An alternative view of the Look Elsewhere Effect

G. Ranucci
Istituto Nazionale di Fisica Nucleare, 20133 Milano, Italy

Abstract
The Look Elsewhere Effect, which influences the significance of a potential signal of a particle with *a priori* unknown mass, has a striking counterpart in searches for the presence of a modulation of unknown frequency in a time series of experimental data. In this work, the formulation of the problem in the frequency domain is outlined and the methodology is illustrated using the time series data of the Super-Kamiokande solar neutrino experiment. The parallelism between the mass and frequency domains will be highlighted.

1 Introduction

It has become common practice in high energy physics to denote as the Look Elsewhere Effect (LEE) [1] the statistical implications, especially for significance, of the search for a new signal when a parameter such as a particle mass is unknown. The search for new signals is at the core of the quest for new physics, especially at the dawn of the exciting LHC era. Where to look for new signals is also an aspect of the searches themselves. In this case, the standard statistical treatment for discovery needs to be suitably modified to account for the multiplicity, in term of possible locations, inherent in a given search. In particular, what is decisively affected is the statistical significance of a claimed detection, which changes drastically from the situation in which the mass of the putative particle is known. In this standard case, assuming the background to be reliably estimated, via auxiliary measurements or through detailed Monte Carlo modeling, the usual statistical procedure of comparing the number of counts with the expected background (typically based on the Poisson distribution for counting experiments) holds.

On the other hand, the *a priori* unknown location of the signal complicates this simple picture, because it effectively enhances the background. In effect, the fixed mass background is replaced with some kind of extreme value distribution in which the background process populates the whole search range. The statistical description of the Look Elsewhere Effect encompasses the mathematical tools and procedures needed to describe and quantify numerically such an occurrence. However, it is not frequently appreciated that the mathematics underling all the aspects of this phenomenon has an immediate correspondence in the methodologies that are exploited in the frequency domain while searching for a modulation of unknown period possibly embedded in a noisy data time series. I fully exploit this analogy here to provide an alternative view of the LEE in the frequency domain, not only specifically unraveling its peculiarities and consequences in the time series scanning and analysis, but also underlining the correspondence and parallelism between the description of the effect in the frequency domain and in the usual context we are interested in of a putative signal in a unknown mass range.

2 General framework for this illustration of the LEE: search for modulations embedded in experimental time series

For the purpose of the present discussion it is enough to recall that a popular method to unravel the presence of possible modulations hidden in noisy time series is the computation of the power density spectrum in the frequency domain of the data under study, which are regarded as originated from the sampling of a random process. By denoting with $H(f)$ the Fourier transform of the series, the corresponding power spectrum is simply defined as $|H(f)|^2$. Periodicities hidden in the data would produce sharp, distinct peaks in the power spectrum, which are in principle easily identified. The calculation of the spectrum depends on the nature of the sampling, which may be regular, when the data points

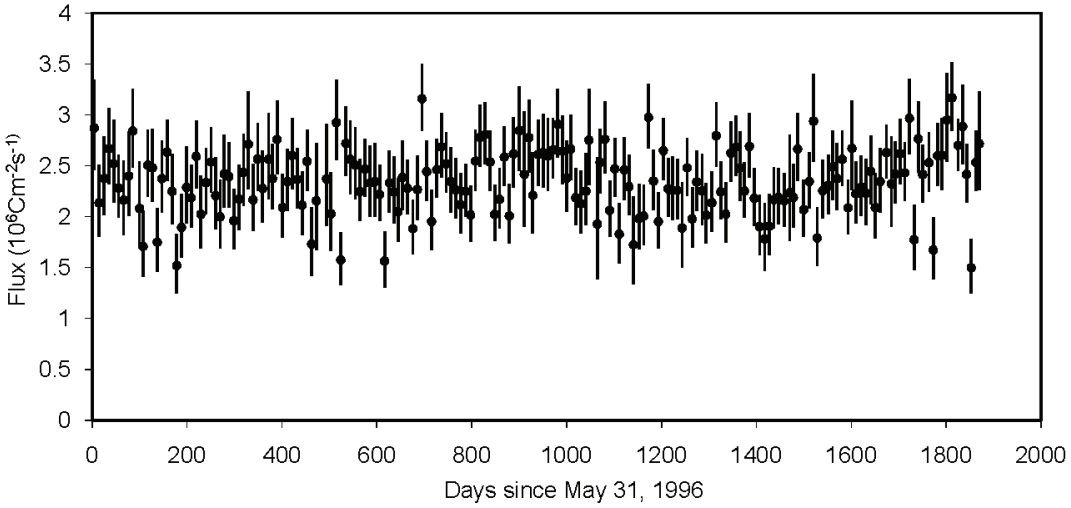

Fig. 1: Time series data of the solar neutrino measurements released by Super-Kamiokande.

are spaced at regular intervals, or as in most practical cases, irregular, i.e. with an unevenly sampled time series. The spectrum implementation in the latter occurrence is performed through the very popular Lomb-Scargle methodology, which produces a power spectrum commonly termed the Lomb-Scargle periodogram [2] [3] (the use of the word periodogram reflects the major emphasis that in this kind of searches is given to the period rather than to the frequency of the searched modulation, especially in astronomical studies of variable phenomena). The Lomb-Scargle periodogram can be viewed as a generalization of the power spectrum estimated via the direct application of the Fourier transform to evenly sampled time series, sometime called Schuster periodogram [3][4][5], not frequently used in practice, but extremely useful to understand the basic statistical features of the spectrum of an experimental time series.

3 A concrete example

To illustrate the method and its statistical implications I use the analysis of the time series data of the Super-Kamiokande solar neutrino experiment [6]. The data officially released by the Collaboration, spanning a 5 year range from April 1996 to July 2001, the so called phase 1 of the experiment, were packed in 10 and 5 day bin format, but I consider here only the 10 day bin series, which for reference is shown in Fig. 1. The analysis of all the released datasets can be found in Ref. [7].

The noise affecting the data can mask a potential low amplitude modulation embedded in the series. The purpose of the frequency analysis is to identify a potential sinusoidal signal, despite the blurring effect of the noise itself.

How a Fourier-related algorithm maps the time series in Fig. 1 into the frequency domain is shown in Fig. 2, specifically illustrating the power spectrum of the series produced by the Lomb-Scargle algorithm (the series indeed is unevenly sampled). The spectrum, naively, is simple to interpret: an unusual high peak should indicate a modulation at that frequency embedded in the data. However, the noise surely present in the time domain has an impact in the frequency domain as well, producing random noise peaks. Therefore, given a high peak in the spectrum, the claim of the detection of a modulation signal at the corresponding frequency has to be confronted with the probability that we are actually dealing with a noise fluctuation. In this context, the Look Elsewhere Effect comes into play because of the a-priori unknown frequency of the sought modulation, therefore the significance of the detection

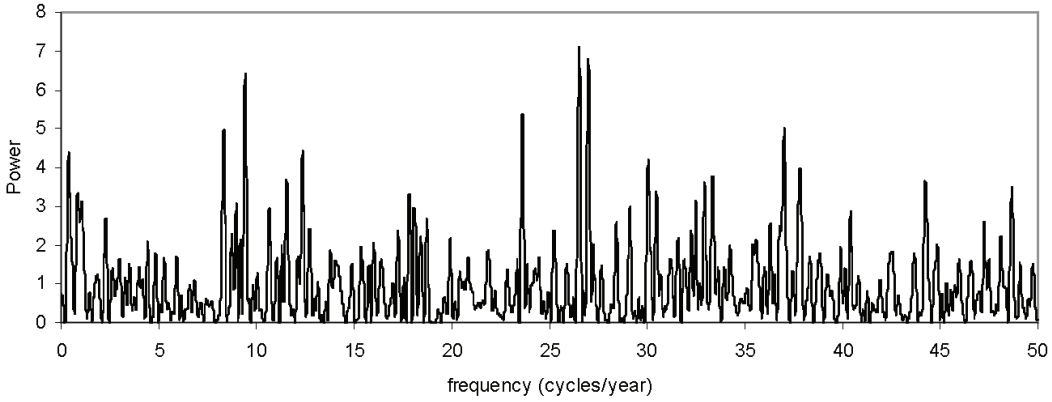

Fig. 2: Lomb-Scargle spectrum of the Super-Kamiokande solar neutrino time series.

is expressed by the probability that a peak as high or higher than the highest peak found in the actual spectrum, can be generated by chance noise fluctuations *at any of the scanned frequencies*. How this can be computed is the subject of the next paragraphs.

4 Formulation and statistical properties of the periodogram

For the sake of brevity the complete expressions of the Lomb-Scargle and Schuster periodograms are not given here, rather the interest reader can refer to the explicit formulation in [2][3][8]. A fundamental statistical property valid for both periodograms is that under the null hypothesis and for gaussian noise affecting the data, the ordinate z of the spectrum at a generic frequency is distributed simply according to e^{-z}, the absence of extra factors in this rather simple noise distribution being due to a normalization term in the periodogram expression containing σ^2, the data variance. It should be highlighted that σ^2 is inferred from the scatter of the data themselves, therefore the errors on the individual data points, even if available, are not taken into account in the standard periodogram framework (a likelihood ratio approach to the spectrum computation, not considered here, would overcome this limitation). Before moving on to the implications of the LEE effect, letÕs consider the search of a signal at a predefined frequency, which occurs when there are reasons to presume the existence of some effect at that specific frequency, with the Look Elsewhere Effect turned off. In this case the ingredients of the detection problem are the usual ones, e.g. the distribution of the ordinate at that frequency in case of absence of the signal (null hypothesis) and the ordinate distribution for the same frequency in case of presence of a signal (alternative hypothesis). The former is simply e^{-z}, while it can be shown that the latter belongs to the family of the non central χ^2 distributions with two degrees of freedom, but is, however, completely defined only if also the presumed amplitude of the sought signal is known. The detection scenario is shown in Fig. 3, where it is possible to distinguish the single frequency exponential noise distribution from three examples of signal distributions for three different signal amplitudes a, the relative amplitude with respect to the average value of the series. Given a threshold, e.g. the straight vertical line in the figure, the significance level of a detection is unambiguously defined as the integral ($p - content$) of the simple exponential noise distribution above it.

5 The Look Elsewhere Effect in action

Now I modify the previous framework moving to a situation in which, lacking the knowledge of where to expect the signal, a whole frequency range is spanned for the search. The usual criterion to decide about the presence of a signal is based upon the height of the largest peak detected in the searched frequency interval, which acts as the test statistics.

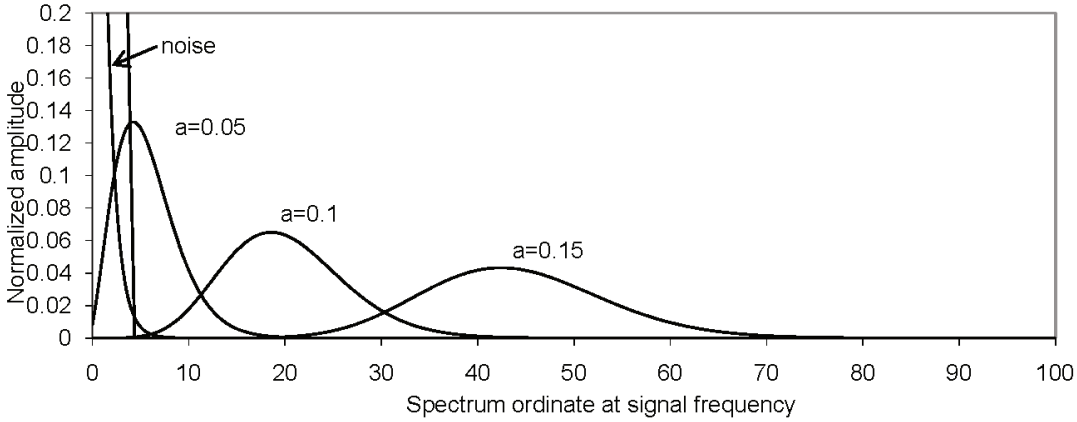

Fig. 3: Single frequency noise distribution together with three different amplitude signal distributions.

The LEE implies the need to modify the detection scenario depicted in Fig. 3 by replacing the single frequency exponential noise PDF with the PDF of the highest peak generated over the entire search band by a pure noisy series (null hypothesis).

The paradigmatic Schuster periodogram is very useful to describe the derivation of this PDF, by exploiting the important frequency analysis result that the direct application of the DFT (Discrete Fourier Transform) to an evenly sampled series with N number of samples generates a spectrum, i.e. the Schuster periodogram, meaningfully computed up to the Nyquist frequency $1/(2T)$, where T is the sampling interval, and which comprises only $N/2$ independent frequencies [5]. Therefore, the spectrum is made of $M = N/2$ independent ordinates each distributed according to e^{-z} under the null hypothesis.

As a consequence, the desired PDF of the height z of the largest among M peaks is given by

$$M\left(1-e^{-z}\right)^{M-1}e^{-z}, \tag{1}$$

where the third factor represents the highest peak, the second factor the occurrence that the remaining $M-1$ peaks do not exceed z and the first factor M the number of possible choices of the highest peak among the M spectral ordinates.

Hence, if we denote with H the ordinate of the largest peak actually detected in the experimental spectrum, the probability to have a chance noise fluctuation as high or higher than H is

$$\int_H^\infty M\left(1-e^{-z}\right)^{M-1}e^{-z}\,dz = 1-\left(1-e^{-H}\right)^M, \tag{2}$$

which can be immediately recognized as a Šidàk-Bonferroni-type [9] formula, typical of Multiple Hypothesis Testing problems.

Following the standard terminology, the above formula gives the $p-value$ of the highest detected peak. If, however, the threshold th for the detection is chosen in advance, then the same formula, with H replaced by th, is more precisely interpreted as the significance (or equivalently the *type I error*) of the detection procedure.

It should be highlighted that this formalism is extendable to peaks of any rank, exploiting the explicit formulation of the PDF of a generic peak of rank i, given by [7]

$$p_i(z|M) = \frac{M!}{(i-1)!(M-i)!}[1-F(z)]^{M-i}\,[F(z)]^{-1}\,p(z), \tag{3}$$

where i is the order of the peak, $i = 1$ being the lowest peak and so on, up to $i = M$, i.e. the highest peak, and

$$F(z) = \int_0^\infty p(\lambda)\, d\lambda. \tag{4}$$

Expression (3) is valid for any form of $p(z)$ and not only for $p(z) = e^{-z}$ as in the problem under study. Furthermore, it is easily verified that Eq. (3) reduces to the Eq. (4) for $i = M$.

6 Application to the Super-Kamiokande time series

The validation of the model in Section 5 and its application to real data time series proceed usually via Monte Carlo (MC) techniques. Indeed, the previous formalism is valid not only for the reference Schuster periodogram, but also in the cases of real practical interest of unevenly sampled data when the Lomb-Scargle periodogram is used. Nevertheless, in such an occurrence the M parameter in the formulae is not derivable by the number of sampling points, but rather can be inferred via simulation. A thorough account of how the MC methodology is applied, via toy models examples, is given in [7]. Here the Monte Carlo estimation is directly employed for the Lomb-Scargle analysis of the measured Super-Kamiokande series; the final goal is to perform a quantitatively statistical inference about the possible presence of a modulation, through the assessment of the p-values of the largest peak(s) in the spectrum.

In general, the Monte Carlo approach envisages the simulation of many synthetic time series generated reproducing the characteristics of the specific time series under study, in particular the data scatter (i.e. the variance). For each simulated series the Lomb-Scargle periodogram is computed and the height of the four highest peaks recorded. At the end of the simulation cycles (10000) the resulting histograms are compared with the respective PDF from Eq. (3).

The MC evaluated distributions of the four highest peaks in a search frequency interval ranging from 0 to 50 cycles/year are shown in Fig. 4. They follow very well the model for a parameter M equal to 529, that in analogy with the paradigmatic Schuster periodogram we can identify as the ÒeffectiveÓ number of independently scanned frequencies in this specific case (see [10] for a clear definition of this concept). Remarkably, not only the highest peak, but also the others are in fairly good agreement with the model.

The MC inferred distributions just obtained are the basis to assess the significance of the signal search: indeed their integral above the height of the peak of corresponding rank (the highest, the second largest, the third largest peak, and so on; the example here is limited to the first four largest peaks) in the spectrum in Fig. 2 gives the desired p-values. The numerical values of the ordinates of the four peaks of interest are such that these integrals can be conveniently evaluated both directly by the MC distributions or through the corresponding model with the M parameter value, 529, inferred from the fit. Anyhow, the latter method would be the more convenient in the occurrence of a very high peak, since in that case in order to get a meaningful p-value from the MC distribution one should run much more than the 10000 simulations used here. The resulting p-values are summarized in the following Table 1; collectively they demonstrate that there is no hint whatsoever of a signal embedded in the series, which thus appears to be perfectly compatible with a pure noise series.

Table 1: Significance of the 4 highest peaks in the spectrum of the Super-Kamiokande time series.

Rank	Frequency (Hz)	Ordinate	Significance
1^{st}	26.51	7.1	34.7%
2^{nd}	26.99	6.8	15%
3^{rd}	9.4	6.41	8%
4^{th}	23.6	5.36	27%

The proper account of the Look Elsewhere Effect via the described MC procedure is essential

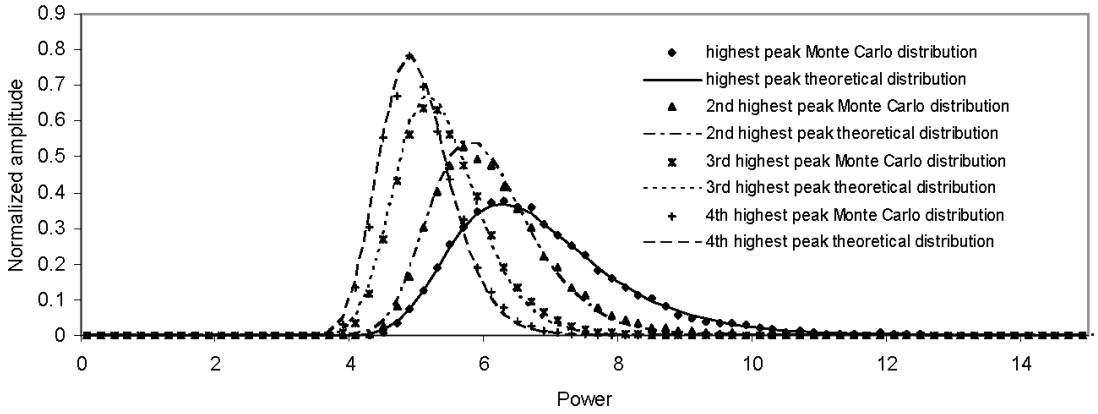

Fig. 4: Comparison of model and Monte Carlo concerning the distributions of the four highest peaks in the Lomb-Scargle periodogram of the Super-Kamiokande data series.

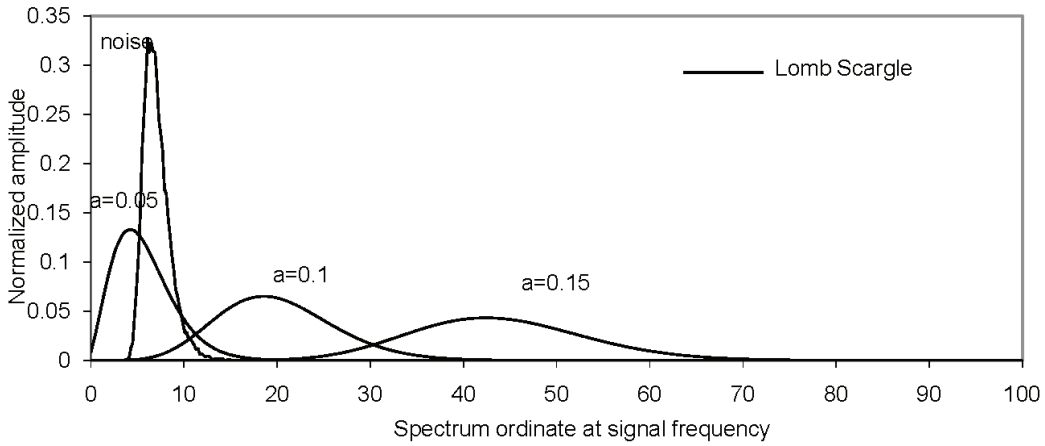

Fig. 5: Effective noise distribution generated by the LEE together with three different amplitude signal distributions.

to reach such a conclusion: for example for the highest 7.1 peak ordinate of the spectrum the p-value for a specific-frequency signal detection from the e^{-z} noise distribution would be 0.08251%, leading to a suspect that something is going on at that frequency. The incorporation of the LEE washes out completely this potential erroneous interpretation.

The overall picture of the impact of the Look Elsewhere Effect can be appreciated in Fig. 5, which reproduced the detection scenario of Fig. 3, but with the single frequency noise distribution now replaced by the effective noise distribution. i.e. the PDF of the highest peak over the search band. We can consider it as the effective noise distribution since it is such PDF that in practice limits the capability to detect small periodic signals embedded in the series. This fact is especially clear in the figure, by noting that the PDF of the highest peak overlaps almost completely the signal induced distribution when the signal amplitude a is small (5% in the figure), hence severely hindering its detection. A detailed description of the methodology and results can be found in Ref. [7], and a similar analysis for the time series data of the solar neutrino experiment SNO is reported in Ref. [11].

7 Parallelism with the search for a unknown mass particle standard problem

The procedure illustrated in the previous Sections to address the LEE in the frequency domain can be easily adapted to describe the LEE effect in the more usual scenario of search for a particle when its mass is unknown. To illustrate this I depict a simple prototype example, described by a toy model consisting of an experimental mass range from 0 to 100, affected by a Poisson background distributed with mean value 500, uniform over the entire mass range. The range is explored for a signal through a set of windows covering the whole interval.

The very basic case is that of a set of W non overlapped, contiguous windows of equal width, depending upon the presumed resolution of the sought signal; in each of them the noise is independently Poisson distributed with mean value $B = 500/W$.

The identification of a signal is denoted by a "significantÓ excess of events above background in any of the search windows. But, how to compute the significance? It is just in this calculation that the parallelism with the frequency search problem appears, establishing a correspondence between the number of explored windows W and the number of independent scanned frequencies M, as shown in the following.

Given the Poisson count rate in each window,

$$p(n|B) = \frac{e^{-B}B^n}{n!}, \tag{5}$$

what limits the detection of a signal is actually the distribution of the largest detected count N over the whole set of searched windows, that stems from that Poisson process. Such a distribution can be inferred by considering all the configurations that produce each specific realization of the random variable N. In particular, a given value N is obtained when in all the windows but one the counts are less than N, while in the residual window the count is exactly N. The probability associated with such a configuration is clearly

$$W \left(\sum_{n=0}^{N-1} \frac{e^{-B}B^n}{n!} \right)^{W-1} \frac{e^{-B}B^N}{N!}, \tag{6}$$

where the factor W takes into account the number of combinations of one window out of the total number W. The parallelism with the frequency case is clearly manifest in Eq. (6), whose three terms are in perfect one to one correspondence with the three factors in Eq. (1). However, in the discrete case there are additional configurations that must be accounted for explicitly. How they modify expression Eq. (6) is not reported here, but a complete derivation can be found in Ref. [12], leading to

$$P_{\max}(N) = \sum_{k=1}^{W} \binom{W}{k} \left(\sum_{n=0}^{N-1} \frac{e^{-B}B^n}{n!} \right)^{W-k} \left(\frac{e^{-B}B^N}{N!} \right)^{k}. \tag{7}$$

Equation (7) is the complete transposition in this context of the Eq. (1).

A graphical representation of the toy model described above shows clearly the implication of this formula. Considering 25 non overlapping windows of width 4, Fig. 6 displays both the individual window background probability function, i.e. the simple Poisson distribution, and the probability function of the highest peak of the 25 scanned windows. As in the frequency case, it is the latter that acts as the effective noise distribution affecting the capability to detect very low level signals, and from which significance and p-values calculations can be inferred. It shows the striking effect of the LEE in action. As in the frequency case, the problem can also be fully addressed via MC, through which, in particular, it has been possible to cross check very accurately the validity of the Eq. (7). Incidentally, we note that this formulation describes also the statistics of the highest bin in a finely binned histogram (under the assumption of constant background). In a histogram like this when one of the bins is anomalously high its $p - value$ is usually given as the single bin $p - value$, from the Poisson function, multiplied by the

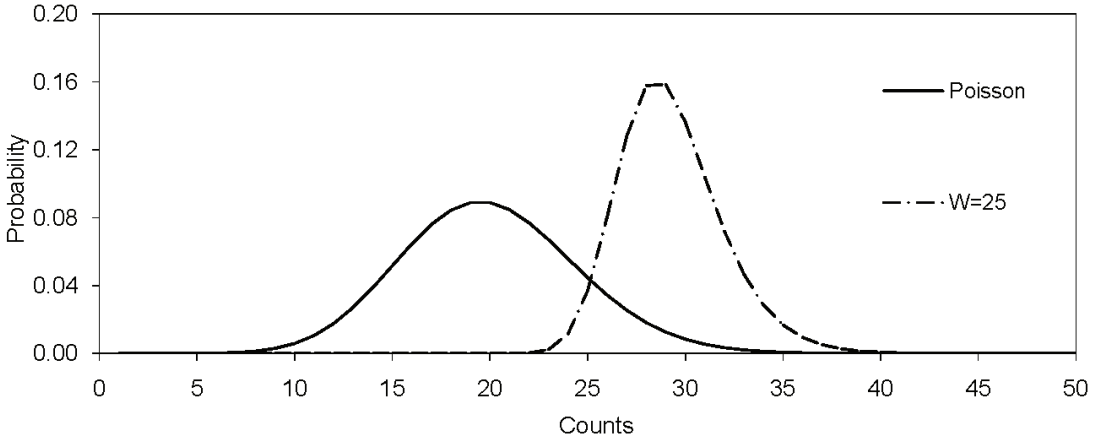

Fig. 6: Comparison of original Poisson distribution with the largest count distribution over 25 non overlapping observation windows, exploring a mass range 0-100 for an excess of events anywhere in that range. The plotted largest count distribution is the concrete manifestation of the LEE in this example.

number of bins. This approximation indeed can be demonstrated to be the asymptotic behavior of Eq. (7) when the single bin Poisson $p - value$ is very low. Finally, the parallelism with the frequency case can be further extended to the concept of effectively scanned windows and to the distributions of the second highest peak, the third, etc., in the explored mass range, as explained in detail in Ref. [12].

8 Conclusion

The search for a signal of unknown location, either in mass or frequency, is affected by noise fluctuations larger than those pertaining to a fixed location search, as described by the Look Elsewhere Effect. In this work a thorough illustration of the LEE in the frequency analysis has been given, showing its implications in the search for modulations embedded in an experimental time series, using as a concrete example the Super-Kamiokande solar neutrino data. Furthermore, a parallelism has been established between the approach in the frequency domain and a description of the search for particles of unknown mass, which led to the identification of interesting correspondences between the manifestation of the LEE in both cases, providing useful, additional insights to this rather peculiar effect.

Acknowledgement

I wish to thank the organizers who allowed me to contribute to such an enlightening workshop.

References

[1] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics", The Eur. Phys. J. C, **70**, Issue 1-2, 525-530 (2010).

[2] N.R. Lomb, “Least-squares frequency analysis of unequally spaced data", Astrophysics and Space Science, **39**, 447-462 (1976).

[3] J.D. Scargle, “Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data", Astrophys. J. **263**, 835-853 (1982).

[4] A. Schuster, “On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena", Terr. Mag. Atoms. Elect., **3**, 13-41,(1898).

[5] W.H. Press *et al.*, “Numerical recipes in Fortran ", Sect. 12.1 (Cambridge University Press,1991).

[6] J.Yoo *et al.*, “Search for periodic modulations of the solar neutrino flux in Super-Kamiokande-I", Phys. Rev. D **68**, 092002 (2003).

[7] G. Ranucci, “Likelihood scan of the Super-Kamiokande I time series data", Phys. Rev. D, **73**, 103003 (2006).

[8] W.H. Press *et al.*, “Numerical recipes in Fortran ", Sect. 13.8 (Cambridge University Press, 1991).

[9] Z. Šidàk, “Rectangular confidence region for the means of multivariate normal distributions", J. Am. Stat. Assoc. **62**, 626Ð633 (1967).

[10] J.H. Horne and S.L. Baliunas, “A prescription for period analysis of unevenly sampled time series", Astrophys. J., **302**, 757-763 (1986).

[11] G. Ranucci and M. Rovere, “Periodogram and likelihood periodicity search in the SNO solar neutrino data", Phys. Rev. D **75**, 013010 (2007).

[12] G. Ranucci, “On the significance of signal search through the ‘sliding window’ algorithm", Nucl. Instrum. Methods Phys. Res. A, **562**, 433-438 (2006).