Moments d'échantillons superposés

par Jörg W. Müller

Bureau International des Poids et Mesures, F-92310 SEVRES

Summary : For two samples, where the moments are known, explicit formulae are derived for the evaluation of expectation value, variance and third central moment of the pooled sample.

1. Introduction et notation

Il est bien connu que, pour un échantillon de n valeurs x_i qui ont toutes le même poids statistique, les estimations non biaisées des paramètres espérance mathématique, variance et troisième moment centré de la population correspondante sont données par exemple par les expressions

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{pour } n \geqslant 1$$

$$v = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 , \qquad \text{"} \quad n \geqslant 2 \quad \text{et} \tag{1}$$

$$z = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - m)^3 , \qquad \text{"} \quad n \geqslant 3 .$$

Nous utilisons donc les mêmes notations que dans une note précédente sur un problème analogue [1] , à laquelle ce qui suit peut servir de complément. Pour distinguer l'appartenance des quantités à l'un ou à l'autre des deux échantillons, nous ajoutons (en bas à gauche) un indice s qui peut prendre les valeurs 1 ou 2.

Les moments ordinaires d'ordre r sont donnés par

$${}_s m_r = \frac{1}{{}_s n} \sum_{i=1}^{{}_s n} {}_s x_i^r , \qquad r = 1, 2, 3, \ldots , \tag{2}$$

où l'indice r = 1 est normalement supprimé.

---

[1] J.W. Müller: "Mise à jour de moments empiriques", WPN-214

Pour la combinaison des deux échantillons, de taille

$$N = {}_1n + {}_2n \,, \tag{3}$$

les moments respectifs seront notés avec des lettres majuscules. On a donc

$$M_r = \frac{1}{N} \sum_{s=1}^{2} \sum_{i=1}^{{}_sn} {}_sx_i^r \,, \tag{4}$$

et les quantités recherchées sont M, V et Z.

2. Evaluation des nouveaux moments

a) La valeur moyenne

La nouvelle espérance mathématique est estimée par la valeur moyenne qui s'obtient directement à partir de (4). Elle est

$$M = \frac{1}{N} \Big(\sum_{i=1}^{{}_1n} {}_1x_i + \sum_{i=1}^{{}_2n} {}_2x_i \Big) = \frac{1}{N} ({}_1n \; {}_1m + {}_2n \; {}_2m) \,, \tag{5}$$

formule qui ne soulève aucun problème. Il s'ensuit que pour ${}_1m \geqslant {}_2m$, on a toujours l'inégalité

$${}_2m \leqslant M \leqslant {}_1m \,.$$

b) La variance

Par analogie avec (1) la variance des N mesures s'exprime maintenant par

$$V = \frac{1}{N-1} \left\{ \sum_{i=1}^{{}_1n} ({}_1x_i - M)^2 + \sum_{i=1}^{{}_2n} ({}_2x_i - M)^2 \right\} . \tag{6}$$

Un simple réarrangement donne

$$\begin{aligned} V &= \frac{1}{N-1} \left\{ \sum ({}_1x_i - {}_1m + {}_1m - M)^2 + \sum ({}_2x_i - {}_2m + {}_2m - M)^2 \right\} \\ &= \frac{1}{N-1} \left\{ \sum \left[({}_1x_i - {}_1m)^2 + ({}_1m - M)^2 + 2({}_1m - M)\,({}_1x_i - {}_1m) \right] \right. \\ &\quad \left. + \sum \left[({}_2x_i - {}_2m)^2 + ({}_2m - M)^2 + 2({}_2m - M)\,({}_2x_i - {}_2m) \right] \right\} . \end{aligned}$$

Avec la variance ${}_sv$ de l'échantillon s, on a d'après (1)

$$\sum ({}_sx_i - {}_sm)^2 = ({}_sn - 1)\, {}_sv \,,$$

et puisque $\sum ({}_s x_i - {}_s m) = 0$, on peut écrire

$$V = \frac{1}{N-1}\left[({}_1n-1)\,{}_1v + ({}_2n-1)\,{}_2v + {}_1n({}_1m-M)^2 + {}_2n({}_2m-M)^2\right] .$$

M est remplacé par ${}_1m$ et ${}_2m$, par exemple à l'aide de

$${}_1n({}_1m-M)^2 = \frac{{}_1n}{N^2}\left[N\,{}_1m - ({}_1n\,{}_1m + {}_2n\,{}_2m)\right]^2$$

$$= \frac{{}_1n\,{}_2n^2}{N^2}({}_1m - {}_2m)^2 ,$$

d'où l'on dérive la formule finale

$$V = \frac{1}{N-1}\left[({}_1n-1)\,{}_1v + ({}_2n-1)\,{}_2v + \frac{{}_1n\,{}_2n}{N}({}_1m-{}_2m)^2\right] . \qquad (7)$$

On peut en déduire l'inégalité (pour ${}_1v \geqslant {}_2v$)

$$\left(\frac{N-2}{N-1}\right){}_2v \leqslant V < {}_1v + \frac{N}{4(N-1)}({}_1m - {}_2m)^2 ,$$

qui, cependant, est peu utile. Pour ${}_1n = {}_2n \gg 1$, on a l'approximation

$$V \cong \frac{1}{2}({}_1v + {}_2v) + \frac{1}{4}({}_1m - {}_2m)^2 .$$

c) Le troisième moment centré

La démarche est tout à fait analogue à celle qui est décrite pour la variance et l'on se contentera d'esquisser les étapes principales du calcul. Avec (1) nous partons, pour les N mesures, de l'expression

$$Z = \frac{N}{(N-1)(N-2)}\left\{\sum_{i=1}^{{}_1n}({}_1x_i - M)^3 + \sum_{i=1}^{{}_2n}({}_2x_i - M)^3\right\}$$

$$= \dots \left\{\frac{({}_1n-1)({}_1n-2)}{{}_1n}\,{}_1z + 3({}_1m-M)({}_1n-1)\,{}_1v + {}_1n({}_1m-M)^3\right.$$

$$\left. + \frac{({}_2n-1)({}_2n-2)}{{}_2n}\,{}_2z + 3({}_2m-M)({}_2n-1)\,{}_2v + {}_2n({}_2m-M)^3\right\} .$$

De nouveau, on élimine M à l'aide de la relation

$${}_1m - M = \frac{{}_2n}{N}({}_1m - {}_2m) , \qquad \text{etc.}$$

En réarrangeant les termes on trouve la formule finale

$$Z = \frac{N}{(N-1)(N-2)}\left\{\frac{({}_1n-1)({}_1n-2)}{{}_1n}\,{}_1z + \frac{({}_2n-1)({}_2n-2)}{{}_2n}\,{}_2z\right.$$

$$\left. + \frac{({}_1m-{}_2m)}{N}\left[3({}_1n-1)\,{}_2n\,{}_1v - 3({}_2n-1)\,{}_1n\,{}_2v - \frac{{}_1n\,{}_2n}{N}({}_1n-{}_2n)({}_1m-{}_2m)^2\right]\right\} . \quad (8)$$

Pour des échantillons de taille égale $({}_1n = {}_2n)$ et $N \gg 1$, on a la formule approchée

$$Z \cong \frac{1}{2}({}_1z + {}_2z) + \frac{3}{4}({}_1m - {}_2m)({}_1v - {}_2v) .$$

Il va sans dire que les expressions principales, qui sont (7) et (8), auraient pu être obtenues à partir des formules analogues à (5) et (6) de [1], mais le remplacement des moments M_2 et M_3 pour l'ensemble par les grandeurs ${}_sv$ et ${}_sz$ se serait avéré plus pénible.

Un programme d'ordinateur, écrit en Fortran IV, et disponible sur demande, calcule les grandeurs M, V et Z à partir de ${}_sn$, ${}_sm$, ${}_sv$ et ${}_sz$ comme données d'entrée. Si ${}_1n = 1$ (et donc ${}_1v$ et ${}_1z$ non définis), on applique automatiquement les formules établies dans [1] pour cette situation.

(Mai 1980)