概率论与数理统计知识点整理

初等概率论部分

KujoStar-hmc

2022 年 1 月 11 日

前言

作者于 2022-2023 学年秋季学期上了唐宏岩老师的概率论与数理统计课程，一学期颇受折磨，但是唐老师教学水平高超、教学内容丰富，若是用心理解钻研，确实会学到不少东西。前几天刚刚考完期末考试，想着可以趁着得闲将一学期所学知识精简整理，希望可以对以后选修概率论与数理统计课程（不局限于唐老师）的同学有一点微薄的帮助。

除了本人笔记外，此文档也参考了 lambda 的听课笔记以及 2099 的初概复习讲义进行整理归纳，特此致谢。

一、常用的计算恒等式

指数函数的级数求和表示

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

与二项分布相关的常见积分

$$\int_{-\infty}^{+\infty} e^{-\frac{x^2+a^2}{2}} \mathrm{d}x = \sqrt{2\pi} e^{-\frac{a^2}{2}}$$

$$\int_{-\infty}^{+\infty} e^{-\frac{(x+a)^2}{2\sigma^2}} \mathrm{d}x = \sqrt{2\pi}\sigma$$

与 Gamma 函数相关的计算式

$$\int_0^1 u^{a-1}(1-u)^{b-1} \mathrm{d}u = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} \mathrm{d}t$$

组合数相关恒等式

$$k\binom{n}{k} = n\binom{n-1}{k-1}$$

$$k^2\binom{n}{k} = n\binom{n-1}{k-1} + n(n-1)\binom{n-2}{k-2}$$

$$\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k} = (p+q)^n$$

另外，打好微积分和线性代数的基础、掌握这两门数学基础课中的一些思想对概统的学习非常重要。

二、事件的概率

1. 事件和与差的概率

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

对于事件和的概率推广到一般性，我们有

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{r=1}^{n}(-1)^{r+1}\sum_{i_1<i_2<\cdots i_r}^{i_k\in\{1,\cdots,n\}} P\left(\prod_{k=1}^{r} A_{i_k}\right)$$

2. 条件概率

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

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

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

3.Bayes 公式

设 $B_i, i=1,2,\cdots,n$ 为全事件 Ω 的一个划分，则有求解条件概率的贝叶斯公式：

$$\forall i, P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

三、随机变量及其分布

如无特殊说明，本节涉及的 $F(x)$ 均表示随机变量 X 的 cdf，$f(x)$ 表示随机变量 X 的 pmf/pdf，$E(X)$ 表示随机变量 X 的期望，$Var(X)$ 表示随机变量 X 的方差。

首先，我们有 $\sum_{i=1}^{\infty} f(x) = 1$（离散）或 $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$（连续）。这是由 pmf 和 pdf 的定义保证的。

1. 常见的离散分布

Bernouli 分布（两点分布）

若随机变量 X 满足

$$X = \begin{cases} 1, & p \\ 0, & 1-p \end{cases}, p \in (0,1)$$

则称 X 服从概率为 p 的 Bernouli 分布，记作

$$X \sim B(p)$$

我们有

$$E(X) = p, Var(X) = p(1-p)$$

二项分布

二项分布其实说明了 n 次独立的 Bernouli 实验的成功次数。记 X 为这一次数，我们有

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, \cdots, n$$

则 X 服从二项分布，记作

$$X \sim B(n, p)$$

我们有

$$E(X) = np, Var(X) = np(1-p)$$

Poisson 分布

若随机变量 X 满足

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \lambda > 0, k \in \mathbb{N}$$

则称 X 服从参数为 λ 的 Poisson 分布，记作

$$X \sim P(\lambda)$$

我们有

$$E(X) = Var(X) = \lambda$$

Remark：二项分布可以用 Poisson 分布来近似。具体来说，若 $X \sim B(n,p)$，则可以近似认为 $X \sim P(\lambda)$，其中参数 $\lambda = np$。当 p 很小，n 很大，λ 不太大时，近似的效果很好，其误差上限为 $\min\{p, np^2\}$。

2. 常见的连续分布

均匀分布

若随机变量 X 的 pdf 为

$$f(x)=\begin{cases}\frac{1}{b-a}, & x\in[a,b]\\ 0, & \text{otherwise}\end{cases}$$

则称 X 服从区间 $[a,b]$ 上的均匀分布，记作

$$X\sim U(a,b)$$

我们有

$$E(X)=\frac{a+b}{2},Var(X)=\frac{1}{12}(a-b)^2$$

Remark：$X\sim U(0,1)$ 通常称为随机数。

正态分布

若随机变量 X 的 pdf 为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},x\in\mathbb{R}$$

则称 X 服从正态分布，记作

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

我们有

$$E(X)=\mu,Var(X)=\sigma^2$$

Remark1：$N(0,1)$ 称为标准正态分布。

Remark2：正态分布标准化：

$$X\sim N(\mu,\sigma^2)\Leftrightarrow Y=\frac{X-\mu}{\sigma^2}\sim N(0,1)$$

指数分布

若随机变量 X 的 pdf 为

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

则称 X 服从指数分布，记作

$$X\sim Exp(\lambda)$$

我们有

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

Remark：此时随机变量 X 的 cdf 为

$$F(x)=\begin{cases}1-e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}$$

Beta 分布

若随机变量 X 的 pdf 为

$$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},0\leq x\leq 1$$

其中

$$\Gamma(x)=\int_0^{+\infty}e^{-t}t^{x-1}\mathrm{d}t$$

则称 X 服从 Beta 分布，记作

$$X\sim \mathrm{Beta}(a,b)$$

我们有

$$E(X)=\frac{a}{a+b},Var(X)=\frac{ab}{(a+b)^2(a+b+1)}$$

Remark：Gamma 函数有如下性质：

$$\Gamma(x+1)=x\Gamma(x)$$

利用 Gamma 函数的定义以及分部积分法易证。

卡方分布

设随机变量 $X_i(i=1,2,\cdots,n)$ 相互独立，且 $X_i\sim N(0,1)$，设随机变量 $Q=\sum\limits_{i=1}^{n}X_i^2$，则有 Q 服从自由度为 n 的卡方分布，记作

$$Q\sim\chi^2(n)$$

3. 随机变量的函数

设 X 为一个随机变量，$g(x)$ 为在 X 取值范围内有定义的函数，若随机变量 $Y=g(X)$，则称 Y 为随机变量 X 的函数。

pdf 公式推导

令 X 为连续型随机变量，其 pdf 为 $f(x)$，假设 $g(x)$ 为严格单调可微函数，则 $Y=g(X)$ 的 pdf 为

$$f_Y(y)=\begin{cases} f(h(y))h'(y), & \alpha\leq y\leq\beta \\ 0, & \text{otherwise}\end{cases}$$

其中

$$\alpha=\min\{g(-\infty),g(+\infty)\},\beta=\max\{g(-\infty),g(+\infty)\},h(x)=g^{-1}(x)$$

利用分布函数的定义易证。

典例

对数正态分布的概率密度函数。设随机变量 $Y>0$，$X=\ln Y\sim N(\mu,\sigma^2)$，则 Y 的 pdf 为

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

四、联合分布

本节研究的对象为随机向量：

$$(X_1,\cdots,X_n),\Omega\to\mathbb{R}^n$$

随机向量的每个分量都是同一个 Ω 上的随机变量。大部分时候主要研究二元分布，此时的随机向量记为 (X,Y)，联合 cdf 记为 $F(x,y)$，联合 pdf 记为 $f(x,y)$。

设 n 元分布的 pdf 为 $f(x_1,x_2,\cdots,x_n)$，由概率的定义，我们首先有

$$\sum_{\mathbf{x}\in\mathbb{R}^n}f(\mathbf{x})=1$$

或者

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

1. 常见的联合分布

多项分布

设某个随机试验有 n 个互斥的可能结果：$A_1,A_2,\cdots,A_n$，$\{A_i\}$ 是全事件 Ω 的一个划分，A_i 发生的概率是 p_i，A_i 出现的次数为 X_i。设 $\sum_{i=1}^n k_i=N$，N 为总实验次数，则我们有

$$P(X_1=k_1,\cdots,X_n=k_n)=\binom{N}{k_1,\cdots,k_n}\prod_{i=1}^n p_i^{k_i}$$

其中

$$\binom{N}{k_1,\cdots,k_n}=\frac{N!}{\prod_{i=1}^n k_i!}$$

二元均匀分布

pdf 为

$$f(x,y)=\begin{cases}\frac{1}{(b-a)(d-c)}, & (x,y)\in(a,b)\times(c,d)\\ 0, & \text{otherwise}\end{cases}$$

Remark：二元分布又被称为矩形分布。

二元正态分布

$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，其中 $|\rho|<1$，则 (X,Y) 服从二元正态分布，其联合 pdf 为

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

2. 边际分布

边际分布的 cdf

我们考虑随机向量 $(X_1, X_2, \cdots, X_n)$，则我们定义 X_i 的边际 cdf:

$$F_i(x) := P(X_i \le x) = P(X_i \le x, \forall j \neq i, X_j \in \mathbb{R})$$

一般研究二元分布 (X, Y)，设联合 cdf 为 $F(x, y)$，则我们有 X 的边际 cdf 为

$$\begin{aligned} F_X(x) :=& P(X \le x) \\ =& P(X \le x, -\infty < y < +\infty) \\ =& \lim_{y \to +\infty} P(X \le x, -\infty < Y \le y) \\ =& \lim_{y \to +\infty} F(x, y) \end{aligned}$$

Y 的边际 cdf 可以类似得到。

边际分布的 pdf

边际分布描述了随即向量每个分量单独的分布情况。此处以连续型二元分布 (X, Y) 为研究的对象。设 (X, Y) 的联合 pdf 为 $f(x, y)$，由上面对边际 cdf 的讨论，我们有

$$F_X(x) = \lim_{y \to +\infty} F(x, y) = \int_{-\infty}^{x} \int_{\mathbb{R}} f(t, y) \mathrm{d}y \mathrm{d}t$$

我们熟知，随机变量 X 的 cdf$F(x)$ 与 pdf$f(x)$ 之间有如下关系:

$$F'(x) = f(x)$$

则对上面得到的式子对 x 求导，我们有

$$f_X(x) = \int_{\mathbb{R}} f(x, y) \mathrm{d}y$$

这就是 X 的边际 pdf。同理我们得到 Y 的边际 pdf 为

$$f_Y(y) = \int_{\mathbb{R}} f(x, y) \mathrm{d}x$$

典例

我们设随机向量 $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则我们有如下结论:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, x \in \mathbb{R}$$

也即

$$X \sim N(\mu_1, \sigma_1^2)$$

同理，我们可以得到

$$Y \sim N(\mu_2, \sigma_2^2)$$

结论的推导对二元正态分布的联合 pdf 直接单侧积分即可，读者可以检验自己的微积分水平（逃）。

Remark：联合分布可以确定边际分布，但是反之未必成立。

3. 条件分布

条件分布描述了在一些分量的值已经确定的情况下，随机向量中另外的分量的分布情况。绝大部分时候，我们仅仅以二元随机向量 (X,Y) 为研究对象。

离散型随机向量

我们设

$$P(X=a_i,Y=b_j)=p_{ij},\sum_{i,j}p_{ij}=1$$

则我们有

$$P(X=a_i|Y=b_j)=\frac{P(X=a_i,Y=b_j)}{P(Y=b_j)}=\frac{p_{ij}}{\sum_k p_{kj}}$$

Remark：

$$\sum_i P(X=a_i|Y=b_j)=1$$

连续型随机向量

我们设 (X,Y) 的 pdf 为 $f(x,y)$，则我们有

$$\begin{aligned}&P(X\le x|y\le Y\le y+\mathrm{d}y)\\=&\frac{P(X\le x,y\le Y\le y+\mathrm{d}y)}{P(y\le Y\le y+\mathrm{d}y)}\\=&\frac{\int_{-\infty}^{x}\left(\int_y^{y+\mathrm{d}y}f(s,t)\mathrm{d}t\right)\mathrm{d}s}{\int_y^{y+\mathrm{d}y}f_Y(s)\mathrm{d}s}\end{aligned}$$

对 x 求导可以得到 X 此时的条件概率密度为

$$f_X(x|y\le Y\le y+\mathrm{d}y)=\frac{\int_y^{y+\mathrm{d}y}f(x,t)\mathrm{d}t}{\int_y^{y+\mathrm{d}y}f_Y(s)\mathrm{d}s}$$

令 $\mathrm{d}y\to 0$，我们可以得到

$$f_X(x|Y=y)=\frac{f(x,y)}{f_Y(y)}$$

于是可以定义 X 的条件密度函数：

$$f_X(x|y):=\frac{f(x,y)}{f_Y(y)}$$

类似地，我们有

$$f_Y(y|x):=\frac{f(x,y)}{f_X(x)}$$

典例

设 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则根据条件分布的定义，简单计算可得

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

也即当 $X=x$ 时，我们得到 Y 的分布：

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

4. 条件分布下的全概率公式

X 离散，Y 离散

$$P(X=x_i)=\sum_j P(X=x_i|Y=y_j)P(Y=y_j)$$

X 离散，Y 连续

$$P(X=x_i)=\int_{\mathbb{R}}P(X=x_i|Y=y)f_Y(y)\mathrm{d}y$$

X 连续，Y 离散

$$f_X(x)=\sum_i f_X(x|y_i)P(Y=y_i)$$

X 连续，Y 连续

$$f_X(x)=\int_{\mathbb{R}}f(x,y)\mathrm{d}y=\int_{\mathbb{R}}f_X(x|y)f_Y(y)\mathrm{d}y$$

5. 条件分布下的 Bayes 公式

X 离散，Y 离散

$$P(Y=y_i|X=x_i)=\frac{P(X=x_i|Y=y_i)P(Y=y_i)}{P(X=x_i)}=\frac{P(X=x_i|Y=y_i)P(Y=y_i)}{\sum_j P(X=x_i|Y=y_j)P(Y=y_j)}$$

X 离散，Y 连续

$$P(Y=y|X=x_i)=\frac{P(X=x_i|Y=y)f_Y(y)}{P(X=x_i)}=\frac{P(X=x_i|Y=y)f_Y(y)}{\int_{\mathbb{R}}P(X=x_i|Y=y)f_Y(y)\mathrm{d}y}$$

X 连续，Y 离散

$$P(Y=y_i|X=x)=\frac{f_X(x|y_i)P(Y=y_i)}{f_X(x)}=\frac{f_X(x|y_i)P(Y=y_i)}{\sum_i f_X(x|y_i)P(Y=y_i)}$$

X 连续，Y 连续

$$f_Y(y|x)=\frac{f_X(x|y)f_Y(y)}{f_X(x)}=\frac{f_X(x|y)f_Y(y)}{\int_{\mathbb{R}} f_X(x|y)f_Y(y)\mathrm{d}y}$$

6. 随机向量分量之间的独立性

设随机向量 (X,Y) 的 cdf 为 $F(x,y)$，若 $F(x,y)=F_X(x)F_Y(y)$，则称 X,Y 独立。
Remark：X,Y 独立 $\Leftrightarrow f(x,y)=f_X(x)f_Y(y),\forall x,y\in\mathbb{R}$

五、多个随机变量的函数

设有 n 个随机变量 $X_i,i=1,2,\cdots,n$，函数 $g(x_1,x_2,\cdots,x_n)$ 为 n 元函数，设有随机变量 Y，满足 $Y=g(X_1,X_2,\cdots,X_n)$，则称 Y 为随机变量 $X_1,X_2,\cdots,X_n$ 的函数。本节主要研究 Y 的相关性质。

1. 随机变量的和的分布

相互独立的二项分布的和

设 $X_i\sim B(n_i,p)$，$i=1,2$，X_1,X_2 独立，设 $Y=X_1+X_2$，则有结论：

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

Remark：利用二项分布的概率表达式，对其中一个变量进行求和即可。也即利用

$$P(Y=k)=P(X_1+X_2=k)=\sum_{k_1=0}^{k}P(X_1=k_1)P(X_2=k-k_1)$$

相互独立的 Poisson 分布的和

设 $X_i\sim P(\lambda_i)$，$i=1,2$，X_1,X_2 独立，设 $Y=X_1+X_2$，则有结论：

$$Y\sim P(\lambda_1+\lambda_2)$$

二元正态分布分量的和

设 $(X_1,X_2)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则有

$$X_1+X_2\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2)$$

2. 随机变量的商的分布

设 (X_1, X_2) 有联合 pdf$f(x_1, x_2)$，设 $Y = \dfrac{X_2}{X_1}$，其中 $X_1 > 0$，研究 Y 的分布。

根据定义，对 $\forall y$，我们有

$$
\begin{aligned}
& P(Y \leq y) \\
=& P\left(\frac{X_2}{X_1} \leq y\right) \\
=& P(X_2 \leq yX_1) \\
=& \iint_{x_2 \leq yx_1} f(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2 \\
=& \int_0^{+\infty} \left(\int_{-\infty}^{yx_1} f(x_1, x_2) \mathrm{d}x_2\right) \mathrm{d}x_1 \\
\xlongequal{x_2 = tx_1}& \int_{-\infty}^{y} \left(\int_0^{+\infty} x_1 f(x_1, x_1 t) \mathrm{d}x_1\right) \mathrm{d}t
\end{aligned}
$$

对 y 求导，得到 Y 的 pdf 为

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

3. 其他特殊分布

t 分布

设随机变量 $X \sim N(0,1)$，随机变量 $Y \sim \chi^2(n)$，设随机变量 $Z = \dfrac{X}{\sqrt{\frac{Y}{n}}}$，则有 Z 服从自由度为 n 的 t 分布，记作

$$
Z \sim t(n)
$$

F 分布

设随机变量 $X \sim \chi^2(n_1)$，随机变量 $Y \sim \chi^2(n_2)$，设随机变量 $F = \frac{\frac{X}{n_1}}{\frac{Y}{n_2}}$，则有 F 服从自由度为 n_1 和 n_2 的 F 分布，记作

$$
F \sim F(n_1, n_2)
$$

3. 密度函数变换法

设随机向量 (X_1, X_2) 有联合 pdf$f(x_1, x_2)$，设 $(Y_1, Y_2) = (g_1(X_1, X_2), g_2(X_1, X_2))$，且该变换可逆可微，也即 $\exists h_1, h_2$，使得 $(X_1, X_2) = (h_1(Y_1, Y_2), h_2(Y_1, Y_2))$，则 (Y_1, Y_2) 的联合 pdf 为

$$
f(h_1(y_1, y_2), h_2(y_1, y_2)) \left|\det J\right|
$$

其中矩阵 J 为变换的 Jacobi 矩阵，也即

$$
J = \frac{\partial(h_1, h_2)}{\partial(y_1, y_2)} = \begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{bmatrix}
$$

典例

设 (X_1, X_2) 的 pdf 为 $f(x_1, x_2)$，利用密度函数变换法求 $Y = X_1 + X_2$ 的 pdf。
我们令

$$Y = X_1 + X_2, Z = X_1$$

则有逆变换为

$$X_1 = Z, X_2 = Y - Z$$

对应的 $\det J$ 为

$$\begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$$

则 Y, Z 的联合 pdf 为

$$l(y,z) = f(z, y-z)|-1| = f(z, y-z)$$

Y 的 pdf 为

$$l_Y(y) = \int_{\mathbb{R}} l(y,z)\mathrm{d}z = \int_{\mathbb{R}} f(z, y-z)\mathrm{d}z$$

Remark：若 X_1, X_2 独立，则有

$$l_Y(y) = \int_{\mathbb{R}} f_{X_1}(z) f_{X_2}(y-z)\mathrm{d}z$$

这是卷积和积分变换的一种直观解释。

4. 次序统计量

设 $X_i (i = 1, 2, \cdots, n)$ 为来自同一总体的样本，总体 cdf 为 $F(x)$，总体 pdf 为 $f(x)$，$X_{(k)}$ 表示从小到大排列后次序为 k 的随机变量，例如 $X_{(1)}$ 表示 $\min\{X_1, X_2, \cdots, X_n\}$，$X_{(n)}$ 表示 $\max\{X_1, X_2, \cdots, X_n\}$。

单个次序统计量的分布

次序统计量 $X_{(k)}$ 的 pdf 为

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k} f(x)$$

两个次序统计量的联合分布

任意两个次序统计量 $X_{(i)}, X_{(j)} (i < j)$ 的联合 pdf 为

$$f_{i,j}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x)]^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j} f(x) f(y), x < y$$

以上结论均采用 pdf 的极限形式进行证明，可以参考作者的知乎专栏：前往奇妙世界

六、随机变量的数字特征

1. 期望

定义

离散型随机变量，$f(x)$ 代表 pmf：

$$E(X)=\sum_{x\in\mathbb{R}}xf(x)$$

连续型随机变量，$f(x)$ 代表 pdf：

$$E(x)=\int_{\mathbb{R}}xf(x)\mathrm{d}x$$

Remark：由期望不能推得 $f(x)$。

相关性质

1. 随机变量的函数的期望。设 f 代表 pmf 或 pdf，则有

$$E\left(g(X_1,X_2,\cdots,X_n)\right)=\begin{cases}\sum_{\mathbb{R}^n}g(x_1,\cdots,x_n)f(x_1,\cdots,x_n)\\ \int_{\mathbb{R}^n}g(x_1,\cdots,x_n)f(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n\end{cases}$$

2. 期望的线性性。设 $X_i(i=1,2,\cdots,n)$ 为随机变量，$c_i(i=1,2,\cdots,n)$ 为实数，则我们有

$$E\left(\sum_{i=1}^{n}c_iX_i\right)=\sum_{i=1}^{n}c_iE(X_i)$$

3. 相互独立事件的期望。设 $X_i(i=1,2,\cdots,n)$ 为随机变量，若 X_i 之间相互独立，则我们有

$$R\left(\prod_{i=1}^{n}X_i\right)=\prod_{i=1}^{n}E(X_i)$$

4. Cauchy-Schwarts 不等式。设 U,V 为随机变量，则成立

$$E^2(UV)\leq E(U^2)E(V^2)$$

等号成立当且仅当 $\exists t_0$，使得 $P(V=t_0U)=1$。

2. 分位数

定义

设 X 为随机变量，实数 $\alpha\in(0,1)$。若成立

$$P(X\leq a)\geq\alpha,P(X\geq a)\geq1-\alpha$$

则称 $X=a$ 为 X 的下侧 $\alpha-$ 分位数。

相关性质

1. 若随机变量 X 的 cdf$F(x)$ 连续，则有 $F(a)=\alpha$。
2. $F^{-1}(\alpha)$ 为 $\alpha-$ 分位数。
3. $\alpha=0.5$ 可以得到中位数，是随机变量集中趋势的一种刻画。

3. 方差

定义

设 X 为随机变量，则定义 X 的方差：

$$Var(X)=E\left[(X-E(X))^2\right]=E(X^2)-E^2(X)$$

同时可以定义标准差：

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

相关性质

1. 方差的极小性质。对 $\forall c\in\mathbb{R}$，我们有

$$Var(X)=E\left[(X-E(X))^2\right]\leq E\left[(X-c)^2\right]$$

2. 方差的线性性与平移不变性。对 $\forall a,b\in\mathbb{R}$，我们有

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

3. 多个随机变量的和的方差。设 $X_i(i=1,2,\cdots,n)$ 为随机变量，则有

$$Var\left(\sum_{i=1}^n X_i\right)=\sum_{i=1}^n Var(X_i)+\sum_{i\neq j}Cov(X_i,X_j)$$

其中 $Cov(X_i,X_j)$ 代表两个随机变量的协方差。若上述所有随机变量都相互独立，则有

$$Var\left(\sum_{i=1}^n X_i\right)=\sum_{i=1}^n Var(X_i)$$

4. 协方差与相关系数

定义

设随机变量 X,Y，记 $E(X)=\mu_1,E(Y)=\mu_2,Var(X)=\sigma_1^2,Var(Y)=\sigma_2^2$，则我们定义 X,Y 的协方差：

$$Cov(X,Y)=E\left[(X-\mu_1)(Y-\mu_2)\right]$$

定义 X,Y 的相关系数：

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sigma_1\sigma_2}$$

Remark：我们考虑随机向量 $(X_1, X_2, \cdots, X_n)$，期望向量记为 $(\mu_1, \mu_2, \cdots, \mu_n)$，则协方差定义为

$$E\left[\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \cdots \\ X_n-\mu_n \end{pmatrix}(X_1-\mu_1, X_2-\mu_2, \cdots, X_n-\mu_n)\right] = (Cov(X_i, X_j))_{n\times n}$$

这被称为随机向量的协方差矩阵。

相关性质

1. 协方差的另一种计算式。对定义式展开可得
$$Cov(X,Y) = E(XY) - E(X)E(Y)$$
2. 协方差与期望的关系。我们有
$$Cov(X,X) = Var(X)$$
3. 协方差的交换性。我们有
$$Cov(X,Y) = Cov(Y,X)$$
4. 协方差的双线性性与平移不变性。对 $\forall a, b, c \in \mathbb{R}$，我们有
$$Cov(aX + bY + c, Z) = aCov(X,Z) + bCov(Y,Z)$$
5. 协方差的展开。设 X_i, Y_j 分别为一系列随机变量，我们有
$$Cov\left(\sum_i X_i, \sum_j Y_j\right) = \sum_{i,j} Cov(X_i, Y_j)$$
这个公式用于形如 $Var\left(\sum_i X_i\right)$ 的方差计算之中。
6. 相关系数的有界性。设 X, Y 为随机变量，则我们有 $|Corr(X,Y)| \le 1$，等号成立当且仅当 $\exists a, b \in \mathbb{R}$，使得 $P(Y = aX + b) = 1$。

5. 矩

定义

设 X 为随机变量，则 $E\left[(X-c)^k\right]$ 称为 X 关于 c 的 k 阶矩。其中 $c \in \mathbb{R}, k \in \mathbb{N}_+$。
特别地，当 $c = 0$ 时，称为原点矩。当 $c = E(X)$ 时，称为中心距。
设随机变量 X 的期望为 μ，方差为 σ^2，则称 $E\left[\left(\frac{X-\mu}{\sigma}\right)^k\right]$ 为 k 阶标准矩。

一些特殊的矩

1. 期望是一阶原点矩，一阶中心距恒为 0。
2. 方差是二阶中心矩。
3. 一阶标准矩恒为 0，二阶标准矩恒为 1。
4. 三阶标准矩 $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ 称为偏度系数，四阶标准矩 $E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ 称为峰度系数。
5. 正态分布的峰度系数恒为 3。

6. 矩母函数

定义

设 X 为随机变量，设

$$M_X(t) = E\left(e^{tX}\right)$$

若在 $t=0$ 的某个邻域内，$M_X(t)$ 存在，则称 $M_X(t)$ 为 X 的矩母函数。

相关性质

1. 在原点处，任何矩母函数值均为 1，也即 $M_X(0)=1$。
2. 设 X 为随机变量，矩母函数为 $M_X(t)$，设 $Y=aX+b$，其中 $a,b\in\mathbb{R}$，设 Y 的矩母函数为 $M_Y(t)$，则我们有
$$M_Y(t) = e^{tb}M_X(at)$$
3. 矩母函数确定矩。我们有 $E(X^n)=M_X^{(n)}(0)$。结论的证明利用泰勒展开：
$$M_X(t) = \sum_{n=0}^{\infty} M_X^{(n)}(0)\frac{t^n}{n!}$$
$$M_X(t) = E(e^{tX}) = E\left(\sum_{n=0}^{\infty} X^n\frac{t^n}{n!}\right) = \sum_{n=0}^{\infty} E(X^n)\frac{t^n}{n!}$$
通过对比即可得到结论。
4. 矩母函数确定分布。设随机变量 X,Y，矩母函数分别为 $M_X(t),M_Y(t)$，若存在实数 $a>0$，使得 $M_X(t)=M_Y(t),\forall t\in(-a,a)$，则可得 X,Y 同分布。
5. 独立随机变量和的分布。设随机变量 $X_i(i=1,2,\cdots,n)$ 相互独立，矩母函数分别为 $M_{X_i}(t)$，设 $Y=\sum_{i=1}^{n}X_i$，则 Y 的矩母函数为
$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$$
结论的证明利用矩母函数的定义即可得到。

常见分布的矩母函数

1. 指数分布的矩母函数。设随机变量 $X \sim Exp(\lambda)$，则 X 的矩母函数为

$$M_X(t) = \frac{\lambda}{\lambda - t}, t < \lambda$$

2. Poisson 分布的矩母函数。设随机变量 $X \sim P(\lambda)$，则 X 的矩母函数为

$$M_X(t) = e^{\lambda(e^t - 1)}$$

3. 均匀分布的矩母函数。设随机变量 $X \sim U(a, b)$，则 X 的矩母函数为

$$M_X(t) = \frac{e^t - 1}{t(b-a)}$$

4. 正态分布的矩母函数。设随机变量 $X \sim N(\mu, \sigma^2)$，则 X 的矩母函数为

$$M_X(t) = e^{\frac{\sigma^2}{2}t^2 + \mu t}$$

以上结论都可以通过矩母函数的定义计算得出。

7. 条件期望

定义

设有随机变量 X, Y，考虑 $X = x$ 时 Y 的期望：

$$E(Y|x) := E(Y|X = x) = \begin{cases} \sum_i y_i P(Y = y_i | X = x) \\ \int_{\mathbb{R}} y f_Y(y|x) \mathrm{d}y \end{cases}$$

则我们可以得到一个函数 $m(x) = E(Y|x)$，进一步得到随机变量 $m(X) = E(Y|X)$，称为给定 X 时 Y 的条件期望。

相关性质

1. 全期望公式。设 X, Y 为随机变量，则有

$$E(Y) = E\left[E(Y|X)\right]$$

这是由条件期望的定义自然得出的结论。

2. 条件期望的极小性。设 X, Y 为随机变量，g 为实函数，则有

$$E\left[(Y - g(X))^2\right] \geq E\left[(Y - E(Y|X))^2\right]$$

利用方差的极小性质，两边对 X 作条件期望可以证明。

3. 设 g, h 为实函数，X, Y 为随机变量，则有

$$E\left[h(X)g(Y)|X\right] = h(X)E\left[g(Y)|X\right]$$

4. 设 g 为实函数，X,Y 为随机变量，则有

$$E\left[E(g(Y)|X)\right] = E(g(Y))$$

5. 设随机变量 X,Y 独立，g 为实函数，则有

$$E\left[g(Y)|X\right] = E(g(Y))$$

6. 设随机变量 $X_i(i=1,2,\cdots,n),Y,c,c_1,c_2,\cdots,c_n \in \mathbb{R}$，则有

$$E\left[c+\sum_{i=1}^{n} c_i X_i|Y\right] = c+\sum_{i=1}^{n} c_i E(X_i|Y)$$

7. 设 X,Y 为随机变量，h 为实函数，则有

$$E\left[(X-E(X|Y))h(Y)\right] = 0$$

可以从期望的几何意义来直观理解这个结论。$E(X|Y)$ 是 X 在 Y 的张成空间上的投影，根据矢量减法，$X-E(X|Y)$ 表示的是垂直于 Y 张成空间的“矢量”，而 $h(Y)$ 是张成空间内的任意一个组合，自然两者是垂直的，从而乘积的期望为 0.

引申：条件方差

定义条件方差为

$$Var(Y|X) = E\left[(Y-E(Y|X))^2|X\right]$$

我们有全方差公式：

$$Var(Y) = E\left[Var(Y|X)\right] + Var\left[E(Y|X)\right]$$

8. 最优线性预测问题

实际应用中常常依据随机变量 X 的观测值对随机变量 Y 的值进行预测。我们假设仅仅知道 X 和 Y 的期望分别为 μ_1 和 μ_2，方差分别为 σ_1^2,σ_2^2，相关系数为 ρ。

均方意义下 Y 的最优线性预测

也即我们要选择合适的系数 a,b 使得均方误差 $E\left[(Y-(aX+b))^2\right]$ 最小。我们直接将目标计算式展开，得到

$$E\left[(Y-(aX+b))^2\right] = E(Y^2)+a^2E(X^2)+2abE(X)-2aE(XY)-2bE(Y)+b^2$$

又因为

$$E(X^2) = Var(X)+E^2(X) = \mu_1^2+\sigma_1^2$$

$$E(Y^2) = Var(Y)+E^2(Y) = \mu_2^2+\sigma_2^2$$

$$E(XY) = E(X)E(Y)+Cov(X,Y) = E(X)E(Y)+Corr(X,Y)\sigma_1\sigma_2 = \rho\sigma_1\sigma_2+\mu_1\mu_2$$

代入原式，得到均方误差为

$$f(a,b)=\sigma_2^2+\mu_2^2+a^2(\mu_1^2+\sigma_1^2)+2ab\mu_1-2a(\rho\sigma_1\sigma_2+\mu_1\mu_2)-2b\mu_2+b^2$$

对 a,b 分别求偏导，令

$$\frac{\partial f}{\partial a}=\frac{\partial f}{\partial b}=0$$

得到

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

则选取如上的 a,b，可以使均方误差最小。

针对正态分布

设随机变量 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则 Y 的最优线性预测就是 $E(Y|X)$。我们可以考虑 Y 的条件分布，当 $X=x$ 时，我们有

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

也即

$$E(Y|X=x)=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$$

于是我们有

$$E(Y|X)=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(X-\mu_1)=aX+b$$

从而针对正态分布，联系条件期望，线性预测有非常简洁的结论。

9. 同分布的样本均值与方差

结论提出

设随机变量 $X_i(i=1,2,\cdots,n)$ 独立同分布，这样的序列也称为来自同一分布的样本。设公共期望为 μ，公共方差为 σ^2，定义样本均值：$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，样本方差：$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$，则我们有如下结论：

1. 样本均值的方差为
$$Var(\overline{X})=\frac{\sigma^2}{n}$$
2. 样本方差的期望为
$$E(S^2)=\sigma^2$$

证明过程

1. 由于 $X_i(i=1,2,\cdots,n)$ 相互独立，所以

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

2. 显然，我们有

$$E(\overline{X}) = \mu$$

则有

$$\begin{aligned}
&E(S^2)\\
=&\frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]\\
=&\frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2-2\overline{X}\sum_{i=1}^{n}X_i+n\overline{X}^2\right]\\
=&\frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right]\\
=&\frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i^2)-nE(\overline{X}^2)\right]
\end{aligned}$$

又因为

$$E(X_i^2) = Var(X_i)+E^2(X_i) = \sigma^2+\mu^2$$
$$E(\overline{X}^2) = Var(\overline{X})+E^2(\overline{X}) = \frac{\sigma^2}{n}+\mu^2$$

则有

$$E(S^2) == \frac{1}{n-1}\left(n\sigma^2+n\mu^2-\sigma^2-n\mu^2\right) = \sigma^2$$

10. 简单随机抽样的均值与方差

结论提出

设总体大小为 N，总体均值为 μ，总体方差为 σ^2，$X_i(i=1,2,\cdots,n)$ 为从总体中无放回抽取的 n 个简单随机样本（认为这 n 个样本相互独立）。设 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，则我们有如下结论：

1. 每个抽取的样本的期望与方差为

$$E(X_i) = \mu, Var(X_i) = \sigma^2$$

2. 样本均值的期望与方差为

$$E(\overline{X}) = \mu, Var(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

由于此处明确了从总体中抽取的过程，因此样本均值的数字特征会受到总体大小的影响。

证明过程

1. 由题意，我们可以得到

$$\mu = \frac{1}{N}\sum_{i=1}^{n} X_i, \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2$$

选取每个 X_i 的过程中，我们可以得到 X_i 的概率分布：

$$P(X_i = X_k) = \frac{1}{N}, k = 1, 2, \cdots, N$$

于是我们有

$$E(X_i) = \sum_{k=1}^{N} X_k P(X_i = X_k) = \frac{1}{N}\sum_{k=1}^{N} X_k = \mu$$

$$Var(X_i) = \frac{1}{N}\sum_{k=1}^{N}(X_k-\mu)^2 = \sigma^2$$

2. 由期望的运算法则，我们显然有

$$E(\overline{X}) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \mu$$

对于方差，我们有

$$\begin{aligned}
&Var(\overline{X})\\
=&Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\\
=&\frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)\right)\\
=&\frac{1}{n^2}\left(\sum_{i=1}^{n} Var(X_i) + n(n-1)Cov(X_i, X_j)\right)\\
=&\frac{\sigma^2}{n} + \frac{n-1}{n}Cov(X_i, X_j), \forall i, j \in [1, n]\cap\mathbb{Z}, i\neq j
\end{aligned}$$

下求 $Cov(X_i, X_j)$。我们有

$$Cov(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j) = E(X_iX_j) - \mu^2$$

考虑从总体中抽取两个不同的样本，可以得到分布

$$P(X_iX_j = X_mX_n) = \frac{1}{N(N-1)}$$

则有

$$E(X_iX_j) = \frac{1}{N(N-1)}\sum_{m\neq n} X_mX_n, m, n \in [1, n]\cap\mathbb{Z}$$

又因为

$$\left(\sum_{i=1}^{N} X_i\right)^2 = \sum_{i=1}^{N} X_i^2 + \sum_{m\neq n} X_m X_n = N^2\mu^2$$

$$E(X^2) = Var(X) + E^2(X) = \sigma^2 + \mu^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2$$

可得

$$\sum_{m\neq n} X_m X_n = N(N-1)\mu^2 - N\sigma^2$$

所以

$$E(X_i X_j) = \mu^2 - \frac{1}{N-1}\sigma^2$$

$$Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

于是得到

$$Var(\overline{X}) = \frac{\sigma^2}{n} - \frac{n-1}{n}\cdot\frac{\sigma^2}{N-1} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

七、概率不等式

1.Markov 不等式

结论

设随机变量 $Y \geq 0$，则 $\forall a > 0$，我们有

$$P(Y \geq a) \leq \frac{E(Y)}{a}$$

证明

考虑设示性变量：

$$I = \begin{cases} 1, & Y \geq a \\ 0, & Y < a \end{cases}$$

则恒成立

$$I \leq \frac{Y}{a}$$

两边取期望，可得

$$E(I) \leq E\left(\frac{Y}{a}\right)$$

也即

$$P(Y \geq a) \leq \frac{E(Y)}{a}$$

2.Chebyshev 不等式

结论

设随机变量 Y，方差 $Var(Y)$ 存在，则 $\forall a>0$，我们有

$$P\left(|Y-E(Y)|\geq a\right)\leq\frac{Var(Y)}{a^2}$$

证明

利用 Markov 不等式。我们有

$$\begin{aligned}&P\left(|Y-E(Y)|\geq a\right)\\=&P\left(\left(Y-E(Y)\right)^2\geq a^2\right)\\\leq&\frac{E\left(\left(Y-E(Y)\right)^2\right)}{a^2}\\=&\frac{Var(Y)}{a^2}\end{aligned}$$

Chernoff 不等式

结论

设随机变量 Y，则 $\forall a>0,\forall t>0$，我们有

$$P(Y\geq a)\leq\frac{E(e^{tY})}{e^{ta}}$$

证明

利用 Markov 不等式。我们有

$$P(Y\geq a)=P(e^{tY}\geq e^{ta})\leq\frac{E(e^{tY})}{e^{ta}}$$

八、大数定律（Law of Large Numbers,LLN）

1.Khinchin 弱大数定律（WLLN）

结论

设 $X_i(i=1,2,\cdots,n)$ 为独立同分布、方差有限的随机变量序列，$E(X_i)=\mu$，$Var(X_i)=\sigma^2$，设 $\overline{X}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i$，则 $\forall\varepsilon>0$，我们有

$$\lim_{n\to\infty}P\left(\left|\overline{X}-\mu\right|\geq\varepsilon\right)=0$$

也称为 $\overline{X}$ 依概率收敛到 μ，记作

$$\overline{X}\xrightarrow{P}\mu$$

证明

利用 Chebyshev 不等式，当 $n \to \infty$ 时，我们有

$$P\left(\left|\overline{X}-\mu\right| \geq \varepsilon\right) \leq \frac{Var(\overline{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

相关补充

1. 依概率收敛一般形式可以写作：

$$Y_n \xrightarrow{P} Y \Longleftrightarrow \lim_{n\to\infty} P\left(|Y_n - Y| \geq \varepsilon\right) = 0, \forall \varepsilon > 0$$

   其中 Y_n 和 Y 均为随机变量。此时并不要求 Y_n 的收敛性，考虑的是随机变量之间偏差值的概率测度。

2. 有与弱大数定律结论相似的估计概率的结论：$\forall \varepsilon > 0$，$\forall \alpha > 0$，均 $\exists N \in \mathbb{N}_+$，使得当 $n \geq N$ 时，恒成立

$$P\left(\left|\overline{X}-\mu\right| \geq \varepsilon\right) \leq \alpha$$

   其中 ε 体现了精度，α 体现了置信度。

3. 若 $X_i \sim B(p)$，则特殊地得到 Bernoulli 大数定律。

4. 关于方差的条件并不是必要的，$Var(X_i)$ 不存在时，定律依然成立。

5. 弱大数定律告诉我们，当 n 很大时，我们有很大的把握（但并非 100%）断言 $\overline{X}$ 的值会接近 μ。这是使用 Chebyshev 不等式对大样本性质的描绘，计算误差也是使用的 Chebyshev 不等式。但是弱大数定律对误差的估计很保守，依概率收敛的结论也不够强，成立条件也显得有些过于紧凑。

2.Kolmogorov 强大数定律（SLLN）

结论

设 $X_i(i=1,2,\cdots,n)$ 为独立同分布的随机变量序列，$E(X_i)=\mu$，设 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，则有

$$P\left(\lim_{n\to\infty} \overline{X} = \mu\right) = 1$$

也称为 $\overline{X}$ 几乎处处收敛到 μ，记作

$$\overline{X} \xrightarrow{a.s.} \mu$$

相关补充

1. 几乎必然收敛的一般形式可以写作：

$$Y_n \xrightarrow{a.s.} Y \Longleftrightarrow P\left(\lim_{n\to\infty} Y_n = Y\right) = 1 \Longleftrightarrow \forall \varepsilon > 0, P\left(\lim_{n\to\infty} |Y_n - Y| < \varepsilon\right) = 1$$

   此时，几乎必然收敛事先要求 Y_n 的收敛，考虑的是概率本身的测度。

2. 若 $X_i \sim B(p)$，则利用强大数定律，可以解释概率的频率解释是合理的。（回归本心）

3. 强大数定律的条件和结论是充分必要的。也即下述结论依然成立：

   设随机变量 $X_i(i=1,2,\cdots,n)$ 为独立同分布的随机变量序列，$E(X_i=\mu)$，设 $\overline{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$，若成立

$$\overline{X} \xrightarrow{a.s.} C$$

   则我们可得 $C=\mu$。

3. 分布的收敛模式

本节中出现的符号约定：设 $\{X_n\}$ 为一列随机变量，X 为一个随机变量。

依分布收敛（弱收敛）

设 X 的 cdf 为 $F(x)$，若在 $F(x)$ 的连续点 x 处，均有

$$\lim_{n\to\infty} P(X_n \leq x) = F(x)$$

则称 X_n 依分布收敛于 X，记作

$$X_n \xrightarrow{d} X$$

依分布收敛仅仅是分布的收敛，而不是随机变量的收敛。此时并不考虑随机变量之间的相互联系，只需要分布函数的极限相等即可。

依概率收敛

若 $\forall\varepsilon>0$，均成立

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \varepsilon\right) = 0$$

则称 X_n 依概率收敛到 X，记作

$$X_n \xrightarrow{P} X$$

依概率收敛是随机变量的收敛，只有当随机变量之间的差值趋于 0 时才可以称之为依概率收敛，但是对变量本身的收敛性并无要求。

几乎处处收敛

若成立

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

则称 X_n 几乎处处收敛到 X，记作

$$X_n \xrightarrow{a.s.} X$$

几乎处处收敛不仅考虑了随机变量的概率特征，还考虑了随机变量对应的样本空间的特征，也即随机变量作为一个映射的性质，比依概率收敛更加严格。

几乎处处收敛的一个等价表述是：

$$P(\{\omega|\lim_{n\to\infty}X_n(\omega)=X(\omega)\})=1$$

这样的表述更加明显地显示出了几乎处处收敛中考虑的随机变量作为样本空间上映射的性质。

L_p 收敛（仅作了解）

对 $p>0$，若成立

$$\lim_{n\to\infty}E\left(|X_n-X|^p\right)=0$$

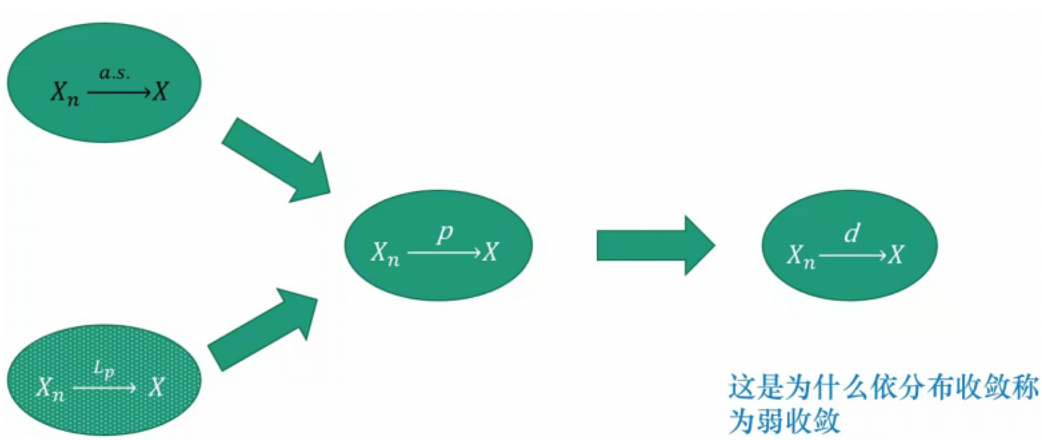
则称 X_n 在 L_p 下收敛到 X，记作

$$X_n\xrightarrow{L_p}X$$

特别地，当 $p=2$ 时称为均方收敛。

收敛模式之间的关系

这里参考初等概率论课程课件上的示意图：

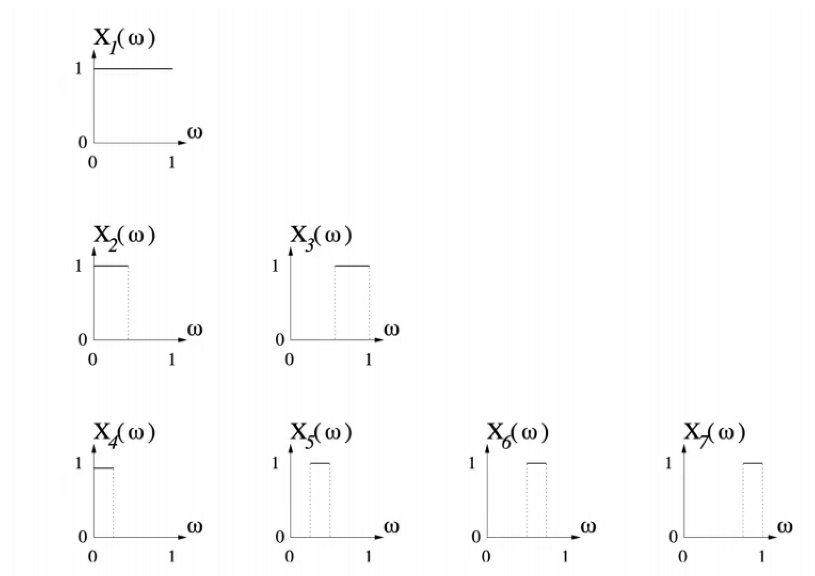


由图可以看出，依分布收敛最弱，依概率收敛其次，几乎处处收敛和 L_p 收敛可以推出前两者，但是不能相互推出。鉴于 L_p 收敛不做要求，我们给出依概率收敛不能推出几乎处处收敛的示例。

依概率收敛弱于 $a.s.$ 的示例

考虑如下定义的一组随机变量 $\{X_n(\omega)\}$：

对 $i=1,2,\cdots$，考虑 $k\in[2^{i-1},2^i-1]\cap\mathbb{Z}$，随机变量 $X_k(\omega)$ 在区间 $\left[\frac{k}{2^{i-1}}-1,\frac{k}{2^{i-1}}-1+\frac{1}{2^{i-1}}\right]$ 上取值为 1，在区间 $[0,1]/\left[\frac{k}{2^{i-1}}-1,\frac{k}{2^{i-1}}-1+\frac{1}{2^{i-1}}\right]$ 上取值为 0，也即如下图所示：



很容易可以得到

$$\lim_{n\to\infty} P\left(|X_n - 0| \geq \varepsilon\right) = 0$$

也就是依概率收敛。但是考察整个样本空间 $\Omega = [0,1]$ 上的任意一点 ω，这一点的概率总会在某个区间内变为 1，因此以下结论不成立：

$$\lim_{n\to\infty} X_n(\omega) = 0$$

由 ω 的任意性，我们得到集合 $\{\omega | \lim_{n\to\infty} X_n(\omega) = 0\}$ 是一个空集，于是我们有

$$P(\{\omega | \lim_{n\to\infty} X_n(\omega) = 0\}) = 0$$

因此几乎处处收敛不成立。

4. 收敛模式常用结论

连续映射定理

设 $\{X_n\}$ 为一随机变量序列，X 为随机变量，g 为连续实函数，则有

1. $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$
2. $X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)$
3. $X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$

注意上述结论对 L_p 收敛并不成立。

Slutsky 定理

设随机变量列 $X_n \xrightarrow{d} X$，$Y_n \xrightarrow{d} c$，c 为常数，则有

1. $X_n + Y_n \xrightarrow{d} X + c$

2. $X_nY_n \xrightarrow{d} Xc$
3. $\dfrac{X_n}{Y_n} \xrightarrow{d} \dfrac{X}{c}, Y_n \neq 0, c \neq 0$

利用这个定理可以证明 t 分布收敛到正态分布。

依概率收敛的四则运算性质

设随机变量列 $X_n \xrightarrow{P} a$，$Y_n \xrightarrow{P} b$，c 为常数，则有

1. $cX_n \xrightarrow{P} ca$
2. $X_n \pm Y_n \xrightarrow{P} a \pm b$
3. $X_nY_n \xrightarrow{P} ab$
4. $\dfrac{X_n}{Y_n} \xrightarrow{P} \dfrac{a}{b}, Y_n \neq 0, b \neq 0$

依概率收敛的四则运算性质在相关证明题中较为常用。

九、中心极限定理（CLT）

这是初等概率论部分与统计部分结合最紧密的知识点之一，常用于大样本的统计推断。

1. 结论

设 $X_i(i = 1, 2, \cdots, n)$ 为独立同分布的随机变量序列，其期望和方差均存在，期望为 μ，方差为 σ^2，则为排名，设 $S_n = \sum_{i=1}^{n} X_i$，则我们有

$$\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} \xrightarrow{d} N(0, 1)$$

也即

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0, 1)$$

在计算概率时，我们有

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x)$$

其中 $\Phi(x)$ 是 $N(0, 1)$ 的 cdf。

通过中心极限定理，我们可以将随机变量依分布收敛到标准正态分布。

2. 均值的近似分布

依照上述约定，设 $M_n = \dfrac{S_n}{n}$，也即样本均值，我们有

$$\frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0, 1)$$

通过中心极限定理，我们可以近似认为样本均值服从分布 $N\left(\mu, \dfrac{\sigma^2}{n}\right)$。

3. 二项分布下 CLT 的连续性修正

中心极限定理将分布近似为正态分布，这会导致单点概率近似为 0，对离散型随机变量的分布来说显然是不成立的。针对离散分布，可以有一种连续性修正的办法。

设 $S_n \sim B(n,p)$，n 充分大，k,m 为非负整数，则有 CLT 的修正形式：

$$P(k \leq S_n \leq m) \approx \Phi\left(\frac{m+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-0.5-np}{\sqrt{np(1-p)}}\right)$$

这样的修正形式还可以用来计算单点 $P(S_n = k)$ 的概率，同时对于其他离散变量也同样适用。