Multivariate distributions (again)

Recap

Section 4.6 of Casella & Berger discusses multivariate distributions (more than 2 random variables at a time). While often focusing on the bivariate case (X, Y), we've already seen the main distributional ideas for $(X_1, \ldots, X_k)$:

1. Joint cdfs: for $x_1, \ldots, x_k \in \mathbb{R}$

$$\longrightarrow F(x_1, \ldots, x_k) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_k \le x_k)$$

2. Jointly Discrete: joint pmf $f(x_1, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$,

$$\longrightarrow P((X_1, \ldots, X_k) \in A) = \sum_{(x_1, \ldots, x_k) \in A} f(x_1, \ldots, x_k), \quad A \subset \mathbb{R}^k$$

3. Jointly Continuous: joint pdf $f(x_1, \ldots, x_k)$ where

$$P((X_1, \ldots, X_k) \in A) = \int \cdots \int_A f(x_1, \ldots, x_k) \, dx_1 \, dx_2 \cdots dx_k, \quad A \subset \mathbb{R}^k$$

4. Expectations: let $g : \mathbb{R}^k \to \mathbb{R}$, define

$$\mathrm{E}g(X_1, \ldots, X_k) = \begin{cases} \sum_{(x_1, \ldots, x_k)} g(x_1, \ldots, x_k) f(x_1, \ldots, x_k) & \text{Discrete} \\ \int \cdots \int g(x_1, \ldots, x_k) f(x_1, \ldots, x_k) dx_1 \cdots dx_k & \text{Continuous} \end{cases}$$

fix

$(X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_k)$

5. Marginal distribution of $X_1, \ldots, X_p$ obtained by summing or integrating out over values $x_{p+1}, \ldots, x_k$ of other variables: fix $x_1, \ldots, x_p \in \mathbb{R}$,

$$f_{X_1, \ldots, X_p}(x_1, \ldots, x_p) = \begin{cases} \sum_{(x_{p+1}, \ldots, x_k)} f(x_1, \ldots, x_p, x_{p+1}, \ldots, x_k) & \text{discrete} \\ \int \cdots \int f(x_1, \ldots, x_p, x_{p+1}, \ldots, x_k) dx_{p+1} \cdots dx_k & \text{jcontinuous} \end{cases}$$

6. Conditionals of $X_1, \ldots, X_p$ given $X_{p+1} = x_{p+1}, \ldots, X_k = x_k$

$$f_{X_1, \ldots, X_p | X_{p+1}, \ldots, X_k}(x_1, \ldots, x_p) = \frac{f(x_1, \ldots, x_k)}{f_{X_{p+1}, \ldots, X_k}(x_{p+1}, \ldots, x_k)}$$

Multivariate Distributions

Common Multivariate Models

Multivariate distributions so far

- textbook examples of $f(x, y)$
- hierarchical models: $f(x, y) = f_X(x) f_{Y|X=x}(y|x)$
- $X_1, \ldots, X_n$ independent with common $f_X(x)$

There are really not so many standard families of multivariate distributions.

Most often, the commonly referenced distributions are

1. Multiple Hypergeometric (generalizing the hypergeometric)
2. Multinomial (generalizing the Binomial)
3. Dirichlet (generalizing the Beta)
4. Multivariate Normal Distribution (generalizing the normal)

not so important for us.

Out of these, the multivariate normal distribution is by far the most important in statistics.

Multivariate Distributions

Blue Green
N_1 N_2

Multiple Hypergeometric Distribution
without replacement

N

- Sample K at random from a population containing N_i items of type i, $i = 1, \ldots, \ell$.

$N_1 = $ # of items in Population of type 1
$N_2 = $ # of " " " " type 2
⋮
$N_\ell = $ # of " " " " type ℓ

$\underbrace{N = N_1 + N_2 + \cdots + N_\ell}$

- Let X_i=# of items in the sample of type i.

- Then $X = (\underbrace{X_1, \ldots, X_\ell})$ has a multiple hypergeometric distribution with joint pmf

items $\to$ 1 2 … ℓ with N_1 N_2 … N_ℓ; $N = N_1 + N_2 + \cdots + N_\ell$; K

$$\begin{aligned}
\longrightarrow\ & P(X_1 = x_1, \ldots, X_\ell = x_\ell) \\
&= f(x_1, \ldots, x_\ell) \\
&= \begin{cases} \dfrac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_\ell}{x_\ell}}{\binom{N_1 + N_2 + \cdots + N_\ell}{K}} & \text{integer } x_1, \ldots, x_\ell \geq 0,\ \underbrace{\sum_{i=1}^{\ell} x_i = K} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

$\binom{N}{K} =$

- Here individual marginal distributions are hypergeometric; conditionals of some given others are multiple hypergeometric

- Example ($\ell = 4$): $\underbrace{N_1 = 5}_{X_1}$, $\underbrace{N_2 = 7}_{X_2}$, $\underbrace{N_3 = 9}_{X_3}$, $\underbrace{N_4 = 11}_{X_4}$, $K = 4$

total $N = 32$

$\boxed{\sum_{i=1}^{4} x_i = 4}$

$X_1 \sim \text{Hypergeometric}(K=4,\ M=5=N_1,\ N=32)$

$$\to \mathbb{P}(X_1 = x_1) = \sum_{\underbrace{(x_2, x_3, x_4)}} \mathbb{P}(X_2 = x_2, X_3 = x_3, X_4 = x_4, X_1 = x_1)$$

$\underbrace{X_1, X_2, X_3 \mid X_4 = 2} \sim$ Multiple Hypergemetric dist.

$N_1 = 5$, $N_2 = 7$, $N_3 = 9$

$K - X_4 = 4 - x_4 = 2$

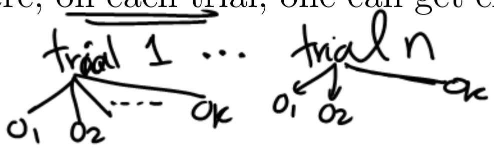
$$\frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = 2)}{\mathbb{P}(X_4 = 2)}$$

Multivariate Distributions

Multinomial Distribution

- Suppose $0 \le p_1, p_2, \dots, p_k$ are probabilities such that $\sum_{i=1}^k p_i = 1$.

- Consider a series of n identical trials where, on each trial, one can get exactly one of k possible outcomes $o_1, \dots, o_k$.



- Let X_i=# of trials resulting in outcome o_i.

- Then $X = (X_1, \dots, X_k)$ has a multinomial$(n, p_1, \dots, p_k)$ distribution with joint pmf

$$x_1 + x_2 + x_3 + \underbrace{x_4}_{2} = 4 \qquad x_1, x_2, x_3 \mid x_4 = 2$$

$$
\begin{aligned}
&P(X_1 = x_1, \dots, X_\ell = x_\ell) \\
&= f(x_1, \dots, x_\ell) \\
\binom{n}{x_1\, x_2 \dots x_k} \to &= \begin{cases} \dfrac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} & \text{integer } x_1, \dots, x_k \ge 0,\ \sum_{i=1}^k x_i = n \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

- $\dfrac{n!}{x_1! x_2! \cdots x_k!}$ is the multinomial coefficient, counting number of ways/arrangements that n trials could result in x_1 outcomes o_1, x_2 outcomes o_2,...,x_k outcomes o_k.

- Here individual marginal distributions are Binomial(n, p_i); conditionals of some given others are multinomial.

k=# of outcomes on a single trial

- Example ($k = 4$, $n = 4$):

\# of trials

$$(X_1, X_2, X_3, X_4)$$

$$X_1 \sim \text{Bin}(4, p_1)$$

$$\mathbb{P}(X_1 = x_1) = \sum_{x_2, x_3, x_4} \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = \cdots$$

$$X_3 \sim \text{Bin}(\underbrace{n}_{4=}, p_3)$$

$$X_1, X_2, X_3 \mid X_4 = 2 \sim \text{multinomial}(n^*, p_1^*, p_2^*, p_3^*) \quad \text{with } p_1^* + p_2^* + p_3^* = 1$$

$$p_1^* = \frac{p_1}{p_1 + p_2 + p_3}, \quad p_2^* = \frac{p_2}{p_1 + p_2 + p_3}$$

$$p_3^* = \frac{p_3}{p_1 + p_2 + p_3}$$

$\mathbb{P}(X_1=x_1, X_2=x_2, X_3=x_3 \mid X_4=2) = \dfrac{\mathbb{P}(X_1=x_1, \dots, X_4=x_4)}{\mathbb{P}(X_4=2)}$

Multivariate Distributions

(STAT 5440)

Dirichlet Distribution

- Suppose $Y_1, \dots, Y_k$ are independent and $Y_i \sim$Gamma$(\alpha_i, 1)$, $i = 1, \dots, k$

$0 < X_i < 1$, $\sum_{i=1}^{k} X_i = 1$

- Let $X_i = Y_i / \sum_{j=1}^{k} Y_j$

- Then $X = (X_1, \dots, X_k)$ has a Dirichlet$(\alpha_1, \dots, \alpha_k)$ distribution

- Note that $X_1, \dots, X_k$ are random variables which sum to 1

- Here individual marginal distributions are Beta$(\alpha_i, \sum_{j=1, j\neq i}^{k} \alpha_j)$; conditionals of some given others are scalar multiples of Dirichlets

- Example $(k = 4)$: $\alpha_1 = 3$, $\alpha_2 = 2$, $\alpha_3 = 4$, $\alpha_4 = 1$