随机变量及其概率分布

本节着重复习单维随机变量的情况。

离散型随机变量的分布律

两点分布

二项分布

Poisson分布

【Poisson定理】设$\lambda>0$是常数，n是任意正整数，且$\lambda=np$，则对于任意取定的非负整数k，有

$$\lim_{n\to\infty}C_n^kp^k(1-p)^{n-k}=\frac{\lambda^k}{k!}e^{-\lambda}$$

由Poisson定理，若n很大p很小时，设$\lambda=np$，则有

$$P\{X=k\}=C_n^kp^k(1-p)^{n-k}\approx\frac{\lambda^k}{k!}e^{-\lambda}$$

近似条件：$n\geq 20$，$p\leq 0.05$。

【Poisson分布】设随机变量X所有可能取的值为0，1，2，$\cdots$，而取各个值的概率为

$$p_k=P\{X=k\}=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda},\quad k=0,1,2,\cdots$$

其中$\lambda>0$是常数。则称X服从参数为λ的泊松分布，记为$X\sim P(\lambda)$.

> 取一个很大的自然数n，把时间段$[0，1)$分为等长的n段：
>
> $$l_1=\left[0,\frac{1}{n}\right),l_2=\left[\frac{1}{n},\frac{2}{n}\right),\cdots,l_i=\left[\frac{i-1}{n},\frac{i}{n}\right),\cdots,l_n=\left[\frac{n-1}{n},1\right)$$
>
> 假定：
>
> 1°在每段l_i内，恰发生一个事故的概率，近似地与这段时间的长$\frac{1}{n}$成正比，即可取为λ/n.又假定在n很大因而$1/n$很小时，在l_i这么短暂的一段时间内，要发生两次或更多的事故是不可能的. 因此在l_i时段内不发生事故的概率为$1-\lambda/n$.
>
> 2°　l_1，$\cdots$，l_n各段是否发生事故是独立的.
>
> 把在$[0，1)$时段内发生的事故数X视作在n个小时段l_1，$\cdots$，l_n内有事故的 时段数，则按上述1°，2°两条假定，X应服从二项分布$B(n，\lambda/n)$. 于是
>
> $$P(X=i)=\binom{n}{i}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i}$$
>
> 再对此式取极限即可。

负二项分布

【例题】为了检查某厂产品的废品率p大小，有两个试验方案可采取:一是从该厂产品中抽出若干个，检查其中的废品数X，这一方案导致二项分布，已于前述；另一个方案是先指定一个自然数r，一个一个地从该厂产品中抽样检查，直到发现第r个废品为止. 以X记到当时为止已检出的合格品个数. 显然，若废品率p小，则X倾向于取较大的值；反之，当p大时，则X倾向于取小值. 故X可用于考究p的目的。为计算X的分布，假定各次抽取的结果是独立的，且每次抽得废品的概率保持固定为p.

考察$\{X=i\}$这个事件，这个事件发生需要以下两个事件同时发生：

(1) 在前$i+r-1$次抽取中，恰有$r-1$个废品;

(2) 第$i+r$次抽出废品.

按所作假定，这两个事件的概率分别为$b(r-1;i+r-1，p)$和p. 再由独立性，即得

$$\begin{aligned}P(X=i)&=b(r-1;i+r-1,p)p\\&=\binom{i+r-1}{r-1}p^r(1-p)^i\quad(i=0,1,2,\cdots).\end{aligned}$$

这个分布称为负二项分布。这个名称的来由：在“负指数二项展开式”

$$\begin{aligned}(1-x)^{-r}&=\sum_{i=0}^{\infty}\binom{-r}{i}(-x)^i=\sum_{i=0}^{\infty}\binom{i+r-1}{i}x^i\\&=\sum_{i=0}^{\infty}\binom{i+r-1}{r-1}x^i\end{aligned}$$

中令$x=1-p$，并两边乘以p^r，得

$$1=p^r[1-(1-p)]^{-r}=\sum_{i=0}^{\infty}\binom{i+r-1}{r-1}p^r(1-p)^i$$

几何分布

$r=1$时的负二项分布。

随机变量的分布函数

设X为一随机变量，则函数$P(X\leqslant x)=F(x)\quad(-\infty<x<\infty)$称为$X$的分布函数.

随机变量的概率密度函数

设连续型随机变量X有概率分布函数$F(x)$，则$F(x)$的导数$f(x)=F'(x)$称为X的概率密度函数。连续型随机变量X的密度函数$f(x)$都具有以下三条基本性质：

(1) $f(x)\geq 0$;

(2) $\int_{-\infty}^{\infty}f(x)\mathrm{d}x=1$

(3) 对任何常数$a<b$，有$P(a\leqslant X\leqslant b)=F(b)-F(a)=\int_a^b f(x)\mathrm{d}x$

连续型随机变量的分布律

均匀分布

设随机变量 X 有概率密度函数

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant x\leqslant b\\ 0, & \text{else}\end{cases}$$

则称X服从区间 $[a,b]$ 上的均匀分布，并常记为 $X\sim\mathcal{U}(a,b)$。均匀分布 $\mathcal{U}(a,b)$ 的分布函数是

$$F(x)=\begin{cases}0, & x\leqslant a\\ \dfrac{x-a}{b-a}, & a<x<b\\ 1, & x\geqslant b\end{cases}$$

QQ截图20211030090703

> 借助均匀分布容易实现对分布的模拟。首先，若以某种方法产生"随机数"，则如取 n 足够大，而独立地产生 n 个随机数字 $a_1,\cdots,a_n$ 时，$X=0.a_1a_2\cdots a_n$ 就很接近于$[0,1]$均匀分布。
>
> 对一般分布函数 $F(x)$，若 $F(x)$ 处处连续且严格上升，则其反函数 G 存在，这时易见，若$X\sim\mathcal{U}(0,1)$，则 $G(X)\sim F$。事实上，$\{G(X)\leqslant x\}$ 这个事件就是 $\{F(G(x))\leqslant F(x)\}$, 即 $\{X\leqslant F(x)\}$，因而（注意到$\mathcal{U}(0,1)$的分布函数为 $F(x)=x(0<x<1)$）
>
> $$P(G(X)\leqslant x)=P(X\leqslant F(x))=F(x)$$
>
> 这证明了 $G(X)\sim F$。这样，产生 X 的模拟值后代入 G 中即得分布 F 的模拟值。

正态分布

如果一个随机变量具有概率密度函数

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]\quad(-\infty<x<\infty)$$

则称X为正态随机变量，并记为$X\sim N\left(\mu，\sigma^2\right)$.

> 当$\mu=1$，$\sigma^2=1$时，$f(x)$成为
>
> $$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{x^2}{2}\right)$$
>
> 它是正态分布$N(0，1)$的密度函数.
>
> $N(0，1)$称为"标准正态分布"。在概率论著作中，其密度函数和分布函数常分别记为$\varphi(x)$和$\Phi(x)$，并造有很仔细的表。任意的正态分布$N\left(\mu，\sigma^2\right)$的计算很容易转化为标准正态分布$N(0，1)$：若$X\sim N\left(\mu，\sigma^2\right)$，则$Y=\frac{x-\mu}{\sigma}\sim N(0，1)$.

> 利用$\varphi(x)$的对称性得到
>
> $$\Phi(x)+\Phi(-x)=1,\quad x\in\mathcal{R}$$
>
> 并且，只要$X\sim N\left(\mu，\sigma^2\right)$，则有

> $$\begin{aligned} P(X \leqslant a) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{a}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(a-\mu)/\sigma}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \\ &= \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$
>
> 于是对任何$a<b$，当$X\sim N\left(\mu，\sigma^2\right)$，用
>
> $$P(a<X\leqslant b)=P(X\leqslant b)-P(X\leqslant a)$$
>
> 得到公式
>
> $$P(a<X\leqslant b)=\phi\left(\frac{b-\mu}{\sigma}\right)-\phi\left(\frac{a-\mu}{\sigma}\right)$$

指数分布

若随机变量X有概率密度函数

$$f(x)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & \text{当 } x>0 \text{ 时}\\ 0, & \text{当 } x\leqslant 0 \text{ 时}\end{cases}$$

则称X服从指数分布。其中$\lambda>0$为参数。由于当$x\leqslant 0$时$f(x)=0$，表示随机变量取负值的概率为0，故X只取正值.

> 【无后效性，无记忆性】设X是连续型非负随机变量，则X服从指数分布的充分必要条件是对任何$s，t\geqslant 0$，有
>
> $$P(X>s+t\mid X>s)=P(X>t)$$

> 【λ的实际意义】设想一种大批生产的电子元件，其寿命X是随机变量。以$F(x)$记X的分布函数。我们来证明：在一定的条件下，$F(x)$就是$\int_{-\infty}^{x}f_{\mathcal{E}}(t)\mathrm{d}t$。
>
> 我们要作的假定，从技术上说就是“无老化”。失效率就是单位长度时间内失效的概率。用条件概率的形式，上述假定可表为
>
> $$\frac{P(x\leqslant X\leqslant x+h\mid X>x)}{h}=\lambda\quad(h\to 0)$$
>
> 未取极限前，表示在x时刻的平均失效率。令$h\to 0$，得瞬时失效率，按假定，它应为常数λ。**按条件概率的定义**，注意到$P(X>x)=1-F(x)$，又
>
> $$\{X>x\}\{x\leqslant X\leqslant x+h\}=\{x<X\leqslant x+h\}$$
>
> 有
>
> $$\begin{aligned} \frac{P(x\leqslant X\leqslant x+h\mid X>x)}{h} &= \frac{P(x<X\leqslant x+h)}{h[1-F(x)]} \\ &= \frac{F(x+h)-F(x)}{h[1-F(x)]} \\ &\to \frac{F'(x)}{1-F(x)}=\lambda \end{aligned}$$
>
> 这是一个微分方程，其通解经计算是$\int_{-\infty}^{x}f_{\mathcal{E}}(t)\mathrm{d}t$形式的。

威布尔分布

若考虑老化，则应取失效率随时间而上升，比如取为一个 x 的增函数 λx^m，其中 $\lambda>0,m>0$ 为常数。在这个条件下，寿命分布 $F(x)$ 满足微分方程

$$F'(x)/[1-F(x)]=\lambda x^m$$

此与初始条件 $F(0)=0$ 结合，得出

$$F(x)=1-\mathrm{e}^{-\frac{\lambda}{m+1}x^{m+1}}$$

取 $\alpha=m+1(\alpha>1)$，并把 $\frac{\lambda}{m+1}$ 记为 λ，得出

$$F(x)=1-\mathrm{e}^{-\lambda x^\alpha}\quad(x>0)$$

而当 $x\leqslant 0$ 时 $F(x)=0$。此分布的密度函数为

$$f(x)=\begin{cases}\lambda\alpha x^{\alpha-1}\mathrm{e}^{-\lambda x^\alpha}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

Gamma分布

设 α, β 是正常数，如果 X 的概率密度是

$$f(x)=\begin{cases}\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

则称 X 服从参数是 (α,β) 的Gamma分布，记做 $X\sim\Gamma(\alpha,\beta)$.

> 伽马函数$\Gamma(\alpha)$由积分
>
> $$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}\mathrm{e}^{-x}\,\mathrm{d}x,\alpha>0$$
>
> 定义。对正数 α 和正整数 n，容易验证如下的基本性质:
>
> $$\Gamma(1+\alpha)=\alpha\Gamma(\alpha)，\Gamma(n)=(n-1)!，\Gamma(\frac{1}{2})=\sqrt{\pi}$$

计算连续型随机变量的概率密度

如果开集 D 使得 $P(X\in D)=1$，$g(x)$ 在 D 中连续，使得

$$P(X=x)=g(x)\mathrm{d}x,x\in D$$

则 X 有概率密度

$$f(x)=g(x),x\in D$$

如果 $h(y)$ 在 y 可微，$F(x)$ 在 $x=h(y)$ 有连续的导数 $f(h(y))$，则对 $h(y)=x$

$$P(X=h(y))=|f(h(y))\mathrm{d}h(y)|=f(h(y))\,|h'(y)|\mathrm{d}y$$

> 【对数正态分布】设 $X\sim N\left(\mu,\sigma^2\right)$，计算 $Y=\mathrm{e}^X$ 的概率密度。
>
> 易见 $P(Y>0)=1$, 对 $y>0$, 利用
>
> $$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$
>
> 得到 Y 的概率密度
>
> $$\begin{aligned}f_Y(y)=\frac{P(Y=y)}{\mathrm{d}y}&=\frac{P\left(\mathrm{e}^X=y\right)}{\mathrm{d}y}\\&=\frac{P(X=\ln y)}{\mathrm{d}y}\\&=\frac{|f_X(\ln y)\mathrm{d}(\ln y)|}{\mathrm{d}y}\\&=\frac{1}{y}f_X(\ln y)\\&=\frac{1}{\sqrt{2\pi}\sigma y}\exp\left[-\frac{(\ln y-\mu)^2}{2\sigma^2}\right],y>0\end{aligned}$$
>
> 这时称 Y 服从参数为 $\left(\mu,\sigma^2\right)$ 的对数正态分布。