几种离散型随机变量与连续型随机变量的关系探讨

由一道作业题引发的思考

摘要：本文受一道课后习题启发，分析探讨了几类离散型随机变量和连续型随机变量的联系。
关键词：泊松分布；指数分布；Erlang 分布；伽玛分布；均匀分布

1 引 言

在做概率论与数理统计的第三章作业时，有一道这样的题目：[1]（下称题 1）

设一大型设备在任何长为 t 的时间内发生故障的次数 $N(t)$ 服从参数为 λt 的泊松分布。求：

（1）相继两次故障之间时间间隔 T 的分布；

（2）在设备已经无故障工作了 8h 的情形下，再无故障运行 8h 的概率。

起初我看到这个问题感到无从下手——本题的风格与其他的习题都迥然不同：其他习题基本是给出连续型随机变量 X 和 Y，再告知 X 服从的概率分布及 Y 与 X 的关系，要求 Y 的概率密度，此时我们可以用分布函数法求解；或是告知了两离散型随机变量间的关系，要求由其中一者的分布列求另一者的分布列，此类题分析起来也不困难。可本题中，发生故障的次数是离散的，而相继两次故障之间的时间间隔显然有不可列无穷多种情况，这就要求我们建立一个离散型随机变量到非离散型随机变量的联系。这让我感到新奇，并启发我寻找更多此类联系。此外，本题问的是“相继两次故障”，若将其推广到多次，结果又会如何？若考察某个发生过 n 次故障的时间段，在这时间段内发生故障的时刻的分布情况又如何？会不会建立起新的由一种随机变量概率分布到另一种概率分布的联系呢？以上就是本文探讨的主要问题。

2 题 1 的解答

对于上述题 1 的第（1）问，我们设定一个初始时刻（比如机器刚开机时，此时未发生过故障）。记 X_1 为机器第 1 次发生故障的时刻与初始时刻的时间间隔，X_n 为机器第 $n-1$ 次发生故障与第 n 次发生故障的时间间隔 $(n \geq 2)$。现在分析 X_i 服从什么分布。首先分析 X_1，易知 $t \leq 0$ 时，$F_{X_1}(t) = 0$；当 $t > 0$ 时，$X_1 > t$ 表示 $[0, t]$ 时间内，机器一次故障也没发生，也就是事件 $N(t) = 0$ 发生，根据泊松分布的公式得

$$P(X_1 > t) = P(N(t) = 0) = \frac{(\lambda t)^0}{0!}\mathrm{e}^{-\lambda t} = \mathrm{e}^{-\lambda t}.$$

因此我们有

$$P(X_1 \leq t) = 1 - P(X_1 > t) = 1 - \mathrm{e}^{-\lambda t}.$$

因此，X_1 的分布函数为

$$F_{X_1}(t)=\begin{cases}0 & t\le 0,\\ 1-\mathrm{e}^{-\lambda t} & t>0.\end{cases}$$

即 X_1 服从参数为 λ 的指数分布。由此也容易知道 X_1 的概率密度：

$$f_{X_1}(t)=\begin{cases}0 & t\le 0,\\ \lambda\mathrm{e}^{-\lambda t} & t>0.\end{cases}$$

接着我们在已知 X_1 的分布的情况下分析 X_2。设 $S_n=\sum_{i=1}^{n}X_i$，我们先求出 (S_1,S_2) 的联合密度函数，再利用变换法求出 (X_1,X_2) 的概率密度函数。

选 $s_1<s_1+\Delta s_1<s_2<s_2+\Delta s_2$，考虑事件 $s_1\le S_1<s_1+\Delta s_1, s_2\le S_2<s_2+\Delta s_2$，此事件发生等价于以下四个事件同时发生：

①$0<t<s_1$ 时，机器未发生故障；②$s_1<t<s_1+\Delta s_2$ 时，机器恰发生故障一次；

③$s_1+\Delta s_1<t<s_2$ 时，机器未发生故障；④$s_2<t<s_2+\Delta s_2$ 时，机器恰发生故障一次。

又由于这四个时段里发生故障的次数是相互独立的（只与该时段的时间长短和固定参数有关），我们有：

$$\begin{aligned}&P(s_1\le S_1<s_1+\Delta s_1, s_2\le S_2<s_2+\Delta s_2)\\ &=\mathrm{e}^{-\lambda s_1}\times\frac{\lambda\Delta s_1}{1!}\mathrm{e}^{-\lambda\Delta s_1}\times\mathrm{e}^{-\lambda(s_2-s_1-\Delta s_1)}\times\frac{\lambda\Delta s_2}{1!}\mathrm{e}^{-\lambda\Delta s_2}\\ &=\lambda^2e^{-\lambda s_2}\Delta s_1\Delta s_2\times\mathrm{e}^{-\lambda\Delta s_2}\end{aligned}\tag{1}$$

当 $\Delta s_2\to 0$ 时，$\mathrm{e}^{-\lambda\Delta s_2}-1\sim-\lambda\Delta s_2$，所以(1)改写为

$$\begin{aligned}&P(s_1\le S_1<s_1+\Delta s_1, s_2\le S_2<s_2+\Delta s_2)\\ &=\lambda^2e^{-\lambda s_2}\Delta s_1\Delta s_2\times(\mathrm{e}^{-\lambda\Delta s_2}-1+1)\\ &=\lambda^2e^{-\lambda s_2}\Delta s_1\Delta s_2+\lambda^2e^{-\lambda s_2}\Delta s_1\Delta s_2\times(-\lambda\Delta s_2)\\ &=\lambda^2e^{-\lambda s_2}\Delta s_1\Delta s_2-\lambda^3e^{-\lambda s_2}\Delta s_1\Delta s_2^2\\ &=\lambda^2e^{-\lambda s_2}\Delta s_1\Delta s_2+o(\Delta s_1\Delta s_2)\end{aligned}$$

当 $\Delta s_1\to 0,\Delta s_2\to 0$ 时，我们得到 (S_1,S_2) 的联合密度函数 $(s_1,s_2>0)$ 为

$$\begin{aligned}&f_{(S_1,S_2)}(s_1,s_2)\\ &=\lim_{\Delta s_1\to 0,\Delta s_2\to 0}\frac{P(s_1\le S_1<s_1+\Delta s_1, s_2\le S_2<s_2+\Delta s_2)}{\Delta s_1\Delta s_2}\\ &=\lambda^2e^{-\lambda s_2}\end{aligned}$$

由变换法 [1][3] 可得，X_1,X_2 的概率密度函数 $(t_1,t_2>0)$ 为

$$f_{(X_1,X_2)}(t_1,t_2)=\lambda^2e^{-\lambda(t_1+t_2)}=\lambda\mathrm{e}^{-\lambda t_1}\times\lambda\mathrm{e}^{-\lambda t_2}$$

类似可得，对于任意 $n \geq 2, s_1, ..., s_n > 0$，均有

$$f_{(S_1,...,S_n)}(s_1, ..., s_n) = \lambda^n e^{-\lambda s_n}$$

作变换后均有

$$f_{(X_1,...,X_n)}(t_1, ..., t_n) = \lambda \mathrm{e}^{-\lambda t_1} \times ... \times \lambda \mathrm{e}^{-\lambda t_n}.$$

做积分可求得 $f_{X_n}(t_n) = \lambda \mathrm{e}^{-\lambda t_n}, t_n > 0, n > 1$，进而可知边缘概率密度的乘积等于联合概率密度，因此可以得到，所有的 X_i 都是独立同分布的。**因此，任意两次故障的时间间隔 T 都服从参数为 λ 的指数分布。**

（2）问在第（1）问基础上即很容易求解。设 $A =$ “在设备已经无故障工作了 8h 的情形下，再无故障运行 8h”，则

$$P(A) = \frac{P(T \geq 16)}{P(T \geq 8)} = \frac{\mathrm{e}^{-16\lambda}}{\mathrm{e}^{-8\lambda}} = \mathrm{e}^{-8\lambda}.$$

3 题 1 的延伸

3.1 第 n 次故障到来时已等待的时间

刚才我们的分析中用到了 S_n。S_n 也有实际意义：比如说设备在某时刻启动了，则 S_n 就表示从此时刻到设备发生 n 次故障经过的时间。那么这个量的分布有什么规律吗？这里，我们尝试通过求解 $P(S_n \leq t)$ 来得出 S_n 的分布函数。此时，$S_n \leq t$ 与 $N(t) \geq n$ 等价。所以

$$F_{S_n}(t) = P(S_n \leq t) = P(N(t) \geq n) = \sum_{i=n}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^i}{i!}$$

易知上述表达式对 t 是可导的。所以对其求导，我们得到 S_n 的概率密度函数：

$$\begin{aligned} f_{S_n}(t) &= \sum_{i=n}^{\infty} (-\lambda) \mathrm{e}^{-\lambda t} \frac{(\lambda t)^i}{i!} + \sum_{i=n}^{\infty} \mathrm{e}^{-\lambda t} (i\lambda) \frac{(\lambda t)^{i-1}}{i!} \\ &= -\sum_{i=n}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^i}{i!} + \sum_{i=n}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^{i-1}}{(i-1)!} \\ &= -\sum_{i=n}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^i}{i!} + \sum_{i=n-1}^{\infty} \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^i}{i!} = \lambda \mathrm{e}^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \end{aligned}$$

这就是 **Erlang（埃尔朗）分布**的概率密度函数，参数为 $n (n \in N)$ 和 λ。[3] 我们还可从推导过程中看出，Erlang 分布不仅对于从初始时刻开始到发生第 n 次故障的总时间成立，也对从任意一次故障后再发生 n 次故障所用的总时间成立。更一般地说，它是 n 个相互独立且服从指数分布的随机变量之和服从的概率分布。若参数 n 不局限于整数，而是可取任意正数

r，我们就由此得到**伽玛分布 (Γ 分布)** 的概率密度函数 [1]，参数为 r 和 λ。此时 Γ 分布的概率密度函数为

$$f(x)=\begin{cases}\frac{\lambda^r}{\Gamma(r)}x^{r-1}\mathrm{e}^{-\lambda x} & ,x>0;\\ 0 & ,x\leq 0.\end{cases}$$

其中 $\Gamma(r)=\int_0^{+\infty}t^{r-1}\mathrm{e}^{-t}\mathrm{d}t(r>0)$.

3.2 发生 n 次故障的时段内发生故障的时刻

我们仍回到题 1 的情境中。假设我们开机后不去管这设备，过了 t 时间我们去查看设备的运行情况，发现设备自动记录了 n 次故障记录。那么，在这 t 时间内，这 n 次故障记录出现的时刻的分布情况如何？

我们先考虑最简单的情况：$n=1$，此时有

$$\begin{aligned}F_X(s)&=P(X\leq s|N(t)=1)\\&=\frac{P(X\leq s,N(t)=1)}{P(N(t)=1)}\\&=\frac{P(N(s)=1,N(t-s)=0)}{P(N(t)=1)}\\&=\frac{P(N(s)=1)P(N(t-s)=0)}{P(N(t)=1)}\\&=\frac{\lambda se^{-\lambda s}\mathrm{e}^{-\lambda(t-s)}}{\lambda te^{-\lambda t}}=\frac{s}{t}.\end{aligned}$$

也就是说，若在某一时间段内只发生一次故障，那么该次故障记录在该时间段内发生的具体时刻服从**均匀分布**。

对于一般的情况，我们要分析的就是 $S_1,...,S_n$ 的分布情况了。令 $0<t_1<...<t_{n+1}=t$，再取 $h_1,...,h_n$ 充分小使得 $t_i+h_i<t_{i+1}$，考虑事件 $t_1\leq S_1<t_1+h_1,...,t_n\leq S_n<t_n+h_n$，此事件发生等价于以下 $2n$ 个事件同时发生：

1.0 $\sim t_1$ 时间内，机器未发生故障；

2.$t_1\sim t_1+h_1$ 时间内，机器恰发生故障一次；

3.$t_1+h_1\sim t_2$ 时间内，机器未发生故障；

4.$t_2\sim t_2+h_2$ 时间内，机器恰发生故障一次；

...

2n−1.$t_n\sim t_n+h_n$ 时间内，机器恰发生故障一次；

2n.$t_n+h_n\sim t_{n+1}$(即t) 时间内，机器不发生故障。

于是

$$
\begin{aligned}
&P(t_1 \le S_1 < t_1 + h_1, ..., t_n \le S_n < t_n + h_n | N(t) = n) \\
&= \frac{\mathrm{e}^{-\lambda t_1} \lambda h_1 e^{-\lambda h_1} \mathrm{e}^{-\lambda(t_2 - t_1 - h_1)} \lambda h_2 e^{-\lambda h_2} \cdots \mathrm{e}^{-\lambda(t_n - t_{n-1} - h_{n-1})} \lambda h_n e^{-\lambda h_n} \mathrm{e}^{-\lambda(t - t_n - h_n)}}{\mathrm{e}^{-\lambda t}(\lambda t)^n / n!} \\
&= \frac{\lambda^n h_1 \cdots h_n \mathrm{e}^{-\lambda t}}{\mathrm{e}^{-\lambda t}(\lambda t)^n / n!} = \frac{n!}{t^n} h_1 \cdots h_n.
\end{aligned}
$$

当 $h_i \to 0$ 时，

$$
\begin{aligned}
&f_{(S_1,...,S_n)}(t_1, ..., t_n) \\
&= \lim_{h_i \to 0, i=1,...,n} \frac{P(t_1 \le S_1 < t_1 + h_1, ..., t_n \le S_n < t_n + h_n | N(t) = n)}{h_1 \cdots h_n} \\
&= \frac{n!}{t^n}.
\end{aligned}
$$

我们得到 $S_1, ..., S_n$ 在已知 $N(t) = n$ 时的条件概率密度函数

$$
f(t_1, ..., t_n | N(t) = n) = \begin{cases} \frac{n!}{t^n} & , 0 < t_1 < ... < t_n < t; \\ 0 & , \text{其他}. \end{cases}
$$

可见，$n = 1$ 的情形也与之吻合。因此，发生故障的时刻（或从启动设备到发生故障用时）的分布与相应于 n 个 $[0, t]$ 上均匀分布的独立随机变量的顺序统计量有相同的分布。[4]

4 结 语

本文受一道课后习题启发，分析探讨了几类离散型随机变量和连续型随机变量的联系。我写作这篇文章的最大体会是：善于发现、勤于思考，才能不断进步。限于作者水平，本文还很稚嫩，推导不够深入，也可能有不严谨之处甚至错误，恳请老师斧正。

参考文献

[1] 王勇, 等. 概率论与数理统计 [M]. 第 2 版, 北京: 高等教育出版社,2014.

[2] 李贤平. 概率论基础学习指导书 [M]. 北京: 高等教育出版社,2011:197-198.

[3] 李贤平. 概率论基础 [M]. 第 3 版, 北京: 高等教育出版社,2010.

[4] 刘次华. 随机过程 [M]. 第 4 版, 武汉: 华中科技大学出版社,2008.