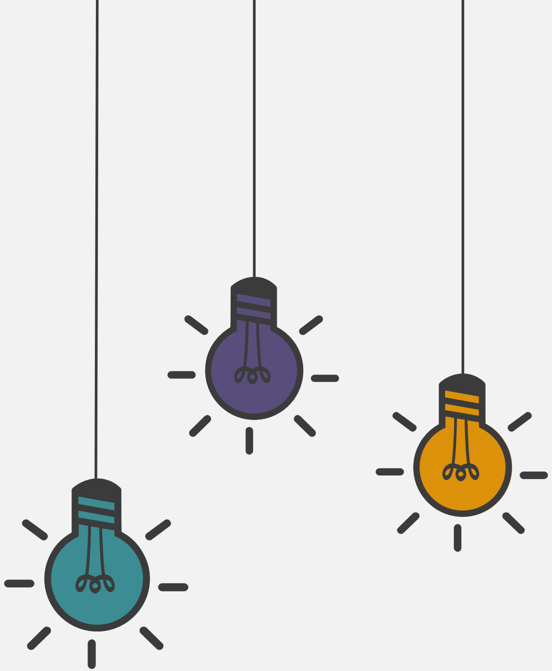


独立性 VS 不相关

Independence versus Zero Correlation

公选课《可计量的社会》

何濯羽

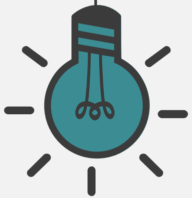
2024 年 9 月 30 日

目 录

CONTENTS

01 联合分布

Joint Distribution



联合分布的定义

之前，我们只考虑单个随机变量的概率分布和矩。本节课中，我们将研究两个（离散型）随机变量共同的概率分布以及它们之间的特殊关系。

两个变量共同的概率分布被称为“**联合分布**”（joint distribution）。对于两个离散型随机变量，它们所有可能的取值以及相应的概率可以由一张“联合分布列”展示出来。

			Y			
		y_1	y_2	$\cdots$	y_m	总和
X	x_1	p_{11}	p_{12}	$\cdots$	p_{1m}	$p_{1\cdot}$
	x_2	p_{21}	p_{22}	$\cdots$	p_{2m}	$p_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	x_n	p_{n1}	p_{n2}	$\cdots$	p_{nm}	$p_{n\cdot}$
	总和	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot m}$	

联合分布的CDF与PMF

$F_{X,Y}(x,y) = \Pr(X \le x, Y \le y)$ 被称为“联合累积分布函数” (joint CDF) 。

$f_{X,Y}(x,y) = \Pr(X = x, Y = y)$ 被称为“联合概率质量函数” (joint PMF) 。

例：

X \ Y	0	2	4	6	总和
0	0.1	0.05	0.08	0.02	0.25
1	0.15	0.1	0.05	0.05	0.35
2	0.13	0.1	0	0.07	0.3
3	0.02	0.05	0.02	0.01	0.1
总和	0.4	0.3	0.15	0.15	**1**

X 的边际分布

Y 的边际分布

符合概率公理的要求

$$
\begin{aligned}
F_{X,Y}(2,2) &= \Pr(X \le 2, Y \le 2) \\
&= \Pr(X=0,Y=0) + \Pr(X=0,Y=2) \\
&\quad + \Pr(X=1,Y=0) + \Pr(X=1,Y=2) \\
&\quad + \Pr(X=2,Y=0) + \Pr(X=2,Y=2) \\
&= 0.1 + 0.05 + 0.15 + 0.1 + 0.13 + 0.1 \\
&= 0.65
\end{aligned}
$$

02 统计独立性

Statistical Independence

定义统计独立性

如果事件 A 和 事件 B 满足

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

那么我们称 A 和 B 为**独立事件**（independent events）。

如果离散型随机变量 X 和 Y 满足

$$\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y), \text{即 } f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y), \qquad \forall x, y \in \mathbb{R}$$

或

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y), \qquad \forall x, y \in \mathbb{R}$$

那么我们称 X 和 Y 相互**统计独立**（statistically independent），或相互**随机独立**（stochastically independent），简称“相互独立”。

例：独立性的应用

已知随机变量 X 和 Y 相互独立，你能根据它们的边际分布填满整张联合分布列吗?

	Y 0	1	2	总和
X 0	0.08	0.12	0.2	0.4
1	0.02	0.03	0.05	0.1
2	0.1	0.15	0.25	0.5
总和	0.2	0.3	0.5	

原理： 独立 $\Longrightarrow \Pr(X=x, Y=y) = \Pr(X=x) \cdot \Pr(Y=y)$

强大的独立性

如果事件 A 和 B **相互独立**，那么 $\Pr(A|B) = \Pr(A)$ 且 $\Pr(B|A) = \Pr(B)$。

如果随机变量 X 和 Y **相互独立**，那么 $\Pr(X = x|Y = y) = \Pr(X = x)$ 且 $\Pr(Y = y|X = x) = \Pr(Y = y)$，$\forall x, y$。

如果随机变量 X 和 Y **相互独立**且 $g(\cdot)$ 和 $h(\cdot)$ 是两个函数，那么 $g(X)$ 和 $h(Y)$ 也相互独立。

如果随机变量 X 和 Y **相互独立**且存在两个函数 $g: \mathbb{R} \to \mathbb{R}$ 和 $h: \mathbb{R} \to \mathbb{R}$，那么

$$E(XY) = E(X) \cdot E(Y)$$

且

$$E[g(X) \cdot h(Y)] = E[g(X)] \cdot E[h(Y)]$$

注意：一般情况下，$E(XY) \neq E(X) \cdot E(Y)$。

独立 VS 互斥

事件 A 和 B **相互独立** $\Leftrightarrow \Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$

事件 A 和 B **互斥** $\Leftrightarrow A \cap B = \emptyset$

从数学定义上看，“独立”和“互斥”是两个完全不同的概念。

如果 $\Pr(A) > 0$ 且 $\Pr(B) > 0$，那么

- 事件 A 和 B 相互独立 $\Rightarrow$事件 A 和 B 不互斥
- 事件 A 和 B 互斥 $\Rightarrow$ 事件 A 和 B 不相互独立

证明：

- 当 A 和 B 相互独立时，则必有 $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$。【反证法】假设 A 和 B 互斥，则 $A \cap B = \emptyset$，这意味着 $\Pr(A \cap B) = \Pr(\emptyset) = 0$。这与 $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B) > 0$ 相悖，因此，假设错误，即 A 和 B 不互斥。
- 当 A 和 B 互斥时，则必有 $A \cap B = \emptyset$。因为 $\Pr(A \cap B) = \Pr(\emptyset) = 0$ 和 $\Pr(A) \cdot \Pr(B) > 0$，所以 $\Pr(A \cap B) \neq \Pr(A) \cdot \Pr(B)$，即 A 和 B 不相互独立。

03 线性不相关

Uncorrelatedness

协方差的定义与作用

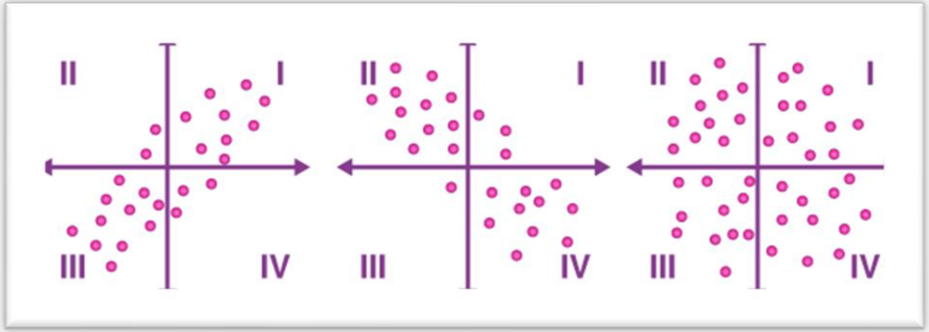
随机变量 X 与 Y 的**协方差**（covariance）是

$$\sigma_{X,Y} = Cov(X,Y) = E\{[X - E(X)] \cdot [Y - E(Y)]\} = E(XY) - E(X) \cdot E(Y)$$

我们可以把 $Cov(X,Y)$ 理解为 X 与 Y 所服从的**联合分布的二阶中心矩。**

协方差可以衡量两个随机变量的线性相关程度（linear relatedness）：

- $Cov(X,Y) > 0$ 意味着 X 与 Y 线性**正相关**。
- $Cov(X,Y) < 0$ 意味着 X 与 Y 线性**负相关**。
- $Cov(X,Y) = 0$ 意味着 X 与 Y 线性**不相关**。



协方差的缺陷

协方差的范围是 $(-\infty,+\infty)$。可是，协方差具有与方差相同的问题：因为随机变量的单位变化势必导致协方差的大小变化，所以协方差的数值大小并不能有效反映线性相关程度的大小。

为了化解单位变量带来的影响，我们创造了“标准差”以对应方差。相似地，我们也可以创造新的统计工具以对应协方差——“相关系数”！

随机变量 X 与 Y 的**相关系数**（correlation coefficient）是

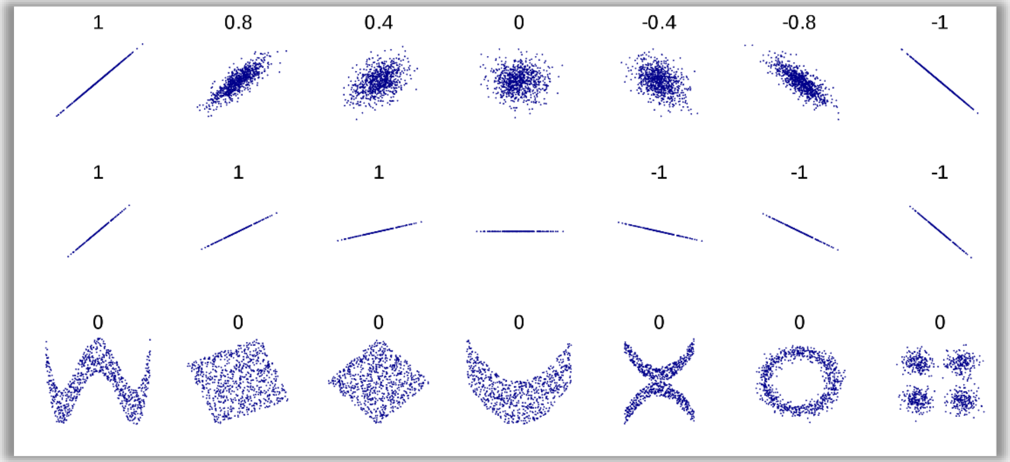
$$\rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

相关系数的含义

该结论的证明需要使用柯西-施瓦茨不等式，在此不做展示。

相关系数 $Corr(X,Y)$ 的范围是 $[-1,1]$，其含义为

- $Corr(X,Y)=1$ 意味着 X 与 Y **完美**线性**正相关**。
- $Corr(X,Y)\in(0,1)$ 意味着 X 与 Y 线性**正相关**。
- $Corr(X,Y)=0$ 意味着 X 与 Y 线性**不相关**。
- $Corr(X,Y)\in(-1,0)$ 意味着 X 与 Y 线性**负相关**。
- $Cov(X,Y)=-1$ 意味着 X 与 Y **完美**线性**负相关**。



线性不相关

如果 $Cov(X,Y)=0$ 或 $Corr(X,Y)=0$，那么我们称随机变量 X 与 Y **线性独立**（linear independent）或**线性不相关**（uncorrelated）。

当 Y 与 X 相互独立时， Y **必定**与 X 线性不相关。

证： $Cov(X,Y)=E\{[X-E(X)]\cdot[Y-E(Y)]\}$

$$
\begin{aligned}
&= E[XY-X\cdot E(Y)-Y\cdot E(X)+E(X)\cdot E(Y)] \\
&= E(XY)-E(X)\cdot E(Y)-E(Y)\cdot E(X)+E(X)\cdot E(Y) \\
&= E(XY)-E(X)\cdot E(Y) \\
&= E(X)\cdot E(Y)-E(X)\cdot E(Y) \quad \longleftarrow \text{独立性} \\
&= 0
\end{aligned}
$$

线性不相关

当 Y 与 X 线性不相关时，Y **不一定** X 与相互独立。

例：

		$X=-1$	$X=0$	$X=1$	总和
Y	0	0.1	0.4	0.1	0.6
	1	0.2	0	0.2	0.4
	总和	0.3	0.4	0.3	

计算可得

$$E(X)=0,\qquad E(Y)=0.4$$

$$Cov(X,Y)=E[(X-0)(Y-0.4)]=0$$

但是，

$$\Pr(\mathrm{X}=-1,\mathrm{Y}=0)=0.1\neq 0.3\times 0.6=\Pr(X=-1)\cdot\Pr(Y=0)$$

因此，Y 与 X 线性不相关，但 Y 并不与 X 相互独立。

总结

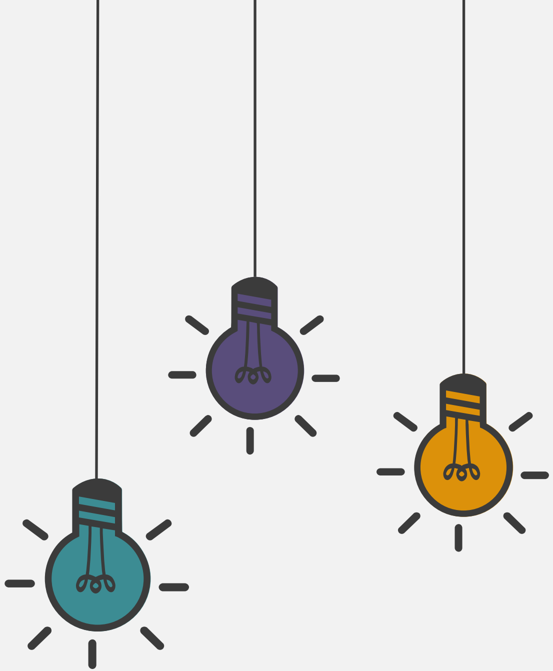
统计独立（简称“独立”）是一个非常强劲的变量关系，**强于线性不相关**。许多统计学的定理都建立在“独立”假设上。可是在实际研究中，独立是难以严格证明的。因此，许多研究者退而求其次，只证明线性不相关（或“均值独立”），用以说明独立“可能”成立。

X 与 Y 相互统计独立

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

X 与 Y 线性不相关

$$Cov(X,Y) = 0$$



感谢聆听！

Thanks for your listening!

公选课《可计量的社会》

何濯羽　　2024 年 9 月 30 日