Statistics for Astronomers: Lecture 6, 2020.10.14

Prof. Sundar Srinivasan

IRyA/UNAM

Review

Frequentist statistical inference:

- Parametric (specify model, compute likelihood) vs.
  - nonparametric (performed on rank-ordered data).
- Estimation (point/interval) or hypothesis testing.
- Bayesian vs frequentist inference.
- Statistics and their desired properties.
- Estimators, estimates. Bias-variance tradeoff.
- Point estimates: likelihood.

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE)

A method of point estimation.

> “[T]he most probable set of values for the [model parameters] will make [the likelihood] a maximum.”
> “The likelihood that [the parameters] should have [an assigned set of values] is proportional to the probability that if this were so, the totality of observation should be that observed.”
>
> — R. A. Fisher, quoted in Feigelsen & Babu

Maximum Likelihood Estimation (MLE)

A method of point estimation.

> "[T]he most probable set of values for the [model parameters] will make [the likelihood] a maximum."
> "The likelihood that [the parameters] should have [an assigned set of values] is proportional to the probability that if this were so, the totality of observation should be that observed."
>
> — R. A. Fisher, quoted in Feigelsen & Babu

Procedure: For a vector of parameters $\vec{\theta}$, write down the functional form of the likelihood. Find the value of $\vec{\theta}$ at which this likelihood is maximum.

Maximum Likelihood Estimation (MLE)

A method of point estimation.

> "[T]he most probable set of values for the [model parameters] will make [the likelihood] a maximum."
> "The likelihood that [the parameters] should have [an assigned set of values] is proportional to the probability that if this were so, the totality of observation should be that observed."
>
> — R. A. Fisher, quoted in Feigelsen & Babu

Procedure: For a vector of parameters $\vec{\theta}$, write down the functional form of the likelihood. Find the value of $\vec{\theta}$ at which this likelihood is maximum.

1D example: $N = 10$ coin tosses result in $X = 8$ heads. Estimate $P(H)$.

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 (1-\theta)^2, \text{ with } 0 < \theta < 1.$$

Maximum Likelihood Estimation (MLE)

A method of point estimation.

> "[T]he most probable set of values for the [model parameters] will make [the likelihood] a maximum."
> "The likelihood that [the parameters] should have [an assigned set of values] is proportional to the probability that if this were so, the totality of observation should be that observed."
> — R. A. Fisher, quoted in Feigelsen & Babu

Procedure: For a vector of parameters $\vec{\theta}$, write down the functional form of the likelihood. Find the value of $\vec{\theta}$ at which this likelihood is maximum.

1D example: $N = 10$ coin tosses result in $X = 8$ heads. Estimate $P(H)$.

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 \,(1-\theta)^2, \text{ with } 0 < \theta < 1.$$

Use log-likelihood for convenience: $\ell(\theta) \equiv \ln \mathscr{L}(\theta) = \mathrm{constant} + 8 \ln \theta + 2 \ln (1-\theta)$.

Maximum Likelihood Estimation (MLE)

A method of point estimation.

"[T]he most probable set of values for the [model parameters] will make [the likelihood] a maximum."
"The likelihood that [the parameters] should have [an assigned set of values] is proportional to the probability that if this were so, the totality of observation should be that observed."
— R. A. Fisher, quoted in Feigelsen & Babu

Procedure: For a vector of parameters $\vec{\theta}$, write down the functional form of the likelihood. Find the value of $\vec{\theta}$ at which this likelihood is maximum.

1D example: $N = 10$ coin tosses result in $X = 8$ heads. Estimate $P(H)$.

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 \, (1-\theta)^2, \text{ with } 0 < \theta < 1.$$

Use log-likelihood for convenience: $\ell(\theta) \equiv \ln \mathscr{L}(\theta) = \text{constant} + 8 \ln \theta + 2 \ln (1 - \theta)$.

$$\frac{\partial}{\partial \theta} \ln \mathscr{L}(\theta) = \frac{8}{\theta} - \frac{2}{1-\theta};$$

Maximum Likelihood Estimation (MLE)

A method of point estimation.

"[T]he most probable set of values for the [model parameters] will make [the likelihood] a maximum."
"The likelihood that [the parameters] should have [an assigned set of values] is proportional to the probability that if this were so, the totality of observation should be that observed."
— R. A. Fisher, quoted in Feigelsen & Babu

Procedure: For a vector of parameters $\vec{\theta}$, write down the functional form of the likelihood. Find the value of $\vec{\theta}$ at which this likelihood is maximum.

1D example: $N = 10$ coin tosses result in $X = 8$ heads. Estimate $P(H)$.

$$\mathscr{L}(\theta) = P(X = 8, N = 10 \mid \theta) = \binom{10}{8} \theta^8 \, (1-\theta)^2, \text{ with } 0 < \theta < 1.$$

Use log-likelihood for convenience: $\ell(\theta) \equiv \ln \mathscr{L}(\theta) = \text{constant} + 8 \ln \theta + 2 \ln (1-\theta)$.

$$\frac{\partial}{\partial \theta} \ln \mathscr{L}(\theta) = \frac{8}{\theta} - \frac{2}{1-\theta}; \qquad \text{vanishes at } \theta = \hat{\theta}_{\mathrm{MLE}} \implies \hat{\theta}_{\mathrm{MLE}} = 0.8.$$

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathscr{L}(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right]$$

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

$$\Longrightarrow \ell \equiv \ln \mathcal{L}(\mu, \sigma^2) = \text{constant} - \frac{N}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma}\right)^2$$

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathscr{L}(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right] = \left(\frac{1}{2\pi\sigma^2} \right)^{N/2} \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right]$$

$$\implies \ell \equiv \ln \mathscr{L}(\mu, \sigma^2) = \mathrm{constant} - \frac{N}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right)$$

@ MLE: $\displaystyle \sum_{i=1}^{N} \left(\frac{x_i - \hat{\mu}}{\widehat{\sigma^2}} \right) = 0$

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathscr{L}(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

$$\implies \ell \equiv \ln \mathscr{L}(\mu, \sigma^2) = \text{constant} - \frac{N}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2$$

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma^2}\right)$$

@ MLE: $\sum_{i=1}^{N}\left(\frac{x_i - \hat{\mu}}{\widehat{\sigma^2}}\right) = 0$

$$\implies \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \equiv \bar{x}.$$

MLE of μ is the sample mean!

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathscr{L}(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

$$\implies \ell \equiv \ln \mathscr{L}(\mu, \sigma^2) = \text{constant} - \frac{N}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma}\right)^2$$

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N}\left(\frac{x_i - \mu}{\sigma^2}\right)$$

@ MLE: $\displaystyle\sum_{i=1}^{N}\left(\frac{x_i - \hat{\mu}}{\widehat{\sigma^2}}\right) = 0$

$$\implies \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \equiv \bar{x}.$$

MLE of μ is the sample mean!

$$\frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma}\left(-N + \sum_{i=1}^{N}\frac{(x_i - \hat{\mu})^2}{\sigma^2}\right)$$

@ MLE: $\displaystyle -N + \sum_{i=1}^{N}\frac{(x_i - \hat{\mu})^2}{\widehat{\sigma^2}} = 0$

MLE for iid Gaussian random variables

$\vec{\theta} = (\mu, \sigma^2)$. N observations $X_i (i = 1, \cdots, N) \sim \mathcal{N}(\mu, \sigma^2)$.

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp\left[-\frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right] = \left(\frac{1}{2\pi\sigma^2} \right)^{N/2} \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right]$$

$$\implies \ell \equiv \ln \mathcal{L}(\mu, \sigma^2) = \text{constant} - \frac{N}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right)$$

@ MLE: $\sum_{i=1}^{N} \left(\frac{x_i - \hat{\mu}}{\widehat{\sigma^2}} \right) = 0$

$$\implies \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \equiv \bar{x}.$$

MLE of μ is the sample mean!

$$\frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma} \left(-N + \sum_{i=1}^{N} \frac{(x_i - \hat{\mu})^2}{\sigma^2} \right)$$

@ MLE: $-N + \sum_{i=1}^{N} \frac{(x_i - \hat{\mu})^2}{\widehat{\sigma^2}} = 0$

$$\implies \widehat{\sigma^2} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

MLE of σ^2 is the (biased) sample variance!

What is the uncertainty on the MLE?

Coin-toss problem: assume that the true θ is $\theta_0 = 0.8$, unknown to observer.
Each round of ten tosses: different value of $\hat{\theta}_{\rm MLE}$ (*e.g.*, $[0.9, 0.7, 0.9, 0.8, 1., 0.9, 1., 0.9, 0.9, 0.8]$).
With finite # experiments, not enough to just quote $\hat{\theta}_{\rm MLE}$. What is the variance on the MLE?

What is the uncertainty on the MLE?

Coin-toss problem: assume that the true θ is $\theta_0 = 0.8$, unknown to observer.
Each round of ten tosses: different value of $\hat{\theta}_{\rm MLE}$ (*e.g.*, $[0.9, 0.7, 0.9, 0.8, 1., 0.9, 1., 0.9, 0.9, 0.8]$).
With finite # experiments, not enough to just quote $\hat{\theta}_{\rm MLE}$. What is the variance on the MLE?

Expand $\ln \mathscr{L}$ around θ_0:

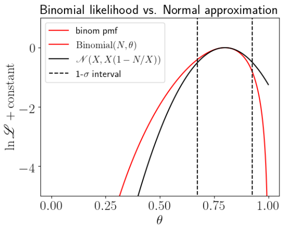
$$\ln\left[\frac{\mathscr{L}(\theta)}{\mathscr{L}(\theta_0)}\right] = \left(\frac{\partial^2}{\partial\theta^2}\ln\mathscr{L}(\theta)\right)_{\theta_0}\frac{(\theta-\theta_0)^2}{2!} + \cdots$$

$\ln \mathscr{L}$ "regular" if we can ignore higher-order terms.

What is the uncertainty on the MLE?

Coin-toss problem: assume that the true θ is $\theta_0 = 0.8$, unknown to observer.
Each round of ten tosses: different value of $\hat{\theta}_{\rm MLE}$ (*e.g.*, $[0.9, 0.7, 0.9, 0.8, 1., 0.9, 1., 0.9, 0.9, 0.8]$).
With finite # experiments, not enough to just quote $\hat{\theta}_{\rm MLE}$. What is the variance on the MLE?

Code for plot available here



Expand $\ln \mathscr{L}$ around θ_0:

$$\ln \left[\frac{\mathscr{L}(\theta)}{\mathscr{L}(\theta_0)}\right] = \left(\frac{\partial^2}{\partial\theta^2} \ln \mathscr{L}(\theta)\right)_{\theta_0} \frac{(\theta-\theta_0)^2}{2!} + \cdots$$

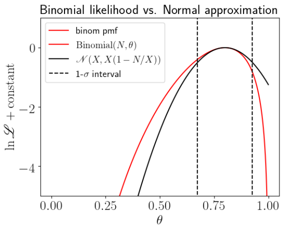
$\ln \mathscr{L}$ "regular" if we can ignore higher-order terms.
$\ln \mathscr{L}$ quadratic $\Longrightarrow \mathscr{L}$ Gaussian. Usually assumed.

Can describe $\ln \mathscr{L}$ with location θ_0 and curvature of $\ln \mathscr{L}$ at θ_0.

What is the uncertainty on the MLE?

Coin-toss problem: assume that the true θ is $\theta_0 = 0.8$, unknown to observer.
Each round of ten tosses: different value of $\hat{\theta}_{\rm MLE}$ (*e.g.*, $[0.9, 0.7, 0.9, 0.8, 1., 0.9, 1., 0.9, 0.9, 0.8]$).
With finite # experiments, not enough to just quote $\hat{\theta}_{\rm MLE}$. What is the variance on the MLE?

Code for plot available here



Expand $\ln \mathscr{L}$ around θ_0:

$$\ln \left[\frac{\mathscr{L}(\theta)}{\mathscr{L}(\theta_0)}\right] = \left(\frac{\partial^2}{\partial\theta^2} \ln \mathscr{L}(\theta)\right)_{\theta_0} \frac{(\theta-\theta_0)^2}{2!} + \cdots$$

$\ln \mathscr{L}$ "regular" if we can ignore higher-order terms.
$\ln \mathscr{L}$ quadratic $\Longrightarrow \mathscr{L}$ Gaussian. Usually assumed.

Can describe $\ln \mathscr{L}$ with location θ_0 and curvature of $\ln \mathscr{L}$ at θ_0.

Curvature defined as the negative second derivative of $\ln \mathscr{L}$ at location of maximum:

$$I(\theta) \equiv -\frac{\partial^2}{\partial\theta^2} \log \mathscr{L} \text{ (1-D)} \qquad I_{ij}(\vec{\theta}) \equiv -\frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j} \log \mathscr{L} \text{ (N-D)}$$ Fisher information matrix.

Large curvature near $\vec{\theta}_0$: less uncertainty (more information) about location of maximum.

What is the uncertainty on the MLE? (contd.)

N-D Taylor Expansion: $\ln\left[\frac{\mathscr{L}(\vec{\theta})}{\mathscr{L}(\vec{\theta}_0)}\right] = -\frac{1}{2}(\vec{\theta}-\vec{\theta}_0)\overbrace{\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\ln\mathscr{L}(\theta)\right]_{\theta_0}}^{\text{Fisher matrix}}(\vec{\theta}-\vec{\theta}_0)^{\mathrm{T}}$

What is the uncertainty on the MLE? (contd.)

N-D Taylor Expansion: $\ln\left[\frac{\mathscr{L}(\vec{\theta})}{\mathscr{L}(\vec{\theta}_0)}\right] = -\frac{1}{2}(\vec{\theta}-\vec{\theta}_0)\overbrace{\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\ln\mathscr{L}(\theta)\right]_{\theta_0}}^{\text{Fisher matrix}}(\vec{\theta}-\vec{\theta}_0)^{\mathrm{T}}$

The observer produces estimates for the Fisher matrix (random variable!) with every experiment.

Observed Fisher information: Fisher matrix evaluated at $\hat{\theta}_{\mathrm{MLE}}$.

What is the uncertainty on the MLE? (contd.)

N-D Taylor Expansion: $\ln\left[\frac{\mathscr{L}(\vec{\theta})}{\mathscr{L}(\vec{\theta}_0)}\right] = -\frac{1}{2}(\vec{\theta}-\vec{\theta}_0)\overbrace{\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\ln\mathscr{L}(\theta)\right]_{\theta_0}}^{\text{Fisher matrix}}(\vec{\theta}-\vec{\theta}_0)^{\mathrm{T}}$

The observer produces estimates for the Fisher matrix (random variable!) with every experiment.
Observed Fisher information: Fisher matrix evaluated at $\hat{\theta}_{\mathrm{MLE}}$.

To compare with the true value, define:

Average/Expected Fisher information: $\mathcal{I}(\vec{\theta}) \equiv \mathbb{E}[I(\vec{\theta})] = \mathbb{E}\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\log\mathscr{L}\right]$.

What is the uncertainty on the MLE? (contd.)

N-D Taylor Expansion: $\ln\left[\frac{\mathscr{L}(\vec{\theta})}{\mathscr{L}(\vec{\theta}_0)}\right] = -\frac{1}{2}(\vec{\theta}-\vec{\theta}_0)\overbrace{\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\ln\mathscr{L}(\theta)\right]_{\theta_0}}^{\text{Fisher matrix}}(\vec{\theta}-\vec{\theta}_0)^{\mathrm{T}}$

The observer produces estimates for the Fisher matrix (random variable!) with every experiment.
Observed Fisher information: Fisher matrix evaluated at $\hat{\theta}_{\mathrm{MLE}}$.

To compare with the true value, define:

Average/Expected Fisher information: $\mathcal{I}(\vec{\theta}) \equiv \mathbb{E}[I(\vec{\theta})] = \mathbb{E}\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\log\mathscr{L}\right]$.

The MLE is distributed around its expected value (= true value if MLE is unbiased) with a spread described by the Fisher matrix.

What is the uncertainty on the MLE? (contd.)

N-D Taylor Expansion: $\ln\left[\frac{\mathscr{L}(\vec{\theta})}{\mathscr{L}(\vec{\theta}_0)}\right] = -\frac{1}{2}(\vec{\theta}-\vec{\theta}_0)\overbrace{\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\ln\mathscr{L}(\theta)\right]_{\theta_0}}^{\text{Fisher matrix}}(\vec{\theta}-\vec{\theta}_0)^{\mathrm{T}}$

The observer produces estimates for the Fisher matrix (random variable!) with every experiment.
Observed Fisher information: Fisher matrix evaluated at $\hat{\theta}_{\mathrm{MLE}}$.

To compare with the true value, define:
Average/Expected Fisher information: $\mathcal{I}(\vec{\theta}) \equiv \mathbb{E}[I(\vec{\theta})] = \mathbb{E}\left[-\frac{\partial}{\partial\vec{\theta}}\frac{\partial}{\partial\vec{\theta}}\log\mathscr{L}\right]$.

The MLE is distributed around its expected value (= true value if MLE is unbiased) with a spread described by the Fisher matrix.

The inverse of the Expected Fisher matrix is the covariance matrix of the parameters:

$$\Sigma(\vec{\theta}) = \mathcal{I}^{-1}(\vec{\theta})$$

Fisher information for the ten coin-toss problem

Experiment: Ten coin tosses with unknown probability θ of obtaining a head.

$\mathscr{L}(\theta) \propto \theta^X (1-\theta)^{(N-X)}$, and $\mathbb{E}[X] = N\theta$.

Fisher information for the ten coin-toss problem

Experiment: Ten coin tosses with unknown probability θ of obtaining a head.

$\mathscr{L}(\theta) \propto \theta^X (1-\theta)^{(N-X)}$, and $\mathbb{E}[X] = N\theta$.

The Fisher Information is

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial^2\theta}\log\mathscr{L}\right] = \frac{N}{\theta(1-\theta)}.$$

Fisher information for the ten coin-toss problem

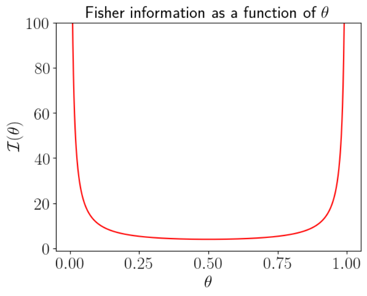
Experiment: Ten coin tosses with unknown probability θ of obtaining a head.

$\mathscr{L}(\theta) \propto \theta^X (1-\theta)^{(N-X)}$, and $\mathbb{E}[X] = N\theta$.

The Fisher Information is

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial^2\theta} \log \mathscr{L}\right] = \frac{N}{\theta(1-\theta)}.$$

Information highest near $\theta = 0$ and $\theta = 1$.



Code for plot available here.

Fisher information for the ten coin-toss problem

Experiment: Ten coin tosses with unknown probability θ of obtaining a head.

$\mathscr{L}(\theta) \propto \theta^X (1-\theta)^{(N-X)}$, and $\mathbb{E}[X] = N\theta$.

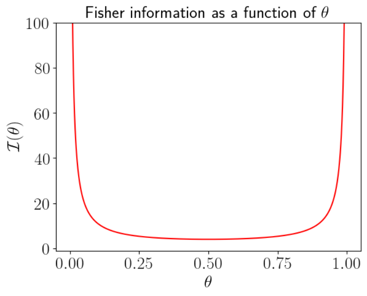
The Fisher Information is

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial^2\theta} \log \mathscr{L} \right] = \frac{N}{\theta(1-\theta)}.$$

Information highest near $\theta = 0$ and $\theta = 1$.

Variance of $\mathrm{Binomial}(N, \theta) = N\ \theta\ (1-\theta)$.

$$= \frac{1}{\mathcal{I}(\theta)} \text{ in this case!}$$



Code for plot available here.

Fisher information for the ten coin-toss problem

Experiment: Ten coin tosses with unknown probability θ of obtaining a head.

$\mathscr{L}(\theta) \propto \theta^X (1-\theta)^{(N-X)}$, and $\mathbb{E}[X] = N\theta$.

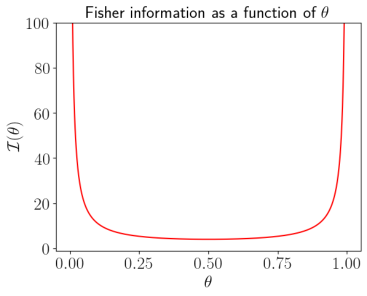
The Fisher Information is

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial^2\theta}\log\mathscr{L}\right] = \frac{N}{\theta(1-\theta)}.$$

Information highest near $\theta = 0$ and $\theta = 1$.

Variance of $\mathrm{Binomial}(N, \theta) = N\,\theta\,(1-\theta)$.

$$= \frac{1}{\mathcal{I}(\theta)} \text{ in this case!}$$



Code for plot available here.

Is this always true? Cramér-Rao Lower Bound.

Variance of unbiased estimators: Cramér-Rao Lower Bound

If $\vec{\mathbf{T}}(X)$ is an unbiased estimator of a function $\vec{\mathbf{g}}(\vec{\theta})$ of the parameters $\vec{\theta}$ (*i.e.*, $\mathbb{E}[\vec{\mathbf{T}}(X)] = \vec{\mathbf{g}}(\vec{\theta})$), and $\mathcal{I}(\vec{\theta})$ is the expected Fisher information matrix, then

Variance of unbiased estimators: Cramér-Rao Lower Bound

If $\vec{\mathbf{T}}(X)$ is an unbiased estimator of a function $\vec{\mathbf{g}}(\vec{\theta})$ of the parameters $\vec{\theta}$ (*i.e.*, $\mathbb{E}[\vec{\mathbf{T}}(X)] = \vec{\mathbf{g}}(\vec{\theta})$), and $\mathcal{I}(\vec{\theta})$ is the expected Fisher information matrix, then

$$\mathrm{Var}[\vec{\mathbf{T}}(X)] \geq \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right) \mathcal{I}^{-1}(\vec{\theta}) \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right)^{\mathrm{T}} \qquad \text{Cramér-Rao Lower Bound (CRLB)}$$

Variance of unbiased estimators: Cramér-Rao Lower Bound

If $\vec{\mathbf{T}}(X)$ is an unbiased estimator of a function $\vec{\mathbf{g}}(\vec{\theta})$ of the parameters $\vec{\theta}$ (*i.e.*, $\mathbb{E}[\vec{\mathbf{T}}(X)] = \vec{\mathbf{g}}(\vec{\theta})$), and $\mathcal{I}(\vec{\theta})$ is the expected Fisher information matrix, then

$$\mathrm{Var}[\vec{\mathbf{T}}(X)] \geq \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right) \mathcal{I}^{-1}(\vec{\theta}) \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right)^{\mathrm{T}} \qquad \text{Cramér-Rao Lower Bound (CRLB)}$$

In particular, if we set $g(\vec{\theta}) = \vec{\theta}$, so that $\vec{\mathbf{T}}(X)$ is an unbiased estimator for $\vec{\theta}$,

$$\mathrm{Var}[\vec{\mathbf{T}}(X)] \geq \mathcal{I}^{-1}(\vec{\theta})$$

The inverse of the Fisher Information ($\equiv$ covariance) of a parameter is a lower bound on the variance of any unbiased estimator of that parameter.

Variance of unbiased estimators: Cramér-Rao Lower Bound

If $\vec{\mathbf{T}}(X)$ is an unbiased estimator of a function $\vec{\mathbf{g}}(\vec{\theta})$ of the parameters $\vec{\theta}$ (*i.e.*, $\mathbb{E}[\vec{\mathbf{T}}(X)] = \vec{\mathbf{g}}(\vec{\theta})$), and $\mathcal{I}(\vec{\theta})$ is the expected Fisher information matrix, then

$$\mathrm{Var}[\vec{\mathbf{T}}(X)] \geq \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right) \mathcal{I}^{-1}(\vec{\theta}) \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right)^{\mathrm{T}}$$

Cramér-Rao Lower Bound (CRLB)

In particular, if we set $g(\vec{\theta}) = \vec{\theta}$, so that $\vec{\mathbf{T}}(X)$ is an unbiased estimator for $\vec{\theta}$,

$$\mathrm{Var}[\vec{\mathbf{T}}(X)] \geq \mathcal{I}^{-1}(\vec{\theta})$$

The inverse of the Fisher Information ($\equiv$ covariance) of a parameter is a lower bound on the variance of any unbiased estimator of that parameter.

Does not tell us if the estimator $\vec{\mathbf{T}}(X)$ exists, or how we can find it.

We can compute the variance for various $\vec{\mathbf{T}}(X)$ and choose the one with variance closest to the CRLB.

Variance of unbiased estimators: Cramér-Rao Lower Bound

If $\vec{\mathbf{T}}(X)$ is an unbiased estimator of a function $\vec{\mathbf{g}}(\vec{\theta})$ of the parameters $\vec{\theta}$ (*i.e.*, $\mathbb{E}[\vec{\mathbf{T}}(X)] = \vec{\mathbf{g}}(\vec{\theta})$), and $\mathcal{I}(\vec{\theta})$ is the expected Fisher information matrix, then

$$\mathrm{Var}[\vec{\mathbf{T}}(X)] \geq \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right) \mathcal{I}^{-1}(\vec{\theta}) \left(\frac{\partial \vec{\mathbf{g}}(\vec{\theta})}{\partial \vec{\theta}}\right)^{\mathrm{T}} \qquad \text{Cramér-Rao Lower Bound (CRLB)}$$

In particular, if we set $g(\vec{\theta}) = \vec{\theta}$, so that $\vec{\mathbf{T}}(X)$ is an unbiased estimator for $\vec{\theta}$,

$$\mathrm{Var}[\vec{\mathbf{T}}(X)] \geq \mathcal{I}^{-1}(\vec{\theta})$$

The inverse of the Fisher Information ($\equiv$ covariance) of a parameter is a lower bound on the variance of any unbiased estimator of that parameter.

Does not tell us if the estimator $\vec{\mathbf{T}}(X)$ exists, or how we can find it.

We can compute the variance for various $\vec{\mathbf{T}}(X)$ and choose the one with variance closest to the CRLB.

For biased estimators: If $\mathbb{E}[\vec{\mathbf{T}}(X) - \vec{\theta}] = \vec{B}(\vec{\theta}) \neq 0$, set $\vec{g}(\vec{\theta}) = \vec{B}(\vec{\theta}) + \vec{\theta}$ and apply CRLB.

Covariance matrix for MLE of Gaussian random variables

Recall:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right) \qquad \frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma} \left(-N + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right) \qquad \mathbb{E}\left[\sum_{i=1}^{N} (x_i - \mu)^2 \right] = N\,\sigma^2$$

Covariance matrix for MLE of Gaussian random variables

Recall:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right) \qquad \frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma} \left(-N + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right) \qquad \mathbb{E}\left[\sum_{i=1}^{N} (x_i - \mu)^2 \right] = N\,\sigma^2$$

Compute all three second derivatives:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{N}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 \qquad \frac{\partial^2 \ell}{\partial \sigma \partial \mu} = -\frac{2}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)$$

Covariance matrix for MLE of Gaussian random variables

Recall:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right) \qquad \frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma} \left(-N + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right) \qquad \mathbb{E}\left[\sum_{i=1}^{N} (x_i - \mu)^2 \right] = N\,\sigma^2$$

Compute all three second derivatives:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{N}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 \qquad \frac{\partial^2 \ell}{\partial \sigma \partial \mu} = -\frac{2}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)$$

Compute expectation values:

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \mu^2} \right] = -\frac{N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2} \right] = \frac{N}{\sigma^2} - \frac{3}{\sigma^2} N = -\frac{2N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma \partial \mu} \right] = 0 \text{ (uncorrelated!)}$$

Covariance matrix for MLE of Gaussian random variables

Recall:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right) \qquad \frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma} \left(-N + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right) \qquad \mathbb{E}\left[\sum_{i=1}^{N} (x_i - \mu)^2 \right] = N\,\sigma^2$$

Compute all three second derivatives:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{N}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 \qquad \frac{\partial^2 \ell}{\partial \sigma \partial \mu} = -\frac{2}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)$$

Compute expectation values:

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \mu^2} \right] = -\frac{N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2} \right] = \frac{N}{\sigma^2} - \frac{3}{\sigma^2} N = -\frac{2N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma \partial \mu} \right] = 0 \ (\text{uncorrelated!})$$

Expected Fisher matrix: $\mathcal{I}(\vec{\theta}) \equiv -\mathbb{E}\left[\frac{\partial}{\partial \vec{\theta}} \frac{\partial}{\partial \vec{\theta}} \ln \mathscr{L} \right] = \frac{1}{\sigma^2} \begin{bmatrix} N & 0 \\ 0 & 2N \end{bmatrix}$

Covariance matrix for MLE of Gaussian random variables

Recall:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right) \qquad \frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma} \left(-N + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right) \qquad \mathbb{E}\Big[\sum_{i=1}^{N} (x_i - \mu)^2 \Big] = N\,\sigma^2$$

Compute all three second derivatives:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{N}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 \qquad \frac{\partial^2 \ell}{\partial \sigma \partial \mu} = -\frac{2}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)$$

Compute expectation values:

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \mu^2} \right] = -\frac{N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2} \right] = \frac{N}{\sigma^2} - \frac{3}{\sigma^2} N = -\frac{2N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma \partial \mu} \right] = 0 \ (\text{uncorrelated!})$$

Expected Fisher matrix: $\mathcal{I}(\vec{\boldsymbol{\theta}}) \equiv -\mathbb{E}\left[\frac{\partial}{\partial \vec{\boldsymbol{\theta}}} \frac{\partial}{\partial \vec{\boldsymbol{\theta}}} \ln \mathscr{L} \right] = \frac{1}{\sigma^2} \begin{bmatrix} N & 0 \\ 0 & 2N \end{bmatrix}$

Covariance matrix: $\boldsymbol{\Sigma}(\vec{\boldsymbol{\theta}}) \equiv \mathcal{I}^{-1}(\vec{\boldsymbol{\theta}}) = \frac{\sigma^2}{N} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

Covariance matrix for MLE of Gaussian random variables

Recall:

$$\frac{\partial \ell}{\partial \mu} = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma^2} \right) \qquad \frac{\partial \ell}{\partial \sigma} = \frac{1}{\sigma} \left(-N + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2} \right) \qquad \mathbb{E}\left[\sum_{i=1}^{N} (x_i - \mu)^2 \right] = N\,\sigma^2$$

Compute all three second derivatives:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{N}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2 \qquad \frac{\partial^2 \ell}{\partial \sigma \partial \mu} = -\frac{2}{\sigma^3} \sum_{i=1}^{N} (x_i - \mu)$$

Compute expectation values:

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \mu^2} \right] = -\frac{N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2} \right] = \frac{N}{\sigma^2} - \frac{3}{\sigma^2} N = -\frac{2N}{\sigma^2} \qquad \mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma \partial \mu} \right] = 0 \ (\text{uncorrelated!})$$

Expected Fisher matrix: $\mathcal{I}(\vec{\boldsymbol{\theta}}) \equiv -\mathbb{E}\left[\frac{\partial}{\partial \vec{\boldsymbol{\theta}}} \frac{\partial}{\partial \vec{\boldsymbol{\theta}}} \ln \mathscr{L} \right] = \frac{1}{\sigma^2} \begin{bmatrix} N & 0 \\ 0 & 2N \end{bmatrix}$

Covariance matrix: $\boldsymbol{\Sigma}(\vec{\boldsymbol{\theta}}) \equiv \mathcal{I}^{-1}(\vec{\boldsymbol{\theta}}) = \frac{\sigma^2}{N} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

Variances = CRBL!

Computational MLE

Log-likelihood is fed to routine by user.

Routine optimises this function using a variety of techniques.

The output will include the MLE as well as the covariance matrix.

Example: fitting a line to data with uncertainties.

Point estimation: caveats (Feigelsen & Babu, ch. 3)

> "It is worth checking any piece of remembered statistics, as it is almost certain to be based on the Gaussian distribution."
>
> — Wall & Jenkins, Sec. 3.2

Point estimation: caveats (Feigelsen & Babu, ch. 3)

> "It is worth checking any piece of remembered statistics, as it is almost certain to be based on the Gaussian distribution."
>
> — Wall & Jenkins, Sec. 3.2

Point estimation requires two decisions:

Point estimation: caveats (Feigelsen & Babu, ch. 3)

> "It is worth checking any piece of remembered statistics, as it is almost certain to be based on the Gaussian distribution."
>
> — Wall & Jenkins, Sec. 3.2

Point estimation requires two decisions:

1. Model specification: required to compute the likelihood. How do we know it is correct?
   - Model validation (goodness-of-fit).
   - Model selection.

Point estimation: caveats (Feigelsen & Babu, ch. 3)

> "It is worth checking any piece of remembered statistics, as it is almost certain to be based on the Gaussian distribution."
>
> — Wall & Jenkins, Sec. 3.2

Point estimation requires two decisions:

1. **Model specification**: required to compute the likelihood. How do we know it is correct?
   - Model validation (goodness-of-fit).
   - Model selection.
2. **Estimation method**: which estimator do we pick?
   - The MLE is not always unbiased.
   - Minimum Variance Unbiased Estimator (MVUE) – among unbiased estimators, pick the one with the least variance.