作业1

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

在过去很长一段时间有关生态学的研究方面，主要采用的就是“假说-验证”范式。其基本探索逻辑为：针对某个存在的科学问题，构思一个假说，界定相关变量，假设变量之间存在着某些因果关系，然后根据假说推演出可以验证的预言，再设计实验，收集数据，验证预言，最后根据验证结果，判断这个假说正确与否。但是后来经过生态学方面的科学家不断探索，逐渐发现对于提出的假说，在实验验证阶段可能会出现很多问题，比如Ronald Fisher所提出的“假说-验证”思想就是证伪主义，构造的假说一般为小概率发生事件，在一次的实验中，基本上不太可能发生，但是不能如此轻易的将原假说证伪。因为在生态学的实验中很多假说的答案并不能单独的从一个方面去考虑，而是会涉及到很多因素，如此便是“假说-验证”范式在应用时的一个困境。

“假说-范式”的另外一个困境就是生态学会涉及很多在自然条件下进行的实验，这些实验的条件和各类影响因子并不能被完全解析，因此所得到的实验结果并不完全符合重复化、随机化及区组化，很多干扰因素不能完全排除，因此通过实验并不能很好的验证原有提出的假说。

此外，在“假说-验证”范式的应用中，还会存在着对于P值的错误理解、假阳性及样本不具有代表性等困境影响生态学家对于假说的理解与解答。

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

数据科学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。主要分为收集数据、数据预处理、探索性分析、数据建模、实证与评级及扩展等方面，涉及到传统的一些统计学方法以及基于机器学习的一些分类、聚类等方法。

其优势主要体现在三个方面：

1. 相关性分析：与其他经典的研究范式不同，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为重要。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。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，完全可以解释过去的现象和问题，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。因此基于大数据的相关性的预测更为准确，不易受偏见。
2. 全数据归纳，样本=总体，避免经验科学范式相关≠因果：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，克服了小样本不完全归纳法的局限性，利于发现异常值。
3. 科学发现的逻辑起点不同：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。