STAT 24400 Lecture 14
Section 8.3 Parameter Estimation
Section 8.4 The Method of Moments
Section 8.5 The Method of Maximum Likelihood

Yibi Huang
Department of Statistics
University of Chicago

Section 8.3 Parameter Estimation

Parameter Estimation

Suppose that we observe data $X_1, X_2, \ldots, X_n$ generated from a known distribution with unknown parameter(s), e.g., the data is from

- $N(\mu, \sigma^2)$, with μ unknown (& σ^2 known)
- $N(\mu, \sigma^2)$, with μ & σ^2 unknown
- Exponential(λ), with λ unknown
- Binomial(n, p), with n known and p unknown

How can we estimate the unknown parameter(s)?
How can we perform inference on the unknown parameter(s)?

General Notation

- $X_1, \ldots, X_n$ = data drawn i.i.d. from the distribution
- θ = the unknown parameter(s)
- θ lies in Θ = subspace of $\mathbb{R}$ (or $\mathbb{R}^2$ if two parameters, etc)

General Notation

- $X_1, \ldots, X_n$ = data drawn i.i.d. from the distribution
- θ = the unknown parameter(s)
- θ lies in Θ = subspace of $\mathbb{R}$ (or $\mathbb{R}^2$ if two parameters, etc)
- We will write $f(x \mid \theta)$ for the PDF or PMF of the distribution, e.g.,
  - Exponential(λ) $\rightsquigarrow$ PDF $f(x \mid \lambda) = \lambda e^{-\lambda x}$
  - Poisson(λ) $\rightsquigarrow$ PMF $f(x \mid \lambda) = \dfrac{\lambda^x e^{-\lambda}}{x!}$

Parameter Estimation (Point Estimate)

Given data $X_1, \ldots, X_n$ i.i.d. $\sim f(x \mid \theta)$, would like to estimate the unknown θ

The *point estimate* or *estimator* of a parameter θ, is a function

$$\widehat{\theta} = g(X_1, \ldots, X_n)$$

computed from the observed data $\{X_1, \ldots, X_n\}$ that is a sensible guess for the unknown θ.

Note: any estimator $\widehat{\theta}$ must be a function of $X_1, \ldots, X_n$ only
it cannot involve any unknown parameter, e.g.,

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}$$

is not a estimator since it involves the unknown μ.

Examples of Point Estimates

Example 1: If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, the point estimate for the population mean μ can be

- the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- the median of $X_1, \ldots, X_n$
- the average of $X_1, \ldots, X_n$ after excluding the minimum & maximum

The point estimate for the population variance σ^2 can be

- the sample variance $S^2 = \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$
- an alternative estimator would result from using divisor n instead of $n-1$

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$$

Examples of Point Estimates

Example 2: If $X \sim Bin(n, p)$ is Binomial, the point estimate for the success probability p can be

- the sample proportion $\widehat{p} = \dfrac{X}{n}$
- Wilson's plus-four estimate $\tilde{p} = \dfrac{X+2}{n+4}$
  - adding successes and two failures to the sample and then calculate the sample proportion of successes

Mean Squared Error

With many possible point estimates $\widehat{\theta}$'s for a parameter θ, how to choose a good one among them?

A population criterion is to compare their *Mean Squared Error (MSE)*, defined as

$$\text{Mean Squared Error (MSE)} = \mathrm{E}[(\widehat{\theta} - \theta)^2]$$

MSE = $(\text{Bias})^2 +$ Variance

Recall the shortcut formula for the variance of any variable Y

$$\text{Var}(Y) = \text{E}(Y^2) - (\text{E}(Y))^2,$$

Rearranging the terms, we get

$$\text{E}(Y^2) = (\text{E}(Y))^2 + \text{Var}(Y).$$

Plugging in $Y = \widehat{\theta} - \theta$, then $\text{E}(\widehat{\theta} - \theta) = \text{E}(\widehat{\theta}) - \theta$, we get

$$\begin{array}{ccccc} \text{E}[(\widehat{\theta} - \theta)^2] & = & [\text{E}(\widehat{\theta}) - \theta]^2 & + & \text{Var}(\widehat{\theta} - \theta) \\ \| & & \| & & \| \\ \text{MSE} & = & (\text{Bias})^2 & + & \text{Var}(\widehat{\theta}) \end{array}$$

where the *bias* of an point estimate $\widehat{\theta}$ for θ is defined to be the difference between the expected value of the estimate and the true value of the parameter

$$\text{Bias} = \text{E}(\widehat{\theta}) - \theta$$

Unbiased Estimators

A point estimator $\widehat{\theta}$ is said to be an *unbiased estimator* of θ if

$$\mathsf{E}(\widehat{\theta}) = \theta$$

for every possible value of θ.

Unbiased Estimators

A point estimator $\widehat{\theta}$ is said to be an *unbiased estimator* of θ if

$$\mathsf{E}(\widehat{\theta}) = \theta$$

for every possible value of θ.

For unbiased estimators, $\boxed{\text{MSE} = \text{Variance}}$.

Examples of MSE

If $X_1, \ldots, X_n$ are i.i.d. with population mean μ and population variance σ^2, using the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ the point estimate for the population mean μ

- the bias is $\mathrm{E}(\overline{X}) - \mu = \mu - \mu = 0$
- the variance is $\mathrm{Var}(\overline{X}) = \sigma^2/n$

The MSE for $\overline{X}$ is hence

$$\mathrm{MSE} = (\mathrm{Bias})^2 + \mathrm{Variance} = 0^2 + \frac{\sigma^2}{n} = \frac{\sigma^2}{n}$$

MSE of Sample Variance S^2

In L13, we have shown that if $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$, then S^2 is an unbiased estimate for σ^2.

$$\mathrm{E}[S^2] = \sigma^2$$

To obtain the MSE, we need to calculate $\mathrm{Var}(S^2)$. From that

$$T = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

and the variance for $T \sim \chi^2_{n-1}$ is $2(n-1)$, it follows that

$$\mathrm{Var}(S^2) = \mathrm{Var}\left(\frac{\sigma^2 T}{n-1}\right) = \left(\frac{\sigma^2}{n-1}\right)^2 \underbrace{\mathrm{Var}(T)}_{=2(n-1)} = \frac{2\sigma^4}{n-1}.$$

The MSE of S^2 is hence

$$\mathrm{MSE} = (\mathrm{Bias})^2 + \mathrm{Variance} = 0^2 + \frac{2\sigma^4}{n-1} = \frac{2\sigma^4}{n-1}$$

A Biased Estimator for σ^2 w/ a Smaller MSE

Consider an alternative estimator for σ^2 that using divisor $n+1$ instead of $n-1$

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n+1} = \frac{(n-1)S^2}{n+1}$$

The expected value and variance of $\widehat{\sigma}^2$ are respectively

$$\mathrm{E}(\widehat{\sigma}^2) = \frac{(n-1)\,\mathrm{E}(S^2)}{n+1} = \frac{(n-1)\sigma^2}{n+1},$$

$$\mathrm{Var}(\widehat{\sigma}^2) = \left(\frac{n-1}{n+1}\right)^2 \mathrm{Var}(S^2) = \left(\frac{n-1}{n+1}\right)^2 \frac{2\sigma^4}{(n-1)} = \frac{2(n-1)\sigma^4}{(n+1)^2}$$

Hence, $\widehat{\sigma}^2$ is a **biased** estimator for σ^2 with

$$\mathrm{Bias} = \mathrm{E}(\widehat{\sigma}^2) - \sigma^2 = \frac{(n-1)\sigma^2}{n+1} - \sigma^2 = \frac{-2\sigma^2}{n+1}.$$

The MSE of $\widehat{\sigma}^2$ is

$$\begin{aligned}\text{MSE} &= (\text{Bias})^2 + \text{Variance} \\ &= \left(\frac{-2\sigma^2}{n+1}\right)^2 + \frac{2(n-1)\sigma^4}{(n+1)^2} = \frac{2n\sigma^4}{(n+1)^2}\end{aligned}$$

which is lower than the MSE of $\dfrac{2\sigma^4}{n-1}$ for the sample variance S^2.

The MSE of $\widehat{\sigma}^2$ is

$$\begin{aligned}\text{MSE} &= (\text{Bias})^2 + \text{Variance} \\ &= \left(\frac{-2\sigma^2}{n+1}\right)^2 + \frac{2(n-1)\sigma^4}{(n+1)^2} = \frac{2n\sigma^4}{(n+1)^2}\end{aligned}$$

which is lower than the MSE of $\dfrac{2\sigma^4}{n-1}$ for the sample variance S^2.

A biased estimator might have a smaller MSE if it has a smaller variance.

MSE of the Sample Proportion $\hat{p} = \dfrac{X}{n}$

If $X \sim Bin(n, p)$ is Binomial, a point estimate for the success probability p is the sample proportion $\hat{p} = \dfrac{X}{n}$. As X is Binomial,

$$
\begin{aligned}
\mathrm{E}(X) &= np & \Rightarrow \quad \mathrm{E}(\hat{p}) &= \frac{\mathrm{E}(X)}{n} = \frac{np}{n} = p \\
\mathrm{Var}(X) &= np(1-p) & \Rightarrow \quad \mathrm{Var}(\hat{p}) &= \frac{\mathrm{Var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}
\end{aligned}
$$

Thus the sample proportion $\hat{p}$ is **unbiased** with the MSE

$$
\mathrm{MSE} = (\mathrm{Bias})^2 + \mathrm{Variance} = 0^2 + \frac{p(1-p)}{n} = \frac{p(1-p)}{n}.
$$

MSE for Wilson's "Plus-Four" Estimate for Proportions

Recall Wilson's plus-four estimate is

$$\tilde{p} = \frac{X+2}{n+4}.$$

It's expected value and variance are respectively,

$$E(\tilde{p}) = \frac{E(X)+2}{n+4} = \frac{np+2}{n+4}, \text{ and } Var(\tilde{p}) = \frac{Var(X)}{(n+4)^2} = \frac{np(1-p)}{(n+4)^2}.$$

Its bias and MSE are respectively

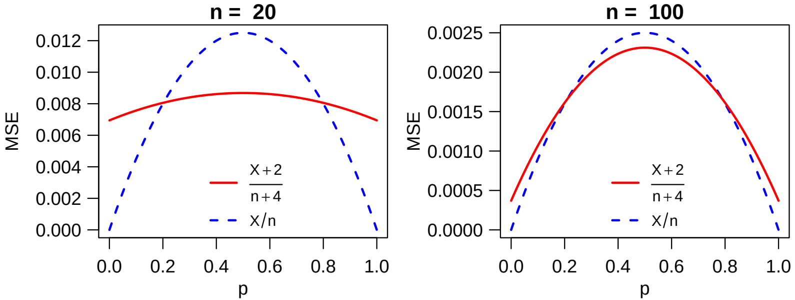
$$\text{Bias} = E(\tilde{p}) - p = \frac{np+2}{n+4} - p = \frac{2-4p}{n+4}$$

$$\text{MSE} = (\text{Bias})^2 + \text{Variance} = \left(\frac{2-4p}{n+4}\right)^2 + \frac{np(1-p)}{(n+4)^2}$$

MSE's for Sample Proportion & Wilson's "Plus-Four"

Below are the graphs of the MSE for $\widehat{p} = X/n$ and $\tilde{p} = \dfrac{X+2}{n+4}$

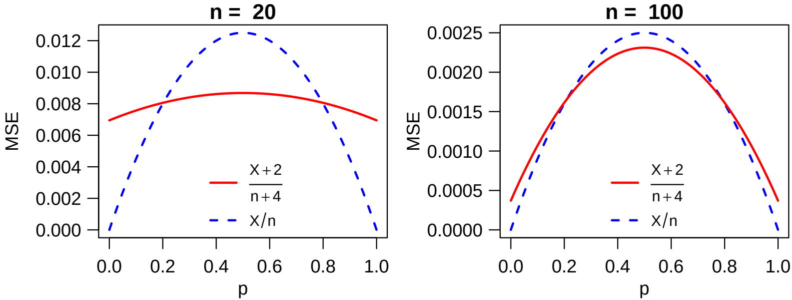
$$\mathrm{MSE}(\widehat{p}) = \frac{p(1-p)}{n}, \quad \mathrm{MSE}(\tilde{p}) = \left(\frac{2-4p}{n+4}\right)^2 + \frac{np(1-p)}{(n+4)^2}$$



MSE's for Sample Proportion & Wilson's "Plus-Four"

Below are the graphs of the MSE for $\widehat{p} = X/n$ and $\tilde{p} = \dfrac{X+2}{n+4}$

$$\text{MSE}(\widehat{p}) = \frac{p(1-p)}{n}, \quad \text{MSE}(\tilde{p}) = \left(\frac{2-4p}{n+4}\right)^2 + \frac{np(1-p)}{(n+4)^2}$$



- $\widehat{p} = X/n$ has a smaller MSE only when p is close to 0 or 1
- $\tilde{p} = \frac{X+2}{n+4}$ has a smaller MSE when p is NOT close to 0 or 1
- The two MSE's are close when n is large

Sampling Distributions

The *sampling distribution* of a point estimate $\widehat{\theta}$ is simply its probability distribution, e.g.,

Ex1. If $X_1, \ldots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$, the sampling distribution for $\widehat{\mu} = \overline{X}$ is

$$\widehat{\mu} = \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and the sampling distribution for S^2 is that

$$S^2 = \frac{\sigma^2 T}{n-1}, \quad \text{where } T \sim \chi^2_{n-1}.$$

Note: The sampling distribution generally depends on some unknown parameter θ.

Ex2: If $X_1, \ldots, X_n$ are i.i.d. from some distribution with mean μ and variance σ^2 (not necessarily normal),

- the exact sampling distribution would depend on the distribution of X_i
- CLT asserts that

$$\widehat{\mu} = \overline{X} \text{ is approx. } \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Standard Error

The *standard error (SE)* of a point estimate $\widehat{\theta}$ refers to any estimate of the standard deviation of $\widehat{\theta}$.

Ex1. If $X_1, \ldots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$,

- the standard deviation for $\widehat{\mu} = \overline{X}$ is

$$\mathrm{SD}(\overline{X}) = \sqrt{\mathrm{Var}(\overline{X})} = \sqrt{\frac{\sigma^2}{n}}$$

  which involves the unknown σ^2
- the standard error for $\overline{X}$ is

$$\mathrm{SE}(\overline{X}) = \sqrt{\frac{S^2}{n}}$$

  which replaces the unknown σ^2 by its estimate S^2.

Ex2. If $X \sim \text{Bin}(n, p)$,

- the standard deviation for $\widehat{p} = X/n$ is

$$\text{SD}(\widehat{p}) = \sqrt{\text{Var}(\widehat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

  which involves the unknown p

- the standard error for $\widehat{p}$ is

$$\text{SE}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

  which replaces the unknown p by its estimate $\widehat{p}$.

Note: The true SD may depend on θ, while SE depends on the data but not on θ

Section 8.4 The Method of Moments

Sample Moments

Recall the kth moment of a random variable X is $\mathrm{E}[X^k]$.

If $X_1, \ldots, X_n$ are i.i.d. from some distribution $f(x \mid \theta)$, the *kth sample moment* is defined to be

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

The Method of Moments (MME)

The *method of moments* is a strategy for finding an estimator $\widehat{\theta}$.

If there is only one parameter θ,

1. Compute $\mathrm{E}(X)$ as a function of θ
2. Compute the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
3. Choose $\widehat{\theta}$ as the value of θ so that $\mathrm{E}(X) = \overline{X}$

If there are k parameters $\theta_1, \ldots, \theta_k$

1. Compute $\mathrm{E}(X)$, $\mathrm{E}(X^2)$, ..., $\mathrm{E}(X^k)$ as functions of θ_i's
2. Compute the sample moments

$$\widehat{\mu}_1 = \frac{1}{n}\sum\nolimits_{i=1}^{n} X_i, \quad \widehat{\mu}_2 = \sum\nolimits_{i=1}^{n} X_i^2, \ldots, \widehat{\mu}_k = \frac{1}{n}\sum\nolimits_{i=1}^{n} X_i^k.$$

3. Choose $(\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ as the value of θ_i so that

$$\mathrm{E}(X^j) = \frac{1}{n}\sum\nolimits_{i=1}^{n} X_i^j \quad \text{for } 1 \le j \le k.$$

(solving a system of k equations, for k unknowns)

Examples (1 Parameter)

Ex1: $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Poisson(λ) for unknown $\lambda > 0$

- PMF: $f(x \mid \lambda) = e^{-\lambda}\lambda^x/x!$, $x = 0, 1, 2, \ldots$
- $\text{E}(X) = \lambda$
- The method of moment estimate (MME) for λ is $\widehat{\lambda} = \overline{X}$

Ex2: $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Geometric(p) for unknown p

- PMF: $f(x \mid p) = (1 - p)^{x-1}p$, $x = 1, 2, 3, \ldots$
- $\text{E}(X) = 1/p$
- MME for p is $\widehat{p} = 1/\overline{X}$

Ex3: $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Exponential(λ) for unknown $\lambda > 0$

- PDF: $f(x \mid \lambda) = \lambda e^{-\lambda x}$, $x > 0$
- $\text{E}(X) = 1/\lambda$
- MME for λ is $\widehat{\lambda} = 1/\overline{X}$.

Example 4 — Uniform$[0, \theta]$

$X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Uniform$[0, \theta]$ for unknown $\theta > 0$

- PDF: $f(x \mid \theta) = \frac{1}{\theta}$, $0 \leq x \leq \theta$
- $\mathrm{E}(X) = \theta/2$
- MME for θ is $\widehat{\theta} = 2\overline{X}$.

Example 5 — MME for Gamma

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda)$ for unknown $\alpha, \lambda > 0$

- PDF: $f(x \mid \alpha, \lambda) = \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad x > 0$
- $\text{E}(X) = \alpha/\lambda$
- $\text{Var}(X) = \alpha/\lambda^2 \;\Rightarrow\; \text{E}[X^2] = \text{Var}(X) + (\text{E}(X))^2 = \dfrac{\alpha(\alpha+1)}{\lambda^2}$
- The MMEs for α and λ must satisfy

$$\overline{X} = \frac{\widehat{\alpha}}{\widehat{\lambda}}, \quad \widehat{\mu}_2 = \frac{\widehat{\alpha}(\widehat{\alpha}+1)}{\widehat{\lambda}^2} \qquad \left(\text{Recall } \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2\right)$$

From the second equation

$$\widehat{\mu}_2 = \overline{X}^2 + \frac{\overline{X}}{\lambda} \quad \Rightarrow \quad \widehat{\lambda} = \frac{\overline{X}}{\widehat{\mu}_2 - \overline{X}^2}$$

and from the first equation,

$$\widehat{\alpha} = \widehat{\lambda}\overline{X} = \frac{\overline{X}^2}{\widehat{\mu}_2 - \overline{X}^2}.$$

Section 8.5 Likelihood & Maximum Likelihood Estimation

A Probability Question

Let p be the proportion of US adults that are willing to get the latest flu shot.

A sample of 20 subjects are randomly selected. Let X be the number of them that are willing to get the latest flu shot. What is $P(X = 8)$?

Answer: X is Binomial $(n = 20,\ p)$ (Why?)

$$P(X = x \mid p) = \binom{20}{x} p^x (1-p)^{n-x}.$$

If p is known to be 0.3, then

$$\mathrm{P}(X = 8 \mid p) = \binom{20}{8} (0.3)^8 (0.7)^{12} \approx 0.1144.$$

A Statistics Question

Suppose 8 of 20 randomly selected U.S. adults said they are willing to get the latest flu shot.

What can we infer about the value of

$$p = \text{proportion of U.S. adults that are willing to get a flu shot?}$$

The chance to observe $X = 8$ in a random sample of size $n = 20$ is

$$\mathrm{P}(X = 8 \mid p) = \begin{cases} \binom{20}{8}(0.3)^8(0.7)^{12} \approx 0.1144 & \text{if } p = 0.3 \\ \binom{20}{8}(0.6)^8(0.4)^{12} \approx 0.0355 & \text{if } p = 0.6 \end{cases}$$

It appears that $p = 0.3$ is **more likely** to be true value p than $p = 0.6$, since the former gives a higher prob. to observe the outcome $X = 8$.

We say the *likelihood* of $p = 0.3$ is higher than that of $p = 0.6$.

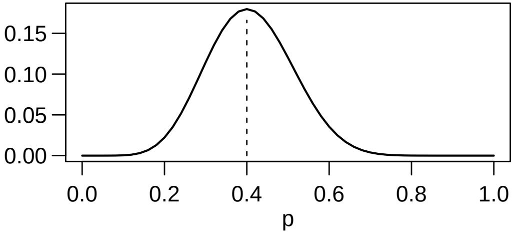
Maximum Likelihood Estimate (MLE)

The *maximum likelihood estimate* (*MLE*) of a parameter θ is the value at which the likelihood function is maximized.

Example. If 8 of 20 randomly selected U.S. adults are comfortable getting the flu shot, the likelihood function

$$L(p \mid x = 8) = \binom{20}{8} p^8 (1-p)^{12}$$

reaches its max at $p = 0.4$,
the MLE for p is $\widehat{p} = 0.4$ given the data $X = 8$.



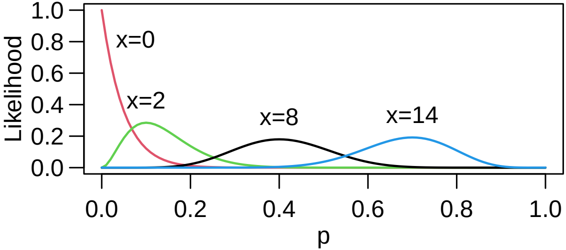
Maximum Likelihood Estimate (MLE)

The probability

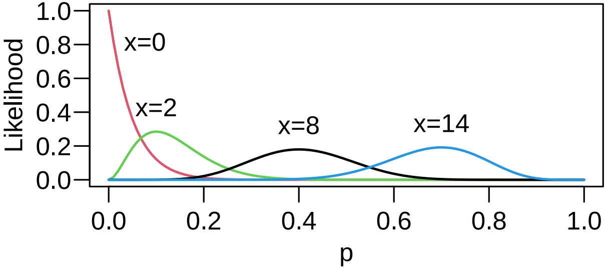
$$P(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x} = L(p \mid x)$$

viewed as a function of p, is called the *likelihood function*, (or just the **likelihood**) of p, denoted as $L(p \mid x)$.

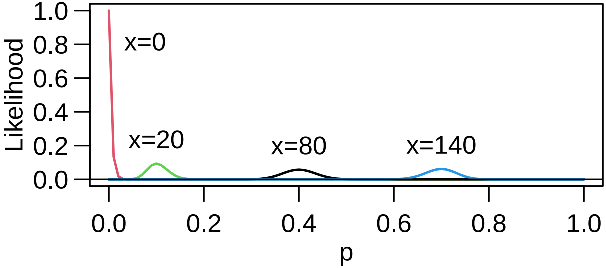
It measure the “plausibility” of a value being the true value of p.



Likelihood functions $L(p \mid x)$ at different values of x for $n = 20$.



Likelihood functions $L(p \mid x)$ for various values of x when $n = 20$.



Likelihood functions $L(p \mid x)$ at various values of x when $n = 200$.

Likelihood in General

In general, suppose the observed data $(X_1, X_2, \ldots, X_n)$ have a joint PDF or PMF with some parameter(s) called θ

$$f(x_1, x_2, \ldots, x_n \mid \theta)$$

The *likelihood function* for the parameter θ is

$$L(\theta) = L(\theta \mid X_1, X_2, \ldots, X_n) = f(X_1, X_2, \ldots, X_n \mid \theta).$$

- Note the likelihood function regards the probability as a function of the parameter θ rather than as a function of the data $X_1, X_2, \ldots, X_n$.
- If
  $$L(\theta_1 \mid x_1, \ldots, x_n) > L(\theta_2 \mid x_1, \ldots, x_n),$$
  then θ_1 appears more plausible to be the true value of θ than θ_2 does, given the observed data $x_1, \ldots, x_n$.

Maximizing the Log-likelihood

Rather than maximizing the likelihood, it is often computationally easier to maximize its natural logarithm, called the *log-likelihood*, denoted as

$$\ell(\theta) = \log L(\theta)$$

which results in the same answer since logarithm is strictly increasing,

$$x_1 > x_2 \quad \iff \quad \log(x_1) > \log(x_2).$$

So

$$L(\theta_1) > L(\theta_2) \quad \iff \quad \log L(\theta_1) > \log L(\theta_2).$$

Here, log() is always the natural log.

Notation:

- **upper case** $L(\theta) =$ **likelihood**
- **lower case** $\ell(\theta) = \log L(\theta) =$ **log-likelihood**

Example (MLE for Binomial)

If the observed data $X \sim$ Binomial (n, p) but p is unknown, the likelihood of p is

$$L(p \mid x) = p(X = x \mid p) = \binom{n}{x} p^x (1-p)^{n-x}$$

and the *log-likelihood* is

$$\ell(p) = \log L(p \mid x) = \log \binom{n}{x} + x \log(p) + (n - x) \log(1 - p).$$

From Calculus, we know a function $g(u)$ reaches its max at $u = u_0$ if

$$\frac{d}{\mathrm{d}u} g(u) = 0 \text{ at } u = u_0 \quad \text{and} \quad \frac{d^2}{du^2} g(u) < 0 \text{ at } u = u_0.$$

Example — MLE for Binomial

$$\frac{d}{dp}\ell(p \mid x) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x-np}{p(1-p).}$$

equals 0 when

$$\frac{x-np}{p(1-p)} = 0$$

That is, when $x - np = 0$.

Solving for p gives the ML estimator (MLE) $\boxed{\widehat{p} = \frac{x}{n}}$.

$$\text{and } \frac{d^2}{dp^2}\ell(p \mid x) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} < 0 \text{ for any } 0 < p < 1$$

Thus, we know $\ell(p \mid x)$ reaches its max when $p = x/n$.

So MLE of p is $\widehat{p} = \dfrac{X}{n}$ = sample proportion of successes.

Likelihood Based on i.i.d. Observations

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \theta)$ for an unknown parameter θ

The joint PDF or PMF of $(X_1, \ldots, X_n)$ is the product of the marginal PDF/PMF since they are i.i.d.

$$\prod_{i=1}^{n} f(x_i \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta) \times \cdots \times f(x_n \mid \theta)$$

The likelihood is then

$$L(\theta) = L(\theta \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i \mid \theta).$$

The log likelihood then has the summation form

$$\ell(\theta) = \log L(\theta \mid X_1, \ldots, X_n) = \log \left(\prod_{i=1}^{n} f(X_i \mid \theta) \right) = \sum_{i=1}^{n} \log \big(f(X_i \mid \theta) \big).$$

Example — MLE for Exponential

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$ for unknown $\lambda > 0$

- PDF: $f(x \mid \lambda) = \lambda e^{-\lambda x}$
- likelihood: $L(\lambda) = \prod_{i=1}^{n} f(X_i \mid \lambda) = \lambda^n \exp(\lambda \sum_{i=1}^{n} X_i)$
- log likelihood:

$$\ell(\lambda) = \log L(\lambda) = n\log(\lambda) - \lambda \sum_{i=1}^{n} X_i = n\log(\lambda) - n\lambda\overline{X}$$

- Solve for MLE:

$$0 = \frac{d}{d\lambda}\ell(\lambda) = \frac{n}{\lambda} - n\overline{X} \quad \Rightarrow \quad \widehat{\lambda} = \frac{1}{\overline{X}} \quad \text{(same as MME)}$$

The likelihood indeed reaches its max at $\lambda = 1/\overline{X}$ since

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{n}{\lambda^2} < 0.$$

Example — MLE for Poisson

$X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$ for unknown $\lambda > 0$

- PMF: $f(x \mid \lambda) = e^{-\lambda}\lambda^x / x!$
- likelihood: $L(\lambda) = \prod_{i=1}^n f(X_i \mid \lambda) = e^{-n\lambda}\lambda^{\sum_{i=1}^n X_i} \Big/ \prod_{i=1}^n X_i!$
- log likelihood:

$$\ell(\lambda) = \log L(\lambda) = -n\lambda + \sum_{i=1}^n X_i \log(\lambda) - \sum_{i=1}^n \log(X_i!)$$

- Solve for MLE:

$$0 = \frac{d}{d\lambda}\ell(\lambda) = -n + \frac{\sum_{i=1}^n X_i}{\lambda} = -n + \frac{n\overline{X}}{\lambda}$$

$$\Rightarrow \quad \widehat{\lambda} = \overline{X} \quad \text{(same as MME)}$$

The likelihood indeed reaches its max at $\lambda = \overline{X}$ since

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{n\overline{X}}{\lambda^2} \leq 0.$$

Example — Negative Binomial

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{NegBin}(r, p)$, r is known, but p is unknown

The PMF is $f(x \mid p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$.

$$\text{likelihood } L(p) = \prod_{i=1}^{n} f(X_i \mid p) = \left[\prod_{i=1}^{n} \binom{X_i - 1}{r - 1}\right] p^{nr} (1-p)^{(\sum_{i=1}^{n} X_i) - nr}$$

$$= \left[\prod_{i=1}^{n} \binom{X_i - 1}{r - 1}\right] p^{nr} (1-p)^{n\overline{X} - nr}$$

The log likelihood is

$$\ell(p) = \sum_{i=1}^{n} \log \binom{X_i - 1}{r - 1} + nr \log(p) + n(\overline{X} - r) \log(1-p)$$

Solve for MLE:

$$0 = \frac{d}{dp} \ell(p) = \frac{nr}{p} - \frac{n(\overline{X} - r)}{1-p} = \frac{n(r - p\overline{X})}{p(1-p)} \quad \Rightarrow \quad \widehat{p} = \frac{r}{\overline{X}}.$$

To see if log likelihood indeed reaches its max at $p = r/\overline{X}$, we check

$$\frac{d^2}{dp^2}\ell(p) = -\frac{nr}{p^2} - \frac{n(\overline{X} - r)}{(1-p)^2}$$

As $X_i \geq r$ and hence $\overline{X} \geq r$, the second derivative above is indeed ≤ 0.

This shows $\widehat{p} = \dfrac{r}{\overline{X}}$ is indeed the MLE.

MLE for Two Parameters

From Calculus, we know a function $g(u, v)$ reaches its maximum at $(u, v) = (u_0, v_0)$ if the following 3 conditions are met

1. $\frac{\partial}{\partial u} g(u, v) = \frac{\partial}{\partial v} g(u, v) = 0$ at $(u, v) = (u_0, v_0)$;
2. $\frac{\partial^2}{\partial u^2} g(u, v) < 0$ at $(u, v) = (u_0, v_0)$;
3. the Hessian matrix

$$\begin{vmatrix} \frac{\partial^2}{\partial u^2} g(u, v) & \frac{\partial^2}{\partial uv} g(u, v) \\ \frac{\partial^2}{\partial vu} g(u, v) & \frac{\partial^2}{\partial v^2} g(u, v) \end{vmatrix}$$

has a *positive* determinant at $(u, v) = (u_0, v_0)$.

Example — MLE for Normal

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ for unknown μ, σ^2

- PDF: $f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
- likelihood:
$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(X_i \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2\right)$$
- log likelihood:
$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$
- Solve for MLE:
$$\begin{cases} 0 = \frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = \frac{n}{\sigma^2}(\overline{X} - \mu) \\ 0 = \frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (X_i - \mu)^2 \end{cases}$$

$$\begin{cases} 0 = \frac{\partial}{\partial \mu}\ell(\mu, \sigma^2) = \frac{n}{\sigma^2}(\overline{X} - \mu) \\ 0 = \frac{\partial}{\partial \sigma^2}\ell(\mu, \sigma^2) = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \mu)^2 \end{cases}$$

The first equation immediately gives $\widehat{\mu} = \overline{X}$.
Plugging $\mu = \overline{X}$ into the second equation, we get

$$0 = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \Rightarrow \quad \widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}.$$

Note the MLE for σ^2 is not $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$.

To check the log likelihood indeed reach its max when $\mu = \overline{X}$ and $\widehat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}$, we calculate the second derivative of the log likelihood:

$$\frac{\partial^2}{\partial \mu^2}\ell(\mu, \sigma^2) = -\frac{n}{\sigma^2} < 0$$

$$\frac{\partial^2}{\partial \sigma^2 \mu}\ell(\mu, \sigma^2) = -\frac{n}{\sigma^4}(\overline{X} - \mu)$$

$$\frac{\partial^2}{\partial (\sigma^2)^2}\ell(\mu, \sigma^2) = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^n (X_i - \mu)^2$$

When $\mu = \overline{X}$ and $\sigma^2 = \widehat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}$, the Hessian matrix is

$$\begin{vmatrix} -\frac{n}{\widehat{\sigma^2}} & 0 \\ 0 & -\frac{n}{2(\widehat{\sigma^2})^2} \end{vmatrix}$$

which has a positive determinant. This shows the MLE for μ and σ^2 are

$$\mu = \overline{X} \quad \text{and} \quad \widehat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n}.$$

Example — MLE for Gamma

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \lambda)$ for unknown α, λ

- PDF: $f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x > 0$
- likelihood:
$$L(\alpha, \lambda) = \prod_{i=1}^{n} f(X_i \mid \alpha, \lambda) = \frac{\lambda^{n\alpha}}{(\Gamma(\alpha))^n} \left(\prod_{i=1}^{n} X_i \right)^{\alpha-1} e^{-\lambda \sum_{i=1}^n X_i}$$
- log likelihood:
$$\ell(\alpha, \lambda) = n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log X_i - \lambda \sum_{i=1}^{n} X_i$$
- Solve for MLE:
$$0 = \frac{\partial}{\partial \alpha} \ell(\alpha, \lambda) = n \log \lambda - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log X_i$$
$$0 = \frac{\partial}{\partial \lambda} \ell(\alpha, \lambda) = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i = \frac{n\alpha}{\lambda} - n\overline{X}$$

The second equation gives

$$\widehat{\lambda} = \frac{\widehat{\alpha}}{\overline{X}},$$

plugging it into the first equation we get

$$n \log(\widehat{\alpha}) - n \log(\overline{X}) - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} + \sum_{i=1}^{n} \log X_i = 0$$

This equation cannot be solved in closed form.
Numerical tools are required to compute the value of the MLE.

Example — Uniform$[0, \theta]$

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}[0, \theta]$ for unknown $\theta > 0$

- PDF: $f(x \mid \theta) = \frac{1}{\theta}$, $0 \leq x \leq \theta$
- Joint PDF:

$$\prod_{i=1}^{n} f(X_i \mid \theta) = \begin{cases} \theta^{-n} & \text{if } 0 \leq X_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

  This means the joint PDF is non-zero only if $\theta \geq X_{(n)} = \max_{1 \leq i \leq n} X_i$.
- Likelihood: $L(\theta) = \theta^{-n}$
- Solve for MLE: Note the smaller the value of θ, the greater the likelihood, but θ cannot fall below $X_{(n)}$. Thus the MLE for θ is

$$\widehat{\theta} = X_{(n)} \quad \text{(Different from MME.)}$$

Comparison of MME and MLE for Uniform$[0, \theta]$

MME $\widehat{\theta}_{\text{MME}} = 2\overline{X}$:

- For each X_i, $\text{E}(X_i) = \frac{\theta}{2}$, $\text{Var}(X_i) = \frac{\theta^2}{12}$
- $\text{E}(\widehat{\theta}_{\text{MME}}) = 2\,\text{E}(\overline{X}) = 2 \cdot \frac{\theta}{2} = \theta$
- $\text{Bias}(\widehat{\theta}_{\text{MME}}) = \text{E}(\widehat{\theta}_{\text{MME}}) - \theta = \theta - \theta = 0$
- Variance:

$$\text{Var}(\widehat{\theta}_{\text{MME}}) = \text{Var}(2\overline{X}) = 2^2\,\text{Var}(\overline{X}) = 2^2\frac{\text{Var}(X_i)}{n} = \frac{\theta^2}{3n}$$

- $\text{MSE} = \text{bias}^2 + \text{Var}(\widehat{\theta}_{\text{MME}}) = \dfrac{\theta^2}{3n}$

Comparison of MME & MLE for Uniform$[0, \theta]$:

MLE $\widehat{\theta}_{\text{MLE}} = X_{(n)}$:

- Using tools in L07, one can obtain the PDF of $X_{(n)}$:

$$f(x) = \frac{nx^{n-1}}{\theta^n}, \quad 0 \le x \le \theta$$

- Bias:

$$\mathrm{E}(\widehat{\theta}_{\text{MLE}}) = \int_{x=0}^{\theta} x \cdot \frac{nx^{n-1}}{\theta^n} \, \mathrm{d}x = \frac{n\theta}{n+1} \;\Rightarrow\; \text{bias} = -\frac{\theta}{n+1}$$

- Variance:

$$\mathrm{E}((\widehat{\theta}_{\text{MLE}})^2) = \int_{x=0}^{\theta} x^2 \cdot \frac{nx^{n-1}}{\theta^n} \, \mathrm{d}x = \frac{n\theta^2}{n+2}$$

$$\Rightarrow \quad \mathrm{Var}(\widehat{\theta}_{\text{MLE}}) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)}$$

- $\mathrm{MSE}(\widehat{\theta}_{\text{MLE}}) = \text{bias}^2 + \mathrm{Var}(\widehat{\theta}_{\text{MLE}}) = \dfrac{2\theta^2}{(n+1)(n+2)}$
  - far smaller than $\mathrm{MSE}(\widehat{\theta}_{\text{MME}}) = \dfrac{\theta^2}{3n}$

Properties of MLE for Exponential

For $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$ with the PDF

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \quad x > 0,$$

The MLE (and MME) for λ is $\widehat{\lambda} = \dfrac{1}{\overline{X}}$.

Since $Y = n\overline{X} = \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda)$ has the PDF

$$f_Y(y) = \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y}, \quad y > 0,$$

we can find the expected value and variance for $\widehat{\lambda} = 1/\overline{X} = n/Y$ as follows,

$$\mathrm{E}[\widehat{\lambda}] = \mathrm{E}\left(\frac{n}{Y}\right) = \int_{y=0}^{\infty} \frac{n}{y} \cdot \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y = \frac{n\lambda}{n-1}$$

$$\mathrm{E}[\widehat{\lambda}^2] = \mathrm{E}\left(\frac{n^2}{Y^2}\right) = \int_{y=0}^{\infty} \frac{n^2}{y^2} \cdot \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y = \frac{n^2\lambda^2}{(n-1)(n-2)}$$

$$\mathrm{Var}(\widehat{\lambda}) = \mathrm{E}[\widehat{\lambda}^2] - (\mathrm{E}[\widehat{\lambda}])^2 = \frac{n^2\lambda^2}{(n-1)(n-2)} - \left(\frac{n\lambda}{n-1}\right)^2 = \frac{n^2\lambda^2}{(n-1)^2(n-2)}$$

The bias is

$$\text{Bias} = \text{E}[\widehat{\lambda}] - \lambda = \frac{n\lambda}{n-1} - \lambda = \frac{\lambda}{n-1}.$$

The MSE of $\widehat{\lambda}$ is

$$\begin{aligned}\text{MSE} &= \text{Bias}^2 + \text{Var}(\widehat{\lambda}) \\ &= \left(\frac{\lambda}{n-1}\right)^2 + \frac{n^2\lambda^2}{(n-1)^2(n-2)} = \frac{(n+2)\lambda^2}{(n-1)(n-2)}.\end{aligned}$$