大数据的统计学基础——第7周

法律声明

【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。

关注炼数成金企业微信

■提供全面的数据资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

余额宝与大数定律

◆ http://finance.sina.com.cn/money/bank/hykx/20140128/074718105416.shtml

大数定律告诉你：余额宝单个账户设限100万的秘密

2014年01月28日 07:47 证券时报网 我有话说(37人参与)

证券时报记者 朱凯

网络金融领头羊“余额宝”1月15日宣布，总规模突破2500亿元，它所对接的天弘基金“增利宝”因此成为国内最大的货币基金。

近日，微信与华夏基金联合推出“理财通”，亦展现出强劲的吸金之势。

值得注意的是，投资者随后发现，余额宝单个账户的资金上限为100万元，每月转入资金不能超过20万元；理财通对每个账户的资金上限也作出了完全一致的规定，即不能超过100万元。

“100万元”这样一个普通数字，为何成为了阿里、腾讯等巨头的共同选择？根据证券时报记者调查，“大数定律”应该是其最核心的因素。

“大数定律”属于概率论中的随机变量序列范畴。简单说就是某些“有规律的随机事件”大量重复出现后，往往会呈现出几乎必然的统计特性。正如当你向上抛硬币达到数百万次后，其正面与反面朝上的概率应各占一半。

业内人士认为，这些网络理财账户100万元上限的规定，其实显示了阿里、腾讯等对“大数定律”的敬畏和担心。

频率——概率

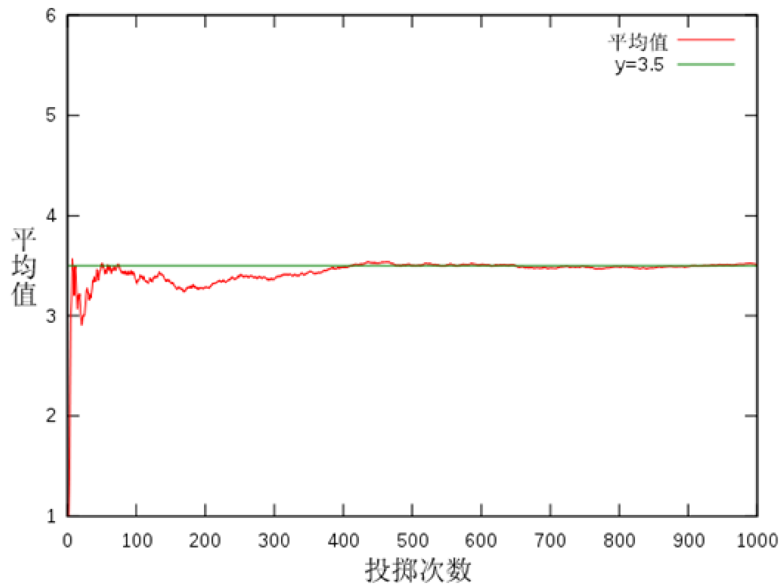
- 在相同的条件下，重复n次试验，事件A发生的次数n_A称为A发生的频数， $\frac{n_A}{n}$称为事件A发生的频率。
- 大量的试验证明，当试验的重复次数n逐渐增大时，事件A发生的频率会逐渐稳定于某个常数p。这个p就是事件A发生的概率
- 重复试验中事件的频率的稳定性，是大量随机现象的统计规律性的典型表现

实验者	n	n_H	$f_n(H)$
德摩根	2 048	1 061	0. 518 1
蒲 丰	4 040	2 048	0. 506 9
K·皮尔逊	12 000	6 019	0. 501 6
K·皮尔逊	24 000	12 012	0. 500 5

随着试验次数的增加，事件H的频率与0.5之间的差距越来越小

统计规律性

- 在随机事件的大量重复出现中，往往呈现几乎必然的规律，这类规律就是大数定律。
- 人口男女比例接近1:1
- 多次抛掷硬币，正面向上出现的频率接近1/2
- 一个精密钳工在测量一个工件时，由于具有随机误差，他总是反复测量多次，然后用各次的平均值来作为测量的结果.而且经验表明：只要测量的次数足够多，总可以达到要求的精度.

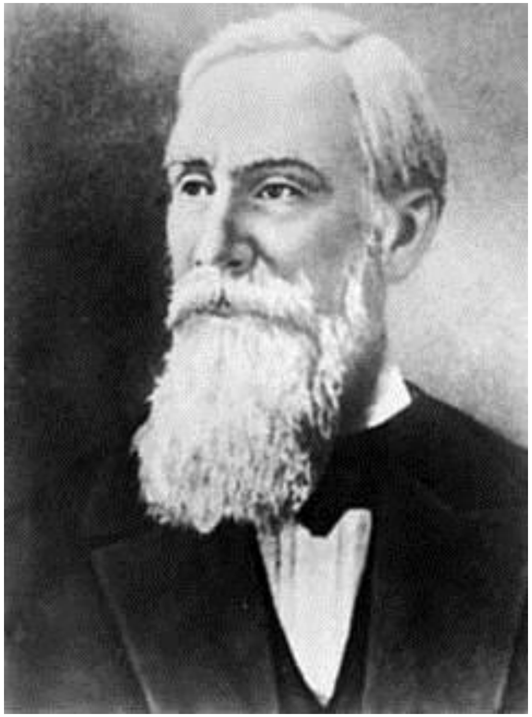


切比雪夫不等式

- 设随机变量X具有数学期望E(X)=μ，方差D(X)=σ^2，则对任意正数ε，不等式

$$P\{|X-\mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

- 都成立。



- $P\{|X-\mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$ 等价于$P\{|X-\mu| < \varepsilon\sigma\} > 1-\frac{1}{\varepsilon^2}$
- 所有数据中，至少有3/4的数据位于平均数2个标准差范围内。
- 所有数据中，至少有8/9的数据位于平均数3个标准差范围内。
- 所有数据中，至少有15/16的数据位于平均数4个标准差范围内

弱大数定律

弱大数定理(辛钦大数定理)　设 $X_1, X_2, \cdots$ 是相互独立[1],服从同一分布的随机变量序列,且具有数学期望 $E(X_k)=\mu\ (k=1,2,\cdots)$. 作前 n 个变量的算术平均 $\frac{1}{n}\sum_{k=1}^{n}X_k$, 则对于任意 $\varepsilon>0$,有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}=1. \tag{1.1}$$

证　我们只在随机变量的方差 $D(X_k)=\sigma^2\ (k=1,2,\cdots)$ 存在这一条件下证明上述结果. 因为

$$E\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n}\sum_{k=1}^{n}E(X_k)=\frac{1}{n}(n\mu)=\mu,$$

又由独立性得

$$D\left(\frac{1}{n}\sum_{k=1}^{n}X_k\right)=\frac{1}{n^2}\sum_{k=1}^{n}D(X_k)=\frac{1}{n^2}(n\sigma^2)=\frac{\sigma^2}{n},$$

由切比雪夫不等式(见第四章(2.9)式)得

$$1\geqslant P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}\geqslant 1-\frac{\sigma^2/n}{\varepsilon^2}.$$

在上式中令 $n\to\infty$,即得 $\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}=1.$

弱大数定律的意义

- 对于独立同分布且具有相同均值μ的随机变量X1,X2,……Xn，当n很大时，它们的算术平均数$\frac{1}{n}\sum_{i=1}^{n}X_i$很接近于μ。→ 可以使用样本的均值去估计总体均值。
- 例：设Xi是赌场某一台老虎机第i局的赢利，易知Xi独立同分布，且具有相同的均值μ(μ>0)。根据弱大数定律，只要n足够大，老虎机的每一局的平均赢利$\frac{1}{n}\sum_{i=1}^{n}X_i$会很接近于μ。也就是说，即使这台老虎机前面几局都赔钱了，只要不断地有人投注到这个老虎机中，最终都是会赢利的。

伯努利大数定理

伯努利大数定理 设 f_A 是 n 次独立重复试验中事件 A 发生的次数，p 是事件 A 在每次试验中发生的概率，则对于任意正数 $\varepsilon>0$，有

$$\lim_{n\to\infty} P\left\{\left|\frac{f_A}{n}-p\right|<\varepsilon\right\}=1 \tag{1.2}$$

或

$$\lim_{n\to\infty} P\left\{\left|\frac{f_A}{n}-p\right|\geqslant\varepsilon\right\}=0. \tag{1.2$'$}$$

证 因为 $f_A\sim b(n,p)$，由第四章§2例6，有

$$f_A=X_1+X_2+\cdots+X_n,$$

其中，$X_1,X_2,\cdots,X_n$ 相互独立，且都服从以 p 为参数的(0−1)分布，因而 $E(X_k)=p\ (k=1,2,\cdots,n)$，由(1.1)式即得

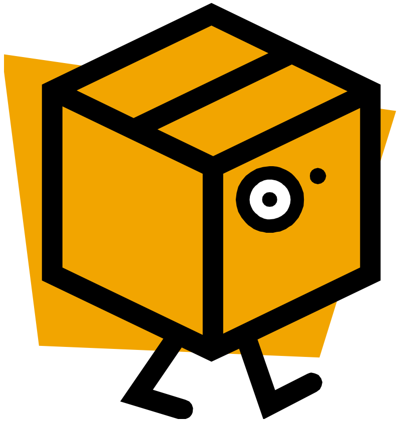
$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-p\right|<\varepsilon\right\}=1,$$

即

$$\lim_{n\to\infty} P\left\{\left|\frac{f_A}{n}-p\right|\geqslant\varepsilon\right\}=0.$$

伯努利大数定理的意义

- 伯努利大数定律的结论虽然简单,但其意义却是相当深刻的.它告诉我们当试验次数趋于无穷时,事件A发生的频率依概率收敛于A发生的概率,这样,频率接近于概率这一直观的经验就有了严格的数学意义.
- 在实际应用中，当试验次数很大时，便可以用事件的频率来代替事件的概率
- 某个箱子里装有若干个白球和红球，具体比例不知道。若从中做1000次有放回抽样，抽出红球100个，白球900个，则我们可以说抽出红球的概率是100/1000=0.1



大数定律的应用

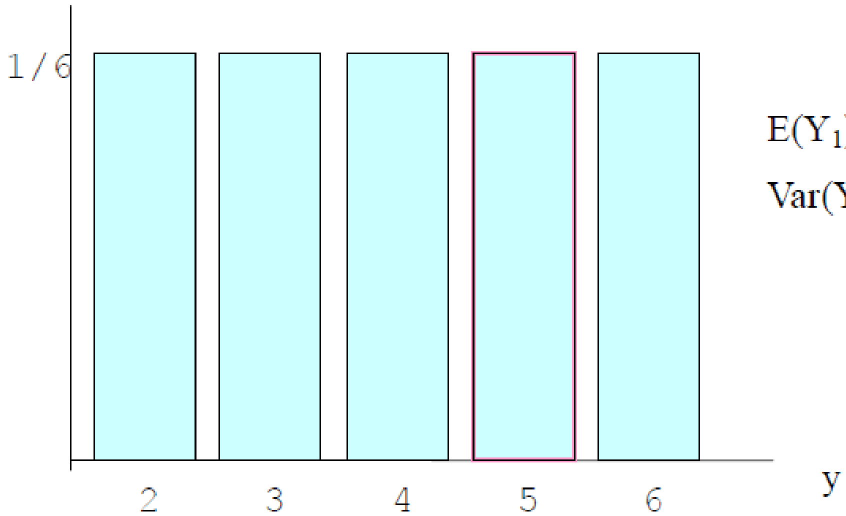
- 赌场的盈利
- 保险公司的保障
- 彩票：http://www.ycwb.com/ePaper/ycwb/html/2014-03/26/content_400215.htm?div=0

中心极限定理

- 一颗均匀的骰子连掷n次，问点数之和Yn是怎样的分布？
- 显然，Yn是n个独立同分布的随机变量之和：$Yn = X1 + X2 + \ldots\ldots + Xn$，其中Xi有着共同的分布律：

X_i	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

- 当n=1时，Y1的分布律与X1的分布律一样



$E(Y_1)=3.5$

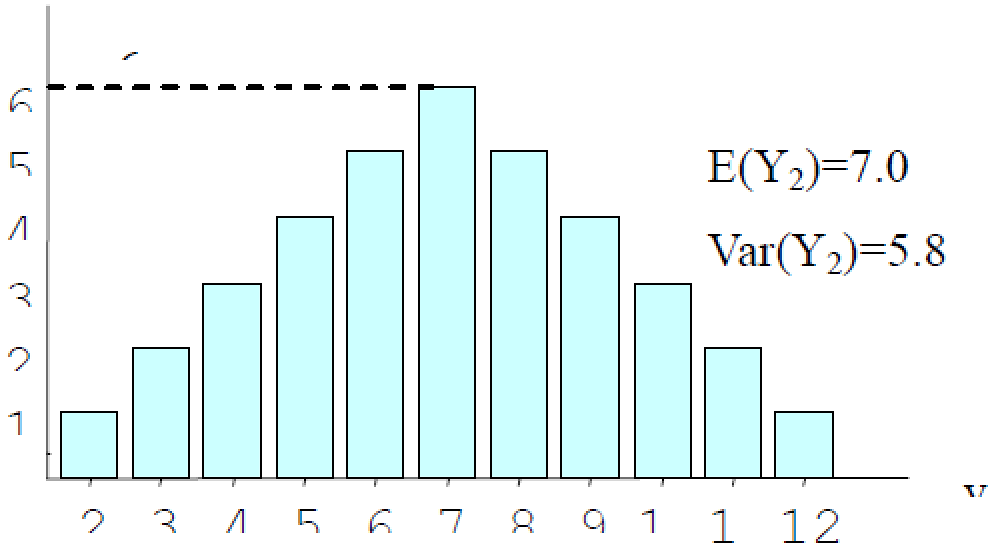
$Var(Y_2)=2.91$

中心极限定理

◆ 当n=2时，Y2的分布律如下：

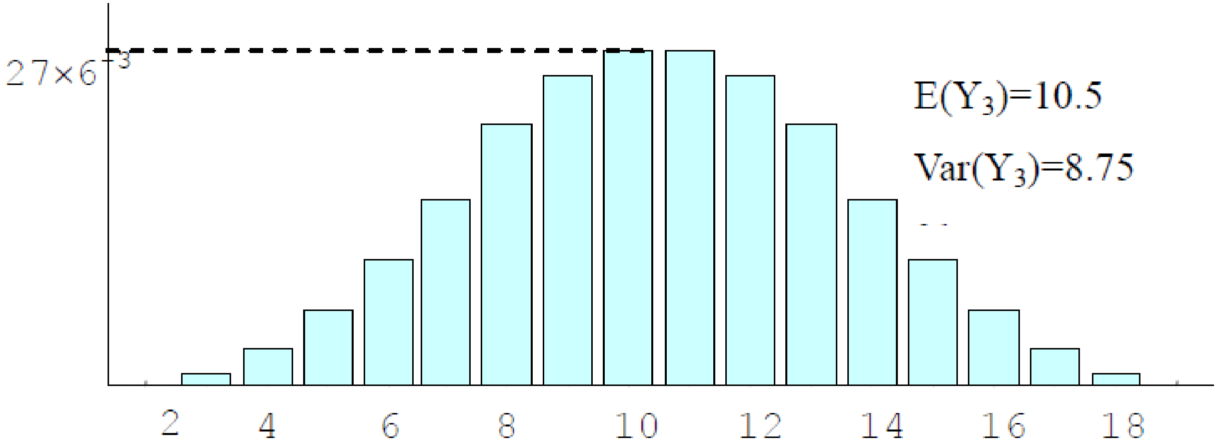
$Y_2 = X_1 + X_2$	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	8/36	5/36	6/36	5/36	8/36	3/36	2/36	1/36

◆ 这时Y2的概率直方图呈单峰对称的阶梯型

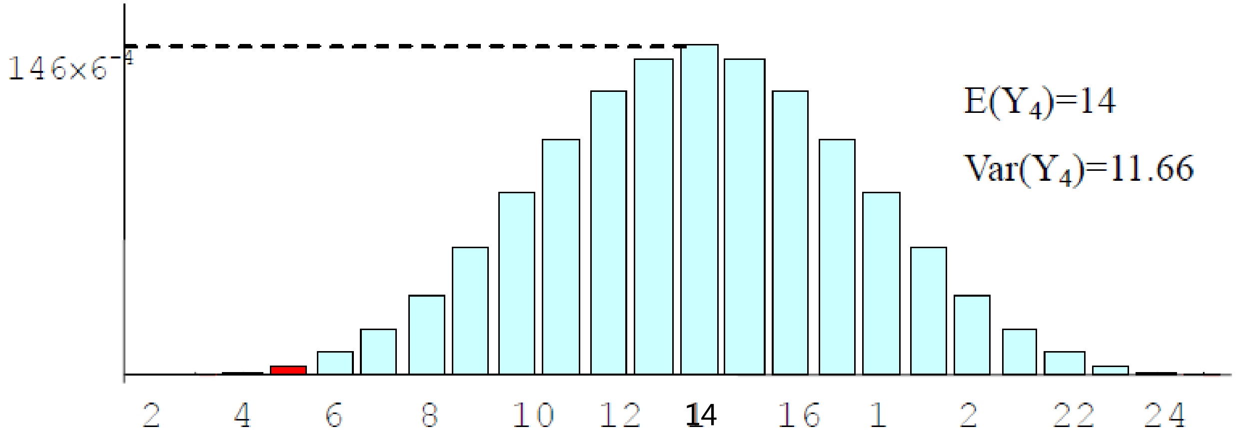


中心极限定理

- n=3时，Y3的概率直方图



- n=4时，Y4的概率直方图
-



独立同分布的中心极限定理

- *n* 个相互独立同分布的随机变量之和的分布近似于正态分布，*n* 愈大，此种近似程度愈好
- 使用严格地数学定义上述定理：

定理一（独立同分布的中心极限定理） 设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，服从同一分布，且具有数学期望和方差：$E(X_k)=\mu, D(X_k)=\sigma^2>0$ $(k=1,2,\cdots)$，则随机变量之和 $\sum_{k=1}^{n} X_k$ 的标准化变量

$$Y_n = \frac{\sum_{k=1}^{n} X_k - E\left(\sum_{k=1}^{n} X_k\right)}{\sqrt{D\left(\sum_{k=1}^{n} X_k\right)}} = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\,\sigma}$$

的分布函数 $F_n(x)$ 对于任意 x 满足

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P\left\{\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\,\sigma} \leqslant x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x). \tag{2.1}$$

定理说明

- 对于均值为μ，方差为$\sigma^2>0$的独立同分布的随机变量X1，X2,……，Xn之和$\sum_{i=1}^{n}X_i$,当n足够大时，有

$$\frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\mu}{\sigma/\sqrt{n}}\overset{\text{近似于}}{\sim}N(0,1)$$

- 一般情况下，$\sum_{i=1}^{n}X_i$的精确分布很难计算出来，但有了上述定理，我们可以求出它的近似正态分布，从而可以计算一些近似概率。

例子

- 设X1，X 2 ,……, Xn是n个独立同分布的随机变量，其共同分布为区间（0，1）上的均匀分布，即诸 Xi ~U(0,1).若取n = 100，求概率P(X1 + X 2 +……+X n ≤ 60) 的近似值。
- E(Xi)=1/2,D(Xi)=1/12；记Y=X1 + X2 +……+X n
- 根据定理，有$\frac{Y-nE(X_i)}{\sqrt{n}\sigma}$近似地服从N(0,1)
- 故P(Y ≤ 60)=P($\frac{Y-100*\frac{1}{2}}{\sqrt{100*\frac{1}{12}}} \le \frac{60-100*\frac{1}{2}}{\sqrt{100*\frac{1}{12}}}$) ≈Φ（3.464）=0.9997

例子

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随机变量，且都在区间(0,10)上服从均匀分布. 记 $V=\sum_{k=1}^{20}V_k$，求 $P\{V>105\}$ 的近似值.

解 易知 $E(V_k)=5, D(V_k)=100/12$ $(k=1,2,\cdots,20)$. 由定理一，随机变量

$$Z=\frac{\sum_{k=1}^{20}V_k-20\times 5}{\sqrt{100/12}\sqrt{20}}=\frac{V-20\times 5}{\sqrt{100/12}\sqrt{20}}$$

近似服从正态分布 $N(0,1)$，于是

$$P\{V>105\}=P\left\{\frac{V-20\times 5}{(10/\sqrt{12})\sqrt{20}}>\frac{105-20\times 5}{(10/\sqrt{12})\sqrt{20}}\right\}$$

$$=P\left\{\frac{V-100}{(10/\sqrt{12})\sqrt{20}}>0.387\right\}$$

$$=1-P\left\{\frac{V-100}{(10/\sqrt{12})\sqrt{20}}\leqslant 0.387\right\}$$

$$\approx 1-\int_{-\infty}^{0.387}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=1-\Phi(0.387)=0.348.$$

即有 $$P\{V>105\}\approx 0.348.$$ □

Lyapunov定理

定理二（李雅普诺夫(Lyapunov)定理） 设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，它们具有数学期望和方差

$$E(X_k) = \mu_k, \quad D(X_k) = \sigma_k^2 > 0, k = 1, 2, \cdots,$$

记
$$B_n^2 = \sum_{k=1}^{n} \sigma_k^2.$$

若存在正数 δ，使得当 $n \to \infty$ 时，

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} E\{\mid X_k - \mu_k \mid^{2+\delta}\} \to 0,$$

则随机变量之和 $\sum_{k=1}^{n} X_k$ 的标准化变量

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

的分布函数 $F_n(x)$ 对于任意 x，满足

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

Lyapunov定理

- 当n很大时，无论各个随机变量X_k服从什么分布，只要相互独立而且满足定理条件

  若存在正数 δ，使得当 $n \to \infty$ 时，

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} E\{|X_k - \mu_k|^{2+\delta}\} \to 0,$$

- 则它们的和$\sum_{k=1}^{n} X_k$就近似服从正态分布。
- 即 $$Z_n = \frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{B_n}$$ 近似服从标准正态分布。

- 如，在任一指定时刻，一个城市的耗电量是大量用户耗电量的总和，从而可以知道这个城市的耗电量服从正态分布。

二项分布近似正态分布

定理三(棣莫弗一拉普拉斯(De Moivre-Laplace)定理) 设随机变量 $\eta_n(n=1,2,\cdots)$服从参数为 n,p $(0<p<1)$的二项分布,则对于任意 x,有

$$\lim_{n\to\infty}P\left\{\frac{\eta_n-np}{\sqrt{np(1-p)}}\leqslant x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=\Phi(x). \qquad (2.5)$$

证 由第四章§2例6知可以将 η_n 分解成为 n 个相互独立、服从同一(0—1)分布的诸随机变量 $X_1,X_2,\cdots,X_n$ 之和,即有

$$\eta_n=\sum_{k=1}^{n}X_k,$$

其中 X_k $(k=1,2,\cdots,n)$的分布律为

$$P\{X_k=i\}=p^i(1-p)^{1-i},\quad i=0,1.$$

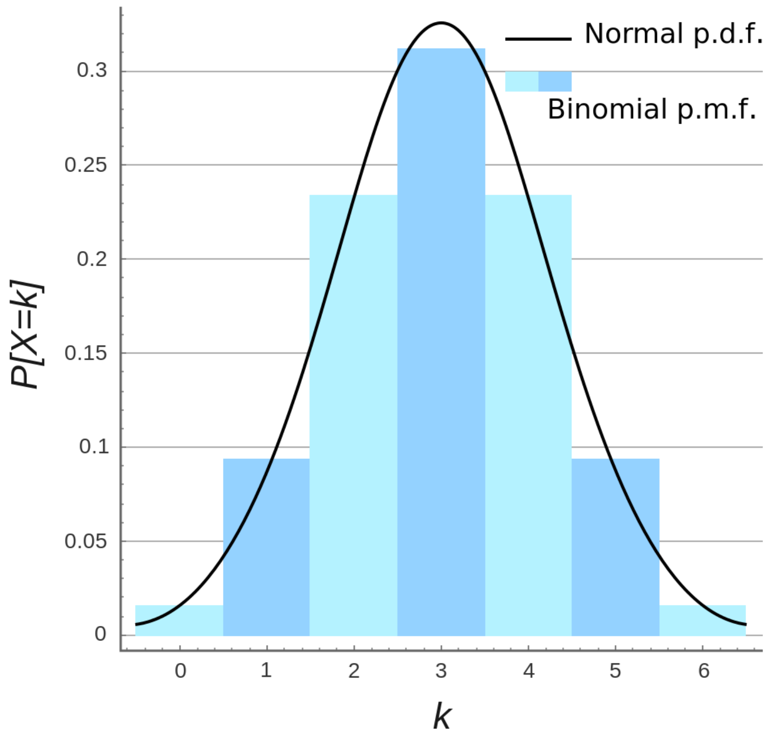
由于 $E(X_k)=p,D(X_k)=p$ $(1-p)$ $(k=1,2,\cdots,n)$,由定理一得

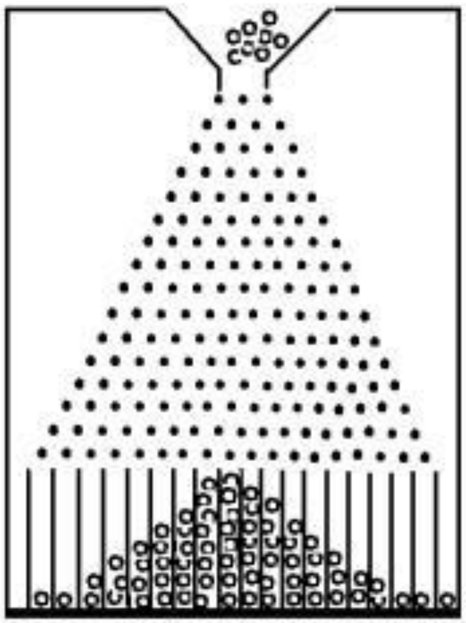
$$\lim_{n\to\infty}P\left\{\frac{\eta_n-np}{\sqrt{np(1-p)}}\leqslant x\right\}=\lim_{n\to\infty}P\left\{\frac{\sum_{k=1}^{n}X_k-np}{\sqrt{np(1-p)}}\leqslant x\right\}$$

$$=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt=\Phi(x). \qquad \square$$

这个定理表明,正态分布是二项分布的极限分布.当 n 充分大时,我们可以利用(2.5)式来计算二项分布的概率.下面举几个关于中心极限定理应用的例子.

二项分布近似正态分布





例子

- 假如某保险公司10000个同阶层的人参加人寿保险，每人每年付12元保险费，在一年内一个人死亡的概率为0.006，死亡时，其家属可向保险公司领得1000元。试问：平均每户支付赔偿金5.9元至6.1元的概率是多少？保险公司亏本的概率有多大？保险公司每年利润大于4万元的概率是多少？

- 设X_i表示保险公司支付给第i户的赔偿金，则

	0	1000
P	0.994	0.006

$$E(X_i)=6, D(X_i)=5.964 (i=1,2,\ldots 10000)$$

- 设X_i相互独立，i=1,2，……，10000.则 $\bar{X}=\frac{1}{10000}\sum_{i=1}^{10000} X_i$ 表示保险公司平均对每户的赔偿金。

$$E(\bar{X})=6, D(\bar{X})=5.964\times 10^{-4},$$

- 由中心极限定理， $\bar{X} \sim N(6, 0.0244^2)$

例子

$$P\{5.9 < \bar{X} < 6.1\} = \Phi(\frac{6.1-6}{0.0244}) - \Phi(\frac{5.9-6}{0.0244}) = 2\Phi(4.09) - 1 = 0.99996$$

- 虽然每一家的赔偿金差别很大，但保险公司平均对每户的支付计划恒等于6万元，在5.9元至6.1元内的概率接近于1，几乎是必然的。所以，对保险公司来说，只关心这个平均数。
- 保险公司亏本，也就是赔偿金额大于1000*120=12（万元），即死亡人数大于120人的概率。死亡人数为Y~B(10000,0.006),则E(Y)=60,D(Y)=59.64.由中心极限定理，Y近似服从正态分布N(60,59.64),那么
- P{Y>120}=1-P{Y<=120}=1-Φ (0.77)=0

例子

- 如果保险公司每年利润大于4万元，即赔偿人数小于80人。则

$$P\{Y<80\}=\Phi(\frac{80-60}{\sqrt{59.64}})=\Phi(2.59)=0.9952$$

- 可见，保险公司每年利润大于4万元的概率接近100%。
- 在保险市场的竞争过程中，由两个可以采用的策略，一是降低保险费3元，另一个是提高赔偿金500元，那种做法更有可能吸纳更多的投保者，哪一种效果更好？对保险公司来说，收益是一样的，而采用提高赔偿金比降低3元保险费更能吸引投保户。

总体与样本

- 普查：人口普查；考察某所高中高三学生成绩，将所有学生的成绩都统计出来……
- 抽样调查：考察某个电视节目的受欢迎程度，随机采访1000名观众；考察1000个产品的质量，从中抽取10个产品检查……
- 总体(population)——有限总体、无限总体
- 个体
- 样本(sample)
- 总体容量N
- 样本容量n

简单随机抽样样本

- 简单随机抽样：总体中每个个体被抽中的概率都相等
- 设X是具有分布函数F的随机变量，若X1,X2,……Xn是具有同一分布函数F的、相互独立的（独立同分布，也记作i.i.d）的随机变量，则称X1,X2,……Xn为从分布函数F（或总体F、或总体X）得到的容量为n的简单随机样本。它们的观察者x1，x2,……,xn称为样本值，又称为X的n个独立的观察值。
- 假设某批灯泡的寿命X（小时）服从U（3000，5000）。从这批灯泡中随机抽出10个做测试，发现这10个灯泡的寿命分别为3125，3692，4297，4172，3186，4852，3946，4286，3912，3364。
- 再从这批灯泡中抽取10个测试，它们的寿命分别为3645，4482，4617，3594，4287，3641，3289，3791，4982，4236.

统计量

- 把样本所包含的关于我们所关心的事物的信息集中起来，这便是针对不同的问题构造出样本的某种函数，这种函数在统计学中称为统计量。
- 例如：样本均值，样本方差，样本标准差
- 数学定义：
- 设X1,X2,……,Xn是来自总体的一个样本，g(X1,X2,……,Xn)是X1,X2,……,Xn的函数，若g中不含未知参数，则称g(X1,X2,……,Xn)是一个统计量
- 只利用已知的总体信息与样本信息就可以求出来的

样本均值

- 设X1,X2,……,Xn是来自总体X（E(X)=μ，D(X)= σ^2 ）的一个样本，其观察值为x1,x2,……,xn。
- 样本均值：$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 样本均值观察值： $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- 样本均值是总体均值的无偏估计量——样本均值的期望等于总体均值
- $\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$
- 一般使用样本均值估计总体均值
- 样本均值的方差： $\mathrm{D}(\bar{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}$

样本方差

- 设X1,X2,……,Xn是来自总体X的一个样本，其观察值为x1,x2,……,xn。
- 样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2)$
- 样本方差观察值：$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2)$
- 样本方差是总体方差的无偏估计量
- 假设总体的方差为σ^2
- $E\left[\frac{1}{n-1}(\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2)\right] = \frac{1}{n-1}\sum_{i=1}^{n} E({X_i}^2) - \frac{n}{n-1}E(\bar{X}^2) = \frac{1}{n-1}\sum_{i=1}^{n}\{D(X_i) + [E(X_i)]^2\} -$

  $\frac{n}{n-1}\{D(\bar{X}) + [E(\bar{X})]^2\} = \frac{1}{n-1}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2$

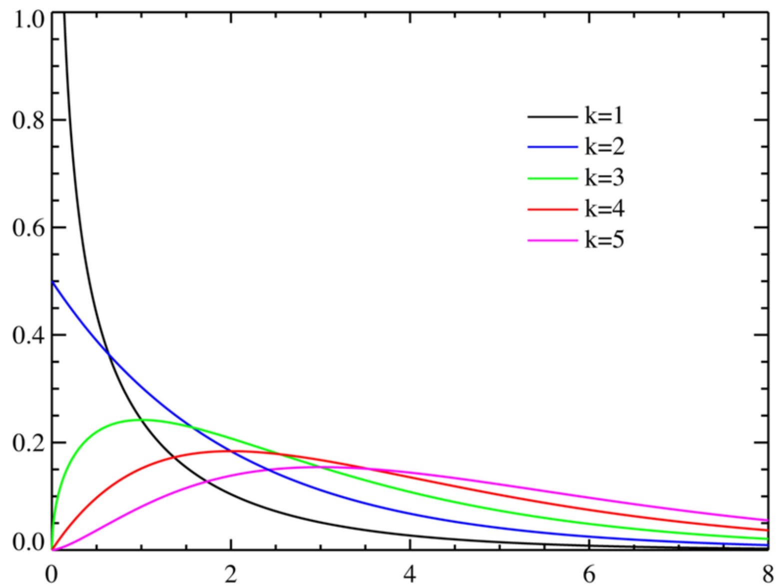
χ^2分布

- 设X_1，X_2，……，X_n是来自总体N(0,1)的样本，则称统计量

$$\chi^2 = {X_1}^2 + {X_2}^2 + \cdots \ldots + {X_n}^2$$

- 服从自由度为n的χ^2分布，记为$\chi^2 \sim \chi^2(\mathrm{n})$

- 卡方分布的概率密度函数为$f(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-y/2}, y > 0 \\ 0, \text{其他} \end{cases}$



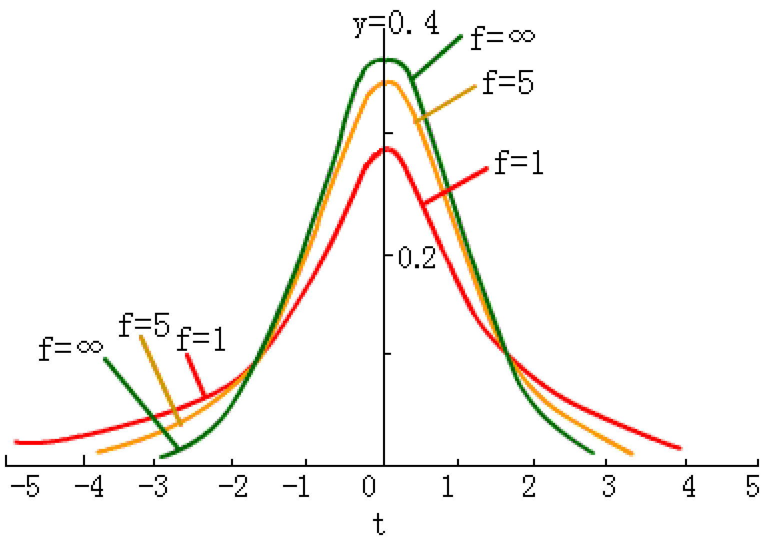
χ^2分布

- 可加性：
- 设$\chi_1{}^2 \sim \chi^2(n_1)$, $\chi_2{}^2 \sim \chi^2(n_2)$,并且$\chi_1{}^2$与$\chi_2{}^2$相互独立，则有$\chi_1{}^2 + \chi_2{}^2 \sim \chi^2(n_1 + n_2)$
- 数学期望与方差：
- 若$\chi^2 \sim \chi^2(\mathrm{n})$，则$E(\chi^2) = E\left(\sum_{i=1}^n X_i{}^2\right) = \sum_{i=1}^n E(X_i{}^2) = \sum_{i=1}^n [D(X_i) + [E(X_i)]^2] = n$
- $D\left(X_i{}^2\right) = E\left(X_i{}^4\right) - \left[E\left(X_i{}^2\right)\right]^2 = 3 - 1 = 2$
- 故$D(\chi^2) = D\left(\sum_{i=1}^n X_i{}^2\right) = \sum_{i=1}^n D(X_i{}^2) = 2n$

t分布——student 分布

- 设X~N(0,1),Y~χ^2(n),且X,Y相互独立，则称随机变量$t=\frac{X}{\sqrt{Y/n}}$服从自由度为n的t分布。记为t~t(n).

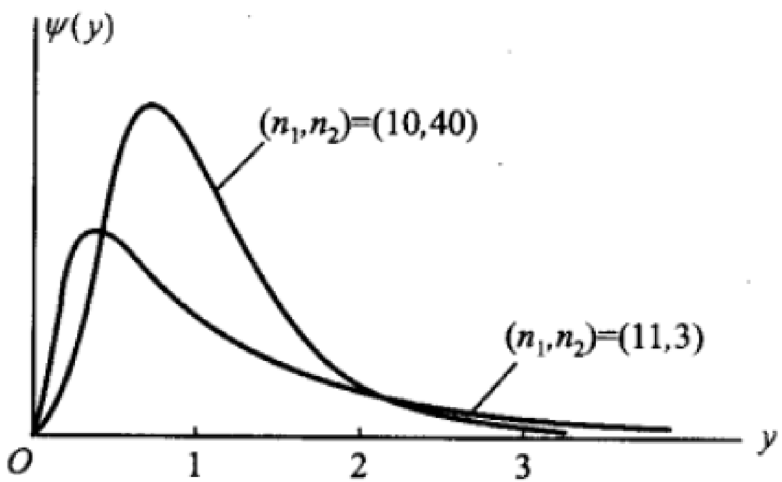
- t分布的概率密度函数：$h(t)=\frac{\Gamma\left[\frac{n+1}{2}\right]}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}$,$-\infty<t<\infty$



F分布

- 设U~$\chi^2(n_1)$，V~$\chi^2(n_2)$，且U,V相互独立，则称随机变量$F=\frac{U/n_1}{V/n_2}$服从自由度为(n_1,n_2)的F分布。记F~ F(n_1,n_2)
- F分布的概率密度函数：

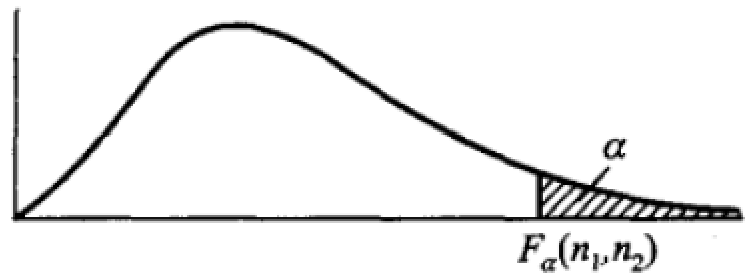
$$\psi(y)=\begin{cases}\dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}y^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1y/n_2)]^{(n_1+n_2)/2}}, & y>0,\\ 0, & \text{其他}.\end{cases}$$



F分布的分位点

- 对于F分布上的α分位点，有

$$F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}$$

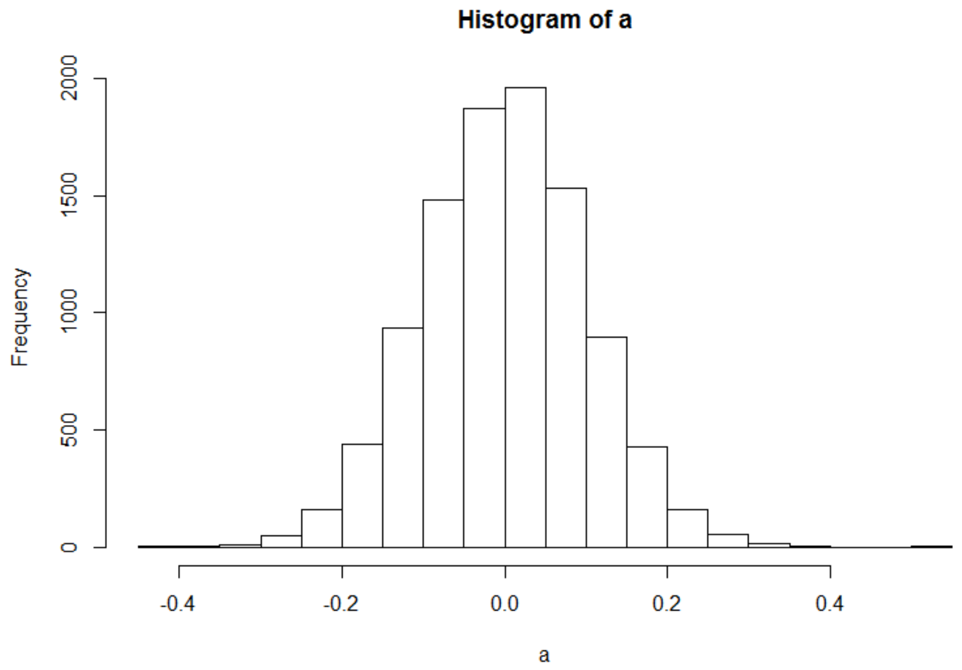


- 例：$F_{0.95}(12,9) = \frac{1}{F_{0.05}(9,12)} = \frac{1}{2.80} = 0.357$

抽样分布

◆ 10000个正态分布的样本均值的分布

```
a=c();
for( i in 1:10000){ a[i]=mean(rnorm(100))}
hist(a,breaks=10)
```



正态总体的样本分布

- 设X1，X2，……，Xn是来自正态总体N(μ，σ^2)，$\bar{X}$是样本均值， S^2是样本方差，则有
- (1)$\bar{X} \sim N(\mu，\sigma^2/n)$
- (2)$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- (3) $\bar{X}$与S^2相互独立。
- (4)$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

两个正态总体的样本均值与样本方差

- 设X1，X2，……Xn1与Y1,Y2,……，Yn2分别来自正态总体$N(\mu_1,\sigma_1{}^2)$，$N(\mu_2,\sigma_2{}^2)$的样本，且这两个样本相互独立。其样本均值分别为$\bar{X}$，$\bar{Y}$，其方差分别为$S_1{}^2$, $S_2{}^2$,则有
- (1)$\frac{S_1{}^2/S_2{}^2}{\sigma_1{}^2/\sigma_2{}^2}\sim F(n_1-1,n_2-1)$
- (2)当$\sigma_1{}^2=\sigma_2{}^2=\sigma^2$时，有

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim \mathrm{t}(n_1+n_2-2)$$

其中，$S_w{}^2=\frac{(n_1-1)S_1{}^2+(n_2-1)S_2{}^2}{n_1+n_2-2}$

炼数成金逆向收费式网络课程

- 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