STAT 330 Mathematical Statistics

University of Waterloo - Winter 2025

26th April 2025

Instructor: Yeying Zhu

LaTeX : Xing Liu

Contents

1 Probability

Lecture 1, 2025/01/06

Definition 1.1 (Sample Space). A set of all possible outcomes from a random experiment, S, is called the **sample space**.

Example. Rolling a die twice: $S = \{(1,1),(1,2),\dots,(6,6)\}$. Each of the (i,j) is called an elementarty event.

Remark. Types of sample spaces:

(1) Finite: $S = \{\omega_1.\dots,\omega_n\}$.

(2) Countable: $S = \{\omega_1,\omega_2,\dots\}$.

Example. Rolling a die until a 6 is obtained.

(3) Uncountable:

Example. Lifetime of a light bulb, $S = \{x : x \geq 0\} = [0,\infty)$.

Definition 1.2 (Event). An **event** is a subset of the sample space, $A \subseteq S$.

Remark.

(1) We say that an event A occurs if the outcome ω of a random experiment is in A.

(2) One goal of probability theory is to study how likely (probability) an event occurs.

Example. Rolling a die twice. We can define $A = \{(x,y) : x \leq y\}$. Then the # of elementary events in $A = 21$.

Definition 1.3 (Probability Set Function).
Let $\mathcal{B} = \{A_1, A_2, \dots\}$ be a suitable class of subsets of S. We call $\mathcal{B}$ a **σ-algebra**. A **probability set function** (p.s.f) is a function $\mathbb{P} : \mathcal{B} \to [0,1]$ such that:

(1) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{B}$.

(2) $\mathbb{P}(S) = 1$.

(3) If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise mutually exclusive, i.e. $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Definition 1.4 (**σ-algebra**). (**not tested**)

Given sample space S, a **σ-algebra** $\mathcal{B}$ on S is a collection of subsets of S that satisfies:

(1) $S \in \mathcal{B}$.

(2) If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$.

(3) If $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

Example. Let $\mathcal{B} = \{\varnothing, S\}$. Then $\mathcal{B}$ is a σ-algebra.

Proposition 1.1 (**Properties of p.s.f**).

If $\mathbb{P}$ is a p.s.f and A, B are any set in $\mathcal{B}$, then:

(1) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(2) $\mathbb{P}(\varnothing) = 0$.

(3) $\mathbb{P}(A) \leq 1$.

(4) $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$.

(5) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

(6) If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof.

(1) $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(S) = 1$.

(2) $\mathbb{P}(\varnothing) = \mathbb{P}(S^c) = 1 - P(S) = 0$.

(3) $\mathbb{P}(S) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) \geq \mathbb{P}(A)$, since $\mathbb{P}(A^c) \geq 0$.

(4) We have the following two equations:

$$\mathbb{P}\left((A \cap B^c) \cup (A \cap B)\right) = \mathbb{P}\left(A \cap (B^c \cup B)\right) = \mathbb{P}(A \cap S) = \mathbb{P}(A)$$
$$\mathbb{P}\left((A \cap B^c) \cup (A \cap B)\right) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B).$$

Combining to get $\mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) = \mathbb{P}(A)$.

(5) Exercise.

(6) Let $B^* = B \setminus A = B \cap A^c$. Then, $\mathbb{P}(B) = \mathbb{P}(A \cup B^*) = \mathbb{P}(A) + \mathbb{P}(B^*) \geq \mathbb{P}(A)$.

□

Lecture 2, 2025/01/08

Definition 1.5 (Conditional Probability).

Suppose A and B are subesets of S. The **conditional probability** of event A given event B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \mathbb{P}(B) > 0.$$

Example. Rolling a die twice. Let $S = \{(1,1), \dots, (6,6)\}$ with 36 elements. Then $\mathbb{P}(i,j) = \frac{1}{36}$ for all $(i,j) \in S$. This is a probability set function.

Let A be the event that the sum is ≤ 4, and B be the event that the sum ≥ 10. Then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) = \frac{1}{3}$, since A and B are mutually exclusive.

Let $C = \{i = j\}$. Suppose event C occured, what is $\mathbb{P}(A \cup B)$?

Proof. Now, our sample space is $S^* = \{(1,1), (2,2), \dots, (6,6)\}$ with 6 elements. Then

$$\mathbb{P}(A \cup B \mid C) = \frac{4}{6} = \frac{2}{3}.$$ □

Remark. $\mathbb{P}(\cdot \mid B)$ is a probability set function!

Proof. We will show the three conditions of a p.s.f:

(1) For any $A \in S$, $\mathbb{P}(A \mid B) \geq 0$.

(2) $\mathbb{P}(\underbrace{B}_{S^*} \mid B) = 1$.

(3) If $A_1, A_2, \ldots$ are pairwise mutually exclusive, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \,\middle|\, B\right) = \frac{\mathbb{P}\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid B).$$

□

Proposition 1.2 (**Law of Total Probability**).

Suppose $B_1, B_2, \ldots, B_n$ is a collection of mutually exclusive and exhaustive events (i.e. $\bigcup_{i=1}^{n} B_i = S$). Then,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A \mid B_i).$$

Proof. Since events $A \cap B_1, A \cap B_2, \ldots, A \cap B_n$ mutually exclusive. Then,

$$\begin{aligned} \mathbb{P}(A) = \mathbb{P}(A \cap S) &= \mathbb{P}\left(A \cap \left(\bigcup_{i=1}^{n} B_i\right)\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) \\ &= \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) \\ &= \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A \mid B_i). \end{aligned}$$

□

Example. An insurance company divides people into two groups, accident-prone and those who are not. For those accident-prone people, the chance of having an accident in a year is 0.4. For those who are not, the chance is 0.2. Suppose 30% of customers are accident-prone, what is the probability that a new policy holder will have a claim within a year?

Proof. Define $A = \{\text{have a claim}\}$, $B_1 = \{\text{accident-prone}\}$, and $B_2 = \{\text{not accident-prone}\}$. Note that $B_1 \cap B_2 = \varnothing$ and $B_1 \cup B_2 = S$, so B_1 and B_2 is a partition of S. Then,

$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1) \cdot \mathbb{P}(B_1) + \mathbb{P}(A \mid B_2) \cdot \mathbb{P}(B_2) = 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26.$$

□

Definition 1.6 (**Independent Events**).

Suppose A and B are events defined on S. A and B are **independent events** if

$$\mathbb{P}(A \mid B) = P(A) \quad \text{OR} \quad \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Definition 1.7 (**Mutually Independent Events**).

Suppose $A_1, \dots, A_n$ are events defined on S. We say that $A_1, \dots, A_n$ are **mutually independent** if for any $i_1, \dots, i_k$ from $\{1, \dots, n\}$, we have

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Example. Tossing a fair coin twice. Let $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ and that $\mathbb{P}_i = \frac{1}{4}$ for $i = 1, 2, 3, 4$. Note that this is a p.s.f. Let $A_1 = \{\text{H on 1st toss}\}$, $A_2 = \{\text{H on 2nd toss}\}$, and $B = \{\text{exactly 1 H and 1 T}\}$.

$$\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(B) = \frac{1}{2}.$$

Next,

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap B) = \mathbb{P}(A_2 \cap B) = \frac{1}{4}.$$

However, $\mathbb{P}(A_1 \cap A_2 \cap B) = 0 \neq \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(B)$, so A_1, A_2, B are pairwise independent but not mutually independent.

Theorem 1.3 (**Bayes' Theorem**).

Suppose $B_1, \dots, B_n$ is a partition of S, then for any event A, we have

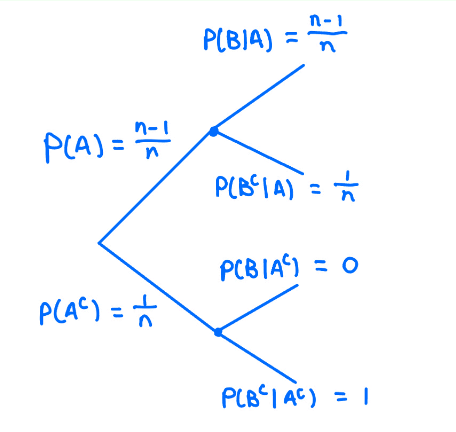
$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A \mid B_i)}{\sum\limits_{j=1}^{n} \mathbb{P}(B_j)\mathbb{P}(A \mid B_j)}.$$

Proof. We have $\mathbb{P}(B_i \mid A) = \dfrac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \dfrac{\mathbb{P}(B_i)\mathbb{P}(A \mid B_i)}{\sum\limits_{j=1}^{n} \mathbb{P}(B_j)\mathbb{P}(A \mid B_j)}$. □

Example. John and Michelle communicate through emails. They agree that they will reply on the same day once they receive an email from each other. Due to a bad server, out of n emails, there will be one that cannot reach the destination on the same day. Now, John sends an email to Michelle, but did not receive a reply on the same day. What is the probability that Michelle receives John's email?

Proof. Use a tree diagram. Let $A = \{\text{Michelle receives email}\}$ and $B = \{\text{John receives response}\}$. Then,

$$\mathbb{P}(A \mid B^c) = \frac{\mathbb{P}(A)\mathbb{P}(B^c \mid A)}{\mathbb{P}(A)\mathbb{P}(B^c \mid A) + \mathbb{P}(A^c)\mathbb{P}(B^c \mid A^c)} = \frac{\frac{n-1}{n} \cdot \frac{1}{n}}{\frac{n-1}{n} \cdot \frac{1}{n} + \frac{1}{n} \cdot 1} = \frac{n-1}{2n-1}. \qquad \square$$



2 Random Variables

2.1 Random Variables

Definition 2.1 (**Random Variable**).

A **random variable** is a function from a sample space S to $\mathbb{R}$, i.e. $X : S \to \mathbb{R}$, such that $\mathbb{P}(X \le x)$ is defined for all $x \in \mathbb{R}$.

Remark. $X \le x$ is an abbreviation for the event $\{\omega \in S : X(\omega) \le x\}$.

Example. Tossing a fair coin three times. Let X = number of heads. Then,

$$X : S \to \mathbb{R}$$

where $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Consider $A = \{X \le 1\} = \{\omega \in S : X(\omega) \le 1\}$. Then,

$$\begin{aligned}\mathbb{P}_X(A) &= \mathbb{P}_S(\{\omega \in S : X(\omega) \le 1\}) \\ &= \mathbb{P}_S(\{HTT, TTH, THT.TTT\}) \\ &= \frac{4}{8} = \frac{1}{2}.\end{aligned}$$

Remark. The sample space S and the probability set function induce the probability on the random variable X. From above example,

$$\begin{aligned}\mathbb{P}(X = 0) = \mathbb{P}(\{\omega \in S : X(\omega) = 0\}) = \frac{1}{8} \\ \mathbb{P}(X = 1) = \mathbb{P}(\{\omega \in S : X(\omega) = 1\}) = \frac{3}{8} \\ \mathbb{P}(X = 2) = \mathbb{P}(\{\omega \in S : X(\omega) = 2\}) = \frac{3}{8} \\ \mathbb{P}(X = 3) = \mathbb{P}(\{\omega \in S : X(\omega) = 3\}) = \frac{1}{8}.\end{aligned}$$

The probability mass function:

x	0	1	2	3
$\mathbb{P}(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

.

Definition 2.2 (**Cumulative Distribution Function (CDF)**).

The **cumulative distribution function** (CDF) of a random variable X is defined as

$$F(x) = \mathbb{P}(X \leq x) \quad \text{for all } x \in \mathbb{R}$$

and $F(x) : \mathbb{R} \to [0, 1]$.

Proposition 2.1 (**Properties of CDF**).

(1) $F(x)$ is non-decreasing, i.e., $F(x_1) \leq F(x_2), \forall x_1 < x_2$.

(2) $\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1$.

(3) $F(x)$ is a right-continuous function, i.e., $\lim_{x \to a^+} F(x) = F(a)$.

Proof.

(1) Let $x_1 < x_2$. Then, $\{\omega : X(\omega) \leq x_1\} \subseteq \{\omega : X(\omega) \leq x_2\}$. Thus, by the property of probability set function,

$$\mathbb{P}(\{\omega : X(\omega) \leq x_1\}) \leq \mathbb{P}(\{\omega : X(\omega) \leq x_2\}) \iff \mathbb{P}(X \leq x_1) \leq \mathbb{P}(X \leq x_2) \iff F(x_1) \leq F(x_2).$$

(2) Not a rigorous proof: As $x \to -\infty$, $\{\omega : X(\omega) \leq x\} \to \varnothing \implies \mathbb{P}(X \leq x) \to 0$. As $x \to \infty$, $\{\omega : X(\omega) \leq x\} \to S \implies \mathbb{P}(X \leq x) \to 1$.

(3) See Prof's notes.

□

Remark.

(1) If we define $F(x) = \mathbb{P}(X < x)$, then F is left-continuous.

(2) Any funciton $F(x)$ satisfying (1), (2), (3) is a valid CDF of some random variable.

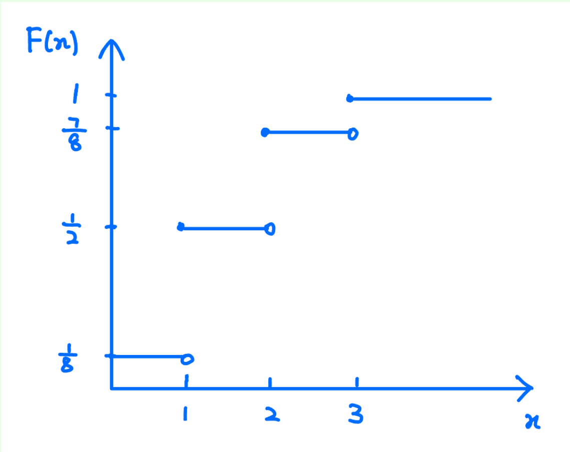
Definition 2.3 (**Discrete Random Variable**).

If S is discrete (i.e. finite or countable), then X is a **discrete random variable**.

Remark. $F(x)$ is a right-continuous step function.

Example. Tossing a fair coin three times with X = number of heads. Then

$$F(x) = \mathbb{P}(X \le x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8} & 0 \le x < 1 \\ \frac{1}{2} & 1 \le x < 2 \\ \frac{7}{8} & 2 \le x < 3 \\ 1 & x \ge 3 \end{cases}$$



Definition 2.4 (**Probability Mass Function (PMF)**).

If X is a discrete random variable, then the **probability mass function** (PMF) of X is given by

$$f(x) = \mathbb{P}(X = x) = F(x) - \lim_{\epsilon \to 0^+} F(x - \epsilon) = F(x) - \lim_{a \to x^-} F(a).$$

Remark. If we can write all possible values of X in an increasing order, i.e. $x_1 < x_2 < \cdots$, then

$$f(x_1) = F(x_1) \quad \text{and for any } i > 1,\ f(x_i) = F(x_i) - F(x_{i-1}).$$

Remark. The set $A = \{x : f(x) > 0\}$ is called the **support** of X.

Example. In the above example, $A = \{0, 1, 2, 3\}$.

Proposition 2.2 (Properties of PMF).

Let f be a PMF of a discrete random variable $X \iff$ the following hold.

(1) $f(x) \geq 0$ for all $x \in \mathbb{R}$.

(2) $\sum_{x \in A} f(x) = 1$.

Proof. Show: If f is a PMF, then $\sum_{x \in A} f(x) = 1$.

$$\sum_{x \in A} f(x) = \sum_{x_i \in A} \mathbb{P}(X = x_i) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} \{X = x_i\}\right) = \mathbb{P}(S) = 1.$$

□

Example. Show that $f(x) = \dfrac{\mu^x e^{-\mu}}{x!}$, $x = 0, 1, \ldots$ and $\mu > 0$ is a PMF.

Proof.

(1) $\dfrac{\mu^x e^{-\mu}}{x!} \geq 0$ for all $x = 0, 1, \ldots$.

(2) $\displaystyle\sum_{x=0}^{\infty} \frac{\mu^x e^{-\mu}}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{\mu^x}{x!} = e^{-\mu} e^{\mu} = 1.$ □

Definition 2.5 (Continuous Random Variable).

Suppose X is a random variable with CDF F. If F is a continuous function for all $x \in \mathbb{R}$ and F is differentiable except possibly at countably many points, then X is a **continuous random variable**.

Lecture 4, 2025/01/15

Example. Recall the example of tossing a coin three times with X = number of heads. We have

$$\mathbb{P}(X = 1) = \mathbb{P}(X \leq 1) - \mathbb{P}(X < 1) = F(1) - \lim_{a \to 1^-} F(a) = \frac{1}{2} - \frac{1}{8} = \frac{3}{8}.$$

Definition 2.6 (**Probability Density Function (PDF)**).

If X is a continuous random variable with CDF $F(x)$, then the **probability density function** (**PDF**) of X is defined as

$$f(x) = F'(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x)$$

if F is differentiable at x, and otherwise, we define $f(x) = 0$.

Example (**Uniform Distribution**).

Let $F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \ge b \end{cases}$. Then, $f(x) = F'(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$.

In this case, the support of X is $A = (a, b)$.

Note. F is not differentiable at $x = a$ and $x = b$.

Proposition 2.3 (**Properties of f(x)**).

$f(x)$ is the PDF for some continuous random variable $X \iff$ (1) and (2) hold.

(1) $f(x) \ge 0$ for all $x \in \mathbb{R}$.

(2) $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = 1$.

(3) $f(x) = \displaystyle\lim_{h\to 0} \frac{F(x+h) - F(x)}{h}$ if the limit exists.

(4) $F(x) = \displaystyle\int_{-\infty}^{x} f(t)\,dt,\ x \in \mathbb{R}$.

(5) $\mathbb{P}(a < X \le b) = \mathbb{P}(X \le b) - \mathbb{P}(X \le a) = \displaystyle\int_{-\infty}^{b} f(t)\,dt - \int_{-\infty}^{a} f(t)\,dt = \int_{a}^{b} f(t)\,dt$.

(6) $\mathbb{P}(X = b) = F(b) - \displaystyle\lim_{a\to b^-} F(a) = F(b) - F(b) = 0$ since the CDF is continuous everywhere.

(7) $\mathbb{P}(a < X \le b) = \mathbb{P}(a \le X < b) = \mathbb{P}(a < X < b) = \mathbb{P}(a \le X \le b)$. That is, $\mathbb{P}(X \in A) = \displaystyle\int_A f(x)\,dx$.

Example. Consider the function $f(x) = \frac{\theta}{x^{\theta+1}}$ for $x \ge 1$. For what values of θ is $f(x)$ a PDF?

Proof. We check the first two conditions.

(1) $f(x) \geq 0$ for all $x \geq 1$.

(2) $\displaystyle\int_1^\infty \frac{\theta}{x^{\theta+1}}\,dx = -x^{-\theta}\Big|_1^\infty = -\lim_{b\to\infty}\frac{1}{b^\theta} - (-1) = 1 \implies \theta > 0.$

$\square$

2.2 Functions of Random Variables

Distribution of Functions of a Random Variable

- CDF technique.

Suppose X is a continuous random variable with PDF $f(x)$ and CDF $F(x)$ and we wish to find the PDF of $Y = h(X)$ where h is a real-valued function.

Example. If $Z \sim \mathrm{N}(0,1)$, find the PDF of $Y = Z^2$.

Proof. Let

$$\begin{aligned}
G(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Z^2 \leq y) \\
&= \mathbb{P}(-\sqrt{y} \leq Z \leq \sqrt{y}) \\
&= F(\sqrt{y}) - F(-\sqrt{y}) \\
&= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz \\
&= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz.
\end{aligned}$$

Then,

$$\begin{aligned}
g(y) = G'(y) &= \frac{2}{\sqrt{2\pi}} \frac{\mathrm{d}}{\mathrm{d}y}\left(\int_0^{\sqrt{y}} e^{-\frac{z^2}{2}}\,dz\right) \\
&= \frac{2}{\sqrt{2\pi}} e^{-\frac{(\sqrt{y})^2}{2}} \frac{1}{2\sqrt{y}} \\
&= \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}, \quad y > 0.
\end{aligned}$$

$\square$

Theorem 2.4 (**One-to-One Transformation of a Discrete Random Variable**).
Suppose X is a discrete random variable with PMF $f(x)$ and $Y = h(X)$ is a one-to-one transformation of X ($X = h^{-1}(Y)$). Then, the PMF of Y is given by

$$g(y) = f(h^{-1}(y)) \quad \text{for } y \in B$$

where $B = \{y : g(y) > 0\}$.

Proof. Note that $g(y) = \mathbb{P}(Y = y) = \mathbb{P}(h(X) = y) = \mathbb{P}(X = h^{-1}(y)) = f(h^{-1}(y))$ for $y \in B$. □

Example. Let $X \sim \mathrm{NB}(r, p)$ be the number of trials required to obtain r successes in repeated independent Bernoulli trials. Then

$$f(x) = \mathbb{P}(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad x = r, r+1, \dots .$$

Let $Y = h(X) = X - r$ be the number of failures before the rth success. Then

$$g(y) = \mathbb{P}(Y = y) = \mathbb{P}(X = y + r) = \binom{y+r-1}{r-1} p^r (1-p)^y, \quad y = 0, 1, \dots .$$

Lecture 5, 2025/01/20

Theorem 2.5 (**One-to-One Transformation of a Continuous Random Variable**).
Suppose X is a continuous random variable with PDF $f(x)$ and support $A = \{x : f(x) > 0\}$ and $Y = h(X)$, where h is one-to-one. Let g be the PDF of Y, then

$$g(y) = f(h^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} h^{-1}(y) \right| \quad \text{for } y \in B$$

where $B = \{y : g(y) > 0\}$ is the support of Y.

Proof. Since $h(x)$ is one-to-one, it is either monotonically increasing or monotonically decreasing.

(1) If h is increasing for $x \in A$, then h^{-1} is also increasing for $y \in B$ and $\frac{\mathrm{d}}{\mathrm{d}y} h^{-1}(y) > 0$. Then,

$$G(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X \le h^{-1}(y)) = F(h^{-1}(y)).$$

Then,

$$g(y) = G'(y) = f(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y) = f(h^{-1}(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y)\right| \quad \text{for } y \in B.$$

(2) If h is decreasing for $x \in A$, then h^{-1} is also decreasing for $y \in B$ and $\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y) < 0$. Then,

$$G(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \geq h^{-1}(y)) = 1 - F(h^{-1}(y)).$$

And,

$$g(y) = G'(y) = -f(h^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y) = f(h^{-1}(y))\left|\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y)\right| \quad \text{for } y \in B.$$

□

Example. Find the PDF of $Y = \ln(X)$ where X is a continuous random variable with $f(x) = \frac{\theta}{x^{\theta+1}}$ for $x \geq 1$ and $\theta > 0$.

Proof. Let $h(X) = \ln(X)$ and $h^{-1}(Y) = e^Y$ for $y \geq 0$. Then,

$$g(y) = f(e^y)\left|\frac{\mathrm{d}}{\mathrm{d}y}e^y\right| = \frac{\theta}{(e^y)^{\theta+1}}e^y = \theta e^{-(\theta+1)y}e^y = \frac{\theta}{e^{y\theta}} \quad \text{for } y \geq 0.$$

□

Theorem 2.6 (**Probability Integral Transformation**).

If X is a continuous random variable with CDF F and F is strictly increasing, then $Y = F(X) \sim \mathrm{Unif}(0, 1)$. $Y = F(X)$ is called the **probability integral transformation**.

Proof. Note that

$$\mathbb{P}(Y \leq y) = \mathbb{P}(F(X) \leq y) = \mathbb{P}(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y \quad \text{for } 0 \leq y \leq 1.$$

Therefore, $Y \sim \mathrm{Unif}(0, 1)$. □

Remark. Suppose we generate $U_1, \dots, U_n$ independently from $\mathrm{Unif}(0, 1)$ by a computer. Then, $F^{-1}(U_1), \dots, F^{-1}(U_n)$ are independent observations from F.

Example. If we want to generate random variables from $\text{Exp}(\lambda)$, where λ is the rate parameter, and $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. So, $F^{-1}(u) = -\frac{\ln(1-u)}{\lambda}$ for $u \in [0, 1]$. The steps are:

1. Generate $U_1, \dots, U_n$ independently and identically from $\text{Unif}(0, 1)$.
2. Compute $X_i = -\frac{\ln(1-U_i)}{\lambda}$ for $i = 1, \dots, n$, where $X_i \sim \text{Exp}(\lambda)$ for $i = 1, \dots, n$ independently and identically.

3 Expectation and Moment Generating Functions

3.1 Expectation

Definition 3.1 (**Expectation**). If X is a discrete random variable with PMF $f(x)$ and support A, then the **expectation** of X is

$$\mathbb{E}[X] = \sum_{x \in A} x f(x)$$

provided that the sum converges absolutely, i.e. $\mathbb{E}|X| = \sum_{x \in A} |x| f(x) < \infty$.

If X is a continuous random variable with PDF $f(x)$, then the **expectation** of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

provided that the integral converges absolutely, i.e. $\mathbb{E}|X| = \displaystyle\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$.

Note. Intuitively, if the sum or integral does not converge absolutely, then the extreme values of x will dominate the expectation and there will be no meaningful interpretation of the central tendency.

Example. Let $f(x) = \frac{\theta}{x^{\theta+1}}$ for $x \geq 1$ and $\theta > 0$. Find $\mathbb{E}[X]$. For what values of θ does the expectation exist?

Proof. Note that

$$\mathbb{E}|X| = \mathbb{E}[X] = \int_1^{\infty} \frac{\theta x}{x^{\theta+1}}\, dx = \theta \int_1^{\infty} \frac{1}{x^{\theta}}\, dx = \theta \left[\frac{x^{1-\theta}}{1-\theta} \right]_1^{\infty} = \frac{\theta}{1-\theta} \lim_{b \to \infty} (b^{1-\theta} - 1).$$

We want this to be $< \infty$, so we need $1 - \theta < 0 \implies \theta > 1$. Hence, the expectation exists if $\theta > 1$ and $\mathbb{E}[X] = -\frac{\theta}{1-\theta}$. □

Example (**Standard Cauchy Distribution does not have an expectation**).
Let X follow a standard Cauchy distribution, i.e. $f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$. Find $\mathbb{E}[X]$.

Proof. Note that

$$\mathbb{E}|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi} \frac{1}{1+x^2}\,dx = 2\int_0^{\infty} \frac{1}{\pi} \frac{x}{1+x^2}\,dx = \frac{2}{\pi} \frac{\ln(1+x^2)}{2}\bigg|_0^{\infty} = \infty.$$

Therefore, $\mathbb{E}[X]$ does not exist! □

Example. Suppose X is a non-negative continuous random variable with CDF $F(x)$ and $\mathbb{E}[X] < \infty$. Show that $\mathbb{E}[X] = \int_0^{\infty} [1 - F(x)]\,dx$.

Proof. Note that

$$\begin{aligned}
\int_0^{\infty} [1 - F(x)]\,dx &= \int_0^{\infty} \mathbb{P}(X > x)\,dx \\
&= \int_0^{\infty} \int_x^{\infty} f_X(y)\,dy\,dx \\
&= \int_0^{\infty} \int_0^{y} f_X(y)\,dx\,dy \\
&= \int_0^{\infty} (y - 0) f_X(y)\,dy \\
&= \int_0^{\infty} y f_X(y)\,dy = \mathbb{E}[X].
\end{aligned}$$

□

Lecture 6, 2025/01/22

Definition 3.2 (Expectation of a Function of X).

If X is discrete with PMF $f(x)$ and support A, then,

$$\mathbb{E}[h(X)] = \sum_{x \in A} h(x) f(x)$$

provided that the sum converges absolutely, i.e. $\mathbb{E}|h(X)| = \sum_{x \in A} |h(x)| f(x) < \infty$.

If X is continuous with PDF $f(x)$, then

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\,dx$$

provided that the integral converges absolutely, i.e. $\mathbb{E}|h(X)| = \int_{-\infty}^{\infty} |h(x)|f(x)\,dx < \infty$.

Proposition 3.1 (**Linearity of Expectation**).
$\mathbb{E}[ag(X) + bh(X)] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)]$.

Proof. Assume that X is continuous for illustration. Then

$$\begin{aligned}\mathbb{E}[ag(X) + bh(X)] &= \int_{-\infty}^{\infty} (ag(x) + bh(x))f(x)\,dx \\ &= a\int_{-\infty}^{\infty} g(x)f(x)\,dx + b\int_{-\infty}^{\infty} h(x)f(x)\,dx \\ &= a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)].\end{aligned}$$

$\square$

Remark. $\mathbb{E}\left[\frac{g(X)}{h(X)}\right] \neq \frac{\mathbb{E}[g(X)]}{\mathbb{E}[h(X)]}$.

Proposition 3.2 (**Special Expectations**).

(1) Variance: $\mathrm{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[X^2\right] - \mu^2$, where $\mu = \mathbb{E}[X]$.

(2) k^{th} moment (about the origin): $\mathbb{E}\left[X^k\right]$.

(3) k^{th} moment (about the mean): $\mathbb{E}\left[(X-\mu)^k\right]$.

Note. $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$ and $\mathbb{E}\left[X^2\right] = \mathrm{Var}(X) + \mu^2$.

Theorem 3.3 (**Markov's Inequality**).
Suppose X is a random variable. Then

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}|X|^k}{c^k} \quad \text{for all } k, c > 0.$$

Proof. We have

$$\begin{aligned}
\frac{\mathbb{E}|X|^k}{c^k} &= \int_{-\infty}^{\infty} \left|\frac{x}{c}\right|^k f(x)\,dx \\
&= \int_{\left|\frac{x}{c}\right| \geq 1} \left|\frac{x}{c}\right|^k f(x)\,dx + \int_{\left|\frac{x}{c}\right| < 1} \left|\frac{x}{c}\right|^k f(x)\,dx \\
&\geq \int_{\left|\frac{x}{c}\right| \geq 1} \left|\frac{x}{c}\right|^k f(x)\,dx \\
&\geq \int_{\left|\frac{x}{c}\right| \geq 1} f(x)\,dx \\
&= \mathbb{P}(|X| \geq c) \quad \text{for } c > 0.
\end{aligned}$$

□

Theorem 3.4 (**Chebyshev's Inequality**).

Suppose X is a random variable with finite mean μ and finite variance σ^2. Then for any $k > 0$,

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. By Markov's Inequality, we have

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{\mathbb{E}|X - \mu|^2}{(k\sigma)^2} = \frac{\mathrm{Var}\,(X)}{k^2\sigma^2} = \frac{1}{k^2}.$$

□

Note. In particular,

$$\begin{aligned}
\mathbb{P}(|X - \mu| \leq 2\sigma) &\geq 1 - \frac{1}{2^2} = \frac{3}{4}. \\
\mathbb{P}(|X - \mu| \leq 3\sigma) &\geq 1 - \frac{1}{3^2} = \frac{8}{9}.
\end{aligned}$$

Example. From below table of PMF,

x	-1	0	1
$f(x)$	$\frac{1}{8}$	$\frac{6}{8}$	$\frac{1}{8}$

we have $\mu = \mathbb{E}[X] = 0$ and $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mu^2 = \frac{1}{4}$. By Chebyshev's Inequality,

$$\mathbb{P}(|X| \geq 1) = \mathbb{P}(|X - 0| \geq 2 \cdot \frac{1}{2}) \leq \frac{1}{2^2} = \frac{1}{4}.$$

Proposition 3.5 (**Degenerate Distribution**).
If $\mu = \mathbb{E}[X]$ and $\mathrm{Var}(X) = 0$, then we have $\mathbb{P}(X = \mu) = 1$ and X is said to have a **degenerate distribution**.

Proof. Look at the event $\{X \neq \mu\} = \bigcup_{i=1}^{\infty}\{|X - \mu| \geq \frac{1}{i}\}$. Then,

$$\begin{aligned} \mathbb{P}(X \neq \mu) = \mathbb{P}\left(\bigcup_{i=1}^{\infty}\left\{|X - \mu| \geq \frac{1}{i}\right\}\right) &\leq \sum_{i=1}^{\infty} \mathbb{P}\left(|X - \mu| \geq \frac{1}{i}\right) \quad \text{by Boole's Inequality.} \\ &\leq \sum_{i=1}^{\infty} \frac{\mathbb{E}|X - \mu|^2}{\left(\frac{1}{i}\right)^2} \quad \text{by Markov's Inequality} \\ &= \sum_{i=1}^{\infty} i^2 \,\mathrm{Var}(X) = 0. \end{aligned}$$

□

Proposition 3.6 (**Variance Stabilizing Transformation**).
Suppose X is a random variable with $\mathbb{E}[X] = \theta$, $\mathrm{Var}(X) = \sigma^2(\theta)$ (a function of $\mathbb{E}[X]$). We aim to find $Y = g(X)$ such that $\mathrm{Var}(Y)$ is a constant.

Let $Y = g(X)$, where g is differentiable. By the linear approximation,

$$Y = g(X) \approx g(\theta) + g'(\theta)(X - \theta).$$

Therefore, $\mathbb{E}[Y] \approx g(\theta) + g'(\theta)(\mathbb{E}[X] - \theta) = g(\theta)$ and $\mathrm{Var}(Y) \approx g'(\theta)^2 \,\mathrm{Var}(X) = (g'(\theta)\sigma(\theta))^2$. If we want $\mathrm{Var}(Y)$ to be a constant, then we need $g'(\theta)\sigma(\theta) = k$ for some constant k. That is,

$$g'(\theta) = \frac{k}{\sigma(\theta)} \quad \text{this is how we pick g.}$$

Example. If $X \sim \text{Exp}(\theta)$, where $\theta = \frac{1}{\lambda}$ is a scale parameter, then show that $Y = g(X) = \ln(X)$ has approximately constant variance.

Proof. Note that $f(x) = \frac{1}{\theta}e^{-x/\theta}$ for $x, \theta > 0$. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x\frac{1}{\theta}e^{-x/\theta}\,dx \quad \text{let } y = \frac{x}{\theta} \\
&= \int_0^\infty y\theta\frac{1}{\theta}e^{-y}\theta\,dy \\
&= \theta\int_0^\infty ye^{-y}\,dy \\
&= \theta\Gamma(2) = \theta.
\end{aligned}$$

Recall that $\Gamma(\alpha) = \displaystyle\int_0^\infty x^{\alpha-1}e^{-x}\,dx = (\alpha-1)!$, for $\alpha = 1, 2, \ldots$. Also,

(1) $\Gamma(1) = 1$.
(2) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
(3) $\Gamma(m) = (m-1)\Gamma(m-1)$ for $m > 1$.

Then,

$$\mathbb{E}\left[X^2\right] = \int_0^\infty x^2\frac{1}{\theta}e^{-x/\theta}\,dx = \int_0^\infty y^2\theta^2\frac{1}{\theta}e^{-y}\theta\,dy = \theta^2\Gamma(3) = \theta^2\cdot 2! = 2\theta^2.$$

Therefore, $\text{Var}(X) = \mathbb{E}\left[X^2\right] - \mu^2 = 2\theta^2 - \theta^2 = \theta^2$. Hence, $\sigma(\theta) = \theta$ and $g'(\theta) = \frac{1}{\theta}$. Therefore, $\text{Var}(Y) \approx g'(\theta)^2\,\text{Var}(X) = \frac{1}{\theta^2}\cdot\theta^2 = 1$. $\square$

Lecture 7, 2025/01/27

Example. If $X \sim \text{Poisson}(\theta)$, then show that $Y = g(X) = \sqrt{X}$ has approximately constatnt variance.

Proof. Note that $\mathbb{E}[X] = \theta = \text{Var}(X)$. Also, $g'(x) = \frac{1}{2\sqrt{x}} \implies g'(\theta) = \frac{1}{2\sqrt{\theta}}$. Therefore, $\text{Var}(Y) \approx g'(\theta)^2\,\text{Var}(X) = \frac{1}{4\theta}\cdot\theta = \frac{1}{4}$. $\square$

3.2 Moment Generating Functions

Definition 3.3 (Moment Generating Function).
If X is a random variable, then $M(t) = \mathbb{E}\left[e^{tX}\right]$ is the **moment generating function** (MGF) of X if the expectation exists for all $t \in (-h, h)$ for some $h > 0$.

Remark. The value of t for which the expectation exists should always be stated.

Example. If $X \sim \text{Gamma}(\alpha, \beta)$, then find $M(t)$.

Proof. Note that $f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}}$ for $x, \alpha, \beta > 0$ where α is the shape parameter and β is the rate parameter.

Now,

$$
\begin{aligned}
M(t) = \mathbb{E}\left[e^{tX}\right] &= \int_0^\infty e^{tx} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}}\, dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\left(\frac{1}{\beta}-t\right)x}\, dx \quad \text{let } y = \left(\frac{1}{\beta} - t\right)x \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \left(\frac{y}{\frac{1}{\beta} - t}\right)^{\alpha-1} e^{-y} \frac{1}{\frac{1}{\beta} - t}\, dy \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\frac{1}{\frac{1}{\beta} - t}\right)^{\alpha} \int_0^\infty y^{\alpha-1} e^{-y}\, dy \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\frac{1}{\frac{1}{\beta} - t}\right)^{\alpha} \Gamma(\alpha) \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \cdot \frac{\beta^\alpha}{(1-\beta t)^\alpha} \cdot \Gamma(\alpha) \\
&= \frac{1}{(1-\beta t)^\alpha}, \quad \frac{1}{\beta} - t > 0 \implies t < \frac{1}{\beta}.
\end{aligned}
$$

□

Example. Find $M_Z(t)$ for $Z \sim \text{N}(0, 1)$.

Proof. We have

$$\begin{aligned}
M_Z(t) = \mathbb{E}\left[e^{tZ}\right] &= \int_{-\infty}^{\infty} e^{tz}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\,dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}+tz}\,dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-t)^2}e^{\frac{t^2}{2}}\,dz \\
&= e^{\frac{t^2}{2}}\underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-t)^2}\,dz}_{\text{PDF of N}(t,1)} \\
&= e^{\frac{t^2}{2}}\cdot 1 = e^{\frac{t^2}{2}}, \quad t \in \mathbb{R}.
\end{aligned}$$

□

Example. Let $X \sim \mathrm{NB}(r,p)$ be the number of failures before obtaining r successes. Note that $\mathbb{P}(X=x) = \binom{x+r-1}{r-1}p^r(1-p)^x$ for $x = 0,1,2,\ldots$, $0<p<1$. Find $M(t)$.

Proof. X is a discrete random variable. Then

$$\begin{aligned}
M(t) = \mathbb{E}\left[e^{tX}\right] &= \sum_{x=0}^{\infty} e^{tx}\binom{x+r-1}{r-1}p^r(1-p)^x \\
&= \sum_{x=0}^{\infty}\binom{x+r-1}{r-1}p^r(e^t(1-p))^x
\end{aligned}$$

Note. Power series fact: $\sum_{i=0}^{\infty}\binom{i+r-1}{r-1}p^i = (1-p)^{-r}$, $0<p<1$.

$$\begin{aligned}
&= p^r(1-e^t(1-p))^{-r} \\
&= \left(\frac{p}{1-e^t(1-p)}\right)^r
\end{aligned}$$

where $0 < e^t(1-p) < 1 \implies t < -\ln(1-p)$. □

Theorem 3.7. Suppose X has MGF $M_X(t)$ for $t \in (-h, h)$. Let $Y = aX + b$ where $a \neq 0$. Then, the MGF of Y is

$$M_Y(t) = e^{bt} M_X(at), \quad |t| < \frac{h}{|a|}.$$

Proof. We have $M_Y(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = e^{bt}\mathbb{E}\left[e^{atX}\right] = e^{bt}M_X(at)$, for $|at| < h \implies |t| < \frac{h}{|a|}$. □

Example. Let $Z \sim \mathrm{N}(0,1)$ and $X \sim \mathrm{N}(\mu, \sigma^2)$. Find $M_X(t)$.

Proof. $X = \sigma Z + \mu$. Then, $M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, for $t \in \mathbb{R}$. □

Lecture 8, 2025/01/29

Note. Cauchy distribution does not have an MGF.

Theorem 3.8. Suppose X has an MGF $M(t)$ defined for $t \in (-h, h)$. Then, $M(0) = 1$ and

$$M^{(k)}(0) = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M(t)\bigg|_{t=0} = \mathbb{E}\left[X^k\right]$$

for $k = 1, 2, \ldots$.

Note. We use this theorem to compute moments of X.

Proof. We will prove the case when X is a continuous random variable, the discrete case is similar.

$$\begin{aligned}
M(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\,dx \\
\implies M^{(k)}(t) &= \frac{\mathrm{d}^k}{\mathrm{d}t^k} \int_{-\infty}^{\infty} e^{tx} f(x)\,dx \\
&= \int_{-\infty}^{\infty} \frac{\mathrm{d}^k}{\mathrm{d}t^k} e^{tx} f(x)\,dx \\
&= \int_{-\infty}^{\infty} x^k e^{tx} f(x)\,dx \\
\implies M^{(k)}(0) &= \int_{-\infty}^{\infty} x^k f(x)\,dx = \mathbb{E}\left[X^k\right].
\end{aligned}$$

□

Example. If $X \sim \text{Gamma}(\alpha, \beta)$, then $M(t) = (1-\beta t)^{-\alpha}$ for $t < \frac{1}{\beta}$. Find $\mathbb{E}\left[X^k\right]$.

Proof. Note that $M(0) = 1$ and

$$\begin{aligned}
\mathbb{E}[X] &= M'(0) = -\alpha(1-\beta t)^{-\alpha-1}(-\beta)\Big|_{t=0} = \alpha\beta(1-\beta t)^{-\alpha-1}\Big|_{t=0} = \alpha\beta \\
\mathbb{E}\left[X^2\right] &= M''(0) = \alpha\beta(-\alpha-1)(1-\beta t)^{-\alpha-2}(-\beta)\Big|_{t=0} = (\alpha+1)\alpha\beta^2 = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}\beta^2 \\
&\vdots \\
\mathbb{E}\left[X^k\right] &= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\beta^k.
\end{aligned}$$

□

Proposition 3.9 (Maclaurin Series).
If we can obtain a Taylor series expansion for $M(t)$ of X, then

$$M(t) = \sum_{k=0}^{\infty} \frac{M^{(k)}(0)}{k!} t^k = \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[X^k\right]}{k!} t^k.$$

The coefficient of t^k is $\frac{\mathbb{E}\left[X^k\right]}{k!}$. Then, we can obtain $\mathbb{E}\left[X^k\right]$ by

$$\mathbb{E}\left[X^k\right] = k! \times \text{coefficient of } t^k \text{ in the Maclaurin series for } M(t).$$

Example. Recall that the MGF of $X \sim \text{Gamma}(\alpha, \beta)$ is $(1-\beta t)^{-\alpha}$ for $t < \frac{1}{\beta}$. Then

$$\begin{aligned}
M(t) &= (1-\beta t)^{-\alpha} \\
&= \sum_{k=0}^{\infty} \binom{k+\alpha-1}{\alpha-1} (\beta t)^k \quad \text{power series fact} \\
&= \sum_{k=0}^{\infty} \binom{k+\alpha-1}{\alpha-1} \beta^k t^k. \\
\implies \mathbb{E}\left[X^k\right] &= k! \binom{k+\alpha-1}{\alpha-1} \beta^k \\
&= k! \frac{(k+\alpha-1)!}{k!(\alpha-1)!} \beta^k \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \beta^k.
\end{aligned}$$

Example. Let $M(t) = \frac{1+t}{1-t}$ for $|t| < 1$. Find $\mathbb{E}\left[X^k\right]$.

Proof. We have

$$\begin{aligned} M(t) &= (1+t)\frac{1}{1-t} \\ &= (1+t)\sum_{k=0}^{\infty} t^k \\ &= \sum_{k=0}^{\infty} t^k + \sum_{k=0}^{\infty} t^{k+1} \\ &= 1 + \sum_{k=1}^{\infty} 2t^k \\ \implies \mathbb{E}\left[X^k\right] &= 2k! \quad \text{for } k = 1, 2, \ldots \end{aligned}$$

And, $\mathbb{E}\left[X^0\right] = 1$. □

Remark. We talked about three approaches to find $\mathbb{E}\left[X^k\right]$: definition, $M^{(k)}(0)$, and Maclaurin series.

Theorem 3.10 (Uniqueness Theorem for MGFs). Suppose the random variable X has MGF $M_X(t)$ and Y has MGF $M_Y(t)$. Suppose also that $M_X(t) = M_Y(t)$ for all $t \in (-h, h)$ for some $h > 0$. Then, X and Y have the same distribution. That is,

$$\mathbb{P}(X \leq s) = F_X(s) = F_Y(s) = \mathbb{P}(Y \leq s) \quad \text{for all } s \in \mathbb{R}.$$

Remark. The MGF uniquely determines the distribution of a random variable.

Example. If $X \sim \text{Exp}(1)$, then find the distribution of $Y = \beta x$ where $\beta > 0$.

Proof. Note that $X \sim \text{Gamma}(1,1)$. Then $M_X(t) = (1-t)^{-1}$ for $t < 1$. Also,

$$M_Y(t) = M_X(\beta t) = (1-\beta t)^{-1} \quad \beta t < 1 \implies t < \frac{1}{\beta}$$

which is the MGF of $\text{Gamma}(1,\beta)$ or $\text{Exp}(\beta)$. By the Uniqueness Theorem, we have $Y \sim \text{Gamma}(1,\beta)$ or $Y \sim \text{Exp}(\beta)$. □

Example (**Example Proof of the Unique Theorem**).

A naive example. Suppose X has the following PMF.

X	0	1
$f(x)$	$\frac{4}{10}$	$\frac{6}{10}$

Then, $M_X(t) = \mathbb{E}\left[e^{tX}\right] = \frac{4}{10} + \frac{6}{10}e^t$ for $t \in \mathbb{R}$.
Now suppose we know $M_X(t) = \frac{4}{10} + \frac{6}{10}e^t$ for $t \in \mathbb{R}$. We want to get $f(0)$ and $f(1)$. Then

$$M_X(t) = f(0) \times 1 + f(1) \times e^t = \frac{4}{10} + \frac{6}{10}e^t, \quad t \in \mathbb{R}.$$

Question: what are $f(0)$ and $f(1)$?

$$\left(\frac{6}{10} - f(1)\right)e^t - \left(\frac{4}{10} - f(0)\right) = 0, \quad t \in \mathbb{R}.$$

Then, $f(0) = \frac{4}{10}$ and $f(1) = \frac{6}{10}$.

Example (**MGF and Moments for Uniform Distribution**).

Let $X \sim \text{Unif}(a, b)$ with $f(x) = \frac{1}{b-a}$ for $x \in (a, b)$. Then,

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \int_a^b e^{tx}\frac{1}{b-a}\,dx \\ &= \begin{cases} \frac{1}{b-a}\left[\frac{e^{tx}}{t}\right]_a^b & \text{if } t \neq 0 \\ \frac{1}{b-a}\int_a^b 1\,dx & \text{if } t = 0 \end{cases} \\ &= \begin{cases} \frac{1}{b-a}\left[\frac{e^{tb}-e^{ta}}{t}\right] & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases} \end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}[X] = M_X'(0) &= \lim_{t\to 0} \frac{M_X(t) - M_X(0)}{t} \\
&= \lim_{t\to 0} \frac{\frac{1}{b-a}\left[\frac{e^{tb}-e^{ta}}{t}\right] - 1}{t} \\
&= \lim_{t\to 0} \frac{e^{bt} - e^{at} - (b-a)t}{t^2(b-a)} \\
&= \lim_{t\to 0} \frac{be^{bt} - ae^{at} - (b-a)}{2t(b-a)} \quad \text{L'Hopital's Rule} \\
&= \lim_{t\to 0} \frac{b^2e^{bt} - a^2e^{at}}{2(b-a)} \quad \text{L'Hopital's Rule} \\
&= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.
\end{aligned}$$

To calculate $\mathbb{E}\left[X^2\right]$, we have

$$\begin{aligned}
\mathbb{E}\left[X^2\right] = M''(0) &= \lim_{t\to 0} \frac{M'(t) - M'(0)}{t} \\
&\vdots \\
&= \frac{a^2 + ab + b^2}{3}.
\end{aligned}$$

Thus, $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{(b-a)^2}{12}$.

4 Joint Distributions

Lecture 9, 2025/02/03

4.1 Bivariate Joint & Marginal Distributions

Definition 4.1 (Random Vector). An n-dimentional **random vector** is a function from $S \to \mathbb{R}^n$, where $\mathbb{R}^n$ is the n-dimensional Euclidean space.

Example. Toss a fair die twice. Then $S = \{(1,1), (1,2), \ldots, (6,6)\}$ with 36 elementary events. Let X = sum of the two tosses and Y = absolute value of difference of the two tosses. Note that $\mathbb{P}(i, j) = \frac{1}{36}$ for $i = 1, \ldots, 6$ and $j = 1, \ldots, 6$. The joint distribution of X and Y is given by

$Y \backslash X$	2	3	4	$\cdots$	12
0	1/36	0	1/36		
1	0	1/18	0		
$\vdots$					
5					

Note that

$$\begin{aligned}\mathbb{P}(X = 2, Y = 0) &= \mathbb{P}(\{\omega \in S : X(\omega) = 2, Y(\omega) = 0\}) \\ &= \mathbb{P}(\{(1,1)\}) = \frac{1}{36}.\end{aligned}$$

And, $\mathbb{P}(X = 3, Y = 0) = \mathbb{P}(\varnothing) = 0$...

Definition 4.2 (Joint CDF).
Suppose (X, Y) is a random vector on a sample S. The **joint CDF** of (X, Y) is

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) \quad \text{for } (x, y) \in \mathbb{R}^2.$$

Proposition 4.1 (**Properties of CDF**).

(1) $F(x, y)$ is non-decreasing for both x,y, i.e. $\forall a < b, F(a, y) \leq F(b, y), F(x, a) \leq F(x, b)$.

(2) $\lim_{x\to-\infty} F(x, y) = 0,\ \lim_{y\to-\infty} F(x, y) = 0,\ \lim_{\substack{x\to-\infty\\ y\to-\infty}} F(x, y) = 0,\ \lim_{\substack{x\to\infty\\ y\to\infty}} F(x, y) = 1.$

(3) $F(x, y)$ is right-continuous for both x, y, that is,

$$\lim_{h\to 0^+} F(x + h, y) = F(x, y) \qquad \lim_{h\to 0^+} F(x, y + h) = F(x, y).$$

Definition 4.3 (**Marginal CDF**).

The **marginal CDF** of X (or Y) is

$$F_1(x) = \mathbb{P}(X \leq x) = \lim_{y\to\infty} F(x, y)$$

$$F_2(y) = \mathbb{P}(Y \leq y) = \lim_{x\to\infty} F(x, y).$$

Definition 4.4 (**Joint Discrete Random Variable: Joint PMF**).

If S is discrete, then X and Y are discrete random variables. The **joint PMF** of (X, Y) is

$$f(x, y) = \mathbb{P}(X = x, Y = y) \quad (x, y) \in \mathbb{R}^2.$$

$A = \{(x, y) : f(x, y) > 0\}$ is called the **support set** of (X, Y).

Example. In the previous example, we have $A = \{(2, 0), (4, 0), \dots , (12, 0), (3, 1), \dots , \}$.

Proposition 4.2 (**Properties of Joint PMF**).

$f(x, y)$ is the PMF of (X, Y) $\iff$ the following hold:

(1) $f(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$.

(2) $\sum_x \sum_y f(x, y) = 1$.

Also, for any set $D \subset \mathbb{R}^2$,

$$\mathbb{P}((X, Y) \in D) = \sum\sum_{(x,y)\in D} f(x, y).$$

Definition 4.5 (**Marginal PMF**).

The **marginal PMF** of X (or Y) is

$$f_1(x) = \mathbb{P}(X = x) = \sum_y f(x, y) \quad x \in \mathbb{R}$$
$$f_2(y) = \mathbb{P}(Y = y) = \sum_x f(x, y) \quad y \in \mathbb{R}.$$

Example. In a fourth year stat course, there are 10 acturial science students, 9 stat students and 6 math students. 5 students are selected at random without replacement. Define the following random variables:

$$X = \text{number of acturial science students selected}$$
$$Y = \text{number of stat students selected.}$$

(a) The joint PMF of (X, Y) is

$$f(x, y) = \mathbb{P}(X = x, Y = y) = \frac{\binom{10}{x}\binom{9}{y}\binom{6}{5-x-y}}{\binom{25}{5}}.$$

(b) The marginal PMF of X is

$$f_1(x) = \mathbb{P}(X = x) = \sum_{y=0}^{5-x} f(x, y) = \frac{\binom{10}{x}\sum_{y=0}^{5-x}\binom{9}{y}\binom{6}{5-x-y}}{\binom{25}{5}} = \frac{\binom{10}{x}\binom{15}{5-x}}{\binom{25}{5}} \quad x = 0, \ldots, 5.$$

Then, $X \sim \text{Hypergeometric}(10, 9, 5)$. Moreover, $0 \le x+y \le 5$ and $0 \le x \le 5, 0 \le y \le 5$.

Note. $\sum_{y=0}^{5-x}\binom{9}{y}\binom{6}{5-x-y} = \binom{15}{5-x}$ is by the hypergeometric identity.

(c) The marginal PMF of Y is

$$f_2(y) = \sum_{x=0}^{5-y} f(x, y) = \frac{\binom{9}{y}\sum_{x=0}^{5-y}\binom{10}{x}\binom{6}{5-x-y}}{\binom{25}{5}} = \frac{\binom{9}{y}\binom{16}{5-y}}{\binom{25}{5}} \quad y = 0, \ldots, 5.$$

Then, $Y \sim \text{Hypergeometric}(9, 6, 5)$.

(d) Note that

$$\begin{aligned}\mathbb{P}(X > Y) = \sum_{x>y} f(x,y) = \sum_{x=1}^{5}\sum_{y=0}^{x-1} f(x,y) \\ &= f(1,0) + f(2,0) + f(2,1) + f(3,0) + \\ &\quad f(3,1) + f(3,2) + f(4,0) + f(4,1) + f(5,0).\end{aligned}$$

Definition 4.6 (Joint Continuous Random Variables: Joint PDF).
Suppose $F(x,y)$ is continuous and $f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$ exists except possibly along a finite number of curves. Then, X and Y are continuous random variables with **joint PDF** $f(x,y)$. The set $A = \{(x,y) : f(x,y) > 0\}$ is called the **support set** of (X,Y).
We arbitrarily define $f(x,y) = 0$ when $\frac{\partial^2}{\partial x \partial y} F(x,y)$ does not exist.

Proposition 4.3 (Properties of Joint PDF).
$f(x,y)$ is the PDF of (X,Y) $\iff$ the following hold:

(1) $f(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$.

(2) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1$.

Also, for any set $D \subset \mathbb{R}^2$,

$$\mathbb{P}((X,Y) \in D) = \iint_D f(x,y)\,\mathrm{d}x\,\mathrm{d}y\,.$$

Definition 4.7 (Marginal PDF).
The **marginal PDF** of X (or Y) is

$$\begin{aligned} f_1(x) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \quad x \in \mathbb{R} \\ f_2(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \quad y \in \mathbb{R}. \end{aligned}$$

Example. Let $f(x,y) = x + y$, for $0 \leq x \leq 1$ and $0 \leq y \leq 1$.

(1) Show that $f(x,y)$ is a joint PDF.

Proof.

(1) $f(x,y) = x + y \geq 0$ for $0 \leq x \leq 1$ and $0 \leq y \leq 1$.

(2) Check the following.

$$\begin{aligned}\int_0^1 \int_0^1 f(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_0^1 \int_0^1 (x+y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^1 \left[\frac{x^2}{2} + xy\right]_{x=0}^{x=1} \mathrm{d}y \\ &= \int_0^1 \left(\frac{1}{2} + y\right) \mathrm{d}y \\ &= \left[\frac{y}{2} + \frac{y^2}{2}\right]_0^1 = 1.\end{aligned}$$

□

(2) Find $\mathbb{P}(X \leq 1/3, Y \leq 1/2)$.

Proof. We have

$$\begin{aligned}\int_0^{1/2} \int_0^{1/3} (x+y)\,\mathrm{d}x\,\mathrm{d}y &= \int_0^{1/2} \left[\frac{x^2}{2} + xy\right]_{x=0}^{x=1/3} \mathrm{d}y \\ &= \int_0^{1/2} \left(\frac{1}{18} + \frac{y}{3}\right) \mathrm{d}y \\ &= \left[\frac{y}{18} + \frac{y^2}{6}\right]_0^{1/2} = \frac{1}{36} + \frac{1}{24} = \frac{5}{72}.\end{aligned}$$

□

Lecture 10, 2025/02/05

Example (**Above Continued**).

(3) Find $\mathbb{P}(X \leq Y)$.

Proof. Note that $\mathbb{P}((X,Y) \in D) = \iint_D f(x,y)\,\mathrm{d}x\,\mathrm{d}y$. Thus,

$$\begin{aligned}
\mathbb{P}(X \le Y) &= \int_0^1 \int_0^y (x+y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^1 \left[\frac{x^2}{2} + xy\right]_{x=0}^{x=y} \mathrm{d}y \\
&= \int_0^1 \left(\frac{y^2}{2} + y^2\right) \mathrm{d}y \\
&= \left[\frac{y^3}{6} + \frac{y^3}{3}\right]_0^1 = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}.
\end{aligned}$$

□

(4) Find $\mathbb{P}(X + Y \le \frac{1}{2})$.

Proof. We have

$$\begin{aligned}
\mathbb{P}(X + Y \le 1/2) &= \int_0^{1/2} \int_0^{1/2-y} (x+y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^{1/2} \int_0^{1/2-x} (x+y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_0^{1/2} \left[xy + \frac{y^2}{2}\right]_{y=0}^{y=1/2-x} \mathrm{d}x \\
&= \int_0^{1/2} \left(\frac{1}{8} - \frac{x^2}{2}\right) \mathrm{d}x \\
&= \left[\frac{x}{8} - \frac{x^3}{6}\right]_0^{1/2} = \frac{1}{16} - \frac{1}{48} = \frac{1}{24}.
\end{aligned}$$

□

(5) Find $\mathbb{P}(XY \le \frac{1}{2})$.

Proof. We need to split the region into two parts.

$$\begin{aligned}
\mathbb{P}(XY \le 1/2) &= \int_0^1 \int_0^{1/2} (x+y)\,\mathrm{d}x\,\mathrm{d}y + \int_{1/2}^1 \int_0^{1/2x} (x+y)\,\mathrm{d}y\,\mathrm{d}x \\
&\vdots \\
&= \frac{3}{4}.
\end{aligned}$$

There is an easier way to do this.

$$
\begin{aligned}
\mathbb{P}(XY \leq 1/2) &= 1 - \mathbb{P}(XY > 1/2) \\
&= 1 - \int_{1/2}^{1} \int_{1/2y}^{1} (x+y) \, \mathrm{d}x \, \mathrm{d}y \\
&\vdots \\
&= \frac{3}{4}.
\end{aligned}
$$

□

(6) Find the marginal PDF of X and Y.

Proof. We have

$$
\begin{aligned}
f_1(x) &= \int_0^1 (x+y) \, \mathrm{d}y = x + \frac{1}{2} \quad 0 \leq x \leq 1 \\
f_2(y) &= \int_0^1 (x+y) \, \mathrm{d}x = y + \frac{1}{2} \quad 0 \leq y \leq 1.
\end{aligned}
$$

□

(7) Find the joint CDF of (X, Y).

Proof. For $x \in (0,1)$ and $y \in (0,1)$, we have

$$
\begin{aligned}
F(x,y) = \mathbb{P}(X \leq x, Y \leq y) &= \int_0^y \int_0^x (x+y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_0^y \left(\frac{x^2}{2} + xy \right)_{y=0}^{y=y} \mathrm{d}y \\
&= \frac{1}{2} x^2 y + \frac{1}{2} y^2.
\end{aligned}
$$

For $x \geq 1$ and $y \in (0,1)$, we have

$$
\begin{aligned}
F(x,y) &= \int_0^y \int_0^1 (x+y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_0^y [x + xy]_{x=0}^{x=1} \, \mathrm{d}y \\
&= \int_0^y (1+y) \, \mathrm{d}y = \frac{1}{2} y + \frac{1}{2} y^2.
\end{aligned}
$$

For $x \in (0,1)$ and $y \geq 1$, we have

$$\begin{aligned} F(x,y) &= \int_0^1 \int_0^x (x+y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{x}x + \frac{1}{2}x^2. \end{aligned}$$

For $x \leq 0$ <u>or</u> $y \leq 0$, we have $F(x,y) = 0$. For $x \geq 1$ and $y \geq 1$, we have $F(x,y) = 1$. □

(8) Find the marginal CDF of X (and Y).

Proof. We have

$$F_1(x) = \begin{cases} 0 & x \leq 0 \\ \int_0^1 \int_0^x (x+y)\,\mathrm{d}y\,\mathrm{d}x = \frac{1}{2}x^2 + \frac{1}{2}x & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}.$$

□

4.2 Independence

Theorem 4.4 (**Independent Random Variables**).

Suppose X and Y are random variables with joint CDF $F(x,y)$ and marginal CDFs $F_1(x)$ and $F_2(y)$, joint PDF or PMF $f(x,y)$ and marginal PDFs or PMFs $f_1(x)$ and $f_2(y)$. Let

$$\begin{aligned} A_1 &= \{x : f_1(x) > 0\} \\ A_2 &= \{y : f_2(y) > 0\} \\ A &= \{(x,y) : f(x,y) > 0\} \end{aligned}$$

be the support sets of X, Y and (X,Y) respectively. Then, X and Y are **independent** $\iff$ either of the following holds.

(1) $f(x,y) = f_1(x)f_2(y)$ for all $(x,y) \in A$, where $A = A_1 \times A_2 = \{(x,y) : x \in A_1, y \in A_2\}$. In other words, the support A is a rectangular region in $\mathbb{R}^2$.

(2) $F(x,y) = F_1(x)F_2(y)$ for all $x \in \mathbb{R}, y \in \mathbb{R}$.

Remark.

(1) The support set A is a Cartesion product (rectangular region).

> *Example.* Let $f(x,y) = 8xy$, with $0 < x < y < 1$. Are X and Y independent?
>
> Answer: No, because $0 < x < y < 1$ is a triangular region. This is not a Cartesian product.

(2) If X and Y are independent, then $h(X)$ and $g(Y)$ are also independent.

Theorem 4.5 (**Factorization Theorem for Independence**).
X and Y are independet random variables $\iff$ $A = A_1 \times A_2$ and $\exists$ non-negative functions $g(x)$ and $h(y)$ such that

$$f(x,y) = g(x)h(y) \quad \forall (x,y) \in A.$$

Proof.
($\Rightarrow$): If X and Y are independent, then $f(x,y) = f_1(x)f_2(y)$. Take $g(x) = f_1(x)$ and $h(y) = f_2(y)$.

($\Leftarrow$): Suppose $f(x,y) = g(x)h(y)$ for $(x.y) \in [a,b] \times [c,d]$ (WLOG). Then

$$f_1(x) = \int_c^d f(x,y)\,\mathrm{d}y = \int_c^d g(x)h(y)\,\mathrm{d}y = k_1 g(x) \quad \text{where } k_1 = \int_c^d h(y)\,\mathrm{d}y$$
$$f_2(y) = \int_a^b f(x,y)\,\mathrm{d}x = \int_a^b g(x)h(y)\,\mathrm{d}x = k_2 h(y) \quad \text{where } k_2 = \int_a^b g(x)\,\mathrm{d}x\,.$$

Also, we know that

$$\begin{aligned} 1 &= \int_c^d \int_a^b f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_c^d \int_a^b g(x)h(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_c^d h(y)\,\mathrm{d}y \int_a^b g(x)\,\mathrm{d}x = k_1 k_2. \end{aligned}$$

Thus, since $k_1 k_2 = 1$, we have

$$f(x,y) = g(x)h(y) = k_1 k_2 g(x)h(y) = f_1(x)f_2(y) \quad \text{for } (x,y) \in [a,b] \times [c,d].$$

This implies that $X \perp\!\!\!\perp Y$. $\square$

Example. Let $f(x,y) = \frac{\theta^{x+y}e^{-2\theta}}{x!y!}$ for $x, y = 0, 1, 2, \ldots$. Are X and Y independent?

Proof. Note that

$$f(x,y) = \underbrace{\frac{\theta^x e^{-\theta}}{x!}}_{g(x)} \cdot \underbrace{\frac{\theta^y e^{-\theta}}{y!}}_{h(y)} \quad \text{for } A = \{(x,y) : x, y = 0, 1, 2, \ldots\} = A_1 \times A_2.$$

Thus, X and Y are independent. $\square$

Lecture 11, 2025/02/10

4.3 Conditional Distributions

Definition 4.8 (Conditional PMF/PDF).

The **conditional PMF/PDF** of X given $Y = y$ is

$$f_1(x|y) = \frac{f(x,y)}{f_2(y)} \quad \text{for } (x,y) \in A \text{ provided } f_2(y) > 0.$$

Similarly, the conditional PMF/PDF of Y given $X = x$ is

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)} \quad \text{for } (x,y) \in A \text{ provided } f_1(x) > 0.$$

Remark.

(1) $f(x|y)$ is a valid PMF/PDF, i.e. $f(x|y) \geq 0$ and $\sum_x f(x|y) = 1$ or $\int_{-\infty}^{\infty} f(x|y)\,\mathrm{d}x = 1$. Note that

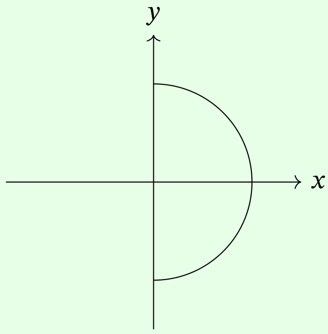
$$\sum_x f(x|y) = \sum_x \frac{f(x,y)}{f_2(y)} = \frac{\sum_x f(x,y)}{f_2(y)} = \frac{f_2(y)}{f_2(y)} = 1$$

$$\int f(x|y)\,\mathrm{d}x = \int \frac{f(x,y)}{f_2(y)}\,\mathrm{d}x = \frac{1}{f_2(y)} \int f(x,y)\,\mathrm{d}x = \frac{f_2(y)}{f_2(y)} = 1.$$

(2) X and Y are independent $\iff$ $f(x|y) = f_1(x)$ for all $x \in A_1$ or $f(y|x) = f_2(y)$ for all $y \in A_2$.

Example. Let $f(x,y) = \frac{2}{\pi}$ for $0 < x < \sqrt{1-y^2}$, $-1 < y < 1$. Find $f_1(x|y)$ and $f_2(y|x)$.

Proof. The support looks like the following.



Then,

$$\begin{aligned} f_1(x) &= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{2}{\pi}\,\mathrm{d}y = \frac{4}{\pi}\sqrt{1-x^2} \quad \text{for } 0 < x < 1 \\ f_2(y) &= \int_0^{\sqrt{1-y^2}} \frac{2}{\pi}\,\mathrm{d}x = \frac{2}{\pi}\sqrt{1-y^2} \quad \text{for } -1 < y < 1. \end{aligned}$$

Thus,

$$\begin{aligned} f_1(x|y) &= \frac{f(x,y)}{f_2(y)} = \frac{2/\pi}{2/\pi\sqrt{1-y^2}} = \frac{1}{\sqrt{1-y^2}} \qquad -1 < y < 1, 0 < x < \sqrt{1-y^2} \\ f_2(y|x) &= \frac{f(x,y)}{f_1(x)} = \frac{2/\pi}{4/\pi\sqrt{1-x^2}} = \frac{1}{2\sqrt{1-x^2}} \qquad 0 < x < 1, \ -\sqrt{1-x^2} < y < \sqrt{1-x^2}. \end{aligned}$$

Hence, $X \mid Y \sim \mathrm{Unif}(0, \sqrt{1-y^2})$ and $Y \mid X \sim \mathrm{Unif}(-\sqrt{1-x^2}, \sqrt{1-x^2})$ and $X \not\perp\!\!\!\perp Y$. □

Proposition 4.6 (**Product Rule**).

$f(x,y) = f_1(x|y)f_2(y) = f_2(y|x)f_1(x)$.

Example. Find the marginal PMF of X if $Y \sim \text{Poi}(\mu)$ and $X \mid Y = y \sim \text{Bin}(y, p)$.

Proof. We will eventually show that $X \sim \text{Poi}(p\mu)$. Now,

$$\begin{aligned}
f_1(x|y) &= \binom{y}{x} p^x (1-p)^{y-x} \quad x = 0, 1, 2, \dots, y \\
f_2(y) &= \frac{e^{-\mu}\mu^y}{y!} \quad y = 0, 1, 2, \dots. \\
f(x,y) = f_1(x|y) f_2(y) &= \binom{y}{x} p^x (1-p)^{y-x} \frac{e^{-\mu}\mu^y}{y!} \quad x = 0, 1, 2, \dots, y \text{ and } y = 0, 1, 2, \dots.
\end{aligned}$$

Thus, we know that X is not independent of Y because the support is not a Cartesian product.

$$\begin{aligned}
f_1(x) = \sum_{y=x}^{\infty} f(x,y) &= \sum_{y=x}^{\infty} \frac{y!}{x!(y-x)!} p^x (1-p)^{y-x} \frac{e^{-\mu}\mu^y}{y!} \\
&= \frac{e^{-\mu} p^x \mu^x}{x!} \sum_{y=x}^{\infty} \frac{y!}{(y-x)! y!} (1-p)^{y-x} \mu^{y-x} \\
&= \frac{(p\mu)^x e^{-\mu}}{x!} \sum_{y-x=0}^{\infty} \frac{[(1-p)\mu]^{y-x}}{(y-x)!} \\
&= \frac{(p\mu)^x e^{-\mu}}{x!} e^{(1-p)\mu} = \frac{(p\mu)^x e^{-p\mu}}{x!} \quad x = 0, 1, 2, \dots.
\end{aligned}$$

Therefore, $X \sim \text{Poi}(p\mu)$. $\square$

4.4 Conditional Expectation

Definition 4.9 (Conditional Expectation).

The **conditional expectation** of $g(Y)$ given $X = x$ is

$$\mathbb{E}\left[g(Y) \mid X = x\right] = \begin{cases} \displaystyle\sum_{y} g(y) f_2(y|x) & \text{if } Y \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} g(y) f_2(y|x)\,\mathrm{d}y & \text{if } Y \text{ is continuous} \end{cases}$$

provided that the sum/integral converges absolutely. Consider the following.

1. $\mathbb{E}\left[Y \mid x\right]$ is the conditional mean of Y given $X = x$.
2. $\operatorname{Var}(Y \mid x)$ is the conditional variance of Y given $X = x$, which is given by

$$\begin{aligned} \operatorname{Var}(Y \mid x) &= \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y \mid x\right]\right)^2 \,\middle|\, x\right] \\ &= \mathbb{E}\left[Y^2 \mid x\right] - \left(\mathbb{E}\left[Y \mid x\right]\right)^2. \end{aligned}$$

Example. Let $f(x, y) = \frac{2}{\pi}$ for $0 < x < \sqrt{1 - y^2}$, $-1 < y < 1$. Find $\mathbb{E}\left[Y \mid x\right]$, $\mathbb{E}\left[Y^2 \mid x\right]$, and $\operatorname{Var}(Y \mid x)$.

Proof. Note that

$$\begin{aligned} f_2(y|x) &= \frac{1}{2\sqrt{1 - x^2}} \quad 0 < x < 1, -\sqrt{1 - x^2} < y < \sqrt{1 - x^2} \\ \implies \mathbb{E}\left[Y \mid x\right] &= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \underbrace{y \frac{1}{2\sqrt{1 - x^2}}}_{\text{odd}}\,\mathrm{d}y = 0 \quad 0 < x < 1. \\ \implies \mathbb{E}\left[Y^2 \mid x\right] &= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \underbrace{y^2 \frac{1}{2\sqrt{1 - x^2}}}_{\text{even}}\,\mathrm{d}y \\ &= 2\int_{0}^{\sqrt{1-x^2}} y^2 \frac{1}{2\sqrt{1 - x^2}}\,\mathrm{d}y = \frac{1 - x^2}{3} \quad 0 < x < 1. \\ \implies \operatorname{Var}(Y \mid x) &= \mathbb{E}\left[Y^2 \mid x\right] - \left(\mathbb{E}\left[Y \mid x\right]\right)^2 = \frac{1 - x^2}{3} \quad 0 < x < 1. \end{aligned}$$

□

Theorem 4.7. If $X \perp\!\!\!\perp Y$, then

$$\mathbb{E}\left[g(Y) \mid X = x\right] = \mathbb{E}\left[g(Y)\right] \quad \text{and} \quad \mathbb{E}\left[h(X) \mid y\right] = \mathbb{E}\left[h(X)\right].$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}\left[g(Y) \mid X = x\right] &= \begin{cases} \sum_y g(y) f_2(y|x) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} g(y) f_2(y|x)\,\mathrm{d}y & \text{if } Y \text{ is continuous} \end{cases} \\
&= \begin{cases} \sum_y g(y) f_2(y) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} g(y) f_2(y)\,\mathrm{d}y & \text{if } Y \text{ is continuous} \end{cases} = \mathbb{E}\left[g(Y)\right].
\end{aligned}$$

□

Lecture 12, 2025/02/12

Theorem 4.8 (Law of Total Expectation).
$\mathbb{E}\left[g(Y)\right] = \mathbb{E}\left[\mathbb{E}\left[g(Y) \mid X\right]\right]$.

Proof. We start from the RHS and look at the continuous case.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[g(Y) \mid X\right]\right] &= \int_{-\infty}^{\infty} \mathbb{E}\left[g(Y) \mid X = x\right] f_1(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(y) f_2(y|x)\,\mathrm{d}y \right) f_1(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y) f(x, y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} g(y) \left(\int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}x \right) \mathrm{d}y \\
&= \int_{-\infty}^{\infty} g(y) f_2(y)\,\mathrm{d}y = \mathbb{E}\left[g(Y)\right].
\end{aligned}$$

□

Remark (**Special Cases**). In particular,

(1) $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$ if g is the identity map.

(2) Variance Decomposition: $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y \mid X)] + \text{Var}(\mathbb{E}[Y \mid X])$.

Proof of (2). Note that $\text{Var}(Y \mid X) = \mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2$. Then,

$$\begin{aligned}
\mathbb{E}[\text{Var}(Y \mid X)] &= \mathbb{E}\left[\mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2\right] = \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[\mathbb{E}[Y \mid X]^2\right] \\
\text{Var}(\mathbb{E}[Y \mid X]) &= \mathbb{E}\left[\mathbb{E}[Y \mid X]^2\right] - (\mathbb{E}[\mathbb{E}[Y \mid X]])^2 = \mathbb{E}\left[\mathbb{E}[Y \mid X]^2\right] - (\mathbb{E}[Y])^2.
\end{aligned}$$

Thus, $\mathbb{E}[\text{Var}(Y \mid X)] + \text{Var}(\mathbb{E}[Y \mid X]) = \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 = \text{Var}(Y)$. □

Example. Suppose $P \sim \text{Unif}(0, 0.1)$ and $Y \mid P = p \sim \text{Bin}(10, p)$. Find $\mathbb{E}[Y]$ and $\text{Var}(Y)$.

Proof. Note that

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid P]] = \mathbb{E}[10P] = 10\mathbb{E}[P] = 10 \times \frac{0.1}{2} = 0.5 \\
\text{Var}(Y) &= \mathbb{E}[\text{Var}(Y \mid P)] + \text{Var}(\mathbb{E}[Y \mid P]) = \mathbb{E}[10P(1-P)] + \text{Var}(10P) \\
&= 10\mathbb{E}[P] - 10\mathbb{E}\left[P^2\right] + 10^2\,\text{Var}(P) = 10 \times \frac{0.1}{2} - 10 \times \frac{0.1^2}{3} + 10^2 \times \frac{0.1^2}{12} = \frac{11}{20}.
\end{aligned}$$ □

4.5 Joint Expectation

Definition 4.10 (**Joint Expectation**).
Suppose X and Y are discrete random variables with joint PMF $f(x, y)$ and support A. Then, the **joint expectation** of $g(X, Y)$ is

$$\mathbb{E}[h(X, Y)] = \sum\sum_{(x,y)\in A} h(x, y) f(x, y).$$

Similarly, if X and Y are continuous, then the joint expectation of $g(X, Y)$ is

$$\mathbb{E}[h(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y) f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

provided that the sum/integral converges absolutely.

Remark (**Independence**). If $X \perp\!\!\!\perp Y$, then $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$.

Proof of Remark. We start with the LHS and look at the continuous case.

$$\begin{aligned}\mathbb{E}[g(X)h(Y)] &= \iint_{A_1\times A_2} g(x)h(y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{A_1} g(x)f_1(x)\,\mathrm{d}x \int_{A_2} h(y)f_2(y)\,\mathrm{d}y \\ &= \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]\,.\end{aligned}$$

□

Remark. More generally, if $X_1, X_2, \ldots, X_n$ are independent, then

$$\mathbb{E}\left[\prod_{i=1}^{n} h_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}[h_i(X_i)]\,.$$

4.6 Joint Moment Generating Functions

Definition 4.11 (Joint MGFs).

If X and Y are random variables, then

$$M(t_1, t_2) = \mathbb{E}\left[e^{t_1X + t_2Y}\right]$$

is called the **joint moment generating function** of X and Y if the expectation exists for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for some $h_1, h_2 > 0$.

Remark (**Marginal MGFs**). The marginal MGF of X and Y are

$$\begin{aligned}M_X(t) &= \mathbb{E}\left[e^{tX}\right] = M(t, 0) \quad t \in (-h_1, h_1) \\ M_Y(t) &= \mathbb{E}\left[e^{tY}\right] = M(0, t) \quad t \in (-h_2, h_2).\end{aligned}$$

Theorem 4.9 (**Independence Theorem for MGFs**).

Suppose X and Y are random variables with joitn MGF $M(t_1, t_2)$. Then

$$X \perp\!\!\!\perp Y \iff M(t_1, t_2) = M_X(t_1)M_Y(t_2)$$

for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$.

Proof.

($\Rightarrow$): If $X \perp\!\!\!\perp Y$, then

$$\begin{aligned} M(t_1, t_2) &= \iint_{A_1 \times A_2} e^{t_1 x + t_2 y} f(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_{A_1} e^{t_1 x} f_1(x)\, \mathrm{d}x \int_{A_2} e^{t_2 y} f_2(y)\, \mathrm{d}y = M_X(t_1)M_Y(t_2). \end{aligned}$$

($\Leftarrow$): If $M(t_1, t_2) = M_X(t_1)M_Y(t_2)$, then there exists (S, T) s.t. $S \perp\!\!\!\perp T$ and S has the same distribution as X and T has the same distribution as Y. Then,

$$M_{(S,T)}(t_1, t_2) = M_X(t_1)M_Y(t_2) = M_{(X,Y)}(t_1, t_2) = M(t_1, t_2).$$

By the Unique Theorem for MGFs, we have (X, Y) and (S, T) have the same joint distribution. Thus, $X \perp\!\!\!\perp Y$. □

Example. Let $f(x, y) = e^{-y}$ for $0 < x < y < \infty$. Find the joint MGF of X and Y. Are X and Y independent? What is the marginal distribution of X and Y?

Proof. First, note that X and Y are not independent because the support is not a Cartesian product. Then,

$$\begin{aligned} M(t_1, t_2) &= \int_0^\infty \int_0^y e^{t_1 x + t_2 y} e^{-y}\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_0^\infty e^{(t_2 - 1)y} \frac{1}{t_1} e^{t_1 x}\Big|_0^{x=y}\, \mathrm{d}y \\ &= \frac{1}{t_1}\int_0^\infty e^{(t_1 + t_2 - 1)y}\, \mathrm{d}y - \frac{1}{t_1}\int_0^\infty e^{(t_2 - 1)y}\, \mathrm{d}y \\ &= \frac{1}{t_1(t_1 + t_2 - 1)} e^{(t_1 + t_2 - 1)y}\Big|_0^\infty - \frac{1}{t_1(t_2 - 1)} e^{(t_2 - 1)y}\Big|_0^\infty = \frac{1}{(1 - t_1 - t_2)(1 - t_2)} \end{aligned}$$

for $t_1 + t_2 < 1$ and $t_2 < 1$. Also, the marginal distribution of X and Y are

$$M_X(t_1) = M(t_1, 0) = \frac{1}{1-t_1} \quad t_1 < 1$$
$$M_Y(t_2) = M(0, t_2) = \frac{1}{(1-t_2)^2} \quad t_2 < 1.$$

Thus, $X \sim \text{Gamma}(1,1) = \text{Exp}(1)$ and $Y \sim \text{Gamma}(2,1)$. □

Lecture 13, 2025/02/24

Theorem 4.10 (**Computing Joint Moments**).

Let $M(t_1, t_2)$ be the joint MGF of X and Y. Then

$$\mathbb{E}\left[X^j Y^k\right] = \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M(t_1, t_2)\bigg|_{(t_1,t_2)=(0,0)}$$

Proof. Note that

$$\begin{aligned}
\frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M(t_1, t_2) &= \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} \mathbb{E}\left[e^{t_1 X + t_2 Y}\right] \\
&= \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{t_1 x + t_2 y} f(x,y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} e^{t_1 x + t_2 y} f(x,y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^j y^k e^{t_1 x + t_2 y} f(x,y)\, \mathrm{d}x\, \mathrm{d}y\,.
\end{aligned}$$

Thus, $\displaystyle \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M(t_1, t_2)\bigg|_{(t_1,t_2)=(0,0)} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^j y^k f(x,y)\, \mathrm{d}x\, \mathrm{d}y = \mathbb{E}\left[X^j Y^k\right].$ □

Definition 4.12 (**Covariance and Correlation Coefficient**).

The **covariance** of X and Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. The **correlation coefficient** of X and Y is

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\mathbb{E}[(X-\mu_X)^2]}$ and $\sigma_Y = \sqrt{\mathrm{Var}(Y)} = \sqrt{\mathbb{E}[(Y-\mu_Y)^2]}$.

Remark. ρ only measures the linear relationship between X and Y. Also, we have

(1) $-1 \le \rho(X,Y) \le 1$.

(2) $\rho(X,Y) = 1 \iff Y = aX + b$ for some $a > 0$. Similarly, $\rho(X,Y) = -1 \iff Y = aX + b$ for some $a < 0$.

Proof of Remark.

(1) Let $W = \frac{Y}{\sigma_Y} - \rho\frac{X}{\sigma_X}$. We know that

$$\begin{aligned} 0 \le \mathrm{Var}(W) &= \mathrm{Var}\left(\frac{Y}{\sigma_Y} - \rho\frac{X}{\sigma_X}\right) \\ &= \frac{\sigma_Y^2}{\sigma_Y^2} + \rho^2\frac{\sigma_X^2}{\sigma_Y^2} - 2\rho\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} \\ &= 1 + \rho^2 - 2\rho^2 = 1 - \rho^2 \ge 0. \end{aligned}$$

Thus, $\rho^2 \le 1 \implies -1 \le \rho \le 1$.

(2) ($\Rightarrow$): If $|\rho| = 1$, then $\mathrm{Var}(W) = 0$, W is degenerated at its mean with probability 1. That is, $\mathbb{P}(W = \mu_W) = 1$. Thus,

$$\begin{aligned} \frac{Y}{\sigma_Y} - \rho\frac{X}{\sigma_X} &= \frac{\mu_Y}{\sigma_Y} - \rho\frac{\mu_X}{\sigma_X} \\ \implies Y &= \underbrace{\rho\frac{\sigma_Y}{\sigma_X}}_{a} X + \underbrace{\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X}_{b}. \end{aligned}$$

($\Leftarrow$): If $Y = aX + b$, then $\sigma_Y^2 = a^2\sigma_X^2$ and

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathbb{E}[(X-\mu_X)(aX - a\mu_X)] \\ &= a\mathbb{E}\left[(X-\mu_X)^2\right] = a\sigma_X^2. \end{aligned}$$

Thus, $\rho = \dfrac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} = \dfrac{a\sigma_X^2}{\sigma_X |a| \sigma_X} = \dfrac{a}{|a|} = \pm 1.$

$\square$

Remark (**Some Results**).

(1) $\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

(2) $\rho(X, X^2) = 0$ if X has a symmetric distribution about its mean.

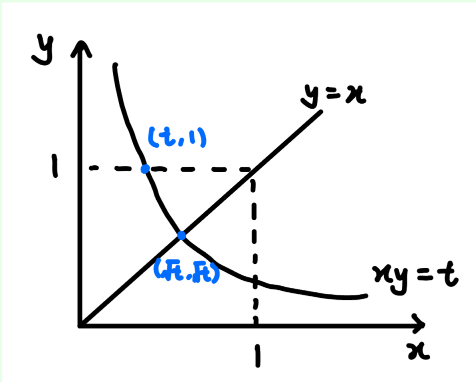
5 More on Joint Distributions

5.1 Distributions of Functions of Multiple Random Variables

First, we look at the CDF technique.

Example. Let $f(x, y) = 3y$ with $0 < x \leq y < 1$. Find the PDF of $T = XY$.

Proof. The plot of this region is shown below.



We have

$$\begin{aligned} G(t) = \mathbb{P}(T \leq t) &= \mathbb{P}(XY \leq t) \\ &= \mathbb{P}(Y \leq \frac{t}{X}) \\ &= 1 - \int_{\sqrt{t}}^{1} \int_{t/y}^{y} 3y \, dx \, dy \\ &= 3t - 2t\sqrt{t} \quad \text{for } 0 < t \leq 1. \end{aligned}$$

Thus, $g(t) = G'(t) = 3 - 3\sqrt{t}$ for $0 < t < 1$. □

Example (Order Statistics).
Suppose $X_1, \dots, X_n$ are i.i.d. continuous random variables with PDF $f(x)$ and CDF $F(x)$. Find the PDF of $Y = \max(X_1, \dots, X_n) = X_{(n)}$ and $T = \min(X_1, \dots, X_n) = X_{(1)}$.

Proof. Let $Y = X_{(n)}$. We have

$$
\begin{aligned}
& G(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X_{(n)} \le y) = \mathbb{P}(X_1 \le y, \dots, X_n \le y) = F(y)^n \\
\implies & g(y) = G'(y) = nF(y)^{n-1} f(y).
\end{aligned}
$$

Also, let $T = X_{(1)}$. We have

$$
\begin{aligned}
H(t) = \mathbb{P}(T \le t) = 1 - \mathbb{P}(T > t) &= 1 - \mathbb{P}(X_{(1)} > t) \\
&= 1 - \mathbb{P}(X_1 > t, \dots, X_n > t) = 1 - (1 - F(t))^n \\
\implies h(t) = H'(t) &= nf(t)(1 - F(t))^{n-1}.
\end{aligned}
$$

□

Definition 5.1 (**Jacobian of One-to-One Transformation**).

Let $S : (x, y) \to (u, v)$ be a one-to-one transformation for all $(x, y) \in \mathbb{R}_{xy}$ where $u = h_1(x, y)$ and $v = h_2(x, y)$. There exists an inverse transformation $T = S^{-1} : (u, v) \to (x, y)$ for all $(u, v) \in \mathbb{R}_{uv}$ where $x = w_1(u, v)$ and $y = w_2(u, v)$. The **Jacobian** of the transformation T is

$$
\frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \left[\frac{\partial(u, v)}{\partial(x, y)} \right]^{-1}
$$

where $\frac{\partial(u,v)}{\partial(x,y)}$ is the Jacobian of the transformation S.

Remark (**Inverse Mapping Theorem: How do we check if S is one-to-one?**).
If $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}$ are continuous functions and $\frac{\partial(u,v)}{\partial(x,y)} \neq 0$ for all $(x, y) \in \mathbb{R}_{xy}$, then S is one-to-one and $T = S^{-1}$ exists.

Theorem 5.1 (**One-to-One Bivariate Transformation Theorem**).

Let X and Y be continuous random variables with joint PDF $f(x, y)$ and support $\mathbb{R}_{xy} = \{(x, y) : f(x, y) > 0\}$. Suppose $S : U = h_1(X, Y), V = h_2(X, Y)$ is a one-to-one transformation with inverse $T = S^{-1} : X = w_1(U, V), Y = w_2(U, V)$. Suppose also that S maps $\mathbb{R}_{xy}$ into $\mathbb{R}_{uv}$. Then, the joint PDF of U and V is given by

$$
g(u, v) = f(w_1(u, v), w_2(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| \quad \text{for all } (u, v) \in \mathbb{R}_{uv}.
$$

Example. Suppose that $X \sim \text{Gamma}(a, 1)$ and $Y \sim \text{Gamma}(b, 1)$ are independent. Find the joint PDF of $U = X + Y$ and $V = \frac{X}{X+Y}$. Show that $U \sim \text{Gamma}(a + b, 1)$ and $V \sim \text{Beta}(a, b)$.

Proof. Let $S : U = X + Y, V = \frac{X}{X+Y}$. We have $T = S^{-1} : X = w_1(U, V) = UV, Y = w_2(U, V) = U(1 - V)$. The Jacobian of the transformation T is

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} v & u \\ 1 - v & -u \end{vmatrix} = -uv - u(1 - v) = -u.$$

Note that $\mathbb{R}_{uv} = (0, \infty) \times (0, 1)$. Then,

$$\begin{aligned} f_1(x) &= \frac{x^{a-1}e^{-x}}{\Gamma(a)} \quad \text{for } x > 0, \\ f_2(y) &= \frac{y^{b-1}e^{-y}}{\Gamma(b)} \quad \text{for } y > 0 \\ \implies f(x, y) &= f_1(x)f_2(y) = \frac{x^{a-1}y^{b-1}e^{-(x+y)}}{\Gamma(a)\Gamma(b)} \quad \text{for } x, y > 0 \\ \implies g(u, v) &= f(uv, u(1 - v)) \cdot |-u| \\ &= \frac{(uv)^{a-1}(u(1 - v))^{b-1}e^{-u}}{\Gamma(a)\Gamma(b)} \cdot |-u| \\ &= \underbrace{\frac{u^{a+b-1}e^{-u}}{\Gamma(a + b)}}_{U\sim\text{Gamma}(a+b,1)} \cdot \underbrace{\frac{v^{a-1}(1 - v)^{b-1}\Gamma(a + b)}{\Gamma(a)\Gamma(b)}}_{V\sim\text{Beta}(a,b)} \quad u > 0, 0 < v < 1. \end{aligned}$$

Note that $U \perp\!\!\!\perp V$ by Factorization Theorem. $\square$

Lecture 14, 2025/02/26

Example. Back to a previous example, $f(x, y) = 3y$ with $0 < x \leq y < 1$. Find the PDF of $U = XY$.

Proof. We want to construct a one-to-one transformation. Let $S : U = XY, V = X$. Then, $T = S^{-1} : X = V, Y = \frac{U}{V}$. Also, $\mathbb{R}_{uv} = \{(u, v) : 0 < v \leq \frac{u}{v} < 1\} = \{(u, v) : v^2 < u < v, 0 < v < 1\}$. The Jacobian of the transformation T is

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}.$$

Then,

$$
\begin{aligned}
f(u,v) &= 3\frac{u}{v} \cdot \left|-\frac{1}{v}\right| \\
&= \frac{3u}{v^2} \quad (u,v) \in \mathbb{R}_{uv}.
\end{aligned}
$$

Thus, $f_1(u) = \int_u^{\sqrt{u}} \frac{3u}{v^2}\, dv = 3 - 3\sqrt{u}$ for $0 < u < 1$. □

5.2 Moment Generating Function Technique

Theorem 5.2 (MGF of Sum of Independent Random Variables).
Suppose that $X_1, \ldots, X_n$ are independent random variables and X_i has MGF $M_i(t)$ which exists for some $t \in (-h, h)$ for some $h > 0$. The MGF of $Y = \sum_{i=1}^n X_i$ is given by

$$
M_Y(t) = \prod_{i=1}^n M_i(t) \quad \text{for } t \in (-h, h).
$$

Remark. If X_i's are i.i.d. random variables each with MGF $M(t)$, then $M_Y(t) = M(t)^n$.

Proof. The MGF of $Y = \sum_{i=1}^n X_i$ is

$$
\begin{aligned}
M_Y(t) &= \mathbb{E}\left[e^{t\sum_{i=1}^n X_i}\right] = \mathbb{E}\left[e^{tX_1} \cdots e^{tX_n}\right] \\
&= \mathbb{E}\left[e^{tX_1}\right] \cdots \mathbb{E}\left[e^{tX_n}\right] \\
&= M_1(t) \cdots M_n(t) = \prod_{i=1}^n M_i(t).
\end{aligned}
$$

□

Proposition 5.3 (Special Results).

(1) If $X \sim \text{Gamma}(\alpha, \beta)$, where α is a positive integer, then $\frac{2X}{\beta} \sim \chi^2(2\alpha)$.

(2) If $X_i \sim \text{Gamma}(\alpha_i, \beta)$ independently, $i = 1, \ldots, n$, then $\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$.

(3) If $X_i \sim \text{Gamma}(1, \beta) = \text{Exp}(\beta)$ independently for $i = 1, \ldots, n$, then $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \beta)$.

(4) If $X_i \sim \text{Gamma}\left(\frac{k_i}{2}, 2\right) = \chi^2(k_i)$ independently for $i = 1, \ldots, n$, then $\sum_{i=1}^n X_i \sim \chi^2\left(\sum_{i=1}^n k_i\right)$.

(5) If $X_i \sim \mathrm{N}(\mu, \sigma^2)$ independently for $i = 1, \ldots, n$, then $\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$.
(6) If $X_i \sim \mathrm{Poi}(\mu_i)$ independently for $i = 1, \ldots, n$, then $\sum_{i=1}^n X_i \sim \mathrm{Poi}\left(\sum_{i=1}^n \mu_i\right)$.
(7) If $X_i \sim \mathrm{Bin}(n_i, p)$ independently for $i = 1, \ldots, n$, then $\sum_{i=1}^n X_i \sim \mathrm{Bin}\left(\sum_{i=1}^n n_i, p\right)$.
(8) If $X_i \sim \mathrm{NB}(k_i, p)$ independently for $i = 1, \ldots, n$, then $\sum_{i=1}^n X_i \sim \mathrm{NB}\left(\sum_{i=1}^n k_i, p\right)$.

Proof.

(1) We know that $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < \frac{1}{\beta}$. The MGF of $\frac{2X}{\beta}$ is

$$\begin{aligned} M_{\frac{2X}{\beta}}(t) &= \mathbb{E}\left[e^{t\frac{2X}{\beta}}\right] = M_X\left(\frac{2t}{\beta}\right) \\ &= \left(1 - \beta\frac{2t}{\beta}\right)^{-\alpha} = (1-2t)^{-\alpha} \quad \text{for } t < \frac{1}{2}. \end{aligned}$$

This is the MGF of $\mathrm{Gamma}(\alpha, 2)$, i.e. $\chi^2(2\alpha)$.

(2) By Theorem 5.2, we have

$$\begin{aligned} M_{\sum_{i=1}^n X_i}(t) &= \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n (1-\beta t)^{-\alpha_i} \\ &= (1-\beta t)^{-\sum_{i=1}^n \alpha_i} \quad \text{for } t < \frac{1}{\beta} \end{aligned}$$

which is the MGF of $\mathrm{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$. Thus, $\sum_{i=1}^n X_i \sim \mathrm{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$ by the Uniqueness Theorem for MGFs.

(5) Suppose that $X_i \sim \mathrm{N}(\mu, \sigma^2)$ independently for $i = 1, \ldots, n$. Then, $\frac{X_i-\mu}{\sigma} \sim \mathrm{N}(0, 1)$ independently. Then, $\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(1)$ independently. Therefore, $\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$.

(6) Suppose that $X_i \sim \mathrm{Poi}(\mu_i)$ independently for $i = 1, \ldots, n$. We know that $M_{X_i}(t) = e^{\mu_i(e^t-1)}$. Then, by Theorem 5.2, we have

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n e^{\mu_i(e^t-1)} = e^{\left(\sum_{i=1}^n \mu_i\right)(e^t-1)} \quad \text{for } t \in \mathbb{R}.$$

This is the MGF of $\mathrm{Poi}\left(\sum_{i=1}^n \mu_i\right)$. □

Note. Proofs for (3), (4), (7), and (8) are left as exercises.

Theorem 5.4 (**Linear Combination of Independent Normal Random Variables**).
If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ independently for $i = 1, \dots, n$, then

$$\sum_{i=1}^n a_i X_i \sim \mathrm{N}\left(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2\right).$$

Proof. Suppose that $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ independently for $i = 1, \dots, n$. Then, $M_{X_i}(t) = e^{\mu_i t + \frac{1}{2}\sigma_i^2 t^2}$ for $t \in \mathbb{R}$. By Theorem 5.2, we have

$$\begin{aligned} M_{\sum_{i=1}^n a_i X_i}(t) &= \prod_{i=1}^n M_{X_i}(a_i t) = \prod_{i=1}^n e^{a_i\mu_i t + \frac{1}{2}a_i^2\sigma_i^2 t^2} \\ &= e^{\left(\sum_{i=1}^n a_i\mu_i\right)t + \frac{1}{2}t^2\left(\sum_{i=1}^n a_i^2\sigma_i^2\right)} \quad \text{for } t \in \mathbb{R} \end{aligned}$$

which is the MGF of $\mathrm{N}\left(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2\right)$. □

Lecture 15, 2025/03/03

5.3 Bivariate Normal Distribution

Definition 5.2 (**Bivariate Normal Distribution**).
If (X_1, X_2) is a random vector with joint PDF:

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right)$$

where $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ and Σ is a non-singular matrix.
Then, (X_1, X_2) is said to have a **bivariate normal distribution**, $X \sim \mathrm{BVN}(\mu, \Sigma)$.

Note. The $(1,1)$ entry of Σ is $\mathrm{Var}(X_1)$, the $(2,2)$ entry of Σ is $\mathrm{Var}(X_2)$, and the $(1,2)$ entry of Σ is $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$.

Proposition 5.5 (**Properties of Bivariate Normal Distribution**).

Suppose that $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{BVN}(\mu, \Sigma)$. Then,

(1) X has joint MGF $M(t_1, t_2) = \exp\left(\mu^\top t + \frac{1}{2} t^\top \Sigma t\right)$ where $t = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \in \mathbb{R}^2$.

(2) $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$.

(3) $\mathrm{Cov}\,(X_1, X_2) = \rho\sigma_1\sigma_2$ and $\mathrm{Corr}\,(X_1, X_2) = \rho$, where $-1 \leq \rho \leq 1$.

(4) $X_1 \perp\!\!\!\perp X_2 \iff \rho = 0$.

(5) Let $c = (c_1, c_2)^\top \in \mathbb{R}^2$ be a non-zero vector of constants, then

$$c^\top X = c_1 X_1 + c_2 X_2 \sim \mathrm{N}(c^\top \mu, c^\top \Sigma c).$$

(6) If A is a 2×2 non-singular matrix and b is a 2×1 vector, then

$$Y = AX + b \sim \mathrm{BVN}(A\mu + b, A\Sigma A^\top).$$

Proof.

(1) See Prof's written notes.

(2) We look at X_1 first.

$$M_{X_1}(t_1) = M(t_1, 0) = e^{\mu_1 t_1 + \frac{1}{2}\sigma_1^2 t_1^2} \quad t_1 \in \mathbb{R}$$

which is the MGF of $\mathrm{N}(\mu_1, \sigma_1^2)$. By the Uniqueness Theorem for MGFs, $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$. Similarly, we can show that $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$.

(3) Note that

$$\begin{aligned} \mathbb{E}\,[X_1 X_2] &= \frac{\partial^2}{\partial t_1 \partial t_2} M(t_1, t_2)\bigg|_{(t_1,t_2)=(0,0)} \\ &= \rho\sigma_1\sigma_2 + \mu_1\mu_2. \\ \mathrm{Cov}\,(X_1, X_2) &= \mathbb{E}\,[X_1 X_2] - \mu_1\mu_2 \\ &= \rho\sigma_1\sigma_2 + \mu_1\mu_2 - \mu_1\mu_2 = \rho\sigma_1\sigma_2. \end{aligned}$$

(4) Note that $M_{(X_1,X_2)}(t_1,t_2) = M_{X_1}(t_1)M_{X_2}(t_2) \iff \rho = 0$.

(5) We have

$$\begin{aligned} M_{c^\top X}(t) &= \mathbb{E}\left[e^{t^\top c^\top X}\right] = \mathbb{E}\left[e^{(ct)^\top X}\right] \\ &= \exp\left(\underbrace{\mu^\top c}_{(\mu^*)^\top}\, t + \frac{1}{2}t^\top \underbrace{c^\top \Sigma c}_{\Sigma^*}\, t\right). \end{aligned}$$

By the Uniqueness Theorem for MGFs, $c^\top X \sim \mathrm{N}(\mu^\top c, c^\top \Sigma c)$.

(6) Consider the MGF of $Y = AX + b$:

$$\begin{aligned} M_Y(t) &= \mathbb{E}\left[e^{t^\top (AX+b)}\right] = e^{t^\top b}\mathbb{E}\left[e^{t^\top AX}\right] \\ &= e^{t^\top b}\mathbb{E}\left[e^{(A^\top t)^\top X}\right] \\ &= e^{t^\top b}\exp\left(\mu^\top A^\top t + \frac{1}{2}(A^\top t)^\top \Sigma A^\top t\right) \\ &= \exp\left(\underbrace{(A\mu + b)}_{\mu^*}{}^\top t + \frac{1}{2}t^\top \underbrace{A\Sigma A^\top}_{\Sigma^*}\, t\right) \end{aligned}$$

which is the MGF of $\mathrm{BVN}(A\mu + b, A\Sigma A^\top)$. $\square$

5.4 Multinomial Distribution

Definition 5.3 (Multinomial Distribution).

Suppose that $(X_1, \ldots, X_k)$ is a discrete random vector with joint PMF:

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

for $x_i = 0, 1, \ldots, n$, $i = 1, \ldots, k$, $\sum_{i=1}^k x_i = n$, $0 < p_i < 1\ \forall i$ and $\sum_{i=1}^k p_i = 1$. Then, $(X_1, \ldots, X_k)$ is said to have a **multinomial distribution**, $(X_1, \ldots, X_k) \sim \mathrm{Multinomial}(n; p_1, \ldots, p_k)$.

Example. Recall a previous example where we select 5 students from 10 actsci students, 9 stat students and 6 math students.

- If we sample without replacement: Extended Hypergeometric.
- If we sample with replacement: Multinomial.

Remark.

(1) If we conduct n independent trials with each trial resulting in one of k categories (outcomes) with probabilities $p_1, \dots, p_k$ (and $\sum_{i=1}^{k} p_i = 1$), then the number of times each category occurs in the n trials follows a multinomial distribution.

(2) $x_k = n - \sum\limits_{i=1}^{k-1} x_i$ and $p_k = 1 - \sum\limits_{i=1}^{k-1} p_i$. Thus, X_k is redundant.

Lecture 16, 2025/03/05

Example. Rolling a fair die 10 times, the number of times each number appears follow Multinomial$\left(10; \frac{1}{6}, \dots, \frac{1}{6}\right)$.

Proposition 5.6 (**Properties of Multinomial Distribution**).

Suppose that $X = (X_1, \dots, X_k) \sim \text{Multinomial}(n; p_1, \dots, p_k)$. Then,

(1) $(X_1, \dots, X_k)$ has joint MGF:

$$M(t_1, \dots, t_k) = \left(p_1 e^{t_1} + \cdots + p_k e^{t_k}\right)^n.$$

(2) Any subset of $(X_1, \dots, X_k)$ also has a multinomial distribution.

In particular, $X_i \sim \text{Binomial}(n, p_i)$ for $i = 1, \dots, k$.

(3) If $T = X_i + X_j$ for $i \neq j$, then $T \sim \text{Binomial}(n, p_i + p_j)$.

(4) $\text{Cov}\left(X_i, X_j\right) = -np_i p_j$ for $i \neq j$.

(5) The conditional distribution of $X_i \mid X_j = x_j$ with $i \neq j$, is Binomial$\left(n - x_j, \frac{p_i}{1-p_j}\right)$.

(6) The conditional distribution of $X_i \mid T = X_i + X_j = t$, with $i \neq j$, is Binomial$\left(t, \frac{p_i}{p_i+p_j}\right)$.

Proof.

(1) We have

$$\begin{aligned}
M(t_1, \dots, t_k) &= \mathbb{E}\left[e^{t_1 X_1 + \cdots t_k X_k}\right] \\
&= \sum_{x_1 + \cdots + x_k = n} e^{t_1 x_1 + \cdots + t_k x_k} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \\
&= \sum_{x_1 + \cdots + x_k = n} \frac{n!}{x_1! \cdots x_k!} (p_1 e^{t_1})^{x_1} \cdots (p_k e^{t_k})^{x_k} \\
&= \left(p_1 e^{t_1} + \cdots + p_k e^{t_k}\right)^n \quad (t_1, \dots, t_k) \in \mathbb{R}^k \text{ by Multinomial Theorem.}
\end{aligned}$$

(2) Note that

$$\begin{aligned}
M_{X_i}(t_i) = M(0, \dots, t_i, \dots, 0) &= \left(p_1 e^0 + \cdots + p_k e^{t_i} + \cdots + p_k e^0\right)^n \\
&= (p_i e^{t_i} + 1 - p_i)^n \quad \text{for } t_i \in \mathbb{R}
\end{aligned}$$

which is the MGF of Binomial(n, p_i).

(3) Note that

$$\begin{aligned}
M_T(t) = \mathbb{E}\left[e^{t(X_i + X_j)}\right] &= M(0, \dots, \underbrace{t}_{i}, \dots, \underbrace{t}_{j}, \dots, 0) \\
&= \left(p_i e^t + p_j e^t + 1 - p_i - p_j\right)^n \\
&= ((p_i + p_j)e^t + 1 - (p_i + p_j))^n \quad \text{for } t \in \mathbb{R}
\end{aligned}$$

which is the MGF of Binomial$(n, p_i + p_j)$.

(4) We have

$$
\begin{aligned}
\mathbb{E}\left[X_i X_j\right] &= \frac{\partial^2}{\partial t_i \partial t_j} M(0, \dots, 0, t_i, \dots, t_j, 0 \dots, 0)\Big|_{t_i=0=t_j} \\
&= \frac{\partial^2}{\partial t_i \partial t_j} \left(p_i e^{t_i} + p_j e^{t_j} + 1 - p_i - p_j\right)^n \Big|_{t_i=0=t_j} \\
&= \frac{\partial}{\partial t_i} n p_j e^{t_j} \left(p_i e^{t_i} + p_j e^{t_j} + 1 - p_i - p_j\right)^{n-1} \Big|_{t_i=0=t_j} \\
&= n(n-1) p_i p_j \left(p_i e^{t_i} + p_j e^{t_j} + 1 - p_i - p_j\right)^{n-2} \Big|_{t_i=0=t_j} \\
&= n(n-1) p_i p_j.
\end{aligned}
$$

Also, $\mathbb{E}[X_i] = np_i$ and $\mathbb{E}\left[X_j\right] = np_j$. Thus, $\operatorname{Cov}\left(X_i, X_j\right) = \mathbb{E}\left[X_i X_j\right] - \mathbb{E}[X_i]\,\mathbb{E}\left[X_j\right] = -np_i p_j$.

(5) Exercise. Hint: use the definition of conditional probability, i.e.

$$
\mathbb{P}(X_i = x_i \mid X_j = x_j) = \frac{\mathbb{P}(X_i = x_i, X_j = x_j)}{\mathbb{P}(X_j = x_j)}.
$$

(6) Exercise, similar to (5). $\square$

6 Asymptotic Distributions

Suppose that we want to measure the average height of women in Canada. Then, we take a sample from $\mathrm{N}(\mu, \sigma^2)$.

Goal: Estimate μ.

Observations	Statistic: Sample Mean
Z_1	$X_1 = Z_1$
Z_1, Z_2	$X_2 = \frac{Z_1+Z_2}{2}$
Z_1, Z_2, Z_3	$X_3 = \frac{Z_1+Z_2+Z_3}{3}$
$\vdots$	$\vdots$
$Z_1, \dots, Z_n$	$X_n = \frac{Z_1+\dots+Z_n}{n}$

Question: What is the limiting (or asymptotic) distribution of $\underline{X_n}$ when $n \to \infty$?

6.1 Convergence in Distribution

Definition 6.1 (**Convergence in Distribution**).
Let $X_1, \dots, X_n$ be a sequence of random variables such that X_n has CDF $F_n(x)$ for $n = 1, 2, \dots$. Let X be a random variable with CDF $F(x)$. We say that X_n **converges in distribution** to X and denote this by $X_n \xrightarrow{d} X$ if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at all points x at which $F(x)$ is continuous. We call F the **limiting/asymptotic distribution** of X_n.

Note. $X_n \xrightarrow{d} X$ means $\mathbb{P}(X_n \leq a) \approx \mathbb{P}(X \leq a)$ for large n but does not mean $X_n(\omega) \approx X(\omega)$.

Example. Let $X_i \sim \mathrm{Exp}(1)$, $i = 1, 2, \dots$, independently. Consider $Y_1, \dots, Y_n$ where $Y_n = \max(X_1, \dots, X_n) - \log(n) = X_{(n)} - \log(n)$. Find the limiting distribution of Y_n.

Proof. Note that $F_i(x) = 1 - e^{-x}$ for $x \geq 0$. Let $G_n(y)$ be the CDF of Y_n. Then,

$$
\begin{aligned}
G_n(y) = \mathbb{P}(Y_n \leq y) &= \mathbb{P}(X_{(n)} - \log(n) \leq y) \\
&= \mathbb{P}(X_{(n)} \leq y + \log(n)) \\
&\overset{\text{ind}}{=} \prod_{i=1}^{n} \mathbb{P}(X_i \leq y + \log(n)) \quad \text{for } \log(n) + y > 0 \\
&= \left(1 - e^{-(y+\log(n))}\right)^n \\
&= \left(1 - \frac{e^{-y}}{n}\right)^n \quad y > -\log(n).
\end{aligned}
$$

Then, $\lim_{n\to\infty} G_n(y) = \lim_{n\to\infty} \left(1 - \frac{e^{-y}}{n}\right)^n = e^{-e^{-y}}$ for $y \in \mathbb{R}$. Thus, $Y_n \xrightarrow{d} Y$ where Y has CDF $F(y) = e^{-e^{-y}}$, $y \in \mathbb{R}$ and Y follows a Gumbel distribution. □

Question; How do we show that $F(y) = e^{-e^{-y}}$ is a valid CDF?

Solution:

1. $\lim_{y\to-\infty} e^{-e^{-y}} = 0$ and $\lim_{y\to\infty} e^{-e^{-y}} = 1$.
2. $(e^{-e^{-y}})' = e^{-y}e^{-e^{-y}} > 0$ for all $y \in \mathbb{R}$. Thus, $F(y)$ is increasing.
3. $F(y)$ is continuous $\implies$ $F(y)$ is right-continuous.

Remark (**Limits Review**).

(1) $\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e$.

(2) $\lim_{n\to\infty} \left(1 + \frac{b}{n}\right)^{cn} = e^{bc}$.

(3) $\lim_{n\to\infty} \left(1 + \frac{b}{n} + \frac{\psi(n)}{n}\right)^{cn} = e^{bc}$ if $\psi(n) \to 0$.

Example. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Pareto}(1, 1)$ with $F(x) = \frac{x}{1+x}$ for $x > 0$. Let $Y_n = nX_{(1)} = n \cdot \min(X_1, \dots, X_n)$. Find the limiting distribution of Y_n.

Proof. We have

$$\begin{aligned}
F_n(y) = \mathbb{P}(Y_n \le y) = \mathbb{P}(nX_{(1)} \le y) \\
&= \mathbb{P}\left(X_{(1)} \le \frac{y}{n}\right) \\
&= 1 - \mathbb{P}\left(X_{(1)} > \frac{y}{n}\right) \\
&\overset{\text{ind}}{=} 1 - \prod_{i=1}^{n} \left(\frac{1}{1+\frac{y}{n}}\right) \\
&= 1 - \left(1 + \frac{y}{n}\right)^{-n} \qquad y > 0.
\end{aligned}$$

Then, $\lim_{n\to\infty} F_n(y) = 1 - e^{-y}$ for $y > 0$. Thus, $Y_n \xrightarrow{d} Y$ where $Y \sim \text{Exp}(1)$. □

Question: How about $Y_n = X_{(n)} = \max(X_1, \dots, X_n)$?
Note that $F_n(y) = \mathbb{P}(Y_n \le y) = \mathbb{P}(X_{(n)} \le y) = \left(\frac{y}{1+y}\right)^n$ for $y > 0$. Then, $\lim_{n\to\infty} F_n(y) = 0$ which is not a valid CDF.

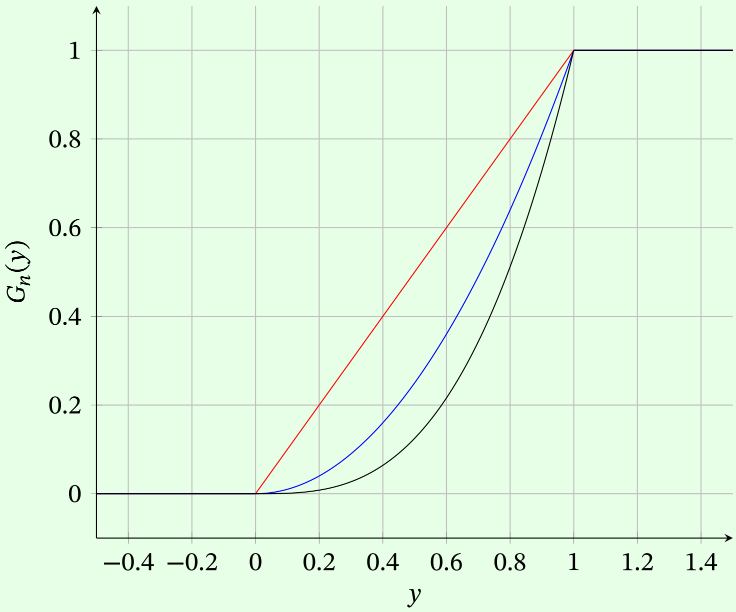
Remark. The limit of a sequence of CDF's is not necessarily a CDF.

Lecture 17, 2025/03/10

Example. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. Let $Y_n = \max(X_1, \dots, X_n) = X_{(n)}$. Then,

$$\begin{aligned}
G_n(y) = \mathbb{P}(Y_n \le y) = \mathbb{P}(X_{(n)} \le y) \\
&= \begin{cases} 0 & \text{if } y \le 0 \\ y^n & \text{if } 0 < y < 1 \\ 1 & \text{if } y \ge 1 \end{cases}
\end{aligned}$$

The plot of CDF looks like the following for $n = 1, 2, 3$.



Then, $\lim_{n\to\infty} G_n(y) = G(y) = \begin{cases} 0 & \text{if } y < 1 \\ 1 & \text{if } y \geq 1. \end{cases}$

6.2 Convergence in Probability

Definition 6.2 (Convergence in Probability).
A sequence of random variables $X_1, \dots, X_n$ **converges in probability** to a random variable X if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0.$$

Equivalently,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \epsilon) = 1.$$

We write $X_n \xrightarrow{P} X$.

Theorem 6.1. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.

Remark. $X_n \xrightarrow{d} X$ does not imply $X_n \xrightarrow{P} X$.

Example (**Converge in D does not imply Converge in P**).

Consider the sample space $\Omega = \{\omega_1, \omega_2\}$ and a probability distribution on Ω defined by $\mathbb{P}(\omega_1) = \frac{1}{2} = \mathbb{P}(\omega_2)$. Define the random variables:

$$X_n(\omega) = \begin{cases} 0 & \text{if } \omega = \omega_1 \\ 1 & \text{if } \omega = \omega_2 \end{cases} \qquad \forall n \in \mathbb{N}$$

and

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \omega_1 \\ 0 & \text{if } \omega = \omega_2 \end{cases}.$$

Then, $X_n \sim \text{Ber}(1/2)$ and $X \sim \text{Ber}(1/2)$. Then, $F_n(x) = F(x) \implies \lim_{n\to\infty} F_n(x) = F(x)$. Thus, $X_n \xrightarrow{d} X$. However,

$$|X_n - X| = \begin{cases} 1 & \text{if } \omega = \omega_1 \\ 1 & \text{if } \omega = \omega_2 \end{cases} \qquad \forall n \in \mathbb{N}.$$

Therefore, $\forall\, 0 < \epsilon < 1$, we have $\mathbb{P}(|X_n - X| \geq \epsilon) = 1$, i.e. $X_n \overset{P}{\nrightarrow} X$.

Definition 6.3 (**Convergence in Probability to a Constant**).

A sequence of random variables $X_1, \dots, X_n$ **converges in probability to a constant** b if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - b| \geq \epsilon) = 0$$

or equivalently, $\lim_{n\to\infty} \mathbb{P}(|X_n - b| < \epsilon) = 1$. We write $X_n \xrightarrow{P} b$.

Theorem 6.2 (**Convergence in Probability to a Constant**).

Suppose $X_1, \dots, X_n$ is a sequence of random variables such that X_n has CDF $F_n(x)$. If

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0 & \text{if } x < b \\ 1 & \text{if } x > b \end{cases}$$

then, $X_n \xrightarrow{P} b$ (no mention of what happens at $x = b$, and it is a point of discontinuity).

Proof. For all $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}(|X_n - b| \geq \epsilon) &= \mathbb{P}(X_n - b \geq \epsilon \text{ or } X_n - b \leq -\epsilon) \\ &= \mathbb{P}(X_n \leq b - \epsilon) + \mathbb{P}(X_n \geq b + \epsilon) \\ &= \mathbb{P}(X_n \leq b - \epsilon) + [1 - \mathbb{P}(X_n < b + \epsilon)] .\end{aligned}$$

Then,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|X_n - b| \geq \epsilon) &= \lim_{n\to\infty} \mathbb{P}(X_n \leq b - \epsilon) + \left[1 - \lim_{n\to\infty} \mathbb{P}(X_n < b + \epsilon)\right] \\ &= 0 + 1 - 1 = 0.\end{aligned}$$

Thus, $X_n \xrightarrow{P} b$. □

Example. Let $X_i \sim \text{Exp}(1, \theta)$, $i = 1, 2, \dots$ independently, where 1 is a rate parameter and θ is a shift parameter. Consider $Y_1, \dots, Y_n$ where $Y_n = \min(X_1, \dots, X_n)$. Show that $Y_n \xrightarrow{P} \theta$.

Proof. Since $X_i \sim \text{Exp}(1, \theta)$, we have

$$f_i(x) = e^{-(x-\theta)} \quad x > \theta$$

$$F_i(x) = \begin{cases} 1 - e^{-(x-\theta)} & \text{if } x > \theta \\ 0 & \text{if } x \leq \theta. \end{cases}$$

Let $G_n(y)$ be the CDF of Y_n. Then,

$$\begin{aligned}G_n(y) = \mathbb{P}(Y_n \leq y) &= \mathbb{P}(X_{(1)} \leq y) \\ &= 1 - \mathbb{P}(X_{(1)} > y) \\ &\overset{\text{ind}}{=} 1 - \prod_{i=1}^{n} \mathbb{P}(X_i > y) \\ &= 1 - \left[1 - \left(1 - e^{-(y-\theta)}\right)\right]^n \\ &= 1 - e^{-n(y-\theta)} \quad y > \theta.\end{aligned}$$

Then, $\lim_{n\to\infty} G_n(y) = \begin{cases} 1 & \text{if } y > \theta \\ 0 & \text{if } y \leq \theta. \end{cases}$, i.e. $Y_n \xrightarrow{P} \theta$, by Theorem 6.2. □

6.3 Moment Generating Funcion Technique for Limiting Distributions

Theorem 6.3 (**Limit Theorem for MGFs or Lévy's Continuity Theorem**).
Let $X_1, \dots, X_n$ be a sequence of random variables such that X_n has MGF $M_n(t)$. Let X be a random variable with MGF $M(t)$. If $\exists h > 0$ such that

$$\lim_{n\to\infty} M_n(t) = M(t) \quad \forall t \in (-h, h),$$

then $X_n \xrightarrow{d} X$.

Example. Suppose $X_n \sim$ Binomial(n, p). If $n \to \infty$, $p \to 0$ s.t. $np = \lambda$ for some $\lambda > 0$, find the limiting distribution of X_n.

Proof. Let's consider the MGF of X_n:

$$\begin{aligned} M_{X_n}(t) &= \left(pe^t + 1 - p\right)^n \\ &= \left(\frac{\lambda e^t}{n} + 1 - \frac{\lambda}{n}\right)^n \\ &= \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \to e^{\lambda(e^t-1)} \quad t \in \mathbb{R}, \text{as } n \to \infty. \end{aligned}$$

This is the MGF of Poisson(λ). By Lévy's Continuity Theorem, $X_n \xrightarrow{d} X \sim$ Poisson(λ). □

Example. Suppose $Y_k \sim$ NB(k, p), $k = 1, 2, \dots$. Find the limiting distribution of Y_k as $k \to \infty$, $p \to 1$ s.t. $\frac{kq}{p} = \mu$ remains constant where $q = 1 - p$.

Proof. Since $Y_k \sim$ NB(k, p), we have

$$M_{Y_k}(t) = \left(\frac{p}{1 - qe^t}\right)^k \quad \text{for } t < -\log(q).$$

Note that $p = \frac{k}{k+\mu}$, then,

$$\begin{aligned}
\lim_{k\to\infty} M_{Y_k}(t) &= \lim_{k\to\infty}\left(\frac{p}{1-qe^t}\right)^k \\
&= \lim_{k\to\infty}\left(\frac{\frac{k}{k+\mu}}{1-\left(\frac{\mu}{k+\mu}\right)e^t}\right)^k \\
&= \lim_{k\to\infty}\left(\frac{\frac{k}{k+\mu}}{\frac{k}{k+\mu}+\frac{\mu}{k+\mu}(1-e^t)}\right)^k \\
&= \lim_{k\to\infty}\left(1+\frac{\mu(1-e^t)}{k}\right)^{-k} \to e^{\mu(e^t-1)} \quad t\in\mathbb{R}, \text{as } k\to\infty.
\end{aligned}$$

This is the MGF of Poisson(μ). By Lévy's Continuity Theorem, $Y_k \xrightarrow{d} Y \sim \text{Poisson}(\mu)$. □

Theorem 6.4 (**Weak Law of Large Numbers (WLLN)**).

Suppose $X_1, \dots, X_n$ is a sequence of independent random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Then, $\overline{X}_n \xrightarrow{P} \mu$.

Note. The sequence $X_1, \dots, X_n$ is not necessarily identically distributed.

Proof. We have

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \overset{Markov's}{\leq} \frac{\mathbb{E}\left[|\overline{X}_n - \mu|^2\right]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty.$$

Thus, $\overline{X}_n \xrightarrow{P} \mu$. □

Lecture 18, 2025/03/12

Theorem 6.5 (**Central Limit Theorem (CLT)**).

Suppose $X_1, \dots, X_n$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Then,

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim \text{N}(0,1).$$

Remark.

(1) This means, for large n, the distribution of $\overline{X}_n$ is approximately $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. But we cannot write it as $\overline{X}_n \xrightarrow{d} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ because for the limiting distribution, we have $n \to \infty$, so $\frac{\sigma^2}{n} \to 0$.

(2) $\overline{X}_n \xrightarrow{P} \mu$ with a rate $\frac{1}{\sqrt{n}}$, i.e. $\overline{X}_n$ is $\sqrt{n}$-consistent.

Proof. First, we can write $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)$. Consider the MGF of Z_n:

$$
\begin{aligned}
M_{Z_n}(t) &= \mathbb{E}\left[e^{t\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)\right)}\right] \\
&\overset{iid}{=} \left[M_{\frac{X_i-\mu}{\sigma}}\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 + \frac{\mathbb{E}\left[\frac{X_i-\mu}{\sigma}\right]}{1!} \frac{t}{\sqrt{n}} + \frac{\mathbb{E}\left[\left(\frac{X_i-\mu}{\sigma}\right)^2\right]}{2!} \left(\frac{t}{\sqrt{n}}\right)^2 + \underbrace{o\left(\left(\frac{t}{\sqrt{n}}\right)^2\right)}_{\to 0}\right]^n \quad \text{by Maclaurin series} \\
&= \left[1 + 0 \cdot \frac{t}{\sqrt{n}} + \frac{1}{2} \cdot \frac{t^2}{n} + o\left(\frac{t^2}{n}\right)\right]^n \\
&= \left[1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \to e^{\frac{t^2}{2}} \quad \text{as } n \to \infty.
\end{aligned}
$$

This is the MGF of $Z \sim \mathrm{N}(0,1)$. Thus, $Z_n \xrightarrow{d} Z$ by Lévy's Continuity Theorem. □

Example. Suppose that $Y_n \sim \chi^2(n)$, $n = 1, 2, \ldots$. Show that $Z_n = \frac{Y_n - n}{\sqrt{2n}} \xrightarrow{d} Z \sim \mathrm{N}(0,1)$.

Proof. Note that $\mathbb{E}[Y_n] = n$ and $\mathrm{Var}(Y_n) = 2n$. Let $X_1, \ldots, X_n$ be i.i.d. $\chi^2(1)$ random variables. We have $\mathbb{E}[X_i] = \mu = 1$ and $\mathrm{Var}(X_i) = \sigma^2 = 2$. Also, $Y_n = \sum_{i=1}^n X_i$. Then,

$$
\begin{aligned}
Z_n = \frac{Y_n - n}{\sqrt{2n}} &= \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i - \sqrt{n}}{\sqrt{2}} \\
&= \sqrt{n}\left(\frac{\overline{X}_n - 1}{\sqrt{2}}\right) \\
&= \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sqrt{\sigma}}\right) \xrightarrow{d} Z \sim \mathrm{N}(0,1) \quad \text{by the CLT.}
\end{aligned}
$$

□

Theorem 6.6 (**Additional Limit Theorems**).

(1) (Continuous Mapping Theorem): If $X_n \xrightarrow{P} a$ and g is continuous at $x = a$, then

$$g(X_n) \xrightarrow{P} g(a).$$

(2) (Extension of above): If $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y$ and $g(x, y)$ is continuous at (a, b), then

$$g(X_n, Y_n) \xrightarrow{P} g(a, b).$$

(3) (Slutsky's Theorem): If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} b$, and $g(x, b)$ is continuous for all x, then

$$g(X_n, Y_n) \xrightarrow{d} g(X, b).$$

(4) If $X_n \xrightarrow{d} X$ and $g(x)$ is a continuous function, then

$$g(X_n) \xrightarrow{d} g(X).$$

Proof. (1) Suppose that $X_n \xrightarrow{P} a$ and g is continuous at $x = a$. Then, $\forall \epsilon > 0$, $\exists \delta > 0$ s.t.

$$|x - a| < \delta \implies |g(x) - g(a)| < \epsilon.$$

Thus,

$$\begin{aligned} &\mathbb{P}\left(|g(X_n) - g(a)| < \epsilon\right) \geq \mathbb{P}\left(|X_n - a| < \delta\right) \\ \implies &\lim_{n\to\infty} \mathbb{P}\left(|g(X_n) - g(a)| < \epsilon\right) \geq \lim_{n\to\infty} \mathbb{P}\left(|X_n - a| < \delta\right) = 1. \end{aligned}$$

Thus, $g(X_n) \xrightarrow{P} g(a)$. □

Example. If $X_n \xrightarrow{P} a > 0$, $Y_n \xrightarrow{P} b \neq 0$, and $Z_n \xrightarrow{d} Z \sim \mathrm{N}(0, 1)$, then confirm the following limiting distributions.

(1) $X_n^2 \xrightarrow{P} a^2$ (6) $2Z_n \xrightarrow{d} 2Z \sim N(0,4)$

(2) $\sqrt{X_n} \xrightarrow{P} \sqrt{a}$ (7) $Z_n + Y_n \xrightarrow{d} Z + b \sim N(b,1)$ (by Slutsky)

(3) $X_nY_n \xrightarrow{P} ab$ (8) $X_nZ_n \xrightarrow{d} aZ \sim N(0,a^2)$ (by Slutsky)

(4) $X_n + Y_n \xrightarrow{P} a + b$ (9) $Z_n^2 \xrightarrow{d} Z^2 \sim \chi^2(1)$

(5) $\frac{X_n}{Y_n} \xrightarrow{P} \frac{a}{b}$ (10) $\frac{1}{Z_n} \xrightarrow{d} \frac{1}{Z}$ (1 over the normal distribution).

Lecture 19, 2025/03/17

Theorem 6.7 (**Delta Method**).
Let $X_1, \ldots, X_n, \ldots$ be a sequence of random variables such that

$$\sqrt{n}(X_n - a) \xrightarrow{d} X \sim \mathrm{N}(0, \sigma^2).$$

Suppose that the function $g(x)$ is differentiable and $g'(a) \neq 0$. Then,

$$\sqrt{n}(g(X_n) - g(a)) \xrightarrow{d} W \sim \mathrm{N}(0, [g'(a)]^2\sigma^2).$$

Proof. By Taylor's theorem, we have $g(X_n) = g(a) + g'(\theta_n)(X_n - a)$ where θ_n is some value between X_n and a. Then,

$$g(X_n) - g(a) = g'(\theta_n)(X_n - a)$$
$$\sqrt{n}(g(X_n) - g(a)) = g'(\theta_n)\sqrt{n}(X_n - a).$$

Since $X_n \xrightarrow{P} a$ and $\theta_n \xrightarrow{P} a$, we have

$$g'(\theta_n) \xrightarrow{P} g'(a) \quad \text{by the Continuous Mapping Theorem.}$$

Then, we have $\sqrt{n}(g(X_n) - g(a)) \xrightarrow{d} g'(a)X \sim \mathrm{N}(0, [g'(a)]^2\sigma^2)$ by Slutsky's Theorem. □

Theorem 6.8. Let $X_1, \ldots, X_n, \ldots$ be a sequence of random variables such that

$$n^b(X_n - a) \xrightarrow{d} X \quad \text{for some } b > 0.$$

Suppose that the function $g(x)$ is differentiable at a and $g'(a) \neq 0$. Then,

$$n^b(g(X_n) - g(a)) \xrightarrow{d} g'(a)X.$$

Example. Suppose that $X_i \sim \text{Poi}(\mu)$, $i = 1, 2 \ldots$ independently. Consider the sequence of random variables $Z_1, \ldots, Z_n \ldots$ where $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\overline{X}_n}}$. Find the limiting distribution of Z_n.

Proof. Note that $\mathbb{E}[X_i] = \mu = \text{Var}(X_i)$. Then by WLLN, we have $\overline{X}_n \xrightarrow{P} \mu$. Also, by the CLT, we have $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\mu}} \xrightarrow{d} Z \sim \text{N}(0, 1)$. Then, $\frac{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\mu}}}{\sqrt{\overline{X}_n/\mu}} \xrightarrow{d} Z \sim \text{N}(0, 1)$ by the Slutsky Theorem. $\square$

Example. Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu \neq 0$, $\text{Var}(X_i) = \sigma^2 < \infty$ and $\mathbb{E}\left[X_i^4\right] < \infty$. Show that

(1) $S_n^2 \xrightarrow{P} \sigma^2$ where $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$.

(2) $T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} \text{N}(0, 1)$.

(3) Find the limiting distribution of $\sqrt{n}\frac{(\overline{X}_n^2 - \mu^2)}{S_n}$.

Proof.

(1) Note that $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{n}{n-1}\overline{X}_n^2$. From WLLN, we have

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \mathbb{E}\left[X_i^2\right] = \sigma^2 + \mu^2 \quad \text{and} \quad \overline{X}_n \xrightarrow{P} \mu.$$

From the Continuous Mapping Theorem, we have, $\overline{X}_n^2 \xrightarrow{P} \mu^2$. Therefore,

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) - \frac{n}{n-1}\overline{X}_n^2 \xrightarrow{P} \mathbb{E}\left[X_i^2\right] - \mu^2 = \sigma^2.$$

Also, we have $S_n \xrightarrow{P} \sigma$.

(2) Note that

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} = \frac{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}{\frac{S_n}{\sigma}} \xrightarrow{d} \mathrm{N}(0,1)$$

since the denominator $\frac{S_n}{\sigma} \xrightarrow{P} 1$. By Slutsky's Theorem, $T_n \xrightarrow{d} \mathrm{N}(0,1)$.

(3) Since $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1)$ by the CLT, then by the Delta Method with $g(x) = x^2$, we have

$$\sqrt{n}\frac{(\overline{X}_n^2 - \mu^2)}{\sigma} \xrightarrow{d} 2\mu\mathrm{N}(0,1) = \mathrm{N}(0,4\mu^2).$$

Therefore, $\sqrt{n}\frac{(\overline{X}_n^2 - \mu^2)}{S_n} = \frac{\sqrt{n}\frac{(\overline{X}_n^2 - \mu^2)}{\sigma}}{\frac{S_n}{\sigma}} \xrightarrow{d} \mathrm{N}(0,4\mu^2)$ from Slutsky's Theorem.

□

Example. Let $X_n \sim \text{Binomial}(n,p)$. Show that $Z_n = \frac{\sqrt{n}\left(\frac{X_n}{n} - p\right)}{\sqrt{\frac{X_n}{n}\left(1 - \frac{X_n}{n}\right)}} \xrightarrow{d} \mathrm{N}(0,1)$.

Proof. Note that $X_n = \sum_{i=1}^n Y_i$ where $Y_1, \dots, Y_n \overset{iid}{\sim} \mathrm{Ber}(p)$. By WLLN, we have

$$\frac{X_n}{n} = \overline{Y}_n \xrightarrow{P} \mathbb{E}[Y_i] = p.$$

By Continuous Mapping Theorem, we have

$$\sqrt{\frac{X_n}{n}\left(1 - \frac{X_n}{n}\right)} \xrightarrow{P} \sqrt{p(1-p)}.$$

Since $\mathrm{Var}(Y_i) = p(1-p)$, then by CLT, we have

$$\frac{\sqrt{n}\left(\frac{X_n}{n} - p\right)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathrm{N}(0,1).$$

Therefore,

$$Z_n = \frac{\sqrt{n}\left(\frac{X_n}{n} - p\right)}{\sqrt{\frac{X_n}{n}\left(1 - \frac{X_n}{n}\right)}} = \frac{\frac{\sqrt{n}\left(\frac{X_n}{n} - p\right)}{\sqrt{p(1-p)}}}{\frac{\sqrt{\frac{X_n}{n}\left(1-\frac{X_n}{n}\right)}}{\sqrt{p(1-p)}}} \xrightarrow{d} \mathrm{N}(0, 1)$$

by Slutsky's Theorem. □

7 Estimation

7.1 Likelihood Function and MLE

Basic Setup

Suppose that $X = (X_1, \dots, X_n)$ are i.i.d. random variables (an i.i.d. sample) from the distribution with PMF/PDF $f(x; \theta)$. Suppose also θ is unknown and $\theta \in \Omega$ where Ω is the set of all possible values of θ, i.e. the parameter space.

Example. Let $X_i \sim \mathrm{N}(\mu, \sigma^2)$. Then $\theta = (\mu, \sigma^2)$ and $\Omega = (-\infty, \infty) \times [0, \infty)$.

We are interested in making inference about θ:

(1) Find estimates (point and interval) of θ.

(2) Test hypothesis about θ.

Definition 7.1 (Statistic). A **statistic** $T = T(X) = T(X_1, \dots, X_n)$ is a function of the data, which does not depend on any unknown parameters.

Example. Suppose $X_1, \dots, X_n$ are i.i.d. with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$. Then,

- $\overline{X}_n = \frac{\sum X_i}{n}$ is a statistic.
- $S^2 = \frac{\sum (X_i - \overline{X}_n)^2}{n-1}$ is a statistic.
- $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is NOT a statistic.

Definition 7.2 (Estimator and Estimate).
A statistic $T = T(X) = T(X_1, \dots, X_n)$ that is used to estimate $\tau(\theta)$, a function of θ, is called an **estimator** of $\tau(\theta)$, and an observed value of T, i.e. $t = t(x) = t(x_1, \dots, x_n)$ is called an **estimate** of $\tau(\theta)$.

Example. $\overline{X}_n$ is an estimator of μ and for a given set of observations $x_1, \dots, x_n$, the number $\overline{x}$ is an estimate of μ.

Definition 7.3 (**Likelihood Function**).

Suppose that X is a discrete random variable with PMF $f(x;\theta)$, where θ is a scalar and $\theta \in \Omega$. If x is the observed data, then the **likelihood function** for θ based on x is

$$L(\theta) = L(\theta; x) = \mathbb{P}(X = x; \theta) = f(x;\theta) \quad \text{for } \theta \in \Omega.$$

Suppose $X_1, \dots, X_n$ is an i.i.d. sample with PMF/PDF $f(x;\theta)$ and $x_1, \dots, x_n$ are the observed data. The **likelihood function** for θ based on $x_1, \dots, x_n$ is

$$\begin{aligned} L(\theta) &= L(\theta; x_1, \dots, x_n) \\ &= \mathbb{P}(\text{observing the data } x_1, \dots, x_n; \theta) \\ &= \mathbb{P}(X_1 = x_1, \dots, X_n = x_n; \theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta) \quad \text{for } \theta \in \Omega. \end{aligned}$$

Definition 7.4 (**log-likelihood Function**).

The **log-likelihood function** is defined as

$$\ell(\theta) = \log L(\theta) = \log L(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \log f(x_i;\theta) \quad \text{for } \theta \in \Omega.$$

Definition 7.5 (**Maximum Likelihood Estimate/Estimator**).

The value of θ that maximizes $L(\theta)$ or $\ell(\theta)$ is called the **maximum likelihood estimate** (**MLE**) of θ, denoted by

$$\hat{\theta} = \hat{\theta}(x).$$

The corresponding **ML estimator** is denoted by

$$\hat{\theta}_n = \hat{\theta}_n(X).$$

Remark. In the absense of any other information, it seems logical that we should estimate θ using a value most compatible with the data.

Example. Let $X \sim \text{Binomial}(n, \theta)$. Then,

$$
\begin{aligned}
L(\theta) &= \mathbb{P}(\text{observing x successes in } n \text{ Bernoulli trials}; \theta) \\
&= \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } 0 \le \theta \le 1 \\
\ell(\theta) &= \log n! - \log(n-x)! - \log x! + x \log \theta + (n-x) \log(1-\theta) \\
\ell'(\theta) &= \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \implies \hat{\theta} = \frac{x}{n}.
\end{aligned}
$$

First derivative test: note that

$$
\begin{aligned}
\ell'(\theta) > 0 \quad & \text{if } 0 < \theta < \frac{x}{n} \\
\ell'(\theta) < 0 \quad & \text{if } \frac{x}{n} < \theta < 1.
\end{aligned}
$$

Therefore, $\hat{\theta} = \frac{x}{n}$ is the MLE of θ.

Example. Suppose that we collected data $x_1, \dots, x_n$ which we believe they are independent observations from a Poisson(θ) distribution. Then,

$$
\begin{aligned}
L(\theta) &= \mathbb{P}(\text{observing } (x_1, \dots, x_n); \theta) \\
&= \prod_{i=1}^{n} f(x_1; \theta) \\
&= \frac{e^{-n\theta} \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \quad \text{for } \theta > 0 \\
\ell(\theta) &= -n\theta + \sum_{i=1}^{n} x_i \log(\theta) - \sum_{i=1}^{n} \log(x_i!) \\
\ell'(\theta) &= -n + \frac{\sum_{i=1}^n x_i}{\theta} = 0 \implies \hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}.
\end{aligned}
$$

First derivative test:

$$
\begin{aligned}
\ell'(\theta) > 0 \quad & \text{if } 0 < \theta < \hat{\theta} \\
\ell'(\theta) < 0 \quad & \text{if } \theta > \hat{\theta}.
\end{aligned}
$$

Thus, $\hat{\theta} = \overline{x}$ is the MLE of θ.

7.2 Score Function and Information Function

Definition 7.6 (Score Function and Information Function).

The **score function** is defined as

$$S(\theta) = S(\theta; x) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log L(\theta) \quad \text{for } \theta \in \Omega.$$

The **information function** is defined as

$$I(\theta) = I(\theta; x) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell(\theta) = -\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log L(\theta) = -\frac{\mathrm{d}}{\mathrm{d}\theta}S(\theta) \quad \text{for } \theta \in \Omega.$$

Note that $S(\hat{\theta}) = 0$ and $I(\hat{\theta})$ is called the **observed information**.

Remark. $I(\theta)$ tells us about the concavity of $\ell(\theta)$.

Second derivative test:

- If $\ell''(\theta) = -I(\theta) < 0$ for all $\theta \in \Omega$, then $\ell(\theta)$ is concave down $\implies$ $\hat{\theta}$ is the global maximum.
- If $\ell''(\theta) = -I(\theta) > 0$ for all $\theta \in \Omega$, then $\ell(\theta)$ is concave up $\implies$ $\hat{\theta}$ is the global minimum.

Definition 7.7 (Expected/Fisher Information).

If θ is a scalar, then the **expected or Fisher Information** function is given by

$$J(\theta) = \mathbb{E}\left[I(\theta; X)\right] = \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell(\theta; X)\right] \quad \text{for } \theta \in \Omega.$$

Remark. If $X_1, \ldots, X_n$ is a random sample from $f(x; \theta)$, then

$$J(\theta) = \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell(\theta; X)\right] = n\mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X; \theta)\right]$$

Example. Find the Fisher Information of the distributions in the previous two examples and compare it with the variance of the ML estimator of θ.

(1) For $X \sim \text{Binomial}(n, \theta)$. Recall that $\ell'(\theta) = \frac{x - n\theta}{\theta(1-\theta)}$. Then,

$$
\begin{aligned}
\ell''(\theta) &= \frac{-n\theta(1-\theta) - (x - n\theta)(1-2\theta)}{\theta^2(1-\theta)^2} \\
\implies J(\theta) &= \mathbb{E}\left[-\ell''(\theta; X)\right] = -\frac{-n\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.
\end{aligned}
$$

Note that $\text{Var}\left(\hat{\theta}_n\right) = \text{Var}\left(\frac{X}{n}\right) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} = \frac{1}{J(\theta)}$.

(2) For $X_1, \ldots, X_n \sim \text{Poisson}(\theta)$ independently. Recall that $\ell'(\theta) = -n + \frac{\sum_{i=1}^n x_i}{\theta}$. Then,

$$
\begin{aligned}
\ell''(\theta) &= -\frac{\sum_{i=1}^n x_i}{\theta^2} \\
\implies J(\theta) &= \mathbb{E}\left[-\ell''(\theta; X)\right] = \frac{\mathbb{E}\left[\sum_{i=1}^n x_i\right]}{\theta^2} = \frac{n\theta}{\theta^2} = \frac{n}{\theta}.
\end{aligned}
$$

Note that $\text{Var}\left(\hat{\theta}_n\right) = \text{Var}\left(\overline{X}\right) = \frac{\theta}{n} = \frac{1}{J(\theta)}$.

Example (**Two Special Examples**).

(1) Suppose that $X_1, \ldots, X_n$ is a random sample from $f(x; \theta) = \frac{1}{\theta}$, for $0 \leq x \leq \theta$. Find the MLE of θ.

Proof. Note that

$$
L(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } 0 \leq x_1, x_2, \ldots, x_n \leq \theta \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \frac{1}{\theta^n} & \text{if } 0 \leq x_{(1)} \leq x_{(n)} \leq \theta \\ 0 & \text{otherwise.} \end{cases}
$$

Then, $S(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}(-n \log \theta) = -\frac{n}{\theta} = 0$, which has no solutions. □

Remark. The support set of X depends on θ. Also, $\hat{\theta} = x_{(n)}$ is the MLE of θ.

Lecture 21, 2025/03/24

(2) Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ where $\theta \in \mathbb{R}$ and

$$
f(x; \theta) = \begin{cases} 1 & \text{if } \theta - \frac{1}{2} \leq x \leq \theta + \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}
$$

We have

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \begin{cases} 1 & \text{if } \theta - \frac{1}{2} \le x_1, \dots, x_n \le \theta + \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

$$= \begin{cases} 1 & \text{if } \theta - \frac{1}{2} \le x_{(1)} \le x_{(n)} \le \theta + \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Then, any statistic $\hat{\theta} = t(X_1, \dots, X_n)$ satisfying $x_{(n)} - \frac{1}{2} \le \hat{\theta} \le x_{(1)} + \frac{1}{2}$ is an MLE of θ.

Remark (**Conclusion about Finding MLEs**).

(1) If $X_1, \dots, X_n$ is a random sample from a distribution where the support set does not depend on θ, then we usually find $\hat{\theta}$ by solving $S(\theta) = 0$.

(2) It is important to verify that $\hat{\theta}$ is the value of θ which maximizes $\ell(\theta)$ (need first/second derivative test).

(3) Often $S(\theta) = 0$ must be solved numerically. We can use the Newton's Method.

Example. Suppose $X_1, \dots, X_n \overset{iid}{\sim}$ Weibull$(1, \theta)$ where 1 is ths scale parameter and θ is the shape parameter. When $\theta = 1$, Weibull$(1, 1)$ is Exp(1). We have

$$f(x;\theta) = \left(\frac{\theta}{1}\right)\left(\frac{x}{1}\right)^{\theta-1} e^{-\left(\frac{x}{1}\right)^{\theta}} = \theta x^{\theta-1} e^{-x^{\theta}} \quad \text{for } x > 0 \text{ and } \theta > 0.$$

Then,

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1} e^{-\sum_{i=1}^{n} x_i^{\theta}} \quad \theta > 0$$

$$\ell(\theta) = n\log(\theta) + (\theta - 1)\sum_{i=1}^{n} \log(x_i) - \sum_{i=1}^{n} x_i^{\theta} \quad \theta > 0.$$

$$S(\theta) = \ell'(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) - \sum_{i=1}^{n} x_i^{\theta} \log(x_i).$$

Note that we cannot solve $S(\theta) = 0$ explicitly. We will need to use the Newton's Method (or Newton-Raphson Method).

Theorem 7.1 (**Newton's Method**).

Let $\theta^{(0)}$ be an initial estimate of θ. The estimate $\theta^{(i)}$ can be updated using

$$\theta^{(i+1)} = \theta^{(i)} + \frac{S(\theta^{(i)})}{I(\theta^{(i)})} \quad \text{for } i = 0, 1, 2, \ldots$$

until $\left|\theta^{(i+1)} - \theta^{(i)}\right| < a$, where a is a very small number, say $a = e^{-10}$.

Note. Recall for solving $f(x) = 0$, $x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$.

Example (**Above Continued**).

Back to Weibull$(1, \theta)$, we have

$$\begin{aligned}
\ell''(\theta) &= -\frac{n}{\theta^2} - \sum_{i=1}^{n} x_i^{\theta} \log^2(x_i) \\
\theta^{(i+1)} &= \theta^{(i)} + \frac{S(\theta^{(i)})}{I(\theta^{(i)})} \\
&= \theta^{(i)} + \frac{\frac{n}{\theta^{(i)}} + \sum_{i=1}^{n} \log(x_i) - \sum_{i=1}^{n} x_i^{\theta^{(i)}} \log(x_i)}{\frac{n}{\theta^{(i)2}} + \sum_{i=1}^{n} x_i^{\theta^{(i)}} \log^2(x_i)}.
\end{aligned}$$

Theorem 7.2 (**Invariance of MLE**).

Suppose that $\tau = h(\theta)$ is a one-to-one function θ and $\hat{\theta}$ is the MLE of θ. Then, $\hat{\tau} = h(\hat{\theta})$ is the MLE of τ.

Example. Suppose $X_1, \ldots, X_n$ is from $f(x, \theta) = \theta x^{\theta-1}$ for $0 < x < 1$ and $\theta > 0$. We know that $L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}$ for $\theta > 0$. Then

$$\begin{aligned}
\ell(\theta) &= n \log(\theta) + (\theta - 1) \sum_{i=1}^{n} \log(x_i) \quad \theta > 0 \\
S(\theta) &= \ell'(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i) = 0 \\
\implies \hat{\theta} &= -\frac{n}{\sum_{i=1}^{n} \log(x_i)}.
\end{aligned}$$

Find the MLE of the median τ.

Proof. We can find τ from $\displaystyle\int_0^{\tau} \theta x^{\theta-1}\, dx = \frac{1}{2} \implies \tau = \left(\frac{1}{2}\right)^{\frac{1}{\theta}}$. Thus, the MLE of τ is

$$\hat{\tau}_{\text{MLE}} = \left(\frac{1}{2}\right)^{\frac{1}{\hat{\theta}}}.$$

□

7.3 Limiting Distribution of Maximum Likelihood Estimator

Proposition 7.3 (**Asymptotic Properties of MLE**).

Suppose that $X = (X_1, \dots, X_n)$ be a random sample from $f(x; \theta)$, and $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ be the ML extimator of θ. Then, under regularity conditions:

A0 If $\theta \neq \theta'$, then $f(x; \theta) \neq f(x; \theta')$.

A1 The support of X does not depend on θ.

A2 θ_0, the true unknown value of θ, is an interior point in Ω.

A3 $f(x_i; \theta)$ is twice differentiable as a function of θ.

A4-A6 We have:

(1) $\hat{\theta}_n \xrightarrow{P} \theta_0$.

(2) $[J(\theta_0)]^{\frac{1}{2}} (\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{N}(0, 1)$.

Remark.

1. (1) implies that $\hat{\theta}_n$ is a consistent estimator of θ or an asymptotically unbiased estimator of θ. Note that
   - Asymptotic unbiased estimator: $\hat{\theta}_n \xrightarrow{P} \theta_0$.
2. (2) indicates that the asymptotic variance $\hat{\theta}_n$ is $J^{-1}(\theta_0)$. Note that

$$J(\theta_0) = n\mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(X; \theta)\right]\Big|_{\theta=\theta_0}$$

   And $\hat{\theta}_n$ is $\sqrt{n}$-consistent.

3. By Limit Theorems (CMT, Slutsky), (1) and (2) implies that

$$[J(\hat{\theta}_n)]^{\frac{1}{2}} (\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{N}(0, 1)$$

for sufficiently large n, $\operatorname{Var}\left(\hat{\theta}_n\right) \approx J^{-1}(\hat{\theta}_n)$.

4. By WLLN, we have

$$\frac{1}{n}I(\theta;X) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X_i;\theta) \xrightarrow{P} \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X;\theta)\right] = \frac{1}{n}J(\theta)$$

which implies that $\left[I(\hat{\theta}_n;X)\right]^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{N}(0,1)$. That is, $\operatorname{Var}\left(\hat{\theta}_n\right) \approx I^{-1}(\hat{\theta}_n)$ when n is large.

5. By Delta Method, we have

$$[J(\theta_0)]^{\frac{1}{2}}\left(\tau(\hat{\theta}_n) - \tau(\theta_0)\right) \xrightarrow{d} \tau'(\theta_0)Z$$

i.e. the asymptotic variance of $\tau(\hat{\theta}_n)$ is

$$\frac{(\tau'(\theta_0))^2}{J(\theta_0)} \to \frac{\left(\tau'(\hat{\theta}_n)\right)^2}{J(\hat{\theta}_n)}.$$

Lecture 22, 2025/03/26

From above remarks, we have

- $\operatorname{Var}\left(\hat{\theta}_n\right) \approx J^{-1}(\theta_0) \approx J^{-1}(\hat{\theta}_n) \approx I^{-1}(\hat{\theta}_n)$.
- $\operatorname{Var}\left(\tau(\hat{\theta}_n)\right) \approx \frac{(\tau'(\theta_0))^2}{J(\theta_0)} \approx \frac{\left(\tau'(\hat{\theta}_n)\right)^2}{I(\hat{\theta}_n)}$.

Example. Suppose that $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Weibull}(\theta, 2)$. We have

$$f(x;\theta) = \left(\frac{2}{\theta}\right)\left(\frac{x}{\theta}\right)^{2-1} e^{-\left(\frac{x}{\theta}\right)^2} = \frac{2}{\theta^2}xe^{-\left(\frac{x}{\theta}\right)^2} \quad \text{for } x > 0,\ \theta > 0.$$

Then,

$$L(\theta) = \left(\frac{2}{\theta^2}\right)^n \left(\prod_{i=1}^n x_i\right) e^{-\sum_{i=1}^n \left(\frac{x_i}{\theta}\right)^2} \quad \theta > 0$$

$$\ell(\theta) = n \log 2 - 2n \log \theta + \sum_{i=1}^n \log(x_i) - \sum_{i=1}^n \left(\frac{x_i}{\theta}\right)^2 \quad \theta > 0$$

$$S(\theta) = -\frac{2n}{\theta} + \frac{2\sum_{i=1}^n x_i^2}{\theta^3} = \frac{2\left(\sum_{i=1}^n x_i^2 - n\theta^2\right)}{\theta^3}.$$

Let $S(\theta) = 0 \implies \hat{\theta} = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}$ and $\hat{\theta}_n = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}$.

Exercise: Use the First (or Second) Derivative Test to check that $\hat{\theta}$ is the MLE (global maximum) of θ.

$$\ell''(\theta) = S'(\theta) = -\frac{4n}{\theta^2} - \frac{6\left(\sum_{i=1}^n x_i^2 - n\theta^2\right)}{\theta^4}$$

$$I(\theta) = \frac{4n}{\theta^2} + \frac{6\left(\sum_{i=1}^n x_i^2 - n\theta^2\right)}{\theta^4}$$

$$I(\hat{\theta}_n) = \frac{4n}{\hat{\theta}_n^2} = \frac{4n}{\frac{\sum_{i=1}^n X_i^2}{n}}$$

$$J(\theta) = \mathbb{E}\left[I(\theta; X)\right] = \frac{4n}{\theta^2} + \frac{6\left[n\mathbb{E}\left[X_i^2\right] - n\theta^2\right]}{\theta^4}$$

$$= \frac{4n}{\theta^2} + \frac{6\left(n\theta^2 - n\theta^2\right)}{\theta^4} = \frac{4n}{\theta^2}.$$

Note that $\mathbb{E}\left[X^k\right] = \theta^k \Gamma\left(\frac{k}{2} + 1\right)$ where 2 is the shape parameter. We also have

$$J(\hat{\theta}_n) = \frac{4n}{\hat{\theta}_n^2} = \frac{4n}{\frac{\sum_{i=1}^n X_i^2}{n}}.$$

In this particular case, $I(\hat{\theta}_n) = J(\hat{\theta}_n)$.

(1) Show that $[J(\theta_0)]^{\frac{1}{2}} (\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{N}(0, 1)$.
(2) Show that $\left[I(\hat{\theta}_n)\right]^{\frac{1}{2}} (\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{N}(0, 1)$.

Proof.

(1) By CLT,

$$\frac{\sqrt{n}\left(\frac{\sum X_i^2}{n} - \mathbb{E}\left[X_i^2\right]\right)}{\sqrt{\operatorname{Var}\left(X_i^2\right)}} \xrightarrow{d} Z$$

$$\implies \frac{\sqrt{n}\left(\hat{\theta}_n^2 - \theta_0^2\right)}{\sqrt{\mathbb{E}\left[X_i^4\right] - (\theta_0^2)^2}} \xrightarrow{d} Z$$

$$\implies \frac{\sqrt{n}\left(\hat{\theta}_n^2 - \theta_0^2\right)}{\sqrt{\theta_0^4\Gamma(3) - \theta_0^4}} = \frac{\sqrt{n}\left(\hat{\theta}_n^2 - \theta_0^2\right)}{\theta_0^2} \xrightarrow{d} Z.$$

Then, $[J(\theta_0)]^{\frac{1}{2}}\left(\hat{\theta}_n - \theta_0\right) = \frac{2\sqrt{n}}{\theta_0}\left(\hat{\theta}_n - \theta_0\right)$. By Delta Method with $g(x) = \sqrt{x}$, we have

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\theta_0^2} \xrightarrow{d} \frac{1}{2}\theta_0^{-1}Z$$

$$\implies \frac{2\sqrt{n}}{\theta_0}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z.$$

That is, $[J(\theta_0)]^{\frac{1}{2}}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} Z \sim \mathrm{N}(0,1)$. We also have $\hat{\theta}_n \xrightarrow{P} \theta_0$.

(2) We have

$$LHS = \left(\frac{4n}{\frac{\sum X_i^2}{n}}\right)^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) = \frac{2\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{\sum X_i^2}{n}}}.$$

Since $\frac{2\sqrt{n}(\hat{\theta}_n - \theta_0)}{\theta_0^2} \xrightarrow{d} Z$, we have $\frac{\sum X_i^2}{n} \xrightarrow{P} \mathbb{E}\left[X_i^2\right] = \theta_0^2$ by WLLN. Then,

$$\sqrt{\frac{\sum X_i^2}{n}} \xrightarrow{P} \theta_0 \quad \text{by CMT.}$$

By Slutsky's Theorem, we have $\frac{2\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{\sum X_i^2}{n}}} \xrightarrow{d} Z$ as desired. □

Theorem 7.4 (Cramer-Rao Lower Bound).

Suppose that $X_1, \ldots, X_n$ is a random sample with PDF $f(x;\theta)$. For any unbiased estimator $T(X) = T(X_1, \ldots, X_n)$ of $\tau(\theta)$, i.e. $\mathbb{E}[T(X)] = \tau(\theta)$, under regularity conditions, we have

$$\operatorname{Var}(T(X)) \geq \frac{(\tau'(\theta))^2}{J(\theta)}.$$

In particular, if $\tau(\theta) = \theta$, then $\operatorname{Var}(T(X)) \geq \dfrac{1}{J(\theta)}$.

Remark.

(1) There is a lower bound on the variance of any unbiased estimator of $\tau(\theta)$, which we can call it Cramer-Rao lower bound.

(2) $\tau(\hat{\theta}_{\text{MLE}})$ has asymptotically the smallest variance among all asymptotically unbiased estimators. This explains the popularity of MLEs.

Proof. Let $S(\theta;X) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta;X)$. We can prove that $\mathbb{E}[S(\theta;X)] = 0$ and $\operatorname{Var}(S(\theta;X)) = J(\theta)$. Note that

$$\tau(\theta) = \mathbb{E}[T(X)] = \int \cdots \int T(x)\, L(\theta;x)\, dx_1 \cdots dx_n$$

Then,

$$\begin{aligned}
\tau'(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta} \int \cdots \int T(x)\, L(\theta;x)\, dx_1 \cdots dx_n \\
&= \int \cdots \int T(x) \frac{\mathrm{d}L(\theta;x)}{\mathrm{d}\theta}\, dx_1 \cdots dx_n \\
&= \int \cdots \int T(x) \underbrace{\frac{\mathrm{d}\log L(\theta;x)}{\mathrm{d}\theta}}_{S(\theta;x)} L(\theta;x)\, dx_1 \cdots dx_n \\
&= \mathbb{E}[T(X)S(\theta;X)] \\
&= \operatorname{Cov}(T(X), S(\theta;X)).
\end{aligned}$$

Since $-1 \leq \frac{\operatorname{Cov}(T(X), S(\theta;X))}{\sqrt{\operatorname{Var}(T(X))}\sqrt{\operatorname{Var}(S(\theta;X))}} \leq 1$, we have

$$\begin{aligned}
&[\operatorname{Cov}(T(X), S(\theta;X))]^2 \leq \operatorname{Var}(T(X)) \operatorname{Var}(S(\theta;X)) \\
&\implies \operatorname{Var}(T(X)) \geq \frac{[\operatorname{Cov}(T(X), S(\theta;X))]^2}{\operatorname{Var}(S(\theta;X))} = \frac{(\tau'(\theta))^2}{J(\theta)}.
\end{aligned}$$

□

7.4 Confidence Intervals and Pivotal Quantities

Definition 7.8 (**Confidence Interval, Confidence Coefficient, Pivotal Quantity**).
Suppose that $L(X)$ and $U(X)$ are both statistics. If $\mathbb{P}(L(X) \leq \theta \leq U(X)) = p$ with $0 < p < 1$, then $[L(X), U(X)]$ is called a $100p\%$ **confidence interval** for θ and p is called the **confidence coefficient**.
The random variable $Q(X, \theta)$ is called a **pivotal quantity** if the distribution of Q does not depend on θ.

Example. Suppose that $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, 1^2)$. Then,

- $\overline{X}_n = \frac{\sum X_i}{n} \sim \mathrm{N}(\mu, \frac{1}{n})$ is not a pivotal quantity.
- $\sqrt{n}(\overline{X}_n - \mu) \sim \mathrm{N}(0, 1)$ is a pivotal quantity.

To construct a 95% CI for μ:

$$\begin{aligned}
\mathbb{P}(Z_{0.025} \leq \sqrt{n}(\overline{X}_n - \mu) \leq Z_{0.975}) &= 0.95 \\
\mathbb{P}(-1.96 \leq \sqrt{n}(\overline{X}_n - \mu) \leq 1.96) &= 0.95 \\
\mathbb{P}\left(\overline{X}_n - \frac{1.96}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{1.96}{\sqrt{n}}\right) &= 0.95.
\end{aligned}$$

Therefore, the 95% CI for μ is $\left[\overline{x} - \frac{1.96}{\sqrt{n}}, \overline{x} + \frac{1.96}{\sqrt{n}}\right]$.

Lecture 23, 2025/03/31

Example. Suppose that $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Exp}(\theta)$, where θ is a scale parameter.

(1) Show that $\hat{\theta}_n = \overline{X}_n$ and $Q = \frac{\hat{\theta}_n \cdot 2n}{\theta}$ is a pivotal quantity.

(2) For the data $n = 15$, and $\sum_{i=1}^{15} x_i = 36$. Find the 95% equal-tailed CI for θ.

Proof.

(1) Note that $f(x_i) = \frac{1}{\theta} e^{-\frac{x_i}{\theta}}$ for $x_i > 0$. Also note that $\mathrm{Exp}(\theta) = \mathrm{Gamma}(1, \theta)$. Then,

$$M_{X_i}(t) = \frac{1}{1 - \theta t} \quad \text{for } t < \frac{1}{\theta}.$$

Then, $Q = \frac{2n\overline{X}_n}{\theta} = \frac{2\sum_{i=1}^n X_i}{\theta}$. We have

$$M_Q(t) = \prod_{i=1}^n M_{X_i}\left(\frac{2t}{\theta}\right) = \left(\frac{1}{1-2t}\right)^n \quad \text{for } t < \frac{1}{2}$$

which is the MGF of Gamma$(n, 2) = \chi^2(2n)$. Thus, $Q \sim \chi^2(2n)$.

(2) We have $\mathbb{P}\left(q_1 \leq \frac{2n\overline{X}_n}{\theta} \leq q_2\right) = 0.95$ where

$$q_1 = \chi^2_{0.025}(30) = 16.79$$
$$q_2 = \chi^2_{0.975}(30) = 46.98.$$

Then, $\mathbb{P}\left(\frac{2n\overline{X}_n}{q_2} \leq \theta \leq \frac{2n\overline{X}_n}{q_1}\right) = 0.95$. The 95% equal-tailed CI for θ is

$$\left[\frac{72}{q_2}, \frac{72}{q_1}\right] = [1.53, 4.29]\,.$$

□

Definition 7.9 (**Asymptotic Pivotal Quantity**).

$Q(X;\theta)$ is called an **asymptotic pivotal quantity** if the limiting distribution of Q as $n \to \infty$ does not depend on θ.

Example. Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poi}(\theta)$. Note that $\mathbb{E}[X_i] = \text{Var}(X_i) = \theta$. Show that $\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\overline{X}_n}}$ is an asymptotic pivotal quantity and find the approximate 95% CI for θ.

Proof. By CLT, we have $\frac{\sqrt{n}(\overline{X}_n-\theta)}{\sqrt{\theta}} \xrightarrow{d} Z \sim \text{N}(0,1)$. By WLLN, $\overline{X}_n \xrightarrow{P} \theta$. Then, we have

$$\frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\overline{X}_n}} = \frac{\frac{\sqrt{n}(\overline{X}_n-\theta)}{\sqrt{\theta}}}{\sqrt{\frac{\overline{X}_n}{\theta}}} \xrightarrow{d} Z \sim \text{N}(0,1) \quad \text{by Slutsky's Theorem.}$$

Note that $\sqrt{\frac{\overline{X}_n}{\theta}} \xrightarrow{P} 1$ by CMT and that the limiting distribution does not depend on θ. To find the approximate 95% CI for θ:

$$\mathbb{P}\left(-1.96 \le \frac{\sqrt{n}(\overline{X}_n - \theta)}{\sqrt{\overline{X}_n}} \le 1.96\right) = 0.95$$

$$\mathbb{P}\left(-1.96\sqrt{\overline{X}_n} \le \sqrt{n}(\overline{X}_n - \theta) \le 1.96\sqrt{\overline{X}_n}\right) = 0.95$$

$$\mathbb{P}\left(\overline{X}_n - \frac{1.96}{\sqrt{n}}\sqrt{\overline{X}_n} \le \theta \le \overline{X}_n + \frac{1.96}{\sqrt{n}}\sqrt{\overline{X}_n}\right) = 0.95.$$

Therefore, the approximate 95% CI for θ is $\overline{X}_n \pm 1.96\sqrt{\frac{\overline{X}_n}{n}}$.

Then, for the data $n = 30$, $\sum_{i=1}^{30} x_i = 36$. We have

$$\begin{aligned}\overline{X}_n \pm 1.96\sqrt{\frac{\overline{X}_n}{n}} &= 1.2 \pm 1.96\sqrt{\frac{1.2}{30}} \\ &= [0.808, 1.592].\end{aligned}$$

□

Remark. Since under regularity conditions, we have

$$\left[J(\hat{\theta}_n)\right]^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{N}(0,1)$$

then, $\left[J(\hat{\theta}_n)\right]^{\frac{1}{2}}(\hat{\theta}_n - \theta_0)$ is an asymptotic pivotal quantity. For an approximate $100p\%$ CI for θ, we have

$$\mathbb{P}\left(-a \le \left[J(\hat{\theta}_n)\right]^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \le a\right) = p \quad \text{where } a = Z_{\frac{1+p}{2}}$$

$$\mathbb{P}\left(\hat{\theta}_n - a\left[J(\hat{\theta}_n)\right]^{-\frac{1}{2}} \le \theta_0 \le \hat{\theta}_n + a\left[J(\hat{\theta}_n)\right]^{-\frac{1}{2}}\right) = p.$$

Then, the approximate $100p\%$ CI for θ is $\left[\hat{\theta}_n - a\left[J(\hat{\theta}_n)\right]^{-\frac{1}{2}}, \hat{\theta}_n + a\left[J(\hat{\theta}_n)\right]^{-\frac{1}{2}}\right]$.

Similarly, $\left[I(\hat{\theta}_n; X)\right]^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{N}(0,1)$ is an asymptotic pivotal quantity. An approximate $100p\%$ CI for θ is $\left[\hat{\theta}_n - a\left[I(\hat{\theta}_n)\right]^{-\frac{1}{2}}, \hat{\theta}_n + a\left[I(\hat{\theta}_n)\right]^{-\frac{1}{2}}\right]$.

Example. Let $X \sim \text{Bin}(n, \theta)$. Then, $\hat{\theta}_n = \frac{X}{n}$, $J(\theta) = \frac{n}{\theta(1-\theta)}$ and $J(\hat{\theta}_n) = I(\hat{\theta}_n) = \frac{n}{\hat{\theta}_n(1-\hat{\theta}_n)}$. Then, the asymptotic pivotal quantity is

$$[J(\hat{\theta}_n)]^{\frac{1}{2}} (\hat{\theta}_n - \theta) = \sqrt{\frac{n}{\hat{\theta}_n(1-\hat{\theta}_n)}} (\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \text{N}(0,1).$$

An approximate 95% CI is

$$\mathbb{P}\left(-1.96 \le \sqrt{\frac{n}{\hat{\theta}_n(1-\hat{\theta}_n)}} (\hat{\theta}_n - \theta_0) \le 1.96\right) = 0.95$$
$$\implies \hat{\theta}_n \pm 1.96\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}.$$

For example, for $n = 50$, $x = 20$. The 95% CI is

$$0.4 \pm 1.96\sqrt{\frac{0.4 \cdot 0.6}{50}} = [0.2642, 0.5358].$$

Example. Suppose $X_1, \ldots, X_n$ is a random sample with CDF

$$F(x) = 1 - \left(\frac{\theta}{x}\right)^2 \quad \text{for } x \ge \theta > 0.$$

Let $Y_n = n\left(\frac{X_{(1)}}{\theta} - 1\right)$.

(1) Show that $Y_n \xrightarrow{d} Y$ with $G(y) = 1 - e^{-2y}$ for $y > 0$.

(2) Construct an approximate 90% equal-tail CI for θ when $n = 30$ and $X_{(1)} = 0.4$.

Proof.

(1) Let's find the CDF of Y_n:

$$\begin{aligned}
G_n(y) = \mathbb{P}(Y_n \le y) &= \mathbb{P}\left(n\left(\frac{X_{(1)}}{\theta} - 1\right) \le y\right) \\
&= \mathbb{P}\left(X_{(1)} \le \frac{y}{n} + \theta\right) = \mathbb{P}\left(X_{(1)} \le \frac{y\theta}{n} + \theta\right) \\
&= 1 - \mathbb{P}\left(X_{(1)} > \frac{y\theta}{n} + \theta\right) \\
&= 1 - \left(\frac{\theta}{\frac{y\theta}{n} + \theta}\right)^{2n} = 1 - \left(1 + \frac{y}{n}\right)^{-2n} \to 1 - e^{-2y}, \quad y > 0
\end{aligned}$$

as $n \to \infty$. So, $Y_n = n\left(\frac{X_{(1)}}{\theta} - 1\right)$ is an asymptotic pivotal quantity.

(2) Note that

$$\mathbb{P}\left(q_1 \le n\left(\frac{X_{(1)}}{\theta} - 1\right) \le q_2\right) = 0.9$$

where q_1 is the 5th percentile of Y and q_2 is the 95th percentile of Y. Then, we have

$$\begin{aligned}
1 - e^{-2q_1} = 0.05 \quad &\text{and} \quad 1 - e^{-2q_2} = 0.95 \\
\Longrightarrow \quad q_1 = 0.0256 \quad &\text{and} \quad q_2 = 1.4979.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\frac{X_{(1)}n}{q_2 + n} \le \theta \le \frac{X_{(1)}n}{q_1 + n}\right) = 0.9.$$

The 90% CI for θ is $[0.381, 0.3997]$. □

Lecture 24, 2025/04/02

7.5 Maximum Likelihood Method for Multiparameter Cases

Definition 7.10 (Log-likelihood Function).

$\ell(\theta) = \ell(\theta_1, \dots, \theta_k) = \log L(\theta_1, \dots, \theta_k; x_1, \dots, x_n)$.

Definition 7.11 (**Maximum Likelihood Estimate**).
$\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ which maximizes $L(\theta)$ or $\ell(\theta)$ is called the **maximum likelihood estimate** (**MLE**) of $\theta = (\theta_1, \dots, \theta_k)$.

Note. $\hat{\theta}$ is found by solving $\frac{\partial \ell}{\partial \theta_j} = 0$ for $j = 1, \dots, k$ simultaneously.

Question: How do we check if $\hat{\theta}$ is a global max point?

Answer: Check that $\left[\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right]_{k\times k}$ is negative definite at $\hat{\theta}$.

Definition 7.12 (**Score Vector**).
$S(\theta) = S(\theta; x) = \left(\frac{\partial \ell}{\partial \theta_1}, \dots, \frac{\partial \ell}{\partial \theta_k}\right)^\top$ for $\theta \in \Omega$.

Note. $\hat{\theta}$ is found by solving $S(\theta) = 0$.

Definition 7.13 (**Information Matrix, Observed Information**).
$I(\theta) = I(\theta; x) = \left[-\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right]_{k\times k}$ for $\theta \in \Omega$. Also, $I(\hat{\theta})$ is the **observed information**.

Theorem 7.5 (**Newton's Method**).
$\theta^{(i+1)} = \theta^{(i)} + \left[I(\theta^{(i)})\right]^{-1} S(\theta^{(i)})$ for $i = 0, 1, \dots$ until $\left\|\theta^{(i+1)} - \theta^{(i)}\right\|_2 < a$, say $a = 10^{-8}$.

Definition 7.14 (**Expected/Fisher Information Matrix**).
$J(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j}\ell(\theta; X)\right]_{k\times k}$ for $\theta \in \Omega$.

Example. Suppose that $X_1, \dots, X_n$ is a random sample form $\mathrm{N}(\mu, \sigma^2)$. Note that $\theta = (\mu, \sigma^2)$ and $\Omega = (-\infty, \infty) \times [0, \infty)$. Then, the likelihood function is

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

Then,

$$\ell(\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\big(\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

The MLE is found by solving

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0\\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0.\end{aligned}$$

We have

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}\\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n} = \frac{n-1}{n}S^2\end{aligned}$$

where $S^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$ is the sample variance.
For the information matrix, we have

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\sigma^2}, \quad \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = -\frac{\sum_{i=1}^{n}(x_i-\mu)}{(\sigma^2)^2}, \quad \frac{\partial^2 \ell}{(\partial \sigma^2)^2} = \frac{n}{2(\sigma^4)} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Then,

$$\begin{aligned}I(\mu,\sigma) &= -\begin{bmatrix}\frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2}\\ \frac{\partial^2 \ell}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \ell}{\partial (\sigma^2)^2}\end{bmatrix} = \begin{bmatrix}\frac{n}{\sigma^2} & \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^4}\\ \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^4} & -\frac{n}{2(\sigma^4)} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2\end{bmatrix}\\ \implies I(\hat{\mu},\hat{\sigma}) &= \begin{bmatrix}\frac{n}{\hat{\sigma}^2} & \frac{\overbrace{\sum_{i=1}^{n}(x_i-\hat{\mu})}^{0}}{\sigma^4}\\ \frac{\sum_{i=1}^{n}(x_i-\overline{x})}{\sigma^4} & -\frac{n}{2(\hat{\sigma}^4)} + \frac{1}{\hat{\sigma}^6}\underbrace{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}_{n\hat{\sigma}^2}\end{bmatrix} = \begin{bmatrix}\frac{n}{\hat{\sigma}^2} & 0\\ 0 & \frac{n}{2(\hat{\sigma}^4)}\end{bmatrix}.\end{aligned}$$

The Fisher information matrix is

$$
\begin{aligned}
J(\mu,\sigma) &= \mathbb{E}\left[I(\mu,\sigma;X)\right] \\
&= \begin{bmatrix} \frac{n}{\sigma^2} & \frac{\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)\right]}{\sigma^4} \\ \frac{\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)\right]}{\sigma^4} & -\frac{n}{2\sigma^4}+\frac{1}{\sigma^6}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \\
\implies J^{-1}(\mu,\sigma) &= \begin{bmatrix} \overbrace{\frac{\sigma^2}{n}}^{\widehat{\mathrm{Var}}(\hat{\mu})} & \overbrace{0}^{\widehat{\mathrm{Cov}}(\hat{\mu},\hat{\sigma}^2)} \\ 0 & \underbrace{\frac{2\sigma^4}{n}}_{\widehat{\mathrm{Var}}(\hat{\sigma})} \end{bmatrix}.
\end{aligned}
$$

Proposition 7.6 (**Asymptotic Properties of MLE - Multiparameter Case**).

Under regularity conditions, we have

(1) $\hat{\theta}_n \xrightarrow{P} \theta_0$, where θ_0 is a $k \times 1$ vector of true but unknown values of θ.

(2) $[J(\theta_0)]^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim \mathrm{MVN}(\vec{0}_k, I_k)$, where $\vec{0}_k$ is a $k \times 1$ vector of zeros and I_k is the $k \times k$ identity matrix.

Remark. Note that (1) means $\forall \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(\left\|\hat{\theta}_n - \theta_0\right\|_2 \geq \epsilon\right) = 0$, where $\|\cdot\|_2$ is the L^2 (Euclidean) norm. For (2), we can replace $J(\theta_0)$ with $J(\hat{\theta}_n)$ or $I(\hat{\theta}_n; X)$. For sufficiently large n, we have

$$
\mathrm{Var}\left(\hat{\theta}_n\right) \approx \underbrace{J^{-1}(\theta_0)}_{\text{asymptotic variance}} \approx J^{-1}(\hat{\theta}_n) \approx I^{-1}(\hat{\theta}_n).
$$

Example. Back to the previous example for $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$.

(1) from above implies that $\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \overline{x} \\ \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n} \end{pmatrix} \xrightarrow{P} \begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix}$.

(2) from above implies that $\begin{bmatrix} \frac{\sqrt{n}}{\sigma_0} & 0 \\ 0 & \frac{\sqrt{n}}{\sqrt{2}\sigma_0^2} \end{bmatrix}\left(\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix}\right) \xrightarrow{d} \mathrm{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$. Equivalently,

$$
\sqrt{n}\left(\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \sigma_0^2 \end{pmatrix}\right) \xrightarrow{d} \begin{bmatrix} \frac{1}{\sigma_0} & 0 \\ 0 & \frac{1}{\sqrt{2}\sigma_0^2} \end{bmatrix}^{-1} \mathrm{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right).
$$

In other words, $\sqrt{n}\left(\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}^2\end{pmatrix}-\begin{pmatrix}\mu_0\\ \sigma_0^2\end{pmatrix}\right)\xrightarrow{d}\text{BVN}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{bmatrix}\sigma_0^2 & 0\\ 0 & 2\sigma_0^4\end{bmatrix}\right)$.

Note (**Final Exam Information**). About 20% for the first three chapters. About 40-50% for joint distribution and limiting distribution. About 30-40% for the last chapter.

Update: about 50% for ch6 and ch7.

END OF STAT 330!