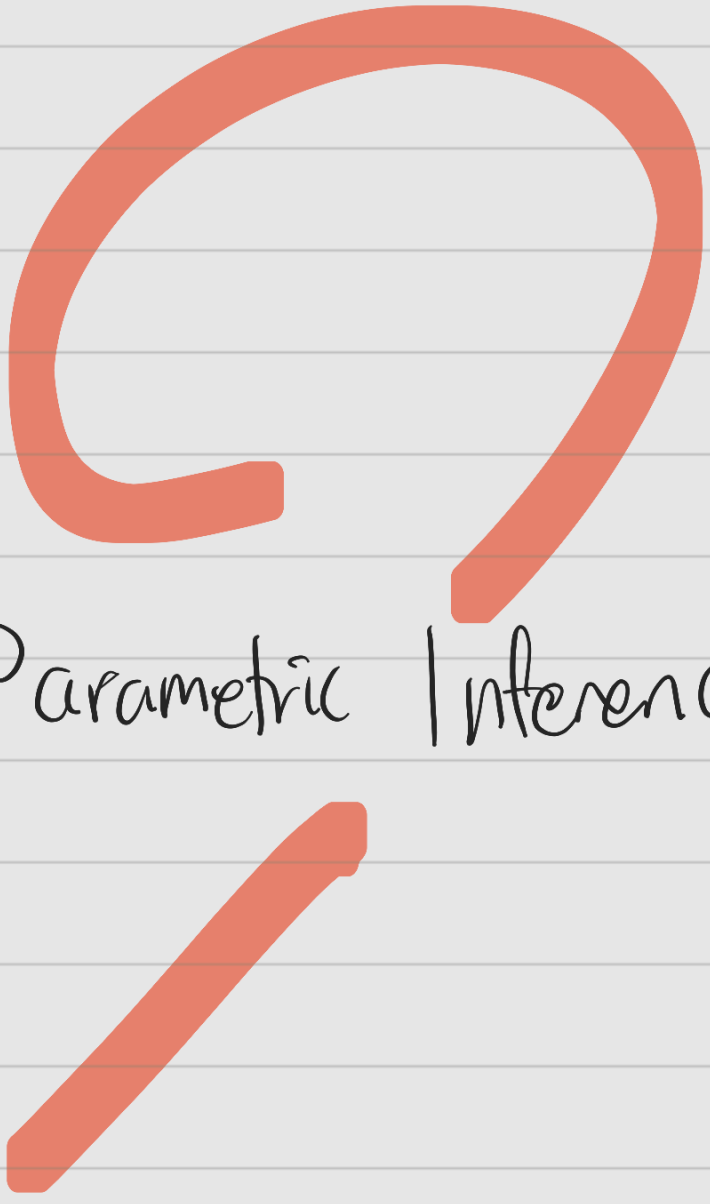


Parametric Inference

§9.1 Parameter of Interest

For $X \sim N(\mu, \sigma^2)$, if we're interested in estimating μ, $\mu = T(\theta)$ is parameter of interest. σ is nuisance parameter.

Example $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. $\theta = (\mu, \sigma)$.

$\Theta = \{(\mu, \sigma) \mid \mu \in \mathbb{R}, \sigma > 0\}$. X_i: outcome of blood test

T: fraction of population with test score > 1.

Then, $T = P(X > 1) = 1 - P(X < 1) = 1 - P\left(Z < \frac{1-\mu}{\sigma}\right)$

$= 1 - \Phi\left(\frac{1-\mu}{\sigma}\right)$.

§9.2 The Method of Moments

Definition (Moment).

$\theta = (\theta_1, \cdots, \theta_k)$, $1 \le j \le k$.

- $\alpha_j \equiv \alpha_j(\theta) = E_\theta(X^j) = \int x^j dF_\theta(x)$ is jth moment.
- $\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$ is jth sample moment

Definition (Method of Moments Estimator $\hat{\theta}_n$).

$\hat{\theta}_n$: ——— is defined to be the value of θ s.t.

$$\begin{cases} \alpha_1(\hat{\theta}_n) = \hat{\alpha}_1 \\ \vdots \\ \alpha_k(\hat{\theta}_n) = \hat{\alpha}_k \end{cases}$$

Example $X_1, \dots, X_n \sim N(\mu, \sigma^2)$.

$\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$,

$\alpha_2 = \mathbb{E}_\theta(X_1^2) = \sigma^2 + \mu^2$.

$$\Rightarrow \quad \alpha_1(\hat{\theta}_n) = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\alpha}_1.$$

$$\alpha_2(\hat{\theta}_n) = \hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\alpha}_2.$$

- Although method of moments is consistent and asymptotically normal, it is not optimal. More often use maximum likelihood estimator

Theorem Let $\hat{\theta}_n$ be the method of moments estimator.

1. $\hat{\theta}_n$ exists with probability tending to 1.
2. The estimate is consistent: $\hat{\theta}_n \xrightarrow{P} \theta$.
3. The estimate is asymptotically normal:

$\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$, with $\Sigma = g\,\mathbb{E}_\theta(YY^T)\,g^T$.

$Y = (X, X^2, \dots, X^k)^T$,

$g = (g_1, \dots, g_k)$, $g_j = \partial \alpha_j^{-1}(\theta)/\partial\theta$.

(pf)

§9.3 Maximum Likelihood

Definition (Likelihood Function)

The likelihood fn is $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$.

The log-likelihood fn is $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

- Likelihood fn is joint density of data except we

View it as a ftn of θ.

Definition (Maximum Likelihood Estimator (MLE)).

MLE is $\hat{\theta}_n = \arg\max \mathcal{L}_n(\theta)$.

$= \arg\max \ell_n(\theta)$.

Example (Bernoulli MLE)

$X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. $f(x;p) = p^x (1-p)^{1-x}$.

$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i;p) = p^S (1-p)^{n-S}$, where $S = \sum X_i$.

$\ell_n(p) = S \log p + (n-S) \log(1-p)$.

$\frac{d\ell_n}{dp} = \frac{S}{p} - \frac{n-S}{1-p} = 0 \rightarrow S(1-p) = (n-S)p$

$\rightarrow \hat{p}_n = \frac{S}{n}$.

Example (Normal MLE)

$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. $\theta = (\mu, \sigma)$.

$$\mathcal{L}_n(\mu,\sigma) = \prod_i \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(X_i-\mu)^2}$$

$$= \sigma^{-n} e^{-\frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2}$$

$$= \sigma^{-n} e^{-\frac{nS^2}{2\sigma^2}} e^{-\frac{n(\bar{X}-\mu)^2}{2\sigma^2}},$$

where $\bar{X} = n^{-1}\sum_i X_i$, $S^2 = n^{-1}\sum_i (X_i - \bar{X})^2$.

$$\ell(\mu,\sigma) = -n\log\sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X}-\mu)^2}{2\sigma^2}.$$

$$\frac{\partial \ell}{\partial \mu} = -\frac{n}{\sigma^2}(\mu-\bar{X}) = 0 \to \hat{\mu} = \bar{X}.$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{nS^2}{\sigma^3} + \frac{n(\bar{X}-\mu)^2}{\sigma^3} = 0 \to \hat{\sigma}^2 = \hat{S}^2.$$

Example (Uniform MLE).

$X_1, \cdots, X_n \sim \text{Unif}(0,\theta)$. $f(x;\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta \\ 0, & \text{otherwise.} \end{cases}$

Suppose $\theta < X_i$ for some i. Then $f(X_i;\theta) = 0$. thus, $\mathcal{L}_n(\theta) = 0$.

Hence $\mathcal{L}_n(\theta) = 0$ if $X_i > \theta$ for any i.

Let $X_{(n)} = \max\{X_1, \cdots, X_n\}$. Then,

$$\mathcal{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } \theta \ge X_{(n)} \\ 0 & \theta < X_{(n)}. \end{cases}$$

$\mathcal{L}_n(\theta)$ is strictly decreasing over $[X_{(n)}, \infty)$. Thus, $\hat{\theta}_n = X_{(n)}$.

Theorem (Properties of MLE)

1. MLE is consistent: $\hat{\theta}_n \xrightarrow{P} \theta_*$, where θ_* is the true value of the parameter θ
2. MLE is equivariant: if $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$
3. MLE is asymptotically Normal: $(\hat{\theta} - \theta_*)/\widehat{se} \rightsquigarrow N(0,1)$.

4. MLE is asymptotically optimal.
   i.e. among all well-behaved estimators, MLE has the smallest variance, at least for large samples.
5. MLE is approximately the Bayes estimator.

§9.5 Consistency of Maximum Likelihood Estimator

Definition (Kullback-Leibler Divergence (KL-Divergence)).
KL-divergence btwn f, g is $D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$.

- By Jensen, $D(f,g) \geq 0$. Moreover, $D(f,f) = 0$.
  for $\theta, \psi \in \Theta$, $D(\theta, \psi) := D(f(x;\theta), f(x;\psi))$.

- $\mathcal{F}$ is identifiable if $\theta \neq \psi \Rightarrow D(\theta, \psi) > 0$.
  From now on, assume the model is identifiable.

Note Maximizing $\ell_n(\theta)$ is equivalent to maximizing
$M_n(\theta) = \frac{1}{n} \sum_i \log \frac{f(X_i;\theta)}{f(X_i;\theta_*)}$. ($\because M_n(\theta) = n^{-1}(\ell_n(\theta) - \ell_n(\theta^*))$.)

By WLLN, $E_{\theta_*}\left(\log \frac{f(X_i;\theta)}{f(X_i;\theta_*)}\right) = \int \log\left(\frac{f(x;\theta)}{f(x;\theta_*)}\right) f(x;\theta_*) dx$

$M_n(\theta)$ converges to

$$= -\int \log\left(\frac{f(x;\theta_*)}{f(x;\theta)}\right) f(x;\theta_*)\,dx$$

$$= -D(\theta_*, \theta).$$

Hence, $M_n(\theta) \approx -D(\theta_*, \theta)$. ($M_n(\theta) \xrightarrow{P} -D(\theta_*, \theta)$.

As $-D(\theta_*, \theta_*) = 0$, $-D(\theta_*, \theta) < 0 \quad \forall\, \theta \neq \theta_*$,

we expect $\operatorname{argmax}_\theta M_n(\theta) \longrightarrow \theta_*$.

To prove this, we need $M_n(\theta) \xrightarrow{P} -D(\theta_*, \theta)$ uniformly over θ.

Theorem Let $M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(x_i;\theta)}{f(x_i;\theta_*)}$.

$M(\theta) = -D(\theta_*, \theta)$.

Suppose $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$, and ($M_n(\theta) \to M(\theta)$ uniformly $/\theta$).

$\forall \varepsilon > 0$, $\sup_{\theta : |\theta-\theta_*|\geq\varepsilon} M(\theta) < M(\theta_*)$. ($\theta_*$ is isolated maximum of M).

Let $\hat{\theta}_n$ denote the MLE. Then, $\hat{\theta}_n \xrightarrow{P} \theta_*$.

(Pf) Let $\delta > 0$ be given.

$\hat{\theta}_n = \operatorname{argmax}_\theta M_n(\theta)$. Let $\varepsilon = M(\theta_*) - \sup_{\theta : |\theta-\theta_*|\geq\delta} M(\theta) > 0$.

$\exists N$ s.t. $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| < \frac{\varepsilon}{2} \quad \forall\, n \geq N$.

$\sup_{\theta : |\theta-\theta_*|\geq\delta} M(\theta) < M(\theta_*)$. Let $n \geq N$

If $|\hat{\theta}_n - \theta_*| \geq \delta \Rightarrow M(\theta_*) - \sup_{\theta : |\theta - \theta_*| \geq \delta} M(\theta) < M(\theta_*) - M(\hat{\theta}_n) \Rightarrow M(\theta_*) > M(\hat{\theta}_n) + \varepsilon$

$|M_n(\theta_*) - M(\theta_*)| < \varepsilon = M(\theta_*) - \sup_{\theta : |\theta - \theta_*| \geq \varepsilon} M(\theta)$

$\leq M(\theta_*) - M(\hat{\theta}_n)$

$|M_n(\hat{\theta}_n) - M(\hat{\theta}_n)| < \varepsilon$

$\Rightarrow M_n(\hat{\theta}_n) = (M_n(\hat{\theta}_n) - M(\hat{\theta}_n)) + (M(\hat{\theta}_n) - M(\theta_*))$

$+ (M(\theta_*) - M_n(\theta_*)) + M_n(\theta_*)$

$< \frac{\varepsilon}{2} - \varepsilon + \frac{\varepsilon}{2} + M_n(\theta_*)$

$\rightarrow \leftarrow$ to $\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} M_n(\theta)$.

Thus, $|\hat{\theta}_n - \theta_*| < \delta$. $\forall n \geq N$.

§9.6 Equivariance of the MLE.

Theorem Let $\tau = g(\theta)$ be a function of θ, $\hat{\theta}_n$ be the MLE of θ.
Then, $\hat{\tau}_n = g(\hat{\theta}_n)$ is the MLE of τ.

(pf) Let $h = g^{-1}$. Then $\hat{\theta}_n = h(\hat{\tau}_n)$.

$$\mathcal{L}(\tau) = \prod f(x_i; h(\tau)) = \prod f(x_i;\theta) = \mathcal{L}(\theta), \quad \theta = h(\tau).$$

Thus, $\mathcal{L}_n(\tau) = \mathcal{L}(\theta) \le \mathcal{L}(\hat{\theta}) = \mathcal{L}_n(\tilde{\tau})$

§9.7 Asymptotic Normality

Definition (Score Function / Fisher information)

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta} \qquad \text{(score fn)}$$

$$I_n(\theta) = \mathbb{V}_\theta\left(\sum_{i=1}^n s(X_i;\theta)\right) = \sum_{i=1}^n \mathbb{V}_\theta(s(X_i;\theta)). \qquad \text{(Fisher info)}$$

(Denote $I = I_1$)

Theorem $\mathbb{E}_\theta(s(X;\theta)) = 0$. $\quad I_n(\theta) = nI(\theta)$.

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right) = -\int \left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right) f(x;\theta)\, dx.$$

(pf) $\mathbb{E}_\theta(s(X;\theta)) = \int \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)\, dx$

$$= \int \frac{\partial}{\partial\theta} f(x;\theta)\,dx = \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \frac{\partial}{\partial\theta} 1 = 0.$$

$$I(\theta) = \mathbb{V}_\theta(s(X;\theta)) = \mathbb{E}(s(X;\theta)^2)$$

$$= \int \left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)\,dx$$

$$= \int \frac{1}{f(x;\theta)^2}\left(\frac{\partial}{\partial\theta} f(x;\theta)\right)^2 f(x;\theta)\,dx$$

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{\partial}{\partial\theta}\left(\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta} f(x;\theta)\right)$$

$$= -\frac{1}{f(x;\theta)^2}\left(\frac{\partial}{\partial\theta} f(x;\theta)\right)^2 + \frac{1}{f(x;\theta)}\frac{\partial^2}{\partial\theta^2} f(x;\theta).$$

$$-\mathbb{E}_\theta\left(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right) = \int \frac{1}{f(x;\theta)^2}\left(\frac{\partial}{\partial\theta} f(x;\theta)\right)^2 f(x;\theta)\,dx$$

$$- \int \frac{\partial^2}{\partial\theta^2} f(x;\theta)\,dx.$$

Theorem (Asymptotic Normality of MLE)

Let $se = \sqrt{V(\hat{\theta}_n)}$. Under regularity condition

1. $se \approx \sqrt{1/I_n(\theta)}$ and $\frac{(\hat{\theta}_n - \theta)}{se} \rightsquigarrow N(0,1)$.

2. $\hat{se} = \sqrt{1/I_n(\hat{\theta}_n)}$. Then, $\frac{(\hat{\theta}_n - \theta)}{\hat{se}} \rightsquigarrow N(0,1)$.

(pf).

Theorem Let $C_n = (\hat{\theta}_n - z_{\alpha/2}\hat{se},\ \hat{\theta}_n + z_{\alpha/2}\hat{se})$.

Then, $P_\theta(\theta \in C_n) \to 1-\alpha$ as $n \to \infty$

(pf) $P_\theta(\theta \in C_n) = P(\hat{\theta}_n - z_{\alpha/2}\hat{se} < \theta < \hat{\theta}_n + z_{\alpha/2}\hat{se})$

$$= P\left(-z_{\alpha/2} < \frac{\theta - \hat{\theta}_n}{\hat{se}} < z_{\alpha/2}\right)$$

$$\to P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha.$$

Example $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. MLE is $\hat{p}_n = \sum X_i / n$

$f(x;p) = p^x (1-p)^{1-x}$, $\log f(x;p) = x \log p + (1-x)\log(1-p)$,

$$s(X;p) = \frac{X}{p} - \frac{1-X}{1-p},$$

$$-s'(X;p) = \frac{X}{p^2} + \frac{1-X}{(1-p)^2}.$$

$$I(p) = E_p(-s'(X;p)) = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

$$\hat{se} = \frac{1}{\sqrt{I_n(\hat{p}_n)}} = \frac{1}{\sqrt{n I(\hat{p}_n)}} = \sqrt{\frac{p(1-p)}{n}}.$$

Example $X_1, \cdots, X_n \sim N(\theta, \sigma^2)$, with σ^2 known.

$\hat{\theta}_n = \bar{X}_n$. $\log f(x;\theta) = \log \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(x-\theta)^2}$

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

$$s(X;\theta) = \frac{1}{\sigma^2}(X-\theta).$$

$$s'(X;\theta) = -\frac{1}{\sigma^2}$$

$$I_1(\theta) = \frac{1}{\sigma^2}. \quad I_n(\theta) = \frac{n}{\sigma^2}. \quad \hat{se} = \frac{1}{\sqrt{I_n(\theta)}} = \frac{\sigma}{\sqrt{n}}.$$

$$\hat{\theta}_n = \bar{X}_n \approx N\left(\theta, \frac{\sigma^2}{n}\right)$$

Example $X_1, \cdots, X_n \sim \text{Poisson}(\lambda)$. $P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$.

$$\hat{\lambda}_n = \bar{X}_n.$$

$$\mathcal{L}_n(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{X_i}}{X_i!} = e^{-n\lambda}\frac{\lambda^{\sum X_i}}{\prod X_i!}$$

$$l_n(\lambda) = -n\lambda + \sum X_i \log\lambda + c.$$

$$\frac{\partial l_n}{\partial \lambda}(\lambda) = -n + \frac{1}{\lambda}\sum X_i = 0 \Rightarrow \hat{\lambda} = \bar{X}_n.$$

$$s(X;\theta) = \frac{\partial}{\partial\theta}\log\left(e^{-\theta}\frac{\theta^X}{X!}\right) = \frac{\partial}{\partial\theta}(-\theta + X\log\theta + c)$$

$$= -1 + \frac{X}{\theta}.$$

$$s'(X;\theta) = -\frac{X}{\theta^2}$$

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

$$\hat{se} = \frac{1}{\sqrt{nI(\hat{\lambda}_n)}} = \sqrt{\frac{\hat{\lambda}_n}{n}}.$$

§9.8 Optimality.

Definition (Asymptotic Relative Efficiency).

$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, t^2)$

$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, u^2)$.

$ARE(U, T) = t^2/u^2$.

Theorem (Asymptotic Optimality of MLE)

If $\hat{\theta}_n$ is MLE and $\tilde{\theta}_n$ is any other estimators, then $ARE(\tilde{\theta}_n, \hat{\theta}_n) \leq 1$.

(pf) $\sqrt{n}(\hat{\theta}_n - \theta) \approx N(0, \frac{1}{I(\hat{\theta}_n)})$

$I(\theta) = -E_\theta(S(X;\theta))$

$S(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}$.

$\sum \log f(x_i;\hat{\theta}_n) \geq \sum \log f(x_i;\theta)$.

§9.9 The Delta Method.

$\tau = g(\theta)$. g : smooth. $\hat{\tau} = g(\hat{\theta})$.

Theorem (The Delta Method)

If $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$,

then $\dfrac{\hat{\tau}_n - \tau}{\hat{se}(\hat{\tau})} \rightsquigarrow N(0,1)$.

where $\hat{\tau}_n = g(\hat{\theta}_n)$, $\hat{se}(\hat{\tau}_n) = |g'(\hat{\theta})| \, \hat{se}(\hat{\theta}_n)$.

Hence, for $C_n = (\hat{\tau}_n - z_{\alpha/2} \hat{se}(\hat{\tau}_n), \hat{\tau}_n + z_{\alpha/2} \hat{se}(\hat{\tau}_n))$,

$P_\theta(\tau \in C_n) \to 1 - \alpha$ as $n \to \infty$.

(pf) By Taylor, $g(\hat{\theta}_n) = g(\theta_*) + (\hat{\theta}_n - \theta_*) g'(\theta_*) + o((\hat{\theta}_n - \theta_*)^2)$

$$\frac{(\hat{\theta}_n - \theta_*)}{\hat{se}} \rightsquigarrow N(0,1).$$

$$\frac{g(\hat{\theta}_n) - g(\theta_*)}{\hat{se}(\hat{\tau})} = \frac{(\hat{\theta}_n - \theta_*) g'(\theta_*) + o((\hat{\theta}_n - \theta_*)^2)}{|g'(\hat{\theta})| \, \hat{se}(\hat{\theta}_n)}$$

$$\rightsquigarrow N(0,1).$$

Example $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$, $\psi = g(p) = \log(p/(1-p))$.

$$\mathcal{L}(p) = p^x (1-p)^{1-x} \quad \ell(p) = x \ln p + (1-x)\ln(1-p)$$

$$S(X;p) = \frac{\partial \ell}{\partial p} = \frac{X}{p} - \frac{1-X}{1-p}$$

$$-S'(X;p) = \frac{X}{p^2} + \frac{1-X}{(1-p)^2} \quad I(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

$$\widehat{se} = \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \quad \text{MLE of } \psi: \hat{\psi} = \log \hat{p}/(1-\hat{p})$$

$$g'(p) = \frac{1-p}{p} \cdot \frac{1-p+p}{(1-p)^2} = \frac{1}{p(1-p)}$$

$$\widehat{se}(\hat{\psi}_n) = |g'(\hat{p}_n)| \widehat{se}(\hat{p}_n) = \frac{1}{\sqrt{n\hat{p}_n(1-\hat{p}_n)}}$$

Example $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ with μ known, σ unknown.

$\psi = \log \sigma$.

$$\mathcal{L}_n(\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(X_i-\mu)^2} \quad \ell_n(\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\frac{\partial \ell}{\partial \sigma} = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum(X_i-\mu)^2 = 0 \quad \sigma = \sqrt{\frac{\sum(X_i-\mu)^2}{n}}$$

$$\log f(X;\sigma) = -\log\sigma - \frac{(X-\mu)^2}{2\sigma^2}$$

$$S(X;\sigma) = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(X-\mu)^2$$

$$s'(X;\sigma) = \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(X-\mu)^2.$$

$$I(\sigma) = \mathbb{E}(-s'(X;\sigma)) = -\frac{1}{\sigma^2} + \frac{3}{\sigma^2} = \frac{2}{\sigma^2}$$

$$\hat{se} = \hat{\sigma}_n / \sqrt{2n}$$

$$\psi = g(\sigma) = \log \sigma. \quad \hat{\psi}_n = \log \hat{\sigma}_n \cdot \quad g' = \frac{1}{\sigma}.$$

$$\Rightarrow \hat{se}(\hat{\psi}_n) = \frac{1}{\hat{\sigma}_n} \frac{\hat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}.$$

§9.10 Multiparameter Models

Note. Let $\theta = (\theta_1, \cdots, \theta_k)$, $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)$.

$$l_n = \sum_{i=1}^{n} \log f(X_i;\theta)$$

$$H_{jk} = \frac{\partial^2 l_n}{\partial\theta_j \partial\theta_k}, \quad I_n(\theta) = -\begin{bmatrix} \mathbb{E}_\theta(H_{11}) & \cdots & \mathbb{E}_\theta(H_{1k}) \\ \vdots & \ddots & \vdots \\ \mathbb{E}_\theta(H_{k1}) & \cdots & \mathbb{E}_\theta(H_{kk}) \end{bmatrix}$$

$$J_n(\theta) = I_n^{-1}(\theta).$$

Theorem. Under appropriate regularity conditions, $\hat{\theta} - \theta \approx N(0, J_n)$

Also, if $\hat{\theta}_j$ is the j^{th} component of $\hat{\theta}$, then

$$\frac{\hat{\theta}_j - \theta_j}{\hat{se}_j} \rightsquigarrow N(0,1),$$

where $\widehat{se}_j^n = J_n(j,j) \cdot Cov(\hat{\theta}_j, \hat{\theta}_k) \approx J_n(j,k)$.

Theorem (Multiparameter Delta Method).

Suppose $\nabla g(\hat{\theta}) \neq 0$. Let $\hat{\tau} = g(\hat{\theta})$. Then,

$\frac{(\hat{\tau} - \tau)}{\hat{se}(\hat{\tau})} \rightsquigarrow N(0,1)$ where

$\hat{se}(\hat{\tau}) = \sqrt{(\hat{\nabla g})^T \hat{J}_n (\hat{\nabla g})}$, $\hat{J}_n = J_n(\hat{\theta}_n)$, $\hat{\nabla} g = \nabla g(\hat{\theta})$.

Example $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. $\tau = g(\mu, \sigma) = \sigma/\mu$.

$$\log f(x; \mu, \sigma) = -\log\sigma - \frac{1}{2\sigma^2}(x-\mu)^2.$$

$$\frac{\partial \log f}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}(x-\mu)^2 \cdot \frac{\partial^2 \log f}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3}{\sigma^4}(x-\mu)^2.$$

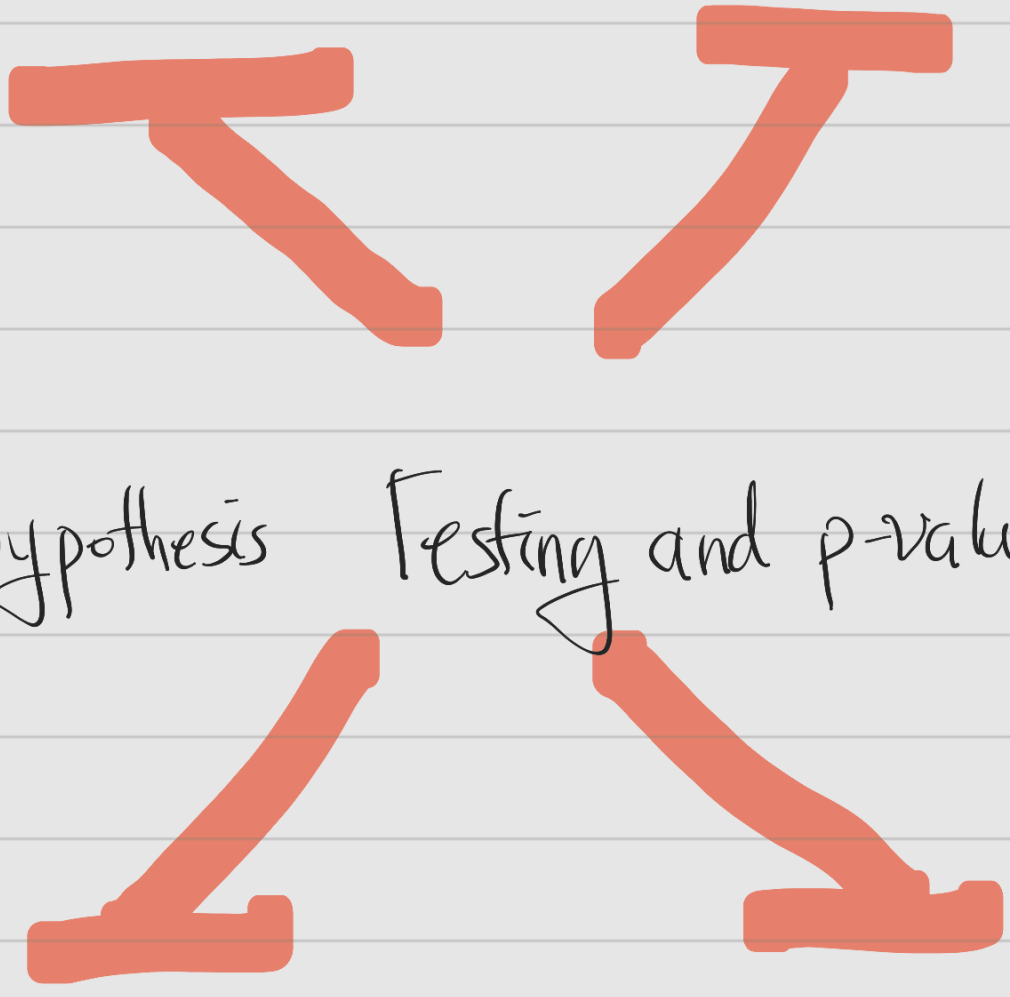
$$\frac{\partial^2 \log f}{\partial \mu \partial \sigma} = \frac{2}{\sigma^3}(\mu - x).$$

$$\frac{\partial \log f}{\partial \mu} = -\frac{1}{\sigma^2}(\mu - x) \cdot \frac{\partial^2 \log f}{\partial \mu^2} = -\frac{1}{\sigma^2}.$$

$$\Rightarrow I(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \Rightarrow I_n(\mu, \sigma) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}.$$

$$J_n = \frac{1}{n}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}. \quad \nabla g = \begin{pmatrix} \frac{\sigma}{-\mu^2} \\ \frac{1}{\mu} \end{pmatrix}. \Rightarrow \hat{se}(\hat{\tau}) = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{\hat{\mu}^4} + \frac{\hat{\sigma}^2}{2\hat{\mu}^2}}.$$

§9.11 The Parametric Bootstrap.


Hypothesis Testing and p-values.

Hypothesis Testing

Partition Θ into Θ_0 and Θ_1.

$H_0: \theta \in \Theta_0$ is the null hypothesis

$H_1: \theta \in \Theta_1$ is the alternative hypothesis.

$\mathcal{X}$: range of X.

$R \subset \mathcal{X}$: rejection region.

$X \in R \longrightarrow$ reject H_0

$X \notin R \longrightarrow$ retain (do not reject) H_0

"Do not reject H_0 unless the evidence is sufficient"

Cases of H.T.

	Retain Null	Reject Null
H_0 true	✓	Type I error ✓
H_1 true	Type II error.	

If try to ↓ type I error, ↑ type II error and vise versa.

Unless increasing sample size.

$R = \{x : T(x) > c\}$. T: test statistic

c: critical value

Definition (Power Function)

The power function of a test with rejection region R is $\beta(\theta) = P_\theta(X \in R)$. → Probability of rejecting H_0 if θ.

The size of a test is $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$ → Largest probability H_0 when H_0 is true

A test is said to have level α if its size $\leq \alpha$,

- Simple hypothesis: Hypothesis of the form $\theta = \theta_0$
  Composite hypothesis: Hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$.

H.T. for simple hypothesis: Two sided test

" composite " : One sided test.

Example $X_1, \dots, X_n \sim N(\mu, \sigma)$ with known σ.

$H_0: \mu \leq 0$, $H_1: \mu > 0$. $\Theta_0 = (-\infty, 0]$, $\Theta_1 = (0, \infty)$.

Test: reject H_0 if $T > c$, with $T = \bar{X}$.

$$\beta(\mu) = P_\mu(\bar{X} > c) = P_\mu\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} > \frac{\sqrt{n}(c-\mu)}{\sigma}\right)$$

$$= P\left(Z > \frac{\sqrt{n}(c-\mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right)$$

$$\text{size} = \sup_{\mu \le 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}\,c}{\sigma}\right).$$

For size α test, $c = \dfrac{\sigma \Phi^{-1}(1-\alpha)}{\sqrt{n}}$.

Thus, reject when $\bar{X} > \sigma\Phi^{-1}(1-\alpha)/\sqrt{n}$.

$$\left(\Leftrightarrow \frac{\sqrt{n}(\bar{X}-0)}{\sigma} > z_\alpha = \Phi^{-1}(1-\alpha)\right).$$

- Test의 size가 α.

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \in \Theta_0} P_\theta(X \in R) = \alpha.$$

Power fn 무슨 의미?

§10.1 The Wald Test.

Definition (The Wald Test)

Consider $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_1$

Assume $\hat{\theta}$ is asymptotically normal:

$$\frac{(\hat{\theta} - \theta_0)}{\hat{se}} \rightsquigarrow N(0,1).$$

The size α Wald test is: reject H_0 when $|W| > z_{\alpha/2}$,

where $W = \dfrac{\hat{\theta} - \theta_0}{\hat{se}}$.

Theorem Asymptotically, Wald test has size α. i.e.

$$P_{\theta_0}(|W| > z_{\alpha/2}) \longrightarrow \alpha \text{ as } n \to \infty.$$

(pf) Under $\theta = \theta_0$, $(\hat{\theta} - \theta_0)/\hat{se} \rightsquigarrow N(0,1)$.

The probability of rejecting when $\theta = \theta_0$ is true is

$$P_{\theta_0}(|W| > z_{\alpha/2}) = P_{\theta_0}\left(\frac{|\hat{\theta} - \theta_0|}{\hat{se}} > z_{\alpha/2}\right)$$

$$\longrightarrow P(|Z| > z_{\alpha/2}) = \alpha.$$

where $Z \sim N(0,1)$

Theorem Suppose the true value of θ is $\theta_* \neq \theta_0$.

The power $\beta(\theta_*)$ - the probability of correctly rejecting the null hypothesis is approximately given by

$$1 - \Phi\left(\frac{\theta_0 - \theta_*}{\hat{se}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_*}{\hat{se}} - z_{\alpha/2}\right).$$

(pf) $\beta(\theta_*) = P_{\theta_*}\left(\left|\frac{\hat{\theta} - \theta_0}{\hat{se}}\right| > z_{\alpha/2}\right)$

$$= P_{\theta_*}\left(\frac{\hat{\theta} - \theta_0}{\hat{se}} > z_{\alpha/2}\right) + P_{\theta_*}\left(\frac{\hat{\theta} - \theta_0}{\hat{se}} < -z_{\alpha/2}\right)$$

$$= P_{\theta_*}\left(\frac{\hat{\theta} - \theta_*}{\hat{se}} > \frac{\theta_0 - \theta_*}{\hat{se}} + z_{\alpha/2}\right) + P_{\theta_*}\left(\frac{\hat{\theta} - \theta_*}{\hat{se}} < \frac{\theta_0 - \theta_*}{\hat{se}} - z_{\alpha/2}\right)$$

$$= 1 - \Phi\left(\frac{\theta_0 - \theta_*}{\hat{se}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_*}{\hat{se}} - z_{\alpha/2}\right)$$

Example (Comparing Two Prediction Algorithm)

Example (Comparing Two Prediction Algorithms)

Two algorithms. Test a predic. alg. 1 on test set size m

2 n

X: nbr of incorrect predictions for alg. 1

Y " 2.

$X \sim Bin(m, p_1)$, $Y \sim Bin(n, p_2)$.

Test null hyp. $p_1 = p_2$.

$H_0: \delta = 0$, $H_1: \delta \neq 0$, $\delta = p_1 - p_2$.

$$\begin{cases} \hat{\delta} = \hat{p}_1 - \hat{p}_2 \\ \hat{se} = \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{m} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n}} \end{cases}$$

size α Wald test reject H_0 when $|W| > z_{\alpha/2}$.

$$W = \frac{\hat{\delta} - 0}{\hat{se}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{m} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n}}}.$$

(Paired Comparison)

If same test used for both alg. $\rightarrow$ Two samples no longer indep.

$X_i = \begin{cases} 1 & \text{alg1 corr. on test case } i \\ 0 & \text{otherwise} \end{cases}$ $Y_i = \begin{cases} 1 & \text{alg2 corr. on test case } i \\ 0 & \text{otherwise.} \end{cases}$

$D_i = X_i - Y_i$.

$\delta = \mathbb{E}[D_i] = \mathbb{E}(X_i) - \mathbb{E}(Y_i) = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1)$

$\hat{\delta} = \bar{D} = n^{-1} \sum_{i=1}^{n} D_i$. $\hat{se}(\hat{\delta}) = S/\sqrt{n}$, $S^2 = n^{-1} \sum_{i=1}^{n} (D_i - \bar{D})^2$.

$W = \hat{\delta} / \hat{se}$.

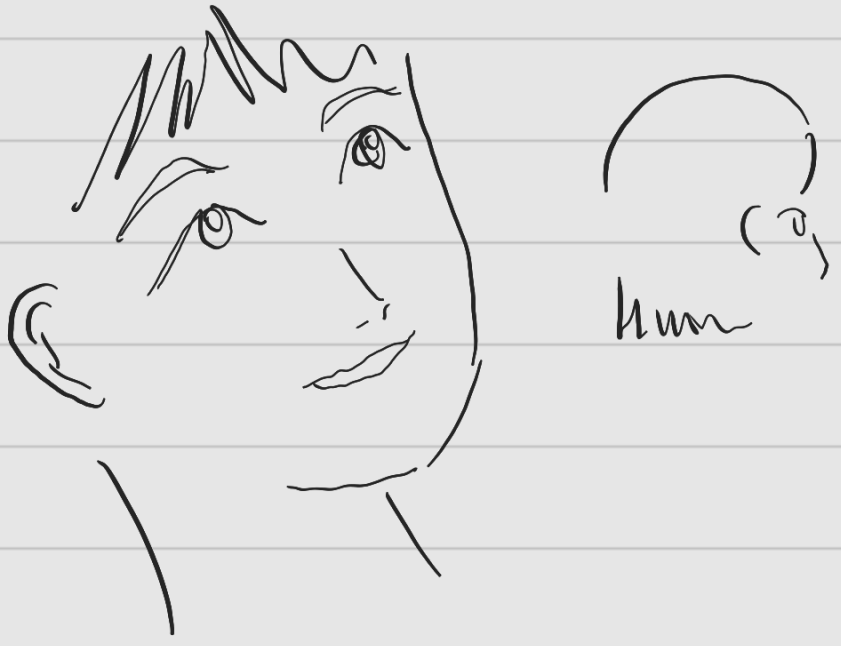
Example (Comparing Two Means)

Let $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$ be two ind. samples from population with means μ_1 and μ_2 respectively.

$H_0: \delta = 0$, $H_1: \delta \neq 0$, $\delta = \mu_1 - \mu_2$.

$\hat{\delta} = \bar{X} - \bar{Y}$, $\hat{se} = \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}$

$$W = \frac{\hat{\delta} - 0}{\hat{se}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}.$$



§10.2 p-values

Definition (p-value)

Suppose given $\alpha \in (0,1)$. We have a size α test with rejection region R_α. Then,

$$p\text{-value} = \inf\{\alpha : T(X^n) \in R_\alpha\}.$$

i.e. p-value is the smallest level at which we can reject H_0.

§10.1

Theorem The size α Wald test rejects $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ with

$$C = (\hat{\theta} - \hat{se}\, z_{\alpha/2},\ \hat{\theta} + \hat{se}\, z_{\alpha/2})$$

Given α, $\exists C$ s.t. $\mathbb{P}(T > C) = \alpha$

$\alpha \uparrow \quad \beta(\theta) \uparrow \quad R \uparrow \quad C \downarrow$

$$\mathbb{P}(T(X^n) \geq C_\alpha) = \alpha$$

Theorem Suppose that the size α test is of the form

reject H_0 iff $T(X^n) \geq C_\alpha$.

Then, $p\text{-value} = \sup_{\theta \in \Theta_0} P_\theta(T(X^n) \geq T(x^n))$,

where x^n is the observed value of X^n. If $\Theta_0 = \{\theta_0\}$, then

$p\text{-value} = P_{\theta_0}(T(X^n) \geq T(x^n))$

(pf) $p\text{-value} := \inf\{\alpha \mid T(X^n) \in R_\alpha\}$

$= \inf\{\alpha \mid T(X^n) \geq C_\alpha\}$

$\beta(\theta) = P_\theta(X \in R)$. $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \in \Theta_0} P_\theta(T(X^n) \geq C_\alpha)$.

↳ holds by def of size of a test.

$p\text{-value} = \inf_\alpha \{\alpha \mid T(X^n) \geq C_\alpha\}$

$= \inf_\alpha \{\alpha = \sup_{\theta \in \Theta_0} P_\theta(T(X^n) \geq C_\alpha) \mid T(x^n) \geq C_\alpha\}$

$= \sup_{\theta \in \Theta_0} P_\theta(T(X^n) \geq C_\alpha)\Big|_{C_\alpha = T(x^n)}$.

$(C_\alpha \uparrow \Rightarrow \alpha \downarrow)$

$= \sup_{\theta \in \Theta_0} P_\theta(T(X^n) \geq T(x^n))$

Theorem Let $w = (\hat{\theta} - \theta_0)/\hat{se}$ be the observed value of Wald statistic W.

p-value $= P_{\theta_0}(|W| > |w|) \approx P(|Z| > |w|) = 2\Phi(-|w|)$

where $Z \sim N(0,1)$.

Theorem If the test statistic has a continuous distribution, then under $H_0: \theta = \theta_0$, p-value has uniform $(0,1)$ distribution.

Hence, if we reject H_0 when p-value $< \alpha$, probability of type I error (reject when H_0 true) is α.

(pf) p-value $\overset{\text{Theorem 10.12}}{=} P_{\theta_0}(T(X^n) \geq T(x^n)) = 1 - F(T(x^n))$

F_T: CDF of $T(X^n)$ under H_0. x^n: observed data $\rightarrow T(x^n)$ has CDF F.

$$\begin{aligned} F_{\text{p-value}}(p) &= P(\text{p-value} < p) \\ &= P(1 - F(T(x^n)) < p) \\ &= P(1-p < F(T(x^n))) \\ &= 1 - P(1-p \geq F(T(x^n))) \\ &= 1 - P(F^{-1}(1-p) \geq T(x^n)) \\ &= 1 - F(F^{-1}(1-p)) = 1 - (1-p) = p. \end{aligned}$$

Thus, p-value $\sim$ Uniform$(0,1)$.

§10.3 χ^2 Distribution

Note χ^2 distribution with k-degree of freedom.

$V \sim \chi_k^2$. if $V = \sum_{i=1}^{k} Z_i^2$ for indep. st.normal $Z_1, \cdots, Z_k$.

$$f(v) = \frac{v^{(k/2)-1} e^{-v/2}}{2^{k/2}\,\Gamma(k/2)}, \text{ for } v>0 \text{ and } E(V)=k,\ V(V)=2k$$

∵) i) For $k=1$,

$$P(V \le v) = P(Z^2 \le v) = P(-v^{1/2} \le Z \le v^{1/2})$$
$$= 1 - \Phi(-v^{1/2}) + \Phi(+v^{1/2})$$
$$f(v) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v} \cdot \left(-\frac{1}{2\sqrt{v}}\right) + \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v} \left(+\frac{1}{2\sqrt{v}}\right)$$
$$= \frac{v^{-\frac{1}{2}}\, e^{-\frac{v}{2}}}{\sqrt{2\pi}}$$

ii) For k,

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$E(V) = \int_0^\infty v f(v)\,dv$$

$$= \int_0^\infty \frac{v^{(k/2)}\, e^{-v/2}}{2^{k/2}\Gamma(k/2)}\,dv$$

$$= \frac{1}{\Gamma(k/2)} \int_0^\infty \left(\frac{v}{2}\right)^{(k/2)+1-1} e^{-v/2}\,dv$$

$$= \frac{1}{\Gamma(k/2)} \int_0^\infty u^{(k/2)+1-1}\, e^{-u} \cdot 2\,du$$

$$= \frac{2\Gamma((k/2)+1)}{\Gamma(k/2)} = 2 \cdot \frac{k}{2} \frac{\Gamma(k/2)}{\Gamma(k/2)} = k.$$

$$E(V^2) = \int_0^\infty v^2 f(v)\,dv$$

$$= \int_0^\infty \frac{v^{(k/2)+1}\, e^{-v/2}}{2^{k/2}\Gamma(k/2)}\,dv$$

$$= \frac{1}{\Gamma(k/2)} \int_0^\infty v \left(\frac{v}{2}\right)^{(k/2)} e^{-v/2}\,dv$$

$$= \frac{2 \cdot \Gamma((k+2)/2)}{\Gamma(k/2)} \int_0^\infty v \frac{v^{((k+2)/2)-1}\, e^{-v/2}}{2^{(k+2)/2}\Gamma((k+2)/2)}\,dv$$

$$= 2 \cdot \frac{k}{2} \cdot E(V_{k+2}) = k(k+2) = k^2 + 2k.$$

Hence, $\mathbb{V}(V) = \mathbb{E}(V^2) - \mathbb{E}(V)^2$

$= k^2 + 2k - k^2 = 2k$.

§10.4 Pearson's χ^2 Test for Multinomial Data

Let $X = (X_1, \cdots, X_k) \sim \text{Multinomial}(n, p)$.

MLE $\hat{p} = (\hat{p}_1, \cdots, \hat{p}_k) = (X_1/n, \cdots, X_k/n)$.

$p_0 = (p_{01}, \cdots, p_{0k})$.

$H_0 : p = p_0$, $H_1 : p \neq p_1$

Definition (Pearson's χ^2 statistic)

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}.$$

where $E_j = \mathbb{E}(X_j) = np_{0j}$ is expected value of X_j under H_0.

Theorem Under H_0, $T \rightsquigarrow \chi^2_{k-1}$.

Thus, the test of rejecting H_0 if $T > \chi^2_{k-1,\alpha}$ has asymptotic level α.

P-value $= \mathbb{P}(\chi^2_{k-1} > t)$ where t is the observed test statistic.

(pf). It suffices to show $\frac{(X_j - E_j)^2}{E_j} \rightsquigarrow \chi^2$.

$X_j = \sum_{i=1}^{n} Y_i$, for $Y_i \sim \text{Binomial}(p_j)$.

Thus, $\mathbb{V}(X_j) = \frac{1}{n} p_{oj}(1 - p_{oj})$, $\mathbb{E}(X_j) = n p_{oj}$.

$$\mathbb{P}\left(\frac{(X_j - E_j)^2}{E_j} \le \alpha^2 \right) = \mathbb{P}\left(\frac{(\sum_{i=1}^{n} Y_i - n p_{oj})^2}{n p_{oj}} \le \alpha^2 \right)$$

$$= \mathbb{P}\left(\frac{\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - p_{oj}\right)^2}{\frac{p_{oj}}{n}} \le \alpha^2 \right)$$

$$T = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j} = \sum_{j=1}^{k} \frac{(X_j - n p_{oj})^2}{n p_{oj}}$$

$\mathbb{P}$

Example (Mendel's Peas).

$\alpha = 0.05$. $\chi^2_3 = 7.815 > \chi^2 = 0.47$.

p-value $= \mathbb{P}(\chi^2_3 > 0.47) = 0.93$

§10.5 The Permutation Test.

For $X_1, \cdots, X_m \sim F_X$, $Y_1, \cdots, Y_n \sim F_Y$,

$H_0: F_X = F_Y$, $H_1: F_X = F_Y$.

T: test statistic. eg:

$T(X_1, \cdots, X_m, Y_1, \cdots, Y_n) = |\bar{X}_m - \bar{Y}_n|$.

$N = m+n$, Permutes $X_1, \cdots, Y_n$ and compute $T \to$ index $T_1, \cdots, T_N$.

P_0: distribution equally $\frac{1}{N!}$ for $T_1, \cdots, T_N$.

t_{obs}: observed value of test statistic

Reject H_0 if $T > t_{obs}$.

$$p\text{-value} = \underbrace{P_0(T > t_{obs})}_{\text{Under } H_0, \text{ probability of rejecting } H_0.} = \frac{1}{N!} \sum_{j=1}^{N!} I(T_j > t_{obs})$$

Example $(X_1, X_2, Y_1) = (1, 9, 3)$, $T(X_1, X_2, Y_1) = |\bar{X} - \bar{Y}| = 2$

permutation	T	probability
(1, 9, 3)	2	1/6
(9, 1, 3)	2	〃
(1, 3, 9)	7	〃

(3,1,9)	7	1/
(3,9,1)	5	1/
(9,3,1)	5	1/

$p\text{-value} = \mathbb{P}(T \geq 2) = 4/6$

§10.6 The Likelihood Ratio Test

Definition (Likelihood Ratio Statistic)

Consider testing $H_0: \theta \in \Theta_0$ vs $H_1: \theta \notin \Theta_0$.

The likelihood ratio statistic is

$$\lambda = 2\log\left(\frac{\sup_{\theta\in\Theta} L(\theta)}{\sup_{\theta\in\Theta_0} L(\theta)}\right) = 2\log\left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)}\right),$$

where $\hat{\theta}$ is the MLE

and $\hat{\theta}_0$ is the MLE when θ is restricted to lie in Θ_0.

Theorem Suppose $\theta = (\theta_1, \cdots, \theta_q, \theta_{q+1}, \cdots, \theta_r)$.

Let $\Theta_0 = \{\theta \mid (\theta_{q+1}, \cdots, \theta_r) = (\theta_{0,q+1}, \cdots, \theta_{0,r})\}$,

Let λ be the likelihood ratio statistic. Under $\theta \in \Theta_0$,

$\lambda(x^n) \rightsquigarrow \chi^2_{r-q,\alpha}$.

where $r = \dim \Theta$, $q = \dim \Theta_0$.

p-value for test is $\mathbb{P}(\chi^2_{r-q} > \lambda)$.

(pf) $\lambda(x^n) = 2 \log\left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)}\right)$.

$\mathbb{P}(\lambda(x^n) > \chi^2_{r-q}) = \mathbb{P}\left(2 \log\left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)}\right) > \chi^2_{r-q}\right)$

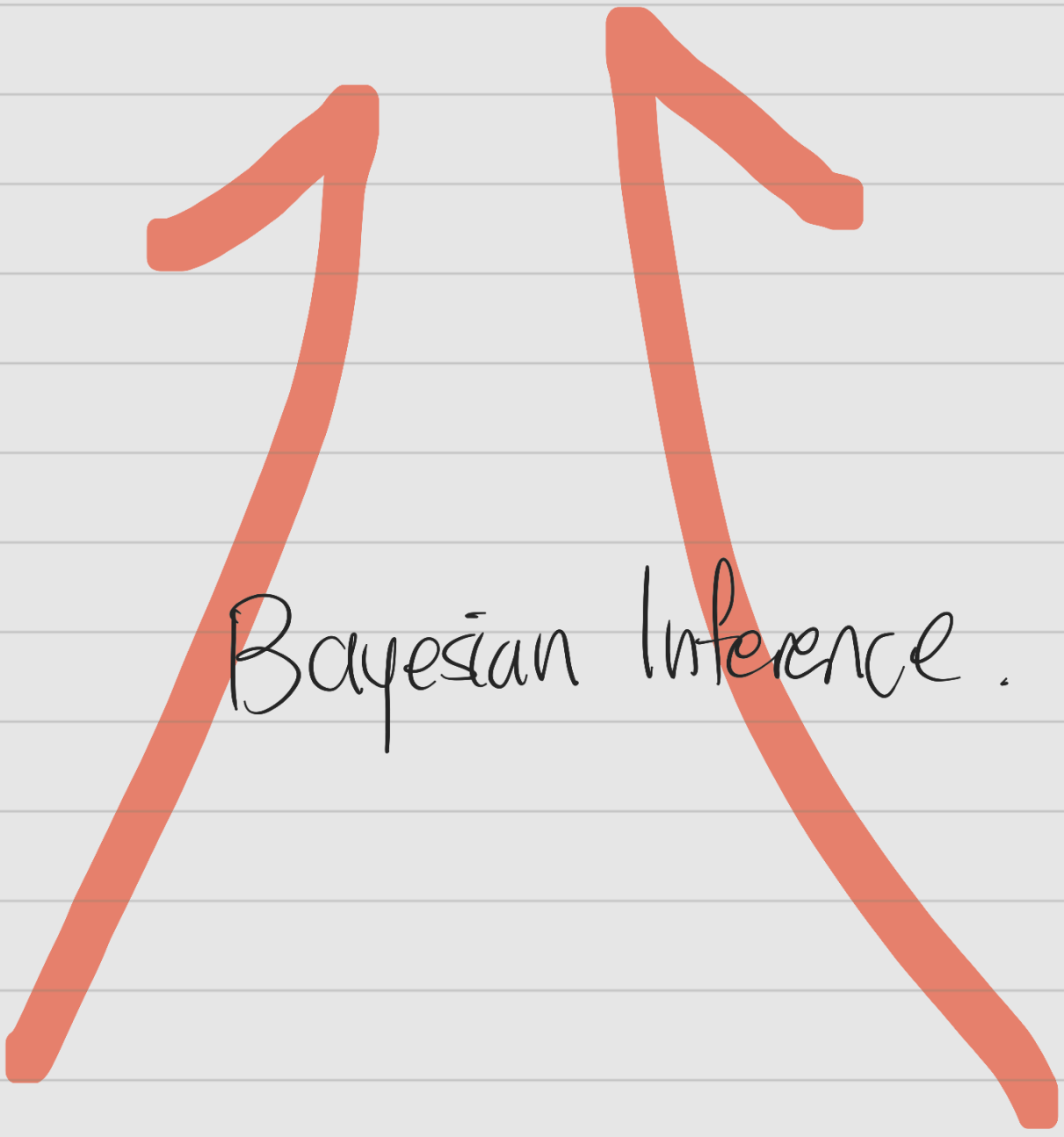
Theorem (Neyman-Pearson)

Suppose we test $H_0: \theta = \theta_0$ versus $H_1: \theta = \theta_1$.

Let $T = \dfrac{L(\theta_1)}{L(\theta_0)} = \dfrac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)}$.

Suppose we reject H_0 when $T > k$. Take k s.t. $\mathbb{P}_{\theta_0}(T > k) = \alpha$. Then this is the most powerful size α test.

(pf


Bayesian Inference.

Note (Postulates on Bayesian Inference)

(B1) Probability describes degree of belief.

(B2) We can make probability statements about parameters.

(B3) Inference about parameter θ.

§11.2 The Bayesian Method

The Bayesian Method

1. $f(\theta)$: prior distribution, express our beliefs about θ before seeing data.
2. $f(x|\theta)$. Belief about x given θ.
3. Posterior distribution $f(\theta|x_1, \cdots, x_n)$.

$$P(\Theta=\theta|X=x) = \frac{P(X=x|\Theta=\theta)P(\Theta=\theta)}{\sum_{\theta} P(X=x|\Theta=\theta)P(\Theta=\theta)}$$

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

$$f(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = L(\theta)$$

Notation $X^n = (X_1, \cdots, X_n)$, $x^n = (x_1, \cdots, x_n)$

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{c_n} \propto \mathcal{L}_n(\theta)f(\theta).$$

$$c_n = \int \mathcal{L}_n(\theta)f(\theta)d\theta.$$

→ Posterior is proportional to Likelihood times prior:

$$f(\theta|x^n) \propto \mathcal{L}(\theta)f(\theta).$$

Definition (Posterior Mean / Bayesian Interval Estimate).

$$\bar{\theta}_n = \int \theta f(\theta|x^n)d\theta = \frac{\int \theta \mathcal{L}_n(\theta)f(\theta)d\theta}{\int \mathcal{L}_n(\theta)f(\theta)d\theta}.$$

Take a, b s.t. $\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_{b}^{\infty} f(\theta|x^n)d\theta = \alpha/2$. $C = (a, b)$.

$$P(\theta \in C|x^n) = \int_a^b f(\theta|x^n)d\theta = 1-\alpha.$$

Example $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. Prior: uniform $f(p) = 1$. ①

By Bayes' $f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^s(1-p)^{n-s} = p^{s+1-1}(1-p)^{n-s+1-1}$

for $s = \sum_{i=1}^{n} x_i$.

$$f(p|x^n) = \frac{\Gamma(n+2)}{\Gamma(S+1)\Gamma(n-S+1)} p^{(S+1)-1}(1-p)^{(n-S+1)-1}.$$

$p|x^n \sim \text{Beta}(S+1, n-S+1)$

Bayes estimator: $\bar{p} = \frac{S+1}{n+2} = \lambda_n \hat{p} + (1-\lambda_n)\tilde{p}$,

with MLE $\hat{p} = \frac{S}{n}$,

prior mean $\tilde{p} = \frac{1}{2}$.

$\lambda_n = \frac{n}{n+2} \approx 1$.

95% posterior interval (a,b) with $\int_a^b f(p|x^n)\,dp = 0.95$.

② Prior $p \sim \text{Beta}(\alpha, \beta)$

$$f(p|x^n) \propto f(p)\mathcal{L}(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \cdot p^S(1-p)^{n-S}$$

$$\propto p^{\alpha+S+1}(1-p)^{\beta+n-S-1}$$

$p|x^n \sim \text{Beta}(\alpha+S, \beta+n-S)$ $\quad \bar{p} = \frac{\alpha+S}{\alpha+\beta+n}$

③ Flat prior $p \sim \text{Beta}(1,1)$

° When prior and the posterior are in the same family, we say the prior is conjugate with respect to the model.

Example $X_1, \cdots, X_n \sim N(\theta, \sigma^2)$, with known σ.

prior: $\theta \sim N(a, b^2)$.

$$f(\theta|x^n) = \frac{f(x^n|\theta) f(\theta)}{\int f(x^n|\theta) f(\theta) d\theta}$$

$$= \frac{\prod_{i=1}^n f(x_i|\theta) \cdot \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2b^2}(\theta-a)^2}}{\int \prod_{i=1}^n \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2}\right) \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2b^2}(\theta-a)^2} d\theta}$$

$$\propto \frac{1}{\sigma b^n} e^{-\left\{\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2 + \frac{1}{2b^2}(\theta-a)^2\right\}}$$

$\propto$

$$\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2 + \frac{1}{2b^2}(\theta-a)^2$$

$$= \frac{1}{2}\left(\frac{n}{\sigma^2}(\theta-\bar{x})^2 + \frac{1}{b^2}(\theta-a)^2\right) + k$$

$$= \frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)\left(\theta - \left(\frac{\frac{1}{se^2}}{\tau^2}\bar{x} + \frac{\frac{1}{b^2}}{\tau^2}a\right)\right)^2\right) + k'$$

$$= \frac{1}{2}\frac{1}{\tau^2}(\theta-\bar{\theta})^2 + k', \text{ with } \bar{\theta} = w\bar{x} + (1-w)a,$$

where $\frac{1}{\tau^2} = \frac{1}{se^2} + \frac{1}{b^2}$, $w = \frac{\frac{1}{se^2}}{\frac{1}{\tau^2}}$, $se = \sigma/\sqrt{n}$

i.e. $\theta|X^n \sim N(\bar{\theta}, \tau^2)$

$w \to 1$, $\tau/se \to 1$ as $n \to \infty$.

Hence, posterior is approximately $N(\hat{\theta}, se^2)$.

§ 11.3 Functions of Parameters.

If $\tau = g(\theta)$, the posterior CDF of τ is

$$H(\tau|x^n) = P(g(\theta) \le \tau | x^n) = \int_A f(\theta|x^n)d\theta,$$

where $A = \{\theta : g(\theta) \le \tau\}$. The posterior density is $h(\tau|x^n) = H'(\tau|x^n)$.

Example $X_1, \cdots, X_n \sim Bernoulli(p)$, $f(p) = 1$. $\Rightarrow p|X^n \sim Beta(s+1, n-s+1)$

for $s = \sum_{i=1}^n x_i$. Let $\psi = \log(p/(1-p))$. Then,

$$\begin{aligned} H(\psi|x^n) &= P(\Psi \le \psi | x^n) \\ &= P\left(\log\left(\frac{p}{1-p}\right) \le \psi | x^n\right) \\ &= P\left(p \le \frac{e^\psi}{1+e^\psi} | x^n\right) \\ &= \int_0^{e^\psi/(1+e^\psi)} f(p|x^n)dp \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \int_0^{e^\psi/(1+e^\psi)} p^s(1-p)^{n-s}dp \end{aligned}$$

$$\Rightarrow h(\psi|x^n) = H'(\psi|x^n)$$

$$= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \left(\frac{\partial\left(\frac{e^\psi}{1+e^\psi}\right)}{\partial\psi}\right)$$

$$\frac{e^\psi(1+e^\psi)}{(e^\psi)^2}$$

$$= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \frac{e^\psi}{(1+e^\psi)^2}$$

$$= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^{s+1} \left(\frac{1}{1+e^\psi}\right)^{n-s+1}.$$

for $\psi \in \mathbb{R}$

§11.4 Simulation

§11.5 Large Sample Properties of Bayes' Procedures

$$(1-\frac{1}{n})^n = e^{-1}$$

$$(1-c^2\frac{\log n}{n})$$

Example - Bernoulli(p) $\rightarrow$

$$I(p) = \frac{1}{p(1-p)}.$$

$$f(p) \propto \sqrt{I(p)} = p^{-1/2}(1-p)^{-1/2}$$

$\Rightarrow p \sim Beta(1/2, 1/2)$, close to uniform

11.7 §Multiparameter Problem

$\theta = (\theta_1, \cdots, \theta_p)$.

$f(\theta|x^n) \propto L_n(\theta) f(\theta)$.

Marginal posterior distribution for θ_i:

$$f(\theta_1|x^n) = \int \cdots \int f(\theta_1, \cdots, \theta_p|x^n)\, d\theta_2 \cdots d\theta_p$$

Example (Comparing Two Binomials).

$f(p_1, p_2) = 1$.

$f(p_1, p_2|x_1, x_2) \propto p_1^{x_1}(1-p_1)^{n_1-x_1} p_2^{x_2}(1-p_2)^{n_2-x_2}$.

$f(p_1|x_1) \propto p_1^{x_1}(1-p_1)^{n_1-x_1}$, $f(p_2|x_2) \propto p_2^{x_2}(1-p_2)^{n_2-x_2}$.

$f(p_1, p_2|x_1, x_2) = f(p_1|x_1) f(p_2|x_2)$.

$\Rightarrow$ p_1, p_2 are independent under the posterior.

$p_1|x_1 \sim \text{Beta}(x_1+1, n_1-x_1+1)$

$p_2|x_2 \sim \text{Beta}(x_2+1, n_2-x_2+1)$.

12

Statistical Decision Theory

§12.1 Preliminaries

Note (Loss function)

Loss ftn measure discrepancy between θ and $\hat{\theta}$.

$$L(\theta,\hat{\theta}) = \begin{cases} (\theta-\hat{\theta})^2 & \text{(squared error loss)} \\ |\theta-\hat{\theta}| & \text{(absolute loss)} \\ |\theta-\hat{\theta}|^p & (L_p \text{ loss}) \\ 0 \text{ if } \theta=\hat{\theta},\ 1 \text{ if } \theta\neq\hat{\theta} & \text{(zero-one loss)} \\ \int \log\left(\frac{f(x;\theta)}{f(x;\hat{\theta})}\right) f(x;\theta)dx & \text{(KL-loss)} \end{cases}$$

Definition (Risk of an Estimator)

The risk of an estimator $\hat{\theta}$ is

$$R(\theta,\hat{\theta}) = \mathbb{E}_\theta(L(\theta,\hat{\theta})) = \int L(\theta,\hat{\theta}(x)) f(x;\theta)dx.$$

When the loss ftn is squared-error, the risk is MLE and so

$$R(\theta,\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}-\theta)^2 = MSE = V_\theta(\hat{\theta}) + \text{bias}^2_\theta(\hat{\theta}).$$

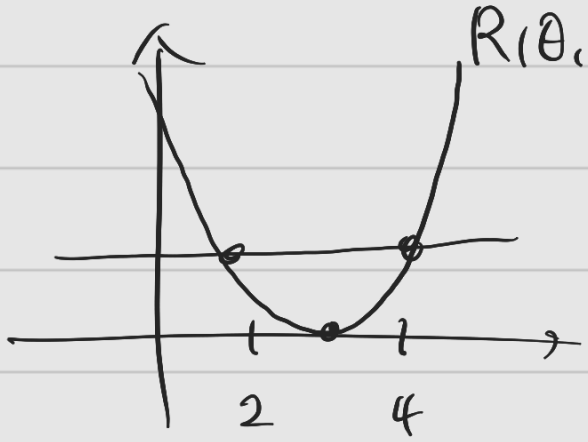
- For rest of chapter, if not mentioned, loss ftn is squared error

§12.2 Comparing Risk Functions

Example $X \sim N(\theta, 1)$, with squared error loss.

$\hat{\theta}_1 = X$, $\hat{\theta}_2 = 3$. $R(\theta, \hat{\theta}_1) = \mathbb{E}_\theta(X-\theta)^2 = 1$

$R(\theta, \hat{\theta}_2) = \mathbb{E}_\theta(3-\theta)^2 = (3-\theta)^2$.



Neither estimator uniformly dominates the other.

Example $X_1, \dots, X_n \sim \text{Bernoulli}(p)$

Let $\hat{p}_1 = \bar{X}$. As $\hat{p}_1$ is unbiased,

$R(p, \hat{p}_1) = V(\bar{X}) = \frac{p(1-p)}{n}$, $Y = \sum_{i=1}^n X_i$

$\hat{p}_2 = \frac{Y+\alpha}{\alpha+\beta+n}$, with $Y = \sum_{i=1}^n X_i$, α, β: constants.

(Posterior mean using $\text{Beta}(\alpha, \beta)$ prior).

↳?

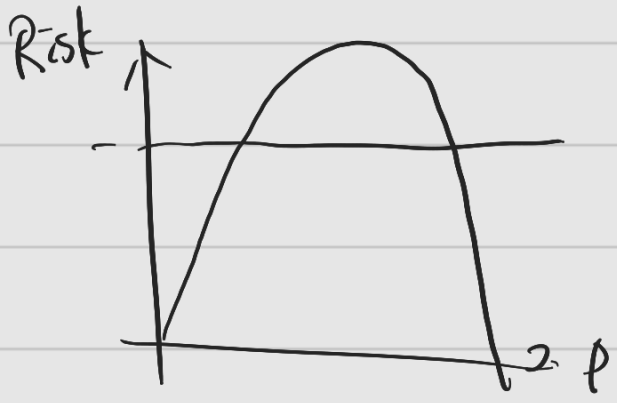
$R(p, \hat{p}_2) = V_p(\hat{p}_2) + (\text{bias}_p \hat{p}_2)^2$

$= V_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2$

$$= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2 .$$

Let $\alpha = \beta = \frac{\sqrt{n}}{4}$. then, $\hat{p}_2 = \frac{Y+\sqrt{n}/4}{n+\sqrt{n}}$

$$R(p, \hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2} \text{ (constant).}$$



Neither estimator uniformly dominates the other.

Definition (Maximum Risk)

The maximum risk is $\bar{R}(\hat{\theta}) = \sup_\theta R(\theta, \hat{\theta})$

The Bayes risk is $r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta) d\theta$,

where $f(\theta)$ is a prior for θ.

Example On previous Bernoulli example,

$$\bar{R}(\hat{p}_1) = \max_{0 \le p \le 1} \frac{p(1-p)}{n} = \frac{1}{4n}.$$

$$\bar{R}(\hat{p}_2) = \max_p \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on max risk, $\bar{R}(p_2) < \bar{R}(p_1)$. $\rightarrow p_2$ is a better estimator.

However, for large n, $\hat{p}_1$ has smaller risk except for a small region near $p = 1/2$.

Bayes Risk, Take $f(p) = 1$.

$$r(f, \hat{p}_1) = \int R(p, \hat{p}_1)\, dp = \int \frac{p(1-p)}{n}\, dp = \frac{1}{6n}.$$

$$r(f, \hat{p}_2) = \int R(p, \hat{p}_2)\, dp = \frac{n}{4(n+\sqrt{n})^2}.$$

For $n \geq 20$, $r(f, \hat{p}_2) > r(f, \hat{p}_1)$, suggesting $\hat{p}_1$ is a better estimator.

Definition (Bayes Rule / Minimax Rule)

A decision rule that minimizes the Bayes risk is a Bayes rule. $\hat{\theta}$ is a Bayes rule wrt prior f if $r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta})$, infimum taken over all estimators $\tilde{\theta}$.

An estimator that minimizes the maximum risk is called a

minimax rule. $\hat{\theta}$ is minimax if $\sup_\theta R(\theta,\hat{\theta}) = \inf_{\tilde{\theta}} \sup_\theta R(\theta,\tilde{\theta})$

§12.3 Bayes Estimator

f: prior. By Bayes' thm, $f(\theta|x) = \frac{f(x|\theta)f(\theta)}{m(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$

$m(x) = \int f(x,\theta)d\theta = \int f(x|\theta)f(\theta)d\theta$: marginal distribution.

$r(\hat{\theta}|x) = \int L(\theta,\hat{\theta}(x))f(\theta|x)d\theta$: posterior risk.

Theorem $r(f,\hat{\theta}) = \int r(\hat{\theta}|x)m(x)dx$.

($r(f,\hat{\theta})$: Bayes' risk.)

Let $\hat{\theta}(x)$ be the minimizer of $r(\hat{\theta}|x)$. Then, $\hat{\theta}$ is the Bayes estimator

(pf) $r(f,\hat{\theta}) = \int R(\theta,\hat{\theta})f(\theta)d\theta = \int\left(\int L(\theta,\hat{\theta}(x))f(x|\theta)dx\right)f(\theta)d\theta$

$= \int\int L(\theta,\hat{\theta}(x))f(x,\theta)\,dx\,d\theta$

$= \int\int L(\theta,\hat{\theta}(x))f(\theta|x)m(x)dx\,d\theta$

$$= \int \left(\int L(\theta, \hat{\theta}(x)) f(\theta|x) d\theta \right) m(x) dx$$

$$= \int r(\hat{\theta}|x) m(x) dx$$

If $\hat{\theta}(x)$ minimize $r(\hat{\theta}|x)$, then minimize $r(\hat{\theta}|x)\, \forall x$.

Thus, minimize $\int r(\hat{\theta}(x)) m(x) dx$.

Theorem If $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then the Bayes estimator is

$$\hat{\theta}(x) = \int \theta f(\theta|x) d\theta = E(\theta | X = x).$$

If $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, then the Bayes' estimator is the median of the posterior $f(\theta|x)$.

If $L(\theta, \hat{\theta})$ is zero-one loss, the Bayes estimator is the mode of the posterior $f(\theta|x)$ (?)

(pf) $\hat{\theta}(x)$ minimize $r(\hat{\theta}|x) = \int (\theta - \hat{\theta}(x))^2 f(\theta|x) d\theta$.

$$\frac{\partial}{\partial \hat{\theta}} r(\hat{\theta}|x) = \int 2(\hat{\theta}(x) - \theta) f(\theta|x) d\theta = 0.$$

$$\Rightarrow \hat{\theta}(x) = \int \theta f(\theta|x) d\theta = E(\theta | X = x).$$

$$\frac{\partial}{\partial\hat\theta} r(\hat\theta|x) = \int \frac{\partial}{\partial\hat\theta} |\theta - \hat\theta(x)| f(\theta|x) d\theta = 0.$$

사람 사랑 인간

Example $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known.

$N(a, b^2)$: prior for μ.

$$\mu | x^n \sim \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2b^2}(\mu-a)^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu)^2}$$

$$\hat\theta(x^{(n)}) = \int \theta f(\theta|x^{(n)}) d\theta$$

$$= w\bar{X} + (1-w)a, \quad w = \frac{\frac{1}{se^2}}{\frac{1}{se^2} + \frac{1}{b^2}}. \quad se = \frac{\sigma}{\sqrt{n}}.$$

(Example 11.2)

$$w = \frac{\frac{1}{se^2}}{\frac{se^2+b^2}{se^2 b^2}} = \frac{b^2}{se^2+b^2} = \frac{b^2}{b^2+\frac{\sigma^2}{n}}.$$

$$1-w = \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}.$$

§12.4 Minimax Rules

Theorem Let $\hat{\theta}^f$ be the Bayes rule for some prior f:

$$r(f, \hat{\theta}^f) = \inf_{\theta} r(f, \hat{\theta})$$

$$\left(r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta) d\theta \right)$$

Suppose $R(\theta, \hat{\theta}^f) \le r(f, \hat{\theta}^f) \ \forall \theta$.

then, $\hat{\theta}^f$ is minimax and f is called a least favorable prior.

(pf) $\bar{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta})$

$$R(\theta, \hat{\theta}^f) \le r(f, \hat{\theta}^f) \ \forall \theta$$

$$\Rightarrow \bar{R}(\hat{\theta}^f) \le r(f, \hat{\theta}^f)$$

$$r(f, \hat{\theta}^f) = \int R(\theta, \hat{\theta}^f) f(\theta) d\theta \le \int \bar{R}(\hat{\theta}^f) f(\theta) d\theta = \bar{R}(\hat{\theta}^f).$$

Thus, $\bar{R}(\hat{\theta}^f) = r(f, \hat{\theta}^f)$

For any $\hat{\theta}$,