概率统计基础知识

第一节　概率基础知识

一、事件与概率

（一）随机现象：抛硬币，投掷骰子，测量误差

- 随机现象

在一定条件下，并不总是出现相同结果的现象。

- 特点

——结果至少有两个

——哪一个结果出现，事先并不知道

认识一个随机现象，首要的是能罗列出它的一切可能发生的基本结果。

- 样本点

随即现象的基本结果称为样本点。

- 样本空间：记为Ω

随机现象可能样本点的全部称为这个随机现象的样本空间。

（二）随机事件

- 事件（随机事件）：随机现象的某些样本点组成的集合。用大写英文字母A、B、C……表示。

- 随机事件的特征

—— 任一事件A是相应样本空间Ω中的一个子集。

—— 事件A发生当且仅当（$\Leftrightarrow$）A 中某一样本点发生。

—— 事件A的表示可用集合，也可用语言，但所用语言要大家明白无误。

—— 任一样本空间Ω有一个最大子集即Ω；它对应的事件称为必然事件，仍用Ω表示。

—— 任一样本空间Ω都有一个最小子集即空集，它对应的事件称为不可能事件，记为Φ

例1：若批产品既有合格品也有不合格品，记抽到合格品为“0”，抽到不合格品为“1”；从中顺序抽取2件，则抽到产品结果的样本空间为（所有可能结果的全体）：

Ω={(0,0),(0,1),(1,0)(1,1)}

事件A=“至少有一件合格品”={(0,0),(0,1),(1,0)}

事件B=“没有一件合格品”={(1,1)}

事件C=“至多有一件合格品”={0,1),(1,0)(1,1)}

事件D=“至多两件合格品” ={(0,0),(0,1),(1,0)(1,1)}= Ω=“至多两件不合格品”=事件E

事件F=“抽到3件合格品”= Φ

事件G= “没有一件不合品格”={(0,0)}

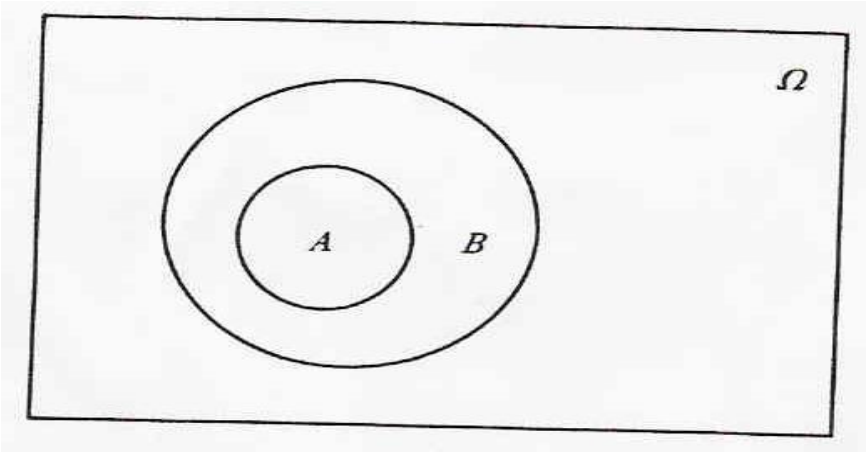
事件H=“两次抽到的结果一致” ={(0,0), (1,1)}

- 若只抽1件，则样本空间？
- A和B什么关系？
- A和C什么关系？
- B和G什么关系？

- **随机事件的关系**

—— 包含：$A \subset B$或$B \supset A$

在一个随机现象中有两个事件A与B，若事件A中任一个样本点必在B中，则称A被包含在B中，或B包含A。



抽2次产品的例子中：

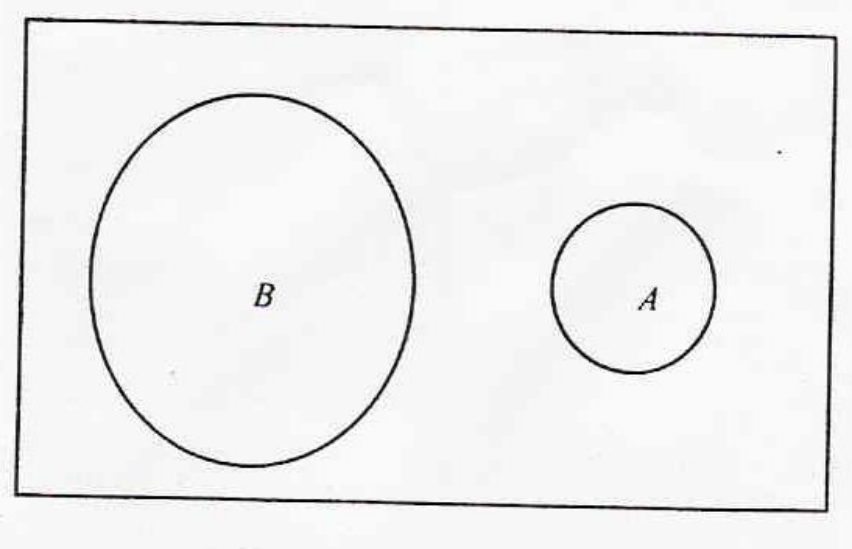
事件B=“没有一件合格品”={(1, 1)}

事件C=“至多有一件合格品”={0, 1), (1, 0) (1, 1)}

—— 互不相容

在一个随机现象中有两个事件A与B，若事件A与B没有相同的样本点，则称A与B互不相容。

- 可推广到三个或更多个事件间的互不相容



—— 相等：A=B即A⊂B且B ⊂ A

两个随机事件A与B，若样本A与B含有相同的样本点，则称事件A与B相等。

投掷骰子2次：A={（x，y）：x + y =奇数}

B={（x，y）：x与y的奇偶性不同}

则：$A=B=\left\{\begin{matrix}(1,2),(1,4),(1,6),(2.1),(2,3),(2,5)\\(3,2),(3,4),(3,6)\cdots\end{matrix}\right\}$

事件D=“至多两件合格品” ={(0,0),(0,1),(1,0)(1,1)}=
Ω=“至多两件不合格品”=事件E

（三）事件的运算

● 事件运算

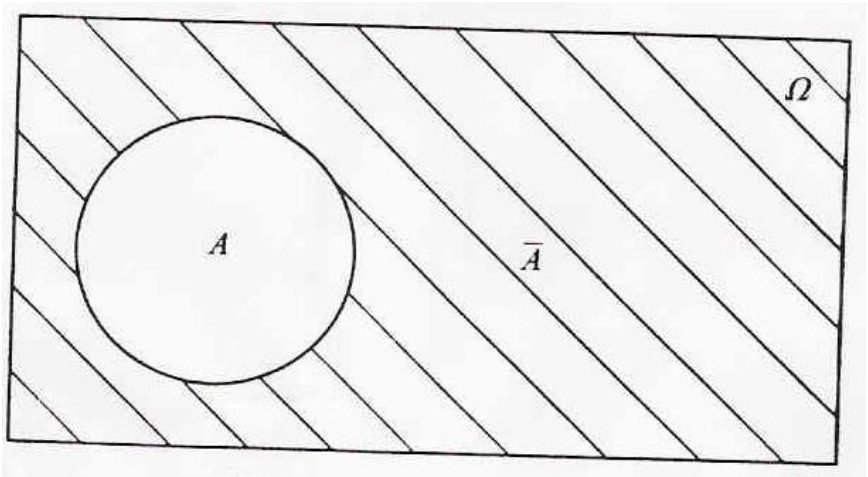
—— 对立事件（余事件）：A→ $\overline{A}$

抽产品例子中：

Ω={(0,0),(0,1),(1,0)(1,1)}

事件A=“至少有一件合格品”={(0,0),(0,1),(1,0)}

事件B=“没有一件合格品”={(1,1)}= $\overline{A}$



则 $A=\overline{\overline{A}}$ ，$\overline{\Omega}=\Phi$ ，$\overline{\Phi}=\Omega$

—— 事件A与B的并：$A \cup B$

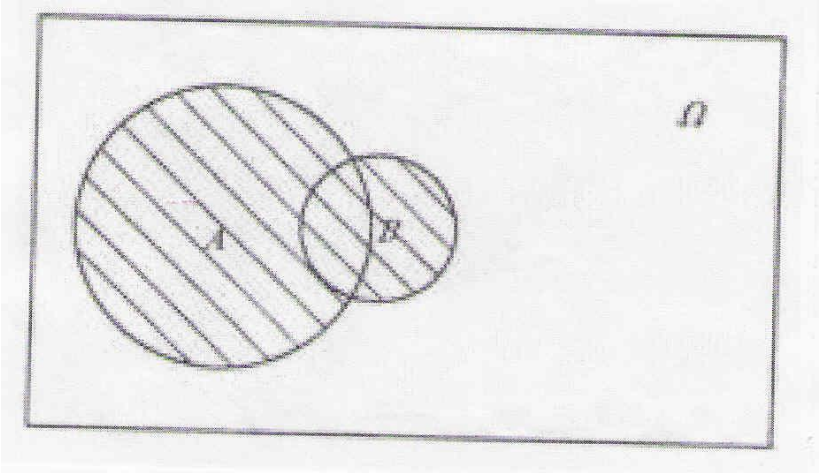
由事件A与B中所有样本点（相同的只计入一次）组成的新事件。称为A与B的并，$A \cup B$ 发生意味着“事件A与B至少一个发生”

例如用样本空间表示：

$A = \{1, 3, 5\}$

$B = \{1, 2, 3\}$

则 $A \cup B = \{1, 2, 3, 5\}$



—— 事件A与B的交：A ∩B或AB

由事件A与B中公共的样本点组成的新事件称为事件A与B的交。　　发生意味着“事件A与B同时发生”

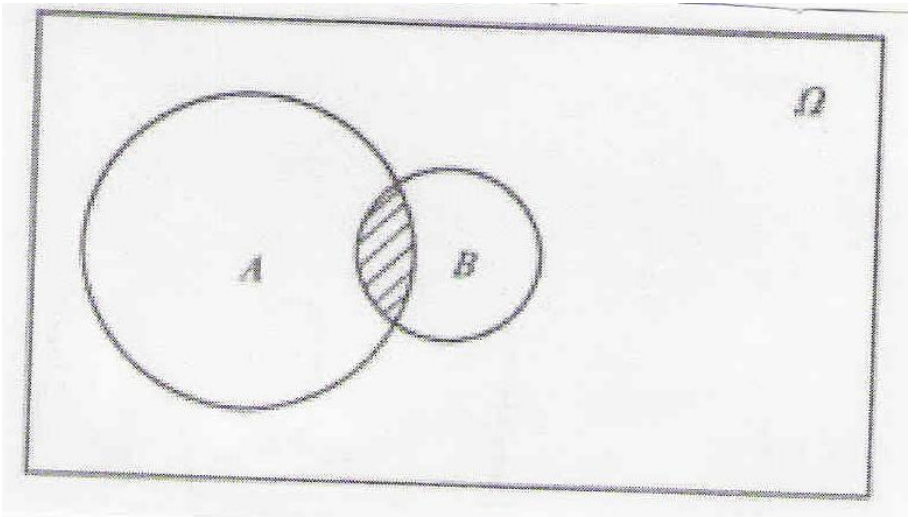
$$A \cap B$$

在北京市随机抽取一个人

A=抽到的是60岁以上的老人

B=抽到的是男性

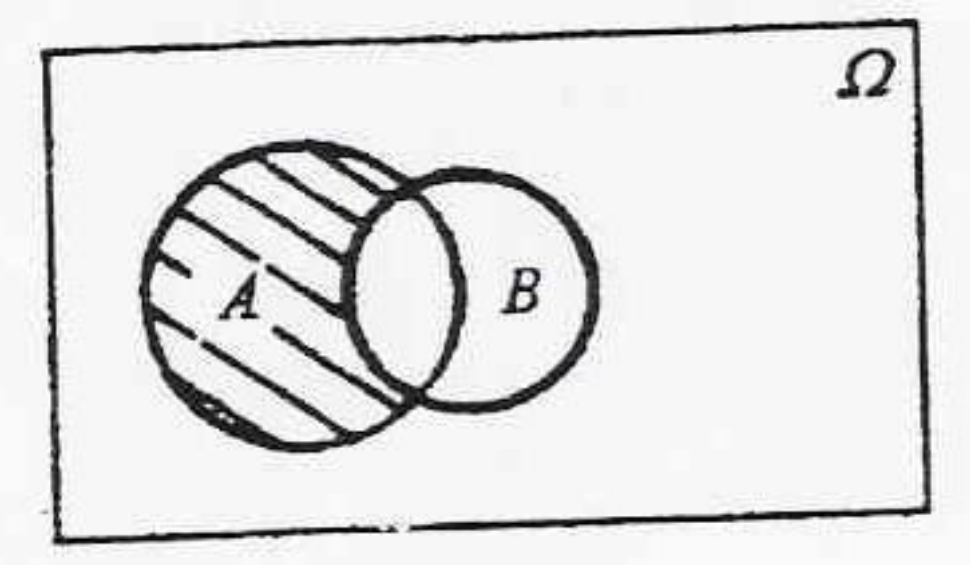
A∩B表示：抽到60岁以上男性



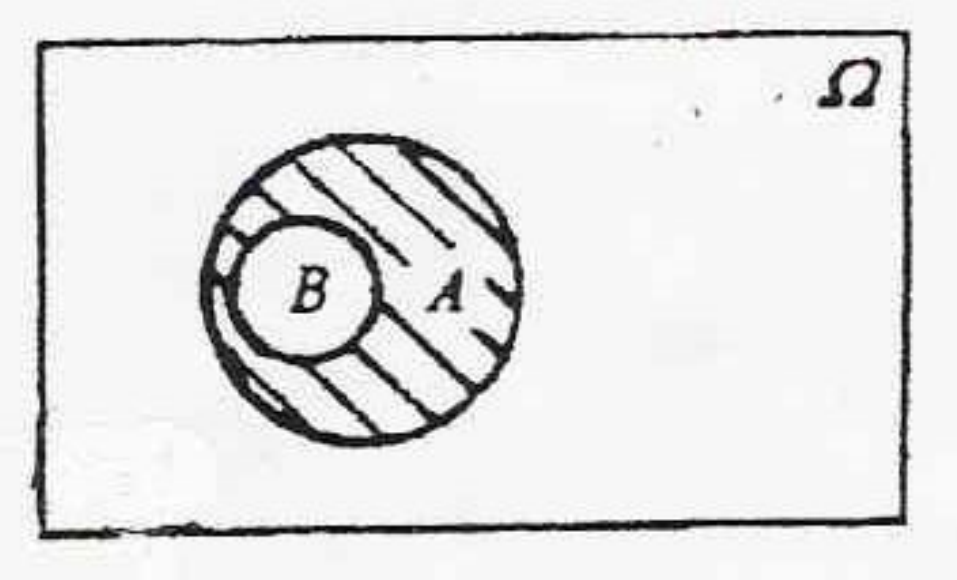
- 事件的并和交可推广到更多个事件上去。

—— 事件A对B的差：A-B

由在事件A中而不在B中的样本点组成的新事件，称为A对B的差。



（a）A-B



（b）A-B（$A \supset B$）

事件运算性质：

—— 交换律：$A\cup B=B\cup A$ ，$A\cap B=B\cap A$

—— 结合律 $A\cup(B\cup C)=(A\cup B)\cup C$ 运算相同：

$A\cap(B\cap C)=(A\cap B)\cap C$

—— 分配律 $A\cup(B\cap C)=(A\cup B)\cap(A\cup C)$ 运算不同：

$A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$

—— 对偶律：$\overline{A\cup B}=\overline{A}\cap\overline{B}$ （并之余等于余之交）

$\overline{A\cap B}=\overline{A}\cup\overline{B}$ （交之余等于余之并）

可用维恩图验证，可推广到三个或三个以上事件的运算。

（四）事件的概率

- **概率——事件发生可能性大小的度量**

记为P（A）。

降水概率：

成功的概率：

中奖的概率：

风险：不希望发生事件发生的可能性

- 概率是一个介于0和1之间的数，即$0 \le P(A) \le 1$;
- 必然事件的概率等于1，即$P(\Omega)=1$；
- 不可能事件的概率等于0，即$P(\Phi)=0$。

二、概率的古典定义与统计定义

（一）古典定义

—— 所涉及的随机现象只有有限个样本点。如共有n个样本点；

—— 每个样本点出现的可能性是相同的（等可能性）；

—— 假如被考察事件A含有K个样本点，则事件 A的概率定义为：

$$P(A)=\frac{K}{n}=\frac{A\text{中含样本点的个数}}{\Omega\text{中样本点的总数}}$$

（二）统计定义

—— 与考察事件A有关的随机现象是可以大量重复试验的；

—— 若在n次重复试验中，事件A发生K_n次，则事件A发生的频率为：

$$f_n(A)=\frac{K_n}{n}=\frac{\text{事件}A\text{发生次数}}{\text{重复试验数}}$$

—— $f_n(A)$将会随着重复试验次数不断增加而趋于稳定，这个频率的稳定值就是事件A的概率。一般用重复次数n较大时的频率去近似。

——概率是频率的稳定值（极限值），例子见教材P10

例1：若批产品既有合格品也有不合格品，记抽到合格品为“0”，抽到不合格品为“1”；从中顺序抽取2件，则抽到产品结果的样本空间为（所有可能结果的全体）：

Ω={(0,0),(0,1),(1,0)(1,1)}

事件A=“至少有一件合格品”={(0,0),(0,1),(1,0)}

事件B=“没有一件合格品”={(1,1)}

事件C=“至多有一件合格品”={0,1),(1,0)(1,1)}

事件H=“两次抽到的结果一致” ={(0,0), (1,1)}

若这批产品10000件中合格品与不合格品各占一半，且产品分布均匀随机，则

- P（A）=?
- P（B）=?
- P（C）=?
- P（H）=?

若批产品总数10000件中不合格品有2000件，结果会怎样呢？

二、概率的古典定义与统计定义

（三）古典概率的计算

$$P(A)=\frac{K}{n}=\frac{A\text{中含样本点的个数}}{\Omega\text{中样本点的总数}}$$

附件、　排列与组合

1. 乘法原理：

如果做某件事需经k步才能完成，其中做第一步有m_1种方法，第二步有m_2种方法，……第k步有m_k种方法，那么完成这件事共有$m_1 \times m_2 \times \cdots\cdots \times m_k$种方法。

例：甲城到乙城有3条线路，乙城到丙城有2条线路，那么从甲城到丙城有3×2=6条线路。

例：用0,~,9这十个数字，可以写出多少个不重复的四位数？

2. 加法原理：如果做某件事可由k类不同方法之一完成，其中在第一类方法中又有m_1 种完成方法，在第二类方法中又有m_2 种完成方法，……在第k类方法中又有m_k 种完成方法，那么完成这件事共有$m_1 + m_2 + \cdots\cdots + m_k$ 种方法。

例如，由甲城到乙城有三类交通工具：汽车、火车和飞机，而汽车有5个班次，火车有3个班次，飞机有2个班次，那么从甲城到乙城共有5+3+2=10个班次供选择。

3、（选）排列定义（不放回抽样，不可重复）

从n个不同的元素中任取r（r不超过n）个元素并按一定顺序排成一列，称为一个排列。

则不同的排列数共有

$$P_n^r = n(n-1).....(n-r+1),$$

若r=n，称为全排列，全排列数共有 $n!$ 个，

记为 P_n ，即

$$P_n = n!$$

4、重复排列定义（放回抽样）

从n个不同的元素中选出r个，排成一列，每个元素可以重复出现，这种排列称为有重复排列。按乘法原理，此种重复排列种数共有 n^r 个。

组合数

从n个不同的元素a_1，a_2，……，a_n中任取r个为一组（两组元素有不同时才看成不同的组，即不考虑其排列顺序），所能得出的全部不同的组数，称为从n个元素中取r个的组合数，记作

$$C_n^r = \binom{n}{r} = \frac{P_n^r}{r!} = \frac{n!}{r!(n-r)!}$$

三、概率的性质及其运算法则

概率的性质：（可由概率的定义看出）

—— 性质1：对任意事件A，有0≤P(A)≤1；

—— 性质2：$P(\overline{A})=1-P(A)$

—— 性质3：若A⊃B

则P（A-B）=P（A）-P（B）

三、概率的性质及其运算法则

概率的性质：（可由概率的定义看出）

—— 性质4：$P(A \cup B)=P(A)+P(B)-P(AB)$

若A与B互不相容$P(A \cup B)=P(A)+P(B)$

—— 性质5：对于多个互不相容事件A_1，A_2，……，

有$P(A_1 \cup A_2 \cup A_3 \cup \ldots\ldots)=P(A_1)+P(A_2)+p(A_3)+\ldots\ldots$；

例1：若批产品既有合格品也有不合格品，记抽到合格品为“0”，抽到不合格品为“1”；从中顺序抽取2件，则抽到产品结果的样本空间为（所有可能结果的全体）：

Ω={(0,0),(0,1),(1,0)(1,1)}

事件A=“至少有一件合格品”={(0,0),(0,1),(1,0)}

事件B=“没有一件合格品”={(1,1)}= $\overline{A}$

事件C=“至多有一件合格品”={0,1),(1,0)(1,1)}

事件F=“抽到3件合格品”= Φ

事件G= “没有一件不合品格”={(0,0)}

事件H=“两次抽到的结果一致” ={(0,0), (1,1)}

事件R=“两次抽到的结果不一致”={0,1),(1,0)}

- R=C-B=A-G： 有且只有一件合格品
- H=B ∪ G
- R= $A \cap C$
- Ω =A ∪C

四、条件概率与概率的乘法法则

（1）条件概率

两个事件A与B，在事件B已发生的条件下，事件A再发生的概率称为条件概率，记P（A/B）。

计算公式：

$$P(A|B)=\frac{P(AB)}{P(B)} \qquad (P(B)>0)$$

- **性质6：对任意二个事件A与B，有**

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

$P(B)>0$ $\qquad$ $P(A)>0$

（2）独立性和独立事件的概率

相互独立：

设有两个事件A与B，假如其中一个事件的发生不影响另一个事件的发生与否，则称A事件与B事件相互独立。

- **性质7：**

  假如二个事件A与B相互独立，则A与B同时发生的概率为P(AB)=P(A)P(B)

- **性质8：**

  假如二个事件A与B相互独立，则在事件B发生条件下，事件A发生的条件概率P(A|B)等于事件A的（无条件）概率p(A)

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

- 事件的相互独立可推广到三个或更多的事件上去。

  P(ABC)=P(A)P(B) P©

注意：两两独立不等于相互独立，即：

P(AB)=P(A)P(B)， P(BC)=P(B) P(C)， P(AC)=P(A)P(C)

不能推出： P(ABC)=P(A)P(B) P(C)

第二节　随机变量及其分布

一、随机变量

- **随机变量**

用来表示随机现象结果的变量称为随机变量。

随机变量常用大写字母X、Y、Z……表示。

取值带有**随机性**的变量称作随机变量（非唯一，不确定）

- **随机变量类型**

—— 离散随机变量

一个随机变量的所有可能值仅取数轴上有限个点或可列个点，称为为离散（型）随机变量。

—— 连续随机变量

如一个随机变量的所有可能取值充满数轴上一个范围（a，b）或整个数轴，则此随机变量为连续（型）随机变量。

二、随机变量的分布

- **随机变量的分布**

  随机变量取值的统计规律性。

- **随机变量X的分布内容：**

  —— **X可能取哪些值或在哪个区间上取值**

  —— **X取这些值的概率各是多少？或X在任一小区间上取值的概率是多少？**

（一）离散随机变量分布的表示

1、用分布列表示（离散分布）

分布列

X	X_1	X_2	……	X_n
P	p_1	p_2	……	p_n

或用数学式表达：

$P(X=Xi)=p_i$　i=1，2……n（$p_1+\ldots+p_n=1$）

- p_i也称为分布的概率函数

（一）离散随机变量的分布

例1：设Y表示“掷两颗骰子的点数之和”，它的分布列为：

Y	2	3	4	5	6	7	8	9	10	11	12
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

则 $P(3 \le Y \le 6) = P(Y=3)+P(Y=4)+P(Y=5)+P(Y=6)$

= 2/36 +3/36 + 4/36 + 5/36

=14/36=7/18

例2：10个产品中，有2个不合格品，若随机抽取4个产品，抽到不合格品的个数X的分布：

解 （不放回抽样）X可能取值仅为0, 1, 2

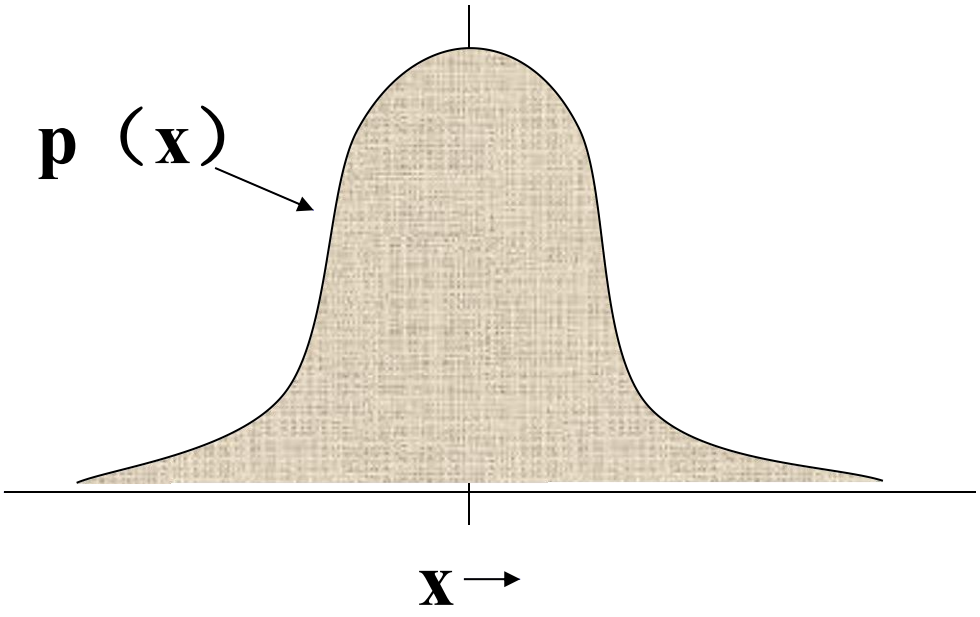
且 $$P(X=m)=\frac{\binom{2}{m}\binom{8}{4-m}}{\binom{10}{4}} \qquad m=0,1,2$$

（二）连续随机变量的分布表示

用概率密度函数p（x）表示，其中：

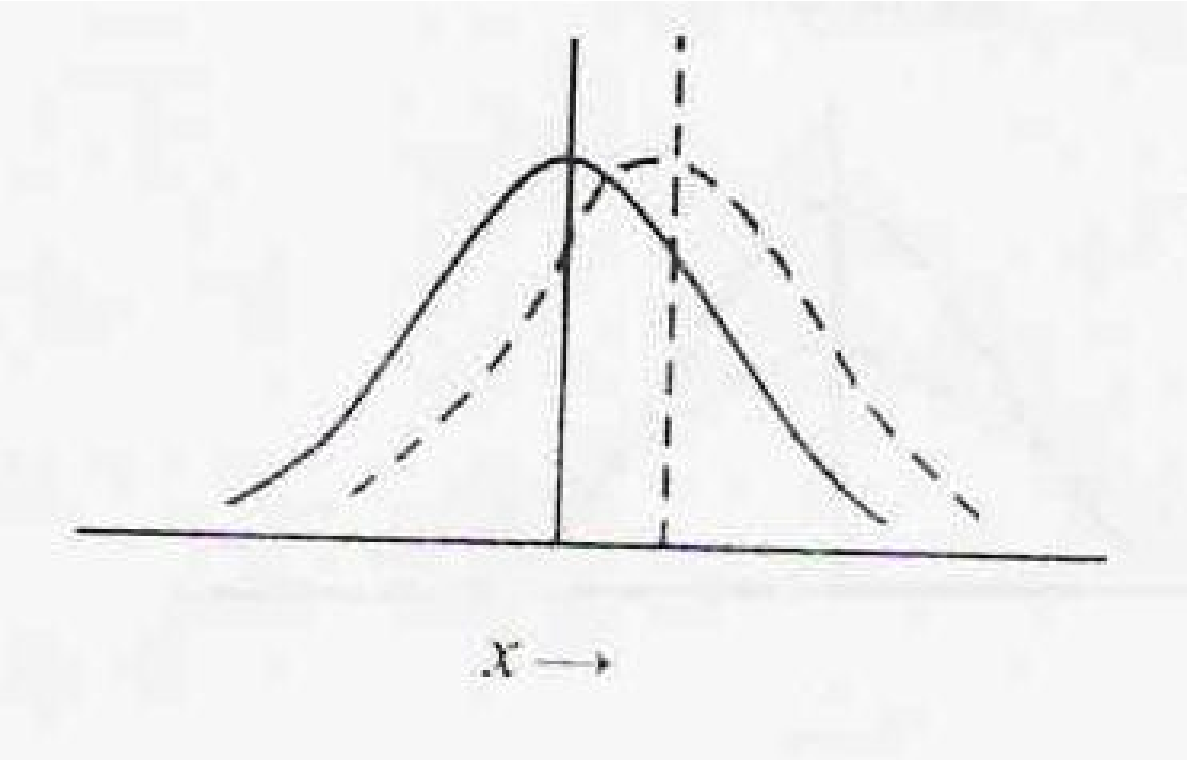
① p（x）≥0

② $\int_{-\infty}^{+\infty} p(x)dx = 1$

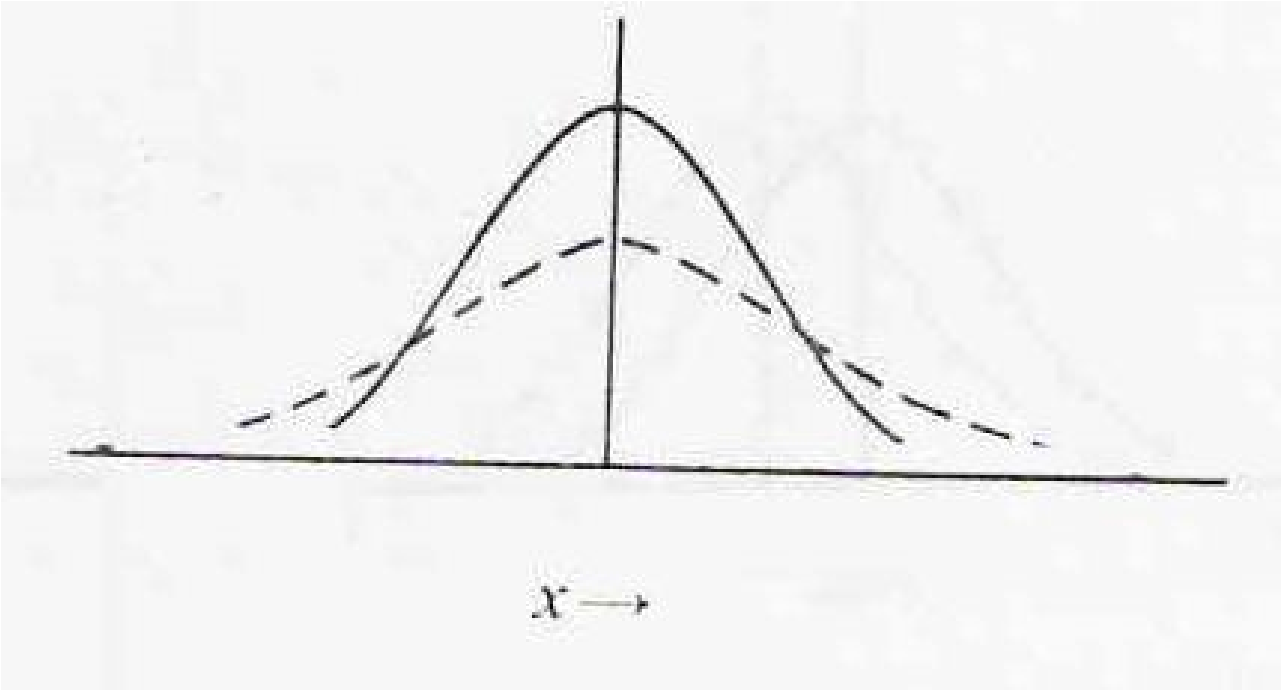


- **概率密度函数p(x)的各种形式**

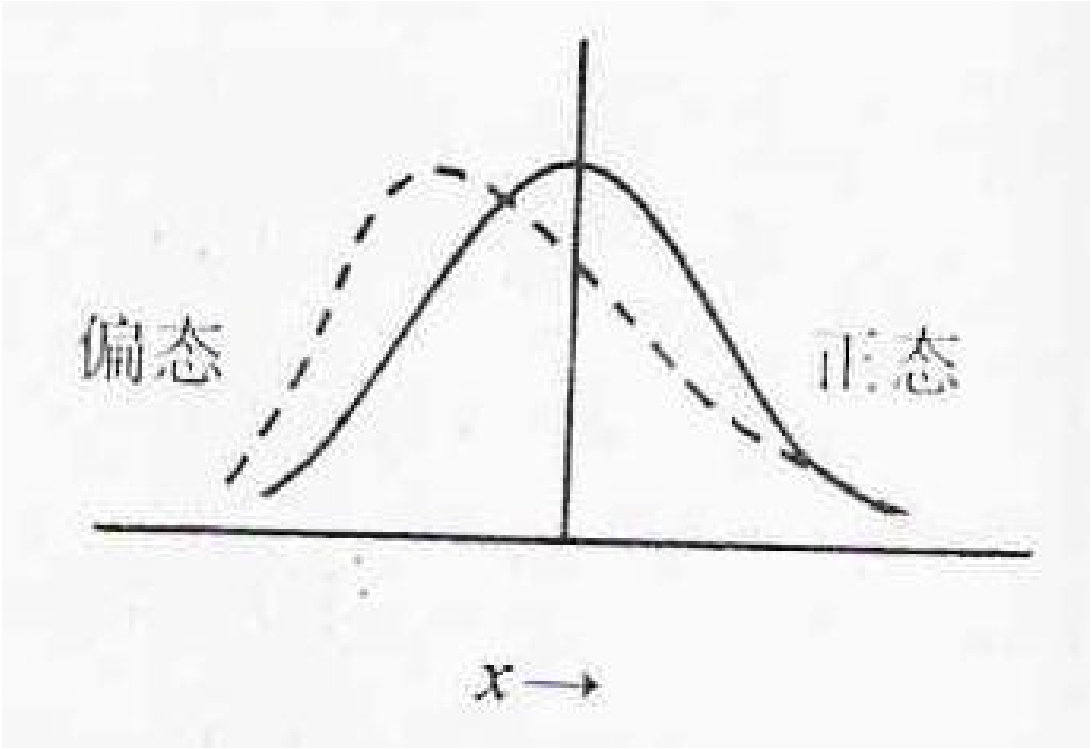
  —— 位置不同



—— 散布不同

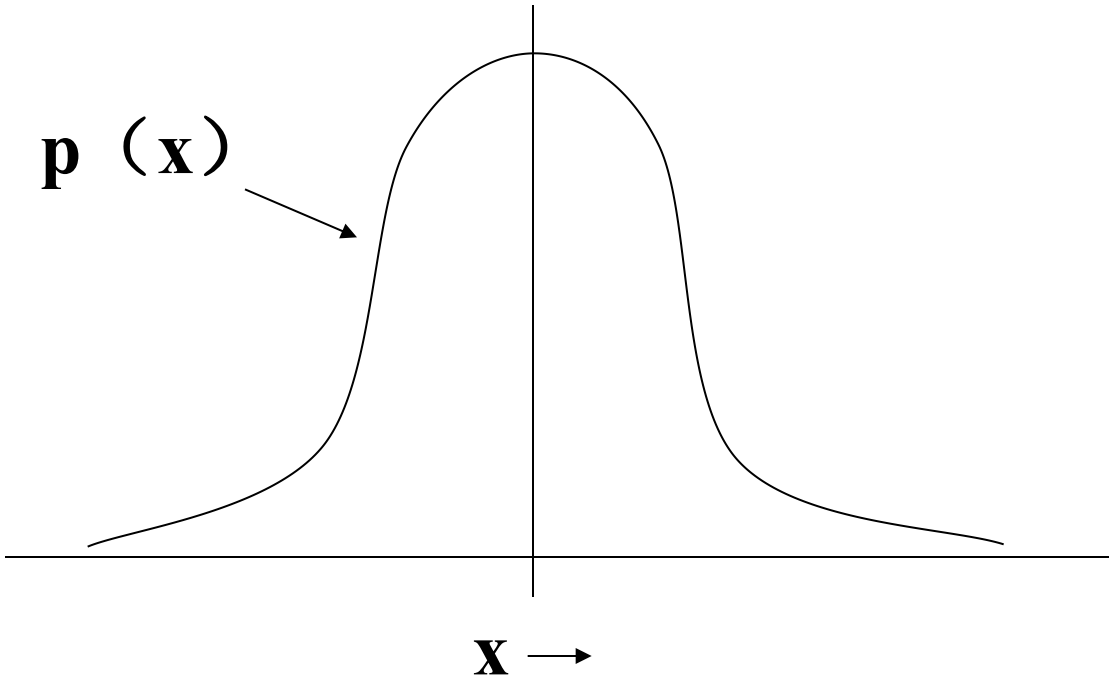


—— 形状不同



其中p(x)在x_0点的值p(x)不是概率，是高度。

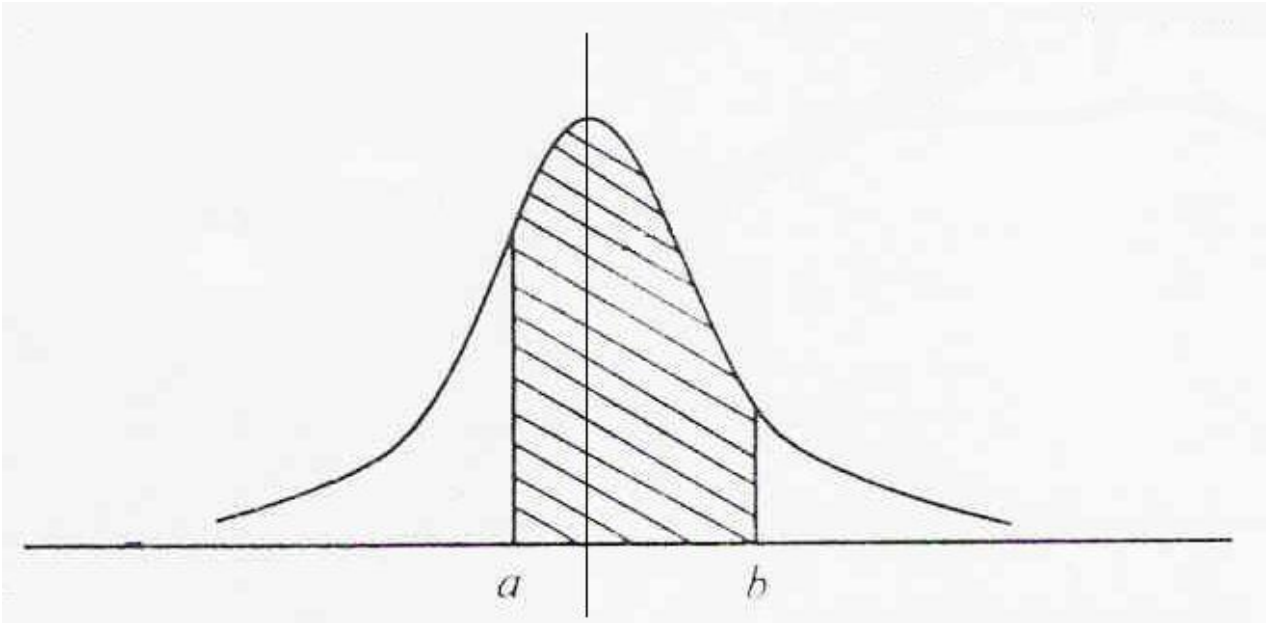
注：纵轴原为“单位长度上的频率”，由频率的稳定性，可用概率代替频率，纵轴就成为“单位长度上的概率”即概率密度的概念，故最后形成的曲线称为概率密度曲线。



- **重要结论：**

1．X在区间（a，b）上取值的概率P(a＜X＜b)为概率密度曲线以下区间（a，b）上的面积，即

$$P(a<X<b)=\int_a^b p(x)dx$$



2. X在一点取值的概率为零，即

$$P(X=a)=0$$

故： $P(a<x<b)=P(a\le x\le b)$

$$=P(a\le X<b)$$

$$=P(a<X\le b)$$

三、随机变量分布的均值、方差与标准差

- **均值（位置参数）：**

  用来表示分布的中心位置，用E(X)表示

$$E(X)=\begin{cases}\sum x_i p_i & \text{X是离散随机变量}\\ \int_{-\infty}^{+\infty} xp(x)dx & \text{X是连续随机变量}\end{cases}$$

- **方差（形状参数）：**

  用来表示分布的散布大小，用$V_{ar}(x)$表示

$$Var(X)=\begin{cases}\sum[x_i-E(x)]^2P_i & \text{X是离散随机变量}\\ \int_{-\infty}^{+\infty}[x-E(x)]^2P(x)dx & \text{X是连续随机变量}\end{cases}$$

- **标准差：用σ表示**

$$\sigma=\sigma(X)=\sqrt{Var(X)}$$

表示分布散布大小。

- **均值与方差的运算性质**

—— 对任意二个随机变量X_1和X_2，有

$$E(X_1+X_2)=E(X_1)+E(X_2)$$

—— 设X为随机变量，a与b为任意常数，有

$$E(ax+b)=aE(x)+b$$

$$Var(aX+b)=a^2Var(X)$$

—— 设X_1与X_2相互独立

$$Var(X_1 \pm X_2) = Var(X_1) + Var(X_2)$$

（和的方差等于方差之和）

- 这个性质可推广到三个或更多个相互独立随机变量场合

—— 方差的这个性质不能推广到标准差场合，对任意两个相互独立的随机变量X_1与X_2，　　$\sigma(X_1+X_2) \neq \sigma(X_1)+\sigma(X_2)$

而应为：$$\sigma(X_1 + X_2) = \sqrt{Var(X_1) + Var(X_2)}$$

- 方差具有可加性，标准差不具有可加性。

例1，离散随机变量分布列为:

X	-1	0	1
P	0.1	0.2	0.7

求该随机变量的均值、方差和标准差

$$\mathbf{E(X)=}\sum_i x_i p_i \mathbf{=-1\times 0.1+0\times 0.2+1\times 0.7=0.6}$$

$$Var(X)=\sum_i [x_i - E(X)]^2 p_i$$

$$=(-1-0.6)^2\times 0.1+(0-0.6)^2\times 0.2+(1-0.6)^2\times 0.7$$

$$=1.6^2\times 0.1+0.6^2\times 0.2+0.4^2\times 0.7$$

$$=2.56\times 0.1+0.36\times 0.2+0.16\times 0.7$$

$$=0.256+0.072+0.112$$

$$=0.44$$

四、常用分布

（一）常用的离散分布

- **二项分布**

$$P(X=x)=\binom{n}{x}p^{x}(1-p)^{n-x}$$

x =0，1，……，n

其中 $\binom{n}{x}=\frac{n!}{x!(n-x)!}$ 表示从n个不同元素取出x个的组合数。

记为b（n，p）或B（n，p）

四、常用分布

（一）常用的离散分布

- 二项分布 （用于无限总体 抽样中 或有限总体放回抽样，抽样检验中N≥250近似看做无限总体）
- 条件:

1) 重复试验n次

2）结果相互独立

3）仅有2个可能结果

4）2试验结果的概率分别为p和1-p

（一）离散随机变量的分布

例2：10个产品中，有2个不合格品，若每次抽完检验后再放回去，随机抽取4次，抽到不合格品的个数Y的分布：

解 （放回抽样）Y可能取值为0, 1, 2，3, 4

$$P(Y=m)=\binom{4}{m}0.2^{m}0.8^{4-m} \qquad m=0,1,2,3,4$$

- **二项分布均值、方差和标准差**

  —— **均值E(x)=np**

  —— **方差：Var(x)=np(1-p)**

  —— **标准差：** $\sigma = \sqrt{np(1-p)}$

- **两点分布：n=1时的二项分布**

$$P(X = x) = p^x (1-p)^{1-x} \qquad x = 0, 1$$

  —— **均值E(x)=p**

  —— **方差：Var(x)=p(1-p)**

  —— **标准差：** $\sigma = \sqrt{p(1-p)}$

- 泊松分布：（常用于无限总体计点过程）

例子：

1）一定时间内，电话接错的次数

2）一定时间内，某操作系统故障次数

3）一块布匹上瑕点的个数

4）一页书上错字个数

5）一幅地图上标注的错误数

前提：一定时间、一定区域或特定单位内的计点次数

用 λ（$\lambda>0$）表示特定单位内的平均计点数

- **泊松分布：（常用于计点过程）**

$$P(X=x)=\frac{\lambda^x}{x!}e^{-\lambda}$$ **x =0，1，2，……**

记为P(λ)　　其中e=2.71828

- **泊松分布均值、方差和标准差**

—— **均值：E(X)=λ**

—— **方差：$Var(X)=\lambda$**

—— **标准差：$\sigma=\sqrt{\lambda}$**

例1，根据记录统计（仅考虑weekday），某人在一天内接到的电话次数平均为8次；

1）一天“没有接到电话”的概率。

2）一天 “接到7次电话”的概率。

3）一天 “接到至少2次电话”的概率。

1) $$P(X=0)=\frac{8^0}{0!}e^{-8}=e^{-8}$$

2) $$P(X=7)=\frac{8^7}{7!}e^{-8}$$

3) $$P(X\geq 2)=1-P(X<2)=1-P(X=0)-P(X=1)$$

$$=1-e^{-8}-8e^{-8}=1-9e^{-8}$$

- **超几何分布：（有限总体不放回抽样）**

$$P(X=x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad x=1，2……，r$$

式中r=min（n，M）

M为批量N中所含不合格品数，　n为样本量

记为h（n，N，M）

例2：10个产品中，有2个不合格品，若随机抽取4个产品，抽到不合格品的 个数X的分布：

$$P(X=m)=\frac{\binom{2}{m}\binom{8}{4-m}}{\binom{10}{4}} \qquad m=0,1,2$$

- 超几何分布：（有限总体不放回抽样）

（了解即可）

- 超几何分布均值、方差、标准差

—— 均值： $E(X)=\dfrac{nM}{N}$

—— 方差：$Var(X)=\dfrac{n(N-n)}{N-1}\bullet\dfrac{M}{N}\left(1-\dfrac{M}{N}\right)$

（二）连续型随机变量的分布

- **正态分布：能描述很多质量特性X随机取值的统计分布。**

正态分布概率密度函数：

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

正态分布含两个参数μ和σ，常记：$N(\mu, \sigma^2)$。其中μ为分布均值（即分布中心）；σ^2为分布方差；$\sigma > 0$为分布标准差。

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

p(x)函数的特性：

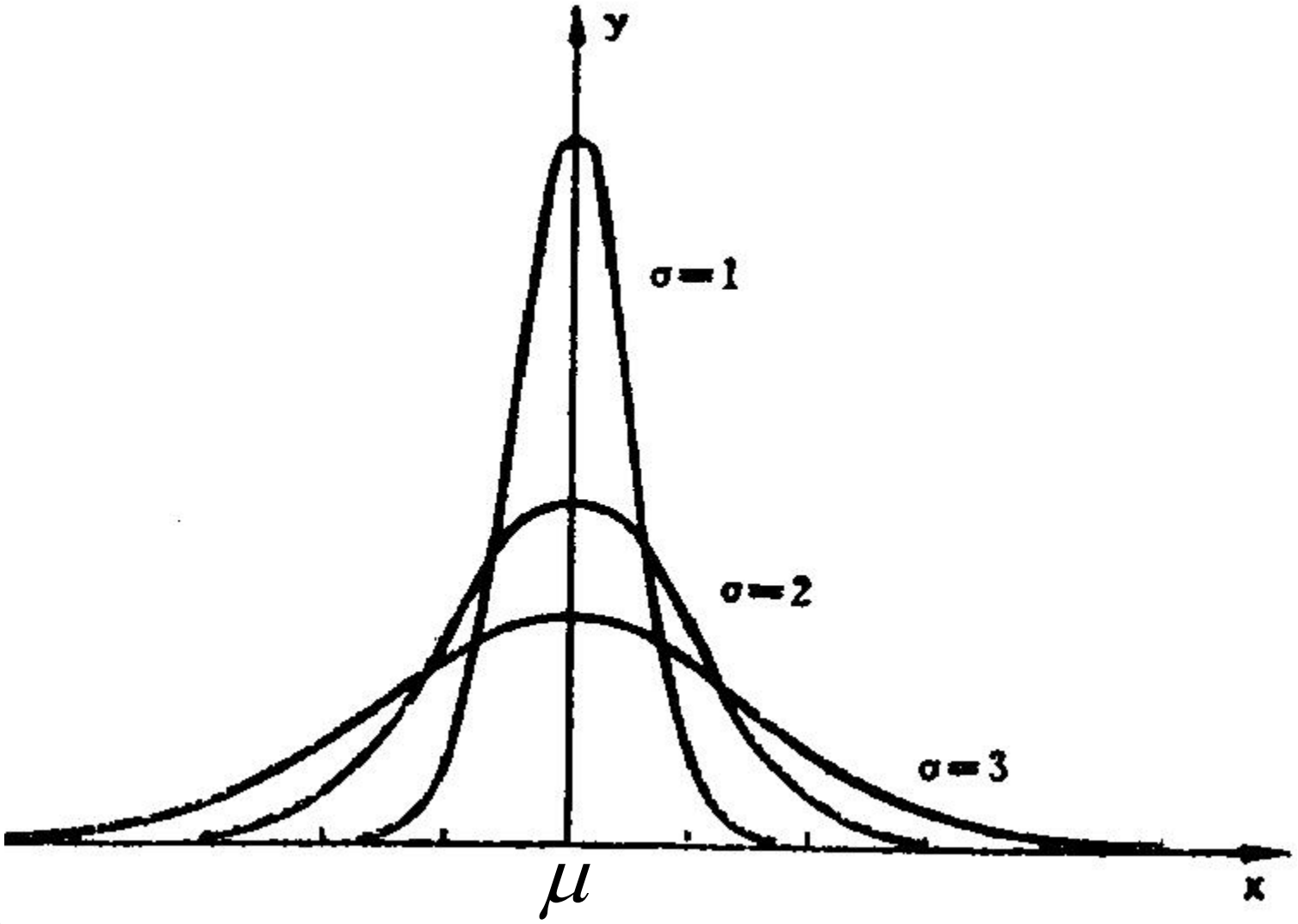
① 对称性，以x=μ为轴对称，

即对于任意实数a有： $p(\mu+a)=p(\mu-a)$

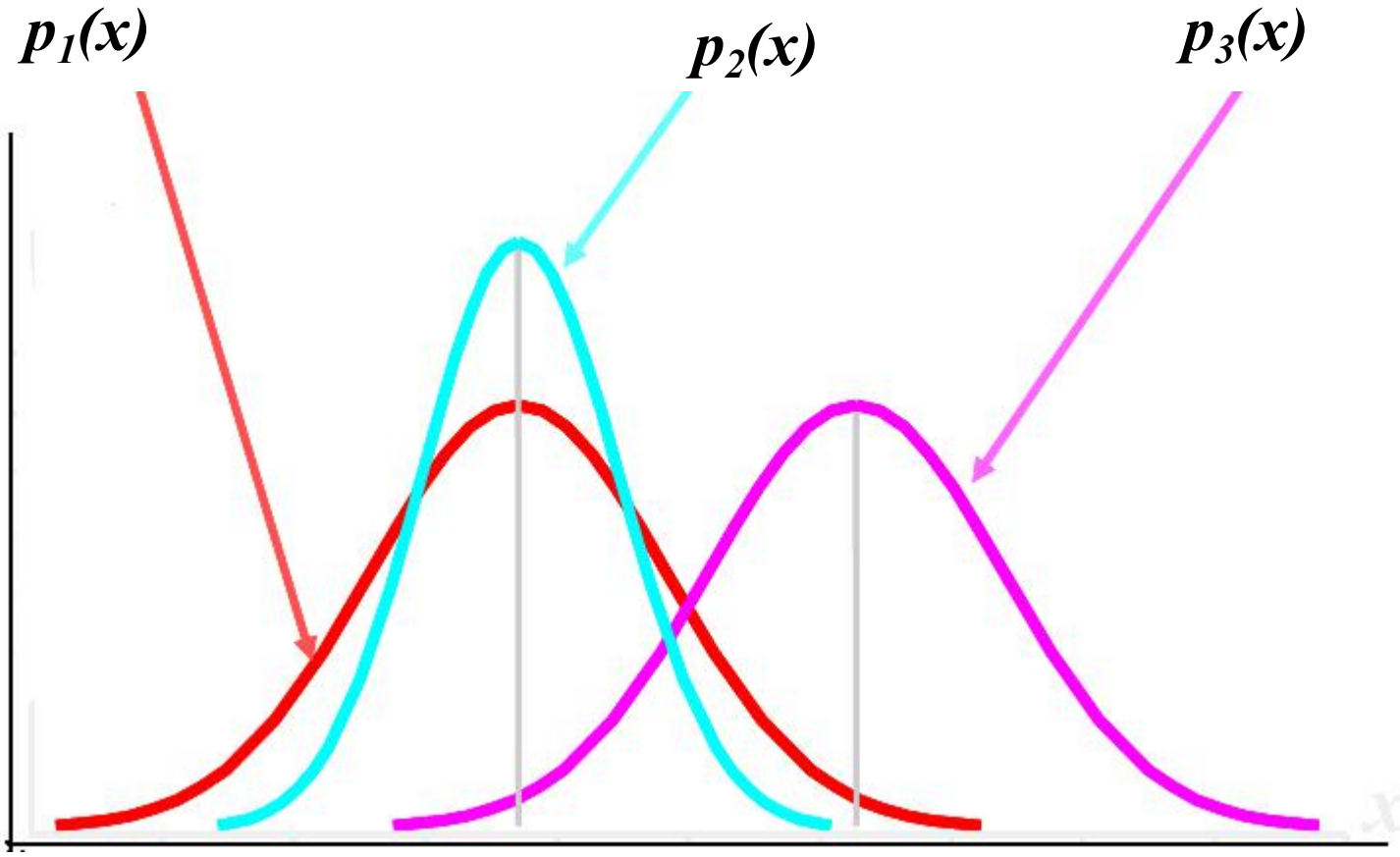
②单峰性，当x=μ时，p(x)取得最大值。 $p(\mu)=\dfrac{1}{\sigma\sqrt{2\pi}}$

③若σ越大，曲线越平缓；若σ越小，曲线越陡峭。

P（x）函数曲线

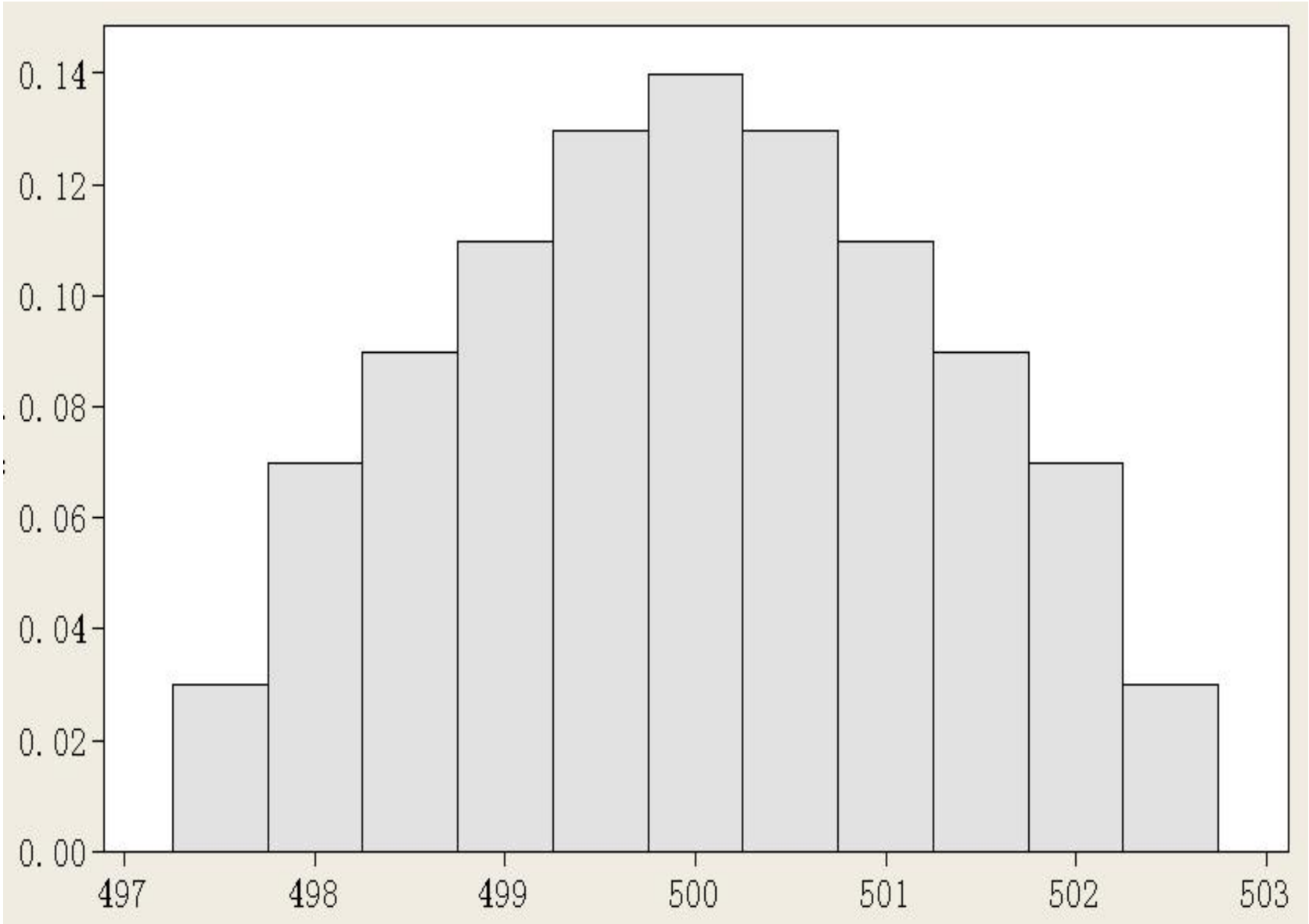


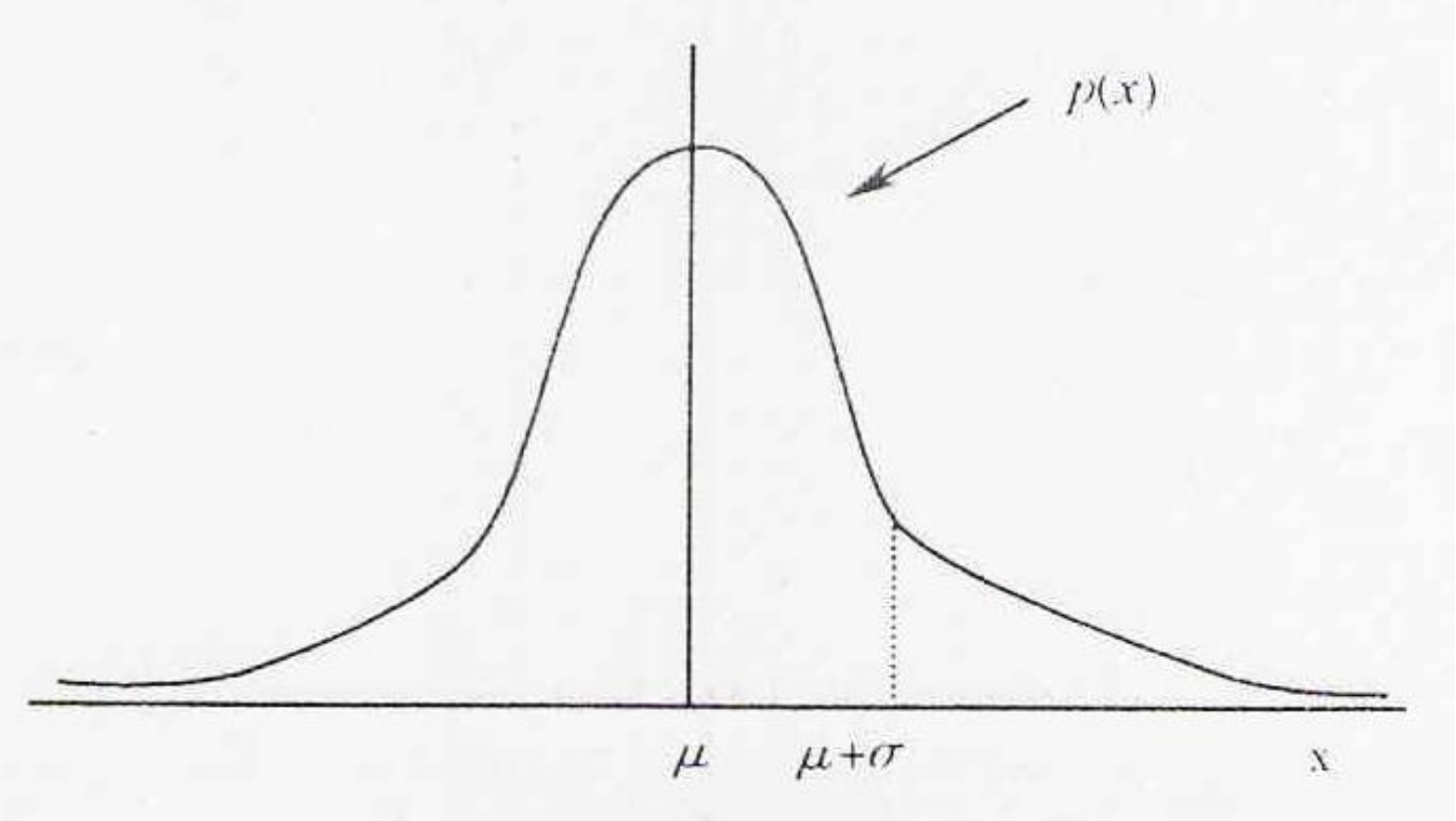
μ和σ对p（x）曲线的影响



$$\mu_1 = \mu_2; \qquad \mu_1 < \mu_3$$

$$\sigma_1 > \sigma_2; \qquad \sigma_1 = \sigma_3$$

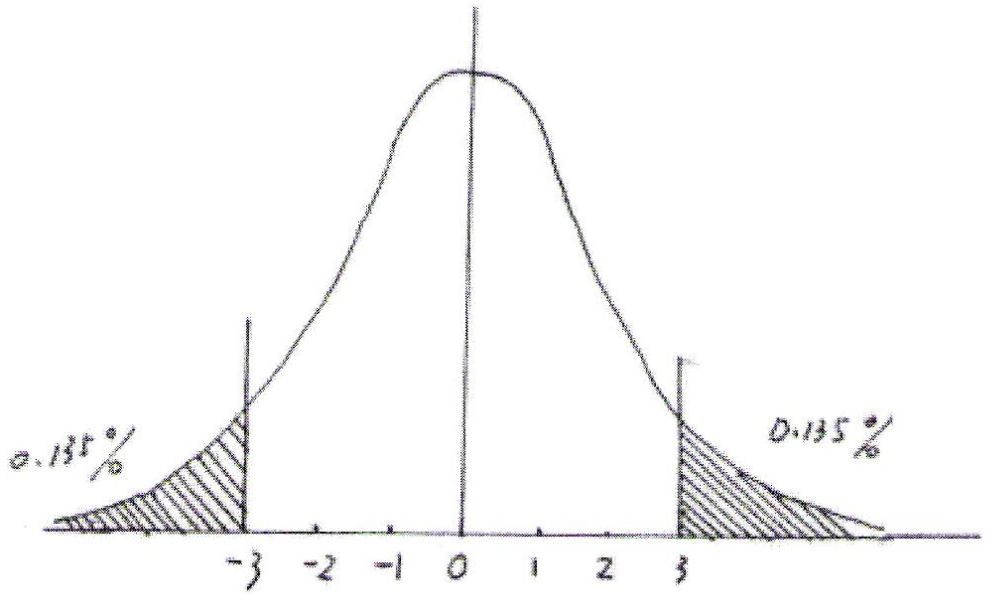

0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
497
498
499
500
501
502
503


p(x)
μ
μ+σ
x

- **标准正态分布：**

  μ=0且σ=1的正态分布，称为标准正态分布，记N（0，1），其变量记为U，概率密度函数记为ϕ(u)

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



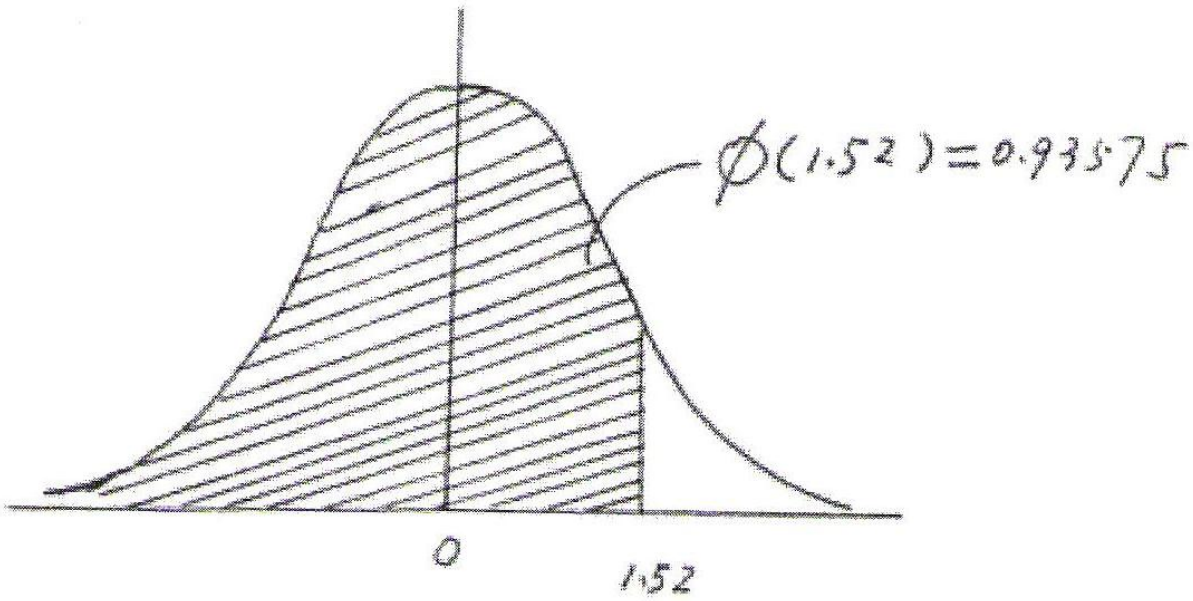
- **标准正态分布表及其应用**

—— **标准正态分布表**

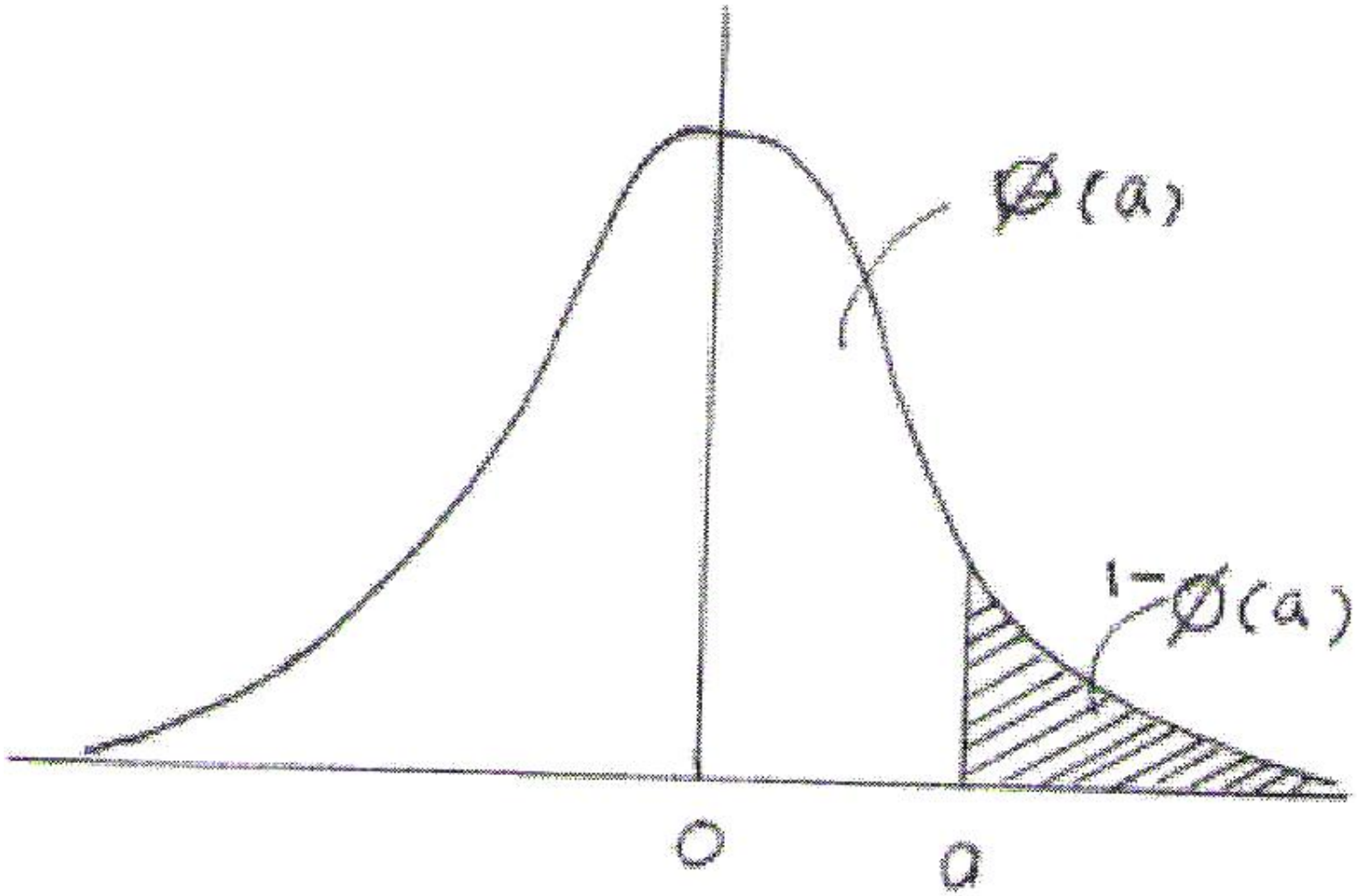
可用于计算形如“U≤u”随机事件发生的概率。

如：$P(U \le 1.52) = \Phi(1.52)$ 查附表得0.93575

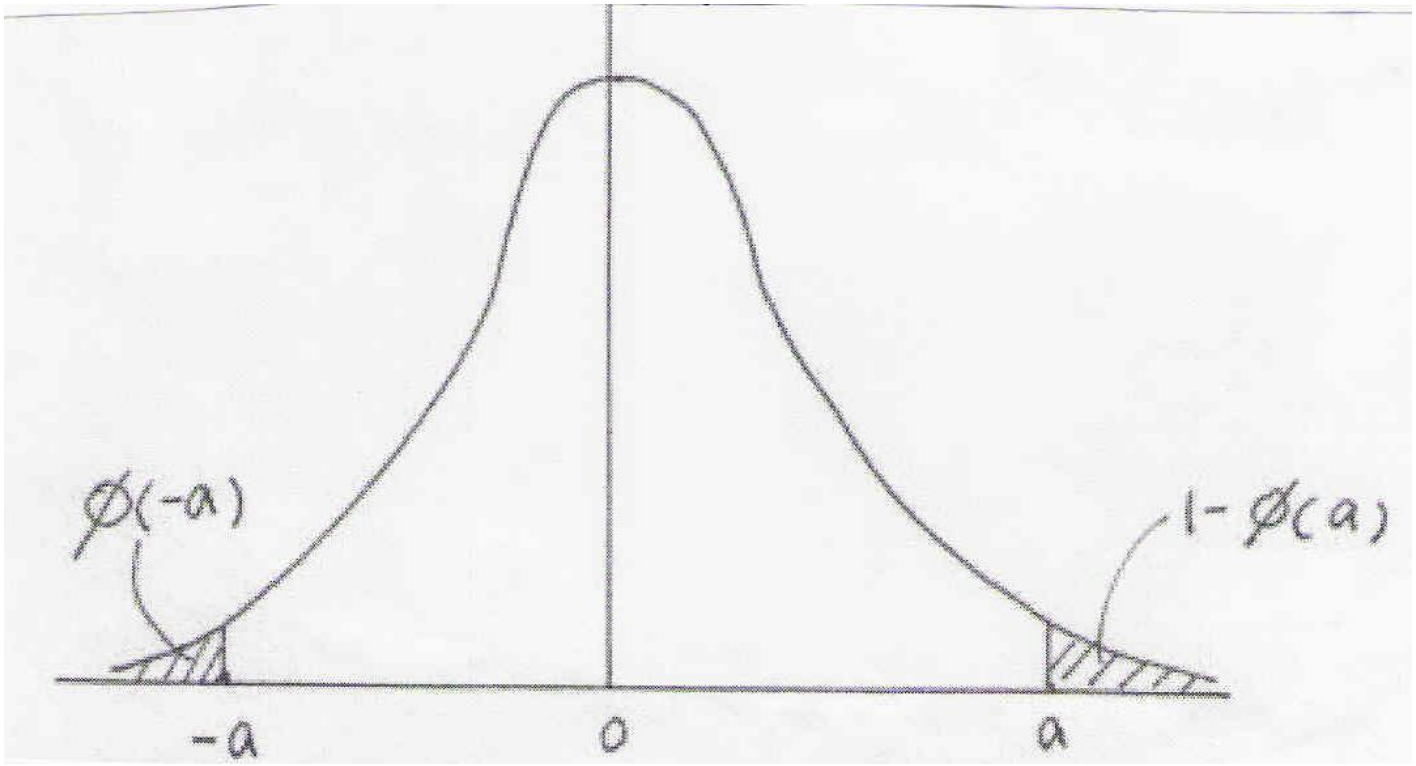
—— $P(U \le a) = p(U < a) = \Phi(a)$



—— $P(U > a) = 1 - \Phi(a)$



—— $\Phi(-a) = 1 - \Phi(a)$



——— $P(a \le U \le b) = \Phi(b) - \Phi(a)$

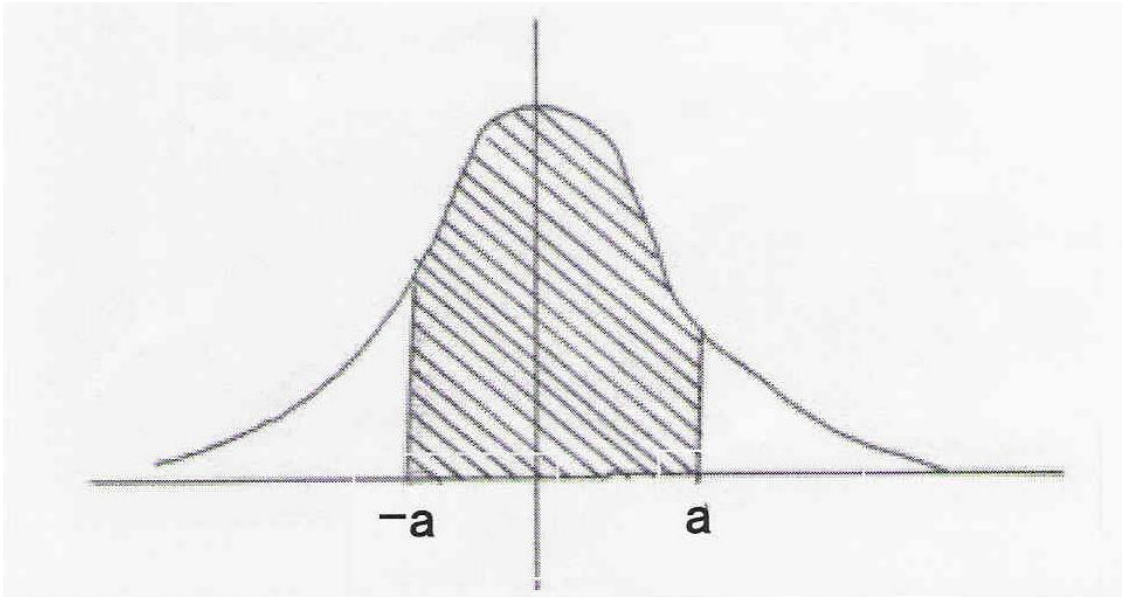
——— $P(|U| \le a) = 2\Phi(a) - 1$

$$P(|U| \le a) = P(-a \le U \le a)$$

$$= \Phi(a) - \Phi(-a)$$

$$= \Phi(a) - 1 + \Phi(a)$$

$$= 2\Phi(a) - 1$$



- 标准正态分布N(0，1)的分位数

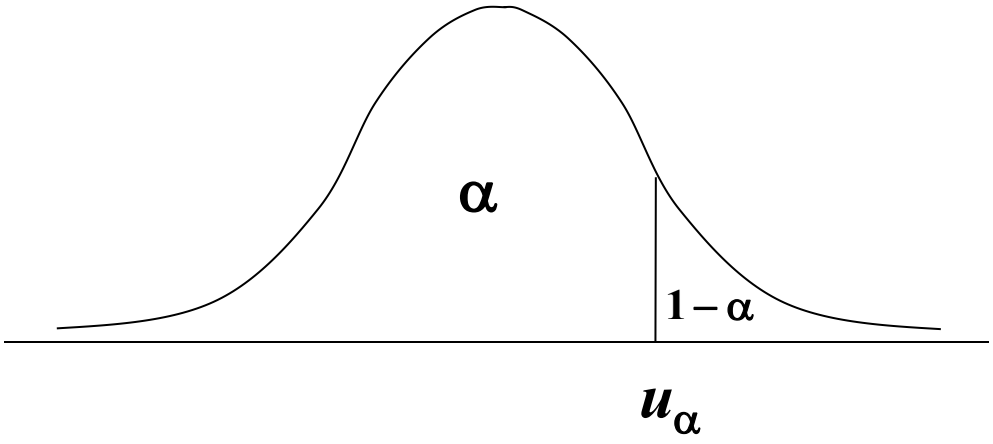
—— α分位数（α为0~1间实数）

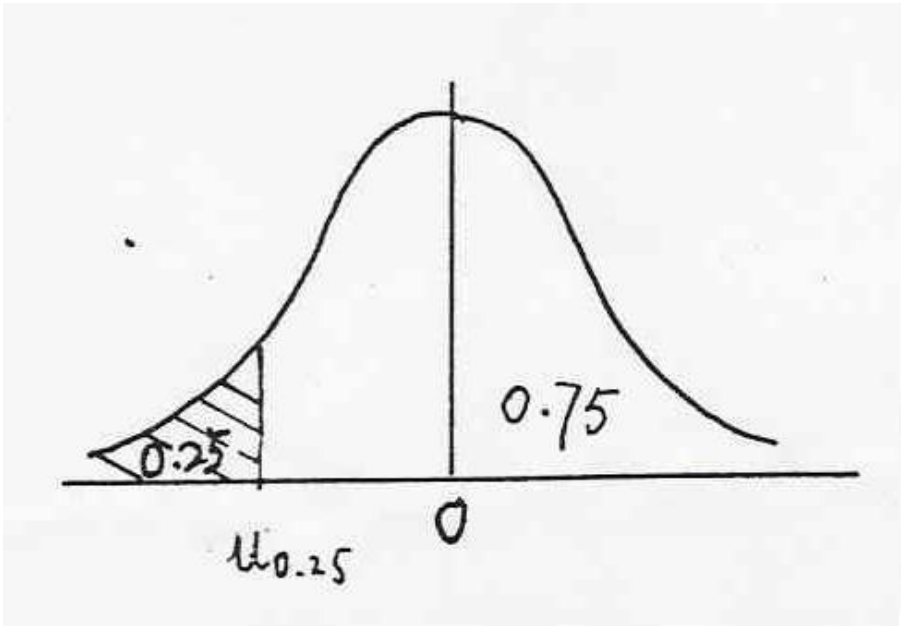
指它的左侧面积恰好为α，右侧面积恰好为1-α，即用概率表达 $P(U \le u_\alpha) = \alpha$

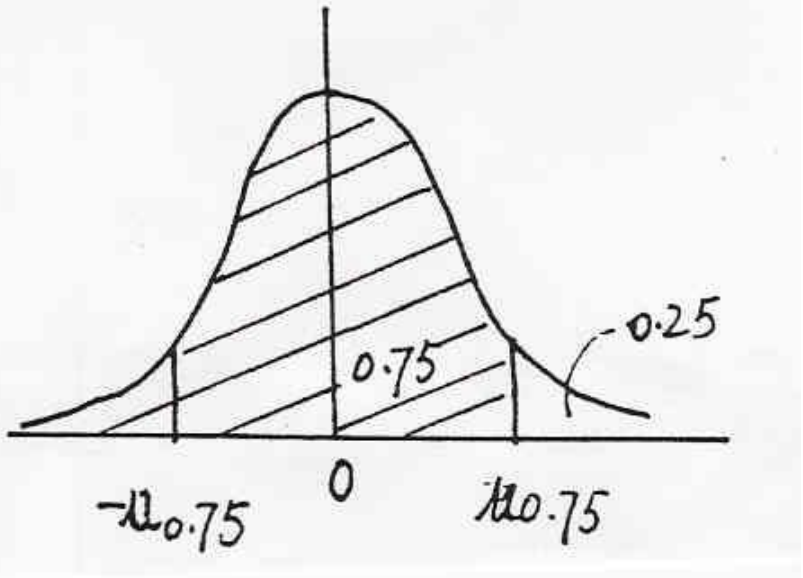
当α=0.5时，称为中位数，N(0，1)分布中$u_{0.5} \equiv 0$

α＜0.5时，如α=0.25则$u_{0.25} = -u_{0.75}$

—— 查附表 $u_{0.75}=0.675$，故$u_{0.25}=-0.675$


α
1 – α
u_α


0.25
0.75
0
$u_{0.25}$


0.25
0.75
0
$-u_{0.75}$
$u_{0.75}$

- **正态分布的计算**

性质1：设 $X \sim N(\mu,\sigma^2)$ **，则** $U=\dfrac{X-\mu}{\sigma}\sim N(0,1)$

性质2：设 $X \sim N(\mu,\sigma^2)$**，则对任意实数a，b有**

—— $P(X<b)=\Phi\left(\dfrac{b-\mu}{\sigma}\right)$

—— $P(X>a)=1-\Phi\left(\dfrac{a-\mu}{\sigma}\right)$

—— $P(a<X<b)=\Phi\left(\dfrac{b-\mu}{\sigma}\right)-\Phi\left(\dfrac{a-\mu}{\sigma}\right)$

例1. 设 $X \sim N(10,4)$

求 $P(8 < X < 14)$

$$P(8 < X < 14) = \Phi(\frac{14-10}{2}) - \Phi(\frac{8-10}{2}) = \Phi(2) - \Phi(-1)$$

$$= \Phi(2) - [1 - \Phi(1)] = \Phi(2) + \Phi(1) - 1$$

- **不合格品率**

为产品质量特性X超出规范限（T_L，T_U）的概率

—— **X超出T_U（上规范限）的概率记P_U**

$$p_U = P(X > T_U)$$

—— **X超出T_L（下规范限）的概率记P_L**

$$p_L = P(X < T_L)$$

—— **X的不合格品率$P = P_U + P_L$**

1. 10　正态分布的分位数的概念

设我国成年人身高大约服从正态分布，有95%的人身高低于185cm，那么185cm就是该正态分布的95%分位数。

记为：

$$q_{0.95}=185cm$$

设我国成年人体重大约服从正态分布，有99%的人体重低于100kg，那么100kg就是该正态分布的99%分位数。

记为：

$$q_{0.99}=100kg$$

- 计算不合格品率要知道两件事：

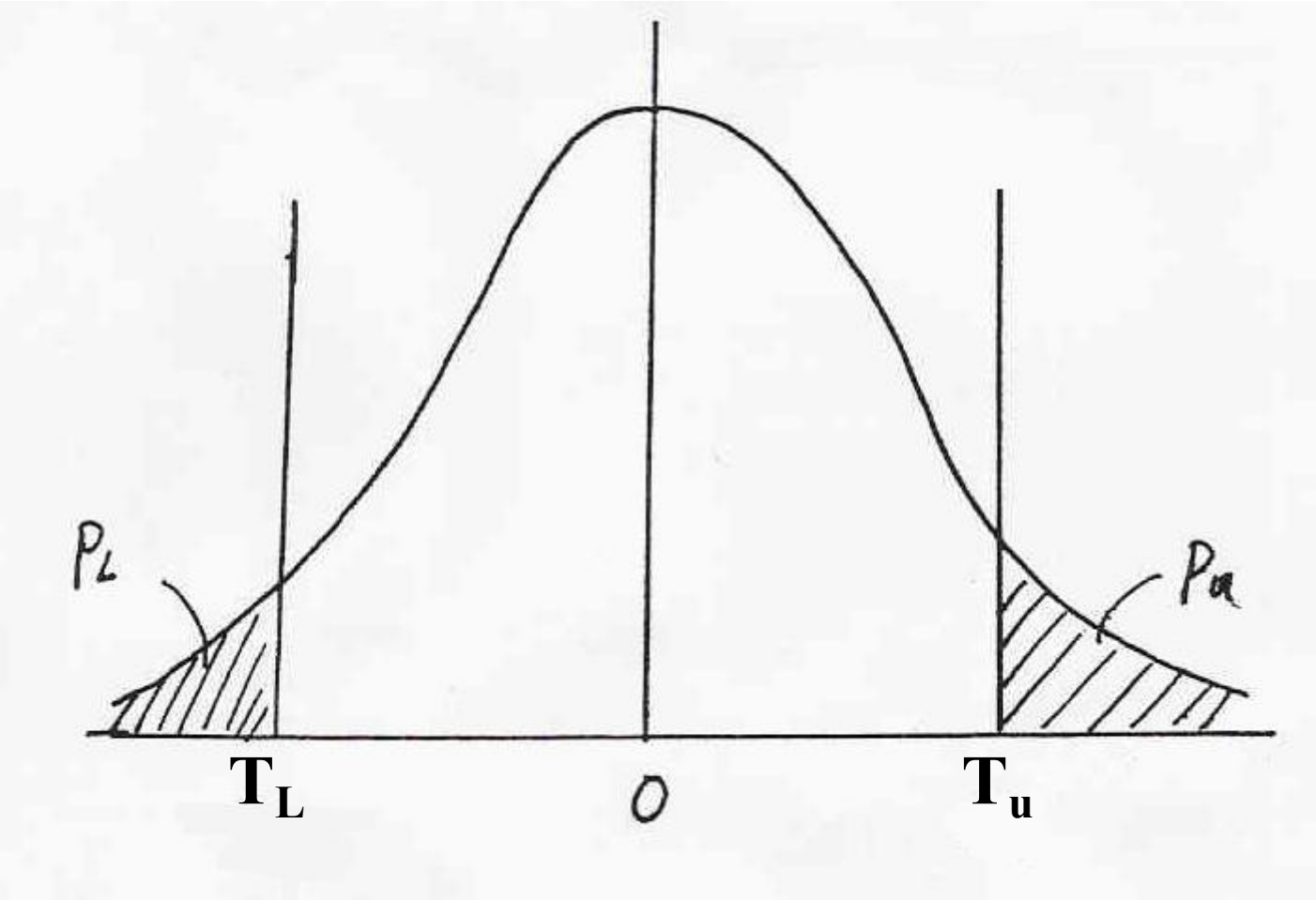
—— 质量特性X的分布，在过程受控情况下，常为正态分布$N(\mu,\sigma^2)$

—— 产品规范限，是对产品质量特性所作的要求，这些要求可能是顾客要求；可能是标准；可能是企业规定的技术要求。

则：$$p_U = P(X > T_U) = 1 - \Phi(\frac{T_U - \mu}{\sigma})$$

$$p_L = P(X < T_L) = \Phi(\frac{T_L - \mu}{\sigma})$$

其中$\Phi(\bullet)$可查标准正态分布函数表


P_L
P_u
T_L
0
T_u

● 当正态分布中心μ=规范中心 $M=\left(\frac{T_L+T_U}{2}\right)$ 时产品质量特性X超出规范μ±3σ的不合格率

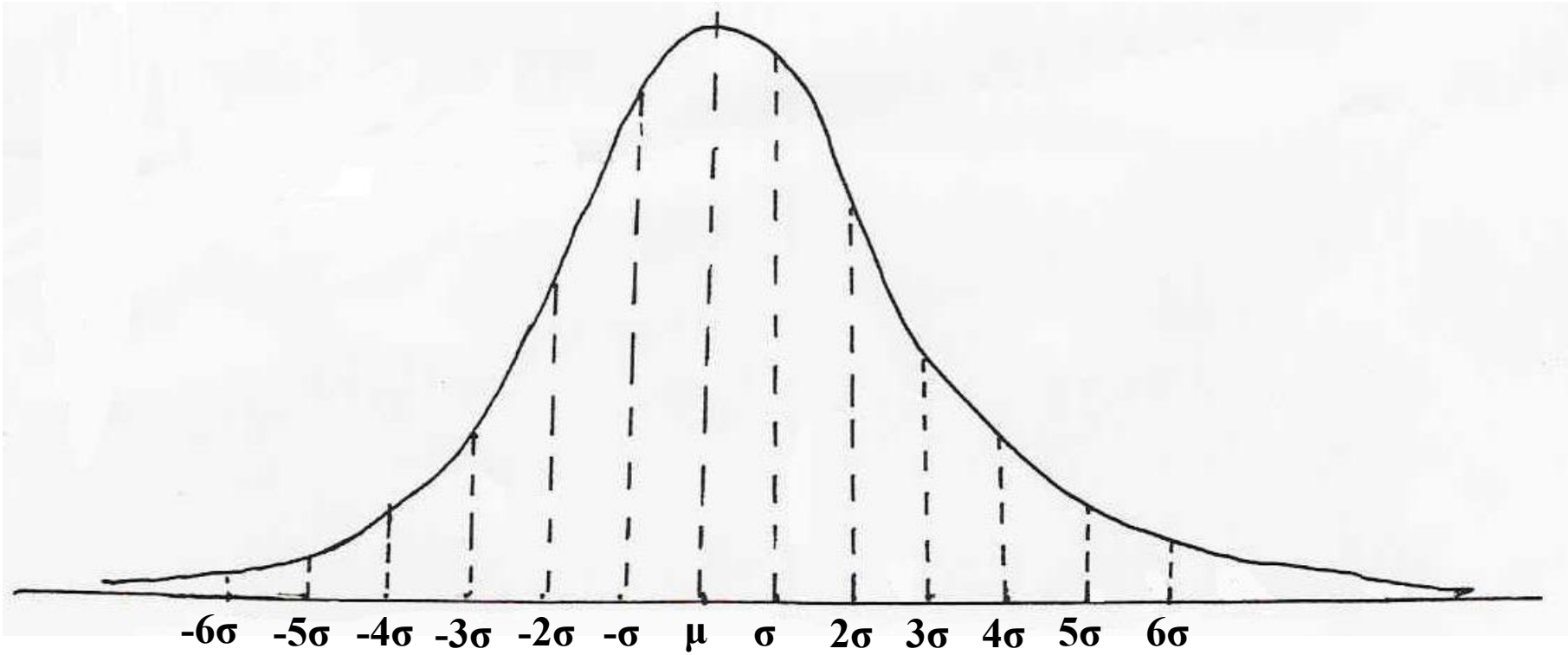
$p_L=P(x<\mu-3\sigma)=\Phi(-3)=1-\Phi(3)$

$=1-0.99865=0.00135=1350PPm$

$p_U=P(x<\mu+3\sigma)=1-\Phi(3)$

$=0.00135=1350PPm$

$p=p_L+p_U=0.00135+0.\ 00135=0.0027=2700PPm$



规范限	合格品率（%）	不合格品率（ppm）
±1σ	68.27	317300
±2σ	95.45	45500
±3σ	99.73	2700
±4σ	99.9937	63
±5σ	99.999943	0.57
±6σ	99.9999998	.002

- 指数分布：**重要的质量特性分布（可靠性）**

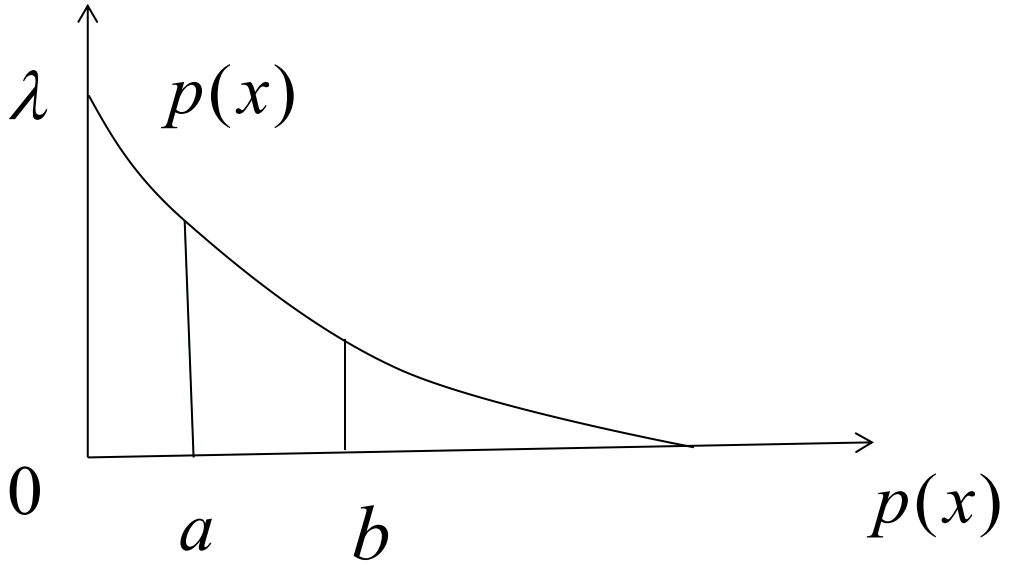
例子：

1）设备维修时间

2）排队等候服务时间

3）电子元器件寿命

4）一次通话时间



4.1 指数分布的记法 $X \sim Exp(\lambda)$

4.2 指数分布的分布函数 $1-e^{-\lambda x} \qquad x \geq 0$

4.2 指数分布的密度函数

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ \\ 0, x < 0 \end{cases}$$

4.5 指数分布的均值、方差与标准差

$$E(X)=\frac{1}{\lambda}$$

$$Var(X)=\frac{1}{\lambda^2}$$

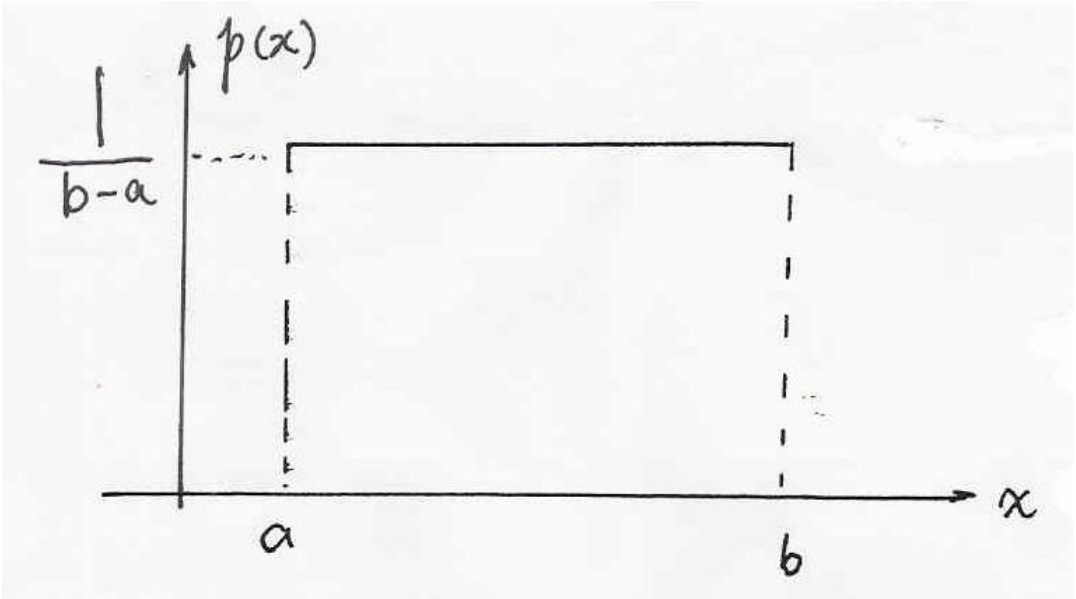
$$\sigma(X)=\frac{1}{\lambda}$$

（三）其他连续分布

- **均匀分布**

—— 在区间（a，b）上的均匀分布，记U（a、b）

$$p(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b\\ 0, & \text{其它}\end{cases}$$



—— 均值、方差、标准差

均值 $E(X)=\dfrac{a+b}{2}$

方差 $Var(X)=\dfrac{(b-a)^2}{12}$

标准差 $\sigma=\sqrt{\dfrac{(b-a)^2}{12}}$

- 对数正态分布

3.1对数正态分布的记法

$$\ln X \sim N(\mu, \sigma^2)$$

3.2 对数正态分布的密度函数（了解即可）

$$p(x) = \begin{cases} \dfrac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

3.3 对数正态分布的均值、方差

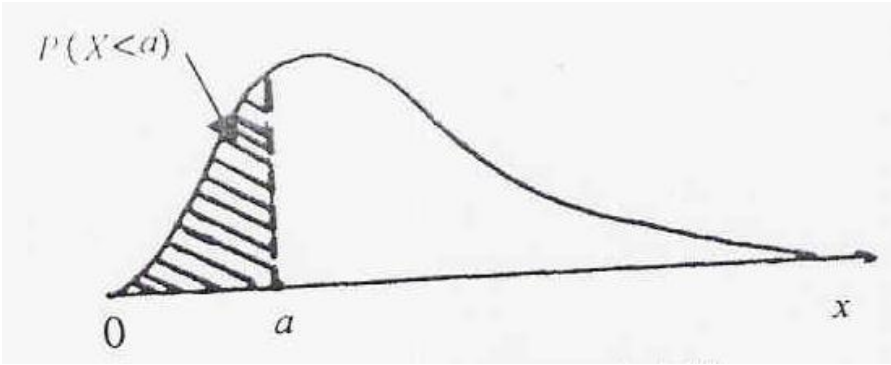
$$E(X) = e^{\mu+\sigma^2/2}$$

$$Var(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$$

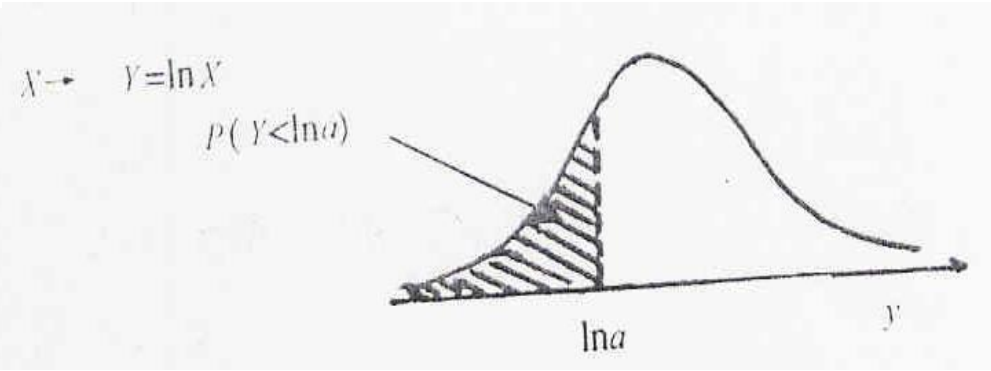
- 对数正态分布（特点）

—— 随机变量都在正半轴（0,+∞）上取值

—— 大量取值在左边，少量取值在右边，且很分散，这样的分布称之为右偏分布。（曲线的尾巴在右边）



对数正态分布密度函数



正态分布的密度函数

五、中心极限定理

- 随机变量的独立性

随机变量X_1与X_2相互独立是指其中一个取什么值不影响另一个的取值，或者说是指两个随机变量独立的取值，互不影响。

随机变量的独立性可以推广到3个或更多个随机变量。

● 中心极限定理

定理1、 即n个相互独立同分布的随机变量X_1，X_2， ……X_n，若 $X_i \sim N(\mu, \sigma^2)$

则

其样本均值 $\bar{X}$ 服从正态分布N（μ，$\frac{\sigma^2}{n}$）。

即：独立同分布正态变量X的样本均值 $\bar{X}$ 服从正态分布N（μ，$\frac{\sigma^2}{n}$）

- **中心极限定理**

在统计中，多个相互独立随机变量的平均值（仍然是一个随机变量）将服从或近似服从正态分布。

定理2， n个相互独立同分布的随机变量X_1，X_2，……X_n，均值μ和方差σ^2都存在，则在n较大时，其样本均值$\bar{X}$服从或近似服从正态分布$N(\mu, \frac{\sigma^2}{n})$。

第三节　统计基础知识

一、总体、个体与样本

（一）总体与个体

总体：在一个统计问题中，我们把研究对象的全体成为总体。

—— 当研究产品某个特定的质量特性X时，也常把全体产品的特性看做为总体。

个体：构成总体的每个成员。

—— 当研究产品的某个特定的质量特性X时，把一个具体产品的特性值x视为个体。

（二）随机样本

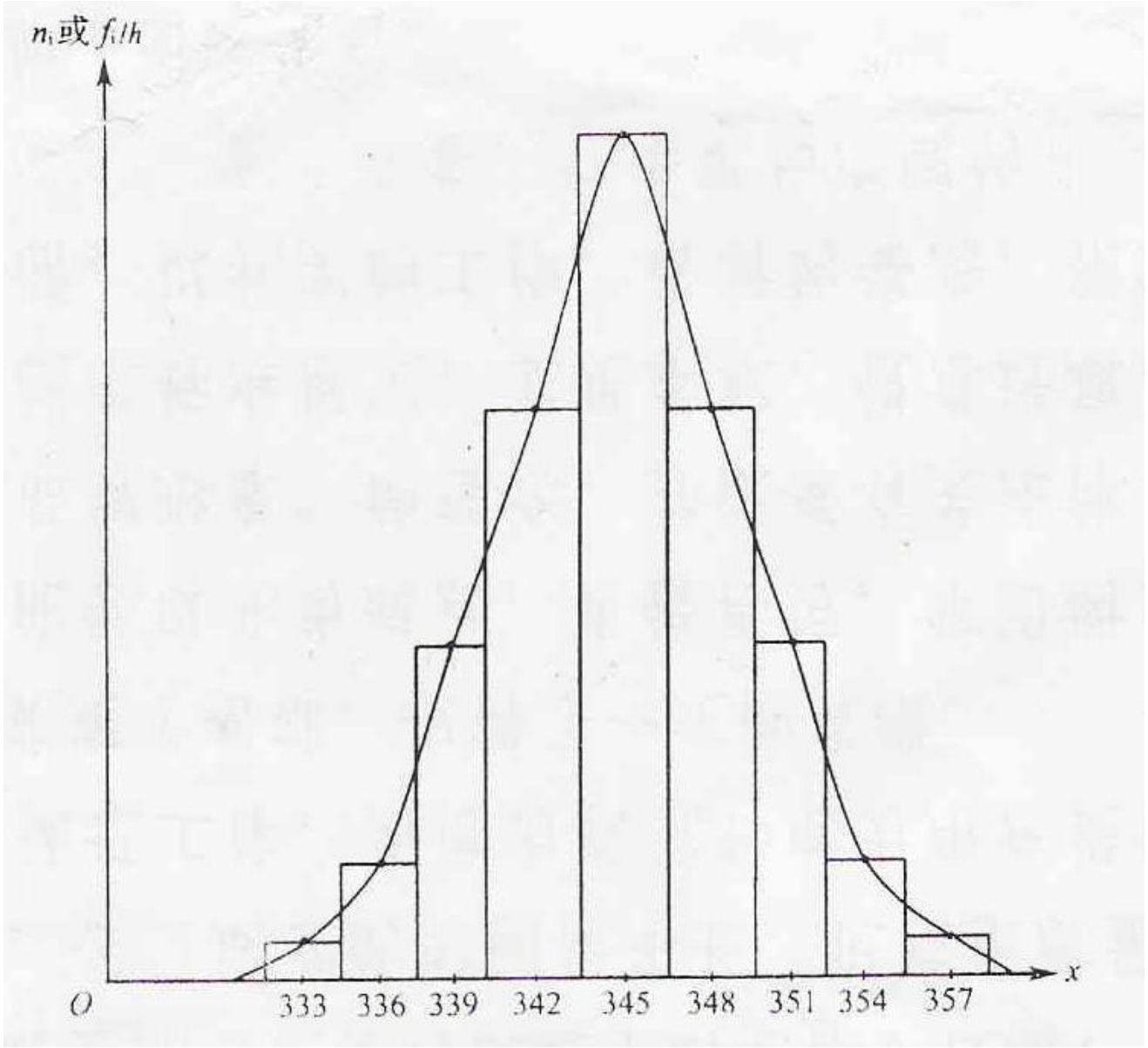
满足下面两个条件的样本称为简单随机样本，简称随机样本：

1. 随机性。总体中每个个体都有相同的机会入样。
2. 独立性。从总体中抽取的每个样品对其它样本的的抽取无任何影响。

- 随机样本可看做n个相互独立的、同分布的随机变量，其分布与总体分布相同。
- 下面所述的样本都是指满足这两个要求的简单随机样本。

二、频数（频率）直方图

为了研究数据的变化规律，需要对数据进行一定的加工整理。直方图是为研究数据变化规律而对数据进行加工整理的一种基本方法。



频数（频率）直方图

（二）直方图的观察与分析

a. 对称型

b. 偏态型

c. 孤岛型

d. 锯齿型

e. 平顶型

f. 双峰型

三、统计量与抽样分布

1．统计量的概念

不含未知参数的样本函数

- 样本均值、样本中位数、样本极差、样本方差、样本标准差及样本变异系数等都是统计量，只有众数除外。

2．抽样分布

统计量的分布称为抽样分布

（一）样本数据集中位置的统计量

（1）样本均值 $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

（2）样本中位数M_e(或 $\tilde{x}$)

$$Me(\tilde{x}) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{，n为奇数} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right] & \text{，n为偶数} \end{cases}$$

（3）众数（Mod）

数据中出现频率最高的值。

（二）描述样本数据分散程度的统计量

（1）样本极差

$$R = x_{(n)} - x_{(1)}$$

（2）样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

—— 因为n个离差（$x_i - \bar{x}$）的总和为零，所以对于n个独立数据，独立的离差个数只有n-1个，称n-1为离差（或离差平方和）的自由度。故方差用离差平方和除以n-1。

简化计算公式：$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$

或 $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$

（3）样本标准差

$$S = \sqrt{S^2}$$

- 标准差的量纲与数据的量纲一致

（4）样本变异系数

$$C_\upsilon = \frac{s}{\bar{x}}$$

四、常用抽样分布

1．$\overline{X}$ 的分布

（1）$\overline{X}$ 的精确分布

设X服从N（μ，σ^2），（x_1，x_2，……，x_n）是由总体X中抽取的一个样本，则服从

N（μ，σ^2/n）

（2）$\overline{X}$ 的渐进分布

设X为任意分布，（x_1，x_2，……，x_n）是由总体X中抽取一个样本，若 $E(x_i)=\mu$ ，$Var(x_i)=\sigma^2\neq 0$ ，则当n→∞时，$\overline{X}$ 近似服从N（μ，σ^2/n）。

（3）χ^2 —— 分布

设X服从N(0，1)，且设（x_1，x_2，……，x_n）是由总体X中抽取的一个样本，则

$$\chi^2 = \chi_1^2 + \chi_2^2 + \cdots + \chi_n^2$$

服从自由度为n的 χ^2分布，记作 $\chi^2 \sim \chi^2(n)$。

- 设X服从N（μ，σ^2），则 $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

（3）t ——分布

设随机变量X，Y相互独立，X~N(0，1)，Y~ x^2(n)则 $t=\frac{X}{\sqrt{Y/n}}$ 服从自由度为n的t—分布记作t~t(n)

- 设X~N（μ， σ^2），（x_1，x_2，……，x_n）是由总体X中抽取的一个样本，则

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$$

- 设X和Y相互独立，且X~N（μ，σ^2），Y~N（μ，σ^2），（x_1，x_2，……，x_{n1}）与（y_1，y_2，……，y_{n2}）分别由总体X和Y中抽取的样本，则

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t(n_1+n_2-2)$$

（4）F—— 分布

设X与Y相互独立，且X～$\chi^2(N_1)$，Y～$\chi^2(N_2)$

则 $F=\frac{X/N}{Y/N_2}$服从自由度为（N_1，N_2）的F—— 分布。

记作 F～F(N_1，N_2)。

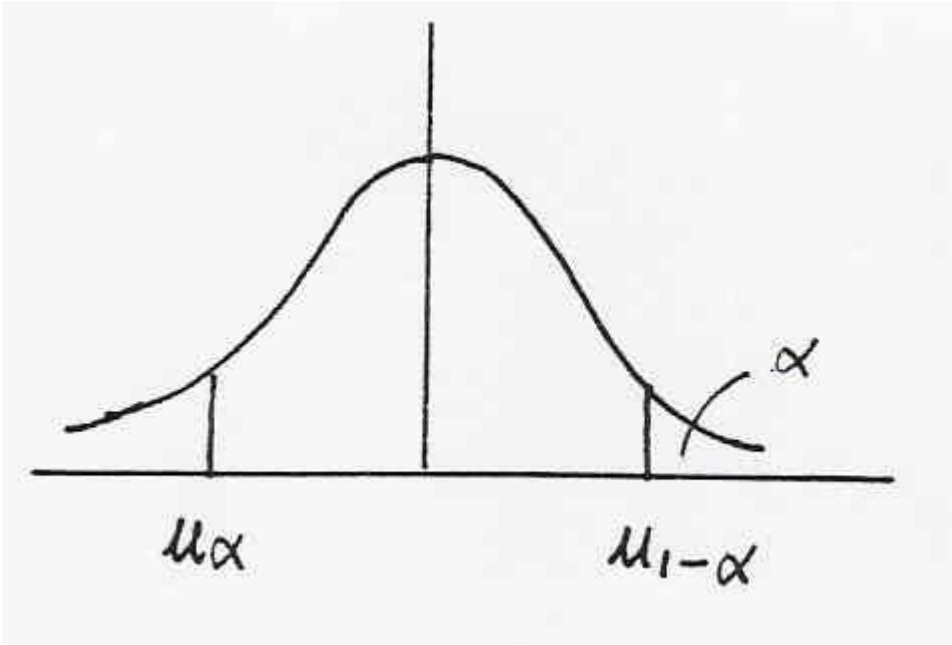
- 设X和Y相互独立，$X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，$(x_1，x_2，\cdots\cdots，x_n)$与$(y_1，y_2，\cdots\cdots，y_m)$分别由X和Y中抽取的样本，则

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F(n-1，m-1)$$

当$\sigma_1^2=\sigma_2^2=\sigma^2$时，则

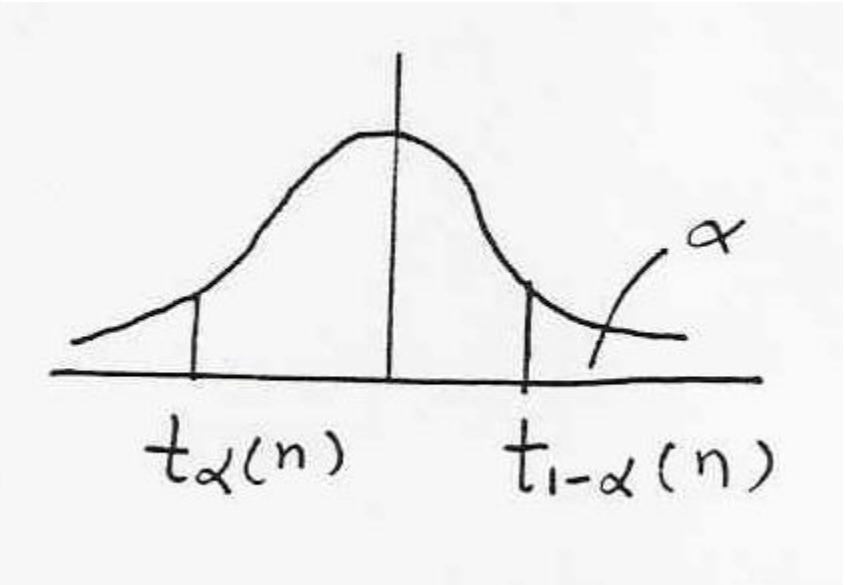
$$\frac{S_1^2}{S_2^2}\sim F(n-1,m-1)$$

正态分布



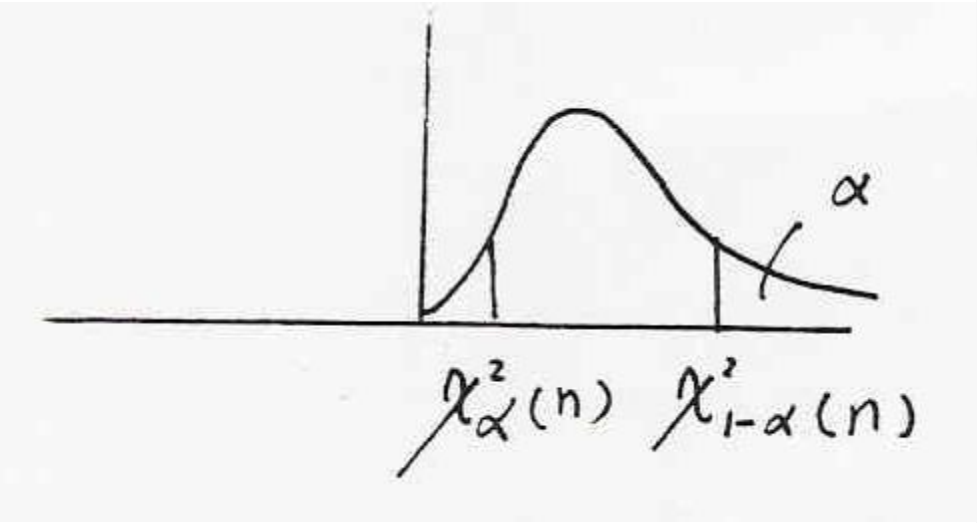
$$\mu_{\alpha} = -\mu_{1-\alpha}$$

t 分布



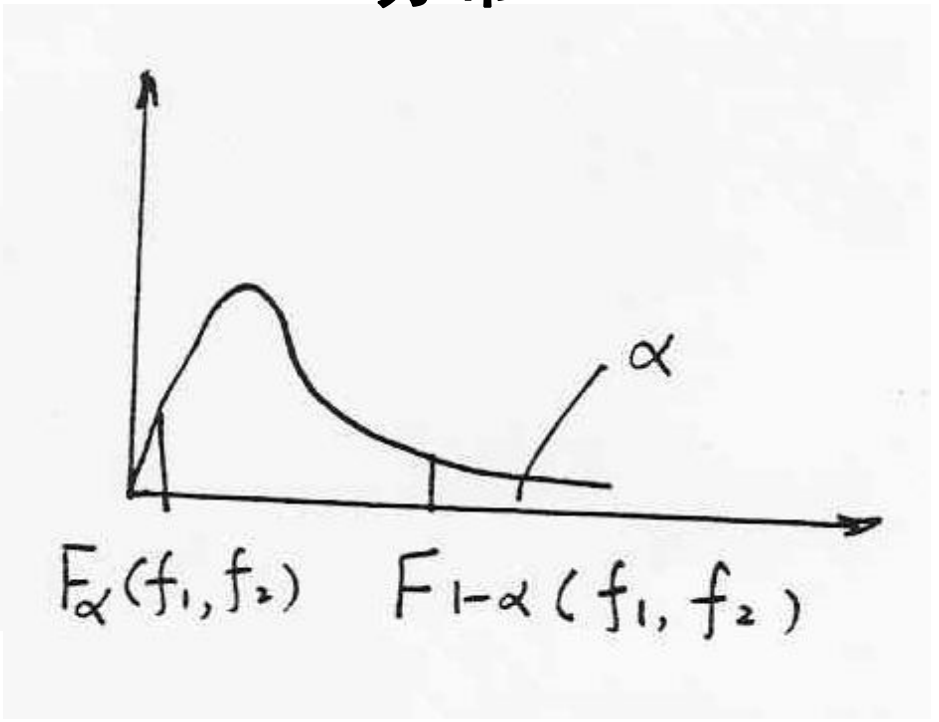
$$t_{1-\alpha(n)} = -t_{\alpha(n)}$$

χ^2 分布



$$\chi^2_{1-\alpha(n)} \neq \chi^2_{\alpha(n)}$$

F分布



$$F_{1-\sigma}(f_1, f_2) \neq -F_{\alpha}(f_1, f_2)$$

第四节　参数估计

一、点估计

1．概念

设 θ 是一个未知参数，$(X_1, X_2, \cdots\cdots, X_n)$由总体X中抽取的样本，则用 $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots\cdots, X_n)$来估计 θ，则称 $\hat{\theta}$ 为 θ 的估计量（或称估计）。

2．矩法估计

（1）用样本矩估计相应总体矩；

（2）用样本矩的函数估计相应总体矩的函数。

例如用样本均值估计总体均值；用样本方差（标准差）来估计总体方差（标准差）。

3. 点估计优劣的评选标准

（1）无偏性

设 $\hat{\theta}$ 是θ的一个估计量，若 $E(\hat{\theta})=\theta$，则称 $\hat{\theta}$ 是θ的无偏估计。

（2）有效性

设 $\hat{\theta}_1,\hat{\theta}_2$ 都是θ的无偏估计量，若对一切θ的可能取值有：

$Var(\hat{\theta}_1)\le Var(\hat{\theta}_2)$，且至少有一个 θ_0，严格不等号成立，则 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

（3）正态总体参数的无偏估计

① μ的无偏估计有两个，即 $\bar{x}$ 和 $\tilde{x}$ 。

② σ^2的无偏估计常用的只有一个，即 S^2 。

③ σ的无偏估计有两个，即 R/d_2 和 S/C_4

二、区间估计

（一）区间估计的概念

设θ是总体分布中的未知参数，其一切可能取值组成的参数空间为Θ，从总体中抽取一个样本$(x_1, x_2, \cdots\cdots, x_n)$，对给定的 $\alpha(0<\alpha\leqslant 1)$确定两个统计量：$\theta_L=\theta_L(x_1,x_2,\cdots,x_n)$与

$$\theta_u=\theta_u(x_1,x_2,\cdots,x_n)$$

对任意的$\theta\in\Theta$有

$$P(\theta_L\leq\theta\leq\theta_u)\geq 1-\alpha$$

则称［θ_L，θ_u］是θ的置信水平为$1-\alpha$的置信区间。

- 1-α 置信区间的含义：

所构造的一个随机区间$[\theta_L,\theta_U]$能包含未知参数θ 的概率为1-α 。由于这个随机区间会随样本观察值的不同而不同，它有时包含了参数θ，有时没有包含 θ，但是用这种方法作区间估计时，100次中大约有100(1- α)个区间能包含未知参数θ 。

（二）一个正态总体均值与方差的置信区间

（1） σ^2已知，求 μ 的置信区间

μ 的1- α置信区间为：

$$\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

（2） σ^2未知，求 μ 的置信区间

$$\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

（3）方差 σ^2 的1- α 的置信区间（ μ 未知）

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}$$

（4）标准差 σ 的1- α 的置信区间（未知）

$$\frac{S\sqrt{n-1}}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}} \le \sigma \le \frac{S\sqrt{n-1}}{\sqrt{\chi^2_{\alpha/2}(n-1)}}$$

（三）比例p的置信区间（大样本场合）

设总体 $X \sim b(1, p)$ ，样本为x_1，x_2，…，x_n，样本之和为K，样本均值为 $\bar{x} = \frac{K}{n}$ 则

$$\hat{p} = \frac{K}{n} \quad \text{（点估计）}$$

当n相当大时，$\bar{x} \sim N\left(p, p(1-p)/n\right)$ ，故p的 $1-\alpha$ 置信区间。

$$\bar{x} - u_{1-\frac{\alpha}{2}} \sqrt{\bar{x}\left(1-\bar{x}\right)/n} \le p \le \bar{x} + u_{1-\frac{\alpha}{2}} \sqrt{\bar{x}\left(1-\bar{x}\right)/n}$$

其中 $u_{1-\frac{\alpha}{2}}$ 是标准正态分布的 $1-\frac{\alpha}{2}$ 分位数。