Random variables (隨機變數), random samples (隨機樣本) and population distributions (母體分布)

- We consider X a random variable if it makes sense to talk about the probability that X is in some interval A (denoted by $P(X \in A)$).
  - Examples of random variables: waiting time, number of occurrences of certain event during a period of time (event should occur with uncertainty).
- A sample $(X_1, \ldots, X_n)$ can be thought as a vector of n random variables which we can observe.
- The distribution (分布) of a random variable X is a function that maps an interval A to $P(X \in A)$. Thus if we know the distribution of X, then we know
$$P(X \in A) \text{ for every interval } A.$$
  - Example. If $P(X = 1) = p$ and $P(X = 0) = 1 - p$, then the distribution of X can be characterized by p.
- Common types of random variables
  - For a random variable X, if the possible values of X can be listed as a sequence, then X is a discrete(離散型的) random variable.
  - If $P(X = x) = 0$ for every x, then X is called a continuous (連續型的) random variable.
- For a discrete random variable X,
$$P(X \in A) = \sum_{x \in A, P(X=x)>0} P(X = x).$$
In such case, the distribution of X can be described by the function p_X: $p_X(x) = P(X = x)$ for all x. The function p_X is called the probability mass function (PMF) of X.

  Example 1. Toss a fair coin twice and let X be the number of heads in the two tosses. Let p_X be the PMF of X, then $p_X(x) = P(X = x)$, so
$$p_X(x) = \begin{cases} 0.25 & \text{if } x = 0; \\ 0.5 & \text{if } x = 1; \\ 0.25 & \text{if } x = 2; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$
  Or, one can use the following table to describe the PMF of X:

x	$P(X = x)$
0	0.25
1	0.5
2	0.25
any value not in { 0, 1, 2 }	0

Example 2. Suppose X is a random variable and the PMF of X is given in (1). Find $P(1 \le X \le 2)$.

Sol. From (1), the possible values for X are 0, 1, 2, so

$$P(1 \le X \le 2) = P(X = 1) + P(X = 2) = 0.25 + 0.5 = 0.75.$$

- Independence of events. We say that events A_1, …, A_n are independent (獨立) if for any m events A_{n_1}, …, A_{n_m} in $\{A_1, \ldots, A_n\}$, we have

$$P(A_{n_1} \cap A_{n_2} \cap \cdots \cap A_{n_m}) = P(A_{n_1}) \cdot P(A_{n_2}) \cdots P(A_{n_m}).$$

- Another equivalent definition of independence of events. Events A_1, …, A_n are independent if

$$P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2) \cdots P(B_n)$$

for all B_1, …, B_n such that $B_k \in \{A_k, A_k^c\}$ for $k = 1$, …, n.

- We say n random variables X_1, …, X_n are independent if the events $\{X_1 \in A_1\}$, …, $\{X_n \in A_n\}$ are independent for all A_1, …, A_n.
- We say n random variables X_1, …, X_n are IID (independently and identically distributed; 獨立同分布) if X_1, …, X_n are independent have the same distribution.
- If X_1, …, X_n are discrete random variables, then X_1, …, X_n have same distribution means that they have the same PMF.
- Random sample (隨機樣本). A sample $(X_1, \ldots, X_n)$ is called a random sample if X_1, …, X_n are IID. The distribution for each X_i is called the population distribution (母體分布).
- Suppose that $(X_1, \ldots, X_n)$ is a random sample and each X_i is discrete with PMF p_X. Then the population distribution is characterized by p_X and p_X is called the population PMF.

Example 3. 丟銅板. Consider the experiment of tossing a fair coin twice. Suppose we run the experiment n times (independently) and let X_i be the number of heads in the i-th trial. Then $(X_1, \ldots, X_n)$ is a random sample with population PMF given in (1).

Example 4. 抽樣取出放回. Suppose that we have a group of N scores, of which 25% are 0's, 50% are 1's and 25% are 2's. Consider the experiment of randomly selecting a score from the N scores and then putting it back. Suppose we run the experiment n times and let X_i be the number selected in the i-th trial. Then $(X_1, \ldots, X_n)$ is a random sample with population PMF given in (1).

Example 5. 抽樣取出不放回. In Example 4, if the selected numbers are not put back, then X_1, ..., X_n have the same distribution but are not independent. However, if N is very large comparing to n, X_1, ..., X_n are "approximately" independent in the sense that

$$P((X_1, \ldots, X_n) = (x_1, \ldots, x_n)) \approx P(X_1 = x_1) \cdots P(X_n = x_n).$$

Therefore, the sample $(X_1, \ldots, X_n)$ can be viewed as an "approximate" random sample.

- Fact (from the law of large numbers 大數法則). Suppose that $(X_1, \ldots, X_n)$ is a random sample. Then for every interval A and for large n,

$$\text{percentage of } X_1, \ldots, X_n \text{ that are in } A \approx P(X_1 \in A).$$