DAM3405 Statistics, Fall 2011*

Chang, Yung-Hsuan

February 12, 2024

*This lecture note is a summarize of the course: https://reurl.cc/OGRELR.

Contents

1 Random Variables and Distribution Functions

1.1 Review on Probability

Definition 1.1.1 (Sample Space).

Sample space is the set of all possible outcomes of a random experiment. We usually use the S to represent the sample space.

Definition 1.1.2 (Event).

Every subset of S is an event.

Definition 1.1.3 (Probability Set Function).

A function $P : \mathcal{P}(S) \to [0, 1]$ is said to be a probability set function if it satisfies all the following:

(a) $P(A) \geq 0$ for all $A \subseteq S$;

(b) $P(S) = 1$;

(c) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Definition 1.1.4 (Random Variable).

Every real-valued function defined on $\mathcal{P}(S)$ is a random variable, i.e., every $X : \mathcal{P}(S) \to \mathbb{R}$ is a random variable.

Remark 1.1.5.

Our interests of probabilities are usually divided into the following two groups:

(a) The probability of a certain event. That is, given $B \in \mathbb{R}$, we want $P(X \in B)$. Note that $X \in B$ is an event of S and represents the set $\{s \in S \mid X(s) \in B\} = X^{-1}(B)$.

(b) Distribution function of a random variable, i.e., the function $F(x) = P(X \leq x)$ defined on $\mathbb{R}$.

Definition 1.1.6 (Discrete Random Variables).

A random variables of which range is countable is called a discrete random variable.

Definition 1.1.7 (Continuous Random Variables).

A random variables is a continuous random variable if it is not a discrete random variable.

1.2 Some Distributions

Definition 1.2.1 (Bernoulli Distribution).

An experiment with two possible outcomes is called a Bernoulli experiment. The sample space is denoted by $S = \{\mathrm{S}, \mathrm{F}\}$, where S denotes "success" and F denotes "Failure". The probability set function is $P(\{\mathrm{S}\}) = p$ and $P(\{\mathrm{F}\}) = 1 - p$, where $p \in (0, 1)$. The random variable X on $S = \{\mathrm{S}, \mathrm{F}\}$ is defined by $X(\mathrm{S}) = 1$ and $X(\mathrm{F}) = 0$. In this case, X has a Bernoulli distribution with probability p, and we write $X \sim \mathrm{Bernoulli}(p)$. The probability mass function is

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

Definition 1.2.2 (Binomial Distribution).

We perform the Bernoulli experiment n times independently, and we let X be the number of successes in the n Bernoulli experiments. Then, $X : \mathcal{P}(S) \to \{0, 1, 2, \ldots, n\}$ is a random variable, where

$$S = \{(x_1, x_2, \ldots, x_n) \mid x_i \in \{\mathrm{S}, \mathrm{F}\} \text{ for all } i = 1, 2, \ldots, n\}.$$

In this case, X has a binomial distribution with parameters n and p, and we write $X \sim B(n, p)$. The probability mass function is

$$f_X(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x}, & \text{if } x = 0, 1, 2, \ldots, n; \\ 0, & \text{otherwise.} \end{cases}$$

Definition 1.2.3 (Normal Distribution).

We say that a random variable X has a normal distribution if its probability density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

for some $\mu \in \mathbb{R}$ and $\sigma > 0$. In this case, we write $X \sim \mathcal{N}(\mu, \sigma^2)$. If X has a normal distribution with $\mu = 0$ and $\sigma^2 = 1$, we say that X has a standard normal distribution.

Theorem 1.2.4.

Let $X \sim B(n, p)$. Let $\lambda = np$. Then, the probability mass function of X

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \to \frac{\lambda^x e^{-\lambda}}{x!}$$

as $n \to \infty$.

Proof.

$$\begin{aligned}
f_X(x) &= \binom{n}{x} p^x (1-p)^{1-x} \\
&= \frac{n!}{(n-x)!x!} p^x (1-p)^{1-x} \\
&= \frac{1}{x!} \cdot \frac{n!}{(n-x)!} \cdot \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \cdot \frac{n!}{(n-x)!} \cdot \frac{1}{n^x} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \frac{n \cdot (n-1) \cdot \cdots \cdot (n-x+1)}{n^x} \cdot \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&\to \frac{\lambda^x}{x!} \cdot e^{-\lambda} \cdot 1 \cdot 1 \\
&= \frac{\lambda^x e^{-\lambda}}{x!}
\end{aligned}$$

as $n \to \infty$. $\square$

Remark 1.2.5.

In the previous theorem, although n tends to infinity, λ is a fixed number. That is, p has to be extremely small so that np is not diverge. It is a little bit paradoxical.

Definition 1.2.6 (Poisson Distribution).

We say that a random variable X has a Poisson distribution if its probability mass function is

$$f_X(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & \text{if } x = 0, 1, 2, \ldots; \\ 0, & \text{otherwise.} \end{cases}$$

In this case, we write $X \sim \text{Poisson}(\lambda)$.

Definition 1.2.7 (Gamma Function).

We define the gamma function Γ on $\mathbb{R}^+$ by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, \mathrm{d}x.$$

Theorem 1.2.8.

We have some properties of the gamma function:

(a) $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$;

(b) $\Gamma(1) = \displaystyle\int_0^\infty e^{-x} \, \mathrm{d}x = 1$;

(c) $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$;

(d) $\Gamma\left(\dfrac{1}{2}\right) = \sqrt{\pi}$.

Definition 1.2.9 (Gamma Distribution).

We say that a random variable X has a gamma distribution if its probability density function is

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, & \text{if } x > 0; \\ 0, & \text{otherwise,} \end{cases}$$

for some $\alpha, \beta > 0$. In this case, we write $X \sim \text{Gamma}(\alpha, \beta)$.

Definition 1.2.10 (Chi-Squared Distribution).

If X has a gamma distribution with $\beta = 2$ and $\alpha = \frac{r}{2}$, we say that X has a chi-squared distribution with degrees of freedom r. In this case, we write $X \sim \chi^2(r)$. The probability density function is

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma\left(\dfrac{r}{2}\right) 2^{\frac{r}{2}}}\, x^{\frac{r}{2}-1} e^{-\frac{x}{2}}, & \text{if } x > 0; \\ 0, & \text{otherwise,} \end{cases}$$

for some $r > 0$.

1.3 Expectation and Variance

Remark 1.3.1.

Let $g : \mathbb{R} \to \mathbb{R}$ be a function and let X be a random variable. Then, $g(X)$ is still a random variable.

Definition 1.3.2 (Expectation).

The expectation stands for mean. Let g be a real-valued function on $\mathbb{R}$ and let X be a random variable. We define the expectation of $g(X)$ to be

$$\mathrm{E}(g(X)) = \begin{cases} \displaystyle\sum_{x \in X(S)} g(x)\, f(x), & \text{if } X \text{ is a discrete random variable;} \\ \displaystyle\int_{-\infty}^{\infty} g(x)\, f(x)\, \mathrm{d}x, & \text{if } X \text{ is a continuous random variable.} \end{cases}$$

Notation 1.3.3.

We use the Greek letter μ to represent the expectation $\mathrm{E}(X)$ if there is no confusion.

Theorem 1.3.4.

We have some properties of expectations:

(a) $\mathrm{E}(c) = c$ for all $c \in \mathbb{R}$;

(b) $\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$ for any random variable X and $a, b \in \mathbb{R}$.

Proof.

(a)

$$\begin{aligned}\mathrm{E}(c) &= \int_{-\infty}^{\infty} c f(x)\,\mathrm{d}x \\ &= c\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= c.\end{aligned}$$

(b)

$$\begin{aligned}\mathrm{E}(aX+b) &= \int_{-\infty}^{\infty} (ax+b)\, f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} axf(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} bf(x)\,\mathrm{d}x \\ &= a\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x + b\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= a\,\mathrm{E}(X) + b.\end{aligned}$$

□

Definition 1.3.5 (Variance).

The variance is a measure of dispersion. Let X be a random variable. We define the variance of X to be $\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2)$.

Notation 1.3.6.

We use σ^2 to represent the variance $\mathrm{Var}(X)$ if there is no confusion.

1.4 Moment Generating Function

Definition 1.4.1 (Moment Generating Function).

The moment generating function of a random variable X is defined as $M_X(t) = \mathrm{E}\left(e^{tX}\right)$.

Theorem 1.4.2.

If there exists a $\delta > 0$ such that $M_X(t)$ exists for $t \in (-\delta, \delta)$, then $\frac{\mathrm{d}^k}{\mathrm{d}t^k}\left(\mathrm{E}\left(e^{tX}\right)\right) = \mathrm{E}\left(\frac{\mathrm{d}^k}{\mathrm{d}t^k}\left(e^{tX}\right)\right)$ for all $k \in \mathbb{N}$.

Theorem 1.4.3.

Let X be a random variable. Then, ${M_X}^{(k)}(t) = \mathrm{E}\left(X^k\right)$ for all $k \in \mathbb{N}$.

Proof. We focus on continuous random variables. Let X be a continuous random variable. Then,

$$\begin{aligned} M_X(t) &= \mathrm{E}\left(e^{tX}\right) \\ &= \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x. \end{aligned}$$

Differentiating both sides yields

$${M_X}'(t) = \int_{-\infty}^{\infty} x e^{tx} f(x)\,\mathrm{d}x.$$

Differentiating both sides again yields

$${M_X}''(t) = \int_{-\infty}^{\infty} x^2 e^{tx} f(x)\,\mathrm{d}x.$$

Continuing in this fashion, we have

$${M_X}^{(k)}(t) = \int_{-\infty}^{\infty} x^k e^{tx} f(x)\,\mathrm{d}x.$$

Hence,

$$\begin{aligned} {M_X}^{(k)}(0) &= \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}x \\ &= \mathrm{E}(X^k). \end{aligned}$$

□

Example 1.4.4.

Let $X \sim \text{Bernoulli}(p)$. Find the moment generating function of X.

Solution.

$$\begin{aligned} M_X(t) &= \mathrm{E}(e^{tX}) \\ &= \sum_{x=0}^{1} e^{tx} p^x (1-p)^{1-x} \\ &= 1 - p + pe^t. \end{aligned}$$

■

Example 1.4.5.

Let $X \sim \text{Bernoulli}(p)$. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Solution. By the previous example, the moment generating function is $M_X(t) = 1 - p + pe^t$. The expectation

$$\begin{aligned} \mathrm{E}(X) &= \left[pe^t\right]_{t=0} \\ &= p \end{aligned}$$

and the variance

$$\begin{aligned} \mathrm{Var}(X) &= \left[pe^t\right]_{t=0} - p^2 \\ &= p(1-p). \end{aligned}$$

■

Example 1.4.6.

Let $X \sim B(n, p)$. Find the moment generating function of X.

Solution.

$$\begin{aligned} M_X(t) &= \mathrm{E}(e^{tX}) \\ &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{1-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} \left(pe^t\right)^x (1-p)^{1-x} \\ &= \left(pe^t + 1 - p\right)^n. \end{aligned}$$

■

Example 1.4.7.

Let $X \sim B(n, p)$. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Solution. By the previous example, the moment generating function is $M_X(t) = (1 - p + pe^t)^n$. The expectation

$$\begin{aligned} \mathrm{E}(X) &= \left[n\left(pe^t + 1 - p\right)^{n-1} \cdot pe^t\right]_{t=0} \\ &= np \end{aligned}$$

and the variance

$$\begin{aligned}
\mathrm{Var}(X) &= \left[n(n-1)\left(pe^t+1-p\right)^{n-2}\cdot\left(pe^t\right)^2 + n\left(pe^t+1-p\right)^{n-1}\cdot pe^t\right]_{t=0} - (np)^2 \\
&= n^2p^2 - np^2 + np - (np)^2 \\
&= np(1-p). \qquad \blacksquare
\end{aligned}$$

Example 1.4.8.

Let $X \sim \mathrm{Poisson}(\lambda)$. Find the moment generating function of X.

Solution.

$$\begin{aligned}
M_X(t) &= \mathrm{E}(e^{tX}) \\
&= \sum_{x=0}^{\infty} e^{tx}\frac{\lambda^x e^{-\lambda}}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty} e^{tx}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty} \left(\lambda e^t\right)^x \frac{1}{x!} \\
&= e^{-\lambda}e^{\lambda e^t} \\
&= e^{\lambda\left(e^t-1\right)}. \qquad \blacksquare
\end{aligned}$$

Example 1.4.9.

Let $X \sim \mathrm{Poisson}(\lambda)$. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Solution. By the previous example, the moment generating function is $M_X(t) = e^{\lambda\left(e^t-1\right)}$. The expectation

$$\begin{aligned}
\mathrm{E}(X) &= \left[\lambda e^{\lambda\left(e^t-1\right)}e^t\right]_{t=0} \\
&= \lambda
\end{aligned}$$

and the variance

$$\begin{aligned}
\mathrm{Var}(X) &= \left[\left(\lambda e^t\right)^2 e^{\lambda\left(e^t-1\right)} + \lambda e^{\lambda\left(e^t-1\right)}e^t\right]_{t=0} - \lambda^2 \\
&= \lambda^2 + \lambda - \lambda^2 \\
&= \lambda. \qquad \blacksquare
\end{aligned}$$

Example 1.4.10.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Find the moment generating function of X.

Solution.

$$
\begin{aligned}
M_X(t) &= \mathrm{E}(e^{tX}) \\
&= \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2 + 2\mu x - \mu^2 - 2\sigma^2 t x}{2\sigma^2}\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2 + 2\left(\mu - \sigma^2 t\right) x - \left(\mu - \sigma^2 t\right)^2 + \left(\mu - \sigma^2 t\right)^2 - \mu^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x - \left(\mu - \sigma^2 t\right)\right)^2}{2\sigma^2} + \frac{\left(\mu - \sigma^2 t\right)^2 - \mu^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \exp\left(\frac{\left(\mu - \sigma^2 t\right)^2 - \mu^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(x - \left(\mu - \sigma^2 t\right)\right)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \exp\left(\frac{\left(\mu - \sigma^2 t\right)^2 - \mu^2}{2\sigma^2}\right) \\
&= \exp\left(\mu t + \frac{\sigma^2}{2} t^2\right).
\end{aligned}
$$

■

Example 1.4.11.

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Solution. By the previous example, the moment generating function is $M_X(t) = \exp\left(\mu t + \frac{\sigma^2}{2} t^2\right)$. The expectation

$$
\begin{aligned}
\mathrm{E}(X) &= \left[\exp\left(\mu t + \frac{\sigma^2}{2} t^2\right) \cdot \left(\mu + \sigma^2 t\right)\right]_{t=0} \\
&= \mu
\end{aligned}
$$

and the variance

$$
\begin{aligned}
\mathrm{Var}(X) &= \left[\exp\left(\mu t + \frac{\sigma^2}{2} t^2\right) \cdot \left(\mu + \sigma^2 t\right)^2 + \exp\left(\mu t + \frac{\sigma^2}{2} t^2\right) \cdot \sigma^2\right]_{t=0} - \mu^2 \\
&= \mu^2 + \sigma^2 - \mu^2 \\
&= \sigma^2.
\end{aligned}
$$

■

Example 1.4.12.

Let $X \sim \text{Gamma}(\alpha, \beta)$. Find the moment generating function of X.

Solution.

$$
\begin{aligned}
M_X(t) &= \mathrm{E}(e^{tX}) \\
&= \int_0^\infty e^{tx} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot x^{\alpha-1} e^{-\frac{x}{\beta}} \, \mathrm{d}x \\
&= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot x^{\alpha-1} e^{x \cdot \left(-\frac{1-\beta t}{\beta}\right)} \, \mathrm{d}x \\
&= (1-\beta t)^{-\alpha} \int_0^\infty \frac{1}{\Gamma(\alpha)\left(\frac{\beta}{1-\beta t}\right)^\alpha} \cdot x^{\alpha-1} e^{x \cdot \left(-\frac{1-\beta t}{\beta}\right)} \, \mathrm{d}x \\
&= (1-\beta t)^{-\alpha}, \quad t \in \left(-\infty, \frac{1}{\beta}\right).
\end{aligned}
$$
■

Example 1.4.13.

Let $X \sim \text{Gamma}(\alpha, \beta)$. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Solution. By the previous example, the moment generating function is $M_X(t) = (1-\beta t)^{-\alpha}$. The expectation

$$
\begin{aligned}
\mathrm{E}(X) &= \left[-\alpha\,(1-\beta t)^{-\alpha-1} \cdot (-\beta)\right]_{t=0} \\
&= \alpha\beta
\end{aligned}
$$

and the variance

$$
\begin{aligned}
\mathrm{Var}(X) &= \left[\alpha\beta(-\alpha-1)\,(1-\beta t)^{-\alpha-2} \cdot (-\beta)\right]_{t=0} - (\alpha\beta)^2 \\
&= \alpha^2\beta^2 + \alpha\beta^2 - (\alpha\beta)^2 \\
&= \alpha\beta^2.
\end{aligned}
$$
■

Example 1.4.14.

Let $X \sim \chi^2(r)$. Find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

Solution. By the previous example, $\mathrm{E}(X) = r$ and $\mathrm{Var}(X) = 2r$. ■

1.5 Probability Density Function of Composite Functions

Remark 1.5.1.

Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a function. Then, $Y = g(X)$ is also a random variable and must have a probability distribution function.

Theorem 1.5.2 (Distribution Function Method).

Let X be a continuous random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a function. Suppose the distribution function of X is $F(x)$. Then, the distribution function of $Y = g(X)$ is $G(y) = P(g(X) \leq y)$. If $G(y)$ is attainable, then the probability density function of Y is $f_Y(y) = G'(y)$.

Proposition 1.5.3.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Proof. The distribution function of Z is

$$\begin{aligned}
F_Z(z) &= P(Z \leq z) \\
&= P(X \leq \sigma z + \mu) \\
&= \int_{-\infty}^{\sigma z+\mu} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right) \, \mathrm{d}x,
\end{aligned}$$

which implies

$$\begin{aligned}
f_Z(z) &= \sigma \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{(\sigma z)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{z^2}{2}\right).
\end{aligned}$$

Hence, $Z \sim \mathcal{N}(0,1)$. □

Proposition 1.5.4.

If $X \sim \text{Gamma}(\alpha, \beta)$, then $Y = \dfrac{2X}{\beta} \sim \chi^2(r)$, where $r = 2\alpha$.

Proof. The distribution function of Y is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\left(X \leq \frac{\beta y}{2}\right) \\ &= \int_0^{\frac{\beta y}{2}} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \, \mathrm{d}x, \end{aligned}$$

which implies

$$\begin{aligned} f_Y(y) &= \frac{\beta}{2} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\frac{\beta y}{2}\right)^{\alpha-1} e^{-\frac{x}{\beta}} \\ &= \frac{1}{\Gamma(\alpha)2^\alpha} y^{\alpha-1} e^{-\frac{y}{2}} \\ &= \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{\frac{r}{2}}} y^{\frac{r}{2}-1} e^{-\frac{y}{2}}. \end{aligned}$$

Hence, $Y \sim \chi^2(r)$. □

Theorem 1.5.5 (Moment Generating Function Method).

The moment generating function and its distribution forms an injection.

Example 1.5.6.

If the probability density function of X is $f(x) = \dfrac{3^x e^{-3}}{x!}$ for $x = 0, 1, 2, \ldots$, then $M_X(t) = e^{3(e^t-1)}$.

Proof. This follows by Theorem 1.5.5. ■

Example 1.5.7.

If the moment generating function of X is $M_X(t) = e^{100(e^t-1)}$, then $X \sim \text{Poisson}(100)$.

Proof. This follows by Theorem 1.5.5. ■

Recall 1.5.8.

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Proof. The moment generating function of Z is

$$
\begin{aligned}
M_Z(t) &= \mathrm{E}\left(e^{tZ}\right) \\
&= \mathrm{E}\left(\exp\left(\frac{tX - t\mu}{\sigma}\right)\right) \\
&= \mathrm{E}\left(\exp\left(\frac{tX}{\sigma}\right)\right) \cdot \mathrm{E}\left(\exp\left(\frac{-t\mu}{\sigma}\right)\right) \\
&= \mathrm{E}\left(\exp\left(\frac{tX}{\sigma}\right)\right) \cdot \exp\left(\frac{-t\mu}{\sigma}\right) \\
&= M_X\left(\frac{t}{\sigma}\right) \cdot \exp\left(\frac{-t\mu}{\sigma}\right) \\
&= \exp\left(\mu \cdot \frac{t}{\sigma} + \frac{\sigma^2}{2}\left(\frac{t}{\sigma}\right)^2\right) \cdot \exp\left(\frac{-t\mu}{\sigma}\right) \\
&= \exp\left(\frac{t^2}{2}\right).
\end{aligned}
$$

By the moment of generating function method, $Z \sim \mathcal{N}(0,1)$. ☐

Recall 1.5.9.

If $X \sim \text{Gamma}(\alpha, \beta)$, then $Y = \dfrac{2X}{\beta} \sim \chi^2(r)$, where $r = 2\alpha$.

Proof. The moment generating function of Y is

$$
\begin{aligned}
M_Y(t) &= \mathrm{E}\left(e^{tY}\right) \\
&= \mathrm{E}\left(e^{t \cdot \frac{2X}{\beta}}\right) \\
&= M_X\left(\frac{2t}{\beta}\right) \\
&= \left(1 - \beta \cdot \frac{2t}{\beta}\right)^{-\alpha} \\
&= (1 - 2t)^{-\alpha}.
\end{aligned}
$$

By the moment of generating function method, $Y \sim \chi^2(r)$. ☐

1.6 Probability Density Functions with Multiple Random Variables

Definition 1.6.1 (Random Vector).

If $X_1, X_2, \ldots, X_n$ are random variables, we call $(X_1, X_2, \ldots, X_n)$ a <u>random vector</u>.

Definition 1.6.2 (Joint Probability Distribution Function).

Let $(X_1, X_2, \dots, X_n)$ be a random vector. If they are discrete, the joint probability mass function is $f(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$. If they are continuous, there exists a $f \geq 0$ such that the joint distribution function

$$\begin{aligned} F(x_1, x_2, \dots, x_n) &= P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \\ &= \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f(t_1, t_2, \dots, t_n) \, \mathrm{d}t_1 \, \mathrm{d}t_2 \cdots \mathrm{d}t_n. \end{aligned}$$

We call $f(x_1, x_2, \dots, x_n)$ the joint probability distribution function of $X_1, X_2, \dots, X_n$.

Theorem 1.6.3.

If X_1, X_2 has a joint probability distribution function $f(x_1, x_2)$, then the marginal probability distribution functions are

$$f_{X_1}(x_1) = \begin{cases} \displaystyle\sum_{x_2 \in X_2(S)} f(x_1, x_2), & \text{if } (X_1, X_2) \text{ are discrete;} \\ \displaystyle\int_{-\infty}^{\infty} f(x_1, x_2) \, \mathrm{d}x_2, & \text{if } (X_1, X_2) \text{ are continuous,} \end{cases}$$

and

$$f_{X_2}(x_2) = \begin{cases} \displaystyle\sum_{x_1 \in X_1(S)} f(x_1, x_2), & \text{if } (X_1, X_2) \text{ are discrete;} \\ \displaystyle\int_{-\infty}^{\infty} f(x_1, x_2) \, \mathrm{d}x_1, & \text{if } (X_1, X_2) \text{ are continuous,} \end{cases}$$

Definition 1.6.4 (Independent).

Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.

Definition 1.6.5 (Independent).

Let X and Y be two random variables with joint probability distribution function $f(x, y)$ and marginal probability distribution functions $f_X(x)$ and $f_Y(y)$. We say X and Y are independent if $f(x, y) = f_X(x) \, f_Y(y)$ for $(x, y) \in \mathbb{R}^2$.

Definition 1.6.6 (Identically Distributed).

Random Variables X and Y are identically distributed if two marginal probability distribution functions f_X and f_Y share the same domain D and satisfy $f_X = f_Y$, i.e., $f_X(u) = f_Y(u)$ for all $u \in D$.

Theorem 1.6.7.

Let X be a continuous random variable with probability density function $f(x)$. If g is an injection, then the probability density function of Y is

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\dfrac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right|, & y \in g(A); \\ 0, & \text{otherwise.} \end{cases}$$

Proof. The distribution function of Y is $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$. Suppose g is increasing. Then, g^{-1} is also increasing. Thus, the probability density function of Y is

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y}\left(P(X \leq g^{-1}(y))\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\left(\int_{-\infty}^{g^{-1}(y)} f_X(x)\,\mathrm{d}x\right) \\ &= f_X(g^{-1}(y))\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \\ &= f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right|. \end{aligned}$$

Now, suppose g is decreasing. Then, g^{-1} is also decreasing. Thus, the probability density function of Y is

$$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y}\left(P(X \geq g^{-1}(y))\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\left(1 - P(X \leq g^{-1}(y))\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\left(1 - \int_{-\infty}^{g^{-1}(y)} f_X(x)\,\mathrm{d}x\right) \\ &= -f_X(g^{-1}(y))\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \\ &= f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right|. \end{aligned}$$

□

Definition 1.6.8 (Uniform Distribution).

We say that a random variable X has a uniform distribution if its probability density function is

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leq x \leq b; \\ 0, & \text{otherwise.} \end{cases}$$

In this case, we write $X \sim U(a, b)$.

Example 1.6.9.

Let $X \sim U(0, 1)$. Find the distribution of $Y = -2\ln(X)$.

Solution. The probability density function of Y is

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right) \left|\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right| \\ &= 1 \cdot \left|-\frac{1}{2}e^{-\frac{y}{2}}\right| \\ &= \frac{1}{\Gamma\left(\frac{2}{2}\right)2^{\frac{2}{2}}} y^{\frac{2}{2}-1}e^{-\frac{y}{2}}. \end{aligned}$$

Hence, $Y \sim \chi^2(2)$. □

Theorem 1.6.10.

Let $y_i = g_i(\mathbf{x})$ be an injection for $i = 1, 2, \ldots, n$ with inverse function $x_i = w_i(\mathbf{y})$ for $i = 1, 2, \ldots, n$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. Then, the distribution function

$$\begin{aligned} P(\mathbf{x} \in A) &= \int \cdots \int_{A \subseteq \mathbb{R}^n} f_{X_1, X_2, \ldots, X_n}(\mathbf{x}) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int \cdots \int_{(g_1, \ldots, g_n)(A)} f_{X_1, X_2, \ldots, X_n}(w_1(\mathbf{y}), w_2(\mathbf{y}), \ldots, w_n(\mathbf{y})) \, |J| \, \mathrm{d}y_1 \cdots \mathrm{d}y_n \\ &= P(\mathbf{y} \in (g_1, \ldots, g_n)(A)), \end{aligned}$$

where the Jacobian matrix $J = \dfrac{\partial(\mathbf{x})}{\partial(\mathbf{y})}$. Hence, $f_{Y_1Y_2\ldots Y_n}(\mathbf{y}) = f_{X_1X_2\ldots X_n}(w_1(\mathbf{y}), w_2(\mathbf{y}), \ldots, w_n(\mathbf{y})) \, |J|$ for $\mathbf{y} \in (g_1, \ldots, g_n)(A)$.

Remark 1.6.11.

If we want to find m random variables $Y_1, Y_2, \ldots, Y_m$ consist of $X_1, X_2, \ldots, X_n$:

Step 1: Find the joint probability density function of $X_1, X_2, \ldots, X_n$ and the domain A.

Step 2: Make sure that there is an injection between $\mathbf{x}$ and $\mathbf{y}$. Find the inverse $x_i = w_i(\mathbf{y})$ for $i = 1, 2, \ldots, n$.

Step 3: Check the range $(y_1, \ldots, y_n)(A)$, which is $(g_1, \ldots, g_n)(A)$.

Example 1.6.12.

Suppose that X_1 and X_2 are independent and identically distributed random variables with distribution $U(0,1)$. Let $Y_1 = X_1 + X_2$ and let $Y_2 = X_1 - X_2$. Find the marginal probability density functions of Y_1 and Y_2.

Solution. The joint probability density function is

$$f_{X_1,X_2}(x_1,x_2) = \begin{cases} 1, & \text{if } (x_1,x_2) \in (0,1)\times(0,1); \\ 0, & \text{otherwise.} \end{cases}$$

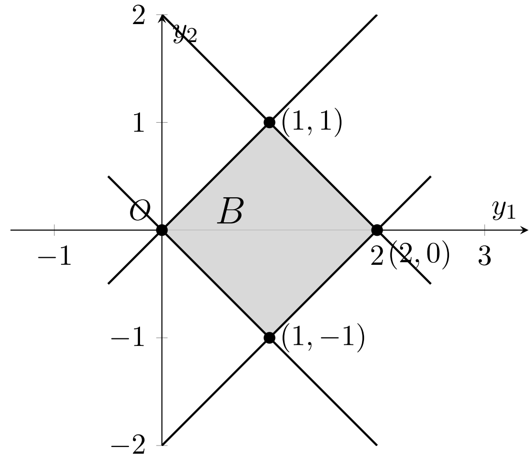
The domain A is the set $(0,1)\times(0,1)$. Notice that $X_1 = \dfrac{Y_1+Y_2}{2}$ and $X_2 = \dfrac{Y_1 - Y_2}{2}$ is an injection. Hence, the Jacobian determinant is

$$\begin{aligned} \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} &= \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} \\ &= -\frac{1}{2}, \end{aligned}$$

and the joint probability density function of Y_1, Y_2 is

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \begin{cases} 1\cdot\frac{1}{2}, & \text{if } \left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right) \in (0,1)\times(0,1); \\ 0\cdot\frac{1}{2}, & \text{otherwise,} \end{cases} \\ &= \begin{cases} 1\cdot\frac{1}{2}, & \text{if } (y_1,y_2)\in B; \\ 0\cdot\frac{1}{2}, & \text{otherwise,} \end{cases} \end{aligned}$$

where B is the shaded area in the picture below.



The marginal probability density functions are

$$f_{Y_1}(y_1) = \begin{cases} \displaystyle\int_{-y_1}^{y_1} \frac{1}{2}\,\mathrm{d}y_2, & \text{if } y_1 \in (0,1); \\ \displaystyle\int_{2-y_1}^{y_1-2} \frac{1}{2}\,\mathrm{d}y_2, & \text{if } y_1 \in (1,2); \\ 0, & \text{otherwise,} \end{cases}$$

$$= \begin{cases} y_1, & \text{if } y_1 \in (0,1); \\ 1 - y_1, & \text{if } y_1 \in (1,2); \\ 0, & \text{otherwise,} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} \displaystyle\int_{-y_2}^{2+y_2} \frac{1}{2}\,\mathrm{d}y_2, & \text{if } y_2 \in (-1,0); \\ \displaystyle\int_{y_2}^{2-y_2} \frac{1}{2}\,\mathrm{d}y_2, & \text{if } y_2 \in (0,1); \\ 0, & \text{otherwise,} \end{cases}$$

$$= \begin{cases} 1 + y_2, & \text{if } y_2 \in (-1,0); \\ 1 - y_2, & \text{if } y_2 \in (0,1); \\ 0, & \text{otherwise.} \end{cases}$$

□

1.7 Statistic and Independence

Definition 1.7.1 (Random Sample).

If random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed, then we call them a random sample.

Proposition 1.7.2.

If $X_1, X_2, \ldots, X_n$ is a random sample from $f_0(x)$, then the joint probability density function of $X_1, X_2, \ldots, X_n$ is

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_0(x_i)$$

for all $\mathbf{x} \in \mathbb{R}^n$.

Proof. This follows by definitions. $\square$

Definition 1.7.3 (Statistic).

Any function $g(X_1, X_2, \ldots, X_n)$ of a random sample $X_1, X_2, \ldots, X_n$, which is not dependent on parameter θ, is called a statistic.

Example 1.7.4.

(a) The sample mean $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$ is a statistic.

(b) The sample variance $s^2 = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2$ is a statistic.

(c) The indicator function $I(X_1 \geq 0)$ is a statistic.

Remark 1.7.5.

If a random variable X has a probability density function $f(x, \theta)$, where θ is an unknown constant, then we call θ a parameter.

Example 1.7.6.

(a) If $X \sim \mathcal{N}(\mu, \sigma^2)$, then μ and σ^2 are parameters.

(b) If $X \sim \text{Poisson}(\lambda)$, then λ is a parameter.

Definition 1.7.7 (Joint Moment Generating Function).

Let $X_1, X_2, \ldots, X_n$ be random variables. The joint moment generating function of $X_1, X_2, \ldots, X_n$ is

$$M_{X_1,X_2,\ldots,X_n}(t_1, t_2, \ldots, t_n) = \mathrm{E}\left(e^{t_1 X_1 + \cdots + t_n X_n}\right).$$

Lemma 1.7.8.

Two random variables X_1 and X_2 are independent if and only if

$$M_{X_1,X_2}(t_1, t_2) = M_{X_1}(t_1) M_{X_2}(t_2).$$

Proof.

($\Longrightarrow$) Suppose X_1 and X_2 are independent. Then,

$$\begin{aligned}
M_{X_1,X_2}(t_1, t_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 X_1 + t_2 X_2} f_{X_1,X_2}(x_1, x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 X_1} e^{t_2 X_2} f_{X_1}(x_1) \, f_{X_2}(x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} e^{t_1 X_1} f_{X_1}(x_1) \, \mathrm{d}x_1 \int_{-\infty}^{\infty} e^{t_2 X_2} f_{X_2}(x_2) \, \mathrm{d}x_2 \\
&= M_{X_1}(t_1) M_{X_2}(t_2).
\end{aligned}$$

($\Longleftarrow$) Suppose $M_{X_1,X_2}(t_1, t_2) = M_{X_1}(t_1) M_{X_2}(t_2)$. Then,

$$\begin{aligned}
M_{X_1,X_2}(t_1, t_2) &= \mathrm{E}\left(e^{t_1 X_1 + t_2 X_2}\right) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 X_1 + t_2 X_2} f_{X_1,X_2}(x_1, x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2,
\end{aligned}$$

and

$$\begin{aligned}
M_{X_1}(t_1) M_{X_2}(t_2) &= \int_{-\infty}^{\infty} e^{t_1 X_1} f_{X_1}(x_1) \, \mathrm{d}x_1 \int_{-\infty}^{\infty} e^{t_2 X_2} f_{X_2}(x_2) \, \mathrm{d}x_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 X_1} e^{t_2 X_2} f_{X_1}(x_1) \, f_{X_2}(x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2.
\end{aligned}$$

Hence, $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \, f_{X_2}(x_2)$, which means X_1 and X_2 are independent. $\square$

Definition 1.7.9 (Independent).

Let $\mathbf{X}$ be a random vector with n component and let $\mathbf{Y}$ be a random vector with m component. We say $\mathbf{X}$ and $\mathbf{Y}$ are independent if

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})\, f_{\mathbf{Y}}(\mathbf{y}).$$

Theorem 1.7.10.

Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors. Let g be a function of $\mathbf{X}$ and let h be a function of $\mathbf{Y}$. If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $g(\mathbf{X})$ and $h(\mathbf{Y})$ are independent.

Theorem 1.7.11.

Let X and Y be random variables. Let g be a function of X and let h be a function of Y. If X and Y are independent, then $\mathrm{E}(g(X)h(Y)) = \mathrm{E}(g(X))\,\mathrm{E}(h(Y))$.

Proof.

$$\begin{aligned}
\mathrm{E}(g(X)h(Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\, f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\, f_X(x)\, f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} g(x)\, f_X(x)\,\mathrm{d}x \int_{-\infty}^{\infty} h(y)\, f_Y(y)\,\mathrm{d}y \\
&= \mathrm{E}(g(X))\,\mathrm{E}(h(Y)).
\end{aligned}$$

□

Theorem 1.7.12.

The joint moment generating function $M_{X_1,X_2,\dots,X_n}(t_1,t_2,\dots,t_n)$ of random variables $X_1,X_2,\dots,X_n$ at point $(0,\dots,0,t,0,\dots,0)$ (the ith component is t and 0 elsewhere) is the moment generating function $M_{X_i}(t)$ of X_i.

Proof. Let $\mathbf{t}_i^* = (0,0,\dots,0,t,0,\dots,0)$ (the ith component is t and 0 elsewhere). Then,

$$M_{X_1,X_2,\dots,X_n}(t_1,t_2,\dots,t_n) = \mathrm{E}\left(\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right),$$

which implies

$$M_{X_1,X_2,\ldots,X_n}(\mathbf{t}_i^*) = \mathrm{E}\left(\exp\left(tX_i\right)\right)$$

$$= M_{X_i}(t). \qquad \square$$

Theorem 1.7.13.

If $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$ are independent, then $X + Y \sim \chi^2(r_1 + r_2)$.

Proof. The moment generating function of $X + Y$ is

$$\begin{aligned} M_{X+Y}(t) &= \mathrm{E}\left(e^{tX+tY}\right) \\ &= \mathrm{E}\left(e^{tX}\right)\mathrm{E}\left(e^{tY}\right) \\ &= M_X(t) + M_Y(t) \\ &= (1+2t)^{-\frac{r_1}{2}}(1+2t)^{-\frac{r_2}{2}} \\ &= (1+2t)^{-\frac{r_1+r_2}{2}}. \end{aligned}$$

Hence, $X + Y \sim \chi^2(r_1 + r_2)$. $\square$

Theorem 1.7.14.

If $Z \sim \mathcal{N}(0, 1)$, then $Z^2 \sim \chi^2(1)$.

Proof. Let $Y = Z^2$. The distribution function of Y is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(Z^2 \leq y) \\ &= P(-\sqrt{y} \leq Z \leq \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x. \end{aligned}$$

Thus, the probability density function of Y is

$$
\begin{aligned}
f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y}\left(2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x\right) \\
&= \frac{2}{\sqrt{2\pi}} e^{-\frac{y}{2}} \cdot \frac{1}{2\sqrt{y}} \\
&= \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{\frac{1}{2}}} y^{\frac{1}{2}-1} e^{-\frac{y}{2}}.
\end{aligned}
$$

Hence, $Y \sim \chi^2(1)$. □

Theorem 1.7.15.

If $X_1, X_2, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, then

(a) $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$;

(b) $\overline{X}$ and s^2 are independent;

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

where $\overline{X}$ is the sample mean, and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is the sample variance.

Proof.

(a) The moment generating function of $\overline{X}$ is

$$
\begin{aligned}
M_{\overline{X}}(t) &= \mathrm{E}\left(e^{t\overline{X}}\right) \\
&= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{X_i}{n}\right)\right) \\
&= \prod_{i=1}^{n} \mathrm{E}\left(\exp\left(t\frac{X_i}{n}\right)\right) \\
&= \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right) \\
&= \prod_{i=1}^{n} \exp\left(\mu\frac{t}{n} + \frac{\sigma^2}{2}\left(\frac{t}{n}\right)^2\right) \\
&= \exp\left(\mu t + \frac{\frac{\sigma^2}{n}}{2} t^2\right).
\end{aligned}
$$

Hence, $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

(b) We want to show that $\overline{X}$ and $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ are independent. The joint moment generating function of $\overline{X}$ and $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ is

$$
\begin{aligned}
&M_{X,X_1-\overline{X},X_2-\overline{X},\ldots,X_n-\overline{X}}(t,t_1,t_2,\ldots,t_n)\\
=&\,\mathrm{E}\left(\exp\left(t\overline{X}+\sum_{i=1}^{n}t_i(X_i-\overline{X})\right)\right)\\
=&\,\mathrm{E}\left(\exp\left(\sum_{i=1}^{n}\frac{t}{n}X_i+\sum_{i=1}^{n}t_i\left(X_i-\sum_{i=1}^{n}\frac{X_i}{n}\right)\right)\right)\\
=&\,\mathrm{E}\left(\exp\left(\sum_{i=1}^{n}\frac{t}{n}X_i+\sum_{i=1}^{n}t_iX_i-\sum_{i=1}^{n}\bar{t}X_i\right)\right)\\
=&\,\mathrm{E}\left(\exp\left(\sum_{i=1}^{n}\left(\frac{t}{n}+t_i-\bar{t}\right)X_i\right)\right)\\
=&\,\mathrm{E}\left(\exp\left(\prod_{i=1}^{n}\left(\frac{t}{n}+t_i-\bar{t}\right)X_i\right)\right)\\
=&\prod_{i=1}^{n}\mathrm{E}\left(\exp\left(\left(\frac{t}{n}+t_i-\bar{t}\right)X_i\right)\right)\\
=&\prod_{i=1}^{n}M_X\left(\frac{t}{n}+t_i-\bar{t}\right)\\
=&\prod_{i=1}^{n}M_X\left(\frac{n(t_i-\bar{t})+t}{n}\right)\\
=&\prod_{i=1}^{n}\exp\left(\mu\frac{n(t_i-\bar{t})+t}{n}+\frac{\sigma^2}{2}\left(\frac{n(t_i-\bar{t})+t}{n}\right)^2\right)\\
=&\exp\left(\sum_{i=1}^{n}\left(\mu\frac{n(t_i-\bar{t})+t}{n}+\frac{\sigma^2}{2}\left(\frac{n(t_i-\bar{t})+t}{n}\right)^2\right)\right)\\
=&\exp\left(\sum_{i=1}^{n}\mu\frac{n(t_i-\bar{t})+t}{n}+\sum_{i=1}^{n}\frac{\sigma^2}{2}\left(\frac{n(t_i-\bar{t})+t}{n}\right)^2\right)\\
=&\exp\left(\mu\sum_{i=1}^{n}\left(t_i-\bar{t}+\frac{t}{n}\right)+\frac{\sigma^2}{2}\left(\sum_{i=1}^{n}\frac{n^2(t_i-\bar{t})^2+2nt(t_i-\bar{t})+t^2}{n^2}\right)\right)\\
=&\exp\left(\mu t+\frac{\sigma^2}{2}\sum_{t=1}^{n}(t_i-\bar{t})^2+\frac{\sigma^2}{2}\sum_{t=1}^{n}\left(\frac{t}{n}\right)^2\right)\\
=&\exp\left(\mu t+\frac{\sigma^2}{2}\sum_{t=1}^{n}(t_i-\bar{t})^2+\frac{\frac{\sigma^2}{n}}{2}t^2\right)\\
=&\exp\left(\mu t+\frac{\frac{\sigma^2}{n}}{2}t^2\right)\exp\left(\frac{\sigma^2}{2}\sum_{t=1}^{n}(t_i-\bar{t})^2\right)\\
=&M_{\overline{X}}(t)M_{X,X_1-\overline{X},X_2-\overline{X},\ldots,X_n-\overline{X}}(0,t_1,t_2,\ldots,t_n)\\
=&M_{\overline{X}}(t)M_{X_1-\overline{X},X_2-\overline{X},\ldots,X_n-\overline{X}}(t_1,t_2,\ldots,t_n).
\end{aligned}
$$

Hence, $\overline{X}$ and $(X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ are independent, which implies $\overline{X}$ and s^2 are independent.

(c) Let $g(X) = \dfrac{X - \mu}{\sigma}$. Then, $g(X_1), g(X_2), \ldots, g(X_n)$ are independent and identically distributed, and $g(X_1), g(X_2), \ldots, g(X_n) \sim \mathcal{N}(0,1)$. Let $h(X) = X^2$. Then, $h(g(X_1)), h(g(X_2)), \ldots, h(g(X_n))$ are independent and identically distributed, and $h(g(X_1)), h(g(X_2)), \ldots, h(g(X_n)) \sim \chi^2(1)$. Notice that $h(g(X)) = \dfrac{(X-\mu)^2}{\sigma^2}$. Hence, $\displaystyle\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$. The moment generating function of $Z_1 = \displaystyle\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$ is

$$
\begin{aligned}
M_{Z_1}(t) &= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}\right)\right) \\
&= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{(X_i-\overline{X}+\overline{X}-\mu)^2}{\sigma^2}\right)\right) \\
&= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2+2(X_i-\overline{X})(\overline{X}-\mu)+(\overline{X}-\mu)^2}{\sigma^2}\right)\right) \\
&= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}+t\sum_{i=1}^{n}\frac{(\overline{X}-\mu)^2}{\sigma^2}\right)\right) \\
&= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}+t\frac{n(\overline{X}-\mu)^2}{\sigma^2}\right)\right) \\
&= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}+t\frac{(\overline{X}-\mu)^2}{\frac{\sigma^2}{n}}\right)\right) \\
&= \mathrm{E}\left(\exp\left(t\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2}\right)\exp\left(t\sum_{i=1}^{n}\frac{(\overline{X}-\mu)^2}{\frac{\sigma^2}{n}}\right)\right) \\
&= M_Z(t)M_{Z_2}(t),
\end{aligned}
$$

where $Z = \displaystyle\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$ and $Z_2 = \displaystyle\sum_{i=1}^{n}\frac{(\overline{X}-\mu)^2}{\frac{\sigma^2}{n}}$. Notice that $\dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$, and thus $Z_2 \sim \chi^2(1)$. Hence,

$$
\begin{aligned}
(1-2t)^{-\frac{n}{2}} &= M_{Z_1}(t) \\
&= M_Z(t)M_{Z_2}(t) \\
&= M_Z(t)(1-2t)^{-\frac{1}{2}} \\
\iff \qquad M_Z(t) &= (1-2t)^{-\frac{n-1}{2}},
\end{aligned}
$$

which implies $Z = \dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$. $\square$

Remark 1.7.16.

We now have

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n); \\ \displaystyle\sum_{i=1}^{n} \frac{\left(X_i-\overline{X}\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1). \end{cases}$$

2 Statistical Inference: Point Estimation

2.1 Introduction to Statistical Inference

Problem 2.1.1 (Problem in Statistics).

We have a random variable X with probability distribution function $f(x,\theta)$, where the function f is know but the parameters θ are unknown. How can we infer θ?

Solution. In our real world, it is almost always the case that the parameters are unknown, and that is why Statistics exists. We must obtain a random sample $X_1, X_2, \ldots, X_n$ from $f(x,\theta)$. Basically, there are two main ways for inferences: estimation and hypothesis testing. Estimation is essentially about finding the value of θ. The most intuitive way for estimation is point estimation, i.e., finding the best $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$. Point estimation is crucial in Statistics, and many theories about point estimation are clear and beautiful. However, there are drawbacks to point estimation. It is nearly impossible for us to find the correct parameters since the probability

$$\begin{aligned} P(\hat{\theta}(X_1, X_2, \ldots, X_n) = \theta) &= \int_{\theta}^{\theta} f_{\hat{\theta}}(u)\,\mathrm{d}u \\ &= 0. \end{aligned}$$

Therefore, statisticians developed interval estimation to find two statistics $T_1 = t_1(X_1, X_2, \ldots, X_n)$ and $T_2 = t_2(X_1, X_2, \ldots, X_n)$ such that

$$1-\alpha = P(T_1 \le \theta \le T_2),$$

where $1-\alpha$ is usually called the level of confidence, also denoted by γ. The most important part of Statistics is hypothesis testing. Statistics is developed from hypothesis testing. Most of the time, we do not aim to find θ, but we would like to know whether it exceeds a specific value or not. □

2.2 Estimation

Definition 2.2.1 (Estimator).

We call a statistic $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ an estimator of parameter θ if it is used to estimate θ. If $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ are observed, we call $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ an estimate of θ.

Problem 2.2.2.

Usually, there are many estimators available; how can we choose one from them?

Solution. We need criterion of good or the best estimation. □

Problem 2.2.3.

Are there general rules in deriving estimators?

Solution. Yes. We will introduce two methods. □

Definition 2.2.4 (Unbiased Estimator).

We call an estimator $\hat{\theta}$ an unbiased estimator if

$$\mathrm{E}_\theta\left(\hat{\theta}(X_1, X_2, \ldots, X_n)\right) = \theta$$

for all $\theta \in \Theta$, where Θ is the set of all possible θ's, also known as the parameter space. The expectation

$$\begin{aligned}
&\mathrm{E}_\theta\left(\hat{\theta}(X_1, X_2, \ldots, X_n)\right) \\
&= \begin{cases} \displaystyle\int_{-\infty}^{\infty} \theta^* f_{\hat{\theta}}(\theta^*)\,\mathrm{d}\theta^*, & \text{if the probability density function } f_{\hat{\theta}} \text{ of } \hat{\theta} \text{ is available;} \\ \displaystyle\int \cdots \int_{\mathbb{R}^n} \hat{\theta}(\mathbf{X})\, f_{\mathbf{X}}(\mathbf{X})\,\mathrm{d}x_1 \cdots \mathrm{d}x_n, & \text{otherwise.} \end{cases}
\end{aligned}$$

Example 2.2.5.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Our interest is μ. Justify whether the following is unbiased or not:

(a) X_1; (b) $\dfrac{X_1 + X_2}{2}$; (c) $\overline{X}$.

Solution.

(a) Since $\mathrm{E}_\mu(X_1) = \mu$, X_1 is an unbiased estimator for μ.

(b) Since

$$\begin{aligned}\mathrm{E}_\mu\left(\frac{X_1 + X_2}{2}\right) &= \mathrm{E}_\mu\left(\frac{X_1}{2}\right) + \mathrm{E}_\mu\left(\frac{X_2}{2}\right) \\ &= \frac{\mu}{2} + \frac{\mu}{2} \\ &= \mu,\end{aligned}$$

$\dfrac{X_1 + X_2}{2}$ is an unbiased estimator for μ.

(c) Since

$$\begin{aligned}\mathrm{E}_\mu\left(\overline{X}\right) &= \mathrm{E}_\mu\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) \\ &= \sum_{i=1}^{n} \mathrm{E}_\mu\left(\frac{X_i}{n}\right) \\ &= \sum_{i=1}^{n} \frac{\mu}{n} \\ &= \mu,\end{aligned}$$

$\overline{X}$ is an unbiased estimator for μ. ■

Definition 2.2.6 (Converge in Probability).

We say that a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ <u>converges</u> to a random variable or a constant X in probability if for any $\varepsilon > 0$, we have

$$P\left(|X_n - X| \geq \varepsilon\right) \to 0$$

as $n \to \infty$. In this case, we write $X_n \xrightarrow{P} X$.

Theorem 2.2.7 (Markov's Inequality).

If $X \geq 0$, then $P(X \geq u) \leq \dfrac{\mathrm{E}(X)}{u}$ for all $u > 0$.

Proof.

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \\
&= \int_{0}^{\infty} x f(x)\,\mathrm{d}x \\
&\leq \int_{u}^{\infty} x f(x)\,\mathrm{d}x \\
&\leq \int_{u}^{\infty} u f(x)\,\mathrm{d}x \\
&= u \int_{u}^{\infty} f(x)\,\mathrm{d}x \\
&= uP(X \geq u) \\
\iff \quad P(X \geq u) &\leq \frac{\mathrm{E}(X)}{u}.
\end{aligned}$$

□

Theorem 2.2.8 (Chebyshev's Inequality).

If $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then $P\left(|x-\mu| \geq k\right) \leq \dfrac{\sigma^2}{k^2}$, i.e., $P\left(|x-\mu| \geq k\sigma\right) \leq \dfrac{1}{k^2}$.

Proof.

$$\begin{aligned}
P\left(|x-\mu| \geq k\right) &= P\left(|x-\mu|^2 \geq k^2\right) \\
&\leq \frac{\mathrm{E}\left(|x-\mu|^2\right)}{k^2} \\
&= \frac{\sigma^2}{k^2}.
\end{aligned}$$

□

Definition 2.2.9 (Asymptotically Unbiased Estimator).

We say that an estimator $\hat{\theta}$ an <u>asymptotically unbiased estimator</u> if

$$\mathrm{E}_\theta\left(\hat{\theta}(X_1, X_2, \ldots, X_n)\right) \to \theta$$

as $n \to \infty$.

Theorem 2.2.10.

If $\hat{\theta}$ is unbiased or asymptotically unbiased, and $\mathrm{Var}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta} \xrightarrow{P} \theta$.

Proof.

$$\begin{aligned}
\mathrm{E}\left(\left(\hat{\theta}-\theta\right)^2\right) &= \mathrm{E}\left(\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)+\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2\right) \\
&= \mathrm{E}\left(\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)^2+2\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)+\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2\right) \\
&= \mathrm{E}\left(\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)^2\right)+\mathrm{E}\left(2\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)\right)+\mathrm{E}\left(\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2\right) \\
&= \mathrm{Var}\left(\hat{\theta}\right)+2\,\mathrm{E}\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)\mathrm{E}\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)+\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2 \\
&= \mathrm{Var}\left(\hat{\theta}\right)+2\cdot 0\cdot \mathrm{E}\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)+\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2 \\
&= \mathrm{Var}\left(\hat{\theta}\right)+\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2.
\end{aligned}$$

Let $\varepsilon > 0$. Then, by Markov's Inequality,

$$\begin{aligned}
P\left(\left|\hat{\theta}-\theta\right| \geq \varepsilon\right) &= P\left(\left(\hat{\theta}-\theta\right)^2 \geq \varepsilon^2\right) \\
&\leq \frac{\mathrm{E}\left(\left(\hat{\theta}-\theta\right)^2\right)}{\varepsilon^2} \\
&= \frac{\mathrm{Var}\left(\hat{\theta}\right)+\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2}{\varepsilon^2} \\
&\to \frac{0+0^2}{\varepsilon^2} \\
&= 0
\end{aligned}$$

as $n \to \infty$. Hence, $\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right| \geq \varepsilon\right) = 0$ for any $\varepsilon > 0$. Therefore, $\hat{\theta} \xrightarrow{P} \theta$. □

Theorem 2.2.11 (Weak Law of Large Numbers).

If $X_1, X_2, \ldots, X_n$ is a random sample with mean μ and finite variance σ^2, then $\overline{X} \xrightarrow{P} \mu$.

Proof. We would like to apply the previous theorem to prove this theorem. We first check that whether $\overline{X}$ is unbiased or asymptotically unbiased or not.

$$
\begin{aligned}
\mathrm{E}\left(\overline{X}\right) &= \mathrm{E}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) \\
&= \frac{1}{n}\,\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(X_i\right) \\
&= \frac{1}{n}\cdot n \cdot \mu \\
&= \mu.
\end{aligned}
$$

Thus, $\overline{X}$ is unbiased. We now check whether $\operatorname{Var}\left(\overline{X}\right) \to 0$.

$$
\begin{aligned}
\operatorname{Var}\left(\overline{X}\right) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) \\
&= \frac{1}{n^2}\cdot n \cdot \sigma^2 \\
&= \frac{\sigma^2}{n} \\
&\to 0
\end{aligned}
$$

as $n \to \infty$. Thus, $\operatorname{Var}\left(\overline{X}\right) \to 0$. By Theorem 2.2.10, $\overline{X} \xrightarrow{P} \mu$. □

Definition 2.2.12 (Consistent Estimator).

We say that $\hat{\theta}$ is a consistent estimator of θ if $\hat{\theta} \xrightarrow{P} \theta$.

Example 2.2.13.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with mean μ and variance σ^2. Is X_1 consistent for μ?

Solution. No. Since $\mathrm{E}\left(X_1\right) = \mu$, X_1 is unbiased. Let $\varepsilon > 0$. Then,

$$
\begin{aligned}
P\left(|X_1-\mu|\geq\varepsilon\right) &= 1-P\left(|X_1-\mu|\leq\varepsilon\right)\\
&= 1-P\left(\mu-\varepsilon\leq X_1\leq\mu+\varepsilon\right)\\
&= 1-\int_{\mu-\varepsilon}^{\mu+\varepsilon} f_{X_1}(x)\,\mathrm{d}x\\
&> 0
\end{aligned}
$$

and is not dependent on n. Hence, $X_1 \overset{P}{\not\to} \mu$. ■

Remark 2.2.14.

We usually use $\overline{X}$ as a consistent estimator of μ since $\mathrm{E}\left(\overline{X}\right)=\mu$ and $\mathrm{Var}\left(\overline{X}\right)=\dfrac{\sigma^2}{n}\to 0$ as $n\to\infty$.

Definition 2.2.15 (Moment).

Let X be a random variable with probability distribution function $f(x,\theta)$. The population kth moment is defined by

$$
\mathrm{E}_\theta\left(X^k\right)=\begin{cases}\displaystyle\sum_{x\in X(S)} x^k f(x,\theta), & \text{if } X \text{ is discrete;}\\ \displaystyle\int_{-\infty}^{\infty} x^k f(x,\theta)\,\mathrm{d}x, & \text{if } X \text{ is continuous.}\end{cases}
$$

Problem 2.2.16.

Let $X_1,X_2,\ldots,X_n$ be independent and identically distributed with mean μ and variance σ^2. Is the sample kth moment $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}{X_i}^k$ consistent for the population kth moment $\mathrm{E}_\theta\left(X^k\right)$?

Solution. Yes. Since $X_1,X_2,\ldots,X_n$ are independent and identically distributed, ${X_1}^k,{X_2}^k,\ldots,{X_n}^k$ are also independent and identically distributed with mean $\mathrm{E}\left(X^k\right)$ and variance $\mathrm{Var}\left(X^k\right)$. By the weak law of large numbers, $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}{X_i}^k\xrightarrow{P}\mathrm{E}_\theta\left(X^k\right)$. □

Remark 2.2.17.

By definitions, we have the equality $\mathrm{E}\left(X^2\right)=\sigma^2+\mu^2$.

Problem 2.2.18.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with mean μ and variance σ^2. Is the sample variance $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2$ unbiased or consistent for the population variance σ^2?

Solution. The sample variance is both unbiased and consistent for the population variance. We first show that s^2 is unbiased for σ^2.

$$
\begin{aligned}
\mathrm{E}\left(s^2\right) &= \mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= \frac{1}{n-1}\,\mathrm{E}\left(\sum_{i=1}^{n}\left({X_i}^2 - 2X_i\overline{X} + \overline{X}^2\right)\right) \\
&= \frac{1}{n-1}\,\mathrm{E}\left(\sum_{i=1}^{n}{X_i}^2 - 2\overline{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\overline{X}^2\right) \\
&= \frac{1}{n-1}\,\mathrm{E}\left(\sum_{i=1}^{n}{X_i}^2 - 2n\overline{X}^2 + n\overline{X}^2\right) \\
&= \frac{1}{n-1}\,\mathrm{E}\left(\sum_{i=1}^{n}{X_i}^2 - n\overline{X}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathrm{E}\left({X_i}^2\right) - n\,\mathrm{E}\left(\overline{X}^2\right)\right) \\
&= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n\left(\sigma^2 + \frac{\sigma^2}{n}\right)\right) \\
&= \frac{1}{n-1}\left((n-1)\sigma^2\right) \\
&= \sigma^2.
\end{aligned}
$$

Hence, s^2 is unbiased for σ^2. We now show that s^2 is consistent for σ^2.

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n}{X_i}^2 - n\overline{X}^2\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}{X_i}^2 - \overline{X}^2\right) \\
&\xrightarrow{P} 1 \cdot \left(\mathrm{E}\left(X^2\right) - \left(\mathrm{E}\left(X\right)\right)^2\right) \\
&= \mathrm{Var}(X)
\end{aligned}
$$

as $n \to \infty$. Hence, s^2 is consistent for σ^2. $\square$

2.3 Estimation Theory

Definition 2.3.1 (Method of Moment Estimator).

The method of moment estimator is the solution to estimation of θ with estimating population moments by sample moments, i.e., for $k \in \{1, 2, \ldots, |\theta|\}$, set

$$\mathrm{E}_\theta\left(X^k\right) = \frac{1}{n}\sum_{i=1}^{n} {X_i}^k.$$

Example 2.3.2.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Bernoulli(p). Find the method of moment estimator $\hat{p}$ for p by $X_1, X_2, \ldots, X_n$.

Solution. Set

$$\mathrm{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, $\hat{p} = \overline{X}$. It is clear that $\hat{p} = \overline{X}$ is unbiased. We now check it is consistent. By the weak law of large numbers, $\hat{p} = \overline{X} \xrightarrow{P} \mu$, and hence $\hat{p} = \overline{X}$ is consistent. ■

Example 2.3.3.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Poisson(λ). Find the method of moment estimator $\hat{\lambda}$ for λ by $X_1, X_2, \ldots, X_n$.

Proof. Set

$$\mathrm{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, $\hat{\lambda} = \overline{X}$. It is clear that $\hat{\lambda} = \overline{X}$ is unbiased. We now check it is consistent. By the weak law of large numbers, $\hat{\lambda} = \overline{X} \xrightarrow{P} \mu$, and hence $\hat{\lambda} = \overline{X}$ is consistent. ■

Example 2.3.4.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Find the method of moment estimators $\hat{\mu}$ for μ and an $\hat{\sigma^2}$ for σ^2 by $X_1, X_2, \ldots, X_n$.

Solution. Set

$$\mathrm{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i;$$
$$\mathrm{E}\left(X^2\right) = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2.$$

Then, the first equation implies $\hat{\mu} = \overline{X}$. It is clear that $\hat{\mu} = \overline{X}$ is unbiased and consistent. For the second equation,

$$\begin{aligned} \mu^2 + \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 \\ \Longrightarrow \qquad \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \hat{\mu}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2. \end{aligned}$$

We now check whether $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ is unbiased or not.

$$\begin{aligned} \mathrm{E}\left(\hat{\sigma^2}\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right) \\ &= \frac{n-1}{n}\,\mathrm{E}\left(s^2\right) \\ &= \frac{n-1}{n}\sigma^2 \\ &\neq \sigma^2, \end{aligned}$$

which implies $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ is not unbiased. We now check whether $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ is consistent or not.

$$\begin{aligned} \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \\ &\xrightarrow{P} \mathrm{E}\left(X^2\right) - \mu^2 \\ &= \left(\mu^2 + \sigma^2\right) - \mu^2 \\ &= \sigma^2. \end{aligned}$$

Hence, $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ is consistent. ■

Remark 2.3.5.

Let $X_1, X_2, \dots, X_n$ be a random sample from $f(x, \theta)$. The joint probability density function of $X_1, X_2, \dots, X_n$ is

$$f(x_1, x_2, \dots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

for $x_i \in \mathbb{R}$. As a joint probability density function, it satisfies

$$\int \cdots \int_{\mathbb{R}^n} f(x_1, x_2, \dots, x_n, \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1$$

for all $\theta \in \Theta$, where Θ is the parameter space.

Definition 2.3.6 (Likelihood Function).

The likelihood function of a random sample is its joint probability density function viewed as a function L of the parameters θ

$$L(\theta) = L(\theta, x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n, \theta)$$

for all $\theta \in \Theta$. With $(x_1, x_2, \dots, x_n)$ fixed, the value $L(\theta, x_1, x_2, \dots, x_n)$ is called the likelihood at θ.

Remark 2.3.7.

(a) If $L(\theta_1, \mathbf{x}) > L(\theta_2, \mathbf{x})$, we consider θ_1 is more reliable than θ_2 when $\mathbf{x}$ is observed.

(b) The value $L(\theta, \mathbf{x})$ is considered as the probability that $\mathbf{X} = \mathbf{x}$ occurs when θ is true.

Definition 2.3.8 (Maximum Likelihood Estimator).

Let $\hat{\theta} = \hat{\theta(\mathbf{x})}$ be any value of θ that maximizes $L(\theta, \mathbf{x})$. Then, we call $\hat{\theta} = \hat{\theta(\mathbf{X})}$ the maximum likelihood estimator of θ. When $\mathbf{X} = \mathbf{x}$ is observed, we call $\hat{\theta} = \hat{\theta(\mathbf{x})}$ the maximum likelihood estimate of θ.

Theorem 2.3.9 (Derivation of the maximum likelihood estimator of θ).

We utilize the positive monotone transformation $\ln \cdot$ to look for the maximum likelihood estimator. If $\hat{\theta} = \hat{\theta}(\mathbf{x})$ is the maximum likelihood estimator, then $L(\hat{\theta}, \mathbf{x}) = \max_{\theta \in \Theta} L(\theta, \mathbf{x})$, which is equivalent to $\ln L(\hat{\theta}, \mathbf{x}) = \max_{\theta \in \Theta} \ln L(\theta, \mathbf{x})$. We have two cases for solving the maximum likelihood estimator:

(a) on one of two ends if $L(\theta)$ is monotone;

(b) $\dfrac{\partial \ln L(\theta)}{\partial \theta} = 0$.

Definition 2.3.10 (Order Statistic).

Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with probability density function $f(x, \theta)$. Let $Y_1, Y_2, \ldots, Y_n$ be the permutation of $X_1, X_2, \ldots, X_n$ such that $Y_i \leq Y_{i+1}$ for $i = 1, 2, \ldots, n-1$. We call $(Y_1, Y_2, \ldots, Y_n)$ the order statistics of $X_1, X_2, \ldots, X_n$.

Theorem 2.3.11 (Largest Order Statistic).

The probability density function of the largest order statistic $Y_n = \max\{X_1, X_2, \ldots, X_n\}$ is

$$f_{Y_n}(y) = n \left(F(y, \theta)\right)^{n-1} f(y, \theta),$$

where $F(\cdot, \theta)$ is the distribution function of X_i and $f(\cdot, \theta)$ is the probability density function of X_i.

Proof. The distribution function of Y_n is

$$\begin{aligned}
F_{Y_n}(y) &= P(Y_n \leq y) \\
&= P(\max\{X_1, X_2, \ldots, X_n\} \leq y) \\
&= P(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y) \\
&= \prod_{i=1}^{n} P(X_i \leq y) \\
&= \left(F(y, \theta)\right)^n .
\end{aligned}$$

Hence, the probability density function of Y_n is

$$
\begin{aligned}
f_{Y_n}(y) &= \frac{\partial}{\partial y}\left(F(y,\theta)\right)^n \\
&= n\left(F(y,\theta)\right)^{n-1}\cdot\left(\frac{\partial}{\partial y}F(y,\theta)\right) \\
&= n\left(F(y,\theta)\right)^{n-1} f(y,\theta).
\end{aligned}
$$

□

Theorem 2.3.12 (Smallest Order Statistic).

The probability density function of the largest order statistic $Y_1 = \min\{X_1, X_2, \ldots, X_n\}$ is

$$f_{Y_1}(y) = n\left(1 - F(y,\theta)\right)^{n-1} f(y,\theta),$$

where $F(\cdot,\theta)$ is the distribution function of X_i and $f(\cdot,\theta)$ is the probability density function of X_i.

Proof. The distribution function of Y_n is

$$
\begin{aligned}
F_{Y_1}(y) &= P(Y_1 \le y) \\
&= 1 - P(Y_1 \ge y) \\
&= 1 - P(\min\{X_1, X_2, \ldots, X_n\} \ge y) \\
&= 1 - P(X_1 \ge y, X_2 \ge y, \ldots, X_n \ge y) \\
&= 1 - \prod_{i=1}^{n} P(X_i \ge y) \\
&= 1 - \left(1 - F(y,\theta)\right)^n.
\end{aligned}
$$

Hence, the probability density function of Y_n is

$$
\begin{aligned}
f_{Y_n}(y) &= \frac{\partial}{\partial y}\left(1 - \left(1 - F(y,\theta)\right)^n\right) \\
&= -n\left(1 - F(y,\theta)\right)^{n-1}\cdot\left(\frac{\partial}{\partial y}\left(-F(y,\theta)\right)\right) \\
&= n\left(1 - F(y,\theta)\right)^{n-1} f(y,\theta).
\end{aligned}
$$

□

Example 2.3.13.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $U(0,\theta)$. Find the maximum likelihood estimator for θ and check whether it is unbiased or consistent or not.

Solution. The probability density function of X is

$$f(x,\theta) = \frac{1}{\theta} I(0 \le x \le \theta).$$

The likelihood function of $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i,\theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \le x_i \le \theta). \end{aligned}$$

Let $Y_n = \max\{X_1, X_2, \ldots, X_n\}$. Then,

$$\begin{aligned} L(\theta) &= \frac{1}{\theta^n} I(0 \le y_n \le \theta) \\ &= \frac{1}{\theta^n} I(y_n \le \theta < \infty) \\ &= \begin{cases} \dfrac{1}{\theta^n}, & \text{if } y_n \le \theta < \infty; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, $L(\theta)$ is maximized when $\theta = y_n$, and the maximum likelihood estimator $\hat{\theta} = Y_n$. The distribution function of X is

$$\begin{aligned} F_X(x) &= \int_0^x \frac{1}{\theta}\,\mathrm{d}t \\ &= \frac{x}{\theta} \end{aligned}$$

for $x \in (0,\theta)$. Thus, the probability density function of $\theta = Y_n$ is

$$\begin{aligned} f_{Y_n}(y) &= n\left(\frac{y}{\theta}\right)^{n-1} \frac{1}{\theta} \\ &= n\frac{y^{n-1}}{\theta^n} \end{aligned}$$

for $y \in (0,\theta)$. Hence, the expected value of $\theta = Y_n$ is

$$\begin{aligned} \mathrm{E}(Y_n) &= \int_0^\theta y \cdot n\frac{y^{n-1}}{\theta^n}\,\mathrm{d}y \\ &= \frac{n}{n+1}\theta, \end{aligned}$$

which implies $\theta = Y_n$ is not unbiased but is asymptotically unbiased. Moreover, the second population moment of Y_n is

$$
\begin{aligned}
\mathrm{E}({Y_n}^2) &= \int_0^\theta y^2 \cdot n \frac{y^{n-1}}{\theta^n}\,\mathrm{d}y \\
&= \frac{n}{n+2}\theta^2,
\end{aligned}
$$

which implies

$$
\begin{aligned}
\mathrm{Var}(Y_n) &= \mathrm{E}({Y_n}^2) - (\mathrm{E}(Y_n))^2 \\
&= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 \\
&\to \theta^2 - \theta^2 \\
&= 0
\end{aligned}
$$

as $n \to \infty$. Thus, the maximum likelihood estimator $\hat{\theta}$ is consistent for θ. ■

Example 2.3.14.

Let $Y \sim B(n,p)$. Find the maximum likelihood estimator for p and check whether it is unbiased or consistent or not.

Solution. Since there is only one sample, the likelihood function of Y is $L(p) = \binom{n}{y} p^y (1-p)^{n-y}$. Set $\dfrac{\partial \ln L(p)}{\partial p} = 0$. Then,

$$
\begin{aligned}
\frac{\partial \ln L(p)}{\partial p} &= 0 \\
\frac{\partial}{\partial p}\left(y \ln p + (n-y)\ln(1-p)\right) &= 0 \\
\frac{y}{p} - \frac{n-y}{1-p} &= 0 \\
\frac{y}{p} &= \frac{n-y}{1-p} \\
\implies \qquad p &= \frac{y}{n}.
\end{aligned}
$$

Hence, $\hat{p} = \dfrac{Y}{n}$. It is clear that $\hat{p}$ is unbiased. Moreover, $\mathrm{Var}(\hat{p}) = \dfrac{1}{n^2}\mathrm{Var}(Y) \to 0$. Therefore, $\hat{p}$ is consistent for p. ■

Example 2.3.15.

Let $X_1, X_2, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Find the maximum likelihood estimator for μ and for σ^2 and check whether they are unbiased or consistent or not.

Solution. The likelihood function for $X_1, X_2, \dots, X_n$ is

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

We first look for $\hat{\mu}$. Set $\dfrac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = 0$. Then,

$$\begin{aligned} \frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} &= 0 \\ \frac{\partial}{\partial \mu}\left(-\frac{n}{2}\ln(2\pi) - n\ln\sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right) &= 0 \\ -2\sum_{i=1}^{n}(x_i - \mu) &= 0 \\ \\ \mu &= \overline{x_i}. \end{aligned}$$

Hence, $\hat{\mu} = \overline{X}$. We now look for $\hat{\sigma^2}$. Set $\dfrac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} = 0$. Then,

$$\begin{aligned} \frac{\partial \ln L(\mu, \sigma^2)}{\partial \sigma^2} &= 0 \\ \frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2}\ln(2\pi) - n\ln\sigma - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right) &= 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 &= 0 \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2. \end{aligned}$$

Hence, $\hat{\sigma^2} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. We now check whether $\hat{\mu}$ is unbiased or consistent or not. It is clear that $\hat{\mu}$ is unbiased. Moreover, $\mathrm{Var}(\hat{\mu}) = \dfrac{\sigma^2}{n} \to 0$ as $n \to \infty$. Therefore, $\hat{\mu}$ is consistent for μ. We now check whether $\hat{\sigma^2}$ is unbiased or consistent or not. The expectation

$$
\begin{aligned}
\mathrm{E}\left(\hat{\sigma^2}\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= \frac{n-1}{n}\,\mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&\to \frac{n-1}{n}\sigma^2,
\end{aligned}
$$

which implies $\hat{\sigma^2}$ is not unbiased but is asymptotically unbiased. The variance

$$
\begin{aligned}
\mathrm{Var}\left(\hat{\sigma^2}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= \mathrm{Var}\left(\frac{\sigma^2}{n}\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= \frac{\sigma^4}{n^2}\,\mathrm{Var}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\right),
\end{aligned}
$$

where $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 \sim \chi^2(n-1)$. Hence,

$$
\begin{aligned}
\mathrm{Var}\left(\hat{\sigma^2}\right) &= \frac{\sigma^4}{n^2}\,\mathrm{Var}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= \frac{\sigma^4}{n^2}\cdot 2(n-1) \\
&\to 0
\end{aligned}
$$

as $n \to \infty$. Therefore, $\hat{\sigma^2}$ is consistent for σ^2. ∎

Theorem 2.3.16.

Suppose that we have the maximum likelihood estimator of θ as $\hat{\theta} = \hat{\theta}(\mathbf{X})$. Let τ be an injective function of θ. Then, the maximum likelihood estimator of $\tau(\theta)$ is $\tau\left(\hat{\theta}\right)$.

Proof. Define $T = \tau(\Theta)$ as the space of $\tau(\theta)$. If the likelihood function for θ is $L(\theta, \mathbf{x})$, the likelihood function for $\tau(\theta)$ can be derived as follows:

$$
\begin{aligned}
L(\theta, \mathbf{x}) &= L(\tau^{-1}\left(\tau(\theta)\right), \mathbf{x}) \\
&= L^*(\tau(\theta), \mathbf{x}) \\
&= L^*(\tau, \mathbf{x})
\end{aligned}
$$

for all $\tau \in T$, where $L^*(\cdot, \mathbf{x}) = L(\tau^{-1}(\cdot), \mathbf{x})$. Notice that $L^*(\tau, \mathbf{x})$ is a likelihood function for $\tau(\theta)$ at τ. We now substitute $\tau = \tau\left(\hat{\theta}\right)$ and see what happens:

$$\begin{aligned} L^*(\tau\left(\hat{\theta}\right), \mathbf{x}) &= L(\hat{\theta}, \mathbf{x}) \\ &\geq L(\theta, \mathbf{x}) \\ &= L(\tau^{-1}\left(\tau(\theta)\right), \mathbf{x}) \\ &= L^*(\tau\left(\theta\right), \mathbf{x}) \\ &= L^*(\tau, \mathbf{x}), \end{aligned}$$

for all $\tau \in T$, where θ is any element in Θ. Hence, $\tau\left(\hat{\theta}\right)$ is the maximum likelihood estimator of $\tau(\theta)$. $\square$

Example 2.3.17.

If $Y \sim B(n, p)$, then the maximum likelihood estimator for p is $\hat{p} = \dfrac{Y}{n}$. Moreover, we can obtain the following table by the previous theorem:

$\tau(p)$	MLE $\widehat{\tau(p)}$
p^2	$\left(\dfrac{Y}{n}\right)^2$
$\sqrt{p}$	$\sqrt{\dfrac{Y}{n}}$
e^p	$e^{\frac{Y}{n}}$

Example 2.3.18.

If $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, then the maximum likelihood estimator for (μ, σ^2) is

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

By the previous theorem, the maximum likelihood estimator for (μ, σ) is

$$(\hat{\mu}, \hat{\sigma}) = \left(\overline{X}, \sqrt{\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n}}\right).$$

Remark 2.3.19.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $f(x, \theta)$. If $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is an estimator of θ, then its mean

$$E_\theta\left(\hat{\theta}\right) = \int \cdots \int_{\mathbb{R}^n} \hat{\theta}(X_1, X_2, \ldots, X_n)\, f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

is a function of θ, and its variance

$$E_\theta\left(\left(\hat{\theta} - \mathrm{E}\left(\hat{\theta}\right)\right)^2\right) = \int \cdots \int_{\mathbb{R}^n} \left(\hat{\theta}(X_1, X_2, \ldots, X_n)\right)^2 f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

is also a function of θ.

2.4 UMVUE: One Concept of the Best Estimation

Definition 2.4.1 (Uniformly Minimum Variance Unbiased Estimator).

Let $\theta \in \Theta$. If $\hat{\theta}$ is unbiased and satisfies

$$\mathrm{Var}_{\theta_0}\left(\hat{\theta}\right) \leq \mathrm{Var}_{\theta_0}\left(\hat{\theta}^*\right)$$

for any unbiased estimator $\hat{\theta}^*$, then $\hat{\theta}$ is the <u>minimum variance unbiased estimator</u> when $\theta = \theta_0$ is true. We call $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ a <u>uniformly minimum variance unbiased estimator</u> of θ or the best estimation of θ if for all $\theta \in \Theta$, we have

$$\mathrm{Var}_{\theta}\left(\hat{\theta}\right) \leq \mathrm{Var}_{\theta}\left(\hat{\theta}^*\right).$$

Definition 2.4.2 (Regularity Conditions).

We say the <u>regularity conditions</u> hold if all of the following are true:

(a) the parameter space Θ is an open interval;

(b) the set $\{x \mid f(x, \theta) = 0\}$ is independent of θ;

(c) $\displaystyle\int_{-\infty}^{\infty} \frac{\partial f(x, \theta)}{\partial \theta}\, \mathrm{d}x = \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x, \theta)\, \mathrm{d}x = 0$;

(d) if $T = t(\mathbf{X})$ is an unbiased estimation of $\tau(\theta)$, then

$$\int \cdots \int_{\mathbb{R}^n} t(\mathbf{x}) \frac{\partial f(\mathbf{x})}{\partial \theta}\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = \frac{\partial}{\partial \theta}\left(\int \cdots \int_{\mathbb{R}^n} t(\mathbf{x})\, f(\mathbf{x})\, \mathrm{d}x_1 \cdots \mathrm{d}x_n\right).$$

Theorem 2.4.3 (Cauchy-Schwarz Inequality).

For any two random variables X and Y, we have

$$(\mathrm{E}\,(XY))^2 \geq \mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right).$$

Theorem 2.4.4 (Cramér–Rao Inequality).

Let $X, X_1, X_2, \ldots, X_n$ be a random sample with distribution $f(x,\theta)$. Suppose that the regularity conditions hold. If $\hat{\tau(\theta)} = t(\mathbf{X})$ is unbiased for $\tau(\theta)$, then

$$\mathrm{Var}_\theta\left(\hat{\tau(\theta)}\right) \geq \frac{(\tau'(\theta))^2}{n\,\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(X,\theta)\right)^2\right)} = \frac{(\tau'(\theta))^2}{-n\,\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right)}.$$

The expression on the right hand side $\dfrac{(\tau'(\theta))^2}{n\,\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln f(X,\theta)\right)^2\right)}$ is called the Cramér–Rao lower bound.

Proof. We only prove the case for continuous random variables. Note that the expectation

$$\begin{aligned}
\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right) &= \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right) f(x,\theta)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\frac{1}{f(X,\theta)}\cdot\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)\cdot f(x,\theta)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}f(x,\theta)\right)\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}f(x,\theta)\,\mathrm{d}x \\
&= 0.
\end{aligned}$$

Since $\hat{\tau(\theta)} = t(\mathbf{X})$ is unbiased for $\tau(\theta)$, we have $\mathrm{E}_\theta\left(\hat{\tau(\theta)}\right) = \tau(\theta)$. Moreover, we have

$$\begin{aligned}
\tau(\theta) &= \mathrm{E}_\theta\left(\hat{\tau(\theta)}\right) \\
&= \mathrm{E}_\theta\left(t(\mathbf{x})\right) \\
&= \int\cdots\int_{\mathbb{R}^n} t(\mathbf{x})\, f(\mathbf{x},\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= \int\cdots\int_{\mathbb{R}^n} t(\mathbf{x})\prod_{i=1}^{n} f(x_i,\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n.
\end{aligned}$$

Notice that the integral

$$\int \cdots \int_{\mathbb{R}^n} \prod_{i=1}^{n} f(x_i, \theta) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n = 1.$$

Thus, differentiating both sides yields

$$\begin{aligned}
\tau'(\theta) &= \frac{\partial}{\partial \theta} \int \cdots \int_{\mathbb{R}^n} t(\mathbf{x}) \prod_{i=1}^{n} f(x_i, \theta) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \frac{\partial}{\partial \theta} \int \cdots \int_{\mathbb{R}^n} t(\mathbf{x}) \prod_{i=1}^{n} f(x_i, \theta) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n - \tau(\theta) \cdot \frac{\partial}{\partial \theta} \int \cdots \int_{\mathbb{R}^n} \prod_{i=1}^{n} f(x_i, \theta) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int \cdots \int_{\mathbb{R}^n} t(\mathbf{x}) \frac{\partial}{\partial \theta} \left(\prod_{i=1}^{n} f(x_i, \theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n - \int \cdots \int_{\mathbb{R}^n} \tau(\theta) \frac{\partial}{\partial \theta} \left(\prod_{i=1}^{n} f(x_i, \theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int \cdots \int_{\mathbb{R}^n} (t(\mathbf{x}) - \tau(\theta)) \frac{\partial}{\partial \theta} \left(\prod_{i=1}^{n} f(x_i, \theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int \cdots \int_{\mathbb{R}^n} (t(\mathbf{x}) - \tau(\theta)) \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} f(x_i, \theta) \right) \prod_{j \neq i} f(x_j, \theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int \cdots \int_{\mathbb{R}^n} (t(\mathbf{x}) - \tau(\theta)) \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right) \cdot f(x_i, \theta) \cdot \prod_{j \neq i} f(x_j, \theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int \cdots \int_{\mathbb{R}^n} (t(\mathbf{x}) - \tau(\theta)) \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right) \cdot \prod_{j=1}^{n} f(x_j, \theta) \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int \cdots \int_{\mathbb{R}^n} (t(\mathbf{x}) - \tau(\theta)) \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right) \right) \prod_{j=1}^{n} f(x_j, \theta) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \mathrm{E} \left((t(\mathbf{x}) - \tau(\theta)) \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right) \right) \right).
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\left(\tau'(\theta)\right)^2 &= \left(\mathrm{E} \left((t(\mathbf{x}) - \tau(\theta)) \left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right) \right) \right) \right)^2 \\
&\leq \mathrm{E} \left((t(\mathbf{x}) - \tau(\theta))^2 \right) \cdot \mathrm{E} \left(\left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right) \right)^2 \right)
\end{aligned}$$

$$\implies \mathrm{E} \left((t(\mathbf{x}) - \tau(\theta))^2 \right) \geq \frac{\left(\tau'(\theta)\right)^2}{\mathrm{E} \left(\left(\sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta} \ln f(x_i, \theta) \right) \right)^2 \right)},$$

where $\mathrm{E} \left((t(\mathbf{x}) - \tau(\theta))^2 \right) = \mathrm{Var} \left(\tau(\hat{\theta}) \right)$. Now,

$$
\begin{aligned}
\mathrm{E}\left(\left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)\right)^2\right) &= \mathrm{E}\left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)^2\right) + \mathrm{E}\sum_{j\neq i}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\cdot\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\right) \\
&= \mathrm{E}\left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)^2\right) + \sum_{j\neq i}\mathrm{E}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\cdot\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\right) \\
&= \mathrm{E}\left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)^2\right) \\
&\quad + \sum_{j\neq i}\mathrm{E}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)\mathrm{E}\left(\frac{\partial}{\partial\theta}\ln f(x_j,\theta)\right) \\
&= \mathrm{E}\left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)^2\right) + \sum_{j\neq i}0\cdot 0 \\
&= \mathrm{E}\left(\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)^2\right) \\
&= n\,\mathrm{E}\left(\left(\frac{\partial}{\partial\theta}\ln f(x_i,\theta)\right)^2\right).
\end{aligned}
$$

For the equality, we know that $\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta)\,\mathrm{d}x = \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)\cdot f(x,\theta)\,\mathrm{d}x = 0$ by the regularity conditions. Taking the partial derivative with respect to θ on both sides yields

$$
\begin{aligned}
\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)\cdot f(x,\theta)\,\mathrm{d}x &= 0 \\
\frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)\cdot f(x,\theta)\,\mathrm{d}x\right) &= \frac{\partial}{\partial\theta}0 \\
\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right)\cdot f(x,\theta) + \left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)\left(\frac{\partial}{\partial\theta}f(x,\theta)\right)\,\mathrm{d}x &= 0 \\
\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right)\cdot f(x,\theta) + \left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)^2 f(x,\theta)\,\mathrm{d}x &= 0 \\
-\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right)\cdot f(x,\theta)\,\mathrm{d}x &= \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)^2 f(x,\theta)\,\mathrm{d}x \\
-n\,\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\right) &= n\,\mathrm{E}\left(\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)^2\right).
\end{aligned}
$$

□

Example 2.4.5.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Poisson(λ). Show that the maximum likelihood estimator $\hat{\lambda}$ is a uniform minimum variance unbiased estimator.

Solution. The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda} \frac{\lambda^{n\overline{x}}}{\prod_{i=1}^{n} x_i!}. \end{aligned}$$

Set $\dfrac{\partial \ln L(\lambda)}{\partial \lambda} = 0$. Then,

$$\begin{aligned} \frac{\partial \ln L(\lambda)}{\partial \lambda} &= 0 \\ \frac{\partial}{\partial \lambda}\left(-n\lambda + n\overline{x}\ln\lambda - \sum_{i=1}^{n} \ln(x_i!)\right) &= 0 \\ -n + \frac{n\overline{x}}{\lambda} &= 0 \\ \frac{\overline{x}}{\lambda} &= 1. \end{aligned}$$

Hence, $\hat{\lambda} = \overline{X}$. It is clear that $\hat{\lambda} = \overline{X}$ is unbiased. Moreover, the variance of $\hat{\lambda} = \overline{X}$

$$\begin{aligned} \operatorname{Var}\left(\hat{\lambda}\right) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}(X_i) \\ &= \frac{1}{n^2}\cdot n \cdot \lambda \\ &= \frac{\lambda}{n} \end{aligned}$$

The Cramér–Rao bound

$$\begin{aligned} \frac{(\tau'(\theta))^2}{-n\,\mathrm{E}_\theta\left(\dfrac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right)} &= \frac{1}{-n\,\mathrm{E}\left(\dfrac{\partial^2}{\partial\lambda^2}\left(\ln\dfrac{e^{-\lambda}\lambda^X}{X!}\right)\right)} \\ &= \frac{1}{-n\,\mathrm{E}\left(\dfrac{\partial^2}{\partial\lambda^2}(-\lambda + X\ln\lambda - \ln(X!))\right)} \\ &= \frac{1}{-n\,\mathrm{E}\left(\dfrac{\partial}{\partial\lambda}\left(-1 + \dfrac{X}{\lambda}\right)\right)} \\ &= \frac{1}{-n\,\mathrm{E}\left(-\dfrac{X}{\lambda^2}\right)} \\ &= \frac{\lambda}{n}, \end{aligned}$$

which equals $\operatorname{Var}\left(\hat{\lambda}\right)$. Therefore, $\hat{\lambda} = \overline{X}$ is a uniform minimum variance unbiased estimator of λ. ■

Example 2.4.6.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Bernoulli(p). Find a uniform minimum variance unbiased estimator of p.

Solution. We first look for the Cramér–Rao bound:

$$\begin{aligned}
\frac{(\tau'(\theta))^2}{-n\,\mathrm{E}_\theta\left(\dfrac{\partial^2}{\partial\theta^2}\ln f(X,\theta)\right)} &= \frac{1}{-n\,\mathrm{E}\left(\dfrac{\partial^2}{\partial p^2}\ln\left(p^X(1-p)^{1-X}\right)\right)} \\
&= \frac{1}{-n\,\mathrm{E}\left(\dfrac{\partial^2}{\partial p^2}\left(X\ln p + (1-X)\ln(1-p)\right)\right)} \\
&= \frac{1}{-n\,\mathrm{E}\left(\dfrac{\partial}{\partial p}\left(\dfrac{X}{p} - \dfrac{1-X}{1-p}\right)\right)} \\
&= \frac{1}{-n\,\mathrm{E}\left(-\dfrac{X}{p^2} - \dfrac{1-X}{(1-p)^2}\right)} \\
&= \frac{1}{n\,\mathrm{E}\left(\dfrac{X}{p^2}\right) + n\,\mathrm{E}\left(\dfrac{1-X}{(1-p)^2}\right)} \\
&= \frac{1}{\dfrac{n}{p} + \dfrac{n}{1-p}} \\
&= \frac{p(1-p)}{n}.
\end{aligned}$$

We guess that the maximum likelihood estimator of p is a uniform minimum variance unbiased estimator of p. Let's check whether it is true. The likelihood function

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

Set $\dfrac{\partial \ln L(p)}{\partial p} = 0$. Then,

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

and

$$
\begin{aligned}
\frac{\partial \ln L(p)}{\partial p} &= 0 \\
\sum_{i=1}^{n} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p} \right) &= 0 \\
\sum_{i=1}^{n} ((1-p)x_i - p(1-x_i)) &= 0 \\
\sum_{i=1}^{n} x_i &= np \\
p &= \overline{x}.
\end{aligned}
$$

Hence, $\hat{p} = \overline{X}$. It is clear that $\hat{p} = \overline{X}$ is unbiased. Moreover, the variance

$$
\begin{aligned}
\operatorname{Var}(\hat{p}) &= \operatorname{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X_i) \\
&= \frac{1}{n^2} \cdot n \cdot p(1-p) \\
&= \frac{p(1-p)}{n},
\end{aligned}
$$

which equals the Cramér–Rao bound. Therefore, $\hat{p} = \overline{X}$ is a uniform minimum variance unbiased estimator of p. ■

Definition 2.4.7 (Conditional Probability).

Let A, B be two events. The conditional probability of $A \subseteq S$ given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

One may show that $P(\cdot \mid B)$ is a probability set function.

Definition 2.4.8 (Conditional Probability Distribution Function).

Let X, Y be two random variables with join probability distribution function $f(x, y)$ and marginal probability distribution functions $f_X(x)$ and $f_Y(y)$. The conditional probability distribution function of Y given $X = x$ is

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}.$$

Remark 2.4.9.

The conditional probability distribution function is still a probability distribution function.

Remark 2.4.10.

In estimation of parameter θ, we have a random sample $X_1, X_2, \ldots, X_n$ from a probability distribution function $f(x, \theta)$. The information we may have about θ is contained in $X_1, X_2, \ldots, X_n$.

Definition 2.4.11 (Conditional Probability Distribution Function).

Suppose that $U = u(X_1, X_2, \ldots, X_n)$ is a statistic (free of parameters θ) with probability distribution function $f_U(u, \theta)$ (possibly with parameters θ). The conditional probability distribution function of $X_1, X_2, \ldots, X_n$ given $U = u$ is

$$f(\mathbf{x}, \theta \mid u) = \frac{f(\mathbf{x}, u, \theta)}{f_U(u, \theta)} = \begin{cases} \dfrac{f(\mathbf{x}, \theta)}{f_U(u, \theta)}, & \text{if } u(\mathbf{x}) = u; \\ 0, & \text{otherwise.} \end{cases}$$

Definition 2.4.12 (Sufficient Statistic).

Let $X_1, X_2, \ldots, X_n$ be a random sample from a probability distribution function $f(x, \theta)$, where $\theta \in \Theta$. We call a statistic $U = u(X_1, X_2, \ldots, X_n)$ a sufficient statistic if for any $U = u$, both the conditional probability distribution function $f(\mathbf{x} \mid u)$ and its domain are not dependent on the parameter θ.

Proposition 2.4.13.

Let $U = \mathbf{X}$ be a random sample as a statistic. Then, U is a sufficient statistic.

Proof. The conditional probability distribution function

$$f(\mathbf{x}, \theta \mid \mathbf{x}') = \frac{\mathbf{x}, \mathbf{x}', \theta}{f(\mathbf{x}', \theta)} = \begin{cases} \dfrac{f(\mathbf{x}, \theta)}{f(\mathbf{x}', \theta)} = 1, & \text{if } \mathbf{x} = \mathbf{x}'; \\ 0, & \text{otherwise,} \end{cases}$$

which is independent of θ. Hence, U is a sufficient statistic. $\square$

Question 2.4.14.

Why do we need sufficiency?

Answer. We want a sufficient statistic with dimension as small as possible. □

Definition 2.4.15 (Minimal Sufficient Statistic).

If $U = u(\mathbf{X})$ is a sufficient statistic with smallest dimension, then it is called the minimal sufficient statistic.

Proposition 2.4.16.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with probability density function $f(x, \theta)$. Consider the order statistics $Y_1 = \min\{X_i\}_{i=1}^n, \ldots, Y_n = \max\{X_i\}_{i=1}^n$. The order statistic $(Y_1, Y_2, \ldots, Y_n)$ is also a sufficient statistic.

Proof. If $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ is observed, the sample $X_1, X_2, \ldots, X_n$ has equal chance being in the set

$$\{\mathbf{x} \mid \mathbf{y} \text{ is a permutation of } \mathbf{x}\}.$$

Then, the conditional probability density function of $X_1, X_2, \ldots, X_n$ given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ is

$$f(\mathbf{x}, \theta \mid \mathbf{y}) = \begin{cases} \dfrac{1}{n!}, & \text{if } \mathbf{y} \text{ is a permutation of } \mathbf{x}; \\ 0, & \text{otherwise,} \end{cases}$$

which is independent of θ. □

Proposition 2.4.17.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Bernoulli(p). Then, the random variable $Y = \sum_{i=1}^n X_i$ is a sufficient statistic.

Proof. Notice that $Y \sim B(n, p)$ with probability mass function $f_Y(y, p) = \binom{n}{y}(p)^y(1-p)^{n-y}$ for $y = 0, 1, 2, \ldots, n$. If $Y = y$, then the space of $\mathbf{X}$ is

$$\left\{\mathbf{x} \;\middle|\; \sum_{i=1}^n x_i = y\right\}.$$

Thus, the conditional probability mass function of $X_1, X_2, \ldots, X_n$ given $Y = y$ is

$$f(\mathbf{x}, p \mid y) = \begin{cases} \dfrac{p^{n\overline{x}}(1-p)^{n-n\overline{x}}}{\dbinom{n}{y}(p)^y(1-p)^{n-y}}, & \text{if } \displaystyle\sum_{i=1}^{n} x_i = y; \\ 0, & \text{otherwise,} \end{cases}$$

which is independent of p. □

Proposition 2.4.18.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $U(0, \theta)$. Then, the largest order statistic Y_n is a sufficient statistic.

Proof. The joint probability density function of $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned} f(\mathbf{x}, \theta) &= \prod_{i=1}^{n} \frac{1}{\theta} I(0 \le x_i \le \theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \le x_i \le \theta) \\ &= \begin{cases} \dfrac{1}{\theta^n}, & \text{if } 0 \le x_i \le \theta \text{ for all } i = 1, 2, \ldots, n; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

The distribution function of X is

$$\begin{aligned} F(x) &= \int_0^x \frac{1}{\theta}\,\mathrm{d}t \\ &= \frac{x}{\theta} \end{aligned}$$

for $x \in (0, \theta)$. Hence, the probability density function of Y_n is

$$\begin{aligned} f_{Y_n}(y) &= n\left(F_{Y_n}(y)\right)^{n-1} f(y, \theta) \\ &= \frac{n x^{y-1}}{\theta^n}, \end{aligned}$$

where $y \le \theta$. Hence, the conditional probability density function of $X_1, X_2, \ldots, X_n$ given $Y_n = y$ is

$$f(\mathbf{x}, \theta \mid y) = \begin{cases} \dfrac{f(\mathbf{x}, \theta)}{f_{Y_n}(y)} = \dfrac{1}{n y^{n-1}}, & \text{if } 0 \le x_i \le y \text{ for all } i = 1, 2, \ldots, n; \\ 0, & \text{otherwise,} \end{cases}$$

which is independent of θ. □

Theorem 2.4.19 (Factorization Theorem).

Let $X_1, X_2, \ldots, X_n$ be a random sample from $f(x, \theta)$. A statistic $U = u(\mathbf{X})$ is sufficient for θ if and only if there exists functions $k_1, k_2 \geq 0$ such that the joint probability distribution function of $X_1, X_2, \ldots, X_n$ can be re-written as

$$f(\mathbf{x}, \theta) = k_1(u(\mathbf{x}), \theta) k_2(\mathbf{x}),$$

where k_2 is independent of θ.

Proof. We only consider continuous random variables.

($\Longrightarrow$) If $U = u(\mathbf{X})$ is sufficient for θ, then

$$f(\mathbf{x} \mid u) = \frac{f(\mathbf{x}, \theta)}{f_U(u, \theta)}$$

is free of θ. Hence,

$$\begin{aligned} f(\mathbf{x} \mid u)\, f_U(u, \theta) &= f(\mathbf{x}, \theta) \\ \Longleftrightarrow \qquad f(\mathbf{x}, \theta) &= f_U(u, \theta)\, f(\mathbf{x} \mid u) \\ &= k_1(u(\mathbf{x}), \theta) k_2(\mathbf{x}). \end{aligned}$$

($\Longleftarrow$) Suppose that $f(\mathbf{x}, \theta) = k_1(u(\mathbf{x}), \theta) k_2(\mathbf{x})$. Let $Y_1 = u(\mathbf{X}), Y_2 = u_2(\mathbf{X}), \ldots, Y_n = u_n(\mathbf{X})$ be an injection with an inverse function $X_1 = w_1(\mathbf{Y}), X_2 = w_2(\mathbf{Y}), \ldots, X_n = w_n(\mathbf{Y})$ and with Jacobian determinant

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \cdots & \dfrac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial x_n}{\partial y_1} & \cdots & \dfrac{\partial x_n}{\partial y_n} \end{vmatrix},$$

which is independent of θ. Thus, the joint probability density function of $Y_1, Y_2, \ldots, Y_n$ is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}, \theta) &= f(\mathbf{x}, \theta)|J| \\ &= k_1(u(\mathbf{x}), \theta) k_2(\mathbf{x})|J| \\ &= k_1(y_1, \theta) k_2(w_1(\mathbf{y}), w_2(\mathbf{y}), \ldots, w_n(\mathbf{y}))|J|. \end{aligned}$$

Hence, the marginal probability density function of $U = Y_1$ is

$$\begin{aligned} f_U(y_1,\theta) &= \int\cdots\int_{\mathbb{R}^{n-1}} k_1(y_1,\theta)k_2(w_1(\mathbf{y}),w_2(\mathbf{y}),\ldots,w_n(\mathbf{y}))|J|\,\mathrm{d}y_2\cdots\mathrm{d}y_n \\ &= k_1(y_1,\theta)\int\cdots\int_{\mathbb{R}^{n-1}} k_2(w_1(\mathbf{y}),w_2(\mathbf{y}),\ldots,w_n(\mathbf{y}))|J|\,\mathrm{d}y_2\cdots\mathrm{d}y_n. \end{aligned}$$

Then, the conditional probability density function of $X_1, X_2, \ldots, X_n$ given $U = y_1$ is

$$\begin{aligned} f(\mathbf{x},\theta \mid y_1) &= \frac{f(\mathbf{x},\theta)}{f_U(y_1,\theta)} \\ &= \frac{k_1(u(\mathbf{x}),\theta)k_2(\mathbf{x})}{k_1(y_1,\theta)\displaystyle\int\cdots\int_{\mathbb{R}^{n-1}} k_2(w_1(\mathbf{y}),w_2(\mathbf{y}),\ldots,w_n(\mathbf{y}))|J|\,\mathrm{d}y_2\cdots\mathrm{d}y_n} \\ &= \frac{k_2(\mathbf{x})}{\displaystyle\int\cdots\int_{\mathbb{R}^{n-1}} k_2(w_1(\mathbf{y}),w_2(\mathbf{y}),\ldots,w_n(\mathbf{y}))|J|\,\mathrm{d}y_2\cdots\mathrm{d}y_n}, \end{aligned}$$

which is independent of θ. Therefore, $U = u(\mathbf{X})$ is sufficient for θ. □

Example 2.4.20.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Poisson(λ). Find a sufficient statistic of λ.

Solution. The joint probability mass function of $X_1, X_2, \ldots, X_n$ is

$$\begin{aligned} f(\mathbf{x},\theta) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda}\lambda^{n\overline{x}} \cdot \frac{1}{\displaystyle\prod_{i=1}^{n} x_i!} \\ &= k_1(n\overline{x},\lambda)\cdot k_2(\mathbf{x}), \end{aligned}$$

where $n\overline{x} = \displaystyle\sum_{i=1}^{n} x_i$, which implies $\displaystyle\sum_{i=1}^{n} X_i$ is sufficient for λ. We can also have

$$\begin{aligned} f(\mathbf{x},\theta) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda}\lambda^{n\overline{x}} \cdot \frac{1}{\displaystyle\prod_{i=1}^{n} x_i!} \\ &= k_1(\overline{x},\lambda)\cdot k_2(\mathbf{x}). \end{aligned}$$

Hence, $\overline{X}$ is sufficient for λ. We further have

$$
\begin{aligned}
f(\mathbf{x}, \theta) &= \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \\
&= e^{-n\lambda} \lambda^{n(\overline{x}^k)^{\frac{1}{k}}} \cdot \frac{1}{\prod_{i=1}^{n} x_i!} \\
&= k_1(\overline{x}^k, \lambda) \cdot k_2(\mathbf{x}),
\end{aligned}
$$

which implies $\overline{X}^k$ is sufficient for λ for any positive integer k. ■

Example 2.4.21.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Find a sufficient statistic for (μ, σ^2).

Solution. The joint probability distribution function is

$$
\begin{aligned}
f(\mathbf{x}, \mu, \sigma^2) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right) \cdot 1 \\
&= k_1(\mathbf{x}, \mu, \sigma^2) \cdot k_2(\mathbf{x}).
\end{aligned}
$$

Hence, as previously derived, $(X_1, X_2, \ldots, X_n)$ is sufficient for (μ, σ^2). Moreover, we have

$$
\begin{aligned}
f(\mathbf{x}, \mu, \sigma^2) &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\overline{x}+\overline{x}-\mu)^2}{2\sigma^2} \right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\overline{x})^2 + 2(x_i-\overline{x})(\overline{x}-\mu) + (\overline{x}-\mu)^2}{2\sigma^2} \right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\overline{x})^2 + (\overline{x}-\mu)^2}{2\sigma^2} \right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\frac{(n-1)s^2 + n(\overline{x}-\mu)^2}{2\sigma^2} \right) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left(-\frac{(n-1)s^2 + n(\overline{x}-\mu)^2}{2\sigma^2} \right) \cdot 1 \\
&= k_1(\overline{x}, s^2, \mu, \sigma^2) \cdot k_2(\mathbf{x}).
\end{aligned}
$$

Hence, $(\overline{X}, S^2)$ is sufficient for (μ, σ^2). ■

2.5 UMVUE: Continuing to Generalization

Recall 2.5.1 (Conditional Probability Distribution Function).

Let X, Y be two random variables with joint probability distribution function $f(x, y)$. The conditional probability distribution functions are

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

and

$$f(x \mid y) = \frac{f(x, y)}{f_Y(y)}.$$

Definition 2.5.2 (Conditional Expectation).

The conditional expectation of Y given $X = x$ is

$$\mathrm{E}(Y \mid x) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} y\, f(y \mid x)\,\mathrm{d}y, & \text{if } Y \text{ is continuous;} \\ \displaystyle\sum_{-\infty}^{\infty} y\, f(y \mid x), & \text{if } Y \text{ is discrete.} \end{cases}$$

Definition 2.5.3 (Conditional Expectation).

The conditional expectation $\mathrm{E}(Y \mid X)$ of Y given X is $\mathrm{E}(Y \mid x)$ with x replaced by X.

Definition 2.5.4 (Conditional Variance).

The conditional variance of Y given $X = x$ is

$$\mathrm{Var}(Y \mid x) = \mathrm{E}\left((Y - \mathrm{E}(Y \mid x))^2 \mid x\right) = \mathrm{E}\left(Y^2 \mid x\right) - (\mathrm{E}(Y \mid x))^2.$$

Definition 2.5.5.

The conditional variance $\mathrm{Var}(Y \mid X)$ of Y given X is $\mathrm{Var}(Y \mid x)$ with x replaced by X.

Remark 2.5.6.

Both the conditional expectation $\mathrm{E}(Y \mid X)$ and the conditional variance $\mathrm{Var}(Y \mid X)$ are random variables and are functions of a random variable X.

Theorem 2.5.7.

Let X, Y be two random variables. Then,

(a) $\mathrm{E}\left(\mathrm{E}\left(Y \mid X\right)\right) = \mathrm{E}\left(Y\right)$;

(b) $\mathrm{Var}\left(Y\right) = \mathrm{E}\left(\mathrm{Var}\left(Y \mid X\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(Y \mid X\right)\right)$

Proof.

(a)

$$
\begin{aligned}
\mathrm{E}\left(\mathrm{E}\left(Y \mid X\right)\right) &= \int_{\infty}^{\infty} \mathrm{E}\left(Y \mid x\right) \cdot f_X(x)\,\mathrm{d}x \\
&= \int_{\infty}^{\infty} \int_{\infty}^{\infty} y\, f\left(y \mid x\right)\,\mathrm{d}y \cdot f_X(x)\,\mathrm{d}x \\
&= \int_{\infty}^{\infty} \int_{\infty}^{\infty} y\, f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{\infty}^{\infty} y \int_{\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{\infty}^{\infty} y\, f_Y(y)\,\mathrm{d}y \\
&= \mathrm{E}\left(Y\right).
\end{aligned}
$$

(b) By the definition,

$$
\mathrm{Var}\left(Y \mid X\right) = \mathrm{E}\left(Y^2 \mid X\right) - \left(\mathrm{E}\left(Y \mid X\right)\right)^2.
$$

Taking expectation on both sides yields

$$
\begin{aligned}
\mathrm{E}\left(\mathrm{Var}\left(Y \mid X\right)\right) &= \mathrm{E}\left(\mathrm{E}\left(Y^2 \mid X\right)\right) - \mathrm{E}\left(\left(\mathrm{E}\left(Y \mid X\right)\right)^2\right) \\
&= \mathrm{E}\left(Y^2\right) - \mathrm{E}\left(\left(\mathrm{E}\left(Y \mid X\right)\right)^2\right). \qquad (2.5.7.1)
\end{aligned}
$$

Also,

$$
\begin{aligned}
\mathrm{Var}\left(\mathrm{E}\left(Y \mid X\right)\right) &= \mathrm{E}\left(\left(\mathrm{E}\left(Y \mid X\right)\right)^2\right) - \left(\mathrm{E}\left(\mathrm{E}\left(Y \mid X\right)\right)\right)^2 \\
&= \mathrm{E}\left(\left(\mathrm{E}\left(Y \mid X\right)\right)^2\right) - \left(\mathrm{E}(Y)\right)^2. \qquad (2.5.7.2)
\end{aligned}
$$

Combining equation 2.5.7.1 and equation 2.5.7.2 yields

$$\begin{aligned}\mathrm{E}\left(\mathrm{Var}\left(Y \mid X\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(Y \mid X\right)\right) &= \mathrm{E}\left(Y^2\right) - \left(\mathrm{E}(Y)\right)^2 \\ &= \mathrm{Var}(Y).\end{aligned}$$

□

Lemma 2.5.8.

Let $\hat{\tau}(\mathbf{X})$ be an unbiased estimator of $\tau(\theta)$ and let $U = u(\mathbf{X})$ be a statistic. Then,

(a) $\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right)$ is unbiased for $\tau(\theta)$;

(b) $\mathrm{Var}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right) \leq \mathrm{Var}_\theta\left(\hat{\tau}(\mathbf{X})\right)$.

Proof.

(a)

$$\begin{aligned}\mathrm{E}\left(\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right)\right) &= \mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X})\right) \\ &= \tau(\theta).\end{aligned}$$

(b)

$$\begin{aligned}\mathrm{Var}_\theta\left(\hat{\tau}(\mathbf{X})\right) &= \mathrm{E}\left(\mathrm{Var}\left(\hat{\tau}(\mathbf{X}) \mid U\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(\hat{\tau}(\mathbf{X}) \mid U\right)\right) \\ &\geq \mathrm{Var}\left(\mathrm{E}\left(\hat{\tau}(\mathbf{X}) \mid U\right)\right)\end{aligned}$$

□

Remark 2.5.9.

We have some conclusions:

(a) If $\hat{\tau}(\mathbf{X})$ is unbiased for $\tau(\theta)$ and U is a statistic, then $\mathrm{E}\left(\hat{\tau}(\mathbf{X}) \mid U\right)$ is unbiased for $\tau(\theta)$ with variance smaller or equal to $\hat{\tau}(\mathbf{X})$.

(b) The random variable $\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right)$ may not be a statistic; so, it may not be an estimator.

(c) If U is a sufficient statistic, then $f(\mathbf{x} \mid u)$ is independent of θ. Then,

$$\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid u\right) = \int \cdots \int_{\mathbb{R}^n} \hat{\tau}(\mathbf{x})\, f(\mathbf{x} \mid u)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

is independent of θ, and $\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right)$ is an estimator.

Theorem 2.5.10 (Rao-Blackwell Theorem).

Let $\hat{\tau}(\mathbf{X})$ be an unbiased estimator of $\tau(\theta)$ and let U be a sufficient statistic. Then,

(a) $\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right)$ is a statistic;

(b) $\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right)$ is an unbiased estimator of $\tau(\theta)$;

(c) $\mathrm{Var}_\theta\left(\mathrm{E}_\theta\left(\hat{\tau}(\mathbf{X}) \mid U\right)\right) \leq \mathrm{Var}\left(\hat{\tau}(\mathbf{X})\right)$ for all $\theta \in \Theta$.

Corollary 2.5.11.

If $\hat{\tau} = \hat{\tau}(\mathbf{X})$ is an unbiased estimator of $\tau(\theta)$ and $U_1, U_2, U_3, \ldots$ are sufficient statistics, then

$$\begin{aligned}
\mathrm{Var}_\theta\left(\hat{\tau}\right) &\geq \mathrm{Var}_\theta\left(\mathrm{E}\left(\hat{\tau} \mid U_1\right)\right) \\
&\geq \mathrm{Var}_\theta\left(\mathrm{E}\left(\mathrm{E}\left(\hat{\tau} \mid U_1\right) \mid U_2\right)\right) \\
&\geq \mathrm{Var}_\theta\left(\mathrm{E}\left(\mathrm{E}\left(\mathrm{E}\left(\hat{\tau} \mid U_1\right) \mid U_2\right) \mid U_3\right)\right) \\
&\ \ \vdots
\end{aligned}$$

Proof. This follows from the Rao-Blackwell Theorem. □

Question 2.5.12.

In Corollary 2.5.11, will this process achieve the Cramér–Rao bound?

Answer. With one special condition, the answer is affirmative, and it just need one step to do so. □

Remark 2.5.13.

Let U be a statistic and let h be a function of U.

(a) If $h(U) = 0$, then $\mathrm{E}_\theta\left(h(U)\right) = \mathrm{E}_\theta(0) = 0$ for all $\theta \in \Theta$.

(b) If $P_\theta\left(h(U) = 0\right) = 1$ for all $\theta \in \Theta$, then the random variable $H = h(U)$ has a probability mass function

$$f_H(h) = \begin{cases} 1, & \text{if } h = 0; \\ 0, & \text{if } h \neq 0. \end{cases}$$

Then, $\mathrm{E}_\theta\left(h(U)\right) = 0 \cdot 1 + \sum_{h \neq 0} h \cdot 0 = 0$.

Remark 2.5.14.

In the previous remark, if $\mathrm{E}_\theta\left(h(U)\right) = 0$, then $P_\theta(h(U) = 0)$ may be 1 or not.

Definition 2.5.15 (Complete Statistic).

Let $X_1 n X_2, \ldots, X_n$ be a random sample from $f(x, \theta)$. A statistic $U = u(\mathbf{X})$ is called a <u>complete statistic</u> if $P_\theta\left(h(U) = 0\right) = 1$ for any function $h(U)$ such that $\mathrm{E}_\theta(h(U)) = 0$.

Question 2.5.16.

How can we verify whether a statistic U is complete or not?

Answer.

(a) To prove completeness, one needs to show that for any function $h(U)$ with $\mathrm{E}_\theta(h(U))$ for all $\theta \in \Theta$, the following is true:

$$P_\theta(h(U) = 0) = 1, \quad \text{for all } \theta \in \Theta.$$

This is hard.

(b) To prove incompleteness, one only needs to find a function $h^*(U)$ with $\mathrm{E}_\theta(h^*(U))$ for all $\theta \in \Theta$ and $P_{\theta_0}(h^*(U) = 0) = 1$ for some $\theta_0 \in \Theta$. □

Proposition 2.5.17.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Bernoulli(p). Then, $Y = \sum_{i=1}^{n} X_i$ is a complete statistic.

Proof. Notice that $Y \sim B(n, p)$. Let $h(\cdot)$ be a function such that $\mathrm{E}_p(h(Y)) = 0$ for all $p \in (0, 1) = \Theta$.

Now,

$$\begin{aligned}
0 &= \mathrm{E}_p(h(Y)) \\
&= \sum_{y=0}^{n} h(y) \binom{n}{y} p^y (1-p)^{n-y} \\
&= (1-p)^n \sum_{y=0}^{n} h(y) \binom{n}{y} \left(\frac{p}{1-p}\right)^y \\
\iff 0 &= \sum_{y=0}^{n} h(y) \binom{n}{y} \left(\frac{p}{1-p}\right)^y
\end{aligned}$$

for all $p \in (0,1)$. Let $\theta = \dfrac{p}{1-p}$. Then, $p \in (0,1) \iff \theta \in (0,\infty)$. Thus,

$$0 = \sum_{y=0}^{n} h(y) \binom{n}{y} \theta^y$$

for all $\theta \in (0,\infty)$. An order $n+1$ polynomial equation cannot have infinite solution except that all coefficients are zeros. As a consequence, $h(y) = 0$ for all $y = 0,1,2,\ldots,n$, and

$$P_p(Y = 0,1,2,\ldots,n) = 1. \tag{2.5.17.1}$$

By the axiom of probability, we have

$$1 \geq P_p(h(Y) = 0). \tag{2.5.17.2}$$

Since $Y = 0,1,2,\ldots,n$ is an event that is concluded in $h(Y) = 0$, we have

$$P_p(h(Y) = 0) \geq P_p(Y = 0,1,2,\ldots,n). \tag{2.5.17.3}$$

Combining equation 2.5.17.2, equation 2.5.17.3, and equation 2.5.17.1, we have

$$P_p(h(Y) = 0) = 1$$

for all $p \in (0,1)$. Therefore, $Y = \sum_{i=1}^{n} X_i$ is a complete statistic. □

Proposition 2.5.18.

Let X_1, X_2 be a random sample from Bernoulli(p). Then, $Z = X_1 - X_2$ is not a complete statistic.

Proof. The probability

$$\begin{aligned}
P_p(Z = 0) &= P_p(X_1 - X_2 = 0) \\
&= P_p(X_1 = X_2 = 0 \ \vee \ X_1 = X_2 = 1) \\
&= P_p(X_1 = X_2 = 0) + P_p(X_1 = X_2 = 1) \\
&= (1-p)^2 + p^2 \\
&< 1
\end{aligned}$$

for all $p \in (0,1)$. Therefore, $Z = X_1 - X_2$ is not a complete statistic. □

Proposition 2.5.19.

Let $X_1, X_2, \dots, X_n$ be a random sample from $U(0, \theta)$. Then, the largest order statistic Y_n is a complete statistic.

Proof. By Proposition 2.4.18, the probability density function of Y_n is $f_{Y_n}(y) = \dfrac{ny^{n-1}}{\theta^n}$. Suppose that $h(Y_n)$ satisfies $\mathrm{E}_\theta(h(Y_n)) = 0$ for $\theta \in (0, \infty)$. Then,

$$\begin{aligned}
0 &= \mathrm{E}_\theta(h(Y_n)) \\
&= \int_0^\theta h(y)\, \frac{ny^{n-1}}{\theta^n}\, \mathrm{d}y \\
&= \frac{n}{\theta^n} \int_0^\theta h(y)\, y^{n-1}\, \mathrm{d}y \\
\iff\ 0 &= \int_0^\theta h(y)\, y^{n-1}\, \mathrm{d}y
\end{aligned}$$

for all $\theta \in (0, \infty)$. Taking partial derivative with respect to θ on both sides yields

$$0 = h(\theta)\, \theta^{n-1}$$

for all $\theta \in (0, \infty)$, which implies $h(\theta) = 0$ for all $\theta \in (0, \infty)$. Hence, $h(y) = 0$ for all $y \in (0, \theta)$ for any $\theta > 0$, which further implies

$$P_\theta(h(Y_n) = 0) = P_\theta(0 < Y_n < \theta) = 1$$

for all $\theta > 0$. This holds for arbitrary function $h(\cdot)$ with $\mathrm{E}_\theta(h(Y_n)) = 0$. Therefore $Y_n = \max\{X_i\}_{i=1}^n$ is a complete statistic. $\square$

Definition 2.5.20 (Exponential Family).

If the probability distribution function of a random variable has the form

$$f(x, \theta) = e^{f_1(x)\, f_2(\theta) + f_3(x) + f_4(\theta)}$$

for all $x \in (a, b)$, where a and b are constants independent of θ, then we say that the probability distribution function f belongs to an <u>exponential family</u>.

Theorem 2.5.21.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $f(x, \theta)$, where f belongs to the exponential family as

$$f(x, \theta) = e^{f_1(x)\, f_2(\theta) + f_3(x) + f_4(\theta)}.$$

Then, $\sum_{i=1}^{n} f_1(X_i)$ is a complete and sufficient statistic.

Remark 2.5.22.

We say two random variables are equal $X = Y$ if $P(X = Y) = 1$.

Theorem 2.5.23 (Lehmann–Scheffé Theorem).

Let $X_1, X_2, \ldots, X_n$ be a random sample from $f(x, \theta)$. Suppose that $U = u(\mathbf{X})$ is a complete and sufficient statistic. If $\hat{\tau} = t(U)$ is unbiased for $\tau(\theta)$, then $\hat{\tau}$ is the unique function of U unbiased for $\tau(\theta)$ and is the uniform minimum variance unbiased estimator of $\tau(\theta)$.

Proof. We first show that $\hat{\tau}$ is unique. Suppose there exists another $\hat{\tau}^* = t^*(Y)$ that is also unbiased for $\tau(\theta)$. Then,

$$\begin{aligned}
\mathrm{E}_\theta\left(\hat{\tau} - \hat{\tau}^*\right) &= \mathrm{E}_\theta\left(\hat{\tau}\right) - \mathrm{E}_\theta\left(\hat{\tau}^*\right) \\
&= \tau(\theta) - \tau(\theta) \\
&= 0
\end{aligned}$$

for all $\theta \in \Theta$. By completeness, we have

$$\begin{aligned}
1 &= P_\theta\left(\hat{\tau} - \hat{\tau}^* = 0\right) \\
&= P_\theta\left(\hat{\tau}^* = \hat{\tau}\right)
\end{aligned}$$

for all $\theta \in \Theta$, which implies $\hat{\tau}^* = \hat{\tau}$. Hence, $\hat{\tau} = t(U)$ is the unique function of U unbiased for $\tau(\theta)$. We now show that $\hat{\tau}$ is the uniform minimum variance unbiased estimator of $\tau(\theta)$. Suppose T is an unbiased estimator of $\tau(\theta)$. Then, the Rao-Blackwell theorem gives

(a) $\mathrm{E}(T \mid U)$ is an unbiased estimator of $\tau(\theta)$. By uniqueness, $\hat{\tau} = \mathrm{E}(T \mid U)$.

(b) $\mathrm{Var}_\theta(\hat{\tau}) = \mathrm{Var}_\theta(\mathrm{E}(T \mid U)) \le \mathrm{Var}_\theta(T)$ for all $\theta \in \Theta$.

Since (b) holds for arbitrary unbiased estimator T, $\hat{\tau}$ is the uniform minimum variance unbiased estimator of $\tau(\theta)$. □

Remark 2.5.24.

We have two ways to construct the uniform minimum variance unbiased estimator with a complete and sufficient statistic U:

(a) If T is unbiased for $\tau(\theta)$, then $\mathrm{E}(T \mid U)$ is the uniform minimum variance unbiased estimator of $\tau(\theta)$. This may be difficult to find an explicit form of $E(T \mid U)$.

(b) If there exists a constant c such that $\mathrm{E}(U) = c\,\tau(\theta)$, then $\dfrac{U}{c}$ is the uniform minimum variance unbiased estimator of $\tau(\theta)$.

Example 2.5.25.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $U(0,\theta)$. Find the uniform minimum variance unbiased estimator of θ.

Solution. By Proposition 2.5.19 and Proposition 2.4.18, the largest order statistic $Y_n = \max\{X_i\}_{i=1}^n$ is a complete and sufficient statistic. The probability density function of Y_n is

$$\begin{aligned} f_{Y_n}(y) &= n\left(F_{Y_n}(y)\right)^{n-1} f_X(y) \\ &= \frac{ny^{n-1}}{\theta^n}, \end{aligned}$$

for all $y \in (0,\theta)$. The expectation of Y_n is

$$\begin{aligned} \mathrm{E}_\theta(Y_n) &= \int_0^\theta y \frac{ny^{n-1}}{\theta^n}\,\mathrm{d}y \\ &= \left[\frac{n}{n+1}\frac{y^{n+1}}{\theta^n}\right]_0^\theta \\ &= \frac{n}{n+1}\theta. \end{aligned}$$

Hence, $\dfrac{n+1}{n}Y_n$ is unbiased for θ, and therefore $\dfrac{n+1}{n}Y_n$ is the uniform minimum variance unbiased estimator of θ. ■

Example 2.5.26.

Let $X_1, X_2, \dots, X_n$ be a random sample from Bernoulli(p). Find the uniform minimum variance unbiased estimator of p.

Solution. The probability mass function of X is

$$\begin{aligned} f(x,p) &= x^p(1-x)^{1-p} \\ &= (1-p)\left(\frac{p}{1-p}\right)^x \\ &= \exp\left(x\ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right) \\ &= \exp\left(x\ln\left(\frac{p}{1-p}\right) + \ln(1-p) + 0\right) \\ &= \exp\left(f_1(x)\, f_2(p) + f_3(p) + f_4(x)\right). \end{aligned}$$

Hence, $Y = \sum_{i=1}^{n} X_i$ is a complete statistic.[1] By Proposition 2.4.17, Y is a sufficient statistic. The expectation

$$\mathrm{E}\left(Y\right) = np.$$

Hence, $\hat{p} = \dfrac{Y}{n} = \overline{X}$ is unbiased for p, and therefore $\overline{X}$ is the uniform minimum variance unbiased estimator of p. ■

Example 2.5.27.

Let $X_1, X_2, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, 1)$. Find the uniform minimum variance unbiased estimator of μ.

Solution. The probability density function of X is

$$\begin{aligned} f(x,\mu) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2-2x\mu+\mu^2}{2}} \\ &= \exp\left(x\mu - \frac{x^2}{2} - \frac{\mu^2}{2} - \ln\left(\sqrt{2\pi}\right)\right) \\ &= \exp\left(f_1(x)\, f_2(\mu) + f_3(\mu) + f_4(x)\right). \end{aligned}$$

[1] This is also proved in Proposition 2.5.17

Hence, $Y = \sum_{i=1}^{n} X_i$ is a complete statistic. By Example 2.4.21, Y is sufficient. The expectation

$$\mathrm{E}(Y) = n\mu.$$

Hence, $\hat{\mu} = \frac{Y}{n} = \overline{X}$ is unbiased for μ, and therefore $\overline{X}$ is the uniform minimum variance unbiased estimator of μ. ■

Example 2.5.28.

Let $X_1, X_2, \ldots, X_n$ be a random sample from Poisson(λ). Find the uniform minimum variance unbiased estimator of λ.

Solution. The probability mass function of X is

$$\begin{aligned} f(x, \lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \exp\left(x \ln \lambda - \lambda + \ln(x!)\right) \\ &= \exp\left(f_1(x)\, f_2(\lambda) + f_3(\lambda) + f_4(x)\right). \end{aligned}$$

Hence, $Y = \sum_{i=1}^{n} X_i$ is a complete statistic. By Example 2.4.20, Y is sufficient. The expectation

$$\mathrm{E}(Y) = n\lambda.$$

Hence, $\hat{\lambda} = \frac{Y}{n} = \overline{X}$ is unbiased for λ, and therefore $\overline{X}$ is the uniform minimum variance unbiased estimator of λ. ■

Alphabetical Index