数学模型 Lecture Notes

Zhennan Zhou

2018 年 5 月 8 日

PRELIMINARY DRAFT. NOT FOR WIDE CIRCULATION.

6 初等概率模型

It is scientific only to say what's more likely or less likely, and not to be proving all the time what's possible or impossible.

—— Richard Feynman

6.1 果壳中的概率（Probability in a nutshell）离散部分

概率空间(Ω, F, P)是一个总测度为1的测度空间（即$P(\Omega) = 1$）。

- Ω是一个非空集合，称为样本空间（Sample Space），他的元素称为样本输出（Outcome）。
- F是样本空间Ω的幂集的一个非空子集 (Ω的子集的集合)，它的元素称为事件（Event），事件是样本空间的子集。
- P称为概率(测度)。$P: F \to \mathbb{R}$。每个事件都被P赋予一个0和1之间的概率值。

随机变量$X: \Omega \to E$是从样本空间到可测空间E的可测函数。这门课里面，我们只考虑$E = \mathbb{R}$。随机变量取值$S \subset E$的概率我们记为

$$\Pr\big(X \in S\big) = P\big(\{\omega \in \Omega | X(\omega) \in S\}\big).$$

离散的随机变量可以被离散的概率密度刻画:

$$\mathbf{f} = \{f_i\}; \quad f_i \geq 0, \quad , i \in \mathbb{N}; \quad \sum_{i \in \mathbb{N}} f_i = 1.$$

如果X服从离散概率密度$\mathbf{f}$，我们记为$X \sim \mathbf{f}$。这个离散变量的（累积）分布函数$\mathbf{F} = \{F_i\}$定义为$F_n = \sum_{i \leq n} f_i$。我们易知，$0 \leq F_i \leq 1$, 而且 $F_n = \Pr(X \leq n)$。

关于记号做一点说明：虽然事件是样本空间的子集，但是，我们也习惯的用随机变量相对应的表示，比如事件$\{\omega \in \Omega | u < X(w) \leq v\}$，这个事件也简写为$u < X \leq v$。

条件概率$P(A|B)$是指事件A在另外一个事件B已经发生条件下的发生概率，定义为

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

两个事件A和B是（统计）独立的，当且仅当 $P(A \cap B) = P(A)P(B)$。易知，如果A和B是独立事件，$P(A|B) = P(A), P(B|A) = P(B)$。

一般的，根据 $P(A\cap B)=P(A)P(B|A)=P(B)P(A|B)$，我们得到贝叶斯公式

$$P(A|B)=\frac{P(A)P(B|A)}{P(B)}.$$

（离散）随机变量的期望（或称均值），是随机变量在概率分布下的平均值。我们用 $\mathbf{E}$ 表示期望，

$$\mathbf{E}X=\sum_i X_i f_i.$$

有时候，我们也用 $\bar{X}$ 表示期望。我们也可以对随机变量的函数求期望。比如，方差定义为

$$\mathrm{Var}(X)=D(X)=\mathbf{E}(X-\mathbb{E}X)^2=\mathbf{E}(X^2)-(\mathbb{E}X)^2.$$

最后，随机过程是指如下的一族的随机变量

$$\{X(t):t\in T\}.$$

这里，T 是一个指标集，可以是连续的，也可以离散的。历史上来看，$t\in T$ 常被理解为时间，而 $X(t)$ 是某个可观测量在时间 t 时对应的随机变量。有时候，人们也会把一个随机过程写成 $\{X(t,\omega):t\in T\}$, 表明它其实是 $t\in T$ 和 $\omega\in\Omega$ 的二元函数。

6.2 随机人口模型

时刻 t 的人口用随机变量 $X(t)$ 表示，$X(t)$ 只取整数值。记 $P_n(t)$ 是 $X(t)=n$ 的概率，$n=0,1,2,\cdots$。下面我们对出生和死亡的概率做出适当的假设，寻求 $P_n(t)$ 的变化规律，并由此得到 $X(t)$的期望和方差。

若 $X(t)=n$，对于充分小的时间Δt，我们对人口在 t 到 $t+\Delta t$ 的出生和死亡做如下的假设：

- 出生一人的概率与 Δt 成正比，记为 $b_n\Delta t$，出生两人及以上的概率是$o(\Delta t)$。且 b_n 与 n 成正比，记为 $b_n=\lambda n$。
- 死亡一人的概率与 Δt 成正比，记为 $d_n\Delta t$，出生两人及以上的概率是$o(\Delta t)$。且 d_n 与 n 成正比，记为 $b_n=\mu n$。
- 出生死亡是相互独立的随机事件。

于是我们得到，

$$P_n(t+\Delta t)=P_{n-1}(t)b_{n-1}\Delta t+P_{n+1}(t)d_{n+1}\Delta t+P_n(t)(1-b_n\Delta t-d_n\Delta t)+o(\Delta t).$$

于是，我们得到如下的微分方程

$$\frac{dP_n}{dt}=\lambda(n-1)P_{n-1}+\mu(n+1)P_{n+1}-(\lambda+\mu)nP_n. \tag{6.1}$$

如果，在初始时刻（t=0）人口为确定的数量n_0，则 $P_n(t)$ 的初始条件为

$$P_{n_0}(0)=1,\quad P_n(0)=0,\quad n\neq n_0. \tag{6.2}$$

求解这些方程非常复杂，但是如果我们只关心 $X(t)$ 的期望（以下简记 $E(t)$ ）和方差(以下简记 $D(t)$)，则我们可以由(6.1)和(6.2)直接得到。根据期望的定义，$E(t)=\sum_n nP_n(t)$。我们可以得到 $E(t)$ 满足的方程

$$\frac{dE}{dt}=\lambda\sum_{n=1}^{\infty}n(n-1)P_{n-1}+\mu\sum_{n=1}^{\infty}n(n+1)P_{n+1}-(\lambda+\mu)\sum_{n=1}^{\infty}n^2P_n.$$

经过化简，我们得到

$$\frac{dE}{dt} = (\lambda - \mu)E.$$

而它的初始条件为$E(0) = n_0$。所以，我们得到方程的解

$$E(t) = n_0 e^{(\lambda-\mu)t}.$$

注意，这个形式就和非随机的模型完全一致了。

对于方差$D(t)$, 按照定义$D(t) = \sum_{n=1}^{\infty} n^2 P_n(t) - E^2(t)$. 可以推出（作业题）

$$D(t) = n_0 \frac{\lambda+\mu}{\lambda-\mu} e^{(\lambda-\mu)t}[e^{(\lambda-\mu)t} - 1].$$

6.3 敏感问题的调查和估计

统计调查中，可能会遇到一些因涉及个人隐私或利害关系的所谓敏感性问题。这时候，即使做无记名的直接调查，也很难消除被调查者的顾虑，从而难以保证数据的真实性。本节中，我们以考试作弊为例，介绍这类敏感问题的调查方法。

调查方案设计的基本思想是：让被调查者从包含是否作弊的若干问题中，随机的选答其中一个，同时让调查者也不知道被调查者回答的是哪一个问题，从而保护被调查者的隐私，消除他们的顾虑，能够真实作答。下面以400名学生对作弊问题的问卷调查为依据，通过若干模型估计考试中有过作弊的比例，并对这些模型进行分析比较。

Warner 模型 Warner在1965年提出了正反问题选答法，要调查的问题在问卷上以正、反两种形式叙述。比如，对于考试作弊，涉及下面两个陈述：

A. I have cheated during an exam.

B. I have never cheated during an exam.

对于选择的问题，学生只需回答“真”或者“假”。调查者准备一套数字为1到13各一张的一套扑克牌。在回答问题时，学生先随机抽一张扑克牌，看后归还，调查者不知道学生抽取的结果。如果抽到1到10的扑克牌，学生对A作答；如果抽到11到13，学生对B作答。我们假定学生都将真实作答。

假定调查结果是，收回了 $n = 400$ 张有效答卷，其中有$n_1 = 112$个学生回答“真”，$n_2 = 288$个学生回答“假”。所以对问题A、B两题选答“真”的概率 π 的估计值为 $\hat{\pi} = n_1/n = 7/25$。

我们要估计的有作弊行为的同学的比例，可以看作一个学生作弊的概率，即对A回答为“真”和对B回答为“假”的概率，记为 π_A。另外，记回答A的概率为 $p = 10/13$。我们引入随机变量 X_i, $i = 1, \cdots, n$: $X_i = 1$, 如果第i个学生回答“真”； $X_i = 0$, 如果第i个学生回答“假”。

易知，$X_1, \cdots, X_n$ 独立同分布，$\mathbb{E}X = \pi$, $D(X_i) = \pi(1-\pi)$。对于n个学生，对回答为“真”的概率 π 的估计值为

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

注意，$\hat{\pi}$ 为无偏估计，即 $\mathbb{E}\hat{\pi} = \pi$. 而且，我们可知

$$\pi = p\pi_A + (1-p)(1-\pi_A).$$

如果 $p \neq \frac{1}{2}$, 可得π_A 的估计值

$$\hat{\pi}_A = \frac{\hat{\pi} - (1-p)}{2p-1}.$$

我们可以验证，$\hat{\pi}_A$ 也是无偏估计，即 $\mathbb{E}\hat{\pi}_A = \pi_A$. 而且，通过直接的计算，我们可以得到

$$D(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} + \frac{p(1-p)}{(2p-1)^2 n}.$$

可以看出，$\hat{\pi}_A$ 的方差由两部分构成：

- 直接调查并得到真实数据时，$\hat{\pi}_A$ 的方差（$p=0$ 或者 $p=1$）。
- 随机回答机制引入的方差。

为了使得估计达到预先确定的方差δ, 我们可以考虑

$$D(\hat{\pi}_A) \le \frac{1}{4n} + \frac{p(1-p)}{(2p-1)^2 n} \le \delta.$$

在考虑合适的 p 后只需要取 n 为大于 $\frac{1}{4\delta} + \frac{p(1-p)}{\delta(2p-1)^2}$ 的整数即可。

最后，对于这个例子中，我们得到作弊学生比例的估计值与其标准差是 $\hat{\pi}_A = 0.091$，$\sqrt{D(\hat{\pi}_A)} = 0.042$。如果用2倍标准差作为估计值的精度，那么可以说有作弊行为的学生比例为$9.1\% \pm 8.4\%$。

注意，Warner模型对p的选择有限制：当$p \approx \frac{1}{2}$，调查失去了精度，当p接近0或者1，对被调查者的保护程度会降低。因此，一般p取值在0.7和0.8之间。

Simmons 模型 下面介绍Simmons的无关问题选答技术（1967），调查人员提出的两个问题，其中一个为敏感性问题，另一个为非敏感性的问题。假设被调查学生的人数仍为 $n = 400$，问题选答规则同Warner模型，只不过两个陈述变为

A. I have cheated during an exam.

B'. My birth month is an even number.

假定调查结果是，收回了 $n = 400$ 张有效答卷，其中有$n_1 = 80$个学生回答“真”，$n_2 = 320$个学生回答“假”。所以对问题A、B两题选答“真”的概率 π 的估计值为 $\hat{\pi} = n_1/n = 1/5$。

由于调查人数较大，可以认为对问题B’中回答“真”的概率为$\pi'_B = \frac{1}{2}$。记对问题A回答为“真”的概率是π'_A，π、p和$X_i\,(i = 1, \cdots, n)$ 的定义同Warner模型相同。

根据模型，我们知道

$$\pi = p\pi'_A + (1-p)\pi'_B.$$

于是，我们得到对π'_A的估计

$$\hat{\pi}'_A = \frac{\hat{\pi} - (1-p)\pi'_B}{p}.$$

我们可以验证，$\hat{\pi}'_A$ 也是无偏估计，即 $\mathbb{E}\hat{\pi}'_A = \pi'_A$. 而且，通过直接的计算，我们可以得到

$$D(\hat{\pi}'_A) = \frac{\pi'_A(1-\pi'_A)}{n} + \frac{1-p^2}{4np^2}.$$

通过比较 $D(\hat{\pi}_A)$ 和 $D(\hat{\pi}'_A)$，我们得知，当且仅当 $p > \frac{1}{3}$，我们就有 $D(\hat{\pi}'_A) < D(\hat{\pi}_A)$。进一步，可以证明，对任意给定的 π'_B，只要$p > \frac{1}{3}$，就可以保证 $D(\hat{\pi}'_A) < D(\hat{\pi}_A)$。所以从估计值的方差的角度，可以说Simmons模型的精度比Warner模型高。

对于这个例子中，我们得到作弊学生比例的估计值与其标准差是 $\hat{\pi}'_A = 0.11$，$\sqrt{D(\hat{\pi}'_A)} = 0.026$。如果用2倍标准差作为估计值的精度，那么可以说有作弊行为的学生比例为$11\% \pm 5.2\%$。

Christofides 模型 Christofides 在2003年设计了如下的模型：

调查者准备一套外观相同的卡牌，每张卡片写着$1,\cdots,L$中的某一个数字，数字为k的卡牌所占比例为p_k，并且p_k不全相同。调查时，学生先抽一张卡片，然后按照下面的规则作答：若学生曾经作弊，就回答$L+1$与他抽的数字的差，否则，回答他抽的数字。

为了方便分析，我们引入随机变量 $Z_i, i=1,\cdots,n$，使得：$Z_i=L+1$，若第i个学生曾经作弊；$Z_i=0$，若第i个学生未曾作弊。仍然记π_A为学生作弊的概率。记随机变量 Y_i 为第i个学生抽到的数字（$i=1,\cdots,n$）。易知，Y_i（$i=1,\cdots,n$）是独立同分布的，而且与 Z_i 是相互独立的。

根据调查的机制，我们知道第i个学生回答的数字为

$$d_i = |Y_i - Z_i|, \quad i=1,\cdots,n.$$

易知，

$$\Pr(d_i=k) = (1-\pi_A)p_k + \pi_A p_{L+1-k}, \quad k=1,\cdots,L.$$

因此，我们可以计算

$$\mathbb{E}(d_i) = \sum_{k=1}^{L} k\Pr(d_i=k) = \sum_{k=1}^{L} kp_k + \pi_A\left(L+1-2\sum_{k=1}^{L} kp_k\right).$$

注意到，$\mathbb{E}Y_i = \sum_{k=1}^{L} kp_k$，我们简记为 $E(Y)$。于是我们有

$$\mathbb{E}(d_i) = E(Y) + \pi_A(L+1-2\mathbb{E}(Y)).$$

如果简记 Y_i 的方差为 $D(Y)$,我们可以计算得到

$$D(d_i) = D(Y) + \pi_A(1-\pi_A)(L+1-2E(Y))^2.$$

计算n个学生回答的数字的均值$d_1,d_2,\cdots,d_n$ 的均值 $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$。于是我们可得 π_A 的估计值

$$\hat{\pi}_A^* = \frac{\bar{d}-E(Y)}{L+1-2E(Y)}.$$

易知，$\hat{\pi}_A^*$ 为无偏估计，而且具有方差

$$D(\hat{\pi}_A^*) = \frac{\pi_A^*(1-\pi_A^*)}{n} + \frac{D(Y)}{n(L+1-2E(Y))^2}.$$

可以验证，Warner模型其实是Christofides模型在$L=2$时的特例。