多元随机变量

司继春

[1]上海对外经贸大学

2023年10月

概览

SHANGHAI UNIVERSITY OF INTERNATIONAL BUS 1960

多元随机变量

多元随机变量的定义

（随机向量）给定一个概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，一个$n$维的随机向量$X$即从样本空间到$n$维欧几里得空间的函数，$X : \Omega \to \mathbb{R}^n$。

向量表达

注意以上定义我们使用了向量的表达方式，即：

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

多元随机变量

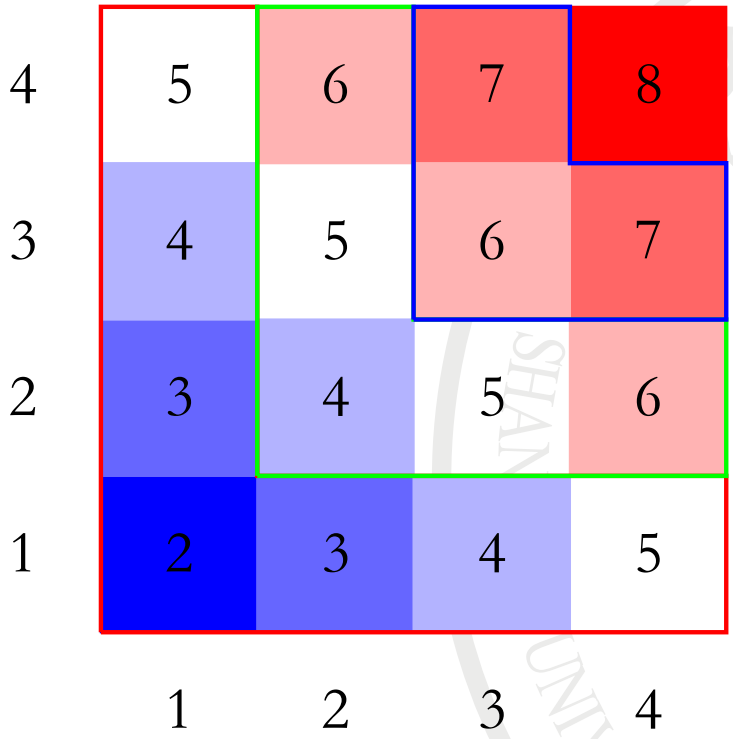


Figure: 四面骰子

多元随机变量

多元随机变量

投两个均匀的四面骰子，则$\Omega = \{(1,1),(1,2),\ldots(4,4)\}$:
1、定义随机变量Y为两个骰子的数值之和
2、定义Z为两个骰子中较小的骰子的数值，如上图所示。
那么向量$(Y,Z)' = X : \Omega \to \mathbb{R}^2$为一个随机向量，其可能的取值为$\{(y,z), y \in \{2,...,8\}, z \in \{1,2,3,4\}\}$。例如，$X^{-1}(\{(5,2)\}) = \{(2,3),(3,2)\}$

联合分布函数

联合分布函数

由$(\Omega, \mathscr{F}, \mathscr{P})$导出的概率空间$(\mathbb{R}^n, \mathscr{B}^n, P)$的**联合分布函数（joint c.d.f.）定义为：**

$$\begin{aligned} F(x) &= F(x_1, x_2, ..., x_n) \\ &= P(X_1 \leq x_1, ..., X_n \leq x_n) \\ &= P((-\infty, x_1] \times (-\infty, x_2] \times \cdots (-\infty, x_n]) \\ &= \mathscr{P}\left(X^{-1}((-\infty, x_1] \times (-\infty, x_2] \times \cdots (-\infty, x_n])\right) \end{aligned}$$

$\forall x \in \mathbb{R}^n$。

联合分布函数为单调递增

且$F(-\infty, -\infty, ..., -\infty) = 0$，$F(\infty, \infty, ..., \infty) = 1$

联合密度函数

联合密度函数

1. 如果随机向量X的每个分量都是离散型随机变量，那么可以定义联合概率质量函数p.m.f为：

$$f\left(x_1, x_2, ..., x_n\right) = P\left(\left\{X_1 = x_1, ..., X_n = x_n\right\}\right)$$

2. 如果随机变量X的联合分布函数连续，如果函数$f\left(x\right)$满足：$P\left(X \in A\right) = \int_A f\left(x\right) dx, x \in \mathbb{R}^n, A \in \mathscr{B}^n$那么我们称$f\left(x\right)$为其联合概率密度函数p.d.f。特别的，如果联合分布函数$F\left(x\right)$可微那么：

$$f\left(x_1, x_2, ..., x_n\right) = \frac{\partial^n F\left(x_1, x_2, ..., x_n\right)}{\partial x_1 \partial x_2 \cdots \partial x_n}$$

概率质量函数

概率质量函数

四面骰子例子中的概率质量函数可以用下表描述:

$Z\backslash Y$	2	3	4	5	6	7	8
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0
4	0	0	0	0	0	0	$\frac{1}{16}$

概率密度函数

概率密度函数

如果随机向量$X=(X_1,X_2)$的两个分量分别服从正态分布，且相互独立，那么其概率密度函数为：

$$f(x)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\}$$

边缘分布

- 如果$X=(X_1,...,X_n)$为随机向量，那么$\widetilde{X}=(X_{i_1},X_{i_2},...,X_{i_k}),1\leq i_1<i_2<...<i_k\leq n$也是一个随机向量。$\widetilde{X}$的联合分布函数可以通过$F(x)$来定义，即令$F(x)$中满足$j\notin\{i_1,...i_k\}$的分量为$\infty$。
- 如对于三维随机变量$X=(X_1,X_2,X_3)$，则$\widetilde{X}=(X_1,X_2)$的分布函数为：$F_{\widetilde{X}}(\widetilde{x})=F(\widetilde{x}_1,\widetilde{x}_2,\infty)$。
- 特别的，对于随机向量X的每个分量X_i，我们可以定义其边缘分布函数（marginal c.d.f.）为：

$$F_{X_i}(x_i)=F(\infty,...,x_i,...,\infty)$$

- 意边缘分布函数对应着一元随机变量X_i的分布函数。

边缘分布

边缘质量函数

四面骰子例子中，$X=(Y,Z)$，Y和Z的边缘概率质量函数如下表所示：

$Z\backslash Y$	2	3	4	5	6	7	8	F_Z	f_Z
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_Y	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_Z$ $\|\|$
f_Y	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_Y=$	1

边缘密度函数

对于随机向量$X=(X_1,...,X_n)$，其联合分布函数为$F(x)=F(x_1,x_2,...,x_n)$，那么：

- 联合密度函数为：

$$f(x)=\frac{\partial^n F(x_1,x_2,...,x_n)}{\partial x_1\partial x_2\cdots\partial x_n}$$

- 边缘密度函数：

$$f_{X_i}(x_i)=\frac{\partial F_{X_i}(x_i)}{\partial x_i}=\int_{\mathbb{R}}\cdots\int_{\mathbb{R}}f(x_1,x_2,...,x_n)\,dx_1\cdots dx_{i-1}dx_{i+1}$$

- 边缘分布函数：

$$\begin{aligned}F_{X_i}(a)&=F(\infty,...,x_i=a,...,\infty)\\&=\int_{-\infty}^{a}f_{X_i}(x_i)\,dx_i\end{aligned}$$

边缘分布

边缘密度函数

上例中的联合正态分布，其边缘分布函数为：

$$
\begin{aligned}
F_{X_1}(t) &= \int_{\mathbb{R}} \int_{-\infty}^{t} f(x_1, x_2)\, dx_1 dx_2 \\
&= \frac{1}{2\pi\sigma_1\sigma_2} \int_{\mathbb{R}} \int_{-\infty}^{t} \exp\left\{ -\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2} \right\} dx_1 dx_2 \\
&= \frac{1}{2\pi\sigma_1\sigma_2} \int_{\mathbb{R}} \exp\left\{ -\frac{(x_2-\mu_2)^2}{2\sigma_2^2} \right\} dx_2 \int_{-\infty}^{t} \exp\left\{ -\frac{(x_1-\mu_1)^2}{2\sigma_1^2} \right\} dx_1 \\
&= \frac{1}{\sqrt{2\pi}\sigma_1} \int_{-\infty}^{t} \exp\left\{ -\frac{(x_1-\mu_1)^2}{2\sigma_1^2} \right\} dx_1
\end{aligned}
$$

则其边缘密度函数为：

$$
f_{X_1}(t) = \frac{dF_{X_1}(t)}{dt} = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{(t-\mu_1)^2}{2\sigma_1^2} \right\}
$$

边缘分布

注意如果只确定了边缘分布，联合分布并不能唯一确定。

联合分布与边缘分布

以下两个联合质量函数具有相同的边缘分布，然而其联合质量函数并不相同：

$Z\backslash Y$	0	1	f_Z
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
f_Y	$\frac{1}{2}$	$\frac{1}{2}$	1

$Z\backslash Y$	0	1	f_Z
0	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{1}{2}$
1	$\frac{5}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
f_Y	$\frac{1}{2}$	$\frac{1}{2}$	1

边缘分布

联合分布于边缘分布

如果随机向量(U,V)，$0 \leq U,V \leq 1$，其分布函数为：

$$F_{U,V}(u,v) = \min\{u,v\}$$

其边缘分布：

$$F_U(u) = F_{U,V}(u,\infty) = u$$
$$F_V(v) = F_{U,V}(\infty,v) = v$$

即其边缘分布为均匀分布。如果另一分布函数为：

$$F_{U,V}(U,V) = u \cdot v$$

其边缘分布也为均匀分布。因而如果只知道边缘分布，不能确定其联合分布。

多元随机变量的期望

- 多元随机变量分量的期望与一元随机变量的期望定义相同
- 我们通常吧随机向量的期望写为向量形式：

$$\mathbb{E}(X) = \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_n) \end{bmatrix}$$

- 对于常数矩阵A，有：$\mathbb{E}(AX) = A\mathbb{E}(X)$
- 在这里期望的线性性仍然成立，比如，如果$\iota = (1, 1, ..., 1)'$为全部由1构成的向量，那么：

$$\mathbb{E}(\iota' X) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i)$$

协方差

如果对于两个一元随机变量Y, Z，如果$\mathbb{E}|Y|^2 < \infty, \mathbb{E}|Z|^2 < \infty$，根据Cauchy-Schwarz 不等式，$\mathbb{E}|YZ| \leq \sqrt{\mathbb{E}|Y|^2 \mathbb{E}|Z|^2} < \infty$，即$YZ$可积，我们可以定义两个随机变量的**协方差（Covariance）**：

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(Y, Z) &= \mathbb{E}[(Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))] \\
&= \mathbb{E}[YZ - \mathbb{E}(Y)Z - Z\mathbb{E}(Y) + \mathbb{E}(Y)\mathbb{E}(Z)] \\
&= \mathbb{E}(YZ) - 2\mathbb{E}(Y)\mathbb{E}(Z) + \mathbb{E}(Y)\mathbb{E}(Z) \\
&= \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z)
\end{aligned}$$

当$Y = Z$时，$\mathbb{C}\mathrm{ov}(Y, Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \mathbb{V}(Y)$。

相关系数

进而可以使用协方差定义相关系数（correlation coefficient）：

$$\rho_{Y,Z} = \frac{\mathbb{C}\mathrm{ov}(Y,Z)}{\sqrt{\mathbb{V}(Y)\mathbb{V}(Z)}}$$

由于：

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(Y,Z) &= \mathbb{E}\left[(Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))\right] \\
&\leq \mathbb{E}\left|(Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))\right| \\
&\leq \sqrt{\mathbb{E}\left|(Y-\mathbb{E}(Y))\right|^2 \mathbb{E}\left|Z-\mathbb{E}(Z)\right|^2} \\
&= \sqrt{\mathbb{V}(Y)\mathbb{V}(Z)}
\end{aligned}$$

可知$-1 \leq \rho_{Y,Z} \leq 1$。

相关系数

- 如果$\rho_{Y,Z}=\pm 1$，那么$P\left(Y=c_1Z+c_2\right)=1, c_1\neq 0$;
- 如果$\rho_{Y,Z}>0$，我们称随机变量Y和Z正相关，反之称为负相关;
- 如果$\rho_{Y,Z}=0$，我们称随机变量Y和Z不相关（uncorrelated）。
- 这里所谓的「相关系数」特指皮尔森相关系数（Pearson correlation coefficient），实际上只度量了随机变量之间的线性相关性。

相关系数

联合分布于边缘分布

如果随机变量$Y = Z^2$，$Z \sim N(0,1)$，那么：

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Z,Y) &= \mathbb{E}ZY - \mathbb{E}Z\mathbb{E}Y \\ &= \mathbb{E}Z^3 \\ &= 0\end{aligned}$$

两者相关系数为0，然而显然两者存在着非线性的函数关系。

协方差

联合密度函数

如果a, b为任意实数，Y和Z为一元随机变量，那么：

$$\begin{aligned}\mathbb{V}(aY+bZ) &= \mathbb{E}(aY+bZ)^2 - [a\mathbb{E}(Y)+b\mathbb{E}(Z)]^2 \\ &= \mathbb{E}\left(a^2Y^2+b^2Z^2+2abYZ\right) \\ &\quad - \left[a^2(\mathbb{E}(Y))^2+b^2(\mathbb{E}(Z))^2+2ab\mathbb{E}(Y)\mathbb{E}(Z)\right] \\ &= a^2\mathbb{V}(Y)+b^2\mathbb{V}(Z)+2ab\mathbb{C}\mathsf{ov}(Y,Z)\end{aligned}$$

如果Y, Z不相关，那么

$$\mathbb{V}(aY+bZ) = a^2\mathbb{V}(Y)+b^2\mathbb{V}(Z)$$

协方差矩阵

如果对于一个随机向量：$X=(X_1,X_2,...,X_n)'$，我们可以定义矩阵：

$$\Sigma=\mathbb{V}(X)=[\mathbb{C}\mathrm{ov}(X_i,X_j)]$$

$$=\begin{bmatrix}\mathbb{V}(X_1) & \mathbb{C}\mathrm{ov}(X_1,X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_1,X_n)\\ \mathbb{C}\mathrm{ov}(X_2,X_1) & \mathbb{V}(X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_2,X_n)\\ \vdots & \vdots & \ddots & \vdots\\ \mathbb{C}\mathrm{ov}(X_n,X_1) & \mathbb{C}\mathrm{ov}(X_n,X_2) & \cdots & \mathbb{V}(X_n)\end{bmatrix}$$

易知协方差矩阵为实对称矩阵。

协方差矩阵的计算

- 根据协方差矩阵的定义，协方差矩阵可以如下计算：

$$\mathbb{V}(X)=\mathbb{E}\left([X-\mathbb{E}(X)][X-\mathbb{E}(X)]'\right)$$

注意X为列向量，从而：

$$X-\mathbb{E}(X)=\begin{bmatrix} X_1-\mathbb{E}(X_1) \\ \vdots \\ X_n-\mathbb{E}(X_n) \end{bmatrix}$$

从而：$[X-\mathbb{E}(X)][X-\mathbb{E}(X)]'=$

$$\begin{bmatrix} (X_1-\mathbb{E}(X_1))^2 & (X_1-\mathbb{E}(X_1))(X_2-\mathbb{E}(X_2)) & \cdots & (X_1-\mathbb{E}(X_1))(X_n-\mathbb{E} \\ (X_2-\mathbb{E}(X_2))(X_1-\mathbb{E}(X_1)) & (X_2-\mathbb{E}(X_2))^2 & \cdots & (X_2-\mathbb{E}(X_2))(X_n-\mathbb{E} \\ \vdots & \vdots & \ddots & \vdots \\ (X_n-\mathbb{E}(X_n))(X_1-\mathbb{E}(X_1)) & (X_n-\mathbb{E}(X_n))(X_2-\mathbb{E}(X_2)) & \cdots & (X_n-\mathbb{E}(X_n))^2 \end{bmatrix}$$

求期望即可。

协方差矩阵的性质

与方差、协方差的性质类似，协方差有如下性质：

- $\mathbb{V}(X) = \mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')$
- $\mathbb{V}(AX + b) = A\mathbb{V}(X)A'$，其中$A$为常数矩阵，$b$为常数向量
- 协方差矩阵为实对称矩阵，且为半正定矩阵
  - 对于任意常数向量a，有：$a'\mathbb{V}(X)a = \mathbb{V}(a'X) \geq 0$
- 如果存在一个向量a和常数b使得$a'X = b$，那么$\mathbb{V}(X)$不满秩。
  - 对于一个实对称半正定矩阵，其特征值一定大于等于0，如果$a'X = b$，那么一定有特征值等于0，从而不满秩。

随机变量的独立性

随机变量的独立性

如果$\{X_i, 1 \leq i \leq n\}$是定义在概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一系列随机变量，如果对于任意的Borel集$\{B_i, 1 \leq i \leq n\}$，有：

$$\mathscr{P}\left(\bigcap_{i=1}^{n}\left(X_i\left(\omega\right) \in B_i\right)\right) = \prod_{i=1}^{n}\mathscr{P}\left(X_i\left(\omega\right) \in B_i\right)$$

那么我们称随机变量$\{X_i, 1 \leq i \leq n\}$相互独立。

随机向量的独立性

对于随机向量，以上定义等价于：

随机向量的独立性

随机向量$(X_1, ..., X_n)$各分量相互独立的充要条件是其联合分布函数等于边缘分布乘积：

$$\begin{aligned} F(x_1, ...x_n) &= P(X_1 \leq x_1, ..., X_n \leq x_n) \\ &= \prod_{i=1}^{n} P(X_i \leq x_i) = \prod_{i=1}^{n} F_{X_i}(x_i) \end{aligned}$$

如果密度（质量）函数存在，那么：

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

随机向量的独立性

独立性

若概率质量函数为:

$Z\backslash Y$	2	3	4	5	6	7	8	F_Z	f_Z
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_Y	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_Z$ $\|\|$
f_Y	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_Y =$	1

可见$f_{Z,Y} \neq f_Z \cdot f_Y$，所以随机变量$(Y, Z)$不独立。

随机向量的独立性

独立性

两个联合分布函数：

$$F^1_{U,V}(u,v) = \min\{u,v\}$$
$$F^2_{U,V}(u,v) = u \cdot v$$

其边缘分布都为均匀分布，即$F_U(u) = u, F_V(v) = v$，然而由于：

$$F^1_{U,V}(u,v) = \min\{u,v\} \qquad \neq F_U(u) \cdot F_C(v)$$
$$F^2_{U,V}(u,v) = u \cdot v \qquad = F_U(u) \cdot F_C(v)$$

因而联合分布服从$F^1_{U,V}$的随机变量不是相互独立的，而服从$F^2_{U,V}$的随机变量是相互独立的。

随机变量函数的独立性

随机变量函数的独立性

$\{X_j, 1 \leq j \leq n\}$为一系列相互独立的随机变量，$1 \leq n_1 \leq n_2 \leq \cdots \leq n_k = n$，那么对于Borel可测函数$f_1, f_2, ...f_k$，那么：

$$\left\{f_1\left(X_1, ...X_{n_1}\right), f_1\left(X_{n_1+1}, ...X_{n_2}\right), \cdots, f_k\left(X_{n_{k-1}+1}, ...X_{n_k}\right)\right\}$$

也为相互独立的随机变量

独立与不相关

独立与不相关

如果概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的随机向量$X = (Y, Z)'$，$Y$和$Z$相互独立且可积，那么：

$$\mathbb{E}(YZ) = \mathbb{E}(Y)\mathbb{E}(Z)$$

因而，如果两个随机变量相互独立，那么其协方差$\mathbb{C}\mathsf{ov}(Y, Z) = \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z) = 0$。然而反之并不成立。

条件期望

- 令(Y, X)为一个二元随机向量，如何使用随机变量X预测随机变量Y?
- 在统计中，我们把这类问题成为回归（regression）。
- 如果我们观察到了随机变量X的值，那么X的何种函数形式可以更好的预测Y呢?
- 比较常见的做法是最小化均方误差（mean squared error）：

$$\min_{h \in \mathscr{L}^2} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

其中

$$\mathscr{L}^2 = \left\{ h | h : \mathbb{R} \to \mathbb{R}, \mathbb{E}\left[(h(X))^2 \right] < \infty \right\}$$

条件期望

- 定义误差项$\epsilon = Y - h_0(X)$，对于随机变量X的任意函数$g(X)$，我们有：

$$\mathbb{E}[\epsilon \cdot g(X)] = 0$$

- 如果令$g(X) = 1$，那么我们有

$$\mathbb{E}[\epsilon \cdot g(X)] = \mathbb{E}[\epsilon] = \mathbb{E}[Y - h_0(X)] = 0$$

因而$\mathbb{E}(Y) = \mathbb{E}(h_0(X))$。

  - 从而$\mathbb{E}[\epsilon \cdot g(X)] = 0$意味着

$$\mathbb{C}\mathrm{ov}(\epsilon, g(X)) = \mathbb{E}[\epsilon \cdot g(X)] - \mathbb{E}(\epsilon)\mathbb{E}[g(X)] = 0$$

即ϵ与X的任意函数都不相关。

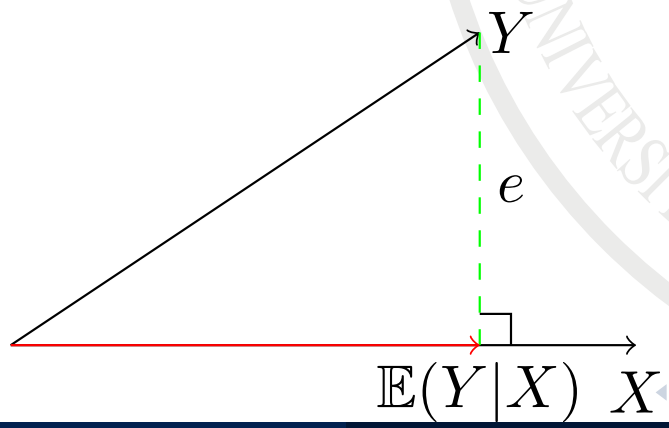
条件期望

- 通过反证法证明，如果存在$g(X)$使得$\mathbb{E}[\epsilon \cdot g(X)] \neq 0$，那么我们令

$$h(X) = h_0(X) + \frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]} g(X)$$

根据这一构造，
有：$\mathbb{E}\left[(Y - h(X))^2\right] < \mathbb{E}\left[(Y - h_0(X))^2\right]$，与$h_0(X)$最小化了均方误差矛盾。

- 我们称$h(X)$为Y在X上的**正交投影（orthogonal projection）**。



条件期望

- 我们知道，

$$\mathbb{E}(Y) = \arg\min_{c \in \mathbb{R}} \left\{ \mathbb{E}(Y-c)^2 \right\}$$

- 仿照上式，我们可以定义随机变量Y给定X的条件期望（conditional expectation）：

$$\mathbb{E}(Y|X) = h_0(X) = \arg\min_{h \in \mathscr{L}^2} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

因而随机变量Y给定X的条件期望$\mathbb{E}(Y|X)$是一个关于X的函数。

条件期望与期望

- 期望可以看做是没有任何其他信息时的最优预测，即只用常数对Y进行预测，是条件期望的特例：

$$\mathbb{E}(Y|c) = c^* = \arg\min_{h \in \mathbb{H}} \left\{ \mathbb{E}\left[(Y-c)^2 \right] \right\}$$

- 这也就意味着：
  - 期望本身是对一个随机变量的最优预测
  - 具体的一个实现与期望之间的差异为误差项
  - 比如：
    - 如果全国所有人的平均体重为60公斤，那么随机从人群中选取一个人，对其体重的最优预测为60公斤
    - 单独每个人的体重与60公斤之间的差距为误差项

条件期望：离散情形

现在，进一步，如果我们能看到一个变量$D \in \{0,1\}$，那么：

$$\mathbb{E}(Y|D) = \arg\min_{h \in \mathbb{H}} \left\{ \mathbb{E}\left[(Y - h(D))^2 \right] \right\} = \begin{cases} \mathbb{E}(Y|D=1) = h(1) & D=1 \\ \mathbb{E}(Y|D=0) = h(0) & D=0 \end{cases}$$

- 比如，现在我们可以观察到性别（$D=1$代表男性），全国所有男性的平均体重为70公斤，所有女性平均体重为50公斤
- 那么：

$$\begin{cases} \mathbb{E}(Y|D=1) = 70 \\ \mathbb{E}(Y|D=0) = 50 \end{cases}$$

即条件期望为分组平均

条件期望：连续情形

或者，如果我们能看到一个变量X为连续型变量，那么

$$\mathbb{E}(Y|X) = \arg\min_{h\in\mathbb{H}} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

为一个未知的函数：

- 比如，如果我们现在可以观察到身高（X）
- 可以假想如果有无数个身高一样的人的平均体重，如：

$$\mathbb{E}(Y|X = 170)$$

即为条件期望。

条件期望的性质

条件期望的性质

对于任意的可测函数$g(X)$，条件期望有如下性质：

1. $\mathbb{E}[g(X)|X] = g(X)$；
2. $\mathbb{E}[(Y - \mathbb{E}(Y|X)) \cdot g(X)] = 0$；
3. $\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}(Y)$，$\mathbb{E}[Y - \mathbb{E}(Y|X)] = 0$；
4. $\mathbb{E}[(g(X) \cdot Y)|X] = g(X) \cdot \mathbb{E}(Y|X)$；
5. $\mathbb{E}(aY_1 + bY_2|X) = a\mathbb{E}(Y_1|X) + b\mathbb{E}(Y_2|X)$。

条件期望的性质

银行到达人数

假设每天到达银行的人数服从泊松分布$N \sim P(\lambda)$，而每个到达银行的人，办理外汇业务的概率为p。那么给定到达人数N，办理外汇业务的人数M服从二项分布，
即$M|N \sim Bi(N,p), N \sim P(\lambda)$。那么每天来银行办理外汇业务的人数的期望：

$$\mathbb{E}(M) = \mathbb{E}[\mathbb{E}(M|N)] = \mathbb{E}(Np) = p\mathbb{E}(N) = p\lambda$$

均值独立

- 注意到如果我们没有任何信息，因而只能用常数c去预测Y，那么以上最小化问题：

$$\mathbb{E}(Y|c) = c^* = \arg\min_{h \in \mathscr{L}^2} \left\{ \mathbb{E}\left[(Y-c)^2 \right] \right\}$$

对以上最优化问题求解，即：

$$\frac{\partial \mathbb{E}\left[(Y-c)^2 \right]}{\partial c} = \mathbb{E}\left[\frac{\partial (Y-c)^2}{\partial c} \right] = 0$$

从而得到：$\mathbb{E}(Y|c) = \mathbb{E}(Y)$

- 即如果我们没有任何其他随机变量的信息，只能用常数预测Y，那么我们将得到Y的期望。
- 如果有其他随机变量X，但是$\mathbb{E}(Y|X) = \mathbb{E}(Y)$，那么$X$对$Y$的均值没有预测能力，因而我们称$Y$对$X$是均值独立（mean independence）的。

均值独立

- 如果随机变量Y对X是均值独立的，那么：

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(X,Y\right) &= \mathbb{E}\left(XY\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)\\ &= \mathbb{E}\left(\mathbb{E}\left(XY|X\right)\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)\\ &= \mathbb{E}\left(X\mathbb{E}\left(Y|X\right)\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)\\ &= \mathbb{E}\left(X\mathbb{E}\left(Y\right)\right)-\mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)=0\end{aligned}$$

  因而随机变量Y和X必然是不相关的。反之则不成立，不相关并不一定意味着均值独立。
- 实际上，可以证明，$\mathbb{C}\mathrm{ov}\left(g\left(X\right),Y\right)=0$，即$Y$对$X$是均值独立的意味着$Y$与$X$的任意函数都不相关。
- 反过来不一定正确，即$\mathbb{C}\mathrm{ov}\left(g\left(Y\right),X\right)=0$不一定成立（why?）

条件方差

- 相应的，我们还可以定义条件方差

$$\mathbb{V}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2|X\right]$$

- 根据条件期望的性质：

$$\begin{aligned}
\mathbb{V}(Y|X) &= \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2|X\right] \\
&= \mathbb{E}\left\{\left[Y^2 + [\mathbb{E}(Y|X)]^2 - 2Y\mathbb{E}(Y|X)\right]|X\right\} \\
&= \mathbb{E}\left(Y^2|X\right) + \mathbb{E}\left\{[\mathbb{E}(Y|X)]^2|X\right\} - 2\mathbb{E}\left[Y\mathbb{E}(Y|X)|X\right] \\
&= \mathbb{E}\left(Y^2|X\right) + [\mathbb{E}(Y|X)]^2 - 2\mathbb{E}(Y|X)\,\mathbb{E}[Y|X] \\
&= \mathbb{E}\left(Y^2|X\right) - [\mathbb{E}(Y|X)]^2
\end{aligned}$$

其中第4个等号由于$\mathbb{E}(Y|X)$也是X的函数

条件方差与方差

根据条件期望的性质，可以证明：

$$\mathbb{V}(Y) = \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}[\mathbb{E}(Y|X)]$$

即条件方差的期望与条件期望的方差之和为方差。

条件方差

银行到达人数的方差

在银行到达人数的例子中，可以计算每天办理外汇业务的人数的方差：

$$\mathbb{V}(M) = \mathbb{V}(\mathbb{E}(M|N)) + \mathbb{E}(\mathbb{V}(M|N))$$

其中$\mathbb{E}(M|N) = Np$，因而

$$\mathbb{V}[\mathbb{E}(M|N)] = \mathbb{V}(Np) = p^2\mathbb{V}(N) = p^2\lambda$$

而$\mathbb{V}(M|N) = Np(1-p)$，从而

$$\mathbb{E}(\mathbb{V}(M|N)) = \mathbb{E}(Np(1-p)) = \lambda p(1-p)$$

从而

$$\mathbb{V}(M) = p^2\lambda + \lambda p - \lambda p^2 = \lambda p$$

条件分布

- 如果对于随机向量(X,Y)，我们取$1_A(x)=1\ if\ x\in A\ else\ =0$，这是一个随机变量$X$的函数，因而

$$\mathbb{E}(Y\cdot 1_A(X))=\mathbb{E}[\mathbb{E}(Y|X)\cdot 1_A(X)]=\mathbb{E}[h_0(X)\cdot 1_A(X)]$$

- 若X,Y是离散型随机变量，那么我们令$A=\{X=x_i\}$有：

$$\mathbb{E}(Y\cdot 1\{X=x_i\})=h_0(x_i)\cdot P(X=x_i)$$

从而：

$$\begin{aligned}\mathbb{E}(Y|X=x_i)=h_0(x_i)&=\frac{\mathbb{E}(Y\cdot 1\{X=x_i\})}{P(X=x_i)}\\&=\frac{\sum_{k=0}^{\infty}[y_k\cdot P(Y=y_k,X=x_i)]}{P(X=x_i)}\\&=\sum_{k=0}^{\infty}y_k\cdot\frac{P(Y=y_k,X=x_i)}{P(X=x_i)}\end{aligned}$$

条件分布

- 对于连续型随机向量(X,Y)，可以证明：

$$\mathbb{E}\left(Y|X=x\right)=h_0\left(x\right)=\frac{\int_{\mathbb{R}}yf\left(x,y\right)dy}{f_X\left(x\right)}=\int_{\mathbb{R}}y\frac{f\left(x,y\right)}{f_X\left(x\right)}dy$$

  - 由于$\mathbb{E}\left[Y-\mathbb{E}\left(Y|X\right)\right]=0$，从而：

$$\int_{\mathbb{R}}\left[y-h_0\left(x\right)\right]f\left(x,y\right)dy=0$$

  - 固定x，那么以上条件意味着：

$$\int_{\mathbb{R}}yf\left(x,y\right)dy=h_0\left(x\right)\int_{\mathbb{R}}f\left(x,y\right)dy=h_0\left(x\right)f_X\left(x\right)$$

条件分布

条件密度函数

对于离散型随机变量，定义

$$f_{Y|X}(y|x) = \frac{P(Y=y, X=x)}{P(X=x)} = \frac{P(Y=y, X=x)}{\sum_y P(Y=y, X=x)}$$

对于连续型随机变量，定义

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dy}$$

我们把$f_{Y|X}(y|x)$定义为**条件密度函数（conditional denstiy function）**。

条件密度函数

- 条件期望可以写为条件密度函数的积分

$$\mathbb{E}\left(Y|X\right)=\int_{\mathbb{R}}y\cdot f_{Y|X}\left(y|x\right)dy$$

- 注意：对于离散型随机变量：

$$\sum_{y}f_{Y|X}\left(y|x\right)=\sum_{y}\frac{P\left(Y=y,X=x\right)}{\sum_{y}P\left(Y=y,X=x\right)}=1$$

而对于连续型随机变量：

$$\int_{\mathbb{R}}f_{Y|X}\left(y|x\right)dy=\int_{\mathbb{R}}\frac{f\left(x,y\right)}{\int_{\mathbb{R}}f\left(x,y\right)dy}dy=1$$

因而条件密度函数也是密度函数。

独立与均值独立

- 如果随机变量X和Y是独立的，那么：

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

即两个随机变量独立的充要条件是$f_{Y|X} = f_Y$。

- 在独立的条件下：

$$\mathbb{E}(Y|X) = \int_{\mathbb{R}} y \cdot f_{Y|X}(y|x)\, dy = \int_{\mathbb{R}} y \cdot f_Y(y)\, dy = \mathbb{E}(Y)$$

因而如果随机变量X和Y是独立的，那么其一定是均值独立的，反之则不成立。

- 独立、均值独立、不相关三者的强弱关系如下：

$$X \perp\!\!\!\perp Y \underset{\not\Leftarrow}{\Rightarrow} \mathbb{E}(Y|X) = \mathbb{E}(Y) \underset{\not\Leftarrow}{\Rightarrow} X \perp Y$$

其中$X \perp\!\!\!\perp Y$代表X与Y独立，$X \perp Y$代表X与Y不相关。

条件密度函数

四面骰子

四面骰子的例子中，其条件密度可以如下计算：

$Z\backslash Y$	2	3	4	5	6	7	8
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0
4	0	0	0	0	0	0	$\frac{1}{16}$
$f_{Y\|Z}(y\|Z=1)$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	0	0	
$f_{Y\|Z}(y\|Z=2)$	0	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	0	

条件密度函数

二元正态分布

对于联合正态

$$(X,Y)' \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

密度函数：

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\}$$

其中$-1 < \rho < 1$为X, Y的相关系数。

条件密度函数

四面骰子

其边际密度函数为:

$$
\begin{aligned}
f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy \\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \\
&\cdot \int_{\mathbb{R}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(1-\rho^2)(x-\mu_X)^2}{\sigma_X^2}+\left(\frac{y-\mu_Y}{\sigma_Y}-\frac{\rho(x-\mu_X)}{\sigma_X}\right)^2\right]\right\} dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}
\end{aligned}
$$

或者$X \sim N(\mu_X, \sigma_X^2)$

条件密度函数

四面骰子

其条件密度函数为:

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2}\left(\frac{y-\left[\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\}
\end{aligned}
$$

或者$Y|X \sim N\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right)$，也服从正态分布，进而:

- 条件期望$\mathbb{E}(Y|X=x) = \mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$
- 条件方差$\mathbb{V}(Y|X) = \sigma_Y^2(1-\rho^2)$。

贝叶斯公式

- 使用条件密度函数的定义，我们还可以得到随机变量的贝叶斯公式。
- 由于：$f(x,y)=f_X(x)\cdot f_{Y|X}(y|x)=f_Y(y)\cdot f_{X|Y}(x|y)$从而条件密度：

$$f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)}=\frac{f_{X|Y}(x|y)\cdot f_Y(y)}{\int_{\mathbb{R}} f(x,y)\,dy}$$
$$=\frac{f_{X|Y}(x|y)\cdot f_Y(y)}{\int_{\mathbb{R}} f_{X|Y}(x|y)\cdot f_Y(y)\,dy}$$

以上方程即随机变量的贝叶斯公式，在贝叶斯统计中有大量的应用。

条件分布与统计模型

- 统计模型通常研究几个随机变量的联合分布，如$F(X,Y)$
- 经常我们不需要对联合分布进行建模，而是通过条件分布对联合分布进行建模。由于：

$$f(x,y) = f_X(x) \cdot f_{Y|X}(y|x)$$

因而我们可以通过对X的边缘分布以及$Y|X$的条件分布对X,Y的联合分布进行建模。

- 比如，分层模型（hierarchical model）就通过分层次的假设条件分布对数据的分布进行建模。
- 有时X的边缘分布不是关注的核心问题，甚至可能只对$Y|X$的边缘分布进行建模。

高斯混合模型

高斯混合模型

如果我们关注某一项疾病指标X，该指标对于患者和健康人群具有不同的分布。记$D=1$为患者，$D=0$为健康人群，记患者该项指标为X_1，健康人群该项指标为X_0，假设：

$$\begin{cases} X_1 \sim N\left(\mu_1, \sigma_1^2\right) \\ X_0 \sim N\left(\mu_0, \sigma_0^2\right) \end{cases}$$

即分别假设了患者和健康人群该项指标的分布，那么观察到的指标：$X = DX_1 + (1-D)X_0$。该模型可以写为：

$$\begin{cases} X|D=1 \sim N\left(\mu_1, \sigma_1^2\right) \\ X|D=0 \sim N\left(\mu_0, \sigma_0^2\right) \\ D \sim Ber\left(p\right) \end{cases}$$

高斯混合模型

高斯混合模型

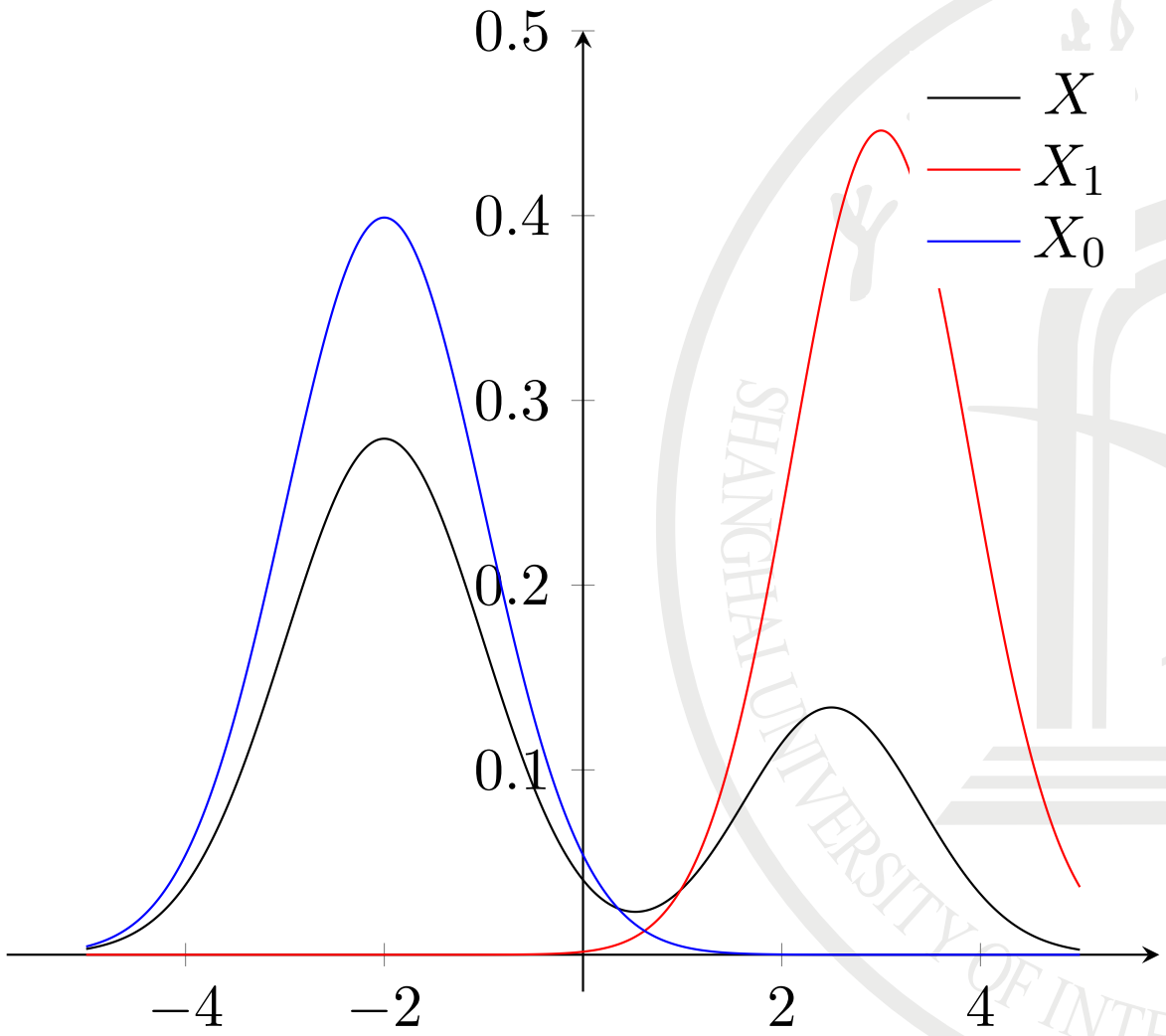
为了得到X的密度函数，注意到：

$$\begin{aligned}
F_X(x) &= P(X \le x) = \mathbb{E}[1(X \le x)] = \mathbb{E}\{\mathbb{E}[1(X \le x)|D]\} \\
&= \mathbb{E}[1(X \le x)|D=1] \cdot P(D=1) \\
&+ \mathbb{E}[1(X \le x)|D=0] \cdot P(D=0) \\
&= \Phi\left(\frac{x-\mu_1}{\sigma_1}\right) \cdot p + \Phi\left(\frac{x-\mu_0}{\sigma_0}\right) \cdot (1-p)
\end{aligned}$$

从而：

$$\begin{aligned}
f_X(x) &= p\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-p)\frac{1}{\sigma_0}\phi\left(\frac{x-\mu_0}{\sigma_0}\right) \\
&= pf_{X_1}(x) + (1-p)f_{X_0}(x)
\end{aligned}$$

高斯混合模型



高斯混合模型

高斯混合模型

我们可以使用条件期望计算X的期望:

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}[\mathbb{E}(X|D)] = \mathbb{E}\{\mathbb{E}[DX_1 + (1-D)X_0|D]\} \\ &= \mathbb{E}\{D\mathbb{E}(X_1|D) + (1-D)\mathbb{E}(X_0|D)\} \\ &= \mathbb{E}\{D\mu_1 + (1-D)\mu_0\} \\ &= \mu_1\mathbb{E}(D) + \mu_0\mathbb{E}(1-D) \\ &= p\mu_1 + (1-p)\mu_0\end{aligned}$$

高斯混合模型

高斯混合模型

此外，如果我们观察到了X，也可以使用贝叶斯公式计算其患病的概率：

$$
\begin{aligned}
f_{D|X}\left(d=1|x\right) &= \frac{f_{X|D}\left(x|d=1\right)f_D\left(d=1\right)}{\int_{\mathbb{R}} f_{X|D}\left(x|\tilde{d}\right)f_D\left(d\right)d\tilde{d}} \\
&= \frac{\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right)p}{\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right)p+\frac{1}{\sigma_0}\phi\left(\frac{x-\mu_0}{\sigma_0}\right)(1-p)}
\end{aligned}
$$

迭代期望公式

- 条件期望可以很方便的扩充到多个X的情形，比如$\mathbb{E}(Y|X_1,X_2)$可以定义为：

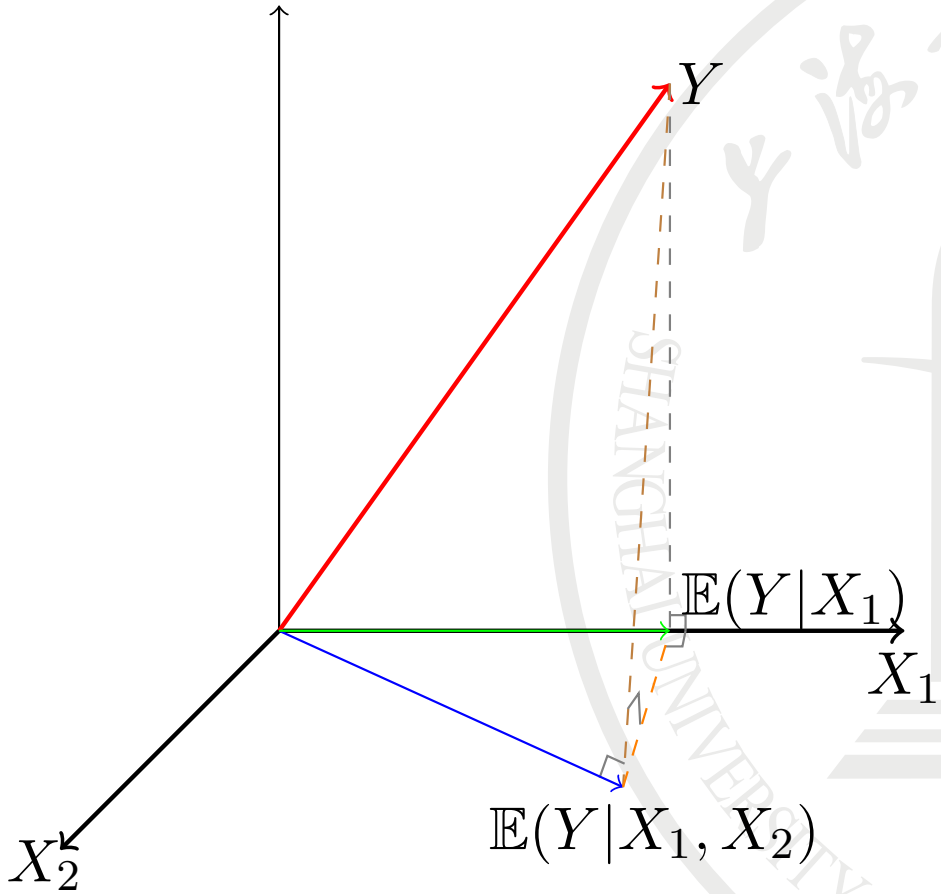
$$\mathbb{E}(Y|X_1,X_2)=h_0(X_1,X_2)=\arg\min_{h\in\mathscr{L}^2}\left\{\mathbb{E}\left[(Y-h(X_1,X_2))^2\right]\right\}$$

- 条件期望有如下性质（**迭代期望公式，law of iterated expectation**）：

$$\mathbb{E}\left[\mathbb{E}(Y|X_1,X_2)|X_1\right]=\mathbb{E}(Y|X_1)$$

即如果我们对随机变量Y，先在大的空间上投影，再在这个大的空间上的一个小的子空间上进行投影，与直接在这个小的空间上进行投影是相等的。

迭代期望公式



条件期望

条件期望与$\sigma-$代数

在四面骰子的例子中，随机变量Z可能取值为：$\{1,2,3,4\}$，因而：

$$\sigma\langle Z\rangle=\sigma\left\langle Z^{-1}(A):A\in\mathcal{B}\right\rangle=$$
$$\sigma<\{(1,1),(2,1),(3,1),(4,1),(1,2),(1,3),(1,4)\},$$
$$\{(2,2),(2,3),(2,4),(3,2),(4,2)\},$$
$$\{(3,3),(3,4),(4,3)\},\{(4,4)\}>$$

例如，如果我们只知道$Z=3$，我们知道实际发生的情况应该是$\{(3,3),(3,4),(4,3)\}$中的某一种。因而如果给定$Z=3$，我们把之前的16种情况降低到了3种情况。如果我们对Y，即两个骰子的和感兴趣，如果我们没有任何信息，那么我们对Y的最优预测应该是$\mathbb{E}(Y)=5$。而如果我们观察到了$Z=3$，那么此时最优预测应该为$\mathbb{E}(Y|Z=3)=\frac{6+7+7}{3}=\frac{20}{3}$。

条件期望

条件期望与$\sigma-$代数

- 在上例中，Z总共有4种可能的取值，在每种Z的可能取值的情况下，都可以把16种情况降低为更少的情况，因而增大了信息量。
- 而如果我们使用随机变量Y，Y共有7种可能的取值，给定Y也会增大我们的信息量。
- 而如果给定(Z,Y)两个随机变量，可以更加细分为10种情况，我们可以得到

$$\sigma\langle Z\rangle \subset \sigma\langle Z,Y\rangle, \sigma\langle Y\rangle \subset \sigma\langle Z,Y\rangle$$

即两个随机变量提供了比单独一个随机变量更多的信息。

  - 例如，如果我们不仅仅观察到$Z=3$，还观察到$Y=7$，那么我们此时知道，实际发生的情况应该是$\{(3,4),(4,3)\}$两种情况下的一种，比只观察到Z时更加准确。

条件期望定义

因而我们通常把条件期望的概念推广到$\sigma-$代数上。对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，我们可以对$\mathscr{F}$的一个子$\sigma-$代数$\mathscr{G} \subset \mathscr{F}$定义条件期望如下：

条件期望

对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，$\mathcal{I} \subset \mathscr{F}$为一个$\sigma-$代数，对于随机变量$Y$满足：$\mathbb{E}(|Y|) < \infty$，如果对于任意的$A \in \mathcal{I}$，随机变量$H$满足：

$$\mathbb{E}(Y \cdot 1_A) = \mathbb{E}(H \cdot 1_A)$$

那么我们称H为给定$\mathcal{I}$随机变量Y的条件期望，记为$\mathbb{E}(Y|\mathcal{I})$。令$B \in \mathscr{F}$，定义$\mathscr{P}(B|\mathcal{I}) = \mathbb{E}(1_B|\mathcal{I})$为条件概率。

迭代期望公式

- 注意以上定义的$\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma\langle X\rangle)$。
- 特别的，令$\mathcal{I} = \{\emptyset, \Omega\}$，$\mathbb{E}(Y|\{\emptyset, \Omega\}) = \mathbb{E}(Y)$，即信息量最小的条件期望即为期望本身。
- 而以上的迭代期望公式也可以相应推广，即如果$\mathcal{I}_1 \subset \mathcal{I}_2 \subset \mathscr{F}$，那么：

$$\mathbb{E}(Y|\mathcal{I}_1) = \mathbb{E}\{[\mathbb{E}(Y|\mathcal{I}_2)]|\mathcal{I}_1\}$$

即先在大的信息集上做投影，再将其投影到小的信息集上，等价于直接投影在小的信息集上。

指数分布族

- 在上一节中，我们学习了单参数指数分布族，即如果一个密度函数可以写为：

$$f(x|\theta) = h(x) \cdot \exp\{\eta(\theta) \cdot T(x) - B(\theta)\}$$

的形式，那么我们称符合该密度函数的所有分布为指数分布族。

- 在这里我们对单参数指数分布族做进一步扩展，即多参数指数分布族：

指数分布族

多参数指数分布族

对于一个参数族$\{P_\theta, \theta \in \Theta\}$，如果其概率密度（质量）函数可以写成如下形式：

$$f(x|\theta) = h(x) \cdot \exp\left\{\sum_{i=1}^{k} [\eta_i(\theta) \cdot T_i(x)] - B(\theta)\right\}$$

那么我们称$\{P_\theta, \theta \in \Theta\}$为**指数分布族**（Exponential family）。

注意其中θ可以为向量，$\theta = (\theta_1, ..., \theta_k)'$，即可以不止依赖于一个参数。

指数分布族

多参数指数分布族

正态分布的密度函数：

$$f\left(x|\mu,\sigma\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

其密度函数可以写为：

$$\begin{aligned}f\left(x|\mu,\sigma\right)&=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}=\frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}-\ln\left(\sigma\right)\right\}\\&=\frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{x^2-2\mu x+\mu^2}{2\sigma^2}-\ln\left(\sigma\right)\right\}\\&=\frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\ln\left(\sigma\right)\right\}\end{aligned}$$

因而也是指数分布族

指数分布族

与单参数指数分布族类似，我们也会将多参数指数分布族改写为规范形式：

多参数指数分布族的规范形式

对于指数分布族

$$f(x|\theta) = h(x) \cdot \exp\left\{\sum_{i=1}^{k} [\eta_i(\theta) \cdot T_i(x)] - B(\theta)\right\}$$

我们令k维向量$\lambda = \eta(\theta)$，那么指数分布族可以写为：

$$f(x|\theta) = h(x) \cdot \exp\left\{\sum_{i=1}^{k} [\lambda_i \cdot T_i(x)] - C(\lambda)\right\}$$

我们称以上形式为规范形式。

指数分布族

多参数指数分布族

在正态分布的例子中：

$$f\left(x|\mu,\sigma\right)=\frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\ln\left(\sigma\right)\right\}$$

令$\lambda_1=-\frac{1}{2\sigma^2},\lambda_2=\frac{\mu}{\sigma^2}$，则$\mu=-\frac{\lambda_2}{2\lambda_1},\sigma^2=-\frac{1}{2\lambda_1}$，
而$C\left(\lambda\right)=-\frac{\lambda_2^2}{4\lambda_1}-\frac{\ln(-2\lambda_1)}{2}$，如此我们写出了正态分布密度函数的规范形式。

作业

- 3.2，3.3，3.5，3.8，3.11