浙江大學 心理与行为科学系
DEPARTMENT OF PSYCHOLOGY AND BEHAVIORAL SCIENCES
ZHEJIANG UNIVERSITY

2.概率论和
数理统计回顾

戴俊毅

研究员/长聘副教授

随机试验

- 可以在相同的条件下重复进行
- 每次试验的可能结果不止一个，并且事先知道试验的所有可能结果
- 每次进行试验前不能确定哪一个结果会出现

样本空间和随机事件

- 样本空间：随机试验的所有可能结果组成的集合
- 随机事件：样本空间的子集
- 基本事件：包含唯一的可能结果的随机事件

随机变量

- 随机变量是由随机试验的结果到实数的函数（映射）
- 例如，就投掷硬币这个随机试验而言，样本空间为{正面、反面}
- 假定正面映射到实数1，反面映射到实数0，那么这样的一个映射，就是一个随机变量。其输入为随机试验的结果，输出为实数1或者0。其随机性来源于试验结果的随机性，且取1的概率，等于试验结果为正面的概率，取0的概率，等于试验结果为反面的概率。
- 一般用大写字母代表随机变量（映射），小写字母代表随机变量的可能取值

离散vs.连续随机变量

- 离散：可能取值的个数有限，或者可列无限多的随机变量
- 例如，新生儿的性别
- 连续：可能取值的个数无限多且不可列的随机变量
- 例如，新生儿的体重

累积分布函数

- 对于随机变量X，函数$F(x) = \Pr(X \leq x)$称为X的（累积）分布函数(CDF; cumulative distribution function)
- 累积分布函数的输入是某一实数值，输出是随机变量的取值不大于该实数值的概率

概率质量函数（分布律）

- 对离散随机变量，函数$f(x) = \Pr(X = x)$称为概率质量函数
- $\Pr(X = x_k) = p_k, k = 1, 2, 3, \ldots$称为随机变量X的分布律，其中$x_k$代表X的可能取值
- $p_k > 0, \sum_k p_k = 1$

概率密度函数

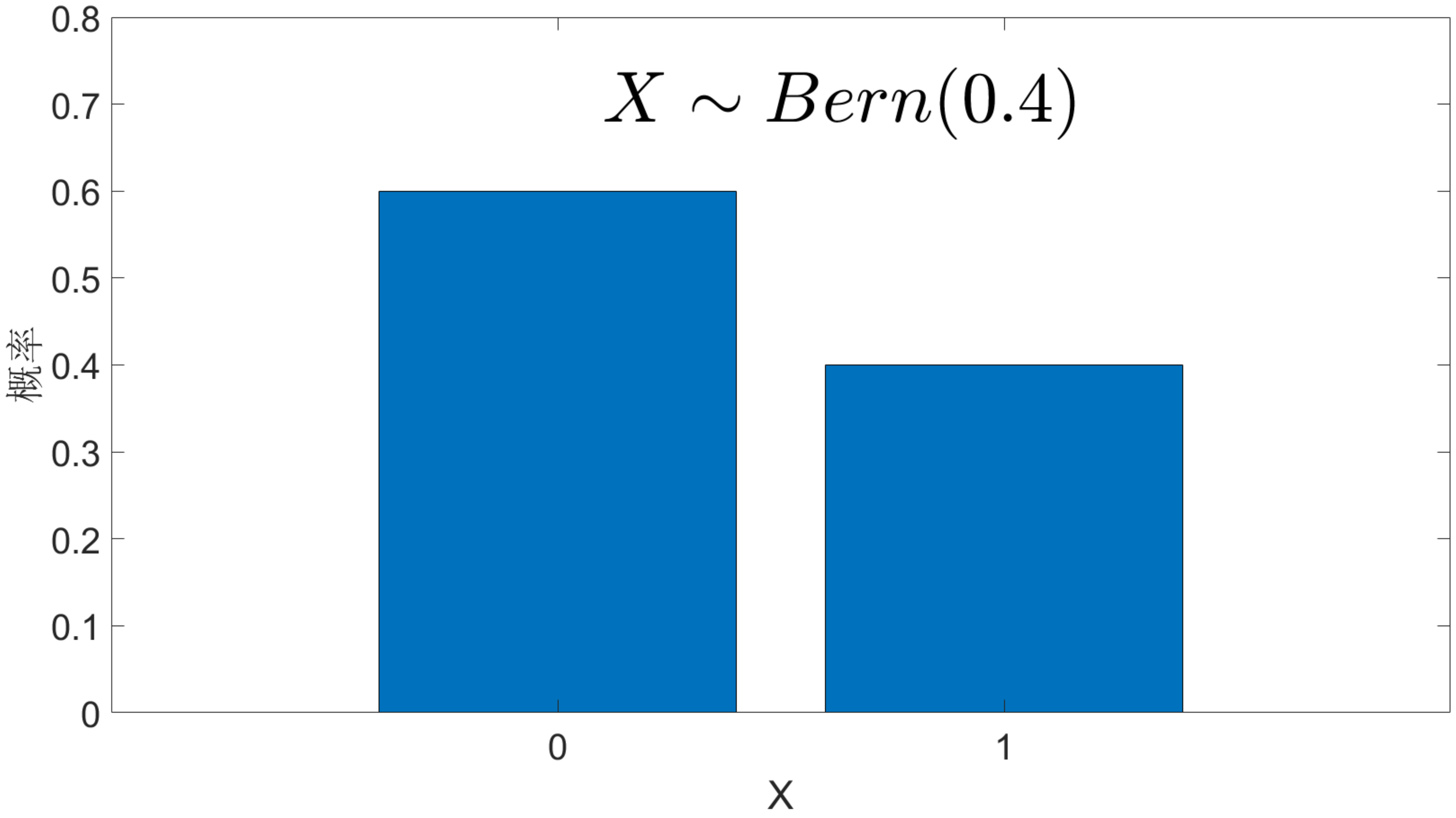
- 对连续随机变量X，f(x) = dF(x)/dx 称为X的概率密度函数，其中F(x)是X的累积分布函数
- 由于累积分布函数的输入是某一实数值，输出是随机变量的取值不大于该实数值的概率，概率密度可以理解为累积概率随着取值范围扩大而增长的速度
- 概率不可能超过1，但概率密度可能超过1

常用离散概率分布

- 伯努利分布
- 二项分布
- 泊松分布

伯努利分布

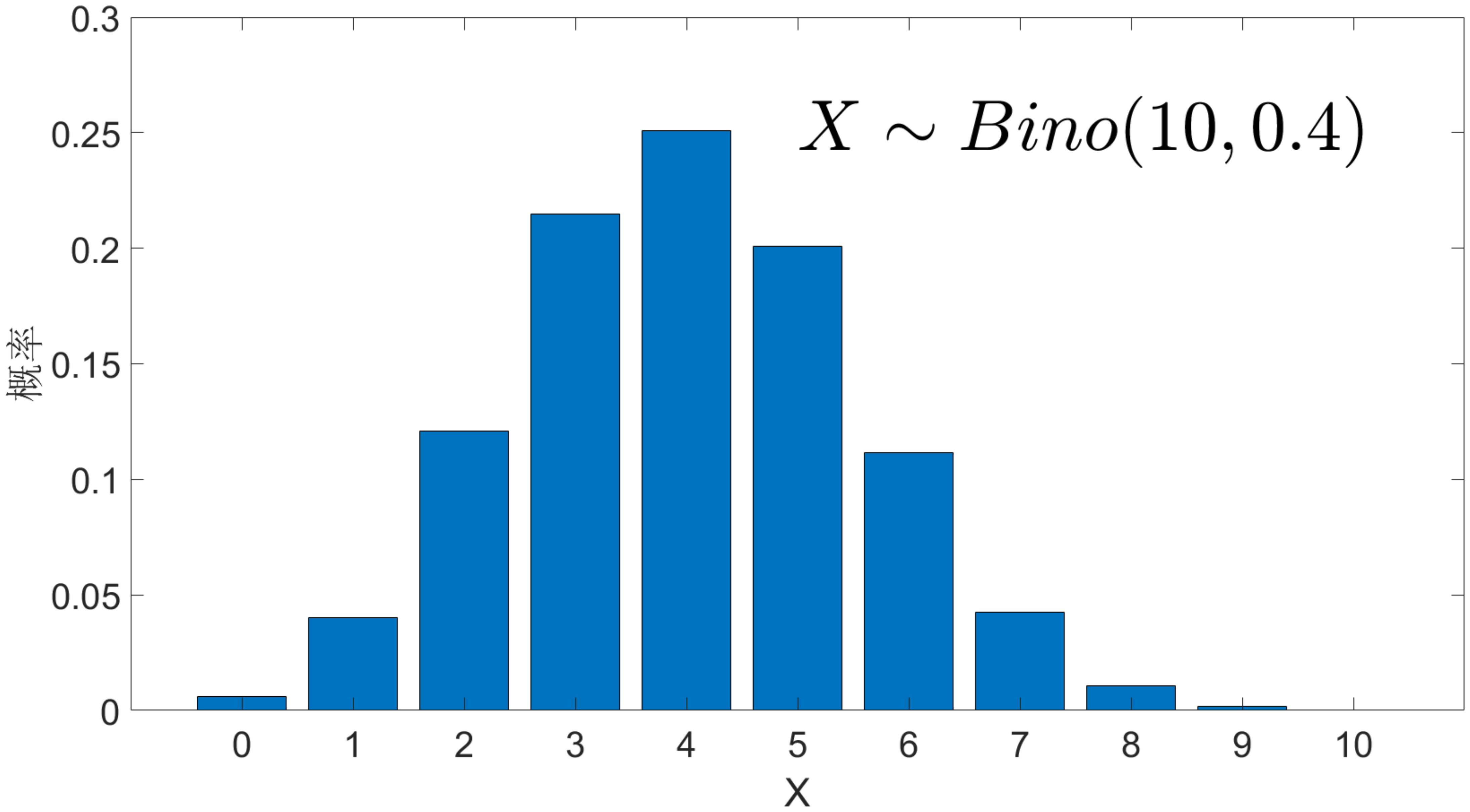
- 伯努利试验：只有两种结果的随机试验
- 伯努利分布：$Pr\{X = z\} = p^z(1-p)^{1-z}, z = 0$或1
- 其中X=0对应伯努利试验的一种结果，X=1对应伯努利试验的另一种结果，0<p<1代表每次试验出第一种结果的概率
- 记作X ~ Bern(p)



$X \sim Bern(0.4)$
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
概率
0
1
X

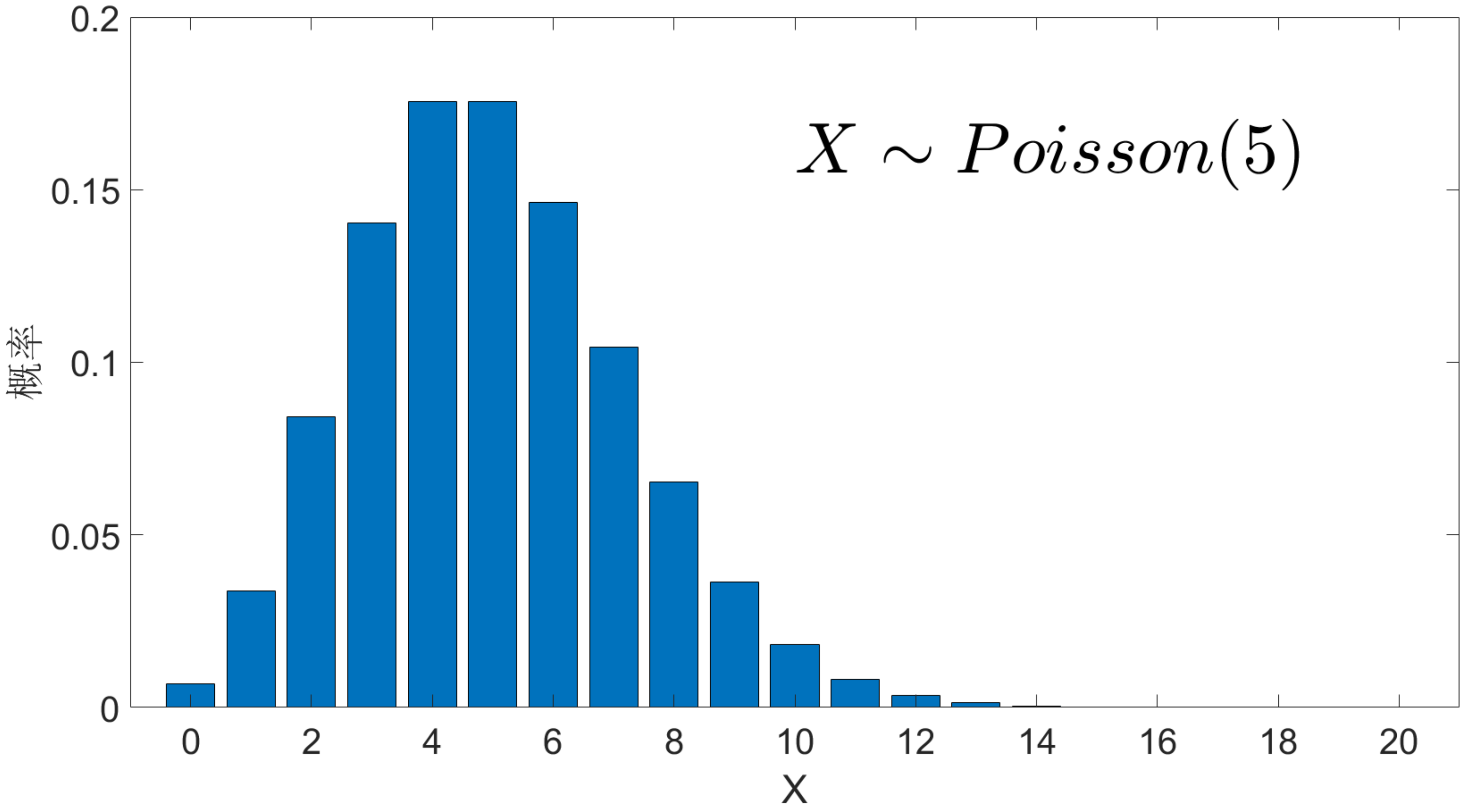
二项分布

- n重伯努利试验：由独立的n次伯努利试验构成的随机试验
- 二项分布：$Pr\{X=z\}=\binom{n}{z}p^z(1-p)^{n-z}, z=0,1,\ldots n$
- 其中n代表独立伯努利试验的次数，z代表在这些独立试验中出现第一种结果的次数，0<p<1代表每次独立试验出第一种结果的概率
- 记作X ~ Bino(n,p)


$X \sim Bino(10, 0.4)$
概率
0.3
0.25
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
X

泊松分布

- $Pr\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,2,\ldots,\lambda>0$
- 泊松分布常被用于描述一段时间内的计数数据，比如某一医院一天内的病人数
- 记作X ~ Poisson(λ)

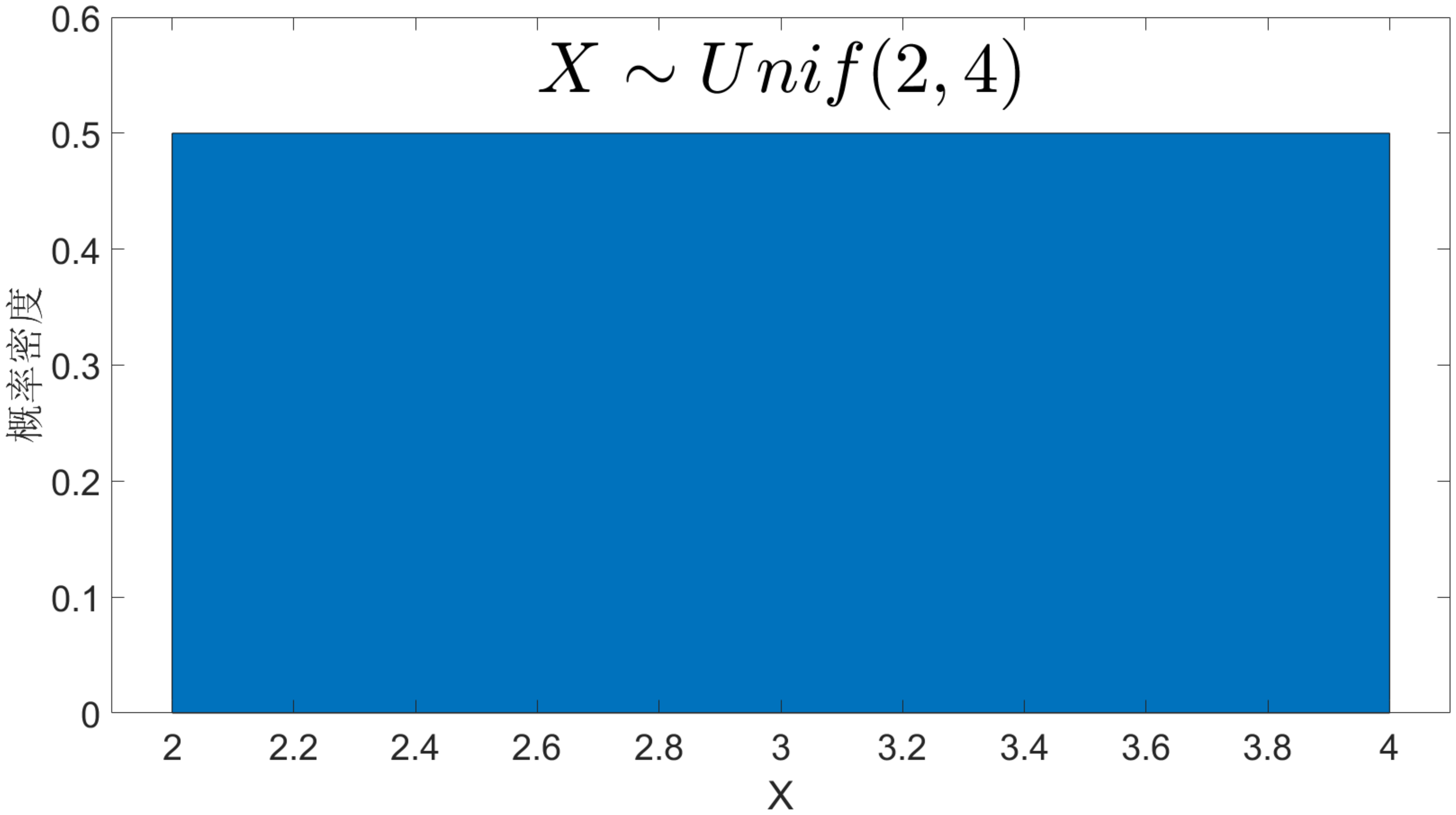

$X \sim Poisson(5)$
0.2
0.15
0.1
0.05
0
概率
0
2
4
6
8
10
12
14
16
18
20
X

常用连续概率分布

- 均匀分布
- 指数分布
- 伽马分布
- 正态分布
- 贝塔分布

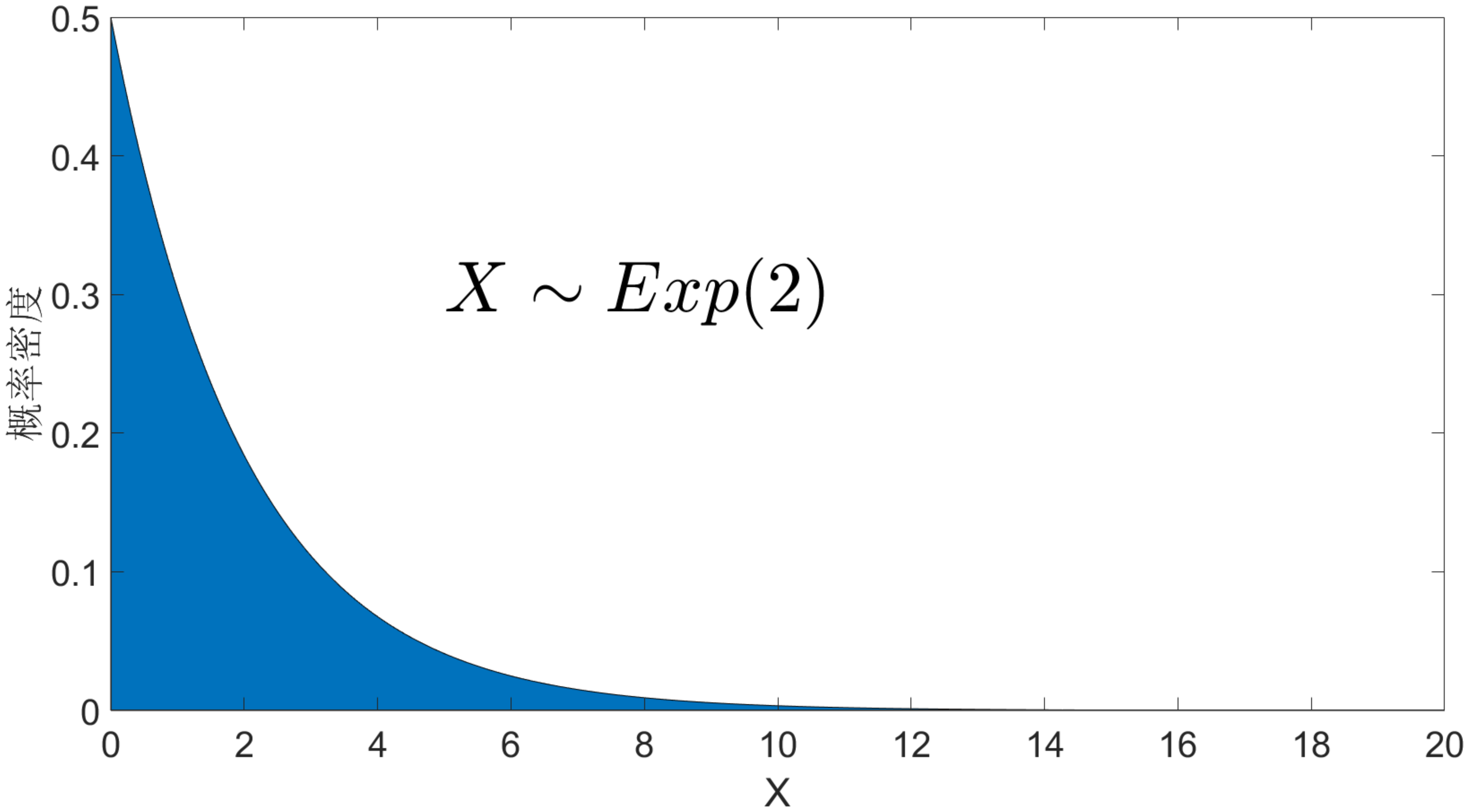
均匀分布

- f(x) = 1/(b-a), 如果 $a \leq x \leq b$，否则f(x)=0
- 记作X ~ Unif(a,b)


$X \sim Unif(2,4)$
概率密度
0.6
0.5
0.4
0.3
0.2
0.1
0
2
2.2
2.4
2.6
2.8
3
3.2
3.4
3.6
3.8
4
X

指数分布

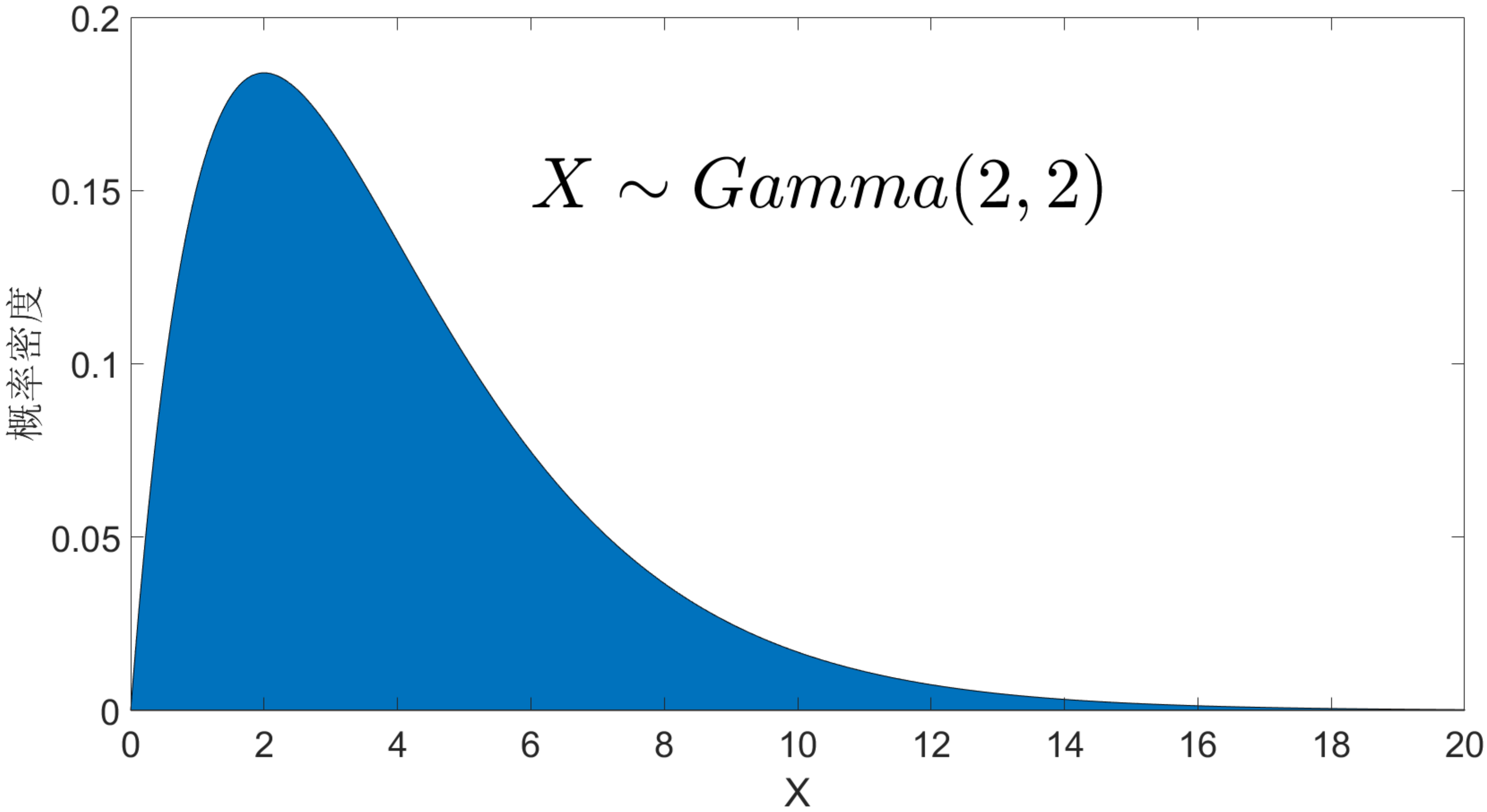
- $f(x) = \lambda e^{-\lambda x}, x \geq 0, \lambda > 0$
- 记作X ~ exp(λ)
- 指数分布常被用于描述非负连续随机变量，比如某一心理加工所需的时间



0.5
0.4
0.3
0.2
0.1
0
概率密度
$X \sim Exp(2)$
0
2
4
6
8
10
12
14
16
18
20
X

伽玛分布

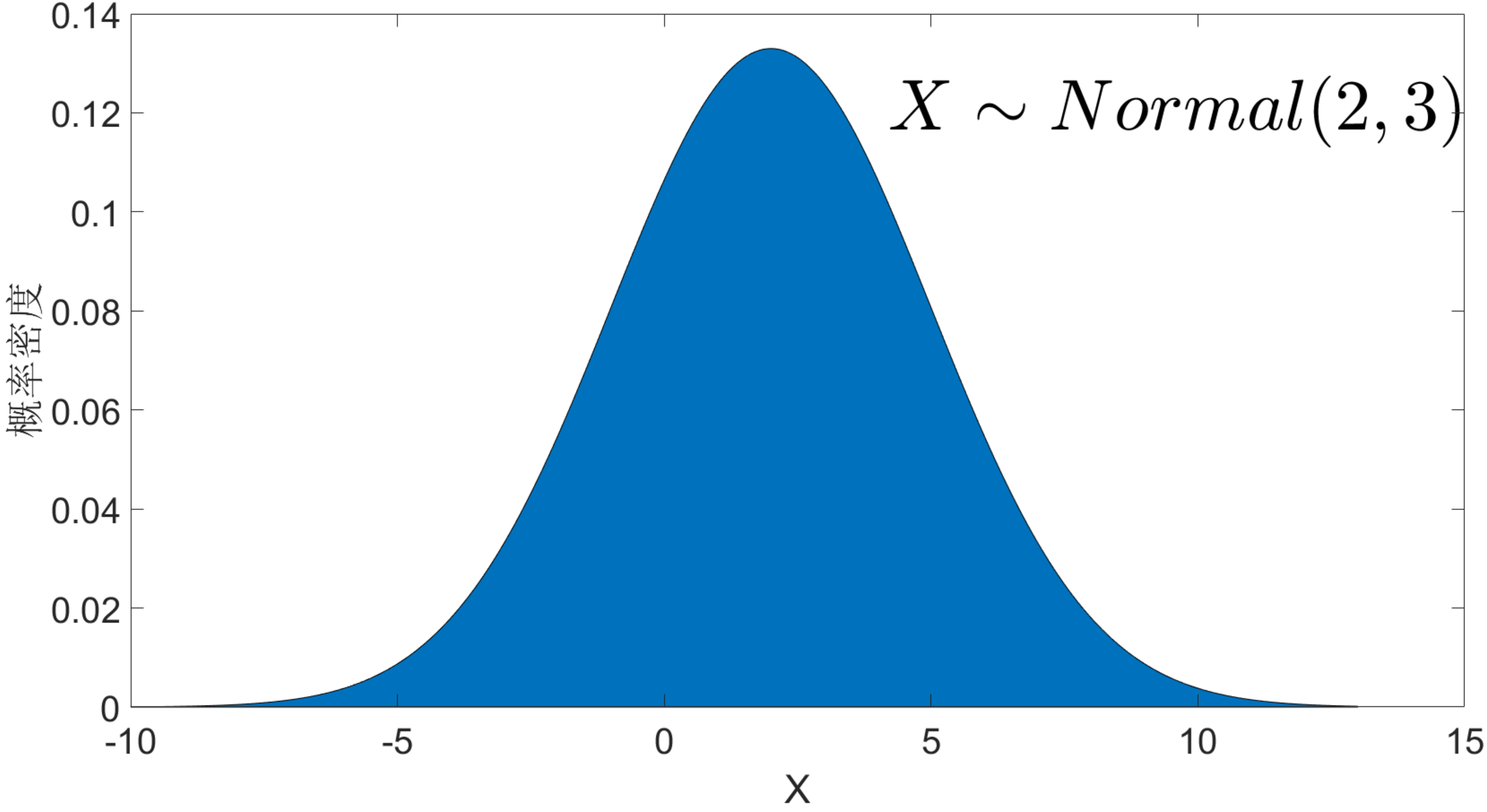
- $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}, x \geq 0, k > 0, \theta > 0$
- 记作X ~ Gamma(k, θ), k称为形状参数， θ称为尺度参数
- Γ(x)称为伽马函数，定义为$\int_0^{\infty} s^{x-1} e^{-s} ds$
- 当x为正整数时，Γ(x)=(x-1)!
- 指数分布是伽马分布当k=1时的特例，此时λ=1/θ



$X \sim Gamma(2,2)$
0.2
0.15
0.1
0.05
0
概率密度
0
2
4
6
8
10
12
14
16
18
20
X

正态分布

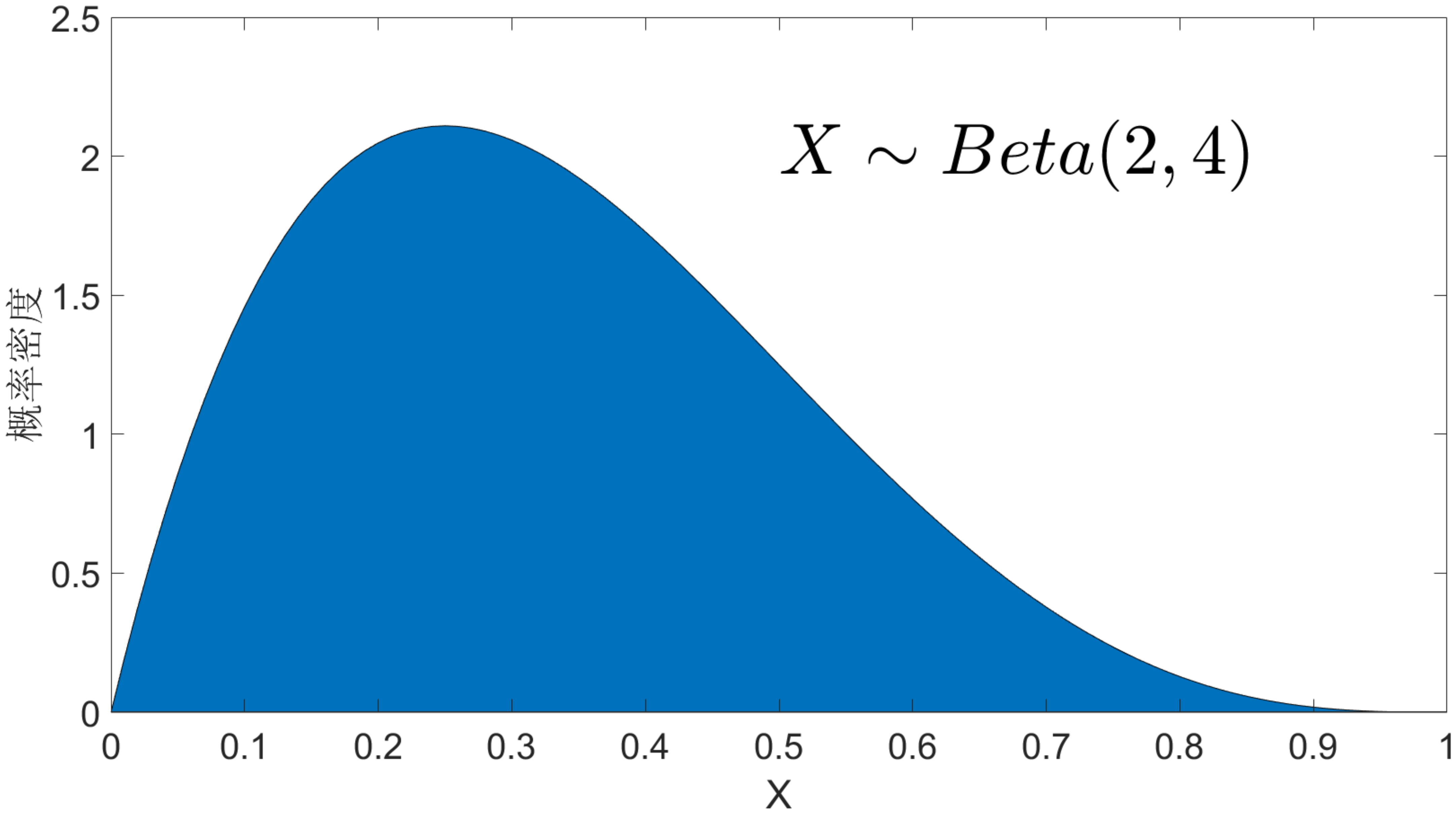
- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in R, \mu \in R, \sigma > 0$
- 记作X ~ Normal(μ, σ)



$X \sim Normal(2, 3)$
概率密度
X
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-10
-5
0
5
10
15

贝塔分布

- $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 \le x \le 1, \alpha > 0, \beta > 0$,否则$f(x) = 0$
- 记作X ~ beta(α, β)
- beta(1,1)就是[0,1]上的均匀分布



$X \sim Beta(2,4)$
概率密度
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
X

联合概率

- 某些随机试验的结果可能涉及多个离散随机变量
- 例如，对于掷色子这一随机试验，
- 定义随机变量X代表结果是否为偶数，Y代表结果是否大于3，是记为1，否记为0，那么
- $\Pr(X=0, Y=0) = 2/6$ （对应可能结果1和3）
- $\Pr(X=0, Y=1) = 1/6$ (对应可能结果5）
- $\Pr(X=1, Y=0) = 1/6$(对应可能结果2）
- $\Pr(X=1, Y=1) = 2/6$(对应可能结果4和6）

联合概率

联合概率密度

- 某些随机试验的结果涉及多个连续随机变量
- 例如，在人群中随机抽取个体，测量其身高和体重
- 定义随机变量X代表随机个体的身高，Y代表同一随机个体的体重，那么
- $f_{X,Y}(x, y)$代表随机个体的身高为x，体重为y的概率密度
- 一般而言，当至少有一个随机变量为连续变量时，$f_{X,Y}(x, y)$仍然代表概率密度。

联合分布

- $F_{X,Y}(x,y) = Pr(X <= x, Y <= y)$称为随机变量X和Y的联合分布函数，代表X的取值不大于x且Y的取值不大于y的概率
- 对于两个离散随机变量，$F_{X,Y}(x,y) = \sum_{u \le x, v \le y} \Pr(X = u, Y = v)$
- 对于两个连续随机变量，$F_{X,Y}(x,y) = \int_{-\infty}^{x} [\int_{-\infty}^{y} f(u,v)dv]du$

边缘概率

- 离散随机变量的边缘概率，是在给定该随机变量的取值，且将其他随机变量的所有可能取值都考虑在内时的概率
- 例如，在掷色子的例子中
- $\Pr(X=0)=\Pr(X=0,Y=0)+\Pr(X=0,Y=1)=2/6+1/6=1/2$
- $\Pr(X=1)=\Pr(X=1,Y=0)+\Pr(X=1,Y=1)=1/6+2/6=1/2$

X的边缘概率

- $\Pr(Y=0)=\Pr(X=0,Y=0)+\Pr(X=1,Y=0)=2/6+1/6=1/2$
- $\Pr(Y=1)=\Pr(X=0,Y=1)+\Pr(X=1,Y=1)=1/6+2/6=1/2$

Y的边缘概率

	X = 0	X = 1	
Y = 0	1/3	1/6	1/3+1/6=1/2
Y = 1	1/6	1/3	1/6+1/3=1/2
	1/3+1/6=1/2	1/6+1/3=1/2	1/2+1/2=1/2+1/2=1

边缘概率密度

- 连续随机变量的边缘概率密度，是在给定该随机变量的取值，且将其他随机变量的所有可能取值都考虑在内时的概率密度
- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$ （当Y为连续随机变量时）
- $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$ （当X为连续随机变量时）
- 通过联合概率（概率密度）计算边缘概率（概率密度）时，是用求和式还是积分式，取决于涉及所有可能取值的随机变量是离散的还是连续的

条件概率

- 离散随机变量的条件概率，是在其他随机变量取特定值这一条件下的概率
- 例如，在掷色子的例子中
- $Pr(X = 0|Y = 0) = Pr(X = 0, Y = 0)/Pr(Y = 0) = (2/6)/(1/2) = 2/3$
- $Pr(X = 1|Y = 0) = Pr(X = 1, Y = 0)/Pr(Y = 0) = (1/6)/(1/2) = 1/3$
- $Pr(X = 0|Y = 1) = Pr(X = 0, Y = 1)/Pr(Y = 1) = (1/6)/(1/2) = 1/3$
- $Pr(X = 1|Y = 1) = Pr(X = 1, Y = 1)/Pr(Y = 1) = (2/6)/(1/2) = 2/3$

X的条件概率

条件概率密度

- 连续随机变量的条件概率密度，是在其他随机变量取特定值这一条件下的概率密度
- $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = f_{X,Y}(x,y) / \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$
- 和计算边缘概率（概率密度）一样，通过联合概率（概率密度）计算条件概率（概率密度）时，是用求和式还是积分式，取决于涉及所有可能取值的随机变量是离散的还是连续的

随机变量的独立性

- 如果对于任意x和y, $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ ，则称对应的随机变量X和Y相互独立
- 对于离散随机变量，这意味着$\Pr(x) = Pr_{X|Y}(x|y)$对任何x和Pr(y)>0的y都成立，反之亦然
- 对于连续随机变量，这意味着$f_X(x) = f_{X|Y}(x|y)$对任何x和$f_Y(y) > 0$的y都成立，反之亦然

随机变量的独立性

- 当两个离散随机变量相互独立时，$Pr_{XY}(x,y) = Pr_X(x)Pr_Y(y)$
- 当两个连续随机变量相互独立时，$f_{XY}(x,y) = f_X(x)f_Y(y)$

随机变量的独立性

- 在掷色子的例子中，X的条件分布和边缘分布分别为
- $\Pr(X = 0|Y = 0) = \Pr(X = 0, Y = 0)/\Pr(Y = 0) = (2/6)/(1/2) = 2/3$
- $\Pr(X = 1|Y = 0) = \Pr(X = 1, Y = 0)/\Pr(Y = 0) = (1/6)/(1/2) = 1/3$
- $\Pr(X = 0|Y = 1) = \Pr(X = 0, Y = 1)/\Pr(Y = 1) = (1/6)/(1/2) = 1/3$
- $\Pr(X = 1|Y = 1) = \Pr(X = 1, Y = 1)/\Pr(Y = 1) = (2/6)/(1/2) = 2/3$
- $\Pr(X = 0) = 1/2$, $\Pr(X = 1) = 1/2$
- 所以，X和Y不独立：相比于边缘概率，当Y=0时，X=0的（条件）概率更高；当Y=1时，X=1的（条件）概率更高

数学期望

- 数学期望(expectation)代表随机变量的平均可能值
- 对离散随机变量X，$E(X) = \sum_k x_k \cdot \Pr(X = x_k)$
- 对连续随机变量X，$E(X) = \int_{-\infty}^{\infty} xf(x)dx$
- E(X)可以不是X的一个可能取值，比如在掷色子的例子中,E(X)=3.5

数学期望实例

- 满足泊松分布的随机变量X，其分布律为$Pr\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,2,\ldots, \lambda>0$，因此

$$E(X)=\sum_{k=0}^{\infty} k\cdot \Pr\{X=k\}=\sum_{k=0}^{\infty} k\cdot\frac{\lambda^k}{k!}e^{-\lambda}=\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}e^{-\lambda}$$

$$=\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda}=\lambda\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=\lambda$$

数学期望实例

- 满足指数分布的随机变量X，其概率密度函数为

$$\mathrm{f(x)} = \lambda e^{-\lambda x}, x \geq 0, \lambda > 0$$

- 因此，

$$E(X) = \int_0^\infty xf(x)dx = \int_0^\infty \lambda x e^{-\lambda x} dx = \frac{1}{\lambda}\int_0^\infty \lambda x e^{-\lambda x} d(\lambda x)$$

$$= \frac{1}{\lambda}\int_0^\infty u e^{-u} du = \frac{1}{\lambda}$$

中位数和众数

- 随机变量X的中位数是使得其累积概率达到50%的实数，即F(median) = .5
- 随机变量X的众数是使其概率或者概率密度取到最大值的实数，即Pr(X=mode) = max(Pr(X=x)) 或者 f(mode) = max(f(x))
- 例如，满足正态分布的随机变量的数学期望、中位数和众数都是μ

原点矩和中心矩

- 对连续随机变量X，定义M(k) = $\int_{-\infty}^{\infty} x^k f(x)dx$，称为X的k阶原点矩
- 对连续随机变量X，定义N(k) = $\int_{-\infty}^{\infty}[x-E(x)]^k f(x)dx$，称为X的k阶中心矩
- 对离散随机变量可以给出类似的定义
- 数学期望是1阶原点矩；方差是2阶中心矩
- 3阶中心矩与偏度(skewness)有关
- 4阶中心矩与峰度(kurtosis)有关

方差

- 方差(variance)是2-阶中心距，因此，对于离散随机变量

$$Var(X) = \sum_k [x_k - E(X)]^2 \cdot Pr\{X = x_k\}$$

- 对于连续随机变量

$$Var(X) = \int_{-\infty}^{\infty} [x - E(x)]^2 f(x) dx$$

- 方差的其他计算公式

$$Var(X) = E(X^2) - [E(X)]^2$$

方差实例

- 满足泊松分布的随机变量X，其分布律为$Pr\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,2,\dots,\lambda>0$，且$E(X)=\lambda$，因此

$$Var(X)=\sum_{k=0}^{\infty}(k-\lambda)^2\frac{\lambda^k}{k!}e^{-\lambda}$$

$$=\sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}e^{-\lambda}-2\lambda\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}+\lambda^2\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}$$

$$=\sum_{k=0}^{\infty}k(k-1)\frac{\lambda^k}{k!}e^{-\lambda}+\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}-2\lambda\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}+\lambda^2\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}$$

$$=\lambda^2+\lambda-2\lambda\cdot\lambda+\lambda^2=\lambda$$

方差实例

$$Var(X) = E(X^2) - [E(x)]^2$$

$$= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

方差实例

- 满足指数分布的随机变量X，其概率密度函数为

$$f(x) = \lambda e^{-\lambda x}, x \geq 0, \lambda > 0$$

且$E(X) = 1/\lambda$，因此

$$Var(X) = E(X^2) - [E(x)]^2$$

$$= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx - \frac{1}{\lambda^2} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

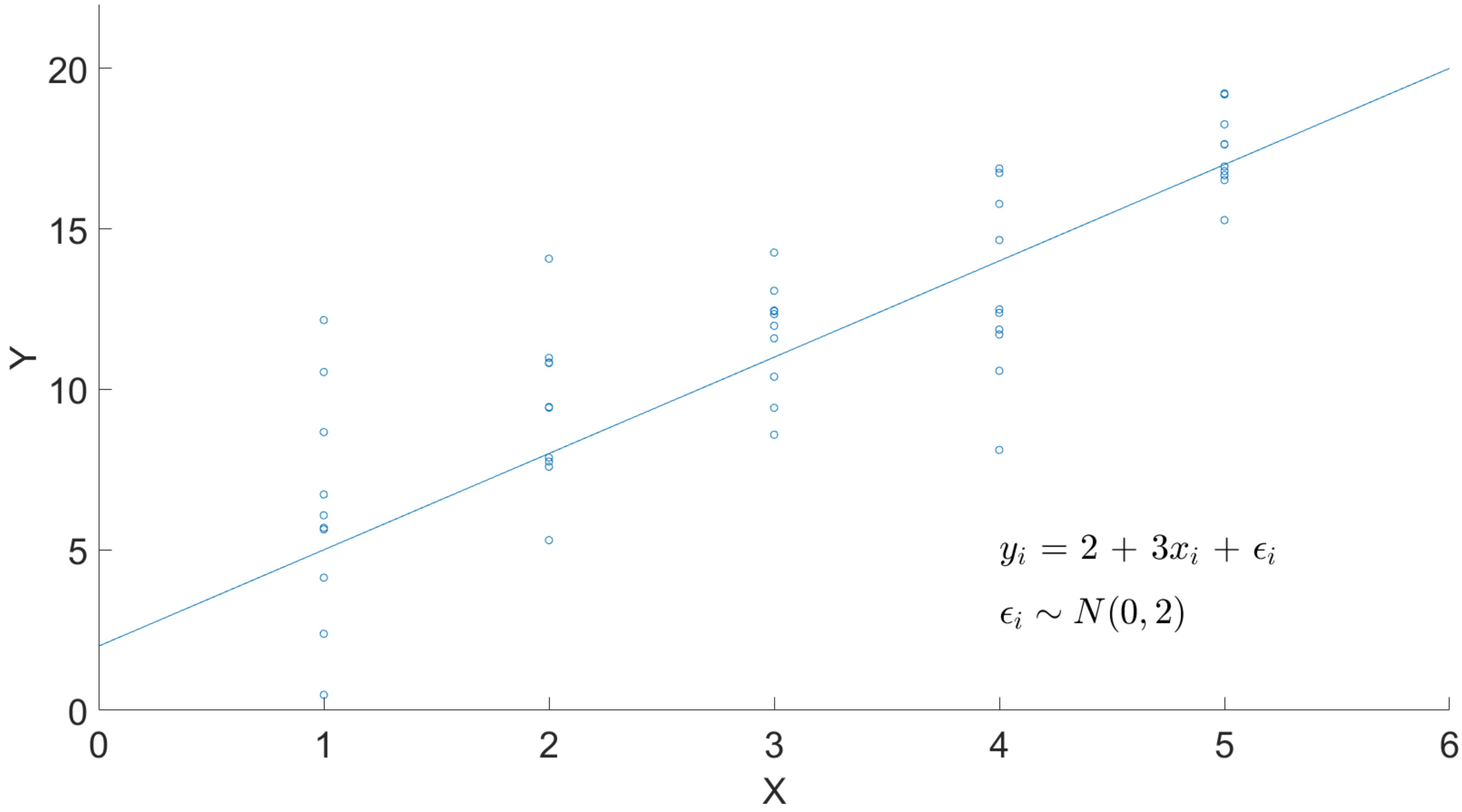
概率论相关概念的贝叶斯解释

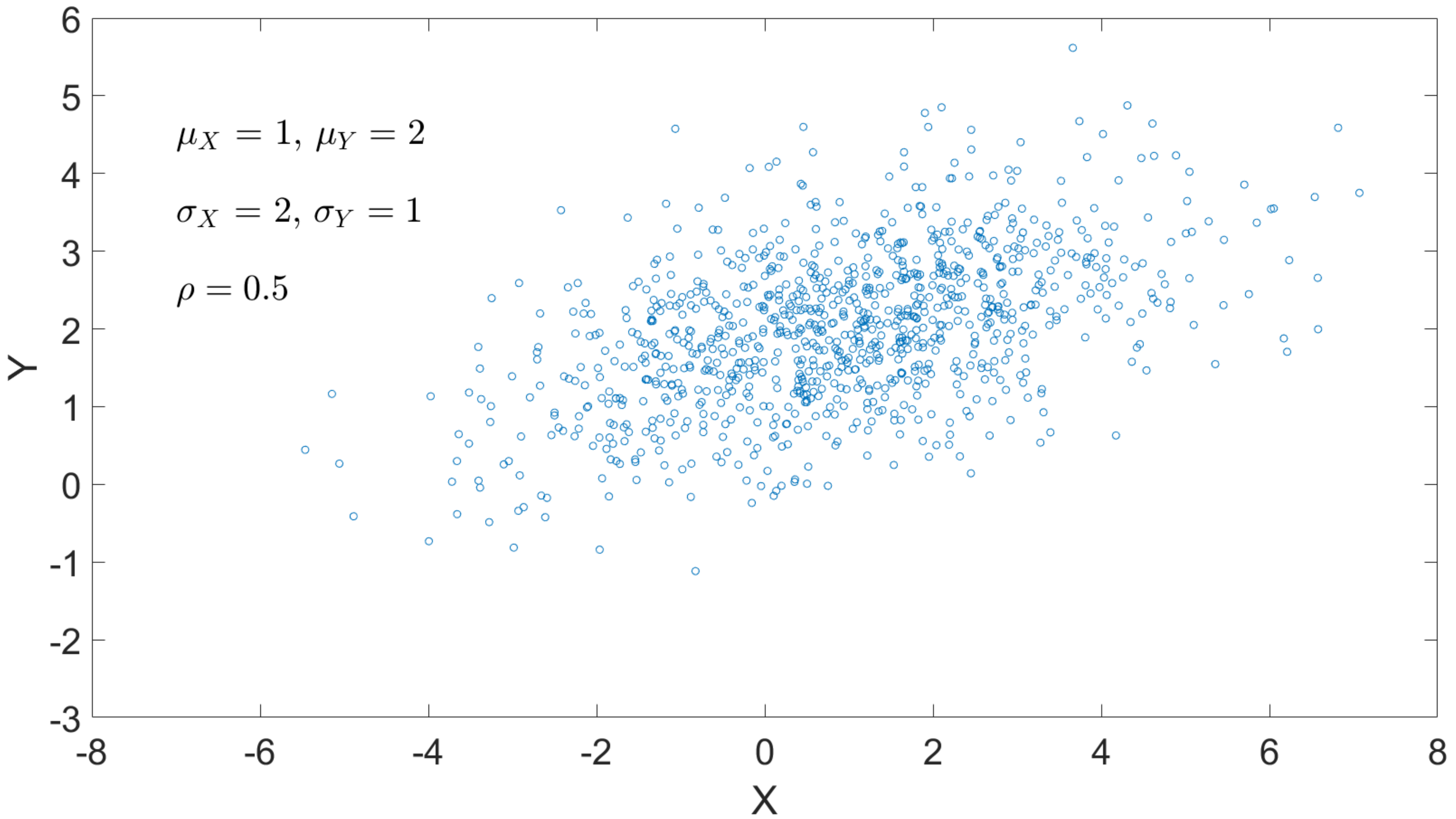
- 贝叶斯统计学沿用了经典概率论的术语体系，但它对随机事件和随机变量的看法，不依赖于无数次重复试验。
- 因此，贝叶斯统计学下的概率，可以不表示无数次重复试验条件下的相对频次，而是代表主观可能性大小。
- 同样的情形也适用于其他与概率有关的定义，比如边缘概率、条件概率、联合概率等传统概率论中的概念。

统计模型

- 任何统计分析都是建立在一定的统计模型之上的，这样的统计模型既反映了数据中稳定的具有规律性的方面，又反映了数据中随机或者不确定的方面，例如

1. 针对总体均值的统计分析通常假定总体中的数据点（即每个样本点）x_i满足
$$x_i = \mu + \varepsilon_i, \varepsilon_i \sim N(0, \sigma)\text{且各}\varepsilon_i\text{相互独立}$$
2. 一元线性回归分析通常假定每一对样本点(x_i, y_i)满足
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma)\text{且各}\varepsilon_i\text{相互独立}$$
3. 线性相关分析通常假定每一对样本点(x_i, y_i)出自一个二元正态分布，其中包括5个参数，随机变量X的数学期望μ_X和标准差σ_X，随机变量Y的数学期望μ_Y和标准差σ_Y，以及随机变量X和Y之间的相关系数ρ


20
15
10
5
0
Y
0
1
2
3
4
5
6
X
$y_i = 2 + 3x_i + \epsilon_i$
$\epsilon_i \sim N(0, 2)$


$\mu_X = 1,\ \mu_Y = 2$
$\sigma_X = 2,\ \sigma_Y = 1$
$\rho = 0.5$
Y
X
6
5
4
3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8

统计推断

- 在频率学派统计学中，大多数统计分析，都是以样本信息为基础，对于相关统计模型中的参数进行的估计和推断
- 例如，当我们从某一正态总体中抽取了样本之后，可以根据样本信息，估计总体平均数，并且对有关假设进行零假设显著性检验进而做出统计推断
- 类似的，我们可以根据样本中两个变量取值组合的情况，对总体中这两个变量的相关程度进行估计，并且对有关假设进行零假设显著性检验进而做出统计推断

极大似然估计

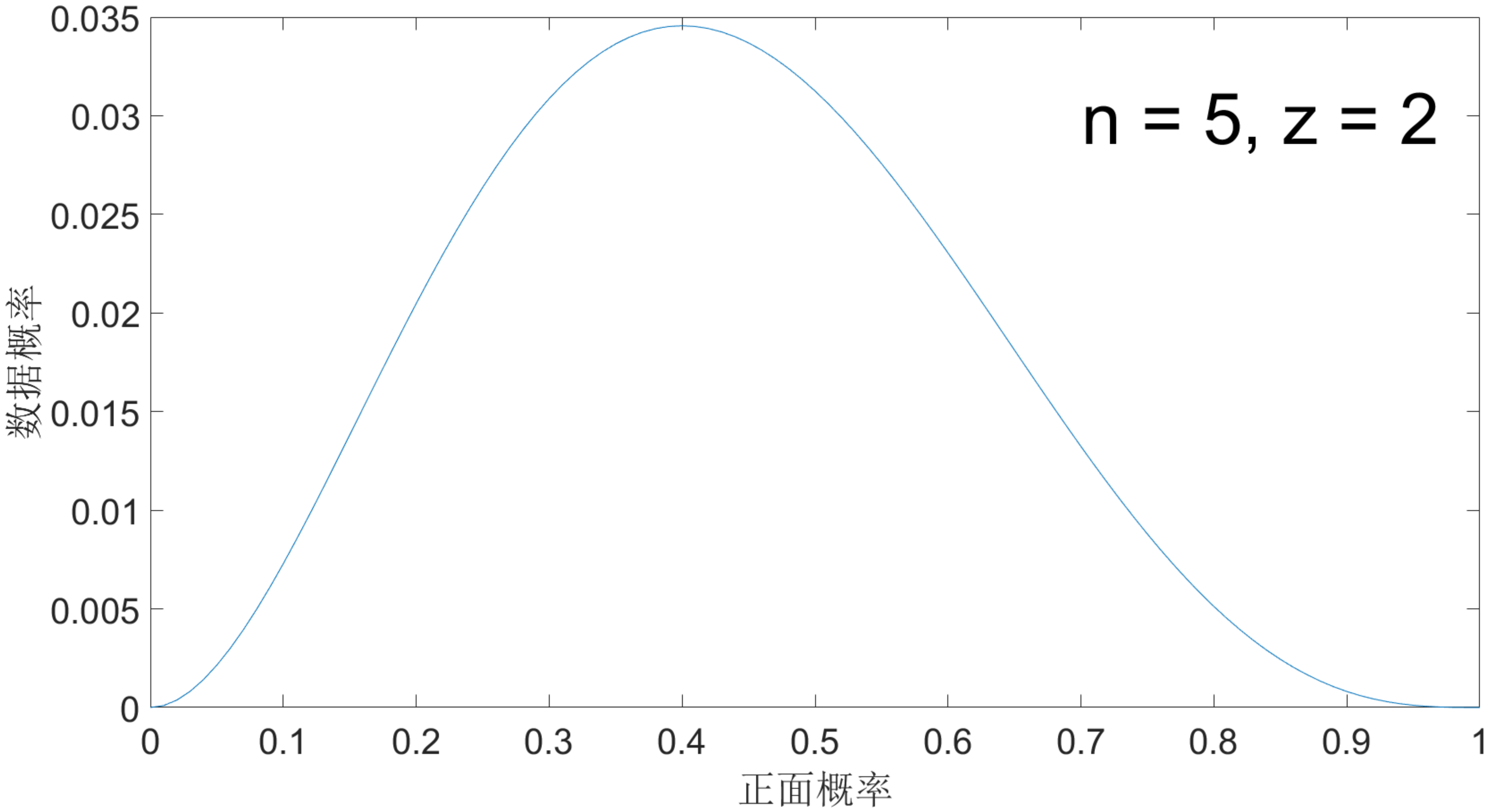
- 在频率学派统计学中，经常使用极大似然法来估计模型参数，其目标是选取合适的模型参数值，使得模型预测的实际观测数据的似然值最大化

1. 似然值：衡量得到实际观测数据的可能性的指标
2. 似然函数：以模型参数值作为输入，对应的似然值作为输出的函数
3. 极大似然估计，就是寻找使得似然函数值最大化的模型参数值

举例

- 假设我们需要估计一枚硬币每次抛掷结果为正面的概率p，再假定我们试验了5次，结果为HTTHT(H = 正面，T = 反面)。

p	似然值
0.1	$0.1^2 \times (1-0.1)^3 = 0.0073$
0.2	$0.2^2 \times (1-0.2)^3 = 0.0205$
0.3	$0.3^2 \times (1-0.3)^3 = 0.0309$
0.4	$0.4^2 \times (1-0.4)^3 = 0.0346$
0.5	$0.5^2 \times (1-0.5)^3 = 0.0312$
0.6	$0.6^2 \times (1-0.6)^3 = 0.0230$
0.7	$0.7^2 \times (1-0.7)^3 = 0.0132$



n = 5, z = 2
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
数据概率
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
正面概率

举例

- 在此情况下，数学上可以证明，极大似然参数估计值 = 样本比率。

$$L(p) = p^a(1-p)^b$$

a = 正面次数，b = 反面次数

$$\frac{dL}{dp} = ap^{a-1}(1-p)^b - bp^a(1-p)^{b-1}$$

$$= p^{a-1}(1-p)^{b-1}[a(1-p) - bp]$$

$$\frac{dL}{dp} = 0 \to a(1-p) - bp = 0 \to p = \frac{a}{a+b}$$

当 $\mathrm{p} < \mathrm{a}/(\mathrm{a}+\mathrm{b})$ 时，$\frac{\mathrm{dL}}{\mathrm{dp}} > 0$; 当 $\mathrm{p} > \mathrm{a}/(\mathrm{a}+\mathrm{b})$ 时，$\frac{\mathrm{dL}}{\mathrm{dp}} < 0$。

所以，$\mathrm{p} = \frac{\mathrm{a}}{\mathrm{a}+\mathrm{b}}$ 时L取最大值，即样本中出现正面的比率，是对每次出现正面的概率的极大似然估计。

对应的极大似然值 $= \frac{\mathrm{a}^{\mathrm{a}}\mathrm{b}^{\mathrm{b}}}{(\mathrm{a}+\mathrm{b})^{\mathrm{a}+\mathrm{b}}}$

极大似然估计的问题

- 在频率学派统计学下，大多数情况下可以得到对于参数的唯一的极大似然估计值
- 由于频率学派关注参数的唯一真值，所以该估计值就被作为参数真值的估计值
- 除极大似然估计值以外的其他值，一般不在频率学派统计学的考虑范围之内