2.10. Joint Probability Distributions

Consider a vector of k random variables, $\mathbf{X} = (X_1, ..., X_k)$, (representing the outcomes of k different experiments carried out once each, or of one experiment carried out k times). The probability distribution of $\mathbf{X}$ is described by a **joint** probability mass or density function.

e.g. Consider the particular case $k = 2$, $\mathbf{X} = (X_1, X_2)$. Then the following functions are used to specify the probability distribution of $\mathbf{X}$;

2.10.1. Joint probability mass/density function

The joint mass/density function is denoted

$$f_{X_1,X_2}(x_1, x_2)$$

that is, a function of **two** variables.

- This function assigns probability to the joint space of outcomes

- in the discrete case,

$$f_{X_1,X_2}(x_1, x_2) = \mathrm{P}[(X_1 = x_1) \cap (X_2 \cap x_2)]$$

- which implies that we need

(i) $f_{X_1,X_2}(x_1, x_2) \geq 0$ for all possible outcomes x_1, x_2.

(ii) $\sum\sum f_{X_1,X_2}(x_1, x_2) = 1$ or $\int\int f_{X_1,X_2}(x_1, x_2)\ dx_1 dx_2 = 1$

where the double summation/integration is over all possible values of (x_1, x_2).

Typically, such a specification is represented by a **probability table**; for example for **discrete** random variables X_1 and X_2, we could have

		X_1			
		1	2	3	4
	1	0.100	0.200	0.000	0.000
	2	0.200	0.250	0.050	0.000
X_2	3	0.000	0.050	0.050	0.025
	4	0.000	0.000	0.025	0.050

where the entry in column i, row j is

$$f_{X_1,X_2}(i, j) = \mathrm{P}[X_1 = i) \cap (X_2 = j)] = \mathrm{P}[X_1 = i, X_2 = j],$$

Here we only study joint distributions for two variables, but the extension to more than two variables is straightforward.

2.10.2. Marginal probability mass/density functions

The joint mass function automatically defines the probability distribution of the individual random variables. For example, if $k = 2$, then we have the two marginal mass/density functions are $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$

- in the discrete case,

$$f_{X_1}(x_1) = \sum_{x_2} f_{X_1,X_2}(x_1, x_2)$$

and

$$f_{X_2}(x_2) = \sum_{x_1} f_{X_1,X_2}(x_1, x_2)$$

- in the continuous case,

$$f_{X_1}(x_1) = \int f_{X_1,X_2}(x_1, x_2)\ dx_2$$

$$f_{X_2}(x_2) = \int f_{X_1,X_2}(x_1, x_2)\ dx_1.$$

so the marginal mass/density function for random variable X_1 is obtained by summing/integrating out the joint mass/density function for X_1 and X_2 over all possible values of random variable X_2. In the discrete case

$$\mathrm{P}[X_1 = x_1] = \sum_{x_2} \mathrm{P}[(X_1 = x_1) \cap (X_2 = x_2)]$$

which is a result that is justified by the Theorem of Total Probability.

In the table above, the marginal mass functions can be computed easily

		X_1				$\downarrow$
		1	2	3	4	$f_{X_2}(x_2)$
X_2	1	0.100	0.200	0.000	0.000	**0.300**
	2	0.200	0.250	0.050	0.000	**0.500**
	3	0.000	0.050	0.050	0.025	**0.125**
	4	0.000	0.000	0.025	0.050	**0.075**
$\rightarrow$	$f_{X_1}(x_1)$	**0.300**	**0.500**	**0.125**	**0.075**	

so that the marginal mass functions are formed by the column and row sums respectively. In this case, it turns out that

$$f_{X_1}(x) = f_{X_2}(x) \qquad x = 1, 2, 3, 4$$

but this will not always be the case.

2.10.3. Conditional mass/density functions

In the discrete two variable case, consider the probability

$$\mathrm{P}[\ X_1 = x_1 \mid X_2 = x_2\ ]$$

that is, the conditional probability distribution of X_1, given that $X_2 = x_2$. This conditional distribution is easily computed from the conditional probability definition, that is

$$\mathrm{P}[X_1 = x_1|X_2 = x_2] = \frac{\mathrm{P}[\ X_1 = x_1, X_2 = x_2\ ]}{\mathrm{P}[\ X_2 = x_2\ ]} = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}$$

that is, proportional to the x_2 row of the table.

By extending these concepts, we may define the conditional probability distributions for both variables in the discrete and continuous cases; The two conditional mass/density functions are $f_{X_1|X_2}(x_1|x_2)$ and $f_{X_2|X_1}(x_2|x_1)$

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &= \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} \qquad && f_{X_2}(x_2) > 0 \\ f_{X_2|X_1}(x_2|x_1) &= \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \qquad && f_{X_1}(x_1) > 0 \end{aligned}$$

In the discrete case, this result becomes

$$f_{X_1|X_2}(x_1|x_2) = \mathrm{P}[X_1 = x_1|X_2 = x_2] = \frac{\mathrm{P}[(X_1 = x_1) \cap (X_2 = x_2)]}{\mathrm{P}[X_2 = x_2]}$$

if $\mathrm{P}[X_2 = x_2] > 0$, which is justified by the definition of conditional probability.

For example, consider the table

		X_1				
		1	2	3	4	$f_{X_2}(x_2)$
X_2	1	0.100	**0.200**	0.000	0.000	0.300
	2	0.200	**0.250**	0.050	0.000	0.500
	3	0.000	**0.050**	0.050	0.025	0.125
	4	0.000	**0.000**	0.025	0.050	0.075
	$f_{X_1}(x_1)$	0.300	**0.500**	0.125	0.075	

The highlighted column gives the conditional mass function for X_2 given that $X_1 = 2$; from the definition,

$$f_{X_2|X_1}(1|2) = \frac{0.200}{0.500} = 0.400 \qquad f_{X_2|X_1}(1|2) = \frac{0.250}{0.500} = 0.500$$

$$f_{X_2|X_1}(3|2) = \frac{0.050}{0.500} = 0.100 \qquad f_{X_2|X_1}(4|2) = \frac{0.000}{0.500} = 0.000$$

Note that

$$\sum_{x=1}^{4} f_{X_2|X_1}(x|2) = 0.400 + 0.500 + 0.100 + 0.000 = 1$$

which we must have for a conditional mass function.

Note that, in general, the conditional mass functions will be **different** for different values of the conditioning variable.

SUMMARY :

Suppose that X_1 and X_2 are discrete random variables that take values $\{1, 2, ..., n\}$ and $\{1, 2, ..., m\}$ respectively. Then the joint mass function can be displayed as a table with n columns and m rows, where

- the (i, j)th cell contains $\mathrm{P}[(X_1 = i) \cap (X_2 = j)]$
- the marginal mass function for X_1 is given by the **column** totals
- the marginal mass function for X_2 is given by the **row** totals
- the conditional mass function for X_1 given $X_2 = j$ is given by the jth row divided by the sum of the jth row
- the conditional mass function for X_2 given $X_1 = i$ is given by the ith column divided by the sum of the ith column

for $i = 1, ..., n$ and $j = 1, ..., m$.

CONTINUOUS EXAMPLE

If the joint density of continuous variables X_1 and X_2 is given by

$$f_{X_1,X_2}(x_1, x_2) = x_2^2 e^{-x_2(1+x_1)}$$

for $x_1, x_2 \geq 0$ and zero otherwise. It can be shown that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_1 dx_2 = 1$$

and that the marginal pdf for X_1 is given by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_2 = \int_0^\infty x_2^2 e^{-x_2(1+x_1)}\ dx_2 = \frac{2}{(1 + x_1)^3}$$

for $x_1 \geq 0$, and zero otherwise.

EXERCISE Check these calculations, and compute the marginal pdf for X_2.

INDEPENDENCE

Random variables X_1 and X_2 are **independent** if

(i) the joint mass/density function of X_1 and X_2 factorizes into the product of the two marginal pdfs, that is,

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

(ii) the range of X_1 does not conflict/influence/depend on the range of X_2 (and *vice versa*).

The concept of independence for random variables is closely related to the concept of independence for events.

EXERCISE A point is to be selected from the interior of the unit circle. Let X and Y be the continuous random variables corresponding to the x- and y-coordinates respectively. Suppose that all points within the circle are equally likely to be selected.

(i) write down the joint pdf of X and Y.

(ii) find the marginal pdfs of X and Y.

(iii) state whether X and Y are independent.

2.10.4. MULTIVARIATE EXPECTATIONS AND COVARIANCE

For random variables $X_1, ..., X_k$ with range $\mathbb{X}^{(k)}$ with mass/density function $f_{X_1,...,X_k}$, the **expectation** of $g(X_1, ..., X_k)$ is defined in the discrete and continuous cases by

$$\mathrm{E}_{f_{X_1,...,X_k}}[\ g(X_1, ..., X_k)\ ] = \begin{cases} \displaystyle\sum_{x_1} ... \sum_{x_k} g(x_1, ..., x_k) f_{X_1,...,X_k}(x_1, ..., x_k) \\ \displaystyle\int_{\mathbb{X}_1} ... \int_{\mathbb{X}_k} g(x_1, ..., x_k) f_{X_1,...,X_k}(x_1, ..., x_k)\ dx_1 ... dx_k \end{cases}$$

COVARIANCE AND CORRELATION

The **covariance** of two random variables X_1 and X_2 is denoted $\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2] = \mathrm{E}_{f_{X_1,X_2}}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathrm{E}_{f_{X_1,X_2}}[X_1 X_2] - \mu_1 \mu_2$$

where $\mu_i = \mathrm{E}_{f_{X_i}}[X_i]$ is the expectation of X_i, for $i = 1, 2$, and where $\mathrm{E}_{f_{X_1,X_2}}[X_1 X_2]$ is the expectation of function $g(x_1, x_2) = x_1 x_2$ with respect to the joint distribution f_{X_1,X_2}.

The **correlation** of X_1 and X_2 is denoted $\mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$\mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2] = \frac{\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2]}{\sqrt{\mathrm{Var}_{f_{X_1}}[X_1]\ \mathrm{Var}_{f_{X_2}}[X_2]}}$$

If

$$\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2] = \mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2] = 0$$

then variables X_1 and X_2 are **uncorrelated**. Note that if random variables X_1 and X_2 are independent then

$$\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2] = \mathrm{E}_{f_{X_1,X_2}}[X_1 X_2] - \mathrm{E}_{f_{X_1}}[X_1]\mathrm{E}_{f_{X_2}}[X_2] = \mathrm{E}_{f_{X_1}}[X_1]\mathrm{E}_{f_{X_2}}[X_2] - \mathrm{E}_{f_{X_1}}[X_1]\mathrm{E}_{f_{X_2}}[X_2] = 0$$

and so X_1 and X_2 are also uncorrelated (the converse does not hold).

Key interpretation

COVARIANCE AND CORRELATION ARE MEASURES OF THE
DEGREE OF $\boxed{\text{ASSOCIATION}}$ BETWEEN VARIABLES

that is, two variables for which the correlation is **large** in magnitude are strongly associated, whereas variables that have low correlation are **weakly** associated.

NOTES:
(i) For random variables X_1 and X_2, with (marginal) expectations μ_1 and μ_2 respectively, and (marginal) variances σ_1^2 and σ_2^2 respectively, if random variables Z_1 and Z_2 are defined $Z_1 = (X_1 - \mu_1)/\sigma_1$ and $Z_2 = (X_2 - \mu_2)/\sigma_2$ so that Z_1 and Z_2 are *standardized* variables. Then

$$\mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2] = \mathrm{Cov}_{f_{Z_1,Z_2}}[Z_1, Z_2].$$

(ii) The extension to k variables: covariances can only be calculated for *pairs* of random variables, but if k variables have a joint probability structure it is possible to construct a $k \times k$ *matrix*, $\mathbf{C}$ say, of covariance values, whose (i,j)th element is

$$\mathrm{Cov}_{f_{X_i,X_j}}[X_i, X_j] = \mathrm{Cov}_{f_{X_i,X_j}}[X_i, X_j]$$

for $i, j = 1,..,k$, (so $\mathbf{C}$ is *symmetric)* that captures the complete covariance structure in the joint distribution. If $i = j$,

$$\mathrm{Cov}_{f_{X_i,X_i}}[X_i, X_i] \equiv \mathrm{Var}_{f_{X_i}}[X_i]$$

The matrix $\mathbf{C}$ is referred to as the **variance-covariance** matrix.

(iii) If random variable X is defined by

$$X = \sum_{i=1}^{k} a_i X_i$$

for random variables $X_1, ..., X_k$ and constants $a_1, ..., a_k$, then

$$\begin{aligned} \mathrm{E}_{f_X}[X] &= \sum_{i=1}^{k} a_i \mathrm{E}_{f_{X_i}}[X_i] \\ \mathrm{Var}_{f_X}[X] &= \sum_{i=1}^{k} a_i^2 \mathrm{Var}_{f_{X_i}}[X_i] + 2\sum_{i=1}^{k}\sum_{j=1}^{i-1} a_i a_j \mathrm{Cov}_{f_{X_i,X_j}}[X_i, X_j] \end{aligned}$$

(iv) Combining (i) and (iii) when $k = 2$, and defining standardized variables Z_1 and Z_2,

$$\begin{aligned} 0 \leq \mathrm{Var}_{f_{Z_1,Z_2}}[Z_1 \pm Z_2] &= \mathrm{Var}_{f_{Z_1}}[Z_1] + \mathrm{Var}_{f_{Z_2}}[Z_2] \pm 2\ \mathrm{Cov}_{f_{Z_1,Z_2}}[Z_1, Z_2] \\ &= 1 + 1 \pm 2\ \mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2] \\ &= 2(1 \pm \mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2]) \end{aligned}$$

and hence we have the key result that

$$-1 \leq \mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2] \leq 1.$$

that is, the correlation is **bounded** between -1 and 1.

We will see later how to compute covariance and correlation for sample data; there is a close relationship between theoretical and sample covariances and correlations.

2.10.5 THE MULTINOMIAL DISTRIBUTION

The **multinomial** distribution is a multivariate generalization of the **binomial** distribution. One interpretation of the binomial distribution is that it is the probability distribution for the random variable that counts the number of "successes" in a sequence of Bernoulli experiments. Let us label the "successes" (1) as **Type I** outcomes and "failures" (0) as **Type II** outcomes. Suppose as usual that the probability of Type I outcomes is θ (so $0 \leq \theta \leq 1$) and hence the probability of Type II outcomes in the urn is $1-\theta$.

If n experiments are carried out, and X is the random variable corresponding to the number of Type I outcomes, then $X \sim Binomial(n,\theta)$, and

$$f_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad \text{for } x \in \{0,1,2,...,n\}$$

Now consider a generalization; suppose that there are $k+1$ types of outcomes ($k = 1,2,...$), with

$$\text{"Probability of type } i\text{" outcome} \;=\; \theta_i$$

for $i = 1,...,k+1$. Let X_i be the random variable corresponding to the number of type i outcomes in n repeats of the experiment, for $i = 1,...,k$. Then the joint distribution of vector $\mathbf{X} = (X_1,...,X_k)$ is given by

$$f_{X_1,...,X_k}(x_1,...,x_k) = \frac{n!}{x_1!...x_k!x_{k+1}!}\theta_1^{x_1}....\theta_k^{x_k}\theta_{k+1}{}^{x_{k+1}} = \frac{n!}{\left\{\prod\limits_{i=1}^{k+1} x_i!\right\}}\prod_{i=1}^{k+1}\theta_i^{x_i} = \binom{n}{x_1,x_2,...,x_k}\prod_{i=1}^{k+1}\theta_i^{x_i}$$

where $0 \leq \theta_i \leq 1$ for all i, and $\theta_1 + ... + \theta_k + \theta_{k+1} = 1$, and where x_{k+1} is defined by $x_{k+1} = n - (x_1 + ... + x_k)$.

This is the mass function for the **MULTINOMIAL DISTRIBUTION** which reduces to the binomial if $k = 1$. It can also be shown that the marginal distribution of X_i is given by

$$X_i \sim Binomial(n,\theta_i).$$

EXAMPLE Recall the previous example of a nucleotide sequence seen on page 7.: Consider the sequence

$$ATAGTAGATACGCACCGAGGA$$

For the probability of seeing such a sequence, we let

$$\mathrm{P}(A) = p_A \qquad \mathrm{P}(C) = p_C \qquad \mathrm{P}(G) = p_G \qquad \mathrm{P}(T) = p_T$$

for some suitable probabilities satisfying

$$0 \leq p_A, p_C, p_G, p_T \leq 1 \qquad p_A + p_C + p_G + p_T = 1$$

and assume independence so that

$$\mathrm{P}(ATAGTAGATACGCACCGAGGA) = p_A \times p_T \times p_A \times ... \times p_G \times p_A$$

which simplifies to

$$\mathrm{P}(ATAGTAGATACGCACCGAGGA) = p_A^8 p_C^4 p_G^6 p_T^3$$

However, for if we merely wish to identify the probability that in a sequence of 21 bases, we observe 8 A, 4 C, 6 G and 3 T in **any order** then the multinomial distribution is used, that is the probability is given by

$$\binom{21}{8,4,6,3}p_A^8 p_C^4 p_G^6 p_T^3 = \frac{21!}{8! \times 4! \times 6! \times 3!}p_A^8 p_C^4 p_G^6 p_T^3$$

2.10.6 ORDER STATISTICS: MAXIMA AND MINIMA

Order statistics are a special kind of transformed random variable that give rise to a particular multivariate distribution. For n random variables $X_1, ..., X_n$, the order statistics, $Y_1, ..., Y_n$, are defined by

$$Y_i = X_{(i)} - \text{"the } i\text{th smallest value in } X_1, ..., X_n\text{"}$$

for $i = 1, ..., n$, so that

$$Y_1 = X_{(1)} = \min\{X_1, ..., X_n\} \qquad\qquad Y_n = X_{(n)} = \max\{X_1, ..., X_n\}$$

For n independent, identically distributed random variables $X_1, ..., X_n$, with marginal density function f_X, there are two main results to consider;

RESULT 1 The joint density function of the order statistics $Y_1, ..., Y_n$ is given by

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = n! f_X(y_1)...f_X(y_n) \qquad y_1 < ... < y_n$$

RESULT 2 The marginal pdf of the jth order statistic Y_j for $j = 1, ..., n$ has the form

$$f_{Y_j}(y_j) = \frac{n!}{(j-1)!(n-j)!}\{F_X(y_j)\}^{j-1}\{1 - F_X(y_j)\}^{n-j} f_X(y_j)$$

SPECIAL CASES: MAXIMUM AND MINIMUM:

To derive the marginal pdf of the **maximum** Y_n, first consider the marginal cdf of Y_n;

$$\begin{aligned}
F_{Y_n}(y_n) &= \text{P}[Y_n \leq y_n] = \text{P}[\max\{X_1, ..., X_n\} \leq y_n] \\
&= \text{P}[X_1 \leq y_n, X_2 \leq y_n, ..., X_n \leq y_n] \\
&= \prod_{i=1}^{n} \text{P}[X_i \leq y_n] = \prod_{i=1}^{n}\{F_X(y_n)\} \\
&= \{F_X(y_n)\}^n \\
\Longrightarrow f_{Y_n}(y_n) &= n\{F_X(y_n)\}^{n-1} f_X(y_n) \qquad \text{differentiating using the chain rule}
\end{aligned}$$

By a similar calculation, we can find the marginal pdf/cdf for the **minimum** Y_1;

$$\begin{aligned}
F_{Y_1}(y_1) &= \text{P}[Y_1 \leq y_1] = 1 - \text{P}[Y_1 > y_1] = 1 - \text{P}[\min\{X_1, ..., X_n\} > y_1] \\
&= 1 - \text{P}[X_1 > y_1, X_2 > y_1, ..., X_n > y_1] \\
&= 1 - \prod_{i=1}^{n} \text{P}[X_i > y_1] = 1 - \prod_{i=1}^{n}\{1 - F_X(y_1)\} \\
&= 1 - \{1 - F_X(y_1)\}^n \\
\Longrightarrow f_{Y_1}(y_1) &= n\{1 - F_X(y_1)\}^{n-1} f_X(y_1) \qquad \text{differentiating using the chain rule}
\end{aligned}$$

Hence

$$F_{Y_1}(y_1) = 1 - \{1 - F_X(y_1)\}^n \qquad f_{Y_1}(y_1) = n\{1 - F_X(y_1)\}^{n-1} f_X(y_1)$$