关于二维 poisson 过程与 K-S 检验的简单说明

2024Spring 空间统计与分析

童川博

中国地质大学未来技术学院

2024 年 3 月 11 日

随机实验, 随机变量, 随机事件, 随机过程

一个也许必要的回顾

- 随机实验: 在相同的条件下可以多次重复进行的过程, 其结果是不确定的, 无法预测的.
- 随机事件: 随机实验的某种结果或一组结果的集合, 也就是样本空间中的一个子集. 每个随机事件都有一个相应的概率与之关联.
- **随机过程 (Random Process): 在一定的时间范围内随机变化的数学模型, 是随机事件随时间的演变. 随机过程可以是离散的 (如随机游走) 或连续的 (如布朗运动).**
- 随机分布: 随机变量可能取得不同值的概率分布, 描述了随机变量的可能取值以及每个取值的概率.

When comes to random

当我们在讨论‘随机‘的时候，其本质内涵是讨论一个随机实验，而非结果呈现随机性.

随机事件是一种" 随机" **模式**，随机过程是一种" 随机" **过程**.

在之前的课程中比较少接触到随机过程，但是这其实是一个很重要的视角.

现在**包含着点信息的地图就是一个模式**，我们能不能回答他的过程?

转换视角: 从二维平面到一维时间

完全随机的点模式分布就像是在地图上随机地撒点, 我们自然而然地认为这是一次随机事件, 然而回想随机事件的定义, 会发现**此处安置一个点是随机事件, 安置地图上的全部点是一个随机过程.**

请务必注意, 虽然地图是事先拿到的, 但是点的数量是我们事先绝对不知道的, 即这个过程不是抓一把点再撒, 而是拿起一个点时再决定是 (按某种约定) 放下去还是扔掉. 请与确定性过程相区分

设想一个扫描线, 从二维空间中的左上角开始, 逐行扫描同时记录经过的点的数量, 如此我们就将二维平面上点的数量统计转换为随扫描线推进的数量统计.

> 请原谅我没有给出二维泊松过程进一步的表述和扫描线的相关细节, 因为他在后续的讨论中无关紧要.

IRP 的两个假设条件

Definition of IRP

Formally, the IRP postulates two conditions:

- **The condition of equal probability.** This states that any event has an equal probability of being in any position or, equivalently, that each small subarea of the map has an equal chance of receiving an event.
- **The condition of independence.**This states that the positioning of any event is independent of the positioning of any other event.

实际上有一些课程和教材会在概率论的后半段介绍 *Markov* 链, 我觉得这其实是统计一个必要的前置视角.

请同样原谅我没有给出 *Markov chain* 的定义, 因为其基本思想已经被阐明.

一个被概率论课程忽略的内容---泊松过程

Definition of Poisson process

A Poisson process is a stochastic process which counts the number of events in a fixed interval of time or space.

- The number of events in non-overlapping intervals are independent.
- The probability of exactly one event occurring in a small interval is proportional to the length of the interval.
- The probability of more than one event occurring in a small interval is negligible.

与前文概念对比，我们可以发现，泊松过程的两个假设条件与 IRP 的两个假设条件是一致的. 所以我们可以断言，完全随机的点模式是 poisson 过程，其概率分布满足 poisson 分布.

K-S test is a nonparametric test

我们之前在课堂上接触的都是参数检验, 比如 t 检验, 卡方检验 etc.
K-S test 是一种非参数检验, 用于检验样本是否来自某一特定的分布而不是事先假定其服从某种 (如正态) 分布.
此处我们要检验这个样本是否服从 poisson 分布.

如果实在好奇定义的话...

The Kolmogorov-Smirnov test (K-S test or KS test) is a nonparametric test of the equality of continuous (or discontinuous, one-dimensional probability distributions that can be used to test whether a sample came from a given reference probability distribution

K-S 检验的统计量

此处给出表达式只是提供一个直觉和印象, 摘自 Wikipedia.
The empirical distribution function F_n for n independent and identically distributed(i.i.d.) ordered observations X_i is defined as

$$F_n(x) = \frac{}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} 1_{(-,x]}(X_i) \tag{1}$$

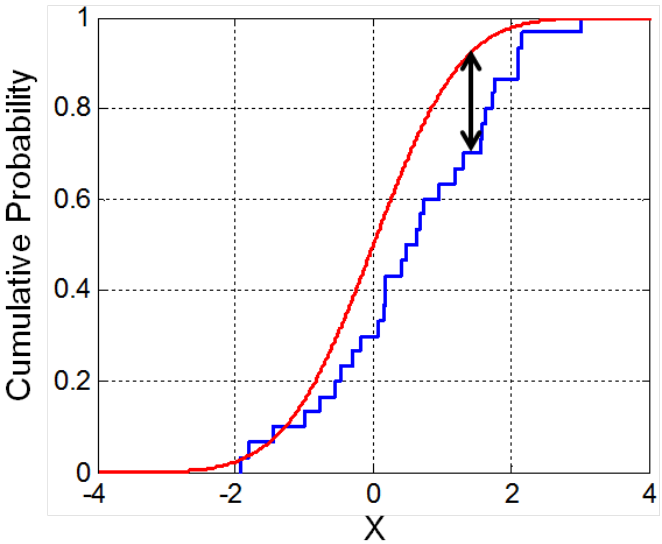
where $1_{(-,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and 0 otherwise.
The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)| \tag{2}$$

where $\sup_x$ is the supermum of the set of distances.

一个滥用: 非参检验是一种 duck test

If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.



K-S 检验的统计量 D 用于检查这样一个假设是否正确: **在随机事件不断发生, 即事件集合扩大的过程中, 当前事件集合的 CDF 与假设分布的 CDF 一致.**
如果整个随机过程中,model CDF 和 empirical CDF 的最大差值都小于某个阈值, 那么我们就可以认为这个样本是来自这个分布的.

P.S. 这大概绝对不是一个严肃的理解, 大佬轻喷 QAQ.

回头看 K-S test 是如何达成目标的

由 *section 1* 我们知道, 完全随机的点模式分布是 poisson 分布.
由 *section 2* 我们知道,K-S 检验可以用于检测样本是否来自某一特定的分布.
我们在做的是这么一件事: 根据检验随机过程的分布断言随机过程
因此, 我们可以用 K-S 检验来检测点模式是否符合 poisson 分布, 进而检验点模式是否符合随机分布.

泊松分布一定由泊松过程产生吗---关于 MAUP 的问题

这个逻辑有一个无法忽视的缺陷: **泊松分布一定由泊松过程产生吗?**

有没有可能其实其产生是遵循某种确定性原理, 但是在关心的尺度上呈现出随机性?

如果我们统计整个城市的五金店分布, 在街道尺度我们也许很容易能分辨他的模式, 但是在城市尺度上我们只能得到一个几乎随机的模式.

如果我们只从统计的角度, 而不参考道路, 建筑等地理要素, 这样的认识将会进一步模糊.

或者相反的, 一些随机过程在另一个尺度上可能呈现出确定性模式 (在现代物理学中有丰富的例子).

有时候可能不能要求太多...

气候学家常常为此受到诟病. 众所周知, 任何气象过程都是完全受到理化过程控制的, 但是面对非常复杂的系统, 其内部交织的复杂机制使得其模式辨别变得非常困难.

一套基于数值模拟数值统计的方法不一定能有助于我们理解其过程, 但是在此基础上发展起来的预报系统具有非常重要的意义.

虽然天气预报经常受到诟病, 但你就说要不要吧.

所以, 我们也许不能对统计能完成的任务要求太多, 其提供的信息就足以启发我们进一步的地理认知.