Measures of Significance in HEP and Astrophysics

James T. Linnemann
Michigan State University, E. Lansing, MI 48840, USA and
Los Alamos National Laboratory, Los Alamos, NM 87545, USA

I compare and discuss critically several measures of statistical significance in common use in astrophysics and in high energy physics. I also exhibit some relationships among them.

1. INTRODUCTION

Significance testing for a possible signal in counting experiments centers on the probability that an observed count in a signal region, or one more extreme, could have been produced solely by fluctuations of the background source(s) in that region. Statisticians refer to this probability as a p-value. The traditions for calculating signal significance differ between High Energy Physics (HEP) and High Energy Gamma Ray Astrophysics (GRA). Both fields often quote significance in terms of equivalent standard deviations of the normal distributions (statisticians sometimes refer to this as a Z-value).

I will present several of the commonly used methods in HEP and GRA, apply them to examples from the literature, then discuss the results. Here I will concentrate on observed significance, the significance of a particular observation, rather than predictions of significance for a given technique as a function of exposure. The prediction problem is slightly different, involving the power of the test, or the probability of making an observation at a given significance level.

GRA has emphasized simple, quickly-evaluated analytical formulae for calculating Z directly (choosing asymptotically normal variables), while HEP has typically calculated probabilities (p-values) and then translated into a Z-value by $p = P(s \geq \text{observed} \mid \text{assume only background})$;

$$Z = \Phi^{-1}(p); \quad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt$$

This relation can be written[1] for large $Z > 1.5$ as

$$Z \approx \sqrt{u - Ln\ u}; \quad u = -2Ln(p\sqrt{2\pi})$$

giving a rough dependence of $Z \sim \sqrt{-Ln\ p}$. While more general than the search for a simple formula for Z-values, the HEP approach loses track of the analytic structure of the problem.

Observations in GRA typically consist of a count of gamma rays when pointing directly at a potential source, called an on-source count, N_{on}. The analogous quantity in HEP is the number of counts in a signal region. The background relevant to an observation of a source is typically estimated in GRA by an off-source observation. The relative exposure of the two observations is denoted by $\alpha = T_{on}/T_{off}$, often less than unity. Then the background count mean's estimate is $b = \alpha N_{off}$, its (Poisson) uncertainty $\delta b = \alpha\sqrt{N_{off}}$, and thus one derives

$$\alpha = (\delta b)^2/b \tag{1}$$

GRA expressions are couched in terms of α. I will also use $x = N_{on},\ y = N_{off},\ k = x + y$ for compactness.

In HEP, sometimes a side-band method of background estimation is used, rather like in a GRA measurement; or b may be estimated as a sum of contributions from Monte Carlo and data-based side-band estimates, so that often $b \pm \delta b$ is quoted, where δb is derived from adding uncertainties in quadrature. One can use Eq.1 to *define* α when comparing HEP results with GRA expressions. Non-integer values for effective N_{off} result, but usually cause no problems.

2. Z-VALUE VARIABLES

Many expressions for Z are of the form of a ratio of estimates of signal to its variance, where the signal is estimated by $s = N_{on} - b = x - \alpha y$. Then $Z = s/\sqrt{V}$, where V is a variance estimate for s. A standard GRA reference[2] gives as an example (their Equation 5) $V_5 = N_{on} + \alpha^2 N_{off}$. The authors note that this expression treats N_{on} and N_{off} as independent; this does not consistently calculate V under the null hypothesis, $\mu_{on} = \alpha\mu_{off}$ and in fact biases against signals for $\alpha < 1$ by overestimating V. I have derived a related formula, $V_{5'} = \alpha(1+\alpha)N_{off}$, by using only the background to estimate the mean and variance: while not optimal, it at least is consistent with the null. They also provide $V_9 = \alpha(N_{on} + N_{off})$, which better implements the null hypothesis. However, their widely-used recommendation is likelihood ratio $L(\mu_s, \mu_b)/L(\mu_b)$,

$$Z_L = \sqrt{2}\ \left(x\ Ln\frac{x(1+\alpha)}{k\alpha} + y\ Ln\frac{y(1+\alpha)}{k} \right)^{\frac{1}{2}}.$$

Z_L derives from the standard likelihood ratio test for a composite hypothesis, and Wilks' Theorem, giving its asymptotic normal behavior. The numerator and denominator likelihoods are each separately maximized:

one for a signal + background model, the other for a background-only (null) model.

One may instead seek an asymptotically normal variable with nearly constant variance[3],

$$Z_0 = \frac{2}{\sqrt{1+\alpha}}(\sqrt{x+3/8} - \sqrt{\alpha(y+3/8)}\ ).$$

The 3/8 speeds convergence to normality from the underlying discreteness.

2.1. Other Frequentist Methods

One widely used form is $Z_{sb} = s/\sqrt{b}$ (sometimes[4] called the "signal to noise ratio"). This entirely ignores the uncertainty in the background estimate. It is often used for optimizing selection criteria, because of its simplicity. Slightly better is a Z_P calculated from the Poisson probability p-value:

$$p_P = P(\geq x|b) = \sum_{j=x}^{\infty} e^{-b}b^j/j! = \Gamma(x,0,b)/\Gamma(x).$$

here written[6] in terms of an incomplete Γ function. Z_P still ignores uncertainty in b. Occasionally one sees substitutions of $b \to b + \delta b$ as a feeble attempt to incorporate the uncertainty in b.

Finally, one may view a significance calculation directly as a p-value calculation which one could use as a test of the null hypothesis. Z_L use the standard (non-optimal) test of a composite hypothesis against a null. However, the relationship of the Poisson means, whether $\mu_{on} > \alpha\mu_{off}$, is a special case of a composite hypothesis test that admits a more optimal solution. There exists a Uniformly Most Powerful test among the class of Unbiased tests for this case, in the form of a binomial proportion test for the *ratio* of the two Poisson means[5]. The UMPU properties are, strictly speaking, derived only with an assumption of randomization, that is, hiding the underlying discreteness by adding a random number to the data. This test yields a binomial probability p-value (using $k = x + y$):

$$p_{Bi} = P_{Bi}(\geq x|\ w,k) = \sum_{j=x}^{k} \frac{k!}{j!(k-j)!} w^j (1-w)^{k-j},$$

where $w = \alpha/(1+\alpha)$ is the expected ratio of the Poisson means for x and $x+y$. After some manipulation, this can be written in terms of incomplete and complete beta functions[1, 6], which is convenient for numerical evaluation:

$$p_{Bi} = B(w,x,1+y)/B(x,1+y)$$

This test is conditional on $x + y$ fixed because of the existence of a nuisance parameter: there are two Poisson means, but the quantity of interest is their ratio. While this test is known to both the GRA[3] and HEP[7] communities, it is common practice in neither, and its optimality properties are not common knowledge.

Given the (restricted) optimality of the test, and the lack of a UMP test for this class of composite hypotheses, this test ought to be more frequently used to calculate significance, even though it is clearly a longer calculation than Z_L. For moderate $x,\ y$, closed forms in terms of special functions are available, while some care is required for larger n. For $Z_B < 3$, the Z-values reported may be somewhat too small[3, 8], but for typical applications one is more interested in $Z_B > 4$.

It is interesting to note that taking a normal approximation to the binomial test (that is, comparing the difference of binomial proportion from its expected value, to the square root of its normal-approximation variance) yields $(x/k - w)/\sqrt{w(1-w)/k}$, which can be shown to be identical to $Z_9 = s/V_9$.

A different approach attempts to move directly from likelihood to significance by using a 3rd-order expansion[9]. The mathematics is interesting, combining two first order estimates (which give significance to order $1/\sqrt{n}$) to yield a $1/\sqrt{n^3}$ result. Typically, the first-order estimates are of the form of a normal deviation, Z_t (like Z_9), and a likelihood ratio like Z_L; of these, the likelihood ratio is usually a better first-order estimate. The two are then combined into the third order estimate by a formula such as

$$Z_3 = Z_L + \frac{1}{Z_L} Ln(Z_t/Z_L).$$

Generically, $Z_t = \Delta/\sqrt{V}$ is a Student t-like variable, where Δ is the difference of the maximum likelihood value of θ (the parameter of interest) from its value under the null hypothesis, and V is a variance estimate derived from the Fisher Information $\partial^2 L/\partial^2\theta$. The attraction of the method is to achieve simple formulae with accurate results. However, the mathematics becomes more complex[10] when nuisance parameters are included, as is needed when the background is imperfectly known. Here I will only compare the approximate calculation for a perfectly known background to the corresponding exact calculation, p_P.

3. BAYESIAN METHODS

HEP common practice often involves Bayesian methods of incorporating "systematic" uncertainties for quantities such as efficiencies[11]. These methods are also used for calculating significance, particularly when the background b is a sum of several contributions, since the method naturally extends to complex situations where components of δb are correlated. The typical calculation represents the lack of knowledge of b by a posterior density function $p(b|y)$; it is referred to as a posterior density because it is posterior to the off-source measurement y. The usual way of proceeding is to calculate Poisson p-values $p_P = P(\geq x|b)$ as was done above, but this time taking into account the uncertainty in b by performing an average of p-values weighted by the Bayesian posterior $p(b|y)$, that is

$$p_{Ba} = \int p_P(\geq x|b)\ p(b|y)\ d\,b.$$

This can be evaluated by Monte Carlo integration, or by a mixture of analytical and numerical methods. I will pursue the latter course here. The most common usage in HEP is to represent $p(b|y)$ as a truncated normal distribution

$$p_N(b|y) = \frac{1}{\delta b\sqrt{2\pi}} \exp\frac{-(b-\alpha y)^2}{2(\delta b)^2},\ \ b > 0.$$

If b is a sum of many contributions, its distribution should asymptotically approach a normal. An alternative I have advocated in HEP[12], and which is also known to the GRA communitity[13], is to start from a flat prior for b and derive the $p(b|y)$ in the usual Bayesian fashion, leading to a Gamma posterior:

$$p_\Gamma(b|y) = \beta^y e^{-\beta}/y!\ ,\ \beta = b/\alpha.$$

This is most appropriate when a single contribution to b dominates and its uncertainty is actually due to counting statistics. I will refer to the Z-values which result from these two choices as Z_N for the normal posterior, and Z_Γ for the Gamma function posterior. Choosing to represent p_P as a sum, and performing the b integration first gives the p-value for the Gamma posterior[13]

$$p_\Gamma = \sum_{j=x}^{\infty} \frac{(y+j)!}{j!y!} \frac{\alpha^j}{(1+\alpha)^{1+y+j}}.$$

Despite appearances, p_Γ is identical to p_{Bi}. The Beta function representation of p_{Bi} is much more suitable for large values of $x,\ y$. The two expressions can be made somewhat closer by using $w = \alpha/(1+\alpha)$.

Bayesian practice typically focuses on direct comparison of specific hypotheses through the odds ratio. However, predictive inference[14] is commonly used in model checking (significance testing is just checking the background-only model). Predictive inference in our case is directly related to calculating $p(x|y)$, that is, averaging over the unknown parameter b.

$$p(j|y) = \int\ p(j|b)\ p(b|y)\ d\,b$$

Interestingly, some Bayesian practitioners go farther, and are willing to calculate a "Bayesian p-value"[14],

$$p_{Bayes} = \sum_{j=x}^{\infty} p(j|x)$$

which is precisely the p_{Ba} given above (there we summed before integrating).

4. COMPARISON OF RESULTS: RELATIVE PERFORMANCE

I have taken several interesting test cases from the HEP and GRA literature. The input values and Z-value calculation results are shown in Table 1. For the HEP cases, the values reported in the papers are N_{on}, b, and δb, while in the GRA case, the reported values are N_{on}, N_{off}, and α. I have also included a few artificial cases in order to sample the parameter space reasonably.

It is worth remarking that there are numerical issues to be faced in evaluation of the more complex methods. These remarks apply–at a minimum–to a Mathematica implementation. The Binomial is straightforward in its Beta function representation. The Bayes p-value methods may involve an infinite sum, and are touchy and slow for large n; [13] suggests approximating the summation by an integral. Fraser-Reid and the Bayes p-value summation results may be sensitive to whether integers are floating point values are used. An alternative attack is to leave the p_P as a Γ function ratio and trade an integration for the infinite sum. Doing so in the Bayes Gaussian case is less unstable than summing, but for large n requires hints on the location of the peak of the integrand.

For the purposes of the present section, I will take the Frequentist UMPU Binomial ratio test as a reference standard, because of its optimality properties. I will have more to say on this later.

None of these examples from the recent literature was published with a seriously wrong significance level. To me, the most striking result in the table is that the Bayes Gamma prior method produces results *identical* to the Binomial result (MSU graduate student Hyeong Kwan Kim has proven the identity).

The method most used in HEP, Bayes with a normal posterior for b, produces Z's always larger than those from Bayes Gamma. Viewing the calculation as averaging the Poisson p-value $p_P(b)$ over the posterior for b, the shorter tails of the normal compared to the gamma place less weight on the larger probabilities (smaller p-values) obtained when the off-source measurement happens to underestimate the true value of b. The difference is most striking for large values of α, that is, when the background estimate is performed with less sensitivity than the signal estimate; in this case, results differing in significance by over .5 σ can occur. The most common method in GRA, the simple Log Likelihood ratio formula, produces comparable or slightly higher estimates of significance, but seems less vulnerable to problems at large α. It appears to claim the highest significance of these methods at small n. The variance stabilization method Z_0 presented in [3] does not appear to be in general use in GRA, but produces results of similar quality to the other two mainstay methods. All methods agree for $N > 500$, where the normal approximations are good, even out to 3-6 σ tails.

The "not recommended" methods all produce results off by more than .5 σ for several low-statistics cases. Z_9, which approximates Z_{Bi}, does best; Z_5 is indeed biased against real signals compared to other measures, and its alleged improvement $Z_5\prime$, while curing that problem, overestimates significance as the price for its less efficient use of information compared to Z_9.

As expected, ignoring the uncertainty in the background estimate leads to overestimates of the significance. $s/\sqrt{b}$ is much more over-optimistic than an exact Poisson calculation, particularly for small n, or $\alpha > 1$, where the background uncertainty is most important. The best that can be said for $s/\sqrt{b}$ is that it is mostly monotonic in the true significance, at least as it is typically used (for comparing two selection criteria with N varying by an order of magnitude at most). The 3rd order Fraser-Reid approximation is fast and accurate up to moderate n, suggesting it is worth pursuing the full nuisance parameter case. However, the approximation fails for one large Z, and is very slow for the largest n.

Of the ad-hoc corrections for signal uncertainty, none are reliable; the "corrected" Poisson calculation is less biased than the un-corrected, but still widely overestimates significance for $\alpha > 1$, and can't be used for serious work. The $s/\sqrt{b + \delta b}$ isn't much better than its "un-corrected" version.

To summarize, most bad formulae overestimate significance (the only exceptions are Z_5 for $\alpha < 1$ and Poisson with $b \to b + \delta b$). Thus, prudence demands using a formula with good properties. The Binomial test seems best for simple Poisson backgrounds. For backgrounds with several components, compare Bayes MC with Γ or Normal posteriors.

5. CALIBRATION OF ABSOLUTE SIGNIFICANCE: MONTE CARLO

In the previous section, results of significance calculations were compared to a reference calculation, the UMPU Binomial Test. That method produces the lowest reported significance among the methods with a sound theoretical basis. This alone could justify its use (on grounds of conservatism)[3], but would beg the question of whether the Binomial test is actually "correct." This has been studied by Monte Carlo simulation[1] in [3].

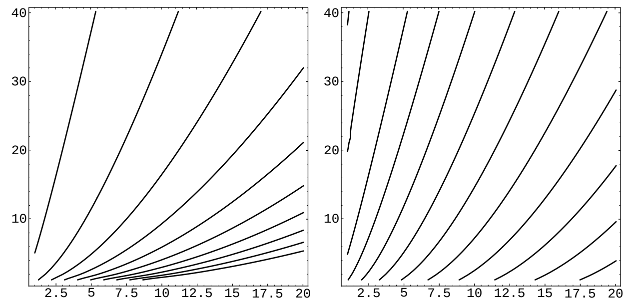

Figure 1: Contours of equal Z, case [18], for Z_{SB} (left) and Z_L (right).

A few observations on MC testing are useful. One might imagine simply generating instances of Poisson variables x, y with means μ, μ/α, and calculating Z^{MC} from p^{MC} = the fraction of events "more signal-like" than (N_{on}, N_{off}). Instead, [2, 3] a separate MC is done *for each individual measure*, because there is no unique "correct" Z-value for a given observation. The best that can be done is to ask that a method produce a Z value consistent with MC probabilities when the observation is analyzed by that method. The problem is that there is no unique definition of "more signal-like". One is essentially trying to find a unique ordering of points on the x, y plane to define those which are similarly far from the observed point N_{on}, N_{off}.

Each variable introduces its own metric, and contours of equal Z do not coincide for different Z variables, as seen in Figure 1.

The p-value for an observation (x_0, y_0) depends on these contours:

$$p^{MC}(x_0, y_0) = \int_{Z > Z_0} p(x, y)\; dx\; dy$$

where the integration is over the region beyond the contour line Z_0 passing through the observation: $Z(x, y) > Z_0(x_0, y_0)$.

For small n, the contours are markedly different, so that two *different* Z-values could both be correct if each agreed with their respective Z^{MC}. Still, the situation is not catastrophic, as values of Z are not wildly different, and presumably the Z^{MC} differ somewhat less than the reported values in Table 1. For larger n, the contours become straighter and more similar, and more importantly, the probability becomes more peaked, so that a smaller region contributes. Thus, the central limit forces convergence to a unique Z value for large n.

[1] There may have been typographical errors in the results for Z_{Bi}, identical to Z_9, but described as having different deviations from the true MC result. If the Z's were, by coincidence, identical, this might be an instance of the measure-dependence described below. Alas, the paper was published without the MC comparisons figures.

Reference	[15]	[16]	[17]	[18]	[19]	[19]	[20]	[21]	[22]	[23]	[22]	RMS
Non = x	4	6	9	17	50	67	200	523	167589	498426	2119449	
Noff = y	5	18.78	17.83	40.11	55	15	10	2327	1864910	493434	23671193	
α	0.2	0.0692	0.2132	0.0947	0.5	2.0	10.0	0.167	0.0891	1.000	0.0891	
$b = \alpha y$	1.0	1.3	3.8	3.8	27.5	30.0	100.0	388.6	166213	493434	2109732	
s = Non - b	3.0	4.7	5.2	13.2	22.5	37	100	134.4	1376	4992	9717	
δb	0.45	0.3	0.9	0.6	3.71	7.75	31.6	8.1	121.7	702.4	433.6	
δb/b	0.447	0.231	0.237	0.158	0.135	0.258	0.316	0.0207	0.000732	0.00142	0.000206	
Reported p		.0030	.027	2.0E-06								
Reported Z		2.7	1.9	4.6	3.0	3.0		5.9	3.2	5.0	6.4	
Recommended:												
Z_{Bi} Binomial	**1.66**	**2.63**	**1.82**	**4.46**	**2.93**	**2.89**	**2.20**	**5.93**	**3.23**	**5.01**	**6.40**	0
Z_Γ Bayes Gamma	**1.66**	**2.63**	**1.82**	**4.46**	**2.93**	**2.89**	**2.20**	**5.93**	*	*	*	0
Reasonable:												
Z_N Bayes Gauss (HEP)	1.88	2.71	1.94	4.55	3.08	*3.44*	*2.90*	**5.93**	**3.23**	**5.02**	**6.40**	.28
Z_0 $\sqrt{}+3/8$	1.93	2.66	1.98	4.22	3.00	3.07	2.39	5.86	**3.23**	**5.01**	**6.40**	.15
Z_L L Ratio (GRA)	1.95	2.81	1.99	4.57	3.02	3.04	2.38	**5.93**	**3.23**	**5.01**	**6.41**	.14
Not Recommended:												
$Z_9 = s/\sqrt{\alpha(N_{on}+N_{off})}$	*2.24*	*3.59*	2.17	*5.67*	3.11	**2.89**	**2.18**	6.16	**3.23**	**5.01**	**6.41**	.52
$Z_5 = s/\sqrt{N_{on}+\alpha^2 N_{off}}$	1.46	*1.90*	1.66	*3.17*	2.82	3.28	*2.89*	*5.54*	**3.22**	**5.01**	**6.40**	.93
$Z_{5'} = s/\sqrt{\alpha(1+\alpha)N_{off}}$	*2.74*	*3.99*	*2.42*	*6.47*	*3.50*	*3.90*	*3.02*	6.31	**3.23**	**5.03**	**6.41**	.53
Ignore δb:												
Z_P Poisson: ignore δb	2.08	2.84	2.14	4.87	*3.80*	*5.76*	*7.72*	*6.44*	3.37	*7.09*	6.69	1.9
Z_3 Fraser-Reid $\approx Z_P$	2.07	2.84	2.14	4.87	*3.80*	*5.76*	*(8.95)*	*6.44*	3.37	6.09	6.69	2.2
$Z_{sb} = s/\sqrt{b}$	*3.00*	*4.12*	*2.67*	*6.77*	*4.29*	*6.76*	*10.00*	*6.82*	3.38	*7.11*	6.69	2.9
Unsuccessful Hacks:												
Poisson: Nb $\to$ b + δb	1.56	2.46	1.64	**4.47**	3.04	*4.24*	*5.51*	6.01	3.07	*6.09*	**6.39**	1.1
s / $\sqrt{b+\delta b}$	*2.49*	*3.72*	*2.40*	*6.29*	*4.03*	*6.02*	*8.72*	*6.75*	3.37	*7.10*	6.69	2.4

Table I : Test Cases and Significance Results: Inputs are at top; α deduced from Eq.1 for HEP examples. The test cases are ordered in data counts; [19]; [20], and [23] have large values of α, troublesome for some methods. Z-values in **bold** are nearly equal the Binomial values; Z-values in *italics* differ by more than .5 . * indicates convergence failure. The last column gives the un-weighted RMS difference of the Z-values from to the Binomial values.

Although Monte Carlo studies can never explore the entire parameter space, the general conclusion of [3] is that Z_{Bi} is the best of the alternatives. Z_{Bi} is only slightly conservative for $Z > 3$. There, p_{Bi} is a bit larger than p^{MC} and thus $Z_{Bi} < Z^{MC}$ by 3% or less on the Z scale when $\min(N_{on}, N_{off}) < 20$, and Z_{Bi} performs even better for larger n. They found the deviations of other methods from Z^{MC} are typically larger. They also cite work[8] which finds larger fractional deviations[2] for Z_{Bi} for smaller Z. Since $Z > 3$ is the lower edge of the region where claims are liable to be made, and the degree of conservatism is small, this would also justify accepting Z_{Bi} as the reference standard, and as the recommended method of evaluating significance when there is any concern about the validity of other methods–at least when a single counting uncertainty dominates the knowledge of the background.

[2]It is not clear whether these limitations (originally studied in the purely-binomial setting) are due to discreteness; or whether the conditioning on $N_{on} + N_{off}$ causes the differences from Monte Carlo.

Acknowledgments

The author wish to thank LANL for hospitality and financial support during his sabbatical, Milagro and D0 colleagues for information and references, and Tom Loredo for reference [3]. This work was supported in part by NSF contract NSF0140106. The calculations were performed with the assistance (mostly) of *Mathematica*.

References

[1] Abramowitz & Segun, Handbook of Mathematical Functions, Dover (1968)
[2] Li & Ma, Astroph. J. 272 (1983) 314-324
[3] Zhang & Ramsden, Exp. Astro. 1 (1990) 145-163
[4] Babu & Feigelson, Astrostatistics (1996), Chapman & Hall
[5] Lehman, Testing Statistical Hypotheses, 2nd edition, Wiley (1986)
[6] Stuart & Ord, Kendall's Advanced Theory of Statistics, Vol 1 & 2
[7] James & Roos, Nuc Phys B172 (1980) 475-480
[8] D'Agostino et. al, Am. Statistician (1988) 198
[9] Fraser, "Statistical Inference: Likelihood to Significance", JASA 86 (1990) 258-65
[10] Fraser, Reid, & Wu, ftp://utstat.toronto.edu/pub/reid/research/general/196rev3.ps.Z
[11] Cousins & Highland, NIM A320 (1992) 331
[12] conferences.fnal.gov/cl2k/copies/linnemann1.pdf
[13] Alexandreas et. al., NIM A328 (1993) 570-577
[14] Gelman, Carlin, Stern, & Rubin, Bayesian Data Analysis, Chapman & Hall (1998)
[15] Artificial case suggested by an example in [13]
[16] Abe et. al., PRL 74 (1995) 2626-31: Top quark: (HEP CDF collab.; I chose one of many results)
[17] Abachi et. al., PRL 74 (1995) 2422-6: Top quark: (HEP D0 collab.)
[18] Abachi et. al., PRL 74 (1995) 2632-7: Top quark: (HEP D0 collab.)
[19] Two artificial examples from [3]
[20] An artificial example with large α
[21] icrr.u-tokyo.ac.jp/can/Symp2002/Presentations/S08-Rowell.pdf: Cyg. OB2 (GRA: Hegra collab.)
[22] Atkins, et al. (2003), Astroph. J., 595, 803-811: Crab Pulsar (GRA: Milagro collab)
[23] Reynolds et. al., Astroph. J. 404 (1993) 206-218: Crab Pulsar (GRA: Whipple collab)