Significativité statistique

Introduction

Par le biais de la fluctuation d'échantillonnage, même les échantillons aléatoires peuvent être faussés et ne pas représenter la réalité de ce qui se passe dans la population cible. Dès lors, si on observe une association entre une exposition et un événement de santé, comment savoir si cette association est simplement le fruit du hasard ? Tout l'intérêt des échantillons aléatoires est qu'il est possible de raisonner de manière probabiliste, grâce aux lois de distribution statistiques.
Par exemple, si la prévalence du tabagisme en France est de 20 %, on sait qu'il est possible d'obtenir au hasard un échantillon de 100 fumeurs. Mais grâce aux lois de distribution, on c'est aussi qu'un tel échantillon est peu probable[1]. Nous pouvons utiliser ces lois de distribution pour tenter de quantifier la probabilité que les associations observées soient dues au hasard, ou, en d'autres termes, que ces associations sont **statistiquement significatives**. Nous nous contentons bien de *tenter de quantifier* cette probabilité, car son estimation directe est bien souvent impossible. Nous allons voir les deux méthodes de quantification les plus répandues : les intervalles de confiance et le test d'hypothèse.

Intervalles de confiance

La méthode préférée par les épidémiologistes pour déterminer la significativité des associations est l'utilisation des intervalles de confiance. La question fondamentale est la suivante : si je suppose que les estimations faites dans mon échantillon sont vraies, quel est l'intervalle de valeurs qui contiendra *X* % des valeurs observées dans des échantillons de même taille ? Dans la grande majorité des cas, on utilise arbitrairement *X = 95 %*, et on dit que les valeurs en question forment *l'intervalle de confiance à 95 %*. Mais on pourrait tout aussi bien estimer l'intervalle de confiance à 90 %, etc.
Reprenons l'exemple du tabagisme. Si dans un échantillon aléatoire de 100 individus ont observe 20 fumeurs, l'intervalle de confiance à 95 % peut être estimé par : $p \pm 1{,}96\times \sqrt{\frac{p\times(1-p)}{n}}$ où p est la prévalence et n la taille de l'échantillon. Dans l'exemple précédent, l'intervalle de confiance correspondrait à [0,12 – 0,29]. Donc si nos estimations sont vraies, la plupart des échantillons aléatoires auront une prévalence comprise entre ces deux valeurs. On voit également qu'un échantillon composé uniquement de fumeur ne fait pas partie de cet intervalle est est donc potentiellement rare. L'intervalle de confiance peut être envisagé comme un reflet de la **précision** de notre estimation : plus l'intervalle de

---

[1] Pour cet exemple, la probabilité d'un tel échantillon est en effet très faible, à peu près égale à 1.27×10^{-70}...

confiance est large, moins notre estimation est précise et donc moins nous sommes sûr que ce que nous avons observé est la véritable valeur. Pour une moyenne, en supposant la distribution normale, l'intervalle de confiance est estimée par : $m \pm 1{,}96 \times \frac{s}{n}$ où m et s correspondent respectivement à la moyenne et à l'écart-type estimés dans l'échantillon.

Le même raisonnement peut également être appliqué à tout autre statistique observée dans un échantillon, notamment pour des mesures d'association (risque relatifs, odds ratio, etc.). C'est pour ce type de mesure que l'intervalle de confiance permet de déterminer la significativité d'un effet. **Pour un RR ou un OR, par exemple, on parlera d'un effet significatif de l'exposition en question quand l'intervalle de confiance à 95 % ne contient pas la valeur 1.** Cela signifiera qu'on a suffisament de précision dans notre estimation pour dire que 1 ne fait pas partie des valeurs probables. L'effet observé n'est donc probablement pas dû au hasard.

Le petit p (*p-value*)

Mais la façon la plus répandue en recherche biomédicale d'aborder la significativité statistique passe par le cadre du test d'hypothèse. Bien que présentée différemment cette solution est toutefois tout à fait équivalente à celle utilisant les intervalles de confiance. Dans ce cadre théorique, on commence par définir deux hypothèses : l'hypothèse nulle (H_0), selon laquelle l'exposition ou le traitement en question n'a pas d'effet, et l'hypothèse alternative (H_1) qui stipule l'inverse. Le principe du test d'hypothèse est de déterminer si l'hypothèse nulle est plausible. Le cheminement est très proche de celui de suivi pour l'intervalle de confiance : on calcule une statistique reflétant l'effet observé, qu'on appelle **statistique de test**. À la différence avec les mesures d'association qu'on utilise généralement dans le cadre des intervalles de confiance, ces statistiques sont choisies car elle possède une loi de distribution connue. Ainsi, la statistique χ^2 utilisé dans le test « du Khi2 » est moins humainement interprétable qu'un risque relatif, mais sa loi de distribution statistique est bien connue. Cela permettra de déterminer facilement une probabilité particulière nommée le **petit p**, ou valeur p (*p value* en anglais), qui correspond à la probabilité d'observer un effet au moins aussi important si on accepte que l'hypothèse nulle est vraie. Si cette probabilité est faible, la plausibilité de l'hypothèse nulle est alors faible elle aussi. Tout comme dans l'intervalle de confiance, on garde 95 % des valeurs les plus plausibles. En d'autres mots, si le p est inférieur à 5 %, l'hypothèse nulle est peu « vraisemblable » et on conclut donc à une différence significative[2]. Dans le cas contraire, l'hyptothèse nulle n'est pas rejetée, et il est alors impossible de conclure. En effet, ce n'est pas parce qu'un phénomène n'a pas été observé qu'il n'existe pas.

---

[2] Si les formulations sont différentes, nous rappelons que les deux approches sont en fait tout à fait équivalente : un risque relatif dont l'intervalle de confiance à 95 % ne contient pas 1 correspond à un test du Khi2 significatif au seuil de 5 %, et inversement.

Facteurs influençant la significativité

Les mesures de significativité présentent en fait un mélange de deux informations : la taille de l'effet (ou force d'association) et la taille de l'échantillon. Ainsi, avec des échantillons de très grande taille, une différence insignifiante peut apparaître comme statistiquement significative. Il convient dès lors de se souvenir que la significativité statistique d'un effet n'a rien à voir avec sa pertinence médicale ou clinique. Prenons l'exemple d'un traitement antitumoral, qui permet un survie significativement plus longue des patients que le traitement standard. La différence de survie est cependant faible, en moyenne de 5 jours. Bien que significatif, la pertinence de ce résultat devrait être considéré : est-ce que 5 jours de vie gagnés peuvent être considérés comme la preuve d'une réelle supériorité du nouveau traitement ? Les effets secondaires et la qualité de vie du patient ont-ils été pris en considération ? Etc.

Le mélange d'information sur les tailles d'effets et d'échantillons est ce qui amènent les épidémiologistes à préférer les intervalles de confiance aux petits p. Prenons l'exemple de deux risques relatifs associés à un petit p inférieur à 0,0001. La seule information qu'on peut tirer du petit p est que les RR sont significativement différents de 1. Regardons maintenant les intervalles de confiance à 95 % : l'un est égale à [1,001 - 1,002] et l'autre à [3,752 - 17,687]. Avec les intervalles de confiance sous les yeux, il n'est pas difficile de voir que le premier RR est de toute évidence très proche de 1, puisque les deux bornes de l'intervalle de confiance sont proches de 1. La significativité ici est plus probablement due à une taille d'échantillon élevée et l'effet mesuré est probablement nul, et sinon négligeable. En revanche, le second RR est estimé moins précisément mais l'effet semble ici être important, et la significativité est moins due à la taille de l'échantillon que dans le premier cas.

Depuis que les progrès informatiques ont permis l'essor d'autres approches statistiques, les méthodes dites « fréquentistes » que nous venons de voir ont été critiquées, plus particulièrement le petit p pour les raisons que nous venons de voir. Ces méthodes restent cependant de loin les plus répandues malgré leurs défauts. Il est toutefois possible que les autres approches statistiques, notamment les statistiques bayésiennes, prennent petit à petit plus de place dans la recherche biomédicales.

Puissance et risques d'erreur

Comment trouver le bon compromis entre taille de l'effet et taille de l'échantillon ? Tous dépend des risques d'erreur que nous sommes prêts à prendre. Il existe en effet 2 manières de se tromper avec les tests statistiques : soit on considère comme significatif un effet qui n'existe pas, soit on considère comme non significatif un effet réel. Ces erreurs correspondent respectivement aux erreurs de type I, ou erreur de première espèce et aux erreurs de type II, ou erreurs de seconde espèce. Le risque de commettre une erreur de type I est appelé **risque α**, le risque de commettre une erreur de type II est appelé **risque β**. À ce

dernier, on préfère souvent utilisé la probabilité complémentaire qu'on appelle la **puissance** : c'est la probabilité de considérer comme significatif un effet réel[3].

L'enjeux est alors de trouver le bon compromis entre risque α et β. Si on veut être sûr que les effets considérés comme significatifs sont bien réels, il faut un risque α faible. On risque dans ce cas de ne considérer comme significatifs que des associations fortes. Si des effets plus faibles existent, ils ne seront pas considérés comme significatifs et le risque β est donc augmenté (et la puissance diminuée). Au contraire, si on cherche à mettre en évidence même les plus petits effets, on voudra un risque β faible (et une puissance élevée). De petites fluctuations aléatoires pourront alors être considérées comme significatives, et le risque α augmentera. En recherche, on fait souvent le choix de considérer une résultat faussement significatif comme plus grave qu'un résultat faussement négatif, et on choisit donc le seuil en fonction du risque α. Comme nous l'avons vu précédemment, le seuil de significativité d'un test est souvent choisi pour pour correspondre à un risque α inférieur à 5 %.

Une fois le risque α fixé, ne reste plus qu'à déterminer deux paramètres parmi la puissance, la taille de l'effet et la taille de l'échantillon pour estimer le troisième paramètre. Ainsi, si on veut déterminer combien de patients devraient être inclus dans un essai thérapeutique, il faudra déterminer quelle puissance on cherche à obtenir (généralement autour de 80 %) mais aussi quelle différence entre les traitements on s'attend à observer (la taille de l'effet). Dans les études épidémiologiques plus exploratoires, dans lesquelles la taille de l'échantillon n'est pas connue a priori, la puissance statistique peut être déterminé à la fin de l'analyse. Elle permettra de juger si l'absence de significativité d'un effet peut être due à une puissance statistique trop faible.

---

[3] En d'autres termes la puissance est égale à 1 - β.