统计整理

杨弘毅

创建：2020 年 4 月 9 日
修改：2021 年 10 月 27 日

目录

1 TODO **3**

2 基础 **3**

2.1 期望 . . . 3
2.2 方差 . . . 4
2.3 协方差 . . . 5
2.4 相关系数 . . . 5
2.5 假设检验 . . . 5
2.6 t 检验 . . . 6
2.7 参数与非参数检验 . . . 8
2.8 回归模型分类 . . . 9

3 矩 **10**

3.1 理解 . . . 10
3.2 定义 . . . 11
3.3 分类 . . . 11
3.4 矩母函数 . . . 14
3.4.1 定义 . . . 14
3.4.2 性质 . . . 15

4 条件概率 **16**

4.1 条件概率 16
4.2 条件概率分布 18
4.2.1 链式法则 19
4.2.2 条件期望 19
4.2.3 条件方差 20
4.3 贝叶斯定理 21
4.4 贝叶斯因子 24
4.4.1 后验因子估计 24
4.4.2 发生率与贝叶斯因子 24

5 似然 25
5.1 似然与概率 25
5.2 似然函数 26
5.3 似然比 26
5.4 最大似然估计 27

6 Chi-square distribution 27

7 时间序列 27
7.1 平稳性 27
7.2 自相关函数 28
7.2.1 相关系数 28
7.2.2 自相关函数 29
7.3 自回归模型 30
7.4 移动平均模型 31
7.5 自回归移动平均模型 31
7.6 差分自回归移动平均模型 31
7.7 格兰杰因果关系 31
7.8 单位根 31

8 条件异方差模型 32
8.1 ARCH 效应检验 32

9 卡尔曼滤波 33

1 TODO

- likelihood, log-likelihood, goodness-of-fit, quasi-maximum likelihood, ratio test
- Chi-square, joint hypothesis
- Newey West 1987
- Durbin Watson

2 基础

2.1 期望

对于随机变量 X，其概率空间为 $(\Omega, \mathcal{F}, P)$，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）。

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

其性质有：

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$
$$\mathbb{E}[aX] = a\mathbb{E}[X]$$
$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \quad \text{(X,Y are independent)}$$

2.2 方差

对于方差（Variance）或 σ^2，定义有：

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

同理，或其连续积分形式有：

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

其性质有：

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

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：

$$\begin{aligned}\mathrm{Cov}(X,a) &= 0 \\ \mathrm{Cov}(X,X) &= \mathrm{Var}(X) \\ \mathrm{Cov}(X,Y) &= \mathrm{Cov}(Y,X) \\ \mathrm{Cov}(aX,bY) &= ab\,\mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X+a,Y+b) &= \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(aX+bY,cW+dV) &= ac\,\mathrm{Cov}(X,W) + ad\,\mathrm{Cov}(X,V) \\ &\quad + bc\,\mathrm{Cov}(Y,W) + bd\,\mathrm{Cov}(Y,V)\end{aligned}$$

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：

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

2.5 假设检验

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**（Althernative 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，可以认为我们并没办法举出反例。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。

2.6 t 检验

t 检验（或称 Student's t test），有如下常见的四个用途：

- 单样本均值检验（One-sample t-test）: 用于检验总体方差未知、正态或近似正态的单样本的均值是否与已知的总体均值相等

- 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
- 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

z 检验（z-test）检验，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），用于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，因此有：

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

同样以单样本检验样本均值与总体均值是否有差异为例，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，用于样本较小（样本量小于 30）或总体标准差未知。t-分布描述了的是标准化（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。其中分母为样本标准误（Standard error）为 $\sigma/\sqrt{n}$，$\hat{\sigma}$ 为样本标准差。

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

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1，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。

备注 2.1. *关于标准误，当统计量为样本均值时，也称为均值标准误（Standard error of the mean, SEM）。假设 $x_1, x_2, \ldots, x_n$ 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 $T = x_1 + x_2 + \cdots + x_n$，样本均值为 $\bar{x} = T/n$，此时有：*

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

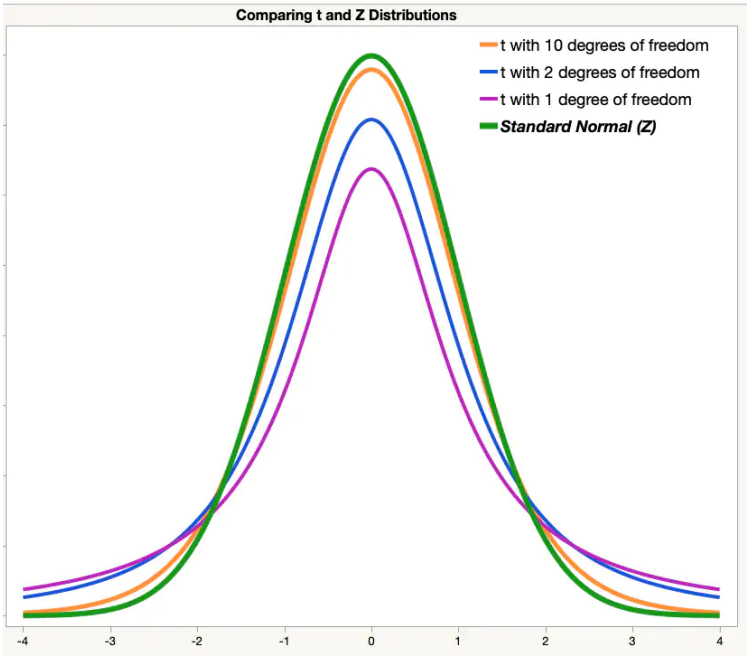


图 1: z 分布与 t 分布的差异

对于样本均值的方差有：

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

即样本均值服从 $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ 的正态分布。此时标准误，即样本均值的标准差为：

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

2.7 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 t 检验（配对或非配对），方差分析（ANOVA）。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。

2.8 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）

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：$\mu_u = \int x^n f(x) dr$，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。

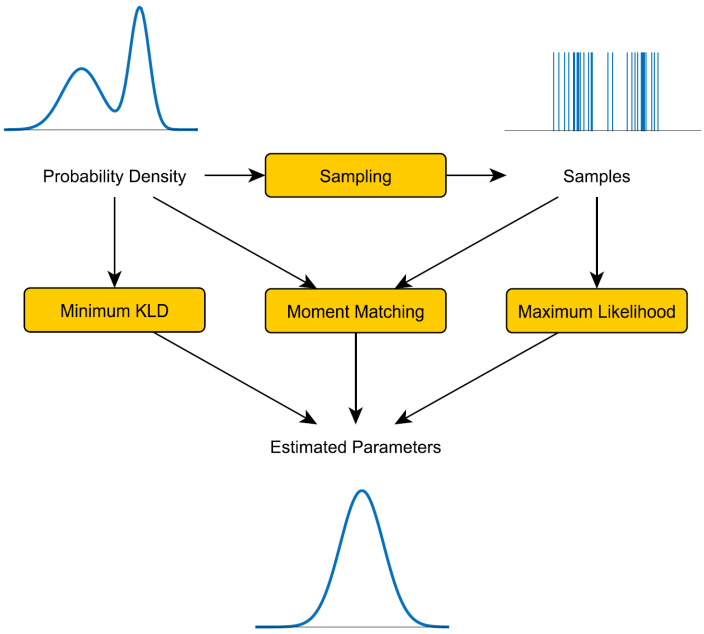


图 2: 矩匹配

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，即称为中心距（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 $\text{Skewness}(X) = \frac{\mathbb{E}[(X-\mu)^4]}{\sigma^3}$ 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 $\sigma = \left(\mathbb{E}[(X-\mu)^2]\right)^{\frac{1}{2}}$。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：

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

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：

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

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

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

其离散形式则有：

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

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：

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

此时，第 n 阶标准矩有：

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

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：

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

偏度（skewness）

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

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

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**：

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

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：

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

- 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 fuction，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：

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

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 $\mathbb{E}(X^n)$*

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

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：

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

□

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：

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

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 条件概率

4.1 条件概率

条件概率（Conditional probability）记为 $P(A \mid B)$，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。

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

备注 4.1. 同理 $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$，即有 $P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B)$。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。

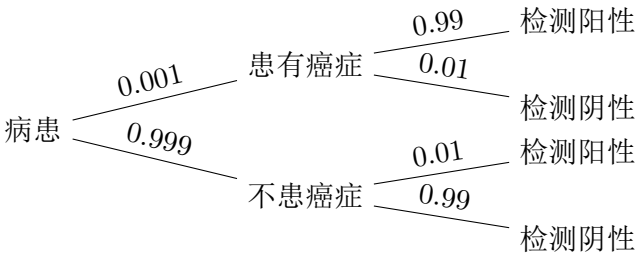
例子 4.2. 假设有两个小碗，分别为甲碗与乙碗，每个小碗中有若干小球，共有蓝色与黄色两种颜色，具体数量如下。

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 $P(\text{蓝色} \mid \text{碗} = \text{甲碗}) = \frac{1}{5}$，此时样本空间改变，被限制在了甲碗中，此时称为条件概率。

	蓝色	黄色
甲碗	1	4
乙碗	3	2

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 $P(\text{甲碗} \mid \text{颜色} = \text{蓝色}) = \frac{1}{4}$，注意 $P(\text{甲碗} \mid \text{蓝色})$ 与 $P(\text{蓝色} \mid \text{甲碗})$ 的概率并不相同。

例子 4.3. 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：

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

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 4.4. 在医学中将 $P(\text{患病} \mid \text{检测阳性})$ 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 Positve*），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 Positve*）。

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

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。

4.2 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 $P(Y = y \,|\, X = x)P(X = x) = P\left[(X = x) \cap (Y = y)\right] = P(X = x \,|\, Y = y)P(Y = y)$。

例子 4.5. 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）：

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

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）。

4.2.1 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，如上所述对于时间 A 与事件 B：

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

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

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

如当 $n=3$ 时，应有:

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

4.2.2 条件期望

与条件概率原理相同，条件期望（Conditional expectation），记为 $\mathbb{E}(X\mid Y)$ 或 $\mathbb{E}(X\mid Y=y)$，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。

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

例子 4.6. 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 $\mathbb{E}(X \mid Y = 1) = \frac{1+0+0}{3} = \frac{1}{3}$，条件期望 $\mathbb{E}(X \mid Y = 0) = \frac{0+1+1}{3} = \frac{2}{3}$

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

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

4.2.3 条件方差

条件方差（Conditional variance）定义为：

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

命题 4.7. 总期望定律（*Law of total expectation*）或称为双重期望定理（*Double expectation theorem*）有：

$$\mathbb{E}(X) = \mathbb{E}\left[\mathbb{E}(X \mid Y)\right]$$

证明.

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

□

备注 4.8. 或可以先展开内层期望进行证明 $\mathbb{E}\left[\mathbb{E}(X \mid Y)\right] = \mathbb{E}\left[\sum_x xP(X = x \mid Y)\right]$。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 4.9. 总方差定律（*Law of total variance*）有：

$$\operatorname{Var}(X) = \mathbb{E}\left[\operatorname{Var}(X \mid Y)\right] + \operatorname{Var}\left[\mathbb{E}(X \mid Y)\right]$$

证明. 计算等式右边第一项，并且根据总期望定律：

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

对于等式右边第二项，根据方差的定义有：

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

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

4.3 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：

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

$P(A)$ 为先验概率（Prior probability），即不考虑任何 B 方面的因素。而 $P(A \mid B)$ 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 $P(B \mid A)$ 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 $P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B)$，由此变形可推导推导贝叶斯定理，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 $P(B \mid A) = P(B)$，A 不受 B 的影响。

例子 4.10. 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 $P(\text{甲碗} \mid \text{蓝色})$，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目，最直观的计算方法为：

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

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音表达，则有：

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

例子 4.11. 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：

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

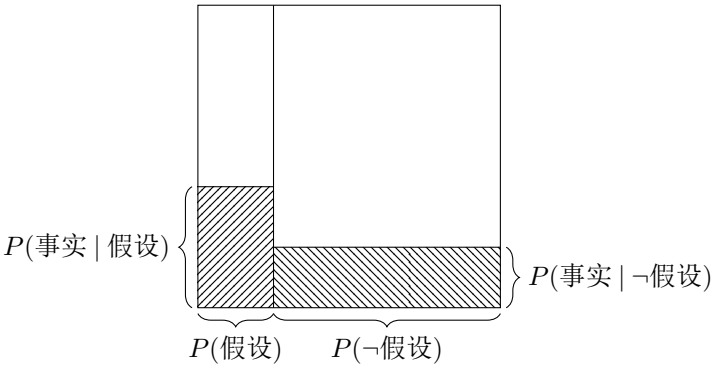
假设该病人进行了第二次同样的检测，此时再次使用贝叶斯定理，可以继续更新该病患真正患有癌症的概率，为了简便将条件事件 (患有癌症 | 第一次阳性) 记为 (患癌首阳) 其概率近似为 9% 代入计算。可以看到第二次检测阳性之后，实际患有癌症的几率大大提高。

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

备注 4.12. 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：

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

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 $P(\text{事实} \mid \text{假设}) = P(\text{事实} \mid \neg\text{假设})$ 时，即上图两个长方体高度相同。此时易知 $P(\text{假设} \mid \text{事实}) = P(\text{假设})$。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。

4.4 贝叶斯因子

4.4.1 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：

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

贝叶斯定理可以进行如下估算：

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

例子 4.13. 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：

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

因此 $P(\text{患有癌症} \mid \text{检测阳性}) \approx P(\text{患有癌症}) \times 99 = 0.1\% \times 99 = 9.9\%$，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。

4.4.2 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 $P(\text{假设})P(\text{事实} \mid \text{假设})$ 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 4.14. 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：

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

5 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 $X = x_1, x_2, \ldots, x_n$，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 $f(\cdot \mid \theta)$，对于任意 x_i 其概率 $f(x_i \mid \theta)$ 若非常低，则说明 x_i 出现在分布 $f(\cdot \mid \theta)$ 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 $f(\cdot \mid \theta)$，这其中的可能性或概率，即称之为似然。

5.1 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：

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

对于连续随机变量：

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

例子 5.1. 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：

$$P(32 < w < 34 \mid \mu = 32, \sigma = 2.5) = 0.29$$
$$P(w = 34 \mid \mu = 34, \sigma = 2.5) = 0.16$$

而对于似然，则为给定数据，求参数 θ 对应的概率：

$$L(\mu = 32, \sigma = 2.5 \mid w = 34) = 0.12$$
$$L(\mu = 34, \sigma = 2.5 \mid w = 34) = 0.21$$

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。

5.2 似然函数

似然函数（Likelihood function）

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

5.3 似然比

似然比（Likelihood ratio）

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

备注 5.2. 在 *Python* 中有如下模块：

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

5.4 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，miximizes the likelihood 找到最大化所有观察重量 likelihood 的分布，找到 mean 和 standard deviation

6 Chi-square distribution

假设有随机变量 X 服从标准正态分布，即有 $X \sim N(0,1)$，此时有随机变量 $Q_1 = X^2$，则有随机变量 Q_1 服从卡方分布（χ^2-distribution），由于此时只有一个随机变量，因此卡方分布自由度（degree of freedom）为 1，即 $Q_1 \sim \chi^2(1)$。如随机变量 $Q_2 = X_1^2 + X_2^2$，且 X_1 与 X_2 同时服从标准正态分布。则此时 Q_2 服从自由度为 2 的卡方分布，即 $Q_2 \sim \chi^2(2)$。

Goodness of fit

Pearson's chi-squared test

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

- O_i the number of observations of type i
- E_i the expected(theoretical) number of type i

7 时间序列

7.1 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 $(r_{t_1}, \ldots, r_{t_K})$ 与 $(r_{t_1+t}, \ldots, r_{t_k+t})$ 的联合分布相同，即将 $(t_1, \ldots, t_k)$ 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：

$$\mathbb{E}[r_t] = \mu$$
$$\mathrm{Cov}(r_t, r_{t-1}) = \gamma_l$$

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

$$\gamma_0 = \mathrm{Var}(r_t)$$
$$\gamma_{-l} = \gamma_l$$

证明. 性质二：

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

□

7.2 自相关函数

7.2.1 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：

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

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 $\bar{y} = \sum_{t=1}^T y_t/T$：

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

7.2.2 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 的相关系数称为 r_t 间隔为 l 的自回归系数：

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

由于 $\{r_t\}$ 为弱平稳收益率时间序列，应有 $\mathrm{Var}(r_t) = \mathrm{Var}(r_{t-l})$。根据定义，$\rho_0 = 1$ 且有 $-1 \le \rho_l \le 1$。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 l 的样本自相关系数为：

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

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 $\hat{p}_l$ 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 $r_t = \mu \sum_{i=0}^{q} \phi_i a_{t-i}$，其中 $\phi_0 = 1$，$\{a_i\}$ 为均值为 0 的独立同分布任意变量序列，对于 $l > q$，则有 $\hat{p}_l$ 渐进的服从如下正态分布，称为 Bartlett 公式：

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

检验单个 ACF

对于正整数 l，可以使用 t 检验（t-statistics）$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$ 来检验 $H_0 : \rho_l = 0$ 与 $H_a : \rho_l \neq 0$ 检验统计量为：

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

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，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。$Z_{\alpha/2}$ 为临界值（Critical value），临界值根据统计量的分布决定，由于统计量 $t_{\hat{\rho}}$ 渐进服从标准正态分布，则此时具体为标准正态分布的 $100(1-\alpha/2)$ 分位点。如此时 $\alpha = 5\%$，则有临界点为使得正态分布左侧阴影区域面积为 2.5% 与 97.5% 的两个 z 值，为 $z = \pm 1.96$。

自相关（Autocorrelation）或称序列相关（Serial correlation）

自相关函数（Autocorrelation function，ACF）

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF), $Corr(x_t, x_{t-k}), k = 1, 2, 3, \ldots$

Durbin-Watson test

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

Test statistic - resitual 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

Ljung-Box test

Test the null hypothesis that a series of residuals exhibits no autocorrelation for a fixed number of lags L. (See Box & Pierce 1970, Q test)

- H0: No residual autocorrelation - H1: There is residual autocorrelation

Test statistic

$$Q = T(T+2)\sum_{k=1}^{L}\frac{\rho(k)^2}{T-k} > \chi_L^2$$

- Q is chi-square with L degrees of freedom

7.3 自回归模型

自回归模型（Autoregressive model，AR）

向量自回归模型（Vector autoregressive model，VAR）

more than one random variable

7.4 移动平均模型

Moving-average (MA) model

7.5 自回归移动平均模型

自回归移动平均模型（Autoregressive moving-average model，ARMA）

7.6 差分自回归移动平均模型

差分自回归移动平均模型（Autoregressive integrated moving average，ARIMA）

7.7 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。

7.8 单位根

Dickey-Fuller test H0: there is unit root, $\delta = \rho - 1 = 0$, no stationary, random walk H1: stationay, mean and variance do not change over time

A simple AR(1) model $y_t = \alpha + \rho y_{t-1} + u_t$, then we have $\Delta y_t = \alpha + (\rho - 1)y_{t-1} + u_t = \alpha + \delta y_{t-1} + u_t$,

Augmented Dickey-Fuller

H0: there is unit root, $\delta = 0$ H1: stationary, $\delta < 0$

ADF test: $\Delta y_t = \alpha + \delta y_{t-1} + \beta_1 \Delta y_{t-1} + \cdots + \beta_p \Delta y_{t-p} + u_t$

AR(1) model: $\Delta y_t = \alpha + \delta y_{t-1} + u_t$ AR(2) model: $\Delta y_t = \alpha + \delta y_{t-1} + \beta \Delta y_{t-1} + u_t$

Test statistics: (negative, more negatvie -> reject H0)

- $DF_\delta = \frac{\hat{\delta}}{SE(\hat{\delta})}$

ADF-Test: 检验结果为负，越负越拒绝原假设，即有单位根。单位根存在，即 $y_t = a + by_{t-1} + e_t$，$|b| = 1$，为随机。

8 条件异方差模型

8.1 ARCH 效应检验

若将 $a_t = r_t - \mu_t$ 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。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 $\alpha_i = 0 (i = 1, \ldots, m)$，即原假设有：

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

线性回归为：

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

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：

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

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

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：

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

Ljung-Box Q-test residual autocorrelation

Breusch-Godfrey test

test for autocorrelation in the errors in a regression model

Breusch-Pagan test

used to test for heteroskedasticity in a linear regression model

备注 8.1. *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）

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

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

9 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：

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

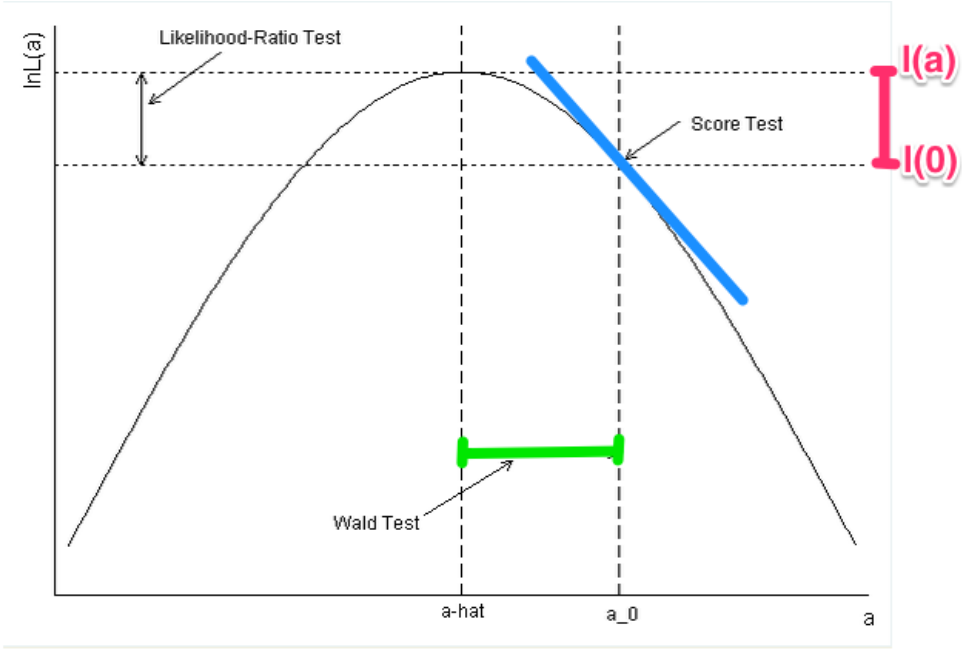


Figure based on a figure in Fox (1997, p. 570); used with authors permission.

图 3: Likelihood/Wald/Score Test

$$E[(\xi_t - \hat{\xi}_{t\,|\,t-1})(\xi_t - \hat{\xi}_{t\,|\,t-1})'] = \mathbf{P_{t\,|\,t-1}}$$
$$E[(\mathbf{y_t} - \hat{\mathbf{y}}_{t\,|\,t-1})(\mathbf{y_t} - \hat{\mathbf{y}}_{t\,|\,t-1})'] = \mathbf{H'P_{t\,|\,t-1}H + R}$$
$$E\{(\xi_t - \xi_{t\,|\,t-1})(\mathbf{y_t} - \hat{\mathbf{y}}_{t\,|\,t-1})'\} = \mathbf{P_{t\,|\,t-1}H}$$

计算步骤

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

迭代

平滑

$$\mathbf{J_t} \equiv \mathbf{P_{t\,|\,t}F'P^{-1}_{t+1\,|\,t}}$$
$$\hat{\xi_{t\,|\,T}} = \hat{\xi}_{t\,|\,t} + J(\hat{\xi}_{t+1\,|\,T} - \hat{\xi}_{t+1\,|\,t})$$