

统计整理

杨弘毅

创建：2020 年 4 月 9 日
修改：2021 年 10 月 21 日

目录

1	TODO	2
2	基础	2
2.1	期望	2
2.2	方差	3
2.3	协方差	4
2.4	相关系数	5
2.5	假设检验	5
3	矩	6
3.1	理解	6
3.2	定义	6
3.3	分类	7
3.4	矩母函数	10
3.4.1	定义	10
3.4.2	性质	11
4	条件概率	12
4.1	条件概率	12
4.2	条件概率分布	14
4.2.1	条件期望	15
4.2.2	条件方差	15

4.3	贝叶斯定理	17
4.4	贝叶斯因子	19
4.4.1	后验因子估计	19
4.4.2	发生率与贝叶斯因子	19
4.5	似然函数	20
5	Chi-square distribution	20
6	Probability vs Likelihood	21
6.1	Probability	21
6.2	Likelihood	21
6.3	Maximum likelihood	21
7	Time series	21

1 TODO

- 参数与非参数方法
- conditional probability / projection
- 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 (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：

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

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：

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

其性质有：

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

2.3 协方差

对于协方差（Covariance）其定义有：

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

性质有：

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

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:

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

2.5 假设检验

在假设检验 (Statistical hypothesis testing) 中, 原假设 (H_0 , null hypothesis), 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 (H_a , alternative hypothesis)。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 (α , significant level), 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 $1 - \alpha$ 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) ，那么有 $P(a < \text{随机变量} < b) = 1 - \alpha$ 。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 。

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。

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：

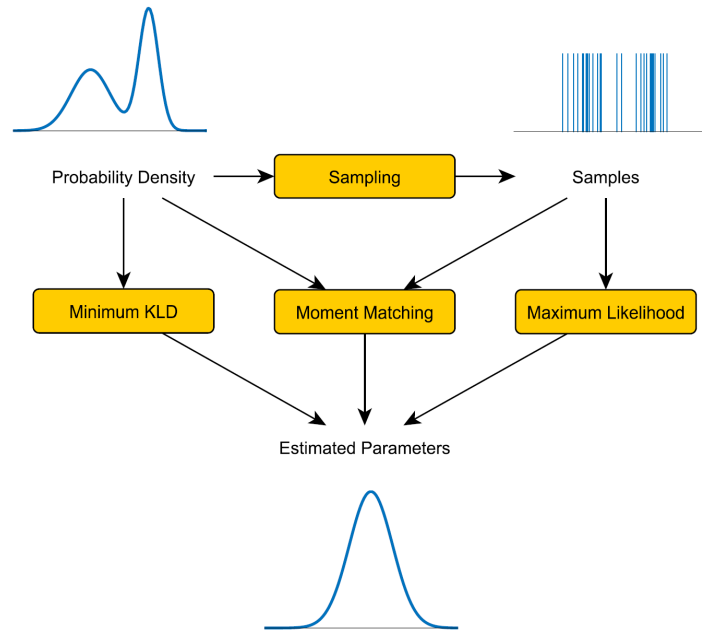


图 1: 矩匹配

- 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 $\sigma = (\mathbb{E}[(x - \mu)^2])^{\frac{1}{2}}$ 。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)**

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

其离散形式则有:

$$\text{Var}(x) = \sigma^2 = \sum p_i (x_i - \mu)^2$$

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**：

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

定义**超值峰度 (excess kurtosis)**为峰度 -3 ，使得正态分布的峰度为 0：

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

- 如果超值峰度为正，即峰度值大于 3，称为高狭峰（leptokurtic）
- 如果超值峰度为负，即峰度值小于 3，称为低阔峰（platykurtic）

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：

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

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 条件概率

4.1 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 4.1. 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。

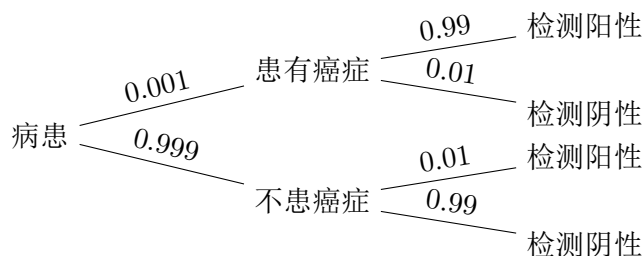
例子 4.2. 假设有两个小碗, 分别为甲碗与乙碗, 每个小碗中有若干小球, 共有蓝色与黄色两种颜色, 具体数量如下。

	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 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：

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

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 $P(\text{患病}|\text{检测阳性})$ 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。

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 $P(Y = y|X = x)P(X = x) = P[(X = x) \cap (Y = y)] = P(X = x|Y = y)P(Y = y)$ 。

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)：

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

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 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。

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

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 $\mathbb{E}(X|Y = 1) = \frac{1+0+0}{3} = \frac{1}{3}$ ，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 $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ ：

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

4.2.2 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 4.7. 总期望定律 (Law of total expectation) 或称为双重期望定理 (Double expectation theorem) 有：

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

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 4.8. 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 4.9. 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

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：

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

$P(A)$ 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 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 $P(\text{甲碗}|\text{蓝色})$ ，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：

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

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：

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

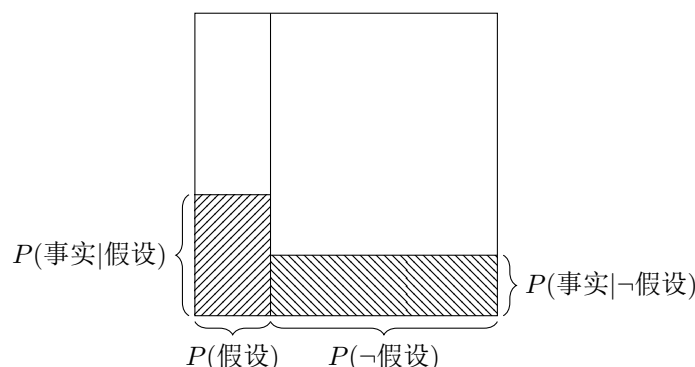
例子 4.11. 上述癌症诊断例子中，患病率为 0.1%，并且该诊断的正确率为 99%。即在病患中检测出阳性的概率为真阳性率或敏感率为 99%，而在健康人群中检测为阴性的概率同为 99%，或称为真阴性率，或特异性。根据贝叶斯定理有：

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

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：

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

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。

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：

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

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 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：

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

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。

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 $P(\text{假设})P(\text{事实}|\text{假设})$ 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：

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

4.5 似然函数

似然函数 (Likelihood function) likelihood ratio

5 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

6 Probability vs Likelihood

6.1 Probability

$P(\text{data} \mid \text{distribution}) = \text{area under curve}$

$P(\text{weight between 32g and 34g} \mid \text{mean} = 32 \text{ and standard deviation} = 2.5) = 0.29$

$P(\text{weight} > 34\text{g} \mid \text{mean} = 32 \text{ and standard deviation} = 2.5) = 0.21$

6.2 Likelihood

$L(\text{distribution} \mid \text{data}) = \text{value of the curve (y)}$

$L(\text{mean} = 32 \text{ and standard deviation} = 2.5 \mid \text{mouse weights 34g}) = 0.12$

$L(\text{mean} = 34 \text{ and standard deviation} = 2.5 \mid \text{mouse weights 34g}) = 0.21$

在调整了分布的 mean 之后, likelihood 最大, 在 mean=34 sigma=2.5 的正态分布中, 抽中一只 34g 的老鼠的概率最大

6.3 Maximum likelihood

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

7 Time series

Autoregressive (AR) model

vector autoregressive model (VAR) (more than one random variable)

Moving-average (MA) model

ARMA / ARIMA

autoregressive-moving-average (ARMA) / autoregressive integrated moving average (ARIMA)

TODO: Autocorrelation (serial correlation) - $cov(u_i, u_j) \neq 0$, for $i \neq j$ - some other estimator

will have a lower variance, no longer best estimate - Unit root processes, autoregressive processes, and moving average processes are specific forms of processes with autocorrelation.

Autocorrelation and Partial Autocorrelation

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

$$\rho_k = \frac{cov(x_t, x_{t-k})}{\sigma_{x_t} \sigma_{x_{t-k}}} = \frac{\gamma_k}{\gamma_0}$$

$$\gamma_k = \sum_{t=1}^{T-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) / T$$

$$\gamma_0 = \sum_{t=1}^T (x_t - \bar{x})^2 / T$$

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 - 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

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

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

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 negative \rightarrow reject H0)

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