

第五章

大数定律及中心极限定理



在第一章中我们曾经提到过事件发生的频率具有稳定性，即随着试验次数的增加，事件发生的频率逐渐稳定于某个常数。在实践中，人们还认识到，测量值的算术平均值具有稳定性，这种稳定性就是我们要讨论的大数定律的背景。

正态分布是概率论中的一个重要的分布，它有着广泛的应用。中心极限定理将阐明，原本不是正态分布的一般随机变量和的分布，在一定条件下可以渐进服从正态分布。



本章要解决的问题

答复

1. 为何能以某事件发生的频率作为该事件的 概率的估计?
2. 为何能以样本均值作为总体期望的估计?
3. 为何正态分布在概率论中占有极其重要的地位?
4. 大样本统计推断的理论基础是什么?

大数
定律

中心极
限定理

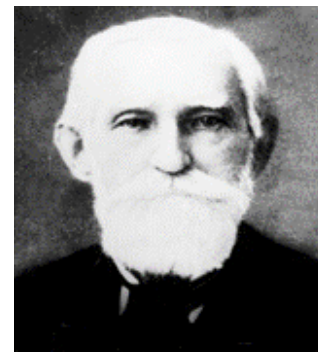


§5.1 大数定律

● 切贝雪夫(chebyshev)不等式

设随机变量 X 的方差 $D(X)$ 存在, 则对于任意实数 $\varepsilon > 0$,

$$P\{|X - E(X)| \geq \varepsilon\} \leq \frac{D(X)}{\varepsilon^2}$$



切比雪夫, П. Л.

切比雪夫

说明: 1、粗略估计 \mathbf{X} 与均值之间范围的关系,

2、实际中很少用

切贝雪夫不等式等价于

$$P\{|X - E(X)| < \varepsilon\} \geq 1 - \frac{D(X)}{\varepsilon^2}$$

上式表明, 随机变量 \mathbf{X} 的方差越小, 事件 $\{|\mathbf{X} - E(\mathbf{X})| < \varepsilon\}$ 发生的概率越大, 即 \mathbf{X} 的取值基本上集中在它的数学期望附近。由此可见, 方差刻画了随机变量取值的分散程度。



切比雪夫 (Chebyshev) 不等式的应用

例 已知随机变量 X 的数学期望 $E(X)=\mu$, 方差 $D(X)=\sigma^2$, 当 $\varepsilon=2\sigma$ 和 $\varepsilon=3\sigma$ 时, 用切比雪夫不等式求 $P\{|X-\mu|<\varepsilon\}$ 的值至少是多少?

解 利用切比雪夫不等式, 当 $\varepsilon=2\sigma$ 和 $\varepsilon=3\sigma$ 时, 分别有

$$P\{|X-\mu|<2\sigma\} \geq 1 - \frac{\sigma^2}{(2\sigma)^2} = \frac{3}{4} = 0.75$$

$$P\{|X-\mu|<3\sigma\} \geq 1 - \frac{\sigma^2}{(3\sigma)^2} = \frac{8}{9} \approx 0.8889$$

从上例可以看出, 当随机变量的分布未知时, 利用它的数学期望和方差可以知道 $P\{|X-\mu|<\varepsilon\}$ 的值至少是多少, 从而可以粗略地估计某些事件发生的概率。



从上例可以看出，当随机变量的分布未知时，利用它的数学期望和方差可以知道 $P\{|X - \mu| < \varepsilon\}$ 的值至少是多少，从而可以粗略地估计某些事件发生的概率。

但是如果已知随机变量 $X \sim N(\mu, \sigma^2)$ ，由P41可知

$$P\{|X - \mu| < 2\sigma\} = 0.9544$$

$$P\{|X - \mu| < 3\sigma\} = 0.9974$$





例1 设有一大批种子，其中良种占1/6. 试估计在任选的 6000 粒种子中，良种所占比例与1/6 比较上下小于1%的概率.

解 设 X 表示 6000 粒种子中的良种数，

$$X \sim B(6000, 1/6)$$

$$E(X) = 1000, D(X) = \frac{5000}{6}$$

$$P\left\{\left|\frac{X}{6000} - \frac{1}{6}\right| < 0.01\right\} = P\{|X - 1000| < 60\} \geq 1 - \frac{\frac{5000}{6}}{60^2} = \frac{83}{108} = 0.7685$$


$$P(|X - E(X)| < \varepsilon) \geq 1 - \frac{D(X)}{\varepsilon^2}$$


$$X \sim B(6000, 1/6)$$

$$P\left\{\left|\frac{X}{6000} - \frac{1}{6}\right| < 0.01\right\} = P\{|X - 1000| < 60\} \geq 1 - \frac{5000}{60^2} = \frac{83}{108} = 0.7685$$

实际精确计算

$$\begin{aligned} P\left(\left|\frac{X}{6000} - \frac{1}{6}\right| < 0.01\right) &= P(940 < X < 1060) \\ &= \sum_{k=941}^{1059} C_{6000}^k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{6000-k} = 0.959036 \end{aligned}$$



大数定律

定义

设 $Y_1, Y_2, \dots, Y_n, \dots$ 是一系列随机变量, a 是一常数,

若 $\forall \varepsilon > 0$ 有 $\lim_{n \rightarrow \infty} P\{|Y_n - a| < \varepsilon\} = 1$

则称随机变量序列

$Y_1, Y_2, \dots, Y_n, \dots$ 依概率收敛于 a ,

记作

$$Y_n \xrightarrow[n \rightarrow \infty]{P} a$$



辛钦大数定律

设随机变量序列 $X_1, X_2, \dots, X_n, \dots$ 相互独立, 服从同一分布, 且具有数学期望

$$E(X_k) = \mu, \quad k = 1, 2, \dots$$

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

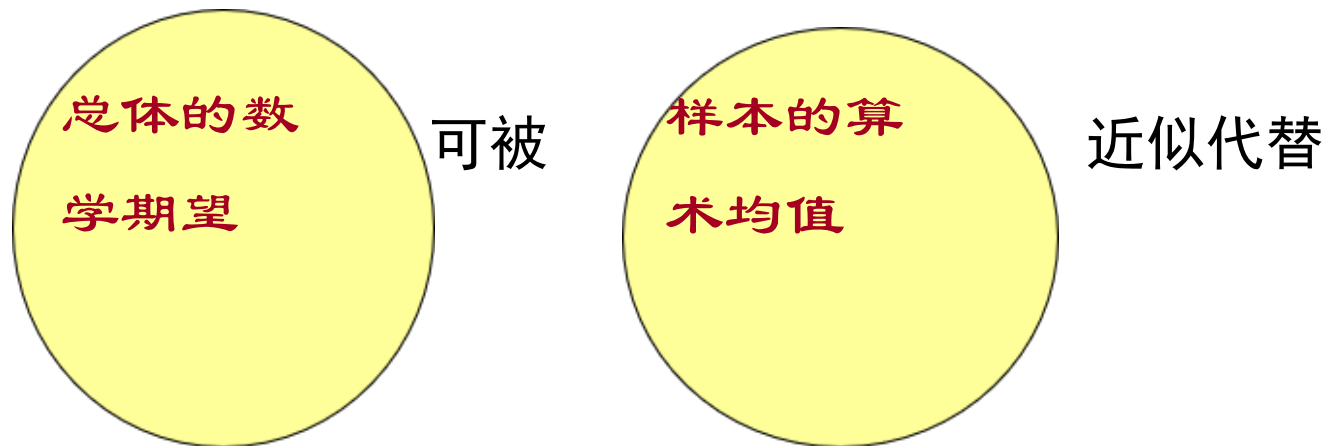


辛钦



定理的意义

具有相同数学期望和方差的独立 r. v. 序列的算术平均值依概率收敛于数学期望. 当 n 足够大时, 算术平均值几乎是一常数.



● 大数定律

贝努里 (Bernoulli) 大数定律

设 n_A 是 n 次独立重复试验中事件 A 发生的次数,
 p 是每次试验中 A 发生的概率, 则

$$\forall \varepsilon > 0 \quad \text{有} \quad \lim_{n \rightarrow \infty} P \left\{ \left| \frac{n_A}{n} - p \right| < \varepsilon \right\} = 1$$

故

$$\frac{n_A}{n} \xrightarrow[n \rightarrow \infty]{P} p$$



雅各布第一·伯努利

贝努里



贝努里(Bernoulli)大数定律的意义

在概率的统计定义中,事件 A 发生的频率

$$\frac{n_A}{n}$$

依概率收敛于事件 A 发生的概率。

它揭示了“事件发生的频率具有稳定性”。

因此,在实际问题的应用中,当试验次数很大时,
可以用频率近似代替概率是合理的。



思考题

电视台需作节目A收视率的调查.

每天在播电视的同时,随机地向当地居民打电话询问是否在看电视.若在看电视,再问是否在看节目A.设回答看电视的居民户数为 n .若要保证以95%的概率使调查误差在10%之内, n 应取多大?



每晚节目A播出一小时,调查需同时进行,设每小时每人能调查20户,每户居民每晚看电视的概率为70%,电视台需安排多少人作调查.又,若使调查误差在1%之内, n 应取多大?



§ 5.2 中心极限定理

定理一

林德伯格-列维中心极限定理
(Lindberg-levi)

[独立同分布的中心极限定理]

定理二

棣莫弗-拉普拉斯中心极限定理
(De Moivre-Laplace)

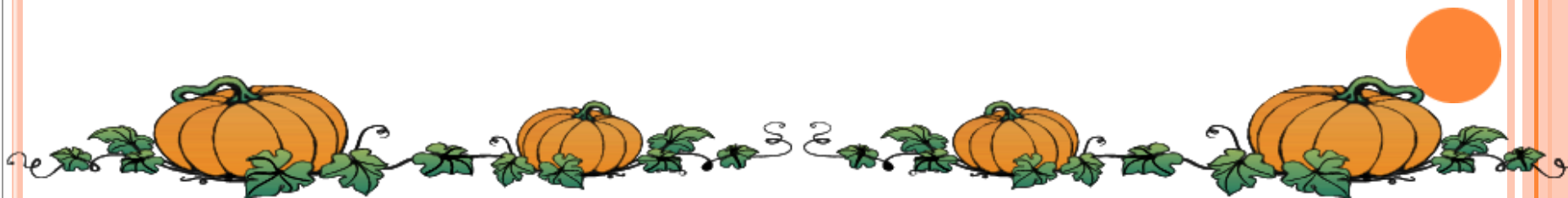
[二项分布以正态分布为极限分布]



中心极限定理 (*Central limit theorem*)

客观背景：客观实际中，许多随机变量是由大量相互独立的偶然因素的综合影响所形成，每一个微小因素，在总的影响中所起的作用是很小的，但总起来，却对总和有显著影响，这种随机变量往往近似地服从正态分布。

研究在什么条件下，大量独立随机变量和的分布以正态分布为极限，这一类定理称为中心极限定理。



定理 1 独立同分布的中心极限定理

设随机变量序列 $X_1, X_2, \dots, X_n, \dots$ 独立同分布,

且有期望和方差:

$$E(X_k) = \mu, D(X_k) = \sigma^2 > 0, k = 1, 2, \dots$$

则对于任意实数 x ,

$$\lim_{n \rightarrow \infty} P\left\{\frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma} \leq x\right\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

注

记 $Y_n = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma}$ 则 Y_n 为 $\sum_{k=1}^n X_k$ 的标准化随机变量.

$$\lim_{n \rightarrow \infty} P\{Y_n \leq x\} = \Phi(x)$$



$$\lim_{n \rightarrow \infty} P\{Y_n \leq x\} = \Phi(x)$$

$$Y_n = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma}$$

即 n 足够大时, Y_n 的分布函数近似于标准正态随机变量的分布函数

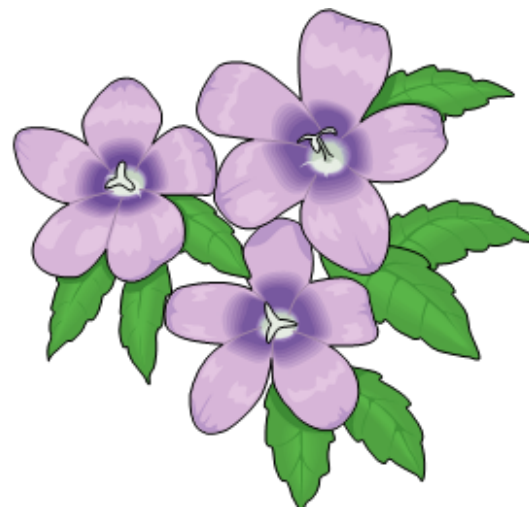
$$\sum_{k=1}^n X_k = \sqrt{n}\sigma Y_n + n\mu \quad Y_n \sim N(0,1) \quad \text{近似服从} \quad N(n\mu, n\sigma^2)$$

在第二章曾讲过有许多随机现象服从正态分布是由于许多彼此没有什么相依关系、对随机现象谁也不能起突出影响, 而均匀地起到微小作用的随机因素共同作用(即这些因素的叠加)的结果. 若联系于此随机现象的随机变量为 X , 则它可被看成为许多相互独立的起微小作用的因素 X_k 的总和, 而这个总和服从或近似服从正态分布.



定理的应用：对于独立的随机变量序列 $\{X_n\}$ ，不管 $X_i (i=1, 2, \dots, n)$ 服从什么分布，只要它们是独立同分布，且数学期望和方差分别为 μ 和 σ^2 ，那么，当 n 充分大时，这些随机变量之和

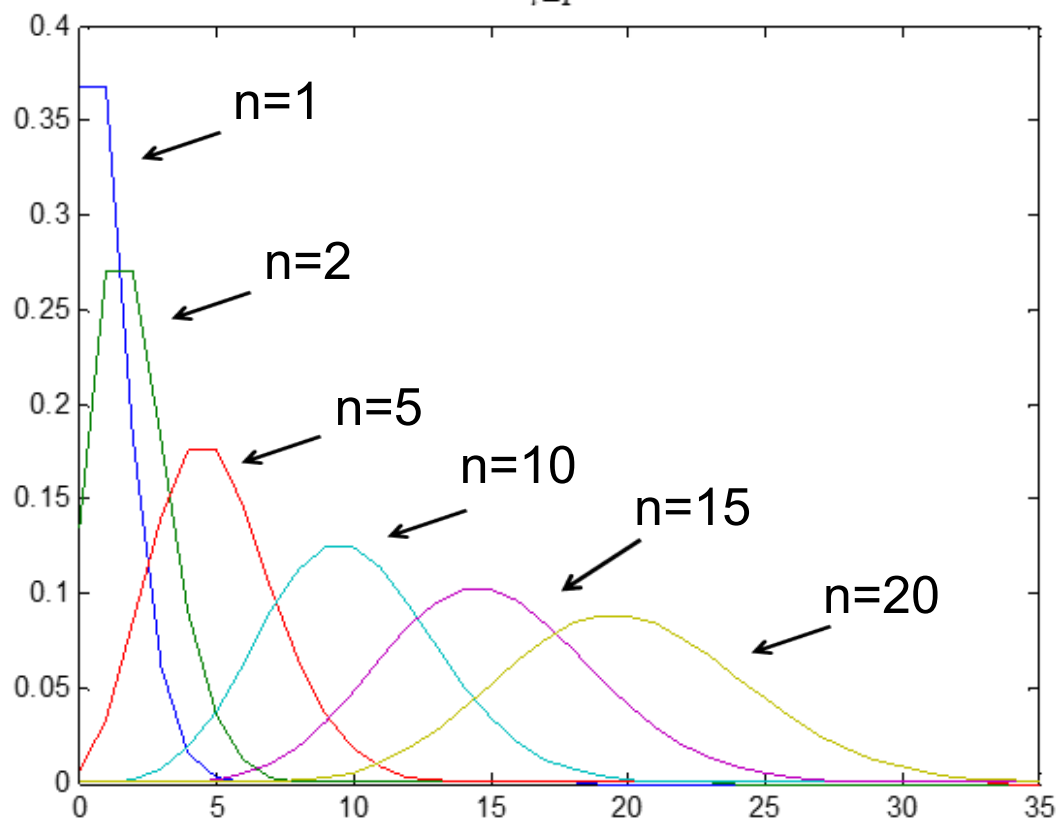
$$\sum_{i=1}^n X_i \overset{\text{近似}}{\sim} N(n\mu, n\sigma^2)$$



例、设 $\{X_i\}$ 是一些相互独立同分布的随机变量，且他们都服从参数为 λ 的泊松分布，可以证明

$$\sum_{i=1}^n X_i \sim P(n\lambda)$$

当 $\lambda=1$ 时,随着 n 的增加, $\sum_{i=1}^n X_i$ 的图像将如何变化?



例 在一个超市中，结账柜台为顾客服务的时间（以min计）是相互独立的随机变量且服从相同的分布，均值为1.5，方差为1。（1）求对100位顾客服务时间不超过2h的概率。（2）要求总的时间不超过1h的概率大于0.95，问至少能为多少位顾客服务？

解 （1）设 $X_i(i=1,2,\cdots,100)$ 表示对第i位顾客的服务时间。根据题意 $X_1, X_2, \cdots, X_{100}$ 相互独立的随机变量且服从相同的分布

$$E(X_i) = 1.5, D(X_i) = 1$$

设 X 表示对100位顾客服务的时间，则

$$X = \sum_{k=1}^{100} X_k, E(X) = 150, D(X) = 100,$$

由独立同分布中心极限定理，有

$$X \overset{\text{近似}}{\sim} N(150, 100)$$

$$P\left\{\sum_{i=1}^{100} X_i \leq 120\right\} = P\{X \leq 120\} \approx \Phi\left(\frac{120-150}{10}\right)$$



$$P\left\{\sum_{i=1}^{100} X_i \leq 120\right\} = P\{X \leq 120\} \approx \Phi\left(\frac{120-150}{10}\right) \\ = \Phi(-3) = 1 - \Phi(3) = 0.0044$$

(2) 设1小时内能对N为顾客服务，并设 $X_i (i=1,2,\dots,N)$ 表示对第i位顾客的服务时间，根据题意，要确定最大

的N，使 $P\left\{\sum_{i=1}^N X_i \leq 60\right\} > 0.95$

则 $X = \sum_{i=1}^N X_i \sim N(1.5 \times N, 1 \times N)$

$$P\left\{\sum_{i=1}^N X_i \leq 60\right\} = P\{X \leq 60\} \approx \Phi\left\{\frac{60 - N \times 1.5}{\sqrt{N} \times 1}\right\} > 0.95$$

查表的 $\frac{60 - N \times 1.5}{\sqrt{N} \times 1} > 1.645 \quad \therefore N < 33.64 \quad \therefore N = 33$

例 炮火轰击敌方防御工事 **100** 次, 每次轰击命中的炮弹数服从同一分布, 其数学期望为 **2**, 均方差为 **1.5**. 若各次轰击命中的炮弹数是相互独立的, 求 **100** 次轰击中

- (1) 至少命中 **180** 发炮弹的概率;
- (2) 命中的炮弹数不到 **200** 发的概率.

解 设 X_k 表示第 k 次轰击命中的炮弹数

$$E(X_k) = 2, \quad D(X_k) = 1.5^2, \quad k = 1, 2, \dots, 100$$

$$X_1, X_2, \dots, X_{100} \quad \text{相互独立,}$$

设 X 表示 **100** 次轰击命中的炮弹数, 则

$$X = \sum_{k=1}^{100} X_k, \quad E(X) = 200, \quad D(X) = 225,$$



设 X 表示100次轰击命中的炮弹数, 则

$$X = \sum_{k=1}^{100} X_k, \quad E(X) = 200, \quad D(X) = 225,$$

由独立同分布中心极限定理, 有

$$X \overset{\text{近似}}{\sim} N(200, 225)$$

$$\begin{aligned} (1) \quad P\{X \geq 180\} &\approx 1 - \Phi\left(\frac{180 - 200}{15}\right) \\ &\approx 1 - \Phi(-1.33) = \Phi(1.33) = 0.9082 \end{aligned}$$

$$\begin{aligned} (2) \quad P\{0 \leq X < 200\} &\approx \Phi\left(\frac{200 - 200}{15}\right) - \Phi\left(\frac{0 - 200}{15}\right) \\ &= \Phi(0) - \Phi(-13.33) \approx 0.5 \end{aligned}$$



例3 检验员逐个检查某产品,每查一个需用10秒钟.但有的产品需重复检查一次,再用去10秒钟.若产品需重复检查的概率为 0.5,求检验员在 8 小时内检查的产品多于 1900个的概率.

解 若在 8 小时内检查的产品多于1900个,即检查1900个产品所用的时间小于 8 小时.

设 X 为检查1900 个产品所用的时间(秒)

设 X_k 为检查第 k 个产品所用的时间(单位: 秒), $k = 1, 2, \dots, 1900$

X_k	10	20
P	0.5	0.5

$$E(X_k) = 15, \quad D(X_k) = 25$$

$$X_1, X_2, \dots, X_{1900} \text{ 相互独立同分布, } X = \sum_{k=1}^{1900} X_k$$

$$E(X) = 1900 \times 15 = 28500, \quad D(X) = 1900 \times 25 = 47500$$

$$X \overset{\text{近似}}{\sim} N(28500, 47500)$$

$$P\{10 \times 1900 \leq X \leq 3600 \times 8\} = P\{19000 \leq X \leq 28800\}$$

$$\approx \Phi\left(\frac{28800 - 28500}{\sqrt{47500}}\right) - \Phi\left(\frac{19000 - 28500}{\sqrt{47500}}\right)$$

$$\approx \Phi(1.376) - \Phi(-43.589)$$

$$\approx 0.9162$$



定理2

棣莫佛—拉普拉斯中心极限定理

(DeMoivre-Laplace)

设 $Y_n \sim b(n, p)$, $0 < p < 1$, $n = 1, 2, \dots$ 则对任一实

数 x , 有

$$\lim_{n \rightarrow \infty} P\left\{\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x\right\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

二项分布的极限分布是正态分布

说明：当 n 很大， p 不是很小时，不能再用泊松定理，用中心极限定理。



设 $Y_n \sim b(n, p)$, $0 < p < 1$, $n = 1, 2, \dots$ 则对任一实

数 x , 有

$$\lim_{n \rightarrow \infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt = \Phi(x)$$

\therefore 当 n 很大时, 二项分布的标准化变量 $\frac{Y_n - np}{\sqrt{np(1-p)}} \overset{\text{近似}}{\sim} N(0, 1)$

即 $Y_n \sim N(np, np(1-p))$ (近似)

一般地, 如果 $Y_n \sim b(n, p)$, 则

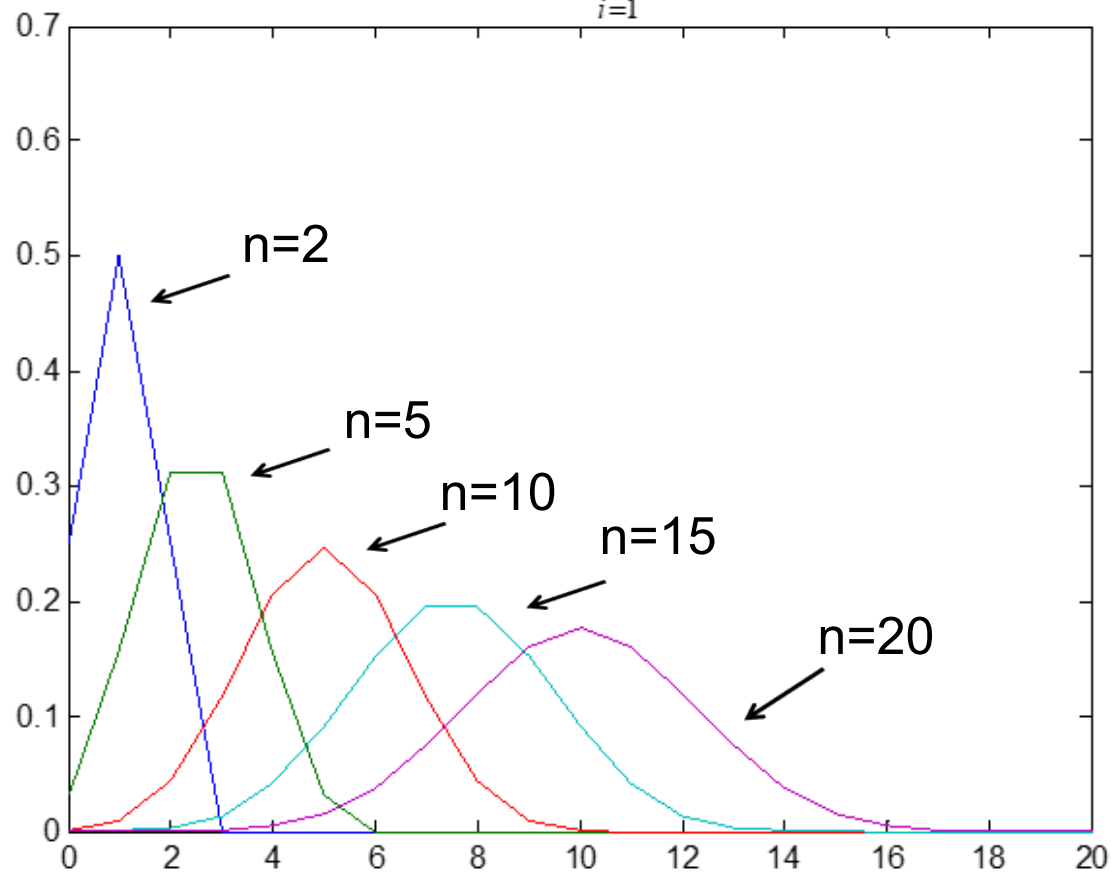
$$P\{a < Y_n \leq b\} \approx \Phi\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - np}{\sqrt{np(1-p)}}\right)$$



例、设 $\{X_i\}$ 是一些相互独立同分布的随机变量，且他们都服从0-1分布的 $b(1, p)$ ，我们知道

$$\sum_{i=1}^n X_i \sim b(n, p)$$

当 $p=0.5$ 时，随着 n 的增加， $\sum_{i=1}^n X_i$ 的图像将如何变化？



例 设在某保险公司的索赔户中，因被盗索赔者占**20%**，求在**200**个索赔户中因被盗而索赔的户数在**25**到**55**的概率。

解、 用**X**表示**200**个索赔户中因被盗而索赔的户数, 则

$$X \sim b(200, 0.2), \quad E(X)=40, D(X)=32=5.66^2,$$

由德莫佛—拉普拉斯中心极限定理, 有

近似

$$X \sim N(40, 32)$$

所求概率为

$$\begin{aligned} P\{25 \leq X \leq 55\} &\approx \Phi\left(\frac{55-40}{\sqrt{32}}\right) - \Phi\left(\frac{25-40}{\sqrt{32}}\right) \\ &= \Phi(2.65) - \Phi(-2.65) \\ &\approx 2\Phi(2.65) - 1 \approx 0.9920 \end{aligned}$$



例 一船舶在某海域内航行，已知每遭受一次波浪的冲击，纵摇角大于 3° 的概率为 $p=1/3$ ，若船舶遭受了**90000**次波浪冲击，问其中有**29500~30500**次纵摇角度大于 3° 的概率是多少？

例 对于一个学生而言，来参加家长会的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，设每个学生参加会议的家长人数相互独立，且服从同一分布。

(1) 求参加会议的人数超过**450**的概率；

(2) 求有**1**名家长来参加会议的学生人数不多于**340**的概率。





有一大批建筑房屋用的木柱，其中**80%**的长度不小于**3**米，现从这批木材中任取**100**根，试求其中至少有**30**根短于**3**米的概率。

解 设**100**根木材中长度短于**3**米的根数为**X**，则

$$X \sim b(100, 0.2)$$

$$\text{则 } E(X) = 20, D(X) = 16$$

由德莫佛—拉普拉斯中心极限定理，有

近似

$$X \sim N(20, 16)$$

所求概率为

$$P\{X \geq 30\} \\ \approx 1 - \Phi\left\{\frac{30 - 20}{\sqrt{16}}\right\} = 1 - \Phi(2.5) = 0.0062$$

