Deux outils d'évaluation des énoncés en statistique: les tests d'hypothèses et les valeurs de p

Ulrike Held
Horten-Zentrum, UniversitätsSpital, Zürich

Rares sont les publications scientifiques qui ne font pas recours aux tests d'hypothèses statistiques pour répondre aux multiples questionnements de la recherche médicale. Après avoir brièvement décrit l'un ou l'autre test dans les articles précédents, nous allons en expliquer plus à fond les concepts et les appliquer à une série de données sur le poids de naissance des nouveau-nés.

A côté de l'estimation des effets, le test statistique représente *l'instrument central* de toute évaluation de données. Un test statistique permet de répondre à un questionnement, par ex. dans le domaine médical, en appliquant un schéma alternatif de type acceptée/rejetée à une hypothèse dite «nulle». Supposons par ex. que l'on veuille savoir si le statut tabagique de la mère pendant la grossesse exerce une influence sur le poids de naissance du nouveau-né. A cet effet, on émet d'abord l'hypothèse nulle (H_0) comme un énoncé que l'on voudrait réfuter, par ex. «le statut tabagique de la mère n'exerce pas d'influence sur le poids de naissance de l'enfant». L'hypothèse alternative (H_1) correspond alors exactement à l'hypothèse scientifique que l'on voudrait valider, c'est-à-dire dans notre exemple «le statut tabagique de la mère pendant la grossesse exerce une influence sur le poids de naissance de l'enfant». Le test d'hypothèse consiste à départager H_0 et H_1 à partir de données statistiques appropriées.

Il existe deux risques d'erreurs attachés aux résultats du test d'hypothèse: l'erreur α, ou erreur de première espèce, et l'erreur β, ou erreur de deuxième espèce, chacune d'entre elles étant évaluée de manière différente. L'erreur de première espèce consiste à rejeter H_0 alors qu'elle est vraie, ce qui dans notre exemple revient à conclure à tort que H_1 est vraie et que le statut tabagique exerce une influence, alors qu'en réalité il n'en exercerait pas; l'erreur de deuxième espèce survient lorsque l'on accepte H_0 alors qu'elle est fausse, ce qui revient à conclure à tort que H_0 est vraie et que le statut tabagique n'exerce pas d'influence. On considère que l'erreur de première espèce est plus grave: il faut éviter de déclarer à tort qu'une influence existe, qu'un médicament est efficace, etc. Par conséquent, on «maîtrise» l'erreur de première espèce en fixant le seuil de signification α, qui représente la probabilité avec laquelle on est disposé à accepter un risque de première espèce, c'est-à-dire à rejeter H_0 alors qu'elle est vraie. Habituellement, la valeur de ce seuil est fixée à 5% ou à 1%. Si l'on veut en outre que la probabilité de l'erreur de deuxième espèce, ou β, soit faible, c'est-à-dire que la valeur de 1 – β, que l'on nomme puissance *(power)* du test, soit élevée, il faut prélever des échantillons de taille suffisante. Pour un échantillon de taille donnée, la décroissance de α entraîne cependant la croissance de β et vice versa; on ne peut donc pas minimiser conjointement les erreurs α et β et il faut trouver un compromis.

Pour expliquer cette procédure, nous allons effectuer un test t bilatéral sur deux groupes indépendants à partir d'une base de données comprenant le poids de naissance de 189 nouveau-nés mis au monde en 1986 au Baystate Medical Centre à Springfield, Massachusetts. Cette base de données documente également le comportement tabagique des mères de ces enfants et montre que 115 d'entre elles (61%) étaient non-fumeuses.

Nous voulons déterminer s'il existe une distinction entre le poids de naissance des nouveau-nés selon que leur mère était tabagique ou non pendant la grossesse. Formulons tout d'abord l'hypothèse nulle (que nous voulons rejeter): «la moyenne $\mu_{fumeuses}$ des poids de naissance au sein du groupe des nouveau-nés dont la mère a fumé pendant la grossesse est égale à la moyenne $\mu_{non\text{-}fumeuses}$ des poids de naissance au sein du groupe des nouveau-nés dont la mère n'a pas fumé pendant cette période», soit

$$H_0: \mu_{fumeuses} = \mu_{non\text{-}fumeuses}$$

ou encore

$$H_0: \mu_{fumeuses} - \mu_{non\text{-}fumeuses} = 0$$

L'hypothèse alternative s'énonce alors comme suit: «la moyenne des poids de naissance au sein du groupe des nouveau-nés dont la mère a fumé pendant la grossesse est différente de celle au sein du groupe des nouveau-nés dont la mère n'a pas fumé pendant cette période», soit

$$H_1: \mu_{fumeuses} \neq \mu_{non\text{-}fumeuses}$$

Ce test est dit bilatéral car les valeurs moyennes en faveur de l'hypothèse alternative peuvent être indifféremment supérieures ou inférieures à la valeur moyenne caractérisant l'hypothèse nulle; la différence peut donc être positive ou négative.

Ulrike Held

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

Tableau 1. Paramètres statistiques des poids de naissance (en grammes) selon le statut tabagique.

Statut tabagique de la mère	Minimum	1er quartile	Médiane	Moyenne	3e quartile	Maximum
Fumeuse	709	2370	2776	2773	3246	4238
Non-fumeuse	1021	2509	3100	3055	3622	4990

Nous effectuons le test ayant pour seuil de signification $\alpha = 0{,}05$, c'est-à-dire que nous acceptons au plus un risque de 5% de rejeter à tort l'hypothèse nulle. A cet effet, il faut d'abord calculer la valeur d'une statistique de test à partir des données, puis la comparer avec la distribution de la statistique de test sous hypothèse H_0. Si la valeur calculée dépasse une limite critique qui dépend du choix de α et du caractère unilatéral ou bilatéral du test, nous pouvons rejeter l'hypothèse nulle.

Le résultat du test, c'est-à-dire le rejet ou non de l'hypothèse nulle, est fréquemment accompagné de la *valeur de p*, une mesure souvent employée qui représente la probabilité d'observer des résultats au moins aussi en désaccord avec l'hypothèse nulle que ceux qui ont été calculés à partir des données. L'hypothèse nulle est donc rejetée lorsque valeur de $p \leq \alpha$; ce rejet est d'autant plus net que la valeur de p est petite, ce qui explique qu'on utilise la valeur de p pour mesurer le degré de rejet.

Revenons à nos séries de données citées en exemple: le tableau 1 compare quelques paramètres statistiques obtenus à partir des données sur le poids de naissance relevées dans chacun des deux groupes (mère tabagique/non tabagique pendant la grossesse).

De façon descriptive, on constate qu'au sein du groupe des nouveau-nés dont la mère était tabagique, la médiane, le 1er et le 3e quartile du poids de naissance sont moins élevés que les valeurs correspondantes du groupe de comparaison. L'utilisation d'un test t pour échantillons indépendants permet de passer de la comparaison descriptive des valeurs moyennes à leur comparaison statistique. Dans notre exemple, les échantillons indépendants sont justement les deux groupes de nouveau-nés.

Le test t donne une valeur de p égale à 0,007. Cette valeur est inférieure au seuil de signification α que nous avions fixé à 0,05 (ou 5%), et par conséquent nous pouvons rejeter l'hypothèse nulle et affirmer qu'en moyenne, le poids de naissance des nouveau-nés diffère selon le statut tabagique de leur mère. Le poids de naissance moyen se situe à 2773,2 grammes dans le groupe dont les mères sont tabagiques et à 3055,0 grammes dans celui dont les mères ne le sont pas. La différence s'élève à 281,7 grammes, et l'intervalle de confiance à 95% de cette différence est (76,47; 486,96).

Dans notre exemple, les deux groupes comparés sont *non appariés* ou indépendants. Si les groupes étaient dépendants, il faudrait utiliser un *test t pour échantillons appariés* ou dépendants. Par exemple, on utiliserait ce genre de test pour comparer la moyenne des poids de naissance d'un groupe de nouveau-nés avec la moyenne de leurs poids au sixième mois. Dans ce cas, les groupes comparés sont appariés et l'on s'attend à trouver une variabilité plus faible que s'ils ne l'étaient pas.

Dans notre exemple, le test est dit *bilatéral* car il confronte une hypothèse nulle de type «=» et une hypothèse alternative de type «≠». Un test *unilatéral*, quant à lui, oppose une hypothèse nulle «≥» à une hypothèse alternative «<» (ou vice versa). Notre exemple pourrait être reformulé dans le sens d'un test unilatéral:

$$H_0: \mu_{\text{fumeuses}} \geq \mu_{\text{non-fumeuses}}$$

versus

$$H_1: \mu_{\text{fumeuses}} < \mu_{\text{non-fumeuses}}$$

Les revues médicales spécialisées préfèrent généralement refuser les tests unilatéraux, car elles soupçonnent que les données étudiées aient déjà fait l'objet d'un test bilatéral sans donner de résultats probants, et que le test unilatéral n'ait été effectué que par la suite pour obtenir malgré tout un résultat significatif.

Lorsque la *taille des échantillons* est suffisamment grande et/ou que les données suivent approximativement une distribution normale, il est avantageux d'utiliser des tests dits paramétriques fondés sur une hypothèse de distribution. Deux exemples permettant de comparer des moyennes sont le test t décrit plus haut et le test de Gauss. Si la taille des échantillons est petite et que l'on ne sait pas si les données suivent une distribution normale, il vaut mieux recourir à des tests non paramétriques qui ne reposent pas sur une hypothèse de distribution. Dans ce cas, on utilise fréquemment le test de Wilcoxon (comme test de la somme des rangs si les groupes à comparer sont indépendants, et comme test des rangs signés si les groupes sont dépendants) et le test des signes.

Pour terminer, il faut encore veiller à ne pas confondre la significativité statistique d'une étude et son importance dans la pratique (clinique). Lorsqu'un test statistique est conduit de façon correcte sur le plan formel et donne un résultat significatif, il faut toujours en évaluer l'utilité dans la pratique. Il se peut par ex. qu'une grande étude montre qu'il existe une différence significative entre deux groupes comprenant de nombreux patients, et que malgré tout cette différence soit trop faible pour avoir une importance pratique sur le plan médical.

Glossaire

Taille d'un échantillon

Pour répondre à un questionnement scientifique, il n'est en général pas possible d'observer la population entière et l'on observe par conséquent un échantillon de population. Dans la mesure du possible, il faut sélectionner cet échantillon de façon randomisée. Au préalable, une planification formelle doit évaluer le nombre de cas à observer et déterminer la taille des échantillons nécessaire pour permettre à l'analyse statistique de répondre aux objectifs fixés dans l'essai.

Observations appariées/non appariées

Lorsqu'on veut comparer deux groupes d'observations au moyen d'un test statistique, il faut d'abord déterminer si les deux groupes d'observations ont été saisis sur des personnes différentes ou non. Les observations relevées seront dites non appariées ou indépendantes dans le premier cas et appariées ou dépendantes dans le deuxième. Cette différenciation est importante pour le test statistique, parce qu'il faut tenir compte du fait que les observations appariées ont en général une variabilité plus faible que celle escomptée pour des observations indépendantes et qu'il faut les ajuster.

Hypothèses bilatérales/unilatérales

Un couple d'hypothèses bilatérales se présente sous la forme H_0: … = … et H_1: … ≠ … Si le départage des hypothèses tend vers le rejet de l'hypothèse H_0, les valeurs d'échantillonnage pourraient se trouver de part et d'autre de celles caractérisant l'hypothèse nulle. La zone de rejet liée au seuil de signifiance se trouve elle aussi scindée en deux zones égales, par conséquent plus petites et plus éloignées de la valeur caractérisant l'hypothèse H_0 que s'il n'y avait qu'une seule zone.

Un couple d'hypothèses unilatérales cherche à départager H_0: … ≤ … et H_1: … > …, resp. H_0: … ≥ … et H_1: … < … Dans ce cas, la zone de rejet liée au seuil de signification est d'un seul tenant et se situe d'un seul côté de la valeur limite commune aux deux hypothèses. Les hypothèses unilatérales ne sont toutefois que rarement utilisées.

Correspondance:
Dr 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.
– 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.