# Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu
chenle02@gmail.com

Emory University
Atlanta, GA

Last updated on Spring 2021
Last compiled on January 15, 2023

2021 Spring

# Chapter 5. Estimation

# Plan

# Chapter 5. Estimation

**Motivating example**: Given an unfair coin, or $p$-coin, such that

$$X = \begin{cases} 1 & \text{head with probability } p, \\ 0 & \text{tail with probability } 1-p, \end{cases}$$

how would you determine the value $p$?

**Solutions:**

1. You need to try the coin several times, say, three times. What you obtain is "HHT".
2. Draw a conclusion from the experiment you just made.

**Motivating example**: Given an unfair coin, or $p$-coin, such that

$$X = \begin{cases} 1 & \text{head with probability } p, \\ 0 & \text{tail with probability } 1 - p, \end{cases}$$

how would you determine the value $p$?

**Solutions:**

1. You need to try the coin several times, say, three times. What you obtain is "HHT".
2. Draw a conclusion from the experiment you just made.

**Rationale:** The choice of the parameter $p$ should be the value that maximizes the probability of the sample.

$$\begin{aligned}\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0) =& P(X_1 = 1)P(X_2 = 1)P(X_3 = 0)\\ =& p^2(1-p).\end{aligned}$$

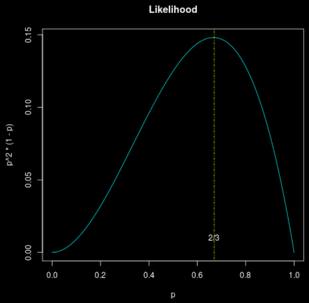
```r
# Hello, R.
p <- seq(0,1,0.01)
plot(p,p^2*(1-p),
      type="l",
      col="red")
 title("Likelihood")
# add a vertical dotted (4) blue line
abline(v=0.67, col="blue", lty=4)
# add some text
text(0.67,0.01, "2/3")
```

Maximize $f(p) = p^2(1-p)$ ....

**Rationale:** The choice of the parameter $p$ should be the value that maximizes the probability of the sample.

$$\begin{aligned}\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0) =& P(X_1 = 1)P(X_2 = 1)P(X_3 = 0)\\ =& p^2(1-p).\end{aligned}$$

```r
# Hello, R.
p <- seq(0,1,0.01)
plot(p,p^2*(1-p),
        type="l",
        col="red")
 title ("Likelihood")
# add a vertical dotted (4) blue line
abline(v=0.67, col="blue", lty=4)
# add some text
text(0.67,0.01, "2/3")
```




Maximize $f(p) = p^2(1-p)$ ....

**A random sample of size $n$ from the population – Bernoulli($p$):**

- $X_1, \cdots, X_n$ are i.i.d.[1] random variables, each following Bernoulli($p$).
- Suppose the outcomes of the random sample are: $X_1 = k_1, \cdots, X_n = k_n$.
- What is your choice of $p$ based on the above random sample?

$$p = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}.$$

[1] independent and identically distributed

**A random sample of size $n$ from the population – Bernoulli($p$):**

- $X_1, \cdots, X_n$ are i.i.d.[1] random variables, each following Bernoulli($p$).
- Suppose the outcomes of the random sample are: $X_1 = k_1, \cdots, X_n = k_n$.
- What is your choice of $p$ based on the above random sample?

$$p = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}.$$

[1] independent and identically distributed

**A random sample of size $n$ from the population – Bernoulli$(p)$:**

- $X_1, \cdots, X_n$ are i.i.d.[1] random variables, each following Bernoulli$(p)$.
- Suppose the outcomes of the random sample are: $X_1 = k_1, \cdots, X_n = k_n$.
- What is your choice of $p$ based on the above random sample?

$$p = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}.$$

[1] independent and identically distributed

**A random sample of size $n$ from the population – Bernoulli$(p)$:**

- $X_1, \cdots, X_n$ are i.i.d.[1] random variables, each following Bernoulli$(p)$.
- Suppose the outcomes of the random sample are: $X_1 = k_1, \cdots, X_n = k_n$.
- What is your choice of $p$ based on the above random sample?

$$p = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}.$$

[1] independent and identically distributed

**A random sample of size $n$ from the population with given pdf**:

- $X_1, \cdots, X_n$ are i.i.d. random variables, each following the same given pdf.

- a **statistic** or an **estimator** is a function of the random sample.

  Statistic/Estimator is a random variable!

  e.g.,

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- The outcome of a statistic/estimator is called an **estimate**. e.g.,

$$p_e = \frac{1}{n}\sum_{i=1}^{n} k_i.$$

**A random sample of size $n$ from the population with given pdf**:

- $X_1, \cdots, X_n$ are i.i.d. random variables, each following the same given pdf.

- a **statistic** or an **estimator** is a function of the random sample.

  Statistic/Estimator is a random variable!

  e.g.,

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- The outcome of a statistic/estimator is called an **estimate**. e.g.,

$$p_e = \frac{1}{n} \sum_{i=1}^{n} k_i.$$

**A random sample of size $n$ from the population with given pdf**:

- $X_1, \cdots, X_n$ are i.i.d. random variables, each following the same given pdf.

- a **statistic** or an **estimator** is a function of the random sample.

  Statistic/Estimator is a random variable!

  e.g.,

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- The outcome of a statistic/estimator is called an **estimate**. e.g.,

$$p_e = \frac{1}{n} \sum_{i=1}^{n} k_i.$$

# Plan

§ 5.1 Introduction

§ 5.2 Estimating parameters: MLE and MME

§ 5.3 Interval Estimation

§ 5.4 Properties of Estimators

§ 5.5 Minimum-Variance Estimators: The Cramér-Rao Lower Bound

§ 5.6 Sufficient Estimators

§ 5.7 Consistency

§ 5.8 Bayesian Estimation

# Chapter 5. Estimation

| **Two methods** for estimating parameters | Corresponding estimator |
| --- | --- |
| 1. Method of maximum likelihood. | MLE |
| 2. Method of moments. | MME |

| **Two methods** for estimating parameters | Corresponding estimator |
|---|---|
| 1. Method of maximum likelihood. | MLE |
| 2. Method of moments. | MME |

**Definition 5.2.1.** For a random sample of size $n$ from the discrete (resp. continuous) population/pdf $p_X(k;\theta)$ (resp. $f_Y(y;\theta)$), the likelihood function, $L(\theta)$, is the product of the pdf evaluated at $X_i = k_i$ (resp. $Y_i = y_i$), i.e.,

$$L(\theta) = \prod_{i=1}^{n} p_X(k_i;\theta) \qquad \left(\text{resp. } L(\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta)\right).$$

**Definition 5.2.2.** Let $L(\theta)$ be as defined in Definition 5.2.1. If $\theta_e$ is a value of the parameter such that $L(\theta_e) \geq L(\theta)$ for all possible values of $\theta$, then we call $\theta_e$ the maximum likelihood estimate for $\theta$.

**Definition 5.2.1.** For a random sample of size $n$ from the discrete (resp. continuous) population/pdf $p_X(k;\theta)$ (resp. $f_Y(y;\theta)$), the likelihood function, $L(\theta)$, is the product of the pdf evaluated at $X_i = k_i$ (resp. $Y_i = y_i$), i.e.,

$$L(\theta) = \prod_{i=1}^{n} p_X(k_i;\theta) \qquad \left(\text{resp. } L(\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta)\right).$$

**Definition 5.2.2.** Let $L(\theta)$ be as defined in Definition 5.2.1. If $\theta_e$ is a value of the parameter such that $L(\theta_e) \geq L(\theta)$ for all possible values of $\theta$, then we call $\theta_e$ the maximum likelihood estimate for $\theta$.

# Examples for MLE

Often but not always MLE can be obtained by setting the first derivative equal to zero:

E.g. 1. Poisson distribution: $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^{k} k_i}\left(\prod_{i=1}^{n} k_i!\right)^{-1}.$$

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} k_i\right)\ln\lambda - \ln\left(\prod_{i=1}^{n} k_i!\right).$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} k_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = 0 \quad\Longrightarrow\quad \boxed{\lambda_e = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}}.$$

Comment: The critical point is indeed global maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln L(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} k_i < 0.$$

# Examples for MLE

Often but not always MLE can be obtained by setting the first derivative equal to zero:

E.g. 1. Poisson distribution: $p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda} \lambda^{\sum_{i=1}^{k} k_i} \left( \prod_{i=1}^{n} k_i! \right)^{-1}.$$

$$\ln L(\lambda) = -n\lambda + \left( \sum_{i=1}^{n} k_i \right) \ln \lambda - \ln \left( \prod_{i=1}^{n} k_i! \right).$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} k_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{1}{n} \sum_{i=1}^{n} k_i =: \bar{k}}.$$

Comment: The critical point is indeed global maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \ln L(\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} k_i < 0.$$

# Examples for MLE

Often but not always MLE can be obtained by setting the first derivative equal to zero:

E.g. 1. Poisson distribution: $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^{k} k_i}\left(\prod_{i=1}^{n} k_i!\right)^{-1}.$$

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} k_i\right)\ln\lambda - \ln\left(\prod_{i=1}^{n} k_i!\right).$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} k_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = 0 \quad\Longrightarrow\quad \boxed{\lambda_e = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}}.$$

Comment: The critical point is indeed global maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln L(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} k_i < 0.$$

# Examples for MLE

Often but not always MLE can be obtained by setting the first derivative equal to zero:

E.g. 1. Poisson distribution: $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^{k} k_i}\left(\prod_{i=1}^{n} k_i!\right)^{-1}.$$

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} k_i\right)\ln\lambda - \ln\left(\prod_{i=1}^{n} k_i!\right).$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} k_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = 0 \quad\Longrightarrow\quad \boxed{\lambda_e = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}}.$$

Comment: The critical point is indeed global maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln L(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} k_i < 0.$$

## Examples for MLE

Often but not always MLE can be obtained by setting the first derivative equal to zero:

E.g. 1. Poisson distribution: $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^{k} k_i}\left(\prod_{i=1}^{n} k_i!\right)^{-1}.$$

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} k_i\right)\ln\lambda - \ln\left(\prod_{i=1}^{n} k_i!\right).$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} k_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}}.$$

Comment: The critical point is indeed global maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln L(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} k_i < 0.$$

# Examples for MLE

Often but not always MLE can be obtained by setting the first derivative equal to zero:

E.g. 1. Poisson distribution: $p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda} \lambda^{\sum_{i=1}^{k} k_i} \left( \prod_{i=1}^{n} k_i! \right)^{-1}.$$

$$\ln L(\lambda) = -n\lambda + \left( \sum_{i=1}^{n} k_i \right) \ln \lambda - \ln \left( \prod_{i=1}^{n} k_i! \right).$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} k_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{1}{n} \sum_{i=1}^{n} k_i =: \bar{k}}.$$

Comment: The critical point is indeed global maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \ln L(\lambda) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} k_i < 0.$$

The following two cases are related to waiting time:

E.g. 2. Exponential distribution: $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\bar{y}}}.$$

The following two cases are related to waiting time:

E.g. 2. Exponential distribution: $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\bar{y}}}.$$

The following two cases are related to waiting time:

E.g. 2. Exponential distribution: $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\bar{y}}}.$$

The following two cases are related to waiting time:

E.g. 2. Exponential distribution: $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\bar{y}}}.$$

The following two cases are related to waiting time:

E.g. 2. Exponential distribution: $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\bar{y}}}.$$

A random sample of size $n$ from the following population:

E.g. 3. Gamma distribution: $f_Y(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$ with $r > 1$ known.

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^r}{\Gamma(r)} y_i^{r-1} e^{-\lambda y_i} = \lambda^{r\,n} \Gamma(r)^{-n} \left( \prod_{i=1}^{n} y_i^{r-1} \right) \exp\left( -\lambda \sum_{i=1}^{n} y_i \right)$$

$$\ln L(\lambda) = r\,n \ln \lambda - n \ln \Gamma(r) + \ln \left( \prod_{i=1}^{n} y_i^{r-1} \right) - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{r\,n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{r\,n}{\sum_{i=1}^{n} y_i} = \frac{r}{\bar{y}}}.$$

Comment:

- When $r = 1$, this reduces to the exponential distribution case.
- If $r$ is also unknown, it will be much more complicated.
  No closed-form solution. One needs numerical solver[2].
  Try MME instead.

---

[2] [DW, Example 7.2.25]

A random sample of size $n$ from the following population:

E.g. 3. Gamma distribution: $f_Y(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$ with $r > 1$ known.

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^r}{\Gamma(r)} y_i^{r-1} e^{-\lambda y_i} = \lambda^{r\,n} \Gamma(r)^{-n} \left( \prod_{i=1}^{n} y_i^{r-1} \right) \exp\left( -\lambda \sum_{i=1}^{n} y_i \right)$$

$$\ln L(\lambda) = r\,n \ln \lambda - n \ln \Gamma(r) + \ln \left( \prod_{i=1}^{n} y_i^{r-1} \right) - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{r\,n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{r\,n}{\sum_{i=1}^{n} y_i} = \frac{r}{\bar{y}}}.$$

Comment:

– When $r = 1$, this reduces to the exponential distribution case.

– If $r$ is also unknown, it will be much more complicated.
  No closed-form solution. One needs numerical solver[2].
  Try MME instead.

[2] [DW, Example 7.2.25]

A random sample of size $n$ from the following population:

E.g. 3. Gamma distribution: $f_Y(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$ with $r > 1$ known.

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^r}{\Gamma(r)} y_i^{r-1} e^{-\lambda y_i} = \lambda^{r\,n} \Gamma(r)^{-n} \left(\prod_{i=1}^n y_i^{r-1}\right) \exp\left(-\lambda \sum_{i=1}^n y_i\right)$$

$$\ln L(\lambda) = r\,n \ln \lambda - n \ln \Gamma(r) + \ln\left(\prod_{i=1}^n y_i^{r-1}\right) - \lambda \sum_{i=1}^n y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{r\,n}{\lambda} - \sum_{i=1}^n y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{r\,n}{\sum_{i=1}^n y_i} = \frac{r}{\bar{y}}}.$$

Comment:

– When $r = 1$, this reduces to the exponential distribution case.
– If $r$ is also unknown, it will be much more complicated.
No closed-form solution. One needs numerical solver[2].
Try MME instead.

[2] [DW, Example 7.2.25]

A random sample of size $n$ from the following population:

E.g. 3. Gamma distribution: $f_Y(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$ with $r > 1$ known.

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^r}{\Gamma(r)} y_i^{r-1} e^{-\lambda y_i} = \lambda^{r\,n} \Gamma(r)^{-n} \left( \prod_{i=1}^{n} y_i^{r-1} \right) \exp\left( -\lambda \sum_{i=1}^{n} y_i \right)$$

$$\ln L(\lambda) = r\,n \ln \lambda - n \ln \Gamma(r) + \ln \left( \prod_{i=1}^{n} y_i^{r-1} \right) - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{r\,n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{r\,n}{\sum_{i=1}^{n} y_i} = \frac{r}{\bar{y}}}.$$

Comment:

- When $r = 1$, this reduces to the exponential distribution case.
- If $r$ is also unknown, it will be much more complicated.
  No closed-form solution. One needs numerical solver[2].
  Try MME instead.

[2] [DW, Example 7.2.25]

A random sample of size $n$ from the following population:

E.g. 3. Gamma distribution: $f_Y(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$ with $r > 1$ known.

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^r}{\Gamma(r)} y_i^{r-1} e^{-\lambda y_i} = \lambda^{r\,n} \Gamma(r)^{-n} \left( \prod_{i=1}^{n} y_i^{r-1} \right) \exp\left( -\lambda \sum_{i=1}^{n} y_i \right)$$

$$\ln L(\lambda) = r\,n \ln \lambda - n \ln \Gamma(r) + \ln \left( \prod_{i=1}^{n} y_i^{r-1} \right) - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{r\,n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{r\,n}{\sum_{i=1}^{n} y_i} = \frac{r}{\bar{y}}}.$$

Comment:

- When $r = 1$, this reduces to the exponential distribution case.
- If $r$ is also unknown, it will be much more complicated.
  No closed-form solution. One needs numerical solver[2].
  Try MME instead.

[2] [DW, Example 7.2.25]

A random sample of size $n$ from the following population:

E.g. 3. Gamma distribution: $f_Y(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$ with $r > 1$ known.

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^r}{\Gamma(r)} y_i^{r-1} e^{-\lambda y_i} = \lambda^{r\,n} \Gamma(r)^{-n} \left( \prod_{i=1}^{n} y_i^{r-1} \right) \exp\left( -\lambda \sum_{i=1}^{n} y_i \right)$$

$$\ln L(\lambda) = r\,n \ln \lambda - n \ln \Gamma(r) + \ln \left( \prod_{i=1}^{n} y_i^{r-1} \right) - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{r\,n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{r\,n}{\sum_{i=1}^{n} y_i} = \frac{r}{\bar{y}}}.$$

Comment:

- When $r = 1$, this reduces to the exponential distribution case.
- If $r$ is also unknown, it will be much more complicated.
  No closed-form solution. One needs numerical solver[2].
  Try MME instead.

[2] [DW, Example 7.2.25]

A detailed study with data:

E.g. 4. Geometric distribution: $p_X(k;p) = (1-p)^{k-1}p$, $k = 1, 2, \cdots$.

$$L(p) = \prod_{i=1}^{n} (1-p)^{k_i - 1} p = (1-p)^{-n+\sum_{i=1}^{k} k_i} p^n.$$

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

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln L(p) = -\frac{-n + \sum_{i=1}^{n} k_i}{1-p} + \frac{n}{p}.$$

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln L(p) = 0 \quad \Longrightarrow \quad \boxed{p_e = \frac{n}{\sum_{i=1}^{n} k_i} = \frac{1}{\bar{k}}}.$$

Comment: Its cousin distribution, the negative binomial distribution can be worked out similarly (See Ex 5.2.14).

A detailed study with data:

E.g. 4. Geometric distribution: $p_X(k;p) = (1-p)^{k-1}p$, $k = 1, 2, \cdots$.

$$L(p) = \prod_{i=1}^{n}(1-p)^{k_i-1}p = (1-p)^{-n+\sum_{i=1}^{k} k_i} p^n.$$

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

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = -\frac{-n + \sum_{i=1}^{n} k_i}{1-p} + \frac{n}{p}.$$

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = 0 \quad \Longrightarrow \quad \boxed{p_e = \frac{n}{\sum_{i=1}^{n} k_i} = \frac{1}{\bar{k}}}.$$

Comment: Its cousin distribution, the negative binomial distribution can be worked out similarly (See Ex 5.2.14).

A detailed study with data:

E.g. 4. Geometric distribution: $p_X(k;p) = (1-p)^{k-1}p$, $k = 1, 2, \cdots$.

$$L(p) = \prod_{i=1}^{n}(1-p)^{k_i-1}p = (1-p)^{-n+\sum_{i=1}^{k} k_i} p^n.$$

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

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = -\frac{-n+\sum_{i=1}^{n} k_i}{1-p} + \frac{n}{p}.$$

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = 0 \quad \Longrightarrow \quad \boxed{p_e = \frac{n}{\sum_{i=1}^{n} k_i} = \frac{1}{\bar{k}}}.$$

Comment: Its cousin distribution, the negative binomial distribution can be worked out similarly (See Ex 5.2.14).

A detailed study with data:

E.g. 4. Geometric distribution: $p_X(k;p) = (1-p)^{k-1}p$, $k = 1, 2, \cdots$.

$$L(p) = \prod_{i=1}^{n}(1-p)^{k_i-1}p = (1-p)^{-n+\sum_{i=1}^{k} k_i} p^n.$$

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

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = -\frac{-n+\sum_{i=1}^{n} k_i}{1-p} + \frac{n}{p}.$$

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = 0 \quad \Longrightarrow \quad \boxed{p_e = \frac{n}{\sum_{i=1}^{n} k_i} = \frac{1}{\bar{k}}}.$$

Comment: Its cousin distribution, the negative binomial distribution can be worked out similarly (See Ex 5.2.14).

A detailed study with data:

E.g. 4. Geometric distribution: $p_X(k;p) = (1-p)^{k-1}p$, $k = 1, 2, \cdots$.

$$L(p) = \prod_{i=1}^{n} (1-p)^{k_i - 1} p = (1-p)^{-n+\sum_{i=1}^{k} k_i} p^n.$$

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

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln L(p) = -\frac{-n + \sum_{i=1}^{n} k_i}{1-p} + \frac{n}{p}.$$

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln L(p) = 0 \quad \Longrightarrow \quad \boxed{p_e = \frac{n}{\sum_{i=1}^{n} k_i} = \frac{1}{\bar{k}}}.$$

Comment: Its cousin distribution, the negative binomial distribution can be worked out similarly (See Ex 5.2.14).

A detailed study with data:

E.g. 4. Geometric distribution: $p_X(k;p) = (1-p)^{k-1}p$, $k = 1, 2, \cdots$.

$$L(p) = \prod_{i=1}^{n}(1-p)^{k_i-1}p = (1-p)^{-n+\sum_{i=1}^{k} k_i} p^n.$$

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

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = -\frac{-n+\sum_{i=1}^{n} k_i}{1-p} + \frac{n}{p}.$$

$$\frac{\mathrm{d}}{\mathrm{d}p}\ln L(p) = 0 \quad \Longrightarrow \quad \boxed{p_e = \frac{n}{\sum_{i=1}^{n} k_i} = \frac{1}{\bar{k}}}.$$

Comment: Its cousin distribution, the negative binomial distribution can be worked out similarly (See Ex 5.2.14).

| k | Observed frequency | Predicted frequency |
|---|---|---|
| 1 | 72 | 74.14 |
| 2 | 35 | 31.2 |
| 3 | 11 | 13.13 |
| 4 | 6 | 5.52 |
| 5 | 2 | 2.32 |
| 6 | 2 | 0.98 |

```r
# The example from the book.
 library (pracma) # Load the library  "Practical  Numerical Math Functions"
k<-c(72, 35, 11, 6, 2, 2) # observed freq.
a=1:6
pe=sum(k)/dot(k,a) # MLE for p.
f=a
for (i in 1:6) {
   f[i] = round((1-pe)^(i-1) * pe * sum(k),2)
}
#  Initialize  the table
d <-matrix(1:18, nrow = 6, ncol = 3)
# Now adding the column names
colnames(d) <- c("k",
                 "Observed freq.",
                 "Predicted freq.")
d[1:6,1]<-a
d[1:6,2]<-k
d[1:6,3]<-f
grid.table(d) # Show the table
PlotResults("unknown", pe, d, "Geometric.pdf") # Output the results  using a user defined function
```

| k | Observed frequency | Predicted frequency |
|---|---|---|
| 1 | 42 | 40.96 |
| 2 | 31 | 27.85 |
| 3 | 15 | 18.94 |
| 4 | 11 | 12.88 |
| 5 | 9 | 8.76 |
| 6 | 5 | 5.96 |
| 7 | 7 | 4.05 |
| 8 | 2 | 2.75 |
| 9 | 1 | 1.87 |
| 10 | 2 | 1.27 |
| 11 | 1 | 0.87 |
| 13 | 1 | 0.59 |
| 14 | 1 | 0.4 |

```r
# Now let's generate random samples from a Geometric distribution with p=1/3 with the same size
        of the sample.
p = 1/3
n = 128
gdata<-rgeom(n, p)+1 # Generate random samples
g<- table(gdata) # Count frequency of your data.
g<- t(rbind(as.numeric(rownames(g)), g)) # Transpose and combine two columns.
pe=n/dot(g[,1], g[,2])  # MLE for p.
f <- g[,1] #  Initialize  f
for (i in 1:nrow(g)) {
  f[i] = round((1-pe)^(i-1) * pe * n,2)
} # Compute the expected frequency
g<-cbind(g,f) # Add one columns to your matrix.
colnames(g) <- c("k",
                 "Observed freq.",
                 "Predicted freq.") # Specify the column names.
d_df <- as.data.frame(d) # One can use data frame to store data
d_df # Show data on your terminal
PlotResults(p, pe, g, "Geometric2.pdf") # Output the results using a user defined function
```

| k | Observed frequency | Predicted frequency |
|---|---|---|
| 1 | 99 | 105.88 |
| 2 | 69 | 68.51 |
| 3 | 47 | 44.33 |
| 4 | 28 | 28.69 |
| 5 | 27 | 18.56 |
| 6 | 9 | 12.01 |
| 7 | 8 | 7.77 |
| 8 | 5 | 5.03 |
| 9 | 5 | 3.25 |
| 10 | 3 | 2.11 |

In case we have several parameters:

E.g. 5. Normal distribution: $f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y \in \mathbb{R}$.

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

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

$$\begin{cases} \frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\ \frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 \end{cases}$$

$$\begin{cases} \frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = 0 \\ \frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = 0 \end{cases} \Longrightarrow \boxed{\begin{cases} \mu_e = \bar{y} \\ \sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \end{cases}}$$

In case we have several parameters:

E.g. 5. Normal distribution: $f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y \in \mathbb{R}$.

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

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

$$\begin{cases} \frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\ \frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 \end{cases}$$

$$\begin{cases} \frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = 0 \\ \frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = 0 \end{cases} \implies \boxed{\begin{cases} \mu_e = \bar{y} \\ \sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \end{cases}}$$

In case we have several parameters:

E.g. 5. Normal distribution: $f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y \in \mathbb{R}$.

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

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

$$\begin{cases} \frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\ \frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 \end{cases}$$

$$\begin{cases} \frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = 0 \\ \frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = 0 \end{cases} \implies \boxed{\begin{cases} \mu_e = \bar{y} \\ \sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \end{cases}}$$

In case we have several parameters:

E.g. 5. Normal distribution: $f_Y(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y\in\mathbb{R}$.

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

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

$$\begin{cases}\dfrac{\partial}{\partial\mu}\ln L(\mu,\sigma^2)=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(y_i-\mu)\\[2ex]\dfrac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2)=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu)^2\end{cases}$$

$$\begin{cases}\dfrac{\partial}{\partial\mu}\ln L(\mu,\sigma^2)=0\\[2ex]\dfrac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2)=0\end{cases}\implies\boxed{\begin{cases}\mu_e=\bar{y}\\[1ex]\sigma_e^2=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i-\bar{y})^2\end{cases}}$$

In case we have several parameters:

E.g. 5. Normal distribution: $f_Y(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y\in\mathbb{R}$.

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

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

$$\begin{cases}\dfrac{\partial}{\partial\mu}\ln L(\mu,\sigma^2)=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(y_i-\mu)\\ \dfrac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2)=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu)^2\end{cases}$$

$$\begin{cases}\dfrac{\partial}{\partial\mu}\ln L(\mu,\sigma^2)=0\\ \dfrac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2)=0\end{cases}\Longrightarrow\boxed{\begin{cases}\mu_e=\bar{y}\\ \sigma_e^2=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i-\bar{y})^2\end{cases}}$$

In case when the parameters determine the support of the density: (Non regular case)

E.g. 6. Uniform distribution on $[a, b]$ with $a < b$: $f_Y(y; a, b) = \frac{1}{b-a}$ if $y \in [a, b]$.

$$L(a, b) = \begin{cases} \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n} & \text{if } a \le y_1, \cdots, y_n \le b, \\ 0 & \text{otherwise.} \end{cases}$$

$L(a, b)$ is monotone increasing in $a$ and decreasing in $b$. Hence, in order to maximize $L(a, b)$, one needs to choose

$$a_e = y_{min} \quad \text{and} \quad b_e = y_{max}.$$

E.g. 7. $f_Y(y; \theta) = \frac{2y}{\theta^2}$ for $y \in [0, \theta]$.

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{2y_i}{\theta^2} = 2^n \theta^{-2n} \prod_{i=1}^{n} y_i & \text{if } 0 \le y_1, \cdots, y_n \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Downarrow$$

$$\theta_e = y_{max}.$$

In case when the parameters determine the support of the density:
(Non regular case)

E.g. 6. Uniform distribution on $[a, b]$ with $a < b$: $f_Y(y; a, b) = \frac{1}{b-a}$ if $y \in [a, b]$.

$$L(a, b) = \begin{cases} \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n} & \text{if } a \le y_1, \cdots, y_n \le b, \\ 0 & \text{otherwise.} \end{cases}$$

$L(a, b)$ is monotone increasing in $a$ and decreasing in $b$. Hence, in order to maximize $L(a, b)$, one needs to choose

$$a_e = y_{min} \quad \text{and} \quad b_e = y_{max}.$$

E.g. 7. $f_Y(y; \theta) = \frac{2y}{\theta^2}$ for $y \in [0, \theta]$.

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{2y_i}{\theta^2} = 2^n \theta^{-2n} \prod_{i=1}^{n} y_i & \text{if } 0 \le y_1, \cdots, y_n \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Downarrow$$

$$\theta_e = y_{max}.$$

In case when the parameters determine the support of the density:
(Non regular case)

E.g. 6. Uniform distribution on $[a, b]$ with $a < b$: $f_Y(y; a, b) = \frac{1}{b-a}$ if $y \in [a, b]$.

$$L(a, b) = \begin{cases} \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n} & \text{if } a \leq y_1, \cdots, y_n \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

$L(a, b)$ is monotone increasing in $a$ and decreasing in $b$. Hence, in order to maximize $L(a, b)$, one needs to choose

$$a_e = y_{min} \quad \text{and} \quad b_e = y_{max}.$$

E.g. 7. $f_Y(y; \theta) = \frac{2y}{\theta^2}$ for $y \in [0, \theta]$.

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{2y_i}{\theta^2} = 2^n \theta^{-2n} \prod_{i=1}^{n} y_i & \text{if } 0 \leq y_1, \cdots, y_n \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Downarrow$$

$$\theta_e = y_{max}.$$

In case when the parameters determine the support of the density:
(Non regular case)

E.g. 6. Uniform distribution on $[a,b]$ with $a<b$: $f_Y(y;a,b)=\frac{1}{b-a}$ if $y\in[a,b]$.

$$L(a,b)=\begin{cases}\prod_{i=1}^n \frac{1}{b-a}=\frac{1}{(b-a)^n} & \text{if } a\le y_1,\cdots,y_n\le b,\\ 0 & \text{otherwise.}\end{cases}$$

$L(a,b)$ is monotone increasing in $a$ and decreasing in $b$. Hence, in order to maximize $L(a,b)$, one needs to choose

$$a_e = y_{min} \quad \text{and} \quad b_e = y_{max}.$$

E.g. 7. $f_Y(y;\theta)=\frac{2y}{\theta^2}$ for $y\in[0,\theta]$.

$$L(\theta)=\begin{cases}\prod_{i=1}^n \frac{2y_i}{\theta^2}=2^n\theta^{-2n}\prod_{i=1}^n y_i & \text{if } 0\le y_1,\cdots,y_n\le \theta,\\ 0 & \text{otherwise.}\end{cases}$$

$$\Downarrow$$

$$\theta_e = y_{max}.$$

In case when the parameters determine the support of the density:
(Non regular case)

E.g. 6. Uniform distribution on $[a, b]$ with $a < b$: $f_Y(y; a, b) = \frac{1}{b-a}$ if $y \in [a, b]$.

$$L(a, b) = \begin{cases} \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n} & \text{if } a \leq y_1, \cdots, y_n \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

$L(a, b)$ is monotone increasing in $a$ and decreasing in $b$. Hence, in order to maximize $L(a, b)$, one needs to choose

$$a_e = y_{min} \quad \text{and} \quad b_e = y_{max}.$$

E.g. 7. $f_Y(y; \theta) = \frac{2y}{\theta^2}$ for $y \in [0, \theta]$.

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{2y_i}{\theta^2} = 2^n \theta^{-2n} \prod_{i=1}^{n} y_i & \text{if } 0 \leq y_1, \cdots, y_n \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Downarrow$$

$$\theta_e = y_{max}.$$

In case when the parameters determine the support of the density: (Non regular case)

E.g. 6. Uniform distribution on $[a, b]$ with $a < b$: $f_Y(y; a, b) = \frac{1}{b-a}$ if $y \in [a, b]$.

$$L(a, b) = \begin{cases} \prod_{i=1}^n \frac{1}{b-a} = \frac{1}{(b-a)^n} & \text{if } a \le y_1, \cdots, y_n \le b, \\ 0 & \text{otherwise.} \end{cases}$$

$L(a, b)$ is monotone increasing in $a$ and decreasing in $b$. Hence, in order to maximize $L(a, b)$, one needs to choose

$$a_e = y_{min} \quad \text{and} \quad b_e = y_{max}.$$

E.g. 7. $f_Y(y; \theta) = \frac{2y}{\theta^2}$ for $y \in [0, \theta]$.

$$L(\theta) = \begin{cases} \prod_{i=1}^n \frac{2y_i}{\theta^2} = 2^n \theta^{-2n} \prod_{i=1}^n y_i & \text{if } 0 \le y_1, \cdots, y_n \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Downarrow$$

$$\theta_e = y_{max}.$$

In case of discrete parameter:

E.g. 8. **Wildlife sampling.** Capture-tag-recapture.... In the history, $a$ tags have been put. In order to estimate the population size $N$, one randomly captures $n$ animals, and there are $k$ tagged. Find the MLE for $N$.

**Sol.** The population follows hypergeometric distr.: $p_X(k;N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$.

$$L(N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$$

How to maximize $L(N)$?

```
> a=10
> k=5
> n=20
> N=seq(a,a+100)
> p=choose(a,k)*choose(N-a,n-k)/
      choose(N,n)
> plot(N,p,type = "p")
> print(paste("The MLE is", n*a/k))
[1] "The MLE is 40"
```

In case of discrete parameter:

E.g. 8. **Wildlife sampling.** Capture-tag-recapture.... In the history, $a$ tags have been put. In order to estimate the population size $N$, one randomly captures $n$ animals, and there are $k$ tagged. Find the MLE for $N$.

**Sol.** The population follows hypergeometric distr.: $p_X(k;N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$.

$$L(N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$$

How to maximize $L(N)$?

```
> a=10
> k=5
> n=20
> N=seq(a,a+100)
> p=choose(a,k)*choose(N-a,n-k)/
      choose(N,n)
> plot(N,p,type = "p")
> print(paste("The MLE is", n*a/k))
[1] "The MLE is 40"
```

In case of discrete parameter:

E.g. 8. **Wildlife sampling.** Capture-tag-recapture.... In the history, $a$ tags have been put. In order to estimate the population size $N$, one randomly captures $n$ animals, and there are $k$ tagged. Find the MLE for $N$.

**Sol.** The population follows hypergeometric distr.: $p_X(k; N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$.

$$L(N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$$

How to maximize $L(N)$?

```
> a=10
> k=5
> n=20
> N=seq(a,a+100)
> p=choose(a,k)*choose(N-a,n-k)/
      choose(N,n)
> plot(N,p,type = "p")
> print(paste("The MLE is", n*a/k))
[1] "The MLE is 40"
```

In case of discrete parameter:

E.g. 8. **Wildlife sampling.** Capture-tag-recapture.... In the history, $a$ tags have been put. In order to estimate the population size $N$, one randomly captures $n$ animals, and there are $k$ tagged. Find the MLE for $N$.

**Sol.** The population follows hypergeometric distr.: $p_X(k; N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$.

$$L(N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$$

How to maximize $L(N)$?

```
> a=10
> k=5
> n=20
> N=seq(a,a+100)
> p=choose(a,k)*choose(N-a,n-k)/
     choose(N,n)
> plot(N,p,type = "p")
> print(paste("The MLE is", n*a/k))
[1] "The MLE is 40"
```

In case of discrete parameter:

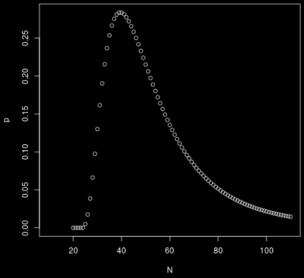
E.g. 8. **Wildlife sampling.** Capture-tag-recapture.... In the history, $a$ tags have been put. In order to estimate the population size $N$, one randomly captures $n$ animals, and there are $k$ tagged. Find the MLE for $N$.

**Sol.** The population follows hypergeometric distr.: $p_X(k;N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$.

$$L(N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$$

How to maximize $L(N)$?

```
> a=10
> k=5
> n=20
> N=seq(a,a+100)
> p=choose(a,k)*choose(N-a,n-k)/
     choose(N,n)
> plot(N,p,type = "p")
> print(paste("The MLE is", n*a/k))
[1] "The MLE is 40"
```



The graph suggests to sudty the following quantity:

$$r(N) := \frac{L(N)}{L(N-1)} = \frac{N-n}{N} \times \frac{N-a}{N-a-n+k}$$

$$r(N) < 1 \quad \Longleftrightarrow \quad na < Nk \quad \text{i.e., } N > \frac{na}{k}$$

$$\boxed{N_e = \arg\max\left\{L(N) : N = \left\lfloor\frac{na}{k}\right\rfloor, \left\lceil\frac{na}{k}\right\rceil\right\}}.$$

□

The graph suggests to sudty the following quantity:

$$r(N) := \frac{L(N)}{L(N-1)} = \frac{N-n}{N} \times \frac{N-a}{N-a-n+k}$$

$$r(N) < 1 \quad \Longleftrightarrow \quad na < Nk \quad \text{i.e., } N > \frac{na}{k}$$

$$\boxed{N_e = \arg\max\left\{L(N) : N = \left\lfloor \frac{na}{k} \right\rfloor, \left\lceil \frac{na}{k} \right\rceil\right\}}.$$

$\square$

The graph suggests to sudty the following quantity:

$$r(N) := \frac{L(N)}{L(N-1)} = \frac{N-n}{N} \times \frac{N-a}{N-a-n+k}$$

$$r(N) < 1 \quad \Longleftrightarrow \quad na < Nk \quad \text{i.e., } N > \frac{na}{k}$$

$$\boxed{N_e = \arg\max \left\{ L(N) : N = \left\lfloor \frac{na}{k} \right\rfloor, \left\lceil \frac{na}{k} \right\rceil \right\}}.$$

$\square$

The graph suggests to sudty the following quantity:

$$r(N) := \frac{L(N)}{L(N-1)} = \frac{N-n}{N} \times \frac{N-a}{N-a-n+k}$$

$$r(N) < 1 \quad \Longleftrightarrow \quad na < Nk \quad \text{i.e., } N > \frac{na}{k}$$

$$\boxed{N_e = \arg\max \left\{ L(N) : N = \left\lfloor \frac{na}{k} \right\rfloor, \left\lceil \frac{na}{k} \right\rceil \right\}}.$$

$\square$

# Method of Moments Estimation

**Rationale:** The population moments should be close to the sample moments, i.e.,

$$\mathbb{E}(Y^k) \approx \frac{1}{n}\sum_{i=1}^{n} y_i^k, \quad k = 1, 2, 3, \cdots .$$

**Definition 5.2.3.** For a random sample of size $n$ from the discrete (resp. continuous) population/pdf $p_X(k;\theta_1,\cdots,\theta_s)$ (resp. $f_Y(y;\theta_1,\cdots,\theta_s)$), solutions to

$$\begin{cases} \mathbb{E}(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i \\ \quad\vdots \\ \mathbb{E}(Y^s) = \frac{1}{n}\sum_{i=1}^{n} y_i^s \end{cases}$$

which are denoted by $\theta_{1e},\cdots,\theta_{se}$, are called the **method of moments estimates** of $\theta_1,\cdots,\theta_s$.

# Examples for MME

MME is often the same as MLE:

E.g. 1. Normal distribution: $f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y \in \mathbb{R}$.

$$\begin{cases} \mu = \mathbb{E}(Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i = \bar{y} \\ \sigma^2 + \mu^2 = \mathbb{E}(Y^2) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^2 \end{cases} \Rightarrow \begin{cases} \mu_e = \bar{y} \\ \sigma_e^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^2 - \mu_e^2 \\ \quad = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{cases}$$

More examples when MLE coincides with MME: Poisson, Exponential, Geometric.

MME is often much more tractable than MLE:

E.g. 2. Gamma distribution[3]: $f_Y(y; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$.

$$\begin{cases} \dfrac{r}{\lambda} = \mathbb{E}(Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i = \bar{y} \\ \dfrac{r}{\lambda^2} + \dfrac{r^2}{\lambda^2} = \mathbb{E}(Y^2) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^2 \end{cases} \quad \Rightarrow \quad \begin{cases} r_e = \dfrac{\bar{y}^2}{\hat{\sigma}^2} \\ \lambda_e = \dfrac{\bar{y}}{\hat{\sigma}^2} = \dfrac{r_e}{\bar{y}} \end{cases}$$

where $\bar{y}$ is the sample mean and $\hat{\sigma}^2$ is the sample variance: $\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Comments: MME for $\lambda$ is consistent with MLE when $r$ is known.

[3] Check Theorem 4.6.3 on p. 269 for mean and variance

Another tractable example for MME, while less tractable for MLE:

E.g. 3. Neg. binomial distribution: $p_X(k;p,r) = \binom{k+r-1}{k}(1-p)^k p^r$, $k = 0, 1, \cdots$.

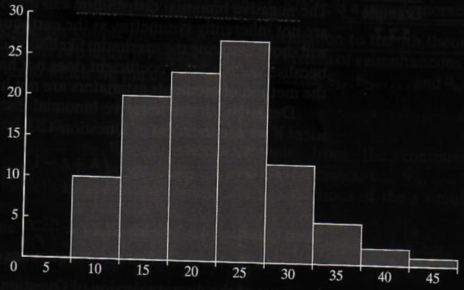
$$\begin{cases} \dfrac{r(1-p)}{p} = \mathbb{E}(X) = \bar{k} \\ \dfrac{r(1-p)}{p^2} = \mathrm{Var}(X) = \hat{\sigma}^2 \end{cases} \qquad \Rightarrow \qquad \begin{cases} p_e = \dfrac{\bar{k}}{\hat{\sigma}^2} \\ \\ r_e = \dfrac{\bar{k}^2}{\hat{\sigma}^2 - \bar{k}} \end{cases}$$

(Case Study 5.2.2 continued)

Table 5.2.4

| Number | Observed Frequency | Expected Frequency |
|---|---|---|
| 0–5 | 0 | 0 |
| 6–10 | 10 | 7.7 |
| 11–15 | 20 | 21.4 |
| 16–20 | 23 | 28.4 |
| 21–25 | 27 | 22.4 |
| 26–30 | 12 | 12.3 |
| 31–35 | 5 | 5.3 |
| 36–40 | 2 | 1.8 |
| > 40 | 1 | 0.7 |

*Data from:* http://www.seattlecentral.edu/qelp/sets/039/039.html




$r_e = 12.74$ and $p_e = 0.391$.

E.g. 4. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$.

$$\overline{y} = \mathbb{E}[Y] = \int_0^\theta \frac{2y^2}{\theta^2}\mathrm{d}y = \frac{2}{3}\frac{y^3}{\theta^2}\bigg|_{y=0}^{y=\theta} = \frac{2}{3}\theta.$$

$$\Downarrow$$

$$\boxed{\theta_e = \frac{3}{2}\overline{y}.}$$

# Plan

# Chapter 5. Estimation

**Rationale.** Point estimate doesn't provide precision information.

By using the variance of the estimator, one can construct an interval such that with a high probability that interval will contain the unknown parameter.

- The interval is called **confidence interval**.
- The high probability is **confidence level**.

**Rationale.** Point estimate doesn't provide precision information.

By using the variance of the estimator, one can construct an interval such that with a high probability that interval will contain the unknown parameter.

- The interval is called **confidence interval**.
- The high probability is **confidence level**.

E.g. 1. A random sample of size $4$, ($Y_1 = 6.5,\ Y_2 = 9.2,\ Y_3 = 9.9,\ Y_4 = 12.4$), from a normal population:

$$f_Y(y;\mu) = \frac{1}{\sqrt{2\pi}\,0.8} e^{-\frac{1}{2}\left(\frac{y-\mu}{0.8}\right)^2}.$$

Both MLE and MME give $\mu_e = \bar{y} = \frac{1}{4}(6.5 + 9.2 + 9.9 + 12.4) = 9.5$. The estimator $\widehat{\mu} = \overline{Y}$ follows normal distribution.

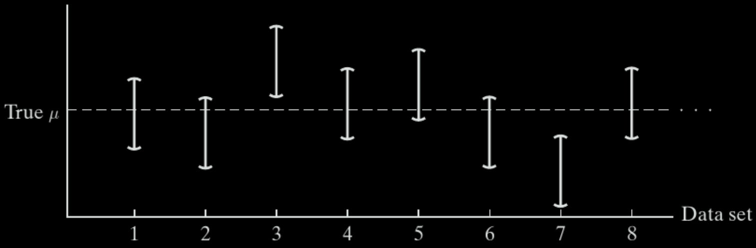
Construct $95\%$-confidence interval for $\mu$ ...

*“The parameter is an unknown constant and no probability statement concerning its value may be made."*

*–Jerzy Neyman, original developer of confidence intervals.*

*"The parameter is an unknown constant and no probability statement concerning its value may be made."*

*–Jerzy Neyman, original developer of confidence intervals.*




Possible 95% confidence intervals for $\mu$

In general, for a normal population with $\sigma$ known, the $100(1-\alpha)\%$ **confidence interval** for $\mu$ is

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Comment: There are many variations

1. One-sided interval such as

$$\left(\bar{y} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \bar{y}\right) \quad \text{or} \quad \left(\bar{y}, \bar{y} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

2. $\sigma$ is unknown and sample size is small: $z$-score $\to$ $t$-score
3. $\sigma$ is unknown and sample size is large: $z$-score by CLT
4. Non-Gaussian population but sample size is large: $z$-score by CLT

In general, for a normal population with $\sigma$ known, the $100(1-\alpha)\%$ **confidence interval** for $\mu$ is

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Comment: There are many variations

1. One-sided interval such as

$$\left(\bar{y} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \bar{y}\right) \quad \text{or} \quad \left(\bar{y}, \bar{y} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

2. $\sigma$ is unknown and sample size is small: $z$-score $\to$ $t$-score
3. $\sigma$ is unknown and sample size is large: $z$-score by CLT
4. Non-Gaussian population but sample size is large: $z$-score by CLT

In general, for a normal population with $\sigma$ known, the $100(1-\alpha)\%$ **confidence interval** for $\mu$ is

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Comment: There are many variations

1. One-sided interval such as

$$\left(\bar{y} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \bar{y}\right) \quad \text{or} \quad \left(\bar{y}, \bar{y} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

2. $\sigma$ is unknown and sample size is small: $z$-score $\to$ $t$-score
3. $\sigma$ is unknown and sample size is large: $z$-score by CLT
4. Non-Gaussian population but sample size is large: $z$-score by CLT

In general, for a normal population with $\sigma$ known, the $100(1-\alpha)\%$ **confidence interval** for $\mu$ is

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Comment: There are many variations

1. One-sided interval such as

$$\left(\bar{y} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \bar{y}\right) \quad \text{or} \quad \left(\bar{y}, \bar{y} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

2. $\sigma$ is unknown and sample size is small: $z$-score $\rightarrow$ $t$-score
3. $\sigma$ is unknown and sample size is large: $z$-score by CLT
4. Non-Gaussian population but sample size is large: $z$-score by CLT

In general, for a normal population with $\sigma$ known, the $100(1-\alpha)\%$ **confidence interval** for $\mu$ is

$$\left(\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Comment: There are many variations

1. One-sided interval such as

$$\left(\bar{y} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \bar{y}\right) \quad \text{or} \quad \left(\bar{y}, \bar{y} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$$

2. $\sigma$ is unknown and sample size is small: *z*-score $\rightarrow$ *t*-score
3. $\sigma$ is unknown and sample size is large: *z*-score by CLT
4. Non-Gaussian population but sample size is large: *z*-score by CLT

**Theorem.** Let $k$ be the number of successes in $n$ independent trials, where $n$ is large and $p = \mathbb{P}(\textit{success})$ is unknown. An approximate $100(1-\alpha)\%$ confidence interval for $p$ is the set of numbers

$$\left( \frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}} \right).$$

Proof: It follows the following facts:

- $X \sim$binomial$(n, p)$ iff $X = Y_1 + \cdots + Y_n$, while $Y_i$ are i.i.d. Bernoulli$(p)$:

$$\mathbb{E}[Y_i] = p \quad \text{and} \quad \mathrm{Var}(Y_i) = p(1-p).$$

- **Central Limit Theorem**: Let $W_1, W_2, \cdots, W_n$ be an sequence of i.i.d. random variables, whose distribution has mean $\mu$ and variance $\sigma^2$, then

$$\frac{\sum_{i=1}^{n} W_i - n\mu}{\sqrt{n\sigma^2}} \quad \textit{approximately} \text{ follows} \quad N(0,1), \quad \text{when } n \text{ is large.}$$

**Theorem.** Let $k$ be the number of successes in $n$ independent trials, where $n$ is large and $p = \mathbb{P}(success)$ is unknown. An approximate $100(1-\alpha)\%$ confidence interval for $p$ is the set of numbers

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

Proof: It follows the following facts:

- $X \sim$binomial$(n, p)$ iff $X = Y_1 + \cdots + Y_n$, while $Y_i$ are i.i.d. Bernoulli$(p)$:

$$\mathbb{E}[Y_i] = p \quad \text{and} \quad \mathrm{Var}(Y_i) = p(1-p).$$

- **Central Limit Theorem**: Let $W_1, W_2, \cdots, W_n$ be an sequence of i.i.d. random variables, whose distribution has mean $\mu$ and variance $\sigma^2$, then

$$\frac{\sum_{i=1}^{n} W_i - n\mu}{\sqrt{n\sigma^2}} \quad \textit{approximately}\text{ follows} \quad N(0,1), \quad \text{when } n \text{ is large.}$$

**Theorem.** Let $k$ be the number of successes in $n$ independent trials, where $n$ is large and $p = \mathbb{P}(\textit{success})$ is unknown. An approximate $100(1-\alpha)\%$ confidence interval for $p$ is the set of numbers

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

Proof: It follows the following facts:

- $X \sim$binomial$(n, p)$ iff $X = Y_1 + \cdots + Y_n$, while $Y_i$ are i.i.d. Bernoulli$(p)$:

$$\mathbb{E}[Y_i] = p \quad \text{and} \quad \mathrm{Var}(Y_i) = p(1-p).$$

- **Central Limit Theorem**: Let $W_1, W_2, \cdots, W_n$ be an sequence of i.i.d. random variables, whose distribution has mean $\mu$ and variance $\sigma^2$, then

$$\frac{\sum_{i=1}^{n} W_i - n\mu}{\sqrt{n\sigma^2}} \quad \textit{approximately} \text{ follows} \quad N(0,1), \quad \text{when } n \text{ is large.}$$

- When the sample size $n$ is large, by the central limit theorem,

$$\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}} \overset{\text{ap.}}{\sim} N(0,1)$$

$$\|$$

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\frac{X}{n} - p}{\sqrt{\frac{p_e(1-p_e)}{n}}}$$

- Since $p_e = \frac{k}{n}$, we see that

$$\mathbb{P}\left( -z_{\alpha/2} \le \frac{\frac{X}{n} - p}{\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}} \le z_{\alpha/2} \right) \approx 1 - \alpha$$

i.e., the $100(1-\alpha)\%$ confidence interval for $p$ is

$$\left( \frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}} \right).$$

- When the sample size $n$ is large, by the central limit theorem,

$$\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}} \quad \overset{\text{ap.}}{\sim} \quad N(0,1)$$

$$\|$$

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\frac{X}{n} - p}{\sqrt{\frac{p_e(1-p_e)}{n}}}$$

- Since $p_e = \frac{k}{n}$, we see that

$$\mathbb{P}\left(-z_{\alpha/2} \le \frac{\frac{X}{n} - p}{\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}} \le z_{\alpha/2}\right) \approx 1-\alpha$$

i.e., the $100(1-\alpha)\%$ confidence interval for $p$ is

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}, \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

- When the sample size $n$ is large, by the central limit theorem,

$$\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}} \overset{\text{ap.}}{\sim} N(0,1)$$

$$\|$$

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\frac{X}{n} - p}{\sqrt{\frac{p_e(1-p_e)}{n}}}$$

- Since $p_e = \frac{k}{n}$, we see that

$$\mathbb{P}\left(-z_{\alpha/2} \le \frac{\frac{X}{n} - p}{\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}} \le z_{\alpha/2}\right) \approx 1 - \alpha$$

i.e., the $100(1-\alpha)\%$ confidence interval for $p$ is

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}, \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

- When the sample size $n$ is large, by the central limit theorem,

$$\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}} \quad \overset{\text{ap.}}{\sim} \quad N(0,1)$$

$$\|$$

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\frac{X}{n} - p}{\sqrt{\frac{p_e(1-p_e)}{n}}}$$

- Since $p_e = \frac{k}{n}$, we see that

$$\mathbb{P}\left( -z_{\alpha/2} \leq \frac{\frac{X}{n} - p}{\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}} \leq z_{\alpha/2} \right) \approx 1 - \alpha$$

i.e., the $100(1-\alpha)\%$ confidence interval for $p$ is

$$\left( \frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}} \right).$$

$\square$

- When the sample size $n$ is large, by the central limit theorem,

$$\frac{\sum_{i=1}^{n} Y_i - np}{\sqrt{np(1-p)}} \overset{\text{ap.}}{\sim} N(0,1)$$

$$\|$$

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \frac{\frac{X}{n} - p}{\sqrt{\frac{p_e(1-p_e)}{n}}}$$

- Since $p_e = \frac{k}{n}$, we see that

$$\mathbb{P}\left(-z_{\alpha/2} \le \frac{\frac{X}{n} - p}{\sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}} \le z_{\alpha/2}\right) \approx 1 - \alpha$$

i.e., the $100(1-\alpha)\%$ confidence interval for $p$ is

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

$\square$

E.g. 1. Use *median test* to check the randomness of a random generator.

*Suppose $y_1, \cdots, y_n$ denote measurements presumed to have come from a continuous pdf $f_Y(y)$. Let $k$ denote the number of $y_i$'s that are less than the median of $f_Y(y)$. If the sample is random, we would expect the difference between $\frac{k}{n}$ and $\frac{1}{2}$ to be small. More specifically, a 95% confidence interval based on $k$ should contain the value 0.5.*

Let $f_Y(y) = e^{-y}$. The median is $m = 0.69315$.

E.g. 1. Use *median test* to check the randomness of a random generator.

*Suppose $y_1, \cdots, y_n$ denote measurements presumed to have come from a continuous pdf $f_Y(y)$. Let $k$ denote the number of $y_i$'s that are less than the median of $f_Y(y)$. If the sample is random, we would expect the difference between $\frac{k}{n}$ and $\frac{1}{2}$ to be small. More specifically, a 95% confidence interval based on $k$ should contain the value 0.5.*

Let $f_Y(y) = e^{-y}$. The median is $m = 0.69315$.

E.g. 1. Use *median test* to check the randomness of a random generator.

*Suppose $y_1, \cdots, y_n$ denote measurements presumed to have come from a continuous pdf $f_Y(y)$. Let $k$ denote the number of $y_i$'s that are less than the median of $f_Y(y)$. If the sample is random, we would expect the difference between $\frac{k}{n}$ and $\frac{1}{2}$ to be small. More specifically, a 95% confidence interval based on $k$ should contain the value 0.5.*

Let $f_Y(y) = e^{-y}$. The median is $m = 0.69315$.

E.g. 1. Use *median test* to check the randomness of a random generator.

*Suppose $y_1, \cdots, y_n$ denote measurements presumed to have come from a continuous pdf $f_Y(y)$. Let $k$ denote the number of $y_i$'s that are less than the median of $f_Y(y)$. If the sample is random, we would expect the difference between $\frac{k}{n}$ and $\frac{1}{2}$ to be small. More specifically, a 95% confidence interval based on $k$ should contain the value 0.5.*

Let $f_Y(y) = e^{-y}$. The median is $m = 0.69315$.

```r
#! /usr/bin/Rscript
main <- function() {
  args <- commandArgs(trailingOnly = TRUE)
  n <- 100 # Number of random samples.
  r <- as.numeric(args[1]) # Rate of the exponential
  # Check if the rate argument is given.
  if (is.na(r)) return("Please provide the rate and try again.")

  # Now start computing ...
  f <- function (y) pexp(y, rate = r)-0.5
  m <- uniroot(f, lower = 0, upper = 100, tol = 1e-9)$root
  print(paste("For rate ", r, "exponential distribution ,",
              "the median is equal to ", round(m,3)))
  data <- rexp(n,r) # Generate n random samples
  data <- round(data,3) # Round to 3 digits after decimal
  data <- matrix(data, nrow = 10,ncol = 10) # Turn the data to a matrix
  prmatrix(data) # Show data on terminal
  k <- sum(data > m) # Count how many entries is bigger than m
  lowerbd = k/n - 1.96 * sqrt((k/n)*(1-k/n)/n);
  upperbd = k/n + 1.96 *sqrt((k/n)*(1-k/n)/n);
  print(paste("The 95% confidence interval is (",
              round(lowerbd,3), ",",
              round(upperbd,3), ")"))
}
main()
```

Try commandline ...

```
Math362:./Example-5-3-2.R 1
[1] "For rate  1 exponential distribution, the median is equal to  0.693"
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] 1.324 1.211 0.561 0.640 2.816 2.348 0.788 2.243 1.759 0.103
 [2,] 0.476 2.288 0.106 0.079 0.636 1.941 0.801 3.838 0.612 0.030
 [3,] 1.085 0.305 0.354 1.013 0.687 1.656 1.043 0.389 1.476 2.158
 [4,] 1.267 1.031 0.917 0.681 0.912 0.236 0.054 0.862 0.065 0.402
 [5,] 0.957 1.003 1.665 1.137 0.378 1.182 0.659 1.923 1.127 0.364
 [6,] 0.307 0.127 0.203 0.394 1.392 2.378 4.192 0.365 3.227 0.337
 [7,] 0.707 0.049 0.391 1.967 1.220 2.605 0.887 1.749 1.479 1.526
 [8,] 0.662 0.141 0.318 0.523 0.646 1.202 0.442 0.174 1.178 0.177
 [9,] 0.397 0.493 0.214 0.522 2.024 4.109 1.268 1.041 0.948 0.382
[10,] 2.260 0.292 0.437 0.962 0.224 4.221 0.594 0.218 0.601 0.941
[1] "The 95% confidence interval is ( 0.422 , 0.618 )"
Math362:./Example-5-3-2.R 10
[1] "For rate  10 exponential distribution, the median is equal to  0.069"
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] 0.199 0.069 0.013 0.025 0.000 0.107 0.068 0.116 0.066 0.146
 [2,] 0.027 0.076 0.044 0.458 0.052 0.127 0.100 0.100 0.014 0.061
 [3,] 0.014 0.078 0.044 0.072 0.028 0.141 0.038 0.022 0.037 0.093
 [4,] 0.042 0.015 0.250 0.132 0.292 0.072 0.105 0.244 0.046 0.054
 [5,] 0.134 0.074 0.182 0.057 0.021 0.038 0.095 0.196 0.004 0.048
 [6,] 0.016 0.021 0.163 0.030 0.139 0.063 0.054 0.006 0.023 0.051
 [7,] 0.227 0.055 0.091 0.121 0.066 0.114 0.004 0.021 0.035 0.211
 [8,] 0.113 0.083 0.129 0.338 0.160 0.008 0.014 0.167 0.050 0.127
 [9,] 0.053 0.073 0.054 0.098 0.004 0.036 0.274 0.276 0.004 0.159
[10,] 0.045 0.469 0.152 0.003 0.129 0.017 0.084 0.072 0.162 0.007
[1] "The 95% confidence interval is ( 0.392 , 0.588 )"
Math362:
```

Instead of the C.I. $\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\ \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right)$.

One can simply specify the mean $\dfrac{k}{n}$ and

the **margin of error**: $$d := z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}.$$

$$\max_{p\in(0,1)} p(1-p) = p(1-p)\Big|_{p=1/2} = 1/4 \quad\Longrightarrow\quad d \leq \frac{z_{\alpha/2}}{2\sqrt{n}} =: d_m.$$

Comment:

1. When $p$ is close to $1/2$, $d \approx \frac{z_{\alpha/2}}{2\sqrt{n}}$, which is equivalent to $\sigma_p \approx \frac{1}{2\sqrt{n}}$.

   E.g., $n = 1000$, $k/n = 0.48$, and $\alpha = 5\%$, then

$$d = 1.96\sqrt{\frac{0.48 \times 0.52}{1000}} = 0.0309\underline{7} \quad \text{and} \quad d_m = \frac{1.96}{2\sqrt{1000}} = 0.0309\underline{9}$$

$$\sigma_p = \sqrt{\frac{0.48 \times 0.52}{1000}} = 0.015\underline{7}9873 \quad \text{and} \quad \sigma_p \approx \frac{1}{2\sqrt{1000}} = 0.015\underline{8}1139.$$

2. When $p$ is away from $1/2$, the discrepancy between $d$ and $d_m$ becomes big....

Comment:

1. When $p$ is close to $1/2$, $d \approx \frac{z_{\alpha/2}}{2\sqrt{n}}$, which is equivalent to $\sigma_p \approx \frac{1}{2\sqrt{n}}$.

   E.g., $n = 1000$, $k/n = 0.48$, and $\alpha = 5\%$, then

$$d = 1.96\sqrt{\frac{0.48 \times 0.52}{1000}} = 0.0309\underline{7} \quad \text{and} \quad d_m = \frac{1.96}{2\sqrt{1000}} = 0.0309\underline{9}$$

$$\sigma_p = \sqrt{\frac{0.48 \times 0.52}{1000}} = 0.015\underline{7}9873 \quad \text{and} \quad \sigma_p \approx \frac{1}{2\sqrt{1000}} = 0.015\underline{8}1139.$$

2. When $p$ is away from $1/2$, the discrepancy between $d$ and $d_m$ becomes big....

E.g. Running for presidency. Max and Sirius obtained 480 and 520 votes, respectively. What is probability that Max will win?

What if the sample size is $n = 5000$, and Max obtained 2400 votes.

# Choosing sample sizes

$$d \le z_{\alpha/2}\sqrt{p(1-p)/n} \quad \Longleftrightarrow \quad n \ge \frac{z_{\alpha/2}^2 p(1-p)}{d^2} \qquad \text{(When } p \text{ is known)}$$

$$d \le \frac{z_{\alpha/2}}{2\sqrt{n}} \quad \Longleftrightarrow \quad n \ge \frac{z_{\alpha/2}^2}{4d^2} \qquad \text{(When } p \text{ is unknown)}$$

E.g. Anti-smoking campaign. Need to find an $95\%$ C.I. with a margin of error equal to $1\%$. Determine the sample size?

Answer: $n \ge \frac{1.96^2}{4\times 0.01^2} = 9640.$

E.g.' In order to reduce the sample size, a small sample is used to determine $p$. One finds that $p \approx 0.22$. Determine the sample size again.

Answer: $n \ge \frac{1.96^2 \times 0.22 \times 0.78}{\times 0.01^2} = 6592.2.$

# Choosing sample sizes

$$d \le z_{\alpha/2}\sqrt{p(1-p)/n} \quad \Longleftrightarrow \quad n \ge \frac{z_{\alpha/2}^2 p(1-p)}{d^2} \qquad \text{(When } p \text{ is known)}$$

$$d \le \frac{z_{\alpha/2}}{2\sqrt{n}} \quad \Longleftrightarrow \quad n \ge \frac{z_{\alpha/2}^2}{4d^2} \qquad \text{(When } p \text{ is unknown)}$$

E.g. Anti-smoking campaign. Need to find an $95\%$ C.I. with a margin of error equal to $1\%$. Determine the sample size?

Answer: $n \ge \frac{1.96^2}{4\times 0.01^2} = 9640.$

E.g.' In order to reduce the sample size, a small sample is used to determine $p$. One finds that $p \approx 0.22$. Determine the sample size again.

Answer: $n \ge \frac{1.96^2 \times 0.22 \times 0.78}{\times 0.01^2} = 6592.2.$

# Plan

# Chapter 5. Estimation

# § 5.4 Properties of Estimators

**Question:** Estimators are not in general unique (MLE or MME ...). How to select one estimator?

Recall: For a random sample of size $n$ from the population with given pdf, we have $X_1, \cdots, X_n$, which are i.i.d. r.v.'s. The estimator $\hat{\theta}$ is a function of $X_i's$:

$$\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n).$$

**Criterions:**

1. Unbiased. (Mean)
2. Efficiency, the minimum-variance estimator. (Variance)
3. Sufficency.
4. Consistency. (Asymptotic behavior)

# § 5.4 Properties of Estimators

**Question:** Estimators are not in general unique (MLE or MME ...). How to select one estimator?

Recall: For a random sample of size $n$ from the population with given pdf, we have $X_1, \cdots, X_n$, which are i.i.d. r.v.'s. The estimator $\hat{\theta}$ is a function of $X_i's$:

$$\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n).$$

**Criterions:**

1. Unbiased. (Mean)
2. Efficiency, the minimum-variance estimator. (Variance)
3. Sufficency.
4. Consistency. (Asymptotic behavior)

# § 5.4 Properties of Estimators

**Question:** Estimators are not in general unique (MLE or MME ...). How to select one estimator?

Recall: For a random sample of size $n$ from the population with given pdf, we have $X_1, \cdots, X_n$, which are i.i.d. r.v.'s. The estimator $\hat{\theta}$ is a function of $X_i's$:

$$\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n).$$

**Criterions:**

1. Unbiased. (Mean)
2. Efficiency, the minimum-variance estimator. (Variance)
3. Sufficency.
4. Consistency. (Asymptotic behavior)

# § 5.4 Properties of Estimators

**Question:** Estimators are not in general unique (MLE or MME ...). How to select one estimator?

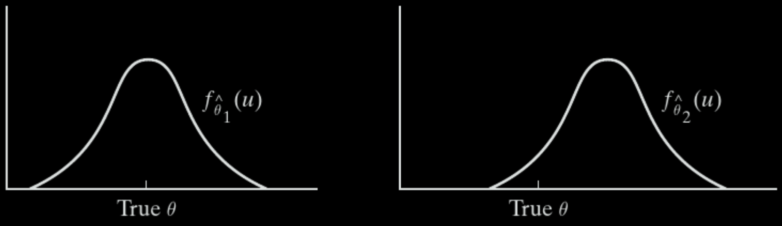
Recall: For a random sample of size $n$ from the population with given pdf, we have $X_1, \cdots, X_n$, which are i.i.d. r.v.'s. The estimator $\hat{\theta}$ is a function of $X_i' s$:

$$\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n).$$

**Criterions:**

1. Unbiased. (Mean)
2. Efficiency, the minimum-variance estimator. (Variance)
3. Sufficency.
4. Consistency. (Asymptotic behavior)

# Unbiasedness




**Definition 5.4.1.** Given a random sample of size *n* whose population distribution dependes on an unknown parameter $\theta$, let $\hat{\theta}$ be an estimator of $\theta$.

Then $\hat{\theta}$ is called **unbiased** if $\mathbb{E}(\hat{\theta}) = \theta$;

and $\hat{\theta}$ is called **asymptotically unbiased** if $\lim_{n\to\infty} \mathbb{E}(\hat{\theta}) = \theta$.

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$.

- $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$
- $\hat{\theta}_2 = Y_{max}$.
- $\hat{\theta}_3 = \frac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$. Show that for any constants $a_i$'s,

$$\hat{\theta} = \sum_{i=1}^{n} a_i X_i \quad \text{is unbiased} \quad \Longleftrightarrow \quad \sum_{i=1}^{n} a_i = 1.$$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$.

- $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$
- $\hat{\theta}_2 = Y_{max}$.
- $\hat{\theta}_3 = \frac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$. Show that for any constants $a_i$'s,

$$\hat{\theta} = \sum_{i=1}^{n} a_i X_i \quad \text{is unbiased} \quad \iff \quad \sum_{i=1}^{n} a_i = 1.$$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$.

- $\hat{\theta}_1 = \dfrac{3}{2}\overline{Y}$
- $\hat{\theta}_2 = Y_{max}$.
- $\hat{\theta}_3 = \dfrac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$. Show that for any constants $a_i$'s,

$$\hat{\theta} = \sum_{i=1}^{n} a_i X_i \quad \text{is unbiased} \quad \Longleftrightarrow \quad \sum_{i=1}^{n} a_i = 1.$$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$.

- $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$
- $\hat{\theta}_2 = Y_{max}$.
- $\hat{\theta}_3 = \dfrac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$. Show that for any constants $a_i$'s,

$$\hat{\theta} = \sum_{i=1}^{n} a_i X_i \quad \text{is unbiased} \quad \iff \quad \sum_{i=1}^{n} a_i = 1.$$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$.

- $\hat{\theta}_1 = \dfrac{3}{2}\overline{Y}$
- $\hat{\theta}_2 = Y_{max}$.
- $\hat{\theta}_3 = \dfrac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$. Show that for any constants $a_i$'s,

$$\hat{\theta} = \sum_{i=1}^{n} a_i X_i \quad \text{is unbiased} \quad \Longleftrightarrow \quad \sum_{i=1}^{n} a_i = 1.$$

E.g. 3. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\sigma^2 = \mathrm{Var}(X)$.

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

– $S^2 = \text{Sample Variance} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

– $S = \text{Sample Standard Deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$. (Biased for $\sigma$!)

E.g. 3. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\sigma^2 = \mathrm{Var}(X)$.

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

- $S^2 = \text{Sample Variance} = \dfrac{1}{n-1} \sum_{i=1}^{n} \left( X_i - \overline{X} \right)^2$

- $S = \text{Sample Standard Deviation} = \sqrt{\dfrac{1}{n-1} \sum_{i=1}^{n} \left( X_i - \overline{X} \right)^2}.$ (Biased for $\sigma$!)

E.g. 3. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\sigma^2 = \mathrm{Var}(X)$.

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

- $S^2 = \text{Sample Variance} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

- $S = \text{Sample Standard Deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$. (Biased for $\sigma$!)

E.g. 3. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\sigma^2 = \mathrm{Var}(X)$.

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

- $S^2 = \text{Sample Variance} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

- $S = \text{Sample Standard Deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$ (Biased for $\sigma$!)

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^{n} Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y \\
&= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^{n} Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y \\
&= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^{n} Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n \int_0^\infty \frac{1}{y} \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y \\
&= \frac{n\lambda}{n-1} \int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^n Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y}\mathrm{d}y \\
&= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)}\mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^n \theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^n Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) = n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)}y^{n-1}e^{-\lambda y}\mathrm{d}y$$

$$= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)}y^{(n-1)-1}e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)}\mathrm{d}y$$

$$= \frac{n}{n-1}\lambda.$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^n \theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^n Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y \\
&= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta}e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^n \theta = \theta.$$

**E.g. 4.** Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^{n} Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y \\
&= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

**E.g. 4'.** Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^n Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y \\
&= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^n \theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^{n} Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) = n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y$$

$$= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y$$

$$= \frac{n}{n-1}\lambda.$$

Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \theta = \theta.$$

E.g. 4. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. $\widehat{\lambda} = 1/\overline{Y}$ is biased.

$n\overline{Y} = \sum_{i=1}^n Y_i \sim$ Gamma distribution$(n, \lambda)$. Hence,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\lambda}\right) = \mathbb{E}\left(1/\overline{Y}\right) &= n\int_0^\infty \frac{1}{y}\frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y} \mathrm{d}y \\
&= \frac{n\lambda}{n-1}\int_0^\infty \underbrace{\frac{\lambda^{n-1}}{\Gamma(n-1)} y^{(n-1)-1} e^{-\lambda y}}_{\text{pdf for Gamma distr. } (n-1,\lambda)} \mathrm{d}y \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

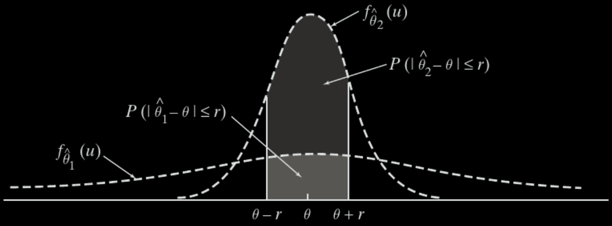
Biased! But $\mathbb{E}(\widehat{\lambda}) = \frac{n}{n-1}\lambda \to \lambda$ as $n \to \infty$. (Asymptotically unbiased.)

Note: $\widehat{\lambda}^* = \frac{n-1}{n\overline{Y}}$ is unbiased.

E.g. 4'. Exponential distr.: $f_Y(y;\theta) = \frac{1}{\theta} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ is unbiased.

$$\mathbb{E}\left(\widehat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^n \theta = \theta.$$

# Efficiency




**Definition 5.4.2.** Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators for a parameter $\theta$. If $\text{Var}(\widehat{\theta}_1) < \text{Var}(\widehat{\theta}_2)$, then we say that $\widehat{\theta}_1$ is **more efficient** than $\widehat{\theta}_2$. The **relative efficiency** of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_2$ is the ratio $\text{Var}(\widehat{\theta}_1)/\text{Var}(\widehat{\theta}_2)$.

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$. Which is more efficient? Find the relative efficiency of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_3$ .

- $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$
- $\hat{\theta}_3 = \frac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$ (suppose $\sigma^2 = \mathrm{Var}(X) < \infty$).

Among all possible unbiased estimators $\hat{\theta} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$. Find the most efficient one.

Sol:

$$\mathrm{Var}(\widehat{\theta}) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X) = \sigma^2 \sum_{i=1}^n a_i^2 \geq \sigma^2 \frac{1}{n}\left(\sum_{i=1}^n a_i\right)^2 = \frac{1}{n}\sigma^2,$$

with equality iff $a_1 = \cdots = a_n = 1/n$.

Hence, the most efficient one is the sample mean $\widehat{\theta} = \overline{X}$. $\square$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$. Which is more efficient? Find the relative efficiency of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_3$ .

- $\hat{\theta}_1 = \dfrac{3}{2}\overline{Y}$
- $\hat{\theta}_3 = \dfrac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$ (suppose $\sigma^2 = \mathrm{Var}(X) < \infty$).

Among all possible unbiased estimators $\hat{\theta} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$. Find the most efficient one.

Sol:

$$\mathrm{Var}(\hat{\theta}) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X) = \sigma^2 \sum_{i=1}^n a_i^2 \geq \sigma^2 \frac{1}{n} \left( \sum_{i=1}^n a_i \right)^2 = \frac{1}{n}\sigma^2,$$

with equality iff $a_1 = \cdots = a_n = 1/n$.

Hence, the most efficient one is the sample mean $\widehat{\theta} = \overline{X}$. $\square$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$. Which is more efficient? Find the relative efficiency of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_3$ .

- $\hat{\theta}_1 = \dfrac{3}{2}\overline{Y}$
- $\hat{\theta}_3 = \dfrac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$ (suppose $\sigma^2 = \mathrm{Var}(X) < \infty$).

Among all possible unbiased estimators $\hat{\theta} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$. Find the most efficient one.

Sol:

$$\mathrm{Var}(\widehat{\theta}) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X) = \sigma^2 \sum_{i=1}^n a_i^2 \geq \sigma^2 \frac{1}{n} \left( \sum_{i=1}^n a_i \right)^2 = \frac{1}{n}\sigma^2,$$

with equality iff $a_1 = \cdots = a_n = 1/n$.

Hence, the most efficient one is the sample mean $\widehat{\theta} = \overline{X}$. $\square$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$. Which is more efficient? Find the relative efficiency of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_3$ .

- $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$
- $\hat{\theta}_3 = \frac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$ (suppose $\sigma^2 = \mathrm{Var}(X) < \infty$).

Among all possible unbiased estimators $\hat{\theta} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$. Find the most efficient one.

Sol:

$$\mathrm{Var}(\widehat{\theta}) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X) = \sigma^2 \sum_{i=1}^n a_i^2 \geq \sigma^2 \frac{1}{n}\left(\sum_{i=1}^n a_i\right)^2 = \frac{1}{n}\sigma^2,$$

with equality iff $a_1 = \cdots = a_n = 1/n$.

Hence, the most efficient one is the sample mean $\widehat{\theta} = \overline{X}$. $\square$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$. Which is more efficient? Find the relative efficiency of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_3$ .

- $\hat{\theta}_1 = \dfrac{3}{2}\overline{Y}$
- $\hat{\theta}_3 = \dfrac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$ (suppose $\sigma^2 = \mathrm{Var}(X) < \infty$).

Among all possible unbiased estimators $\hat{\theta} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$. Find the most efficient one.

Sol:

$$\mathrm{Var}(\widehat{\theta}) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X) = \sigma^2 \sum_{i=1}^n a_i^2 \geq \sigma^2 \frac{1}{n}\left(\sum_{i=1}^n a_i\right)^2 = \frac{1}{n}\sigma^2,$$

with equality iff $a_1 = \cdots = a_n = 1/n$.

Hence, the most efficient one is the sample mean $\widehat{\theta} = \overline{X}$. $\square$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$. Which is more efficient? Find the relative efficiency of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_3$ .

- $\hat{\theta}_1 = \frac{3}{2}\overline{Y}$
- $\hat{\theta}_3 = \frac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$ (suppose $\sigma^2 = \text{Var}(X) < \infty$).

Among all possible unbiased estimators $\hat{\theta} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$. Find the most efficient one.

Sol:

$$\text{Var}(\widehat{\theta}) = \sum_{i=1}^n a_i^2 \text{Var}(X) = \sigma^2 \sum_{i=1}^n a_i^2 \ge \sigma^2 \frac{1}{n}\left(\sum_{i=1}^n a_i\right)^2 = \frac{1}{n}\sigma^2,$$

with equality iff $a_1 = \cdots = a_n = 1/n$.

Hence, the most efficient one is the sample mean $\widehat{\theta} = \overline{X}$. $\square$

E.g. 1. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ if $y \in [0,\theta]$. Which is more efficient? Find the relative efficiency of $\widehat{\theta}_1$ w.r.t. $\widehat{\theta}_3$ .

- $\hat{\theta}_1 = \dfrac{3}{2}\overline{Y}$
- $\hat{\theta}_3 = \dfrac{2n+1}{2n} Y_{max}$.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample of size $n$ with the unknown parameter $\theta = \mathbb{E}(X)$ (suppose $\sigma^2 = \mathrm{Var}(X) < \infty$).

Among all possible unbiased estimators $\hat{\theta} = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$. Find the most efficient one.

Sol:

$$\mathrm{Var}(\widehat{\theta}) = \sum_{i=1}^n a_i^2 \mathrm{Var}(X) = \sigma^2 \sum_{i=1}^n a_i^2 \geq \sigma^2 \frac{1}{n}\left(\sum_{i=1}^n a_i\right)^2 = \frac{1}{n}\sigma^2,$$

with equality iff $a_1 = \cdots = a_n = 1/n$.

Hence, the most efficient one is the sample mean $\widehat{\theta} = \overline{X}$. $\square$

# Plan

# Chapter 5. Estimation

## § 5.5 MVE: The Cramér-Rao Lower Bound

**Question:** Can one identify the unbiased estimator having the *smallest* variance?

**Short answer:** In many cases, yes!

We are going to develop the theory to answer this question in details!

**Regular Estimation/Condition:** The set of $y$ (resp. $k$) values, where $f_Y(y;\theta) \neq 0$ (resp. $p_X(k;\theta) \neq 0$), does not depend on $\theta$.

i.e., the domain of the pdf does not depend on the parameter (so that one can differentiate under integration).

**Definition.** The **Fisher's Information** of a continous (resp. discrete) random variable $Y$ (resp. $X$) with pdf $f_Y(y;\theta)$ (resp. $p_X(k;\theta)$) is defined as

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln f_Y(Y;\theta)}{\partial\theta}\right)^2\right] \qquad \left(\text{resp.} \quad \mathbb{E}\left[\left(\frac{\partial \ln p_X(X;\theta)}{\partial\theta}\right)^2\right]\right).$$

**Regular Estimation/Condition:** The set of $y$ (resp. $k$) values, where $f_Y(y;\theta) \neq 0$ (resp. $p_X(k;\theta) \neq 0$), does not depend on $\theta$.

i.e., the domain of the pdf does not depend on the parameter (so that one can differentiate under integration).

**Definition.** The **Fisher's Information** of a continous (resp. discrete) random variable $Y$ (resp. $X$) with pdf $f_Y(y;\theta)$ (resp. $p_X(k;\theta)$) is defined as

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln f_Y(Y;\theta)}{\partial \theta}\right)^2\right] \qquad \left(\text{resp. } \quad \mathbb{E}\left[\left(\frac{\partial \ln p_X(X;\theta)}{\partial \theta}\right)^2\right]\right).$$

**Lemma.** Under regular condition, let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the continuous population pdf $f_Y(y;\theta)$. Then the Fisher Information in the random sample $Y_1, \cdots, Y_n$ equals $n$ times the Fisher information in $X$:

$$\mathbb{E}\left[\left(\frac{\partial \ln f_{Y_1,\ldots,Y_n}(Y_1,\cdots,Y_n;\theta)}{\partial\theta}\right)^2\right] = n\,\mathbb{E}\left[\left(\frac{\partial \ln f_Y(Y;\theta)}{\partial\theta}\right)^2\right] = n\,I(\theta). \quad (1)$$

(A similar statement holds for the discrete case $p_X(k;\theta)$).

Proof. Based on two observations:

$$LHS = \mathbb{E}\left[\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_{Y_i}(Y_i;\theta)\right)^2\right]$$

$$\mathbb{E}\left(\frac{\partial}{\partial\theta}\ln f_{Y_i}(Y_i;\theta)\right) = \int_{\mathbb{R}}\frac{\frac{\partial}{\partial\theta}f_Y(y;\theta)}{f_Y(y;\theta)}f_Y(y;\theta)\mathrm{d}y = \int_{\mathbb{R}}\frac{\partial}{\partial\theta}f_Y(y;\theta)\mathrm{d}y$$

$$\overset{\text{R.C.}}{=} \frac{\partial}{\partial\theta}\int_{\mathbb{R}}f_Y(y;\theta)\mathrm{d}y = \frac{\partial}{\partial\theta}1 = 0.$$

$\square$

**Lemma.** Under regular condition, let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the continuous population pdf $f_Y(y;\theta)$. Then the Fisher Information in the random sample $Y_1, \cdots, Y_n$ equals $n$ times the Fisher information in $X$:

$$\mathbb{E}\left[\left(\frac{\partial \ln f_{Y_1,\dots,Y_n}(Y_1,\cdots,Y_n;\theta)}{\partial\theta}\right)^2\right] = n\,\mathbb{E}\left[\left(\frac{\partial \ln f_Y(Y;\theta)}{\partial\theta}\right)^2\right] = n\,I(\theta). \quad (1)$$

(A similar statement holds for the discrete case $p_X(k;\theta)$).

Proof. Based on two observations:

$$LHS = \mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f_{Y_i}(Y_i;\theta)\right)^2\right]$$

$$\begin{aligned}\mathbb{E}\left(\frac{\partial}{\partial\theta} \ln f_{Y_i}(Y_i;\theta)\right) &= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta} f_Y(y;\theta)}{f_Y(y;\theta)} f_Y(y;\theta)\mathrm{d}y = \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f_Y(y;\theta)\mathrm{d}y \\ &\overset{\text{R.C.}}{=} \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f_Y(y;\theta)\mathrm{d}y = \frac{\partial}{\partial\theta} 1 = 0.\end{aligned}$$

$\square$

**Lemma.** Under regular condition, if $\ln f_Y(y;\theta)$ is twice differentiable in $\theta$, then

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta)\right]. \tag{2}$$

(A similar statement holds for the discrete case $p_X(k;\theta)$).

Proof. This is due to the two facts:

$$\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta) = \frac{\frac{\partial^2}{\partial\theta^2}f_Y(Y;\theta)}{f_Y(Y;\theta)} - \underbrace{\left(\frac{\frac{\partial}{\partial\theta}f_Y(Y;\theta)}{f_Y(Y;\theta)}\right)^2}_{=\left(\frac{\partial}{\partial\theta}\ln f_Y(Y;\theta)\right)^2}$$

$$\mathbb{E}\left(\frac{\frac{\partial^2}{\partial\theta^2}f_Y(Y;\theta)}{f_Y(Y;\theta)}\right) = \int_{\mathbb{R}}\frac{\frac{\partial^2}{\partial\theta^2}f_Y(y;\theta)}{f_Y(y;\theta)}f_Y(y;\theta)\mathrm{d}y = \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}f_Y(y;\theta)\mathrm{d}y.$$

$$\overset{R.C.}{=} \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}}f_Y(y;\theta)\mathrm{d}y = \frac{\partial^2}{\partial\theta^2}1 = 0.$$

$\square$

**Lemma.** Under regular condition, if $\ln f_Y(y;\theta)$ is twice differentiable in $\theta$, then

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta)\right]. \tag{2}$$

(A similar statement holds for the discrete case $p_X(k;\theta)$).

Proof. This is due to the two facts:

$$\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta) = \frac{\frac{\partial^2}{\partial\theta^2}f_Y(Y;\theta)}{f_Y(Y;\theta)} - \underbrace{\left(\frac{\frac{\partial}{\partial\theta}f_Y(Y;\theta)}{f_Y(Y;\theta)}\right)^2}_{=\left(\frac{\partial}{\partial\theta}\ln f_Y(Y;\theta)\right)^2}$$

$$\mathbb{E}\left(\frac{\frac{\partial^2}{\partial\theta^2}f_Y(Y;\theta)}{f_Y(Y;\theta)}\right) = \int_{\mathbb{R}}\frac{\frac{\partial^2}{\partial\theta^2}f_Y(y;\theta)}{f_Y(y;\theta)}f_Y(y;\theta)\mathrm{d}y = \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}f_Y(y;\theta)\mathrm{d}y.$$

$$\overset{R.C.}{=} \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}}f_Y(y;\theta)\mathrm{d}y = \frac{\partial^2}{\partial\theta^2}1 = 0.$$

□

**Theorem** (Cramér-Rao Inequality) Under regular condition, let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the continuous population pdf $f_Y(y;\theta)$. Let $\widehat{\theta} = \widehat{\theta}(Y_1, \cdots, Y_n)$ be any unbiased estimator for $\theta$. Then

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{n\, I(\theta)}.$$

(A similar statement holds for the discrete case $p_X(k;\theta)$).

Proof. If $n = 1$, then by Cauchy-Schwartz inequality,

$$\mathbb{E}\left[(\widehat{\theta} - \theta)\frac{\partial}{\partial\theta}\ln f_Y(Y;\theta)\right] \leq \sqrt{\mathrm{Var}(\widehat{\theta}) \times I(\theta)}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\left[(\widehat{\theta} - \theta)\frac{\partial}{\partial\theta}\ln f_Y(Y;\theta)\right] &= \int_{\mathbb{R}} (\widehat{\theta} - \theta)\frac{\frac{\partial}{\partial\theta} f_Y(y;\theta)}{f_Y(y;\theta)} f_Y(y;\theta)\mathrm{d}y \\
&= \int_{\mathbb{R}} (\widehat{\theta} - \theta)\frac{\partial}{\partial\theta} f_Y(y;\theta)\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}\underbrace{\int_{\mathbb{R}} (\widehat{\theta} - \theta) f_Y(y;\theta)\mathrm{d}y}_{= \mathbb{E}(\widehat{\theta} - \theta) = 0} + 1 = 1.
\end{aligned}$$

For general $n$, apply for (1). $\square$.

**Theorem** (Cramér-Rao Inequality) Under regular condition, let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the continuous population pdf $f_Y(y;\theta)$. Let $\widehat{\theta} = \widehat{\theta}(Y_1, \cdots, Y_n)$ be any unbiased estimator for $\theta$. Then

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{n\, I(\theta)}.$$

(A similar statement holds for the discrete case $p_X(k;\theta)$).

Proof. If $n = 1$, then by Cauchy-Schwartz inequality,

$$\mathbb{E}\left[(\widehat{\theta} - \theta)\frac{\partial}{\partial\theta}\ln f_Y(Y;\theta)\right] \leq \sqrt{\mathrm{Var}(\widehat{\theta}) \times I(\theta)}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\left[(\widehat{\theta} - \theta)\frac{\partial}{\partial\theta}\ln f_Y(Y;\theta)\right] &= \int_{\mathbb{R}} (\widehat{\theta} - \theta)\frac{\frac{\partial}{\partial\theta} f_Y(y;\theta)}{f_Y(y;\theta)} f_Y(y;\theta)\mathrm{d}y \\
&= \int_{\mathbb{R}} (\widehat{\theta} - \theta)\frac{\partial}{\partial\theta} f_Y(y;\theta)\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}\underbrace{\int_{\mathbb{R}} (\widehat{\theta} - \theta) f_Y(y;\theta)\mathrm{d}y}_{=\mathbb{E}(\widehat{\theta} - \theta) = 0} + 1 = 1.
\end{aligned}$$

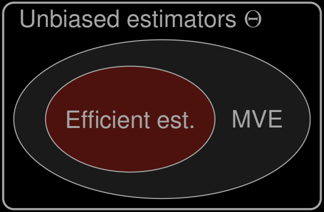
For general $n$, apply for (1). $\square$.

**Definition.** Let $\Theta$ be the set of all estimators $\widehat{\theta}$ that are unbiased for the parameter $\theta$. We say that $\widehat{\theta}^*$ is a **best** or **minimum-variance** esimator (MVE) if $\widehat{\theta}^* \in \Theta$ and

$$\mathrm{Var}(\widehat{\theta}^*) \leq \mathrm{Var}(\widehat{\theta}) \qquad \text{for all } \widehat{\theta} \in \Theta.$$

**Definition.** An unbiased estimator $\widehat{\theta}$ is **efficient** if $\mathrm{Var}(\widehat{\theta})$ is equal to the Cramér-Rao lower bound, i.e., $\mathrm{Var}\widehat{\theta} = (n\, I(\theta))^{-1}$.

The **efficiency** of an unbiased estimator $\widehat{\theta}$ is defined to be $\left(nI(\theta)\mathrm{Var}(\widehat{\theta})\right)^{-1}$.



E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k;p) = p^k(1-p)^{1-k}.$$

$$\ln p_X(k;p) = k\ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p}\ln p_X(k;p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$\frac{\partial^2}{\partial^2 p}\ln p_X(k;p) = -\frac{k}{p^2} - \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p}\ln p_X(X;p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute $\text{Var}(\widehat{p})$.

$$\text{Var}(\widehat{p}) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Conclusion. Because $\widehat{p}$ is unbiased and $\text{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k;p) = p^k(1-p)^{1-k}.$$

$$\ln p_X(k;p) = k \ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p} \ln p_X(k;p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p} \ln p_X(k;p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p} \ln p_X(X;p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute $\mathrm{Var}(\widehat{p})$.

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Conclusion. Because $\widehat{p}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k; p) = p^k (1-p)^{1-k}.$$

$$\ln p_X(k; p) = k \ln p + (1-k) \ln(1-p)$$

$$\frac{\partial}{\partial p} \ln p_X(k; p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p} \ln p_X(k; p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p} \ln p_X(X; p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute $\mathrm{Var}(\widehat{p})$.

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} npq = \frac{pq}{n}$$

Conclusion. Because $\widehat{p}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k;p) = p^k(1-p)^{1-k}.$$

$$\ln p_X(k;p) = k \ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p} \ln p_X(k;p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p} \ln p_X(k;p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p} \ln p_X(X;p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute $\mathrm{Var}(\widehat{p})$.

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} npq = \frac{pq}{n}.$$

Conclusion. Because $\widehat{p}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k; p) = p^k (1-p)^{1-k}.$$

$$\ln p_X(k; p) = k \ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p} \ln p_X(k; p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p} \ln p_X(k; p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p} \ln p_X(X; p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute Var$(\widehat{p})$.

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} npq = \frac{pq}{n}$$

Conclusion. Because $\widehat{p}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k;p) = p^k(1-p)^{1-k}.$$

$$\ln p_X(k;p) = k\ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p}\ln p_X(k;p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p}\ln p_X(k;p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p}\ln p_X(X;p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute $\text{Var}(\widehat{p})$.

$$\text{Var}(\widehat{p}) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}npq = \frac{pq}{n}.$$

Conclusion. Because $\widehat{p}$ is unbiased and $\text{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k;p) = p^k(1-p)^{1-k}.$$

$$\ln p_X(k;p) = k\ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p}\ln p_X(k;p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p}\ln p_X(k;p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p}\ln p_X(X;p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute $\mathrm{Var}(\widehat{p})$.

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}npq = \frac{pq}{n}$$

Conclusion. Because $\widehat{p}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k;p) = p^k(1-p)^{1-k}.$$

$$\ln p_X(k;p) = k \ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p}\ln p_X(k;p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p}\ln p_X(k;p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p}\ln p_X(X;p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute Var($\widehat{p}$).

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}npq = \frac{pq}{n}$$

Conclusion Because $\widehat{p}$ is unbiased and Var($\widehat{p}$) $= (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 1. $X \sim$Bernoulli(p). Check whether $\widehat{p} = \overline{X}$ is efficient?

Step 1. Compute Fisher's Information:

$$p_X(k;p) = p^k(1-p)^{1-k}.$$

$$\ln p_X(k;p) = k\ln p + (1-k)\ln(1-p)$$

$$\frac{\partial}{\partial p}\ln p_X(k;p) = \frac{k}{p} - \frac{1-k}{1-p}$$

$$-\frac{\partial^2}{\partial^2 p}\ln p_X(k;p) = \frac{k}{p^2} + \frac{1-k}{(1-p)^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 p}\ln p_X(X;p)\right] = \mathbb{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{pq}.$$

$$\boxed{I(p) = \frac{1}{pq}, \quad q = 1-p.}$$

Step 2. Compute Var($\widehat{p}$).

$$\mathrm{Var}(\widehat{p}) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}npq = \frac{pq}{n}$$

Conclusion Because $\widehat{p}$ is unbiased and Var($\widehat{p}$) $= (nI(p))^{-1}$, $\widehat{p}$ is efficient.

E.g. 2. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$\ln f_Y(y;\lambda) = \ln \lambda - \lambda y$$

$$\frac{\partial}{\partial \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda} - y$$

$$-\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(Y;\lambda)\right] = \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

E.g. 2. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$\ln f_Y(y;\lambda) = \ln \lambda - \lambda y$$

$$\frac{\partial}{\partial \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda} - y$$

$$-\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(Y;\lambda)\right] = \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

## E.g. 2. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$\ln f_Y(y;\lambda) = \ln\lambda - \lambda y$$

$$\frac{\partial}{\partial\lambda}\ln f_Y(y;\lambda) = \frac{1}{\lambda} - y$$

$$-\frac{\partial^2}{\partial^2\lambda}\ln f_Y(y;\lambda) = \frac{1}{\lambda^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\lambda}\ln f_Y(Y;\lambda)\right] = \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

## E.g. 2. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$\ln f_Y(y;\lambda) = \ln \lambda - \lambda y$$

$$\frac{\partial}{\partial \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda} - y$$

$$-\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(Y;\lambda)\right] = \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

**E.g. 2.** Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer: No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$\ln f_Y(y;\lambda) = \ln\lambda - \lambda y$$

$$\frac{\partial}{\partial\lambda}\ln f_Y(y;\lambda) = \frac{1}{\lambda} - y$$

$$-\frac{\partial^2}{\partial^2\lambda}\ln f_Y(y;\lambda) = \frac{1}{\lambda^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\lambda}\ln f_Y(Y;\lambda)\right] = \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

E.g. 2. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$\ln f_Y(y;\lambda) = \ln \lambda - \lambda y$$

$$\frac{\partial}{\partial \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda} - y$$

$$-\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(y;\lambda) = \frac{1}{\lambda^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2 \lambda} \ln f_Y(Y;\lambda)\right] = \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

E.g. 2. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$\ln f_Y(y;\lambda) = \ln\lambda - \lambda y$$

$$\frac{\partial}{\partial\lambda}\ln f_Y(y;\lambda) = \frac{1}{\lambda} - y$$

$$-\frac{\partial^2}{\partial^2\lambda}\ln f_Y(y;\lambda) = \frac{1}{\lambda^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\lambda}\ln f_Y(Y;\lambda)\right] = \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

E.g. 2. Exponential distr.: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Is $\widehat{\lambda} = 1/\overline{Y}$ efficent?

Answer No, because $\widehat{\lambda}$ is biased. Nevertheless, we can still compute Fisher's Information as follows

Fisher's Inf.

$$
\begin{aligned}
\ln f_Y(y;\lambda) &= \ln\lambda - \lambda y \\
\frac{\partial}{\partial\lambda}\ln f_Y(y;\lambda) &= \frac{1}{\lambda} - y \\
-\frac{\partial^2}{\partial^2\lambda}\ln f_Y(y;\lambda) &= \frac{1}{\lambda^2} \\
-\mathbb{E}\left[\frac{\partial^2}{\partial^2\lambda}\ln f_Y(Y;\lambda)\right] &= \mathbb{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.
\end{aligned}
$$

$$\boxed{I(\lambda) = \lambda^{-2}}$$

Try: $\widehat{\lambda}^* := \frac{n-1}{n}\frac{1}{\overline{Y}}$. It is unbiased. Is it efficient?

E.g. 2'. Exponential distr.: $f_Y(y;\theta) = \theta^{-1} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{1}{n^2}n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficent.

E.g. 2'. Exponential distr.: $f_Y(y;\theta) = \theta^{-1} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta} \ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{1}{n^2} n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficient.

**E.g. 2'.** Exponential distr.: $f_Y(y;\theta) = \theta^{-1} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{1}{n^2}n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficient.

E.g. 2'. Exponential distr.: $f_Y(y;\theta) = \theta^{-1} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta} \ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{1}{n^2} n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficient.

E.g. 2'. Exponential distr.: $f_Y(y;\theta) = \theta^{-1} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{1}{n^2}n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficient.

E.g. 2'. Exponential distr.: $f_Y(y;\theta) = \theta^{-1} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta} \ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(Y_i) = \frac{1}{n^2} n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and $\mathrm{Var}(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficient.

E.g. 2'. Exponential distr.: $f_Y(y;\theta) = \theta^{-1}e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute Var$(\widehat{\theta})$:

$$\mathrm{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{1}{n^2}n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and Var$(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficient.

E.g. 2'. Exponential distr.: $f_Y(y;\theta) = \theta^{-1} e^{-y/\theta}$ for $y \geq 0$. $\widehat{\theta} = \overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = -\ln\theta - \frac{y}{\theta}$$

$$\frac{\partial}{\partial\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta} + \frac{y}{\theta^2}$$

$$-\frac{\partial^2}{\partial^2\theta} \ln f_Y(y;\theta) = -\frac{1}{\theta^2} + \frac{2y}{\theta^3}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta} \ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{1}{\theta^2} + \frac{2Y}{\theta^3}\right] = -\frac{1}{\theta^2} + \frac{2\theta}{\theta^3} = \theta^{-2}.$$

$$\boxed{I(\theta) = \theta^{-2}}$$

Step 2. Compute Var$(\widehat{\theta})$:

$$\text{Var}(\overline{Y}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(Y_i) = \frac{1}{n^2} n\theta^2 = \frac{\theta^2}{n}.$$

Conclusion. Because $\widehat{\theta}$ is unbiased and Var$(\widehat{p}) = (nI(p))^{-1}$, $\widehat{\theta}$ is efficient.

E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = \frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: $(\dagger)$ and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} \leq \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}.$$

This is because this is not a reguler estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: $(\dagger)$ and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} < \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}.$$

This is because this is not a regular estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: $(\dagger)$ and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} < \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}.$$

This is because this is not a regular estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: $(\dagger)$ and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} < \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}.$$

This is because this is not a regular estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: ($\dagger$) and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} < \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}.$$

This is because this is not a regular estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: (†) and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} < \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}$$

This is because this is not a regular estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta} \ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: $(\dagger)$ and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} \le \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}$$

This is because this is not a regular estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta} \ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta} \ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta} \ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: ($\dagger$) and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} \leq \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}$$

This is because this is not a regular estimation!

## E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: (†) and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} \leq \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}$$

This is because this is not a regular estimation!

E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: (†) and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} \leq \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}$$

This is because this is not a regular estimation!

E.g. 3. $f_Y(y;\theta) = 2y/\theta^2$ for $y \in [0,\theta]$. $\widehat{\theta} = \frac{3}{2}\overline{Y}$ efficent?

Step. 1. Compute Fisher's Information:

$$\ln f_Y(y;\theta) = \ln(2y) - 2\ln\theta$$

$$\frac{\partial}{\partial\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta}$$

By the definition of Fisher's information,

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] = \mathbb{E}\left[\left(-\frac{2}{\theta}\right)^2\right] = \frac{4}{\theta^2}.$$

However, if we compute

$$-\frac{\partial^2}{\partial^2\theta}\ln f_Y(y;\theta) = -\frac{2}{\theta^2}$$

$$-\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\ln f_Y(Y;\theta)\right] = \mathbb{E}\left[-\frac{2}{\theta^2}\right] = -\frac{2}{\theta^2} \neq \frac{4}{\theta^2} = I(\theta). \qquad (\dagger)$$

Step 2. Compute $\mathrm{Var}(\widehat{\theta})$:

$$\mathrm{Var}(\widehat{\theta}) = \frac{9}{4n}\mathrm{Var}(Y) = \frac{9}{4n}\frac{\theta^2}{18} = \frac{\theta^2}{8n}.$$

Discussion. Even though $\widehat{\theta}$ is unbiased, we have two discripencies: (†) and

$$\mathrm{Var}(\widehat{\theta}) = \frac{\theta^2}{8n} \leq \frac{\theta^2}{4n} = \frac{1}{nI(\theta)}$$

This is because this is not a regular estimation!

# Plan

# Chapter 5. Estimation

§ 5.1 Introduction

§ 5.2 Estimating parameters: MLE and MME

§ 5.3 Interval Estimation

§ 5.4 Properties of Estimators

§ 5.5 Minimum-Variance Estimators: The Cramér-Rao Lower Bound

§ 5.6 Sufficient Estimators

§ 5.7 Consistency

§ 5.8 Bayesian Estimation

# § 5.6 Sufficient Estimators

**Rationale:** Let $\widehat{\theta}$ be an estimator to the unknown parameter $\theta$. Whether does $\widehat{\theta}$ contain all information about $\theta$?

Equivalently, how can one reduce the random sample of size $n$, denoted by $(X_1, \cdots, X_n)$, to a function without losing any information about $\theta$?

E.g., let's choose the function $h(X_1, \cdots, X_n) := \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean. In many cases, $h(X_1, \cdots, X_n)$ contains all relevant information about the true mean $\mathbb{E}(X)$. In that case, $h(X_1, \cdots, X_n)$, as an estimator, is sufficient.

# § 5.6 Sufficient Estimators

**Rationale:** Let $\widehat{\theta}$ be an estimator to the unknown parameter $\theta$. Whether does $\widehat{\theta}$ contain all information about $\theta$?

Equivalently, how can one reduce the random sample of size $n$, denoted by $(X_1, \cdots, X_n)$, to a function without losing any information about $\theta$?

E.g., let's choose the function $h(X_1, \cdots, X_n) := \frac{1}{n}\sum_{i=1}^{n} X_i$, the sample mean. In many cases, $h(X_1, \cdots, X_n)$ contains all relevant information about the true mean $\mathbb{E}(X)$. In that case, $h(X_1, \cdots, X_n)$, as an estimator, is sufficient.

# § 5.6 Sufficient Estimators

**Rationale:** Let $\widehat{\theta}$ be an estimator to the unknown parameter $\theta$. Whether does $\widehat{\theta}$ contain all information about $\theta$?

Equivalently, how can one reduce the random sample of size $n$, denoted by $(X_1, \cdots, X_n)$, to a function without losing any information about $\theta$?

E.g., let's choose the function $h(X_1, \cdots, X_n) := \frac{1}{n}\sum_{i=1}^n X_i$, the sample mean. In many cases, $h(X_1, \cdots, X_n)$ contains all relevant information about the true mean $\mathbb{E}(X)$. In that case, $h(X_1, \cdots, X_n)$, as an estimator, is sufficient.

**Definition.** Let $(X_1, \cdots, X_n)$ be a random sample of size $n$ from a discrete population with a unknown parameter $\theta$, of which $\widehat{\theta}$ (resp. $\theta_e$) be an estimator (resp. estimate). We call $\widehat{\theta}$ and $\theta_e$ **sufficient** if

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{\theta} = \theta_e\right) = b(k_1, \cdots, k_n) \qquad \text{(Sufficency-1)}$$

is a function that does not depend on $\theta$.

In case for random sample $(Y_1, \cdots, Y_n)$ from the continuous population, (Sufficency-1) should be

$$f_{Y_1,\cdots,Y_n|\widehat{\theta}=\theta_e}\left(y_1, \cdots, y_n \,\middle|\, \widehat{\theta} = \theta_e\right) = b(y_1, \cdots, y_n)$$

Note: $\widehat{\theta} = h(X_1, \cdots, X_n)$ and $\theta_e = h(k_1, \cdots, k_n)$.
or $\widehat{\theta} = h(Y_1, \cdots, Y_n)$ and $\theta_e = h(y_1, \cdots, y_n)$.

Equivalently,

**Definition.** ... $\widehat{\theta}$ (or $\theta_e$) is **sufficient** if the likelihood function can be factorized as:

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} p_X(k_i;\theta) = g(\theta_e,\theta)\; b(k_1,\cdots,k_n) & \text{Discrete} \\ \prod_{i=1}^{n} f_Y(y_i;\theta) = g(\theta_e,\theta)\; b(y_1,\cdots,y_n) & \text{Continous} \end{cases} \quad \text{(Sufficency-2)}$$

where $g$ is a function of two arguments only and $b$ is a function that does not depend on $\theta$.

E.g. 1. A random sample of size $n$ from Bernoulli(P). $\widehat{p} = \sum_{i=1}^{n} X_i$. Check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-1):

Case I: If $k_1, \cdots, k_n \in \{0,1\}$ such that $\sum_{i=1}^{n} k_i \neq c$, then

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) = 0.$$

Case II: If $k_1, \cdots, k_n \in \{0,1\}$ such that $\sum_{i=1}^{n} k_i = c$, then

E.g. 1. A random sample of size $n$ from Bernoulli(P). $\widehat{p} = \sum_{i=1}^{n} X_i$. Check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-1):

Case I: If $k_1, \cdots, k_n \in \{0,1\}$ such that $\sum_{i=1}^{n} k_i \neq c$, then

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) = 0.$$

Case II: If $k_1, \cdots, k_n \in \{0,1\}$ such that $\sum_{i=1}^{n} k_i = c$, then

E.g. 1. A random sample of size $n$ from Bernoulli(P). $\widehat{p} = \sum_{i=1}^{n} X_i$. Check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-1):

Case I: If $k_1, \cdots, k_n \in \{0, 1\}$ such that $\sum_{i=1}^{n} k_i \neq c$, then

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) = 0.$$

Case II: If $k_1, \cdots, k_n \in \{0, 1\}$ such that $\sum_{i=1}^{n} k_i = c$, then

E.g. 1. A random sample of size $n$ from Bernoulli(P). $\widehat{p} = \sum_{i=1}^{n} X_i$. Check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-1):

Case I: If $k_1, \cdots, k_n \in \{0, 1\}$ such that $\sum_{i=1}^{n} k_i \neq c$, then

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) = 0.$$

Case II: If $k_1, \cdots, k_n \in \{0, 1\}$ such that $\sum_{i=1}^{n} k_i = c$, then

$$\begin{aligned}
&\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, \widehat{p} = c\right)}{\mathbb{P}(\widehat{p} = c)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, X_n + \sum_{i=1}^{n-1} X_i = c\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_{n-1} = k_{n-1}, X_n = c - \sum_{i=1}^{n-1} k_i\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\left(\prod_{i=1}^{n-1} p^{k_i}(1-p)^{1-k_i}\right) \times p^{c - \sum_{i=1}^{n-1} k_i}(1-p)^{1-c+\sum_{i=1}^{n-1} k_i}}{\binom{n}{c} p^c (1-p)^{n-c}} \\
&\quad = \frac{1}{\binom{n}{c}}.
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, \widehat{p} = c\right)}{\mathbb{P}(\widehat{p} = c)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, X_n + \sum_{i=1}^{n-1} X_i = c\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_{n-1} = k_{n-1}, X_n = c - \sum_{i=1}^{n-1} k_i\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\left(\prod_{i=1}^{n-1} p^{k_i}(1-p)^{1-k_i}\right) \times p^{c-\sum_{i=1}^{n-1} k_i}(1-p)^{1-c+\sum_{i=1}^{n-1} k_i}}{\binom{n}{c} p^c (1-p)^{n-c}} \\
&\quad = \frac{1}{\binom{n}{c}}.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, \widehat{p} = c\right)}{\mathbb{P}(\widehat{p} = c)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, X_n + \sum_{i=1}^{n-1} X_i = c\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_{n-1} = k_{n-1}, X_n = c - \sum_{i=1}^{n-1} k_i\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\left(\prod_{i=1}^{n-1} p^{k_i}(1-p)^{1-k_i}\right) \times p^{c-\sum_{i=1}^{n-1} k_i}(1-p)^{1-c+\sum_{i=1}^{n-1} k_i}}{\binom{n}{c} p^c (1-p)^{n-c}} \\
&\quad = \frac{1}{\binom{n}{c}}.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) \\
&= \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, \widehat{p} = c\right)}{\mathbb{P}(\widehat{p} = c)} \\
&= \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, X_n + \sum_{i=1}^{n-1} X_i = c\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&= \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_{n-1} = k_{n-1}, X_n = c - \sum_{i=1}^{n-1} k_i\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&= \frac{\left(\prod_{i=1}^{n-1} p^{k_i}(1-p)^{1-k_i}\right) \times p^{c-\sum_{i=1}^{n-1} k_i}(1-p)^{1-c+\sum_{i=1}^{n-1} k_i}}{\binom{n}{c} p^c (1-p)^{n-c}} \\
&= \frac{1}{\binom{n}{c}}.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, \widehat{p} = c\right)}{\mathbb{P}(\widehat{p} = c)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, X_n + \sum_{i=1}^{n-1} X_i = c\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_{n-1} = k_{n-1}, X_n = c - \sum_{i=1}^{n-1} k_i\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\left(\prod_{i=1}^{n-1} p^{k_i}(1-p)^{1-k_i}\right) \times p^{c-\sum_{i=1}^{n-1} k_i}(1-p)^{1-c+\sum_{i=1}^{n-1} k_i}}{\binom{n}{c} p^c (1-p)^{n-c}} \\
&\quad = \frac{1}{\binom{n}{c}}.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, \widehat{p} = c\right)}{\mathbb{P}(\widehat{p} = c)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n, X_n + \sum_{i=1}^{n-1} X_i = c\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\mathbb{P}\left(X_1 = k_1, \cdots, X_{n-1} = k_{n-1}, X_n = c - \sum_{i=1}^{n-1} k_i\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = c\right)} \\
&\quad = \frac{\left(\prod_{i=1}^{n-1} p^{k_i}(1-p)^{1-k_i}\right) \times p^{c-\sum_{i=1}^{n-1} k_i}(1-p)^{1-c+\sum_{i=1}^{n-1} k_i}}{\binom{n}{c} p^c (1-p)^{n-c}} \\
&\quad = \frac{1}{\binom{n}{c}}.
\end{aligned}
$$

In summary,

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) = \begin{cases} \frac{1}{\binom{n}{c}} & \text{if } k_i \in \{0,1\} \text{ s.t. } \sum_{i=1}^{n} k_i = c, \\ 0 & \textit{otherwise}. \end{cases}$$

Hence, by (Sufficency-1), $\widehat{p} = \sum_{i=1}^{n} X_i$ is a sufficient estimator for $p$.

In summary,

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_n = k_n \,\middle|\, \widehat{p} = c\right) = \begin{cases} \frac{1}{\binom{n}{c}} & \text{if } k_i \in \{0,1\} \text{ s.t. } \sum_{i=1}^n k_i = c, \\ 0 & \textit{otherwise}. \end{cases}$$

Hence, by (Sufficency-1), $\widehat{p} = \sum_{i=1}^n X_i$ is a sufficient estimator for $p$.

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^{n} p_X(k_i; p) &= \prod_{i=1}^{n} p^{k_i} (1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^{n} k_i} (1-p)^{n - \sum_{i=1}^{n} k_i} \\
&= p^{p_e} (1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$
g(p_e, p) = p^{p_e} (1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.
$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^{n} p_X(k_i; p) &= \prod_{i=1}^{n} p^{k_i} (1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^{n} k_i} (1-p)^{n - \sum_{i=1}^{n} k_i} \\
&= p^{p_e} (1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$
g(p_e, p) = p^{p_e} (1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.
$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^{n} p_X(k_i; p) &= \prod_{i=1}^{n} p^{k_i}(1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^{n} k_i}(1-p)^{n-\sum_{i=1}^{n} k_i} \\
&= p^{p_e}(1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^n k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^n p_X(k_i; p) &= \prod_{i=1}^n p^{k_i}(1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^n k_i}(1-p)^{n-\sum_{i=1}^n k_i} \\
&= p^{p_e}(1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$
g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.
$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^{n} p_X(k_i; p) &= \prod_{i=1}^{n} p^{k_i}(1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^{n} k_i}(1-p)^{n-\sum_{i=1}^{n} k_i} \\
&= p^{p_e}(1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^{n} p_X(k_i; p) &= \prod_{i=1}^{n} p^{k_i}(1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^{n} k_i}(1-p)^{n-\sum_{i=1}^{n} k_i} \\
&= p^{p_e}(1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$
g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.
$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

**E.g. 1'.** As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^{n} p_X(k_i; p) &= \prod_{i=1}^{n} p^{k_i}(1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^{n} k_i}(1-p)^{n-\sum_{i=1}^{n} k_i} \\
&= p^{p_e}(1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$
g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.
$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$
\begin{aligned}
L(p) = \prod_{i=1}^{n} p_X(k_i; p) &= \prod_{i=1}^{n} p^{k_i}(1-p)^{1-k_i} \\
&= p^{\sum_{i=1}^{n} k_i}(1-p)^{n-\sum_{i=1}^{n} k_i} \\
&= p^{p_e}(1-p)^{n-p_e}
\end{aligned}
$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$
g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.
$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^{n} k_i$. Then

$$L(p) = \prod_{i=1}^{n} p_X(k_i; p) = \prod_{i=1}^{n} p^{k_i}(1-p)^{1-k_i}$$
$$= p^{\sum_{i=1}^{n} k_i}(1-p)^{n-\sum_{i=1}^{n} k_i}$$
$$= p^{p_e}(1-p)^{n-p_e}$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^n k_i$. Then

$$L(p) = \prod_{i=1}^n p_X(k_i; p) = \prod_{i=1}^n p^{k_i}(1-p)^{1-k_i}$$
$$= p^{\sum_{i=1}^n k_i}(1-p)^{n-\sum_{i=1}^n k_i}$$
$$= p^{p_e}(1-p)^{n-p_e}$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 1'. As in E.g. 1, check sufficiency of $\widehat{p}$ for $p$ by (Sufficency-2):

Notice that $p_e = \sum_{i=1}^n k_i$. Then

$$L(p) = \prod_{i=1}^n p_X(k_i; p) = \prod_{i=1}^n p^{k_i}(1-p)^{1-k_i}$$
$$= p^{\sum_{i=1}^n k_i}(1-p)^{n-\sum_{i=1}^n k_i}$$
$$= p^{p_e}(1-p)^{n-p_e}$$

Therefore, $p_e$ (or $\widehat{p}$) is sufficient since (Sufficency-2) is satisfied with

$$g(p_e, p) = p^{p_e}(1-p)^{n-p_e} \quad \text{and} \quad b(k_1, \cdots, k_n) = 1.$$

Comment 1. The estimator $\widehat{p}$ is sufficient but not unbiased since $\mathbb{E}(\widehat{p}) = np \neq p$.

2. Any one-to-one function of a sufficient estimator is again a sufficient estimator. E.g., $\widehat{p}_2 := \frac{1}{n}\widehat{p}$, which is a unbiased, sufficient, and MVE.

3. $\widehat{p}_3 := X_1$ is not sufficient!

E.g. 2. Poisson($\lambda$), $p_X(k;\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \cdots$. Show that $\widehat{\lambda} = (\sum_{i=1}^n X_i)^2$ is sufficient for $\lambda$ for a sample of size $n$.

Sol: The Corresponding estimate is $\lambda_e = (\sum_{i=1}^n k_i)^2$.

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
&= e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i}\left(\prod_{i=1}^n k_i!\right)^{-1} \\
&= \underbrace{e^{-n\lambda}\lambda^{\sqrt{\lambda_e}}}_{g(\lambda_e,\lambda)} \times \underbrace{\left(\prod_{i=1}^n k_i!\right)^{-1}}_{b(k_1,\cdots,k_n)}.
\end{aligned}
$$

Hence, $\widehat{\lambda}$ is sufficient estimator for $\lambda$. $\square$

E.g. 2. Poisson($\lambda$), $p_X(k;\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \cdots$. Show that $\widehat{\lambda} = (\sum_{i=1}^n X_i)^2$ is sufficient for $\lambda$ for a sample of size $n$.

Sol: The Corresponding estimate is $\lambda_e = (\sum_{i=1}^n k_i)^2$.

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
&= e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i}\left(\prod_{i=1}^n k_i!\right)^{-1} \\
&= \underbrace{e^{-n\lambda}\lambda^{\sqrt{\lambda_e}}}_{g(\lambda_e,\lambda)} \times \underbrace{\left(\prod_{i=1}^n k_i!\right)^{-1}}_{b(k_1,\cdots,k_n)}.
\end{aligned}
$$

Hence, $\widehat{\lambda}$ is sufficient estimator for $\lambda$. $\square$

E.g. 2. Poisson($\lambda$), $p_X(k;\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \cdots$. Show that $\widehat{\lambda} = (\sum_{i=1}^n X_i)^2$ is sufficient for $\lambda$ for a sample of size $n$.

Sol: The Corresponding estimate is $\lambda_e = (\sum_{i=1}^n k_i)^2$.

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
&= e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i}\left(\prod_{i=1}^n k_i!\right)^{-1} \\
&= \underbrace{e^{-n\lambda}\lambda^{\sqrt{\lambda_e}}}_{g(\lambda_e,\lambda)} \times \underbrace{\left(\prod_{i=1}^n k_i!\right)^{-1}}_{b(k_1,\cdots,k_n)}.
\end{aligned}
$$

Hence, $\widehat{\lambda}$ is sufficient estimator for $\lambda$. $\square$

E.g. 2. Poisson($\lambda$), $p_X(k;\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \cdots$. Show that $\widehat{\lambda} = (\sum_{i=1}^n X_i)^2$ is sufficient for $\lambda$ for a sample of size $n$.

Sol: The Corresponding estimate is $\lambda_e = (\sum_{i=1}^n k_i)^2$.

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
&= e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i}\left(\prod_{i=1}^n k_i!\right)^{-1} \\
&= \underbrace{e^{-n\lambda}\lambda^{\sqrt{\lambda_e}}}_{g(\lambda_e,\lambda)} \times \underbrace{\left(\prod_{i=1}^n k_i!\right)^{-1}}_{b(k_1,\cdots,k_n)}.
\end{aligned}
$$

Hence, $\widehat{\lambda}$ is sufficient estimator for $\lambda$. $\square$

E.g. 2. Poisson($\lambda$), $p_X(k;\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \cdots$. Show that $\widehat{\lambda} = (\sum_{i=1}^n X_i)^2$ is sufficient for $\lambda$ for a sample of size $n$.

Sol: The Corresponding estimate is $\lambda_e = (\sum_{i=1}^n k_i)^2$.

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
&= e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i}\left(\prod_{i=1}^n k_i!\right)^{-1} \\
&= \underbrace{e^{-n\lambda}\lambda^{\sqrt{\lambda_e}}}_{g(\lambda_e,\lambda)} \times \underbrace{\left(\prod_{i=1}^n k_i!\right)^{-1}}_{b(k_1,\cdots,k_n)}.
\end{aligned}
$$

Hence, $\widehat{\lambda}$ is sufficient estimator for $\lambda$. $\square$

E.g. 2. Poisson($\lambda$), $p_X(k;\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \cdots$. Show that $\widehat{\lambda} = (\sum_{i=1}^n X_i)^2$ is sufficient for $\lambda$ for a sample of size $n$.

Sol: The Corresponding estimate is $\lambda_e = (\sum_{i=1}^n k_i)^2$.

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\
&= e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i}\left(\prod_{i=1}^n k_i!\right)^{-1} \\
&= \underbrace{e^{-n\lambda}\lambda^{\sqrt{\lambda_e}}}_{g(\lambda_e,\lambda)} \times \underbrace{\left(\prod_{i=1}^n k_i!\right)^{-1}}_{b(k_1,\cdots,k_n)}.
\end{aligned}
$$

Hence, $\widehat{\lambda}$ is sufficient estimator for $\lambda$. $\square$

E.g. 2. Poisson($\lambda$), $p_X(k;\lambda) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, \cdots$. Show that $\widehat{\lambda} = (\sum_{i=1}^n X_i)^2$ is sufficient for $\lambda$ for a sample of size $n$.

Sol: The Corresponding estimate is $\lambda_e = (\sum_{i=1}^n k_i)^2$.

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} \\ &= e^{-n\lambda}\lambda^{\sum_{i=1}^n k_i}\left(\prod_{i=1}^n k_i!\right)^{-1} \\ &= \underbrace{e^{-n\lambda}\lambda^{\sqrt{\lambda_e}}}_{g(\lambda_e,\lambda)} \times \underbrace{\left(\prod_{i=1}^n k_i!\right)^{-1}}_{b(k_1,\cdots,k_n)}. \end{aligned}$$

Hence, $\widehat{\lambda}$ is sufficient estimator for $\lambda$. $\square$

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) &= 2^n \theta^{-2n} \left(\prod_{i=1}^{n} y_i\right) \times \prod_{i=1}^{n} I_{[0,\theta]}(y_i) \\
&= 2^n \theta^{-2n} \left(\prod_{i=1}^{n} y_i\right) \times I_{[0,\theta]}(y_{max}) \\
&= \underbrace{2^n \theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^{n} y_i}_{=b(y_1,\cdots,y_k)} .
\end{aligned}
$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. $\square$

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) &= 2^n \theta^{-2n} \left( \prod_{i=1}^{n} y_i \right) \times \prod_{i=1}^{n} I_{[0,\theta]}(y_i) \\
&= 2^n \theta^{-2n} \left( \prod_{i=1}^{n} y_i \right) \times I_{[0,\theta]}(y_{max}) \\
&= \underbrace{2^n \theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^{n} y_i}_{=b(y_1,\cdots,y_k)} .
\end{aligned}
$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. $\square$

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^n \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) &= 2^n\theta^{-2n}\left(\prod_{i=1}^n y_i\right) \times \prod_{i=1}^n I_{[0,\theta]}(y_i) \\
&= 2^n\theta^{-2n}\left(\prod_{i=1}^n y_i\right) \times I_{[0,\theta]}(y_{max}) \\
&= \underbrace{2^n\theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^n y_i}_{=b(y_1,\cdots,y_k)} .
\end{aligned}
$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. □

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$L(\theta) = \prod_{i=1}^{n} \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) = 2^n \theta^{-2n} \left( \prod_{i=1}^{n} y_i \right) \times \prod_{i=1}^{n} I_{[0,\theta]}(y_i)$$

$$= 2^n \theta^{-2n} \left( \prod_{i=1}^{n} y_i \right) \times I_{[0,\theta]}(y_{max})$$

$$= \underbrace{2^n \theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^{n} y_i}_{=b(y_1,\cdots,y_k)} .$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. □

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) &= 2^n\theta^{-2n}\left(\prod_{i=1}^{n} y_i\right) \times \prod_{i=1}^{n} I_{[0,\theta]}(y_i) \\
&= 2^n\theta^{-2n}\left(\prod_{i=1}^{n} y_i\right) \times I_{[0,\theta]}(y_{max}) \\
&= \underbrace{2^n\theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^{n} y_i}_{=b(y_1,\cdots,y_k)}.
\end{aligned}
$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. $\square$

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) &= 2^n \theta^{-2n} \left(\prod_{i=1}^{n} y_i\right) \times \prod_{i=1}^{n} I_{[0,\theta]}(y_i) \\
&= 2^n \theta^{-2n} \left(\prod_{i=1}^{n} y_i\right) \times I_{[0,\theta]}(y_{max}) \\
&= \underbrace{2^n \theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^{n} y_i}_{=b(y_1,\cdots,y_k)} .
\end{aligned}
$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. $\square$

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) &= 2^n \theta^{-2n} \left(\prod_{i=1}^{n} y_i\right) \times \prod_{i=1}^{n} I_{[0,\theta]}(y_i) \\
&= 2^n \theta^{-2n} \left(\prod_{i=1}^{n} y_i\right) \times I_{[0,\theta]}(y_{max}) \\
&= \underbrace{2^n \theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^{n} y_i}_{=b(y_1,\cdots,y_k)} .
\end{aligned}
$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. $\square$

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

E.g. 3. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$. Whether is the MLE $\widehat{\theta} = Y_{max}$ sufficient for $\theta$?

Sol: The corresponding estimate is $\theta_e = y_{max}$.

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{n} \frac{2y}{\theta^2} I_{[0,\theta]}(y_i) &= 2^n \theta^{-2n} \left( \prod_{i=1}^{n} y_i \right) \times \prod_{i=1}^{n} I_{[0,\theta]}(y_i) \\
&= 2^n \theta^{-2n} \left( \prod_{i=1}^{n} y_i \right) \times I_{[0,\theta]}(y_{max}) \\
&= \underbrace{2^n \theta^{-2n} I_{[0,\theta]}(\theta_e)}_{=g(\theta_e,\theta)} \times \underbrace{\prod_{i=1}^{n} y_i}_{=b(y_1,\cdots,y_k)} .
\end{aligned}
$$

Hence, $\widehat{\theta}$ is a sufficient estimator for $\theta$. □

Note: MME $\widehat{\theta} = \frac{3}{2}\overline{Y}$ is NOT sufficient for $\theta$!

# Plan

§ 5.1 Introduction

§ 5.2 Estimating parameters: MLE and MME

§ 5.3 Interval Estimation

§ 5.4 Properties of Estimators

§ 5.5 Minimum-Variance Estimators: The Cramér-Rao Lower Bound

§ 5.6 Sufficient Estimators

§ 5.7 Consistency

§ 5.8 Bayesian Estimation

# Chapter 5. Estimation

**Definition.** An estimator $\widehat{\theta}_n = h(W_1, \cdots, W_n)$ is said to be consistent if it converges to $\theta$ *in probability*, i.e., for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\theta}_n - \theta| < \epsilon\right) = 1.$$

Comment: In the $\epsilon$-$\delta$ language, the above convergence in probability says

$$\forall \epsilon > 0,\ \forall \delta > 0,\ \exists n(\epsilon,\delta) > 0,\ s.t.\ \forall n \geq n(\epsilon,\delta),$$

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| < \epsilon\right) > 1 - \delta.$$

**Definition.** An estimator $\widehat{\theta}_n = h(W_1, \cdots, W_n)$ is said to be consistent if it converges to $\theta$ *in probability*, i.e., for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\theta}_n - \theta| < \epsilon\right) = 1.$$

Comment: In the $\epsilon$-$\delta$ language, the above convergence in probability says

$$\forall\epsilon > 0,\ \forall\delta > 0,\ \exists n(\epsilon,\delta) > 0,\ s.t.\ \forall n \geq n(\epsilon,\delta),$$

$$\mathbb{P}\left(|\widehat{\theta}_n - \theta| < \epsilon\right) > 1 - \delta.$$

A useful tool to check convergence in probability is

**Theorem.** (Chebyshev's inequality) Let $W$ be any r.v. with finite mean $\mu$ and variance $\sigma^2$. Then for any $\epsilon > 0$

$$\mathbb{P}\left(|W-\mu|<\epsilon\right) \geq 1-\frac{\sigma^2}{\epsilon^2},$$

or, equivalently,

$$\mathbb{P}\left(|W-\mu|\geq\epsilon\right) \leq \frac{\sigma^2}{\epsilon^2}.$$

Proof. ... $\square$

As a consequence of Chebyshev's inequality, we have

**Proposition.** The sample mean $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n W_i$ is consistent for $\mathbb{E}(W) = \mu$, provided that the population $W$ has finite mean $\mu$ and variance $\sigma^2$.

Proof.

$$\mathbb{E}(\widehat{\mu}_n) = \mu \qquad \text{and} \qquad \text{Var}(\widehat{\mu}_n) = \frac{\sigma^2}{n}.$$

$$\forall \epsilon > 0, \quad \mathbb{P}\left(|\widehat{\mu}_n - \mu| \leq \epsilon\right) \geq 1 - \frac{\sigma^2}{n\epsilon^2} \to 1.$$

$\square$

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\
&= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\
&= 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

$\square$

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\ &= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\ &= 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n \\ &\to 0 \quad \text{as } n \to \infty.\end{aligned}$$

$\square$

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\
&= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\
&= 1 - \left(\frac{\theta - \epsilon}{\theta}\right)^n \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

$\square$

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\
&= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\
&= 1 - \left(\frac{\theta - \epsilon}{\theta}\right)^n \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

□

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\
&= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\
&= 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

$\square$

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\
&= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\
&= 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

$\square$

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}
\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\
&= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\
&= 1 - \left(\frac{\theta - \epsilon}{\theta}\right)^n \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

□

E.g. 1. Let $Y_1, \cdots, Y_n$ be a random sample of size $n$ from the uniform pdf $f_Y(y;\theta) = 1/\theta$, $y \in [0,\theta]$. Let $\widehat{\theta}_n = Y_{max}$. We know that $Y_{max}$ is biased. Is it consistent?

**Sol.** The c.d.f. of $Y$ is equal to $F_Y(y) = y/\theta$ for $y \in [0,\theta]$. Hence,

$$f_{Y_{max}}(y) = nF_Y(y)^{n-1} f_Y(y) = \frac{ny^{n-1}}{\theta^n}, \qquad y \in [0,\theta].$$

Therefore,

$$\begin{aligned}\mathbb{P}(|\widehat{\theta}_n - \theta| < \epsilon) &= \mathbb{P}(\theta - \epsilon < \widehat{\theta}_n < \theta + \epsilon) \\ &= \int_{\theta-\epsilon}^{\theta} \frac{ny^{n-1}}{\theta^n} \mathrm{d}y + \int_{\theta}^{\theta+\epsilon} 0 \mathrm{d}y \\ &= 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n \\ &\to 0 \quad \text{as } n \to \infty.\end{aligned}$$

$\square$

E.g. 2. Suppose $Y_1, Y_2, \cdots, Y_n$ is a random sample from the exponential pdf, $f_Y(y; \lambda) = \lambda e^{-\lambda y}$, $y > 0$. Show that $\widehat{\lambda}_n = Y_1$ is not consistent for $\lambda$.

**Sol.** To prove $\widehat{\lambda}_n$ is not consistent for $\lambda$, we need only to find out one $\epsilon > 0$ such that the following limit does not hold:

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \epsilon\right) = 1. \tag{3}$$

We can choose $\epsilon = \lambda/m$ for any $m \geq 1$. Then

$$|\widehat{\lambda}_n - \lambda| \leq \frac{\lambda}{m} \iff \left(1 - \frac{1}{m}\right)\lambda \leq \widehat{\lambda}_n \leq \left(1 + \frac{1}{m}\right)\lambda$$

$$\implies \widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda.$$

E.g. 2. Suppose $Y_1, Y_2, \cdots, Y_n$ is a random sample from the exponential pdf, $f_Y(y;\lambda) = \lambda e^{-\lambda y}$, $y > 0$. Show that $\widehat{\lambda}_n = Y_1$ is not consistent for $\lambda$.

**Sol.** To prove $\widehat{\lambda}_n$ is not consistent for $\lambda$, we need only to find out one $\epsilon > 0$ such that the following limit does not hold:

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \epsilon\right) = 1. \tag{3}$$

We can choose $\epsilon = \lambda/m$ for any $m \geq 1$. Then

$$\begin{aligned} |\widehat{\lambda}_n - \lambda| \leq \frac{\lambda}{m} \quad &\Longleftrightarrow \quad \left(1 - \frac{1}{m}\right)\lambda \leq \widehat{\lambda}_n \leq \left(1 + \frac{1}{m}\right)\lambda \\ &\Longrightarrow \quad \widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda. \end{aligned}$$

E.g. 2. Suppose $Y_1, Y_2, \cdots, Y_n$ is a random sample from the exponential pdf, $f_Y(y;\lambda) = \lambda e^{-\lambda y}$, $y > 0$. Show that $\widehat{\lambda}_n = Y_1$ is not consistent for $\lambda$.

**Sol.** To prove $\widehat{\lambda}_n$ is not consistent for $\lambda$, we need only to find out one $\epsilon > 0$ such that the following limit does not hold:

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \epsilon\right) = 1. \tag{3}$$

We can choose $\epsilon = \lambda/m$ for any $m \geq 1$. Then

$$|\widehat{\lambda}_n - \lambda| \leq \frac{\lambda}{m} \quad\Longleftrightarrow\quad \left(1 - \frac{1}{m}\right)\lambda \leq \widehat{\lambda}_n \leq \left(1 + \frac{1}{m}\right)\lambda$$

$$\Longrightarrow \quad \widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda.$$

E.g. 2. Suppose $Y_1, Y_2, \cdots, Y_n$ is a random sample from the exponential pdf, $f_Y(y; \lambda) = \lambda e^{-\lambda y}$, $y > 0$. Show that $\widehat{\lambda}_n = Y_1$ is not consistent for $\lambda$.

**Sol.** To prove $\widehat{\lambda}_n$ is not consistent for $\lambda$, we need only to find out one $\epsilon > 0$ such that the following limit does not hold:

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \epsilon\right) = 1. \tag{3}$$

We can choose $\epsilon = \lambda/m$ for any $m \geq 1$. Then

$$\begin{aligned} |\widehat{\lambda}_n - \lambda| \leq \frac{\lambda}{m} \quad &\iff \quad \left(1 - \frac{1}{m}\right)\lambda \leq \widehat{\lambda}_n \leq \left(1 + \frac{1}{m}\right)\lambda \\ &\implies \quad \widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda. \end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \frac{\lambda}{m}\right) &\leq \mathbb{P}\left(\widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \mathbb{P}\left(Y_1 \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \int_{\left(1-\frac{1}{m}\right)\lambda}^{\infty} \lambda e^{-\lambda y} \mathrm{d}y \\
&= e^{-\left(1-\frac{1}{m}\right)\lambda^2} < 1.
\end{aligned}$$

Therefore, the limit in (3) cannot hold. $\square$

Hence,

$$
\begin{aligned}
\mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \frac{\lambda}{m}\right) &\leq \mathbb{P}\left(\widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \mathbb{P}\left(Y_1 \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \int_{\left(1-\frac{1}{m}\right)\lambda}^{\infty} \lambda e^{-\lambda y} \mathrm{d}y \\
&= e^{-\left(1-\frac{1}{m}\right)\lambda^2} < 1.
\end{aligned}
$$

Therefore, the limit in (3) cannot hold. $\square$

Hence,

$$\begin{aligned}
\mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \frac{\lambda}{m}\right) &\leq \mathbb{P}\left(\widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \mathbb{P}\left(Y_1 \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \int_{\left(1-\frac{1}{m}\right)\lambda}^{\infty} \lambda e^{-\lambda y} \mathrm{d}y \\
&= e^{-\left(1-\frac{1}{m}\right)\lambda^2} < 1.
\end{aligned}$$

Therefore, the limit in (3) cannot hold. □

Hence,

$$\begin{aligned}
\mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \frac{\lambda}{m}\right) &\leq \mathbb{P}\left(\widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \mathbb{P}\left(Y_1 \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \int_{\left(1-\frac{1}{m}\right)\lambda}^{\infty} \lambda e^{-\lambda y} \mathrm{d}y \\
&= e^{-\left(1-\frac{1}{m}\right)\lambda^2} < 1.
\end{aligned}$$

Therefore, the limit in (3) cannot hold. $\square$

Hence,

$$\begin{aligned}
\mathbb{P}\left(|\widehat{\lambda}_n - \lambda| < \frac{\lambda}{m}\right) &\leq \mathbb{P}\left(\widehat{\lambda}_n \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \mathbb{P}\left(Y_1 \geq \left(1 - \frac{1}{m}\right)\lambda\right) \\
&= \int_{\left(1-\frac{1}{m}\right)\lambda}^{\infty} \lambda e^{-\lambda y} \mathrm{d}y \\
&= e^{-\left(1-\frac{1}{m}\right)\lambda^2} < 1.
\end{aligned}$$

Therefore, the limit in (3) cannot hold. $\square$

# Plan

# Chapter 5. Estimation

**Rationale:** Let $W$ be an estimator dependent on a parameter $\theta$.

1. Frequentists view $\theta$ as a parameter whose exact value is to be estimated.

2. Bayesians view $\theta$ is the value of a random variable $\Theta$.

   One can incorporate our knowledge on $\Theta$ — the **prior distribution** $p_\Theta(\theta)$ if $\Theta$ is discrete and $f_\Theta(\theta)$ if $\Theta$ is continuous — and use Bayes' formula to update our knowledge on $\Theta$ upon new observation $W = w$:

$$g_\Theta(\theta|W = w) = \begin{cases} \dfrac{p_W(w|\Theta = \theta)p_\Theta(\theta)}{\mathbb{P}(W = w)} & \text{if } W \text{ is discrete} \\[2ex] \dfrac{f_W(w|\Theta = \theta)f_\Theta(\theta)}{f_W(w)} & \text{if } W \text{ is continuous} \end{cases}$$

   where $g_\Theta(\theta|W = w)$ is called **posterior distribution** of $\Theta$.

**Rationale:** Let $W$ be an estimator dependent on a parameter $\theta$.

1. Frequentists view $\theta$ as a parameter whose exact value is to be estimated.

2. Bayesians view $\theta$ is the value of a random variable $\Theta$.

   One can incorporate our knowledge on $\Theta$ — the **prior distribution** $p_\Theta(\theta)$ if $\Theta$ is discrete and $f_\Theta(\theta)$ if $\Theta$ is continuous — and use Bayes' formula to update our knowledge on $\Theta$ upon new observation $W = w$:

$$g_\Theta(\theta|W = w) = \begin{cases} \dfrac{p_W(w|\Theta = \theta)p_\Theta(\theta)}{\mathbb{P}(W = w)} & \text{if } W \text{ is discrete} \\[2ex] \dfrac{f_W(w|\Theta = \theta)f_\Theta(\theta)}{f_W(w)} & \text{if } W \text{ is continuous} \end{cases}$$

   where $g_\Theta(\theta|W = w)$ is called **posterior distribution** of $\Theta$.

**Rationale:** Let $W$ be an estimator dependent on a parameter $\theta$.

1. Frequentists view $\theta$ as a parameter whose exact value is to be estimated.

2. Bayesians view $\theta$ is the value of a random variable $\Theta$.

   One can incorporate our knowledge on $\Theta$ — the **prior distribution** $p_\Theta(\theta)$ if $\Theta$ is discrete and $f_\Theta(\theta)$ if $\Theta$ is continuous — and use Bayes' formula to update our knowledge on $\Theta$ upon new observation $W = w$:

$$g_\Theta(\theta|W = w) = \begin{cases} \dfrac{p_W(w|\Theta = \theta)p_\Theta(\theta)}{\mathbb{P}(W = w)} & \text{if } W \text{ is discrete} \\[2ex] \dfrac{f_W(w|\Theta = \theta)f_\Theta(\theta)}{f_W(w)} & \text{if } W \text{ is continuous} \end{cases}$$

   where $g_\Theta(\theta|W = w)$ is called **posterior distribution** of $\Theta$.

**Rationale:** Let $W$ be an estimator dependent on a parameter $\theta$.

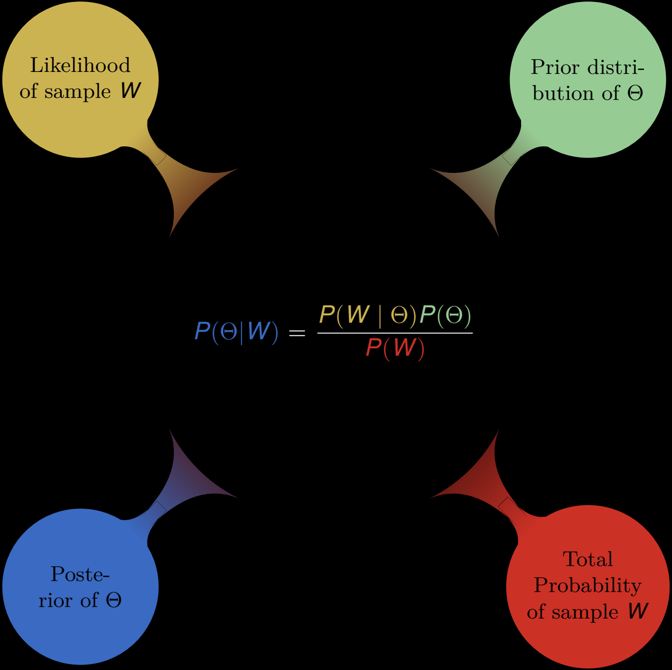
1. Frequentists view $\theta$ as a parameter whose exact value is to be estimated.

2. Bayesians view $\theta$ is the value of a random variable $\Theta$.

   One can incorporate our knowledge on $\Theta$ — the **prior distribution** $p_\Theta(\theta)$ if $\Theta$ is discrete and $f_\Theta(\theta)$ if $\Theta$ is continuous — and use Bayes' formula to update our knowledge on $\Theta$ upon new observation $W = w$:

$$g_\Theta(\theta|W = w) = \begin{cases} \dfrac{p_W(w|\Theta = \theta)p_\Theta(\theta)}{\mathbb{P}(W = w)} & \text{if } W \text{ is discrete} \\[2ex] \dfrac{f_W(w|\Theta = \theta)f_\Theta(\theta)}{f_W(w)} & \text{if } W \text{ is continuous} \end{cases}$$

   where $g_\Theta(\theta|W = w)$ is called **posterior distribution** of $\Theta$.




$$P(\Theta|W) = \frac{P(W \mid \Theta)P(\Theta)}{P(W)}$$

# Four cases for computing posterior distribution

| $g_\Theta(\theta \mid W = w)$ | $W$ discrete | $W$ continuous |
|---|---|---|
| $\Theta$ discrete | $\dfrac{p_W(w \mid \Theta = \theta) p_\Theta(\theta)}{\sum_i p_W(w \mid \Theta = \theta_i) p_\Theta(\theta_i)}$ | $\dfrac{f_W(w \mid \Theta = \theta) p_\Theta(\theta)}{\sum_i f_W(w \mid \Theta = \theta_i) p_\Theta(\theta_i)}$ |
| $\Theta$ continuous | $\dfrac{p_W(w \mid \Theta = \theta) f_\Theta(\theta)}{\int_{\mathbb{R}} p_W(w \mid \Theta = \theta') f_\Theta(\theta') \mathrm{d}\theta'}$ | $\dfrac{f_W(w \mid \Theta = \theta) f_\Theta(\theta)}{\int_{\mathbb{R}} f_W(w \mid \Theta = \theta') f_\Theta(\theta') \mathrm{d}\theta'}$ |

# Gamma distributions

$$\Gamma(r) := \int_0^\infty y^{r-1} e^{-y} \mathrm{d}y, \quad r > 0.$$

Two parametrizations for **Gamma distributions**:

1. With a shape parameter $r$ and a scale parameter $\theta$:

$$f_Y(y; r, \theta) = \frac{y^{r-1} e^{-y/\theta}}{\theta^r \Gamma(r)}, \qquad y > 0, r, \theta > 0.$$

2. With a shape parameter $r$ and a rate parameter $\lambda = 1/\theta$,

$$f_Y(y; r, \lambda) = \frac{\lambda^r y^{r-1} e^{-\lambda y}}{\Gamma(r)}, \qquad y > 0, r, \lambda > 0.$$

$$\mathbb{E}[Y] = \frac{r}{\lambda} = r\theta \quad \text{and} \quad \mathrm{Var}(Y) = \frac{r}{\lambda^2} = r\theta^2$$

# Gamma distributions

$$\Gamma(r) := \int_0^\infty y^{r-1} e^{-y} \mathrm{d}y, \quad r > 0.$$

Two parametrizations for **Gamma distributions**:

1. With a shape parameter $r$ and a scale parameter $\theta$:

$$f_Y(y; r, \theta) = \frac{y^{r-1} e^{-y/\theta}}{\theta^r \Gamma(r)}, \qquad y > 0, r, \theta > 0.$$

2. With a shape parameter $r$ and a rate parameter $\lambda = 1/\theta$,

$$f_Y(y; r, \lambda) = \frac{\lambda^r y^{r-1} e^{-\lambda y}}{\Gamma(r)}, \qquad y > 0, r, \lambda > 0.$$

$$\mathbb{E}[Y] = \frac{r}{\lambda} = r\theta \quad \text{and} \quad \mathrm{Var}(Y) = \frac{r}{\lambda^2} = r\theta^2$$

# Gamma distributions

$$\Gamma(r) := \int_0^\infty y^{r-1} e^{-y} \mathrm{d}y, \quad r > 0.$$

Two parametrizations for **Gamma distributions**:

1. With a shape parameter $r$ and a scale parameter $\theta$:

$$f_Y(y; r, \theta) = \frac{y^{r-1} e^{-y/\theta}}{\theta^r \Gamma(r)}, \qquad y > 0, r, \theta > 0.$$

2. With a shape parameter $r$ and a rate parameter $\lambda = 1/\theta$,

$$f_Y(y; r, \lambda) = \frac{\lambda^r y^{r-1} e^{-\lambda y}}{\Gamma(r)}, \qquad y > 0, r, \lambda > 0.$$

$$\mathbb{E}[Y] = \frac{r}{\lambda} = r\theta \quad \text{and} \quad \mathrm{Var}(Y) = \frac{r}{\lambda^2} = r\theta^2$$

# Gamma distributions

$$\Gamma(r) := \int_0^\infty y^{r-1} e^{-y} \mathrm{d}y, \quad r > 0.$$

Two parametrizations for **Gamma distributions**:

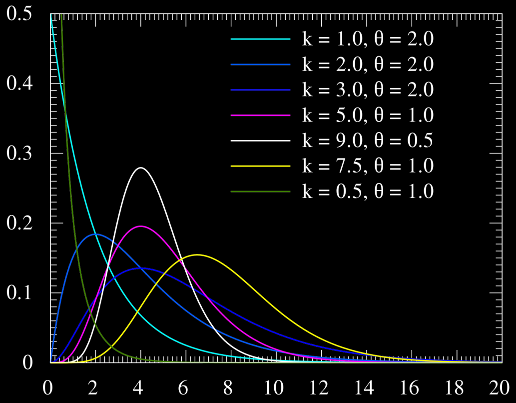
1. With a shape parameter $r$ and a scale parameter $\theta$:

$$f_Y(y; r, \theta) = \frac{y^{r-1} e^{-y/\theta}}{\theta^r \Gamma(r)}, \qquad y > 0, r, \theta > 0.$$

2. With a shape parameter $r$ and a rate parameter $\lambda = 1/\theta$,

$$f_Y(y; r, \lambda) = \frac{\lambda^r y^{r-1} e^{-\lambda y}}{\Gamma(r)}, \qquad y > 0, r, \lambda > 0.$$

$$\mathbb{E}[Y] = \frac{r}{\lambda} = r\theta \quad \text{and} \quad \mathrm{Var}(Y) = \frac{r}{\lambda^2} = r\theta^2$$




```r
# Plot  gamma distributions
x = seq(0,20,0.01)
k= 3 # Shape parameter
theta = 0.5 # Scale parameter
plot(x,dgamma(x, k, scale = theta),
      type="l",
      col="red")
```

# Beta distributions

$$B(\alpha, \beta) := \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\mathrm{d}y, \quad \alpha, \beta > 0.$$

$$\vdots \qquad \vdots$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$ (see Appendix)

Beta distribution

$$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}, \quad y \in [0,1], \alpha, \beta > 0.$$

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

# Beta distributions

$$B(\alpha, \beta) := \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\mathrm{d}y, \quad \alpha, \beta > 0.$$

$$\vdots \qquad \vdots$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \qquad \text{(see Appendix)}$$

Beta distribution

$$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}, \quad y \in [0,1], \alpha, \beta > 0.$$

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

# Beta distributions

$$B(\alpha, \beta) := \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\mathrm{d}y, \quad \alpha, \beta > 0.$$

$$\vdots \qquad \vdots$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

(see Appendix)

Beta distribution

$$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}, \quad y \in [0, 1], \alpha, \beta > 0.$$

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

# Beta distributions

$$B(\alpha, \beta) := \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\mathrm{d}y, \quad \alpha, \beta > 0.$$

$$\vdots \qquad \vdots$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \qquad \text{(see Appendix)}$$

Beta distribution

$$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}, \quad y \in [0,1], \alpha, \beta > 0.$$

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

# Beta distributions

$$B(\alpha, \beta) := \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\mathrm{d}y, \quad \alpha, \beta > 0.$$

$$\vdots \qquad \vdots$$

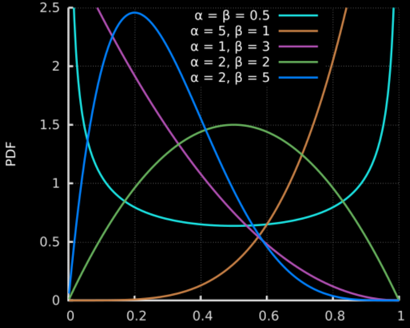
$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

(see Appendix)

Beta distribution

$$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)}, \quad y \in [0, 1], \alpha, \beta > 0.$$

$$\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$




```
    # Plot Beta distributions
x = seq(0,1,0.01)
a = 13
b = 2
plot (x,dbeta(x,a,b),
     type="l",
     col="red")
```

E.g. 1. Let $X_1, \cdots, X_n$ be a random sample from Bernoulli$(\theta)$: $p_{X_i}(k;\theta) = \theta^k(1-\theta)^{1-k}$ for $k = 0, 1$.

Let $X = \sum_{i=1}^n X_i$. Then $X$ follows binomial$(n, \theta)$.

Prior distribution: $\Theta \sim$beta$(r, s)$, i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$ for $\theta \in [0, 1]$.

$$\begin{aligned} X_1, \cdots, X_n \,\Big|\, \theta \quad &\sim \quad \text{Bernoulli}(\theta) \\ \Theta \quad &\sim \quad \text{Beta}(r, s) \\ & \quad\;\; r \,\&\, s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^n X_i \,\Big|\, \theta \quad &\sim \quad \text{Binomial}(n, \theta) \\ \Theta \quad &\sim \quad \text{Beta}(r, s) \\ & \quad\;\; r \,\&\, s \text{ are known} \end{aligned}$$

E.g. 1. Let $X_1, \cdots, X_n$ be a random sample from Bernoulli$(\theta)$: $p_{X_i}(k;\theta) = \theta^k(1-\theta)^{1-k}$ for $k = 0, 1$.

Let $X = \sum_{i=1}^n X_i$. Then $X$ follows binomial$(n, \theta)$.

Prior distribution: $\Theta \sim$beta$(r, s)$, i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$ for $\theta \in [0, 1]$.

$$\begin{aligned} X_1, \cdots, X_n \,\Big|\, \theta \quad &\sim \quad \text{Bernoulli}(\theta) \\ \Theta \quad &\sim \quad \text{Beta}(r, s) \\ & \quad\;\; r \;\&\; s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^n X_i \,\Big|\, \theta \quad &\sim \quad \text{Binomial}(n, \theta) \\ \Theta \quad &\sim \quad \text{Beta}(r, s) \\ & \quad\;\; r \;\&\; s \text{ are known} \end{aligned}$$

E.g. 1. Let $X_1, \cdots, X_n$ be a random sample from Bernoulli$(\theta)$: $p_{X_i}(k;\theta) = \theta^k (1-\theta)^{1-k}$ for $k = 0, 1$.

Let $X = \sum_{i=1}^n X_i$. Then $X$ follows binomial$(n, \theta)$.

Prior distribution: $\Theta \sim$beta$(r, s)$, i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \theta^{r-1} (1-\theta)^{s-1}$ for $\theta \in [0, 1]$.

$$\begin{aligned} X_1, \cdots, X_n \Big| \theta &\sim \text{Bernoulli}(\theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \ \&\ s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^n X_i \Bigg| \theta &\sim \text{Binomial}(n, \theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \ \&\ s \text{ are known} \end{aligned}$$

E.g. 1. Let $X_1, \cdots, X_n$ be a random sample from Bernoulli$(\theta)$: $p_{X_i}(k;\theta) = \theta^k(1-\theta)^{1-k}$ for $k = 0, 1$.

Let $X = \sum_{i=1}^n X_i$. Then $X$ follows binomial$(n, \theta)$.

Prior distribution: $\Theta \sim$beta$(r, s)$, i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$ for $\theta \in [0, 1]$.

$$\begin{aligned} X_1, \cdots, X_n \,\Big|\, \theta &\sim \text{Bernoulli}(\theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \,\&\, s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^n X_i \,\Bigg|\, \theta &\sim \text{Binomial}(n, \theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \,\&\, s \text{ are known} \end{aligned}$$

E.g. 1. Let $X_1, \cdots, X_n$ be a random sample from Bernoulli$(\theta)$: $p_{X_i}(k;\theta) = \theta^k(1-\theta)^{1-k}$ for $k = 0, 1$.

Let $X = \sum_{i=1}^n X_i$. Then $X$ follows binomial$(n,\theta)$.

Prior distribution: $\Theta \sim$beta$(r,s)$, i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$ for $\theta \in [0,1]$.

$$
\begin{aligned}
X_1, \cdots, X_n \,\big|\, \theta &\sim \text{Bernoulli}(\theta) \\
\Theta &\sim \text{Beta}(r,s) \\
& \quad r \,\&\, s \text{ are known}
\end{aligned}
\qquad
\begin{aligned}
X = \sum_{i=1}^n X_i \,\Big|\, \theta &\sim \text{Binomial}(n,\theta) \\
\Theta &\sim \text{Beta}(r,s) \\
& \quad r \,\&\, s \text{ are known}
\end{aligned}
$$

E.g. 1. Let $X_1, \cdots, X_n$ be a random sample from Bernoulli$(\theta)$: $p_{X_i}(k;\theta) = \theta^k(1-\theta)^{1-k}$ for $k = 0, 1$.

Let $X = \sum_{i=1}^n X_i$. Then $X$ follows binomial$(n, \theta)$.

Prior distribution: $\Theta \sim$beta$(r, s)$, i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$ for $\theta \in [0, 1]$.

$$\begin{aligned} X_1, \cdots, X_n \,\middle|\, \theta &\sim \text{Bernoulli}(\theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \text{ \& } s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^n X_i \,\Big|\, \theta &\sim \text{Binomial}(n, \theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \text{ \& } s \text{ are known} \end{aligned}$$

**Example 5.8.2**

Max, a video game pirate (and Bayesian), is trying to decide how many illegal copies of *Zombie Beach Party* to have on hand for the upcoming holiday season. To get a rough idea of what the demand might be, he talks with $n$ potential customers and finds that $X = k$ would buy a copy for a present (or for themselves). The obvious choice for a probability model for $X$, of course, would be the binomial pdf. Given $n$ potential customers, the probability that $k$ would actually buy one of Max's illegal copies is the familiar

$$p_X(k \mid \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}, \quad k = 0, 1, \ldots, n$$

where the maximum likelihood estimate for $\theta$ is given by $\theta_e = \frac{k}{n}$.

It may very well be the case, though, that Max has some additional insight about the value of $\theta$ on the basis of similar video games that he illegally marketed in previous years. Suppose he suspects, for example, that the percentage of potential customers who will buy *Zombie Beach Party* is likely to be between 3% and 4% and probably will not exceed 7%. A reasonable prior distribution for $\Theta$, then, would be a pdf mostly concentrated over the interval 0 to 0.07 with a mean or median in the 0.035 range.

One such probability model whose shape would comply with the restraints that Max is imposing is the *beta pdf*. Written with $\Theta$ as the random variable, the (two-parameter) beta pdf is given by

$$f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \theta^{r-1}(1-\theta)^{s-1}, \quad 0 \leq \theta \leq 1$$

The beta distribution with $r = 2$ and $s = 4$ is pictured in Figure 5.8.1. By choosing different values for $r$ and $s$, $f_\Theta(\theta)$ can be skewed more sharply to the right or to the left, and the bulk of the distribution can be concentrated close to zero or close to one. The question is, if an appropriate beta pdf is used as a *prior* distribution for $\Theta$, and if a random sample of $k$ potential customers (out of $n$) said they would buy the video game, what would be a reasonable *posterior* distribution for $\Theta$?

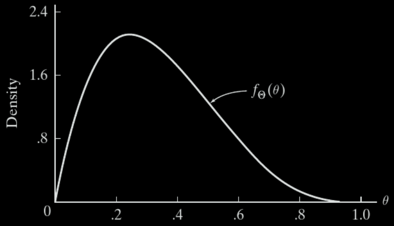



**Figure 5.8.1**

$X$ is discrete and $\Theta$ is continuous.

$$g_\Theta(\theta|X=k) = \frac{p_X(k|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_X(k|\Theta=\theta)f_\Theta(\theta) &= \binom{n}{k}\theta^k(1-\theta)^{n-k} \times \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1} \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \quad \theta \in [0,1]. \end{aligned}$$

$$\begin{aligned} p_X(k) &= \int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^1 \theta'^{k+r-1}(1-\theta')^{n-k+s-1}\mathrm{d}\theta' \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))} \end{aligned}$$

$X$ is discrete and $\Theta$ is continuous.

$$g_\Theta(\theta|X=k) = \frac{p_X(k|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned}
p_X(k|\Theta=\theta)f_\Theta(\theta) &= \binom{n}{k}\theta^k(1-\theta)^{n-k} \times \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1} \\
&= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \quad \theta \in [0,1].
\end{aligned}$$

$$\begin{aligned}
p_X(k) &= \int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\
&= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^1 \theta'^{k+r-1}(1-\theta')^{n-k+s-1}\mathrm{d}\theta' \\
&= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))}
\end{aligned}$$

$X$ is discrete and $\Theta$ is continuous.

$$g_\Theta(\theta|X=k) = \frac{p_X(k|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_X(k|\Theta=\theta)f_\Theta(\theta) &= \binom{n}{k}\theta^k(1-\theta)^{n-k} \times \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1} \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \quad \theta \in [0,1]. \end{aligned}$$

$$\begin{aligned} p_X(k) &= \int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^1 \theta'^{k+r-1}(1-\theta')^{n-k+s-1}\mathrm{d}\theta' \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))} \end{aligned}$$

$X$ is discrete and $\Theta$ is continuous.

$$g_\Theta(\theta|X=k) = \frac{p_X(k|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_X(k|\Theta=\theta)f_\Theta(\theta) &= \binom{n}{k}\theta^k(1-\theta)^{n-k} \times \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1} \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \quad \theta \in [0,1]. \end{aligned}$$

$$\begin{aligned} p_X(k) &= \int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^1 \theta'^{k+r-1}(1-\theta')^{n-k+s-1}\mathrm{d}\theta' \\ &= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))} \end{aligned}$$

$X$ is discrete and $\Theta$ is continuous.

$$g_\Theta(\theta | X = k) = \frac{p_X(k|\Theta = \theta) f_\Theta(\theta)}{\int_{\mathbb{R}} p_X(k|\Theta = \theta') f_\Theta(\theta') \mathrm{d}\theta'}$$

$$\begin{aligned} p_X(k|\Theta = \theta) f_\Theta(\theta) &= \binom{n}{k} \theta^k (1-\theta)^{n-k} \times \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \theta^{r-1} (1-\theta)^{s-1} \\ &= \binom{n}{k} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \theta^{k+r-1} (1-\theta)^{n-k+s-1}, \quad \theta \in [0,1]. \end{aligned}$$

$$\begin{aligned} p_X(k) &= \int_{\mathbb{R}} p_X(k|\Theta = \theta') f_\Theta(\theta') \mathrm{d}\theta' \\ &= \binom{n}{k} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \int_0^1 \theta'^{k+r-1} (1-\theta')^{n-k+s-1} \mathrm{d}\theta' \\ &= \binom{n}{k} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))} \end{aligned}$$

$X$ is discrete and $\Theta$ is continuous.

$$g_\Theta(\theta|X=k) = \frac{p_X(k|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned}
p_X(k|\Theta=\theta)f_\Theta(\theta) &= \binom{n}{k}\theta^k(1-\theta)^{n-k} \times \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1} \\
&= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \quad \theta \in [0,1].
\end{aligned}$$

$$\begin{aligned}
p_X(k) &= \int_{\mathbb{R}} p_X(k|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\
&= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\int_0^1 \theta'^{k+r-1}(1-\theta')^{n-k+s-1}\mathrm{d}\theta' \\
&= \binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))}
\end{aligned}$$

$$g_\Theta(\theta|X=k) = \frac{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \theta^{k+r-1}(1-\theta)^{n-k+s-1}}{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))}}$$

$$= \frac{\Gamma(n+r+s)}{\Gamma(k+r)\Gamma(n-k+s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \qquad \theta \in [0,1]$$

Conclusion: the posterior $\sim$ beta distribution$(k+r, n-k+s)$.

Recall that the prior $\sim$ beta distribution$(r, s)$.

$$
\begin{aligned}
g_\Theta(\theta|X=k) &= \frac{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \theta^{k+r-1}(1-\theta)^{n-k+s-1}}{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))}} \\
&= \frac{\Gamma(n+r+s)}{\Gamma(k+r)\Gamma(n-k+s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \qquad \theta \in [0,1]
\end{aligned}
$$

Conclusion: the posterior $\sim$ beta distribution$(k+r, n-k+s)$.

Recall that the prior $\sim$ beta distribution$(r, s)$.

$$
\begin{aligned}
g_\Theta(\theta|X=k) &= \frac{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \theta^{k+r-1}(1-\theta)^{n-k+s-1}}{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \times \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma((k+r)+(n-k+s))}} \\
&= \frac{\Gamma(n+r+s)}{\Gamma(k+r)\Gamma(n-k+s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}, \qquad \theta \in [0,1]
\end{aligned}
$$

Conclusion: the posterior $\sim$ beta distribution$(k+r, n-k+s)$.

Recall that the prior $\sim$ beta distribution$(r, s)$.

It remains to determine the values of $r$ and $s$ to incorporate the prior knowledge:

PK 1. Mean is about 0.035.

$$\mathbb{E}(\Theta) = 0.035 \quad \Longrightarrow \quad \frac{r}{r+s} = 0.035 \quad \Longleftrightarrow \quad \frac{r}{s} = \frac{7}{193}$$

PK 2. The pdf mostly concentrated over $[0, 0.07]$. ... trial ...

It remains to determine the values of $r$ and $s$ to incorporate the prior knowledge:

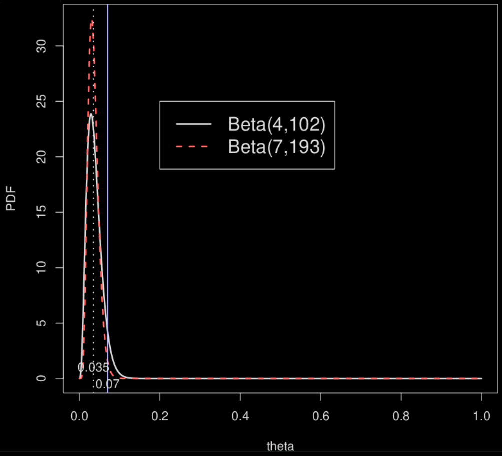
PK 1. Mean is about 0.035.

$$\mathbb{E}(\Theta) = 0.035 \quad \Longrightarrow \quad \frac{r}{r+s} = 0.035 \quad \Longleftrightarrow \quad \frac{r}{s} = \frac{7}{193}$$

PK 2. The pdf mostly concentrated over $[0, 0.07]$. ... trial ...

```r
x <- seq(0, 1, length = 1025)
plot(x,dbeta(x,4,102),
      type="l")
plot(x,dbeta(x,7,193),
      type="l")
dev.off()

pdf=cbind(dbeta(x,4,102),dbeta(x,7,193))
matplot(x,pdf,
        type="l",
         lty = 1:2,
        xlab = "theta", ylab = "PDF",
        lwd = 2 # Line width
         )
legend(0.2, 25, # Position of legend
       c("Beta(4,102)", "Beta(7,193)"),
       col = 1:2, lty = 1:2,
       ncol = 1, # Number of columns
       cex = 1.5, # Fontsize
       lwd=2 # Line width
       )
abline(v=0.07, col="blue", lty=1,lwd=1.5)
text(0.07, -0.5, "0.07")
abline(v=0.035, col="gray60", lty=3,lwd=2)
text(0.035, 1, "0.035")
```



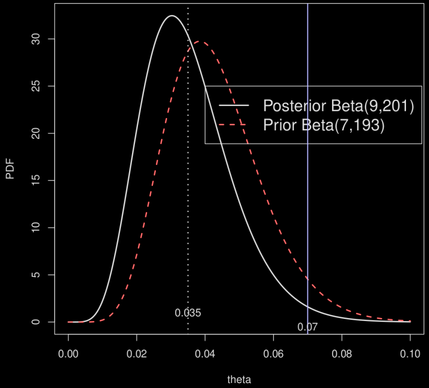
If we choose $r = 7$ and $s = 193$:

$$g_\Theta(\theta|X = k) = \frac{\Gamma(n + 200)}{\Gamma(k + 7)\Gamma(n - k + 193)} \theta^{k+6}(1 - \theta)^{n-k+192}, \qquad \theta \in [0, 1]$$

Moreover, if $n = 10$ and $k = 2$,

$$g_\Theta(\theta|X = k) = \frac{\Gamma(210)}{\Gamma(9)\Gamma(201)} \theta^8(1 - \theta)^{200}, \qquad \theta \in [0, 1]$$

```r
x <- seq(0, 0.1, length = 1025)
pdf=cbind(dbeta(x,7,193),dbeta(x,9,201))
matplot(x,pdf,
          type="l",
          lty  = 1:2,
          xlab = "theta", ylab = "PDF",
          lwd = 2 # Line width
)
legend(0.05, 25, # Position of legend
        c("Posterior Beta(9,201)", "Prior
            Beta(7,193)"),
        col = 1:2, lty = 1:2,
        ncol = 1, # Number of columns
        cex = 1.5, # Fontsize
        lwd=2 # Line width
)
abline(v=0.07,col="blue", lty=1,lwd=1.5)
text(0.07, -0.5, "0.07")
abline(v=0.035,col="black", lty=3,lwd=2)
text(0.035, 1, "0.035")
```



**Definition.** If the posterior distributions $p(\Theta|X)$ are in the same probability distribution family as the prior probability distribution $p(\Theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

1. Beta distributions are conjugate priors for Bernoulli, binomial, nega. binomial, geometric likelihood.
2. Gamma distributions are conjugate priors for Poisson and exponential likelihood.

**Definition.** If the posterior distributions $p(\Theta|X)$ are in the same probability distribution family as the prior probability distribution $p(\Theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

1. Beta distributions are conjugate priors for Bernoulli, binomial, nega. binomial, geometric likelihood.

2. Gamma distributions are conjugate priors for Poisson and exponential likelihood.

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample from Poisson$(\theta)$: $p_X(k;\theta) = \frac{e^{-\theta}\theta^k}{k!}$ for $k = 0, 1, \cdots$.

Let $W = \sum_{i=1}^{n} X_i$. Then $W$ follows Poisson$(n\theta)$.

Prior distribution: $\Theta \sim$ Gamma$(s,\mu)$, i.e., $f_\Theta(\theta) = \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta}$ for $\theta > 0$.

$$
\begin{aligned}
X_1, \cdots, X_n \,\big|\, \theta &\sim \text{Poisson}(\theta) \\
\Theta &\sim \text{Gamma}(s,\mu) \\
& \quad s \text{ \& } \mu \text{ are known}
\end{aligned}
\qquad\qquad
\begin{aligned}
W = \sum_{i=1}^{n} X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\
\Theta &\sim \text{Gamma}(s,\mu) \\
& \quad s \text{ \& } \mu \text{ are known}
\end{aligned}
$$

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample from Poisson$(\theta)$: $p_X(k;\theta) = \frac{e^{-\theta}\theta^k}{k!}$ for $k = 0, 1, \cdots$.

Let $W = \sum_{i=1}^n X_i$. Then $W$ follows Poisson$(n\theta)$.

Prior distribution: $\Theta \sim$ Gamma$(s, \mu)$, i.e., $f_\Theta(\theta) = \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta}$ for $\theta > 0$.

$$\begin{aligned} X_1, \cdots, X_n \,\big|\, \theta &\sim \text{Poisson}(\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \,\&\, \mu \text{ are known} \end{aligned}$$

$$\begin{aligned} W = \sum_{i=1}^n X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \,\&\, \mu \text{ are known} \end{aligned}$$

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample from Poisson$(\theta)$: $p_X(k;\theta) = \frac{e^{-\theta}\theta^k}{k!}$ for $k = 0, 1, \cdots$.

Let $W = \sum_{i=1}^n X_i$. Then $W$ follows Poisson$(n\theta)$.

Prior distribution: $\Theta \sim$ Gamma$(s, \mu)$, i.e., $f_\Theta(\theta) = \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta}$ for $\theta > 0$.

$$\begin{aligned} X_1, \cdots, X_n \,\Big|\, \theta &\sim \text{Poisson}(\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & s \,\&\, \mu \text{ are known} \end{aligned}$$

$$\begin{aligned} W = \sum_{i=1}^n X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & s \,\&\, \mu \text{ are known} \end{aligned}$$

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample from Poisson$(\theta)$: $p_X(k;\theta) = \frac{e^{-\theta}\theta^k}{k!}$ for $k = 0, 1, \cdots$.

Let $W = \sum_{i=1}^n X_i$. Then $W$ follows Poisson$(n\theta)$.

Prior distribution: $\Theta \sim$ Gamma$(s,\mu)$, i.e., $f_\Theta(\theta) = \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta}$ for $\theta > 0$.

$$\begin{aligned} X_1,\cdots,X_n \,\big|\, \theta &\sim \text{Poisson}(\theta) \\ \Theta &\sim \text{Gamma}(s,\mu) \\ & \quad s \text{ \& } \mu \text{ are known} \end{aligned}$$

$$\begin{aligned} W = \sum_{i=1}^n X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\ \Theta &\sim \text{Gamma}(s,\mu) \\ & \quad s \text{ \& } \mu \text{ are known} \end{aligned}$$

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample from Poisson$(\theta)$: $p_X(k;\theta) = \frac{e^{-\theta}\theta^k}{k!}$ for $k = 0, 1, \cdots$.

Let $W = \sum_{i=1}^n X_i$. Then $W$ follows Poisson$(n\theta)$.

Prior distribution: $\Theta \sim$ Gamma$(s, \mu)$, i.e., $f_\Theta(\theta) = \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta}$ for $\theta > 0$.

$$\begin{aligned} X_1, \cdots, X_n \,\big|\, \theta &\sim \text{Poisson}(\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \,\&\, \mu \text{ are known} \end{aligned}$$

$$\begin{aligned} W = \sum_{i=1}^n X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \,\&\, \mu \text{ are known} \end{aligned}$$

E.g. 2. Let $X_1, \cdots, X_n$ be a random sample from Poisson$(\theta)$: $p_X(k;\theta) = \frac{e^{-\theta}\theta^k}{k!}$ for $k = 0, 1, \cdots$.

Let $W = \sum_{i=1}^n X_i$. Then $W$ follows Poisson$(n\theta)$.

Prior distribution: $\Theta \sim$ Gamma$(s, \mu)$, i.e., $f_\Theta(\theta) = \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta}$ for $\theta > 0$.

$$
\begin{aligned}
X_1, \cdots, X_n \,\big|\, \theta \quad &\sim \quad \text{Poisson}(\theta) \\
\Theta \quad &\sim \quad \text{Gamma}(s, \mu) \\
& \qquad s \;\&\; \mu \text{ are known}
\end{aligned}
$$

$$
\begin{aligned}
W = \sum_{i=1}^n X_i \,\Big|\, \theta \quad &\sim \quad \text{Poisson}(n\theta) \\
\Theta \quad &\sim \quad \text{Gamma}(s, \mu) \\
& \qquad s \;\&\; \mu \text{ are known}
\end{aligned}
$$

$$g_\Theta(\theta|W=w) = \frac{p_W(w|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_W(w|\Theta=\theta)f_\Theta(\theta) &= \frac{e^{-n\theta}(n\theta)^w}{w!} \times \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta} \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \theta^{w+s-1}e^{-(\mu+n)\theta}, \quad \theta > 0. \end{aligned}$$

$$\begin{aligned} p_W(w) &= \int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)}\int_0^\infty \theta'^{w+s-1}e^{-(\mu+n)\theta'}\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \frac{\Gamma(w+s)}{(\mu+n)^{w+s}} \end{aligned}$$

$$g_\Theta(\theta|W=w) = \frac{p_W(w|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_W(w|\Theta=\theta)f_\Theta(\theta) &= \frac{e^{-n\theta}(n\theta)^w}{w!} \times \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta} \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \theta^{w+s-1}e^{-(\mu+n)\theta}, \quad \theta > 0. \end{aligned}$$

$$\begin{aligned} p_W(w) &= \int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)}\int_0^\infty \theta'^{w+s-1}e^{-(\mu+n)\theta'}\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \frac{\Gamma(w+s)}{(\mu+n)^{w+s}} \end{aligned}$$

$$g_\Theta(\theta|W = w) = \frac{p_W(w|\Theta = \theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_W(w|\Theta = \theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_W(w|\Theta = \theta)f_\Theta(\theta) &= \frac{e^{-n\theta}(n\theta)^w}{w!} \times \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta} \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \theta^{w+s-1}e^{-(\mu+n)\theta}, \quad \theta > 0. \end{aligned}$$

$$\begin{aligned} p_W(w) &= \int_{\mathbb{R}} p_W(w|\Theta = \theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)}\int_0^\infty \theta'^{w+s-1}e^{-(\mu+n)\theta'}\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \frac{\Gamma(w+s)}{(\mu+n)^{w+s}} \end{aligned}$$

$$g_\Theta(\theta|W=w) = \frac{p_W(w|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_W(w|\Theta=\theta)f_\Theta(\theta) &= \frac{e^{-n\theta}(n\theta)^w}{w!} \times \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta} \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \theta^{w+s-1}e^{-(\mu+n)\theta}, \quad \theta > 0. \end{aligned}$$

$$\begin{aligned} p_W(w) &= \int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)}\int_0^\infty \theta'^{w+s-1}e^{-(\mu+n)\theta'}\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \frac{\Gamma(w+s)}{(\mu+n)^{w+s}} \end{aligned}$$

$$g_\Theta(\theta|W=w) = \frac{p_W(w|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_W(w|\Theta=\theta)f_\Theta(\theta) &= \frac{e^{-n\theta}(n\theta)^w}{w!} \times \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta} \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \theta^{w+s-1}e^{-(\mu+n)\theta}, \quad \theta > 0. \end{aligned}$$

$$\begin{aligned} p_W(w) &= \int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)}\int_0^\infty \theta'^{w+s-1}e^{-(\mu+n)\theta'}\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \frac{\Gamma(w+s)}{(\mu+n)^{w+s}} \end{aligned}$$

$$g_\Theta(\theta|W=w) = \frac{p_W(w|\Theta=\theta)f_\Theta(\theta)}{\int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta'}$$

$$\begin{aligned} p_W(w|\Theta=\theta)f_\Theta(\theta) &= \frac{e^{-n\theta}(n\theta)^w}{w!} \times \frac{\mu^s}{\Gamma(s)}\theta^{s-1}e^{-\mu\theta} \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \theta^{w+s-1}e^{-(\mu+n)\theta}, \quad \theta > 0. \end{aligned}$$

$$\begin{aligned} p_W(w) &= \int_{\mathbb{R}} p_W(w|\Theta=\theta')f_\Theta(\theta')\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)}\int_0^\infty \theta'^{w+s-1}e^{-(\mu+n)\theta'}\mathrm{d}\theta' \\ &= \frac{n^w}{w!}\frac{\mu^s}{\Gamma(s)} \times \frac{\Gamma(w+s)}{(\mu+n)^{w+s}} \end{aligned}$$

$$
\begin{aligned}
g_\Theta(\theta|X=k) &= \frac{\dfrac{n^w}{w!}\dfrac{\mu^s}{\Gamma(s)} \times \theta^{w+s-1} e^{-(\mu+n)\theta}}{\dfrac{n^w}{w!}\dfrac{\mu^s}{\Gamma(s)} \times \dfrac{\Gamma(w+s)}{(\mu+n)^{w+s}}} \\
&= \frac{(\mu+n)^{w+s}}{\Gamma(w+s)} \theta^{w+s-1} e^{-(\mu+n)\theta}, \qquad \theta > 0.
\end{aligned}
$$

Conclusion: the posterior of $\Theta \sim$ gamma distribution$(w+s, n+\mu)$.

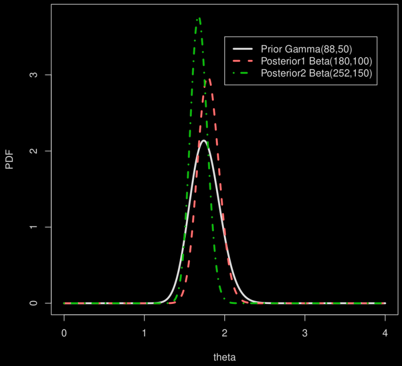
Recall that the prior of $\Theta \sim$ gamma distribution$(s, \mu)$.

# Case Study 5.8.1

```r
x <- seq(0, 4, length = 1025)
pdf=cbind(dgamma(x, shape=88, rate=50),
            dgamma(x, shape=88+92, 100),
            dgamma(x, 88+92+72, 150))
matplot(x,pdf,
         type="l",
          lty = 1:3,
         xlab = "theta",  ylab = "PDF",
         lwd = 2 # Line width
)
legend(2, 3.5, # Position of legend
        c("Prior Gamma(88,50)",
          "Posterior1 Beta(180,100)",
          "Posterior2 Beta(252,150)"),
        col = 1:3,  lty = 1:3,
        ncol = 1, # Number of columns
        cex = 1.5, # Fontsize
        lwd=2 # Line width
)
```

**Table 5.8.1**

| Years | Number of Hurricanes |
|---|---|
| 1851–1900 | 88 |
| 1901–1950 | 92 |
| 1951–2000 | 72 |



# Bayesian Point Estimation

**Question.** Can one calculate an appropriate *point estimate* $\theta_e$ given the posterior $g_\Theta(\theta|W = w)$?

**Definitions.** Let $\theta_e$ be an estimate for $\theta$ based on a statistic $W$. The loss function associated with $\theta_e$ is denoted $L(\theta_e, \theta)$, where $L(\theta_e, \theta) \geq 0$ and $L(\theta, \theta) = 0$.

Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$. Then the risk associated with $\widehat{\theta}$ is the expected value of the loss function with respect to the posterior distribution of $\Theta$:

$$\text{risk} = \begin{cases} \displaystyle\int_{\mathbb{R}} L(\widehat{\theta}, \theta) g_\Theta(\theta|W = w) \mathrm{d}\theta & \text{if } \Theta \text{ is continuous} \\ \displaystyle\sum_i L(\widehat{\theta}, \theta_i) g_\Theta(\theta_i|W = w) & \text{if } \Theta \text{ is discrete} \end{cases}$$

# Bayesian Point Estimation

**Question.** Can one calculate an appropriate *point estimate* $\theta_e$ given the posterior $g_\Theta(\theta|W = w)$?

**Definitions.** Let $\theta_e$ be an estimate for $\theta$ based on a statistic $W$. The loss function associated with $\theta_e$ is denoted $L(\theta_e, \theta)$, where $L(\theta_e, \theta) \geq 0$ and $L(\theta, \theta) = 0$.

Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$. Then the risk associated with $\widehat{\theta}$ is the expected value of the loss function with respect to the posterior distribution of $\Theta$:

$$\text{risk} = \begin{cases} \displaystyle\int_{\mathbb{R}} L(\widehat{\theta}, \theta) g_\Theta(\theta|W = w)\mathrm{d}\theta & \text{if } \Theta \text{ is continuous} \\ \displaystyle\sum_i L(\widehat{\theta}, \theta_i) g_\Theta(\theta_i|W = w) & \text{if } \Theta \text{ is discrete} \end{cases}$$

# Bayesian Point Estimation

**Question.** Can one calculate an appropriate *point estimate* $\theta_e$ given the posterior $g_\Theta(\theta|W = w)$?

**Definitions.** Let $\theta_e$ be an estimate for $\theta$ based on a statistic $W$. The loss function associated with $\theta_e$ is denoted $L(\theta_e, \theta)$, where $L(\theta_e, \theta) \geq 0$ and $L(\theta, \theta) = 0$.

Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$. Then the risk associated with $\widehat{\theta}$ is the expected value of the loss function with respect to the posterior distribution of $\Theta$:

$$\text{risk} = \begin{cases} \displaystyle\int_{\mathbb{R}} L(\widehat{\theta}, \theta) g_\Theta(\theta|W = w) \mathrm{d}\theta & \text{if } \Theta \text{ is continuous} \\ \displaystyle\sum_i L(\widehat{\theta}, \theta_i) g_\Theta(\theta_i|W = w) & \text{if } \Theta \text{ is discrete} \end{cases}$$

**Theorem.** Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$.

1. If $L(\theta_e, \theta) = |\theta_e - \theta|$, then the Bayes point estimate for $\theta$ is the median of $g_\Theta(\theta|W = w)$.
2. If $L(\theta_e, \theta) = (\theta_e - \theta)^2$, then the Bayes point estimate for $\theta$ is the mean of $g_\Theta(\theta|W = w)$.

Remarks

1. Median usually does not have a closed form formula.
2. Mean usually has a closed formula.

**Theorem.** Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$.

1. If $L(\theta_e, \theta) = |\theta_e - \theta|$, then the Bayes point estimate for $\theta$ is the median of $g_\Theta(\theta|W = w)$.

2. If $L(\theta_e, \theta) = (\theta_e - \theta)^2$, then the Bayes point estimate for $\theta$ is the mean of $g_\Theta(\theta|W = w)$.

Remarks

1. Median usually does not have a closed form formula.
2. Mean usually has a closed formula.

**Theorem.** Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$.

1. If $L(\theta_e, \theta) = |\theta_e - \theta|$, then the Bayes point estimate for $\theta$ is the median of $g_\Theta(\theta|W = w)$.

2. If $L(\theta_e, \theta) = (\theta_e - \theta)^2$, then the Bayes point estimate for $\theta$ is the mean of $g_\Theta(\theta|W = w)$.

Remarks

1. Median usually does not have a closed form formula.
2. Mean usually has a closed formula.

**Theorem.** Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$.

1. If $L(\theta_e, \theta) = |\theta_e - \theta|$, then the Bayes point estimate for $\theta$ is the median of $g_\Theta(\theta|W = w)$.
2. If $L(\theta_e, \theta) = (\theta_e - \theta)^2$, then the Bayes point estimate for $\theta$ is the mean of $g_\Theta(\theta|W = w)$.

Remarks

1. Median usually does not have a closed form formula.
2. Mean usually has a closed formula.

**Theorem.** Let $g_\Theta(\theta|W=w)$ be the posterior distribution of the random variable $\Theta$.

1. If $L(\theta_e,\theta)=|\theta_e-\theta|$, then the Bayes point estimate for $\theta$ is the median of $g_\Theta(\theta|W=w)$.
2. If $L(\theta_e,\theta)=(\theta_e-\theta)^2$, then the Bayes point estimate for $\theta$ is the mean of $g_\Theta(\theta|W=w)$.

Remarks

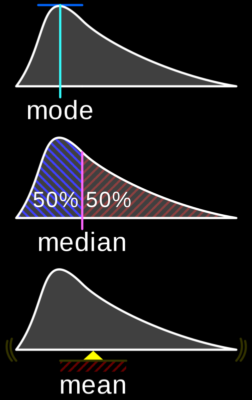
1. Median usually does not have a closed form formula.
2. Mean usually has a closed formula.

**Theorem.** Let $g_\Theta(\theta|W = w)$ be the posterior distribution of the random variable $\Theta$.

1. If $L(\theta_e, \theta) = |\theta_e - \theta|$, then the Bayes point estimate for $\theta$ is the median of $g_\Theta(\theta|W = w)$.
2. If $L(\theta_e, \theta) = (\theta_e - \theta)^2$, then the Bayes point estimate for $\theta$ is the mean of $g_\Theta(\theta|W = w)$.

Remarks

1. Median usually does not have a closed form formula.
2. Mean usually has a closed formula.




https://en.wikipedia.org

Proof. ( of Part 1. )

*Let m be the median of the random variable W. We first claim that*

$$\mathbb{E}(|W - m|) \leq \mathbb{E}(|W|). \qquad (\star)$$

*For any constant $b \in \mathbb{R}$, because*

$$\frac{1}{2} = \mathbb{P}(W \leq m) = \mathbb{P}(W - b \leq m - b)$$

*we see that $m - b$ is the median of $W - b$. Hence, by $(\star)$,*

$$\mathbb{E}\left(|W - m|\right) = \mathbb{E}\left(|(W - b) - (m - b)|\right) \leq \mathbb{E}\left(|W - b|\right), \quad \text{for all } b \in \mathbb{R},$$

*which proves the statement.*

## Proof. ( of Part 1. continued )

*It remains to prove* $(\star)$. *Without loss of generality, we may assume* $m > 0$. *Then*

$$
\begin{aligned}
\mathbb{E}(|W-m|) &= \int_{\mathbb{R}} |w-m| f_W(w)\,dw \\
&= \int_{-\infty}^{m} (m-w) f_W(w)\,dw + \int_{m}^{\infty} (w-m) f_W(w)\,dw \\
&= -\int_{-\infty}^{m} w f_W(w)\,dw + \int_{m}^{\infty} w\, f_W(w)\,dw + \frac{1}{2}(m-m) \\
&= -\int_{-\infty}^{0} w f_W(w)\,dw - \underbrace{\int_{0}^{m} w f_W(w)\,dw}_{\geq 0} + \int_{m}^{\infty} w\, f_W(w)\,dw \\
&\leq -\int_{-\infty}^{0} w f_W(w)\,dw + \int_{0}^{\infty} w\, f_W(w)\,dw \\
&= \int_{\mathbb{R}} |w| f_W(w)\,dw \\
&= \mathbb{E}(|W|).
\end{aligned}
$$

∎

Proof. ( of Part 2. )

*Let $\mu$ be the* mean *of $W$. Then for any $b \in \mathbb{R}$, we see that*

$$\begin{aligned}
\mathbb{E}\left[(W-b)^2\right] &= \mathbb{E}\left[([W-\mu]+[\mu-b])^2\right] \\
&= \mathbb{E}\left[(W-\mu)^2\right] + 2(\mu-b)\underbrace{\mathbb{E}(W-\mu)}_{=0} + [\mu-b]^2 \\
&= \mathbb{E}\left[(W-\mu)^2\right] + [\mu-b]^2 \\
&\geq \mathbb{E}\left[(W-\mu)^2\right],
\end{aligned}$$

*that is,*

$$\mathbb{E}\left[(W-\mu)^2\right] \leq \mathbb{E}\left[(W-b)^2\right], \quad \text{for all } b \in \mathbb{R}.$$

■

E.g. 1'.

$$X_1, \cdots, X_n \,|\theta \sim \text{Bernoulli}(\theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

$$X = \sum_{i=1}^{n} X_i \,\Big|\theta \sim \text{Binomial}(n, \theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

Prior Beta$(r, s) \to$ posterior Beta$(k + r, n - k + s)$ upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned}\theta_e &= \text{mean of Beta}(k + r, n - k + s) \\ &= \frac{k + r}{n + r + s} \\ &= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}}\end{aligned}$$

E.g. 1'.

$$\begin{aligned} X_1, \cdots, X_n \,\big|\theta &\sim \text{Bernoulli}(\theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \,\&\, s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^{n} X_i \,\Big|\theta &\sim \text{Binomial}(n, \theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \,\&\, s \text{ are known} \end{aligned}$$

Prior Beta$(r, s)$ $\rightarrow$ posterior Beta$(k + r, n - k + s)$ upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Beta}(k + r, n - k + s) \\ &= \frac{k + r}{n + r + s} \\ &= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}} \end{aligned}$$

E.g. 1'.

$$X_1, \cdots, X_n \,\big|\, \theta \sim \text{Bernoulli}(\theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

$$X = \sum_{i=1}^{n} X_i \,\Big|\, \theta \sim \text{Binomial}(n, \theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

Prior Beta$(r, s)$ $\rightarrow$ posterior Beta$(k + r, n - k + s)$ upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned}
\theta_e &= \text{mean of Beta}(k + r, n - k + s) \\
&= \frac{k + r}{n + r + s} \\
&= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}}
\end{aligned}$$

E.g. 1'.

$$\begin{aligned} X_1, \cdots, X_n \,\big|\theta &\sim \text{Bernoulli}(\theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \,\&\, s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^{n} X_i \,\Big|\theta &\sim \text{Binomial}(n, \theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \,\&\, s \text{ are known} \end{aligned}$$

Prior Beta$(r, s)$ $\rightarrow$ posterior Beta$(k + r, n - k + s)$
upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Beta}(k + r, n - k + s) \\ &= \frac{k + r}{n + r + s} \\ &= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}} \end{aligned}$$

E.g. 1'.

$$X_1, \cdots, X_n \,\middle|\, \theta \sim \text{Bernoulli}(\theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

$$X = \sum_{i=1}^{n} X_i \,\middle|\, \theta \sim \text{Binomial}(n, \theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

Prior Beta$(r, s) \rightarrow$ posterior Beta$(k + r, n - k + s)$
upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned}
\theta_e &= \text{mean of Beta}(k + r, n - k + s) \\
&= \frac{k + r}{n + r + s} \\
&= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}}
\end{aligned}$$

E.g. 1'.

$$X_1, \cdots, X_n \,|\theta \sim \text{Bernoulli}(\theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

$$X = \sum_{i=1}^{n} X_i \,\Big|\theta \sim \text{Binomial}(n, \theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

Prior Beta$(r, s) \rightarrow$ posterior Beta$(k + r, n - k + s)$ upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned}\theta_e &= \text{mean of Beta}(k + r, n - k + s)\\ &= \frac{k + r}{n + r + s}\\ &= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}}\end{aligned}$$

E.g. 1'.

$$\begin{aligned} X_1, \cdots, X_n \,\middle|\, \theta &\sim \text{Bernoulli}(\theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \;\&\; s \text{ are known} \end{aligned}$$

$$\begin{aligned} X = \sum_{i=1}^{n} X_i \,\middle|\, \theta &\sim \text{Binomial}(n, \theta) \\ \Theta &\sim \text{Beta}(r, s) \\ & \quad r \;\&\; s \text{ are known} \end{aligned}$$

Prior Beta$(r, s)$ $\rightarrow$ posterior Beta$(k + r, n - k + s)$
upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Beta}(k + r, n - k + s) \\ &= \frac{k + r}{n + r + s} \\ &= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}} \end{aligned}$$

E.g. 1'. $$X_1, \cdots, X_n \,\middle|\, \theta \sim \text{Bernoulli}(\theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

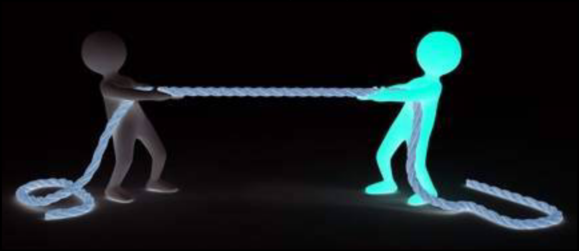
$$X = \sum_{i=1}^{n} X_i \,\Big|\, \theta \sim \text{Binomial}(n, \theta)$$
$$\Theta \sim \text{Beta}(r, s)$$
$r$ & $s$ are known

Prior Beta$(r, s) \to$ posterior Beta$(k + r, n - k + s)$
upon observing $X = k$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned}
\theta_e &= \text{mean of Beta}(k + r, n - k + s) \\
&= \frac{k + r}{n + r + s} \\
&= \frac{n}{n + r + s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r + s}{n + r + s} \times \underbrace{\left(\frac{r}{r + s}\right)}_{\text{Mean of Prior}}
\end{aligned}$$

# MLE vs. Prior



$$\theta_e$$

$$\|$$

$$\frac{n}{n+r+s} \times \underbrace{\left(\frac{k}{n}\right)}_{\text{MLE}} + \frac{r+s}{n+r+s} \times \underbrace{\left(\frac{r}{r+s}\right)}_{\text{Mean of Prior}}$$

E.g. 2'.

$$X_1, \cdots, X_n \,\big|\theta \sim \text{Poisson}(\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$s$ & $\mu$ are known

$$W = \sum_{i=1}^{n} X_i \,\Big|\theta \sim \text{Poisson}(n\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$s$ & $\mu$ are known

Prior Gamma$(s, \mu) \to$ Posterior Gamma$(w + s, \mu + n)$ upon observing $W = w$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned}\theta_e &= \text{mean of Gamma}(w + s, \mu + n) \\ &= \frac{w + s}{\mu + n} \\ &= \frac{n}{\mu + n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu + n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}}\end{aligned}$$

E.g. 2'.

$$X_1, \cdots, X_n \,\Big|\, \theta \sim \text{Poisson}(\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$$s \,\&\, \mu \text{ are known}$$

$$W = \sum_{i=1}^{n} X_i \,\Big|\, \theta \sim \text{Poisson}(n\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$$s \,\&\, \mu \text{ are known}$$

Prior Gamma$(s, \mu) \to$ Posterior Gamma$(w + s, \mu + n)$ upon observing $W = w$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned}
\theta_e &= \text{mean of Gamma}(w + s, \mu + n) \\
&= \frac{w + s}{\mu + n} \\
&= \frac{n}{\mu + n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu + n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}}
\end{aligned}$$

E.g. 2'.

$$X_1, \cdots, X_n \,\middle|\, \theta \sim \text{Poisson}(\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$s$ & $\mu$ are known

$$W = \sum_{i=1}^{n} X_i \,\middle|\, \theta \sim \text{Poisson}(n\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$s$ & $\mu$ are known

Prior Gamma$(s, \mu) \to$ Posterior Gamma$(w + s, \mu + n)$
upon observing $W = w$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Gamma}(w + s, \mu + n) \\ &= \frac{w + s}{\mu + n} \\ &= \frac{n}{\mu + n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu + n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}} \end{aligned}$$

E.g. 2'.

$$\begin{aligned} X_1, \cdots, X_n \,\big|\, \theta &\sim \text{Poisson}(\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \ \&\ \mu \text{ are known} \end{aligned}$$

$$\begin{aligned} W = \sum_{i=1}^{n} X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \ \&\ \mu \text{ are known} \end{aligned}$$

Prior Gamma$(s, \mu) \to$ Posterior Gamma$(w + s, \mu + n)$
upon observing $W = w$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Gamma}(w + s, \mu + n) \\ &= \frac{w + s}{\mu + n} \\ &= \frac{n}{\mu + n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu + n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}} \end{aligned}$$

E.g. 2'.

$$\begin{aligned} X_1, \cdots, X_n \,\big|\, \theta &\sim \text{Poisson}(\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \;\&\; \mu \text{ are known} \end{aligned}$$

$$\begin{aligned} W = \sum_{i=1}^{n} X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & \quad s \;\&\; \mu \text{ are known} \end{aligned}$$

Prior Gamma$(s, \mu) \to$ Posterior Gamma$(w + s, \mu + n)$
upon observing $W = w$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Gamma}(w + s, \mu + n) \\ &= \frac{w+s}{\mu+n} \\ &= \frac{n}{\mu+n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu+n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}} \end{aligned}$$

E.g. 2'.

$$\begin{aligned} X_1, \cdots, X_n \,\big|\, \theta &\sim \text{Poisson}(\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & s \text{ \& } \mu \text{ are known} \end{aligned}$$

$$\begin{aligned} W = \sum_{i=1}^{n} X_i \,\Big|\, \theta &\sim \text{Poisson}(n\theta) \\ \Theta &\sim \text{Gamma}(s, \mu) \\ & s \text{ \& } \mu \text{ are known} \end{aligned}$$

Prior Gamma$(s, \mu) \to$ Posterior Gamma$(w + s, \mu + n)$ upon observing $W = w$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Gamma}(w + s, \mu + n) \\ &= \frac{w + s}{\mu + n} \\ &= \frac{n}{\mu + n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu + n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}} \end{aligned}$$

E.g. 2'.

$$X_1, \cdots, X_n \,\Big|\, \theta \sim \text{Poisson}(\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$s$ & $\mu$ are known

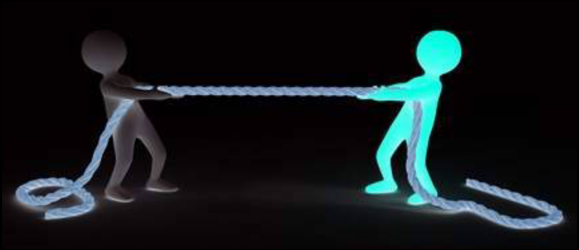
$$W = \sum_{i=1}^{n} X_i \,\Big|\, \theta \sim \text{Poisson}(n\theta)$$
$$\Theta \sim \text{Gamma}(s, \mu)$$
$s$ & $\mu$ are known

Prior Gamma$(s, \mu) \to$ Posterior Gamma$(w + s, \mu + n)$ upon observing $W = w$ for a random sample of size $n$.

Consider the $L^2$ loss function.

$$\begin{aligned} \theta_e &= \text{mean of Gamma}(w + s, \mu + n) \\ &= \frac{w + s}{\mu + n} \\ &= \frac{n}{\mu + n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu + n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}} \end{aligned}$$

# MLE vs. Prior



$$\theta_e$$

$$||$$

$$\frac{n}{\mu + n} \times \underbrace{\left(\frac{w}{n}\right)}_{\text{MLE}} + \frac{\mu}{\mu + n} \times \underbrace{\left(\frac{s}{\mu}\right)}_{\text{Mean of Prior}}$$

# Appendix: Beta integral

**Lemma.** $$B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

**Proof.** Notice that

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \quad \text{and} \quad \Gamma(\beta) = \int_0^\infty y^{\beta-1}e^{-y}dy.$$

Hence,

$$\Gamma(\alpha)\Gamma(\beta) = \int_0^\infty \int_0^\infty x^{\alpha-1}y^{\beta-1}e^{-(x+y)}dxdy.$$

The key in the proof is the following change of variables:

$$\begin{cases} x = r^2 \cos^2(\theta) \\ y = r^2 \sin^2(\theta) \end{cases}$$

$$\implies \quad \frac{\partial(x,y)}{\partial(r,\theta)} = \begin{pmatrix} 2r\cos^2(\theta) & 2r\sin^2(\theta) \\ -2r^2\cos(\theta)\sin(\theta) & 2r^2\cos(\theta)\sin(\theta) \end{pmatrix}$$

$$\implies \quad \left| \det\left( \frac{\partial(x,y)}{\partial(r,\theta)} \right) \right| = 4r^3 \sin(\theta)\cos(\theta).$$

Therefore,

$$\Gamma(\alpha)\Gamma(\beta) = \int_0^{\frac{\pi}{2}} d\theta \int_0^{\infty} dr\, r^{2(\alpha+\beta)-4} e^{-r^2} \cos^{2\alpha-2}(\theta) \sin^{2\beta-2}(\theta) \times \underbrace{4r^3 \sin(\theta)\cos(\theta)}_{\text{Jacobian}}$$

$$=4 \left( \int_0^{\frac{\pi}{2}} \cos^{2\alpha-1}(\theta) \sin^{2\beta-1}(\theta) d\theta \right) \left( \int_0^{\infty} r^{2(\alpha+\beta)-1} e^{-r^2} dr \right).$$

Now let us compute the following two integrals separately:

$$I_1 := \int_0^{\frac{\pi}{2}} \cos^{2\alpha-1}(\theta) \sin^{2\beta-1}(\theta) d\theta$$

$$I_2 := \int_0^{\infty} r^{2(\alpha+\beta)-1} e^{-r^2} dr$$

For $I_2$, by change of variable $r^2 = u$ (so that $2rdr = du$),

$$
\begin{aligned}
I_2 &= \int_0^\infty r^{2(\alpha+\beta)-1} e^{-r^2} dr \\
&= \frac{1}{2} \int_0^\infty r^{2(\alpha+\beta)-2} e^{-r^2} \underbrace{2rdr}_{=du} \\
&= \frac{1}{2} \int_0^\infty u^{\alpha+\beta-1} e^{-u} du \\
&= \frac{1}{2} \Gamma(\alpha+\beta).
\end{aligned}
$$

For $I_1$, by the change of variables $\sqrt{x} = \cos(\theta)$ (so that $-\sin(\theta) d\theta = \frac{1}{2\sqrt{x}} dx$),

$$
\begin{aligned}
I_1 &= \int_0^{\frac{\pi}{2}} \cos^{2\alpha-1}(\theta) \sin^{2\beta-1}(\theta) d\theta \\
&= \int_0^{\frac{\pi}{2}} \cos^{2\alpha-1}(\theta) \sin^{2\beta-2}(\theta) \times \underbrace{\sin(\theta) d\theta}_{=-\frac{1}{2\sqrt{x}} dx} \\
&= \int_1^0 x^{\alpha-\frac{1}{2}} (1-x)^{\beta-1} \, \frac{-1}{2\sqrt{x}} dx \\
&= \frac{1}{2} \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \, dx \\
&= \frac{1}{2} B(\alpha, \beta)
\end{aligned}
$$

Therefore,

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= 4I_1 \times I_2 \\ &= 4 \times \frac{1}{2}\Gamma(\alpha+\beta) \times \frac{1}{2}B(\alpha,\beta)\end{aligned}$$

i.e.,

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

$\square$