王慈俭 SA23008216 生命科学学院生态学

1. **What difficulties does the old paradigm of the hypothesis test encounter?**

答：

旧范式的假设检验的逻辑是：针对某个科学问题，构思一个假说（有关变量），假设变量之间存在因果关系，然后根据假说推演可验证的预言，再设计实验，收集数据，验证预言，最后依据验证结果，拒绝或接受假说。

①二分思维：传统假设检验是基于二元逻辑思维（接受或拒绝零假设），忽略了中间地带的信息，不能提供全面的分析。

Fisher的证伪主义：其假设检验的基本思想是小概率反证法，即构造一个小概率事件或原假说(H0)，同时提出另一个与原假说对立的备择假说(H1)。根据这个思想设计实验，即从总体中抽出样本，依据样本计算统计量(t、x2或F)的值，根据预设的概率水平(p<0.01或p<0.05)检验和选择的统计量的概率分布，确定原假说成立的可能性大小，做出拒绝或接受H0判断。

Fisher主张的H0假说是小概率事件，小概率事件在一次实验中不可能发生。但生态过程不断演化，无法预见原假说是小概率事件。因此，生态学的所有试验验证逻辑上应倾向于证实假说(H0)成立。

②假设条件限制：传统假设检验通常需要满足一些假设条件，如正态分布、独立性等，而现实数据往往不完全符合这些假设条件，导致结果的可靠性受到影响。

与物理、化学等学科不同，生态学问题的答案不是简单的“是”或“不是”，通过H0显著性检验来验证生态学假说并不严谨。生态学中本身可能无法确定H0的初始条件，甚至可能不存在H0，构建单一主导因子的H0很难。例如检验群落中竞争的作用，这无法通过一个简单实验进行证伪，因为群落中同时存在竞争、捕食和寄生、干扰等多因素作用。

③操控实验：操控实验包括正交实验设计法与析因法，无论哪种实验都需要遵循重复、随机化、区组化的原则。但是由于研究对象所处的环境是开放的，其中影响因子及相互关系难以被发现和确定。另外，在大尺度环境下，要采取严谨和一致操作才能排除无关干扰。

但生态学实验研究过程中，部分研究对象研究尺度过大，研究时间过长，在此之中，生态过程在不断演化，生态规律也在不断变化，很难进行实验的重复和一致。

④过度依赖于显著性水平：传统的假设检验通常只关注结果是否显著，而忽略了效应的大小和实际意义。这可能导致对结果的片面解释。

⑤样本量要求：传统假设检验通常对样本量有一定的要求，特别是在小效应或复杂模型下，可能需要大量样本才能得到可靠的结果。

⑥多重比较问题：在进行多个假设检验时，存在多重比较问题，容易导致统计推断的偏差，需要采取措施来控制错误发现率。

1. **What advances does the paradigm of data science used for ecological data?**

答：

①相关性分析：科学研究的目的是寻找现象之间的因果关系。

分析变量之间的相关性比探寻因果更重要，基于相关分析的预测是大数据研究的核心。按照科学哲学观点，如果一个命题能够解释以往出现的现象，又能预测未来可能出现的新现象和新问题，其科学性就得到了检验。一方面数据规律本身是从过去所积累的数据中挖掘出来的，完全可解释过去的现象或问题，另一方面，大数据包含了海量的各种现实数据，通过机器学习过去的经验来推测未来。因此，与因果律相比，基于大数据相关性的预测更准确，而且不易受偏见的影响。

②全数据归纳：样本≠总体，避免经验科学范式≠因果。

数据科学范式科学发现的逻辑起点不同于其他范式，主要表现在如下三个方面：第一，经验科学范式认为“科学始于观察”，即在自然和实验观察的基础上，通过归纳提炼出科学理论；假说-验证范式主张科学发现始于科学问题；而数据科学颠覆了原来的科学发现模式，从数据出发，利用数据挖掘方法发现数据中蕴含的规律性，形成了“科学始于数据”的新模式。

③利用非参数或半参数模型。

生态学研究强调生物与环境之间的关系，因此确定影响某个生态过程或现象的关键因子是首要任务。基于传统的统计分析方，探索数据分布特征，结合决策树、随机森林等分析特征的相关性，计算特征重要性，就可筛选出关键特征。利用提升回归树、朴素贝叶斯等建立分类与回归模型，用于生态预测或预警。基于关键特征，利用聚类分析方法，发现如空间上的hotspots、时间上的Motifs等模式。

④将数据科学范式与理论科学范式融合，避免归纳偏见。

数据科学采用归纳方法，而不是演绎逻辑。经验科学采用的是不完全归纳。数据科学范式沿袭了经验科学的归纳逻辑，不同的是数据科学采用的是全数据模式，即“样本=整体”的完全归纳法，克服了小样本不完归纳法的局限性，利于发现异常值。