第四部分：统计学方法

参数估计
(Parameter estimation)

什么是参数估计：

- 假定我们对某一物理系统进行了测量，得到了容量为**n**的事例样本；
- 我们想从这一有限样本中获取有关该物理系统的信息；
- 描述该物理系统的概率分布数学形式是已知的，但其中包含了某些未知的参数；
- 参数估计的任务是通过对观测到的事例样本的统计分析来最大限度地获取有关这些未知参数的信息。

例：共振峰参数的估计：

概率密度函数： $f(m;m_0) \propto \dfrac{\sqrt{\Gamma}}{(m-m_0)^2+\frac{1}{4}\Gamma^2}$ ←**Breit-Wigner**公式

未知参数：

$\mathbf{m_0}$=共振峰的质量

Γ=共振峰的宽度

观测量：

m=共振峰衰变产物的不变质量

参数估计的内容：

1、点估计(**Point Estination**)：估计未知参数的值

2、区间估计(**Interval Estimation**)：估计未知参数的估计值的精确性和可靠性

$$\theta = \hat{\theta} + \Delta\theta$$

本章介绍统计学中的参数估计的一些基本概念和方法

总体的概率密度函数(pdf)：f(x|θ)

θ：未知参数

x：实验可测量量

随机样本(容量为n)：

$$x_1, x_2, \cdots, x_n$$

x_i：独立的随机变量

一、基本定义

、似然函数(Likelihood Function, LF):

由于x_i是相互独立的随机变量，因而在给定的θ值下获得测量量$x_1,x_2,\dots,x_n$的联合条件概率为**(Joint Conditional Probability)**

$$L(x_1,x_2,\cdots,x_n\mid\theta)=\prod_{i=1}^{n}f(x_i,\theta)$$

(1)似然值(Likelihood)：

如果θ和x_i都为固定值，则称L 为在特定的θ值下，观测量$x_1,x_2,\dots,x_n$的似然值；

(2)似然函数(LF)：

如果将L看成是θ的函数，而x_i固定，则称L为似然函数；

(3)可测量量x_i得pdf：

θ固定，L是x_i的函数

2、统计量(Statistic)：

如果$t=t(x_1,x_2,\ldots,x_n)$是样本变量x_i的函数，且不依赖于任何的未知参数θ，则称t为统计量

例：样本的平均值和方差：

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \quad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

、估计式(Estimator)：

如果统计量t给出了未知参数θ的估计值，则称t为θ的估计式，即

$$\hat{\theta}=t$$

例：样本平均值 $\bar{X}$和方差s^2分别是总体平均值μ和方差σ^2的估计式。

参数估计的目标之一就是求出未知参数的估计式

二、估计式的特性

由估计式t得到的参数θ的估计值 $\hat{\theta}$是随机变量，将满足某种分布，这种分布的特性将反映该估计式的好坏

判断估计式好坏的标准：

(1)一致性(**Consistency**)：样本容量为无限大时估计式的特性

(2)无偏性(**Unbiassedness**)：样本容量为有限时估计式的特性

(3)最小方差(**Minimum variance**)
有效性(**Efficiency**)
估计式的分布特性

(4)充分性(**Sufficiency**)：估计式是否包含了样本中所包含的有关θ的所有信息

、一致性(Consistency):

如果一个估计式的值当样本容量增加时收敛到待估参数的真值，则称该估计式具有一致性

概率语言的一致性描述：

如果估计值θ_n是从容量为n的样本得到的，则对于给定的正数ε和η，存在着正整数N，使得对所有的n>N，$|\theta_n - \theta| > \varepsilon$的概率小于$\eta$

$$P(|\theta_n - \theta| > \varepsilon) < \eta$$

即，当$n \to \infty$时，$\theta_n \to \theta$

例: 样本平均值$\bar{x}$是总体平均值μ的一致性估计式

根据大数定理：当$n \to \infty$时, $\bar{x} \to \mu$

2、无偏性(Unbiassedness):

对于任意大的样本，如果估计式t的期望值都等于参数的真值θ

$$E(t)=\int t(x_1,x_2,\cdots,x_n)L(\underline{x}\mid\theta)d\,\underline{x}=\theta$$

则称t是θ的无偏估计式

注：

- 无偏性保证了估计式的值不会系统地偏离参数θ的真值；
- 一致性和无偏性是不相关的，具有一致性并不等于具有无偏性
- 一致性和无偏性是对参数估计式的基本要求，因为参数估计的目的就是求θ的真值。

3、最小方差和有效性

估计值 是随机变量，服从一定的分布，好的估计式给出的估计值的方差应尽可能地小。

假定：(1)对所有的θ，L($\underline{x}$| θ)对θ的一、二阶导数存在；
(2)变量x的定义域与θ无关；

则由估计式得到的估计值的方差存在着一个下限

设t是τ(θ)的估计式， τ(θ)为 θ的函数，估计值的偏差为b(θ)

$$E(t)=\int\cdots\int t(x_1,x_2,\cdots,x_n)L(\underline{x}\,|\,\theta)dx_1dx_2\cdots dx_n=\tau(\theta)+b(\theta)$$

估计式t的方差V(t)满足下列Cramer-Rao不等式：

$$V(t)\ge\left(\frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}\right)^2\Big/E\left[\left(\frac{\partial\ln L}{\partial\theta}\right)^2\right]=\left(\frac{\partial\tau}{\partial\theta}+\frac{\partial b}{\partial\theta}\right)^2\Big/E\left(\frac{\partial^2\ln L}{\partial\theta^2}\right)$$

➔最小方差限(Minimum Variance Bound, MVB)

有效估计式：方差等于最小方差限的估计式

t为有效估计式的充分必要条件：

$$\frac{\partial \ln L}{\partial \theta} = A(\theta)[t-\tau(\theta)-b(\theta)]$$

MVB： $V(t) = \left(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}\right) / A(\theta)$

在实际应用中，有效估计式只是在有限的几类参数估计问题中存在。

例如：泊松分布样本的样本平均值是泊松总体平均值的有效估计式

4、充分性(Sufficiency)

设t是参数θ的估计式，如果测量量中所包含的有关θ的信息都包含在t内，则称t为θ的充分估计式

充分估计式的存在有利于数据的浓缩**(Data Reduction)**：

t中所包含的有关θ的信息与原始数据中的一样多；或者：任何其它的原始数据的函数都给不出更多的有关参数θ的信息

R.A.Fisher的信息的定义：

由观测量**x**给出的有关未知参数θ的信息量的定义：

$$I_x(\theta) = E\left[\left(\frac{\partial \ln L(x|\theta)}{\partial \theta}\right)^2\right] = \int_\Omega \left(\frac{\partial \ln L(x|\theta)}{\partial \theta}\right)^2 L(x|\theta)dx$$

如果θ是**k**维的

$$[I_x(\theta)]_{ij} = E\left(\frac{\partial \ln L(x|\theta)}{\partial \theta_i} \cdot \frac{\partial \ln L(x|\theta)}{\partial \theta_i}\right)$$

根据此定义，若**t**为θ的充分估计式，则

$$\mathbf{I_t(\theta)=I_x(\theta)}$$

一、基本概念

t 是参数θ的充分估计式的充分必要条件：似然函数**L(x| θ)**可分解为如下的形式：

$$L(\underline{x}\mid\theta)=\prod_{i=1}^{n} f(x_i,\theta)=G(t\mid\theta)H(\underline{x})$$

其中：

i) **H(x)**与参数θ无关；

ii) **G(t| θ)**是估计式**t**的函数，表示在θ一定的条件下**t**得**pdf**

可证：有效估计式总是具有充分性

注：充分统计量只对某些特殊类型的**pdf**存在；如果**f(x, θ)**为指数形式：

$$f(x,\theta)=\exp[B(\theta)C(x)+D(\theta)+E(x)]$$

则充分统计量**t**一定存在，且

$$t=\sum^{n} C(x_i)$$

例：在σ²已知的情况下，样本平均值 是正态分布N(μ, σ²)中μ的充分估计式

$$L(\underline{x};\sigma^2\mid\mu)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma}}\exp\left[-\tfrac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]$$

$$=\underbrace{\left\{\frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}}\exp\left(-\tfrac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2\right)\right\}}_{G(\underline{x}\mid\mu)}\underbrace{\left\{\frac{n^{-\frac{1}{2}}}{\left(\sqrt{2\pi}\sigma\right)^{n-1}}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2\right)\right\}}_{H(\underline{x})}$$

$$N(\mu,\sigma^2)=\exp\left[\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\frac{x^2}{2\sigma^2}-\tfrac{1}{2}\ln(2\pi\sigma^2)\right]$$

$$B(\mu)=\frac{\mu}{\sigma^2}\quad C(x)=x\quad D(\mu)=-\frac{\mu^2}{2\sigma^2}\quad E(x)=-\frac{\mu^2}{2\sigma^2}-\tfrac{1}{2}\ln(2\pi\sigma^2)$$

$$\therefore t=\sum_{i=1}^{n}C(x_i)=\sum_{i=1}^{n}x_i\sim\bar{x}$$

二、 参数的区间估计

区间估计的目的：

找出未知参数θ的一个变化范围

$$\theta_a \le \theta \le \theta_b$$

使得θ的真值落入该范围的概率为γ

一、区间估计的基本概念

1、置信区间(**Confidence Interval**)

若参数θ的真值落入闭区间[θ_a, θ_b]内的概率为γ，则称该区间为参数θ的**100**γ**%**置信区间

$$p(\theta_a \le \theta \le \theta_b) = \gamma$$

γ：置信系数（置信水平）

θ_a, θ_b ：置信限(**Confidence Limits**)

在实验上，置信区间对应于θ的估计值的误差

特性：

1) 是随机的：由两个容量相同的样本得到的置信区间一般是不同的

2) 置信区间可能包含θ的真值，也可能不包含；对于一个特定样本

$$p(\theta_a \le \theta \le \theta_b) = 1 \quad \Rightarrow \theta \in [\theta_a, \theta_b]$$

$$p(\theta_a \le \theta \le \theta_b) = 0 \quad \Rightarrow \theta \notin [\theta_a, \theta_b]$$

γ反映了不等式$\theta_a \le \theta \le \theta_b$的可靠性

3) 两难性(**Dilemma**)：

$\theta_b - \theta_a$大，γ大，但参数θ的不确定性大；

$\theta_b - \theta_a$小，γ小，但对参数θ的确定具有较高的精度；

实验上一般取γ=68.3%或95.5%，分别对应一个和二个标准偏差的置信区间；

2、区间估计的基本方法

区间估计就是：给定置信系数γ，根据参数θ的分布，求出置信区间

设统计量t是参数θ的估计式，t的pdf为f(t)

$$p(\theta_a \le \theta \le \theta_b) = \gamma = \int_{\theta_a}^{\theta_b} f(t)dt$$

$[\theta_a, \theta_b]$即为欲求的置信区间

1) 如果f(t)与θ无关，则可通过求解上述积分方程求出θ_a和θ_b

2) 如果f(t)与θ有关，则上式中的积分将无法计算

z=z(t, θ) ➔pdf f(z) 与θ无关

$$\gamma = \int_{z_a}^{z_b} f(z)dz \Rightarrow z_a, z_b \Rightarrow \theta_a, \theta_b$$

假设检验
(Hypothesis Testing)

一．假设检验的基本概念

二．假设检验的一般方法

三．假设检验的一个例子：**Li－Ma**显著性（**Significance**）

一. 假设检验的基本概念

1. 什么是假设检验

实验的目的：验证一个科学论断的正确性

假设检验：利用概率和统计的语言，根据实验的结果来验证一个理论模型是否可接受。

统计假设：待检验的理论模型

例：Ξ^0粒子的衰变。

实验结果：测量衰变时间求Ξ^0粒子的平均寿命τ。

理论模型：$\Delta I=1/2$规则，Ξ^0的寿命是Ξ^-的两倍，

Ξ^-的寿命 ➔ Ξ^0的寿命τ_0

问题：$\tau \overset{?}{=} \tau_0$

由于τ有测量误差，对该问题的回答：$\tau=\tau_0$概率是多少？

2．假设检验的分类

（1）参数检验：如果欲检验的统计假设只包括某些参数的特定值，如：$\tau \overset{?}{=} \tau_0$

（2）非参数检验：被观测的随机变量的分布是否符合一个特定的函数形式？两个给定的实验分布是否具有相同分布形式？……

3．原假设和备择假设（**Null Hypothesis**，**Alternative Hypothesis**）

原假设：欲检验的统计假设，如

$$H_0: \tau = \tau_0$$

备择假设：实验结果有可能支持原假设，也可能支持别的假设而拒绝原假设，与原假设不同的其它假设称为备择假设，如

$$H_1: \tau \neq \tau_0$$

一般情况下，是否接收原假设依赖于与备择假设的比较结果。

4．简单假设和复合假设（**Simple Hypothesis**，**Composite Hypothesis**）

简单假设：假设中参数的值是一常数，如

$$H_0：\tau=\tau_0$$

复合假设：假设中的某一参数的值不是完全确定的，如

$$H：\tau\neq\tau_0、\quad H_1：\tau\geq\tau_0$$

如何选择原假设和备择假设，要根据所要解决的实际问题决定

二．假设检验的一般方法

参数检验

随机变量x

p.d.f.: $f(x, \theta)$， θ为未知参量

观测结果：容量为n的样本， $(x_1, x_2, \ldots, x_n)$

检验 θ是否取某一值

原假设 H_0： $\theta = \theta_0$

备择假设 H_1： $\theta = \theta_1$

定义通过观测结果来接收原假设或拒绝原假设的标准

检验统计量：$t = t(x_1, x_2, \ldots, x_n)$

t 的定义域：ω

$f(t|H_0)$：H_0为真时，t 的p.d.f.

$f(t|H_1)$：H_1为真时，t 的p.d.f.

R： ω中的子域

α：t 落入R中的概率。 $0 \le \alpha \le 1$ （H_0为真时）

$$\alpha = P(t \in R \mid H_0) = \int_R f(t \mid H_0)dt$$

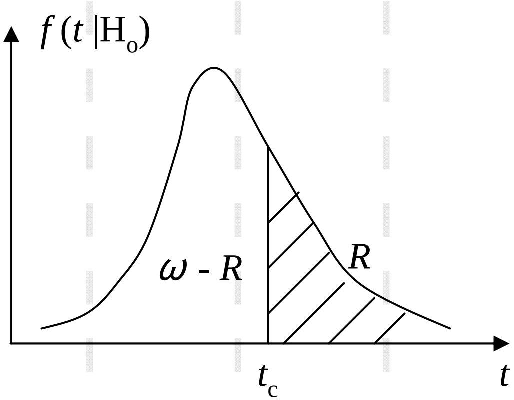
R： H_0 的拒绝域

$\omega - R$： H_0 的接收域

即：若t 的观测量t_{obs}，

落入R，则拒绝H_0

否则，接收H_0



α：显著性水平（Significance Level），t_c：临界值

第一类错误（弃真错误）：当H_0为真时，t_{obs}有 α的概率落入R

当$t_{obs} > t_c$时， H_0被拒绝，而实验上H_0为真

I类错误的概率：

$$\alpha = \int_R f(t \mid H_0)dt$$

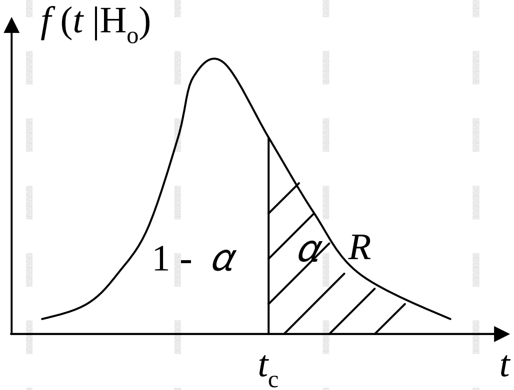
➔ α应尽可能地小

第二类错误（取伪错误）：H_0不为真，但却接收了H_0

II类错误的概率：

$$\beta = P(t \in \omega - R \mid H_1) = \int_{\omega - R} f(t \mid H_1)dt$$

$$1 - \beta = P(t \in R \mid H_1) = \int_R f(t \mid H_1)dt$$



1- β：H_0 对H_0的检验势

1- β大，II类错误的概率小

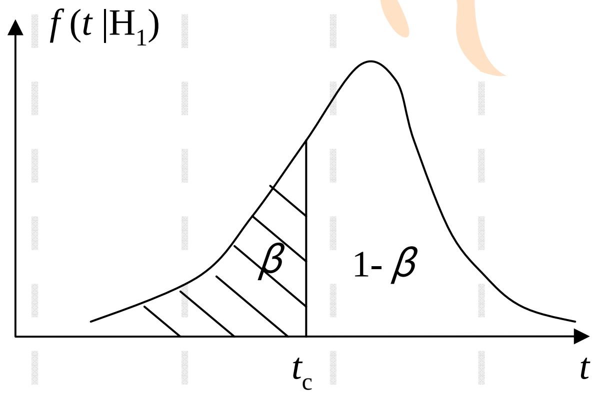
假设检验的方法：

1. 选择合适的检验统计量t

2. 选择适当的临界值t_c

$$\alpha = \int_R f(t \mid \mathrm{H}_0)dt$$

➔ α应尽可能地小



标准：α尽可能地小，1- β尽可能大。

三. 复合假设的检验：似然比（Likelihood Ratio）

设x的p.d.f.：$f(t \mid \vec{\theta})$，x样本：$(x_1, x_2, \ldots, x_n)$

$$\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$$

Ω：$\vec{\theta}$的取值空间

ω：Ω的子空间，即$\vec{\theta}$的分量中只有一个受到某种约束

原假设 $\mathrm{H}_0: \vec{\theta} \in \omega$

备择假设 $\mathrm{H}_1: \vec{\theta} \in \Omega - \omega$

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

设 $L(\hat{\Omega})$ ：L 在Ω中的极大值

$L(\hat{\omega})$ ：在H_0为真时，L 在ω中的极大值

定义： $\lambda = \dfrac{L(\hat{\omega})}{L(\hat{\Omega})}$ ：似然比（Likelihood Ratio）

$\because$ $L(\hat{\omega})$ 不可能比 $L(\hat{\Omega})$ 大 $\therefore$ $0 \le \lambda \le 1$

$\lambda \sim 1$ $L(\hat{\omega}) \sim L(\hat{\Omega})$ ➔H_0 为真的可能性较大

$\lambda \sim 0$ $L(\hat{\omega}) << L(\hat{\Omega})$ ➔H_0 为真的可能性较小

$\therefore$ λ可作为原假设H_0的检验统计量

对H_0的检验 ➔求在给定的显著性水平 α下，λ的临界值 λ_α：

$$L = \int_0^{\lambda_\alpha} g(\lambda | H_0) d\lambda$$ ：L 在Ω中的极大值

$g(\lambda | H_0)$：在H_0 为真时，λ的p.d.f.

如果$g(\lambda | H_0)$的函数形式未知：

$$y=y(\lambda) \quad ➔ \quad h(y | H_0)$$

$$L = \int_0^{\lambda_\alpha} g(\lambda | H_0) d\lambda = \int_{y(0)}^{y(\lambda_\alpha)} h(y | H_0) d\lambda$$

$$y(\lambda_\alpha) \rightarrow \lambda_\alpha$$

一般情况下，$g(\lambda | H_0)$很难找到 ➔ 采用近似方法：

设 $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$中当$H_0$ 为真时有 γ个参数取固定值，则

当样本容量n很大时，统计量$-2\ln\lambda$趋近于自由度为 γ的 χ^2分布

➔ $\chi^2(\gamma)$分布求λ_α

令 $\mu = -2\ln\lambda$ $\quad L = \int_0^{\mu_\alpha} \chi^2(\mu, \gamma) d\mu$ ➔ λ_α

例：γ点源寻找中的Li-Ma显著性

问题：来自某一天体方向的事例数的超出是γ点源信号还是背景涨落，如果是信号，如何用统计学的方法描述这种超出？

实验测量：

N_{on}：向源事例数

N_{off}：离源事例数

t_{on}：向源测量时间

t_{off}：离源测量时间

未知量：γ事例数N_s，源方向的背景事例数N_B

背景事例：强子和核引起的簇射，是各向同性的，

可用似然比检验解决上述问题。

原假设 $H_0: N_s = 0$

备择假设 $H_1: N_s \neq 0$

N_{on}、N_{off}服从什么分布？

假定：事例率为常数，则N_{on}、N_{off}服从Possion分布

$$P(N,\mu) = \frac{1}{N!}\mu^N e^{-\mu}$$

μ：平均值

N_S、N_B的估计值：

令L=t_{on}/t_{off}

$$\hat{N}_B = \frac{N_{off}}{t_{off}} \cdot t_{on} = \alpha N_{off}$$

$$\hat{N}_S = N_{on} - \hat{N}_B = N_{on} - \alpha N_{off}$$

如果H_0 为真，N_s=0，$\hat{N}_B = \frac{N_{on} + N_{off}}{t_{on} + t_{off}} \cdot t_{on} = \frac{\alpha}{1+\alpha}(N_{on} + N_{off})$

N_{on}和N_{off}的平均值：

（1）H_0 为真时，N_{on}全为本底事例。

$$\langle \hat{N}_{on} \rangle^0 = \frac{N_{on} + N_{off}}{t_{on} + t_{off}} \cdot t_{on} = \frac{\alpha}{1+\alpha}(N_{on} + N_{off})$$

$$\langle \hat{N}_{off} \rangle^0 = \frac{N_{on} + N_{off}}{t_{on} + t_{off}} \cdot t_{off} = \frac{1}{1+\alpha}(N_{on} + N_{off})$$

（2）H_1 为真时：，$N_S \neq 0$

$$\langle N_{on} \rangle_1 = \hat{N}_S + \hat{N}_B = N_{on}$$

$$\langle N_{off} \rangle_1 = N_{off}$$

参数空间：N_B：x轴，N_S：y轴

H_0 ：$N_S=0$ ： $\omega =$ x轴

H_1 ：$N_S \neq 0$ ： $\Omega = N_S > 0$，$N_B \geq 0$

H_0为真时，有一个参数取定值：自由度为1

似然函数：

由于N_{on}和N_{off}是独立的随机变量，故一次实验获得测量量N_{on}和N_{off}概率为

$$L = P(N_{on}) \cdot P(N_{off})$$

H_0：

$$P(N_{on}) = \frac{1}{N_{on}!}\langle N_{on}\rangle_0^{N_{on}} e^{-\langle N_{on}\rangle_0}$$

$$P(N_{off}) = \frac{1}{N_{off}!}\langle N_{off}\rangle_0^{N_{off}} e^{-\langle N_{off}\rangle_0}$$

$$L(N_{on}, N_{off} \mid H_0) =$$

H_1：

$$P(N_{on}) = \frac{1}{N_{on}!} N_{on}^{N_{on}} e^{-N_{on}}$$

$$P(N_{off}) = \frac{1}{N_{off}!} N_{off}^{N_{off}} e^{-N_{off}}$$

$$L(N_{on}, N_{off} \mid H_1) =$$

似然比：

$$\lambda=\frac{L(N_{on},N_{off}\mid H_0)}{L(N_{on},N_{off}\mid H_1)}=\left[\frac{\alpha}{1+\alpha}\left(\frac{N_{on}+N_{off}}{N_{on}}\right)\right]^{N_{on}}\left[\frac{1}{1+\alpha}\left(\frac{N_{on}+N_{off}}{N_{off}}\right)\right]^{N_{off}}$$

如果N_{on}、N_{off}不是很小， $-2\ln\lambda \rightarrow \chi^2(1)$

χ^2分布的定义：如果$\chi \rightarrow N(\mu,\sigma^2)$

$$u=\left(\frac{x-\mu}{\sigma}\right)^2 \rightarrow \chi^2$$

$\sqrt{u}$：x偏离平均值μ多少标准偏差

$$u \sim -2\ln\lambda$$

∴ 在H_0假设为真时， $S=\sqrt{-2\ln\lambda}$ =观测结果偏离H_0多少个标准偏差。

$$S = \sqrt{-2\ln\lambda}$$

$$= \sqrt{2}\left\{ N_{on} \ln\left[\frac{1+\alpha}{\alpha}\left(\frac{N_{on}}{N_{on}+N_{off}}\right)\right] + N_{off} \ln\left[(1+\alpha)\left(\frac{N_{off}}{N_{on}+N_{off}}\right)\right]\right\}^{\frac{1}{2}}$$

➔李-马显著性(Li-Ma significance)