- Consider a single observation $Y \sim \text{Binomial}(n, \theta)$. The pmf of Y is

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \text{for } y \in \{0, 1, \ldots, n\}.$$

The second derative of log likelihood is

$$\partial^2 \log f(y|\theta)/\partial^2\theta = -y/\theta^2 - (n-y)/(1-\theta)^2.$$

The Fisher information of Y, is

$$I_Y(\theta) = -E_\theta\left[-\frac{Y}{\theta^2} - \frac{n-Y}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Note that $I_Y(\theta)$ is the same as $I_{X_1,\cdots,X_n}(\theta)$.

Theorem 6.5 (consistency of MLE, TBp. 275)

Under appropriate smoothness conditions of f, the MLE from an i.i.d. sample is consistent.

Proof (sketch, for pdf case): Denote the true value of θ by θ_0. The MLE maximizes $\frac{l(\theta)}{n} = \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$.

The weak law of large numbers implies

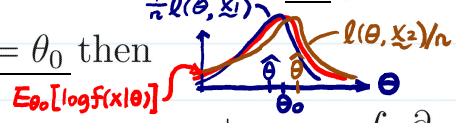
$$\forall\, \theta, \quad \frac{l(\theta)}{n} \xrightarrow{\mathcal{P}} \mathrm{E}_{\theta_0}[\log f(X|\theta)] = \int [\log f(x|\theta)]\, f(x|\theta_0)\, dx \quad \text{as } n \to \infty.$$

What is the difference between $\int [\log f(x|\theta)] f(x|\theta_0) dx$ and $\int [\log f(x|\theta)] f(x|\theta) dx$?

For large n, the θ value that maximizes $l(\theta)$ should be close to the θ value that maximizes $\mathrm{E}_{\theta_0}[\log f(X|\theta)]$. To maximize $\mathrm{E}_{\theta_0}[\log f(X|\theta)]$, consider

$$\frac{\partial}{\partial\theta}\mathrm{E}_{\theta_0}[\log f(X|\theta)] = \frac{\partial}{\partial\theta}\int [\log f(x|\theta)] f(x|\theta_0)\, dx = \int \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)} f(x|\theta_0)\, dx.$$

If $\theta = \theta_0$ then



When n is large, $\frac{1}{n}l(\theta, \underline{x}) \approx \mathrm{E}_{\theta_0}[\log f(X|\theta)]$

$$\frac{\partial}{\partial\theta}\mathrm{E}_{\theta_0}[\log f(X|\theta)]\Big|_{\theta=\theta_0} = \int \frac{\partial}{\partial\theta} f(x|\theta)\, dx\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\int f(x|\theta)\, dx\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta} 1 = 0.$$

Thus θ_0 is a stationary point and hopefully a maximizer.

Theorem 6.6 (asymptotic normality of MLE for one-dimensional parameter, TBp. 277)

Under some regularity conditions on f, the probability distribution of

$$\sqrt{nI_{X_1}(\theta_0)}(\hat\theta_{\text{MLE}} - \theta_0) = \frac{\sqrt{n}(\hat\theta_{\text{MLE}} - \theta_0)}{\sqrt{I_{X_1}(\theta_0)^{-1}}}$$

tends to a standard Normal distribution as n tends to infinity, where $\hat\theta_{\text{MLE}}$ is MLE and θ_0 is the true value of θ.

Proof (sketch): Denote $\hat\theta_{\text{MLE}}$ by $\hat\theta$. By Taylor expansion,

$$0 = l'(\hat\theta) \approx l'(\theta_0) + (\hat\theta - \theta_0)l''(\theta_0) + \cdots$$

$$\sqrt{n}(\hat\theta - \theta_0) \approx \frac{-l'(\theta_0)/\sqrt{n}}{l''(\theta_0)/n} \xrightarrow{d} N(0, I_{X_1}(\theta_0)) \to N\left(0, \frac{1}{I_{X_1}(\theta_0)}\right)$$

$$\xrightarrow{p} -I_{X_1}(\theta_0)$$

Consider the numerator,

$$\mathrm{E}_{\theta_0}\left[l'(\theta_0)\right] = \sum_{i=1}^{n} \mathrm{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right] = 0,$$

$$\mathrm{Var}_{\theta_0}\left[l'(\theta_0)\right] = \sum_{i=1}^{n} \mathrm{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right]^2 = nI_{X_1}(\theta_0).$$

And, since $l'(\theta_0) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(X_i|\theta_0)$, Central Limit Theorem implies that $l'(\theta_0)/\sqrt{nI_{X_1}(\theta_0)}$ converges in distribution to a standard Normal random variable. For the denominator,

$$\frac{l''(\theta_0)}{n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0) \xrightarrow{P} \mathrm{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0)\right] = -I_{X_1}(\theta_0).$$

Hence

$$\sqrt{nI_{X_1}(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N(0,1).$$

Also,

$$\mathrm{E}_{\theta_0}\left[\sqrt{n}(\hat{\theta} - \theta_0)\right] \approx 0, \quad \mathrm{Var}_{\theta_0}(\hat{\theta} - \theta_0) \approx \frac{1}{nI_{X_1}(\theta_0)}.$$

Thus, MLE ($\hat{\theta}$) is asymptotically unbiased and its asymptotic variance is

$$\left[nI_{X_1}(\theta_0)\right]^{-1} = \left[I_{X_1,\ldots,X_n}(\theta_0)\right]^{-1} = \left[-\mathrm{E}_{\theta_0}(l''(\theta_0))\right]^{-1}.$$

Notes (TBp. 277)

1. Theorem 6.6 (LNp.39) says the large sample distribution of an MLE is approximately normal with
   - mean θ_0 and ($\Rightarrow$ **asymptotically unbiased**)
   - variance $1/(nI_{X_1}(\theta_0))$. ($\Rightarrow 1/(nI_{X_1}(\theta_0))$ referred to as **asymptotic variance**)
2. Regularity conditions for Theorem 6.6 (LNp.39):
   (a) True value θ_0 is an interior point of the set of all parameter values (e.g., the theorem would not be expected to apply in Ex 6.16, LNp.27, if $\alpha_0 = 1$.)
   (b) Support of $f(x|\theta)$, i.e., the set of x's for which $f(x|\theta) > 0$, does not depend on θ (e.g., the theorem would not be expected to apply to estimating θ from a sample that are uniformly distributed on the interval $[0, \theta]$ (Ex 6.13, LNp.22).)

Theorem 6.7 (Fisher information under reparameterization)

Under the reparameterization $\tau(\theta)$, the (Fisher) information of $\tau(\theta)$ is

$$I_{X_1}(\tau(\theta)) \equiv \mathrm{E}\left[\frac{\partial}{\partial\tau(\theta)}\log f(X_1|\theta)\right]^2 = \frac{I_{X_1}(\theta)}{\tau'(\theta)^2}.$$

Proof:

$$I_{X_1}(\tau(\theta)) = E\left[\frac{\partial}{\partial \tau(\theta)} \log f(X_1|\theta)\right]^2 = E\left[\frac{\partial \theta}{\partial \tau(\theta)} \frac{\partial}{\partial \theta} \log f(X_1|\theta)\right]^2$$

Fisher information of n i.i.d. observations = $n \cdot I_{X_1}(\tau(\theta))$

Chain rule

$$= \left(\frac{1}{\tau'(\theta)}\right)^2 \cdot E\left[\frac{\partial}{\partial \theta} \log f(X_1|\theta)\right]^2 = \frac{I_{X_1}(\theta)}{\tau'(\theta)^2}$$

$\mathrm{Var}(\tau(\hat\theta)) \approx [\tau'(\theta)]^2 \times \mathrm{Var}(\hat\theta)$

Theorem 6.8 (asymptotic normality of MLE under reparameterization)

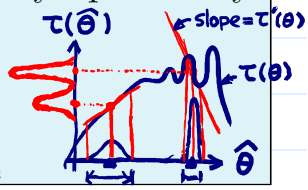
Under the reparameterization $\tau(\theta)$, the MLE of $\tau(\theta)$, $\tau(\hat\theta)$, is asymptotically Normal with mean $\tau(\theta)$ and variance

$$\frac{1}{nI_{X_1}(\tau(\theta))} = \frac{1}{n} \cdot \frac{\tau'(\theta)^2}{I_{X_1}(\theta)}.$$

$\hat\theta \xrightarrow{d} N\left(\theta, \frac{1}{nI_{X_1}(\theta)}\right)$

C.-R. bound (LNp.65)

(Ec) Hint. By δ method (item6, CH1~6, P.90)



Example 6.19 (information and asymptotic distribution of MLE for Poisson mean, TBp.282)

Let $X_1, \ldots, X_n$ be i.i.d $\sim P(\lambda)$. The MLE of λ is $\hat\lambda = \overline{X}$. The information of Poisson distribution is:

statistical modeling

LNp.19

$$I_{X_1}(\lambda) = \mathrm{E}\left[\frac{\partial}{\partial \lambda} \log f(X|\lambda)\right]^2 = \mathrm{E}\left[\frac{\partial}{\partial \lambda} \log\left(\frac{\lambda^X e^{-\lambda}}{X!}\right)\right]^2$$

$$= \mathrm{E}\left[\frac{\partial}{\partial \lambda}(X \log \lambda - \lambda - \log X!)\right]^2 = \mathrm{E}\left(\frac{X}{\lambda} - 1\right)^2 = \frac{1}{\lambda}.$$

$E\left(\frac{X}{\lambda} - 1\right) = \frac{\lambda}{\lambda} - 1 = 0$

$\mathrm{Var}\left(\frac{X}{\lambda} - 1\right) = \mathrm{Var}\left(\frac{X}{\lambda}\right) = \frac{1}{\lambda^2}\mathrm{Var}(X) = \frac{1}{\lambda}$

Smaller λ ⇒ larger $I(\lambda)$ ⇒ more information ⇒ MLE has smaller variance

Or,

$$I_{X_1}(\lambda) = -\mathrm{E}\left[\frac{\partial^2}{\partial \lambda^2} \log f(X|\lambda)\right] = -\mathrm{E}\left[\frac{-X}{\lambda^2}\right] = \frac{1}{\lambda}.$$

$I_{X_1,\ldots,X_n}(\lambda) = \frac{n}{\lambda}$

Hence, by theorem 6.6, the asymptotic distribution of $\overline{X}$ is Normal with mean λ and variance λ/n.

cf. → CLT for Poisson (Note 1, LN, CH1-8, LNp.97)

→ Normal approximation to Poisson (LN, CH1-8, p.92)

Exercise: Suppose that we are interested in the parameter $\tau = 1/\lambda$, 單位時間/次

- what is the (Fisher) information of τ? (Ans: τ^{-3}) ← by Thm 6.7 (LNp.41)
- what is the MLE of τ? (Ans: $\hat\tau_{\mathrm{MLE}} = 1/\overline{X}$)
- what is the asymptotic distribution of the MLE? ← use Thm 6.8 (LNp.42)
- what is its asymptotic variance?

FYI

Theorem 6.9 (multidimensional parameters, information and asymptotic normality of MLE, TBp.279)

Suppose $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)'$, a k-dimensional vector. Then, the asymptotic joint distribution of the MLE $\hat\Theta = (\hat\theta_1, \hat\theta_2, \ldots, \hat\theta_k)'$ is multivariate normal with mean vector Θ_0 and covariance matrix $\frac{1}{n} I^{-1}(\Theta_0)$ where $I(\Theta)$ is the $k \times k$ matrix with ij-th component

Fisher information matrix of n i.i.d. observations = $nI(\Theta)$

$$\mathrm{E}_\Theta\left[\frac{\partial}{\partial \theta_i} \log f(X_1|\Theta) \frac{\partial}{\partial \theta_j} \log f(X_1|\Theta)\right] = -\mathrm{E}_\Theta\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X_1|\Theta)\right].$$

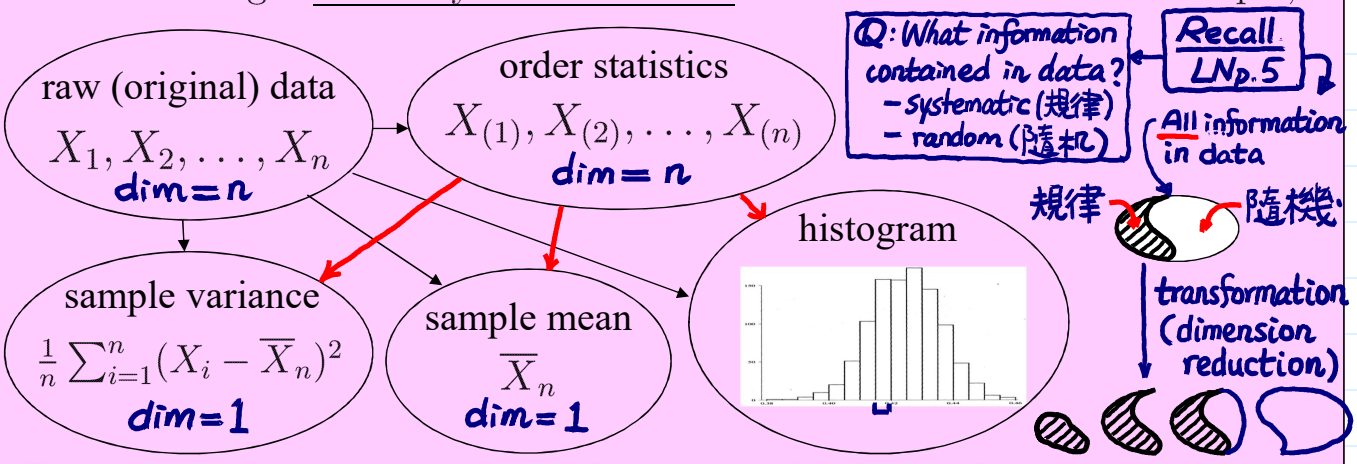
Here, $I(\Theta)$ is called the **(Fisher) information matrix** for Θ. ← one observation

❖ **Reading**: textbook, 8.5.2; **Further reading**: Hogg et al., 6.1, 6.2

• Data reduction --- the concepts of sufficiency, minimal sufficiency, and completeness

Question 6.1 (information and data reduction)

(numerical or graphical) transformations of data appear all the time in statistics for offering a summary of information contained in data. For example,



To present or extract concrete information, the data reduction through useful transformations is required. However, non-invertible transformations can cause the loss of information. (Example?) The lost information can be important or worthless to the objective of studying the data.

depend on statistical modeling → to understand the (unknown) systematic pattern

Q: How to examine whether the important information lost in transformation? Furthermore, what is *important information*?

Summary (formulation of information and data reduction problem, TBp. 305)

- Let $X_1, X_2, \ldots, X_n$ be a sample with joint pdf/pmf $f(\mathbf{x}|\Theta)$, where Θ is unknown parameter. ← statistical modeling ⇒ introduce Θ (unknown systematic pattern)
  - $X_1, X_2, \ldots, X_n$ contains two types of information:
    - information related to Θ ← "important" (useful) information
    - information irrelevant to Θ ← useless information
  - For example, toss a coin n times, i.e., $X_1, X_2, \ldots, X_n$ are i.i.d. from Bernoulli $B(\theta)$, ← What is important information?
    - $\overline{X}_n$ or $T = \sum_{i=1}^n X_i$ contains information about θ 3/25
    - When T is known, say $T = t$, the information that at which trials the t head's occur is irrelevant to θ ← place heads in t out of n positions; # of heads $\geq t$
    - n=5, consider the following possible results:
      - ▷ (0, 1, 1, 1, 1), $T = 4$; (1, 0, 1, 1, 1), $T = 4$; (1, 1, 0, 1, 1), $T = 4$; (1, 1, 1, 0, 1), $T = 4$; (1, 1, 1, 1, 0), $T = 4$
      - ▷ (1, 0, 0, 0, 0), $T = 1$; (0, 1, 0, 0, 0), $T = 1$; (0, 0, 1, 0, 0), $T = 1$; (0, 0, 0, 1, 0), $T = 1$; (0, 0, 0, 0, 1), $T = 1$

Fisher information — cf.

estimator of θ

useless information

$P(X_1, \ldots, X_5 | T=4) = 1/5$ ← irrelevant to θ

$P(X_1, \ldots, X_5 | T=1) = 1/5$ ← irrelevant to θ

distribution of $X_1, \ldots, X_5 | T$

T — $X_1, \ldots, X_n | T$ — all information

$T \sim \text{Binomial}(5, \theta)$ / distribution of T

cf.