Rapport BIPM-84/7

A propos d'un abus de poids statistiques

par Jörg W. Müller

Bureau International des Poids et Mesures, F-92310 Sèvres

Abstract

For results which follow a Poisson distribution it is frequently assumed that the estimated variance corresponding to an observation can be set equal to the measured value. It is shown, however, that by using these values for the determination of statistical weights, the ensuing mean value is systematically biased (essentially by one unit), whereas the corresponding variance remains practically unaffected.

1. Introduction

Il est généralement admis que l'utilisation de poids statistiques, supposés connus, améliore les estimations et qu'elle est donc à recommander. Leur emploi permet de tenir compte de la différence de qualité ou de fiabilité que l'on peut attribuer aux résultats de mesure dont on veut tirer le meilleur parti possible.

Pour déterminer les poids on se sert normalement des écarts-types respectifs obtenus à partir de mesures répétées. Ainsi, en disposant de n résultats de la forme

$$x_j = \bar{x}_j \pm s_j, \qquad \text{avec } j = 1, 2, \ldots, n , \tag{1}$$

on choisit habituellement pour leurs poids

$$g_j = c/s_j^2 , \tag{2}$$

où c est une constante arbitraire (par exemple c = 1).

Ce choix est normalement justifié par l'observation suivante: à la valeur moyenne $\bar{y}$ de m résultats y_k, ayant tous la même variance σ_o^2, correspond la variance σ_o^2/m, ce qui permet d'interpréter le poids statistique (2) comme une grandeur qui correspond à un nombre "effectif" de mesures supposées "élémentaires". Ainsi, on propose comme moyenne pondérée des résultats (1) la valeur

$$\bar{x} = \frac{\sum_j g_j \bar{x}_j}{\sum_j g_j} , \text{ avec } g_j = s_j^{-2} . \tag{3}$$

En pratique, cependant, le résultat d'une telle pondération est souvent loin d'être satisfaisant. Ceci est tout d'abord le cas si l'on ne dispose pas d'un ensemble homogène d'écarts-types, comme cela se produit habituellement pour des résultats de provenances très diverses (par exemple lors d'une comparaison internationale). Mais il existe d'autres cas où l'origine des ennuis est moins évidente, et cela est vrai, en particulier, si les incertitudes sont corrélées. Dans de telles situations il est souvent prudent - en l'absence d'une théorie fiable, susceptible de nous renseigner sur le procédé correct à suivre - de renoncer à l'emploi de poids statistiques.

2. Cas d'une distribution de Poisson

Il arrive souvent qu'on soit amené à observer des événements dont le nombre est supposé suivre la loi de Poisson. Dans ce cas, le résultat x d'une mesure est un nombre naturel (0, 1, 2, ..., k, ...) et la probabilité d'observer (dans un intervalle de temps fixe) x = k événements est alors donnée par la formule

$$P_k = \frac{\mu^k}{k!} e^{-\mu} , \qquad (4)$$

où μ est une constante positive, mais inconnue. Or, on sait que pour un processus de Poisson l'espérance mathématique E(x), ainsi que la variance V(x), sont égales à μ. L'expérimentateur en déduit habituellement que pour une mesure $x_j = k$ on peut supposer que l'écart-type est approximativement donné par $s_j \cong \sqrt{k}$, ce qui donne pour le poids statistique $g_j \cong 1/k$. Il semble qu'on fasse souvent ce raisonnement, comme le montrent de nombreuses représentations graphiques.

Supposons maintenant que l'on s'intéresse à la détermination de μ. Pour son estimation on prendra la valeur moyenne des mesures individuelles et, puisqu'on dispose également de poids statistiques approximatifs, on sera amené à former

$$E(x) = \frac{\sum g_k \, k \, P_k}{\sum g_k \, P_k} \equiv \overline{x} . \qquad (5)$$

De manière analogue, on déterminera le deuxième moment

$$E(x^2) = \frac{\sum g_k \, k^2 \, P_k}{\sum g_k \, P_k} \equiv \overline{x^2} . \qquad (6)$$

Puisque la variance est définie par

$$V(x) = E(x^2) - E^2(x) ,$$

on formera

$$\overline{x^2} - \overline{x}^2 \equiv \sigma^2(x) . \tag{7}$$

Dans ce qui suit on va essayer de déterminer ces deux premiers moments, en admettant que k suive la loi (4).

La plupart des sommes à évaluer peuvent être ramenées au type suivant (avec r = 0, 1, 2, ...)

$$m_r = \sum_{k=0}^{\infty} k^r P_k = e^{-\mu} \sum_k \frac{k^r \mu^k}{k!} . \tag{8a}$$

Il s'agit donc de moments ordinaires d'ordre r d'une distribution de Poisson. Or, ceux-ci sont bien connus et s'expriment (voir par exemple [1]) à l'aide des nombres de Stirling de seconde espèce, par la formule

$$m_r = \sum_{j=1}^{r} S(r,j) \mu^j . \tag{8b}$$

Toutes les expressions du type (8) peuvent donc être supposées connues, et en particulier on a

$$\begin{aligned} m_o &= 1 , \\ m_1 &= \mu , \\ m_2 &= \mu + \mu^2 . \end{aligned} \tag{8c}$$

Pour disposer dès le début d'une certaine vue d'ensemble sur les résultats auxquels on peut s'attendre, nous avons calculé numériquement, à l'aide d'un ordinateur, les sommes intervenant dans les relations (5) à (7). Ce calcul a été effectué pour deux choix différents des poids g_k, comme on l'expliquera plus loin en détail. Les résultats ainsi obtenus sont rassemblés dans le tableau 1 et suggèrent nettement pour les moments l'existence de simples valeurs limites (pour $\mu \gg 1$). Il nous sera possible d'expliquer cette observation par une évaluation exacte des sommes en question.

μ	$\overline{x}$ avec		$\overline{x^2}$ avec		$\sigma^2(x)$ avec	
	1^g	2^g	1^g	2^g	1^g	2^g
1	1,304	0,582	2,063	1,000	0,363	0,661
2	1,734	1,313	4,012	3,313	1,004	1,589
3	2,311	2,157	7,297	7,314	1,955	2,661
4	3,034	3,075	12,361	13,224	3,158	3,771
5	3,879	4,034	19,529	21,136	4,479	4,863
6	4,813	5,015	28,947	31,075	5,786	5,925
7	5,798	6,006	40,620	43,038	7,008	6,962
8	6,808	7,003	54,481	57,019	8,134	7,981
9	7,827	8,001	70,455	73,009	9,188	8,991
10	8,848	9,000	88,483	91,004	10,198	9,996
11	9,866	10,000	108,531	111,002	11,187	10,998
12	10,882	11,000	130,583	133,001	12,169	11,999
13	11,895	12,000	154,630	157,000	13,149	13,000
14	12,905	13,000	180,669	183,000	14,132	14,000
15	13,914	14,000	208,703	211,000	15,117	15,000
16	14,921	15,000	238,730	241,000	16,105	16,000
17	15,927	16,000	270,754	273,000	17,094	17,000
18	16,932	17,000	304,773	307,000	18,086	18,000
19	17,936	18,000	340,790	343,000	19,079	19,000
20	18,940	19,000	378,804	381,000	20,073	20,000
25	23,954	24,000	598,853	601,000	25,053	25,000
30	28,963	29,000	868,882	871,000	30,042	30,000
35	33,969	34,000	1188,902	1191,000	35,035	35,000
40	38,973	39,000	1558,916	1561,000	40,029	40,000
45	43,976	44,000	1978,926	1981,000	45,026	45,000
50	48,979	49,000	2448,934	2451,000	50,023	50,000
valeurs limites	$\mu - 1$		$\mu(\mu - 1) \mp 1$		μ	

Tableau 1 - Résultats d'une évaluation directe des deux premiers moments d'une variable aléatoire x qui est supposée suivre une distribution de Poisson, avec deux choix différents pour les poids ($_1g$ ou $_2g$; voir texte). Les valeurs limites sont pratiquement atteintes avec $\mu \gtrsim 4$ pour $\overline{x}$ et avec $\mu \gtrsim 7$ pour $\sigma^2(x)$.

3. Premier choix des poids

Si l'on choisit pour le résultat k le poids $g_k = 1/k$, on rencontre immédiatement une difficulté pour l'observation k = 0. Elle peut être surmontée (bien qu'un peu artificiellement) en excluant ce cas, c'est-à-dire en ne sommant que sur les valeurs de k à partir de 1. Pour $\mu \gg 1$ le problème disparaît pratiquement puisque P_o est alors négligeable.

Si l'on procède de cette manière, il nous reste à évaluer pour les deux premiers moments, d'après (5) et (6), les expressions

$$ {}_1\overline{x} = \frac{\sum_1^\infty \frac{k}{k}\frac{\mu^k}{k!}e^{-\mu}}{\sum_1^\infty \frac{1}{k}\frac{\mu^k}{k!}e^{-\mu}} \equiv \frac{e^\mu - 1}{A} \quad \text{et} \tag{9} $$

$$ {}_1\overline{x^2} = \frac{\sum_1^\infty \frac{k^2}{k}\frac{\mu^k}{k!}e^{-\mu}}{\sum_1^\infty \frac{1}{k}\frac{\mu^k}{k!}e^{-\mu}} \equiv \frac{\mu\, e^\mu}{A}, \tag{10} $$

où la somme

$$ A = \sum_{k=1}^\infty \frac{\mu^k}{k\,k!} \tag{11} $$

est difficile à déterminer. On reviendra sur ce point plus tard (section 5).

4. Deuxième choix des poids

Pour éviter le problème précédent avec k = 0, on peut essayer de modifier légèrement les poids, par exemple en choisissant

$$ {}_2g_k = 1/(k+1). $$

On s'attendra alors, au moins pour $\mu \gg 1$, à peu de changements dus à cette modification. D'autre part, s'il arrive que les sommes correspondantes puissent être évaluées rigoureusement, on disposera pour ce domaine de résultats fiables.

Les nouveaux poids donnent lieu aux expressions qui suivent. D'abord, on a pour le premier moment

$$\overline{2^x} = \frac{\sum_0^\infty \frac{k}{k+1}\frac{\mu^k}{k!}e^{-\mu}}{\sum_0^\infty \frac{1}{k+1}\frac{\mu^k}{k!}e^{-\mu}} \equiv \frac{B}{C}, \qquad (12)$$

avec $B = \sum_1^\infty \frac{k\,\mu^k}{(k+1)!}$ et $C = \sum_0^\infty \frac{\mu^k}{(k+1)!}$.

On trouve facilement que

$$C = \frac{1}{\mu}\sum_{k=1}^\infty \frac{\mu^k}{k!} = \frac{1}{\mu}(e^\mu - 1). \qquad (13)$$

Par contre, l'évaluation de B est un peu plus difficile, mais on peut, par exemple, procéder comme suit:

$$B = \frac{1}{\mu}\sum_1^\infty \frac{k\,\mu^{k+1}}{(k+1)!} = \frac{1}{\mu}\sum_2^\infty \frac{(k-1)\,\mu^k}{k!}$$

$$= \frac{1}{\mu}\left[\sum_0^\infty \frac{(k-1)\,\mu^k}{k!} - \left\{\frac{(-1)\,\mu^0}{0!} + \frac{0\,\mu^1}{1!}\right\}\right]$$

$$= \frac{1}{\mu}\left[\sum_0^\infty \frac{k\,\mu^k}{k!} - \sum_0^\infty \frac{\mu^k}{k!} + 1\right] = \frac{1}{\mu}\left[e^\mu(\mu-1) + 1\right]. \qquad (14)$$

Une autre possibilité de déterminer B m'a été indiquée aimablement par Mme M. Boutillon qui utilise l'astuce suivante. En remarquant que

$$\frac{d}{d\mu}\left[\frac{k\,\mu^{k+1}}{(k+1)!}\right] = \frac{k(k+1)\,\mu^k}{(k+1)!} = \frac{\mu^k}{(k-1)!},$$

on peut aussi écrire (en intervertissant somme et intégration)

$$B = \frac{1}{\mu}\sum_1^\infty \int \frac{\mu^k}{(k-1)!}\,d\mu = \frac{1}{\mu}\int\mu \sum_1^\infty \frac{\mu^{k-1}}{(k-1)!}\,d\mu$$

$$= \frac{1}{\mu}\int\mu\,d\mu \sum_0^\infty \frac{\mu^k}{k!} = \frac{1}{\mu}\int_0^\mu \mu\, e^\mu\,d\mu.$$

Or, cette intégrale est bien connue et l'on obtient

$$B = \frac{1}{\mu}\left[e^{\mu}(\mu-1) + 1\right], \qquad \text{comme précédemment.}$$

Cela nous permet donc d'écrire

$$2^{\overline{x}} = \frac{B}{C} = \frac{e^{\mu}(\mu-1) + 1}{e^{\mu} - 1} = \frac{\mu}{1 - e^{-\mu}} - 1 , \tag{15}$$

d'où l'on tire, pour $\mu \gg 1$,

$$2^{\overline{x}} \to \mu - 1 . \tag{16}$$

Pour le deuxième moment, on a maintenant

$$2^{\overline{x^2}} = \frac{\sum_{0}^{\infty} \frac{k^2}{k+1} \frac{\mu^k}{k!} e^{-\mu}}{\sum_{0}^{\infty} \frac{1}{k+1} \frac{\mu^k}{k!} e^{-\mu}} \equiv \frac{D}{C} , \tag{17}$$

où C a été déterminé plus haut et

$$D = \sum_{0}^{\infty} \frac{k^2\ \mu^k}{(k+1)!} = \frac{1}{\mu} \sum_{1}^{\infty} \frac{(k-1)^2\ \mu^k}{k!} ,$$

d'où (en appliquant (4) et (8))

$$D\ e^{-\mu} = \frac{1}{\mu} \sum_{1}^{\infty} (k^2 - 2k + 1)\ P_k$$

$$= \mu - 1 + \frac{1}{\mu}(1 - e^{-\mu}) . \tag{18}$$

Il en résulte pour le deuxième moment

$$2^{\overline{x^2}} = \frac{\mu - 1 + \frac{1}{\mu}(1 - e^{-\mu})}{\frac{1}{\mu}(1 - e^{-\mu})} = \frac{\mu(\mu-1)}{1 - e^{-\mu}} + 1 . \tag{19}$$

Pour $\mu \gg 1$ on a donc

$$2\overline{x^2} \to \mu(\mu - 1) + 1 \ . \tag{20}$$

Enfin, il en découle, d'après (7), pour la variance, en utilisant (15) et (19),

$$\begin{aligned} 2\sigma^2(x) &= \frac{\mu(\mu-1)}{1 - e^{-\mu}} + 1 - \left[\frac{\mu}{1 - e^{-\mu}} - 1\right]^2 \\ &= \frac{\mu^2 - \mu + 2\mu}{1 - e^{-\mu}} - \frac{\mu^2}{(1 - e^{-\mu})^2} \\ &= \frac{\mu}{1 - e^{-\mu}} \left(\mu + 1 - \frac{\mu}{1 - e^{-\mu}}\right) \ . \end{aligned} \tag{21}$$

Il s'ensuit pour $\mu \gg 1$ que

$$2\sigma^2(x) \to \mu \ . \tag{22}$$

Les relations (15), (19) et (21) peuvent être utilisées pour vérifier les valeurs numériques du tableau 1, obtenues par un calcul direct.

De plus, les relations asymptotiques (16), (20) et (22) expliquent bien les valeurs limites suggérées dans ce tableau.

5. Retour au premier choix

Les évaluations exactes pour le premier choix des poids n'ont pas pu se faire, car la détermination de la somme (11) a résisté à nos tentatives. On peut maintenant se rendre compte qu'il y a de bonnes raisons pour cela.

En utilisant l'astuce due à Mme Boutillon, décrite auparavant, c'est-à-dire

$$\frac{d}{d\mu}\left[\frac{\mu^k}{k\ k!}\right] = \frac{\mu^{k-1}}{k!} \ ,$$

on peut aussi écrire pour (11)

$$A = \sum_1^\infty \int \frac{\mu^{k-1}}{k!}\, d\mu = \int_\mu^1 \sum_1^\infty \frac{\mu^k}{k!}\, d\mu = \int \frac{e^\mu - 1}{\mu}\, d\mu \ .$$

Or, en utilisant une relation connue (voir [2], par exemple), on a, pour $\mu > 0$,

$$A = \int_0^{\mu} \frac{e^x - 1}{x}\, dx = Ei(\mu) - \ln \mu - \gamma , \qquad (23)$$

où $Ei(\mu)$ est une intégrale exponentielle et $\gamma \cong 0{,}577\ 216$ est la constante d'Euler.

A l'aide de tables numériques pour $Ei(\mu)$ - ou des fonctions qui y sont associées [2] - on peut dresser le tableau 2.

μ	$Ei(\mu)$	A
1	1,895 1	1,317 9
2	4,964 2	3,683 8
3	9,933 8	8,258 0
4	16,630 9	17,667 4
5	40,185 3	37,998 6
6	85,989 8	83,620 8
7	191,505	188,982
8	440,380	437,723
9	1 037,88	1 035,10
10	2 492,23	2 489,35

Tableau 2 - Quelques valeurs numériques pour l'intégrale exponentielle et la somme A donnée par (23).

A l'aide de ces valeurs de A il est facile de vérifier (au moins pour $\mu \leqslant 10$) que l'application des expressions (9) et (10) permet de retrouver les valeurs correspondantes dans le tableau 1.

Il nous reste, cependant, à comprendre leur comportement asymptotique. Pour y arriver, on peut utiliser la série semi-convergente [3]

$$Ei(\mu) = e^{\mu} \sum_{j=0}^{\infty} \frac{j!}{\mu^{j+1}} \qquad (24)$$

de laquelle on peut tirer, pour $\mu \gg 1$, l'approximation

$$A \cong \frac{e^{\mu}}{\mu}\left(1 + \frac{1}{\mu} + \frac{2}{\mu^2}\right) \cong \frac{e^{\mu}}{\mu - 1 - 1/\mu} .$$

Cela permet donc de parvenir, avec (9) et (10), pour les deux premiers moments, aux valeurs limites

$$ {}_1\overline{x} = \frac{e^{\mu} - 1}{A} \to \left(\frac{e^{\mu} - 1}{e^{\mu}}\right)(\mu - 1 - 1/\mu) \cong \mu - 1 , \qquad (25) $$

et

$$ {}_1\overline{x^2} = \frac{\mu\, e^{\mu}}{A} \to \frac{\mu\, e^{\mu}}{e^{\mu}}(\mu - 1 - 1/\mu) \cong \mu(\mu - 1) - 1 , \qquad (26) $$

et pour la variance

$$ {}_1\sigma^2(x) = {}_1\overline{x^2} - {}_1\overline{x}^2 \to \mu(\mu - 1) - 1 - (\mu - 1 - 1/\mu)^2 $$

$$ \cong \mu^2 - \mu - 1 - (\mu^2 - 2\mu - 1) = \mu , \qquad (27) $$

toujours dans le cas $\mu \gg 1$. Les relations (25) à (27) expliquent bien le comportement asymptotique des moments que nous avons observé dans le tableau 1 pour le premier choix des poids statistiques.

6. Conclusions

L'exemple que nous venons de discuter de façon assez détaillée ne présente pas beaucoup d'importance en lui-même, mais il permet d'illustrer quelques points qu'il conviendra peut-être de retenir.

Tout d'abord, il nous montre que l'application de poids statistiques est pleine de pièges car, même dans le cas étudié ci-dessus, où l'on dispose d'estimations homogènes et non corrélées des variances, leur application systématique sous forme de poids peut induire en erreur. Puis, il nous rappelle que le bon sens ne peut pas être remplacé par un excès de zèle. Car finalement l'origine de l'erreur réside dans un mauvais raisonnement qui oublie que les poids devraient se fonder sur la meilleure estimation de la variance dont on dispose, et celle-ci, s'appuyant sur l'ensemble des observations, est évidemment la même pour tous les résultats observés. Dans cette optique, l'emploi de poids se révèle superflu et l'on a tout simplement

$$ E(x) = \sum_k k\, P_k = \mu , $$

comme on pouvait s'y attendre.

Il n'est pas surprenant que l'emploi (non justifié) de poids qui favorisent les valeurs basses des observations donne lieu à un premier moment $\overline{x}$ qui est inférieur à μ; par contre, le décalage constant d'une unité est moins évident. De plus, le fait que la pondération n'a pratiquement aucune influence sur la variance n'était certainement pas prévisible.

Le résultat de cette petite étude a déjà été mentionné dans [4], où il nous a servi d'illustration. Une version très abrégée en a été présentée, le 13 novembre 1984, lors d'une conférence intitulée "Faut-il toujours pondérer les résultats de mesures?" au personnel du BIPM.

Références

[1] J.W. Müller: "Quelques remarques sur une 'double' distribution de Poisson", BIPM WPN-224 (1982)

[2] "Handbook of Mathematical Functions" (ed. by M. Abramowitz and I.A. Stegun) NBS, AMS 55 (GPO, Washington, 1964), p. 230

[3] E. Jahnke, F. Emde: "Tables of Functions" (Dover, New York, 1945^4), p. 3

[4] J.W. Müller: "The treatment of measurement uncertainties", in Proceedings of IMEKO Training Course, Vol. 1 (Seibersdorf, 1984).

(Novembre 1984)