第2章

概　率

与

概率分布

第2章 概率与概率分布

2.1 概率的基本概念

问题的提出

- 前两章对总体和样本的基本概念以及样本数据的处理方法做了一般介绍。样本仅仅是总体的一部分，统计学的目的不在于研究样本，而是通过样本去推断总体。
- 从同一总体中随机抽取样本，各次所得的样本不会完全相同。用不同的样本去推断同一总体将得出不同的结论。这些结论不可能都是正确的。用某个样本推断总体时，推断错误的可能性有多大？置信度有多高？这是对总体进行推断时发须回答的基本问题。这了回答这些问题首先对总体的分布有所了解。总体分布是建立在概率(probability)这一概念之上的，因此在研究总体分布之前首先对概率的基本知识有所了解。

2.1 概率的基本概念

必然现象与随机现象

在自然界与生产实践和科学试验中，人们会观察到各种各样的现象，把它们归纳起来，大体上分为两大类：

- 一类是可预言其结果的，即在保持条件不变的情况下，重复进行观察，其结果总是确定的，必然发生（或必然不发生）。这类现象称为**必然现象**或**确定性现象**。
- 另一类是事前不可预言其结果的，即在保持条件不变的情况下，重复进行观察，其结果未必相同。这种在一次或少数几次试验中其结果呈现偶然性、不确定性现象，称为**随机现象**或**不确定性现象**。

2.1 概率的基本概念

随机现象不存在简单的因果关系。支配这些现象出现的因素很多，各因素起所起的作用不尽相同，作用的程度也不一样。对于一个个体来说，这些因素的配合方式是偶然的，或者说是随机的。

随机现象似乎是无规律的，但也并非不可以认识，当对某一随机现象做了大量的研究之后，就能从其偶然性中提示出内在的规律。

研究偶然现象本身规律性的科学就是概率论。基于实际观测结果，利用概率论得出规律，提示偶然性中所寄寓的必然性的科学就是统计学。概率论与统计学是研究随机现象规律性的科学，概率论是统计学的基础，而统计学是根据概率论得出的规律在各领域中的实际应用。

2.1 概率的基本概念

随机现象或不确定性现象，有如下特点：

在一定的条件实现时，有多种可能的结果发生，事前人们不能预言将出现哪种结果；对一次或少数几次观察或试验而言，其结果呈现偶然性、不确定性；

但在相同条件下进行大量重复试验时，其试验结果却呈现出某种固有的特定的规律性——**频率的稳定性**，通常称之为**随机现象的统计规律性**。

2.1 概率的基本概念

例. 掷币实验的结果列于下表中

实验者	掷币次数	正面次数	频率
蒲丰（C. D. Buffon）	4040	2048	0.5069
皮尔逊（Pearson）	12000	6019	0.5016
皮尔逊（Pearson）	24000	12012	0.5005

随着投掷次数的增加，正面出现的次数越来越接近一个常数：0.5. 这一结果很好的反映了多次重复的随机实验中频率的稳定性。

随机试验与随机事件

试验：通常把根据某一研究目的，在一定条件下对自然现象所进行的观察或试验统称为**试验**。

随机试验：一个试验如果满足下述三个特性，则称其为一个**随机试验**，简称**试验**：

1）试验可以在相同条件下多次重复进行；

2）每次试验的可能结果不止一个，并且事先知道会有哪些可能的结果；

3）每次试验总是恰好出现这些可能结果中的一个，但在一次试验之前却不能肯定这次试验会出现哪一个结果。

样本空间与随机事件

- 样本空间：一个试验的所有可能结果的集合，称为试验的样本空间，一般用Ω表示。而试验的任何一个可能结果称为一个样本点. 一般用ω表示。称样本空间的一个子集称为一个随机事件，简称事件. 通常用A、B、C 等来表示. 称事件A在一次试验中发生，当且仅当试验中出现的样本点 $\omega \in A$
- 基本事件：试验的每个基本结果，称为基本事件. 即只含一个样本点的事件称为基本事件.
- 必然事件 、不可能事件： Ω 本身是 Ω 的子集，它包含所有的样本点，称为必然事件，因为任何一次试验 Ω 都会发生. 记∅为 Ω 的空了集. 它不包含任何样本点，称为不可能事件，因为任何一次试验Φ都不可能发生

例. 以ω、Ω表示样本点与样本空间

	ω	Ω
投1格硬币	{正}，{反}	{正，反}
投2枚硬币	{正正}，{正反}，{反正}，{反反}	{正正，正反，反正，反反}

样本点和样本空间是严格依赖于实验设计的，不同的实验设计可能有不同的样本点和样本空间。每一个最基本、最简单的结果称为样本点，所有可能的样本点构成样本空间，可以是有限集也可以是无穷集，可以是一维或多维的数集，也可以是抽象的集合。

概 率

刻划事件发生可能性大小的数量指标，称为概率。事件A的概率记为$P(A)$。

频率（frequency）

若在相同的条件下，进行了n次试验，在这n次试验中，事件A出现的次数m称为事件A出现的频数，比值m/n称为事件A出现的频率(frequency)，记为W(A)=m/n。

$$0 \le W(A) \le 1$$

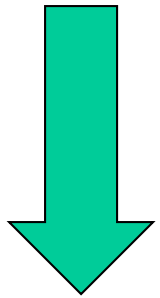
例：

表3-1 玉米种子发芽试验结果

种子总数(n)	10	20	50	100	200	500	1000
发芽种子数(m)	9	19	47	91	186	458	920
种子发芽率(m/n)	0.900	0.950	0.940	0.910	0.930	0.918	0.920

种子发芽与否是不能事先确定的，但从表中可以看出，试验随着n值的不同，种子发芽率也不相同，当n充分大时，发芽率在0.92附近摆动。

频率表明了事件频繁出现的程度，因而其稳定的值说明了随机事件发生的可能性大小，是其本身固有的客观属性，提示了隐藏在随机现象中的规律性。



概 率

概率的统计定义

在相同条件下进行n次重复试验，如果随机事件A发生的次数为m，那么m/n称为随机事件A的**频率**，当试验重复数n逐渐增大时，随机事件A的频率越来越稳定地接近某一数值p，那么就把p称为随机事件A的**概率**.

这样定义的概率称为**统计概率**(statistics probability)或后验概率(posterior probability)

2.1 概率的基本概念

统计概率

抛掷一枚硬币发生正面朝上的试验记录

实验者	投掷次数	发生正面朝上的次数	频率(m/n)
蒲丰	4040	2048	0.5069
K 皮尔逊	12000	6019	0.5016
K 皮尔逊	24000	12012	0.5005

随着实验次数的增多，正面朝上这个事件发生的频率稳定接近0.5，我们称0.5作为这个事件的概率。

概率（probability,P)

$$P(A) = p=\lim \frac{m}{n} \approx \frac{m}{n}$$

在一般情况下，随机事件的概率P是不可能准确得到的。通常以试验次数n充分大时，随机事件A的频率作为该随机事件概率的近似值。

概率的古典定义

对于某些随机事件，不用进行多次重复试验来确定其概率，而是根据随机事件本身的特性直接计算其概率。

随机事件

(1)试验的所有可能结果只有有限个，即样本空间中的基本事件只有有限个;

(2)各个试验的可能结果出现的可能性相等，即所有基本事件的发生是等可能的;

(3)试验的所有可能结果两两互不相容。

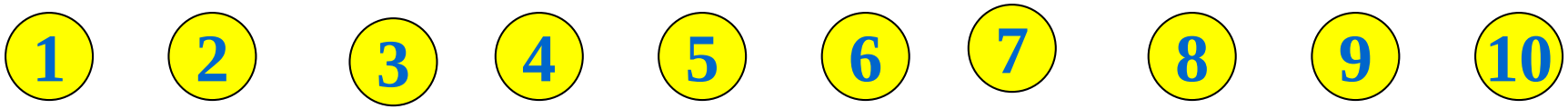
概率的古典定义

具有上述特征的随机试验，称为古典概型（classical model).

设样本空间有n个等可能的基本事件所构成，其中事件A包含有m个基本事件，则事件A的概率为m/n，即P(A)=m/n。

古典概率(classical probability)

先验概率(prior probability)



随机抽取一个球，求下列事件的概率;

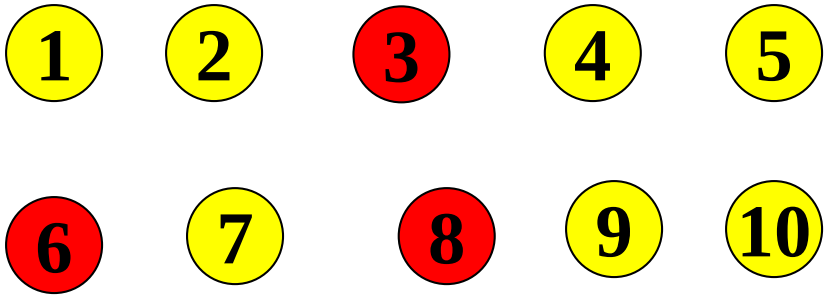
（1)事件A＝抽得一个编号< 4

（2)事件B =抽得一个编号是2的倍数

该试验样本空间由10个等可能的基本事件构成，即n=10，而事件A所包含的基本事件有3个，即抽得编号为1、2、3中的任何一个，事件A便发生。

P(A)=3/10=0.3

P(B)=5/10=0.5

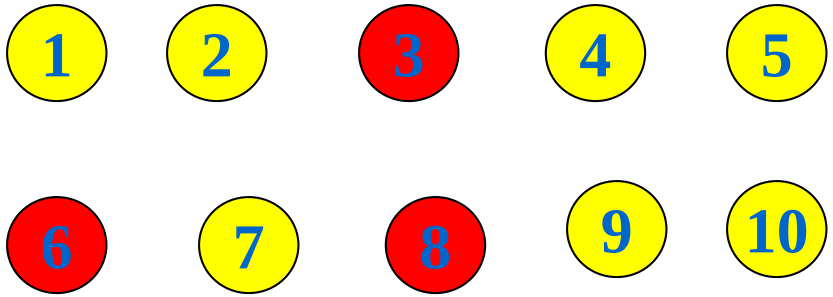


A＝“一次取一个球，取得红球的概率”

10个球中取一个球，其可能结果有10个基本事件（即每个球被取到的可能性是相等的），即n=10

事件A：取得红球，则A事件包含3个基本事件，即m=3

$$P(A)=3/10=0.3$$



B＝ “一次取5个球，其中有2个红球的概率”

10个球中任意取5个，其可能结果有C_{10}^{5}个基本事件，

即n= C_{10}^{5}

事件B =5个球中有2个红球，则B包含的基本事件数m= $C_3^2 C_7^3$

$$P(B) = C_3^2 C_7^3 / C_{10}^5 = 0.417$$

概率的基本性质

任何事件 $0 \leqslant P(A) \leqslant 1$

必然事件 $P(\Omega)=1$

不可能事件 $P(\varnothing)=0$

随机事件 $0<P(A)<1$

2.2 概率的计算

1) 事件的相互关系

和事件

积事件

互斥事件

对立事件

独立事件

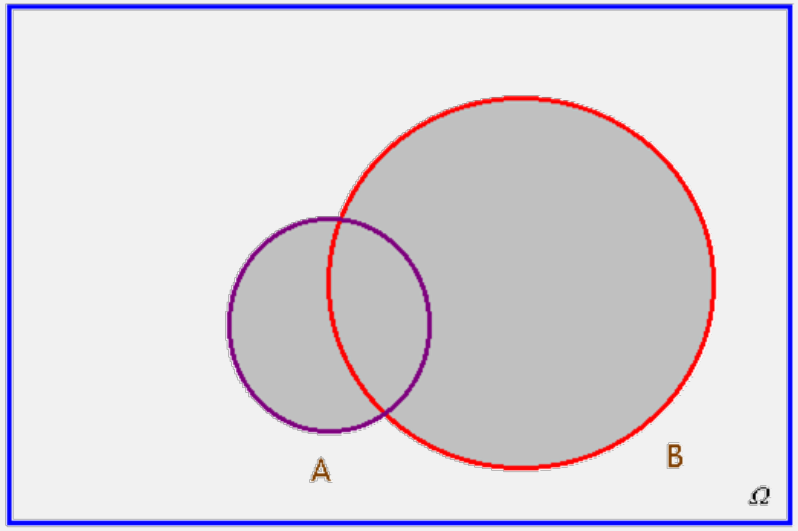
完全事件系

1 和事件

事件A和事件B中至少有一个发生而构成的新事件称为事件A和事件B的和事件或并，记作A+B或者AUB。

n个事件的和，
可表示为$A_1+A_2+\ldots+A_n$
或$A_1 \cup A_2 \cup \ldots \cup A_n$

$A \cup B$

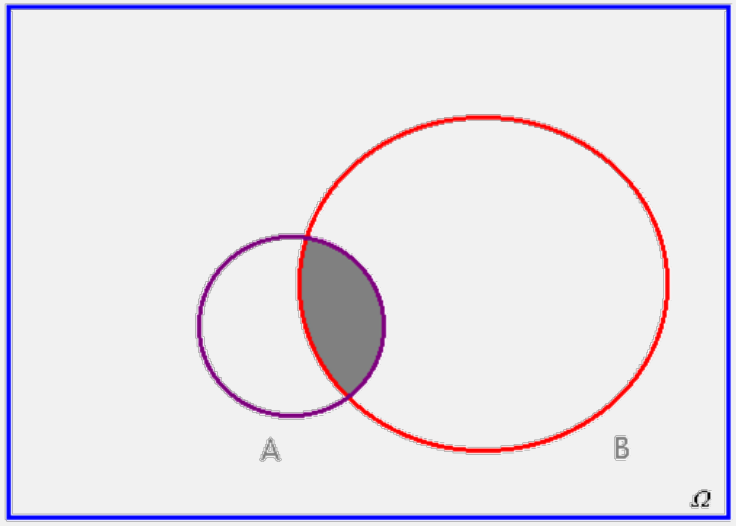


2 积事件（事件的交）

事件A和事件B中同时发生而构成的新事件称为事件A和事件B的积事件，记作A•B或$A \cap B$。

n个事件的积，可表示为A1 • A2 • … • An

$A \cap B$

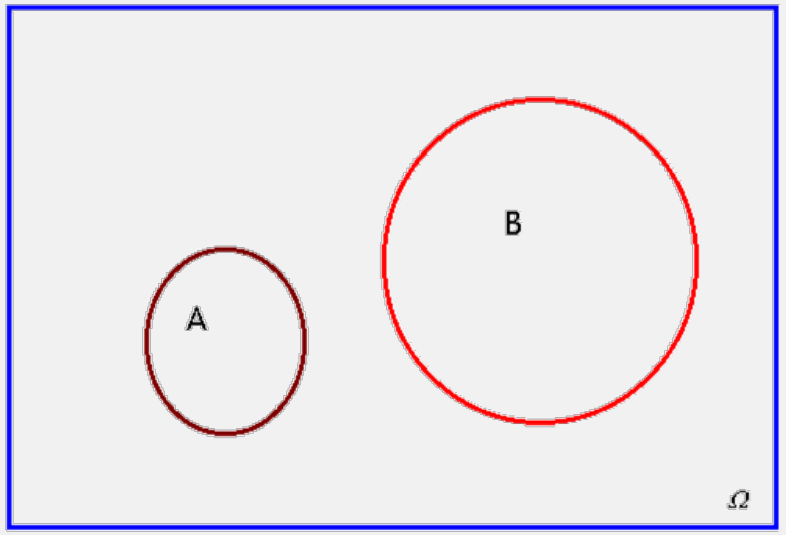


3 互斥事件（互不相容事件）

事件A和事件B不能同时发生，则称这两个事件A和B互不相容或互斥。 A•B=V

n个事件两两互不相容，则称这n个事件互斥。

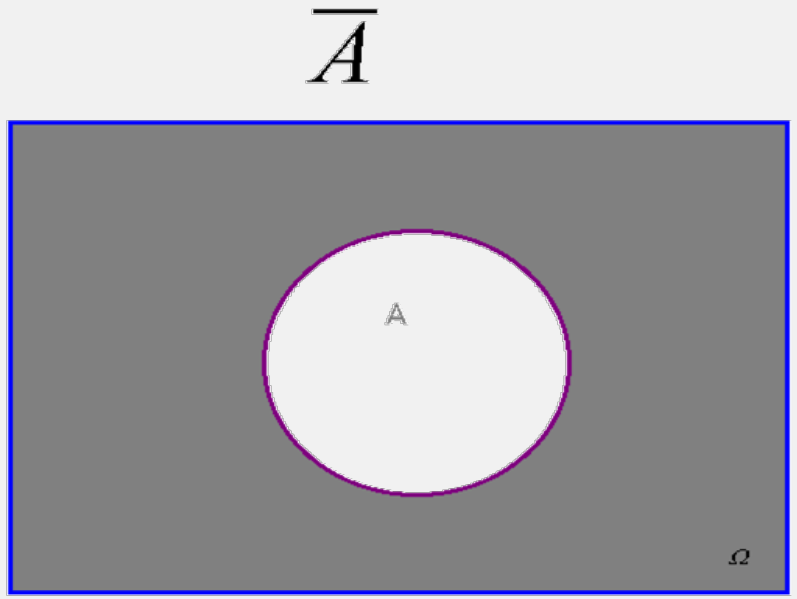
如血型：A\B\O\AB\



4 对立事件

事件A和事件B必有一个发生，但二者不能同时发生，且A和B的和事件组成整个样本空间。即A+B=U，AB=V。我们称事件B为事件A的对立事件。如：新生儿男或女。

$B=\overline{A}$



5 独立事件

事件A和事件B的发生无关，事件B的发生与事件A的发生无关，则事件A和事件B为独立事件。

例如，事件A为“花的颜色为黄色”，事件B为“产量高”，显然如果花的颜色与产量无关,则事件A与事件B相互独立。

如果多个事件A1、A2、A3、…、An 彼此独立，则称之为独立事件群。

6 完全事件系

如果多个事件A_1、A_2、A_3、…、A_n两两互斥，且每次试验结果必然发生其一，则称事件A_1、A_2、A_3、…、A_n为完全事件系。

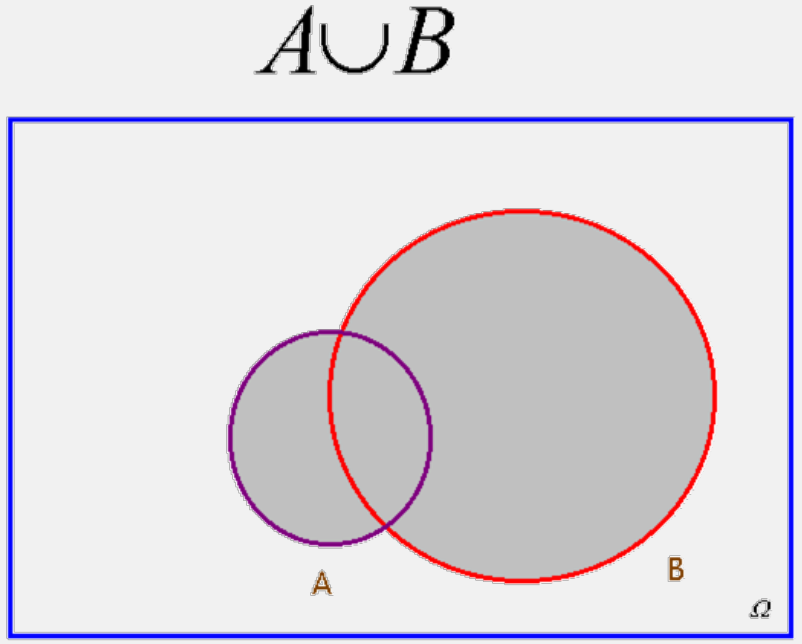
例如，仅有三类花色：黄色、白色和红色，则取一朵花，“取到黄色”、“取到白色”和“取到红色”就构成完全事件系。

完全事件系的和事件概率为 1，任何一个事件发生的概率为1/n。即：

$$P(A_1+A_2+\ldots+A_n)=1$$

2) 概率的加法法则

$P(A \cup B)=P(A)+P(B)- P(A \cap B)$



互斥事件加法定理

定理：若事件A与B互斥，则 $P(A+B)=P(A)+P(B)$

试验的全部结果包含n个基本事件，事件A包含其中m_1个基本事件，事件B包含其中m_2个基本事件。由于A和B互斥，因而它们各包含的基本事件应该完全不同。所以事件A＋B所包含的基本事件数为m_1+m_2。

$$P(A+B)=m_1+m_2/n=m_1/n+m_2/n=P(A)+P(B)$$

推理1　$P(A_1+A_2+\ldots+A_n)=P(A_1)+P(A_2)+\ldots+P(A_n)$

推理2　$P(\bar{A})=1-P(A)$

推理3　完全事件系的和事件的概率为1。

2) 条件概率

已知事件B发生的条件下，事件A发生的条件概率，记为P(A|B)。

条件概率计算：

$$P(A|B) = \frac{P(AB)}{P(B)}$$

	死亡（A）	存活（$\overline{A}$）	和
甲药物（B）	96	24	120
乙药物（$\overline{B}$）	64	16	80
和	160	40	200

现在计算以下个概率:

1）在200只虫子中任取一个，该虫子是死的概率?

2)在200只虫子中任取一个，该虫子接收了甲药物的概率?

3)在200只虫子中任取一个，该虫子接收了甲药物且死亡的概率?

3)在200只虫子中任取一个，该虫子死亡下，接收了甲药物的条件概率?

$$P(A|B) = \frac{P(AB)}{P(B)}$$

3) 概率乘法法则

$$P(AB) = P(A|B)P(B)$$

$$P(AB) = P(B|A)P(A)$$

独立事件乘法定理

定理：事件A和事件B为独立事件，则事件A与事件B同时发生的概率为各自概率的乘积。

$$P(AB)=P(A)P(B)$$

推理：A_1、A_2、…A_n彼此独立，则

$$P(A_1A_2A_3\ldots A_n)=P(A_1)P(A_2)P(A_3)\ldots P(A_n)$$

例:播种玉米，种子的发芽率为90%，每穴两粒，则：

A:第一粒种子发芽，P (A) = 0.9， $P(\bar{A}) = 0.1$

B:第二粒种子发芽，P (B) = 0.9， $P(\bar{B}) = 0.1$

求：

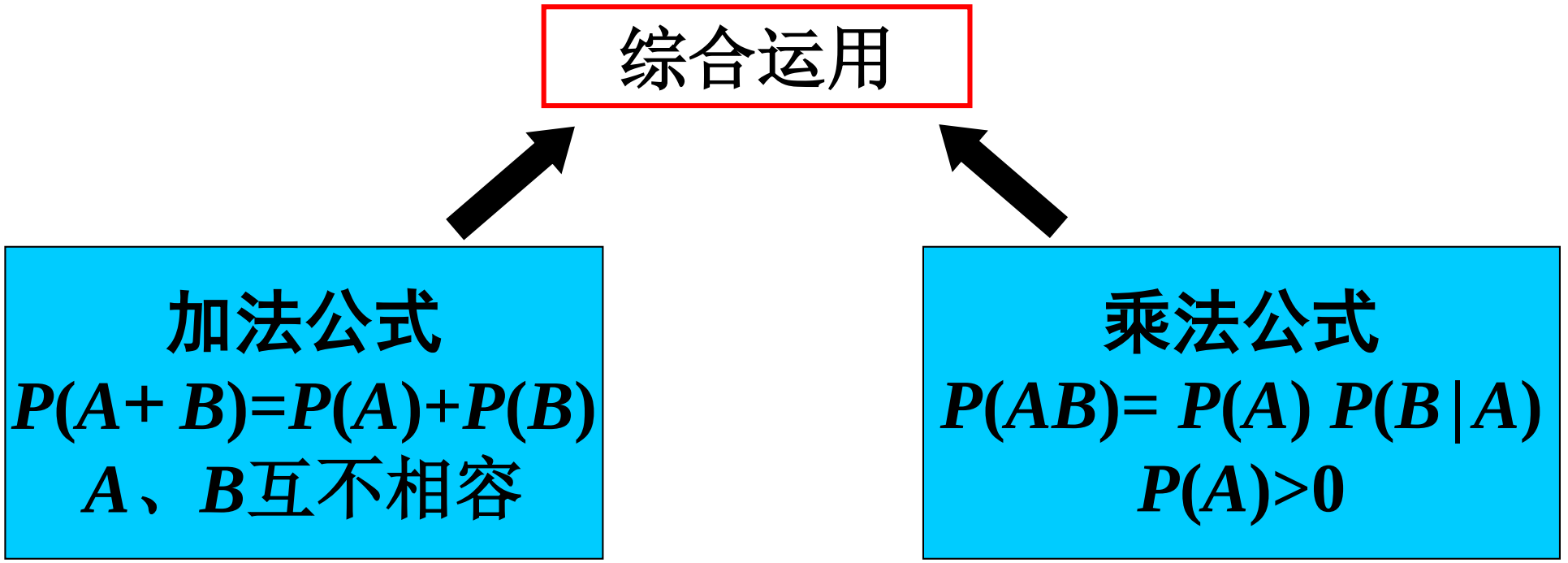
C:两粒种子均发芽，C = AB，P（C) = P(A) P(B) = 0.81

D:一粒种子发芽：$D= A\bar{B} + \bar{A}B$，P(D)＝0.9*0.1+ 0.1*0.9=0.18

E:两粒种子均不发芽：$E= \bar{A}\,\bar{B}$，$P(E)＝P(\bar{A})P(\bar{B})$=0.1*0.1=0.01

4) 全概率公式和贝叶斯(Bayes)公式

全概率公式和贝叶斯公式主要用于计算比较复杂事件的概率，它们实质上是加法公式和乘法公式的综合运用.

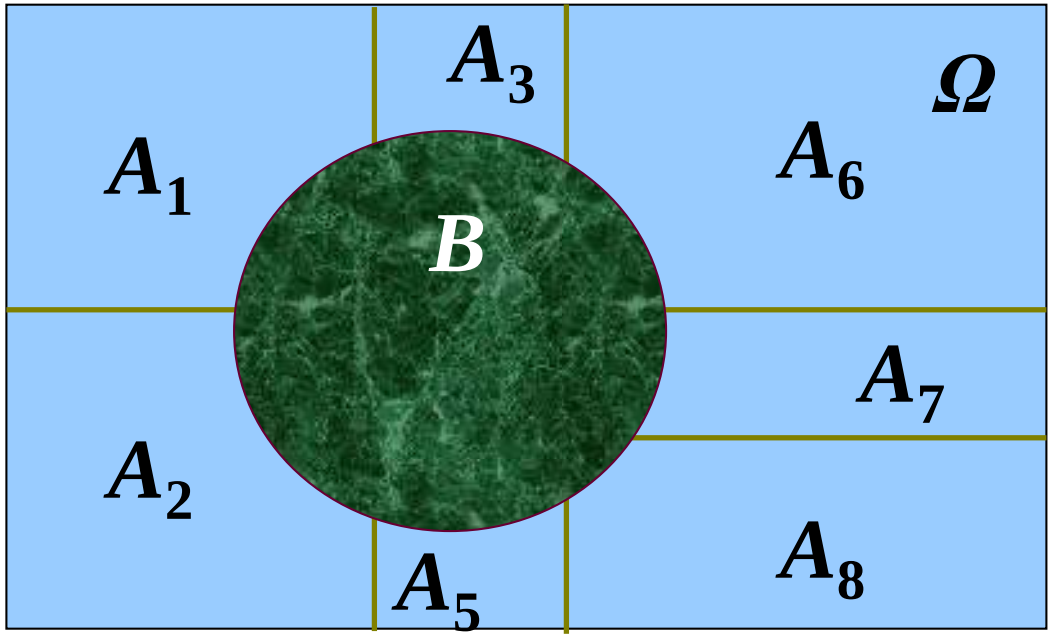


设 $A_1, A_2, \cdots, A_n$ 为一个完备事件组，对任一事件 B，有

$$B = \Omega B = (A_1 + A_2 + \cdots + A_n)B$$

$$= A_1B + A_2B + \cdots + A_nB\ ,$$

显然 $A_1B, A_2B, \cdots, A_nB$ 也两两互不相容，



设 $A_1, A_2, \cdots, A_n$ 为一个完备事件组, 对任一事件 B, 有

$$B = \Omega B = (A_1 + A_2 + \cdots + A_n)B$$

$$= A_1B + A_2B + \cdots + A_nB\,,$$

显然 $A_1B, A_2B, \cdots, A_nB$ 也两两互不相容，

由概率的**可加性**及**乘法公式**，有

$$P(B) = P(A_1B + A_2B + \cdots + A_nB) = \sum_{i=1}^{n} P(A_iB)$$

$$= \sum_{i=1}^{n} P(A_i)\, P(B \mid A_i)\,.$$

这个公式称为**全概率公式**，它是概率论的基本公式.

全概率公式

$$P(B)=\sum_{i=1}^{n}P(A_i)\,P(B\mid A_i)$$

利用全概率公式，可以把较复杂事件概率的计算问题，化为若干互不相容的较简单情形，分别求概率然后求和.

例1 市场上有甲、乙、丙三家工厂生产的同一品牌产品，已知三家工厂的市场占有率分别为30%、20%、50%，且三家工厂的次品率分别为3%、3%、1%，试求市场上该品牌产品的次品率.

解 设A_1、A_2、A_3分别表示买到一件甲、乙、丙的产品；B表示买到一件次品，显然A_1、A_2、A_3构成一个完备事件组，由题意有

$$P(A_1)=0.3,\ P(A_2)=0.2,\ P(A_3)=0.5,$$

$$P(B\mid A_1)=0.03,\ P(B\mid A_2)=0.03,\ P(B\mid A_3)=0.01,$$

由全概率公式，

$$P(B)=\sum_{i=1}^{3}P(A_i)P(B\mid A_i)$$

$$=0.3\times0.03+0.2\times0.03+0.5\times0.01=0.02.$$

加权平均

例2 袋中有a个白球b个黑球，不放回摸球两次，问第二次摸出白球的概率为多少？

解 分别记A, B为第一次、第二次摸到白球，

由全概率公式，

$$P(B)=P(A)P(B|A)+P(\overline{A})P(B|\overline{A})$$

$$=\frac{a}{a+b}\cdot\frac{a-1}{a+b-1}+\frac{b}{a+b}\cdot\frac{a}{a+b-1}=\frac{a}{a+b}.$$

可以想见，第三次、第四次…摸出白球的概率仍为$\frac{a}{a+b}$，这体现了抽签好坏与先后次序无关的公平性.

例3 袋中有a个白球b个黑球，分别以A, B记第一次、第二次摸得白球，(1)采用有放回摸球；(2)采用无放回摸球，试分别判断A, B的独立性.

解 (1) 有放回摸球,

$$P(A)=\frac{a}{a+b},\ P(AB)=\frac{a^2}{(a+b)^2},\ P(\overline{A}B)=\frac{ab}{(a+b)^2},$$

$$\therefore\quad P(B\mid A)=\frac{P(AB)}{P(A)}=\frac{a}{a+b}.$$

全概率公式

$$\text{而}\quad P(B)=P(AB)+P(\overline{A}B)=\frac{a^2}{(a+b)^2}+\frac{ab}{(a+b)^2}=\frac{a}{a+b}.$$

由于 $P(B\mid A)=P(B)$，所以A, B相互独立.

(2) 无放回摸球, $P(A)=\dfrac{a}{a+b}$,

$$P(AB)=\frac{a(a-1)}{(a+b)(a+b-1)},\quad P(\overline{A}B)=\frac{ab}{(a+b)(a+b-1)},$$

$$\therefore\quad P(B\mid A)=\frac{P(AB)}{P(A)}=\frac{a-1}{a+b-1}.$$

而 $P(B)=P(AB)+P(\overline{A}B)$

$$=\frac{a(a-1)}{(a+b)(a+b-1)}+\frac{ab}{(a+b)(a+b-1)}=\frac{a}{a+b},$$

由于 $P(B\mid A)\neq P(B)$, 所以A, B不相互独立.

在上面例1中，如买到一件次品，问它是甲厂生产的概率为多大？这就要用到贝叶斯公式.

定理(贝叶斯公式) 设 $A_1, A_2, \cdots, A_n$ 为一个完备事件组，$P(A_i) > 0$，$i = 1, \cdots, n$，对任一事件 B，若 $P(B) > 0$，有

$$P(A_k \mid B) = \frac{P(A_k B)}{P(B)}$$

$$= \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}, \qquad (k = 1, 2, \cdots, n)$$

贝叶斯公式

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}$$

$$(k = 1,2,\cdots,n)$$

该公式于1763年由贝叶斯(Bayes)给出．它是在观察到事件B已发生的条件下，寻找导致B发生的每个原因A_k的概率.

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)} \quad (k = 1,2,\cdots,n)$$

贝叶斯 **Thomas Bayes**，英国数学家，**1702**年出生于伦敦，做过神甫. **1742**年成为英国皇家学会会员. **1763**年**4**月**7**日逝世. 贝叶斯在数学方面主要研究概率论. 他对统计推理的主要贡献是使用了“逆概率”这个概念，在**1763**年提出了著名的贝叶斯公式.

例4 已知三家工厂的市场占有率分别为30%、20%、50%，次品率分别为3%、3%、1%. 如果买了一件商品，发现是次品，问它是甲、乙、丙厂生产的概率分别为多少?

解 $P(A_1 \mid B)=\dfrac{P(A_1)P(B \mid A_1)}{P(B)}=\dfrac{0.3\times 0.03}{0.02}=0.45$,

$P(A_2 \mid B)=\dfrac{0.2\times 0.03}{0.02}=0.3$, $P(A_3 \mid B)=\dfrac{0.5\times 0.01}{0.02}=0.25$.

所以这件商品最有可能是甲厂生产的.

$P(A_i)$: 0.3, 0.2, 0.5

$P(A_i \mid B)$: 0.45, 0.3, 0.25

解释：事件 $A_1, A_2, \cdots, A_n$ 看作是导致事件 B 发生的“原因”，在不知事件 B 是否发生的情况下，它们的概率为 $P(A_1), P(A_2), \cdots, P(A_n)$，通常称为**先验概率**.

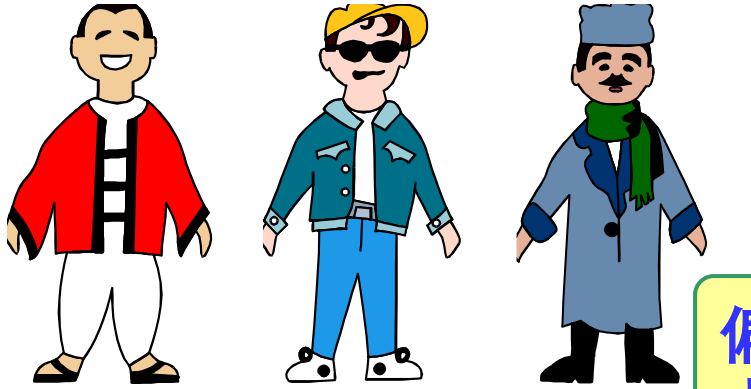
现在有了新的信息已知（B 发生），我们对 $A_1, A_2, \cdots, A_n$ 发生的可能性大小 $P(A_1 \mid B), P(A_2 \mid B), \cdots, P(A_n \mid B)$ 有了新的估价，称为“**后验概率**”.

全概率公式可看成“由原因推结果”，而贝叶斯公式的作用在于“由结果推原因”：现在一个“结果” A 已经发生了，在众多可能的“原因”中，到底是哪一个导致了这一结果？

故贝叶斯公式也称为“**逆概公式**”.

例如，某地发生了一个案件，怀疑对象有甲、乙、丙三人.

在不了解案情细节(事件A)之前，侦破人员根据过去的前科，对他们作案的可能性有一个估计，设为



甲　乙　丙

偏小

$P(A_1)$　$P(A_2)$　$P(A_3)$

但在知道案情细节后, 这个估计就有了变化.

知道B发生后

$P(A_1 \mid B)$　$P(A_2 \mid B)$　$P(A_3 \mid B)$

最大

比如原来认为作案可能性较小的某丙，现在变成了重点嫌疑犯.

在医疗诊断中，为了诊断病人到底患了毛病 $A_1, A_2, \cdots, A_n$ 中的哪一种，对病人进行检查，确定了某个指标 B(比如体温).

根据以往资料可知 $P(A_1), P(A_2), \cdots, P(A_n)$，

依靠医疗知识可知 $P(B|A_1), P(B|A_2), \cdots, P(B|A_n)$，

再利用贝叶斯公式算出 $P(A_i|B)$，显然对较大的 $P(A_i|B)$ 的“病因” A_i 应多加考虑.

在实际工作中检查的指标 B 一般有多个，综合这些后验概率，当然会对诊断有很大帮助，在实现计算机自动诊断或辅助诊断中，这一方法是有实用价值的.

下面举一个实际的医学例子，说明贝叶斯公式在解决实际问题中的作用.

例5 用血清甲胎蛋白法诊断肝癌，A 表示被检验者患肝癌，B 表示判断被检验者患肝癌. 由于种种原因使检验方法带有误差. 假定 $P(B|A)=0.95$，$P(\overline{B}|\overline{A})=0.90$，又设人群中患肝癌的比例为 $P(A)=0.0004$. 现在若有一人被此法诊断为患肝癌，求此人真正患肝癌的概率 $P(A|B)$.

解 $$P(A|B)=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\overline{A})P(B|\overline{A})}$$

$$=\frac{0.0004\times 0.95}{0.0004\times 0.95+0.9996\times 0.1}=0.0038\,.$$

因此，虽然检验法相当可靠，但被诊断为患肝癌的人真正患病的概率并不大，其主要原因是人群中患 肝癌的比例相当小. 当然，医生在公布某人患肝癌之前，是不会只做一次或一种检验，还会辅以其他检验手段.

一个不懂概率的人可能会这样推理：一个没有患肝癌的人被诊断为患肝癌的机会才 $P(B|\overline{A})=0.1$，现在我被诊断为患肝癌，说明我患肝癌的概率为 0.9，其实大相径庭. 正确的概率思维是人们正确地思考问题而必备的文化修养的一个成分.

思考：诊断为无病，而确实没有患病的概率为多少？

大数据背后的神秘公式：贝叶斯公式

贝叶斯公式在商业决策及其他企业管理学科中也有重要应用. 有人依据贝叶斯公式的思想发展了一整套统计推断方法, 叫作“贝叶斯统计”. 可见贝叶斯公式的影响. 与其他统计学方法不同，贝叶斯方法建立在主观判断的基础上，你可以先估计一个值，然后根据客观事实不断修正。

大数据、人工智能、海难搜救、生物医学、邮件过滤，这些看起来彼此不相关的领域之间有什么联系？答案是，它们都会用到同一个数学公式——贝叶斯公式。它虽然看起来很简单、很不起眼，但却有着深刻的内涵。如：2014年初马航失联航班MH370的搜救和1968年5月美国海军天蝎号核潜艇在大西洋亚速海海域失踪后的搜救.

贝叶斯公式为我们提供了一些决策原则

- 平时注意观察和思考，建立自己的思维框架，这样在面临选择时就容易形成一个接近实际情况的先验概率，这样经过少量的试错和纠错的迭代循环就可能得到理想的结果；在经过很多次选择和实践的历练后就能够形成自己的直觉，在面对陌生情况时，根据自己的经验和少量信息就能够快速地做出比较准确的判断。
- 大数据时代获得信息的成本越来越低，社会也变得更加开放和包容，初始状态（先验概率）的重要性下降了，即使最初选择不理想，只要根据新情况不断进行调整，仍然可以取得成功。所以如果当下觉得很难做出选择，那就倾听内心的声音，让直觉来选择，这有利于治疗选择恐惧症。
- 以开发App的例子来说，先按照自己的想法弄个可用的原型出来，然后充分利用互联网的力量，让活跃的用户社区帮助它快速迭代，逐渐使它的功能和体验越来越好。
- 对新鲜事物保持开放的心态，愿意根据新信息对自己的策略和行为进行调整。“大胆假设，小心求证”，“不断试错，快速迭代”，这些都可以看成贝叶斯公式的不同表述。英国哲学家以赛亚·伯林（Isaish Berlin）曾经援引古希腊诗人的断简残片“狐狸多知而刺猬有一大知”，将人的策略分为狐狸和刺猬两类。刺猬用一个宏大的概念解释所有现象，而狐狸知道很多事情，用多元化的视角看待问题，它也愿意包容新的证据以使得自己的模型与之相适应。在这个快速变化的时代，固守一个不变的信条的刺猬很难适应环境的变化，而使用贝叶斯公式的灵活的狐狸才更容易生存。

例6 10个乒乓球有7个新球3个旧球. 第一次比赛时随机取出2个，用过后放回. 现在第二次比赛 又取出 2个，问第二次取到几个新球的概率最大?

解 设 A_i 为第一次取到 i 个新球， $i=0,1,2$ ，

B_j 为第二次取到 i 个新球， $j=0,1,2$ ，

A_0, A_1, A_2 构成一个完备事件组，

$$P(A_i)=\frac{C_7^i C_3^{2-i}}{C_{10}^2}, \quad i=0,1,2,$$

$$P(B_j \mid A_i)=\frac{C_{7-i}^j C_{3+i}^{2-j}}{C_{10}^2}, \quad i, j=0,1,2,$$

$$P(A_i)=\frac{C_7^i C_3^{2-i}}{C_{10}^2},\quad P(B_j\mid A_i)=\frac{C_{7-i}^j C_{3+i}^{2-j}}{C_{10}^2},\quad i,j=0,1,2,$$

具体计算得

$$P(A_0)=\frac{1}{15},\quad P(A_1)=\frac{7}{15},\quad P(A_2)=\frac{7}{15},$$

$$P(B_0\mid A_0)=\frac{1}{15},\quad P(B_0\mid A_1)=\frac{2}{15},\quad P(B_0\mid A_2)=\frac{2}{9},$$

$$P(B_1\mid A_0)=\frac{7}{15},\quad P(B_1\mid A_1)=\frac{8}{15},\quad P(B_1\mid A_2)=\frac{5}{9},$$

$$P(B_2\mid A_0)=\frac{7}{15},\quad P(B_2\mid A_1)=\frac{1}{3},\quad P(B_2\mid A_2)=\frac{2}{9},$$

由全概率公式,

$$P(B_0)=\sum_{i=0}^{2}P(A_i)P(B_0\mid A_i)$$

$$=\frac{1}{15}\times\frac{1}{15}+\frac{7}{15}\times\frac{2}{15}+\frac{7}{15}\times\frac{2}{9}=0.17\,,$$

$$P(B_1)=\frac{1}{15}\times\frac{7}{15}+\frac{7}{15}\times\frac{8}{15}+\frac{7}{15}\times\frac{5}{9}=0.54\,,$$

$$P(B_2)=\frac{1}{15}\times\frac{7}{15}+\frac{7}{15}\times\frac{1}{3}+\frac{7}{15}\times\frac{2}{9}=0.29\,,$$

所以第二次取到一个新球的概率最大.

如果发现第二次取到的是两个新球，问第一次没有取到新球的概率为多大？

由贝叶斯公式，

$$P(A_0 \mid B_2) = \frac{P(A_0)P(B_2 \mid A_0)}{\sum_{i=0}^{2} P(A_i)P(B_2 \mid A_i)}$$

$$= \frac{\frac{1}{15} \times \frac{7}{15}}{\frac{1}{15} \times \frac{7}{15} + \frac{7}{15} \times \frac{1}{3} + \frac{7}{15} \times \frac{2}{9}} = 0.11 .$$

例7 甲、乙、丙三人独立地向同一飞机射击，设击中的概率分别为0.4, 0.5, 0.7．如果只有一人击中，则飞机被击落的概率为0.2；如果有两人击中，则飞机 被击落的 概率为0.6；如果三人都击中，则飞机一定被 击落.

(1)求飞机被击落的概率；(2)若飞机被击落，求是 三人同时击中的概率.

解 (1) 以 $A_i(i=1,2,3)$ 分别表示甲、乙、丙击中飞机，$B_i(i=0,1,2,3)$ 表示有 i 个人击中飞机，C 表示飞机被击落，则

$$P(B_0)=P(\overline{A}_1\overline{A}_2\overline{A}_3)=P(\overline{A}_1)P(\overline{A}_2)P(\overline{A}_3)$$

$$=0.6\times0.5\times0.3=0.09,$$

由独立性

$$P(B_0)=0.09,$$

$$P(B_1)=P(A_1\overline{A}_2\overline{A}_3)+P(\overline{A}_1A_2\overline{A}_3)+P(\overline{A}_1\overline{A}_2A_3)$$

$$=0.4\times0.5\times0.3+0.6\times0.5\times0.3+0.6\times0.5\times0.7=0.36$$

$$P(B_2)=P(A_1A_2\overline{A}_3)+P(A_1\overline{A}_2A_3)+P(\overline{A}_1A_2A_3)$$

$$=0.4\times0.5\times0.3+0.4\times0.5\times0.7+0.6\times0.5\times0.7=0.41,$$

$$P(B_3)=P(A_1A_2A_3)=0.4\times0.5\times0.7=0.14,$$

由全概率公式

$$P(C)=\sum_{i=0}^{3}P(C\mid B_i)P(B_i)$$

$$=0\times0.09+0.2\times0.36+0.6\times0.41+1\times0.14=0.458.$$

$$P(C)=\sum_{i=0}^{3}P(C\mid B_i)P(B_i)$$

$$=0\times0.09+0.2\times0.36+0.6\times0.41+1\times0.14=0.458\,.$$

---

(2) 由贝叶斯公式有

$$P(B_3\mid C)=\frac{P(B_3)P(C\mid B_3)}{\sum_{i=0}^{3}P(B_i)P(C\mid B_i)}=\frac{0.14}{0.458}=0.306\,.$$

类似有： $P(B_2\mid C)=0.537$， $P(B_1\mid C)=0.157$ 。

2.3 概 率 分 布

概率分布

随机变量

随机变量：在随机试验中被测定的量。
随机变量所取得值为观测值。

例1. 抛硬币试验中$S=\{H,T\}$, H与T不是数量,不便于计算及理论的研究,因而引入以下变量X,

$$X=X(e)=\begin{cases}0, e=T,\\ 1, e=H.\end{cases}$$

即X(e)是定义在样本空间S上的一个实函数,对于不同的试验结果e, X取不同的值, 由于试验前不能预料e的取值, 因而X取1还是取0也是随机的, 故称X(e)为随机变量。

例：每10只新生动物中，雄性动物的只数.

记雄性动物的只数为X, 则X是定义在样本空间

S={0,1,2,3,4,5,6,7,8,9,10}上的函数, 即

$$X=t,\ t\in S.$$

例3. 测试灯泡寿命试验, 其结果是用数量表示的. 记灯泡的寿命为X, 则X是定义在样本空间

$S=\{e\}=\{t|t\geq 0\}$上的函数, 即

$$X=X(e)=t,\ e=t\in S.$$

注

(1) 可用随机变量X描述事件.

例掷一颗骰子, 设出现的点数记为X, 事件A为“掷出的点 数大于 3”， 则A可表示为“X>3”.

反过来, X的一个变化范围表示一个随机事件:
“2<X<5”表示事件“掷出的点数大于2且小于5”.

(2) 随机变量随着试验的结果而取不同的值,在试验之前不能确切知道它取什么值, 但是随机变量的取值有一定的统计规律性—概率分布.

分类:

(1) 离散型随机变量;

(2) 非离散型随机变量

1、离散型随机变量

离散型随机变量

若随机变量全部可能取到的值是有限多个或可列无限多个。

将这种变量的所有可能取值及其对应的概率一一列出所形成的分布，称为**离散型随机变量的概率分布**：

变量x_i	x_1	x_2	x_3	…	x_n
概率$P(X=x_i)$	P_1	P_2	P_3	…	P_n

求分布律的步骤:

(1) 明确X的一切可能取值;

(2) 利用概率的计算方法计算X取各个确定值的概率, 即可写出X的概率分布律.

例1. 设一汽车在开往目的地的道路上需经过四盏信号灯,每盏信号灯以概率p禁止汽车通过, 以X表示汽车首次停下时已通过信号灯的盏数, 求X的分布律.(设各信号灯的工作是相互独立的).

X	0	1	2	3
p_k	p	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$

例2. 袋中装有4只红球和2只白球,从袋中不放回地逐一地摸球, 直到第一次摸出红球为止,设X表示到第一次摸出红球时所摸的次数, 求X的分布律.

X	1	2	3
p_k	$\frac{4}{6}$	$\frac{2}{6}*\frac{4}{5}$	$\frac{2}{6}*\frac{1}{5}$

例：

表3-2某鱼群的年龄组成

年龄(x)	1	2	3	4	5	6	7
频率(W)	0.4597	0.3335	0.1254	0.0507	0.0215	0.0080	0.0012

此表给出了该鱼群年龄构成的全部，我们称之为该鱼群年龄的概率分布。

例：

表 婴儿的性别情况表

性别(x)	0（男）	1（女）
概率(P)	0.517	0.483

此表列出了性别变量的取值及相应值的概率，揭示了观察婴儿性别试验的统计规律。

用随机变量的可能取值及取相应值的概率来表示随机试验的规律称为随机变量的分布律或概率函数。

表3-3 离散型变量的概率分布

变量(x)	x_1	x_2	x_3	x_4	……	x_n
概率(P)	p_1	p_2	p_3	p_4	…….	p_n

$$P\ (x=x_i) = p_i,\ \ i=1,2,3\ldots n$$

设离散型变量x的所有一切可能值x_i(i=1,2,3…n),取相应值的概率为p_i，则P (x=x_i)称为离散型随机变量x的概率函数。

离散型变量的概率分布的特点

$$P_i \geqslant 0 \qquad (i=1,2,\ldots)$$

$$\sum_{i=1}^{\infty} Pi = 1$$

离散型随机变量的分布函数

离散型随机变量的分布函数指随机变量小于等于某一可能值（y_0）的概率。

$$F(y_0) = \sum_{y_i \le y_0} p(y_i) = P(Y \le y_0)$$

（二）连续型变量的概率分布

当试验资料为连续型变量，一般通过分组整理成频率分布表。如果从总体中抽取样本的容量n相当大，则频率分布就趋于稳定，我们将它近似地看成总体概率分布。

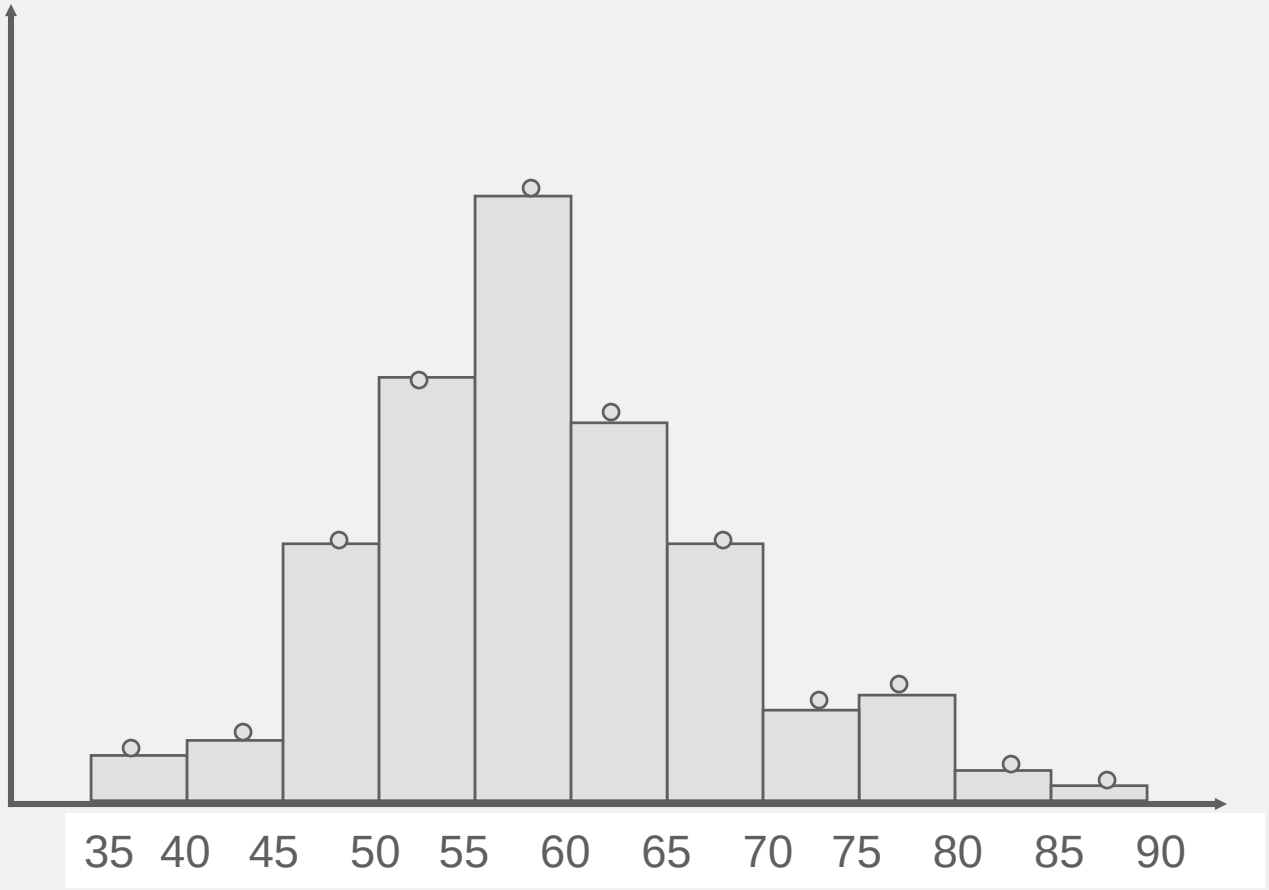


图3.1 鲢鱼体长的频率分布图

直方图中同一组内的频率是相等的。

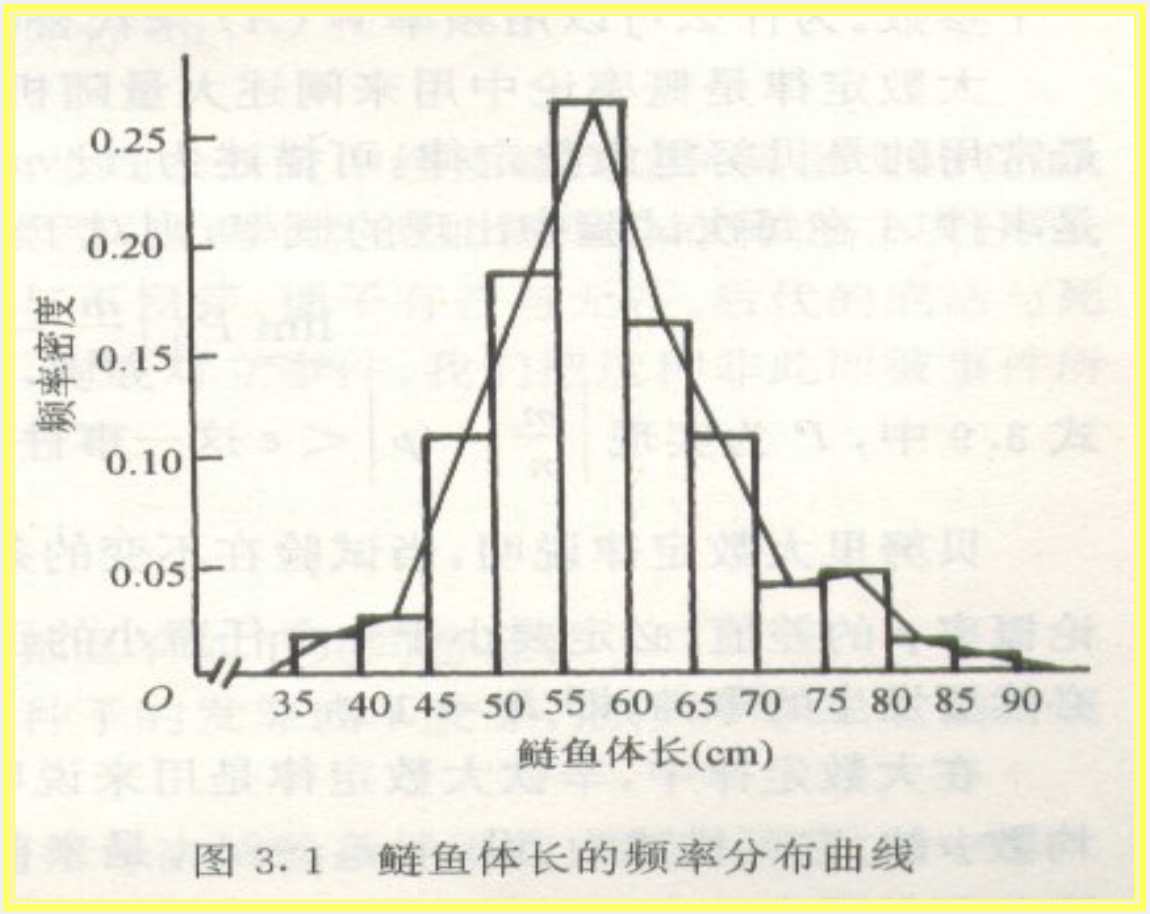


图 3.1　鲢鱼体长的频率分布曲线

直方图中每一矩形的面积就表示该组的频率。

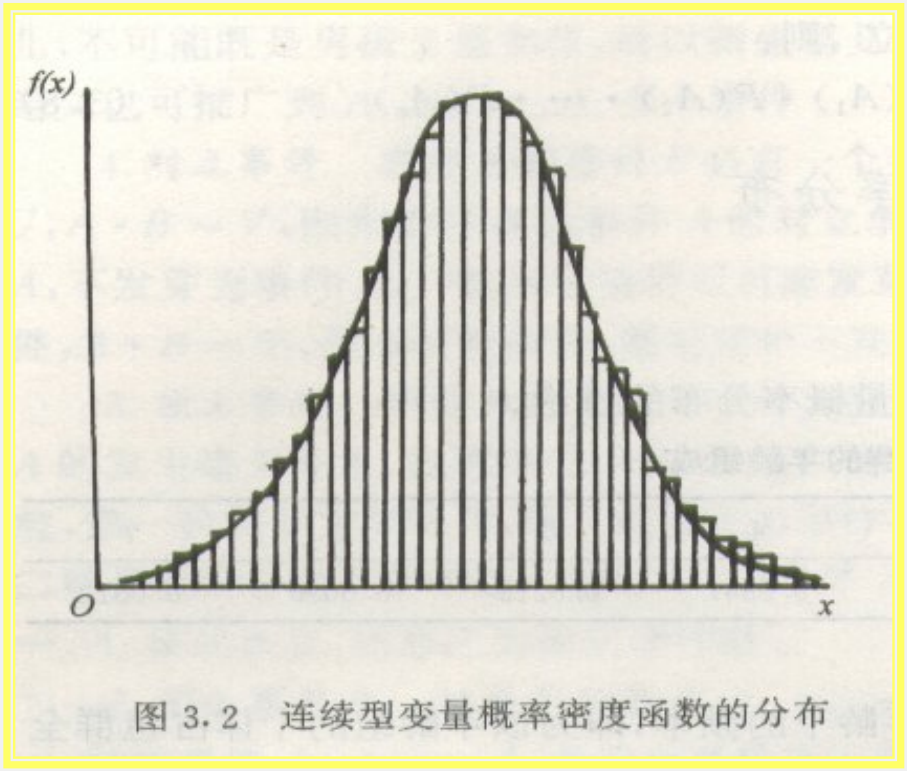


图 3.2　连续型变量概率密度函数的分布

当n无限大时，频率转化为概率，频率密度也转化为概率密度，阶梯形曲线也就转化为一条光滑的连续曲线，这时**频率分布也就转化为概率分布**了，此曲线为总体的概率密度曲线，曲线函数f(x)称为概率密度函数。

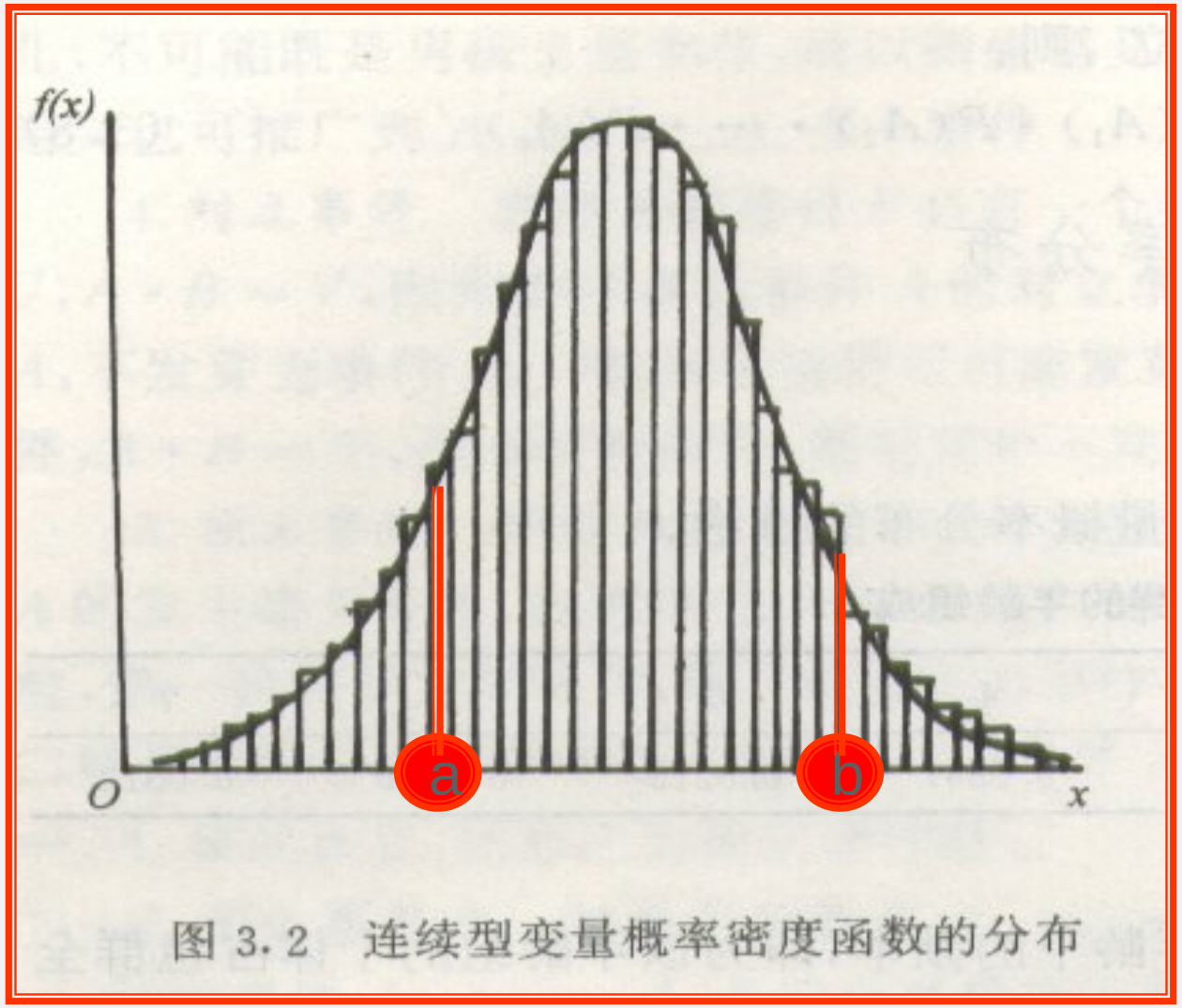


图 3.2　连续型变量概率密度函数的分布

对于一个连续型随机变量x，取值于区间[a,b]内的概率为函数f(x)从a到b的积分，即：

$$P(a \le x \le b) = \int_a^b f(x)dx$$

连续型随机变量的概率由概率分布密度函数所确定。

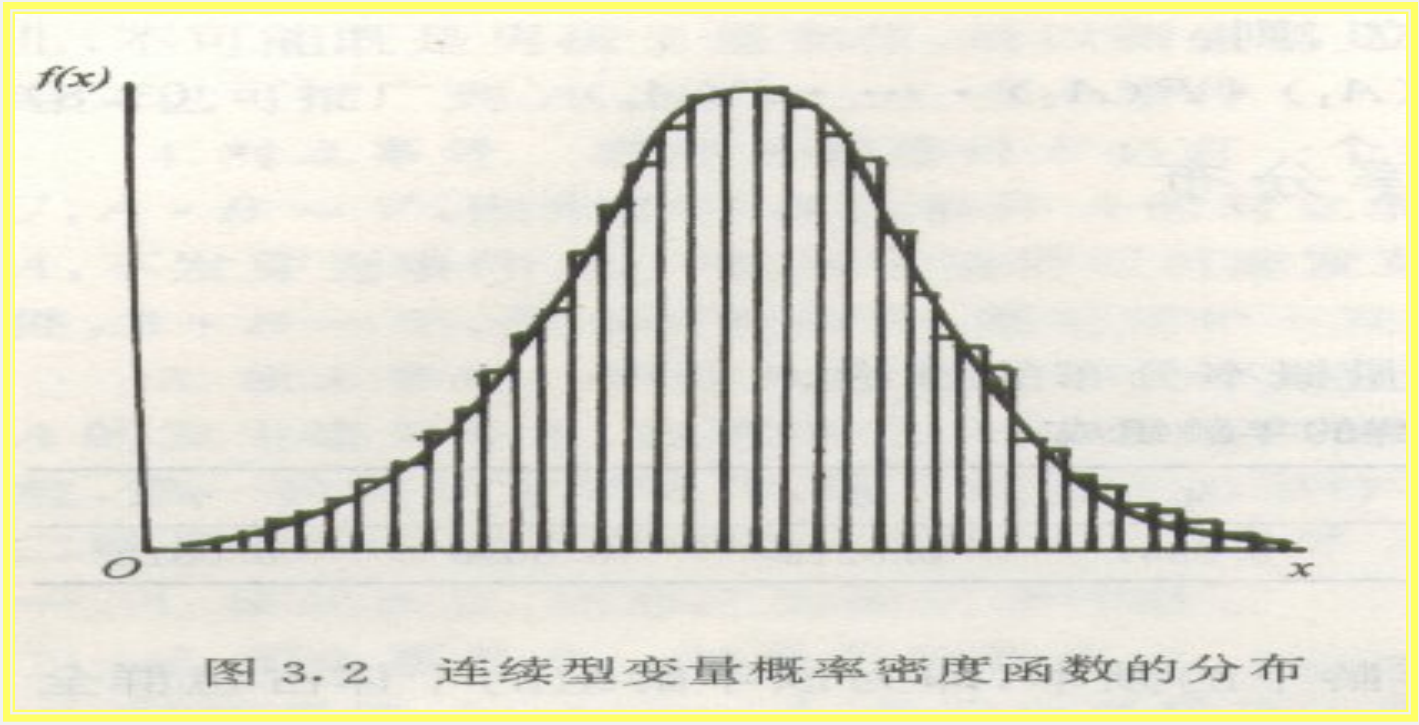


图 3.2　连续型变量概率密度函数的分布

$$P(-\infty \le x \le \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

概率密度函数f(x)曲线与x轴所围成的面积为1。

连续型随机变量的概率分布

- 连续型随机变量可以取某一区间或整个实数轴上的任意一个值
- 它取任何一个特定的值的概率都等于0
- 不能列出每一个值及其相应的概率
- **通常研究它取某一区间值的概率**
- 用数学函数的形式和分布函数的形式来描述

概率密度函数(probability density function)

设X为一连续型随机变量，x 为任意实数，
X的**概率密度函数**记为$f(x)$，它满足条件

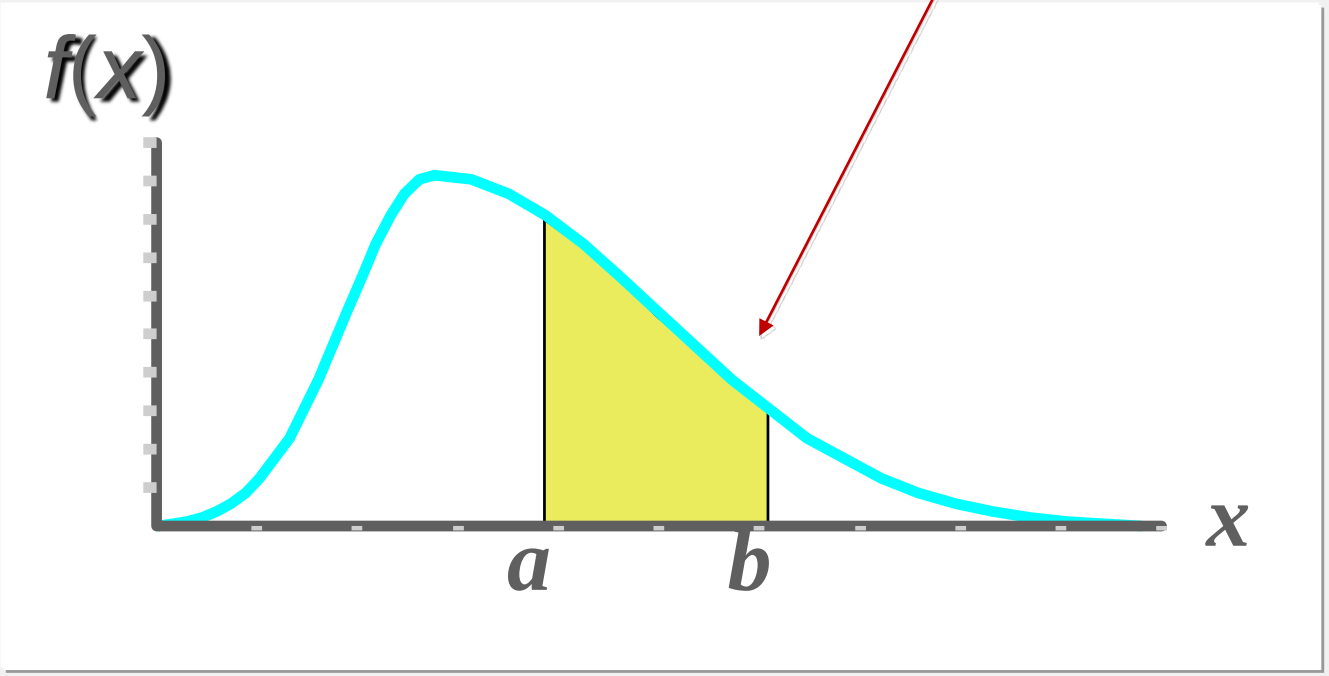
$$(1) f(x) \geq 0$$

$$(2) \int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$$

注：$f(x)$不是概率

- 在平面直角坐标系中画出$f(x)$的图形，则对于任何实数 a < b，$P(a< X\leq b)$是该曲线下从a到 b的面积

$$P(a < X \leq b) = \int_a^b f(x)\mathrm{d}x$$



分布函数 (distribution function)

1. 连续型随机变量的概率也可以用分布函数$F(x)$来表示
2. 分布函数定义为

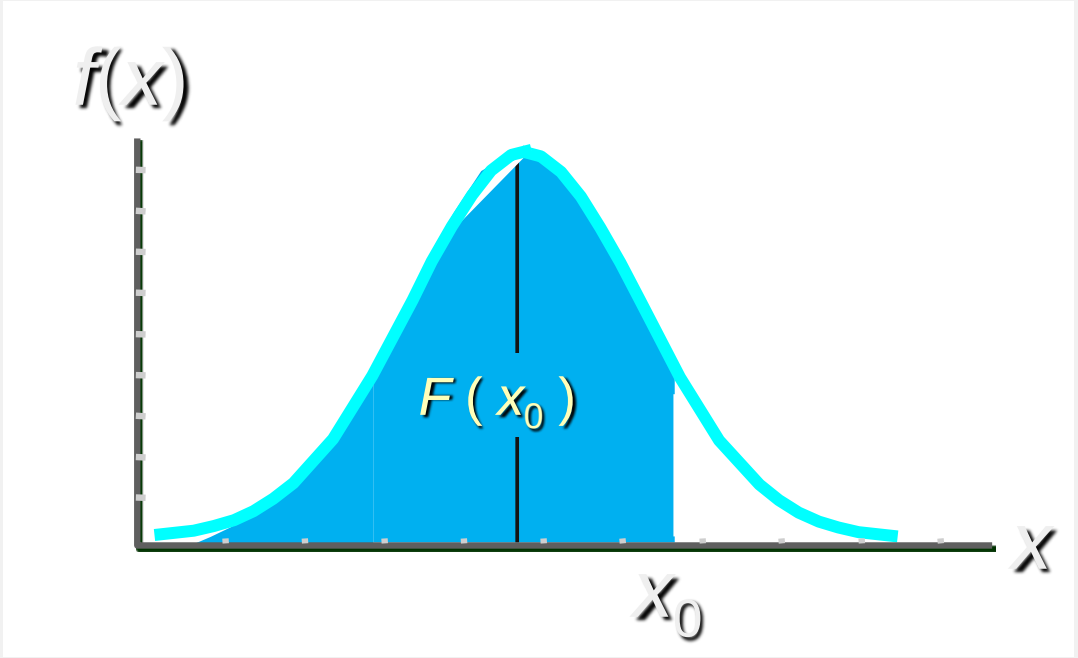
$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \qquad (-\infty < x < +\infty)$$

3. 根据分布函数，P(a<X<b)可以写为

$$P(a < X < b) = \int_{a}^{b} f(x)\mathrm{d}x = F(b) - F(a)$$

分布函数与密度函数的图示

- 密度函数曲线下的面积等于1
- 分布函数是曲线下 $x \le x_0$ 的面积



2.4 总体特征数

离散型随机变量的数学期望(expected value)

- 在离散型随机变量X的一切可能取值的完备组中，各可能取值x_i与其取相对应的概率p_i乘积之和
- 描述离散型随机变量取值的集中程度
- 计算公式为

$$E(X)=\sum_{i=1}^{n}x_ip_i \qquad (X取有限个值)$$

$$E(X)=\sum_{i=1}^{\infty}x_ip_i \qquad (X取无穷个值)$$

离散型随机变量的方差(variance)

- 随机变量X的每一个取值与期望值的离差平方和的数学期望，记为$D(X)$
- 描述离散型随机变量取值的分散程度
- 计算公式为

$$D(X) = E[X - E(X)]^2$$

若X是离散型随机变量，则

$$D(X) = \sum_{i=1}^{\infty} \left[x_i - E(X)\right]^2 \cdot p_i$$

投掷一枚骰子，出现的点数是个离散型随机变量，其概率分布为如下。计算数学期望和方差

$X = x_i$	1	2	3	4	5	6
$P(X = x_i) = p_i$	1/6	1/6	1/6	1/6	1/6	1/6

解：数学期望为：

$$E(X) = \sum_{i=1}^{6} x_i p_i = 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5$$

方差为：

$$D(X) = \sum_{i=1}^{6} \left[x_i - E(X) \right]^2 \cdot p_i$$

$$= (1 - 3.5)^2 \times \frac{1}{6} + \cdots + (6 - 3.5)^2 \times \frac{1}{6} = 2.9167$$

连续型随机变量的期望和方差

1. **连续型随机变量的数学期望为**

$$E(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \mu$$

2. **方差为**

$$D(X) = \int_{-\infty}^{+\infty} \left[x - E(X)\right]f(x)\mathrm{d}x = \sigma^2$$

数学期望和方差的运算

$$E(c) = c$$

$$E(cY) = cE(Y)$$

$$E(c + Y) = c + E(Y)$$

$$E(cY + A) = cE(Y) + A$$

$$var(Y)=E[(Y-\mu)^2]$$

$$var(Y+c)= var(Y)$$

$$var(cY)= c^2var(Y)$$

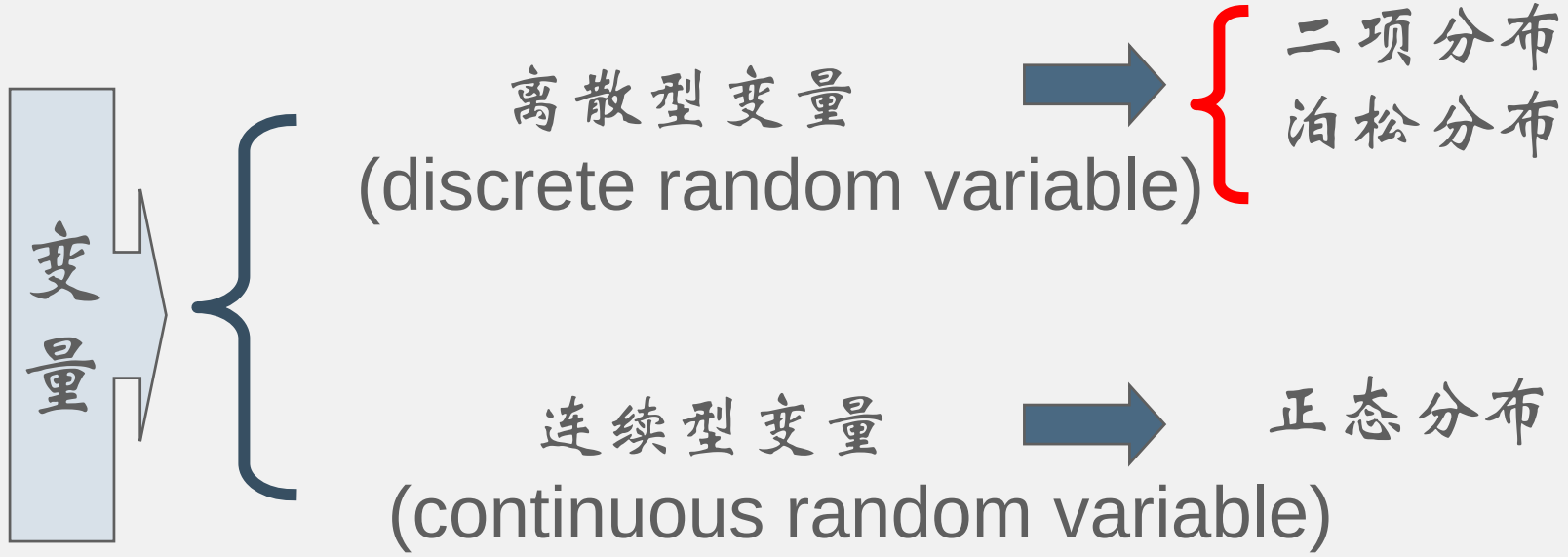
$$var(cY + A)= c^2var(Y)$$

2.5 几种常见的概率分布

- 二项分布(Binomial distribution)
- 泊松分布(Poisson distribution)
- 另外几种离散型概率分布
  - 超几何分布(hypergeometric distribution)
  - 负二项分布(negative binomial distribution)
- 正态分布(Normal distribution)
- 另外几种连续型概率分布
  - 指数分布(exponential distribution)
  - Γ分布(gamma distribution)
  - β分布（beta distribution)
  - 威布尔分布(Weibull distribution)
  - 均匀分布(uniform distribution)

几种常见的理论分布

随机变量的概率分布 (probability distribution)



二项分布

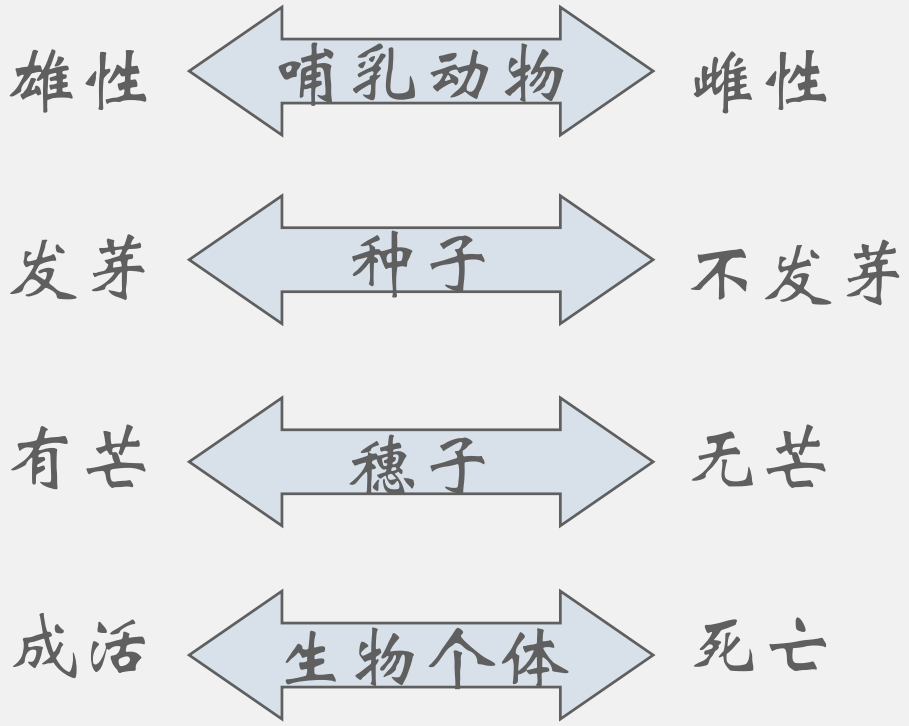
一、二项分布

二项分布是一种离散型随机变量 的分布，对于某个性状，常常可以把其资料分为两个类型。试验结果只能是“非此即彼”构成对立事件，将这种事件构成的总体称为二项总体，其概率分布称为二项分布。

一、二项分布的概率函数

离散型随机变量的分布

对立事件



非此即彼

设有一随机试验，每次试验结果出现且只出现对立事件 A 与 $\overline{A}$ 之一，这两种结果是互不相容的，在每次试验中出现 A 的概率是 p $(0<p<1)$，则出现对立事件$\overline{A}$的概率是$1-p=q$，则称这一串重复的独立试验为**n重贝努里试验，简称贝努里试验**(Bernoulli trials)。

在贝努里试验中，独立将此试验重复n次，求在n次试验中，一种结果A出现x次的概率P(x)是多少?

在种子发芽试验中，设事件A为“种子发芽”，则$\overline{A}$为“种子不发芽”。取4粒种子（n=4）来做试验，求有2粒种子发芽(x=2)的概率。

在4次试验中，事件A发生2次的方式有以下 C_4^2 种：

$$A_1A_2\overline{A}_3\overline{A}_4 \qquad A_1\overline{A}_2A_3\overline{A}_4 \qquad A_1\overline{A}_2\overline{A}_3A_4$$

$$\overline{A}_1A_2A_3\overline{A}_4 \qquad \overline{A}_1A_2\overline{A}_3A_4 \qquad \overline{A}_1\overline{A}_2A_3A_4$$

其中A_x (x=1，2，3，4)表示第x粒种子发芽，p为种子发芽的概率；$\overline{A_x}$ (x=1，2，3，4)表示第x粒种子不发芽，q为种子不发芽的概率，所以q=1-p。

由于试验是独立的，按概率的乘法法则，于是有：

$$P(A_1A_2\,\overline{A_3}\,\overline{A_4})=P(A_1)\cdot P(A_2)\cdot P(\overline{A_3})\cdot P(\overline{A_4})=p^2q^2$$

又由于以上各种方式中，任何二种方式都是互不相容的，按**概率的加法法则**，在4 粒种子中正好有2粒种子发芽的概率为：

$$P_4(2) = P(A_1A_2\overline{A}_3\overline{A}_4) + P(A_1\overline{A}_2A_3\overline{A}_4) + \cdots$$

$$+ P(\overline{A}_1\overline{A}_2A_3A_4) = C_4^2 p^2 q^{4-2}$$

一般，在*n*重贝努利试验中，事件*A*恰好发生*k*(0≤*k*≤n)次的概率为

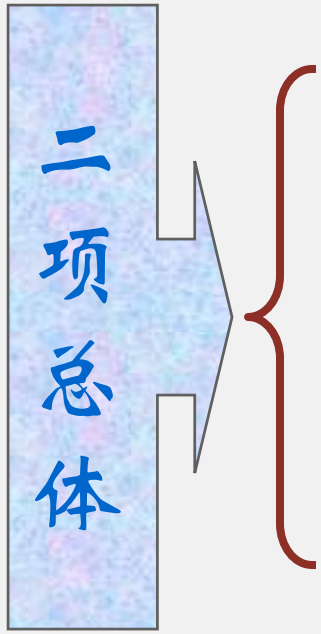
$$P_n(\mathrm{x}) = C_{\mathrm{n}}^{x} p^{x} q^{n-x} \qquad x=0，1，2\ldots，n$$

若把上式与二项展开式

$$(q+p)^n = \sum_{x=0}^{n} C_n^x p^x q^{n-x}$$

相比较就可以发现，在*n*重贝努里试验中，事件*A*发生*x*次的概率恰好等于展开式中的第*x*+1项，所以把P(x)称为随机变量X服从参数为n和p的**二项分布**(binomial distribution)，也称为贝努里分布，记作B(n，p)。这种“非此即彼”的事件所构成的总体称为**二项总体**。

二项分布的两个条件：



试验只有两个对立结果，记为A和$\bar{A}$，出现概率分别为p和q=1-p。

重复性：每次试验条件不变时，事件A出现为恒定概率p；

独立性：任何一次试验中事件A的出现与其余各次试验结果无关。

公式

$$p(x) = C_n^x \cdot p^x \cdot q^{n-x}$$

称作**二项分布概率函数**，其中 $C_n^x = \frac{n!}{x!(n-x)!}$， p>0，q>0，

p+q=1，x是一个离散型随机变量，取值为0，1，2，…，n。

$$(p+q)^n = C_n^0 \cdot p^0 \cdot q^n + C_n^1 \cdot p^1 \cdot q^{n-1} + C_n^2 \cdot p^2 \cdot q^{n-2} + \cdots + C_n^x \cdot p^x \cdot q^{n-x} + \cdots + C_n^n \cdot p^n \cdot q^0$$

$$= \sum_{x=0}^{n} C_n^x \cdot p^x \cdot q^{n-x}$$

例：n=试验次数（或样本含量） n=4

x=在n次试验中事件A出现的次数 x=2

p=事件A发生的概率（每次试验是恒定的） p=0.9

1-p=事件A不发生的概率 1-p=0.1

p(x)=X的概率函数=P(X=x) P(2)

则4粒种子有两粒发芽的概率为：

$P(x)=C_4^2 p^2 q^2=6\times0.9^2\times0.1^2=0.0486$

由于二项式中p+q=1，

$$(p+q)^n = 1$$

或者n个事件构成一个完全事件系，所以有：

$$p(0) + p(1) + p(2) + \ldots + p(x) + \ldots + p(n) = 1$$

$$\sum_{x=0}^{n} P(x) = 1$$

现已求出某事件发生的概率，若试验N次，则该事件发生的理论次数为：

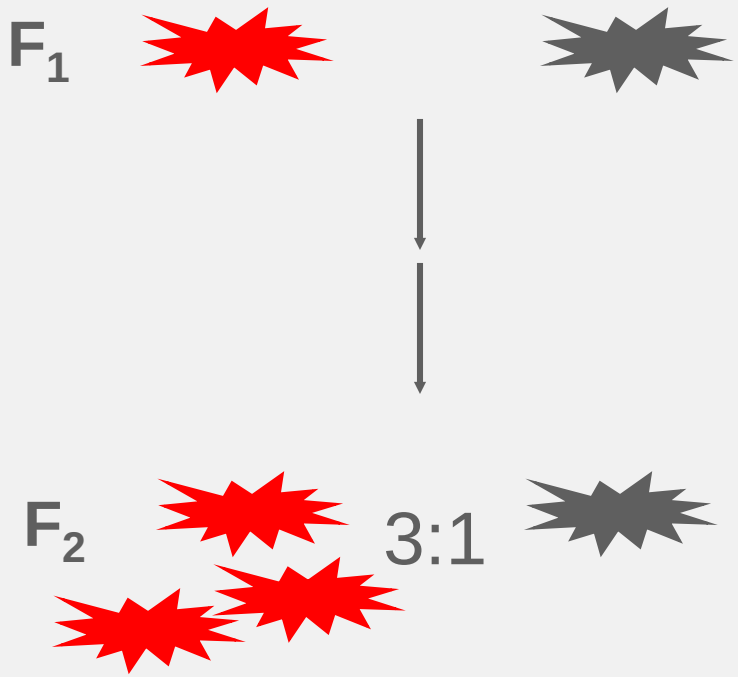
理论次数＝NP(x)

二项分布的概率累积函数为：

$$F(x)=\sum P(x)=1$$

二项分布的计算

例：豌豆红花和白花杂交后， 在F_2红花：白花＝3：1



若每次观察4株，共观察100次，问得红花为0、1、2、3、4株的概率各为多少？

观察4株出现红花的概率分布表 (p=0.75 q=1-p=0.25)

概率函数	$C_n^x p^x q^{n-x}$	P(x)	F(x)	NP(x)
P(0)	$C_4^0 p^0 q^4$	0.0039	0.0039	0.39
P(1)	$C_4^1 p^1 q^3$	0.0469	0.0508	4.69
P(2)	$C_4^2 p^2 q^2$	0.2109	0.2617	21.09
P(3)	$C_4^3 p^3 q^1$	0.4219	0.6836	42.19
P(4)	$C_4^4 p^4 q^0$	0.3164	1.000	31.64
合计		1.000		100

例2：鸡蛋孵化率为p=0.9，每次选5个进行孵化，试求孵出小鸡的各种可能概率，若做1000次试验，其理论次数分别为多少？

孵化小鸡的概率分布表(p= 0.90 q=0.10)

概率函数	$C_n^x p^x q^{n-x}$	P(x)	F(x)	NP(x)
P(0)	$C_5^0 p^0 q^5$	0.00001	0.00001	0.01
P(1)	$C_5^1 p^1 q^4$	0.00045	0.00046	0.45
P(2)	$C_5^2 p^2 q^3$	0.0081	0.00856	8.1
P(3)	$C_5^3 p^3 q^2$	0.0729	0.08046	72.9
P(4)	$C_5^4 p^4 q^1$	0.32805	0.40951	328.05
P(5)	$C_5^5 p^5 q^0$	0.59049	1.0000	590.49

二项分布的形状和参数

二项分布的**形状**由n和p两个参数决定。B(n，p)

(1)当p值较小且n不大时，分布是偏倚的。随n的增大，分布趋于对称；

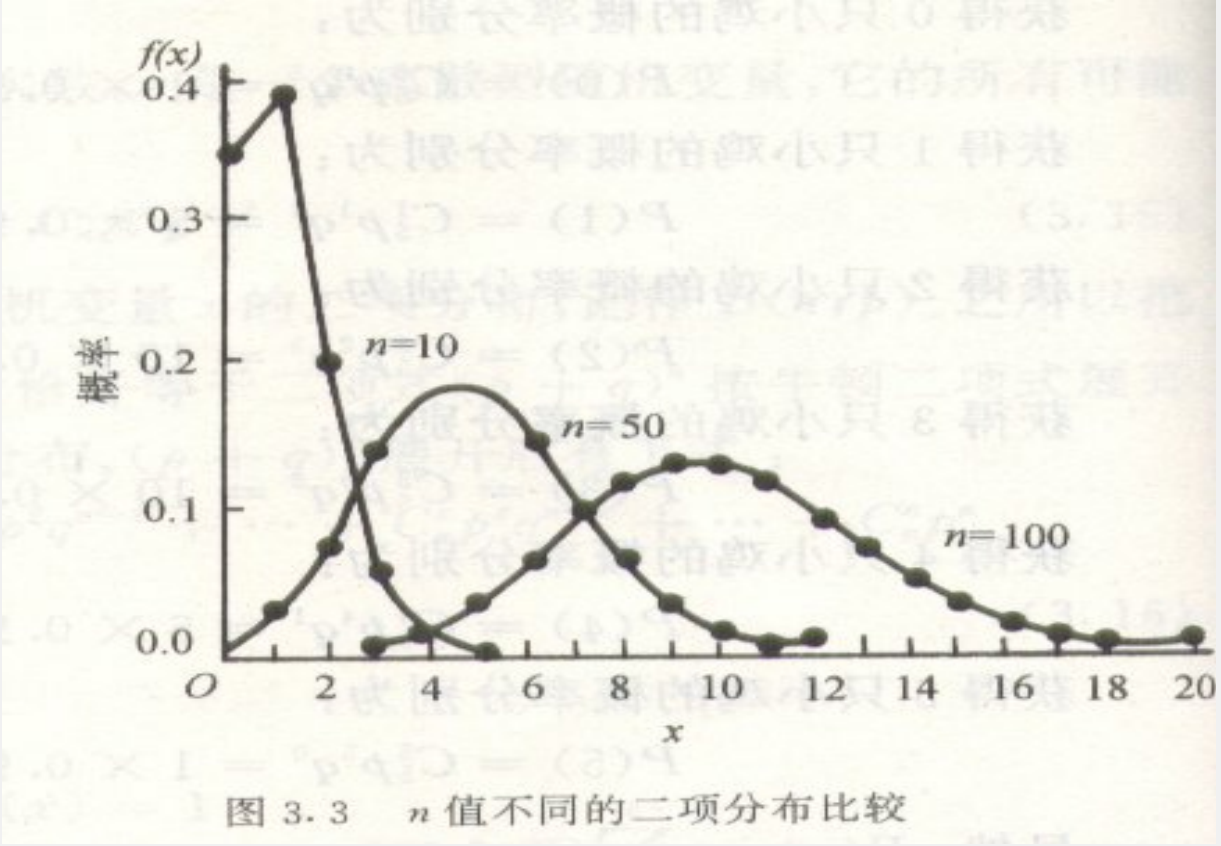


图 3.3　n 值不同的二项分布比较

二项分布的形状和参数

（2）当p值趋于0.5时，分布趋于对称。

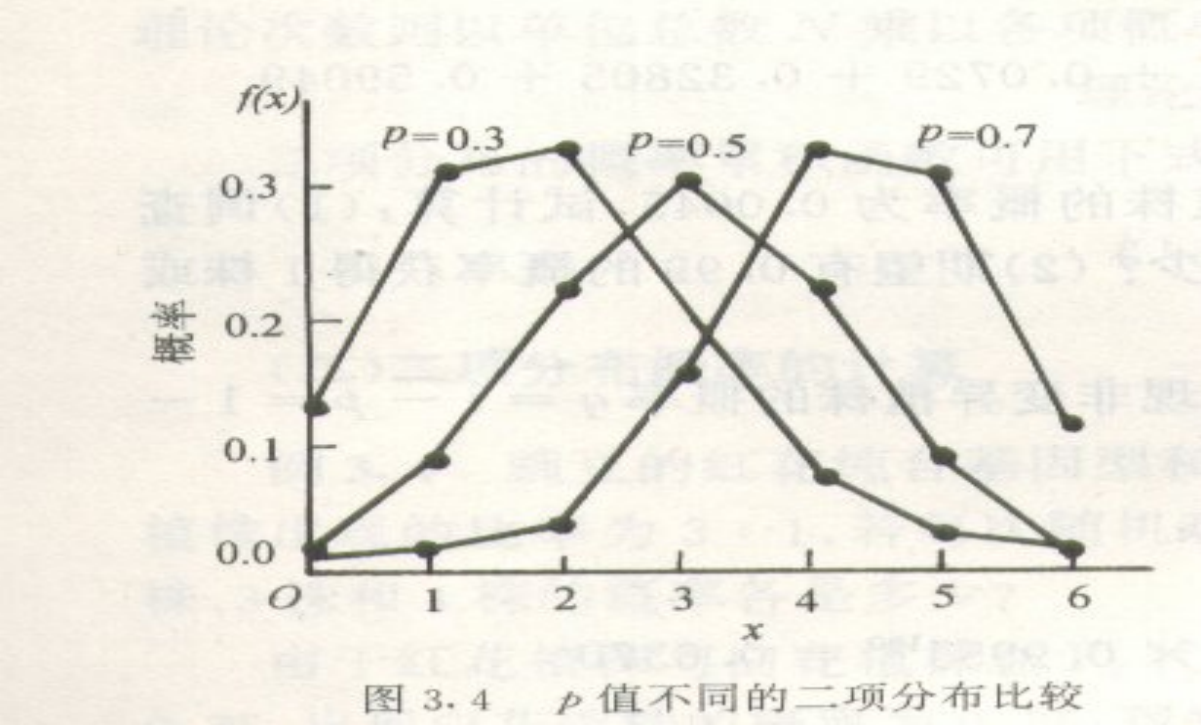


图 3.4　p 值不同的二项分布比较

二项分布的形状和参数

统计学证明，服从二项分布B(n，p)的随机变量所构成的总体的**平均数μ**、**标准差σ**与n、p这两个参数有关。

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

例：豌豆红花纯合基因型和白花纯合基因型杂交后，在F2代红花植株与白花植株出现的比例为3:1。每次观察4株，**n=4，红花出现概率为p=3／4＝0.75。**

红花出现株数　**0，1，2，3，4**

（1）红花出现的平均株数 $\mu = np = 3.0$ (株）

（2）标准差 $\sigma = \sqrt{np(1-p)} = 0.8660$（株）

泊松分布

泊松分布(Poisson distribution)

波松分布是一种 可以用来描述和分析随机地发生在单位空间或 时间里的**稀有事件**的概率分布。要观察到这类事件，样本含量 *n* 必须很大 。

在生物 、医学研究中，服从波松分布的随机变量是常见的。如， 一定畜群中某种患病率很低的非传染性疾病患病数或死亡数，畜群中遗传的畸形怪胎数， 每升饮水中大肠杆菌数，计数器小方格中血球数， 单位空间中某些野生动物或昆虫数等，都是服从波松分布的。

泊松分布的意义

若随机变量 $x(x=k)$ 只取零和正整数值 0，1，2，…，且其概率分布为

$$P(x=k)=\frac{\lambda^k}{k!}e^{-\lambda}\ ,\qquad k=0，1，\ldots\ldots \quad (4\text{-}23)$$

其中 λ＞0；e=2.7182… 是自然对数的底数，则 称 x 服 从 参 数 为 λ 的 波 松分布(Poisson‘s distribution)，记 为 $x\sim P(\lambda)$。

$$P(x=k)=\frac{\lambda^k}{k!}e^{-\lambda},$$ **k=0，1，……**

1 对于小概率事件，可用泊松分布描述其概率分布。

2 二项分布当p<0.1和np<5时，可用泊松分布来近似。

泊松分布的特点

平均数和方差相等，都等于常数 λ，即

$$\mu=\sigma^2=\lambda$$

调查某种猪场闭锁育种群仔猪畸形数，共记录200窝， 畸形仔猪数的分布情况如表4-3所示。试判断畸形仔猪数是否服从波松分布。

表4-3 畸形仔猪数统计分布

每窝畸形数k	0	1	2	3	>3	合计
窝数f	120	62	15	2	1	200

表4-3 畸形仔猪数统计分布

每窝畸形数k	0	1	2	3	>3	合计
窝数f	120	62	15	2	1	200

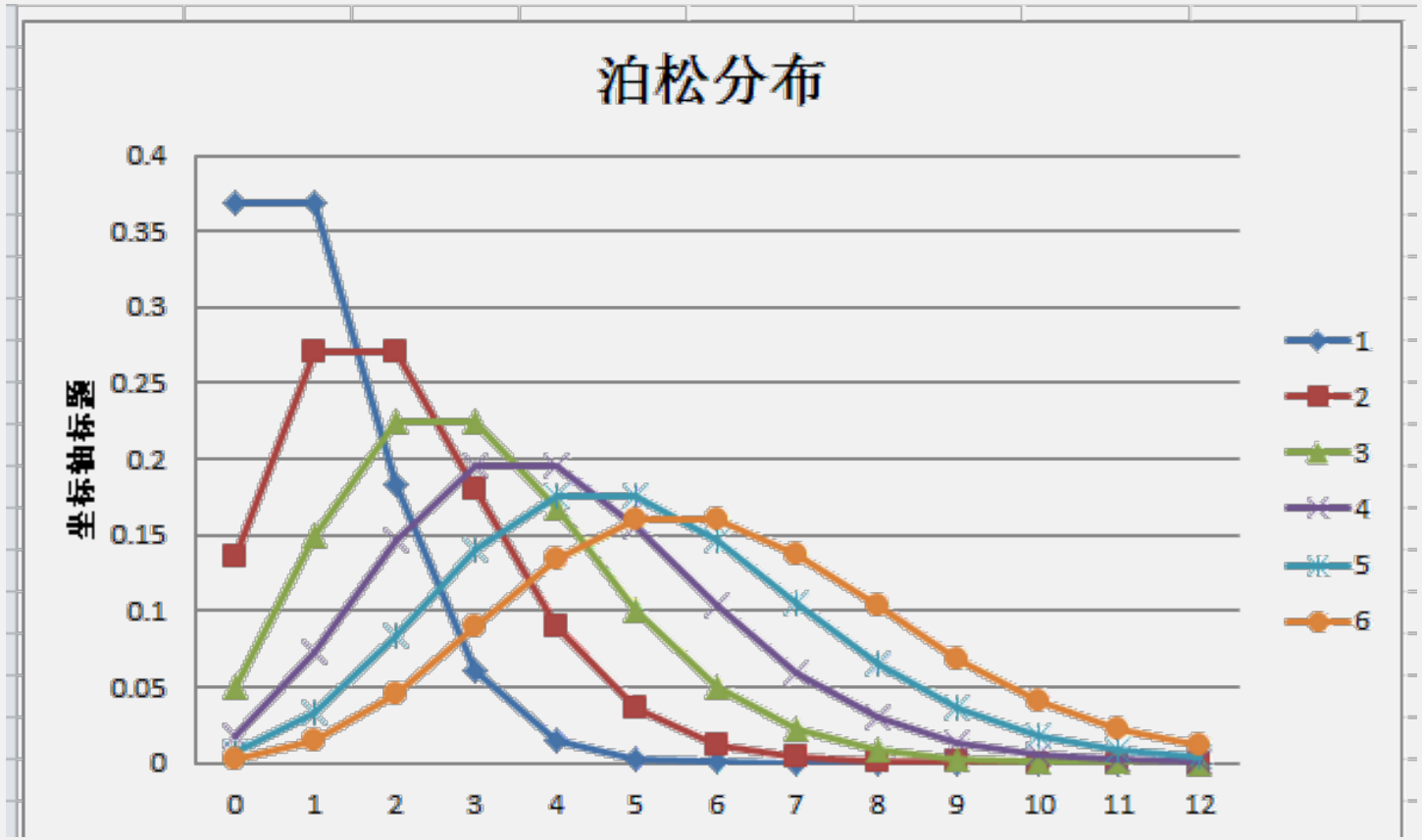
样本均数 ($\bar{x}$)和方差 (S^2)计算结果如下：

$\bar{x}$ =Σfk/n =(120×0+62×1+15×2+2×3+1×4)/200=0.51

$$S^2=\frac{\sum fk^2-(\sum fk)^2/n}{n-1}=0.52$$

$\bar{x}$ =0.51，S²=0.52,这两个数是相当接近的，因此可以认为畸形仔猪数服从波松分布。

P(λ)的形状由λ确定



- λ 较小时，泊松分布偏倚。
- λ 增大时，泊松分布趋于对称。
- λ 无限增大时，泊松分布接近正态分布。

λ是波松分布所依赖的唯一参数。 λ值愈小分布愈偏倚，随着λ的增大 ，分布趋于对称。当λ=20时分布接近于正态分布；当λ=50时， 可以认为波松分布呈正态分布。 所以在实际工作中，**当λ≥20时就可以用正态分布来近似地处理波松分布的问题**。

波松分布的概率计算

波松分布的概率计算，依赖于参数 λ的确定，只要参数λ确定了，把k=0，1，2，… 代入公式即可求得各项的概率。

但是在大多数服从波松分布的实例中，分布参数λ往往是未知的，只能从所观察的随机样本中计算出相应的样本平均数作为 λ 的 估计值，将其代替公式中的λ，计算出 k = 0，1，2，… 时的各项概率。

如例中已判断畸形仔猪数服从波松分布，并已算出样本平均数=0.51。将0.51代替公式中的λ得：

$$P(x=k)=\frac{0.51^{k}}{k!}e^{-0.51},\qquad (k=0,1,2,\ldots)$$

因为$e^{-0.51}=1.6653$，所以畸形仔猪数各项的概率为：

$P(x=0)=0.510 / (0!\times 1.6653)=0.6005$

$P(x=1)=0.511 / (1!\times 1.6653)=0.3063$

$P(x=2)=0.512 / (2!\times 1.6653)=0.0781$

$$P(x=3)=0.513\ /\ (3!\times 1.6653)=0.0133$$

$$P(x=4)=0.514\ /\ (4!\times 1.6653)=0.0017$$

$$P(x>4)=1-\sum_{k=0}^{4}p(x=k)=1\quad 0.9999=0.0001$$

将实际计算得的频率与根据λ=0.51的泊松分布计算的概率相比较， 发现畸形仔猪的频率分布与 λ=0.51的 波松分布是吻合得很好的 。这进一步说明了畸形仔猪数是服从波松分布的。

把上面各项概率乘以总观察窝数(*n*=200)即得各项按波松分布的

窝数。 波松分布与相应的频率分布列于表中。

每窝畸形数 k	0	1	2	3	≥4	合计
窝　数	120	62	15	2	1	200
频　率	0.6000	0.3100	0.0750	0.0100	0.0050	1.00
概　率	0.6005	0.3063	0.0781	0.0133	0.0018	1.00
理论窝数	120.12	61.26	15.62	2.66	0.34	200

为监测饮用水的污染情况， 现检验某社区每毫升饮用水中细菌数， 共得400个记录如下：

1ml水中细菌数	0	1	2	≥3	合　计
次数f	243	120	31	6	400

试分析饮用水中细菌数的分布是否服从波松分布。若服从，按波松分布计算每毫升水中细菌数的概率及理论次数并将频率分布与波松分布作直观比较。

经计算得每毫升水中平均细菌数 $\bar{x}$ =0.500,方差S^2=0.496。两者很接近

故可认为每毫升水中细菌数服从波松分布。以=0.500代替公式

$$P(x=k)=\frac{0.5^k}{k!}e^{0.5},$$

计算结果如表所示。

可见细菌数的频率分布与λ=0.5的波松分布是相当吻合的， 进一步说明用波松分布描述单位容积(或面积)中细菌数的分布是适宜的。

表 细菌数的波松分布

1ml水中细菌数	0	1	2	≥3	合　计
实际次数	243	120	31	6	400
频　率	0.6075	0.3000	0.0775	0.0150	1.00
概　率	0.6065	0.3033	0.0758	0.0144	1.00
理论次数	242.60	121.32	30.32	5.76	400

注意，二项分布的应用条件也是波松分布的应用条件。比如二项分布要求n 次试验是相互独立的，这也是波松分布的要求。然而一些具有传染性的罕见疾病的发病数，因为首例发生之后可成为传染源，会影响到后续病例的发生，所以不符合波松分布的应用条件。对于在单位时间、单位面积或单位容积内，所观察的事物由于某些原因分布不随机时，如细菌在牛奶中成集落存在时，亦不呈波松分布。

对于波松分布，当$\lambda \to \infty$时 ，波松分布以正态分布为极限。

在实际计算中， 当 $\lambda \geq 20$ (也有人认为$\lambda \geq 6$)时，用波松分布中的λ代替正态分布中的μ及σ^2 ，即可由后者对前者进行近似计算。

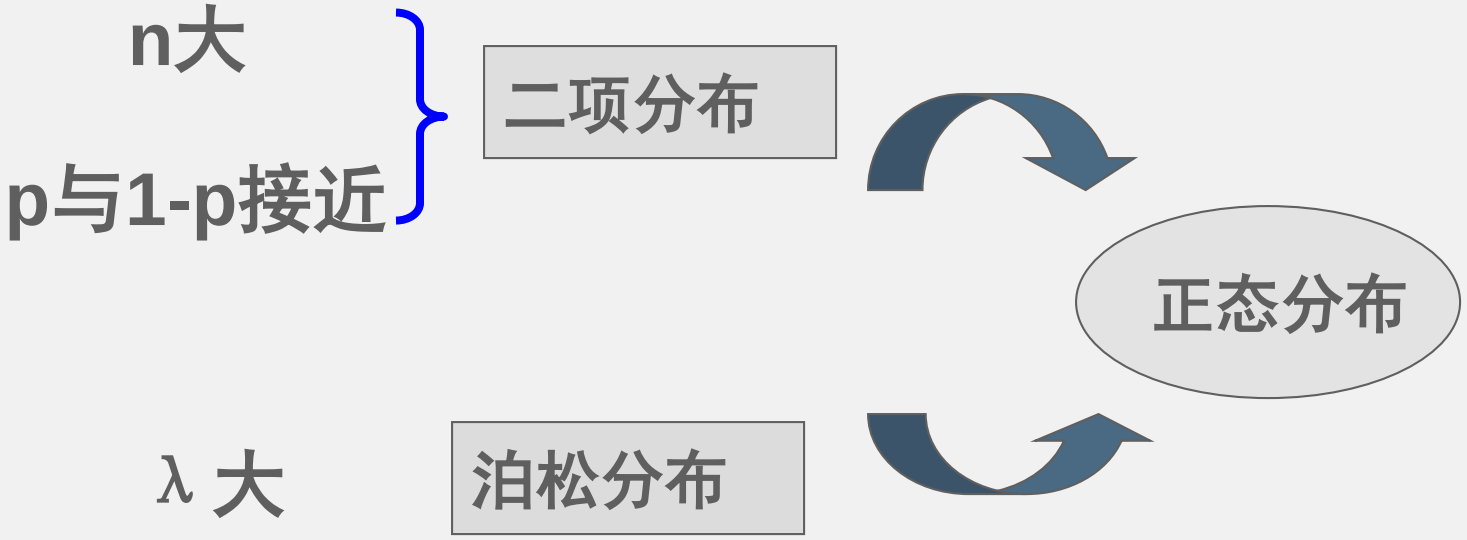
正 态 分 布

正态分布（normal distribution）

正态分布也称为高斯分布(Gauss distribution)。

特点

围绕在平均值左右，由平均值到分布的两侧，
变量数减少，即**两头少，中间多，两侧对称。**



正态分布是生物统计学的重要基础。

正态分布的概率函数

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(-\infty<x<\infty)$$

记N (μ，σ²)

f(x) 为正态分布的概率密度函数，表示某一定x值出现的概率密度函数值。

μ 总体平均数　　σ 总体标准差

π 圆周率，3.14159　　e为自然对数底，2.71828

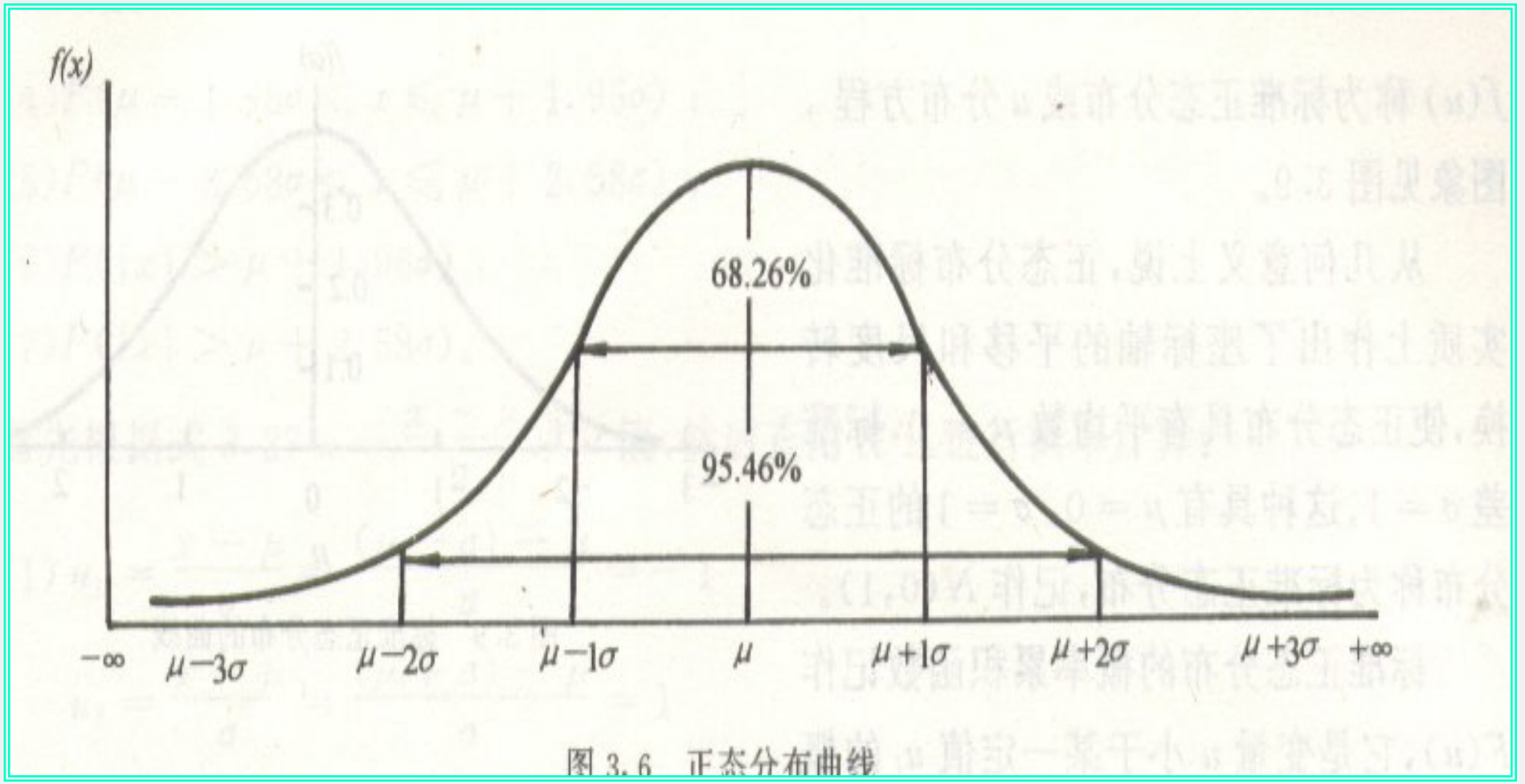


图 3.6 正态分布曲线

图3.1 正态分布曲线

正态分布的特征

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(-\infty<x<\infty)$$

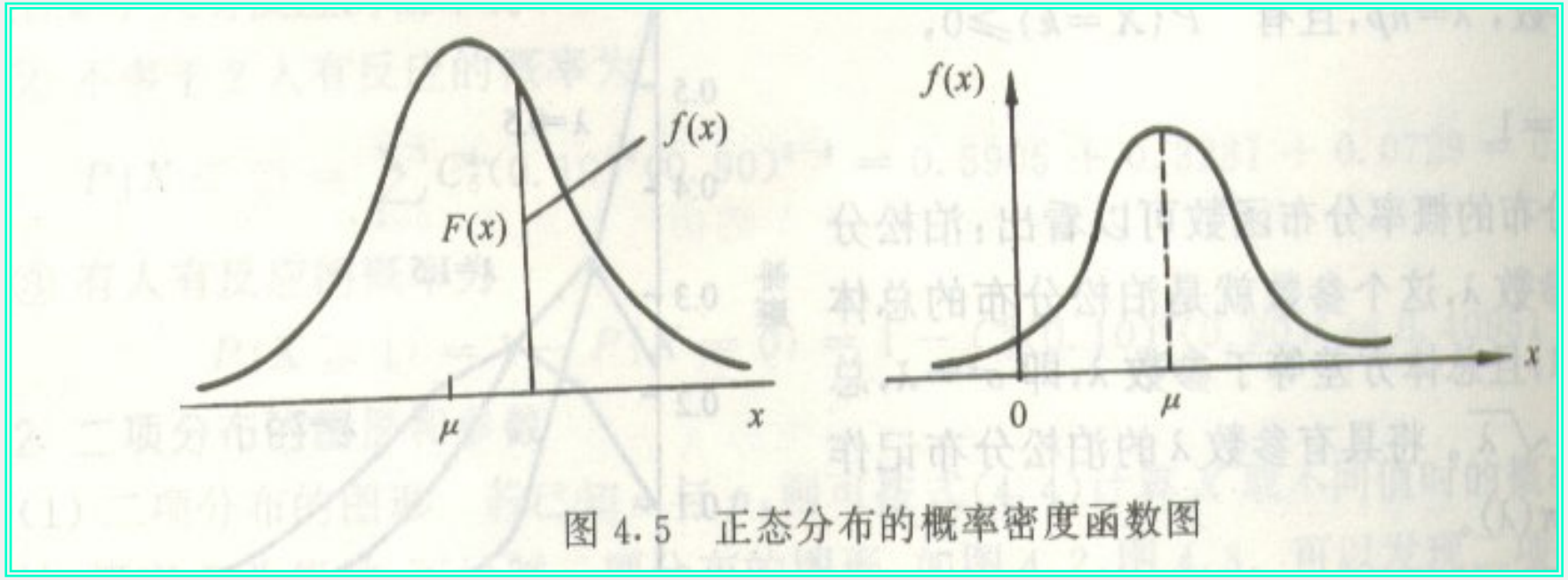


图 4.5　正态分布的概率密度函数图

1　x= μ 时，f(x)值最大，正态分布曲线以平均数 μ 为中心的分布。

正态分布的特征

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(-\infty<x<\infty)$$

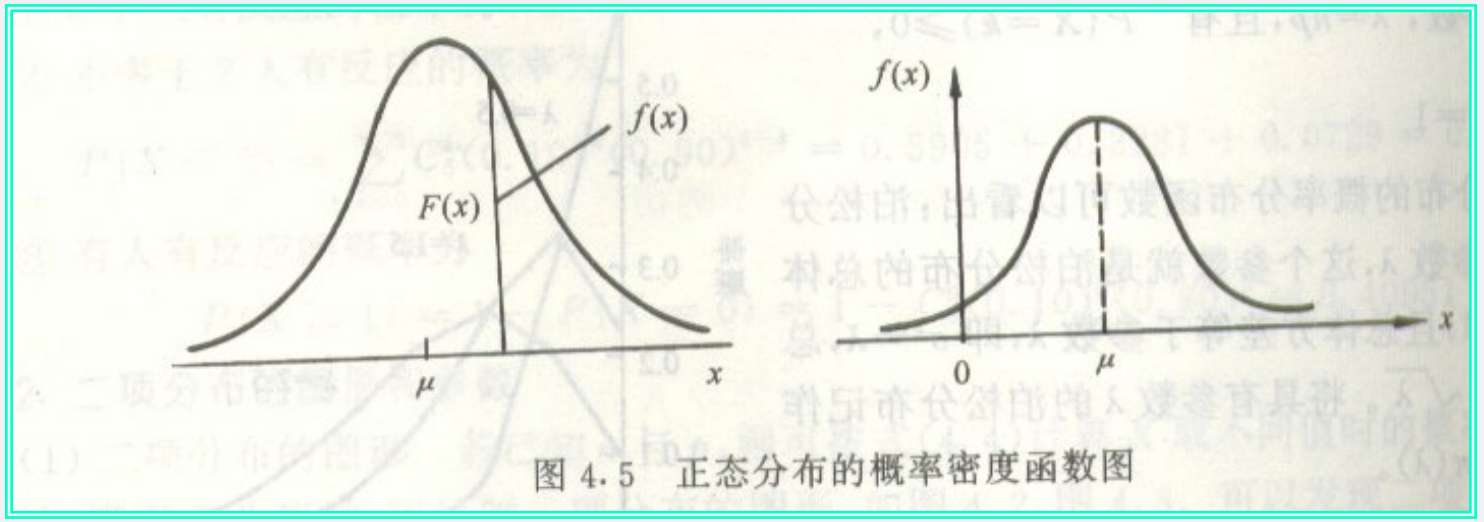


图 4.5　正态分布的概率密度函数图

2　x－μ的绝对值相等时，f(x)也相等，正态分布密度曲线以μ为中心向左右两侧对称。

正态分布的特征

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(-\infty<x<\infty)$$

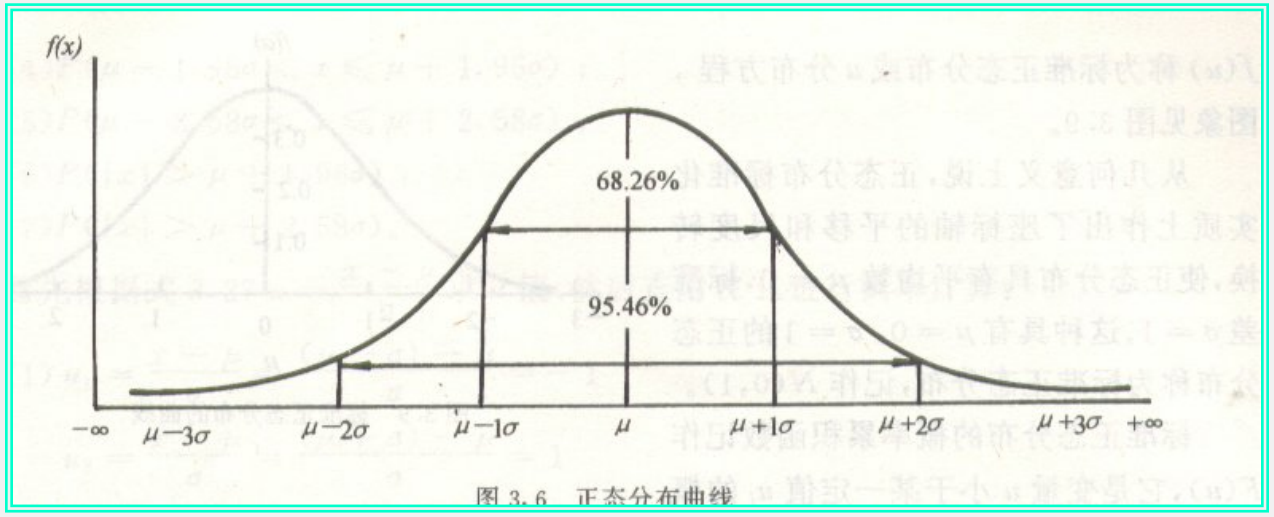


图 3.6 正态分布曲线

3 f(x)是非负函数，以x轴为渐近线，x的取值区间为(−∞，+∞)。

正态分布的特征

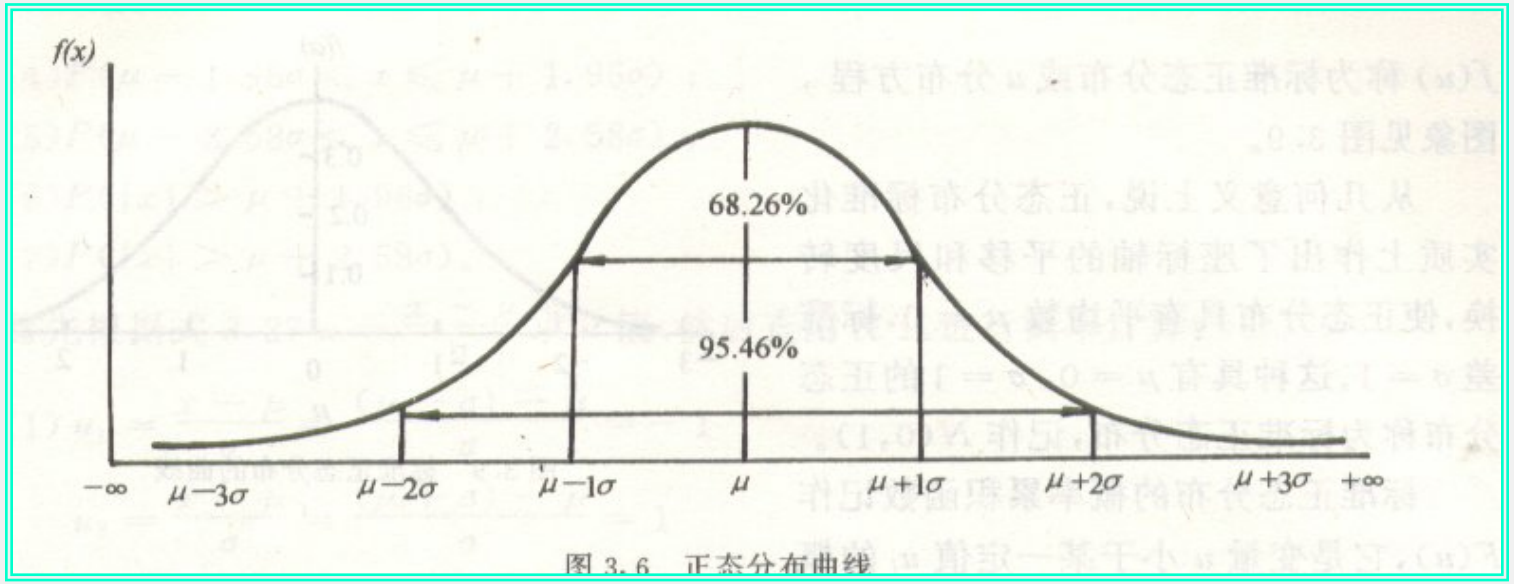


图 3.6 正态分布曲线

4

正态分布曲线在x= μ ± σ 处各有一个拐点，曲线通过拐点时改变弯曲度。

正态分布的特征

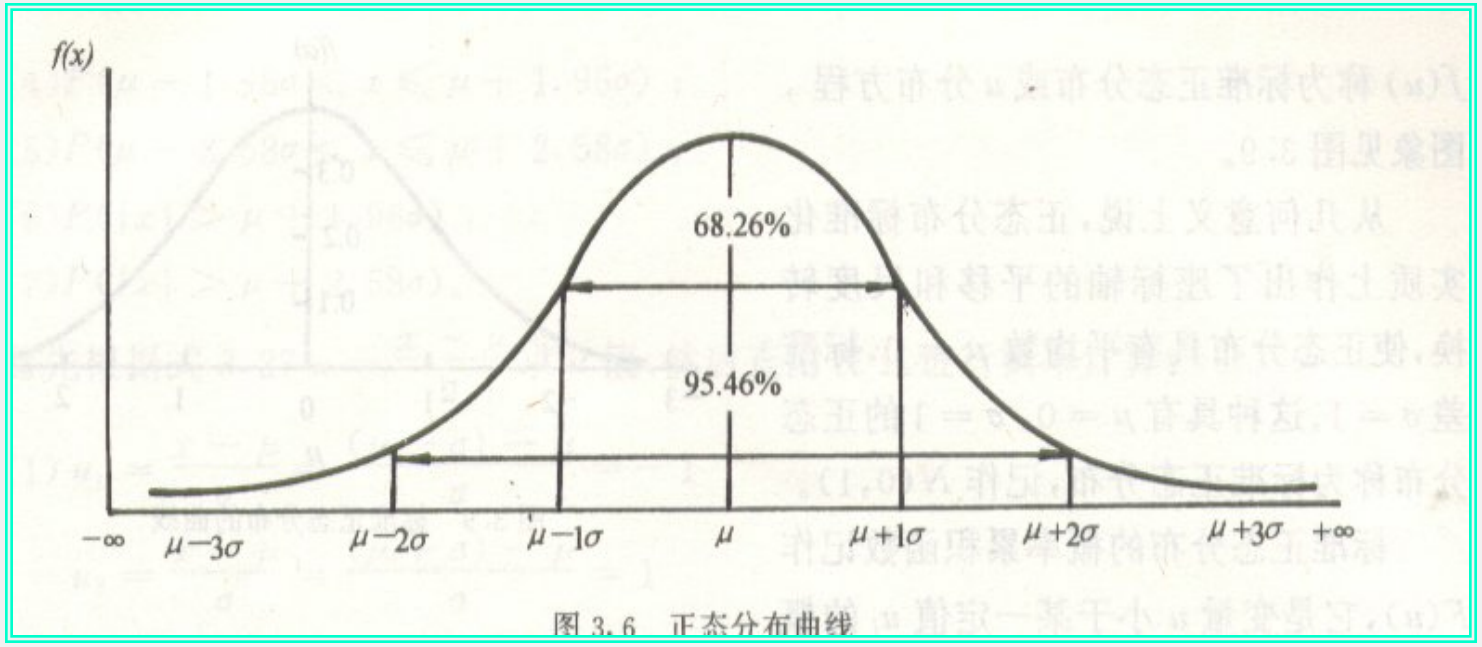


图 3.6　正态分布曲线

5　分布曲线与x轴围成的全部面积为1

正态分布的特征

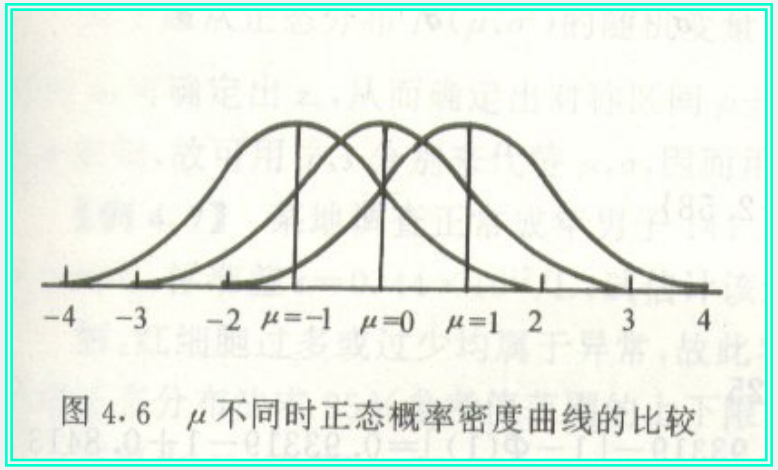


图 4.6　μ不同时正态概率密度曲线的比较

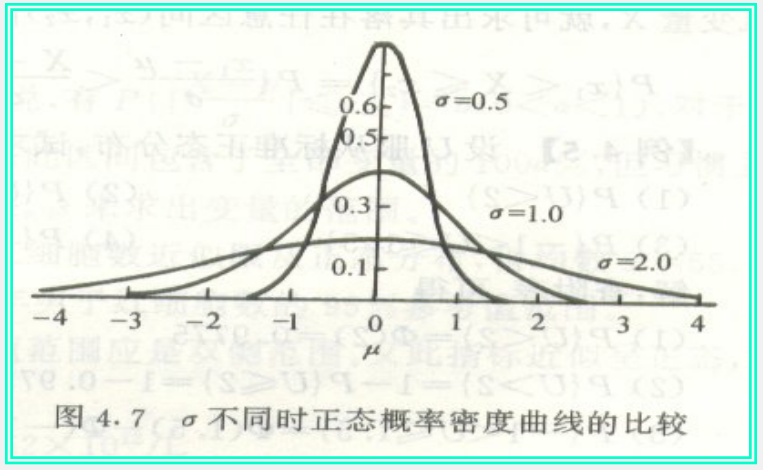


图 4.7　σ不同时正态概率密度曲线的比较

6

正态分布曲线由参数μ，σ决定，μ确定正态分布曲线在X轴上的中心位置，σ确定正态分布的变异度。

（三）标准正态分布

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(-\infty<x<\infty)$$

$$N(\mu,\ \sigma^2)$$

正态分布是依赖于参数$(\mu,\ \sigma^2)$的一个曲线系，正态曲线的位置及形态随$(\mu,\ \sigma^2)$的不同而不同，这就给研究具体的正态分布总体带来了困难，我们现将其标准化。

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ (-\infty<x<\infty)$$

$N(\mu,\ \sigma^2)$

$$f(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$$

f(u)称为标准正态分布(standard normal distribution)或u分布方程。

$$u=\frac{x-\mu}{\sigma}$$

u表示标准正态离差（standard normal deviate），它表示离开平均数 μ 有几个标准差 σ 。

$N(0,\ 1)$

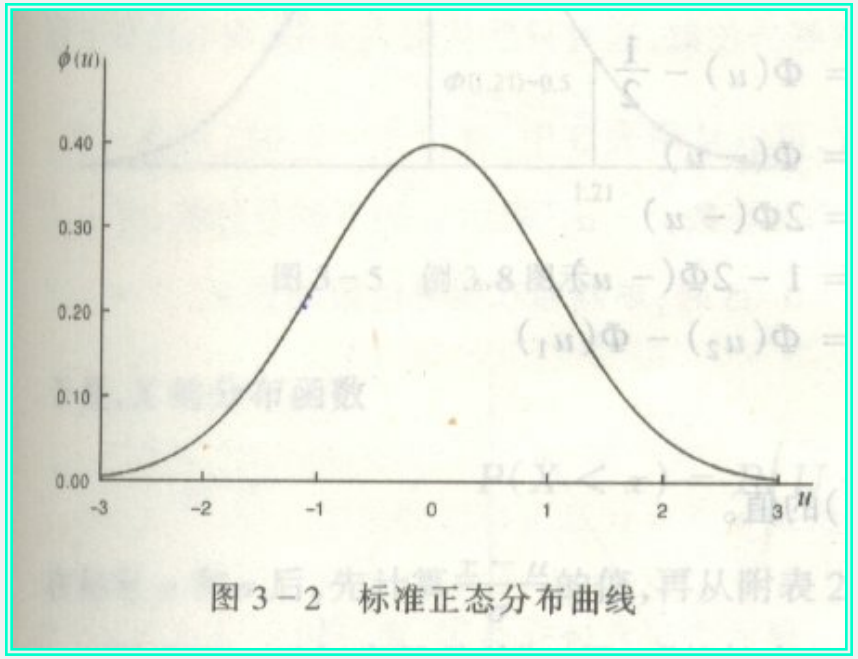


图 3－2　标准正态分布曲线

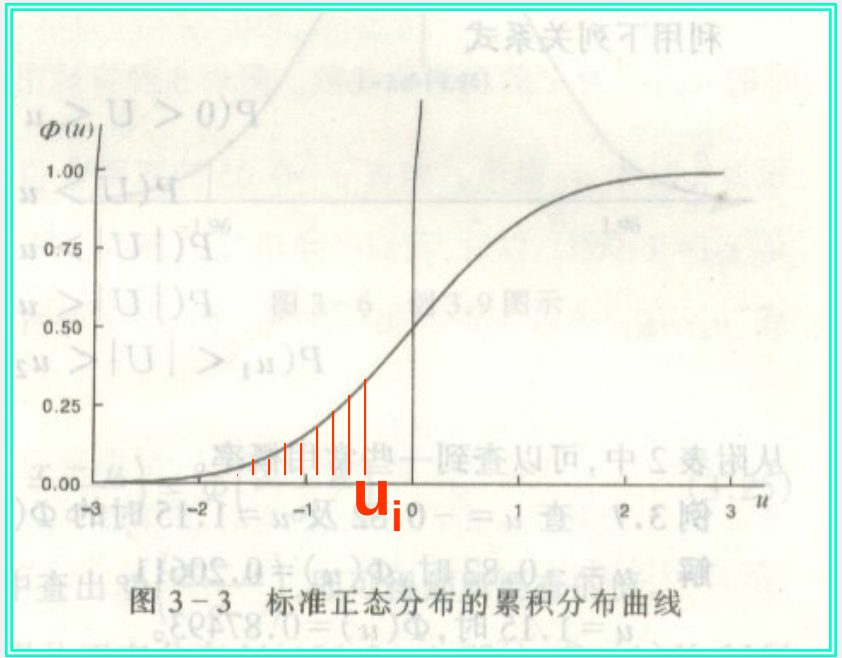


图 3－3　标准正态分布的累积分布曲线

标准正态分布的概率累积函数记作F(u)，它是变量u小于某一定值的概率。

$$F(u_i) = P(u < u_i) = \int_{-\infty}^{u_i} f(u)du$$

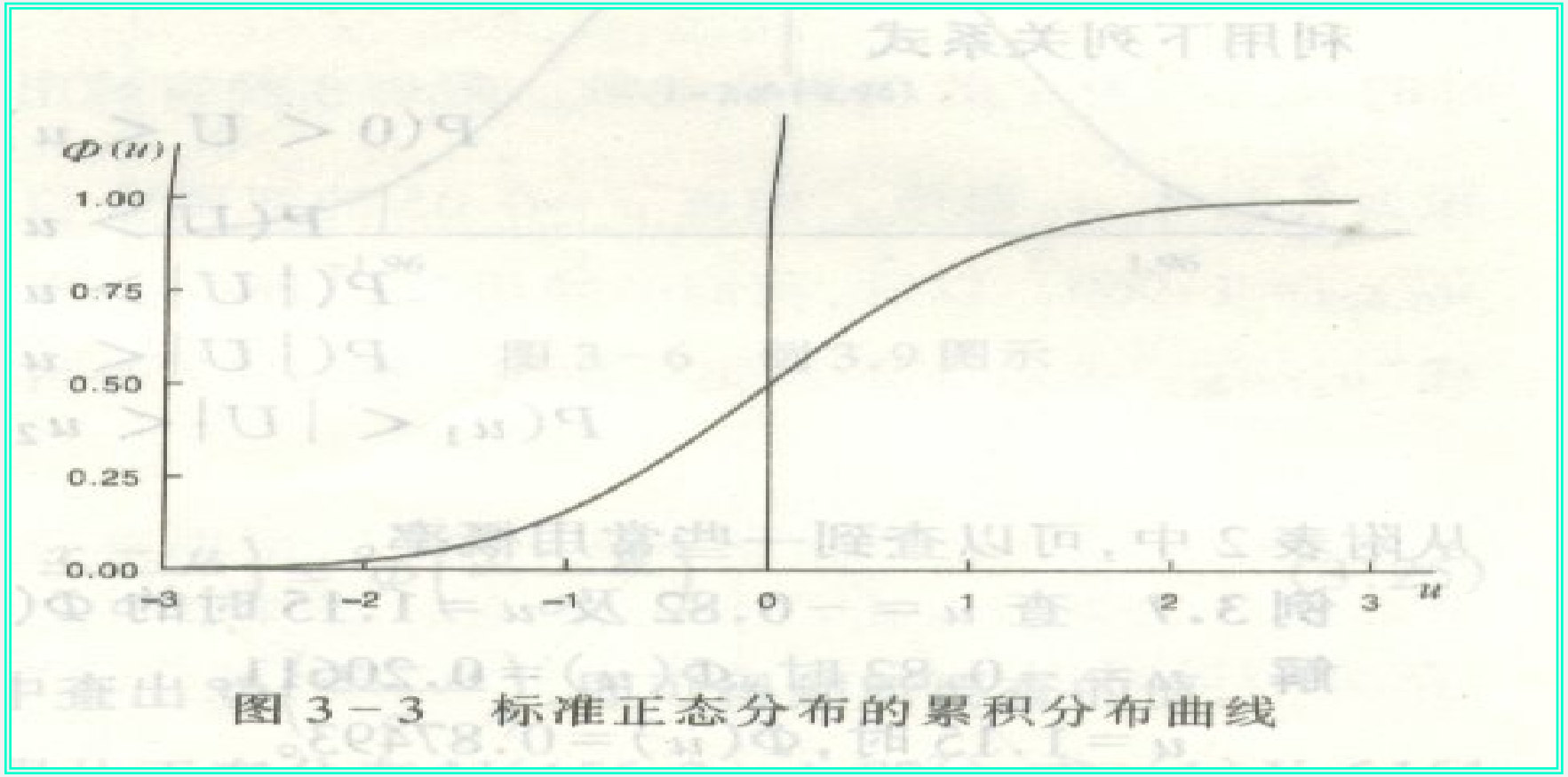


图 3－3　标准正态分布的累积分布曲线

为了计算方便，对于不同的u值，计算出不同的F(u)，编成函数表，称为正态分布表，从中可以查到u任意一个区间内取值的概率。

标准正态分布u落在区间[a，b]的概率

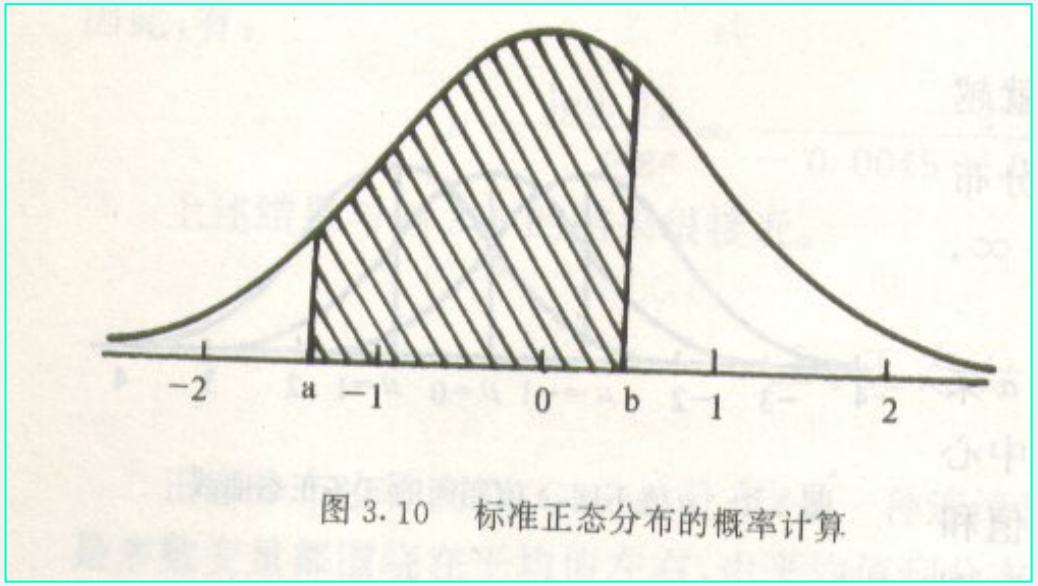


图 3.10　标准正态分布的概率计算

$$P(a \le u \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

$$= \int_{-\infty}^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du - \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

$$= F(b) - F(a)$$

（四）正态分布的概率计算

1 一般正态分布的概率计算

$$P(0 \le u < a) = F(a) - 0.5$$

$$P(u \ge a) = 1 - F(a) = F(-a)$$

$$P(|u| \ge a) = 2F(-a)$$

$$P(|u| < a) = 1 - 2F(-a)$$

$$P(a \le u < b) = F(b) - F(a)$$

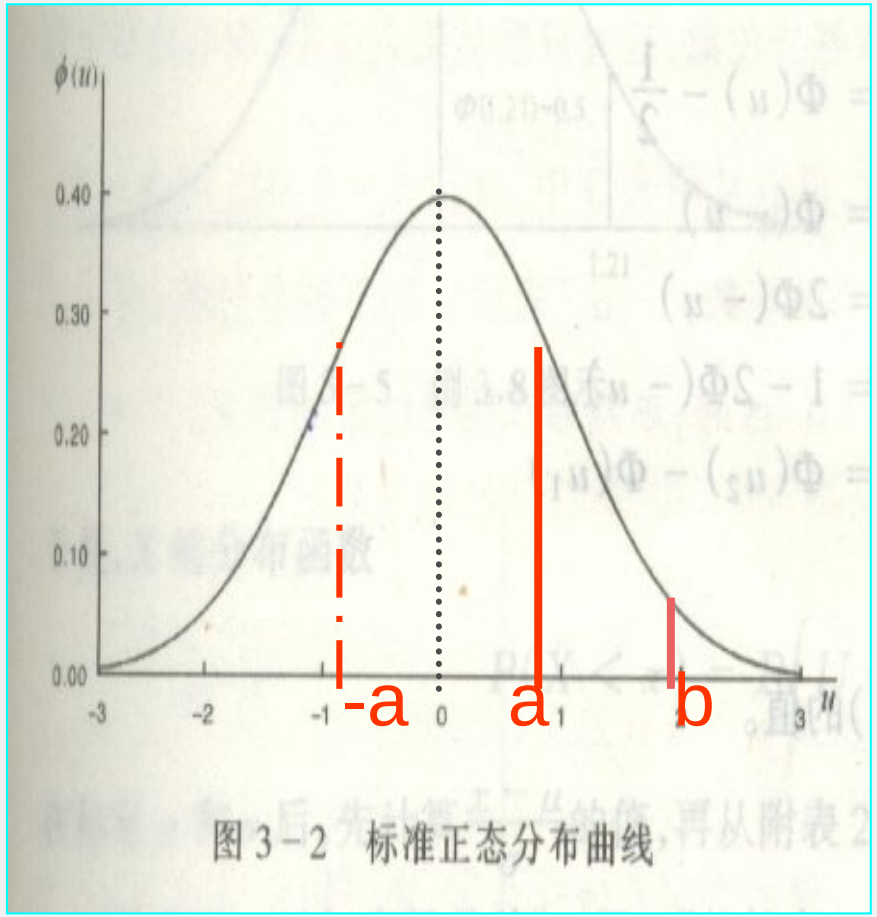


图 3-2 标准正态分布曲线

正态分布的概率计算

1一般正态分布的概率计算

计算一般正态分布的概率时，只要将区间的上下限作适当变换（标准化），就可用查标准正态分布的概率表的方法求得概率了。

服从正态分布$N(\mu, \sigma^2)$的随机变量，x的取值落在区间$[x_1, x_2]$的概率，记作$P(x_1 \leqslant x < x_2)$，等于服从标准正态分布的随机变量u在$[(x_1-\mu)/\sigma, (x_2-\mu)/\sigma]$内取值的概率。

1 一般正态分布的概率计算

$$P(x_1 \le x < x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$u = \frac{x-\mu}{\sigma} \qquad dx = \sigma du$$

$$P(x_1 \le x \le x_2) = P\left(\frac{x_1-\mu}{\sigma}_1 \le \frac{x-\mu}{\sigma} \le \frac{x_2-\mu}{\sigma}\right) = P(u_1 \le u \le u_2)$$

$$= \int_{u_1}^{u_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = F(u_2) - F(u_1)$$

$$P(\mu-\sigma<x\leqslant\mu+\sigma)=P(-1\leqslant u\leqslant 1)=0.6826$$

$$P(\mu-2\sigma<x\leqslant\mu+2\sigma)=P(-2\leqslant u\leqslant 2)=0.9545$$

$$P(\mu-3\sigma<x\leqslant\mu+3\sigma)=P(-3\leqslant u\leqslant 3)=0.9973$$

$$P(|x| \leqslant \mu + 1.96\sigma) = P(-1.96 \leqslant u \leqslant 1.96) = 0.95$$

$$P(|x| \leqslant \mu + 2.58\sigma) = P(-2.58 \leqslant u \leqslant 2.58) = 0.99$$

$$P(|x| \geqslant \mu + 1.96\sigma) = 0.05$$

$$P(|x| \geqslant \mu + 2.58\sigma) = 0.01$$

$$P(|u| > 1.96) = 0.05$$

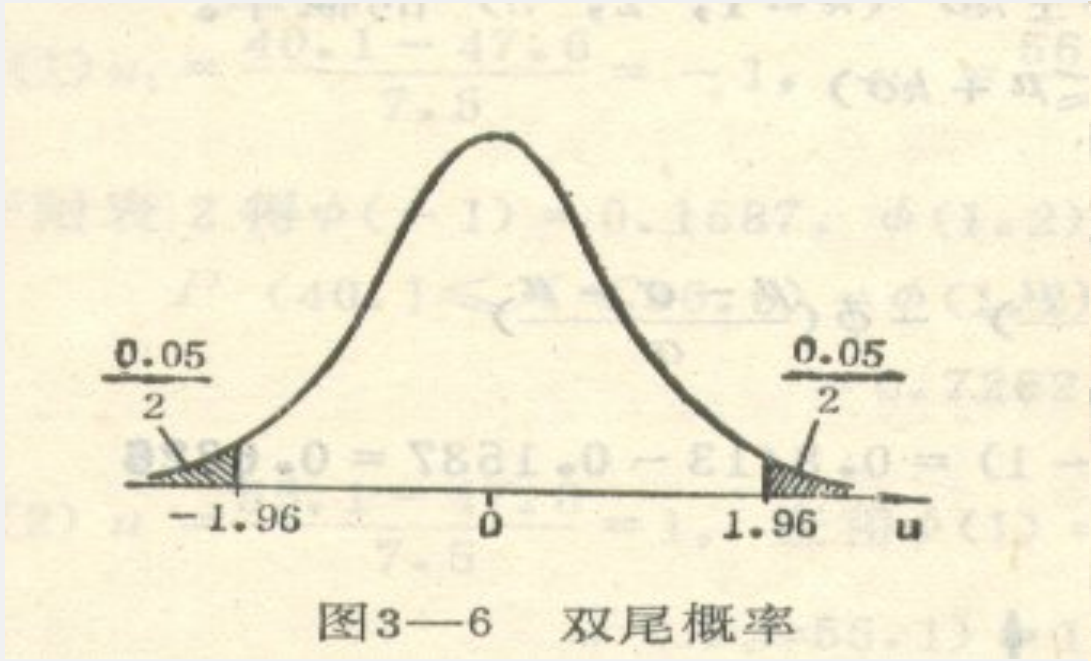


图3—6　双尾概率

(two-tailed probability)

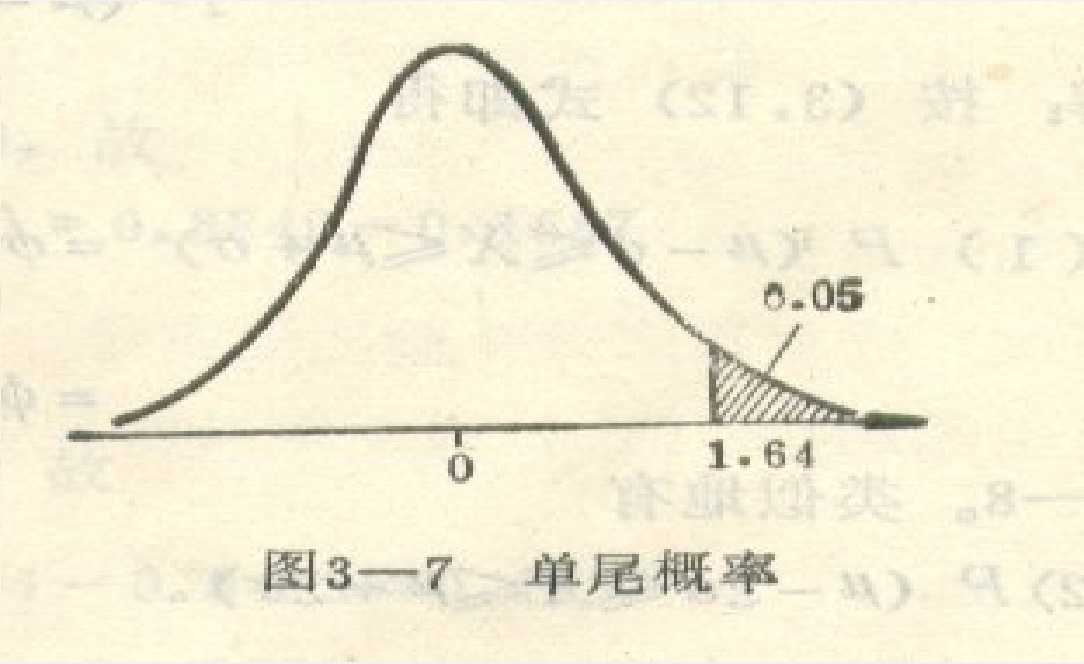


图3—7　单尾概率

(one-tailed probability)

$$P(u > 1.64) = 0.05$$

$$P(-1 \leqslant u \leqslant 1) = 0.6826$$

$$P(-2 \leqslant u \leqslant 2) = 0.9545$$

$$P(-3 \leqslant u \leqslant 3) = 0.9973$$

$$P(-1.96 \leqslant u \leqslant 1.96) = 0.95$$

$$P(-2.58 \leqslant u \leqslant 2.58) = 0.99$$

正态分布的应用

1 估计参考值范围

20株小麦株高(cm) 为82，79，85，84，86，84，83，82，83，83，84，81，80，81，82，81，82，82，82，80。其平均值为82.3cm，标准差为1.7502cm。问：小麦株高95%的正常范围值。

小麦株高服从正态分布。总体平均数 μ 和标准差 σ 未知，可以用样本平均数 $\bar{x}$ 和标准差 s 来估计总体 μ 和 σ 。

$$\bar{x} = 82.3(cm) \Rightarrow \mu$$

$$s = 1.7502(cm) \Rightarrow \sigma$$

$$u = \frac{x-\mu}{\sigma} \rightarrow x = \mu + u_{\alpha}\sigma$$

95% → $\alpha = 0.05$ → $u_{0.05} = 1.96$

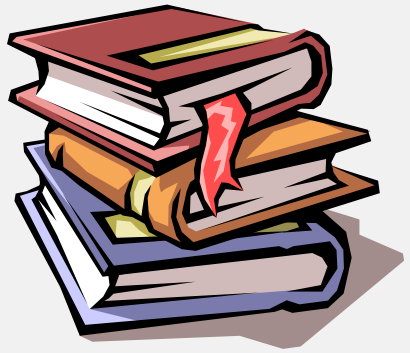
$$\mu \pm 1.96\sigma$$

[78.57， 85.73]

问2：x≥85(cm) 的概率？

$$u=\frac{x-\mu}{\sigma}=\frac{85-82.3}{1.7502}=1.54$$

$$\begin{aligned}P(x\geqslant 85)&=P(u\geqslant 1.54)\\&=1-F(u=1.54)\\&=1-0.9328=0.0618\end{aligned}$$



正态分布的应用

2 质量控制

服从正态分布的变量落在 $\mu \pm 2\sigma$ 及 $\mu \pm 3\sigma$ 的概率为95.45%和99.73%，在试验中，为了控制检测误差，常以 $\bar{x} \pm 2s$作为上下警戒线，以$\bar{x} \pm 3s$作为上下控制线。

正态分布的应用

3 正态分布是很多统计方法的理论基础。

二项分布，泊松分布的极限均为正态分布，在一定条件下，均可按正态分布的原理来处理。后面的t检验，方差分析，相关回归分析等多种统计方法均要求分析的指标服从正态分布。对于非正态分布资料，实施统计处理的一个重要途径是先作变量的转换，使转换后的资料近似正态分布，然后按正态分布的方法作统计处理。

另外几种连续型概率分布

指数分布(exponential distribution)

Γ分布(gamma distribution)

β分布（beta distribution)

威布尔分布(Weibull distribution)

均匀分布(uniform distribution)

小结

前面讨论的三个重要的概率分布中，前一个属连续型随机变量的概率分布，后两个属离散型随机变量的概率分布 。 三 者间的关系如下：

对于二项分布，在$n\to\infty, p\to 0$， 且 $np=\lambda$(较小常数)情况下 ，二项分布 趋于 波 松布。在这种场合，波松分布中的参数 λ用二项分布的np代之；在$n\to\infty, p\to 0.5$时 ， 二项分布趋于正态分布。在这种场合 ，正态分布中的 μ、σ^2用二项分布的np、npq代之。在实际计算中，当$p<0.1$且n 很大时 ， 二项分布可由波松分布近似；当$p>0.1$且n很大时 ，二项分布可由正态分布近似。

2.6 大数定律

大数定律

大数定律：是概率论中用来阐述大量随机现象平均结果稳定性的一系列定律的总称。

主要内容：样本容量越大，样本统计量与总体参数之差越小。

大数定律

(1) 贝努里大数定律

设m是n次独立试验中事件A出现的次数，而p是事件A在每次试验中出现的概率，则对于任意小的正数 ε ，有如下关系：

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

大数定律

(2) 辛钦大数定律

设$x_1, x_2, x_3, \cdots, x_n$是来自同一总体的变量，对于任意小的正数 ε，有如下关系：

$$\lim_{n \to \infty} P\{|\bar{x} - \mu| < \varepsilon\} = 1$$