

Lecture Notes 4: Multinomial Distribution

Professor: Zhihua Zhang

2.5 More About Mixture Distribution

Definition 2.1. In probability theory and statistics, the moment-generating function of a random variable X is

$$M_X(t) = \mathbb{E}[e^{tX}] = \int e^{tx} f_X(x) dx$$

One property about moment-generating function is that we can get $\mathbb{E}[X^k]$ from $M_X^{(k)}(0)$, as we can see $M_X^{(k)}(t) = \int x^k e^{tx} f_X(x) dx$, where we assume we can put the derivation inside. So $M_X^{(k)}(0) = \mathbb{E}[X^k]$.

Definition 2.2. A function $f : (0, \infty) \rightarrow \mathbb{R}$ is completely monotone function if and only if f is of class C^∞ (infinitely derivable), and $(-1)^n f^{(n)}(\lambda) \geq 0$ for all $n \in \mathbb{N} \cup \{0\}$, and $\lambda > 0$.

Theorem 2.1. (Bernstein) Let $g : (0, \infty) \rightarrow \mathbb{R}$ be a completely monotone function. Then it is the Laplace transform of an unique measure μ on $[0, \infty]$, i.e. for all $\lambda > 0$,

$$g(\lambda) = \mathcal{L}(\mu; \lambda) = \int_{[0, \infty)} e^{-\lambda t} \mu(dt)$$

. Conversely, whenever $\mathcal{L}(\mu; \lambda) < \infty$ for every $\lambda > 0$, $\lambda \mapsto \mathcal{L}(\mu; \lambda)$ is a completely monotone function.

Proof. Assume $g(0+) = 1$ and $g(+\infty) = 0$. By Taylor's formula

$$\begin{aligned} f(\lambda) &= \sum_{k=0}^{n-1} \frac{f^{(k)}(a)}{k!} (\lambda - a)^k + \int_a^\lambda \frac{f^{(n)}(s)}{(n-1)!} (\lambda - s)^{n-1} ds \\ &= \sum_{k=0}^{n-1} \frac{(-1)^k f^{(k)}(a)}{k!} (a - \lambda)^k + \int_\lambda^a \frac{(-1)^n f^{(n)}(s)}{(n-1)!} (s - \lambda)^{n-1} ds \end{aligned} \quad (1)$$

where $a > 0$ and $n \in \mathbb{N}$. Let $a \rightarrow \infty$, then

$$\begin{aligned} \lim_{a \rightarrow \infty} \int_\lambda^a \frac{(-1)^n f^{(n)}(s)}{(n-1)!} (s - \lambda)^{n-1} ds &= \int_\lambda^\infty \frac{(-1)^n f^{(n)}(s)}{(n-1)!} (s - \lambda)^{n-1} ds \\ &\leq f(\lambda). \end{aligned}$$

So the sum in (1) converges for every $n \in \mathbb{N}$ as $a \rightarrow \infty$. Let

$$\rho_n(\lambda) = \lim_{a \rightarrow \infty} \frac{(-1)^n f^{(n)}(a)}{n!} (a - \lambda)^n$$

. This limit doesn't depend on $\lambda > 0$. Indeed, for $k > 0$,

$$\begin{aligned}\rho_n(k) &= \lim_{a \rightarrow \infty} \frac{(-1)^n f^{(n)}(a)}{n!} (a - k)^n \\ &= \lim_{a \rightarrow \infty} \frac{(-1)^n f^{(n)}(a)}{n!} (a - \lambda)^n \frac{(a - k)^n}{(a - \lambda)^n} \\ &= \rho_n(\lambda).\end{aligned}$$

So we can get

$$f(\lambda) = \sum_{k=0}^{n-1} \rho_k(\lambda) + \int_{\lambda}^{\infty} \frac{(-1)^n f^{(n)}(s)}{(n-1)!} (s - \lambda)^{n-1} ds$$

Let $\lambda \rightarrow \infty$, since $f(+\infty) = 0$, so $\rho_k(\lambda) = 0$. Then we can get

$$f(\lambda) = \int_{\lambda}^{\infty} \frac{(-1)^n f^{(n)}(s)}{(n-1)!} (s - \lambda)^{n-1} ds \quad (2)$$

. And since $f(0+) = 1$, we can get:

$$1 = \lim_{\lambda \rightarrow 0+} f(\lambda) = \int_0^{\infty} \frac{(-1)^n f^{(n)}(s)}{(n-1)!} s^{n-1} ds$$

And (2) can also be written as:

$$f(\lambda) = \int_0^{\infty} \left(1 - \frac{\lambda}{s}\right)_+^{n-1} \frac{(-1)^n f^{(n)}(s)}{(n-1)!} s^{n-1} ds.$$

Let $t = \frac{n}{s}$, then

$$f(\lambda) = \int_0^{\infty} \left(1 - \frac{\lambda t}{n}\right)_+^{n-1} \frac{(-1)^n}{n!} f^{(n)}\left(\frac{n}{t}\right) \left(\frac{n}{t}\right)^{n+1} dt$$

. Since $\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda t}{n}\right)_+^{n-1} = e^{-\lambda t}$. So

$$f(\lambda) = \int_0^{\infty} e^{-\lambda t} \frac{(-1)^n}{n!} f^{(n)}\left(\frac{n}{t}\right) \left(\frac{n}{t}\right)^{n+1} dt.$$

For the converse, let $f(\lambda) = \mathcal{L}(\mu; \lambda) = \int_0^{\infty} e^{-\lambda t} \mu(dt)$. So

$$(-1)^n f^{(n)}(\lambda) = \int_0^{\infty} t^n e^{-\lambda t} \mu(dt) \geq 0$$

□

Corollary Let $g(t)$ be a function that is symmetric about the origin, integrable, convex and twice differentiable on $(0, \infty)$ and $g(0+) = 1$, $g(+\infty) = 0$ then

$$g(t) = \int_0^{\infty} \frac{1}{s} \left(1 - \frac{t}{s}\right)_+ s^2 g''(s) ds$$

Theorem 2.2. A function $f(x)$ can be represented as a Gaussian scale mixture iff $f(\sqrt{x})$ is completely monotone on $(0, \infty)$.

Proof.

Let $g(x) = f(\sqrt{x})$.

$f(\sqrt{x})$ is completely monotone,

$\iff g(x)$ is completely monotone.

By Bernstein :

$$\iff g(x) = \int_0^\infty e^{-xt} \mu(dt)$$

$$\iff f(\sqrt{x}) = \int_0^\infty e^{-xt} \mu(dt)$$

$$\iff f(x) = \int_0^\infty e^{-x^2 t} \mu(dt) = C \int_0^\infty N(x \mid 0, \frac{1}{2t}) \mu(dt), \text{ and } \int_0^\infty \mu(dt) = 1$$

$\iff f(x)$ can be represented as a Gaussian scale mixture.

□

Theorem 2.3. If $f(x) > 0$, then $e^{-uf(x)}$ is completely monotone for every $u > 0$ iff $f'(x)$ is completely monotone.

Proof. If $e^{-uf(x)}$ is completely monotone for every $u > 0$:

$$e^{-\mu f(x)} = \sum_{j=0}^{\infty} \frac{(-1)^j \mu^j}{j!} [f(x)]^j$$

and all of its formal derivatives converge uniformly, so we can calculate $\frac{d^n}{dx^n} e^{-\mu f(x)}$ by termwise differentiation. Since $e^{-\mu f}$ is completely monotone, we have:

$$0 \leq (-1)^n \frac{d^n}{dx^n} e^{-\mu f(x)} = \sum_{j=1}^{\infty} \frac{\mu^j}{j!} (-1)^{n+j} \frac{d^n}{dx^n} [f(x)]^j$$

As $\mu > 0$, dividing μ , there is:

$$0 \leq (-1)^{n+1} \frac{d^n}{dx^n} f(x) + \sum_{j=2}^{\infty} \frac{\mu^{j-1}}{j!} (-1)^{n+j} \frac{d^n}{dx^n} [f(x)]^j$$

Then let $\mu \rightarrow 0$:

$$0 \leq (-1)^{n-1} \frac{d^{n-1}}{dx^{n-1}} f'(x)$$

Eventually, $f'(x)$ is completely monotone.

If $f'(x)$ is completely monotone:

$$(-1)^{n-1} \frac{d^n}{dx^n} f(x) \geq 0$$

Let $g(\lambda) = e^{-\lambda}$, $\lambda = f(x)$:

$$h(x) = e^{-f(x)} = g(\lambda) \circ f(x)$$

And there is a formula for the n -th derivative of the composition $h = g \circ f$:

$$h^{(n)}(\lambda) = \sum_{(m, i_1, \dots, i_l)} \frac{n!}{i_1! \dots i_l!} g^{(m)}(f(\lambda)) \prod_{j=1}^l \left(\frac{f^{(j)}(\lambda)}{j!} \right)^{i_j},$$

where $\sum_{j=1}^l j \cdot i_j = n$ and $\sum_{j=1}^l i_j = m$.

We can see that $n = m + \sum_{j=1}^l (j-1) \cdot i_j$.

We have $(-1)^m g^{(m)}(f(x)) \geq 0$ and $(-1)^{j-1} f^{(j)} \lambda \geq 0$.

So $(-1)^n h^{(n)}(x) \geq 0$ which means $e^{-f(x)}$ is completely monotone.

And $e^{-\mu f(x)}$ is completely monotone.

□

3 Multivariable Distribution

4 Multinomial Distribution

4.1 Bivariate Distribution

Given a pair of discrete random variable X and Y , define the joint mass distribution by $f_{X,Y}(X = x, Y = y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x \text{ and } Y = y)$.

Definition 4.1. In the continuous case, we call a function $f(x, y)$ a probability density function, if

1. $f(x, y) \geq 0$ for all x, y .
2. $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1$.
3. for any set $A \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}((X, Y) \in A) = \iint_A f(x, y) dx dy$.

The cumulative distribution function of joint (X, Y) is given by $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$.

Definition 4.2. If random variable X and Y have joint probability density function $f_{X,Y}(x, y)$, then the marginal distribution function is given by $f_X(x) = \int f_{X,Y}(x, y) dy$.

Definition 4.3. Random variables X and Y are independent, if for every A and B , $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B)$.

Theorem 4.1. Random variables X and Y have joint probability density function $f_{X,Y}$, then X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all x and y .

Definition 4.4. If $f_Y(y) > 0$, then the conditional density function given Y is $f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(Y=y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$.

Definition 4.5. Let $X = (X_1, X_2, \dots, X_n)$ where X_i is a random variable. We call X a random vector, its probability density function is $f_{X_1, \dots, X_n}(x_1, x_2, \dots, x_n)$, and the marginal is $f(x_i) = \sum_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n} f(x_1, \dots, x_n)$ for discrete case. For continuous case, we will use integral instead.

Definition 4.6. Let $f(x_1, x_2, \dots, x_n)$ be the joint density function of X_1, X_2, \dots, X_n , $\pi_1, \pi_2, \dots, \pi_n$ is a permutation of $\{1, 2, \dots, n\}$. If $f(x_1, x_2, \dots, x_n) = f(x_{\pi_1}, x_{\pi_2}, \dots, x_{\pi_n})$, then X_1, \dots, X_n are exchangeable.

Theorem 4.2. (de Finetti) Let $X_i \subset X$ for all $i \in \{1, 2, \dots\}$. Suppose that for any n , x_1, x_2, \dots, x_n are exchangeable. Then we have

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

for some parameter θ with prior distribution $f(\theta)$.

Theorem 4.3. If $\theta \sim f(\theta)$ and X_1, X_2, \dots, X_n are conditionally iid given θ , then marginally X_1, X_2, \dots, X_n are exchangeable.

4.2 Transformation

Random variable X has pdf f_X and cmf F_X . Let $Y = g(X)$ be a function of X . In the discrete case, the pmf of Y is $f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \mathbb{P}(x \in g^{-1}(y))$.

Example 4.1. Suppose $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = \frac{1}{4}$ and $\mathbb{P}(X = 0) = \frac{1}{2}$. Let $Y = X^2$. So $\mathbb{P}(Y = 0) = \frac{1}{2}$, $\mathbb{P}(Y = 1) = \frac{1}{2}$.

In the continuous case, the steps to find density of transformation variable is given by:

1. For each y , find set $A_y = \{x : g(x) \leq y\}$.
2. Find CDF, $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(x) \leq y) = \mathbb{P}(\{x : g(x) \leq y\}) = \int_{A_y} f_X(x) dx$.
3. $f_Y(y) = F'_Y(y)$.

Example 4.2. $f_X(x) = e^{-x}$ for $x > 0$, and $Y = g(X) = \log X$. Then $F_X(x) = \int_0^x f_X(u) du = 1 - e^{-x}$. $A_Y = \{x : x \leq e^y\}$. $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\log x \leq y) = \mathbb{P}(x \leq e^y) = F_X(e^y) = 1 - e^{-e^y}$. $f_Y(y) = (1 - e^{-e^y})' = e^y e^{-e^y}$.

Example 4.3. $X \sim \text{Uniform}(-1, 3)$, $Y = X^2$. $f_X(x) = \begin{cases} \frac{1}{4} & x \in (-1, 3) \\ 0 & \text{o.w.} \end{cases}$. Now let us think about the distribution density of Y . Y can take value in $(0, 9)$.

1. $0 < Y < 1$. $A_y = \{X : X^2 \leq y\} = [-\sqrt{y}, \sqrt{y}]$. $F_Y(y) = \int_{A_y} f_X(x) dx = \frac{1}{2} \sqrt{y}$.
2. $1 \leq Y < 9$. $A_y = [-1, -\sqrt{y}] \cup [\sqrt{y}, 3]$. $F_Y(y) = \int_{A_y} \frac{1}{4} dx = \frac{1}{4}(1 + \sqrt{y})$.

So, $f_Y(y) = \begin{cases} \frac{1}{4\sqrt{y}} & 0 < y < 1 \\ \frac{1}{8\sqrt{y}} & 1 \leq y < 9 \end{cases}$

If random variable $Z = g(X, Y)$, then the way to find density of Z is given by:

1. For each z , find $A_z = \{(x, y) : g(x, y) \leq z\}$.
2. Find CDF $F_Z(z) = \mathbb{P}(Z \leq z) = \iint_{A_z} f_{X,Y}(x, y) dx dy$.
3. $f_Z(z) = F'_Z(z)$.

Example 4.4. Let $X_1, X_2 \stackrel{iid}{\sim} \text{Uniform}(0, 1)$, $Y = X_1 + X_2$. $f_{X_1, X_2}(x_1, x_2) = \begin{cases} 1 & 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{o.w.} \end{cases}$.

$$F_Y(y) = \mathbb{P}(\{(x_1, x_2) : (x_1 + x_2) \leq y\}) = \iint_{A_y} f(x_1, x_2) dx_1 dx_2 = \begin{cases} \frac{1}{2}y^2 & 0 < y < 1 \\ 1 - \frac{(1-y)^2}{2} & 1 \leq y \leq 2 \\ 1 & y > 2 \\ 0 & y \leq 0 \end{cases}. \text{ So,}$$

$$f_Y(y) = \begin{cases} y & 0 \leq y \leq 1 \\ 1 - y & 1 < y \leq 2 \\ 0 & \text{o.w.} \end{cases}$$

Theorem 4.4. Let X have CDF $F_X(x)$ and $Y = g(X)$, and let $\mathcal{X} = \{x : f_X(x) > 0\}$, $\mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}$

1. if g is a strictly increasing function on \mathcal{X} , $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$.
2. if g is a strictly decreasing function on \mathcal{X} and X is a continuous random variable.
 $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$

Theorem 4.5. Let X have continuous pdf $f_X(x)$, $Y = g(X)$, and g is strictly monotone function, then $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$

Proof. According to two case in theorem 3.4.

1. g is a strictly increasing function on \mathcal{X} , then $f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$
2. g is a strictly decreasing function on \mathcal{X} , then $f_Y(y) = \frac{dF_Y(y)}{dy} = -f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$.

So, we can combine them to $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$. □

Theorem 4.6. (Probability integral transformation) Let X has a continuous cdf $F_X(x)$, $Y = F_X(x)$. Then Y has uniform distribution on $(0, 1)$, i.e. $\mathbb{P}(Y \leq y) = y$ where $0 \leq y \leq 1$.

Proof. $\mathbb{P}(Y \leq y) = \mathbb{P}(F_X(x) \leq y) = \mathbb{P}(F_X^{-1}(F_X(x)) \leq F_X^{-1}(y)) = \mathbb{P}(x \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$. □