CORRECTION OF FREQUENCY FUNCTIONS FOR OBSERVATIONAL ERRORS OF THE VARIABLES

ROBERT J. TRUMPLER
UNIVERSITY OF CALIFORNIA

1. Introduction

In astronomy statistical distributions or frequency functions are often established empirically from observational data that are affected by appreciable measuring errors. The problem of correcting frequency functions for the effects of observational errors is therefore of great importance and has received considerable attention. In the following we shall briefly summarize the solutions applicable under various conditions.

2. One variable directly measured

For a given population or sample of stars we want to study the distribution according to one variable x. We assume that for each individual of the population a measured value ξ of the variable is available and that we have established the frequency function $F_0(\xi)$ from these data. To find the frequency function $F_t(x)$ of the true values x we have to know the statistical distribution of the measuring errors ϵ. In the most general case the error distribution may vary with x, the quantity measured; it must be considered as an array distribution

$$\Phi\,(\epsilon|x)$$

where x is to be treated like a parameter.

The three variables x, ξ, ϵ are subject to the condition

$$\xi = x + \epsilon\,,$$

and the corresponding relation between the frequency functions can be written in the two forms

$$(1)\qquad F_0(\xi) = \int_{-\infty}^{+\infty} F_t(x)\Phi(\xi - x\,|\,x)\,dx$$

$$= \int_{-\infty}^{+\infty} F_t(\xi - \epsilon)\Phi(\epsilon\,|\,\xi - \epsilon)\,d\epsilon\,.$$

When $F_0(\xi)$ and $\Phi(\epsilon|x)$ are known this is an integral equation (Fredholm's first kind) for the determination of $F_t(x)$.

The most important special cases are:

(a) The error distribution is independent of x.

(1) All measures are of equal accuracy and the errors have a Gaussian distribution

$$\Phi(\epsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\epsilon/\sigma)^2/2} \tag{2}$$

where σ is the so called mean error of the observations. The solution of the integral equation (1) is accomplished by means of interpolation formulae or interpolation series or by numerical methods. When the errors are relatively small, the Eddington series

$$F_t(\xi) = F_0(\xi) - \frac{\sigma^2}{2} F_0''(\xi) + \frac{\sigma^4}{8} F_0^{\mathrm{IV}}(\xi) - \dots, \tag{3}$$

where $F_t(\xi)$ is the value of $F_t(x)$ for $x = \xi$, generally gives a good approximation.

(2) The measurements ξ are not of the same accuracy and are assigned different weights. The individuals of the population are divided into n groups according to the weight w_i or the corresponding mean error σ_i. The fraction of the population having ξ measurements of a mean error σ_i is designated as $\nu(\sigma_i)$ and we assume that the weight distribution $\nu(\sigma_i)$ is independent of x. The error law

$$\Phi(\epsilon) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{\nu(\sigma_i)}{\sigma_i} e^{-(\epsilon/\sigma_i)^2/2} \tag{4}$$

is no longer of the Gaussian form, but the series (3) can still be used with average values of σ^2 and σ^4

$$\overline{\sigma^2} = \sum_{i=1}^{n} \nu(\sigma_i)\,\sigma_i^2; \qquad \overline{\sigma^4} = \sum_{i=1}^{n} \nu(\sigma_i)\,\sigma_i^4. \tag{5}$$

(b) When the error distribution varies with x, the variation is usually of the following two types, the second being more general and including the first as a special case:

(1) The mean error of measurement changes with x; in the error law (2) the dispersion σ must be replaced by a function $\sigma(x)$.

(2) The weight distribution changes with x; in the error law (4) we have to replace $\nu(\sigma_i)$ by $\nu(\sigma_i | x)$.

The solution of the integral equation (1) in these two cases is somewhat more complicated but can be obtained by numerical methods. When the variation of σ or $\nu(\sigma_i)$ with x is very slow so that the derivatives $\sigma'(x)$, $\frac{d}{dx}\sum \nu(\sigma_i | x)$, . . . are negligible, the series (3) can still be used by substituting variable values $\sigma^2(\xi)$, $\sigma^4(\xi)$, $\sum \sigma_i^2 \nu(\sigma_i | \xi)$, $\sum \sigma_i^4 \nu(\sigma_i | \xi)$ for σ^2, σ^4, ...

3. Two variables directly measured

Between the true values x, y of two variables, the measured values ξ, η and the errors ϵ_1 of ξ and ϵ_2 of η, we have two conditions

$$\xi = x + \epsilon_1, \qquad \eta = y + \epsilon_2.$$

In the most general case we have to consider a bivariate error distribution which may vary with x and y

$$\Phi\,(\epsilon_1,\ \epsilon_2\,|\,x,\ y)\,. \tag{6}$$

The true distribution $F_t(x, y)$ of the two variables is obtained from the distribution $F_0(\xi, \eta)$ of the measured values by solution of the double integral equation

$$F_0(\xi,\ \eta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} F_t(x,\ y)\Phi(\xi - x,\ \eta - y\,|\,x,\ y)\,dxdy\,. \tag{7}$$

a) The simplest case, most frequent in practice, is that where the two errors ϵ_1, ϵ_2 are independent of each other and independent of the variables x, y. The error function (6) is then reduced to

$$\Phi_1\,(\epsilon_1)\ \Phi_2\,(\epsilon_2)$$

and the double integral equation (7) can be solved in two steps

$$F_0(\xi,\ \eta) = \int_{-\infty}^{+\infty} H(x,\ \eta)\Phi_1(\xi - x)\,dx\,, \tag{8}$$

$$H(x,\ \eta) = \int_{-\infty}^{+\infty} F_t(x,\ y)\Phi_2(\eta - y)\,dy\,.$$

By the first equation we correct the ξ-distribution of each η-array for the errors in ξ and obtain the distribution $H(x, \eta)$ of the true x-values and the measured η-values. By the second equation we correct the η-distribution of each x-array for the errors of η. The solution of the bivariate problem is thus resolved into a series of solutions each of which involves only one variable.

For small errors with Gaussian distributions the solution can be given in the form of a series similar to the series (3).

$$F_t(\xi,\ \eta) = F_0(\xi,\ \eta) - \frac{\sigma_1^2}{2}\frac{\partial^2 F_0}{\partial \xi^2} - \frac{\sigma_2^2}{2}\frac{\partial^2 F_0}{\partial \eta^2} + \ldots\,. \tag{9}$$

A generalization of this series using the even order moments of the bivariate error distribution will also cover the cases where ϵ_1 and ϵ_2 are correlated but independent of x, y, and where the observations are of unequal weight.

b) When the two errors are independent of each other while the distribution of ϵ_1 varies only with x, that of ϵ_2 with y or with both x and y the frequency function of the errors has the form

$$\Phi_1\,(\epsilon_1\,|\,x)\ \Phi_2\,(\epsilon_2\,|\,x,\ y)$$

and the correction for observational errors can still be made in two separate steps similar to equations (8).

4. One variable calculated from several observed quantities

We are often interested in the distribution of a variable t which is not directly observable but is a known function

$$t = f\,(x, y\,.\,.\,.)$$

of several observable quantities $x, y, \ldots$. The problem then is to find the true frequency function $G_t(t)$ from measured values ξ, η, . . . of these quantities, the measurements ξ, η, . . . being affected by observational errors ϵ_1, ϵ_2, ...

a) The rigorous method consists in first establishing the distribution $F_0(\xi, \eta, \ldots)$ of the measured values and correcting this distribution for observational errors by one of the methods discussed under section 3. The true distribution $F_t(x, y, \ldots)$ is then transformed to a new set of variables which includes t defined by $t = f(x, y, \ldots)$; the most convenient is usually to eliminate one of the variables—say x—and express it in terms of t, y, ... By integration over all variables other than t we obtain

$$(10) \qquad G_t(t) = \int_{-\infty}^{\infty} F_t\left[x(t, y, \ldots), y, \ldots\right] \frac{\partial x(t, y, \ldots)}{\partial t}\, dy \ldots .$$

The difficulty with this method is that it requires a very large population, since we have to establish a bivariate or multivariate distribution.

b) More direct but approximate methods based on series developments may be used, provided the observational errors are relatively small and have independent Gaussian distributions with mean errors σ_1, σ_2, ... From the measurements ξ, η, . . . we calculate

$$\tau = f(\xi, \eta, \ldots)$$

for each individual and establish the frequency function $G_0(\tau)$. If we neglect terms of fourth and higher orders in σ_1, σ_2, . . . the true distribution $G_t(t)$ is found from $G_0(\tau)$ by

$$(11) \quad G_t(\tau) = G_0(\tau)\left\{1 - \frac{\sigma_1^2}{2}\left(\frac{d^2A_1}{d\tau^2} - \frac{dB_1}{d\tau}\right) - \frac{\sigma_2^2}{2}\left(\frac{d^2A_2}{d\tau^2} - \frac{dB_2}{d\tau}\right) - \ldots\right\}$$
$$- G_0'(\tau)\left\{\frac{\sigma_1^2}{2}\left(2\frac{dA_1}{d\tau} - B_1\right) + \frac{\sigma_2^2}{2}\left(2\frac{dA_2}{d\tau} - B_2\right) + \ldots\right\}$$
$$- G_0''(\tau)\left\{\frac{\sigma_1^2}{2}A_1 + \frac{\sigma_2^2}{2}A_2 + \ldots\right\}.$$

The functions $A_1(\tau)$, $A_2(\tau)$, . . . , $B_1(\tau)$, $B_2(\tau)$, . . . are means formed for successive intervals of τ:

$$A_1(\tau) = E\left[\left(\frac{\partial f}{\partial \xi}\right)^2 \middle| \tau\right], \qquad A_2(\tau) = E\left[\left(\frac{\partial f}{\partial \eta}\right)^2 \middle| \tau\right], \ldots ,$$

$$B_1(\tau) = E\left[\frac{\partial^2 f}{\partial \xi^2} \middle| \tau\right], \qquad B_2(\tau) = E\left[\frac{\partial^2 f}{\partial \eta^2} \middle| \tau\right], \ldots .$$

The regressions A_1, B_1, . . . can be established empirically from the observational data even for a population of moderate size. Often, however, it is possible to derive some of them from theoretical considerations.