

Chi-Square Random Variables Involving Differences Between Data Points and the Mean of the Data

John W. Fowler

25 April 2008; revised 6 August 2008

Given a set of N measurements x_i with uncorrelated Gaussian uncertainties σ_i , we apply Gaussian refinement:

$$x_{ref} = \frac{\sum_{i=1}^N \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^N \frac{1}{\sigma_i^2}} = \sigma_{ref}^2 \sum_{i=1}^N \frac{x_i}{\sigma_i^2}$$
$$\sigma_{ref}^2 = \left(\sum_{i=1}^N \frac{1}{\sigma_i^2} \right)^{-1}$$

Now we define the following variable for the i^{th} data point:

$$\xi_i^2 \equiv \frac{(x_i - x_{ref})^2}{\sigma_i^2 + \sigma_{ref}^2}$$

This is the squared difference between two Gaussian random variables divided by the sum of their uncertainty variances, so it seems to satisfy the rigorous definition of a chi-square random variable. It does *not*, however, because x_i and x_{ref} are correlated. Using circumflex to indicate true quantities, we have

$$\hat{x}_{ref} + \varepsilon_{ref} = \hat{x} + \varepsilon_{ref} = \sigma_{ref}^2 \sum_{i=1}^N \frac{\hat{x} + \varepsilon_i}{\sigma_i^2} = \sigma_{ref}^2 \left(\sum_{i=1}^N \frac{\hat{x}}{\sigma_i^2} + \sum_{i=1}^N \frac{\varepsilon_i}{\sigma_i^2} \right)$$
$$\hat{x} = \sigma_{ref}^2 \sum_{i=1}^N \frac{\hat{x}}{\sigma_i^2}$$

Subtracting the second equation from the first, and changing the dummy summation index to j ,

$$\varepsilon_{ref} = \sigma_{ref}^2 \sum_{j=1}^N \frac{\varepsilon_j}{\sigma_j^2}$$

Multiplying both sides by ε_i and taking expectation values,

$$\langle \varepsilon_i \varepsilon_{ref} \rangle = \left\langle \sigma_{ref}^2 \sum_{j=1}^N \frac{\varepsilon_i \varepsilon_j}{\sigma_j^2} \right\rangle = \sigma_{ref}^2 \sum_{j=1}^N \frac{\langle \varepsilon_i \varepsilon_j \rangle}{\sigma_j^2} = \sigma_{ref}^2 \frac{\sigma_i^2}{\sigma_i^2} = \sigma_{ref}^2$$

So the refined value's error has a covariance with each data error equal to the former's own final uncertainty variance. Now we can compute the variance of the difference between any given data value and the mean:

$$\begin{aligned} z &\equiv x_i - x_{ref} \\ \hat{z} + \varepsilon_z &= (\hat{x} + \varepsilon_i) - (\hat{x} + \varepsilon_{ref}) \\ \hat{z} &= \hat{x} - \hat{x} = 0 \\ \varepsilon_z &= \varepsilon_i - \varepsilon_{ref} \\ \sigma_z^2 &= \langle \varepsilon_z^2 \rangle = \langle \varepsilon_i^2 - 2\varepsilon_i \varepsilon_{ref} + \varepsilon_{ref}^2 \rangle \\ &= \sigma_i^2 - 2\langle \varepsilon_i \varepsilon_{ref} \rangle + \sigma_{ref}^2 \\ &= \sigma_i^2 - 2\sigma_{ref}^2 + \sigma_{ref}^2 \\ &= \sigma_i^2 - \sigma_{ref}^2 \end{aligned}$$

But z^2 divided by its own variance is chi-square with one degree of freedom, and so

$$\chi_i^2 \equiv \frac{(x_i - x_{ref})^2}{\sigma_i^2 - \sigma_{ref}^2}$$

is chi-square with one degree of freedom. The same is true if the data errors are correlated with each other, as long as the full data error covariance matrix is used throughout (the proof is left as an exercise to the reader).

Note that the expression above is indeterminate for $N=1$; hence it may be used only for sample sizes greater than 1. If the data errors are uncorrelated, then for $N=2$, and in the limit of $N \rightarrow \infty$, the sum over i from $i=1$ to N yields a value equal to $N/(N-1)$ times the value of the following sum:

$$\sum_{i=1}^N \frac{(x_i - x_{ref})^2}{\sigma_i^2}$$

which is the usual expression for χ^2 when the data errors are uncorrelated; this sum is distributed as χ^2 with $N-1$ degrees of freedom in the limit $N \rightarrow \infty$ and for the special case of $N=2$, and also in the special case in which all the data errors have the same variance, i.e., $\sigma_i = \text{constant}$. If none of these conditions is met, then the summation is only approximately distributed as χ^2 with $N-1$ degrees of freedom.