NOTES ON THE STATISTICS OF SMALL COUNTS OF NUCLEAR OR PARTICLE EVENTS

Ivan V. Anicin

Faculty of Physics, University of Belgrade, Belgrade
Serbia and Montenegro
e-mail: anicin@ff.bg.ac.yu

Abstract: Reconciliation of frequentist and Bayesian approaches to elementary treatment of data in nuclear and particle physics is attempted. Unique procedure to express the significance of small count in presence of background is discussed in some detail.

It has been for more than fifty years now that nuclear and particle physicists worry about the unique way of expressing the significance of their conclusions which rely on small number of registered events of given signature, and especially so when some of these events are suspected to be of origin different than the one of current interest. This worry has in recent years been only aggravated, for new physics beyond Standard models is, if at all, necessarily represented by only a small signal immersed in the high background of the prevailing phenomena already embraced by the Standard models. From the early days of Regener (1951) [1], till PhyStat2003, physicists were exploring the rich heritage of the mathematicians in search of the final solution to the problem. Distressingly diversified body of literature on the subject has henceforth emerged, while great number of recent workshops and conferences devoted to this matter [2-6] speaks of the annoying situation. The results of all these efforts are accessible via the links found at the CDF Statistics Committee site [7,8].

Most of what follows is, of course, *deja vu*; we shall at best try to give a somewhat different interpretation of common situations with counting statistics (for as Barlow well emphasized, "everything is a counting experiment" [9]), including the important case when some phenomena other than the one under scrutiny are suspected to contribute to the overall number of indistinguishable counts. Invoking this interpretation we shall hopefully be able to reconcile the two opposed views, the frequentist and the Bayesian, at least when applied to elementary problems of counting statistics, and perhaps make life easier for a practicing nuclear and particle physicist, as well as contribute to more homogeneous presentation of our results.

To start with, we consider the inferences which we are allowed to make in the simplest of cases, when we have, in a given measurement time, counted a certain number of events of given signature, and when we either do not care about their origin or are 100% sure that they are all of the same origin, and when we believe that they satisfy the conditions required for their distribution to be Poissonian.

Consider first the case when we have counted nothing, what is, of course, how every experiment begins (and some even end). Let us recall, for future purposes, that this eventless interval is exponentially distributed, and that on the average it lasts longer when the average counting rate is lower. If we assume that the process is Poissonian, such that the average count in measurement time τ is N (or the average counting rate R is N/τ), the probability to obtain n counts in this measurement time is $P(n|N) = N^n e^{-N}/n!$ The probability to obtain zero counts is thus exponential, $P(0|N) = e^{-N}$, and this is ALL that we can possibly KNOW about N without further elaboration (we shall not consider here the alternative approach of Neyman, which is not easily compared with the Bayesian). Granting our initial assumptions were correct, this function represents the total information content of our result. In usual frequentist terminology this is called the likelihood function, and is used for parameter estimation in the manner similar to the one we shall be using here. However, it is this function that we shall try to interpret somewhat differently, more in the spirit of Bayesian attitude, hoping to bring the two confronted views closer together. In doing so we shall not imply that the average count is a stochastic variable, what certainly it is not, what otherwise constitutes one of the main objections of frequentists to the Bayesians, and prevents the former from integrating the likelihood function in order to find confidence intervals for the average count. We rather imply that it is our KNOWLEDGE of the true average count which is now specified by the above distribution. If we denote the degree of our knowledge of N with K(N) then in the case of zero recorded counts $dK=K(N)dN=e^{-N}dN$. We thus imply that this function can be integrated, as in the Bayesian tradition (where it is interpreted as a pdf), in order to quantify our current knowledge about the average count being in a certain interval. For these purposes the function even need not be normalized. Our KNOWLEDGE about the average count is in this situation thus concentrated around zero; we are 90% convinced that the average count is smaller than 2.3 $\{\int_0^{2.3} K(N)dK=0.9\}$; we are only 10% convinced that it is greater than 2.3, etc., and, quite uselessly, we KNOW for SURE that the average count has a certain value between zero and infinity. To stress that in our view this function is neither the distribution of average counts N, nor the likelihood function in the classical sense, from now on we call it the "knowledge density function", or "kdf" for short. The interval within which given percentage of our conviction about the value of the average count resides, we shall, in good frequentist tradition, still call the confidence interval and the confidence level (CL) respectively.

For a non-zero recorded count n, the kdf is $K(N) = N^n e^{-N}/n!$ This has the well known properties: our knowledge of the average count peaks at n, it gets relatively more concentrated with increasing n (the dispersion (and the mean, which is of no consequence here) of the kdf are n+1), and once any non-zero count is obtained we know for sure that the average count cannot be zero ($K(0)=0$ for $n\neq0$). Also, what is usually disregarded, this kdf is at the same time the probability density function of time intervals (since $N=\tau R$) between (n+1) scaled counts. This is good, for it allows the treatment of the results of preset time and preset count measurements on equal footing.

(A comment is perhaps appropriate here. If we were to repeat the measurement, or if the first one was of double duration, and if we were still left with the null result, the kdf of the average count will remain the same, but our knowledge of the average counting *rate* (which is always our final objection) will improve. For a non-zero result repetition of identical measurements is good for control purposes, but is in every other respect identical to a single measurement performed during the same overall measurement time.)

The degree of arbitrariness as to how to position the confidence interval, at a given confidence level, remains. This is still a matter of convention. It is perhaps only plausible to always choose the narrowest one, which necessarily contains the maximum of the kdf.

Now, all that was a purely frequentist reasoning, which had nothing to do with Bayesian interpretation of probability. To the Bayesians, however, the function which we have here dubbed "kdf" is known as the posterior density function, and is interpreted along the similar lines as we did it here. To obtain this particular function, which is to our belief the only justified kdf, the Bayesians have to use the uniform prior function. Any other prior will simply yield a wrong kdf. We have thus arrived at a heuristic proof that the two approaches are, at least in the case of elementary counting statistics, equivalent, assuming our interpretation of the likelihood function, and provided the Bayesian prior is uniform. (Formal identity of the two functions is of course well known. We just do not believe this to be a mere coincidence).

Now that we have, under the conditions stated above, hopefully agreed upon the equivalence of frequentist and Bayesian views, we may proceed with the analysis of realistic cases of small number of counts, in the presence of background, what is a typical two parameter problem most frequently met with in both the nuclear and particle physics.

To this end we shall deal in some detail with the procedure which is usually pursued by the Bayesians for the analysis of weak spectral lines situated on the continuous constant background, which may, under our terms, be adopted by the frequentists as well. The procedure is nicely presented by Sivia [10], and was recently used by Klapdor et. al. [11] to support their first ever hint of a positive result for the neutrinoless double beta decay. As we shall see it will turn out to be quite useful generally, for the definition of significance of arbitrary count in the presence of background which is known with arbitrary precision.

We assume that the spectral data consist of a certain number of channels x_i (i=1,2,...,m) with counts n_i which are distributed as $P(n_i|N_i)=(N_i)^{n_i} e^{-N_i} /n_i!$, with average values N_i which are in turn supposed to satisfy the equation $N_i(A,B)=C\{A\exp[-(x_i-x_0)^2/2w^2] + B\}$. The position x_0 and the width w of the Gaussian spectral line are supposed to be known in advance with negligible uncertainty, while the height of the line A (or its intensity) is the parameter of interest, and the height of the constant spectral background B is a nuisance parameter which must be determined only in order to find the value of A (the value of C is of no consequence here). This situation is common in high resolution gamma-ray spectroscopy, and has been as such in practically the same way analyzed in real life by Klapdor, ref.11. Assuming that the counts in the channels are independent, the whole set of data $\{n_i\}$ now has the probability to appear, or likelihood, equal to the product of probabilities for individual channel counts: $P[\{n_i\}|N_i(A,B)]=\prod_i P[n_i|N_i(A,B)]$. Now, this is at the same time equal to both our (frequentist) kdf for the parameters A and B and the (Bayesian) posterior pdf which is obtained with the constant prior function. According to the interpretation of both approaches, as agreed above, this function determines completely our knowledge about the parameters A and B in the light of the measured spectrum $\{n_i\}$.

Following the tradition of both the frequentist maximum likelihood approach to parameter estimation, and the Bayesian analysis of the posterior pdf, we shall also analyze not the kdf itself, but its logarithm, which is in this case equal to: $\ln[K(A,B)]=\ln P[\{n_i\}|N_i(A,B)]=$ const + $\sum_i \{n_i \ln[N_i(A,B)]-N_i(A,B)\}$. Our knowledge of the parameters is concentrated around the maximum of both the kdf and ln(kdf), and sharpness of both functions determines the size of the confidence intervals for the parameters. The choice of the logarithm of the kdf, instead of the kdf itself, is motivated mostly by the ease with which the limits of confidence intervals are derived in this case. As it turns out (and is seen clearly when the kdf is normal, or is approximated by the second term of Taylor expansion around the maximum) the limits of confidence intervals which correspond to confidence levels of n standard deviations for the normal distribution, are the contours of equal lnK-values which are obtained when $n^2/2$ is subtracted from the value of the ln(kdf) in its maximum, $\ln K_{max}$. Thus, the 1σ interval, or the confidence interval at the 68% confidence level, is the one which is enclosed by the iso-lnK line which is obtained when 0.5 is subtracted from $\ln K_{max}$, the 2σ, or the 95.4% confidence level interval, is the one within the iso-lnK line which is by 2 lower than the maximum, for the 3σ level it is 4.5, we are 90% convinced that the parameters are enclosed within the iso-lnK line which is 1.35 lower than the maximum lnK value, etc. The 5σ level, which is near our complete conviction that it contains the true value of the average count, and which is nowadays recommended for the results of extreme importance (like the Higgs), requires that the large confidence interval which is bound within the iso-lnK line 12.5 below the maximum, does not comprise zero. Some subtleties we shall discuss later on, in the examples.

The two parameter problem is convenient for it can be solved graphically [minding the accuracy of our statistical inferences at low statistics (or statistical errors of our statistical errors, which hardly justify more than a single significant figure in the quantification of our knowledge), the accuracy of the graphical method is quite appropriate]. The easiest way is to find numerically, and then inspect, the values of lnK(A,B) for a sufficiently fine grid of A and B values around the maximum, so as to cover the wanted confidence interval. This is best illustrated by an example.

We first analyze in some detail a typical germanium gamma-ray spectrum in the close surrounding of an expected spectral line. In the first numerical example the counts in the consecutive 11 channels are: 1,2,0,1,4,3,2,1,0,2,0. The Gaussian line of unit width, w=1, is expected in channel $x_0=6$. The spectral background is low and the line is, if there is any, weak. Upon inspecting the numbers we do not expect the background to be higher than, say, 1, and the line to be higher than 3, what would be the anticipated coordinates (A_{max},B_{max}) of the maximum of lnK. Since the line is weak we are most probably interested in the maximum confidence level at which it may be considered existent, and this is why we would like to know the values of lnK for A=0, so we include zero in the range of values for A. On the other hand, we might want to know the confidence interval for the intensity at a certain confidence level, which is why we have to extend the range of values for A well past the maximum of lnK, perhaps up to 6 in this case. Though we are not interested in the value of background B, we still have to determine the interval for its values, where we shall investigate the values of lnK. This is somewhat tricky because the two parameters, A and B, are in principle correlated, most probably in such a way that to the positive deviation of one parameter from its average there corresponds a negative deviation of the second one from its average (what is called anticorrelation). Correlation between the parameters is judged by the tilt of symmetry axes of the iso-lnK contours projected onto the A,B plane. It may thus happen, as we shall see, that the confidence limits at a given confidence level for A correspond to the values of B well away from its value at $\ln K_{max}$. This is difficult to predict, and if this graphical method is adopted, one often has to work in a number of steps, until all wanted information is obtained. We shall take this interval somewhat wider that it might appear necessary, say from zero to 2. The steps of 0.1 in both parameters are sufficiently fine, and a small program will easily find the values of [A, B, lnK(A,B)] in such a grid.

If the resulting three columns are imported into ORIGIN®, and converted to matrix form, one among the Plot3D routines will draw a contour plot of wanted iso-lnK lines projected onto the (A,B) plane. The result of this procedure for our example is presented in Fig.1.

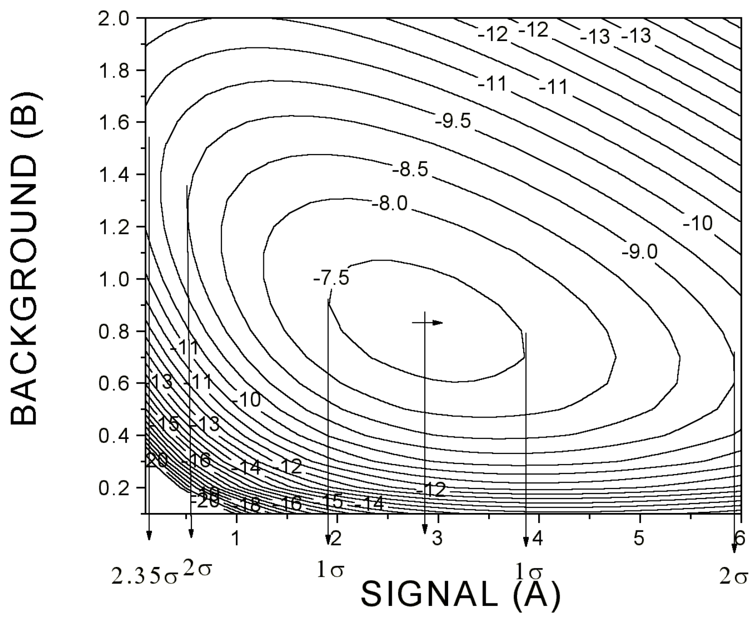

Fig.1. Analysis of the kdf for the spectrum 1,2,0,1,4,3,2,1,0,2,0, assuming a Gaussian line one channel wide in channel 6, and flat background elsewhere.

The iso-lnK lines are drawn at 0.5 intervals, starting from $\ln K_{max}$. The coordinates of the maximum of the kdf are: $A_{max} \approx 2.8$ and $B_{max} \approx 0.8$. The limits of the 1σ and 2σ intervals are denoted in the figure, and it may be seen that they correspond to positions of vertical tangents (for background these are the horizontal tangents) to the $\ln K_{max}-0.5$ and $\ln K_{max}-2$ iso-lnK lines. The aforementioned effects of (anti)correlation between A and B may also now be inspected. We may thus at the 68%CL quote the result as $A \approx 2.8^{+1.0}_{-1.0}$, or at the 95.4%CL as $A \approx 2.8^{+3.0}_{-2.4}$, etc. Alternatively, we may say, for instance, that we are 95.4% convinced that the height of the supposed line is somewhere in the interval from 0.5 to 6. Another possibility is to express the confidence level at which we are still convinced that there is a line at all. To find this confidence level we have to find the last iso-lnK level which lies completely in the region of positive A values, which is the one that touches the vertical axis. If we have determined the confidence level which corresponds to this iso-lnK line to be $(CL)_{max}$, then the confidence level (expressed in %) at which we are still convinced that there is a line at all is $(CL)_{max}+[100-(CL)_{max}]/2$. In our case the iso-lnK line which touches the vertical axis is the one which is 2.35σ below $\ln K_{max}$, so that $(CL)_{max}$ is 98%. We are thus 99% convinced that there is a line at all. Finally, at the 3σ level we may not claim the line any more, since the $\ln K_{max}-4.5$ level has the left vertical tangent well into the region of negative A values.

We see that every positive result for the signal can *always* be interpreted in two basically different ways. First, it can be expressed as a two-sided confidence interval at a given confidence level (or, what is equivalent, as a definite value with definite errors, so as to encompass the same interval). Secondly, it can be expressed, always at the higher confidence level than in the first case, as a maximum level at which we are convinced that the result may be considered non-zero (what is sometimes called the one-sided interval). Which of the possibilities we shall use depends primarily on the value of the confidence level, $(CL)_{max}$, at which the two ways of expressing the result start to differ. If the highest confidence level at which the two-sided confidence interval can be still stated is too low, smaller than one sigma for instance, then we shall rather state the maximum level at which the signal can still be considered non-zero.

When the maximum of the kdf occurs for the negative value of the signal, what we shall illustrate in the examples which follow, we can only state the maximum confidence level (which is always smaller than 50%) at which the signal can still be considered positive.

We now examine the transitional case between the spectral analysis, which we performed above, and the simple event counting case, which is of more interest in particle physics. We consider the same spectral situation as before, but with the width of the Gaussian line degenerated to a single channel (what is the infamous case of a "single-point Gaussian", which occurs at low dispersions). The width of such a "line" may be taken anything smaller than or equal to, say, w=0.1. If other channels contain only background, the problem is then equivalent to testing the hypothesis that this single count belongs to the background or not, what amounts to determination of the confidence level at which it may be considered a fluctuation of the background, or that at which it may not. Consider the following sequence of counts: 0,1,3,0,2,7,1,2,0,2,2, and suppose that there is a "single-point Gaussian" in channel 6, where the count is 7. Same procedure as above now yields the result which is presented in Fig.2.

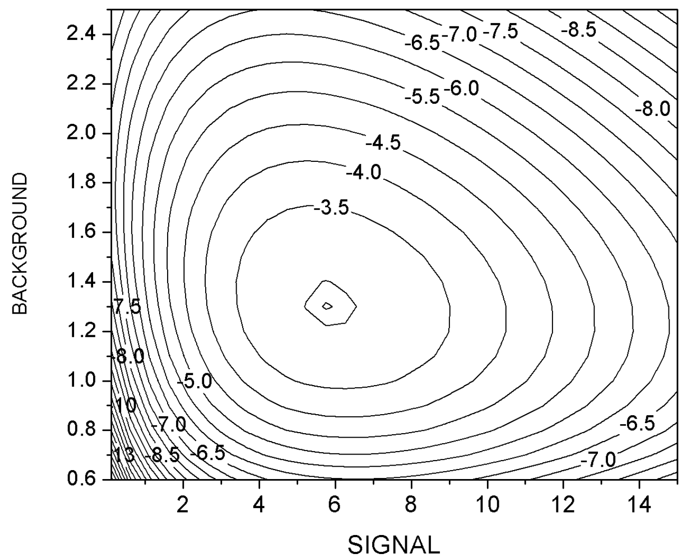

Fig.2. Analysis of the kdf for the spectrum 0,1,3,0,2,7,1,2,0,2,2 assuming a "single-point Gaussian" line 0.1 channel wide in channel 6, and flat background elsewhere.

We see that our knowledge of average background now peaks around 1.3 and the height of the "line" therefore peaks around the remaining 5.7, and that this can be considered a "line" even at the 3σ confidence level, but not higher.

Next example will take us to the case of a simple counting problem. If we now let not only the line, but also the background to degenerate to a single channel, then we are left with two counts only (one of which is supposed to be background and the other signal+background), and the problem of finding out the significance of the difference between them. Our algorithm will, without any changes, provide us with the answer. Let the two counts be 3 and 6. Let us first suppose that background is 3 and background plus signal is 6. We have to remember that all the counts are Poissonian, so that we expect the kdf for background to be inverse Poisson, while for the signal it will be something different, but with the dispersion which is the sum of dispersions of both counts. The result of the procedure which is performed under these assumptions is presented in Fig.3.

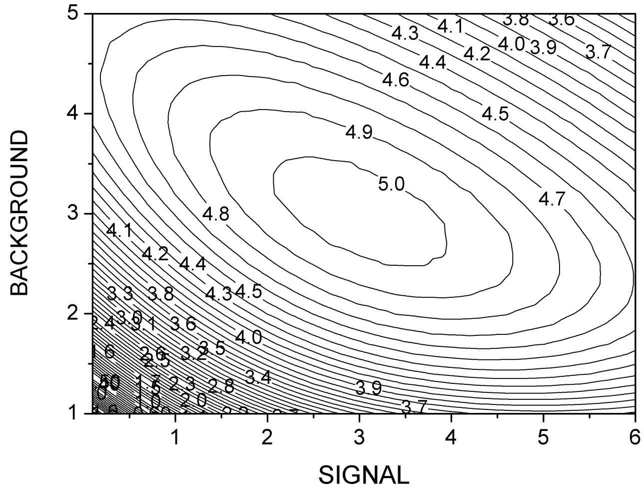

Fig.3. Analysis of the kdf for the two counts only, 3 and 6, first of which is supposed to be Poissonian background, and second signal+background. The iso-lnK lines now differ by 0.1.

We see that the kdf behaves as expected. The signal may be considered significant at the level of 1σ only, what amounts to the statement that we are at maximum 68% convinced that count 6 belongs to a different population than count 3.

Next we invert the situation. We keep the same counts as in the previous example, but exchange their roles; we now suppose that predicted background is Poissonian 6 and that signal+background turned out in the experiment to be 3. Result of the analysis is presented in Fig.4.

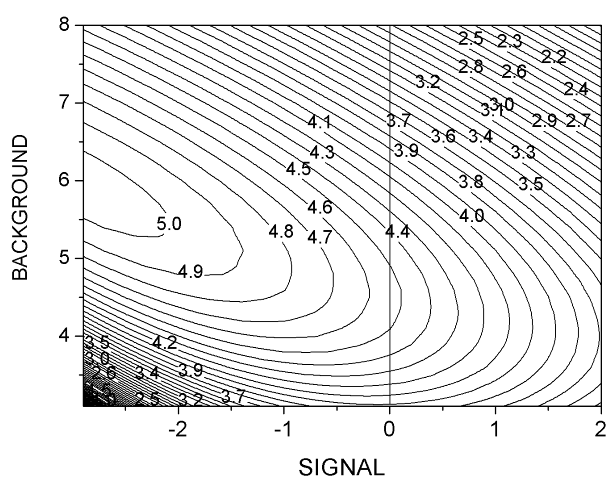

Fig.4. Analysis of the kdf for counts 3 and 6, as above, but with the roles of the counts interchanged.

We see that the signal becomes positive at the confidence level around 70%, so that we may be about 30% convinced that the signal is still non-zero.

Our algorithm is easily adapted to deal with the cases when background count in a given measurement time is in advance predicted with any given precision, either from separate measurement of the duration different than that of the actual experiment which measures signal+background, or from an appropriate Monte Carlo simulation. Since in our algorithm only the average value of counts in the channels which are declared as background determines its properties, irrespective of the actual distribution of counts in these channels, we are entitled to structure the background in such a way so as to satisfy our needs. This can be done in the following way.

If in the experiment which is supposed to measure signal+background the predicted background at the 68%CL is $n_B \pm \delta_B$, and if $\delta_B = s\sqrt{(n_B)}$ (with $s \leq 1$), then we construct the background so as to be distributed in $\eta = 1/s^2 = \mathrm{CINT}(n_B/\delta_B^2)$ channels, which contain integer counts which all add up to $\mathrm{CINT}(\eta n_B)$, where CINT denotes operation of taking the closest integer of the expression in the parentheses (we have accepted a certain inconsistency here - a symmetric confidence interval at a given confidence level for the background estimate, instead of an asymmetric one - but the difference is quite small at low CL, and absolute rigor is out of place). The way in which we shall distribute these counts is arbitrary. To illustrate how this works let us work out an example.

Suppose the predicted background at the 68%CL in a certain experiment is 3.2 ± 0.5. If it was a Poissonian prediction its dispersion would have been 3.2, and 1σ confidence interval would be $\sqrt{3.2} = 1.79$, instead of 0.5 in our case. The confidence interval is thus by the factor $1/s = 1.79/0.5 = 3.6$ narrower than it would have been if the background were determined on the basis of the measurement of the same duration as that of the actual experiment. That means that the duration of the background measurement is $\eta = 1/s^2 = \mathrm{CINT}(n_B/\delta_B^2) = 13$ times longer than that of the signal+background measurement, and that during that time $\mathrm{CINT}(\eta n_B) = 41$ background counts must have been observed (or simulated). To construct the background with such properties for the purposes of our analysis we thus have to distribute 41 counts in 13 channels, in an arbitrary way. Let us take this to be $12 \times 3 + 5$, or: 3,3,3,3,3,3,3,3,3,3,3,3,5. If in the actual experiment 6 counts have been observed, which we potentially attribute to signal+background, we then have to analyze the following "spectrum": 3,3,3,3,3,3,3,3,3,3,3,3,5,6, assuming the existence of a "single-point line" in channel 14, and considering all the rest as background. When we do this we obtain the result presented in Fig.5.

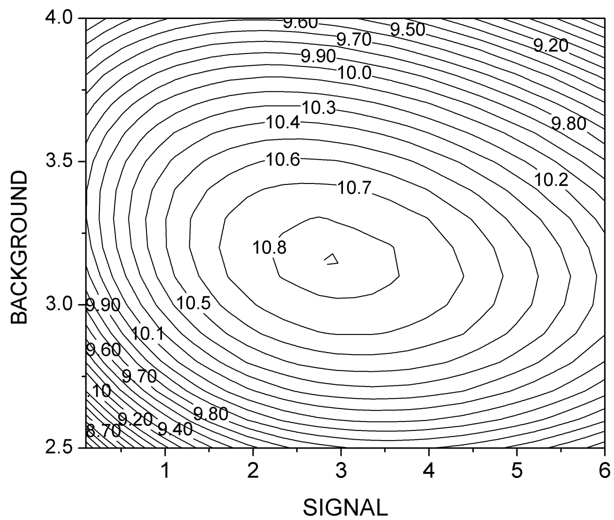

Fig.5. Analysis of the kdf for the counting experiment in which the expected background at 68%CL is 3.2(5), which is represented by the counts 3,3,3,3,3,3,3,3,3,3,3,3,5 in background channels, and the signal+background measured count is 6, which is represented by the count in a separate channel, where a single-point Gaussian line 0.1 channels wide is supposed to exist.

We see that the characteristics of the background are as required, and that increase of its accuracy over the Poissonian accuracy resulted in the increased confidence level at which the signal may be considered non-zero, as compared to our example in Fig. 3 (though perhaps not as great an increase as might have been expected).

Finally, when the counts in the channels get bigger than twenty, or maybe even ten, both the count pdf and its inverse, the kdf, get close to normal, and it becomes irrelevant which algorithm one adopts. The results of our analysis then become equal to those of the standard least squares analysis, which in principle cannot be applied when the counts are Poissonian (though at low confidence levels the LSF method, even at low counts, produces virtually the same results, at higher confidence levels, where the asymmetry of the distributions is more apparent, they start to differ significantly from those of the method applied here). To demonstrate this we apply our algorithm to the counting experiment where the expected background is Poissonian 10 and the signal+background count is 30. We expect the confidence interval for the signal at the 68%CL to be $2\sqrt{(30+10)}$ wide, and in Fig.6, where this situation is analyzed assuming the existence of the single-point line in the channel which contains signal+background, we see that it is indeed so.

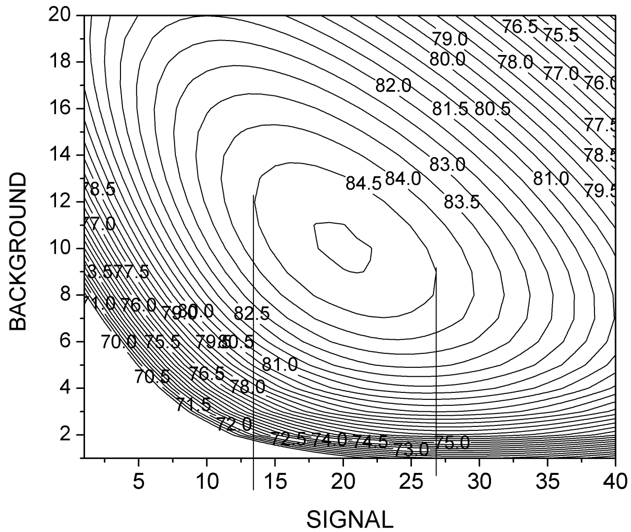

Fig.6. Analysis of the kdf for the two counts, 10 and 30, first of which is supposed to be Poissonian background, and second signal+background. The iso-lnK lines differ by 0.5.

To conclude, we believe that the examples, which were chosen to be representative of elementary counting experiments in nuclear and particle physics, demonstrate that the significance of results from these experiments may be meaningfully assessed by always using one and the same algorithm, which we have here elaborated and, hopefully, justified.

This work is partly supported by the Serbian Ministry of Science and Environment, under Projects No.1451 and No.1859.

References:

1. V.H.Regener: "Statistical Significance of Small Samples of Cosmic-Ray Counts", Phys. Rev. **84**(1951)161
2. F.James, L.Lyons, Y.Perrin, Eds.: “Workshop on Confidence Limits”, CERN 17-18 January 2000, CERN Yellow Report 2000-005, Geneva 2000, available on-line at: http://ph-dep.web.cern.ch/ph-dep/Events/CLW/papers.html
3. “Workshop on Confidence Limits”, Fermilab, 27-28 March 2000 http://conferences.fnal.gov/cl2k/fredjames_lectures.pdf
4. “Advanced Statistical Techniques in Particle Physics” Conference, Durham 2002, http://www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings.shtml
5. “Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology”, PhyStat2003, Stanford 2003 http://www.slac.stanford.edu/econf/C030908/proceedings.html

6. PHYSTAT "Workshop on Statistical Software for Physics and Astronomy", March 1-2, 2004 Michigan State University http://user.pa.msu.edu/linnemann/public/workshop/program.htm
7. http://www-cdf.fnal.gov/physics/statistics/statistics_home.html
8. http://www.nevis.columbia.edu/~evans/references/statistics.html
9. R. Barlow, Introduction to Statistical Issues in Particle Physics, Presented at PHYSTAT 2003, ref.5
10. D.S.Sivia: "Data Analysis – A Bayesian Tutorial", Clarendon, Oxford 1996
11. H.V. Klapdor-Kleingrothaus, A. Dietz, H.V. Harney and I.V.Krivosheina, "Evidence for Neutrinoless Double Beta Decay" Modern Physics Letters **A 16**, No. 37 (2001) 2409-2420,