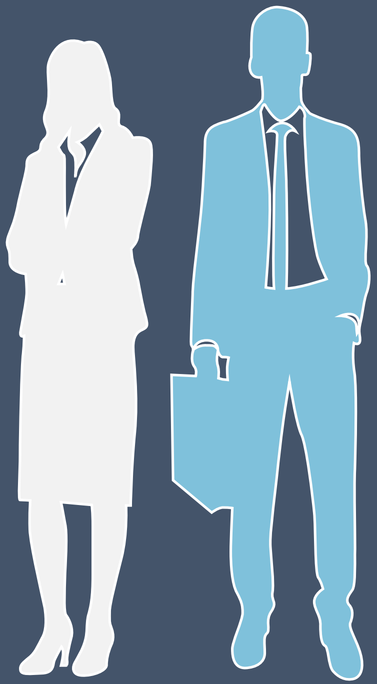


金融基础-数量分析

严心宇

CONTENTS

AQF Analyst of Quantitative Finance

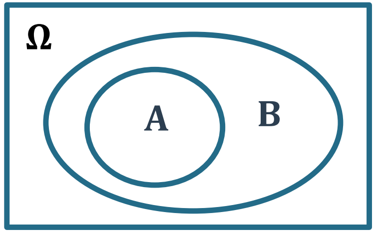
随机事件

- **随机试验**
  - 试验可以在相同条件下重复进行
  - 试验的结果不止一个，且事先可以明确试验的所有可能结果
  - 试验之前无法预知会出现哪一个结果
- **样本空间**
  - 随机试验所有可能结果组成的集合
  - 样本点：样本空间的元素，即随机试验的每个结果
- **随机事件**
  - 随机试验的结果，样本空间的子集，简称事件
  - 当且仅当这一子集中的一个样本点出现时，称这一事件发生
  - 基本事件：由一个样本点组成的单点集
  - 复合事件：由多个样本点组成的集合

事件间的关系

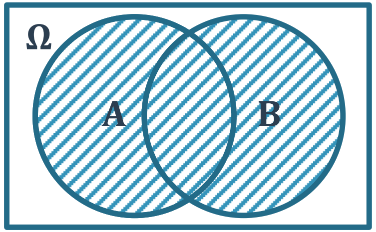
- **包含关系：** $A \subseteq B$ 表示"A 发生则 B 发生"
- **相等事件：** 如果 $A \subseteq B$ 且 $B \subseteq A$，则 $A = B$
- **和（并）事件：** $A \cup B$，表示"A、B 中至少有一个发生"的事件
- **积（交）事件：** $A \cap B$（或 AB），表示"A、B 同时发生"的事件

$A \subseteq B$



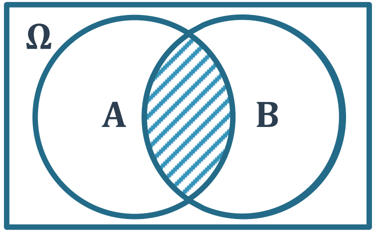
包含关系

$A \cup B$



和（并）事件

$A \cap B$

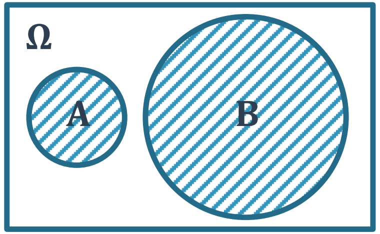


积（交）事件

事件间的关系

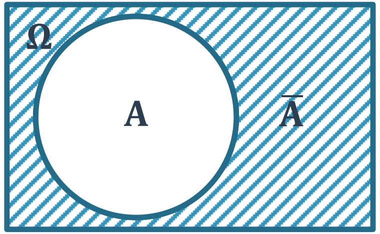
- **互斥事件：** 若 $A\cap B=\emptyset$，则称事件 A 与 B“互斥”或“互不相容”，表示“A、B 不能同时发生”
- **对立事件：** 若 $A\cup B=\Omega$ 且 $A\cap B=\emptyset$，则称事件 A 与事件 B 互为“逆事件”或“对立事件”，表示“每次试验中，事件 A、B 中必有一个发生，且仅有一个发生”；A 的对立事件记为 $\bar{A}$
- **差事件：** $A-B=A\cap\bar{B}$，表示“A 发生但 B 不发生”的事件

$\mathbf{A\cap B=\emptyset}$



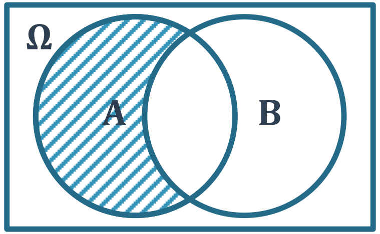
互斥事件

$\mathbf{A\cup\bar{A}=\Omega, A\cap\bar{A}=\emptyset}$



对立事件

$\mathbf{A-B}$



差事件

概率的定义和性质

Analyst of Quantitative Finance

- **概率的定义**
  - 在一次试验中，某事件发生的可能性的大小
- **概率的性质**
  - $P(\emptyset)=0$，$0\le P(A)\le 1$，$P(\Omega)=1$
  - 差事件的概率：$P(A-B)=P(A)-P(AB)$
  - 逆事件的概率：$P(\bar{A})=1-P(A)$
  - 如事件 $A_1,A_2,\ldots,A_n$ 互不相容，则：$P(\bigcup_{k=1}^{n}A_k)=\sum_{k=1}^{n}P(A_k)$
  - 如有任意事件 A、B，且 $A\subseteq B$，则：$P(A)\le P(B)$，$P(B-A)=P(B)-P(A)$
  - 加法公式：$P(A\cup B)=P(A)+P(B)-P(AB)$
    - ✓ 若 A、B 互斥，则：$P(A\cup B)=P(A)+P(B)$
    - ✓ 若 A、B 独立，则：$P(A\cup B)=P(A)+P(B)-P(A)P(B)$

概率模型

古典概型

- 试验的样本空间只包含有限个元素，且每个基本事件发生的可能性相等
- 事件 A 的概率为：

$$P(A) = \frac{\text{事件 A 包含的样本点的个数}}{\text{样本空间中样本点的总数}}$$

- 例：抛硬币，掷骰子

排列组合

- 排列：从 n 个不同元素中取出 $m(m \le n)$ 个元素的所有排列的个数，用符号 A_n^m 表示

$$A_n^m = n(n-1)\ldots(n-m+1) = \frac{n!}{(n-m)!}$$

- 组合：从 n 个不同元素中取出 $m(m \le n)$ 个元素的所有组合的个数，用符号 C_n^m 表示

$$C_n^m = \frac{A_n^m}{m!} = \frac{n!}{m!\,(n-m)!}$$

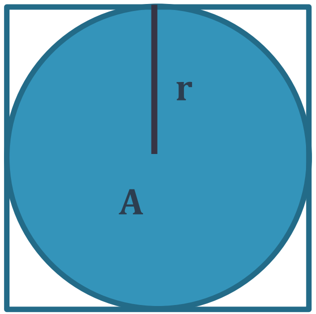
概率模型

几何概型

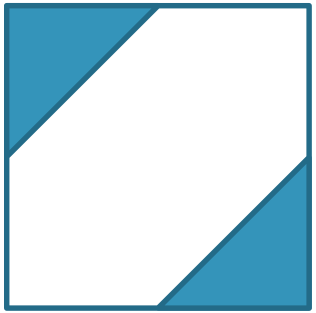
- 试验的样本空间包含无限个元素，且每个基本事件发生的可能性相等
- 如果每个事件发生的概率只与构成该事件区域的测度（长度、面积、体积等）成比例，而与该区域的位置和形状无关，则称这样的概率模型为几何概率模型，简称为几何概型
- 事件 A 的概率为：

$$P(A) = \frac{A\text{的测度}}{\text{样本空间的测度}}$$

- 例：扔石子问题，见面问题



$$P(A) = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}$$



边际、联合、条件概率

- **边际概率**：指一个事件发生的概率，而不考虑过去或者将来其他事件发生的情况，一般记为 $P(A)$
- **联合概率**：指两个事件同时发生的概率，一般记为 $P(AB)$
- **条件概率**：指在一个事件已经发生的条件下，考虑另一个事件发生的概率，一般记为 $P(A|B)$，表示在事件 B 发生的情况下，事件 A 发生的概率

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

- **乘法公式**：
  - $P(AB) = P(A|B) \cdot P(B)$
  - 若 A、B 互斥，则：$P(AB) = 0$
  - 若 A、B 独立，则：$P(AB) = P(A) \cdot P(B)$

边际、联合、条件概率

➢ **例：**已知 A 和 B 联合概率矩阵：

A	B			
		B_1	B_2	A_i
	A_1	14%	6%	20%
	A_2	20%	30%	50%
	A_3	6%	24%	30%
	B_j	40%	60%	1

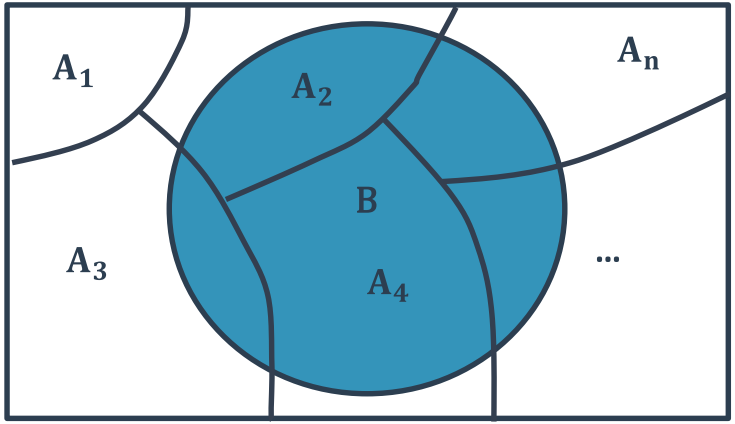
- 事件 A_2 与事件 B_1 的联合概率
- 事件 A_2 发生的边际概率
- 在事件 A_2 发生的情况下，事件 B_1 发生的概率

全概率公式

- 如果样本空间 A 划分为 A_1、A_2、 …、A_n，并且相互间为互斥事件，那么事件 B 发生的概率为：

$$P(B) = P(A_1B) + P(A_2B) + \cdots + P(A_nB)$$

$$= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$



贝叶斯公式

➢ **贝叶斯公式**

- 如果样本空间 A 划分为 A_1、A_2、 …、 A_n，并且相互间为互斥事件，那么事件 B 发生的情况下事件 A_i 发生的概率为：

后验概率

$$P(A_i|B)=\frac{P(A_iB)}{P(B)}=\frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{n}P(A_j)P(B|A_j)}$$

$$=P(A_i)\cdot\boxed{\frac{P(B|A_i)}{\sum_{j=1}^{n}P(A_j)P(B|A_j)}}$$

先验概率

信息调整因子

随机变量

- **离散型随机变量**
  - 随机变量全部可能取到的值是有限个或可列无限多个
- **概率质量函数（PMF）**
  - 设离散型随机变量 X 所有可能取值为 x_1，$x_2, \ldots, x_n$，则 X 的概率质量函数为：

$$P(X = x_i) = p_k\text{，其中 } i = 1, 2, \ldots, n$$

  - 离散型随机变量的概率质量函数的性质：$\sum_{i=1}^{n} P(x_i) = 1$
- **累积分布函数（CDF）**
  - 离散型随机变量 X 的累积分布函数记为：

$$F(x) = P(X \leq x)$$

  - 表示离散型随机变量 X 取值小于等于 x 时的概率

随机变量

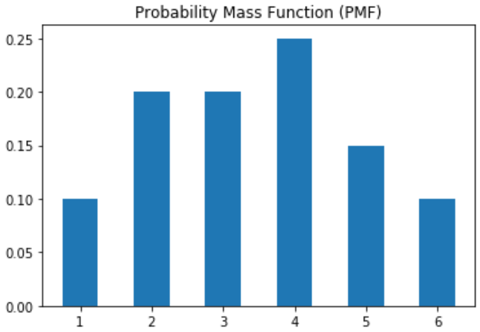
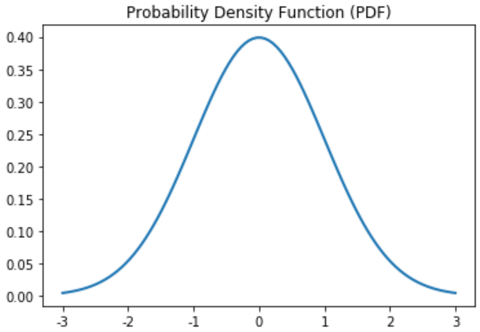
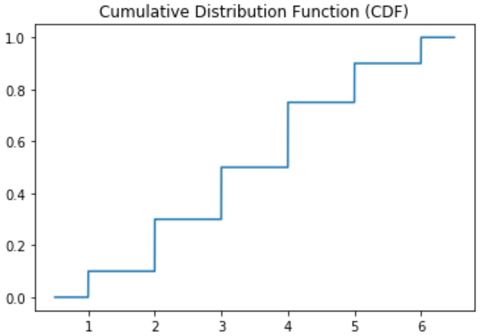
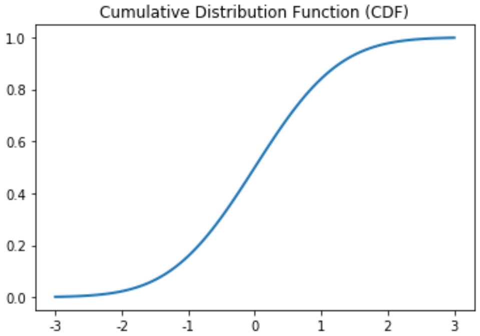
- **连续型随机变量**
  - 随机变量的所有可能值不可以逐个列举出来
- **概率密度函数（PDF）**
  - 连续型随机变量 X 的概率密度函数通常用 $f(x)$ 来表示
  - 连续型随机变量的点概率等于零：$P(X = x_i) = 0$
  - 连续型随机变量的概率密度函数的性质：$\int_{-\infty}^{+\infty} f(x)dx = 1$
- **累积分布函数（CDF）**

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

  - 连续型随机变量 X 取值落在常数 a 与常数 b 之间的概率可以表示为

$$P(a < X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

概率函数和分布函数

	离散型	连续型
概率函数		
分布函数		

CONTENTS

期望

➢ **期望**

- 离散变量：$E(X)=\sum_{i=1}^{i=n}P(x_i)x_i$
- 连续变量：$E(X)=\int_{-\infty}^{\infty}xf(x)dx$

➢ **期望的性质（a、b、c 均为常数）**

- $E(c)=c$
- $E(aX)=aE(X)$，$E(X+b)=E(X)+b$
- $E(aX+b)=aE(X)+b$
- $E(X\pm Y)=E(X)\pm E(Y)$
- 一般来说，$E(XY)\neq E(X)E(Y)$；如果随机变量 X 与 Y 相互独立，则 $E(XY)=E(X)E(Y)$
- $E(X^2)\neq[E(X)]^2$，只有当 X 为常数时等号才成立

方差

方差

- $\mathrm{Var}(X) = \sigma^2 = E[(X - E(X))^2]$

方差的性质（a、b、c 均为常数）

- $\mathrm{Var}(c) = 0$
- $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$，$\mathrm{Var}(X + b) = \mathrm{Var}(X)$
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$
- 如果随机变量 X 与 Y 相互独立，则：
  - ✓ $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
  - ✓ $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$
- $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$

协方差

- **协方差**

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

  - 衡量两个变量的总体误差
  - 协方差的取值范围为负无穷到正无穷

- **协方差的性质**

  - $\text{Cov}(X,X) = \text{Var}(X)$
  - $\text{Cov}(a + bX, c + dY) = bd\text{Cov}(X,Y)$
  - $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\text{Cov}(X,Y)$
  - $\text{Var}(aX \pm bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) \pm 2ab\text{Cov}(X,Y)$
  - 如果随机变量 X 与 Y 相互独立，则：$\text{Cov}(X,Y) = 0$

相关系数

相关系数

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

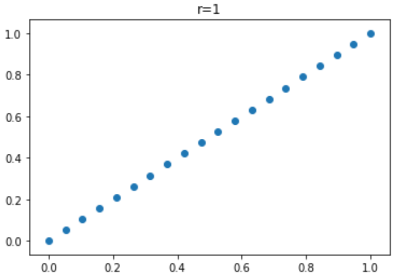
相关系数的性质

- $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\rho\sigma_X\sigma_Y$
- 相关系数衡量两个变量之间的线性关系
- 相关系数没有单位，取值从-1到1之间
- 相关系数不表明因果关系
- 如果两个变量相互独立，则相关系数为0
- 相关系数为0，两个变量不一定相互独立，例如：$X \sim U(-1,1), Y = \sqrt{1 - X^2}$

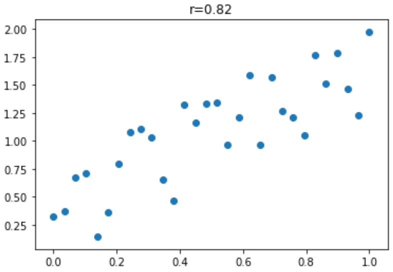
相关系数

➢ 线性相关程度

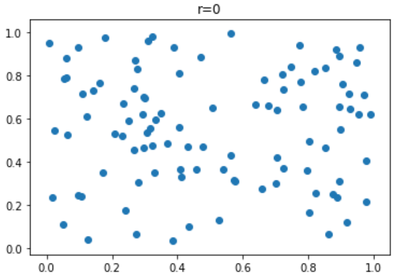
完全线性正相关



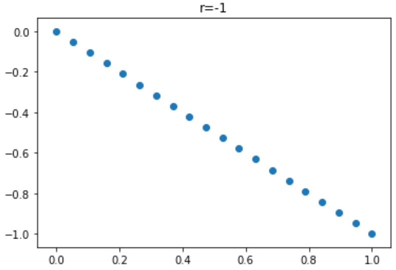
线性正相关



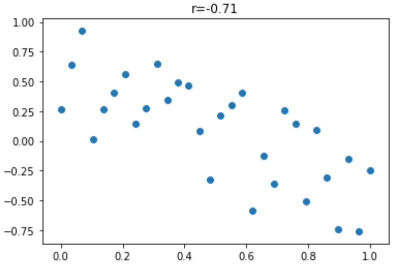
线性不相关



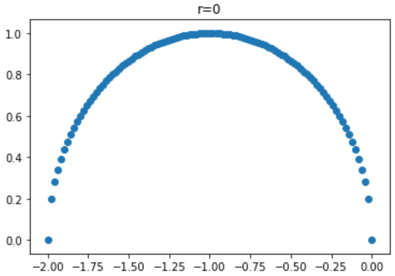
完全线性负相关



线性负相关



线性不相关



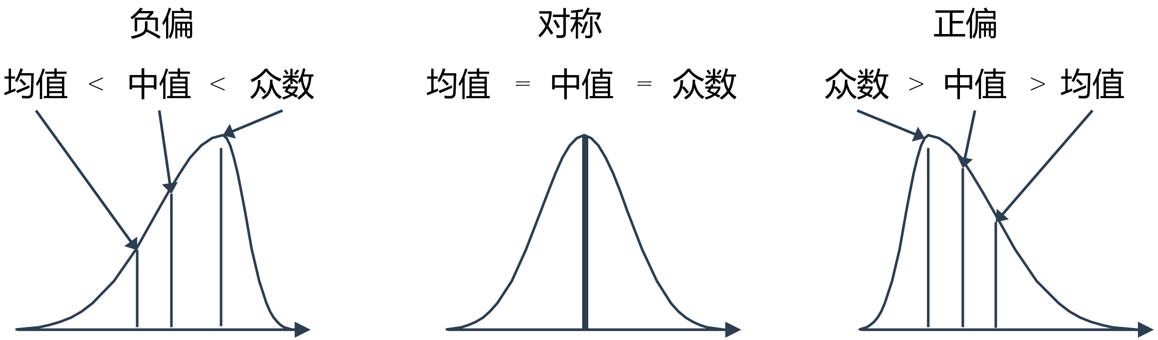
偏度

➢ 偏度

- 衡量概率密度函数的不对称性：

$$\text{Skewness} = \frac{E[(X-\mu)^3]}{\sigma^3}$$

- 正偏（右偏）与负偏（左偏）

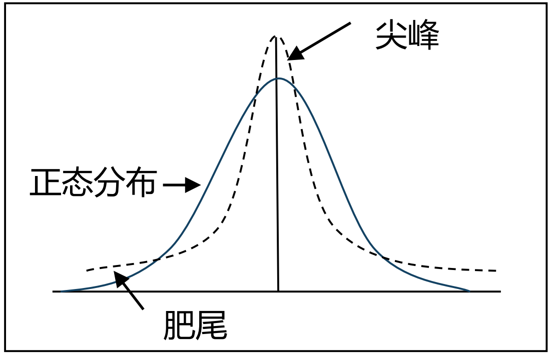


峰度

➢ 峰度

- 衡量概率密度函数峰部的尖度：

$$K = \frac{E[(X-\mu)^4]}{\sigma^4}$$



- **超额峰度** = 峰度 – 3

	尖峰	常峰态 (正态分布)	低峰
峰度	>3	$=3$	<3
超额峰度	>0	$=0$	<0

K阶距和中心距

K阶距

- 随机变量 X 的 k 阶距为：$m_k = E[X^k]$
- 当 $k = 1$ 时，有 $m_1 = E[X]$，可以看出 m_1 为数学期望

中心距

- 随机变量 X 的 k 阶中心距为：$\mu_k = E[(X - \mu)^k]$
- 如果 $k = 1$，一阶中心距为0
- 如果 $k = 2$，二阶中心距为方差
- 如果 $k = 3$，三阶中心距除以标准差的立方，为偏度
- 如果 $k = 4$，四阶中心距除以方差的平方，为峰度

CONTENTS

伯努利分布

➢ 伯努利试验

- 试验只有两种可能的结果 A 及 $\bar{A}$

➢ 伯努利分布（0-1分布）

- 随机变量 X 只取 0 和 1 两个值：

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

- 期望和方差：

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

Analyst of Quantitative Finance

二项分布

- n 次独立重复的伯努利试验中，成功次数为 k 的概率为：

$$P(X=k)=C_n^k p^k(1-p)^{n-k}$$

- 期望和方差：

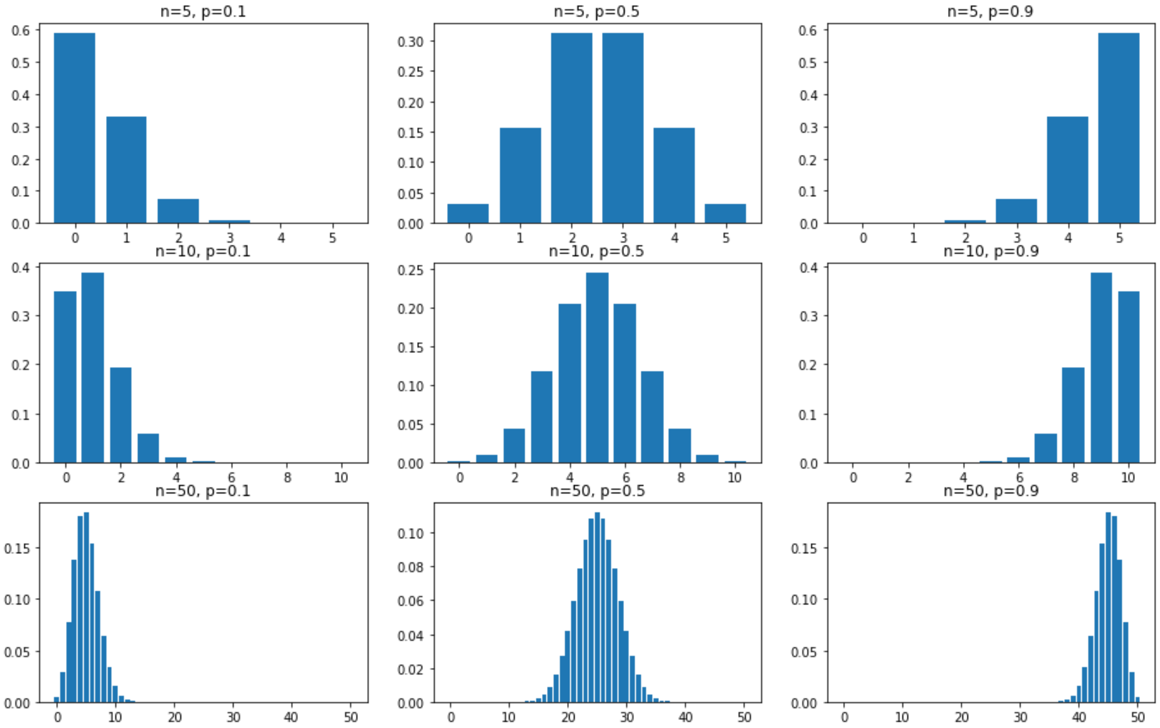
$$E(X)=np,\qquad Var(X)=np(1-p)$$

✓ $kC_n^k=nC_{n-1}^{k-1}$

✓ 二项展开式：$(a+b)^n=C_n^0a^0b^n+C_n^1a^1b^{n-1}+\cdots+C_n^na^nb^0$

Analyst of Quantitative Finance

➢ 二项分布



- 随着 n 的增大，二项分布趋近于正态分布

泊松分布

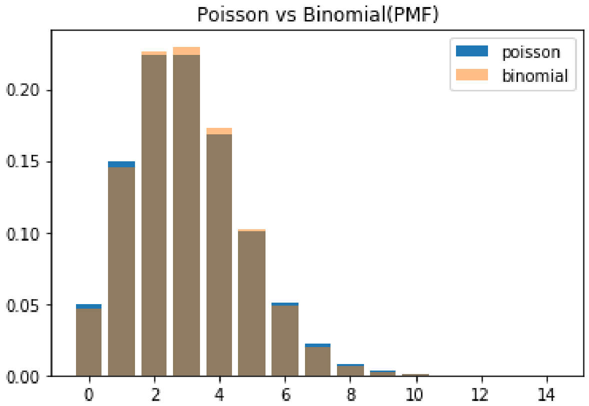
➢ 泊松分布

- 泊松分布适合于描述单位时间内随机事件发生的次数

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,\dots$$

✓ λ 是单位时间内随机事件的平均发生率

✓ 当二项分布 $n\to\infty$，$p\to 0$，而 np 比较稳定时，泊松分布可作为二项分布的逼近（$np=\lambda$）



- 期望和方差：

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

均匀分布

➢ 均匀分布

- 概率密度函数：

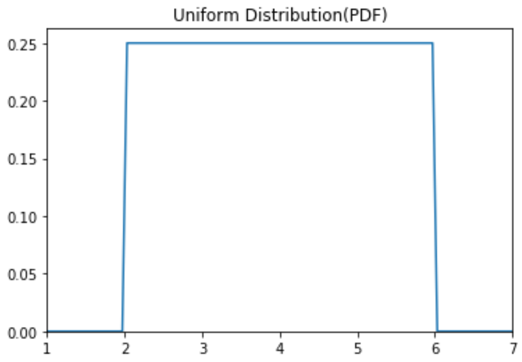
$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b\\ 0, & \text{其他}\end{cases}$$

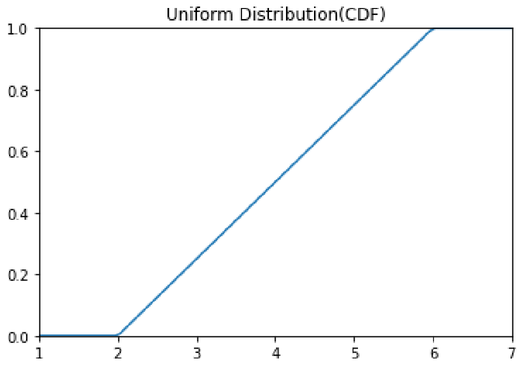
- 累积分布函数：

$$F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\le x<b\\ 1, & x\ge b\end{cases}$$

- 期望和方差：

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





指数分布

➢ **指数分布**

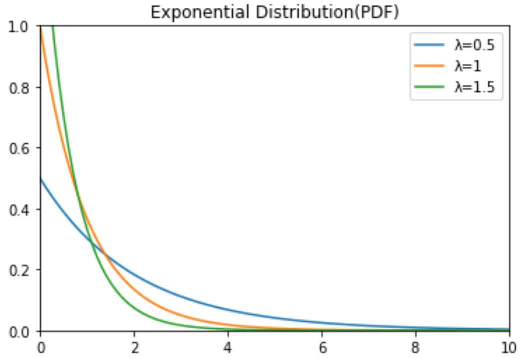
- 描述泊松过程中事件之间的间隔时间
- 概率密度函数：

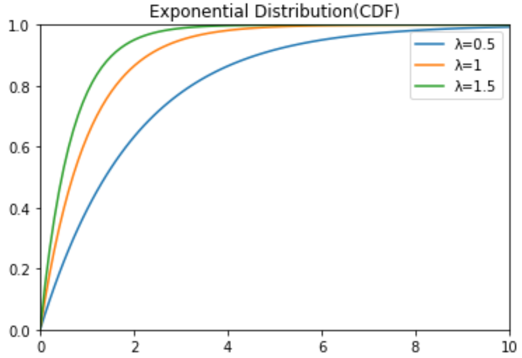
$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}$$

其中$\lambda>0$

- 累积分布函数：

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





➢ 指数分布

- 期望和方差：

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

- 无记忆性：

  ✓ 对于任意 $s,t>0$，有：

$$P(X>s+t|X>s)=\frac{P(X>s+t)}{P(X>s)}=\frac{e^{-\lambda(s+t)}}{e^{-\lambda s}}=e^{-\lambda t}=P(X>t)$$

  ✓ 如果 X 是某元件的寿命，已知元件已使用了 s 小时，它总共能使用至少 $s+t$ 小时的条件概率，与从开始使用时算起至少能使用 t 小时的概率相等，即元件对它已使用过 s 小时没有记忆

➢ 正态分布

- 概率密度函数:

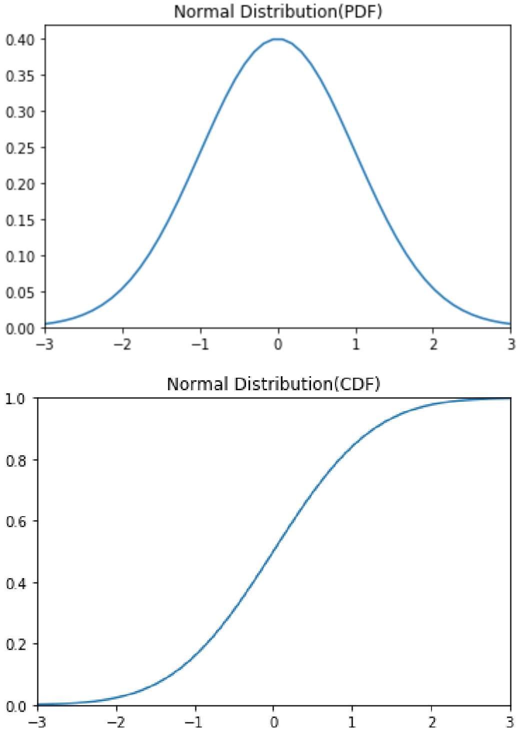
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

- 累积分布函数:

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

- 期望和方差:

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



正态分布

➢ 正态分布的可加性

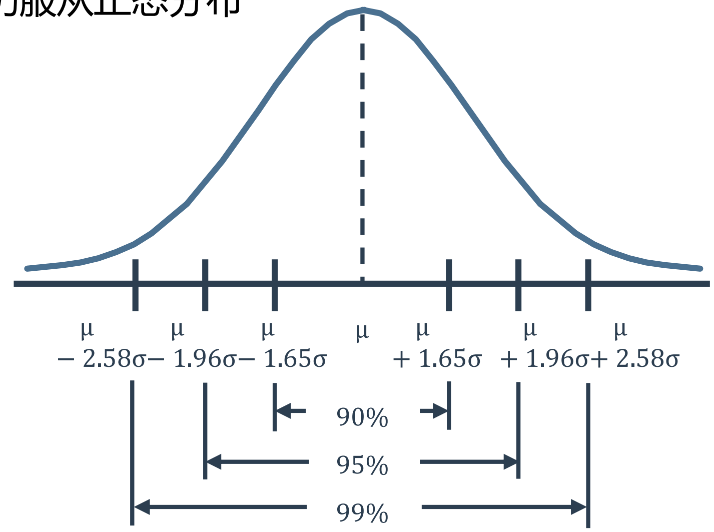
- 两个（或多个）满足正态分布的随机变量，经过线性组合构成的新的随机变量仍满足正态分布
  - ✓ 如果 $X \sim N(\mu, \sigma^2)$，则 $Y = mX + b \sim N(m\mu + b, m^2\sigma^2)$
  - ✓ 如果 $X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_2, \sigma_2^2)$，则 $Z = X + Y$ 仍服从正态分布

➢ 标准正态分布

- $N(0,1)$ 或 Z分布
- 标准化：如果 $X \sim N(\mu, \sigma^2)$，那么：

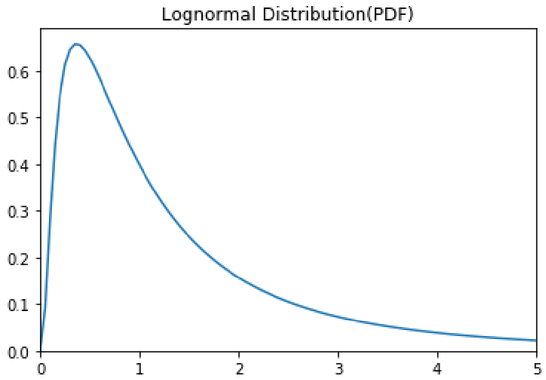
$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

- $\Phi(-x) = 1 - \Phi(x)$



对数正态分布

➢ 对数正态分布



- $\ln X \sim N(\mu, \sigma^2)$
- 右偏（正偏）

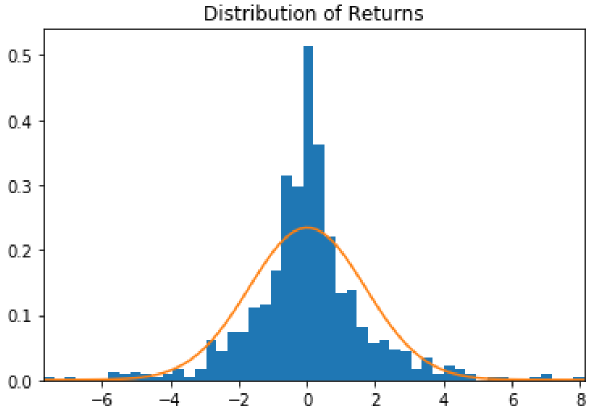
- 概率密度函数:

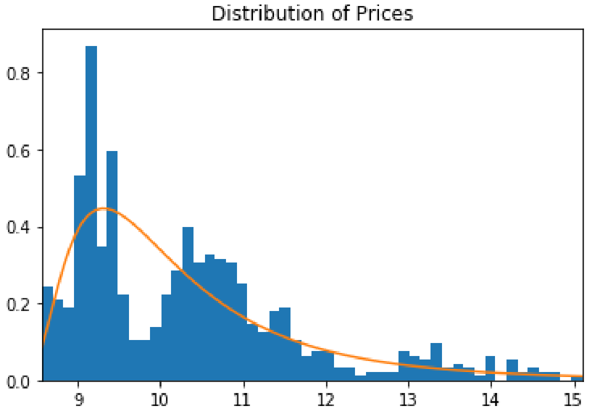
$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0$$

- 如果 $\ln X$ 满足正态分布，那么 X 满足对数正态分布；反之亦然
- 该分布在对资产价格建模时非常有用，例如：BSM模型假设标的资产价格满足对数正态分布

对数正态分布

➢ 例：某股票的日收益率和股价分布





- 资产收益率近似服从正态分布
- 资产价格近似服从对数正态分布

CONTENTS

切比雪夫不等式

➢ 切比雪夫不等式

- 设服从任意分布的随机变量 X 的数学期望 $E(X)=\mu$，方差 $D(X)=\sigma^2$，则：

$$P(|X-\mu| < k\sigma) \geq 1-\frac{1}{k^2},\quad k>1$$

- 利用切比雪夫不等式，可以在随机变量 X 的分布未知的情况下，对事件 $|X-\mu| \leq k\sigma$ 的概率作出估计

➢ 例：对于任意一个分布而言，观测值落在偏离均值正负3个标准差内的概率最小为多少?

- 解析：根据切比雪夫不等式：

$$P(|X-\mu| \leq 3\sigma) \geq 1-\frac{1}{3^2} \approx 89\%$$

大数定律

- 设随机变量 $X_1, X_2, \dots, X_n$ 独立同分布，期望为 μ，$S_n = X_1 + X_2 + \cdots + X_n$，则 $\frac{S_n}{n}$ 收敛到 μ：

$$\lim_{n \to \infty} \overline{X} = \mu$$

中心极限定理

- 设随机变量 $X_1, X_2, \dots, X_n$ 独立同分布，且具有有限的数学期望和方差：$E(X_k) = \mu$，$D(X_k) = \sigma^2 > 0$。当 n 充分大时，样本均值近似服从正态分布，即：

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

总体和样本

总体和样本

$$
\begin{tikzcd}
\text{总体} \arrow[r, dashed] \arrow[d, "\text{抽样}"'] & \text{总体参数} \quad \mu，\sigma^2 \\
\text{样本} \arrow[r] & \text{样本统计量} \quad \overline{X}，S^2 \arrow[u, "\text{估计}"']
\end{tikzcd}
$$

样本统计量

- 样本均值：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 总体均值：$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$
- 样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- 总体方差：$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$

样本统计量

➢ 样本统计量

- **标准差（SD）：** 衡量数据分散程度
  - ✓ 总体标准差： $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$ ✓ 样本标准差： $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

- **标准误（SE）：** 样本统计量的标准差，是衡量样本统计量抽样误差大小的尺度
  - ✓ 样本均值的标准误： $SEM = \frac{\sigma}{\sqrt{n}}$
  - ✓ 由于通常 σ 未知，可以用样本标准差 S 替代，则： $SEM = \frac{S}{\sqrt{n}}$

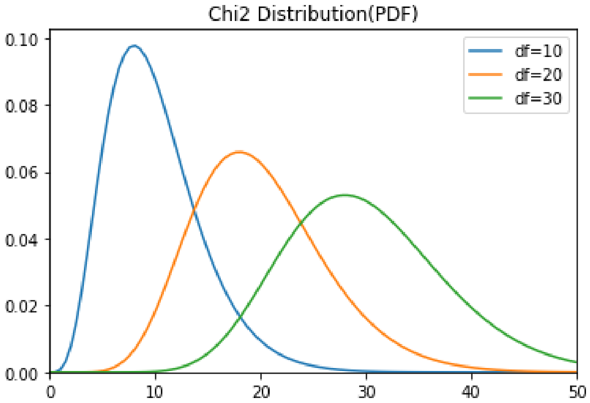
Analyst of Quantitative Finance

卡方分布

- 若 n 个相互独立的随机变量 $X_1, X_2, \dots, X_n$ 均服从标准正态分布，则：

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

服从自由度为 n 的卡方分布，记为 $Y \sim \chi^2(n)$



性质

- 右偏（正偏）
- 期望和方差：$E(\chi^2) = n$，$D(\chi^2) = 2n$
- 随着自由度的增加，卡方分布接近正态分布
- 可加性：设 $\chi_1^2 \sim \chi^2(n_1)$，$\chi_2^2 \sim \chi^2(n_2)$，并且 χ_1^2，χ_2^2 相互独立，则有 $\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$

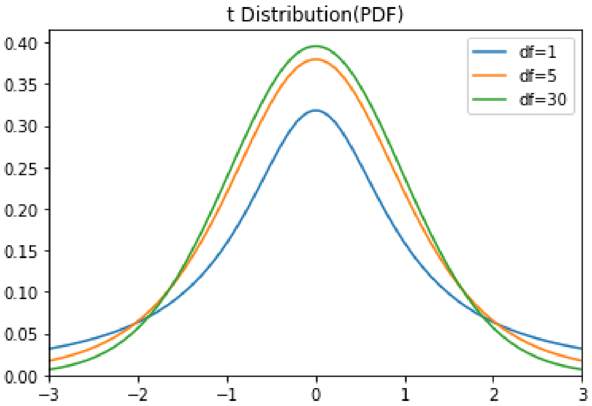
抽样分布

t 分布

- 设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，且 X，Y 相互独立，则称随机变量

$$T = \frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布，记为 $T \sim t(n)$



性质

- 对称，肥尾
- 期望和方差： $E(X) = 0$，$D(X) = \frac{n}{n-2}$
- 随着自由度的增加，t 分布接近标准正态分布

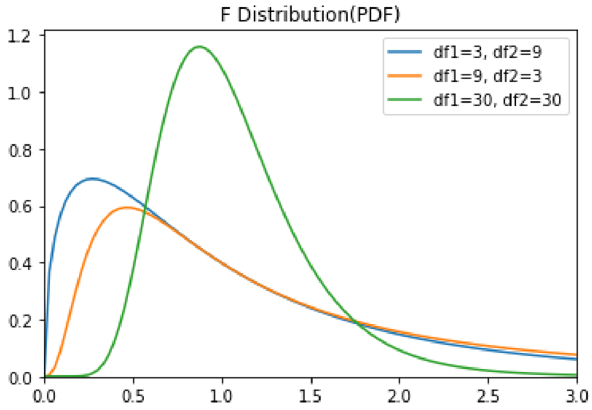
抽样分布

➢ F 分布

- 设 X、Y 为两个独立的随机变量，且 $X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，则：

$$F = \frac{X/m}{Y/n}$$

服从自由度 (m, n) 的 F 分布，记为 $F \sim F(m, n)$



➢ 性质

- 右偏（正偏）
- 如果随机变量 $T \sim t(n)$，则 $T^2 \sim F(1,\ n)$

抽样分布定理

- **定理**1：设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，$\overline{X}, S^2$ 分别是样本均值和样本方差，则有：

$$Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

  - $\overline{X}, S^2$ 相互独立

- **定理**2：设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本，$\overline{X}, S^2$ 分别是样本均值和样本方差，则有：

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

- **定理**3：设 $X_1, X_2, \dots, X_n$ 与 $Y_1, Y_2, \dots, Y_n$ 分别是来自正态总体 $N(\mu_1, \sigma_1^2)$ 与 $N(\mu_2, \sigma_2^2)$ 的样本，且这两个样本相互独立。设 $\overline{X}, \overline{Y}, S_1^2, S_2^2$ 分别表示这两个样本的样本均值与样本方差，则有：

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

点估计与区间估计

Analyst of Quantitative Finance

点估计

$$\begin{tikzcd}
\text{总体} \arrow[r, dashed] \arrow[d, "\text{抽样}"'] & \text{总体参数}\quad \mu,\ \sigma^2 \\
\text{样本} \arrow[r] & \text{样本统计量}\quad \bar{X},\ S^2 \arrow[u, "\text{估计}"']
\end{tikzcd}$$

区间估计

- 显著性水平：α；置信度：$1 - \alpha$
- 当置信度为95%时，我们可以说，置信区间包含真实总体参数的概率为95%
- 置信区间 = [点估计 ± 关键值 × 标准误]

- **估计量的评价标准**
  - 无偏性：估计量的数学期望等于需要估计的总体参数值

$$E[\hat{\theta}] = \theta$$

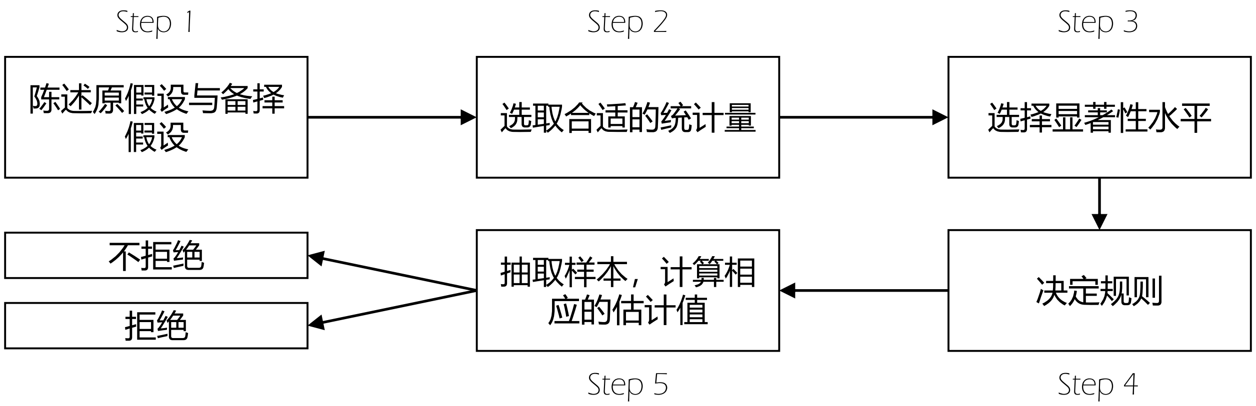
  - 有效性：对于同一个参数的多个无偏估计量中方差最小
  - 一致性：随着样本量的增大，该估计量越接近总体参数真实值

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

- **最优线性无偏估计（BLUE）**
  - 如果一个参数的估计量是样本观察值的线性函数，且具有无偏性和有效性，那么这个估计量就被称为是最优线性无偏估计量。

假设检验

➢ **假设检验的步骤**



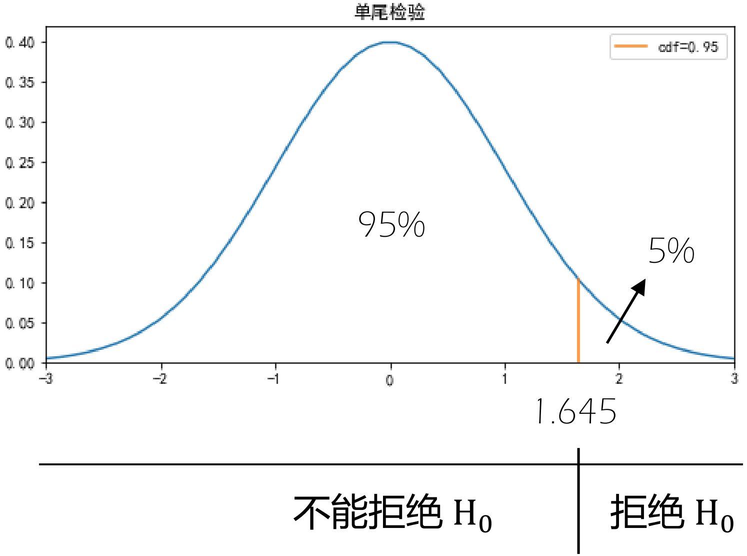
➢ **单尾检验** vs. **双尾检验**

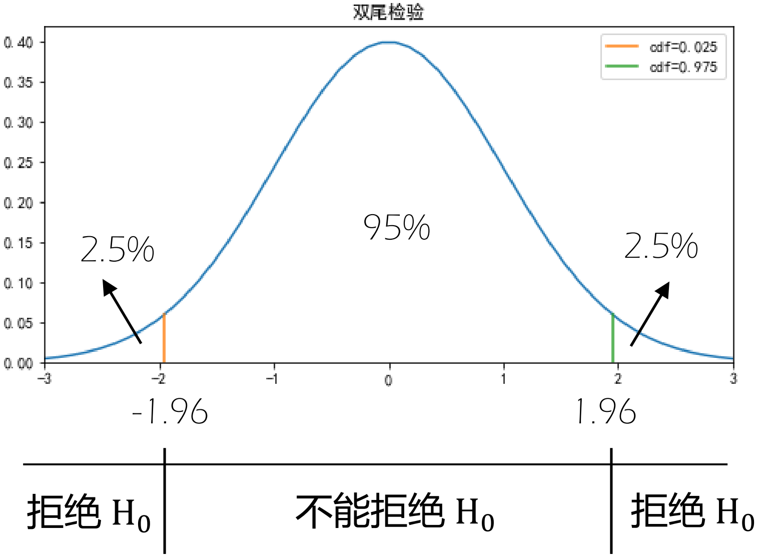
- 单尾检验：H_0: $\mu = \mu_0$ H_a: $\mu > \mu_0$（或 H_a: $\mu < \mu_0$）
- 双尾检验：H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$

决定规则

➢ 拒绝域

- 单尾检验：如果 样本估计值 > 关键值，则拒绝原假设 H_0
- 双尾检验：如果 |样本估计值| > 关键值，则拒绝原假设 H_0



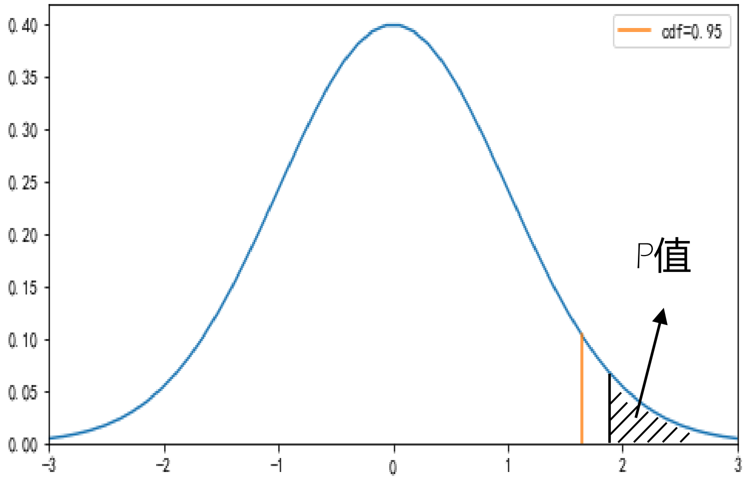


决定规则

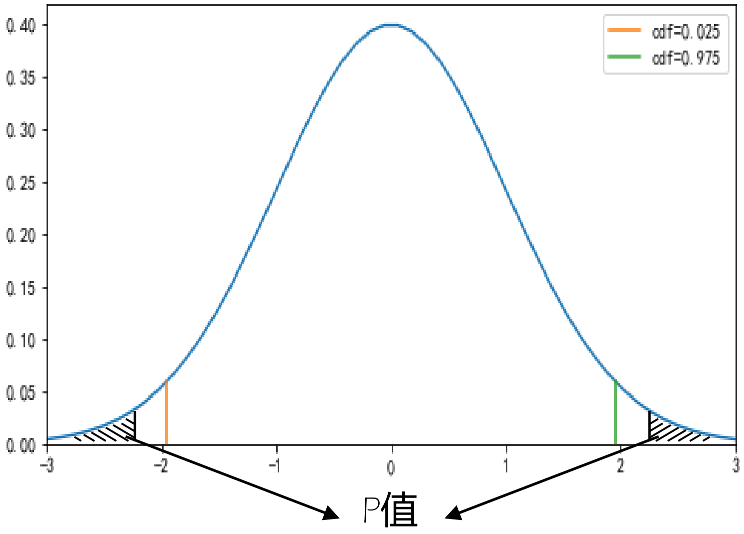
➢ P值

- $p - \text{value} < \alpha$，则拒绝原假设

单尾检验



双尾检验



总体均值的假设检验

➢ $\mathbf{H_0}$: $\boldsymbol{\mu} = \boldsymbol{\mu_0}$

	正态总体, $n < 30$	$n \geq 30$
方差已知 (σ^2)	Z 检验	Z 检验
方差未知	t 检验	t 检验 或 Z 检验

- **Z 检验:** $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

- **t 检验:** $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$

总体方差的假设检验

- $\mathbf{H_0}$: $\boldsymbol{\sigma}^2 = \boldsymbol{\sigma_0}^2$
  - **卡方检验**

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

- $\mathbf{H_0}$: $\boldsymbol{\sigma_1}^2 = \boldsymbol{\sigma_2}^2$
  - **F 检验**

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1)$$

  - 通常将较大的方差放在分子的位置，即：$S_1^2 > S_2^2$
  - 无论是单尾还是双尾，拒绝域通常都在右尾

假设检验汇总

➢ **常见检验类型**

检验类型	假设	H_0	统计量	临界值
均值检验	正态分布总体 总体方差已知	$\mu = \mu_0$	$Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$
	正态分布总体 总体方差未知	$\mu = \mu_0$	$t = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$t(n-1)$
方差检验	正态分布总体	$\sigma^2 = {\sigma_0}^2$	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$	$\chi^2(n-1)$
	两个独立的 正态分布总体	${\sigma_1}^2 = {\sigma_2}^2$	$F = \dfrac{S_1^2}{S_2^2}$	$F(n_1-1, n_2-1)$

第一类错误与第二类错误

AQF Analyst of Quantitative Finance

- **第一类错误：**原假设正确，但被拒绝（**拒真**）
- **第二类错误：**原假设错误，但却没有拒绝原假设（**存伪**）
- **显著性水平** (α)：第一类错误发生的概率
- **检验的势：**当原假设错误时，正确拒绝原假设的概率

<table>
<tr><th rowspan="2">决策</th><th colspan="2">真实情况</th></tr>
<tr><th>H_0 正确</th><th>H_0 错误</th></tr>
<tr><td>不拒绝 H_0</td><td>决策正确</td><td>第二类错误</td></tr>
<tr><td>拒绝 H_0</td><td>第一类错误
P(Type I error) = α</td><td>决策正确
Power of the test
= 1 − P(Type II error)</td></tr>
</table>

CONTENTS

简单线性回归

➢ 总体回归函数

$$E(Y|X_i) = \beta_0 + \beta_1 X_i$$

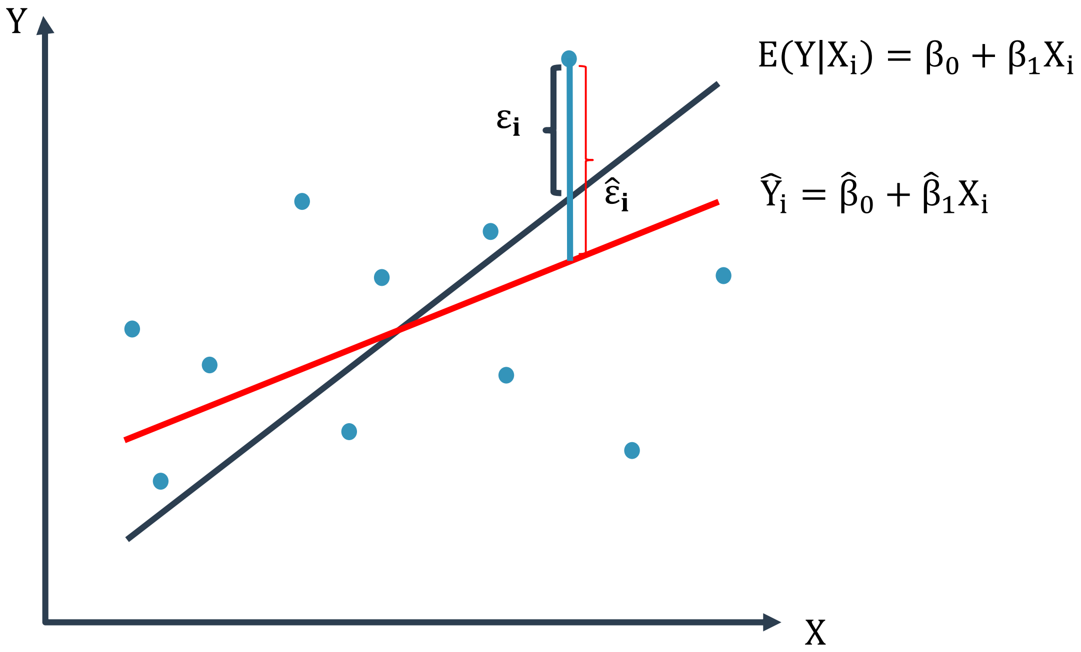
- 对于任何一个观测点，有：$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

➢ 样本回归函数

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- $\hat{Y}$ 为 $E(Y|X_i)$ 的估计量
- $\hat{\beta}_0$、$\hat{\beta}_1$ 为 β_0、β_1 的估计量
- 对于任何一个观测点，有：$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$

➢ **总体回归线和样本回归线**



普通最小二乘法

- **普通最小二乘法（OLS）**
  - 使样本回归方程中残差项的平方最小，即：

$$\min_{\hat{\beta}_0,\hat{\beta}_1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2$$

  - 求解可得：

$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}$$

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\frac{Cov(X,Y)}{Var(X)}$$

- Gauss-Markov **定理：**在线性回归模型中，如果随机误差项期望为零，方差相等，且互不相关，则OLS估计量是最佳线性无偏估计（BLUE）。

简单线性回归模型假设

- **简单线性回归模型（SLR）假设：**
  - 线性回归模型：参数线性
  - 自变量 X 非随机
  - 随机误差项的期望为零：$E(\varepsilon_i) = 0, i = 1,2,\ldots,n$
  - 随机误差项的方差相同：$Var(\varepsilon_i) = \sigma_\varepsilon^2, i = 1,2,\ldots,n$
  - 随机误差项之间彼此不相关：$Cov(\varepsilon_i, \varepsilon_j) = 0$，$i \neq j$
  - 自变量 X 和随机误差项 ε 不相关：$Cov(X_i, \varepsilon_i) = 0$
  - 随机误差项服从正态分布：$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

回归系数显著性检验

回归系数显著性检验

- **原假设与备择假设**

$$H_0:\ \beta_1 = 0 \quad H_a: \beta_1 \neq 0$$

- **Z 检验 – 方差已知**

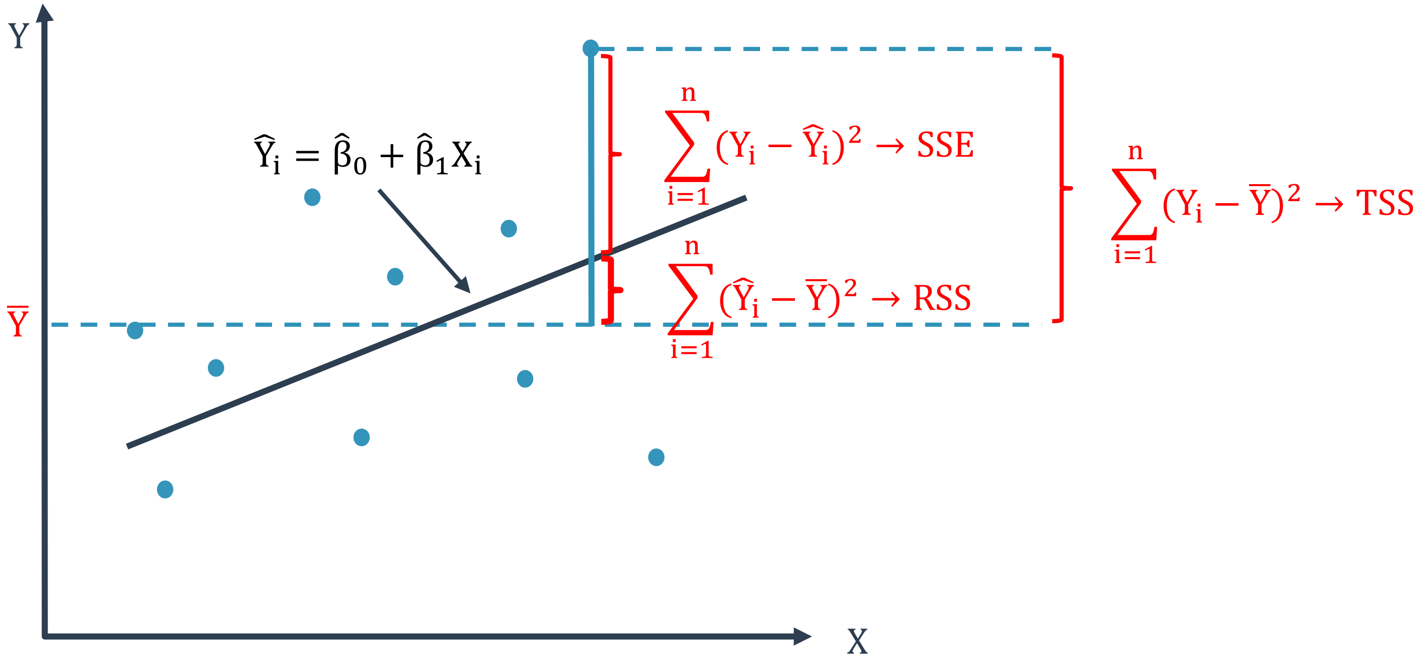
$$Z = \frac{\hat{\beta}_1 - 0}{SD(\hat{\beta}_1)} \sim N(0,1)$$

- **t 检验 – 方差未知**

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

- 也可以检验回归系数是否等于某个特定假设的值

样本回归方程



Y
$\bar{Y}$
X
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \rightarrow SSE$
$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \rightarrow RSS$
$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \rightarrow TSS$

方差分析

AQF Analyst of Quantitative Finance

- **ANOVA表**

	df	SS	MSS
回归	1	RSS	RSS/1
残差	$n-2$	SSE	$SSE/(n-2)$
总值	$n-1$	TSS	–

- **TSS、RSS、SSE**

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2$$

- Total sum of squares (TSS) = Sum of squares total (SST)
- Regression sum of squares (RSS) = Explained sum of squares (ESS)
- Sum of squared errors (SSE) = Sum of squared residual (SSR)

拟合优度

➢ R^2（决定系数）

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS}$$

- 表示因变量的变化有多少是由自变量解释的
- 取值范围为 $0 \le R^2 \le 1$
- 在一元线性回归中：$R^2 = r^2$

➢ SER（回归标准误）

$$SER = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}}$$

- SER衡量的是，真实的Y值偏离回归线的程度
- SER越小，拟合度越好

Python 实现线性回归

➢ **例：**工资水平与工作年限的关系

OLS Regression Results

Dep. Variable:	Salary	**R-squared:**	0.957
Model:	OLS	**Adj. R-squared:**	0.955
Method:	Least Squares	**F-statistic:**	622.5
Date:	Tue, 23 Jun 2020	**Prob (F-statistic):**	1.14e-20
Time:	13:16:09	**Log-Likelihood:**	-301.44
No. Observations:	30	**AIC:**	606.9
Df Residuals:	28	**BIC:**	609.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.579e+04	2273.053	11.347	0.000	2.11e+04	3.04e+04
YearsExperience	9449.9623	378.755	24.950	0.000	8674.119	1.02e+04

Omnibus:	2.140	**Durbin-Watson:**	1.648
Prob(Omnibus):	0.343	**Jarque-Bera (JB):**	1.569
Skew:	0.363	**Prob(JB):**	0.456
Kurtosis:	2.147	**Cond. No.**	13.2

多元线性回归模型假设

➢ **多元线性回归模型（MLR）假设：**

- 线性回归模型：参数线性
- 自变量 X 非随机
- 随机误差项的期望为零：$E(\varepsilon_i) = 0, i = 1,2,\ldots,n$
- 随机误差项的方差相同：$Var(\varepsilon_i) = \sigma_\varepsilon^2, i = 1,2,\ldots,n$
- 随机误差项之间彼此不相关：$Cov(\varepsilon_i,\varepsilon_j) = 0$，$i \neq j$
- 自变量 X 和随机误差项 ε 不相关：$Cov(X_i,\varepsilon_i) = 0$
- 随机误差项服从正态分布：$\varepsilon_i \sim N(0,\sigma_\varepsilon^2)$
- **自变量之间不存在完全的线性相关，即完全共线性**

回归系数显著性检验

➢ 单个回归系数显著性检验

- 原假设与备择假设

$$H_0\text{: } \beta_j = 0, j = 1,2, \dots, k$$

- Z 检验

$$Z = \frac{\hat{\beta}_j - 0}{SD(\hat{\beta}_j)} \sim N(0,1)$$

- t 检验

$$T = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{n-k-1}$$

回归系数显著性检验

➢ 联合假设检验

- 原假设与备择假设

$$H_0:\ \beta_1 = \beta_2 = \cdots = \beta_k = 0, j = 1,2,\dots,k$$

$$H_a: \text{至少一个} \beta_j \neq 0, j = 1,2,\dots,k$$

- F 检验

$$F = \frac{RSS/k}{SSE/(n-k-1)}$$

- 决策规则：拒绝 H_0，如果 F（检验统计量）>Fc（关键值）

拟合优度

> **调整后的 R^2**

- 在多元线性回归中，R^2会随着自变量的加入而增大，甚至新加入的变量并不满足统计上的显著性检验

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS}$$

- 调整后的 R^2 不一定随着自变量的加入而变大

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{TSS/(n-1)}$$

- 调整后的 $R^2 \leq R^2$
- 调整后的 R^2也许会小于 0

多重共线性

- **多重共线性**
  - 多元线性回归模型中的自变量之间存在高度线性相关关系

- **多重共线性对统计推断的影响**
  - 多重共线性不影响 OLS 估计量的一致性
  - OLS 估计量的标准误会被高估，t 检验失效
  - 很难区分各自变量对因变量的影响

- **诊断多重共线性**
  - 现实中，我们常常关注的是多重共线性的程度，而非它是否存在
  - 诊断多重共线性最常用的方法是：回归模型的 R^2很高，但是斜率系数的 t 统计量都不显著

Thank you!