# 模型符號介紹與相關文獻回顧

## 符號定義

在常見的生態資料的蒐集上，依抽樣單位分為兩種資料型態：其一為以個體為抽樣單位的豐富度數據 (abundance data)；其次為依照區塊為抽樣單位的出現型數據 (incidence data)。本文主要以區塊為抽樣單位的出現型數據來進行共同物種數估計。本文使用的符號如下表示：

#### 母體群落

* ：第*j*群落的物種數，。
* ：兩群落的共同物種數。
* ：混合群落中的總相異物種數，。
* ：第*j*群集的總區塊數量，。
* ：物種*i*在第 *j*群落區塊中所出現的比例，，。

#### 樣本

* ：第一群落樣本中第*i*物種出現的區塊數量。
* ：第二群落樣本中第*i*物種出現的區塊數量。
* ：第*j*群落樣本中出現的物種數，，。
* ：兩樣本中出現的共同物種數，。
* ：第*j*個集的抽樣區塊數量，。
* ：抽樣比例，。
* ：第*j*群落樣本中出現*k*個區塊的物種數，即物種的區塊計數。，。
* ：在第一群落樣本出現*k*個區塊，並在第二群落樣本出現*l*個區塊的物種數。，， 。
* ：在第一群落樣本出現*k*個區塊，並在第二群落樣本出現至少一個區塊的物種數。， 。
* ：在第一群落樣本出現至少一個區塊，並在第二群落樣本出現*l*個區塊的物種數。， 。

## 相關文獻回顧

### 出現型數據

文獻中，大多數的物種數估計研究，其理論依據皆假設抽樣單位為隨機且獨立，但在真實野外的情境下，由於物種會因不同的習性以及特性而聚集，導致在某些特定情況下，更容易觀測到相同物種。因此以個體為抽樣單位的方式進行調查較難符合該假設。而相較於個體抽樣，僅記錄物種在該區塊出現與否的區塊抽樣較能符合該理論假設。此時，區塊為抽樣單位，如陷阱、區塊、穿越線或固定時間的調查。在出現型數據的抽樣中，可先將研究區域劃分為多個面積大致相同的區塊，並從中隨機選擇區塊做為抽樣樣本進行調查。對於不同類型的物種，準確計算每個抽樣區塊中出現的個體數往往是一件相對困難的。因此在多數情況下，調查時僅記錄該物種抽樣區塊中的物種出現與否。

出現型數據由個抽樣區塊的樣本所組成，其中每個抽樣區塊只紀錄物種的出現或是未出現，以形成一個以物種為列區塊為行的矩陣。其中，若是*i*物種出現在*j*區塊中，無論個體總數為何，皆計為1 ()；反之若尚未觀測到該物種在區塊中則計為0 ()。

以第一群落為例，為落樣本中第物種出現的區塊數量)，表示在該樣本中第*i*物種出現的區塊數量，故 。

因此，可得到第一樣本中剛好出現*k*個區塊的物種數，，且。其中， 為在該樣本中僅出現在一個區塊的物種數，為在該樣本中出現在兩個區塊的物種數，並依此類推。除此之外，將被定義為在該樣本並未被觀測到的物種數。而真實的物種數，應為被觀測到的物種數與未被觀測到的物種數之和 (， Figure 2.1)。

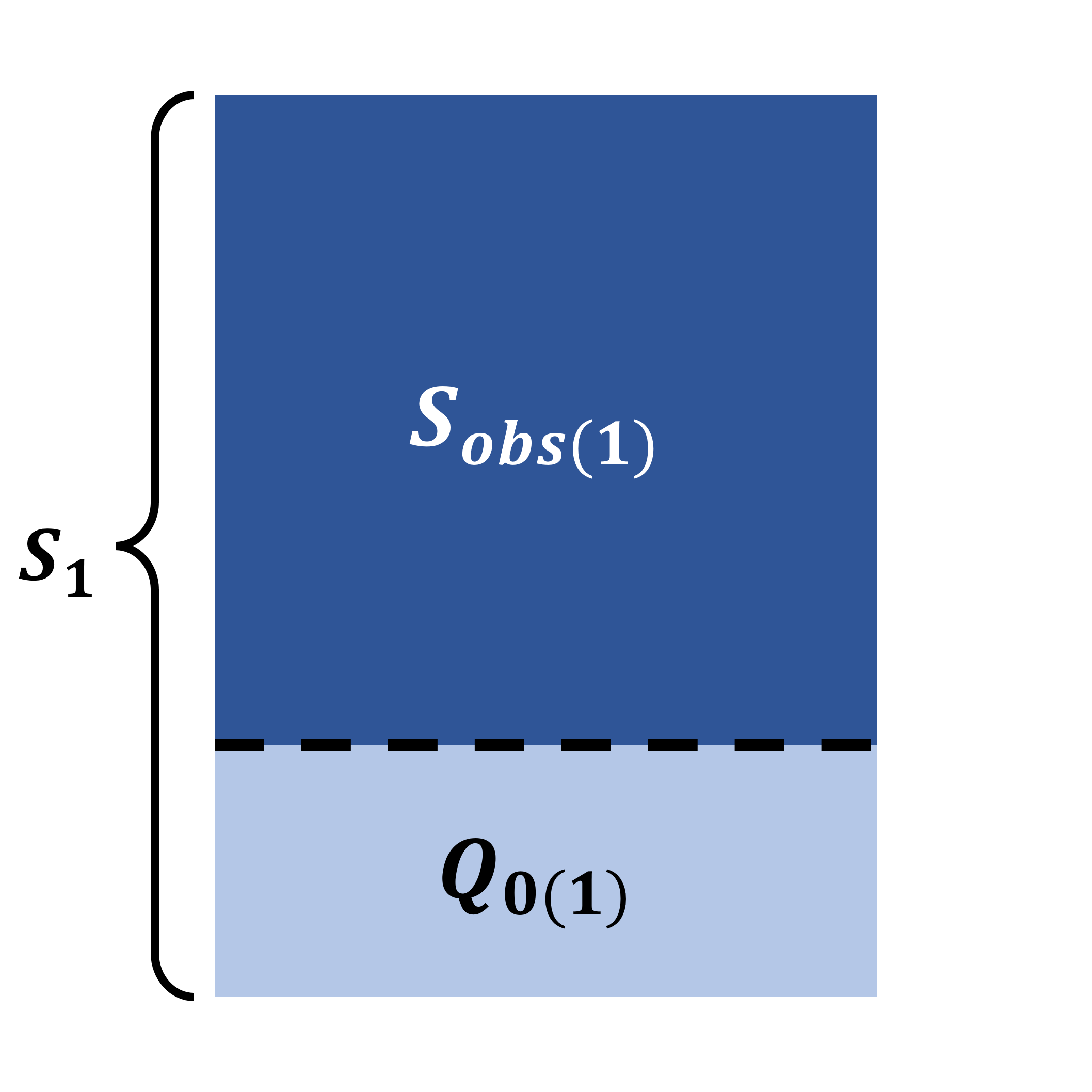


Figure 2.1 單一群落的物種分配示意圖。

### 取後放回之抽樣方式

在多數生態調查的研究中，計算物種數在群落之中的占比，是最直觀量化多樣性的指標之一。但普查物種數量往往需要消耗大量的人力、經費與時間等成本，這使得在抽樣的結果中能看見所有物種皆出現之狀況的機率大幅降低，因此在大部分的生態調查結果中，皆可能存在部分未被觀測到的物種。故需針對這些未被觀測到的物種進行估計，以獲取到更接近於真實物種數的結果。

對於出現型數據所開發之物種數估計的模型多數皆是依據捉放法 (capture-recapture) 的理論架構為基礎所建立。而在物種數的估計中，可將捉放法「個體」對應至「物種」，以估計群落中的物種數。

在物種數的調查結果中，可依據物種在樣本中出現的區塊計數，將物種大致分為豐富物種與稀有物種。當樣本中皆為豐富物種時 (出現頻率較高)，通常會被認為該群落之物種皆已被調查完全；相反，當樣本中的多數物種為稀有物種 (出現頻率較低) 時，在直覺上會認為群落中仍存在更多未被觀測到的物種，因此樣本中稀有物種對於未被觀測到的物種數提供較多的資訊。在過去許多研究中，最廣為應用估計方法是藉由稀有物種提供的資訊進行物種數估計。

依據2.2.1所述，並以第一群落為例，出現型矩陣可整理為，應為服從伯努力分佈 (Bernoulli distribution) 的隨機變數，且當時機率為 ，而時機率為。則發生率矩陣的機率分佈為：

而 服從二項分佈 (Binomial distribution)，則其機率分佈為：

#### 單群落物種數估計

在所有估計方法中，最被廣為利用的估計方法是Chao (1987) 針對出現型數據建立物種數的無母數估計模型 *Chao2*。無母數的物種數估計是一個基本且直接的觀念，所謂無母數估計意旨在該估計方法中，不對物種數或者物種出現機率的分布進行假設。以第一群落作為單群落的例子，根據樣本中物種出現區塊計數的機率分佈，可以表示為：

又可依據柯西-施瓦茨不等式 (Cauchy-Schwarz inequality)，利用稀有物種中所含的資訊以估計真實的物種數。最終求得*Chao2*估計式為：

並針對*Chao2* 估計式，Chiu (2022) 使用動差法 (moment method)，推導出新的有母數方法的估計式*BB*：

在上述式子中，；且皆表示為：且在上述等式中皆表示為：若 時，則；而若 時，則表示為。

#### 兩群落的共同物種數估計

在生態統計中，群落之間的共同物種在樣本中所佔的比例，可以表示兩群落間的物種多樣性，同時也能表現兩群落間的相似性 (Chao et al., 2000)。在兩群落的抽樣樣本中除了共同種之外，也會分別存在只出現於其中單一群落的特有物種 (Figure 2.2)。

為此，當在比較兩群落之間的物種數時，並非僅考慮單一群落或是混合群落的物種數，也必須針對群落之間所存在的共同物種數進行估計。與單群落的物種數估計相似，在大多數情況下，抽樣樣本無法觀測到所有存在的共同種。因此需針對未被抽樣觀測到的共同種進行估計，並加上於樣本中已被觀測到的共同物種數，作為修正的共同物種數所使用。

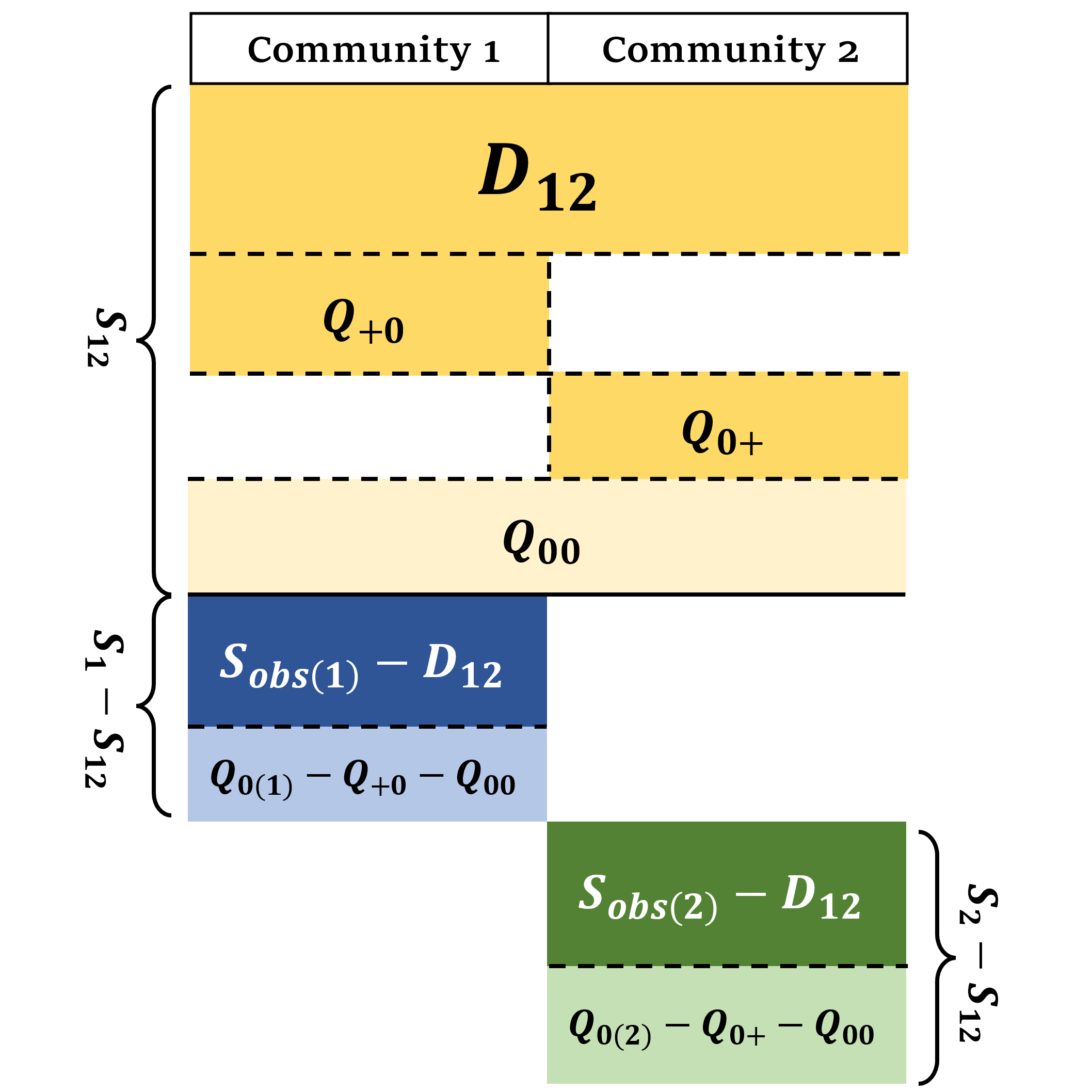


Figure 2.2 兩群落的物種分配示意圖。

Pan等人(2009) 將*Chao2*的方法推廣至兩群落，建立一估計兩群落間存在的共同物種數之估計式。根據樣本中物種出現區塊計數的機率分佈，可以表示為：

同樣可根據柯西-施瓦茨不等式，最終可以推導出估計式*Pan*：

### 取後不放回之抽樣方式

相對於取後放回的抽樣方式，另一種在生態資料中常見的抽樣方法為取後不放回，該抽樣方法廣泛使用在林業調查中，例如依照所選區塊對樹木進行不重複取樣，或是使用陷阱或誘捕器的抽樣方式等需要殺死個體的抽樣方法中。

在這種類型抽樣方法的單群落情況下，以第一群落的為例，假設將欲調查地區大致分為個相等的區塊，又每個區塊物種存在於該群落的比例為，而，且為未知參數。並假設在區塊中，物種 僅能在的目標區塊中被檢驗到，亦為未知參數，且。則在給定的條件下遵循參數 和的零截尾二項分佈 (zero-truncated binomial distribution)：

當從 個區塊中，針對群落進行取後不放回之隨機抽樣，抽取的區塊數，若在取樣區塊中發現該物種出現與否的情況。最終將各物種出現區塊的總和整理整理成物種出現的區塊數量。且在給定 的情況下，應服從超幾何分佈 (hypergeometric distribution)：

#### 單群落物種數估計

Chao and Lin (2012) 基於*Chao2*進行修正，針對取後不放回的樣本資料開發新的估計方法。以第一群落作為單一群落的情況下，在該估計方法中可以表示為：

隨後根據柯西-施瓦茨不等式的概念推導出估計式，建構出針對取後不放回的樣本資料新的估計方法*wChao2*為：

其中，，。

而Chiu (2023) 使用動差法，推導出新的有母數方法的估計式*MoRE*：

在上述式子中；此外表示為若時，則等於；若，該式則表示為。

#### 兩群落的共同種估計

與取後放回的估計方法相似，在取後不放回的估計中也存在兩群落間的共同種估計需求。Chao與Lin (2012) 將估計由單群落的物種數估計延伸至兩群落的共同物種估計，建構一個針對取後不放回的抽樣方式，估計兩群落共同種的估計方法。在給定與的情況下，可以表示為：

同理於取後不放回的單群落物種數估計方法，藉由樣本中分別未出現於兩群落的期望值計算兩群落的共同種，可得最終估計式為：

其中，，，

### 標準差估計

根據的漸近分布，其服從大小為以及機率為的多項分布 (multinomial distribution) (Chao & Lee, 1992)。所提出的物種數估計量的變異數估計量可以使用 multivariate delta-method導出，表示為

其中

### 95%信賴區間

物種豐富的信賴區間通過假設 符合對數常態分佈 (log normal distribution) (Chiu et al., 2014)，為此確保了信賴區間之下限值大於觀察到的物種數。故物種數之95%信賴區間為：

### 樣本涵蓋率

樣本涵蓋率 (sample coverage) 的概念起源於第二次世界大戰，艾倫圖靈 (Alan Turing) 與其同事I.J 古德 (I. J. Good.) 在密碼分析中所開發之估計方法。於出現率樣本中，樣本涵蓋率被定義為樣本中已檢測物種的總出現的比例，即物種的相對出現率總和。在隨機抽樣的假設下，物種出現區塊數為 ()，而對抽樣區塊中所觀測到的物種機率估計為 ，則樣本覆蓋率可表示為：

由上述式子可以得知，物種的相對出現率總和，其值應界於0至1之間。並且， 可被解釋為再多觀測一個新的區塊時，該區塊出現新物種的機率。由於該區塊出現的為新物種，固可使用樣本中僅出現一個區塊 () 的資訊估計該指標，樣本涵蓋率的估計式為：

### Jaccard指數

在量化兩群落之間的相似性時，其中一項常見的指標為Jaccard指數 (Jaccard index)。該指標為Jaccard (1901) 提出，利用共同物種數在兩群落中所佔的比例，定義群落之間的相似程度。

Jaccard指數範圍由0至1，0表示兩群落物種組成完全相異；反之，1表示為0表示兩群落物種組成完全相同。

並可將其轉換為Jaccard距離 (Jaccard distance) 用來表示兩群落間的相異性：

Jaccard距離為0至1，0表示兩群落物種組成完全相同；反之，1表示為0表示兩群落物種組成完全相異。