1. **生态学主要采用的是“假说-验证”范式，面临的困境是什么？**

（1）通过H0显著性检验来验证生态学假说不严谨。例如，检验群落中竞争的作用，不是通过一个简单实验就可证伪的，因为群落中同时存在竞争，捕食和寄生、干扰等多因素作用。在生态学上构建单一主导因子的H0是很艰难的。

1. 生态过程是始终演化的，生态规律（因果关系）在不断变化。在逻辑上，生态学的所有实验验证倾向于“证实”假说（H0）成立，而不是严格的Fisher证伪结论。
2. 操控实验包括正交实验设计法与析因法，无论哪种实验，都遵循重复，随机化，区组化的原则。但研究对象所处的环境是开放的，影响因子及相互关系难以被发现和确定。
3. 在大尺度环境下，要采取严谨和一直操作才可能排除无关干扰。但是Anja Jaeschke 等认为无论如何精心设计和操控，都“不可能包括所有的多样性和空间”。Samuel Scheiner等认为野外实验存在难以重复的困难，不能作为假说的判决性实验。

**二．何为“数据科学”范式，与“假设-验证”范式主要区别是什么？**

1. ”数据科学”范式是指利用超级计算能力，直接分析海量数据发现相关关系，获得新知识。
2. 主要区别：
3. 科学发现的逻辑起点不同。经验范式认为“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。
4. 寻找相关性，而不是因果律。无论是经验科学范式，还是假说-验证范式，都认为科学研究的目的就是寻找现象之间的因果关系。与其他研究范式不同，数据科学范式认为分析变量之间的相关性比探寻因果律更重要，基于相关分析的预测是大数据研究的核心。
5. 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。例如，从达尔文的生物进化论到种群增长的Logistic方程都采用了归纳方法。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，克服了小样本不完全归纳法的局限性，利于发现异常值。

**三．什么是数据驱动的方法？与传统上哪种范式采用的方法相似？**

1. 数据驱动是指通过数据挖掘方法，找到数据或变量之间关联性，并作出科学发现。从数据中发现有意义的见解和知识，并将这种发现表示为模型。
2. 与传统的统计分析相似，数据挖掘采用的朴素贝叶斯分类就是统计理论的发展和延伸。常用于数据挖掘的主成分分析和回归分析也属于统计学范畴。但是二者有本质上的区别，传统数据分析需要对数据分布和变量间的关系做假设，确定用什么样的概率函数来描述变量间的关系做假设，确定用什么样的概率函数来描述变量间的关系，然后建立参数模型，并依据统计推断，确定模型的合理性，而数据挖掘并不需要对数据做任何假设，而是侧重利用机器学习方法，自动寻找变量间的关系。具体区别如下：

|  |  |  |
| --- | --- | --- |
|  | 数据挖掘 | 传统数据分析 |
| 前提条件 | 事先不清楚预测变量和响应变量之间关系 | 事先对预测变量和响应变量之间的关系作出假设 |
| 数据质量 | 对自变量多元共线性问题不敏感，不 考虑交互和非线性作用，含缺失值 | 对自变量多元共线性问题敏感，要求变量 的独立性，正态分布 |
| 主要目标 | 发现数据模式，提出新的假设并建模 | 对模型和假设进行验证 |
| 建模方法 | 机器学习，人工神经网络 | 模型参数估计，假说检验 |
| 模型评估 | 基于训练集和测试集的交叉验证 （预测百分率或AUC），通过高预测性检验模型 | 模型拟合（似然比检验或AIC值比较），基于理论验证 |
| 预测结果 | 预测准确性高 | 受到多种因素影响，不稳定 |
| 可解释性 | 不容易解释 | 参数有明确含义 |