Quand recourt-on à l'analyse de variance pour comparer des moyennes?

Ulrike Held
Horten-Zentrum, UniversitätsSpital, Zürich

L'analyse de variance, appelée également ANOVA (*an*alysis *o*f *va*riance) est une des techniques les plus utilisées en statistique, et la recherche médicale y recourt fréquemment. Elle permet de tester s'il existe une relation (causale) entre deux phénomènes apparaissant fréquemment de façon simultanée ou présentant des variations similaires sous contexte changeant. Un exemple: l'analyse de variance permet d'étudier si les résultats d'analyse d'hémoglobine varient avec les différentes formes de drépanocytose. Si nous ne faisons intervenir qu'une seule variable explicative, admettons par exemple qu'il s'agisse de la forme de drépanocytose, nous allons effectuer une analyse de variance à un seul facteur *(one-way ANOVA)*. Si au contraire nous y ajoutons un autre facteur, comme par exemple le sexe, nous ferons une ANOVA à deux facteurs *(two-way ANOVA)*. L'analyse de variance représente un cas spécial de la régression linéaire multiple, présentée au début de cette série d'articles.

Notre exemple comporte les résultats d'hémoglobine de 41 patients atteints de drépanocytose (tab. 1). Les données proviennent d'une étude de Anionwu et al. parue en 1981 dans le *BMJ* (dans Kirkwood et Sterne, regardez les références recommandées en bas, p. 82).

Nous allons examiner s'il existe une différence entre les moyennes des trois groupes de patients atteints de drépanocytose. Il faut connaître en premier quelle proportion de la variabilité totale de la variable dépendante (valeur d'hémoglobine) s'explique par des différences entre les groupes (c.-à-d. entre les formes de drépanocytose). En analyse de variance à un seul facteur, on calcule la somme des carrés des écarts:

$$SQA = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad i = 1, \ldots, n$$

que l'on décompose en: 1) la somme des carrés des écarts *entre les moyennes de groupes et la moyenne totale* (SCE intergroupes) et 2) la somme des carrés des écarts *entre les observations et la moyenne* au sein de chaque groupe (SCE intragroupe). Ensuite, nous testons l'*hypothèse nulle* selon laquelle il n'y a pas de différence entre les moyennes intragroupes, en comparant (en langage simplifié) la variabilité *inter*groupe avec la variabilité *intra*groupe. Si la variabilité intergroupe est nettement plus élevée que la variabilité intragroupe, alors l'appartenance au groupe (forme de drépanocytose) a une importance statistiquement significative.

La question qui se pose dans notre exemple est de savoir si les trois groupes drépanocytaires se distinguent par leurs concentrations d'hémoglobine. Par l'ANOVA, il ressort que dans ce cas la variation intergroupe (= 49,945) est beaucoup plus grande que la variation intragroupe (= 0,999), ce qui se reflète par ailleurs dans la très faible valeur de p ($p < 0{,}001$). Nous pouvons donc partir du principe que les concentrations d'hémoglobine à l'état d'équilibre se distinguent selon la forme de drépanocytose du patient. En calculant les valeurs moyennes d'hémoglobine de chaque groupe, nous constatons que la moyenne des observations du groupe Hb-SS est la plus faible (8,713 g/dl), celle du groupe de la thalassémie Hb-S/β se situe à 10,630 g/dl et celle du groupe Hb-SC est la plus élevée avec 12,300 g/dl.

L'analyse des variances repose sur deux hypothèses fondamentales qu'il faudrait toujours vérifier: la première est que la variable dépendante doit suivre une *distribution normale* au sein de chaque groupe, la seconde est que les *écarts types* des données des différents groupes ne doivent pas sensiblement différer. Les déviations par rapport à l'hypothèse de distribution normale ont moins d'influence sur la robustesse du résultat que les déviations par rapport à l'hypothèse de la similarité des écarts types.

Dans l'exemple cité, nous vérifions l'hypothèse de la distribution normale des échantillons à l'aide d'un diagramme quantile-quantile (ou *qq-plot*). En général, nous déconseillons plutôt les tests statistiques de distribution normale. La figure 1 illustre les écarts entre les quantiles de la variable dépendante (observations) et ceux d'une distribution normale (théorie).

La figure 1 montre également que l'hypothèse de la distribution normale des résultats d'hémoglobine est bien vérifiée pour chacun des groupes, car le nuage de points expérimentaux suit à peu près une ligne droite théorique. Les écarts types des trois échantillons sont estimés à 0,844 (Hb-SS), 1,284 (thalassémie Hb-S/β) et 0,942 (Hb-SC). Nous pouvons donc admettre qu'ils sont statistiquement égaux. Si les écarts types avaient été très différents, nous aurions dû faire appel à d'autres méthodes pour contourner le problème, par exemple une transformation de la variable dépendante.

Avec cet exemple, nous avons présenté un cas d'analyse de variance à un seul facteur. Si l'on veut analyser des variances multifactorielles, il faut prêter attention au nombre d'individus de l'échantillon (ou groupe), qui ne doit pas être trop petit.

Au début de cet article, nous avons laissé entendre que l'analyse de variance représentait un cas spécial de la régression linéaire présentée dans le premier article de cette série. Plus précisément, la régression linéaire s'utilise par exemple pour établir la relation entre un facteur d'influence constant, caractérisé par une variable explicative, et l'effet de ce facteur sur la valeur dépendante afin de pouvoir en évaluer l'impact. Si au contraire nous sommes en présence d'un facteur d'influence catégoriel comprenant au moins trois catégories (si les catégories sont au nombre de deux, on peut également recourir au test de t), alors on peut utiliser un modèle de régression linéaire. Cependant l'analyse de variance convient également. Il faut se décider entre la recherche d'explication de la variabilité (analyse

Ulrike Held

L'auteur certifie qu'aucun conflit d'intérêt n'est lié à cet article.

Tableau 1. Concentrations d'hémoglobine à l'état d'équilibre.

Forme de drépanocytose	Hémoglobine (g/dl)
Hb-SS	7,2; 7,7; 8,0; 8,1; 8,3; 8,4; 8,4; 8,5; 8,6; 8,7; 9,1; 9,1; 9,1; 9,8; 10,1; 10,3
Thalassémie Hb-S/β	8,1; 9,2; 10,0; 10,4; 10,6; 10,9; 11,1; 11,9; 12,0; 12,1
Hb-SC	10,7; 11,3; 11,5; 11,6; 11,7; 11,8; 12,0; 12,1; 12,3; 12,6; 12,6; 13,3; 13,3; 13,8; 13,9

de variance) et celle d'une estimation de l'impact du facteur d'influence (modèle de régression linéaire). Nous pouvons nous demander quelle est la raison de l'importance accordée à l'analyse de variance dans la recherche médicale, alors que tous les problèmes traités par analyse de variance pourraient l'être également par régression. L'explication est probablement d'ordre historique: l'analyse de variance fut proposée dans les années 20 par R. A. Fisher, et elle est utilisée encore telle quelle de nos jours; la tendance aux modèles de régression, linéaire ou généralisée, n'est apparue que bien plus tard, en parallèle à l'invention de l'ordinateur qui a facilité le calcul. C'est pourquoi, face à un questionnement sur l'examen de l'influence sur une variable dépendante d'un facteur comprenant plus de deux catégories différentes, la recherche médicale recourt presque toujours au modèle statistique d'analyse de variance plutôt qu'à celui de régression.

Glossaire

Distribution normale

On dit que des données suivent une distribution normale (ou gaussienne) lorsque leur diagramme de fréquence rappelle la forme d'une cloche, qui se caractérise par le fait que les données équidistantes du maximum central ont la même fréquence.

Hypothèse nulle

Dans de nombreuses circonstances, on pose une hypothèse statistique uniquement dans le but de la rejeter. C'est l'hypothèse nulle, ou H_0. En utilisant les observations et un modèle statistique, on peut faire appel à des calculs de tests d'hypothèses permettant d'accepter ou de rejeter l'hypothèse nulle. Si le test d'hypothèses rejette l'hypothèse nulle avec suffisamment de robustesse, on peut partir du principe que l'hypothèse alternative (H_1) a une forte probabilité d'être correcte.

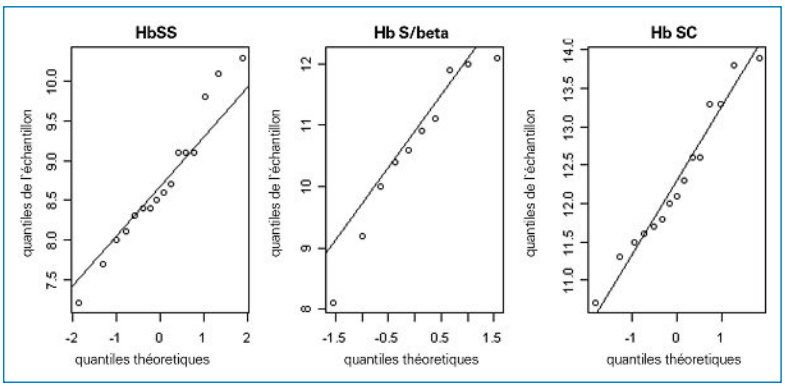

Figure 1
Diagramme quantile-quantile des concentrations d'hémoglobine selon la forme de drépanocytose.

Quantile

Les quantiles sont obtenus par subdivision en parts égales des données triées par ordre croissant ou décroissant. Un quantile indique le nombre de données d'un échantillon se situant de part et d'autre du quantile. Les quantiles les plus couramment utilisés sont les quartiles (les données sont réparties en quatre parts égales), les quintiles (cinq parts égales) ou les percentiles (cent parts égales).

Diagramme quantile-quantile

Le diagramme quantile-quantile (ou *qq-plot*) est la représentation graphique conjointe des quantiles de deux variables. On se sert du *qq-plot* pour comparer la distribution de ces deux variables. Fréquemment, on compare les quantiles d'observations de l'échantillon avec ceux d'une distribution théorique (une distribution normale par exemple) afin de vérifier la correspondance entre observation et théorie.

Ecart type

L'écart type sert à mesurer la dispersion d'une série d'observations. L'écart type est la racine carrée de la variance, et son avantage par rapport à la variance réside dans sa facilité d'interprétation, car il représente un écart exprimé dans les mêmes unités que les observations elles-mêmes.

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

Références recommandées

- 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.
- Kirkwood B, Sterne J. Essential Medical Statistics. 2. Auflage. Massachussets: Blackwell Science; 2003.
- 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.