

I) La recherche clinique : un impossible obligatoire ?

Participer, prolonger, initier, diriger des travaux de recherche pour répondre aux multiples questions qui restent encore en suspens peut sembler irréaliste pour le clinicien confronté à ses difficultés quotidiennes. Cela peut même paraître impossible s'il travaille dans une structure d'Urgence, synonyme d'immédiateté, de réponse en temps réel « Time is life » dans un contexte où les flux des patients, les motifs de recours et les pathologies ainsi que la charge en soins sont par définition non contrôlables.

Et pourtant, comment être médecin professionnel sans comprendre et participer à la recherche médicale ? Aussi chargée soit la barque, on ne peut tenir la barre en regardant ses pieds. Il faut regarder derrière, regarder sur les côtés et regarder devant. Regarder derrière, c'est analyser sa pratique. Regarder sur les côtés, c'est suivre l'évolution du paysage scientifique en perpétuelle évolution. Regarder devant, c'est « pré-voir » les difficultés, envisager les différentes solutions, prendre un critère de jugement, mesurer, analyser le résultat et le communiquer aux autres pour trouver ensemble le chemin le plus sûr. C'est participer activement à des travaux de recherche.

L'implication des cliniciens dans la recherche est une orientation nationale et internationale qui se décline à tous les niveaux :

- Dans la formation initiale où l'analyse de travaux de recherche par la lecture critique d'article scientifique est un enseignement obligatoire sanctionné par une épreuve lors de l'examen national classant en France.
- Dans les formations spécialisées où chaque DES ou DESC se termine par la présentation d'un mémoire et/ou d'une thèse qui doit suivre les règles de bonne pratique de recherche scientifique à la fois dans sa réalisation et dans sa présentation au mieux réalisée sous la forme d'un article scientifique.
- Dans la formation médicale continue qui, plus qu'une obligation légale, est une obligation morale et professionnelle : offrir aux patients ce qui est le plus adapté à leurs besoins.
- Dans l'évaluation et l'amélioration des pratiques professionnelles qui conduit à confronter la pratique quotidienne avec les données actuelles du savoir médical. Pour cela il nous faut rechercher, comprendre et analyser les données scientifiques issues de la recherche.
- Dans l'évaluation des cliniciens et des centres à orientation universitaire où l'analyse des publications scientifiques est prise en compte, y compris sur le plan financier.

Ce guide prend pour exemple la recherche en Médecine d'Urgence, non pas parce que les règles et conseils pratiques de recherche clinique qui y sont décrits sont spécifiques à cette discipline médicale et ne s'adressent qu'aux urgentistes, bien au contraire. En montrant que la recherche en médecine d'Urgence est aujourd'hui effective et reconnue par des publications internationales, il prouve que la recherche par et pour les cliniciens est possible quelles que soient leurs conditions d'exercice.

Il a pour but de nous aider à y participer, chacun à son niveau, en évitant les pièges pour aller jusqu'au bout du chemin, là où on est récompensé des efforts accomplis.

II) Bases du raisonnement scientifique et de l'analyse statistique

1) Introduction

La démarche médicale scientifique est basée sur deux notions fondamentales : l'incertitude et l'estimation du risque. Elle s'oppose ainsi à la démarche dogmatique qui nie l'incertitude et à la démarche empirique où le risque n'est pas estimé. Pour le scientifique, la quantification du risque d'erreur est un préalable à la décision diagnostique, thérapeutique voire tout simplement descriptive. C'est lorsqu'elle s'appuie sur une comparaison du risque calculé avec le risque acceptable, qu'une décision médicale devient une décision basée sur des preuves.

L'appréciation du risque s'appuie sur une analyse précise et répétée de données afin d'appréhender au plus juste une réalité inconnue, c'est-à-dire sur un travail de recherche.

On peut définir par activité de recherche médicale toute évaluation ou expérimentation comportant une quantification d'un risque et dont le but ou l'un des buts est d'élargir les connaissances médicales et de participer à l'amélioration de la qualité des soins. Il s'agit donc de répondre à une question, éventuellement plusieurs, en analysant les caractéristiques des sujets ou les effets d'une ou plusieurs interventions, actives ou passives, réalisées sur des malades, des personnes saines, des soignants, des structures, des animaux, du matériel vivant ou inerte et en précisant le niveau de confiance qu'il est possible d'accorder à la réponse (quantification du risque).

La recherche fait appel à un langage (déterminants, variables) et à des techniques spécifiques (tests statistiques). Ces termes et tests épidémiologiques et statistiques sont devenus les compagnons quasi-obligatoires des médecins. Actuellement, de nombreux logiciels de statistiques sont disponibles et permettent, grâce à la puissance des ordinateurs, un accès de plus en plus facile à des outils statistiques de plus en plus complexes. La maîtrise de ces outils par des utilisateurs non avertis est souvent approximative.¹ Pourtant, des connaissances simples sur les statistiques permettent d'appréhender de nombreux problèmes courants et de fixer la limite au delà de laquelle le recours à un professionnel de la biostatistique est indispensable. En outre, ces connaissances élémentaires sont indispensables au clinicien pour comprendre et évaluer les articles scientifiques publiés.

POINTS-CLEFS

La démarche scientifique consiste à prendre des décisions dans une situation d'incertitude en évaluant le risque d'erreur.

Elle s'appuie sur des travaux de recherche dont la caractéristique est de répondre à une question en mesurant la précision de la réponse et/ou le niveau de confiance à lui accorder (risque d'erreur).

La recherche fait appel à un langage et des techniques spécifiques : les tests statistiques.

2) Les principes généraux d'un test d'hypothèse

Les tests statistiques sont utilisés chaque fois que l'on veut comparer un paramètre à une valeur de référence ou plusieurs paramètres entre eux. On s'intéresse alors à l'écart existant entre le paramètre et la référence ou entre les différents paramètres. La vraie valeur de cet écart est inconnue mais on peut formuler une hypothèse la concernant, l'hypothèse nulle étant que cet écart n'existe pas. Les tests statistiques permettent d'estimer les risques d'erreur face à l'acceptation ou au rejet de cette hypothèse nulle.

Un test d'hypothèse constitue un outil d'aide à la décision permettant de choisir entre deux hypothèses (Tableau 1).^{2, 3} Formuler une hypothèse est ainsi la première étape de toute analyse statistique.

Soit un analgésique A couramment utilisé et un nouvel agent B et nous nous posons la question de savoir si la proportion de patients soulagés est différente entre A et B. L'effet de l'agent A est-il différent de celui de B ?

Notons Δ la différence des effets des agents A et B. On peut définir une hypothèse dite hypothèse nulle, notée H_0 , pour laquelle Δ est égal à 0. L'hypothèse nulle est souvent le complémentaire d'une hypothèse alternative que l'on veut démontrer (hypothèse H_1). Celle-ci stipule, par exemple, que Δ est différent de zéro : les effets des deux agents sont différents sur la proportion de patients soulagés. On remarque que l'hypothèse alternative correspond à une multitude de situations couvrant toutes les possibilités où la différence Δ des effets entre A et B est distincte de 0.

Dans l'exemple précédent, l'hypothèse nulle concerne le pourcentage P de patients soulagés par le traitement B. Elle consiste à supposer que P est égal au pourcentage de patients soulagés par A, que nous noterons P_{H_0} . Si ce pourcentage vaut ici 60 %, l'hypothèse nulle s'écrit : $H_0 : P = P_{H_0} = 0,6$.

L'hypothèse alternative (H_1) est la nouvelle hypothèse que l'on propose pour décrire la réalité si H_0 est fausse. Elle correspond à une valeur de P que nous noterons P_{H_1} .

Selon ce que l'on sait du pourcentage P_{H1} , l'écriture peut prendre l'une des trois formes suivantes :

H_1 :

- $P \neq P_{H_0}$ (test de différence)
- $P < P_{H_0}$ (test d'infériorité)
- $P > P_{H_0}$ (test de supériorité)

Le choix de l'une de ces alternatives dépend de la façon dont le problème se pose.

Choisir entre H_0 et H_1 c'est prendre le risque de se tromper par rapport à la vérité que l'on cherche à découvrir.² En effet, nous sous-entendons que soit H_0 est vraie, soit H_1 est vraie. Notre décision de choix entre H_0 et H_1 est donc associée à un risque de se tromper lui-même caractérisé par deux probabilités (Tableau 1) :

- le risque de première espèce, α , qui est la probabilité de rejeter H_0 alors que H_0 est vraie ;
- le risque de deuxième espèce, β , qui est la probabilité de ne pas rejeter H_0 alors que H_1 est vraie.

Tableau 1 : Types d'erreur et test d'hypothèses.

Décision	Réalité	
	H_1 est vraie	H_0 est vraie
Rejet de H_0	$1 - \beta$	α
Non rejet de H_0	β	$1 - \alpha$

H_0 est l'hypothèse nulle, H_1 l'hypothèse alternative.

α est le risque de première espèce, β le risque de deuxième espèce, $1 - \beta$ la puissance du test.

POINTS-CLEFS

La comparaison d'un paramètre à une valeur de référence ou la comparaison de plusieurs paramètres entre eux est à la base de la recherche.

La valeur réelle du paramètre étudié étant inconnue, une hypothèse doit être formulée.

Lors de la comparaison entre 2 interventions, on teste l'hypothèse nulle correspondant à l'absence de différence entre elles. L'hypothèse alternative est l'existence d'une différence entre les 2 interventions.

Les tests statistiques permettent d'estimer les risques d'erreur face à l'acceptation ou le rejet de l'hypothèse nulle.

3) Le risque α de première espèce

3.1. Définition

Le risque α de première espèce correspond à la probabilité de rejeter l'hypothèse H_0 (et donc d'accepter l'hypothèse alternative H_1), alors que H_0 est vraie (tableau 1). C'est la pierre angulaire des tests statistiques.⁴ En effet, la crainte principale de l'investigateur dans une démarche expérimentale est de conclure à tort à l'hypothèse alternative H_1 .

La valeur seuil de 0,05 pour le risque α est communément admise pour trancher entre H_0 et H_1 . Elle est arbitraire. Il n'existe pas en effet de différence "fondamentale" entre $p=0,06$ et $p=0,04$. Un risque α inférieur à 0,01 est parfois choisi.

Le risque α étant considéré comme essentiel dans les tests d'hypothèses, il est important de s'assurer que le risque consenti lors de l'analyse est effectivement celui qui était prévu initialement. Le risque alpha augmente avec la multiplication des analyses et la valeur seuil doit être définie en prenant en compte l'ensemble des analyses réalisées.

3.2. Situations justifiant le recours à un biostatisticien

Une bonne analyse statistique doit être définie *a priori* et non *a posteriori*. La méthodologie statistique d'une investigation clinique ou expérimentale doit être définie lors de la conception initiale du projet. L'avis d'un statisticien est envisagé selon le degré de complexité du problème abordé dès la conception du projet.

Cinq situations à risque d'erreur justifient de faire appel à un biostatisticien.

3.2.1. Comparaisons entre plusieurs traitements

Prenons l'exemple de la comparaison de trois agents analgésiques (B, C et D) à l'agent A. La proportion de patients soulagés est de 60 % avec B, 70 % avec C, et 80 % avec D. L'analyse statistique réalisée en comparant A vs B, A vs C, A vs D conduit à conclure que seul D est différent de A. Du fait des comparaisons multiples, il faut bien noter que le risque de première espèce consenti n'est plus de 0,05 mais de 0,11 (3 comparaisons). Si l'on comparait 100 agents analgésiques à l'agent A, avec un risque $\alpha = 5\%$, on pourrait attendre que 5 d'entre eux soient différents de A, seulement par hasard ! L'analyse statistique doit tenir compte des comparaisons multiples afin de garantir *in fine* un risque global de première espèce de 0,05.⁵

3.2.2. Analyses intermédiaires

Dans un protocole où l'on compare deux agents analgésiques, une analyse est réalisée tous les 50 patients inclus. Ce type d'analyse séquentielle groupée permet de ne pas prolonger indûment une étude dont les résultats apparaissent significatifs avant que l'ensemble de l'effectif total prévu ait été inclus. Après 200 patients inclus, il y a 60 patients soulagés ayant reçu

A et 75 patients ayant reçu B. On rejette l'hypothèse nulle et on conclut que B est meilleur que A. Cependant, pour cette analyse finale, les trois analyses précédentes infructueuses ont été oubliées. Tous calculs faits, le risque de première espèce consenti n'est pas de 0,05 mais en fait de 0,13.⁶ L'analyse statistique aurait dû tenir compte des comparaisons multiples effectuées dans cet essai. Ces analyses intermédiaires doivent être prévues *a priori* dans le protocole d'un essai pour garantir *in fine* un risque α de 0,05.

3.2.3. Mesures répétées

Imaginons que nous comparions le pourcentage de patients soulagés par A ou B toutes les heures pendant les 24 premières heures. Là encore, si on désire maintenir un risque α de 0,05, il faut tenir compte du fait que 24 comparaisons successives ont été effectuées, ou se contenter d'une analyse globale sur l'ensemble des 24 heures. Plusieurs stratégies d'analyse sont envisageables dans le cadre de mesures répétées.⁷

3.2.4. Analyse par sous-groupes

Nous avons comparé la proportion de patients soulagés entre A et B et conclu à l'absence de différence significative. Il vient alors l'idée de répartir les patients en différents sous-groupes (hommes et femmes, patients de plus ou moins de 75 ans, suivant la pathologie douloureuse). Là encore, l'analyse par sous-groupes expose au risque de conclure à tort à une différence significative si des tests séparés sont réalisés par sous-groupes et doit donc tenir compte de la multiplicité des comparaisons. De plus, une analyse par sous-groupes suppose que ceux-ci soient comparables (selon A et B), que ceci ait été prévu à l'avance dans l'élaboration du protocole et que les interactions soient prises en compte dans l'analyse. Une interaction, au sens statistique, décrit une situation pour laquelle l'impact d'un facteur sur la réponse mesurée dépend de la valeur d'un autre facteur.⁸⁻¹⁰

3.2.5. Critères de jugement multiples

Dans une même étude comparant les agents analgésiques A et B, nous étudions plusieurs variables différentes (dont la proportion de patients soulagés) au risque $\alpha = 0,05$. La réalisation de tests séparés pour chaque critère de jugement augmente le risque de faux résultats positifs.⁶ C'est pourquoi une bonne étude s'efforce de ne répondre qu'à une seule question principale (B soulage-t-il plus de patient que A ?) avec un seul critère de jugement (la proportion de patient soulagé), dit critère de jugement principal.

3.3. La valeur de p et sa signification

Une fois l'hypothèse nulle définie, on peut évaluer la probabilité d'obtenir les résultats observés si l'hypothèse nulle est vraie. Cette probabilité est dénommée p.

Il faut bien comprendre la différence entre le risque d'erreur α et le degré de signification p . Le risque d'erreur est une caractéristique du test qui fixe le pourcentage de cas où on conclura au rejet de H_0 alors que H_0 est vraie. Il reste identique d'un échantillon à l'autre. α est défini *a priori* et p est calculé *a posteriori* à partir des observations recueillies sur un échantillon particulier (il y a autant de degrés de signification que d'échantillons différents). La valeur p mesure, d'une certaine manière, l'écart entre cet échantillon et l'hypothèse H_0 . Lorsque p est plus petit ou égal au risque α , on rejette l'hypothèse H_0 et l'hypothèse H_1 est acceptée.

La valeur de p est souvent considérée à tort comme la probabilité que l'hypothèse nulle soit vraie. Cette opinion est erronée : p est la probabilité d'observer une différence au moins aussi importante que celle observée dans l'essai, sous l'hypothèse nulle.

La valeur de p est souvent perçue comme une mesure du poids de l'évidence d'un test statistique : "plus p est petit plus la différence entre A et B est grande". C'est faire abstraction de la nature des hypothèses testées et du contexte de l'étude. Considérons un essai où les sujets reçoivent un traitement A et un traitement B. On leur demande ensuite quel traitement ils ont préféré. Les deux essais suivants aboutissent à un même p de 0,041. Dans un premier cas 15 sujets préfèrent A et 5 préfèrent B, dans un deuxième cas 1.001.445 préfèrent A et 998.555 préfèrent B. Dans le premier cas le taux de préférence est de 75 % mais l'échantillon est trop restreint pour considérer la conclusion comme fiable. Dans le deuxième cas on considérera que les deux traitements sont très proches car le taux de préférence n'est que de 50,07 %. Plus un essai comprend un nombre élevé de sujets plus la chance de montrer une certaine différence statistiquement significative augmente. La pertinence clinique du résultat doit bien sûr être examinée. La valeur de p doit être replacée dans son contexte, en rappelant que statistiquement significatif n'est pas synonyme de cliniquement pertinent. La pertinence clinique d'un résultat doit être envisagée selon l'amplitude de la différence observée. Nous verrons plus bas l'intérêt de l'intervalle de confiance dans ce cadre.

POINTS-CLEFS

Le risque α de première espèce correspond à la probabilité de rejeter l'hypothèse nulle alors que celle-ci est vraie. Ce risque est déterminé *a priori*.

La valeur seuil (arbitraire) du risque α est habituellement de 0,05. Elle correspond au risque maximum accepté pour rejeter l'hypothèse nulle. Le risque α augmente avec le nombre d'analyses.

La probabilité d'obtenir les résultats observés si l'hypothèse nulle est vraie est dénommée p . Elle est calculée *a posteriori* et comparée à la valeur α .

La pertinence clinique sera à mettre en rapport avec l'amplitude de la différence observée.

Les situations complexes requièrent l'aide d'un biostatisticien.

4) Le risque β de deuxième espèce

Le risque β correspond à la probabilité d'accepter l'hypothèse nulle H_0 , alors que H_1 est vraie (tableau 1). Or l'hypothèse H_1 est que l'effet de l'agent analgésique B diffère de celui de A, ce qui correspond à une infinité de possibilités. Il n'est donc pas possible de calculer β dans l'absolu. On calcule β pour une différence donnée (Δ) entre les agents A et B.³ La valeur $1-\beta$ (Δ) est appelée puissance du test.

Dans la présentation des résultats d'un test statistique, il est habituel de présenter α et p et d'oublier β (Δ). En effet, il est considéré comme plus grave d'accepter à tort une nouvelle hypothèse (B différent de A) que de conserver la situation B non différent de A à tort. Or, lorsque le test statistique ne permet pas de rejeter l'hypothèse nulle, le risque β (Δ) devient essentiel. Les résultats négatifs doivent en effet être analysés avec beaucoup de précautions. Le principe général est de considérer que l'on a rejeté l'hypothèse H_1 provisoirement (faute de mieux) et qu'il est alors important de considérer la force de conviction de ce résultat négatif. Dans l'exemple des analgésiques, si 5 patients ont été inclus dans chaque groupe, 4 (80 %) ont été soulagés dans le groupe A et 2 (40 %) dans le groupe B, le test statistique conclut à l'absence de différence significative entre ces deux groupes. Pour autant, il est évident que ce résultat négatif est intuitivement peu convaincant du fait du faible nombre de patients inclus. L'absence d'évidence d'un effet n'est pas l'évidence de l'absence de cet effet. Ne pas mettre en évidence un effet dans un échantillon ne signifie pas qu'aucun effet n'existe en réalité. Tout essai thérapeutique doit faire état du calcul préalable du nombre de patients nécessaire et de la puissance de l'essai. C'est un des critères de bonne pratique méthodologique.¹¹

Il est devenu habituel d'utiliser un risque β inférieur ou égal à 0,20. Il faut néanmoins souligner que dans certaines situations (maladies rares, difficulté de réalisation de l'étude) on peut sacrifier délibérément β en raison du faible nombre de patients qu'il est possible d'étudier. C'est un choix parfaitement acceptable mais qui doit être explicite dans la rédaction de l'article et les problèmes inhérents à la fragilité de l'acceptation de l'hypothèse nulle (puissance insuffisante) doivent être clairement exposés dans la discussion.

POINTS-CLEFS

Le risque β de deuxième espèce correspond à la probabilité d'accepter l'hypothèse nulle alors que l'hypothèse alternative est vraie. On le calcule pour une différence donnée entre les deux interventions thérapeutiques analysées.

Il permet de déterminer la puissance du test.

Il est habituel d'utiliser un risque β inférieur ou égal à 0,20.

5) Les types de variable

On distingue plusieurs types de variables, schématiquement les variables qualitatives et les variables quantitatives. Les variables qualitatives comprennent les variables nominales et ordinales. Les variables nominales ont deux modalités (homme/femme), ou plus de deux modalités (insuffisance rénale aiguë pré-rénale / rénale / post-rénale). S'il y a une structure d'ordre entre les catégories d'une variable on parle de variable ordinale (grade de l'insuffisance cardiaque NYHA). La différence entre deux catégories adjacentes n'est pas forcément homogène sur toute l'étendue de la variable (score de Glasgow). Les variables qualitatives sont généralement représentées sous forme de pourcentages ou de proportions avec leurs intervalles de confiance.

Les variables quantitatives sont de deux types, soient continues lorsqu'elles peuvent prendre toutes les valeurs d'un continuum (âge, glycémie, pression artérielle), soient discrètes lorsqu'elles prennent des valeurs entières (nombre de transfusions, nombre de grossesses). Les variables ont une distribution normale lorsque leur répartition suit une courbe de Gauss. Elles sont alors représentées par leur moyenne et l'écart-type. Les variables ne suivant pas une distribution normale, sont représentées par la médiane, les extrêmes et/ou les interquartiles.

POINTS-CLEFS

Les Variables Qualitatives sont des variables nominales ou ordinales. Elles sont décrites sous forme de pourcentages ou de proportions avec leur intervalle de confiance.

Les Variables Quantitatives sont des variables continues ou discrètes. Si la distribution est normale (courbe de Gauss), elles sont décrites sous forme de moyenne avec écart-type. Sinon elles sont décrites par leur médiane avec les extrêmes ou interquartiles.

6) Les types de test

6.1. Test unilatéral ou bilatéral

Le plus souvent, l'hypothèse nulle H_0 correspond à l'égalité entre A et B et l'hypothèse alternative H_1 correspond à l'inégalité entre A et B. Toutefois, si B est différent A, rien ne dit si B est supérieur ou inférieur à A. La plupart du temps, les tests statistiques considèrent les deux éventualités ($B < A$ ou $B > A$) et on dit que le test est effectué en situation bilatérale.¹¹ On peut imaginer d'autres situations : A est un placebo et B un antalgique par exemple. Dans ce cas, on considérera uniquement l'éventualité $B > A$, le test réalisé est dit alors unilatéral. Dans certaines circonstances, l'utilisation de tests unilatéraux est judicieuse. L'industriel qui effectue un criblage de nombreuses molécules potentiellement actives réalisera ces études de manière unilatérale. L'objectif est ici de ne pas ignorer une

molécule potentiellement active. Cependant, d'une manière générale, il est recommandé d'utiliser les tests statistiques en situation bilatérale, à moins d'avoir des raisons précises de choisir une situation unilatérale qu'il faut alors justifier dans le protocole.¹²

Le choix d'un test bilatéral ou unilatéral doit toujours être fait a priori, jamais au vu des résultats. C'est la condition pour que le risque d'erreur α reste effectivement fixé à 5 % et ne devienne pas égal à 10 %.

POINTS-CLEFS

Il est recommandé d'utiliser en général des tests statistiques en situation bilatérale.

Les tests statistiques en situation unilatérale peuvent être utilisés lorsque le sens de la différence entre 2 interventions est connu (traitement versus placebo).

6.2. Test d'équivalence

Dans certains cas, l'hypothèse que l'on souhaite tester n'est pas l'efficacité différentielle de deux (ou plusieurs) traitements mais leur équivalence.^{13, 14} Par exemple, on veut montrer qu'un nouveau traitement B d'emploi plus aisé, présentant moins d'effets secondaires ou moins coûteux est équivalent à un traitement de référence A. Dans ces essais d'équivalence, il convient de spécifier δ , la plus grande différence cliniquement acceptable entre les deux traitements. Le nouveau traitement B sera considéré comme non équivalent au traitement A de référence si l'hypothèse nulle est vraie ($H_0 : |B-A| \geq \delta$). Le traitement B sera considéré comme équivalent à A si l'hypothèse alternative ($H_1 : |B-A| < \delta$) est retenue.

On réalise également des tests unilatéraux dans les essais d'équivalence.^{13, 14} On parle alors de test de non infériorité. Le nouveau traitement B sera considéré comme non équivalent au traitement A de référence si l'hypothèse nulle ($H_0 : B \geq A + \delta$) est vraie. Le traitement B sera considéré comme équivalent à A si l'hypothèse alternative ($H_1 : B < A + \delta$) est retenue.

POINTS-CLEFS

Pour démontrer qu'un nouveau traitement mieux toléré est équivalent en termes d'efficacité avec un traitement validé, on réalise un test d'équivalence et on détermine δ comme la plus grande différence acceptable entre les deux traitements.

Il y a équivalence si la différence entre les deux traitements est inférieure à δ .

Le test de non infériorité est un test unilatéral dans les essais d'équivalence.

6.3. Analyse des intervalles de confiance

Le plus fréquemment, on ignore la valeur vraie du paramètre étudié dans la population. On ne connaît que les observations faites sur les sujets *d'un seul échantillon*. L'estimation vise alors à calculer, à partir de ces observations, la vraie valeur du paramètre dans la population. L'existence des fluctuations d'échantillonnage ne permet pas de remplir totalement cet objectif. On sait, en effet, que des échantillons de composition différente peuvent être observés dans une même population. On ne peut donc obtenir qu'une valeur approchée (moyenne ou pourcentage). On tient compte des fluctuations d'échantillonnage en estimant l'intervalle dans lequel la vraie valeur du paramètre est contenue avec une probabilité fixée a priori. On l'appelle l'intervalle de confiance. Par exemple, l'intervalle de confiance à 95 % de la moyenne d'un échantillon représente l'étendue des valeurs qui contient la vraie valeur de la moyenne dans la population avec une probabilité de 95 %.

Il faut savoir que la taille de l'intervalle de confiance est d'autant plus petite, c'est-à-dire la précision avec laquelle le paramètre est estimé est d'autant plus grande, que le nombre de sujets de l'échantillon est grand. Ainsi, le rôle joué par les effectifs est facilement appréciable avec les intervalles de confiance. Ce n'est pas toujours le cas avec un test d'hypothèse. Certains journaux privilégient la présentation des résultats sous la forme d'intervalles de confiance.¹⁵

POINTS-CLEFS

La valeur vraie d'un paramètre dans une population est inconnue.

A partir d'un échantillon, on ne peut obtenir qu'une valeur approchée (moyenne ou pourcentage) contenue dans un intervalle de confiance qui représente l'intervalle dans lequel se trouve cette valeur avec une probabilité fixée a priori (en général 95%).

La précision de l'estimation du paramètre est directement dépendante de la taille de l'échantillon.

7) Choix d'un test statistique

Décider du choix du test le plus approprié est un aspect important des statistiques (Tableau 2).² Il est nécessaire de définir préalablement les conditions dans lesquelles le ou les tests statistiques seront employés, donc de définir *a priori* les hypothèses testées, les risques α et β consentis et le caractère unilatéral ou bilatéral de l'analyse.

Pour guider son choix parmi les nombreux tests statistiques disponibles, il sera nécessaire de prendre en compte plusieurs éléments.¹⁶

Tableau 2 : Etapes préalables à la réalisation d'un test statistique.

1	Type de variable étudiée (quantitative, qualitative) ?
2	Distribution sous-jacente de la variable (normale, binomiale, poisson, autres) ?
3	Définition des conditions du test d'hypothèse, risque de première espèce, puissance, différence attendue ?
4	Unilatéralité ou bilatéralité du test ?
5	Petit ou grand échantillon ?
6	Séries appariées ?
7	Test paramétrique ou non paramétrique ?
8	Nombre de traitements comparés ?
9	Analyses intermédiaires prévues ?

7.1. Type de variables considérées

La nature de la variable analysée est le premier paramètre. Pour les variables quantitatives, on fera appel à des tests de comparaison de moyennes ou de comparaison de variance (test t de Student, test de Wilcoxon,...). Pour les variables qualitatives, on fera appel à des tests de comparaison de répartition (test du Chi^2 : X^2 , test exact de Fisher,...).

7.2. Conditions d'application du test

La plupart des tests statistiques ne sont utilisables que dans des conditions bien définies. Il s'agit de la nature de la distribution de la variable, des effectifs, ou de conditions plus particulières comme l'égalité des variances l'indépendance ou non des variables, etc. (tableau 2).

Pour les variables qualitatives, le test du Chi^2 (X^2) n'est utilisable que lorsque le nombre de patients est suffisant. En effet, chaque case du tableau croisé doit avoir un effectif théorique calculé ≥ 5 . Dans les autres cas, le test exact de Fisher est utilisé.

Pour les variables quantitatives, les tests paramétriques comme le test de Student ne peuvent être utilisés que lorsque leur distribution est normale. Lorsque la distribution n'est pas normale, seuls les tests non paramétriques peuvent être utilisés. Il convient ainsi de vérifier préalablement si l'hypothèse de normalité de la distribution est acceptable. Ceci est possible lorsque l'effectif est suffisamment important (> 20). Plusieurs procédures permettent de vérifier la normalité.¹ De nombreux auteurs conseillent de se contenter de vérifications graphiques de la normalité.^{3, 4}

Il est probablement possible de limiter cette vérification aux situations où la nature même de la variable fait suspecter une distribution non normale.

¹ Des transformations de variables, par exemple de type logarithmique ($\log x$) ou inverse ($1/x$), peuvent permettre de se ramener à une distribution normale. Un exemple de tests paramétriques et des équivalents non paramétriques est donné dans le tableau 3.

Tableau 3 : Exemples de tests paramétriques et de leurs équivalents non paramétriques pour les variables quantitatives.

Test paramétrique	Test non paramétrique
Test t de Student non apparié	Test de Mann et Whitney ou de Wilcoxon
Test t de Student apparié	Test de Wilcoxon apparié
Analyse de variance	Test de Kruskal et Wallis*
Corrélation linéaire ou coefficient de corrélation de Pearson	Coefficient de corrélation de Spearman

* cas particulier de deux variables

7.3. Séries appariées

La comparaison de deux moyennes (de lois normales) entre deux groupes de patients peut faire appel au test t de Student. Que se passe-t-il si ces moyennes proviennent du même groupe, mesurées à deux temps différents ? On dit que ces mesures sont appariées. L'utilisation du test t de Student usuel n'est plus appropriée car les deux séries de mesure ne sont pas indépendantes sur le plan statistique. Il convient d'utiliser un test qui prenne en compte le fait que la mesure a été effectuée deux fois sur les mêmes patients comme un test t de Student pour séries appariées (ou un test non paramétrique de Wilcoxon pour séries appariées). Dans le cas de plusieurs mesures (>2) répétées au cours du temps, des stratégies variées peuvent être retenues.⁷

7.4. Comparaisons multiples

A partir du moment où la comparaison statistique est effectuée sur plusieurs moyennes (plus de 2), il n'est pas possible d'utiliser un test prévu pour comparer deux moyennes une seule fois ou deux médianes. Il convient d'utiliser une méthode statistique qui garantisse la conservation du risque α . Sans entrer dans le détail des nombreux choix possibles, on peut indiquer quelques pistes. L'analyse de variance permet de tester si n moyennes sont différentes entre elles, sous certaines conditions (mesures indépendantes, données de distribution gaussienne, même variance entre groupes). Si c'est le cas on peut s'interroger alors pour savoir quelle moyenne diffère de quelle autre. On comprendra qu'en fonction du nombre de groupes testés, le nombre de comparaisons peut être important. Aussi des tests particuliers ont été proposés pour la réalisation de ces comparaisons dites a posteriori ou post hoc. Si on compare plusieurs groupes à un groupe de référence on utilisera le test de Dunnett. Si on compare plusieurs groupes entre eux, on utilisera un test de Tukey ou un test de Newman-Keuls. Il est toujours possible aussi d'utiliser la correction dite de Bonferroni. Elle consiste à diviser le risque α par le nombre n de comparaisons à tester. Si on compare 3 valeurs de pression artérielle entre elles (1 avec 2, 1 avec 3 et 2 avec 3), le risque α sera de $0,05/3 = 0,016$. Si une valeur de p est inférieure à 0,016, on écrira alors que la différence est significative au risque de 0,05.

POINTS-CLEFS

Pour choisir le test approprié, il importe de connaître les conditions dans lesquels le test sera appliqué (risque α et β , unilatéral ou bilatéral) et de vérifier les points suivants :

- Le type de variable étudiée (quantitative ou qualitative).
- La taille de l'échantillon ou des groupes (conditionnant l'applicabilité des tests). Pour utiliser un test χ^2 , le nombre de patients doit être suffisant pour que chaque effectif théorique calculé soit > 5 . Sinon, on utilise un test de Fischer. Pour les variables quantitatives, on peut faire l'hypothèse de normalité lorsque l'effectif est important.
- La distribution des variables quantitatives (normale, binomiale, poisson,...) Si la normalité de la distribution est vérifiée, on pourra utiliser un test de Student, sinon on devra utiliser des tests non paramétriques.
- Le caractère apparié ou non apparié des données : des tests spécifiques doivent être utilisés pour les séries appariées (test de Student apparié, test de Wilcoxon pour séries appariées).
- La réalisation éventuelle de comparaisons multiples : Il faut pouvoir garantir que le risque α reste à 0,05, ceci nécessite de faire appel à des tests spécifiques (analyse de variance, test de Dunnett, test de Tukey ou Newman-Keuls) et des mesures correctives (correction de Bonferroni = division du risque α par le nombre de comparaisons à tester).

D'une manière générale, le recours à un biostatisticien est souhaitable pour toute analyse complexe.