CRCHUM

EXCELLENCE ■ INNOVATION ■ VALORISATION

Faire et analyser un graphique de Bland-Altman pour évaluer la concordance entre deux instruments ou plus

Par *Marie-Pierre Sylvestre*

Contexte

On désire comparer deux instruments qui mesurent le même concept. Par exemple : deux façons de mesurer la glycémie. Plus précisément, on veut savoir si les deux instruments concordent (agreement).

Données

Nous avons dix sujets sur lesquels les mesures sont prises. Pour chaque sujet de 1 à 20, on a une mesure de chaque instrument à comparer.

Sujets	Instrument 1	Instrument 2
1	13	12
2	14	10
3	11	15
4	10	10
5	12	14
6	14	14
7	12	12
8	11	13
9	9	10
10	11	9

Un graphique de Bland-Altman a comparé les moyennes de mesures à leur différences. On la moyenne entre les deux instruments pour chaque sujet (colonne 3). On calcule ensuite la différence entre les deux instruments pour chaque sujet (colonne 4).

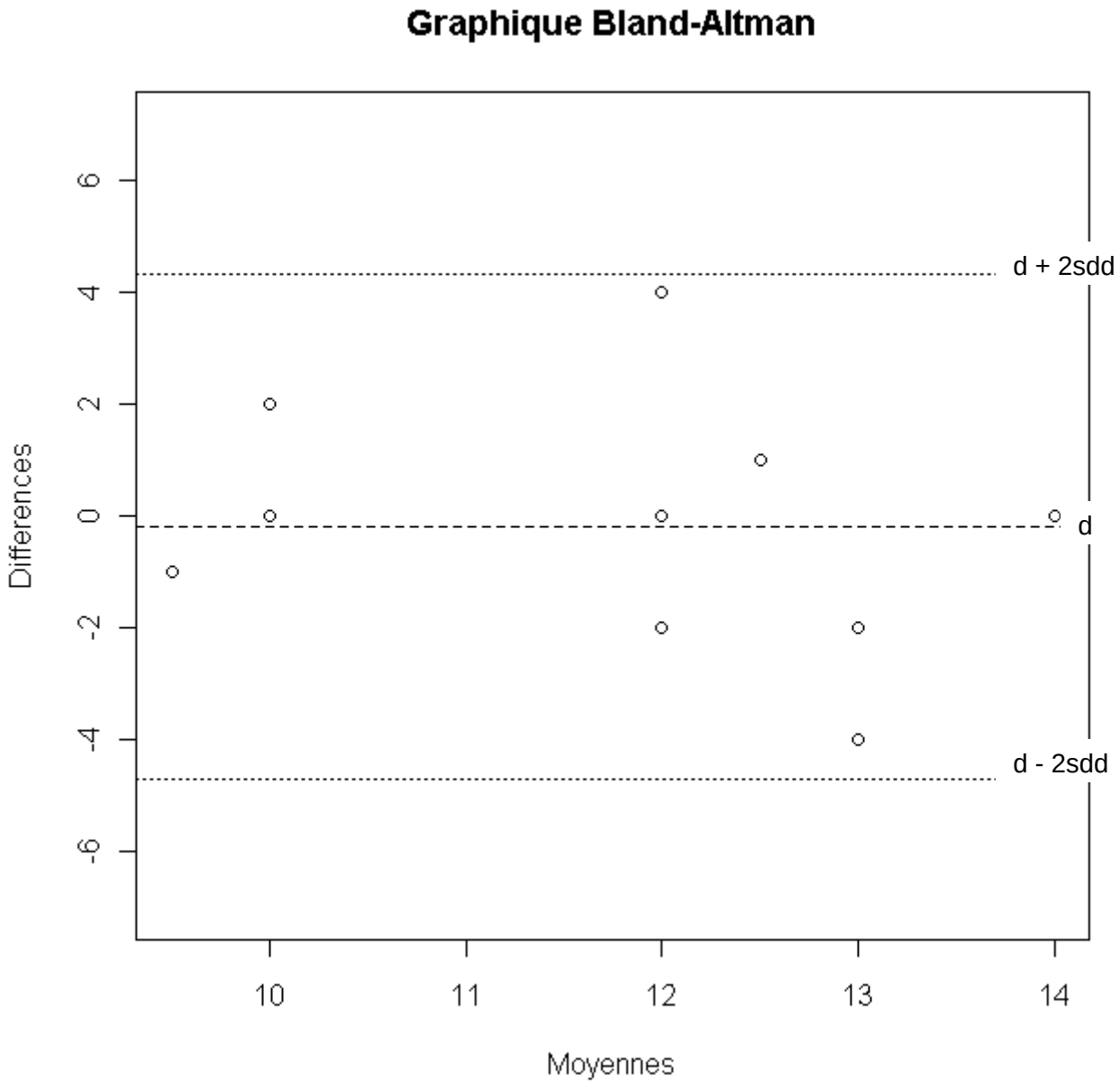
Sujets	Instrument 1	Instrument 2	Moyennes	Différences
1	13	12	(12+13)/2 = 12.5	13 -12 = 1
2	14	10	(10+14)/2 = 12	14 -10 = 4
3	11	15	13	-4
4	10	10	10	0
5	12	14	13	-2
6	14	14	14	0
7	12	12	12	0
8	11	13	11	-2
9	9	10	9	-1
10	11	9	11	2

On fait un graphique dans lequel l'axe des x est la moyenne et l'axe des y est la différence.

Pour calculer ce qu'on appelle les limites d'agrément (*limits of agreement*), il y a trois étapes :

1. calculer d = moyenne des Différences (ici d = -0.2)
2. calculer sdd = l'écart-type des Différences (ici sdd = 2.25)
3. calculer la limite inférieure et supérieure = d ± 2 sdd (ici d ± 2 sdd = -4.7 et 4.3)

FIGURE 1



Interprétation

La moyenne des Différences, d, nous indique si un de nos deux instruments tend à produire des valeurs systématiquement plus basses ou plus élevées que l'autre. Par exemple, ici, d = -0.2. Il semble donc que l'instrument 1 tende à produire des valeurs un peu plus petites que l'instrument 2. Cependant, il faut se questionner si 0.2 est une différence significative (voir la section « Interprétation des limites d'agrément» plus bas).

d est souvent présenté comme le biais. Pourquoi ? On ne connaît pas la vraie mesure pour chaque sujet, tout ce qu'on a, pour chaque sujet, ce sont deux valeurs qui comportent probablement des erreurs de mesure. La moyenne entre ces deux valeurs est notre meilleur estimé de ce que la vraie mesure est. d est la moyenne des Différences, ce qui correspond à la différence entre les moyennes de mesures de l'instrument 1 et celles de l'instrument 2. Si un des instruments est plus fiable que l'autre ou « standard », on peut considérer d comme la mesure du biais car d mesure comment en moyenne les valeurs des instruments diffèrent. Si aucun des deux instruments n'est « standard », alors on peut considérer d davantage comme une différence systématique entre deux instruments qu'un biais.

Interprétation des limites d'agrément

On s'attend à ce que la plupart de nos points se situent dans l'intervalle donné par les limites d'agrément d ± 2 sdd. Il faut donc savoir interpréter ces limites.

Important : les procédures reliées au graphe de Bland-Altman ne sont pas associées à des tests statistiques et ne produisent pas de valeur p. L'interprétation des limites d'agrément se fait en lien avec le contexte de recherche et non en comparaison avec des valeurs « étalon ». Avant de faire le graphe de Bland-Altman, il faut se demander quelle différence de mesure entre nos deux instruments considérons-nous comme acceptable ?

Par exemple, je pourrais considérer que si la différence entre deux mesures obtenues par deux instruments est de plus ou moins 3 unités, alors je considère ces mesures comme similaires ou interchangeables. Sinon, je les considère comme différentes. Si par la suite mes limites d'agrément se situent entre -2 et 2 unités (donc à l'intérieur de mes propres limites de -3 et 3), alors je conclurai que mes deux instruments sont concordants. Si mes limites d'agrément sont de -4 à 4, je conclurai qu'il n'y a pas de concordance (ou faible concordance).

Par conséquent, des limites d'agrément de -10 à 10, par exemple, mèneront à une conclusion de concordance dans certaines applications et de discordance dans d'autres, tout dépendant des unités de mesures utilisées et du contexte de recherche dans lequel les instruments sont comparés.

Intervalles de confiance

On oublie facilement à quel point un graphe de Bland-Altman dépend des données et qu'un autre échantillon produira d'autres limites. Si on veut évaluer le degré de confiance qu'on peut avoir dans notre biais et nos limites d'agrément calculés, on peut calculer des intervalles de confiance autour de ceux-ci.

On vérifie d'abord si nos Différences ont une distribution normale, à l'aide d'un histogramme. Cela peut être difficile si on a peu de mesures. En général, les Différences auront une distribution normale même si les mesures elles-mêmes ne sont pas normalement distribuées. Si, et seulement si, la distribution des Différences semble normale, on peut calculer les intervalles de confiance.

On calcule les valeurs suivantes :

1. l'écart-type empirique de d est $\sqrt{\frac{sdd^2}{n}}$, ou n représente le nombre de sujets (pas le nombre de mesures !). Ici $\sqrt{\frac{sdd^2}{n}}$ = 0.71
2. t_{n-1} = la valeur t correspondant à n-1 degrés de liberté. Ici t_{n-1} = 2.26

On peut ensuite construire l'intervalle de confiance autour de d en calculant $d \pm t_{n-1} \sqrt{\frac{sdd^2}{n}}$. On obtient (-1.81; 1.41) comme intervalle de confiance à 95 % pour d.

Cet intervalle ne correspond pas à celui des limites d'agrément ! Cet intervalle sert à quantifier la confiance que nous pouvons avoir dans notre estimation de d.

On peut faire la même chose pour calculer des intervalles autour des bornes des limites d'agrément mais cette fois, on utilisera l'approximation suivante pour l'écart-type empirique de (d + 2sdd) = $\sqrt{\frac{3 * sdd^2}{n}}$.

Comme c'est généralement le cas, on s'attend à des intervalles de confiance plus étroits si on a un grand nombre de sujets que si on en a peu !

<u>Question connexes</u>

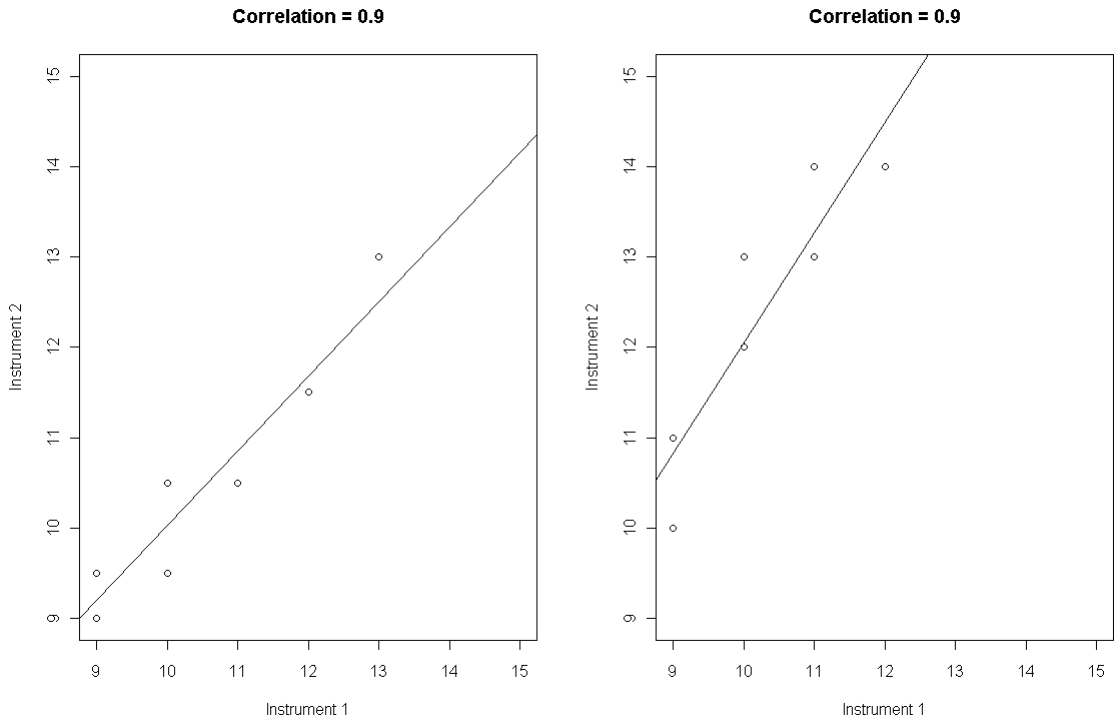
Pourquoi l'utilisation de la corrélation de Pearson pour établir la concordance entre deux instruments peut être trompeuse ?

La corrélation de Pearson est une indication du degré de relation linéaire entre deux variables. Si deux variables sont très corrélées, alors non seulement on peut facilement tracer une ligne entre les points représentant les deux variables, mais ces points seront près de la ligne.

Dans la figure 2, on montre deux graphiques. Chacun d'eux montre les valeurs d'un instrument *vs* celles de l'autre. La seule différence entre les deux graphiques se trouve au niveau des valeurs utilisées pour chacun des instruments. Dans les deux cas, la corrélation

entre les deux instruments est élevée, soit 0.9. Cependant, la ligne qui traverse les points n'a pas la même pente (inclinaison).

FIGURE 2



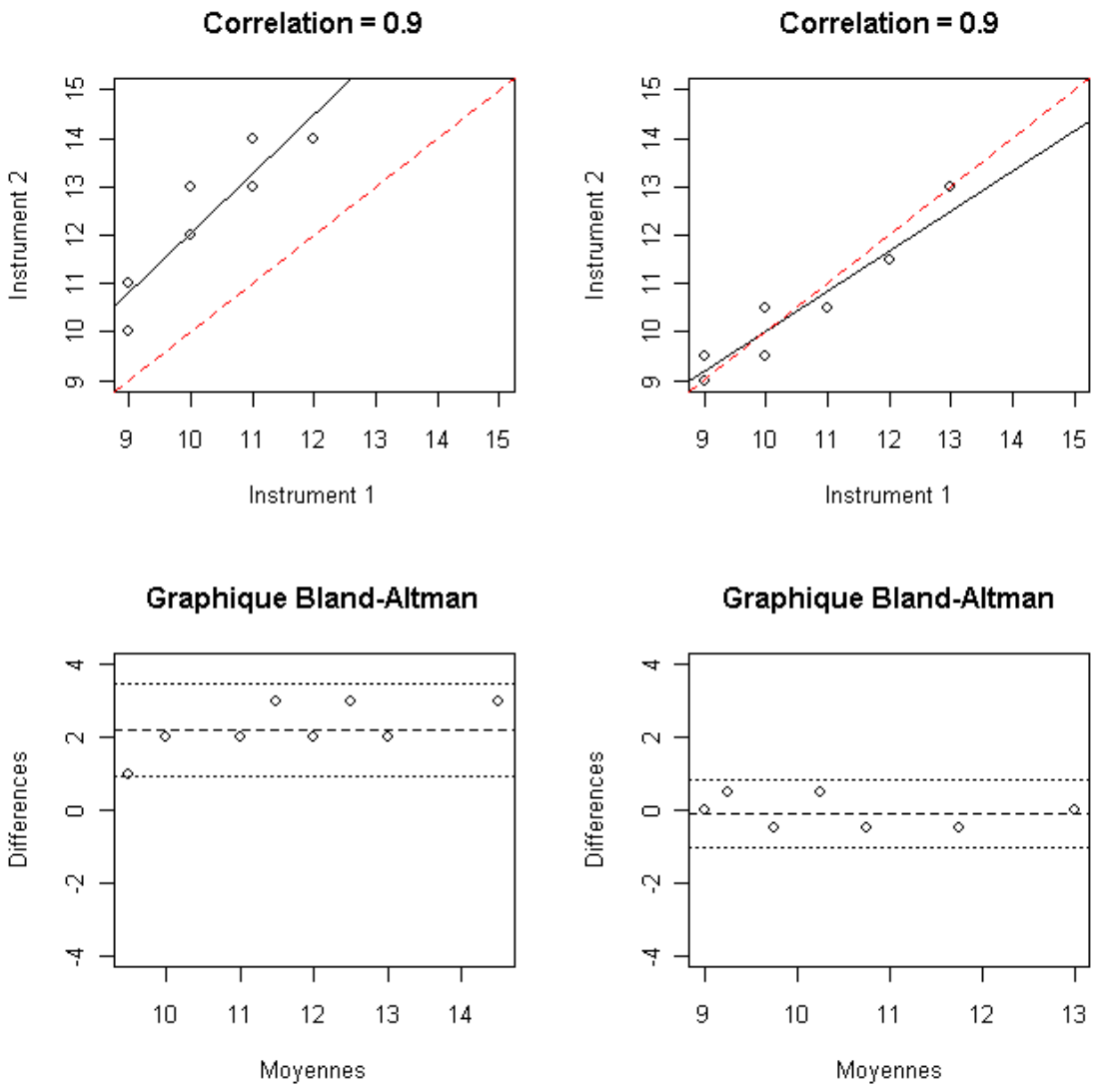
Quand on parle de concordance entre deux instruments ou deux variables, on veut non seulement que ces deux variables soient corrélées, mais aussi que la ligne qui passe entre les points soit une ligne la plus près possible de celle de 45º. Pourquoi ? La ligne de 45º est celle qui passe par les points dont la valeur de l'instrument 1 et 2 sont identiques. C'est la concordance parfaite entre deux mesures. On veut donc s'en approcher le plus possible.

Voyons le lien entre la corrélation et la concordance telle que représentée par les graphes de Bland-Altman. La figure 3 reprend les données de la figure 2. Cette fois, j'ai ajouté les lignes à 45º en rouge. En dessous, j'ai mis les graphes de Bland-Altman correspondant aux graphiques de corrélation.

On voit que dans l'exemple de droite, la ligne de corrélation (en noir) est près de la ligne de 45º (en rouge). Le graphe de Bland-Altman correspondant montre une valeur d très près de 0 et des limites d'agrément assez étroites. À l'inverse, dans l'exemple de gauche, la ligne de

corrélation est assez loin de la ligne de 45º et le graphe de Bland-Altman correspondant montre une valeur d assez loin de 0 et des limites d'agrément assez larges.

FIGURE 3



On voit ainsi que la corrélation n'est pas une mesure adéquate pour la concordance entre deux instruments. De plus, les graphes de Bland-Altman sont utiles car ils nous indiquent s'il y a une différence systématique entre les deux mesures (d près de 0 ou non ?).

Que faire si j'ai plus de deux mesures pour le même instrument ? Parfois, on obtient des mesures répétées sur un même sujet, avec le même instrument, pour évaluer la répétabilité de celui-ci. On a alors des données du type :

Sujets	Instrument 1		Instrument 2	
	Mesure 1	Mesure 2	Mesure 1	Mesure 2
1	12	14	11	10
2	10	14	12	12
3	15	11	12	11
…	…	…	…	…

Pour évaluer la concordance entre l'instrument 1 et l'instrument 2, il suffit de faire la moyenne, pour chaque sujet, entre la mesure 1 et la mesure 2 pour chaque instrument.

Sujets	Instrument 1 Moyenne	Instrument 2 Moyenne
1	13	10.5
2	12	12
3	13	11.5
…	...	…

On refait ensuite la même procédure que pour le Bland-Altman habituel MAIS on doit corriger les sdd pour tenir compte du fait que chaque instrument a été mesuré deux fois par sujet. Si on utilise les formules décrites plus haut pour sdd, ceux-ci seront sous-estimés. Il faut utiliser sdd*, calculé comme suit :

1. on calcule sdd_1 = écart-type des différences entre la mesure 1 de l'instrument 1 et la mesure 2 de l'instrument 1 pour chaque individu.
2. on calcule sdd_2 = écart-type des différences entre la mesure 1 de l'instrument 2 et la mesure 2 de l'instrument 2 pour chaque individu.
3. on calcule sdd_3 = écart-type des différences entre les moyennes des deux mesures pour chaque instrument.
4. on calcule sdd* = $\sqrt{sdd_3^2 + \frac{1}{4}sdd_1^2 + \frac{1}{4}sdd_2^2}$
5. La suite est la même que dans la présentation initiale de la procédure.

Que faire si les points sur mon graphe Bland-Altman sont distribués comme un cône qui serait étroit à gauche et large à droite?

Il n'est pas rare de voir que les différences associées aux petites moyennes sont plus petites que celles associées aux grandes moyennes. Si c'est le cas, les limites d'agrément seront trop larges pour les petites moyennes. On peut considérer appliquer une transformation de log ou exprimer les données en terme de % de différence (axe des y). On refait ensuite toutes les procédures sur les données transformées.

Note historique : le nom de Bland-Altman est notoire mais un des premiers à avoir proposé et popularisé de tels graphiques est John Tukey dans les années 70.

Références :

Bland JM, Altman DG. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, i, 307-10.

Altman DG, Bland JM. (1983). Measurement in medicine: the analysis of method comparison studies. The Statistician 32, 307-17.

Tukey JW (1977). Exploratory data analysis. Addison-Wesley, Reading, MA.