第18卷 第10期　　高能物理与核物理　　Vol. 18, No. 10
1994年10月　　HIGH ENERGY PHYSICS AND NUCLEAR PHYSICS　　Oct., 1994

小样本时皮尔森 χ^2 量概率分布的讨论*

吴 枚　张春生　李惕碚　程凌翔
（中国科学院高能物理研究所　北京　100039）

1993年6月15日收到

摘　要

以高能 γ 天文数据分析中寻找周期信号的相位折迭法为例，详细研究了小样本时皮尔森 χ^2 量的概率分布，指出通常所用的皮尔森 χ^2 量分布的渐近表达式低估了大 χ^2 量事件的统计涨落概率，给出了严格的计算皮尔森 χ^2 量的概率分布式，并讨论了这一修正对超高能 γ 天文观测结果的影响。同时给出一些蒙特卡罗计算的数值表格，以便实际应用。

关键词　概率分布，皮尔森 χ^2 量，小样本，超高能 γ 天文，显著性估计。

1 引　言

在科学实验和工业生产中，人们经常使用统计量皮尔森 χ^2 量(以下简称 χ^2 量)做各种统计推断，χ^2 量的渐近分布函数称为 χ^2 分布。在高能、甚高能和超高能天体物理观测中，由于有用信号往往很微弱，观测数据必须经过统计处理，才能提炼出信息，实验工作者必须根据数据处理的方法构造相应的皮尔森 χ^2 量，并根据其值的大小来推断所得结果的置信水平。例如，为了观测脉冲星所发射的周期性的X射线或 γ 射线，实验工作者探测该脉冲星的X射线或 γ 射线的辐射，记录下每个光子的到达时间，由于信噪比太小或由于信号光子数太少，从原始记录中不能直接看出是否存在周期结构。常用的一种方法是按一个给定的周期，把原始数据折迭到一起，分该周期为 m 道，全部光子将落入各道之内，形成一个分布，称为相图。利用从相图算出的皮尔森 χ^2 量及 χ^2 分布，来估计数据中含有周期成分的概率。1983年，西德 Kiel 组[1]报道，首次观测到来自 Cyg X-3 方向的超高能 γ 射线和它的4.8小时的周期结构 (χ^2 量为45，自由度为9)，总光子数为41个，他们用 χ^2 分布估计出，如果此周期现象是由于本底的涨落造成，则其几率只有 10^{-6}。

然而，应当注意，皮尔森 χ^2 量并不精确服从通常的 χ^2 分布，只是在大样本时渐近服从 χ^2 分布。当样本较小时，皮尔森 χ^2 量的真实分布与 χ^2 分布函数有很大差异，这时如果还利用 χ^2 分布函数做统计推断，往往会高估显著性水平。本文详细研究了这一问题，讨论了天体测量的周期分析中可以近似应用通常的 χ^2 分布函数的条件。本文也给

* 国家自然科学基金资助。

出了一个精确计算皮尔森 χ^2 量概率分布的公式，并作了蒙特卡罗计算进行比较.

2 χ^2 分布函数及其在小样本情况下的近似性

设有随机变量 ζ 定义在 $[a, b]$ 区间，把 $[a, b]$ 分成 m 道，对 ζ 做 N 次观测，设 ζ 的观测值出现在第 i 道的频数为 n_i，而 ζ 的值出现在第 i 道的概率为 p_i，则统计量

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - Np_i)^2}{Np_i} \tag{1}$$

称为皮尔森 χ^2 量(以下简称 χ^2 量)，它也是一个随机变量. 而表达式

$$p(\chi^2, \nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} (\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}} \tag{2}$$

即 χ^2 分布概率密度函数，是皮尔森 χ^2 量的概率密度函数的渐近表达式. ν 称为自由度，$\nu = m - 1$.

在天体物理观测数据中搜寻周期信号是皮尔森 χ^2 量应用的一个典型例子. 设观测到来自某一天体的光子的到达时间为 $t_i (j = 1, \cdots, N)$，对给定的周期 T 做折迭，按下式可得到每个光子的相位值 φ_j

$$\varphi_j = \frac{t_j}{T} - \left[\frac{t_j}{T}\right], \tag{3}$$

其中 $[t_j/T]$ 表示 t_j/T 的整数部分. 显然 φ_j 定义在[0,1]区间，如果没有周期信号，φ_j 应均匀分布在[0,1]区间. 由(1)式可以算出折迭得到的相图的 χ^2 量. 如果观测数据中存在周期信号，则按此周期来折迭数据所得到的相图的 χ^2 量应该较大(显著大于 χ^2 量的平均值 $m - 1$). 反之，如果得到了较大的 χ^2 量，则可以用(2)式估计由于均匀本底的统计涨落出现如此大的 χ^2 量的概率，也就是说，可以估计观测数据中含有该周期信号的置信水平.

为了明确地显示出(2)式的近似性，用一个简单方法来推导(2)式，该方法的普遍性稍差，但它可明确指出，用(2)式做为 χ^2 量的密度函数时，近似在何处. 假定，观测数据中只有本底光子，没有周期信号，则(1)式中 $P_i = P = \frac{1}{m}, (i = 1, \cdots, m)$，设 $\omega_i = \frac{n_i - NP}{\sqrt{NP}}$，如果 n_i 足够大，则 ω_i 渐近服从标准正态分布，因此，随机变量 $y_i = \omega_i^2$ 的分布密度函数为

$$p(y_i) = \frac{1}{\sqrt{2\pi y_i}} e^{-\frac{y_i^2}{2}}, \tag{4}$$

其特征函数为 $(1 - 2it)^{-\frac{1}{2}}$. 所以，$\chi^2 = \sum_{i=1}^{m} y_i$ 的特征函数为 $(1 - 2it)^{-\frac{m}{2}}$，由此，易于求出 χ^2 的分布密度函数如(2)式所示. 从推导过程可知，仅仅总光子数 N 很大还不能得到(4)式，只有当总光子数 N 及每道的光子数 n_i 都足够大时，(1)式所定义的 χ^2 量才渐近服从(2)式所定义的 χ^2 密度函数.

在超高能 γ 天文中，通常 γ 光子的流强都很小，上述要求往往不能满足，如引言中提到的工作[1]的观测，总事例数 41，平均每道只有 4.1 个光子，服从很不对称的泊松分布，与正态分布相差甚远．如果用(2)式做显著性估计，可能会高估置信水平．

3　用于周期折迭的 χ^2 量的精确分布及蒙特卡罗结果

对于上述周期折迭，光子之间相互独立，每个光子落入任何一道的概率为 p_i，则第 1 道包含 n_1 个光子，……，第 m 道包含 n_m 个光子这一事件的概率为

$$p(n_1,\cdots,n_m)=\frac{N!}{n_1!\cdots n_m!}p_1^{n_1}\cdots p_m^{n_m}, \tag{5}$$

如果背景光子服从均匀分布，则有 $p_1=p_2=\cdots=p_m=\frac{1}{m}$，(5)式变为

$$p(n_1,\cdots,n_m)=\frac{N!}{n_1!\cdots n_m!}\left(\frac{1}{m}\right)^N. \tag{6}$$

对于给定的 χ^2 值，例如 $\chi^2=\chi_0^2$，可能会有多种 $(n_1,\cdots,n_m)$ 的组合满足之．因此，对于给定的 χ_0^2 出现的概率为

$$p(\chi^2=\chi_0^2)=\sum_{\chi^2=\chi_0^2}p(n_1,\cdots,n_m), \tag{7}$$

求和是对满足 $\chi_0^2=\sum_{i=1}^{m}\frac{\left(n_i-\frac{N}{m}\right)^2}{\frac{N}{m}}$ 所有可能的 $(n_1,\cdots,n_m)$ 的组合进行的．用蒙特卡罗方法验证了(7)式，做法如下：(I) 产生包含 N 个粒子的时间序列，粒子之间的到达时间之差服从指数分布．(II) 选取 10^4 个相互独立的周期做折迭．(III) 计算折迭所得相位图的 χ^2 量．(IV) 重复步骤 (I)—(III)．最后，统计不同的 χ^2 量的值出现的数目，并由此计算 χ^2 量取各种值的概率．

图 1 给出自由度 $\nu=9$，$N=30,60,100,200,400$ 时，蒙特卡罗及(2)式和(6)式以及 (7) 式的计算结果．横坐标为约化 χ^2 量 $\left(\frac{\chi^2}{\nu}\right)$，纵坐标为微分概率的对数值 $(\log_{10}P(\chi^2))$．为方便比较，从 $N=60$ 到 400，纵坐标依次下降三个量级．直方图是 $N=30$ 时，(6)式以及(7)式的计算结果．图 2 给出 $\nu=19$，$N=200$ 时的蒙特卡罗及(2)式的计算结果，较之图 1 的 $N=200$ 那条曲线，它与(2)式的差别更大，原因在于此图是针对 $m=20$ 计算的，每道的光子数更少．表 1 给出 $\chi^2>40$ 时，由(7)式与(2)式所计算的积分概率的比较．从图明显看出，蒙特卡罗模拟与(6)式加上(7)式的计算结果符合得很好，而(2)式的结果与蒙特卡罗有明显偏离，尤其当 N 较小及 χ^2 量较大时，此种偏离将变得很大．例如，$N=30,\nu=9,\chi^2>60$ 时，(2)式算出的值比蒙特卡罗的结果小 100 倍，而 30—40 个观测光子，正是超高能天文观测中常遇到的．此时如果还用(2)式做显著性检验，将会高估信号的置信水平．

以上讨论是在背景光子为均匀分布的假设下作出的．在地面进行的超高能 γ 观测往

往遇到非均匀分布的本底．此时可用(5)式加上(7)来作显著性的估计．对非均匀本底，(5)式的正确性可简单论证如下．因为任何一种连续的概率分布，都可通过一个反函数变换变成一个均匀分布[2]．而代表均匀分布的(6)式已经通过验证是正确的，所以(5)式也是正确的．在地面进行的超高能 γ 射线观测，由于观测仪器随地球转动，使得从任何一个固定天区入射的 γ 射线在不同的时间经过不同厚度的大气吸收，其本底计数是沿时间轴以一天为周期的非均匀分布．如果观测时间相当长(与折迭的周期相比)，这种非均匀性可以忽略，反之则需要认真地考虑．但是在实验工作中，为简单起见，人们往往没有考虑这种效应．对于非均匀本底直接用(5)式及(7)式计算 χ^2 量的概率很不方便．由于非均匀本底的具体形状又因实验而异，难以对各种各样的情况给出类似于表(1)那样的数值结果．所以，较方便的办法是根据本底的实测值，用蒙特卡罗直接计算各种 χ^2 值出现的概率．现以 Kiel 组的工作[1]为例，用蒙特卡罗方法来讨论小样本和非均匀性可能引起多大的误差，过程如下：(1)考虑实验的有效观测时间，以及背景光子由于大气吸收的衰减，计算出光子的记录效率，进而抽样出观测到的 41 光子的时间序列；(2)按原文中所用的 Cyg X-3 的轨道周期值和相位原点进行周期折迭，计算皮尔森 χ^2 量．重复以上步骤(改变随机数抽样种子)并统计 χ^2 量的分布．显然，以上计算同时包含了小样本和本底非均匀性的效应．结果表明，$\chi^2 > 45$ 的事件出现的概率为 5×10^{-5}，这比 Kiel 组报道的值约大 50 倍．从表 1

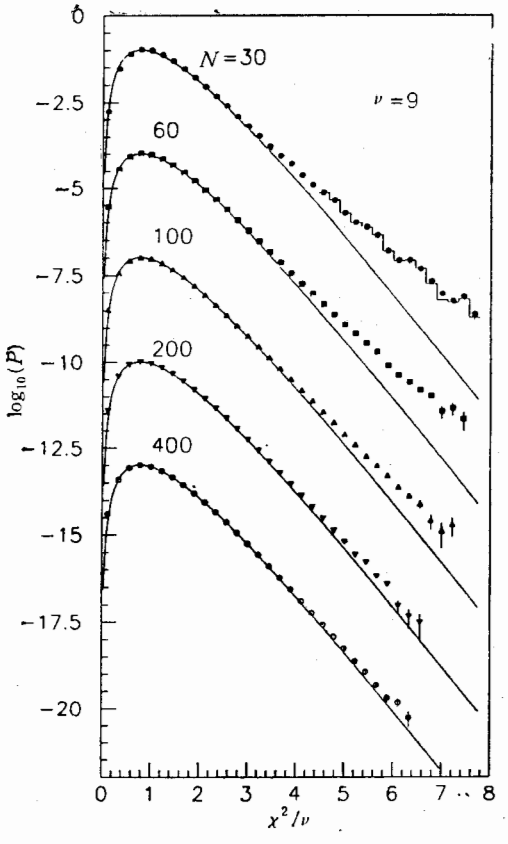

图 1 自由度 $\nu = 9$，样本数 $N = 30, 60, 100, 200, 400$ 时，蒙特卡罗及(2)式和(7)式(见正文)的计算结果

其中实曲线由(2)式算出，直方图由(7)式算出，“◆”“●”“○”“▲”“■”符号是蒙特卡罗结果。为方便比较，从 $N = 60$ 到 400，纵坐标依次下降三个量级．

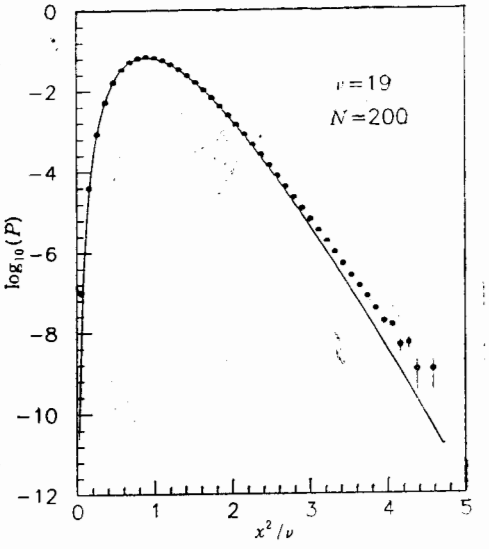

图 2 $\nu = 19, N = 200$ 时，蒙特卡罗结果(圆点)与(2)式(曲线)的比较

可以估计出 $N=41$ 时均匀分布的背景光子产生 $\chi^2>45$ 的概率约为 5.8×10^{-6}，因此，对 Kiel 组的观测来讲，由于既没有考虑本底的非均匀性，又没有考虑到小样本时不能使用(2)式，所以大大高估了结果的显著性，其中，本底的非均匀造成的影响更大。

表 1　自由度 $\nu=9$，$\chi^2>40$ 时，通常用的 χ^2 分布渐近式与真实 χ^2 分布式所计算出的积分概率的比较

χ^2 量	>40	>45	>50	>55	>60	>65
普通 χ^2 分布式	9.4×10^{-6}	1.0×10^{-6}	1.1×10^{-7}	1.1×10^{-8}	1.2×10^{-9}	1.3×10^{-10}
真实 χ^2 分布式 $N=30$	2.8×10^{-5}	7.4×10^{-6}	2.0×10^{-6}	5.1×10^{-7}	1.3×10^{-7}	3.5×10^{-8}
$N=60$	1.8×10^{-5}	3.9×10^{-6}	8.6×10^{-7}	1.9×10^{-7}	4.3×10^{-8}	9.4×10^{-9}
$N=100$	1.2×10^{-5}	2.3×10^{-6}	4.4×10^{-7}	8.4×10^{-8}	1.6×10^{-8}	3.0×10^{-9}
$N=200$	1.0×10^{-5}	1.6×10^{-6}	2.5×10^{-7}	4.0×10^{-8}	6.2×10^{-9}	9.9×10^{-10}

从本文对皮尔森 χ^2 量的讨论可以看出，应用任何一种统计量的渐近分布时，都应该注意渐近分布的近似性。尤其应用它来估计小概率事件时，更应注意，因为渐近分布往往在大概率部分与精确分布一致，而在小概率部分会呈现较大差异。

参　考　文　献

[1] M. Samorski and W. Stamm, *Astrophys. J.*, **268** (1983) L17.
[2] 李惕碚，《实验的数学处理》，科学出版社，1980.

Probability Distribution of the Pearson χ^2 of Small Samples

Wu Mei　Zhang Chunsheng　Li Tipei　Cheng Lingxiang
(*Institute of High Energy Physics, Academia Sinica Beijing* 100039)
Received on June 15, 1993

Abstract

The probability distribution of the Pearson χ^2 of small samples was investigated in detail, by use of the period analysis of high energy astrophysical observations as an example. It is pointed out that the commonly used formula of χ^2 distribution leads to underestimating fluctuation probabilities of large Pearson χ^2 events if one uses that formula as the probability distribution of the Pearson χ^2. A new formula is presented for more precise calculation of the Pearson χ^2 probability distribution and compared with extensive Monte Carlo simulation results.

Key words probability distribution, statistic Pearson χ^2, small samples, UHE gamma-ray astronomy, significance estimation.