Bull. Astr. Soc. India (2000) **28**, 363-365

# Some home-truths about small samples and counting statistics

C.J. Salter[1], D.G. Banhatti[2]

[1]*Arecibo Observatory, NAIC, P.O. Box 995, Puerto Rico, PR 00613, USA*
[2]*School of Physics, Madurai Kamaraj University, Madurai 625 021, India*

**Abstract.** Assessing a fluctuation in the sky density of astronomical objects uses counting (i.e., Poisson) statistics for estimating the errors. For a finite sample, there is a limit to the maximum possible fluctuation from the mean in units of the rms. This maximum is derived, and an example is used to illustrate these concepts.

## 1. Introduction

The significance and strength of a fluctuation from the mean must be assessed in many astronomical contexts. Large fluctuations may be excluded from mean and rms calculations, but fluctuations comparable to the rms must be included to avoid bias. For a sample of size N, the maximum any instance can deviate from the mean in units of the rms is $N^{1/2}$–$N^{-1/2}$. For example, a $3\sigma$ deviation for any instance in a sample of 10 is not possible, since the maximum possible deviation is only $2.85\sigma$.

The density of a class of discrete astronomical objects on the sky is determined by counting the number of objects in a given area and dividing by the area. Each such measurement has an associated error, estimated from the Poisson distribution, which is applicable to counting statistics, and according to which a count of n has an associated error $n^{1/2}$ (Rao 1973).

We give an example of a deep radio search, in which a possible deficiency of radio scintillators is (mis) interpreted as a radio void, due to ignoring the errors from counting statistics and excluding the deficient field from mean and rms calculation.

## 2. Maximum deviation

Consider N instances of a quality forming the sample. Starting with N = 2 (which is the minimum needed for calculating rms) and considering successively larger samples N = 3, 4 ..., some thought will show that the maximum deviation for an instance will obtain when all the other instances are clustered far away from it. For a sample of size N, taking the maximally deviant value to be A, and the other N – 1 to be all equal to B, we calculate below the deviation of A from the mean in units of the rms.

Mean $m = [A + (N - 1)B]/N$,

variance $= [1/(N - 1)][(A - m)^2 + (N - 1)(B - m)^2] = (A - B)^2/N$, rms $= |A - B|/N^{1/2}$.

So, maximum deviation $= (A - m) / (|A - B|/N^{1/2}) = (N - 1)/N^{1/2} = N^{1/2} - N^{-1/2}$.

As expected, the maximum deviation in units of rms is independent of the values of A, B and the spread of the sample $|A - B|$. A paper giving a proof that this value $N^{1/2} - N^{-1/2}$ is indeed the maximum for any arbitrary sample of N quantities, and the relation of this result to the Chebyshev-Bienayme inequality (Rao 1973) is in preparation.

## 3. Example and discussion

Banhatti (1990) found that the suggestion by Artyukh and Ogannisyan (1988b) (AOb) of a radio void toward the direction of the giant radio galaxy DA 240 was not borne out by data from the Texas survey. Here we present the relevant numbers from AOb (see Table) and show that even from *their* data, the null hypothesis that the sky density of scintillators is uniform cannot be ruled out.

For the 4 fields selected by AOb, their areas, the number of scintillators in each and quantities derived from these are tabulated (see Table). $\bar{x}_3$ and $\bar{x}_4$ are calculated as the total number N of scintillators divided by the total area A of 3 and 4 fields respectively, excluding the DA 240 field (for 3) or including it (for 4). The error $\varepsilon_f$ on the source density for each field having area $a_f$ and $N_f$ scintillators is calculated as $\varepsilon_f = N_f^{1/2}/a_f$, while the rms $\sigma_f$ for each field is $\sigma_f = [\Sigma N_f/A)a_f]^{1/2}./a_f = (N/Aa_f)^{1/2}$. That $\varepsilon_f \approx \sigma_f$ for each field (except the DA 240 field) shows consistency with Poisson statistics.

**Table 1.** Pushchino scintillator data from AOb and derived quantities

| Field | M 33 | M 31 | 3C 236 | DA 240 |
|---|---|---|---|---|
| Area ($deg^2$) | 18 | 24 | 20 | 18 |
| No. of scntltrs | 18 | 25[a] | 16[b] | 4[c] |
| Scntltr density ($deg^{-2}$) | 1.00±0.24 | 1.04±0.21 | 0.80±0.20 | 0.22±0.11 |
| $\sigma_3(deg^{-2})$ | 0.23 | 0.20 | 0.22 | 0.23 |
| $(x-\bar{x}_3) / \sigma_3$ | +0.21 | +0.45 | –0.70 | –3.18 |
| $\sigma_4(deg^{-2})$ | 0.21 | 0.18 | 0.20 | 0.21 |
| $(x-\bar{x}_4) / \sigma_4$ | +1.02 | +1.40 | +0.07 | –2.70 |

$\bar{x}_3 =$ 0.95±0.12 deg $^{-2}$, $\bar{x}_4 =$ 0.79±0.10 deg $^{-2}$.

[a] AOb claim one of the 25 to be a supernova remnant. However, detailed examination (Banhatti 1990) puts this claim in doubt.

[b] AOb give 17, including the core of 3C 236, which is a scintillator. For an unbiased count, this must be discounted, resulting in 16.

[c] AOb give 5, 2 at the edge of the field. After examining their observations (AOa), we decided to retain one of these outliers, but exclude the other.

AOb claim what is equivalent to a 6σ deficiency of radio scintillators in the DA 240 field relative to the average for the three other fields and interpret it as a radio void. The 6σ result implied by them comes about by erroneously using the scintillator densities 1.00, 1.04 and 0.80 $deg^{-2}$ to compute the mean and rms, and then using these to find the deviation of 0.22 $deg^{-2}$, the density for the DA 240 field, from this mean in units of this rms. (The result is –6.18.) The fact that the densities result from counts, which come with the usual Poisson statistics, i.e., root-N errors, is ignored in such a calculation, which takes the densities, rather than the counts for the fields, as the primary data. As seen in the Table, using counts as the primary data, the deficiency is only about 3σ, a little higher (3.18σ) if calculated using the mean for the 3 other fields, and a little lower (2.70σ) if the average for all 4 fields is used instead. Note that the maximum deviation possible for a sample of 4 is ($4^{1/2} - 4^{-1/2}$ =) 1.50σ, while the DA 240 field has a 2.70σ deficiency! However the 1.50σ applies only if the densities are the primary data. Indeed, the deviation for the DA 240 field is –1.45 in that case. Taking account of counting stastistics, the deficiency is 2.70σ rather than 1.45σ.

## 4. Conclusion

In assessing the significance of a deviation from the mean, the sample size N determines the maximum possible deviation to be $N^{1/2}$–$N^{-1/2}$ in units of the rms, i.e., the deviation is ($N^{1/2}$–$N^{-1/2}$)σ. For a quantity like the sky density of a class of astronomical objects, counting (i.e., Poisson) statistics must be used to estimate the error: $\sqrt{N}$ on a count N.

## Acknowledgements

The financial support provided by UGC is gratefully acknowledged by DGB. We thank the referee for constructive comments. N. Krishnan of Gravitation, TIFR brought to our notice the relation of our result on the maximum deviation to the Chebyshev-Bienayme inequality.

## References

Artyukh V.S., Ogannisyan M.A., 1988a, Sov. Astr. Let. 14 301.

Artyukh V.S., Ogannisyan M.A., 1988b, Sov. Astr. Let. 14 377-8.

Banhatti D.G., 1990, MNRAS, 246 7P-10P.

Rao C.R., 1973, *Linear Statistical Inference and Its Applications*, Wiley.