复习与提示
大作业安排
probe 大作业
泊松点过程
probe 历史经验
机器学习
可微编程
概率编程
SQL
总复习
现实世界的大数据方法
下一步学习

大作业与未来方向

续本达

清华大学 工程物理系

2024-07-28 清华

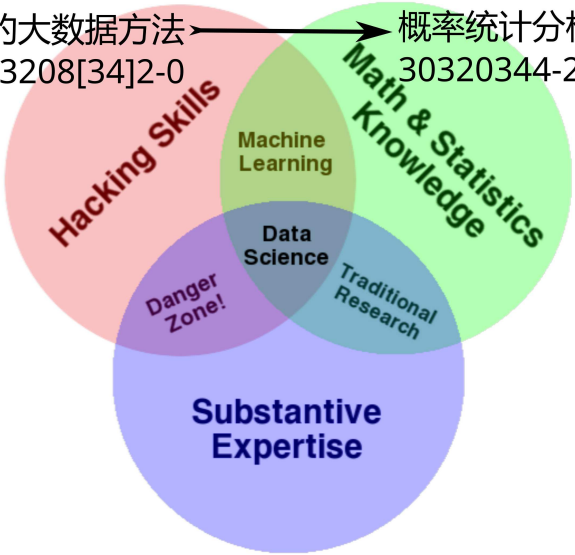
广义线性回归

1. 关系代数让回归分析变得极其直观，让我们专注于问题的本质
2. 模型的选择极其重要，应当使用客观标准
   - AIC 是最简单直接的客观标准
   - 找到“最佳模型”的过程充满曲折，是实验“研究”的主体过程
3. 广义线性回归，把误差分布从高斯替换为其它指数族分布
   - 把连接函数从恒等替换为非线性函数
   - 几乎可以解决所有日常工作中的非线性问题，惊喜！

数据时代的物理技能

实验物理的大数据方法 → 概率统计分析及量测技术

夏 403208[34]2-0 30320344-2 秋



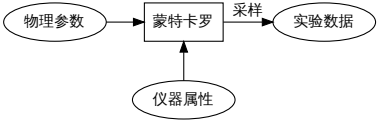
概率是逻辑的扩展 – Cox 定理

- Laplace: probability theory is nothing but common sense reduced to calculations.
- R. T. Cox, E. T. Jaynes，两位对统计学有重大贡献的物理学家
- Logical interpretation of probability
  1. Divisibility and comparability – The plausibility of a proposition is a real number and is dependent on information we have related to the proposition.
  2. Common sense – Plausibilities should vary sensibly with the assessment of plausibilities in the model.
  3. Consistency – If the plausibility of a proposition can be derived in many ways, all the results must be equal.
- 概率论的唯一性：Any system for plausible reasoning that satisfies certain qualitative requirements intended to ensure consistency with classical deductive logic and correspondence with commonsense reasoning is isomorphic to probability theory.

参考：Van Horn, K.S., 2003. Constructing a logic of plausible inference: a guide to Cox’s theorem. International Journal of Approximate Reasoning 34, 3 – 24.

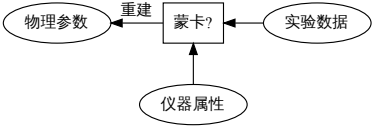
大作业安排

正向第一阶段



- 2024-07-11 – 2024-08-02
- 模拟实验测量

逆向第二阶段



- 2024-07-28 – 2024-08-18
- 分析数据
- 测量模型参数
- 黑盒分数按排名

实验测量的分析

输入 （模拟的）实验测量原始数据

- 如果不是模拟的，则无法评分
- 但是助教会尽可能把它做得和真的一样

输出 物理对象的数学描述，实现

采分 与助教手中的模拟输入相比

第二阶段分组

- 每队至多三人
  - 单人队：大作业得分 $\times 1.03$
  - 三人队：每人大作业得分 $=$ 队伍得分 $\times 0.95$
  - 不同队伍间请勿直接交换代码
- 如果大作业结果含有学术突破，总评保送 A+。

实验测量的分析

输入 （模拟的）实验测量原始数据

- 如果不是模拟的，则无法评分
- 但是助教会尽可能把它做得和真的一样

输出 物理对象的数学描述，实现

采分 与助教手中的模拟输入相比

第二阶段分组

- 每队至多三人
  - 单人队：大作业得分 $\times 1.03$
  - 三人队：每人大作业得分 $=$ 队伍得分 $\times 0.95$
  - 不同队伍间请勿直接交换代码
- 如果大作业结果含有学术突破，总评保送 A+。

大作业的可复现性

复现 原则的要求

大作的结果必须可复现，否则无效。

思路和要点

- 将获得 `https://git.tsinghua.edu.cn/physics-data/2024/project_2` 之下的仓库一份，使用 GNU Make 构建整个分析流程，连同报告、程序整理到仓库中。
  - 注意小组分工中 Git 使用的规范
  - 善于使用 Git `branch`, Gitlab `merge request` 等团队协作功能
- 把流程系统化成输入、输出与过程三要素。
  - 而向数据编程，data-driven programming
- 系统表达输入数据、输出数据和中间结果的依赖关系，
  - 成为“可执行的说明文档”

具体说明

https://git.tsinghua.edu.cn/physics-data/projects/tpl_junoprobe

- 正在完善
  - 大家可以开始 clone 大作业模板，再 push 到自己的仓库中。
  - 后续如果模板有更新，通过 git merge 整合。

复习与提示
大作业安排
probe 大作业
泊松点过程
probe 历史经验
机器学习
可微编程
概率编程
SQL
总复习
现实世界的大数据方法
下一步学习

泊松点过程

- 作业的目标是取得 JUNO 探测器泊松点过程的条件均值函数。

likelihood-free inference

- "likelihood-free inference" somewhat misleading
- 从一个没有 likelihood 的模型中，提炼出 likelihood。

随机过程概念

- 随机过程的研究对象是 随时间演变的随机现象 。
- 不能用随机变量或多维随机变量来合理表达，而需要用一族无限多个随机变量来描述。

定义

随机过程是一族随机变量 $\{X(t) : t \in T\}$, 其中 t 是 参数 ,T 称为 参数集 .

- 一般地, t 表示时间。对于每一个 $t \in T$ ，$X(t)$ 是一个随机变量，称 $X(t)$ 为时刻 t 时 过程的状态 。
- $X(t)$ 所有可能取值的全体称为随机过程的 状态空间 。

随机过程概念

- 随机过程的研究对象是 随时间演变的随机现象 。
- 不能用随机变量或多维随机变量来合理表达，而需要用一族无限多个随机变量来描述。

定义

随机过程是一族随机变量 $\{X(t) : t \in T\}$, 其中 t 是 参数 ,T 称为 参数集 .

- 一般地, t 表示时间。对于每一个 $t \in T$ ，$X(t)$ 是一个随机变量，称 $X(t)$ 为时刻 t 时 过程的状态 。
- $X(t)$ 所有可能取值的全体称为随机过程的 状态空间 。

随机过程概念

- 随机过程的研究对象是 随时间演变的随机现象 。
- 不能用随机变量或多维随机变量来合理表达，而需要用一族无限多个随机变量来描述。

定义

随机过程是一族随机变量 $\{X(t) : t \in T\}$, 其中 t 是 参数 ,T 称为 参数集 .

- 一般地, t 表示时间。对于每一个 $t \in T$ ，$X(t)$ 是一个随机变量，称 $X(t)$ 为时刻 t 时 过程的状态 。
- $X(t)$ 所有可能取值的全体称为随机过程的 状态空间 。

随机过程概念

- 随机过程的研究对象是 随时间演变的随机现象 。
- 不能用随机变量或多维随机变量来合理表达，而需要用一族无限多个随机变量来描述。

定义

随机过程是一族随机变量 $\{X(t) : t \in T\}$, 其中 t 是 参数 ,T 称为 参数集 .

- 一般地, t 表示时间。对于每一个 $t \in T$ ，$X(t)$ 是一个随机变量，称 $X(t)$ 为时刻 t 时 过程的状态 。
- $X(t)$ 所有可能取值的全体称为随机过程的 状态空间 。

样本函数

定义

对随机过程 $\{X(t) : t \in T\}$ 进行一次试验（即在 T 上进行一次全程观测），其结果是 t 的函数，记为 $\{x(t) : t \in T\}$，称为随机过程的一个 样本函数 。

随机过程观测获得样本函数，如同总体观测获得个体样本

符号化

把随机过程 $\{X(t), t \in T\}$ 写成

$$\{X(\omega, t) : \omega \in \Omega, t \in T\}$$

的形式，其中 ω, Ω 分别是随机试验的 样本点 和 样本空间 。

- 固定一个时间 t_0，随机过程对应于一个随机变量 $X(t_0)$。
- 固定 $\omega_0 \in \Omega$ 让 t 在 T 中变化，$X(\omega_0, t)$ 是定义 在 T 上的一个实函数，称之为对应于 ω_0 的一个 样本函数 或者 样本轨道 。

样本函数

定义

对随机过程 $\{X(t) : t \in T\}$ 进行一次试验（即在 T 上进行一次全程观测），其结果是 t 的函数，记为 $\{x(t) : t \in T\}$，称为随机过程的一个 样本函数 。

随机过程观测获得样本函数，如同总体观测获得个体样本

符号化

把随机过程 $\{X(t), t \in T\}$ 写成

$$\{X(\omega, t) : \omega \in \Omega, t \in T\}$$

的形式，其中 ω, Ω 分别是随机试验的 样本点 和 样本空间 。

- 固定一个时间 t_0，随机过程对应于一个随机变量 $X(t_0)$。
- 固定 $\omega_0 \in \Omega$ 让 t 在 T 中变化，$X(\omega_0, t)$ 是定义 在 T 上的一个实函数，称之为对应于 ω_0 的一个 样本函数 或者 样本轨道 。

样本函数

定义

对随机过程 $\{X(t) : t \in T\}$ 进行一次试验（即在 T 上进行一次全程观测），其结果是 t 的函数，记为 $\{x(t) : t \in T\}$，称为随机过程的一个 样本函数 。

随机过程观测获得样本函数，如同总体观测获得个体样本

符号化

把随机过程 $\{X(t), t \in T\}$ 写成

$$\{X(\omega, t) : \omega \in \Omega, t \in T\}$$

的形式，其中 ω, Ω 分别是随机试验的 样本点 和 样本空间 。

- 固定一个时间 t_0，随机过程对应于一个随机变量 $X(t_0)$。
- 固定 $\omega_0 \in \Omega$ 让 t 在 T 中变化，$X(\omega_0, t)$ 是定义 在 T 上的一个实函数，称之为对应于 ω_0 的一个 样本函数 或者 样本轨道 。

样本函数

定义

对随机过程 $\{X(t) : t \in T\}$ 进行一次试验（即在 T 上进行一次全程观测），其结果是 t 的函数，记为 $\{x(t) : t \in T\}$，称为随机过程的一个 样本函数 。

随机过程观测获得样本函数，如同总体观测获得个体样本

符号化

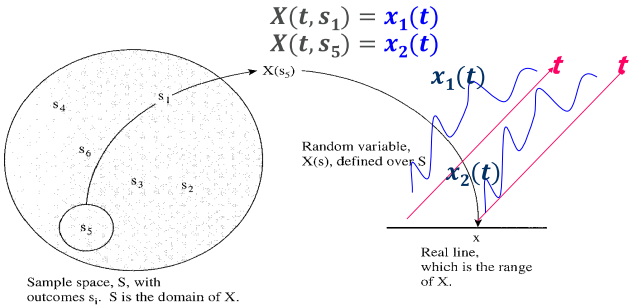
把随机过程 $\{X(t), t \in T\}$ 写成

$$\{X(\omega, t) : \omega \in \Omega, t \in T\}$$

的形式，其中 ω, Ω 分别是随机试验的 样本点 和 样本空间 。

- 固定一个时间 t_0，随机过程对应于一个随机变量 $X(t_0)$。
- 固定 $\omega_0 \in \Omega$ 让 t 在 T 中变化，$X(\omega_0, t)$ 是定义 在 T 上的一个实函数，称之为对应于 ω_0 的一个 样本函数 或者 样本轨道 。

图示

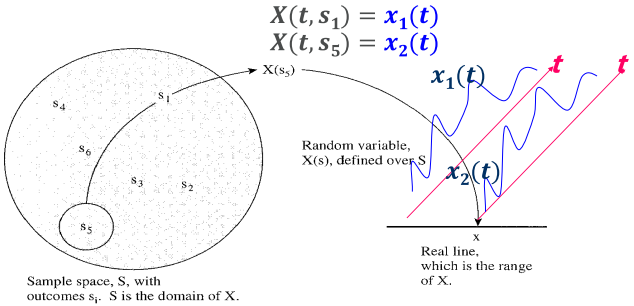


$$X(\omega, t) \equiv \Omega \ni \omega \to X(t)$$

随机过程 $\{X(\omega, t) : \omega \in \Omega, t \in T\}$ 四种不同情况下的意义：

$X(\omega, t)$	t 固定	t 可变
ω 固定	确定值	样本函数
ω 可变	随机变量	随机过程

图示



$$X(\omega, t) \equiv \Omega \ni \omega \to X(t)$$

随机过程 $\{X(\omega, t) : \omega \in \Omega, t \in T\}$ 四种不同情况下的意义：

$X(\omega, t)$	t 固定	t 可变
ω 固定	确定值	样本函数
ω 可变	随机变量	随机过程

分布函数族

定义

设随机过程 $\{X(t): t \in T\}$，对每一个固定的 $t \in T$，

$$F_X(x,t) = P(X(t) \leq x), x \in \mathbb{R}$$

称为随机过程 $\{X(t): t \in T\}$ 的一维分布函数，$\{F_X(x,t): t \in T\}$ 称为 一维分布函数族 。

随机过程的数字特征

给定随机过程 $\{X_\omega(t), \omega \in \Omega, t \in T\}$:

均值函数 $\mu_X(t) = \mathrm{E}_\omega[X_\omega(t)]$

均方值函数 $\Psi_X^2(t) = \mathrm{E}_\omega[X_\omega^2(t)]$

方差函数 $\sigma_X^2(t) = \mathrm{Var}_X(t) = \mathrm{E}_\omega[X_\omega(t) - \mu_X(t)]^2$

标准差函数 $\sigma_X(t) = \sqrt{\mathrm{Var}_X(t)}$

相关函数 $R_X(s,t) = \mathrm{E}_\omega[X_\omega(s)X_\omega(t)]$

$$R_X(t,t) = \Psi_X^2(t)$$

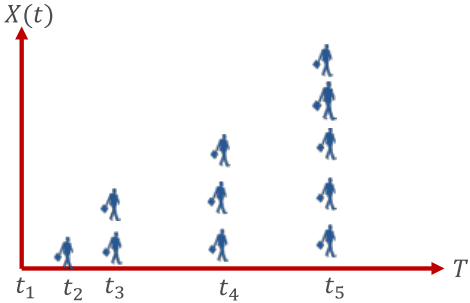
协方差函数 $C_X(s,t) = \mathrm{Cov}[X(s), X(t)]$

例：独立增量过程

阿茶扫过的人数

设 $X(t)$ 为截至 t 时刻，阿茶通过自动测温仪的累计总人数。

- 增量 $X(t_2)-X(t_1)$ 表示在时间区间 $(t_1, t_2]$ 测量体温的人数。
- 不相交的时间区间的增量是相互独立的。

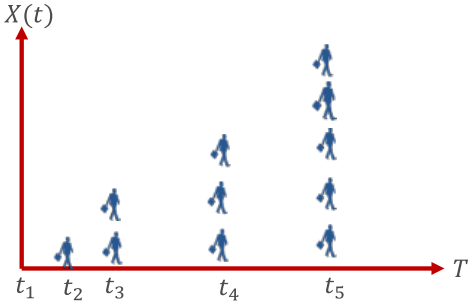


例：独立增量过程

阿茶扫过的人数

设 $X(t)$ 为截至 t 时刻，阿茶通过自动测温仪的累计总人数。

- 增量 $X(t_2)-X(t_1)$ 表示在时间区间 $(t_1, t_2]$ 测量体温的人数。
- 不相交的时间区间的增量是相互独立的。



独立增量过程：定义

独立增量过程

设 $\{X(t), t \in T\}$ 是一随机过程，若 $\forall t_1 < \cdots < t_n (n \geq 2, t_i \in T)$ 诸增量

$$X(t_2) - X(t_1), X(t_3) - X(t_2), \cdots, X(t_n) - X(t_{n-1})$$

相互独立，则称 $\{X(t), t \in T\}$ 是一个 独立增量过程 。

性质

若 $\{X(t) : t \in T\}$ 是独立增量过程，且 $X(0) = 0$，则：

1. $X(t)$ 的有限维分布函数族可以由增量 $X(t) - X(s) (0 \leq s \leq t)$ 的分布所确定
2. 设 $\mathrm{Var}_X(t)$ 已知，则 $C_X(s,t) = \mathrm{Var}_X[\min(s,t)]$

独立增量过程：定义

独立增量过程

设 $\{X(t), t \in T\}$ 是一随机过程，若 $\forall t_1 < \cdots < t_n (n \geq 2, t_i \in T)$ 诸增量

$$X(t_2) - X(t_1), X(t_3) - X(t_2), \cdots, X(t_n) - X(t_{n-1})$$

相互独立，则称 $\{X(t), t \in T\}$ 是一个 独立增量过程 。

性质

若 $\{X(t) : t \in T\}$ 是独立增量过程，且 $X(0) = 0$，则：

1. $X(t)$ 的有限维分布函数族可以由增量 $X(t) - X(s) (0 \leq s \leq t)$ 的分布所确定
2. 设 $\mathrm{Var}_X(t)$ 已知，则 $C_X(s,t) = \mathrm{Var}_X[\min(s,t)]$

泊松过程案例特点

随机点过程 如“来到银行要求服务的顾客流”，“在一段时间内机器故障产生的故障流”等等。

计数过程 $N(t)$ 表示在 $[0, t]$ 内随机点 (事件) 发生的数目，$N(t)$ 即为一计数过程。如，阿荼测体温的例子中，$X(t)$ 表示截至 t 时刻累计测体温的人数。

泊松过程案例特点

随机点过程 如“来到银行要求服务的顾客流”，“在一段时间内机器故障产生的故障流”等等。

计数过程 $N(t)$ 表示在 $[0, t]$ 内随机点 (事件) 发生的数目，$N(t)$ 即为一计数过程。如，阿荼测体温的例子中，$X(t)$ 表示截至 t 时刻累计测体温的人数。

泊松过程定义

泊松过程

计数过程 $\{N(t), t \geq 0\}$ 称为强度 λ (λ >0) 的 时齐泊松过程 (homogeneous Poisson process)，若满足：

1. $N(0) = 0$
2. $\{N(t), t \geq 0\}$ 是 时齐 的 独立增量过程
3. 对任意 $t > 0$ 和充分小的 $\Delta t > 0$, 有
   - $P[N(t + \Delta t) - N(t) = 1] = \lambda \Delta t + o(\Delta t)$
   - $P[N(t + \Delta t) - N(t) \geq 2] = o(\Delta t)$

   其中 $o(\Delta t)$ 是 Δt 的无穷小，即 $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$

上述过程称为 泊松过程 ，因为任意长度为 t 的区间中时间个数服从均值 λt 的泊松分布。

泊松过程定义

泊松过程

计数过程 $\{N(t), t \geq 0\}$ 称为强度 λ (λ >0) 的 时齐泊松过程 (homogeneous Poisson process)，若满足：

1. $N(0) = 0$
2. $\{N(t), t \geq 0\}$ 是 时齐 的 独立增量过程
3. 对任意 $t > 0$ 和充分小的 $\Delta t > 0$, 有
   - $P[N(t + \Delta t) - N(t) = 1] = \lambda \Delta t + o(\Delta t)$
   - $P[N(t + \Delta t) - N(t) \geq 2] = o(\Delta t)$

   其中 $o(\Delta t)$ 是 Δt 的无穷小，即 $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$

上述过程称为 泊松过程 ，因为任意长度为 t 的区间中时间个数服从均值 λt 的泊松分布。

泊松过程定义

泊松过程

计数过程 $\{N(t), t \geq 0\}$ 称为强度 λ (λ >0) 的 时齐泊松过程 (homogeneous Poisson process)，若满足：

1. $N(0) = 0$
2. $\{N(t), t \geq 0\}$ 是 时齐 的 独立增量过程
3. 对任意 $t > 0$ 和充分小的 $\Delta t > 0$, 有
   - $P[N(t + \Delta t) - N(t) = 1] = \lambda \Delta t + o(\Delta t)$
   - $P[N(t + \Delta t) - N(t) \geq 2] = o(\Delta t)$

   其中 $o(\Delta t)$ 是 Δt 的无穷小，即 $\lim\limits_{\Delta t \to 0} \dfrac{o(\Delta t)}{\Delta t} = 0$

上述过程称为 泊松过程 ，因为任意长度为 t 的区间中时间个数服从均值 λt 的泊松分布。

定理

例

设 $\{N(t): t \geq 0\}$ 是一个强度为 λ 的泊松过程，则 $N(t)$ 的一维分布是参数为 λt 的泊松分布, 即对任意 $t > 0$ 有

$$P[N(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, k = 0, 1, 2, \cdots$$

数字特征：

$$\begin{aligned} \mu_N(t) &= \lambda t \\ \mathrm{Var}_N(t) &= \lambda t \\ C_N(s,t) &= \lambda \min(s,t) \end{aligned}$$

定理

例

设 $\{N(t): t \geq 0\}$ 是一个强度为 λ 的泊松过程，则 $N(t)$ 的一维分布是参数为 λt 的泊松分布, 即对任意 $t > 0$ 有

$$P[N(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, k = 0, 1, 2, \cdots$$

数字特征：

$$\begin{aligned} \mu_N(t) =& \lambda t \\ \mathrm{Var}_N(t) =& \lambda t \\ C_N(s,t) =& \lambda \min(s,t) \end{aligned}$$

二项分布视角

- 将区间 $[0, t]$ 分为 k 份。
- 当 $k \to \infty$ 时，根据定义有 $P[N(s+\Delta t) - N(s) \geq 2] = o(\Delta t)$，即任意一个区间内包含两个或者两个以上的事件概率趋于 0.
- $N(t)$（以趋于 1 的概率）等于恰好含有一个事件的子区间的个数，服从二项分布 (k, p) 且 $p = \dfrac{\lambda t}{k} + o\left(\dfrac{t}{k}\right)$
- $N(t)$ 服从参数为 (k, p) 的二项分布，均值函数

$$\mu_N(t) = \lim_{k \to \infty} kp = \lim_{k \to \infty} k \left[\frac{\lambda t}{k} + o\left(\frac{t}{k}\right)\right] = \lambda t + o(t)$$

二项分布视角

- 将区间 $[0, t]$ 分为 k 份。
- 当 $k \to \infty$ 时，根据定义有 $P[N(s+\Delta t) - N(s) \geq 2] = o(\Delta t)$，即任意一个区间内包含两个或者两个以上的事件概率趋于 0.
- $N(t)$（以趋于 1 的概率）等于恰好含有一个事件的子区间的个数，服从二项分布 (k, p) 且 $p = \frac{\lambda t}{k} + o\left(\frac{t}{k}\right)$
- $N(t)$ 服从参数为 (k, p) 的二项分布，均值函数

$$\mu_N(t) = \lim_{k \to \infty} kp = \lim_{k \to \infty} k \left[\frac{\lambda t}{k} + o\left(\frac{t}{k}\right)\right] = \lambda t + o(t)$$

二项分布视角

- 将区间 $[0,t]$ 分为 k 份。
- 当 $k \to \infty$ 时，根据定义有 $P[N(s+\Delta t)-N(s) \geq 2] = o(\Delta t)$，即任意一个区间内包含两个或者两个以上的事件概率趋于 0.
- $N(t)$（以趋于 1 的概率）等于恰好含有一个事件的子区间的个数，服从二项分布 (k,p) 且 $p = \dfrac{\lambda t}{k} + o\left(\dfrac{t}{k}\right)$
- $N(t)$ 服从参数为 (k,p) 的二项分布，均值函数

$$\mu_N(t) = \lim_{k\to\infty} kp = \lim_{k\to\infty} k\left[\frac{\lambda t}{k} + o\left(\frac{t}{k}\right)\right] = \lambda t + o(t)$$

二项分布视角

- 将区间 $[0,t]$ 分为 k 份。
- 当 $k \to \infty$ 时，根据定义有 $P[N(s+\Delta t)-N(s) \geq 2] = o(\Delta t)$，即任意一个区间内包含两个或者两个以上的事件概率趋于 0.
- $N(t)$（以趋于 1 的概率）等于恰好含有一个事件的子区间的个数，服从二项分布 (k,p) 且 $p = \dfrac{\lambda t}{k} + o\left(\dfrac{t}{k}\right)$
- $N(t)$ 服从参数为 (k,p) 的二项分布，均值函数

$$\mu_N(t) = \lim_{k\to\infty} kp = \lim_{k\to\infty} k\left[\frac{\lambda t}{k} + o\left(\frac{t}{k}\right)\right] = \lambda t + o(t)$$

指数分布视角

- 将泊松过程第 n 个事件发生时刻记为 T_n。
- 记 $W_n = T_n - T_{n-1}$ 为第 n 个事件的等待时间，特别地 $W_1 = T_1$。

定理 (泊松分布的指数间隔)

$W_n (n = 1, 2, \cdots)$ 独立地服从参数为 λ 的指数分布。

- 考虑 事件 $\{W_1 > t\}$，即区间 $[0, t]$ 中无事件，根据定义有

$$P(W_1 > t) = P(N(t) = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

- 由全概率公式，$P(W_2 > t) = \mathrm{E}_{T_1}[P(T_2 > t | T_1)]$

$$\begin{aligned} P(W_2 > t | T_1 = s) &= P[(s, s+t] \text{中无事件} | T_1 = s] \\ &= P[(s, s+t] \text{中无事件}] (\text{由独立增量}) = e^{-\lambda t} \end{aligned}$$

指数分布视角

- 将泊松过程第 n 个事件发生时刻记为 T_n。
- 记 $W_n = T_n - T_{n-1}$ 为第 n 个事件的等待时间，特别地 $W_1 = T_1$。

定理 (泊松分布的指数间隔)

$W_n (n = 1, 2, \cdots)$ 独立地服从参数为 λ 的指数分布。

- 考虑 事件 $\{W_1 > t\}$，即区间 $[0, t]$ 中无事件，根据定义有

$$P(W_1 > t) = P(N(t) = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

- 由全概率公式，$P(W_2 > t) = \mathrm{E}_{T_1}[P(T_2 > t | T_1)]$

$$\begin{aligned} P(W_2 > t | T_1 = s) &= P[(s, s+t] \text{中无事件} | T_1 = s] \\ &= P[(s, s+t] \text{中无事件}] (\text{由独立增量}) = e^{-\lambda t} \end{aligned}$$

指数分布视角

- 将泊松过程第 n 个事件发生时刻记为 T_n。
- 记 $W_n = T_n - T_{n-1}$ 为第 n 个事件的等待时间，特别地 $W_1 = T_1$。

定理 (泊松分布的指数间隔)

$W_n (n = 1, 2, \cdots)$ 独立地服从参数为 λ 的指数分布。

- 考虑 事件 $\{W_1 > t\}$，即区间 $[0, t]$ 中无事件，根据定义有

$$P(W_1 > t) = P(N(t) = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

- 由全概率公式，$P(W_2 > t) = \mathrm{E}_{T_1}[P(T_2 > t | T_1)]$

$$\begin{aligned} P(W_2 > t | T_1 = s) &= P[(s, s+t]\text{中无事件} | T_1 = s] \\ &= P[(s, s+t]\text{中无事件}]\text{(由独立增量)} = e^{-\lambda t} \end{aligned}$$

指数分布视角

- 将泊松过程第 n 个事件发生时刻记为 T_n。
- 记 $W_n = T_n - T_{n-1}$ 为第 n 个事件的等待时间，特别地 $W_1 = T_1$。

定理 (泊松分布的指数间隔)

$W_n (n = 1, 2, \cdots)$ 独立地服从参数为 λ 的指数分布。

- 考虑 事件 $\{W_1 > t\}$，即区间 $[0, t]$ 中无事件，根据定义有

$$P(W_1 > t) = P(N(t) = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}$$

- 由全概率公式，$P(W_2 > t) = \mathrm{E}_{T_1}[P(T_2 > t | T_1)]$

$$\begin{aligned} P(W_2 > t | T_1 = s) =& P[(s, s+t]\text{中无事件} | T_1 = s] \\ =& P[(s, s+t]\text{中无事件}]\text{(由独立增量)} = e^{-\lambda t} \end{aligned}$$

逆命题成立

定理 (泊松过程判定)

给定参数为 λ 的独立同分布指数随机变量列 $\{W_n\}$。称第 n 个事件在时间 $T_n = W_1 + W_2 + \cdots + W_n$ 发生 $(S_0 = 0)$，得到

$$N(t) = \max(n|T_n \leq t), t \geq 0$$

是强度为 λ 的时齐泊松过程。

例：泊松过程的二维分布

例

设 $\{N(t), t \geq 0\}$ 是一个强度为 λ 的泊松过程，求 $N(t)$ 的二维分布。

对任意 $t \geq s > 0$，

$$
\begin{aligned}
& P[N(s)=i, N(t)=j] \\
=& P[N(s)=i, N(t)-N(s)=j-i] \\
=& P[N(s)=i]P[N(t)-N(s)=j-i]\text{：独立增量性} \\
=& P[N(s)=i]P[N(t-s)=j-i] \\
=& \frac{(\lambda s)^i}{i!}e^{-\lambda s}\frac{[\lambda(t-s)]^{j-i}}{(j-i)!}e^{-\lambda(t-s)}
\end{aligned}
$$

例：泊松过程的二维分布

例

设 $\{N(t), t \geq 0\}$ 是一个强度为 λ 的泊松过程，求 $N(t)$ 的二维分布。

对任意 $t \geq s > 0$，

$$
\begin{aligned}
& P[N(s)=i, N(t)=j] \\
=& P[N(s)=i, N(t)-N(s)=j-i] \\
=& P[N(s)=i]P[N(t)-N(s)=j-i]\text{：独立增量性} \\
=& P[N(s)=i]P[N(t-s)=j-i] \\
=& \frac{(\lambda s)^i}{i!}e^{-\lambda s}\frac{[\lambda(t-s)]^{j-i}}{(j-i)!}e^{-\lambda(t-s)}
\end{aligned}
$$

非齐次泊松过程

从常数的增量，拓展为任意非负函数的增量。

$$\mu_N(t) = \lambda t = \int_0^t \lambda \mathrm{d}s$$

$$\to \mu_N(t) \int_0^t R(s)\mathrm{d}s$$

- 方差函数等于均值函数。

$$\mathrm{Var}_N(t) = \mu_N(t)$$

Probe 的回顾

- 理念（张爱强）：使用 Poisson 点过程，统一探测器的时间和能量响应

流派

- 套用既有轮子的 Poisson 回归（王宇逸）
  - 优点：易于理解
  - 缺点：算不出来
- 函数基拟合（续本达）
  - 种类：指数型（窦威）、平方型（武益阳）、混合型（刘学伟）、棍子状（徐闯）
  - 拟合方法：Markov Chain Monte Carlo（续本达）、凸优化（武益阳）
  - 优点：自由参数个数可控
  - 缺点：有正交多项式的 Gibbs 波动残留，效果不够理想
  - 参考：`https://gitlab.airelinux.org/tjjk/proposal/-/tree/master/probe`
- 直方图（窦威）
  - 优点：最有效的模型，还原度高
  - 缺点：自由参数过多，刻度困难
  - 参考：`https://gitlab.airelinux.org/douwei/thesis`

流派（二）

- 手调反射（窦威）
  - 优点：符合物理世界，模型精准
  - 缺点：手扶拖拉机，探测器变了得重调，不调会翻车
  - 参考：`https://gitlab.airelinux.org/douwei/thesis`
- 光（武益阳）
  - 优点：对物理更深入一层，以光为单元建模，与发光曲线解耦
  - 缺点：做不出来，呈现出半个手扶拖拉机
  - 参考：`https://gitlab.airelinux.org/wuyy/bachelorthesis`
    - `https://gitlab.airelinux.org/tjjk/proposal/-/tree/master/light_probe`
- Normalized flow（武益阳）
  - 优点：上深度神经网络，理论上能拟合一切函数
  - 缺点：实际上啥也没拟合出来
  - 参考：没找到，在哪里？

流派（三）

- 光线（续本达）
  - 优点：自然产生平方反比
  - 缺点：没人做，不知道行不行
  - 参考：http://dpcraid.g.airelinux.org/meetings/TJJK/20220521-%E5%85%89probe.mp4
- 模拟退火（续本达）
  - 优点：不必显式写出 likelihood 函数，用模拟代替
  - 缺点：没人做，不知道行不行
  - 参考：https://dpcg.g.airelinux.org/user/xubd/lab/workspaces/auto-i/tree/annealing.ipynb

可用 sklearn

- Debian

```
apt install python3-sklearn-pandas
```

- 下面的算法可以理解为无解析形式的函数逼近。

支持向量机

```
### SVM regression
from sklearn.svm import SVR
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
regr = make_pipeline(StandardScaler(), SVR(C=1.0, epsilon=0.2))
regr.fit(X, y)
```

决策树

```
### GBoost
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, random_state=0)
reg = GradientBoostingRegressor(random_state=0)
reg.fit(X_train, y_train)
reg.predict(X_test[1:2])
reg.score(X_test, y_test)
```

神经网络

```
### 神经网络
from sklearn.neural_network import MLPRegressor
regr = MLPRegressor(random_state=1, max_iter=500).fit(X_train, y_train)
regr.predict(X_test[:2])
regr.score(X_test, y_test)
```

可微编程

把程序的函数看成数学函数，让函数输出的返回值相对于输入参数可微。

$$y = f(x) \to \frac{\mathrm{d}f}{\mathrm{d}x}$$

自动进行解析微分。

- PyTorch

```
apt install python3-torch
```

  - CPU 版本，GPU 版本在 WSL 之内不容易安装，建议申请科协服务器账号。

概率编程

$$y = f(x) \to Y = F(X)$$

把普通函数关系，自动换成随机变量之间的函数关系。

自动推理机

- 利用第一阶段大作业成果，直接生成第二阶段大作业的解答程序
- 语言为 Anglican ，为 Clojure 语言的一种
- Clojure 语言是 Scheme 语言在 Java 平台上的实现

duckdb

- 效率非常高的 SQL 数据分析引擎，可与 dbplyr 联动。

黑客技能

科学数据处理的原则

黑客的审美：复现 透明 一次 最佳工具

版本控制

Git 与队友分工协作，与明天的自己协作
Git 是“搬砖工地安全帽”，无头盔禁止上岗

关系代数

数据表示成关系，数据的操作表示成关系代数运算

数据格式

透明 CSV, HDF5, JSON, 数据库 SQL

黑客技能（二）

数据流水线

GNU Make 管理数据的依赖与转换，实现错误恢复和并行计算
实现数据层次的 Python/R/Bash/Scheme/SQL 多语言融合

正则表达式

描述字符串的微型语言，数学模型

命令环境

POSIX 环境中强大的小工具组合，开发与使用相融合

计算语言

Python 语法友好，工具丰富，统领 C/C++/Fortran/R/SQL 库

永远留在古老的计算环境

- SuperK 质子衰变和中微子实验，XMASS 暗物质实验
- 问题
  1. 数据处理技术发展停滞，违反“最佳工具”原则
  2. 新成员需要花精力学习旧技术，人力浪费
- 解决方案：test-driven development

永远留在古老的计算环境

- SuperK 质子衰变和中微子实验，XMASS 暗物质实验
- 问题
  1. 数据处理技术发展停滞，违反“最佳工具”原则
  2. 新成员需要花精力学习旧技术，人力浪费
- 解决方案：test-driven development

自制 Python 驱动的批量处理流水线

- 症状问题
  1. 使用 Python 调用大量 shell 命令，程序可读性差，违反“最佳工具”原则

```
for name in args.name:
  if name not in productions:
    print('Unknown production: ' + name)
    continue
  for script in productions[name]:
    parg = arguments + ' --name %s ' % name
    print('python %s %s' % (script, parg))
    os.system('python %s %s' % (script, parg))
```

  2. 流水线验证 flag 文件是否存在来确定是否成功执行，误判多，难以 debug
- 解决方案：使用 GNU Make 构建流水线
  - make 默认使用 /bin/sh 执行命令。SHELL=/bin/sh
  - 把 SHELL 换成提交任务给超级计算机集群的脚本

```
SHELL=lsf
export MAKE_TARGET=$@
export MAKE_SOURCE=$^
```

自制 Python 驱动的批量处理流水线

- 症状问题
  1. 使用 Python 调用大量 shell 命令，程序可读性差，违反“最佳工具”原则

```
for name in args.name:
  if name not in productions:
    print('Unknown production: ' + name)
    continue
  for script in productions[name]:
    parg = arguments + ' --name %s ' % name
    print('python %s %s' % (script, parg))
    os.system('python %s %s' % (script, parg))
```

  2. 流水线验证 flag 文件是否存在来确定是否成功执行，误判多，难以 debug
- 解决方案：使用 GNU Make 构建流水线
  - make 默认使用 /bin/sh 执行命令。SHELL=/bin/sh
  - 把 SHELL 换成提交任务给超级计算机集群的脚本

```
SHELL=lsf
export MAKE_TARGET=$@
export MAKE_SOURCE=$^
```

make 对接集群的脚本

```
#!/home/jinping/gentoo/bin/bash -e
# Platform LSF wrapper to be used as GNU Make shell.

# GNU Make convention for the first argument.
[[ ${1} = '-c' ]] && shift

for j in ${MAKE_SOURCE}; do
    [[ -z $(bjobs -J ${j}) ]] && continue
    DEP+=" && done(${j})" # 把 make 中的依赖关系传递给调度系统
done

cat << EOF > ${MAKE_TARGET}.sh
#!/home/jinping/gentoo/bin/bash
$@
EOF

chmod +x ${MAKE_TARGET}.sh

bsub -q normal"${DEP}" -J ${MAKE_TARGET} -o ${MAKE_TARGET}.log ${MAKE_TARGET}.sh
```

以本课程为起点

竞赛

- probe 大作业 → Ghost Hunter 2024 中微子数据分析排位赛
  - 概率统计分析及量测技术课赛结合

技术问题：TUNA 协会

- 清华大学学生开源软件与网络技术协会
- TUNA 主页 `https://tuna.moe/`
- TUNA 技术群，黑客（广义）技术问题探讨
- 为课程提供了
  - Debian 的镜像支持 `https://mirrors.tuna.tsinghua.edu.cn/debian/`
- 为大家提供了
  - 清华大学学位论文 LaTeX 模版 `thuthesis`

暑期实习

如果你觉得配环境是一个非常快乐的事，并且经常帮助同学配环境。
计算环境的可复现性是而容易被忽视，但也最重要的环节。

- Google Summer of Code：Google
  - Debian、Gentoo 操作系统相关项目
- 开源之夏活动：中科院软件所、华为、TUNA
  - Debian、Gentoo 操作系统相关项目

暑期实习

如果你觉得配环境是一个非常快乐的事，并且经常帮助同学配环境。计算环境的可复现性是而容易被忽视，但也最重要的环节。

- Google Summer of Code：Google
  - Debian、Gentoo 操作系统相关项目
- 开源之夏活动：中科院软件所、华为、TUNA
  - Debian、Gentoo 操作系统相关项目

SRT 与大创：Scheme

如果你认同实验物理与形式逻辑是文明的两大支柱，并喜欢课程的内容，可以考虑与我继续探索：

- 分析力学的 Scheme 描述：
  - Structure and Interpretation of Classical Mechanics, MIT 本科课程
  - 分析力学完成后，计划拓展至量子力学和电动力学

SRT 与大创：Scheme

如果你认同实验物理与形式逻辑是文明的两大支柱，并喜欢课程的内容，可以考虑与我继续探索：

- 分析力学的 Scheme 描述：
  - Structure and Interpretation of Classical Mechanics, MIT 本科课程
  - 分析力学完成后，计划拓展至量子力学和电动力学

致谢

顾问 陈晟祺、陈嘉杰
助教 武益阳、王宇逸、陶嘉燊、徐闯、郝传晖、刘逸祺
小助教 何宇峰、刘骥安、吕先艨、张珺铫、张贺、杨哲涵、王驰、程幸、胡溟崴、胡磊、陈威、黄一浩
科协 物理系科协、工物系科协；

积极对课程贡献的同学们 谢谢！

致谢

顾问 陈晟祺、陈嘉杰
助教 武益阳、王宇逸、陶嘉燊、徐闯、郝传晖、刘逸祺
小助教 何宇峰、刘骥安、吕先艨、张珺铫、张贺、杨哲涵、王驰、程幸、胡溟崴、胡磊、陈威、黄一浩
科协 物理系科协、工物系科协；

积极对课程贡献的同学们 谢谢！

致谢

顾问 陈晟祺、陈嘉杰

助教 武益阳、王宇逸、陶嘉燊、徐闯、郝传晖、刘逸祺

小助教 何宇峰、刘骥安、吕先艨、张珺铫、张贺、杨哲涵、王驰、程幸、胡淏崴、胡磊、陈威、黄一浩

科协 物理系科协、工物系科协；

- 提供算力支持

积极对课程贡献的同学们 谢谢！

致谢

顾问 陈晟祺、陈嘉杰

助教 武益阳、王宇逸、陶嘉燊、徐闯、郝传晖、刘逸祺

小助教 何宇峰、刘骥安、吕先艨、张珺铫、张贺、杨哲涵、王驰、程幸、胡淏崴、胡磊、陈威、黄一浩

科协 物理系科协、工物系科协；

- 提供算力支持
- 环境配置培训

积极对课程贡献的同学们 谢谢！

致谢

顾问 陈晟祺、陈嘉杰

助教 武益阳、王宇逸、陶嘉燊、徐闯、郝传晖、刘逸祺

小助教 何宇峰、刘骥安、吕先艨、张珺铫、张贺、杨哲涵、王驰、程幸、胡淏崴、胡磊、陈威、黄一浩

科协 物理系科协、工物系科协；

- 提供算力支持
- 环境配置培训

积极对课程贡献的同学们 谢谢！

致谢

顾问 陈晟祺、陈嘉杰

助教 武益阳、王宇逸、陶嘉燊、徐闯、郝传晖、刘逸祺

小助教 何宇峰、刘骥安、吕先艨、张珺铫、张贺、杨哲涵、王驰、程幸、胡淏崴、胡磊、陈威、黄一浩

科协 物理系科协、工物系科协；

- 提供算力支持
- 环境配置培训

积极对课程贡献的同学们 谢谢！

致谢

顾问 陈晟祺、陈嘉杰

助教 武益阳、王宇逸、陶嘉燊、徐闯、郝传晖、刘逸祺

小助教 何宇峰、刘骥安、吕先艨、张珺铫、张贺、杨哲涵、王驰、程幸、胡淏崴、胡磊、陈威、黄一浩

科协 物理系科协、工物系科协；

- 提供算力支持
- 环境配置培训

积极对课程贡献的同学们 谢谢！