第四章第一部分
概率与概率分布

1.试验结果及其关系

2.试验结果随机不确定性的度量

3.离散型随机变量的概率分布

4.连续型随机变量的概率分布

概率论的由来

- 1654年,一个名叫梅累的骑士就“两个赌徒约定赌若干局, 且谁先赢c局便算赢家, 若在一赌徒胜 a局 （$a<c$）,另一赌徒胜b局（$b<c$）时便终止赌博,问应如何分赌本” 为题求教于帕斯卡, 帕斯卡与费尔马通信讨论这一问题。1657年，荷兰数学家、物理学家惠更斯也试图解决帕斯卡和费尔马通信中所提出的问题，在他撰写的《论赌博中的计算》一书中，第一次建立了概率和数学期望的概念。

在生活中经常会遇到这样一些关于不确定性的问题，比如：

- （1）在新产品上市前，经销商需要知道顾客是否会购买这种产品。为了降低风险，经销商通常进行市场调查，如随机抽取500人进行调查，询问他们对新产品的反应以及可能的建议。
- （2）汽车行业最近竞争很激烈。企业为确保达到汽车整车性能检测标准，需要从生产的汽车中选取一些样本进行检测。显然，如果所有汽车都接受检测，那么就没有用于销售的汽车了。于是随机抽取若干辆汽车进行检测，根据检测结果，可得知汽车性能是否达标。
- （3）在经济危机的情况下，是否可以购买新的汽车或房子呢？

一、试验结果及其关系

1.概率论的研究对象

- 随机现象：在一定条件下，并不总是出现相同结果的现象。如抛一枚硬币、掷一颗骰子等。
- 特点：

  （1）结果不止一个

  （2）哪一个结果出现，人们事先并不知道
- 确定性现象：只有一个结果的现象。

2.概率论的研究目的

- 探寻随机现象在一定概率意义下的变化规律，即统计规律

3.概率论的研究方法

- 随机试验：是导致所有可能观测中有且仅有一个出现的过程。其特点表现为

  （1）可以在相同的条件下重复地进行（重复性）

  （2）每次试验的可能结果不止一个，并且能事先明确试验的所有可能结果（明确性）

  （3）进行一次试验之前不能确定哪一个结果会出现（随机性）

- 概率论通过观察试验结果，探寻随机现象的统计规律

4.概率论的逻辑思路

- 首先，描述试验的结果及其关系
- 其次，度量试验结果的随机不确定性
- 最后，描述出随机现象变化的统计规律

描述试验结果的若干概念

- 基本事件（样本点）
  - 随机试验的每一个“不可能再分”的可能结果
- 样本空间
  - 随机试验的所有基本事件构成的集合，用Ω表示
- 随机事件
  - 常简称事件，指每次试验可能出现也可能不出现的试验结果，样本空间中的某些样本点组成的集合，是样本空间的一个子集，用大写英文字母A、B、C等表示
  - 例如，在掷一枚骰子的试验中，有6种可能的结果（样本点），因为有1到6个点数，但是有很多可能的事件（点数为偶数等）。

- 必然事件
  - 每次试验一定出现的试验结果，样本空间的最大子集，用Ω表示
- 不可能事件
  - 每次试验一定不出现的事件，样本空间的最小子集，用Φ表示
- 随机变量
  - 取值不确定的变量，试验的每一个“不可能再分”的可能结果，即样本点的数值性描述，用英文大写字母X、Y、Z等表示

试验结果的三种表述方法

- 语言
- 概念（集合）
- 随机变量的取值

二、试验结果随机不确定性的度量

- 基本度量方法------概率
- 什么是概率？
  - 概率与某事件发生的机会、可能性或确定程度有关。
  - 概率(probability)就是一个数字。介于0和1之间，描述一个事件发生的可能性大小。
  - 事件*A*的概率表示为*P*(*A*)
  - 小概率(接近零)的事件很少发生，而大概率(接近1)的事件则经常发生。
  - 例如：一个人中一次彩票一等奖的概率很小；一年中至少有一场飓风袭击我国沿海地区的概率就很大，因为在大部分年份中都多于一场飓风发生。

- 概率的公理化定义
  - 设Ω为一个样本空间，Θ为Ω上的某些子集组成的一个事件域（可测域）。如果对任一事件A（A属于Θ），定义在Θ上的一个实值函数P(A)满足：
    - 1、非负性公理 若A属于Θ，则P(A) ≥0
    - 2、正则性公理 P(Ω)=1
    - 3、可列可加性公理 若$A_1,A_2,\ldots,A_n,\ldots$互不相容，有

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right)=\sum_{i=1}^{+\infty} P(A_i)$$

则称P(A)为事件A的概率。

概率的加法公式

- ➡ 公式一

1. 两个互斥事件之和的概率，等于两个事件概率之和。设A和B为两个互斥事件，则

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

2. 事件A_1，A_2，…，A_n两两互斥，则有

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

- ➡ 公式二
- 对任意两个随机事件A和B，它们和的概率为两个事件各自概率的和减去两个事件交的概率，即

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

条件概率
(conditional probability)

- ➡ 在事件B已经发生的条件下，求事件A发生的概率，称这种概率为事件B发生条件下事件A发生的条件概率，记为

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

- 由于样本空间的改变而产生了条件概率

【例】100件产品中，有80件正品，20件次品；而80件正品中有50件一等品，30件二等品。现从这100件产品中任取1件，问：它是正品的概率？它是一等品的概率？已知它是正品，是一等品的概率？

解：设事件A表示“取到一等品”，事件B表示“取到正品”。

$$P(B)= 80/100=0.8$$

$$P(AB)= 50/100=0.5$$

$$P(A\mid B)=\frac{50}{80}=\frac{50/100}{80/100}=\frac{P(AB)}{P(B)}$$

概率的乘法公式

1. 用来计算两事件交的概率
2. 以条件概率的定义为基础
3. 设 A、B 为两个事件，若 $P(B)>0$，则 $P(AB)=P(B)P(A|B)$，或 $P(AB)=P(A)P(B|A)$

概率的乘法公式(例题分析)

【例】设有1000件产品，其中850件是正品，150件是次品，从中依次抽取2件，两件都是次品的概率是多少？

解： 设 A_i 表示“第 i 次抽到的是次品”(i=1,2)，所求概率为$P(A_1A_2)$

$$P(A_1A_2) = P(A_1)P(A_2 \mid A_1)$$

$$= \frac{150}{1000} \cdot \frac{149}{999} = 0.0224$$

事件的独立性(independence)

1. 一个事件的发生与否并不影响另一个事件发生的概率，则称两个事件独立
2. 若事件 A 与 B 独立，则 $P(B|A)=P(B)$，$P(A|B)=P(A)$
3. 此时概率的乘法公式可简化为

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

4. 推广到多个独立事件，对于任何的n<N

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

事件的独立性(例题分析)

【例】某工人同时看管三台机床，每单位时间(如30分钟)内机床不需要看管的概率：甲机床为0.9，乙机床为0.8，丙机床为0.85。若机床是自动且独立地工作，求

（1）在30分钟内三台机床都不需要看管的概率

（2）在30分钟内甲、乙机床不需要看管，且丙机床需要看管的概率

解：设 A_1，A_2，A_3为甲、乙、丙三台机床不需要看管的事件，$\bar{A}_3$ 为丙机床需要看管的事件，依题意有

(1) $P(A_1A_2A_3)= P(A_1)\cdot P(A_2)\cdot P(A_3)=0.9\times0.8\times0.85=0.612$

(2) $P(A_1A_2\ \bar{A}_3)= P(A_1)\cdot P(A_2)\cdot P(\ \bar{A}_3)$

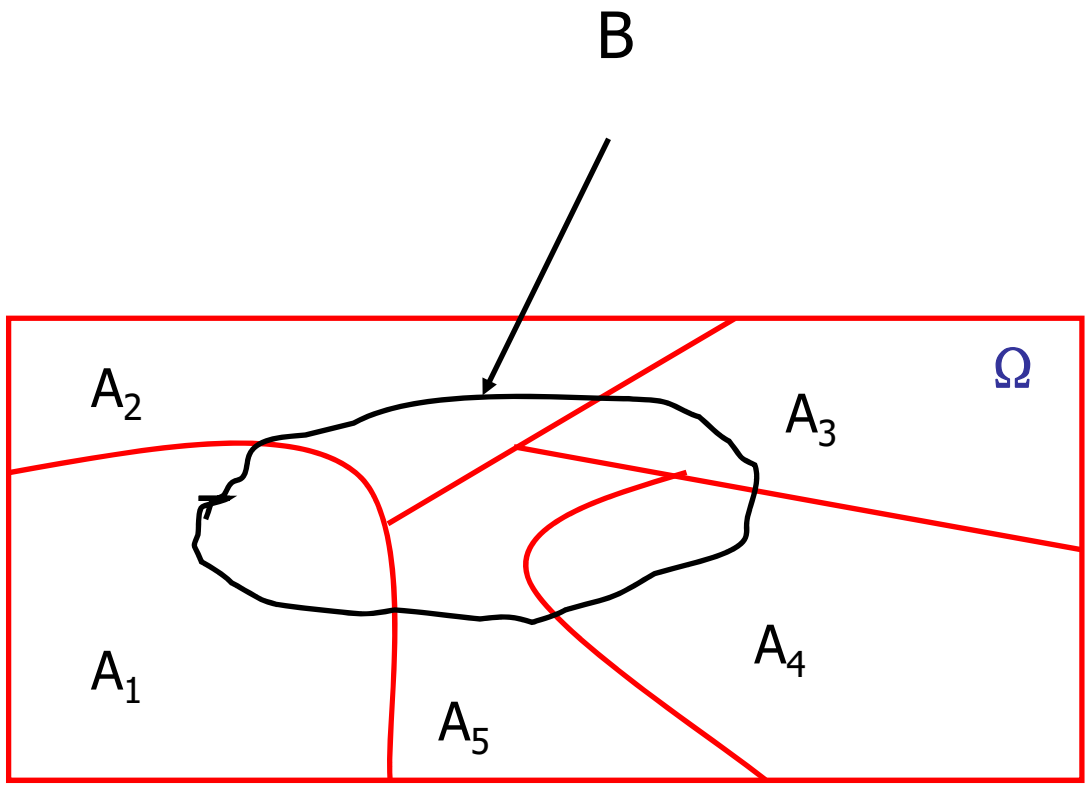
$= 0.9\times0.8\times(1-0.85)=0.108$

全概率公式

- ➡ 设事件A_1，A_2，…，A_n 两两互斥，$A_1+A_2+\ldots+A_n=\Omega$（满足这两个条件的事件组称为一个完备事件组），且$P(A_i)>0(i=1,2,\ldots,n)$，则对任意事件$B$，有

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

我们把事件A_1，A_2，…，A_n 看作是引起事件B发生的所有可能原因，事件B 能且只能在原有A_1，A_2，…，A_n 之一发生的条件下发生的结果，求事件B 的概率就是上面的全概率公式


B
Ω
A_2
A_3
A_1
A_5
A_4

全概率公式(例题分析)

【例】某车间用甲、乙、丙三台机床进行生产，各种机床的次品率分别为5%、4%、2%，它们各自的产品分别占总产量的25%、35%、40%，将它们的产品组合在一起，求任取一个是次品的概率。

解：设 A_1表示“产品来自甲台机床”， A_2表示“产品来自乙台机床”， A_3表示“产品来自丙台机床”， B表示“取到次品”。根据全概率公式有

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

$$=0.25\times0.05+0.35\times0.04+0.40\times0.02$$

$$=0.0345$$

贝叶斯公式(逆概公式)

1. 与全概公式解决的问题相反，贝叶斯公式是建立在条件概率基础上的寻找事件发生的原因
2. 设 n 个事件 A_1，A_2，…，A_n 两两互斥，$A_1+A_2+\ldots+A_n=\Omega$ (满足这两个条件的事件组称为一个完备事件组)，且$P(A_i)>0(i=1,2,\ldots,n)$，则

$$P(A_i \mid B)=\frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{n} p(A_j)P(B \mid A_j)}$$

贝叶斯公式 (例题分析)

【例】某车间用甲、乙、丙三台机床进行生产，各种机床的次品率分别为5%、4%、2%，它们各自的产品分别占总产量的25%、35%、40%，将它们的产品组合在一起，如果取到的一件产品是次品，分别求这一产品是甲、乙、丙生产的概率

解：设 A_1表示“产品来自甲台机床”， A_2表示“产品来自乙台机床”， A_3表示“产品来自丙台机床”， B表示“取到次品”。根据贝叶斯公式有：

$$P(A_1 \mid B) = \frac{0.25 \times 0.05}{0.0345} = 0.3623$$

$$P(A_2 \mid B) = \frac{0.35 \times 0.04}{0.0345} = 0.406$$

$$P(A_3 \mid B) = \frac{0.4 \times 0.02}{0.0345} = 0.232$$

随机变量

1. 一次试验的基本结果（样本点）的数值性描述，是试验结果的函数。
2. 一般用 X、Y、Z 来表示
3. 例如: 投掷两枚硬币出现正面的数量
4. 根据取值情况的不同分为离散型随机变量和连续型随机变量

离散型随机变量

1. 随机变量 X 取有限个值或所有取值都可以逐个列举出来 $X_1, X_2, \ldots$
2. 以确定的概率取这些不同的值
3. 离散型随机变量的一些例子

试验	随机变量	可能的取值
抽查100个产品	取到次品的个数	0,1,2, …,100
一家餐馆营业一天	顾客数	0,1,2, …
电脑公司一个月的销售	销售量	0,1, 2,…
销售一辆汽车	顾客性别	男性为0,女性为1

连续型随机变量

1. 随机变量 X 取无限个值
2. 所有可能取值不可以逐个列举出来，而是取数轴上某一区间内的任意点
3. 连续型随机变量的一些例子

试验	随机变量	可能的取值
抽查一批电子元件	使用寿命(小时)	$X \geq 0$
新建一座住宅楼	半年后工程完成的百分比	$0 \leq X \leq 100$
测量一个产品的长度	测量误差(cm)	$X \geq 0$

随机变量的概率分布

- 概率分布的含义及意义
  - 含义
    - 随机变量在其取值范围内，取值与取值概率之间一一对应的关系，称之为随机变量的概率分布，简称分布。
    - 注意：概率分布是就样本空间中所有样本点而言的
  - 意义
    - 描述随机变量变化的统计规律。
    - 方便地计算任一事件发生的概率。

三、离散型随机变量的概率分布

1. 列出离散型随机变量X的所有可能取值
2. 列出随机变量取这些值的概率
3. 通常用下面的表格来表示

$X=x_i$	x_1，x_2，…，x_n
$P(X=x_i)=p_i$	p_1，p_2，…，p_n

4. $P(X=x_i)=p_i$称为离散型随机变量的概率函数

- $p_i \geq 0$　　$\sum_{i=1}^{n} p_i = 1$

离散型随机变量的概率分布
(例题分析)

【例】如规定打靶中域Ⅰ得3分，中域Ⅱ得2分，中域Ⅲ得1分，中域外得0分。今某射手每100次射击，平均有30次中域Ⅰ，55次中域Ⅱ，10次中Ⅲ，5次中域外。则考察每次射击得分为0,1,2,3这一离散型随机变量，其概率分布为

$X=x_i$	0	1	2	3
$P(X=x_i)=p_i$	0.05	0.10	0.55	0.30

离散型随机变量的概率分布
(0—1分布)

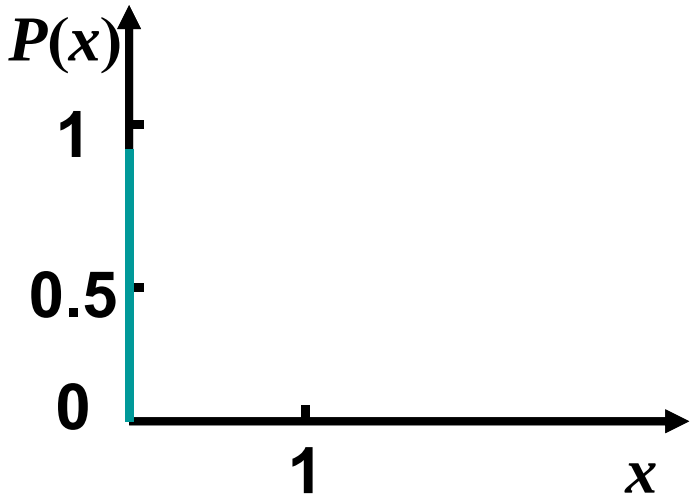
1. 一个离散型随机变量*X*只取两个可能的值
   - 例如，男性用 1表示，女性用0表示；
     合格品用 1 表示，不合格品用0表示
2. 列出随机变量取这两个值的概率

$$P(X=k)=p^{k}(1-p)^{1-k},k=0,1$$

(0—1分布)例题分析

【例】已知一批产品的不合格品率为$p=0.05$，合格率为$q=1-p=1-0.05=0.95$。并指定不合格品用1表示，合格品用0表示。则任取一件为不合格品或合格品这一离散型随机变量，其概率分布为

$X=x_i$	1	0
$P(X=x_i)=p_i$	0.05	0.95



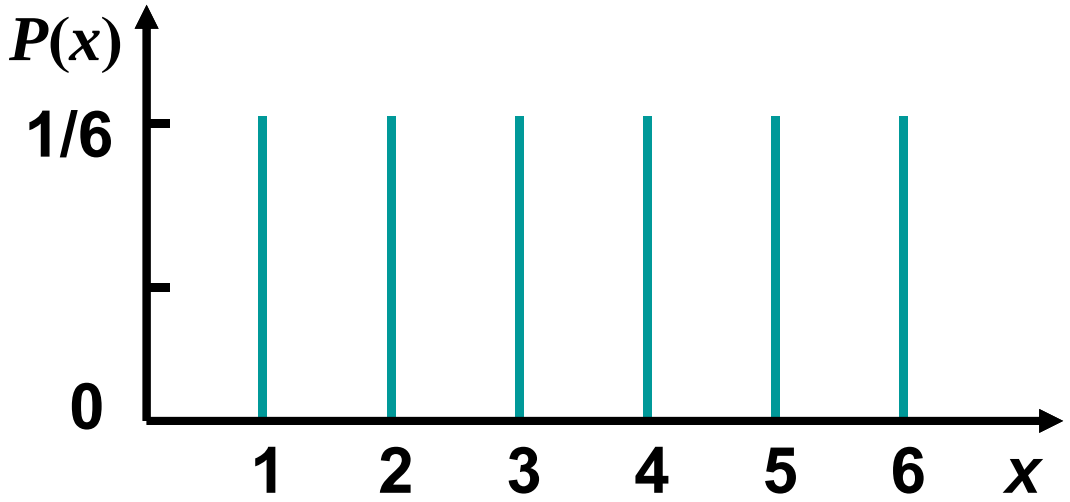
离散型随机变量的概率分布
(均匀分布)

1. 一个离散型随机变量取各个值的概率相同
2. 列出随机变量取值及其取值的概率
3. 例如，投掷一枚骰子，出现的点数及其出现各点的概率

(均匀分布)例题分析

【例】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

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



离散型随机变量的数学期望 (expected value)

1. 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
2. 描述离散型随机变量取值的集中程度
3. 计算公式为

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

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

离散型随机变量的方差 (variance)

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

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

若X是离散型随机变量，则

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

离散型随机变量的期望与方差
(例题分析)

【例】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：$D(X)=\sum_{i=1}^{6}[x_i-E(X)]^2\cdot p_i$

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

几种常见的离散型随机变量的概率分布

二项试验(贝努里试验)

1. 二项分布与贝努里试验有关
2. 贝努里试验具有如下属性
   - 试验包含了n个相同的试验
   - 每次试验只有两个可能的结果，即“成功”和“失败”
   - 出现“成功”的概率p对每次试验结果是相同的；“失败”的概率q也相同，且$p+q=1$
   - 试验是相互独立的
   - 试验“成功”或“失败”可以计数

二项分布(Binomial distribution)

1. 进行 n 次重复试验，出现“成功”的次数的概率分布称为二项分布
2. 设X为 n 次重复试验中事件A出现的次数，X 取 x 的概率为

$$P\{X=x\}=C_n^x p^x q^{n-x} \qquad (x=0,1,2,\cdots,n)$$

式中：$C_n^x=\dfrac{n!}{x!(n-x)!}$

3. 二项分布简记为：X~B(n,p)

4. 显然， 对于$P\{X=x\}\geq 0$， $x=1,2,\ldots,n$，有

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

5. 同样有

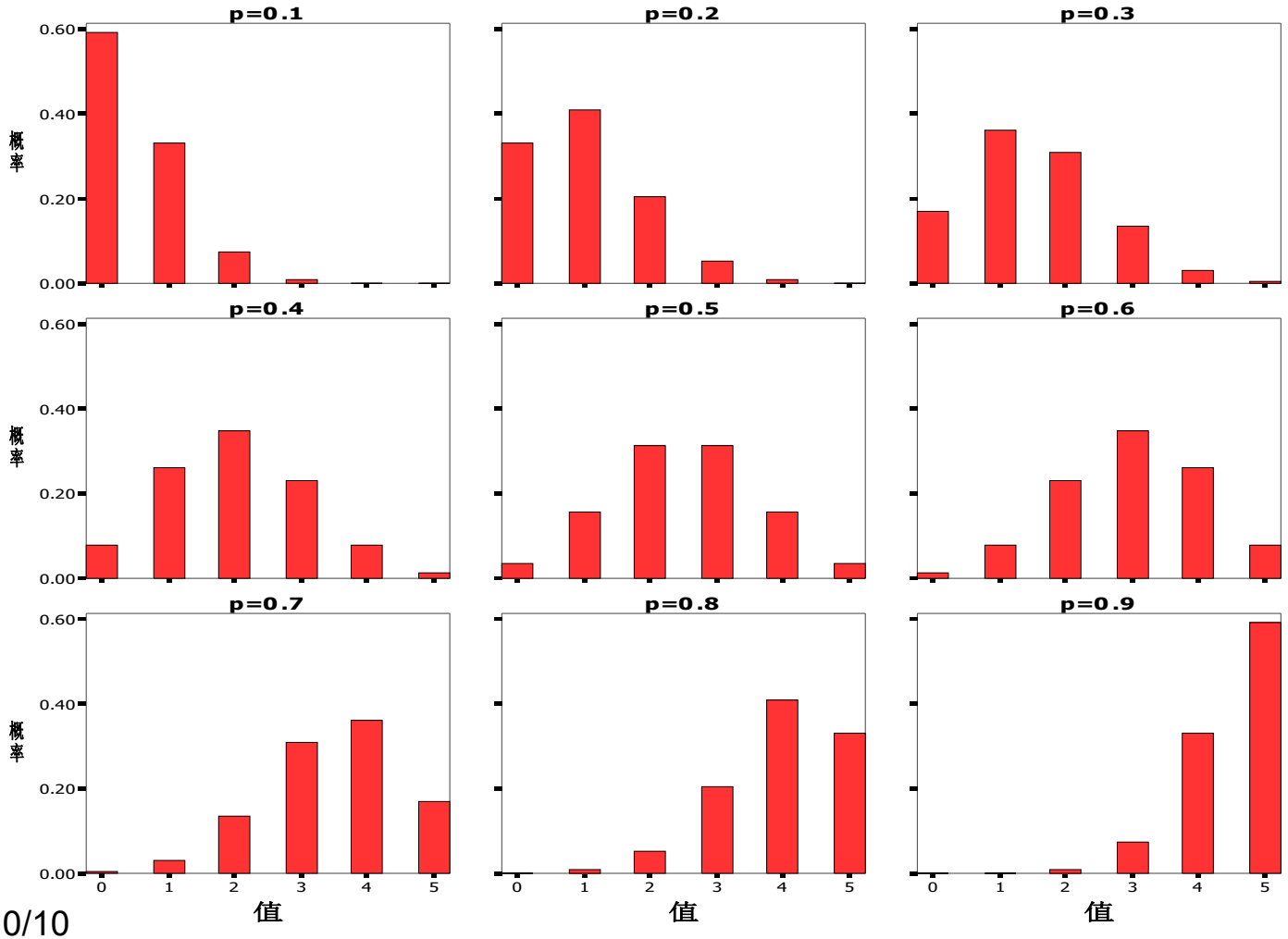
$$P\{0 \leq X \leq m\} = \sum_{x=0}^{m} C_n^x p^x q^{n-x}$$

$$P\{m \leq X \leq n\} = \sum_{x=m}^{n} C_n^x p^x q^{n-x}$$

6. 当 $n=1$ 时，二项分布化简为

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

九个二项分布B(5,p)
(p＝0.1到0.9)的概率分布图



二项分布的数学期望和方差

1. 二项分布的数学期望为

$$E(X) = np$$

2. 方差为

$$D(X) = npq$$

二项分布(例题分析)

【例】已知100件产品中有5件次品，现从中任取一件，有放回地抽取3次。求在所抽取的3件产品中恰好有2件次品的概率

解：设 X 为所抽取的3件产品中的次品数，则 $X \sim B(3, 0.05)$，根据二项分布公式有

$$P\{X=2\}=C_3^2(0.05)^2(0.95)^{3-2}=0.007125$$

泊松分布(Poisson distribution)

1. 用于描述在一指定时间范围内或在一定的长度、面积、体积之内每一事件出现次数的分布
2. 泊松分布的例子
   - 一个城市在一个月内发生的交通事故次数
   - 消费者协会一个星期内收到的消费者投诉次数
   - 人寿保险公司每天收到的死亡声明的人数

泊松分布的概率函数

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

λ— 给定的时间间隔、长度、面积、体积内“成功”的平均数

e = 2.71828

x —给定的时间间隔、长度、面积、体积内“成功”的次数

泊松分布简记为：$X\sim P(\lambda)$

泊松分布的期望和方差

1. 泊松分布的数学期望为

$$E(X)=\lambda$$

2. 方差为

$$D(X)=\lambda$$

泊松分布(例题分析)

【例】假定某企业的职工中在周一请假的人数X服从泊松分布，且设周一请事假的平均人数为2.5人。求

（1）X的均值及标准差

（2）在给定的某周一正好请事假是5人的概率

解：(1) $E(X)=\lambda=2.5$, $\sqrt{D(X)}=\sqrt{2.5}=1.581$

(2) $$P\{X=5\}=\frac{(2.5)^5 e^{-2.5}}{5!}=0.067$$

泊松分布
(作为二项分布的近似)

1. 当试验的次数 n 很大，成功的概率 p 很小时，可用泊松分布来近似地计算二项分布的概率，即

$$C_n^x p^x q^{n-x} \approx \frac{\lambda e^{-\lambda}}{x!}$$

2. 实际应用中，当 $P \le 0.25$，$n > 20$，$np \le 5$时，近似效果良好

超几何分布（hyper geometric distribution）

- 设有一批产品500个，其中次品有5个。假定该产品的质量检查采取随机抽取20个产品进行检查。如果抽到的20个产品中含有2个或更多不合格产品，则整个500个产品将会被退回。
- 这时，人们想知道，该批产品被退回的概率是多少？
- 这种概率就满足超几何分布。

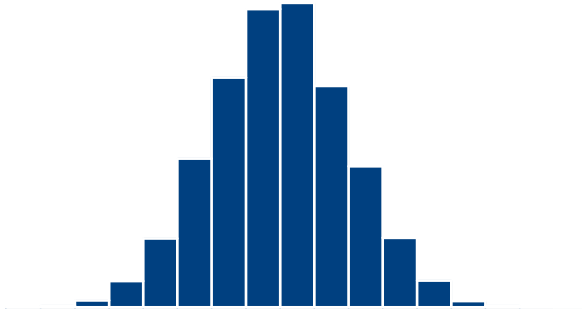
四、连续型随机变量的概率分布

1. 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
2. 因为它取任何一个特定的值的概率都等于0，所以不能列出每一个值及其相应的概率。因此，通常研究它取某一区间值的概率
3. 用密度函数和分布函数的形式来描述

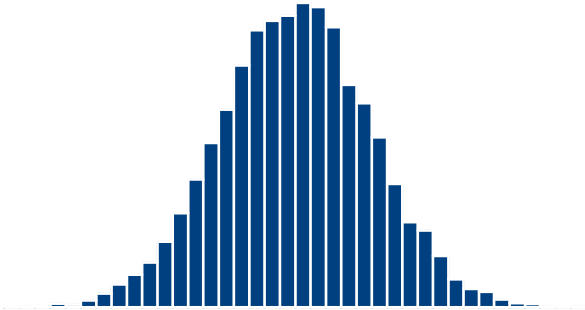
- 想象连续变量观测值的直方图；如果其纵坐标为相对频数，那么所有这些矩形条的高度和为1；完全可以重新设置量纲，使得这些矩形条的面积和为1。
- 不断增加观测值及直方图的矩形条的数目，直方图就会越来越像一条光滑曲线，其下面的面积和为1。
- 该曲线即所谓概率密度函数(probability density function，pdf)，简称密度函数或密度。下图为这样形成的密度曲线。

逐渐增加矩形条数目的直方图和一个形状类似的密度曲线。

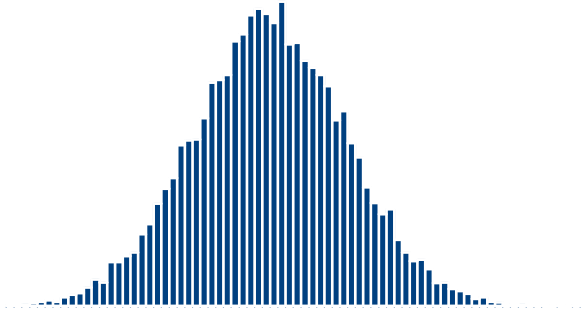
(1)



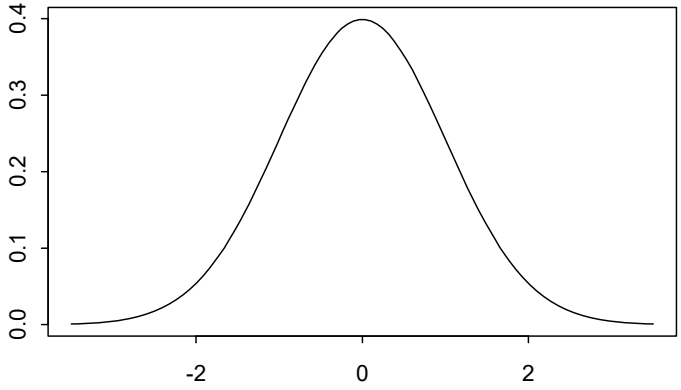
(2)



(3)



(4)



概率密度函数
(probability density function)

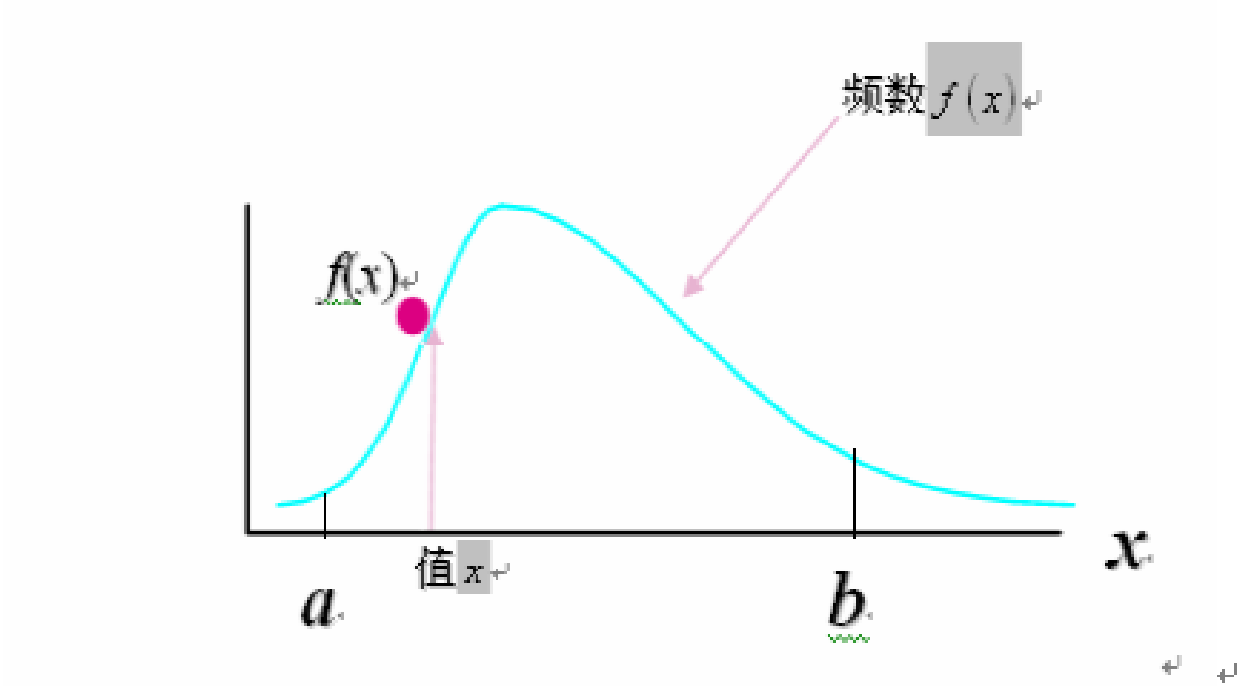
1. 设X为一连续型随机变量，x 为任意实数，X的概率密度函数记为$f(x)$，它满足条件

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

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

2. 注意：$f(x)$本身不是概率，是连续型随机变量取某一区间值的一种间接表述形式

概率密度函数

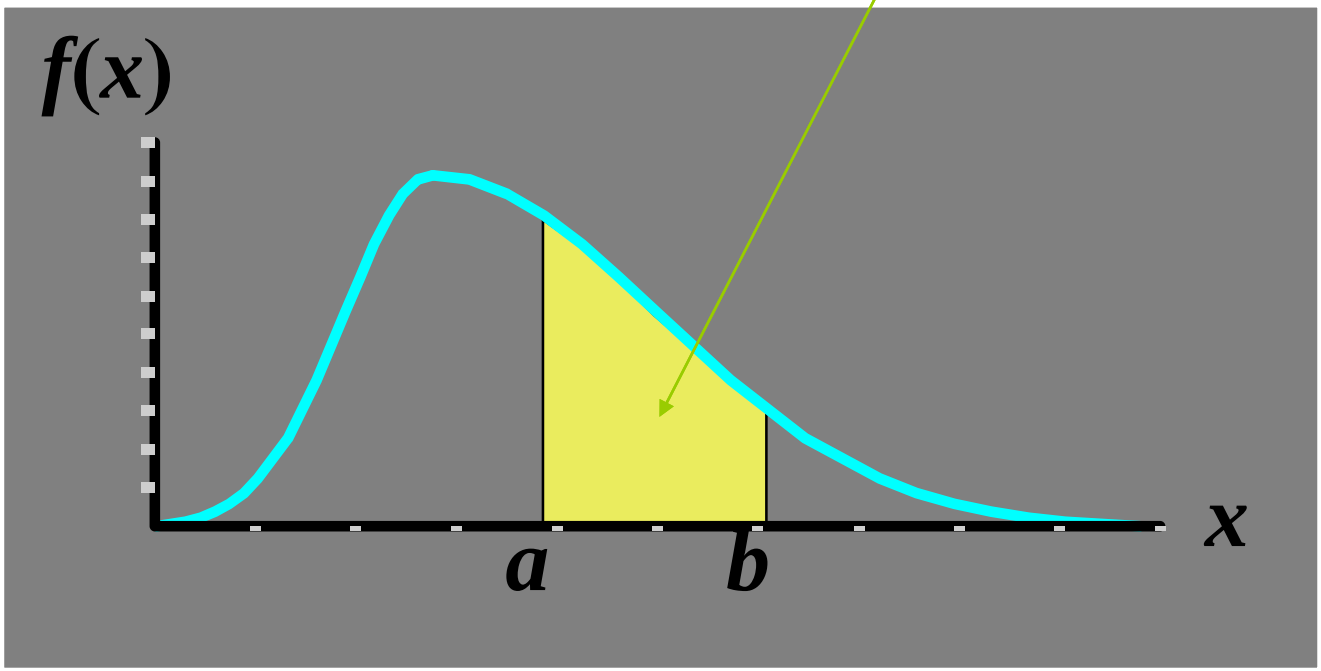


密度函数 $f(x)$ 的几何意义

连续型随机变量的概率

- ➡ 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)

- 分布函数的来源
  - 如前所述，离散型随机变量的分布用概率函数来描述，连续型随机变量的分布用密度函数来描述，两者形式不同，表现各异。为了更方便地表现随机变量的分布，下面引入分布函数。
- 分布函数定义为

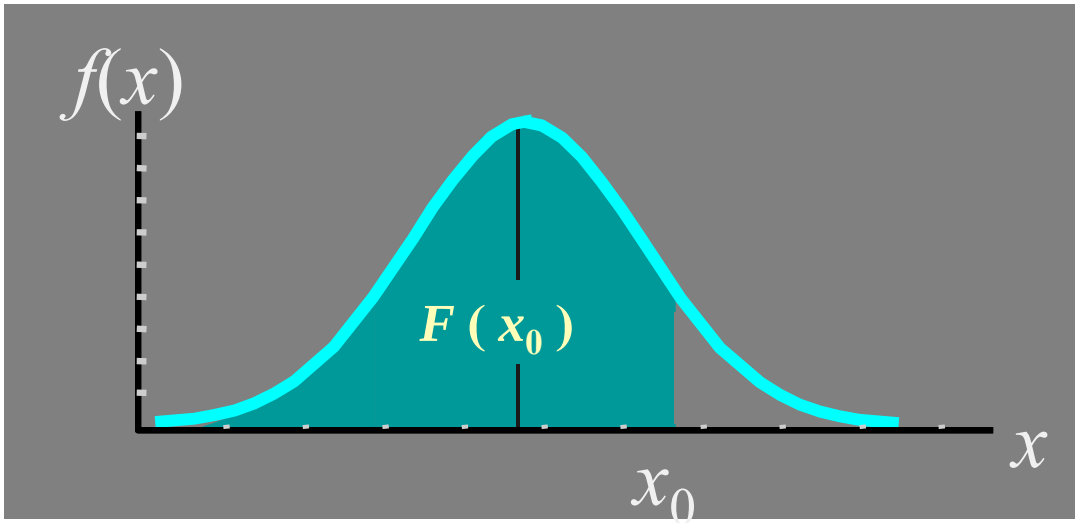
$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\mathrm{d}t \qquad (-\infty < x < +\infty)$$

- 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

1. 密度函数曲线下的面积等于1
2. 分布函数是曲线下小于 x_0 的面积



3. 分布函数本身是概率，即事件 “$X<x_0$”发生的概率

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

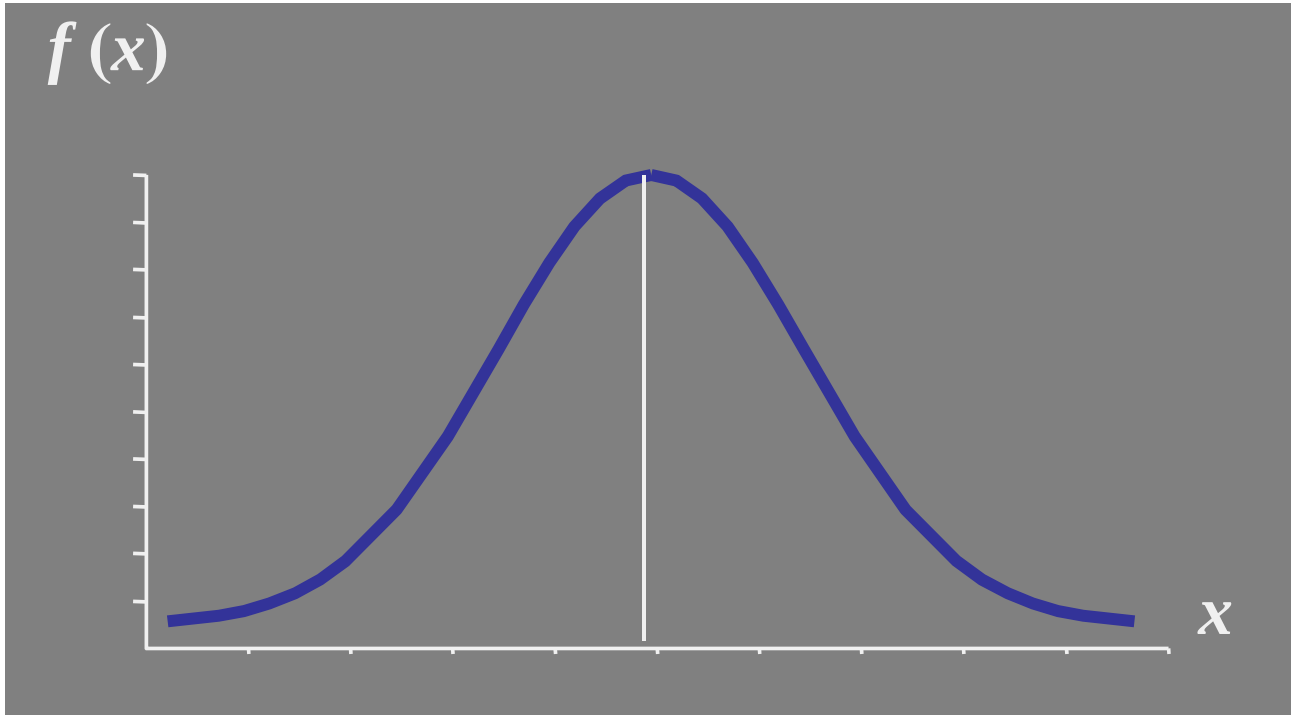
$$D(X)=\int_{-\infty}^{+\infty}\left[x-E(X)\right]^2 f(x)\mathrm{d}x=\sigma^2$$

几种常见的连续型随机变量的概率分布

- 均匀分布
- 指数分布
- 正态分布

正态分布(normal distribution)

- 描述连续型随机变量分布中最重要的分布



正态分布的重要意义

- 在随机理论中，正态分布是最重要的一种分布，理由如下：

  (1) 它是最常见的一种分布，现实中许多随机变量服从或近似服从正态分布。

  (2) 在一定的条件下，正态分布是其他分布的近似分布。

  (3) 许多有用的分布，特别是小样本的精确分布是由正态分布推导出来的。

  （4）是经典统计推断的基础。

正态分布的概率密度函数

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

$f(x)$ = 随机变量 X 的频数

σ^2 =随机变量 X 的方差

π =3.14159; e = 2.71828

x =随机变量 X 的取值 ($-\infty < x < +\infty$)

μ =随机变量 X 的均值

正态分布密度函数的性质（1）

1. 概率密度函数在x 轴的上方，即$f(x)>0$
2. 正态曲线的最高点在均值μ处，它也是分布的中位数和众数
3. 正态分布是一个分布族，每一特定正态分布通过均值μ和标准差σ来区分。 μ 决定了图形的中心位置，σ决定曲线的平缓程度，即宽度

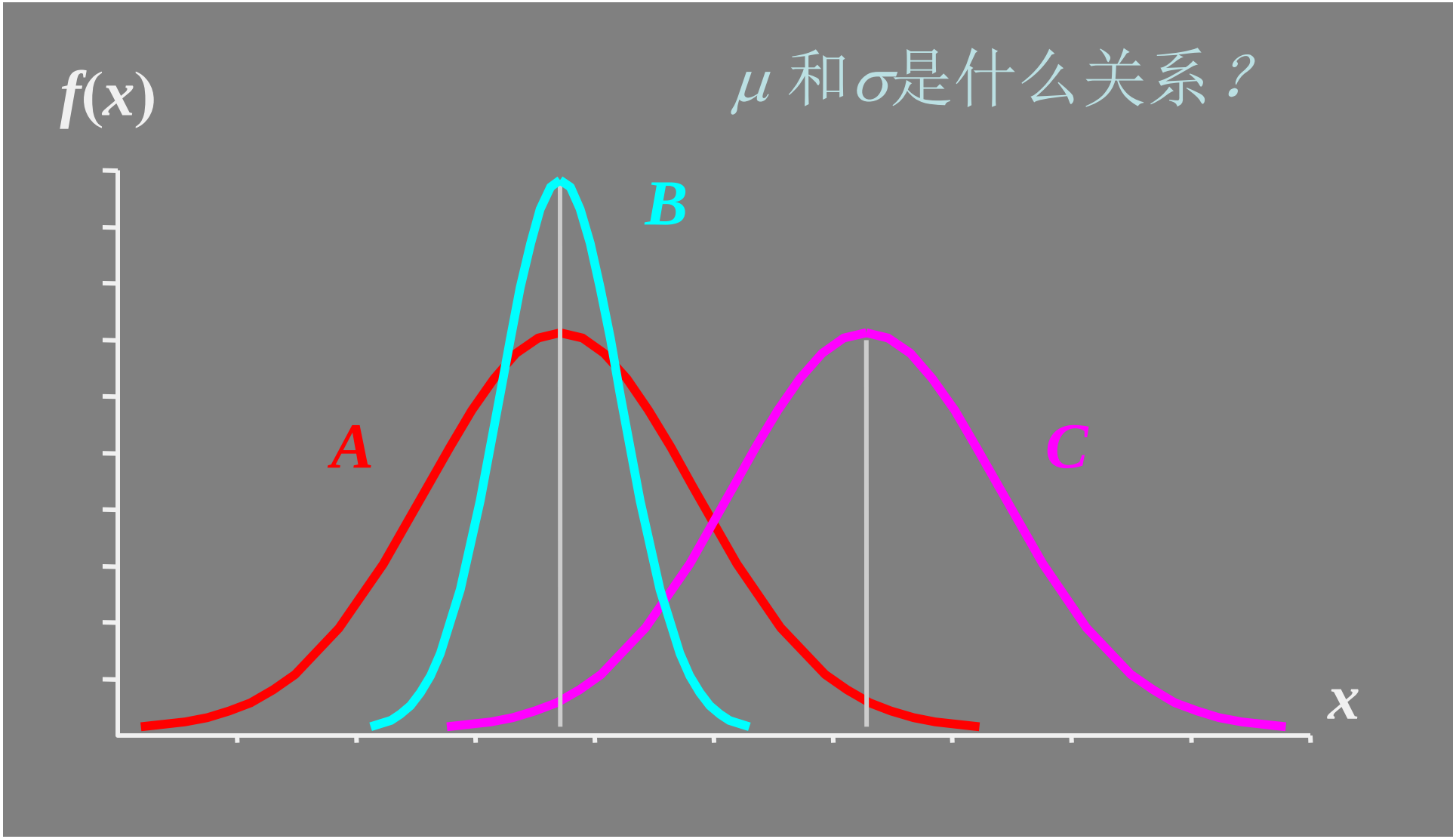
正态分布密度函数的性质（2）

4. 曲线$f(x)$相对于均值μ对称，尾端向两个方向无限延伸，且理论上永远不会与横轴相交

5. 正态曲线下的总面积等于1

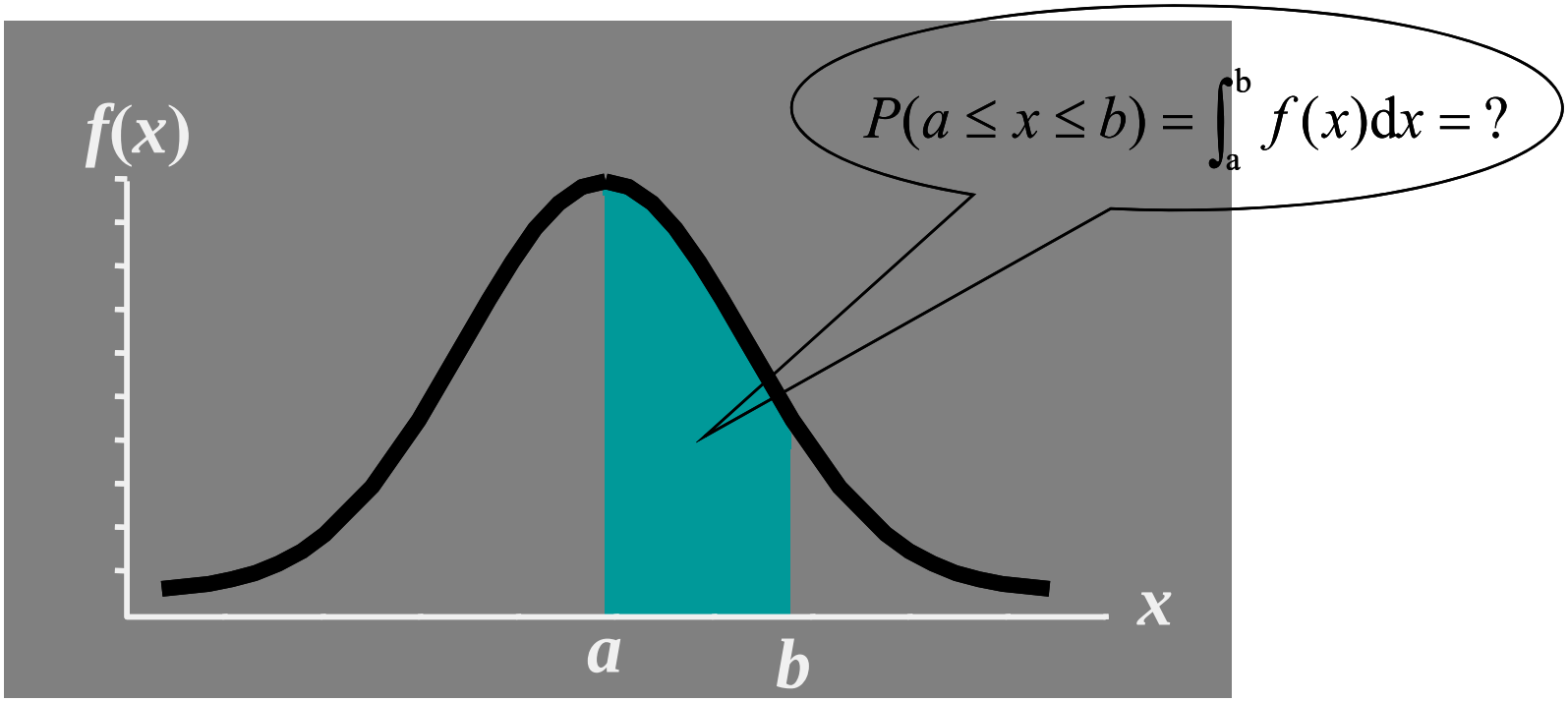
6. 随机变量的概率由曲线下的面积给出

μ 和 σ 对正态曲线的影响



正态分布的概率

概率是曲线下的面积!



标准正态分布
(standard normal distribution)

1. 一般的正态分布取决于均值μ和标准差 σ
2. 计算概率时 ，每一个正态分布都需要有自己的正态概率分布表，这种表格是无穷多的
3. 若能将一般的正态分布转化为标准正态分布，计算概率时只需要查一张表

标准正态分布的密度函数

1. 任何一个一般的正态分布，均可通过下面的线性变换转化为标准正态分布

$$Z=\frac{X-\mu}{\sigma}\sim N(0,1)$$

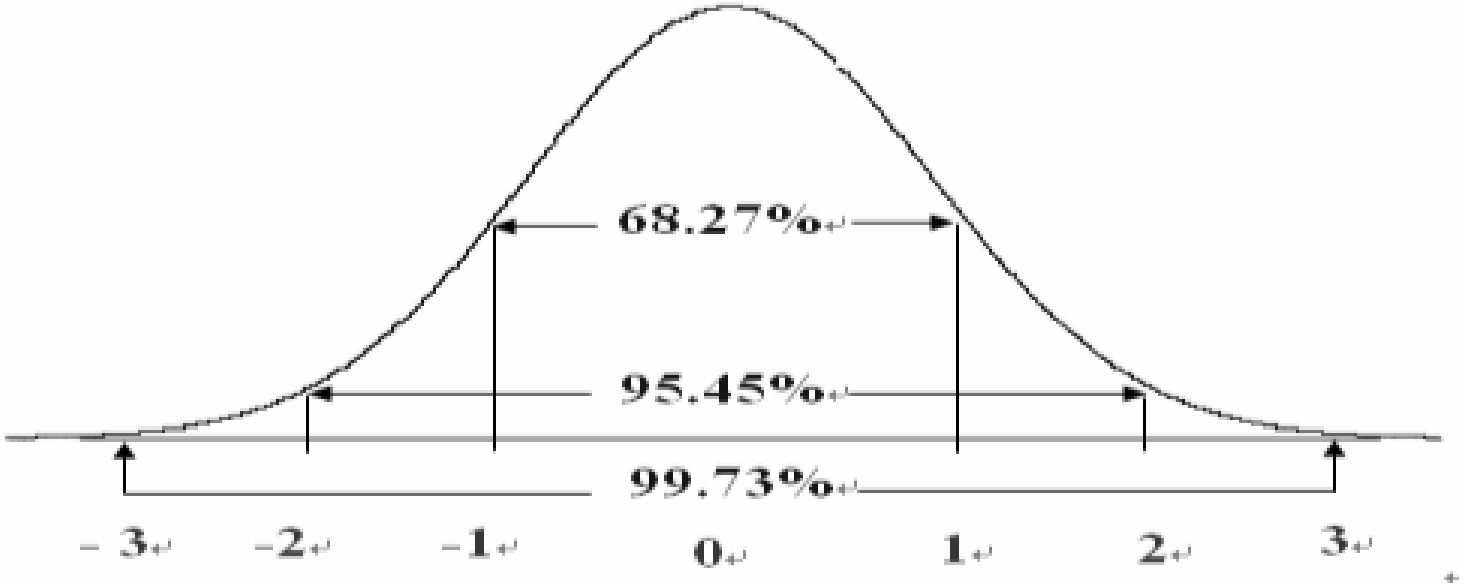
2. 标准正态分布的概率密度函数

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\quad -\infty<x<+\infty$$

3. 标准正态分布的分布函数

$$\Phi(x)=\int_{-\infty}^{x}\varphi(t)\mathrm{d}t=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$

3σ准则



标准正态分布表的使用

1. 将一个一般的转换为标准正态分布
2. 计算概率时，查标准正态概率分布表
3. 对于负的 x，可由 $\Phi(-x)=1-\Phi(x)$ 得到
4. 对于标准正态分布，即 $X\sim N(0,1)$，有
   - $P(a\le X\le b)=\Phi(b)-\Phi(a)$
   - $P(|X|\le a)=2\Phi(a)-1$
5. 对于一般正态分布，即 $X\sim N(\mu,\sigma)$，有

$$P(a\le X\le b)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

正态分布(例题分析)

【例】设$X\sim N(0，1)$，求以下概率：

(1) $P(X<1.5)$ ； (2) $P(X>2)$； (3) $P(-1<X\leq 3)$ ； (4) $P(|X|\leq 2)$

解：(1) $P(X<1.5)=\Phi(1.5)=0.9332$

(2) $P(X>2)=1-P(X\leq 2)=1-0.9973=0.0227$

(3) $P(-1<X\leq 3)=P(X\leq 3)-P(X<-1)$

$=\Phi(3)-\Phi(-1)=\Phi(3)-[1-\Phi(1)]$

$=0.9987-(1-0.8413)=0.84$

(4) $P(|X|\leq 2)=P(-2\leq X\leq 2)=\Phi(2)-\Phi(-2)$

$=\Phi(2)-[1-\Phi(2)]=2\Phi(2)-1=0.9545$

正态分布(例题分析)

【例】设$X\sim N(5,\ 3^2)$，求以下概率

(1) $P(X\le 10)$ ； (2) $P(2<X<10)$

解：(1) $P(X\le 10)=P\left(\dfrac{X-5}{3}\le\dfrac{10-5}{3}\right)$

$$=P\left(\frac{X-5}{3}\le 1.67\right)=\Phi(1.67)=0.9525$$

(2) $P(2<X<10)=P\left(\dfrac{2-5}{3}<\dfrac{X-5}{3}<\dfrac{10-5}{3}\right)$

$$=P\left(-1<\frac{X-5}{3}<1.67\right)$$

$$=\Phi(1.67)-\Phi(-1)=0.7938$$

二项分布的正态近似

1. 当n 很大时，二项随机变量X近似服从正态分布$N\{np\ ,\ np(1\text{-}p)\}$
2. 对于一个二项随机变量X，当n很大时，求$P(x_1 \le X \le x_2)$时可用正态分布近似为

$$P\{x_1 \le X \le x_2\} = \sum_{x=x_1}^{x_2} C_n^x p^x q^{n-x} = \int_a^b \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t$$

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

式中：$a = \dfrac{x_1 - np}{\sqrt{npq}}$ ， $b = \dfrac{x_2 - np}{\sqrt{npq}}$ ， $q = 1 - p$

二项分布的正态近似
（实例）

【例】100台机床彼此独立地工作，每台机床的实际工作时间占全部工作时间的80%。求

(1)任一时刻有70～86台机床在工作的概率

(2)任一时刻有80台以上机床在工作的概率

解：设X表示100台机床中工作着的机床数，则$X\sim B(100,0.8)$。现用正态分布近似计算，$np=80$，$npq=16$

(1) $$P(70\le X\le 86)=P\left(\frac{70-80}{4}\le\frac{X-80}{4}\le\frac{86-80}{4}\right)$$

$$=\Phi(1.5)-\Phi(-2.5)=0.927$$

(2) $$P(X\ge 80)=P\left(\frac{X-80}{4}\ge 0\right)=1-\Phi(0)=0.5$$

本章小结

1. 随机事件及其概率
2. 概率的性质与运算法则
3. 离散型随机变量的分布
4. 连续型随机变量的分布

1. 样本的定义

设 X 是具有分布函数 F 的随机变量, 若 X_1, $X_2, \cdots, X_n$ 是具有同一分布函数 F、相互独立的随机变量, 则称 $X_1, X_2, \cdots, X_n$ 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, 简称样本.

当n次观察一完成, 我们就得到一组实数 $x_1, x_2, \cdots, x_n$, 它们依次是随机变量 $X_1, X_2, \cdots, X_n$ 的观察值, 称为样本值.

注意:样本是一组独立且和总体同分布的随机变量.

简单随机样本满足以下三个条件:

(1) 随机性: 抽样应随机地进行, 每个个体被抽到的机会均等;

(2) 独立性: 每次抽样应独立进行, 其结果相互不受影响.

(3) 同分布: 样本与总体服从同一分布.

2. 样本概率分布

通过样本来研究总体的概率分布问题.

1. 若 $X_1, X_2, \cdots, X_n$ 为F的一个样本, 则 $X_1, X_2, \cdots, X_n$ 相互独立, 它们的分布函数都是F,

所以 $X_1, X_2, \cdots, X_n$ 的分布函数为

$$F(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)$$

2. 若X是离散型总体，且具有分布率

$P(x)$，则$X_1, X_2, \cdots, X_n$ 的联合分布率为

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_{i=1}^{n} P(x_i)$$

3. 若X是连续型总体，且具有概率密度

$f(x)$，则 $\boldsymbol{X_1, X_2, \cdots, X_n}$ 的概率密度为

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)$$

例1 设总体 X 服从参数为 $\lambda\,(\lambda>0)$ 的指数分布，$(X_1,X_2,\cdots,X_n)$ 是来自总体的样本，求样本 $(X_1,X_2,\cdots,X_n)$ 的概率密度.

解 总体 X 的概率密度为 $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0,\\ 0, & x\le 0,\end{cases}$

因为 $X_1,X_2,\cdots,X_n$ 相互独立，且与 X 有相同的分布，所以 $(X_1,X_2,\cdots,X_n)$ 的概率密度为

$$f_n(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i)=\begin{cases}\lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}, & x_i>0,\\ 0, & \text{其他}.\end{cases}$$

例2 设总体 X 服从两点分布 $B(1, p)$, 其中 $0 < p < 1$, $(X_1, X_2, \cdots, X_n)$ 是来自总体的样本, 求样本 $(X_1, X_2, \cdots, X_n)$ 的分布律.

解 总体 X 的分布律为

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

因为 $X_1, X_2, \cdots, X_n$ 相互独立,

且与 X 有相同的分布,

所以 $(X_1, X_2, \cdots, X_n)$ 的分布律为

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$$

$$= P\{X_1 = x_1\}P\{X_2 = x_2\} \cdots P\{X_n = x_n\}$$

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}$$

其中 $x_1, x_2, \cdots, x_n$ 在集合 $\{0,1\}$ 中取值.

一. 统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$g(X_1, X_2, \cdots, X_n)$ 是 $X_1, X_2, \cdots, X_n$ 的函数，若 g 中不含未知参数，则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量.

设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值，则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值.

显然，统计量是一个随机变量.
它不含任何未知参数.

例1　设 X_1, X_2, X_3 是来自总体 $N(\mu, \sigma^2)$ 的一个样本, 其中 μ 为已知, σ^2 为未知, 判断下列各式哪些是统计量, 哪些不是?

$T_1 = X_1,$　　$T_2 = X_1 + X_2 e^{X_3},$

$T_3 = \frac{1}{3}(X_1 + X_2 + X_3),$

$T_4 = \max(X_1, X_2, X_3),$　　$T_5 = X_1 + X_2 - 2\mu,$

是

$T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2).$

不是

二. 几个常用统计量

设 $X_1, X_2, \cdots, X_n$ 是来自总体的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观察值.

(1)样本均值 $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

其观察值 $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

(2)样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

修正样本方差

其观察值

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

(3)样本标准差

$$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2};$$

其观察值 $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

(4) 样本 k 阶(原点)矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \ , k = 1, 2, \cdots;$

其观察值 $\alpha_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \ , k = 1, 2, \cdots .$

(5)样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \ , k = 2, 3, \cdots;$$

未修正样本方差

其观察值 $b_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^k \ , k = 2, 3, \cdots .$

注:1阶原点矩$A1$ 就是样本均值 $\overline{X}$,

2阶中心矩$B2$ 不是样本方差 . S^2

三. 经验分布函数

定义： 设 $X_1, X_2, \cdots, X_n$ 是取自总体 X 的样本，它的顺序统计量为 $X_1^*, X_2^*, \cdots, X_n^*$。对于任意实数 x，令

$$F_n(x) = \begin{cases} 0, & \text{若} x < X_1^*, \\ \dfrac{k}{n}, & \text{若} X_k^* \le x < X_{k+1}^*, \\ 1, & \text{若} X_n^* \le x. \end{cases}$$

则称 $F_n(x)$ 为总体 X 的经验分布函数。

由定义知，对每一给定的 x，经验分布函数是样本 $X_1, X_2, \cdots, X_n$ 的函数，所以它是一个统计量，若 $x_1, x_2, \cdots, x_n$ 是样本观察值，称

$$F_n(x)=\begin{cases}0, & \text{若} x < x_1^*, \\ \dfrac{k}{n}, & \text{若} x_k^* \le x < x_{k+1}^*, \\ 1, & \text{若} x_n^* \le x. \end{cases}$$

为经验分布函数的观察值。

实例　设总体 F 具有一个样本值 $1, 2, 3,$

则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{1}{3}, & 1\le x<2,\\ \dfrac{2}{3}, & 2\le x<3,\\ 1, & x\ge 3.\end{cases}$$

实例　设总体 F 具有一个样本值 $1,1,2,$

则经验分布函数 $F_3(x)$ 的观察值为

$$F_3(x)=\begin{cases}0, & x<1,\\ \dfrac{2}{3}, & 1\le x<2,\\ 1, & x\ge 2.\end{cases}$$

格里汶科定理

格里汶科

对于任一实数 x,当 $n \to \infty$ 时, $F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$, 即

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<+\infty} |F_n(x)-F(x)| = 0\right\} = 1.$$

对于任一实数 x 当 n 充分大时, 经验分布函数的任一个观察值 $F_n(x)$ 与总体分布函数 $F(x)$ 只有微小的差别, 从而在实际上可当作 $F(x)$ 来使用.

一、χ^2分布

χ^2分布是由正态分布派生出来的一种分布.

定义: 设 $X_1, X_2, \cdots, X_n$ 相互独立, 都服从正态分布 $N(0,1)$, 则称随机变量:

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

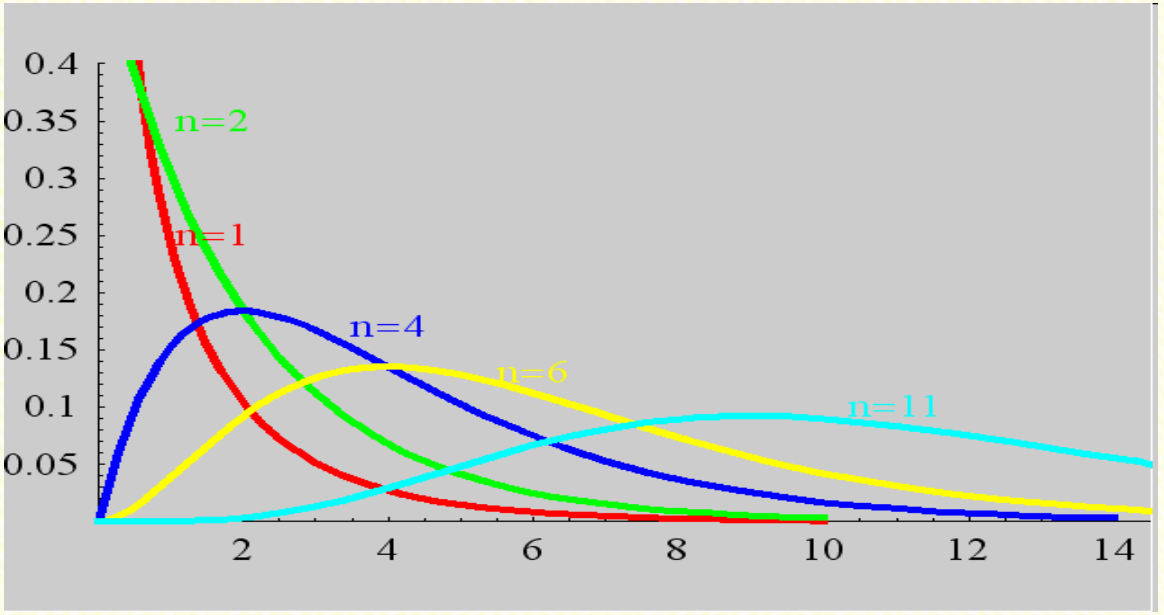
所服从的分布为自由度为 n 的 χ^2分布.

记为 $\chi^2 \sim \chi^2(n)$

$\chi^2(n)$分布的概率密度为

$$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}, & y>0\\ 0, & \text{其他.}\end{cases}$$

$\chi^2(n)$分布的概率密度曲线如图.



$\chi^2(n)$分布的概率密度为

$$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}, & y>0\\ 0 & \text{其他.}\end{cases}$$

证明　因为$\chi^2(1)$分布即为$\Gamma\left(\frac{1}{2},2\right)$分布,

又因为$X_i\sim N(0,1)$,　由定义$X_i^2\sim\chi^2(1)$,

即$X_i^2\sim\Gamma\left(\frac{1}{2},2\right)$,　$i=1,2,\cdots,n.$

因为 $X_1, X_2, \cdots, X_n$ 相互独立,

所以 $X_1^2, X_2^2, \cdots, X_n^2$也相互独立,

根据Γ分布的可加性知 $\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \Gamma\left(\frac{n}{2}, 2\right).$

$\chi^2(n)$分布的概率密度曲线如图.

χ^2 分布的性质

性质1 (χ^2 分布的可加性)

设 $\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$, 并且 χ_1^2, χ_2^2 独立, 则 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$.

(此性质可以推广到多个随机变量的情形.)

设 $\chi_i^2 \sim \chi^2(n_i)$, 并且 χ_i^2 $(i = 1, 2, \cdots, m)$ 相互独立, 则 $\sum_{i=1}^{m} \chi_i^2 \sim \chi^2(n_1 + n_2 + \cdots + n_m)$.

性质2 （χ^2分布的数学期望和方差）

若 $\chi^2 \sim \chi^2(n)$, 则 $E(\chi^2)=n, \quad D(\chi^2)=2n.$

事实上，由 $X_i \sim N(0,1)$, 故 $E(X_i^2)=D(X_i)=1$

$$D(X_i^2)=E(X_i^4)-[E(X_i^2)]^2=3-1=2$$

$$E(\chi^2)=\sum_{i=1}^{n}E(X_i^2)=n, D(\chi^2)=\sum_{i=1}^{n}D(X_i^2)=2n.$$

性质3

设 $X_1, X_2, \cdots, X_n$ 相互独立, 都服从正态分布

$N(\mu, \sigma^2)$, 则 $\chi^2 = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n)$

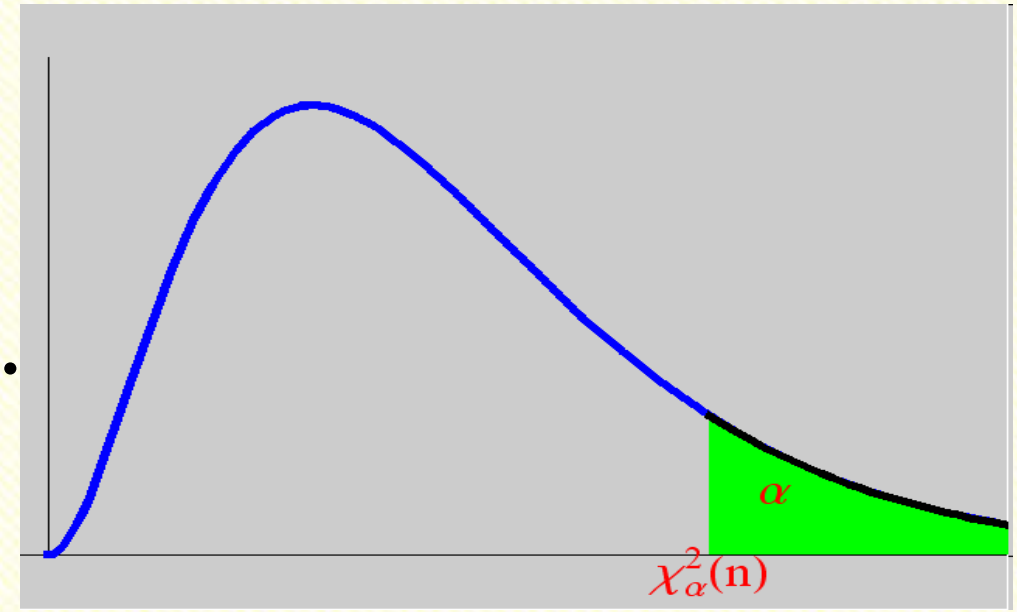
χ^2 分布的分位点

对于给定的正数 α, $0<\alpha<1$, 称满足条件

$$P\{\chi^2 > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{\infty} f(y)\mathrm{d}y = \alpha$$

的点 $\chi_\alpha^2(n)$ 为 $\chi^2(n)$ 分布的上 α 分位点.

对于不同的 α, n,
可以通过查表求
得上 α 分位点的值.



例1　设 X 服从标准正态分布 $N(0,1)$, $N(0,1)$ 的上 α 分位点 z_α 满足 $P\{X > z_\alpha\} = \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{z_\alpha}^{+\infty} e^{-\frac{x^2}{2}}\,dx = \alpha,$

求 z_α 的值, 可通过查表完成.

$z_{0.05} = 1.645,$　附表2-1

$z_{0.025} = 1.96,$　附表2-2

根据正态分布的对称性知

性质　$z_{1-\alpha} = -z_\alpha.$

附表2-1 标准正态分布表

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0			0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0			0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0			0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545

1.645

例2　设 $Z \sim \chi^2(n)$, $\chi^2(n)$ 的上 α 分位点满足

$$P\{Z > \chi_\alpha^2(n)\} = \int_{\chi_\alpha^2(n)}^{+\infty} \chi^2(y;n)\mathrm{d}y = \alpha,$$

求 $\chi_\alpha^2(n)$ 的值, 可通过查表完成 .

$\chi_{0.025}^2(8) = 17.535,$ 附表4-1

$\chi_{0.975}^2(10) = 3.247,$

$\chi_{0.1}^2(25) = 34.382.$

附表4只详列到 n=40 为止.

附表4-1　　χ^2分布表

n	α=0.25	0.10	0.05	0.025	0.01	0.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.7		11.143	13.277	14.860
5	6.626	9.2		12.833	15.086	16.750
6	7.841	10.6		14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.757
12	14.845	18.549	21.026	23.337	26.217	28.299
13	15.984	19.812	22.362	24.736	27.688	29.891
14	17.117	20.064	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801
16	19.369	23.542	26.296	28.845	32.000	34.267

17.535

费舍尔(R.A.Fisher)证明:

$$\text{当 } n \text{ 充分大时, } \chi_\alpha^2(n) \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2 .$$

其中 z_α 是标准正态分布的上 α 分位点.

利用上面公式,

可以求得 $n > 40$, 上 α 分位点值。

例如 $\chi_{0.05}^2(50) \approx \frac{1}{2}(1.645 + \sqrt{99})^2 = 67.221.$

而查更详细的表可得 $\chi_{0.05}^2(50) = 67.505$.

二. t 分布

设 $X \sim N(0,1)$, $Y \sim \chi^2(n)$, 且 X, Y 独立,

则称随机变量 $t = \dfrac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布, 记为 $t \sim t(n)$.

t 分布又称学生氏(Student)分布.

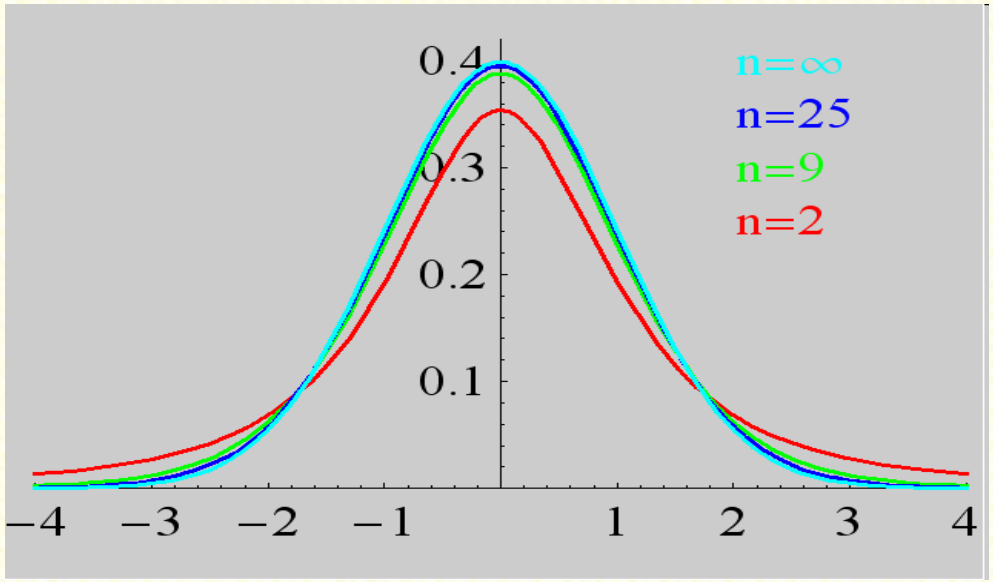
$t(n)$ 分布的概率密度函数为

$$h(t) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\dfrac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < +\infty$$

t 分布的概率密度曲线如图

显然图形是关于

$t=0$对称的.



当 n 充分大时, 其图形类似于标准正态变量概率密度的图形.

因为 $\lim\limits_{n\to\infty} h(t)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$,

所以当 n 足够大时 t 分布近似于 $N(0,1)$ 分布,

但对于较小的 n, t分布与$N(0,1)$分布相差很大.

性质 $E(t(n))=0, D(t(n))=n/(n-2), n>2$

t 分布的分位点

对于给定的 α, $0<\alpha<1$, 称满足条件

$$P\{t>t_\alpha(n)\}=\int_{t_\alpha(n)}^{\infty}h(t)\mathrm{d}t=\alpha$$

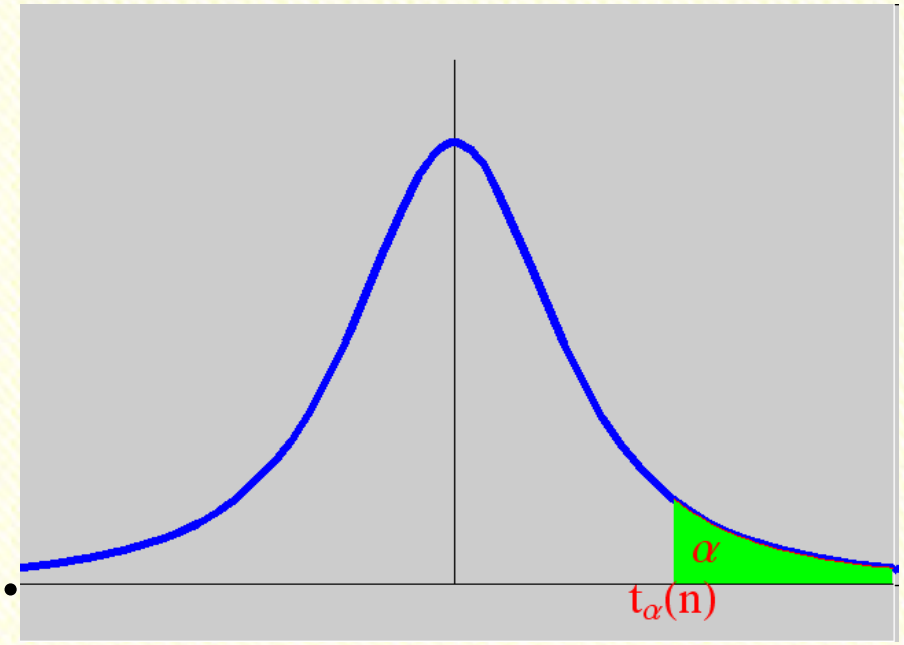
的点 $t_\alpha(n)$ 为 $t(n)$ 分布的上 α 分位点.

可以通过查表求得上α分位点的值.

由分布的对称性知

性质 $t_{1-\alpha}(n)=-t_\alpha(n)$.

当 $n>45$ 时, $t_\alpha(n)\approx z_\alpha$.



例3　设 $T \sim t(n)$, $t(n)$ 的上 α 分位点满足

$$P\{T > t_\alpha(n)\} = \int_{t_\alpha(n)}^{+\infty} t(y;n)\mathrm{d}y = \alpha,$$

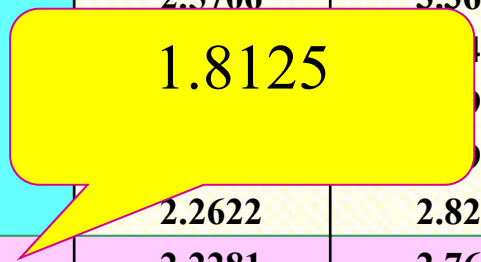
求 $t_\alpha(n)$ 的值, 可通过查表完成.

$t_{0.05}(10) = 1.8125,$ 附表3-1

$t_{0.025}(15) = 2.1315.$

附表3-1　　t 分布表

n	α =0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
6	0.7176	1.4398	1.9432		427	3.7074
7	0.7111	1.4149	1.8946		980	3.4995
8	0.7064	1.3968	1.8595		065	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208



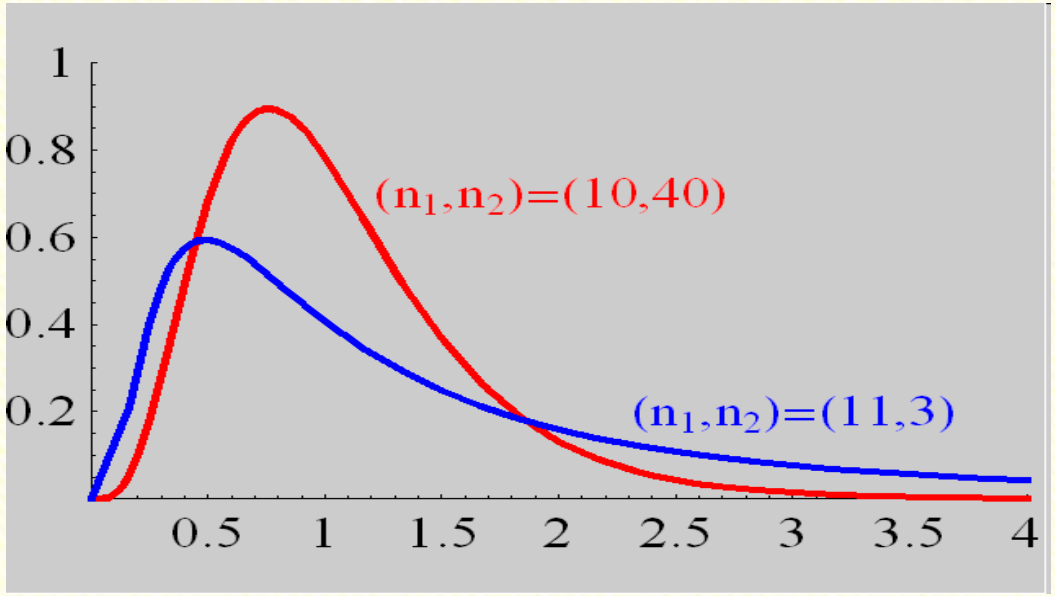
三. F分布

设$U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U, V 独立, 则称随机变量 $F = \dfrac{U/n_1}{V/n_2}$服从自由度为(n_1, n_2)的 F 分布, 记为 $F \sim F(n_1, n_2)$.

$F(n_1, n_2)$分布的概率密度为

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

F分布的概率密度曲线如图



根据定义可知,

性质 若$F \sim F(n_1, n_2)$,

则$\dfrac{1}{F} \sim F(n_2, n_1)$.

F 分布的分位点

对于给定的α, $0<\alpha<1$, 称满足条件

$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1, n_2)}^{+\infty} \psi(y)\mathrm{d}y = \alpha$$

的点$F_\alpha(n_1, n_2)$为$F(n_1, n_2)$分布的上α分位点.

例4　设 $F(n_1, n_2)$分布的上α分位点满足

$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1, n_2)}^{+\infty} \psi(y)\mathrm{d}y = \alpha,$$

求 $F_\alpha(n_1, n_2)$ 的值, 可通过查表完成 .

$F_{0.025}(7,8) = 4.90,$

$F_{0.05}(14,30) = 2.31$.

F 分布的上 α 分位点具有如下性质：

性质　$F_{1-\alpha}(n_1,n_2)=\dfrac{1}{F_{\alpha}(n_2,n_1)}.$

证明　因为 $F\sim F(n_1,n_2),$

所以 $1-\alpha=P\{F>F_{1-\alpha}(n_1,n_2)\}$

$$=P\left\{\frac{1}{F}<\frac{1}{F_{1-\alpha}(n_1,n_2)}\right\}=1-P\left\{\frac{1}{F}\geq\frac{1}{F_{1-\alpha}(n_1,n_2)}\right\}$$

$$=1-P\left\{\frac{1}{F}>\frac{1}{F_{1-\alpha}(n_1,n_2)}\right\},$$

故 $P\left\{\dfrac{1}{F}>\dfrac{1}{F_{1-\alpha}(n_1,n_2)}\right\}=\alpha,$

因为 $\frac{1}{F} \sim F(n_2, n_1)$, 所以 $P\left\{\frac{1}{F} > F_\alpha(n_2, n_1)\right\} = \alpha,$

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

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

用来求分布表中未列出的一些上 α 分位点.

例 $F_{0.95}(12,9) = \frac{1}{F_{0.05}(9,12)} = \frac{1}{0.28} = 0.357$.

四. 正态总体的样本均值与样本方差的分布

定理一

设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}$ 是样本均值 , 则有 $\overline{X} \sim N(\mu, \sigma^2/n)$.

正态总体 $N(\mu,\sigma^2)$ 的样本均值和样本方差有以下两个重要定理.

定理二(证明略)

设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本，$\overline{X}, S^2$ 分别是样本均值和样本方差，则有

(1) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$;

(2) $\overline{X}$ 与 S^2 独立.

(3) $\chi^2 = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n)$

(1)' $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 \sim \chi^2(n-1)$

定理三 设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}, S^2$ 分别是样本均值和样本方差, 则有

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1).$$

证明 因为 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),\ \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$

且两者独立, 由 t 分布的定义知

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1).$$

定理四　设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是具有相同方差的两正态总体 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ 的样本, 且这两个样本互相独立, 设 $\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ 分别是这两个样本的均值,

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

分别是这两个样本的方差, 则有

(1) $\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1);$

(2) 当$\sigma_1^2 = \sigma_2^2 = \sigma^2$时,

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

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

证明 (1) 由定理二

$$\frac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1),\quad \frac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1),$$

由假设 S_1^2, S_2^2 独立, 则由 F 分布的定义知

$$\frac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2}\Big/\frac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2}\sim F(n_1-1,n_2-1),$$

即 $$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2}\sim F(n_1-1,n_2-1).$$

(2) 因为 $\overline{X}-\overline{Y}\sim N\left(\mu_1-\mu_2,\dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2}\right)$

所以 $U=\dfrac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim N(0,1),$

由 $\dfrac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1),\quad\dfrac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1),$

且它们相互独立, 故由 χ^2 分布的可加性知

$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2),$$

由于 U 与 V 相互独立，按 t 分布的定义.

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}}$$

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

第二节 大数定律与中心极限定理

一、大数定律

1.切比雪夫大数定律

2.贝努里大数定律

二、中心极限定理

1.林德贝格-勒维中心极限定理

2.棣莫弗-拉普拉斯中心极限定理

一、大数定律

大数定律又称作大数法则，是关于“均值具有稳定性”的一类定理。个别事物因偶然因素的影响而产生变异，有各自不同的表现，但是，对总体进行大量观察后平均，就能使偶然因素的影响相互抵消，消除由个别偶然因素引起的极端性影响，从而使总体均值稳定下来，反映出事物变化的一般规律。

（一）切比雪夫大数定律

设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，且具有相同的有限期望和方差：$E(X_i)=\mu$，$D(X_i)=\sigma^2, (i=1,2,\cdots,n)$则对于任意正数$\varepsilon$。都有

$$\lim_{x\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right\} = 1$$

该定律说明，当n充分大时。独立同分布的一系列随机变量的均值与它们共同的期望值之间的偏差，可以很大的把握被控制在任意给定的范围之内。

（一）切比雪夫大数定律

当随机变量 $X_1, X_2, \cdots, X_n$ 看做是从总体中随机抽出的一个容量为n的简单随机样本时，$\frac{1}{n}\sum_{i=1}^{n} X_i$ 正是该样本的均值 $\overline{X}$，μ正是该总体的期望。大数定律证明，当样本容量n充分大时，样本均值与总体均值间的误差可以有很大的把握被控制在任意给定的要求之内，这正是样本总体的理论依据。

（二）贝努里大数定律

设m是n次独立重复试验中事件A发生的次数，记事件A发生的频率为p，即 $P=\frac{m}{n}$，而 π 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

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

即当n充分大时，时间A发生的频率具有稳定性，并依概率收敛于事件A出现的概率。

实际上，事件A发生的频率P在统计学中被称为样本成数（一个特殊的平均数），事件A在每次试验中出现的概率π被称为总体成数。贝努力大树定律一方面提供了赝本成数推断总体成数的理论依据，另一方面也论证了“频率代概率”的概率统计定义。

二、中心极限定理

大数定律说明了当样本容量n充分大时，样本均值趋于总体均值，但并不等于总体均值，说明样本推断总体时存在误差。若要控制推断误差，显然须知样本均值这一随机变量的概率分布，可惜大数定律只提供了推断方法，并未给出推断误差的概率分布。而中心极限定理正好弥补了大数定律的这一不足。

（一） 林德贝格-勒维中心极限定理

设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，且服从相同的分布 $E(X_i)=\mu, D(X_i)=\sigma^2, (i=1,2,\cdots,n)$，记 $W_n=\dfrac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)$ 则对于任意实数x，都有

$$\lim_{n\to\infty}F_n(x)=\lim_{n\to\infty}P\{W_n\le x\}=\Phi(x)=\int_{-\infty}^{x}\varphi(x)dx$$

其中：

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

（一）林德贝格-勒维中心极限定理

该定理告诉我们，无论总体服从何种分布，只要它的平均数与标准差客观存在，我们就可以通过增大样本容量n的方式，保证样本均值$\bar{X}$近似服从正态分布，样本容量越大，样本均值的分布就越接近正态分布，一般认为样本容量n不少于30，即 $n \geq 30$ 的样本为大样本，大样本的均值 $\bar{X}$ 近似服从正态分布，即有

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

（二） 棣莫弗-拉普拉斯（De Moivre – Laplace ）中心极限定理

设n次独立重复试验中，事件A在每次试验中出现的概率为π（ $0<\pi<1$），记μ_n为n次试验中事件A出现的次数，且记 $Y_n^* = \dfrac{\mu_n - n\pi}{\sqrt{n\pi(1-\pi)}}$

则对任意实数x，都有 $\lim\limits_{n\to\infty} F_n(x) = \lim\limits_{n\to\infty} P\left(Y_n^* \le x\right) = \Phi(x) = \int_{-\infty}^{x} \varphi(x)dx$

其中： $\varphi(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

两类极限定理的意义

- 1.如果说大数定律是关于“均值具有稳定性”的一类定律，它提供了样本估计总体的方法，那么中心极限定理则是关于“估计误差概率分布”的一类定理，它不仅提供了估计方法，而且还提供了控制估计误差的方法。
- 2.中心极限定理还揭示了正态分布形成的机制，即如果某一个量是许多随机因素综合影响迭加形成的，在这许多影响因素中没有任何一个起着主导作用，那么这个量就是一个服从正态分布的正态随机变量。回归模型中的随机误差项常假定服从正态分布，其依据便在于此。