

统计整理

杨弘毅

创建：2020 年 4 月 9 日

修改：2021 年 11 月 5 日

目录

1	TODO	3
2	基础定义	4
2.1	期望	4
2.2	方差	4
2.3	协方差	5
2.4	相关系数	6
3	矩	6
3.1	理解	6
3.2	定义	7
3.3	分类	8
3.4	矩母函数	11
3.4.1	定义	11
3.4.2	性质	12
4	假设检验	13
4.1	整体思想	13
4.2	参数与非参数检验	14
4.3	标准差与标准误	14
4.4	Z 检验	15

4.5	T 检验	15
4.5.1	单样本检验	16
4.5.2	两独立样本	17
4.5.3	一元回归系数检验	17
4.5.4	多元线性回归系数	18
4.6	F 检验	18
4.6.1	方差齐次 F 检验	19
4.6.2	方差分析	20
4.6.3	线性回归整体显著性检验	23
4.7	χ^2 检验	24
4.7.1	皮尔逊卡方检验	25
5	条件概率	28
5.1	条件概率	28
5.2	条件概率分布	30
5.2.1	链式法则	31
5.2.2	条件期望	32
5.2.3	条件方差	32
5.3	贝叶斯定理	34
5.4	贝叶斯因子	36
5.4.1	后验因子估计	36
5.4.2	发生率与贝叶斯因子	37
6	似然	38
6.1	似然与概率	38
6.2	似然函数	39
6.3	似然比	39
6.4	最大似然估计	39
7	回归模型	40
7.1	回归模型分类	40
7.2	最小二乘估计	41
8	时间序列	42

8.1	平稳性	42
8.2	自相关函数	42
8.2.1	相关系数	42
8.2.2	自相关函数	43
8.3	白噪声与线性时间序列	45
8.3.1	白噪声	45
8.3.2	线性时间序列	45
8.4	自回归模型 (AR)	46
8.5	残差自相关检验	51
8.6	移动平均模型 (MA)	52
8.7	自回归移动平均模型 (ARMA)	52
8.8	自回归移动平均模型 (ARIMA)	52
8.9	单位根检验	53
8.10	一致的协方差矩阵估计	54
8.11	格兰杰因果关系	55
9	条件异方差模型	55
9.1	ARCH 效应检验	55
9.2	ARCH 模型	57
9.2.1	ARCH 模型的性质	58
9.3	GARCH 模型	58
10	卡尔曼滤波	58

1 TODO

- degree of freedom
- 最大似然法 (条件似然法), 精确似然法
- likelihood, log-likelihood, goodness-of-fit, quasi-maximum likelihood, ratio test
- Newey West 1987

2 基础定义

2.1 期望

对于随机变量 X ，其概率空间为 (Ω, \mathcal{F}, P) ，期望值 $\mathbb{E}[X]$ 或 μ ，应有：

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) dP(\omega)$$

在离散以及连续情形下有如下定义，其中 $f(x)$ 为变量 X 的概率密度函数（PDF）。

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^n x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n \\ \mathbb{E}[X] &= \int x f(x) dx\end{aligned}$$

其性质有：

$$\begin{aligned}\mathbb{E}[X \pm Y] &= \mathbb{E}[X] \pm \mathbb{E}[Y] \\ \mathbb{E}[aX] &= a\mathbb{E}[X] \\ \mathbb{E}[XY] &= \mathbb{E}[X]\mathbb{E}[Y] \quad (X, Y \text{ are independent})\end{aligned}$$

2.2 方差

对于方差（Variance）或 σ^2 ，定义有：

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mu)^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \text{Cov}(X, X)\end{aligned}$$

同理，或其连续积分形式有：

$$\begin{aligned}\mathrm{Var}(X) &= \int (X - \mu)^2 f(x) dx \\ &= \int x^2 f(x) dx - 2\mu \int x f(x) dx + \mu^2 \int f(x) dx \\ &= \int x^2 f(x) dx - \mu^2\end{aligned}$$

其性质有：

$$\begin{aligned}\mathrm{Var}(X + a) &= \mathrm{Var}(X) \\ \mathrm{Var}(aX) &= a^2 \mathrm{Var}(X) \\ \mathrm{Var}(aX \pm bY) &= a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) \pm 2ab \mathrm{Cov}(X, Y) \\ \mathrm{Var}\left(\sum_{i=1}^N X_i\right) &= \sum_{i,j=1}^N \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^N \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \\ \mathrm{Var}\left(\sum_{i=1}^N a_i X_i\right) &= \sum_{i,j=1}^N a_i a_j \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^N a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^N a_i^2 \mathrm{Var}(X_i) + 2 \sum_{1 \leq i < j \leq N} a_i a_j \mathrm{Cov}(X_i, X_j)\end{aligned}$$

2.3 协方差

对于协方差（Covariance）其定义有：

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - E(X))(Y - E(Y))] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

性质有：

$$\text{Cov}(X, a) = 0$$

$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

$$\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$$

$$\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$$

$$\begin{aligned} \text{Cov}(aX + bY, cW + dV) &= ac \text{Cov}(X, W) + ad \text{Cov}(X, V) \\ &\quad + bc \text{Cov}(Y, W) + bd \text{Cov}(Y, V) \end{aligned}$$

2.4 相关系数

相关系数（Correlation Coefficient），为研究变量间线性相关程度的量。最早由统计学家卡尔·皮尔逊设计，也称为皮尔逊积矩相关系数（Pearson product-moment correlation coefficient），或皮尔逊相关系数：

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sigma_X \sigma_Y}$$

3 矩

3.1 理解

在物理学中，矩（Moment）源于阿基米德的杠杆原理，可简单认为是物理量与参照点距离的乘积，如力与力臂（参考点的距离）的乘积，得到的是力矩（或扭矩）。如一杆“秤”，“秤”的平衡的两边重量与距离的乘积相同，则能保持平衡。

具体而言， n 阶矩 μ_n 为物理量 Q 与某参考点 x 的 n 次方的乘积，即 $\mu_n = x^n Q$ 。常见的物理量如力或电荷等，若物理量并非集中在单点上，矩就应该是在物理量在空间上的积分，因有： $\mu_u = \int x^n f(x) dx$ ，其中 $f(x)$ 为物理量的密度分布函数。

而物理中的矩与数学中的矩概念相通，而在概率论上，如一端秤砣重量为中奖金额 500 元，中奖概率为百分之一，即离中心点距离为 0.01，那么其期望应为 5 元。可以理解为了使得秤保持平衡，则另一端，在距离中心距离为 1，对应其秤砣重量中奖金额应为 5 元。

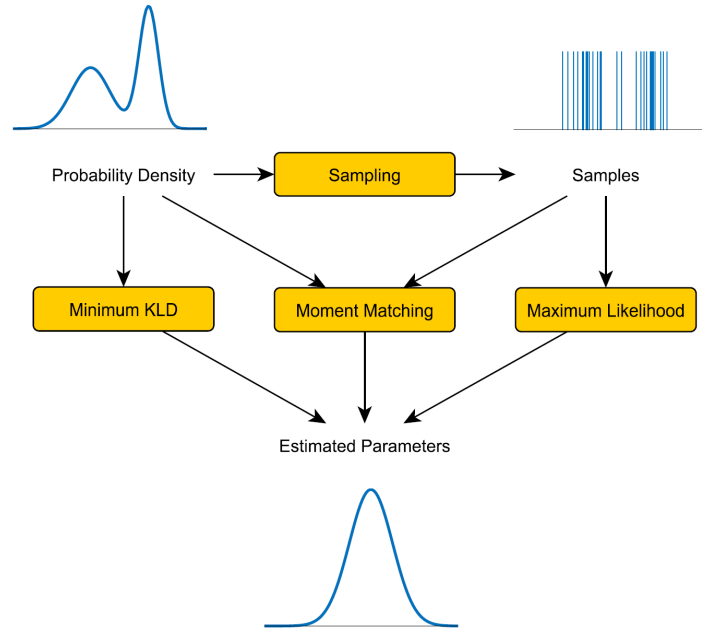


图 1: 矩匹配

3.2 定义

根据上述理解，物理学中与数学中的矩概念相通，即距离（概率）乘以物理量（随机变量）的大小。 p_i 为概率质量函数（Probability mass function, PMF），则对于 n 阶矩的离散形式有：

$$\mathbb{E}[x^n] = \sum_i x_i^n p_i$$

在连续形式下， n 阶矩可以表示为 $(x - c)^n$ 的期望，其中 $f(x)$ 为概率密度函数（Probability density function, PDF），其中 c 为均值。当 c 为 0 时，即称为中心矩（Central moment）。相反，则称为非中心矩，或原始矩（Raw moment）：

$$\mathbb{E}[x^n] = \mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx$$

除了根据 c 是否为零，根据是否进行标准化处理，可细分为标准矩。常用的矩有：

- 均值 $\text{Mean}(X) = \mathbb{E}(X)$ 为一阶非中心矩
- 方差 $\text{Variance}(X) = \mathbb{E}(X - \mu)^2$ 为二阶中心矩

- 偏度 $\text{Skewness}(X) = \frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3}$ 为三阶标准矩
- 峰度 $\text{Kurtosis}(X) = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$ 为四阶标准矩

3.3 分类

根据如上定义，从零阶至四阶的原始矩与中心矩有如下定义，其中定义 $\sigma = (\mathbb{E}[(X - \mu)^2])^{\frac{1}{2}}$ 。正态分布由它的前两阶矩决定，而对于其他的分布，需要了解更高阶矩。同时注意到三阶矩以上都称标准矩，如同方差要去除均值的影响，偏度和峰度也要去除方差的影响。

阶	原始矩	中心矩	标准矩
0	$\mathbb{E}(x^0) = 1$	$\mathbb{E}[(X - \mu)^0] = 1$	$\frac{\mathbb{E}[(X-\mu)^0]}{\sigma^0} = 1$
1	$\mathbb{E}(x^1) = \mu$ (均值)	$\mathbb{E}[(X - \mu)^1] = 0$	$\frac{\mathbb{E}[(X-\mu)^1]}{\sigma^1} = 0$
2	$\mathbb{E}(x^2)$	$\mathbb{E}[(X - \mu)^2] = \sigma^2$ (方差)	$\frac{\mathbb{E}[(X-\mu)^2]}{\sigma^2} = 1$
3	$\mathbb{E}(x^3)$	$\mathbb{E}[(X - \mu)^3]$	$\frac{\mathbb{E}[(X-\mu)^3]}{\sigma^3}$ (偏度)
4	$\mathbb{E}(x^4)$	$\mathbb{E}[(X - \mu)^4]$	$\frac{\mathbb{E}[(X-\mu)^4]}{\sigma^4}$ (峰度)

原始矩 (Raw/crude moment)

当 $c = 0$ 时，称为原始矩。此时则有平均数 (mean) 或期望 (expected value) 的连续形式为：

$$\mathbb{E}(X) = \mu = \int_{-\infty}^{\infty} (x - 0)^1 f(x) dx = \int_{-\infty}^{\infty} x f(x) dx$$

其离散形式为：

$$\mu = \mathbb{E}(X) = \sum_i x_i p_i$$

中心矩 (Central moment)

期望值可以成为随机变量的中心，即当 $c = \mathbb{E}(X)$ 时

$$\mu_n = \mathbb{E}[(x - \mathbb{E}(X))^n] = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^n f(x) dx$$

同时可知任何变量的一阶中心矩为 0:

$$\begin{aligned}
 \mu_1 &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^1 f(x) dx \\
 &= \int_{-\infty}^{\infty} x f(x) dx - \int_{-\infty}^{\infty} \mathbb{E}(X) f(x) dx \\
 &= \mathbb{E}(X) - \mathbb{E}(X) \int_{-\infty}^{\infty} f(x) dx \\
 &= \mathbb{E}(X) - \mathbb{E}(X) \times 1 = 0
 \end{aligned}$$

而二阶中心矩 (second central moment) 为**方差 (Variance)**

$$\begin{aligned}
 \mu_2 &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx \\
 &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mathbb{E}(X) \int_{-\infty}^{\infty} x f(x) dx + \mathbb{E}^2(X) \int_{-\infty}^{\infty} f(x) dx \\
 &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}^2(X) \times 1 \\
 &= \int_{-\infty}^{\infty} x^2 f(x) dx - \mathbb{E}^2(X) \\
 &= \mathbb{E}(X^2) - \mathbb{E}^2(X) = \sigma^2
 \end{aligned}$$

其离散形式则有:

$$\text{Var}(X) = \sigma^2 = \sum p_i (x_i - \mathbb{E}(X))^2$$

标准矩 (Standardized moment)

标准矩为标准化 (除以标准差) 后的中心矩, 第 n 阶中心矩 (standardized moment of degree n) 有:

$$\mu_n = \mathbb{E}[(X - \mu)^n] = \int_{-\infty}^{\infty} (x - \mu)^n f(x) dx$$

已知标准差的 n 次方有:

$$\sigma^n = \left(\sqrt{\mathbb{E}[(X - \mu)^2]} \right)^n = \left(\mathbb{E}[(X - \mu)^2] \right)^{\frac{n}{2}}$$

此时, 第 n 阶标准矩有:

$$\tilde{\mu}_n = \frac{\mu_n}{\sigma^n} = \frac{\mathbb{E}[(X - \mu)^n]}{\sigma^n}$$

由一阶中心矩为 0, 可知一阶标准矩 (first standardized moment) 也为 0。而二阶标准矩 (second standardized moment) 则有:

$$\tilde{\mu}_2 = \frac{\mu_2}{\sigma^2} = \frac{\mathbb{E}[(X - \mu)^2]}{(\mathbb{E}[(X - \mu)^2])^{2/2}} = 1$$

偏度 (skewness)

三阶标准矩 (third standardized moment) 为**偏度**:

$$\tilde{\mu}_3 = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}[(X - \mu)^3]}{(\mathbb{E}[(X - \mu)^2])^{3/2}}$$

偏度分为两种:

- 负偏态或左偏态: 左侧的尾部更长, 分布的主体集中在右侧
- 正偏态或右偏态: 右侧的尾部更长, 分布的主体集中在左侧

峰度 (kurtosis)

四阶标准矩 (third standardized moment) 为**峰度**:

$$\tilde{\mu}_4 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X - \mu)^4]}{(\mathbb{E}[(X - \mu)^2])^{4/2}}$$

由于正态分布的峰度 $K(X) = 3$, 因此定义**超额峰度 (Excess kurtosis)**为峰度 $K(X) - 3$, 那么就有正态分布的超额峰度为 0:

$$\text{Excess kurtosis} = \tilde{\mu}_4 - 3$$

- 若超额峰度为正, 称为高狭峰 (Leptokurtic), 此时有尖峰厚尾。即相比正态分布, 其“质量”更集中于中心
- 若超额峰度为负, 称为低阔峰 (Platykurtic), 此时有低峰轻尾。相比于正态分布, 其“质量”在中心位置更分散

3.4 矩母函数

3.4.1 定义

矩母函数或称为矩生成函数 (Moment generating function, MGF) 或动差生成函数, 顾名思义就是产生矩的函数。对于随机变量 X , 其矩生成函数定义为:

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

离散形式下有:

$$\mathbb{E}[e^{tX}] = \sum e^{tx} p(x)$$

而在连续形势下有:

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

定理 3.1. 将矩母函数进行 n 次求导, 并令 $t = 0$ 则可得到 $\mathbb{E}(X^n)$

$$\mathbb{E}(X^n) = \left. \frac{d^n}{dt^n} M_X(t) \right|_{t=0}$$

证明. 对于 e^x 使用泰勒展开有:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!}$$

那么 e^{tx} 的期望为:

$$\begin{aligned} \mathbb{E}[e^{tX}] &= \mathbb{E} \left[1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots + \frac{(tx)^n}{n!} \right] \\ &= \mathbb{E}(1) + t\mathbb{E}(x) + \frac{t^2}{2!}\mathbb{E}(x^2) + \frac{t^3}{3!}\mathbb{E}(x^3) + \cdots + \frac{t^n}{n!}\mathbb{E}(x^n) \end{aligned}$$

对其求一阶导:

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[e^{tX}] &= \frac{d}{dt} \left[\mathbb{E}(1) + t\mathbb{E}(x) + \frac{t^2}{2!}\mathbb{E}(x^2) + \frac{t^3}{3!}\mathbb{E}(x^3) + \cdots + \frac{t^n}{n!}\mathbb{E}(x^n) \right] \\ &= 0 + \mathbb{E}(x) + t\mathbb{E}(x^2) + \frac{t^2}{2}\mathbb{E}(x^3) + \cdots + \frac{t^{n-1}}{(n-1)!}\mathbb{E}(x^n) \\ &\quad (\text{代入 } t = 0) \\ &= 0 + \mathbb{E}(x) + 0 + 0 + \cdots + 0 \\ &= \mathbb{E}(x) \end{aligned}$$

□

3.4.2 性质

对于标准正态分布 $N \sim (0, 1)$ 的矩母函数，则有：

$$\begin{aligned}
 M_X(t) &= \mathbb{E}(e^{tX}) = \int e^{tx} \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}x^2} dx \\
 &= \int \frac{1}{\sqrt{\pi}} e^{tx - \frac{1}{2}x^2} dx \\
 &= \int \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}(x^2 - 2xt + t^2 - t^2)} dx \\
 &= \int \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}(x-t)^2 + \frac{1}{2}t^2} dx \\
 &= e^{\frac{1}{2}t^2} \int \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}(x-t)^2} dx \\
 &= e^{\frac{1}{2}t^2}
 \end{aligned}$$

对于正态分布 $N \sim (\mu, \sigma)$ 的矩母函数，则有：

$$M_X(t) = \mathbb{E}(e^{xt}) = \int e^{xt} \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

此时代换 $z = \frac{x-\mu}{\sigma}$ ，即 $x = \sigma z + \mu$ ，并有 $dx = \sigma dz$ ：

$$\begin{aligned}
 M_X(t) &= \int e^{(\sigma z + \mu)t} \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}z^2} dx \\
 &= e^{\mu t} \int e^{\sigma z t} \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}z^2} dx \\
 &= e^{\mu t} \int \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}(z^2 - 2\sigma t z + (\sigma t)^2 - (\sigma t)^2)} dx \\
 &= e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} \int \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}(z - \sigma t)^2} dx \\
 &= e^{\mu t + \frac{1}{2}\sigma^2 t^2}
 \end{aligned}$$

4 假设检验

4.1 整体思想

在假设检验 (Statistical hypothesis testing) 中, **原假设** (Null hypothesis, H_0), 也称为零假设或虚无假设。而与原假设相反的假设称为**备择假设** (Alternative hypothesis, H_a)。假设检验的核心为**反证法**。在数学中, 由于不能穷举所有可能性, 因此无法通过举例的方式证明一个命题的正确性。但是可以通过举一个反例, 来证明命题的错误。在掷骰子的例子中, 在每次掷的过程相当于一次举例, 假设进行了上万次的实验, 即便实验结果均值为 3.5, 也无法证明总体的均值为 3.5, 因为无法穷举。

可以理解为原假设为希望拒绝的假设, 或反证法中希望推翻的命题。我们先构造一个小概率事件作为原假设 (H_0), 并假设其正确。如样本均值等于某值, 两个样本均值是否相等, 样本中的不同组直接是否等概率发生, 一般使用等式 (小概率) 作为原假设。如果抽样检验中小概率事件发生, 则说明原假设的正确性值得怀疑。如此时假设实验的结果 (样本) 远大于或小于理论计算结果 3.5, 即发生了小概率事件, 那么就有理由相信举出了一个反例, 这时就可以否定原命题 (reject the null hypothesis)。而相反, 如果原假设认为均值为 3.5, 在实验的过程中结果大概率不会偏离这个理论值太多, 可以认为我们没办法举出反例。由于不能直接证明原命题为真, 只能说 “We can not(fail to) reject the null hypothesis”, 无法拒绝原命题。

在需要评估总体数据的时候, 由于经常无法统计全部数据, 需要从总体中抽出一部分样本进行评估。假设掷骰子一个骰子, 其期望为 3.5, 但假设掷骰子了 100 次, 计算均值为 3.47, 由于总体的理论值和样本呢的实验值可能存在偏差, 误差永远存在, 无法避免。那么是否可以认为么 3.47 “等于” 3.5? 这时候就需要要界定一个**显著水平** (Significant level, α), 相当于设定一个等于的阈值范围。即多小概率的事情发生, 是 10% 还是 5% 的概率, 使我们认为举出了一个反例, 值得去怀疑原命题的正确性。当我们知道随机变量的分布时候, 根据所进行的检验, 我们可以根据计算出的**统计量** (Test statistic), 由于分布已知, 统计量对应了一个 **p 值** (p-value), 即小概率 (极端) 事件发生的概率, 因此在图形上表示为统计量向两侧延申的线下区域。如果这个概率足够低, 如小于 $\alpha = 5\%$, 那么就有理由拒绝原假设。

用 1-显著水平 ($1 - \alpha$), 得到值称为**置信水平** (Confidence level) (概率大小)。置信水平越大, 对应的置信区间也越大 (随机变量范围)。此时有置信水平为 $1 - \alpha$, 假设置信区间为 (a, b) , 对于随机变量 x 有 $P(a < x < b) = 1 - \alpha$ 。对于双侧检验, 有置信水平为 $1 - \alpha$ (概率大小), 两侧拒

绝域分别为 $\alpha/2$ 。对于单侧检验，则有单侧拒绝域大小为 α 。

4.2 参数与非参数检验

参数检验 (Parametric test) 指总体分布服从正态分布或总体分布已知的情况下的统计检验，如 z 检验、 t 检验、方差分析 (ANOVA) 等。单因素方差分析 (One-way ANOVA) 是检验由单一因素影响的两组样本均值是否存在显著差异，同时有双因素方差分析 (Two-way ANOVA) 或多因素方差分析 (Multi-way ANOVA)，单因素与多因素是对于自变量（或称为因子）而言的，两者都是检验对于单因变量或单变量 (Univariate) 的影响。如上所述，若因变量只有一个，称为单变量方差分析。而多变量方差分析 (Multivariate analysis of variance, MANOVA)，指当因变量为两个或两个以上的情形。

非参数检验 (Nonparametric test) 指总体分布不要求服从正态分布或总体分布情况不明时（未知分布、样本太少），用来检验数据是否来自同一个总体的统计检验方法。如卡方检验 (Chi-squared test, χ^2 test)，其中最著名的为皮尔逊卡方检验 (Pearson's χ^2)，其他的卡方检验还有如费希尔精确检验 (Fisher's exact test) 二项检验 (Binomial test)。其他的非参数检验还有曼-惠特尼 U 检验 (Mann-Whitney U-test)、K-S 检验 (Kolmogorov-Smirnov test)、K-W 检验 (Kruskal-Wallis test) 等等。

4.3 标准差与标准误

标准差与标准误都衡量的是离散程度。标准差的对象为一次抽样里样本个体（样本量为 n ）间的离散程度。而标准误的对象，为从同一总体多次抽样得到的多个样本（每个样本量为 n ），每个的样本的某种统计量（均值、标准差、中位数，分位数等等），为该统计量的标准差。因此当统计量为样本均值时，此时样本均值标准误 (Standard error of the mean, SEM) 衡量的为样本均值的离散程度。

假设 x_1, x_2, \dots, x_n 为 n 个独立观察值，并且有总体均值为 μ ，总体标准差为 σ ，并令 $T = x_1 + x_2 + \dots + x_n$ ，样本均值为 $\bar{x} = T/n$ ，此时有：

$$\text{Var}(T) = \text{Var}(x_1) + \text{Var}(x_2) + \dots + \text{Var}(x_n) = n\sigma^2$$

对于样本均值的方差有：

$$\text{Var}(\bar{x}) = \text{Var}\left(\frac{T}{n}\right) = \frac{1}{n * 2} \text{Var}(T) = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

即样本均值服从 $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ 的正态分布。此时样本均值标准误，具体而言即样本均值的标准差（Standard deviation of the sample means）为：

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

对于样本均值的标准误可以用如上简单公式计算得到，当样本量太小、正态分布假设不满足，或没有简单公式可以计算标准误时，可以采用自助法（Bootstrapping）重抽样的方法模拟出一个抽样分布。具体而言，对于一个样本量为 N 的样本，重复进行多次有放回的随机抽样，每次抽样时，样本量也均为 N 。注意此时为有放回的抽样，因此样本中的个体可能被抽样多次。对每次抽样都计算所关注的统计量（如均值），从而可以直接获得一个关于该统计量的抽样分布。

4.4 Z 检验

当总体均值与标准差已知，或样本量大于 30 时，此时的样本标准差 $\hat{\sigma}$ 将接近总体标准差 σ 。Z 检验（Z-test）是一种参数检验，用于检验样本与总体均值是否有差异，z 统计量（z-statistic）服从 z-分布或标准正态分布（查表用 z-table），其中分母为样本标准误（Standard error）为 σ/\sqrt{n} ， $\hat{\sigma}$ 为样本标准差。因此有：

$$z\text{-statistic} = \frac{\bar{X} - \mu}{SE(\bar{X})} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

4.5 T 检验

T 检验（或称 Student's t test）是一种参数检验，即假设样本服从或近似服从正态分布。用于样本较小（样本量小于 30）或总体标准差未知。有如下常见的四个用途：

- 单样本均值检验（One-sample t-test）：用于检验单样本的均值是否与已知的总体均值相等
- 回归系数的显著性检验（t-test for regression coefficient significance）：用于检验回归模型的解释变量对被解释变量是否有显著影响

- 两独立样本均值检验 (Independent two-sample t-test) : 用于检验两对独立的正态或近似正态的样本的均值是否相等, 这里可根据总体方差是否相等分类讨论
- 配对样本均值检验 (Dependent t-test for paired samples) : 用于检验一对配对样本均值的差是否等于某一个值

4.5.1 单样本检验

对于检验单样本样本均值与总体均值是否有差异 (One-sample t-test), t 统计量 (t-statistics) 服从 t-分布 (查表用 t-table, 自由度为 $n - 1$, n 为样本数量), 基于样本标准差。t-分布描述的是标准化 (Standardized) 的样本均值 \bar{X} 与总体均值 μ 的距离, 即 t 统计量描述了此时样本均值距离总体均值多少个标准差。

$$t\text{-statistic} = \frac{\bar{X} - \mu}{SE(\bar{X})} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

如下所示为 t-分布的概率密度函数 (PDF), 其中 $\nu = n - 1$ 为自由度, Γ 为 Gamma 函数 (Gamma function):

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

而 z 分布与 t 分布的差异可见图2, 并且根据自由度不同而不同, 以下图为例, 可以看到当自由度变大, 当样本量 n 变大时 t 分布不断接近标准正态分布。z 分布与 t 分布两者都有均值为 0。

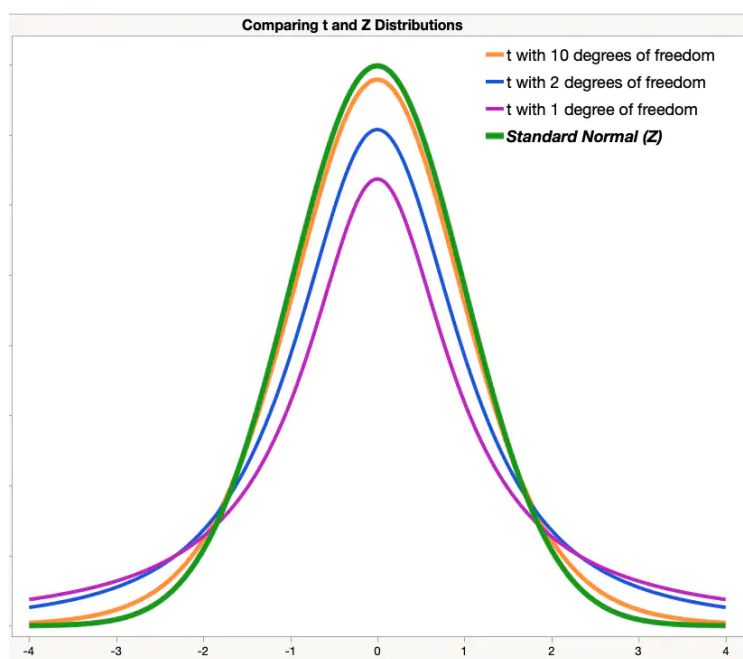


图 2: z 分布与 t 分布的差异

4.5.2 两独立样本

检验两个正态总体均值是否相等 (Independent two-sample t-test)

【待补充】

4.5.3 一元回归系数检验

假设有如下线性回归模型, x 已知, α 与 β 为待估参数, 且 ϵ 服从均值为 0, 方差未知假设为 σ^2 的正态分布:

$$y = \alpha + \beta x + \epsilon$$

此时检验回归系数 β 是否等于某 β_0 , 一般假设 $\beta_0 = 0$, 即原假设为 X 与 Y 两者不存在线性关系。t-统计量 (或 t-score) 服从自由度为 $n - 2$ 的 t 分布:

$$t\text{-statistic} = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} \sim T_{n-2}$$

其中有 $\hat{\beta}$ 的标准误为：

$$SE_{\hat{\beta}} = \frac{\sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

对于残差，有：

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$$

由于残差平方和（Sum of squared residuals）为 $SSR = \sum_{i=1}^n \epsilon_i^2$ ，可改写 t-统计量为：

$$t\text{-statistic} = \frac{(\hat{\beta} - \beta_0)\sqrt{n-2}}{\sqrt{\frac{SSR}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

4.5.4 多元线性回归系数

检验在多元线性回归中， x_1, x_2, \dots, x_m 是否存在线性关系：

$$H_0 : \beta_i = 0, \quad i = 1, 2, \dots, m$$

【待补充】

4.6 F 检验

F 检验（F-test），即检验统计量（Test statistic）在原假设下服从 F 分布（F-distribution）

主要的作用有如下三大类，而方差分析有可以分为许多类，如按照因素区分可分为单因素和多因素：

- 方差齐性 F 检验（F-test of equality of variances）
- 方差分析（Analysis of Variance, ANOVA）
- 线性回归整体显著性检验（F-test of overall significance in regression）

具体而言， S_1 与 S_2 为独立随机变量，且服从 χ^2 分布（多个独立服从标准正态分布随机变量的平方和），那么自由度为 ν_1 与 ν_2 的服从 F 分布的随机变量 F 为：

$$F = \frac{S_1/\nu_1}{S_2/\nu_2}$$

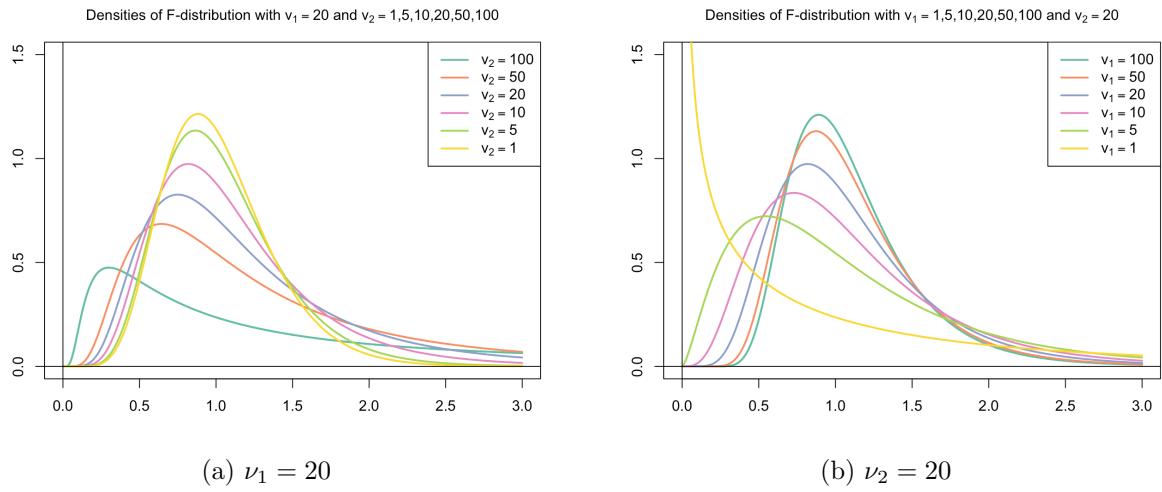


图 3: F 分布 PDF

F 分布的概率密度函数为，注意有 $x \geq 0$ ，因为其为两个方差的比率：

$$F\text{-distribution} = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right) \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2} - 1}}{\Gamma\left(\frac{\nu_1}{2}\right) \Gamma\left(\frac{\nu_2}{2}\right) \left(1 + \frac{\nu_1}{\nu_2}x\right)^{\frac{\nu_1 + \nu_2}{2}}} \quad \text{for } x \geq 0$$

n 为样本数， K 为变量数目，F 统计量服从以自由度为 ν_1 与 ν_2 的 F 分布（查表用 F-table）：

$$F\text{-statistic} \sim F_{\nu_1, \nu_2}$$

F 分布概率密度函数（PDF）如下图所述，为右偏（Right skew）或正偏（Postive skew）。当 $\nu_2 > 2$ 时，期望有 $\frac{\nu_2}{\nu_2 - 2}$ 。

4.6.1 方差齐次 F 检验

检验两个正态总体之间的方差是否相等，原假设为两个正态分布总体的方差相等。假设 X 与 Y 来自两个独立的服从正态分布的整体，即有 $X \sim N(\mu_X^2, \sigma_X^2)$ 和 $Y \sim N(\mu_Y^2, \sigma_Y^2)$ ，从中抽取 m 与 n 个样本，即 X_1, X_2, \dots, X_m 和 Y_1, Y_2, \dots, Y_n 。

因此两者均值和标准差为：

$$\begin{aligned}\bar{X} &= \frac{1}{m} \sum_{i=1}^m X_i & S_X^2 &= \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2 \\ \bar{Y} &= \frac{1}{n} \sum_{i=1}^n Y_i & S_Y^2 &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2\end{aligned}$$

此时双侧的原假设为假设两总体方差相等，备择假设为假设两总体方差不等：

$$H_0: \sigma_X^2 = \sigma_Y^2$$

$$H_a: \sigma_X^2 \neq \sigma_Y^2$$

此时 F 统计量应有，一般约定以较大的方差作为分子，较小的方差作为分母，使得计算出来的 $F > 1$ 方便进行查表。假设显著性水平为 α ，此时临界值为 $F_{\alpha, (m-1, n-1)}$ ，当 $F > F_{\alpha, (m-1, n-1)}$ 时，拒绝原假设：

$$F = \frac{S_X^2}{S_Y^2} \sim F(m-1, n-1)$$

4.6.2 方差分析

方差分析 (Analysis of variance, ANOVA)，可以用于一次性检验多个（两个或两个以上）总体的均值是否相等，而 T 检验弱面对多个总体，只能两两进行比较。以单因素方差分析 (One-way ANOVA) 为例，比较不同组之间的均值是否相同。例如研究不同的猪饲料对猪的增重效果是否不同，此时猪饲料为（单）因素，即不同的配方为不同组（水平），随机取样，观察猪增重效果是否相同。又或研究客户满意程度在不同航空公司之间是否不同，此时航空公司为（单）因素，即不同的航空公司为不同组（水平），随机抽样，观察不同航空公司客户满意度的均值是否相同。

例子 4.1. 航空公司客户满意度 (0-10 分) 例子

甲航空	乙航空	丙航空	丁航空
8	3	7	8
6	7	8	9
7	5	9	6
5	9		7
9			

注意虽然 ANOVA 叫做方差分析，但是他的目的是检验每个组的均值是否相同。ANOVA 的原假设为 K 个总体的均值相等， $H_0: \mu_1 = \mu_2 = \dots = \mu_K$ 。可以换个角度考虑这个问题，如果几家公司的客户满意度没有显著的差别，那么可以认为是一个总体的四次随机抽样。相反，若有一家航空公司表现与众不同，那么这家航空公司不是来自同一个总体。ANOVA 主要有以下 3 个假设：

- 方差的同质性，可以理解为每组样本背后的总体的方差相同
- 总体遵循正态分布
- 每一次抽样都是独立的

组间偏差平方和（Sum of square between groups, SSB 或 SSA）与组内偏差平方和（Sum of square error, SSE）都是偏差的平方和，而两者计算的偏差个数不一样，无法直接相除，因此除以各自的自由度，使得两者在平均的意义下可比。因此有 $MSB = SSB/df_1$ 和 $MSE = SSE/df_2$ ，其中 $df_1 = K - 1$ 和 $df_2 = N - K$ 为两者自由度。

$$F = \frac{SSB/df_1}{SSE/df_2} = \frac{MSB}{MSE} \sim F(K - 1, N - K)$$

可以从图4中看到，组内均方（Mean squared error, MSE）对应着为每个样本分布自身的方差。组间均方（Mean squared between, MSB）对应着相当于样本对于总体的方差。可以想象，当两者之间相差越大时，不是来自同一总体（假设）的概率就越大

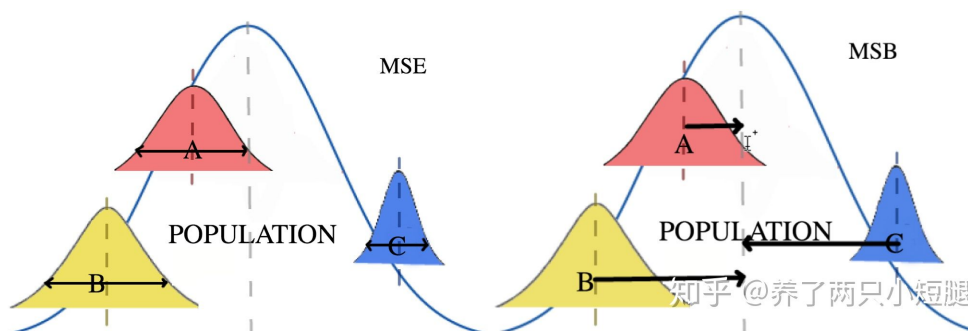


图 4: MSE 和 MSB

方差分析表如下，假设显著性水平为 α ，那么有当 $F > F_{\alpha, (K-1, N-K)}$ 时，可以认为这一次抽样几乎不可能，从而拒绝原假设，认为该因素显著：

偏差来源	偏差平方和	自由度	均方	F 值
组间	SSB	$K - 1$	$MSB = \frac{SSB}{K-1}$	$\frac{MSA}{MSE}$
组内	SSE	$N - K$	$MSE = \frac{SSE}{N-K}$	
总计	SST	$N - 1$		

具体而言，假设实验共有 K 个分类（水平），总样本量为 N ，每个分类下有 N_i 个体，并有 $N = \sum_{i=1}^K N_i$ ，令 $x_{i,j}$ 为第 i 个分类第 j 个样本。

第 i 组的组内均值为：

$$\bar{x}_{i,\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j}$$

总体均值为：

$$\bar{x} = \frac{1}{N} \sum_{i=1}^K \sum_{j=1}^{N_i} x_{i,j}$$

因此，组间偏差平方和应为，组均值与总体均值偏差的平方和。一共有 K 组，均值确定后，只有 $K - 1$ 可自由变化，因此 SSB 的自由度为 $K - 1$ ：

$$SSB = \sum_{i=1}^K \sum_{j=1}^{N_i} (\bar{x}_{i,\cdot} - \bar{x})^2 = \sum_{i=1}^K N_i (\bar{x}_{i,\cdot} - \bar{x})^2$$

组内偏差平方和应有，即个体与组内均值偏差的平方和。 N 个个体，每个组的平均值 $\bar{x}_{i,\cdot}$ 确定后，只有 $N - K$ 可自由变动，因此 SSE 的自由度为 $N - K$ ：

$$SSE = \sum_{i=1}^K \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_{i,\cdot})^2$$

那么总偏差平方和（Sum of square total, SST）为个体与总体平均值偏差的平方和。当 \bar{x} 确定后，有 $N - 1$ 可自由变换，因此 SST 自由度为 $N - 1$ ：

$$SST = \sum_{i=1}^K \sum_{j=1}^{N_i} (x_{i,j} - \bar{x})^2 = SSB + SSE$$

证明. 对于 SST 的分解有:

$$\begin{aligned}
 SST &= \sum_{i=1}^K \sum_{j=1}^{N_i} (x_{i,j} - \bar{x})^2 \\
 &= \sum_{i=1}^K \sum_{j=1}^{N_i} [(x_{i,j} - \bar{x}_{i,\cdot}) + (\bar{x}_{i,\cdot} - \bar{x})]^2 \\
 &= \sum_{i=1}^K \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_{i,\cdot})^2 + \sum_{i=1}^K \sum_{j=1}^{N_i} (\bar{x}_{i,\cdot} - \bar{x})^2 + 2 \sum_{i=1}^K \sum_{j=1}^{N_i} (\bar{x}_{i,\cdot} - \bar{x})(x_{i,j} - \bar{x}_{i,\cdot}) \\
 &= \sum_{i=1}^K \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_{i,\cdot})^2 + \sum_{i=1}^K \sum_{j=1}^{N_i} (\bar{x}_{i,\cdot} - \bar{x})^2 + 2 \sum_{i=1}^K \left[(\bar{x}_{i,\cdot} - \bar{x}) \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_{i,\cdot}) \right]
 \end{aligned}$$

由于:

$$\sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_{i,\cdot}) = \sum_{j=1}^{N_i} x_{i,j} - N_i \bar{x}_{i,\cdot} = 0$$

因此 SST 推导中最后一项为 0, 可以消去, 得到最终结果:

$$SST = SSB + SSE$$

□

4.6.3 线性回归整体显著性检验

F 检验还可以用于检验变量被解释变量 y 与解释变量 x_1, x_2, \dots, x_k , 两者之间的整体线性关系是否显著, 检验的是线性方程整体的显著性。假设有如下回归方程:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

若各回归系数 β_i ($i = 1, 2, \dots, k$) 均为 0, 则被解释变量 y 与解释变量 x_i ($i = 1, 2, \dots, k$) 没有任何关系, 回归模型变为 $y = \beta_0$, 只有截距项。反之, 若各回归系数 β_i 中仅有部分为 0, 则对于这部分回归系数不为 0 的 x_i 而言, 该回归模型还是有意义的。因此原假设与备择假设如下:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_a : \beta_i (i = 1, 2, \dots, k) \text{ 不全为 } 0$$

方差分析表如下，假设显著性水平为 α ，那么有当 $F > F_{\alpha, (p, n-p-1)}$ 时，可以认为这一次抽样几乎不可能，从而拒绝原假设，认为该因素显著：

偏差来源	偏差平方和	自由度	均方	F 值
回归	SSR	p	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
误差	SSE	$n - p - 1$	$MSE = \frac{SSE}{n-k-1}$	
总计	SST	$n - 1$		

4.7 χ^2 检验

卡方检验（Chi-squared test）为非参数检验，不存在具体参数和总体正态分布的假设。最初由统计天王 Karl Pearson 于 1900 年提出，是三大抽样（T 检验、F 检验与卡方检验）分布的检验里历史最悠久的。

卡方检验是比较数据的实测（实际观测值）分布与数的预期（理论预期值）分布的假设检验，检验两者之间的偏离程度，实际观测值与理论预期值之间的偏离程度就决定卡方的大小。如果卡方越大，二者偏差程度越大。相反，二者偏差越小。若两个值完全相等时，卡方值就为 0，表明理论值完全符合。常见的卡方检验有：

- 皮尔逊卡方检验（Pearson's chi squared test）
- 耶茨的连续性修正（Yates's correction for continuity）
- Fisher 确切概率法（Fisher's exact test）

假设有随机变量 x_1, x_2, \dots, x_k 为独立随机变量，服从 $N(0, 1)$ 的标准正态分布。其令平方和为 Q ，注意此为随机变量的平方和，并非分布的平方和。则有 Q 服从自由度为 $k > 0$ 的卡方分布，即：

$$Q = \sum_{i=1}^k x_i^2 \quad Q \sim \chi^2(k) \quad \text{or} \quad Q \sim \chi_k^2$$

对于自由度为 k 的卡方分布的概率密度函数（PDF）有：

$$f(x) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

卡方检验为单尾检验，拒绝域在右尾。当卡方分布的自由度 $k < 2$ 时，曲线呈现倒 L 形，当 $k > 2$ 时趋向于对称，而当 $k \rightarrow \infty$ 时，则趋向于正态分布。

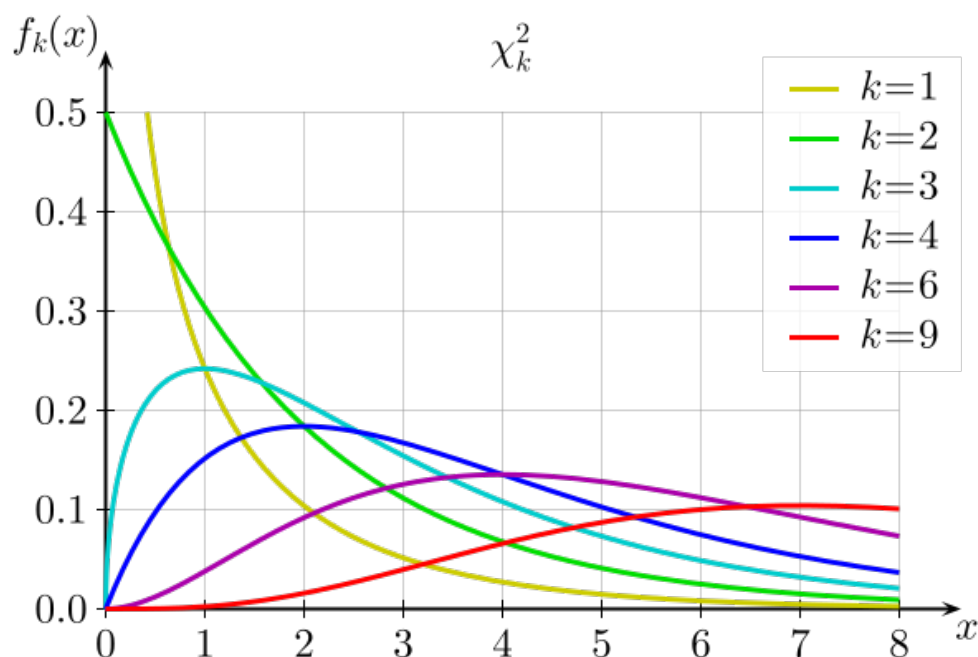


图 5: 卡方分布 PDF

4.7.1 皮尔逊卡方检验

皮尔逊卡方检验 (Pearson's chi-squared test)，当提及卡方检验，通常即指皮尔逊卡方检验。其中 O_i 指实际观察频数， E_i 指理论预期频数：

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

皮尔逊卡方检验有多种用途：

- 卡方拟合优度检验 (Chi-square goodness-of-fit tests)
 - 使用此分析检验分类数据样本与总体理论分布的拟合程度
 - 例如，通过多次掷骰子并使用卡方拟合优度检验来确定结果是否服从均匀分布，可以检验骰子是否公平。
- 卡方关系检验 (Chi-square tests for relationships)
 - 独立性检验 (Test of independence)：可以使用独立性检验确定一个变量的观测值是否取决于另一个变量的观测值。例如，确定总

统候选人党派是否与投票人的性别无关（同一人群样本，两个分类变量，性别与党派）。

– 同质性检验（Test of homogeneity）：可以使用同质性检验不同样本中同一个变量的分布是否相同。例如，确定不同颜色的汽车的销量是否取决于汽车售出的城市（随机选择城市不同样本，汽车颜色为单一变量）

卡方拟合优度检验

卡方拟合优度检验或称拟合度的卡方检验、最佳拟合度卡方检验，用于检验实际数据是否符合理论分布，即度量实际分布与理论分布的拟合优度（Goodness of fit）。对于离散情形下，假设随机抽样的总样本量为 n ，分为互斥的 k 类，每一类的个体数为 x_i , $i = 1, 2, \dots, k$ 为实际频数（个数），同时有 k 理论频数为，原假设为：

$$H_0 : \text{第 } k \text{ 类的占比为 } p_i \quad i = 1, 2, \dots, k$$

此时有卡方统计量，分子为实际频数与理论频数的平方。服从自由度为 $k - 1$ 的卡方分布：

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - np_i)^2}{np_i} \sim \chi^2(k - 1)$$

例子 4.2. 孟德尔提出遗传第二定律，他发现四种不同种类的豌豆种子：黄色圆形的、黄色褶皱的、绿色圆形的、绿色褶皱的。这些表现型类别（由可观察到的特征定义类别）以大约 9:3:3:1 的比例出现。假设实际频数为 315:108:101:32，总频数为 556。那对于理论频数应有：

$$556 \frac{9}{16} : 556 \frac{3}{16} : 556 \frac{3}{16} : 556 \frac{1}{16} = 312.75 : 104.25 : 104.25 : 34.75$$

因此计算卡方统计量有：

$$\chi^2 = \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.47$$

对于显著性水平 $\alpha = 0.05$, $\chi_{0.05,3}^2 = 7.815 > 0.47$ ，因此无法拒绝原假设，实验结果符合理论预期。

卡方独立性与同质性检验

卡方相关性与独立性的检验，检验的方式并无不同，主要差异在于实验的设计和如何解读实验结果。若将不同样本理解为分类变量，也可认为是独立性检验，可以理解为同质性检验为独立性检验的延伸，用于一些特殊情形，例如检验待检测产品的特性、小白鼠的体重等，产品和小白鼠并无实质的差异，当然可以设置为待测产品或小白鼠以不同批次划分，为随机抽取的结果，而非本质上的划分。具体而言，在独立检测中有两个根据其本质划分的分类变量，而在同质性检验中，其中一个变量是随机产生的（随机选取批次），不能视为一个分类变量，因此只能进行同质性检验。

	研究对象	研究目的
独立性	单一样本下的两个分类变量	同一个样本下，两个分类变量是否独立
同质性	多个样本下度量同一分类变量	不同的样本是否同质，即分类变量的分布是否相同

例子 4.3. 假设在赌场中，老板认为以为负责二十一点的庄家赔付的金额高于平均值，三位庄家观察的数据如下：

	庄家甲	庄家乙	庄家丙	合计
胜	43	49	22	114
平	8	2	5	15
负	47	44	30	121
合计	98	95	57	250

此时原假设为庄家和赌局结果没有关系，即两者独立，而备择假设为有关系，两者不独立。对于庄家甲胜场的情况下，假设庄家与独具相互独立，应有理论频率为：

$$\begin{aligned}
 P(\text{庄家甲胜场}) &= \text{总场数} \times P(\text{庄家甲参加比例}) \times P(\text{胜场比例}) \\
 &= 250 \times \frac{114}{250} \times \frac{98}{250}
 \end{aligned}$$

计算所有理论频数为：

	庄家 A	庄家 B	庄家 C
胜	$(114 \times 98)/250$	$(114 \times 95)/250$	$(114 \times 57)/250$
平	$(15 \times 98)/250$	$(15 \times 95)/250$	$(15 \times 57)/250$
负	$(121 \times 98)/250$	$(121 \times 95)/250$	$(121 \times 57)/250$

计算卡方统计量：

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i} = 5.62$$

对于显著性水平 $\alpha = 0.05$, $\chi_{0.05,4}^2 = 9.488 > 5.62$, 因此无法拒绝原假设, 赌局的胜负与庄家没有关系。

例子 4.4. 以泰坦尼克号的生还者与遇难者的数据为例, 计算可计算出生还的概率约为 32%, 因此头等舱乘客的生还的理论概率为 $32\% \times 325 \approx 104$, 继续计算所有理论频数。得到 $\chi^2 = 192.2$, 此时自由度有 $df = (4 - 1)(2 - 1) = 3$, 因此在显著水平 $\alpha = 0.05$ 下临界值 $\chi_{0.05,3}^2 = 7.815 < 192.2$, 因此拒绝四个类别的乘客生还率相同的原假设。

	头等舱	二等舱	三等舱	船员	合计
生还	202	118	178	215	713
遇难	123	167	528	698	1516
合计	325	285	796	913	2229

5 条件概率

5.1 条件概率

条件概率 (Conditional probability) 记为 $P(A | B)$, 指已知事件 B 发生的情况下事件 A 发生的概率, 即把原本的样本空间, 缩小为只有 B 发生的样本空间, 从中再计算 A 发生的概率。而联合概率分布 (Joint probability distribution) 或简称为联合分布, 表示两个事件同时发生的概率, 记为 $P(A, B)$ 或 $P(A \cap B)$, 其样本空间为原本未缩小的空间。

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

备注 5.1. 同理 $P(B|A) = \frac{P(A \cap B)}{P(A)}$, 即有 $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$ 。易知同时满足 AB 条件的概率, 即为已知满足 A 条件的子集中, 再满足条件 B 的概率 (为条件概率, 即其样本空间为缩小后满足条件 A 的空间)。或满足条件 B 的子集中, 再满足条件 A 的概率。

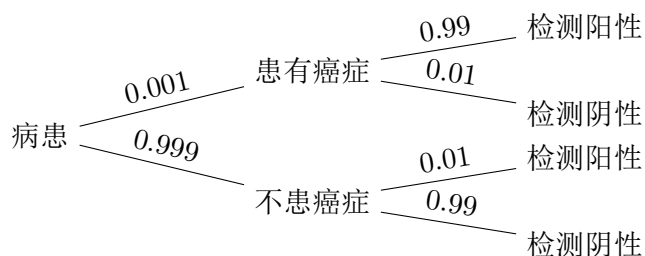
例子 5.2. 假设有两个小碗, 分别为甲碗与乙碗, 每个小碗中有若干小球, 共有蓝色与黄色两种颜色, 具体数量如下。

	蓝色	黄色
甲碗	1	4
乙碗	3	2

那么任意选取一个碗, 并从中选取一个小球, 颜色为蓝色的概率应为 $P(\text{蓝色}) = \frac{4}{10}$, 同时有 $P(\text{黄色}) = \frac{6}{10}$ 。若已知取的碗为甲碗, 那么选取蓝色小球的概率应为 $P(\text{蓝色} | \text{碗} = \text{甲碗}) = \frac{1}{5}$, 此时样本空间改变, 被限制在了甲碗中, 此时称为条件概率。

若此时任意从两个碗中选取一个小球, 小球颜色是蓝色, 那么此时该小球是从甲碗中选取的概率为 $P(\text{甲碗} | \text{颜色} = \text{蓝色}) = \frac{1}{4}$, 注意 $P(\text{甲碗} | \text{蓝色})$ 与 $P(\text{蓝色} | \text{甲碗})$ 的概率并不相同。

例子 5.3. 假设某病患被检测出了某癌症阳性, 该癌症在人群中的患病率 (*Prevalence*) 为 0.1%, 并且该诊断的正确率为 99%。



该病人检测为阳性, 因此有关的概率为 $P(\text{患有癌症} \cap \text{检测阳性})$ 与 $P(\text{不患癌症} \cap \text{检测阳性})$,

而该病人在检测为阳性的前提下，真正患有癌症的概率应为：

$$\begin{aligned}
 P(\text{患有癌症} | \text{检测阳性}) &= \frac{P(\text{患有癌症} \cap \text{检测阳性})}{P(\text{检测阳性})} \\
 &= \frac{P(\text{患有癌症} \cap \text{检测阳性})}{P(\text{患有癌症} \cap \text{检测阳性}) + P(\text{不患癌症} \cap \text{检测阳性})} \\
 &= \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.999 \times 0.01} \approx 9.016\%
 \end{aligned}$$

或使用表格表示：

	患有癌症（0.001）	不患癌症（0.999）
检测正确（0.99）	检测结果：阳性	检测结果：阴性
检测错误（0.01）	检测结果：阴性	检测结果：阳性

备注 5.4. 在医学中将 $P(\text{患病} | \text{检测阳性})$ 的概率称为阳性预测值 (*Positive Predictive Value*, *PPV*)，即在检测为阳性的前提下，有多大概率是真正患病。具体为检测阳性并真实患病的人数 (*True Positive*)，除以全体检测阳性的人数。全体检测阳性的人数中，除了正确诊断的病患之外，还包含检测阳性但并不患病的人数 (*False Positive*)。

$$PPV = \frac{TP}{TP + FP}$$

如上所述，在医学中敏感性 (*Sensitivity*) 或真阳性率，指在病患中检测结果为阳性的概率。同样在病患中，检测为阴性的概率称为假阴性率，或漏诊率。而特异性 (*Specificity*) 或真阴性率，指在健康人群检测为阴性的概率。同样，在健康人群中，检测出阳性的概率称为假阳性率，或误诊率。

5.2 条件概率分布

继续推广条件概率的概念，条件概率分布 (Conditional probability distribution)，在离散形势下有称为条件概率质量函数 (Conditional probability mass function)：

$$P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

与条件概率相同，有等价关系 $P(Y = y | X = x)P(X = x) = P[(X = x) \cap (Y = y)] = P(X = x | Y = y)P(Y = y)$ 。

例子 5.5. 有一个骰子，假设当掷出来的数字为偶数（如：2, 4, 6）时 $X = 1$ ，而奇数时 $X = 0$ 。同时假设当掷出来的数字为质数（如：2, 3, 5）时 $Y = 1$ ，其他情况下 $Y = 0$ 。

	1	2	3	4	5	6
X	0	1	0	1	0	1
Y	0	1	1	0	1	0

对于无条件概率 $P(X = 1) = \frac{1}{2}$ ，而条件概率 $P(X = 1 | Y = 1) = \frac{1}{3}$ ，具体计算如下：

$$\begin{aligned}
 P(X = 1 | Y = 1) &= \frac{P(X = 1)P(Y = 1 | X = 1)}{P(Y = 1)} \\
 &= \frac{P(X = 1)P(Y = 1 | X = 1)}{P(X = 1)P(Y = 1 | X = 1) + P(X = 0)P(Y = 1 | X = 0)} \\
 &= \frac{1}{1 + 2} = \frac{1}{3}
 \end{aligned}$$

对于连续情形下，有条件概率密度函数（Conditional probability density function）：

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

如上所述其中 $f_{X,Y}(x,y)$ 为联合分布（Joint probability distribution）。对于离散随机变量而言，称为联合分布概率质量函数（Joint probability mass function），对于连续随机变量也称为联合分布概率密度函数（joint probability density function）。而 $f_Y(y)$ 为边缘分布（Marginal distribution），同样分为边缘概率质量函数（Marginal probability mass function）与边缘概率密度函数（Marginal probability density function）。

5.2.1 链式法则

对于概率论，同时也有链式法则（Chain rule）或称为一般乘法法则（General product rule），提供了使用条件概率分布计算联合概率分布的方法，如上所述对于事件 A 与事件 B ：

$$P(A \cap B) = P(A) \cdot P(B | A)$$

对于多个事件 A_1, \dots, A_n 的联合分布, 其中有 A_0 为全集, 即 $P(A_1 | A_0) = P(A_1)$:

$$\begin{aligned} P(A_1 \cap \dots \cap A_n) &= P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_2, A_1) \dots P(A_n | A_{n-1}, \dots, A_1) \\ &= \prod_{i=1}^n P\left(A_i \left| \bigcap_{j=1}^{i-1} A_j \right.\right) \end{aligned}$$

如当 $n = 3$ 时, 应有:

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_3 | A_2, A_1) \cdot P(A_2, A_1) \\ &= P(A_3 | A_2, A_1) \cdot P(A_2 | A_1) \cdot P(A_1) \end{aligned}$$

5.2.2 条件期望

与条件概率原理相同, 条件期望 (Conditional expectation), 记为 $\mathbb{E}(X | Y)$ 或 $\mathbb{E}(X | Y = y)$, 即限制条件 $Y = y$ 缩小样本空间后, 计算 X 的期望。在离散的情形下有如下表达式, 其中 $P(X = x, Y = y)$ 为联合概率密度函数。

$$\mathbb{E}(X | Y = y) = \sum_x x P(X = x | Y = y) = \sum_x x \frac{P(X = x, Y = y)}{P(Y = y)}$$

例子 5.6. 在上述投掷骰子的例子, 对于无条件概率 $\mathbb{E}(X) = \frac{1}{2}$, 而条件期望 $\mathbb{E}(X | Y = 1) = \frac{1+0+0}{3} = \frac{1}{3}$, 条件期望 $\mathbb{E}(X | Y = 0) = \frac{0+1+1}{3} = \frac{2}{3}$

对于连续随机变量, 其中 $f_{X,Y}(x, y)$ 为 X 与 Y 的联合概率密度函数, $f_Y(y)$ 为 Y 的概率密度函数, 令 $f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$:

$$\mathbb{E}(X | Y = y) = \int_x x f_{X|Y}(x | y) dx = \frac{1}{f_Y(y)} \int x f_{X,Y}(x, y) dx$$

5.2.3 条件方差

条件方差 (Conditional variance) 定义为:

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

命题 5.7. 总期望定律 (*Law of total expectation*) 或称为双重期望定理 (*Double expectation theorem*) 有:

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

证明.

$$\begin{aligned} \mathbb{E}[\mathbb{E}(X | Y)] &= \sum_y \mathbb{E}(X | Y = y)P(Y = y) \\ &= \sum_y \left[\sum_x xP(X = x | Y = y) \right] P(Y = y) \\ &= \sum_y \sum_x xP(X = x | Y = y)P(Y = y) \\ &= \sum_y \sum_x xP(Y = y | X = x)P(X = x) \\ &= \sum_x \sum_y xP(Y = y | X = x)P(X = x) \\ &= \sum_x xP(X = x) \left[\sum_y P(Y = y | X = x) \right] \\ &= \sum_x xP(X = x) = \mathbb{E}(X) \end{aligned}$$

□

备注 5.8. 或可以先展开内层期望进行证明 $\mathbb{E}[\mathbb{E}(X | Y)] = \mathbb{E}[\sum_x xP(X = x | Y)]$ 。由此可以发现内层条件期望求得的结果为关于 Y 的函数, 因此外层的期望作用于随机变量 Y , 而内层的期望作用于随机变量 X 。积分先后顺序可以对调, 可以理解为一个矩形面积, 可以由积分底再积分高获得, 或先积分高再积分底, 两者结果相同。且倒数第二步, 可以将与随机变量 Y 无关的, 只关于随机变量 X 的部分提出至括弧外。

命题 5.9. 总方差定律 (*Law of total variance*) 有:

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

证明. 计算等式右边第一项, 并且根据总期望定律:

$$\begin{aligned} \mathbb{E}[\text{Var}(X | Y)] &= \mathbb{E}[\mathbb{E}(X^2 | Y) - (\mathbb{E}(X | Y))^2] \\ &= \mathbb{E}(X^2) - \mathbb{E}[(\mathbb{E}(X | Y))^2] \end{aligned}$$

对于等式右边第二项，根据方差的定义有：

$$\begin{aligned}
 \text{Var}[\mathbb{E}(X|Y)] &= \mathbb{E}[\mathbb{E}(X|Y) - \mathbb{E}[\mathbb{E}(X|Y)]]^2 \\
 &= \mathbb{E}[\mathbb{E}(X|Y) - \mathbb{E}(X)]^2 \\
 &= \mathbb{E}[\mathbb{E}(X|Y)^2 - 2\mathbb{E}(X|Y)\mathbb{E}(X) + \mathbb{E}^2(X)] \\
 &= \mathbb{E}[\mathbb{E}(X|Y)^2] - \mathbb{E}^2(X)
 \end{aligned}$$

将两项相加，再次根据方差定义 $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$ 替换，得证。 \square

5.3 贝叶斯定理

通常而言，事件 A 在给定事件 B 已发生的条件下发生的概率，与事件 B 在给定事件 A 已发生的条件下发生的概率是不一样的。然而这两者是有确定的关系的，贝叶斯定理 (Bayes' theorem) 就是这种关系的陈述，具体而言有：

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

$P(A)$ 为先验概率 (Prior probability)，即不考虑任何 B 方面的因素。而 $P(A|B)$ 为已知 B 发生后，A 发生的概率，也称为 A 的后验概率 (Posterior probability) 或似然 (Likelihood)。同理，所要求的条件概率 $P(B|A)$ 也称为 B 的后验概率，因为是在已知 A 发生的情况下，B 发生的概率。

已知 $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$ ，由此变形可推导贝叶斯定理，即两个条件概率之间的关系。**注意：**只有当 A 与 B 互相独立时，才有 $P(A \cap B) = P(A) \times P(B)$ 。由于此时 $P(B|A) = P(B)$ ，A 不受 B 的影响。

例子 5.10. 继续使用小球的例子，并计算相应概率有：

	蓝色 (0.4)	黄色 (0.6)
甲碗 (0.5)	1	4
乙碗 (0.5)	3	2

要计算 $P(\text{甲碗} | \text{蓝色})$ ，将样本空间限制在蓝色球的范围内，即应知道甲乙两个碗中蓝球的数量

目，最直观的计算方法为：

$$P(\text{甲碗} | \text{蓝色}) = \frac{P(\text{甲碗中蓝色球的数目})}{P(\text{甲碗中蓝色球的数目}) + P(\text{乙碗中蓝色球的数目})}$$

其中分子甲碗中的蓝色球数目，应有甲碗的概率再乘以，已知甲碗蓝色球的概率（即在甲碗这个缩小的样本空间内，蓝色球的概率），若使用较为严谨的数学语言表达，则有：

$$\begin{aligned} P(\text{甲碗} | \text{蓝色}) &= \frac{P(\text{甲碗}) \times P(\text{蓝色} | \text{甲碗})}{P(\text{甲碗}) \times P(\text{蓝色} | \text{甲碗}) + P(\text{乙碗}) \times P(\text{蓝色} | \text{乙碗})} \\ &= \frac{P(\text{甲碗}) \times P(\text{蓝色} | \text{甲碗})}{P(\text{蓝色})} \\ &= \frac{0.5 \times 0.2}{0.4} = \frac{1}{4} \end{aligned}$$

例子 5.11. 上述癌症诊断例子中，患病率为 0.1%，并且该诊断的正确率为 99%。即在病患中检测出阳性的概率为真阳性率或敏感率为 99%，而在健康人群中检测为阴性的概率同为 99%，或称为真阴性率，或特异性。根据贝叶斯定理有：

$$\begin{aligned} P(\text{患有癌症} | \text{检测阳性}) &= \frac{P(\text{患有癌症}) \times P(\text{检测阳性} | \text{患有癌症})}{P(\text{检测阳性})} \\ &= \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.999 \times 0.01} \approx 9.016\% \end{aligned}$$

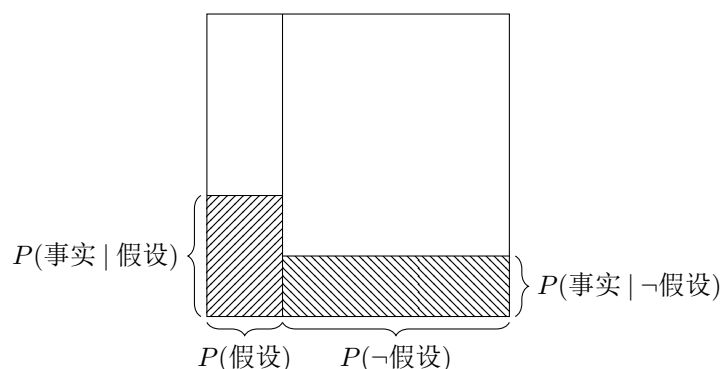
假设该病人进行了第二次同样的检测，此时再次使用贝叶斯定理，可以继续更新该病患真正患有癌症的概率，为了简便将条件事件（患有癌症 | 第一次阳性）记为（患癌首阳）其概率近似为 9% 代入计算。可以看到第二次检测阳性之后，实际患有癌症的几率大大提高。

$$\begin{aligned} P(\text{患癌首阳} | \text{第二次阳性}) &= \frac{P(\text{患癌首阳}) \times P(\text{第二次阳性} | \text{患癌首阳})}{P(\text{第二次阳性})} \\ &= \frac{P(\text{患癌首阳}) \times P(\text{第二次阳性} | \text{患癌首阳})}{P(\text{患癌首阳}) \times P(\text{第二次阳性} | \text{患癌首阳}) + P(\neg \text{患癌首阳}) \times P(\text{第二次阳性} | \neg \text{患癌首阳})} \\ &= \frac{0.09 \times 0.99}{0.09 \times 0.99 + 0.91 \times 0.01} \approx 90.73\% \end{aligned}$$

备注 5.12. 假设集合 A 为某假设或模型，而集合 B 为某事实或数据，那么有：

$$P(\text{假设} | \text{事实}) = \frac{P(\text{假设})P(\text{事实} | \text{假设})}{P(\text{假设})P(\text{事实} | \text{假设}) + P(\neg \text{假设})P(\text{事实} | \neg \text{假设})}$$

贝叶斯定理可以通过图形进行理解与记忆，假设如下为边长为 1 的正方形，此时正方形面积为 1 代表整体概率为 1。



假设 $P(\text{假设})$ 不变, 当 $P(\text{事实} | \text{假设}) = P(\text{事实} | \neg \text{假设})$ 时, 即上图两个长方体高度相同。此时易知 $P(\text{假设} | \text{事实}) = P(\text{假设})$ 。此时, 事实为真的概率, 在假设条件为真与假设条件为假的两个子集中, 概率相同, 因此其概率就应该等于全样本下的先验概率, 即后验概率不发生改变。由此可知, 当事实为真的概率, 在假设为真与假设为假的概率差别越大时, 后验概率的变化也越大。

5.4 贝叶斯因子

5.4.1 后验因子估计

关于贝叶斯定理, 可以理解为根据已事实对认知进行更新。如在上述癌症诊断的例子中, 原本的患病率为先验概率为 0.1%, 通过已知的事实, 原本的先验概率为后验概率, 上升约为 9%。而这个上升的概率可以使用贝叶斯因子 (Bayes factor) 进行估算, 是一种似然比 (Likelihood ratio)。对于贝叶斯定理有:

$$\text{贝叶斯因子} = \frac{P(\text{事实} | \text{假设})}{P(\text{事实} | \neg \text{假设})}$$

贝叶斯定理可以进行如下估算:

$$\begin{aligned} P(\text{假设} | \text{事实}) &= P(\text{假设}) \times \frac{P(\text{事实} | \text{假设})}{P(\text{事实})} \\ &= P(\text{假设}) \times \frac{P(\text{事实} | \text{假设})}{P(\text{假设})P(\text{事实} | \text{假设}) + P(\neg \text{假设})P(\text{事实} | \neg \text{假设})} \\ &\approx P(\text{假设}) \times \frac{P(\text{事实} | \text{假设})}{P(\text{事实} | \neg \text{假设})} \\ &= P(\text{假设}) \times \text{贝叶斯因子} \end{aligned}$$

例子 5.13. 在上述癌症诊断例子中，患病率为 0.1%，并且该诊断的正确率为 99%。此时使用贝叶斯因子对后验概率进行估计：

$$\frac{P(\text{检测阳性} | \text{患有癌症})}{P(\text{检测阳性} | \text{不患癌症})} = \frac{\text{True Positive Rate}}{\text{False Positive Rate}} = \frac{0.99}{0.01} = 99$$

因此 $P(\text{患有癌症} | \text{检测阳性}) \approx P(\text{患有癌症}) \times 99 = 0.1\% \times 99 = 9.9\%$ ，与实际计算结果 9.016% 接近。此方法只在于帮助理解，贝叶斯定理实际上是一个更新概率的过程，实际的更新因子已由贝叶斯定理给出。由上式对比贝叶斯定理与贝叶斯因子估计，可以看出，当 $P(\text{假设})$ 较小时，贝叶斯定理分母第一项趋近于 0，此时 $P(\neg \text{假设})$ 趋近于 1，使得两者较为接近。

5.4.2 发生率与贝叶斯因子

概率 (Probability) 与发生比 (Odds) 或俗称赔率的差别在于，就单一事件而言，概率中可包含多种结果，如上涨、不变、下跌，但发生率只能表示发生与不发生，两种结果。概率的分子为单一结果，但分母是全体事件为 1，而发生率分子与概率相同为单一结果，为发生，分母也为单一结果，即不发生。使得在发生率定义下，分母中不再包含 $P(\text{假设})P(\text{事实} | \text{假设})$ 项，使用贝叶斯因子能准确计算后验概率，而非估计，且形式更为简洁。

例子 5.14. 同样上述癌症诊断的例子中，患病率 0.1% 转化为发生比为 1 : 999，定义为先验发生比。此时贝叶斯因子的作用，是将分子患有癌症的人群中检测为阳性与分母中不患癌症中检测为阳性的人群分别挑选出来。因此结果就是检测为阳性的人群中，实际患有癌症的发生率。由上文计算贝叶斯因子为 99，那么后验发生比为：

$$\begin{aligned} O(\text{患有癌症} | \text{检测阳性}) &= \frac{\text{检测阳性且患有癌症的人数}}{\text{检测阳性但不患癌症的人数}} \\ &= \frac{P(\text{患有癌症}) \times P(\text{检测阳性} | \text{患有癌症})}{P(\text{不患癌症}) \times P(\text{检测阳性} | \text{不患癌症})} \\ &= O(\text{患有癌症}) \times \frac{P(\text{检测阳性} | \text{患有癌症})}{P(\text{检测阳性} | \text{不患癌症})} \\ &= 1 : 999 \times 99 = 99 : 999 \\ &\rightarrow \frac{99}{99 + 999} \approx 9.016\% \end{aligned}$$

6 似然

对于似然 (Likelihood) 的简单理解, 若已有数据 X , 想要猜测是来自分布 f 或分布 g 。那么将 X 的分布作图为直方图 (Histogram), 并同时对比两个分布, 看分布 f 或分布 g 更为接近 X 的直方图, 而似然就是对比的定量值。

对于数据集 $X = x_1, x_2, \dots, x_n$, 并假设变量 x_i 均为独立同分布, 即取值只与自己有关系, 而与别的 $x_j, j \neq i$ 无关。对于分布 $f(\cdot | \theta)$, 对于任意 x_i 其概率 $f(x_i | \theta)$ 若非常低, 则说明 x_i 出现在分布 $f(\cdot | \theta)$ 的概率非常低, 而反之则说明概率高。即 x_i 有多大的可能性或概率, 是来自分布 $f(\cdot | \theta)$, 这其中的可能性或概率, 即称之为似然。

6.1 似然与概率

如上所述, 似然和概率, 其本质都是可能性。对于由一组参数 θ 确定的分布, 概率是给定参数 θ , 求某一结果的可能性。而相反而言, 似然 (Likelihood) 是给定某一结果 X , 求参数 θ 的可能性。具体而言, 如对于只需要 μ 与 σ 就能确定的正态分布, 对于离散随机变量:

$$\mathcal{L}(\theta | x) = P(X = x | \theta)$$

对于连续随机变量:

$$\mathcal{L}(\theta | x) = f(x | \theta)$$

例子 6.1. 对于随机抽取实验室小白鼠, 并测量其体重 w 的试验中, 假设小白鼠的体重服从正态分布 $N \sim (\mu, \sigma)$, 此时概率为给定分布参数 θ 的情况下, 求小白鼠体重对应的概率:

$$P(32 < w < 34 | \mu = 32, \sigma = 2.5) = 0.29$$

$$P(w = 34 | \mu = 34, \sigma = 2.5) = 0.16$$

而对于似然, 则为给定数据, 求参数 θ 对应的概率:

$$L(\mu = 32, \sigma = 2.5 | w = 34) = 0.12$$

$$L(\mu = 34, \sigma = 2.5 | w = 34) = 0.21$$

在调整了分布的 μ 之后, 似然变大。显然对于单次的检验, 单一数据, 使得抽取一只小白鼠体重为 34g 的概率最大的分布, 为 $\mu = 34$ 的正态分布。

6.2 似然函数

似然函数 (Likelihood function)

$$\mathcal{L}(\theta | X) = \prod_{i=1}^n f(x_i | \theta)$$

6.3 似然比

似然比 (Likelihood ratio)

$$\Lambda(\theta_1, \theta_2 | x) = \frac{\mathcal{L}(\theta_1 | x)}{\mathcal{L}(\theta_2 | x)}$$

备注 6.2. 在 *Python* 中有如下模块:

- *Matlab: lratiotest* (*Likelihood ratio* 检验)

6.4 最大似然估计

最大似然估计 (Maximum likelihood estimation, MLE)

测量了数只老鼠的重量, 尝试找到其分布, maximizes the likelihood 找到最大化所有观察重量 likelihood 的分布, 找到 mean 和 standard deviation

得分检验 (Score test), score (informant) is the gradient of the log-likelihood function with respect to the parameter vector.

$$s(\theta) = \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}$$

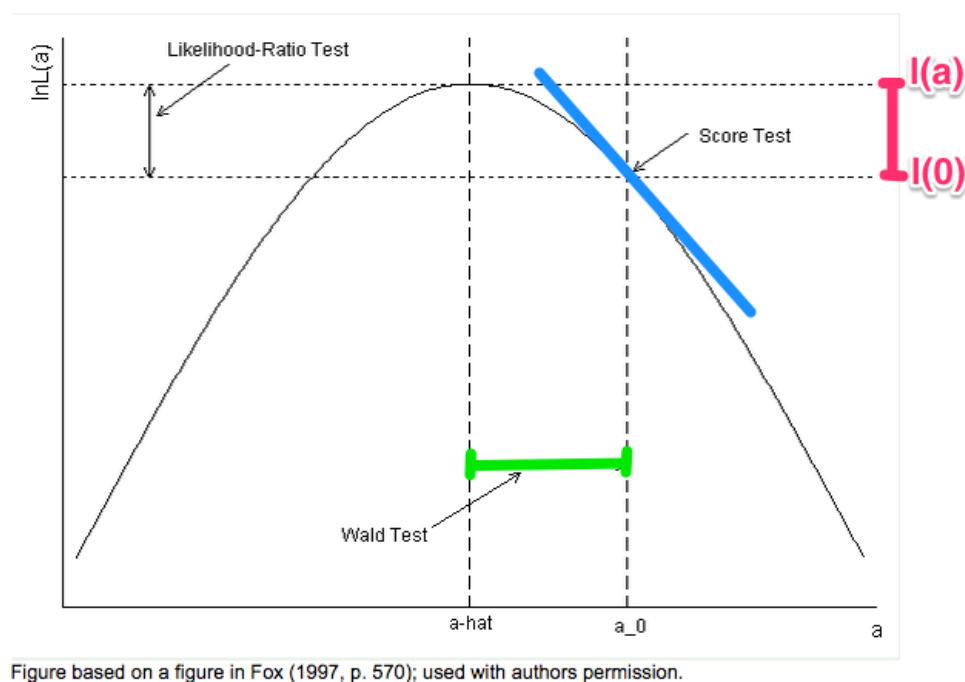


图 6: Likelihood/Wald/Score Test

7 回归模型

7.1 回归模型分类

线性回归模型

线性回归模型（Linear regression）包含有：

- 最小二乘法（Ordinary least squares）
- 广义最小二乘法（Generalized least squares）
- 加权最小二乘法（Weighted least squares）
- 广义线性模型（General linear model）

参数回归与半参数非参数回归

参数回归或参数模型 (Parametric regression), 如各类线性回归, 其回归函数形式为已知, 参数待定, 外延容易, 但形式呆板。而对于半参或非参回归 (Semi/Non-parametric regression), 为非线性回归, 其回归函数不确定, 外延困难, 但拟合效果却较好。其中非参数回归基本方法如:

- 最近邻函数法 (Nearest-neighbor 或 K-nearest neighbors)
- 核函数回归 (Kernel regression)
- 神经网络 (Neural networks)
- 支持向量机 (Support vector regression)
- 光滑样条 (Smoothing spline)

7.2 最小二乘估计

有 n 个观察值 $\{x_i, y_i\}_{i=1}^n$, 其线性回归模型有如下形式:

$$y_i = \beta_0 + x_{i,0} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

或改写为向量形式为如下所示, 其中 $\mathbf{x}_i = [x_{i,0}, x_{i,2}, \dots, x_{i,p-1}]^T$ 为 $p \times 1$ 的列向量, $\boldsymbol{\beta}$ 也是 $p \times 1$ 的列向量:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

若写成矩阵形式, 则有 $\mathbf{Y} = [y_1, \dots, y_n]^T$ 、 $\boldsymbol{\beta} = [\beta_1, \dots, \beta_n]^T$ 与 $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]^T$ 都为 $n \times 1$ 的列向量, \mathbf{X} 为 $n \times p$, 即 n 行 p 列的矩阵 ($x_{\cdot,0} = 1$):

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p-1} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p-1} \end{bmatrix}$$

此时矩阵形式有:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

【待补充】时间序列的估计, 使用 OLS, 也需要满足 OLS 的假设, 否则估计有偏

8 时间序列

8.1 平稳性

对于时间序列 $\{r_t\}$ ，**强平稳** (Strongly stationary) 或**严格平稳** (Strictly stationary)，是指对于任意 t 和任意正整数 k ，资产收益率时间序列 $(r_{t_1}, \dots, r_{t_k})$ 与 $(r_{t_1+t}, \dots, r_{t_k+t})$ 的联合分布相同，即将 (t_1, \dots, t_k) 平移 t 个单位之后，分布不变。而**弱平稳** (Weakly stationary) 指对于时间序列 $\{r_t\}$ ，当 r_t 的均值与 r_{t-l} 的协方差不随时间改变，即：

$$\begin{aligned}\mathbb{E}[r_t] &= \mu \\ \text{Cov}(r_t, r_{t-l}) &= \gamma_l\end{aligned}$$

μ 为常数，不随时间改变，而 γ_l 指依赖时间间距 l 的大小。平稳性的强弱区别，在于弱平稳只要求期望与方差不变，即前两阶矩不变，而强平稳要求各阶距都不变。由于正态分布只需要两阶矩确认，因此时间序列 $\{r_t\}$ 若是正态分布的，那么此时弱平稳与强平稳等价。

协方差 $\gamma_l = \text{Cov}(r_t, r_{t-l})$ 称为 r_t 的间隔为 l 的**自协方差** (Autocovariance)，其中有两个性质：

$$\begin{aligned}\gamma_0 &= \text{Var}(r_t) \\ \gamma_{-l} &= \gamma_l\end{aligned}$$

证明. 性质二：

$$\gamma_{-l} = \text{Cov}(r_t, r_{t-(-l)}) = \text{Cov}(r_{t-(-l)}, r_t) = \text{Cov}(r_{t+l}, r_t) = \text{Cov}(r_t, r_{t-l}) = \gamma_l$$

□

8.2 自相关函数

8.2.1 相关系数

根据定义，两个随机变量 X 与 Y 的相关系数定义为：

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\mathbb{E}[(X - \mu_x)(Y - \mu_y)]}{\sqrt{\mathbb{E}(X - \mu_x)^2 \mathbb{E}(Y - \mu_y)^2}}$$

对于样本 $\{(x_t, y_t)\}_{t=1}^T$, 其中样本均值为 $\bar{x} = \sum_{t=1}^T x_t/T$ 与 $\bar{y} = \sum_{t=1}^T y_t/T$:

$$\hat{\rho}_{x_t, y_t} = \frac{\sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^T (x_t - \bar{x})^2 \sum_{t=1}^T (y_t - \bar{y})^2}}$$

8.2.2 自相关函数

假设 $\{r_t\}$ 为若平稳收益率时间序列, 使用相关系数的概念推广至 r_t 与其过去值 r_{t-l} , 那么 r_t 与 r_{t-l} 的相关系数称为 r_t 间隔为 l 的**自相关系数** (Autoregressive coefficient), 即一个时间序列中当前节点, 与其自身在不同时间节点的相关关系 (相关系数):

$$\rho_l = \frac{\text{Cov}(r_t, r_{t-l})}{\sqrt{\text{Var}(r_t) \text{Var}(r_{t-l})}} = \frac{\text{Cov}(r_t, r_{t-l})}{\text{Var}(r_t)} = \frac{\gamma_l}{\gamma_0}$$

由于 $\{r_t\}$ 为弱平稳收益率时间序列, 应有 $\text{Var}(r_t) = \text{Var}(r_{t-l})$ 。根据定义, $\rho_0 = 1$ 且有 $-1 \leq \rho_l \leq 1$ 。一个弱平稳序列, 当且仅当对所有 $l > 0$ 都有 $\rho_l = 0$ 时, 其为序列不相关。

对于 $\{r_t\}$, 其样本均值为 $\bar{r} = \sum_{t=1}^T r_t/T$, 间隔为 1 的样本自相关系数为:

$$\hat{\rho}_l = \frac{\sum_{t=l+1}^T (r_t - \bar{r})(r_{t-l} - \bar{r})}{\sum_{t=1}^T (r_t - \bar{r})^2} \quad 0 \leq l < T-1$$

此时 $\hat{\rho}_1, \hat{\rho}_2, \dots$ 称为**样本自相关函数** (Autocorrelation Function, ACF)。若 $\{r_t\}$ 为独立同分布 (i.i.d.) 序列, 满足 $\mathbb{E}(r_t^2) < \infty$, 那么对于任意正整数 l , 有 $\hat{\rho}_l$ 渐进的服从均值为 0、方差为 $1/T$ 的 $N \sim (0, \frac{1}{T})$ 的正态分布。

若 $\{r_t\}$ 为弱平稳序列, 满足 $r_t = \mu + \sum_{i=0}^q \psi_i a_{t-i}$, 其中 $\phi_0 = 1$, $\{a_t\}$ 为均值为 0 的独立同分布变量序列, 对于 $l > q$, 则有 $\hat{\rho}_l$ 渐进的服从如下正态分布, 称为 Bartlett 公式:

$$\hat{\rho}_l \sim N \left(0, \frac{(1 + 2 \sum_{i=1}^q \rho_i^2)}{T} \right)$$

检验单个 ACF

对于正整数 l , 可以使用 t 检验 (t-statistics) 来检验 $H_0: \rho_l = 0$ 与 $H_a: \rho_l \neq 0$ 检验统计量为:

$$t_{\hat{\rho}} = \frac{\hat{\rho}_l}{\sqrt{(1 + 2 \sum_{i=1}^q \hat{\rho}_i^2) / T}}$$

如果 $\{r_t\}$ 为平稳的高斯序列并满足当 $j > l$ 时, $\rho_j = 0$, 则有他检验渐进的服从标准正态分布。此时为双侧检验, 假设显著性水平为 α , 当 $|t_{\hat{\rho}}| > z_{\alpha/2}$ 时拒绝 H_0 , 即在正态分布左右两端面积为 $\alpha/2$ 的区域内拒绝 H_0 , 而中间 $1 - \alpha$ (置信水平) 的区域不拒绝 H_0 , 其中 $Z_{\alpha/2}$ 为临界值 (Critical value)。

临界值根据统计量的具体分布决定, 由于统计量 $t_{\hat{\rho}}$ 渐近服从标准正态分布。对于双侧检验临界值 $\pm Z_{\alpha/2}$ 为, 使得标准正态分布 PDF 曲线下阴影面积部分, 左右两侧面积相加为 α 的 Z 值。具体而言为标准正态分布 CDF 的 $100(\alpha/2)$ 与 $100(1 - \alpha/2)$ 的两个分位点。例如当临界值为 $\alpha = 5\%$ 时, 查表 (Z-table) 得知, 为使得标准正态分布左右两侧阴影区域面积为 2.5% 与 97.5% 的两个临界值 Z 值, 为 $Z = \pm 1.96$ 。

混成检验

在时间序列研究中, 有两个著名的用于检测残差中自相关的混成检验 (Portmanteau test), 为 Ljung-Box 检验和 Box-Pierce 检验, 其中 Ljung-Box 是 Box-Pierce 的扩展。目的在于检验 r_t 的几阶自相关系数是否同时为零。可以采用的方法还包括计算 ACF 和 PACF 并观察其图像, 但是无论是 ACF 还是 PACF 都只能考虑某一特定滞后阶数的自相关, BP 检验与 LB 检验则是基于一系列滞后阶数。Box 和 Pierce (1970) 检验统计量为:

$$Q^*(m) = T \sum_{i=1}^m \hat{\rho}_i^2$$

$Q^*(m)$ 渐进地服从自由度为 m 的 χ^2 分布, 检验原假设为无显著序列相关 (自相关), 而备择假设为有显著的序列相关性。即当拒绝原假设 H_0 时, 则代表该时间序列有显著的序列相关性:

$$H_0 : \rho_1 = \cdots = \rho_m = 0$$

$$H_1 : \text{对于某 } i \in \{1, \dots, m\}, \rho_i \neq 0$$

Ljung 和 Box (1978) 提高有限样本时的检验功效, 修改后的统计量如下所示。临界值为 α , 则有当 $Q(m) > \chi_{\alpha}^2$ 时拒绝 H_0 , χ_{α}^2 为自由度为 m 的 χ^2 分布的 $100 - (1 - \alpha)$ 的分位点, 或当 p 值小于显著性水平 $1 - \alpha$ 时拒绝 H_0 。

$$Q(m) = T(T+2) \sum_{i=1}^m \frac{\hat{\rho}_i^2}{T-l}$$

8.3 白噪声与线性时间序列

8.3.1 白噪声

当 $\{r_t\}$ 为一个有限均值和有限方差的**独立同分布** (Independent and identically distributed, iid) 随机变量序列, 则称其为**白噪声** (White noise) 序列。对于白噪声序列, 有其所有的自相关函数 (ACF) 都为零。若 r_t 服从均值为 0 和方差为 α^2 的正态分布, 则称其为**高斯白噪声** (Gaussian white noise)。独立同分布指一组随机变量 X_t 中, 每个变量的概率分布都相同 (均值方差等相同), 但随机变量相互独立, 即 X_i 的取值不影响 X_j 的取值。

8.3.2 线性时间序列

如下时间序列称为线性时间序列 (Linear time series), 则有:

$$r_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i} = \mu + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$$

其中 μ 为 r_t 的均值, $\pi_0 = 1$, 并且 $\{a_t\}$ 为均值为零的独立同分布随机变量序列, 即 $\{a_t\}$ 为白噪声序列。可以认为 a_t 表示该时间序列 t 时刻出现的新息 (Innovation) 或扰动 (Shock)。其中 r_t 系数被称为 ψ -权重, 若 r_t 为弱平稳, 则可以利用 $\{a_t\}$ 的独立性得到:

$$\mathbb{E}(r_t) = \mu \quad \text{Var}(r_t) = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i^2$$

σ_a^2 为 a_t 的方差, 由于方差有限, 应有 $\text{Var}(r_t) < +\infty$, 所以 $\{\psi_i^2\}$ 必须是收敛序列, 当 $i \rightarrow \infty$ 时, $\psi_i^2 \rightarrow 0$, 即随着 i 的增大, 越早的历史扰动对当前时点的 r_t 影响越小, 直至消失。 r_t 与间隔 l 的自协方差为:

$$\begin{aligned} \gamma_l &= \text{Cov}(r_t, r_{t-l}) = \mathbb{E}[(r_t - \mu)(r_{t-l} - \mu)] \\ &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \psi_i a_{t-i}\right)\left(\sum_{j=0}^{\infty} \psi_j a_{t-l-j}\right)\right] \\ &= \mathbb{E}[(\psi_0 a_t + \psi_1 a_{t-1} + \dots + \psi_l a_{t-l} + \psi_{l+1} a_{t-l-1} + \dots)(\psi_0 a_{t-l} + \psi_1 a_{t-l-1} + \dots)] \\ &= \mathbb{E}(\psi_l \phi_0 a_{t-l} a_{t-l} + \psi_{l+1} \phi_1 a_{t-l-1} a_{t-l-1} + \dots) \\ &= \sum_{i=0}^{\infty} \psi_{i+l} \psi_i \mathbb{E}(a_{t-l-i}^2) = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+l} \end{aligned}$$

此时 ψ -权重与 r_t 自相关系数有如下关系:

$$\rho_l = \frac{\text{Cov}(r_t, r_{t-l})}{\text{Var}(r_t)} = \frac{\gamma_l}{\gamma_0} = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+l}}{1 + \sum_{i=0}^{\infty} \psi_i^2} \quad l \geq 0$$

8.4 自回归模型 (AR)

假设收益率 r_t 具有统计显著的间隔为 1 的自相关系数, 说明在预测 r_t 时, r_{t-1} 也发挥了较大的作用, 因此可以建立简单的线性回归模型:

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t$$

其中 $\{a_t\}$ 为均值为 0, 方差为 σ_a^2 的白噪音, 此时由于 r_t 为因变量, r_{t-1} 为自变量, 因此如上模型也被称为自回归模型 (Autoregressive model, AR), 由于与自身滞后一阶自回归, 因此称为一阶自回归模型或 AR(1) 模型。在过去收益率 r_{t-1} 已知条件下。有其均值与方差 (波动来源于白噪声) 有:

$$\begin{aligned} \mathbb{E}(r_t | r_{t-1}) &= \phi_0 + \phi_1 r_{t-1} \\ \text{Var}(r_t | r_{t-1}) &= \mathbb{E}[(r_t - \mathbb{E}(r_t))^2] \\ &= \mathbb{E}(a_t^2) = \text{Var}(a_t) = \sigma_a^2 \end{aligned}$$

对 AR(1) 模型进行推广, 因此 p 阶的 AR 模型表示在给定过去数据时, 那么滞后 1 阶至 p 阶联合决定了 r_t 的条件期望:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \cdots + \phi_p r_{t-p} + a_t$$

假设 AR(1) 是弱平稳的, 则有 $\mathbb{E}(r_t) = \mu$ 为常数, $\text{Var}(r_t) = \gamma_0$ 为常数, $\text{Cov}(r_t, r_{t-l}) = \gamma_l$ 是 l 的函数与 t 无关。对 AR(1) 模型两边取均值, 得到:

$$\mathbb{E}(r_t) = \phi_0 + \phi_1 \mathbb{E}(r_{t-1})$$

在平稳性的条件下 r_t 的期望为常数 μ , 即 $\mathbb{E}(r_t) = \mathbb{E}(r_{t-1}) = \mu$, 因此有:

$$\mathbb{E}(r_t) = \mu = \frac{\phi_0}{1 - \phi_1}$$

改写上式, 利用 $\phi_0 = \mu(1 - \phi_1)$, 代入 AR(1) 中有:

$$r_t = \mu(1 - \phi_1) + \phi_1 r_{t-1} + a_t = \mu + \phi_1(r_{t-1} - \mu) + a_t$$

将 $r_{t-1} - \mu = \phi_1(r_{t-2} - \mu) + a_{t-1}$ 代入上式, 得:

$$\begin{aligned} r_t - \mu &= \phi_1(\phi_1(r_{t-2} - \mu) + a_{t-1}) \\ &= \phi_1^2(r_{t-2} - \mu) + \phi_1 a_{t-1} + a_t \end{aligned}$$

重复上述流程, 重复代入, 可以得到:

$$r_t - \mu = a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \cdots = \sum_{i=0}^{\infty} \phi_1^i a_{t-i}$$

可以看到, AR(1) 模型为一般线性时间序列 $r_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i}$, 当 $\psi_i = \phi_1^i$ 时的特殊形式。由上式可知, $r_t - \mu$ 为 a_t 的线性组合, 且 a_t 互相独立, 那么应有 $\mathbb{E}((r_t - \mu)a_{t+1}) = 0$, 因此有:

$$\text{Cov}(r_{t-1}, a_t) = \mathbb{E}[(r_{t-1} - \mu)(a_t - 0)] = 0$$

对 AR(1) 模型的两边取方差, 常数的方差为零, 注意包含协方差项, 则有:

$$\begin{aligned} \text{Var}(r_t) &= \text{Var}(\phi r_{t-1} + a_t) \\ &= \phi_1^2 \text{Var}(r_{t-1}) + \sigma_a^2 + \phi_1 \text{Cov}(r_{t-1}, a_t) \\ &= \phi_1^2 \text{Var}(r_{t-1}) + \sigma_a^2 \end{aligned}$$

此时由于平稳性, 则有方差相等, 即 $\text{Var}(r_t) = \text{Var}(r_{t-1})$, 因此:

$$\text{Var}(r_t) = \frac{\sigma_a^2}{1 - \phi_1^2}$$

此时由于方差是非负且有限的, 那么则应有 $\phi_1^2 < 1$ 。这样由 AR(1) 模型的弱平稳性可推得 $-1 < \phi_1 < 1$, 为充分条件。反之, 由 $\phi_1 < 1$ 可推出 AR(1) 是弱平稳的, 为必要条件。综上所述, AR(1) 模型是弱平稳的充分必要条件是 $|\phi_1| < 1$ 。

AR(1) 的自相关系数

已知当 AR(1) 时, 有 $\psi_i = \phi_1^i$, 则 r_t 的方差应有:

$$\gamma_0 = \text{Var}(r_t) = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i^2 = \sigma_a^2 \sum_{i=0}^{\infty} (\phi_1^i)^2$$

将 ϕ_1^{i+l} 分解为 $\phi_1^i \phi_1^l$, 同时有自协方差 γ_l :

$$\gamma_l = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+l} = \sigma_a^2 \sum_{i=0}^{\infty} \phi_1^i \phi_1^{i+l} = \sigma_a^2 \phi_1^l \sum_{i=0}^{\infty} \phi_1^i \phi_1^i = \sigma_a^2 \phi_1^l \sum_{i=0}^{\infty} (\phi_1^i)^2$$

因此有当 $l > 0$ 时, 有 $\gamma_l = \phi_1 \gamma_{l-1}$ 或 $\gamma_l = \phi_1^l \gamma_0$. 同时 AR(1) 模型的自相关系数为 ($\rho_0 = 1$):

$$\rho_l = \frac{\text{Cov}(r_t, r_{t-l})}{\text{Var}(r_t)} = \phi_1^l$$

AR(2) 模型

AR(2) 模型, 如同 AR(1) 模型, 为 r_t 自身与其滞后一阶与滞后两阶的函数:

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + a_t$$

同理可以得到其期望:

$$\mathbb{E}(r_t) = \mu = \frac{\phi_0}{1 - \phi_1 - \phi_2}$$

将 $\phi_0 = (1 - \phi_1 - \phi_2)\mu$ 代回 AR(2) 模型中, 可得:

$$r_t - \mu = \phi_1(r_{t-1} - \mu) + \phi_2(r_{t-2} - \mu) + a_t$$

两边同乘以 $(r_{t-l} - \mu)$, 得到:

$$(r_{t-l} - \mu)(r_t - \mu) = \phi_1(r_{t-l} - \mu)(r_{t-1} - \mu) + \phi_2(r_{t-l} - \mu)(r_{t-2} - \mu) + (r_{t-l} - \mu)a_t$$

由于 $\mathbb{E}[(r_{t-l} - \mu)a_t]$, 对等式两边取期望则有:

$$\gamma_l = \phi_1 \gamma_{l-1} + \phi_2 \gamma_{l-2} \quad l > 0$$

对上式同除 γ_0 可得 r_t 的 ACF 性质有:

$$\rho_l = \phi_1 \rho_{l-1} + \phi_2 \rho_{l-2} \quad l > 0$$

并且有 $\rho_1 = \phi_1 \rho_0 + \phi_2 \rho_{-1} = \phi_1 \rho_0 + \phi_2 \rho_1$, 因此对于平稳的 AR(2) 序列 r_t 应有:

$$\begin{aligned} \rho_0 &= 1 \\ \rho_1 &= \frac{\rho_1}{1 - \rho_2} \\ \rho_l &= \phi_1 \rho_{l-1} + \phi_2 \rho_{l-2} \quad l \geq 2 \end{aligned}$$

令 B 为向后推移算子 (back-shift operator), 即有 $B \times \rho_l = \rho_{l-1}$, 有时也用 L 表示 (Lag operator, 滞后算子)。可以写成如下形式:

$$(1 - \phi_1 B - \phi_2 B^2) \rho_l = 0$$

该方程对应二次多项式, 或称为逆特征方程, 并且方程的解, 称为逆特征解:

$$\begin{aligned} 1 - \phi_1 x - \phi_2 x^2 &= 0 \\ x &= \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \end{aligned}$$

假设方程的解 (逆特征解) 为 ω_1 和 ω_2 , 在时间序列文献中, 称这两个解的倒数即 $1/\omega_1$ 与 $1/\omega_2$ 为 AR(2) 模型的特征根 (Characteristic roots), 即逆特征解与特征解为互为倒数。AR(2) 序列平稳的条件就是两个特征根的模都小于 1, 或逆特征根大于 1, 保证模型的自相关函数随着滞后 l 长度的增加, 而趋向于 0, 即历史冲击的影响越来越小, 为序列平稳的必要条件。

AR(p) 模型

同样将 AR(1) 与 AR(2) 的结果推广至 AR(p) 模型, 对于平稳的 AR(p) 序列, 其均值为:

$$\mathbb{E}(r_t) = \mu = \frac{\phi_0}{1 - \phi_1 - \dots - \phi_p}$$

对应的逆特征方程有:

$$1 - \phi_1 x - \phi_2 x^2 - \dots - \phi_p x^p = 0$$

该方程的所有解得模都大于 1，则有序列 $\{r_t\}$ 是平稳的。或换言之，解得倒数为该模型的特根，平稳性要求所有特征根的模都小于 1。对于 $AR(p)$ 的平稳性有如下充分必要条件：

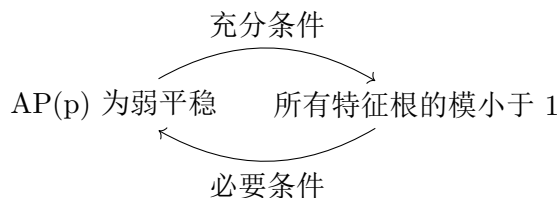


图 7: $AR(p)$ 为弱平稳的充分必要条件

特征方程与特征根

例子 8.1. 斐波拉契 (Fibonacci) 数列

斐波拉契数列满足 $a_1 = 1$, $a_2 = 1$, 且有 $a_{n+2} = a_{n+1} + a_n$, $n = 1, 2, \dots$ 。试图将 a_n 改写为函数形式 $f(n)$, 则有 $a_{n+2} = a_{n+1} + a_n$ 改写为 $f(n+2) = f(n+1) + f(n)$ 。假设已求得该方程的两个特解 f_1 与 f_2 , 对于任何实数 c_1 与 c_2 , $f = c_1 f_1 + c_2 f_2$ 也是方程的解:

$$f_1(n+2) = f_1(n+1) + f_1(n)$$

$$f_2(n+2) = f_2(n+1) + f_2(n)$$

$$c_1 f_1(n+2) + c_2 f_2(n+2) =$$

两式相加, 可以看到 $f = c_1 f_1 + c_2 f_2$ 满足原式 $f(n+2) = f(n+1) + f(n)$:

$$(c_1 f_1(n+1) + c_2 f_2(n+1)) + (c_1 f_1(n) + c_2 f_2(n))$$

代入初始条件 $a_1 = a_2 = 1$, 即 $f(1) = f(2) = 1$ 可以确定两个待定系数。将原式同除 $f(n)$ 得到:

$$\frac{f(n+2)}{f(n+1)} \cdot \frac{f(n+1)}{f(n)} - \frac{f(n+1)}{f(n)} - 1 = 0$$

若 $|f(n)|$ 为等比数列, 则有上式变为 $x^2 - x - 1 = 0$ 为特征方程, 其解为 $x_{1,2} = \frac{1 \pm \sqrt{5}}{2}$, 两个特征特解应为:

$$f_1(n) = \left(\frac{1 + \sqrt{5}}{2} \right)^{n-1} \quad f_2(n) = \left(\frac{1 - \sqrt{5}}{2} \right)^{n-1}$$

该方程的通解应有：

$$a_n = f(n) = c_1 f_1(n) + c_2 f_2(n) = c_1 \left(\frac{1+\sqrt{5}}{2} \right)^{n-1} + c_2 \left(\frac{1-\sqrt{5}}{2} \right)^{n-1}$$

由初始条件 $f(1) = f(2) = 1$ 可得：

$$\begin{aligned} c_1 + c_2 &= 1 \\ c_1 \left(\frac{1+\sqrt{5}}{2} \right) + c_2 \left(\frac{1-\sqrt{5}}{2} \right) &= 1 \end{aligned}$$

联立求解可得：

$$c_1 = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right) \quad c_2 = -\frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)$$

最终得证：

$$a_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^n - B \left(\frac{1-\sqrt{5}}{2} \right)^n \right]$$

【待补充】偏自相关系数（PACF）

8.5 残差自相关检验

To check the serial correlations of residuals, we recommend that the Ljung-Box statistics be used instead of the Durbin-Watson (DW) statistic because the latter only considers the lag-1 serial correlation. There are cases in which serial dependence in residuals appears at higher order lags. This is particularly so when the time series involved exhibits some seasonal behavior.

Durbin-Watson test

- H0: $\rho = 0$, no autocorrelation / serial correlation in residual - H1: $\rho \neq 0$, autocorrelation in residual, follow first order autoregressive process

$$\text{Test statistic - residual at lag 1, } \epsilon_t = \rho \epsilon_{t-1} + u_t - DW = \frac{\sum_{t=2}^T (\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^T \epsilon_t^2}$$

2 -> no autocorrelation 0-2 -> positive autocorrelation 2-4 -> negative autocorrelation

【待补充】

8.6 移动平均模型 (MA)

移动平均模型 (Moving-average, MA)

$$r_t = \phi_0(1 - \theta_1) + a_t - \theta_1 a_{t-1}$$

r_t 为两个扰动项 a_t 与 a_{t-1} 的加权平均, 因此模型称为 1 阶 MA 模型, 简称 MA(1)。其中 $\{a_t\}$ 为白噪声序列, 且 c_0 为常数:

$$r_t = c_0 + a_t - \theta_1 a_{t-1}$$

类似的有 MA(2) 为:

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

MA(q) 模型为:

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

8.7 自回归移动平均模型 (ARMA)

自回归移动平均模型 (Autoregressive moving-average model, ARMA)。ARMA(1,1) 模型有 ($\phi_1 \neq \theta_1$):

$$r_t - \phi_1 r_{t-1} = \phi_0 + a_t - \phi_1 a_{t-1}$$

对于一般的 ARMA(p,q) 模型而言, 即该模型有 p 阶滞后自回归项, 与 q 阶滞后残差项:

$$r_t = \phi_0 + \sum_{i=1}^p \phi_i r_{t-i} + a_t - \sum_{i=1}^q \theta_i a_{t-i}$$

平稳时有均值为:

$$\mathbb{E}(r_t) = \frac{\phi_0}{1 - \phi_1 - \cdots - \phi_p}$$

8.8 自回归移动平均模型 (ARIMA)

将 ARMA 推广至允许其 AR 多项式以 1 作为它的特征根, 则该模型变成了自回归求和移动平均模型 (Autoregressive integrated moving average, ARIMA)。因为 A 多项式有单位根 1, 因此 ARIMA 模型称为是单位根非平稳的。

8.9 单位根检验

假设模型有：

$$p_t = \phi_1 p_{t-1} + e_t$$

$$p_t = \phi_0 + \phi_1 p_{t-1} + e_t$$

$$p_t = \phi_1 p_{t-1} + e_t$$

著名的单位根检验 Dickey 和 Fuller (1979) 的原假设为存在单位根，非平稳。而备择假设为该序列为平稳的，有：

$$H_0 : \phi_1 = 1$$

$$H_a : \phi_1 < 1$$

使用最小二乘法估计 T 检验：

$$\hat{\phi}_1 = \frac{\sum_{t=1}^T p_{t-1} p_t}{\sum_{t=1}^T p_{t-1}^2}$$

$$\hat{\sigma}_e^2 = \frac{\sum_{t=1}^T (p_t - \hat{\phi}_1 p_{t-1})^2}{T - 1}$$

假设 $p_0 = 0$ ，T 为样本容量，此时 **Dickey-Fuller 检验** 的统计量为：

$$DF = \frac{\hat{\phi}_1 - 1}{s.d.(\hat{\phi}_1)} = \frac{\sum_{t=1}^T p_{t-1} e_t}{\hat{\sigma}_a \sqrt{\sum_{t=1}^T p_{t-1}^2}}$$

对于计量常用的 AR(p) 模型而言，为了检验 AR(p) 模型是否存在单位根，可以使用扩展的 **Dickey-Fuller 检验** (Augmented Dickey-Fuller)。其中原假设为 $H_0 : \beta = 1$ ，而备择假设为 $H_a : \beta < 1$ ，模型有：

$$x_t = c_t + \beta x_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta x_{t-i} + e_t$$

$\Delta x_i = x_i - x_{i-1}$ 为 x_t 的差分数列，且 c_t 可以使零或者常数，或 $c_t = \omega_0 + \omega_1 t$ 。ADF 检验的统计量如下，其中 $\hat{\beta}$ 为 β 的最小二乘估计：

$$ADF = \frac{\hat{\beta} - 1}{s.e.(\hat{\beta})}$$

8.10 一致的协方差矩阵估计

对于同时研究两个时间序列 y_t 与 x_t 的关系，考虑如下线性回归：

$$y_t = \alpha + \beta x_t + e_t$$

误差项存在着序列相关或者条件异方差性，而分析的目标为回归系数 α 与 β 做推断。在系数的最小二乘估计依然满足一致性的情况下，比较广泛的有两种方法能给出系数协方差的一致性估计（Consistent Covariance Matrix Estimation），有 Eicker（1967）和 White（1980）提出的异方差一致估计（HC），与 Newey 和 West（1987）提出的异方差及自相关一致估计（HAC）。使用向量将回归模型改写：

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + e_t \quad t = 1, 2, \dots, T$$

其中 y_t 为因变量， $\mathbf{x}_t = (x_{1t}, \dots, x_{kt})'$ 为包含常数项在内的 k 维向量， $\boldsymbol{\beta} = (\beta_{1t}, \dots, \beta_{kt})'$ 为参数向量， $\boldsymbol{\beta}$ 那么最小二乘估计与协方差矩阵为：

$$\boldsymbol{\beta} = \left[\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \sum_{t=1}^T \mathbf{x}_t y_t \quad \text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma_e^2 \left[\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right]^{-1}$$

当存在序列相关或条件异方差时，上述的协方差矩阵估计就不再是一致的，使得 $\hat{\boldsymbol{\beta}}$ 的 T 统计量偏大。使用 White（1980）的方法调整之后有，其中 $\hat{e}_t = y_t - \mathbf{x}_t' \hat{\boldsymbol{\beta}}$ 为 t 时刻的残差：

$$\text{Cov}(\hat{\boldsymbol{\beta}})_{HC} = \left[\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \left[\sum_{t=1}^T \hat{e}_t^2 \mathbf{x}_t \mathbf{x}_t' \right] \left[\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right]^{-1}$$

而 Newey 和 West（1987）本质上是使用非参数的方法估计了 $\left\{ \sum_{t=1}^T \hat{e}_t \mathbf{x}_t \right\}$ 的协方差矩阵，有：

$$\text{Cov}(\hat{\boldsymbol{\beta}})_{HAC} = \left[\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \hat{\mathbf{C}}_{HAC} \left[\sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right]^{-1}$$

其中：

$$\hat{\mathbf{C}}_{HAC} = \sum_{t=1}^T \hat{e}_t^2 \mathbf{x}_t \mathbf{x}_t' + \sum_{j=1}^l \omega_j \sum_{t=j+1}^T \left(\mathbf{x}_t \hat{e}_t \hat{e}_{t-j} \mathbf{x}_{t-j}' + \mathbf{x}_{t-j} \hat{e}_{t-j} \hat{e}_t \mathbf{x}_t' \right)$$

ω_j 为权重函数，Newey 和 West 建议 l 取 $4(T/100)^{2/9}$ 的整部：

$$\omega_j = 1 - \frac{j}{l+1}$$

8.11 格兰杰因果关系

格兰杰因果关系检验 (Granger causality test) 是一种假设检验的统计方法, 检验一组时间序列 x 是否为另一组时间序列 y 的原因。它的基础是回归分析当中的自回归模型。

9 条件异方差模型

假设 r_t 服从平稳的 ARMA(p,q) 模型, 即:

$$r_t = \phi_0 + \sum_{i=1}^p \phi_i r_{t-i} + a_t - \sum_{i=1}^q \theta_i a_{t-i}$$

对于增加了解释变量 (外生变量) $x_{i,t}$ 的模型, 其中 y_t 只是记号, 表示去除解释变量影响后的调整收益率序列。

$$\begin{aligned} r_t &= \mu_t + a_t \\ \mu_t &= \sum_{i=1}^p \phi_i y_{t-i} - \sum_{i=1}^q \theta_i a_{t-i} \\ y_t &= r_t - \phi_0 - \sum_{i=1}^k \beta_i x_{it} \end{aligned}$$

对于给定 $t-1$ 时刻, 已知信息集合 F_{t-1} 时, r_t 的条件均值与条件方差为:

$$\begin{aligned} \mu_t &= \mathbb{E}(r_t | F_{t-1}) \\ \sigma_t^2 &= \text{Var}(r_t | F_{t-1}) \\ &= \mathbb{E}[(r_t - \mu_t)^2 | F_{t-1}] \\ &= \mathbb{E}[a_t^2 | F_{t-1}] \\ &= \text{Var}(a_t | F_{t-1}) \end{aligned}$$

9.1 ARCH 效应检验

$a_t = r_t - \mu_t$ 为均值方程的残差, 则可以用平方序列 a_t^2 来检验条件异方差性, 即所谓的 **ARCH** 效应, 即为残差平方的序列相关性, 或白噪声检测。第一种方式, 可以使用 Ljung-Box 统计量 $\{Q(m)\}$

应用于序列 $\{a_t^2\}$, 参见 Mcleod 和 Li (1983)。该检验的原假设是 $\{a_t^2\}$ 序列的前 m 个间隔的 ACF 值都为零。第二种方式, 可以使用 Engle (1982) 中的拉格朗日乘子检验。对于拉格朗日乘子检验 (Lagrange Multiplier Test), 等价于对如下线性回归中用 F 统计量检验 $\alpha_i = 0$ ($i = 1, \dots, m$), 即原假设有:

$$H_0 : \alpha_1 = \dots = \alpha_m = 0$$

线性回归为:

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \dots + \alpha_m a_{t-m}^2 + e_t \quad t = m+1, \dots, T$$

m 为正整数, T 为样本容量, 其中 e_t 为误差项, \hat{e}_t 为上述线性回归最小二乘法估计的残差, $\bar{\omega}$ 为 a_t^2 的样本平均值。令:

$$SSR_0 = \sum_{t=m+1}^T (a_t^2 - \bar{\omega})^2 \quad \bar{\omega} = \frac{1}{T} \sum_{t=1}^T a_t^2$$

$$SSR_1 = \sum_{t=m+1}^T \hat{e}_t^2$$

F 统计量服从自由度为 m 的 χ^2 分布, 若 $F > \chi_m^2(\alpha)$ 或 F 的 p 值小于 α 则拒绝原假设, 其中 $\chi_m^2(\alpha)$ 为 χ_m^2 上的 $100(1-\alpha)$ 点:

$$F = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(T - 2m - 1)}$$

备注 9.1. ARCH 效应检验或称为 *Engle's ARCH test*, 如上所述应平方序列检验异方差性, 详情见 Engle (1982), 与其他检验如 *Ljung-box* 检验, 拉格朗日检验, 在 *Python* 与 *Matlab* 中有如下模块可进行计算:

- *Python: acorr_lm* (拉格朗日乘子序列相关检验)
- *Python: acorr_ljungbox* (*Ljung-Box* 序列相关检验)
- *Python: het_arch* (*Engle* 异方差检验)
- *Python: het_breuschpagan* (*Breusch-Pagan* 拉格朗日乘子异方差检验)
- *Matlab: lbqtest* (*Ljung-Box* 序列相关检验)
- *Matlab: archtest* (*Engle* 异方差检验)

9.2 ARCH 模型

第一个尝试给波动率 (σ_t , 收益率的条件标准差) 建模并提供系统框架的是由 Engle (1982) 提出的 ARCH 模型。基本思想为:

- 资产收益率的扰动项 $a_t = r_t - \mathbb{E}(r_t | F_{t-1})$ 是序列不相关的, 但不是独立的, 若独立则不需要建模条件方差模型
- a_t 的不独立性, 描述为 $\text{Var}(r_t | F_{t-1}) = \text{Var}(a_t | F_{t-1})$, 即可以用 a_t^2 的滞后项的线性组合表示

具体而言, ARCH(m) 模型如下所示。其中 $\{\varepsilon_t\}$ 为均值为 0, 方差为 1 的独立同分布随机变量序列:

$$\begin{aligned} a_t &= \sigma_t \varepsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2 \end{aligned}$$

证明. 由于 $a_t = r_t - \mathbb{E}(r_t | F_{t-1})$, 因此 $\mathbb{E}(a_t) = \mathbb{E}(r_t) - \mathbb{E}[\mathbb{E}(r_t | F_{t-1})] = \mathbb{E}(r_t) - \mathbb{E}(r_t) = 0$ 。并且有 $\mathbb{E}(a_t | F_{t-1}) = \mathbb{E}(\sigma_t \varepsilon_t | F_{t-1}) = \varepsilon_t \mathbb{E}(\sigma_t | F_{t-1}) = 0$ 。

由于 ε_t 独立, 因此有 $\text{Var}(\varepsilon_t) = \text{Var}(\varepsilon_t | F_{t-1}) = \mathbb{E}(\varepsilon_t^2 | F_{t-1}) = 1$ 。对于条件方差有性质提取已知量, 即对于已知量, 可以直接从条件期望中提出。具体而言, 若 X 是 \mathcal{H} 可测的, 应有:

$$\mathbb{E}[XY | \mathcal{H}] = X \mathbb{E}[Y | \mathcal{H}]$$

由于 σ_t 是可以通过 F_{t-1} 信息集获得, 为已知量。因此可以证明:

$$\begin{aligned} \text{Var}(r_t | F_{t-1}) &= \mathbb{E} \left[(r_t - \mathbb{E}[r_t | F_{t-1}])^2 | F_{t-1} \right] \\ &= \mathbb{E}(a_t^2 | F_{t-1}) = \mathbb{E}(\sigma_t^2 \varepsilon_t^2 | F_{t-1}) \\ &= \sigma_t^2 \mathbb{E}(\varepsilon_t^2 | F_{t-1}) = \sigma_t^2 \\ &= \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2 \end{aligned}$$

同时其 a_t 的无条件方差为：

$$\begin{aligned}
 \text{Var}(a_t) &= \mathbb{E}[(a_t - E[a_t])^2] \\
 &= \mathbb{E}(a_t^2) = \mathbb{E}[\mathbb{E}(a_t^2 | F_{t-1})] = \mathbb{E}[\mathbb{E}(\sigma_t^2 \varepsilon_t^2 | F_{t-1})] \\
 &= \mathbb{E}[\sigma_t^2 \mathbb{E}(\varepsilon_t^2 | F_{t-1})] = \mathbb{E}[\sigma_t^2] \\
 &= \mathbb{E}[\alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2]
 \end{aligned}$$

□

9.2.1 ARCH 模型的性质

9.3 GARCH 模型

10 卡尔曼滤波

卡尔曼滤波 (Kalman Filter) 的 State Space Representation 有：

$$\begin{aligned}
 \xi_{t+1} &= \mathbf{F} \xi_t + \mathbf{v}_{t+1} \\
 \mathbf{y}_t &= \mathbf{A}' \mathbf{x}_t + \mathbf{H}' \xi_t + \mathbf{w}_t
 \end{aligned}$$

定义有：

$$\begin{aligned}
 \mathbf{P}_{t|t-1} &\equiv E[(\xi_t - \hat{\xi}_{t|t-1})(\xi_t - \hat{\xi}_{t|t-1})'] \\
 \mathbf{K}_t &= \mathbf{F} \mathbf{P}_{t|t-1} \mathbf{H} (\mathbf{H}' \mathbf{P}_{t|t-1} \mathbf{H} + \mathbf{R})^{-1}
 \end{aligned}$$

$$\begin{aligned}
 E[(\xi_t - \hat{\xi}_{t|t-1})(\xi_t - \hat{\xi}_{t|t-1})'] &= \mathbf{P}_{t|t-1} \\
 E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'] &= \mathbf{H}' \mathbf{P}_{t|t-1} \mathbf{H} + \mathbf{R} \\
 E\{(\xi_t - \hat{\xi}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'\} &= \mathbf{P}_{t|t-1} \mathbf{H}
 \end{aligned}$$

计算步骤

$$\begin{aligned}\hat{\xi}_{1|0} &= E(\xi_1) \\ P_{1|0} &= E\{[\xi_1 - E(\xi_1)][\xi_1 - E(\xi_1)]'\}\end{aligned}$$

迭代

平滑

$$\begin{aligned}\mathbf{J}_t &\equiv \mathbf{P}_{t|t} \mathbf{F}' \mathbf{P}_{t+1|t}^{-1} \\ \hat{\xi}_{t|T} &= \hat{\xi}_{t|t} + J(\hat{\xi}_{t+1|T} - \hat{\xi}_{t+1|t})\end{aligned}$$