二项分布/possion分布

二项分布

一、二项分布的概念与特征

（一）成败型实验（Bernoulli实验）

在医学卫生领域的许多实验或观察中，人们感兴趣的是某事件是否发生。如用白鼠做某药物的毒性实验，关心的是白鼠是否死亡；某种新疗法临床实验观察患者是否治愈；观察某指标的化验结果是否呈阳性等。将我们关心的事件A出现称为成功，不出现称为失败，这类试验就称为成-败型实验。指定性资料中的二项分类实验。

成-败型（Bernoulli）实验序列：

满足以下三个条件的n次实验构成的序列称为成-败型实验序列。

1）每次实验结果，只能是两个互斥的结果之一（A或非A）。

2）相同的实验条件下，每次实验中事件A的发生具有相同的概率π。（非A的概率为1-π）。

实际工作中要求π是从大量观察中获得的较稳定的数值。

3）各次实验独立。各次的实验结果互不影响。

（二）二项分布的概率函数

二项分布是指在只能产生两种可能结果（如“阳性”或“阴性”）之一的n次独立重复实验中，当每次试验的“阳性”概率保持不变时，出现“阳性”的次数X=0, 1, 2, …, n的一种概率分布。

若从阳性率为π的总体中随机抽取大小为n的样本，则出现“阳性”数为X的概率分布即呈现二项分布，记作

B(X; n, π)或B(n, π)。

举例 设实验白鼠共3只，要求它们同种属、同性别、体重相近，且他们有相同的死亡概率，即事件“白鼠用药后死亡”为A，相应死亡概率为π。记事件“白鼠用药后不死亡”为 $\overline{A}$ ，相应不死亡概率为1-π。设实验后3只白鼠中死亡的白鼠数为X，则X的可能取值为0，1，2和3，则死亡鼠数为X的概率分布即表现为二项分布。

表 7.1　3 只白鼠各种试验结果及其发生概率

死亡数	未死亡数	试验结果		
X	$3-X$	甲	乙	丙
0	3	生	生	生
1	2	死	生	生
		生	死	生
		生	生	死
2	1	死	死	生
		死	生	死
		生	死	死
3	0	死	死	死

互不相容事件的加法定理

立事件的法定理

构成成-败型实验序列的n次实验中，事件A出现的次数X的概率分布为：

$$P(X)=C_n^X\pi^X(1-\pi)^{n-X}$$

其中X=0，1，2...，n。

n，π是二项分布的两个参数 。

$$C_n^X=\frac{n!}{X!(n-x)!}$$

对于任何二项分布，总有 $\sum_{x=0}^{n}P(X)=1$

P40 例：3.1

- 应用条件：每一种结果在每次试验中都有恒定的概率，试验之间应是独立的。
- N=10,x=3

$$P(mmmfffffff) = \varphi^{3}(1-\varphi)^{7}$$

$$p(x) = C_n^x \varphi^x (1-\varphi)^{n-x}, x = 0,1,2,\cdots,n$$

$$p(x)=C_n^x\varphi^x(1-\varphi)^{n-x},x=0,1,2,\cdots,n$$

$$\begin{aligned}&[\varphi+(1-\varphi)]^n\\&=C_n^0\varphi^0(1-\varphi)^n+C_n^1\varphi^1(1-\varphi)^{n-1}+\cdots+C_n^x\varphi^x(1-\varphi)^{n-x}+\cdots+C_n^n\varphi^n(1-\varphi)^0\\&=p(0)+p(1)+p(2)+\cdots+p(x)+\cdots+p(n)\\&=\sum_{x=0}^{n}p(x)\end{aligned}$$

因为：$\varphi+(1-\varphi)=1$　　所以：

$$\sum_{x=0}^{n}p(x)=[\varphi+(1-\varphi)]^n=1$$

$$p(0)=\frac{10!}{0!(10-1)!}(\frac{1}{2})^0(\frac{1}{2})^{10}=2^{-10}=0.009766$$

$$p(1)=\frac{10!}{1!(10-1)!}(\frac{1}{2})(\frac{1}{2})^9=10(2^{-10})=0.0097656$$

$$p(2)=\frac{10!}{2!(10-1)!}(\frac{1}{2})^2(\frac{1}{2})^8=45(2^{-10})=0.0439453$$

$$p(3)=\frac{10!}{3!(10-1)!}(\frac{1}{2})^3(\frac{1}{2})7=120(2^{-10})=0.1171876$$

抽到3只和3只以下雄性动物的概率：

$$\begin{aligned}F(3)&=P(0)+P(1)+P(2)+P(3)\\&=2^{-10}+10(2^{-10})+45(2^{-10})+120(2^{-10})\\&=176(2^{-10})\\&=0.1718751\end{aligned}$$

杨辉三角

n	系数
0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1

$$[\varphi+(1-\varphi)]^5$$

$$=\varphi^5+5\varphi^4(1-\varphi)+10\varphi^3(1-\varphi)^2+10\varphi^2(1-\varphi)^3+5\varphi(1-\varphi)^4+(1-\varphi)^5$$

例4-2　临床上用针灸治疗某型头疼，有效的概率为60%，现以该疗法治疗3例，其中2例有效的概率是多大？

分析：治疗结果为有效和无效两类，每个患者是否有效不受其他病例的影响，有效概率均为0.6，符合二项分布的条件。

$$P(X)=C_n^X\pi^X(1-\pi)^{n-X}$$

$$P(2)=C_3^2\pi^2(1-\pi)^{3-2}=\frac{3!}{2!(3-2)!}0.6^2(1-0.6)^{3-2}=0.432$$

2例有效的概率是0.432

一例以上有效的概率为：

$$P(X \geq 1) = P(1) + P(2) + P(3)$$

$$= \frac{3!}{1!(3-1)!} 0.61(1-0.6)^{3-1} + \frac{3!}{2!(3-2)!} 0.6^2(1-0.6)^{3-2}$$

$$+ \frac{3!}{3!(3-3)!} 0.6^3(1-0.6)^{3-3} = 0.288 + 0.432 + 0.216$$

$$= 0.936$$

或

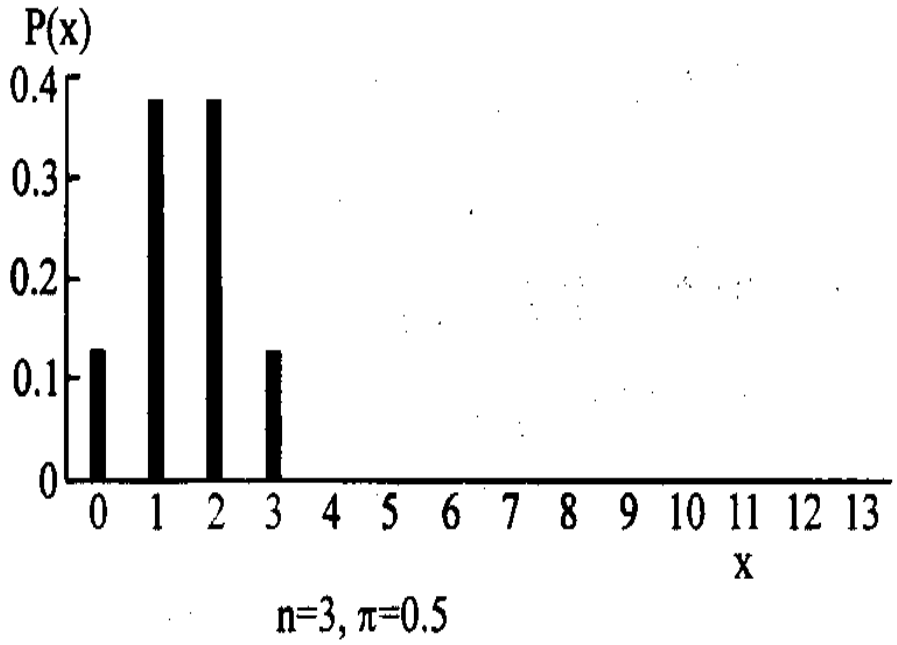
$$P(X \geq 1) = 1 - P(0)$$

$$= 1 - 0.064 = 0.936$$

（三）二项分布的特征

1. 二项分布的图形特征

n，π是二项分布的两个参数，所以二项分布的形状取决于n，π。可以看出，当π =0.5时分布对称，近似对称分布。当π ≠0.5时，分布呈偏态，特别是n较小时，π偏离0.5越远，分布的对称性越差，但只要不接近1和0时，随着n 的增大，分布逐渐逼近正态。因此，π或1- π不太小，而n足够大，我们常用正态近似的原理来处理二项分布的问题。



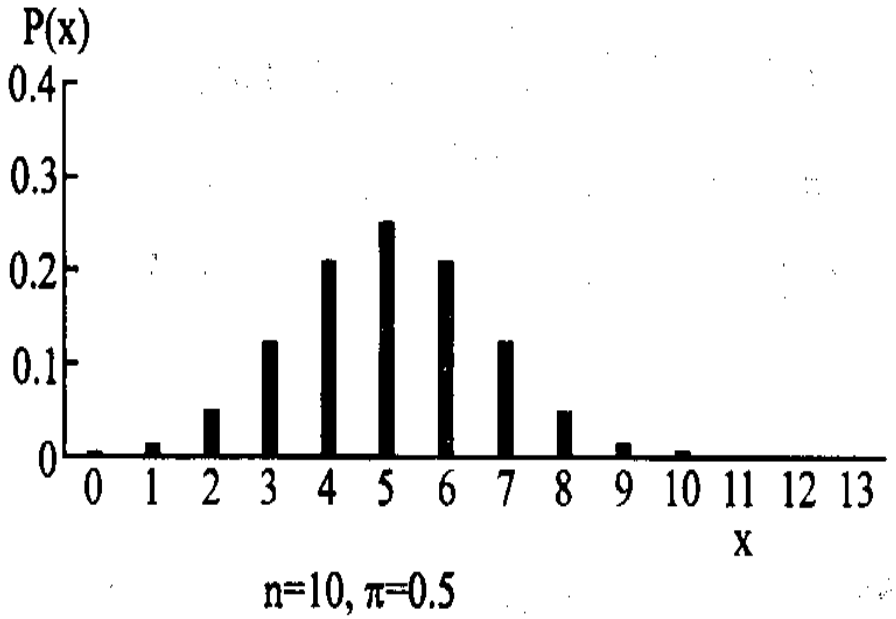
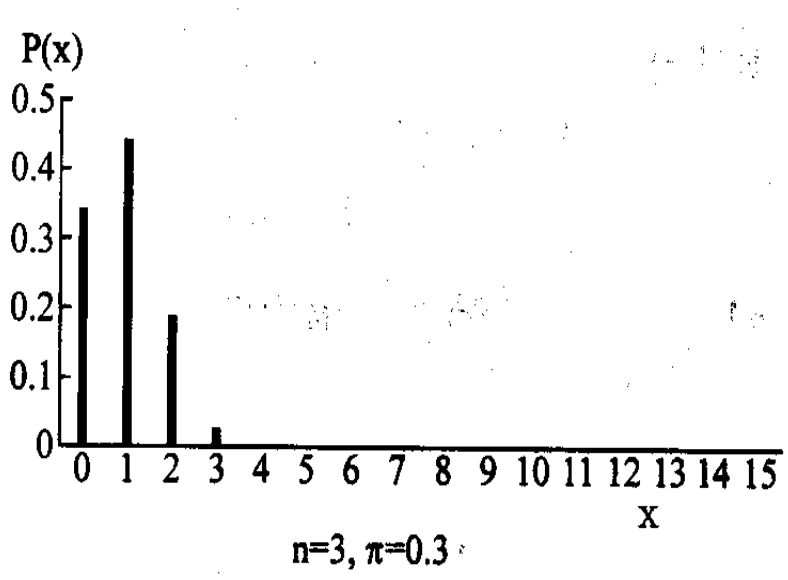
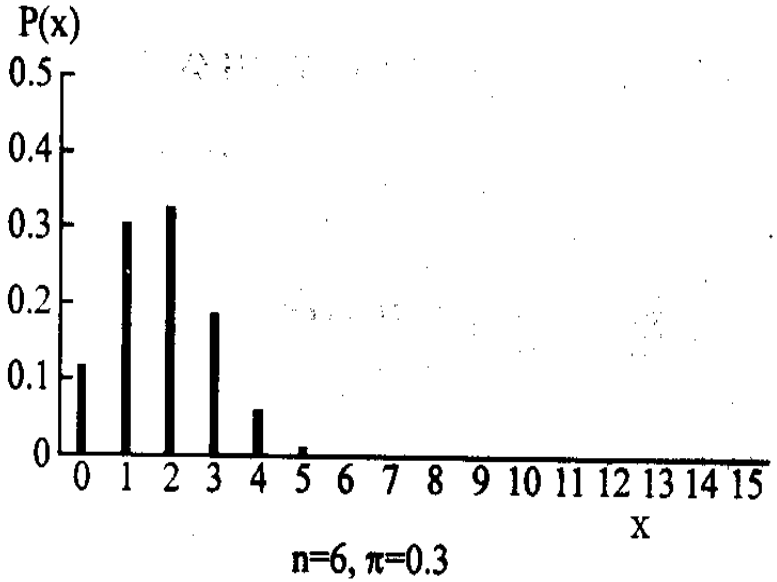
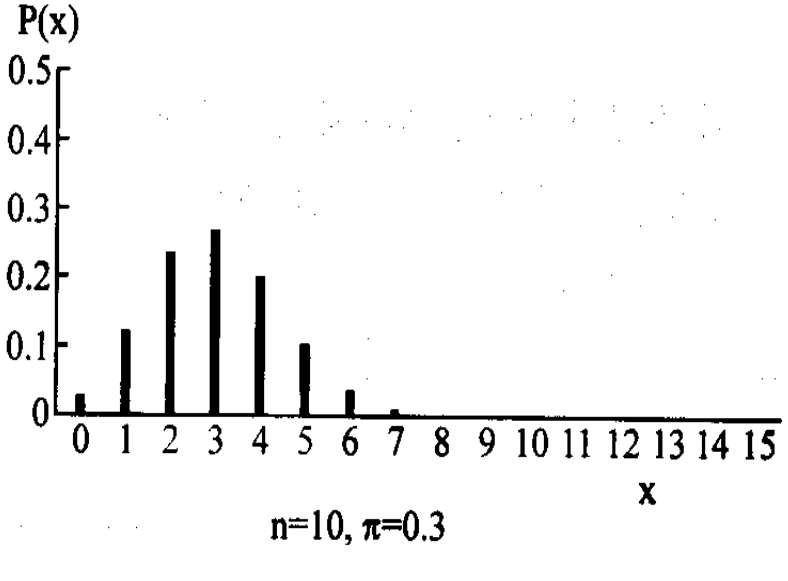


图 4-1　$\pi = 0.5$ 时，不同 n 值对应的二项分布







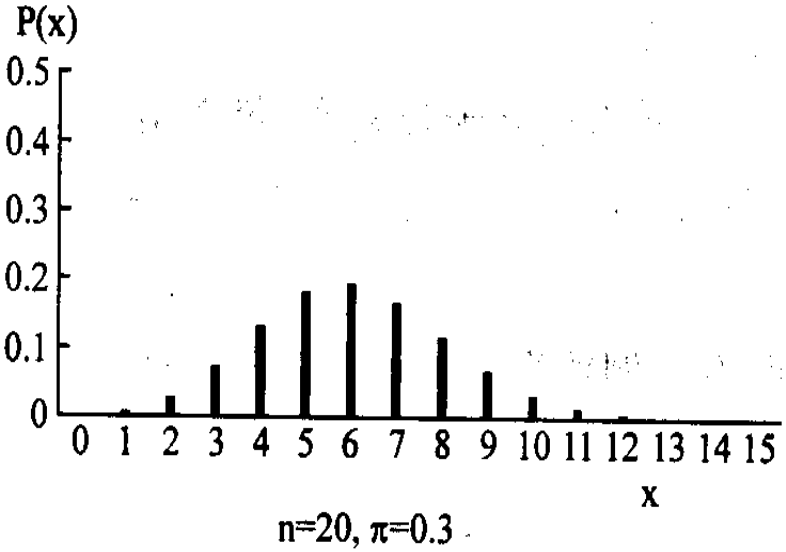


图 4-2　$\pi = 0.3$ 时，不同 n 值对应的二项分布

2. 二项分布的均数和标准差

对于任何一个二项分布B(X;n,π)，如果每次试验出现“阳性”结果的概率均为π，则在n次独立重复实验中，出现阳性次数

X的总体均数为 $\mu = n\pi$

方差为 $\sigma^2 = n\pi(1-\pi)$

标准差为 $\sigma = \sqrt{n\pi(1-\pi)}$

例　实验白鼠3只，白鼠用药后死亡的死亡概率π=0.6，则3只白鼠中死亡鼠数X的总体均数　$\mu = n\pi$

=3×0.6=1.8（只）

方差为　$\sigma^2 = n\pi(1-\pi)$

$= 3\times 0.6\times 0.4 = 0.72$（只）

标准差为　$\sigma = \sqrt{n\pi(1-\pi)}$

$= \sqrt{3\times 0.6\times 0.4} = 0.85$(只）

如果以率表示，将阳性结果的频率记为 $p=\frac{X}{n}$ ，则P的总体均数 $\mu_p=\pi$

总体方差为 $\sigma_p^{\ 2}=\frac{\pi(1-\pi)}{n}$

总体标准差为 $\sigma_p=\sqrt{\frac{\pi(1-\pi)}{n}}$

式中 σ_p 是频率**p的标准误**，反映阳性频率的抽样误差的大小。

例4-4 如果某地钩虫感染率为6.7%,随机观察当地150人,样本钩虫感染率为p,求p的抽样误差 σ_p。

$$n = 150, \pi = 0.067$$

$$\sigma_p = \sqrt{\frac{0.067(1-0.067)}{150}} = 0.02$$

二、二项分布的应用

（一） 概率估计

例4-5 如果某地钩虫感染率为13%，随机观察当地150人，其中有10人感染钩虫的概率有多大?

$$P(X)=C_n^X\pi^X(1-\pi)^{n-X}$$

$$P(10)=C_{150}^{10}0.13^{10}(1-0.13)^{150-10}$$

$$=0.0055$$

(二)单侧累计概率计算

二项分布出现阳性次数至少为K次的概率为

$$P(X \geq K) = \sum_{x=k}^{n} P(X) = \sum_{x=k}^{n} \frac{n!}{X!(n-X)!} \pi^{X} (1-\pi)^{n-X}$$

阳性次数至多为K次的概率为

$$P(X \leq K) = \sum_{x=0}^{k} P(X) = \sum_{x=0}^{k} \frac{n!}{X!(n-X)!} \pi^{X} (1-\pi)^{n-X}$$

例4-6　如果某地钩虫感染率为13%，随机观察当地150人，其中至多有2人感染钩虫的概率有多大?至少有2人感染钩虫的概率有多大?至少有20人感染钩虫的概率有多大?

至多有2名感染的概率为:

$$P(X \le 2) = \sum_{x=0}^{2} P(X) = \sum_{x=0}^{2} \frac{n!}{X!(n-X)!} \pi^{X} (1-\pi)^{n-X}$$

$$P(0) + P(1) + P(2)$$

$$= \frac{150!}{0!150!} 0.13^{0} (1-0.13)^{150} + \frac{150!}{1!149!} 0.13^{1} (1-0.13)^{149}$$

$$+ \frac{150!}{2!148!} 0.132 (1-0.13)^{148}$$

$$= 2.31 \times 10^{-7}$$

至少有2名感染的概率为:

$$P(X \geq 2) = \sum_{x=2}^{n} P(X) = 1 - \sum_{0}^{1} P(X)$$

$$= 1 - [P(0) + P(1)] \approx 1$$

至少有20名感染的概率为:

$$P(X \geq 20) = \sum_{x=20}^{n} P(X) = 1 - \sum_{0}^{19} P(X)$$

$$= 1 - [P(0) + P(1) + P(3) + \ldots . P(19)]$$

$$= 0.4879$$

Poisson分布的概念与特征

一、Poisson分布的概念

Poisson分布也是一种离散型分布，用以描述罕见事件发生次数的概率分布。Poisson分布也可用于研究单位时间内(或单位空间、容积内)某罕见事件发生次数的分布，如分析在单位面积或容积内细菌数的分布，在单位空间中某种昆虫或野生动物数的分布，粉尘在观察容积内的分布，放射性物质在单位时间内放射出质点数的分布等。Poisson分布一般记作 $\Pi(\lambda)$或$\mathbf{P}(\lambda)$ 。

- **Poisson分布作为二项分布的一种极限情况**

Poisson分布可以看作是发生的概率π很小，而观察例数很大时的二项分布。除要符合二项分布的三个基本条件外，Poisson分布还要求π或1-π接近于0和1。有些情况π和n都难以确定，只能以观察单位(时间、空间、容积、面积)内某种稀有事件的发生数X等来表示，如每毫升水中大肠杆菌数，每个观察单位中粉尘的记数，单位时间内放射性质点数等，只要细菌、粉尘、放射性脉冲在观察时间内满足以上条件，就可以近似看为Poisson分布。

二、Poisson分布的特征

1. Poisson分布的概率函数为:

$$P(X) = e^{-\lambda}\frac{\lambda^{X}}{X!}$$

式中 $\lambda = n\pi$ 为Poisson分布的总体均数，X为观察单位时间内某稀有事件的发生次数；e为自然对数的底，为常数，约等于2.71828。

如某地20年间共出生短肢畸形儿10名，平均每年0.5名。就可用 $\lambda = 0.5$ 代入Poisson分布的概率函数来估计该地每年出生此类短肢畸形儿的人数为0，1，2...的概率P(X)。

$$P(X) = e^{-\lambda}\frac{\lambda^X}{X!}$$

$$P(0) = e^{-0.5}\frac{0.5^0}{0!} = 0.607$$

$$P(1) = e^{-0.5}\frac{0.5^1}{1!} = 0.303$$

$$P(2) = e^{-0.5}\frac{0.5^2}{2!} = 0.076$$

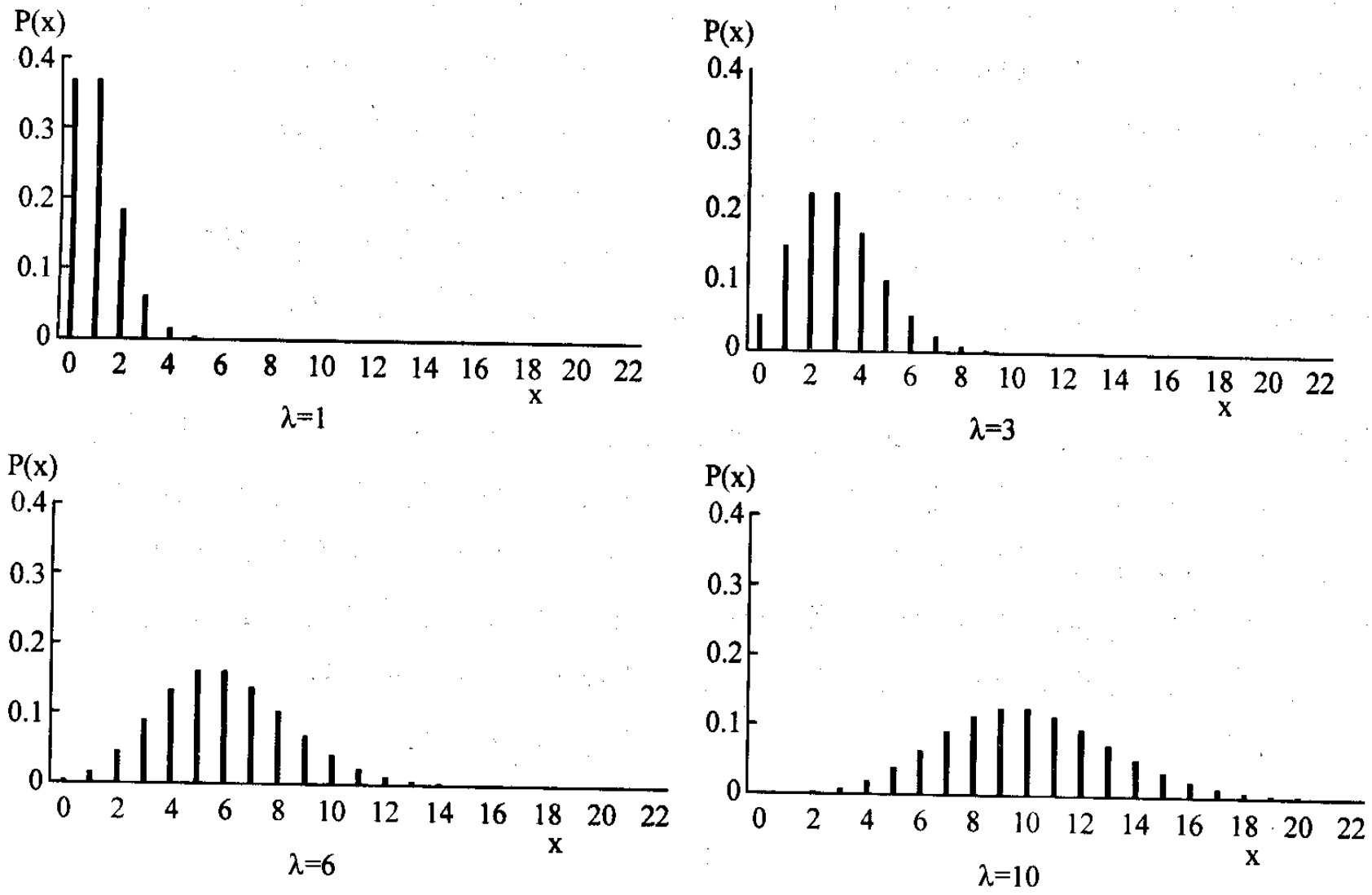


图 4-3　λ 取不同值时的 Poisson 分布图

2. Poisson分布的特性：

（1）Poisson分布的的总体均数与总体方差相等，均为 λ 。

（2）Poisson分布的观察结果有可加性。即对于服从Poisson分布的m个互相独立的随机变量$X_1, X_2 \ldots X_M$，它们之和也服从Poisson分布，其均数为这m个随机变量的均数之和。

从总体均数为 λ_1 的服从Poisson分布总体中随机抽出一份样本，其中稀有事件的发生次数为X_1，再独立地从总体均数为 λ_2 的Poisson分布总体中随机抽出另一份样本，其中稀有事件的发生次数为X_2，则他们的合计发生数$T=X_1+X_2$也服从Poisson分布，总体均数为 $\lambda_1+\lambda_2$ 。

Poisson分布的这些性质还可以推广到多个Poisson分布的情形。例如，从同一水源独立地取水样5次，进行细菌培养，每次水样中的菌落数分别为 $X_i, i=1,2,...5$，均服从Poisson分布，分别记为 $\Pi(\lambda_i), i=1,2,...,5$，把5份水样混合，其合计菌落数 $\sum X_i$ 也服从Poisson分布，记为 $\Pi(\lambda_1+\lambda_2+...+\lambda_5)$ 其均数为 $(\lambda_1+\lambda_2+...+\lambda_5)$ 。

医学研究中常利用Poisson分布的可加性，将小的观察单位合并以增大发生次数X，以便用正态近似法进行统计推断。

二、Poisson分布的应用

（一） 概率估计

例4-7 如果某地新生儿先天性心脏病的发病概率为$8^0/_{00}$，那么该地120名新生儿中有4人患先天性心脏病的概率有多大?

$$\lambda = n\pi = 120 \times 0.008 = 0.96$$

$$P(X) = e^{-\lambda}\frac{\lambda^X}{X!}$$

$$P(4) = e^{-0.96}\frac{0.96^4}{4!} = 0.014$$

(二)单侧累计概率计算

Poisson分布出现阳性次数至多为K次的概率为

$$P(X \le K) = \sum_{x=0}^{k} P(X) = \sum_{x=0}^{k} e^{-\lambda} \frac{\lambda^X}{X!}$$

阳性次数至少为K次的概率为

$$P(X \ge K) = 1 - P(X \le k-1)$$

例4-8　如果某地新生儿先天性心脏病的发病概率为$8^0/_{00}$，那么该地120名新生儿中至多有4人患先天性心脏病的概率有多大?至少有5人患先天性心脏病的概率有多大?

至多有4人患先天性心脏病的概率：

$$P(X \le 4) = \sum_{x=0}^{4} P(X) = \sum_{x=0}^{4} e^{-0.96} \frac{0.96^X}{X!}$$

$$= P(0) + P(1) + P(2) + P(3) + P(4) = 0.977$$

至少有5人患先天性心脏病的概率

$$P(X \ge 5) = 1 - P(X \le 4) = 1 - 0.997 = 0.003$$

例4-9 实验显示某100cm^{2}培养皿平均菌落数为6个，试估计该培养皿菌落数小于3个的概率，大于1个的概率。

该培养皿菌落数小于3个的概率

$$P(X \prec 3) = \sum_{x=0}^{2} P(X) = \sum_{x=0}^{2} e^{-6} \frac{6^X}{X!}$$

$$= P(0) + P(1) + P(2) = 0.062$$

该培养皿菌落数大于1个的概率

$$P(X \succ 1) = 1 - [P(0) + P(1)] = 0.983$$

三、二项分布、Poisson分布的的正态近似

1. 二项分布的正态近似

二项分布的形状取决于n,π，当π=0.5时分布对称，当π≠0.5时，分布呈偏态，特别是n较小时，π偏离0.5越远，分布的对称性越差，随着n的增大，分布逐渐趋向于对称。理论上可以证明，不管π如何，当n相当大时，只要π不接近1和0时，**特别是当nπ或n（1- π ）都大于5时**，二项分布B(X;n,π)近似正态分布N(nπ,nπ(1-π))。

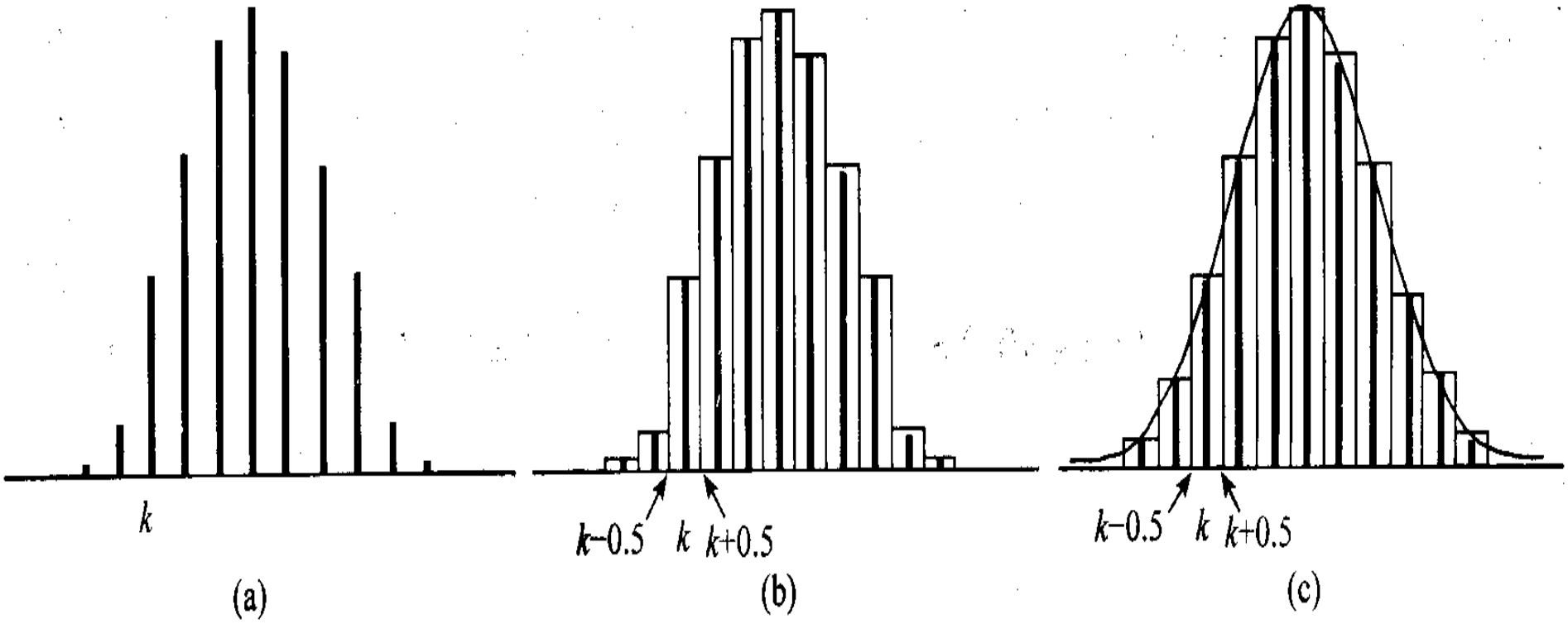


图 4-11　二项分布连续性校正和正态近似示意图

(a)概率函数直条图;(b)连续性校正直方图;(c)正态近似图

二项分布累积概率的正态近似公式为：

$$P(X \le K) = \sum_{x=0}^{k} C_n^X p_x q^{n-x} \approx \Phi\left(\frac{k+0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

$$P(X \ge K) = \sum_{x=k}^{n} C_n^X p_x q^{n-x} \approx 1-\Phi\left(\frac{k-0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

$$P(k_1 \le X \le K_2) \approx \Phi\left(\frac{k_2+0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right) - \Phi\left(\frac{k_1-0.5-n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

为标准正态分布的分布函数

例4-14　如果某地钩虫感染率为13%，随机观察当地150人, 其中至少有20人感染钩虫的概率有多大?

$n\pi = 150 \times 0.13 = 19.5$

$n(1-\pi) = 150 \times (1-0.13) = 130.5$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{150 \times 0.13 \times (1-0.13)} = 4.12$$

$$P(X \geq 20) \approx 1 - \Phi\left(\frac{k - 0.5 - n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

$$P(X \geq 20) \approx 1 - \Phi\left(\frac{20 - 0.5 - 19.5}{4.12}\right) = 1 - \Phi(0) = 0.5$$

至少有20人感染钩虫的概率为50%。

2. Poisson分布的正态近似

Poisson分布，当总体均数 λ 小于5时，λ 越小，分布越呈偏态，随着 λ 的增大，分布逐渐趋向于对称。理论上可以证明，随着 $\lambda \to \infty$，Poisson分布也渐近为正态分布。当 $\lambda \geq 20$ 时，Poisson分布资料可按正态分布处理。

Poisson分布累积概率的正态近似公式为：

$$P(X \le K) \approx \Phi\left(\frac{k+0.5-\lambda}{\sqrt{\lambda}}\right)$$

$$P(X \ge K) = 1 - P(X \langle K) \approx 1 - \Phi\left(\frac{k-0.5-\lambda}{\sqrt{\lambda}}\right)$$

$$P(k_1 \le X \le K_2) \approx \Phi\left(\frac{k_2+0.5-\lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{k_1-0.5-\lambda}{\sqrt{\lambda}}\right)$$

Φ为标准正态分布的分布函数

例4-15　实验显示某放射性物质半小时内发出的脉冲数服从Poisson分布，平均为360个，试估计该放射性物质半小时内发出的脉冲数大于400个的概率。

$$P(X>K)=1-P(X<K)\approx 1-\Phi\left(\frac{k+0.5-\lambda}{\sqrt{\lambda}}\right)$$

$$P(X>400)=1-P(X\le 400)\approx 1-\Phi\left(\frac{400+0.5-360}{\sqrt{\lambda}}\right)$$

$$=1-\Phi(2.13)=0.0166$$

试估计该放射性物质半小时内发出的脉冲数大于400个的概率为1.66%。