_1, \dots, X_n$ be real-valued random sample with $f_\theta \in \{f_\theta : \theta \in \Theta\}$. Define $X := (X_1, \dots, X_n)$ for convenience. Let $Y := t(X)$ where $t : \mathbb{R}^n \to \mathbb{R}$. For all $c \in \mathbb{R}$, consider the hypothesis test with rejection region $\{x \in \mathbb{R}^n : t(x) \geqslant c\}$. Let $p : \mathbb{R}^n \to [0,1]$ be defined by

$$p(x) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(t(X) \geqslant t(x)) \qquad \text{for all } x \in \mathbb{R}^n.$$

The ***p*-value** for the hypothesis test is defined to be the statistic $p(X)$.

Remark. Intuitively, p-value stands for the probability that our test statistic evaluates to some thing *at least as extreme as* the observed value.

Example: (HW5 P3). Suppose you flip a coin 1000 times, resulting in 560 heads and 440 heads. Is it reasonable to conclude that the coin is fair?

Solution. Flipping a coin can be thought of as a Bernoulli random variable with parameter $\lambda \in [0,1]$. Let $X_i \sim \text{Bernoulli}(\lambda)$ and define $S := \sum_{i=1}^{1000} X_i = \#\text{heads}$. Define the rejection region to be $\{(x_1, \ldots, x_{1000}) \in \mathbb{R}^{1000} : |S - 500| > c\}$ so that $t(X) := |\#\text{heads} - 500|$. Θ_0 is simply $\{1/2\}$ as we assume that H_0 states the coin is fair. Then,

$$p(\text{560 heads}) = \mathbb{P}_{1/2}(S < 440 \text{ or } S > 560).$$

By CLT, S is approximately $\mathcal{N}(500, \sqrt{1000 \cdot (1/2) \cdot (1 - 1/2)}) = \mathcal{N}(500, 5\sqrt{10})$. Hence

$$p(\text{560 heads}) = \mathbb{P}(\mathcal{N}(500, 5\sqrt{10}) < 440 \text{ or } > 560) \approx \mathbb{P}\left(Z < \frac{60}{5\sqrt{10}} \text{ or } > \frac{60}{5\sqrt{10}}\right) = 2\Phi(12/\sqrt{10})$$

which is extremely small. (Here Z denotes the standard normal.) Therefore we reject the null hypothesis, i.e., we claim the coin is biased.

Example: (HW5 P4). Suppose the number of typos in [Heilman's] notes in a given year follows a Poisson distribution. In the last few years, the average number of typos was 15, and this year, [Heilman] had 10 typos in [his] notes. Is it reasonable to conclude that the rate of typos has dropped this year?

Solution. Let $X \sim \Pr(\lambda)$ be the random variable describing Heilman's typos. Since we are only interested in whether the rate has dropped, $\Theta = (0, 15]$, $\Theta_0 = \{15\}$, and $\Theta_1 = (0, 15)$. Hence H_0 is $\{\lambda = 15\}$ and H_1 is $\{0 < \lambda < 15\}$. Let the rejection region be defined by $\{x : t(x) \geqslant c\}$ where $t(X) := 15 - X$. Then if 10 typos are observed, the p-value is

$$p(10) = \mathbb{P}_{15}(15 - X \geqslant 15 - 10) = \mathbb{P}_{15}(X \leqslant 10) = \sum_{k=0}^{10} \frac{15^k e^{-15}}{k!} \approx 0.118,$$

so it is likely that Heilman indeed made some improvements in avoiding typos.

5.4 Generalized Likelihood Ratio Tests

Definition: (5.22) Generalized Likelihood Ratio Test

Let $k > 1$. The **generalized likelihood ratio test** of a hypothesis H_0 that $\{\theta \in \Theta_0\}$ is defined by the following region:

$$C := \{x \in \mathbb{R}^n : \sup_{\theta \in \Theta} f_\theta(x) \geqslant k \sup_{\theta \in \Theta_0} f_\theta(x)\}.$$

If $0 < k \leqslant 1$ *then* $C = \mathbb{R}^n$ *since* $\sup_{\theta \in \Theta}$ *is taken over a largest set than* $\sup_{\theta \in \Theta_0}$, *so* $\sup_{\theta \in \Theta} \geqslant \sup_{\theta \in \Theta_0}$. *The claim becomes trivial in this case.*

Example: (5.24). Let $X_1, \ldots, X_n$ be i.i.d. Gaussians with known $\sigma^2 > 0$ but unknown $\mu \in \mathbb{R}$. Fix $\mu_0 \in \mathbb{R}$. Suppose H_0 is $\mu = \mu_0$ and H_1 is $\mu \neq \mu_0$. Hence $\Theta = \mathbb{R}, \Theta_0 = \{\mu_0\}$, and $\Theta_1 = \mathbb{R} - \{\mu_0\}$. Also, for $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$f_\mu(x) = f_\mu(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Our null region contains μ_0 only so $\sup_{\mu \in \Theta_0} f_\mu(x) = f_{\mu_0}(x)$. Also recall that the MLE in this case is the sample mean. That is,

$$\sup_{\mu \in \Theta} f_\mu(x) = f_{\overline{\mu}}(x)$$

where $\overline{\mu} = (x_1 + \ldots + x_n)/n$. Therefore,

$$C = \{x \in \mathbb{R}^n : f_{\overline{\mu}}(x) \geqslant k f_{\mu_0}(x)\}.$$

Therefore

$$\begin{aligned}
C &= \left\{x \in \mathbb{R}^n : \prod_{i=1}^{n} \exp\left(\frac{-(x_i - \overline{x})^2 + (x_i - \mu_0)^2}{2\sigma^2}\right) \geqslant k\right\} \\
&= \left\{x \in \mathbb{R}^n : \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n}((x_i - \overline{x})^2 - (x_i - \mu_0)^2)\right) \geqslant k\right\} \\
&= \left\{x \in \mathbb{R}^n : \sum_{i=1}^{n} \left[(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j)^2 - (x_i - \mu_0)^2\right] \leqslant -2\sigma^2 \log k\right\} \\
&= \left\{x \in \mathbb{R}^n : -n(\frac{1}{n}\sum_{j=1}^{n} x_j - \mu_0)^2 \leqslant -2\sigma^2 \log k\right\} \\
&= \left\{x \in \mathbb{R}^n : \left|\frac{1}{n}\sum_{j=1}^{n} x_j - \mu_0\right| \geqslant \sqrt{2n^{-1}\sigma^2 \log k}\right\}.
\end{aligned}$$

Intuitively, the rejection region consists of points where the sample mean is far from μ_0.