统计学复习

司继春

[1] 上海对外经贸大学

2023年9月

概览

1. 随机变量
2. 多元随机变量
3. 条件期望
4. 大样本理论
5. 参数估计的基本概念
6. 区间估计
7. 矩估计
8. 极大似然估计
9. 假设检验的基本概念
10. 假设检验的一般步骤
11. 评价检验的方法
12. 构造假设检验的一般方法

一元随机变量

对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，样本空间$\Omega$可能是$\{\heartsuit, \spadesuit, \clubsuit, \diamondsuit\}$这样的集合，不便于研究。为了方便研究，我们通常将样本空间Ω映射到实轴$\mathbb{R}^*$上，并研究概率空间$(\mathbb{R}^*, \mathscr{B}, P)$。为此我们需要引入随机变量：

随机变量

对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，映射

$$X : \Omega \to \mathbb{R}^*$$

满足：对于任意的$B \in \mathscr{B}$，有：

$$X^{-1}(B) \triangleq \{\omega : X(\omega) \in B\} \in \mathscr{F}$$

那么我们称X为**随机变量**（**random variable, r.v.**）。

分布函数

分布函数

对于一个随机变量X，函数

$$F_X(x) = P_X(X \leq x) = \mathscr{P}\left(X^{-1}((-\infty, x])\right), \forall x \in \mathbb{R}$$

为一个分布函数（满足分布函数定义的要求），我们称其为**累积分布函数**（cumulative distribution function, c.d.f.）。

- 对于随机变量来说，**累积分布函数包含了所有概率函数P_X的信息**，因而使用P_X和使用累积分布函数F_X是等价的。
- 因而我们通常使用标记$X \sim F_X(x)$表示随机变量X服从F_X分布。
- 如果两个随机变量的累积分布函数$F_X(x) = F_Y(x)$，则我们称两个随机变量**同分布**（identically distributed），记为$X \sim Y$。

分布函数

泊松分布的分布函数

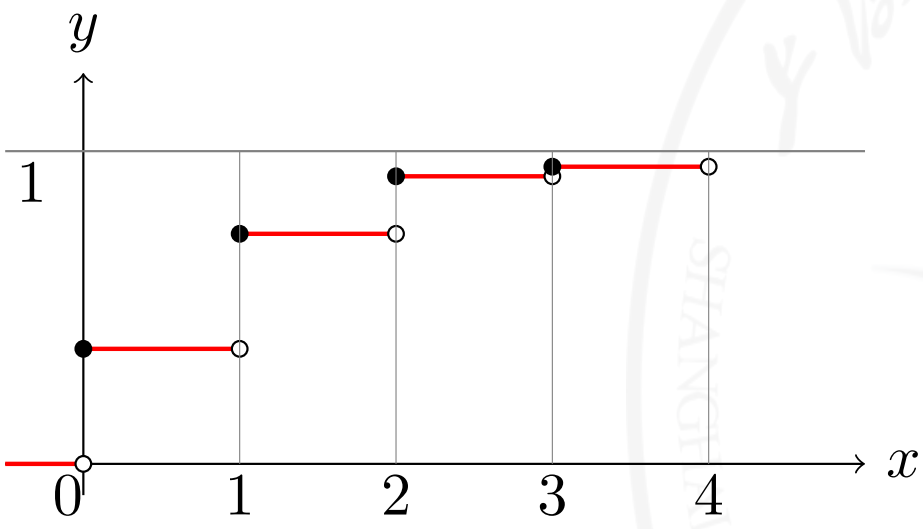
在建模一段时间内的到达次数时，我们经常使用所谓的泊松分布，即到达次数只可能取自然数值，或者$P(X \in \mathbb{Z}) = 1$，由于自然数集$\mathbb{Z}$为可数集，因而该分布是一个离散分布，且其概率函数为：

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

其中λ为这段时间的平均到达次数。其分布函数为：

$$F_X(x) = P(X \leq x) = \sum_{i=0}^{x} \frac{\lambda^i}{i!} e^{-\lambda}$$

泊松分布的分布函数



分布函数

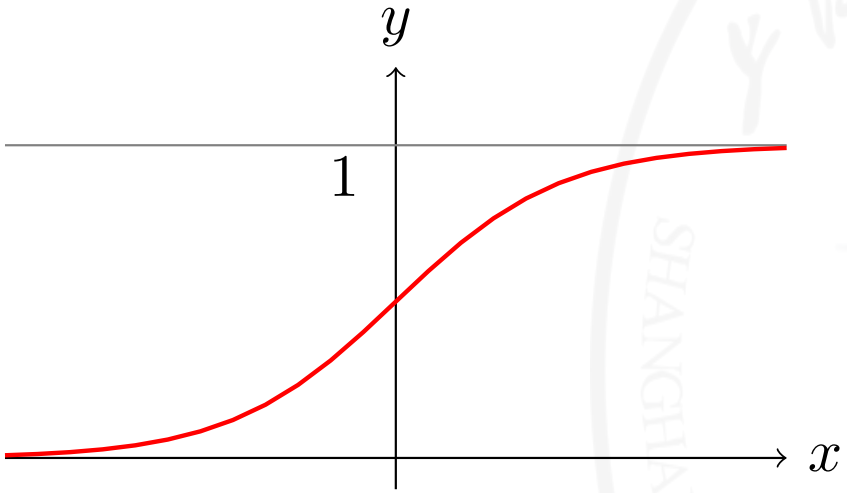
Logistic分布

如果一个随机变量X的分布函数为:

$$F_X(x) = \frac{e^x}{1+e^x}$$

那么我们称X服从Logistic分布

Logistic分布的分布函数



概率密度函数

概率密度函数

对于连续型随机变量，**概率密度函数**（**probability density function, p.d.f**），$f_X(x)$定义为：

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt, \text{for all } x$$

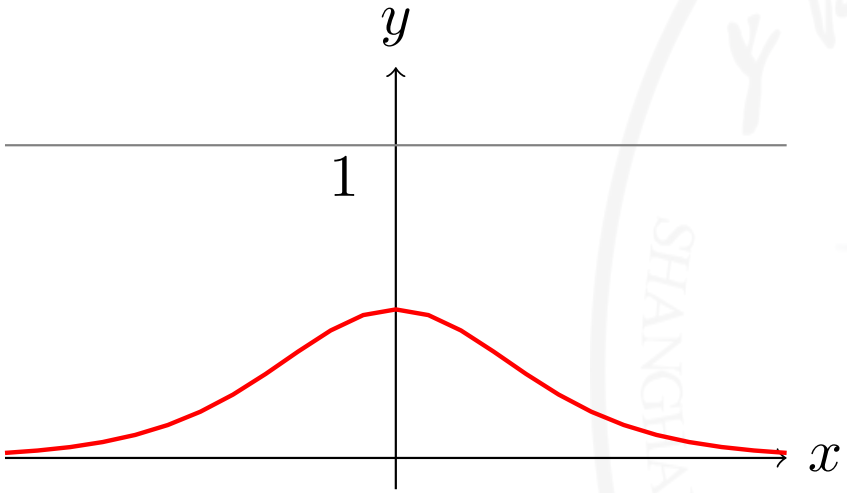
注意如果分布函数$F_X(x)$可导，那么其密度函数即其导函数。

Logistic分布的密度函数

Logistic分布的p.d.f为其分布函数的导函数：

$$f_X(x) = \frac{e^x}{(1+e^x)^2}$$

Logistic分布的密度函数



离散型随机变量的期望

离散型随机变量期望

对于离散型随机变量$X \in B$，其数学期望（mathematical expectation）定义为：

$$\mathbb{E}(X) = \sum_{b \in B} b \cdot P(X = b)$$

离散型随机变量期望

以上定义可以使用Riemann-Stieltjes积分的形式表达：

$$\mathbb{E}(X) = \int_{\mathbb{R}} x dF_X(x)$$

离散型随机变量的期望

泊松分布的期望

泊松分布随机变量的期望为:

$$
\begin{aligned}
\mathbb{E}(X) &= \sum_{j=0}^{\infty} j \cdot P(X=j) \\
&= \sum_{j=0}^{\infty} j \cdot \frac{\lambda^j}{j!} e^{-\lambda} \\
&= \lambda \sum_{j=1}^{\infty} \frac{\lambda^{j-1}}{(j-1)!} e^{-\lambda} \\
&= \lambda
\end{aligned}
$$

期望

期望

如果概率空间$(\Omega, \mathscr{F}, \mathscr{P})$导出了概率空间$(\mathbb{R}, \mathscr{B}, P)$，令$g$为一个可测函数，则期望：

$$\mathbb{E}\left(g\left(X\right)\right) = \int_{\mathbb{R}} g\left(x\right) P\left(dx\right) = \int_{\mathbb{R}} g\left(x\right) dF_X\left(x\right)$$

<u>如果等式两边积分都存在</u>。如果额外的，密度函数存在，则：

$$\mathbb{E}\left(g\left(X\right)\right) = \int_{\mathbb{R}} g\left(x\right) f_X\left(x\right) dx$$

期望的线性性

对于两个$(\Omega, \mathscr{F}, \mathscr{P})$上的随机变量$X, Y$，如果期望存在，有：

$$\mathbb{E}\left(aX + bY\right) = a\mathbb{E}\left(X\right) + b\mathbb{E}\left(Y\right)$$

可积性

可积性

积分存在要求：

$$\mathbb{E}\left(|X|\right) < \infty$$

如果上式不成立，则积分不存在，从而期望也不存在。

期望的存在性

积分不存在的例子：Cauchy分布

Cauchy的密度函数为：$f(x)=\frac{1}{\pi(1+x^2)}$如果一个随机变量服从Cauchy分布，则：

$$\mathbb{E}(|X|)=\int_{\mathbb{R}}\frac{|x|}{\pi}\frac{1}{1+x^2}dx=\frac{2}{\pi}\int_{[0,\infty)}\frac{x}{1+x^2}dx$$

对于任意正数M:

$$\int_{[0,M]}\frac{x}{1+x^2}dx=\left.\frac{\log\left(1+x^2\right)}{2}\right|_0^M=\frac{\log\left(1+M^2\right)}{2}$$

因此：$\mathbb{E}(|X|)=\frac{1}{\pi}\lim_{M\to\infty}\log\left(1+M^2\right)=\infty$，因而该随机变量是不可积的，期望不存在。

矩

矩和中心矩

- 对于任意的正数p，如果$\mathbb{E}\left(|X|^p\right) < \infty$，则记$X \in L^p = L^p\left(\Omega, \mathscr{F}, \mathscr{P}\right)$。
- 对于整数$r < p$，随机变量X的r阶矩被定义为$\mathbb{E}\left(X^r\right)$。一阶矩即为随机变量$X$的期望。
- 此外，随机变量X的r阶中心矩被定义为$\mathbb{E}\left([X - E(X)]^r\right)$。

方差和标准差

方差和标准差

2阶中心矩：

$$\mathbb{V}(X) = \mathbb{E}[X - \mathbb{E}(X)]^2$$

即为随机变量的方差（variance），记为$\mathbb{V}(X)$或者$\sigma^2(X)$，标准差（standard deviation）定义为$\sigma(X) = \sqrt{\mathbb{V}(X)}$。

方差的性质

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}\left([X-\mathbb{E}(X)]^2\right) = \mathbb{E}\left(X^2 - 2\mathbb{E}(X)\cdot X + \mathbb{E}(X)^2\right)\\ &= \mathbb{E}\left(X^2\right) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2\end{aligned}$$

$$\begin{aligned}\mathbb{V}(aX+b) &= \mathbb{E}[aX+b-\mathbb{E}(aX+b)]^2\\ &= \mathbb{E}[aX+b-a\mathbb{E}(X)-b]^2\\ &= a^2\mathbb{E}(X-\mathbb{E}X)^2\\ &= a^2\mathbb{V}(X)\end{aligned}$$

偏度

偏度（skewness）为随机变量的三阶中心矩，即定义：

$$\gamma = \mathbb{E}\left(\left[\frac{X - \mathbb{E}(X)}{\sigma(X)}\right]^3\right) = \frac{\mathbb{E}\left[X - \mathbb{E}(X)\right]^3}{\sigma^3(X)}$$

- 如果X为对称分布，那么必然有$\gamma = 0$。
- 当$\gamma > 0$时，分布函数右边的尾巴比较厚，我们称其分布为右偏（positive skew）分布
- 反之为左偏（negative skew）分布。

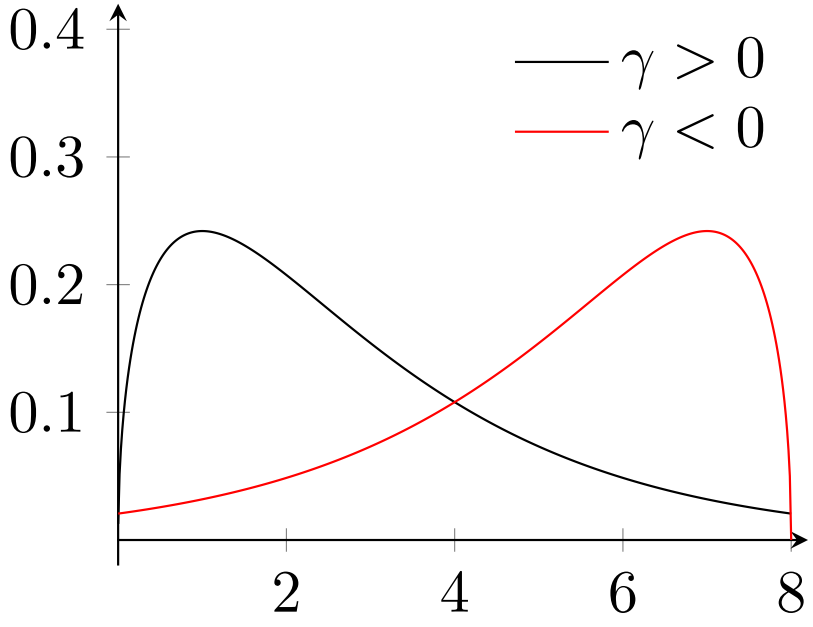
峰度

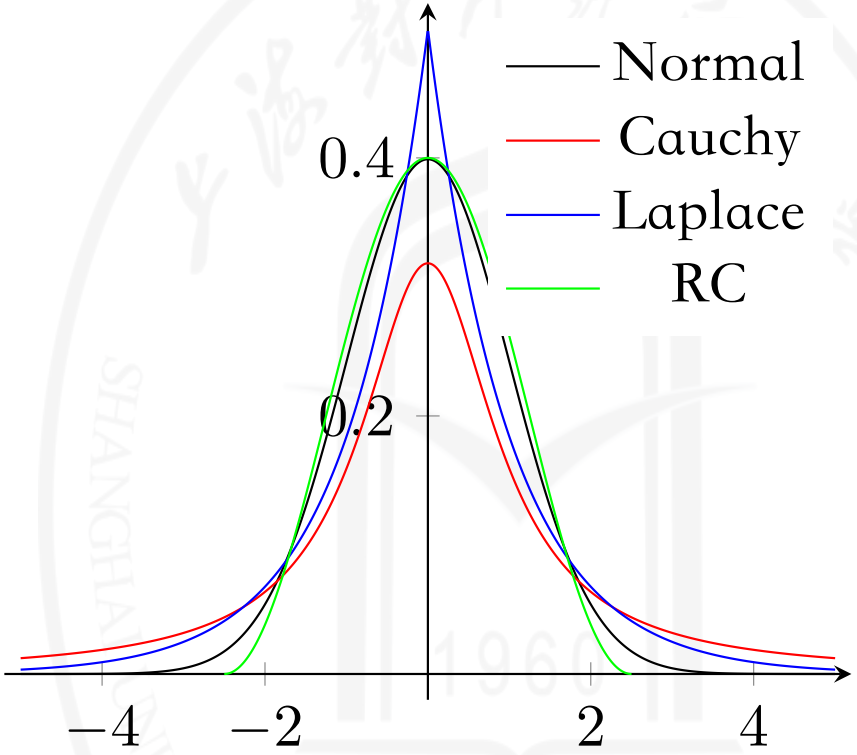
峰度是随机变量的四阶中心矩，即：

$$\text{Kurt}(X) = \mathbb{E}\left(\left[\frac{X - \mathbb{E}(X)}{\sigma(X)}\right]^4\right) = \frac{\mathbb{E}\left([X - \mathbb{E}(X)]^4\right)}{[\mathbb{V}(X)]^2}$$

- 尽管我们一般将其称为「峰度」，然而这一称呼并不准确，更准确的称呼应该为「尾厚度」。
- 正态分布的峰度为3。
- 我们一般会把峰度与正态分布的峰度相比，定义超额峰度（excess kurtosis）为峰度减3
- 因而如果超额峰度>0，那么其尾巴比正态分布的尾巴要厚，而如果超额峰度<0，那么其尾巴要比正态分布的尾巴要薄。

偏度和峰度


0.4
0.3
0.2
0.1
2
4
6
8
$\gamma > 0$
$\gamma < 0$


0.4
0.2
−4
−2
2
4
Normal
Cauchy
Laplace
RC

Chebyshev不等式

Chebyshev不等式

如果函数ψ满足：$\psi(u)=\psi(-u)\geq 0$，且在$(0,\infty)$上单调递增，X为随机变量，且$\psi(X)<\infty$，则对于$u>0$，有：

$$P(|X|\geq u)\leq \frac{\mathbb{E}[(\psi(X))]}{\psi(u)}$$

证明

$$\begin{aligned}\mathbb{E}[\psi(X)] &= \int_{\mathbb{R}} \psi(X)\, P(dX) \\ &\geq \int_{\{|X|\geq u\}} \psi(X)\, P(d\omega) \\ &\geq \psi(u)\, P(|X|\geq u)\end{aligned}$$

Chebyshev不等式

Chebyshev不等式应用

令

$$Y = \frac{[X - \mathbb{E}(X)]}{\sigma(X)}$$

则$\mathbb{E}(Y) = 0$，$\mathbb{E}\left(Y^2\right) = 1$。令$\psi(x) = x^2$，有：

$$P(|Y| \geq u) \leq \frac{1}{u^2}$$

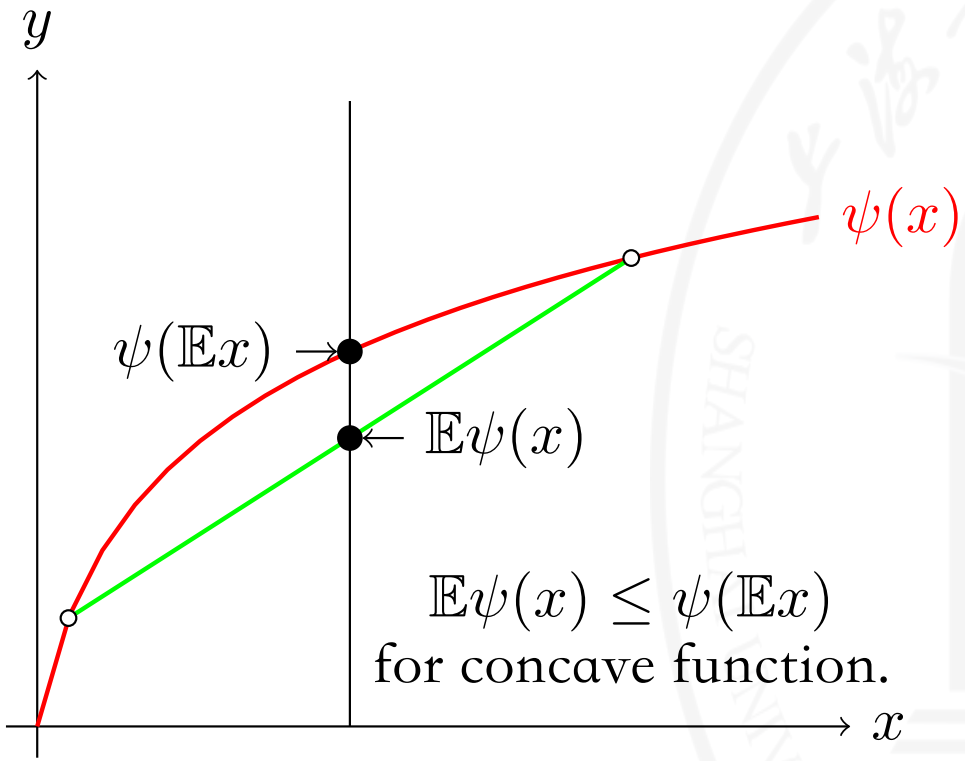
Jensen不等式

Jensen不等式

如果ψ为凹函数，且随机变量X和$\psi(X)$可积，则：

$$\psi(\mathbb{E}(X)) \geq \mathbb{E}[\psi(X)]$$

Jensen不等式


y
$\psi(x)$
$\psi(\mathbb{E}x)$
$\mathbb{E}\psi(x)$
$\mathbb{E}\psi(x) \leq \psi(\mathbb{E}x)$
for concave function.
x

伯努利分布

伯努利分布即伯努利试验（Bernoulli trial）的结果。伯努利试验即试验结果只有两种可能性（成功，失败），定义为$X = 1\ if$ 成功, $= 0\ if$ 失败。如果成功的概率为p，那么伯努利分布（**Bernoulli Distribution**）$X \sim Ber\,(p)$即：

$$X = \begin{cases} 1 & with\ probability\ p \\ 0 & with\ probability\ 1-p \end{cases} \quad 0 \le p \le 1$$

- 期望：$\mathbb{E}\,(X) = p$
- 方差：$\mathbb{V}\,(X) = p\,(1-p)$

正态分布

如果一个随机变量X潜在受到非常多的独立因素的影响，即$X = f_1 + f_2 + \cdots$，而每个f_i又不能单独对X有非常大的影响，那么一般来说X将会服从**正态分布**（**normal distribution**）或者**高斯分布**（**Gaussian distribution**）。正态分布的密度函数为

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

记为$X \sim N\left(\mu,\sigma^2\right)$。

- 期望：$\mathbb{E}\left(X\right) = \mu$
- 方差：$\mathbb{V}\left(X\right) = \sigma^2$

正态分布

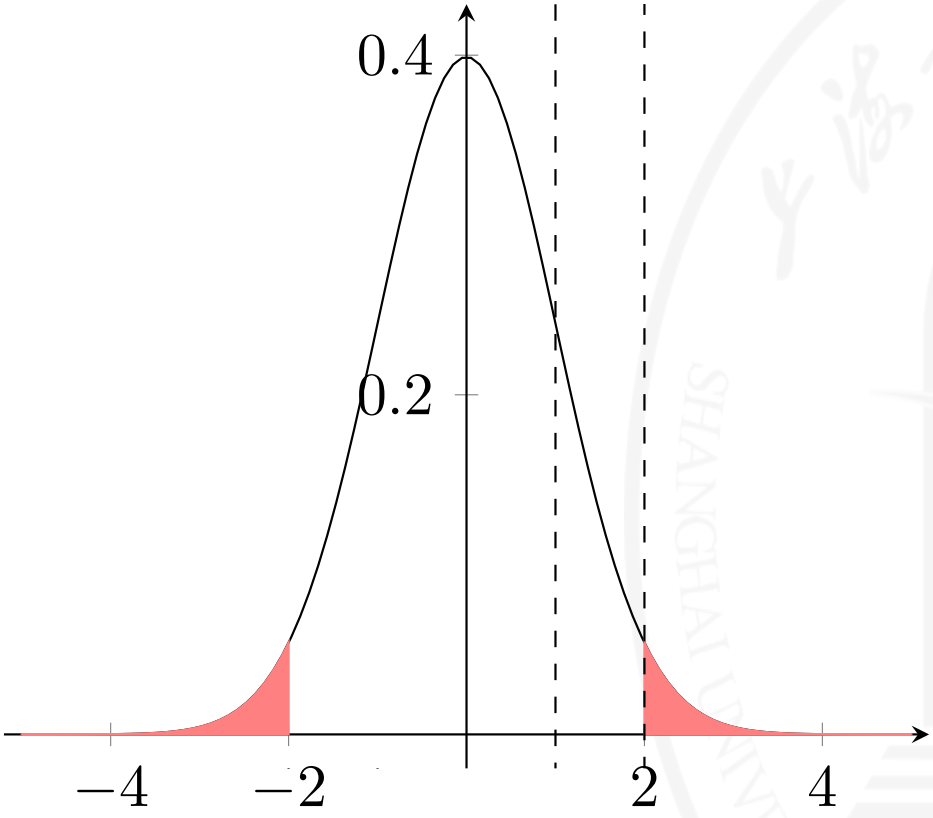
- 如果$X \sim N\left(\mu, \sigma^2\right)$，如果令$Z = \frac{X-\mu}{\sigma}$，则$E\left(Z\right) = \frac{E(X)-\mu}{\sigma} = 0$，$\mathbb{V}\left(Z\right) = \frac{1}{\sigma^2} \cdot \mathbb{V}\left(X\right) = 1$，随机变量$Z$服从$\mu = 0, \sigma = 1$的正态分布，即$Z \sim N\left(0, 1\right)$，我们称之为**标准正态分布**（standard normal distribution）。
- 标准正态分布的密度函数为：

$$\phi\left(x\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- 而其分布函数为：

$$\Phi\left(x\right) = \int_{-\infty}^{x} \phi\left(t\right) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$ 由于$\Phi\left(x\right)$没有初等函数表示，因而一般我们用$\Phi\left(z\right)$来代表标准正态的分布函数。

正态分布



正态分布

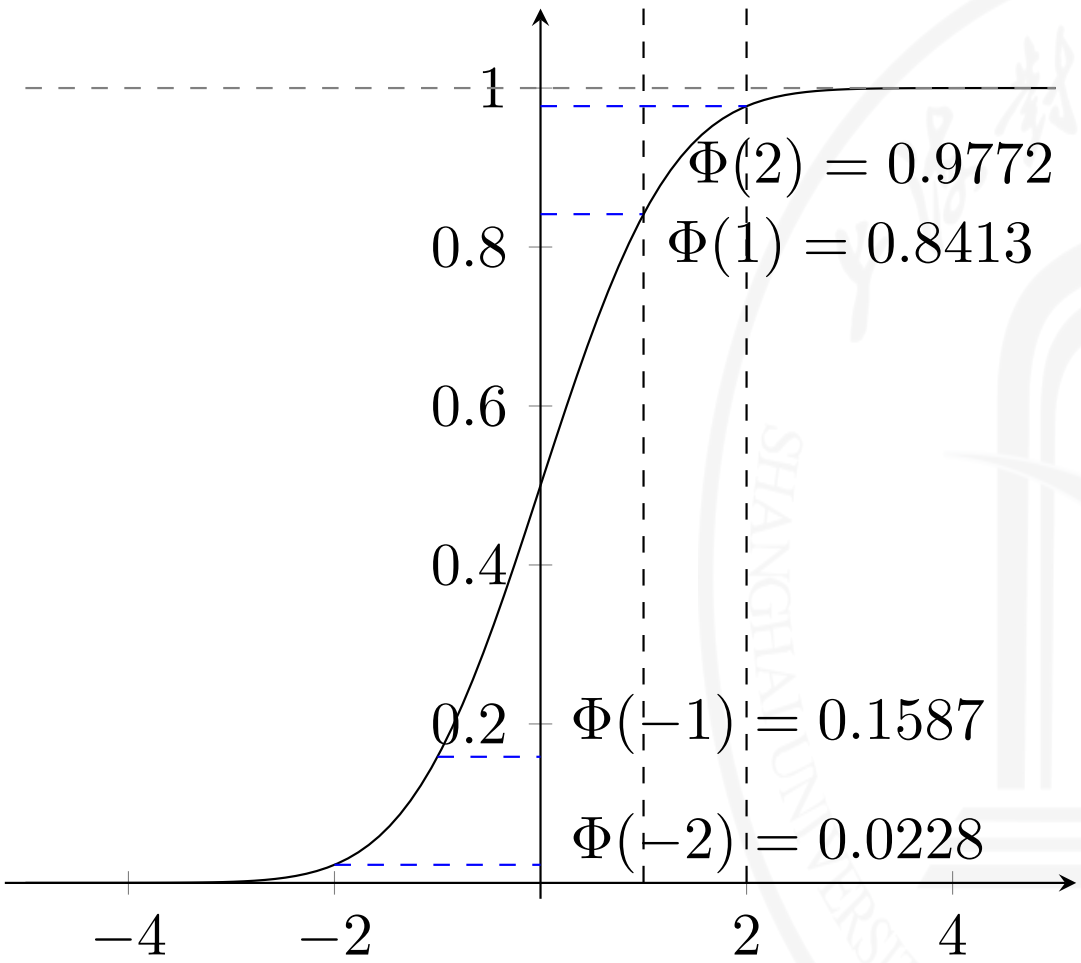
- 由于正态分布为对称分布，即$\phi(x)=\phi(-x)$，因而分布函数$\Phi(x)=1-\Phi(-x)$。
- 如果$X\sim N\left(\mu,\sigma^2\right)$，那么$Z=\frac{X-\mu}{\sigma}\sim N(0,1)$，根据标准正态的分布函数$\Phi(x)$可以计算$X$在区间内取值的概率，比如：

$$P(|X-\mu|\leq\sigma)=P(|Z|<1)=\Phi(1)-\Phi(-1)=2\Phi(1)-1=0.6827$$

同理，有：

$$\begin{aligned}
P(|X-\mu|\leq 1.65\sigma)&=P(|Z|<1.65)\approx 0.90\\
P(|X-\mu|\leq 1.96\sigma)&=P(|Z|<1.96)\approx 0.95\\
P(|X-\mu|\leq 2\sigma)&=P(|Z|<2)=0.9545\\
P(|X-\mu|\leq 2.58\sigma)&=P(|Z|<2.58)\approx 0.99\\
P(|X-\mu|\leq 3\sigma)&=P(|Z|<3)=0.9973\\
P(|X-\mu|\leq 5\sigma)&=P(|Z|<5)\geq 1-10^{-6}\\
P(|X-\mu|\leq 6\sigma)&=P(|Z|<6)\geq 1-10^{-8}
\end{aligned}$$

正态分布



对数正态分布

若随机变量$Y = e^X, X \sim N\left(\mu, \sigma^2\right)$，则我们称随机变量$Y$服从对数正态分布（lognormal distribution），记为$Y \sim LN\left(\mu, \sigma^2\right)$。

- 期望：$\mathbb{E}\left(Y\right) = e^{\mu + \frac{\sigma^2}{2}}$
- 方差：$\mathbb{V}\left(Y\right) = e^{2\left(\mu + \sigma^2\right)} - e^{2\mu + \sigma^2}$

χ^2分布

K个独立的标准正态分布的平方和的分布被称为卡方分布（Chi-square distribution）或者χ^2分布。即，如果$X_1, X_2, ..., X_K$为K个独立的标准正态分布，那么

$$X = \sum_{i=1}^{K} X_i^2 \sim \chi^2(K)$$

其中参数K为卡方分布的自由度（degrees of freedom）。

- 期望：$\mathbb{E}(X) = K$
- 方差：$\mathbb{V}(X) = 2K$

F分布

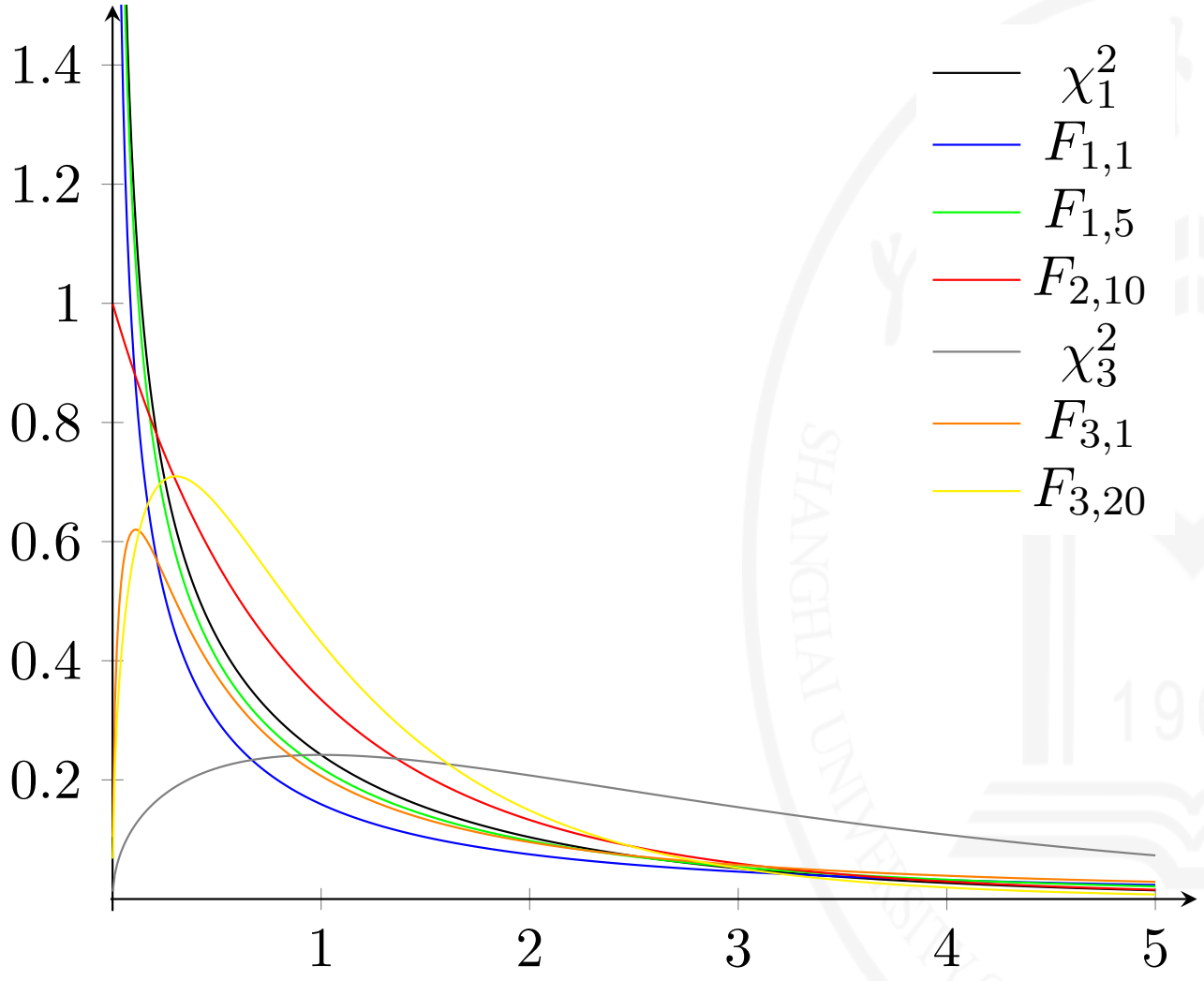
如果存在两个独立的χ^2分布$X_1 \sim \chi^2(n), X_2 \sim \chi^2(m)$，那么随机变量

$$F = \frac{X_1/n}{X_2/m}$$

即服从F分布（F-distribution），记为$F \sim F(n, m)$，参数n, m为自由度。

- 期望：$\mathbb{E}(X) = \frac{m}{m-2}, m > 2$
- 方差：$\mathbb{V}(X) = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}, m > 4$

χ^2分布与F分布



t分布

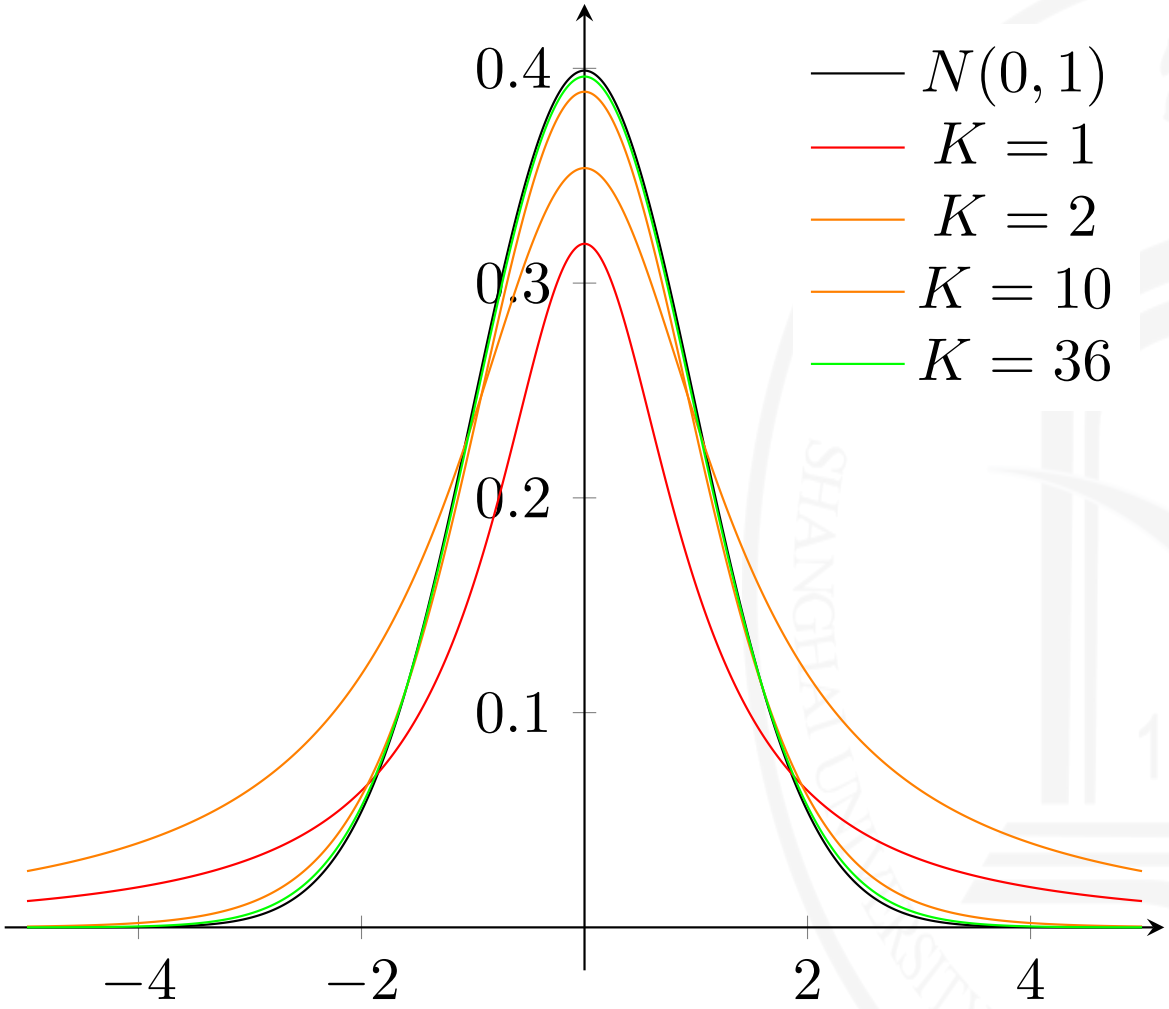
如果存在一个标准正态分布$Z \sim N(0,1)$，以及一个χ^2分布$X \sim \chi^2(K)$，且Z和X独立，那么随机变量

$$T = \frac{Z}{\sqrt{X/K}}$$

即服从学生氏t分布（Students' t-distribution）或者简称t分布（t-distribution），记为$T \sim t(K)$，参数K为自由度。

- 期望：$\mathbb{E}(X) = 0, K > 1$
- 方差：$\mathbb{V}(X) = \frac{K}{K-2}, K > 2$

标准正态分布与t分布



柯西分布

若随机变量X的概率密度函数为:

$$f\left(x|\mu,\sigma\right)=\frac{1}{\pi\sigma}\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}$$

那么我们称X服从**柯西分布**（**Cauchy distribution**），记为$X\sim C\left(\mu,\sigma\right)$。

- 期望：不存在
- 方差：不存在

由于柯西分布的一阶矩不存在，因而通常作为不可积的分布的例子。此外，可以得到，两个独立的标准正态分布X,Y，其比例$U=\frac{Y}{X}$刚好服从柯西分布。此外，$C\left(0,1\right)\sim t\left(1\right)$。

逻辑斯蒂分布

若随机变量X的分布函数为：

$$F\left(x|\mu,\sigma\right)=\frac{e^{\frac{x-\mu}{\sigma}}}{1+e^{\frac{x-\mu}{\sigma}}}$$

那么我们称X服从**逻辑斯蒂分布**（logistic distribution），其密度函数为：

$$f\left(x|\mu,\sigma\right)=\frac{1}{\sigma}\frac{e^{-\frac{x-\mu}{\sigma}}}{\left[1+e^{-\frac{x-\mu}{\sigma}}\right]^2}$$

记为：$X\sim LG\left(\mu,\sigma\right)$。

- 期望：$\mathbb{E}\left(X\right)=\mu$
- 方差：$\mathbb{V}\left(X\right)=\sigma^2\frac{\pi^2}{3}$

多元随机变量

多元随机变量的定义

（随机向量）给定一个概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，一个$n$维的随机向量$X$即从样本空间到$n$维欧几里得空间的函数，$X : \Omega \to \mathbb{R}^n$。

向量表达

注意以上定义我们使用了向量的表达方式，即：

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}, x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

联合分布函数

联合分布函数

由$(\Omega, \mathscr{F}, \mathscr{P})$导出的概率空间$(\mathbb{R}^n, \mathscr{B}^n, P)$的**联合分布函数**（joint c.d.f.）定义为：

$$
\begin{aligned}
F(x) &= F(x_1, x_2, ..., x_n) \\
&= P(X_1 \le x_1, ..., X_n \le x_n) \\
&= P((-\infty, x_1] \times (-\infty, x_2] \times \cdots (-\infty, x_n]) \\
&= \mathscr{P}\left(X^{-1}((-\infty, x_1] \times (-\infty, x_2] \times \cdots (-\infty, x_n])\right)
\end{aligned}
$$

$\forall x \in \mathbb{R}^n$。

联合分布函数为单调递增

且$F(-\infty, -\infty, ..., -\infty) = 0$，$F(\infty, \infty, ..., \infty) = 1$

联合密度函数

联合密度函数

1. 如果随机向量X的每个分量都是离散型随机变量，那么可以定义联合概率质量函数p.m.f为：

$$f\left(x_1, x_2, ..., x_n\right) = P\left(\left\{X_1 = x_1, ..., X_n = x_n\right\}\right)$$

2. 如果随机变量X的联合分布函数连续，如果函数$f\left(x\right)$满足：$P\left(X \in A\right) = \int_A f\left(x\right) dx, x \in \mathbb{R}^n, A \in \mathscr{B}^n$那么我们称$f\left(x\right)$为其联合概率密度函数p.d.f。特别的，如果联合分布函数$F\left(x\right)$可微那么：

$$f\left(x_1, x_2, ..., x_n\right) = \frac{\partial^n F\left(x_1, x_2, ..., x_n\right)}{\partial x_1 \partial x_2 \cdots \partial x_n}$$

边缘分布

- 如果$X=(X_1,...,X_n)$为随机向量，那么$\widetilde{X}=(X_{i_1},X_{i_2},...,X_{i_k}),1\leq i_1<i_2<...<i_k\leq n$也是一个随机向量。$\widetilde{X}$的联合分布函数可以通过$F(x)$来定义，即令$F(x)$中满足$j\notin\{i_1,...i_k\}$的分量为$\infty$。
- 如对于三维随机变量$X=(X_1,X_2,X_3)$，则$\widetilde{X}=(X_1,X_2)$的分布函数为：$F_{\widetilde{X}}(\widetilde{x})=F(\widetilde{x}_1,\widetilde{x}_2,\infty)$。
- 特别的，对于随机向量X的每个分量X_i，我们可以定义其边缘分布函数（marginal c.d.f.）为：

$$F_{X_i}(x_i)=F(\infty,...,x_i,...,\infty)$$

- 意边缘分布函数对应着一元随机变量X_i的分布函数。

边缘密度函数

对于随机向量$X=(X_1,...,X_n)$，其联合分布函数为$F(x)=F(x_1,x_2,...,x_n)$，那么：

- 联合密度函数为：

$$f(x)=\frac{\partial^n F(x_1,x_2,...,x_n)}{\partial x_1\partial x_2\cdots\partial x_n}$$

- 边缘密度函数：

$$f_{X_i}(x_i)=\frac{\partial F_{X_i}(x_i)}{\partial x_i}=\int_{\mathbb{R}}\cdots\int_{\mathbb{R}}f(x_1,x_2,...,x_n)\,dx_1\cdots dx_{i-1}dx_{i+1}$$

- 边缘分布函数：

$$\begin{aligned}F_{X_i}(a)&=F(\infty,...,x_i=a,...,\infty)\\&=\int_{-\infty}^{a}f_{X_i}(x_i)\,dx_i\end{aligned}$$

多元随机变量的期望

- 多元随机变量分量的期望与一元随机变量的期望定义相同
- 我们通常吧随机向量的期望写为向量形式：

$$\mathbb{E}(X) = \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_n) \end{bmatrix}$$

- 对于常数矩阵A，有：$\mathbb{E}(AX) = A\mathbb{E}(X)$
- 在这里期望的线性性仍然成立，比如，如果$\iota = (1, 1, ..., 1)'$为全部由1构成的向量，那么：

$$\mathbb{E}(\iota' X) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i)$$

协方差

如果对于两个一元随机变量Y, Z，如果$\mathbb{E}|Y|^2 < \infty, \mathbb{E}|Z|^2 < \infty$，根据Cauchy-Schwarz 不等式，$\mathbb{E}|YZ| \leq \sqrt{\mathbb{E}|Y|^2 \mathbb{E}|Z|^2} < \infty$，即$YZ$可积，我们可以定义两个随机变量的**协方差**（**Covariance**）：

$$
\begin{aligned}
\mathbb{C}\mathsf{ov}(Y, Z) &= \mathbb{E}[(Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))] \\
&= \mathbb{E}[YZ - \mathbb{E}(Y)Z - Z\mathbb{E}(Y) + \mathbb{E}(Y)\mathbb{E}(Z)] \\
&= \mathbb{E}(YZ) - 2\mathbb{E}(Y)\mathbb{E}(Z) + \mathbb{E}(Y)\mathbb{E}(Z) \\
&= \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z)
\end{aligned}
$$

当$Y = Z$时，$\mathbb{C}\mathsf{ov}(Y, Y) = \mathbb{E}\left(Y^2\right) - [\mathbb{E}(Y)]^2 = \mathbb{V}(Y)$。

相关系数

进而可以使用协方差定义相关系数（correlation coefficient）：

$$\rho_{Y,Z} = \frac{\mathbb{C}\mathsf{ov}\left(Y,Z\right)}{\sqrt{\mathbb{V}\left(Y\right)\mathbb{V}\left(Z\right)}}$$

由于：

$$\begin{aligned}
\mathbb{C}\mathsf{ov}\left(Y,Z\right) &= \mathbb{E}\left[\left(Y-\mathbb{E}\left(Y\right)\right)\left(Z-\mathbb{E}\left(Z\right)\right)\right] \\
&\leq \mathbb{E}\left|\left(Y-\mathbb{E}\left(Y\right)\right)\left(Z-\mathbb{E}\left(Z\right)\right)\right| \\
&\leq \sqrt{\mathbb{E}\left|\left(Y-\mathbb{E}\left(Y\right)\right)\right|^2\mathbb{E}\left|Z-\mathbb{E}\left(Z\right)\right|^2} \\
&= \sqrt{\mathbb{V}\left(Y\right)\mathbb{V}\left(Z\right)}
\end{aligned}$$

可知$-1 \leq \rho_{Y,Z} \leq 1$。

相关系数

- 如果$\rho_{Y,Z} = \pm 1$，那么$P\left(Y = c_1 Z + c_2\right) = 1, c_1 \neq 0$;
- 如果$\rho_{Y,Z} > 0$，我们称随机变量Y和Z正相关，反之成为负相关;
- 如果$\rho_{Y,Z} = 0$，我们称随机变量Y和Z不相关（uncorrelated）。
- 这里所谓的「相关系数」特指皮尔森相关系数（Pearson correlation coefficient），实际上只度量了随机变量之间的线性相关性。

相关系数

联合分布于边缘分布

如果随机变量$Y = Z^2$，$Z \sim N(0,1)$，那么：

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Z,Y) &= \mathbb{E}ZY - \mathbb{E}Z\mathbb{E}Y \\ &= \mathbb{E}Z^3 \\ &= 0\end{aligned}$$

两者相关系数为0，然而显然两者存在着非线性的函数关系。

协方差

联合密度函数

如果a, b为任意实数，Y和Z为一元随机变量，那么：

$$\begin{aligned}\mathbb{V}(aY+bZ) &= \mathbb{E}(aY+bZ)^2 - [a\mathbb{E}(Y)+b\mathbb{E}(Z)]^2 \\ &= \mathbb{E}\left(a^2Y^2+b^2Z^2+2abYZ\right) \\ &\quad - \left[a^2(\mathbb{E}(Y))^2+b^2(\mathbb{E}(Z))^2+2ab\mathbb{E}(Y)\mathbb{E}(Z)\right] \\ &= a^2\mathbb{V}(Y)+b^2\mathbb{V}(Z)+2ab\mathbb{C}\mathsf{ov}(Y,Z)\end{aligned}$$

如果Y, Z不相关，那么

$$\mathbb{V}(aY+bZ) = a^2\mathbb{V}(Y)+b^2\mathbb{V}(Z)$$

协方差矩阵

如果对于一个随机向量：$X=(X_1,X_2,...,X_n)'$，我们可以定义矩阵：

$$\Sigma=\mathbb{V}(X)=[\mathbb{C}\mathrm{ov}(X_i,X_j)]$$
$$=\begin{bmatrix} \mathbb{V}(X_1) & \mathbb{C}\mathrm{ov}(X_1,X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_1,X_n) \\ \mathbb{C}\mathrm{ov}(X_2,X_1) & \mathbb{V}(X_2) & \cdots & \mathbb{C}\mathrm{ov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}(X_n,X_1) & \mathbb{C}\mathrm{ov}(X_n,X_2) & \cdots & \mathbb{V}(X_n) \end{bmatrix}$$

易知协方差矩阵为实对称矩阵。

协方差矩阵的计算

- 根据协方差矩阵的定义，协方差矩阵可以如下计算：

$$\mathbb{V}(X) = \mathbb{E}\left([X - \mathbb{E}(X)][X - \mathbb{E}(X)]'\right)$$

注意X为列向量，从而：

$$X - \mathbb{E}(X) = \begin{bmatrix} X_1 - \mathbb{E}(X_1) \\ \vdots \\ X_n - \mathbb{E}(X_n) \end{bmatrix}$$

从而：$[X - \mathbb{E}(X)][X - \mathbb{E}(X)]' =$

$$\begin{bmatrix} (X_1 - \mathbb{E}(X_1))^2 & (X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)) & \cdots & (X_1 - \mathbb{E}(X_1))(X_n - \mathbb{E} \\ (X_2 - \mathbb{E}(X_2))(X_1 - \mathbb{E}(X_1)) & (X_2 - \mathbb{E}(X_2))^2 & \cdots & (X_2 - \mathbb{E}(X_2))(X_n - \mathbb{E} \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \mathbb{E}(X_n))(X_1 - \mathbb{E}(X_1)) & (X_n - \mathbb{E}(X_n))(X_2 - \mathbb{E}(X_2)) & \cdots & (X_n - \mathbb{E}(X_n))^2 \end{bmatrix}$$

求期望即可。

协方差矩阵的性质

与方差、协方差的性质类似，协方差有如下性质：

- $\mathbb{V}(X) = \mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')$
- $\mathbb{V}(AX+b) = A\mathbb{V}(X)A'$，其中$A$为常数矩阵，$b$为常数向量
- 协方差矩阵为实对称矩阵，且为半正定矩阵
  - 对于任意常数向量a，有：$a'\mathbb{V}(X)a = \mathbb{V}(a'X) \geq 0$
- 如果存在一个向量a和常数b使得$a'X = b$，那么$\mathbb{V}(X)$不满秩。
  - 对于一个实对称半正定矩阵，其特征值一定大于等于0，如果$a'X = b$，那么一定有特征值等于0，从而不满秩。

随机向量的独立性

随机向量的独立性

随机向量$(X_1, ..., X_n)$各分量相互独立的充要条件是其联合分布函数等于边缘分布乘积：

$$F(x_1, ...x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$$
$$= \prod_{i=1}^{n} P(X_i \leq x_i) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

如果密度（质量）函数存在，那么：

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

随机变量函数的独立性

随机变量函数的独立性

$\{X_j, 1 \le j \le n\}$为一系列相互独立的随机变量，$1 \le n_1 \le n_2 \le \cdots \le n_k = n$，那么对于Borel可测函数$f_1, f_2, ...f_k$，那么：

$$\left\{f_1\left(X_1, ...X_{n_1}\right), f_1\left(X_{n_1+1}, ...X_{n_2}\right), \cdots, f_k\left(X_{n_{k-1}+1}, ...X_{n_k}\right)\right\}$$

也为相互独立的随机变量

独立与不相关

独立与不相关

如果概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的随机向量$X = (Y, Z)'$，$Y$和$Z$相互独立且可积，那么：

$$\mathbb{E}(YZ) = \mathbb{E}(Y)\mathbb{E}(Z)$$

因而，如果两个随机变量相互独立，那么其协方差$\mathbb{C}\mathrm{ov}(Y, Z) = \mathbb{E}(YZ) - \mathbb{E}(Y)\mathbb{E}(Z) = 0$。然而反之并不成立。

条件期望

- 令(Y, X)为一个二元随机向量，如何使用随机变量X预测随机变量Y？
- 在统计中，我们把这类问题成为回归（regression）。
- 如果我们观察到了随机变量X的值，那么X的何种函数形式可以更好的预测Y呢？
- 比较常见的做法是最小化均方误差（mean squared error）：

$$\min_{h \in \mathscr{L}^2} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

其中

$$\mathscr{L}^2 = \left\{ h | h : \mathbb{R} \to \mathbb{R}, \mathbb{E}\left[(h(X))^2 \right] < \infty \right\}$$

- 令

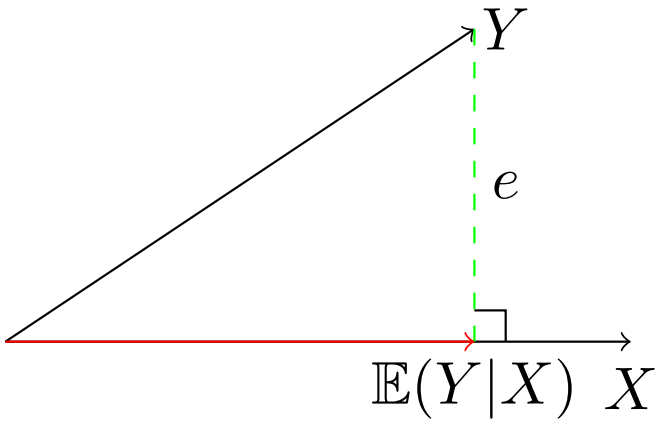
$$h_0(X) = \arg\min_{h \in \mathbb{H}} \left\{ \mathbb{E}\left[(Y - h(X))^2 \right] \right\}$$

条件期望

- 定义误差项$\epsilon = Y - h_0(X)$，对于随机变量X的任意函数$g(X)$，我们有：

$$\mathbb{E}[\epsilon \cdot g(X)] = 0$$

  - 通过反证法证明，如果存在$g(X)$使得$\mathbb{E}[\epsilon \cdot g(X)] \neq 0$，那么我们令$h(X) = h_0(X) + \frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]} g(X)$，根据这一构造，有：$\mathbb{E}\left[(Y - h(X))^2\right] < \mathbb{E}\left[(Y - h_0(X))^2\right]$.
- 我们称$h(X)$为Y在X上的正交投影（orthogonal projection）。



条件期望

- 如果令$g(X)=1$，那么我们有$\mathbb{E}[\epsilon\cdot g(X)]=\mathbb{E}[\epsilon]=\mathbb{E}[Y-h_0(X)]=0$，因而$\mathbb{E}(Y)=\mathbb{E}(h_0(X))$。
- 我们知道，

$$\mathbb{E}(Y)=\arg\min_{c\in\mathbb{R}}\left\{\mathbb{E}(Y-c)^2\right\}$$

- 仿照上式，我们可以定义随机变量Y给定X的条件期望（conditional expectation）：

$$\mathbb{E}(Y|X)=h_0(X)=\arg\min_{h\in\mathscr{L}^2}\left\{\mathbb{E}\left[(Y-h(X))^2\right]\right\}$$

因而随机变量Y给定X的条件期望$\mathbb{E}(Y|X)$是一个关于X的函数。

条件期望与期望

- 期望可以看做是没有任何其他信息时的最优预测，即只用常数对Y进行预测，是条件期望的特例：

$$\mathbb{E}\left(Y|c\right) = c^* = \arg\min_{h\in\mathbb{H}} \left\{\mathbb{E}\left[(Y-c)^2\right]\right\}$$

- 这也就意味着：
  - 期望本身是对一个随机变量的最优预测
  - 具体的一个实现与期望之间的差异为误差项
  - 比如：
    - 如果全国所有人的平均体重为60公斤，那么随机从人群中选取一个人，对其体重的最优预测为60公斤
    - 单独每个人的体重与60公斤之间的差距为误差项

条件期望：离散情形

现在，进一步，如果我们能看到一个变量$D \in \{0,1\}$，那么：

$$\mathbb{E}(Y|D) = \arg\min_{h \in \mathbb{H}} \left\{ \mathbb{E}\left[(Y - h(D))^2 \right] \right\} = \begin{cases} \mathbb{E}(Y|D=1) = h(1) & D=1 \\ \mathbb{E}(Y|D=0) = h(0) & D=0 \end{cases}$$

- 比如，现在我们可以观察到性别（$D=1$代表男性），全国所有男性的平均体重为70公斤，所有女性平均体重为50公斤
- 那么：

$$\begin{cases} \mathbb{E}(Y|D=1) = 70 \\ \mathbb{E}(Y|D=0) = 50 \end{cases}$$

即条件期望为分组平均

条件期望：连续情形

或者，如果我们能看到一个变量X为连续型变量，那么

$$\mathbb{E}\left(Y|X\right) = \arg\min_{h\in\mathbb{H}}\left\{\mathbb{E}\left[\left(Y-h\left(X\right)\right)^2\right]\right\}$$

为一个未知的函数：

- 比如，如果我们现在可以观察到身高（X）
- 可以假想如果有无数个身高一样的人的平均体重，如：

$$\mathbb{E}\left(Y|X=170\right)$$

即为条件期望。

条件期望的性质

条件期望的性质

对于任意的可测函数$g(X)$，条件期望有如下性质：

1. $\mathbb{E}[g(X)|X] = g(X)$;
2. $\mathbb{E}[(Y - \mathbb{E}(Y|X)) \cdot g(X)] = 0$;
3. $\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}(Y)$，$\mathbb{E}[Y - \mathbb{E}(Y|X)] = 0$;
4. $\mathbb{E}[(g(X) \cdot Y)|X] = g(X) \cdot \mathbb{E}(Y|X)$;
5. $\mathbb{E}(aY_1 + bY_2|X) = a\mathbb{E}(Y_1|X) + b\mathbb{E}(Y_2|X)$。

均值独立

- 注意到如果我们没有任何信息，因而只能用常数c去预测Y，那么以上最小化问题：

$$\mathbb{E}\left(Y|c\right)=c^{*}=\arg\min_{h\in\mathscr{L}^{2}}\left\{\mathbb{E}\left[\left(Y-c\right)^{2}\right]\right\}$$

对以上最优化问题求解，即：

$$\frac{\partial\mathbb{E}\left[\left(Y-c\right)^{2}\right]}{\partial c}=\mathbb{E}\left[\frac{\partial\left(Y-c\right)^{2}}{\partial c}\right]=0$$

从而得到：$\mathbb{E}\left(Y|c\right)=\mathbb{E}\left(Y\right)$

- 即如果我们没有任何其他随机变量的信息，只能用常数预测Y，那么我们将得到Y的期望。
- 如果有其他随机变量X，但是$\mathbb{E}\left(Y|X\right)=\mathbb{E}\left(Y\right)$，那么$X$对$Y$的均值没有预测能力，因而我们称$Y$对$X$是**均值独立**（mean independence）的。

均值独立

如果随机变量Y对X是均值独立的，那么：

$$
\begin{aligned}
\mathbb{C}\mathsf{ov}\left(X,Y\right) &= \mathbb{E}\left(XY\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(XY|X\right)\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right) \\
&= \mathbb{E}\left(X\mathbb{E}\left(Y|X\right)\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right) \\
&= \mathbb{E}\left(X\mathbb{E}\left(Y\right)\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right) = 0
\end{aligned}
$$

因而随机变量Y和X必然是不相关的。反之则不成立，不相关并不一定意味着均值独立。

条件方差

- 相应的，我们还可以定义随机变量的条件方差

$$\mathbb{V}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2 |X\right]$$

- 根据条件期望的性质：

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

其中第4个等号由于$\mathbb{E}(Y|X)$也是X的函数

条件方差与方差

根据条件期望的性质，可以证明：

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

即条件方差的期望与条件期望的方差之和为方差。

条件分布

- 如果对于随机向量(X,Y)，我们取$1_A(x)=1\ if\ x\in A\ else\ =0$，这是一个随机变量$X$的函数，因而

$$\mathbb{E}\left(Y\cdot 1_A(X)\right)=\mathbb{E}\left[\mathbb{E}\left(Y|X\right)\cdot 1_A(X)\right]=\mathbb{E}\left[h_0(X)\cdot 1_A(X)\right]$$

- 若X,Y是离散型随机变量，那么我们令$A=\{X=x_i\}$有：

$$\mathbb{E}\left(Y\cdot 1\{X=x_i\}\right)=h_0(x_i)\cdot P(X=x_i)$$

从而：

$$\begin{aligned}\mathbb{E}\left(Y|X=x_i\right)=h_0(x_i)&=\frac{\mathbb{E}\left(Y\cdot 1\{X=x_i\}\right)}{P(X=x_i)}\\&=\frac{\sum_{k=0}^{\infty}\left[y_k\cdot P(Y=y_k,X=x_i)\right]}{P(X=x_i)}\\&=\sum_{k=0}^{\infty}y_k\cdot\frac{P(Y=y_k,X=x_i)}{P(X=x_i)}\end{aligned}$$

条件分布

- 对于连续型随机向量(X,Y)，可以证明：

$$\mathbb{E}\left(Y|X=x\right)=h_0\left(x\right)=\frac{\int_{\mathbb{R}}yf\left(x,y\right)dy}{f_X\left(x\right)}=\int_{\mathbb{R}}y\frac{f\left(x,y\right)}{f_X\left(x\right)}dy$$

  - 由于$\mathbb{E}\left[Y-\mathbb{E}\left(Y|X\right)\right]=0$，从而：

$$\int_{\mathbb{R}}\left[y-h_0\left(x\right)\right]f\left(x,y\right)dy=0$$

  - 固定x，那么以上条件意味着：

$$\int_{\mathbb{R}}yf\left(x,y\right)dy=h_0\left(x\right)\int_{\mathbb{R}}f\left(x,y\right)dy=h_0\left(x\right)f_X\left(x\right)$$

条件分布

条件密度函数

对于离散型随机变量，定义

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

对于连续型随机变量，定义

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)\,dy}$$

我们把$f_{Y|X}(y|x)$定义为条件密度函数（conditional denstiy function）。

条件密度函数

- 条件期望可以写为条件密度函数的积分

$$\mathbb{E}\left(Y|X\right)=\int_{\mathbb{R}} y\cdot f_{Y|X}\left(y|x\right)dy$$

- 注意：对于离散型随机变量：

$$\sum_{y} f_{Y|X}\left(y|x\right)=\sum_{y}\frac{P\left(Y=y,X=x\right)}{\sum_{y}P\left(Y=y,X=x\right)}=1$$

而对于连续型随机变量：

$$\int_{\mathbb{R}} f_{Y|X}\left(y|x\right)dy=\int_{\mathbb{R}}\frac{f\left(x,y\right)}{\int_{\mathbb{R}}f\left(x,y\right)dy}dy=1$$

因而条件密度函数也是密度函数。

独立与均值独立

- 如果随机变量X和Y是独立的，那么：

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

即两个随机变量独立的充要条件是$f_{Y|X} = f_Y$。

- 在独立的条件下：

$$\mathbb{E}(Y|X) = \int_{\mathbb{R}} y \cdot f_{Y|X}(y|x)\, dy = \int_{\mathbb{R}} y \cdot f_Y(y)\, dy = \mathbb{E}(Y)$$

因而如果随机变量X和Y是独立的，那么其一定是均值独立的，反之则不成立。

- 独立、均值独立、不相关三者的强弱关系如下：

$$X \perp\!\!\!\perp Y \underset{\not\Leftarrow}{\Rightarrow} \mathbb{E}(Y|X) = \mathbb{E}(Y) \underset{\not\Leftarrow}{\Rightarrow} X \perp Y$$

其中$X \perp\!\!\!\perp Y$代表X与Y独立，$X \perp Y$代表X与Y不相关。

贝叶斯公式

- 使用条件密度函数的定义，我们还可以得到随机变量的贝叶斯公式。
- 由于：$f(x,y) = f_X(x) \cdot f_{Y|X}(y|x) = f_Y(y) \cdot f_{X|Y}(x|y)$从而条件密度：

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{f(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{\int_{\mathbb{R}} f(x,y)\,dy} \\
&= \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{\int_{\mathbb{R}} f_{X|Y}(x|y) \cdot f_Y(y)\,dy}
\end{aligned}
$$

以上方程即随机变量的贝叶斯公式，在贝叶斯统计中有大量的应用。

迭代期望公式

- 条件期望可以很方便的扩充到多个X的情形，比如$\mathbb{E}(Y|X_1,X_2)$可以定义为：

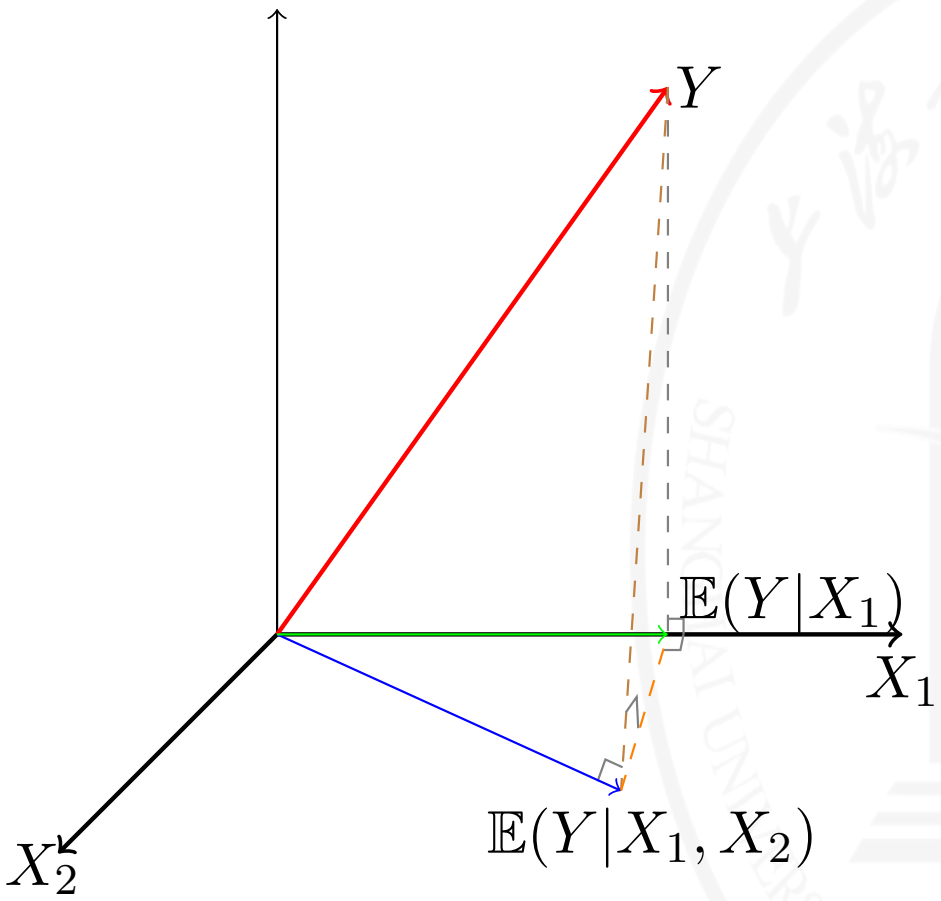
$$\mathbb{E}(Y|X_1,X_2)=h_0(X_1,X_2)=\arg\min_{h\in\mathscr{L}^2}\left\{\mathbb{E}\left[(Y-h(X_1,X_2))^2\right]\right\}$$

- 条件期望有如下性质（**迭代期望公式**，**law of iterated expectation**）：

$$\mathbb{E}\left[\mathbb{E}(Y|X_1,X_2)|X_1\right]=\mathbb{E}(Y|X_1)$$

即如果我们对随机变量Y，先在大的空间上投影，再在这个大的空间上的一个小的子空间上进行投影，与直接在这个小的空间上进行投影是相等的。

迭代期望公式



条件期望的一般定义

因而我们通常把条件期望的概念推广到$\sigma-$代数上。对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，我们可以对$\mathscr{F}$的一个子$\sigma-$代数$\mathscr{G} \subset \mathscr{F}$定义条件期望如下：

条件期望

对于概率空间$(\Omega, \mathscr{F}, \mathscr{P})$，$\mathcal{I} \subset \mathscr{F}$为一个$\sigma-$代数，对于随机变量$Y$满足：$\mathbb{E}(|Y|) < \infty$，如果对于任意的$A \in \mathcal{I}$，随机变量$H$满足：

$$\mathbb{E}(Y \cdot 1_A) = \mathbb{E}(H \cdot 1_A)$$

那么我们称H为给定$\mathcal{I}$随机变量Y的条件期望，记为$\mathbb{E}(Y|\mathcal{I})$。令$B \in \mathscr{F}$，定义$\mathscr{P}(B|\mathcal{I}) = \mathbb{E}(1_B|\mathcal{I})$为条件概率。

迭代期望公式

- 注意以上定义的$\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma\langle X\rangle)$。
- 特别的，令$\mathcal{I} = \{\emptyset, \Omega\}$，$\mathbb{E}(Y|\{\emptyset, \Omega\}) = \mathbb{E}(Y)$，即信息量最小的条件期望即为期望本身。
- 而以上的迭代期望公式也可以相应推广，即如果$\mathcal{I}_1 \subset \mathcal{I}_2 \subset \mathscr{F}$，那么：

$$\mathbb{E}(Y|\mathcal{I}_1) = \mathbb{E}\{[\mathbb{E}(Y|\mathcal{I}_2)]|\mathcal{I}_1\}$$

即先在大的信息集上做投影，再将其投影到小的信息集上，等价于直接投影在小的信息集上。

收敛的概念

收敛的定义

若$\{a_n, n = 1, 2, ...\}$为实数序列，如果对于任意的$\epsilon > 0$，存在$n_0 = n_0(\epsilon)$使得：

$$|a_n - a| < \epsilon, \forall n > n_0$$

那么我们称数列$\{a_n\}$的极限为a，或者$\{a_n\}$收敛到（converges to）a，记为

$$\lim_{n\to\infty} a_n = a$$

或者

$$a_n \to a \ as \ n \to \infty$$

收敛的概念

有界的定义

若$\{a_n, n = 1, 2, ...\}$为实数序列，如果存在常数$b < \infty$，使得

$$|a_n| < b$$

那么我们称数列$\{a_n\}$为有界的（bounded），否则称之为无界的（unbounded）。

小o符号

小o符号的定义

对于两个序列$\{a_n\},\{b_n\}$，如果随着$n \to \infty$，有：

$$\frac{a_n}{b_n} \to 0$$

那么我们记为$a_n = o(b_n)$。特别的，如果令$b_n = 1$，那么$a_n = o(1)$等价为$a_n \to 0$。

小o符号

小o实例

如果$a_n = \frac{1}{n^2}, b_n = \frac{1}{n}$，那么：

$$\frac{a_n}{b_n} = \frac{\frac{1}{n^2}}{\frac{1}{n}} = \frac{1}{n} \to 0$$

因而$\frac{1}{n^2} = o\left(\frac{1}{n}\right)$，即$\frac{1}{n^2}$以更快的速度收敛到0。如果两个序列$a_n \to 0, b_n \to 0$，且$a_n = o\left(b_n\right)$，那么我们称$a_n$为比$b_n$高阶的无穷小量。

小o符号的应用：泰勒展开

当$x \to a$时，$(x-a)=o(1)$，同时我们有$(x-a)^{k+1}=o\left((x-a)^k\right)$，即当$x \to a$时，$(x-a)$的高阶幂是低阶幂的无穷小量。对于一个单变量实值函数$f(x)$且k阶可微，那么有：

$$f(x)=f(a)+f'(a)(x-a)+\frac{f''(a)}{2!}(x-a)^2+\cdots$$
$$+\frac{f^{(k)}(a)}{k!}(x-a)^k+o\left(|x-a|^k\right)$$

因而对于一个难以计算的函数f，我们经常使用其前k阶泰勒多项式对其进行逼近。

泰勒展开

泰勒展开

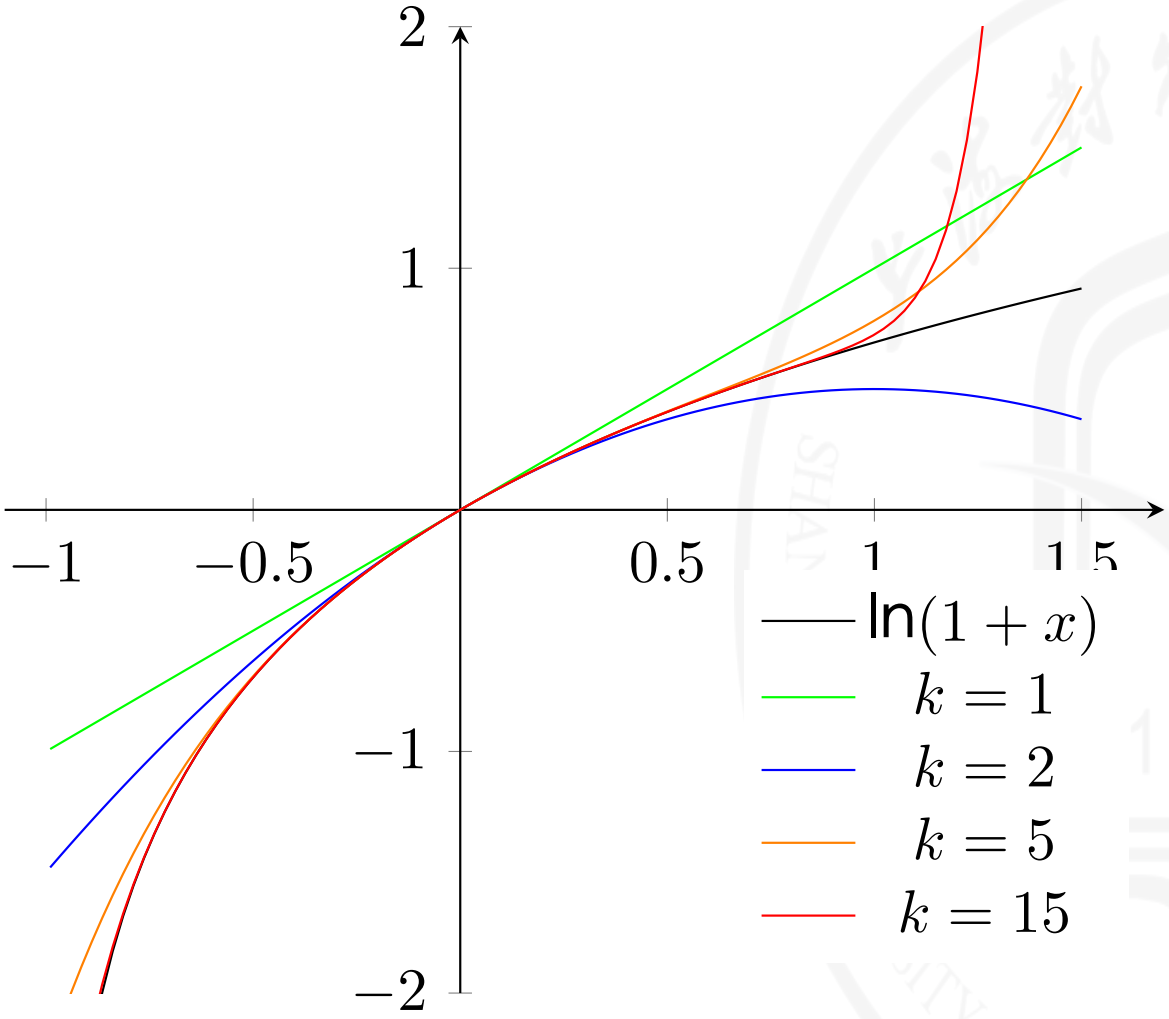
函数$f(x)=\ln(1+x)$在$x=0$处的泰勒展开为:

$$f(x)=\sum_{n=1}^{\infty}(-1)^{n+1}\frac{x^n}{n}=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots$$

因而当x充分靠近0时，我们可以使用前k阶泰勒展开对其进行逼近。特别的，如果令

$$k=1,\ \ \ln(1+x)=x+o(x)\approx x$$

泰勒展开



多元函数的泰勒展开

更一般的，如果$f:\mathbb{R}^n\to\mathbb{R}$为多元实值函数，那么其泰勒级数为：

$$f(x)=f(a)+\frac{\partial f}{\partial x'}(a)(x-a)+\frac{1}{2!}(x-a)'\frac{\partial^2 f}{\partial x\partial x'}(a)(x-a)+o\left(\|x-a\|^2\right)$$

其中x和a为$n\times 1$向量。

多元函数的泰勒展开

多元函数泰勒展开示例

令$f(x)=e^{x_1}\ln(1+x_2)$，其中$x=(x_1,x_2)'$。那么：

$$\frac{\partial f}{\partial x}=\begin{bmatrix} e^{x_1}\ln(1+x_2) \\ \frac{e^{x_1}}{1+x_2} \end{bmatrix},\frac{\partial^2 f}{\partial x\partial x'}=\begin{bmatrix} e^{x_1}\ln(1+x_2) & \frac{e^{x_1}}{1+x_2} \\ \frac{e^{x_1}}{1+x_2} & -\frac{e^{x_1}}{(1+x_2)^2} \end{bmatrix}$$

因而其在$a=(0,0)'$处的二阶泰勒展开：

$$\begin{aligned} p_2(x) &= [0,1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}+\frac{1}{2}[x_1,x_2]\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= x_2+\frac{1}{2}\left(2x_1x_2-x_2^2\right) \end{aligned}$$

小o符号

小o的性质

1. 若$a_n = o(b_n), b_n = o(c_n)$，那么$a_n = o(c_n)$
2. 对于任意的常数$c \neq 0$，及$a_n = o(b_n)$，有$ca_n = o(b_n)$
3. 对于任意的数列$c_n \neq 0$，及$a_n = o(b_n)$，有$c_n a_n = o(c_n b_n)$
4. 如果$d_n = o(b_n), e_n = o(c_n)$，那么$d_n e_n = o(b_n c_n)$
5. 如果$a_n, b_n > 0, c_n, d_n > 0$，$a_n = o(b_n), c_n = o(d_n)$，那么$a_n + c_n = o(b_n + d_n)$。

大O符号

大O符号的定义

对于两个序列$\{a_n\},\{b_n\}$，如果随着$n\to\infty$，$\left|\frac{a_n}{b_n}\right|$是有界的，即存在一个M使得：

$$\left|\frac{a_n}{b_n}\right| < M$$

那么我们记为$a_n = O(b_n)$。特别的，如果令$b_n = 1$，那么$a_n = O(1)$等价为a_n是有界的。

同阶的定义

对于两个序列$\{a_n\},\{b_n\}$，如果$a_n = O(b_n)$，且同时$b_n = O(a_n)$，那么我们称两个序列是同阶的，简记为$a_n \asymp b_n$。

大O符号

大O符号示例

对于序列$a_n = \frac{1}{n} + \frac{b}{n\sqrt{n}} + \frac{c}{n^2} + \frac{d}{n^2\sqrt{n}}$，同时定义$R_n = \frac{b}{n\sqrt{n}} + \frac{c}{n^2} + \frac{d}{n^2\sqrt{n}}$那么：

1. $a_n \sim \frac{1}{n}$
2. 若$b = 0$，$R_n = O\left(\frac{1}{n^2}\right)$
3. 若$b = 0$，$R_n \asymp \frac{1}{n^2}$
4. 若$b \neq 0$，$R_n \sim \frac{b}{n\sqrt{n}}$
5. 若$b = c = 0$，$R_n = o\left(\frac{1}{n^2}\right)$

大O符号

大O符号的性质

1. 若$a_n = O(b_n), b_n = O(c_n)$，那么$a_n = O(c_n)$
2. 对于任意的常数$c \neq 0$，及$a_n = O(b_n)$，有$ca_n = O(b_n)$
3. 对于任意的数列$c_n \neq 0$，及$a_n = O(b_n)$，有$c_n a_n = O(c_n b_n)$
4. 如果$d_n = O(b_n), e_n = O(c_n)$，那么$d_n e_n = O(b_n c_n)$
5. 如果$a_n = o(b_n)$，$c_n = O(b_n)$，那么$a_n c_n = o(b_n)$
6. 如果$a_n = o(b_n)$，$c_n = O(b_n)$，那么$a_n + c_n = O(b_n)$

依概率收敛

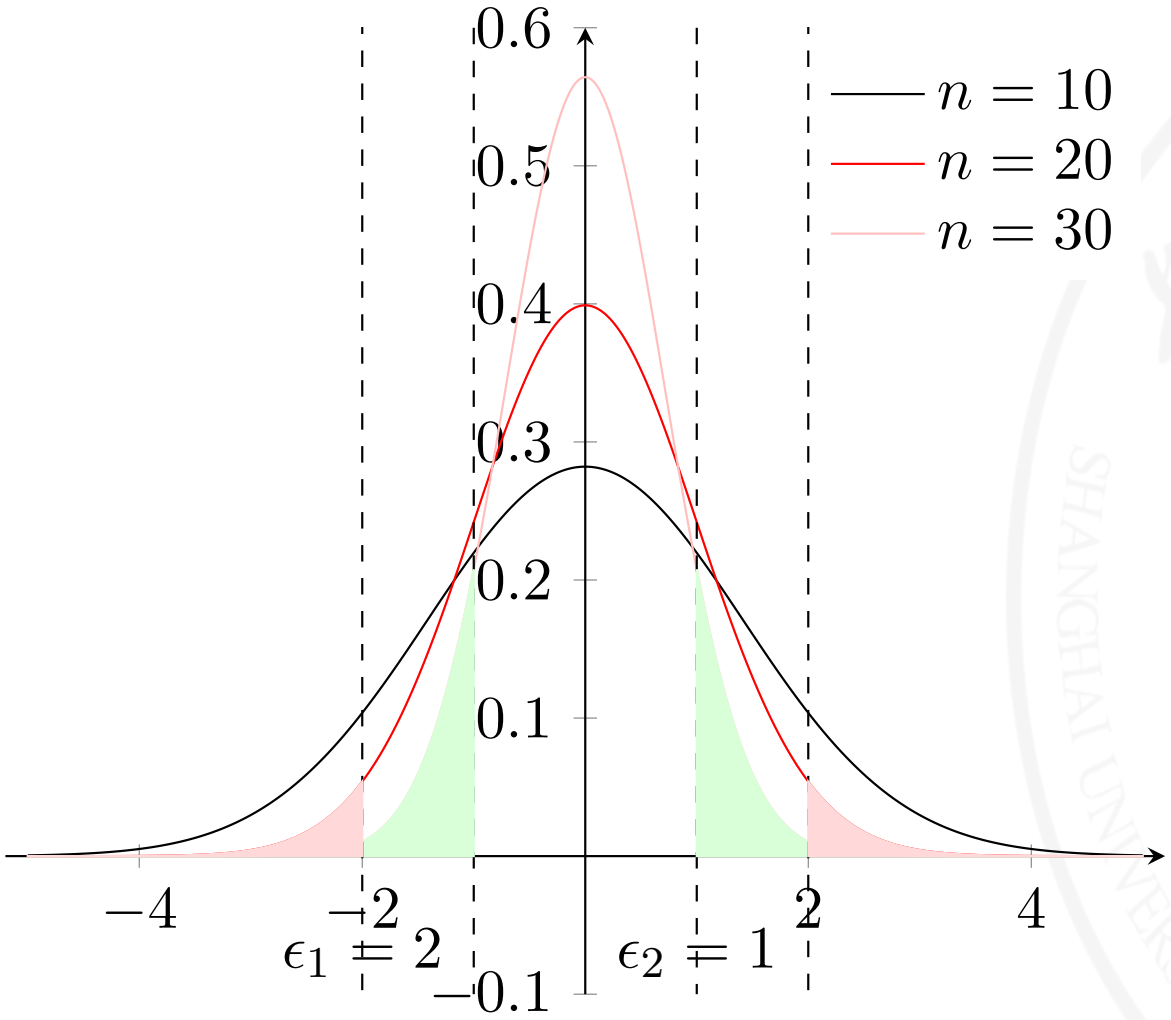
依概率收敛

如果对于任意的$\epsilon > 0$，概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一系列随机变量序列$\{X_n\}$ 满足：

$$\mathscr{P}\left(|X_n - X| > \epsilon\right) \to 0$$

那么我们称X_n依概率收敛于X，记为$X_n \xrightarrow{p} X$，
或$\text{plim} X_n = X$。

依概率收敛



o_p符号

o_p符号

$\{X_n\}$与$\{Y_n\}$为定义在概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的两个随机变量序列，如果

$$\frac{X_n}{Y_n} \xrightarrow{p} 0$$

那么我们记为$X_n = o_p(Y_n)$。特别的，当$Y_n = 1$时，即$X_n = o_p(1)$，等价于$X_n \xrightarrow{p} 0$。

依概率收敛

简单的大数定律

现在假设有一组独立同分布的样本$x_i, i = 1, 2, ..., N$，其方差有界：$\mathbb{V}(x_i) < M$，其样本均值为$\bar{x}_N$。根据切比雪夫不等式，随着样本量$N \to \infty$，有：

$$
\begin{aligned}
P\left(\left|\bar{x}_N - \mathbb{E}(x_i)\right| > \epsilon\right) &\leq \frac{\mathbb{E}\left[\left(\bar{x}_N - \mathbb{E}(x_i)\right)^2\right]}{\epsilon^2} \\
&= \frac{\mathbb{V}(\bar{x}_N)}{\epsilon^2} = \frac{\frac{\mathbb{V}(x_i)}{N}}{\epsilon^2} \\
&< \frac{M}{N\epsilon^2} \to 0
\end{aligned}
$$

从而$\bar{x}_N \xrightarrow{p} \mathbb{E}(x_i)$，或者$\bar{x}_N - \mathbb{E}(x_i) = o_p(1)$，或者$\bar{x}_N = \mathbb{E}(x_i) + o_p(1)$。

O_p符号

O_p符号

$\{X_n\}$与$\{Y_n\}$为定义在概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的两个随机变量序列，如果对于任意的$\epsilon > 0$，存在一个$C_\epsilon$使得：

$$\sup_n \mathscr{P}\left(|X_n| \geq C_\epsilon |Y_n|\right) < \epsilon$$

那么我们记$X_n = O_p(Y_n)$。特别的，当$Y_n = 1$时，我们称X_n依概率有界（bounded in probability）。

O_p符号

O_p符号是对O符号在概率意义上的扩展。注意如果$\mathbb{V}(X_n) < M$，即随机变量序列$\{X_n\}$的方差有界，那么对于任意的$\epsilon > 0$，取$C_\epsilon = \sqrt{M/\epsilon + 1}$，那么：

$$\mathscr{P}(|X_n| \geq C_\epsilon) \leq \frac{\mathbb{E}\left(X_n^2\right)}{C_\epsilon^2} = \frac{\mathbb{E}\left(X_n^2\right)}{M/\epsilon + 1} < \epsilon$$

因而$X_n = O_p(1)$。因而如果X_n的方差有界，那么必然是$O_p(1)$。

依概率收敛

样本均值的阶数

现在假设有一组独立同分布的样本$x_i, i = 1, 2, ..., N$，其方差有界：$\mathbb{V}(x_i) < M$，其样本均值为$\bar{x}_N$。可以计算：

$$\mathbb{V}(\bar{x}_N) = \frac{\mathbb{V}(x_i)}{N}$$

因而$\bar{x}_N$的方差是有界的，$\mathbb{V}(\bar{x}_N) = O_p(1)$。此外：

$$\mathbb{V}\left(\sqrt{N}\bar{x}_N\right) = \mathbb{V}(x_i)$$

因而$\sqrt{N}\bar{x}_N = O_p(1)$。但是，如果我们考虑样本和：$S_N = \sum_{i=1}^{N} x_i$由于$\mathbb{V}(S_N) = N\mathbb{V}(x_i)$，该方差是无界的，因而不是$O_p(1)$。

O_p和o_p符号的性质

O_p和o_p的性质

如果$X_n = o_p(1), Y_n = o_p(1), Z_n = O_p(1), W_n = O_p(1)$，那么：

1. $X_n + Y_n = o_p(1)$
2. $X_n + Z_n = O_p(1)$
3. $Z_n + W_n = O_p(1)$
4. $X_n Y_n = o_p(1)$
5. $X_n Z_n = o_p(1)$
6. $Z_n W_n = O_p(1)$

均方收敛

均方收敛的定义

如果概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一系列随机变量序列$\{X_n\}$随着$n \to \infty$满足：

$$\mathbb{E}\left(|X_n - X|^2\right) \to 0$$

那么我们称X_n均方收敛于X，记为$X_n \xrightarrow{L^2} X$。

均方收敛

均方收敛与依概率收敛

如果随机变量序列$X_n \xrightarrow{L^2} X$，那么$X_n \xrightarrow{p} X$。

Proof.

据切比雪夫不等式，对于任意的$\epsilon > 0$，有：

$$\mathscr{P}\left(|X_n - X| > \epsilon\right) \leq \frac{\mathbb{E}\left(|X_n - X|^2\right)}{\epsilon^2} \to 0$$

□

依分布收敛

依分布收敛的定义

令F_n,F为分布函数，如果对于每一个$F(x)$连续的点x，有：

$$\lim_{n\to\infty} F_n(x) = F(x)$$

那么我们称$F_n(x)$弱收敛于$F(x)$，记为$F_n \xrightarrow{w} F$。

如果一系列随机变量$\{X_n\}$的分布函数$F_{X_n}(x) \xrightarrow{w} F_X$，我们称$X_n$依分布收敛于$X$，记为$X_n \xrightarrow{D} X$或者：$X_n \overset{a}{\sim} F$，其中$a$代表渐进的（asymptotically），即$X_n$渐进服从分布函数为$F$的分布。

依分布收敛

依分布收敛与O_p的关系

1. 如果$X_n \neq O_p(1)$，那么分布函数极限$\lim_{n\to\infty} F_n(x) = F(x)$不是一个分布函数。
2. 如果X_n依分布收敛，那么$X_n = O_p(1)$。

因而当我们讨论依分布收敛时，一定要保证我们讨论的$X_n = O_p(1)$。

大数定律

大数定律（Law of Large Numbers, LLN）讨论样本均值的极限，即在何种条件下，以下结论：

$$\frac{S_N - \mathbb{E}\left(S_N\right)}{N} = \bar{x} - \mathbb{E}\left(\bar{x}\right) \xrightarrow{p} 0$$

成立，其中：

$$S_N = \sum_{i=1}^{N} x_i$$

大数定律

实际上，$S_N - \mathbb{E}(S_N) = \sum_{i=1}^{N}[x_i - \mathbb{E}(x_i)]$，因而：

$$
\begin{aligned}
\mathbb{E}[S_N - \mathbb{E}(S_N)]^2 &= \mathbb{E}\left(\sum_{i=1}^{N}[x_i - \mathbb{E}(x_i)]\right)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{N}[x_i - \mathbb{E}(x_i)]^2\right) \\
&+ \mathbb{E}\left(2\sum_{1\le j<i\le N}^{N}[x_i - \mathbb{E}(x_i)][x_j - \mathbb{E}(x_j)]\right) \\
&= \sum_{i=1}^{N}\mathbb{V}(x_i) + 2\sum_{1\le j<i\le N}^{N}\mathbb{C}\mathrm{ov}(x_i, x_j)
\end{aligned}
$$

大数定律

$$\mathbb{E}\left[S_N - \mathbb{E}\left(S_N\right)\right]^2 = \sum_{i=1}^{N} \mathbb{V}\left(x_i\right) + 2 \sum_{1 \le j < i \le N}^{N} \mathbb{C}\mathsf{ov}\left(x_i, x_j\right)$$

如果我们假设$\mathbb{V}\left(x_i\right) < M$，那么：

1. $\sum_{i=1}^{N} \mathbb{V}\left(x_i\right) < NM = O\left(N\right)$
2. 而根据Cauthy-Schwartz不等式，$\mathbb{C}\mathsf{ov}\left(x_i, x_j\right) \le \sqrt{\mathbb{V}\left(x_i\right)\mathbb{V}\left(x_j\right)} \le M$，而上式中有$\frac{N(N-1)}{2}$个协方差，所以

$$\sum_{1 \le j < i \le N}^{n} \mathbb{C}\mathsf{ov}\left(x_i, x_j\right) = O\left(N^2\right)$$

大数定律

$$\mathbb{E}\left[S_N - \mathbb{E}\left(S_N\right)\right]^2 = O\left(N\right) + O\left(N^2\right)$$

从而：

$$\mathbb{E}\left[\frac{S_N - \mathbb{E}\left(S_N\right)}{N}\right]^2 = \frac{1}{N^2}O\left(N\right) + \frac{1}{N^2}O\left(N^2\right) = o\left(1\right) + O\left(1\right)$$

根据均方收敛定义，如果希望$\frac{S_N}{N} - \frac{\mathbb{E}(S_N)}{N} \xrightarrow{L^2} 0$，必须

$$\mathbb{E}\left[\frac{S_N - \mathbb{E}\left(S_N\right)}{N}\right]^2 = o\left(1\right)$$

大数定律

$$\mathbb{E}\left[\frac{S_N-\mathbb{E}\left(S_N\right)}{N}\right]^2=o\left(1\right)+O\left(1\right)$$

其中的$O\left(1\right)$来源于$\frac{N(N-1)}{2}$个协方差，如果$\mathbb{C}\mathsf{ov}\left(x_i,x_j\right)=0$，那么自然有：

$$\mathbb{E}\left[\frac{S_N-\mathbb{E}\left(S_N\right)}{N}\right]^2=o\left(1\right)$$

从而

$$\frac{S_N}{N}-\frac{\mathbb{E}\left(S_N\right)}{N}\xrightarrow{L^2}0\Rightarrow\frac{S_N}{N}-\frac{\mathbb{E}\left(S_N\right)}{N}\xrightarrow{p}0$$

大数定律

大数定律1

如果概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的一个随机变量序列$\{x_i\}$两两不相关，且存在一个M使得对于所有的$i = 1, 2, \ldots$，都有$\mathbb{V}(x_i) < M$，那么：

$$\bar{x} - \mathbb{E}(\bar{x}) = \frac{S_N - \mathbb{E}(S_N)}{N} \xrightarrow{L^2} 0$$

从而

$$\bar{x} - \mathbb{E}(\bar{x}) = \frac{S_N - \mathbb{E}(S_N)}{N} \xrightarrow{p} 0$$

如果额外假设$\{x_i\}$是同分布的，那么$\mathbb{E}(\bar{x}) = \mathbb{E}(x_i) = \mu$，从而：

$$\bar{x} \xrightarrow{p} \mu$$

大数定律

以下的定理放松了二阶矩有限的假定以及独立的假定，保留了独立同分布的假定：

大数定律2

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的两两独立且同分布的随机变量序列，若

$$\mathbb{E}\left|x_i\right| < \infty$$

那么

$$S_N/N = \bar{x} \xrightarrow{p} \mu$$

其中$\mu = \mathbb{E}\left(x_i\right)$。

大数定律

以下的定理则同时放宽了同分布的假定以及二阶矩的假定。

大数定律3

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上的相互独立的随机变量序列，如果存在一个常数$p \in [1, 2]$，随着$N \to \infty$，使得：

$$\frac{1}{N^p} \sum_{i=1}^{N} \mathbb{E} \left| x_i \right|^p \to 0$$

那么$S_N / N = \bar{x} \xrightarrow{p} \mathbb{E}(\bar{x})$

大数定律

大数定律的应用

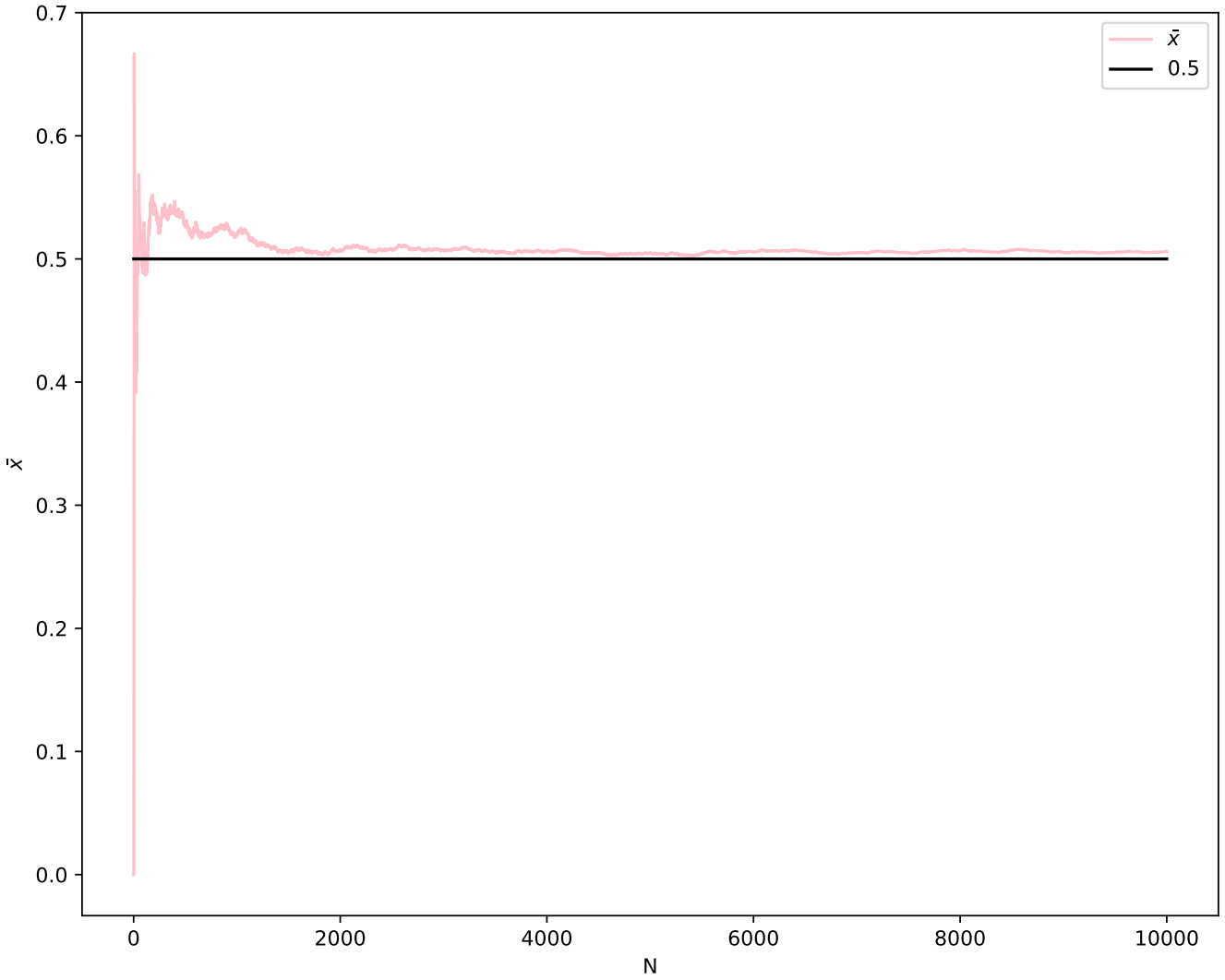
如果令$\{x_i\}$为一系列i.i.d的随机变量，且$x_i \sim Ber(p)$，那么$\mathbb{E}(x_i) = p, \mathbb{V}(x_i) = p(1-p) < \infty$，定义：

$$\hat{p} = \frac{\sum_{i=1}^{N} x_i}{N}$$

即成功的比例，那么根据上例，可以得到

$$\hat{p} \xrightarrow{p} p$$

大数定律



中心极限定理

中心极限定理（Central Limit Theorem, CLT）讨论样本均值的极限分布，即在大样本条件下，样本均值的分布情况，通常中心极限定理可以得到如下结论：

$$\sqrt{N}\left(\bar{x}-\mu\right)=\frac{S_N-N\mu}{\sqrt{N}}\overset{a}{\sim}N\left(0,\mathbb{V}\left(x_i\right)\right)$$

其中：

$$S_N=\sum_{i=1}^{N}x_i$$

即样本均值的极限分布为正态分布。

中心极限定理

前面我们提到，如果希望讨论$\sqrt{N}(\bar{x}-\mu)$的极限分布，需要保证$\sqrt{N}(\bar{x}-\mu)=O_p(1)$，实际上：如果假设$\{x_i\}$之间两两不相关，那么：

$$\begin{aligned}\mathbb{V}(S_N) &= \mathbb{E}[S_N-\mathbb{E}(S_N)]^2 \\ &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mathbb{E}(X_i))\right]^2 \\ &= \sum_{i=1}^{n}\mathbb{V}(X_i) \\ &= O(N)\end{aligned}$$

从而$\mathbb{V}\left(\sqrt{N}\frac{S_N}{N}\right)=O(1)$，或者其方差有界，因而$\sqrt{N}\frac{S_N}{N}=\sqrt{N}\bar{x}=O_p(1)$。

中心极限定理

中心极限定理（标量）

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上$i.i.d$的随机变量序列，
且$\mathbb{E}(x_i) = \mu, \mathbb{V}(x_i) = \sigma^2$，那么：

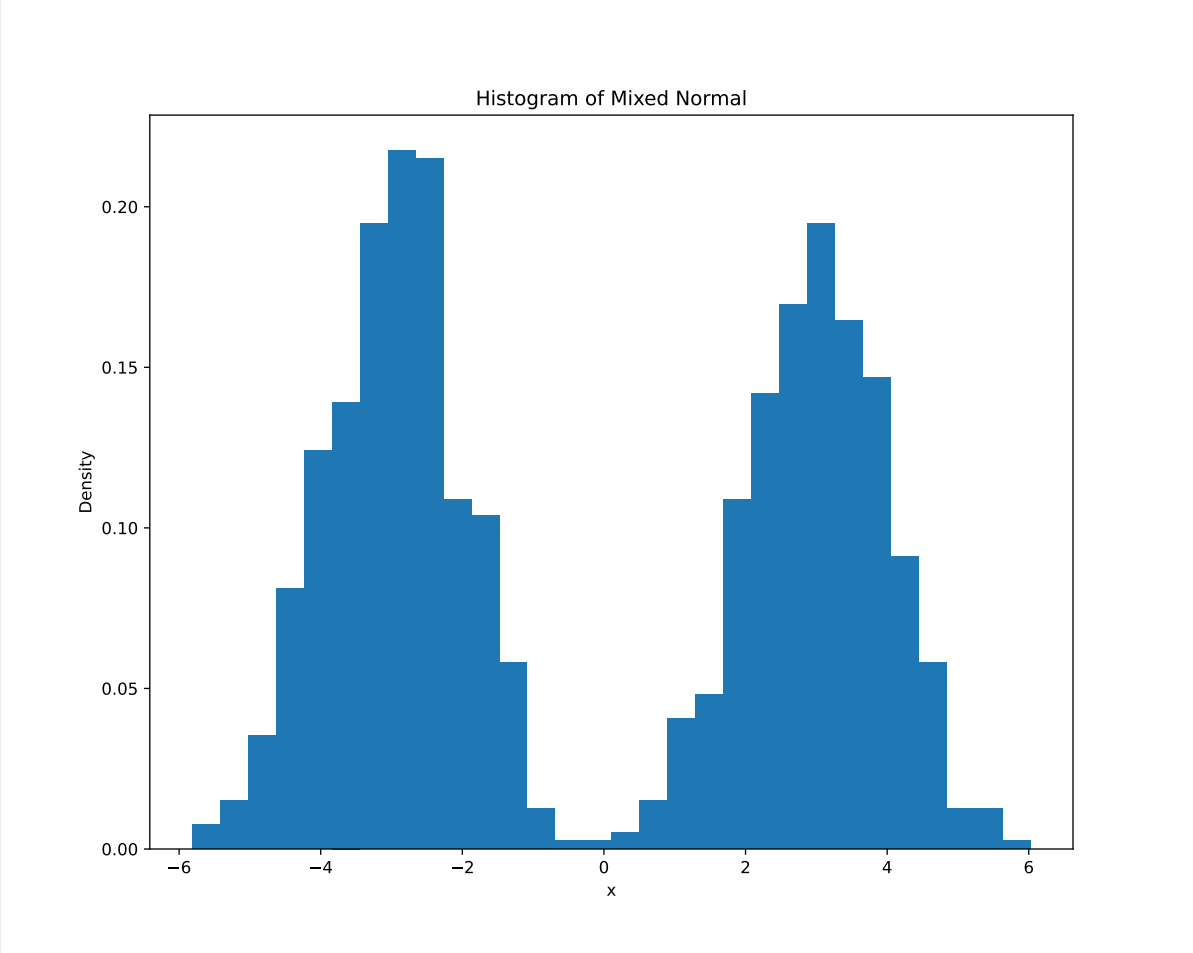
$$\sqrt{N}(\bar{x}_N - \mu) \overset{a}{\sim} N(0, \sigma^2)$$

或：

$$\sqrt{N}\left(\frac{\bar{x}_N - \mu}{\sigma}\right) \overset{a}{\sim} N(0, 1)$$

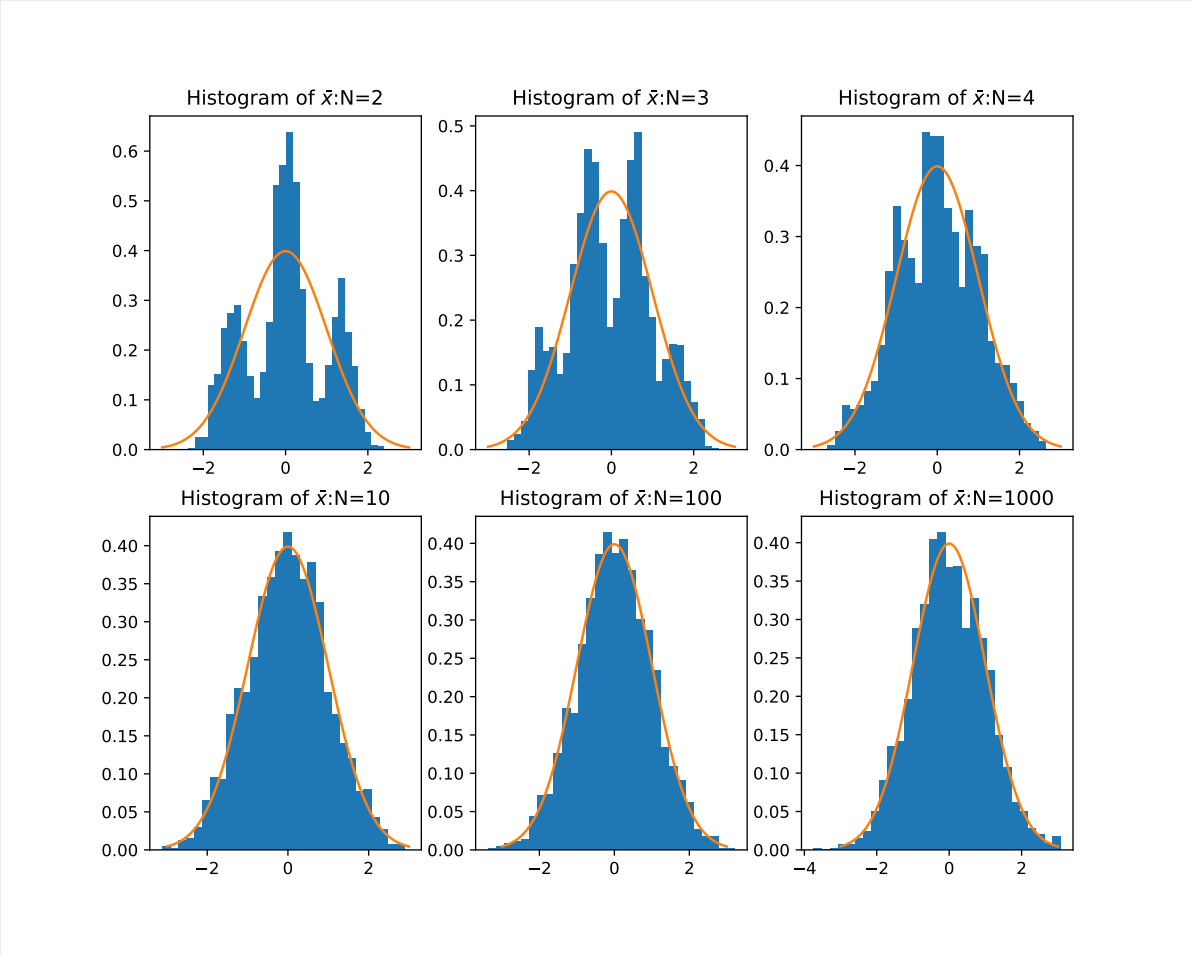
中心极限定理

x_i的分布



中心极限定理

$\sqrt{N}\left(\bar{x}-\mu\right)/\sigma$的分布



中心极限定理

中心极限定理示例

如果$\{x_i\}$为$i.i.d$的随机变量，且$x_i \sim Ber(p)$，令$\hat{p}_N$如前定义，那么：

$$\sqrt{N}\left(\hat{p}_N - p\right) \overset{a}{\sim} N\left(0, p\left(1-p\right)\right)$$

如果$\{x_i\}$为$i.i.d$的随机变量，且$x_i \sim N(0,1)$，那么可知$\mathbb{E}\left(x_i^2\right) = 1, \mathbb{E}\left(x_i^4\right) = 3$，因而：

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - 1\right) \overset{a}{\sim} N(0,2)$$

中心极限定理

中心极限定理（向量）

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上$i.i.d$的随机向量序列，
且$\mathbb{E}(x_i) = \mu, \mathbb{V}(x_i) = \Sigma$，那么：

$$\sqrt{N}(\bar{x}_N - \mu) \overset{a}{\sim} N(0, \Sigma)$$

或：

$$\sqrt{N}\Sigma^{-\frac{1}{2}}(\bar{x}_N - \mu) \overset{a}{\sim} N(0, I)$$

随机变量连续函数的收敛

连续函数的收敛

令$\{X_i\}$为k维随机向量，$g(x):\mathbb{R}^k \to \mathbb{R}^l$为连续函数，那么：

1. $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$
2. $X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$
3. $X_n \xrightarrow{D} X \Rightarrow g(X_n) \xrightarrow{D} g(X)$

随机变量连续函数的收敛

样本相关系数的极限

对于二维随机向量(x_i, y_i)，令(x_i, y_i)为i.i.d的样本，那么在可积性条件下，

$$\begin{cases} \frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{p} \mathbb{E}\left(x_i\right) & \frac{1}{N}\sum_{i=1}^{N} y_i \xrightarrow{p} \mathbb{E}\left(y_i\right) \\ \frac{1}{N}\sum_{i=1}^{N} x_i y_i \xrightarrow{p} \mathbb{E}\left(x_i y_i\right) & \\ \frac{1}{N}\sum_{i=1}^{N} x_i^2 \xrightarrow{p} \mathbb{E}\left(x_i^2\right) & \frac{1}{N}\sum_{i=1}^{N} y_i^2 \xrightarrow{p} \mathbb{E}\left(y_i^2\right) \end{cases}$$

从而

$$\frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \frac{1}{N}\sum_{i=1}^{N} y_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2}} \xrightarrow{p} \mathbb{C}\mathrm{orr}\left(x_i, y_i\right)$$

Slutsky定理

Slutsky定理

如果随机变量$X_n \xrightarrow{D} X$，$R_n = o_p(1)$，那么

$$X_n + R_n \xrightarrow{D} X$$

同时如果$Y_n \xrightarrow{p} a \neq 0$，那么

$$\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{a}$$

如果$Y_n \xrightarrow{p} a$，那么

$$X_n Y_n \xrightarrow{D} aX$$

Slutsky定理

t统计量的大样本分布

之前曾讨论过，如果$x_i \sim N\left(\mu, \sigma^2\right)$ $i.i.d$，那么

$$\frac{\sqrt{N}\left(\bar{x}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}{N-1}}} \sim t\left(N-1\right)$$

现在我们不假设x_i服从正态分布，而是假设其独立同分布且具有有限的二阶矩，那么我们有$\bar{x} \xrightarrow{p} \mathbb{E}\left(x_i\right), \frac{1}{N}\sum_{i=1}^{n} x_i^2 \xrightarrow{p} \mathbb{E}\left(x_i^2\right)$

Slutsky定理

t统计量的大样本分布

因而：

$$\begin{aligned}\frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}{N-1} &= \frac{\sum_{i=1}^{N}x_i^2}{N-1}-\frac{N}{N-1}\bar{x}^2 \\ &= \frac{\sum_{i=1}^{N}x_i^2-N\bar{x}^2}{N-1} \\ &\xrightarrow{p} \mathbb{E}\left(x_i^2\right)-\left[\mathbb{E}\left(x_i\right)\right]^2=\mathbb{V}\left(x_i\right)\end{aligned}$$

进而：

$$\frac{\sqrt{N}\left(\bar{x}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}{N-1}}} \xrightarrow{p} \frac{\sqrt{N}\left(\bar{x}-\mu\right)}{\sqrt{\mathbb{V}\left(x_i\right)}}=\sqrt{N}\left(\frac{\bar{x}-\mu}{\sqrt{\mathbb{V}\left(x_i\right)}}\right)\overset{a}{\sim}N\left(0,1\right)$$

delta方法

delta方法示例

令$\{x_i\}$为概率空间$(\Omega, \mathscr{F}, \mathscr{P})$上$i.i.d$的随机变量序列，
且$\mathbb{E}(x_i) = \mu, \mathbb{V}(x_i) = \sigma^2$，那么根据中心极限定理：

$$\sqrt{N}(\bar{x} - \mu) \overset{a}{\sim} N(0, \sigma^2)$$

delta方法

delta方法示例

如果我们关心$Y = \exp(\bar{x})$的分布，那么可以对其进行泰勒展开：

$$
\begin{aligned}
\sqrt{N}Y - \sqrt{N}\exp(\mu) =& \sqrt{N}\exp(\bar{x}) - \sqrt{N}\exp(\mu) \\
=& \sqrt{N}\exp(\mu)(\bar{x}-\mu) + \frac{1}{2}\sqrt{N}\exp(\mu)(\bar{x}-\mu)^2 + \cdots \\
=& \sqrt{N}\exp(\mu)(\bar{x}-\mu) + \frac{1}{2}\exp(\mu)O_p(1)o_p(1) \\
=& \sqrt{N}\exp(\mu)(\bar{x}-\mu) + o_p(1) \\
\xrightarrow{D}& \sqrt{N}\exp(\mu)(\bar{x}-\mu) \\
\overset{a}{\sim}& N\left(0, \exp(2\mu)\sigma^2\right)
\end{aligned}
$$

因而$Y \overset{a}{\sim} N\left(e^{\mu}, \frac{\exp(2\mu)\sigma^2}{N}\right)$

参数估计

- 在参数模型中，我们假设总体P属于某一个参数族$\{P_\theta, \theta \in \Theta\}$，从而推断总体等价于找到一个参数$\theta_0$，使得$P_{\theta_0} = P$。
- 我们一般把θ_0称为真值（true value）。
- 而由于总体是不可观测的，我们只能通过样本对总体进行推断，因而我们不可能得到θ_0的精确值，只能对其进行估计，即**参数估计**（Estimation）。

参数估计

- 参数估计包含两部分，即:
  - 点估计（point estimation）：找到一个统计量$\hat{\theta}(x)$，对总体参数θ_0进行推断，而统计量$\hat{\theta}$我们一般称为估计量（estimator）。
  - 区间估计（interval estimation）：找到一组统计量$L(x), U(x)$，使得由其组成的区间包含总体参数θ_0的概率为已知的，即

$$P(L(x) \leq \theta_0 \leq U(x)) = p$$

  其中统计量$L(x), U(x)$被称为区间估计量（interval estimator）。
- 估计量即样本的一个函数，即用于估计参数的统计量，而估计（estimate）是对于某一个样本，估计量的实现。

评价估计量的标准

对于同一个参数，经常我们有不同的估计量，比如：

方差的估计

- 对于正态总体$x_i \sim N\left(\mu, \sigma^2\right)$ $i.i.d$，自然地，σ^2的一个估计量为：

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

而类似的，我们也可以使用：

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \text{作为}\sigma^2$$

的一个估计。

- 以上两个估计量的差别在于分母的不同，很显然，以上两个估计量具有不同的抽样分布。

均方误差

- 那么，在有很多统计量可供选择时，该如何评价这些统计量呢？
- 一个常用的标准是**均方误差**（mean squared error, MSE），即对于一个参数θ和它的估计量$\hat{\theta}$，其误差平方的期望$\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2$为估计量$\hat{\theta}$的均方误差。
- 注意由于：

$$
\begin{aligned}
\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2 &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)^2+\mathbb{E}\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2+2\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)\left(\mathbb{E}\hat{\theta}-\theta_0\right) \\
&= \mathbb{V}\left(\hat{\theta}\right)+\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2+2\left(\mathbb{E}\hat{\theta}-\theta_0\right)\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right) \\
&= \mathbb{V}\left(\hat{\theta}\right)+\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathbb{V}\left(\hat{\theta}\right)+\left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2
\end{aligned}
$$

- 其中定义偏差（bias）：$\mathrm{Bias}\left(\hat{\theta}\right)=\mathbb{E}\left(\hat{\theta}-\theta_0\right)$

均方误差

- 从而$\text{MSE}\left(\hat{\theta}\right) = \mathbb{V}\left(\hat{\theta}\right) + \left[\text{Bias}\left(\hat{\theta}\right)\right]^2$，即均方误差等于估计量的方差与偏差平方的和。
  - 因而，降低均方误差有两种途径：降低估计的方差以及降低偏差。
- 此外，根据均方收敛的定义，只要

$$\mathbb{E}\left(\hat{\theta} - \theta_0\right)^2 = \mathbb{V}\left(\hat{\theta}\right) + \left[\text{Bias}\left(\hat{\theta}\right)\right]^2 \to 0$$

那么$\hat{\theta} \xrightarrow{L^2} \theta_0$，从而$\hat{\theta} \xrightarrow{p} \theta_0$。尽管小样本情况下估计量的偏差不为0，但是我们希望当样本量趋向于无穷时，估计量收敛到真值，也是可以接受的。

评价估计量的三个标准

- 这就引出了评价点估计量的三条标准：
  - 无偏性（unbiasedness）
  - 有效性（efficiency）
  - 一致性（consistency）。

无偏性

- 无偏性要求估计量的偏差为0。
- 当估计量的偏差为0，即$\mathbb{E}\left(\hat{\theta}\right)=\theta_0$时，我们称估计量$\hat{\theta}$为**无偏的**（**unbiased**）。
- 无偏性意味着，尽管对于每个样本，$\hat{\theta}$对θ_0的估计不可能完全准确而是有误差的，但是估计量$\hat{\theta}$总是围绕在真值θ_0的周围，平均而言（其期望）误差为0。

无偏性

均值的偏差

如果$\mathbb{E}x_i = \mu_0$，那么样本均值的期望：

$$\mathbb{E}\left(\bar{x}\right) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \mu_0$$

因而$\bar{x}$是总体均值μ_0的无偏估计。

无偏性

方差的偏差

根据之前的计算，我们知道对于$x_i \sim \left(\mu_0, \sigma_0^2\right)\ i.i.d$:

$$\mathbb{E}\left(s^2\right) = \sigma_0^2$$

因而$\hat{\sigma}^2$的期望为:

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\frac{N-1}{N}s^2\right) = \frac{N-1}{N}\sigma_0^2$$

因而$\text{Bias}\left(s^2\right) = 0, \text{Bias}\left(\hat{\sigma}^2\right) = \frac{1}{N}\sigma_0^2$，只有$s^2$是$\sigma_0^2$的无偏估计量。

有效性

- 为了降低MSE，除了降低偏差以外，降低估计量的方差$\mathbb{V}\left(\hat{\theta}\right)$也是非常重要的手段。
- 一般而言，如果两个估计量$\hat{\theta}_1$和$\hat{\theta}_2$，如果

$$\mathbb{V}\left(\hat{\theta}_1\right) < \mathbb{V}\left(\hat{\theta}_2\right)$$

那么我们称$\hat{\theta}_1$相对于$\hat{\theta}_2$是有效的（efficient）。

有效性

样本方差的有效性

在样本方差的例子中，因为$\hat{\sigma}^2 = \frac{N-1}{N}s^2$，从而

$$\mathbb{V}\left(\hat{\sigma}^2\right) = \left(\frac{N-1}{N}\right)^2 \mathbb{V}\left(s^2\right) < \mathbb{V}\left(s^2\right)$$

因而$\hat{\sigma}^2$是相对于s^2更有效的估计量。

偏差-方差权衡

- 我们注意到，尽管s^2比$\hat{\sigma}^2$的偏差更小，但是s^2比$\hat{\sigma}^2$的方差更大。
- 在很多应用问题，比如非参数回归或者监督学习（supervised learning）中，我们都会碰到类似的偏差-方差权衡（bias-variance tradeoff），即很多时候，同时降低偏差和方差是不可能的。

一致性

- 很多时候，尽管在有限样本下，一个估计量的偏差不为零，但是如果样本量足够大时，估计量与真值之间的误差充分的小，那么在样本量比较大时，我们也可以得到一个足够好的估计量。
- 如果一个估计量$\hat{\theta}$依概率收敛到真值θ_0，即$\hat{\theta} \xrightarrow{p} \theta_0$，那么我们称估计量$\hat{\theta}$为一致（consistent）估计量。
- 如果一个估计量是不一致的，也就是说即便我们拥有无限多的样本，我们也不能获得真值θ_0的估计，因而一致性是对一个估计量的最低要求。

一致性

样本均值的一致性

在样本均值的例子中，如果对样本$\{x_i\}$做额外的假设（比如x_i独立同分布且可积），那么根据大数定律，有

$$\bar{x} \xrightarrow{p} \mathbb{E}\left(x_i\right) = \mu_0$$

因而样本均值是总体均值的一致估计量。

一致性

样本方差的一致性

在样本方差的例子中，由于

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2$$

其中$\bar{x} \xrightarrow{p} \mu_0$，而

$$\frac{1}{N}\sum_{i=1}^{N}x_i^2 \xrightarrow{p} \mathbb{E}\left(x^2\right) = \mu_0^2 + \sigma_0^2$$

从而$\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$，即$\hat{\sigma}^2$是$\sigma_0^2$的一致估计量。而

$$s^2 = \frac{N}{N-1}\hat{\sigma}^2 \xrightarrow{p} \sigma_0^2$$

因而s^2也是σ_0^2的一致估计量。

无偏性与一致性

- 需要注意的是:
  - 无偏性关注的是估计量的期望
  - 而一致性则是当样本足够大时估计量的性质
- 两者并没有任何必然联系，无偏性和一致性并不是彼此的充分或者必要条件。

区间估计

- 在上一节中，无偏性、一致性都是使用单个估计量（如样本均值、样本方差）对未知参数进行估计，这种估计被称为点估计（point estimation）。
- 尽管我们可以使用点估计方法对参数值进行推断，然而我们知道，参数的点估计值$\hat{\theta}$与真值θ_0相等的概率一般为0，即$P\left(\hat{\theta}=\theta_0\right)=0$。
- 因而更进一步的，我们很多时候希望得到一个区间，使得这个区间能够以正的概率包含真值θ_0。这就诞生了**区间估计（interval estimation）**的概念。
- 区间估计即对于样本$x=(x_1,...,x_N)$，通过一对统计量$L(x)$和$U(x)$，满足$L(x)\leq U(x)$，我们可以使用区间$[L(x),U(x)]$对未知参数θ_0进行推断。

区间估计

区间包含真值的概率

如果样本$x_i \sim N(\mu_0, 1)\ i.i.d, i = 1, ..., N$，那么区间$[\bar{x} - 0.5, \bar{x} + 0.5]$包含真值$\mu_0$的概率为：

$$P(\mu_0 \in [\bar{x} - 0.5, \bar{x} + 0.5]) = P(\mu_0 - 0.5 \leq \bar{x} \leq \mu_0 + 0.5)$$
$$= P\left(-\frac{0.5}{\sqrt{\frac{1}{N}}} \leq \frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \leq \frac{0.5}{\sqrt{\frac{1}{N}}}\right)$$

由于$\bar{x} \sim N\left(\mu_0, \frac{1}{N}\right)$，因而

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{1}{N}}} \sim N(0, 1)$$

区间估计

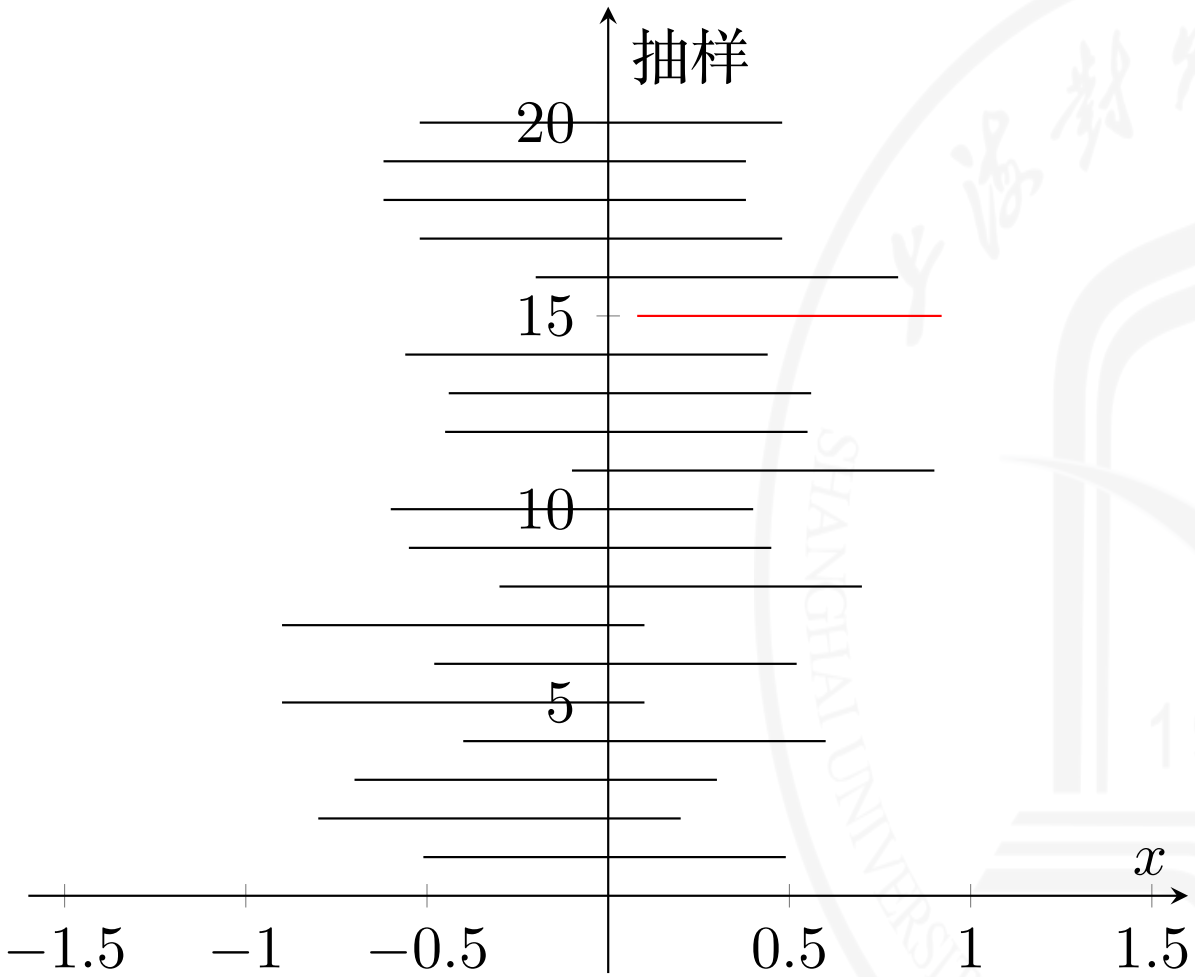
区间包含真值的概率

因而

$$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - \Phi\left(-\frac{0.5}{\sqrt{\frac{1}{N}}}\right)$$
$$= 2\Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - 1$$

例如，当$N=16$时，查表可得，$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = 2\Phi(2) - 1 \approx 2 \times 0.9772 - 1 = 0.9544$，即区间$[\bar{x}-0.5, \bar{x}+0.5]$包含真值$\mu_0$的概率为95.44%。

区间估计



置信区间

- 注意由于未知参数θ_0是一个未知的常数，而统计量$L(x)$和$U(x)$是随着抽样的变化而变化的，因此我们不能说「θ_0落入区间$[\bar{x}-0.5,\bar{x}+0.5]$的概率是多少」，而只能说「区间$[\bar{x}-0.5,\bar{x}+0.5]$包含$\theta_0$的概率是多少」。
- 我们把概率$P(\theta_0 \in [L(x), U(x)])$称为覆盖概率（coverage probability）。
- 注意由于总体参数θ_0未知，因而概率$P(\mu_0 \in [L(x), U(x)])$可能依赖于未知的参数$\theta_0$，因而我们通常将而覆盖概率的下界，即$\inf_\theta P_\theta(\theta_0 \in [L(x), U(x)])$称为置信度（confidence coefficient）或者置信水平，通常用$1-\alpha$表示。
- 在某一置信度下，区间$[L(x), U(x)]$又被称为置信区间（confidence interval）。
  - 因而上例中，我们可以说在95.44%的置信水平下，置信区间为$[\bar{x}-0.5,\bar{x}+0.5]$。

基准统计量

- 此外还需要注意的是，在上例中，为了求得置信区间和覆盖概率，我们首先将统计量$\bar{x}$做了标准化处理，即使用$(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$推算概率，而不是直接使用$\bar{x}$。
- 使用$(\bar{x}-\mu_0)/\sqrt{\frac{1}{N}}$的好处是，此统计量的抽样分布不依赖于任何未知参数，因而其分布不会随着未知参数的变化而变化，即服从一个“标准的”分布，这样一来，我们得到的覆盖概率不依赖于任何未知参数。
- 一般的，我们把分布不依赖于未知参数的统计量成为基准统计量（pivotal statistic）。

基准统计量

样本均值与基准统计量

如果样本$x_i \sim N\left(\mu_0, \sigma_0^2\right)\ i.i.d, i = 1, ..., N$，那么：

1. 统计量$\bar{x} \sim N\left(\mu_0, \frac{\sigma_0^2}{N}\right)$，其分布依赖于两个未知参数；
2. 统计量$\bar{x} - \mu_0 \sim N\left(0, \frac{\sigma_0^2}{N}\right)$，其分布仍然依赖于未知参数$\mu_0$；
3. 统计量$\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma_0^2}{N}}} \sim N(0,1)$分布不依赖于任何未知参数，因而是基准统计量；
4. 统计量$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \sim t(N-1)$分布不依赖于任何未知参数，因而是基准统计量；

基准统计量

样本方差与基准统计量

如果样本$x_i \sim N\left(\mu_0, \sigma_0^2\right)\ i.i.d, i=1,...,N$，那么$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi^2(N-1)$，其分布不依赖于任何未知参数，因而是基准统计量。

置信区间的构造

- 之前我们首先给出了区间，进而计算了该区间的置信度。
- 然而现实中，我们经常希望得到在一定置信水平下的置信区间，即一般的区间估计过程。
- 由于基准统计量的分布不依赖任何未知参数，因而一般情况下，使用基准统计量可以很方便地构造置信区间。

置信区间的构造

正态分布均值的置信区间

如果样本$x_i \sim N\left(\mu_0, \sigma_0^2\right)\ i.i.d, i = 1, ..., N$，为了得到$\mu_0$的95%的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有μ_0是未知的，其他都是已知的（包括已知常数以及已知统计量）。在上例中，只有统计量

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim t\left(N - 1\right)$$

满足以上条件。

置信区间的构造

正态分布均值的置信区间

如果令$t_{\alpha/2} = F_t^{-1}\left(1-\frac{\alpha}{2}\right)$，我们有：

$$
\begin{aligned}
P\left(-t_{\alpha/2} \le \frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \le t_{\alpha/2}\right) &= F_t\left(t_{\alpha/2}\right) - F_t\left(-t_{\alpha/2}\right) \\
&= 1 - 2F_t\left(t_{\alpha/2}\right) \\
&= 1 - 2F_t\left(F_t^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \alpha
\end{aligned}
$$

因而我们可以得到：

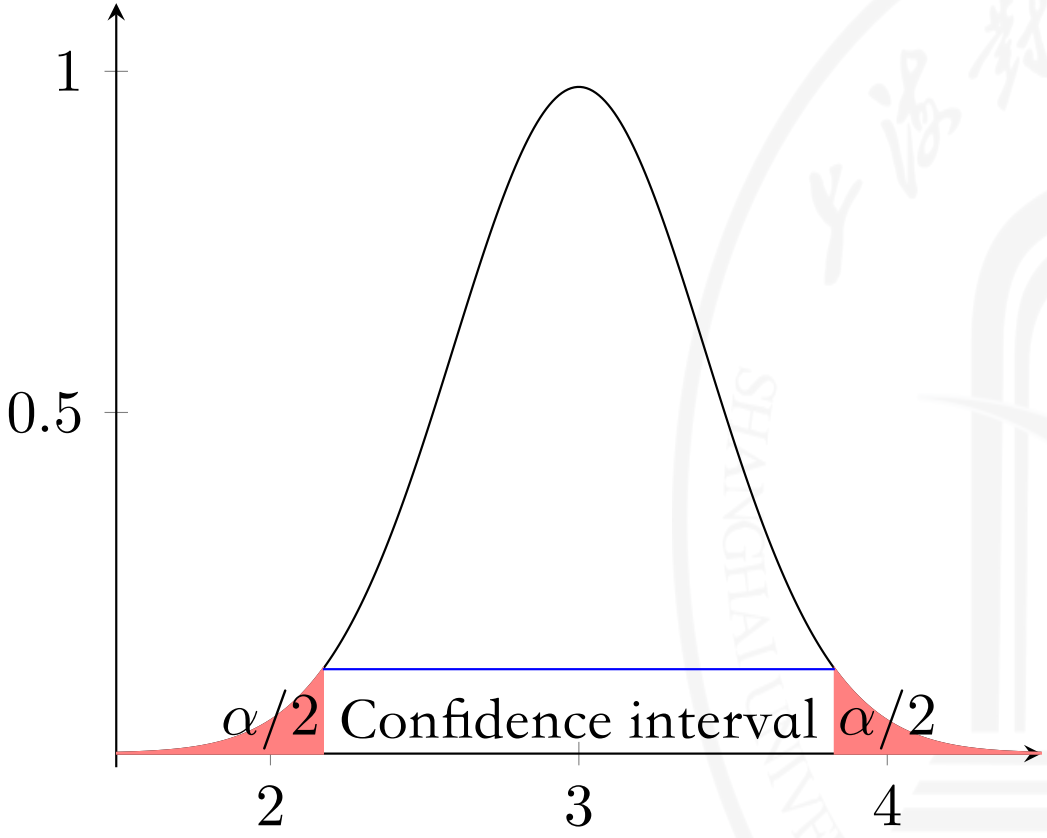
$$
P\left(\bar{x} - t_{\alpha/2}\sqrt{\frac{s^2}{N}} \le \mu_0 \le \bar{x} + t_{\alpha/2}\sqrt{\frac{s^2}{N}}\right) = 1 - \alpha
$$

置信区间的构造

正态分布均值的置信区间

从而$\left[\bar{x}-t_{\alpha/2}\sqrt{\frac{s^2}{N}},\bar{x}+t_{\alpha/2}\sqrt{\frac{s^2}{N}}\right]$就是我们想要的置信区间。例如，对于一个$N=30$的正态样本，$\bar{x}=3$，$s^2=5$，如果我们想要得到95%置信水平下的置信区间，查表（d.f.=29）得到$t_{\alpha/2}=2.0452$，因而置信下界为$3-2.0452\times\sqrt{5/30}\approx 2.17$，置信上界为$3+2.0452\times\sqrt{5/30}\approx 3.83$。如下图所示，其中红色区域为左右两个概率为$\alpha/2$的区域，中间的一块即为所要求的置信区间。

置信区间的构造



置信区间的构造

正态分布样本方差的置信区间

如果样本$x_i \sim N\left(\mu_0, \sigma_0^2\right)\ i.i.d, i = 1, ..., N$，为了得到$\sigma_0^2$的95%的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有σ_0^2是未知的，其他都是已知的。在上例中，统计量

$$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi^2(N-1)$$

满足以上条件。

置信区间的构造

正态分布样本方差的置信区间

如果令

$$\chi^2_{\alpha/2} = F^{-1}_{\chi^2}\left(\frac{\alpha}{2}\right), \quad \chi^2_{1-\alpha/2} = F^{-1}_{\chi^2}\left(1-\frac{\alpha}{2}\right)$$

我们有:

$$P\left(\chi^2_{\alpha/2} \le (N-1)\frac{s^2}{\sigma_0^2} \le \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

因而:

$$P\left(\frac{(N-1)\,s^2}{\chi^2_{1-\alpha/2}} \le \sigma_0^2 \le \frac{(N-1)\,s^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha\sigma_0^2$$

的95%的置信区间为: $\left[\frac{(N-1)s^2}{\chi^2}, \frac{(N-1)s^2}{\chi^2}\right]$。

置信区间的构造步骤

总结上述两个置信区间的计算，一般而言我们得到置信区间的步骤如下：

1. 给定置信度$1-\alpha$；
2. 找到一个基准统计量，其中只有所要求的参数是未知的，其他都是已知的；
3. 找到这个基准统计量的分布函数$F(\cdot)$；
4. 查表或使用计算机计算$F^{-1}\left(\frac{\alpha}{2}\right)$以及$F^{-1}\left(1-\frac{\alpha}{2}\right)$；
5. 通过不等式变换得到置信区间。

置信区间的模拟

- 我们可以使用程序验证以上步骤得到的置信区间包含真值的概率刚好为置信度$1-\alpha$。
- 代码CI_small_sample.do给出了正态均值的区间估计中小样本均值的区间估计的模拟。
- 在以上程序中，我们首先定义了一个抽样过程，即从$N(5,100)$的正态总体中进行抽样，接下来使用公式

$$\left[\bar{x}-t_{\alpha/2}\sqrt{\frac{s^2}{N}},\bar{x}+t_{\alpha/2}\sqrt{\frac{s^2}{N}}\right]$$

计算置信区间，并验证置信区间是否包含了真值（5）。重复抽样10000次，我们就得到了10000个置信区间，最终计算出置信区间包含真值的比率为94.93%，结果显示与与95%相差无几。

大样本下的置信区间

- 尽管上两例给出了正态总体的均值和方差的置信区间的计算方法，然而很多时候我们的总体并不是一定来自于正态总体，很多时候我们很难计算在非正态总体下样本均值的精确分布。
- 然而根据中心极限定理，在一定条件下，有：

$$\sqrt{N}\left(\bar{x}-\mu_0\right) \overset{a}{\sim} N\left(0, \mathbb{V}\left(x\right)\right)$$

- 根据上式，我们有：

$$\frac{\sqrt{N}\left(\bar{x}-\mu_0\right)}{\sqrt{\mathbb{V}\left(x\right)}} \overset{a}{\sim} N\left(0,1\right)$$

然而以上公式中，$\mathbb{V}\left(x\right)$是未知的，因而不能直接用于假设检验。

大样本下的置信区间

- 根据Slutsky定理，我们可以使用样本方差s^2代替$\mathbb{V}(x)$，由于$s^2 \xrightarrow{p} \mathbb{V}(x)$，因而不改变分子上的渐近分布，即有：

$$\frac{\sqrt{N}(\bar{x}-\mu_0)}{\sqrt{s^2}} \overset{a}{\sim} N(0,1)$$

因而我们可以使用上式进行区间估计。

- 使用类似的技巧，有：

$$P\left(\left|\frac{\sqrt{N}(\bar{x}-\mu_0)}{s}\right| \leq z_{\alpha/2}\right) = 1-\alpha$$

其中$z_{\alpha/2} = \Phi_t^{-1}\left(1-\frac{\alpha}{2}\right)$为标准正态分布的上$\alpha/2$分位数，从而置信区间为$\left[\bar{x} - z_{\alpha/2}\sqrt{\frac{s^2}{N}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{s^2}{N}}\right]$。

大样本下的置信区间

收入的置信区间

根据2009年中国城镇住户调查，在37480户家庭中，已知家庭年收入均值为54157.63 元，标准差为38533.96元，那么全国家庭家庭平均收入的95%置信区间是多少?

- 一般而言，收入不服从正态分布，但是在大样本条件下，我们知道样本均值近似服从正态分布
- 如果取$1-\alpha=95\%$，查表知$z_{2.5\%}=1.96$，因而:
  - 置信下界为：$54157.63-1.96\times\sqrt{\frac{38533.96^2}{37480}}\approx 53767.50$
  - 置信上界约为54547.75因而全国家庭家庭平均收入的95%置信区间为$[53767.50, 54547.75]$。

大样本下的置信区间

比率的置信区间

根据2013年中国家庭金融调查，样本7711户家庭中，有6%的家庭有信用卡，请问全国持有信用卡的家庭比例的95%置信区间是多少?

- 同样的，比例一般不服从正态分布，但是如果把每个家庭是否持有信用卡假设为独立同分布的伯努利分布，
  即$x_i \sim Ber\left(p_0\right)$，那么$x_i^2 = x_i$，因而$\overline{x^2} = \bar{x}$，从而

$$\frac{s^2}{N} = \frac{N}{N-1}\frac{\overline{x^2} - \bar{x}^2}{N} \approx \frac{\overline{x^2} - \bar{x}^2}{N} = \frac{\bar{x} - \bar{x}^2}{N} = \frac{\bar{x}\left(1 - \bar{x}\right)}{N}$$

  其中比例$\hat{p} = \bar{x}$，从而

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \approx \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}}$$

大样本下的置信区间

比率的置信区间

- 由于$\hat{p} \xrightarrow{p} p_0$，从而$\hat{p}(1-\hat{p}) \xrightarrow{p} p_0(1-p_0) = \mathbb{V}(x_i)$，从而根据Slutsky定理：

$$\frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \overset{a}{\sim} N(0,1)$$

- 那么
  - 置信下界为

$$\hat{p}-z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = 0.06-1.96\times\sqrt{\frac{0.06\times(1-0.06)}{7711}} \approx 5.47\%$$

  - 同理置信上界约为6.53%
  - 因而全国家庭持有信用卡比例的95%置信区间为$[5.47\%, 6.53\%]$。

大样本下的置信区间的模拟

- 我们同样可以使用程序验证以上大样本条件下均值的区间估计。
- CI_large_sample.do给出了一个模拟
- 以上程序与之前小样本的区间估计模拟类似，区别在于我们放弃了正态性的假设，转而生成了一个$N(-3,1)$和$N(3,1)$的混合正态，并在不同的样本量下观察其覆盖率。
- 结果发现，当样本量仅为10时，覆盖率为91.37%，的确不能保证置信区间95%的覆盖了真值，但是当样本量逐渐增大时，覆盖率会接近95%。

两个样本均值比较

- 此外，很多时候我们还对两个样本的差值感兴趣。如果假设两个独立的样本x_1和x_2，其均值分别为$\bar{x}_1$和$\bar{x}_2$，
- 样本均值差值的期望为

$$\mathbb{E}\left(\bar{x}_1 - \bar{x}_2\right) = \mu_1 - \mu_2$$

由于独立性，方差为

$$\mathbb{V}\left(\bar{x}_1 - \bar{x}_2\right) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

两个样本均值比较

- 由于$\bar{x}_k, k=1,2$在大样本条件下分别渐近服从正态分布，从而其差值经过标准化后：

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}} \overset{a}{\sim} N(0,1)$$

- 同样，由于σ_1^2和σ_2^2不可观测，我们分别用其样本方差代替，即：

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}} \overset{a}{\sim} N(0,1)$$

从而置信区间为$\left[\bar{x}_1-\bar{x}_2-z_{\alpha/2}\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}, \bar{x}_1-\bar{x}_2+z_{\alpha/2}\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}\right]$。

两个样本均值比较

不同性别收入比较

在2009年中国城镇住户调查中，共有23440位20-50岁的男性，以及21184位20-50岁的女性。已知男性年平均收入为28367.96元，标准差为21811.88元；女性年平均收入为20145.77元，标准差为16541.08元。如果假设男女收入独立，请问男女收入差异的95%置信区间是多少?

- 同上，尽管收入不服从正态分布，但是大样本情况下可以使用正态分布近似。其中$\bar{x}_1 - \bar{x}_2 = 28367.96 - 20145.77 = 8222.19$，

$$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} = \sqrt{\frac{21811.88^2}{23440} + \frac{16541.08^2}{21184}} = 182.24$$

  因而其差值的置信区间的下界为$8222.19 - 1.96 \times 182.24 = 7864.99$，同理得到置信上界，最终置信区间为：$[7864.99, 8579.38]$。

样本量的确定

- 区间估计中区间的长度实际上度量了精度：如果区间估计的长度越短，那么意味着在一定的置信水平下，我们可以把真值的范围缩的更小。
- 而反过来，根据区间估计，我们还能确定为了达到某一精度所需要样本量的大小。
- 比如，为了确定给定精度下样本均值对总体均值的样本量，根据中心极限定理，在一定条件下
有$\sqrt{N}\left(\bar{x}-\mu_0\right) \overset{a}{\sim} N\left(0, \mathbb{V}\left(x\right)\right)$

样本量的确定

- 因而大样本条件下，均值在$1-\alpha$置信水平下的置信区间为：$\left[\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{N}},\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right]$区间长度应为：$\frac{2z_{\alpha/2}\sigma}{\sqrt{N}}$。
- 可以看到，区间大小随着样本量的增加而减小。
- 如果我们要求在$1-\alpha$置信水平下的置信区间的长度为l，那么样本量应为$N=\left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2$，即样本量与精度成二次方关系。

样本量的确定

比率问题中样本量的确定

在信用卡比例的例子中，为了使95%置信区间长度不超过1%，需要的样本量为：

$$N = \left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2 = \left[\frac{2 \times 1.96 \times \sqrt{0.06 \times (1-0.06)}}{0.01}\right]^2 \approx 8667$$

即需要8667户样本。

样本量的确定

比率问题中样本量的确定

然而实际情况中，我们不太可能知道σ，所以有时预调研是非常重要的。不过在这个例子中，我们可以得到一个样本量的上界：

$$N = \left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2 = \left[\frac{2z_{\alpha/2}\sqrt{p(1-p)}}{l}\right]^2$$
$$\leq \left[\frac{2z_{\alpha/2}\sqrt{0.5(1-0.5)}}{l}\right]^2 = \left[\frac{z_{\alpha/2}}{l}\right]^2$$

即在这个例子中，如果我们不知道持有信用卡的家庭约为6%，那么需要的样本数量上界为38416户家庭。

矩估计

- 接下来我们介绍一些经典的点估计方法。
- 矩估计（method of moments）是使用历史最长的参数估计方法，其思路是使用样本矩代替总体矩对参数进行估计。
- 接下来我们将介绍经典的矩估计方法，并对此方法做进一步推广。

矩估计的思想

- 如果样本$x=(x_1,...,x_N)'$是来自于总体P_{θ_0}的独立同分布的样本
- 其一阶样本矩和一阶总体矩可以分别定义为

$$\begin{cases} m_1(x)=\frac{1}{N}\sum_{i=1}^{N}x_i \\ \mu_1(\theta)=\mathbb{E}_\theta x_i \end{cases}$$

其中$\mathbb{E}_\theta$表示给定一个参数θ，使用总体P_θ计算得到的理论的总体期望。

  - 由于真值为θ_0，因而真实的期望$\mathbb{E}x_i=\mathbb{E}_{\theta_0}x_i$。

矩估计的思想

- 我们知道，在一定比较宽松的条件下，根据大数定律有：

$$m_1(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{p} \mathbb{E}_{\theta_0} x_i = \mu_1(\theta_0)$$

- 如果$\mu_1(\cdot)$是一个连续且可逆的函数，那么真实参数θ_0可以写为：

$$\theta_0 = \mu_1^{-1}(\mu_1(\theta_0))$$

- 那么我们可以使用样本矩$m_1(x)$代替上式中的总体矩$\mu_1(\theta_0)$，由于$m_1(x) \xrightarrow{p} \mu_1(\theta_0)$，而$\mu_1^{-1}(\cdot)$为连续函数，从而估计量：

$$\hat{\theta} \triangleq \mu_1^{-1}(m_1(x)) \xrightarrow{p} \mu_1^{-1}(\mu_1(\theta_0)) = \theta_0$$

从而$\hat{\theta}$是θ_0的一致估计。

矩估计的思想

- 形象的理解是，给定任何一个θ，总体P_θ是一个确定的概率函数，因而可以计算在θ情况下的样本矩$\mathbb{E}_\theta x_i$。
- 理论上，样本矩$m_1(x)$和总体矩$\mu_1(\theta)$在样本量足够大的情况下应该是充分接近的
- 那么我们可以找到一个$\hat{\theta}$使得$\mu_1\left(\hat{\theta}\right)$与$m_1(x)$的差距最小，从而得到对真值$\theta_0$的估计。
- 以上就是矩估计的思想。

矩估计

泊松分布的矩估计

- 如果样本$x_i \sim P(\lambda_0)\ i.i.d$，我们知道样本矩$m_1(x) = \bar{x}$，比如，如果我们的样本观测值为$x = (3, 5, 7, 2, 3)$，那么样本矩为$m_1(x) = \bar{x} = \frac{3+5+7+2+3}{5} = 4$
- 假如任意给定一个λ，比如令$\lambda = 2$，总体的期望为$\mathbb{E}_\lambda x_i = \lambda = 2 \neq 4$，因而如果认为$\lambda_0 = 2$，那么总体$P(2)$所产生的总体矩与样本矩仍然有差异。
- 只有当$\lambda = 4$时，总体矩$\mathbb{E}_\lambda x_i = 4 = m_1(x)$，总体矩与我们观察到的样本矩相等，因而我们可以推断$\hat{\lambda} = 4$。

矩估计

泊松分布的矩估计

一般的，对于泊松分布总体，我们可以直接令总体矩等于样本矩得到估计，即：$\hat{\lambda} = m_1(x) = \bar{x}$下面我们分别讨论该估计量的无偏性和一致性。

- 首先，对于无偏性，由于：

$$\mathbb{E}\hat{\lambda} = \mathbb{E}\bar{x} = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}x_i = \lambda_0$$

  因而$\hat{\lambda}$是λ_0的无偏估计。
- 而对于一致性，根据大数定理：$\hat{\lambda} = \bar{x} \xrightarrow{p} \mathbb{E}x_i = \lambda_0$因而$\hat{\lambda}$是$\lambda_0$的一致估计。

矩估计

泊松分布的矩估计

- 当然，一致性还可以通过分析$\hat{\lambda}$的偏差与方差来证明。
  - 根据以上讨论，该估计量的偏差为$\text{Bias}\left(\hat{\lambda}\right)=\mathbb{E}\left(\hat{\lambda}\right)-\lambda_0=0$，而其方差为：

$$\mathbb{V}\left(\hat{\lambda}\right)=\mathbb{V}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)=\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}\left(x_i\right)=\frac{\lambda_0}{N}$$

从而

$$\mathbb{E}\left(\hat{\lambda}-\lambda_0\right)^2=\mathbb{V}\left(\hat{\lambda}\right)+\left[\text{Bias}\left(\hat{\lambda}\right)\right]^2\to 0$$

从而$\hat{\lambda}\xrightarrow{L^2}\lambda_0$，从而$\hat{\lambda}\xrightarrow{p}\lambda_0$。

- 进一步，根据中心极限定理，有：

$$\sqrt{N}\left(\hat{\lambda}-\lambda_0\right)=\sqrt{N}\left(\bar{x}-\lambda_0\right)\overset{a}{\sim}N\left(0,\lambda_0\right)$$

矩估计

对数正态分布的矩估计

- 如果样本$x_i \sim LN(\mu_0, 2)\ i.i.d$，即总体为对数正态分布，且一个参数$\sigma^2 = 2$已知。类似的，样本矩$m_1(x) = \bar{x}$，而总体矩$\mathbb{E}_\mu x_i = e^{\mu+1}$。
- 根据矩估计的思想，令总体矩等于样本矩，即：

$$e^{\hat{\mu}+1} = m_1(x) = \bar{x}$$

可以得到μ_0的矩估计值：$\hat{\mu} = \ln \bar{x} - 1$

矩估计

对数正态分布的矩估计

- 现在讨论该估计量的无偏性和一致性。首先根据Jensen不等式：

$$\mathbb{E}\left(\hat{\mu}\right) = \mathbb{E}\left(\ln \bar{x}\right) - 1 \leq \ln\left(\mathbb{E}\bar{x}\right) - 1 = \ln e^{\mu_0+1} - 1 = \mu_0$$

  因而$\hat{\mu}$并不是μ_0的无偏估计。
- 而根据大数定律，$\bar{x} \xrightarrow{p} \mathbb{E}x_i = e^{\mu_0+1}$，由于ln为连续函数，从而：

$$\hat{\mu} = \ln \bar{x} - 1 \xrightarrow{p} \ln e^{\mu_0+1} - 1 = \mu_0$$

  因而$\hat{\mu}$是μ_0的一致估计。

矩估计

对数正态分布的矩估计

- 此外，我们还可以使用delta方法计算$\hat{\mu}$的极限分布。
- 根据中心极限定理：

$$\sqrt{N}\left(\bar{x}-e^{\mu_0+1}\right)\xrightarrow{D}N\left(0,\mathbb{V}\left(x_i\right)\right)$$

其中$\mathbb{V}\left(x_i\right)=e^{2(\mu_0+2)}-e^{2\mu_0+2}$。

- 因而，对估计量在e^{μ_0+1}处进行泰勒展开：

$$\begin{aligned}\sqrt{N}\left(\hat{\mu}-\mu_0\right)&=\sqrt{N}\left(\ln\bar{x}-1-\ln e^{\mu_0+1}+1\right)\\&=\sqrt{N}\left(\ln\bar{x}-\ln e^{\mu_0+1}\right)\\&=\sqrt{N}\left(\ln e^{\mu_0+1}+\frac{1}{e^{\mu_0+1}}\left(\bar{x}-e^{\mu_0+1}\right)+O\left(\left(\bar{x}-e^{\mu_0+1}\right)^2\right)-\ln e^{\mu_0+1}\right)\\&=\sqrt{N}\frac{1}{e^{\mu_0+1}}\left(\bar{x}-e^{\mu_0+1}\right)+o_p\left(1\right)\xrightarrow{D}\frac{1}{e^{\mu_0+1}}\sqrt{N}\left(\bar{x}-e^{\mu_0+1}\right)\\&\overset{a}{\sim}N\left(0,e^{-2(\mu_0+1)}\mathbb{V}\left(x_i\right)\right)\end{aligned}$$

矩估计与区间估计

- 实际上，在以上两个例子中，我们不仅使用矩估计得到了未知参数的点估计，并使用大数定律验证了其一致性，还使用了中心极限定理和delta方法计算了矩估计量的大样本分布。
- 比如，直接使用中心极限定理可以得到，在泊松分布中:

$$\sqrt{N}\left(\hat{\lambda}-\lambda_0\right) \overset{a}{\sim} N\left(0, \lambda_0\right)$$

使用这一结论，我们可以方便的构建真值的区间估计。比如在上面的例子中，我们知道:

$$P\left(\left|\frac{\hat{\lambda}-\lambda_0}{\sqrt{\frac{\lambda_0}{N}}}\right| \leq 1.96\right)=95\%$$

- 因而对以上不等式稍加改变，即得到:

$$P\left(\hat{\lambda}-1.96 \times \sqrt{\frac{\lambda_0}{N}} \leq \lambda_0 \leq \hat{\lambda}+1.96 \times \sqrt{\frac{\lambda_0}{N}}\right)=95\%$$

因而λ_0的95%的置信区间应该

多个参数的矩估计

如果我们有k个未知参数，即θ为k维向量，那么我们可以联立前k个样本矩和总体矩对θ进行估计，其中前k个样本矩和总体矩定义为：

$$\begin{cases} m_1(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^1 & \mu_1(\theta) = \mathbb{E}_\theta x_i^1 \\ m_2(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 & \mu_2(\theta) = \mathbb{E}_\theta x_i^2 \\ \vdots & \vdots \\ m_k(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^k & \mu_k(\theta) = \mathbb{E}_\theta x_i^k \end{cases}$$

多个参数的矩估计

一般而言，如果我们有k个参数，$\theta = (\theta_1, ..., \theta_k)'$，那么我们使用前$k$个矩，解方程：

$$\begin{cases} m_1(x) = \mu_1\left(\hat{\theta}\right) \\ m_2(x) = \mu_2\left(\hat{\theta}\right) \\ \vdots \\ m_k(x) = \mu_k\left(\hat{\theta}\right) \end{cases}$$

如果该联立方程有解，即可得到参数θ_0的估计。

多个参数的矩估计

正态分布的矩估计

- 对于正态总体$x_i \sim N\left(\mu_0, \sigma_0^2\right)\ i.i.d$，其中未知总体参数$\theta = \left(\mu, \sigma^2\right)$，其一阶样本矩为$m_1(x) = \bar{x}$，二阶样本矩为$m_2(x) = \overline{x^2}$。
- 我们知道对于正态分布，$\mu_1(\theta) = \mu, \mu_2(\theta) = \mu^2 + \sigma^2$，从而矩估计为：

$$\begin{cases} m_1(x) = \bar{x} = \hat{\mu} \\ m_2(x) = \overline{x^2} = \hat{\mu}^2 + \hat{\sigma}^2 \end{cases}$$

解得：

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2 \end{cases}$$

多个参数的矩估计

正态分布的矩估计

下面分析其无偏性和一致性。

- 根据之前的结论，$\mathbb{E}\hat{\mu} = \mu_0, \mathbb{E}\left(\hat{\sigma}^2\right) = \frac{N-1}{N}\sigma_0^2$，因而$\hat{\mu}$是无偏估计量而$\hat{\sigma}^2$并非无偏估计量。
- 而由于$\bar{x} \xrightarrow{p} \mu_0$，$\overline{x^2} \xrightarrow{p} \mu_0^2 + \sigma_0^2$，从而$\hat{\sigma}^2 \xrightarrow{p} \mu_0^2 + \sigma_0^2 - \mu_0^2 = \sigma_0^2$，因而$\hat{\mu}$和$\hat{\sigma}^2$都是一致估计量。

更加一般的矩估计

接下来，我们将讨论矩估计的一般形式。

- 对于一个统计模型，如果我们关心真实参数$\theta_0 \in \Theta \subset \mathbb{R}^K$
- 只要我们可以找到K个矩条件，使得θ_0为矩条件（moment condition）方程：

$$\mathbb{E}\left[g\left(w_i, \theta\right)\right] = \mathbb{E}\left(\begin{bmatrix} g_1\left(w_i, \theta\right) \\ \vdots \\ g_K\left(w_i, \theta\right) \end{bmatrix}\right) = 0$$

的唯一解，那么我们就可以解以上总体矩方程组的样本方程等价形式：

$$\frac{1}{N}\sum_{i=1}^{N} g\left(w_i, \hat{\theta}\right) = 0$$

解得$\hat{\theta}$

- 可以证明，在一些额外比较宽松的条件下，$\hat{\theta} \xrightarrow{p} \theta_0$。

更加一般的矩估计

对数正态分布的矩估计

- 之前我们已经计算了对数正态分布的矩估计，我们使用了矩条件$\mathbb{E}_\mu x_i = e^{\mu+1}$进行估计
- 而实际上我们也可以使用x_i其他函数的期望进行估计
- 比如，我们可以使用：$\mathbb{E}_\mu\left(\ln x_i\right) = \mu$，从而一个自然的估计为：$\hat{\mu} = \overline{\ln x} = \frac{1}{N}\sum_{i=1}^{N}\ln\left(x_i\right)$
- 可以验证以上估计是一个无偏、一致估计。

识别

- 注意我们要求θ_0为矩条件方程组的唯一解，即模型是**可识别的**（identifiable），这要求矩条件方程不仅有解，而且解唯一。
  - 如果矩条件方程有不止一组解，那么我们无法区分真实参数究竟是哪一组解，导致该统计问题无法准确回答。
- **识别**（identification）问题是计量经济学的核心问题。

识别问题

Beta分布的识别问题

如果$x_i \sim Beta\left(\alpha, \beta\right)$，我们知道其前两阶矩的矩条件为：

$$\begin{cases} \mathbb{E}\left(x_i\right) = \frac{\alpha}{\alpha+\beta} \\ \mathbb{E}\left(x_i^2\right) = \frac{\alpha\beta+\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{cases}$$

因而其矩估计为：

$$\begin{cases} \frac{\alpha}{\alpha+\beta} = \bar{x} \\ \frac{\alpha\beta+\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} = \overline{x^2} \end{cases}$$

识别问题

Beta分布的识别问题

- 然而在估计以上问题之前，我们必须首先证明识别问题，即真值(α_0, β_0)为矩条件的唯一解，或者不存在另外的(α_1, β_1)使得：

$$\begin{cases} \frac{\alpha_1}{\alpha_1+\beta_1} = \frac{\alpha_0}{\alpha_0+\beta_0} \\ \frac{\alpha_1\beta_1+\alpha_1^2(\alpha_1+\beta_1+1)}{(\alpha_1+\beta_1)^2(\alpha_1+\beta_1+1)} = \frac{\alpha_0\beta_0+\alpha_0^2(\alpha_0+\beta_0+1)}{(\alpha_0+\beta_0)^2(\alpha_0+\beta_0+1)} \end{cases}$$

- 如果存在着这样的(α_1, β_1)，那么使用(α_1, β_1)和使用真值(α_0, β_0)所代表的总体有相同的一、二阶矩，因而我们无法判断哪一个是对的。

识别问题

Beta分布的识别问题

为了证明以上结论，用反证法，现在假设存在这样的(α_1, β_1)，根据第一个等式，那么必然有：$\frac{\alpha_1}{\beta_1} = \frac{\alpha_0}{\beta_0} \triangleq c_0$带入到第二个式子，有：

$$\begin{aligned}
\frac{\alpha_1\beta_1 + \alpha_1^2(\alpha_1+\beta_1+1)}{(\alpha_1+\beta_1)^2(\alpha_1+\beta_1+1)} &= \frac{\beta_1 c_0\beta_1 + \beta_1^2 c_0^2(\beta_1 c_0+\beta_1+1)}{(\beta_1 c_0+\beta_1)^2(\beta_1 c_0+\beta_1+1)} \\
&= \frac{c_0\beta_1^2 + \beta_1^2 c_0^2(\beta_1 c_0+\beta_1+1)}{\beta_1^2(c_0+1)^2(\beta_1 c_0+\beta_1+1)} \\
&= \frac{c_0 + c_0^2(\beta_1 c_0+\beta_1+1)}{(c_0+1)^2(\beta_1 c_0+\beta_1+1)}
\end{aligned}$$

如果存在这样的(α_1, β_1)，那么上式必然可以被整理为仅仅是c_0的函数才能使得等式左右两边相等，然而显然并不存在这样的形式。因而，使用前两阶矩是完全可以识别(α_0, β_0)的。

识别问题

外汇业务人数的识别问题

- 如果$M|N \sim Bi\left(N,p\right)$，$N \sim P\left(\lambda\right)$，$(p,\lambda)$为未知参数。
- 现在，如果我们只观察到M，而N观察不到，如果我们试图使用M的前两阶矩估计p和λ，有：

$$\begin{cases} \mathbb{E}\left(M\right) & = \mathbb{E}\left[\mathbb{E}\left(M|N\right)\right] = \mathbb{E}\left(Np\right) = \lambda p \\ \mathbb{E}\left(M^2\right) & = \mathbb{V}\left(M\right) + \left[\mathbb{E}\left(M\right)\right]^2 \\ & = \mathbb{V}\left[\mathbb{E}\left(M|N\right)\right] + \mathbb{E}\left[\mathbb{V}\left(M|N\right)\right] + \lambda^2 p^2 \\ & = \mathbb{V}\left(Np\right) + \mathbb{E}\left[Np\left(1-p\right)\right] + \lambda^2 p^2 \\ & = p^2\lambda + p\left(1-p\right)\lambda + \lambda^2 p^2 \\ & = \lambda p + \lambda^2 p^2 \end{cases}$$

识别问题

外汇业务人数的识别问题

- 联立以上方程得到了一个恒等式：$\mathbb{E}\left(M^2\right)=\mathbb{E}\left(M\right)+\left[\mathbb{E}\left(M\right)\right]^2$或者$\mathbb{V}\left(M\right)=\mathbb{E}\left(M\right)$，无法单独解出$\lambda$或者$p$。
- 实际上，在这种情况下，由于$\mathbb{V}\left(M\right)=\mathbb{E}\left(M\right)$，因而第二个方程是多余的，而如果只使用第一个方程，一个方程无法解除两个参数，即有无穷多组解。因而如果只使用前两阶矩估计p和λ，是无法识别的。

一元线性回归

在保证识别的前提下，很多统计学问题都可以使用矩估计进行参数估计。

一元线性回归

- 给定数据$w_i = (y_i, x_i)'$，我们希望估计条件期望：$\mathbb{E}(y_i|x_i)$。如果我们假定条件期望为线性函数形式，即

$$\mathbb{E}(y_i|x_i) = \alpha_0 + \beta_0 x_i$$

  那么只要估计得到α和β就得到了条件期望的估计。
- 注意在这里我们并没有对x_i和y_i的联合分布做任何假定，只是假设条件期望$\mathbb{E}(y_i|x_i)$为线性函数形式。

一元线性回归

一元线性回归

- 如果我们令$u_i = y_i - \mathbb{E}(y_i|x_i) = y_i - \alpha_0 - \beta_0 x_i$那么根据条件期望的性质，有：

$$\mathbb{E}(u_i|x_i) = \mathbb{E}(y_i - \alpha_0 - \beta_0 x_i|x_i) = \mathbb{E}(y_i|x_i) - (\alpha_0 + \beta_0 x_i) = 0$$

从而真值α_0, β_0满足：

$$\begin{cases} \mathbb{E}(y_i - \alpha_0 - \beta_0 x_i) = \mathbb{E}(u_i) = \mathbb{E}[\mathbb{E}(u_i|x_i)] = 0 \\ \mathbb{E}[x_i(y_i - \alpha_0 - \beta_0 x_i)] = \mathbb{E}(x_i u_i) = \mathbb{E}[\mathbb{E}(x_i u_i|x_i)] = \mathbb{E}[x_i \mathbb{E}(u_i|x_i)] = 0 \end{cases}$$

一元线性回归

一元线性回归

- 因而(α_0, β_0)是矩条件方程：

$$\begin{cases} \mathbb{E}\left(y_i - \alpha - \beta x_i\right) = 0 \\ \mathbb{E}\left[x_i\left(y_i - \alpha - \beta x_i\right)\right] = 0 \end{cases}$$

的解。

一元线性回归

一元线性回归

- 在使用以上矩条件之前还需要讨论识别问题，也就是虽然(α_0, β_0)是以上矩条件方程的解，但是我们还要求该解时唯一解。
- 使用上式解得

$$\begin{cases} \alpha_0 = \mathbb{E}\left(y_i\right) - \beta_0 \mathbb{E}\left(x_i\right) \\ \beta_0 = \frac{\mathbb{E}(x_i y_i) - \mathbb{E}(x_i)\mathbb{E}(y_i)}{\mathbb{E}\left(x_i^2\right) - [\mathbb{E}(x_i)]^2} = \frac{\mathbb{C}\mathrm{ov}(x_i, y_i)}{\mathbb{V}(x_i)} \end{cases}$$

  - 如果$\mathbb{V}\left(x_i\right) \neq 0$，即$x_i$不是常数，那么方程组有唯一解；
  - 如果$x_i = c$为一个常数，只要满足$\alpha - c\beta = \mathbb{E}\left(y_i\right)$的所有$(\alpha, \beta)$都是解，因而以上问题时不可识别的。

一元线性回归

一元线性回归

- 假设x_i不是常数，那么我们可以用样本的等价形式

$$\begin{cases} \frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)=0 \\ \frac{1}{N}\sum_{i=1}^{N}\left[x_i\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)\right]=0 \end{cases}$$

解得：

$$\begin{cases} \hat{\alpha}=\bar{y}-\hat{\beta}\bar{x} \\ \hat{\beta}=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \end{cases}$$

使用大数定律可以得到，以上为一致估计量。

矩估计的大样本性质

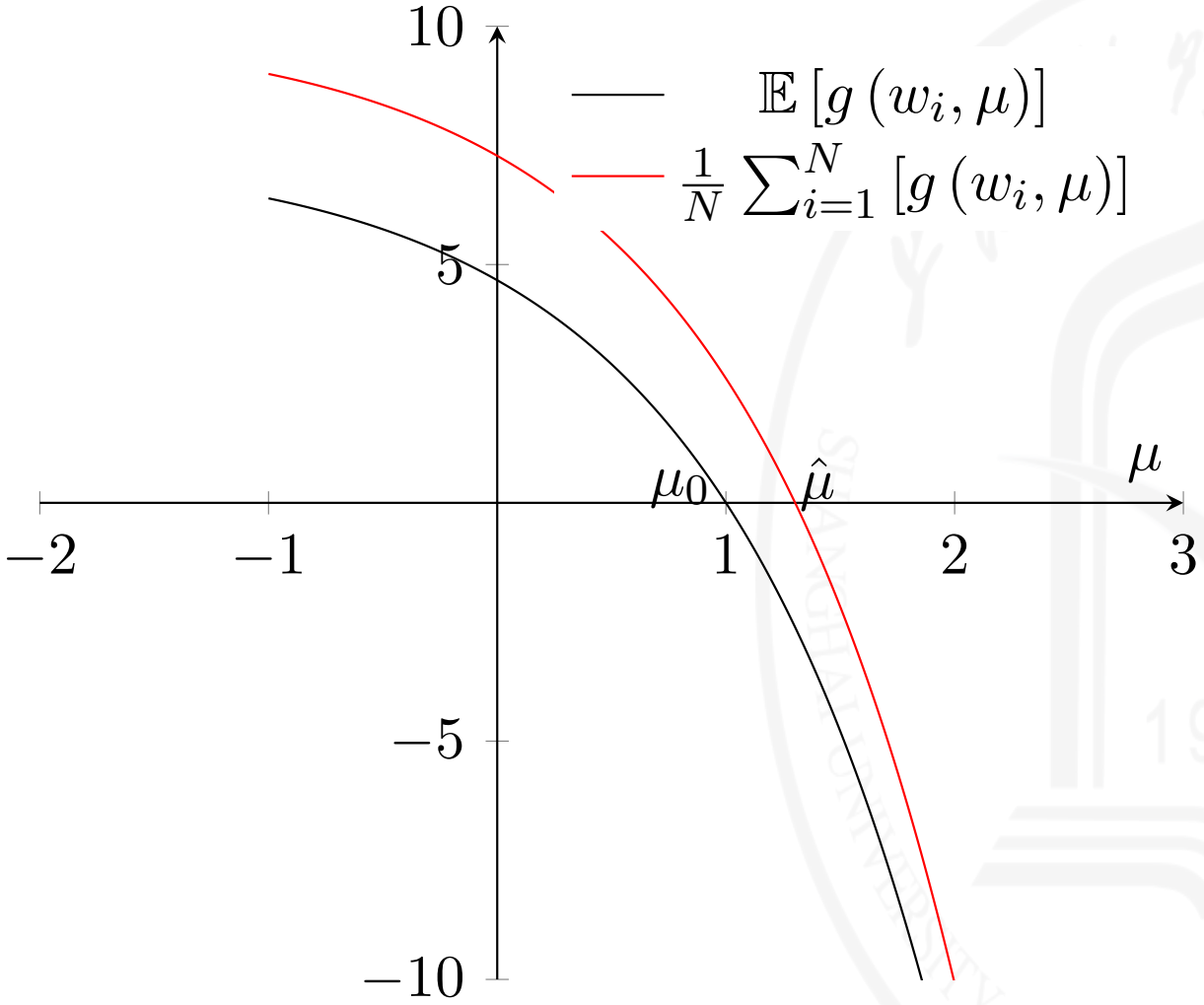
矩估计的一致性

如果$w_i \in \mathbb{R}^p$为一系列独立同分布的随机向量，假设：

- $\theta_0 \in \Theta \subset \mathbb{R}^K$，其中$\Theta$为紧集；
- （连续性条件）函数$g(w,\theta) \in \mathbb{R}^K$为Borel可测函数，且对于任意的$w$，$g(w,\theta)$在$\Theta$上对$\theta$为连续函数；
- （收敛性条件）存在一个函数$K(w)$，使得对于任意的$\theta \in \Theta$，$|g_j(w,\theta)| \leq K(w)$，其中$\mathbb{E}[K(w)] < \infty$；
- （识别条件）θ_0为方程：$\mathbb{E}[g(w_i,\theta)] = 0$的唯一解

那么方程：$\frac{1}{N}\sum_{i=1}^{N} g\left(w_i,\hat{\theta}\right) = 0$的解$\hat{\theta} \xrightarrow{p} \theta_0$。

矩估计的大样本性质


10
5
−5
−10
−2
−1
1
2
3
μ
μ_0
$\hat{\mu}$
$\mathbb{E}\left[g\left(w_i,\mu\right)\right]$
$\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\mu\right)\right]$

矩估计的渐近分布

- 在得到未知参数θ的估计量$\hat{\theta}$之后，我们经常还需要知道估计量$\hat{\theta}$的抽样分布，比如其大样本分布，才能够在此基础之上完成区间估计、假设检验等任务。
- 估计量的精确分布一般是非常难以计算的，因而我们经常诉诸于估计量的大样本分布进行近似。

矩估计的渐近分布

- 首先考虑一维情形。我们假设$\theta \in \mathbb{R}$，那么在矩条件：$\mathbb{E}\left[g\left(w_i, \theta_0\right)\right] = 0$成立的条件下，矩估计量即解如下方程：

$$\sum_{i=1}^{N}\left[g\left(w_i, \hat{\theta}\right)\right] = 0$$

- 如果我们假设函数$g\left(\cdot, \cdot\right)$对$\hat{\theta}$是可微的，那么我们可以对其在$\theta_0$处进行泰勒展开：

$$0 = \frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i, \theta_0\right)\right] + \frac{1}{N}\sum_{i=1}^{N}\frac{dg}{d\theta}\left(w_i, \theta_0\right)\left(\hat{\theta} - \theta_0\right) + O\left(\left(\hat{\theta} - \theta_0\right)^2\right)$$

矩估计的渐近分布

- 两边乘以$\sqrt{N}$，得到：

$$-\sqrt{N}\frac{\sum_{i=1}^{N}\left[g\left(w_i,\theta_0\right)\right]}{N}=\left[\frac{\sum_{i=1}^{N}\frac{dg}{d\theta}\left(w_i,\theta_0\right)}{N}\right]\left[\sqrt{N}\left(\hat{\theta}-\theta_0\right)\right]+O\left(\sqrt{N}\left(\hat{\theta}-\theta_0\right)^2\right)$$

- 其中，根据一致性，有$\hat{\theta}-\theta_0=o_p\left(1\right)$，而$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=O_p\left(1\right)$，从而$\sqrt{N}\left(\hat{\theta}-\theta_0\right)^2=O_p\left(1\right)\cdot o_p\left(1\right)=o_p\left(1\right)$，因而最后一项可以忽略

矩估计的渐近分布

- 而根据中心极限定理，由于$\mathbb{E}\left[g\left(w_i,\theta_0\right)\right]=0$，因而有

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}\left[g\left(w_i,\theta_0\right)\right]\overset{a}{\sim}N\left(0,\mathbb{V}\left(g\left(w_i,\theta_0\right)\right)\right)$$

其中

$$\mathbb{V}\left(g\left(w_i,\theta_0\right)\right)=\mathbb{E}\left(g\left(w_i,\theta_0\right)^2\right)-\left[\mathbb{E}\left(g\left(w_i,\theta_0\right)\right)\right]^2=\mathbb{E}\left(g\left(w_i,\theta_0\right)^2\right)\triangleq B$$

此外，根据大数定律：

$$\frac{1}{N}\sum_{i=1}^{N}\frac{dg}{d\theta}\left(w_i,\theta_0\right)\xrightarrow{p}\mathbb{E}\left(\frac{dg}{d\theta}\left(w_i,\theta_0\right)\right)\triangleq A$$

矩估计的渐近分布

- 从而：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \overset{a}{\sim} N\left(0, \frac{B}{A^2}\right)$$

- 注意到A和B都依赖于未知参数θ_0，因而如果需要计算$\hat{\theta}$的渐近方差，可以将$\hat{\theta}$带入到A和B的表达式中进行计算，由于$\hat{\theta}$是θ_0的一致估计量，所以大样本条件下对渐近方差的估计仍然是准确的。

矩估计的渐近分布

对数正态分布的矩估计的渐近分布

在对数正态分布的例子中，矩条件为：

$$\mathbb{E}\left[g\left(x_i,\mu\right)\right]=\mathbb{E}\left(x_i-e^{\mu+1}\right)=0$$

从而其中：

$$\begin{cases} A=\mathbb{E}\left(\frac{dg}{d\mu}\left(x_i,\mu_0\right)\right)=-e^{\mu_0+1} \\ B=\mathbb{E}\left(g\left(x_i,\mu_0\right)^2\right)=\mathbb{E}\left[\left(x_i-e^{\mu_0+1}\right)^2\right] \\ =\mathbb{V}\left(x_i\right)=e^{2(\mu_0+2)}-e^{2\mu_0+2} \end{cases}$$

因而

$$\frac{B}{A^2}=\frac{\mathbb{V}\left(x_i\right)}{e^{2\mu_0+2}}$$

矩估计的渐近分布

对数正态分布的矩估计的渐近分布

根据以上结论，有：

$$\sqrt{N}\left(\hat{\mu}-\mu_0\right) \overset{a}{\sim} N\left(0, \frac{\mathbb{V}\left(x_i\right)}{e^{2\mu_0+2}}\right)$$

与之前计算的结论一致。在实际计算中，注意到$\sqrt{N}\left(\hat{\theta}-\theta_0\right)$的渐近方差$\frac{\mathbb{V}(x_i)}{e^{2\mu_0+2}}$为未知参数$\mu_0$的函数，因而为了计算渐近方差，我们可以将$\mu_0$的估计值$\hat{\mu}$带入到渐近方差公式中进行计算。

多元的情形

结论:

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \overset{a}{\sim} N\left(0, G_0^{-1}\mathbb{E}\left[g\left(w_i,\theta_0\right)g\left(w_i,\theta_0\right)'\right]\left(G_0^{-1}\right)'\right)$$

其中:

$$G_0=\mathbb{E}\left[\frac{\partial}{\partial\theta}g\left(w_i,\theta_0\right)\right]$$

Stata中的矩估计

- 在一些简单的例子中，我们可以推导出矩估计的封闭接（closed-form solution）
- 而在另外一些更加复杂的例子中，封闭接是无法通过推导得到的，此时我们需要一些数值方法对其进行求解
  - 比如，在Beta分布的例子中，$\hat{\alpha}, \hat{\beta}$就无法通过解非线性方程将其表示为$\bar{x}, \overline{x^2}$的函数
  - 为了获得$\hat{\alpha}, \hat{\beta}$，必须通过数值方法。

广义矩估计

- 在Stata中并没有简单的矩估计的命令使用，不过我们可以使用其推广形式，即**广义矩估计**（**generalized method of moments，GMM**）的命令来完成矩估计。
- 实际上解方程组：

$$\frac{1}{N}\sum_{i=1}^{N} g\left(w_i, \hat{\theta}\right) = 0$$

的问题可以转化为一个优化问题：

$$\min_{\theta} \sum_{k=1}^{K}\left[\frac{1}{N}\sum_{i=1}^{N} g_k\left(w_i, \hat{\theta}\right)\right]^2$$

当且仅当以上目标函数等于0时，方程组成立，又由于$\left[\frac{1}{N}\sum_{i=1}^{N} g_k\left(w_i, \hat{\theta}\right)\right]^2 \geq 0$，从而当以上目标函数达到最小时，就可以得到方程组的解了。

Stata中的矩估计

基于这一原理，在Stata中，我们只需要使用gmm命令，将每个$g_k\left(w_i,\hat{\theta}\right)$表达出来，Stata会自动帮助我们进行以上的最优化问题求解，从而得到估计值，并计算标准误等。

Stata中的矩估计

Beta分布的矩估计

Beta分布的矩估计问题，我们可以使用如下代码进行估计：

```
clear
set obs 100
gen x=rbeta(1.2,0.5)
gen x2=x^2
gmm (x-({alpha=1}/({alpha}+{beta=1}))) (x2-({alpha
    }*{beta}+{alpha}^2*({alpha}+{beta}+1))/(({alpha
    }+{beta})^2*({alpha}+{beta}+1))), winit(
    identity)
```

Stata中的矩估计

Beta分布的矩估计

- 其中gmm命令后面的两个括号分别代表了两个矩条件：

$$\begin{cases} \mathbb{E}\left(x_i - \frac{\alpha}{\alpha+\beta}\right) = 0 \\ \mathbb{E}\left(x_i^2 - \frac{\alpha\beta+\alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)}\right) = 0 \end{cases}$$

- 而winit(identity)用于控制gmm的权重矩阵，在目前的矩估计中无需关注，按照如此设置即可；
- 第一次使用“{alpha}”、“{beta}”时大括号中的“=1”的作用是设置一个初始值。

Stata中的矩估计

Beta分布的矩估计

- 我们可以使用模拟的方法验证以上矩估计：
  - MM_beta.do
- 以上代码将矩估计封装在一个program中，重复了1000次，每次产生100个不同来自于$\alpha = 1.2, \beta = 0.5$的Beta分布随机数，然后使用矩估计的方法对其进行参数估计。
- 得到的结果显示1000次估计的$\hat{\alpha}$的均值为1.23，$\hat{\beta}$的均值为0.51，与真值几乎无异；
- 而1000次估计的$\hat{\alpha}$和$\hat{\beta}$的标准差分别为0.21和0.08，这实际上是对$\hat{\alpha}$和$\hat{\beta}$标准误的估计，可以比较单次使用gmm命令计算的标准误与该数值相差不大。

极大似然估计的思想

- 极大似然估计量（maximum likelihood estimator，MLE）是目前为止最常见的得到估计量的方法
- 其思想是，如果我们要对未知参数总体P_θ做推断，估计θ_0，那么我们就寻找一个$\hat{\theta}$，使得这组数据出现的概率最高，则$\hat{\theta}$理应是θ_0的一个合理估计。

极大似然估计的思想

极大似然估计的思想

在抛硬币的例子中，如果我们观察到在100次抛硬币的独立试验中，有55次正面以及45次反面，我们需要对抛一枚硬币得到正面的结果p进行估计。

- 首先猜测$p = 0.1$，那么得到55次正面以及45次反面这一结果的概率为：$0.1^{55} \times 0.9^{45} = 8.73 \times 10^{-58}$，显然该概率过低，即在$p = 0.1$的情况下，得到该结果的可能性很低。
- 如果猜测$p = 0.5$，那么得到该结果的概率为$0.5^{55} \times 0.5^{45} = 7.89 \times 10^{-31}$，可能性大了很多；
- 当猜测$p = 0.55$的时候，得到该结果的概率为$0.55^{55} \times 0.45^{45} = 1.3 \times 10^{-30}$达到最大值，因而很自然的，我们应该更倾向于认为，这枚硬币得到正面的概率最有可能等于0.55。

极大似然估计

- 如果假设一组独立同分布的样本$x=(x_1,...,x_N)$来自于参数总体P_θ，且密度函数为$f(x_i|\theta)$，那么样本的联合分布函数为

$$f(x|\theta)=\prod_{i=1}^{N} f(x_i|\theta)$$

- 现在，将未知参数θ视为变量，x为给定的样本，由于对数函数为单调函数，因而可以将联合分布函数取对数，得到对数似然函数（log-likelihood function）：

$$L(\theta|x)=\ln f(x|\theta)=\sum_{i=1}^{N}\ln f(x_i|\theta)$$

- 极大似然估计即找到一个$\hat{\theta}$使得对数似然函数最大化

$$\hat{\theta}=\arg\max_{\theta} L(\theta|x)$$

从而我们得到了极大似然估计量$\hat{\theta}$。

极大似然估计

伯努利分布的极大似然估计

如果$x_i \sim Ber\left(p_0\right)\ i.i.d$，那么其联合密度函数为：

$$f\left(x|p\right) = \prod_{i=1}^{N} p^{x_i}\left(1-p\right)^{1-x_i}$$

对数似然函数为：

$$\begin{aligned} L\left(p|x\right) &= \sum_{i=1}^{N}\left[x_i \ln p + \left(1-x_i\right)\ln\left(1-p\right)\right] \\ &= \left(\sum_{i=1}^{N} x_i\right)\ln p + \left(N - \sum_{i=1}^{N} x_i\right)\ln\left(1-p\right) \end{aligned}$$

极大似然估计

伯努利分布的极大似然估计

- 现在欲得到p的极大似然估计值，只要对上述对数似然函数求最大值，即：

$$\frac{\partial L\left(p|x\right)}{\partial p}=\frac{\sum_{i=1}^{N}x_i}{p}-\frac{\left(N-\sum_{i=1}^{N}x_i\right)}{1-p}=0$$

从而得到：$\hat{p}=\frac{1}{N}\sum_{i=1}^{N}x_i$

- 无偏性和一致性讨论略

极大似然估计

正态分布的极大似然估计

如果$x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d$，其中$\theta = \left(\mu, \sigma^2\right)'$，那么其联合密度函数为

$$f\left(x|\theta\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right\}$$

对数似然函数为：

$$\begin{aligned}
L\left(\theta|x\right) &= \sum_{i=1}^{N}\left[-\frac{1}{2}\ln\left(2\pi\right) - \ln\sigma - \frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right] \\
&= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i - \mu\right)^2 \\
&= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i^2 + \mu^2 - 2\mu x_i\right)
\end{aligned}$$

极大似然估计

正态分布的极大似然估计

对其求极大值，得到：

$$\frac{\partial L(\theta|x)}{\partial \theta} = \begin{pmatrix} -\frac{\mu}{\sigma^2} + \frac{\bar{x}}{\sigma^2} \\ -\frac{1}{\sigma} + \frac{\mu^2}{\sigma^3} + \frac{\overline{x^2}}{\sigma^3} - \frac{2\mu\bar{x}}{\sigma^3} \end{pmatrix} = 0$$

解得：

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2 \end{cases}$$

这与矩估计的估计量是一样的。我们之前已经证明了，$\hat{\mu}$是μ_0的无偏、一致估计量，而$\hat{\sigma}^2$是σ_0^2的一致估计量，$\frac{N}{N-1}\hat{\sigma}^2$是$\sigma_0^2$的无偏估计量。

极大似然估计

截尾数据

现在正在进行一项调查，其中一项调查为收入（y_i）调查，其中关于收入的问题为：

- 请问您的收入是多少?
  - 小于1000
  - 大于10000
  - 其他_____（请填写具体数值）

如果假设收入的对数（$x_i^* = \log_{10} y_i$）服从正态分布，即$x_i^* \sim N\left(\mu, \sigma^2\right)$ $i.i.d$，那么我们观察到的数据为：

$$x_i = \begin{cases} 3 & y_i \leq 1000 \\ 4 & y_i \geq 10000 \\ x_i^* & otherwise \end{cases}$$

我们称数据存在截尾（censoring）现象。

极大似然估计

截尾数据

- dgp_mle_censor.ado对以上过程进行了模拟
  - 为了后续程序调用，写成了.ado文件
- 可以观察到很多生成的数据值都是1000或者10000。

极大似然估计

截尾数据

为了估计以上问题，我们可以计算

$$
\begin{aligned}
P\left(x_i=3\right) &= P\left(x_i^* \leq 3\right) \\
&= P\left(\frac{x_i^*-\mu}{\sigma} \leq \frac{3-\mu}{\sigma}\right) = \Phi\left(\frac{3-\mu}{\sigma}\right)
\end{aligned}
$$

同理$P\left(x_i=4\right)=1-\Phi\left(\frac{4-\mu}{\sigma}\right)$。因而$x_i$的密度函数为

$$
f\left(x_i|\theta\right)=\left[\Phi\left(\frac{3-\mu}{\sigma}\right)\right]^{1\{x_i=3\}}\left[1-\Phi\left(\frac{4-\mu}{\sigma}\right)\right]^{1\{x_i=4\}}\left[f\left(x_i\right)\right]^{1\{3<x_i<4\}}
$$

其中$f\left(x_i\right)$为$3<x_i<4$时x_i的密度函数，我们知道$z_i=\frac{x_i-\mu}{\sigma}\sim N\left(0,1\right)$，从而$F\left(x\right)=P\left(\frac{x_i-\mu}{\sigma}\leq\frac{x-\mu}{\sigma}\right)=\Phi\left(\frac{x-\mu}{\sigma}\right)$，从而密度函数：

$$
f\left(x_i\right)=\frac{1}{\sigma}\phi\left(\frac{x_i-\mu}{\sigma}\right)
$$

极大似然估计

截尾数据

因而其对数似然函数为

$$L(\theta|x) = N_3 \ln \Phi\left(\frac{3-\mu}{\sigma}\right) + N_4 \ln\left[1 - \Phi\left(\frac{4-\mu}{\sigma}\right)\right]$$
$$+ \sum_{i=1}^{N} 1\{3 < x_i < 4\} \ln\left[\frac{1}{\sigma}\phi\left(\frac{x_i-\mu}{\sigma}\right)\right]$$

最大化以上对数似然函数，我们就得到了正态分布总体的极大似然估计。

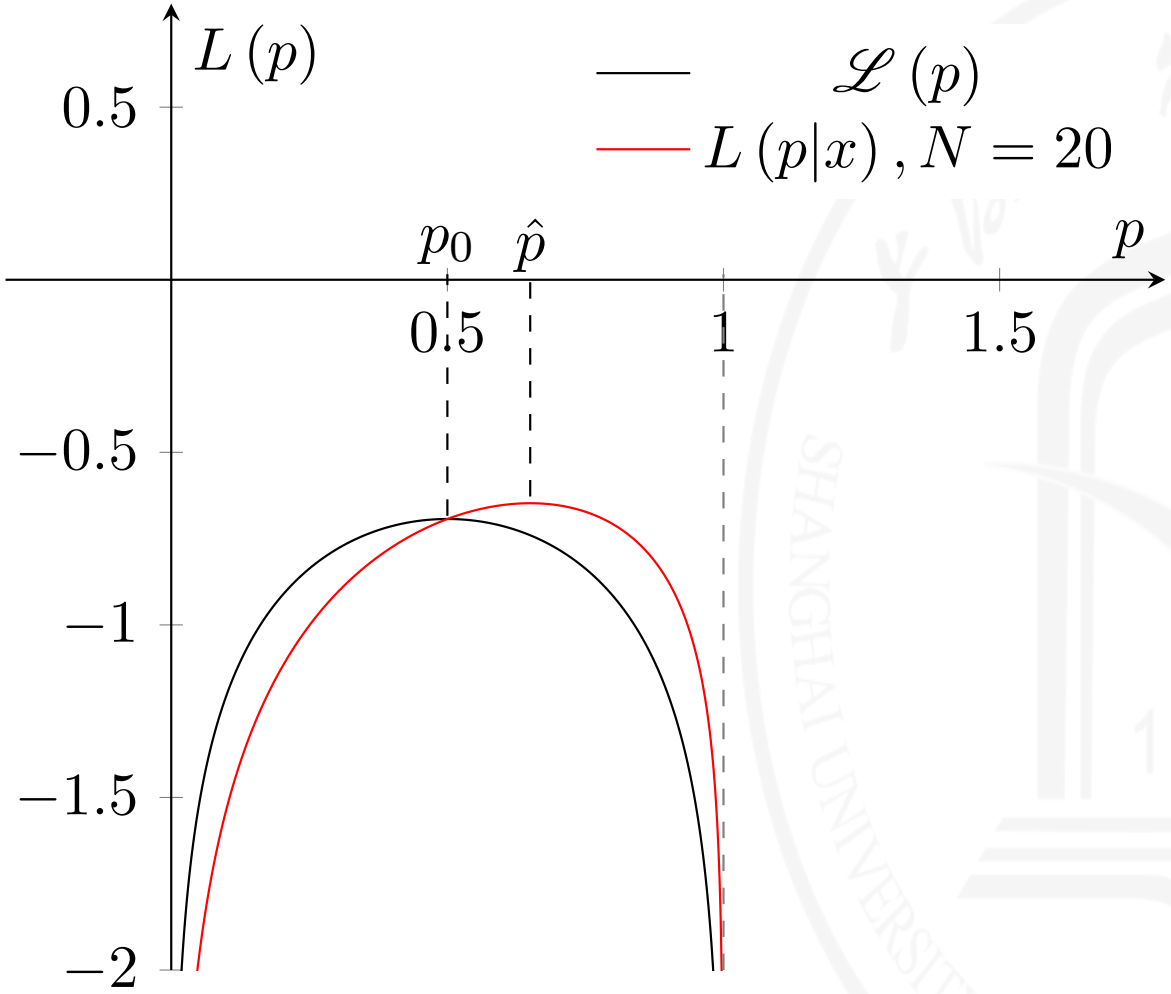
极大似然估计的一致性

- 我们知道，对数似然函数：$\frac{1}{N}L(\theta|x)=\frac{1}{N}\sum_{i=1}^{N}\ln f(x_i|\theta)$，对于任意一个给定的$\theta$（不一定是真值$\theta_0$），在一定的条件下，根据大数定律，有：

$$\frac{1}{N}L(\theta|x)\xrightarrow{p}\mathbb{E}\ln f(x_i|\theta)\triangleq\mathscr{L}(\theta)$$

  即样本似然函数收敛到总体的似然函数。在一定条件下，这个收敛是一致收敛的。
- 由于$\frac{1}{N}L(\theta|x)$一致收敛到$\mathscr{L}(\theta)$，样本似然函数$L(\theta|x)$（红线）随着样本量增大不断逼近总体似然函数$\mathscr{L}(\theta)$（黑线），而极大似然估计的方法是最大化$L(\theta|x)$获得估计，即$\hat{\theta}=\arg\max_\theta L(\theta|x)$那么如果真值$\theta_0=\arg\max_\theta\mathscr{L}(\theta)$那么可以想象，样本似然函数最大值$\hat{\theta}$应该也会收敛到总体似然函数最大值$\theta_0$。

极大似然估计的一致性



极大似然估计的一致性

伯努利分布的总体似然函数

若$x_i \sim Ber\left(p_0\right)\ i.i.d$，那么其联合密度函数为：

$$f\left(x|p\right) = \prod_{i=1}^{N} p^{x_i}\left(1-p\right)^{1-x_i}$$

对数似然函数为：

$$L\left(p|x\right) = \sum_{i=1}^{N}\left[x_i \ln p + \left(1-x_i\right)\ln\left(1-p\right)\right]$$

根据大数定律，对于任意的p，上述似然函数

$$\begin{aligned}\frac{1}{N}L\left(p|x\right) \xrightarrow{p} \mathscr{L}\left(p\right) &\triangleq \mathbb{E}\left[x_i \ln p + \left(1-x_i\right)\ln\left(1-p\right)\right] \\ &= \mathbb{E}\left(x_i\right)\ln p + \mathbb{E}\left(1-x_i\right)\ln\left(1-p\right)\end{aligned}$$

极大似然估计的一致性

伯努利分布的总体似然函数

是不是只有当$p = p_0$时，$\mathscr{L}(p)$达到了最大值呢?

- 为求最大值，我们对$\mathscr{L}(p)$求导数并令其等于0得到

$$\frac{\partial \mathscr{L}(p)}{\partial p} = \frac{p_0}{p} - \frac{1-p_0}{1-p} = 0$$

- 从而只有当$p = p_0$时，以上导数等于0。因而真值p_0最大化了总体似然函数$\mathscr{L}(p)$。

极大似然估计的一致性

以上伯努利分布的例子并不是个例，实际上，我们可以证明，真值θ_0总是可以最大化总体似然函数。

- 为了证明这一点，我们可以从总体似然函数出发：

$$
\begin{aligned}
\mathscr{L}(\theta) &= \mathbb{E} \ln f(x_i|\theta) \\
&= \mathbb{E}_{\theta_0} \ln f(x_i|\theta) \\
&= \int_{\mathbb{R}} \ln f(x|\theta) \cdot f(x|\theta_0)\, dx
\end{aligned}
$$

极大似然估计的一致性

- 求其最大值，得到：

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} \ln f(x|\theta) \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \ln f(x|\theta) \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} \cdot f(x|\theta_0)\, dx
\end{aligned}
$$

当$\theta = \theta_0$时，有：

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta_0)}{\partial \theta} &= \int_{\mathbb{R}} \frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial \theta} \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \frac{\partial f(x|\theta_0)}{\partial \theta} dx \\
&= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(x|\theta_0)\, dx = 0
\end{aligned}
$$

极大似然估计的一致性

- 我们通常称对数似然函数的一阶导数

$$\frac{\partial}{\partial\theta}\ln f\left(x_i|\theta\right)$$

为得分函数（score function），记为$s_i\left(\theta\right)=\frac{\partial}{\partial\theta}\ln f\left(x_i|\theta\right)$。

- 以上结论意味着得分函数的期望等于0，即$\mathbb{E}s_i\left(\theta_0\right)=0$。
- 极大似然估计可以视为矩条件$\mathbb{E}s_i\left(\theta_0\right)=0$下的矩估计，如果得分函数$s_i\left(\theta\right)$满足矩估计一致性的要求，那么必然有极大似然估计量$\hat{\theta}\xrightarrow{p}\theta_0$。

极大似然估计的一致性

- 然而如果使用矩估计的思路证明极大似然的一致性，仍然需要更多的假设。为了建立极大似然方法的一致性，我们仍然需要讨论真值θ_0是否的确最大化了总体似然函数$\mathscr{L}(\theta)$。
- 我们知道，一阶导数等于0是最大化的必要条件而非充分条件，因而以上的推论并不一定能够得到真值θ_0最大化了总体似然函数$\mathscr{L}(\theta)$这一结论。为了更进一步得到真值θ_0是否最大化了总体似然函数，我们引入Kullback-Leiber信息的概念。

Kullback-Leiber信息

Kullback-Leiber信息的定义

令P和Q为同一概率空间中的两个概率函数，p和q分别为其概率密度函数。Kullback-Leiber信息被定义为

$$\mathscr{K}(P,Q)=\int_{\mathbb{R}} p(x)\ln\frac{p(x)}{q(x)}dx$$

当P和Q属于参数族P_θ和Q_η时，Kullback-Leiber信息即

$$\mathscr{K}(\theta,\eta)=\int_{\mathbb{R}} f(x|\theta)\ln\frac{f(x|\theta)}{g(x|\eta)}dx=\mathbb{E}_\theta\left[\ln\frac{f(x|\theta)}{g(x|\eta)}\right]$$

Kullback-Leiber信息

- 实际上，Kullback-Leiber信息度量的是两个概率函数的「距离」，可以证明，Kullback-Leiber信息$\mathscr{K}(P,Q)\geq 0$，当且仅当$P=Q$时等号成立。
- 对于极大似然函数，给定任意一个θ，其代表的概率函数与真值代表的概率函数之间的距离，即Kullback-Leiber信息为：

$$\mathscr{K}(\theta_0,\theta)=\mathbb{E}_{\theta_0}\left[\ln\frac{f(x|\theta_0)}{f(x|\theta)}\right]\geq 0$$

当且仅当$\theta=\theta_0$时等式成立，因而θ_0最小化了：

$$\mathbb{E}_{\theta_0}\left[\ln\frac{f(x|\theta_0)}{f(x|\theta)}\right]=\mathbb{E}_{\theta_0}\left[\ln f(x|\theta_0)-\ln f(x|\theta)\right]$$

或者等价的，最大化了$\mathbb{E}_{\theta_0}\left[\ln f(x|\theta)\right]=\mathscr{L}(\theta)$。

极大似然估计的一致性

一致性

如果$w_i \in \mathbb{R}^p$为一系列独立同分布的随机向量，假设：

- $\theta_0 \in \Theta \subset \mathbb{R}^K$，其中$\Theta$为紧集；
- （连续性条件）函数$\ln f(w_i|\theta) \in \mathbb{R}$为Borel可测函数，且对于任意的$w$，$\ln f(w|\theta)$在$\Theta$上对$\theta$为连续函数；
- （收敛性条件）存在一个函数$K(w)$，使得对于任意的$\theta \in \Theta$，$\left|\ln f(w|\theta)\right| \leq K(w)$，其中$\mathbb{E}[K(w)] < \infty$；
- （识别条件）$f(w_i|\theta_0)$为真实的密度函数，且θ_0为$\mathbb{E}\left[\ln f(w_i|\theta)\right]$的唯一最大值解。

那么估计量：$\hat{\theta} = \arg\max_\theta \sum_{i=1}^N \ln f(w_i|\theta)$满足：$\hat{\theta} \xrightarrow{p} \theta_0$。

极大似然估计的极限分布

- 由于我们计算极大似然估计量时最大化了极大似然函数，其一阶条件为

$$\frac{\partial L\left(\hat{\theta}|x\right)}{\partial\theta}=\frac{\partial}{\partial\theta}\sum_{i=1}^{N}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=0$$

即样本得分函数的均值等于0。

- 我们对上式在$\theta=\theta_0$处进行泰勒展开，得到：

$$0=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\theta_0\right)+\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)+O\left(\left(\hat{\theta}-\theta_0\right)^2\right)$$

- 我们记

$$H_i\left(\theta\right)=\frac{\partial}{\partial\theta'}s_i\left(\theta\right)=\frac{\partial}{\partial\theta\partial\theta'}\ln f\left(x_i|\theta\right)$$

为对数似然函数的海塞矩阵，我们有

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\xrightarrow{p}\mathbb{E}H_i\left(\theta_0\right)\triangleq-H_0$$

极大似然估计的极限分布

- 而由于$\mathbb{E}s_i(\theta_0)=0$，因而

$$\mathbb{V}(s_i(\theta_0))=\mathbb{E}\left[s_i(\theta_0)s_i'(\theta_0)\right]\triangleq\mathcal{I}_0$$

因而根据中心极限定理：

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i(\theta_0)\overset{a}{\sim}N(0,\mathcal{I}_0)$$

- 因而

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=H_0^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i(\theta_0)+o_p(1)$$
$$\overset{a}{\sim}N\left(0,H_0^{-1}\mathcal{I}_0H_0^{-1}\right)$$

极大似然估计的极限分布

注意其中:

$$
\begin{aligned}
-H_0 = \mathbb{E}H_i(\theta_0) &= \int_{\mathbb{R}} \frac{\partial}{\partial\theta\partial\theta'} \ln f(x|\theta_0) \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\theta'} \left[\frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \right] \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \left[\frac{1}{f(x|\theta_0)} \frac{\partial^2 f(x|\theta_0)}{\partial\theta\partial\theta'} - \frac{1}{f^2(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \frac{\partial f(x|\theta_0)}{\partial\theta'} \right] \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 f(x|\theta_0)}{\partial\theta\partial\theta'} - \frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \frac{\partial f(x|\theta_0)}{\partial\theta'} dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 f(x|\theta_0)}{\partial\theta\partial\theta'} dx - \int_{\mathbb{R}} \left(\frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \frac{\partial f(x|\theta_0)}{\partial\theta'} \right)^2 f(x|\theta_0)\, dx \\
&= \frac{\partial^2}{\partial\theta\partial\theta'} \int_{\mathbb{R}} f(x|\theta_0)\, dx - \mathbb{E}_{\theta_0} \frac{\partial \ln f(x|\theta_0)}{\partial\theta} \frac{\partial \ln f(x|\theta_0)}{\partial\theta'} \\
&= -\mathbb{E}_{\theta_0} \left[s_i(\theta_0)\, s_i(\theta_0)' \right] \\
&= -\mathbb{V}\left[s_i(\theta_0) \right] = -\mathcal{I}_0
\end{aligned}
$$

极大似然估计的极限分布

- 因而我们有：$H_0 = \mathcal{I}_0$，即对数似然函数海塞矩阵的期望等于得分函数的方差。
- 将以上等式带入，可以得到：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \overset{a}{\sim} N\left(0, \mathcal{I}_0^{-1}\right)$$

即大样本条件下，极大似然估计量的极限分布为正态分布，且其渐近方差为对数似然函数海塞矩阵倒数的逆矩阵。

- 特别的，当θ为标量时，

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \overset{a}{\sim} N\left(0, -\frac{1}{\mathbb{E}\left(\frac{d^2}{d\theta^2}\ln f\left(x|\theta_0\right)\right)}\right)$$

极大似然估计的极限分布

伯努利分布极大似然估计的极限分布

- 在前例中，我们已经得到伯努利分布的极大似然估计为：$\hat{p}=\bar{x}$根据中心极限定理，$\hat{p}=\bar{x}\xrightarrow{D}N\left(p_0,\frac{p_0(1-p_0)}{N}\right)$。
- 另一方面，直接使用以上结论也可以得到$\hat{p}$的极限分布。注意对数似然函数为：

$$L\left(p|x\right)=\sum_{i=1}^{N}\left[x_i\ln p+\left(1-x_i\right)\ln\left(1-p\right)\right]$$

因而得分函数为：$s_i\left(p\right)=\frac{x_i}{p}-\frac{1-x_i}{1-p}$进而海塞矩阵（二阶导）：

$$-H_i\left(p\right)=-\frac{x_i}{p^2}-\frac{1-x_i}{\left(1-p\right)^2}$$

极大似然估计的极限分布

伯努利分布极大似然估计的极限分布

- 带入真值之后其期望为:

$$\mathcal{I}_0 = H_0 = \mathbb{E}H_i\left(p_0\right) = \frac{1}{p_0} + \frac{1}{1-p_0} = \frac{1}{p_0\left(1-p_0\right)}$$

从而$H_0^{-1} = p_0\left(1-p_0\right)$，从而

$$\sqrt{N}\left(\hat{p}-p\right) \sim N\left(0, p_0\left(1-p_0\right)\right)$$

。

Cramér-Rao下界

- 实际上，可以证明，以上的$\mathcal{I}_0^{-1}$是渐近无偏估计量（asymptotically unbiased）所能达到的最小方差
  - 渐近无偏：随着样本量趋向于正无穷，偏差趋向于0：$\lim_{N\to\infty}\mathbb{E}\left(\hat{\theta}-\theta\right)=0$
- 我们称方差为$\mathcal{I}_0^{-1}$的估计量为渐近有效（asymptotic efficient）估计量。
- 显然极大似然估计是一个渐近有效估计量。

Cramér-Rao下界

- 为了说明以上结果，现在假设存在一个无偏估计量$\tilde{\theta}(x)$，即$\mathbb{E}\left(\tilde{\theta}(x)\right)=\theta_0$，我们有：

$$
\begin{aligned}
\int_{\mathbb{R}} \tilde{\theta}(x)\, s(\theta_0)\, f(x|\theta_0)\, dx &= \int_{\mathbb{R}} \tilde{\theta}(x) \frac{\partial \ln f(x|\theta_0)}{\partial \theta} f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \tilde{\theta}(x) \frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial \theta} \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \tilde{\theta}(x) \frac{\partial f(x|\theta_0)}{\partial \theta} dx \\
&= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} \tilde{\theta}(x)\, f(x|\theta_0)\, dx \\
&= \frac{\partial}{\partial \theta} \mathbb{E}\left(\tilde{\theta}(x)\right) \\
&= \frac{\partial}{\partial \theta} \theta_0 = 1
\end{aligned}
$$

Cramér-Rao下界

- 且由于$\int_{\mathbb{R}} s(\theta_0)\cdot f(x|\theta_0)\,dx = 0$从而

$$\int_{\mathbb{R}} \tilde{\theta}(x)\, s(\theta_0)\cdot f(x|\theta_0)\,dx - \theta_0 \int_{\mathbb{R}} s(\theta_0)\cdot f(x|\theta_0)\,dx$$
$$= \int_{\mathbb{R}} \left[\tilde{\theta}(x) - \theta_0\right] s(\theta_0)\cdot f(x|\theta_0)\,dx = 1$$

- 根据Cauthy-Schwarz不等式，有

$$\begin{aligned}
1 &= \left[\int_{\mathbb{R}} \left[\tilde{\theta}(x) - \theta_0\right] s(\theta_0)\cdot f(x|\theta_0)\,dx\right]^2 \\
&\le \left[\int_{\mathbb{R}} \left[\tilde{\theta}(x) - \theta_0\right]^2 f(x|\theta_0)\,dx\right]\left[\int_{\mathbb{R}} s(\theta_0)\, s(\theta_0)'\cdot f(x|\theta_0)\,dx\right] \\
&= \mathbb{V}\left(\tilde{\theta}(x)\right)\mathbb{E}\left(s(\theta_0)\, s(\theta_0)'\right) \\
&= \mathbb{V}\left(\tilde{\theta}(x)\right)\cdot \mathcal{I}_0
\end{aligned}$$

- 最终我们得到了Cramér-Rao下界：

$$\mathbb{V}\left(\tilde{\theta}(x)\right) \ge \mathcal{I}_0^{-1}$$

费雪信息矩阵

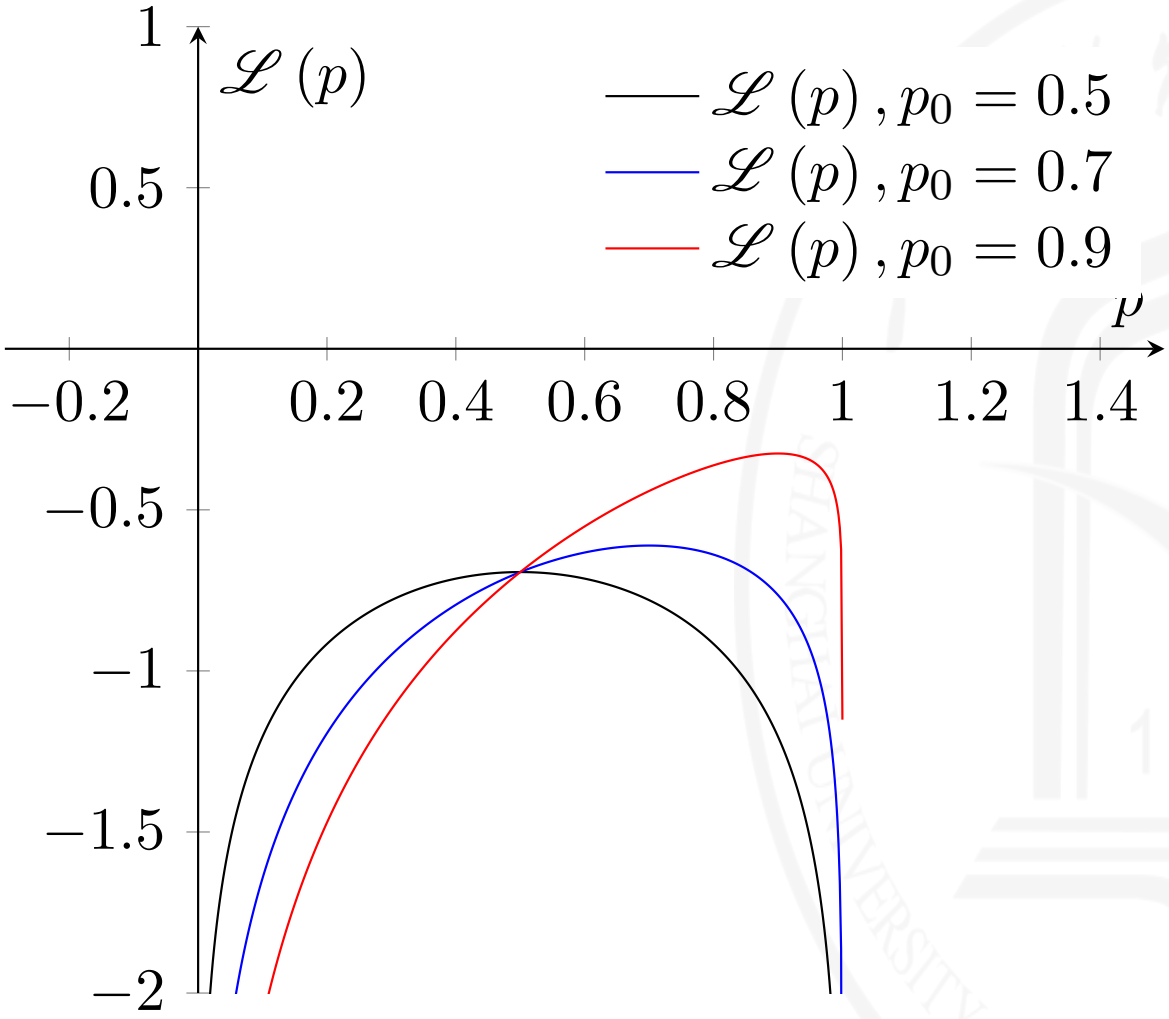
- 基于此，$\mathcal{I}_0$实际上度量了数据中所包含的「信息量」的大小，因而我们称$\mathcal{I}_0$为费雪信息矩阵（Fisher information matrix）。
  - $\mathcal{I}_0$越大，意味着所包含的信息越多，极大似然估计所得到的方差也越小，（渐近）无偏估计量所能达到的最小方差也越小。

费雪信息矩阵

伯努利分布的信息矩阵

- 如图所示，当$p_0 = 0.5$时，信息矩阵$\mathcal{I}_0$达到了最小值，而$\hat{p}$的方差达到了最大值，表现在图中即对数似然函数在真值p_0处非常平缓。
- 当真值p_0逐渐接近0或者1时，信息矩阵$\mathcal{I}_0$逐渐变大，$\hat{p}$的方差也逐渐变小，图中对数似然函数在真值p_0处也更加尖锐。
- 因而同样是伯努利分布，真值越接近于0或者1的伯努利分布实际上携带了更多的信息。

费雪信息矩阵


$\mathscr{L}(p)$
$\mathscr{L}(p), p_0 = 0.5$
$\mathscr{L}(p), p_0 = 0.7$
$\mathscr{L}(p), p_0 = 0.9$
p
1
0.5
−0.5
−1
−1.5
−2
−0.2
0.2
0.4
0.6
0.8
1
1.2
1.4

条件极大似估计

- 以上介绍了极大似然估计法，需要设定数据x的完整的分布情况才能得到估计。
- 然而很多时候，我们观察到一系列数据$w_i \in \mathbb{R}^k, i = 1, ..., N$，其中$w_i = (y_i', x_i')', y_i \in \mathbb{R}^{k_1}, x_i \in \mathbb{R}^{k_2}$，很多时候我们仅仅希望研究$x$和$y$之间的关系，而不关心随机向量$x$自身的分布，如果使用极大似然估计，我们就必须设定$x$的联合分布。
- 然而，设定x的联合分布很多时候是多余的，实际上，如果我们能够找到y给定x的条件分布，即$f(y|x, \theta)$，那么基于条件分布的极大似然估计：

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln f(y_i|x_i, \theta)$$

仍然能够得到参数θ_0的一致估计。

条件极大似估计

条件极大似然估计的应用

如果$x_i|d_i \sim N(\mu, 1+d_i), d_i \sim Ber(p)$，这个模型中有两个未知参数$\mu, p$，如果我们能够观察到$x_i$和$d_i$，且我们只关心参数$\mu$，那么我们仅仅需要最大化条件对数似然函数

$$
\begin{aligned}
L(\mu|x_i, d_i) &= \sum_{i=1}^{N} \ln f(x_i|d_i, \mu) \\
&= \sum_{i=1}^{N} \ln \left[\frac{1}{\sqrt{2\pi}\sqrt{1+d_i}} \exp\left\{ -\frac{(x_i-\mu)^2}{1+d_i} \right\} \right] \\
&= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(1+d_i) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{1+d_i}
\end{aligned}
$$

从而得到$\hat{\mu} = \bar{x}$。

条件极大似估计

条件极大似然估计的应用

如果我们对两个未知参数μ, p都感兴趣，那么：

$$f\left(x_i, d_i|\mu, p\right) = f\left(x_i|d_i, \mu, p\right) f\left(d_i|\mu, p\right)$$

注意其中$f\left(x_i|d_i, \mu, p\right) = f\left(x_i|d_i, \mu\right)$与$p$无关，$f\left(d_i|\mu, p\right) = f\left(d_i|p\right)$与$\mu$无关，从而最大化无条件的对数似然函数：

$$\max_{\mu,p}\left[\sum_{i=1}^{N}\ln f\left(x_i|d_i, \mu\right) + \sum_{i=1}^{N}\ln f\left(d_i|p\right)\right]$$

与直接最大化：

$$\max_{\mu}\sum_{i=1}^{N}\ln f\left(x_i|d_i, \mu\right)$$

对于估计，两方是等价的

条件极大似估计

线性回归

- 如果$(y_i, x_i')', i = 1, ..., N, x_i \in \mathbb{R}^K$为一系列独立同分布的随机向量。为了使用$x_i$预测或者拟合$y_i$，我们可以假设$y_i|x_i \sim N\left(x_i'\beta_0, \sigma^2\right)$，即给定$x$，$y$服从正态分布。
- 或者等价的，以上模型也可以写成：

$$y_i = x_i'\beta_0 + u_i$$

其中$u_i|x_i \sim N\left(0, \sigma^2\right)$。

条件极大似估计

线性回归

- 在上述条件下，条件密度函数为：

$$f\left(y_i|x_i,\beta\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma^2}\right\}$$

因而似然函数为：

$$L\left(\beta|y,x\right)=-\frac{N}{2}\ln\left(2\pi\right)-N\ln\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-x_i'\beta\right)^2$$

如果对β求导，可以得到：

$$-\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i-x_i'\hat{\beta}\right)x_i=0$$

条件极大似估计

线性回归

- 解得：

$$\hat{\beta} = \left[\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1} \sum_{i=1}^{N}\left(x_i y_i\right) = \left(X'X\right)^{-1} X'Y$$

  我们再次得到了最小二乘估计（Ordinary least squares, OLS）。

- 注意在以上设定中，虽然我们得到了与矩估计相同的结果，但是极大似然估计的假设是更强的，因为矩估计中我们只使用了假设$\mathbb{E}\left(y_i|x_i\right) = x_i'\beta_0$，而这里不仅仅假设了条件期望，还假设了条件分布：$y_i|x_i \sim N\left(x_i'\beta_0, \sigma^2\right)$，因而假设更强。

Stata中的极大似然估计

- 同样，在Stata中，也可以进行极大似然估计。我们只需要向Stata提供极大似然估计的目标函数即可，必要时也可以提供对数似然函数一阶导、二阶导的信息以供Stata能够更好的进行估计计算。
- Stata通过ml命令进行极大似然估计
  - 为了使用该命令，需要首先写一个program以定义对数似然函数，ml命令允许不同的“model”，这其中有的需要提供一阶导数，有的需要提供二阶导数。
  - 此外，还需要指定极大似然估计中需要估计的参数以及数据。

Stata中的极大似然估计

- MLE_censor.do展示了截尾数据例子中的极大似然估计
- 其中：
  - mle_obj_censor为极大似然估计的目标函数
  - 而mle_censor将极大似然估计包装为一个program
  - “ml model”命令定义了极大似然问题的model以及需要的参数和数据，为了方便起见，我们选择“lf”这个model，在这个model中我们需要要计算出$\ln f(w_i|\theta)$并放在变量‘lnfj’中，Stata就可以自动对该变量求和，并求最大化从而找到极大似然估计；
  - “ml search”用于寻找最优化的初始值；
  - “ml maximize”即寻找能使得目标函数的最大化的参数。
- 可以发现最终估计结果与真实值相差无几。

假设检验

- 上一节中，我们讨论了对于未知总体参数θ的参数估计问题，包括点估计和区间估计。
- 很多时候，我们不仅仅需要回答总体参数θ是多少的问题，或者θ在什么区间范围以内的问题，还要回答「是不是」的问题，比如$\theta = \theta_0$是不是成立？
  - 在上一节中，我们发现在中国城镇住户调查的数据中，男性收入平均比女性年收入多8222元，那么我们是不是可以推断总体男性收入的确比女性高呢？还是仅仅因为抽样的巧合导致了我们计算的男性收入比女性收入高？
- 我们可以使用假设检验（hypothesis testing）的方法回答此类问题。

假设

- 为了讨论假设检验的问题，我们首先介绍“假设（hypothesis）”的概念。
- 在假设检验中，假设指的是关于总体参数θ的一个命题。
- 比如，对于不同总体，我们可能有如下假设：
  - 山东成年男性的平均身高为175cm（$\theta = \theta_0$）
  - 某生产线次品率控制在0.1%范围内（$\theta \leq \theta_0$）
  - 北方人平均身高高于南方人（$\theta_1 \geq \theta_2$）
- 以上的例子都是关于未知总体参数的一些猜想。由于总体参数是未知的，我们只能观察到样本，因而我们不能确切的知道以上命题究竟是否成立，而只能使用样本对以上命题进行推断。
- 需要注意的是，这里的假设（hypothesis）与数学命题中的假设（assumption）是不同的。假设检验中的假设是我们要验证或者推翻的某个命题，而数学命题中的假设则是结论的前提条件。

原假设和备择假设

- 假设检验中有两个互补的假设：原假设（null hypothesis）和备择假设（alternative hypothesis），分别用H_0和H_1来表示。
- 如果参数的范围为Θ，即总体参数$\theta \in \Theta$，而原假设为$\theta \in \Theta_0$，那么备择假设即为原假设的补集，即$\theta \in \Theta_0^c$。
  - 比如，如果$\Theta = \mathbb{R}$，若原假设是$\theta = \theta_0$，那么备择假设就是$\theta \neq \theta_0$；
  - 若原假设是$\theta \geq \theta_0$，那么备择假设就是$\theta < \theta_0$。
- 注意原假设一般包含等号。

原假设和备择假设

- 假设检验的过程就是使用数据作为“证据”试图推翻原假设的过程，这个过程与法官判案的过程类似。
- 在法律中，有所谓的“无罪推定原则（presumption of innocence）”，即对于犯罪嫌疑人，必须预先假设其无罪，原告方有义务提出证据证明其犯罪，而不得强迫嫌疑人自证其罪。
- 使用以上术语，即原假设（H_0）为被告无罪，备择假设（H_1）为被告有罪，假设检验的目的就是使用证据（样本数据）试图推翻原假设（无罪）。
  - 如果现有证据可以推翻原假设，那么我们称为拒绝原假设（rejecting H_0），即可以认为原假设为假；
  - 而如果现有证据不能推翻原假设，即没有充足的证据证明原假设为假，那么我们称不能拒绝原假设（not rejecting H_0）。

原假设和备择假设

- 注意“接受原假设（accepting H_0）”的说法与“不能拒绝原假设”的说法有细微差别，如果不能拒绝原假设，可能是由于我们的证据不够充分，而并非原假设的确就是对的，因而“不能拒绝原假设”的说法更加准确。
- 基于上述原因，我们一般会把想要搜集证据去推翻的结论放在原假设上。

两类错误

- 由于统计方法总会存在误差，因而基于以上两类假设的推断也会存在犯错的可能性。在假设检验中，有两种错误可能会发生：
  - 第I类错误（type-I error）：原假设为真，但是拒绝原假设，即“弃真错误”；
  - 第II类错误（type-II error）：备择假设为真，但是接受原假设，即“取伪错误”。
- 比如:
  - 如果一个被告本来无罪，但是错误地判其有罪，那么就犯了第I类错误；
  - 而如果一个被告的确犯罪，但是却判其无罪，那么就犯了第II类错误。

检验统计量

- 假设检验一般通过设定一个检验统计量$T(x)$，以及一个拒绝域R，当$T(x) \in R$时拒绝原假设。
- 如果记

$$H_0 : \theta \in \Theta_0$$

而

$$H_1 : \theta \in \Theta_0^c$$

那么第I类错误，即原假设为真但是拒绝原假设的概率为

$$P_{\theta \in \Theta_0}(T(x) \in R)$$

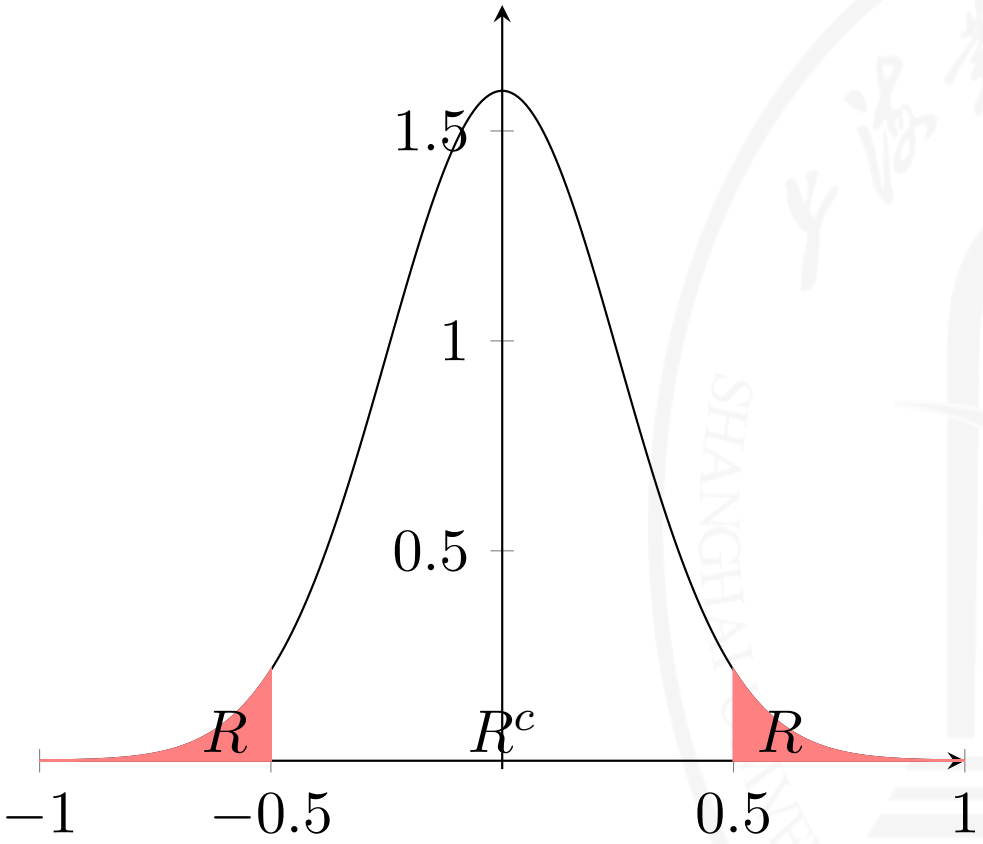
第I类错误

正态总体假设检验

如果样本$x_i \sim N(\mu, 1)\ i.i.d, i = 1, ..., N$，$\mu$为未知总体参数。

- 如果原假设为
$$H_0 : \mu = 0$$
备择假设为
$$H_1 : \mu \neq 0$$
那么$\Theta_0 = \{0\}$。
- 令检验统计量$T(x) = \bar{x}$，如果在原假设条件下，即假设$\mu = 0$，那么
$$\bar{x} \sim N\left(0, \frac{1}{N}\right)$$
即如果原假设成立，那么样本均值应该分布在0附近。而如果我们看到了样本均值距离0比较远，那么说明原假设有可能是不成立的。

第I类错误



第I类错误

正态总体假设检验

- 如果取拒绝域为$R=(-\infty,-0.5)\cup(0.5,\infty)$，那么：

$$
\begin{aligned}
P_{\theta=0}(\bar{x}\in R) &= P_{\theta=0}(|\bar{x}|>0.5) \\
&= P_{\theta=0}\left(\left|\frac{\bar{x}}{\sqrt{\frac{1}{N}}}\right|>\frac{0.5}{\sqrt{\frac{1}{N}}}\right) \\
&= 2\left(1-\Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right)\right)
\end{aligned}
$$

- 如果令$N=16$，那么

$$P_{\theta=0}(\bar{x}\in R)=2(1-\Phi(2))\approx 4.56\%$$

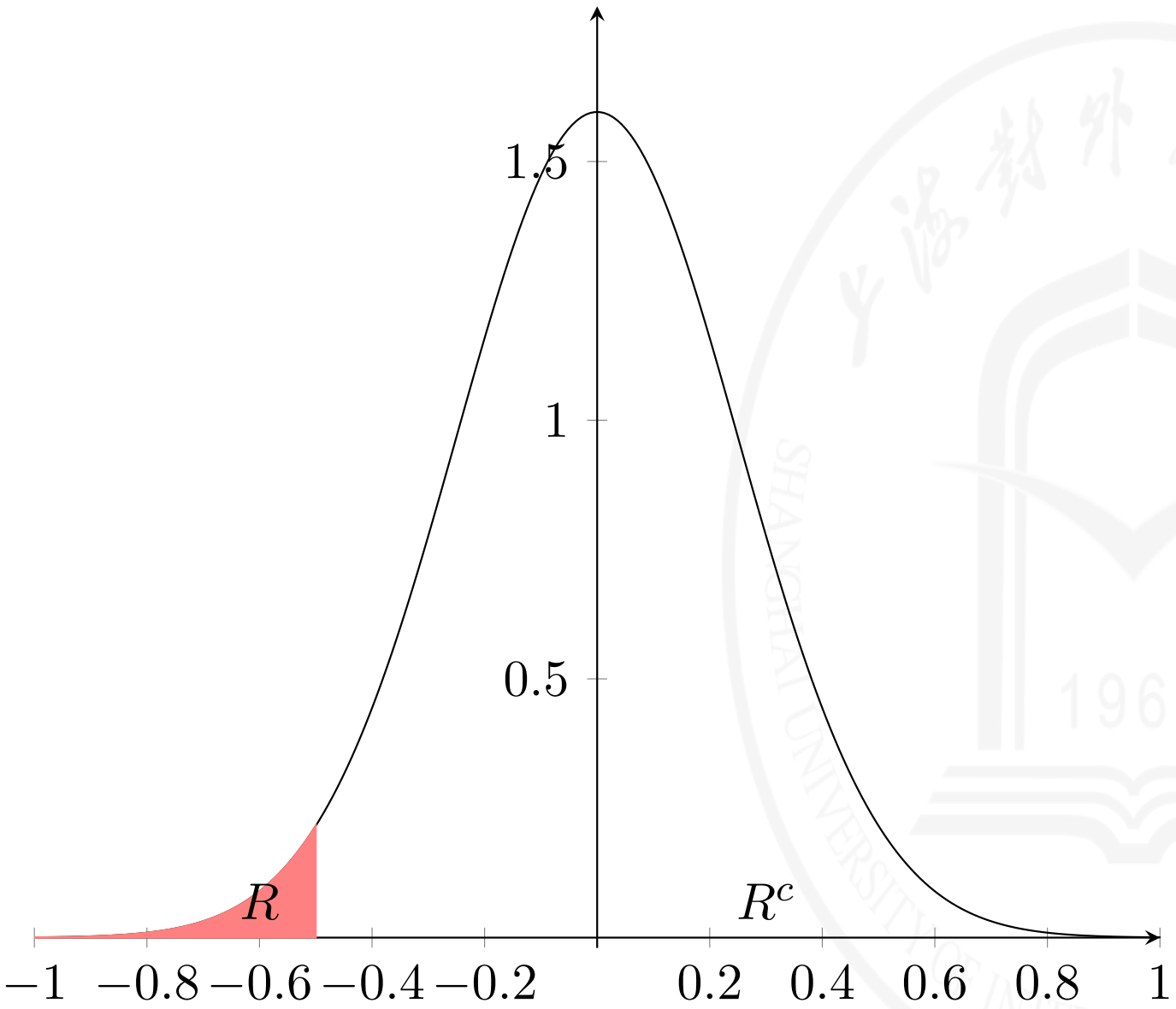
  - 以上概率是我们在原假设的假设上进行计算的，这意味着，如果原假设成立，那么得到均值落在拒绝域，
    即$\bar{x}\in R=(-\infty,-0.5)\cup(0.5,\infty)$的概率为4.56%。
  - 在原假设成立的条件下，我们犯错的概率就是4.56%，即犯第I类错

单侧检验和双侧检验

根据原假设的不同，检验还可以分为单侧检验（one-tailed test）和双侧检验（two-tailed test）

- 如果拒绝域同时出现在分布的左侧和右侧，那么该检验称为双侧检验；
- 如果拒绝域仅仅出现在分布的左侧，那么检验称为左侧检验；
- 而相应的，如果拒绝域出现在分布的右侧，则称为右侧检验。
- 左侧检验和右侧检验合称为单侧检验。

单侧检验和双侧检验



单侧检验

正态分布的单侧检验

如果样本$x_i \sim N(\mu, 1)\ i.i.d, i = 1, ..., N$，$\mu$为未知总体参数。

- 如果原假设为

$$H_0 : \mu \leq 0$$

  那么当$\bar{x}$足够大时，可以拒绝原假设。
- 由于拒绝域在分布的右侧，因而该检验为一个右侧检验。
- 在原假设成立的条件下，当$\mu \leq 0$时，如果取拒绝域为$R = (0.5, \infty)$，那么：

$$\begin{aligned} P_\mu(\bar{x} \in R) &= P_\mu(\bar{x} > 0.5) \\ &= P_\mu\left(\frac{\bar{x} - \mu}{\sqrt{\frac{1}{N}}} > \frac{0.5 - \mu}{\sqrt{\frac{1}{N}}}\right) \\ &= 1 - \Phi\left(\frac{0.5 - \mu}{\sqrt{\frac{1}{N}}}\right) \end{aligned}$$

单侧检验

正态分布的单侧检验

- 注意以上概率随着μ的增加而增加，由于$\mu \in \Theta_0 = (-\infty, 0]$，因而其上界：

$$\sup_{\theta \in \Theta_0} P_\theta \left(T(x) \in R\right) = 1 - \Phi\left(\frac{0.5 - 0}{\sqrt{\frac{1}{N}}}\right) = 1 - \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right)$$

- 当$N = 16$时，上式为$1 - \Phi(2) \approx 2.28\%$，即在原假设$H_0 : \mu \leq 0$的假设下，错误拒绝原假设的概率上界为2.28%，或者等价的，犯第I类错误的概率上界为2.28%。

显著性水平

- 在假设检验中，我们通常希望控制犯第I类错误的概率在某一个水平以内，即给定一个α，找到一个拒绝域R_α，使得

$$\sup_{\theta\in\Theta_0} P_\theta\left(T\left(x\right)\in R_\alpha\right)\leq\alpha$$

- 如此，我们便保证了使用拒绝域R_α进行假设检验，犯第I类错误的概率不超过α。
- 我们称α为**显著性水平**（level of significance）
  - 如果我们能够保证$\sup_{\theta\in\Theta_0} P_\theta\left(T\left(x\right)\in R_\alpha\right)=\alpha$，我们通常称这个检验的**水准**（size）为$\alpha$。
  - 否则，如果只能保证$\sup_{\theta\in\Theta_0} P_\theta\left(T\left(x\right)\in R_\alpha\right)\leq\alpha$，我们称这个检验的**水平**（level）为$\alpha$。
  - 一般不做明确区分，我们接下来统一成为显著性水平。

显著性水平

- 一般取$\alpha = 0.01, 0.05, 0.1$，而R_α为一个区间，区间的临界点成为临界值（critical value）。
  - 在应用中，经常会使用“*”号代表在不同的显著性水平下假设检验是否显著。一般的惯例是，
  - 如果在10%的显著性水平下显著，则标记一个“*”符号
  - 如果在5%的显著性水平下显著，则标记两个“*”符号，即“**”
  - 如果在1%的显著性水平下显著，则标记“***”符号。
- 更小的α代表我们对犯第I类错误更加不能容忍，因而我们会更大的概率接受原假设，即拒绝域也会越小。

p值

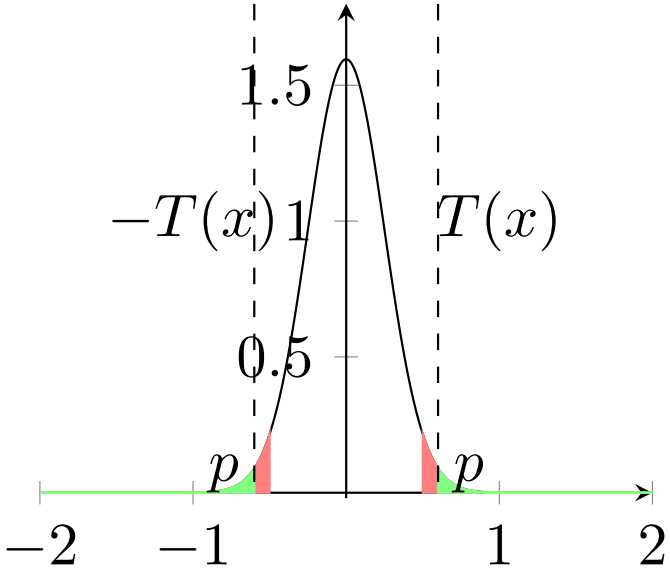
此外，我们还可以定义p值。

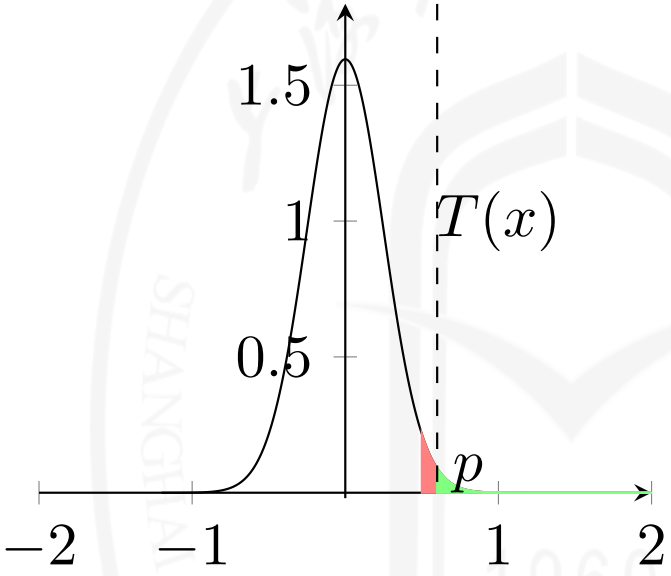
- p值指的是，给定检验统计量$T(x)=t$，在原假设的条件下，能够取到t或者比t更加极端的值的概率。
- 或者等价的，p值可以被定义为**能够拒绝原假设的最小的显著性水平**，即给定$T(x)=t$：

$$p=\inf\{\alpha\in(0,1):t\in R_\alpha\}$$

- 由于检验统计量$T(x)$为随机变量，因而$p(x)=\inf\{\alpha\in(0,1):T(x)\in R_\alpha\}$也是一个随机变量。
- 当$p\leq 0.01,0.05,0.1$时，可以在相应的显著性水平下拒绝原假设。
- 如下图所示，绿色区域面积即p值。
- 需要特别注意的是，对于双侧检验，需要同时计算两侧的绿色面积。

p值





假设检验的步骤

综上，一般的假设检验的步骤可以总结如下：

1. 确定原假设H_0和备择假设H_1；
2. 找到一个检验统计量$T(x)$（通常为基准统计量）；
3. 确定检验统计量$T(x)$<u>在原假设H_0下的</u>分布；
4. 设定显著性水平α，并根据α确定拒绝域R_α，若$T(x) \in R_\alpha$则在α的显著性水平下拒绝原假设；或者根据$T(x)$计算p值，若$p < \alpha$则拒绝原假设。

常见的假设检验

总体均值的检验（小样本正态总体）

如果样本$x_i \sim N\left(\mu, \sigma^2\right)\ i.i.d, i = 1, ..., N$，为了检验

$$H_0 : \mu = \mu_0$$

在H_0的假定下，

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma^2}{N}\right)$$

因而可以构建统计量：

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim t\left(N - 1\right)$$

由于是双侧检验，因而临界值为$t_{\alpha/2}$（$t\left(N-1\right)$的$1 - \alpha/2$分位数），拒绝域为$R_\alpha = \left(-\infty, -t_{\alpha/2}\right) \cup \left(t_{\alpha/2}, \infty\right)$，当$|t| > t_{\alpha/2}$时拒绝原假设，即认为$\mu \neq \mu_0$，否则不能拒绝原假设。

常见的假设检验

总体均值的检验（小样本正态总体）

- 在代码test_under_null.do中，我们在原假设成立时，重复了10000次以上的假设检验，并计算了拒绝原假设的次数
  - 在test_normal_mean程序中，生成了10个$x_i \sim N(0,1)$，并在$H_0: \mu = 0$的原假设以及5%的显著性水平下进行假设检验
  - 如果拒绝原假设则返回rejected=1，否则为0。
  - 我们将以上步骤通过simulate命令重复了10000次
- 最终有5.1%的假设检验拒绝了原假设，与我们5%的显著性水平一致。

常见的假设检验

总体均值的检验（大样本）

- 在上例中，如果不假设样本的正态性，但是在大样本条件下，有

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \overset{a}{\sim} N(0,1)$$

- 因而在原假设：$H_0 : \mu = \mu_0$的条件下，可知当$|z| \geq 1.96$时，则在5%的显著性水平下是显著的。
- 由于$\sqrt{\frac{s^2}{N}}$为$\bar{x}$的标准误的估计，即$s.e.(\bar{x})$，因而经验法则是当$\bar{x} - \mu_0$超过$1.96 \times s.e.$时，即可以在5%的显著性水平下拒绝原假设
- 不同学科中该标准可能存在差异，比如在物理学中，要求超过5个标准误才可以认定为证据，比一般社会科学的要求更加严格。

t检验和z检验

- 由于t分布的极限分布是标准正态分布，因而在大样本条件下，使用t分布去近似标准正态分布是完全可以的
- 由于这一点，软件中一般不会区分小样本或大样本情况，而是全都使用t分布的分为点作为临界值（即都当做t检验来做）。

实践中的总体均值检验

CFPS中收入的检验

在2014年的CFPS数据中，36706个样本每个人的收入均值为9026.73元，标准差为18893.67元，请问若假设每个个体的收入相互独立，在1%、5%的显著性水平下，是否可以认为我国的人均收入超过了8000元?

- 我们选取$H_0: \mu \leq 8000$，那么

$$z = \frac{\bar{x} - 8000}{\sqrt{\frac{s^2}{36706}}} \overset{a}{\sim} N(0,1)$$

- 带入得到：$z = 10.41 > z_\alpha = 2.58$，从而在0.5%的显著性水平下可以拒绝原假设，认为我国的人均收入超过了8000元。

实践中的总体均值检验

CFPS中收入的检验

- 该检验的p值应为$p = 1 - \Phi(10.41)$，即10.41右侧的面积。
- 如果原假设为$H_0 : \mu = 8000$，这是一个双侧检验，那么p值应为$p = 2(1 - \Phi(10.41))$。
- 在Stata中，可以使用“ttest”命令进行该检验：

```
use datasets/cfps_adult.dta, clear
//
drop if p_income<0
//    t
ttest p_income=8000
```

总体比例的假设检验

- 当然，如果$x_i \sim Ber(p)$，中心极限定理也适用，从而对于原假设$H_0: p = (\geq / \leq) p_0$，可以类似的构建检验统计量

$$z = \frac{\bar{x} - p_0}{\sqrt{\frac{s^2}{N}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \overset{a}{\sim} N(0,1)$$

- 以上检验统计量虽然成立，不过一个更加常用的办法是使用原假设中的p_0直接计算方差，即在H_0成立的假设下：

$$z = \frac{\bar{x} - p_0}{\sqrt{\frac{\sigma^2}{N}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} \overset{a}{\sim} N(0,1)$$

- 由于检验统计量的分布是在原假设成立的条件下得到的，所以分母上使用p_0是完全可以的，结论仍然不变。

实践中的总体均值检验

CFPS中劳动参与率的检验

- 在2014年的CFPS数据中，35547个人中有7893个退出劳动力市场。那么是否可以说2014年我国的劳动参与率低于80%？
- 为了回答这一问题，首先可以计算劳动参与率为$\hat{p} = 1 - {}^{7893}/_{35547} = 77.8\%$，原假设为$H_0 : p \geq 80\%$
- 检验统计量为$z = \frac{\hat{p}-0.8}{\sqrt{\frac{0.8(1-0.8)}{35547}}} \overset{a}{\sim} N(0,1)$带入计算，得到：$z = -10.37$
- 由于该检验为左侧检验，$z = -10.37 < -z_{0.05} = -1.65$，从而拒绝原假设，可以认为我国的劳动参与率低于80%。

实践中的总体均值检验

CFPS中劳动参与率的检验

- 在Stata中可以如下计算:

```
use datasets/chfs_ind.dta, clear
//
drop if employ2014 <0
//      3
gen employ_partici=employ2014!=3
//
ttest employ_partici=0.8
```

势函数

- 根据以上的假设检验步骤，我们知道假设检验可以控制犯第I类错误的概率
- 然而对于假设检验而言，还有犯第II类错误的可能性，即当我们无法拒绝原假设的时候，备择假设可能是成立的。为
- 了研究第II类错误的概率，我们引入假设检验的势（power）的概念。

势函数

对于一个检验统计量$T(x)$及其拒绝域R，检验的**势函数**（power function）即给定真实参数θ时，拒绝原假设的概率，即

$$\beta(\theta) = P_\theta(T(x) \in R)$$

势函数

- 注意在以上定义中我们并没有限定$\theta \in \Theta_0$或者$\theta \in \Theta_0^c$，因而理想的情况下，当$\theta \in \Theta_0$时，$\beta(\theta)$应该接近于0，而当$\theta \in \Theta_0^c$时，$\beta(\theta)$应该接近于1。
- 然而，根据我们假设检验的过程，我们设定了在原假设成立，即当$\theta \in \Theta_0$时，拒绝原假设的概率为显著性水平α，因而实际上我们要求当$\theta \in \Theta_0$时，$\beta(\theta)$应该接近或者等于α。
- 而当$\theta \in \Theta_0^c$，即备择假设为真时，没有拒绝原假设的概率，即$1-\beta(\theta)$，即犯第II类错误的概率。

势函数

正态总体均值单侧检验的势函数

若样本$x_i \sim N\left(\mu, \sigma^2\right)\ i.i.d, i = 1, ..., N$，且$\sigma^2$已知，设原假设为$H_0 : \mu \leq \mu_0$，备择假设为$H_1 : \mu > \mu_0$，那么可以构建检验统计量为

$$T\left(x\right) = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{N}}} \sim N\left(0, 1\right)$$

在原假设的条件下，若给定$\alpha = 0.05$，那么$R_{0.05} = \left(z_{0.05}, \infty\right) = \left(1.65, \infty\right)$。

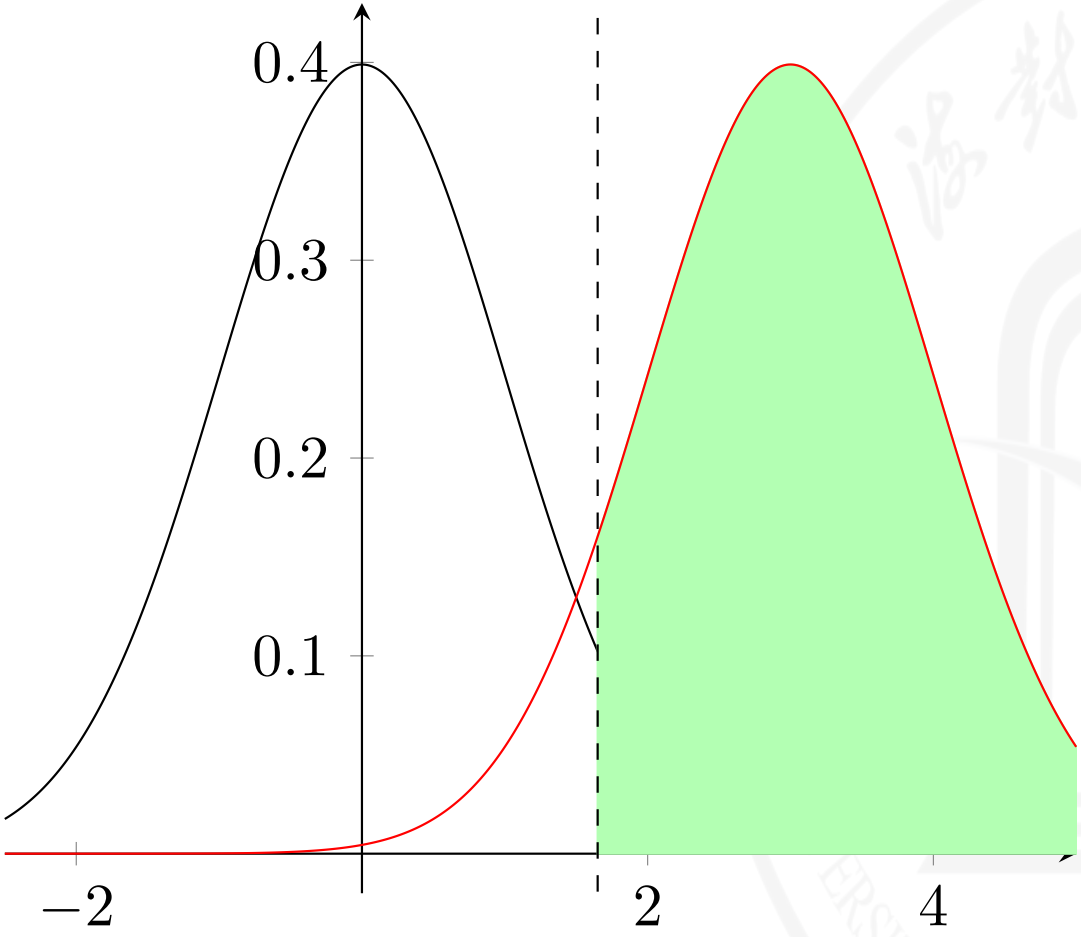
势函数

正态总体均值单侧检验的势函数

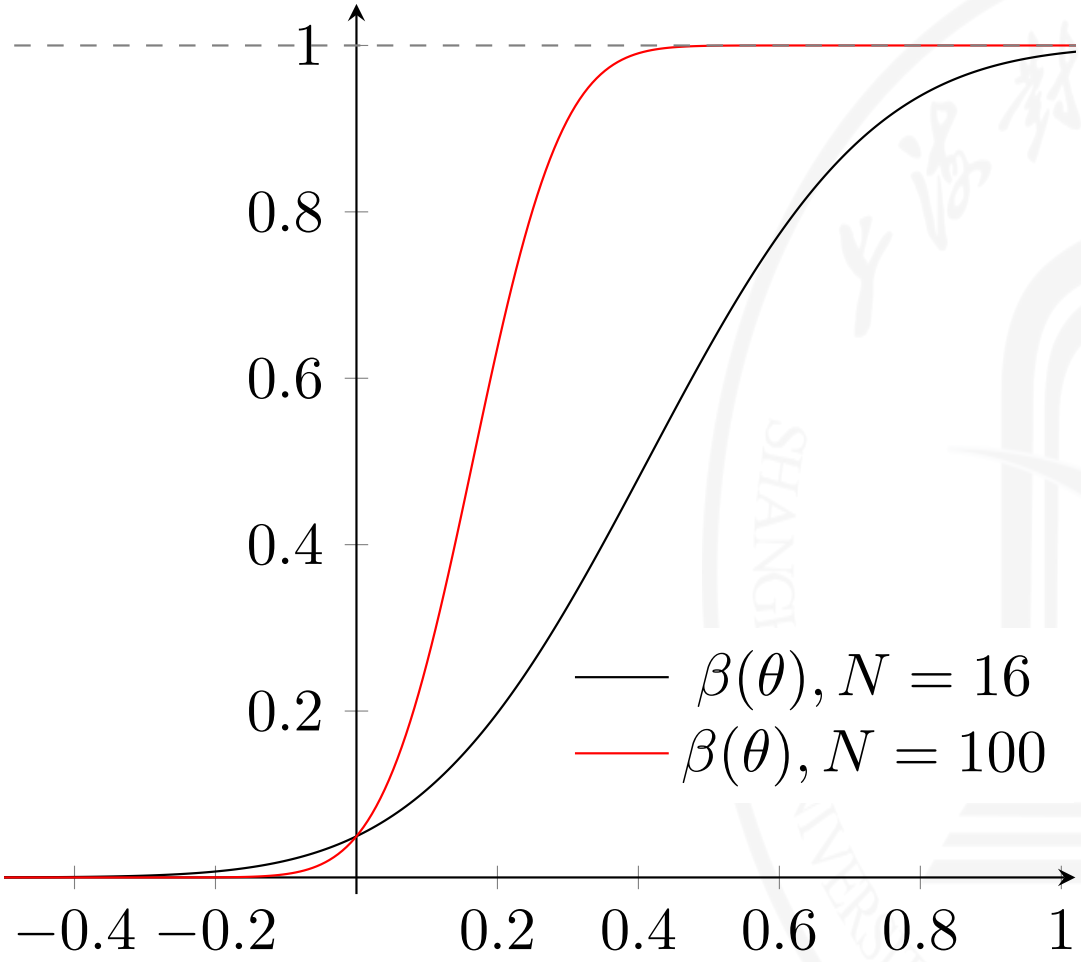
因而，给定μ，势函数为

$$
\begin{aligned}
\beta(\mu) &= P_\mu\left(T(x) \in R_{0.05}\right) \\
&= P_\mu\left(\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma^2}{N}}} > z_{0.05}\right) \\
&= P_\mu\left(\frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{N}}} > z_{0.05} + \frac{\mu_0-\mu}{\sqrt{\frac{\sigma^2}{N}}}\right) \\
&= 1-\Phi\left(z_{0.05} + \frac{\mu_0-\mu}{\sqrt{\frac{\sigma^2}{N}}}\right)
\end{aligned}
$$

势函数



势函数



势函数的模拟

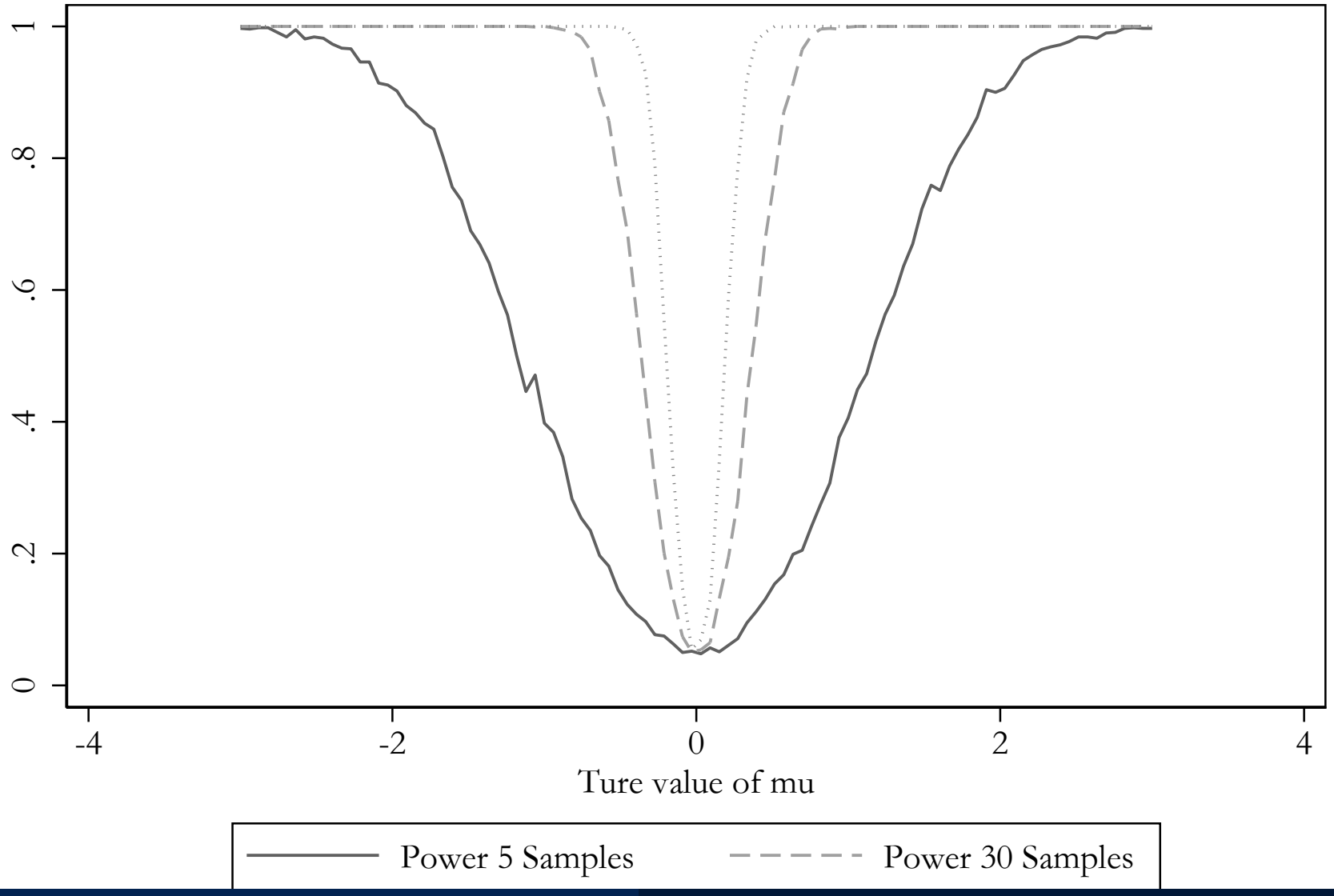
正态总体均值单侧检验的势函数

针对正态总体，检验原假设$H_0: \mu = 0$，检验统计量及其在原假设成立的条件下的抽样分布为：

$$t = \frac{\bar{x}}{\sqrt{\frac{s^2}{N}}} \sim t\left(N-1\right)$$

test_power_function.do对以上检验的势函数进行了模拟，
从$\mu \in [-3, 3]$均匀的选取了100个μ，对于每一个μ，都重复进行一次1000次假设检验的模拟，并计算1000次假设检验中拒绝的比例，并将拒绝的比率与真实的μ画图

势函数


1
.8
.6
.4
.2
0
-4
-2
0
2
4
Ture value of mu
Power 5 Samples
Power 30 Samples

势函数的模拟

正态总体均值单侧检验的势函数

可以发现:

- 当真实的μ与原假设越来越远时，拒绝原假设的概率越高;
- 在原假设成立的条件下，拒绝原假设的概率不变，都是5%;
- 随着样本量的增大，原假设不成立时，拒绝原假设的概率也越高。也就是说，犯第II类错误的概率随着样本量的增大而减小。

第II类错误

- 当我们进行假设检验并且可以拒绝原假设时，此时只有可能犯第I类错误而非第II类错误
  - 而第I类错误的概率不超过α，因而我们可以有比较强的自信认为原假设的确不成立。
- 但是当我们不能拒绝原假设时，我们只有可能犯第II类错误而非第I类错误
  - 然而犯第II类错误的概率依赖于样本量以及真实参数与原假设之间差距的绝对大小，因而一般情况下我们并不知道犯第II类错误的概率
  - 因而如果不能拒绝原假设，一般情况下我们也不能说原假设就是对的，因为很有可能原假设并不成立，仅仅是样本量不足以使得我们收集到足够多的证据拒绝原假设而已。

第II类错误

- 不管样本量N再大，当真值$\theta > \theta_0$，但是差异很小时，第II类错误概率$1-\beta(\theta)$仍然会无限趋向于$1-\alpha$。
- 因而当备择假设为真，但是真值与原假设非常接近时，仍然需要样本量非常大才能够正确的拒绝原假设。

无差异区域与样本量

- 介于此种情况，我们可以引入无差异区域（indifference region）
  - 即虽然备择假设为真，但是θ与θ_0的差异足够的小，我们认为在这个区域里面错误接受原假设也是可以接受的
  - 此时我们可以计算出一个为了能够在一定概率下区分出无差异区域以外的差异所需的样本量。

无差异区域与样本量

势函数与样本量

- 如果我们设计实验研究wifi会不会致癌，可以通过照射组和非照射组的实验对象患癌症的概率差异是否显著大于0，即$H_0: \mu \leq 0$，其中μ为两组之间的患癌症概率的差异。
- 假设wifi的确会致癌，但是致癌的概率充分的接近于0，如果我们想要正确的拒绝原假设，需要非常大的样本量才能保证以一个比较大的概率β拒绝原假设。
- 然而如果我们认为，致癌概率在一定的范围内，比如$[0, \Delta)$内，是可以接受的，那么我们可以据此设计样本量，保证以一个确定的概率拒绝原假设。
- 如果我们希望，当$\mu = \Delta$时，至少以β的概率拒绝原假设，那么

$$\beta = 1 - \Phi\left(z_\alpha + \frac{0 - \Delta}{\sqrt{\frac{\sigma^2}{N}}}\right)$$

从而

$$N = \left[\frac{\sigma}{\Delta}\left(z_\alpha - z_\beta\right)\right]^2$$

无差异区域与样本量

势函数与样本量

- 记实验组（wifi照射）得癌症的概率为p_1，对照组（无wifi照射）得癌症的概率为p_2，我们关注的关键变量为$\mu = p_1 - p_2$，原假设为$H_0 : \mu \leq 0$。
- 如果拒绝原假设，即可以认为wifi照射会致癌。如果假设两组的样本量相同，那么：

$$\mathbb{V}(\hat{\mu}) = \mathbb{V}(\hat{p}_1 - \hat{p}_2) = \mathbb{V}(\hat{p}_1) + \mathbb{V}(\hat{p}_2) = \frac{p_1(1-p_1) + p_2(1-p_2)}{N}$$

- 在原假设下，可以认为$p_1 = p_2$，那么$\mathbb{V}(\hat{\mu}) = \frac{2p_1(1-p_1)}{N}$。
- 如果取$\Delta = 0.0001, \beta = 0.8$，即我们希望当wifi致癌的概率为千分之一时，以80%的概率拒绝原假设，取$\alpha = 0.05$，那么所需样本量约为：

$$N = \left[\frac{2\sqrt{p_1(1-p_1)}}{\Delta}(1.65 - 0.15)\right]^2 = 4\left(\frac{1.65 - 0.15}{0.0001}\right)^2 [p_1(1-p_1)]$$

- 如果自然条件下，癌症发病率为0.2%，即当$p_1 = 0.002$时，上述样本量大概需要$N = 1796400$个。

经济显著性

- 而反过来，如果样本量足够大，那么μ与μ_0的一些非常细微的差别也足以导致统计上的显著性，尽管很多时候这种细微的差别几乎没有现实意义。
- 因而，即使数据中可以得到统计上显著的结果，特别是在样本量非常大的情况下，我们仍然要注意这些结果是不是有“经济显著性（economic significancy）”，即这些差别在现实中是不是足以引起重视。
  - 比如，在上面的例子中，如果真实的概率增加仅仅为0.0001，即万分之一的差异，这一差异在现实中究竟是否值得我们注意就比较值得怀疑了。

一般的原假设

- 在以上两节中我们介绍了假设检验的一般概念和思路。我们知道，如果在原假设H_0的条件下得到检验统计量及其抽样分布，我们就可以使用上节中给出的步骤进行假设检验。
- 尽管上一节中我们讨论了一些初步的假设检验，然而很多时候我们可能希望对不止一个参数进行假设检验，或者对参数的函数进行假设检验。
- 一般的，记我们的原假设为：

$$H_0 : C(\theta) = 0$$

其中$\theta \in \mathbb{R}^k$，$C(\theta) \in \mathbb{R}^r$，且$C(\theta)$为$\theta$的连续可微函数。

单个参数的检验

- 经常在得到了某个参数θ的点估计以后，需要对这个参数进行假设检验。
- 在最简单的情况中，我们仅仅希望检验该参数是否等于（或者大于、小于）某一个具体数值。
  - 比如在回归分析中，我们经常需要使用$\theta_0 = 0$作为原假设进行假设检验。
- 此时的原假设应该为：

$$H_0 : C(\theta) = \theta - \theta_0 = 0$$

与区间估计类似，如果我们可以得到点估计量$\hat{\theta}$的（渐近）分布，那么我们可以通过其渐进分布构建检验统计量，进而对原假设进行检验。

单个参数的检验

泊松分布参数估计的假设检验

- 若$x_i \sim P(\lambda)$，我们希望在5%显著性水平下，对于原假设：$H_0 : \lambda = 10$进行检验。
- 首先得到点估计，其矩估计（极大似然估计）为：$\hat{\lambda} = \bar{x}$。
- 根据中心极限定理，有

$$\sqrt{N}\left(\hat{\lambda} - \lambda\right) \overset{a}{\sim} N(0, \lambda)$$

因而

$$\frac{\sqrt{N}\left(\hat{\lambda} - 10\right)}{\sqrt{10}} \overset{a}{\sim} N(0, 1)$$

- 如果$\left|\frac{\sqrt{N}\left(\hat{\lambda}-10\right)}{\sqrt{10}}\right| > 1.96$，则可以拒绝原假设。

Wald检验

- 如果对于θ的估计，我们已经有

$$\sqrt{N}\left(\hat{\theta}-\theta\right) \overset{a}{\sim} N\left(0,\Sigma\right)$$

，那么使用Delta方法：

$$C\left(\hat{\theta}\right)=C\left(\theta\right)+\ddot{C}\left(\hat{\theta}-\theta\right)+O\left(\left|\hat{\theta}-\theta\right|^{2}\right)$$

其中$\ddot{C}=\frac{\partial C(\theta)}{\partial\theta}$为在真值$\theta$处的$r\times k$的雅克比矩阵，且假设rank$\left(\ddot{C}\right)=r$，那么在原假设$H_0:C\left(\theta\right)=0$的条件下，带入$C\left(\theta\right)=0$，得到：

$$\sqrt{N}C\left(\hat{\theta}\right)=\ddot{C}\sqrt{N}\left(\hat{\theta}-\theta_0\right)+o_p\left(1\right)\overset{a}{\sim} N\left(0,\ddot{C}\Sigma\ddot{C}'\right)$$

Wald检验

- 进而我们可以构建检验统计量：

$$C'\left(\hat{\theta}\right)\left[\frac{\ddot{C}\Sigma\ddot{C}'}{N}\right]^{-1}C\left(\hat{\theta}\right)\overset{a}{\sim}\chi^2\left(r\right)$$

可以使用以上检验统计量对原假设进行假设检验。

Wald检验

- 上式中$\ddot{C}$是在真值处取值的雅克比矩阵，然而真值无法观测，在实际计算时，可以直接使用其一致估计量$\hat{\ddot{C}} = \frac{\partial C(\hat{\theta})}{\partial \theta}$，即在$\hat{\theta}$处计算该雅克比矩阵然后带入即可。
- 注意当原假设成立时，以上检验统计量应该趋向于0，而备择假设成立时，检验统计量应该显著大于0，因而以上检验应为右侧检验，即当

$$C'\left(\hat{\theta}\right)\left[\frac{\ddot{C}\Sigma\ddot{C}'}{N}\right]^{-1}C\left(\hat{\theta}\right) > \chi^2_{0.05}(r)$$

时，在5%的显著性水平下拒绝原假设。

Wald检验

Wald检验示例

- 假设$x_i \sim (\mu, \sigma^2)$，我们需要检验：$H_0 : \mu^2 = 1$，或者：$H_0 : \mu^2 - 1 = 0$。
  - 该原假设相当于同时检验了$\mu = \pm 1$。
- 由于

$$\hat{\mu} = \bar{x} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{N}\right)$$

从而使用Delta方法，有

$$\begin{aligned}\sqrt{N}\left[\left(\hat{\mu}^2 - 1\right) - \left(\mu^2 - 1\right)\right] &= 2\mu\sqrt{N}\left(\hat{\mu} - \mu\right) + o_p(1) \\ &\overset{a}{\sim} N\left(0, 4\mu^2\sigma^2\right)\end{aligned}$$

Wald检验

Wald检验示例

- 将原假设$H_0: \mu^2 - 1 = 0$带入，得到：

$$\left(\frac{4\mu^2\sigma^2}{N}\right)^{-1/2}\left(\hat{\mu}^2 - 1\right) \overset{a}{\sim} N(0,1)$$

从而：

$$\frac{\left(\hat{\mu}^2 - 1\right)^2}{\frac{4\mu^2\sigma^2}{N}} \overset{a}{\sim} \chi^2(1)$$

Wald检验

Wald检验示例

- 由于分母中μ, σ^2未知，我们可以使用其一致估计代替，从而：

$$\frac{N\left(\hat{\mu}^2-1\right)^2}{4\hat{\mu}^2\hat{\sigma}^2} \overset{a}{\sim} \chi^2(1)$$

- 当原假设成立时，以上检验统计量应该趋向于0，因而拒绝域在χ^2分布的右侧，因而该检验应为右侧检验。
- 即，当

$$\frac{N\left(\hat{\mu}^2-1\right)^2}{4\hat{\mu}^2\hat{\sigma}^2} > \chi^2_{0.05}(1)$$

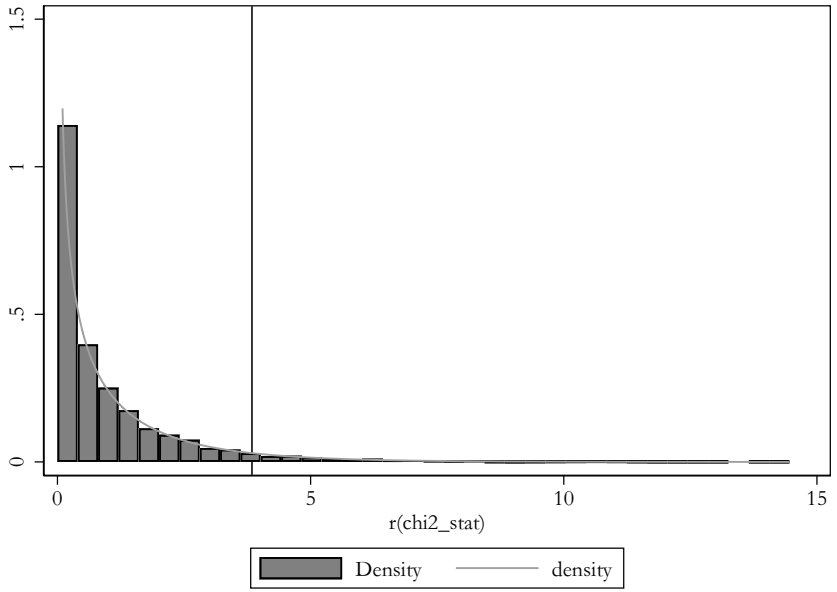
时，拒绝原假设。

Wald检验

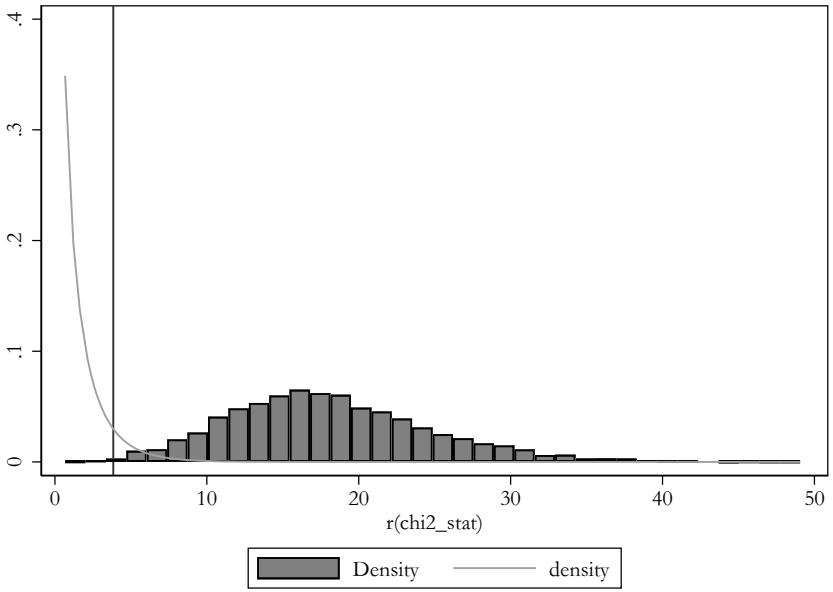
Wald检验示例

- 我们使用simulation_wald_test.do对以上检验在原假设和备择假设下分别进行了模拟，结果如下图所示
- 其中：
  - (a)展示了原假设下模拟的检验统计量的直方图以及$\chi^2(1)$的密度函数
  - 而(b)展示了原假设下模拟的检验统计量的直方图以及$\chi^2(1)$的密度函数。
  - 可以发现，虽然样本量比较少，原假设下检验统计量的分布与$\chi^2(1)$非常接近，而备择假设下两者有相当区别。
  - 原假设下检验统计量大于临界值的模拟次数大约占5.58%，与5%相差较小；而备择假设下大于临界值的模拟大约占99.74%。

Wald检验


1.5
1
.5
0
0
5
10
15
r(chi2_stat)
Density
density

(a)


.4
.3
.2
.1
0
0
10
20
30
40
50
r(chi2_stat)
Density
density

(b)

似然比检验

- 极大似然估计通过最大化对数似然函数获得，即

$$\hat{\theta} = \arg\max_{\theta} L(\theta|x)$$

- 而与此同时，我们可以事先假设原假设成立，在H_0的约束下对θ进行估计，即

$$\tilde{\theta} = \arg\max_{\theta} L(\theta|x)$$
$$s.t. C(\theta) = 0$$

如此我们估计出的$\tilde{\theta}$必然满足原假设。

似然比检验

- 由于$\tilde{\theta}$是在H_0的约束条件下得到的，而$\hat{\theta}$是在无约束的条件下得到的，因而$L\left(\hat{\theta}|x\right) \geq L\left(\tilde{\theta}|x\right)$。
- 如果原假设成立，即$C(\theta)=0$，那么是否对极大似然不及施加原假设并不会对参数估计造成很大影响，从而$\tilde{\theta}$和$\hat{\theta}$应该充分接近，因而$L\left(\hat{\theta}|x\right)$和$L\left(\tilde{\theta}|x\right)$也应该充分接近；
- 但是如果原假设不成立，那么约束就会起作用，从而得到的$\tilde{\theta}$和$\hat{\theta}$应该有相当的差距，因而$L\left(\hat{\theta}|x\right)$和$L\left(\tilde{\theta}|x\right)$也应该有足够大的差距。
- 根据此逻辑，反过来，如果$L\left(\hat{\theta}|x\right) \gg L\left(\tilde{\theta}|x\right)$成立，那么我们可以认为$C(\theta)=0$是不成立的。

似然比检验

- 实际上，我们可以得到两个对数似然函数值在取得最大时，其差距的两倍服从一个卡方分布，即

$$LR = 2\left[L\left(\hat{\theta}|x\right) - L\left(\tilde{\theta}|x\right)\right] \overset{a}{\sim} \chi^2(r)$$

  因而我们可以根据以上结论对$C(\theta) = 0$进行检验。
- 注意原假设成立的条件下，$LR \approx 0$，而备择假设成立的条件下，$LR > 0$，所以以上检验同样是一个右侧检验。
- 以上检验我们通常成为**似然比检验**（likelihood-ratio test）。

似然比检验

泊松分布的似然比检验

如果$x_i \sim P(\lambda)$，$i=1,...,N$，如果原假设为$H_0:\lambda=1$

- 无约束时，$\hat{\lambda}=\bar{x}$，
- 有约束时，$\tilde{\lambda}=1$。
- 其似然函数分别为：

$$L\left(\hat{\lambda}|x\right)=\sum_{i=1}^{N}\left[x_i \ln\left(\hat{\lambda}\right)-\ln\left(x_i!\right)-\hat{\lambda}\right]=\sum_{i=1}^{N}\left[x_i \ln\left(\bar{x}\right)-\ln\left(x_i!\right)-\bar{x}\right]$$

$$L\left(\tilde{\lambda}|x\right)=\sum_{i=1}^{N}\left[x_i \ln\left(\tilde{\lambda}\right)-\ln\left(x_i!\right)-\tilde{\lambda}\right]=\sum_{i=1}^{N}\left[x_i \ln\left(1\right)-\ln\left(x_i!\right)-1\right]$$

似然比检验

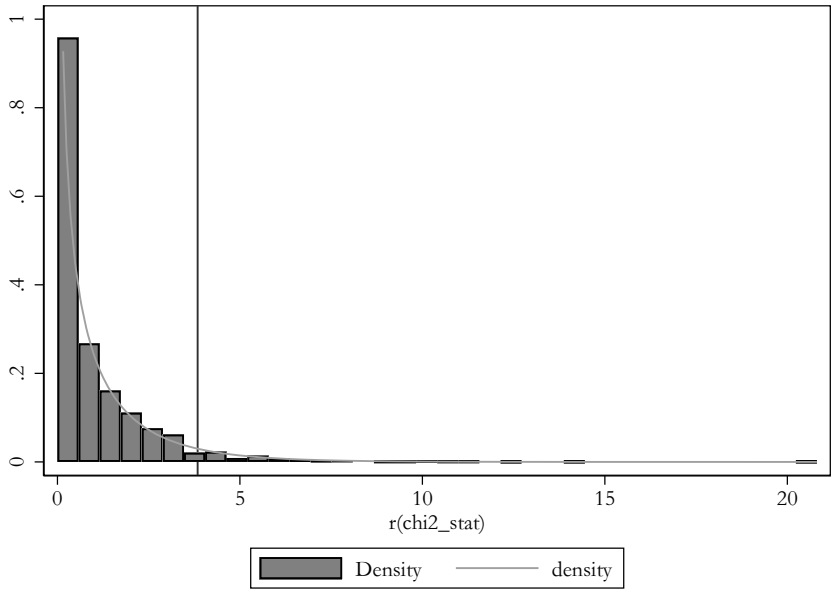
泊松分布的似然比检验

- 因而检验统计量为：

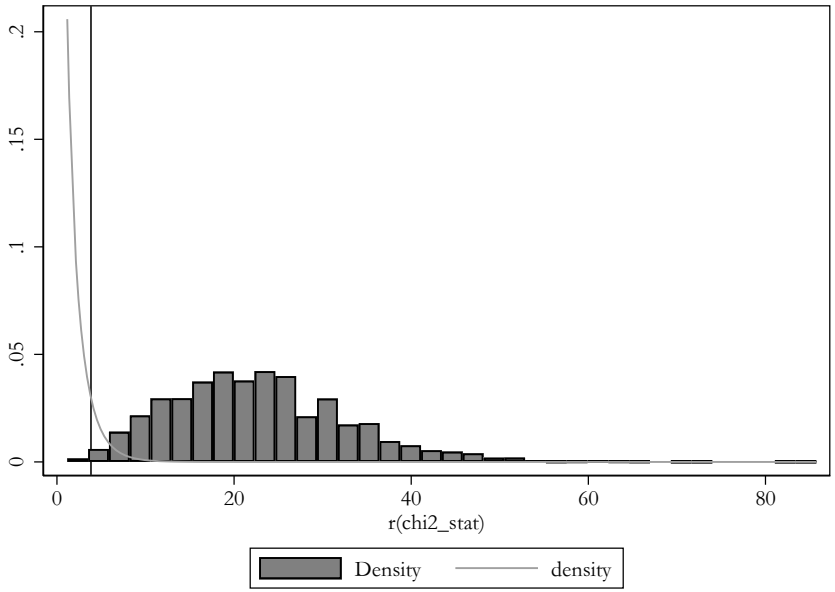
$$
\begin{aligned}
LR &= 2\left[\sum_{i=1}^{N}\left[x_i \ln(\bar{x}) - \ln(x_i!) - \bar{x}\right] - \sum_{i=1}^{N}\left[x_i \ln(1) - \ln(x_i!) - 1\right]\right] \\
&= 2\left[\sum_{i=1}^{N}\left[x_i \ln(\bar{x}) - \bar{x} + 1\right]\right] \\
&= 2N\left[\ln(\bar{x})\,\bar{x} - \bar{x} + 1\right] \\
&\overset{a}{\sim} \chi^2(1)
\end{aligned}
$$

- 模拟：simulation_lr_test.do

似然比检验


0
.2
.4
.6
.8
1
0
5
10
15
20
r(chi2_stat)
Density
density

(a)


0
.05
.1
.15
.2
0
20
40
60
80
r(chi2_stat)
Density
density

(b)

似然比检验

Beta分布的似然比检验

如果$x_i \sim B(\alpha, \beta)$，原假设为：$H_0: \alpha = 1, \beta = 1$此时我们需要同时检验两个假设，我们仍然可以通过似然比检验对其进行检验。

- 如果$\left(\hat{\alpha}, \hat{\beta}\right)$为极大似然估计，那么无约束时的似然函数为：

$$L\left(\hat{\alpha}, \hat{\beta}|x\right) = \sum_{i=1}^{N}\left[-\ln B\left(\hat{\alpha}, \hat{\beta}\right) + (\hat{\alpha}-1)\ln x_i + \left(\hat{\beta}-1\right)\ln(1-x_i)\right]$$

- 而有约束时的似然函数：

$$L\left(\hat{\alpha}, \hat{\beta}|x\right) = L(1,1|x) = -N\ln B(1,1) = 0$$

- 从而似然比检验统计量为：

$$LR = 2\sum_{i=1}^{N}\left[-\ln B\left(\hat{\alpha}, \hat{\beta}\right) + (\hat{\alpha}-1)\ln x + \left(\hat{\beta}-1\right)\ln(1-x)\right] \overset{a}{\sim} \chi^2(2)$$

拉格朗日乘子检验

- 似然比检验构造虽然简单，但是其有一个缺点即需要估计两次极大似然估计：无约束的极大似然估计和有约束的极大似然估计。
- 相比之下，Wald检验只需要估计无约束的估计，且不仅限于极大似然估计，似乎更方便。
- 然而有的时候，估计有约束的问题比估计无约束的问题要简单很多。
  - 比如在以上Beta分布的例子中，有约束的估计甚至不需要计算，约束（原假设）本身已经给出了答案，相比之下无约束的估计会复杂很多。
- 那么，是否可以只计算有约束的估计从而进行检验呢？

拉格朗日乘子检验

- 考虑有约束的极大似然估计问题，通常我们可以使用拉格朗日乘子法求解：

$$\mathscr{L}(\theta,\lambda)=L(\theta|x)-\lambda C(\theta)$$

从而一阶条件为

$$\frac{\partial\mathscr{L}(\theta,\lambda)}{\partial\theta}=\frac{\partial L(\theta|x)}{\partial\theta}-\lambda\frac{\partial C(\theta)}{\partial\theta}=0$$
$$\frac{\partial\mathscr{L}(\theta,\lambda)}{\partial\lambda}=-C(\theta)=0$$

从而解得$\tilde{\theta}$。

拉格朗日乘子检验

- 考虑如果原假设成立，即对于真值θ，约束$C(\theta)=0$成立，那么此时$\tilde{\theta}$应该同时可以最大化$L(\theta|x)$，从而：

$$\frac{\partial L\left(\tilde{\theta}|x\right)}{\partial\theta}=\sum_{i=1}^{N}s_i\left(\tilde{\theta}\right)\approx 0$$

  我们可以通过检验以上一阶条件是否成立来检验原假设是否成立。
- Rao（1948）提出了如上检验，由于以上检验实际上等价于检验得分函数的期望$\mathbb{E}s_i\left(\tilde{\theta}\right)=0$，从而以上检验被称为得分检验（score test）。
- 如果取原假设为$H_0:\theta=\theta_0$，从而$\ddot{C}=1$，带入以上拉格朗日乘子法的一阶条件中，就可以得到$\lambda=\sum_{i=1}^{N}s_i\left(\tilde{\theta}\right)$，因而Aitcheson和Silvey（1958）以及Silvey（1959）将称为拉格朗日乘子检验（Lagrangian multiplier test，LM test）。

拉格朗日乘子检验

- 为了构造渐近分布，注意到在原假设的条件下，$\mathbb{E}\left(s_i\left(\tilde{\theta}\right)\right)=0$，从而其渐近分布为

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}s_i\left(\tilde{\theta}\right)\overset{a}{\sim}N\left(0,\mathcal{I}_0\right)$$

从而可以使用

$$\frac{1}{N}\left[\sum_{i=1}^{N}s_i\left(\tilde{\theta}\right)\right]'\mathcal{I}_0^{-1}\left[\sum_{i=1}^{N}s_i\left(\tilde{\theta}\right)\right]\overset{a}{\sim}\chi^2\left(r\right)$$

进行假设检验。

- 由于$\mathcal{I}_0$不可观测，实际使用时可以将有约束的极大似然估计带入，从而得到其估计$\tilde{\mathcal{I}}$，带入上式即可。

拉格朗日乘子检验

Beta分布的LM检验

考虑使用得分检验对似然比检验中Beta分布检验的问题构造检验统计量。

- 极大似然函数为

$$L(\alpha,\beta|x)=\sum_{i=1}^{N}\left[-\ln B(\alpha,\beta)+(\alpha-1)\ln x_i+(\beta-1)\ln(1-x_i)\right]$$

- 首先计算得分函数：

$$s_i(\alpha,\beta)=\frac{\partial L(\alpha,\beta|x_i)}{\partial\begin{pmatrix}\alpha\\ \beta\end{pmatrix}}=\begin{pmatrix}-\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}+\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)}+\ln x_i\\ -\frac{\Gamma'(\beta)}{\Gamma(\beta)}+\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)}+\ln(1-x_i)\end{pmatrix}$$

拉格朗日乘子检验

Beta分布的LM检验

- 记$\psi_0(z)=\frac{\Gamma'(z)}{\Gamma(z)}=\frac{d}{dz}\ln\Gamma(z)$为多伽马函数（polygamma function），根据其性质：$\psi_0(z+1)=\psi_0(z)+\frac{1}{z}$
- 将原假设$\alpha=\beta=1$（从而有约束的估计$\tilde{\alpha}=\tilde{\beta}=1$）带入，得到

$$s_i\left(\tilde{\alpha},\tilde{\beta}\right)=\begin{pmatrix}1+\ln x_i\\ 1+\ln(1-x_i)\end{pmatrix}$$

继续求海塞矩阵有

$$H_i(\alpha,\beta)=\begin{bmatrix}-\psi_1(\alpha)+\psi_1(\alpha+\beta) & \psi_1(\alpha+\beta)\\ \psi_1(\alpha+\beta) & -\psi_1(\alpha)+\psi_1(\alpha+\beta)\end{bmatrix}$$

其中$\psi_1(z)=\frac{d}{dz}\psi_0(z)=\frac{d^2}{dz^2}\ln\Gamma(z)$

拉格朗日乘子检验

Beta分布的LM检验

- 带入原假设得到

$$\tilde{\mathcal{I}} = \tilde{H} = -\begin{bmatrix} -\psi_1(1) + \psi_1(2) & \psi_1(2) \\ \psi_1(2) & -\psi_1(1) + \psi_1(2) \end{bmatrix}$$

从而检验统计量为

$$\frac{1}{N}\left[\sum_{i=1}^{N}\begin{pmatrix} 1 + \ln x_i \\ 1 + \ln(1 - x_i) \end{pmatrix}\right]' \tilde{\mathcal{I}}^{-1}\left[\sum_{i=1}^{N}\begin{pmatrix} 1 + \ln x_i \\ 1 + \ln(1 - x_i) \end{pmatrix}\right] \overset{a}{\sim} \chi^2(2)$$

拉格朗日乘子检验

Beta分布的LM检验

- 其中ψ_0, ψ_1函数可以由计算机语言计算
  - 比如在Stata中，两个函数分别可以通过digamma(x)以及trigamma(x)两个函数计算。
- 以上检验过程甚至无需进行参数估计，只需要将原假设带入以上公式即可，是非常方便的办法。

拉格朗日乘子检验

- 以上拉格朗日乘子检验不仅可以应用在极大似然估计中，在其他检验中也可以类似应用。
- 比如，对于所有通过最大（小）化目标函数得到估计的估计量中，都可以计算无约束最优化的一阶条件，并将有约束估计量带入，检验一阶条件是否为0。
  - 比如我们前面提到，矩估计可以转化为广义矩估计，从而转化为一个求极值的问题，当然可以采用类似手段。

拉格朗日乘子检验

- 或者，在矩估计中，我们可以将约束下的矩估计量带入计算，在原假设的条件下有

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g\left(w_i,\tilde{\theta}\right) \overset{a}{\sim} N\left(0,\mathbb{V}\left[g\left(w_i,\theta\right)\right]\right)$$

其中

$$\mathbb{V}\left[g\left(w_i,\theta\right)\right]=\mathbb{E}\left[g\left(w_i,\theta\right)g\left(w_i,\theta\right)'\right]$$

- 从而检验统计量为：

$$\frac{1}{N}\left[\sum_{i=1}^{N} g\left(w_i,\tilde{\theta}\right)\right]'\left(\mathbb{V}\left[g\left(w_i,\theta\right)\right]\right)^{-1}\left[\sum_{i=1}^{N} g\left(w_i,\tilde{\theta}\right)\right] \overset{a}{\sim} \chi^2\left(r\right)$$

拉格朗日乘子检验

泊松分布的LM检验

- 对于泊松总体，我们知道总体矩条件为$\mathbb{E}(x_i - \lambda) = 0$从而矩估计为$\hat{\lambda} = \bar{x}$。
- 为了检验$H_0 : \lambda = 10$，在原假设的约束下，约束估计量为$\tilde{\lambda} = 10$
- 从而在原假设的条件下：

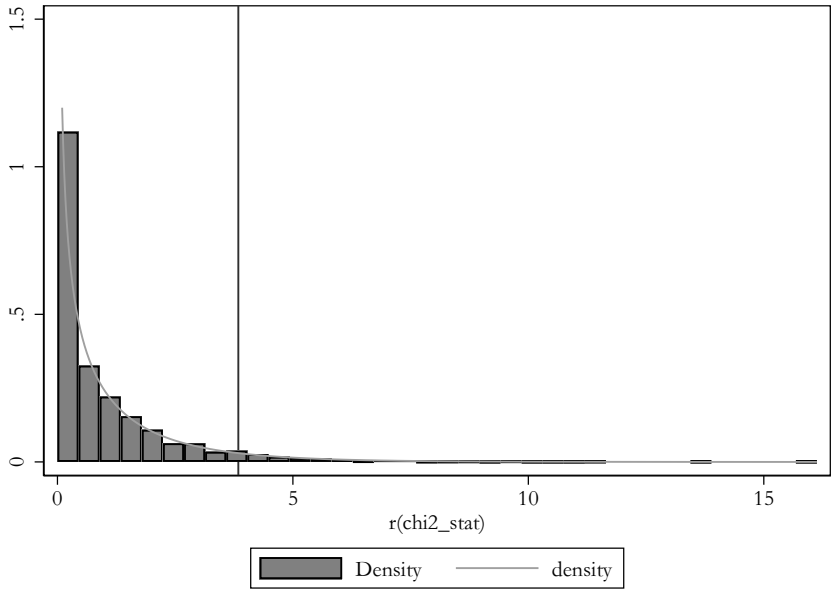
$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (x_i - 10) \overset{a}{\sim} N(0, 10)$$

- 进而可以构建检验统计量：

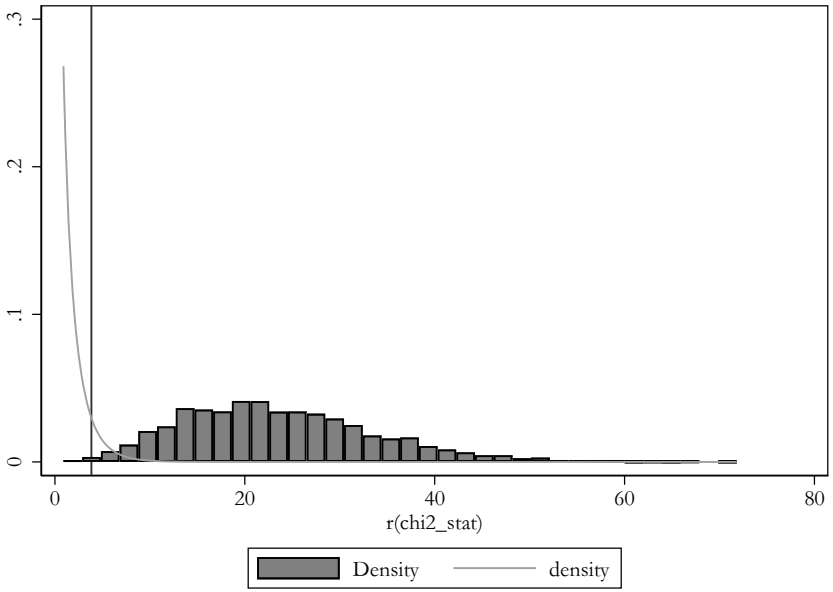
$$\frac{1}{N \times 10} \left[\sum_{i=1}^{N} (x_i - 10) \right]^2 \overset{a}{\sim} \chi^2(1)$$

- 模拟：simulation_lm_test.do

拉格朗日乘子检验


1.5
1
.5
0
0
5
10
15
r(chi2_stat)
Density
density

(a)


.3
.2
.1
0
0
20
40
60
80
r(chi2_stat)
Density
density

(b)

LM检验与模型设定检验

- 由于得分检验通常依赖于有约束的参数估计，而有约束的参数估计通常比较容易计算，所以在很多检验，特别是模型设定检验或者模型诊断中非常有用。
- 通常我们可以先计算一个简单的模型（带约束的），使用得分检验检验是否需要进行更复杂的模型
  - 如果该检验不显著，那么意味着使用复杂模型的必要性不足；
  - 否则如果检验显著，则意味着我们可能需要更复杂的模型。

LM检验与模型设定检验

一元线性回归中的LM检验

- 考虑一元线性回归，假设

$$\mathbb{E}\left(y_i|x_i\right) = \alpha + \beta x_i$$

  我们可以使用矩估计对以上问题进行估计。
- 然而我们可能会猜测x_i对y_i没有均值上的预测能力，即$\beta = 0$
- 如果该假设成立，那么以上问题就变成了一个简单的正态总体参数估计问题。

LM检验与模型设定检验

一元线性回归中的LM检验

- 为此，我们可以设定原假设为：$H_0:\beta=0$从而得到估计：$\tilde{\alpha}=\bar{y}$。
- 为了检验该原假设，注意其总体矩条件为

$$\begin{cases}\mathbb{E}\left(y_i-\alpha-\beta x_i\right)=0\\ \mathbb{E}\left[x_i\left(y_i-\alpha-\beta x_i\right)\right]=0\end{cases}$$

- 注意到原假设成立时，以上矩条件即

$$\begin{cases}\mathbb{E}\left(y_i-\alpha\right)=0\\ \mathbb{E}\left[x_i\left(y_i-\alpha\right)\right]=0\end{cases}$$

LM检验与模型设定检验

一元线性回归中的LM检验

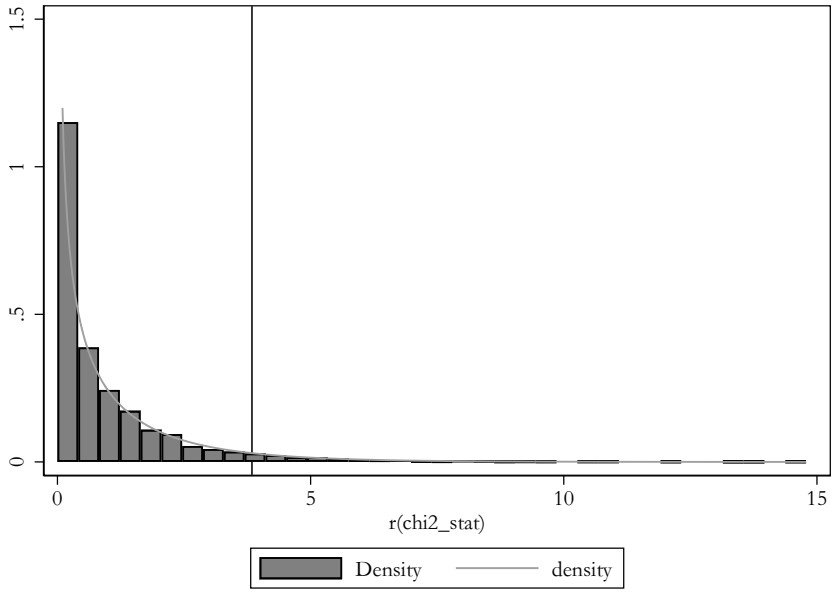
- 然而注意到样本矩条件中，$\sum_{i=1}^{N}(y_i-\tilde{\alpha})=0$必然成立，从而无需对该矩条件进行检验，只需要对$\mathbb{E}[x_i(y_i-\alpha)]=0$进行检验即可。
- 令$e_i=y_i-\tilde{\alpha}$，在原假设条件下

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(x_ie_i)\overset{a}{\sim}N(0,\mathbb{V}(x_ie_i))$$

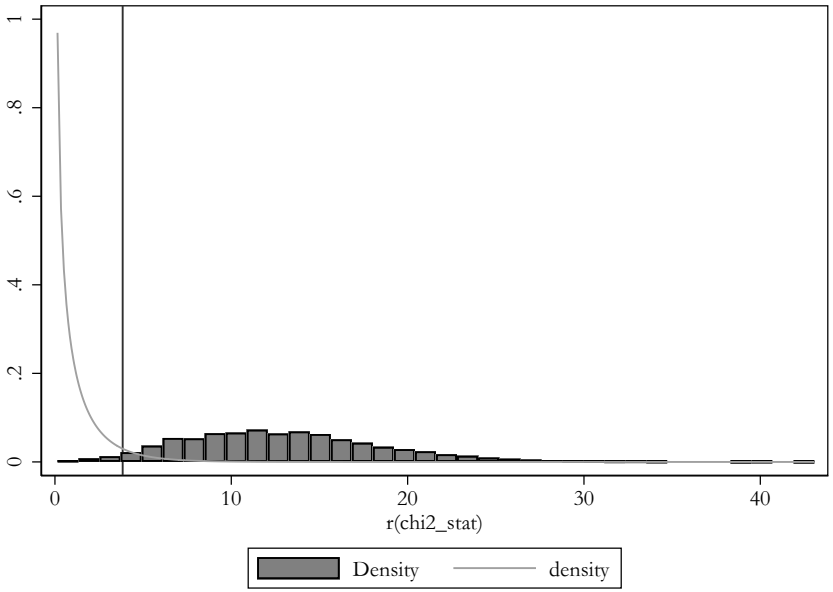
从而构造检验统计量：

$$\frac{\left[\sum_{i=1}^{N}(x_ie_i)\right]^2}{N\mathbb{V}(x_ie_i)}\overset{a}{\sim}\chi^2(1)$$

LM检验与模型设定检验



(a)



(b)

LM检验与模型设定检验

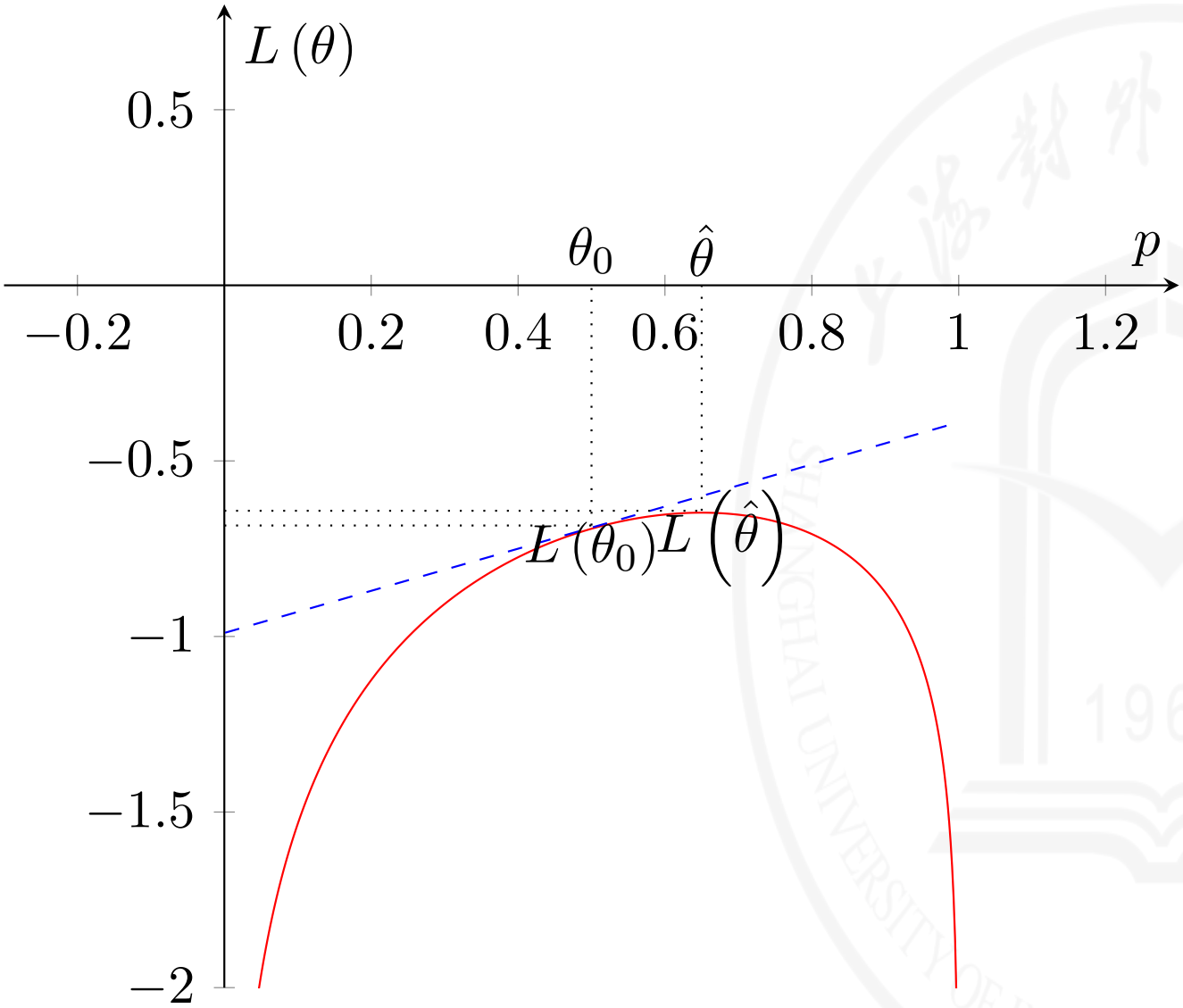
一元线性回归中的LM检验

- 因而，如果我们需要检验是否需要更复杂的线性回归模型，只需要计算以上统计量
  - 如果以上检验拒绝了原假设，那么意味着线性回归统计量是比简单的均值计算更好的模型，或者x对y有预测能力，从而可以进行线性回归的建模；
  - 反之，则代表现有证据对使用线性回归的证据不足，或者x对y有预测能力的证据不足。

三种检验的关系

- 以上我们介绍了三种构造假设检验的方法：Wald检验、似然比（LR）检验以及拉格朗日乘子（LM）检验。
- 其中
  - Wald检验需要估计不受约束的估计量；
  - LM检验需要估计受约束的估计量；
  - LR检验需要同时估计受约束和不受约束的估计量。
- 三种检验具有同样的极限分布（卡方分布）和自由度，实际上这并不是巧合，可以证明，以上三种检验是渐近等价的。

三种检验的关系


$L(\theta)$
0.5
−0.5
−1
−1.5
−2
θ_0
$\hat{\theta}$
p
−0.2
0.2
0.4
0.6
0.8
1
1.2
$L(\theta_0)$
$L(\hat{\theta})$

三种检验的关系

- 观察上图
  - Wald检验重在检验θ_0与$\hat{\theta}$之间的距离；
  - LR检验重在检验似然函数$L(\theta_0)$与$L\left(\hat{\theta}\right)$之间的距离；
  - LM检验则是检验θ_0处切线的斜率。
- 如果原假设成立，那么应该有$\hat{\theta} \approx \theta_0$，从而$L\left(\hat{\theta}\right) \approx L(\theta_0)$，而此时切线斜率应该为0（由于最大化）。
- 因此三者虽然形式不同，但是本质上相同的检验。

作业

1. 如果一个随机变量$X \sim N(0,1)$，现如下定义随机变量Y：

$$Y = \begin{cases} X-2 & with\ prob\ 0.5 \\ X+2 & with\ prob\ 0.5 \end{cases}$$

   求$\mathbb{V}(Y)$。
2. 证明$\mathbb{V}(Y) = \mathbb{V}[\mathbb{E}(Y|X)] + \mathbb{E}[\mathbb{V}(Y|X)]$
3. 如果$\mathbb{E}(Y|X) = \mathbb{E}(Y)$，那么有没有可能$\mathbb{E}(X|Y) \neq \mathbb{E}(X)$？请举反例或者证明。
4. 如果$Y|(X,Z) \sim N(X,Z)$，其中$X \sim N(\mu,\sigma^2)$，$Z \sim \chi^2(K)$，求$\mathbb{E}(Y)$及$\mathbb{V}(Y)$。
5. 证明：
   1. $\mathbb{V}(g(X)+Y|X) = \mathbb{V}(Y|Y)$
   2. $\mathbb{V}(g(X)Y|X) = [g(X)]^2\mathbb{V}(Y|X)$

作业

1. 如果$x_i \sim N(0,1)\ i.i.d$，那么请写出：
   1. $\chi_n^2 = \sum_{i=1}^n x_i^2$的分布
   2. $\frac{1}{n}\chi_n^2$的渐进分布
   3. 根据以上的结论，能否认为χ^2分布的极限分布是正态分布（或者说$\chi_n^2 \overset{a}{\sim} N(n, 2n)$是否成立？）？
2. 已知：如果$x \sim N(\mu, \sigma^2)$，那么

$$\mathbb{E}\left[(x - \mathbb{E}(x))^4\right] = 3\sigma^4$$
$$\mathbb{E}\left[(x - \mathbb{E}(x))^6\right] = 15\sigma^6$$
$$\mathbb{E}\left[(x - \mathbb{E}(x))^8\right] = 105\sigma^8$$

   对于样本偏度系数：

$$b_1 = \frac{N^2}{(N-1)(N-2)} \frac{\frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})^3}{s^3}$$

   请计算：
   1. b_1的概率极限
   2. $\sqrt{N} b_1$的极限分布
3. 若$x_i \sim N(\mu, \sigma^2)$，求$\sqrt{N}\bar{x}^2$的极限分布和$\sqrt{N}\,\overline{x^2}$的极限分布。

作业

1. 如果$x_i|\left(\mu_i,\sigma_i^2\right)\sim N\left(\mu_i,\sigma_i^2\right),\mu_i\sim N\left(a,1\right),\sigma_i^2\sim U\left(0,b\right)$，其中$\left(\mu_i,\sigma_i^2\right)$均不可观测。请求出$a,b$的矩估计，并给出$a$的95%置信区间。
2. 如果$\left(y_i,x_i'\right)',i=1,...,N,x_i\in\mathbb{R}^K$为一系列独立同分布的随机向量，且$y_i\in\mathbb{Z}$，经常使用的模型为泊松回归（Possion regression），即假设：

$$y_i|x_i\sim P\left(e^{x_i'\beta}\right)$$

   请写出其条件对数似然函数。
3. 若$x_i\sim N\left(\mu,\sigma^2\right)$，请构造统计量检验：$H_0:\frac{\mu}{\sigma^2}=1$
4. 若$x_i\sim E\left(0,\beta\right)$，即：$f\left(x_i|\beta\right)=\frac{1}{\beta}e^{-\frac{x}{\beta}},x>0$请：
   1. 求出β的极大似然估计；
   2. 请在$H_0:\beta=10$的原假设下构建检验统计量及拒绝域；
   3. 请写出该假设检验的势函数。