# Limites de tolérance dans le cas d'un mélange de deux populations bivariées: Estimation de la fréquence des malclassifications[1,2]

R. Lang et A. Rieben
Institut de médecine sociale et préventive, Université de Genève

**Les mouvements de population demandent un examen systématique des phénomènes statistiques propres aux mélanges de populations. Les auteurs proposent une approche à ce problème.**

La question du mélange de populations et ses implications sur le contour de *normalité biologique* ont été longuement prises en considération lors de l'élaboration du plan de sondage pour l'«Enquête médico-anthropologique chez les enfants et adolescents de Genève» effectuée en 1972 conjointement par le Département d'anthropologie et l'Institut de médecine sociale et préventive de l'Université ainsi que le Service de santé de la jeunesse de Genève. En effet, l'un des objectifs de cette enquête est précisément la création de normes[3] à jour pour la taille et pour le poids.

Si l'on veut bien admettre, en première approximation du moins, que la taille (x) et le poids (y) d'un sujet suivent conjointement une loi de probabilité (densité) bivariée dite de *Laplace-Gauss,* la théorie statistique indique comment déterminer, dans le plan x, y, un contour elliptique $C_{0,95}$ jouissant de la propriété suivante: la probabilité pour un couple $\langle x ; y \rangle$ de mensurations d'être situé à l'intérieur de ce contour $C_{0,95}$ égale précisément 95 %; cf. [2, 3, 5].

Par analogie avec les limites de tolérance univariées (cf. [1]) au même seuil de probabilité, on peut définir le contour $C_{0,95}$ comme délimitant dans la population l'ensemble des sujets *biologiquement normaux* en ce qui concerne leurs dimensions staturo-pondérales. Graphiquement parlant, cela s'énonce ainsi: selon que le point représentatif du couple $\langle x ; y \rangle$ se situe à l'intérieur ou au contraire à l'extérieur du contour $C_{0,95}$, le sujet correspondant sera par convention jugé *biologiquement normal,* respectivement *biologiquement non-normal.*

Du point de vue calcul, l'élaboration du contour $C_{0,95}$ pour une unique population stable et non mélangée nécessite la connaissance des paramètres $\mu_x$, $\mu_y$ (moyennes), $\sigma_x$, $\sigma_y$ (écarts-types) et $\varrho_{xy}$ (coefficient de corrélation), à défaut, celle de leur estimation en cas d'échantillonnage.

Plus généralement, cependant, on se trouve en présence d'un mélange d'au moins deux populations distinctes. On doit alors élucider la question de savoir comment le contour $C_{0,95}$ établi pour un tel mélange se comporte vis-à-vis des contours relatifs aux populations composantes, notamment en ce qui concerne d'éventuels risques de jugements contradictoires («malclassifications»).

Le cas particulier discuté dans ce qui va suivre est basé sur la simulation d'un mélange, dans un rapport 6:4 ($\lambda' : \lambda''$), de deux populations bivariées du type *Laplace-Gauss* auxquelles on a assigné pour paramètres les valeurs réellement obtenues dans deux enquêtes sur de jeunes adultes, l'une suisse, l'autre italienne; cf. tableau 1. Les paramètres du mélange ont été obtenus par application des formules reproduites au tableau 2. A ces trois populations, Suisse, Italie, Mélange, sont associés trois contours $C_{0,95}$ calculés chacun dans l'hypothèse de la loi de probabilité mentionnée au début.

Tableau 1
*Valeur numérique des paramètres des deux populations d'origine et du mélange*

| | Suisse* | | Italie* | | Mélange | |
|---|---|---|---|---|---|---|
| Taille x [cm] | | | | | | |
| Moyenne | 173,9 | ($\mu_x'$) | 164,3 | ($\mu_x''$) | 170,1 | ($\tilde{\mu}_x$) |
| Ecart-type | 6,086 | ($\sigma_x'$) | 6,430 | ($\sigma_x''$) | 7,803 | ($\tilde{\sigma}_x$) |
| Poids y [kg] | | | | | | |
| Moyenne | 64,5 | ($\mu_y'$) | 57,9 | ($\mu_y''$) | 61,9 | ($\tilde{\mu}_y$) |
| Ecart-type | 7,010 | ($\sigma_y'$) | 7,040 | ($\sigma_y''$) | 7,720 | ($\tilde{\sigma}_y$) |
| Corrélation x y | | | | | | |
| Coefficient de corrélation | 0,59 | ($\varrho'$) | 0,59 | ($\varrho''$) | 0,68 | ($\tilde{\varrho}$) |
| Mélange | | | | | | |
| Proportions | 0,6 | ($\lambda'$) | 0,4 | ($\lambda''$) | .... | |

* Les auteurs remercient le Professeur *P. Moeschler* (Département d'anthropologie de l'Université de Genève) d'avoir bien voulu leur indiquer ces valeurs.

La superposition de ces 3 contours découpe le plan x, y en huit régions théoriquement. On peut désigner par M (respectivement $\overline{M}$) l'intérieur (resp. l'extérieur) du contour relatif au mélange, de même par S, $\overline{S}$ et I, $\overline{I}$ leurs équivalents suisses et italiens, enfin par MSI, M$\overline{S}$I, MS$\overline{I}$, etc., la région commune à M, S, I, à M, $\overline{S}$, I, à M, S, $\overline{I}$, etc. Ainsi, les huit régions correspondent aux combinaisons MSI, M$\overline{S}$I, MS$\overline{I}$, M$\overline{S}$$\overline{I}$, $\overline{M}$SI, $\overline{M}$$\overline{S}$I, $\overline{M}$S$\overline{I}$ et $\overline{M}$$\overline{S}$$\overline{I}$; cf. figure 1. Dire qu'un couple $\langle x ; y \rangle$ se situe dans la région $\overline{M}$S$\overline{I}$ par exemple, revient à juger que le sujet en question est à la fois

[1] Basé sur une présentation lors des Journées d'exposés scientifiques de la Société suisse de médecine sociale et préventive, Bâle, 21—22 juin 1974.

[2] Travail effectué à l'aide du subside de recherche No 4.20.70 du Fonds national suisse de la recherche scientifique (Commission de recherche pour la santé).

[3] Les normes en question paraîtront, d'une part sous la forme usuelle de tables et courbes de percentiles en fonction de l'âge (4 à 20 ans par paliers de 6 mois; garçons et filles à part) pour la taille et le poids indépendamment, d'autre part sous la forme de contours pour la taille et le poids conjointement, à chacun des paliers précités. Vu le caractère actuellement très mélangé de la population genevoise, ces normes, comme l'enquête elle-même du reste, ne concernent pour le moment que les sujets suisses.

Tableau 2
*Formules pour le calcul des paramètres du mélange*
*La signification des symboles ressort du tableau 1.*

| | |
|---|---|
| Taille | |
| Moyenne | $\tilde{\mu}_x = \lambda'\mu_x' + \lambda''\mu_x''$ |
| Ecart-type | $\tilde{\sigma}_x = \sqrt{[\lambda'\sigma_x'^2 + \lambda''\sigma_x''^2 + \lambda'\lambda''\delta_x^2]}$ avec $\delta_x = \mu_x' - \mu_x''$ |
| Poids | |
| Moyenne | $\tilde{\mu}_y = \lambda'\mu_y' + \lambda''\mu_y''$ |
| Ecart-type | $\tilde{\sigma}_y = \sqrt{[\lambda'\sigma_y'^2 + \lambda''\sigma_y''^2 + \lambda'\lambda''\delta_y^2]}$ avec $\delta_y = \mu_y' - \mu_y''$ |
| Corrélation | |
| Coefficient | $\tilde{\varrho} = [\lambda'\sigma_x'\sigma_y'\varrho' + \lambda''\sigma_x''\sigma_y''\varrho'' + \lambda'\lambda''\delta_x\delta_y]/\tilde{\sigma}_x\tilde{\sigma}_y$ |

- *normal* par rapport au contour du mélange,
- *non-normal* par rapport au contour suisse,
- enfin *normal* par rapport au contour italien.

Cet exemple montre que les deux jugements basés, l'un sur le contour du mélange, l'autre sur le contour de la population d'origine, sont compatibles dans le cas d'un sujet de provenance italienne, mais contradictoires dans le cas d'un sujet de provenance suisse.

Le tableau 3 récapitule le classement d'un échantillon de 50 000 sujets composé de 30 000 Suisses et 20 000 Italiens issus des populations d'origine spécifiées au tableau 1. Les 50 000 couples $\langle x ; y \rangle$ ont été engendrés par simulation sur un ordinateur.

Dans l'ensemble, on obtient ainsi une contradiction 3 fois sur 100 environ (3,08 %), plus fréquemment cependant chez les Italiens (4,45 %) que chez les Suisses (2,17 %). La plus forte contribution à ces malclassifications est apportée dans le cas présent par la région $\overline{MSI}$, c'est-à-dire surtout par les 559 sujets de provenance italienne, *normaux* pour le contour de leur population d'origine, mais *non-normaux* par rapport au contour du mélange. Statistiquement parlant, l'écart entre les pourcentages 2,17 et 4,45 est significatif; pour le détail du test appliqué, cf. [6].

Pour conclure, il convient d'observer que si l'on désire juger de la *normalité biologique* d'un sujet en utilisant à la place du contour établi pour la population d'origine un contour calculé sur la base d'un mélange, on devra bien s'accommoder d'un certain risque de malclassification. En ce qui concerne le cas particulier traité ici dans les hypothèses faites, ce risque en termes de fréquence relative est approximativement de l'ordre de grandeur de la probabilité choisie pour définir la *non-normalité biologique* (5 %).

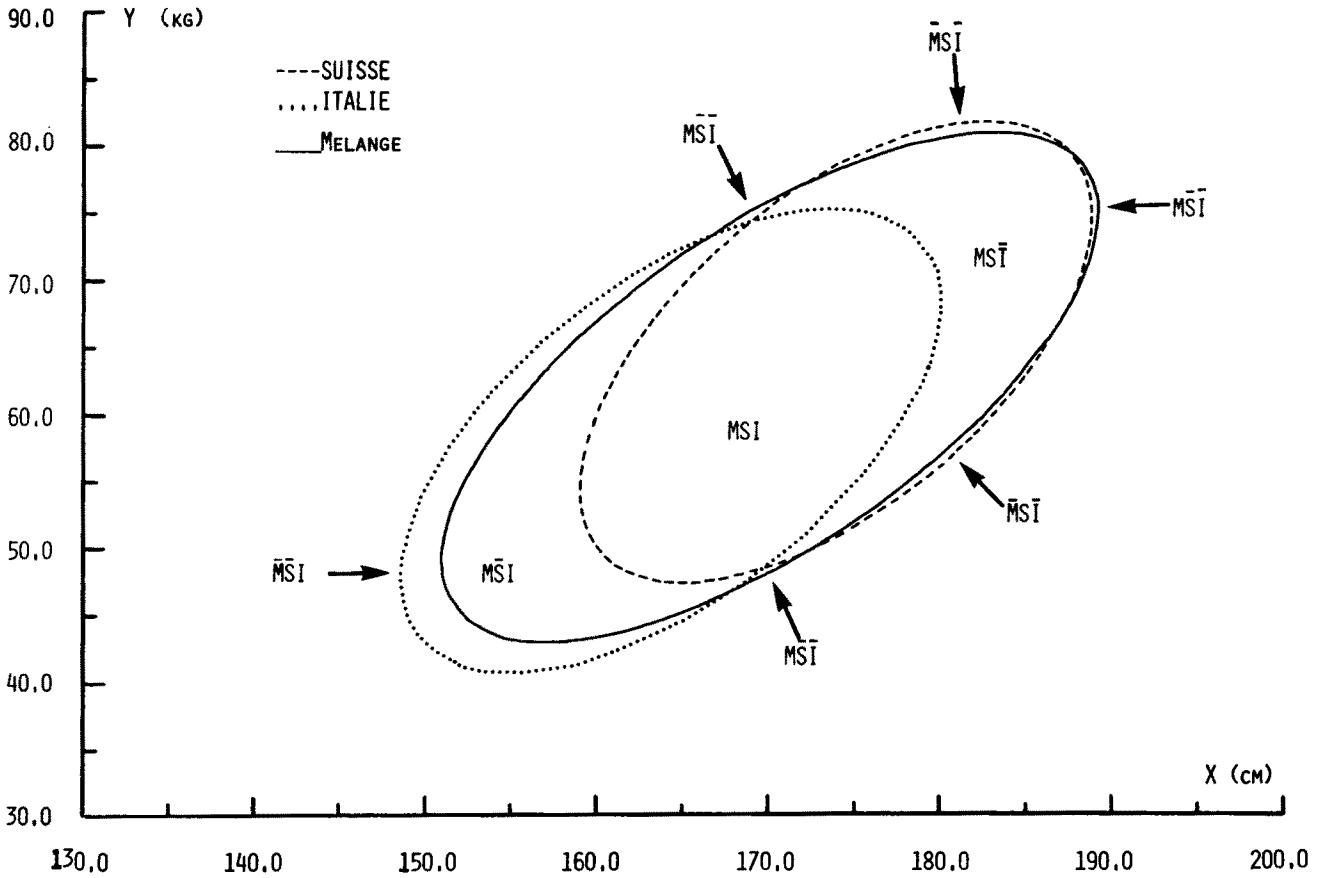
Il ne faut toutefois par perdre de vue qu'il ne s'agit là que d'une situation très particulière, de sorte que pour d'autres populations mélangées dans des proportions différentes, ce risque de malclassification pourra prendre le cas échéant d'autres valeurs. La répercussion, sur cette valeur estimée de la fréquence de malclassifications, de la variation de certains des paramètres $\mu$, $\sigma$, $\varrho$, $\lambda$ ou de la probabilité de tolérance, est analysée plus en détail dans [4].

Tableau 3 *Fréquences absolue et relative des jugements compatibles et contradictoires («malclassifications»)*

| Jugements | | Fréquences | | | | | |
|---|---|---|---|---|---|---|---|
| Relatif à la population d'origine | Relatif au mélange | absolue | relative | absolue | relative | absolue | relative |
| | | Suisse | | Italie | | Ensemble | |
| Normaux | Normaux | 28 369 [MSI, $\overline{MSI}$] | 94,56 % | 18 436 [MSI, $\overline{MSI}$] | 92,18 % | 48 805 | 93,61 % |
| Non-normaux | Non-normaux | 982 [$\overline{MSI}$, $\overline{MSI}$] | 3,27 % | 665 [$\overline{MSI}$, $\overline{MSI}$] | 3,37 % | 1 656 | 3,31 % |
| Total des jugements compatibles | | 29 351 | 97,83 % | 19 110 | 95,55 % | 48 461 | 96,92 % |
| Normaux | Non-normaux | 131 [$\overline{MSI}$, $\overline{MSI}$] | 0,44 % | 559 [$\overline{MSI}$, $\overline{MSI}$] | 2,80 % | 690 | 1,38 % |
| Non-normaux | Normaux | 518 [$\overline{MSI}$, $\overline{MSI}$] | 1,73 % | 331 [$\overline{MSI}$, $\overline{MSI}$] | 1,65 % | 849 | 1,70 % |
| Total des jugements contradictoires | | 649 | 2,17 % | 890 | 4,45 % | 1 539 | 3,08 % |
| Tous les jugements(n) | | 30 000 | 100 % | 20 000 | 100 % | 50 000 | 100 % |

N.B. Les fréquences relatives se rapportent, dans chaque colonne, à l'effectif total n des jugements dans la population concernée. Entre crochets [] figurent les régions correspondantes.

Figure 1 *Contours $C_{0,95}$ pour les deux populations d'origine (Suisse, Italie) et pour le mélange*




N.B. Sur cette figure, la région $\overline{MSI}$, extérieure aux trois contours, n'est pas signalée. La région $\overline{MSI}$ est vide, puisque la région SI est entièrement incluse dans le contour du mélange.

**Zusammenfassung**

*Toleranzgrenzen im Fall einer Mischung von zwei zweidimensionalen Grundgesamtheiten: Schätzung der relativen Häufigkeit von Fehlzuordnungen*

Anhand von Angaben betreffend Körpergrösse und -gewicht aus der anthropologischen Literatur werden die Mittelwerte, Streuungen und der Korrelationskoeffizient für eine Mischung von zwei verschiedenen Grundgesamtheiten in einem gegebenen Verhältnis berechnet. Darauf werden die drei entsprechenden 95 %-Streuungsellipsen als zweidimensionale biologische Normalbereiche überlagert dargestellt. Endlich wird aufgrund einer auf dem Computer simulierten Stichprobe aus derselben Mischung die relative Häufigkeit von Urteilswidersprüchen geschätzt (nämlich ungefähr 3 %), die sich dann ergeben können, wenn beim Urteilen die Mischungsellipse anstelle der Ellipse der richtigen Ursprungspopulation eingesetzt wird.

**Résumé**

A partir de données numériques pour la taille et le poids émanant d'enquêtes anthropologiques, on détermine les moyennes, les écarts-types et le coefficient de corrélation pour le mélange, dans des proportions données, de deux populations distinctes. Ensuite, on présente la superposition des trois ellipses-contours à 95 %, en tant que domaines de normalité biologique à deux dimensions dans les trois populations respectives. Enfin, à l'aide d'un échantillon issu du même mélange mais simulé sur l'ordinateur, on procède à l'estimation (à savoir 3 % approximativement) de la fréquence relative des contradictions qui peuvent apparaître lorsqu'on confronte les inférences reposant d'une part sur le contour du mélange, d'autre part sur celui de la population d'origine appropriée.

**Summary**

*Tolerance limits for a mixture of two two-dimensional populations: estimation of the relative frequency of misclassification*

Numerical data on height and weight quoted in anthropological surveys have been used to compute the means, standard deviations and coefficient of correlation for a mixture (in a given ratio) of two distinct populations. Next, the three corresponding 95 % contour-ellipses are plotted on the same graph as a representation of two-dimensional limits for biological normality. Lastly, a random sample from the same mixture has been simulated on the computer in order to derive an estimate (viz. approximately 3 %) of the relative frequency of contradictory inferences which may arise when classification is based on the ellipse from the mixture instead of the ellipse from the proper component population.

**Bibliographie**

[1] Ciba-Geigy: Tables scientifiques *(Diem K.* et *Lentner C.*, rédacteurs), 7e éd. Ciba Geigy SA, Bâle 1972.

[2] *Hald A.:* Statistical theory with engineering applications, 7th printing. J. Wiley and Sons Inc., New York-London-Sidney 1967.

[3] *Hoyer G.:* Automatisch gesteuerte Streuungsellipsen. Meth. Inform. Med. *11*, 37 (1972).

[4] *Lang R.* and *Rieben A.:* Estimated frequency of misclassification in contour-ellipses for mixed bivariate populations. To be published.

[5] *Linder A.:* Statistische Methoden für Naturwissenschafter, Mediziner und Ingenieure, 3. Aufl. Birkhäuser, Basel-Stuttgart 1960.

[6] *Schwartz D.:* Méthodes statistiques à l'usage des médecins et des biologistes. Editions médicales Flammarion, Paris 1963.

**Adresse des auteurs**

*R. Lang* et *A. Rieben,* Institut de médecine sociale et préventive, Quai Ernest-Ansermet 20, CH-1205 Genève.