**Homework-01 田晓妍 SA23008057 2024.03.03**

**The questions**

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

假说-验证的思想面临困境主要是*H0*不存在或很难以及判决性实验困难，具体包括以下几个方面：

(1) Ronald Fishe的假说-验证的思想是小概率反证法——即构造一个原假说*H0*（小概率事件)，同时提出另一个与原假说对立的备择假说*H1*，然后设计实验，即从总体中抽出样本，依据样本计算统计量的值，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。所以这个范式的假说不是被证实的，而是采纳证伪的结论，即证伪主义。

然而，生态学问题的答案不是简单的“是”或“不是”，所以通过*H0*显著性检验来验证生态学假说并不严谨。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*是很难的。Kristin Shrader - Frechette等认为不能确定*H0*的初始条件，Aaron Ellison 认为生态学中很少存在*H0*。

(2) 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*H0*成立，而不是严格的Fisher证伪结论。

(3) 假说-验证范式的另一个困境是操控实验，遵循重复、随机化、区组化的原则。由于研究对象所处的环境是开放的，鉴于生态学的特性，有限的试验不能严格地验证生态学假说，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可能排除无关干扰。可是， 生态学研究对象所处环境的开放性，很难达到近代物理学试验中的严格控制和排除其他干扰项的试验效果。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。因此，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。所以，生态学假说可能不适合用试验进行验证。

(4) 仅依据观察数据进行相关分析很难发现生态现象的因果关系。即使生态学试验结果支持了假说，也无法确证假说中的因果关系一定成立，即A原因并不一定导致a结果或a结果并不一定是由 A 原因导致。比如一个伪相关的例子，即荷兰出生婴儿数量与鹳雀巢数之间存在相关关系。事实上，两者之间没有关联，只是这两件事都与观测之前9个月的天气有关，天气成了干扰因素。显然，伪相关将导致错误的统计推断甚至误导科学发现。

综上，对假说-演绎模型进行分析，生态学面临很难存在严格被试验验证的原假说的困境。Fisher 的证伪主义与 Neyman-Pearson的非判决性使得统计学原假说不能被严格验证; 而生态过程中存在的不同于经典物理学的原假说*H0* ( α = 1，β = 0) 与不同的备假说*H1* ( α' = 1，β' = 0) 的情况，使得生态学原假说也很难得到严格的实验验证。但统计学的原假说显著性验证不应等同于生态学假说中有关因果关系的逻辑证明方法。因此，现有大量基于原假说显著性验证的生态学假说的方法研究和试验验证的结果与结论都不是绝对的逻辑可靠的。

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

数据科学颠覆了原来的科学发现模式，从长时间，大范围的数据出发，基于机器学习，从大规模、大尺度的数据中利用数据挖掘方法从数以百计的潜在因子中筛选出预测变量或特征，找到数据或变量之间的关联性，发现数据中蕴含的规律性，形成了"科学始于数据"的新模式。

数据科学范式相较于其他范式的优势主要表现在以下三个方面：

(1) 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。

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) 数据科学采用全数据归纳方法，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，克服了小样本不完全归纳法的局限性，利于发现异常值。

(3) 利用非参数或半参数模型。数据挖掘事先不清楚预测变量与响应变量的关系，对自变量和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；而传统的统计分析对多元共线性问题敏感，要求变量的独立性和正态分布，对模型参数估计和假设检验，但数据挖掘利用机器学习、人工神经网络等建模，基于训练集和测试集的交叉验证，通过高预测性检验模型，预测效率高。

(4) 将数据科学范式与理论科学范式融合，避免归纳偏见。在此过程中，首先产生一个假说的理论和知识，然后通过机器学习人类的知识，以此分析新的数据（全球或局域的大数据与本地的小数据）。这样，就可以使用每种范式单独使用的局限性。