Gaia
DPAC
Data Processing & Analysis Consortium

Technical note on non-weighted and weighted means

prepared by:	I. Lecoeur, L. Eyer
approved by:	L. Eyer
reference:	GAIA-C7-TN-GEN-LE-016
issue:	1
revision:	0
date:	2016-03-07
status:	Issued

Abstract

This technical note illustrates two estimators of the mean of a signal with the help of simulations. It mainly focuses on a comparison of non-weighted mean versus weighted mean of a heteroscedastic photometric time series. In some cases, the weighted mean might give a biased result.

1 Introduction

CU7 will receive photometric data as Gaia time series. These ”time series” are defined as sequences of N observations ordered in time, and for each time point t_i, estimates of the magnitude y_i and of the uncertainty on this magnitude δy_i are given. The t_i, y_i can be seen as the sampling of the signal $y(t)$ with each measurement having an uncertainty level δy_i.

CU7 variability processing (PD-006) will be mainly based on the analysis of the time series and thus, the choice of good estimators, and in particular the estimator of the mean, is essential.

This technical note addresses the issue of the **estimator of the mean**. The questions is as simple as: should we use simple means or weighted means when measurements have different uncertainties? On first approximation, the weighted mean should give a better computation of the average value but is it really the case?

We first give simple definitions of the means we have used for our study: the (arithmetic) mean and the weighted mean. Then, we describe our methodology to compare the estimators of the mean and of the weighted mean and we give the results of a concrete computation based on the Hipparcos time series HIP 000008.

Finally, we show a simulation based on the Gaia sampling which clearly shows the discrepancy between the ”true” mean and the computed mean.

2 References

[PD-006], Holl, B., Guy, L., Dubath, P., et al., 2014, *CU7 Top-Level Functional Analysis and Software Requirements Specification*,
GAIA-C7-SP-GEN-PD-006,
URL http://www.rssd.esa.int/cs/livelink/open/2786539

[R1] http://www.r-project.org/

3 Definitions

3.1 Mean

Let us have a signal $y(t)$, sampled times t_i. The resulting time series is t_i, y_i, $i = 1, \ldots, N$, where N is the number of observations. The mean μ of a signal $y(t)$ can be estimated on the sample y_i, $i = 1, \ldots, N$ by the sample mean $\hat{\mu} = \overline{y}$, defined as:

$$\mu \simeq \overline{y} = \frac{\sum_{i=1}^{N} y_i}{N},$$

The sample mean of y ($\overline{y}$) minimises the (squared) Euclidean norm of deviations of the variate values from itself or explicitly stated, it minimises

$$S = \sum_{i=1}^{N} |y_i - c|^2$$

since S attains its minimum when $c = \overline{y}$.

3.2 Weighted mean

Weighted means are typically used to estimate the mean when the uncertainties are changing from one measurement to an other (heteroscedastic data).

The weighted mean is similar to an arithmetic mean, where instead of each of the data points contributing equally to the final average, some data points contribute more than others. Usually, the weights are directly related to the uncertainty of the measurements. Each data point y_i is weighted inversely by its own uncertainty.

Note that if all the weights are equal, then the weighted mean is the same as the arithmetic mean.

Let y_i be a time series of size N, with error bars (uncertainties) δy_i and weights $w_i \equiv 1/(\delta y_i)^2$, the weighted mean $\overline{y}_w$ is defined as

$$\overline{y}_w = \frac{1}{W} \sum_{i=1}^{N} w_i \, y_i,$$

where $W \equiv \sum_{i=1}^{N} w_i$ is the sum of the weights.

4 Comparisons of means versus weighted means

4.1 Data

Let's take one experimental time series from the Hipparcos mission, e.g HIP 000008 (Fig. 1). This star is a well-known Mira with a large amplitude and a period of 327.5 days. This time series consists of 77 observation times, for which the magnitude and the uncertainties are estimated over about 1000 days (so we can assume that we have 3 full pulsation cycles during the observing period).

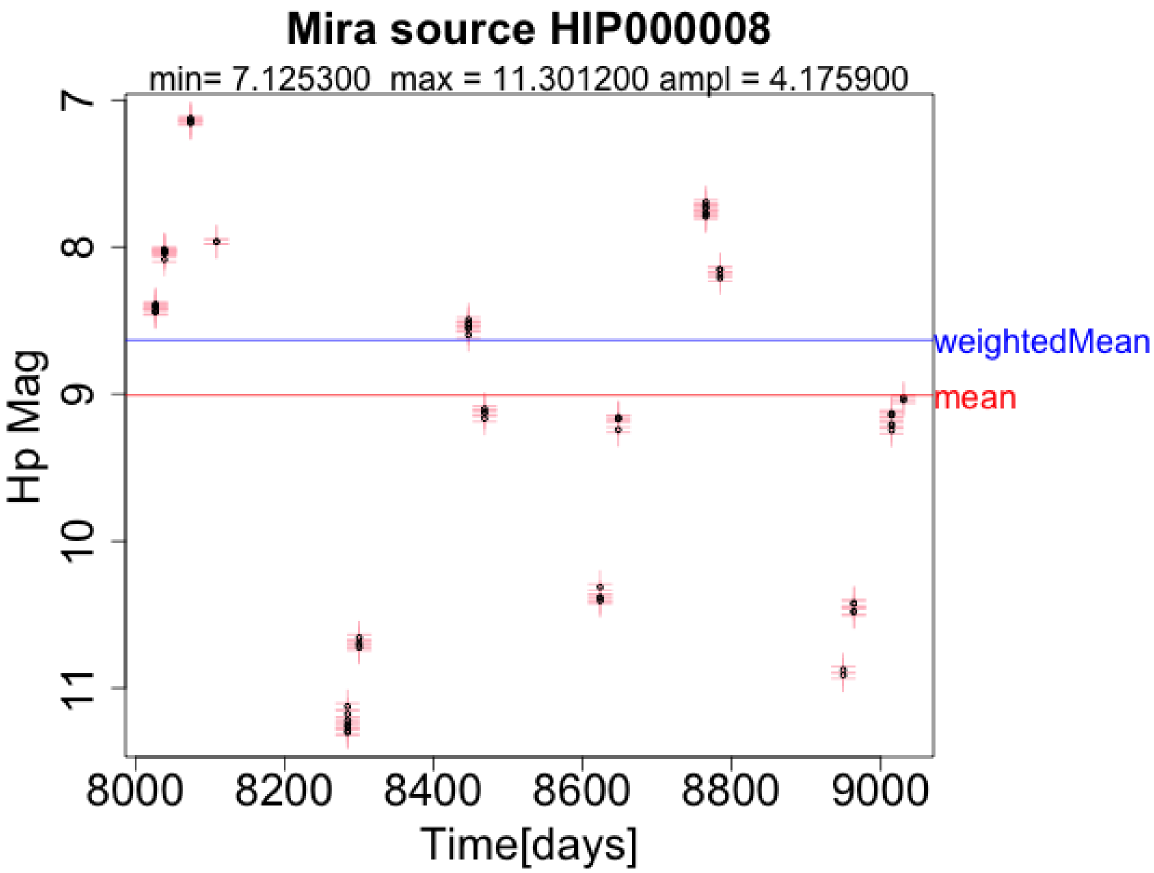

FIGURE 1: Time series of Hipparcos source HIP 000008. The coloured horizontal lines are the two estimated mean: the red line is the mean and the blue line is the weighted mean. We remark that the weighted mean is about 0.4 brighter that the mean.

This time series consists of 77 observation times, for which the magnitude and the uncertainties are estimated. The mean magnitude meanMag is equals to 9.01 and the weighted mean is equal to 8.6 mag, the latter being about 0.4 magnitude brighter.

4.2 Method

We have conducted some Monte Carlo simulations to study the relevance of the weighted mean versus the mean. This has been performed with the R package ([R1]).

Using 80 observation times of a Gaia sampling over 5 years, 10,000 samples (time series) were created with the following characteristics:

$$y_i = meanMag + \frac{ampl}{2} * sin(2\pi\nu t_i + \varphi), \qquad (1)$$

where $meanMag$ is the mean magnitude of HIP 000008 (9.007843), $ampl$ is its amplitude (= 3 magnitudes), ν is the frequency (= 1/period = 1/ 0.1309 d^{-1}), t_i are the observation times and φ is the phase randomly generated in the interval from 0 to 2π.

A gaussian noise has been added to the 10,000 time series, with a mean equals to 0 and a sigma derived from the error curve from HIP 000008. Plotting the error versus the magnitude, we assume for simplicity that the error is linear as a function of magnitude (Fig. 2).

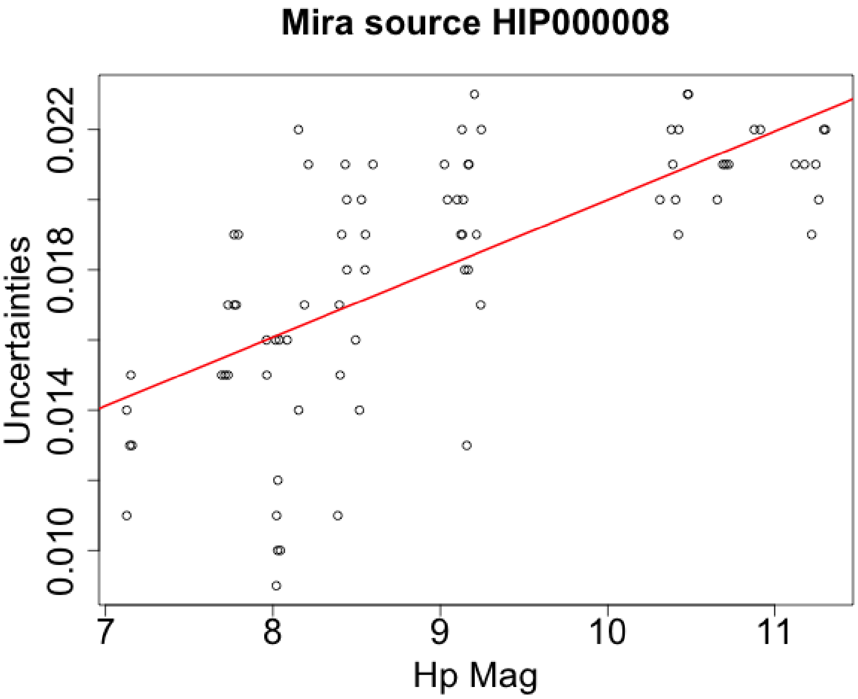

FIGURE 2: Uncertainty as a function of the Hp magnitude (HIP 000008 source).

Thus, the sigma of the gaussian noise is simply computed from the magnitude.

4.3 Results

The superimposition of the weighted and non-weighted means for all 10,000 simulated light curves is shown in Fig. 3.

Over the 10,000 time series, the mean of the arithmetic means is 9.007, while the means of the

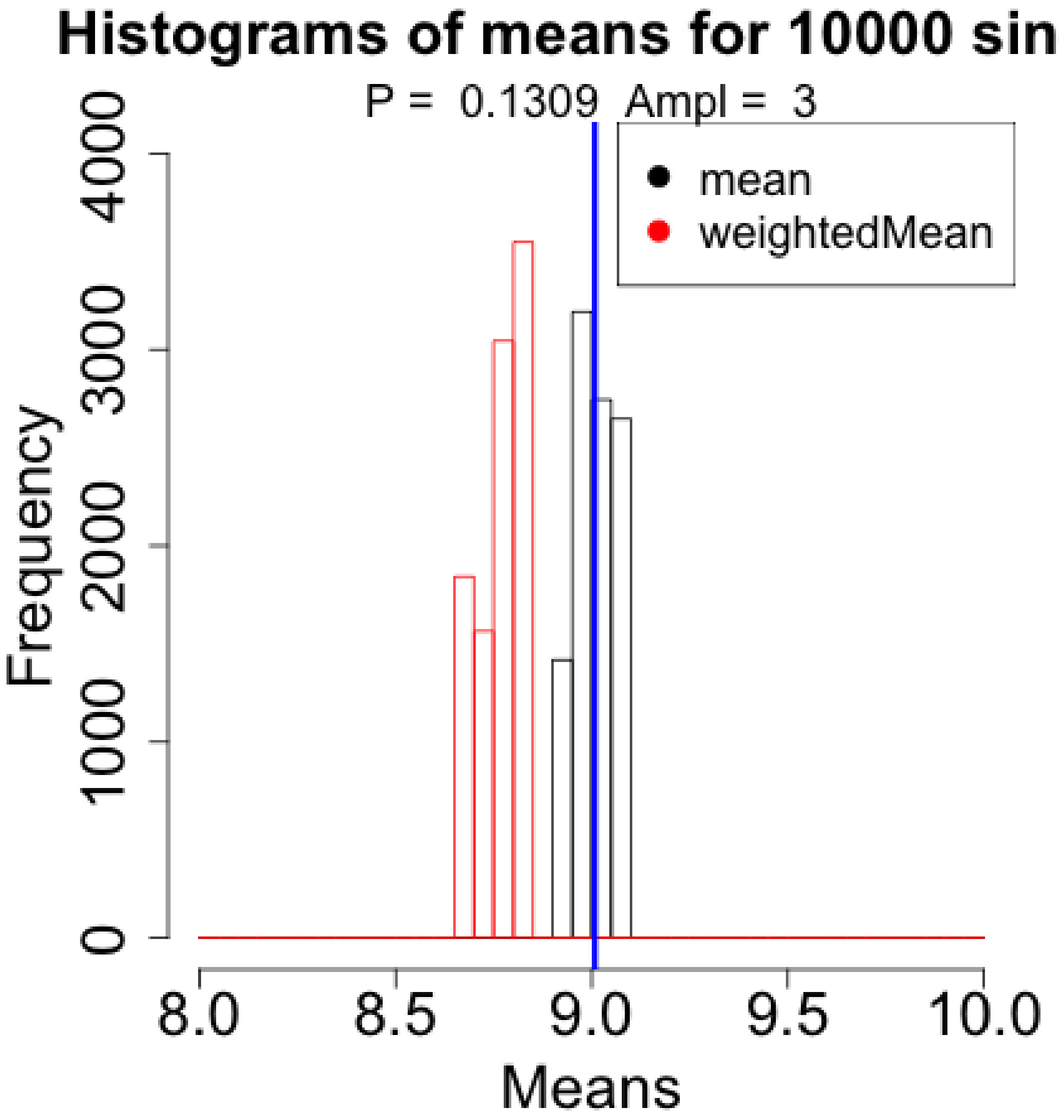

FIGURE 3: Superimposed histograms of means and weighted means for 10,000 sines. In black is the histogram of the mean values and in red the one of the weighted mean. The true value is displayed by a blue vertical line. We clearly remark that the weighted mean has a bias.

weighted means is 8.764. This shows a much closer value of the mean of the arithmetic means to the real mean (equals to 9.0078) than the mean of the weighted means.

4.4 Conclusion

We remark that the two estimators (weighted and non-weighted) of the mean of the signal give two distinct results. Given the simulation conditions, the weighted mean is giving a biased estimation of the mean of the signal. The non-weighted mean gives clearly a better estimate. This very simple simulation on this very simple problem, which is close to Hipparcos or Gaia real data, shows that we cannot adopt statistics as a cookbook without having a deeper understanding of the problem (for example here: we deal with a sinusoidal signal and there are correlations of uncertainties with the magnitude).