高级计量经济学

理论经济学博士课程

Lecture 1-1: Review of Probability

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

Office 粤海校区汇文楼1510
Email huangjp@szu.edu.cn
Website https://huangjp.com

概率论

样本空间与事件

Sample space and events

- 具有随机性的事情称作试验（experiment/trial）：抛硬币、掷骰子、明天的最高气温等。
- 试验的所有可能结果（outcome）的集合称为**样本空间（sample space）**：
  - “抛一次硬币得到的面” → $\{H, T\}$
  - “同时掷两个骰子得到的数字” → $\{11, 12, 21, 13, 31, 22, \ldots, 66\}$
- 样本空间的子集称为**事件（event）**：
  - “掷两个骰子的到的数字之和不大于4” → $\{11, 12, 21, 13, 31, 22\}$
  - A 和 B 是样本空间 S 上的事件，则 $A \cup B, A \cap B, A^c, A - B$ 也都是 S 上的事件
  - 一个事件发生了，指的是该事件的集合中某一个结果实现了
  - 样本空间本身也是事件，其含义为进行了试验（或事情发生了）

概率函数

Probability function

令样本空间 S 上的事件的集合为 $\mathscr{A}$，定义在 $\mathscr{A}$ 上的实数值函数 $\Pr : \mathscr{A} \to \mathbb{R}$ 如果满足下列公理，则称为**概率函数（probability function）**

1. $\Pr(A) \geq 0$ for $A \in \mathscr{A}$.
2. $\Pr(S) = 1$.
3. 如果 $A_1, A_2, \ldots$ 两辆不相交，则 $\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$.

若集合 A 与 B 满足 $A \cap B = \varnothing$，则称它们不相交（disjoint）

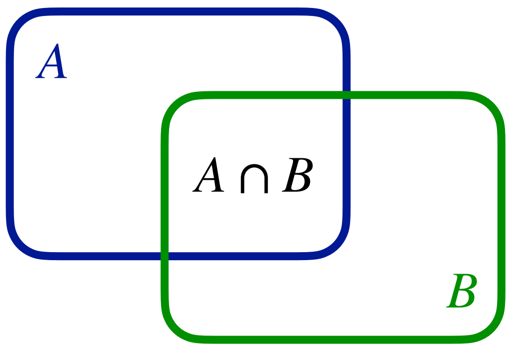
- 事件的概率就是衡量它发生可能性大小的测度。概率一定为非负，试验产生任意结果的概率为 1，不相交的事件发生其一的概率等于各自发生概率之和。
- 以“抛一个公正的骰子”为例：
  - $S = \{1,2,3,4,5,6\}$，$\Pr(\{i\}) = 1/6$
  - 根据概率的公理，$\Pr(\{1,2\}) = \Pr(\{1\}) + \Pr(\{2\}) = 1/3$

概率函数的性质

Properties of the probability function

针对事件 A 和 B，下面的性质成立

1. $\Pr(A^c) = 1 - \Pr(A)$
2. $\Pr(\emptyset) = 0$
3. $\Pr(A) \leq 1$
4. $A \subseteq B \ \Rightarrow \ \Pr(A) \leq \Pr(B)$
5. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
6. Boole's Inequality: $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$
7. Benferroni's Inequality: $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$

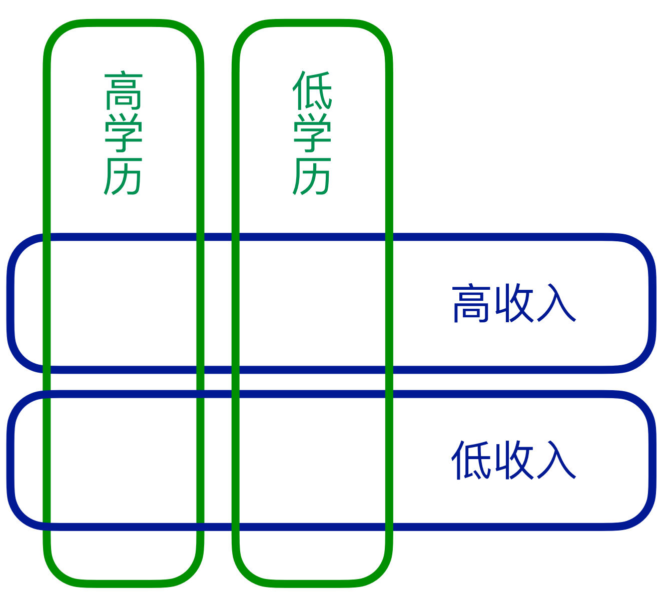


联合事件
Joint Events

- 若两个事件同时发生，我们称之为联合事件
  - 假设你毕业后结识的第一个朋友叫小明，他的收入和学历的样本空间是 $S = \{\{高收入, 高学历\}, \{高收入, 低学历\}, \{低收入, 高学历\}, \{低收入, 低学历\}\}$
  - 令事件
    $H = \{高收入\} = \{高收入, 高学历\} \cup \{高收入, 低学历\}$,
    $C = \{高学历\} = \{高收入, 高学历\} \cup \{低收入, 高学历\}$
  - 则 $H \cap C = \{高收入, 高学历\}$ 是联合事件，其概率称为**联合概率（joint probability）**
- 联合概率的直观理解

联合概率

	高学历	低学历	任意学历
高收入	0.19	0.12	0.31
低收入	0.17	0.52	0.69
任意收入	0.36	0.64	1

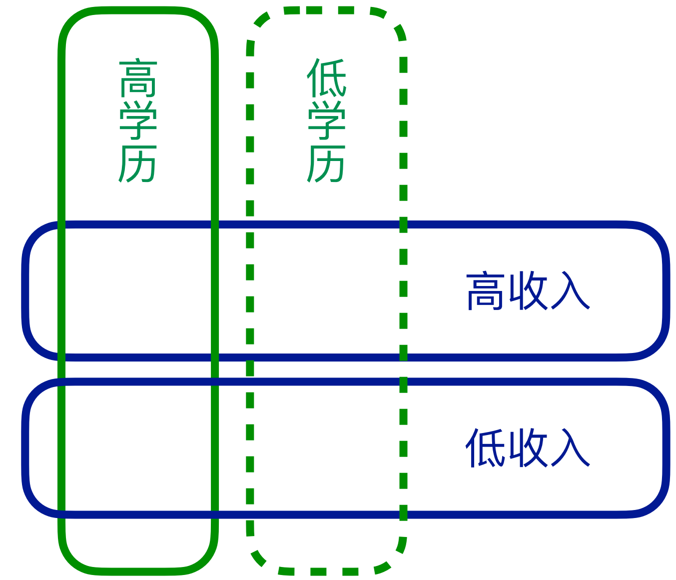


条件概率

Conditional probability

- 当你和小明逐渐熟悉，你了解到他是研究生毕业（高学历），此时，他收入的概率就发生了变化

	高学历	低学历	任意学历
高收入	0.19	0.12	0.31
低收入	0.17	0.52	0.69
任意收入	0.36	0.64	1



  - 任意学历下高收入的概率为 $\Pr(\text{高收入}) = 0.31$
  - 已知高学历时，高收入的**条件概率**为 $\Pr(\text{高收入} \mid \text{高学历}) = 0.19/0.36 = 0.53$

当 $\Pr(B) > 0$ 时，条件概率 $\Pr(A \mid B)$ 定义为

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

独立性

Independence

一个事件发生的概率如果不影响另一个事件发生的条件概率，我们说这两个事件是独立的。因此，独立事件的定义为

$$\left\{\begin{aligned}\Pr(A \mid B) = \Pr(A),\\ \Pr(B \mid A) = \Pr(B)\end{aligned}\right\} \Leftrightarrow \Pr(A \cap B) = \Pr(A)\Pr(B)$$

- 例如两个人分别抛一枚硬币，如果两个人的胳膊没有被连接在一起，一般情况下我们可以认为两枚硬币出现的面是独立的（两者之间没有关联）
- 由定义可知，独立事件的联合概率是各自概率的乘积。若令第一枚硬币的面为 H_1 和 T_1，第二枚硬币的面为 H_2 和 T_2，并令 $\Pr(H_1) = p$，$\Pr(H_2) = q$，则有

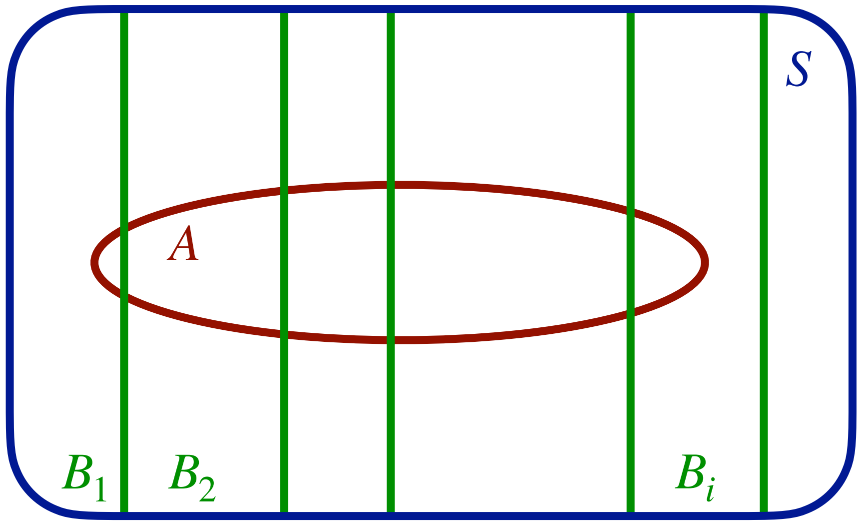
	H_1	T_1	
H_2	pq	$(1-p)q$	q
T_2	$p(q-1)$	$(1-p)(1-q)$	$1-q$
	p	$1-p$	1

- 小明的收入和学历是独立的吗?
- 不相交的事件 A 与 B 不是独立事件，为什么?

全概率公式

Law of total probability

- 若样本空间 S 中的事件 $B_1, B_2, \ldots$ 互不相交，且 $\bigcup_{i=1}^{\infty} B_i = S$，我们说 $B_1, B_2, \ldots$ 是 S 的**分割（partition）**



- 若 A 也是 S 中的事件，则

$$A = \bigcup_{i=1}^{\infty} (A \cap B_i)$$

- $\{A \cap B_i\}$ 互不相交，因此由概率公理第三条可得 $\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \cap B_i)$，带入条件概率公式可得下面的公式

全概率公式：若 $\{B_1, B_2, \ldots\}$ 是 S 的分割，且所有 $\Pr(B_i) > 0$，则

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(A \mid B_i) \Pr(B_i)$$

贝叶斯公式

The Bayes rule

- 贝叶斯公式是关于条件概率计算的便利公式
  - 根据条件概率的定义可知
$$\Pr(A \cap B) = \Pr(A \mid B)\Pr(B) = \Pr(B \mid A)\Pr(A)$$
  - 解第二个等式可得
$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$$
  - A 和 A^c 是样本空间 A 的分割，因此，对 B 使用全概率公式可得
$$\Pr(B) = \Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c)$$

    带入上式可得贝叶斯公式

贝叶斯公式：如果 $\Pr(A) > 0$，$\Pr(B) > 0$，则

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c)}$$

贝叶斯公式

The Bayes rule: an example

贝叶斯公式：

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B \mid A)\Pr(A) + \Pr(B \mid A^c)\Pr(A^c)}$$

- 假设大学毕业生分为认真工作（E）和不认真工作（N）两种，从历年的行业调研可知 $\Pr(E) = 1/4$, $\Pr(N) = 3/4$

- 企业希望能够雇佣更多认真工作的员工，于是开发了一种测试系统，并通过早期实验得知不同员工在该测试中得分高（H）与低（L）的概率分别为

$$\Pr(H \mid E) = 4/5,\ \Pr(L \mid E) = 1/5$$
$$\Pr(H \mid N) = 1/3,\ \Pr(L \mid N) = 2/3$$

- 企业能否通过测试成绩鉴别优秀毕业生呢?

$$\Pr(E \mid H) = \frac{\Pr(H \mid E)\Pr(E)}{\Pr(H \mid E)\Pr(E) + \Pr(H \mid N)\Pr(N)} = \frac{(4/5)\times(1/4)}{(4/5)\times(1/4) + (1/3)\times(3/4)} = \frac{4}{9}$$

$$\Pr(E \mid L) = \frac{\Pr(L \mid E)\Pr(E)}{\Pr(L \mid E)\Pr(E) + \Pr(L \mid N)\Pr(N)} = \frac{(1/5)\times(1/4)}{(1/5)\times(1/4) + (2/3)\times(3/4)} = \frac{1}{11}$$

- 可见，企业在招聘中采用这种测试系统是可以获得一定收益的：在测试正获得高分的毕业生，其认真工作的可能性为 4/9，大于不采用测试时的概率 1/4；而在测试中获得低分的毕业生，其认真工作的可能性为 1/11，远小于不采用测试时的概率 3/4。

随机变量

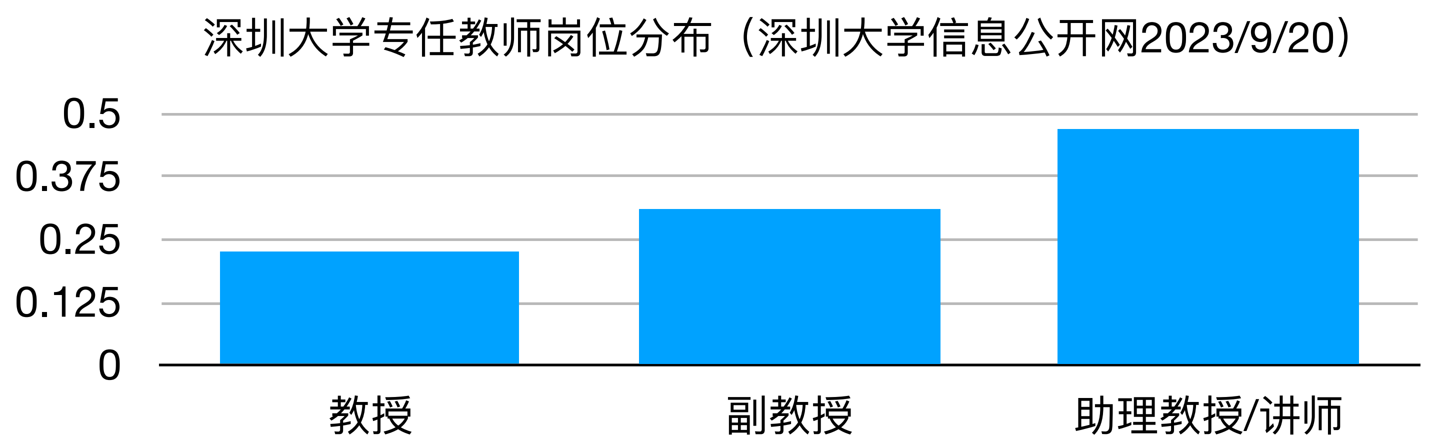
Random variable

随机变量是从样本空间 S 映射向实数 $\mathbb{R}$ 的函数。

- “抛一枚硬币”的样本空间是 $S = \{H, T\}$，我们可以定义随机变量 X：

$$X = \begin{cases} 1 & \text{if } H \\ 0 & \text{if } T \end{cases}$$

- 如果随机变量 X 的取值为离散值，我们称之为**离散随机变量**
- 对于离散随机变量 X，其**概率函数（probability mass function）**为 $\pi(x) = \Pr(X = x)$。这里的概率 $\Pr(X = x)$ 是基于 $X = x$ 时发生的事件定义的。概率为正的取值称为支撑点（support point），支撑点的集合称为**支撑（support）**。
- 概率函数可以通过柱形图描绘



期望值

Expectation/Expected value

支撑为 $\mathscr{X}$ 的离散随机变量 X 的期望值定义为

$$E[X] = \sum_{x\in\mathscr{X}} x\,\pi(x) = \sum_{x\in\mathscr{X}} x\,\Pr(X = x)$$

- 期望值也称均值（average/mean），因为他是随机变量的取值用其发生概率进行加权平均后得到的值
- “掷一个公正骰子”得到的数字的期望值是

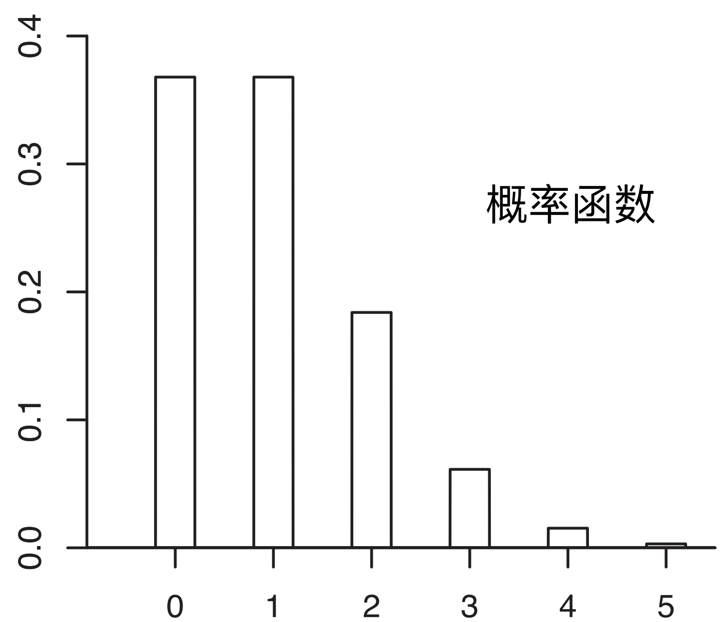
$$\sum_{i=1}^{6} i \times \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{7}{2}$$

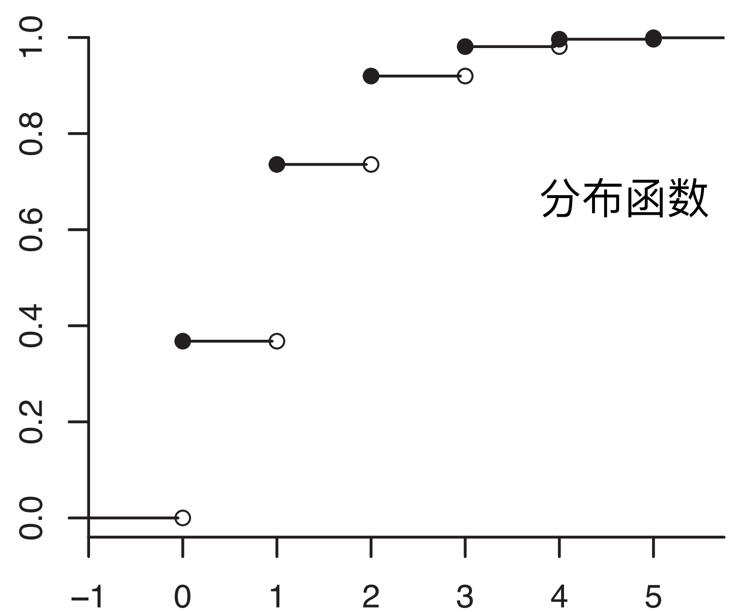
- 当支撑包含无限个点时，期望值可能不存在
  - 圣彼得堡悖论（St. Petersburg Paradox）：反复抛一枚硬币直至出现 H 为止。如果在第 k 轮出现了 H，则你可以获得 2^k 元现金。那么你愿意为参加这个游戏付多少钱呢?
  - 令游戏的持续轮数为 K，则 K 是随机变量，其概率函数为 $\Pr(K = k) = 2^{-k}$
  - 游戏回报的期望值是 $\sum_{i=1}^{\infty} 2^k\, 2^{-k} = \sum_{i=1}^{\infty} 1 = \infty$，即不存在（级数是发散的）

累积分布函数

Cumulative distribution function (CDF)

随机变量 X 的**累积分布函数**定义为 $F(x) = \Pr(X \le x)$，写为 $X \sim F(x)$





- 概率函数的部分和是分布函数：

$$F(x) = \sum_{a \in \mathcal{X}, a \le x} \Pr(X = a)$$

- 分布函数的差分是概率函数：

$$\Pr(X = x) = F(x) - \lim_{\varepsilon \to 0} F(x - \varepsilon)$$

- 分布函数 $F(x)$ 的一般性质
    1. $F(x)$ 是非减少函数
    2. $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$
    3. $F(x)$ 是右连续函数，即 $\lim_{x \downarrow a} F(x) = F(a)$

连续随机变量

Continuous random variable

如果随机变量 X 的取值为连续值，我们称之为**连续随机变量**。下面通过分布函数给出一个等价的定义：

当 $X \sim F(x)$ 且 $F(x)$ 为连续函数时，称 X 为连续随机变量

- 服从均匀分布（uniform distribution）的随机变量是连续的：

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

尝试画出这个函数的图像

- 区间概率：$\Pr(a < X \leq b) = F(b) - F(a)$

- 连续随机变量取任意特定值的概率都为零：

$$\Pr(X = x) = \lim_{\varepsilon \to 0} \Pr(x - \varepsilon < X \leq x) = F(x) - \lim_{\varepsilon \to 0} F(x - \varepsilon) = 0$$

- 因此，$\Pr(X \leq x) = \Pr(X < x) = F(x)$

概率密度函数

Probability density function (PDF)

- 连续随机变量没有对应的概率函数（因为取特定值的概率为零），但是我们可以通过分布函数定义和概率函数相对应的**密度函数**

当连续随机变量 X 的分布函数是 $F(x)$，且可微分（differentiable）时，其密度函数为 $f(x) = \frac{d}{dx}F(x)$

- 任意函数 $f(x)$ 是密度函数的充分必要条件是：

  1. $f(x) \geq x$ for all x
  2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

- 区间概率：

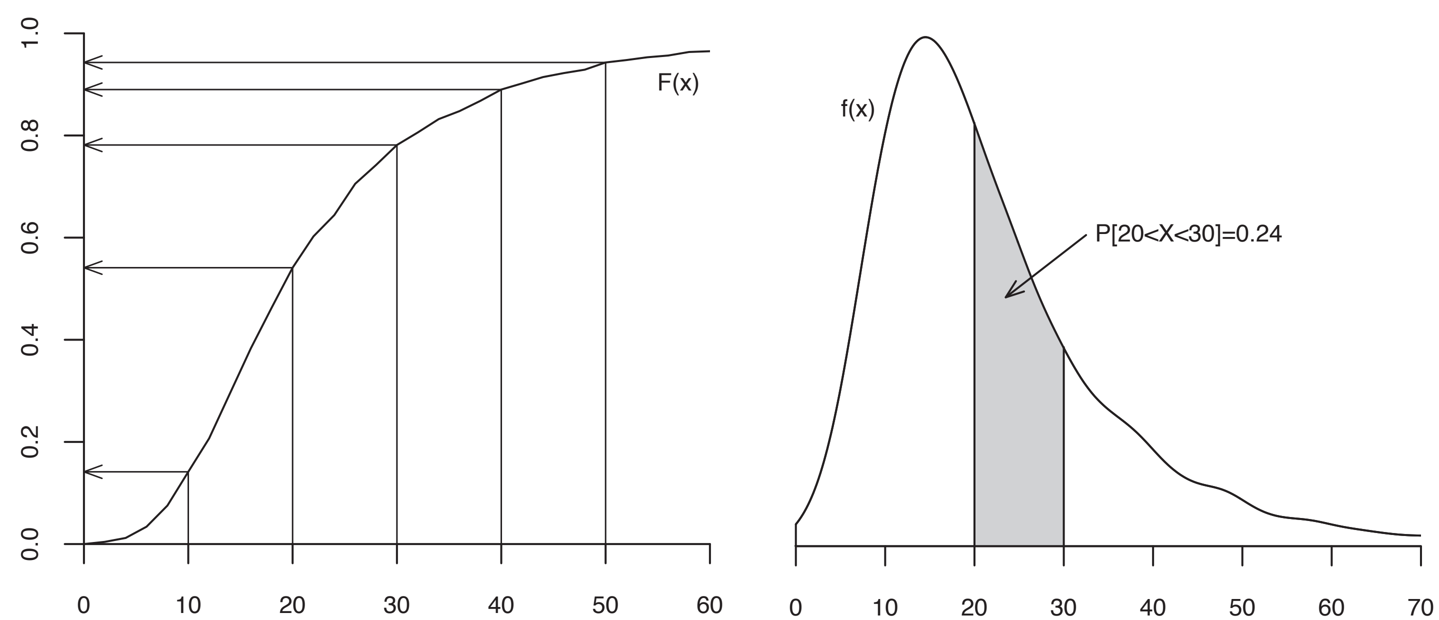
$$\Pr(a < X \leq b) = \int_a^b f(x)\,dx$$

- 期望值：$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$

- 正态分布（normal distribution）的密度函数：

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

2009年美国时薪的分布函数（左）与密度函数（右）



矩

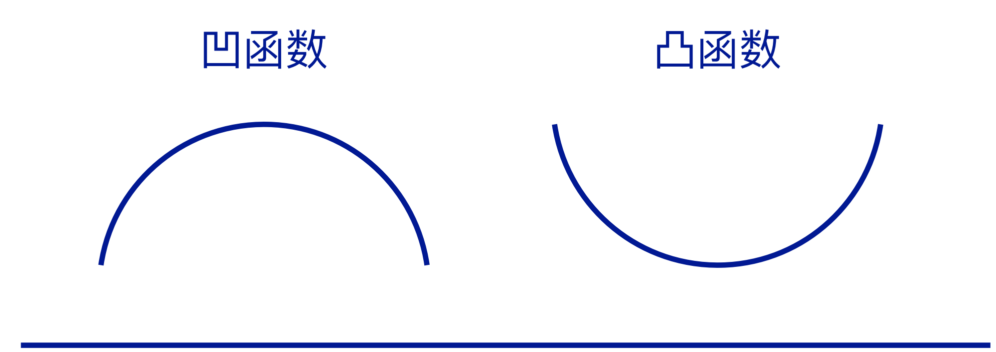
Moments

随机变量 X 的 k **阶矩**（k**-th moment**）是 $E[X^k]$

- 若 X 是离散变量，则 $E[X^k] = \sum_{x \in \mathcal{X}} x^k \Pr(X = x)$
- 若 X 是连续变量，则 $E[X^k] = \int_{-\infty}^{\infty} x^k f(x)\, dx$

当 $k > 1$ 时，随机变量 X 的 k **阶中心矩**（k**-th central moment**）是 $E\big[(X - E[X])^k\big]$，同时定义 $E[X]$ 为一阶中心矩

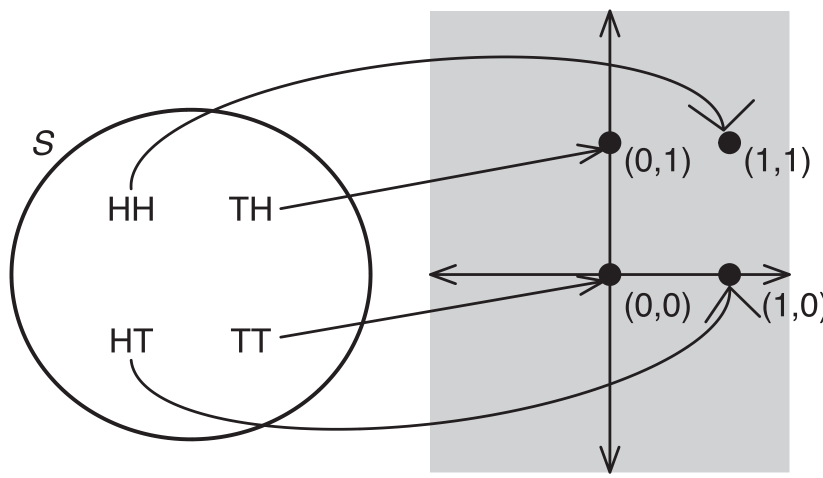
- $2, 3, 4$ 阶中心矩分别称为**方差**（**variance**），**偏度**（**skewness**）和**峰度**（**kurtosis**），其中方差写为 $\mathrm{Var}[X]$ 或 $\mathrm{var}[X]$，其算数平方根称为**标准偏差**（**standard deviation**）
- 矩的一些性质：
  1. $E[a + bX] = a + bE[X]$
  2. $\mathrm{Var}[X] = E[X^2] - (E[X])^2$
  3. $\mathrm{Var}[a + bX] = b^2 \mathrm{Var}[X]$
  4. Jensen's inequality:
     如果 $g(x)$ 是凸函数，则 $g(E[X]) \le E[g(x)]$；
     如果 $g(x)$ 是凹函数，则 $g(E[X]) \ge E[g(x)]$

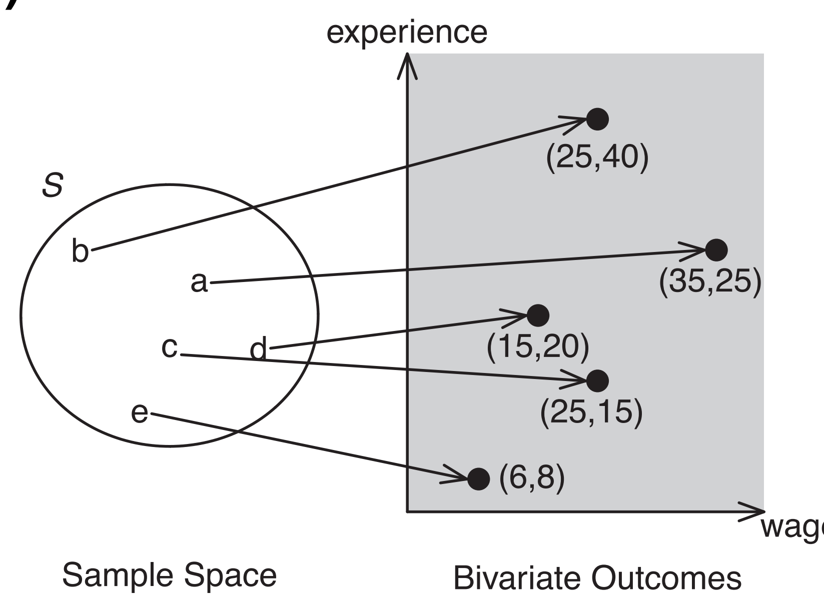


联合分布

Joint distributions

- 在之前的讨论中，我们把样本空间 S 投影到 $\mathbb{R}$ 得到了随机变量。如果我们将 S 投影到多维欧氏空间，则可获得**随机向量（random vector）**





二维随机向量 (X, Y) 的**联合分布（joint distribution）**是

$$F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y) = \Pr\left[\{X \leq x\} \cap \{Y \leq y\}\right]$$

联合分布函数为连续且可导时，可以定义联合密度（joint density）函数

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x, y)$$

在不产生误解的情况下，可简写为 $F(x, y)$ 和 $f(x, y)$

边际分布

Marginal distribution

- 边际分布是指在给出联合分布的前提下，每个随机变量各自的分布

对于 $(X, Y) \sim F(x, y)$，X 的边际分布是

$$F_X(x) = \Pr(X \leq x) = \Pr(X \leq x, Y < \infty) = \lim_{y \to \infty} F(x, y)$$

- 从定义可推出 $F_X(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{x} f(u, v)\, du\, dv$
- 边际密度函数为 $f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$

边际概率函数

	H_1	T_1	
H_2	pq	$(1-p)q$	q
T_2	$p(q-1)$	$(1-p)(1-q)$	$1-q$
	p	$1-p$	1

条件分布

Conditional distribution

- 条件分布是指某一随机变量（Y）在其他随机变量取特定值（$X = x$）时的分布。条件分布的定义因条件中变量是离散还是连续而不同

- X 是离散随机变量时，若 $\Pr(X = x) > 0$，则 Y 关于 $X = x$ 的**条件分布**和**条件密度函数**分别是

$$F_{Y|X}(y \mid x) = \Pr(Y \leq y \mid X = x), \qquad f_{Y|X}(y \mid x) = \frac{\partial}{\partial y} F_{Y|X}(y \mid x)$$

- X 是连续随机变量时，$\Pr(X = x) = 0$。若 $f_X(x) > 0$，则定义 Y 关于 $X = x$ 的**条件密度函数**为

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

如果联合分布函数 $F(x, y)$ 关于 x 可导，且 $f_X(x) > 0$，则 Y 关于 $X = x$ 的**条件分布函数**为

$$F_{Y|X}(y \mid x) = \frac{\frac{\partial}{\partial x} F(x, y)}{f_X(x)}$$

独立性（续）

Independence (cont.)

- 事件 A 和 B 独立 $\Leftrightarrow \Pr(A \cap B) = \Pr(A)\Pr(B)$
- 若令 $A = \{X \leq x\}, B = \{Y \leq y\}$，则可用分布函数定义随机变量的独立性

随机变量 X 和 Y 在**统计学上独立（statistically independent）**的定义是，所有 (x, y) 满足下式

$$F(x, y) = F_X(x)F_Y(y)$$

写作 $X \perp\!\!\!\perp Y$。

- 如果存在概率函数或密度函数，则独立的定义可以写成

$$\pi(x, y) = \pi_X(x)\pi_Y(y) \quad \text{或} \quad f(x, y) = f_X(x)f_Y(y)$$

- 定理：如果 $X \perp\!\!\!\perp Y$ 且两者都是连续变量，则条件密度等于边际密度

$$f_{Y|X}(y \mid x) = f_Y(y), \quad f_{X|Y}(x \mid y) = f_X(x)$$

证明这个定理

协方差与相关系数

Covariance and correlation coefficient

- 协方差和相关系数都是衡量两个随机变量间关系的指标

如果随机变量 X 和 Y 的方差皆为有限，则两者间的**协方差（covariance）**是

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

两者间的相关系数（correlation coefficient）是

$$\mathrm{Corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}$$

$$\boxed{-1 \le \mathrm{Corr}[X,Y] \le 1}$$

- 如果 $\mathrm{Cov}[X,Y]=0$，我们称 X 和 Y 不相关（uncorrelated）
- 定理：$X \perp\!\!\!\perp Y \;\; \overset{\Rightarrow}{\not\Leftarrow} \;\; \mathrm{Cov}[X,Y]=0$
- 定理：$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$

条件期望

Conditional expectation

随机变量 Y 关于 $X = x$ 的**条件期望**是条件分布 $F_{Y|X}(y \mid x)$ 的期望值，记为 $m(x) = E[Y \mid X = x]$

- 如果 X 和 Y 均为离散变量，则 $E[Y \mid X = x] = \dfrac{\sum_i y_i \, \pi(x, y_i)}{\pi_X(x)}$
- 如果 Y 是连续变量，则 $E[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) \, dy$

需注意：X 为离散和连续变量时，条件密度 $f_{Y|X}(y \mid x)$ 的定义不同

- 条件期望可以描述不同组别内的期望值，例如当 X 代表教育程度（高中、本科、硕士、博士），Y 代表收入时，$E[Y \mid X = \text{博士}]$ 就是博士毕业生的平均收入

- 条件期望是计量经济学的核心概念，在回归模型中，回归函数可以解释为条件期望

迭代期望定律

Law of iterated expectation

- 条件期望 $m(x) = E[Y \mid X = x]$ 是 x 的确定函数，但如果我们没有确定 $X = x$ 已经发生，则 $X = x$ 伴随着概率 $\pi_X(x)$
- 如果我们考虑所有可能的 X 的取值及其概率分布，则条件期望可以看作随机变量 X 的函数，其本身也是随机变量，写作 $m(X) = E[Y \mid X]$
- $E[Y \mid X]$ 是随机变量，因此可以计算期望值。一个非常重要的结果是下面的迭代期望定律

迭代期望定律：如果 $E[Y] < \infty$，则 $E\big[E[Y \mid X]\big] = E[Y]$

X 和 Y 均为连续变量时的证明：

$$E\big[E[Y \mid X]\big] = \int_{-\infty}^{\infty} E[Y \mid X = x]\, f_X(x)\, dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x) f_X(x)\, dy\, dx$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x, y)\, dy\, dx = E[Y]$$

均值独立

Mean independence

均值独立性：当 $E[Y \mid X] = E[Y]$ 时，称 Y 均值独立于 X

- 注意：均值独立性并不是对称概念，“Y 均值独立于 X”不等于 “X 均值独立于 Y”

定理：$X \perp\!\!\!\perp Y \quad \begin{matrix}\Rightarrow \\ \not\Leftarrow\end{matrix} \quad \begin{Bmatrix} E[Y \mid X] = E[Y] \\ E[X \mid Y] = E[X] \end{Bmatrix} \quad \begin{matrix}\Rightarrow \\ \not\Leftarrow\end{matrix} \quad \mathrm{Cov}[X, Y] = 0$

- 证明可参考 https://www.econometrics.blog/post/why-econometrics-is-confusing-part-ii-the-independence-zoo/

- 这个定理告诉我们，均值独立性比统计学上的独立性弱。在回归模型中，我们经常会用到均值独立性作为假设条件

条件方差与方差分解公式

Conditional variance and variance decomposition

- Y 关于 $X=x$ 的**条件方差**定义为 $\mathrm{Var}[Y \mid X=x] = E\left[(Y-m(x))^2 \mid X=x\right]$
- 由定义可知 $\mathrm{Var}[Y \mid X=x] = E[Y^2 \mid X=x] - m(x)^2$
- $\mathrm{Var}[Y \mid X]$ 也是随机变量 X 的函数

方差的分解公式：$\mathrm{Var}[Y] = E\left[\mathrm{Var}[Y \mid X]\right] + \mathrm{Var}\left[E[Y \mid X]\right]$

总方差 = 组内方差 + 组间方差

X 和 Y 均为连续变量时的证明：

$$
\begin{aligned}
E\left[\mathrm{Var}[Y \mid X]\right] &= \int_{-\infty}^{\infty} \mathrm{Var}[Y \mid X=x]\, f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \left\{E[Y^2 \mid X=x] - m(x)^2\right\} f_X(x)\, dx = \int_{-\infty}^{\infty} E[Y^2 \mid X=x]\, f_X(x)\, dx - \int_{-\infty}^{\infty} m(x)^2 f_X(x)\, dx \\
&= E\left[E[Y^2 \mid X]\right] - E\left[m(X)^2\right] = E[Y^2] - E\left[m(X)^2\right]
\end{aligned}
$$

迭代期望定律

已知 $\mathrm{Var}[Y] = E[Y^2] - (E[Y])^2$，$\mathrm{Var}[m(X)] = E\left[m(X)^2\right] - \left(E[m(X)]\right)^2 = E\left[m(X)^2\right] - (E[Y])^2$，可得

$$E\left[\mathrm{Var}[Y \mid X]\right] = \mathrm{Var}[Y] - \mathrm{Var}[m(X)]$$