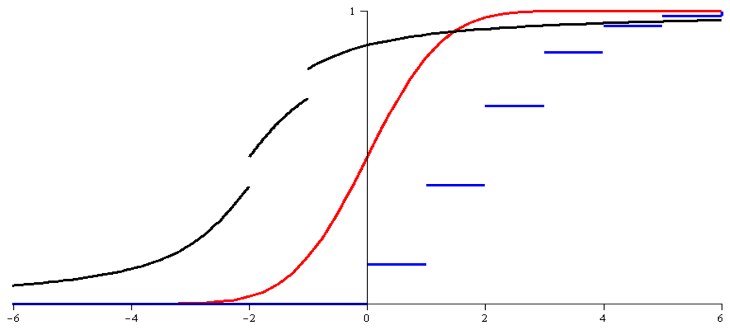
概率理论与数理统计

随机变量及其分布

分布函数



- 红色：标准正态分布的分布函数
- 黑色：在-2点和-1点的狄拉克函数与参数(-2,1)的Cauchy分布的复合。
- 蓝色：参数2的Poisson分布

分布函数

如果随机变量X的取值空间是$\mathbb{R}$或其子集，集合$\{X \le x\} \in \mathcal{F}$ 可测，定义函数

$$F_X(x) = P(X \le x)$$

即$F(x)$给出X小于或等于给定数值x的概率, 称为X的分布函数, 满足

1. 单调性: F_X是单调非降函数，$F(x_1) \le F(x_2)$, $x_1 \le x_2$,
2. 有界性: $0 \le F(x) \le 1$, $\lim_{x\to-\infty} F(x) = 0$; $\lim_{x\to\infty} F(x) = 1$.
3. 右连续性: $\forall x_0 \in \mathbb{R}$, $\lim_{x\to x_0+} F_X(x) = F(x_0)$,

定理

- 由$(\mathbb{R}, \mathcal{B}(\mathbb{R})$上的概率$\mathbf{P}$可以诱导出一个分布函数
$$F(x) = \mathbf{P}((-\infty, x])$$
- 由分布函数F可以确定一个在$(\mathbb{R}, \mathcal{B}(\mathbb{R})$ 上的概率$\mathbf{P}$.
- 若函数F满足定义中的1)-3), 则F为一个在$(\mathbb{R}, \mathcal{B}(\mathbb{R})$上(唯一)确定概率的分布函数.

连续型随机变量-密度函数

- 连续型随机变量: 存在一个$\mathbb{R}$上的非负Lebesgue可测函数$p_X(x)$满足, 对$x \in \mathbb{R}$

  $$F(x) = \int_{-\infty}^{x} p_X(u)du.$$

  称$p_X(x)$为X的概率密度函数.
- 在一个Lebesgue零测集之外, $p_X(x) = \frac{d}{dx}F(x)$.
- $\forall a < b \in \mathbb{R}$, $P(a < X < b) = \int_a^b p_X(x)dx$.

定理

如果$p_X(\cdot)$为非负Lebesgue可积函数，在R上的全积分为1，

$$p_X(x) \geq 0, \int_{-\infty}^{+\infty} p_X(x)dx = 1,$$

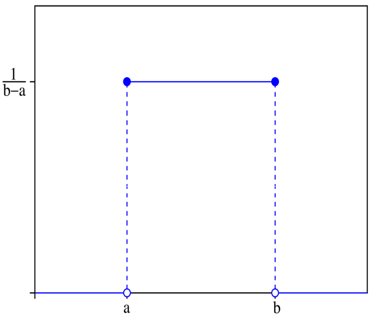
那么p_X为一个概率密度函数.

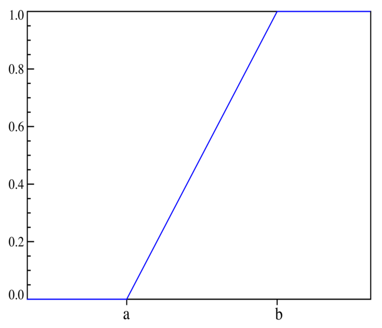
均匀分布

均匀分布$\xi \sim U(a,b)$

- $a, b \in R$, $a < b$, 在$[a,b]$上的几何分布.
- 其密度函数为

$$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases}$$

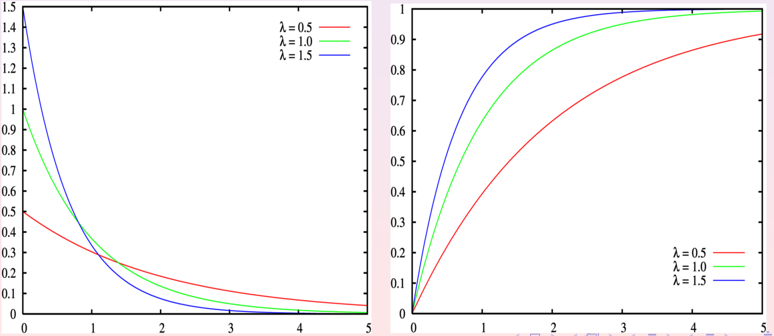




指数分布

指数分布$\xi \sim Exp(\lambda)$

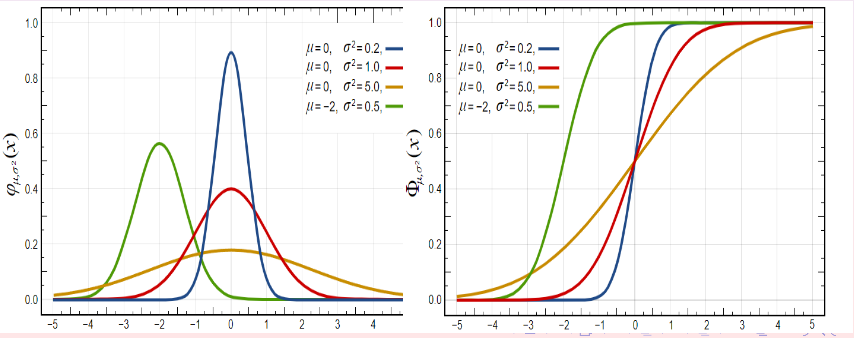
- 密度函数: $p(x) = \lambda e^{-\lambda x}$, $x > 0$, $p(x) = 0$, $x \leq 0$.
- 分布函数: $F(x) = 1 - e^{-\lambda x}$, $x > 0$, $F(x) = 0$, $x \leq 0$.
- 无记忆性: $P(X > s + t | X > s) = P(X > t)$, $s, t > 0$.



正态分布

正态分布$N(\mu,\sigma)$

- 标准正态分布$N(0,1)$: $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $x\in\mathbb{R}$.
- 密度函数: $p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x\in\mathbb{R}$. $\int_{\mathbb{R}}p(x)dx=1$.
- 当$x=\mu$时取最大值, $f(\mu)=\frac{1}{\sqrt{2\pi}\sigma}$. σ越小时图形越尖.



- 统计学家William Youden 在1962年写了1篇解释科学中正态分布的目的和地位的文章。他在书法中把它呈现为一条钟形曲线。

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY ♦ IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING ♦
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

"在人类的经验中，误差的正态分布作为自然哲学最重要的概括之一脱颖而出。它是物理和社会科学以及医学、农业和工程研究的指导性工具。它是分析和解释通过观察和实验获得的基本数据的不可缺少的工具。"

正态分布

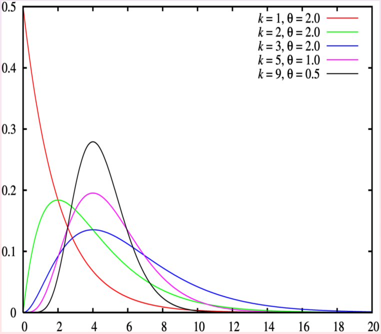
1. 曲线关于$x=\mu$对称, 对于任意$h>0$有
$$P(\mu-h<X\leq\mu)=P(\mu<X\leq\mu+h).$$
2. 标准正态分布: $\phi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, $\Phi(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{x^2}{2}}$,
$$\Phi(0)=\frac{1}{2},\quad \Phi(-x)=1-\Phi(x)$$
$\Phi(x)$为严格增函数, 则其逆存在.
3. 若随机变量$X\sim N(\mu,\sigma)$, 则$Z=\frac{X-\mu}{\sigma}\sim N(0,1)$,
$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt=P(\frac{X-\mu}{\sigma}\leq\frac{x-\mu}{\sigma})=\Phi(\frac{x-\mu}{\sigma})$$
$$P(x_1<X\leq x_2)=\Phi(\frac{x_2-\mu}{\sigma})-\Phi(\frac{x_1-\mu}{\sigma})$$
$$P(\mu-\sigma<X<\mu+\sigma)=\Phi(1)-\Phi(-1)=68.26\%$$
$$P(\mu-2\sigma<X<\mu+2\sigma)=\Phi(2)-\Phi(-2)=95.44\%$$
$$P(\mu-3\sigma<X<\mu+3\sigma)=\Phi(3)-\Phi(-3)=99.74\%$$

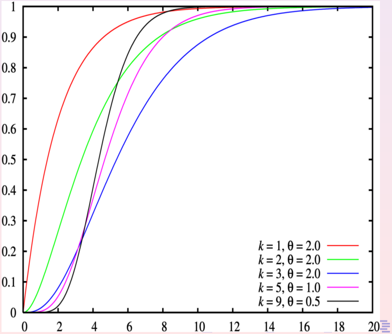
Γ分布

Γ分布$\xi \sim \Gamma(r, \alpha)$

$x > 0, p(x) = \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x}, \quad x \le 0, p(x) = 0.$

- $r = 1$, ξ 服从参数为λ的指数分布，
- $r = \frac{k}{2}$, $\alpha = 2$, ξ 服从参数为k的χ^2分布.
- r为整数时，也称为Erlang分布(同时拨入的电话数目建模的分布).





随机变量函数的分布

定理.

设随机变量X有概率密度函数$p_X(x)$, $x \in \mathbb{R}$, 设函数$g(x)$处处可导且恒有$g'(x) > 0$(或者$g'(x) < 0$), 则$Y = g(X)$为连续型随机变量, 其概率密度为,

$$p_Y(y) = \begin{cases} p_X(h(y))|h'(y)|, & \alpha < y < \beta, \\ 0, & \text{其他}, \end{cases}$$

其中$\alpha = \min\{g(-\infty), g(\infty)\}$, $\beta = \max\{g(-\infty), g(\infty)\}$, $h(x)$是$g(x)$的反函数.

- $X \sim N(0,1)$, $Y = X^2$,
- $X \sim N(\mu, \sigma)$, $Y = aX + b$,
- $X \sim N(\mu, \sigma)$, $Y = e^X$.

概率理论与数理统计

随机向量及其分布

随机向量

随机向量

- 若X_1, X_2, $\cdots$, X_d为定义在同一概率空间中的随机变量，则$X = (X_1, X_2, \cdots, X_d)$ 称为d-维随机向量.

随机向量的分布函数

d-维随机向量$X = (X_1, X_2, \cdots, X_d)$ 的联合分布函数为d元函数，对于实数x_1, x_2, $\cdots$, x_d, 定义

$$F(x_1, x_2, \cdots, x_d) = P(X_1 \le x_1, X_2 \le x_2, \cdots, X_d \le x_d).$$

- 学生的身高X_1, 体重X_2,
- 粒子的位置，x坐标X_1, y坐标X_2, z坐标X_3,
- 市场中d只股票的价格: S_1, S_2, $\cdots$, S_d.

$d=2$时, 随机向量的分布函数

设X与Y为随机变量，(X,Y)取值于$\mathbb{R}^2$, 其联合累积分布函数为$F_{X,Y}(x,y)$, $x,y\in\mathbb{R}$

$$F_{X,Y}(x,y)=P((X\le x)\cap(Y\le y))=P(X\le x,Y\le y).$$

对于任意$x_1<x_2$, $y_1<y_2$ 有

$$\begin{aligned} P(x_1< & X\le x_2,y_1<Y\le y_2)\\ = & F(x_2,y_2)-F(x_1,y_2)-F(x_2,y_1)+F(x_1,y_1)\ge 0. \end{aligned}$$

1. 单调性: $F(x,y)$分别关于x, y单调非减,
   - 若$x_1<x_2$, 则$F(x_1,y)\le F(x_2,y)$,
   - 若$y_1<y_2$, 则$F(x,y_1)\le F(x,y_2)$.
2. 有界性: $0\le F(x,y)\le 1$,
   $$F(-\infty,y)=0,F(x,-\infty)=0,F(\infty,\infty)=1.$$
3. 右连续: 关于每个变量x, y右连续:
   $$F(x+0,y)=F(x,y),F(x,y+0)=F(x,y).$$
4. 非负性: $F(x_2,y_2)-F(x_1,y_2)-F(x_2,y_1)+F(x_1,y_1)\ge 0.$

离散随机向量的分布

离散随机向量与分布律

- 若X, Y取值于有限或可数集合，则(X, Y) 称为2-维离散随机向量.
- 设(x_i, y_j), $i, j = 1, 2 \cdots$, 为(X, Y)的所有可能取值，设

$$P(X = x_i, Y = y_j) = p_{ij},$$

满足

$$p_{ij} \geq 0, \sum_i \sum_j p_{ij} = 1,$$

$\{p_{ij}\}$称为2-维离散随机向量(X, Y)的联合分布律.

- 对于$x, y \in R$,

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{ij}.$$

连续随机向量与分布函数

- 若存在非负可积$p(x,y)$满足: $\forall x, y \in R$
$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} p(u,v)dudv.$$
则称(X,Y) 为二维连续型随机变量, 函数$p(x,y)$称为二维连续型随机变量(X,Y)的联合密度函数.
- 联合密度函数$p(x,y)$满足
  - $p(x,y) \geq 0$,
  - $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x,y)dxdy = F(\infty,\infty) = 1.$
- 由连续型r.v.(X,Y)定义的联合分布为, $\forall G \in \mathcal{B}(\mathbb{R}^2)$
$$\begin{aligned} \mu_{X,Y}(G) &= P((X,Y) \in G), \\ &= \int\int_G p(x,y)dxdy = \int_R\int_R 1_G(x,y)p(x,y)dxdy \end{aligned}$$
- 在p的连续点(x,y): $\frac{\partial^2 F(x,y)}{\partial x \partial y} = p(x,y)$,
$$P(x < X \leq x+\Delta x, y < Y \leq y+\Delta y) \approx p(x,y)\Delta x \Delta y.$$