



大数据的统计学基础 第四周

DATAGURU专业数据分析社区

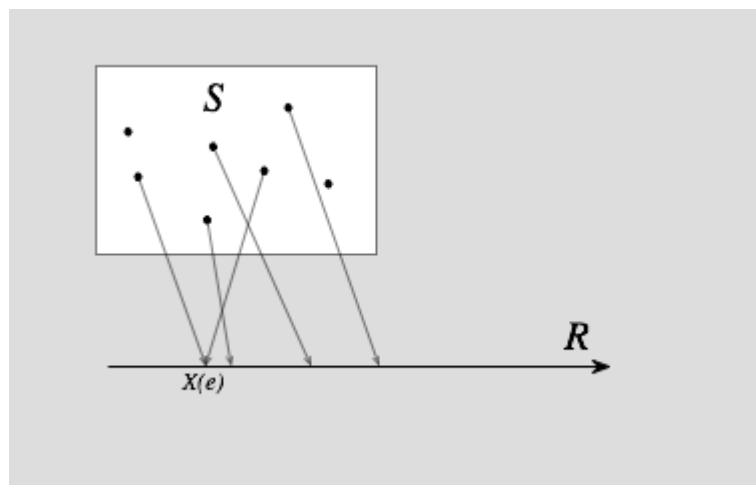
【声明】 本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

<http://edu.dataguru.cn>

随机变量——Random Variable

- ◆ 抛一颗骰子，用 X 记录得到的点数
- ◆ 将一颗硬币抛三次，用 Y 记录三次抛掷得到正面朝上的总数
- ◆ 对于明天的天气，用 $Z = \begin{cases} 1, & \text{下雨} \\ 0, & \text{不下雨} \end{cases}$ 记录明天是否下雨
- ◆ 以上的 X 、 Y 、 Z 都是**随机变量**——一个从样本空间映射到实数域的**函数**
- ◆ **定义**：设随机试验的样本空间为 $S = \{e\}$ ， $X = X(e)$ 是定义在样本空间 S 上的实值单值函数，称 $X = X(e)$ 为随机变量



例 1 在第一章 § 4 例 1 中,将一枚硬币抛掷三次,观察出现正面和反面的情况,样本空间是

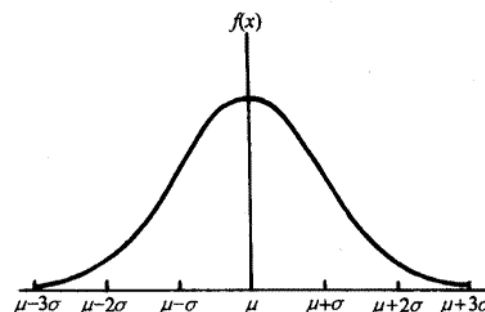
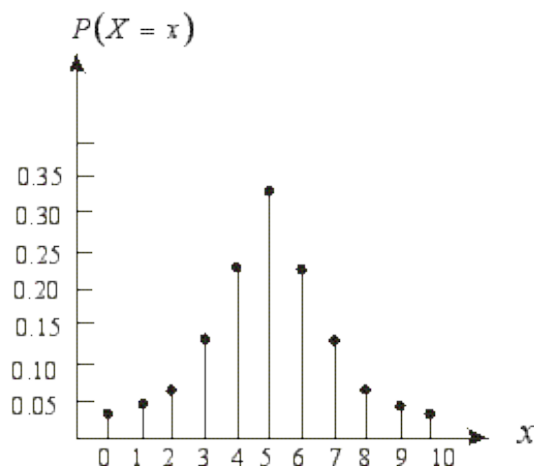
$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}.$$

以 X 记三次投掷得到正面 H 的总数,那么,对于样本空间 $S = \{e\}$ ^①中的每一个样本点 e , X 都有一个数与之对应. X 是定义在样本空间 S 上的一个实值单值函数. 它的定义域是样本空间 S , 值域是实数集合 $\{0, 1, 2, 3\}$. 使用函数记号可将 X 写成

$$X = X(e) = \begin{cases} 3, & e = HHH, \\ 2, & e = HHT, HTH, THH, \\ 1, & e = HTT, THT, TTH, \\ 0, & e = TTT. \end{cases}$$

□

- ◆ 对比X与Y的取值：
- ◆ 1. 用X记录3月份下雨的天数，则 $X=0,1,2,3,\dots,31$ ——X的值可以一一列出
- ◆ 2. 用Y记录3月份降雨量总和，则 $Y \in [0, +\infty)$ ——Y的值不能一一列出
- ◆ 像X这种随机变量，叫做**离散(Discrete)**型随机变量
- ◆ 像Y这种随机变量，叫做**连续(Continuous)**型随机变量
- ◆ **离散型随机变量与连续型随机变量** $\xleftrightarrow{\text{类比于}}$ 自然数与实数



- ◆ 抛一颗骰子，用 X 记录得到的点数
- ◆ 当 $X=1$ 时，意味着得到1点，即事件 $\{X=1\}$ 与事件 $\{\text{得到1点}\}$ 相等，所以 $P(X=1)=P(\text{得到1点})=1/6$
- ◆ 同理可得， $P(X=2)=P(X=3)=P(X=4)=P(X=5)=P(X=6)=1/6$
- ◆ 对于离散型随机变量，随机变量的每一个取值都一定的概率。
- ◆ 如，将一颗硬币抛三次，用 Y 记录三次抛掷得到正面朝上的总数。 $X=2$ 对应于样本点集合 $A=\{HHT, HTH, THH\}$ 。故 $P(X=2)=P(A)=3/8$
- ◆ 求正面向上次数不多于一次的概率：
- ◆ $P(X \leq 1) = P(X=0) + P(X=1) = P\{HTT, THT, TTH, TTT\} = 4/8 = 1/2$

分布律——Distribution law

- ◆ 试验：将一颗硬币抛三次。用X记录硬币在三次抛掷中正面向上的次数。将X的所有可能取值相对应的概率算出来。
- ◆ 样本空间： $S=\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$
- ◆ X所有可能的取值：0, 1, 2, 3
- ◆ $P(X=0)=P\{TTT\}=1/8$
- ◆ $P(X=1)=P\{HTT, THT, TTH\}=3/8$
- ◆ $P(X=2)=P\{HHT, HTH, THH\}=3/8$
- ◆ $P(X=3)=P\{HHH\}=1/8$

X	0	1	2	3
P	1/8	3/8	3/8	1/8

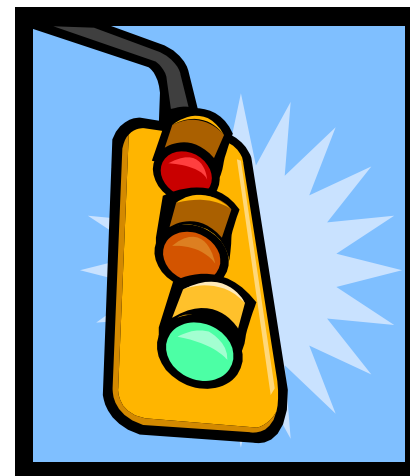


离散型随机变量X的分布律

分布律——Distribution law

- ◆ 例：某人骑自行车从学校到火车站，一路上要经过3个独立的交通灯，设各灯工作独立，且设各灯为红灯的概率为 p ， $0 < p < 1$ ，以 X 表示首次停车时所通过的交通灯数，求 X 的概率分布律。
- ◆ 设 $A_i = \{\text{第}i\text{个灯为红灯}\}$ ，则 $P(A_i) = p$ ， $i = 1, 2, 3$ 。且 A_1, A_2, A_3 相互独立。
- ◆ X 所有可能的取值：0, 1, 2, 3
- ◆ $P(X=0) = P(A_1) = p$
- ◆ $P(X=1) = P(\overline{A_1}A_2) = (1-p)p$
- ◆ $P(X=2) = P(\overline{A_1}\overline{A_2}A_3) = (1-p)(1-p)p$
- ◆ $P(X=3) = P(\overline{A_1}\overline{A_2}\overline{A_3}) = (1-p)(1-p)(1-p)$

X	0	1	2	3
p	p	$p(1-p)$	$(1-p)^2 p$	$(1-p)^3$



分布律——Distribution law

- ◆ 例：从生产线上随机抽产品进行检测，设产品的次品率为 p ， $0 < p < 1$ ，若查到一只次品就得停机检修，设停机时已检测到 X 只产品，试写出 X 的概率分布律。
- ◆ 设 $A_i = \{\text{第}i\text{次抽到次品}\}$ ， $i=1,2,\dots$ ；则 A_1, A_2, \dots 相互独立。
- ◆ $P(X=0) = P(A_1) = p$
- ◆ $P(X=1) = P(\overline{A_1}A_2) = (1-p)p$
- ◆ $P(X=2) = P(\overline{A_1}\overline{A_2}A_3) = (1-p)(1-p)p$
- ◆
- ◆ $P(X=k) = P(\overline{A_1}\overline{A_2} \dots \overline{A_{k-1}}A_k) = (1-p)^{k-1}p$ ——当情况太多时，可以使用统一的公式表示分布律

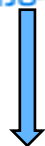
X	0	1	2	k
P	p	$(1-p)p$	$(1-p)^2p$		$(1-p)^{k-1}p$	

(0-1) 分布——the Bernoulli distribution

- ◆ 例：某人骑自行车从学校到火车站，一路上要经过1个交通灯，设该灯为绿灯的概率为 p , $0 < p < 1$, 以 X 表示首次 停车时所通过的交通灯数，求 X 的概率分布律。
- ◆ 如果只经过一个交通灯，那么 X 的取值只能是0或1。分布律变为：
- ◆ $P(X=0)=1-p$
- ◆ $P(X=1)=p$
- ◆ 即 $P(X=k)=p^k(1-p)^{1-k}$, $k=0,1$

X	0	1
P	$1-p$	p

- ◆ 像这种，随机变量 X 只能取0和1的情况，我们称 X 服从以 p 为参数的 (0-1) 分布或是两点分布。



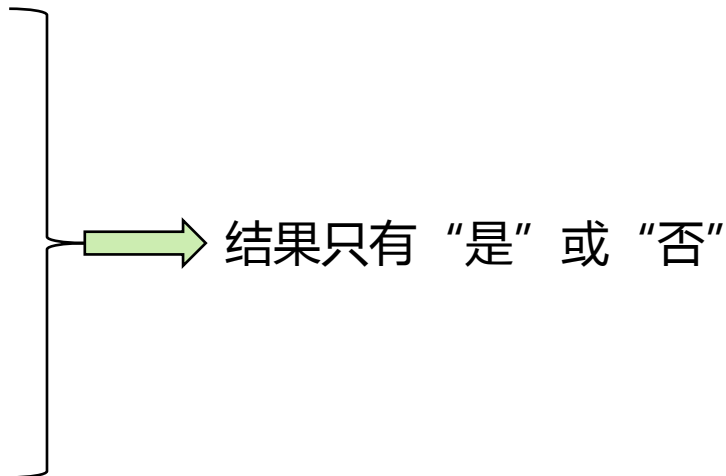
试验的可能结果
只分为两种情况

(0-1) 分布——the Bernoulli distribution

- ◆ 最常见的 (0-1) 分布：
- ◆ 抛一次硬币，用 X 记正面向上的次数，那么
- ◆ $P(X=0)=1/2$; $P(X=1)=1/2$ 。 这里， $p=1/2$.
- ◆ 其他 (0-1) 分布：
- ◆ 对于明天的天气，用 $Z = \begin{cases} 1, & \text{下雨} \\ 0, & \text{不下雨} \end{cases}$ 记录明天是否下雨
- ◆ 若明天下雨的概率是 p ，则 $P(X=0)=1-p$ ； $P(X=1)=p$



伯努利试验——Bernoulli trial

- ◆ 像上面提到的抛硬币，明天是否下雨等试验，可能结果只有两个：{正面向上，正面向下}与{明天下雨，明天不下雨}，我们称这一类试验为伯努利试验
 - ◆ 刚出生的小孩是个女孩吗？
 - ◆ 一个人的双眼是绿色的吗？
 - ◆ 在有蚊子的地方喷洒杀虫剂，蚊子会死掉吗？
 - ◆ 一个可能是顾客的人会买我的产品吗？
 - ◆ 公民（citizen）会投给特定的候选人吗？
 - ◆ 雇员会投票支持工会吗？
- 
- 结果只有“是”或“否”
- ◆ 一般情况下，我们将“是”的情况称为“成功”，“否”的情况称为“失败”。将“成功”的概率记为 p 。
 - ◆ 当“成功”时，记随机变量 $X=1$ ；当“失败”时，记随机变量 $X=0$ 。 $P(X=1)=p$ ，则 X 服从以 p 为参数的（0-1）分布

n重伯努利试验——Bernoulli process

- ◆ 将一个伯努利试验独立地重复n次，称这一串重复的独立试验为n重伯努利试验
- ◆ 试验E是抛一个硬币，观察得到的是正面向上还是正面向下。如果将这个硬币重复抛n次，那么就是一个n重伯努利试验。
- ◆ 试验E是从一个装有a个白球，b个黑球的箱子中任取一个球，观察得到的是白球还是黑球。若连续做10次有放回抽样，则这是一个10重伯努利试验。
- ◆ 从52张牌中有放回地取n次，设 $A = \{\text{取到红牌}\}$ ，则每次只有两个结果： A 或是 \bar{A} 。
- ◆ 若是从52张牌中无放回地取n次，这就不是一个n重伯努利试验了。例如，若第一次取出一张红牌，那么第二次取到红牌的概率就变为 $12/51 \neq 13/52$ ， $P(A)$ 的概率变了，就不是重复的试验了。



二项分布——Binomial distribution

- ◆ 对于n重伯努利试验，如将一个硬币抛掷n次，用X记录正面向上的次数。计算X的分布律。
- ◆ X的可能取值：0, 1, 2.....n
- ◆ 当n=4时：
- ◆ 样本空间： $S=\{HHHH, HHHT.....TTTT\}$ ，共 $2*2*2*2=2^4$ 种情况，每一种情况的可能性为 $\frac{1}{2^4}$
- ◆ $P(X=0)=P(\text{四次反面向上})=\frac{1}{2^4}$
- ◆ $P(X=1)=P(\text{第一次正面向上，其余反面向上})+P(\text{第二次正面向上，其余反面向上})+P(\text{第三次正面向上，其余反面向上})+P(\text{第四次正面向上，其余反面向上})=$
$$4 * \frac{1}{2^4} = \binom{4}{1} \frac{1}{2^4}$$

二项分布——Binomial distribution

- ◆ $P(X=2)=P(\text{第一次正面向上, 第二次正面向上, 第三次反面向上, 第四次反面向上})+P(\text{第一次正面向上, 第二次反面向上, 第三次正面向上, 第四次反面向上})+.....+P(\text{第一次反面向上, 第二次反面向上, 第三次正面向上, 第四次正面向上})$

$$=\binom{4}{2}\frac{1}{2^4}$$

共6种情况

- ◆ $P(X=3)=P(\text{第一次反面向上, 其余正面向上})+P(\text{第二次反面向上, 其余正面向上})+P(\text{第三次反面向上, 其余正面向上})+P(\text{第四次反面向上, 其余正面向上})=$

$$4 \times \frac{1}{2^4} = \binom{4}{3}\frac{1}{2^4}$$

- ◆ $P(X=4)=P(\text{HHHH})=\frac{1}{2^4}$

- ◆ 归纳：

- ◆ X的分布律： $P(X=k)=\binom{n}{k}\frac{1}{2^n}$ ——X服从n，1/2的二项分布

二项分布——Binomial distribution

- ◆ 在一个n重伯努利试验中，事件A（成功）发生的次数记为X，则X是一个随机变量， $P(A)=p$ 。
- ◆ 如：某人骑自行车从学校到火车站，一路上要经过n个独立的交通灯，设各灯工作独立，且设各灯为红灯的概率为p， $0 < p < 1$ ，以X表示一路上遇到红灯的次数，求X的概率分布律。
- ◆ X的可能取值：0，1，2.....，n，A={遇到红灯}
- ◆ 设 $A_i = \{ \text{第}i\text{次}A\text{发生} \}$ ，先设 $n=3$

$$P(X = 0) = P(\bar{A}_1 \bar{A}_2 \bar{A}_3) = (1 - p)^3$$

$$P(X = 1) = P(A_1 \bar{A}_2 \bar{A}_3 \cup \bar{A}_1 A_2 \bar{A}_3 \cup \bar{A}_1 \bar{A}_2 A_3) = C_3^1 p^1 (1 - p)^{3-1}$$

$$P(X = 2) = P(A_1 A_2 \bar{A}_3 \cup A_1 \bar{A}_2 A_3 \cup \bar{A}_1 A_2 A_3) = C_3^2 p^2 (1 - p)^{3-2}$$

$$P(X = 3) = P(A_1 A_2 A_3) = p^3$$

二项分布——Binomial distribution

◆ 通过上面的计算，我们可以归纳出：

一般 $P(X = k) = C_n^k p^k (1 - p)^{n-k}, k = 0, 1, 2, \dots, n$

◆ 这时我们称X服从参数为n，p的**二项分布**。其分布律为 $P(X=k) = \binom{n}{k} p^k (1 - p)^k$ ，记为 $X \sim B(n, p)$

◆ 特别地，当n=1时，二项分布就是（0-1）分布。

◆ 抛10次硬币，用X记录正面向上的次数，则 $X \sim B(10, 0.5)$

◆ 同时抛6颗骰子，可以看做是抛一颗骰子6次。用Y记录6点出现的次数，则 $Y \sim B(6, 1/6)$



- ◆ 某人独立射击400次，设每次命中率为0.02， $0 < p < 1$ ，设命中 X 次，
- ◆ (1) 求 X 的概率分布律；(2) 求至少有两次次命中的概率。

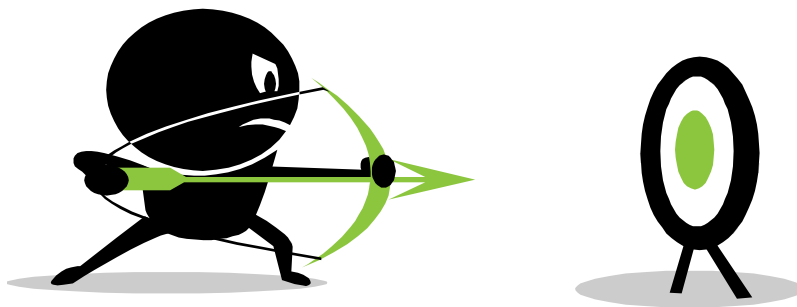
解 将一次射击看成是一次试验. 设击中的次数为 X , 则 $X \sim b(400, 0.02)$.
 X 的分布律为

$$P\{X = k\} = \binom{400}{k} (0.02)^k (0.98)^{400-k}, k = 0, 1, \dots, 400.$$

于是所求概率为

$$\begin{aligned} P\{X \geq 2\} &= 1 - P\{X = 0\} - P\{X = 1\} \\ &= 1 - (0.98)^{400} - 400(0.02)(0.98)^{399} = 0.9972. \end{aligned}$$

□



例 2 按规定,某种型号电子元件的使用寿命超过 1 500 小时的为一级品. 已知某一大批产品的一级品率为 0.2,现在从中随机地抽查 20 只. 问 20 只元件中恰有 k 只($k=0,1,\dots,20$)为一级品的概率是多少?

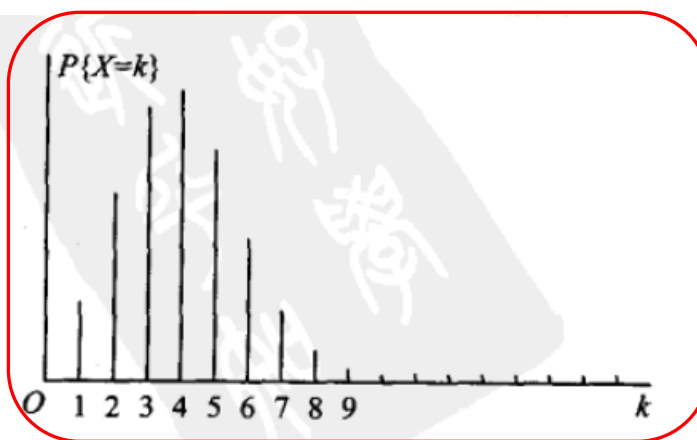
解 这是不放回抽样. 但由于这批元件的总数很大,且抽查的元件的数量相对于元件的总数来说又很小,因而可以当作放回抽样来处理,这样做会有一些误差,但误差不大. 我们将检查一只元件看它是否为一级品看成是一次试验,检查 20 只元件相当于做 20 重伯努利试验. 以 X 记 20 只元件中一级品的只数,那么, X 是一个随机变量,且有 $X \sim b(20, 0.2)$. 由 (2.6) 式即得所求概率为

$$P\{X=k\} = \binom{20}{k} (0.2)^k (0.8)^{20-k}, \quad k=0,1,\dots,20.$$

将计算结果列表如下：

$P\{X=0\}=0.012$	$P\{X=4\}=0.218$	$P\{X=8\}=0.022$
$P\{X=1\}=0.058$	$P\{X=5\}=0.175$	$P\{X=9\}=0.007$
$P\{X=2\}=0.137$	$P\{X=6\}=0.109$	$P\{X=10\}=0.002$
$P\{X=3\}=0.205$	$P\{X=7\}=0.055$	
当 $k \geq 11$ 时, $P\{X=k\} < 0.001$		

从图 2-3 中看到,当 k 增加时,概率 $P\{X=k\}$ 先是随之增加,直至达到最大值(本例中当 $k=4$ 时取到最大值),随后单调减少. 我们指出,一般,对于固定的 n 及 p ,二项分布 $b(n, p)$ 都具有这一性质. \square



概率密度分布图

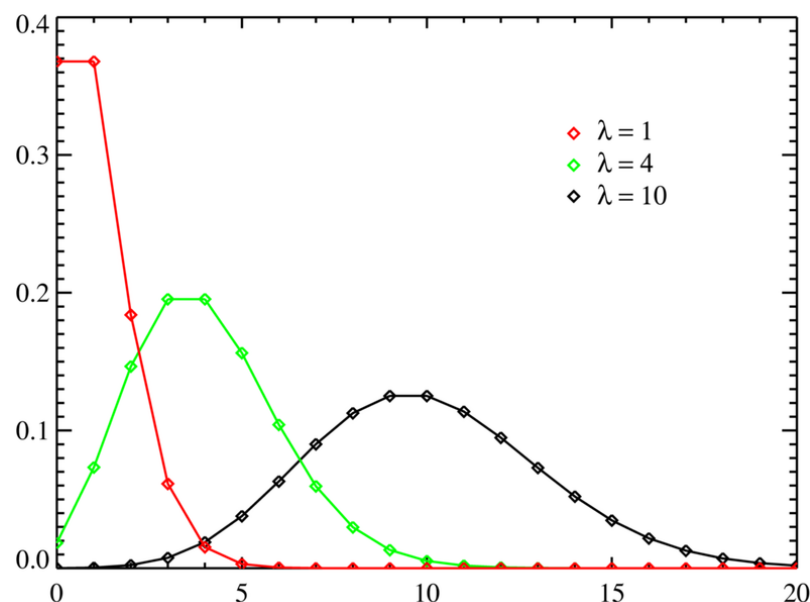
泊松分布——Poisson distribution

◆ 泊松定理—— $\lim_{n \rightarrow \infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$, 其中 $\lambda = np$

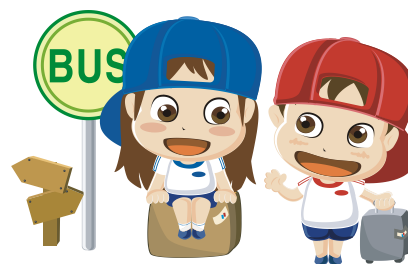
◆ 泊松分布：设随机变量X的所有可能取值为0,1,2,.....X的分布律为

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

◆ 其中 $\lambda > 0$ 是一个常数。则称X服从参数为 λ 的泊松分布，记为 $X \sim \pi(\lambda)$



- ◆ 泊松分布的提出，是作为二项分布的一个近似。
- ◆ 当 p 相当小（一般是 $p \leq 0.1$ ）时，有近似公式 $\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k e^{-\lambda}}{k!}$ ， $\lambda = np$
- ◆ 实际应用场景：
 - ◆ 1. 社会生活，对服务的各种要求：某一医院在一天内的急诊病人数目，某一网站访问数，公共汽车站来到的乘客数等等
 - ◆ 2. 物理科学：放射性分裂落到某区域的质点数，热电子的发射等



例 5 计算机硬件公司制造某种特殊型号的微型芯片,次品率达 0.1%,各芯片成为次品相互独立. 求在 1000 只产品中至少有 2 只次品的概率. 以 X 记产品中的次品数, $X \sim b(1000, 0.001)$.

解 所求概率为

$$\begin{aligned} P\{X \geq 2\} &= 1 - P\{X=0\} - P\{X=1\} \\ &= 1 - (0.999)^{1000} - \binom{1000}{1} (0.999)^{999} (0.001) \\ &\approx 1 - 0.3676954 - 0.3680635 \approx 0.2642411. \end{aligned}$$

利用(2.7)式来计算得, $\lambda = 1000 \times 0.001 = 1$,

$$\begin{aligned} P\{X \geq 2\} &= 1 - P\{X=0\} - P\{X=1\} \\ &= 1 - e^{-1} - e^{-1} \approx 0.2642411. \end{aligned}$$

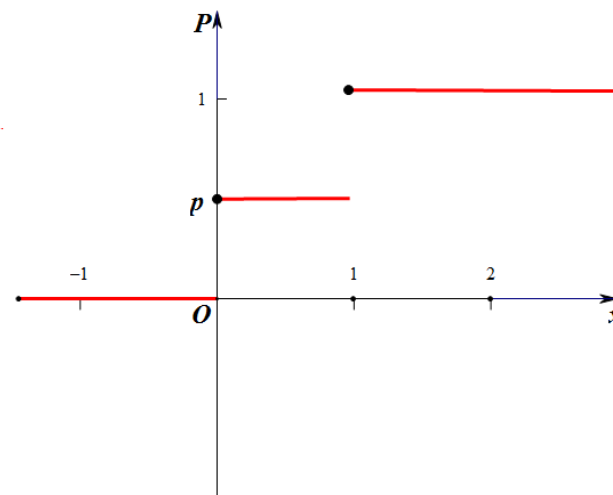
□

显然利用(2.7)式的计算来得方便. 一般, 当 $n \geq 20, p \leq 0.05$ 时用 $\frac{\lambda^k e^{-\lambda}}{k!}$ ($\lambda =$

np) 作为 $\binom{n}{k} p^k (1-p)^{n-k}$ 的近似值效果颇佳.

分布函数——Cumulative Distribution Function

- ◆ 对于连续型随机变量，由于其可能的取值不能一一列出，所以就不能像离散型随机变量那样使用分布律去描述它。这时我们需要更加通用的描述方式——分布函数
- ◆ 设 X 是一个随机变量， x 是任意实数，函数 $F(x) = P\{X \leq x\}$ 称为 X 的分布函数（累积分布函数）（英文简写 CDF）。
- ◆ 分布函数的性质：
 - ◆ 1. $F(x)$ 是一个不减函数。
 - ◆ 2. $0 \leq F(x) \leq 1$ ，且 $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$;
 - ◆ $F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$
 - ◆ 3. $F(x)$ 是右连续的。



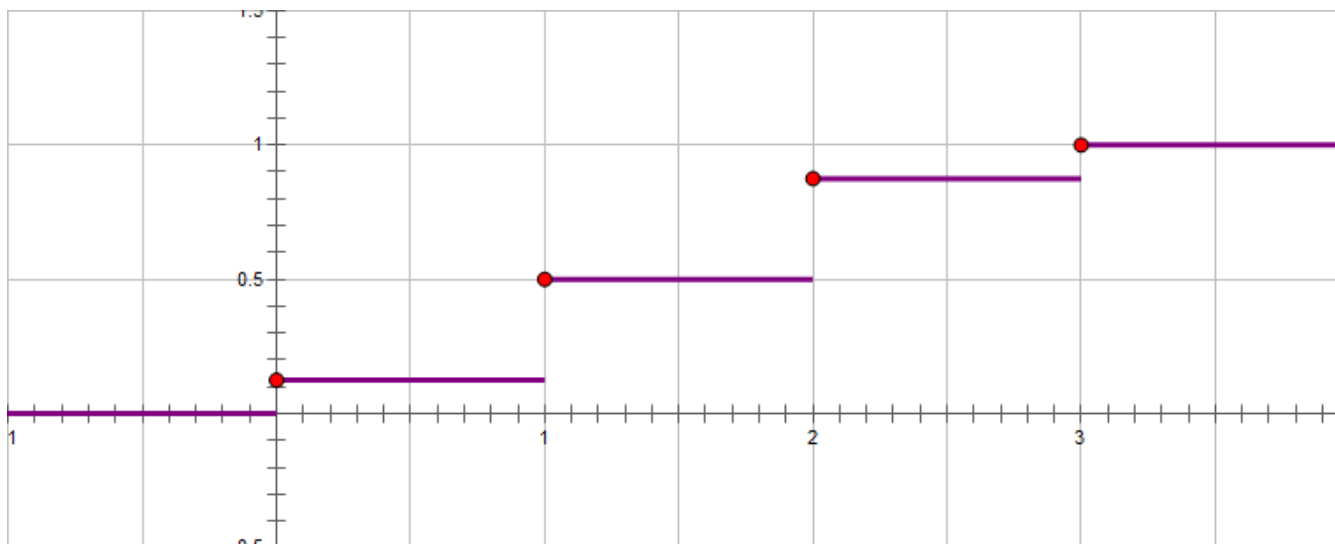
(0-1)分布的分布函数图象

◆ 将一颗硬币重复抛3次，X记录正面向上的次数。求X的分布函数。

◆ 先把X的分布律写出

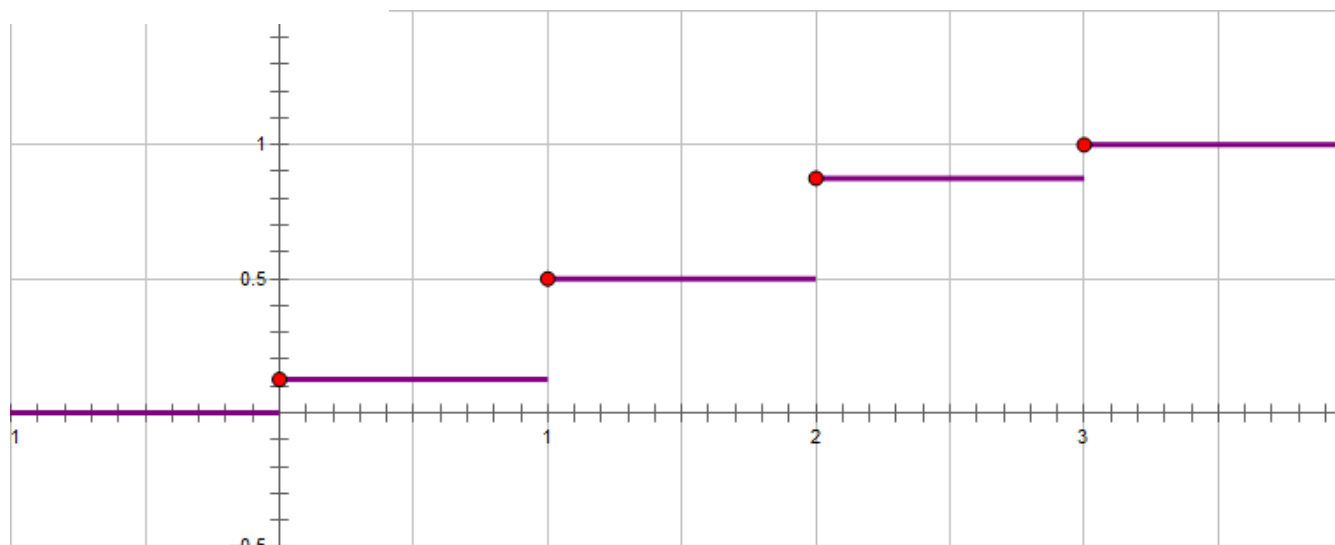
X	0	1	2	3
P	1/8	3/8	3/8	1/8
累积概率	1/8	1/2	7/8	1

◆ 将累积概率图（分布函数画出）



- ◆ 从图上可以看出，分布函数是一个分段函数，其中0,1,2,3（随机变量的可能取值）是断点

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/8, & 0 \leq x < 1 \\ 1/2, & 1 \leq x < 2 \\ 7/8, & 2 \leq x < 3 \\ 1, & x \geq 3 \end{cases}$$



例 1 设随机变量 X 的分布律为

X	-1	2	3
p_k	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

求 X 的分布函数, 并求 $P\left\{X \leq \frac{1}{2}\right\}$, $P\left\{\frac{3}{2} < X \leq \frac{5}{2}\right\}$, $P\{2 \leq X \leq 3\}$.

$$F(x) = \begin{cases} 0 & x < -1, \\ P\{X = -1\}, & -1 \leq x < 2, \\ P\{X = -1\} + P\{X = 2\}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

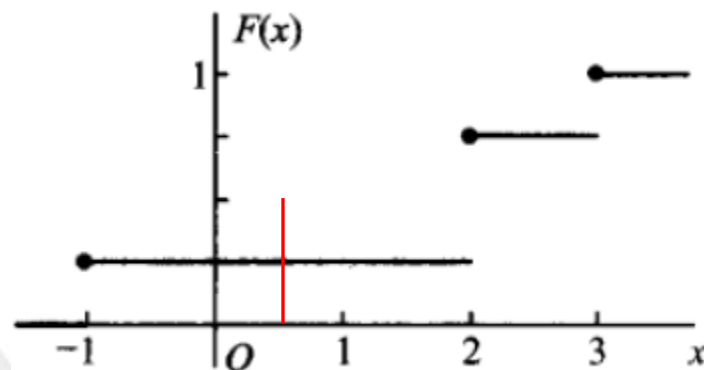
即

$$F(x) = \begin{cases} 0, & x < -1, \\ \frac{1}{4}, & -1 \leq x < 2, \\ \frac{3}{4}, & 2 \leq x < 3, \\ 1, & x \geq 3. \end{cases}$$

$$P\left\{X \leq \frac{1}{2}\right\} = F\left(\frac{1}{2}\right) = \frac{1}{4},$$

$$P\left\{\frac{3}{2} < X \leq \frac{5}{2}\right\} = F\left(\frac{5}{2}\right) - F\left(\frac{3}{2}\right) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}.$$

$$\begin{aligned} P\{2 \leq X \leq 3\} &= F(3) - F(2) + P\{X=2\} \\ &= 1 - \frac{3}{4} + \frac{1}{2} = \frac{3}{4}. \end{aligned}$$



- ◆ 有了分布函数之后，随机变量 X 在某个取值范围的概率很容易求出
- ◆ 离散型随机变量的分布函数都是分段函数，跳跃点在随机变量的可能取值上。
- ◆ 一般地，若离散型随机变量 X 的分布律为 $P(X = x_k) = p_k$ ，那么其分布函数就是
- ◆ $F(x) = P(X \leq x) = \sum_{x_k \leq x} p_k$

例 2 一个靶子是半径为 2 m 的圆盘, 设击中靶上任一同心圆盘上的点的概率与该圆盘的面积成正比, 并设射击都能中靶, 以 X 表示弹着点与圆心的距离. 试求随机变量 X 的分布函数.

解 若 $x < 0$, 则 $\{X \leq x\}$ 是不可能事件, 于是

$$F(x) = P\{X \leq x\} = 0.$$

若 $0 \leq x \leq 2$, 由题意, $P\{0 \leq X \leq x\} = kx^2$, k 是某一常数, 为了确定 k 的值, 取 $x = 2$, 有 $P\{0 \leq X \leq 2\} = 2^2 k$, 但已知 $P\{0 \leq X \leq 2\} = 1$, 故得 $k = 1/4$, 即

$$P\{0 \leq X \leq x\} = \frac{x^2}{4}.$$

于是

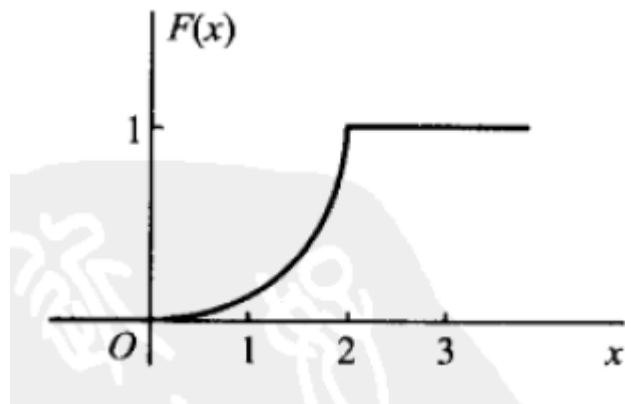
$$F(x) = P\{X \leq x\} = P\{X < 0\} + P\{0 \leq X \leq x\} = \frac{x^2}{4}.$$

若 $x \geq 2$, 由题意 $\{X \leq x\}$ 是必然事件, 于是

$$F(x) = P\{X \leq x\} = 1.$$

综合上述, 即得 X 的分布函数为

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x^2}{4}, & 0 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$



另外,容易看到本例中的分布函数 $F(x)$, 对于任意 x 可以写成形式

$$F(x) = \int_{-\infty}^x f(t) dt,$$

其中

$$f(t) = \begin{cases} \frac{t}{2}, & 0 < t < 2, \\ 0, & \text{其他.} \end{cases}$$

- ◆ 严格定义：
- ◆ 对于随机变量X的分布函数F(x)，存在非负可积函数f(x)，使对于任意实数x有

$$F(x) = \int_{-\infty}^x f(t)dt$$

- ◆ 则称X为连续型随机变量，f(x)称为X的**概率密度函数**
(**Probability Density Function**)，简称概率密度 (PDF)。
- ◆ 公共汽车每15分钟一班，某人在站台等车时间X是个随机变量，X的取值范围是[0,15)，它是一个区间，从理论上说在这个区间内可取任一实数3.5、 $\sqrt{20}$ 等，因而称这随机变量是连续型随机变量。
- ◆ 如：小明每天在7:00到8:00这段时间出门上学，X为小明出门的准确时间，那么X就是一个连续型随机变量。

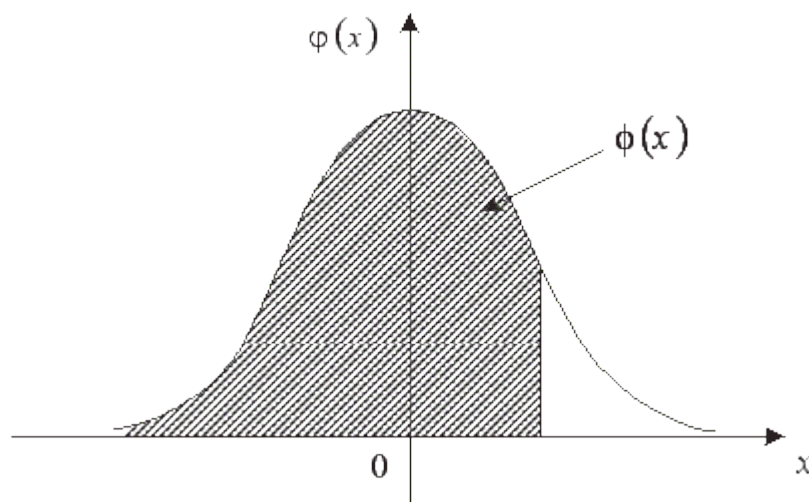
◆ 性质：

◆ 1. $f(x) \geq 0$

◆ 2. $\int_{-\infty}^{\infty} f(x)dx = F(\infty) = 1$

◆ 3. 对于任意实数 $x_1, x_2 (x_1 \leq x_2)$, $P\{x_1 < X \leq x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx$

◆ 4. 若 $f(x)$ 在点 x 处连续，则有 $F'(x) = f(x)$

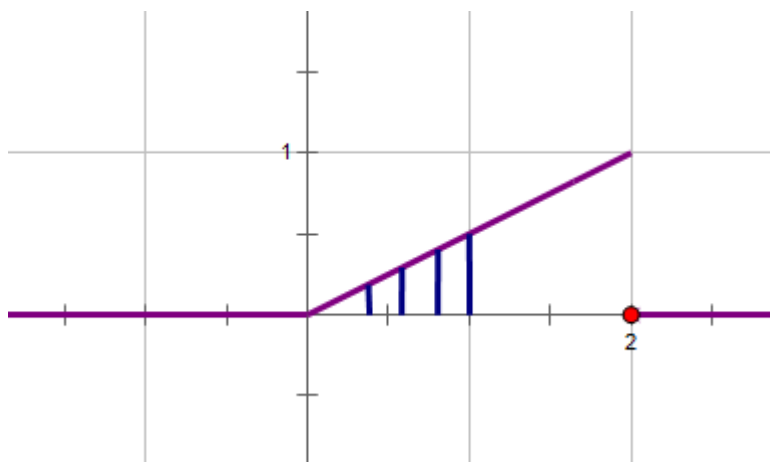


例 2 一个靶子是半径为 2 m 的圆盘,设击中靶上任一同心圆盘上的点的概率与该圆盘的面积成正比,并设射击都能中靶,以 X 表示弹着点与圆心的距离. 试求随机变量 X 的分布函数.



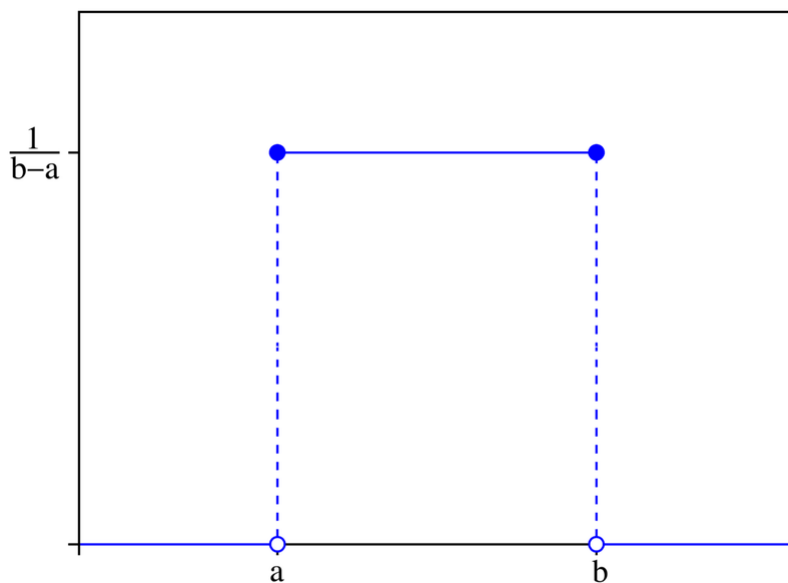
◆ 在这个例子中, X 的概率密度函数是 $f(x) = \begin{cases} \frac{x}{2}, & 0 < x < 2 \\ 0, & \text{其他} \end{cases}$

◆ 根据图象, 若是要求 $P(0 < X \leq 1) = 1/4$ 则是求概率密度曲线与 $x=0, x=1, y=0$ 所围成的图形面积

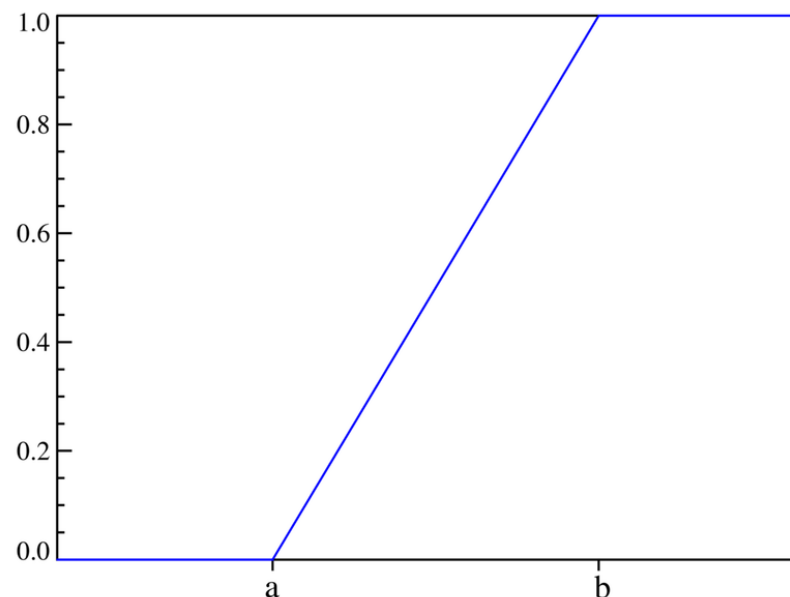


均匀分布——Uniform distribution

- ◆
- ◆ 若连续函数X具有概率密度 $f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{其他} \end{cases}$, 则称X在区间(a,b)上服从均匀分布 , 记为 $X \sim U(a,b)$



概率密度函数

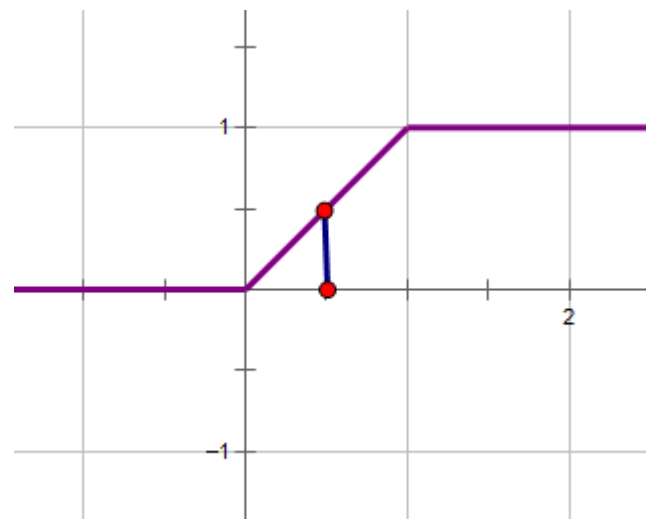


累积分布函数

均匀分布——Uniform distribution

- ◆ 在 $(0, 1)$ 上任意取出一个实数，记为 X 。则 $X \sim U(0,1)$
- ◆ 根据定义，很容易求出 $P(X \leq 0.5) = F(0.5) = \int_0^{0.5} 1 dx = 0.5$
- ◆ $P(X < 0.5) = P(X \leq 0.5) - P(X = 0.5) = 0.5$

对于连续型随机变量 X ， X 等于某个特定值的概率很小，基本可以看做是 0



例 2 设电阻值 R 是一个随机变量, 均匀分布在 $900\ \Omega \sim 1100\ \Omega$. 求 R 的概率密度及 R 落在 $950\ \Omega \sim 1\ 050\ \Omega$ 的概率.

解 按题意, R 的概率密度为

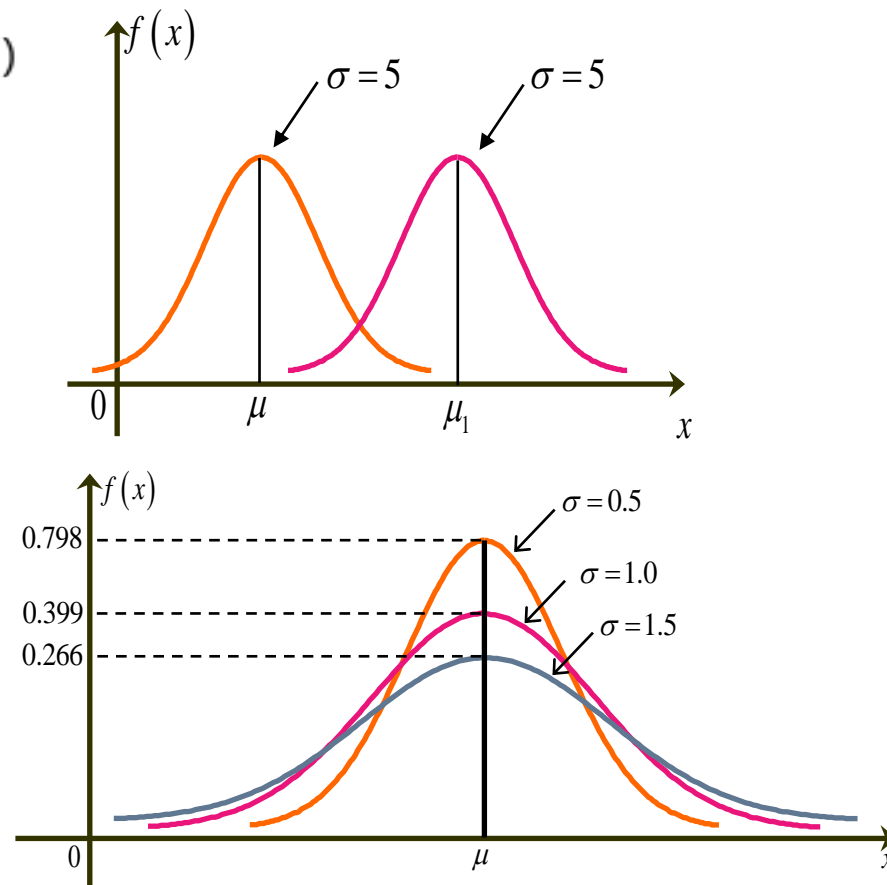
$$f(r) = \begin{cases} \frac{1}{1\ 100 - 900}, & 900 < r < 1100, \\ 0, & \text{其他.} \end{cases}$$

故有

$$P\{950 < R \leq 1\ 050\} = \int_{950}^{1\ 050} \frac{1}{200} dr = 0.5. \quad \square$$

正态分布——Normal distribution

- ◆ 若连续型随机变量X的概率密度为 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x < \infty$, 则称X服从参数为 μ , σ^2 的**正态分布**, 记为 $X \sim N(\mu, \sigma^2)$
- ◆ 正态分布也称为高斯分布
- ◆ 当参数 μ 与 σ^2 的取值不同, 正态分布的概率密度函数图象也有所不同。
- ◆ 其中, μ 是位置参数, 控制图象的对称轴位置所在; σ^2 是尺度参数, 控制钟型曲线的矮胖程度。 σ^2 越大, 曲线越矮胖。



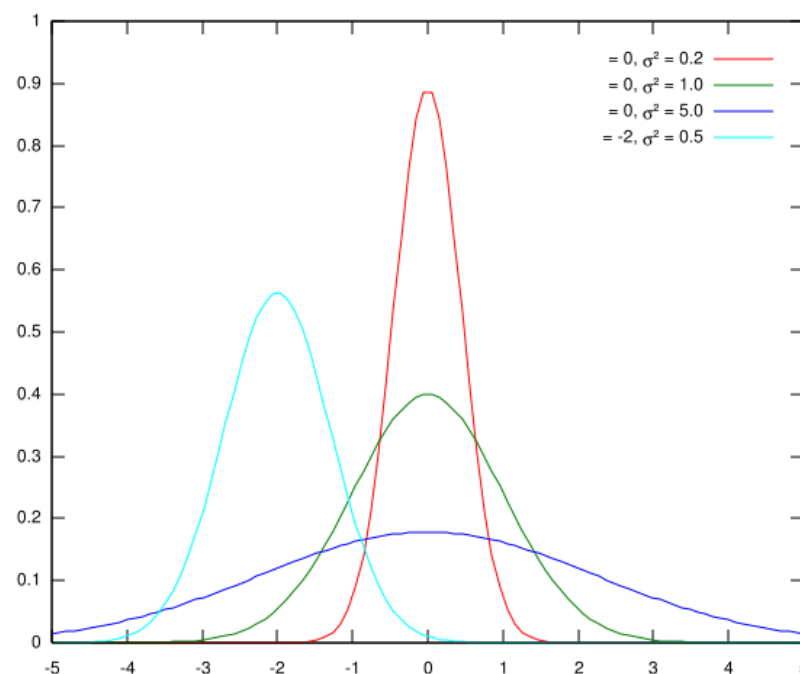
正态分布——Normal distribution

◆ 性质：

◆ 1. 曲线关于 $x = \mu$ 对称

◆ 2. 当 $x = \mu$ 时概率密度函数可以取得最大值 $f(x) = \frac{1}{\sqrt{2\pi}\sigma}$

◆ 3. 在具有同样长度的区间中，当区间离 μ 越远，X落在这个区间的概率越小

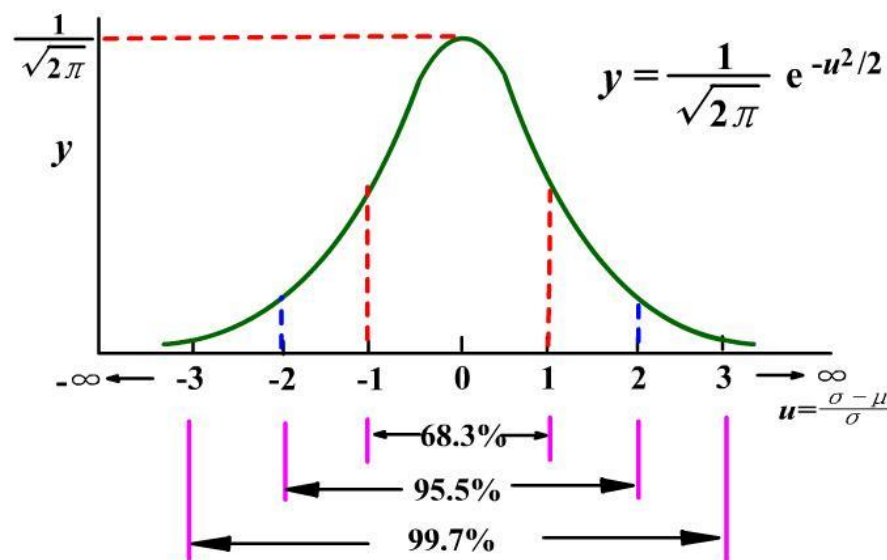


标准正态分布

- ◆ 当 $\mu=0$ ， $\sigma^2=1$ 时，称 X 服从标准正态分布，其概率密度和分布函数分布用 $\varphi(x)$ ， $\Phi(x)$

表示。
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- ◆ 在标准正态分布中，随机变量 X 落
(-1,1) 的概率是68.3%
- ◆ 已知， $\Phi(-x) = 1 - \Phi(x)$
- ◆ 如 $X \sim N(0,1)$ ，则 $P\{X < -2\} = 1 - P\{X < 2\}$
 $= P\{X > 2\}$
- ◆ 对于标准正态分布，人们已经编制了
 $\Phi(x)$ 的函数表，可以直接查表求特定值
- ◆ 见书附表2 P382



标准正态分布曲线

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389

- ◆ $X \sim N(0,1)$, 则
- ◆ $P(X \leq 0.55) = 0.7088$
- ◆ $P(X < -0.98) = 1 - P(X < 0.98) = 1 - 0.8365 = 0.1635$
- ◆ $P(X > 0.4) = 1 - P(X \leq 0.4) = 1 - 0.6554 = 0.3446$

正态分布→标准正态分布

- ◆ 对于一般正态分布，可以变换变为标准正态分布
- ◆ 若 $X \sim N(\mu, \sigma^2)$ ，则 $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- ◆ 所以 $F(x) = P\{X \leq x\} = P\left\{\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right\} = \Phi\left(\frac{x - \mu}{\sigma}\right)$
- ◆ 如，若 $X \sim N(1, 4)$ ，则 $P\{X \leq 1.6\} = \Phi\left(\frac{1.6 - 1}{2}\right) = \Phi(0.3) = 0.6179$

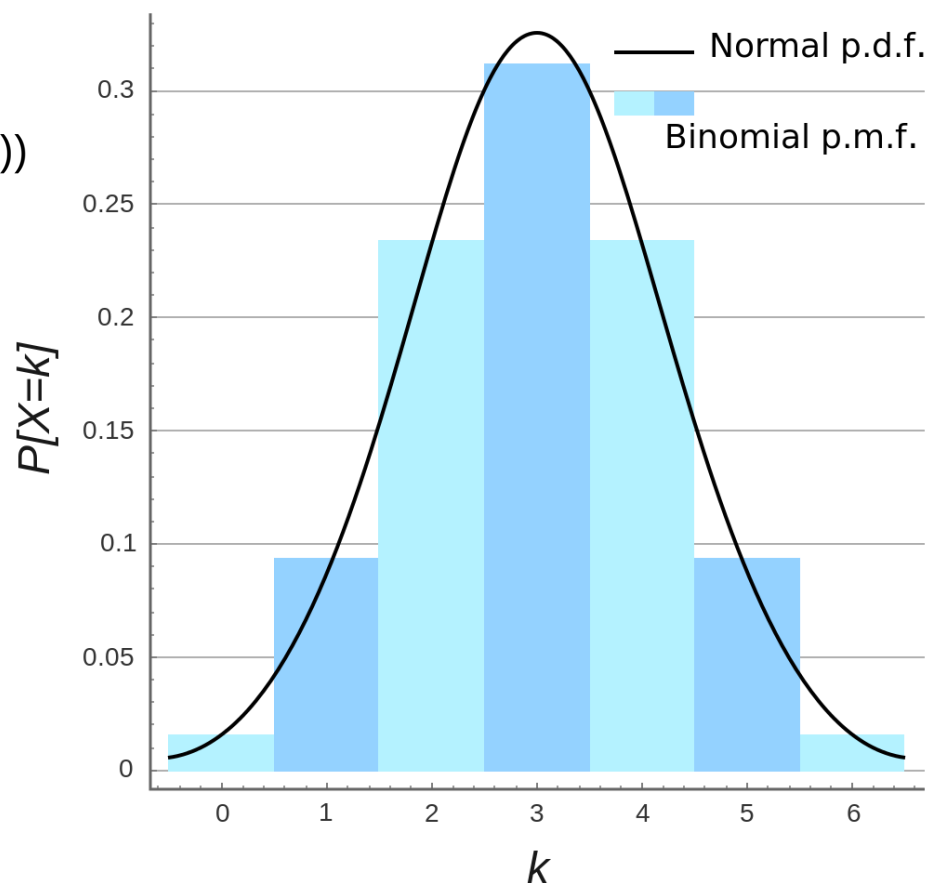
- ◆ 设某地区男子身高 $X(\text{cm}) \sim N(169.7, 4.1^2)$
- ◆ (1) 从该地区随机找一男子测身高，求他的身高大于175cm的概率；
- ◆ (2) 若从中随机找5个男子测身高，问至少有一人身高大于175cm的概率是多少？恰有一人身高大于175cm的概率为多少？
- ◆ (1) $P(X > 175) = 1 - P(X \leq 175) = 1 - \Phi\left(\frac{175-169.7}{4.1}\right) = 1 - \Phi(1.293) = 1 - 0.9015 = 0.0985$
- ◆ (2) 设5人中有 Y 人身高大于175cm，则 $Y \sim B(5, 0.0985)$

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1 - p)^5 = 0.4045$$

$$P(Y = 1) = C_5^1 p^1 (1 - p)^4 = 0.3253$$

二项分布与正态分布

- ◆ 二项分布是离散情况下的正态分布。
- ◆ 当 n 足够大时，可以用正态分布近似二项分布，从而避免二项分布中繁杂的计算。
- ◆ 若 $X \sim B(n, p)$ ，当 n 足够大时，有 X 近似服从正态分布 $N(np, np(1-p))$



- ◆ 正态分布的前世今生
- ◆ <http://cos.name/2013/01/story-of-normal-distribution-1/>
- ◆ <http://cos.name/2013/01/story-of-normal-distribution-2/>



- ◆ **Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。**
- ◆ **关于逆向收费式网络的详情，请看我们的培训网站 <http://edu.dataguru.cn>**



Thanks

FAQ时间