1 多样性

生物多样性、物种多样性、文化多样性、遗传多样性

1.1 物种多样性

也称物种岐异度，是指一定区域内动物、植物、微生物等生物种类的丰富程度

物种多样性是生物多样性的核心，也是生物多样性研究的基础

物种多样性包括物种丰富度和物种均匀度两个方面的含义：

1. 物种丰富度是对一定空间范围内的物种数目的简单描述
2. 物种均匀度则是对不同物种在数量上接近程度的衡量

1.2 多样性测度方案

1.2.1 物种丰富度

- **物种密度**：多用于植物多样性的研究，用每平方米的物种数目表示
- **数量丰度**：一定生物量中的物种数目，多用于水域物种多样性的研究，如1000条鱼中的物种数目。

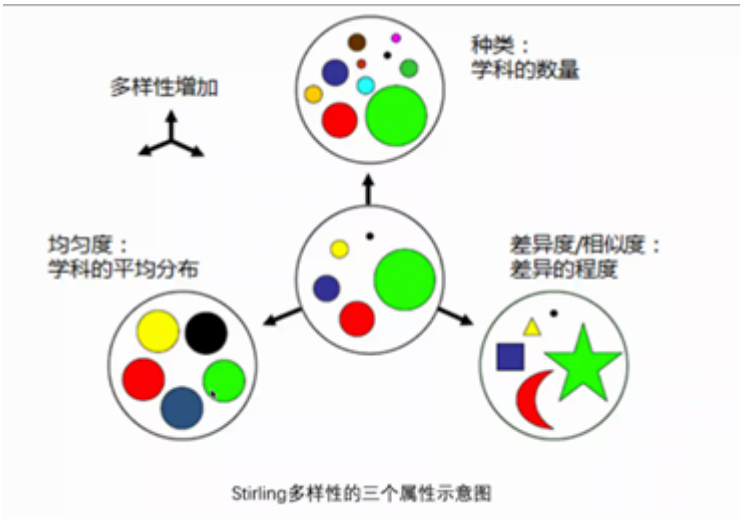
1.2.2 物种多样性指数

多样性指数是物种多样性测定，主要有三个空间尺度：α多样性，β多样性，γ多样性，每个空间尺度的环境不同，测定的数据也不相同

- α多样性：主要关注局域均匀生境下的物种数目，因此也被称为生境内的多样性
- β多样性：指沿环境梯度不同生境群落之间物种组成的相异性或物种沿环境梯度的更替速率，也被称为**生境间的多样性**

    - 控制β多样性的主要生态因子有土壤、地貌及干扰等。
- γ多样性：描述区域或大陆尺度的多样性，是指区域或大陆尺度的物种数量，也被称为**区域多样性**
    - 控制γ多样性的生态过程主要为水热动态，气候和物种形成以及演化的历史
- **α多样性信息度量**
    - **Gleanson指数**：D=S/lnA，A为调查的总面积，S为群落中的物种总数目
    - **Margalef指数**：D=(S−1)/lnN，式中S为群落中的物种总数目，N为观察到的所有物种的个体总数。
    - **Shannon− Wiener多样性指数（H’）**
        - $$H'=-\sum n_i/N\times ln\ (n_i/N)$$
        - 式中：n_i为第i个类群的个体数；N为群落中所有类群的个体总数
    - **Pielou均匀度指数**
        - E=H/Hmax，H为实际观察的物种多样性指数，Hmax为最大的物种多样性指数，Hmax=lnS（S为群落中的总物种数）
    - **多样性概率度量：Simpson优势度指数（C）**
        - $$C=\sum[\ (N_i\ (N_i-1)/N\ (N-1))\ ]$$
        - 式中：N_i/N为第i物种第一次被抽中的概率；(Ni−1)/(N−1)为第i物种第二次抽中的概率

1.3 多样性测度的应用场景——科学



Stirling多样性的三个属性示意图

Stirling多样性的三个属性示意图：用于学科交叉的测度

种类：学科的数量；均匀度：学科的平均分布；差异度/相似度：差异的程度

1.3.1 学科

1. 按照学问的性质，依据学术的性质而划分的学科门类。如自然科学中的物理学、化学；人文学科中的历史学、语言学
2. 学校教学的科目，如语文、数学、英语；
3. 军事训练或体育训练中的各种理论知识性的科目（区别于“术科”）。

1.3.2 学科的种类

学科的种类的计算是基于已有的学科分类体系，根据论文、作者或者期刊等研究层次的学科属性对其进行分类计数

- 学术分类——《学科分类与代码》
- 作用与意义：科研管理——学科分类

如何判断论文所属学科?

1. 基于词：`不太可行`，元数据

2. 基于机构：机构一般只对应少数几种学科
3. 基于作者：可行，作者的学科归属具有稳定性，难度在于大量作者是重名的
4. 基于参考文献：最可行，参考文献基于所属期刊来判断

信息熵[1]

1.3.3 学科间的相似度

1. Rao–Stirling指标
    1. $$广义\ Stirling \quad D = \sum_{i,j} d_{ij}^{\alpha}(p_i p_j)^{\beta}$$

       其中，p_i和p_j是不同学科的概率分布；d_{ij}是学科网络中不同学科之间的距离；α和β是计量参数，分别表示关系较远的学科和大的学科的权重值

> “相似度”如何测度?

> 人员、机构、参考文献重叠的频次、量

> 参考文献表征的是知识的流动

2. 余弦相似度
    1. $$余弦相似度 \quad \cos\theta = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$
    2. $\cos\theta$的取值在0-1之间，$\cos\theta$的值越趋近于1，表明两组向量之间的相似度越低，那么两个学科之间的距离d_{ij}就越大。X和Y分别是两组空间向量，其中x_i和y_i分别是X和Y向量中第i个变量的值，n为空间向量的维度。

学科：教育部学科分类（13个门类，110个一级学科，375个二级学科）

Web of Science学科分类（五大类，约250个小类）

学科交叉：学科交叉概念第一次提出是由哥伦比亚大学心理学家R.S.Woodworth，他于1962年率先指出跨学科活动是打破已知学科边界并涉及两个或两个以上学科的研究活动

> 文理学科交叉对文科发展的影响?
>
> Q：学科应该如何交叉?
>
> - 假设：**文理交叉有助于成果影响力的提升**
> - 变量：
>   - 自变量：人文社会科学论文的参考文献中理工农医论文的比例
>   - 因变量：人文社会科学论文的被引次数
>   - 控制变量：学科自引比例、论文作者的数量、关键词数量、页码数量、参考文献数量等等
> - **结论：文科中使用太多数理统计知识反而会阻碍知识传播**

2 信息熵

信息熵是香农将热力学熵的概念引入信息论而提出来的，是测度不确定性的重要概念

在跨学科性的测度中，信息熵表征的是分类分布的均匀度：**信息熵的值越大，表示一组研究对象中各个分类的平均分布的程度越高**

计算方式：

$$信息熵\ H = -\sum_i p_i \ln p_i$$

其中，p_i表示不同分类的概率分布

布里渊指数：

- 布里渊指数是在1956年，在信息熵计算原理的基础上提出的，用于测度一条信息中所包含的信息量
- 源于信息熵的布里渊指数实质上也是用于测量不确定性的
- 观测对象的总数N越大或者观测对象中分类的数量n越大，并且观测对象在分类中分布越均匀时，布里渊指数H的值就越大，表示该观测对象的多样性程度越高
- 计算方式为：

$$H = \frac{1}{n}\log\frac{n!}{\prod_{i=1}^{k} n_i!} = \frac{\log n! - \sum_{i=1}^{k}\log n_i!}{n} \qquad H = \frac{\log N! - \sum(\log n_i!)}{N}$$

其中，N表示观测对象的总数，n_i表示观测对象归属于分类i的数量

---

1. # 2 信息熵