MILAGRO MEMO # 53
Allen I. Mincer
Peter Nemethy
23 May, 1999

Some Comments on Statistics of BATSE GRB Type Searches

INTRODUCTION

The most difficult part of a statistical/probabilistic analysis seems to be asking the right question. In the following, we present one way to phrase the statistical question about the significance of an excess seen in Milagrito coincident with one of N bursts observed by another experiment. We then calculate the answer.

Definiton of the Problem

Given N GRBs observed by another experiment, each with an uncertainty region of area A_i, $i = 1, 2, ...N$, we search our data to look for an excess number of showers in bins of radius r_i and area $a_i = \pi r_i^2$ contained in A_i. Let m_i be the maximum number of showers detected in any bin when searching over all bins of area a_i for burst i. Let $P_i(b)$ be the probability that the background distribution in burst i gives a value of m_i greater than or equal to b. Let b_i^ϵ be the value of b for which $P_i(b_i^\epsilon) = \epsilon$. Let $F(\epsilon)$ be the probability that for a given ϵ, at least one GRB has background fluctuations giving $m_i > b_i^\epsilon$.

The questions we need to answer are:

1. Given P_i, and a measured least probable fluctuation (equal to the minimum of over all events i of $P_i(m_i)$) what is $F(min_i(P_i(m_i)))$?

2. What is P_i?

Calculation of $F(\epsilon)$

As stated above, $F(\epsilon)$ is the probability that for a given ϵ, at least one GRB has $m_i \geq b_i^\epsilon$. This is equal to one minus the probability that $m_i < b_i^\epsilon$ for all GRBs i. We therefore get

$$
\begin{aligned}
F(\epsilon) &= 1 - \Pi_i^N (1 - P_i(b_i^\epsilon)) && (1)\\
&= 1 - \Pi_i^N (1 - \epsilon) && (2)\\
&= 1 - (1 - \epsilon)^N && (3)\\
&\approx N\epsilon \quad for\ \epsilon \ll 1. && (4)
\end{aligned}
$$

Thus the probability that when searching N bursts we find at least one with an excess events of probability ϵ is approximately $N\epsilon$.

Discussion of result for $F(\epsilon)$

What this says is that for each burst we must calculate the significance of an excess for that burst. Doing many trials for a different burst cannot erase a significant excess for the first burst. A "trials factor" does come in, but only for the number of bursts studied. This result is independent of how P_i is defined and, in particular, is independent of the methods presented in the rest of this memo.

Calculation of P_i

The method of covering the grid with non-independent areas a_i can lead to confusion. Some approaches also clearly do not make sense in the limit: as an extreme example, if we cover the area A_i with a set of areas a_i and then displace the a_i by 10^{-10} degrees and repeat the search, we should not expect to have to pay a penalty for searching again that way. Yet counting trials factors would have us pay a factor of 2 for this additional set of "trials". We therefore develop another method of determing the probability when searching the area A_i for an excess. In this method, we can search an infinite number of ways for an excess in an area a_i but pay only the merited statistical cost.

We do this with the following analysis: Given that we have a total of S_i showers in an area A_i, we draw a circle of radius r_i with area $a_i = \pi r_i^2$ about each shower position and count the total number of showers in the circle. We then pose the statistical question, what is the probability that m_i showers will be contained in at least one circle? This manner of phrasing the question implies some approximations discussed in the next section.

The number of ways of picking the first shower is S_i. An area a_i is now defined as a circle of radius r_i centered on this event. The second shower can be picked $S_i - 1$ ways and falls in the area a_i with probability a_i/A_i. The third shower can be picked in $S_i - 2$ ways and falls in the bin with probability a_i/A_i and so on. The $m_i - 1$ showers (not counting the the one that defined the circle center) can be rearranged among themselves without altering the physical situation, so we divide by $m_i - 1!$ to avoid overcounting. The remaining $S_i - m_i$ showers will each fall outside the circle with probability $(A_i - a_i)/A_i$. The total probability $q_i(m_i, S_i)$ of getting exactly m_i showers in an area a_i is therefore given by

$$q_i(m_i, S_i) = \frac{S_i \cdot (S_i-1)\cdots(S_i-(m_i-1))}{(m_i-1)!} \cdot (a_i/A_i)^{m_i-1} \cdot ((A_i-a_i)/A_i)^{S_i-m_i} \tag{5}$$

$$= S\frac{(S_i-1)!}{(m_i-1)!((S_i-1)-(m_i-1))!} \cdot (a_i/A_i)^{m_i-1} \cdot ((A_i-a_i)/A_i)^{S_i-m_i} \tag{6}$$

$$= SC_{S_i-1}^{m_i-1} \cdot (a_i/A_i)^{m_i-1} \cdot ((A_i-a_i)/A_i)^{S_i-m_i} \tag{7}$$

where C_S^m is the number of combinations of S objects taken m at a time. The probability of getting m_i or more showers in a circle is then given by

$$P_i(m_i, S_i) = \Sigma_{j=m_i}^{\infty} q_i(j, S_i). \tag{8}$$

Discussion of result for P_i

The above merits some discussion, including a few fine points that were glossed over.

We have suggested an alternative way of finding the "significance" for an interesting event. This method is not affected by "trials factor" problem of uncertainty in statistical significance.

The calculation found the probability that at least one circle contains exactly m_i showers. Among the events fulfilling this condition, there will be some with greater than m_i points in another circle. The calculated probability is therefore an overestimate of the quantity we are interested in, namely the probability that the circle containing the most showers contains m_i of them. This difference is negligible for m_i large enough that $P_i(m_i) \ll 1$. Where it is not negligible, we would have to calculate $\Sigma_{j=1}^{\infty} q_i(m_i + j, S_i - m_i)$ and subtract it from the above result.

For circles formed from showers with center nearer to the border of the region A_i than r_i, the full area a_i is no longer available and the calculation becomes more difficult. See the example below for how one can still do an approximate calculation even when the border region is not negligible, such as in the case of interest here.

The method assumes that the expected background is uniform over the search region of interest. Note that the number used for S_i is the measured number of showers in the region, not the expected number. For $a_i \ll A_i$ the two would be identical. When this is not the case, the actual probability may be larger than that obtained using the expected background number instead of S_i. See the example in the next section for further discussion of this point.

Test of the calculation

A monte carlo has been done to test this method. Sets of 50 showers were generated with a flat x and y distribution over a region of area 10 units by 10 units. The number of showers contained in a circle of radius 1.6 units about each shower is determined, and the maximum over the 50 showers is recorded. 10^7 such sets of showers were generated. These numbers are meant to be similar to what one might expect for some search in Milagrito, but are not precisely the same as any particular case we know of.

The border region B in which the full area a_i is not available is equal to 54% of A, so this effect had to be included. Since the fraction of area a varies from 100% to 50% over the border, the calculation was done assuming $0.75a$ as the fractional area available throught the border region. The probability $q(m, S)$ was then calculated as

$$\begin{aligned} q(m, S) &= SC_{S-1}^{m-1} \cdot [((A-B)/A)(a/A)^{m-1}((A-a)/A)^{S-m} \\ &+ (B/A)(0.75a/A)^{m-1}((A-0.75a)/A)^{S-m}] \end{aligned} \quad (9)$$

where the parameters here are $A = 100$, $a = 8.04$, $B = 53.8$, and $S = 50$ as can be determined from the description of the simulation.

The simulation results and calculation are compared in Table I. The table contains the number of monte carlo events out of 10^7 for which the maximum number of showers in a circle is m as a function of m, the fraction and error this number corresponds to, and the calculated probability of getting exactly m.

The agreement for small probability is good. For large probability, one must correct for the probability of getting more events then the stated number. For example, for $m = 10$, the correction would be $q(11, 40) + q(12, 40) + \cdots q(40, 40)$ using the smaller available area. From the size of $q(11, 50)$, $q(12, 50)$, etc., it appears that this correction might explain the observed differences for large probabilities. Calculations for $m = 10$ and $m = 9$ are included to show how quickly the uncorrected calculation breaks down.

TABLE I.

Comparison of monte carlo simulated and calculated probability $q(m, 50)$ of getting a maximum number of events equal to m in a circle of radius 1.6 centered on each of 50 points thrown on a grid of size 10 by 10.

number m	number $= m$ in Monte Carlo C	fraction in Monte Carlo	$q(m, 50)$
1	0		
2	0		
3	0		
4	1		
5	6168		
6	348422		
7	1985404		
8	3165426		
9	2457742	0.24577 ± 0.00016	0.76
10	1272745	0.12728 ± 0.00011	0.29
11	517349	0.05173 ± 0.00007	0.095
12	176455	0.01765 ± 0.00004	0.029
13	52512	$(5.25 \pm 0.02) \times 10^{-3}$	7.7×10^{-3}
14	13602	$(1.36 \pm 0.01) \times 10^{-3}$	1.8×10^{-3}
15	3316	$(3.32 \pm 0.06) \times 10^{-4}$	4.1×10^{-4}
16	709	$(7.09 \pm 0.27) \times 10^{-5}$	8.4×10^{-5}
17	123	$(1.23 \pm 0.11) \times 10^{-5}$	1.5×10^{-5}
18	22	$(2.2 \pm 0.5) \times 10^{-6}$	2.6×10^{-6}
19	4	$(4 \pm 2) \times 10^{-7}$	4.0×10^{-7}
20	0	$< 2.3 \times 10^{-7}$	5.7×10^{-8}

The calculation of the total number equal to or greater than m, $P(m, 50)$ is presented in Table II. For comparison, we have also included the Poisson calculation of probability of getting m or more for an average number of 4 ($= 8.04*50/100$). Note however, that this is a somewhat different calculation $-$ in the present one we have not calculated the probability of an upward fluctuation from background giving a total of S showers, only the uniformity of the distribution given that S showers were found in A. We therfore would not expect the Poisson calculation to give a meaningful result for this case. The last column gives the ratio of $P(m, S)$ divided by the Poisson number which would be the equivalent of the number of trials factors. Naively we would used $100/8.04 = 12.4$ as the number of trials if there were no overlapping areas. The small value of this number for high m relates to the fact that the number of background events was not allowed to fluctuate hence inducing correlations among the various possible areas a.

TABLE II.

Comparison of monte carlo simulated and calculated probability $P(m, 50)$ of getting a maximum number of events equal to m or more in a circle of radius 1.6 centered on each of 50 points thrown on a grid of size 10 by 10.

num	num$\geq m$ in MC	fraction $\geq m$ in MC	$P(m, 50)$	Poisson	$\frac{(MC\geq m)}{Poisson}$
11	764092	0.07641 ± 0.00009	0.134	0.00284	26.9
12	246743	0.02467 ± 0.00005	0.039	9.15×10^{-4}	27.0
13	70288	$(7.02 \pm 0.02) \times 10^{-3}$	1.0×10^{-2}	2.74×10^{-4}	25.6
14	17776	$(1.78 \pm 0.01) \times 10^{-3}$	2.3×10^{-3}	7.63×10^{-5}	23.3
15	4174	$(4.17 \pm 0.06) \times 10^{-4}$	5.1×10^{-4}	1.99×10^{-5}	21.0
16	858	$(8.58 \pm 0.29) \times 10^{-5}$	1.0×10^{-4}	4.89×10^{-6}	17.5
17	149	$(1.49 \pm 0.12) \times 10^{-5}$	1.8×10^{-5}	1.13×10^{-6}	13.2
18	26	$(2.6 \pm 0.5) \times 10^{-6}$	3.0×10^{-6}	2.48×10^{-7}	10.5
19	4	$(4 \pm 2) \times 10^{-7}$	4.6×10^{-7}	5.16×10^{-8}	7.8
20	0	$< 2.3 \times 10^{-7}$	6.5×10^{-8}	1.02×10^{-8}	

In order to study the effect of overall number of background events fluctuations, a 10^7 event monte carlo was run where the number S of events in A was generated using a gaussian distribution with average 50 and $\sigma = \sqrt{50}$. The resulting number of monte carlo events and calculated probabilities is given in table III. What would be the "trials factor" is calculated by dividing the monte carlo probability by the calculated one. This ratio is not constant with m for the Poisson calculation since the method does not have separate uncorrelated areas. Comparison with $P(m, 50)$ shows that one may not use the above calculation in a way it was not intended, with the background average instead of the number of actual events S.

TABLE III.

Comparison of monte carlo simulated and calculated probability of getting maximum number of events equal to m or more in a circle of radius 1.6 centered on each of point thrown on a grid of size 10 by 10. The number of points is generated with a Gaussian of average 50 and σ of $\sqrt{50}$.

num	num$\geq m$ in MC	fraction $\geq m$ in MC	(MC $\geq m$)/Poisson	$P(m.50)$
11	934626	0.09346 ± 0.00010	33	0.134
12	418453	0.04185 ± 0.00006	46	0.039
13	145130	0.01451 ± 0.00004	53	0.01
14	45876	$(4.59 \pm 0.02) \times 10^{-3}$	60	2.3×10^{-3}
15	13270	$(1.33 \pm 0.01) \times 10^{-3}$	67	5.1×10^{-4}
16	3552	$(3.55 \pm 0.06) \times 10^{-4}$	73	1.0×10^{-4}
17	883	$(8.8 \pm 0.3) \times 10^{-5}$	78	1.8×10^{-5}
18	217	$(2.2 \pm 0.1) \times 10^{-5}$	89	3.0×10^{-6}
19	50	$(5.0 \pm 0.7) \times 10^{-6}$	97	4.6×10^{-7}
20	8	$(8 \pm 3) \times 10^{-7}$	80	6.5×10^{-8}

Conclusions

In combining searches of various bursts, we have shown that one accounts for the trials by multiplying the least significant probability, if it is small, by the number of bursts searched. This result is applicable for any statistically well defined significance for individual bursts.

We have presented an alternate method of defining significance which uses the total number of showers found in a region and determines the significance of clustering around a single shower. This method avoids the necessity of determining trials factors. It assumes a flat background distribution in the search area. Since this is not the case in Milagrito data, one needs to do a monte carlo to determine the actual probabilities. The results here are nonetheless useful since they allow a calculation of approximately what significance some observed excess represents.

Acknowledgments

A discussion with Charles Newman, from the Courant Institute at NYU was very helpful.