# 摘要

我自学使用的教材为李贤平的《概率论基础》，在此笔记中主要记录书中的核心内容，配以心得体会。

# 目录

# 1 公理化结构

## 1.1 事件域

**定义 1.1.1** (样本空间). 对于(随机)试验,可能出现的结果称为**样本点**$\omega$,样本点全体构成**样本空间**$\Omega$.

在概率论中一般假定样本空间是给定的,这是必要的抽象,是我们能更好地把我住随机变量的本质,类比于线性空间.

**定义 1.1.2. 事件**定义为样本空间$\Omega$的一个子集,称事件发生当且仅当它所包含的某一个样本点出现.

在此定义下,集合的包含关系诱导了事件的包含关系,补集对应于逆事件,或称对立事件,两个集合的交意味着两个事件的交意味着两个事件同时发生,并集意味着至少发生一个.

一般不把样本空间$\Omega$的一切子集作为事件,这会带来困难,譬如在几何概率中把不可测集也作为事件将会带来不可克服的麻烦. 另一方面,又必须把感兴趣的事件都包括进来,所以要求事件全体$\mathscr{F}$组成一个$\sigma$代数.

**定义 1.1.3** (事件域). 若$\mathscr{F}$是由样本空间$\Omega$的一些子集构成的一个$\sigma$**代数**,则称它为**事件域**, $\mathscr{F}$中的元素称为**事件**,$\Omega$称为必然事件,$\varnothing$称为不可能事件.

需要特别注意的是由给定的$\Omega$的一个非空集族$\mathscr{G}$,必定存在由$\mathscr{G}$生成的$\sigma$代数.这种方法可以定义Borel集.

## 1.2 概率

**定义 1.2.1** (概率). 定义在事件域$\mathscr{F}$上的一个集合函数$P$称为**概率**,如果它满足如下三个要求:

(i) $P(A) \geq 0, \forall A \in \mathscr{F}$

(ii) $P(\Omega) = 1$

(iii) $\forall A_i \in \mathscr{F}, i = 1, 2, \cdots$,若$A_i$两两互不相容,则

$$P(\sum_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

实际上实变函数中的一般测度在全空间的测度为1时就是概率.

有了可数可加性,我们便有**下连续性**.实际上若$A_i \in \mathscr{F}, i = 1, 2, \cdots$且$A_i$两两互不相容,则有

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

两边取极限,右边由于是级数和,1显然是其上界,所以级数和存在,于是有:

$$\lim_{n\to\infty} P(\sum_{i=1}^{n} A_i) = \sum_{i=1}^{\infty} P(A_i) \overset{\text{可数可加}}{=} P(\sum_{i=1}^{\infty} A_i)$$

即对于单调不减的集合列,其**概率的极限等于极限集合的概率**.考虑补集,可以知道概率也是上连续的.另外,有限可加且下连续与可数可加等价.

## 1.3 概率空间

**定义 1.3.1** (概率空间). *$\Omega$是样本空间,$\mathscr{F}$是事件域,$P$是概率,则称三元总体$(\Omega.\mathscr{F}, P)$为* ***概率空间***.

最后这里放个最大似然估计法,因为书上第一章提到了,我觉得比较重要就记下来了,以后有更合适的地方再搬过去吧.

**定义 1.3.2.** *把概率 $p(n)$ 看作未知参数$n$的函数,称为* ***似然函数***,*在通过求其最大值而得到$n$的估计,这就是数理统计中的* ***最大似然估计法***

# 2 条件概率与统计独立性

## 2.1 条件概率

**定义 2.1.1** (条件概率). 设$(\Omega, \mathscr{F}, P)$为一个概率空间,$B \in \mathscr{F}, P(B) > 0$,则对任意$A \in \mathscr{F}$,记

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

并称$P(A|B)$为**在事件$B$发生的条件下事件$A$发生的条件概率**.

概率论的重要课题之一就是通过简单事件的概率推算处复杂事件的概率，这里全概率公式起着重要作用.

**定义 2.1.2** (全概率公式). 设事件$A_1, A_2, \cdots, A_n, \cdots$是样本空间$\Omega$的一个分割,亦称完备事件组,即$A_i$两两互不相容,而且$\sum_{i=1}^{\infty} A_i = \Omega$,这样便有 $B = \sum_{i=1}^{\infty} A_i B$,由概率的可加性与条件概率定义可得

$$P(B) = \sum_{i=1}^{\infty} P(A_i)P(B|A_i)$$

此公式称为**全概率公式**

**定义 2.1.3** (Bayes公式). 若$B$总是与两两互不相容的事件$A_1, A_2, \cdots$之一同时发生,即$B = \sum_{i=1}^{\infty} BA_i$,结合条件概率的定义与全概率公式得

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{\infty} P(A_i)P(B|A_i)}$$

此公式称为Bayes**公式**

Bayes公式得应用场景十分广泛,比如医生为了诊断病人是患了$A_1, \cdots, A_n$这几个疾病中的哪一种,可以先对病人进行检查确定指标$B$,此时利用指标$B$,可以计算相关概率.

Bayes公式中的$P(A_i)$称为**先验概率**,反映了各种原因发生的可能性大小,实际应用中一般是以往经验的总结,试验前便已知道.条件概率$P(A_i|B)$称为**后验概率**, 它反映了试验发生后各种原因发生的可能性大小的条件概率.

## 2.2 事件独立性

**定义 2.2.1** (事件独立性). 对$n$个事件$A_1, A_2, \cdots, A_n$,若对于所有可能的组合$1 \leq i < j < k < \cdots \leq n$成立着

$$\begin{aligned}
&P(A_iA_j) = P(A_i)P(A_j)\\
&P(A_iA_jA_k) = P(A_i)P(A_j)P(A_k)\\
&\cdots\\
&P(A_1A_2\cdots A_n) = P(A_1)P(A_2)\cdots P(A_n)
\end{aligned}$$

则称$A_1, A_2, \cdots, A_n$**相互独立**

事件独立性的每个条件都是必要的,与集合不同,集合两两相交为空集便有任意的交为空集,但是这与事件独立性不同,下面是三个事件两两独立,但三个事件一起并不独立的例子.

考虑一个均匀的正四面体,第一面染成1色,第二面染成2色,第三面染成3色,第四面同时染上1,2,3三种颜色,现在记事件$A, B, C$分别为投一次四面体出现1,2,3色朝下的事件,因此$P(A) = P(B) = P(C) = \frac{1}{2}, P(AB) = P(BC) = P(AC) = \frac{1}{4}$,即此时事件$A, B, C$两两独立,但是$P(ABC) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C)$,即三个事件一起并不一致独立.

有了事件的独立性之后可以定义**试验的独立性**与**重复独立试验**,不在这里写了.

## 2.3 伯努利试验

**定义 2.3.1** (伯努利试验). *把事件域$\mathscr{F}$取为$\{\varnothing, A, \bar{A}, \Omega\}$,并称出现$A$为成功,出现$\bar{A}$为失败,这种只有两个可能结果的试验称为**伯努利试验**. 考虑重复进行$n$次独立的伯努利试验,这种试验称作$n$**重伯努利试验***

**定义 2.3.2** (二项分布). *记伯努利试验中$P(A) = p, P(\bar{A}) = q = 1 - p$, 记$n$重伯努利试验中事件$A$出现$k$次的概率为$b(k; n, p)$, 那么有$b(k; n, p) = \binom{n}{k} p^k q^{n-k}, k = 1, 2, \cdots, n$, $b(k; n, p)$称为**二项分布***

二项分布的结果也可以容易地推广到$n$次重复独立试验且每次试验可能有若干有限个结果的情形,此时称为**多项分布**.

## 2.4 二项分布与泊松分布

二项分布定义如上,我们下面考虑二项分布的性质,由于

$$\frac{b(k; n, p)}{b(k-1;, b, p)} = 1 + \frac{(n+1)p - k}{kq}$$

于是$k = [(n+1)p]$为最可能成功次数,这也是使得$b(k; n, p)$最大的项,此时也称为中心项.

**定理 2.4.1** (二项分布的泊松逼近). *在独立试验中,以$p_n$代表事件$A$在试验中出现的概率,它与试验总数$n$有关,如果 $n \to \infty$ 时 $np_n \to \lambda$,则有*

$$b(k; n, p_n) \to \frac{\lambda^k}{k!} e^{-\lambda}$$

*其中$p(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$称为**泊松分布***

# 3 随机变量与分布函数

## 3.1 随机变量及其分布

**定义 3.1.1** (随机变量). 设$\xi(\omega)$是定义于概率空间$(\Omega, \mathscr{F}, P)$上的单值实函数,其中$\omega$是样本点,如果对于直线上任一$Borel$点集$B$,有

$$\{\omega : \xi(\omega) \in B\} \in \mathscr{F}$$

则称$\xi(\omega)$为**随机变量**,而$P\{\omega : \xi(\omega) \in B\}$称为随机变量$\xi(\omega)$的**概率分布**.

随机变量其实就是测度空间上的可测函数.

**定义 3.1.2** (分布函数). 称

$$F(x) = P\{\omega : \xi(\omega) < x\},\ -\infty < x < \infty$$

为随机变量$\xi(\omega)$的**分布函数**.

为了书写方便,通常把“随机变量$\xi(\omega)$服从分布函数$F(x)$”简记作$\xi(\omega) \sim F(x)$. 由分布函数的定义立刻得到

$$P\{a \le \xi(\omega) < b\} = F(b) - F(a)$$

由此定义的分布函数具有**左连续性**,即$F(x-0) = F(x)$,这里左连续而不一定右连续是因为分布函数的定义中不取等号导致的,具体来说更本质的原因是 $[x_n, x)$ 在$x_n \to x$时会趋近空集.分布函数的定义还能推导出以下等式:

$$\begin{aligned} P\{\xi(\omega) = a\} =& F(a+0) - F(a) \\ P\{\xi(\omega) \le a\} =& F(a+0) \end{aligned}$$

这些公式基本都来源于**概率的下连续性**,即**概率的极限等于极限集合的概率**

**定义 3.1.3** (离散型随机变量). 设$\{x_i\}$为离散型随机变量的所有可能值,而$p(x_i)$是$\xi$取$x_i$的概率,即

$$P\{\xi = x_i\} = p(x_i), i = 1, 2, \cdots$$

那么称$\{p(x_i), i = 1, 2, \cdots\}$为随机变量$\xi$的概率分布.

由此我们可以求出分布函数

$$F(x) = P\{\xi(\omega) < x\} = \sum_{x_k < x} p(x_k)$$

此时分布函数为一个跳跃的阶梯函数.另外,常用**分布列**表出离散型随机变量的概率分布.

**定义 3.1.4** (连续型随机变量). 连续型随机变量$\xi$可取某个区间$[c,d]$或者$(-\infty,\infty)$中的一切值,而且其分布函数$F(x)$是绝对连续函数, 即存在可积函数$p(x)$,使得

$$F(x)=\int_{-\infty}^{x}p(t)\mathrm{d}t$$

此时称$p(x)$为$\xi$的**密度函数**.

这里的绝对连续是实变函数中的概念,$p(x)$在零测集上作改动也不影响分布函数, 所以对于密度函数$p(x)$的论断通常都是在“几乎处处”的意义上成立.这些都是实变函数中老生常谈的内容.

## 3.2 正态分布, 指数分布, $\Gamma$ 分布

**定义 3.2.1** (正态分布). 密度函数为

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

其中$\sigma>0$,$\mu$与$\sigma$均为常数,相应的分布函数为

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t,,-\infty<x<\infty$$

此分布称为**正态分布**,简记为$N(\mu,\sigma^2)$.

为了验证如上定义的$p(x)$确实是密度函数,非负是显然的,于是只需验证其在$\mathbb{R}$上的积分为1, 这是经典数分题目,考虑原积分的平方,将其化成一个累次积分的形式,进而化为二元重积分,再利用极坐标换元便可直接积出来.

特别地,当$\mu=0,\sigma=1$,这时分布称为**标准正态分布**,记为$N(0,1)$, 相应的密度函数与分布函数分别记为$\varphi(x)$与$\Phi(x)$.

可以验证,若随机变量$\xi$服从正态分布$N(\mu,\sigma^2)$, 简记作$\xi\sim N(\mu,\sigma^2)$,则随机变量$\zeta=\frac{\xi-\mu}{\sigma}$服从$N(0,1)$.

有了以上的关系式,一般的正态分布可以化为标准正态分布来处理:
若$\xi\sim N(\mu,\sigma^2)$,则

$$F(x)=P\{\xi<x\}=P\{\frac{\xi-\mu}{\sigma}<\frac{x-\mu}{\sigma}\}=\Phi(\frac{x-\mu}{\sigma})$$

$$P\{a\le\xi<b\}=\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})$$

$$P\{|\xi-\mu|<k\sigma\}=P\{-k<\frac{\xi-\mu}{\sigma}<k\}=\Phi(k)-\Phi(-k)=2\Phi(k)-1$$

**定义 3.2.2** (指数分布). 分布密度函数为

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

分布函数为

$$F(x)=\begin{cases}1-e^{-\lambda x} & x\geq 0\\ 0 & x<0\end{cases}$$

$z$这里$\lambda>0$是参数,这分布称为**指数分布**,简记为$\mathrm{Exp}(\lambda)$.

**定义 3.2.3** ($\Gamma$分布). 称密度函数为

$$f(x)=\begin{cases}\frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x} & x>0\\ 0 & x\leq 0\end{cases}$$

的分布为$\Gamma$**分布**,其中$\lambda>0,r>0$为参数.简记作$\Gamma(r,\lambda)$.

## 3.3 随机向量, 联合分布函数

**定义 3.3.1** (随机向量). 若随机变量$\xi_1(\omega),\cdots,\xi_n(\omega)$定义在同一概率空间$(\Omega,\mathscr{F},P)$上,则称

$$\boldsymbol{\xi}(\omega)=(\xi_1(\omega),\cdots,\xi_n(\omega))$$

构成一个$n$**维随机向量**,亦称$n$**维随机变量**.

对于任意$n$个实数$x_1,\cdots,x_n$,

$$\{\omega:\xi_1(\omega)<x_1,\cdots,\xi_n(\omega)<x_n\}=\bigcap_{i=1}^{n}\{\xi_i(\omega)<x_i\}\in\mathscr{F}$$

即对于$\mathbb{R}^n$中的$n$维矩形$C_n=\prod_{i=1}^n(-\infty,x_i)$,有$\{\boldsymbol{\xi}(\omega)\in C_n\}\in\mathscr{F}$ 于是随机向量也可良好定义(联合)分布函数:

**定义 3.3.2** (联合分布函数). 称$n$元函数

$$F(x_1,\cdots,x_n)=P\{\omega:\xi_1(\omega)<x_1,\cdots,\xi_n(\omega)<x_n\}$$

为随机向量$\boldsymbol{\xi}(\omega)=(\xi_1(\omega),\cdots,\xi_n(\omega))$的***(联合)分布函数***.

多元分布函数的一些性质:

(i) 单调性:关于每个变元是单调不减函数.

(ii) $F(x_1,\cdots,-\infty,\cdots,x_n)=0$
$F(+\infty,\cdots,+\infty)=1$

(iii) 关于每个变元左连续.

特别地,在二元场合,还有: (类似的结论可以推广到$n$元)

(iv) 对任意$a_1<b_1,a_2<b_2$,都有

$$F(b_1,b_2)-F(a_1,b_2)-F(b_1,a_2)+F(a_1,a_2)=P\{a_1\le\xi_1<b_1,a_2\le\xi_2<b_2\}\ge 0$$

随机向量也有不同类型,最常见的也是离散型与连续型两类.

离散型场合中,概率分布集中在有限或可列个点上,其中比较重要的有多项分布与多元超几何分布

在连续型场合,存在非负函数$p(x_1,\cdots,x_n)$,使得

$$F(x_1,\cdots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}p(t_1,\cdots,t_n)\mathrm{d}t_1\cdots\mathrm{d}t_n$$

这里的$p(x_1,\cdots,x_n)$称为(多元分布)**密度函数**

**定义 3.3.3** (多元正态分布). 若$\boldsymbol{\Sigma}=(\sigma_{ij})$是$n$阶正定对称矩阵,以$\boldsymbol{\Sigma}^{-1}=(\gamma_{ij})$表示$\boldsymbol{\Sigma}$的逆矩阵, $\det\boldsymbol{\Sigma}$表示$\boldsymbol{\Sigma}$的行列式的值.$\boldsymbol{\mu}=(\mu_1,\cdots,\mu_n)$是任意实值行向量,则由密度函数

$$p(x_1,\cdots,x_n)=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\boldsymbol{\Sigma})^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}}$$

定义的分布称为$n$**元正态分布**,简记为$N(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

## 3.4 边际分布, 条件分布

下面的讨论将对二维场合进行,$n$维时这些结论仍然成立,设随机向量为$(\xi,\eta)$.

**定义 3.4.1** (离散型随机变量的边际分布). 对于离散型分布,设$\xi$取值$x_1,x_2,\cdots,\eta$取值$y_1,y_2,\cdots$,显然有

$$p_1(x_i):=P\{\xi=x_i\}=\sum_j p(x_i,y_j)$$

$$p_2(y_j):=P\{\eta=y_j\}=\sum_i p(x_i,y_j)$$

这里$p_1(x_i)$与$p_2(y_j)$称为$p(x_i,y_j)$的**边际分布**或**边缘分布**

**定义 3.4.2** (连续型随机变量的边际分布). 若 $(\xi,\eta)$ 是二维随机变量, 其分布函数为 $F(x,y)$, 我们能由 $F(x,y)$ 得出 $\xi$ 和 $\eta$ 的分布函数. 事实上,

$$F_1(x)=P\{\xi<x\}=P\{\xi<x,\eta<+\infty\}=F(x,+\infty)$$

同理

$$F_2(x)=p\{\eta<y\}=F(+\infty,y)$$

$F_1(x)$ 及 $F_2(x)$ 称为 $F(x,y)$ 的**边际分布函数**.
若 $F(x,y)$ 是连续型分布函数, 有密度函数 $p(x,y)$ 那么

$$F_1(x)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty}p(u,y)\mathrm{d}y\mathrm{d}u$$

因此 $F_1(x)$ 是连续型分布函数, 其密度函数为

$$p_1(x)=\int_{-\infty}^{\infty}p(x,y)\mathrm{d}y$$

$p_1(x)$ 称为**边际(分布)密度函数**. $F_2(x),p_2(x)$ 同理.

**定义 3.4.3** (离散型随机变量的条件分布). 对于离散型随机变量, 若已知 $\xi=x_i(p_1(x_i)>0)$, 则事件 $\{\eta=y_i\}$ 的条件概率为

$$P\{\eta=y_j|\xi=x_i\}=\frac{P\{\xi=x_i,\eta=y_j\}}{P\{\xi=x_i\}}=\frac{p(x_i,y_j)}{p_1(x_i)}$$

此式定义了随机变量 $\eta$ 关于随机变量 $\xi$ 的**条件分布**.

**定义 3.4.4** (连续型随机变量的条件分布). 对于连续型随机变量, $p(x,y),p_1(x)$ 为其密度函数与边际密度函数, 那么其**条件分布函数**定义为

$$P\{\eta<y|\xi=x\}=\frac{\int_{-\infty}^{y}p(x,v)\mathrm{d}v}{p_1(x)}=\int_{-\infty}^{y}\frac{p(x,v)}{p_1(x)}\mathrm{d}v$$

其分布密度函数为

$$p(y|x)=\frac{p(x,y)}{p_1(x)}$$

**定义 3.4.5** (二元正态分布). 函数

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\times\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(x-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}$$

称为**二元正态(分布)密度函数**, 其中 $\sigma_1>0,\sigma_2>0,|\rho|<1$, 简记为 $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$.

**定理 3.4.1** (二元正态分布的典型分解)**.** 其实也不是定理, 就是一个重要结论. 二元正态密度函数具有如下分解式:

$$p(x,y)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\times\frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}e^{-\frac{\left[y-\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)\right]^2}{2\sigma_2^2(1-\rho^2)}}$$

对于此分解, 第一部分为 $N(\mu_1,\sigma_1^2)$ 的密度函数, 第二部分为 $N\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\sigma_2^2(1-\rho^2)\right)$ 的密度函数

并且二元正态分布的边际分布为正态分布:

$$p_1(x)=\int_{-\infty}^{\infty}p(x,y)\mathrm{d}y=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

二元正态分布的条件分布仍然是正态分布:

$$p(y|x)=\frac{p(x,y)}{p_1(x)}=\frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}e^{-\frac{\left[y-\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)\right]^2}{2\sigma_2^2(1-\rho^2)}}$$

因此, 此典型分解式的涵义就完全清楚了, 第一部分是边际密度 $p_1(x)$, 第二部分是条件密度 $p(y|x)$, 整个式子形如 $p(x,y)=p_1(x)p(y|x)$

## 3.5 随机变量的独立性

**定义 3.5.1** (随机变量的独立性)**.** 设 $\xi_1,\cdots,\xi_n$ 为 $n$ 个随机变量, 若对于任意的 $x_1,\cdots,x_n$, 成立:

$$P\{\xi_1<x_1,\cdots,\xi_n<x_n\}=P\{\xi_1<x_1\}\cdots P\{\xi_n<x_n\}$$

则称 $\xi_1,\cdots,\xi_n$ 是**相互独立的**.

若 $\xi_i$ 的分布函数为 $F_i(x)$, 联合分布函数为 $F(x_1,\cdots,x_n)$, 则上式等价于对一切 $x_1,\cdots,x_n$ 成立:

$$F(x_1,\cdots,x_n)=F_1(x_1)\cdots F_n(x_n)$$

在这种场合, 由每个随机变量的(边际)分布函数可以唯一地确定联合分布函数,此时条件分布化为无条件分布

$$P\{\eta<y|\xi=x\}=P\{\eta<y\}$$

即由 $\xi$ 的取值不能得出任何关于 $\eta$ 的信息.

对连续型随机变量, 上述独立性条件的等价形式是对 $x_1,\cdots,x_n$ 几乎处处成立

$$p(x_1,\cdots,x_n)=p_1(x_1)\cdots p_n(x_n)$$

这里 $p(x_1,\cdots,x_n)$ 是联合分布密度函数, $p_i(x)$ 是各随机变量的密度函数.

## 3.6 随机变量的函数及其分布

**定义 3.6.1** (Borel 函数)**.** 设 $y=g(x)$ 是 $\mathbb{R}\to\mathbb{R}$ 的一个映射, 若对于一切 $\mathbb{R}$ 中的 Borel 点集 $B$ 均有

$$\{x:g(x)\in B\}\in\mathscr{B}$$

其中 $\mathscr{B}$ 为 $\mathbb{R}$ 上的 Borel $-\sigma$ 域, 则称 $g(x)$ 是**一元 Borel （可测）函数**.

若 $\xi$ 是随机变量, $g(x)$ 是一元 Borel 函数, 则 $g(\xi)$ 也是随机变量, 事实上, 对一切 $B\in\mathscr{B}$ 有:

$$\{\omega:g(\xi(w))\in B\}=\{\omega:\xi(\omega)\in g^{-1}(B)\}\in\mathscr{F}$$

若 $\zeta\sim N(0,1)$, 下面来求 $\eta=\zeta^2$ 的密度函数 $q(y)$:
$y\le 0$ 时, $G(y)=P\{\eta<y\}=0$, 显然此时 $q(y)=0$.
$y>0$ 时,

$$\begin{aligned}G(y)=&P\{\eta<y\}=P\{\zeta^2<y\}=P\{-\sqrt{y}<\zeta<\sqrt{y}\}\\=&\int_{-\sqrt{y}}^{\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x=2\int_0^{\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x\end{aligned}$$

因此 $\eta=\zeta^2$ 的密度函数为

$$q(y)=\frac{1}{\sqrt{2\pi}}y^{-\frac{1}{2}}e^{-\frac{y}{2}}\quad(y>0)$$

此分布是下面 $\chi^2$ 分布 $n=1$ 的特例.

**定义 3.6.2** ($\chi^2$ 分布)**.** 具有密度函数

$$p(x)=\frac{1}{2^{n/2}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{x}}\quad(x>0)$$

的分布称为**具有自由度 $n$ 的 $\chi^2$ 分布**.、、

**定理 3.6.1.** 若 $\xi$ 是连续型随机变量, 其密度函数为 $p(x)$, 而 $\eta=g(\xi)$, 对其密度函数, $q(y)$ 有如下结果:
(1) 若 $g(x)$ 严格单调, 其反函数为 $g^{-1}(y)$ 有连续导函数, 则 $\eta=g(\xi)$ 是具有密度函数

$$p(g^{-1}(y))|g^{-1}(y)'|$$

的连续型随机变量.
(2) 若 $g(x)$ 在不相重叠的区间 $I_1,I_2,\cdots$ 上逐段严格单调, 其反函数分别为 $h_1(y),h_2(y),\cdots$ 而且 $h_1'(x),h_2'(x),\cdots$ 均为连续函数, 那么 $\eta=g(\xi)$ 是连续性随机变量, 其分布密度函数为

$$p(h_1(y))|h_1'(y)|+p(h_2(y))|h_2'(y)|+\cdots$$

证明. $P\{\eta < a\} = P\{g(\xi) < a\} = \int_{-\infty}^{g^{-1}(a)} p(x)\mathrm{d}x = \int_{-\infty}^{a} p(g^{-1}(y))|g^{-1}(y)'|\mathrm{d}y$ 分段的情况类似. □

**定理 3.6.2** (均匀分布的特殊地位). 若随机变量 $\xi$ 的分布函数为 $F(x)$, 因为 $F(x)$ 是非降函数, 对任意 $0 \le y \le 1$, 可定义

$$F^{-1}(y) = \inf\{x : F(x) > y\}$$

作为 $F(x)$ 的反函数.

下面考察随机变量 $\theta = F(\xi)$ 的分布, 这里 $F(x)$ 是连续函数. 对 $0 \le x \le 1$:

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

即 $\theta = F(\xi)$ 服从 $[0,1]$ 均匀分布, 这个结论在统计中起重要作用.

反之, 若 $\theta$ 服从 $[0,1]$ 均匀分布, 对任意分布函数 $F(x)$, 令 $\xi = F^{-1}(\theta)$, 则

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

因此 $\xi$ 是服从分布函数 $F(x)$ 的随机变量.

由此, 只要我们产生 $[0,1]$ 中均匀分布的随机变量的样本（观察值）, 那么我们就可以用如上方法得到分布函数为 $F(x)$ 的随机变量的样本.

## 3.7 随机向量的函数及其分布

**定义 3.7.1** ($n$ 元 Borel 函数). 设 $y = g(x_1, \cdots, x_n)$ 是 $\mathbb{R}^n \to \mathbb{R}^n$ 上的一个映射, 若对一切 $\mathbb{R}$ 中的 Borel 点集 $B$ 均有:

$$\{(x_1, \cdots, x_n) : g(x_1, \cdots, x_n) \in B\} \in \mathscr{B}_n$$

其中 $\mathscr{B}_n$ 为 $\mathbb{R}^n$ 上的 Borel $-\sigma$ 域, 则称 $g(x_1, \cdots, x_n)$ 为 $n$ **元** Borel **（可测）函数**.

若 $(\xi_1, \cdots, \xi_n)$ 是随机向量, $g(x_1, \cdots, x_n)$ 是 $n$ 元 Borel 函数, 则同上节可证 $g(\xi_1, \cdots, \xi_n)$ 是随机变量.

若 $\eta = g(\xi_1, \cdots, \xi_n)$, 而 $(\xi_1, \cdots, \xi_n)$ 的密度函数为 $p(x_1, \cdots, x_n)$, 则有:

$$G(y) = P\{\eta < y\} = \int\cdots\int_{g(x_1,\cdots,x_n)<y} p(x_1, \cdots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

下面看一些具体的例子.

**定理 3.7.1** (和的分布-卷积). 若 $\eta = \xi_1 + \xi_2$, 而 $(\xi_1, \xi_2)$ 的密度函数为 $p(x_1, x_2)$, 则

$$G(y) = P\{\eta < y\} = \int_{-\infty}^{\infty}\int_{-\infty}^{y-x_1} p(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1$$

特别地, 当 $\xi_1,\xi_2$ 相互独立时, 有 $p(x_1,x_2)=p_1(x_1)p_2(x_2)$, 这里 $p_1(x)$ 是 $\xi_1$ 的密度函数, $p_2(x)$ 是 $\xi_2$ 的密度函数, 代入得

$$\begin{aligned}G(y)&=\int_{-\infty}^{\infty}\int_{-\infty}^{y-x_1}p_1(x_1)p_2(x_2)\mathrm{d}x_2\mathrm{d}x_1=\int_{-\infty}^{\infty}\int_{-\infty}^{y}p_1(x_1)p_2(z-x_1)\mathrm{d}z\mathrm{d}x_1\\&=\int_{-\infty}^{y}\int_{-\infty}^{\infty}p_1(x_1)p_2(z-x_1)\mathrm{d}x_1\mathrm{d}z\end{aligned}$$

因此 $\eta$ 的密度函数为

$$q(y)=\int_{-\infty}^{\infty}p_1(u)p_2(y-u)\mathrm{d}u=\int_{-\infty}^{\infty}p_1(y-u)p_2(u)\mathrm{d}u$$

上面两式称为 $p_1$ 与 $p_2$ 的**卷积**.

## 3.8 随机向量的变换

这一节比较难, 没掌握好.

若 $(\xi_1,\cdots,\xi_n)$ 的密度函数为 $p(x_1,\cdots,x_n)$, 求 $\eta_1=g_1(\xi_1,\cdots,\xi_n),\cdots,\eta_m=g_m(\xi_1,\cdots,\xi_n)$ 的分布, 此时有:

$$G(y_1,\cdots,y_m)=P\{\eta_1<y_1,\cdots,\eta_m<y_m\}=\int\cdots\int_{g_1<y_1,\cdots,g_m<y_m}p(x_1,\cdots,x_n)\mathrm{d}x_1\mathrm{d}x_n$$

其中 $g_i=g_i(x_1,\cdots,x_n)$, 上述是最一般的情况.

## 3.9 随机变量的函数的独立性

**定理 3.9.1.** 若 $\xi_1,\cdots,\xi_n$ 是相互独立的随机变量, 则 $f_1(\xi_1),\cdots,f_n(\xi_n)$ 也是相互独立的, 这里 $f_i$ 是任意的一元 Borel 函数.

# 4 数字特征与特征函数

## 4.1 数学期望

**定义 4.1.1** (加权平均值). 给定权 $\omega_i \geq 0, i = 1, \cdots, n$, 满足 $\sum\limits_{i=1}^{n} \omega_i = 1$ 则

$$\bar{x}_\omega = \sum_{i=1}^{n} \omega_i x_i$$

称为 $x_1, \cdots, x_n$ 关于权 $\{\omega_i\}$ 的**加权平均值**.

**定义 4.1.2** (离散型随机变量的数学期望). 设 $\xi$ 为一离散型随机变量, 它取值 $x_1, x_2, \cdots$ 对应的概率为 $p_1, p_2, \cdots$ 如果级数

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

绝对收敛, 则把它称为 $\xi$ 的**数学期望** (mathematical expectation), 简称**期望, 期望值**或**均值** (mean), 记作 $E\xi$. 若级数不绝对收敛, 则说 $\xi$ 的数学期望不存在.

下面考虑连续型随机变量的数学期望, 设随机变量 $\xi$ 有密度函数 $p(x)$, 取划分 $x_0 < x_1 < \cdots < x_n$, 则 $\xi$ 落在 $[x_i, x_{i+1}]$ 的概率近似等于 $p(x_i)(x_{i+1} - x_i)$, 因此 $\xi$ 与以概率 $p(x_i)(x_{i+1} - x_i)$ 取值 $x_i$ 的离散型随机变量近似, 而这离散型随机变量的数学期望为

$$\sum_i x_i p(x_i)(x_{i+1} - x_i)$$

上式是积分 $\int_{-\infty}^{\infty} xp(x)\mathrm{d}x$ 的渐进和式.

**定义 4.1.3** (连续型随机变量的数学期望). 设 $\xi$ 是具有密度函数 $p(x)$ 的连续型随机变量, 当积分 $\int_{-\infty}^{\infty} xp(x)\mathrm{d}x$ 绝对收敛时, 我们称它为 $\xi$ 的**数学期望**（或**均值**）, 记作 $E\xi$, 即

$$E\xi = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$$

下面计算一些重要的连续型分布的数学期望.

**例 4.1.1** (正态分布的期望). 对于正态分布 $N(\mu, \sigma^2)$,

$$\begin{aligned}\int_{-\infty}^{\infty} xp(x)\mathrm{d}x &= \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma z + \mu)e^{-z^2/2}\mathrm{d}z \\ &= \frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\mathrm{d}z = \mu\end{aligned}$$

可见 $N(\mu, \sigma^2)$ 中的参数 $\mu$ 正是它的数学期望.

下面考虑一般场合的数学期望, 需要利用 Stieltjes 积分.

若随机变量 $\xi$ 的分布函数为 $F(x)$, 类似于对连续型随机变量的讨论, 取划分 $x_0 < x_1 < \cdots < x_n$, 则 $\xi$ 落在 $[x_i, x_{i+1})$ 中的概率为 $F(x_{i+1}) - F(x_i)$, 因此 $\xi$ 与以概率 $F(x_{i+1}) - F(x_i)$ 取值 $x_i$ 的离散型随机变量近似, 而后者的数学期望为

$$\sum_i x_i[F(x_{i+1}) - F(x_i)]$$

上式为 Stieltjes 积分的渐进和式. 故引进如下定义:

**定义 4.1.4** (数学期望)**.** *若 $\xi$ 的分布函数为 $F(x)$, 则定义*

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

*为 $\xi$ 的**数学期望**（或**均值**）. 这里我们还是要求上述积分绝对收敛, 否则称数学期望不存在.*

关于 Stieltjes 积分

$$I = \int_{-\infty}^{\infty} g(x)\mathrm{d}F(x)$$

它有如下性质:

(1) 当 $F(x)$ 为跳跃函数, 在 $x_i$ 处具有跃度 $p_i$ 时, 上面积分化为求和级数

$$I = \sum_i g(x_i)p_i$$

(2) 当 $F(x)$ 存在导数 $F'(x) = p(x)$ 时, 上述积分化为普通积分

$$I = \int_{-\infty}^{\infty} g(x)p(x)\mathrm{d}x$$

由此可以知道此定义可以囊括离散型和连续型随机变量的数学期望的定义.

**随机变量函数的数学期望** 下面讨论随机变量的函数 $\eta = g(\xi)$ 的数学期望, 这里 $\xi$ 是分布函数为 $F_\xi(x)$ 的随机变量, $g(x)$ 是一元 Borel 函数, 类似于上节的讨论, 似应定义 $g(\xi)$ 的数学期望为

$$\sum_i g(x_i)[F_\xi(x_{i+1}) - F_\xi(x_i)]$$

的极限, 即

$$Eg(\xi) = \int_{-\infty}^{\infty} g(x)\mathrm{d}F_\xi(x)$$

但是, 另一方面, 因为 $\eta$ 是随机变量, 也有分布函数, 设为 $F_\eta(x)$, 又应有

$$E\eta = \int_{-\infty}^{\infty} y\mathrm{d}F_\eta(y)$$

因此, 这两个积分应该相等, 即:

$$\int_{-\infty}^{\infty} y\mathrm{d}F_\eta(y) = \int_{-\infty}^{\infty} g(x)\mathrm{d}F_\xi(x)$$

事实上这是个定理, 但证明需要用到测度论, 不作讨论. 但这个定理很重要, 因为它跳过了计算 $\eta$ 的分布函数 $F_\eta(y)$ 并求得了 $\eta$ 的数学期望.

**多维场合** 若随机向量 $(\xi_1, \cdots, \xi_n)$ 的分布函数为 $F(x_1, \cdots, x_n)$, 而 $g(x_1, \cdots, x_n)$ 为 $n$ 元 Borel 函数, 则

$$Eg(x_1, \cdots, x_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_n)\mathrm{d}F(x_1, \cdots, x_n)$$

特别地,

$$E\xi_1 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1\mathrm{d}F(x_1, \cdots, x_n) = \int_{-\infty}^{\infty} x_1\mathrm{d}F_1(x_1)$$

其中 $F_1(x_1)$ 是 $\xi_1$ 的分布函数. 一般地引进如下定义:

**定义 4.1.5** (随机向量的数学期望)**.** *随机向量 $(\xi_1, \cdots, \xi_n)$ 的**数学期望**为 $(E\xi_1, \cdots, E\xi_n)$, 其中*

$$E\xi_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i\mathrm{d}F(x_1, \cdots, x_n) = \int_{-\infty}^{\infty} x_i\mathrm{d}F_i(x_i)$$

*这里 $F_i(x_i)$ 是 $\xi_i$ 的分布函数.*

**数学期望的基本性质** 利用随机变量函数的数学期望计算公式, 我们可以得到如下基本性质:

性质(1) 若 $a \le \xi \le b$, 则 $a \le E\xi \le b$. 特别地 $Ec = c$, 这里 $a, b, c$ 为常数.

性质(2) 线性性: 对任意常数 $c_i, i = 1, \cdots, n$ 及 $b$, 有

$$E(\sum_{i=1}^{n} c_i\xi_i + b) = \sum_{i=1}^{n} c_iE\xi_i + b$$

## 4.2 方差

数学期望给出了随机变量的均值, 接下来考虑的便是随机变量对于均值的偏差, 也就是方差.

**定义 4.2.1** (方差)**.** *若 $E(\xi - E\xi)^2$ 存在, 则称它为随机变量 $\xi$ 的**方差**(variance), 并记为 $D\xi$, 而 $\sqrt{D\xi}$ 称为**根方差**, **均方差**, 或更多的称为**标准差**(standard deviation).*

标准差与随机变量具有相同的量纲(齐次), 有时更便于应用, 但方差的数学性质更好, 因此更为常用.

利用数学期望的线性性质, 可以得到方差很方便的计算式:

$$D\xi = E(\xi - E\xi)^2 = E[\xi^2 - 2\xi \cdot E\xi + (E\xi)^2] = E\xi^2 - (E\xi)^2$$

**例 4.2.1** (正态分布的方差)**.**

$$D\xi = \int_{-\infty}^{\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}\mathrm{d}x = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}z^2e^{-z^2/2}\mathrm{d}z$$
$$= \frac{\sigma^2}{\sqrt{2\pi}}\left[(-ze^{-z^2/2})\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty}e^{-z^2/2}\mathrm{d}z\right] = \sigma^2$$

*于是正态分布中的第二个参数 $\sigma$ 就是标准差, 正态分布由它的数学期望及标准差唯一确定.*

### 方差的基本性质

性质(1)　常数的方差为 0.

性质(2)　$D(\xi + c) = D(\xi)$, 这里 $c$ 是常数.

性质(3)　$D(c\xi) = c^2D\xi$, 这里 $c$ 是常数.

对于随机变量 $\xi$, 若它的期望和方差均存在, 而且 $D\xi > 0$, 有时可以考虑**标准化**了的随机变量

$$\xi^* = \frac{\xi - E\xi}{\sqrt{D\xi}}$$

此时 $E\xi^* = 0, D\xi^* = 1$

性质(4)　若 $c \neq E\xi$, 则 $D\xi < E(\xi - c)^2$

此式表明数学期望在方差表达式中的一个极值性质.

### 切比雪夫不等式

**定理 4.2.1** (切比雪夫不等式)**.** *对于任何具有有限方差的随机变量 $\xi$, 都有*

$$P\{|\xi - E\xi| \geq \epsilon\} \leq \frac{D\xi}{\epsilon^2}$$

*其中 $\epsilon$ 是任一正数.*

*证明.* 设 $F(x)$ 为 $\xi$ 的分布函数, 则

$$D\xi = \int_{-\infty}^{\infty}(x-E\xi)^2\mathrm{d}F(x) \geq \int_{|x-E\xi|\geq\epsilon}(x-E\xi)^2\mathrm{d}F(x)$$
$$\geq \int_{|x-E\xi|\geq\epsilon}\epsilon^2\mathrm{d}F(x) = \epsilon^2P\{|\xi - E\xi| \geq \epsilon\}$$

□

切比雪夫不等式断言不管 $\xi$ 的分布是什么, $\xi$ 落在 $(E\xi - \sigma\delta, E\xi + \sigma\delta)$ 中的概率均不小于 $1 - \frac{1}{\delta^2}$

## 4.3 相关系数

对于随机向量 $\xi = (\xi_1, \cdots, \xi_n)$, 定义它的**方差**为 $(D\xi_1, \cdots, D\xi_n)$. 这反映了随机向量各个分量对于各自的数学期望的偏离程度, 但是我们还希望知道各个分量之间的联系, 从而引进相关系数.

由于

$$D(\xi \pm \eta) = E[(\xi \pm \eta) - (E\xi \pm E\eta)]^2 = D\xi + D\eta \pm 2E[(\xi - E\xi)(\eta - E\eta)]$$

可见, 为了计算 $\xi \pm \eta$ 的方差, 需要计算 $E[(\xi - E\xi)(\eta - E\eta)]$, 引入如下定义:

**定义 4.3.1** (协方差)**.** 称

$$\sigma_{ij} = \mathrm{cov}(\xi_i, \xi_j) = E[(\xi_i - E\xi_i)(\xi_j - E\xi_j)] \quad (i, j = 1, \cdots, n)$$

为 $\xi_i$ 与 $\xi_j$ 的**协方差**(covariance).

方差是协方差的特例: $\sigma_{ii} = D\xi_i$

直接验证可以得到如下性质:

$$\mathrm{cov}(\xi_i, \xi_j) = E\xi_i\xi_j - E\xi_i \cdot E\xi_j$$

$$D(\sum_{i=1}^{n} \xi_i) = \sum_{i=1}^{n} D\xi_i + 2 \sum_{1 \le i < j \le n} \mathrm{cov}(\xi_i, \xi_j)$$

如下对称矩阵称为 $\xi$ 的**协方差矩阵**.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

此外，对任何实数 $t_j (j = 1, \cdots, n)$ 有

$$\sum_{j,k} \sigma_{jk} t_j t_k = E\Big[\sum_{j=1}^{n} t_j(\xi - E\xi)\Big]^2 \ge 0$$

因此 $\Sigma$ 是一个**非负定**矩阵, 所以 $\det \Sigma \ge 0$.

更常用的是如下"标准化"之后的协方差:

**定义 4.3.2** (相关系数)**.** 称

$$\rho_{ij} = \frac{\mathrm{cov}(\xi_i, \xi_j)}{\sqrt{D\xi_i}\sqrt{D\xi_j}}$$

为 $\xi_i$ 与 $\xi_j$ 的**相关系数**(correlation coefficient), 这里要求 $D\xi_i, D\xi_j$ 不为零.

补充定义常数与任何随机变量的相关系数为 0.

相关系数为正时, 称两随机变量**正相关**, 为负时则称**负相关**.

由定义, 相关系数也就是标准化的随机变量 $\frac{\xi_i-E\xi_i}{\sqrt{D\xi_i}}$ 与 $\frac{\xi_j-E\xi_j}{\sqrt{D\xi_j}}$ 的协方差.

可以说相关系数时规格化了的协方差, 其优点是排除了随机变量的量纲的影响, 并且在线性变换下保持不变, 即对于 $ac>0$, 则 $a\xi+b$ 与 $c\eta+d$ 的相关系数仍为 $\rho_{\xi\eta}$.

**定义 4.3.3** (Cauchy-Schwarz 不等式)**.** *对任意随机变量 $\xi$ 与 $\eta$ 都有*

$$|E\xi\eta|^2 \le E\xi^2 \cdot E\eta^2$$

*等式成立当且仅当*

$$P\{\eta = c\xi\} = 1$$

*这里 $c$ 是某一个常数.*

*证明.* 对任意实数 $t$, 定义 $u(t)=E(t\xi-\eta)^2$, 显然二次函数 $u(t)\ge 0, \forall t$, 利用判别式可得. □

将这个不等式应用到随机变量 $\frac{\xi-E\xi}{\sqrt{D\xi}}$ 与 $\frac{\eta-E\eta}{\sqrt{D\eta}}$ 上可以得到如下性质:

**定理 4.3.1.** *对相关系数 $\rho=\rho_{\xi\eta}$, 成立 $|\rho|\le 1$, 并且 $\rho=1$ 当且仅当 $P\{\frac{\xi-E\xi}{\sqrt{D\xi}}=\frac{\eta-E\eta}{\sqrt{D\eta}}\}=1$, $\rho=-1$ 当且仅当 $P\{\frac{\xi-E\xi}{\sqrt{D\xi}}=-\frac{\eta-E\eta}{\sqrt{D\eta}}\}=1$*

上述性质表明 $\rho=\pm 1$ 时, $\xi$ 与 $\eta$ 之间存在着完全线性关系. $\rho=1$ 时, 称为**完全正相关**, $\rho=-1$ 时, 称为**完全负相关**.

另一个极端是 $\rho=0$ 的场合:

**定义 4.3.4.** *若随机变量 $\xi$ 与 $\eta$ 的相关系数 $\rho=0$, 则我们称 $\xi$ 与 $\eta$* **不相关**.

根据定义即可验证如下结论:

**定理 4.3.2.** *对随机变量 $\xi$ 与 $\eta$, 下面的事实等价:*

(i) $\operatorname{cov}(\xi,\eta)=0$

(ii) $\xi$ 与 $\eta$ 不相关

(iii) $E\xi\eta=E\xi E\eta$

(iv) $D(\xi+\eta)=D\xi+D\eta$

下面的性质刻画了"独立性"与"不相关性"的联系:

**定理 4.3.3.** *若 $\xi$ 与 $\eta$ 独立, 则 $\xi$ 与 $\eta$ 不相关.*

证明. 我们只对连续型随机变量给出证明:

因为 $\xi$ 与 $\eta$ 独立, 故其密度函数 $p(x,y)=p_1(x)p_2(y)$, 因此

$$\begin{aligned}\operatorname{cov}(\xi,\eta)&=\int_{-\infty}^{\infty}(x-E\xi)(y-E\eta)p(x,y)\mathrm{d}x\mathrm{d}y\\&=\int_{-\infty}^{\infty}(x-E\xi)p_1(x)\mathrm{d}x\cdot\int_{-\infty}^{\infty}(y-E\eta)p_2(y)\mathrm{d}y=0\end{aligned}$$

□

反过来是不一定成立的. 但对于二元正态分布, 其中的参数 $\rho$ 就是其相关系数, 所以二元正态分布场合, 独立性等价于不相关性.

## 4.4 矩, 分位数, 条件数学期望

**定义 4.4.1** (原点矩). 对正整数 $k$, 称

$$m_k=E\xi^k$$

为 $k$ **阶原点矩**. 数学期望是一阶原点矩. 由于 $|\xi|^{k-1}\leq 1+|\xi|^k$, 因此若 $k$ 阶矩存在, 则所有低阶矩都存在.

**定义 4.4.2** (中心矩). 对正整数 $k$, 称

$$c_k=E(\xi-E\xi)^k$$

为 $k$ **阶中心矩**. 方差是二阶中心矩.

显然中心矩可由原点矩表示, 反之在已知数学期望后原点矩也可由中心矩表示.

**定义 4.4.3** (分位数). 对 $0<p<1$, 若

$$F(x_p)\leq p\leq F(x_p+0)$$

则称 $x_p$ 为分布函数 $F(x)$ 的 $p$ **分位数**.

最重要的分位数是 $x_{0.5}$, 称为**中位数** (median).

**定义 4.4.4** (条件数学期望). 在 $\xi=x$ 的条件下, $\eta$ 的**条件数学期望**定义为

$$E\{\eta|\xi=x\}=\int_{-\infty}^{\infty}yp(y|x)\mathrm{d}y$$

今后我们将称 $y = E\{\eta|\xi = x\}$ 是 $\eta$ 关于 $\xi$ 的**回归**.

如果以 $E\{\eta|\xi\}$ 记为随机变量 $\xi$ 的函数: 若 $\xi = x$ 时, 它取值 $E\{\eta|\xi = x\}$, 那么可以考虑它的期望, 并有以下关系式:

$$E\eta = E[E\{\eta|\xi\}]$$

这是条件数学期望的一个十分重要的性质, 称为**重期望公式**, 下面对连续型随机变量的情形进行证明:

$$\begin{aligned} E[E\{\eta|\xi\}] &= \int_{-\infty}^{\infty} E\{\eta|\xi = x\}p_1(x)\mathrm{d}x = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} yp(y|x)\mathrm{d}y\right]p_1(x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yp(x,y)\mathrm{d}x\mathrm{d}y = E\eta \end{aligned}$$

**最佳线性预测** 若 $\xi, \eta$ 是相依的随机变量, 我们要找 $\xi$ 与 $\eta$ 的函数关系, 即要找函数 $h$ 使得 $\eta$ 与 $h(\xi)$ "尽可能地接近", 这里接近的标准最常用的是高斯的最小二乘法, 即要求如下的均方误差达到最小:

$$\min_h \quad E[\eta - h(\xi)]^2$$

因为

$$\begin{aligned} E[\eta - h(\xi)]^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[y - h(x)]^2 p(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty} p_1(x)\left\{\int_{-\infty}^{\infty}[y - h(x)]^2 p(y|x)\mathrm{d}y\right\}\mathrm{d}x \end{aligned}$$

由此知道 $h(x) = E\{\eta|\xi = x\}$ 时, $\int_{-\infty}^{\infty}[y - h(x)]^2 p(y|x)\mathrm{d}y$ 达到最小, 从而使上式均方误差最小, 即当我们观察到 $\xi = x$ 时, $E\{\eta|\xi = x\}$ 是一切对 $\eta$ 的估值中均方误差最小的一个. 称 $y = E\{\eta|\xi = x\}$ 是 $\eta$ 关于 $\xi$ 的**回归**.

通常 $(\xi, \eta)$ 的联合分布函数是不知道的, 或者虽然知道但是不易算出 $E\{\eta|\xi = x\}$. 假定已知 $\xi$ 与 $\eta$ 的数学期望为 $\mu_1, \mu_2$, 标准差 $\sigma_1, \sigma_2$ 及相关系数 $\rho$, 此时可以降低要求, 改为求**最佳线性预测**. 也就是说, 把 $h(x)$ 限定为线性函数 $L(x) = a + bx$, 求 $a, b$ 使

$$e(a,b) = E[\eta - (a + b\xi)]^2$$

达到最小. 通过求偏导可以解得

$$a = \mu_2 - b\mu_1 \quad b = \rho \cdot \frac{\sigma_2}{\sigma_1}$$

于是最佳线性预测为

$$L(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

我们称上式为 $\eta$ 关于 $\xi$ 的**线性回归**. 这个结果与 $E\{\eta|\xi = x\}$ 一般是不同的, 但是在 $(\xi, \eta)$ 是二元正态分布的场合, 两者是重合的, 最佳预测是线性预测.

进一步, 我们可以计算最佳线性预测的均方误差:

$$E[\eta - L(\xi)]^2 = \sigma_2^2(1-\rho^2)$$

最佳线性预测理论中的一个重要事实是: 预测值 $\hat{\eta} = L(\xi)$ 与残差 $\eta - \hat{\eta}$ 是不相关的, 即

$$\mathrm{cov}(\hat{\eta}, \eta - \hat{\eta}) = 0$$

这个事实可以解释为: 残差中已不再包含对预测 $\eta$ 有用的知识. 因此观察值 $\eta$ 被分解为两个不相关的的随机变量之和:

$$\eta = \hat{\eta} + (\eta - \hat{\eta})$$

以上是**二阶矩理论**, 或称**均值-方差理论**, 它以最小二乘法为准则, 研究最佳线性预测, 是概率论中最有实用价值的理论之一.

## 4.5 熵与信息

为了从数值上估计各种随机试验的不确定性程度, 香农(Shannon)推导出满足三个期望基本性质(书上有讲)的唯一函数, 并称其为熵:

**定义 4.5.1** (熵). *若研究的随机试验 $\alpha$ 只有有限个不相容的结果 $A_1, \cdots, A_n$, 它们相应的概率为 $p(A_1), \cdots, p(A_n)$, 称*

$$H(\alpha) = -\sum_{i=1}^{n} p(A_i)\log p(A_i)$$

*为试验 $\alpha$ 的**熵*** (entropy).

下面考虑熵的基本性质:

性质1 在有 $n$ 个可能结果的试验中, 等概试验具有最大熵, 其值为 $\log n$.

*证明.* $\varphi(x) = -x\log x$ 是凹函数, 于是由 Jensen 不等式即得. □

性质2 若试验 $\alpha$ 与试验 $\beta$ 独立, 则

$$H(\alpha\beta) = H(\alpha) + H(\beta)$$

*证明.* 此时 $p(A_kB_l) = p(A_k)p(B_l)$, 因此

$$H(\alpha\beta) = -\sum_{k,l} p(A_k)p(B_l)\log p(A_k)p(B_l) = H(\alpha) + H(\beta)$$

□

为了进一步研究熵的性质, 引进条件熵的概念, 设 $\alpha, \beta$ 是两个试验, 以 $p(B_l|A_k)$ 记试验 $\alpha$ 出现结果 $A_k$ 的条件下, 试验 $\beta$ 出现结果 $B_i$ 的概率, 则:

$$H_{A_k}(\beta) = -\sum_{i=1}^{n} p(B_l|A_k) \log p(B_l|A_k)$$

是在试验 $\alpha$ 出现 $A_k$ 的条件下, 试验 $\beta$ 的熵.

我们称平均值

$$H_\alpha(\beta) = \sum_{k=1}^{m} p(A_k) H_{A_k}(\beta)$$

为在试验 $\alpha$ 实现的条件下试验 $\beta$ 的**条件熵**.

下面指出 $H_\alpha(\beta)$ 的重要性质:

性质1: $H(\alpha\beta) = H(\alpha) + H_\alpha(\beta)$

性质2: $H_\alpha(\beta) \leq H(\beta)$

性质2的含义不难理解, 因为在进行试验 $\alpha$ 后, 一般对试验 $\beta$ 的结果会增加了解, 从而消除部分不确定性, 只有当两个试验独立时, $H_\alpha(\beta) = H(\beta)$, 因此量 $H(\beta) - H_\alpha(\beta)$ 是作了辅助试验 $\alpha$ 之后试验 $\beta$ 不肯定性的减少量, 即是由试验 $\alpha$ 得到的关于试验 $\beta$ 的信息, 此量也称为试验 $\alpha$ 中的有关试验 $\beta$ 的**信息量**.

下面介绍连续型分布的熵:

**定义 4.5.2.** *设随机变量 $\alpha, \beta$ 的密度函数分别为 $p(x), q(x)$, 它们的联合密度函数为 $f(x, y)$. 仿照离散场合定义*熵*:*

$$H(\alpha) = -\int_{-\infty}^{\infty} p(x) \log p(x) \mathrm{d}x$$

$$H(\alpha\beta) = -\int\int f(x, y) \log f(x, y) \mathrm{d}x\mathrm{d}y$$

$$H_\alpha(\beta) = -\int\int f(x, y) \log \frac{f(x, y)}{p(x)} \mathrm{d}x\mathrm{d}y$$

下面是连续型分布的熵的性质:

性质1: 若 $\alpha$ 限制在 $V$ 种变化, 则 $V$ 中的均匀分布有最大熵, 其值等于 $\log |V|$, 此处 $|V|$ 是 $V$ 的测度.

性质2: $H(\alpha\beta) = H(\alpha) + H_\alpha(\beta) = H(\beta) + H_\beta(\alpha)$, 而且 $H_\alpha(\beta) \leq H(\beta)$

性质3: 设 $p(x)$ 是一元密度函数, 其标准差为 $\sigma$, 则当 $p(x)$ 为正态分布时其熵最大, 其值等于 $\log \sqrt{2\pi e}\sigma$

*证明.* 此时要求 $p(x)$ 满足约束条件

$$\int p(x) \mathrm{d}x = 1$$

$$\sigma^2 = \int x^2 p(x) \mathrm{d}x$$

又使

$$H(x) = -\int p(x) \log p(x) \mathrm{d}x$$

达到最大. 根据变分法, 这相当于要求

$$\int \left[ -p(x) \log p(x) + \lambda p(x) + \mu x^2 p(x) \right] \mathrm{d}x$$

达到最大, 即

$$-1 - \log p(x) + \lambda + \mu x^2 = 0$$

选取常数使其满足约束条件, 即得

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

□

性质4: 若密度函数 $p(x)$ 当 $x \leq 0$ 时等于 0, 并且其均值为 $a$, 则指数分布 $\mathrm{Exp}(\frac{1}{a})$, 即

$$p(x) = \frac{1}{a} e^{-x/a} \quad (x > 0)$$

达到最大熵, 其值为 $\log ea$.

熵是**信息论**中的基本概念, 它的引入使得人们能对不肯定性进行度量, 具有重大意义.

## 4.6 母函数

我们称取值为非负整数值的随机变量为**整值随机变量**, 对于整值随机变量, 母函数法十分便于应用.

**定义 4.6.1** (母函数). *若整值随机变量 $xi$ 的分布列为 $P\{\xi = i\} = p_i \quad (i = 0, 1, \cdots)$, 则称*

$$P(s) = \sum_{k=0}^{\infty} p_k s^k$$

*为 $\xi$ 的**母函数*** (generating function).

由随机变量函数的数学期望的计算可知

$$P(s) = Es^{\xi}$$

由于

$$\sum_{k=0}^{\infty} p_k = 1$$

由幂级数的收敛性知道 $P(s)$ 至少在 $|s| \leq 1$ 一致收敛且绝对收敛. 因此母函数对任何整值随机变量都存在.