«بسمه تعالی»

«تمرین سری «1» مباحثی در آمار دکتر علیشاهی»

① ـ الف) دو فرض H_0 و H_1 داریم که فرض H_1 بیان می‌کند که $x \sim P_1$ و فرض H_0 بیان می‌کند که $x \sim P_0$

در واقع این دو فرض با تعریف زیر معادل‌اند:

$$\begin{cases} H_0 : x \sim P_{\theta_0} \\ H_1 : x \sim P_{\theta_1} \end{cases}$$

دقت کنید که هر توزیع را پارامتری در نظر گرفتیم.

می‌خواهیم ثابت کنیم که نسبت درستنمایی این دو توزیع به همراه انتخاب یک کران مناسب، بیش‌ترین توان آزمون را دارد.

$$\text{Power} = \mathbb{P}\left(\text{رد کردن} \;\middle|\; \begin{matrix} H_1 \text{ درست} \\ H_0 \text{ نادرست} \end{matrix}\right) = \Lambda(x) = \frac{\mathcal{L}(\theta_1|x)}{\mathcal{L}(\theta_0|x)}$$

دقت کنید که آزمون فرض که طراحی کردیم: $\quad$ if $\Lambda(x) > t$: H_0 rejects

در ادامه اثبات می‌کنیم این آزمون به ازای انتخاب مناسب t و به ازای سطح معناداری α، بیش‌ترین توان را خواهد داشت.

منطقه‌ی رد کردن به صورت زیر تعریف می‌شود:

$$R_{NP} = \left\{ x : \frac{\mathcal{L}(\theta_1|x)}{\mathcal{L}(\theta_0|x)} > t \right\}$$

به ازای قرار گرفتن x در این مجموعه فرض H_0 رد می‌شود.

حالا فرض کنید t را بگذاریم: $\quad P(R_{NP}(t) \mid \theta_0) = \alpha$

یعنی احتمال اینکه داده در منطقه‌ی رد قرار گیرد به شرط آنکه فرض H_0 درست باشد، α است.

حال یک تابع تست Φ تعریف می‌کنیم که اگر فرض H_0 رد شود $\Phi(x)=1$ و درغیراین صورت $\Phi(x)=0$ است.

توان را هم می‌توانیم به صورت زیر تعریف کنیم:

$$1-\beta=\int \Phi(x) P_{\theta_1}(x)\,dx = E_{\theta_1}[\Phi(x)]$$

در واقع تابع Φ خود یک متغیر تصادفی است زیرا تابعی از داده‌هاست.

$$\int_x \underbrace{\left(\Phi_{NP}(x)-\Phi_A(x)\right)}_{E_1}\underbrace{\left(P_{\theta_1}(x)-tP_{\theta_0}(x)\right)}_{E_2}dx \geq 0$$

اثبات می‌کنیم که این نامساوی در حالت کلی برقرار است:

- اگر هر دو آزمون، H_0 را رد کنند، سمت چپ برابر صفر می‌شود، پس نامساوی برقرار است.
- اگر NP را رد کند و A را رد نکند آنگاه $E_1>0$ است و چون NP فرض را رد کرده، $P_{\theta_1}>P_{\theta_0}$ است پس $E_2>0$
- اگر A رد کند و NP رد نکند، مطابق بالا $E_1<0$, $E_2<0$ پس $E_1E_2>0$

$$\int_x \left(\Phi_{NP}(x)-\Phi_A(x)\right)P_{\theta_1}(x)\,dx \geq t\int\left(\Phi_{NP}(x)-\Phi_A(x)\right)P_{\theta_0}(x)$$

$$= t\int \Phi_{NP}(x)P_{\theta_0}(x)\,dx - t\int \Phi_A(x)P_{\theta_0}(x)\,dx$$

$$\geq 0$$

② $g(\mu|x) = C_0\, g(\mu) f(x|\mu) = C_0 \left(p\,\delta(\mu) + (1-p) \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \right.$

$\delta(\mu) f(\mu) = \delta(\mu) f(0)$

$C_0 \left(\frac{p}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{2\sigma^2}} \delta(\mu) + (1-p) \frac{1}{2\pi s \sigma} e^{\frac{-1}{2}\left(\mu^2\left(\frac{1}{s^2} + \frac{1}{\sigma^2}\right) - \mu\left(\frac{2x}{\sigma^2}\right) + \frac{x^2}{\sigma^2}\right)} \right.$

$C_0 \left(\frac{p}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{2\sigma^2}} \delta(\mu) + (1-p) \frac{1}{2\pi s \sigma} e^{\frac{-A}{2}\left(\mu - \frac{x}{\sigma^2 A}\right)^2 + C_1} \right)$

$C_1 = \frac{x^2}{\sigma^2 A} - \frac{x^2}{\sigma^2 A^2} \Rightarrow \int g(\mu|x)\, d\mu = 1$

$C_0 \left(\frac{p}{\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} + (1-p) \frac{e^{\frac{-AC_1}{2}}}{\sqrt{2\pi A}\, s\sigma} = 1 \right.$

باسمه تعالی

مباحثی در آمار

سری اول (استنباط فراوانی‌گرا و بیزی)

---

۱. لم نیمن ــ پیرسون

آزمون فرض یک روش آماری است برای تصمیم‌گیری راجع به رد کردن یا نکردن یک فرض صفر (H_0) در برابر یک فرض بدیل (H_1)، بر اساس داده مشاهده شده. منظور از خطای نوع یک رد شدن فرض صفر در صورت درست بودن آن است و منظور از خطای نوع دو رد نشدن فرض صفر در صورت غلط بودن این فرض است.

سطح معنی‌داری آزمون ← خطای نوع یک ←

$$\alpha = \mathbb{P}(H_0 \text{ رد شدن} | H_0 \text{ درستی})$$

$$\beta = \mathbb{P}(H_0 \text{ رد نشدن} | H_0 \text{ غلط بودن})$$

توان آزمون $1-\beta$ ← خطای نوع دو ←

به α سطح معنی‌داری آزمون و به $1-\beta$ توان آزمون می‌گویند.

فرض کنید داده x مشاهده شده و فرض صفر و بدیل ما راجع به تابع چگالی توزیع x است:

$\frac{P_1(x)}{P_0(x)} = h(x)$

$$\begin{cases} H_0: & x \sim P_0 \\ H_1: & x \sim P_1 \end{cases}$$

$P_1(x) > P_0(x)\, t$

الف) لم نیمن ــ پیرسون را ثابت کنید:

کمترین β را دارد

عددی مانند t وجود دارد که در بین آزمون‌های با سطح معنی‌داری α، آزمون زیر بیشترین توان را دارد:

$\beta = \mathbb{P}(H_0 \text{ رد نشدن} | H_0 \text{ غلط بودن})$

$$\frac{P_1(x)}{P_0(x)} > t \Longrightarrow H_0 \text{ رد شدن}$$

اگر $h(x) > t$ باشد، فرض صفر رد می‌شود

ب) این لم را برای حالت زیر اعمال کنید و نتیجه را توصیف کنید:

$$\begin{cases} H_0: & x_1, \cdots, x_n \sim \mathcal{N}(0,1) \\ H_1: & x_1, \cdots, x_n \sim \mathcal{N}(\mu,1) \quad (\mu > 0) \end{cases}$$

۲. آزمون فرض بیزی

الف) فرض کنید $x \sim \mathcal{N}(\mu, \sigma^2)$ و برای μ توزیع پیشینی در نظر می‌گیریم که احتمال صفر شدن آن مثبت باشد: $\mu \sim p\delta_0 + (1-p)\mathcal{N}(0, s^2)$. در این‌جا δ_0 اندازه اتمی دلتای دیراک در نقطه صفر است. توزیع پسین $g(\mu|x)$ را محاسبه کنید.

ب) توضیح دهید که چگونه می‌توان یک آزمون فرض بیزی برای بررسی فرض $\mu = 0$ در مقابل فرض $\mu \neq 0$ طراحی کرد.

$g(\mu|x) = g(x)\, g(x|\mu) \to c\, g(x)\, \mathcal{L}(x|\mu) = g(\mu|x)$

$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

۳. **مشاهدات بیشتر و افزایش اطمینان**

الف) فرض کنید $x \sim \mathcal{N}(\mu, \sigma^2)$ و $\mu \sim \mathcal{N}(0, s^2)$. توزیع پسین $g(\mu|x)$ را محاسبه کنید و نشان دهید واریانس توزیع پسین کمتر از واریانس توزیع پیشین است.

ب) آیا می‌توانید مثالی ارائه دهید که واریانس توزیع پسین از توزیع پیشین بیشتر باشد؟

ج) نشان دهید به طور متوسط مشاهده داده عدم اطمینان ما را کاهش می‌دهد! به این معنی که اگر واریانس توزیع پسین را متغیری تصادفی در نظر بگیریم (به خاطر تصادف موجود در داده)، امید ریاضی آن کوچک‌تر یا مساوی واریانس توزیع پیشین است:

$$\mathbb{E}_x[\mathrm{Var}[\theta|x]] \leq \mathrm{Var}[\theta]$$

۴. **آماره بسنده**

در مدل پارامتری $f_\theta(x)$ آماره $s = s(x)$ بسنده است.

الف) نشان دهید آماردان فراوانی‌گرا برای استنباط با روش بیشترین درست‌نمایی فقط به مقدار s نیاز دارد.

ب) نشان دهید آماردان بیزگرا هم برای محاسبه توزیع پسین نیازی به کل داده x ندارد و دانستن s برای او کافی است.

۵. **کمیت لولا**

منظور از کمیت لولا تابعی از داده‌ها و پارامترهای مدل است که توزیع آن به پارامترها بستگی نداشته باشد.

فرض کنید $Y_1, \cdots, Y_n$ نمونه‌ای $i.i.d.$ از تابع چگالی $f_\theta(x) = \frac{1}{\theta} e^{\frac{-x}{\theta}}$ باشد. اگر $\hat{\theta}$ برآوردگر بیشترین درست‌نمایی باشد، ابتدا توضیح دهید چگونه می‌توان با توجه به توزیع مجانبی $\hat{\theta}$ یک بازه اطمینان تقریبی برای θ ارائه داد.

سپس نشان دهید $\frac{\hat{\theta}}{\theta}$ یک کمیت لولاست، توزیع آن را به دست آورید و توضیح دهید چگونه می‌توان به کمک آن یک بازه اطمینان دقیق ساخت.

۶. **توزیع پیشین جفریز**

توزیع پیشین جفریز یک توزیع ناآگاهنده روی فضای پارامتر است که متناسب است با ریشه دوم اطلاع فیشر: $g^{\text{Jeff}}(\theta) \propto \sqrt{I_\theta}$.

الف) نشان دهید توزیع پیشین جفریز تحت تغییر پرمایش خوش رفتار است.

ب) توزیع پیشین جفریز را برای مدل دوجمله‌ای $f_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$ محاسبه کنید.

۷. **اریبی و واریانس برآوردگر بیشترین درست‌نمایی**

الف) برای برآوردگر $\hat{\theta}$ از پارامتر θ، امید مربع خطا، MSE، به صورت $\mathbb{E}(\hat{\theta} - \theta)^2$ تعریف می‌شود. نشان دهید اگر bias و variance اریبی و واریانس تخمین‌گر $\hat{\theta}$ باشند، $\text{MSE} = \text{variance} + \text{bias}^2$

ب) ثابت کنید برای تخمین‌گر بیشترین درست‌نمایی $\text{variance} = O(1/n)$, $\text{bias}^2 = O(1/n^2)$

با اطلاعات فیشر و حد کرامر رائو قابل حل است .

۸. **پدیده استاین در بعد یک**

الف) فرض کنید $\hat{\theta}$ تخمین‌گر دلخواهی از θ باشد. اگر $b(\theta) = \mathbb{E}[\hat{\theta}] - \theta$ اریبی $\hat{\theta}$ باشد، نشان دهید:

$$\mathrm{Var}[\hat{\theta}] \geq \frac{(1+b'(\theta))^2}{I(\theta)}$$

ب) نشان دهید در بعد یک پارادوکس استاین نداریم:
اگر $x_1, \cdots, x_n \sim \mathcal{N}(\theta, 1)$ نشان دهید با معیار امید مربع خطا، هیچ تخمین‌گری اکیداً بهتر از بیشترین درست‌نمایی $\hat{\theta} = \frac{x_1+x_2+\cdots+x_n}{n}$ نیست.

۹. **استنباط شرطی**

فرض کنید $i = 1, 2$ $x_i \sim \mathcal{N}(\mu_i, 1)$ و می‌خواهیم بردار میانگین‌ها، $\mu = (\mu_1, \mu_2)$ را برآورد کنیم. اگر بدانیم که μ در فاصله ۲ از مبدا قرار گرفته است، $\mu = (2\cos\theta, 2\sin\theta)$ و مساله برآورد θ خواهد بود. برای این کار از $\hat{\theta} = \arctan(x_2/x_1)$ استفاده می‌کنیم و قرارداد می‌کنیم که $\hat{\theta} - \theta$ از $-\pi$ تا π تغییر می‌کند.

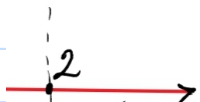
الف) نشان دهید $\hat{\theta}$ نااریب است و امید مربع خطای آن را تا دو رقم اعشار محاسبه کنید.

ب) اگر $(x_1, x_2) = (1, 1)$ دقت $\hat{\theta}$ چقدر است؟ نظرتان درباره محاسبه خطا به شرط $r = \sqrt{x_1^2 + x_2^2}$ چیست؟!

۱۰. **تخمین ماکزیمم**

فرض کنید برای $n = 1, \cdots, N$، داریم $x_i \sim \mathcal{N}(\mu_i, 1)$ و می‌دانیم اکثر μ_iها صفر هستند. چگونه به کمک x_iها $\max_i \mu_i$ را تخمین بزنیم؟ برای پاسخ یک بار خود را آماردان فراوانی‌گرا و بار دیگر آماردان بیزگرا تصور کنید!

$\hat{\theta} = tg^{-1}\left(\frac{x_2}{x_1}\right)$ $\quad x_1 \sim N(\mu_1, 1)$, $x_2 \sim N(\mu_2, 1)$



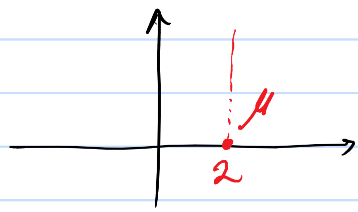
۹. **استنباط شرطی**

فرض کنید $i = 1, 2$ $x_i \sim \mathcal{N}(\mu_i, 1)$ و می‌خواهیم بردار میانگین‌ها $\mu = (\mu_1, \mu_2)$ را برآورد کنیم. اگر بدانیم که μ در فاصله ۲ از مبدا قرار گرفته است، $\mu = (2\cos\theta, 2\sin\theta)$ و مساله برآورد θ خواهد بود. برای این کار از $\hat{\theta} = \arctan(x_2/x_1)$ استفاده می‌کنیم و قرارداد می‌کنیم که $\hat{\theta} - \theta$ از $-\pi$ تا π تغییر می‌کند.

الف) نشان دهید $\hat{\theta}$ نااریب است و امید مربع خطای آن را تا دو رقم اعشار محاسبه کنید.

ب) اگر $(x_1, x_2) = (1, 1)$ دقت $\hat{\theta}$ چقدر است؟ نظرتان درباره محاسبه خطا به شرط $r = \sqrt{x_1^2 + x_2^2}$ چیست؟!

$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$



$\mu = (2\cos\theta, 2\sin\theta)$

$\hat{\theta} = tg^{-1}\left(\frac{x_2}{x_1}\right)$

$\mathbb{E}(\hat{\theta}) = \theta \qquad \mathbb{E}\left(tg^{-1}\left(\frac{x_2}{x_1}\right)\right)$

۷. **اریبی و واریانس برآوردگر بیشترین درست‌نمایی**

الف) برای برآوردگر $\hat{\theta}$ از پارامتر θ، امید مربع خطا، MSE، به صورت $\mathbb{E}(\hat{\theta} - \theta)^2$ تعریف می‌شود. نشان دهید اگر bias و variance اریبی و واریانس تخمین‌گر $\hat{\theta}$ باشند، $\text{MSE} = \text{variance} + \text{bias}^2$

ب) ثابت کنید برای تخمین‌گر بیشترین درست‌نمایی $\text{variance} = O(1/n)$, $\text{bias}^2 = O(1/n^2)$

$$E(\hat{\theta} - \theta)^2 = \mathbb{E}(\hat{\theta} - \bar{x} + \bar{x} - \theta)^2 =$$

$$E(\hat{\theta} - \bar{x})^2 + E(\bar{x} - \theta)^2 + \cancel{2E((\hat{\theta} - \bar{x})(\bar{x} -}$$ 0

$$E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - \bar{x})^2 + E(\bar{x} - \theta)^2$$

دقت کنید که bias میزان دقت مدل روی داده های train هستش، و واریانس میزان عملکرد مدل روی داده های test.

اگر bias زیاد باشه دچار کم برازش شده ایم (underfitting)

اگر variance زیاد باشه یعنی دچار overfitting شده ایم.

۳. مشاهدات بیشتر و افزایش اطمینان

الف) فرض کنید $x \sim \mathcal{N}(\mu, \sigma^2)$ و $\mu \sim \mathcal{N}(0, s^2)$. توزیع پسین $g(\mu|x)$ را محاسبه کنید و نشان دهید واریانس توزیع پسین کمتر از واریانس توزیع پیشین است.

ب) آیا می‌توانید مثالی ارائه دهید که واریانس توزیع پسین از توزیع پیشین بیشتر باشد؟

ج) نشان دهید به طور متوسط مشاهده داده عدم اطمینان ما را کاهش می‌دهد! به این معنی که اگر واریانس توزیع پسین را متغیری تصادفی در نظر بگیریم (به خاطر تصادف موجود در داده)، امید ریاضی آن کوچک‌تر یا مساوی واریانس توزیع پیشین است:

$$\mathbb{E}_x[\mathrm{Var}[\theta|x]] \leq \mathrm{Var}[\theta]$$

$$Var(\theta) = \mathbb{E}\left[Var(\theta|x)\right] + \underbrace{Var\big(\mathbb{E}(\theta|x)\big)}_{\geq 0} \rightarrow$$ اثبات ج

الف

$$g(\mu|x) = c\, g(\mu)\, f(x|\mu)$$

$$\frac{1}{\sqrt{2\pi s}} \exp\left(\frac{-\mu^2}{2s^2}\right) \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$g(\mu|x) = \frac{c}{2\pi\sigma s} \exp\left(\frac{-\mu^2}{2s^2}\right) \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

$$c = \left[\int \frac{1}{2\pi\sigma s} \exp\left(\frac{-1}{2}\left(\frac{\mu^2}{s^2} + \frac{(x-\mu)^2}{\sigma^2}\right)\right) d\mu\right]^{-1}$$

$$g(\mu|x) = \frac{\exp\left(\frac{-1}{2}\left(\frac{\mu^2}{s^2} + \frac{(x-\mu)^2}{\sigma^2}\right)\right)}{\left[\int \frac{1}{2\pi\sigma s} \exp\left(\frac{-1}{2}\left(\frac{\mu^2}{s^2} + \frac{(x-\mu)^2}{\sigma^2}\right)\right) d\mu\right]^{-1}}$$

lemma: $$\int_{-\infty}^{+\infty} \exp\left(\frac{-1}{2}(ax^2+bx+c)\right) dx = \sqrt{\frac{2\pi}{a}} \exp\left(\frac{-1}{2}\left(c-\frac{b^2}{4a}\right)\right)$$

final-form $$= \frac{1}{\sqrt{\frac{2\pi\sigma^2 s^2}{s^2+\sigma^2}}} \exp\left(-\frac{\left(\mu - \frac{\frac{x}{\sigma^2}}{\frac{1}{s^2}+\frac{1}{\sigma^2}}\right)}{2\frac{\sigma^2 s^2}{\sigma^2+s^2}}\right)$$

۴. آماره بسنده

در مدل پارامتری $f_\theta(x)$ آماره $s = s(x)$ بسنده است.

الف) نشان دهید آماردان فراوانی‌گرا برای استنباط با روش بیشترین درست‌نمایی فقط به مقدار s نیاز دارد.

ب) نشان دهید آماردان بیزگرا هم برای محاسبه توزیع پسین نیازی به کل داده x ندارد و دانستن s برای او کافی است.

الف)

$$f_\theta(x) = g(x)\, h_\theta(s(x),\theta) = f(x|\theta)$$

$$\mathcal{L}(\theta, X) = \prod_i f(x_i|\theta) = \prod g(x_i)\, h_\theta(s(x_i),\theta)$$

$$\log(\mathcal{L}(\theta,X)) = \log\left(\prod f(x_i|\theta)\right) = \log(g(x_i)) + \log(h_\theta(s(x_i),\theta))$$

$$\frac{\partial \log(\mathcal{L}(\theta,X))}{\partial\theta} = \boxed{\frac{\partial \log(g(x_i))}{\partial\theta}} + \frac{\partial}{\partial\theta}\left(\log(h_\theta(s(x_i),\theta))\right)$$

(independent from θ) ربطی به θ ندارد

$$\frac{\partial \log(\mathcal{L}(\theta,X))}{\partial\theta} \propto \frac{\partial}{\partial\theta}\left(\log(h_\theta(s(x_i),\theta))\right) \quad \blacksquare$$

۴. **آماره بسنده**

در مدل پارامتری $f_\theta(x)$ آماره $s = s(x)$ بسنده است.

الف) نشان دهید آماردان فراوانی‌گرا برای استنباط با روش بیشترین درست‌نمایی فقط به مقدار s نیاز دارد.

ب) نشان دهید آماردان بیزگرا هم برای محاسبه توزیع پسین نیازی به کل داده x ندارد و دانستن s برای او کافی است.

آماره بسنده

ب)

$$f_\theta(x) = g(x)\, h_\theta(S(x),\theta) = f(x|\theta)$$

$$L(X,\theta) = \prod_i f(x_i|\theta) = g(x)\, h_\theta(S(x),\theta)$$

$$f(x,\theta) = f(x|\theta)\, f(\theta) \qquad g(x) \sim N(\mu,\sigma^2)$$

$$\frac{\partial L(X,\theta)}{\partial\theta} = \prod_i f(x_i,\theta) \Rightarrow \frac{\partial}{\partial\theta}\left(\prod_i f(x_i,\theta)\right)$$

$$\frac{\partial}{\partial\theta}\left(\prod_i (f(x_i,\theta)\right) = \frac{\partial}{\partial\theta}\left(g(x)\, h_\theta(S(x),\theta)\right) = \frac{\partial}{\partial\theta} h_\theta(S(x),\theta)$$

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ربطی به θ ندارد!!

۶. توزیع پیشین جفریز

توزیع پیشین جفریز یک توزیع ناآگاهنده روی فضای پارامتر است که متناسب است با ریشه دوم اطلاع فیشر: $g^{\text{Jeff}}(\theta) \propto \sqrt{I_\theta}$.

الف) نشان دهید توزیع پیشین جفریز تحت تغییر پرمایش خوش رفتار است.

ب) توزیع پیشین جفریز را برای مدل دوجمله‌ای $f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ محاسبه کنید.

① $I_\theta = E\left[\frac{\partial}{\partial \mu}(f(x|\mu))\right]^2$ or $\ell(\theta, x) = f(x|\theta) \to I_\theta = \mathbb{E}[-\ddot{\ell}(\theta)]$

We can implement "Fisher Information" with these two formula.

$$\eta = h(\theta) \to f(x|\theta) \to g(x|\eta)$$

$$f(x|\theta)f(\theta) = g(x|\eta)g(\eta)$$

$$f(x|\theta) = g(x|\eta)$$

$$I_\theta = \left[\mathbb{E}\left(-\frac{\partial^2}{\partial\theta^2}(\log f(x|\theta)\right)\right]$$

$$I_\eta = \left[\mathbb{E}\left(\frac{-\partial^2}{\partial\eta^2}(\log g(x|\eta)\right)\right] = \mathbb{E}\left[\left(\frac{-\partial^2}{\partial\eta^2}(\log f(x|\theta)\right)\right]$$

$$\mathbb{E}\left[\left(\frac{-\partial^2}{\partial\eta^2}(\log f(x|\theta)\right)\right] = \mathbb{E}\left[\left(\frac{-\partial}{\partial\eta}\left(\frac{\partial}{\partial\theta}(\log f(x|\theta))\right)\frac{\partial\theta}{\partial\eta}\right]\right]$$

$$\frac{\partial}{\partial\eta} = \frac{\partial}{\partial\theta} \times \frac{\partial\theta}{\partial\eta} \to \mathbb{E}\left[\left(\frac{-\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}(\log f(x|\theta))\frac{\partial\theta}{\partial\eta}\right)\frac{\partial\theta}{\partial\eta}\right]\right]$$

$$\mathbb{E}\left[\frac{-\partial^2}{\partial\theta^2}(\log f(x|\theta))\right]\left(\frac{\partial\theta}{\partial\eta}\right)^2 + \left(\frac{\partial^2\theta}{\partial\eta^2}\right)\underbrace{\left(\frac{\partial \log f(x|\theta)}{\partial\theta}\right)}_{0}$$

$$I(\eta) = I(\theta)\left(\frac{\partial\theta}{\partial\eta}\right)^2 \to \sqrt{I(\eta)} = \sqrt{I(\theta)} \times \frac{\partial\theta}{\partial\eta}$$

۸. **پدیده استاین در بعد یک**

الف) فرض کنید $\hat{\theta}$ تخمین‌گر دلخواهی از θ باشد. اگر $b(\theta) = \mathbb{E}[\hat{\theta}] - \theta$ اریبی $\hat{\theta}$ باشد، نشان دهید:

$$\mathrm{Var}[\hat{\theta}] \geq \frac{(1+b'(\theta))^2}{I(\theta)}$$

ب) نشان دهید در بعد یک پارادوکس استاین نداریم:

اگر $x_1, \cdots, x_n \sim \mathcal{N}(\theta, 1)$ نشان دهید با معیار امید مربع خطا، هیچ تخمین‌گری اکیداً بهتر از بیشترین درست‌نمایی $\hat{\theta} = \frac{x_1+x_2+\cdots+x_n}{n}$ نیست.

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\left[\theta - \mathbb{E}(\hat{\theta})\right]^2 \quad , \quad I(\theta) = \mathbb{E}(-\ddot{\ell}(\theta))$$

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}(\log(f(x|\theta))\right]^2$$

$$I(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}(\log(f(x|\theta))\right]$$

$$b(\theta) = \mathbb{E}(\hat{\theta}) - \theta \Rightarrow \frac{\partial b(\theta)}{\partial\theta} = \frac{\partial \mathbb{E}(\hat{\theta})}{\partial\theta} - 1$$

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \mathbb{E}(\theta))^2 = \mathbb{E}(\theta^2) + \mathbb{E}(\hat{\theta})^2$$

$$\mathrm{Var}(\hat{\theta}) \geq \frac{(1+b'(\theta))^2}{I_\theta} \Rightarrow \frac{\left(\frac{\partial \mathbb{E}(\hat{\theta})}{\partial\theta}\right)^2}{\mathbb{E}\left[\frac{\partial}{\partial\theta}(\log(f(x|\theta))\right]^2}$$

$$\frac{\mathbb{E}(\theta)^2 + \mathbb{E}(\hat{\theta})^2}{\frac{\partial^2 \mathbb{E}(\hat{\theta})}{\partial\theta^2}} \geq \frac{1}{\mathbb{E}\left[\frac{\partial}{\partial\theta}(\log(f(x|\theta))\right]^2}$$