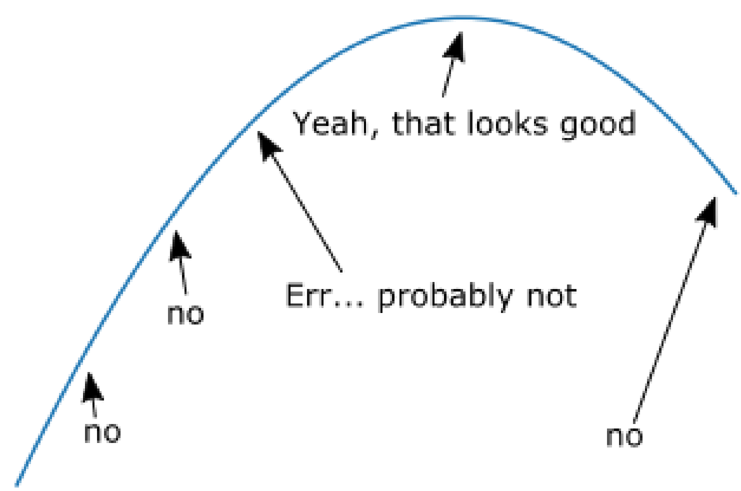
Slides 4: Maximum likelihood estimates

- Likelihood function
- Maximum likelihood
- Sufficient statistics
- Large sample properties of MLE

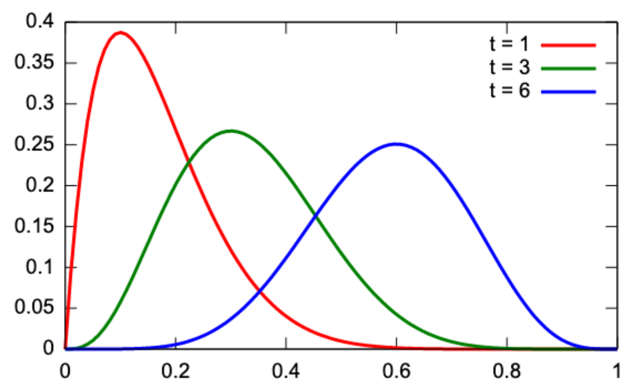


Consider a binomial model $T \sim \mathrm{Bin}(10, p)$. Suppose after observing $n = 10$ binary values we got $t = 1$ successful outcomes. The probability of the observed data

$$L(p) = \mathrm{P}(T = 1) = 10p(1-p)^9, \quad 0 \le p \le 1,$$

treated as a function of the unknown population parameters is called the likelihood function.

For three outcomes $t = 1, 3, 6$ we obtain three likelihood functions



Question. Clearly, the areas under each of the three likelihood curves on the figure are less than 1. Aren't they all supposed to be equal 1?

The parameter value that maximises the likelihood function is called a maximum likelihood estimate.

For the binomial model $T \sim \mathrm{Bin}(n, p)$ if the observed value is $T = t$, then

$$L(p) = \binom{n}{t} p^t (1-p)^{n-t}$$

and to maximise $L(p)$ is equivalent to maximise the log-likelihood

$$\log L(p) = \mathrm{const} + t \log(p) + (n-t) \log(1-p)$$

Take the derivative and put it equal to zero

$$\frac{t}{p} - \frac{n-t}{1-p} = 0$$

The solution gives $\hat{p} = \frac{t}{n}$. We conclude that the sample proportion is the MLE of the population proportion p.

Question. Does the figure above confirm this conclusion for $n = 10$ and $t = 1, 3, 6$?

Let us turn to the normal distribution $\mathrm{N}(\mu, \sigma)$ model. For a given sample $(x_1, \ldots, x_n)$ generated from $\mathrm{N}(\mu, \sigma)$, the likelihood function is

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{t_2 - 2\mu t_1 + n\mu^2}{2\sigma^2}}$$

fully determined by a pair of summary statistics

$$t_1 = \sum_{i=1}^{n} x_i, \qquad t_2 = \sum_{i=1}^{n} x_i^2$$

We can speak of a two-dimensional sufficient statistic (t_1, t_2), since it is sufficient to know this pair of numbers (t_1, t_2) to write down the likelihood. The MLEs for (μ, σ) will be the following functions of (t_1, t_2)

$$\hat{\mu} = \tfrac{t_1}{n}, \quad \hat{\sigma} = \sqrt{\tfrac{t_2}{n} - (\tfrac{t_1}{n})^2}$$

Question. What is the relation between $(\hat{\mu}, \hat{\sigma})$ and $(\bar{x}, s)$?

For a random sample $(x_1, \ldots, x_n)$ from $\mathrm{Gam}(\alpha, \lambda)$,

$$L(\alpha, \lambda) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x_i^{\alpha-1} e^{-\lambda x_i}$$
$$= \frac{\lambda^{n\alpha}}{\Gamma^n(\alpha)} (x_1 \cdots x_n)^{\alpha-1} e^{-\lambda(x_1+\ldots+x_n)} = \frac{\lambda^{n\alpha}}{\Gamma^n(\alpha)} t_2^{\alpha-1} e^{-\lambda t_1},$$

with a pair of sufficient statistics

$$t_1 = x_1 + \ldots + x_n, \quad t_2 = x_1 \cdots x_n.$$

To find the MLE of (α, λ), take two partial derivatives of

$$\log L(\alpha, \lambda) = n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \log t_2 - \lambda t_1$$

set the derivatives equal to zero and numerically the system of equations

$$0 = n \ln(\lambda) - n \tfrac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \ln t_2,$$
$$0 = \tfrac{n\alpha}{\lambda} - t_1.$$

Question. Why the likelihood function is the product of n densities?

More generally, for a random sample $(x_1, \ldots, x_n)$ taken from a population distribution $f(x|\theta)$, the likelihood function is given by the product

$$L(\theta) = f(x_1|\theta) \cdots f(x_n|\theta).$$

This implies that the log-likelihood function can be treated as a sum of independent and identically distributed random variables $\log f(X_i|\theta)$.

Using the CLT argument one can derive a normal approximation for the maximum likelihood estimator $\hat{\theta}$

$$\hat{\Theta} \approx \mathrm{N}(\theta, \frac{1}{\sqrt{n\mathbb{I}(\theta)}}), \text{ as } n \gg 1$$

$\mathbb{I}(\theta)$ is the Fisher information in a single observation, see below.

$$\boxed{\text{Approximate 95\% confidence interval } I_\theta \approx \hat{\theta} \pm 1.96 \cdot \frac{1}{\sqrt{n\mathbb{I}(\hat{\theta})}}}$$

Question. Can you see now that the MLEs are asymptotically unbiased and consistent?

The larger is the value of

$$g(x,\theta) = -\frac{\partial^2}{\partial\theta^2} \ln f(x|\theta)$$

at the top of the log-likelihood curve, the more information on the parameter θ is contained at the single observation x.

The Fisher information in a single observation is the expected value

$$\mathbb{I}(\theta) = \mathrm{E}[g(X,\theta)] = \int g(x,\theta) f(x|\theta) dx.$$

Then $n\mathbb{I}(\theta)$ is the Fisher information in n observations.

MLE is asymptotically efficient (have minimal variance) in the sense of Cramer-Rao inequality:

$$\boxed{\text{If } \theta^* \text{ is an unbiased estimator of } \theta\text{, then } \mathrm{Var}(\Theta^*) \ge \frac{1}{n\mathbb{I}(\theta)}.}$$

Question. Can a biased estimate θ^* have a smaller mean square error $\mathrm{E}[(\Theta^* - \theta)^2]$ than an unbiased estimate?

We illustrate by example. Data: lifetimes of five batteries in hours

$$x_1 = 0.5, \quad x_2 = 14.6, \quad x_3 = 5.0, \quad x_4 = 7.2, \quad x_5 = 1.2.$$

We propose an exponential model $X \sim \text{Exp}(\lambda)$. The likelihood function

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(x_1+\ldots+x_n)} = \lambda^5 e^{-\lambda \cdot 28.5}$$

first grows from 0 to $2.2 \cdot 10^{-7}$ and then falls down towards zero. The maximum is reached at $\hat{\lambda} = 0.175$.

Fisher information for the exponential model is easy to compute:

$$g(x, \lambda) = -\tfrac{\partial^2}{\partial \lambda^2} \ln f(x|\lambda) = \tfrac{1}{\lambda^2}, \qquad \mathbb{I}(\lambda) = \mathrm{E}[g(X, \lambda)] = \tfrac{1}{\lambda^2}.$$

This yields a standard error $s_{\hat{\lambda}} \approx \sqrt{\frac{\hat{\lambda}^2}{n}} = \frac{\hat{\lambda}}{\sqrt{n}}$ and a confidence interval

$$I_\lambda \approx 0.175 \pm 1.96 \cdot \tfrac{0.175}{\sqrt{5}} = 0.175 \pm 0.153.$$

Question. Is $\hat{\lambda}$ a biased estimate of λ?