# Mathematical Statistics and Data Analysis

## Lecture 8: Parameter Estimation


Lyu Ni

DaSE@ECNU
(lni@dase.ecnu.edu.cn)


November 24, 2019

# Outlines

1. Point Estimation
2. Methods of Finding an estimate
   - Method of Moments
   - Method of Maximum Likelihood
3. Property of Estimates
   - Unbiasedness
   - Efficiency
   - Consistency
   - Asymptotic Normality
   - Mean Squared Error
   - Uniform Minimum Variance Unbiased Estimate, UMVUE
   - Cramér-Rao Inequality
4. Bayesian Approach

# Reading Material

Textbook:

- Rice: Chapter 8;
- Mao: Chapter 6;

# Point Estimation

## Example

*On the Error of Counting with a Haemacytometer*(1907) by Student.

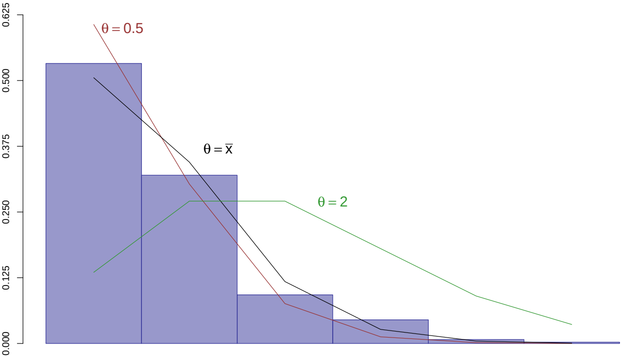
- The famous statistician William Gosset, who worked for Guinness brewery, took measure of the number of yeast cells per square in a hemocytometer. The count of yeast cells could be model with a probability distribution known as 'Poisson distribution' $P(\theta)$.
- This distribution $P(\theta)$ has an unknown parameter $\theta$.
- The data is shown as follows:

| Containing | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Actual | 213 | 128 | 37 | 18 | 3 | 1 |

- Problem: What is a guess of $\theta$?

# Point Estimation

## Example (Con'd)

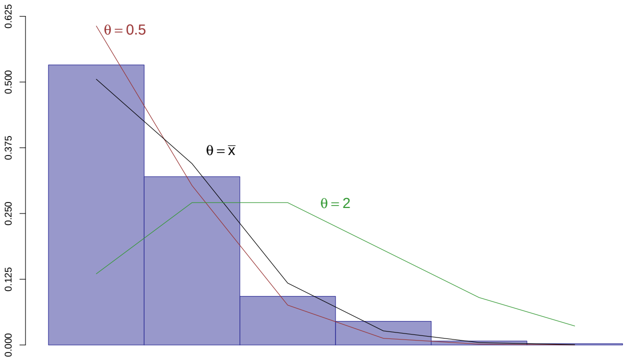



## Definition

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a population with unknown parameter $\theta$. The statistic $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ is called an **point estimate** of $\theta$.

# Point Estimation

## Example (Con'd)




## Definition

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a population with unknown parameter $\theta$. The statistic $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ is called an **point estimate** of $\theta$.

# Method of Moments

The $k$th moment of a random variable $X$ is defined as

$$\mu_k = E(X^k).$$

Suppose that $x_1, x_2, \cdots, x_n$ is a sample. The $k$th sample moment is defined as

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Then, we can view $a_k$ as an estimate of $\mu_k$, and thus let $\hat{\mu}_k = a_k$.

## Idea

The method of moments estimates parameters by finding expressions for them in terms of the lowest moments and then substitution sample moments into the expressions.

# Method of Moments

- The p.d.f. or p.m.f. of the population is $f(x : \theta_1, \cdots, \theta_k)$;
- $(\theta_1, \cdots, \theta_k) \in \Theta$ is an unknown parameter vector;
- $\Theta$ is a parameter space.
- Suppose that the $i$th moment $\mu_i$ exists, $i = 1, 2, \cdots, k$;
- The parameters $\theta_1, \cdots, \theta_k$ can be written as the functions of $\mu_1, \cdots, \mu_k$, that is $\theta_j = \theta_j(\mu_1, \cdots, \mu_k)$;
- The method of moments estimates of $\theta_j$ is

$$\hat{\theta}_j = \theta_j(\hat{\mu}_1, \cdots, \hat{\mu}_k), j = 1, \cdots, k$$

- Furthermore, if $\eta = g(\theta_1, \cdots, \theta_k)$ is to be estimated, the method of moment estimate of $\eta$ is

$$\hat{\eta} = g(\hat{\theta}_1, \cdots, \hat{\theta}_k)$$

# Method of Moments

## Example: Exponential Distribution

The p.d.f. of an exponential distribution is

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

and $x_1, x_2, \cdots, x_n$ is a sample.

- Consider $k = 1$. Since $EX = 1/\lambda$, i.e. $\lambda = 1/EX$, then the method of moment estimate of $\lambda$ is

$$\hat{\lambda} = 1/\bar{x};$$

- Consider $k = 2$. Since $Var(X) = 1/\lambda^2$, i.e. $\lambda = 1/\sqrt{Var(X)}$, then the moment of method estimate of $\lambda$ is

$$\hat{\lambda} = 1/s.$$

# Method of Moments

## Remark

- The method of moment estimate is straight forward.
- The method of moment estimate is **not unique**.
- Problem: Which one is better?

## Rule of thumb

The sample moments used in the method of moment should be as **low** as possible.

# Method of Moments

## Example: Poisson Distribution

The p.d.f. of a Poisson distribution is

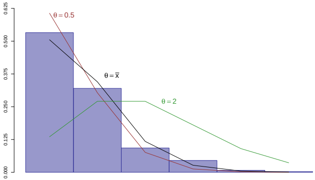
$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}, x = 0, 1, 2, \cdots$$

and $x_1, x_2, \cdots, x_n$ is a sample. Since $E(X) = \lambda$, the method of moment estimate of $\lambda$ is

$$\lambda = \bar{x}$$

The data are shown as follows:

| | | | | | | |
|---|---|---|---|---|---|---|
| Containing | 0 | 1 | 2 | 3 | 4 | 5 |
| Actual | 213 | 128 | 37 | 18 | 3 | 1 |



# Method of Moments

## Example: Uniform Distribution

The p.d.f. of a uniform distribution is

$$f(x;\lambda) = \frac{1}{b-a} I_{(a,b)}(x)$$

with two unknown parameter $a$ and $b$. Suppose that $x_1, x_2, \cdots, x_n$ is a sample. Since

$$E(X) = \frac{a+b}{2} \quad \text{and} \quad Var(X) = \frac{(b-a)^2}{12},$$

it is obvious that $a = EX - \sqrt{3Var(X)}$ and $b = EX + \sqrt{3Var(X)}$. Thus, the method of moment estimates of $a$ and $b$ are

$$\hat{a} = \bar{x} - \sqrt{3}s \quad \text{and} \quad \hat{b} = \bar{x} + \sqrt{3}s.$$

# Method of Maximum Likelihood

## Example One

Suppose that it is difficult to distinguish two urns from the appearance. Urn A contains $99$ white balls and $1$ black ball while Urn B contains $1$ white ball and $99$ black balls. Here we randomly select an urn and then take a ball. If this ball is a white ball, which urn do you select?

**Solution:** Let the event

$$A = \{\text{A white ball is taken}\}.$$

- If Urn A is chosen, the probability $P(A) = 0.99$.
- If Urn B is chosen, the probability $P(A) = 0.01$

If $A$ occurs and then we may think that it is likely that this white ball is taken out of Urn A.

# Method of Maximum Likelihood

## Example Two

We flip a coin and use a random variable $X$ to represent the result. If it heads up, then $X = 1$; otherwise, $X = 0$. Then, $X$ is distributed as a Bernoulli distribution $B(p)$ with a unknown parameter $p$.
Suppose that $x_1, x_2, \cdots, x_n$ is a sample. The joint p.m.f. of $(x_1, x_2, \cdots, x_n)$ is

$$f(x_1, x_2, \cdots, x_n; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

Since $p$ is unknown, this function could be thought to be a likelihood function of $p$, denoted as $L(p)$. That is,

$$L(p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}, p \in (0, 1).$$

# Method of Maximum Likelihood

## Example Two (Con'd)

- How to determine $p$?
- We would like to choose $p$ so that the probability is as large as possible. Equivalently,

$$\hat{p} = \arg\max_{p} L(p)$$

Then,

$$\frac{\partial \ln L(p)}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0.$$

Thus, the maximum likelihood estimate of $p$ is

$$\hat{p} = \hat{p}(x_1, x_2, \cdots, x_n) = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

# Method of Maximum Likelihood

## Definition

Suppose that the p.m.f. or p.d.f. of the population is $p(x;\theta), \theta \in \Theta$, where $\theta$ is a unknown parameter (vector) and $\Theta$ is the parameter space. Let $x_1, x_2, \cdots, x_n$ be a sample. The joint p.m.f. or p.d.f. of $x_1, x_2, \cdots, x_n$ could be thought to be a function of $\theta$, denoted as $L(\theta; x_1, \cdots, x_n)$ or $L(\theta)$.

- This function $L(\theta)$ is called as the **likelihood function**.
- A statistic $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ is called **maximum likelihood estimate (MLE)** if this statistic $\hat{\theta}$ satisfies

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

# Method of Maximum Likelihood

## Example: Normal Distribution

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a normal distribution $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2)$ is a two-dimensional parameter vector. The likelihood function is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} \right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}, \end{aligned}$$

and its log-likelihood function is

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

# Method of Maximum Likelihood

## Example: Normal Distribution (Con'd)

The partials with respect to $\mu$ and $\sigma^2$ are

$$\begin{aligned}\frac{\partial(-l)}{\partial\mu} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\\ \frac{\partial(-l)}{\partial\sigma^2} &= -\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2+\frac{n}{2\sigma^2}.\end{aligned}$$

Setting the first partial equal to zero and solving for the MLE, we obtain

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x} \quad \text{and} \quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=s_*^2.$$

# Method of Maximum Likelihood

## Example: Normal Distribution (Con'd)

The second-order partial deviates are, respectively,

$$\frac{\partial^2(-l)}{\partial \mu^2} = \frac{n}{\sigma^2} \quad \text{and} \quad \frac{\partial^2(-l)}{\partial(\sigma^2)^2} = \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2\sigma^4}$$

$$\frac{\partial^2(-l)}{\partial(\sigma^2)\partial\mu} = \frac{\partial^2(-l)}{\partial\mu\partial(\sigma^2)} = \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu).$$

It is easy to verify the matrix is negative definite since

$$\frac{\partial(-l)}{\partial \mu^2}\Big|_{\mu=\bar{x},\sigma^2=s_*^2} = \frac{n}{s_*^2} > 0$$

$$\left(\frac{\partial(-l)}{\partial \mu^2}\cdot\frac{\partial(-l)}{\partial(\sigma^2)^2} - \left(\frac{\partial(-l)}{\partial(\sigma^2)\partial\mu}\right)^2\right)\Big|_{\mu=\bar{x},\sigma^2=s_*^2} = \frac{n^2}{2s_*^6} > 0$$

# Method of Maximum Likelihood

## Example: Uniform Distribution

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a uniform distribution $U(0, \theta)$. Find the maximum likelihood estimate of $\theta$.

**Solution:** The likelihood function is

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{\{0<x_i\leq\theta\}} = \frac{1}{\theta^n} I_{\{0<x_{(n)}\leq\theta\}}$$

To maximize the likelihood,

- let $I_{\{x_{(n)}\leq\theta\}}$ be 1;
- let $1/\theta^n$ be as large as possible.

Since $\frac{1}{\theta^n}$ is decreasing in $\theta$, the maximum likelihood estimate of $\theta$ is

$$\hat{\theta} = x_{(n)}$$

# Method of Maximum Likelihood

## Theorem: Invariance Property

If $\hat{\theta}$ is the MLE of $\theta$, then for any function of $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta})$.

## Example: Normal Distribution (Revisit)

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from $N(\mu, \sigma^2)$. The MLE of $\mu$ and $\sigma^2$ are respectively

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma^2} = s_*^2.$$

From the invariance property, find the MLE:

- The standard deviation $\sigma$;
- The probability $P(X < 3) = \Phi\left(\frac{3-\mu}{\sigma}\right)$.;
- The $90\%$ quantile $x_{0.90} = \mu + \sigma u_{0.90}$, where $u_{0.90}$ is the $90\%$ quantile of a standard normal r.v.

# Method of Maximum Likelihood

## Theorem: Invariance Property

If $\hat{\theta}$ is the MLE of $\theta$, then for any function of $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta})$.

## Example: Normal Distribution (Revisit)

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from $N(\mu, \sigma^2)$. The MLE of $\mu$ and $\sigma^2$ are respectively

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma^2} = s_*^2.$$

From the invariance property, we have

- The MLE of $\sigma$ is $\hat{\sigma} = s_*$;
- The MLE of $P(X < 3)$ is $\Phi\left(\frac{3-\bar{x}}{s_*}\right)$;
- The MLE of the $90\%$ quantile $x_{0.90}$ is $\bar{x} + s_* u_{0.90}$.

# EM Algorithm

## Example

Suppose that a trial has four results and the probabilities are respectively $\frac{1}{2} - \frac{\theta}{4}$, $\frac{1-\theta}{4}$, $\frac{1+\theta}{4}$ and $\frac{\theta}{4}$, $\theta \in (0,1)$. Among 197 trials, the numbers of four results are $75, 18, 70, 34$. Find the MLE of $\theta$.

**Solution:** Let $y_1, y_2, y_3, y_4$ be the numbers of four results. $\boldsymbol{y} = (y_1, y_2, y_3, y_4)$ is a multinomial distribution and the likelihood function is

$$
\begin{aligned}
L(\theta; \boldsymbol{y}) &\propto \left(\frac{1}{2} - \frac{\theta}{4}\right)^{y_1} \left(\frac{1-\theta}{4}\right)^{y_2} \left(\frac{1+\theta}{4}\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4} \\
&\propto (2-\theta)^{y_1}(1-\theta)^{y_2}(1+\theta)^{y_4}\theta^{y_4}
\end{aligned}
$$

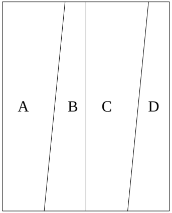
It is difficult to find the maximizer of $L(\theta; \boldsymbol{y})$.

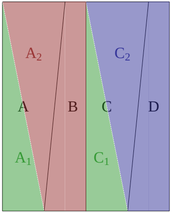
# EM Algorithm

## Example (Con'd)

Two variable $z_i$ and $z_2$ are introduced and are called latent variables.

- Suppose that the first result is divided into two parts with the probabilities $\frac{1-\theta}{4}$ and $\frac{1}{4}$. Let $z_1$ and $y_1 - z_1$ be respectively the number of two parts.
- Suppose that the third result is divided into two parts with the probabilities $\frac{\theta}{4}$ and $\frac{1}{4}$. Let $z_2$ and $y_3 - z_2$ be respectively the number of two parts.



# EM Algorithm

## Example (Con'd)

The likelihood function could be written as

$$\begin{aligned} L(\theta; \boldsymbol{y}) &\propto \left(\frac{1}{4}\right)^{y_1 - z_1} \left(\frac{1-\theta}{4}\right)^{z_1 + y_2} \left(\frac{1}{4}\right)^{y_3 - z_2} \left(\frac{\theta}{4}\right)^{z_2 + y_4} \\ &\propto \theta^{z_2 + y_4} (1-\theta)^{z_1 + y_2} \end{aligned}$$

The log-likelihood function is

$$l(\theta; \boldsymbol{y}, \boldsymbol{z}) = (z_2 + y_4) \ln \theta + (z_1 + y_2) \ln(1 - \theta).$$

Note that

- If $(\boldsymbol{y}, \boldsymbol{z})$ is known, then it is easy to obtain the MLE of $\theta$;
- $\boldsymbol{y}$ is known but $\boldsymbol{z}$ is unknown;
- If $\boldsymbol{y}$ and $\theta$ is known, $z_1 \sim b(y_1, \frac{1-\theta}{2-\theta})$ and $z_2 \sim b(y_3, \frac{\theta}{1-\theta})$.

# EM Algorithm

## Example (Con'd)

We use **Expectation-Maximization (EM)** Algorithm to find the solution.

- **E** step: Given the observed data $\boldsymbol{y}$ and the $i$th estimate $\theta = \theta^{(i)}$, find the expectation of the log-likelihood function, that is,

$$Q(\theta|\boldsymbol{y}, \theta^{(i)}) = E_{\boldsymbol{z}}\left(l(\theta; \boldsymbol{y}, \boldsymbol{z})\right)$$

- **M** step: Find the maximizer of $Q(\theta|\boldsymbol{y}, \theta^{(i)})$, that is

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\boldsymbol{y}, \theta^{(i)})$$

# EM Algorithm

## Example (Con'd)

In this example,

- **E** step:

$$Q(\theta|\boldsymbol{y}, \theta^{(i)}) = \left(E(z_2|\boldsymbol{y}, \theta^{(i)}) + y_4\right) \ln\theta + \left(E(z_1|\boldsymbol{y}, \theta^{(i)}) + y_2\right) \ln(1-\theta)$$

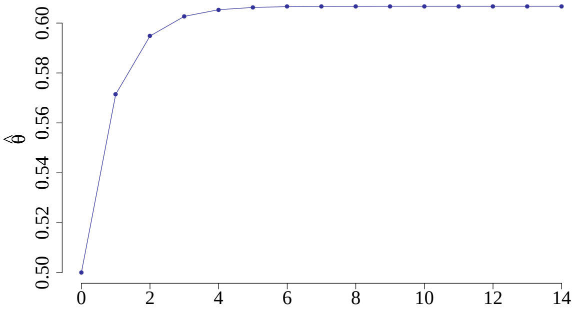
- **M** step: Let the first-order deviate be zero.

$$\frac{\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3 + y_4}{\theta^{(i+1)}} + \frac{\frac{1-\theta^{(i)}}{2-\theta^{(i)}} y_1 + y_2}{1-\theta^{(i+1)}} = 0$$

Thus, the iterative formula is

$$\theta^{(i+1)} = \frac{\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3 + y_4}{\frac{\theta^{(i)}}{1+\theta^{(i)}} y_3 + y_4 + \frac{1-\theta^{(i)}}{2-\theta^{(i)}} y_1 + y_2}$$

## Example (Con'd)




The result is $\hat{\theta} = 0.6067466$.

# Unbiasedness

## Definition

Suppose that $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ is an estimate of $\theta$ and $\Theta$ is the parameter space of $\theta$.

- The **bias** of an estimate $\hat{\theta}$ of the parameter $\theta$ is the difference between the expectation of $\hat{\theta}$ and $\theta$, that is $E(\hat{\theta}) - \theta$;
- The estimate $\hat{\theta}$ is **unbiased** for $\theta$ if

  $$E(\hat{\theta}) = \theta,$$

  for any $\theta \in \Theta$;
- The estimate $\hat{\theta}$ is **asymptotically unbiased** for $\theta$ if

  $$E(\hat{\theta}) \to \theta \quad \text{as} \quad n \to \infty,$$

  for any $\theta \in \Theta$.

# Unbiasedness

## Example

For an unknown population, $\mu$ is the expectation/population mean, $\sigma^2$ is the variance and $\mu_k$ is the $k$th moment. Suppose that $x_1, x_2, \cdots, x_n$ is a sample. We have

- The sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}$ is unbiased for $\mu$;
- The $k$th sample moment $\hat{\mu}_k = a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ is unbiased for $\mu_k$;
- The sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is unbiased for $\sigma^2$;
- The sample variance $s_*^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is asymptotically unbiased for $\sigma^2$ since

$$E(s_*^2) = \frac{n-1}{n}\sigma^2 \to \sigma^2 \quad \text{as} \quad n \to \infty.$$

# Unbiasedness

Is $g(\hat{\theta})$ an unbiased estimate of $g(\theta)$, for any function $g(\cdot)$, if $\hat{\theta}$ is an unbiased estimate of $\theta$?

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from $N(\mu, \sigma^2)$. It is well-known that $s^2$ is an unbiased estimate of $\sigma^2$. We wonder whether $s$ is an unbiased estimate of $\sigma$ or not.
We know

$$Y = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1),$$

and the p.d.f. is

$$f_Y(y) = \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} y^{\frac{n-1}{2}-1} e^{-\frac{y}{2}}, y > 0$$

## Example (Con'd)

Thus,

$$\begin{aligned} E(Y^{1/2}) &= \int_0^\infty y^{1/2} f_Y(y)\mathrm{d}y \\ &= \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}\int_0^\infty y^{\frac{n}{2}-1}e^{-\frac{y}{2}}\mathrm{d}y \\ &= \frac{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} = \sqrt{2}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}. \end{aligned}$$

Therefore,

$$E(s) = \frac{\sigma}{\sqrt{n-1}}E(Y^{1/2}) = \sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\sigma \overset{\text{def}}{=} \frac{\sigma}{c_n}$$

# Unbiasedness

## Remark

- $s$ is not an unbiased estimate of $\sigma$
- $c_n \cdot s$ is an unbiased estimate, where $c_n = \sqrt{\frac{n-1}{2}} \cdot \frac{\Gamma((n-1)/2)}{\Gamma(n/2)}$;
- $s$ is asymptotically unbiased for $\sigma$ since $c_n \to \infty$ as $n \to \infty$.

## Definition

If there exist an unbiased estimate $\hat{\theta}$ for a parameter $\theta$, that is,

$$E(\hat{\theta}) = \theta,$$

the parameter $\theta$ is called as **estimable**; otherwise, this parameter is **inestimable**.

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a Bernoulli distribution $B(p), 0 < p < 1$. We next explain why the parameter $\theta = \frac{1}{p}$ is inestimable.
First, $T = \sum_{i=1}^{n} x_i$ is a sufficient statistic for $p$ and $T \sim b(n, p)$. Suppose that an estimate $\hat{\theta} = \hat{\theta}(t)$ is unbiased for $\theta$. Then,

$$E(\hat{\theta}) = \sum_{i=0}^{n} \binom{n}{i} \hat{\theta}(i) p^i (1-p)^{n-i} = \frac{1}{p}$$

and equivalently,

$$\sum_{i=0}^{n} \binom{n}{i} \hat{\theta}(i) p^{i+1} (1-p)^{n-i} - 1 = 0$$

## Example (Con'd)

Let

$$g(p) = \sum_{i=0}^{n} \binom{n}{i} \hat{\theta}(i) p^{i+1} (1-p)^{n-i} - 1,$$

which is a $n+1$th order polynomial function of $p$. Then, there exist at most $n+1$ roots of $g(p)$. For any $p \in (0,1)$, it is impossible that $p$ is a root of the function $g(p)$. Therefore, $\theta = 1/p$ is inestimable.

# Efficiency

## Definition

Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimate of $\theta$.

- The **efficiency** of $\hat{\theta}_1$ **relative** to $\hat{\theta}_2$ is defined to be

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}.$$

- $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ if

$$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$$

  holds for all $\theta \in \Theta$ with strict inequality holding somewhere.

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from an unknown population with the mean $\mu$ and the variance $\sigma^2$. We know,

- The 1st estimate: $\hat{\mu}_1 = x_1$;
- The 2nd estimate: $\hat{\mu}_2 = \bar{x}$;

Since

$$E(\hat{\mu}_1) = E(x_1) = \mu \quad \text{and} \quad E(\hat{\mu}_2) = E(\bar{x}) = \mu$$

and

$$Var(\hat{\mu}_1) = Var(x_1) = \sigma^2 \quad \text{and} \quad Var(\hat{\mu}_2) = Var(\bar{x}) = \frac{\sigma^2}{n},$$

two estimates $\hat{\mu}_1$ and $\hat{\mu}_2$ are both unbiased for $\mu$ and then $\hat{\mu}_2$ is more efficient than $\hat{\mu}_1$ if $n > 1$.

# Efficiency

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a uniform $U(0, \theta)$. On one hand, we often use the MLE $x_{(n)}$ to estimate $\theta$. Since

$$E(x_{(n)}) = \frac{n}{n+1}\theta,$$

$x_{(n)}$ is not unbiased for $\theta$, but it is asymptotically unbiased for $\theta$. Then, we could obtain an unbiased estimate

$$\hat{\theta}_1 = \frac{n+1}{n}x_{(n)}$$

and

$$Var(\hat{\theta}_1) = \frac{(n+1)^2}{n^2}\frac{n}{(n+2)(n+1)^2}\theta^2 = \frac{1}{n(n+2)}\theta^2$$

## Example (Con'd)

On the other hand, we consider the method of moment estimate. Since $E(x_1) = \frac{\theta}{2}$. Another unbiased estimate of $\theta$ is $\hat{\theta}_2 = 2\bar{x}$ and

$$Var(\hat{\theta}_2) = 4Var(\bar{x}) = \frac{4}{n}Var(x_1) = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta}{3n}$$

Thus, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.

# Consistency

## Definition

Let $\hat{\theta}_n$ be an estimate of a parameter $\theta$ based on a sample of size $n$. Then $\hat{\theta}_n$ is said to be **consistent** in probability if $\hat{\theta}_n$ converges in probability to $\theta$ as $n$ approaches infinity; that is, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| \geq \epsilon) = 0.$$

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from $N(\mu, \sigma^2)$. From the Central Limit Theorem,

- $\bar{x}$ is consistent for $\mu$;
- $s_*^2$ is consistent for $\sigma^2$;
- $s^2$ is consistent for $\sigma^2$;

## Theorem One

Suppose that $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \cdots, x_n)$ is an estimate of $\theta$. If

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta \quad \text{and} \quad \lim_{n\to\infty} Var(\hat{\theta}_n) = 0,$$

then $\hat{\theta}_n$ is consistent for $\theta$.

**Proof:** For any $\epsilon > 0$, from the Chebyshev's Inequality,

$$P\left(|\hat{\theta}_n - E(\hat{\theta}_n)| \geq \frac{\epsilon}{2}\right) \leq \frac{4}{\epsilon^2} Var(\hat{\theta}_n)$$

Since $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$, when $n$ is sufficiently large, we have

$$\left| E(\hat{\theta}_n) - \theta \right| < \frac{\epsilon}{2}.$$

# Consistency

## Theorem One(Con'd)

Note that if $\left|\hat{\theta}_n - E(\hat{\theta}_n)\right| < \frac{\epsilon}{2}$, then

$$\left|\hat{\theta}_n - \theta\right| \leq \left|\hat{\theta}_n - E(\hat{\theta}_n)\right| + \left|E(\hat{\theta}_n) - \theta\right| < \epsilon.$$

Thus,

$$\left\{\left|\hat{\theta}_n - E(\hat{\theta}_n)\right| < \frac{\epsilon}{2}\right\} \subset \left\{\left|\hat{\theta}_n - \theta\right| < \epsilon\right\}.$$

Equivalently,

$$\left\{\left|\hat{\theta}_n - E(\hat{\theta}_n)\right| \geq \frac{\epsilon}{2}\right\} \supset \left\{\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right\}.$$

Therefore, as $n \to \infty$,

$$P\left(\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right) \leq P\left(\left|\hat{\theta}_n - E(\hat{\theta}_n)\right| \geq \epsilon/2\right) \leq \frac{4}{\epsilon^2} Var(\hat{\theta}_n) \to 0.$$

# Consistency

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from $U(0, \theta)$. Prove that $x_{(n)}$ is consistent for $\theta$.
**Solution:** The p.d.f. of $x_{(n)}$ is

$$f(y) = ny^{n-1}/\theta^n, y < \theta$$

Then, as $n \to \infty$, we have

$$\begin{aligned}
E(\hat{\theta}) &= \int_0^\theta ny^n \mathrm{d}y/\theta^n = \frac{n}{n+1}\theta \to \theta \\
E(\hat{\theta}^2) &= \int_0^\theta ny^{n+1} \mathrm{d}y/\theta^n = \frac{n}{n+2}\theta^2 \\
Var(\hat{\theta}) &= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 \to 0.
\end{aligned}$$

## Theorem Two

Suppose that $\hat{\theta}_{n1}, \cdots, \hat{\theta}_{nk}$ are respectively consistent for $\theta_1, \cdots, \theta_k$ and $\eta = g(\theta_1, \cdots, \theta_k)$ is a continuous function of $\theta_1, \cdots, \theta_k$. Then, $\hat{\eta}_n = g(\hat{\theta}_{n1}, \cdots, \hat{\theta}_{nk})$ is consistent for $\eta$.

**Proof:** Since the function $g(\cdot)$ is continuous, for any $\epsilon > 0$ and some $\delta > 0$, when $|\hat{\theta}_j - \theta_j| < \delta, j = 1, 2, \cdots, k$, we have

$$\left| g(\hat{\theta}_{n1}, \cdots, \hat{\theta}_{nk}) \right| < \epsilon. \tag{1}$$

Since $\hat{\theta}_{n1}, \cdots, \hat{\theta}_{nk}$ are consistent, for the $\delta > 0$ and any $\nu > 0$, there exists a positive integer $N$ such that

$$P\Big( |\hat{\theta}_{nj} - \theta_j| \geq \delta \Big) < \frac{\nu}{k}, j = 1, 2, \cdots, k.$$

## Theorem Two (Con'd)

Then,

$$\begin{aligned} P\Big(\cap_{j=1}^{k}\Big\{|\hat{\theta}_{nj}-\theta_j|<\delta\Big\}\Big) &= 1-P\Big(\cup_{j=1}^{k}\Big\{|\hat{\theta}_{nj}-\theta_j|\geq\delta\Big\}\Big) \\ &\geq 1-\sum_{j=1}^{k}P\Big(|\hat{\theta}_{nj}-\theta_j|\geq\delta\Big) \\ &> 1-k\cdot\frac{\nu}{k}=1-\nu \end{aligned}$$

According to the equation (1), we have

$$\cap_{j=1}^{k}\Big\{|\hat{\theta}_{nj}-\theta_j|<\delta\Big\}\subset\{|\hat{\eta}_n-\eta|<\epsilon\}$$

Then, $P(|\hat{\eta}_n-\eta|<\epsilon)>1-\nu$. Since $\nu$ is arbitrary, it is proved.

# Consistency

## Remark

From the CLT,

- The method of moment estimate is consistent;
- The sample mean is consistent for the population mean;
- The sample standard deviation is consistent for the standard deviation;
- The sample coefficient of variation is consistent for the coefficient of variation;

# Asymptotic Normality

## Definition

Suppose that an estimate $\hat{\theta}_n$ is consistent for the parameter $\theta$. The estimate $\hat{\theta}_n$ is called to be **asymptotically normal** if there exists such a sequence of non-negative constants $\sigma_n(\theta)$ which approaches to zero that

$$\frac{\hat{\theta}_n - \theta}{\sigma_n(\theta)} \xrightarrow{L} N(0,1).$$

Then, $\hat{\theta}_n$ is also called to be distributed as **asymptotic normal distribution** $N(\theta, \sigma_n(\theta))$. That is, $\hat{\theta}_n \sim AN(\theta, \sigma_n^2(\theta))$, where $\sigma_n^2(\theta)$ is the asymptotic variance of $\hat{\theta}_n$.

# Asymptotic Normality

## Example: Poisson distribution

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a Poisson distribution $P(\lambda)$.

- The method of moment estimate of $\lambda$?
- The maximum likelihood estimate of $\lambda$?

The estimate is

$$\hat{\lambda}_n = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

From CLT,

$$\frac{\hat{\lambda}_n - \lambda}{\sqrt{\lambda/n}} \xrightarrow{L} N(0,1)$$

The asymptotic distribution of $\hat{\lambda}_n$ is $AN(\lambda, \lambda/n)$.

# Asymptotic Normality

## Example: Poisson distribution

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from a Poisson distribution $P(\lambda)$.

- The method of moment estimate of $\lambda$?
- The maximum likelihood estimate of $\lambda$?

The estimate is

$$\hat{\lambda}_n = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

From CLT,

$$\frac{\hat{\lambda}_n - \lambda}{\sqrt{\lambda/n}} \xrightarrow{L} N(0,1)$$

The asymptotic distribution of $\hat{\lambda}_n$ is $AN(\lambda, \lambda/n)$.

# Asymptotic Normality

## Definition: Fisher Information

Suppose that the p.d.f. of a random variable $X$ is $p(x;\theta), \theta \in \Theta$. The following conditions are satisfied:

- The sample space $\Theta$ is an open interval;
- The support $S = \{x : p(x;\theta) > 0\}$ is not related to $\theta$;
- The deviate $\frac{\partial}{\partial\theta} p(x;\theta)$ exists for all the $\theta \in \Theta$;
- For $p(x;\theta)$,
$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} p(x;\theta)\mathrm{d}x = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta} p(x;\theta)\mathrm{d}x$$
- $E\left(\frac{\partial}{\partial\theta}\ln p(x;\theta)\right)^2$ exist.

Then,

$$I(\theta) = \int_{-\infty}^{\infty}\left(\frac{\partial \ln p(x;\theta)}{\partial\theta}\right)^2 p(x;\theta)\mathrm{d}x = E\left(\frac{\partial}{\partial\theta}\ln p(x;\theta)\right)^2$$

is said to be **Fisher Information**.

# Asymptotic Normality

## Theorem

Suppose that the p.d.f. of a population $X$ is $p(x;\theta)$, $\theta \in \Theta$, where $\Theta$ is a non-degenerate interval. Suppose that

- For any $x$, there exist the partials $\frac{\partial \ln p}{\partial \theta}, \frac{\partial^2 \ln p}{\partial \theta^2}, \frac{\partial^3 \ln p}{\partial \theta^3}$ for all the $\theta \in \Theta$;
- For all the $\theta \in \Theta$, there exist some functions $F_1(x), F_2(x)$ and $F_3(x)$ such that

$$\left|\frac{\partial p}{\partial \theta}\right| < F_1(x), \left|\frac{\partial^2 p}{\partial \theta^2}\right| < F_2(x), \left|\frac{\partial^3 \ln p}{\partial \theta^3}\right| < F_3(x),$$

  where $\int_{-\infty}^{\infty} F_1(x)\mathrm{d}x < \infty,\ \int_{-\infty}^{\infty} F_2(x)\mathrm{d}x < \infty,\ \sup_{\theta\in\Theta} \int_{-\infty}^{\infty} F_3(x)p(x;\theta)\mathrm{d}x < \infty$.
- For all the $\theta \in \Theta$, $0 < I(\theta) < \infty$.

Suppose that $x_1, x_2, \cdots, x_n$ is a sample of the population. Then, there exists the MLE of $\theta$, denoted as $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \cdots, x_n)$, and $\hat{\theta}_n$ is consistent and asymptotically normal, i.e. $\hat{\theta}_n \sim AN\left(\theta, \frac{1}{nI(\theta)}\right)$.

# Asymptotic Normality

## Revisit Example: Poisson Example

Suppose that a random variable $X$ is distributed as a Poisson distribution $P(\lambda)$ with the p.m.f.

$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}, x = 0, 1, 2, \cdots .$$

Then,

$$\ln f(x;\lambda) = x\ln\lambda - \lambda - \ln(x!), \quad \text{and} \quad \frac{\partial}{\partial\lambda}\ln f(x;\lambda) = \frac{x}{\lambda} - 1.$$

Thus,

$$I(\lambda) = E\left(\frac{x-\lambda}{\lambda}\right)^2 = \frac{1}{\lambda}.$$

# Example

A trial has three outcomes with the probabilities are respectively

$$p_1 = \theta^2 \quad p_2 = 2\theta(1-\theta) \quad p_3 = (1-\theta)^2$$

The trials are conducted $n$ times and these three outcomes are respectively $n_1, n_2$ and $n_3$.

- The likelihood function is

  $$L(\theta) \propto (\theta^2)^{n_1}(2\theta(1-\theta))^{n_2}(1-\theta)^{2n_3} = 2^{n_2}\theta^{2n_1+n_2}(1-\theta)^{2n_3+n_2}$$

  The MLE is

  $$\hat{\theta}_{\mathsf{MLE}} = \frac{2n_1+n_2}{2n}$$

  which is also consistent for $\theta$.

# Example (Con'd)

A trial has three outcomes with the probabilities are respectively

$$p_1 = \theta^2 \quad p_2 = 2\theta(1-\theta) \quad p_3 = (1-\theta)^2$$

The trials are conducted $n$ times and these three outcomes are respectively $n_1, n_2$ and $n_3$.

- The method of moment estimates are consistent for $\theta$.

$$\hat{\theta}_1 = \sqrt{n_1/n}, \quad \hat{\theta}_2 = 1 - \sqrt{n_3/n}, \quad \hat{\theta}_3 = (n_1 + n_2/2)/n$$

since

$$\theta = \sqrt{p_1}, \quad \theta = 1 - \sqrt{p_3}, \quad \theta = p_1 + p_2/2$$

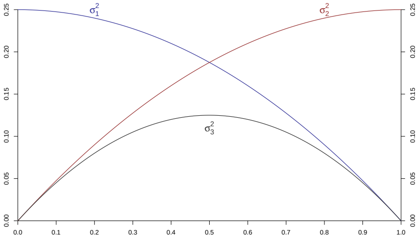
# Example (Con'd)

All these method of moment estimates are asymptotically normal, that is

$$\frac{\sqrt{n}(\hat{\theta}_i - \theta)}{\sigma_i(\theta)} \xrightarrow{\mathsf{L}} N(0,1), i = 1, 2, 3.$$

where

$$\sigma_1^2(\theta) = \frac{1-\theta^2}{4}, \quad \sigma_2^2(\theta) = \frac{1-(1-\theta)^2}{4}, \quad \sigma_3^2(\theta) = \frac{\theta(1-\theta)}{2}$$



# Mean Squared Error

## Definition

The **Mean Squared Error** of $\hat{\theta}$ as an estimate of $\theta$ is

$$\mathsf{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

## Remark

- $\mathsf{MSE} = \mathsf{Variance} + \mathsf{Bias}^2$. Equivalently,

$$\begin{aligned}
& \mathsf{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E((\hat{\theta} - E\hat{\theta}) - (E\hat{\theta} - \theta))^2 \\
= \; & E(\hat{\theta} - E\hat{\theta})^2 + E(E\hat{\theta} - \theta)^2 - 2E((\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta)) \\
= \; & E(\hat{\theta} - E\hat{\theta})^2 + E(E\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E\hat{\theta} - \theta)^2
\end{aligned}$$

- If $\hat{\theta}$ is an unbiased estimate of $\theta$, then $\mathsf{MSE}(\hat{\theta}) = Var(\hat{\theta})$.

# Mean Squared Error

## Example

Suppose that the population is $U(0,\theta)$ and $x_1, x_2, \cdots, x_n$ is a sample. Find an estimate with minimum MSE.

- The maximum likelihood estimate of $\theta$ is $\hat{\theta}_{\text{MLE}} = x_{(n)}$;
- Based on the MLE, an unbiased estimate is $\hat{\theta}_1 = \frac{n+1}{n}x_{(n)}$ and the mean squared error of $\hat{\theta}_1$ is

$$\text{MSE}(\hat{\theta}_1) = Var(\hat{\theta}_1) = \frac{\theta^2}{n(n+2)}.$$

- Consider an estimate $\hat{\theta}_\alpha = \alpha x_{(n)}$. The mean squared error of $\hat{\theta}_\alpha$ is

$$\text{MSE}(\hat{\theta}_\alpha) = Var(\alpha x_{(n)}) + (\alpha E x_{(n)} - \theta)^2$$

# Mean Squared Error

## Example (Con'd)

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}_\alpha) &= Var(\alpha x_{(n)}) + (\alpha E x_{(n)} - \theta)^2 \\
&= \alpha^2 \cdot \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(\alpha \cdot \frac{n}{n+1}\theta - \theta\right)^2 \\
&= \frac{\alpha^2 n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{\alpha n}{n+1} - 1\right)^2 \theta^2
\end{aligned}
$$

Then,

$$\alpha_0 = \arg\min_\alpha \mathrm{MSE}(\hat{\theta}_\alpha) = \frac{n+2}{n+1}.$$

and let $\hat{\theta}_0 = \frac{n+2}{n+1} x_{(n)}$ with the mean squared error

$$\mathrm{MSE}(\hat{\theta}_0) = \frac{\theta^2}{(n+1)^2} < \frac{\theta^2}{n(n+2)} = \mathrm{MSE}(\hat{\theta}_1)$$

# Mean Squared Error

## Remark

- Indeed, there is no one "best MSE" estimate.
- For example, for a certain $\theta_0 \in \Theta$, the estimate $\hat{\theta} = \theta_0$ cannot be beaten in MSE at $\theta = \theta_0$ but is a terrible estimator otherwise;
- One way to make the problem of finding a "best" estimate tractable is to **limit the class of estimates**;
- A popular way of restricting the class of estimates is to consider only **unbiased estimates**.

# UMVUE

## Definition

An estimator $\hat{\theta}$ is a **best unbiased estimate** of $\theta$ if

- it satisfies
$$E(\hat{\theta}) = \theta$$
for all $\theta$;
- for any other unbiased estimate $\tilde{\theta}$ with $E(\tilde{\theta}) = \theta$, we have
$$Var(\hat{\theta}) \leq Var(\tilde{\theta})$$
for all $\theta$.

The estimator $\hat{\theta}$ is also called a **uniform minimum variance unbiased estimate, UMVUE** of $\theta$.

# UMVUE

## Theorem

Suppose that $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ is a sample. Let $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ be an unbiased estimate with $Var(\hat{\theta}) < \infty$. Then, the necessary and sufficient condition for $\hat{\theta}$ to be a UMVUE of $\theta$ is, for every estimate $\varphi(\boldsymbol{x})$ with $E(\varphi(\boldsymbol{x})) = 0$ and $Var(\varphi(\boldsymbol{x})) < \infty$, we have

$$Cov(\hat{\theta}, \varphi) = 0, \quad \forall \theta \in \Theta.$$

**Proof:** $(\Leftarrow)$ For an unbiased estimate $\tilde{\theta}$, let $\varphi = \tilde{\theta} - \hat{\theta}$. Then,

$$E(\varphi) = E(\tilde{\theta}) - E(\hat{\theta}) = 0$$

Thus,

$$\begin{aligned} Var(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 = E\left((\tilde{\theta} - \hat{\theta}) + (\hat{\theta} - \theta)\right)^2 \\ &= E(\varphi^2) + Var(\hat{\theta}) + 2Cov(\varphi, \hat{\theta}) \geq Var(\hat{\theta}). \end{aligned}$$

# UMVUE

## Theorem

Suppose that $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ is a sample. Let $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ be an unbiased estimate with $Var(\hat{\theta}) < \infty$. Then, the necessary and sufficient condition for $\hat{\theta}$ to be a UMVUE of $\theta$ is, for every estimate $\varphi(\boldsymbol{x})$ with $E(\varphi(\boldsymbol{x})) = 0$ and $Var(\varphi(\boldsymbol{x})) < \infty$, we have

$$Cov(\hat{\theta}, \varphi) = 0, \quad \forall \theta \in \Theta.$$

**Proof:** $(\Leftarrow)$ For an unbiased estimate $\tilde{\theta}$, let $\varphi = \tilde{\theta} - \hat{\theta}$. Then,

$$E(\varphi) = E(\tilde{\theta}) - E(\hat{\theta}) = 0$$

Thus,

$$\begin{aligned} Var(\tilde{\theta}) &= E(\tilde{\theta} - \theta)^2 = E\left((\tilde{\theta} - \hat{\theta}) + (\hat{\theta} - \theta)\right)^2 \\ &= E(\varphi^2) + Var(\hat{\theta}) + 2Cov(\varphi, \hat{\theta}) \geq Var(\hat{\theta}). \end{aligned}$$

# UMVUE

## Theorem (Con'd)

($\Rightarrow$) We show the proof by contradiction. Suppose that $\hat{\theta}$ is a UMVUE of $\theta$ and $\varphi(\boldsymbol{x})$ is an unbiased estimate of $0$ with $Var(\varphi(x)) < \infty$. If there exists such $\theta_0 \in \Theta$ that

$$Cov_{\theta_0}(\hat{\theta}, \varphi(\boldsymbol{x})) \stackrel{\text{def}}{=} a \neq 0.$$

- Let $b = -\frac{a}{Var_{\theta_0}(\varphi(\boldsymbol{x}))}$. Then,

$$b^2 Var_{\theta_0}(\varphi(\boldsymbol{x})) + 2ab = b(-a+2a) = ab = -\frac{a^2}{Var_{\theta_0}(\varphi(\boldsymbol{x}))} < 0$$

Let $\tilde{\theta} = \hat{\theta} + b\varphi(\boldsymbol{x})$. Then $E(\tilde{\theta}) = E(\hat{\theta}) + bE(\varphi(\boldsymbol{x})) = \theta$. This means that $\tilde{\theta}$ is an unbiased estimate of $\theta$.

# UMVUE

## Theorem (Con'd)

- The variance is

$$\begin{aligned} Var_{\theta_0}(\tilde{\theta}) &= E_{\theta_0}(\hat{\theta} + b\varphi(\boldsymbol{x}) - \theta)^2 \\ &= E_{\theta_0}(\hat{\theta} - \theta)^2 + b^2 E_{\theta_0}(\varphi^2(\boldsymbol{x})) \\ &\quad +2bE_{\theta_0}((\hat{\theta} - \theta)\varphi(\boldsymbol{x})) \\ &= Var_{\theta_0}(\hat{\theta}) + b^2 Var_{\theta_0}(\varphi(\boldsymbol{x})) + 2ab \\ &< 0. \end{aligned}$$

Thus, $\hat{\theta}$ is not a UMVUE of $\theta$. It is proved that

$$Cov(\hat{\theta}, \varphi(\boldsymbol{x})) = 0$$

for all $\theta \in \Theta$.

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from an exponential distribution $Exp(1/\theta)$. From the factorization theorem,

$$T = x_1 + x_2 + \cdots + x_n$$

is a sufficient statistic for $\theta$. Since $E(T) = n\theta$,

$$\bar{x} = \frac{T}{n}$$

is an unbiased for $\theta$. Suppose that $\varphi = \varphi(x_1, x_2, \cdots, x_n)$ is an unbiased estimate of $0$. Then,

$$E\varphi = \int_0^\infty \cdots \int_0^\infty \varphi(x_1, x_2, \cdots, x_n) \prod_{i=1}^n \left\{ \frac{1}{\theta} e^{-x_i/\theta} \right\} \mathrm{d}x_1 \cdots \mathrm{d}x_n = 0$$

# UMVUE

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from an exponential distribution $Exp(1/\theta)$. From the factorization theorem,

$$T = x_1 + x_2 + \cdots + x_n$$

is a sufficient statistic for $\theta$. Since $E(T) = n\theta$,

$$\bar{x} = \frac{T}{n}$$

is an unbiased for $\theta$. Suppose that $\varphi = \varphi(x_1, x_2, \cdots, x_n)$ is an unbiased estimate of $0$. Then,

$$E\varphi = \int_0^\infty \cdots \int_0^\infty \varphi(x_1, x_2, \cdots, x_n) \prod_{i=1}^n \left\{ \frac{1}{\theta} e^{-x_i/\theta} \right\} \mathrm{d}x_1 \cdots \mathrm{d}x_n = 0$$

## Example (Con'd)

That is,

$$E\varphi = \int_0^\infty \cdots \int_0^\infty \varphi(x_1, x_2, \cdots, x_n) e^{-(x_1+\cdots+x_n)/\theta} \mathrm{d}x_1 \cdots \mathrm{d}x_n = 0,$$

Differentiating both sides with respect to $\theta$, i.e.

$$\int_0^\infty \cdots \int_0^\infty \frac{n\bar{x}}{\theta^2} \varphi(x_1, x_2, \cdots, x_n) e^{-(x_1+\cdots+x_n)/\theta} \mathrm{d}x_1 \cdots \mathrm{d}x_n = 0.$$

This means $E(\bar{x} \cdot \varphi) = 0$. Thus,

$$Cov(\bar{x}, \varphi) = E(\bar{x} \cdot \varphi) - E(\bar{x})E(\varphi) = 0 - 0 = 0$$

Therefore, $\bar{x}$ is a UMVUE of $\theta$.

# UMVUE

## Theorem (Rao-Blackwell Theorem)

Suppose that $f(x;\theta)$ is the population density function and $x_1, x_2, \cdots, x_n$ is a sample. Let $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ be any unbiased estimate of $\theta$, and let $T = T(x_1, x_2, \cdots, x_n)$ be a sufficient statistic for $\theta$. Define $\tilde{\theta} = E(\hat{\theta}|T)$. Then $E(\tilde{\theta}) = \theta$ and

$$Var(\tilde{\theta}) \leq Var(\hat{\theta})$$

for all $\theta$.

**Proof:** We follow the three steps:

- $\tilde{\theta} = E(\hat{\theta}|T)$ **is an estimate of** $\theta$. Since $T = T(x_1, x_2, \cdots, x_n)$ is a sufficient statistic, $\tilde{\theta} = E(\hat{\theta}|T)$ does not depend on $\theta$.
- $\tilde{\theta}$ **is unbiased for** $\theta$. From the property of iterated conditional expectation,

$$E(\tilde{\theta}) = E(E(\hat{\theta}|T)) = E(\hat{\theta}) = \theta.$$

## Theorem (Rao-Blackwell Theorem)

Suppose that $f(x;\theta)$ is the population density function and $x_1, x_2, \cdots, x_n$ is a sample. Let $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ be any unbiased estimate of $\theta$, and let $T = T(x_1, x_2, \cdots, x_n)$ be a sufficient statistic for $\theta$. Define $\tilde{\theta} = E(\hat{\theta}|T)$. Then $E(\tilde{\theta}) = \theta$ and

$$Var(\tilde{\theta}) \leq Var(\hat{\theta})$$

for all $\theta$.

**Proof:** We follow the three steps:

- $\tilde{\theta} = E(\hat{\theta}|T)$ **is an estimate of** $\theta$**.** Since $T = T(x_1, x_2, \cdots, x_n)$ is a sufficient statistic, $\tilde{\theta} = E(\hat{\theta}|T)$ does not depend on $\theta$.
- $\tilde{\theta}$ **is unbiased for** $\theta$**.** From the property of iterated conditional expectation,

$$E(\tilde{\theta}) = E(E(\hat{\theta}|T)) = E(\hat{\theta}) = \theta.$$

## Theorem (Rao-Blackwell Theorem)

Suppose that $f(x;\theta)$ is the population density function and $x_1, x_2, \cdots, x_n$ is a sample. Let $\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$ be any unbiased estimate of $\theta$, and let $T = T(x_1, x_2, \cdots, x_n)$ be a sufficient statistic for $\theta$. Define $\tilde{\theta} = E(\hat{\theta}|T)$. Then $E(\tilde{\theta}) = \theta$ and

$$Var(\tilde{\theta}) \leq Var(\hat{\theta})$$

for all $\theta$.

**Proof:** We follow the three steps:

- $\tilde{\theta} = E(\hat{\theta}|T)$ **is an estimate of** $\theta$**.** Since $T = T(x_1, x_2, \cdots, x_n)$ is a sufficient statistic, $\tilde{\theta} = E(\hat{\theta}|T)$ does not depend on $\theta$.
- $\tilde{\theta}$ **is unbiased for** $\theta$**.** From the property of iterated conditional expectation,

$$E(\tilde{\theta}) = E(E(\hat{\theta}|T)) = E(\hat{\theta}) = \theta.$$

## Theorem (Rao-Blackwell Theorem)(Con'd)

- $Var(\tilde{\theta}) \leq Var(\hat{\theta})$. The variance is

$$\begin{aligned} Var(\hat{\theta}) &= E(\hat{\theta}-\theta)^2 = E(\hat{\theta}-\tilde{\theta}+\tilde{\theta}-\theta)^2 \\ &= E(\hat{\theta}-\tilde{\theta})^2 + E(\tilde{\theta}-\theta)^2 + 2E((\hat{\theta}-\tilde{\theta})(\tilde{\theta}-\theta)) \\ &= E(\hat{\theta}-\tilde{\theta})^2 + E(\tilde{\theta}-\theta)^2 \geq Var(\tilde{\theta}) \end{aligned}$$

where the third equality holds since

$$\begin{aligned} E((\hat{\theta}-\tilde{\theta})(\tilde{\theta}-\theta)) &= E(E((\hat{\theta}-\tilde{\theta})(\tilde{\theta}-\theta)|T)) \\ &= E((\tilde{\theta}-\theta)E((\hat{\theta}-\tilde{\theta})|T)) = 0 \end{aligned}$$

## Example

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from $B(p)$. It is known that $\bar{x}$ is a sufficient for $p$. We would like to estimate $\theta = p^2$. Let

$$\hat{\theta}_1 = \begin{cases} 1, & x_1 = 1, x_2 = 1, \\ 0, & \text{otherwise} \end{cases}$$

Since

$$E(\hat{\theta}_1) = P(x_1 = 1, x_2 = 1) = p \cdot p = \theta,$$

$\hat{\theta}_1$ is an unbiased estimate of $\theta$. However, it is not good enough because it only involves two observations. We would like to improve it.

## Example (Con'd)

Let $T = \sum_{i=1}^{n} x_i$ be a sufficient statistic. From the Rao-Blackwell Theorem,

$$
\begin{aligned}
\hat{\theta} &= E(\hat{\theta}_1 | T = t) = P(\hat{\theta}_1 = 1 | T = t) \\
&= \frac{P(x_1 = 1, x_2 = 1, T = t)}{P(T = t)} \\
&= \frac{P(x_1 = 1, x_2 = 1, \sum_{i=3}^{n} x_i = t - 2)}{P(T = t)} \\
&= \frac{p^2 \binom{n-2}{t-2} p^{t-2} (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{\binom{n-2}{t-2}}{\binom{n}{t}} = \frac{t(t-1)}{n(n-1)}.
\end{aligned}
$$

# Cramér-Rao Inequality

## Motivative

- In searching for a "best" estimate, we might ask whether **there is a lower bound for the MSE of any estimate.**
- If such a lower bound existed, it would function as a **benchmark** against which estimates could be compared.
- If an estimate achieved this lower bound, we would know that it could not be improved upon.
- In the case in which the estimate is unbiased, the Cramér-Rao inequality provides such a lower bound.

# Cramér-Rao Inequality

## Motivative

- In searching for a "best" estimate, we might ask whether **there is a lower bound for the MSE of any estimate.**
- If such a lower bound existed, it would function as a **bench-mark** against which estimates could be compared.
- If an estimate achieved this lower bound, we would know that it could not be improved upon.
- In the case in which the estimate is unbiased, the Cramér-Rao inequality provides such a lower bound.

# Cramér-Rao Inequality

## Motivative

- In searching for a "best" estimate, we might ask whether **there is a lower bound for the MSE of any estimate.**
- If such a lower bound existed, it would function as a **benchmark** against which estimates could be compared.
- If an estimate achieved this lower bound, we would know that it could not be improved upon.
- In the case in which the estimate is unbiased, the Cramér-Rao inequality provides such a lower bound.

# Cramér-Rao Inequality

## Motivative

- In searching for a "best" estimate, we might ask whether **there is a lower bound for the MSE of any estimate.**
- If such a lower bound existed, it would function as a **benchmark** against which estimates could be compared.
- If an estimate achieved this lower bound, we would know that it could not be improved upon.
- In the case in which the estimate is unbiased, the Cramér-Rao inequality provides such a lower bound.

# Cramér-Rao Inequality

## Theorem: Continuous Case

Suppose that the p.d.f. is $f(x;\theta)$ and the Fisher information $I(\theta)$ exists. Let $x_1, x_2, \cdots, x_n$ be a sample, and let $T = T(x_1, x_2, \cdots, x_n)$ be an unbiased estimate of $g(\theta)$, that is,

$$g(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i;\theta) \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

If

- the (partial) derivative $g'(\theta) = \frac{\partial g(\theta)}{\partial \theta}$ exists;
- the integration and differentiation could be interchanged:

$$g'(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(x_1, \cdots, x_n) \frac{\partial}{\partial \theta} \left( \prod_{i=1}^{n} f(x_i;\theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n;$$

# Cramér-Rao Inequality

## Theorem: Continuous Case (Con'd)

Then,

$$Var(T) \geq \frac{(g'(\theta))^2}{nI(\theta)},$$

which is said to be **Cramér-Rao Inequality**.

## Remark

- $\frac{(g'(\theta))^2}{nI(\theta)}$ is called as Cramér-Rao bound on the variance of any unbiased estimate of $g(\theta)$.
- If $Var(T) = \frac{(g'(\theta))^2}{nI(\theta)}$, $T = T(x_1, x_2, \cdots, x_n)$ is an efficient estimate of $g(\theta)$ and it is also a UMVUE.
- Special Case: If $\hat{\theta}$ is an unbiased estimate of $\theta$, then

$$Var(\hat{\theta}) \geq (nI(\theta))^{-1}.$$

# Cramér-Rao Inequality

## Theorem: Continuous Case (Con'd)

**Proof:** Since $\int_{-\infty}^{\infty} f(x_i;\theta)\mathrm{d}x_i = 1, i = 1, \cdots, n$ and take the derivative from both sides,

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(x_i;\theta)\mathrm{d}x_i \\ &= \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\theta} \ln f(x_i;\theta)\right) f(x_i;\theta)\mathrm{d}x_i \\ &= E\left(\frac{\partial}{\partial\theta} \ln f(x_i;\theta)\right), \end{aligned}$$

where the first equality holds since the integration and differentiation could be interchanged, the second equality holds since $\frac{\partial}{\partial\theta} \ln f(x_i;\theta) = (f(x_i;\theta))^{-1} \cdot \frac{\partial}{\partial\theta} f(x_i;\theta)$.

# Cramér-Rao Inequality

## Theorem: Continuous Case (Con'd)

Let $Z = \frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n} f(x_i;\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_i;\theta)$. Then,

$$E(Z) = \sum_{i=1}^{n} E\left(\frac{\partial}{\partial\theta}\ln f(x_i;\theta)\right) = 0$$

and

$$\begin{aligned} E(Z^2) &= Var(Z) = \sum_{i=1}^{n} Var\left(\frac{\partial}{\partial\theta}\ln f(x_i;\theta)\right) \\ &= \sum_{i=1}^{n} E\left(\frac{\partial}{\partial\theta}\ln f(x_i;\theta)\right)^2 = nI(\theta) \end{aligned}$$

# Cramér-Rao Inequality

## Theorem: Continuous Case (Con'd)

Thus,

$$
\begin{aligned}
g'(\theta) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(x_1, \cdots, x_n) \frac{\partial}{\partial \theta} \left( \prod_{i=1}^{n} f(x_i; \theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} T(x_1, \cdots, x_n) \left( \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f(x_i; \theta) \right) \prod_{i=1}^{n} f(x_i; \theta) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= E(T \cdot Z) \\
&= E(T \cdot Z) - E(T)E(Z) \\
&= E((T - g(\theta))Z).
\end{aligned}
$$

From the Schwarz's inequality,

$$[g'(\theta)]^2 \leq E((T - g(\theta))^2) \cdot E(Z^2) = Var(T)Var(Z).$$

# Cramér-Rao Inequality

## Remark

- Since the integration and differentiation could be interchanged, it is proved that $E\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right) = 0$. Then,

$$I(\theta) = E\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 = Var\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right).$$

- If the second derivative $\frac{\partial^2}{\partial\theta^2} f(x;\theta)$ exists for all $\theta \in \Theta$, The Fisher information is

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\right).$$

- Cramér-Rao Inequality still holds in the discrete case.

# Cramér-Rao Inequality

## Example: Bernoulli distribution

Suppose that the population is a Bernoulli distribution $B(\theta)$. The p.m.f. is

$$f(x;\theta) = \theta^x (1-\theta)^{1-x}, x = 0, 1$$

The Fisher's Information is

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

Suppose that $x_1, x_2, \cdots, x_n$ is a sample. Then, Cramér-Rao bound is $(nI(\theta))^{-1} = \theta(1-\theta)/n$. We know that $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ is unbiased for $\theta$ and its variance is $\theta(1-\theta)/n$. Since $\bar{x}$ achieves the Cramér-Rao bound, $\bar{x}$ is efficient for $\theta$ and thus it is also a UMVUE.

# Cramér-Rao Inequality

## Example: Exponential distribution

Suppose that the population is an Exponential distribution $Exp(1/\theta)$. The p.m.f. is

$$f(x;\theta) = \theta^{-1} \exp\left\{-\frac{x}{\theta}\right\}, x > 0$$

The Fisher's Information is

$$I(\theta) = \frac{1}{\theta^2}.$$

Suppose that $x_1, x_2, \cdots, x_n$ is a sample. Then, Cramér-Rao bound is $(nI(\theta))^{-1} = \theta^2/n$. We know that $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ is unbiased for $\theta$ and its variance is $\theta^2/n$. Since $\bar{x}$ achieves the Cramér-Rao bound, $\bar{x}$ is efficient for $\theta$ and thus it is also a UMVUE.

# Cramér-Rao Inequality

## Remark

- An efficient estimate is a UMVUE;
- **A UMVUE may not be an efficient estimate.**

## Example

Suppose that the population is a Normal distribution $N(0, \sigma^2)$. The p.d.f. is

$$f(x; \sigma^2) = (2\pi\sigma)^{-1/2} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}.$$

Note that $\frac{x^2}{\sigma^2} \sim \chi^2(1)$. Then,

$$I(\sigma^2) = E\left(\frac{\partial}{\partial\sigma^2} \ln f(x; \sigma^2)\right)^2 = E\left(\frac{x^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right)^2 = \frac{1}{2\sigma^4}$$

# Cramér-Rao Inequality

## Example (Con'd)

Let $x_1, x_2, \cdots, x_n$ be a sample. Then the Cramér-Rao bound on the variance of $\sigma^2$ is $\frac{2\sigma^2}{n}$. We know that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{n} x_i^2$ is unbiased for $\sigma^2$ and its variance achieves the Cramér-Rao bound. So, $\hat{\sigma}^2$ is a UMVUE of $\sigma^2$.
Let $\sigma = g(\sigma^2) = \sqrt{\sigma^2}$. Then, the Cramér-Rao bound on the variance of $\sigma$ is

$$\frac{g'(\sigma^2)^2}{nI(\sigma^2)} = \frac{(1/(2\sigma))^2}{n/(2\sigma^4)} = \frac{\sigma^2}{2n}$$

The unbiased estimate of $\sigma$ is

$$\hat{\sigma} = \sqrt{\frac{n}{2}} \cdot \frac{\Gamma(n/2)}{\Gamma((n+1)/2)} \sqrt{n^{-1}\sum_{i=1}^{n} x_i^2}$$

# Cramér-Rao Inequality

## Example (Con'd)

It is can be proved that

- $\hat{\sigma}$ is a UMVUE of $\sigma$;
- The variance of $\hat{\sigma}$ is larger than $\frac{\sigma^2}{2n}$. Thus, $\hat{\sigma}$ is not an efficient estimate.

# UMVUE (Con'd)

## Definition

Let $f(t;\theta)$ be p.d.f. or p.m.f for a statistic $T(\boldsymbol{x})$. $T(\boldsymbol{x})$ is called a **complete statistic** if

$$E(g(T)) = 0 \text{ for all } \theta \Rightarrow P(g(T) = 0) = 1 \text{ for all } \theta.$$

## Example

Suppose that $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ is a sample from a Bernoulli distribution $B(p)$ with an unknown parameter $p \in (0, 1)$. The sufficient statistic $T = \sum_{i=1}^{n} x_i$ for the parameter $p$. The distribution of $T$ is binomial distribution $b(n, p)$. Let $g$ be a function such that $E(g(T)) = 0$. Then,

# UMVUE (Con'd)

## Example (Con'd)

$$
\begin{aligned}
0 &= E(g(T)) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} \\
&= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t
\end{aligned}
$$

for all $p, 0 < p < 1$. The factor $(1-p)^n$ is not $0$ for any $p$ in this range. Thus it must be that

$$
0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t
$$

for all $r, 0 < r < \infty$.

# UMVUE (Con'd)

## Example (Con'd)

But the last expression is a polynomial of degree $n$ in $r$, where the coefficient of $r^t$ is $g(t)\binom{n}{t}$. For the polynomial to be $0$ for all $r$, each coefficient must be $0$. Since none of the $\binom{n}{t}$ terms is $0$, this implies that $g(t) = 0$ for $t = 0, 1, \cdots, n$. Since $T$ takes on the values $0, 1, \cdots, n$ with probability $1$, this yields that $P(g(T) = 0) = 1$ for all $p$, the desired conclusion. Hence, $T$ is a complete statistic.

# UMVUE (Con'd)

## Theorem

Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ be a sample from an exponential family with p.d.f. or p.m.f. of the form

$$f(x;\theta) = h(x)\exp\{\eta(\theta)^\tau T(x) - \zeta(\theta)\}\,.$$

The joint p.d.f. or p.m.f. of $\boldsymbol{x}$ is

$$f(\boldsymbol{x};\theta) = \prod_{i=1}^n h(x_i)\cdot\exp\left\{\eta(\theta)^\tau\sum_{i=1}^n T(x_i) - n\zeta(\theta)\right\}.$$

Thus, $\sum_{i=1}^n T(x_i)$ is a complete and sufficient statistic.

# UMVUE (Con'd)

## Remark

There are three methods to find a UMVUE of $g(\theta)$:

- If $E(T)$ is an unbiased estimate of $g(\theta)$ for all $\theta$ and $Var(T)$ achieves Cramér-Rao bound. Then $T$ is a UMVUE of $g(\theta)$;
  Note that a UMVUE may not attain Cramér-Rao bound.
- If $S$ is a complete and sufficient statistic, and $T = h(S)$ satisfies $E(T) = g(\theta)$ for all $\theta$. Then $T$ is a UMVUE of $g(\theta)$;
- If $U$ is an unbiased estimate for $g(\theta)$ and $S$ is a complete and sufficient statistic for $\theta$. Then, $T = E(U|S)$ is a UMVUE of $g(\theta)$.

# UMVUE (Con'd)

## Example: Normal Distribution

Suppose that $x_1, x_2, \cdots, x_n$ is a sample from $N(\mu, \sigma_0^2)$ where $\sigma_0^2$ is known, but $\mu$ is unknown.
First, We would liket to find an UMVUE of $\theta_1 = \mu$. Then

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is sufficient and complete statistics for $\theta_1$. Note that

$$E(\bar{x}) = \mu$$

Since it depends only on the sufficient and complete statistic, $\bar{x}$ is an UMVUE of $\theta_1$.

# UMVUE (Con'd)

## Example: Normal Distribution(Con'd)

Next, we would like to find a UMVUE of

$$\theta_2 = P(X_1 \le c) = \Phi\left(\frac{c-\mu}{\sigma}\right)$$

where $c$ is a known constant. It is a fact that $I_{(-\infty,c]}(x_1)$ is an unbiased estimate of $\theta_2$. Then, the UMVUE of $\theta_2$ is

$$\begin{aligned} & E(I_{(-\infty,c)}(x_1)|\bar{x}) = P(x_1 \le c|\bar{x}) = P(x_1 - \bar{x} \le c - \bar{x}|\bar{x}) \\ \overset{①}{=} \quad & P(x_1 - \bar{x} \le c - \bar{x}) \overset{②}{=} \Phi\left(\frac{c-\bar{x}}{\sqrt{1-1/n}}\right), \end{aligned}$$

- $\overset{①}{=}$ holds since $x_1 - \bar{x}$ and $\bar{x}$ are independent;
- $\overset{②}{=}$ holds since $x_1 - \bar{x} \sim N(0, 1-1/n)$.

# Frequentists' Idea

## Example

- A freshly minted coin has a certain probability of coming up heads if it is spun on its edge, but that probability is not necessarily equal to $\frac{1}{2}$.
- Now suppose that it is spun $n$ times and comes up heads $k$ times. What has been learned about the chance the coin comes up heads?
- How to solve this problem?
  - Let $x_1, x_2, \cdots, x_n$ **be a sample** from **a Bernoulli distribution** $B(\theta)$ with an unknown parameter $\theta$.
  - A reasonable estimate of $p$ is $\hat{\theta} = \frac{k}{n}$.
- Note that
  - The parameter $\theta$ is an unknown, but **fixed** quantity;
  - The sample $x_1, x_2, \cdots, x_n$ are i.i.d **random variables**.

# Frequentists' Idea

## Example

- A freshly minted coin has a certain probability of coming up heads if it is spun on its edge, but that probability is not necessarily equal to $\frac{1}{2}$.
- Now suppose that it is spun $n$ times and comes up heads $k$ times. What has been learned about the chance the coin comes up heads?
- How to solve this problem?
  - Let $x_1, x_2, \cdots, x_n$ **be a sample** from **a Bernoulli distribution** $B(\theta)$ with an unknown parameter $\theta$.
  - A reasonable estimate of $p$ is $\hat{\theta} = \frac{k}{n}$.
- Note that
  - The parameter $\theta$ is an unknown, but **fixed** quantity;
  - The sample $x_1, x_2, \cdots, x_n$ are i.i.d **random variables**.

# Bayesian Approach

In the Bayesian approach,

- $\theta$ is considered to be a **random variable**, which can be described by a probability distribution (called the **prior distribution**).
- A sample is taken from a population indexed by $\theta$;
- Given a prior probability about a hypothesis and the observed information, the Bayes rule is used to obtain the **posterior probability** which is the conditional probability on the observed evidence.

## Review: Bayes rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# Bayesian Approach

## Basic concepts

Given a hypothesis (denoted as $H$), the Bayes theorem is used to update our beliefs about it once the data (denoted as $D$) have been observed:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

where

- $P(H)$: is the **prior probability** on the hypothesis;
- $P(D|H)$: is **likelihood** given that hypothesis is true;
- $P(D)$ is the **marginal likelihood**;
- $P(H|D)$ is the **posterior probability** for $H$ once the data have been observed.

# Bayesian Approach

## Main steps:

- Select a **prior distribution** for $\theta$, denoted as $\pi(\theta), \theta \in \Theta$;
- In the Bayesian view, generate the **sample** $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ in the following two steps:
  - Generate a parameter sample $\theta$ from the prior distribution $\pi(\theta)$, that is, $\theta \sim \pi(\theta)$;
  - Given the parameter $\theta$, generate a sample from $f(\boldsymbol{x}|\theta)$.

$$f(\boldsymbol{x}|\theta) = f(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- Using Bayes' rule, obtain the **posterior probability** of $\theta$:

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int_\Theta f(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta}$$

# Bayesian Approach

## Remark

- The marginal likelihood

$$f(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta$$

  is a constant to ensure that the posterior distribution of $\theta$ integrates up to $1$ and does not depend on $\theta$.
- The posterior distribution is usually expressed by

$$\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)\pi(\theta)$$

  which means that the posterior distribution is proportional to the likelihood times the prior distribution.

# Bayesian Approach

## Example

A freshly minted coin has a certain probability of coming up heads if it is spun on its edge, but that probability is not necessarily equal to $\frac{1}{2}$. Now suppose that it is spun $n$ times and comes up heads $k$ times. What has been learned about the chance the coin comes up heads?

**Solution:** Let $\theta$ be the chance that the coin comes up heads.

- First, we assume the prior distribution on $\theta$ is a Beta distribution with two **hyperparameters** $\alpha$ and $\beta$, i.e. $p \sim Be(\alpha, \beta)$.
- Second, let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ be a sample and $k = \sum_{i=1}^{n} x_i$. It is obvious that $k|\theta \sim b(n, \theta)$.

## Example (Con'd)

- Third, the joint distribution of $k$ and $\theta$ is

$$\begin{aligned} f(k,\theta) &= f(k|\theta)\pi(\theta) \\ &= \binom{n}{k}\theta^k(1-\theta)^{n-k}\cdot\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1} \end{aligned}$$

- Forth, the marginal p.m.f. of $k$ is

$$f(y) = \int_0^1 f(y,\theta)\mathrm{d}\theta = \binom{n}{k}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)}$$

## Example (Con'd)

- Fifth, the posterior distribution, the distribution of $\theta$ given $k$, is

$$\begin{aligned}\pi(\theta|k) &= \frac{f(k,\theta)}{f(y)} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(k+\alpha)\Gamma(n-k+\beta)}\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}\end{aligned}$$

which is also distributed as a Beta distribution. Specifically, $\theta|k \sim Be(k+\alpha, n-k+\beta)$.

# Bayesian Approach

## Definition

If the posterior distributions $\pi(\theta|\boldsymbol{x})$ are in the same probability distribution family as the prior probability distribution $\pi(\theta)$, then

- the prior and posterior are said to be **conjugate distribution**,
- the prior is said to be a **conjugate prior** for the likelihood function.

In the previous example,

| Prior | Likelihood | Posterior |
|---|---|---|
| Beta | binomial | Beta |

# Loss function

## Introduction

- After the data $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ are observed, a decision regarding $\theta$ is made. The set of allowable decisions is the action space, denoted as $\mathcal{A}$.
- The **loss function** in a point estimation problem reflects the fact that
  - if an action $a$ is close to $\theta$, then the decision $a$ is reasonable and little loss is incurred;
  - f $a$ is far from $\theta$, then a large loss is incurred.
- The loss function is a non-negative function that generally increases as the distance between $a$ and $\theta$ increases. For example, a common used loss function is

$$L(\theta, a) = (a - \theta)^2.$$

## Introduction (Con'd)

- In a loss function or decision theoretic analysis, the quality of an estimate is quantified in its **risk function**; that is, for an estimate $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ of $\theta$, the **risk function**, a function of $\theta$, is

  $$R(\theta, \hat{\theta}) = E\left(L(\theta, \hat{\theta}(\boldsymbol{x}))\right)$$

  At a given $\theta$, the risk function is the average loss that will be incurred if the estimate $\hat{\theta}$ is used.
- Since the true value of $\theta$ is unknown, we would like to use an estimate that has a small value of $R(\theta, \hat{\theta})$ for all values of $\theta$. This would mean that, regardless of the true value of $\theta$, the estimate will have a small expected loss.

# Bayes risk

## Introduction

- In Bayesian view, we would use this prior distribution to compute an average risk

$$\int_{\Theta} R(\theta, \hat{\theta})\pi(\theta)\mathrm{d}\theta$$

  known as the **Bayes risk**.

- Averaging the risk function gives us one number of assessing the performance of an estimate with respect to a given loss function.
- Moreover, we can attempt to find the estimate that yields the smallest value of the Bayes risk. Such an estimate is said to be the **Bayes rule with respect to a prior $\pi$**.

## Introduction (Con'd)

For $\boldsymbol{x} \sim f(\boldsymbol{x}|\theta)$ and $\theta \sim \pi(\theta)$, the Bayes risk of a decision rule $\hat{\theta}$ can be written as

$$\int_{\Theta} R(\theta, \hat{\theta})\pi(\theta)\mathrm{d}\theta = \int_{\Theta} \left( \int_{\mathcal{X}} L(\theta, \hat{\theta}(\boldsymbol{x}))f(\boldsymbol{x}|\theta)\mathrm{d}\boldsymbol{x} \right) \pi(\theta)\mathrm{d}\theta$$

Let $f(\boldsymbol{x}|\theta)\pi(\theta) = \pi(\theta|\boldsymbol{\theta})f(\boldsymbol{x})$, where $\pi(\theta|\boldsymbol{x})$ is the posterior distribution of $\theta$ and $f(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{x}$, we can write the Bayes risk as

$$\int_{\Theta} R(\theta, \hat{\theta})\pi(\theta)\mathrm{d}\theta = \int_{\mathcal{X}} \left( \int_{\Theta} L(\theta, \hat{\theta}(\boldsymbol{x}))\pi(\theta|\theta)\mathrm{d}\theta \right) f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

which the quantity in square brackets is said to be the **posterior expected loss**.

## Introduction (Con'd)

Particularly, the loss function of $\theta$ could be written as

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2.$$

which is said to be **quadratic loss or squared error loss**.
In Frequentists' view, the risk function becomes the mean squared error of the estimate, that is,

$$\hat{\theta} = \arg\min_{\hat{\theta}} E(\theta - \hat{\theta})^2$$

The Bayes estimate is the expectation of the posterior distribution of $\theta$, that is,

$$E(\theta|\boldsymbol{x}) = \int_{\Theta} \theta \pi(\theta|\boldsymbol{x}) \mathrm{d}\theta$$

# Bayes Approach

## Example: Revisit

A freshly minted coin has a certain probability of coming up heads if it is spun on its edge, but that probability is not necessarily equal to $\frac{1}{2}$. Now suppose that it is spun $n$ times and comes up heads $k$ times. What has been learned about the chance the coin comes up head?

Let $\theta$ be the chance that coin comes up heads. Assume the prior distribution of $\theta$ is $Be(\alpha, \beta)$. The posterior distribution of $\theta$ given $k$ is $Be(k + a, n - k + \beta)$.

The Bayes estimate of $\theta$ is

$$\hat{\theta}|k = \frac{k + \alpha}{n + \alpha + \beta}.$$

# Bayes Approach

## Example: Normal distribution

Suppose that $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ is sample from $N(\mu, \sigma_0^2)$ where $\sigma_0^2$ is known but $\mu$ is unknown. Assume that the prior distribution of $\mu$ is $N(\theta, \tau^2)$, where $\theta$ and $\tau^2$ are both known. Find the Bayes estimate of $\mu$.

**Solution:** The prior of $\mu$ is

$$\pi(\mu) = (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\tau^2}(\mu-\theta)^2\right\}$$

and the sample distribution of $\boldsymbol{x}$ is

$$f(\boldsymbol{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

## Example: Normal distribution (Con'd)

The posterior distribution of $\mu$ is

$$\begin{aligned}\pi(\mu|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\mu)\pi(\mu)\\ &\propto \exp\left\{-\frac{1}{2}(A\mu^2 - B\mu + C)\right\}\\ &\propto \exp\left\{-\frac{1}{2}\frac{(\mu - B/A)^2}{2/A}\right\}\end{aligned}$$

where $A = \frac{n}{\sigma_0^2} + \frac{1}{\tau^2}$, $B = \frac{n\bar{x}}{\sigma_0^2} + \frac{\theta}{\tau^2}$ and $C = \frac{\sum_{i=1}^n x_i^2}{\sigma_0^2} + \frac{\theta^2}{\tau^2}$. Given $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, the posterior of $\mu$ is distributed as

$$\mu|\boldsymbol{x} \sim N\left(\frac{n\bar{x}{\sigma_0}^{-2} + \theta\tau^{-2}}{n{\sigma_0}^{-2} + \tau^{-2}}, \frac{1}{n{\sigma_0}^{-2} + \tau^{-2}}\right)$$

## Example: Normal distribution (Con'd)

The Bayes estimate of $\mu$ is

$$\hat{\mu} = \frac{n/\sigma_0^2}{n/\sigma_0^2 + 1/\tau^2}\bar{x} + \frac{1/\tau^2}{n/\sigma_0^2 + 1/\tau^2}\theta.$$

- It is a weighted average of the sample mean and prior mean;
- If the population variance $\sigma_0^2$ is small or the sample size $n$ is large, it is dominated by the sample mean $\bar{x}$;
- If the prior variance $\tau^2$ is small, it is dominated by the prior mean $\theta$.