Évaluer les erreurs de mesure en anthropométrie

Frédéric Santos
CNRS, UMR 5199 PACEA
Courriel : frederic.santos@u-bordeaux.fr

30 janvier 2014

Table des matières

Résumé

Ce petit guide vise à donner des procédures, méthodes et indicateurs simples pour évaluer les erreurs de mesure en anthropométrie. Il s'agit donc ici d'accord intra- ou inter-juges, mesurés sur des variables *continues*[1].

Fondamentalement, il n'y a aucune raison de distinguer les erreurs de mesure intra-observateur ou inter-observateurs[2] : les mêmes procédures s'appliquent indistinctement aux deux types d'erreurs. Dans ce document, nous ne différencierons donc pas ces deux cas, et nous nous bornerons, sauf mention spécifique, à considérer *différentes séries d'observations*, sans nous préoccuper du fait qu'elles soient issues d'un même ou de plusieurs observateurs.

1. Cas de deux séries d'observations uniquement

§1.1. Exemple introductif. — On considère qu'à l'aide d'un instrument de mesure donné (pied à coulisse, règle, ruban, etc.), deux séries de mesures ont été effectuées. Typiquement, la même mesure aura été prise deux fois de suite par un même observateur (cas de l'erreur *intra-observateur*) ou par deux observateurs différents (cas de l'erreur *inter-observateurs*) sur un certain nombre d'individus.

1. Le cas de l'accord entre juges sur des critères qualitatifs ou discrets n'est pas traité dans ce document, et appelle des méthodes spécifiques (typiquement, le Kappa de Cohen ou de Fleiss)

2. Avec la nuance d'ordre théorique que, dans le cas de l'erreur intra-observateur, il est difficile de considérer les deux relevés comme indépendants, ce qui n'est pas sans poser parfois quelques problèmes philosophiques et mathématiques — que nous négligerons ici.

Le but est d'évaluer *globalement* la différence de jugement entre les deux séries d'observations, et à ce titre, on doit disposer de suffisamment d'individus. Une quinzaine paraît être le minimum requis, dans un contexte anthropométrique.

Au final, nous obtenons le tableau de mesures suivant :

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Obs. 1	15.3	14.2	18.7	16.9	15.6	12.1	14.5	17.0	16.3	13.4	17.2	15.5	14.2	13.9
Obs. 2	14.9	15.0	18.4	17.1	15.9	12.2	14.5	16.6	16.4	13.3	16.8	15.8	14.3	13.6

Table 1. — Relevés de deux séries de mesures, effectuées par deux observateurs sur 14 individus.

§1.2. Formalisme et objectifs. — Formellement, on considère que le tableau précédent est une matrice à $n = 2$ lignes et $p = 14$ colonnes. Chacune des cases (i, j) du tableau, ou encore chaque valeur $X_{i,j}$ de la matrice, est la mesure effectuée par l'observateur i chez l'individu j. Cette valeur, *a priori*, n'est pas la mesure *exacte* de l'individu, mais une mesure entachée d'erreurs :

— éventuel biais systématique lié à la propension d'un observateur ou d'un instrument de mesure à être plus ou moins « large » dans ses valeurs (une des série de mesures peut surestimer systématiquement, ou sous-estimer systématiquement) ;

— erreur aléatoire, dépendant de la façon dont l'expérimentateur va positionner plus ou moins exactement son instrument de mesure à chaque prise, et de la précision dudit instrument.

En définitive, chaque valeur $X_{i,j}$ du tableau peut s'écrire comme :

$$X_{i,j} = x_j + \varepsilon_{i,j}$$

où x_j est la « vraie valeur » de l'individu j, et $\varepsilon_{i,j}$ une variable aléatoire symbolisant l'erreur commise par l'observateur i, pour laquelle on peut raisonnablement émettre l'hypothèse que $\varepsilon_{i,j} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, c'est-à-dire comportant un biais systématique μ_i, et ayant une variance σ_i^2.

Un cheminement logique pour évaluer l'écart entre les deux relevés peut être le suivant :

— représentation graphique simple des différences entre les estimations, par tout moyen jugé adéquat et un minimum « parlant », pour fixer les idées et avoir un premier aperçu de l'amplitude et de la direction des différences ;

— évaluation du biais systématique : il s'agit de savoir s'il est raisonnable de considérer que $\mu_1 = \mu_2 = 0$, ce qui reviendrait à dire qu'il n'y a aucune sur- ou sous-évaluation systématique de la part d'un des relevés ;

— calcul empirique de « l'amplitude de l'écart » entre les deux relevés ;

— idéalement, on souhaiterait enfin estimer σ_i^2, c'est-à-dire non pas l'écart empirique entre les deux relevés, mais l'erreur technique de mesure liée à l'utilisation de l'instrument.

§1.3. Recherche d'un biais : direction de l'erreur. — Il s'agit ici essentiellement de voir si l'erreur de mesure (Obs 1 – Obs 2) est « toujours dirigée dans le même sens », sans préjuger de son amplitude.

Par exemple, sur la table 2 ci-dessous, l'écart entre les mesures prises par l'observateur 1 et l'observateur 2 est moins important que sur la table 1, mais par contre, et contrairement au tableau 1 cette fois, les erreurs sont toujours « dirigées dans le même sens » : l'observateur

2 a constamment sous-estimé par rapport à l'observateur 1. Même si l'amplitude de l'erreur est faible, il peut s'agir d'un motif d'inquiétude.

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Obs. 1	15.3	14.2	18.7	16.9	15.6	12.1	14.5	17.0	16.3	13.4	17.2	15.5	14.2	13.9
Obs. 2	15.2	14.0	18.6	16.7	15.5	12.0	14.5	16.9	16.1	13.3	17.0	15.4	14.1	13.8

Table 2. — Exemple de biais systématique entre deux séries d'observations.

1.3.1. Aspects graphiques et descriptifs. — Quelques indicateurs élémentaires permettent de se faire une bonne idée de la situation. On reprend l'exemple de la table 1, que l'on complète avec les calculs suivants :

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Obs. 1	15.3	14.2	18.7	16.9	15.6	12.1	14.5	17.0	16.3	13.4	17.2	15.5	14.2	13.9
Obs. 2	14.9	15.0	18.4	17.1	15.9	12.2	14.5	16.6	16.4	13.3	16.8	15.8	14.3	13.6
Diff. brute	0.4	-0.8	0.3	-0.2	-0.3	-0.1	0	0.4	-0.1	0.1	0.4	-0.3	-0.1	0.3
Diff. moy.	0.0													

Table 3. — Résumés statistiques basiques pour les erreurs de mesure de la table 1.

Ici, il ne semble pas y avoir de biais systématique : la différence moyenne entre les deux relevés est nulle, et les différences (Obs. 1 – Obs. 2) sont tantôt positives, tantôt négatives, dans des proportions similaires.

La figure 1 présente ces différences de mesure sous forme graphique sur le barplot de gauche. À titre de comparaison, le barplot des erreurs de mesure de la table 2 est également présent, à droite de la figure. On retrouve graphiquement le constat déjà effectué plus haut : la table 2 présente des écarts de mesure globalement très faibles, mais cette précision est à nuancer par la présence d'un biais systématique.

1.3.2. Tests statistiques. — Pour peu qu'on ait suffisamment d'individus, la présence d'un biais systématique peut être formalisée objectivement par un test *apparié* de Student ou de Wilcoxon entre les mesures prises par les deux observateurs.

À titre d'exemple, ici, la présence d'une différence systématique de jugement entre les observateurs 1 et 2 est attestée pour la table 2 ($p < 0,01$), et il n'y a évidemment aucun biais systématique entre les observateurs 1 et 2 pour la table 1 ($p = 1$ puisque la moyenne des différences brutes est nulle).

Contrairement à une idée fort répandue parmi les utilisateurs de statistique en anthropométrie, un test apparié de Student ou de Wilcoxon employé seul n'est donc *absolument pas adapté* à l'évaluation d'une « significativité » globale des différences de mesures entre observateurs : le test peut être « significatif » pour des erreurs très faibles mais toujours dirigées dans le sens (par exemple) d'une surévaluation de l'observateur 1, et peut être « non significatif » pour des erreurs de mesure énormes mais sans surévaluation constante de la part d'un des deux observateurs. Un test apparié de Student et ou de Wilcoxon ne permettra de tirer aucune autre conclusion que celle de la présence d'un biais systématique.

L'exploration d'un biais systématique s'avère donc être quasi-totalement décorrélée de la question de l'ordre de grandeur de l'erreur de mesure.

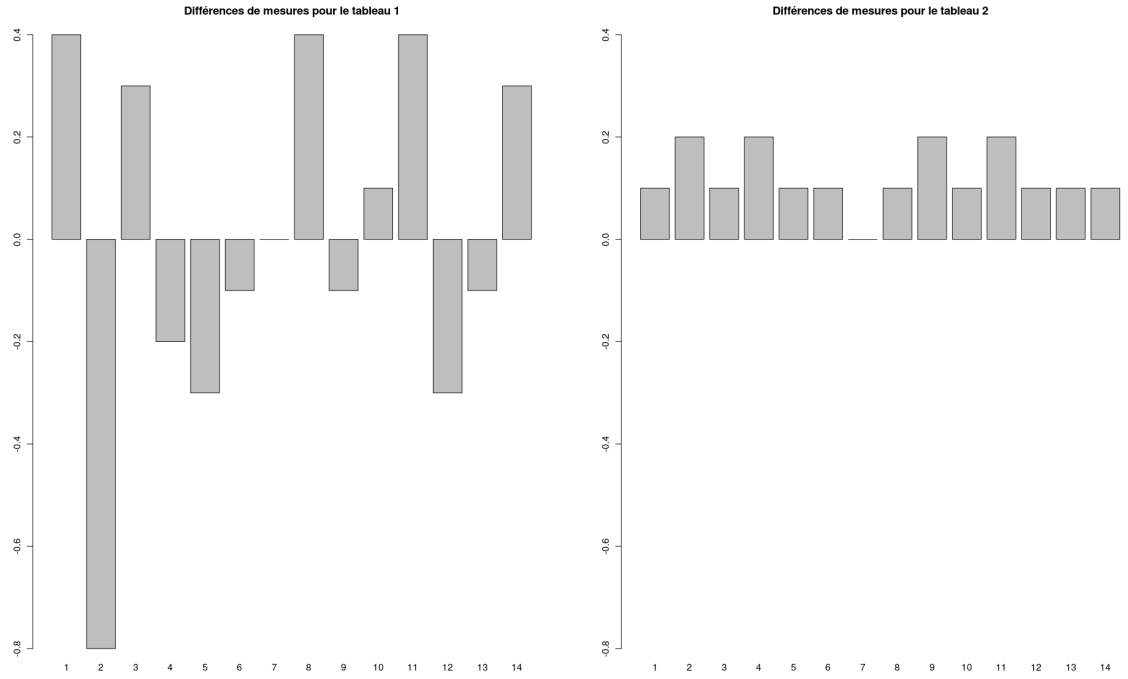

Figure 1. — Barplot pour les erreurs de mesures associées aux tables 1 et 2.

§1.4. Quantifier l'amplitude de l'erreur. — Après la recherche d'un biais, on s'intéresse à la question (décorrélée, répétons-le) de l'amplitude de l'erreur, qui est en réalité ce qui nous intéresse au premier chef : sans nous occuper de sa « direction », l'erreur entre les deux séries d'observations est-elle, en moyenne, trop grande, ou acceptable ?

Une première étape peut être, pour chaque paire de mesure (Obs. 1, Obs. 2), de calculer la « déviation relative » (en %) entre ces deux observations.

Elle peut se calculer comme la différence en valeur absolue, divisée par la moyenne des deux mesures :

$$\text{déviation relative} = \frac{|\text{Obs. 1} - \text{Obs. 2}|}{(\text{Obs. 1} + \text{Obs. 2})/2}$$

Il s'agit bien d'une déviation *relative*, puisqu'exprimée comme le rapport de la différence entre observations sur l'ordre de grandeur de la mesure.

Individu	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Obs. 1	15.3	14.2	18.7	16.9	15.6	12.1	14.5	17.0	16.3	13.4	17.2	15.5	14.2	13.9
Obs. 2	14.9	15.0	18.4	17.1	15.9	12.2	14.5	16.6	16.4	13.3	16.8	15.8	14.3	13.6
Diff. abs. (cm)	0.4	0.8	0.3	0.2	0.3	0.1	0	0.4	0.1	0.1	0.4	0.3	0.1	0.3
Dév. rel. (%)	2.6	5.5	1.6	1.2	1.9	0.8	0	2.3	0.6	0.7	2.3	1.9	0.7	2.2

Table 4. — Déviations relatives pour les mesures de la table 1.

Ici, la moyenne des erreurs absolues est de 0.27 (cm) : cela signifie que, en moyenne, lorsque ces deux observateurs mesurent cette même quantité sur un même individu, on peut s'attendre à ce que l'écart absolu entre leurs deux mesures soit de 0.27 cm. L'expertise du praticien parlera pour savoir s'il s'agit d'une erreur acceptable ou non.

Afin d'aider à l'interprétation, l'indicateur de déviation relative, ici calculé mesure par mesure, peut être étudié de façon globale, en calculant la moyenne des erreurs absolues, et en la divisant par la moyenne de toutes les mesures effectuées par les deux observateurs. Il sera donc ici égal à $0.27/15.34 = 1.76\%$. Cela paraît relativement acceptable en soi qu'en moyenne, les deux observateurs donnent des mesures distantes l'une de l'autre de moins de 2% ; néanmoins, cela pourra peut-être déjà être beaucoup si vous les mesures se font avec un instrument censé être très précis et très simple à utiliser.

On insistera notamment sur le fait qu'ici, c'est l'expertise du praticien qui est censée parler pour commenter l'acceptabilité de la valeur obtenue — et que ce n'est pas une mauvaise chose, loin de là. Comme on le verra plus loin (§3, p. 9), l'objectivité statistique conduit parfois à des conclusions sans intérêt ou signification pratiques. L'information subjective apportée par l'expert se révèle souvent indispensable dans ce contexte.

Repérer les valeurs posant problème. — Il est utile de pouvoir repérer en un coup d'œil si, ponctuellement, une ou plusieurs valeurs présentent une erreur anormalement élevée compte tenu de celles observées sur tous les autres individus. Il existe au moins deux causes possibles à cela : une grosse faute au moment de l'utilisation de l'instrument de mesure, ou une erreur de saisie dans votre fichier de données.

Pour repérer les individus pour lesquels l'erreur de mesure obtenue pose problème, on pourra utiliser un boxplot mettant en valeur les valeurs aberrantes, ou mieux, un « graphique de Bland-Altman ». Il s'agit tout simplement d'un graphique des erreurs brutes où trois lignes horizontales sont tracées : la moyenne des erreurs brutes en trait plein, et en pointillés la moyenne des erreurs brutes $\pm$ deux fois l'écart-type de ces dernières. Les points s'excluant de la « bande centrale », tel que l'individu 2 sur la figure 2 p. 6, sont ceux pour lesquels l'erreur est nettement trop élevée par rapport à celle commise sur les autres individus. Pour plus d'information sur ce type de graphique, on consultera avec profit [Syl11].

§1.5. Coefficient de concordance. — Plutôt que de considérer séparément les questions de l'amplitude et de la direction de l'erreur, le coefficient de concordance offre une approche unifiée.

1.5.1. Aspects graphiques. — Une autre méthode graphique simple pour évaluer la concordance entre les deux relevés est de tracer un graphique bivarié comme en figure 3 p. 7, sur lequel est également représentée la première bissectrice des axes — c'est-à-dire la droite d'équation $y = x$, ou autrement dit, la droite de parfaite concordance entre les deux relevés.

Plus le point représentant un individu est proche de la droite $y = x$, plus la concordance entre les deux observations réalisées sur cette individu est forte. En d'autres termes, plus le nuage de points (Obs.1, Obs. 2) « colle » exactement à la première bissectrice des axes (mais avec des points se répartissant des deux côtés de la bissectrice), et plus l'accord entre les deux observateurs est bon. On introduit ci-dessous un nouvel outil, le coefficient de concordance, basé sur cette idée.

1.5.2. De l'insuffisance du coefficient de corrélation. — Il est à rappeler qu'on ne mesure *jamais* l'accord entre deux séries de mesure en calculant le coefficient de corrélation entre ces deux séries. Certes, plus les deux séries d'observations concorderont, et plus le coefficient de corrélation se rapprochera de 1, mais la réciproque n'est absolument pas vraie : si les mesures prises dans la deuxième série d'observation sont toujours égales au triple de celles de la première série, ce coefficient sera également proche de 1 !

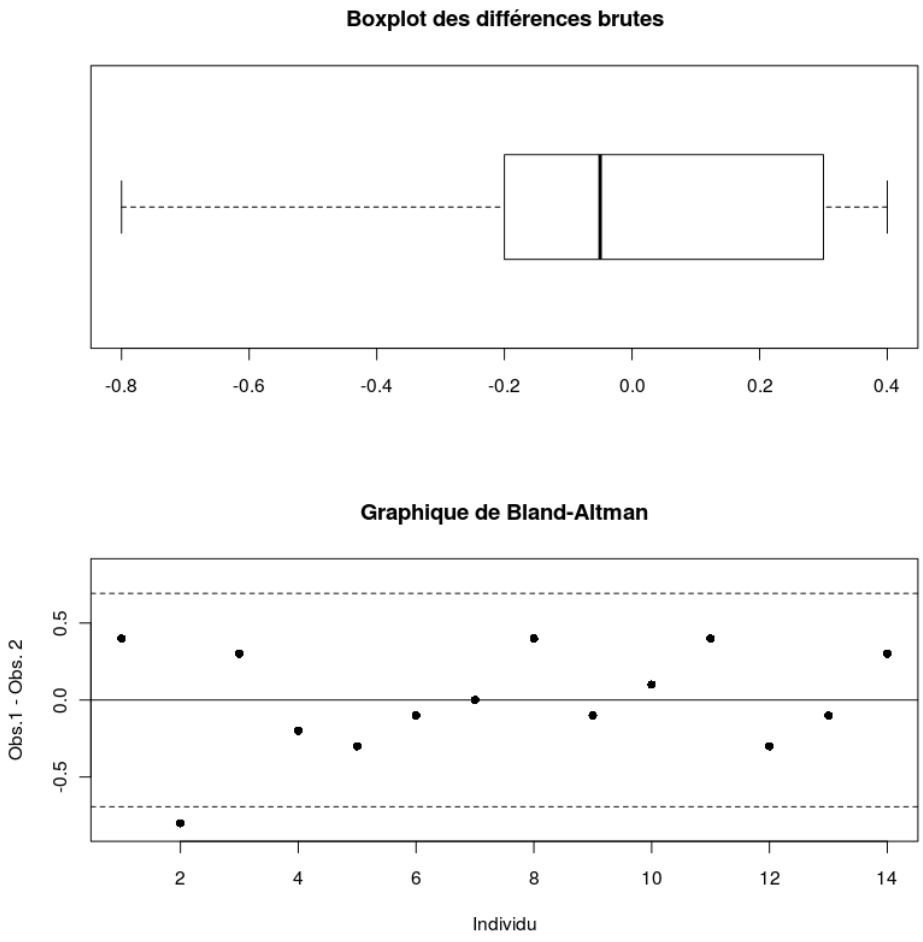

Figure 2. — Boxplot (en haut) et « graphique de Bland-Altman » (en bas) pour les erreurs de mesure de la table 1.

La figure 4 p. 8 présente plusieurs cas de très mauvais accord entre les deux séries d'observations, qui possèdent pourtant un coefficient de corrélation très élevé.

Le coefficient de corrélation entre les deux séries de mesures ne reflète donc pas correctement (dans le sens que l'on souhaite) leur concordance.

1.5.3. Aspects calculatoires. — Nous introduisons donc un indicateur très simple, destiné à corriger les « failles » du coefficient de corrélation dans ce contexte : le coefficient de concordance [Lin89].

Définition. — On appellera *coefficient de concordance* entre deux relevés l'indicateur suivant :

$$\widehat{\rho_c} = \frac{2 \times S_{12}}{S_1^2 + S_2^2 + \left(\overline{Y_1} - \overline{Y_2}\right)^2}$$

où S_{12} est la covariance entre les deux relevés, S_i^2 est la variance du i-ème relevé, et $\overline{Y_i}$ sa moyenne.

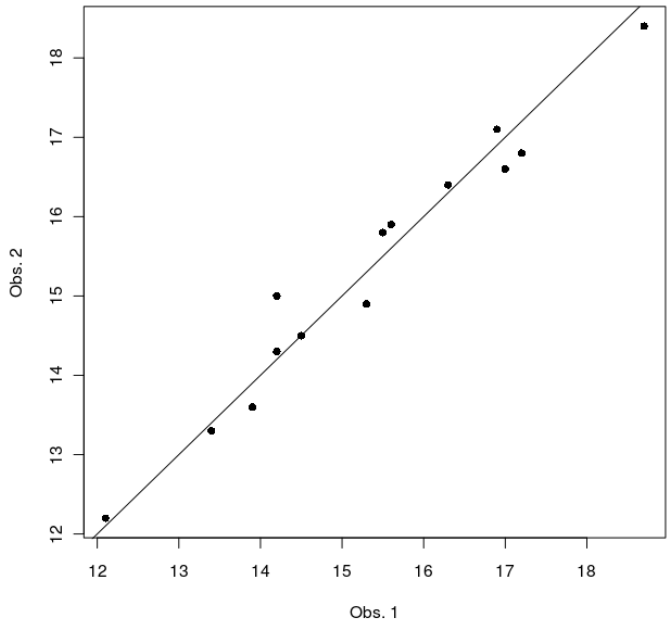

Figure 3. — Graphique bivarié des deux relevés du tableau 1, et droite $y = x$.

Ce coefficient propose un compromis entre amplitude de l'erreur et biais, et incorpore de façon synthétique ces deux composantes de l'erreur. Il possède les propriétés suivantes :

(i) $-1 \leqslant \widehat{\rho}_c \leqslant 1$

(ii) $|\widehat{\rho}_c| \leqslant \widehat{\rho}$; avec $\widehat{\rho}$ désignant le coefficient de corrélation « habituel » : le coefficient de concordance donne toujours une vision moins optimiste (donc plus réaliste !) de la concordance réelle des deux relevés.

(iii) Dans l'inégalité large ci-dessus, le cas d'égalité $\widehat{\rho}_c = \widehat{\rho}$ se produit si et seulement si les deux relevés ont la même moyenne $\left(\overline{Y_1} - \overline{Y_2} = 0\right)$, et la même variance $(S_1^2 = S_2^2)$.

(iv) $\widehat{\rho}_c = 0 \Leftrightarrow \widehat{\rho} = 0$; c.à-d. que le coefficient de concordance est nul si et seulement si les deux relevés sont totalement décorrélés (ce qui serait en soi surprenant !).

(v) Corollaire des points précédents : si les deux relevés sont en parfait accord, alors $\widehat{\rho}_c = 1$.

Ce coefficient peut être calculé afin d'obtenir en un seul indicateur une idée globale de l'accord inter-relevés. On souhaite évidemment qu'il soit le plus proche possible de 1, sachant que toute valeur comprise entre 0.9 et 1 pourra être considérée comme rassurante. Toute valeur inférieure à 0.8 reflétera un désaccord entre observateurs commençant à être une réelle source d'inquiétude : il faudra alors s'interroger sur le protocole, la définition des mesures, l'expertise des personnes ayant effectué les mesures, etc. Il n'existe pas de grille de lecture standardisée et universelle pour le coefficient de concordance, et on pourra faire preuve de plus ou moins de sévérité en fonction de l'instrument de mesure utilisé, et du niveau d'expérience des utilisateurs.

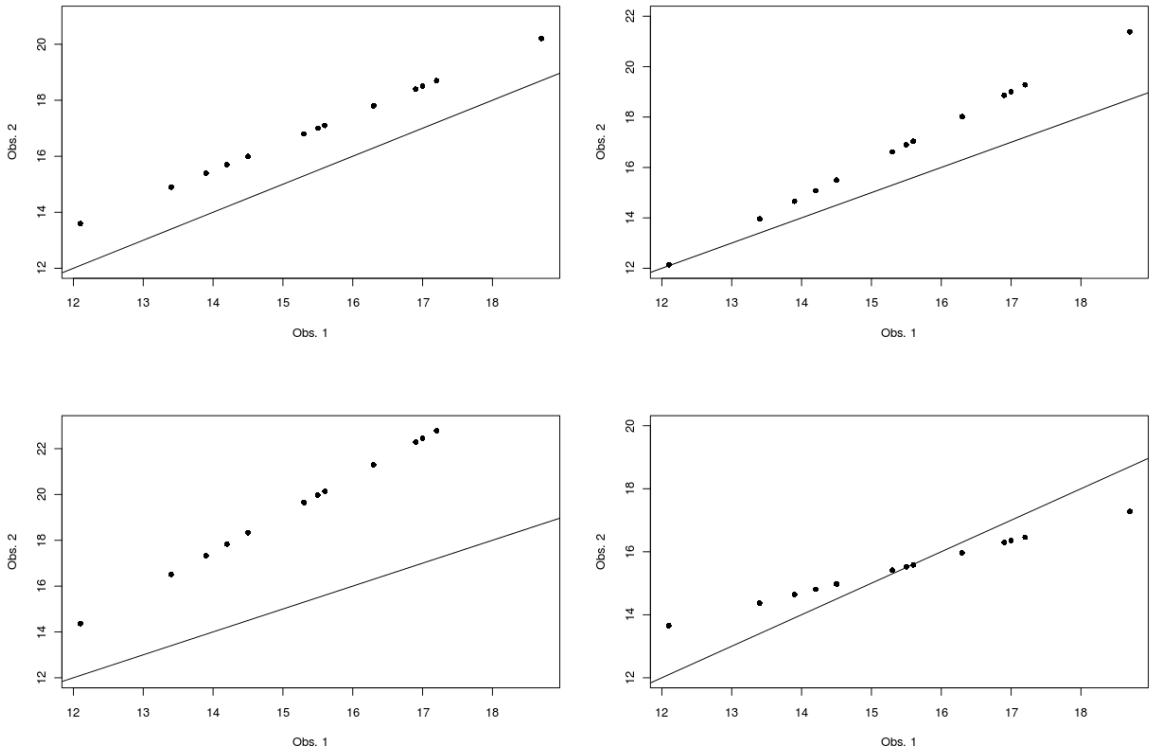

Figure 4. — Graphiques bivariés entre deux relevés, et première bissectrice des axes

§1.6. Bilan et présentation des résultats. — En définitive, voici une démarche possible pour évaluer l'accord entre deux séries d'observations :

(i) Calculer les différences brutes (Obs. 1 – Obs. 2) pour chaque mesure, puis effectuer un premier résumé graphique (cf. figures 2 et 3, ou à la rigueur figure 1) de ces erreurs afin d'avoir un aperçu global de la situation et d'émettre vos premières hypothèses concernant la présence ou pas d'un biais, et l'importance de l'erreur.

(ii) Calculer la moyenne des différences brutes (donc « en gardant les signes ») entre les deux séries d'observations. Idéalement, celle-ci doit être relativement proche de 0 si l'une des deux séries ne surestime pas systématiquement par rapport à l'autre. Un test apparié de Student ou Wilcoxon peut être mené afin de déceler la possible présence d'un biais.

(iii) Calculer les différences absolues |Obs. 1 – Obs. 2| ainsi que leur moyenne, cette dernière constituant un bon indicateur global de précision. Interprétez-la eu égard à vos objectifs initiaux de précision et en utilisant votre expertise disciplinaire. Si besoin, facultativement, les déviations relatives peuvent être calculées et présentées également.

(iv) Calculer et interpréter, comme résumé global, le coefficient de concordance.

2. Généralisation : cas de plus de deux séries d'observations

Dans le cas où l'évaluation de l'erreur inter-observateurs porte sur trois juges ou plus, si l'erreur doit être considérée « globalement » et non « par paires » entre un observateur de référence (celui qui a défini un nouveau protocole de mesure, par exemple) et chacun des autres observateurs, on pourra se cantonner, d'un point de vue descriptif, aux outils suivants :

— pour chaque grandeur étudiée, calculer l'écart-type des valeurs obtenues par les différents juges pour obtenir leur dispersion *absolue*, en unités de mesure d'origine ;

— pour chaque grandeur étudiée, calculer le coefficient de variation des valeurs obtenues par les différents juges pour obtenir leur dispersion *relative*, en pourcentage par rapport à l'ordre de grandeur des mesures.

La présence d'un biais systématique pourra être évaluée grâce à des procédures d'ANOVA appariée, comme le test (non-paramétrique, et donc applicable dans tous les cas, même pour des variables numériques ordinales) de Friedman [3]. Ce test est par exemple disponible dans les logiciels (gratuits) PAST ou R : consulter des manuels dédiés pour plus d'information [HH05, LDL10], ou visiter le site `http://marne.u707.jussieu.fr/biostatgv/`.

3. Vous prendrez bien un petit TOST ?

§3.1. Philosophie des tests d'équivalence. — Selon le contexte et l'instrument de mesure utilisé, pour une mesure donnée, une erreur inter-observateurs (ou une erreur entre deux instruments de mesures) de 5 mm pourra vous paraître très satisfaisante, ou extrêmement élevée. En avoir une petite idée fait partie de l'expertise que chaque praticien est censé posséder quant à l'objet de son étude, et aux objectifs de précision qu'il s'était fixés *a priori*.

Prenons un exemple caricatural : imaginons que l'on veuille comparer la hauteur moyenne de deux espèces de peupliers. Si l'on dispose d'énormément d'arbres dans chacun des deux groupes, la puissance du test sera très élevée. En conséquence, le test pourra devenir significatif même pour de (très) petites différences observées entre ces deux espèces. Typiquement, si l'on dispose de 2.000 arbres de chaque espèce, même une différence de hauteur moyenne de 1 cm entre les deux groupes pourra être décrétée « significative », ce qui est absolument insignifiant « dans la vraie vie » pour des arbres faisant plusieurs dizaines de mètres de hauteur [4].

Dans le cadre de l'évaluation des erreurs de mesure, il peut parfois s'agir du même souci : vous pouvez très bien décider que la question « existe-t-il un biais systématique entre mes deux séries de mesure ? » est mal posée : la « significativité » attestée (par un $p < 0.05$) d'un biais systématique de 0.5 mm peut vous être parfaitement indifférente, en pratique. En d'autres termes, même si le biais de 0.5 mm entre les deux séries de mesures n'est pas imputable au hasard, cela peut vous sembler négligeable, en tout cas pas « significativement diffŕent de 0 » dans la vie concrète !

Vous pouvez donc être tenté(e) de remplacer la question « existe-t-il un biais systématique entre mes deux séries de mesures ? », par la question « existe-t-il un biais systématique *supérieur à 1 mm* [ou tout autre seuil de votre choix !] entre mes deux séries de mesures ? ». Vous

3. Du nom de l'économiste américain Milton Friedman, libéral et chef de file du courant monétariste.

4. On rappelle que la « significativité » statistique dans un test de comparaison signifie que les deux échantillons ne proviennent pas *exactement* de la même distribution de probabilité, mais de deux distributions différentes. Le fait de savoir si l'écart entre ces deux distributions théoriques est suffisamment grand pour être porteur de sens « dans la vraie vie » ne relève par contre absolument pas de la statistique ni de l'objectivité mathématique... ce qui aboutit à de graves contre-sens dans beaucoup d'études.

ne vous intéressez alors plus à l'existence d'une différence « significative » abstraite (« dans l'absolu ») ; mais à l'existence d'une différence supérieure à un seuil que vous fixez vous-même comme *le seuil à partir duquel, selon votre expertise, vous commenceriez à considérer qu'il y a un vrai écart problématique* entre vos deux séries de mesures. Ce qui, dans certains cas, peut s'avérer beaucoup plus pertinent.

Cette procédure est la procédure canonique de certification des médicaments génériques, et est systématiquement employée dans les essais cliniques de bio-équivalence et de non-infériorité [ERJ+08]. Si un médicament générique coûte 50% moins cher à produire (et donc à rembourser !) qu'un médicament classique, une analyse coût-bénéfice raisonnable pourra aisément lui pardonner d'être 2% moins efficace que le médicamnt standard, sachant que ces 2% n'auront aucun impact concret sur la santé du patient. La procédure de certfication imposera seulement de prouver que *la différence d'efficacité entre les deux médicaments n'excède pas 10%* [par exemple !], seuil à partir duquel on considèrerait que le médicament générique ne remplirait plus correctement son office — en dépit de tous ses avantages économiques.

Pour résumer, dans les procédures statistiques exposées dans les sections précédentes, c'était « le test » qui choisissait le seuil à partir duquel une différence était « statistiquement significative » ou non, et cela peut être inadéquat. On propose dans cette section de fixer *nous-mêmes* ce seuil, et de ne laisser au test que la possibilité de trancher par rapport au seuil que l'utilisateur aura fixé.

§3.2. Formalisation. — On pourra consulter [ERJ+08] pour se familiariser avec l'idée de Two One-Sided Test (TOST), même si cet article ne traite que du cas d'échantillons indépendants.

On note $(X_1, \cdots, X_n)$ les valeurs prises par le premier observateur sur les n individus, et $(Y_1, \cdots, Y_n)$ les valeurs prises par le second observateur sur les mêmes individus. On note $(D_1, \cdots, D_n)$ les différences brutes entre ces deux séries de mesures. On pourra une nouvelle fois, à titre d'exemple, se référer aux tables 2 ou 3.

On note $\overline{D}$ la moyenne des différences observées. Le praticien définit alors un seuil Δ au-delà duquel la moyenne des différences lui semblerait *concrètement* significative.

Il s'agit donc désormais de trancher entre les deux hypothèses :
— $\mathcal{H}_0 : |\overline{D}| \geqslant \Delta$, *i.e.* $\overline{D} \geqslant \Delta$ ou $\overline{D} \leqslant -\Delta$ (présence d'un biais significatif) ;
— $\mathcal{H}_1 : |\overline{D}| < \Delta$ (absence de biais significatif).

Remarques. — (i) L'hypothèse nulle est une hypothèse dite *composite* : elle est constituée de la réunion de *deux* inégalités. Sa mise à l'épreuve nécessitera donc *deux* tests unilatéraux, d'où le nom de « Two One-Sided Test ».

(ii) On notera les rôles inhabituels (inversés) joués ici par les hypothèses nulle et alternative : on doit ici rejeter l'hypothèse nulle pour prouver la bonne adéquation des deux séries de mesures.

§3.3. Mise en pratique. — Pour rejeter l'hypothèse nulle et ainsi pouvoir conclure que les deux séries de mesure ne permettent pas de mettre en évidence un biais gênant, on se doit de rejeter, par deux tests unilatéraux « classiques », les deux hypothèses nulles $\mathcal{H}_0' : \overline{D} \geqslant \Delta$ et $\mathcal{H}_0'' : \overline{D} \leqslant -\Delta$.

Reprenons l'exemple de la table 2 p. 3, dont on se rappelle qu'un test apparié de Student avait mis en évidence une différence « significative » (p. 3) :

X	15.3	14.2	18.7	16.9	15.6	12.1	14.5	17.0	16.3	13.4	17.2	15.5	14.2	13.9
Y	15.2	14.0	18.6	16.7	15.5	12.0	14.5	16.9	16.1	13.3	17.0	15.4	14.1	13.8
D	0.1	0.2	0.1	0.2	0.1	0.1	0	0.1	0.2	0.1	0.2	0.1	0.1	0.1

Table 5. — Exemple de biais systématique entre deux séries d'observations.

Ici, on va considérer qu'étant donné l'instrument de mesure utilisé, on fixe $\Delta = 0.3$ comme différence « significative » et gênante. On réalise un premier test unilatéral de Student pour rejeter l'hypothèse nulle $\mathcal{H}_0' : \overline{D} \geqslant 0.3$, et donc accepter l'hypothèse alternative $\mathcal{H}_1' : \overline{D} < 0.3$: on obtient une p-valeur de $1 \cdot 10^{-8}$. On rejette ainsi (fortement !) l'hypothèse selon laquelle le premier observateur surcote systématiquement de plus de 0.3 les valeurs mesurées par le second. Symétriquement, on réalise un second test unilatéral de Student pour rejeter l'hypothèse nulle $\mathcal{H}_0'' : \overline{D} \leqslant -0.3$, et donc accepter l'hypothèse alternative $\mathcal{H}_1'' : \overline{D} > -0.3$: on obtient une p-valeur de $4 \cdot 10^{-13}$. On rejette ainsi (fortement !) l'hypothèse selon laquelle le premier observateur sous-cote systématiquement de plus de 0.3 les valeurs mesurées par le second.

Au final, les deux « composantes » de l'hypothèse nulle $\mathcal{H}_0$ sont rejetées, et on peut donc dire que la différence systématique entre les deux observateurs est cantonnée dans un intervalle $[-0.3, 0.3]$ que l'on a reconnu comme étant « indolore », ou sans importance.

Pour plus d'informations sur cette méthode qui n'a été que brièvement abordée ici, on pourra notamment consulter [Cla] ou [LSS02].

4. Foire aux questions

Voici des réponses possibles à quelques questions non abordées dans ce document et que vous pourriez légitimement vous poser.

Question 1. — Dans ce document, on ne s'occupe que d'évaluer l'erreur intra- ou inter-observateurs pour une *mesure (une grandeur) donnée, c'est-à-dire variable par variable. Or, moi, je dispose de 20 crânes d'hommes modernes ; pour chacun d'entre eux j'ai mesuré 25 grandeurs différentes à 2 reprises, et je souhaiterais évaluer une erreur intra-observateur* globale*, et non mesure par mesure. Je ne veux pas me retrouver avec 25 indicateurs de précision, mais avec un seul ! Comment faire ?*

Prenons les choses dans l'ordre : pourquoi voulez-vous évaluer une erreur intra-observateur globale, et comptez-vous en faire ? Quelques petites choses à avoir à l'esprit :

— Si vous avez autant de mesures différentes ($25 \times 20 \times 2 = 1000$ mesures !), vous ne les avez probablement pas toutes prises dans la même matinée, ni dans la même journée, et peut-être même pas dans la même semaine. Il s'agit d'un premier biais très important, et la première chose que vous risqueriez de mesurer par un indicateur de précision global, c'est qu'il y a un jour où vous étiez mal réveillé ! On ne peut légitimement évaluer les erreurs que pour des mesures prises dans des conditions similaires (c.-à-d. avec le même instrument, durant la même journée, le même état de fatigue, dans le même environnement). Gardez bien à l'esprit que si vous voulez évaluer la reproductibilité d'une nouvelle mesure ou d'un nouveau protocole que vous venez de définir, vous voulez tester l'erreur inhérente au protocole ou à la définition de la mesure, pas l'erreur liée à votre état de fatigue ! Limitez autant que possible les facteurs de confusion dans votre étude...

— Si vos mesures sont d'un ordre de grandeur très différent, à quoi vous servirait réellement un indicateur global ? (De plus, êtes-vous en mesure de supposer que votre erreur de mesure n'est absolument pas dépendante de l'ordre de grandeur de l'objet mesuré ?)

— Plus prosaïquement, en règle générale, on veut connaître les erreurs associées justement à *chaque* mesure. Le but d'évaluer les erreurs intra- ou inter-observateurs est généralement de distinguer les mesures qui seront fiables, facilement reproductibles par d'autres expérimentateurs, de celles qui seront soumises à fortes différences de jugement car elles sont mal définies – et ainsi devront être exclues de l'étude, ou au moins, sévèrement retravaillées ou réexpliquées. Si le calcul de votre potentiel indicateur de précision global n'est associé à aucune prise de décision et n'a qu'un intérêt décoratif, vous pouvez — et devez ! — vous en dispenser. ♢

Question 2. — Pour chaque mesure dont je veux évaluer l'erreur intra ou inter, sur combien d'individus au minimum peut-on se baser pour chaque série d'observations, et/ou combien de réplications effectuer au minimum ?

Sur ce point, le statisticien décrètera évidemment « plus j'ai de mesures, mieux c'est » ; tandis que l'anthropométricien rétorquera « moins j'ai à prendre de mesures, mieux c'est » ! Naturellement, le point de vue de celui qui traite les données s'accorde difficilement avec celui qui doit en baver pour faire les mesures... L'enjeu est d'opérer un compromis entre l'optimalité statistique et la contrainte du temps que prend l'accomplissement d'une série de mesures.

(i) Cas où vous avez un échantillon de base conséquent :

— Si à la base votre travail concerne par exemple un seul os (mettons le tibia) et que votre étude est centrée sur trois grandeurs prises sur le tibia (mettons la longueur, la largeur et le diamètre), vous avez très peu de mesures différentes, et vous pouvez probablement vous permettre de mesurer deux fois de suite 20 individus. Cela ne fait que (?) 120 mesures à prendre en tout et pour tout (2 séries $\times$ 20 individus $\times$ 3 mesures), ce qui semble impliquer un temps de mesure à peu près supportable.

— Si à la base votre travail inclut l'étude d'un très grand nombre de mesures (une quarantaine de mesures différentes, par exemple), il paraît à peu près clair que vous n'aurez pas le temps, l'envie ou la possibilité matérielle de réaliser deux séries de mesures sur 20 individus pour chacune d'entre elles. Un arbitrage sera ici à opérer au cas par cas.

(ii) Cas où vous avez très peu, voire un seul (!) individu à disposition : c'est un cas régulièrement rencontré dans le cadre d'études crâniennes. Si vous n'avez qu'un seul individu à disposition, il suffit d'effectuer beaucoup de réplications de mesures sur cet individu, et/ou de trouver des collaborateurs pour effectuer d'autres séries de mesures. Au bout d'une dizaine de prises de mesures, calculer moyenne, écart-type et coefficient de variation pour chacune des mesures pourra suffire. ♢

Références

[Cla] M. Clark : Equivalence testing. `http://www.unt.edu/rss/class/mike/5700/Equivalence%20testing.ppt`.

[ERJ+08] C. Elie, Y. De Rycke, J.-P. Jais, R. Marion-Gallois et P. Landais : Aspects méthodologiques et statistiques des essais d'équivalence et de non-infériorité. *Revue d'Épidémiologie et de Santé Publique*, (56):267–277, 2008.

[HH05] Ø. Hammer et D. A. T. Harper : *Paleontological Data Analysis*. Wiley-Blackwell, 2005.

[LDL10] P. Lafaye de Micheaux, R. Drouilhet et B. Liquet : *Le logiciel R : Maîtriser le langage, effectuer des analyses statistiques*. Springer-Verlag, 2010.

[Lin89] L. I-Kuei Lin : A concordance correlation coefficient to evaluate reproductibility. *Biometrics*, (45):255–268, 1989.

[LSS02] V. Luzar-Stiffler et C. Stiffler : Equivalence testing the easy way. *Journal of Computing and Information Technology*, (10):233–239, 2002.

[Syl11] Marie-Pierre Sylvestre : Faire et analyser un graphique de Bland-Altman pour évaluer la concordance entre deux instruments ou plus. `http://www.chumtl.qc.ca/userfiles/Image/CENTRE_RECHERCHE/CRCHUM/Documentaions/Services/Janv%202011_Bland-Altman_f.pdf`, 2011.