36-705 07/08/2020

Lecture Notes 9 - Asymptotic Theory - Review

· consistency: $\hat{\theta}_n \xrightarrow{P} \theta$ · Relation between estimator and parameter as more data is collected

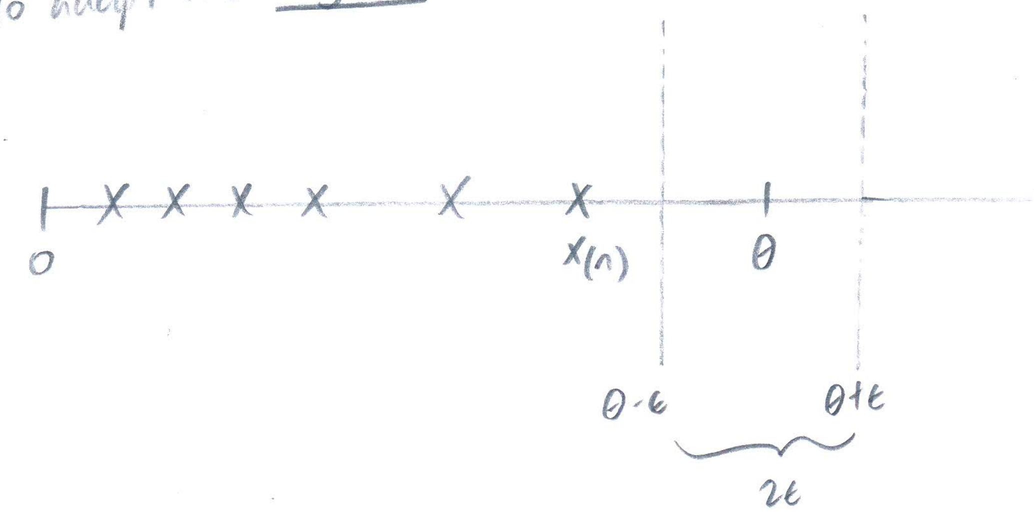
(property of an estimator) $\hat{\theta}_n - \theta = o_P(1)$

· Asymptotic Normality: $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \tau^2)$ · CLT style result for MLE in particular.

· 2 ways of showing consistency

example 2

- clarification of diagram :- $X_1, \dots, X_n \sim \text{Unif}(0, \theta)$
- Recall E7 in (N4) $\hat{\theta}_{n,MLE} = X_{(n)} = \max\{X_1, \dots, X_n\}$
- Gives similar example $\hat{\theta}_n \xrightarrow{P} \theta$ · really tripped up on this.
  but for
  $X_1, \dots, X_n \sim \text{Unif}(0,1)$ and previous proof/example.
· Going to adapt the diagram (which is clear)



· following previous:-

$$P(|\hat{\theta}_n - \theta| > \epsilon) = P(|X_{(n)} - \theta| > \epsilon)$$
$$= P(X_{(n)} - \theta > \epsilon) + P(-(X_{(n)} - \theta) > \epsilon)$$
$$= \underbrace{P(X_{(n)} > \theta + \epsilon)}_{=0 \text{ as } X_i \sim \text{Unif}(0,\theta)} + P(X_{(n)} < \theta - \epsilon)$$

· so $P(|X_{(n)} - \theta| > \epsilon) = P(X_{(n)} < \theta - \epsilon) = P(\text{all } X_i < \theta - \epsilon) = \prod_{i=1}^{n} P(X_i < \theta - \epsilon)$

- And hence,

$$P(|X_{(n)}-\theta|>\epsilon) = \prod_{i=1}^{n} P(X_i<\theta-\epsilon) = (\theta-\epsilon)^n \xrightarrow{n\to\infty} 0$$

- Hence $\hat{\theta}_{n,MLE} = X_{(n)} \xrightarrow{P} \theta$ and estimator using last orde statistic for is consistent

---

(A5) - check convegence in prob statement of $\hat{\sigma}^2$.

(b): spent a while on this:

- That is, we want to show:-

$$\sigma_n^2 \xrightarrow{P} \frac{\sigma^2}{2} \qquad \text{where} \quad \sigma_n^2 = \sum_{i=1}^{n}\sum_{j=1}^{2} \frac{(Y_{ij}-\bar{Y}_i)^2}{2n}$$

- Before this, show σ_n^2 is MLE:- of σ^2

- MLE for μ_i $\quad \hat{\mu}_i = \dfrac{Y_{i1}+Y_{i2}}{2}$

- Likelihood:

$$L(\sigma^2,\mu) = \prod_{i=1}^{n}\left(p(Y_{i1};\mu_i,\sigma^2)\,p(Y_{i2};\mu_i,\sigma^2)\right)$$

$$= \prod_{i=1}^{n}\left(\frac{1}{2\pi\sigma^2}\exp\left\{\frac{-[(Y_{i1}-\mu_i)^2+(Y_{i2}-\mu_i)^2]}{2\sigma^2}\right\}\right)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_{i1}-\mu_i)^2+(Y_{i2}-\mu_i)^2\right\}$$

LL

$$l(\sigma^2,\mu) = -n\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(Y_{i1}-\mu_i)^2+(Y_{i2}-\mu_i)^2\right]$$

To maximise l, estimate μ_i using $\hat{\mu}_i$ (MLE); take derivative of l wrt σ^2 and solve to get $\hat{\sigma}_n^2$.

$\cdot \frac{\partial}{\partial \sigma^2} l(\sigma^2, \hat{\mu}) = \frac{\partial}{\partial \sigma^2}\left[-n\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(Y_{i1}-\hat{\mu}_i)^2 + (Y_{i2}-\hat{\mu}_i)^2\right]\right]$

$= -n\left(\frac{1}{2\pi\sigma^2}\right)2\pi + \frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^{n}\left[(Y_{i1}-\hat{\mu}_i)^2 + (Y_{i2}-\hat{\mu}_i)^2\right]$

$= \frac{-n}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left[(Y_{i1}-\hat{\mu}_i)^2 + (Y_{i2}-\hat{\mu}_i)^2\right] = 0$

$\Rightarrow -n\sigma^2 + \frac{1}{2}\sum_{i=1}^{n}\left[(Y_{i1}-\hat{\mu}_i)^2 + (Y_{i2}-\hat{\mu}_i)^2\right] = 0$

$\Rightarrow \hat{\sigma}_n^2 = \frac{1}{2n}\sum_{i=1}^{n}\left[(Y_{i1}-\hat{\mu}_i)^2 + (Y_{i2}-\hat{\mu}_i)^2\right]$

$\blacksquare$

$\cdot\ \hat{\sigma}_n^2 \xrightarrow{P} \frac{\sigma^2}{2}$ (i.e. $\hat{\sigma}_n^2$ is <u>not</u> a <u>consistent</u> estimator of σ^2)

. <u>strategy</u>:- i) compute $\mathbb{E}[\hat{\sigma}_n^2]$

ii) use WLLN.

. consider:-

$\mathbb{E}[\hat{\sigma}_n^2] = \sum_{i=1}^{n}\sum_{j=1}^{2}\frac{\mathbb{E}[(Y_{ij}^2 - 2Y_{ij}\bar{Y}_i + \bar{Y}_i^2)]}{2n}$

$\cdot\ \mathbb{E}[Y_{ij}^2] = var[Y_{ij}] + \mathbb{E}[Y_{ij}]^2 = (\sigma^2 + \mu_i^2)$

$\cdot\ -2\mathbb{E}[Y_{ij}\bar{Y}_i] = -2\mathbb{E}\left[Y_{ij}\frac{(Y_{i1}+Y_{i2})}{2}\right] = -\mathbb{E}\left[Y_{ij}^2 + Y_{i1}Y_{i2}\right] = -\left[\mathbb{E}[Y_{ij}^2] + \underbrace{\mathbb{E}[Y_{i1}]\mathbb{E}[Y_{i2}]}_{Y_{i1}\ Y_{i2}\ \text{IID.}}\right]$

$= -\left[(\sigma^2 + \mu_i^2) + \mu_i^2\right] = -(\sigma^2 + 2\mu_i^2)$

$\cdot\ \mathbb{E}[\bar{Y}_i^2] = var[\bar{Y}_i] + \mathbb{E}[\bar{Y}_i]^2 = \frac{\sigma^2}{2} + \mu_i^2$

- Hence;

$$\mathbb{E}[\hat{\sigma}_n^2] = \sum_{i=1}^{n}\sum_{j=1}^{2} \frac{1}{2n}\left\{(\sigma^2+\mu_i)^2 - (\sigma^2+2\mu_i^2) + \left(\frac{\sigma^2}{2}+\mu_i^2\right)\right\}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{2} \frac{1}{2n}\left(\frac{\sigma^2}{2}\right) = \sum_{i=1}^{n}\sum_{j=1}^{2} \frac{\sigma^2}{4n} = \sum_{i=1}^{n} \frac{\sigma^2}{2n} = \frac{n\sigma^2}{2n} = \frac{\sigma^2}{2}$$

- Via WLLN; $\hat{\sigma}_n^2 \xrightarrow{P} \mathbb{E}[\hat{\sigma}_n^2]$
- Hence $\hat{\sigma}_n^2 \xrightarrow{P} \frac{\sigma^2}{2}$ ■

- consistency of MoM estimators
- one parameter case:-

$$\mu(\hat{\theta}) = m = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Assuming μ^{-1} exists and is continuous

Then $\hat{\theta} = \mu^{-1}(m)$

- Via WLLN: $m \xrightarrow{P} \mu(\theta)$
- via C.M:- $\hat{\theta}_n = \mu^{-1}(m) \xrightarrow{P} \mu^{-1}(\mu(\theta)) = \theta$

---

4. consistency of MLE

(*) under regularity conditions, MLE is consistent

- prove in special case $\longrightarrow$ reveal MLE and Hellinger distance
- model $\rightarrow$ finitely many densities, distinct:- $\mathcal{P} = \{p_0, p_1, \dots, p_N\}$

$$L(p_j) = \prod_{i=1}^{n} p_j(X_i)$$

- MLE $\hat{p}$ is density p_j that maximises $L(p_j)$; wlog p_0 is true density

Theorem 3

$$P(\hat{p} \neq p_0) \longrightarrow 0 \quad \text{as } n \to \infty$$

Proof

- define $\epsilon_j = h(p_0, p_j)$ i.e. Hellinger distance between MLE $\hat{p}$ and true density p_0
- recall $h(p_0, p_j) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$
- for $j \neq 0$ (i.e. any density other than the true density p_0)
- strategy: - upper bound probability of likelihood ratio / ratio of likelihoods to get $P\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right) \le e^{-n\epsilon_j^2/2}$

(?)

O/S 1: Many steps of this proof not understood.
- cleared many of these.

- consider:

$$P\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right) = P\left(\prod_{i=1}^{n} \frac{p_j(X_i)}{p_0(X_i)} > e^{-n\epsilon_j^2/2}\right)$$

- typo

$$= P\left(\prod_{i=1}^{n} \sqrt{\frac{p_j(X_i)}{p_0(X_i)}} > e^{-n\epsilon_j^2/4}\right)$$

Markov ✓

$$\le e^{n\epsilon_j^2/4}\, \mathbb{E}\left(\prod_{i=1}^{n} \sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\right)$$

✓

independence of $\sqrt{\frac{p_j(X_i)}{p_0(X_i)}}$ as X_i IID.

$$= e^{-n\epsilon_j^2/4} \prod_{i=1}^{n} \mathbb{E}\left(\sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\right)$$

(i)

- see LN2 for sim trick.

$$= e^{n\epsilon_j^2/4} \left(\int \sqrt{p_j\, p_0}\right)^n$$

prop of $h(p_0, p_q)$

$$= e^{n\epsilon_j^2/4} \left(1 - \frac{h^2(p_0, p_j)}{2}\right)^n = e^{n\epsilon_j^2/4} \left(1 - \frac{\epsilon_j^2}{2}\right)^n$$

$$= e^{n\epsilon_j^2/4} \exp \log \left(1 - \frac{\epsilon_j^2}{2}\right)^n$$

✓

$$= e^{n\epsilon_j^2/4} \exp\left\{ n \log\left(1 - \frac{\epsilon_j^2}{2}\right)\right\}$$

$\log(1-x) \leq -x$

$$\leq e^{n\epsilon_j^2/4} e^{-n\epsilon_j^2/2} = e^{-n\epsilon_j^2/2}$$

Hence $P\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right) \leq e^{-n\epsilon_j^2/2}$

- used $h^2(p_0, p_j) = 2 - 2\int\sqrt{p_0 p_j}$
- define $\epsilon = \min\{\epsilon_1, \dots, \epsilon_N\}$

$$P(\hat{p} \neq p_0) \leq P\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2} \text{ for some } j\right)$$

$P(\bigcup_i B_i) \leq \sum_i P(B_i)$

$$\leq \sum_{j=1}^{N} P\left(\frac{L(p_j)}{L(p_0)} > e^{-n\epsilon_j^2/2}\right)$$

$$\leq \sum_{j=1}^{N} e^{-n\epsilon_j^2/2} \leq N e^{-n\epsilon^2/2} \xrightarrow{n\to\infty} 0$$

∎

(i) This step involves taking expectation wrt <u>true density</u> $p_0(X_i)$

$$\mathbb{E}_{p_0(X_i)}\left[\sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\right] = \int \sqrt{\frac{p_j(X_i)}{p_0(X_i)}}\, p_0(X_i)\, dX_i$$

$$= \int \sqrt{\frac{p_j(X_i)}{p_0(X_i)}} \cdot \sqrt{p_0^2(X_i)}\, dX_i$$

$$= \int \sqrt{p_j(X_i)\, p_0(X_i)}\, dX_i$$

Theorem 4

- (?) confused about $P\left(\frac{1}{n}(\ell(\theta_0) - \ell(\theta)) > 0\right) \to 1$

O/S2 - in particular, how this relates to convergence in probability

- not a full proof of consistency of MLE, but heuristic
- KL divergence and MLE related.
- It is uniform convergence that is required for consistency of MLE ($\hat{\theta}_{MLE} \xrightarrow{P} \theta$)
- O/S3: relate to uniform convergence

6. Asymptotic normality

- Focus on MLE.
- Asymptotic normality of MLE allows for approx. of probability statements about MLE using normal distns.
- going to include some wordy ways of remembering definitions.
- score function $S_n(\theta)$; function of data and params $S_n(\theta, X_1, X_2, \dots, X_n) =$ - derivative of log-like.

$I_n(\theta) = \mathrm{Var}_\theta(S_n(\theta))$

- Fisher info $I_n(\theta)$ - variance of score function; function of param

(?) trickiness of this part was the multitude of relationships covered.

- going to briefly state main relations

- integral/expect wrt to joint distn of the data $p(x^n; \theta)$

- $\mathrm{Var}(\hat{\theta}_{MLE}) \approx \frac{1}{I_n(\theta)}$ asymptotic (?) - variance of MLE is inverse of Fisher info.

- $\mathbb{E}_\theta[S_n(\theta)] = 0$ (under reg conditions, counterexample → uniform) - same value of θ. - exp. of score fn is 0 with params 0

- $I_n(\theta) = \mathbb{E}_\theta[S_n^2(\theta)]$ (as $\mathbb{E}_\theta[S_n(\theta)] = 0$ and $I_n(\theta) = \mathrm{Var}_\theta(S_n(\theta))$) - 2nd moment of score function is Fisher info

- $I_n(\theta) = nI(\theta)$ (due to IID r.v.s. in $\ell(\theta)$) - IID Fisher info rel.

- $I_n(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} \ell_n(\theta)\right]$ (under reg conditions) (!) regularity conditions are imp. - Fisher info is -ve expect. of 2nd derivative of log-likelihood

- vector generalisation

(07): parameter is a vector

- score function in vector case is a vector of functions

$$S_n(\theta) = \begin{bmatrix} \frac{\partial \ell_n(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell_n(\theta)}{\partial \theta_n} \end{bmatrix} = \nabla_\theta \ell_n(\theta)$$

$S_n(\theta) \in \mathbb{R}^k$ $\quad (S_n(\theta))_k = \frac{\partial \ell_n(\theta)}{\partial \theta_k}$

likelihood and (log likelihood is a scalar function)

- score function $S_n(\theta)$ is gradient of ~~vec~~ scalar function $\ell_n(\theta)$
- fisher info $I_n(\theta)$ is variance-covariance matrix of score vector
- $I_n(\theta) = \mathrm{Cov}(S_n(\theta))$ $\quad I_n(r,s) = -\mathbb{E}_\theta\left[\frac{\partial^2 \ell(\theta)}{\partial \theta_r \partial \theta_s}\right]$

(11) - example 11

(?) $\quad$ (?) - is μ a vector? or is θ a vector (μ, γ)?

- $X_1, \ldots, X_n \sim N(\mu, \gamma)$ $\quad \gamma = \sigma^2$

$$L_n(\mu,\gamma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\gamma}} \exp\left\{-\frac{1}{2\gamma}(x_i-\mu)^2\right\} \propto \gamma^{-n/2} \exp\left\{-\frac{1}{2\gamma}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$\ell_n(\mu,\gamma) = -\frac{n}{2}\log\gamma - \frac{1}{2\gamma}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$S_n(\mu,\gamma) = \begin{bmatrix} \frac{\partial \ell_n(\theta)}{\partial \mu} \\ \frac{\partial \ell_n(\theta)}{\partial \gamma} \end{bmatrix} = \begin{bmatrix} \frac{1}{\gamma}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{bmatrix}$$

$$I_n(\mu,\gamma) = -\mathbb{E}\begin{bmatrix} \frac{\partial^2 \ell(\theta)}{\partial \mu^2} & \frac{\partial^2 \ell(\theta)}{\partial \gamma \partial \mu} \\ \frac{\partial^2 \ell(\theta)}{\partial \mu \partial \gamma} & \frac{\partial^2 \ell(\theta)}{\partial \gamma^2} \end{bmatrix} = -\mathbb{E}\begin{bmatrix} -\frac{n}{\gamma} & -\frac{1}{\gamma^2}\sum_{i=1}^{n}(x_i-\mu) \\ -\frac{1}{\gamma^2}\sum_{i=1}^{n}(x_i-\mu) & \frac{n}{2\gamma^2} - \frac{1}{\gamma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \end{bmatrix}$$

=

$$= \begin{bmatrix} \frac{n}{\gamma} & 0 \\ 0 & \frac{n}{2\gamma^2} \end{bmatrix}$$

- why is paper I read so pale?

✓

- Arises from $-\frac{1}{\gamma^2}\mathbb{E}\left[\sum_{i=1}^n (x_i-\mu)\right] = 0$

and $-\frac{1}{\gamma^3}\mathbb{E}\left[\sum_{i=1}^n (x_i-\mu)^2\right] = 0$ ✗ (see below)

- Related: checking $\mathbb{E}_\theta(S_n(\underline{\theta})) = \underline{0}$

$$\mathbb{E}_\theta\begin{bmatrix} \frac{1}{\gamma}\sum_{i=1}^n (x_i-\mu) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^n (x_i-\mu)^2 \end{bmatrix}$$

- $\mathbb{E}_\theta\left[\frac{1}{\gamma}\sum_{i=1}^n (x_i-\mu)\right] = \frac{1}{\gamma}\sum_{i=1}^n \mathbb{E}[x_i]-\mu = \frac{1}{\gamma}\sum_{i=1}^n (\mu-\mu) = 0$

- $\mathbb{E}_\theta\left[-\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^n (x_i-\mu)^2\right] = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^n \mathbb{E}[x_i^2] - 2\mathbb{E}[x_i]\mu + \mu^2$

$$= -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^n (\gamma+\mu^2) - 2\mu^2 + \mu^2$$

$$= -\frac{n}{2\gamma} + \frac{n\gamma}{2\gamma^2} = \frac{-n\gamma+n\gamma}{2\gamma^2} = 0$$

Hence

$\mathbb{E}_\theta[S_n(\mu,\gamma)] = \underline{0}$

(*) Note: $\mathbb{E}\left[\sum_{i=1}^n (x_i-\mu)^2\right] = n\gamma$

$$\Rightarrow -\mathbb{E}\left[\frac{n}{2\gamma^2} - \frac{1}{\gamma^3}\sum_{i=1}^n (x_i-\mu)^2\right] = -\left\{\mathbb{E}\left[\frac{n}{2\gamma^2}\right] - \frac{1}{\gamma^3}\sum_{i=1}^n \mathbb{E}[(x_i-\mu)^2]\right\}$$

$= -\frac{n}{2\gamma^2} + \frac{n\gamma}{\gamma^3} = -\frac{n}{2\gamma^2} + \frac{n}{\gamma^2} = -\frac{n}{2\gamma^2} + \frac{2n}{2\gamma^2} = \frac{n}{2\gamma^2}$

Theorem 12 Proof

- Asymptotic Normality of MLE $\quad \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \frac{1}{I(\theta)})$
- Proof strategy: express $\sqrt{n}(\hat{\theta}_n - \theta)$ as a fraction via Taylor exp around true param θ.
  - use identities, CLT.

  - under appropriate regularity conditions.

- $\ell'(\hat{\theta}) = \ell'(\theta) + (\hat{\theta} - \theta)\ell''(\theta) + \dots = 0$ ⊛

$$\Rightarrow \sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \equiv \frac{A}{B}.$$

- $\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n} S(\theta, X_i)$ i.e. a sample mean of individual scores $S(\theta, X_i)$

⊛ each $S(\theta, X_i)$ is IID r.v. with $\mathbb{E}_\theta[S(\theta, X_i)] = 0$ and $\mathrm{Var}_\theta(S(\theta, X_i)) = I(\theta)$

- via CLT $\quad \frac{\sqrt{n}(\bar{S}_n - 0)}{I(\theta)} \xrightarrow{d} N(0,1) \Rightarrow \sqrt{n}(\bar{S}_n - 0) \xrightarrow{d} N(0, I(\theta))$

- So $A \xrightarrow{d} \sqrt{I(\theta)}\, Z$

⊛ $\frac{\ell'(\theta)}{-\ell''(\theta)} \approx (\hat{\theta}_n - \theta) \Rightarrow \frac{\sqrt{n}\,\ell'(\theta)}{-\ell''(\theta)} \approx \sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \approx \sqrt{n}(\hat{\theta}_n - \theta)$

- $B = -\frac{1}{n}\ell''(\theta) = -\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ell(\theta) = -\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log p(X_i, \theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log p(X_i, \theta)$

- Note that due to trick, $-\frac{1}{n}\ell''(\theta)$ is a sample mean (or continuous mapping?)

- via WLLN $B = -\frac{1}{n}\ell''(\theta) \xrightarrow{p} -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] = I(\theta)$

- via Slutsky (i.e. convergence in distn to r.vs. and constants - sum and product)

$$\frac{A}{B} \xrightarrow{d} \frac{\sqrt{I(\theta)}Z}{I(\theta)} = \frac{Z}{I(\theta)} \sim N\left(0, \frac{1}{I(\theta)}\right) \quad \text{as} \quad A \xrightarrow{d} \sqrt{I(\theta)}Z$$

$$B \xrightarrow{P} I(\theta) \Rightarrow B \xrightarrow{d} I(\theta)$$

$I(\theta)$ = constant.

Hence. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right)$

$$\Rightarrow \hat{\theta}_n \approx N\left(0, \frac{1}{nI(\theta)}\right) = N\left(0, \frac{1}{I_n(\theta)}\right)$$

(611) Variance of MLE (approximation) $\quad Var_\theta(\hat{\theta}_n) \approx \frac{1}{I_n(\theta)}$

(611) Approximate standard errors $\quad se(\hat{\theta}_n) \approx \sqrt{\frac{1}{I_n(\theta)}} = \sqrt{\frac{1}{nI(\theta)}}$ $\quad$ (511)

(611) estimated approx. standard errors $\quad \hat{se} = \hat{se}(\hat{\theta}_n) = \sqrt{\frac{1}{I_n(\hat{\theta}_n)}} = \sqrt{\frac{1}{nI(\hat{\theta}_n)}}$

(511) replace θ with $\hat{\theta}_n$ in Fisher info

(?) - How is Slutsky's theorem relevant (in notes) $\quad$ D/S1

- can see following argument instead:-
  - Fisher info is smooth/continuous function $\quad I_n(\hat{\theta}_n): \hat{\theta}_n \longrightarrow I_n(\hat{\theta}_n)$
  - via continuous mapping;

$$\hat{\theta}_n \xrightarrow{P} \theta \Rightarrow I_n(\hat{\theta}_n) \xrightarrow{P} I_n(\theta) \Rightarrow \sqrt{\frac{1}{I_n(\hat{\theta}_n)}} \xrightarrow{P} \sqrt{\frac{1}{I_n(\theta)}}$$

and hence:-

$$\hat{se} \xrightarrow{P} se$$

i.e. estimated (approx) std errors are a consistent estimator of (approx) std. errors.

Theorem 14 (Delta Method)

- conventionally, delta method concerns approx / conv. in distn results / extension of CLT to a function of a sample mean
- in context of asymptotic Normality, delta method concerns approx / conv in distri / extension of CLT to a function of MLE

(D1) for a smooth function of θ, $\tau(\theta)$, then

$$\sqrt{n}\left(\tau(\hat{\theta}_n) - \tau(\theta)\right) \xrightarrow{d} N\left(0, \frac{(\tau'(\theta))^2}{I(\theta)}\right)$$

- multiply by sq. deriv. eval at param.

- from Wasserman ch 9.9:- also impose differentiability

i.e. $\tau'(\theta) \neq 0$ and τ is differentiable.

- MLE estimator of $\tau = \tau(\theta)$ is $\hat{\tau} = \tau(\hat{\theta})$. (confused by not.)
- delta method addresses distribution of $\hat{\tau}$.
- from earlier;

approx std. error $se(\hat{\theta}_n) = \sqrt{\frac{1}{I_n(\theta)}}$ est std error $\hat{se}(\hat{\theta}_n) = \sqrt{\frac{1}{I_n(\hat{\theta}_n)}}$

(N): this part was confusing on review

- for clarity, previously concerned with approx and est. std errors of MLE
- under delta method, now concerned with approx. and est std errors of functions of MLE i.e. $\hat{\tau} = \tau(\hat{\theta}_n)$

- approx std. error $se(\hat{\tau}) = \sqrt{\frac{|\tau'(\theta)|^2}{I_n(\theta)}}$ est std. error $\hat{se}(\hat{\tau}) = \sqrt{\frac{|\tau'(\hat{\theta}_n)|^2}{I_n(\hat{\theta}_n)}}$

(A3) $I_n(\theta) = \mathbb{E}[-\ell''(\theta)] = \mathbb{E}\left[\frac{n}{\theta^2}\right] = \frac{n}{\theta^2}$

Hence $Var_\theta(\hat{\theta}_n) \approx \frac{1}{I_n(\theta)} \Rightarrow se(\hat{\theta}_n) = \sqrt{\frac{1}{I_n(\theta)}} = \frac{\theta}{\sqrt{n}}$

$\hat{se}(\hat{\theta}_n) = \sqrt{\frac{1}{I_n(\hat{\theta}_n)}} = \frac{\hat{\theta}_n}{\sqrt{n}}$ ✓

Example 16

- take opp. to clear up lack of familiarity with manipulating normals

$X_1, \dots, X_n \sim Bern(p)$

$\hat{p}_{MLE} = \bar{X}_n$ Fisher info $I(p) = \dfrac{1}{p(1-p)}$

Hence A.N. $\Rightarrow$ $\sqrt{n}(\hat{p}-p) \xrightarrow{d} N(0, p(1-p))$

$$(\hat{p}-p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{n}\right)$$

$$\hat{p} \xrightarrow{d} N\left(p, \frac{p(1-p)}{n}\right)$$

- Asymptotic variance $\dfrac{p(1-p)}{n}$ estimated via $\dfrac{\hat{p}(1-\hat{p})}{n}$

- Approx. std error $se(\hat{p}) = \sqrt{\dfrac{p(1-p)}{n}}$, est. std error $\hat{se}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

- estimate $\tau = \dfrac{p}{(1-p)} \Rightarrow \hat{\tau} = \dfrac{\hat{p}}{1-\hat{p}}$

- $\dfrac{\partial}{\partial p} \dfrac{p}{1-p} = \dfrac{1}{(1-p)^2}$

- est. std. error $\hat{se}(\hat{\tau}) = \sqrt{\dfrac{|\tau'(\hat{\theta}_n)|^2}{I_n(\hat{\theta}_n)}} = \sqrt{\dfrac{1}{I_n(\hat{\theta}_n)}} \cdot |\tau'(\hat{\theta}_n)|$

$$= \hat{se}(\hat{p}) \cdot |\tau'(\hat{\theta}_n)| \quad \text{(on)}$$

$$= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \frac{1}{(1-\hat{p})^2} = \sqrt{\frac{\hat{p}}{n(1-\hat{p})^3}}$$

(on) crucially, from Wasserman ch 9.

If $\tau = \tau(\theta)$ where τ is diff and $\tau'(\theta) \neq 0$, then ...

and $\hat{\tau} = \tau(\hat{\theta}_n)$ then

$$\hat{se}(\hat{\tau}) = \hat{se}(\hat{\theta}_n) |\tau'(\hat{\theta}_n)|$$

(17) Popularity of MLE due to efficiency
- MLE has lowest asymptotic variance, compared to other well-behaved estimators $\hat{\theta}$.

(15) Example 17
- (?): this was confusing.
- $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$

$\hat{\lambda}_{MLE} = \bar{X}_n$
- interested in estimation of $\tau = P(X_i = 0)$
- under model, $\tau = e^{-\lambda}$
- via equivariance; $\hat{\tau} = e^{-\hat{\lambda}}$, use this to construct mle $V_n = e^{-\hat{\lambda}}$ (*)
- construct a new random variable

$$Y_i = \mathbb{1}(X_i = 0)$$
- From this, use a different estimator:- (*)

$$W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
- Going to list explicitly how delta method is used
- delta method in context of asymptotic Normality is concerns limiting distn of a function of an estimator
- note: $\tau = \tau(\lambda) = e^{-\lambda}$ (function of parameter)

$V_n = \tau(\hat{\lambda}) = e^{-\hat{\lambda}}$ (MLE constructed via equivariance)
- delta method in general:- $\sqrt{n}(\tau(\hat{\theta}_n) - \tau(\theta)) \xrightarrow{d} N\left(0, \frac{(\tau'(\theta))^2}{I(\theta)}\right)$
- Our context:

$$\text{var}(V_n) = \text{var}(\tau(\hat{\lambda})) \approx \frac{(\tau'(\lambda))^2}{I_n(\lambda)} = \frac{(\tau'(\lambda))^2}{n I(\lambda)}$$

· or alternatively

$var(V_n) \approx var(\hat{\lambda}) \cdot (\tau'(\lambda))^2$

$var(\hat{\lambda}) \approx \dfrac{1}{I_n(\lambda)}$

· Fisher info $I_n(\lambda) = var(S_n(\lambda))$ (variance of score)

· score function:- $S_n(\lambda) = l'(\lambda) = \dfrac{\partial}{\partial \lambda} \log L(\lambda)$

$$= \frac{\partial}{\partial \lambda} \log\left(\prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}\right)$$

$$= \frac{\partial}{\partial \lambda} \log\left(\frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)}\right)$$

$$= \frac{\partial}{\partial \lambda}\left(-n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log(x_i!)\right)$$

$$= -n + \left(\frac{1}{\lambda}\right)\sum_{i=1}^{n} x_i$$

Hence $I_n(\lambda) = var(S_n(\lambda))$

$= var\left(-n + \left(\frac{1}{\lambda}\right)\sum_{i=1}^{n} x_i\right)$

$= \frac{1}{\lambda^2} var\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{\lambda^2}(n\lambda) = \frac{n}{\lambda}$

and $I(\lambda) = \frac{1}{\lambda}$

· $(\tau'(\lambda))^2 = (-e^{-\lambda})^2 = e^{-2\lambda}$

so $var(V_n) \approx \dfrac{\lambda e^{-2\lambda}}{n}$

$\therefore \sqrt{n}(V_n - \tau(\lambda)) \xrightarrow{d} N(0, \lambda e^{-2\lambda})$

- Returning to construction of W_n estimator:-

$$Y_i = \mathbb{I}(X_i = 0) \Rightarrow Y_i \sim Bern(\tau) \qquad \tau = P(X_i = 0) = e^{-\lambda}$$

$$\mathbb{E}[Y_i] = \tau \quad \text{and} \quad Var(Y_i) = \tau(1-\tau)$$

- As $W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a sample mean, invoke CLT

$$\sqrt{n}(W_n - \tau) \xrightarrow{d} N\left(0, e^{-\lambda}(1-e^{-\lambda})\right)$$

so $$ARE(W_n, \underset{(mle)}{V_n}) = \frac{\sigma_V^2}{\sigma_W^2} = \frac{\lambda e^{-2\lambda}}{e^{-\lambda}(1-e^{-\lambda})} = \frac{\lambda}{e^{\lambda}(1-e^{-\lambda})} = \frac{\lambda}{e^{\lambda}-1} \leq 1$$

- Illustrating efficiency of MLE: $ARE(W_n, mle) \leq 1$