Statistics for Data Science: Week 5

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

Example (MLE for Gaussian distribution)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(Y_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{2\sigma^2}\right\}.$$

giving loglikelihood

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

All partial second derivatives exist and are

$$\frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu) = 0$$

$$\frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \mu)^2. = 0$$

Example (MLE for Gaussian distribution, continued)

Solving $\nabla_{(\mu,\sigma^2)}\ell(\mu,\sigma^2) = 0$ for (μ,σ^2) gives a system of equations in two unknowns, with unique root

$= \frac{1}{n}\sum Y_i$

$S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$

$$\left(\bar{Y}, \underbrace{n^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right).$$

$-\nabla_\theta^2 \ell \succ 0$

Positive -definite.

Call this $(\hat{\mu},\hat{\sigma}^2)$, and let's verify it's a maximum. Note that

$$\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2) = -\frac{n}{\sigma^2}, \quad \frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(Y_i - \mu)^2$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2) = \frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2) = -\frac{\sum_{i=1}^{n}(Y_i - \mu)}{\sigma^4} = \boxed{\frac{n\mu - n\bar{Y}}{\sigma^4}}.$$

Calculating these derivatives at $(\hat{\mu},\hat{\sigma}^2)$, we get

$$\left.\frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = -\frac{n}{\hat{\sigma}^2}, \quad \left.\frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2)\right|_{(\mu,\sigma^2)=(\hat{\mu},\hat{\sigma}^2)} = -\frac{n}{2\hat{\sigma}^4}$$

Example (MLE for Gaussian distribution, continued)

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat\mu,\hat\sigma^2)} = \frac{\partial^2}{\partial\sigma^2\partial\mu}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat\mu,\hat\sigma^2)} = \frac{n\hat\mu - n\hat\mu}{\hat\sigma^4} = 0.$$

Thus the matrix

$$\left[-\nabla^2_{(\mu,\sigma^2)}\ell(\mu,\sigma^2)\Big|_{(\mu,\sigma^2)=(\hat\mu,\hat\sigma^2)}\right]$$

$$\begin{bmatrix} \frac{\partial^2}{\partial\mu^2} & \underbrace{\frac{\partial}{\partial\sigma^2\partial\mu}}_{=0} \\ \underbrace{\frac{\partial}{\partial\mu\partial\sigma^2}}_{=0} & \frac{\partial^2}{\partial(\sigma^2)^2} \end{bmatrix}$$

is diagonal. If both of its diagonal elements are positive, then it will be positive definite. This is indeed the case since $\hat\sigma^2 > 0$ nd so the unique MLE of (μ,σ^2) is given by

$$(\hat\mu,\hat\sigma^2) = \left(\bar Y\ ,\ \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar Y)^2\right).$$

□

Note that from our Gaussian sampling results we get that σ^2 is biased.

Example (MLE for Poisson Distribution)

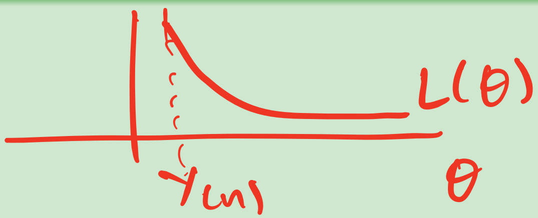
Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Poisson}(\lambda)$. Then

$$L(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{Y_i}}{Y_i!} e^{-\lambda} \right\} \implies \log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^{n} Y_i - \underbrace{\sum_{i=1}^{n} \log(Y_i!)}_{\perp\!\!\!\perp \lambda}$$

Setting $\nabla_\lambda \log L(\lambda) = -n + \lambda^{-1} \sum Y_i = 0$ we obtain $\hat{\lambda} = \bar{Y}$ since $\nabla^2_\lambda \log L(\lambda) = -\lambda^{-2} \sum Y_i < 0$.

Example (MLE for Uniform Distribution – a non-differentiable case)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$. The likelihood is



$$L(\theta) = \theta^{-n} \prod_{i=1}^{n} \mathbf{1}\{0 \leq Y_i \leq \theta\} = \theta^{-n} \mathbf{1}\{\theta \geq Y_{(n)}\}.$$

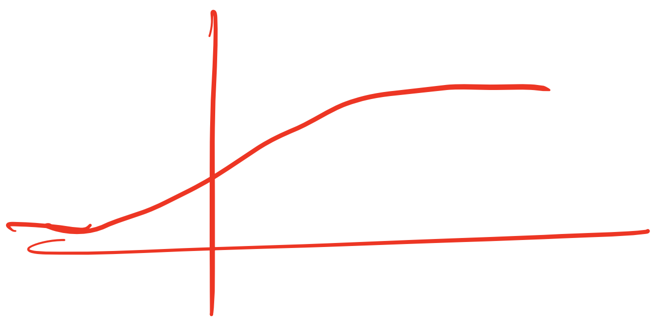
Hence if $\theta < Y_{(n)}$ the likelihood is zero. In the domain $[Y_{(n)}, \infty)$, the likelihood is a decreasing function of θ. Hence $\hat{\theta} = Y_{(n)}$.

Example (Equivariance of the MLE)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$, and suppose we're interested in estimating $\underbrace{\mathbb{P}[Y_1 \leq y]}_{F_Y(y)}$, for a given $y \in \mathbb{R}$. Note that

$$\mathbb{P}[Y_1 \leq y] = \mathbb{P}[\frac{Y_1 - \mu}{\sigma} \leq \frac{y - \mu}{\sigma}] = \Phi(\frac{y - \mu}{\sigma}),$$

where Φ is the standard normal CDF. The mapping $\mu \mapsto \Phi(y - \mu)$ is bijective, since Φ is strictly monotone. So by equivariance, the MLE of $\mathbb{P}[Y_1 \leq y]$ is $\Phi(y - \hat{\mu})$, where $\hat{\mu}$ is the MLE of μ (which by our previous example is $\hat{\mu} = \bar{Y}$).



$\hat{\Phi} = \Phi(y - \hat{\mu})$

MLE

Example (Equivariance and usual vs natural parameterisation)

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, with

$$f(y) = \exp\left\{\phi T(y) - \gamma(\phi) + S(y)\right\}, \qquad y \in \mathcal{Y}$$

where $\phi \in \Phi \subseteq \mathbb{R}$ is the natural parameter. Suppose we can write $\phi = \eta(\theta)$, where $\theta \in \Theta$ is the usual parameter and $\eta : \Theta \to \Phi$ is a differentiable bijection (so that $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, for $d = \gamma \circ \eta$). In this notation, the density/frequency takes the form

$$\exp\left\{\phi T(y) - \gamma(\phi) + S(y)\right\} = \exp\left\{\eta(\theta) T(y) - d(\theta) + S(y)\right\}.$$

Equivariance now implies that if $\hat{\theta}$ is the MLE of θ, then $\eta(\hat{\theta})$ is the MLE of $\phi = \eta(\theta)$. The converse is also true: if $\hat{\phi}$ is the MLE of ϕ, then $\eta^{-1}(\hat{\phi})$ is the MLE of $\theta = \eta^{-1}(\phi)$. $\square$

Examples show that likelihood generally gives sensible estimators – still:

- Beyond intuition, is there a canonical mathematical reason for it?
- What rigorous guarantees can we offer?
  - $\hookrightarrow$ Can we get consistency?
  - $\hookrightarrow$ Can we approach reasonable MSE performance?

To answer these questions, we go back to entropy and Kullback-Leibler divergence.

$\hat{\theta}_{MLE} \xrightarrow{p} \theta$

Consider the random function

$$\hat{\theta}_{MLE} = \arg\max_{u} \Psi_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}[\underbrace{\log f(Y_i;\boldsymbol{u})}_{\ell(u)} - \underbrace{\log f(Y_i;\boldsymbol{\theta})}_{\ell(\theta)}]$$

which is maximized at $\hat{\boldsymbol{\theta}}_n$. By the law of large numbers, for each $\boldsymbol{u} \in \Theta$,

$$\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\frac{f(Y_i;\boldsymbol{u})}{f(Y_i;\boldsymbol{\theta})}\right)\right] = -KL\big(f(Y_i;\boldsymbol{u})\big\|f(Y_i;\boldsymbol{\theta})\big)$$

max $f(x)$ = min $-f(x)$

- The latter is minimised at $\boldsymbol{\theta}$ and so $\Psi(u)$ is maximized at $\boldsymbol{\theta}$.
- Moreover, unless $f(x;\boldsymbol{u}) = f(x;\boldsymbol{\theta})$ for all $x \in \operatorname{supp} f$, we have $\Psi(\boldsymbol{u}) < 0$
- It follows that Ψ is uniquely maximised at $\boldsymbol{\theta}$

MLE can be regarded as a minimiser of an approximate (empirically constructed) KL-divergence from the truth!

Does $\left\{\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) \ \forall\ u \text{ with } \Psi \text{ maximized uniquely at } \boldsymbol{\theta}\right\}$ imply $\left\{\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}\right\}$?

Does $\left\{\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) \ \forall\ u \text{ with } \Psi \text{ maximized uniquely at } \boldsymbol{\theta}\right\}$ imply $\left\{\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}\right\}$?

- Unfortunately, the answer is in general no, without additional information.

Does $\left\{\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) \ \forall \ u \text{ with } \Psi \text{ maximized uniquely at } \boldsymbol{\theta}\right\}$ imply $\left\{\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}\right\}$?

- Unfortunately, the answer is in general no, without additional information.
- If $\theta \in \mathbb{R}$, can prove consistency if f is regular enough & MLE exists uniquely.

Does $\left\{\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) \ \forall \ u \text{ with } \Psi \text{ maximized uniquely at } \boldsymbol{\theta}\right\}$ imply $\left\{\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}\right\}$?

- Unfortunately, the answer is in general no, without additional information.
- If $\theta \in \mathbb{R}$, can prove consistency if f is regular enough & MLE exists uniquely.
- If $\theta \in \mathbb{R}^p$, we need more information on the form of the likelihood function
  - $\hookrightarrow$ For instance concavity and existence will usually give us consistency. We will show consistency in exponential families using this approach.
  - $\hookrightarrow$ More general situations require stronger forms of convergence of $\Psi_n(u) \to \Psi(u)$ plus additional regularity conditions.

Does $\left\{\Psi_n(\boldsymbol{u}) \xrightarrow{p} \Psi(\boldsymbol{u}) \ \forall \ u \text{ with } \Psi \text{ maximized uniquely at } \boldsymbol{\theta}\right\}$ imply $\left\{\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}\right\}$?

- Unfortunately, the answer is in general no, without additional information.
- If $\theta \in \mathbb{R}$, can prove consistency if f is regular enough & MLE exists uniquely.
- If $\theta \in \mathbb{R}^p$, we need more information on the form of the likelihood function
  - $\hookrightarrow$ For instance concavity and existence will usually give us consistency. We will show consistency in exponential families using this approach.
  - $\hookrightarrow$ More general situations require stronger forms of convergence of $\Psi_n(u) \to \Psi(u)$ plus additional regularity conditions.

When we can deduce consistency, though, we get some very nice properties for the (asymptotic) sampling distribution of the MLE...

Example (Consistency of MLE in $\theta \in \mathbb{R}$)

f' is cts.

Let $Y_1, \ldots, Y_n \overset{iid}{\sim} f(y;\theta)$ where f is C^1 with respect to θ. Assume that $\forall$ n, there exists a unique MLE $\hat{\theta}_n$. We will show that $\hat{\theta}_n \xrightarrow{p} \theta$.

Define

$$\Xi_n(u) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial u}\underbrace{\log\left(\frac{f(Y_i;u)}{f(Y_i;\theta)}\right)}_{\Psi(u)}\right] \quad \text{and} \quad \Xi(u) = \mathbb{E}\left[\frac{\partial}{\partial u}\log\left(\frac{f(Y_i;u)}{f(Y_i;\theta)}\right)\right],$$

θ

so that

- $\Xi_n(\hat{\theta}_n) = 0$ uniquely, by uniqueness of the MLE.
- $\Xi(\theta) = 0$ uniquely, assuming regularity allowing interchange of $\mathbb{E}$ and $\frac{\partial}{\partial u}$.

Since f is C^1, we have the inequality $\Rightarrow \Xi_n(\theta) \simeq 0$ $\qquad \varepsilon > 0, \ \varepsilon \to 0$

$$1 \Leftarrow \mathbb{P}[\Xi_n(\theta-\varepsilon) < 0 \ \& \ \Xi_n(\theta+\varepsilon) > 0] \leq \mathbb{P}[\theta-\varepsilon < \hat{\theta}_n < \theta+\varepsilon] \leq 1$$

because the event on the left hand side implies that on the right hand side. Finally, the law of large numbers implies that $\Xi_n(u) \xrightarrow{p} \Xi(u)$ for any u, so that the left hand side converges to 1, yielding consistency.

$$\Xi(u) = \mathbb{E}\left[\frac{\partial}{\partial u} \log\left(\frac{f(u)}{f(\theta)}\right)\right] = \frac{\partial}{\partial u} \mathbb{E}\left[\log\left(\frac{f(u)}{f(\theta)}\right)\right]$$

$$= \frac{\partial}{\partial u} \int \log\left(\frac{f(u)}{f(\theta)}\right) f(\theta)\, dx$$

$$\Xi(\mu) = \frac{\partial}{\partial u} \underbrace{\int \log\left(\frac{f(\mu)}{f(\theta)}\right) f(\theta)}_{\neq 1}$$

Example (Consistency of MLE in $\mathbb{R}^k$ for exponential families)

Consider $Y_1, ..., Y_n \overset{iid}{\sim} f(y; \phi)$ from a k-parameter exponential family

$$f(y) = \exp\left\{\sum_{j=1}^{k} \phi_j T_j(y) - \gamma(\phi_1, ..., \phi_k) + S(y)\right\}, \boldsymbol{\phi} = (\phi_1, ..., \phi_k)^\top \in \Phi \text{ open.}$$

The likelihood and loglikelihood (up to constants w.r.t. ϕ) are given by

$$L(\phi) = \exp\left\{\boldsymbol{\phi}^\top \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi})\right\} \quad \& \quad \ell(\boldsymbol{\phi}) = \boldsymbol{\phi}^\top \boldsymbol{\tau} - n\gamma(\boldsymbol{\phi})$$

where

$$\boldsymbol{\tau} = (\tau_1, ..., \tau_k)^\top, \qquad \tau_j(y_1, \ldots, y_n) = \textstyle\sum_{i=1}^n T_j(y_i).$$

If it exists, the MLE $\hat{\boldsymbol{\phi}}_n$ must thus satisfy

$$\nabla_{\boldsymbol{\phi}} \ell(\hat{\boldsymbol{\phi}}_n) = 0 \implies \nabla_{\boldsymbol{\phi}} \gamma(\hat{\boldsymbol{\phi}}_n) = n^{-1}\boldsymbol{\tau}.$$

Furthermore, existence of the MLE guarantees uniqueness by strict concavity:

$$-\nabla^2_{\boldsymbol{\phi}} \ell(\boldsymbol{\phi}) = n\nabla^2_{\boldsymbol{\phi}} \gamma(\boldsymbol{\phi}) = \mathrm{cov}\{\boldsymbol{\tau}\} \succ 0,$$

Example (Consistency of MLE in $\mathbb{R}^k$ for exponential families, ctd)

Now notice that by the law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} T_j(Y_i) \xrightarrow{p} \mathbb{E}[T_j] \overset{\text{defn.}}{=} \frac{\partial}{\partial \phi_j}\gamma(\phi), \qquad j = 1, \ldots, k.$$

It follows that

$$\nabla_{\boldsymbol{\phi}}\gamma(\hat{\boldsymbol{\phi}}_n) = n^{-1}\boldsymbol{\tau} \xrightarrow{p} \nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}).$$

Now if $\nabla_{\boldsymbol{\phi}}\gamma : \mathbb{R}^k \to \mathbb{R}^k$ were continuously invertible, with inverse map h, then the continuous mapping theorem would give us:

$$\nabla_{\boldsymbol{\phi}}\gamma(\hat{\boldsymbol{\phi}}_n) \xrightarrow{p} \nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}) \implies h\big(\nabla_{\boldsymbol{\phi}}\gamma(\hat{\boldsymbol{\phi}}_n)\big) \xrightarrow{p} h\big(\nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi})\big) \implies \hat{\boldsymbol{\phi}}_n \xrightarrow{p} \boldsymbol{\phi}.$$

In fact, the inverse function theorem tells us that the infinitely differentiable function $\nabla_{\boldsymbol{\phi}}\gamma : \mathbb{R}^k \to \mathbb{R}^k$ must admit a continuously differentiable inverse map h locally.

In summary: provided it exists, the MLE of the natural parameter in a k-parameter natural exponential family with open parameter space Φ is consistent.

Assuming we can get consistency, we can focus on understanding the sampling distribution of the MLE.

$F_{\hat{\theta}}$

For simplicity, assume $X_1, \dots, X_n$ are iid with density/frequency $f(x;\theta)$, $\theta \in \mathbb{R}$.

Introduce the notation:

- $\ell(x_i;\theta) = \log f(x_i;\theta)$
- $\ell'(x_i;\theta)$, $\ell''(x_i;\theta)$ and $\ell'''(x_i;\theta)$ are partial derivatives w.r.t θ.

$\frac{\partial}{\partial\theta}\ell$, $\frac{\partial^2}{\partial\theta^2}\ell$

Assuming we can get consistency, we can focus on understanding the sampling distribution of the MLE.

For simplicity, assume $X_1, \ldots, X_n$ are iid with density/frequency $f(x;\theta)$, $\theta \in \mathbb{R}$.

Introduce the notation:

- $\ell(x_i;\theta) = \log f(x_i;\theta)$
- $\ell'(x_i;\theta)$, $\ell''(x_i;\theta)$ and $\ell'''(x_i;\theta)$ are partial derivatives w.r.t θ.

Regularity Conditions $(*)$

(A1) Θ is an open subset of $\mathbb{R}$. Unif $(0,\theta]$

(A2) The support of f, $\mathrm{supp}(f)$, is independent of θ.

(A3) f is thrice continuously differentiable w.r.t. θ for all $x \in \mathrm{supp}(f)$. $f \in C^3$

(A4) $\mathbb{E}_\theta[\ell'(X_i;\theta)] = 0\ \forall\theta$ and $\mathrm{var}_\theta[\ell'(X_i;\theta)] = \mathcal{I}_1(\theta) \in (0,\infty)\ \forall\theta$.

(A5) $-\mathbb{E}_\theta[\ell''(X_i;\theta)] = \mathcal{J}_1(\theta) \in (0,\infty)\ \forall\theta$. $= \mathbb{E}[(\ell')^2]$

(A6) $\exists\ M(x) > 0$ and $\delta > 0$ such that $\mathbb{E}_{\theta_0}[M(X_i)] < \infty$ and

$$|\theta - \theta_0| < \delta \implies |\ell'''(x;\theta)| \leq M(x) \leq \max_x M(x)$$

Let's demistify these conditions...

- If Θ is open, then for θ the true parameter, it always makes sense for an estimator $\hat{\theta}$ to have a symmetric distribution around θ (e.g. Gaussian).

$\theta \in \Theta$, $|\hat{\theta} - \theta| \leq \varepsilon \Rightarrow \hat{\theta} \in \Theta$

- If Θ is open, then for θ the true parameter, it always makes sense for an estimator $\hat{\theta}$ to have a symmetric distribution around θ (e.g. Gaussian).
- Under condition (A2) we have $\frac{d}{d\theta} \underbrace{\int_{\text{supp } f} f(x;\theta)dx}_{=1} = 0$ for all $\theta \in \Theta$ so that, if we can interchange integration and differentiation,

$$0 = \int \frac{d}{d\theta} f(x;\theta)dx = \int \ell'(x;\theta) f(x;\theta)dx = \mathbb{E}_\theta[\ell'(X_i;\theta)]$$

so that in the presence of (A2), (A4) is essentially a condition that enables differentiation under the integral and asks that the r.v. ℓ' have a finite second moment for all θ.

- If Θ is open, then for θ the true parameter, it always makes sense for an estimator $\hat{\theta}$ to have a symmetric distribution around θ (e.g. Gaussian).
- Under condition (A2) we have $\frac{d}{d\theta}\int_{\text{supp } f} f(x;\theta)dx = 0$ for all $\theta \in \Theta$ so that, if we can interchange integration and differentiation,

$$0 = \int \frac{d}{d\theta} f(x;\theta)dx = \int \ell'(x;\theta)f(x;\theta)dx = \mathbb{E}_\theta[\ell'(X_i;\theta)]$$

  so that in the presence of (A2), (A4) is essentially a condition that enables differentiation under the integral and asks that the r.v. ℓ' have a finite second moment for all θ.
- Similarly, (A5) requires that ℓ'' have a first moment for all θ.

- If Θ is open, then for θ the true parameter, it always makes sense for an estimator $\hat{\theta}$ to have a symmetric distribution around θ (e.g. Gaussian).
- Under condition (A2) we have $\frac{d}{d\theta}\int_{\text{supp } f} f(x;\theta)dx = 0$ for all $\theta \in \Theta$ so that, if we can interchange integration and differentiation,

$$0 = \int \frac{d}{d\theta} f(x;\theta)dx = \int \ell'(x;\theta)f(x;\theta)dx = \mathbb{E}_\theta[\ell'(X_i;\theta)]$$

  so that in the presence of (A2), (A4) is essentially a condition that enables differentiation under the integral and asks that the r.v. ℓ' have a finite second moment for all θ.
- Similarly, (A5) requires that ℓ'' have a first moment for all θ.
- Conditions (A2) and (A6) are smoothness conditions that will allow us to 'linearize' the problem, while the other conditions will allow us to 'control' the random linearization. $|\ell'''| \leq M(x)$ ℓ

- If Θ is open, then for θ the true parameter, it always makes sense for an estimator $\hat{\theta}$ to have a symmetric distribution around θ (e.g. Gaussian).
- Under condition (A2) we have $\frac{d}{d\theta}\int_{\text{supp } f} f(x;\theta)dx = 0$ for all $\theta \in \Theta$ so that, if we can interchange integration and differentiation,

$$0 = \int \frac{d}{d\theta} f(x;\theta)dx = \int \ell'(x;\theta)f(x;\theta)dx = \mathbb{E}_\theta[\ell'(X_i;\theta)]$$

  so that in the presence of (A2), (A4) is essentially a condition that enables differentiation under the integral and asks that the r.v. ℓ' have a finite second moment for all θ.
- Similarly, (A5) requires that ℓ'' have a first moment for all θ.
- Conditions (A2) and (A6) are smoothness conditions that will allow us to 'linearize' the problem, while the other conditions will allow us to 'control' the random linearization.
- Furthermore, if we can differentiate twice under the integral sign

$$0 = \int \frac{d}{d\theta}[\ell'(x;\theta)f(x;\theta)]dx = \underbrace{\int \ell''(x;\theta)f(x;\theta)dx}_{J(\theta)} + \underbrace{\int (\ell'(x;\theta))^2 f(x;\theta)dx}_{I(\theta)}$$

  so that $\mathcal{I}(\theta) = \mathcal{J}(\theta)$.

Theorem (Asymptotic Distribution of the MLE)

Let $X_1, ..., X_n$ be iid random variables with density (frequency) $f(x; \theta)$ and satisfying the stated regularity conditions. If the MLE $\hat{\theta}_n$ exists uniquely and is consistent, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathcal{I}_1(\theta)}{\mathcal{J}_1^2(\theta)}\right).$$

When $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$, we have of course $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right)$.

Theorem (Asymptotic Distribution of the MLE)

Let $X_1, \ldots, X_n$ be iid random variables with density (frequency) $f(x;\theta)$ and satisfying the stated regularity conditions. If the MLE $\hat{\theta}_n$ exists uniquely and is consistent, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathcal{I}_1(\theta)}{\mathcal{J}_1^2(\theta)}\right).$$

When $\mathcal{I}_1(\theta) = \mathcal{J}_1(\theta)$, we have of course $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right)$.

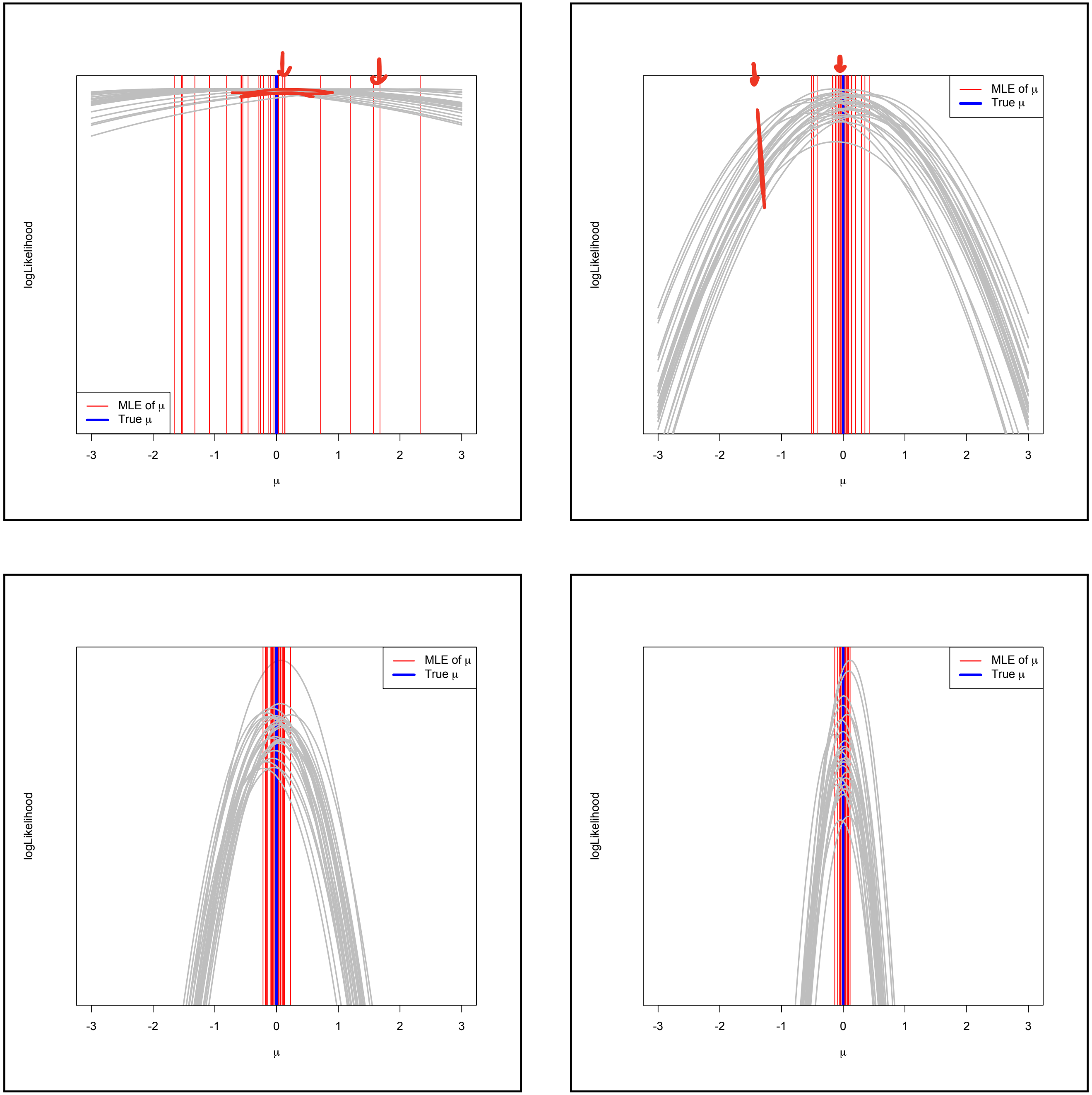
- Note that this can be interpreted as

$$\hat{\theta}_n \overset{d}{\approx} N\left(\theta\,, \frac{1}{n\mathcal{I}_1(\theta)}\right) \equiv N\left(\theta\,, \frac{1}{\mathcal{I}_n(\theta)}\right).$$

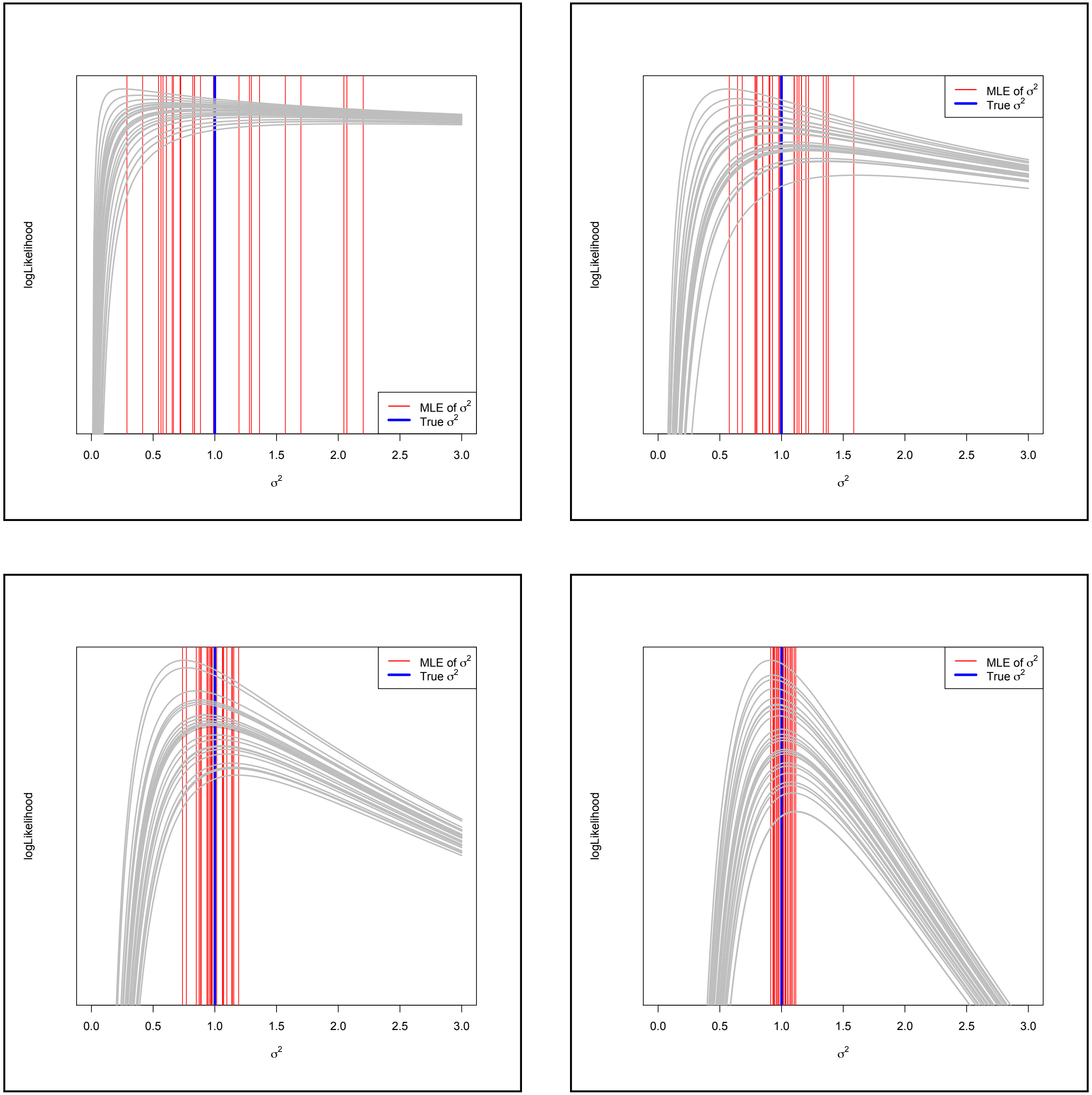
- In order words: the MLE is approximately normally distributed, approximately unbiased, and approximately achieving the Cramér-Rao lower bound!

Why $\mathcal{I}_n(\theta)$? (... curvature)

$I(\theta) = Var(\ell') = \mathbb{E}[(\ell')^2]$



Why $\mathcal{I}_n(\theta)$? (... curvature)



Proof.

Under conditions (A1)-(A3), if $\hat{\theta}_n$ maximizes the likelihood, we have

$$\sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) = 0.$$

$f(x) = f(a)$

$+ \frac{f'(x)(x-a)}{1}$

$+ \frac{f''(\tilde{x})(x-a)^2}{2}$

Expanding this equation in a Taylor series, we get

$\hat{\theta} \xrightarrow{P} \theta$

$|\hat{\theta} - \theta| < \varepsilon$

$$0 = \sum_{i=1}^{n} \ell'(X_i; \hat{\theta}_n) \quad = \sqrt{n} \underline{\sum_{i=1}^{n} \ell'(X_i; \theta)} +$$

$$\sqrt{n} + (\hat{\theta}_n - \theta) \sum_{i=1}^{n} \ell''(X_i; \theta)$$

$$\sqrt{n} + \frac{1}{2}(\hat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(X_i; \theta_n^*)$$

with θ_n^* lying between θ and $\hat{\theta}_n$.

Dividing accross by $\sqrt{n}$ yields $\qquad \sqrt{n}(\hat{\theta}-\theta)$

$$
\begin{aligned}
0 \quad = \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i;\theta) + \boxed{\sqrt{n}(\hat{\theta}_n - \theta)} \frac{1}{n} \sum_{i=1}^{n} \ell''(X_i;\theta) \\
& + \frac{1}{2} \boxed{\sqrt{n}(\hat{\theta}_n - \theta)^2} \frac{1}{n} \sum_{i=1}^{n} \ell'''(X_i;\theta_n^*)
\end{aligned}
$$

$\mathbb{E}[\cdot] = 0$

which suggests that $\sqrt{n}(\hat{\theta}_n - \theta)$ equals

$$
\frac{-n^{-1/2} \sum_{i=1}^{n} \ell'(X_i;\theta)}{n^{-1} \sum_{i=1}^{n} \ell''(X_i;\theta) + (\hat{\theta}_n - \theta)(2n)^{-1} \sum_{i=1}^{n} \ell'''(X_i;\theta_n^*)} .
$$

$\to 0$

Now, from the central limit theorem and condition (A4), it follows that

$$
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \ell'(X_i;\theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1(\theta)).
$$

Next, the weak law of large numbers along with condition (A5) implies

$$\frac{1}{n}\sum_{i=1}^{n} \ell''(X_i;\theta) \xrightarrow{p} -\mathcal{J}(\theta).$$

By Slutsky's lemma, the theorem will follow if we show that $R_n \xrightarrow{p} 0$. This is established in the next lemma, which we appeal to, completing the proof. □

Lemma

In the same context as in the previous theorem,

$$R_n = (\hat{\theta}_n - \theta)\frac{1}{2n}\sum_{i=1}^{n} \ell'''(X_i;\theta_n^*) \xrightarrow{p} 0$$

for any random variable θ_n^* on the segment joining $\hat{\theta}_n$ and θ.

Next, the weak law of large numbers along with condition (A5) implies

$$\frac{1}{n}\sum_{i=1}^{n} \ell''(X_i;\theta) \xrightarrow{p} -\mathcal{J}(\theta).$$

By Slutsky's lemma, the theorem will follow if we show that $R_n \xrightarrow{p} 0$. This is established in the next lemma, which we appeal to, completing the proof. $\square$

Lemma

In the same context as in the previous theorem,

$$R_n = (\hat{\theta}_n - \theta)\frac{1}{2n}\sum_{i=1}^{n} \underbrace{\ell'''(X_i;\theta_n^*)}_{\leq M(x)} \xrightarrow{p} 0$$

for any random variable θ_n^* on the segment joining $\hat{\theta}_n$ and θ.

Proof. $(*)$

We have that for any $\epsilon > 0$

$$\mathbb{P}[|R_n| > \epsilon] = \underbrace{\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| > \delta]}_{\leq \mathbb{P}[|\hat{\theta}_n - \theta| > \delta] \xrightarrow{p} 0} + \mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| \leq \delta]$$

→ consistency of MLE

If $|\hat{\theta}_n - \theta| < \delta$, (A6) implies $|R_n| \leq \frac{\delta}{2n} \sum_{i=1}^n M(X_i) = \bar{M}_n$.
so we may write

$$\mathbb{P}[|R_n| > \epsilon, |\hat{\theta}_n - \theta| \leq \delta] \leq \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \bar{M}_n]$$

$\approx M$

$= \mathbb{P}(\epsilon \leq |R_n| \leq \frac{1}{2}\delta\bar{M}_n)$

and for $\xi > 0$, the last term can be bounded by

$$\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \bar{M}_n, \bar{M}_n \leq M + \xi] +$$
$$+\mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta \bar{M}_n, \bar{M}_n > M + \xi]$$

which in turn is bounded by

$$\begin{aligned} &\leq \quad \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)] + \mathbb{P}[\bar{M}_n > M + \xi] \\ &\leq \quad \mathbb{P}[|R_n| > \epsilon, |R_n| \leq (1/2)\delta(M + \xi)] + \mathbb{P}[|\bar{M}_n - M| > \xi] \end{aligned}$$

But the law of large numbers implies that

$$\bar{M}_n = \frac{1}{n} \sum_{i=1}^n M(X_i) \xrightarrow{p} \mathbb{E}[M(X_1)] < \infty,$$

It follows that

$$\mathbb{P}[|\bar{M}_n - M| > \xi] \to 0.$$

Since we can always choose δ to be as small as we wish, we can make the term

$$\mathbb{P}[|R_n| > \epsilon, |R_n| \leq \underbrace{(1/2)\overset{\downarrow}{\delta}(M+\xi)}_{\varepsilon+\gamma}] \longrightarrow 0$$

equal to zero. In summary, we have established that $R_n \xrightarrow{p} 0$

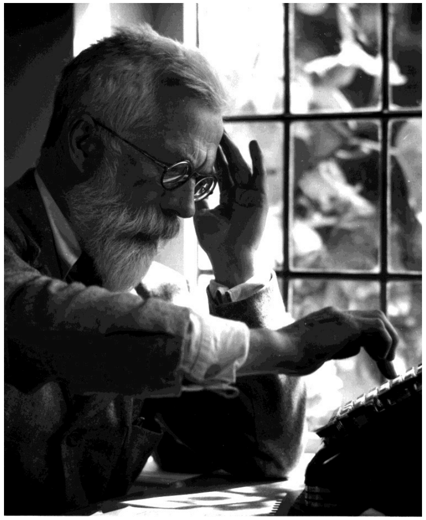
□

Does this mean that likelihood estimators are essentially optimal?

$\hat{\theta} \xrightarrow{P} \theta \Rightarrow \mathrm{Bias}(\hat{\theta}) \to 0$

Does this mean that likelihood estimators are essentially optimal?

- The result holds asymptotically in n, so care must be taken in interpreting it.
- For finite sample size n, the theorem says very little.
- Though bias must vanish asymptotically for consistency to go through...
- ... a little bit of bias can help reduce variance in finite samples.
- The delicate finite-sample tradeoff of bias and variance is decisive.
- Manifested both in parametric and (quite lucidly) nonparametric estimation.



Here's a spectacularly simple (and surprising) counterexample by Charles Stein.

Stein's setup

1. Let $Y_1, \ldots, Y_n$ be independent random variables.
2. Assume that $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
   - Notice that each Y_i has a different mean but same variance.
3. Suppose that σ^2 is known, say $\sigma^2 = 1$ (wlog)
4. Unknown parameter to estimate: $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top \in \mathbb{R}^n$
5. Consider mean squared error to judge quality.

$\hookrightarrow$ Looks like the usual setup, but notice the subtlety: the dimension of the parameter $\dim(\mu)=n$ grows along with the dimension of the sample size.

Is this artificial? No: many modern problems have # parameters comparable to # observations.

$\hookrightarrow$ Will later see other examples with parameter dimension fixed relative to sample size (ridge regression).

By independence, the loglikelihood in Stein's setup is

$$\ell(\boldsymbol{\mu}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu_i)^2$$

and by differentiation and convexity, we have

$$\underline{\hat{\boldsymbol{\mu}} = (Y_1, ..., Y_n)^\top} \quad \frac{1}{n}\Sigma Y_i \text{ for } n=1$$

is the unique MLE of μ.

- Intuition: we essentially have n Gaussian mean separate problems, each of sample size 1.
- Hence separately estimate each of these means by corresponding sample mean (which is Y_i since there is only 1 observation in each sample)

The MSE of this estimator can be easily calculated to be equal to n:

$$\mathrm{MSE}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \underline{\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2} = \sum_{i=1}^{n}\underline{(Y_i - \mu_i)^2} = n.$$

Stein realised that one can always improve this MSE by cleverly introducing bias...

Theorem (James-Stein)

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ be such that $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, I_{n\times n})$, $\boldsymbol{\mu} \in \mathbb{R}^n$ (Stein's setup). Let $\tilde{\boldsymbol{\mu}}_a$ be an estimator defined as

$$\tilde{\boldsymbol{\mu}}_a = \left(1 - \frac{a}{\|\boldsymbol{Y}\|^2}\right)\boldsymbol{Y} = \left(1 - \frac{a}{\|\hat{\boldsymbol{\mu}}\|^2}\right)\hat{\boldsymbol{\mu}},$$

$\leq \|\hat{\boldsymbol{\mu}}\|$

i.e. a *shrunken* version of the MLE $\hat{\boldsymbol{\mu}}$. Then, if $n \geq 3$,

1. for all $a \in (0, 2n-4)$,

$$\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) \leq \mathrm{MSE}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$$

2. for $a = n-2$,

$$\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_{n-2}, 0) < \mathrm{MSE}(\hat{\boldsymbol{\mu}}, 0)$$

3. For all $\boldsymbol{\mu} \in \mathbb{R}^n$ and all $a \in (0, 2n-4)$,

$$\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_{n-2}, \boldsymbol{\mu}) \leq \mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}).$$

$\leq \mathrm{MSE}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$

Comments:

- The result is surprising, not just because the MLE is outperformed.
- The JS estimator takes the MLE and shrinks it towards zero.
- The amount of shrinkage depends on $\|\boldsymbol{Y}\|$
- That is, we take into account the estimate of μ_i in order to estimate μ_j ($i \neq j$), even though these are completely unrelated (no "smoothness" assumptions on $\boldsymbol{\mu}$).
- The performance of the MLE as compared to the JS estimator becomes worse and worse as n grows.
- The proof is surprisingly elementary (once one knows what to look for!)

We'll need a simple lemma first.

$\int u\,dv = uv - \int v\,du$

Lemma (∗).

Let $Y \sim \mathcal{N}(\theta, \sigma^2)$ and $h : \mathbb{R} \to \mathbb{R}$ be differentiable. If

1. $\mathbb{E}|h(Y)| < \infty$,
2. $\lim_{y \to \pm\infty} \left\{ h(y) \exp\left[-\frac{1}{2\sigma^2}(y-\theta)^2 \right] \right\} = 0,$

then

$$\mathbb{E}[h(Y)(Y-\theta)] = \sigma^2 \mathbb{E}\left[h'(Y)\right].$$

Proof (∗).

By definition, $\mathbb{E}[h(Y)(Y-\theta)] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} h(y)(y-\theta) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} dy$.

Integration by parts transforms the right hand side into

$$\underbrace{-\frac{\sigma^2}{\sigma\sqrt{2\pi}} \left(h(y) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} \right)\Big|_{-\infty}^{+\infty}}_{=0} + \underbrace{\frac{\sigma^2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} h'(y) e^{-\frac{1}{2\sigma^2}(y-\theta)^2} dy}_{=\sigma^2 \mathbb{E}[h'(Y)]}$$

□

Proof of the James-Stein Theorem. $(*)$

$(b-c)^2$

$$
\begin{aligned}
\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a, \boldsymbol{\mu}) &= \mathbb{E}\left\| \overbrace{\left(1 - \frac{a}{\|\boldsymbol{Y}\|^2}\right)\boldsymbol{Y}}^{\tilde{\mu}_a} - \boldsymbol{\mu} \right\|^2 = \mathbb{E}\left\| \overbrace{\boldsymbol{Y} - \boldsymbol{\mu}}^{b} - \frac{\overbrace{a\boldsymbol{Y}}^{c}}{\|\boldsymbol{Y}\|^2} \right\|^2 \\
&= \underbrace{\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{\mu}\|^2}_{\mathrm{MSE}(\hat{\mu},\mu)} - 2\mathbb{E}\left(\frac{a\boldsymbol{Y}^\top(\boldsymbol{Y} - \boldsymbol{\mu})}{\|\boldsymbol{Y}\|^2}\right) + \mathbb{E}\left[\frac{a^2\|\boldsymbol{Y}\|^2}{\|\boldsymbol{Y}\|^4}\right] \\
&= n - 2a\sum_{i=1}^{n} \mathbb{E}\left[\frac{Y_i(Y_i - \mu_i)}{\sum_{j=1}^{n} Y_j^2}\right] + a^2\mathbb{E}\left[\frac{1}{\|\boldsymbol{Y}\|^2}\right]
\end{aligned}
$$

Now define n differentiable functions $h_i : \mathbb{R}^n \to \mathbb{R}$ by

$$
\boldsymbol{u} = (u_1, \ldots, u_n) \overset{h_i}{\mapsto} \boxed{\frac{u_i}{u_i^2 + \sum_{j\neq i}^{n} u_j^2}} \longrightarrow = \|u\|^2 = \sum_i u_i^2
$$

and observe that, for all $i \in \{1, \ldots, n\}$ and all $\{u_j\}_{j\neq i} \in \mathbb{R}^{n-1}$,

$$
\lim_{u_i \to \pm\infty} \left\{ h_i(\boldsymbol{u}) \exp\left[-\tfrac{1}{2\sigma^2}(u_i - \mu_i)^2\right] \right\} = 0,
$$

where we note that h_i becomes an $\mathbb{R} \to \mathbb{R}$ function once $\{u_j\}_{j\neq i} \in \mathbb{R}^{n-1}$ is fixed.

We now use the tower property and apply our lemma to re-write $\mathbb{E}\left[\frac{Y_i(Y_i-\mu_i)}{\sum_{j=1}^n Y_j^2}\right]$ as

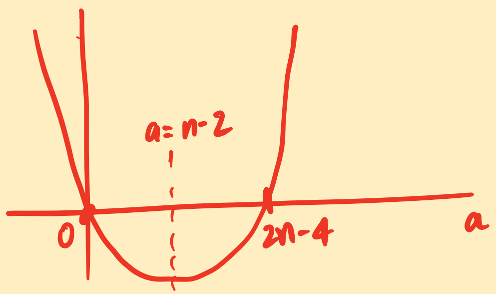
Lemma

$$\mathbb{E}\left\{\mathbb{E}\left[\frac{Y_i}{Y_i^2+\sum_{j\neq i}Y_j^2}(Y_i-\mu_i)\,\middle|\,\{Y_j\}_{j\neq i}\right]\right\} = \mathbb{E}\{\mathbb{E}\left[h_i(\boldsymbol{Y})(Y_i-\mu_i)|\,\{Y_j\}_{j\neq i}\right]\} =$$

$$\overset{\sigma^{-2}}{=}\mathbb{E}\left\{\mathbb{E}\left[\frac{\partial}{\partial u_i}\, h_i(\boldsymbol{u})|_{\boldsymbol{u}=\boldsymbol{Y}}\,|\{Y_j\}_{j\neq i}\right]\right\} = \mathbb{E}\left[\frac{\partial}{\partial u_i}\, h_i(\boldsymbol{u})|_{\boldsymbol{u}=\boldsymbol{Y}}\right] = \mathbb{E}\left[\frac{\|\boldsymbol{Y}\|^2-2Y_i^2}{\|\boldsymbol{Y}\|^4}\right]$$

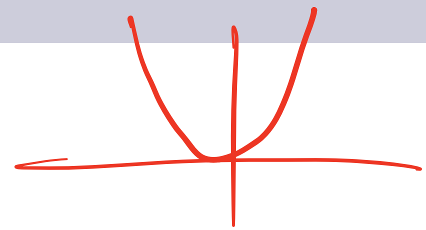
It follows that the MSE can be written as

$$\begin{aligned}\mathrm{MSE}(\tilde{\boldsymbol{\mu}}_a,\boldsymbol{\mu}) &= \overset{\mathrm{MSE}(\hat{\mu},\mu)}{n} - 2a\mathbb{E}\left[\frac{n\|\boldsymbol{Y}\|^2-2\|\boldsymbol{Y}\|^2}{\|\boldsymbol{Y}\|^4}\right] + a^2\mathbb{E}\left[\frac{1}{\|\boldsymbol{Y}\|^2}\right] \\ &= n + \underbrace{[a^2-2a(n-2)]}_{<0}\underbrace{\mathbb{E}\left[\frac{1}{\|\boldsymbol{Y}\|^2}\right]}_{>0}.\end{aligned}$$



Now, the polynomial $p(a) = a^2 - 2a(n-2)$ is strictly negative in the range $(0, 2n-4)$. Therefore, we have proven part (1). Furthermore, on the same range, $p(a)$ has a unique minimum at $a = n-2$, which proves part (3). For part (2), note that if $\boldsymbol{\mu} = 0$, $\|\boldsymbol{Y}\|^2 \sim \chi^2_n$, so $\mathbb{E}[1/\|\boldsymbol{Y}\|^2] = 1/(n-2)$ (recall that $n \geq 3$). Consequently, $\mathrm{MSE}(\underset{\tilde{\mu}_{(n-2)}}{\cancel{\delta_{n-2}}}, 0) = 2$.

□



Much of our discussion can be extended to cases where the MSE is replaced by some other convex measure of performance.

One can formulate a general framework as follows:

- Replace $\|\hat{\theta} - \theta\|$ by different deviation measure $\mathcal{L}(\hat{\theta}, \theta)$ called a loss function.
- The expected loss is then called the risk, $\mathcal{L} = \|\hat{\theta} - \theta\|$

$$R(\hat{\theta}, \theta) = \mathbb{E}[\mathcal{L}(\hat{\theta}, \theta)].$$

- The choice of loss function can be crucial and must be made judiciously.

Example (Exponential Distribution)

Let $Y_1, \dots, Y_n \overset{iid}{\sim} \text{Exponential}(\lambda)$, $n \geq 2$. The MLE of λ is

$$\hat{\lambda} = \frac{1}{\bar{Y}} = \frac{1}{\frac{1}{n}\sum Y_i}$$

with $\bar{Y}$ the empirical mean. We can easily calculate

$$\mathbb{E}[\hat{\lambda}] = \frac{n\lambda}{n-1}.$$

It follows that $\tilde{\lambda} = (n-1)\hat{\lambda}/n$ is an unbiased estimator of λ. Observe now that

$$\mathrm{MSE}(\tilde{\lambda}) < \mathrm{MSE}(\hat{\lambda})$$

since $\tilde{\lambda}$ is unbiased and $\mathrm{var}(\tilde{\lambda}) < \mathrm{var}(\hat{\lambda})$. Hence the $\hat{\lambda}$ is strictly dominated by $\tilde{\lambda}$.

Observe that the parameter space here is $(0, \infty)$:

- In such cases, quadratic loss penalises over-estimation more heavily than under-estimation
- The maximum possible under-estimation is bounded!
- What happens if we change the loss function to account for that?

Example (Exponential Distribution, continued)

Consider a different loss function

$$\mathcal{L}(a, b) = a/b - 1 - \log(a/b)$$

where, for each fixed a, $lim_{b\to 0}\mathcal{L}(a, b) = lim_{b\to\infty}\mathcal{L}(a, b) = \infty$.

Now, for $n > 1$,

$$\begin{aligned} R(\lambda, \tilde{\lambda}) &= \mathbb{E}_\lambda\left[\frac{n\lambda\bar{Y}}{n-1} - 1 - \log\left(\frac{n\lambda\bar{Y}}{n-1}\right)\right] \\ &= \underbrace{\mathbb{E}_\lambda\left[\lambda\bar{Y} - 1 - \log(\lambda\bar{Y})\right]}_{R(\lambda,\hat{\lambda})} + \underbrace{\frac{\mathbb{E}_\lambda(\lambda\bar{Y})}{n-1} - \log\left(\frac{n}{n-1}\right)}_{\substack{g(n) \\ >0}} \end{aligned}$$

where we wrote $\bar{Y} = \frac{n-1}{n}\bar{Y} + \frac{1}{n}\bar{Y}$. Note that $\mathbb{E}_\lambda[\bar{Y}] = \lambda^{-1}$, so

$$g(n) = \frac{1}{n-1} - \log\left(\frac{n}{n-1}\right).$$

We claim that $g(n) > 0$ for $n \geq 2$.

Example (Exponential Distribution, continued)

Using $\log x = \int_1^x t^{-1} dt$, this follows if

$$
\begin{aligned}
\frac{1}{x} &> \log(x+1) - \log x, \qquad x > 1 \\
\iff \frac{1}{x} &> \int_x^{x+1} t^{-1} dt, \qquad x > 1
\end{aligned}
$$

which holds by a rectangle area bound on the integral, as follows:

$$
\frac{1}{x} = [(x+1) - x]\frac{1}{x} = \int_x^{x+1} \frac{1}{x} dt > \int_x^{x+1} \frac{1}{t} dt, \quad \text{when } x > 1
$$

Consequently, $R(\tilde{\lambda}, \lambda) > R(\hat{\lambda}, \lambda)$ and $\hat{\lambda}$ dominates $\tilde{\lambda}$.

Decision Theory

We can push generality even further, and obtain an all encompassing framework.

Called decision theory, it views inference as a game between nature and the statistician.

We can push generality even further, and obtain an all encompassing framework.

Called decision theory, it views inference as a game between nature and the statistician.

Recall our general framework for statistical inference:

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

We can push generality even further, and obtain an all encompassing framework.

Called decision theory, it views inference as a game between nature and the statistician.

Recall our general framework for statistical inference:

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

The decision theory framework formalises step (4) to include estimation, testing, and confidence intervals.

The decision theory framework has the following elements:

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.

- A *parameter space* Θ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves available to Nature*.

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.

- A *parameter space* Θ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves available to Nature*.

- A *data space* $\mathcal{Y}^n$, on which the parametric family is supported. This represents the *space of possible outcomes* following a play by Nature.

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.
- A *parameter space* Θ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves available to Nature*.
- A *data space* $\mathcal{Y}^n$, on which the parametric family is supported. This represents the *space of possible outcomes* following a play by Nature.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* or *decisions* or *plays/moves available to the statistician*.

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.
- A *parameter space* Θ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves available to Nature*.
- A *data space* $\mathcal{Y}^n$, on which the parametric family is supported. This represents the *space of possible outcomes* following a play by Nature.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* or *decisions* or *plays/moves available to the statistician*.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{Y}^n \to \mathcal{A}$. These represent the *possible strategies* available to the statistician.

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.
- A *parameter space* Θ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves available to Nature*.
- A *data space* $\mathcal{Y}^n$, on which the parametric family is supported. This represents the *space of possible outcomes* following a play by Nature.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* or *decisions* or *plays/moves available to the statistician*.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{Y}^n \to \mathcal{A}$. These represent the *possible strategies* available to the statistician.
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. This represents *how much the statistician has to pay* nature when losing.

The decision theory framework has the following elements:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies). This is the *variant of the game we decide to play*.
- A *parameter space* Θ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$. This represents the space of possible *plays/moves available to Nature*.
- A *data space* $\mathcal{Y}^n$, on which the parametric family is supported. This represents the *space of possible outcomes* following a play by Nature.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* or *decisions* or *plays/moves available to the statistician*.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{Y}^n \to \mathcal{A}$. These represent the *possible strategies* available to the statistician.
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. This represents *how much the statistician has to pay* nature when losing.

Choice of $\mathcal{A}$ determines what inference we are making. Choice of $\mathcal{D}$ determines what class of procedures we are willing to entertain. Choice of $\mathcal{L}$ determines how we measure our errors.

The statistician would like to pick strategy δ so as to limit his losses. But the losses are random, which is why risk comes into play.

Given a decision rule $\delta : \mathcal{Y}^n \to \mathcal{A}$, the risk is $R(\delta, \theta) = \mathbb{E}\left[\mathcal{L}(\delta(\boldsymbol{Y}), \theta)\right]$.

δ : MLE estimation.

The statistician would like to pick strategy δ so as to limit his losses. But the losses are random, which is why risk comes into play.

Given a decision rule $\delta : \mathcal{Y}^n \to \mathcal{A}$, the risk is $R(\delta, \theta) = \mathbb{E}\left[\mathcal{L}(\delta(\boldsymbol{Y}), \theta)\right].$

The key principle of decision theory is that

$$\boxed{\text{decision rules should be compared by comparing their risk functions}}$$

The statistician would like to pick strategy δ so as to limit his losses. But the losses are random, which is why risk comes into play.

Given a decision rule $\delta : \mathcal{Y}^n \to \mathcal{A}$, the risk is $R(\delta, \theta) = \mathbb{E}\left[\mathcal{L}(\delta(\boldsymbol{Y}), \theta)\right]$.

The key principle of decision theory is that

$$\boxed{\text{decision rules should be compared by comparing their risk functions}}$$

- Risk varies depending on true state of nature, though.

$\theta \in \Theta$

- So comparisons can be made in different ways:
  1. Uniform (hard). Seek dominance everywhere in Θ.
  2. Minimax (relaxed). Compare worst-case risks over Θ.
  3. Bayes (relaxed). Compare average risk over Θ

Will not go into details, but will give two definitions for educational purposes.

Rather than look at risk at every θ minimax risk concentrates on maximum risk

Definition (Minimax Decision Rule)

Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{f_\theta\}_{\theta \in \Theta}, \mathcal{L})$. If

$$\sup_{\theta \in \Theta} R(\theta, \delta) \leq \sup_{\theta \in \Theta} R(\theta, \delta'), \quad \forall\ \delta' \in \mathcal{D},$$

then δ is called a minimax decision rule.

Rather than look at risk at every θ minimax risk concentrates on maximum risk

Definition (Minimax Decision Rule)

Let $\mathcal{D}$ be a class of decision rules for an experiment $(\{f_\theta\}_{\theta\in\Theta}, \mathcal{L})$. If

$$\sup_{\theta\in\Theta} R(\theta,\delta) \le \sup_{\theta\in\Theta} R(\theta,\delta'), \quad \forall\ \delta' \in \mathcal{D},$$

then δ is called a minimax decision rule.

Rather than look at risk at every θ Bayes risk concentrates on average risk

Definition (Bayes Risk)

Let $\pi(\theta)$ be a probability density (frequency) on Θ and let δ be a decision rule for the experiment $(\{f_\theta\}_{\theta\in\Theta}, \mathcal{L})$. The π-Bayes risk of δ is defined as

$\mathbb{E}_\pi[R]$

$$r(\pi,\delta) = \int_\Theta R(\theta,\delta)\underline{\pi(\theta)}d\theta = \int_\Theta\int_\mathcal{X} \mathcal{L}(\theta,\delta(\mathbf{y}))f_\theta(\mathbf{y})d\mathbf{y}\pi(\theta)d\theta$$

If $\delta \in \mathcal{D}$ is such that $r(\pi,\delta) \le r(\pi,\delta')$ for all $\delta' \in \mathcal{D}$, then δ is called a *Bayes decision rule* with respect to π.

The prior $\pi(\theta)$ places different emphasis for different values of θ based on our prior interest/knowedge.

Comments:

- Minimax rules are useful to establish the fundamental inferential complexity or a statistical experiment.
- But using them for more practical purposes requires caution.
- Motivated as follows: we do not know anything about θ so let us insure ourselves against the worst thing that can happen.
- Makes sense if you are in a zero-sum adversarial game: if your opponent chooses θ to maximize $\mathcal{L}$ then one should look for minimax rules.
- If there is no reason to believe that "nature" is trying to "do her worst", then the minimax approach is overly conservative: it places emphasis on the "bad θ".
- Bayes rules are quite attractive as they can nearly never be uniformly dominated.
- Intuitively, if you can show your rule to be Bayes for a nice prior, you know you're doing reasonably well.

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

$\Theta = \Theta_0 \cup \Theta_1$

The first sort of assertion we wish to make is:

1. **Hypothesis Testing**. Given two disjoint regions Θ_0 and Θ_1, which is more plausible to contain the true θ that generated our observation $(Y_1, \ldots, Y_n)^\top$?

The context:

1. We know that the true parameter lies in one of two subsets: Θ_0 or Θ_1, with $\Theta_0 \cap \Theta_1 = \emptyset$.

2. We need to use the sample $(Y_1, .., Y_n)^\top$ at hand to decide between the two possibilities.

3. This situation presents itself often in science, where two concurrent theories need to be confronted with the empirical evidence.

    1. The null hypothesis H_0 which states that $\theta \in \Theta_0$,

    $$H_0 : \theta \in \Theta_0,$$

    and

    2. The alternative hypothesis that postulates $\theta \in \Theta_1$,

    $$H_1 : \theta \in \Theta_1. \qquad H_a$$

Example (Searching for the Higgs)

- One of the biggest questions of the last quarter century in physics: whether the infamous *Higgs boson* existed or not.
- Using the standard model of particle physics, we can calculate how many diphotons would be produced on average in the absence of Higgs' boson. Call this number $b > 0$.
- Similarly, we can calculate the additional mean number of diphotons produce if the Higgs boson existed. Call this number $s > 0$.
- Diphoton events are well-accounted to be Poissonian with mean (say) μ.

Our null hypothesis (no Higgs) is then

$$H_0 : \mu = b,$$

and the competing alternative is

$$H_1 : \mu = b + s.$$

□

Our decision must be based on the sample, so we need to define:

Definition (Test Function)

$\delta : \mathcal{Y}^n \to \mathcal{A}$

A test function is a map $\delta : \mathcal{Y}^n \to \{0, 1\}$.

Obtaining 0 or 1 must be decided on whether or not the sample satisfies a certain condition:

$$\delta(Y_1, \dots, Y_n) = \begin{cases} 1, & \text{if } T(Y_1, \dots, Y_n) \in C, \\ 0, & \text{if } T(Y_1, \dots, Y_n) \notin C, \end{cases}$$

where

- T is a statistic called a *test statistic* and
- C is a subset of the range of T, called *critical region*.

In compact form

$$\delta(Y_1, \dots, Y_n) = \mathbf{1}\{T(Y_1, \dots, Y_n) \in C\}.$$

- To choose good test functions we need to quantify the performance of a test function.

Remark that, obviously, δ is just a Bernoulli random variable:

$$\delta = \begin{cases} 1, & \text{with probability } \mathbb{P}[T(Y_1, \dots, Y_n) \in C], = p \\ 0, & \text{with probability } \mathbb{P}[T(Y_1, \dots, Y_n) \notin C]. \end{cases}$$

- So a good test function must have a sampling distribution concentrated around the right decision.
- The difference from point estimation is that our action space is discrete.
- Can we get an analogue of mean squared error?

Possible errors[1] to be made?

Action / Truth	H_0	H_1
0	☺ TP	Type II Error
1	Type I Error	☺ TN

By an abuse of terminology, we could define:

$$\mathrm{MSE}(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2], \qquad i \in \{0, 1\}.$$

Since δ is Bernoulli, and i takes values in $\{0, 1\}$, we have

$$\mathrm{MSE}(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2] = \mathbb{E}_\theta[|\delta - i|] = \begin{cases} \mathbb{E}_\theta[\delta], & \text{if } \theta \in \Theta_0, \\ 1 - \mathbb{E}_\theta[\delta], & \text{if } \theta \in \Theta_1. \end{cases}$$

$$= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_0, \\ 1 - \mathbb{P}_\theta[\delta = 1], & \text{if } \theta \in \Theta_1. \end{cases}$$

$$= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \mathbb{1}(\theta \in \Theta_0), \\ \mathbb{P}_\theta[\delta = 0], & \text{if } \theta \in \Theta_1. \end{cases}$$

[1] Potential asymmetry in practice: false positive VS false negative. Will return to this.

In decision theory terms, the action space is $\mathcal{A} = \{0, 1\}$ and the loss function is the so-called "0–1" loss,

$$\mathcal{L}(a, \theta) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \ \& \ a = 1 & \text{(Type I Error)} \\ 1 & \text{if } \theta \in \Theta_1 \ \& \ a = 0 & \text{(Type II Error)} \\ 0 & \text{otherwise} & \text{(No Error)} \end{cases}$$

i.e. we lose 1 unit whenever committing a type I or type II error.

The risk function then becomes

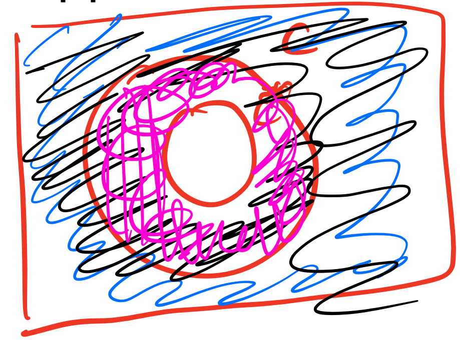
$$R(\delta, \theta) = \begin{cases} \mathbb{E}_\theta[\mathbf{1}\{\delta = 1\}] = \mathbb{P}_\theta[\delta = 1] & \text{if } \theta \in \Theta_0 & \text{(prob of type I error)} \\ \mathbb{E}_\theta[\mathbf{1}\{\delta = 0\}] = \mathbb{P}_\theta[\delta = 0] & \text{if } \theta \in \Theta_1 & \text{(prob of type II error)} \end{cases}$$

In short,

$$R(\delta, \theta) = \mathbb{P}_\theta[\delta = 1]\mathbf{1}\{\theta \in \Theta_0\} + \mathbb{P}_\theta[\delta = 0]\mathbf{1}\{\theta \in \Theta_1\}$$

Can we hope to simultaneously control both type I and II error probabilities? $\hookrightarrow$ Unfortunately the answer is no.

Here's why let $\delta(Y_1, \ldots, Y_n) = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C\}$ and suppose we wish to reduce the type I error probability



$$\mathbb{P}_\theta[\delta = 1], \qquad \theta \in \Theta_0,$$

for all $\theta \in \Theta_0$.

To do this, we must replace C by a subset $C_* \subset C$, obtaining

$$\delta_* = \mathbf{1}\{T(Y_1, \ldots, Y_n) \in C_*\}.$$

Observe that, $\forall\, \theta \in \Theta_0$,

$$\mathbb{P}_\theta[\delta_* = 1] = \mathbb{P}[T(Y_1, \ldots, Y_n) \in C_*] \leq \mathbb{P}[T(Y_1, \ldots, Y_n) \in C] = \mathbb{P}_\theta[\delta = 1]$$

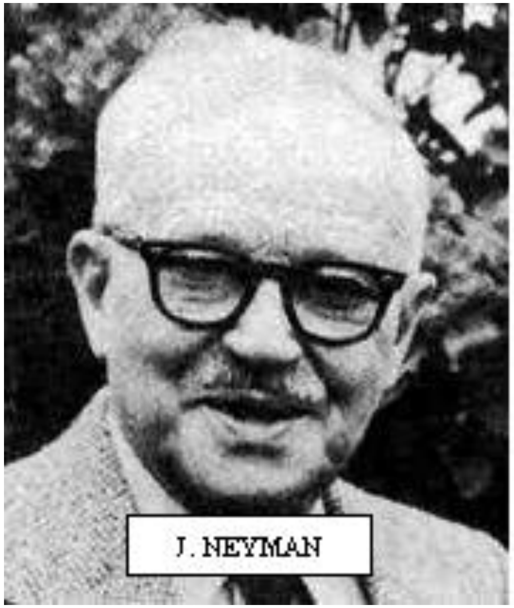
On the other hand $C_* \subset C \implies C_*^c \supset C^c$ and so $\forall\, \theta \in \Theta_1$

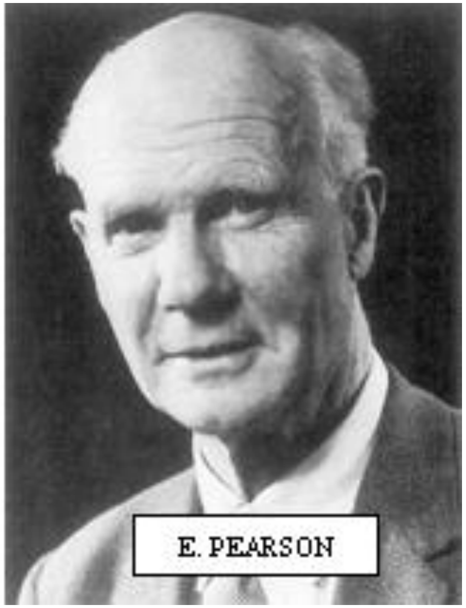
$$\mathbb{P}_\theta[\delta_* = 0] = \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C_*] \geq \mathbb{P}[T(Y_1, \ldots, Y_n) \notin C] = \mathbb{P}_\theta[\delta = 0].$$

$\in C_*^c$ $\qquad\qquad$ $\in C^c$

By reducing the type I error probability we increased the type II error probability

We need to make some concessions...





The fundamental paradigm of *Neyman and Pearson* informally dictates:

1. In applications, one type of error (false positive or negative) is typically more severe.
2. Say this is the type I error, and exploit the asymmetry: fix a tolerance ceiling for the probability of this error.
3. Given this ceiling, consider only test functions that respect it, and focus on minimising type II error (i.e. maximising power).

In mathematical terms:

The Neyman-Pearson Framework

1. We fix an $\alpha \in (0,1)$, usually small (called the significance level)
2. We declare that we only consider test functions $\delta : \mathcal{X} \to \{0,1\}$ such that

$$\delta \in \mathcal{D}(\Theta_0, \alpha) = \left\{\delta : \sup_{\theta\in\Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha\right\}$$

0.05

i.e. rules for which prob of type I error is bounded above by α

$\hookrightarrow$ *Jargon: we fix a significance level for our test*

3. Within this restricted class of rules, choose δ to minimize prob of type II error:

$$\min \ \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 0] = 1 - \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1], \qquad \theta \in \Theta_1$$

4. Equivalently, maximize the *power* $\iff$ max power function.

$$\beta(\theta, \delta) = \mathbb{P}_\theta[\delta(\boldsymbol{X}) = 1] = \mathbb{E}_\theta[\mathbf{1}\{\delta(\boldsymbol{X}) = 1\}] = \mathbb{E}_\theta[\delta(\boldsymbol{X})], \quad \theta \in \Theta_1$$

(since $\delta = 1 \iff \mathbf{1}\{\delta = 1\} = 1$ and $\delta = 0 \iff \mathbf{1}\{\delta = 1\} = 0$)

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0

Example: Biden VS Trump 2024. Pollsters gather iid sample $\boldsymbol{Y}$ from Florida with $Y_i = \mathbf{1}\{\text{vote Trump}\}$. Which pair of hypotheses to test?

$$\begin{cases} H_0: & \text{Trump wins Florida} \\ H_1: & \text{Biden wins Florida} \end{cases} \qquad \text{OR} \qquad \begin{cases} H_0: & \text{Biden wins Florida} \\ H_1: & \text{Trump wins Florida} \end{cases}$$

- Neyman-Pearson setup naturally exploits any asymmetric structure
- But, if natural asymmetry absent, need judicious choice of H_0

Example: Biden VS Trump 2024. Pollsters gather iid sample $\boldsymbol{Y}$ from Florida with $Y_i = \mathbf{1}\{\text{vote Trump}\}$. Which pair of hypotheses to test?

$$\begin{cases} H_0: & \text{Trump wins Florida} \\ H_1: & \text{Biden wins Florida} \end{cases} \quad \text{OR} \quad \begin{cases} H_0: & \text{Biden wins Florida} \\ H_1: & \text{Trump wins Florida} \end{cases}$$

- Which pair to choose to make a prediction? (confidence intervals?)
- If Trump is conducting poll to decide whether he'll spend more money to campaign in Florida, then his possible losses due to errors are:
  - (a) Spend more $'s to campaign in Florida even though he would win anyway: lose $'s
  - (b) Lose Florida to Biden because he thought he would win without any extra effort.
- (b) is much worse than (a) (especially since Trump had lots of $'s)
- Hence Trump would pick $H_0 = \{\text{Biden wins Florida}\}$ as his null

Consider the simplest situation:

$$\Theta_0 = \{\theta_0\} \qquad \& \qquad \Theta_1 = \{\theta_1\}$$

The Neyman-Pearson Lemma - Continuous Case

Let $\boldsymbol{Y}$ have joint density/frequency $f \in \{f_0, f_1\}$ and suppose we wish to test

$$H_0 : f = f_0 \qquad vs \qquad H_1 : f = f_1.$$

If $\Lambda(\boldsymbol{Y}) = f_1(\boldsymbol{Y})/f_0(\boldsymbol{Y})$ is a continuous random variable, then there exists a $k > 0$ such that

$f_1 \geq k f_0$

$$\mathbb{P}_0[\Lambda(\boldsymbol{Y}) \geq k] = \alpha$$

and the test whose test function is given by

$$\delta(\boldsymbol{Y}) = \mathbf{1}\{\Lambda(\boldsymbol{Y}) \geq k\},$$

is a *most powerful (MP)* test of H_0 versus H_1 at significance level α.

Proof.

Use obvious notation $\mathbb{E}_0$, $\mathbb{E}_1$, $\mathbb{P}_0$, $\mathbb{P}_1$ corresponding to H_0 or H_1. Let $G_0(t) = \mathbb{P}_0[\Lambda \leq t]$. By assumption, G_0 is a differentiable distribution function, and so is onto $[0, 1]$. Consequently, the set $\mathcal{K}_{1-\alpha} = \{t : G_0(t) = 1 - \alpha\}$ is non-empty for any $\alpha \in (0, 1)$. Setting $k = \inf\{t \in \mathcal{K}_{1-\alpha}\}$ we will have $\mathbb{P}_0[\Lambda \geq k] = \alpha$ and k is simply the $1 - \alpha$ quantile of the distribution G_0. Consequently,

$\mathbb{P}(\Lambda \leq k) = G(k)$
$= 1-\alpha$

$$\mathbb{P}_0[\delta = 1] = \alpha \qquad (\text{since } \mathbb{P}_0[\delta = 1] = \mathbb{P}_0[\Lambda \geq k])$$

and therefore $\delta \in \mathcal{D}(\{\theta_0\}, \alpha)$ (i.e. δ indeed respects the level α).
To show that δ is also most powerful, it suffices to prove that if ψ is any function with $\psi(\boldsymbol{y}) \in \{0, 1\}$, then

$$\mathbb{E}_0[\psi(\boldsymbol{Y})] \leq \underbrace{\mathbb{E}_0[\delta(\boldsymbol{Y})]}_{=\alpha \text{(by first part of proof)}} \implies \underbrace{\mathbb{E}_1[\psi(\boldsymbol{Y})]}_{\beta_1(\psi)} \leq \underbrace{\mathbb{E}_1[\delta(\boldsymbol{Y})]}_{\beta_1(\delta)}.$$

(recall that $\beta_1(\delta) = 1 - \mathbb{P}_1[\delta = 0] = \mathbb{P}_1[\delta = 1] = \mathbb{E}_1[\delta]$).