应用统计

第2讲　　常见分布与经验分布

总体和样本

- 总体：　一个统计问题研究对象的全体。构成总体的每个成员称为个体。简单说总体即为分布。

- 样本：从总体中随机抽样的部分个体组成的集合称为样本，样本中的个体称为样品，样品的个数称为样本容量或样本量。

  简单随机抽样：（1）样本具有随机性　（2）　样本之间相互独立。

- 例如：　随机抛掷一枚色子，总体是1，2，3，4，5，6的均匀分布。随机地、相互独立地投掷10次，得到

5　6　1　6　4　1　2　4　6　6

即得到10个样品，样本容量为10。

随机变量

定义在样本空间Ω上的函数，就称为随机变量。

$X:\Omega \to R$，常用大写字母X,Y,Z等表示随机变量，

其取值用小写字母x,y,z等表示，$x = X(\omega)$，$\omega \in \Omega$；

$X:\Omega \to R$，随机变量$X(\omega)$一般简记为X。

随机变量的分布函数

设 X 是一个随机变量，对任意实数 x ，定义

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

为随机变量 X 的分布函数，且称 X 服从 $F(x)$，

记为 $X\sim F(x)$，有时也记作 $F_X(x)$。

离散型随机变量

如果随机变量 X 所有可能的取值是有限或可列多个，则其分布可表示为

$$\begin{array}{c|ccccc} X & x_1 & x_2 & \cdots & x_n & \cdots \\ \hline P & p(x_1) & p(x_2) & \cdots & p(x_n) & \cdots \end{array} \quad \text{或} \quad \begin{pmatrix} x_1 & x_2 & \cdots & x_n & \cdots \\ p(x_1) & p(x_2) & \cdots & p(x_n) & \cdots \end{pmatrix}$$

这种表示称为分布列。其中 $p(x_i)=P(X=x_i)\geq 0 \quad (i=1,2,\cdots n,\cdots)$，$\sum_{i=1}^{\infty} p(x_i)=1$，

$F(x)=\sum_{x_i\leq x} p(x_i)$。分布函数图形为阶梯函数。

二项分布，泊松分布，几何分布

连续型随机变量

设随机变量 X 的分布函数为 $F(x)$，如果存在非负可积函数 $p(x)$ $(x \in R)$，使得 $\forall x \in R$，有 $F(x)=\int_{-\infty}^{x} p(t)dt$，则称 X 为连续型随机变量，$p(x)$称为 X 的概率密度函数，简称密度函数（probability density function，常缩写为 pdf）。

$\int_{-\infty}^{+\infty} p(x)dx = 1$。

均匀分布，指数分布，正态分布

卡方分布，t分布，F分布

几种常见的离散型分布

- **二项分布，几何分布，泊松分布**

伯努利(Bernoulli)试验:

一随机试验有两个基本结果，记为事件 A 和 $\overline{A}$，

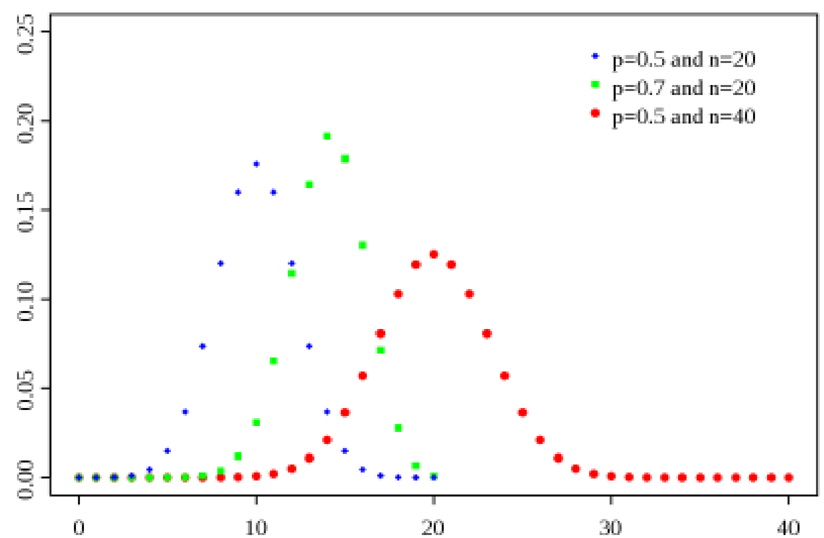
$P(A)=p$，$P(\overline{A})=q$，$p+q=1$。

二项分布 $X \sim b(n,p)$

将伯努利试验独立地重复 n 次，比如连续投掷 n 次硬币、连续 n 次射击等，基本结果（过程）有 2^n 种， X 为 A 出现的次数，则 X 的分布为:

$$\begin{pmatrix} 0 & 1 & 2 & \cdots & k & \cdots & n \\ p_0 & p_1 & p_2 & \cdots & p_k & \cdots & p_n \end{pmatrix},\quad 其中\ \ p_k = \binom{n}{k} p^k q^{n-k},\quad k = 0,1,,n。$$

此分布即称为二项分布，记为 $X \sim b(n,p)$。 p_k 为 $(p+q)^n$ 的二项展开系数。



二项分布的实例

- 甲、乙两棋手约定进行10局比赛，每局棋甲获胜的概率是0.6，乙获胜的概率为0.4。如果各局比赛独立进行，试问甲获胜、战平和失败的概率？

X 表示甲获胜的局数，则 $X \sim b(10,0.6)$

$$P(\text{甲胜}) = P(X > 5) = \sum_{k=6}^{10} \binom{10}{k} 0.6^k 0.4^{10-k} = 0.6330$$

$$P(\text{乙胜}) = P(X < 5) = \sum_{k=0}^{4} \binom{10}{k} 0.6^k 0.4^{10-k} = 0.1663$$

$$P(\text{战平}) = P(X = 5) = \binom{10}{5} 0.6^5 0.4^5 = 0.2007$$

泊松（Poisson）分布： $X \sim P(\lambda)$

描述稀有事件

$$\begin{pmatrix} 0 & 1 & 2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \end{pmatrix}, \quad \text{其中} \quad p_k = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0,1,2,\cdots。$$

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

泊松分布与二项分布的关系

考虑二项分布 $b(n,p)$，当 p 很小 n 很大时， $b(n,p)$ 与 $P(np)$ 非常接近，可相互近似

泊松分布实例一：伦敦飞弹

伦敦飞弹。二战时伦敦遭到很多次炸弹袭击，将整个面积分为 $N=567$ 小块，中 k 枚飞弹的块数为 N_k，共投下 537 枚，$\lambda=\frac{537}{567}\approx 0.9323$。

每一小块遭受到的炸弹数：$X\sim b\left(537,\frac{1}{567}\right)\approx P\left(\frac{537}{567}\right)$

k	0	1	2	3	4	≥ 5
N_k	229	211	93	35	7	1
$N\cdot p(k,0.9323)$	226.7	211.4	98.6	31.6	7.1	1.6

泊松分布实例二：上海暴雨

上海市在 1875—1955 年中间有 63 年的夏季（5 月—9 月）暴雨记录，共计 180 次。每年夏季共有 $n = 31 + 30 + 31 + 31 + 30 = 153$ 天，每次暴雨如果以一天计算，则每天发生暴雨的概率为 $p = \dfrac{180}{63 \times 153} = 0.0187$，其值很小，同时 $n = 153$ 较大。如果暴雨可看成服从二项分布的稀有事件，则一个夏季发生暴雨的次数用泊松分布近似是合理的。$\lambda = np = \dfrac{180}{63} \approx 2.9$

暴雨次数	0	1	2	3	4	5	6	7	≥ 8
实际年份数	4	8	14	19	10	4	2	1	1
理论年数 $(63 \times p_k(2.9))$	3.5	10.1	14.6	14.1	10.2	6.0	2.9	1.2	0.6

几何分布

考虑伯努利试验 $P(A) = p$。

第一次成功时的试验次数 $X \sim Ge(p)$ 几何分布

$$P(X = k) = (1 - p)^{k-1} \cdot p, \quad k = 1,2,\cdots$$

一报贩发现每个路过他的报摊的行人向他买报的概率为1/5，第一个买报纸的人是经过的第几个人？平均见到几个人能够卖出一张报纸？

几何分布

几何分布与指数分布的无记忆性:

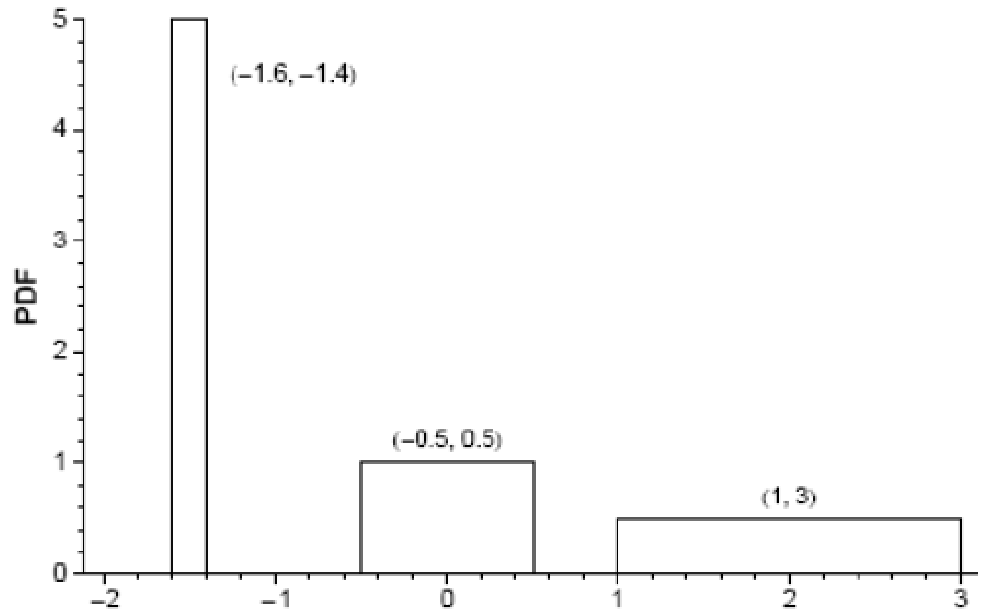
对任意 $s>0$ 和 $t>0$，有 $P(Y>s+t \mid Y>s)=P(Y>t)$。

均匀分布

均匀分布 $-\infty < a < b < +\infty$

$$p(x)=\begin{cases}\dfrac{1}{b-a} & X\in[a,b]\\ 0 & \text{其他}\end{cases},$$

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

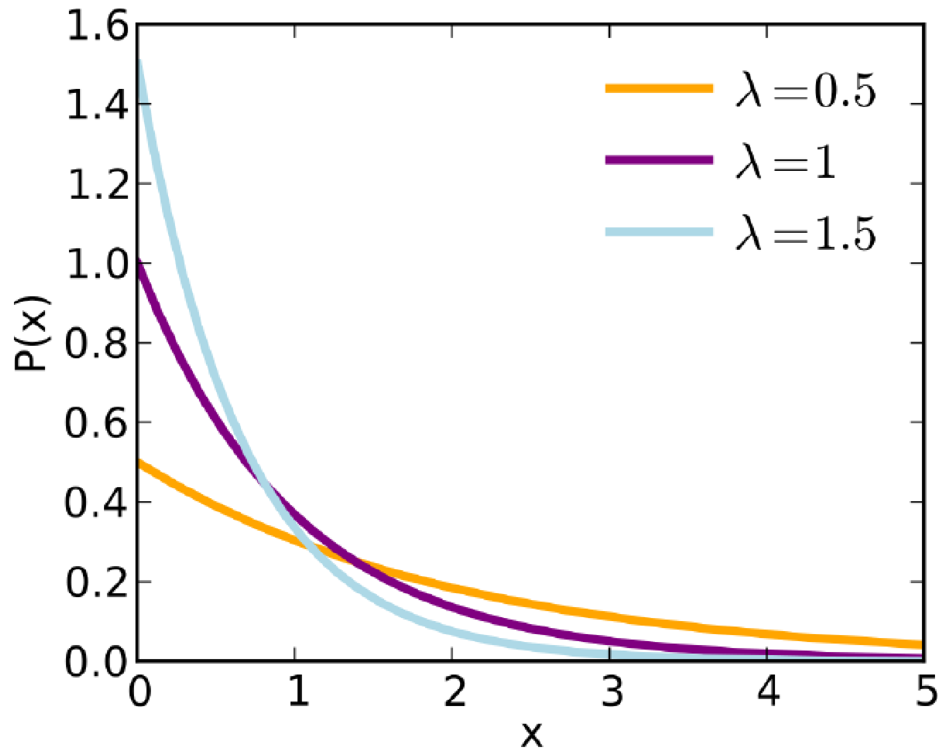


$$X\sim U(a,b),\quad E(X)=\frac{a+b}{2},\quad Var(X)=\frac{(b-a)^2}{12}$$

指数分布 $X \sim Exp(\lambda),\ \lambda > 0$

$$p(x) = \begin{cases} 0 & ,x \le 0 \\ \lambda e^{-\lambda x} & ,x > 0 \end{cases}$$

$$F(x) = \begin{cases} 0 & x \le 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$$



指数分布的无记忆性：$\forall\ s > 0$，$t > 0$，$P(Y > s + t \mid Y > s) = P(Y > t)$。

指数分布的实例：电子元件的寿命

假设一种电子元件的寿命 X 随机变量，对已使用了 t 小时的元件，在以后 Δt 小时内失效的概率为 $\lambda\Delta t+o(\Delta t)$，其中 λ 为不依赖 t 的常数，称为失效率，求该元件寿命的分布函数。

由题设有 $P(X\le t+\Delta t\mid X>t)=\lambda\Delta t+o(\Delta t)$， 记 $f(t)=P(X>t)$

$$
\begin{aligned}
f(t+\Delta t)=P(X>t+\Delta t)&=P(X>t+\Delta t,X>t)\\
&=P(X>t)P(X>t+\Delta t\mid X>t)\\
&=P(X>t)(1-P(X\le t+\Delta t\mid X>t))
\end{aligned}
$$

$$f(t+\Delta t)=P(X>t+\Delta t)=P(X>t)(1-P(X\le t+\Delta t \mid X>t))$$

$$\Rightarrow f(t+\Delta t)=f(t)[1-\lambda\Delta t+o(\Delta t)]$$

$$\Rightarrow \frac{f(t+\Delta t)-f(t)}{\Delta t}=-\lambda f(t)+o(1)$$

$$\Rightarrow \frac{df(t)}{dt}=-\lambda f(t)$$

考虑 $f(0)=1$，有 $f(t)=e^{-\lambda t}$，

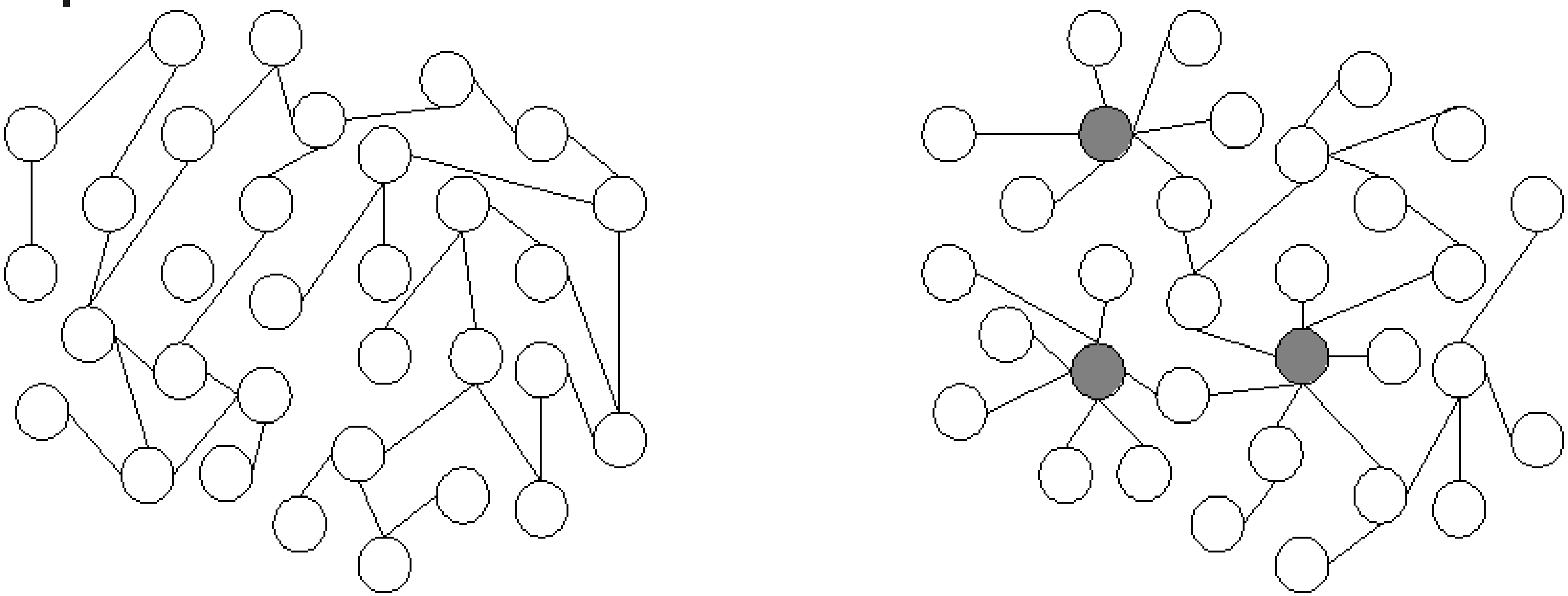
$F(x)=P(X\le x)=1-P(X\ge x)=1-e^{-\lambda x}$。

幂律，80/20法则

- 帕雷托法则（Pareto principle），也称为二八定律或80/20法则
- 此法则指在众多现象中，80%的结果取决于20%的原因，如
- 20%的人做了80%的工作；最初是意大利经济学家帕雷托在1906年对意大利20%的人口拥有80%的财产的观察而得出

- WWW上80%的链接指向15%网页；
- 80%的学术引用出自38%的科学家；
- 好莱坞80%的链接指向30%的演员；
- Microsoft，20%的计算机病毒，造成80%的危害
- ……

幂律：$p(x) = Cx^{-\alpha} \quad x \geq x_{min}$



(a) Random network　　　　(b) Scale-free network

复杂网络

DJ Watts，SH Strogatz, Collective dynamics of 'small-world' networks, Nature, 1998

AL Barabasi，R Albert, Emergence of Scaling in Random Networks Science,1999

R.Albert，AL.Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics, 2001, 74(1)

幂律：$p(x) = Cx^{-\alpha} \quad x \geq x_{min}$

最常用的分布 ： 正态分布

标准正态分布 $X \sim N(0,1)$

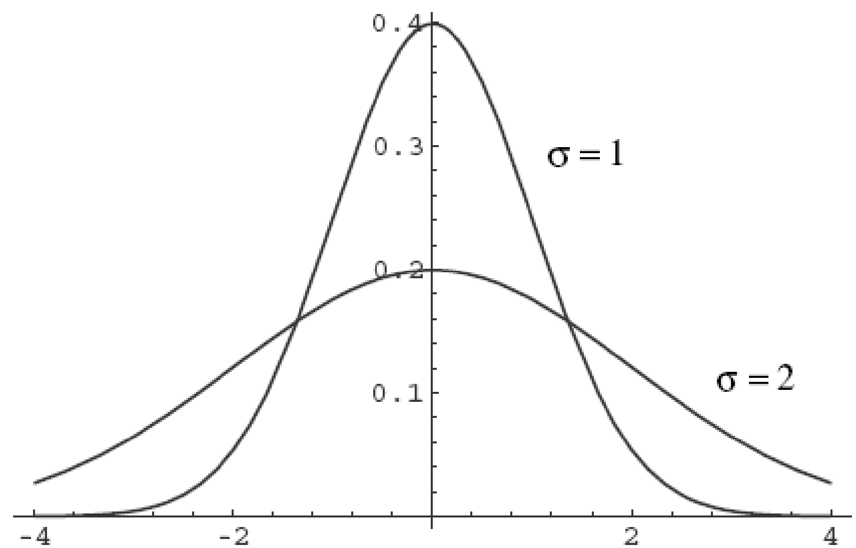
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad x \in R$$

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$，此值可查表得到

一般正态分布 $X \sim N(\mu, \sigma^2)$,

$$p(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad x \in R$$

$$\frac{X-\mu}{\sigma} \sim N(0,1)$$



σ 表示分散程度，
σ 越大则数据分布越散开，
越小则数据分布越集中.

考虑：σ 趋于正无穷或零时?

$$X_1+X_2+\cdots+X_n \overset{\cdot}{\sim} N\left(\sum_{k=1}^{n}E(X_k),\sum_{k=1}^{n}Var(X_k)\right)$$

中心极限定理

中心极限定理（林德伯格-勒维Lindeberg-Levy） 设$\{X_n\}$是独立同分布的随机变量序列。如果其期望$E(X_1)=\mu$，方差$Var(X_1)=\sigma^2$，则对每一个固定的y有

$$\lim_{n\to\infty}\Pr\left\{\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\cdot\sqrt{n}}\le y\right\}=\Phi(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}e^{-\frac{t^2}{2}}dt\text{。}$$

李雅普诺夫（Liapunov）中心极限定理： 设$\{X_n\}$为独立随机变量序列，

若存在$\delta>0$，满足 $\lim\limits_{n\to+\infty}\frac{1}{\sqrt{Var(X_1+\cdots+X_n)}^{2+\delta}}\sum_{k=1}^{n}E\left(\left|X_k-E(X_k)\right|^{2+\delta}\right)=0$，

则 $$\frac{\sum_{k=1}^{n}(X_k-E(X_k))}{\sqrt{\sum_{k=1}^{n}Var(X_k)}}\xrightarrow[n\to\infty]{}N(0,1)$$

中心极限定理

例. 设系统由100个相互独立的部件组成，运行时间每个部件损坏的概率为0.1， 至少有85个部件完好是系统才能正常工作，求系统正常工作的概率。

解：正常工作部件的数目 X 服从二项分布 $b(100,0.9)$

$$E(X)=90,\ Var(X)=100\cdot 0.9\cdot 0.1=9,\ \text{所以}\ X\sim N(90,9)$$

$$\frac{X-90}{3}\sim N(0,1)$$

$$P(\text{正常工作})=P(X\geq 85)=P\left(\frac{X-90}{3}\geq\frac{85-90}{3}\right)\approx 1-\Phi\left(-\frac{5}{3}\right)=\Phi\left(\frac{5}{3}\right)$$

中心极限定理

例．设系统由一些相互独立的部件组成，运行时间每个部件损坏的概率为0.1， 至少有80%个部件完好是系统才能正常工作，问部件数n至少为多少才能使系统正常工作的概率不小于0.95。Φ(1.645)=0.95

解：正常工作部件的数目 X 服从二项分布 $b(n, 0.9)$,

$$X \sim N(0.9n, 0.09n), \quad \frac{X-0.9n}{0.3\sqrt{n}} \sim N(0,1)$$

$$P(X \geq 0.8n) = P\left(\frac{X-0.9n}{0.3\sqrt{n}} \geq \frac{0.8n-0.9n}{0.3\sqrt{n}}\right) \approx 1-\Phi\left(-\frac{0.1n}{0.3\sqrt{n}}\right) = \Phi\left(\frac{\sqrt{n}}{3}\right) \geq 0.95$$

$$\Rightarrow \frac{\sqrt{n}}{3} \geq u_{0.05} = 1.645 \quad \Rightarrow n \geq 25$$

三大统计分布，卡方、t、F

I. $X_1, X_2, \cdots, X_n$ 独立同分布，服从 $N(0,1)$ 则

$$Y = X_1^2 + \cdots + X_n^2 \sim \chi_n^2 \text{ 或 } \chi^2(n)$$，称为自由度为 n 的 χ^2 分布。

II. t 分布 $X_1 \sim N(0,1)$，$X_2 \sim \chi_n^2$，X_1, X_2 相互独立

$$t = \frac{X_1}{\sqrt{X_2 / n}} \sim t(n)$$，称为自由度为 n 的 t 分布。

III. F 分布 X_1, X_2 相互独立，$X_1 \sim \chi_m^2$，$X_2 \sim \chi_n^2$

$$F = \frac{X_1 / m}{X_2 / n} \sim F(m, n)$$，称为自由度为 m 与 n 的 F 分布。

随机变量函数的分布

已知随机变量 X 的分布，求 $Y=g(X)$ 的分布

1. 离散随机变量函数的分布

$$\begin{array}{cccccc} X & x_1 & x_2 & \cdots & x_n & \cdots \\ P & p_1 & p_2 & \cdots & p_n & \cdots \end{array} \quad \overset{Y=g(X)}{\Longrightarrow} \quad \begin{array}{cccccc} Y & g(x_1) & g(x_2) & \cdots & g(x_n) & \cdots \\ P & p_1 & p_2 & \cdots & p_n & \cdots \end{array}$$

2. 连续随机变量函数的分布

$$Y=g(X) \Rightarrow F_Y(y)=P(Y\le y)=P(g(X)\le y)$$

随机变量函数的分布例题

例 1. 已知随机变量

X	-2	-1	0	1	2
P	0.2	0.1	0.1	0.3	0.3

，求 $Y=X^2+X$ 的分布

解:

Y	2	0	0	2	6
P	0.2	0.1	0.1	0.3	0.3

$\Rightarrow$

Y	0	2	6
P	0.2	0.5	0.3

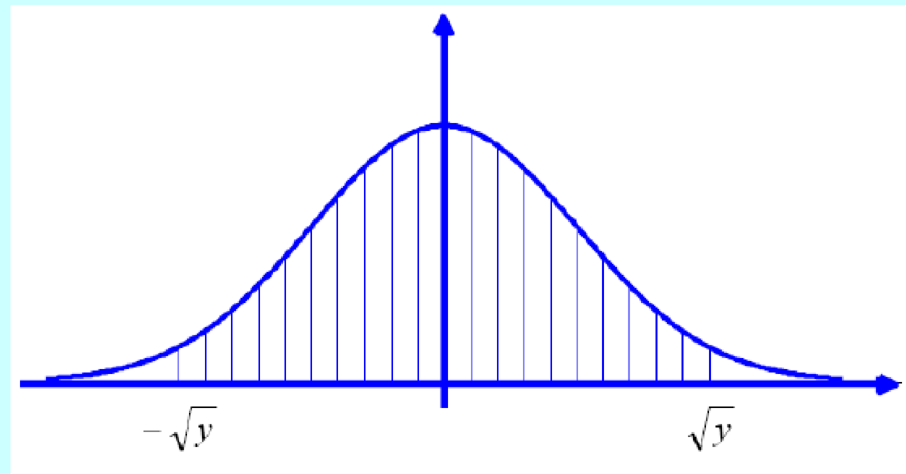
例 2. 设 $X \sim N(\mu,\sigma^2)$，求 $Y=\dfrac{X-\mu}{\sigma}$ 的分布

$$F_Y(y)=P(Y\le y)=P\left(\frac{X-\mu}{\sigma}\le y\right)=P(X\le\sigma y+\mu)=\int_{-\infty}^{\sigma y+\mu}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx$$

做变量代换 $t=\dfrac{x-\mu}{\sigma}$， $F_Y(y)=\displaystyle\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{t^2}{2}\right\}dt$， 即 $Y\sim N(0,1)$。

随机变量函数的分布例题

例 3. 设 $X \sim N(0,1)$，求 $Y = X^2$ 的分布



$$F_Y(y) = P(Y \le y) = P\left(X^2 \le y\right),$$

当 $y \le 0$ 时，$F_Y(y) = 0$

当 $y > 0$ 时，$F_Y(y) = P\left(-\sqrt{y} \le X \le \sqrt{y}\right) = 2\left[\Phi\left(\sqrt{y}\right) - \frac{1}{2}\right] = 2\Phi\left(\sqrt{y}\right) - 1$

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \varphi\left(\sqrt{y}\right) y^{\frac{-1}{2}}, & y > 0 \\ 0, & y \le 0 \end{cases}$$

注：当 $X \sim N(0,1)$ 时，X^2 称为一个自由度的 χ^2 分布，χ^2 分布是统计学中一类有重要应用的分布。

从均匀分布得到一般分布

设随机变量 U 服从 $[0,1]$ 上的均匀分布，函数 F 为定义在实数集合 R 的连续单调递增函数，且对任何 $x \in R$ 有

$F(-\infty) = 0 \le F(x) \le 1 = F(+\infty)$。

则随机变量 $X = F^{-1}(U)$ 的概率分布函数为 $F(x)$。

练习：利用 $U(0,1)$ 分布的随机数生成服从期望为 $\dfrac{1}{\lambda}$ 的指数分布随机数。

连续随机变量的函数的分布

定理： 设 n 维随机变量 $X=(X_1,X_2,\cdots,X_n)$ 的密度函数为 $p_X(x_1,x_2,\cdots,x_n)$，n 元函数 $\{g_i(x_1,x_2,\cdots,x_n)\}_{i=1}^n$ 满足条件

(1) 存在唯一的反函数 $x_i(y_1,y_2,\cdots,y_n)$，即存在 $g_i(x_1,x_2,\cdots,x_n)=y_i$ 的唯一实数解

(2) $g_i(x_1,x_2,\cdots,x_n)$ 与 $x_i(y_1,y_2,\cdots,y_n)$ 都连续

(3) 存在连续的偏导数 $\dfrac{\partial x_i}{\partial y_j}$，$\dfrac{\partial g_i}{\partial x_j}$，记 $J=\det\left(\begin{bmatrix}\dfrac{\partial x_1}{\partial y_1} & \cdots & \dfrac{\partial x_1}{\partial y_n}\\ \vdots & \ddots & \vdots\\ \dfrac{\partial x_n}{\partial y_1} & \cdots & \dfrac{\partial x_n}{\partial y_n}\end{bmatrix}\right)$

则 n 维随机变量 $Y=(Y_1,Y_2,\cdots,Y_n)^T=(g_1(X_1,\cdots,X_n),\cdots,g_n(X_1,\cdots,X_n))^T$ 有密度函数

$$p_Y(y_1,\cdots,y_n)=\begin{cases}p_X(x_1(y_1,\cdots,y_n),\cdots,x_n(y_1,\cdots,y_n))\cdot|J|, & \text{当}\bigcap_{i=1}^{n}x_i(y_1,\cdots,y_n)\neq\Phi\\ 0, & \text{其他}\end{cases}$$

非线性一般情形的线性化

一般情形：映射 $T: R^n \to R^n$，　$(y_1, \cdots, y_n) = T(x_1, \cdots, x_n)$，则

$$\iint_D f(x_1, \cdots, x_n) dx_1 \cdots dx_n = \iint_{D'} f(x_1(y_1, \cdots, y_n), \cdots, x_n(y_1, \cdots, y_n)) \cdot |J_n| dy_1 \cdots dy_n$$

其中 $J_n = \dfrac{\partial(x_1, \cdots, x_n)}{\partial(y_1, \cdots, y_n)} = \det\left(\begin{bmatrix} \dfrac{\partial x_1}{\partial y_1} & \cdots & \dfrac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial x_n}{\partial y_1} & \cdots & \dfrac{\partial x_n}{\partial y_n} \end{bmatrix}\right)$

Jacobi 矩阵 $\left[\dfrac{\partial x_i}{\partial y_j}\right]_{n \times n}$ 为 T 的线性化，即 ***Taylor*** 展开的线性项。

例 5. (X,Y) 的密度函数为 $p_{XY}(x,y)=\dfrac{1+xy}{4}$（$|x|<1,|y|<1$），求 X^2 和 Y^2 的联合概率密度函数。

解：设 $\begin{cases}U=X^2\\V=Y^2\end{cases}\Rightarrow\begin{cases}x^{(1)}=\sqrt{u}\\y^{(1)}=\sqrt{v}\end{cases}$，$\begin{cases}x^{(2)}=-\sqrt{u}\\y^{(2)}=\sqrt{v}\end{cases}$，$\begin{cases}x^{(3)}=\sqrt{u}\\y^{(3)}=-\sqrt{v}\end{cases}$，$\begin{cases}x^{(4)}=-\sqrt{u}\\y^{(4)}=-\sqrt{v}\end{cases}$

$$\left|J^{(1)}\right|=\left|J^{(2)}\right|=\left|J^{(3)}\right|=\left|J^{(4)}\right|=\frac{1}{4\sqrt{uv}},$$

$$p_{UV}(u,v)=\frac{1}{4\sqrt{uv}}\left[p_{XY}\left(\sqrt{u},\sqrt{v}\right)+p_{XY}\left(-\sqrt{u},\sqrt{v}\right)+p_{XY}\left(\sqrt{u},-\sqrt{v}\right)+p_{XY}\left(-\sqrt{u},-\sqrt{v}\right)\right]$$
$$=\frac{1}{4\sqrt{uv}}I_{0<u<1,0<v<1}$$

$\dfrac{1}{4\sqrt{uv}}I_{0<u<1,0<v<1}=\dfrac{1}{2\sqrt{u}}I_{0<u<1}\cdot\dfrac{1}{2\sqrt{v}}I_{0<v<1}$，所以 X^2 和 Y^2 相互独立。

生成正态随机变量

Box-Muller 方法（Ann. Math. Stat. 1958, 29(2), 610-611）顶级学术杂志研究短文

X,Y 相互独立，且均服从 $U(0,1)$，做变换 $\begin{cases} U=(-2\ln X)^{1/2}\cos 2\pi Y \\ V=(-2\ln X)^{1/2}\sin 2\pi Y \end{cases}$，则得到的 U 和 V 相互独立，且均服从 $N(0,1)$。

证明：

$$\begin{cases} u=(-2\ln x)^{1/2}\cos 2\pi y \\ v=(-2\ln x)^{1/2}\sin 2\pi y \end{cases} \Rightarrow \begin{cases} x=\exp\left\{-\dfrac{u^2+v^2}{2}\right\} \\ y=\dfrac{1}{2\pi}\arctan\dfrac{v}{u} \end{cases}$$

$$\frac{\partial(x,y)}{\partial(u,v)}=\frac{1}{2\pi}e^{-\frac{u^2+v^2}{2}}\begin{vmatrix} -u & -v \\ -\dfrac{v}{u^2}\cdot\dfrac{1}{1+(v/u)^2} & \dfrac{1}{u}\cdot\dfrac{1}{1+(v/u)^2} \end{vmatrix}=\frac{-1}{2\pi}e^{-\frac{u^2+v^2}{2}}$$

$$p_{UV}(u,v)=p_{XY}(x(u,v),y(u,v))\cdot\frac{1}{2\pi}e^{-\frac{u^2+v^2}{2}}=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{v^2}{2}}$$

统计抽样定理的证明

设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N\left(\mu, \sigma^2\right)$ 的样本，其样本均值和样本方差分别为：

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \text{ 和 } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2, \text{ 则有}$$

（1）$\bar{X}$ 与 S^2 相互独立

（2）$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

（3）$\frac{(n-1)\cdot S^2}{\sigma^2} \sim \chi^2(n-1)$

$$X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N\left(\mu, \sigma^2\right)$$

$$\begin{cases} Y_1 & = & \dfrac{1}{\sqrt{1\cdot 2}}\left(X_1 - X_2\right) \\ Y_2 & = & \dfrac{1}{\sqrt{2\cdot 3}}\left(X_1 + X_2 - 2X_3\right) \\ \cdots & \cdots & \cdots \\ Y_{n\text{-}1} & = & \dfrac{1}{\sqrt{(n-1)\cdot n}}\left(X_1 + X_2 + \cdots + X_{n-1} - (n-1)X_n\right) \\ Y_n & = & \dfrac{1}{\sqrt{n}}\left(X_1 + X_2 + \cdots + X_n\right) \end{cases}$$

Jacobi矩阵是正交矩阵， $(Y_1, Y_2, \ldots, Y_n)$ 为 n 维正态分布

$Y_1, Y_2, \ldots, Y_n$ 两两不相关，所以 $Y_1, Y_2, \ldots, Y_n$ 相互独立

$$Y_n = \sqrt{n}\,\bar{X}, \quad \sum_{k=1}^{n}(X_k - \bar{X})^2 = \sum_{k=1}^{n} X_k{}^2 - n\bar{X}^2 = \sum_{k=1}^{n} Y_k{}^2 - Y_n{}^2 = \sum_{k=1}^{n-1} Y_k{}^2$$

最大值与最小值的分布

例 8. 设 $X_1,X_2,\cdots,X_n$ 相互独立，且 X_i 的分布函数为 $F_X(x)$， $i=1,\cdots,n$ 。

$Y=\max(X_1,X_2,\cdots,X_n)$ 与 $Z=\min(X_1,X_2,\cdots,X_n)$ 的分布。

解： $$F_Y(y)=P(\max(X_1,\cdots,X_n)\le y)=P(X_1\le y,\cdots,X_n\le y)=\prod_{i=1}^{n}P(X_i\le y)$$

$$=\prod_{i=1}^{n}F_X(y)。$$

若 X_i 的分布函数均为 $p_X(x)$，则 $p_Y(y)=F_Y{}'(y)=\left[F_X(y)^n\right]'=n[F_X(y)]^{n-1}p_X(y)$。

$$F_Z(z)=P(\min(X_1,\cdots,X_n)\le z)=1-P(\min(X_1,\cdots,X_n)>z)$$

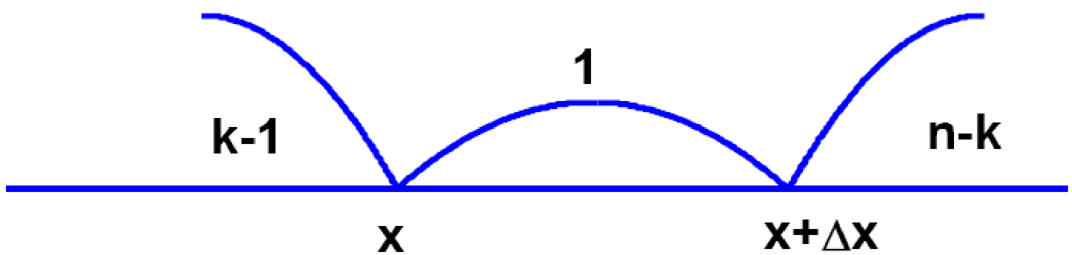
$$=1-\prod_{i=1}^{n}P(X_i>z)=1-\prod_{i=1}^{n}[1-P(X_i\le z)]=1-\prod_{i=1}^{n}[1-F_X(z)]$$

次序统计量

利用差分的方式进行计算

例 9. $X_1, X_2, \cdots, X_n$ 独立同分布，分布函数 $F(x)$，密度函数 $p(x)$，将这 n 个随机变量做升序排列 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$。求 $X_{(k)}$ 的分布，$F_k(x)$，$p_k(x)$。

解：$X_{(k)}$ 落于 $[x, \Delta x]$ 的概率



$$F_k(x+\Delta x) - F_k(x)$$

$$= \frac{n!}{(k-1)!(n-k)!} \cdot [F(x)]^{k-1} \cdot [F(x+\Delta x) - F(x)] \cdot [1 - F(x+\Delta x)]^{n-k}$$

令 $\Delta x \to 0$，得 $p_k(x) = \frac{n!}{(k-1)!(n-k)!} \cdot [F(x)]^{k-1} \cdot p(x) \cdot [1 - F(x)]^{n-k}$。

$$\sup_{x\in R} | F_n(x) - F(x) | \xrightarrow{a.s.} 0.$$

经验分布函数

定义 1.1.5 设 $\xi_1,\xi_2,\cdots,\xi_n$ 为总体 ξ 的样本，$\xi_{(1)},\xi_{(2)},\cdots,\xi_{(n)}$ 为样本 $\xi_1,\xi_2,\cdots,\xi_n$ 的顺序统计量. 对任意实数 x，记

$$F_n(x)=\begin{cases} 0, & x\leqslant \xi_{(1)}, \\ \dfrac{k}{n}, & \xi_{(k)}<x\leqslant \xi_{(k+1)},k=1,2,\cdots,n-1, \\ 1, & x>\xi_{(n)}, \end{cases} \tag{1.1.17}$$

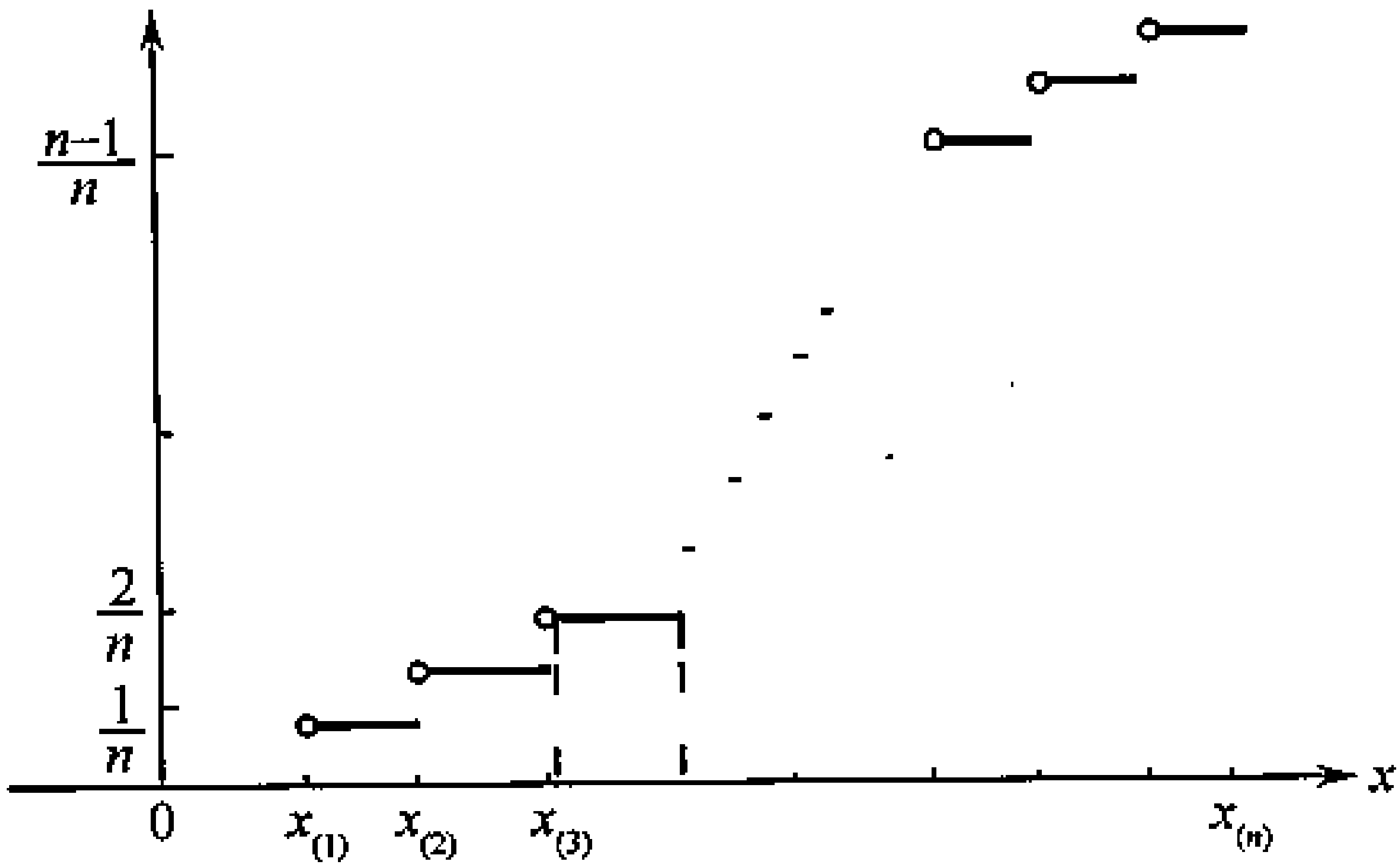
则称 $F_n(x)$为总体 ξ 的经验分布函数. $F_n(x)$是分段函数不便于使用，为此引入单位阶跃函数：

$$\mu(x)=\begin{cases} 1, & x>0, \\ 0, & x\leqslant 0, \end{cases} \tag{1.1.18}$$

$$\mu(x)=\begin{cases}1, & x>0,\\ 0, & x\leqslant 0,\end{cases}\tag{1.1.18}$$

则式(1.1.17)可改写为

$$F_n(x)=\frac{1}{n}\sum_{k=1}^{n}\mu(x-\xi_k),\quad x\in R,\tag{1.1.19}$$



模拟

- n=1000;
- a=randn(n,1);
- b=sort(a);
- c=b(100:100:1000);
- c=b(100:100:900);
- [c,norminv(0.1:0.1:0.9)']
- hist(a,30)

作业

- 习题1
- 4, 6, 33, 34　（10月10日）

- 大作业一 （10月17日）