第6章 大数定律及中心极限定理

6.1 大数定律

- 切比雪夫大数定律
- 伯努利大数定理
- 辛钦大数定理

6.2 中心极限定理

- 独立同分布的中心极限定理
- 李雅普诺夫（Liapunov)定理
- 棣莫佛--拉普拉斯定理
- 林德贝格中心极限定理

6.1 大数定律

(1) 频率具有稳定性

n 次独立试验中，A 发生的频率记为 f_n，则

$$f_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \to p.(= P(A))$$

(2) 平均值具有稳定性

n 次独立测量中，记算术平均为 $\overline{X}_n$,数学期望为 μ，则

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu.(= \mathrm{E}(X))$$

定义1：设 $X_n\ (n=1, 2, \ldots)$是一随机变量序列，若存在随机变量 X，使得对任意的正数ε，恒有 $\lim\limits_{n\to\infty} P\{|X_n - X| \geq \varepsilon\} = 0$

或 $\lim\limits_{n\to\infty} P\{|X_n - X| < \varepsilon\} = 1$ 则称随机变量序列$\{X_n\}$依概率收敛于随机变量 X，记作 $\lim\limits_{n\to\infty} X_n = X(P)$或$X_n \xrightarrow{P} X.$

解释：记 $A_n=\{|X_n- X|<\varepsilon\}$，$p_n=P(A_n)\to 1$，当$n\to\infty$时.

特别地，X 为常数a, 则 $\lim\limits_{n\to\infty} P\{|X_n - a| \geq \varepsilon\} = 0$

一、切比雪夫大数定律:

设 $X_1,\ X_2,\ \ldots, X_n, \ldots,$ 是由相互独立的 r.v. 所构成的序列, $\mathrm{E}(X_k) = \mu_k$,
并且它们的方差有公共的上界 $\mathrm{D}(X_k) \le C\ (k = 1, 2\ \cdots),$

则对 $\forall \varepsilon > 0,$ 都有 $\lim\limits_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n}\mu_k\right| < \varepsilon\right\} = 1.$

$$P\{|X-\mu| \ge \varepsilon\} \le \frac{\sigma^2}{\varepsilon^2} \text{ 或} P\{|X-\mu| < \varepsilon\} \ge 1-\frac{\sigma^2}{\varepsilon^2}$$

这一不等式称为Chebyshev不等式.

证明：因为$\mathrm{E}(\frac{1}{n}\sum_{k=1}^{n}X_k)=\frac{1}{n}\sum_{k=1}^{n}\mu_k$，$\mathrm{D}(\frac{1}{n}\sum_{k=1}^{n}X_k)=\frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^2\le\frac{C}{n}$

由契比雪夫不等式可得：

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\frac{1}{n}\sum_{k=1}^{n}\mu_k\right|<\varepsilon\right\}\ge 1-\frac{\mathrm{D}(\frac{1}{n}\sum_{k=1}^{n}X_k)}{\varepsilon^2}\ge 1-\frac{C}{n\varepsilon^2}$$

令$n\to\infty$ 即得 $\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\frac{1}{n}\sum_{k=1}^{n}\mu_k\right|<\varepsilon\right\}=1.$

二. 伯努利大数定理:

设 n_A 是 n 次独立重复试验中A发生的次数, p 是事件A在每次试验中发生的概率, 则 对于$\forall \varepsilon > 0$, 有 $\lim\limits_{n\to\infty} P\left\{\left|\frac{n_A}{n} - p\right| < \varepsilon\right\} = 1$ 或

$$\lim_{n\to\infty} P\left\{\left|\frac{n_A}{n} - p\right| \geq \varepsilon\right\} = 0, \text{ 即 } \frac{n_A}{n} \xrightarrow{P} p.$$

➢ 伯努利定理说明:

- 事件 A 发生的频率 n_A/n 依概率收敛到事件 A 发生的概率 p, 这就以严格的数学形式表达了频率的稳定性;
- 就是说, 当 n 很大时, 事件A发生的频率与概率有较大差别的可能性很小, 因而在实际中便可以用频率来代替概率.

证明： 因为$n_A \sim b(n,p)$,则有 $n_A = X_1 + X_2 + \cdots + X_n$,

其中，X_1，X_2,…，X_n相互独立，且都服从以p 为参数的（0-1）分布。

因而$\mathrm{E}(X_k) = p$，$\mathrm{D}(X_k) = p(1-p)\ (k = 1,2,\cdots,n)$

由切比雪夫大数定理有:

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}(X_1 + X_2 + \cdots + X_n) - p\right| < \varepsilon\right\} = 1$$

即$$\lim_{n\to\infty} P\left\{\left|\frac{n_A}{n} - p\right| < \varepsilon\right\} = 1$$

三. 辛钦大数定理:

设 r.v. $X_1, X_2, \ldots, X_n, \ldots$ 相互独立，服从同一分布，且具数学期望

$\mathrm{E}(X_k) = \mu, (k = 1, 2, \cdots)$, 则对$\forall \varepsilon > 0$, 有 $\lim\limits_{n\to\infty} P\left\{ \left| \frac{1}{n}\sum\limits_{k=1}^{n} X_k - \mu \right| < \varepsilon \right\} = 1.$

- 为什么可以用样本均值代替总体均值，或者作为总体均值的估计？

6.2 中心极限定理

一. 独立同分布的中心极限定理:

设 r.v. X_k (k=1, 2, …) 相互独立, 服从同一分布(i.i.d.)且具有有限的数学期望和方差: $\mathrm{E}(X_k)=\mu, \mathrm{D}(X_k)=\sigma^2\neq 0, k=1, 2, \cdots$ 则r.v.

$$Y_n=\frac{\sum_{k=1}^{n}X_k-\mathrm{E}(\sum_{k=1}^{n}X_k)}{\sqrt{\mathrm{D}(\sum_{k=1}^{n}X_k)}}=\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}$$ 的分布函数F_n对于$\forall x$

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\{Y_n\le x\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2})\,\mathrm{d}t$$

即 r.v. 序列 $Y_n=\dfrac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}\xrightarrow{L}N(0, 1).$

二. 李雅普诺夫（Liapunov)定理

设随机变量X_1，X_2,…，X_n,…相互独立，它们具有数学期望和方差：

$$\mathrm{E}(X_k)=\mu_k,\ \mathrm{D}(X_k)=\sigma_k^2>0, k=1,2,\cdots,\ \text{记}B_n^2=\sum_{k=1}^{n}\sigma_k^2.$$

若存在正数δ，使得当$n\to\infty$时，$\dfrac{1}{B_n^{2+\delta}}\sum_{k=1}^{n}\mathrm{E}\{|X_k-\mu_k|^{2+\delta}\}\to 0$

则随机变量之和$\sum_{k=1}^{n}X_k$的标准化变量

$$Z_n=\frac{\sum_{k=1}^{n}X_k-\mathrm{E}(\sum_{k=1}^{n}X_k)}{\sqrt{\mathrm{D}(\sum_{k=1}^{n}X_k)}}=\frac{\sum_{k=1}^{n}X_k-\sum_{k=1}^{n}\mu_k}{B_n}$$

Z_n的分布函数$F_n(x)$对于任意x，满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\{\frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n} \le x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x)$$

回顾

- 大数定律：随机变量序列的前一些项的算术平均值依概率收敛到均值的算术平均值；
  - 切比雪夫大数定律：独立，公共上界；
  - 辛钦大数定理：独立同分布；
  - 伯努利大数定理：n重伯努利试验；

三. 棣莫佛--拉普拉斯定理:

设$r.v.X_n(n=1,\ 2,\ \cdots)$服从二项分布$b(n,p)$, 对于$\forall x$, 恒有

$$\lim_{n\to\infty} P\left\{\frac{X_n-np}{\sqrt{np(1-p)}}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t;$$

$$\lim_{n\to\infty} P\left\{a<\frac{X_n-np}{\sqrt{npq}}\le b\right\}=\int_{a}^{b}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\Phi(b)-\Phi(a),$$

因而当n较大时, 我们可以用正态 分布的数值表来近似计算二项分布的概率,又为二项分布找到了一个近似计算公式, 在使用时, 只有当p很小时才能用$Poisson$分布来近似二项分布,而用棣莫佛 – –拉普拉斯定理时则没有这个限制.

中心极限定理实际上讲述 $\overline{X}$ 具有近 似正态性，$\overline{X}=\frac{1}{n}X_1+\frac{1}{n}X_2+\cdots+\frac{1}{n}X_n$, r.v. $\overline{X}$ 受诸多因素的影响, 每个因素对 $\overline{X}$ 的影响均具有权$\frac{1}{n}$是相同的, 这表明这么多的因素 $X_1, X_2, \cdots, X_n$ 中不能指出哪一个因素的影响最大；由此, n 相当大时, $\overline{X}$ 应具有正态性.

四. 林德贝格中心极限定理

(1)（林德贝格条件）设随机变量 $X_1, X_2, \ldots, X_n, \ldots$相互独立，具有有限的数学期望及方差：$\mathrm{E}(X_k)=\mu_k, \mathrm{D}(X_k)=\sigma_k^2 (k=1,2,\cdots,n,\cdots)$

记$B_n^2=\sum_{k=1}^{n}\sigma_k^2$，记$X_k$的概率密度为$f_k(x)(k=1,2,\cdots,n,\cdots)$ 林德贝格条件为：对任意的$\varepsilon>0$,有

$$\lim_{n\to\infty}\frac{1}{B_n^2}\sum_{k=1}^{n}\int_{|x-\mu_k|>\varepsilon B_n}(x-\mu_k)^2 f_k(x)\mathrm{d}x=0$$

(2) (林德贝格定理）对任意实数 x, 有:

$$\lim_{n\to\infty} P\left\{\frac{\sum_{k=1}^{n}(X_k-\mu_k)}{B_n}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t \text{ 且 } \frac{|X_k-\mu_k|}{B_n}\xrightarrow{P}0$$

记 $A_k=\left\{\frac{|X_k-\mu_k|}{B_n}>\varepsilon\right\}(k=1,2,\cdots,n,\cdots)$, 则

$$P\left\{\max_{1\le k\le n}\frac{|X_k-\mu_k|}{B_n}>\varepsilon\right\}=P(\bigcup_{k=1}^{n}A_k)\le\sum_{k=1}^{n}P(A_k)$$

$$=\sum_{k=1}^{n}\int_{|x-\mu_k|>\varepsilon B_n} f_k(x)\mathrm{d}x$$

$$\le\frac{1}{\varepsilon^2B_n^2}\sum_{k=1}^{n}\int_{|x-\mu_k|>\varepsilon B_n}(x-\mu_k)^2f_k(x)\mathrm{d}x$$

令$n \to \infty$,可知上式右端极限为0，故

$$\lim_{n\to\infty} P\left\{\max_{1\le k\le n}\frac{|X_k-\mu_k|}{B_n}>\varepsilon\right\}=0$$

上式表明，当n充分大时，和式中项$\frac{|X_k-\mu_k|}{B_n}$一致地依概率收敛于0

为什么$\left\{\max_{1\le k\le n}\frac{|X_k-\mu_k|}{B_n}>\varepsilon\right\}=\bigcup_{k=1}^{n}A_k$?

- **中心极限定理：**大量随机变量之和的分布逼近于正态分布；
  - 独立同分布的中心极限定理；
  - 李雅普诺夫（Liapunov)定理：独立、$\frac{1}{B_n^{2+\delta}}\sum_{k=1}^{n}\mathrm{E}\{|X_k-\mu_k|^{2+\delta}\}\to 0$
  - 棣莫佛--拉普拉斯定理：二项分布；
  - 林德贝格中心极限定理：

- **中心极限定理的意义：**概率论中最著名的结果之一
  - 提供了计算独立随机变量之和的近似概率的简单方法；
  - 有助于解释为什么很多自然群体的经验频率呈现出钟形曲线这一值得注意的事实；

例1. 一加法器同时收到 20个噪声电压 $V_k(k=1,2,\cdots,20)$, 设它们是相互独立的r.v., 且都在区间 $(0,10)$ 上服从均匀布, 记 $V=\sum_{k=1}^{20}V_k$, 求 $P\{V>105\}$ 的近似值.

解：易知 $\mathrm{E}(V_k)=5$, $\mathrm{D}(V_k)=100/12(k=1,2,\cdots,20)$.

由独立同分布的中心极限定理知, r.v.

$$Z=\frac{\sum_{k=1}^{20}V_k-20\times 5}{\sqrt{100/12}\sqrt{20}}=\frac{V\text{-}20\times 5}{\sqrt{100/12}\sqrt{20}}$$ 近似服从正态分布$N(0,1)$, 于是

$$P\{V>105\}=P\{\frac{V\text{-}20\times 5}{\sqrt{100/12}\sqrt{20}}>\frac{105\text{-}20\times 5}{\sqrt{100/12}\sqrt{20}}\}$$

$$=P\{\frac{V\text{-}20\times 5}{\sqrt{100/12}\sqrt{20}}>0.387\}$$

$$= 1 - P\{\frac{V - 20\times 5}{\sqrt{100/12}\sqrt{20}} \le 0.387\}$$

$$= 1 - \Phi(0.387)$$

$$\approx 0.349$$

即有$P\{V > 105\} \approx 0.349.$

例2. 一船舶在某海区航行, 已知每遭受一次波浪的冲击, 纵摇角大于 3^0 的概率 $p=1/3$, 若船舶遭受了90000次波浪冲击, 问其中有29500~30500次纵摇角大于 3^0概率是多少?

解: 将船舶每遭受一次波浪冲击看成是一次试验, 并假定每次试验是独立的;

在90000次波浪冲击中纵摇角度大于3^0的次数记为 X, 则 X是一个 r.v.且 $X\sim b(90000,1/3)$ 其分布律

$$P\{X=k\}=C_{90000}^{k}\left(\frac{1}{3}\right)^{k}\left(\frac{2}{3}\right)^{90000-k},\ k=0,1,\cdots,90000.$$

所求概率为 $\mathrm{P}\{29500<X\leq 30500\}=\sum_{k=29500}^{30500}C_{90000}^{k}\left(\frac{1}{3}\right)^{k}\left(\frac{2}{3}\right)^{90000-k}$

显然, 直接计算十分麻烦, 利用棣莫佛-拉普拉斯定理来近似求解, 即有:

$$P\{29500 < X \le 30500\} = P\left\{\frac{29500 - np}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} \le \frac{30500 - np}{\sqrt{np(1-p)}}\right\}$$

$$\approx \Phi\left(\frac{30500 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{29500 - np}{\sqrt{np(1-p)}}\right)$$

其中$n = 90000, p = 1/3.$即有

$$P\{29500 < X \le 30500\} \approx \Phi(5\sqrt{2}/2) - \Phi(-5\sqrt{2}/2)$$

$$= 2\Phi(5\sqrt{2}/2) - 1 \approx 2\Phi(3.536) - 1$$

$$= 0.9995.$$

例3. 有240台电话分机，独立使用，每台话机约有5%的时间使用外线。问总机至少需要多少外线才能90%以上的概率保证各分机用外线不必等候。

解：设 X 为240台分机中同时需用外线的台数，显然 $X \sim b(240,0.05)$.

即求最小的 N, 使得 $P\{0 \le X \le N\} \ge 0.9$

由于n =240很大，而 $\mathrm{E}(X) = np = 240 \times 0.05 = 12$

$\mathrm{D}(X) = npq = 12 \times 0.95 = 11.4$，由棣莫佛—拉普拉斯定理知

$$P\{0 \le X \le N\} = P\{\frac{0-12}{\sqrt{11.4}} \le \frac{X-12}{\sqrt{11.4}} \le \frac{N-12}{\sqrt{11.4}}\}$$

$$\approx \Phi(\frac{N-12}{\sqrt{11.4}}) - \Phi(\frac{-12}{\sqrt{11.4}})$$

$$\approx \Phi(\frac{N-12}{3.38}) \ge 0.9$$

查正态分布表得 $\Phi(1.28) = 0.8997 < 0.9,$

$$\Phi(1.29) = 0.9015 > 0.9$$

因而得 $\frac{N-12}{3.38} \ge 1.29,$ 于是 $N \ge 16.36$

故取 $N = 17.$ 即总机至少需要17条外线才可满足要求。

例4　一生产线生产的产品成箱包装，每箱的重量是随机的。假设每箱平均重50kg，标准差为5kg。若用最大载重量为5t的汽车承运，试利用中心极限定理说明每辆车最多可以装多少箱，才能保障不超载的概率大于0. 977（$\Phi(2)=0.977$,其中$\Phi(x)$是标准正态分布函数）。

解：设$X_i=\{$装运的第i箱的重量$\}$，　$(i=1,2,\cdots,n)$(单位：kg)，n是所求箱数。由条件可以把$X_1,X_2,\cdots,X_n$视为独立同分布的随机变量，而n箱总重量

$T_n=X_1+X_2+\cdots+X_n$　是独立同分布的随机变量之和。

由条件知$\mathrm{E}(X_i)=50,\ \sqrt{\mathrm{D}(X_i)}=5$,从而有

根据独立同分布的中心极限定理，$T_n \overset{\text{近似地}}{\sim} N(50n,25n)$,

$$\mathrm{E}(T_n)=50n,\ \sqrt{\mathrm{D}(T_n)}=5\sqrt{n}$$

即$\dfrac{T-50n}{5\sqrt{n}} \overset{\text{近似地}}{\sim} N(0,1)$

由$P\{T_n \le 5000\} = P\{\dfrac{T_n - 50n}{5\sqrt{n}} \le \dfrac{5000 - 50n}{5\sqrt{n}}\}$

$$\approx \Phi(\frac{1000 - 10n}{\sqrt{n}}) > 0.977 = \Phi(2)$$

由此可见$\dfrac{1000 - 10n}{\sqrt{n}} > 2$ 从而$n < 98.02$，即最多可以装98箱。

例5. 对于一个学生而言，来参加家长会的家长人数是一个随机变量，设一个学生无家长、1名家长、2名家长来参加会议的概率分别为0.05, 0.8, 0.15. 若学校共有400名学生，设各学生参加会议的家长数相互独立，且服从同一分布。

（1）求参加会议的家长数 X 超过450人的概率；

（2）求恰有1名家长来参加会议的学生数不多于340的概率。

解：(1) 以$X_k(k=1,2,\cdots,400)$，记第k个学生来参加会议的家长数，则X_k的分布律为

X_k	0	1	2
p_k	0.05	0.8	0.15

易知$\mathrm{E}(X_k)=1.1,\ \mathrm{D}(X_k)=0.19,\ k=1,2,\cdots,400.$

而$X=\sum_{k=1}^{400}X_k$.由独立同分布的中心极限定理，随机变量

$\dfrac{\sum_{k=1}^{400}X_k-400\times1.1}{\sqrt{400}\times\sqrt{0.19}}=\dfrac{X-400\times1.1}{\sqrt{400}\sqrt{0.19}}$ 近似地服从正态分布$N(0,1)$,于是

$$P\{X>450\}=P\left\{\frac{X-400\times1.1}{\sqrt{400}\sqrt{0.19}}>\frac{450-400\times1.1}{\sqrt{400}\sqrt{0.19}}\right\}$$

$$=1-P\left\{\frac{X-400\times1.1}{\sqrt{400}\sqrt{0.19}}\le1.147\right\}$$

$$\approx1-\Phi(1.147)\approx1-0.8743$$

$$=0.1257$$

(2) 以Y记恰有一名家长来参加会议的学生数，则$Y \sim b(400, 0.8)$,

由棣莫弗-拉普拉斯定理得

$$P\{Y \le 340\} = P\{\frac{Y - 400\times 0.8}{\sqrt{400\times 0.8\times 0.2}} \le \frac{340 - 400\times 0.8}{\sqrt{400\times 0.8\times 0.2}}\}$$

$$= P\{\frac{Y - 400\times 0.8}{\sqrt{400\times 0.8\times 0.2}} \le 2.5\}$$

$$\approx \Phi(2.5)$$

$$= 0.9938$$

马尔可夫不等式 设 X 为取非负值的随机变量，则对于任何常数 $a>0$，有

$$P\{X \geq a\} \leq \mathrm{E}(X)/a$$

证明： 对于 $a>0$，令 $I=\begin{cases}1, \text{若} X \geq a \\ 0, \text{其他}\end{cases}$。

并且注意到，由于 $X \geq 0$，有 $I \leq X/a$，

对不等式两边求期望，得 $\mathrm{E}(I) \leq \mathrm{E}(X)/a$。

$\mathrm{E}(I)=P\{X \geq a\}$，所以不等式成立。

命题[单边的切比雪夫不等式]

设X具有0均值和有限方差σ^2，则对任意$a>0$,

$$P\{X \geq a\} \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

证明： 令$b>0$,注意到

$$X \geq a \Leftrightarrow X + b \geq a + b$$

由于$a+b>0$，$X+b \geq a+b$可推知$(X+b)^2 \geq (a+b)^2$

故$P\{X \geq a\} = P\{X + b \geq a + b\} \leq P\{(X+b)^2 \geq (a+b)^2\}$

再利用马尔可夫不等式（$P(|X|\geq a)\leq \mathrm{E}(|X|)/a, a>0$），可得

$$P\{X\geq a\}\leq \frac{\mathrm{E}[(X+b)^2]}{(a+b)^2}=\frac{\sigma^2+b^2}{(a+b)^2}$$

上式中，b 可以取任意正常数，取$b=\sigma^2/a$，便得到本命题的结论。实际上。当$b=\sigma^2/a$时，$(\sigma^2+b^2)/(a+b)^2$达到极小值。

例5a 设某工厂每周的产量是一个随机变量，其均值为$\mu=100$ ，方差为$\sigma^2=400$。计算这一周产量至少为120的概率上界。

解:利用单边切比雪夫不等式

$$P\{X\geq 120\}=P\{X-100\geq 20\}\leq \frac{400}{400+20^2}=1/2$$

这说明本周产量至少为120 的概率不会超过$1/2$ 。

如果直接利用马尔科夫不等式，可得

$$P\{X \geq 120\} \leq \frac{\mathrm{E}(X)}{120} = 5/6$$

这个上界就比较弱，（上界越小，结论越强，若上界为1，这个结论就没有任何意义了）。

- 现在设X具有均值μ，方差σ^2，由于$X-\mu$ 与μ-X 都具有均值为0和方差σ^2，利用单边的切比雪夫不等式可知，对于$a>0$,

$$P\{X-\mu \geq a\} \leq \frac{\sigma^2}{\sigma^2+a^2}$$

$$P\{\mu-X \geq a\} \leq \frac{\sigma^2}{\sigma^2+a^2}$$

因此，得到下面的推论，

推论5.1　若$\mathrm{E}(X)=\mu$，　$\mathrm{Var}(X)=\sigma^2$，则对于$a>0$，下列不等式成立：

$$P\{X\geq\mu+a\}\leq\frac{\sigma^2}{\sigma^2+a^2}$$

$$P\{X\leq\mu-a\}\leq\frac{\sigma^2}{\sigma^2+a^2}$$

例5*b.* 一个由100个男人和100 个女人组成的集合，被随机分成两两一组的100组，试给出最多30个组是由一男一女组成的概率的上界。

解：对所有男人任意地从1 至100 进行编号，对于，$i=1,2,3,\cdots$

令

$$X_i=\begin{cases}1 & \text{男人}i\text{ 所在的组内有女人；}\\ 0 & \text{其他.}\end{cases}$$

这样，男女组的数量X 可以表示为，

$$X=\sum_{i=1}^{100}X_i$$

由已知第i 个男人和其他199 个人配对的概率是相等的，而其中有100个人是女人，我们有

$$\mathrm{E}(X_i)=P\{X_i=1\}=\frac{100}{199}$$

类似地，对于$i\neq j$

$$\mathrm{E}(X_iX_j)=P\{X_i=1,X_j=1\}$$

$$=P\{X_i=1\}P\{X_j=1\big|X_i=1\}=\frac{100}{199}\bullet\frac{99}{197}$$

其中$P\{X_j=1\big|X_i=1\}=\dfrac{99}{197}$。

这是因为当第i 个男人已经和一个女人配对时，男人j只可能跟剩余的197人配对，其中99 人为女人。

因此，我们得到

$$\mathrm{E}(X)=\sum_{i=1}^{100}\mathrm{E}(X_i)=100\cdot\frac{100}{199}\approx 50.25$$

$$\begin{aligned}\mathrm{Var}(X)&=\sum_{i=1}^{100}\mathrm{Var}(X_i)+2\sum\sum_{i<j}\mathrm{Cov}(X_i,X_j)\\&=100\cdot\frac{100}{199}\cdot\frac{99}{199}+2\cdot\binom{100}{2}\cdot\left[\frac{100}{199}\frac{99}{197}-\left(\frac{100}{199}\right)^2\right]\\&\approx 25.126\end{aligned}$$

由切比雪夫不等式可得，

$$P\{X \le 30\} \le P\{|X - 50.25| \ge 20.25\}$$

$$\le \frac{25.126}{(20.25)^2} \approx 0.061$$

由此看出，最多30 对为一男一女的概率上界为0.061 。然而，我们可以利用单边切比雪夫不等式对该上界进行改进, 得到

$$P\{X \le 30\} = P\{X \le 50.25 - 20.25\}$$

$$\le \frac{25.126}{25.126 + (20.25)^2} \approx 0.058$$

当随机变量X的矩母函数已知时，可以得到更加有效的$P\{X \geq a\}$的上界。令

$$M(t) = \mathrm{E}(\mathrm{e}^{tX})$$

为随机变量X的矩母函数，则对于$t > 0$, 有

$$P\{X \geq a\} = P\{\mathrm{e}^{tX} \geq \mathrm{e}^{ta}\} \leq \mathrm{E}(\mathrm{e}^{tX})\mathrm{e}^{-ta} \text{ [利用马尔科夫不等式]}$$

类似地，对于$t < 0$,

$$P\{X \leq a\} = P\{\mathrm{e}^{tX} \geq \mathrm{e}^{ta}\} \leq \mathrm{E}(\mathrm{e}^{tX})\mathrm{e}^{-ta}$$

这样，得到了下列结果，被称为切尔诺夫界。

命题5.2[切尔诺夫界]

$$P\{X \geq a\} \leq \mathrm{e}^{-ta} M(t) \qquad \text{对一切} t > 0$$

$$P\{X \leq a\} \leq \mathrm{e}^{-ta} M(t) \qquad \text{对一切} t < 0$$

由于切尔诺夫界对t 为正数或负数的情况都成立，可以通过找到使$\mathrm{e}^{-ta} M(t)$ 达到最小的t 值，来获得$P\{X \geq a\}$ 的最佳上界。