W. GOSIEWSKI.

Sur un moyen de corriger dans le calcul de la moyenne arithmétique les valeurs suspectes de l'inconnue. [1])

„On accroîtrait beaucoup la précision d'une mesure, si l'on pouvait découvrir les évaluations les meilleures et les adopter de préférence à la moyenne générale; il en serait de même si l'on pouvait prendre la moyenne des observations les meilleures. Elles sont malheuresement confondues avec les autres sans que rien les décèle avec certitude".

J. Bertrand. Cal. des prob. p. 218.

Dans la théorie des probabilités, quand il s'agit de déterminer une inconnue, on obtient, en principe, deux réponses différentes: l'une d'elles — c'est l'espérance mathématique de l'inconnue, nommée aussi sa valeur moyenne ou probable; l'autre — c'est la valeur la plus probable de l'inconnue. Par conséquent, plus ces réponses s'approchent l'une de l'autre, plus le cas est heureux, et il est le plus heureux, si l'une est égale à l'autre.

Tel est précisement le cas qui présente la moyenne arithmétique $\bar{x}$ des valeurs données de l'inconnue x, obtenues dans les mêmes conditions et avec le même soin. Car sous cette restriction la moyenne x est l'espérance mathématique de l'inconnue x, et aussi elle est sa valeur la plus probable, c'est à dire, celle qui parmis toutes les valeurs eventuelles de cette inconnue, alors possibles, possède le plus de chances d'être égale à elle.

---

[1]) Travail communiqué par M. le Prof. L. Silberstein au Congrès de la Società Italiana per il progresso delle scienze, le 16 décembre 1910. Voir „Atti della Società Italiana etc. Quarta riunione, Napoli. Dicembre 1910; Roma 1911 p. 734. Cf. l'article de l'auteur imprimé dans les Comptes rendus de la Societé scientifique de Varsovie, 1910 pp. 151—161.

Mais ni l'un, ni l'autre, ni même l'un et l'autre à la fois ne préjuge point la question de la grandeur de l'erreur eventuelle $x-\bar{x}$ de la moyenne arithmétique x. Et cependant dans la pratique cette question est la plus importante. Car, en effet, quel profit a-t-on de la conviction que $\bar{x}$ est l'espérance mathématique de x, ou bien qu'il possède le plus de chances d'être égal à x, si l'erreur $x-\bar{x}$ est assez grande?

Par l'erreur dont on parle on comprend, comme on sait, sa valeur absolue $|x-\bar{x}|$. Donc encore il s'agit de déterminer l'inconnue $|x-\bar{x}|$, c'est à dire, son espérance mathématique et sa valeur la plus probable. La seconde étant essentiellement égale à zéro, tandis que la première en diffère, on a ainsi à faire avec le cas moins heureux qu' auparavant.

Cependant et alors même on peut diminuer à un certain degré ladite différence, ce qui constitue le moyen de corriger la moyenne arithmétique.

En effet, la restriction enoncée que les valeurs données de l'inconnue x sont obtenues dans les mêmes conditions et avec le même soin, exprime seulement qu'on a fait tout pour qu'elles puissent être également bonnes pour notre but, mais point du tout pour qu'elles soient forcement telles; car il n'y a pas de moyen de s'assurer contre les caprices du hasard.

Malgré donc tous les efforts entrepris dans ce sens, quelques-unes parmis les valeurs données de l'inconnue x peuvent être suspectes, et alors il faut savoir les distinguer des autres, ainsi que les corriger.

Ce problème s'appuie sur un autre problème dont voici l'enoncé et la résolution succincte.

Soient $x_1, x_2, \ldots, x_n$ les valeurs de l'inconnue x, données par n mesures successives, faites avec la même précision, mais qui n'est pas connue, et cherchons à déterminer l'expression de l'inconnue x en fonction des $x_1, x_2, \ldots, x_n$, ainsi que sa probabilité.

Désignons par h la précision, exprimons avec Gauss par $he^{-h^2\Delta^2}d\Delta/\sqrt{\pi}$ la probabilité de l'erreur Δ, et posons

$$\Delta_i = x - x_i, \qquad \psi = \frac{h^n e^{-h^2 \Sigma_i \Delta_i^2}\, d\Delta_1\, d\Delta_2 \ldots d\Delta_n}{\sqrt{\pi^n}},$$

$$(i = 1, 2, \ldots, n)$$

ψ étant la probabilité du concours des erreurs $\Delta_1, \Delta_2, \ldots, \Delta_n$.

Alors, h n'étant pas connu, la probabilité q que la valeur de l'inconnue soit égale à x, s'exprime, en vertu du principe de Bayes, par la formule

$$q = \frac{dx \int_0^\infty \psi\, dh}{\int_0^\infty dh \int_{-\infty}^\infty \psi\, dx}.$$

Cette formule peut être très facilement achevée, car si l'on nomme $\bar{x}$ la moyenne arithmétique des valeurs $x_1, x_2, \ldots, x_n$, et si l'on pose

$$(1) \qquad x = \bar{x} + \eta \sqrt{\frac{1}{n} \Sigma_i (\bar{x} - x_i)^2},$$

$$a = h \sqrt{\Sigma_i (\bar{x} - x_i)^2},$$

η et a étant les variables nouvelles, on aura tout d'abord

$$\psi = \frac{a^n e^{-a^2(1+\eta^2)}\, d\Delta_1\, d\Delta_2 \ldots d\Delta_n}{\sqrt{\pi^n}\, (\sqrt{\Sigma_i (\bar{x} - x_i)^2})^n}, \qquad dx\, dh = \frac{d\eta\, da}{\sqrt{n}},$$

et par conséquent

$$q = \frac{d\eta \int_0^\infty a^n e^{-a^2(1+\eta^2)}\, da}{\int_0^\infty a^n e^{-a^2}\, da \int_{-\infty}^\infty e^{-a^2\eta^2}\, d\eta}.$$

Et comme

$$\int_{-\infty}^\infty e^{-a^2\eta^2}\, d\eta = \frac{\sqrt{\pi}}{a}, \qquad \int_0^\infty a^{n-1} e^{-a^2}\, da = \frac{1}{2}\Gamma\left(\frac{n}{2}\right),$$

$$\int_0^\infty a^n e^{-a^2(1+\eta^2)}\, da = \frac{\Gamma\left(\frac{n+1}{2}\right)}{2(1+\eta^2)^{\frac{n+1}{2}}}$$

on a définitivement

$$(2) \qquad q = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{d\eta}{(1+\eta^2)^{\frac{n+1}{2}}}.$$

Ainsi l'expression cherchée de l'inconnue x en fonction

de $x_1, x_2, \ldots, x_n$, et sa probabilité q, sont données respectivement par les formules (1) et (2).[1]

§ 1.

Occupons nous maintenant de notre problème principal.

Puisque les valeurs $x_1, x_2, \ldots, x_n$ sont données, et leurs poids sont, par hypothèse, les mêmes, par conséquent il existe une précision actuelle, p. e. h, avec laquelle chacune d'elles a été obtenue.

À cette condition, c'est à dire, que la précision h soit constante et actuelle, on a d'après Gauss

(3) $$x = \bar{x} + \frac{t}{h},$$

avec la probabilité

(4) $$p = \frac{\sqrt{n}\, e^{-nt^2}\, dt}{\sqrt{\pi}}.$$

Par conséquent l'espérance mathématique de $|x - \bar{x}|$, et que nous désignons alors par E. M. $|x - \bar{x}|_h$, a l'expression

(5) $$\text{E. M. } |x - \bar{x}|_h = \frac{1}{\sqrt{\pi}\, h \sqrt{n}}.$$

Mais au fond la précision h ne peut être jamais connue, car pour l'obtenir, il faudrait que le même observateur et avec le même instrument accomplisse un nombre infiniment grand des épreuves, ce qui est impossible.

Pour exprimer donc l'espérance mathématique de $|x - \bar{x}|$, et que désignons maintenant par E. M. $|x - \bar{x}|$, il vaut mieux d'employer les formules (1) et (2). On a alors

(6) $$\text{E. M. } |x - \bar{x}| = \xi_n(\bar{x}),$$

où

(7) $$\xi_n(\bar{x}) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\, \Gamma\left(\frac{n}{2}\right)} \sqrt{\frac{1}{n} \Sigma_i (\bar{x} - x_i)^2}.$$

Cela posé, imaginons nous qu'à la suite des valeurs $x_1, x_2, \ldots, x_n$ se joint une nouvelle valeur y, obtenue avec la même précision h. Alors les équations (5) et (6) deviennent respectivement

[1]) J'ai obtenu ce résultat pour la première fois, et je l'ai donné dans mon ouvrage: „Zasady Rachunku prawdopodobieństwa" (Éléments du calcul des probabilités). Varsovie. 1906

(8) $$\text{E. M. } \left| x - \frac{n\bar{x} + y}{n+1} \right|_h = \frac{1}{\sqrt{\pi}\, h \sqrt{n+1}},$$

(9) $$\text{E. M. } \left| x - \frac{n\bar{x} + y}{n+1} \right| = \xi_{n+1}\left(\frac{n\bar{x} + y}{n+1}\right),$$

et vu l'identité

$$\sum_i \left(\frac{n\bar{x} + y}{n+1} - x_i\right)^2 + \left(\frac{n\bar{x} + y}{n+1} - y\right)^2 = \sum_i (\bar{x} - x_i)^2 + \frac{n(\bar{x} - y)^2}{n+1},$$

$$(i = 1, 2, \ldots, n)$$

on aura d'après (7)

(10) $$\xi_{n+1}\left(\frac{n\bar{x} + y}{n+1}\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\, \Gamma\left(\frac{n+1}{2}\right)} \sqrt{\frac{1}{n+1}\left\{\sum_i (x - \bar{x}_i)^2 + \frac{n(\bar{x} - y)^2}{n+1}\right\}}.$$

Les équations (8) et (5) donnent l'inégalité unique

(11) $$\text{E. M. } \left| x - \frac{n\bar{x} + y}{n+1} \right|_h < \text{E. M. } |x - \bar{x}|_h,$$

et l'on voit par là que l'espérance mathématique de l'erreur de la moyenne arithmétique diminue, si à la suite des valeurs données de l'inconnue se joint une nouvelle valeur, obtenue avec la même précision que chacune d'elle.

Cependant les équations (9) et (6) donnent alors alternativement:

1) ou bien l'inégalité

(12) $$\text{E. M. } \left| x - \frac{n\bar{x} + y}{n+1} \right| < \text{E. M. } |x - \bar{x}|$$

si l'on a

(13) $$\xi_{n+1}\left(\frac{n\bar{x} + y}{n+1}\right) < \xi_n(\bar{x});$$

2) ou bien l'inégalité

(14) $$\text{E. M. } \left| x - \frac{n\bar{x} + y}{n+1} \right| \geqslant \text{E. M. } |x - \bar{x}|$$

si l'on a

(15) $$\xi_{n+1}\left(\frac{n\bar{x} + y}{n+1}\right) \geqslant \xi_n(\bar{x}).$$

Qu'est ce que cela peut signifier?

Dans le prémier cas l'inégalité (12) approuve l'inégalité (11), et par conséquent il est permis de présumer que la valeur y est alors indubitablement

bonne; dans le second cas l'inégalité (14) conteste l'inégalité (11), et par conséquent il est permis de présumer que la valeur y est alors douteusement bonne, c'est à dire, suspecte.

Ainsi nous sommes parvenus à définir la valeur y — censée bonne, et la valeur y — censée suspecte: la première exige l'inégalité (13), la seconde — l'inégalité (15).

§ 2.

Les inégalités (13) et (15) s'excluant, on peut se borner à ne considérer qu' une d'elles, p. e. l'inégalité (13). Elle prend en vertu des équations (10) et (7) la forme très simple

$$(\bar{x}-y)^2 < \rho^2, \tag{16}$$

où l' on a posé

$$\rho^2 = \frac{n+1}{n}\left(\frac{n+1}{n}\alpha^2_n - 1\right)\sum_i (\bar{x}-x_i)^2,$$

$$\alpha_n = \frac{\Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)^2}.$$

Mais comme

$$\frac{n}{n-1} < \alpha^2_n < \frac{n-1}{n-2},$$

et par suite

$$\alpha^2_n = \frac{n-\varepsilon}{n-1-\varepsilon_n}, \qquad 0 < \varepsilon_n < 1,$$

on peut aussi écrire

$$\rho^2 = \frac{2+\frac{2-\varepsilon_n}{n}-\frac{\varepsilon_n}{n^2}}{n-1-\varepsilon_n}\sum_i (\bar{x}-x_i)^2.$$

De cette manière pour $n=\infty$ on a simultanement: $\lim \bar{x} = x$, et

$$\lim \rho^2 = \left(\frac{2}{n}\sum_i (\bar{x}-x_i)^2\right)_{n=\infty} = \frac{1}{h^2},$$

h étant la précision actuelle dans le cas considéré, la même dont on a tant parlé au § précédent.

Par conséquent l' inégalité (16) devient alors la suivante:

$$h^2(x-y)^2 < 1, \tag{16'}$$

dont la probabilité est égale, comme on sait, à

$$\frac{2}{\sqrt{\pi}}\int_0^1 e^{-t^2}\,dt = \theta(1) = 0{,}8427008.$$

On conclue de là que parmis n valeurs $y = x_1, x_2, \ldots, x_n$ il y en a probablement m qui sont bonnes, et $n-m$ qui sont suspectes, m étant le nombre entier le plus approché de $n\theta(1)$, de sorte que l'on a grossièrement $m = 5n/6$, $n-m = n/6$.[1]

Il faut cependant faire démêler explicitement les valeurs suspectes de valeurs bonnes parmis les données $x_1, x_2, \ldots, x_n$, et de les faire remplacer par les valeurs meilleures.

§ 3.

Reprenons dans ce but l'inégalité (16), le prototype de l'inégalité (16'), et l'exprimons sous la forme

$$\bar{x}-\rho < y < \bar{x}+\rho. \tag{17}$$

Rémarquons que la valeur y, nouvelle, peut être aussi bien égale à quelques-unes des anciennes

$$x_1, \quad x_2, \ldots, \quad x_n. \tag{18}$$

Par conséquent chaque valeur $y = x_i$ de la suite (18) qui satisfait à l' inégalité (17) est censée bonne, et chacune autre de la même suite qui ne la satisfait pas, c'est à dire, qui est $\leqslant \bar{x}-\rho$ ou $\geqslant \bar{x}+\rho$, est suspecte.

Ainsi nous avons le moyen d'indiquer explicitement les valeurs relativement bonnes et celles suspectes dans la suite (18).

Mais dans la pratique on n'a pas besoin de se servir de l'inégalité (17) dont le terme ρ est difficile à calculer. Il suffit d'écrire la suite des valeurs absolues

$$|\bar{x}-x_1|, \quad |\bar{x}-x_2|, \ldots, \quad |\bar{x}-x_n|,$$

et de faire choisir parmis ses termes ceux les plus grands en nombre $n-m$, $n-m$ étant connu d'avance. Alors à chaque terme ainsi choisi correspond une valeur suspecte de l'inconnue, bien déterminée.

---

[1]) En effet, on a pour les réduites de $\theta(1)$: $^1/_1$, $^5/_6$, $^{11}/_{13}$, $^{16}/_{19}$, etc.

Supposons, pour fixer les idées, qu'en appliquant l'un ou l'autre de ces moyens nous avons trouvé que parmis les valeurs (18) celles

$$x_1, \quad x_2, \ldots, \quad x_m, \tag{19}$$

en nombre m, sont bonnes, et celles réstantes

$$x_{m+1}, \quad x_{m+2}, \ldots, \quad x_n, \tag{20}$$

en nombre $n-m$, sont suspectes.

Cela étant, la correction de la moyenne x ne présente aucune difficulté.

En effet, soit $\overline{x}_m$ la moyenne arithmétique des valeurs censées bonnes (19), et remarquons que toutes ses composantes $y = x_1, x_2, \ldots, x_m$ satisfont, par hypothèse, à l'inégalité (17). Par suite $\overline{x}_m$ lui même la satisfait aussi, et jouit ainsi de caractère d'une valeur bonne dans le cas considéré.

D'autre part, le poids de $\overline{x}_m$ étant en général m fois plus fort que celui d'une observation isolée dans le cas considéré, il est sans contestation m fois plus fort que celui de $x_{m+\lambda}$, $(\lambda = 1, 2, \ldots, n-m)$, car $x_{m+\lambda}$ est censé suspect, et les composantes de x_m sont censées bonnes.

Par conséquent $\overline{x}_m$ s'adapte à remplacer chaque $x_{m+\lambda}$ avec le poids m fois plus fort, d'où il vient

$$\overline{x}\ \text{corr.} = \frac{1}{(1+n-m)\,m}\left(\sum_{i}^{m} x_i + (n-m)\,m\,\overline{x}_m\right), \tag{21}$$

et en même temps

$$\frac{\text{poids de } \overline{x}\ \text{corr.}}{\text{poids de } \overline{x}} = \frac{(1+n-m)\,m}{n} = 1 + \frac{(n-m)(m-1)}{n} > 1. \tag{22}$$

§ 4.

L'équation (21) fait voir que l'on a définitivement

$$\overline{x}\ \text{corr.} = \frac{1}{m}\sum_{1}^{m} x_i, \tag{23}$$

ce qui prouve que pour avoir la valeur correcte de la moyenne arithmétique, il suffit tout simplement d'exclure de son calcul les valeurs suspectes.

D'ailleurs on sait que plusieurs observateurs pratiquent ladite exclusion depuis bien longtemps, sans se servir cependant d'aucune règle propre à indiquer avec la rigueur les valeurs suspectes à exclure.

Mais ce mode de procéder a été questionné par les théoriciens même éminents. On a prétendu que l'on abaisse le poids de la moyenne, si même on rejète de son calcul les valeurs suspectes. Cependant c'e n'est qu'illusoire. À la vérité on ne les rejète alors point, seulement on les remplace par les valeures meilleures, et par conséquent ou relève le poids de la moyenne.

§ 5.

Parmis les valeurs données $x_1, x_2, \ldots, x_n$ de l'inconnue x il y en a une, désignons la par z_1, qui est la plus approchée de x, et puis z_2 qui succède z_1 sous ce rapport, et puis z_3 qui succède z_2, etc. jusqu' à epuiser toutes les valeurs les meilleures de l'inconnue x parmis $x_1, x_2, \ldots, x_n$.

Or, les valeurs de l'inconnue x ainsi définies et classées

$$z_1, \quad z_2, \quad z_3, \ldots$$

ne sont point connues, non plus que leur nombre, mais J. Bertrand[1] a trouvé que leurs erreurs probables sont respectivement les suivantes:

$$\frac{1}{h(n+1)}, \quad \frac{2}{h(n+1)}, \quad \frac{3}{h(n+1)}, \ldots \tag{24}$$

h étant la précision actuelle.

D'autre part, en vertu de la relation (22), la formule (5) donne

$$\text{E. M.}\ |x - \overline{x}\ \text{corr.}|_h = \frac{1}{\sqrt{\pi}\,h\sqrt{(1+n-m)\,m}},$$

et si l'on met encore $n\,\theta(1)$ au lieu de m, on aura

$$\text{E. M.}\ |x - \overline{x}\ \text{corr.}|_h = \frac{1}{\sqrt{\pi}\,h\sqrt{an+bn^2}}, \tag{25}$$

où

$$a = \theta(1), \qquad b = \theta(1)\,[1-\theta(1)],$$

pour l'erreur probable de la moyenne corrigée.

Par conséquent, l'erreur de la moyenne corrigée étant ainsi du même ordre que les erreurs (24), la moyenne corrigée elle même paraît s'approcher de la moyenne dont parle J. Bertrand.

Wołomin.

---

[1]) l. c. p. 216 – 218.