概率论与数理统计

Fan

2023年秋季

目录

摘要

本笔记结合概率论与数理统计课堂内容以及陈希孺所著的概率论与数理统计.

含义	记号
样本空间	Ω
基本事件	ω
事件A发生概率的估计值	$\widehat{P(A)}$

含义	记号
随机变量X的（边缘）分布函数	$F_X(x)$
随机变量X的（边缘）密度函数	$f_X(x)$
已知$\{Y=y\}$发生时X的条件密度函数	$f_{X\|Y}(x\|y)$
随机变量X的期望	EX
随机变量X的方差	$Var(X)$
随机变量X的标准差	$\sigma(X)$
服从参数n,p的二项分布	$B(n,p)$
服从参数λ的泊松分布	$P(\lambda)$
服从参数p的几何分布	$Ge(p)$
服从参数N,M,n的几何分布	$h(N,M,n)$
服从参数a,b的均匀分布	$U(a,b)$
服从参数μ,σ^2的正态分布	$N(\mu,\sigma^2)$
标准正态分布的分布函数	$\Phi(x)$
标准正态分布的密度函数	$\varphi(x)$
服从参数λ的指数分布	$E(\lambda)$
服从参数μ,σ^2的对数正态分布	$LN(\mu,\sigma^2)$
标准正态分布的p分位数	u_p
一般分布的p分位数	x_p

表 1: 常用记号

成绩组成为:

1. 10%考勤
2. 30%平时成绩
    - 作业(对正确率要求不高)
    - 线上(期中)考试
    - 随堂测试
3. 60%期末考试

1 概率与等可能概型

概率,又称或然率,机率,是表示某种情况(事件)出现的可能性大小的一种数量指标,它介于0与1之间.

1.1 主观概率

主观概率,顾名思义,是主观的,是依靠个人经验和感性判断的,是不严谨的. 但是值得注意的是, 主观概率仍有其意义,毕竟人类也并不是完全理性的.

主观概率有广泛的生活基础;能反映认识主体的倾向;能够体现出个体条件的差异所带来的影响.

1.2 试验与事件

在概率论中,'事件'一词的含义是:

1. 有一个明确界定的试验.'试验'一词,有人为,主动的意思,例如抛硬币,掷色子. 特别地,观测某种现象也可以视为是一种试验.

2. 这个试验的全部可能结果,这是在试验前就明确的. 例如'明天下不下雨'就有'下雨'和'不下雨' 两个结果,这一试验也可以记为(下雨,不下雨)

   注. 有些时候试验的全部可能结果虽然确实是在试验前就明确的,但是我们无法确切的知道,此时我们可以将该范围扩大,例如从某个正实数集的有限子集扩大到正实数集.这种操作是允许的,甚至有些时候为了方便起见会故意而为之.

   注. 随机试验的三个特点分别为:"可重复性"、"结果多样性和明确性"、"结果不确定性"。

   注. 这也就意味着,对于用一个试验,其得到的样本空间不是唯一的,是由试验者来决定的.例如扔色子这个试验,其样本空间可以是所有可能的点数,也可以是{点数是素数,点数不是素数}

3. 我们有一个明确的陈述,这个陈述界定了试验的全部可能结果中一个确定的部分. 这个陈述,或者说一个确定的部分,就叫做一个事件.

   注. 特别地,一个空集也算是所有可能结果中的一个确定部分,这一特例在集合论的语言中更好理解.

在概率论中,常称事件为'随机事件'或'偶然事件'.'随机'即事件是否在某次试验中发生取决于机遇,其极端为'必然事件'和'不可能事件'

1.3 概率的统计定义

概率的统计定义是:通过实验去估计事件概率的方法. 用频率去估计概率的直观背景即:一个事件出现的可能性大小,应由在多次重复试验中其出现的频繁程度去刻画.

但是,频率只是概率的估计而非概率本身.频率的重要性不在于能通过它来求出准确的概率,这实际上是不可能的, 而是它不仅提供了一种估计概率的方法,也提供了一种验证概率是否合理的标准.

1.4 概率的公理化

我们将一个实验的所有可能的结果所构成的集合称为样本空间,记为Ω,其元素ω称为基本事件. 由Ω的子集(包括其本身和空集)构成的一个集类$\mathcal{F}$(不必包含Ω的所有可能子集),其中的每一个成员就称为'事件'. 易知,每个事件都是由基本事件所组成的.事件有概率,其大小随事件而异.换句话说,概率是事件的函数,即定义在$\mathcal{F}$上的函数P. 对$\mathcal{F}$中的任一成员A,$P(A)$的值理解为事件A的概率.在公理体系中, 函数P应满足:

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1, P(\varnothing) = 0$
3. 加法公理

注. 1. 在课堂中,基本元素定义为单元集合$\{\omega\}$,其中ω可以是数,也可以不是数.例如它可以是'明天下雨',或是明天的降雨量'100'ml

2. 基本元素即单次试验得到的结果,因此'一次掷一个色子'和'一次掷两个色子'的基本元素的范围不一样
3. 样本空间可以是可列集,也可以是不可列集
4. 某次试验中发生了事件A即该试验对应的结果$\omega \in A$
5. 在课堂中,公理的内容为:$0 \leq P(A) \leq 1$;$P(\Omega) = 1$;概率的可列可加性,即可列个两两互斥的事件之和的概率等于它们的概率之和.在此基础之上我们推$P(\varnothing) = 0$,以及有限可加性.

   注. 概率为1的事件不一定为必然事件，概率为0的事件也不一定为不可能事件，这在连续型随机变量中十分常见。

A是B的子事件	$A \subseteq B$
A与B互为子事件	$A = B$
并事件	$\bigcup \mathcal{F}$
交事件	$\bigcap \mathcal{F}, \prod \mathcal{F}$
A与B为互不相容事件(互斥)	$A \cap B = \varnothing$
A, B之差	A-B
A的对立事件(补事件)	$\bar{A}, A^C$

注. 1. 当$\mathcal{F}$中的元素两两互斥时,$\bigcup \mathcal{F}$可写为$\sum \mathcal{F}$

2. A 与$\bar{A}$互斥,称他们为对立事件.A 与B为对立事件等价于$A \cup B = \Omega, A \cap B = \varnothing$

3. 事件之间的运算与集合之间的运算一致

4. $\overline{\text{至少n个}}$ = 至多$n-1$个.该结论有时可以用于化简题目.

5. 在求解概率时,可以利用集合论中的对偶关系,将并事件与交事件相互转化,以利用题目中的互斥或独立的条件,最后再利用对立事件的概率之间的关系.

6. 在求解对偶事件的概率时,可以通过实际含义求出式子,例如$\overline{AB + BC + AC}$实际上就是ABC至少有两个成立这一事件的对立事件,求其概率时可以直接计算至少有两个没有成立这一事件的概率.

7. 对于事件列$A_1, A_2, \cdots, A_n$,事件"至少有k个A_i"发生的概率为

$$P(\sum_{1 \le i_1 < \cdots < i_k \le n} \prod_{j=1}^{k} A_{i_j}),$$

而非

$$\sum_{t=k}^{\infty} P(\sum_{1 \le i_1 < \cdots < i_t \le t} \prod_{j=1}^{t} A_{i_j}).$$

8. 概率也有单调性，即对于任意两个事件A与B：

$$P(AB) \le P(A)$$

进一步，我们可以推出

$$P(AB) \le \frac{P(A) + P(B)}{2}$$

注. 有些时候题目中涉及到三个事件的交事件的概率，但是只给出了两个事件的事件的概率，或者互斥、互不相容的条件，这个时候往往要利用单调性。

常用的公式有:

- (加法公式)

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

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

- (减法公式)

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

- (结合律)

$$P(A(BC)) = P((AB)C)$$

- (分配律)

$$P((A \cup B)C) = P(AC \cup BC)$$

注. 分配律可以类比加法对于乘法的分配律.

注. 加法公式本质上是容斥原理。

1.5 等可能概型

定义. 若样本空间中的每一个样本点等可能性地发生,则称该模型为等可能概型. 当其为有限集时,称为古典概型,当其为某个空间(一维区间,二维平面或三维空间)时,称其为几何概型.

例. 在投针试验中,有两个维度,一个维度代表着针的中心到平行线的距离,另一个维度代表着针与平行线的夹角. 这两个维度都是等可能的,因此这是一个几何概型.

易知在等可能概型中

$$P(A) = \frac{|A|}{|\Omega|}$$

因此在等可能概型中,求出事件的概率的关键在于计数.

例. 甲,乙两人对赌,约定谁先胜三局谁就拿走奖金.现在已经赌了三局,甲二胜一负,请问此时甲赢得奖金的概率是多少?

答: 设想继续赌两局,则所有可能情况为:甲甲,甲乙,乙甲,乙乙.其中前三种情况中,甲均可以赢得奖金,因此概率为75%.

注. 这是陈书所给的解答,但是前两种情况中,实际上第一局甲获胜后赌局就已经结束了,然而假设继续赌局.这样的处理是否严谨?需不需要给出证明?在之后的题目中,也有使用过这种人为决定何时试验结束的手段,有时选取合适的节点能够大大简化解答的复杂度.

注. 我们可以用树状图的形式来模拟可能的结果,可以约定'向上生长的树枝'代表甲赢,'向下生长的树枝'代表乙赢.当决出胜负时便停止生长,我们可以通过末端树枝的方向来统计甲乙的胜负情况.但是不能简单地认为两种树枝的比值便是获胜概率的比值,因为越靠近树根的树枝代表的事件发生的概率越大. 换一种看法,我们把样本空间视为一个单位正方形,每一个树枝分叉便对应着一次二等分,根据最终结果分别染色为白色或黑色. 此时两种颜色的面积比值便是概率比值. 这样来看,陈书的处理方法便类似于第二种,保留了越靠前的胜利对应越多的概率的特性,且直观易懂.

注. 这实际上就是对于同一个试验，取不同的样本空间。陈书所取的样本空间是等概率的，因此大大简化了分析的过程。

1.6 古典概率计算

n个相异物件取$r(1 \le r \le n)$个的不同排列总数,为

$$P_r^n = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

n个相异物件取$r(1 \le r \le n)$个的不同组合总数,为

$$C_r^n = \binom{n}{r} = \frac{P_r^n}{r!} = \frac{n!}{r!(n-r)!}$$

注. 按公式

$$\binom{n}{r} = n(n-1)\cdots(n-r+1)/r!,$$

只要r为非负整数,n无论为任何实数,上式均成立.

与组合数相关的恒等式有:

1.
$$\binom{-1}{r} = (-1)^r$$

2.
$$\sum_{i=0}^{n} \binom{n}{i} = 2^n$$

3.

$$\sum_{i=0}^{n}(-1)^i\binom{n}{i}=0$$

4.

$$\binom{m+n}{k}=\sum_{i=0}^{k}\binom{m}{i}\binom{n}{k-i}$$

特别地，将$m=k=n$代入第四式可得

$$\binom{2n}{n}=\sum_{i=0}^{n}\binom{n}{i}^2$$

将n个相异物件分成k堆,第i堆有r_i个物件,则分法个数为:

$$\frac{n!}{\prod_{i=1}^{k}(r_i!)}$$

注. 需要指出的是,若有若干堆含有相同数量的物件时,这若干堆实际上是有序的. 该式的分子表明是先排序,分母则只是将每一堆的内部视为无序. 这一过程可以理解为现在平面上依次画上n个格子,然后先将他们按照数量分成k堆, 并进行排序,最终再将物品摆放在上面,并将每一堆的内部视为无序.

例. n个相异物品用绳子串起来,串成一个圆环,请问有多少种可能?

答: 这样形成的一个圆环,我们可以挑其中任一一个物品,从它开始,按照圆环的顺时针方向形成一个直线上的排列. 从而一个圆环对应着n个排列.所以有$\frac{n!}{n}=(n-1)!$个可能.

注. 这种解法是先计算出含有重复的情况,通常较易计算,且是实际情况总数的倍数,然后再除以倍数.

例. N件产品中有M件废品.从中随机挑取n件产品,其中有m件废品的概率是?

答:设该事件为 E,则

$$P(E)=\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

分母是指在n件里随意挑,分子则是指分别在非废品与废品中挑.

注. 严谨起见,还要标明变量的取值范围,一般就是符合常理的范围,例如不要无中生有.这里便要求 $m\leq M, n\leq N$

例. n双相异的鞋共2n只,随机地分成n堆,每堆2只,则‘各堆自成一双鞋’的概率是多少?

答:

$$P(E) = \frac{n!}{(2n)!/2^n}$$

这里分子是指将每双鞋视为整体,一般写作‘将他们捆绑起来’,然后进行排列.分子则是将n件物品分为k堆的问题.

例. n个男孩,m个女孩($m \leq n+1$)随机排成一列,则‘任意两个女孩都不相邻’这一事件E的概率为?

答:

$$P(E) = \frac{n!\binom{n+1}{m}m!}{(m+n)!}$$

分母是全排列,分子则指的是先排列男生,再选取女生所站的位置,然后再排列女生.

例. 有两盒火柴,每盒有n支火柴,每次随机打开一盒然后抽取一支.问‘第一次发现所抽的盒子为空，且这时另一盒中恰有m支火柴’的概率是多少?

答: 事件必然发生在第$2n-m+1$次抽取时,

$$P(E) = \frac{2\binom{2n-m}{n}}{2^{2n-m+1}}$$

分母指的是即使有一盒空了,也仍然会随机抽取. 分子中的因子2是因为对称性,$\binom{2n-m}{n}$是指把一盒抽光的n步在前2n-m中的分布.

注. 注意这里试验结束节点的选取.这里也可以选取抽取$2n-1$次后再结束,但是会十分复杂。然而,不同结束节点所得到的结果是一致的,尽管形式不一致,这也是获取恒等式的一种方法.

例. 有21本不同的书,随机地分给17个人.问“有6人得0本,5人得1本,2人得2本,4人得3本”这一事件E发生的概率是多少?

答:

$$P(E) = \frac{17!}{6!5!2!4!}\frac{21!}{0!^6 1!^5 2!^2 3!^4}\frac{1}{17^{21}}$$

不仅要将人分堆,也要将书分堆,所有可能的情况则是让每本书分别分发.

注. 将人分堆实际上是构造一个完备事件群. 而这种分堆方式背后隐藏着使得对于该完备事件群中的每个事件,不会发生重复,也不会发生遗漏.

注. 在计算排列组合时,尤其要注意是否发生情况重复.在该问题中,如果人或书是全排列,则会出现重复问题.

例 (抽签问题). 假设有N个签，其中M个有奖。假设人们依次无放回抽签，证明无论是第几个抽签的，中奖的概率都为M/N。

证明. 假设第k位中奖，那么我们现在M个有奖的签中挑选他抽中的签，然后在安排其他人的签，即

$$P(\text{第k位中奖}) = M \times \frac{(N-1)(N-2)\cdots(N-(k-1))}{N(N-1)\cdots(N-(k-1))} = M/N$$

□

1.7 加法定理

定理 1.1. 若干个互斥事件之和的概率,等于各事件的概率之和,即:

$$P(\sum \mathcal{F}) = \sum P(\mathcal{F})$$

特别地,

$$P(\bar{A}) = 1 - P(A)$$

1.8 条件概率

条件概率是附加在一定的条件之下所计算的概率.当说到‘条件概率’是,总是指另外附加的条件,其形式可以归结为‘已知某事件发生了’. 一般地,

定义. 设有两个事件A,B,而$P(B) \neq 0$.则‘在给定B发生的条件下A的条件概率’,记作$P(A|B)$,定义为

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

注. 条件概率也可以公理化,即可由条件概率的非负性,规范性以及可列可加性推出其他条件概率的性质.

注. 这个式子在古典概型中是十分显然的,我们将其拓展至一般的模型中.在实际运用中,可以直接考虑给定条件发生的情况下的概率.

从韦恩图的角度,P(A)是A在Ω中所占的‘比例’,而$P(A|B)$则是把B给从韦恩图中裁下来后在测量A所占的‘比例’.

例. 假设有10个色子,事件B为{至少有两个色子点数为1},事件A_1为{色子1点数为1},A_2为{至少有一个色子点数为1},则

$$P(B|A_1) > P(B|A_2)$$

即事件A_1更有利于接下来获得点数1.

直观上可以这样理解,A_1发生后,想要让B发生,需要至少在9个色子里扔出来1个1点,而若以A_2为前提, 则仍需在 10个色子里扔出来2个1点,这从直觉上看应该更难.(可以考虑极端情况来验证正确性,显然从100个色子中扔出来992个1点比9个色子里扔出来1个色子难得多).

1.9 事件的独立性,概率乘法定理

若$P(A) = P(A|B)$,则B的发生与否对A发生的可能性毫无影响.这时,在概率论中就称A,B两事件独立,同时有

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

注. 此时也有

$$P(A|\bar{B}) = P(A)$$

即如果A与B独立,那么A与$\bar{B}$也独立.这就像从一杯糖水中倒出一杯同样浓度的糖水,剩余的糖水浓度不变. 类似地,$\bar{A}$与B独立,$\bar{A}$与$\bar{B}$独立.

定义. 两个事件A,B若满足

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

则称A,B独立

注. 取这一式作为定义是因为当$P(B) = 0$时,该式也成立

注. 这一定义的主要用途是用于验证.一些数据可能在数值上符合'独立'的条件,但只是巧合.

注. 当$P(A) \neq 0$时，A与B相互独立的充要条件为

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

即缩小了样本空间后B发生的概率保持不变。再利用定理1.4，当$0 < P(A) < 1$时，A与B相互独立的充要条件还有

$$P(B|A) + P(\bar{B}|\bar{A}) = 1$$

$$P(B|\bar{A}) + P(\bar{B}|A) = 1$$

等一系列等式，其本质仍然是，若两个事件相互独立，且其中一个非空，则将样本空间缩小为该事件后，另一个事件的概率不变。同时，如果两事件独立，那么它们的互斥事件也相互独立，与对方的互斥事件也相互独立。

定理 1.2. 两独立事件A,B的积AB的概率$P(AB)$等于其各自概率的积$P(A)P(B)$.

定义. 设$A_1, A_2, \cdots$为有限或无限个事件.如果从其中任意取出有限个$A_{i_1}, A_{i_2}, \cdots, A_{i_m}$,都成立

$$P(A_{i_1}A_{i_2}\cdots A_{i_m}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_m})$$

则称事件$A_1, A_2, \cdots$相互独立,或简称独立.

注. 这个定义与由条件概率出发的定义是等价的,后者是说,对任何互不相同的$i_1, i_2, \cdots, i_m$,有

$$P(Ai_1|Ai_2\cdots Ai_m) = P(A_{i_1}).$$

即任意事件A_{i_1}发生的可能性大小,不受其他事件发生的影响.这更接近于独立性的原义. 但是,该式的左边依赖于$P(A_{i_2}\cdots A_{i_m}) > 0$,否则无意义.

注. 多个事件的独立性往往产生于由多个试验构成的复合试验中,每个事件只与其中一个试验有关. 例如,在试验E中,我们先抛硬币,记录是正面还是反面,再扔色子,记录其点数,如果事件A只关心是否为正面,事件B只关心点数是否为素数, 那么,从直觉上,这两个事件是独立的,事实也是如此.

定理 1.3 (乘法定理)**.** 若干个相互独立事件$A_1, A_2, \cdots, A_n$之积的概率,等于各事件概率的乘积:

$$P(A_1\cdots A_n) = P(A_1)\cdots P(A_n)$$

注. 显然,从若干个相互独立事件中取出若干个事件,它们之间也是独立的.

注. 可以证明,若干个相互独立事件分为n堆,每一堆中的独立事件之间进行运算,例如差,并,所得到的n个新的事件之间仍然相互独立.

定理 1.4. 若一列事件$A_1, A_2, \cdots$相互独立,则将其中任一部分改为对立事件时,所得事件仍为相互独立.

注. '相互独立'与'两两独立'并不等价,前者是后者的充分非必要条件.

但是'两两互斥'与'相互互斥'是等价的.

对于n个事件,若要说明它们两两独立,则要验证$\binom{n}{2}$个式子,而若要说明它们相互独立, 则要验证$2^n - n - 1$个式子

定理 1.5 (乘法公式)**.**

$$P(\prod_i A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|\prod_{i\neq n} A_i)$$

注. 该式没有独立性的要求。若A_i之间相互独立,我们可以由此推出乘法定理.

有些时候,若题目中没有明确给出相互独立的条件,我们应该用乘法公式而非乘法定理. 例如如果题目给出第一关不被淘汰的概率为p,第二关不被淘汰的概率为q, 那么这实际上是假设第一关不被淘汰后的第二关不被淘汰的条件概率为q, 通过两关的概率用乘法公式计算.

例 (三门问题). 有一个综艺节目,选手要从三个门中选一个,其中只有一个门后有奖品,主持人知道哪个门后有奖. 选手选择了一个门后,主持人打开一扇既不是选手选的,也不是有奖的门,请问此时如果选手改变主意重新选择, 他获奖的概率会怎么改变?

因为选手刚开始作决定和之后再作决定之间,选手并没有获取有用的信息,毕竟主持人总是能打开一扇既不是选手选的,也不是有奖的门. 具体分析,选手一开始的概率是1/3,假设主持人能打开这样一扇门,这个事件A 的概率其实是1,在此基础之上,选手选中奖的概率$P(B|A)$ 为$P(BA)/P(A) = P(\Omega B)/P(\Omega) = P(B) = 1/3$,所以没有任何变化.也可以用全概率的公式来计算.

1.10 全概率公式与贝叶斯公式

设$B_1, B_2, \cdots$为有限或无限个事件,它们两两互斥且在每次试验中至少发生一个. 此即:

- $B_iB_j = \varnothing$(不可能事件)$(i \neq j)$
- $B_1 + B_2 + \cdots = \Omega$(必然事件)

我们把具有这些性质的一组事件称为一个"完备事件群".特别地,任一事件及其对立事件组成一个完备事件群.

对于任一事件A,因为Ω为必然事件, 有

$$A = A\Omega = AB_1 + AB_2 + \cdots .$$

因$B_1, B_2, \cdots$两两互斥,我们使用加法定理有

$$P(A) = P(A\Omega) = P(AB_1) + P(AB_2) + \cdots ,$$

再由条件概率的定义,我们把$P(AB_i) = P(B_i)P(A|B_i)$代入上式,得

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots$$

上式即“全概率公式”.可以理解为:“全”部概率$P(A)$被分解成了许多部分之和. 它的理论和实际用途在于:再比较复杂的情况下直接计算$P(A)$不易,但A总是随某个B_i伴出,适当去构造这一组B_i往往可以简化计算.

注. 有点类似于分类讨论的思想,多种情况互不重复但又覆盖所有可能.

这一公式还可以从另一个角度去理解:把B_i视为导致A发生的一种可能途径.对不同途径,A发生的概率即条件概率$P(A|B_i)$各不相同,而采取哪种途径却是随机的.

例. 设一个家庭有k个小孩的概率为$p_k(k=0,1,2,\cdots)$.又设各小孩的性别分别独立,且生男,女孩的概率各为$1/2$.试求事件$A=$\{家庭中所有小孩为同一性别\}的概率.

答: 引进事件$B_k=$\{家庭中有k各小孩\},则$B_1,B_2,\cdots$构成完备事件群,$P(B_k)=p_k$.现考虑$P(A|B_k)$.约定当$k=0$时其值为1. 若$k\geq 1$,则k各小孩性别全同的概率为$2*1/2^k=1/2^{k-1}(k\geq 1)$,由此,用全概率公式,得出

$$P(A)=p_0+\sum_{k=1}^{\infty}p_k/2^{k-1}$$

在全概率公式的假设之下,有

$$P(B_i|A)=P(AB_i)/P(A)=P(B_i)P(A|B_i)/\sum_j P(B_j)P(A|B_j)$$

我们可以把B_i视为已知概率的‘因’,把A视为‘果’,但是这是有概率的因果,也就是B_i就算发生了,A也可以不发生,同样的,A发生了,但某些B_i也可以不发生. 那么这个公式就定量描述了,当‘果’发生后,‘因’的发生概率会怎样变化,即从$P(B_i)$变为$P(B_i|A)$. 从另一个角度,该公式告诉我们各原因可能性的大小$P(A|B_i)$与结果发生时该原因也发生的可能性$P(B_i|A)$的大小成比例.

注. 或者说，贝叶斯公式告诉我们，一个因的概率$P(B_i|A)$是这个因本身发生的概率在加权后的占比$P(B_i)P(A|B_i)$，这个权即$P(A|B_i)$，即它导致果的概率。

例. 有三个盒子C_1,C_2,C_3,各有100个球.三个盒子含有白球的数量分别为80,10,10.现从三个盒子中随机抽取一个,然后再从里面抽取一个球,结果抽出白球. 问‘该白球是从C_i盒中抽出的可能性有多大?’$(i=1,2,3)$

分析: 这里的因是$B_i=$\{抽出的为C_i盒\}$(i=1,2,3)$,果为A=\{抽出白球\}.‘果’的原来的概率$P(B_i)$,每个原因导致结果的可能性$P(A|B_i)$均已知, 这

两个量都是可以通过设置的条件推理出来,或者通过试验来估计的,我们要求的则是$P(B_i|A)$,即最终的结果中,各个由各个原因导致的可能性分别是多少,而这个可以使用贝叶斯公式求得. 这也是为什么贝叶斯公式如此重要,它提供给我们执果溯因的能力.与之相对的则是全概率公式.

注. 当试验简单时,$P(B_i|A)$可能也能十分轻松地求出,但是试验复杂时,往往$P(B_i)$和$P(A|B_i)$的获得比$P(B_i|A)$的获得轻松不少. 例如,研究一个死亡病例是由病毒A,B,C致死的概率分别是多少时,$P(B_i)$只需通过统计发病率即可,$P(A|B_i)$也可以通过统计患病病人的死亡率得到,然后通过贝叶斯公式求得$P(B_i|A)$.若想直接求得$P(B_i|A)$,则需要研究大量的死亡病例.这或许便是统计学的意义所在,将复杂的分析用简单的统计来代替.

例. 假设我们认为诚实的孩子说谎的概率为0.1，不诚实的孩子说谎的概率为0.5.有个孩子被认为有80%的概率是诚实的，那么当他说了一次谎后，我们认为他不诚实的概率为

$$P(\text{诚实}|\text{说了一次谎}) = \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.2 \times 0.5} = 0.444.$$

此时，我们对这个小孩的认识得到了更新，认为他诚实的概率变为了44.4%.

当他再一次说谎时，我们认为他诚实的概率为

$$P(\text{诚实}|\text{再一次说谎}) = \frac{0.444 \times 0.1}{0.444 \times 0.1 + 0.556 \times 0.5} = 0.138.$$

如果他这次并没有说谎，我们认为他诚实的概率为

$$P(\text{诚实}|\text{没有再一次说谎}) = \frac{0.444 \times 0.9}{0.444 \times 0.9 + 0.556 \times 0.5} = 0.590.$$

注. 上式体现出来的贝叶斯公式的价值在于：他反映了人们对于未知事物的认识过程。我们一开始有先验概率，在上例中便是孩子是诚实的概率。在得知新的信息后，便对其进行更新，如此循环往复，直到接近其本质属性。

2 随机变量及概率分布

2.1 一维随机变量

2.1.1 随机变量的概念

随机变量可以理解为随机试验的试验结果的函数，例如掷色子这个试验中，色子的点数X便是一个随机变量，它可以取$1, 2, \cdots, 6$等6个值。在试验之前，我们不能预知它将取什么值，但试验之后，取值就确定了。

定义. 随机变量定义为

$$X = X(\omega)$$

其中

$$X : \Omega \to \mathbb{R}^d$$

是一个定义在样本空间Ω上的实值单值函数。

定义. 如果在一个试验中，我们只关心某个事件A 是否发生，则称该试验为伯努利试验，该试验的样本空间为$\{A$发生$, \bar{A}$发生$\}$，对应的随机变量为

$$X(\omega) = \begin{cases} 1, \omega \in A \\ 0, \omega \notin A \end{cases}$$

注. 这样函数也被称为示性函数，记作$\mathbf{1}_A$。对于事件A_i，也可以写为指示变量X_i，这种写法表明事件可以视作随机变量的一种特例。

随机变量的反面便是所谓的 '确定变量'，即其取值遵循某种严格的规律的变量，在试验结束之前我们便可以准确预知其取值。

随机事件的本质实际上便是随机变量。如果我们关心一件事情是否发生，我们可以用0和1来分别表示发生与否。如果我们关心其中的某个量的大小，我们也自然可以将其视为随机变量。

随机变量可以分为离散型随机变量和连续型随机变量两类。前者的特征是，其可能的取值至多有可列个，而后者的可能取值则是不可列个。

2.1.2 离散型随机变量的分布

定义. 设X为离散型随机变量，其全部可能值为$\{a_1, a_2, \cdots\}$，则

$$p_i = P(X = a_i) \quad (i = 1, 2, \cdots)$$

称为X的概率函数。

我们常把上式称为随机变量X的 "概率分布"。它可以列表的方式给出，称该表为X的分布表：

可能值	a_1	a_2	$\cdots$	a_i	$\cdots$
概率	p_1	p_2	$\cdots$	p_i	$\cdots$

由概率的公理化定义可知：

$$p_i \geq 0, \quad p_1 + p_2 + \cdots = 1.$$

注.

$$P(X \in D) = \sum_{x_i \in D} P(X = x_i)$$

$$P(f(x) \in D) = \sum_{f(x_i) \in D} P(X = x_i) = \sum_{x_i \in f^{-1}(D)} P(X = x_i)$$

对离散型变量，用概率函数取表达其概率分布是最方便的，也可以用下面定义的分布函数表示：

定义. 设X为一维随机变量，则函数

$$P(X \leq x) = F(x) \quad (-\infty < x < \infty)$$

称为X的分布函数。

注. 注意，这里并未限定X为离散型随机变量，它对任何随机变量都有定义。

显然，对于任何随机变量X，其分布函数$F(X)$满足

1. 单调非降
2. $\lim_{x \to \infty} F(x) = 1$
3. $\lim_{x \to -\infty} F(x) = 0$
4. $0 \leq F(x) \leq 1 (-\infty < x < +\infty)$
5. 右连续

注. 之所以是右连续，可以理解为形如$(-\infty, x]$的左极限是一个开区间，而右极限则是一个闭区间。

注. 分布函数的这几条性质是其特征性质，这是一个函数为某个随机变量的分布函数的充要条件。

注.

$$P(a) = F(a) - F(a - 0)$$

当随机变量X服从某种分布F时，我们用$X \sim F$来表达这一点。

定义 (二项分布). 如果我们将伯努利试验重复n次，则我们称其为n重伯努利试验。假设：

1. 各次试验的条件稳定，即事件A在各次伯努利试验中发生的概率p保持不变

2. 各次试验相互独立

那么，以X记事件A在这n次试验中发生的次数，则X可取$1, 2, \cdots, n$ 等值。

概率分布为

$$p_i = \binom{n}{i} p^i (1-p)^{n-i} = b(i; n, p) \qquad (i = 1, 2, \cdots, n)$$

X所遵从的概率分布称为二项分布，并常记为$B(n, p)$。

注. 这两个条件符合抽取后又放回的抽样方式，如果样本数量相较于抽取的个数很大，那么即使抽取后不放回，所得到的概率分布也近似于二项分布。

注.
1. 若$p \le 1/(n+1)$，则当k增加时$b(k; n, p)$非增；

2. 若$p \ge 1 - 1/(n+1)$，则当k增加时$b(k; n, p)$非降；

3. 若$1/(n+1) < p < 1 - 1/(n+1)$，则当$k$增加时，$b(k; n, p)$先增后降，最大值当$k = \lceil (n+1)p - 1 \rceil, \lfloor (n+1)p \rfloor$取到。

例. 设随机变量X服从二项分布$B(n, p)$,k为小于n的非负整数，记$f(p) = P(X \le k)$，试用概率方法证明$f(p)$随p增加而下降。

证明. 易得

$$f(p) = \sum_{i=0}^{k} b(n; i, p)$$

任取概率$p_1 < p_2$，我们设想一个试验，有三个结果A_1, A_2, A_3，其概率分别为$p_1, p_2 - p_1, 1 - p_2$, 记$A = A_1 + A_2$，以X_i记n次实验中A_i发生的次数，则$X_1 \sim B(n, p_1), X_1 + X_2 \sim B(n, p_2)$. 所以

$$P(X_1 \le k) = \sum_{i=0}^{k} b(i; n, p_1) \qquad P(X_1 + X_2 \le k) = \sum_{i=0}^{k} b(i; n, p_2).$$

因为当$X_1 + X_2 \le k$时必有$X_1 \le k$，所以我们有$P(X_1 + X_2 \le k) \le P(X_1 \le k)$,即

$$\sum_{i=0}^{k} b(i; n, p_2) \le \sum_{i=0}^{k} b(i; n, p_1)$$

□

定义 (泊松分布). 若随机变量X的可能取值为$1, 2, \cdots$，且概率分布为

$$P(X = i) = e^{-\lambda}\lambda^i/i!,$$

则称X服从泊松分布，常记为$X \sim P(\lambda)$.此处，$\lambda > 0$是某一常数。

这个分布也是重要的离散型分布之一，它多出现在当X表示在一定的事件或空间中出现的事件个数这种场合，并且假设事件发生的期望是与观测的长度称正比。即事件发生在某一时间段或地点内，那么我们观测到该事件发生的次数的期望应该与我们所观测的时间段长度或地点面积大小成正比，这个比值λ，这个比值通常较小。

将该模型应用于单位时间段中，有：

$$P(X = i) = \lim_{x\to\infty} \binom{n}{i} (\frac{\lambda}{n})^i (1 - \frac{\lambda}{n})^{n-i} = e^{-\lambda}\lambda^i/i!$$

从模型的构造可以看出，泊松分布可以作为二项分布的极限而得到，此时二项分布对应的伯努利试验便是在单个时间段内事件是否发生。

一般地说，若$X \sim B(n, p)$,其中n很大，p很小，且$np = \lambda$不太大时，则X 的分布接近于泊松分布$P(\lambda)$。

例. 设随机变量X服从泊松分布$P(\lambda)$。k为正整数。试用概率的方法来证明$P(X \leq k)$随λ增加而下降。

证明. 设$\lambda_1 < \lambda_2$，X_1, X_2独立，分别服从泊松分布$P(\lambda_1)$和$P(\lambda_2-\lambda_1)$，则X_1+X_2服从泊松分布$P(\lambda_2)$。再有$P(X_1 + X_2 \leq k) \leq P(X_1 \leq k)$即推出所要的结果。 □

定义 (超几何分布). 考虑在N件物品中有M件目标物品，从这N个物品中抽n个物品，以X记抽出的物品中所含目标物品的个数，则X的分布为

$$P(X = m) = \binom{M}{m}\binom{N - M}{n - m}/\binom{N}{n}.$$

该分布称为超几何分布。

之所以称为超几何分布，是因为其形式与 “超几何函数”的级数展开式的系数相关。这个过程实际上是无放回的抽样，当抽取的数量与总数相比可以忽略不记时，它与有放回的抽样相似。即当n固定，$M/N = p$固定，$N \to \infty$时，X近似服从二项分布$B(n, p)$。

定义. 概率分布

$$P(X=i)=\binom{i+r-1}{r-1}p^r(1-p)^i, i=0,1,2,\cdots$$

称为负二项分布。

定义.

$$P(X=i)=p(1-p)^{i-1}, i=1,2,\cdots$$

称之为几何分布。

定理 2.1 (几何分布的无记忆性)**.** 设随机变量$X\sim Ge(p)$，则对任意正整数m和n有

$$P(X>m+n|X>m)=P(X>n)$$

2.1.3 连续型随机变量的分布

定义. 若存在非负可积函数$f(x)$，使得对任一实数x，随机变量X的分布函数都有

$$F(x)=\int_{-\infty}^{x}f(x)\mathbf{d}x,$$

则称随机变量X为连续型随机变量。

注. 连续型随机变量的分布函数是一个积分形式，因此不会产生阶跃，即其图像应该是连续的。而离散型随机变量的分布函数应该是由多个阶跃函数叠加而成的。存在着一些随机变量，它既不是离散型随机变量，也不是连续型随机变量，它的分布函数的图像实际上可以分解为阶跃函数和非零的连续函数。

如果一个随机变量的分布函数中存在着不连续点，则该分布函数不存在概率密度函数，因为它在不连续点的概率不为0.

概率分布函数是刻画随机型连续变量的概率分布的一个方法，但是在理论和实用上更加方便的是所谓的 “概率密度函数”，或简称密度函数。

定义. 设连续型随机变量X有概率分布函数$F(x)$，则$F(x)$的导数$f(x)=F'(x)$称为X的概率密度函数。

注. 概率分布函数$F(x)$不一定处处可导。不可导处的概率密度函数按照概率分布函数在该点两侧定义的开闭区间来进行连续延拓，或者统一置零。

概率密度函数$f(x)$反映了在x点处（无穷小区段内）单位长的概率，或者说它反映了概率在x点处的“密集程度”。

显然，连续型随机变量X的密度函数$f(x)$都具有：

1. $f(x) \geq 0$;
2. $\int_{-\infty}^{\infty} f(x)\mathbf{d}x = 1$;
3. 对任何常数$a < b$,有

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(x)\mathbf{d}x.$$

概率分布函数的图像反映出来的是各个点处的之前发生的概率的累计值，而概率密度函数则是能够体现出该处的概率大小。

定义. 如果一个随机变量具有概率密度函数

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

则称X为正态随机变量，并记作$X \sim N(\mu, \sigma^2)$。这里$\mu, \sigma^2 \in \mathbb{R}$且$\sigma^2 \neq 0$。它们分别称为这个分布的位置参数和形状参数。

当$\mu = 1, \sigma^2 = 1$时，上式变为

$$f(x) = e^{-x^2/2}/\sqrt{2\pi},$$

它是正态分布$N(0,1)$的密度函数。$N(0,1)$也被称为标准正态分布，其概率密度函数和分布函数分别记为$\varphi(x), \Phi(x)$。易知$\Phi(x) = 1 - \Phi(-x)$。可以证明，若$X \sim N(\mu, \sigma^2)$,则

$$Y = (X - \mu)/\sigma \sim N(0,1).$$

注. 更一般地，若$X \sim N(\mu, \sigma^2), Y = kX + c, k \neq 0$,则$Y \sim N(k\mu + c, k^2\sigma^2)$，即正态分布经线性变化后仍服从正态分布。变化后的参数可以根据变化后的期望和方差逆推出来。

定义. 设随机变量$X \sim N(0,1)$，若

$$P(X \leq u_p) = \Phi(u_p) = p,$$

则称u_p为标准正态分布的p分位数，又称标准正态分布的下（尾）p分位数。

定义. 设$X \sim N(\mu, \sigma^2)$，则$Y = e^X \sim LN(\mu, \sigma^2)$，即服从对数正态分布。

定义 (指数分布). 若随机变量X有概率密度函数

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{当} x > 0\text{时} \\ 0, & \text{当} x \leq 0\text{时} \end{cases}$$

则称X服从指数分布。其中$\lambda > 0$，为参数。

指数分布常见于寿命分布，条件是无老化，或者说瞬时失效率为常量。事实上，λ 便是失效率。

例. 有一大批元件，其寿命服从指数分布，固定一个时间$T > 0$，让这一个元件从时刻0开始工作。每当这个元件坏了时，便马上用一个新的替换。以X记到时刻T为止的替换次数。证明：X服从泊松分布$P(\lambda T)$。

易得p_0的值。记泊松分布$P(\lambda T)$的概率密度函数为$f(x)$，则

$$p_k = \int_0^T f(t)\mathbf{d}t * p_{k-1}|_{T=T-t}$$

注. 该积分可以理解为求和的极限，是对0 T中的时间微元。对于每一段时间微元$\mathbf{d}t$，$f(t)\mathbf{d}t$是该元件寿命在t附近的概率，也就是第一个元件在t时刻附加损坏的概率，然后再利用乘法原理，乘以p_{k-1}，其中T替换为t之后的时间段，即$T - t$。

上述的归纳法利用的是将n次损坏依次拆分为 1次和$n-1$次。如果将其依次拆分为$n-1$和 1次，我们将无法得到和上式类似的积分式

$$p_k = \int_0^T p_{k-1}|_{T=t} * p_1|_{T=T-t}\mathbf{d}t$$

是错误的，其中的时间微元没法得到解释，其原因在于其中没有密度函数。

定理 2.2 (指数分布的无记忆性). 若随机变量$X \sim E(\lambda)$ ，则对任意的$s, t > 0$，有

$$\frac{P(X > s + t)}{P(X > s)} = e^{-\lambda t},$$

即

$$P(X > s + t | X > s) = P(X > t).$$

定义. 设随机变量X有概率密度函数

$$f(x) = \begin{cases} 1/(b-a), & \text{当} a \leq x \leq b\text{时} \\ 0, & \text{其他} \end{cases}$$

则称X服从区间$[a,b]$上的均匀分布，并常记为$X \sim R(a,b)$, 这里a,b都是常数，$-\infty < a < b < \infty$。

定理 2.3 (分布的可加性)**.** 设随机变量X与Y相互独立，

1. 若$X \sim B(m,p), Y \sim B(n,p)$,则$X+Y \sim B(m+n,p)$;
2. 若$X \sim P(\lambda_1), Y \sim P(\lambda_2)$,则$X+Y \sim B(\lambda_1+\lambda_2)$;

注. 二项分布的可加性很好理解，因为二项分布是多重伯努利试验对应的分布。而泊松分布是二项分布的极限，λ的含义是期望，这样看也就显然了。

定理 2.4. 设$X_1, X_2, \cdots, X_n$独立同分布，X_1有分布函数$F(x)$和密度函数$f(x)$，记

$$Y = \max(X_1, X_2, \cdots, X_n), \qquad Z = \min(X_1, X_2, \cdots, X_n),$$

则Y, Z的概率密度函数分别为

$$f_Y(x) = nF^{n-1}(x)f(x)$$

$$f_Z(x) = n[1-F(x)]^{n-1}f(x)$$

注. $F(x)$随x减少而减少，而min的分布集中在max的左侧，所以两者的概率密度函数形式有如上差别。

2.2 多维随机变量(随机向量)

一般地，设$X = (X_1, X_2, \cdots, X_n)$为一个$n$维向量，其每个分量，即$X_1, X_2, \cdots, X_n$，都是一维随机变量，则称$X$是一个$n$维随机向量或$n$维随机变量。

随机向量也有离散型与连续型之分。

2.2.1 离散型随机向量的分布

一个随机向量$X = (X_1, X_2, \cdots, X_n)$，如果其每一个分量$X_i$都是一维离散型随机变量，则称$X$为离散型的。

定义. 以$\{a_{i1}, a_{i2}, \cdots\}$记$X_i$的全部可能值$(i = 1, 2, \cdots)$,则事件$\{X_1 = a_{1j_1}, X_2 = a_{2j_2}, \cdots, X_n = a_{nj_n}\}$的概率

$$p(j_1, j_2, \cdots, j_n) = P(X_1 = a_{1j_1}, X_2 = a_{2j_2}, \cdots, X_n = a_{nj_n})$$

$$(j_1 = 1, 2, \cdots; j_2 = 1, 2, \cdots; \cdots, j_n = 1, 2, \cdots)$$

称为随机向量$X = (X_1, X_2, \cdots, X_n)$的概率函数或概率分布，概率函数应满足条件

$$p(j_1, j_2, \cdots, j_n) \geq 0, \sum_{j_n} \cdots \sum_{j_2} \sum_{j_1} p(j_1, j_2, \cdots, j_n) = 1.$$

定义 (多项分布)**.** 设$A_1, A_2, \cdots, A_n$ 是某一试验之下的完备事件群，分别以$p_1, p_2, \cdots, p_n$记事件$A_1, A_2, \cdots, A_n$的概率。先将试验独立地重复N次，而以X_i记在这N次试验中事件A_i 出现的次数$(i = 1, 2, \cdots, n)$，则$X = (X_1, X_2, \cdots, X_n)$为一个$n$维随机变量。其概率分布为

$$P(X_1 = k_1, X_2 = k_2, \cdots, X_n = k_n) = \frac{N!}{k_1! k_2! \cdots k_n!} p_1^{k_1} p_2^{k_2} \cdots p_n^{k_n}$$

其中k_i为非负整数，$k_1 + \cdots + k_n = N$。这种分布记为$M(N; p_1, p_2, \cdots, p_n)$

注. 多项分布的名称由来是因多项展开式

$$(x_1 + \cdot + x_n)^N = \sum^{*} \frac{N!}{k_1! \cdots k_n!} x_1^{k_1} \cdots x_n^{k_n}.$$

当总体按某种属性分为几类时，就会涉及到多项分布

2.2.2 连续型随机变量的分布

设$X = (X_1, X_2, \cdots, X_n)$是一个$n$为随机向量，其取值可以视为$n$维欧氏空间$\mathbb{R}^n$中的一个点。如果$X$的全部取值能够充满$\mathbb{R}^n$中某一区域，则称它是连续型的。

定义. 若$f(x_1, x_2, \cdots, x_n)$是定义在$\mathbb{R}^n$上的非负函数，使对$\mathbb{R}^n$ 中的任何集合A，有

$$P(X \in A) = \int_A f(x_1, x_2, \cdots, x_n) \mathbf{d}V,$$

则称f是X的（概率）密度函数。

显然，概率密度函数f必须满足

$$\int_{\mathbb{R}^n} f(x_1, x_2, \cdots, x_n) \mathbf{d}V = 1.$$

注. 对于随机向量的分布函数$F(x_1, x_2, \cdots, x_n)$，有以下性质

1. $0 \leq F(x_1, x_2, \cdots, x_n) \leq 1, x_1, x_2, \cdots, x_n \in (-\infty, \infty)$

2. F关于其中任一个自变量单调不减。

3. F关于其中任一个自变量右连续。

4. 对于任意一个分量x_i，当$x_i \to -\infty$时，$F \to 0$

5. 当所有分量趋向正无穷时，F趋向1。

6. 非负性，即F对应的随机变量取值为$\mathbb{R}^n$中的任意区域的概率非负。

注. 对于高维随机向量，第6条要求是否满足较不直观，需多加注意。特别地，对于二维随机向量，第6条等价于“对任意$a < b, c < d, F(b,d) - F(b,c) - F(a,d) + F(a,c) \geq 0$”。

例. 设

$$F(x,y) = \begin{cases} 0, & x+y<1, \\ 1, & x+y \geq 1, \end{cases}$$

易知该函数满足前5条性质，但是当取一个横跨$x+y=1$的矩形区域时，第6条便不满足。

定义 (二维正态分布).

$$f(x_1,x_2) = (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})\exp[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-a)^2}{\sigma_1^2} - \frac{2\rho(x_1-a)(x_2-b)}{\sigma_1\sigma_2} + \frac{(x_2-b)^2}{\sigma_2^2}\right)].$$

上式为二维正态分布的概率密度。其中$a, b, \sigma_1^2, \sigma_2^2, \rho$为参数，取值范围分别为

$$-\infty < a < \infty, -\infty < b < \infty, \sigma_1^2 > 0, \sigma_2^2 > 0, -1 < \rho < 1.$$

常把这个二维正态分布记为$N(a, b, \sigma_1^2, \sigma_2^2, \rho)$。

二维正态分布的概率密度函数的图像好似一个椭圆切面的钟倒扣在Ox_1x_2平面上，其中心在(a,b)点。

注. 在对正态分布的密度函数进行积分时，常常通过换元来使得被积函数转化为$e^{-x^2/2}$的形式。

我们有

$$\int_{-\infty}^{\infty} e^{-t^2/2}\mathbf{d}t = \sqrt{2\pi}$$

注. 可以证明，若联合分布为正态分布，则其边缘密度也为正态分布，但反之则不成立。

二维情形中的一个反例如下：

记二维正态分布的密度函数为$f(x,y)$，则

$$g(x,y)=\begin{cases}f(x,y)+xy/100, & \text{当}x^2+y^2\leq 1\text{时}\\ f(x,y), & \text{当}x^2+y^2>1\text{时}\end{cases}$$

为一个反例。

不论是一维还是多维情形，在定义连续型随机变量时，实质之点都在于它有概率密度函数存在，这一点可以直接取为连续型随机变量的定义：它就是有密度函数的随机变量。至于它可以在一个区间或区域上连续取值倒不是本质的，甚至也是不确切的。与离散型随机向量的定义不同，连续型随机向量不能简单定义为 “其各分量都是一维连续型随机变量的那种随机向量”。例如：设$X_1\sim R(0,1), X_2=X_1$，则随机向量$X=(X_1,X_2)$的两个分量X_1,X_2都是连续型的，但是其取值范围只是二维平面上的一条线段，因而不可能有函数能够成为其概率密度函数。

注. 与一维情况一样，也可以用概率分布函数取描述多维随机向量的概率分布，其定义为

$$F(x_1,x_2,\cdots,x_n)=P(X_1\leq x_1,X_2\leq x_2,\cdots,X_n\leq x_n).$$

然而，在多维情况下，分布函数极少应用。

2.2.3 边缘分布

设$X=(X_1,X_2,\cdots,X_n)$为一个n维随机向量。X有一定的分布F，这是一个n维分布。因为X的每个分量X_i都是一维随机变量，故它们都有各自的分布$F_i(i=1,2,\cdots,n)$,这些都是一维分布，称为随机向量X或其分布F的 “边缘分布”。边缘分布完全由原分布F确定。

注. 反过来不对，即使知道了所有X_i的边缘分布$F_i(i=1,2,\cdots,n)$，也不足以决定X的分布F，因为边缘分布中不含与其他边缘分布之间的关系这一信息。

对于离散型概率分布，以X_1为例，它的全部可能值为$a_{11},a_{12},\cdots$。例如，我们要求$P(X_1=a_{1k})$。我们有

$$P(X_1=a_{1k})=\sum_{j_2,\cdots,j_n}p(k,j_2,\cdots,j_n)\qquad(k=1,2,\cdots).$$

例. 设$X=(X_1,X_2,\cdots,X_n)$服从多项分布$M(N;p_1,p_2,\cdots,p_n)$，要求其边缘分布。

例如，考虑X_1，我们把事件A_1作为一方，那么$A_2+\cdots+A_n$作为另一方，即$\bar{A_1}$，则易知多项分布退化为了二项分布$B(N,p_1)$。

设$X=(X_1,X_2,\cdots,X_n)$有概率密度函数$f(x_1,x_2,\cdots,x_n)$，为求分量X_i的概率密度函数，只需把$f(x_1,x_2,\cdots,x_n)$中的x_i固定，然后对$x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n$在$-\infty$ 到∞之间做定积分。例如，X_1的密度函数为

$$f_1(x_1)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,x_2,\cdots,x_n)\mathbf{d}x_2\cdots\mathbf{d}x_n.$$

注. x_i对应的边缘分布函数即联合分布函数中的其他分量趋向正无穷后得到的函数。

注. 所谓的“边缘分布”其实就是通常的分布，并无任何特殊的含义，它只不过强调了：这个分布是由于X_i作为随机向量$(X_1,X_2,\cdots,X_n)$的一个分量，从后者的分布中派生出的分布而已。

与此相应，为了强调$(X_1,X_2,\cdots,X_n)$的分布是把$X_1,X_2,\cdots,X_n$作为一个有联系的整体来考虑的，有时我们把它称为$X_1,X_2,\cdots,X_n$的“联合分布”。

另外，边缘分布也可以不只是一维的。例如$X=(X_1,X_2,X_3)$的分布也决定了其任一部分，例如(X_1,X_3)的二维分布，这也称为边缘分布。

2.3 条件概率分布与随机变量的独立性

2.3.1 条件概率分布的概念

条件概率一般形式为：设有两个随机变量或向量X,Y，在给定了Y取某个或某些值的条件下，去求X的条件分布。

2.3.2 离散型随机变量的条件概率分布

设(X_1,X_2)为一个二维离散型随机向量，X_1的全部可能值为$a_1,a_2,\cdots$；X_2的全部可能值为$b_1,b_2,\cdots$；而(X_1,X_2)的联合概率分布为

$$p_{ij}=P(X_1=a_i,X_2=b_j)\quad(i,j=1,2,\cdots)$$

现在考虑X_1在给定$X_2 = b_j$的条件下的条件分布，即寻找条件概率$P(X_1 = a_i|X_2 = b_j), i = 1, 2, \cdots$。依条件概率的定义有，有

$$P(X_1 = a_1|X_2 = b_j) = P(X_1 = a_i, X_2 = b_j)/P(X_2 = b_j) = p_{ij}/\sum_k p_{kj}.$$

注. 分母上的因子实际上是为了归一化。

2.3.3 连续型随机变量的条件分布

设二维随机向量$X = (X_1, X_2)$有概率密度函数$f(x_1, x_2)$，我们先来考虑在限定$a \le x_2 \le b$的条件下X_1的条件分布。有

$$\begin{aligned} & P(X_1 \le x_1|a \le X_2 \le b) \\ = \ & P(X_1 \le x_1, a \le X_2 \le b)/P(a \le X_2 \le b) \\ = \ & \int_{-\infty}^{x_1} \mathbf{d}t_1 \int_a^b f(t_1, t_2)\mathbf{d}t_2 / \int_a^b f_2(t_2)\mathbf{d}t_2 \end{aligned}$$

其中$f_2(x) = \int_{-\infty}^{\infty} f(t, x)\mathbf{d}t$。这便是$X_1$的条件分布函数。对$x_1$求导，可得到条件密度函数为

$$f_1(x_1|a \le X_2 \le b) = \int_a^b f(x_1, t_2)\mathbf{d}t_2 / \int_a^b f_2(t_2)\mathbf{d}t_2$$

注. 离散型随机变量的概率函数对应这连续型随机变量的密度函数，都是描述在一个点处的信息。而分布函数描述连续的信息，对于离散型随机变量而言，我们通过对概率函数求和来得到分布函数，反方向则是差分；对于连续型随机变量而言，我们通过对密度函数积分来得到分布函数，反方向则是求导。对于离散型的随机变量而言，可能分析单一情况比较简单，即概率函数比分布函数更容易求得，一般性先求概率函数，然后在此基础上求得分布函数。但是对于连续型的随机变量而言，情况则是相反，我们一般先求分布函数，在求密度函数。在计算上，求导也比积分要来的简单。

令$a = x_2, b = x_2 + h, h \to 0$，我们可以得到

$$\begin{aligned} f_1(x_1|x_2) \ & = f_1(x_1|X_2 = x_2) \\ & = \lim_{h\to 0} f_1(x_1|x_2 \le X_2 \le x_2 + h) \\ & = \lim_{h\to 0} \tfrac{1}{h} \int_{x_2}^{x_2+h} f(x_1, t_2)\mathbf{d}t_2 / (\lim_{h\to 0} \tfrac{1}{h} \int_{x_2}^{x_2+h} f_2(t_2)\mathbf{d}t_2) \\ & = f(x_1, x_2)/f_2(x_2) \end{aligned}$$

运用高等概率论的知识，我们可以证明上式的成立无需连续性条件。

这个公式告诉我们两个随机变量X_1, X_2的联合概率密度，等于其中一个变量的概率密度乘以在给定这一个变量之下另一个变量的条件概率密度，对应于条件概率公式$P(AB) = P(B)P(A|B)$。同时，该公式可以推广到任意多个变量的场合：设有n维随机向量$(X_1, X_2, \cdots, X_n)$，其概率密度函数为$f(x_1, x_2, \cdots, x_n)$，则

$$f(x_1, x_2, \cdots, x_n) = g(x_1, \cdots, x_k)h(x_{k+1}, \cdots, x_n|x_1, \cdots, x_k),$$

其中g是$(X_1, X_2, \cdots, X_k)$的概率密度，而h是在给定$X_1 = x_1, \cdots, X_k = x_k$的条件下$X_{k+1}, \cdots, X_n$ 的条件概率密度。

例. 通过计算可以得知，正态变量的条件分布仍为正态，这是正态分布的一个重要性质。

设(X_1, X_2)服从二维正态分布$N(a, b, \sigma_1^2, \sigma_2^2, \rho)$。则

$$f_2(x_2|x_1) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} exp[-\frac{(x_2 - (b + \rho\sigma_2\sigma_1^{-1}(x_1 - a)))^2}{2(1-\rho^2)\sigma_2^2}]$$

可以看出该正态分布的位置系数当$\rho > 0$时与x_1 称正比，即随着x_1的增大，x_2取大值的可能性最大，此情况称为“正相关”，反之则称为“负相关”。

对$f(x_1, x_2) = f_2(x_2)f_1(x_1|x_2)$两边同时对$x_2$积分可得

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\mathbf{d}x_2 = \int_{-\infty}^{\infty} f_1(x_1|x_2)f_2(x_2)\mathbf{d}x_2$$

这告诉我们一个联合分布中的边缘分布可以是其条件分布以其他边缘分布为权的加权平均。可以视为连续版本的全概率公式。

2.3.4 随机变量的独立性

定义. 设n维随机向量$(X_1, X_2, \cdots, X_n)$的联合密度函数为$f(x_1, x_2, \cdots, x_n)$，而X_i的边缘密度函数为$f_i(x_i)\ (i = 1, 2, \cdots, n)$。如果

$$f(x_1, x_2, \cdots, x_n) = f_1(x_1) \cdots f_n(x_n),$$

就称随机变量$X_1, X_2, \cdots, X_n$相互独立，或简称独立。

注. 要证明随机变量不相互独立，只需要指出在某一点上上式不成立即可，但是该点必须要是联合密度函数的连续点。

定理 2.5. 如果连续变量$X_1, X_2, \cdots, X_n$独立，则对任何$a_i < b_i\,(i = 1, 2, \cdots, n)$，事件

$$A_1 = \{a_1 \le X_1 \le b_1\}, \cdots, A_n = \{a_n \le X_n \le b_n\},$$

也独立。

反之，若上述事件相互独立，则变量$X_1, X_2, \cdots, X_n$也独立。

定理 2.6. 若连续型随机向量$(X_1, X_2, \cdots, X_n)$的概率密度函数$f(x_1, x_2, \cdots, x_n)$可表为n个函数$g_1, g_2, \cdots, g_n$之积，其中g_i只依赖于x_i，即

$$f(x_1, x_2, \cdots, x_n) = g_1(x_1) \cdots g_n(x_n),$$

则$X_1, X_2, \cdots, X_n$相互独立，且X_i的边缘密度函数$f_i(x_i)$与$g_i(x_i)$只相差一个常数因子。

注. 与之前验证事件之间相互独立不同，这里只需要验证一个关系式成立即可，实际上要证的关系式是对一个区间上的恒等式，相当于验证了无数组关系式。

定理 2.7. 若$X_1, X_2, \cdots, X_n$相互独立，而

$$Y_1 = g_1(X_1, X_2, \cdots, X_m), \qquad Y_2 = g_2(X_{m+1}, X_{m+2}, \cdots, X_n),$$

则Y_1和Y_2独立。

定理 2.8. 设$X_1, X_2, \cdots, X_n$都是离散型随机变量。若对任何常数$a_1, a_2, \cdots, a_n$，都有

$$P(X_1 = a_1, \cdots, X_n = a_n) = P(X_a = a_1) \cdots P(X_n = a_n)$$

则称$X_1, X_2, \cdots, X_n$相互独立。

所有关于独立性的定理全部适用于离散型。唯一的变动时：凡是在这些定理中提到“密度函数”的地方，现在要改为“概率函数”。

定理 2.9. 随机变量$X_1, X_2, \cdots, X_n$独立的充分必要条件是“对于任意数集$S_1, S_2, \cdots, S_n$，有$P(X_1 \in S_1, X_2 \in S_2, \cdots, X_n \in S_n) = P(X_1 \in S_1)P(X_2 \in S_2) \cdots P(X_n \in S_n)$”

实际应用中，随机变量的独立性往往是根据实际背景判断出来的，然后再利用独立性定义中所赋予的性质和独立性相关的定理。

如果用指示变量X_i来代替事件A_i的话，可以证明若事件$A_1, A_2, \cdots, A_n$相互独立，则其指示变量$X_1, X_2, \cdots, X_n$亦独立，反之亦成立。可以用这一观点来简洁的证明关于事件的独立性的一系列定理。

2.4 随机变量的函数的概率分布

在理论和应用上，经常碰到这种情况：已知某个或某些随机变量$(X_1, X_2, \cdots, X_n)$的分布，现另有一些随机变量$Y_1, Y_2, \cdots, Y_m$，它们都是$X_1, X_2, \cdots, X_n$的函数：

$$Y_i = g_i(X_1, X_2, \cdots, X_n) \quad (i = 1, 2, \cdots, m)$$

要求$(Y_1, Y_2, \cdots, Y_m)$的概率分布。下面分别讨论离散型和连续型变量的处理办法。

2.4.1 离散型分布的情况

一般地，把$Y = g(X_1, X_2, \cdots, X_n)$可以取的不同值找出来，把与某个值相应的全部$(X_1, X_2, \cdots, X_n)$值的概率加起来，即得$Y$取这个值的概率。

2.4.2 连续型分布的情况

对于随机变量而言，设X有密度函数$f(x)$。设$Y = g(x)$，g时一个严格单调的函数，则Y的密度函数为

$$l(y) = f(h(y))|h'(y)|$$

其中h是g的反函数。

如果g并不是单调函数，我们一般先将Y的分布函数表示出来，然后再通过求导求得其密度函数。

对于二维随机连续型随机向量，假设(X_1, X_2)到(Y_1, Y_2)是一一变换

$$Y_1 = g_1(X_1, X_2), \quad Y_2 = g_2(X_1, X_2),$$

则有逆变换

$$X_1 = h_1(Y_1, Y_2), \quad X_2 = h_2(Y_1, Y_2).$$

又假设g_1, g_2都有一阶连续偏导数，则h_1, h_2也有一阶连续偏导数，且

$$\frac{\partial(h_1, h_2)}{\partial(y_1, y_2)} \neq 0$$

通过积分换元，我们可以得到(Y_1, Y_2)的密度函数为

$$l(y_1, y_2) = f(h_1(y_1, y_2)), h_2(y_1, y_2))|\frac{\partial(h_1, h_2)}{\partial(y_1, y_2)}|\mathbf{d}y_1\mathbf{d}y_2$$

如果所要求的只是一个函数

$$Y_1 = g_1(X_1, X_2)$$

的密度函数。那么也是先求其分布，再通过求导求得其密度函数。或者可以配上另一个函数$Y_2 = g_2(X_1, X_2)$，使得(X_1, X_2)到(Y_1, Y_2)成一一对应变换；然后求出其联合密度函数$l(y_1, y_2)$，最后Y_1的密度函数按照公式$\int_{-\infty}^{+\infty} l(y_1, y_2)\mathbf{d}y_2$给出。

上面的方法可以推广至n维随机向量的情形，沿用上面的记号，则$(Y_1, Y_2, \cdots, Y_n)$的密度函数为

$$l(y_1, y_2, \cdots, y_n) = f(h_1(y_1, y_2, \cdots, y_n), \cdots, h_n(y_1, y_2, \cdots, y_n))|\frac{\partial(h_1, h_2, \cdots, h_n)}{\partial(y_1, y_2, \cdots, y_n)}|$$

2.4.3 随机变量和的密度函数

设(X_1, X_2)的联合密度函数为$f(x_1, x_2)$，可以用上一节提到的两种方法来求得$Y = X_1 + X_2$ 的密度函数。第一种方法的不足之处在于需要在积分号内求导，在理论上有较高限制。用第二种方式则只需要配上变量$Z = X_1$即可。通过两种方式均可以求出Y的密度函数为

$$l(y) = \int_{-\infty}^{+\infty} f(x, y - x)\mathbf{d}x = \int_{-\infty}^{+\infty} f(y - x, x)\mathbf{d}x$$

当X_1, X_2独立时，有

$$l(y) = \int_{-\infty}^{+\infty} f_1(x)f_2(y - x)\mathbf{x} = \int_{-\infty}^{+\infty} f_1(y - x)f_2(x)\mathbf{d}x.$$

注. 上述公式称为卷积公式。

通过计算可以得到，如果X_1, X_2独立，且分别服从$N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$，则$Y = X_1 + X_2$服从$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$。利用更高阶的知识可以证明其逆命题也成立，即：如果Y服从正态分布，而Y表成两个独立随机变量X_1, X_2之和，则X_1, X_2必都服从正态分布。这一性质称为正态分布的“再生性”。

反复利用这一结论，我们可以将其推广至n维。

更一般地，设$X_1, X_2, \cdots, X_n$相互独立，且$X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \cdots, n, k_1, k_2, \cdots, k_n$是一些不全为0的常数，$c$是常数，则

$$\sum_{i=1}^{n} k_i X_i + c \sim N(\sum_{i=1}^{n} k_i\mu_i + c, \sum_{i=1}^{n} k_i^2\sigma_i^2).$$

2.4.4 随机函数商的密度函数

设(X_1, X_2)有联合密度函数$f(x_1, x_2)$，而$Y = X_1/X_2$，要求Y的密度函数。为了简化讨论，只考虑X_1取正值的情况。与上一节类似，我们可以通过两种方式得到其密度函数，其中引入变量的方式能够避免理论上对于积分号内求导的要求。Y的密度函数为

$$l(y) = \int_0^\infty x_1 f(x_1, x_1 y)\mathbf{d}x_1.$$

若X_1, X_2独立，则$f(x_1, x_2) = f_1(x_1)f_2(x_2)$，上式可以变为

$$l(y) = \int_0^\infty x_1 f_1(x_1) f_2(x_1 y)\mathbf{d}x_1.$$

3 随机变量的数字特征

3.1 数学期望（均值）与中位数

3.1.1 数学期望的定义

定义. 设随机变量X只取有限个可能值$a_1, a_2, \cdots, a_m$，其概率分布为$P(X = a_i) = p_i\,(i = 1, 2, \cdots, m)$。则$X$的数学期望，记为$E(X)$或$EX$，定义为

$$E(X) = a_1 p_1 + a_2 p_2 + \cdots + a_m p_m.$$

定义. 如果X为离散型变量，且有无数个可能取值$a_1, a_2, \cdots$，而概率分布为$P(X = a_i) = p_i\,(i = 1, 2, \cdots)$，如果

$$E(X) = \sum_{i=1}^{\infty} |a_i| p_i < \infty$$

则称

$$\sum_{i=1}^{\infty} a_i p_i$$

为X的数学期望。

注. 这是为了保证期望不随着求和的次序改变而改变。

定义. 设X有概率密度函数$f(x)$.如果

$$\int_{-\infty}^{+\infty} |x| f(x)\mathbf{d}x < \infty,$$

则称

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

为X的数学期望。

下面给出一些常用分布对于的期望：

服从分布	期望
二项分布$B(n,p)$	np
泊松分布$P(\lambda)$	λ
负二项分布	r(1-p)/p
几何分布$Ge(p)$	1/p
超几何分布$h(N,M,n)$	$n\frac{M}{N}$
均匀分布$R(a,b)$	(a+b)/2
指数分布	λ^{-1}
正态分布$N(\mu,\sigma^2)$	μ
卡方分布χ_n^2	n
t分布(自由度大于1)	0
自由度为(m,n)的F分布,n>2	n/(n-2)

注. 对于二项分布，其期望可以理解为n次伯努利实验的期望之和；对于泊松分布$P(\lambda)$，λ即在所指定的时间段中发生某一事件的平均次数。

当我们知道了一个随机变量的分布或者密度后，便能够得到其数学期望。但这不是必要的，实际上，在很多情况下，一个随机变量的期望比其分布或密度更容易估计。

3.1.2 数学期望的性质

定理 3.1. 若随机变量$X_1,X_2,\cdots,X_n$的期望均存在，则

$$E(\sum_i X)=\sum_i E(X_i)$$

定理 3.2. 若随机变量$X_1,X_2,\cdots,X_n$的期望均存在，且相互独立，则

$$E(\prod_i X_i)=\prod_i E(X_i)$$

注. 如果一个离散型随机变量的期望较难计算，可能是因为其概率分布较难计算，那么可以将其拆分为多个分布较容易计算的随机变量，然后再利用上述定理，这种方法常用于有实际背景的题目。

定理 3.3. 设随机变量X为离散型，有分布$P(X=a_i)=p_i(i=1,2,\cdots)$；或者为连续型，有概率密度函数$f(x)$。则

$$E(g(x))=\sum_i g(a_i)p_i \quad \text{当}\sum_i|g(a_i)|p_i<\infty\text{时}$$

或

$$E(g(X))=\int_{-\infty}^{+\infty}g(x)f(x)\mathbf{d}x \quad \text{当}\int_{-\infty}^{+\infty}|g(x)|f(x)\mathbf{d}x<\infty\text{时}$$

这一定理告诉我们，为了计算X的某一函数$g(X)$的期望，并不需要先求出$g(X)$的密度函数，而可以就从X的分布出发。

从中可以推出

1. $E(cX)=cE(X)$
2. $E(kX+c)=kE(x)+c$
3. $E(X^2)=0\Leftrightarrow P(X=0)=1$

定理 3.4 (施瓦兹不等式)**.**

$$E(X^2)E(Y^2)\geq(E(XY))^2.$$

等号成立的条件为X,Y有线性关系，即存在常数c，使得$X=cY$或$Y=cX$。

注. 特别地，对于正随机变量X_1，取$X=\sqrt{X_1},Y=1/\sqrt{X_1}$，我们有

$$E(\frac{1}{X_1})\geq\frac{1}{E(X_1)}$$

等号仅当X_1取值为常数时才成立。

3.1.3 条件数学期望（条件均值）

随机变量Y的条件数学期望就是它在给定的某种附加条件下的数学期望。

如果知道了(X,Y)的联合密度，则$E(Y|x)$的定义就可以具体化为：先定出在给定$X=x$之下Y的条件密度函数$f(y|x)$，然后算出

$$E(Y|x)=\int_{-\infty}^{+\infty}yf(y|x)\mathbf{d}y.$$

如果说条件分布是变量X与Y的相依关系在概率上的完全刻画，那么，条件期望则在一个很重要的方面刻画了两者的关系，它反映了随着X取值x的变化Y的平均变化的情况如何。在统计学上，常把条件期望$E(Y|x)$作为x的函数，称为Y对X的“回归函数”。

回想全概率公式

$$P(A) = \sum_i P(B_i)P(A|B_i),$$

它可以理解为通过事件A的条件概率$P(A|B_i)$去计算其（无条件）概率$P(A)$。更准确地说，$P(A)$就是条件概率$P(A|B_i)$的某种加权平均，权即为事件B_i的概率。以此类推，变量Y的（无条件）期望应等于其条件期望$E(Y|x)$对x取加权平均，x的权与变量X在x点的概率密度$f_1(x)$成比例，即

$$E(Y) = \int_{-\infty}^{+\infty} E(Y|x)f_1(x)\mathbf{d}x.$$

如果记$g(x) = E(Y|x)$，它是x的函数，则上式变为

$$E(Y) = \int_{-\infty}^{+\infty} g(x)f_1(x)\mathbf{d}x.$$

根据定理3.3，上式右边便是$E(g(X)) = E[E(Y|x)]|_{x=X}$，可简写为$E[E(Y|X)]$。可得

$$E(Y) = E[E(Y|X)].$$

可以理解为随机变量Y的期望是其条件期望的期望，即可以把一个很大的范围内的求平均分成两步走，先将这个很大的区域划分为多个较小的区域分别求平均，然后再对这些平均值求平均。

上式可以推广至更一般的情形，例如，X不必为一维的，如果X为n为随机向量$(X_1, X_2, \cdots, X_n)$，有概率密度$f(x_1, x_2, \cdots, x_n)$，则上式有形式

$$E(Y) = \int_{-\infty}^{\infty} E(Y|x_1, x_2, \cdots, x_n)f(x_1, x_2, \cdots, x_n)\mathbf{d}x_1 \cdots \mathbf{d}x_n$$

又若X为一维离散型变量，有分布

$$P(X = a_i) = p_i \quad (i = 1, 2, \cdots)$$

则有形式

$$E(Y) = \sum_{i=1}^{\infty} p_i E(Y|a_i)$$

3.1.4 中位数

定义. 设连续型随机变量X的分布函数为$F(x)$，则满足条件

$$P(X \leq x_p) = F(x_p) = p$$

的数x_p称为X或F的p分位数，也称下（尾）p分位数。特别地，当$p = 0.5$时，称$x_{0.5}$为X的中位数。

中位数和期望相比较，其受极端值的影响较小，且理论上总是存在的。

注. 但是这样定义出来的中位数不一定是唯一的。对于同一个p，我们常取x_p的最小值。

3.2 方差与矩

3.2.1 方差和标准差

为了刻画随机变量较其均值的偏差，我们要为其寻找一个数字特征。$E(X-E(X))$显然不符合要求，因为它恒为零，而绝对平均差$E(|X-E(X)|)$中的绝对值又不够光滑，因此我们选用方差$E(X - E(X))^2$。

注. 方差，即差的方。

定义. 设X为随机变量，分布为F，则

$$Var(X) = E(X - EX)^2$$

称为X（或分布F）的方差，其算数平方根$\sqrt{Var(X)}$称为X（或分布F）的标准差。

定理 3.5.

$$Var(X) = E(X^2) - (EX)^2.$$

这个形式在计算上往往更加方便。

注. 一个随机变量的方差如果存在，则其期望一定存在，反之不成立。这是因为方差存在等价于

$$\int_{-\infty}^{\infty} x^2 f(x) \mathbf{d}x < \infty$$

又因为

$$\int_{-\infty}^{\infty} f(x) \mathbf{d}x = 1$$

我们有

$$\int_{-\infty}^{\infty}|x|f(x)\mathbf{d}x<\int_{-\infty}^{\infty}(x^2+1)f(x)\mathbf{d}x<\infty$$

方差有以下性质：

定理 3.6. 1. 常数的方差为0

2. 若c为常数，则$Var(X+c)=Var(X)$

3. 若c为常数，则$Var(cX)=c^2Var(X)$.

注. 若$k=\frac{1}{\sqrt{Var(X)}}=\frac{1}{\sigma(X)}$，则$Var(kX)=Var(\frac{X}{\sigma(X)})=1$。对于随机变量$X$，我们称

$$X^*=\frac{X-EX}{\sigma(X)}$$

为X的标准化，有

$$EX^*=0\quad Var(X^*)=1$$

定理 3.7. 独立随机变量之和的方差等于各变量的方差之和，即

$$Var(\sum_i X_i)=\sum_i Var(X_i)$$

下面给出一些常用分布的方差：

服从分布	方差
二项分布$B(n,p)$	np(1-p)
泊松分布$P(\lambda)$	λ
几何分布$Ge(p)$	$(1-p)/p^2$
超几何分布$h(N,M,n)$	$nM(N-M)(N-n)/N^2(N-1)$
均匀分布$R(a,b)$	$(b-a)^2/12$
指数分布	λ^{-2}
正态分布$N(\mu,\sigma^2)$	σ^2
卡方分布χ_n^2	2n
t分布(自由度大于1)	n/(n-2)
自由度为(m,n)的F分布,n>4	$2n^2(m+n-2)/[m(n-2)^2(n-4)]$

注. 二项分布的方差可以视为n个相互独立的伯努利实验的方差之和

3.2.2 矩

定义. 设X为随机变量，c为常数，k为正整数。则量$E[(X-c)^k]$称为X关于c点的k阶矩。

特别地，当$c=0$时，$\alpha_k=E(X^k)$称为X的k阶原点矩；当$c=E(X)$时，$\mu_k=E[(X-EX)^k]$称为X的k阶中心距。

k	α_k	μ_k
1	均值、期望	0
2		方差
3		$\mu_3/\mu_2^{3/2}$为偏度系数
4		μ_4/μ_2^2为峰度系数

注. 偏度系数和峰度系数分别描述了分布与正态分布之间的偏离情况和分布密度在均值附近的陡峭程度。由于两者在定义时都进行了齐次化，两个分布在比较峰度系数时，要将它们的方差调整至1后再进行比较。

例. 设有随机变量$X\sim N(0,1)$，则

$$E(X^k)=\begin{cases}(k-1)!!, & k=2,4,6,\cdots\\ 0, & k=1,3,5,\cdots\end{cases}$$

3.3 协方差与相关系数

3.3.1 协方差

定义. 称$E[(X-EX)(Y-EY)]$为X,Y的协方差，并记为 $\mathrm{Cov}(X,Y)$.

协，即“协同”。X的方差的定义即$(X-EX)$与$(X-EX)$的乘积的期望。如今将其中一者替换为$(Y-EY)$，我们便可以得到X,Y之间的协方差。

易得协方差具有以下性质：对于随机变量X,Y:

1. $\mathrm{Cov}(X,Y)$=$\mathrm{Cov}(Y,X)$
2. 对于任意常数c_1,c_2,c_3,c_4，有 $\mathrm{Cov}(c_1X+c_2,c_3Y+c_4)=c_1c_3\mathrm{Cov}(X,Y)$
3. $\mathrm{Cov}(X,Y)$=$E(XY)$-$EXEY$
4. $Var(X\pm Y)=Var(X)+Var(Y)\pm 2Cov(X,Y)$

5. 更一般地，设a, b, c是任意常数，则$Var(aX \pm bY + c) = a^2Var(X) + b^2Var(Y) \pm 2abCov(X, Y)$

6. 若C为任意常数，则$Cov(X, C) = 0$

7. $Cov(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} Cov(X_i, Y_j)$

定理 3.8. 若X, Y独立，则 $\mathrm{Cov}(X, Y) = 0$.

定理 3.9.

$$[Cov(X, Y)]^2 \leq \sigma_1^2 \sigma_2^2,$$

等号当且仅当X, Y之间有严格线性关系，即存在常数a, b使得$Y = a + bX$时成立。

3.3.2 相关系数

定义. 称 $\mathrm{Cov}(X, Y)/(\sigma_1\sigma_2)$为$X, Y$的相关系数，并记为 $\mathrm{Corr}(X, Y)$.

形式上可以理解为“标准尺度下的协方差”。协方差原先依赖于X, Y的度量单位，选择适当单位使得X, Y的方差都为1，则协方差就是相关系数。这样能更好地反映出X, Y之间的关系，不受所用单位的影响。

定理 3.10. 若X, Y独立，则 $\mathrm{Corr}(X, Y)$=0.

定理 3.11. $|Corr(X, Y)| \leq 1$，等号当且仅当X和Y有严格线性关系时达到。

注. 当 $\mathrm{Corr}(X, Y) = 0$时，我们称X, Y“不相关”。当两个随机变量独立时，它们不相关，反之则不成立，单位圆上的二维均匀分布就是一个反例。当X, Y不相关时，$Var(X \pm Y) = Var(X) + Var(Y), E(XY) = E(X)E(Y)$.

注. 相关系数也常称为线性相关系数。为正线性相关时，相关系数为1；为负线性相关时，相关系数为-1。

对于二维正态分布$N(a, b, \sigma_1^2, \sigma_2^2, \rho)$，可以证明 $\mathrm{Corr}(X, Y) = \rho$. 即对于二维正态分布而言，它的两个边缘分布独立等价于不相关。

3.4 大数定理和中心极限定理

3.4.1 大数定理

定理 3.12. 设$X_1, X_2, \cdots, X_n, \cdots$是独立同分布的随机变量，记它们的公共均值为$a$。又设它们的方差存在并记为$\sigma^2$。则对任意给定的$\epsilon > 0$，有

$$\lim_{n\to\infty} P(|\bar{X_n} - a| \geq \epsilon) = 0.$$

其中$p_n = (X_1 + X_2 + \cdots + X_n)/n = \bar{X_n}$。

注. 当X_i是伯努利试验对应的随机变量时，p_n即指定事件发生的频率。上述定理在该背景下的含义即概率的统计定义。即“频率收敛于概率”：

$$\lim_{n\to\infty} P(|p_n - p| \geq \epsilon) = 0.$$

注. 上式中的极限所表示的收敛性在概率论中叫做“$\bar{X_n}$依概率收敛于a”.

注. a是公共均值，即公共期望，是理论上的。而$\bar{X}$则是依概率的。

定理 3.13 (马尔可夫不等式)**.** 若Y为只取负值的随机变量，则对任给常数$\epsilon > 0$，有

$$P(Y \geq \epsilon) \leq E(Y)/\epsilon.$$

证明. 设Y为非负的连续型变量，密度函数为$f(x)$，则

$$E(Y) = \int_0^\infty yf(y)\mathbf{d}y \geq \int_\epsilon^\infty yf(y)\mathbf{d}y \geq \epsilon \int_\epsilon^\infty f(y)\mathbf{d}y = \epsilon P(Y \geq \epsilon)$$

□

定理 3.14 (切比雪夫不等式)**.** 若 Var(Y)存在，则

$$P(|Y - EY| \geq \epsilon) \leq Var(Y)/\epsilon^2.$$

大数定理有许多版本：

定理 3.15 (马尔可夫大数定理)**.** 如果随机变量序列$X_1, X_2, \cdots, X_n, \cdots$满足

$$\lim_{n\to\infty} Var(\sum_{i=1}^{n} X_i)/n^2 = 0,$$

则随机变量序列$X_1, X_2, \cdots, X_n, \cdots$服从大数定律.

定理 3.16 (辛钦大数定律)**.** 对于独立同分布的随机变量序列$X_1, X_2, \cdots, X_n, \cdots$，若$E(X_1)$存在，则随机变量序列$X_1, X_2, \cdots, X_n, \cdots$服从大数定律

最后进行一下总结，大数定律即一系列有关$\overline{X} - E(\overline{X})$依概率收敛于0的定律。按照给定的条件来区分的话，当给定的条件最为宽松时，即所给的随机变量序列中的随机变量没有任何附加条件时，我们需要证明

$$\lim_{n\to\infty} Var(\sum_{i=1}^{n} X_i)/n^2 = 0.$$

进一步，当这些随机变量独立同分布时，由辛钦大数定理可知，只要$E(X_1)$存在，便可以保证大数定理成立，且$\overline{X}$依概率收敛于$E(X_1)$.

3.4.2 中心极限定理

定理 3.17. 设$X_1, X_2, \cdots, X_n, \cdots$为独立同分布的随机变量，$E(X_i) = a$, $Var(X_i) = \sigma^2 (0 < \sigma^2 < \infty)$.则对任何实数$x$，有

$$\lim_{n\to\infty} P(\frac{1}{\sqrt{n}\sigma}(X_1 + \cdots + X_n - na) \le x) = \Phi(x).$$

注. $X_1 + \cdots + X_n$有均值na，方差$n\sigma^2$，故

$$(X_1 + \cdots + X_n - na)/(\sqrt{n}\sigma)$$

就是$X_1 + \cdots + X_n$的标准化。

注. 中心极限定理本质上告诉我们的是，对于独立同分布的随机变量序列$X_1, X_2, \cdots, X_n, \cdots$，由此生成的随机变量$\sum_{i=1}^{n} X_i$的分布函数当$n$足够大时近似等于正态分布，这个正态分布的参数，即期望和方差应该是和$\sum_{i=1}^{n} X_i$一致的。为了统一，上述定理将$\sum_{i=1}^{n} X_i$先进行规范化，这样就近似等于标准正态分布。

上述中心极限定理的一个特例如下。

定理 3.18. 设$X_1, X_2, \cdots, X_n, \cdots$独立同分布，$X_i$的分布是

$$P(X_i = 1) = p, \qquad P(X_i = 0) = 1 - p \qquad (0 < p < 1).$$

则对任何实数x，有

$$\lim_{n\to\infty} P(\frac{1}{\sqrt{np(1-p)}}(X_1 + \cdots + X_n - np) \le x) = \Phi(x).$$

上式即用正态分布去逼近二项分布，条件是p固定，n相当大。当n相当大，$np = \lambda$不太大时，则用泊松分布来逼近二项分布。

4 统计量及其分布

4.1 基本概念

4.1.1 什么是数理统计学？

数理统计学中的两大问题分别是：统计分析和统计推断。举个例子：假设我们有一批寿命服从指数分布的元件，我们想要通过抽样调查其中若干个元件来推出其理论寿命，这便是统计分析的目标。但是，这样得出的结果总是有一定的误差，那么如何在通过这种带有误差的结论来进一步进行决策，这便是统计推断的目标。

4.1.2 总体与个体

总体是值与所研究的问题有关的对象（个体）的全体所构成的集合。例如在上例中，所有元件便是问题的总体，而每一个元件便是个体，所有的个体便组成了总体。总体随着我们所研究的范围而变，当我们只关心一些指标时，例如元件的寿命，此时个体就变为了单个元件的寿命，总体则变为了所有元件的寿命。同时，为了赋予总体一定的含义，我们要赋予总体一定的概率分布，并称这种总体为“统计总体”。当总体服从正态分布时，我们便称其为正态总体，以此类推。

注. 称一个重量为a的物品的重量，则我们得到的总体是我们所测得的重量，而非”a”。

数据分为连续型变量数据和分类变量数据。而分类变量数据又可以进一步分为无序分类变量数据和有序分类变量数据。如果指标是离散型的，我们则将总体的分布律称为总体分布律。如果为连续型，则称为总体密度函数。

如果统计总体所服从的分布有若干个参数——例如正态分布中的位置参数和形状参数——则它实际上是一个概率分布族中的一员。

4.1.3 样本

样本即按一定的规定从总体中抽出的一部分个体，也就是每个个体有同等的概率被抽出。如果我们抽出若干个数据$X_1, X_2, \cdots, X_n$，则称n为样本大小，或样本容量。因为第i个抽到的数据是随机的，所以其指标值X_i实际上是个随机变量。一般地，我们假设总体中有无穷多个或很多个个体，此

时，可以将$X_1, X_2, \cdots, X_n$视为独立同分布的随机变量，其共同分布为总体分布。

4.1.4 经验分布函数

为了从样本中构建与总体分布函数相近的分布函数，我们令

$$F_n(x) = X_1, X_2, \cdots, X_n\text{中不大于}x\text{的个数}/n$$

我们可以利用其数字特征去推测总体分布函数对应的数字特征。

定理 4.1. 经验分布函数各点处依概率收敛于总体分布函数。

4.1.5 统计量

完全由样本决定的量叫做统计量。它只依赖于样本，而不依赖与总体分布的未知参数。统计量可以视为是对样本的一种“加工”。统计量的构造都是有目的的，例如

$$S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$$

便是“样本方差”，用于估计样本的散布程度。

另一类统计量叫做样本据，分为样本原点矩和样本中心距，设$X_1, X_2, \cdots, X_n$为样本，n为正整数，则

$$a_k = (X_1^k + X_2^k + \cdots + X_n^k)/n$$

称为k阶原点样本距。$a_1 = \overline{X}$便是样本均值。而

$$m_k = \sum_{i=0}^{n}(X_i - \overline{X})^k/n$$

称为k阶样本中心距。

注. 分母上的n实际上代表的是在抽到的n个样本中再次抽取时是等概率的，实际上就是计算期望是时的“p”。

之前所定义的样本矩实际上就是经验分布的矩。

定理 4.2. 设$X_1, X_2, \cdots, X_n$是取自总体X的样本，则

1. $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$;

2. $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i - c)^2 - n(\overline{X} - c)^2$，其中$c$为任意实数。特别地，当$c = 0$时，有$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2$；

3. 对总体X而言，如果$E(X^2) < \infty$，则$E(\overline{X}) = E(X), Var(\overline{X}) = \frac{Var(X)}{n}, E(S^2) = Var(X)$。

由样本矩组合而成的函数也可以作为统计量：

定义. 1. 样本变异系数：$\delta_X = \frac{S}{|\overline{X}|}$

2. 样本偏度系数：$\hat{\beta}_S = \frac{B_3}{B_2^{3/2}}$

3. 样本峰度系数：$\hat{\beta}_k = \frac{B_4}{B_2^2} - 3$

定义. 如果总体指标是二维随机变量(X, Y)，抽样得到的样本为$(X_1, Y_1), \cdots, (X_n, Y_n)$，则可以定义样本$(k, l)$阶混合矩为

$$C_{k,l} = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^k (Y_i - \overline{Y})^l, k, l = 1, 2, \cdots$$

特别地，$C_{1,1}$称为样本协方差，可以用来估计总体的协方差$Cov(X, Y)$。

类似地，样本相关系数定义为

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

4.1.6 次序统计量

设$x_1, x_2, \cdots, x_n$为一组样本观测值，将其按照从小到大进行重排，得到$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$，对应的随机变量$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$称为次序统计量，其中$X_{(1)}, X_{(n)}$分别称为最小次序统计量和最大次序统计量。样本极差即$R_X = X_{(n)} - X_{(1)}$。

注. 虽然X_i是随机变量，取值是依概率的，但是当我们获得样本值后，其实都确定下来了，所以可以在此基础上将其加工为次序统计量。

利用次序统计量可以定义样本中位数：

$$M_{0.5} = (X_{(\lceil n/2 \rceil)} + X_{(\lfloor n/2+1 \rfloor)})/2$$

样本的p分为数则定义为

$$M_p = (X_{(\lceil np \rceil)} + X_{(\lfloor np+1 \rfloor)})/2$$

特别地$M_{0.25}, M_{0.75}$分别称为第一四分位数（或下四分位数）和第三四分位数（或上四分位数）。称

$$IQR = M_{0.75} - M_{0.25}$$

样本最小次序统计量值、第一四分位数、样本中位数、第三四分位数和最大统计量值统称为五数总括。

4.1.7 可视化

直方图的频率密度定义为

$$p_k = \frac{n_k}{n \cdot \Delta_k}, k = 1, 2, \cdots, m,$$

即第k个区间的频数n_k除以总样本数n与区间长度Δ_k，表示频率相对于区间长度的密度。

茎叶图，即将数据中的最高位相同的数据集中在一起，只列出其公共的最高位和各自的其他位的一种表示方式。

根据五数总括，我们可以画出对应的箱线图。盒的上下底分别对应着上下四分位数，中间的线对应着中位数。而顶部和底部延申出去的线则代表着最大值和最小值。

4.1.8 充分统计量

我们构造统计量的目的便是把样本观测值通过加工，以便于我们得到有关总体分布的未知参数信息。如果一个统计量可以完全确定总体分布的未知参数，那么我们就称其为充分统计量。也就是说，当我们得到了充分统计量的值后，便可以得到不带未知参数的总体分布。

当给定统计量$T(X_1, X_2, \cdots, X_n) = t$时，$(X_1, X_2, \cdots, X_n)$的联合条件分布函数为

$$\begin{aligned} & F_{(X_1, X_2, \cdots, X_n | T(X_1, X_2, \cdots, X_n))}(x_1, x_2, \cdots, x_n | t) \\ = \quad & P(X_1 \leq x_1, \cdots, X_n \leq x_n | T(X_1, X_2, \cdots, X_n) = t) \end{aligned}$$

当总体分布为连续型随机变量的分布时，假设$(X_1, X_2, \cdots, X_n)$的来拟合密度函数为$f(x_1, x_2, \cdots, x_n; \theta)$，统计量$T(X_1, X_2, \cdots, X_n)$的密度函数为$g(t; \theta)$，则当给定统计量$T(X_1, X_2, \cdots, X_n) = t$时，$(X_1, X_2, \cdots, X_n)$的联合条件密度函数为

$$f_{(X_1, X_2, \cdots, X_n | T(X_1, X_2, \cdots, X_n))}(x_1, x_2, \cdots, x_n | t) = \frac{f(x_1, x_2, \cdots, x_n; \theta) I_{\{T(x_1, x_2, \cdots, x_n) = t\}}}{g(t; \theta)}$$

当总体分布为离散型随机变量的分布时，$(X_1, X_2, \cdots, X_n)$的联合条件分布律为

$$\begin{aligned} & f_{(X_1,X_2,\cdots,X_n|T(X_1,X_2,\cdots,X_n))}(x_1, x_2, \cdots, x_n|t) \\ = \quad & P(X_1 = x_1, \cdots, X_n = x_n|T(X_1, X_2, \cdots, X_n) = t) \end{aligned}$$

定理 4.3 (因子分解定理). 设$X_1, X_2, \cdots, X_n$是取自总体X的样本，总体分布函数为$F(x;\theta)$，其中θ为未知参数，则统计量$T(X_1, X_2, \cdots, X_n)$是θ的充分统计量的充要条件为：$(X_1, X_2, \cdots, X_n)$的联合密度函数（或联合分布律）可以写成如下形式

$$f(x_1, x_2, \cdots, x_n; \theta) = g(T(x_1, x_2, \cdots, x_n), \theta)h(x_1, x_2, \cdots, x_n),$$

其中$g(T(x_1, x_2, \cdots, x_n), \theta)$是关于统计量$T(X_1, X_2, \cdots, X_n)$和$\theta$的函数，$h(x_1, x_2, \cdots, x_n)$是与$\theta$无关的函数。

4.1.9 相合统计量

定义. 如果统计量$T = T(X_1, X_2, \cdots, X_n)$依概率收敛于参数$\theta$，称统计量$T = T(X_1, X_2, \cdots, X_n)$为参数$\theta$的相合统计量，或称该统计量具有相合性。

一个统计量只有具有相合性了以后，才能保证它是一个合格的统计量，因为相合性确保了当我们获得的样本数越多时，对于参数的估计更准确的概率更大。

注. 经验分布函数就是总体分布函数的相合统计量。

定理 4.4. 如果统计量$T = T(X_1, X_2, \cdots, X_n)$满足：

$$\lim_{n\to\infty} E[T(X_1, X_2, \cdots, X_n)] = \theta, \qquad \lim_{n\to\infty} Var(T(X_1, X_2, \cdots, X_n)) = 0,$$

则统计量$T = T(X_1, X_2, \cdots, X_n)$是参数$\theta$的相合统计量。

注. 此处和大数定理的区别在于：大数定理中涉及到的是随机变量列，且最终的结论是关系到随机变量列中的所有随机变量的，也就是它们的均值。而统计量中n只是代表着样本个数，，所给的条件的含义是当样本个数达到一定量后，估计是准确且稳定的，其结论是当样本个数达到一定量后，所获得的估计近似于实际的值的概率可以任意大。

4.1.10 统计学三大分布

定义. 若$X_1, X_2, \cdots, X_n$相互独立，都服从正态分布$N(0,1)$，则$Y = X_1^2 + \cdots + X_n^2$服从自由度为$n$的卡方分布$\chi_n^2$，其概率密度函数为：

$$k_n(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})2^{n/2}} e^{-x/2} x^{(n-2)/2}, & \text{当} x > 0 \text{时} \\ 0, & \text{当} x \leq 0 \text{时} \end{cases}$$

是概率密度函数，称为“自由度为n的皮尔逊卡方密度”（相应的分布称为卡方分布），常记为χ_n^2。

由定义易知，$\chi^2(n)$的期望和方差分别为n和$2n$。

注.

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \mathbf{d}t \quad (x > 0)$$

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} \mathbf{d}t \quad (x > 0, y > 0)$$

注. n实际上不只局限于正整数。

注. “$X_1, X_2, \cdots, X_n$相互独立，都服从正态分布$N(0,1)$”可以称为“$X_1, X_2, \cdots, X_n$相互独立同分布于$N(0,1)$”，也可以称“$X_1, X_2, \cdots, X_n$有共同分布$N(0,1)$”。可以简记为$X_1, X_2, \cdots, X_n \; iid., \sim N(0,1)$。

注. 卡方分布有如下重要性质：

1. 设X_1, X_2独立，$X_1 \sim \chi_m^2, X_2 \sim \chi_n^2$，则$X_1 + X_2 \sim \chi_{m+n}^2$.
2. 若$X_1, X_2, \cdots, X_n$独立，且都服从指数分布，则

$$X = 2\lambda(X_1 + X_2 + \cdots + X_n) \sim \chi_{2n}^2.$$

特别地，令$\lambda = \frac{1}{2}, n = 1$，则$E(1/2) = \chi^2(2)$

定义. 若随机变量X与Y相互独立，且$X \sim N(0,1), Y \sim \chi^2(n)$，则称随机变量$T = \frac{X}{\sqrt{Y/n}}$服从自由度为$n$的$t$分布，记为$T \sim t(n)$。其密度函数为：

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + \frac{x^2}{n})^{-\frac{n+1}{2}}, -\infty < x < +\infty$$

易知，t分布和正态分布一样具有对称性。

注. 当$X, Y \sim N(0,1)$，且X, Y相互独立时，$X/|Y| = X/\sqrt{Y^2/1} \sim t(1)$。

定义. 若随机变量U和V独立，且$U \sim \chi^2(m), V \sim \chi^2(n)$，则称随机变量$F = \frac{U/m}{V/n}$服从自由度为$(m,n)$的$F$分布，记为$F \sim F(m,n)$。其概率密度函数为

$$f_{mn}(y) = m^{m/2}n^{n/2}\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}y^{m/2-1}(my+n)^{-(m+n)/2} \qquad (y>0)$$

当$y \leq 0$时，$f_{mn}(y) = 0$。

由定义可知，当$F \sim F(m,n)$时，$\frac{1}{F} \sim F(n,m)$。

注. $P(F(m,n) < u) = P(F(n,m) > 1/u) = 1 - P(F(n,m) < 1/u)$，所以$F_p(m,n) = \frac{1}{F_{1-p}(n,m)}$

4.2 抽样分布

根据总体分布我们可以很容易的得出样本的密度函数，但是统计量的分布较难得到。

4.2.1 正态总体下的抽样分布

对于正态总体，其统计量的分布相对较容易得到：

定理 4.5. 设$X_1, X_2, \cdots, X_n$是取自正态总体$N(\mu, \sigma^2)$的样本，样本均值为$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$，样本方差为$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$，则有

1. $\overline{X} \sim N(\mu, \sigma^2/n)$，即$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1)$
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
3. $\overline{X}$与S^2相互独立。

定理 4.6. 设$X_1, X_2, \cdots, X_n$为来自正态总体$N(\mu, \sigma^2)$的样本，则有

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

定理 4.7. 设$X_1, X_2, \cdots, X_m$为取自总体$X \sim N(\mu_1, \sigma_1^2)$的样本，$Y_1, Y_2, \cdots, Y_n$为取自总体$Y \sim N(\mu_2, \sigma_2^2)$的样本。假设$X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n$ 相互独立。记

$$\overline{X} = \frac{1}{m}\sum_{i=1}^m X_i,$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i,$$

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2,$$

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2,$$

$$S_w^2 = \frac{1}{m+n-2}[\sum_{i=1}^{m}(X_i-\overline{X})^2+\sum_{i=1}^{n}(Y_i-\overline{Y})^2] = \frac{1}{m+n-2}[(m-1)S_X^2+(n-1)S_Y^2],$$

$$S_w = \sqrt{S_w^2},$$

则有

1. $$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1);$$

2. $$\frac{\frac{S_X^2}{S_Y^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \sim F(m-1, n-1);$$

3. $$\frac{\frac{\frac{1}{m}\sum_{i=1}^{m}(X_i-\mu_1)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_2)^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \sim F(m,n).$$

当$\sigma_1^2 = \sigma_2^2$时，有

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

2. $$F = \frac{S_X^2}{S_Y^2} \sim F(m-1, n-1).$$

4.3 非正态总体下的抽样分布

定理 4.8. 假设总体X的分布函数为$X_{(k)}$的密度函数为

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x), \qquad k = 1, 2, \cdots, n$$

定理 4.9. 设$X_1, X_2, \cdots, X_n$是取自指数分布总体$X \sim E(\lambda)$的样本，样本均值为$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，则$2\lambda n\overline{X} \sim \chi^2(2n)$。

注. 由因子分解定理易知：样本均值是λ的充分统计量。

对于一般总体而言，由于统计量的精确分布很难得到，经常在求样本容量较大时统计量的近似分布。正态分布时最常用的近似分布。

定义. 记总体分布中的未知参数为θ，如果存在$\nu(\theta) > 0$，使得统计量$T(X_1, X_2, \cdots, X_n)$满足$\sqrt{n}[T(X_1, X_2, \cdots, X_n) - g(\theta)]$按分布收敛于$N(0, \nu(\theta))$，则称统计量$T(X_1, X_2, \cdots, X_n)$为渐进正态的，也称$N(g(\theta), \frac{\nu(\theta)}{n})$为统计量$T(X_1, X_2, \cdots, X_n)$的渐进分布。

可以证明，如果统计量$T(X_1, X_2, \cdots, X_n)$为渐进正态的，则统计量$T(X_1, X_2, \cdots, X_n)$为$g(\theta)$的相合统计量。

根据中心极限定律，$N(\mu, \frac{\sigma^2}{n})$为统计量$\overline{X}$的渐进分布。

5 参数估计

5.1 点估计

5.1.1 矩估计法

矩估计的思想比较简单：通过抽样，我们能得到一些样本$X_1, X_2, \cdots, X_n$，从而可以算出它们的样本（中心）矩。同时，总体也有其自身的（中心）矩，它们往往是未知参数的函数。将样本的（中心）矩与总体所对应的（中心）矩进行联立，使得能够联立方程组能够解出未知参数，即得出用$X_1, X_2, \cdots, X_n$所表示的估计量即可。

一般用$\bar{X}$对应着总体的一阶矩，即期望；$S = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$而非样本二阶中心矩对应着总体的方差，$\sqrt{S}$对应着总体的标准差。

步骤可以总结为：

1. 列出所需要估计的参数
2. 选取与这些参数同等数量的总体矩，并将它们表示为这些参数的函数
3. 总体矩近似等于样本矩，得到联立方程组
4. 求解方程组

一般地，在第二步中，尽可能选取阶数较低的总体矩。例如在估计指数总体的参数λ时，我们知道$1/\lambda = EX = \sigma(X)$，通常将选择使用$1/\lambda = EX \approx \bar{X}$来进行联立。当低阶矩中不包含未知参数的信息时，我们才使用高阶矩。

5.1.2 极大似然估计法

设总体有分布$f(x;\theta_1,\theta_2,\cdots,\theta_n)$，$X_1,X_2,\cdots,X_n$为自这个总体中抽出的样本，则样本$(X_1,X_2,\cdots,X_n)$的分布为

$$f(x_1;\theta_1,\theta_2,\cdots,\theta_k)f(x_2;\theta_1,\theta_2,\cdots,\theta_k)\cdots f(x_n;\theta_1,\theta_2,\cdots,\theta_k),$$

记为$L(x_1,x_2,\cdots,x_n;\theta_1,\theta_2,\cdots,\theta_k)$。

当把L视为$x_1,x_2,\cdots,x_n$的函数时，它便是概率密度函数，反映的是给定$\theta_1,\theta_2,\cdots,\theta_k$时样本的分布概率。而当把$L$视为$\theta_1,\theta_2,\cdots,\theta_k$的函数时，它所反映的实际上是对于一个固定的样本，当参数变化时，取到这个样本的概率的变化情况，此时，当L取到最大值时，说明参数为最大值点$\theta_1^*,\theta_2^*,\cdots,\theta_n^*$时，取到给定样本的概率最大。反过来，我们也可以理解为当实际取到的便是给定的样本时，参数为$\theta_1^*,\theta_2^*,\cdots,\theta_n^*$的概率最大，因此我们便把它作为未知参数$\theta_1,\theta_2,\cdots,\theta_n$的估计。

注. 对于正态分布，其参数为μ,σ^2，而非μ,σ。因此在求最大值时要将σ^2视为整体，可以用换元的思想来理解。

该方法的核心便是求L的最大值。当该函数连续时，通常将其取对数后砸进行求导，从而将连乘转化为累加。当该函数不连续时，需要根据函数特性来判断最大值点。

例. 设$X_1,X_2,\cdots,X_n$是从均匀分布$R(0,\theta)$的总体中抽出的样本，求θ的极大似然估计。

现在我们要求的便是给定$X_1,X_2,\cdots,X_n$，θ变化时，取到该样本的概率。当$\theta\le\max\limits_i X_i$时，最大值不可能从该总体中取出，所以对应的$L$取0。当$\theta>\max\limits_i X_i$时，由于每个分量都是均匀分布，所以$L=1/\theta^n$。现在所要求的便是$L$的最大值。易知$L$是个分段函数，最大值为当$\theta=\max\limits_i X_i$时取到。所以用极大似然估计法得到的对于$\theta$的估计便是$\max\limits_i X_i$。

注. 这里实际上在均匀分布的密度函数在边缘处的取值问题上有点问题。这里实际上求的是上确界，取的是达到上确界的极限点。

5.1.3 贝叶斯法

贝叶斯法要求我们对于参数有一个先验知识，也就是在进行预估之前，我们对于该参数的密度函数的估计。该先验知识可以是主观的，甚至该密度函数可以并不满足

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

此时称该先验密度为广义先验估计。

贝叶斯法的核心还是贝叶斯公式。给定未知参数的先验密度$h(\theta)$，以及从总体中得到的样本$X_1, X_2, \cdots, X_n$。设总体有概率函数$f(X, \theta)$，那么这组岩本的密度为$f(X_1, \theta)f(X_2, \theta)\cdots f(X_n, \theta)$，$(\theta, X_1, X_2, \cdots, X_n)$的联合密度为

$$h(\theta)f(X_1, \theta)f(X_2, \theta)\cdots f(X_n, \theta).$$

由此算出$(X_1, X_2, \cdots, X_n)$的边缘密度为

$$p(X_1, X_2, \cdots, X_n) = \int h(\theta)f(X_1, \theta)f(X_2, \theta)\cdots f(X_n, \theta)\mathbf{d}\theta$$

得到在给定$(X_1, X_2, \cdots, X_n)$的条件下，θ的条件密度为

$$h(\theta|X_1, X_2, \cdots, X_n) = h(\theta)f(X_1, \theta)f(X_2, \theta)\cdots f(X_n, \theta)/p(X_1, X_2, \cdots, X_n).$$

从贝叶斯法的观点来看，该函数$h(\theta|X_1, X_2, \cdots, X_n)$综合了$\theta$的先验知识与样本带来的信息，因此称为$\theta$的后验密度。并且，从此之后先验密度就不再使用，被后验密度所替代。通常，取后验密度的期望作为未知参数的估计。

5.1.4 其他

当上述方法均不方便使用时，可以使用其他方法进行估计。例如，如果一个参数是分布的均值，则可以用但是总体的期望并不存在时，我们无法直接使用矩估计法，但是我们可以用样本的中位数来进行估计。

实际上，只要估计方法合理，所得到的结果都是可以接受的，但是到底那种估计方法是最优的，则需要使用点估计的优良性法则进行评价。

5.2 点估计的优良性法则

5.2.1 估计量的无偏性

设某统计总体的分布包含未知参数$\theta_1, \theta_2, \cdots, \theta_k$。$X_1, X_2, \cdots, X_n$则是从该总体中抽出的样本，现要估计$g(\theta_1, \theta_2, \cdots, \theta_n)$。$g$ 为一已知函数。设$\hat{g}(X_1, X_2, \cdots, X_n)$是一个估计量。如果对任何可能的$(\theta_1, \theta_2, \cdots, \theta_k)$，都有

$$E_{\theta_1, \theta_2, \cdots, \theta_k}[\hat{g}(X_1, X_2, \cdots, X_n)] = g(\theta_1, \theta_2, \cdots, \theta_k),$$

则称$\hat{g}$是$g(\theta_1, \theta_2, \cdots, \theta_k)$的一个无偏估计量。其中记号$E_{\theta_1, \theta_2, \cdots, \theta_k}$的含义是给定参数$(\theta_1, \theta_2, \cdots, \theta_k)$的情况下的期望。

一个估计量具有无偏性的含义是：虽然给定样本$X_1, X_2, \cdots, X_n$，带入估计量中所得到的估计值可能与实际有偏差，但是这个偏差对于样本所服从的概率密度而言，平均为0。同时，无论样本的概率密度中的参数怎么变化，均满足前一要求。这一含义可以理解为该估计量没有系统性的偏差。

同时，大数定律告诉我们，当一个估计量具有无偏性时，对于给定的总体密度，只需要测足够多的次数，然后对得到的估计值取平均，我们所得到的值可以无限近似于实际值。

易知，样本均值是总体分布均值的无偏估计；S^2是总体分布方差的无偏估计。因为$\sigma^2 = E(S^2) = E(S)^2 + Var(S)$，所以$S$并不是$\sigma$的无偏估计，会系统性偏小，为此，可以将$S$乘以一个大于1、于样本大小$n$有关的因子$c_n$，得到$c_n S$来使得该估计量为无偏估计。

例. 判断$\hat{\theta} = \max_i X_i$作为均匀总体$R(0, \theta)$的$\theta$的估计量是否无偏。

对于给定的θ，该估计量的分布实际上就是最大值分布，所以$\hat{\theta}$的概率密度函数为

$$g(x, \theta) = \begin{cases} nx^{n-1}/\theta^n, & \text{当} 0 < x < \theta \\ 0, & \text{其他} \end{cases}$$

对于给定的θ：

$$E_\theta(\hat{\theta}) = \int_0^\theta x g(x, \theta) \mathbf{d}x = \frac{n}{n+1}\theta.$$

所以不是无偏估计。若想成为无偏估计，则需要乘以因子$\frac{n+1}{n}$

如果一个估计量并不是无偏的，但是满足

$$\lim_{n \to +\infty} E_\theta[\hat{\theta}(X_1, X_2, \cdots, X_n)] = \theta$$

则称$\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$是$\theta$的一个渐进无偏估计。

5.2.2 最小方差无偏估计

一个参数往往有多个无偏估计，而这些无偏估计中，仍有优劣之分。一个常用的评价标准便是均方误差：

$$M_{\hat{\theta}}(\theta) = E_\theta[\hat{\theta}(X_1, X_2, \cdots, X_n) - \theta]^2.$$

均方误差从整体的角度衡量了估计量与实际值之间的误差。

因

$$M_{\hat{\theta}}(\theta) = E_\theta[\hat{\theta} - \theta]^2 = Var_\theta(\hat{\theta} - \theta) + E_\theta^2(\hat{\theta} - \theta) = Var_\theta(\hat{\theta}) + (E_\theta\hat{\theta} - \theta)^2$$

所以当$\hat{\theta}$为无偏估计，即$E_\theta\hat{\theta}=\theta$时，我们有

$$M_{\hat{\theta}}(\theta)=Var_\theta(\hat{\theta}).$$

所以无偏估计之间的比较归结为了它们的方差之间的比较：方差小者为优。

定义. 设$\hat{\theta}$为$g(\theta)$的一个无偏估计。若对$g(\theta)$的任何一个无偏估计$\hat{\theta_1}$，都有

$$Var_\theta(\hat{\theta})\leq Var_\theta(\hat{\theta_1})$$

对于θ的任何取值均成立，则称$\hat{\theta}$为$g(\theta)$的一个最小无偏估计(MVU)。

注. MVU即 Minimum Variance Unbiased的缩写。

想要直接求一个参数的MVU难度较高，但是我们可以用一定的方法来判断一个无偏估计是否为MVU：只需算出参数的估计的均方误差的下界，如果给定的估计量的均方误差达到该下界，那么它便是一个MVU。

定理 5.1. 在一定的条件下，对$g(\theta)$的任一无偏估计$\hat{g}=\hat{g}(X_1,X_2,\cdots,X_n)$，有

$$Var_\theta(\hat{g})\geq(g;(\theta))^2/(nI(\theta)),$$

其中n为样本大小，$I(\theta)$为费歇尔信息量，定义为

$$I(\theta)=\int\left[\left(\frac{\partial f(x,\theta)}{\partial\theta}\right)^2/f(x,\theta)\right]\mathbf{d}x,$$

积分范围即x可取的范围。如果总体分布是离散型的，则上式改为

$$I(\theta)=\sum_i\left(\frac{\partial f(x,\theta)}{\partial\theta}\right)^2/f(x,\theta)\mathbf{d}x,$$

该不等式给出的便是均方误差的下界。

5.2.3 估计量的相合性

大数定理告诉我们，若$X_1,X_2,\cdots,X_n,\cdots$独立同分布，其公共均值为$\theta$。记$\bar{X_n}=\sum_{i=1}^n X_i/n$，则对于任意$\epsilon>0$，有

$$\lim_{n\to\infty}P(|\bar{X_n}-\theta|\geq\epsilon)=0.$$

从统计量的角度去看，这告诉我们，如果我们采集的样本足够大，则样本均值依概率收敛于总体均值。

更一般地，我们有

定义. 设总体分布依赖于参数$\theta_1, \theta_2, \cdots, \theta_k$，且$g(\theta_1, \theta_2, \cdots, \theta_k)$是$\theta_1, \theta_2, \cdots, \theta_k$的一个给定函数。设$X_1, X_2, \cdots, X_n$为自该总体中抽出的样本，$T(X_1, X_2, \cdots, X_n)$是$g(\theta_1, \theta_2, \cdots, \theta_k)$的一个估计量。如果对任给的$\varepsilon > 0$，有

$$\lim_{n\to\infty} P_{\theta_1,\theta_2,\cdots,\theta_k}(|T(X_1, X_2, \cdots, X_n) - g(\theta_1, \theta_2, \cdots, \theta_k)| \geq \varepsilon) = 0,$$

而且这对$(\theta_1, \theta_2, \cdots, \theta_k)$一切可能取的值都成立，则称$T(X_1, X_2, \cdots, X_n)$是$g(\theta_1, \theta_2, \cdots, \theta_k)$的一个相合估计。

也就是说，当一个估计值是相合的时，只要样本数量足够大，该估计值就能足够精确。

定理 5.2. 若$\hat{\theta}$是θ的一个无偏估计或渐进无偏估计，且$\lim\limits_{n\to+\infty} Var_\theta(\hat{\theta}) = 0$，则$\hat{\theta}$是$\theta$的一个相合估计。

5.2.4 小结

当我们进行参数估计时，所估计的参数是已经确定下来的，但是实际上它是有一个取值范围，这就要求我们的估计方法以及对于估计的优良性评估是要对所有可能的取值都适用的。因此，我们使用E_θ, Var_θ等记号，来表明参数是给定的，但是可以“流动”的。

而估计时，我们利用的是样本。样本本质上是服从总体分布的随机变量，由样本组成的函数也是由随机变量组成的函数，有其自身的分布和矩。

当我们进行参数估计时，未知参数的确定值实际上无所谓。而我们在对估计量进行评估时，就要利用未知参数的确定值来作为标准。无偏性就是在对估计量的期望和未知参数的确定值进行比较，要注意，估计量是样本的函数，而样本是随机变量，其分布又是由未知参数所决定的。在无偏性的基础上，又可以进一步比较该估计值的方差，这本质上也是随机变量的方差。无偏性是从平均的角度去评价一个估计量，而相合性则是从样本大小和准确率的相关性上进行评价，因此前者不涉及样本大小，而后者涉及。

5.3 渐近正态性

许多统计量的分布都难以计算，但是可以证明，大多数统计量的分布当$n \to \infty$时，都渐进于正态分布，这称为统计量的“渐近正态性”。

5.4 区间估计

区间估计，顾名思义，便是利用一个区间来作为一个参数的估计。与点估计不同，区间估计能在估计的同时提供估计的误差这一信息。

设$X_1, X_2, \cdots, X_n$是从总体中抽出的样本，所谓θ的区间估计，就是以满足条件$\hat{\theta_1}(X_1, X_2, \cdots, X_n) \leq \hat{\theta_2}(X_1, X_2, \cdots, X_n)$ 的两个统计量$\hat{\theta_1}, \hat{\theta_2}$为端点的区间$[\hat{\theta_1}, \hat{\theta_2}]$.

评价一个区间估计的好坏主要有两个维度：

1. 参数的真值落在该区间的概率要大
2. 该区间的长度要小

但是这两个要求是相互矛盾的。基本原则是先保证前者，在保证后者。

定义. 给定一个很小的数$\alpha > 0$。如果对参数θ的任何值，

$$P_\theta(\hat{\theta_1}(X_1, X_2, \cdots, X_n) \leq \theta \leq \hat{\theta_2}(X_1, X_2, \cdots, X_n)) = 1 - \alpha$$

均成立，则称区间估计$[\hat{\theta_1}, \hat{\theta_2}]$的置信系数为$1 - \alpha$。

区间估计也常称为“置信区间”，即“对该区间能包含未知参数θ可置信到何种程度”。

如果只能保证上式中的概率小于等于$1 - \alpha$，则称$1 - \alpha$为$[\hat{\theta_1}, \hat{\theta_2}]$的“置信水平”。

5.4.1 枢轴变量法

枢轴变量法是寻找区间估计的方法之一，其步骤为：

1. 找到一个与要估计的参数$g(\theta)$有关的统计量T，一般是其一个良好的点估计
2. 设法找出T和$g(\theta)$的某一函数$S(T, g(\theta))$，其分布F要与θ无关
3. 对任何常数$a < b$，不等式$a \leq S(T, g(\theta)) \leq b$要能改写为等价的形式$A \leq g(\theta) \leq B$，$A, B$只与$T, a, b$有关，而与$\theta$无关
4. 取分布F的分位点$w_{\alpha/2}, w_{1-\alpha/2}$，则有$F(w_{1-\alpha/2}) - F(w_{\alpha/2}) = 1 - \alpha$，因此
$$P(w_{\alpha/2} \leq S(T, g(\theta)) \leq w_{1-\alpha/2}) = 1 - \alpha.$$

5. 改写为等价形式

$$P(A(T) \leq g(\theta) \leq B(T)) = 1 - \alpha.$$

这样，给定一组样本，我们便能得到一个具体的区间。虽然参数仍是未知的，但是未知参数和样本不同，后者可以视为随机变量，但是前者是已经确定了的。所以上述“概率”的含义是指：用该估计方法得到的区间估计，平均100次中有$100 \cdot P$次包含了所要估计的值，而不是有 100·P%的概率包含所要估计的值。

例. 设$X_1, X_2, \cdots, X_n$为抽自指数分布总体的样本，要求其参数λ的区间估计。

由于$2n\lambda\bar{X} \sim \chi^2_{2n}$，所以令$S(\bar{X}, \lambda) = 2n\lambda\bar{X}$，从而得到区间估计。

5.4.2 大样本法

大样本法便是利用极限分布，主要是中心极限定理，来建立枢纽变量。

例. 某事件A在每次试验中发生的概率为p。做n次独立试验，以Y_n记A发生的次数，要求p的区间估计。当n相当大时，由中心极限定理可知，我们近似地有

$$(Y_n - np)/\sqrt{np(1-p)} \sim N(0,1).$$

从而有

$$P(-u_{\alpha/2} \leq (Y_n - np)/\sqrt{np(1-p)} \leq u_{\alpha/2}) \approx 1 - \alpha$$

可以进一步改写为

$$P(A \leq q \leq B) \approx 1 - \alpha.$$

5.4.3 置信界

定义. 设$X_1, X_2, \cdots, X_n$是从某一总体中抽出的样本，总体分布包含未知参数θ，$\bar{\theta} = \bar{\theta}(X_1, X_2, \cdots, X_n)$，和$\underline{\theta} = \underline{\theta}(X_1, X_2, \cdots, X_n)$ 都是统计量，则：

1. 若对θ的一切取值，有

$$P_\theta(\bar{\theta}(X_1, X_2, \cdots, X_n) \geq \theta) = 1 - \alpha,$$

则称$\bar{\theta}$为θ的一个置信系数为$1-\alpha$的置信上界。

2. 若对θ的一切取值，有

$$P_\theta(\underline{\theta}(X_1, X_2, \cdots, X_n) \leq \theta) = 1 - \alpha,$$

则称$\underline{\theta}$为θ的一个置信系数为$1-\alpha$的置信下界。

类似地，我们可以使用枢轴变量法来求参数的置信上界和置信下界。

5.4.4 贝叶斯法

沿用之前的符号，如果$\hat{\theta_1}, \hat{\theta_2}$满足

$$\int_{\hat{\theta_1}}^{\hat{\theta_2}} h(\theta|X_1, X_2, \cdots, X_n)\mathbf{d}\theta = 1 - \alpha,$$

则贝叶斯法将$[\hat{\theta_1}, \hat{\theta_2}]$作为$\theta$的一个区间估计。通常，这样子的区间不是唯一的，此时通常取区间长度最小的那一个。

对于置信上界和置信下界，贝叶斯法也用类似的方法求得。

5.4.5 区间估计的长度

当利用枢纽变量法来求区间估计时，如果区间的长度仅与样本数量有关，而与样本取值无关，则我们可以通过调整样本数量来调整区间估计的区间长度。如果区间的长度与样本取值无关，则需要更高级的理论来进行解决。

当利用贝叶斯方法来求区间估计时，则可以利用分析的方法来求得区间长度的最小值。