Notes in Probability Theory and Mathematical Statistics

Weizheng Zhang
imzwz@qq.com
April 10, 2018

Cover: Galton Board

Preface

由于觉得教材 [1] 太luō (LaTeX编译不出这个字) 嗦且公式排版不好, 甚至有些错误, 于是自己整理了一份总结, 主要参考内容是 [1,2], 所涉及的集合论基础不再赘述. 章节标题用中文编译时会报错, 因此章节标题都用了英文. 附录表和数理统计的部分待补充. 特别鸣谢维基百科. 若读者发现错误或不足, 欢迎邮件联系 imzwz@qq.com 指正, 谢谢.

编者
2018年4月

符号表

$\mathbb{R}$ 实数集 $(-\infty, +\infty)$

$\mathbb{R}_+$ 正实数集 $(0, +\infty)$

$\mathbb{Z}$ 整数集 $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$

$\mathbb{Z}_+$ 正整数集 $\{1, 2, 3, \dots\}$

$\mathbb{N}$ 自然数集 $\{0, 1, 2, 3, \dots\}$

A^c或$\overline{A}$ 集合A(关于全集)的补集

$\mathcal{P}(\Omega)$ 集合Ω的幂集

$\sigma(\mathcal{E})$ 集族$\mathcal{E}$生成的σ域

$\mathcal{B}(\Omega)$ 集合Ω上的Borel σ域

$\overline{\lim}_{n\to\infty} A_n$或$\limsup_{n\to\infty} A_n$ $\{A_n\}$的上极限集$\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$

$\underline{\lim}_{n\to\infty} A_n$或$\liminf_{n\to\infty} A_n$ $\{A_n\}$的下极限集$\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k$

$\binom{n}{k}$ 组合数$C_n^k = \frac{n!}{k!(n-k)!}$

$E(X)$ X的数学期望

$\mathrm{Var}(X)$ X的方差

$\mathrm{Cov}(X, Y)$ X与Y的协方差

ρ_{XY} X与Y的相关系数

$E(X|Y)$ X关于Y的条件期望

$X \perp Y$ X与Y独立

$\mathbf{1}_A$ 集合A的示性函数

$\Gamma(\alpha)$ Gamma函数$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} \mathrm{d}x,\ \alpha > 0;\ \Gamma(n) = (n-1)!,\ n \in \mathbb{Z}_+$

$\mathrm{B}(x, y)$ Beta函数$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} \mathrm{d}t = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},\ x, y > 0$

Contents

1 Events and Probability

Definition 1.1 把随机试验E中每一个可能出现的结果称为**样本点**（用ω表示）, 所有可能的样本点组成的集合称为E的**样本空间**（用Ω表示）. 一些样本点组成的集合称为**（随机）事件**（用大写英文字母表示）.

- 样本空间——全集Ω
- 样本点——元素$\omega \in \Omega$
- 事件——全集的子集$A \subset \Omega$
- 事件发生——事件中的某一元素所对应的现象发生（设ω发生, 如果$\omega \in A$, 则称事件A发生）. "事件A发生必然导致事件B发生"即$A \subset B$, "事件A和B互不相容"即$A \cap B = \varnothing$
- 全集Ω包含所有样本点, 对应必然事件, 空集$\varnothing$不包含任何样本点, 对应不可能事件.

Definition 1.2 如果试验E满足：样本空间只包含有限个样本点（即$\Omega = \{\omega_1, \ldots, \omega_n\}$）, 且每种结果发生的可能性相同（即$P(\{\omega_i\}) = \frac{1}{n}, i = 1, \ldots, n$）, 则称这样的试验模型为古典概率模型（简称**古典概型**） "样本点有限且等可能发生"

排列组合基础（相关的常见练习题看第一章课件79页）

1. 二项式展开:
$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$
2. 多项式组合：把n个不同元素分成k个部分, 各部分包含元素个数分别为$r_1, \ldots, r_k$, 则不同的分配方式共有$\binom{n}{r_1 \cdots r_k} = \frac{n!}{r_1! \cdots r_k!}$种.
3. 从n个物体中抽取r个，总共的组合数由下表所示:

组合总数	无放回	放回
有序	$\frac{n!}{(n-r)!}$	n^r
无序	$\binom{n}{r}$	$\binom{n+r-1}{r}$

Definition 1.3 几何概型：概率与"长度/面积/体积"成正比（略）

Definition 1.4 设集类$\mathcal{F} \subset \mathcal{P}(\Omega)$（全集的幂集）, 若满足:

1. $\Omega \in \mathcal{F}$（包含全集）;

2. $A \in \mathcal{F} \Rightarrow \overline{A} \in \mathcal{F}$（对差运算封闭）;

3. $A_k \in \mathcal{F}(k=1,2,\dots) \Rightarrow \cup_{k=1}^{\infty} A_k \in \mathcal{F}$（对可列并运算封闭）;

则称$\mathcal{F}$为Ω上的一个σ**域**或σ**代数**. 由定义可以推出$\varnothing \in \mathcal{F}$, 且$\mathcal{F}$对可列交运算封闭.

若$\mathcal{F}$是Ω上的σ域, 则称它为**事件域**, 其中的元素（是集合）称为**事件**.

Definition 1.5 （*Borel*集）

- 一维*Borel*集：由$\mathbb{R}$中所有形如$[a,b)$的左闭右开区间构成的集类所生成的σ域称为一维*Borel* σ域, 记为$\mathcal{B}(\mathbb{R})$. $\mathcal{B}(\mathbb{R})$中的元素（是集合）称为一维*Borel*集.
- n维*Borel*集：由$\mathbb{R}^n$中所有n维矩体构成的集类所生成的σ域称为n维*Borel* σ域, 记为$\mathcal{B}(\mathbb{R}^n)$. $\mathcal{B}(\mathbb{R}^n)$中的元素（是集合）称为$n$维*Borel*集.
- 一般地, Ω上的*Borel* σ域是指由Ω的所有开子集构成的集类所生成的σ域, 记为$\mathcal{B}(\Omega)$. $\mathcal{B}(\Omega)$中的元素（是集合）称为Ω上的*Borel*集.

Definition 1.6 （*Kolmogorov*公理系统）
定义在事件域$\mathcal{F}$上的集合函数P如果满足:

1. 非负性: $P(A) \geq 0, \forall A \in \mathcal{F}$

2. 规范性: $P(\Omega) = 1$

3. 可列可加性: $A_1, A_2, \dots \in \mathcal{F}$是两两不相容事件$\Rightarrow P(\sum_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

则称P为可测空间$(\Omega, \mathcal{F})$上的一个概率测度（简称**概率**）, $(\Omega, \mathcal{F}, P)$称为**概率空间**. 由定义可推出$P(\varnothing) = 0$.

若全空间Ω是可数集, 则$(\Omega, \mathcal{F}, P)$称为离散概型的概率空间.

Property 1.7 （概率测度的一些性质）

1. （加法公式，用容斥原理）$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

2. （次可加性）$P(A \cup B) \leq P(A) + P(B)$.

3. （*Bonferroni*不等式）$P(A \cap B) \geq P(A) + P(B) - 1$.

Definition 1.8 （上连续、下连续）
设$P(\cdot)$是$\mathcal{F}$上的集合函数,

1. 若它对$\mathcal{F}$中任何单调递增序列$\{S_n\}$都有

$$\lim_{n\to\infty} P(S_n) = P(\lim_{n\to\infty} S_n) \equiv P(\cup_{n=1}^{\infty} S_n)$$

则称它是**下连续**的;

2. 若它对$\mathcal{F}$中任何单调递减序列$\{S_n\}$都有

$$\lim_{n\to\infty} P(S_n) = P(\lim_{n\to\infty} S_n) \equiv P(\cap_{n=1}^{\infty} S_n)$$

则称它是**上连续**的;

Theorem 1.9 若P为$\mathcal{F}$上满足$P(\Omega)=1$的非负集合函数, 则它有可列可加性的充要条件为它是有限可加的且是下连续的.

Corollary 1.10 概率是下连续的且是上连续的.（证明见 *[2]*第一章课件*145*页）

2 Conditional Probability and Independence

Definition 2.1 设$(\Omega, \mathcal{F}, P)$是概率空间, $B \in \mathcal{F}$且$P(B) > 0$, 则对任意$A \in \mathcal{F}$, 记$P(A|B) = \dfrac{P(AB)}{P(B)}$, 称为事件$B$发生的条件下事件$A$发生的**条件概率**. 易知条件概率也是概率, 具有概率的所有性质.

乘法公式: 设$P(B) > 0$, 则$P(AB) = P(B)P(A|B)$; 设$P(A) > 0$, 则$P(AB) = P(A)P(B|A)$;
设$P(A_1A_2\cdots A_{n-1}) > 0$, 则

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2|A_1)\cdots P(A_n|A_1A_2\cdots A_{n-1}).$$

注: n个事件的概率乘法公式有$n!$个, 这只是其中一个.

全概率公式: 设事件$\{B_i\}_{i=1}^{\infty}$是样本空间Ω的一种分割, 且$P(B_i) > 0 (\forall i \in \mathbb{Z}^+)$, 则有

$$A = \sum_{i=1}^{\infty} AB_i,$$

$$P(A) = \sum_{i=1}^{\infty} P(AB_i) = \sum_{i=1}^{\infty} P(B_i)P(A|B_i).$$

贝叶斯公式: 设事件$\{A_i\}_{i=1}^{\infty}$是样本空间Ω的一种分割, 且$P(B) > 0$, 则有

$$P(A_i|B) = \frac{P(A_iB)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{\infty} P(A_j)P(B|A_j)}.$$

事件的独立性:

1. 称事件A与B**独立**, 若$P(AB) = P(A)P(B)$. 当$P(B) > 0$时, 有$P(A|B) = P(A)$;

2. 称事件A, B与C**相互独立**, 若以下条件同时成立: $P(AB) = P(A)P(B)$, $P(BC) = P(B)P(C)$, $P(AC) = P(A)P(C)$, $P(ABC) = P(A)P(B)P(C)$ (若只满足前面三式子, 则称事件A, B与C**两两独立**. 易知相互独立⇒两两独立, 但逆命题不成立);

3. 称一组事件$\{A_i\}_{i=1}^n$相互独立, 若以下条件同时成立: （略, 共2^n-n-1个等式）;

4. 必然事件Ω与不可能事件∅都与任何事件独立;

5. “相互独立”与“互不相容”是两个无关的概念.

Theorem 2.2

1. 若事件A与B独立，则A与$\bar{B}$独立, $\bar{A}$与B独立, $\bar{A}$与$\bar{B}$独立.

2. 若事件, B与C相互独立，则: $A\cup B$与C独立, $A\cap B$与C独立, $A-B$与C独立.

注: “重复试验”一般蕴含各次试验是相互独立的.

牛顿二项式: 对正整数k和任意实数a, 有$\binom{-a}{k} = (-1)^k\binom{a+k-1}{k}$, 和$(1+x)^a = \sum_{r=0}^{\infty}\binom{a}{r}x^r$.

Theorem 2.3 小概率事件发生的概率为*1:* 设随机事件A在一次试验中发生的概率是$p > 0$, 如果不停地进行独立重复试验, 则事件A最终发生的概率为*1.*

证明*:*

$$P(\cup_{k=1}^{\infty}D_k) = 1 - P(\cap_{k=1}^{\infty}\bar{D}_k) = 1 - \prod_{k=1}^{\infty}P(\bar{D}_k) = 1 - \prod_{k=1}^{\infty}(1-p) = 1 - 0 = 1.$$

直线上的随机游动：考虑x轴上的一个质点, 在时刻$t = 0$时, 它处于初始位置a (整数)，以后每隔单位时间, 分别以概率p及概率$q = 1ﬂp$向正的或负的方向移动一个单位. 用这种方式描述的质点的运动称为随机游动. 质点在时刻 $t = n$ 时位于$k(-n \le k \le n)$等价于在前n次游动中向右移动的次数比向左移动的次数多k次.

若质点可以在整个数轴的整数点上游动，则称这种随机游动为无限制随机游动. 无限制的随机游动——有无穷赌本的赌徒在n局后的输赢

若在d点设有一个吸收壁，质点一到达这点即被吸收而不再游动，因而整个游动就结束，这种随机游动称为在d点有吸收壁的随机游动.

两端带有吸收壁的随机游动——有穷赌本的赌徒的输赢

多项分布: （二项分布的推广）n次重复独立试验且每次试验有若干种结果$A_1,\ldots,A_r$, $P(A_i)=p_i\geq 0(i=1,\ldots,r)$, 且$\sum_{i=1}^r p_i=1$, 则在$n$次试验中$A_i$出现$k_i$次($\forall i=1,\ldots,r$且$\sum_{i=1}^r k_i=n$)的概率为

$$\frac{n!}{k_1!\cdots k_r!}p_1^{k_1}\cdots p_r^{k_r}.$$

对于二项分布$b(k;n,p)$, 当$(n+1)p$不为整数时, $b(k;n,p)$在$k=\lfloor(n+1)p\rfloor$达到最大值; 当$(n+1)p=:m$为整数时, $b(k;n,p)$在$k=m$和$k=m-1$达到最大值.

Theorem 2.4 泊松定理（二项分布的泊松近似）
设常数$\lambda>0$, $n\in\mathbb{Z}_+$. 若$np\to\lambda$, 则对任一固定的非负整数k, 有

$$\lim_{n\to\infty}\binom{n}{k}p_n^k(1-p_n)^{n-k}=\frac{\lambda^k}{k!}e^{-\lambda}.$$

因此，**泊松分布是二项分布的极限分布**, 当n很大而p很小时, 二项分布可以近似看成是参数$\lambda=np$的泊松分布, 即$b(k;n,p)\approx\dfrac{(np)^k}{k!}e^{-np}$.

$b(k;n,p)=b(n-k;n,1-p)$.

对于泊松分布$p(k;\lambda)=\dfrac{\lambda^k}{k!}e^{-\lambda},k=0,1,2,\ldots$, 有$\dfrac{P(X=k)}{P(X=k-1)}=\dfrac{\lambda}{k}$. 因此当$\lambda$是整数时, $k=\lambda$或$\lambda-1$使概率最大; 当λ不是整数时, $k=\lfloor\lambda\rfloor$使概率最大.

Gamma函数$\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x$, $\forall\alpha>0$ 具有以下性质:
$\Gamma(1)=1$, $\Gamma(1/2)=\sqrt{\pi}$, 且$\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$.

Beta积分$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x$, $\forall\alpha,\beta>0$ 具有以下性质:
$B(\alpha,\beta)=\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

泊松分布的尾概率: 对$\forall r\in\mathbb{Z}_+$, 有$\sum_{k=r}^\infty\dfrac{\lambda^k}{k!}e^{-\lambda}=\dfrac{1}{(r-1)!}\int_0^\lambda x^{r-1}e^{-\lambda}\mathrm{d}x$, 即$\Gamma(r,1)$在$\lambda$处的分布函数值.

二项分布的尾概率: 对$\forall r\in\mathbb{Z}_+$, 有$\sum_{k=r}^n\binom{n}{k}p^kq^{n-k}=r\binom{n}{r}\int_0^p x^{r-1}(1-x)^{n-r}\mathrm{d}x$, 即$\beta(r,n-r+1)$在$p$处的分布函数值.

3 Random Variables and Distributions

3.1 Random Variables and Distribution Functions

Definition 3.1 （随机变量）
设$\xi(\omega)$是定义在概率空间$(\Omega, \mathcal{F}, P)$上的单值实函数, 如果对于$\mathbb{R}$上任意$Borel$集B有$\{\omega : \xi(\omega) \in B\} \in \mathcal{F}$, 则称$\xi(\omega)$为**随机变量**（即$Borel$集的原像是事件）, 而$P\{\xi(\omega) \in B\}$称为随机变量$\xi(\omega)$的概率分布.

注: $\{\xi(\omega) < x\}$是$\{\omega : \xi(\omega) < x\}$的缩写.

Definition 3.2 （分布函数） 称$F(x) = P\{\xi(\omega) < x\}, -\infty < x < +\infty$ 为随机变量$\xi(\omega)$的（累计）**分布函数**, 记作$\xi(\omega) \sim F(x)$.

注: 教材 [1]的定义是$F(x) = P\{\xi(\omega) < x\}$, 而大部分主流教材定义是$F(x) = P\{\xi(\omega) \le x\}$, 从而导致一些性质的表述不同（比如左连续还是右连续）. 考试时以教材的定义为准!

Property 3.3 分布函数的性质:

1. 单调非减: $a < b \Rightarrow F(a) \le F(b)$;
2. $0 \le F(x) \le 1$, 且$F(-\infty) = 0$, $F(+\infty) = 1$;
3. 左连续性: $F(x^-) = \lim_{t \to x^-} F(t) = F(x)$.

Theorem 3.4 满足以上三个性质的函数一定是某个随机变量的分布函数. 因此可以把满足这三个性质的函数都称为分布函数.

Theorem 3.5 $P\{\xi(\omega) \le x\} = F(x^+)$; $P\{\xi(\omega) = x\} = F(x^+) - F(x)$;
$P\{\xi(\omega) \ge x\} = 1 - F(x)$; $P\{\xi(\omega) > x\} = 1 - F(x^+)$;
$P\{a \le \xi(\omega) < b\} = F(b) - F(a)$.

两类重要的随机变量: 离散型随机变量、连续型随机变量

Definition 3.6 （离散型随机变量）
若随机变量$\xi(\omega)$的全部可能取值是可数个, 则称$\xi(\omega)$是**离散型随机变量**. 此时设其所有可能取值为$x_k (k = 1, 2, \dots)$, 则$P\{\xi = x_k\} = p_k \ (k = 1, 2, \dots)$称为$\xi$的概率分布或**分布律**.

对于离散型随机变量, 分布函数与分布律可以互相唯一确定.

无记忆性:

1. **几何分布**是**唯**一的具有无记忆性的**离散**型分布. 若随机变量X服从几何分布, 则它具有无记忆性:

$$P\{X > m+n|X > m\} = P\{X > n\}, \forall m, n \geq 1,$$

或

$$P\{X = m+n|X > m\} = P\{X = n\}.$$

证明⇒: $P\{X > n\} = q^n (n = 1, 2, \dots)$, 故$P\{X > m+n|X > m\} = q^n = P\{X > n\}$.

几何分布具有无记忆性的根本原因在于, 进行的是**独立重复试验**.

参数为$r = 1$的帕斯卡分布就是几何分布. 参数为r的帕斯卡分布可以分解为r个独立同分布的几何分布的随机变量的和.

2. **指数分布**是**唯**一的具有无记忆性的**连续**型分布. 若取非负实值的随机变量X服从指数分布, 则它具有无记忆性:

$$P\{X > s+t|X > s\} = P\{X > t\}, \forall s, t > 0.$$

3. 几何分布与指数分布的关系: 伯努利过程可视为泊松过程的离散化, 若每间隔Δt进行一次试验，到时刻$n\Delta t$为止, 共进行n次试验, 伯努利过程中成功次数服从二项分布, 泊松过程中到时刻t的到来数服从泊松过程. 为等待第一次成功，伯努利试验中的等待时间服从几何分布，而泊松过程中的等待时间服从指数分布. 为等待第r次成功，伯努利试验中的等待时间服从帕斯卡分布（负二项分布），而泊松过程中的等待时间服从Erlang分布;

4. 指数分布与泊松过程的关系: 参数为λ_t的泊松过程中第1个跳跃发生的时刻ξ服从参数为λ的指数分布.

$\Gamma(1, \lambda)$是指数分布, $\Gamma(r, \lambda)$是Erlang分布, $\Gamma(\frac{n}{2}, \frac{1}{2})$是自由度为$n$的卡方分布. 指数分布就是$r = 1$时的Erlang分布.

参数为λ_t的泊松过程中第r个跳跃发生的时刻ξ_r服从Erlang分布.

3.2 Independence

Definition 3.7 （连续型随机变量）

若$F(x)$是随机变量$\xi(\omega)$分布函数, 若存在$\mathbb{R}$上的非负可积函数$p(x)$使得

$$F(x) = \int_{-\infty}^{x} p(t)\mathrm{d}t, \quad \forall x \in \mathbb{R},$$

则称$\xi(\omega)$是**连续型随机变量**, 称$p(x)$为ξ的**概率密度函数**（简称概率密度或密度函数, *pdf*）.

$P\{a \leq \xi(\omega) < b\} = F(b) - F(a) = \int_a^b p(x)\mathrm{d}x$.

随机变量ξ是连续型随机变量$\Longleftrightarrow$它的分布函数是**绝对连续**的$\Longleftrightarrow$它的分布函数能写成不定积分的形式.

一个可微函数是绝对连续函数$\Longleftrightarrow$它等于它的导函数的Lebesgue积分.

概率密度函数的性质:

1. 非负性: $p(x) \geq 0$;
2. $\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$;
3. $F(x)$几乎处处可导, 且$p(x) = F'(x)$, a.e. $x \in \mathbb{R}$.

两个连续型随机变量的概率密度函数如果**几乎处处**相等, 则称两者同分布.

Theorem 3.8 *若连续型随机变量ξ有分布函数$F(x)$, 则随机变量$F(\xi) \sim U(0,1)$; 若随机变量$\eta \sim U(0,1)$, 则随机变量$F^{-1}(\eta)$的分布函数为$F(x)$.*

该定理说明了均匀分布在随机模拟中的重要地位.

3σ原则: 设随机变量$X \sim N(\mu, \sigma^2)$, 则

$$P\{|X - \mu| < \sigma\} \approx 0.6826;$$

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

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

Definition 3.9 （随机向量） 若随机变量$\xi_1(\omega), \xi_2(\omega), \ldots, \xi_n(\omega)$定义在同一概率空间$(\Omega, \mathcal{F}, P)$上, 则称$\xi(\omega) = (\xi_1(\omega), \xi_2(\omega), \ldots, \xi_n(\omega))$是一个$n$维**随机向量**.

若B为$\mathbb{R}^n$上任一Borel集,则有$\{\xi(\omega) \in B\} \in \mathcal{F}$. 对于$n$个任意实数$x_1, x_2, \ldots, x_n$, 有

$$\{\xi_1(\omega) < x_1, \xi_2(\omega) < x_2, \ldots, \xi_n(\omega) < x_n\} = \cap_{i=1}^n \{\xi_i(\omega) < x_i\} \in \mathcal{F}.$$

Definition 3.10 （联合分布函数） 称n元函数$F(x_1, x_2, \ldots, x_n) = P\{\xi_1(\omega) < x_1, \xi_2(\omega) < x_2, \ldots, \xi_n(\omega) < x_n\}$为随机向量$\xi_1(\omega), \xi_2(\omega), \ldots, \xi_n(\omega)$的**联合分布函数**.

多元分布函数的性质:

1. 单调性：关于每个变元是单调不减函数（monotonic non-decreasing）;
2. $F(x_1, \ldots, -\infty, \ldots, x_n) = 0$, $F(+\infty, \ldots, +\infty) = 1$;

3. 关于每个变元左连续（主流教材是右连续, 这里以课本为准）;

4. 在二元场合, 对任意$a_1 < b_1, a_2 < b_2$, 有

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0.$$

注: 对于二元分布函数，只需定义满足性质(2)-(4)即可, 由它们可以推出(1).

多项分布（用于有放回抽样）

n次独立重复试验中每次实验可能结果为$A_1, A_2, \ldots, A_r$，$P(A_i) = p_i, i = 1, 2, \ldots, r$且$p_1 + \cdots + p_r = 1$，用$X_1, \ldots, X_r$分别记$A_1, A_2, \ldots, A_r$出现的次数，则

$$P\{X_1 = k_1, X_2 = k_2, \ldots, X_r = k_r\} = \frac{n!}{k_1!k_2!\cdots k_r!} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r},$$

其中$k_i \geq 0$且$k_1 + k_2 + \cdots + k_r = n$（不满足该约束时的概率为0）.

多元超几何分布（用于无放回抽样）

编号为i的球有N_i个，$i = 1, 2, \ldots, r$且$N_1 + N_2 + \cdots + N_r = N$. 从中随机抽取$n$个, 用$X_1, \ldots, X_r$分别记$1, 2, \ldots, r$号球的个数，则

$$P\{X_1 = k_1, X_2 = k_2, \ldots, X_r = k_r\} = \frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_r}{k_r}}{\binom{N}{n}},$$

其中$k_i \geq 0$且$k_1 + k_2 + \cdots + k_r = n$（不满足该约束时的概率为0）.

对于连续型多元分布$F(x_1, \ldots, x_n)$, 存在非负函数$p(x_1, \ldots, x_n)$, 对任意$x_i \in \mathbb{R}, i = 1, 2, \ldots, n$有

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p(y_1, \ldots, y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n.$$

其中$p(x_1, \ldots, x_n)$成为联合密度函数，它满足：(以下两条件是联合密度函数的充要条件)

1. $p(x_1, \ldots, x_n) \geq 0$（非负性）;

2. $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, \ldots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n$.

均匀分布

设E是$\mathbb{R}^n$中的有限区域, 其测度为$S > 0$， 则由以下密度函数得到的分布称为E上的均匀分布:

$$p(x_1, \ldots, x_n) = \begin{cases} \frac{1}{S}, & (x_1, \ldots, x_n) \in E \\ 0, & (x_1, \ldots, x_n) \notin E \end{cases}.$$

多元正态分布

$\mathbf{x}=(\mathbf{x_1},\ldots,\mathbf{x_n})^{\mathrm{T}}$, $\mu=(\mu_1,\ldots,\mu_n)^{\mathrm{T}}$, $\mathbf{\Sigma}=(\sigma_{ij})$是$n$阶正定(从而可逆)对称矩阵. 则$n$元正态分布$\mathrm{N}(\mu,\mathbf{\Sigma})$的密度函数为

$$p(\mathbf{x})=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\mathbf{\Sigma})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}.$$

特别地, 对于二元正态分布, $\mathbf{\Sigma}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_2\sigma_1 & \sigma_2^2\end{pmatrix}$, 有概率密度函数

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{x-\mu_2}{\sigma_2}\right)+\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\right\},$$

则称(X,Y)服从参数为$\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$的二维正态分布, 记作$(X,Y)\sim\mathrm{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$.

Property 3.11 若n元正态分布$\xi\sim\mathrm{N}(\mu,\mathbf{\Sigma})$, A是n阶可逆矩阵, 则

$$A\xi\sim\mathrm{N}(A\mu,A\mathbf{\Sigma}A^{\mathrm{T}}).$$

二维离散型分布的联合分布律: 设ξ取值$x_1,x_2,\ldots$, η取值$y_1,y_2,\ldots$, 记

$$p(x_i,y_j)=P\{\xi=x_i,\eta=y_j\},i=1,2,\ldots$$

$p(x_i,y_j)$称为(ξ,η)的**联合分布律**, 满足（联合分布律的充要条件）$p(x_i,y_j)\geq 0$和$\sum_{i,j}p(x_i,y_j)=1$. $p_1(x_i)=\sum_j p(x_i,y_j)$和$p_2(y_j)=\sum_i p(x_i,y_j)$分别称为$\xi$和$\eta$的**边际分布律**.

$F_1(x)=P\{\xi<x,\eta<+\infty\}=F(x,+\infty)$和$F_2(y)=P\{\xi<+\infty,\eta<y\}=F(+\infty,y)$称为$F(x,y)$的边际分布函数.

显然, 联合分布可以唯一确定边际分布, 但反之不能.

连续型随机变量(ξ,η)的联合分布函数为$F(x,y)$, 联合密度函数为$p(x,y)$, 则

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{x}p(u,v)\mathrm{d}u\mathrm{d}v,$$

$$p(x,y)=\frac{\partial F(x,y)}{\partial x\partial y},\text{a.e.}$$

$p_1(x)=\int_{-\infty}^{\infty}p(x,y)\mathrm{d}y$和$p_2(y)=\int_{-\infty}^{\infty}p(x,y)\mathrm{d}x$称为$p_1(x)$和$p_2(y)$的边际分布密度函数.

二元正态分布的边际分布仍是正态分布. 即

$$(\xi,\eta)\sim\mathrm{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)\Longrightarrow\xi\sim\mathrm{N}(\mu_1,\sigma_1^2),\eta\sim\mathrm{N}(\mu_2,\sigma_2^2).$$

但反之不成立, 有反例

$$p(x,y)=\frac{1+\sin x\sin y}{2\pi}\exp\left[-\frac{x^2+y^2}{2}\right].$$

条件分布函数和条件密度函数

$$F(y|x)=P\{\eta<y|\xi=x\}=\lim_{\Delta x\to 0}P\{\eta<y|x\le\xi<x+\Delta x\}$$
$$=\lim_{\Delta x\to 0}\frac{P\{\eta<y,x\le\xi<x+\Delta x\}}{P\{x\le\xi<x+\Delta x\}}=\lim_{\Delta x\to 0}\frac{F(x+\Delta x,y)-F(x,y)}{F(x+\Delta x,\infty)-F(x,\infty)}.$$

特别地, 对于有连续密度函数的分布, 有

$$F(y|x)=P\{\eta<y|\xi=x\}=\lim_{\Delta x\to 0}\frac{\int_x^{x+\Delta x}\left[\int_{-\infty}^{y}p(u,v)\mathrm{d}v\right]\mathrm{d}u}{\int_x^{x+\Delta x}\left[\int_{-\infty}^{\infty}p(u,v)\mathrm{d}v\right]\mathrm{d}u}.$$

在给定$\xi=x$且$p_1(x)\neq 0$的条件下, η的条件密度函数为$p(y|x)=\dfrac{p(x,y)}{p_1(x)}$, 在给定$\eta=y$且$p_2(y)\neq 0$的条件下, ξ的条件密度函数为$p(x|y)=\dfrac{p(x,y)}{p_2(y)}$.

计算可得二元正态分布的条件分布

$$\eta|\xi=x\sim\mathrm{N}(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\sigma_2^2(1-\rho^2)),$$

$$\xi|\eta=y\sim\mathrm{N}(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2),\sigma_1^2(1-\rho^2)).$$

Definition 3.12 设$\xi_1,\ldots,\xi_n$是概率空间$(\Omega,\mathcal{F},P)$上的n个随机变量, 如果它们的联合分布函数等于各自的边缘分布函数之积, 即

$$F(x_1,\ldots,x_n)=F_1(x_1)\cdots F_n(x_n),\forall x_i\in\mathbb{R},i=1,\ldots,n,$$

则称$\xi_1,\ldots,\xi_n$**相互独立**. 一族**无限**多个随机变量称为是相互独立的, 如果其中**任意有限个**是相互独立的.

独立性的等价定义:

1. 随机变量ξ与η相互独立当且仅当

$$P\{\xi\in B_1,\eta\in B_2\}=P\{\xi\in B_1\}P\{\eta\in B_2\},\quad\forall B_1,B_2\in\mathfrak{B}(\mathbb{R}^1).$$

2. 称n维随机向量ξ和m维随机向量η相互独立, 若

$$P\{\xi\in A,\eta\in B\}=P\{\xi\in A\}P\{\eta\in B\},\quad\forall A\in\mathfrak{B}(\mathbb{R}^n),B\in\mathfrak{B}(\mathbb{R}^m).$$

显然, 若ξ与η相互独立, 则ξ的子向量与η的子向量相互独立.

3. 对于离散型随机变量(ξ,η), ξ与η独立当且仅当

$$p(x_i,y_j)=p(x_i)p(y_j),\quad \forall i,j=1,2,\dots$$

4. 对于连续型随机变量(ξ,η), ξ与η独立当且仅当

$$p(x,y)=p_1(x)p_2(y),\quad \text{a.e.}\ \ x,y\in\mathbb{R}.$$

若ξ与η相互独立, 则条件分布转化为无条件分布, 即$P\{\eta<y|\xi=x\}=P\{\eta<y\}$.

二元正态随机变量相互独立$\Longleftrightarrow\rho=0$.（该结论对其他情况一般不成立）

3.3 Functions of Random Variables

Borel(可测)函数: "Borel集的原象是Borel集". 有n个变元的Borel函数称为n元Borel函数. 连续函数和单调函数一定是Borel函数.

Proposal 3.13 设ξ是概率空间$(\Omega,\mathcal{F},P)$上的随机变量, $g(x)$是一元*Borel*函数, 则$g(\xi)$是$(\Omega,\mathcal{F},P)$上的随机变量.

设$\xi_1,\dots,\xi_n$是概率空间$(\Omega,\mathcal{F},P)$上的n个随机变量, $g(\xi_1,\dots,\xi_n)$是n元*Borel*函数, 则$g(x_1,\dots,x_n)$是$(\Omega,\mathcal{F},P)$上的随机变量.

离散型随机变量的函数的分布: 若ξ的分布列为

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix},$$

则$\varphi(\xi)$的分布列为（相等的$\varphi(x_i)$可合并）

$$\begin{pmatrix} \varphi(x_1) & \varphi(x_2) & \cdots & \varphi(x_n) & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix}.$$

离散卷积公式（独立随机变量和的分布）：设ξ和η是相互独立的随机变量, 都取非负整数值, 则

$$P(\xi+\eta=r)=\sum_{k=0}^{r}P(\xi=k)P(\eta=r-k).$$

二项分布的加法定理: 若$\xi\sim B(m,p)$与$\eta\sim B(n,p)$相互独立, 则$\xi+\eta\sim B(m+n,p)$.（从二项分布的意义很好理解）

泊松分布的加法定理: 若$\xi\sim P(\lambda_1)$与$\eta\sim P(\lambda_2)$相互独立, 则$\xi+\eta\sim P(\lambda_1+\lambda_2)$.

单个连续型随机变量的函数的分布：随机变量ξ的分布函数为$F(x)$或密度函数为$p(x)$，要求$\eta = g(\xi)$的分布函数$G(y)$或密度函数$q(y)$，易知

$$G(y) = P\{\eta < y\} = P\{g(\xi) < y\} = \int\limits_{\{x:g(x)<y\}} p(x)\mathrm{d}x.$$

在求解时通常用直接法或变换法.

直接法：通过把$\{g(\xi) < y\}$直接转化为关于ξ的等价事件而求得η的分布函数或密度函数.

一般的处理方法：设连续型随机变量ξ有密度函数$p(x)$, 定义随机变量$\eta = g(\xi)$, 则η的分布函数为

$$P(\eta < y) = P(g(\xi) < y) = \int\limits_{\{x:g(x)<y\}} p(x)\mathrm{d}x.$$

分两种情况讨论：

1. 若$g(x)$严格单调, 且反函数$g^{-1}(y)$有连续导函数, 则$\eta = g(\xi)$是具有密度函数为 $p\left[g^{-1}(y)\right] \cdot \left|\left[g^{-1}(y)\right]'\right|$ 的连续型随机变量.

2. 若$g(x)$在不重叠的区间$I_1, I_2, \ldots$上逐段严格单调, 反函数分别为$h_1(y), h_2(y), \ldots$, 且都有连续导数, 则$\eta = g(\xi$是连续型随机变量, 且密度函数为

$$p\left[h_1^{-1}(y)\right] \cdot \left|\left[h_1^{-1}(y)\right]'\right| + p\left[h_2^{-1}(y)\right] \cdot \left|\left[h_2^{-1}(y)\right]'\right| + \cdots$$

注; 对于反函数无意义的点y, 密度函数定义为0.

例：

1. 设连续型随机变量ξ有密度函数$p_1(x)$, 定义随机变量$\eta = a + b\xi (b \neq 0)$, 则$\eta$的密度函数为$p_2(y) = \dfrac{1}{|b|} p_1\left(\dfrac{y-a}{b}\right)$.

2. $\xi \sim U(c,d) \Longrightarrow a + b\xi \sim U(a+bc, a+bd)$.

3. $\xi \sim \Gamma(\lambda, \alpha), \quad b > 0 \Longrightarrow b\xi \sim \Gamma(\dfrac{\lambda}{b}, \alpha)$.

4. $\xi \sim \mathrm{N}(\mu, \sigma^2) \Longrightarrow a + b\xi \sim \mathrm{N}(a + b\mu, b^2\sigma^2)$.

5. （随机变量平方的分布）设ξ有密度函数$p_1(x)$, 定义$\eta = \xi^2$， 则η的密度函数为

$$p_2(y) = \frac{1}{2\sqrt{y}}\left[p_1(\sqrt{y}) + p_1(-\sqrt{y})\right], y > 0.$$

分布函数的单调逆函数

Definition 3.14 分布函数$F(x)$的**单调逆**定义为

$$F^{-1}(y) = \inf\{x : F(x) > y\}, \quad y \in [0,1],$$

等价于

$$F^{-1}(y) = \sup\{x : F(x) \le y\}, \quad y \in [0,1].$$

注: 这是以课本对分布函数定义左连续性延伸出来的, 如果本对分布函数定义右连续性, 则上面两式中严格不等号换为非严格不等号, 非严格不等号换为严格不等号.

Property 3.15 （分布函数的单调逆函数的性质）

1. 当$F(x)$时严格递增的连续函数时, 它的单调逆就是它的反函数;
2. 是单调(非严格)递增函数;
3. $F^{-1}(y)$在$\forall y \in (0,1)$处均右连续;
4. $F(F^{-1}(y)) \le y, \quad \forall y \in (0,1)$, 当$F(x)$在点$F^{-1}(y)$连续时等号成立, 但逆命题不成立.
5. $F^{-1}(F(x)) \ge x, \quad \forall x \in \mathbb{R}$;
6. 对$\forall x \in \mathbb{R}$和$\forall y \in (0,1)$, 有

$$F^{-1}(y) < x \Longleftrightarrow y < F(x).$$

均匀分布的特殊地位:

1. 设随机变量ξ的分布函数为$F(x)$, 且$F(x)$为连续函数, 则随机变量$\theta = F(\xi)$服从$[0,1]$上的均匀分布. 因为对$\forall y \in [0,1]$, 有

$$P\{\theta < y\} = P\{F(\xi) < y\} = P\{\xi < F^{-1}(y)\} = F(F^{-1}(y)) = y.$$

2. 设$F(x)$为分布函数, ξ服从$[0,1]$上的均匀分布, 则随机变量$\eta = F^{-1}(\xi)$的分布函数是$F(x)$. 因为对$\forall x \in \mathbb{R}$, 有

$$P\{\xi < x\} = P\{F^{-1}(\theta) < x\} = P\{\theta < F(x)\} = F(x).$$

因此, 对于任意分布函数$F(x)$, 都可以构造一个概率空间$(\Omega, \mathcal{F}, P)$和定义在它上的随机变量η, 使得η以$F(x)$作为分布函数.

Theorem 3.16 （随机变量的存在性定理） 若函数$F(x)$满足分布函数的三个性质, 则存在一个概率空间$(\Omega, \mathcal{F}, P)$和定义在它上的随机变量$\xi(\omega)$, 使ξ的分布函数恰好为$F(x)$. （证明见 *[1]165*页）

Theorem 3.17 （随机变量和差积商的分布）
设(ξ_1, ξ_2)有联合密度函数$p(x_1, x_2)$, 则

1. 和的分布（卷积公式）：$\eta = \xi_1 + \xi_2$的概率密度是
$$p_\eta(z) = \int_{-\infty}^{\infty} p(x_1, z - x_1)\mathrm{d}x_1 = \int_{-\infty}^{\infty} p(z - x_2, x_2)\mathrm{d}x_2;$$
特别地, 当ξ_1与ξ_2独立时有
$$p_\eta(z) = \int_{-\infty}^{\infty} p_1(x_1)p_2(z - x_1)\mathrm{d}x_1 = \int_{-\infty}^{\infty} p_1(z - x_2)p_2(x_2)\mathrm{d}x_2;$$

2. 商的分布：$\eta = \dfrac{\xi_1}{\xi_2}$的概率密度是
$$p_\eta(z) = \int_{-\infty}^{\infty} |x_2| p(zx_2, x_2)\mathrm{d}x_2;$$

3. 差的分布：$\eta = \xi_1 - \xi_2$的概率密度是
$$p_\eta(z) = \int_{-\infty}^{\infty} p(x_1, x_1 - z)\mathrm{d}x_1 = \int_{-\infty}^{\infty} p(x_2 + z, x_2)\mathrm{d}x_2;$$

4. 乘积的分布：$\eta = \xi_1\xi_2$的概率密度是
$$p_\eta(z) = \int_{-\infty}^{\infty} \frac{1}{|x_1|} p\left(x_1, \frac{z}{x_1}\right)\mathrm{d}x_1 = \int_{-\infty}^{\infty} \frac{1}{|x_2|} p\left(\frac{z}{x_2}, x_2\right)\mathrm{d}x_2.$$

Theorem 3.18 （顺序统计量的分布）

1. （最大值和最小值的分布） 设$X_1, \ldots, X_n$是相互独立的n个随机变量, 它们的分布函数为$F_1(x), \ldots, F_n(x)$, 则$\max X_i$和$\min X_i$的分布函数分别为
$$F_{\max}(z) = \prod_{i=1}^{n} F_i(z), \quad F_{\min}(z) = 1 - \prod_{i=1}^{n}[1 - F_i(z)].$$
特别地, 当$X_1, \ldots, X_n$独立同分布*(*分布函数为$F(x)$*)*时, 有
$$F_{\max}(z) = [F(z)]^n, \quad F_{\min}(z) = 1 - [1 - F(z)]^n.$$

2. （一般顺序统计量的分布） 设$\xi_1, \xi_2, \ldots, \xi_n$是定义在同一概率空间$(\Omega, \mathcal{F}, P)$上的**独立同分布**（分布函数为$F(x)$）的$n$个随机变量, $\xi_{(n)}$与$\xi_{(1)}$分别是最大和最小的顺序统计量*(*按升序*)*, 则
$$P\{\xi_{(n)} < x\} = P\{\max_i \xi_i < x\} = P\{\xi_1 < x, \ldots, \xi_n < x\} = \prod_{i=1}^{n} P\{\xi_i < x\} = [F(x)]^n,$$

$$P\{\xi_{(1)} \geq x\} = P\{\min_i \xi_i \geq x\} = P\{\xi_1 \geq x, \ldots, \xi_n \geq x\} = \prod_{i=1}^{n} P\{\xi_i \geq x\} = [1-F(x)]^n,$$

因此

$$P\{\xi_{(1)} < x\} = 1 - [1-F(x)]^n.$$

特别地, 又若$\{\xi_i\}_{i=1}^n$是连续型随机变量, 有相同的密度函数$p(x)$, 则$\xi_{(1)}$与$\xi_{(n)}$的密度函数分别为$p_1(x) = np(x)\,[1-F(x)]^n$和$p_n(x) = np(x)\,[F(x)]^n$.

一般地, $\xi_{(k)}$的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!} p(x)\,[F(x)]^{k-1}\,[1-F(x)]^{n-k}.$$

连续型随机向量的变换

设$(\xi_1, \ldots, \xi_n)$的密度函数为$p(x_1, \ldots, x_n)$,则$\eta_i = f_i(\xi_1, \ldots, \xi_n), i = 1, \ldots, m$的分布为

$$F(y_1, \ldots, y_m) = P\{\eta_1 < y_1, \ldots, \eta_m < y_m\} = \int\cdots\int_{\substack{f_i(x_1,\ldots,x_n)<y_i \\ i=1,\ldots,m}} p(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

$m = 1$时对应随机向量的情形, $m = n = 1$时对应单个随机变量的函数的情形.
以下考虑$m = n$且$(\xi_1, \ldots, \xi_n)$与$(\eta_1, \ldots, \eta_n)$有一一对应的变换时的特殊情形.

设$y_i = g_i(x_1, \ldots, x_n)$存在唯一的反函数$x_i = x_i(y_1, \ldots, y_n), i = 1, \ldots, n$, 且$(\eta_1, \ldots, \eta_n)$的密度函数为$q(y_1, \ldots, y_n)$, 则有

$$F(y_1, \ldots, y_n) = \int\cdots\int_{\substack{u_i<y_i \\ i=1,\ldots,n}} q(u_1, \ldots, u_n)\mathrm{d}u_1 \cdots \mathrm{d}u_n,$$

和

$$q(y_1, \ldots, y_n) = |J| p(x_1(y_1, \ldots, y_n), \ldots, x_n(y_1, \ldots, y_n)),\ \ (y_1, \ldots, y_n) \in \mathrm{Im}(g_1, \ldots, g_n).$$

其中J是坐标变换的Jacobi行列式$\left|\dfrac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)}\right|$（假定偏导数存在且连续）.

通过**增补变量**使用变换法, 见书173页.

随机变量的函数的独立性

Theorem 3.19 若$\xi_1, \ldots, \xi_n$是相互独立的随机变量, 则它们各自的函数$f_1(\xi_1), \ldots, f_n(\xi_n)$也是相互独立的, 其中$f_i (i = 1, \ldots, n)$是任意一元*Borel*函数.（证明见书*174*页）

相同的随机向量构成的不同函数也可能是独立的.（见书175页）

顺序统计量的联合分布：若$X_1, \ldots, X_n$独立同分布, 密度函数为$f(x), a < x < b$, 则其顺序统计量$Y_1, \ldots, Y_n$的联合密度函数为

$$g(y_1, \ldots, y_n) = \begin{cases} n! f(y_1) \cdots f(y_n), & a < y_1 < \cdots < y_n < b \\ 0, & \text{elsewhere} \end{cases}.$$

4 Characteristic Functions

4.1 Mathematical Expectation

数字特征是由随机变量决定的一些常数, 期望与方差是最重要的两个特征, 它们能刻画随机变量的部分性质.

离散型随机变量的数学期望：设X的分布律为$P(X = x_i) = p_i, \quad i = 1,2,\ldots$, 如果级数$\sum_{i=1}^{+\infty}|x_i|p_i$收敛, 则称级数$\sum_{i=1}^{+\infty}x_ip_i$的值为随机变量$X$的数学期望, 记作$E(X)$, 即$E(X) = \sum_{i=1}^{+\infty}x_ip_i$.（要求绝对收敛是为了保证期望值不受求和顺序的影响.）

连续型随机变量的数学期望：设X的密度函数$p(x)$满足$\int_{-\infty}^{+\infty}|x|p(x)\mathrm{d}x < \infty$, 则随机变量$X$的数学期望定义为$E(X) = \int_{-\infty}^{+\infty}xp(x)\mathrm{d}x$.

Definition 4.1 （*Riemann-Stieltjes*积分） 设$F(x)$是$\mathbb{R}$上的单调*(*非严格*)*递增的左*(*右*)*连续函数, $g(x)$是$\mathbb{R}$上的单值实函数, 对于区间$[a,b]$任取分点$a = x_0 < x_1 < \cdots < x_n = b$, $\forall u_i \in [x_{i-1}, x_i), i = 1,2,\ldots,n$, 作和式

$$\sum_{i=1}^{n} g(u_i)\Delta F(x_i) = \sum_{i=1}^{n} g(u_i)[F(x_i) - F(x_{i-1})].$$

令$\lambda = \max_{1\le i\le n}\Delta x_i = \max_{1\le i\le n}(x_i - x_{i-1})$, 若极限$\lim_{\lambda\to 0}\sum_{i=1}^{n} g(u_i)\Delta F(x_i)$存在, 则记

$$\int_a^b g(x)\mathrm{d}F(x) = \lim_{\lambda\to 0}\sum_{i=1}^{n} g(u_i)\Delta F(x_i),$$

称为$g(x)$关于$F(x)$在区间$[a,b]$上的*Riemann-Stieltjes*积分（简称*R-S*积分）. 特别地, 当$F(x) = x$时, *R-S*积分就是*Riemann*积分.

$F(x)$是分布函数时, $\int_a^b 1\mathrm{d}F(x) = F(b) - F(a) = P\{a \le X < b\}$.

若X是离散型随机变量, $P(X = c_i) = p_i, i = 1,2,\ldots$, 则$F(x) = \sum_{c_i<x}p_i$是一个跳跃分布函数, $F(x)$仅在$c_1, c_2, \ldots$点作p_i的变化, $\int_a^b g(x)\mathrm{d}F(x) = \sum_{i=1}^{\infty} g(c_i)p_i$.

Definition 4.2 （数学期望的一般定义） 若随机变量X的分布函数$F(x)$满足

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

则X的数学期望定义为

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

否则称X的数学期望不存在.

若随机变量$g(X)$的数学期望为$I = \int_{-\infty}^{+\infty}g(x)\mathrm{d}F(x)$, 则

1. $F(x)$在x_k处有跳跃度p_k时, 化为级数$I=\sum_{k=-\infty}^{+\infty} g(x_k)p_k$;

2. $F(x)$存在导数$p(x)$时, 化为*Riemann*积分$I=\int_{-\infty}^{+\infty} g(x)p(x)\mathrm{d}x$.

Theorem 4.3 设ξ是随机变量, $g(x)$是一元*Borel*函数, 定义随机变量$\eta=g(\xi)$, 则有

$$E(\eta)=E[g(\xi)]=\int_{-\infty}^{+\infty} y\mathrm{d}F_\eta(y)=\int_{-\infty}^{+\infty} g(x)\mathrm{d}F_\xi(x).$$

注：由此定理, 不必计算新的随机变量的分布. 定理证明要用测度论, 超纲.

根据$E[g(\xi)]=\int_{-\infty}^{+\infty} g(x)\mathrm{d}F_\xi(x)$,

1. 在离散型场合可化为

$$E[g(\xi)]=\sum_{i=1}^{\infty} g(x_i)p_i;$$

2. 在连续型场合, 设ξ有密度函数$p(x)$, 则

$$E[g(\xi)]=\int_{-\infty}^{+\infty} g(x)p(x)\mathrm{d}x.$$

Definition 4.4 （随机向量函数的期望） 设$F(x_1,\dots,x_n)$是随机向量$(\xi_1,\dots,\xi_n)$的联合分布函数, $g(x_1,\dots,x_n)$是n元*Borel*函数, 定义$\eta=g(\xi_1,\dots,\xi_n)$, 则

$$E[g(\xi_1,\dots,\xi_n)]=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} g(x_1,\dots,x_n)\mathrm{d}F(x_1,\dots,x_n).$$

Property 4.5 （独立乘积的期望等于期望的乘积）
若随机变量$\xi_1,\dots,\xi_n$相互独立, 则

$$E(\xi_1\times\cdots\times\xi_n)=E(\xi_1)\times\cdots\times E(\xi_n).$$

4.2 Correlation Coefficients and Moments

Definition 4.6 （方差） 设X是随机变量, 若$E\{[X-E(X)]^2\}$存在, 则称$E\{[X-E(X)]^2\}$为X的方差, 记为$\mathrm{Var}(X)$或$\mathrm{D}(X)$. X的标准差定义为$\sigma(X)=\sqrt{\mathrm{Var}(X)}$. 方差反映了随机变量的取值于平均值的偏离程度.
计算方差的常用公式：$\mathrm{Var}(X)=E(X^2)-[E(X)]^2$. 易知$E(X^2)\geq[E(X)]^2$.

Property 4.7

1. $E(aX+bY)=aE(X)+bE(Y)$;
2. $\mathrm{Var}(aX+b)=a^2\mathrm{Var}(X)$;
3. 若X与Y**独立**, 则$\mathrm{Var}(X\pm Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$;
4. 若$X_1,\ldots,X_n$**相互独立**, 则
$$\mathrm{Var}\left[\sum_{i=1}^{n}a_iX_i\right]=\sum_{i=1}^{n}a_i^2\mathrm{Var}(X_i);$$
5. $\mathrm{Var}(X)=0\iff$ 存在常数C使 $P\{X=C\}=1$;
6. $\mathrm{Var}(X)\le E[(X-c)^2]$, 等号成立当且仅当$c=E(X)$（说明$E(X)$在某种意义下是极小的）.

Bernoulli分布	$E\xi=p$	$D\xi=pq$
二项分布	$E\xi=np$	$D\xi=npq$
几何分布	$E\xi=\frac{1}{p}$	$D\xi=\frac{1-p}{p^2}$
Pascal分布	$E\xi=\frac{r}{p}$	$D\xi=\frac{r(1-p)}{p^2}$
Poisson分布	$E\xi=\lambda$	$D\xi=\lambda$
均匀分布	$E\xi=\frac{a+b}{2}$	$D\xi=\frac{(b-a)^2}{12}$
Gamma分布	$E\xi=\frac{\alpha}{\lambda}$	$D\xi=\frac{\alpha}{\lambda^2}$
指数分布	$E\xi=\frac{1}{\lambda}$	$D\xi=\frac{1}{\lambda^2}$
卡方分布	$E\xi=n$	$D\xi=2n$
正态分布	$E\xi=\mu$	$D\xi=\sigma^2$

Figure 1: 常见分布的期望和方差

Definition 4.8 设 $E(X)=\mu$ 和 $\mathrm{Var}(X)=\sigma^2>0$ 都存在, 则随机变量 $\eta=\dfrac{X-\mu}{\sigma}$ 称为X的**中心标准化**.（效果: 期望化为0, 方差化为1）

Theorem 4.9 （*Chebyshev*不等式） 若随机变量X有有限的方差（从而有有限的期望）, 则对$\forall\varepsilon>0$, 都有
$$P\{|X-E(X)|\ge\varepsilon\}\le\frac{\mathrm{Var}(X)}{\varepsilon^2},$$

或等价地

$$P\{\frac{X-E(X)}{\sqrt{\mathrm{Var}(X)}} \geq \varepsilon\} \leq \frac{1}{\varepsilon^2}.$$

可见, 方差越小, 随机变量的偏离程度也越小, 因此方差适合描述随机变量与其期望值的偏离程度.

推论: 方差为零的随机变量几乎处处是常数.

Definition 4.10 （协方差） (X,Y)是二维随机变量, 若$E\{[X-E(X)][Y-E(Y)]\}$存在, 则称它为X与Y的**协方差**, 记为$\mathrm{Cov}(X,Y)$, 即

$$\mathrm{Cov}(X,Y) = E\{[X-E(X)][Y-E(Y)]\}.$$

满足$a_{ij} = \mathrm{Cov}(X_i, X_j), i,j = 1,\ldots,n$的方阵$A = (a_{ij})$称为$(X_1,\ldots,X_n)$的**协方差矩阵**, 它是半正定对称矩阵.

Definition 4.11 （相关系数） 随机变量X与Y的相关系数定义为

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}},$$

也就是$\dfrac{X-E(X)}{\sqrt{\mathrm{Var}(X)}}$与$\dfrac{Y-E(Y)}{\sqrt{\mathrm{Var}(Y)}}$的协方差. 规定**常数**与任何随机变量的相关系数都为**零**（补充定义）.

Property 4.12

1. 若X与Y独立, 则$\mathrm{Cov}(X,Y)=0$, 逆命题不成立;
2. $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$（常用于计算协方差）;
3. $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X,Y)$, a, b为常数;
4. $\mathrm{Cov}(X_1+X_2, Y) = \mathrm{Cov}(X_1,Y) + \mathrm{Cov}(X_2,Y)$;
5. $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\mathrm{Cov}(X,Y)$（常用于计算协方差）;
6. 根据协方差矩阵（对角线部分和非对角线部分）易知

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{1\leq i<j\leq n} \mathrm{Cov}(X_i, X_j).$$

Theorem 4.13 （*Cauchy-Schwarz*不等式）
若随机变量X与Y满足$E(X^2), E(Y^2) < +\infty$, 则有

$$|E(XY)| \leq E(X^2)E(Y^2),$$

等式成立当且仅当存在不全为零的常数a, b使$P\{aX+bY=0\}=1$.
（注: 教材中等式成立的充要条件是错的）

由Cauchy-Schwarz不等式知, 两个随机变量的相关系数的绝对值小于等于1.

Definition 4.14 （相关性） 若随机变量X与Y的相关系数$\rho_{XY} = 0$, 则称X与Y**不相关**.

Property 4.15 对于随机变量X与Y, 以下命题等价:

1. X与Y不相关;
2. $\mathrm{Cov}(X,Y) = 0$;
3. $E(XY) = E(X)E(Y)$;
4. $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Property 4.16 X与Y独立 $\Longrightarrow$ X与Y不相关. 反之一般不成立（如X与X^2不相关但不独立）. 相关系数ρ_{XY}只反映了X与Y的线性关系, 而无法反映其他关系.

Property 4.17

1. 对于二元正态分布, 不相关性与独立性等价. 可推广到多元场合: 正态随机变量$X_1, \ldots, X_n$相互独立的充要条件是它们两两不相关.
2. 若X和Y都是取二值的随机变量, 则它们之间的不相关性与独立性等价.

Definition 4.18 （矩）设k为正整数, 则 称$m_k = E(X^k)$ 为k阶**原点矩**, 称 $c_k = E[(X - EX)^k]$ 为k阶**中心矩**.

易知, 数学期望是一阶原点矩, 方差是二阶中心矩, 协方差是二阶混合中心矩.

原点矩和中心矩可以相互表示:

$$c_k = \sum_{i=0}^{k} \binom{k}{i} (-m_1)^{k-i} m_i, \quad m_k = \sum_{i=0}^{k} \binom{k}{i} c_{k-i} m_1^i.$$

正态分布$\mathrm{N}(\mu, \sigma^2)$的$k$阶中心矩：$c_k = 0$, 若$k$为奇数; $c_k = \sigma^k (k-1)(k-3)\cdots 3 \cdot 1$, 若$k$为偶数. 其原点矩可由各阶中心矩算出.

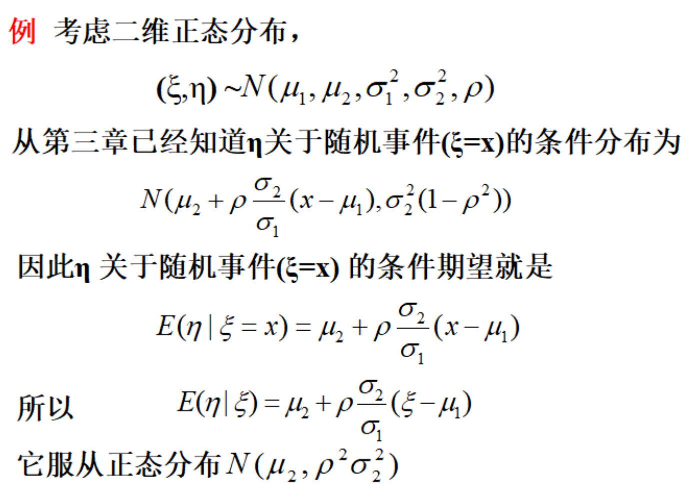
Definition 4.19 （分位数） 对于$\forall p \in (0,1)$, 若$F(x_p) \le p \le F(x_p + 0)$, 则称$x_p$为分布函数$F(x)$的$p$**分位数**. 特别地, $x_{0.5}$称为中位数.
注: 由于分布函数不一定连续, 所以对$\forall p \in (0,1)$, x_p不一定存在, 即使存在也不一定唯一.

Definition 4.20 （条件数学期望）

1. 离散随机变量的条件期望: 随机变量Y关于随机事件$\{X=x\}$的条件期望为
$$E(Y|X=x)=\sum_j y_j P\{Y=y_j|X=x\},$$
它反映了Y的平均值对X的依赖. Y关于X的条件期望$E(Y|X)$也是一个随机变量, 它取值为$E(Y|X=x)$的概率是$P(X=x)$.

2. 连续随机变量的条件期望: 随机变量Y关于随机事件$\{X=x\}$的条件期望为
$$E(Y|X=x)=\int_{-\infty}^{+\infty} yp(y|x)\mathrm{d}y.$$
Y关于X的条件期望$E(Y|X)$也是一个随机变量, 它取值为$E(Y|X=x)$的概率密度是$p_X(x)$.

条件数学期望具有数学期望的所有性质.


例 考虑二维正态分布，

$$(\xi,\eta)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$$

从第三章已经知道η关于随机事件(ξ=x)的条件分布为

$$N(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\sigma_2^2(1-\rho^2))$$

因此η 关于随机事件(ξ=x) 的条件期望就是

$$E(\eta\,|\,\xi=x)=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$$

所以 $E(\eta|\xi)=\mu_2+\rho\dfrac{\sigma_2}{\sigma_1}(\xi-\mu_1)$

它服从正态分布$N(\mu_2,\rho^2\sigma_2^2)$

Figure 2: 二元正态分布的条件期望

由上例可知, 在二元正态分布的场合, $E(\eta|\xi)$是随机变量ξ的线性函数, 因此$E(\eta|\xi)$确实是随机变量.

注意区分$E(Y|X)$和$E(Y|X=x)$, 前者才是随机变量.

Theorem 4.21 （全期望公式） 设二维随机向量(X,Y)的$E(X)$存在, 则
$$E[E(X|Y)]=E(X).$$

注: 在高等概率论中, 全期望公式是作为条件期望的定义.

推论:

$$E(X)=\sum_j E(X|Y=y_j)P(Y=y_j)\text{（离散型）},$$

$$E(X)=\int_{-\infty}^{+\infty}E(X|Y=y)f_Y(y)\mathrm{d}y\text{（连续型）}.$$

Property 4.22 （条件期望的一些性质）

1. $\mathrm{Var}(X)=\mathrm{Var}[E(X|Y)]+E[\mathrm{Var}(X|Y)]$（重要, 很有用）*;*

2. $\mathrm{Var}(X)\geq\mathrm{Var}[E(X|Y)]$ *(Rao-Blackwell).*

4.3 Characteristic Functions

引入特征函数的动机: 能完全决定分布函数, 又比分布函数有更好的分析性质.

Definition 4.23 （复随机变量） 设X与Y都是概率空间$(\Omega,\mathcal{F},P)$上的实值随机变量, 则称$Z=X+iY$为**复随机变量**.

若二维随机向量 (X_1,Y_1)与(X_2,Y_2) 相互独立, 则称复随机变量 $Z_1=X_1+iY_1$与$Z_2=X_2+iY_2$ 相互独立. 复随机变量 $Z=X+iY$ 的数学期望定义为 $E(Z)=E(X)+iE(Y)$.

设$g(x)$是Borel函数, $Y=g(X)$, 则

$$E[e^{itY}]=E[e^{itg(X)}]=\int_{-\infty}^{+\infty}e^{itg(X)}\mathrm{d}F_X(x).$$

Definition 4.24 （特征函数） 设随机变量X的分布函数为$F_X(x)$, 则称

$$f_X(t)=E[e^{itX}]=\int_{-\infty}^{+\infty}e^{itx}\mathrm{d}F_X(x),\quad\forall t\in\mathbb{R}$$

为X或$F_X(x)$的**特征函数**, 它是一个实变量的复值函数. 特别地,

1. 对于离散型随机变量, 若其分布列为

$$\begin{pmatrix}x_1 & x_2 & \cdots & x_n & \cdots\\ p_1 & p_2 & \cdots & p_n & \cdots\end{pmatrix},$$

则其特征函数为

$$f(t)=\sum_{j=1}^{\infty}p_je^{itx_j};$$

2. 对于连续型随机变量, 若其密度函数为$p(x)$, 则其特征函数为

$$f(t) = \int_{-\infty}^{+\infty} e^{itx} p(x) \mathrm{d}x,$$

此时特征函数是密度函数$p(x)$的Fourier变换.

Property 4.25 随机变量X的特征函数$f(t)$有以下性质:

1. $f(t)$为实函数 $\iff$ X与$-X$同分布;

2. $|f(t)| \le f(0) = 1$, $f(-t) = \overline{f(t)}$;

3. $f(t)$在$\mathbb{R}$上一致连续;

4. 对 $\forall n \in \mathbb{Z}_+$ 和 $\forall t_1, \dots, t_n \in \mathbb{R}$ 和 $\forall \lambda_1, \dots, \lambda_n \in \mathbb{C}$, 有

$$\sum_{k=1}^{n} \sum_{j=1}^{n} f(t_k - t_j) \lambda_k \overline{\lambda_j} \ge 0,$$

该性质称为特征函数的**非负定性**;

5. 两个**相互独立**的随机变量之**和**的特征函数等于它们的特征函数之**积**:

$$f_{\xi_1+\xi_2}(t) = E[e^{it(\xi_1+\xi_2)}] = E[e^{it\xi_1}]E[e^{it\xi_2}] = f_{\xi_1}(t) f_{\xi_2}(t),$$

可推广到n个独立随机变量之和:

$$f_{\sum_{i=1}^{n} \xi_i}(t) = \prod_{i=1}^{n} f_{\xi_i}(t).$$

独立和的分布函数要通过卷积运算才能得到, 而用特征函数处理独立和更简单;

6. 设随机变量ξ的n阶矩存在, 则它的特征函数可微分n次, 且当$k \le n$时有

$$f^{(k)}(0) = i^k E(\xi^k).$$

由此性质容易求得随机变量的各阶矩; 此外它的特征函数可作展开:

$$f(t) = 1 + (it)E(\xi) + \frac{(it)^2}{2!} E(\xi^2) + \cdots + \frac{(it)^n}{n!} E(\xi^n) + o(t^n);$$

7. 设$\eta = a\xi + b$, 其中a, b为常数, 则有

$$f_\eta(t) = e^{ibt} f_\xi(at);$$

8. 一元正态分布N(μ,σ^2)的特征函数为

$$f(t)=\exp\left(i\mu t-\frac{1}{2}\sigma^2t^2\right),$$

用特征函数容易求得N(μ,σ^2)的期望和方差分别为μ和σ^2.

Theorem 4.26 （逆转公式）
设分布函数$F(x)$的特征函数为$f(t)$, x_1,x_2是$F(x)$的连续点, 则有

$$F(x_2)-F(x_1)=\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-itx_1}-e^{-itx_2}}{it}f(t)\mathrm{d}t.$$

Theorem 4.27 （唯一性定理） 分布函数由其特征函数唯一确定.
该定理对多元场合也成立. 由此, 特征函数可完整地描述一个随机变量.

Corollary 4.28 设F和G是两个分布函数, 若对 $\forall t\in\mathbb{R}$ 都有

$$\int_{-\infty}^{+\infty}e^{itx}\mathrm{d}F(x)=\int_{-\infty}^{+\infty}e^{itx}\mathrm{d}G(x),$$

则$F=G$.

Theorem 4.29 若特征函数$f(t)$绝对可积（即$\int_{-\infty}^{+\infty}|f(t)|\,\mathrm{d}t<\infty$）, 则相应分布函数$F(x)$的导数存在且连续, 且有

$$p(x)=F'(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-itx}f(t)\mathrm{d}t.$$

在$f(t)$绝对可积的条件下, 密度函数$p(x)$与特征函数$f(t)$通过Fourier变换联系.

Definition 4.30 （分布函数的再生性） 若两个具有同类型分布的独立随机变量之和的分布仍是该类型的分布, 且对应的参数等于这两个随机变量的相应参数之和, 则称该分布具有**再生性**.

例: 证明以下分布的再生性可利用特征函数的性质:

$$\xi_1\perp\xi_2\Longrightarrow f_{\xi_1+\xi_2}(t)=f_{\xi_1}(t)f_{\xi_2}(t).$$

1. 二项分布

$$B(n_1,p)\oplus B(n_2,p)\sim B(n_1+n_2,p),$$

$$\text{因为}\ (pe^{it}+q)^{n_1}\cdot(pe^{it}+q)^{n_2}=(pe^{it}+q)^{n_1+n_2};$$

2. 泊松分布

$$P(\lambda_1)\oplus P(\lambda_2)\sim P(\lambda_1+\lambda_2),$$
$$\text{因为 } e^{\lambda_1(e^{it}-1)}\cdot e^{\lambda_2(e^{it}-1)}=e^{\lambda_1+\lambda_2(e^{it}-1)};$$

3. 正态分布

$$N(\mu_1,\sigma_1^2)\oplus N(\mu_2,\sigma_2^2)\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2),$$
$$\text{因为 } e^{i\mu_1 t-\frac{1}{2}\sigma_1^2t^2}\cdot e^{i\mu_2 t-\frac{1}{2}\sigma_2^2t^2}=e^{i(\mu_1+\mu_2)t-\frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2};$$

4. 伽马分布

$$\Gamma(r_1,\lambda)\oplus\Gamma(r_2,\lambda)\sim\Gamma(r_1+r_2,\lambda),$$
$$\text{因为 } \left(1-\tfrac{it}{\lambda}\right)^{-r_1}\cdot\left(1-\tfrac{it}{\lambda}\right)^{-r_2}=\left(1-\tfrac{it}{\lambda}\right)^{-(r_1+r_2)}.$$

分布函数的分解问题（再生性问题的逆问题）：两个独立随机变量之和服从某一分布, 能否判断这两个随机变量也都服从该分布? 二(多)项分布、泊松分布和正态分布的可分解性已被证明.

Definition 4.31 （多元特征函数） 设随机向量$(\xi_1,\ldots,\xi_n)$的分布函数为$F(x_1,\ldots,x_n)$, 定义它的特征函数为

$$f(t_1,\ldots,t_n)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\exp[i(t_1x_1+\cdots+t_nx_n)]\mathrm{d}F(x_1,\ldots,x_n).$$

Property 4.32 多元特征函数的一些性质*:*

1. $f(t_1,\ldots,t_n)$在 $\mathbb{R}^n$ 中一致连续, 且

$$|f(t_1,\ldots,t_n)|\le f(0,\ldots,0)=1,$$
$$f(-t_1,\ldots,-t_n)=\overline{f(t_1,\ldots,t_n)};$$

2. 若$f(t_1,\ldots,t_n)$是$(\xi_1,\ldots,\xi_n)$的特征函数, 则$\eta=a_1\xi_1+\cdots+a_n\xi_n$的特征函数为

$$f_\eta(t)=f(a_1t,\ldots,a_nt);$$

3. 若矩$E(\xi_1^{k_1}\cdots\xi_n^{k_n})$存在, 则

$$E(\xi_1^{k_1}\cdots\xi_n^{k_n})=i^{\left(-\sum_{j=1}^n k_j\right)}\left[\frac{\partial^{k_1+\cdots+k_n}f(t_1,\ldots,t_n)}{\partial t_1^{k_1}\cdots\partial t_n^{k_n}}\right]_{t_1=\cdots=t_n=0};$$

4. 若$(\xi_1,\ldots,\xi_n)$的特征函数为$f(t_1,\ldots,t_n)$, 则$k(<n)$维随机变量$(\xi_1,\ldots,\xi_k)$的特征函数为

$$f_{1,\ldots,k}(t_1,\ldots,t_k)=f(t_1,\ldots,t_k,0,\ldots,0).$$

Theorem 4.33 （多元特征函数的逆转公式） 若$f(t_1,\dots,t_n)$和$F(x_1,\dots,x_n)$分别是随机向量$(\xi_1,\dots,\xi_n)$的特征函数和分布函数, 则对$\forall a_k \le b_k \in \mathbb{R}, k = 1,\dots,n$, 有

$$P\{a_k \le \xi_k < b_k, k = 1,\dots,n\}$$
$$= \lim_{\substack{T_j\to\infty \\ j=1,\dots,n}} \frac{1}{(2\pi)^n}\int_{-T_1}^{T_1}\cdots\int_{-T_n}^{T_n}\left[f(t_1,\dots,t_n)\cdot\prod_{k=1}^{n}\frac{e^{-it_ka_k}-e^{-it_kb_k}}{it_k}\right]\mathrm{d}t_1\cdots\mathrm{d}t_n.$$

Theorem 4.34 （多元特征函数的唯一性定理） 分布函数$F(x_1,\dots,x_n)$由其特征函数唯一确定.

Theorem 4.35 （独立性的充要条件）

1. 若随机向量$(\xi_1,\dots,\xi_n)$的特征函数为$f(t_1,\dots,t_n)$, 且ξ_i的特征函数为$f_{\xi_i}(t), i = 1,\dots,n$, 则随机变量$\xi_1,\dots,\xi_n$相互独立的充要条件为
$$f(t_1,\dots,t_n) = f_{\xi_1}(t_1)\cdots f_{\xi_n}(t_n);$$
2. 若随机向量$\xi = (\xi_1,\dots,\xi_n)$, $\eta = (\eta_1,\dots,\eta_m)$和$Z = (\xi_1,\dots,\xi_n,\eta_1,\dots,\eta_m)$的特征函数分别为$f_\xi(t_1,\dots,t_n)$, $f_\eta(u_1,\dots,u_m)$和$f_Z(t_1,\dots,t_n,u_1,\dots,u_m)$, 则$\xi = (\xi_1,\dots,\xi_n)$与$\eta = (\eta_1,\dots,\eta_m)$独立的充要条件为:
对$\forall t_i,\dots,t_n,u_1,\dots,u_m \in \mathbb{R}$, 有
$$f_Z(t_1,\dots,t_n,u_1,\dots,u_m) = f_\xi(t_1,\dots,t_n)\cdot f_\eta(u_1,\dots,u_m).$$

Theorem 4.36 （连续性定理） 若特征函数列$\{f_k(t_1,\dots,t_n)\}_{k=1}^{\infty}$收敛于一个连续函数$f(t_1,\dots,t_n)$, 则$f(t_1,\dots,t_n)$是某个分布函数所对应的特征函数.

5 Limit Theorems

5.1 Convergences

5.1.1 Convergences of Distribution Functions

考虑单位质量全部集中在$x = -\frac{1}{n}$处的退化分布

$$F_n(x) = \begin{cases} 0, & x \le -\frac{1}{n}; \\ 1, & x > -\frac{1}{n}. \end{cases}$$

它的极限函数为

$$F(x) = \begin{cases} 0, & x \le 0; \\ 1, & x > 0. \end{cases}$$

$F_n(0) = 1$对$\forall n \in \mathbb{Z}_+$成立, 但$F(0) = 0$, 因此$\lim_{n\to\infty} F_n(0) \ne F(0)$. 而不收敛的点0是极限函数$F(x)$的不连续点, 因此为了放宽“在所有点都收敛”的条件, 引入弱收敛的概念.

Definition 5.1 （分布函数的弱收敛） 对于分布函数列$\{F_n(x)\}_{n=1}^{\infty}$, 若存在一个非降函数$F(x)$使

$$\lim_{n\to\infty} F_n(x) = F(x)$$

在$F(x)$的**每一连续点**上都成立, 则称$F_n(x)$**弱收敛**于$F(x)$, 记为

$$F_n(x) \xrightarrow{W} F(x).$$

注: 这样得到的极限函数$F(x)$不一定是有界非负函数, 也不一定是分布函数. 例如

$$F_n(x) = \begin{cases} 0, & x \le n; \\ 1, & x > n. \end{cases}$$

显然 $\lim_{n\to\infty} F_n(x) = 0$ 对 $\forall x \in \mathbb{R}$ 成立, 但 $F(x) = 0$ 不是分布函数.

当$F(x)$是一个分布函数时, 至多有可数个不连续点, 分布函数的左连续性保证了它在不连续点上的值完全由它在连续点集C_F上的值唯一确定, 因此分布函数列的**弱收敛极限是唯一的**. 若分布函数列弱收敛到一个连续的分布函数, 这种收敛还是一致的(习题20).

Lemma 5.2 设$\{F_n(x)\}_{n=1}^{\infty}$是关于实变量$x$非降的函数序列, D是$\mathbb{R}$上的一个稠密集, C_F是$F(x)$的连续点集. 则有

$$\forall x \in D, \lim_{n\to\infty} F_n(x) = F(x) \quad \Longrightarrow \quad \forall x \in C_F, \lim_{n\to\infty} F_n(x) = F(x).$$

Theorem 5.3 （*Helly*第一定理） 一致有界的非降函数的序列$\{F_n(x)\}_{n=1}^{\infty}$必存在一个子列$\{F_{n_k}(x)\}_{k=1}^{\infty}$, 它弱收敛到一个有界非降函数$F(x)$.

Theorem 5.4 （*Helly*第二定理） 设$f(x)$是$[a,b]$上的连续函数, $\{F_n(x)\}_{n=1}^{\infty}$是$[a,b]$上弱收敛到函数$F(x)$的一致有界的非降函数的序列, 且$a,b$是$F(x)$的连续点, 则有

$$\lim_{n\to\infty} \int_a^b f(x)\mathrm{d}F_n(x) = \int_a^b f(x)\mathrm{d}F(x).$$

Theorem 5.5 （推广的*Helly*第二定理） 设$f(x)$是$(-\infty,+\infty)$上的有界连续函数, $\{F_n(x)\}_{n=1}^{\infty}$ 是$(-\infty,+\infty)$上弱收敛到函数$F(x)$的一致有界的非降函数的序列, 且

$$\lim_{n\to\infty} F_n(-\infty) = F(-\infty), \quad \lim_{n\to\infty} F_n(+\infty) = F(+\infty),$$

则有

$$\lim_{n\to\infty} \int_{-\infty}^{+\infty} f(x)\mathrm{d}F_n(x) = \int_{-\infty}^{+\infty} f(x)\mathrm{d}F(x).$$

5.1.2 Continuity Theorem

Theorem 5.6 （连续性定理）

1. （正极限定理）设分布函数列$\{F_n(x)\}_{n=1}^{\infty}$弱收敛到某一分布函数$F(x)$, 则相应的特征函数列$\{f_n(t)\}_{n=1}^{\infty}$收敛到特征函数$f(t)$, 且在$t$的任一有限区间内收敛是一致的.

2. （逆极限定理）设特征函数列$\{f_n(t)\}_{n=1}^{\infty}$收敛到某一函数$f(t)$, 且$f(t)$在$t=0$处连续, 则相应的分布函数列$\{F_n(x)\}_{n=1}^{\infty}$弱收敛到某一分布函数$F(x)$, 且$f(t)$是$F(x)$的特征函数.

   将正极限定理与逆极限定理统称为连续性定理, 从而有

3. （*Lévy-Cramér* 连续性定理）分布函数列$\{F_n(x)\}_{n=1}^{\infty}$弱收敛到某一分布函数$F(x)$, 当且仅当$F_n(x)$对应的特征函数列$\{f_n(t)\}_{n=1}^{\infty}$在任意有限区间内一致收敛到某个函数$f(t)$.

连续性定理表明了分布函数与特征函数之间的一一对应是连续的.

5.1.3 Convergences of Random Variables

Definition 5.7 （随机变量的几种收敛）

1. （依分布收敛） 设随机变量$\{\xi_n(\omega)\}_{n=1}^{\infty}$和$\xi(\omega)$的分布函数分别为$\{F_n(x)\}_{n=1}^{\infty}$和$F(x)$, 如果
$$F_n(x) \xrightarrow{W} F(x),$$
   则称$\{\xi_n(\omega)\}_{n=1}^{\infty}$**依分布收敛**于$\xi(\omega)$, 记为
$$\xi_n(\omega) \xrightarrow{L} \xi(\omega) \quad 或 \quad \xi_n(\omega) \xrightarrow{d} \xi(\omega).$$

2. （依概率收敛） 如果对$\forall \varepsilon > 0$, 有
$$\lim_{n\to\infty} P\{|\xi_n(\omega) - \xi(\omega)| \geq \varepsilon\} = 0,$$
   则称$\{\xi_n(\omega)\}_{n=1}^{\infty}$**依概率收敛**于$\xi(\omega)$, 记为
$$\xi_n(\omega) \xrightarrow{P} \xi(\omega).$$

3. （几乎处处收敛） 若随机变量$\{\xi_n(\omega)\}_{n=1}^{\infty}$和$\xi(\omega)$满足
$$P\left\{\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)\right\} = 1,$$
   则称$\{\xi_n(\omega)\}_{n=1}^{\infty}$**几乎处处收敛**（或以概率*1*收敛）于$\xi(\omega)$, 记为
$$\xi_n(\omega) \xrightarrow{a.s.} \xi(\omega).$$

4. （r阶矩收敛） 设随机变量$\{\xi_n(\omega)\}_{n=1}^{\infty}$和$\xi(\omega)$满足$E|\xi_n|^r < \infty$和$E|\xi|^r < \infty$, 其中常数$r > 0$, 如果

$$\lim_{n\to\infty} E|\xi_n - \xi|^r = 0,$$

则称$\{\xi_n(\omega)\}_{n=1}^{\infty}$ r阶矩收敛于$\xi(\omega)$, 记为

$$\xi_n(\omega) \xrightarrow{r} \xi(\omega).$$

其中最重要的是 $r = 2$ 的情况, 称为均方收敛.

Property 5.8 设$\{X_n\}_{n=1}^{\infty}$和$\{Y_n\}_{n=1}^{\infty}$是两列随机变量, 如果

$$X_n \xrightarrow{P} X \quad \text{且} \quad Y_n \xrightarrow{P} Y,$$

则有

1.
$$X_n \pm Y_n \xrightarrow{P} X \pm Y, \quad X_nY_n \xrightarrow{P} XY;$$

2. 设A, B是常数且$B \neq 0$, 则

$$X_n \xrightarrow{P} A, Y_n \xrightarrow{P} B \quad \Longrightarrow \quad \frac{X_n}{Y_n} \xrightarrow{P} \frac{A}{B};$$

3.
$$g(x)\text{是}\mathbb{R}\text{上的连续函数} \quad \Longrightarrow \quad g(X_n) \xrightarrow{P} g(X);$$

4. 依概率收敛的极限在几乎处处的意义下是唯一的, 即

$$X_n \xrightarrow{P} X' \quad \Longrightarrow \quad P\{X = X'\} = 1.$$

Theorem 5.9 （几种收敛的关系）

1. 几乎处处收敛蕴含依概率收敛, 即

$$\xi_n(\omega) \xrightarrow{a.s.} \xi(\omega) \quad \Longrightarrow \quad \xi_n(\omega) \xrightarrow{P} \xi(\omega).$$

证明见 *[1]*的*332-334*页. 其逆命题一般不成立, 反例见 *[1]*的*334*页.

2. 依概率收敛蕴含依分布收敛, 即

$$\xi_n(\omega) \xrightarrow{P} \xi(\omega) \quad \Longrightarrow \quad \xi_n(\omega) \xrightarrow{L} \xi(\omega).$$

逆命题一般不成立, 即依分布收敛一般无法推出依概率收敛, 但有以下结论

3. 若C为常数, 则

$$\xi_n(\omega) \xrightarrow{P} C \quad \Longleftrightarrow \quad \xi_n(\omega) \xrightarrow{L} C.$$

4. r阶矩收敛蕴含依概率收敛（从而蕴含依分布收敛）, 即

$$\xi_n(\omega) \xrightarrow{r} \xi(\omega) \quad \Longrightarrow \quad \xi_n(\omega) \xrightarrow{P} \xi(\omega).$$

证明利用$Markov$不等式（$Chebyshev$不等式的推广）*:*

$$P\{|\xi_n - \xi| \geq \varepsilon\} \leq \frac{E\,|\xi_n - \xi|^r}{\varepsilon^r}, \forall \varepsilon > 0.$$

注*:* 逆命题一般不成立, 反例见 *[1]*的*314*页, 它甚至说明了几乎处处收敛也不能蕴含r阶矩收敛*;*

5. 几乎处处收敛与r阶矩收敛互相不能蕴含*.*

小结:
几乎处处收敛 $\Longrightarrow$ 依概率收敛 $\Longrightarrow$ 依分布收敛
矩收敛 $\Longrightarrow$ 依概率收敛 $\Longrightarrow$ 依分布收敛

5.2 Laws of Large Numbers (LNN)

Definition 5.10 （大数定律） 设$\{\xi_n\}_{n=1}^{\infty}$是随机变量序列, 令

$$\eta_n = \frac{\xi_1 + \cdots + \xi_n}{n},$$

若存在一个常数序列$\{a_n\}_{n=1}^{\infty}$, 使得对$\forall \varepsilon > 0$, 都有

$$\lim_{n\to\infty} P\{|\eta_n - a_n| < \varepsilon\} = 1,$$

则称序列$\{\xi_n\}_{n=1}^{\infty}$服从**大数定律**.

大数定律只要求依概率收敛, 若将其条件强化为几乎处处收敛, 则得到强大数定律. 因此依概率收敛意义下的大数定律也称为弱大数定律.

Definition 5.11 （强大数定律） 设$\{\xi_n\}_{n=1}^{\infty}$是**独立**随机变量序列, 若

$$P\left\{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(\xi_i - E\xi_i) = 0\right\} = 1,$$

则称序列$\{\xi_n\}_{n=1}^{\infty}$服从**强大数定律**.

Theorem 5.12 （*Bernoulli*大数定律） 设μ_n是n重*Bernoulli*试验中事件A发生的次数, p是事件A发生的概率, 则对$\forall\varepsilon>0$, 有

$$\lim_{n\to\infty} P\left\{\left|\frac{\mu_n}{n}-p\right|<\varepsilon\right\}=1,\quad \text{或等价地,}\quad \lim_{n\to\infty} P\left\{\left|\frac{\mu_n}{n}-p\right|\geq\varepsilon\right\}=0.$$

Bernoulli大数定律表明, 在相同条件下重复同一随机试验n次, 当n充分大时, "事件A发生的频率接近其概率"是一个大概率事件, 可用事件发生的频率作为其概率的估计, 这为估计随机事件的**概率**提供了可行的途径. Bernoulli大数定律建立了大量重复独立试验中事件出现频率的稳定性, 使得概率的概念具有现实意义.

Theorem 5.13 设μ_n是n重*Bernoulli*试验中事件A发生的次数, p是事件A发生的概率, 则有

$$P\left\{\lim_{n\to\infty}\frac{\mu_n}{n}=p\right\}=1.$$

Lemma 5.14 （*Borel-Cantelli*引理, 非常有用）

1. 若随机事件序列$\{A_n\}_{n=1}^{\infty}$满足

$$\sum_{n=1}^{\infty}P(A_n)<\infty,$$

则有

$$P\left\{\overline{\lim_{n\to\infty}}\,A_n\right\}=0,\quad \text{或等价地,}\quad P\left\{\underline{\lim_{n\to\infty}}\,A_n^c\right\}=1;$$

2. 若$\{A_n\}_{n=1}^{\infty}$是**相互独立**的随机事件序列, 则

$$\sum_{n=1}^{\infty}P(A_n)=\infty$$

成立的**充要条件**为

$$P\left\{\overline{\lim_{n\to\infty}}\,A_n\right\}=1,\quad \text{或等价地,}\quad P\left\{\underline{\lim_{n\to\infty}}\,A_n^c\right\}=0.$$

(2)给出了使(1)的逆命题也成立而需要添加的条件: "相互独立".

Theorem 5.15 （几个大数定律）

1. （*Borel*强大数定律）设μ_n是n重*Bernoulli*试验中事件A发生的次数, p是事件A发生的概率, 则有

$$P\left\{\lim_{n\to\infty}\frac{\mu_n}{n}=p\right\}=1.$$

2. （*Chebyshev*大数定律）设$\{\xi_n\}_{n=1}^{\infty}$是**两两不相关**(不一定相互独立)的随机变量序列, 它们都有有限方差, 且方差有共同上界(一致有界?), 即存在常数C使 $\sup_i \mathrm{Var}(\xi_i) \le C$, 则对$\forall \varepsilon > 0$, 都有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - \frac{1}{n}\sum_{k=1}^{n}E\xi_k\right| < \varepsilon\right\} = 1.$$

3. （独立同分布下的大数定律）设$\{\xi_n\}_{n=1}^{\infty}$是**独立同分布**的随机变量序列, 且$E\xi_i = \mu$, $\mathrm{Var}(\xi_i) = \sigma^2$, $i = 1, 2, \ldots$, 则对 $\forall \varepsilon > 0$, 都有

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

4. （*Poisson*大数定律）在一个独立试验序列中, 事件A在第k次试验中发生的概率为p_k, 前n次试验中事件A发生的次数为μ_n, 则对 $\forall \varepsilon > 0$, 都有

$$\lim_{n\to\infty} P\left\{\left|\frac{\mu_n}{n} - \frac{p_1 + \cdots + p_n}{n}\right| < \varepsilon\right\} = 1.$$

当 $p_k \equiv p$ 时*Poisson*大数定律即为*Bernoulli*大数定律.

*Markov*注意到在*Chebyshev*的论证中, 只需满足

$$\frac{1}{n^2}\mathrm{Var}\left(\sum_{k=1}^{n}\xi_k\right) \to 0,$$

大数定律就能成立, 该条件称为*Markov*条件, 便得到如下的*Markov*大数定律

5. （*Markov*大数定律）若随机变量序列$\{\xi_n\}_{n=1}^{\infty}$满足

$$\lim_{n\to\infty}\frac{1}{n^2}\mathrm{Var}\left(\sum_{k=1}^{n}\xi_k\right) = 0,$$

则对 $\forall \varepsilon > 0$, 都有

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k - \frac{1}{n}\sum_{k=1}^{n}E\xi_k\right| < \varepsilon\right\} = 1.$$

注: Markov大数定律是Chebyshev大数定律的推广, 因为后者要求两两不相关, 而前者没有关于独立性的假定.

前面介绍了通过Chebyshev不等式建立的几种大数定律, 它们都假定方差存在, 但在独立同分布场合, 不需要方差存在的条件, 有如下的辛钦(Khintchin)大数定律.

Theorem 5.16 （*Khintchin*大数定律）设$\{\xi_n\}_{n=1}^{\infty}$是**相互独立**的随机变量序列, 它们服从相同的分布, 且有有限的数学期望$E\xi_n = \mu$, 则对 $\forall \varepsilon > 0$, 有

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

即

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \xrightarrow{P} \mu.$$

Khintchin大数定律表明, 同一量X在相同条件下观测n次, 当n充分大时, "观测值的平均值接近期望值" 是一个大概率事件, 这为估计随机变量的**期望值**提供了可行的途径.

Theorem 5.17 （*Kolmogorov*强大数定律）设$\{\xi_n\}_{n=1}^{\infty}$是**相互独立**的随机变量序列, 且$\sum_{i=1}^{n}\dfrac{\mathrm{Var}(\xi_n)}{n^2} < \infty$, 则有

$$P\left\{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[\xi_i - E(\xi_i)] = 0\right\} = 1.$$

Kolmogorov在Khintchin大数定律的条件下, 利用对随机变量 "截尾" 的技巧把结论强化为强收敛, 即独立同分布时只要期望存在, 序列部分和的算术平均值几乎处处收敛.

Theorem 5.18 （*Kolmogorov*）设$\{\xi_n\}_{n=1}^{\infty}$是**独立同分布**的随机变量序列, 则

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \xrightarrow{a.s.} E$$

当且仅当数学期望$E\xi_i$存在且$E\xi_i = E$, $i = 1, \ldots, n$.

（用蒙特卡洛方法计算定积分）计算积分

$$J = \int_a^b g(x)dx.$$

可以通过下面的概率方法实现。任取一列相互独立的，都具有$[a, b]$中均匀分布的随机变量$\{\xi_i\}$，则$\{g(\xi_i)\}$也是一列相互独立同分布的随机变量，而且

$$Eg(\xi_i) = \frac{1}{b-a}\int_a^b g(x)dx = \frac{J}{b-a},$$

$$\text{即} \qquad J = (b-a)\cdot Eg(\xi_i).$$

因此只要能求得$Eg(\xi_i)$，便能得到J的数值。

为求$Eg(\xi_i)$，自然想到大数定律，因为

$$\frac{g(\xi_1)+g(\xi_2)+\cdots+g(\xi_n)}{n} \xrightarrow{P} Eg(\xi_i).$$

这样一来，只要能生成随机变量序列$\{g(\xi_i)\}$就能对前面的积分进行数值计算。

而生成$\{g(\xi_i)\}$的关键是生成相互独立同分布的$\{\xi_i\}$，这里的$\{\xi_i\}$服从$[a, b]$上的均匀分布。

现在已经可以把上述想法变成现实。这就是在电子计算机上产生服从均匀分布$[a, b]$的随机数$\{\xi_i\}$。
强大数律保证了这种算法失效的概率为0.

Figure 3: [2]大数定律的应用: 用Monte Carlo方法计算定积分

5.3 Central Limit Theorems (CLT)

Bernoulli大数定律只断言$\frac{\mu_n}{n}$接近于p, 而De Moivre-Laplace极限定理则给出了$\frac{\mu_n}{n}$的渐进分布的精确描述.

Theorem 5.19 （*De Moivre-Laplace*极限定理）
设μ_n是n重*Bernoulli*试验中事件A出现的次数, $0 < p < 1$, 则对任意有限区间$[a, b]$, 有

1. （局部极限定理）若$a \leq x_k := \dfrac{k-np}{\sqrt{npq}} \leq b$, 则当$n \to \infty$时, 一致地有
$$P(\mu_n = k) \div \left(\frac{1}{\sqrt{npq}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_k^2} \right) \to 1;$$

2. （积分极限定理）当$n \to \infty$时, 一致地有
$$P\left\{ a \leq \frac{\mu_n - np}{\sqrt{npq}} < b \right\} \to \int_a^b \varphi(x)\mathrm{d}x,$$
其中 $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} (x \in \mathbb{R})$ 是标准正态分布的密度函数, 易知取 $a = -\infty$ 或 $b = +\infty$ 时仍成立.

局部极限定理提供了$P(\mu_n = k)$的渐进表达式, 积分极限定理给出了标准化随机变量 $\dfrac{\mu_n - np}{\sqrt{npq}}$ 的渐进分布, 它是一般中心极限定理的特例. 因此当n充分大时, 二项分布的概率问题可近似用正态分布解决.

Theorem 5.20 （*Lindeberg-Lévy*中心极限定理）

设$\{X_n\}_{n=1}^{\infty}$是**独立同分布**的随机变量序列, 该分布的期望和方差分别为 $\mu \in \mathbb{R}$ 和 $\sigma^2 \in \mathbb{R}_+$, 则对 $\forall x \in \mathbb{R}$, 有
$$\lim_{n\to\infty} P\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} < x \right\} = \Phi(x) = \int_{-\infty}^{x} \varphi(t)\mathrm{d}t,$$
其中 $\Phi(x)$ 和 $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} (x \in \mathbb{R})$ 分别是标准正态分布的分布函数和密度函数. 定理结论蕴含标准化随机变量 $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$ 依分布收敛（好像还是逐点收敛*?*）于N$(0,1)$.

证明见 [1]324页, 证明简短, 可能会考.

Lindeberg-Lévy中心极限定理表明, 不管$\{X_n\}_{n=1}^{\infty}$服从什么类型的分布, **当n充分大时, $\sum_{i=1}^{n} X_i$的标准化近似服从标准正态分布**. 该定理可以推出De Moivre-Laplace积分极限定理. 应用: 只要n足够大, 便可以把独立同分布的随机变量之和当作正态变量来近似处理.

例: 二项分布计算中的应用

由积分极限定理, 当p不太接近0或1（因为若p接近0或1, 则pq接近0, 误差太大）, 而n又不太小时, 对二项分布的近似计算有如下公式:
$$P\{k_1 \leq \mu_n \leq k_2\} = P\left\{ \frac{k_1 - np}{\sqrt{npq}} \leq \frac{\mu_n - np}{\sqrt{npq}} \leq \frac{k_2 - np}{\sqrt{npq}} \right\}$$
$$\approx \Phi\left(\frac{k_2 - np}{\sqrt{npq}} \right) - \Phi\left(\frac{k_1 - np}{\sqrt{npq}} \right).$$
实际计算中通常用以下的修正公式:
$$P\{k_1 \leq \mu_n \leq k_2\} \approx \Phi\left(\frac{k_2 - np + 0.5}{\sqrt{npq}} \right) - \Phi\left(\frac{k_1 - np - 0.5}{\sqrt{npq}} \right).$$

6 Mathematical Statistics

考虑关于参数θ的假设检验问题:

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

其中Θ_0和Θ_1是参数空间Θ的一个划分, H_0和H_1分别称为零假设(**null hypothesis**)和备择假设(**alternative hypothesis**).

假设检验中的两类错误:

第一类错误 (**Type I error**): 零假设被拒绝, 但其实零假设是正确的;

第二类错误 (**Type II error**): 零假设不被拒绝, 但其实零假设是错误的.

The type I error rate or significance level is the probability of rejecting the null hypothesis given that it is true. Often, the significance level is set to 0.05 , implying that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis.

The rate of the type II error is denoted by β and related to the power of a test (which equals $1 - \beta$).

Table of error types		**Null hypothesis (H_0) is**	
		True	**False**
Decision About Null Hypothesis (H_0)	**Fail to reject**	Correct inference (True Positive)	Type II error (False Negative)
	Reject	Type I error (False Positive)	Correct inference (True Negative)

Figure 4: Table of error types

拒绝零假设的概率依赖于参数θ的真实值, 记$\pi(\theta)$为拒绝零假设的概率, 则**显著性水平**α定义为:

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta).$$

Theorem 6.1 *(Cochran) Suppose $U_1, \ldots, U_n$ are i.i.d. standard normally distributed random variables, and an identity of the form*

$$\sum_{i=1}^{N} U_i^2 = \sum_{j=1}^{k} Q_j$$

can be written, where each Q_j is a quadratic form in $U_1, \ldots, U_n$. Further suppose that

$$r_1 + \cdots + r_k = N$$

where r_j is the rank of Q_j. Cochran's theorem states that the Q_j are independent, and each Q_j has a chi-squared distribution with r_j degrees of freedom. Here the rank of Q_j should be interpreted as meaning the rank of the matrix $B^{(j)}$, with elements $B_{k,l}^{(j)}$, in the representation of Q_j as a quadratic form:

$$Q_j = \sum_{j=1}^{N} \sum_{l=1}^{N} U_k B_{k,l}^{(k)} U_l.$$

Less formally, it is the number of linear combinations included in the sum of squares defining Q_j, provided that these linear combinations are linearly independent.

Example 6.2 *(Sample mean and sample variance)*

If $X_1, \ldots, X_n$ are independent normally distributed random variables with mean μ and standard deviation σ then

$$U_i = \frac{X_i - \mu}{\sigma}$$

is standard normal for each i. It is possible to write

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 + n \left(\frac{\overline{X} - \mu}{\sigma} \right)^2$$

(here $\overline{X}$ is the sample mean). To see this identity, multiply throughout by σ^2 and note that

$$\sum (X_i - \mu)^2 = \sum (X_i - \overline{X} + \overline{X} - \mu)^2$$

and expand to give

$$\sum (X_i - \mu)^2 = \sum (X_i - \overline{X})^2 + \sum (\overline{X} - \mu)^2 + 2 \sum (X_i - \overline{X})(\overline{X} - \mu).$$

The third term is zero because it is equal to a constant times

$$\sum (\overline{X} - X_i) = 0,$$

and the second term has just n identical terms added together. Thus

$$\sum(X_i-\mu)^2=\sum(X_i-\overline{X})^2+n(\overline{X}-\mu)^2,$$

and hence

$$\sum\left(\frac{X_i-\mu}{\sigma}\right)^2=\sum\left(\frac{X_i-\overline{X}}{\sigma}\right)^2+n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2=Q_1+Q_2.$$

Now the rank of Q_2 is just 1 (it is the square of just one linear combination of the standard normal variables). The rank of Q_1 can be shown to be $n-1$, and thus the conditions for Cochran's theorem are met.

Cochran's theorem then states that Q_1 and Q_2 are independent, with chi-squared distributions with $n-1$ and 1 degree of freedom respectively. This shows that the sample mean and sample variance are independent.

假设随机变量X的密度为φ，f为一函数，随机变量Y的密度函数为ψ且满足：若$f(x)\varphi(x)>0$，则$\psi(x)>0$．这样有$E(f(X))=\int f(x)\varphi(x)dx=\int f(x)\dfrac{\varphi(x)}{\psi(x)}\psi(x)dx=E\left(f(Y)\dfrac{\varphi(Y)}{\psi(Y)}\right)$．

$E(f(X))$的重要抽样估计定义为$Z_N^{IS}=\frac{1}{N}\sum_{j=1}^N f(Y_j)\dfrac{\varphi(Y_j)}{\psi(Y_j)}$，其中$Y_1,Y_2,\ldots,Y_N$独立同分布且服从$\psi$.

The **standard error** (SE, 标准误差) of a statistic (usually an estimate of a parameter) is the **standard deviation of its sampling distribution** or an estimate of the standard deviation.

If the parameter or the statistic is the mean, the sampling distribution of a population mean is generated by repeated sampling and recording of the means obtained. This forms a distribution of different means, and this distribution has its own mean and variance. The variance of the sampling distribution obtained is equal to the variance of the population divided by the sample size. As the sample size increases, sample means cluster more closely around the population mean.

Therefore, the relationship between the standard error and the standard deviation is such that, for a given sample size, the standard error equals the standard deviation divided by the square root of the sample size. In other words, the standard error of the mean is a measure of the dispersion of sample means around the population mean.

A **sampling distribution** (抽样分布) is the probability distribution of a given statistic based on a random sample. The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n. The sampling distribution depends on the underlying distribution of the population, the statistic being considered, the sampling procedure employed, and the sample size used.

Population	Statistic	Sampling distribution
Normal: $\mathcal{N}(\mu, \sigma^2)$	Sample mean $\bar{X}$ from samples of size *n*	$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ or (In case sigma is not known): $\bar{X} \sim \mathcal{T}\left(\mu, \frac{S^2}{n}\right)$ Where S is the standard deviation of the sample and $\mathcal{T}$ is the Student's t-distribution.
Bernoulli: $\operatorname{Bernoulli}(p)$	Sample proportion of "successful trials" $\bar{X}$	$n\bar{X} \sim \operatorname{Binomial}(n, p)$
Two independent normal populations: $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$	Difference between sample means, $\bar{X}_1 - \bar{X}_2$	$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

Figure 5: 抽样分布的例子

6.0.1 Mean Squared Error (MSE)

The mean squared error (MSE) of an estimator measures the average of the squares of the errors, that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate. The MSE is a measure of the quality of an estimator. It is always non-negative, and values closer to zero are better. An MSE of zero, meaning that the estimator $\hat{\theta}$ predicts observations of the parameter θ with perfect accuracy, is the ideal, but is typically not possible.

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance of the estimator. Taking the square root of MSE yields the root-mean-square error (RMSE), which has the same units as the quantity being estimated; for an unbiased estimator, the RMSE is the square root of the variance, known as the standard deviation.

The MSE assesses the quality of an estimator or a predictor. Definition of an MSE differs according to whether one is describing an estimator or a predictor.

Predictor

If $\hat{Y}$ is a vector of n predictions, and Y is the vector of observed values of the variable being predicted, then the within-sample MSE of the predictor is computed as $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. I.e., the MSE is the mean of the squares of the errors.

Estimator

The MSE of an estimator $\hat{\theta}$ with respect to an unknown parameter θ is defined as $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}_{\hat{\theta}}\left[(\hat{\theta}-\theta)^2\right]$.

The MSE can be written as the sum of the variance of the estimator and the squared bias of the estimator, providing a useful way to calculate the MSE and implying that in the case of unbiased estimators, the MSE and variance are equivalent.

$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}_{\hat{\theta}}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta},\theta)^2$.

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}_{\hat{\theta}}\left[(\hat{\theta}-\theta)^2\right] \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]+\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right)^2 + 2\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right)\left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right) + \left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + \mathrm{E}_{\hat{\theta}}\left[2\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right)\left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)\right] + \mathrm{E}_{\hat{\theta}}\left[\left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + 2\left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)\mathrm{E}_{\hat{\theta}}\left[\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right] + \left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2 && \mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta = \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + 2\left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)\left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right) + \left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2 && \mathrm{E}_{\hat{\theta}}[\hat{\theta}] = \text{cons} \\
&= \mathrm{E}_{\hat{\theta}}\left[\left(\hat{\theta}-\mathrm{E}_{\hat{\theta}}[\hat{\theta}]\right)^2\right] + \left(\mathrm{E}_{\hat{\theta}}[\hat{\theta}]-\theta\right)^2 \\
&= \mathrm{Var}_{\hat{\theta}}(\hat{\theta}) + \mathrm{Bias}_{\hat{\theta}}(\hat{\theta},\theta)^2
\end{aligned}$$

6.0.2 Jensen's Inequality

Jensen's inequality relates the value of a convex function of an integral to the integral of the convex function. In its simplest form the inequality states that the convex transformation of a mean is less than or equal to the mean applied after convex transformation; it is a simple corollary that the opposite is true of concave transformations.

In the context of probability theory, it is generally stated in the following form: if X is a random variable and ϕ is a convex function, then $\varphi\left(\mathrm{E}[X]\right) \le \mathrm{E}\left[\varphi(X)\right]$.

Jensen's inequality generalizes the statement that the secant line of a convex function lies above the graph of the function, which is Jensen's inequality for two points: the secant line consists of weighted means of the convex function (for $t \in$ [0,1]), $tf(x_1) + (1-t)f(x_2)$, while the graph of the function is the convex function of the weighted means, $f\left(tx_1 + (1-t)x_2\right)$. Thus, Jensen's inequality is $f\left(tx_1 + (1-t)x_2\right) \le tf(x_1) + (1-t)f(x_2)$.

1. Finite form

For a real convex function φ, numbers $x_1, x_2, \ldots, x_n$ in its domain, and positive weights a_i, Jensen's inequality can be stated as: $\varphi\left(\frac{\sum a_i x_i}{\sum a_i}\right) \le \frac{\sum a_i \varphi(x_i)}{\sum a_i}$ (1)

and the inequality is reversed if φ is concave, which is $\varphi\left(\frac{\sum a_i x_i}{\sum a_i}\right) \ge \frac{\sum a_i \varphi(x_i)}{\sum a_i}$. (2)

Equality holds if and only if $x_1 = x_2 = \cdots = x_n$ or φ is linear. As a particular case, if the weights a_i are all equal, then (1) and (2) become $\varphi\left(\frac{\sum x_i}{n}\right) \le \frac{\sum \varphi(x_i)}{n}$ (3)

$\varphi\left(\frac{\sum x_i}{n}\right) \ge \frac{\sum \varphi(x_i)}{n}$ (4)

2. Measure-theoretic and probabilistic form

Let (Ω, A, μ) be a probability space, such that $\mu(\Omega) = 1$. If g is a real-valued function that is μ -integrable, and if φ is a convex function on the real line, then: $\varphi\left(\int_\Omega g\, d\mu\right) \le \int_\Omega \varphi \circ g\, d\mu$.

Let $(\Omega, \mathfrak{F}, \mathrm{P})$ be a probability space, X an integrable real-valued random variable and ϕ a convex function. Then: $\varphi\left(\mathrm{E}[X]\right) \le \mathrm{E}\left[\varphi(X)\right]$.

3. General inequality in a probabilistic setting

More generally, let T be a real topological vector space, and X a T-valued integrable random variable. In this general setting, integrable means that there exists an element $\mathrm{E}[X]$ in T, such that for any element z in the dual space of T: $\mathrm{E}\left|\langle z, X\rangle\right| < \infty$, and $\langle z, \mathrm{E}[X]\rangle = \mathrm{E}[\langle z, X\rangle]$. Then, for any measurable convex function ϕ and any sub-σ-algebra $\mathfrak{G}$ of $\mathfrak{F}$: $\varphi\left(\mathrm{E}\left[X \mid \mathfrak{G}\right]\right) \le \mathrm{E}\left[\varphi(X) \mid \mathfrak{G}\right]$. Here $\mathrm{E}[\cdot \mid \mathfrak{G}]$ stands for the expectation conditioned to the σ-algebra $\mathfrak{G}$. This general statement reduces to the previous ones when the topological vector space T is the real axis, and $\mathfrak{G}$ is the trivial σ-algebra $\{\varnothing, \Omega\}$.

7 Elementary Multivariate Statistics

将r个实值随机变量$X_1, \ldots, X_r$写成r维列向量的形式

$$\mathbf{X} = (X_1, \ldots, X_r)^{\mathrm{T}},$$

$\mathbf{X}$称为随机向量, 其分布函数定义为

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, \ldots, x_r) = P\{X_1 \le x_1, \ldots, X_r \le x_r\} = P\{\mathbf{X} \le \mathbf{x}\}.$$

数学期望$\mu_{\mathbf{X}} = E(\mathbf{X}) = (EX_1, \ldots, EX_r)^{\mathrm{T}} = (\mu_1, \ldots, \mu_r)^{\mathrm{T}}$,
$\mathbf{X}$的协方差矩阵$\mathbf{\Sigma_{XX}} = \mathrm{Cov}(\mathbf{X}, \mathbf{X}) = E\{(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^{\mathrm{T}}\} = (\sigma_{ij})_{r\times r}$.
$\mathbf{\Sigma_{XX}} = E(\mathbf{X}\mathbf{X}^{\mathrm{T}}) - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^{\mathrm{T}}$.
$\mathbf{X}$与$\mathbf{Y}$的协方差矩阵$\mathbf{\Sigma_{XY}} = E\{(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})^{\mathrm{T}}\}$.
$\mathbf{Y} = A\mathbf{X} + b \Longrightarrow \mu_{\mathbf{Y}} = A\mu_{\mathbf{X}} + b$, $\mathbf{\Sigma_{YY}} = A\mathbf{\Sigma_{XX}}A^{\mathrm{T}}$.

7.1 Wishart Distribution

Chi-square分布的推广
The Wishart distribution is a generalization to multiple dimensions of the chi-squared distribution. The Wishart distribution arises as the distribution of the sample covariance matrix for a sample from a multivariate normal distribution. It occurs frequently in likelihood-ratio tests in multivariate statistical analysis. It also arises in the spectral theory of random matrices.

设X是$n \times p$矩阵,

7.2 Hotelling's T-squared distribution

A Summary of Probability Distributions

注: $\varphi(\cdot)$表示特征函数.

A shape parameter is any parameter of a probability distribution that is neither a location parameter nor a scale parameter (nor a function of either or both of these only, such as a rate parameter). Such a parameter must affect the shape of a distribution rather than simply shifting it (as a location parameter does) or stretching/shrinking it (as a scale parameter does).

A.1 Univariate Discrete Distributions

- **Bernoulli distribution** $\mathrm{Ber}(p)$, $0 \le p \le 1$.
  for the outcome of a single Bernoulli trial
  $P(X=1)=p$, $P(X=0)=1-p$.
  $E[X]=p$, $\mathrm{Var}(X)=p(1-p)$.
  $\varphi(t)=1-p+pe^{it}$.

- **Binomial distribution** $\mathrm{B}(n,p)$, $0 \le p \le 1$.
  n重Bernoulli试验中成功次数为k的概率
  for the number of "positive occurrences" given a fixed total number of independent occurrences
  $P(X=k)=b(k;n,p)=\binom{n}{k}p^k(1-p)^{n-k}$, $k=0,1,\ldots,n$.
  $E[X]=np$, $\mathrm{Var}(X)=np(1-p)$.
  $\varphi(t)=(1-p+pe^{it})^n$.

- **Geometric distribution** $\mathrm{G}(p)$, $0 < p \le 1$.
  Bernoulli试验中首次成功出现在第k次的概率
  for binomial-type observations but where the quantity of interest is the number of failures before the first success
  describes the number of attempts needed to get the first success in a series of independent Bernoulli trials, or alternatively only the number of losses before the first success
  $P(X=k)=p(1-p)^{k-1}$, $k=1,2,\ldots$.
  $E[X]=\dfrac{1}{p}$, $\mathrm{Var}(X)=\dfrac{1-p}{p^2}$.
  $\varphi(t)=\dfrac{p}{e^{-it}-(1-p)}$.

- **Hypergeometric distribution** $\mathrm{HG}(n,N,M)$, $0 \le n \le M \le N$.
  N件产品中有M件次品, 抽取n件产品(无放回)中的次品个数为k的概率
  for the number of "positive occurrences" given a fixed number of total occurrences, using sampling without replacement
  $P(X=k)=\dfrac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$, $0 \le k \le n$.

$E[X] = n\frac{M}{N}$, $\mathrm{Var}(X) = n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}$.

- **Poisson distribution** Poi(λ), $\lambda > 0$.
  for the number of occurrences of a Poisson-type event in a given period of time
  $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$, $k = 0, 1, \ldots$.
  $E[X] = \lambda$, $\mathrm{Var}(X) = \lambda$.
  $\varphi(t) = e^{\lambda(e^{it}-1)}$.

- **Negative Binomial distribution** NB(p), $0 < p < 1$.
  有很多种版本, 先以教材的为准

- **Pascal distribution** 以教材的为准

A.2 Univariate Continuous Distributions

- **Uniform distribution** U$[a, b]$, $-\infty < a < b < +\infty$.
  $p(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}$.
  $E[X] = \frac{a+b}{2}$, $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.
  $\varphi(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}$.

- **Exponential distribution** Exp(λ)=$\Gamma(1, \frac{1}{\lambda})$, 率参数$\lambda \in \mathbb{R}_+$.
  for the time before the next Poisson-type event occurs
  $p(x) = \lambda e^{-\lambda x},\ x \geq 0,\ F(x) = 1 - e^{-\lambda x},\ x \geq 0$.
  $E[X] = \frac{1}{\lambda}$, $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.
  $\varphi(t) = (1 - it\lambda^{-1})^{-1}$.

- **Normal (Gaussian) distribution** N(μ, σ^2), 位置参数$\mu \in \mathbb{R}$, 尺度参数$\sigma \in \mathbb{R}_+$.
  $p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$.
  $E[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.
  $\varphi(t) = \exp\left(it\mu - \frac{1}{2}\sigma^2 t^2\right)$.

- **Gamma distribution** $\Gamma(\alpha, \beta)$, 形状参数$\alpha \in \mathbb{R}_+$, 尺度参数$\beta \in \mathbb{R}_+$.
  for the time before the next k Poisson-type events occur
  $p(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}},\ x \geq 0$.
  $E[X] = \alpha\beta$, $\mathrm{Var}(X) = \alpha\beta^2$.
  $\varphi(t) = (1 - it\beta)^{-\alpha}$.

- **Beta distribution** Beta(α,β), 形状参数$\alpha,\beta\in\mathbb{R}_+$.
$p(x)=\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}=\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ 0\le x\le 1.$
$E[X]=\dfrac{\alpha}{\alpha+\beta}$, $\mathrm{Var}(X)=\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

- **Central Chi-squared distribution** $\chi^2(r)$=$\Gamma(\frac{r}{2},2)$, 自由度$r\in\mathbb{R}_+$.
the distribution of a sum of the squares of r independent standard normal random variables
$p(x)=\dfrac{1}{\Gamma(\frac{r}{2})2^{\frac{r}{2}}}x^{\frac{r}{2}-1}e^{-\frac{x}{2}},\ x\ge 0.$
$E[X]=r$, $\mathrm{Var}(X)=2r$.

- **Student's *t*-distribution** $t(r)$, 自由度$r\in\mathbb{R}_+$.
the distribution of the ratio of a standard normal variable and the square root of a scaled chi squared variable
$p(x)=\dfrac{\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})\sqrt{\pi r}}\left(1+\dfrac{x^2}{r}\right)^{-\frac{r+1}{2}}.$
$r>1$时$E[X]=0$, $r>2$时$\mathrm{Var}(X)=\dfrac{r}{r-2}$.
$t(r)\to\mathrm{N}(0,1)\ (r\to\infty)$.
$\varphi(t)=(1-2it)^{-\frac{r}{2}}$.

- **Central *F*-distribution** $F(r_1,r_2)$, 分子自由度r_1,分母自由度$r_2\in\mathbb{R}_+$.
the distribution of the ratio of two scaled chi squared variables
$p(x)=\dfrac{1}{\mathrm{B}\left(\frac{r_1}{2},\frac{r_2}{2}\right)}\left(\dfrac{r_1}{r_2}\right)^{\frac{r_1}{2}}x^{\frac{r_1}{2}-1}\left(1+\dfrac{r_1}{r_2}x\right)^{-\frac{r_1+r_2}{2}},\ x\ge 0.$
$r_2>2$时$E[X]=\dfrac{r_2}{r_2-2}$, $r_2>4$时$\mathrm{Var}(X)=2\left(\dfrac{r_2}{r_2-2}\right)^2\dfrac{r_1+r_2-2}{r_1(r_2-4)}$.

- **Cauchy distribution** Cau(α,β), 位置参数$\alpha\in\mathbb{R}$, 尺度参数$\beta\in\mathbb{R}_+$.
$p(x)=\dfrac{1}{\pi\beta}\left[1+\left(\dfrac{x-\alpha}{\beta}\right)^2\right]^{-1}.$
$E[X]$不存在, $\mathrm{Var}(X)$不存在.
$\varphi(t)=\exp\left(it\alpha-\beta|t|\right)$.

- **Logistic distribution** Log(μ), 位置参数$\mu\in\mathbb{R}$, 尺度参数$s\in\mathbb{R}_+$.
$p(x)=\dfrac{\exp\left(-\frac{x-\mu}{s}\right)}{s\left[1+\exp\left(-\frac{x-\mu}{s}\right)\right]^2}.$
$E[X]=\mu$, $\mathrm{Var}(X)=\dfrac{\pi^2s^2}{3}$.
$\varphi(t)=e^{it\mu}\dfrac{\pi st}{\sinh(\pi st)}$.

- **Laplace distribution** Lap(μ, b), 位置参数$\mu \in \mathbb{R}$, 尺度参数$b \in \mathbb{R}_+$.
  $p(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$.
  $E[X] = \mu$, $\mathrm{Var}(X) = 2b^2$.
  $\varphi(t) = \frac{e^{it\mu}}{1+b^2t^2}$.

- **Log-normal distribution**

A.3 Multivariate Distributions

- **Normal (Gaussian) distribution** $\mathrm{N}(\mu, \mathbf{\Sigma})$, $\mathbf{\Sigma}_{n\times n}$正定.
  $p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det\mathbf{\Sigma})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$.
  $\varphi(\mathbf{t}) = \exp\left(i\mu^{\mathrm{T}}\mathbf{t} - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{t}\right)$.

- **Wishart distribution** Chi-square分布的推广
  The Wishart distribution is a generalization to multiple dimensions of the chi-squared distribution. The Wishart distribution arises as the distribution of the sample covariance matrix for a sample from a multivariate normal distribution. It occurs frequently in likelihood-ratio tests in multivariate statistical analysis. It also arises in the spectral theory of random matrices.

- **Hotelling's T^2 distribution** t分布的推广

B Relations of Probability Distributions

1. Exp(λ)=$\Gamma(1, \frac{1}{\lambda})$.
2. $\chi^2(r)$=$\Gamma(\frac{r}{2}, 2)$.
3. 常数$c > 0$, 则$X \sim \chi^2(r) \Longrightarrow cX \sim \Gamma(\frac{r}{2}, 2c)$.
4. $Z \sim \mathrm{N}(0,1)$与$V \sim \chi^2(r)$独立 $\Longrightarrow \frac{Z}{\sqrt{V/r}} \sim t(r)$.
5. $X_1 \sim \chi^2(r_1)$与$X_2 \sim \chi^2(r_2)$独立 $\Longrightarrow \frac{X_1/r_1}{X_2/r_2} \sim F(r_1, r_2)$.
6. $X \sim \mathrm{Beta}\left(\frac{r_1}{2}, \frac{r_2}{2}\right) \Longleftrightarrow \frac{r_2 X}{r_1(1-X)} \sim F(r_1, r_2)$.
7. $X \sim F(r_1, r_2) \Longrightarrow (\lim_{r_2\to\infty} r_1 X) \sim \chi^2(r_1)$.

8. $X \sim F(r_1, r_2) \Longrightarrow X^{-1} \sim F(r_2, r_1)$.
9. $X \sim t(r) \Longrightarrow X^2 \sim F(1, r)$.
10. $\mathrm{Exp}\left(\frac{1}{2}\right)=\chi^2(2)$.
11. $X_1 \sim \chi^2(r_1)$与$X_2 \sim \chi^2(r_2)$独立 $\Longrightarrow \dfrac{X_1}{X_1 + X_2} \sim \mathrm{Beta}\left(\dfrac{r_1}{2}, \dfrac{r_2}{2}\right)$.
12. $X \sim \mathrm{U}(0, 1) \Longrightarrow -2\log(X) \sim \chi^2(2)$.**（是log还是ln？待确认）**
13. $X_i \sim \Gamma(\alpha_i, \beta_i), i = 1, 2$且$X_1 \perp X_2 \Longrightarrow \dfrac{\alpha_2\beta_2 X_1}{\alpha_1\beta_1 X_2} \sim F(2\alpha_1, 2\alpha_2)$.
14. $X_i \sim \Gamma(\alpha_i, \beta), i = 1, 2$且$X_1 \perp X_2 \Longrightarrow \dfrac{X_1}{X_1 + X_2} \sim \mathrm{Beta}(\alpha_1, \alpha_2)$.

References

[1] 李贤平, 概率论基础 (第三版), 北京: 高等教育出版社, 2010.

[2] 张俊玉 & 巫静, 概率论课件, 2017.

[3] Kai Lai Chung & Farid AitSahlia, *Elementary Probability Theory (Fourth Edition)*, Springer-Verlag New York Inc., 2003.

[4] Kai Lai Chung, *A Course in Probability Theory (Third Edition)*, Academic Press, 2001.

[5] Olav Kallenberg, *Foundations of Modern Probability*, Springer-Verlag New York Inc., 1997.

[6] 严加安, 测度论讲义 (第二版), 北京: 科学出版社, 2004.

[7] Rick Durrett, *Probability: Theory and Examples (Fifth Edition)*, Cambridge University Press, 2017.

[8] Patrick Billingsley, *Probability and Measure (Anniversary Edition)*, John Wiley & Sons, Inc., 2011.

[9] Robert Hogg & Joseph McKean & Allen Craig, *Introduction to Mathematical Statistics (Seventh Edition)*, Pearson Education, Inc., 2013.

[10] 高惠璇, 应用多元统计分析, 北京大学出版社, 2014.

[11] Jun Shao, *Mathematical Statistics (Second Edition)*, Springer Science+Business Media, LLC, 2003.

[12] George Casella & Roger Berger, *Statistical Inference (Second Edition)*, Thomson Learning Inc., 2002.

[13] Erich Lehmann & George Casella, *Theory of Point Estimation (Second Edition)*, Springer-Verlag New York Inc., 1998.

[14] Erich Lehmann & Joseph Romano, *Testing Statistical Hypotheses (Third Edition)*, Springer Science+Business Media, LLC, 2005.

[15] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis (Third Edition)*, John Wiley & Sons, Inc., 2003.

[16] 王学钦, 多元统计课件, 2017.