The application here is that if $\theta \in \mathbb{R}^k$, for a nicely behaved "MLE" $\hat{\theta}_n$ with

$$\underset{k\times k}{I_1}(\theta) \text{ the FI in } X_1 \text{ about } \theta$$

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}_\theta} MVN_k\left(0, I_1^{-1}(\theta)\right)$$

means

$$\underbrace{\sqrt{n}\,(\hat{\theta}_n - \theta)' \left(I_1^{-1}(\theta)\right)^{-1} \sqrt{n}\,(\hat{\theta}_n - \theta)}_{(\hat{\theta}_n - \theta)'\, n I_1(\theta)\, (\hat{\theta}_n - \theta)} \xrightarrow{\mathcal{L}_\theta} \chi^2_k$$

and so an (unusable) confidence set for θ is

$$\left\{\theta \,\middle|\, (\hat{\theta}_n - \theta)'\,(n I_1(\theta))(\hat{\theta}_n - \theta) < c\right\}$$

This involves the unknown I_1 — two practical fixes are then

1) Use of the "Expected FI" — i.e. replace $I_1(\theta)$ by $I_n(\hat{\theta}_n)$

2) Use of the "Observed FI" — i.e. think with

$$\underset{k\times k}{H_n}(\theta) = \left(\frac{\partial^2 l_n(\theta)}{\partial \theta_i \partial \theta_j}\right) \quad \text{the Hessian matrix for the loglikelihood}$$

that $H_n(\theta)$ is a sum of iid terms and it's plausible that

$$-\frac{1}{n} H_n(\theta) \xrightarrow{P_\theta} I_1(\theta)$$

and then further that

$$-\frac{1}{n} H_n(\hat{\theta}_n) \xrightarrow{P_\theta} I_1(\theta)$$

so that one might well replace $nI_1(\theta)$ with

$$n\left(-\frac{1}{n} H_n(\hat{\theta}_n)\right) = -H_n(\hat{\theta}_n)$$

to get the approximate confidence set for θ

$$\left\{\theta \mid (\hat{\theta}_n - \theta)\left(-H_n(\hat{\theta}_n)\right)(\hat{\theta}_n - \theta) < c\right\}$$

In this multi-parameter context I might have in mind inference for only some sub-vector of θ, say θ_1 of dimension $l < n$ — i.e. suppose

$$\theta = \begin{pmatrix} \underset{l \times 1}{\theta_1} \\ \underset{(k-l)\times 1}{\theta_2} \end{pmatrix} \quad \text{and} \quad \hat{\theta}_n = \begin{pmatrix} \underset{l \times 1}{\hat{\theta}_{n1}} \\ \underset{(k-l)\times 1}{\hat{\theta}_{n2}} \end{pmatrix}$$

and my primary interest is in θ_1 — I need to be careful as I think about using MVN stuff to do inference here —

$$\begin{aligned} \text{covariance matrix for approximating dsn for } \hat{\theta}_n &= I^{-1}(\theta) = \left(nI_1(\theta)\right)^{-1} \\ &\approx \left(nI_1(\hat{\theta}_n)\right)^{-1} \\ &\approx \left(-H_n(\hat{\theta}_n)\right)^{-1} \end{aligned}$$

Then the covariance matrix for the approximating dsn for $\widehat{\theta}_{n1}$ is

upper left $l \times l$ block of $\underline{I}^{-1}(\theta) \left(= (n\underline{I}_1(\theta))^{-1}\right)$

$\approx$ upper left $l \times l$ block of $(n\underline{I}_1(\widehat{\theta}_n))^{-1}$

$\approx$ upper left $l \times l$ block of $(-H_n(\widehat{\theta}_n))^{-1}$

and these blocks are <u>NOT</u> in general the inverses of the upper left blocks of the matrices inside the ( ) above

---

The folklore is that in all this using observed rather than expected FI does a better job of producing actual coverage probability close to nominal (based on the limiting dsn) —

day 38

Something even better in this regard is based on a different use of limiting dsns for "MLE's" —

<u>Theorem 9</u> ($\theta \in \mathbb{R}^1$ version) (of ML handout)

Under appropriate conditions in an iid model, if $\{\delta_n(X)\}$ is consistent for θ at θ_0 and with θ_0 probability approaching 1 is a root of the likelihood equation

$$l_n'(\theta) = 0$$

Then

$$2\left(l_n(\delta_n(X)) - l_n(\theta_0)\right) \xrightarrow{\mathcal{L}_{\theta_0}} \chi_1^2$$

Note that: the expression

$$2\,(\underbrace{l_n(\delta_n(X))}_{\text{log likelihood at the "MLE"}} - \underbrace{l_n(\theta_0)}_{\text{log likelihood at } \theta_0})$$

is "sort of"

$$2 \log \left(\frac{\sup_\theta L_n(\theta)}{L_n(\theta_0)} \right)$$

← likelihood ratio statistic for testing $H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$

(L_n The likelihood)

So Theorem 9 gives me a way to set critical values for LRTs of point null hypotheses for large n - Far more importantly, it also gives me a way to make confidence sets for θ by inverting tests - This amounts to doing the following - If c is the upper α pt of χ^2_1

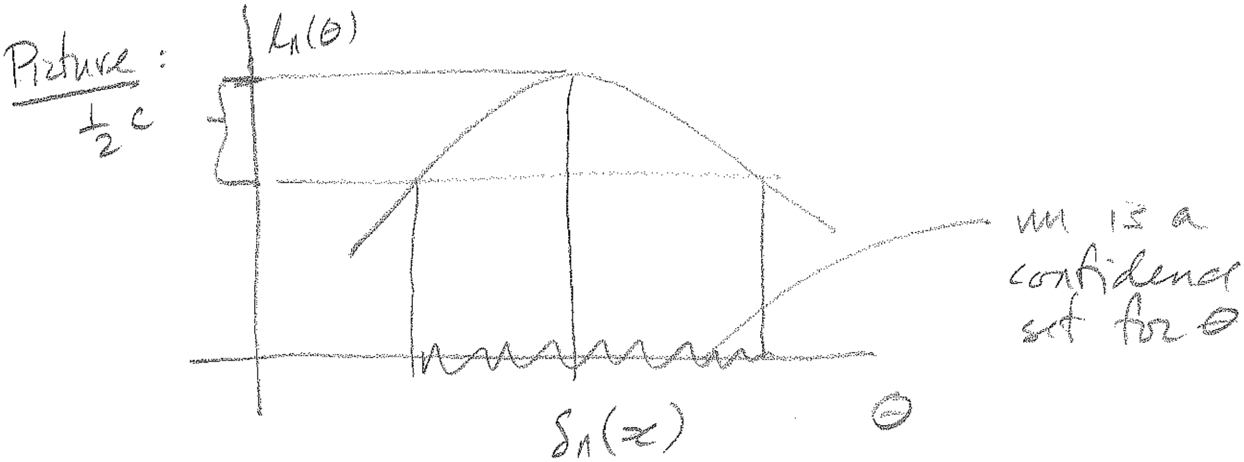
$$\underbrace{P_{\theta_0}\left[\,2\,(l_n(\delta_n(X)) - l_n(\theta_0)) < c\,\right]}_{P_{\theta_0}\left[\, l_n(\delta_n(X)) - \frac{1}{2}c < l_n(\theta_0)\,\right]} \approx 1-\alpha$$

That is, the set of θ's with $l_n(\theta)$ no more than $\frac{1}{2}c$ below the maximum of the loglikelihood functions as a confidence set for θ - That is with data $X = x$

$$\{ \theta \mid l_n(\theta) > l_n(\delta_n(x)) - \tfrac{1}{2} c \}$$

can be used as a confidence set for θ

Picture:



There is an important multi-parameter extension of this, but before doing that I want to give you some idea of how this comes about

Argument for Thm 3 of the handout:

Again write $\hat\theta_n$ instead of $\delta_n(X)$ — (recall BTW that the argument for asymptotic normality for $\hat\theta_n$ is done by expanding $l_n'(\cdot)$ in a Taylor series around θ_0) — here expand $l_n(\cdot)$ around $\hat\theta_n$

$$l_n(\theta_0) = l_n(\hat\theta_n) + (\theta_0 - \hat\theta_n) l'(\hat\theta_n) + \tfrac{1}{2}(\theta_0 - \hat\theta_n)^2 l_n''(\hat\theta_n)$$
$$+ \tfrac{1}{6}(\theta_0 - \hat\theta_n)^3 l_n'''(\theta')$$

for some θ' between $\hat\theta_n$ and θ_0 — So

$$2(l_n(\hat\theta_n) - l_n(\theta_0)) = \underbrace{-(\theta_0 - \hat\theta_n) l_n'(\hat\theta_n)}_{A_n^*}$$
$$\underbrace{- 2(\tfrac{1}{2})(\theta_0 - \hat\theta_n)^2 l_n''(\hat\theta_n)}_{B_n^*}$$
$$\underbrace{- \tfrac{1}{3}(\theta_0 - \hat\theta_n)^3 l_n'''(\theta')}_{C_n^*}$$

$\hat{\theta}_n$ an "MLE" makes $l_n'(\hat{\theta}_n) = 0$, makes $A_n^* = 0$

$$B_n^* = (\theta_0 - \hat{\theta}_n)^2 (-l_n''(\hat{\theta}_n))$$

$$= (\sqrt{n}(\theta_0 - \hat{\theta}_n))^2 (-\tfrac{1}{n} l_n''(\hat{\theta}_n))$$

$\sqrt{n}(\theta_0 - \hat{\theta}_n) \xrightarrow{\mathcal{L}_{\theta_0}} N(0, I_1^{-1}(\theta_0))$, so the first factor goes to $(N(0, I_1^{-1}(\theta_0)))^2$

the second factor: not too surprising if this converges to $I_1(\theta_0)$

So $B_n^* \xrightarrow{\mathcal{L}_{\theta_0}} \underbrace{\left(\sqrt{I_1(\theta_0)} \cdot N(0, I_1^{-1}(\theta_0))\right)^2}_{\chi_1^2}$

(since the square of a standard normal is χ_1^2)

<u>And</u> sure enough, standard regularity conditions are set up (just as for the proof of asymptotic normality of "MLEs") to produce

$$C_n^* \xrightarrow{P_{\theta_0}} 0$$

There is an important multivariate version of the χ^2 limit for a LRT statistic - that goes as follows - again

$$\underset{k\times 1}{\theta} = \begin{pmatrix} \underset{l\times 1}{\theta_1} \\ \underset{(k-l)\times 1}{\theta_2} \end{pmatrix} \qquad \hat{\theta}_n \text{ an "MLE" similarly partitioned}$$

Suppose that for each $\theta_1 \in \mathbb{R}^l$

$\theta_{2n}^*(\theta_1) \in \mathbb{R}^{k-l}$ is a "maximizer" of $l_n(\theta_1, \cdot)$ over choices of θ_2

Then
$$l_n^*(\theta_1) = l_n(\theta_1, \theta_{2n}^*(\theta_1))$$
$$= \max_{\theta_2} l_n(\theta_1, \theta_2)$$
is called the "profile likelihood" for θ_1 and can essentially be used like a likelihood to do inference for θ_1 - There is, e.g. the large sample result

<u>"Thm"</u> Under appropriate regularity conditions in an iid model, if $\theta \in \mathbb{R}^k$

$$2(l_n(\hat{\theta}_n) - l_n^*(\theta_{10})) \xrightarrow{\mathcal{L}_{\theta_{10}}} \chi^2_l$$

↑ max of $l_n(\theta)$ is also the max of $l_n^*(\theta_1)$

(under) any θ with $\theta_1 = \theta_{10}$

I can then use this to set critical values for LRT's of $H_0: \theta_1 = \theta_{10}$ and if c^* is the upper α pt of χ^2_l

$$P_{\theta_0}\left[2(l_n(\hat{\theta}_n) - l_n^*(\theta_{10})) < c^*\right] \approx 1-\alpha$$

$$P_{\theta_0}\left[l_n(\hat{\theta}_n) - \tfrac{1}{2}c^* < l_n^*(\theta_{10})\right]$$

So that with data $X = x$ and MLE $\hat{\theta}_n(x)$,

$$\left\{\theta_1 \,\middle|\, l_n^*(\theta_1) > l_n(\hat{\theta}_n(x)) - \tfrac{1}{2}c^*\right\}$$

can be used as an approximate $(1-\alpha)\times 100\%$ confidence set for θ_1