

# 贝叶斯方法确定 泊松变量的置信上限

朱永生

中科院高能所，北京100049

第7届全国高能物理学术年会

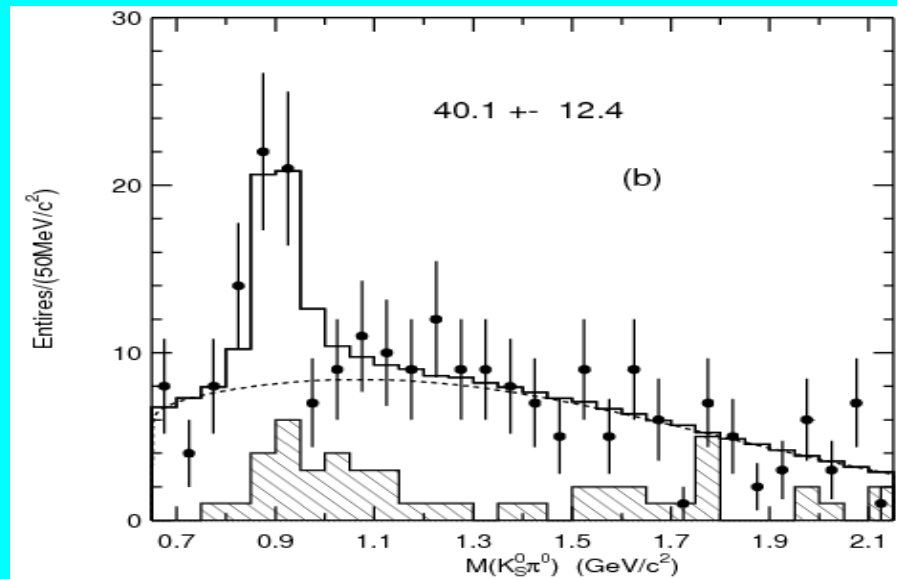
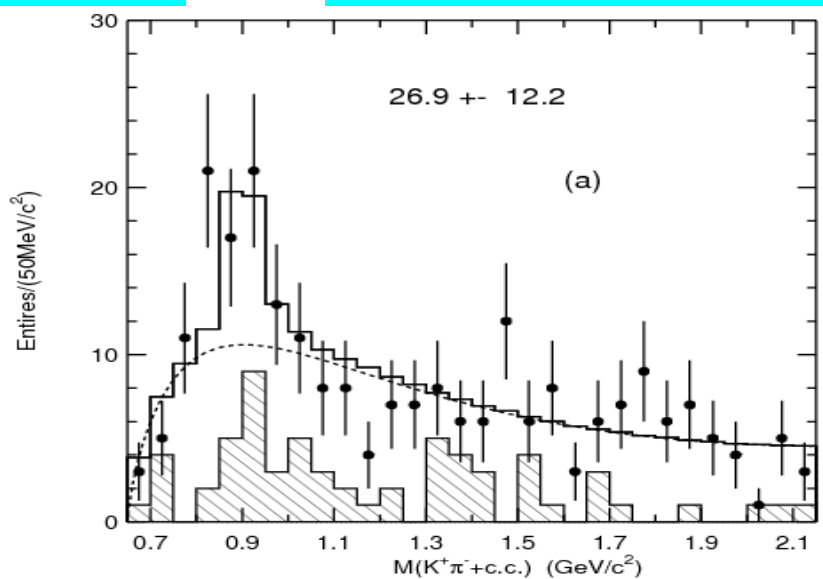
2006.10.28 - 11.1， 中国昆明

- **问题的提出**
- **经典统计方法**
- **Bayes 统计方法**
- **程序包 BPULE**
- **结束语**

# (一) 问题的提出

研究  $\psi' \rightarrow \pi^0 K_S^0 K^+ \pi^- + c.c. + \dots$

测量  $|K^{\pm} \pi^{\mp}|$   $|K^0 \pi^0|$  不变质量谱, 寻找  $K^*(892)^{\pm}$ ,  $K^0(892)$  信号。



信号区内事例数  $n$

既有本底

也有信号  $s$

信号和本底的确定包含系统误差

如何确定一定置信水平  $CL = \gamma = 1 - \alpha$  下的置信区间或置信上限?

# (一) 问题的提出 (续)

信号区内事例数  $n$  服从泊松分布,

$$P(n | s) = e^{-(s+b)} \frac{(s+b)^n}{n!}.$$

本底事例数服从期望值  $b$  的泊松分布,  
信号事例数服从期望值  $s$  的泊松分布。

$b$  的确定存在系统误差

信号事例的探测效率存在系统误差

怎样确定  $s$  的一定置信水平的置信区间或上限?

# (二) 经典统计方法

## Neyman方法

(J.Phil.Trans. Royal Soc. London, Series, A236,333(1937))

### 经典的区间估计方法

设待估计参数为  $s$  , 实验观测值为  $x$  , 所谓区间估计问题, 是要从实验观测值来确定的一个区间, 满足  $P(s \in [s_1, s_2]) = \gamma$  或  $P(s \in [0, s_{UP}]) = \gamma$

缺点—— flip – flopping

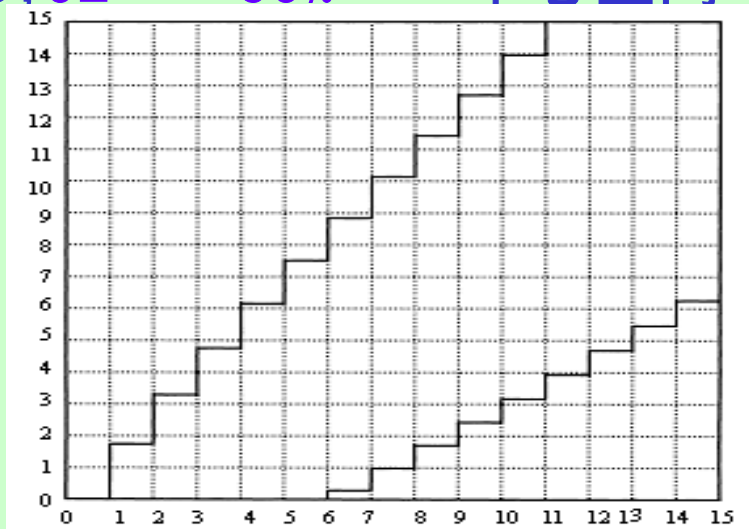
under-coverage (测量值靠近物理边界)

$n=0$  时, 给不出置信区间

$b=3, CL=\gamma=90\%$

中心区间

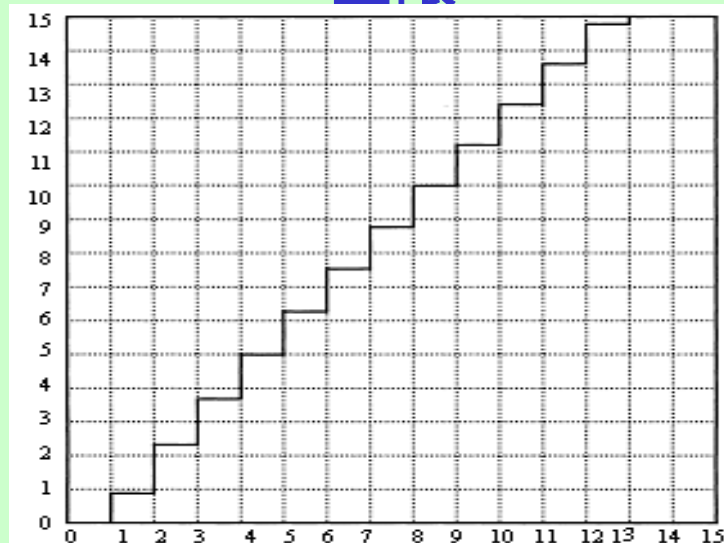
$s$



$x$

上限

$s_{UP}$



$x$

## (二) 经典统计方法 (续)

Feldman-Cousins方法

P.R.D57,3873(1998)

基于似然比求和确定置信区间

优点 - - 克服了Neyman方法的缺点：

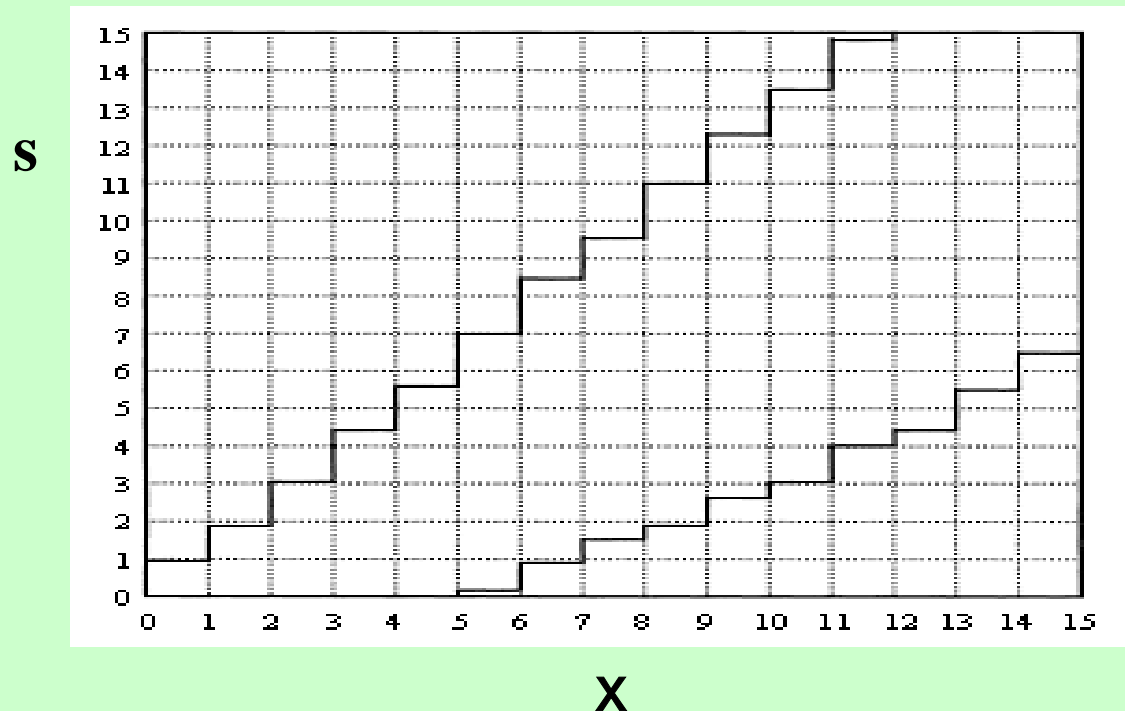
自动确定报道中心区间或上限 ( Unified approach )

有正确的coverage

$n=0$  时，给出置信区间

$CL = \gamma = 90\%$

$b=3$



## (二) 经典统计方法 (续)

### Feldman-Cousins方法

**缺点** - -  $n=0$  时, 置信区间上限依赖于平均本底  $b$  的大小

$b=0$      $S_{UP}=2.44$              $CL= \gamma = 90\%$

1            1.61

2            1.26

3            1.08

4            1.01

从实际出发来考虑问题, 既然总的观测事例数  $n=0$ , 实际的信号事例数和本底事例数的期望值都应当是零, 这时的置信区间基本上不应随预期的平均本底而变化 (预期的平均本底不准确) .

## (二) 经典统计方法 (续)

### 改进的似然比方法

B.Roe & M.Woodroffe: P.R.D60,053009(1999)

基本思想是对于任一特定观测总事例数 $n$ ，本底事例数不可能大于 $n$ 。将这一要求考虑到置信区间的构造上，原来的Poisson概率密度函数

$$p(k|s) = e^{-(s+b)} \frac{(s+b)^k}{k!}$$

要用条件概率密度代替： $q_{s+b}^n(k)$

似然比顺序求和方法(虚线)

和改进方案(实线)

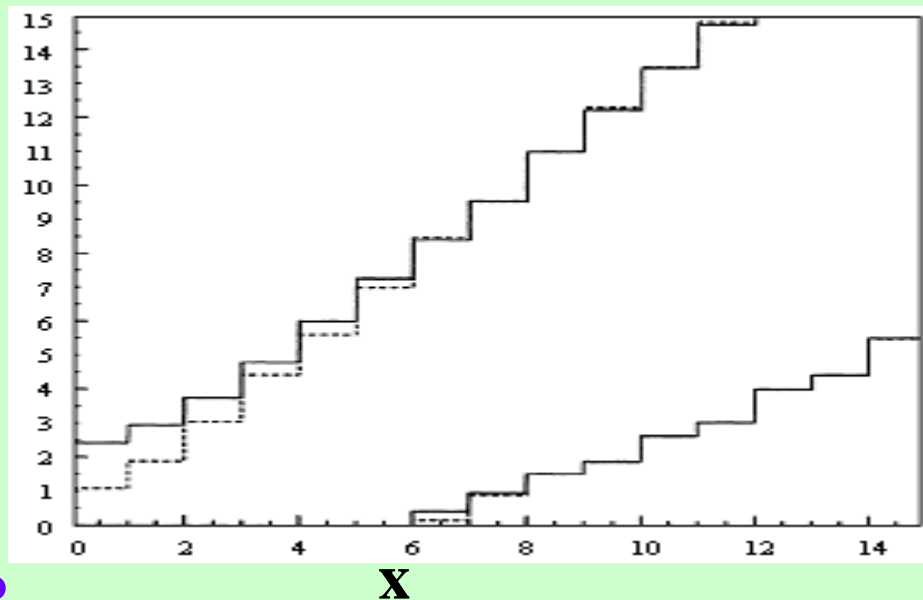
构造的置信带的比较

观测值服从泊松分布，

$$\gamma = 0.90, b = 3$$

$$b=3, S_{UP} = 2.42, CL = \gamma = 90\%$$

s





## (二) 经典统计方法 (续)

### 考虑系统误差

>> B.Cousins & V.Highland: NIM A320,331(1992)

包含信号探测效率的系统误差 (semi-Bayesian approach) :  
将信号探测效率考虑为一种概率分布, 对不同概率的探测效率  
求平均。

>> W.Rolke & A.Lopez: NIM A458,745(2001)

包含平均本底的系统误差(entirely frequentist approach)  
基于似然比假设检验和两维置信带的构建方法。

>> J.Conrad et al PRD67,012002(2003)

包含信号探测效率的系统误差以及平均本底的系统误差  
(semi-Bayesian approach)

程序包POLE 可计算  $n < 100$ ,  $b < 50$  的置信区间。

## (三) Bayes统计方法

### (1) 基本方法

- 关于未知参数（信号期望值  $s$ ）的先验（prior）pdf:  $\pi(s)$   
（反映实验者在测量之前对未知参数的认识）

- 实验测量：信号为  $s$ ，观察到  $n$  个事例的条件 pdf:

Poisson分布

$$p(n|s) = e^{-(s+b)} \frac{(s+b)^n}{n!}. \quad (1)$$

- 后验（posterior）pdf: Bayes 原理  
（先验知识和实验测量知识的综合）

$$h(s|n) = \frac{p(n|s)\pi(s)}{\int_0^{\infty} p(n|s)\pi(s)ds}. \quad (2)$$

- 后验pdf是Bayes统计中进行统计推断的基础。

## (三) Bayes统计方法 (续)

### (1) 基本方法

- 关于未知参数 (信号期望值  $s$ ) 的Bayes置信区间 (置信水平  $CL = \gamma = 1 - \alpha$ ) :

$$1 - \alpha = \int_{s_L}^{s_U} h(s | n) ds. \quad (3)$$

- 唯一确定的置信区间 (Highest Posterior Density) **HPD**置信区间  $I$  :

$$1 - \alpha = \int_I h(s | n) ds, \quad (4)$$

$$h(s_1 | n) \geq h(s_2 | n) \quad \text{for any } s_1 \in I \text{ and } s_2 \notin I.$$

- 上限  $s_{UP}$

$$1 - \alpha = \int_0^{s_{UP}} h(s | n) ds. \quad (5)$$

## (三) Bayes统计方法 (续)

### (2) informative & non-informative prior

- Informative prior

prior pdf 考虑先前实验的知识

- Non-Informative prior (无信息先验密度)

(a) **Bayes '同等无知假设'** :  $\pi(\theta)$ 为均匀分布。

$\pi(\theta) \propto 1$ , 对参数  $\theta$  在其取值区间内取任意值有等可能性。

缺点 :  $\pi(\theta)$ 为均匀分布 ,  $\pi(g(\theta))$  亦应是均匀分布。

事实上不是。

(b) **Jeffreys不变无信息先验密度** :

$$\pi(\theta) \propto |I(\theta)|^{1/2}, \quad I(\theta) \text{ Fisher information}$$

对于 (1) 式的Poisson分布 ,  $\pi(s) \propto 1/\sqrt{s+b}$  (6)

# (三) Bayes统计方法 (续)

## (3) 置信上限

### Poisson分布的后验密度

一般的无信息先验密度

得到相应的后验密度

$$\pi(s) \propto 1/(s+b)^m. \quad (7)$$

$$h(s|n) = \frac{(s+b)^{n-m} e^{-(s+b)}}{\Gamma(n-m+1, b)}, \quad (8)$$

incomplete gamma function

$$\Gamma(x, b) = \int_0^\infty s^{x-1} e^{-s} ds.$$

### $S_{UP}$ (忽略系统误差)

(8) 式代入(5)式得

$$\alpha = \frac{\Gamma(n-m+1, S_{UP} + b)}{\Gamma(n-m+1, b)}. \quad (9)$$

对  $m=0$ ,

(均匀先验分布)

$$\alpha = e^{-S_{UP}} \cdot \frac{\sum_{i=0}^n \frac{(S_{UP} + b)^i}{i!}}{\sum_{i=0}^n \frac{b^i}{i!}}. \quad (10)$$

数值求解(9)、(10)式得到  $S_{UP}$ 。

## (三) Bayes统计方法 (续)

### (3) 置信上限

- $S_{UP}$  (考虑系统误差)

#### (a) 考虑本底期望值的系统误差

本底期望值的分布用pdf  $f_{b'}(b, \sigma_b)$ 表示 (均值**b**, 方差  $\sigma_b^2$ ) ,

条件 pdf 式 (1)  $p(n|s)_b = e^{-(s+b)} \frac{(s+b)^n}{n!}$  需修改为

$$p(n|s)_b = e^{-(s+b)} \frac{(s+b)^n}{n!}$$

$$q(n|s)_b = \int_0^{\infty} p(n|s)_{b'} \cdot f_{b'}(b, \sigma_b) db', \quad (11)$$

式中  $p(n|s)_{b'}$  与 (1) 式形式相同, **b** 用**b'** 代替。

## (三) Bayes统计方法 (续)

### (3) 置信上限

- $S_{UP}$  (考虑系统误差)

(b) 考虑本底期望值的系统误差，

考虑信号(相对)探测效率  $\varepsilon$  的系统误差

$\varepsilon$  的分布用pdf  $f_\varepsilon(1, \sigma_\varepsilon)$  表示(均值1, 方差  $\sigma_\varepsilon^2$ ),

条件 pdf 式 (1)  $p(n|s)_b = e^{-(s+b)} \frac{(s+b)^n}{n!}$  需修改为

$$q(n|s)_b = \int_0^\infty \int_0^\infty p(n|s\varepsilon)_{b'} \cdot f_{b'}(b, \sigma_b) f_\varepsilon(1, \sigma_\varepsilon) db' d\varepsilon, \quad (12)$$

式中  $p(n|s\varepsilon)_{b'}$  与 (1) 式形式相同,  $b$  用  $b'$  代替,  $s$  用  $s\varepsilon$  代替。

利用(11,12)式的  $q(n|s)_b$  构造后验 pdf, 再用(5)式可算  $S_{UP}$  :

$$h(s|n) = \frac{q(n|s)_b \pi(s)}{\int_0^\infty q(n|s)_b \pi(s) ds}. \quad (13)$$

## (四) 程序包BPULE

计算给定置信水平的置信上限  $S_{UP}$  程序包：BPULE

朱永生：<http://psip.ihep.ac.cn/~dusx/zhuys/Bayes/>

- 考虑或不考虑系统误差。

- Prior pdf 可选择， $\pi(s) = 1/(s+b)^m$ ,  $m \in (0, 1)$ .

- 系统误差的 pdf 可选择：

高斯分布  $f_1(x) = G(1, \sigma) \equiv \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(1-x)^2}{2\sigma^2}}$ ,  $x \in (-\infty, \infty)$ ;

对数高斯分布

$$f_2(x) = LG(1, \sigma) \equiv \frac{1}{\sqrt{2\pi}\sigma_x x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma_x^2}\right],$$

$x, \sigma_x \in (0, \infty)$ ,  $\mu \in (-\infty, \infty)$ ,  
 $\mu = -\ln(1 + \sigma^2) / 2$ ,  
 $\sigma_x = \sqrt{\ln(1 + \sigma^2)}$ ;

均匀分布

$$f_3(x) = F(1, \sigma) \equiv \frac{1}{\sqrt{12}\sigma},$$

$x \in [1 - \sqrt{3}\sigma, 1 + \sqrt{3}\sigma]$ .



## (四) 程序包BPULE (续)

用Jeffreys and Flat prior 计算的  $S_{UP}$  @CL=90%的比较。

$$S_{UP}(J) < S_{UP}(F)$$

$\hat{\mu}$	$\hat{\sigma}$	$\sigma_c$	POLE		BPULE	
			Likelihood ratio	Likelihood ratio with conditioning	flat prior	Jeffreys prior
2	2	0	3.90	4.00	3.88	3.41
		0.1			3.97	3.48
		0.2	3.95	4.34	4.30	3.70
		0.3	4.10	4.75	5.37	4.30
3	2	0	5.40	5.30	4.93	4.38
		0.1			5.05	4.47
		0.2	5.70	5.65	5.49	4.77
		0.3	5.95	6.20	6.93	5.57
4	2	0	6.60	6.60	6.09	5.50
		0.1			6.25	5.61
		0.2	7.10	7.30	6.83	6.01
		0.3	7.75	7.85	8.73	7.07

## (四) 程序包BPULE (续)

$n=10$ ,  $b=5$ ,  $S_{UP}@CL=90\%$  与  $\sigma_b, \sigma_\varepsilon$  的关系。

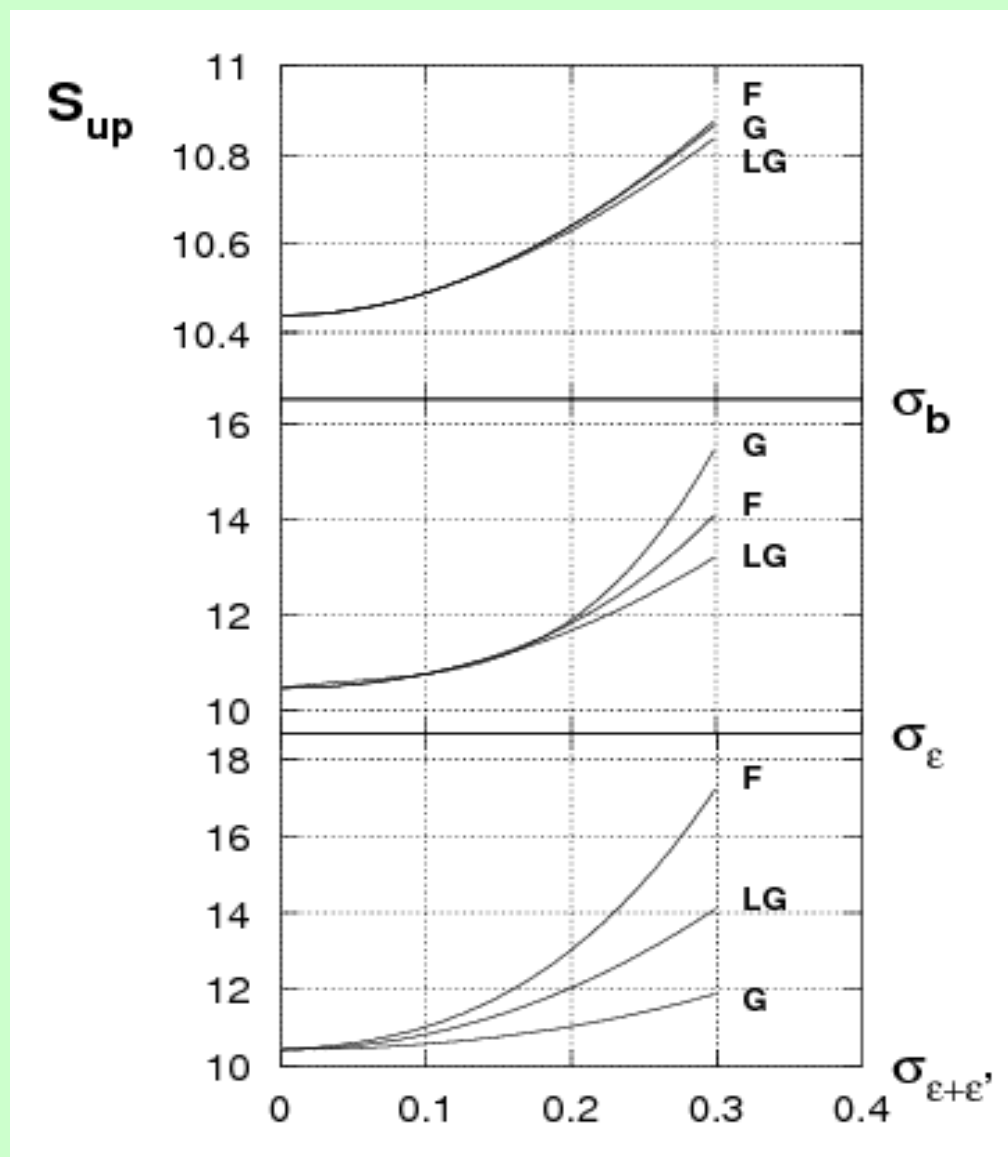
(Flat prior)

三种pdf 对  $S_{UP} - \sigma_b$

差别很小。

三种pdf 对  $S_{UP} - \sigma_\varepsilon$

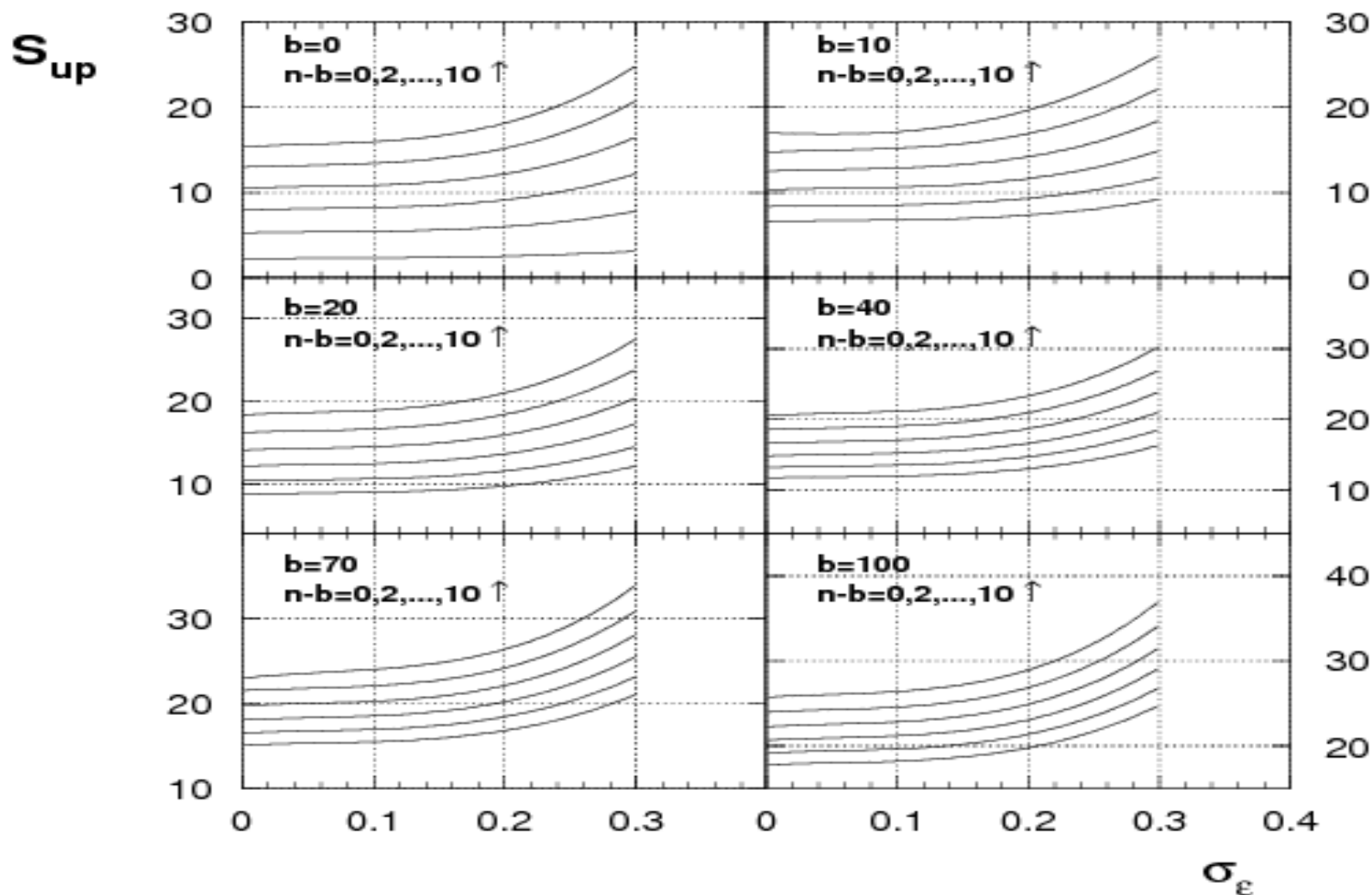
差别较明显 (对大  $\sigma_\varepsilon$ )。



## (四) 程序包BPULE (续)

$S_{UP}@CL=90\%$  与  $\sigma_\varepsilon$  的关系。(Flat prior, Gaussian pdf)

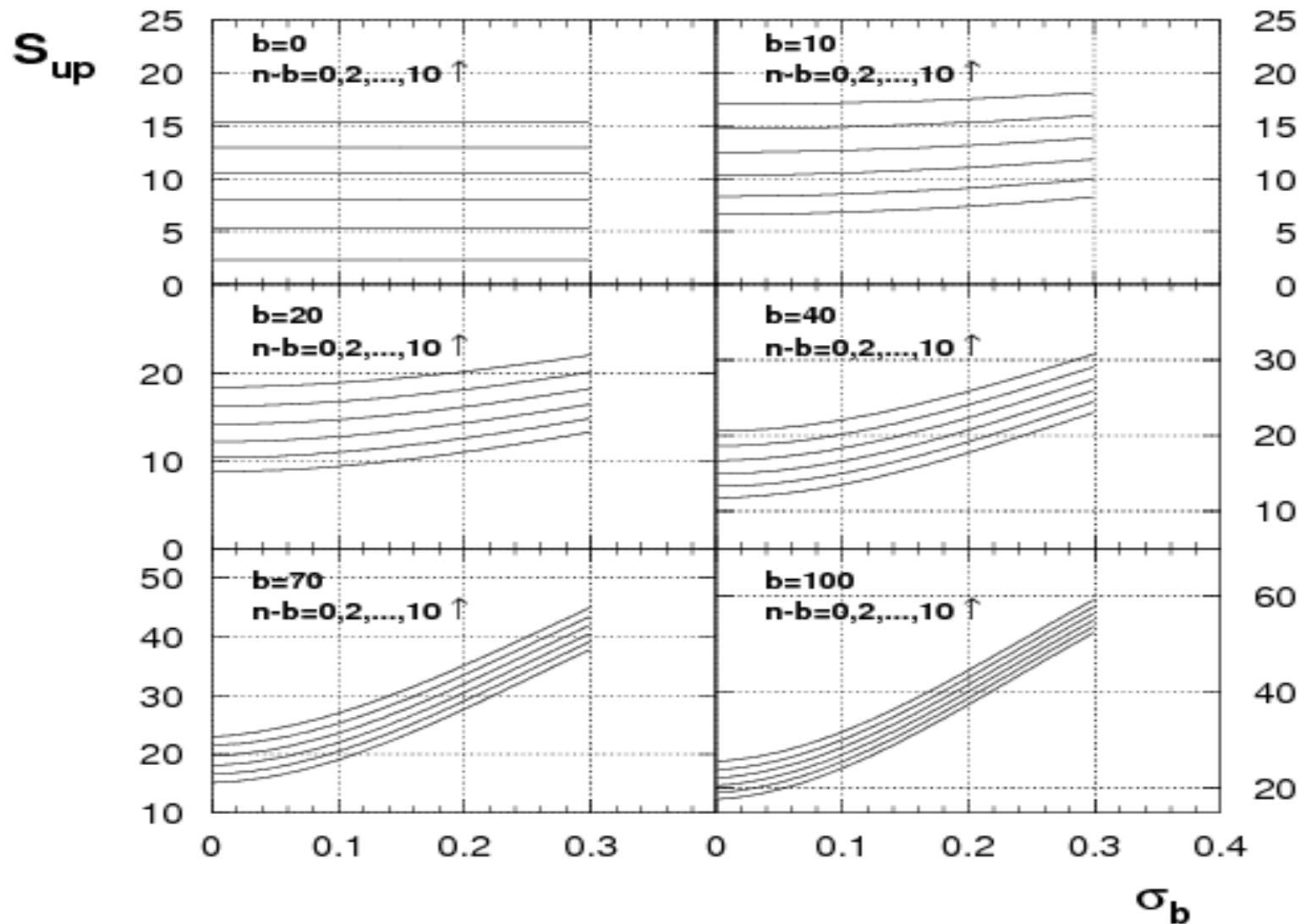
对不同的  $n$  和  $b$  值, 曲线形状类似。



## (四) 程序包BPULE (续)

$S_{UP}$  @CL=90% 与  $\sigma_b$  的关系。(Flat prior, Gaussian pdf)

b小时, 曲线平坦; b大时,  $S_{UP}$  随  $\sigma_b$  变化快。



## (五) 结束语

- 在 Flat prior pdf 情形下，Bayes 统计的置信区间或置信上限与 Neyman 方法相同。
- 在 Bayes 统计下，Jeffreys prior 计算置信区间或置信上限比用 Flat prior pdf 计算合理。
- 一般情形下，不同的统计方法（经典、Bayes 统计）计算出的相同置信水平下的置信区间或置信上限不一定相同，这并不表示它们哪个优、哪个劣，只表示它们的统计含义不相同。
- 在报道置信区间或置信上限时，应明确说明所使用的方法。



*Thanks a lot !*

**谢谢！**