Mise à jour de moments empiriques*

par Jörg W. Müller

Bureau International des Poids et Mesures, F-92310 SEVRES

Summary: Explicit expressions are given for updating the numerical values for mean, variance and third central moment of an empirical sample when an additional measurement becomes available.

1. Introduction

L'importance pratique des moments pour caractériser une distribution statistique n'a plus besoin d'être soulignée: leur emploi est depuis longtemps devenu une nécessité qui va de soi, même dans des domaines qui, à première vue, ne semblent guère avoir de liens avec les sciences physiques. En particulier, l'évaluation d'une valeur moyenne ou d'un écart-type pour une série de mesures est souvent déjà programmée, même sur de petites calculatrices de poche. Cependant, le calcul numérique peut être un peu moins évident s'il convient d'associer différents "poids" aux mesures, ou si l'on s'intéresse aussi à la précision des résultats ainsi obtenus, exprimée normalement par les écarts-types respectifs. Pour ne pas alourdir inutilement notre propos, nous renoncerons dans ce qui suit à ces complications - dont la description est d'ailleurs bien connue.

De plus, l'intérêt accru se manifestant pour le troisième moment centré, qui fournit une mesure commode d'une éventuelle asymétrie, demande des calculs plus fastidieux, dont la simplicité ne saurait garantir l'absence d'erreurs numériques. Notons que, même dans un domaine comme la numismatique [1], que l'on croirait pourtant bien à l'abri de raisonnements d'ordre mathématique, l'évaluation du troisième moment centré pour la répartition du poids de pièces de monnaie est devenue une opération indispensable depuis qu'on a reconnu qu'il peut être utilisé pour l'estimation du poids originel d'une série de pièces qui sont maintenant plus ou moins altérées par l'usure.

Les temps où l'on pouvait, en bonne conscience, se contenter de comparer un peu n'importe quelle distribution empirique à la seule loi normale de Gauss et Laplace, en s'appuyant (avec peu de raison) pour justifier son emploi sur la loi limite, sont bien révolus, et il ne faut pas le regretter.

---

* En hommage au Professeur Julien Guey pour sa contribution importante au développement de la numismatique statistique

Il arrive assez souvent - après avoir effectué tous les calculs nécessaires pour déterminer les moments - qu'une nouvelle donnée expérimentale devienne disponible et qu'on aimerait l'incorporer dans les résultats précédents. Or, cela n'est souvent possible qu'au prix d'une répétition de tous les calculs.
Une telle situation est peu satisfaisante et l'on peut se demander s'il n'y a pas moyen de déterminer directement les corrections à apporter (qui sont d'ailleurs souvent peu importantes), en n'utilisant que la nouvelle donnée et les moments établis antérieurement. C'est ce petit problème que nous nous proposons de traiter dans cette note.

Alors que l'influence sur la valeur moyenne est triviale et que l'effet sur la variance se devine avec un peu d'imagination, au moins de façon approximative, ceci est beaucoup moins évident pour le troisième moment centré μ_3. Nous avions originellement supposé, en toute naïveté, qu'une nouvelle donnée qui se situe, par exemple, bien en dessous de la valeur moyenne, diminuerait sans doute μ_3 (et vice-versa pour une valeur au-delà), car cette contribution accentuerait la "queue à gauche" de la distribution. Or, et à notre surprise, les résultats numériques montrent le plus souvent un comportement opposé. S'il est déjà regrettable de ne pas disposer d'une estimation quantitative de l'influence sur μ_3, il est fort troublant de se tromper même de signe, sans en connaître la raison. Cette situation a été le point de départ pour examiner ces choses d'un peu plus près.

2. Quelques notions de base et notation adoptée

On suppose qu'on dispose d'un échantillon de mesures x_j, avec $j = 1, 2, \ldots, n$. Pour estimer les paramètres de la population de base, on a calculé les quantités

$$m = \frac{1}{n} \sum_{j=1}^{n} x_j = \text{valeur moyenne}, \tag{1}$$

$$v = \frac{1}{n-1} \sum_{1}^{n} (x_j - m)^2 = \text{variance}^* \qquad \text{et} \tag{2}$$

$$z = \frac{n}{(n-1)(n-2)} \sum_{1}^{n} (x_j - m)^3 = \text{troisième moment centré}^{**} . \tag{3}$$

Les facteurs $\frac{1}{n-1}$ pour v et $\frac{n}{(n-1)(n-2)}$ pour z garantissent l'absence de biais pour un échantillon de taille finie. Ils sont d'emploi général [2] et ne supposent (à part l'existence des moments correspondants) en particulier aucune distribution spécifique pour la population.

---

* habituellement désignée par s^2 ou σ^2

** le plus souvent noté μ_3

Les sommations portant sur des différences par rapport à la valeur moyenne m supposent évidemment l'évaluation préalable de m. Or, au fur et à mesure que les résultats x_i deviennent disponibles, m change et tous les calculs pour v et z doivent être refaits. Ceci est peu économique et il convient d'y remédier. En effet, on peut montrer qu'à l'aide des moments "empiriques"

$$m_r \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^r , \qquad r = 1, 2, 3, \ldots, \tag{4}$$

les formules (2) et (3) peuvent être écrites sous la forme équivalente

$$v = \frac{n}{n-1} (m_2 - m^2) \qquad \text{et} \tag{5}$$

$$z = \frac{n^2}{(n-1)(n-2)} (m_3 - 3mm_2 + 2m^3) , \tag{6}$$

où $m \equiv m_1$. Il suffit alors de mettre à jour les trois premiers moments définis par (4) et de les insérer dans (5) et (6) pour obtenir les quantités v et z, ce qui simplifie le travail de calcul.

Admettons maintenant qu'une nouvelle mesure $x_{n+1} \equiv x_0$ soit disponible. Les nouveaux moments "empiriques" seront notés

$$M_r = \frac{1}{n+1} \sum_{i=0}^{n} x_i^r , \qquad r = 1, 2, 3, \ldots,$$

pour les distinguer facilement de (4). On a donc

$$M_r = \frac{1}{n+1} (x_0^r + \sum_{1}^{n} x_i^r) = \frac{1}{n+1} (x_0^r + n \cdot m_r)$$

$$= m_r \left[\frac{n}{n+1} + \frac{x_0^r}{m_r (n+1)} \right] = \frac{m_r}{n+1} (n + \frac{x_0^r}{m_r}) . \tag{8}$$

Il en découle que M_r est plus grand (ou plus petit) que m_r si x_0^r est plus grand (ou plus petit) que m_r. Il est pratique de définir des quantités d_r par

$$x_0^r = m_r + d_r . \tag{9}$$

Par conséquent, on peut aussi écrire

$$M_r = \frac{m_r}{n+1} (n+1 + \frac{d_r}{m_r}) = m_r + \frac{d_r}{n+1} . \tag{10}$$

Les moments correspondant aux n+1 mesures seront notés par des majuscules; on écrira donc M pour la valeur moyenne, V pour la variance et Z pour le troisième moment centré.

3. Evaluation des nouveaux moments

a) La valeur moyenne

La détermination de la nouvelle valeur moyenne est maintenant triviale, car il découle de (10) que

$$M = m + \frac{d}{n+1} , \tag{11}$$

où $d \equiv d_1 = x_0 - m$.

Ce résultat n'a rien de surprenant; notons que l'écart M-m a toujours le même signe que d.

b) La variance

Pour les n+1 mesures la variance est, d'après (5),

$$V = \frac{n+1}{n} (M_2 - M^2) .$$

En utilisant (10) et (5) on peut aussi écrire

$$\begin{aligned} V &= \frac{n+1}{n} \left[m_2 + \frac{d_2}{n+1} - (m + \frac{d}{n+1})^2 \right] \\ &= \frac{n+1}{n} \left[m_2 - m^2 + \frac{d_2}{n+1} - \frac{2md}{n+1} - \frac{d^2}{n+1} \right] \\ &= \frac{n+1}{n} \left[(\frac{n-1}{n}) v + \frac{1}{n+1} (d_2 - 2md - \frac{d^2}{n+1}) \right] . \end{aligned} \tag{12}$$

Pour déterminer d_2 nous formons avec (9)

$$\begin{aligned} x_0^2 &= m_2 + d_2 \\ &= (m+d)^2 = m^2 + 2md + d^2 \\ &= - (\frac{n-1}{n}) v + m_2 + 2md + d^2 , \end{aligned}$$

d'où l'on tire

$$\begin{aligned} d_2 &= - (\frac{n-1}{n}) v + 2md + d^2 , \qquad \text{ou aussi} \\ m_2 &= (\frac{n-1}{n}) v + m^2 . \end{aligned} \tag{13}$$

Il s'ensuit pour (12) que

$$V = \left(\frac{n^2-1}{n^2}\right) v + \frac{1}{n}\left[-\left(\frac{n-1}{n}\right) v + 2md + d^2 - 2md - \frac{d^2}{n+1}\right]$$

$$= \frac{(n^2-1) - (n-1)}{n^2} \cdot v + \frac{d^2}{n}\left(1 - \frac{1}{n+1}\right)$$

$$= \left(\frac{n-1}{n}\right) v + \frac{1}{n+1} d^2 \quad . \tag{14}$$

Il y a donc pour la variance un changement de

$$\Delta V \equiv V - v = \frac{d^2}{n+1} - \frac{v}{n} \quad . \tag{15}$$

Par conséquent, $V \leqslant v$ pour $d^2 \leqslant \left(\frac{n+1}{n}\right) v$, sinon $V > v$.

Par ailleurs, le même résultat peut s'obtenir en partant de (2) au lieu de (5), c'est-à-dire en écrivant

$$V = \frac{1}{n} \sum_{i=0}^{n} (x_i - M)^2$$

$$= \frac{1}{n}\left[\sum_{1}^{n} (x_i - m - M + m)^2 + (x_0 - M)^2\right] .$$

Puisque

$$x_0 - M = (x_0 - m) + (m - M) = d - \frac{d}{n+1} = \left(\frac{n}{n+1}\right) d \ ,$$

on a également

$$V = \frac{1}{n}\left[\sum_{1}^{n} \left(x_i - m - \frac{d}{n+1}\right)^2 + \left(\frac{n}{n+1}\right)^2 d^2\right]$$

$$= \frac{1}{n}\left\{\sum_{1}^{n}\left[(x_i - m)^2 - \frac{2d}{n+1}(x_i - m) + \frac{d^2}{(n+1)^2}\right] + \left(\frac{n}{n+1}\right)^2 d^2\right\}$$

$$= \frac{1}{n}\left[(n-1)\, v - 0 + \frac{n d^2}{(n+1)^2} + \left(\frac{n}{n+1}\right)^2 d^2\right]$$

$$= \frac{1}{n}\left[(n-1)\, v + \left(\frac{n}{n+1}\right) d^2\right] = \left(\frac{n-1}{n}\right) v + \frac{1}{n+1} d^2 \ ,$$

ce qui est identique à (14).

c) Le troisième moment centré

En partant de (6), on a pour l'ensemble des n+1 mesures

$$Z = \frac{(n+1)^2}{n(n-1)} (M_3 - 3\,M\,M_2 + 2\,M^3) ,$$

ou, avec (10),

$$\begin{aligned} Z &= \frac{(n+1)^2}{n(n-1)} \left\{ m_3 + \frac{d_3}{n+1} - 3\,(m + \frac{d}{n+1})\,(m_2 + \frac{d_2}{n+1}) + 2\,(m + \frac{d}{n+1})^3 \right\} \\ &= \ldots \left\{ m_3 - 3mm_2 + 2m^3 + \frac{d_3}{n+1} - 3 \left[m \frac{d_2}{n+1} + \frac{d}{n+1}\, m_2 + \frac{d\,d_2}{(n+1)^2} \right] \right. \\ &\qquad \left. + 2 \left[3m^2 \frac{d}{n+1} + 3m (\frac{d}{n+1})^2 + (\frac{d}{n+1})^3 \right] \right\} \\ &= \ldots \left\{ \frac{(n-1)\,(n-2)}{n^2} z + \frac{1}{n+1} \left[d_3 + 3md_2 - 3m_2 d \right. \right. \\ &\qquad \left. \left. - 3 \frac{dd_2}{n+1} + 6m^2 d + 6m \frac{d^2}{n+1} + 2 \frac{d^3}{(n+1)^2} \right] \right\} . \end{aligned} \tag{16}$$

Il s'agit maintenant d'éliminer toutes les grandeurs comportant un indice. Par conséquent, et avec (13), il nous faut une relation pour d_3. Elle s'obtient à partir de (9) en posant

$$x_0^3 = (m + d)^3 = m_3 + d_3 ,$$

d'où $d_3 = (m + d)^3 - m_3$.

Puisque

$$\begin{aligned} m_3 &= \frac{(n-1)\,(n-2)}{n^2} z + 3mm_2 - 2m^3 \\ &= \frac{(n-1)\,(n-2)}{n^2} z + 3m (\frac{n-1}{n})\, v - m^3 , \end{aligned}$$

on arrive à

$$d_3 = (m+d)^3 - \frac{(n-1)\,(n-2)}{n^2} z - 3m (\frac{n-1}{n})\, v - m^3 . \tag{17}$$

Maintenant, il faut insérer (13) et (17) dans (16). Après quelques arrangements algébriques qu'il serait peu utile de reproduire, on aboutit finalement à

$$Z = z \frac{(n+1)\,(n-2)}{n^2} - \frac{d}{n} (3v - \frac{n}{n+1} d^2) . \tag{18}$$

Ceci correspond à un changement du troisième moment de

$$\Delta Z \equiv Z - z = \frac{1}{n}\left[d\left(\frac{n}{n+1}d^2 - 3v\right) - z\,\frac{n+2}{n}\right]. \qquad (19)$$

Pour $n \gg 1$, on a donc l'approximation

$$\Delta Z \cong \frac{d}{n}(d^2 - 3v) - \frac{z}{n}$$

qui est représentée dans la figure 1. Elle change trois fois de signe en fonction de d, à savoir approximativement aux abscisses $d = 0$ et $\pm\sqrt{3v}$.

Si le nombre n n'est pas très élevé, il convient d'utiliser la formule exacte (18) pour déterminer la nouvelle valeur Z du troisième moment centré, qui ne fait appel qu'à la variance v et au troisième moment centré z valables pour les n mesures précédentes, que l'on suppose connus.

Mentionnons que (18) aurait également pu s'obtenir en partant de (3) au lieu de (6).

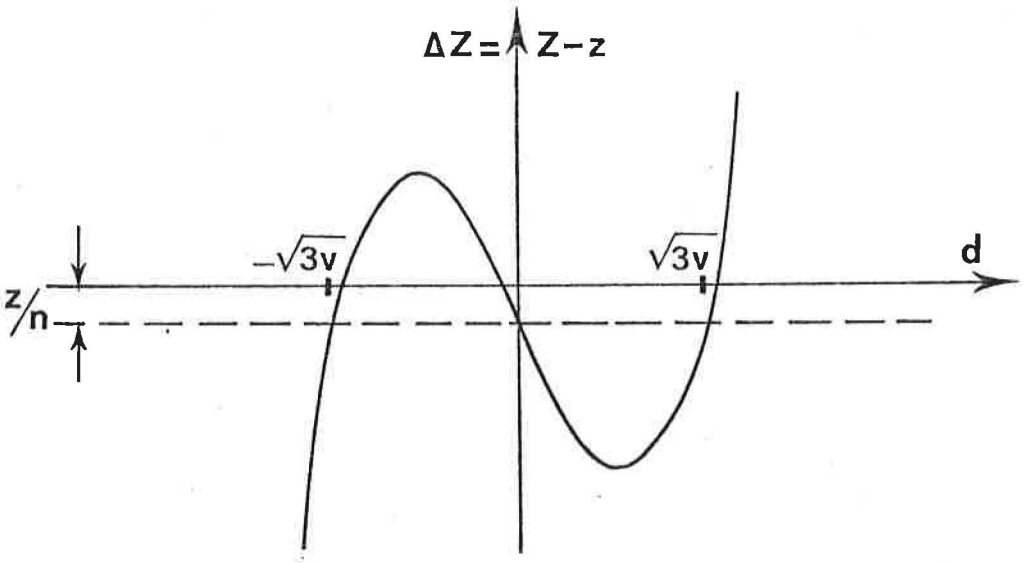

Figure 1 - Représentation schématique du changement ΔZ du troisième moment centré dû à un nouveau résultat $x_{n+1} = m + d$, pour $n \gg 1$.

La Fig. 1 explique pourquoi l'espérance simpliste d'un changement ΔZ du troisième moment centré qui aurait le même signe que d est erronnée pour toutes les valeurs x_{n+1} qui sont à l'intérieur du domaine approximativement délimité par $m \pm \sqrt{3v}$, donc pour la majorité des cas.

Les formules simples données pour la mise à jour des trois premiers moments peuvent rendre service s'il s'agit d'incorporer rapidement une nouvelle mesure qui s'ajoute à un lot déjà étudié. En principe, le procédé peut être répété s'il s'agit de tenir compte de plusieurs nouveaux résultats, mais en pratique le nombre de ceux-ci dépassera rarement deux. La méthode est également applicable s'il convient de

réduire l'effectif d'une (ou de deux) unités, avec de petits changements évidents dans les formules respectives.

Pour un nombre plus élevé de nouvelles données, cependant, l'emploi itératif devient peu pratique. La combinaison de deux échantillons, dont chacun comprend au moins trois mesures, est traitée dans une autre note (WPN-215).

Références

[1] J. Guey: "Symétrie ou dissymétrie d'émission?", dans Comptes rendus de la Table Ronde "Numismatique et Statistique", PACT (Conseil de l'Europe, Strasbourg), à paraître

[2] H. Cramér: "Mathematical Methods of Statistics" (Princeton University Press, Princeton, 1946), chapter 27

(Mai 1980)