Principles of Point Estimation

Dr.Mamatha.H.R
Professor
Department of Computer Science and Engineering
PES University
Bangalore

Course material created using various Internet resources and text book

Parameters

- Populations are described by their probability distributions and/or parameters.
  - For quantitative populations, the location and shape are described by μ and σ.
  - Binomial populations are determined by a single parameter, ***p***.
- If the values of parameters are unknown, we make inferences about them using **sample** information.

- if *X* $\sim$ *Bin(n, p), the* sample proportion *p^ = X/n is often used to estimate the unknown population proportion p .*

- In general, a quantity calculated from data is called a statistic, and a statistic that is used to estimate an unknown constant, or parameter, is called a **point estimator or point estimate.**

- For example, if *X* $\sim$ *Bin(10, p), and we observe X = 3, then the number p^=* 3/10 *is a point estimate of the unknown parameter p.*
- *On the other hand, if no* particular value is specified for *X, the random quantity p ^ = X/10 is often called a point* estimator of *p.*

- the letter ϑ *is used to denote an unknown parameter, and* ϑ ^ *to* denote an estimator of ϑ*.*

Point Estimator of Population Mean

An point estimate of population mean,μ , is the sample mean

$$\bar{x} = \frac{\sum x_i}{n}$$

A sample of weights of 34 male freshman students was obtained.

185	161	174	175	202	178	202	139	177
170	151	176	197	214	283	184	189	168
188	170	207	180	167	177	166	231	176
184	179	155	148	180	194	176		

If one wanted to estimate the true mean of all male freshman students, you might use the sample mean as a point estimate for the true mean.

$$\text{sample mean} = \bar{\text{x}} = 182.44$$

Point Estimation of Population Proportion

An point estimate of population proportion, p, is the sample proportion $\hat{p} = x / n$, where x is the number of successes in the sample.

A sample of 200 students at a large university is selected to estimate the proportion of students that wear contact lens. In this sample 47 wear contact lens.

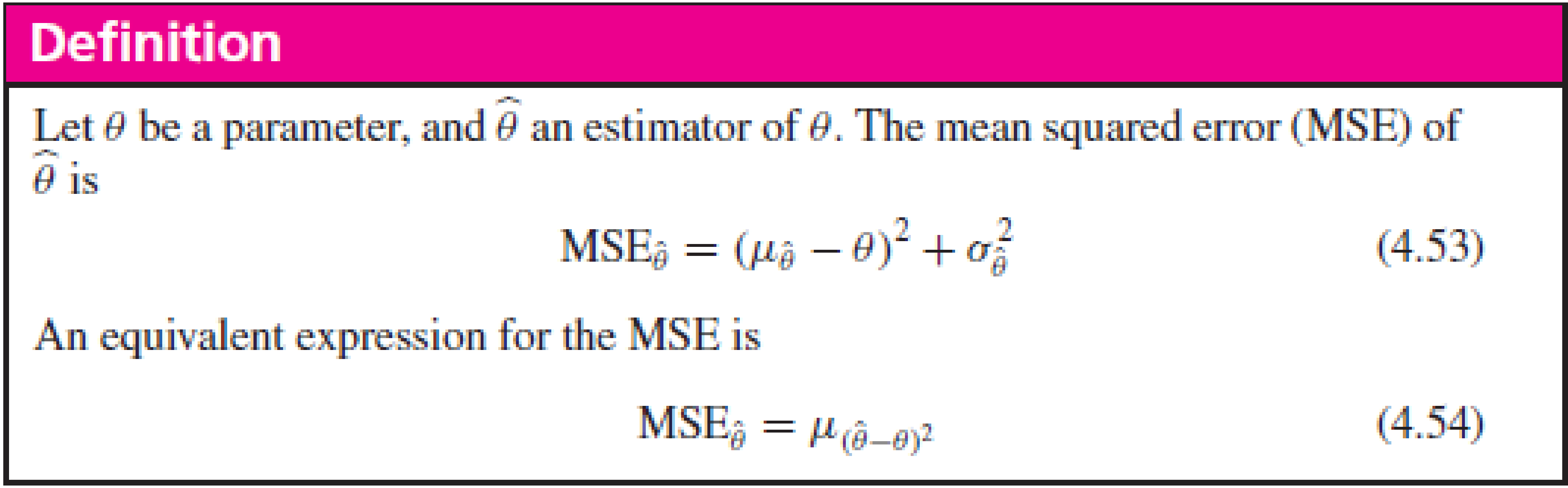
$$\hat{p} = 47/200 = .235$$

1. **Given a point estimator, how do we determine how good it is?**
2. **What methods can be used to construct good point estimators?**

Measuring the Goodness of an Estimator

- an estimator should be both accurate and precise.
- The accuracy of an estimator is measured by its bias,
- the precision is measured by its standard deviation, or uncertainty.
- The quantity most often used to evaluate the overall goodness of an estimator is the **mean squared error (abbreviated MSE), which combines both bias and uncertainty.**

- The bias of the estimator ϑ^ *is* μ ˆ$\vartheta - \vartheta$*, the difference between the mean of the estimator* and the true value.
- The uncertainty is the standard deviation σ ˆϑ *, and is sometimes referred* to as the **standard error of the estimator.**
- The MSE is found by adding the variance to the square of the bias.



Definition

Let θ be a parameter, and $\widehat{\theta}$ an estimator of θ. The mean squared error (MSE) of $\widehat{\theta}$ is

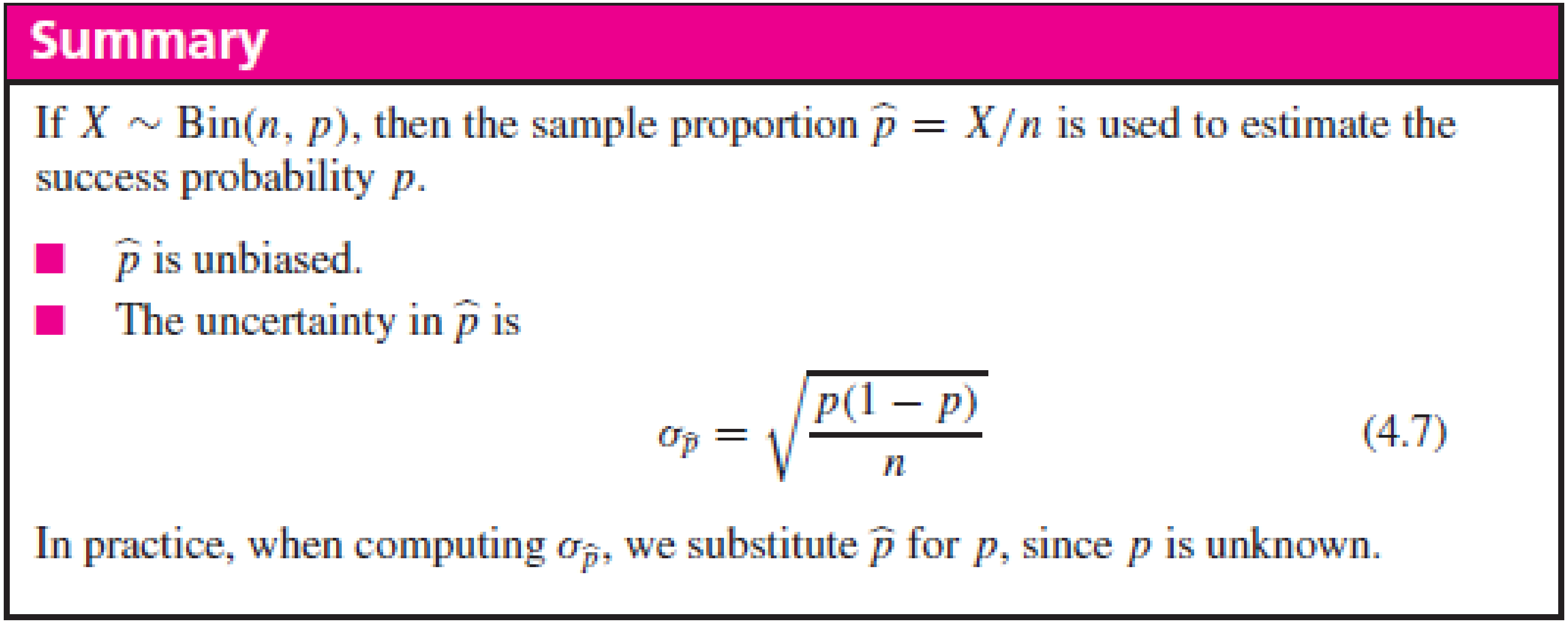
$$\text{MSE}_{\hat{\theta}} = (\mu_{\hat{\theta}} - \theta)^2 + \sigma_{\hat{\theta}}^2 \qquad (4.53)$$

An equivalent expression for the MSE is

$$\text{MSE}_{\hat{\theta}} = \mu_{(\hat{\theta}-\theta)^2} \qquad (4.54)$$

Example

- Let *X ~ Bin(n, p) where p is unknown. Find the MSE of p^ = X/n.*



Summary

If $X \sim \text{Bin}(n, p)$, then the sample proportion $\widehat{p} = X/n$ is used to estimate the success probability p.

- $\widehat{p}$ is unbiased.
- The uncertainty in $\widehat{p}$ is

$$\sigma_{\widehat{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad (4.7)$$

In practice, when computing $\sigma_{\widehat{p}}$, we substitute $\widehat{p}$ for p, since p is unknown.

MSE is 0 + *p(1 – p)/n, or p(1 – p)/n.*

the estimator was unbiased, so the MSE was equal to the variance of the estimator.

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. The sample variance is $s^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$. It can be shown that s^2 has mean $\mu_{s^2} = \sigma^2$ and variance $\sigma_{s^2}^2 = 2\sigma^4/(n-1)$. Consider the estimator $\widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$, in which the sum of the squared deviations is divided by n rather than $n-1$. Compute the bias, variance, and mean squared error of both s^2 and $\widehat{\sigma}^2$. Show that $\widehat{\sigma}^2$ has smaller mean squared error than s^2.

Solution

Since $\mu_{s^2} = \sigma^2$, s^2 is unbiased for σ^2, so the mean squared error is equal to the variance: $\text{MSE}_{s^2} = 2\sigma^4/(n-1)$. To compute the bias and variance of $\widehat{\sigma}^2$, note that

$$\widehat{\sigma}^2 = \frac{n-1}{n}s^2$$

It follows that

$$\mu_{\widehat{\sigma}^2} = \frac{n-1}{n}\mu_{s^2} = \frac{n-1}{n}\sigma^2$$

Therefore

$$\text{Bias of } \widehat{\sigma}^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\sigma^2/n$$

The variance is given by

$$\sigma^2_{\widehat{\sigma}^2} = \frac{(n-1)^2}{n^2}\sigma^2_{s^2} = \frac{2(n-1)}{n^2}\sigma^4$$

The mean squared error of $\widehat{\sigma}^2$ is therefore

$$\begin{aligned} \mathrm{MSE}_{\widehat{\sigma}^2} &= \left(\frac{-\sigma^2}{n}\right)^2 + \frac{2(n-1)}{n^2}\sigma^4 \\ &= \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

To show that $\widehat{\sigma}^2$ has smaller mean squared error than s^2, we subtract:

$$\begin{aligned} \mathrm{MSE}_{s^2} - \mathrm{MSE}_{\widehat{\sigma}^2} &= \frac{2\sigma^4}{n-1} - \frac{(2n-1)\sigma^4}{n^2} \\ &= \frac{3n-1}{n^2(n-1)} \\ &> 0 \qquad \text{(since } n > 1\text{)} \end{aligned}$$

1. **Given a point estimator, how do we determine how good it is?**
2. **What methods can be used to construct good point estimators?**

Maximum Likelihood Estimation

The idea behind the method of maximum likelihood is to estimate a parameter with the value that makes the observed data most likely.

To illustrate the method,

Let $X \sim$ *Bin(20, p) where p is unknown.*

Suppose we observe the value X = 7.

The probability mass function is

- $f(7; p) = (20!/7!13!)p^{7}(1 - p)^{13}$

- the probability mass function is treated as a function of *p, with the* data value 7 being constant.

- When a probability mass function or probability density function is considered to be a function of parameters, it is called a **likelihood function.**

- The **maximum likelihood estimate (MLE) is the value** $\boldsymbol{\hat{p}}$ ***which, when substituted*** for *p, maximizes the likelihood function.*

- Computing the maximum of the likelihood function *f (7; p).*

- *In principle, we could maximize* this function by taking the derivative with respect to *p and setting it equal to 0.*

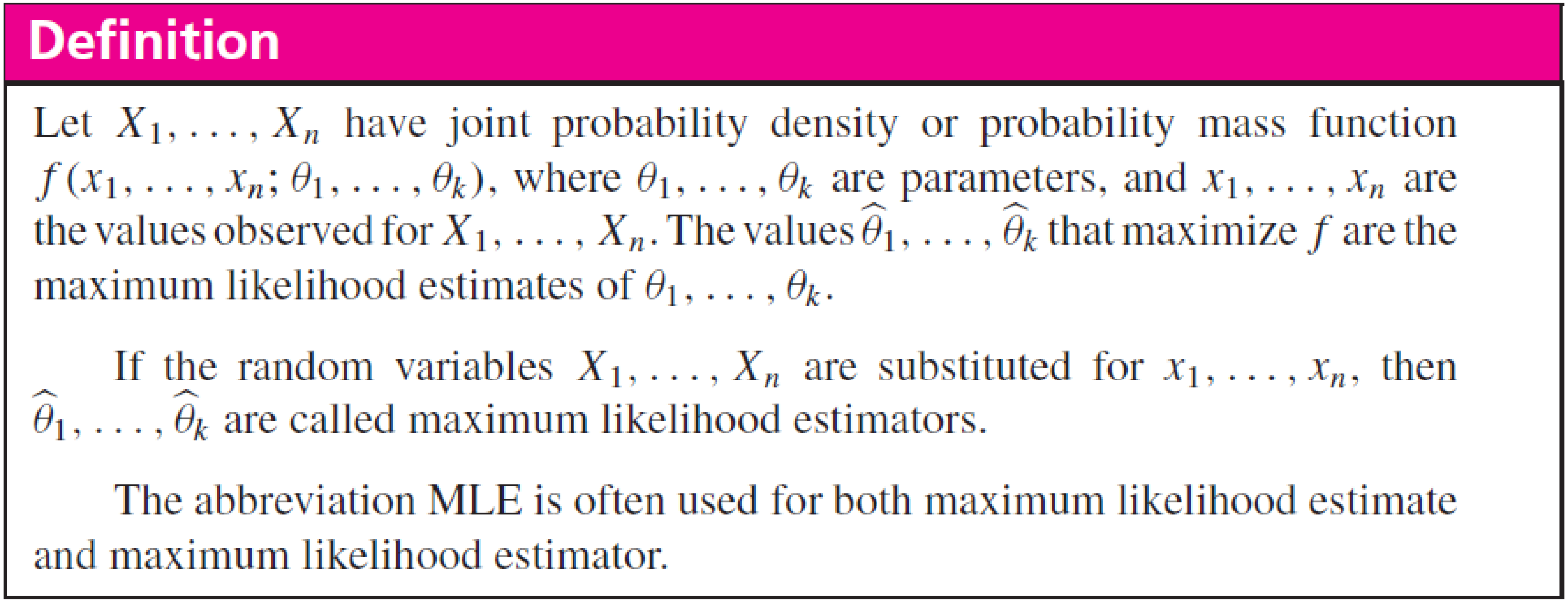
- *It is* easier, however, to maximize ln *f (7; p) instead.*

- *Note that the quantity that maximizes* the logarithm of a function is always the same quantity that maximizes the function itself.

- $\ln f\,(7; p) = \ln 20! - \ln 7! - \ln 13! + 7 \ln p + 13 \ln(1 - p)$
- We take the derivative with respect to *p and set it equal to 0:*
- $d/dp\,(\ln f\,(7; p)) = 7/p - 13/1 - p = 0$

- the maximum likelihood estimate is *p* **^** *= 7/20.*
- It is easy to see that whatever value is observed for *X, the maximizing value would*
- be *X/20. We say, therefore, that the maximum*

  likelihood estimator is p **^** *= X/20*



Definition

Let $X_1, \ldots, X_n$ have joint probability density or probability mass function $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_k)$, where $\theta_1, \ldots, \theta_k$ are parameters, and $x_1, \ldots, x_n$ are the values observed for $X_1, \ldots, X_n$. The values $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$ that maximize f are the maximum likelihood estimates of $\theta_1, \ldots, \theta_k$.

If the random variables $X_1, \ldots, X_n$ are substituted for $x_1, \ldots, x_n$, then $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$ are called maximum likelihood estimators.

The abbreviation MLE is often used for both maximum likelihood estimate and maximum likelihood estimator.

Desirable Properties of Maximum Likelihood Estimators

1. **In most cases, as the sample size *n increases, the bias of the MLE converges to 0.***
2. **In most cases, as the sample size *n increases, the variance of the MLE converges* to a theoretical minimum.**

Together, these two properties imply that when the sample size is sufficiently large,the bias of the MLE will be negligible, and the variance will be nearly as small as is theoretically possible.

Statement of the Problem

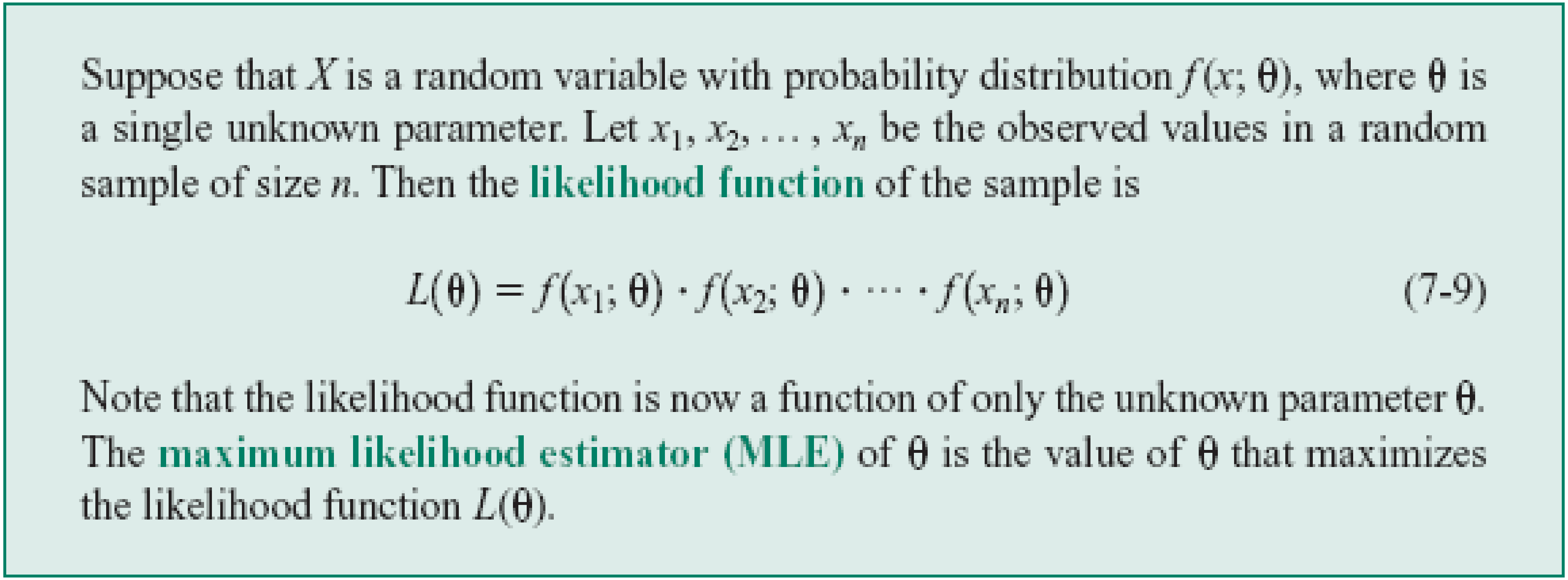
- Suppose we have a random sample *X1, X2,..., Xn whose assumed probability distribution depends on some* unknown parameter *ϑ.*
- *Our primary goal here will be to find a point estimator u(X1, X2,..., Xn), such* that *u(x1, x2,..., xn) is a "good" point estimate of ϑ, where x1, x2,..., xn are the observed values of the random* sample.

- For example, if we plan to take a random sample *X1, X2,..., Xn for which the Xi are assumed to be* normally distributed with mean *μ and variance σ^2, then our goal will be to find a good estimate of μ, say, using* the data *x1, x2,..., xn that we obtained from our specific random sample.*

Methods of Point Estimation

Method of Maximum Likelihood

Definition



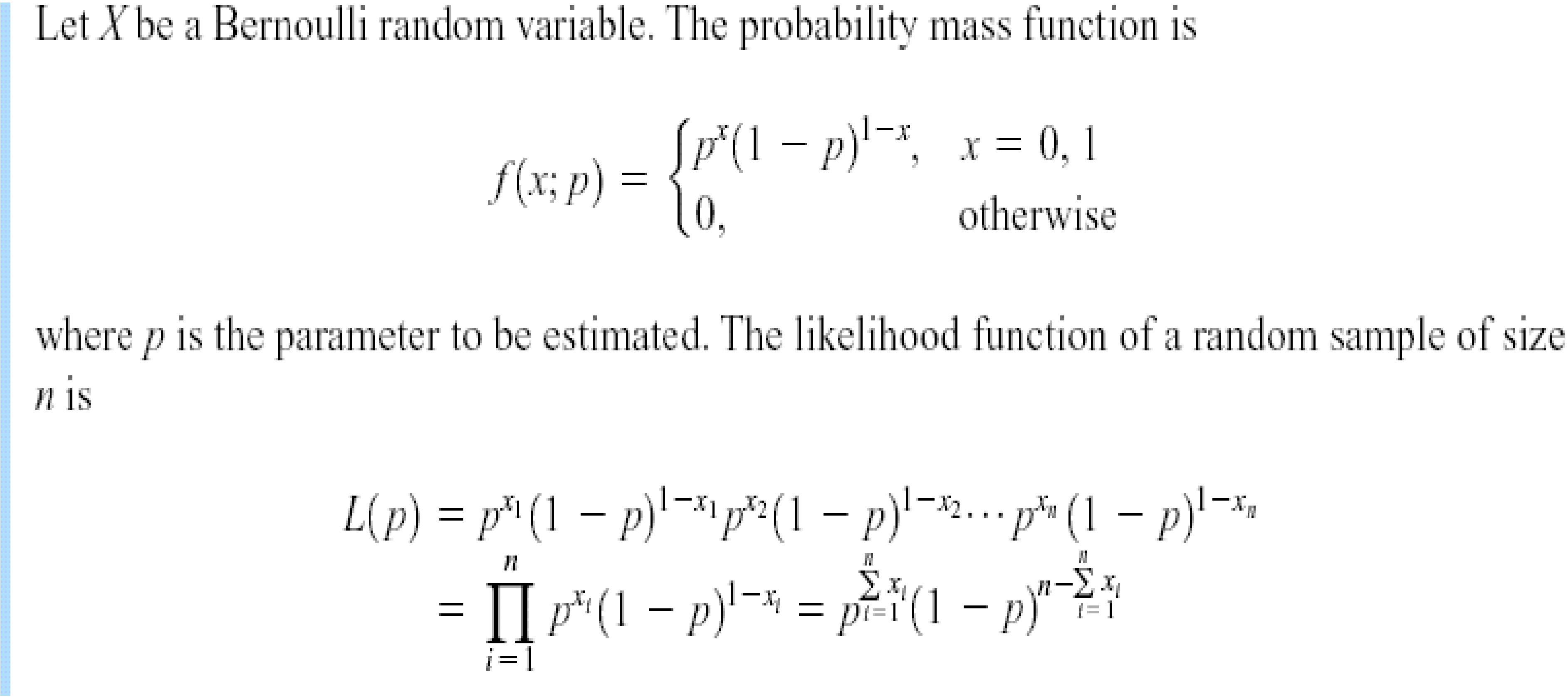
Suppose that X is a random variable with probability distribution $f(x; \theta)$, where θ is a single unknown parameter. Let $x_1, x_2, \ldots, x_n$ be the observed values in a random sample of size n. Then the **likelihood function** of the sample is

$$L(\theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \cdots \cdot f(x_n; \theta) \qquad (7\text{-}9)$$

Note that the likelihood function is now a function of only the unknown parameter θ. The **maximum likelihood estimator (MLE)** of θ is the value of θ that maximizes the likelihood function $L(\theta)$.

Methods of Point Estimation

Example



Let X be a Bernoulli random variable. The probability mass function is

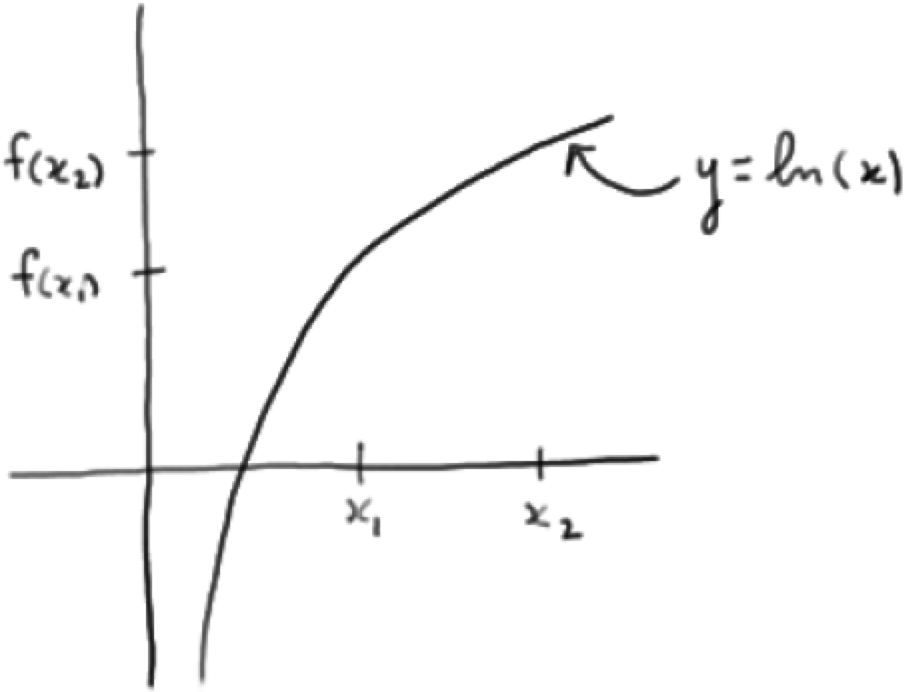
$$f(x;p) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

where p is the parameter to be estimated. The likelihood function of a random sample of size n is

$$\begin{aligned} L(p) &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n} \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

- in order to implement the method of maximum likelihood, we need to find the *p that* maximizes the likelihood *L(p).*
- *in order to maximize* the function, we to need to differentiate the likelihood function with respect to *p.*
- *To* make the differentiation a bit easier we use natural logarithm.

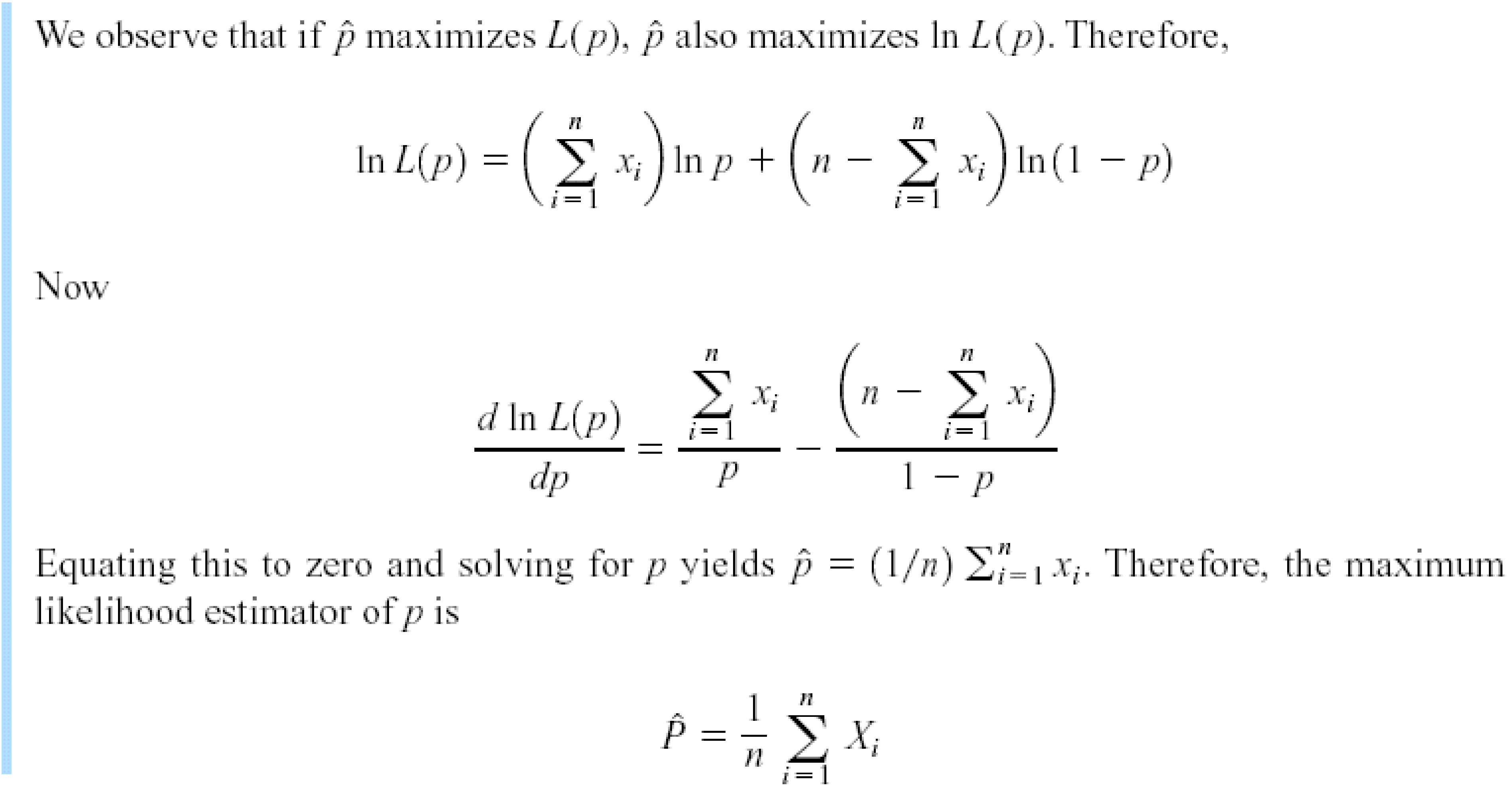
- Note:
- the natural logarithm is an increasing function of *x*



- That is, if *x1 < x2, then f(x1) < f(x2).*
- *That means that the value of p that maximizes the natural* logarithm of the likelihood function *ln(L(p)) is also the value of p that maximizes the likelihood* function *L(p).*
- *So, the "trick" is to take the derivative of ln(L(p)) (with respect to p) rather than taking* the derivative of *L(p).*

Methods of Point Estimation

Example (continued)



We observe that if $\hat{p}$ maximizes $L(p)$, $\hat{p}$ also maximizes $\ln L(p)$. Therefore,

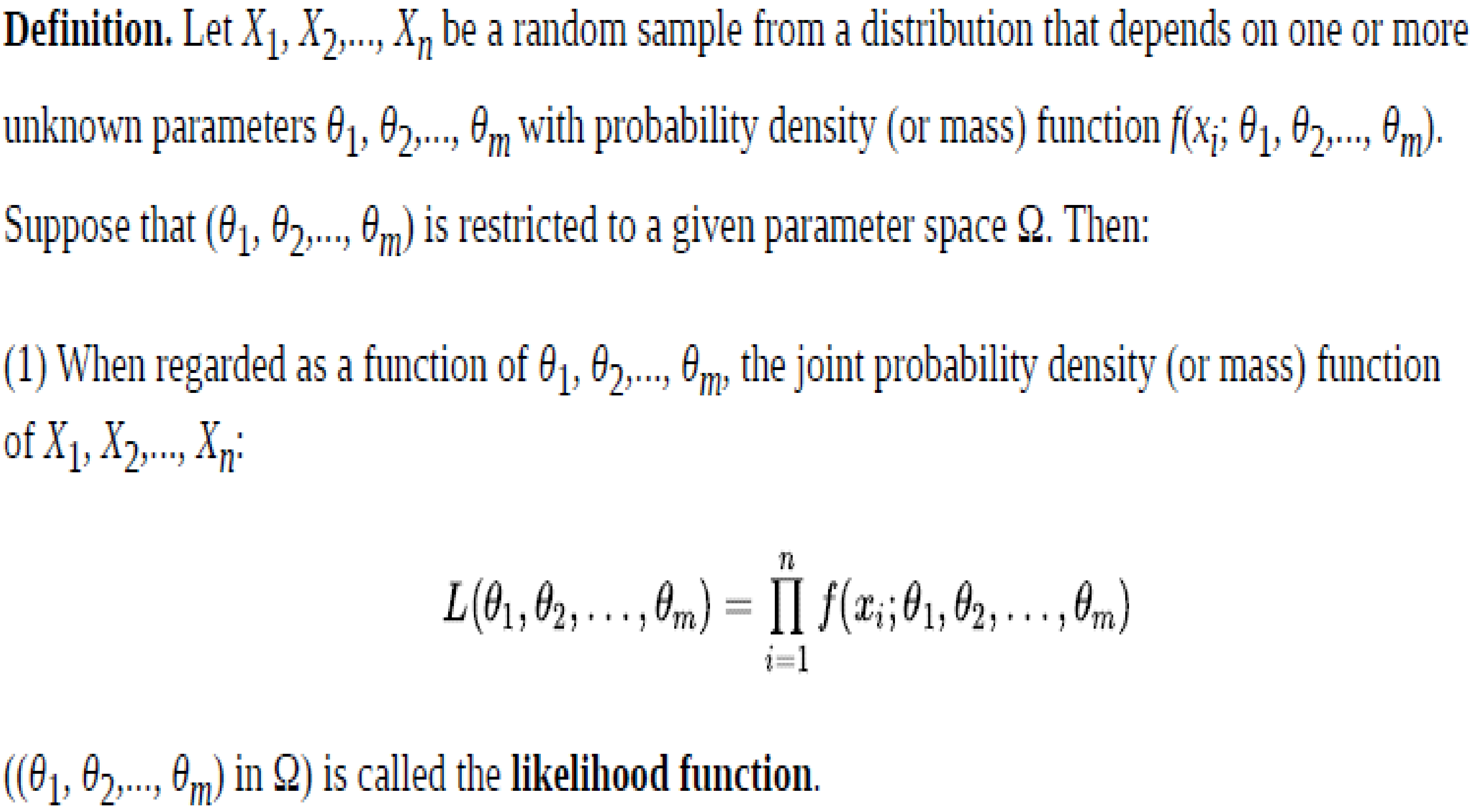
$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1 - p)$$

Now

$$\frac{d \ln L(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1 - p}$$

Equating this to zero and solving for p yields $\hat{p} = (1/n) \sum_{i=1}^{n} x_i$. Therefore, the maximum likelihood estimator of p is

$$\hat{P} = \frac{1}{n} \sum_{i=1}^{n} X_i$$



Definition. Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution that depends on one or more unknown parameters $\theta_1, \theta_2, \dots, \theta_m$ with probability density (or mass) function $f(x_i; \theta_1, \theta_2, \dots, \theta_m)$. Suppose that $(\theta_1, \theta_2, \dots, \theta_m)$ is restricted to a given parameter space Ω. Then:

(1) When regarded as a function of $\theta_1, \theta_2, \dots, \theta_m$, the joint probability density (or mass) function of $X_1, X_2, \dots, X_n$:

$$L(\theta_1, \theta_2, \dots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \dots, \theta_m)$$

($(\theta_1, \theta_2, \dots, \theta_m)$ in Ω) is called the **likelihood function**.

(2) If: $[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)]$

is the *m*-tuple that maximizes the likelihood function, then:

$$\hat{\theta}_i = u_i(X_1, X_2, \ldots, X_n)$$

is the **maximum likelihood estimator** of θ_i, for $i = 1, 2, \ldots, m$.

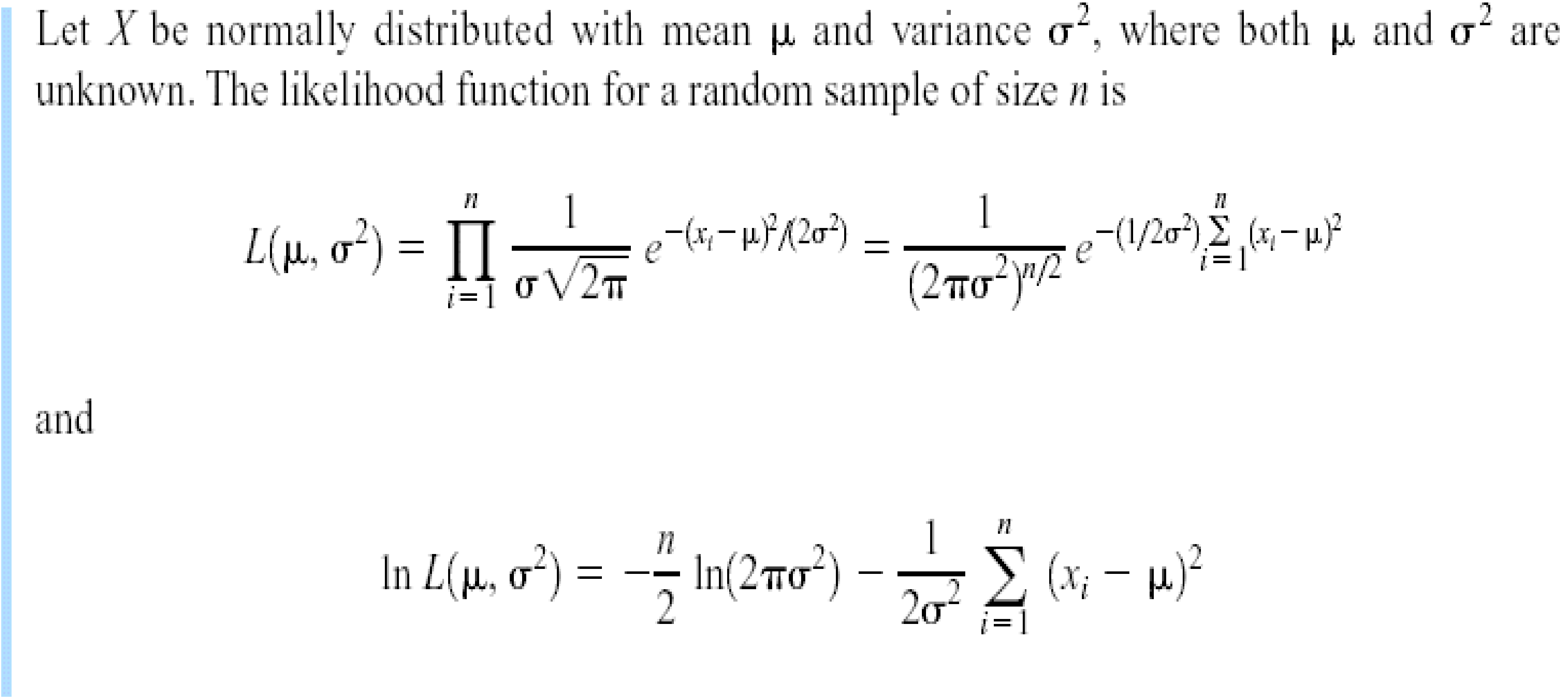
(3) The corresponding observed values of the statistics in (2), namely:

$$[u_1(x_1, x_2, \ldots, x_n), u_2(x_1, x_2, \ldots, x_n), \ldots, u_m(x_1, x_2, \ldots, x_n)]$$

are called the **maximum likelihood estimates** of θ_i, for $i = 1, 2, \ldots, m$.

Methods of Point Estimation

Example



Let X be normally distributed with mean μ and variance σ^2, where both μ and σ^2 are unknown. The likelihood function for a random sample of size n is

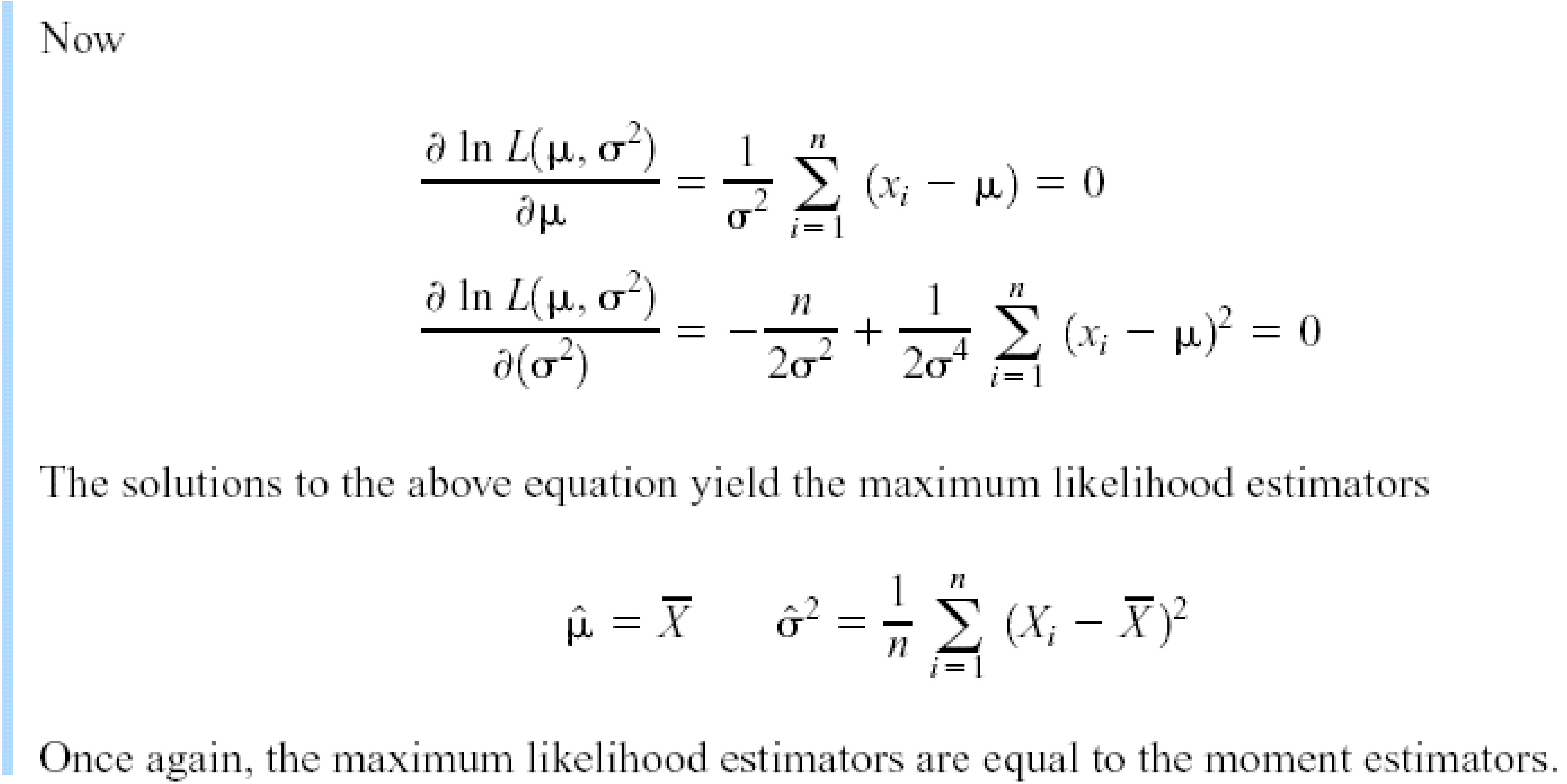
$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}}\, e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

and

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Methods of Point Estimation

Example (continued)



Now

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

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

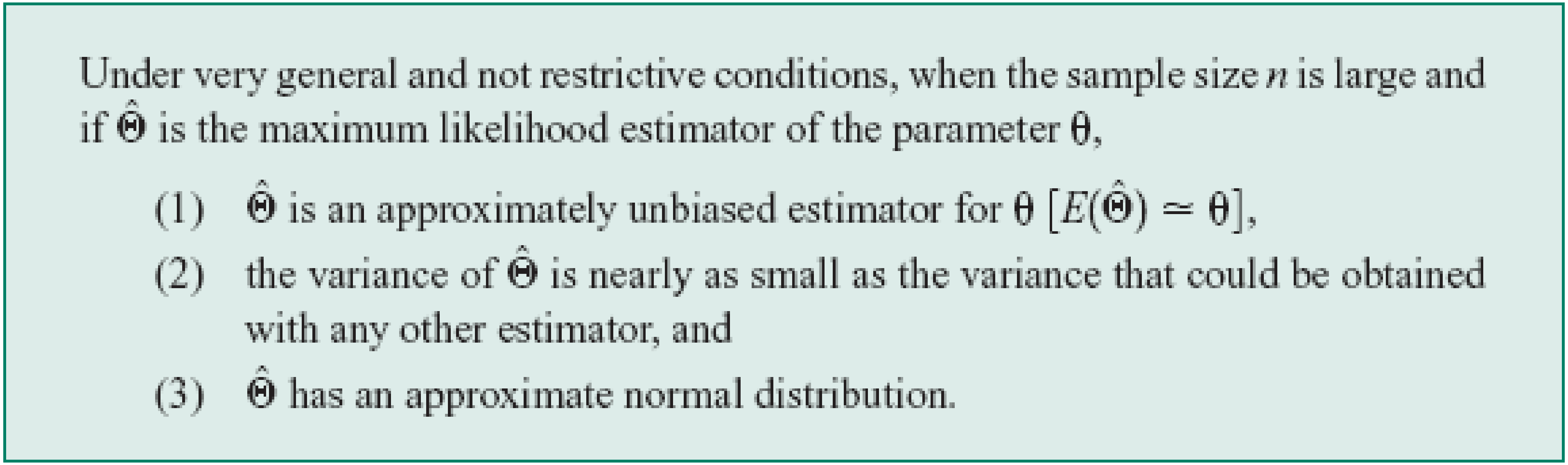
The solutions to the above equation yield the maximum likelihood estimators

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Once again, the maximum likelihood estimators are equal to the moment estimators.

Methods of Point Estimation

Properties of the Maximum Likelihood Estimator



Under very general and not restrictive conditions, when the sample size n is large and if $\hat{\Theta}$ is the maximum likelihood estimator of the parameter θ,

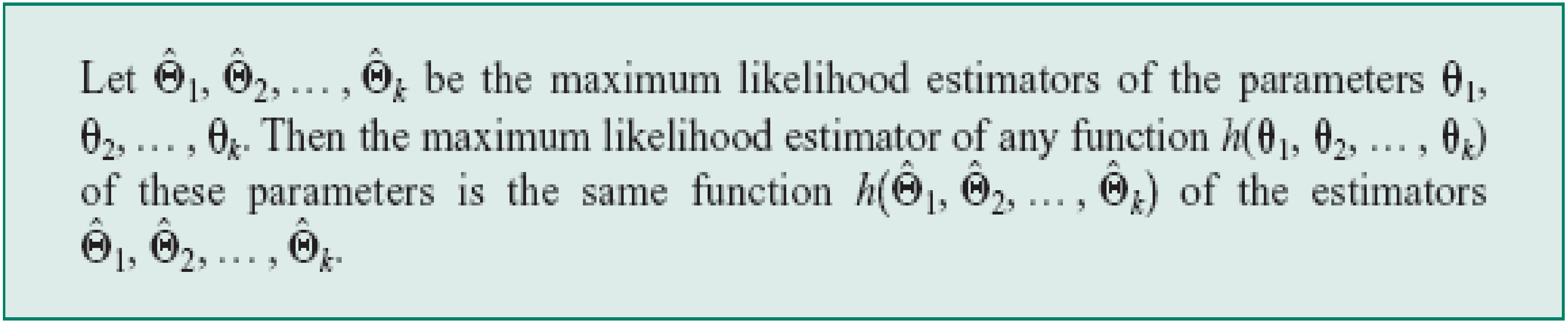
(1) $\hat{\Theta}$ is an approximately unbiased estimator for θ $[E(\hat{\Theta}) \simeq \theta]$,

(2) the variance of $\hat{\Theta}$ is nearly as small as the variance that could be obtained with any other estimator, and

(3) $\hat{\Theta}$ has an approximate normal distribution.

Methods of Point Estimation

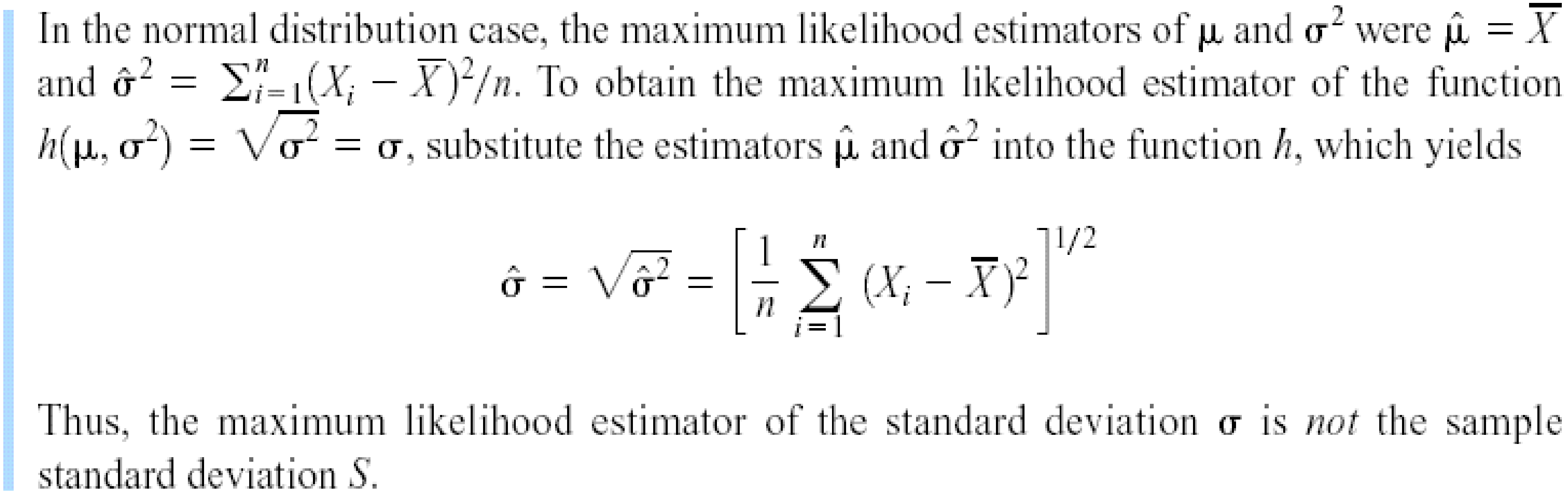
The Invariance Property



Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$ be the maximum likelihood estimators of the parameters $\theta_1, \theta_2, \ldots, \theta_k$. Then the maximum likelihood estimator of any function $h(\theta_1, \theta_2, \ldots, \theta_k)$ of these parameters is the same function $h(\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k)$ of the estimators $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k$.

Methods of Point Estimation

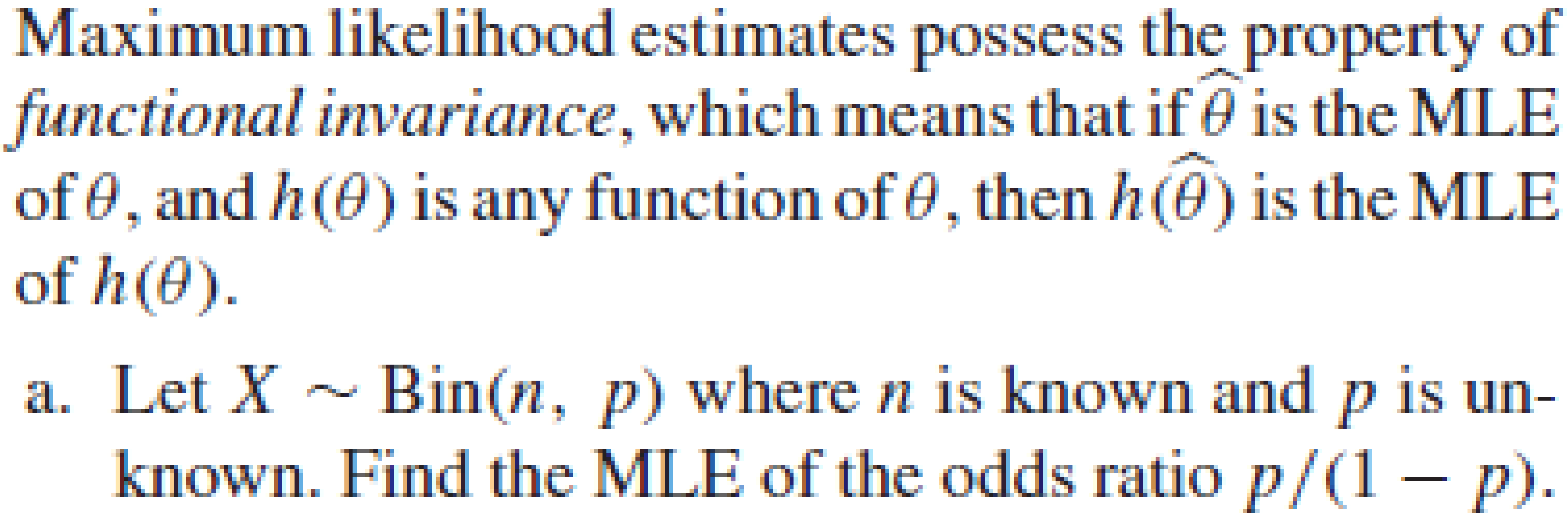
Example



In the normal distribution case, the maximum likelihood estimators of μ and σ^2 were $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$. To obtain the maximum likelihood estimator of the function $h(\mu, \sigma^2) = \sqrt{\sigma^2} = \sigma$, substitute the estimators $\hat{\mu}$ and $\hat{\sigma}^2$ into the function h, which yields

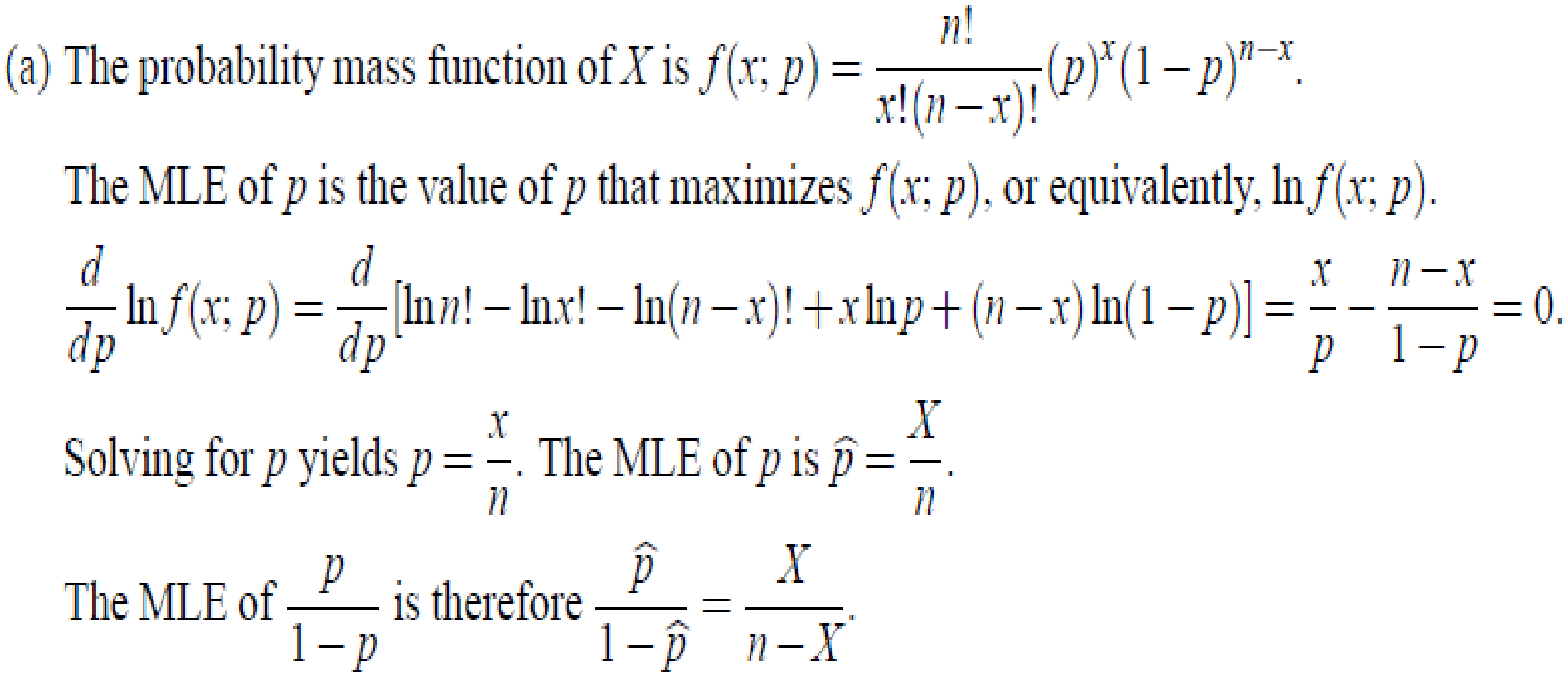
$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{1/2}$$

Thus, the maximum likelihood estimator of the standard deviation σ is *not* the sample standard deviation S.



Maximum likelihood estimates possess the property of *functional invariance*, which means that if $\widehat{\theta}$ is the MLE of θ, and $h(\theta)$ is any function of θ, then $h(\widehat{\theta})$ is the MLE of $h(\theta)$.

a. Let $X \sim \text{Bin}(n,\ p)$ where n is known and p is unknown. Find the MLE of the odds ratio $p/(1-p)$.



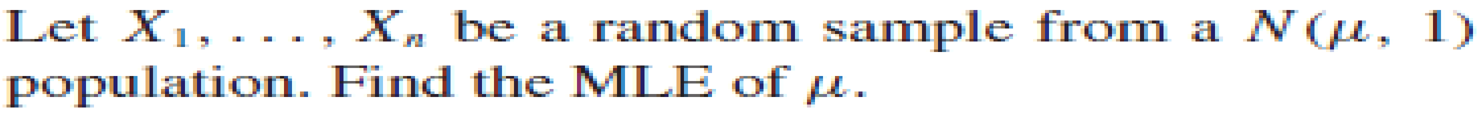
(a) The probability mass function of X is $f(x;\, p) = \dfrac{n!}{x!(n-x)!}(p)^x(1-p)^{n-x}$.

The MLE of p is the value of p that maximizes $f(x;\, p)$, or equivalently, $\ln f(x;\, p)$.

$$\frac{d}{dp}\ln f(x;\, p) = \frac{d}{dp}[\ln n! - \ln x! - \ln(n-x)! + x\ln p + (n-x)\ln(1-p)] = \frac{x}{p} - \frac{n-x}{1-p} = 0.$$

Solving for p yields $p = \dfrac{x}{n}$. The MLE of p is $\widehat{p} = \dfrac{X}{n}$.

The MLE of $\dfrac{p}{1-p}$ is therefore $\dfrac{\widehat{p}}{1-\widehat{p}} = \dfrac{X}{n-X}$.



Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, 1)$ population. Find the MLE of μ.

The joint probability density function of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2} = (2\pi)^{-n/2} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2}.$$

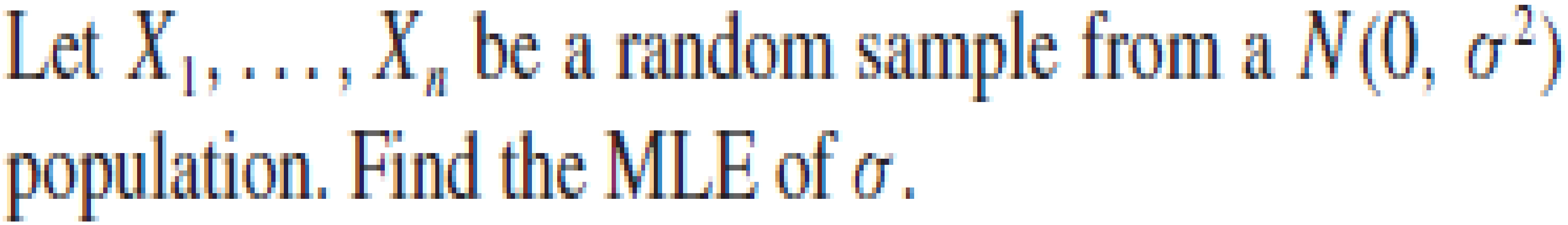
The MLE is the value of μ that maximizes $f(x_1, \ldots, x_n; \mu)$, or equivalently, $\ln f(x_1, \ldots, x_n; \mu)$.

$$\frac{d}{d\mu} \ln f(x_1, \ldots, x_n; \mu) = \frac{d}{d\mu}\left[-(n/2)\ln 2\pi - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2}\right] = \sum_{i=1}^{n}(x_i - \mu) = 0.$$

Now we solve for μ:

$$\sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n} x_i - n\mu = 0, \text{ so } \mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

The MLE of μ is $\widehat{\mu} = \overline{X}$.



Let $X_1, \ldots, X_n$ be a random sample from a $N(0, \sigma^2)$ population. Find the MLE of σ.

The joint probability density function of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-x_i^2/2\sigma^2} = (2\pi)^{-n/2}\sigma^{-n}e^{-\sum_{i=1}^{n} x_i^2/2\sigma^2}.$$

The MLE is the value of σ that maximizes $f(x_1, \ldots, x_n; \sigma)$, or equivalently, $\ln f(x_1, \ldots, x_n; \sigma)$.

$$\frac{d}{d\sigma}\ln f(x_1, \ldots, x_n; \sigma) = \frac{d}{d\sigma}\left[-(n/2)\ln 2\pi - n\ln\sigma - \sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}\right] = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} x_i^2}{\sigma^3} = 0.$$

Now we solve for σ:

$$-\frac{n}{\sigma} + \frac{\sum_{i=1}^{n} x_i^2}{\sigma^3} = 0$$

$$-n\sigma^2 + \sum_{i=1}^{n} x_i^2 = 0$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

The MLE of σ is $\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}}$.