Que faire si les données sont dépendantes?

Analyse d'observations répétées (repeated measurements) et de séries chronologiques

Ulrike Held

Horten Zentrum, UniversitätsSpital Zürich

Dans la plupart des cas, l'analyse de données statistiques présuppose que les observations individuelles sont mutuellement indépendantes. Une méthode d'analyse doit tenir compte du fait que si les données sont dépendantes, leur variabilité est en règle générale plus faible que celle prévue sous hypothèse d'indépendance. On obtient par exemple des données dépendantes en mesurant un paramètre (poids, EEG, tension artérielle, …) chez un seul et même individu de façon répétée et dans des conditions d'observation variables, ou encore en relevant le nombre d'admissions par semaine dans un hôpital, la mortalité mensuelle d'une maladie, la concentration journalière d'un polluant dans l'environnement, le cours des actions ou les données sur le chômage. Si ces données sont datées, on les qualifiera de données chronologiques (ou temporelles). Dans cet article, nous voulons présenter quelques méthodes de représentation graphique et d'analyse applicables aux données dépendantes.

Observations répétées ou *repeated measurements*

Pour obtenir ces données, on réitère en général l'observation de la variable dépendante chez un *même* sujet, par exemple à intervalles répétés ou en variant les conditions d'observation. La variabilité de ces données est alors nettement moins grande que celle d'observations relevées au sein d'un *groupe* d'individus.

A titre d'exemple, considérons une étude croisée d'envergure limitée (cf. U. Held. Quels sont les types de plans d'étude existants et comment les réalise-t-on de façon appropriée?[1]). L'essai randomisé porte sur 10 patients, chez lesquels nous avons déterminé la ligne de base de la tension avant de tester deux médicaments hypotenseurs différents (en réservant une «période de lavage» suffisamment longue). Les données relevées sont exposées dans le tableau 1.

Examinons ces données de plus près dans la représentation de la figure 1.

Pour évaluer l'effet du traitement par rapport à la ligne de base, nous avons utilisé un modèle de régression linéaire avec la pression systolique comme variable dépendante et le type de médicament comme variable explicative (ou indépendante). Les résultats sont énumérés dans le tableau 2.

En fixant le seuil de signification à 5%, la valeur de p de 0,037 pourrait nous faire supposer que l'effet du médicament A est significatif par rapport à la ligne de base. Qu'avons-nous oublié? Nous avons traité l'ensemble des données comme s'il s'agissait d'observations indépendantes relevées sur 30 patients différents. Or il s'agit de 10 patients ayant fait l'objet de 3 mesures chacun, et la variabilité des observations est plus faible si elles sont dépendantes que si elles sont indépendantes. Cela peut conduire à une sous-estimation des *erreurs type*, voire même à des effets faussement significatifs. Dans notre exemple, nous choisirons par conséquent une autre méthode d'analyse, robuste cette fois, qui tienne compte de la nécessité de corriger l'estimation des erreurs type pour des données dépendantes provenant d'un même patient, et nous obtenons les résultats du tableau 3.

A l'évidence, cette dernière méthode montre que l'effet du médicament A n'est plus significatif par rapport à la ligne de base.

Analyse de séries chronologiques

L'analyse d'observations dépendantes porte fréquemment sur des séries chronologiques. En règle générale, les observations ont été répétées régulièrement, à intervalles fixes, pendant une période assez longue. Il est vraisemblable que deux observations soient plus proches l'une de l'autre si elles sont consécutives que si un grand écart de temps les sépare, et nous avons vu plus haut qu'il faut appliquer une méthode d'analyse particulière pour tenir compte d'une telle dépendance. A part cette corrélation, dite sérielle, les données peuvent présenter d'autres tendances temporelles et/ou saisonnières. L'analyse d'une série chronologique peut donc chercher à déceler une variation régulière ou systématique dans les données. Une fois trouvé le modèle statistique (de régression) approprié, on peut par exemple évaluer certains effets comme les tendances temporelles, dans le but de prévoir l'évolution future d'un paramètre en fonction du temps et/ou d'autres variables explicatives. Afin de permettre une approche descriptive des données, il faut, dans un premier temps, représenter la série chronologique dans un graphique ayant le temps pour abscisse.

Prenons l'exemple d'un ensemble d'observations faites en Grande-Bretagne de 1974 à 1979, qui portaient sur la mortalité mensuelle d'hommes et de femmes atteints de bronchite, d'emphysème ou d'asthme (Diggle, 2000). La figure 2 montre le nombre de décès mensuels enregistrés au cours de cette étude; relevons que les

Ulrike Held

L'auteur déclare ne pas être en conflit d'intérêt en relation avec cette contribution.

1 Cet article fait partie de la série sur la biostatistique et sera publié dans Forum à une date ultérieure: 2010;10(41).

Tableau 1. Pression systolique en mm Hg.

Patient nº	Ligne de base	Traitement A	Traitement B
1	144,04	145,92	132,78
2	130,55	102,32	103,93
3	153,03	129,80	155,22
4	165,91	114,47	107,18
5	140,36	111,01	127,86
6	145,52	120,82	107,45
7	120,36	145,62	124,43
8	151,08	103,58	160,31
9	140,67	115,66	145,90
10	124,27	153,11	116,60

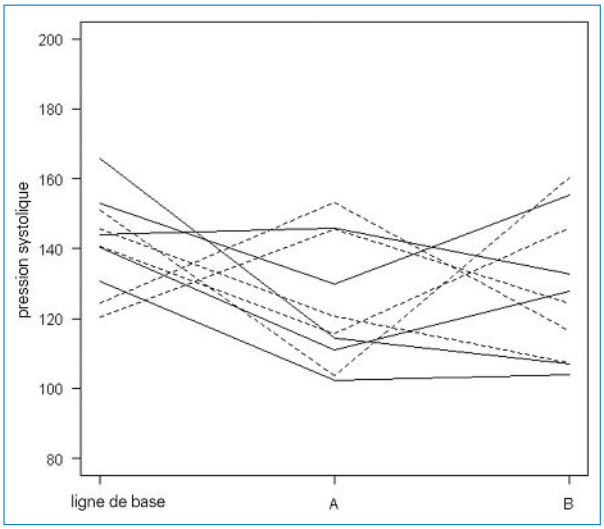

Figure 1
Observations de la pression systolique en ligne de base et sous traitements A et B. Les lignes continues représentent les patients qui ont suivi le traitement dans l'ordre (A, B), les lignes discontinues ceux qui l'ont suivi dans l'ordre (B, A). La répartition était randomisée (c.-à-d. que le choix de l'ordre de traitement était dû au hasard).

Tableau 2. Résultats du modèle de régression linéaire: effet des traitements A et B sur la pression systolique.

	Effet	Erreur type	Valeur de p	IC à 95%
Traitement A	–17,35	7,92	0,037	[–33,60; –1,10]
Traitement B	–13,41	7,92	0,102	[–29,66; 2,84]
Ordonnée à l'origine	141,58	5,60	<0,001	[130,09; 153,07]

Tableau 3. Résultats du modèle de régression linéaire avec erreurs types robustes (estimateur de variance sandwich).

	Effet	Erreur type (robuste)	p-Wert	IC à 95%
Traitement A	–17,35	9,01	0,086	[–37,73; 3,03]
Traitement B	–13,41	7,15	0,093	[–29,59; 2,76]
Ordonnée à l'origine	141,58	4,52	<0,001	[131,36; 151,79]

courbes y représentent le nombre total de décès pendant un mois donné et non les observations individuelles.

Que nous révèlent ces séries chronologiques? Tout d'abord, on voit qu'il existe une différence nette entre hommes et femmes par rapport au nombre absolu de décès mensuels. Ensuite on observe une certaine périodicité annuelle (saisonnière) des données, aussi bien chez les hommes que chez les femmes: il y a plus de décès en hiver qu'en été. Le nombre de décès était particulièrement élevé pendant l'hiver 1975/76 (mois 24 à 26).

On peut «lisser» une série chronologique pour mieux distinguer ses propriétés spécifiques. Par exemple, la figure 3 montre le résultat obtenu en remplaçant, dans l'échantillon féminin, chaque élément de la série (sauf le premier et le dernier) par la moyenne entre l'élément lui-même, celui qui le précède et celui qui lui succède (appelée moyenne mobile d'ordre 3).

On distingue mieux la périodicité saisonnière dans la courbe lissée, car la composante fortuite est plus faible dans une donnée moyenne que dans les données originales (représentées par des étoiles dans la figure 3), qui sont plus dispersées.

Après avoir mis en évidence la composante périodique, intra-annuelle des observations, nous voudrions savoir si ces données présentent également une tendance interannuelle. A cet effet, on peut par exemple remplacer chaque élément de la série par sa moyenne mobile «annuelle». De cette façon, la majeure partie de la périodicité saisonnière est annulée. Dans notre calcul, nous avons utilisé une moyenne mobile pondérée d'ordre 13 dont le facteur de pondération est $w_j = {}^1/_{12}$ pour tous les mois sauf le 1er et le 13e, pour lesquels il se situe à $w_6 = w_{-6} = {}^1/_{24}$.

La figure 4 montre les valeurs moyennes ainsi obtenues. Les deux courbes de mortalité des hommes et des femmes présentent une variation interannuelle indiquant clairement que le nombre de décès diminue avec le temps.

Au-delà de la représentation graphique, les modèles les plus connus pour analyser les séries chronologiques sont les modèles ARMA. Ces modèles supposent qu'une série chronologique est composée d'une partie autorégressive (AR) et d'une partie «moyenne mobile» (*moving average*, MA). La partie AR est une combinaison linéaire d'observations antérieures et la part MA une combinaison linéaire de termes d'erreurs aléatoires antérieurs. On peut par exemple décrire les données sur la létalité de la bronchite, de l'emphysème et de l'asthme au moyen d'un modèle ARMA.

De nos jours, l'analyse des séries chronologiques revêt une grande importance pour la pratique dans l'évaluation de données sur les infections à forte variabilité saisonnière. Par conséquent, il existe des modèles statistiques complexes permettant de détecter le début d'une crise épidémique de grippe ou d'autres maladies infectieuses soumises à l'annonce obligatoire. Par exemple, il est très important de pouvoir prédire l'incidence d'une maladie avec précision pour être en mesure d'évaluer l'efficacité potentielle d'un vaccin.

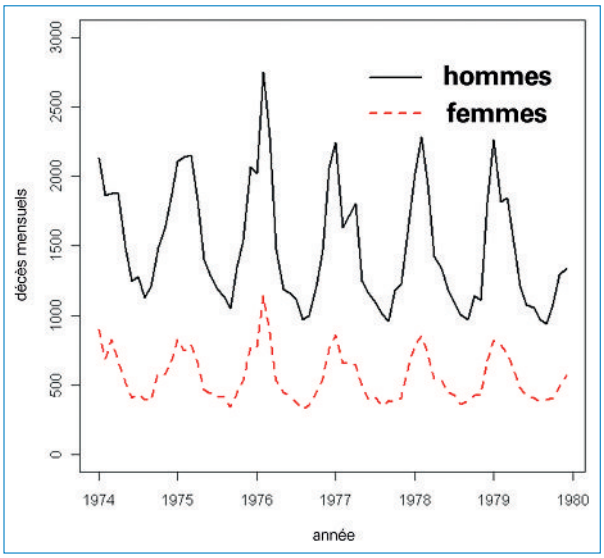

Figure 2
Représentation de 1974 à 1979 de la mortalité mensuelle de la bronchite, de l'emphysème et de l'asthme en Grande-Bretagne.

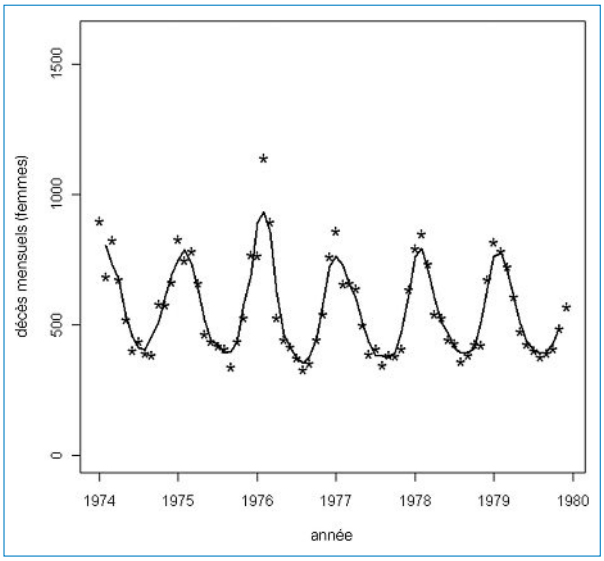

Figure 3
Nombre de décès mensuels chez les femmes: observations (étoiles) et représentation lissée par une moyenne mobile d'ordre 3 (courbe continue).

Glossaire

Etude croisée

Dans ce genre d'étude, chaque sujet suit plusieurs traitements l'un après l'autre. Il faut randomiser l'ordre des traitements et réserver entre eux une période de lavage suffisamment longue pour pouvoir exclure tout effet dû au traitement précédent.

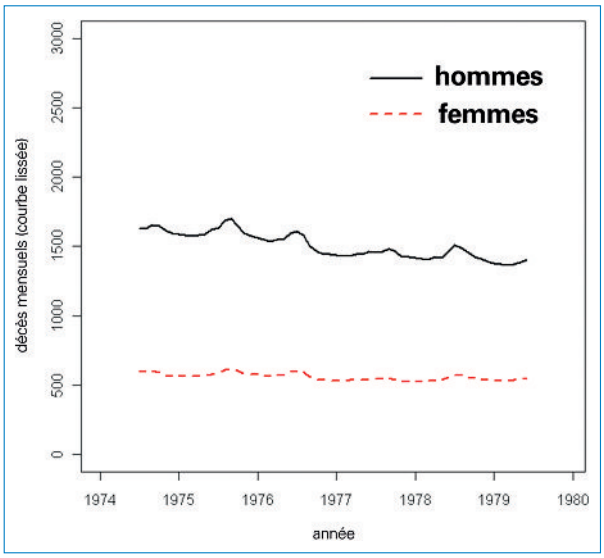

Figure 4
Représentation lissée par une moyenne mobile d'ordre 13 du nombre de décès mensuels chez les hommes et les femmes.

Erreur type

L'erreur type $\sigma_{\bar{x}}$ est l'écart type d'un estimateur de la moyenne. Elle décrit la dispersion d'une distribution d'échantillonnage. Elle est définie par $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, où n représente la taille de l'échantillon et σ l'écart type de la population.

L'erreur type permet de comparer les écarts types de deux séries de données de taille différente, car elle est normée par rapport à la taille des échantillons.

Correspondance:
Dr rer. nat. Ulrike Held
Horten-Zentrum
UniversitätsSpital Zürich
Postfach Nord
CH-8091 Zürich
ulrike.held@usz.ch

Références recommandées

- Diggle P. Time Series – A Biostatistical Introduction. Oxford: Oxford University Press; 2000.
- Everitt B. Time Series. In: Encyclopedia of Biostatistics. P. Armitage P, Colton T, Hrsg. 2. Auflage. New York: Wiley; 2005.
- Held L, Rufibach C, Seifert B. Einführung in die Biostatistik. 4. Auflage. Zürich: Abteilung Biostatistik, Institut für Sozial- und Präventivmedizin der Universität Zürich; Juli 2009. http://www.biostat.uzh.ch.
- Hüsler J, Zimmermann H. Statistische Prinzipien für medizinische Projekte. 4. Auflage. Bern: Huber-Verlag; 2006.
- Kreienbrock L, Schach S. Epidemiologische Methoden. 4. Auflage. München: Elsevier-Verlag; 2005.
- R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN 3–900051–07–0. URL http://www.R-project.org.
- Williams RL. A note on robust variance estimation for cluster-correlated data. Biometrics. 2000:56;645–6.