CHAPITRE 16

NOTIONS DE STATISTIQUES POUR LE CLINICIEN

Pierre Duhaut, Claire Andréjak

Les cliniciens sont souvent réfractaires aux statistiques. Or, elles sont indispensables pour arriver à observer tous les événements non forcément visibles 'à l'œil nu', et arriver à différencier la fréquence des événements auxquels on s'intéresse, par rapport à une survenue aléatoire. Les statistiques sont inutiles pour mettre en évidence l'action des antibiotiques sur la méningite à méningocoque, ou sur la tuberculose. Elles seront nécessaires pour mesurer l'incidence des accidents vasculaires cérébraux, pathologies graves s'il en est, sous anti-hypertenseurs par comparaison à des patients non traités. La différence entre les deux situations est que dans la première, on s'intéresse à une action rapide et individuelle, presque immédiatement visible. Dans la seconde, on s'intéresse avant tout à une tendance sur un groupe, qui se traduira par une amélioration à long terme du pronostic d'une partie seulement de la population traitée : le domaine des statistiques est celui de l'évaluation de tendances perceptibles uniquement au niveau d'échantillons ou de populations, mais pouvant bénéficier également, sur le plan clinique, à certains des patients traités ou observés à défaut de tous.
Ce chapitre passe en revue les questions cliniques les plus fréquentes et leur traduction en langage et en tests statistiques.

Biostatistiques et méthodologie sont indissociables : les données recueillies doivent être analysées… et doivent donc être recueillies de telle sorte à être analysables. La construction d'une étude doit prévoir le type d'analyse à faire en fonction de la question posée, et la façon de colliger les données doit obéir à deux impératifs parfois contradictoires, qu'il faudra concilier :

- Décrire la réalité au plus près,
- Rendre ces données compatibles avec un test statistique existant permettant la comparaison la plus exacte entre les groupes étudiés.

I. Questions posées, variables à colliger, analyse à faire

Partons d'une question clinique banale et d'une base de données effective pour rendre le propos plus intelligible :
Les thrombocytoses sont fréquemment rencontrées en médecine clinique. Un vieil aphorisme affirme qu'une thrombocytose supérieure à $1.10^6/mm^3$ correspond 'à tous les coups' à une thrombocytémie essentielle (TE) (syndrome myéloprolifératif). Le diagnostic étiologique peut cependant être difficile à poser, y compris par la biologie moléculaire (la mutation JAK2 n'est présente que dans 60 % des cas environ), et la biopsie de moelle peut ne pas montrer de myélofibrose évocatrice. Un 'gold standard' pourrait être le diagnostic porté après évolution du patient et élimination -si possible- de toutes les causes réactionnelles.
L'ensemble des thrombocytose supérieures à 600 000/mm^3 d'un hôpital ont été réunies dans une série. Cette base de données doit d'abord être décrite, et les facteurs prédictifs de diagnostics peuvent ensuite être analysés.

II Variables dichotomiques (proportions) :

Mille quarante sept patients ont été inclus : il faut connaître la distribution des sexes, le nombre de patients avec thrombocytose réactionnelle (TRe) ou essentielle (en définissant chaque classe comme exclusive de l'autre). Ces variables sont dites **dichotomiques** car elles ne peuvent prendre que deux valeurs (masculin-féminin, essentielle-réactionnelle, vrai-faux, oui-non). Elles sont également **catégorielles,** sans hiérarchie entre les deux valeurs. Elles permettent de calculer la **proportion** de femmes et d'hommes atteints par l'une, ou l'autre, des grandes causes de thrombocytose.

Application :
Notre série comprend 509 femmes et 538 hommes. Le diagnostic de TRe est posé chez 357 femmes (70,1 % de 509) et 461 hommes (85,7 % de 538). Ce taux est-il *significativement différent* entre les hommes et les femmes ? Autrement dit, 85,7 % - 70,1 % est-il égal à (ou proche de) 0 ?
Nous venons de poser *l'hypothèse nulle* : on conclura, si la différence entre les deux taux n'est pas significativement différente de 0, que les deux taux sont proches l'un de l'autre, similaires, et qu'il n'y a pas de différence entre patients de sexe masculin et féminin. Au contraire, si la différence est significativement différente de 0, les deux proportions, et par conséquent les deux groupes, seront estimés différents.
Les données peuvent s'exprimer dans une table 2×2 :

Table 1 : nombre de patients effectivement observés dans chaque catégorie

	Femmes	Hommes	Total
TE	152 *(29,9% de 509)*	77 *(14,3% de 538)*	229 *(21,9% de 1047)*
TRe	357 *(70,1% de 509)*	461 *(85,7% de 538)*	818 *(78,1% de 1047)*
Total	509 *(48,6% de 1047)*	538 *(51,4% de 1047)*	1047

1. Test de chi-2

Le **test de chi-2** permet la comparaison de proportions. Comme tout test statistique, son principe repose sur l'hypothèse nulle : s'il n'y a pas de différence entre hommes et femmes, alors hommes et femmes ne composent qu'un seul groupe dans lequel 229 patients (21,9 %) présentent une thrombocytémie essentielle, et 818 (78,1 %) une thrombocytose réactionnelle. Le test de chi-2 va mesurer, pour chaque cellule, le nombre de patients d'écart entre les patients *observés* ('O'tels qu'ils sont donnés par l'étude effectuée) et *attendus ('A')* si les proportions dans l'ensemble du groupe, et dans chaque sous-groupe, étaient égales. La table devient donc :

Table 2 : nombre de patients attendus dans chaque cellule)

	Femmes	Hommes	Total
TE	509 x 21,9 % (proportion attendue de TE dans les 1047 patients) = 111,5 (nombre Attendu = A)	538 x 21,9 % (proportion attendue de TE dans les 1047 patients) 117,5 (nombre Attendu = A)	229
TRe	509 x 78,1 % (proportion attendue de TRe dans les 1047 patients) = 397,5 (nombre Attendu = A)	538 x 78,1 % (proportion attendue de TRe dans les 1047 patients) = 420,5 (nombre Attendu = A)	818
Total	509	538	1047

Cette nouvelle table présente plusieurs caractéristiques notables par rapport à la précédente :

- Les totaux n'ont pas changé.
- Lorsque le nombre de patients attendus dans la cellule 1 est fixé, les nombres attendus dans les cellules 2, 3 et 4 sont déterminés sans liberté aucune : la somme des lignes, et la somme des colonnes, doit rester constante (car fixées par le nombre de patients effectivement observés dans chaque ligne et chaque colonne !).
- Donc, la détermination du nombre de patients attendus n'a pu se faire qu'avec *un seul degré de liberté*, celui de la cellule 1.

La somme des écarts entre observés et attendus pourrait s'écrire : **Σ (O - A)**.
On remarquera que cette somme est égale à 0 : effectivement, les patients retirés à une cellule ont été ajoutés à la cellule voisine : l'écart de la première cellule est l'opposé exact de la deuxième, celui de la troisième, l'opposé de la quatrième.

Pour contrer cet inconvénient, chaque écart est élevé au carré. La somme devient donc : **Σ (O - A)^2**.
Un écart absolu ne peut cependant à lui seul, rendre compte d'une différence : l'écart entre 10000 et 10002 est égal à celui entre 2 et 4. L'augmentation est de 1/5000 dans le premier cas, et de 50 % dans le second... il faut donc rapporter l'écart à un dénominateur. Le calcul du chi-2 le ramène au nombre d'événements attendus, et la formule du chi-2 devient :

$$\text{chi-2} = \frac{\Sigma (O - A)^2}{A}$$

Dans notre exemple,

$$\text{chi-2} = \frac{\Sigma (O - A)^2}{A} = \frac{(152-111{,}5)^2}{111{,}5} + \frac{(77-117{,}5)^2}{117{,}5} + \frac{(357-397{,}5)^2}{397{,}5} + \frac{(461-420{,}5)^2}{420{,}5} = 37$$

Il est ensuite possible de consulter les tables de distribution du chi-2 *pour un degré de liberté* et d'estimer quelle est la probabilité pour que 37 soit similaire à 0. Cette probabilité est largement inférieure à 1/1000 (en fait, inférieure à 10^{-7} !)

Le seuil de probabilité pour admettre l'hypothèse nulle est habituellement fixé, par convention, à 5 %. L'hypothèse nulle est acceptée au-dessus du seuil, rejetée en-dessous. L'hypothèse nulle dans notre exemple ayant moins d'une chance sur 10 millions d'être vraie, ($p < 10^{-7}$, *< 0,05)* est rejetée. Les deux groupes n'étant pas similaires, et on en déduit qu'ils sont *probablement* différents. Il y a là une petite contorsion de logique, car toute la base du calcul repose sur l'axiome de l'égalité des deux groupes, et ça n'est que sur la base de cet axiome que la méthode de calcul est appropriée : or, la base de validité du calcul n'est pas respectée… Nous avons simplement montré *que les deux groupes n'étaient probablement pas similaires*, et en *déduisons qu'ils sont probablement différents.*
Les statistiques n'établissent pas de vérité : elles cherchent à circonscrire une incertitude, et ceci doit toujours rester présent à l'esprit lors de l'interprétation des résultats. Cependant, lorsque la valeur de p est aussi faible, on peut très raisonnablement penser que les deux groupes sont différents. Lorsque la valeur de p est proche de 0,05 et que changer quelques patients de groupe la fait passer au-dessus ou au-dessous de 0,05, la discussion reste ouverte !

Attention :

- Le calcul simple du chi-2 décrit ci-dessus n'est valable que si le nombre d'événements attendus est supérieur à 5 dans toutes les cellules de la table. Dans le cas contraire, un calcul exact faisant appel aux distributions géométriques doit être réalisé : c'est le *test de Fisher exact*, qui peut toujours être employé dans les cas douteux pour donner au calcul un maximum de rigueur.
- Plusieurs auteurs ont voulu rendre le calcul plus sévère (donc, à rapprocher la somme du chi-2 de 0), en introduisant des facteurs de correction dans la formule. *C'est le cas du chi-2 de Yates, qui soustrait ½ à chaque (A-O) avant de l'élever au carré.*
- Le test de Mantel-Haenszel est la variante du chi-2 pour les données stratifiées : nous l'aurions appliqué dans notre exemple si les diagnostics avaient été donnés par sexe et par tranche d'âge (20-40 ans, 40-60, 60-80, > 80). Les chi-2 sont alors calculés pour chaque strate séparément, puis additionnées sur l'ensemble des strates.
- On comprend aisément que la valeur du chi-2, et donc la significativité du test, dépend de la taille de la population analysée : refaites le calcul en divisant les nombres de chaque cellule par 10 pour atteindre une taille d'échantillon totale de 105 (fréquente dans les études cliniques) (proportions conservées), et vérifiez la valeur de p !

2. Généralisation du test de chi-2 à une table à n lignes et m colonnes :

La somme des écarts entre O et A peut toujours être calculée. Simplement, le nombre de degrés de liberté change, et devient égal à (n-1)(m-1) : le nombre A est toujours fixé sans liberté pour la dernière cellule de chaque ligne, et la dernière cellule de chaque colonne. La probabilité d'égalité entre la somme du chi-2 et 0 doit alors se lire sur la ligne de distribution du chi-2 au nombre de degrés de liberté correspondant (les logiciels fournissent la valeur exactement calculée de p).

Un chi-2 significativement différent de 0 dans une table n*m n'indiquera pas où se trouve la différence : elle peut se répartir de façon homogène entre les différentes cellules, ou, de façon beaucoup plus fréquente, entre quelques, voire deux, cellules de la table. Le risque d'atteindre un nombre d'événements attendus inférieur à 5 dans une table à multiples cellules augmente avec le nombre de cellules, ce qui rend la validité du calcul incertaine : mieux vaut alors prévoir un autre plan d'analyse.

III. Variables quantitatives

1. Moyenne, variance, écart-type, médiane, percentiles, mode.

Nous voulons maintenant décrire les âges des patients dans chaque groupe, et les comparer. L'âge est une donnée *quantitative continue*. Sa description peut être faite sous la forme de ***moyenne*** (somme de toutes les valeurs divisée par le nombre de sujets), et de ***variance*** (somme de l'ensemble des carrés des écarts entre la moyenne de l'échantillon et l'âge de chaque sujet, rapportée au nombre de sujets : comme pour le chi-2, les écarts sont élevés au carré de telle sorte à ne pas s'annuler).

La variance peut donc s'écrire : $V = \Sigma \frac{(\mu - a)^2}{n}$,

où μ désigne la moyenne des âges pour l'ensemble du groupe, a l'âge de chaque sujet du groupe, et n le nombre total de personnes dans le groupe.

Cette formule exprime la **variance idéale**, d'une grande population. Nous travaillons le plus souvent en médecine sur des échantillons de patients bien plus réduits, même dans les études multicentriques.

Un échantillon ne peut fournir qu'une *estimation, qu'une valeur approchée,* de la moyenne et de la variance de la population totale. Par sécurité, il vaudra mieux sur un échantillon définir cette variance de façon un peu plus large (elle aura ainsi plus de chances de recouvrir la variance de la population totale). Pour ce faire, on corrige le dénominateur en remplaçant n par n-1: la valeur numérique de la variance augmente légèrement, *et sa formule pour un échantillon* devient : $V = \frac{\Sigma (\mu - a)^2}{n - 1}$

Cette correction est d'autant plus importante que n, la taille d'échantillon, est petit, et d'autant plus insignifiante que n est grand : la variance d'une variable sur un échantillon de 1000 personnes a plus de chances d'être proche de celle de la population totale, que la variance de la même variable sur un échantillon de 15 personnes.

De même, on souhaite approcher la moyenne de la population globale à partir de celle de l'échantillon, en sachant cependant que si l'échantillon avait été sélectionné différemment, sa moyenne serait sans doute un peu différente : il se peut que la moyenne d'hémoglobine glyquée d'un groupe de 40 patients diabétiques de type 2 soit *exactement la même* que celle d'un autre groupe de 40 patients. Il est cependant probable qu'elle n'en soit pas trop éloignée. On définit la notion d'*intervalle de confiance à 95 %* pour exprimer le fait que la moyenne se trouverait comprise dans cet intervalle chez 95 % des 100 échantillons potentiels de taille égale que l'on pourrait tirer au sort au sein de la population globale. Autrement dit, la moyenne du taux d'hémoglobine glyquée dans la population globale de diabétiques de type 2 - que l'on cherche à approcher- a sans doute 95 % de chances d'être comprise dans l'intervalle ainsi défini à partir de l'échantillon. On peut définir, de la même façon, un intervalle de confiance à 90 ou 99 ou 99,9 %... en fonction de la précision que l'on souhaite obtenir. L'intervalle de confiance à 95 % est celui le plus souvent retenu en médecine.

La variance est très souvent exprimée sous la forme de sa racine carrée, appelée ***écart-type, ou déviation standard.***

Ecart-type = $\sqrt{v} = \sigma$

L'intérêt de l'écart-type est qu'il permet d'apprécier assez rapidement la distribution de la variable : pour une taille d'échantillon supérieure à 30, 95 % de l'échantillon sera compris entre la moyenne $\pm$ 1,96σ. La valeur correspondant à 1,96 augmente lorsque la taille d'échantillon diminue : là encore, cette augmentation traduit le fait que la précision diminue avec la taille d'échantillon, et donc que la dispersion probable des valeurs, augmente.

Attention :

- Cette façon de décrire une variable quantitative n'est valable que lorsque la distribution des valeurs suit une courbe gaussienne (distribution normale).
- La moyenne n'a de sens que si
  - les valeurs se distribuent de façon symétrique autour d'elle, et si
  - elle correspond à la valeur la plus fréquemment rencontrée (dans notre exemple, la tranche d'âge la plus importante). En effet, d'établir une moyenne d'âge à 40 ans pour une série comportant 20 enfants de 10 ans et 20 adultes de 70 ans ne permettrait pas d'appréhender correctement la réalité du groupe de patients : il n'y a aucun adulte de 40 ans dans ce groupe composé de deux sous-groupes très différents, et calculer une moyenne arithmétique à 40 ans amènerait à une description fausse de la réalité.
- Pour que la moyenne et son écart-type décrivent correctement le groupe considéré, il faut donc que la distribution de la variable examinée soit *symétrique et unimodale,*

autrement dit, que la distribution des valeurs ne reflète pas l'existence de deux groupes différents de patients, d'un taux biologique, ou d'une valeur numérique quelconque. La maladie de Hodgkin a deux pics d'incidence, autour de 20-25 ans, puis autour de 60 à 65 ans. De dire que la moyenne d'âge des patients est de 40 ans et traiter les patients âgés de 75 ans de la même façon que ceux âgés de 20 ans sous prétexte d'une tolérance moyenne égale à celle des 40 ans serait un non-sens médical.

- Il est toujours utile de réaliser un graphe de la distribution des données qui permette d'en apprécier la forme (symétrique, unimodale). La plupart des logiciels statistiques permettent également de réaliser un test de normalité de la distribution, qui pourra guider l'analyse statistique ultérieure.

Lorsque la distribution de la variable quantitative n'est pas *normale (gaussienne)*, il faut privilégier d'autres modes de description de données et d'autres modes d'analyse. Il arrive assez souvent en médecine, que des variables quantitatives aient une distribution gaussienne (*statistiquement normale = gaussienne, à différencier de biologiquement normale = dans les normes biologiques)* chez le sujet sain (exemple : le taux d'hémoglobine, des leucocytes, des plaquettes...). Cette normalité statistique disparaît très souvent chez le sujet malade : les leucocytes peuvent varier de 10 000/mm^3 à plus de 100 000 dans une leucémie myéloïde chronique ou une leucémie aiguë, les plaquettes de 400 000 à plus de 1 500 000 dans une thrombocytémie essentielle ou réactionnelle, et l'on ne peut pas en général extrapoler la distribution *normale* de la variable chez le sujet *sain* à la distribution dissymétrique, parfois logarithmique, parfois difficile à décrire, de la variable chez le sujet *malade.*
Si la distribution n'est pas statistiquement normale à l'appréciation visuelle ou sur le test de normalité, recourir à la *médiane et aux percentiles* donnera une idée plus précise de la population considérée : la médiane définit le seuil en valeur absolue, en-dessous duquel se trouve 50% de l'échantillon et au-dessus duquel, se trouvent les 50% restant. Une leucocytose médiane à 20 000/mm^3 signifie que 50% des patients ont un taux de leucocytes inférieur à 20 000, et 50% d'entre eux, supérieur. Les 5ème, 10ème, ou 75ème percentiles correspondent au taux de leucocytes *en-dessous desquels* se trouvent 5%, 10%, ou 75% des patients.

Le *mode*, enfin, est le troisième descriptif d'une variable quantitative : il correspond à la valeur la plus souvent rencontrée dans l'échantillon. Il est relativement peu utilisé.
Dans une distribution *statistiquement normale,* moyenne, médiane et mode sont confondus. Dans une distribution non normale, ils sont habituellement distincts. Une distribution à plusieurs pics de hauteur égale peut comprendre plusieurs modes, mais ne comprendra qu'une seule médiane... et qu'une seule moyenne, non représentative.

Application :
Dans notre exemple, l'étude de la distribution des âges donne les résultats suivants :

	Moyenne	Ecart-type	Médiane	Extrêmes
Femmes	62,8	19,8	67	18,7-102
Hommes	55,5	17,5	55,5	18,5-96,5

A première vue, les valeurs données peuvent être compatibles avec une distribution normale : retrancher ou additionner deux écarts-types à la moyenne ne conduit pas à des âges aberrants (par exemple : négatifs), mais à des âges relativement voisins des extrêmes.

A deuxième vue, si médiane et moyenne sont confondues chez les hommes, elles s'écartent de plus de 4 ans chez les femmes, ce qui fait suspecter compte tenu de la taille d'échantillon, une distribution non normale. La question est de savoir si cet écart est significatif, ou non.

2. Comparaison de moyennes, ou analyse de variance :
Les hommes et les femmes de notre étude ont-ils le même âge ?
Le principe reste le même que pour le test du chi-2 : l'hypothèse nulle, simplement, devient :

Les deux moyennes sont semblables, ou $\mu_f - \mu_h = 0$ (μ_f représentant la moyenne d'âge des femmes, et μ_h la moyenne d'âge des hommes).
Cette simple soustraction cependant ne suffit pas à assurer une comparaison correcte : on serait prêt à reconnaître que les deux moyennes sont différentes si l'écart-type était très petit :

Par exemple :

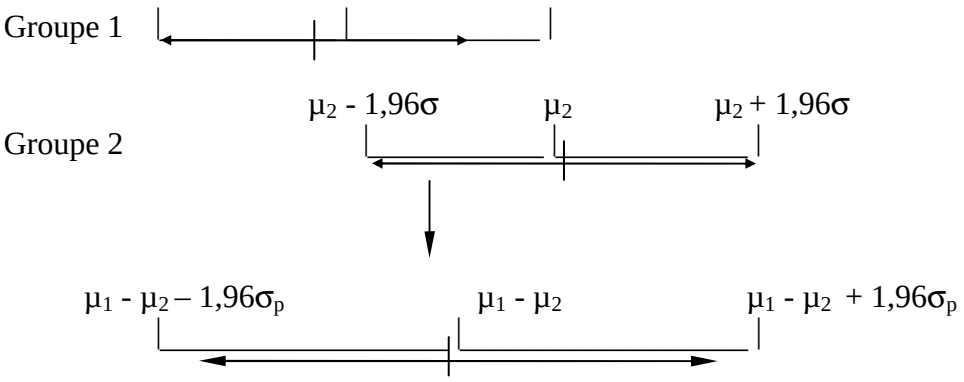
- $62,8 \pm 0,2$ conduirait à un échantillon comprenant 95% des patientes entre 62,4 et 63,2 années, et $55,5 \pm 0,2$ à un échantillon comprenant 95% des patients entre 55,1 et 55,9 ans (moyenne $\pm$ 1,96 σ). Les distributions des âges de ces deux échantillons ne se recouvrent pas, et l'on peut donc admettre que quoique proches en moyenne, les deux échantillons soient différents.
- Dans notre étude, l'écart-type est beaucoup plus important, et conduit à une distribution des âges pour 95% des patients entre 24 et 101,6 années pour les femmes, et 21,2 et 99,8 années pour les hommes : le recouvrement de ces deux distributions est considérable, et quoique les moyennes soient les mêmes que pour l'exemple précédent, l'on ne serait pas enclin spontanément à reconnaître comme différentes ces deux distributions.

La prise en compte de la variance est donc indispensable : le test de comparaison des moyennes est en fait, une *analyse de variance.* L'analyse de variance prend en compte d'une part, la différence entre les moyennes, et d'autre part, la variance combinée des deux distributions comparées (= variance poolée).
Cette variance combinée des deux distributions comparées, permet d'estimer la variance de la distribution de la différence de moyenne : la variance poolée traduit la variance de l'ensemble des échantillons comparés, et la variance de la différence des moyennes, celle de l'écart moyen entre les deux populations comparées.

La variance de la différence des moyennes permet ensuite de calculer un intervalle de confiance autour de la différence de moyennes. Si cet intervalle contient 0, la différence est alors considérée comme similaire à 0, et les deux moyennes semblables. Si l'intervalle de confiance ne contient pas 0, la différence des moyennes est considérée comme éloignée de 0, et les moyennes comme différentes. Selon l'exigence, l'intervalle de confiance peut être calculé à 95% (correspondant à un p de 0,05), à 99% (correspondant à un p de 0,01), à 99,9%... la *significativité de la différence des moyennes peut être donnée pour n'importe quelle valeur de p. p* donne la probabilité pour la différence des moyennes d'être égale à 0. On admet, comme plus haut, que les moyennes peuvent être considérées comme différentes si cette probabilité est inférieure à 5%, *si $p < 0,05$.*

En résumé, une représentation graphique de nos groupes pourrait être :

$\mu_1 - 1,96\sigma$ μ_1 $\mu_1 + 1,96\sigma$



où σ_p représente l'erreur-type de la différence des moyennes, calculée à partir de la variance *poolée* des deux échantillons initiaux, groupe 1 et groupe 2.

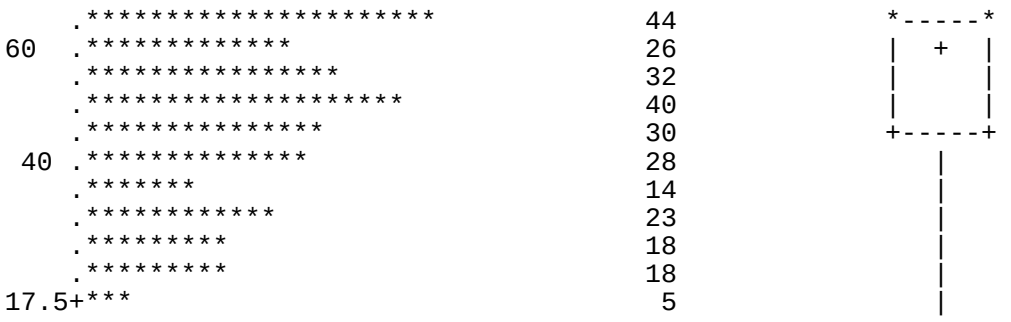
Si l'analyse de variance porte sur une population complète (rarissime en médecine), le test utilisé est le *test Z*, qui prend en compte la variance (avec n en dénominateur). Si l'analyse porte sur des échantillons de patients (cas habituel), le test à utiliser est le *test t*, qui prend en compte la formule de la variance corrigée (avec n-1 en dénominateur). On peut ensuite se référer aux tables du test t, à la ligne correspondant au nombre de patients totaux -2 (nombre de degrés de liberté du test t), et trouver la probabilité pour que la valeur de t soit différente de 0.
Bien sûr, n-1 sera très proche de n lorsque n, la taille d'échantillon, est grande. Dans ces situations, le test Z (prenant n au dénominateur), et le test t (prenant n-1 au dénominateur) donneront des résultats similaires.

Dans notre exemple, la valeur du test t est de 6,7. Nous ne serions pas surpris que sa probabilité d'être proche de 0 soit très faible, compte tenu de la taille d'échantillon : elle est, en effet, de 0,0001. Autrement dit, la probabilité pour que la différence d'âge entre le groupe des hommes et le groupe des femmes soit égale à 0 est de 1/10 000, ou *la différence des moyennes est significativement différente de 0 avec p = 0,0001. L'hypothèse nulle est refusée, et l'on accepte qu'il existe une différence d'âge significative entre les hommes et les femmes :* $\mu_f - \mu_h$ est significativement différente de 0.

3. Test non paramétrique : test de rang de Wilcoxon :

En regardant grossièrement la moyenne, l'écart-type, et les extrêmes des âges des femmes et des hommes, nous avons vu que ces valeurs étaient *compatibles* avec une *distribution normale.* Cet examen sommaire n'est cependant pas suffisant. Regardons les données de plus près (

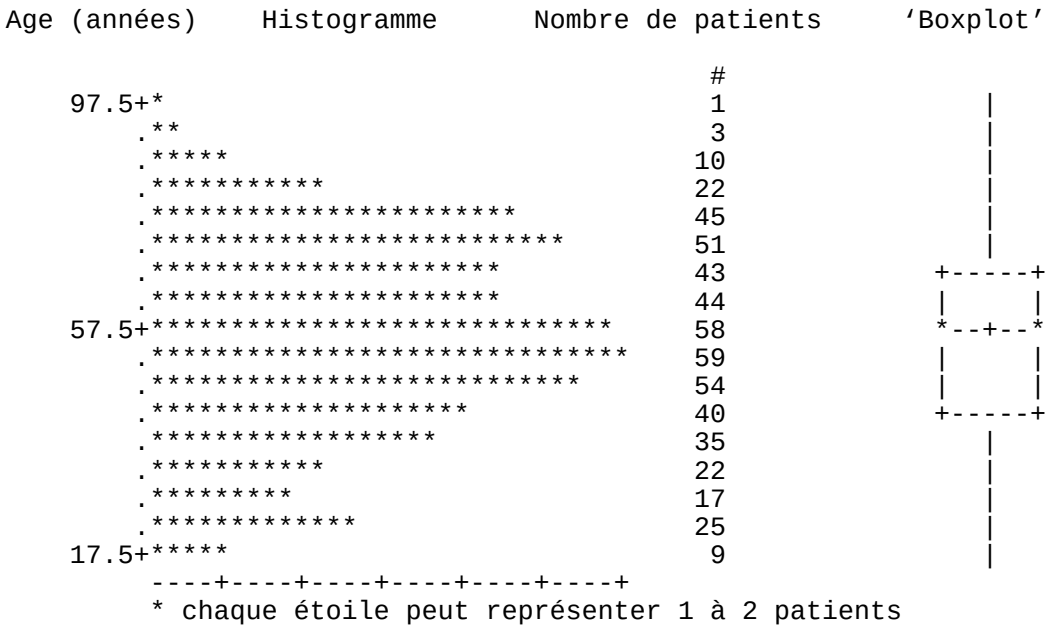
```
Fig X : Pyramide des âges en années: groupe des femmes.
      Age (années)    Histogramme      Nombre de patientes     'Boxplot'
                                                #
        102.5+*                                 1                |
             .***                               6                |
             .*******                          14                |
             .******************              35                |
             .****************************    57                |
             .******************************  62              +-----+
             .**************************      56              |     |
```



L'allure de la pyramide des âges ne paraît pas très gaussienne : il existe trois pics, le premier correspondant à 23 patientes sous la barre des 40 ans, le second à 40 patientes sous la barre des 60 ans, le troisième à 62 patientes au-dessus. Il est possible qu'existent dans cette population 3 groupes de patientes différentes, et il peut être intéressant d'examiner ces trois groupes séparément dans une étude exploratoire. Le schéma dit en 'boxplot' confirme cette impression : la moyenne (+) est différente de la médiane (barre transversale au milieu de la boîte), et la médiane n'est pas située à équidistance du 25ème et 75ème percentile (barres transversales inférieure et supérieure de la 'boîte'.

Le test de normalité effectué par notre programme statistique confirme cette impression : l'hypothèse nulle (celle de la normalité) est rejetée avec $p \leq 0{,}0001$ (la probabilité pour que cette distribution soit normale est inférieure à 1 chance sur 10 000).

Fig. 2 : Pyramide des âges chez les hommes



L'allure de la pyramide des âges des hommes est un peu différente de celle des femmes, mais il existe aussi différents pics. L'hypothèse nulle de normalité est également refusée avec p = 0,0001.

L'analyse de variance repose sur l'hypothèse de la normalité des distributions et de l'égalité des variances entre les groupes. Il existe des tests statistiques ne requérant pas la normalité des distributions ou l'égalité des variances, et permettant de comparer les groupes ainsi inhomogènes. Cela suppose de considérer les variables autrement et de prendre en compte, à la place de la valeur numérique, absolue, de la variable quantitative, *le rang qu'elle occupe dans la distribution.*
Dans notre exemple, les patients seraient ainsi classés dans chaque groupe du plus jeune au plus âgé, et le test consiste à évaluer la tendance générale des âges dans un groupe par rapport à l'autre : la question 'la moyenne d'âge compte tenu de la variance est-elle différente dans les deux groupes ?' devient 'un groupe est-il globalement plus jeune, ou plus âgé, que l'autre ?'

Application :

âge	18	19	20	21	.....	98	99	100	101	102
Femmes : n	2	3	4	2		2	1	3	1	1

âge	18	19	20	21	.....	95	96	97	98	99	100	101	102
Hommes : n	1	3	2	4		3	2	1	1	0	0	0	0

L'âge des femmes est ensuite comparé à l'âge des hommes par la constitution de 3 catégories de paires comprenant chaque fois une seule femme et un seul homme :

1- Les paires pour lesquelles les femmes sont plus jeunes que les hommes
2- Les paires pour lesquelles les femmes et les hommes sont d'âge égal
3- Les paires pour lesquelles les femmes sont plus âgées que les hommes

Dans notre exemple, les deux patientes âgées de 18 ans sont plus jeunes que les 537 hommes sur 538 âgés de plus de 18 ans : il y a donc 2 x 537 paires (= 1074 paires) pour lesquelles une patiente du groupe 1 est plus jeune qu'un patient du groupe 2.
Les 3 patientes de 19 ans sont plus jeunes que les (538 – 4) hommes âgés de plus de 19 ans. Il y a donc 3 x 534 (= 1602) paires supplémentaires pour lesquelles une patiente est plus jeune qu'un patient.
On continue ainsi à compter toutes les paires pour lesquelles une patiente est plus jeune qu'un patient, et on les additionne (nombre total : x).
Les deux patientes de 18 ans ont le même âge que le patient de 18 ans : cela fait donc deux paires d'âge égal. On dénombre ainsi toutes les paires d'âge égal (nombre total : y).
La patiente âgée de 102 est plus âgée que tous les 538 hommes, de même que la patiente âgée de 101 ans, celle âgée de 100 ans et celle âgée de 99 ans : cela donne 4 x 538 paires pour lequelles les patientes sont plus âgées que les patients, et l'on comptabilise toutes les paires 'plus âgées' pour l'ensemble de la série.
Le nombre de paires égales est ensuite partagé à parts égales dans le groupe « paires plus jeunes » et le groupe « paires plus âgées », qui comprennent alors x + y/2 paires pour le premier, et z + y/2 paires pour le second (nombre total : z). Nos deux groupes de patientes et patients sont ainsi transformés en deux groupes de paires, 'plus âgée' et 'moins âgée', que l'on va comparer.

L'hypothèse nulle devient : le nombre de paires est égal dans les groupes 'Plus âgée' et 'Moins âgée', ou (x + y/2) - (z + y/2) = 0, et la comparaison revient à une comparaison de proportions. La nombre de paires 'plus jeunes' rapporté au nombre total de paires est-il similaire, ou différent, du nombre de paires 'plus âgées' rapporté au nombre *total* de paires ? Comparaison de proportions renvoie au type d'analyse effectué par le test de Chi-2 décrit plus haut.

Ce *test de comparaison par rangs* dit de *Wilcoxon ignore* donc la distance existant entre les âges, et ne donne pas de poids supplémentaire aux âges extrêmes, contrairement à l'analyse de variance. Il n'est pas dépendant du caractère normal ou non de la distribution. Il permet l'analyse de petits échantillons : une moyenne calculée sur 10 patients a des chances de ne pas être très précise, la variance importante, et la comparaison avec un autre groupe de 10 patients peu puissante. Le test de comparaison par rang portera sur 10 x 10 = 100 paires, plutôt que sur 20 individus : dans le cas de petits échantillons, le test par rang de Wilcoxon sera probablement plus performant, mais également plus rigoureux qu'une comparaison de moyennes car il est rare qu'une distribution soit normale dans ces conditions, et l'analyse de variance pourrait conduire à une estimation fausse de p.

Le test de *U-Mann-Whitney* est basé sur le même principe et arrive, par une technique de calcul différente, aux mêmes résultats que le test de Wilcoxon. Ils sont interchangeables et les logiciels de statistiques fournissent indifféremment l'une ou l'autre procédure.

Attention :

- Comme tout test, le test de Wilcoxon rencontre des limites. Chaque groupe doit comprendre au moins 10 patients, sinon le test perd de sa précision et de sa valeur… mais vouloir comparer de trop petits groupes nous fait quitter de fait le champ de l'analyse statistique !
- En pratique, l'analyse de variance donnera les mêmes résultats que le test de Wilcoxon en terme de p lorsque la taille de chaque groupe dépasse 30, même si les distributions ne sont pas gaussiennes. Elle aurait été possible dans l'exemple ci-dessus, mais seule l'observation de la distribution des variables permet de se rendre compte qu'il y a peut-être trois groupes d'âge de patients.

4. Variables ordinales :

On utilise souvent en médecine des variables semi-quantitatives d'un type particulier : le cancer du colon est classé en stade A, B, ou C de Duke en fonction du degré d'envahissement de la muqueuse, et la gravité va croissante du stade A vers B puis vers C, mais C n'est pas *trois fois* plus grave que A, ou *deux fois* plus grave que B. Il en est de même des stades I à IV de la dyspnée ou de l'artériopathie des membres inférieurs. Ces stades ne peuvent pas s'additionner, se multiplier, ou se diviser : simplement, ils traduisent une hiérarchie dans la gravité de la maladie plus qu'une quantité strictement mesurable.
Ces variables semi-quantitatives sont dites *ordinales*.
On peut souhaiter comparer, cependant, des données de ce type en conservant leur caractère hiérarchisé, ou *ordonné* : l'administration d'un diurétique chez ce groupe de patients atteints d'insuffisance cardiaque améliore-t-il le stade de la dyspnée, autrement dit, la fait-il passer d'un ordre supérieur (III ou IV par exemple), à un ordre inférieur (II) ?

La comparaison du nombre de patients atteints d'une dyspnée de stade I, II, III, ou IV dans un groupe traité et non traité pourra faire appel au test de Wilcoxon : chacun des stades peut être considéré comme un rang, et le test de Wilcoxon (ou le test de U-Man-Whitney) pourra permettre de dégager la tendance vers le stade moins - ou plus - important de la dyspnée.

5. Généralisation de l'analyse de variance à plusieurs groupes :

Il peut être intéressant de comparer les moyennes et variances d'une variable quantitative entre plusieurs groupes. Dans notre exemple, on peut vouloir comparer l'importance de l'amaigrissement entre les groupes 'origine psychogène', 'origine cancéreuse', 'origine nutritionnelle ou endocrinienne', et 'origine autre'.

Il existe plusieurs façons de considérer le problème :
La première consiste à faire une seule analyse, et à examiner s'il existe, au moyen d'un seul test, une différence *quelque part* au sein de l'échantillon global des 4 sous-groupes. Le principe de base en est le même que pour l'analyse de variance entre deux groupes, et le test employé est le *F-test* (Généralisation du t-test à plusieurs groupes). Si le F-test ne met pas en évidence de différence significative au sein des x échantillons, on peut en conclure qu'il n'existe vraisemblablement pas de différence entre le groupe à plus faible, et le groupe à plus forte, moyenne. Par conséquent, il n'existe sans doute pas de différence entre les groupes dont les moyennes sont comprises entre la plus petite, et la plus grande, moyenne, et notre analyse pourra s'arrêter là.

Si par contre, le F-test objective une différence significative, c'est qu'il existe sans doute une différence entre le groupe à la plus petite, et le groupe à la plus forte, moyenne... sous réserve que ces deux groupes aient une taille d'échantillon permettant d'arriver à une différence significative. Il est possible aussi que la différence, en fait, siège entre deux groupes intermédiaires à forte taille d'échantillon, voire entre plusieurs groupes, voire de façon diffuse entre chaque groupe de patients. Comme un test de chi-2 à multiples cellules, le F-test permet de détecter une différence *quelque part, mais ne permet pas de la localiser.* Il peut pourtant être important de savoir si devant un amaigrissement conséquent, il vaille mieux s'orienter en premier lieu vers une étiologie ou une autre...

On peut envisager, pour répondre à cette question, de comparer les groupes deux à deux au moyen d'un test t. Ainsi, de comparer le groupe 'origine psychogène' au groupe 'origine cancéreuse', puis au groupe 'origine endocrinienne', puis au groupe 'origine somatique autre'... et de continuer par les comparaisons 'cancer'-'endocrinien', 'cancer'-'autre', puis 'endocrinien'-'autre' et d'épuiser toutes les combinaisons logiques possibles. Pour quatre groupes, il existe ainsi six possibilités logiques. Pour 5 groupes, dix. Pour 6, 15. La comparaison systématique de tous les groupes deux à deux pourrait paraître plus rigoureuse, ou plus porteuse d'information intéressante, que l'analyse globale du F-test.
Le problème cependant est qu'elle multiplie les chances de montrer une différence *significative par hasard :* nous acceptons une différence comme significative si elle a moins de 5 chances sur 100, d'être liée au hasard. De ce fait, si 100 comparaisons au hasard sont réalisées, il est très probable que 5 au moins d'entre elles, s'avèrent significatives... par hasard, puisqu'il s'agit de la limitation même du test statistique. Sur 20 comparaisons, une au moins peut être statistiquement significative par hasard, et cette *significativité* serait dépourvue de toute *signification* clinique ou biologique.

Il est donc nécessaire, pour éviter de tomber dans le piège du hasard et de tirer des conclusions fausses de données correctement recueillies, mais mal analysées, d'établir un garde-fou. Une première solution serait de diviser le p exigible par le nombre de comparaisons effectuées : si 6 comparaisons sont réalisées, le risque d'obtenir une différence à p = 0,05 est de 5% x 6, soit de 30 %. Pour ramener ce risque à 5%, le plus simple est de diviser le p exigible par 6, ce qui ramène à 0,05/6 = 0,0083 (6 comparaisons effectuées, 4 sous-groupes comparés deux à deux), le p exigible pour avoir moins de 5% de chances de se tromper en affirmant une différence significative. Cet ajustement s'appelle *ajustement de Bonferroni.*
D'aucuns prétendront cependant que cet ajustement est trop sévère, et ne permettra pas de reconnaître, dans cette démarche exploratoire sans hypothèse a priori sur la localisation de la différence, une différence existant dans la réalité à niveau de p = 0,05 entre deux sous-groupes donnés. D'autres types d'ajustement de p, moins sévères, ont été proposés : l'ajustement selon Tukey (qui exige des groupes de taille identique, ce qui est rare en médecine), selon Scheffé, que l'on peut appliquer même si les groupes sont de taille différentes, selon Neuman-Peul, qui consiste à diminuer le nombre de comparaisons effectuées (on commence par comparer les deux groupes extrêmes : s'il n'existe pas de différence, les calculs s'arrêtent là et une seule comparaison aura été faite. S'il existe une différence, on compare ensuite un des groupes extrêmes (à moyenne la plus basse par exemple) avec le deuxième groupe extrême opposé (à moyenne immédiatement inférieure au groupe ayant la moyenne la plus élevée). S'il n'existe pas de différence, les tests s'arrêtent là et on n'aura effectué que deux comparaisons : on déclare qu'a priori il ne doit pas y avoir d'autres différences significatives entre les sous-groupes. S'il existe une différence, on poursuit les comparaisons entre le groupe à moyenne inférieure et le groupe à moyenne directement inférieure au dernier groupe testé, et ainsi de suite). Ce type de procédure permet de ne pas comparer de façon systématique, tous les sous-groupes formés et d'arrêter les comparaisons de moyennes dès lors que la dernière différence testée n'est plus significative : moins de tests sont effectués, et l'ajustement de p pourra être moins sévère.
Il existe toujours un certain *trade-off*, un certain équilibre, entre rigueur du test et obtention de résultats *statistiquement significatifs :* les chances de déceler une différence significative sont moindres avec un ajustement de type Bonferroni qu'avec un ajustement de type Neuman-Peuls, mais une différence significative en Bonferroni aura plus de chances d'être réellement significative qu'une différence observée d'après Neuman-Peuls...

En tout état de cause, les comparaisons de moyennes multiples ne doivent se faire qu'après ajustement de p, et l'interprétation de ces différences doit tenir compte de la rigueur de l'ajustement : le $p < 0,05$ n'établit pas à lui seul une vérité médicale ou biologique, mais représente une donnée d'analyse devant faciliter la lecture des résultats.

6. Généralisation du test de Wilcoxon à plusieurs sous-groupes :

La comparaison d'une variable quantitative à distribution non normale, ou d'une variable ordinale peut être intéressante entre plusieurs sous-groupes : on peut vouloir comparer l'action d'un diurétique (groupe 1), d'un inhibiteur de l'enzyme de conversion (groupe 2) et d'un beta-bloqueur (groupe 3) dans le traitement de l'insuffisance cardiaque, et évaluer combien de patients dans chacun des trois groupes passent dans le ou les stades de dyspnée NYHA (cotée de I à IV) inférieure.
Une comparaison de moyenne ne sera pas possible, car on ne peut pas calculer de *moyenne* de stade de dyspnée. Un test de Wilcoxon cependant pourrait être réalisé pour des comparaisons deux à deux.

Le test de Kruskall-Wallis représente la généralisation du test de Wilcoxon pour comparaison de variables ordinales ou quantitatives non gaussiennes entre multiples sous-groupes. Il est par ailleurs équivalent au test de Wilcoxon pour la comparaison entre deux groupes.

7. Comparaison entre deux variables quantitatives :

Il ne s'agit plus maintenant de comparer deux moyennes, ou la variance d'une variable quantitative entre deux groupes (l'importance de l'amaigrissement chez les patients à étiologie somatique ou psychogène), mais d'examiner s'il existe une relation entre deux variables quantitatives : le taux de créatinine est-il fonction de la masse corporelle ? Les taux sériques d'un médicament potentiellement néphro- ou myélotoxique sont-ils fonction de la clearance de la créatinine ? Le taux de beta2-microglobuline est-il fonction de la masse tumorale du lymphome ? Existe-t-il une relation entre l'importance de la virémie HIV et le nombre de lymphocytes CD4 circulants ? Le taux d'hémoglobine glycosylée est-il un bon reflet des moyennes glycémiques ?

Ces questions relèvent toutes de la notion de *corrélation,* et la relation peut tout d'abord s'appréhender sur un graphe : le nombre de copies virales est porté sur l'axe des x, le nombre de lymphocytes CD4 circulants sur l'axe des y, et la forme du nuage des points peut être appréciée. Si les points se trouvent tous sur une droite parfaite, il existe très certainement une corrélation nette entre les deux variables, à deux exceptions près : si les points se trouvent tous sur un nuage vertical, cela veut dire que les valeurs des y (le nombre de CD4), ne varient pas en fonction de la virémie, mais peuvent prendre toutes les valeurs pour une virémie constante donnée. *Il n'y a donc pas de corrélation, et la pente de la droite (verticale) est égale à + l'infini.* Si les points se trouvent tous sur une droite horizontale, le nombre de CD4 reste constant quelle que soit la valeur de la virémie : *il n'y a aucune corrélation, là non plus, et la pente de la droite horizontale est égale à 0.*
Pour qu'il y ait corrélation, il faut tout d'abord que la pente de la droite soit significativement différente de 0 et de l'infini : le nombre de CD4 circulants doit varier avec le nombre de copies virales, le taux d'hémoglobine glycosylée doit varier avec la moyenne des glycémies. La pente de la droite sera positive (supérieure à 0), si l'augmentation d'un taux est associée à l'augmentation de l'autre. Elle sera négative (inférieure à 0) si l'augmentation d'un taux est associée à la diminution de l'autre.
Exemples : il y a une corrélation positive entre le taux d'hémoglobine glycosylée et les taux de glycémie, mais la corrélation est négative entre le nombre de copies virales HIV et le nombre de lymphocytes CD4 circulants.

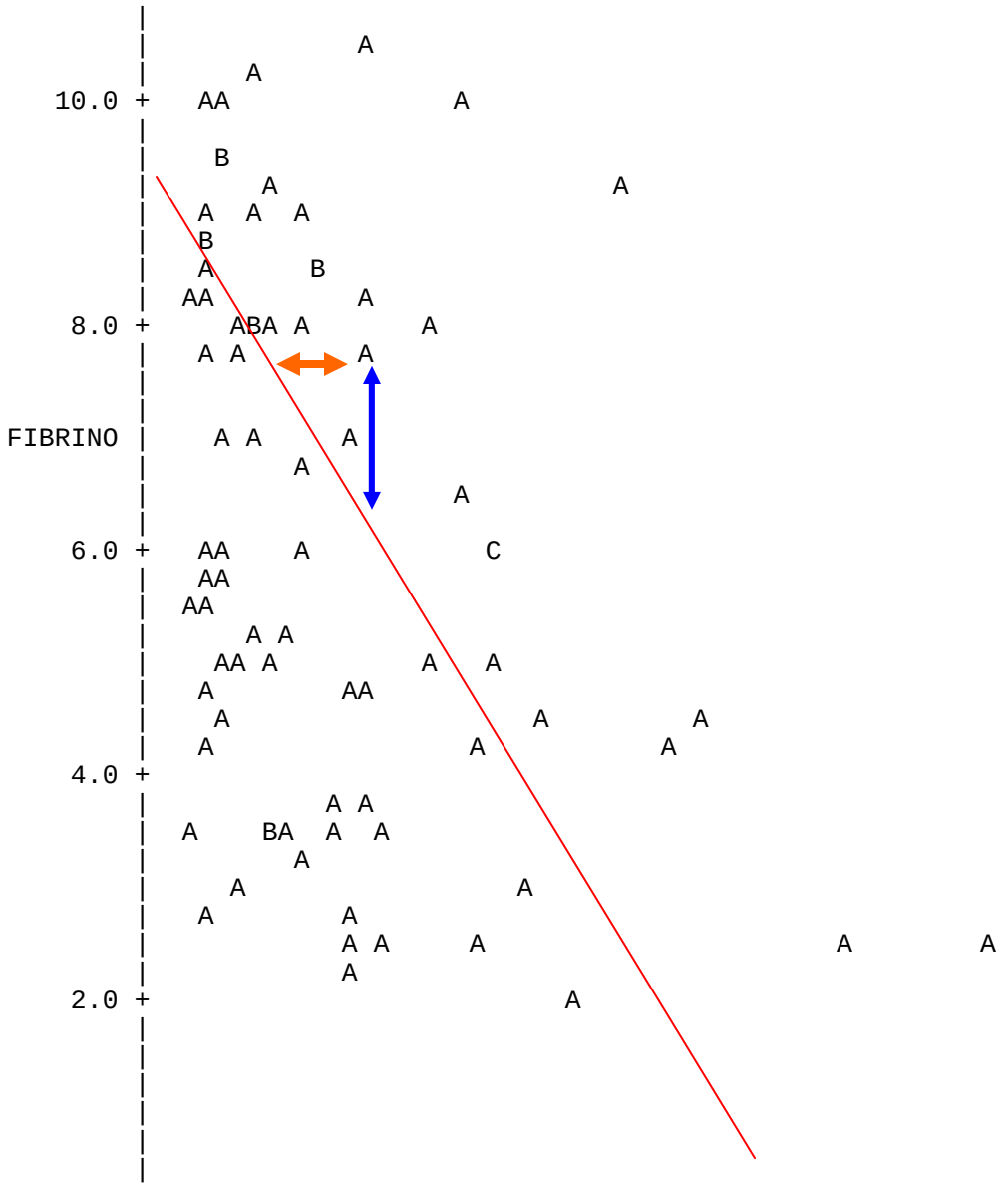
Le modèle de base de l'équation d'une corrélation est donc l'équation d'une droite :
$y = a.x + b$, où a représente la pente de la droite, et b l'intercept, ou la valeur de y lorsque x est égal à 0. Les valeurs biologiques égales à 0 sont rares en médecine, du moins pour les paramètres biologiques de base (NF, ionogramme sanguin, paramètres de la coagulation...) ou lorsque les techniques de dosages sont suffisamment sensibles pour détecter des taux faibles (une TSH rigoureusement égale à 0 est rare avec la technique de dosage ultra-sensible). L'intercept est donc souvent une valeur extrapolée par l'équation, et il n'est pas certain qu'elle corresponde à une valeur *biologiquement observée.*

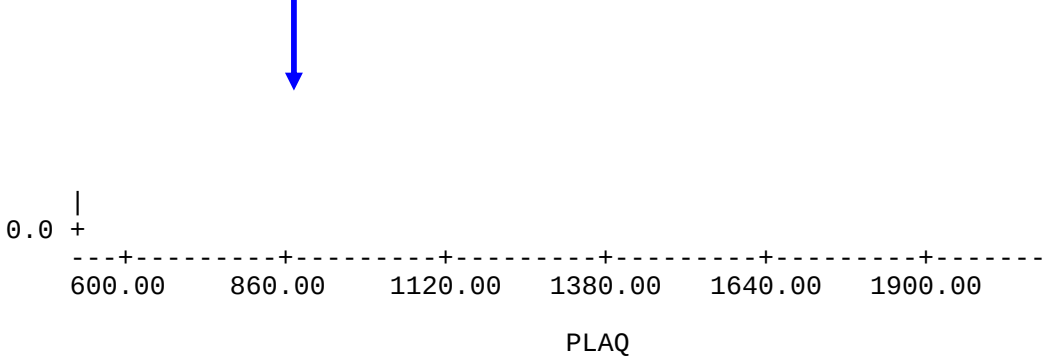
Il est rare cependant qu'en biologie ou en médecine, les points de corrélation entre deux variables puissent être reportés de façon *exacte* sur une droite. Ils forment le plus souvent, un nuage dont la droite de corrélation est la bissectrice. L'équation de la droite ne suffit pas à

décrire de façon satisfaisante le phénomène observé. Comme dans la comparaison de moyennes, où la variance décrit la dispersion des valeurs autour de la moyenne, il existe une variance de chacune des deux variables autour de la droite de régression. Pour un x donné (un nombre de copies virales), l'y correspondant (le nombre de lymphocytes CD4 circulants) peut se trouver à plus ou moins grande distance en-dessous ou au-dessus de la droite. La meilleure droite, celle qui représentera le mieux la corrélation, est celle pour laquelle la somme de l'ensemble de ces distances (la distance de chaque y par rapport à la droite) est la plus petite possible. Certaines de ces distances négatives (lorsque l'y observé est situé sous la droite), diminueraient artificiellement la somme des distances entre l'y observé et la droite (figurant l'y prédit par le modèle) : la parade à ce problème consiste à additionner non pas les distances brutes négatives et positives, mais le carré de ces distances. La meilleure droite décrivant le phénomène est celle pour laquelle la somme de ces carrés sera minimale : la technique de *fabrication* de la droite est appelée technique de la ***somme des moindres carrés (least square sum)*** (Fig. 1).

Figure 1 : CORRELATION ENTRE FIBRINOGENE ET PLAQUETTES

Légende: A = 1 observation, B = 2 observations, etc.





Dans cet exemple, nous avons réalisé un graphe de corrélation entre les valeurs de fibrinogène et celles de la numération de plaquettes dans une population d'hyperthrombocytose réactionnelle (essentiellement inflammatoire), et avons voulu tester l'hypothèse de mécanismes de régulation physiologique du risque de thrombose chez ces patients.
On se rend compte tout d'abord qu'il existe une corrélation négative (plus la thrombocytose est importante, plus le fibrinogène est bas), avec un coefficient de corrélation à – 0,34690. De plus, cette corrélation est significative puisque la valeur de p (égale à 0,0016) est largement inférieure au seuil classiquement admis de 0,05.
La flèche en bleu représente la distance entre la valeur observée de y (le point A sur notre graphe) et sa valeur attendue (calculée) sur la droite de régression linéaire : il s'agit de la distance (y – y'). La flèche en orange représente la distance entre la valeur observée de x et sa valeur calculée sur la droite de régression linéaire : il s'agit de la distance (x- x'). On comprend que si toutes les valeurs observées de y et de x se trouvaient sur la droite, alors la corrélation entre les deux valeurs serait parfaite : y serait toujours exactement prédictible en fonction de x, et vice-versa. Le coefficient de corrélation serait égal à 1 en valeur absolue. Plus les valeurs observées y et x se 'promènent' à distance de la droite, plus la corrélation est lâche : plus les distances (y-y') et (x-x') sont grandes, plus le nuage de points est épars, plus la corrélation sera faible, et plus le coefficient de corrélation s'approchera de 0.

Le ***coefficient de corrélation r (ou rho)*** prend en compte l'ensemble de ces distance y-y' et x-x' en les rapportant au nombre d'observations. Il prend également en compte, par son signe, la pente de la droite de régression : il sera négatif si la pente est descendante (plus les plaquettes sont hautes, plus le fibrinogène est bas) et positif si la pente est montante (plus les plaquettes sont hautes, plus le fibrinogène est haut). *Il traduit donc la somme de l'ensemble des écarts des points obervés, par rapport aux points calculés situés sur la droite.*

La *pente de la droite a (ou alpha)* traduit, elle l'augmentation ou la diminution de la valeur de y lorsque x varie : si alpha = 2, la valeur de y augmente deux fois plus vite que la valeur de x.

L'ensemble de la procédure (calcul de l'équation de la droite, de la variance de x, de la variance de y, du coefficient de corrélation) s'appelle une *régression linéaire.* A nouveau, une corrélation sera dite significative si la pente de la droite a suffisamment de chance de différer de 0 ou de l'infini, et si le coefficient de régression a suffisamment de chances de différer du 0. Autrement dit, si la droite de régression est suffisamment loin d'une droite horizontale ou verticale, et si les points observés sont suffisamment proches de la droite estimée. Le p de la corrélation prend en compte ces deux ingrédients, et l'on pourra voir ainsi des corrélations significatives à moins de 5 % de chances de se tromper, avec une pente faible mais un coefficient de corrélation proche de 1, ou à l'inverse avec une pente significative mais un coefficient de corrélation faible. Le plus souvent, une corrélation significative signe une tendance, mais il est rare qu'en médecine on puisse, à l'instar de ce qui se fait en physique ou en chimie, prédire connaissant x une valeur y à partir de l'équation de régression.

IV. Analyse multivariée : notion de régression logistique

L'ensemble des tests exposés jusqu'à présent constituent les outils de l'analyse univariée : analyse d'une variable en fonction d'un paramètre (groupe ou sous-groupes, autre variable dans la régression linéaire simple). Il existe de nombreuses situations en médecine dans lesquelles l'analyse univariée marque rapidement ses limites : on sait que l'hypertension artérielle, l'hypercholestérolémie, le diabète, le tabagisme, le stress, sont des facteurs de risque de maladie cardiovasculaire ; mais quel est le poids respectif de chacun de ces facteurs de risque dans la survenue d'un infarctus du myocarde ? Ces facteurs de risque sont-ils indépendants les uns des autres, ou certains d'entre eux ne font-ils que 'traduire' les autres (le tabagisme et l'hypertension artérielle par exemple ne sont-ils que des expressions, partiellement ou totalement, du stress) ? Certains de ces facteurs de risque sont-ils synergiques, ou au contraire antagonistes ? Les mêmes questions peuvent se poser pour les facteurs de risque connus du cancer du sein, du cancer du poumon, de la survenue d'une maladie thrombo-embolique, voire des maladies dont la cause objective est connue : le BK est à l'origine de la tuberculose, mais le contact avec le BK seul ne suffit pas à déclencher la maladie... avant la découverte du BK, les facteurs de risque de la tuberculose maladie multifactorielle auraient compté la malnutrition, la promiscuité, les conditions socio-économiques défavorables... on rajouterait à l'heure actuelle toutes les causes d'immuno-dépression, et plus récemment encore, la façon de réagir du système immunitaire, et notamment de l'immunité innée macrophagique, au contact du BK. La maladie multifactorielle avant la découverte du BK, devenue monofactorielle avec sa découverte, redevient de fait multifactorielle avec les progrès de l'immunologie et de la génétique…

Il est possible, pour évaluer le rôle de chacun de ces facteurs de risque, de les 'peser' de façon indépendante, de procéder à une succession d'analyses univariées en stratifiant par chacun d'entre eux : on pourrait comparer, dans une étude de cohorte, l'incidence de l'infarctus du myocarde parmi les fumeurs et les non-fumeurs ; puis, chez les fumeurs d'une part, et les non-fumeurs de l'autre, l'incidence de l'infarctus chez les hypertendus et les non-hypertendus ; puis, chez les fumeur hypertendus, les fumeurs normo-tendus, les non-fumeurs hypertendus, et les non-fumeurs normotendus, l'incidence de l'infarctus chez les diabétiques d'une part, et les non-diabétiques de l'autre ; puis, chez.... et ainsi de suite.

Cette succession d'analyses univariées permet bien sûr, de déterminer si le stress rajoute un risque supplémentaire d'infarctus dans chacune des sous-catégories, et permet de mesurer l'importance de ce risque : s'il est plus élevé dans la sous-catégorie fumeur-hypertendu-diabétique que dans la sous-catégorie fumeur-hypertendu-non-diabétique, c'est que peut-être le stress agit de façon synergique avec l'un des trois premiers facteurs de risque... mais on ne saura pas lequel.

Il va de soi que la puissance des tests diminue avec la taille d'échantillon : la stratification en sous-groupes demande une taille d'échantillon initiale très importante si les derniers sous-groupes doivent encore comprendre un nombre suffisant de patients... en pratique, cette stratégie bien que théoriquement satisfaisante, est rarement possible. Elle est également très lourde.

L'alternative est représentée par la *régression logistique* : sans entrer dans les détails mathématiques, le principe de son équation pourrait s'écrire ainsi (*attention, cette expression est mathématiquement fausse, mais sa signification globale est juste*).

Maladie = intercept + OR1. FR1 + OR2.FR2 + OR3.FR3 + OR4.FR4 + ...ORn.FRn.

Dans cette équation, la maladie s'exprime généralement de façon binaire : elle existe, ou non. Les différents facteurs de risque (FR) peuvent s'exprimer de façon binaire (1 ou 0), de façon ordinale (1, 2, 3, 4...), voire sous la forme d'une variable quantitative continue. Lorsqu'il s'agit d'une variable binaire, on peut extraire du coefficient qui lui est attribué l'odds ratio (OR), représentant le risque relatif lié au facteur de risque considéré *compte tenu du risque correspondant aux autres facteurs de risque.* Autrement dit, OR1, OR2, OR3, OR4, quantifient le risque associé à chaque facteur de risque, sachant que la maladie est aussi expliquée par les autres facteurs de risque gardés dans l'équation.

En pratique, on introduit dans le modèle de régression logistique les facteurs de risque significatifs à 0,1 en analyse univariée (pour lesquels $p < 0,1$). On peut forcer dans le modèle des facteurs de risque non significatifs en analyse univariée si cela paraît justifié sur le plan biologique ou médical. Il existe plusieurs types de procédures en régression logistique, mais le principe consiste à introduire d'abord dans le modèle le facteur de risque le plus significatif ; s'il n'explique pas à lui seul toute la maladie, on introduit dans le modèle le deuxième facteur de risque ; s'il est trouvé significatif, il reste dans le modèle. Si non, il en sort et le troisième est alors introduit et testé. L'introduction du xième facteur de risque s'arrête lorsque plus aucun facteur de risque n'est trouvé significatif, autrement dit, lorsque l'introduction d'un facteur de risque supplémentaire n'apporte plus d'explication supplémentaire à la survenue de la maladie.

Exemple : si dans les maladies cardiovasculaires, l'hypertension ou le tabagisme n'étaient qu'une expression (qu'une traduction) du stress, mais n'expliquaient pas par eux-mêmes une partie de l'incidence de l'infarctus de myocarde, ils seraient éliminés du modèle au profit de la variable stress. Si au contraire le stress ne jouait pas de rôle, et que l'hypertension ou le tabagisme agissaient comme facteurs *confondants* permettant au stress d'apparaître comme significatif en analyse univariée, le stress ne sortirait plus significatif du modèle de régression logistique qui ne garderait que l'hypertension et le tabagisme comme facteurs de risque vrais.

On peut calculer à partir des paramètres de chaque facteur de risque intégré dans l'équation de régression logistique les odds ratio pour chaque facteur de risque testé et leur intervalle de confiance. On pourra également tester l'interaction entre deux facteurs de risque. Imaginons, dans une étude des facteurs de risque du mésothéliome, que le tabagisme soit codé 1 si présent, 0 si absent. L'exposition à l'amiante sera codée de la même façon. Il est facile de créer une variable traduisant l'exposition simultanée aux deux facteurs de risque : tabac.amiante prendra la valeur 1 (1x1) lorsque les deux facteurs de risque seront présents, et 0 (1x0, 0x1, ou 0x0) lorsqu'il n'y aura pas présence simultanée du tabagisme et de l'exposition à l'amiante. On pourra écrire l'équation de régression logistique suivante :

mésothéliome = intercept + OR1.Tabac + OR2. Amiante + OR3. Tabac.amiante

Si la variable tabac.amiante est retenue dans le modèle comme significative avec un odds ratio supérieur à 1, c'est que l'association tabac-amiante rajoute un risque significatif par rapport à l'exposition au tabac d'une part, et à l'amiante d'autre part. C'est donc qu'il y *synergie* entre le tabac et l'amiante. Si l'association tabac-amiante est retenue comme significative, mais avec un odds ratio inférieur à 1, c'est que l'association est *antagoniste* : la présence conjuguée des deux facteurs de risque diminue le risque global de survenue de la maladie (les

antagonistes sont plutôt rares en médecine !). Si l'association des deux facteurs de risque n'est pas retenue dans le modèle, c'est qu'elle ne modifie pas le risque déjà exprimé par la présence du tabac d'une part, et de l'amiante d'autre part : les deux facteurs de risque ne sont ni synergiques, ni antagonistes, mais s'additionnent.

La régression logistique représente donc un outil extrêmement puissant et utile en médecine ou en biologie. Elle permet de réaliser une analyse fine en évitant l'écueil des tailles d'échantillon gigantesques nécessitées par l'analyse univariée stratifiée. Elle permet de *peser chaque facteur de risque, d'en mesurer l'indépendance par rapport aux autres, de tester les interactions, de contrôler les éléments confondants.*

V. Analyse du pronostic. Courbes de survie

Les courbes de survie représentent l'outil indispensable à l'analyse du pronostic : l'événement 'vivant/décédé' peut bien sûr constituer le facteur descriptif utilisé, mais tout autre événement traduisant le pronostic peut l'être également : survenue ou non d'une complication, survenue ou non de la guérison, survenue ou non d'une maladie intercurrente : leur caractéristique commune est qu'à chaque fois, l'événement peut être déjà survenu au moment de l'analyse des données ou peut ne pas l'être (parce qu'il n'est pas encore survenu, ou parce qu'il ne surviendra pas, et l'on parle alors de *donnée censurée*). Les courbes de survie incluent donc des patients pour lesquels on ne sait pas si l'événement étudié va survenir, ou non, un jour. Il peut sembler présomptueux d'analyser des données dont on ne saisit pas la réalité dans le futur : médecins, biologistes, statisticiens sont tous des humains, et l'avenir ne leur appartient pas... c'est vrai. Une courbe de survie ne permet donc jamais de décrire *ce qui va se passer chez un patient : elle permet tout au plus de décrire ce qui s'est passé chez les patients précédemment connus, et une probabilité de survie à un moment de l'évolution de la maladie. Il ne faut jamais se servir d'une courbe de survie pour affirmer à un patient ou à sa famille que son risque de décès, sa chance de guérison, son risque de survenue d'une complication est de x % à un an : pour un patient donné, le risque de décès est soit de 100 %, soit de 0 %. Il n'y a pas d'alternative entre le fait de vivre, ou de ne plus vivre. A l'heure actuelle, aucune science ne permet de prédire le futur, et d'utiliser un vocabulaire scientifique, voire une démarche calculée, pour essayer de l'approcher ne permet jamais d'être affirmatif au niveau d'un individu : le domaine du pronostic est sans doute celui dans lequel la médecine offre le plus d'incertitudes non circonscrites par les techniques d'analyse.*

Comment construit-on une courbe de survie ? Faiblesses et forces :

1- Le temps 0

Les seuls temps fixes dans une vie sont ceux de la naissance, habituellement datée avec précision, et de la mort, dont on peut connaître l'heure... une fois survenue. La seule courbe de survie *fournissant une information indubitable* serait donc celle décomptant le temps entre la naissance et la mort. Il faut, pour analyser une durée, partir si possible d'un temps 0 identique pour tous les patients.

Lorsque l'on s'intéresse au pronostic d'une maladie, le temps 0 est plus difficile à définir. On considère très souvent comme temps 0 le moment du diagnostic de la maladie, et l'on dit, un peu légèrement, que le pronostic du cancer du poumon est de x% de survie *à deux ans.* Les limites apparaissent de façon évidente : le temps 0, celui du diagnostic, n'est pas le même pour le patient dépisté en médecine du travail (petite tache ronde asymptomatique) que pour

celui diagnostiqué devant une altération massive de l'état général avec métastases osseuses et cérébrales. On pourrait, pour plus de précision, stratifier le temps 0 en fonction du stade du cancer. Les études de dépistage ont cependant montré que le temps 0 n'était pas le même, à taille d'image ronde pulmonaire asymptomatique sur une radiographie effectuée de façon systématique, pour une lésion diagnostiquée l'année du premier dépistage effectué dans une entreprise, année au cours de laquelle on dépiste des tumeurs ayant pu être présentes depuis 2, 3, 4... années, et l'année suivante, celle du deuxième dépistage, où l'on ne dépiste plus que les tumeurs apparues dans les douze derniers mois. A 1 centimètre de diamètre, une tumeur développée en 3 mois est sans doute plus agressive qu'une tumeur de même diamètre évoluant depuis 3 ans. Le temps 0 du diagnostic de la tumeur n'est pas le même : une tumeur agressive à trois mois peut déjà être évoluée, une tumeur indolente à trois mois peut encore être dans sa phase pré-clinique…

Le temps 0, préalable à la construction de toute courbe de survie, est donc défini par les moyens d'observation dont nous disposons. Nous retiendrons avant de construire une courbe de survie, que le soleil existe probablement avant l'heure de son lever, que l'heure de son lever change en tout point de la surface du globe, et que, si midi au sein d'un fuseau horaire sonnera au même moment, le vrai midi sera différent pour chaque individu en fonction de sa localisation dans le fuseau horaire. Une approximation de même nature, mais d'amplitude sans doute plus grande en fonction de la pathologie, préside à la construction d'une courbe de survie.

2- Notion de survie conditionnelle

L'idéal serait bien sûr, de disposer de la même période d'observation pour tous les patients inclus dans l'étude : on pourrait ainsi affirmer, en retenant cependant l'incertitude du temps 0, qu'à dix ans du diagnostic la survie de la cohorte atteint, par exemple, 50% (ce qui ne veut pas dire, encore une fois, que les chances de survie de Mr P. Dupont à 10 ans sont de 50%). Les patients cependant ne débutent pas tous leur maladie au même moment : certains seront suivis depuis 10 ans au moment de l'analyse, d'autres seulement depuis 1 an. Les premiers auront eu le temps de guérir, d'entrer en rémission, de décéder, les autres, non. On parle de *données censurées,* lorsque l'événement mesuré n'est pas encore survenu, et de *données non censurées,* lorsqu'il est survenu. Si le décès est l'événement mesuré, les données censurées correspondront aux patients vivants au moment de l'analyse (leur date de décès n'est pas connue).

La courbe de probabilité de survie cumulée basée sur le produit des probabilités conditionnelles (courbe de Kaplan-Meier) est donc bâtie de la façon suivante : tous les patients (100% d'entre eux), sont vivants au moment du diagnostic. La courbe de survie part, en ordonnée, de la valeur 100. Au premier décès (mettons-le à 3 mois du temps 0), 1% de la population initiale disparaît : la courbe descend d'une marche de 1% au niveau de l'abscisse 3 mois. Imaginons que deux décès surviennent à 6 mois du temps 0 : la courbe enregistrera alors une marche descendante de 2% sur la base des 99 survivants, à l'abscisse 6 mois. Si seuls 5 patients ont été suivis plus de 5 ans (soit parce que tous les autres sont morts, soit parce qu'ils ont été diagnostiqués durant les 4 dernières années...) et que deux patients parmi ces 5 meurent à 5 ans, la marche descendante sera de 2/5, soit 40% de la population résiduelle.

Ceci explique que sur une courbe de survie, les marches descendantes deviennent de plus en plus marquées vers les temps importants. Il apparaît également qu'elles deviennent de moins

en moins précises, car elles portent sur une taille d'échantillon de plus en plus réduite : *l'intervalle de confiance autour de la valeur, donc l'incertitude sur la valeur, augmente lorsque la taille d'échantillon diminue, et elle diminue de façon inéluctable au fil de la courbe de survie.*

La prudence s'impose donc lorsque l'on parle de probabilité de survie en médecine, et certains points doivent toujours rester présents à l'esprit :

- Une probabilité de survie s'applique à un groupe de patients, mais ne s'applique pas à chaque patient de façon individuelle. On ne sait pas, car l'outil statistique ne répond pas à cette question, si le patient diagnostiqué tel jour vivra, ou ne vivra pas, dans 5 ans : sa survie sera de 100%, ou ne sera pas. Sa probabilité de survie ne sera pas égale à celle du groupe.

- La probabilité de survie à x années ne vaut pour le groupe, qu'au moment du temps 0 : la probabilité conditionnelle de survie varie en fonction du temps, et pour l'exemple des 5 patients survivants à 5 années parmi 100 patients initialement inclus, elle sera, à ce moment de la courbe, de 40 % pour le temps à venir.

- L'avenir, même entouré de statistiques, reste une donnée très difficile à approcher.

3- Comparaisons de courbes de survie en mode univarié : le test de log-rank.

Deux courbes de survie finissent toujours par se rejoindre, il ne s'agit que d'une question de temps… à terme, la différence ne peut pas être significative. Par conséquent, leur comparaison dépendra de la rapidité avec laquelle chaque événement (le décès, la survenue de telle complication), surviendra en cours de suivi, et du nombre d'événements survenant à un temps donné. A un instant t, x% de patients dans un groupe, y% dans l'autre, survivront. La comparaison de proportions, à cet instant t, pourra donc être réalisée par un test de chi-2. A l'instant suivant, une des proportions pourra avoir varié : un second test de chi-2 devra alors être réalisé. Il en sera de même pour chaque changement de proportion sur l'une, ou l'autre courbe, autrement dit, pour chaque nouvelle marche d'escalier sur l'une, ou l'autre courbe.

La comparaison globale de l'ensemble des courbes de survie dépendra donc de la résultante de l'ensemble des comparaisons de proportions réalisées à chaque changement dans une des courbes analysées, soit, en termes statistiques, de la somme de l'ensemble des chi-2 effectués ajustés pour la taille d'échantillon variant à chaque étape. Le log-rank test réalise cette analyse et résume l'ensemble des différences additionnées : il s'agit d'une forme particulière de Mantel-Haenszel.

En pratique, ce test statistique mesure les écarts entre les deux courbes de survie comparées, à chaque marche d'escalier survenant sur l'une, ou l'autre, des courbes de survie. Il teste ensuite l'hypothèse nulle : la somme de ces distances est égale à 0. Si cela est vrai, les deux courbes ne sont pas très éloignées l'une de l'autre, et elles traduisent une survie similaire. Si cela n'est pas vrai, alors les deux courbes sont éloignées l'une de l'autre, et traduisent des survies différentes.

4- Comparaison de courbes de survie en mode multivarié : le modèle de Cox

Plusieurs éléments peuvent entrer en ligne de compte dans la survie : la présence ou l'absence de maladie, bien sûr, mais aussi le type de traitement reçu, la compliance au traitement, le respect du protocole initial, la présence de co-morbidités, l'existence d'autres facteurs de risque, le stade ou grade de la maladie, l'état clinique du patient mesuré par un score au moment du diagnostic, etc.

Tous ces éléments ne peuvent pas être représentés sur une courbe de survie, mais l'on peut imaginer que certains d'entre eux peuvent être plus importants, pour la survie du patient, que le type de traitement reçu. Si l'on compare deux traitements A et B, deux formes de la maladie X et Y, il sera important d'égaliser les facteurs pronostiques éventuels entre les deux groupes. Cette égalisation est le but de la randomisation dans un essai contrôlé, mais elle peut ne pas toujours être atteinte. Elle est rarement atteinte lorsqu'il n'y a pas de randomisation, et la prise en compte des divers facteurs intervenant dans la survie devrait alors faire appel à une stratification étagée, et la comparaison effectuée dans des sous-groupes de patients homogènes pour chaque facteur pronostique. Cela aboutirait rapidement à une multiplication de sous-groupes à taille d'échantillon réduite, et à une perte importante de puissance pour l'étude. Les résultats deviendraient difficiles à interpréter.

Une autre possibilité consiste à intégrer l'ensemble des variables pouvant jouer un rôle sur le pronostic dans un modèle d'analyse de type régression logistique, adapté à l'analyse de courbes de survie (capable de prendre en compte des données censurées, et non censurées). Ce modèle particulier de régression logistique est le modèle publié par Cox, qui pourra donner une estimation du risque relatif associé à chaque facteur pronostique. Si cet ce risque relatif est significativement différent de 1, le facteur considéré jouera un rôle significatif dans le pronostic. Dans le cas contraire, il pourra être considéré comme non déterminant. Dans tous les cas, le modèle donnera une estimation du poids relatif de chaque facteur pronostique dans la survenue de l'événement mesuré (décès, survenue d'une complication iatrogène ou naturelle,…), sans perte de puissance secondaire à une stratification quelconque. Par contre, seules les observations sans donnée manquante pour toutes les variables analysées seront prises en compte, ce qui peut résulter en une importante diminution de la taille d'échantillon si ces variables n'ont pas été correctement renseignées : comme pour toute étude clinique, la rigueur dans le recueil des données conditionne la fiabilité des résultats, quel que soit le degré de sophistication de l'analyse employée.

Conclusion

L'analyse statistique est indispensable en médecine et en biologie pour tester des tendances au niveau d'échantillons (qu'ils soient des groupes de patients, de colonies cellulaires, des groupes de gènes, des familles de protéines…) ou de population. Les tendances testées seront valables pour les populations testées, mais ne peuvent pas s'appliquer telles quelles à un individu donné, notamment pour les études de survie intégrant, en plus des données connues (non censurées) des données de valeur inconnue au moment de l'analyse (censurées). L'analyse n'est valable que si les données ont été correctement recueillies, sont complètes, et si elles correspondent à une hypothèse de travail *a priori*. L'analyse 'en pêche à la ligne' dans une base de données, à la recherche d'associations ou de corrélations statistiquement significatives sans fondement biologique ou clinique pressenti par l'investigateur, a de grandes chances de ne donner de résultats significatifs *que par hasard et doit donc être proscrite*. Il est possible que ce type d'analyse non planifiée soit à l'origine d'une grande partie des résultats contradictoires rapportés dans la littérature médicale, et les controverses alors engagées pourraient ne reposer, au moins partiellement, que sur des effets de hasard… et

non pas sur des bases d'acquisition de connaissances dites 'scientifiques'. Comme pour toute méthode d'observation, les conditions d'application, les indications et les contrindications des différents tests statistiques doivent être connues par les médecins qui seront de facto, les utilisateurs des résultats. Tout résultat ensuite *doit être interprété* en fonction de l'ensemble de la méthodologie de l'étude en amont du test statistique, et des conditions dans lesquelles le test a été appliqué.

Un test statistique prouve rarement. Il essaie de circonscrire le fait du hasard autour des événements observés, à condition que le modèle sous-jacent au test décrive assez bien la réalité multiforme de la vie (ceci peut parfois être assez difficile à affirmer). Il permet, en revanche, plus facilement d'éviter des affirmations indues (tel traitement est supérieur, ou différent, de tel autre) lorsque la différence de fréquence des événements observés est trop proche de 0.

Pour en savoir plus :

Bouyer J. Méthodes statistiques. Médecine-Biologie. ESTEM, Editions INSERM, 2000.

Bailer JC III, Mosteller F. Medical uses of Statistics, 2nd Edition, 1992, NEJM Books, Boston, Massachusetts.

Hill C, Com-Nougué C, Kramar A, Moreau T, O'Quigley J, Senoussi R, Chastang C. Analyse statistique des données de survie, 2ème Edition, 1996. INSERM Médecine-Sciences, Editions Flammarion.

Tableau I : Récapitulatif des tests statistiques courants, de leurs indications et limites.

Type de variable	Test indiqué	Limites d'applicabilité	Alternative nécessaire si limite du test indiqué atteinte
Dichotomique (proportions), 2 groupes	Chi-2 à 1 degré de liberté Alternatives : Yates (Chi-2 corrigé) Mantel-Haenszel (Chi-2 pour données stratifiées)	Une des cellules au moins de la table contient moins de 5 événements attendus	Test de Fisher exact
Dichotomique (proportions), plusieurs groupes	Chi-2 avec degrés de liberté = (nombre de colonnes -1)(nombre de lignes -1)	Ne permet pas de savoir où se trouve une différence si le test est significatif, mais indique simplement qu'il existe une notion de différence au sein de la table des données Donnera des résultats faussés si une des cellules de la table contient moins de 5 événements attendus	Si l'une de cellules de la table contient moins de 5 événements attendus, un test de Fisher est théoriquement possible, mais peu d'ordinateurs auront une mémoire suffisante pour le réaliser.... savoir que les résultats sont faussés, difficilement interprétables, et envisager un autre type d'analyse des données.
Ordinale, 2 groupes	Wilcoxon rang sum test	Chaque groupe doit contenir au moins 10 observations	
Ordinale, groupes multiples	Kruskall-Wallis	Indique s'il existe une différence quelque part entre les différents groupes. N'indique pas où se trouve la différence.	
Quantitative, 2 groupes	Analyse de variance (comparaison de moyennes)	La distribution des variables doit être normale (gaussienne) dans les deux	Si la distribution n'est pas normale, employer le test de Wilcoxon comme

	Z-test si population très importante t-test si échantillon : situation clinique habituelle	groupes.	pour les variables ordinales. Le test de Wilcoxon et l'analyse de variance donneront des résultats similaires si n > 30
Quantitative, plusieurs groupes	Analyse de variance F-test : analyse globale Analyse en sous-groupes avec ajustement selon Bonferroni, Scheffe, Tukey,	La distribution des variables doit être normale dans les différents sous-groupes. Indique qu'il existe une différence quelque part entre les groupes. N'indique pas où se trouve la différence. Indique quels sous-groupes diffèrent entre eux. La sévérité du test dépend du type d'ajustement (Bonferroni étant le plus rigoureux)	Si la distribution n'est pas normale, utiliser le test de Kruskall-Wallis comme pour les variables ordinales.
Relation entre deux variables quantitatives	Test de corrélation selon Pearson	La distribution des deux variables doit être normale. Test très sensible aux valeurs extrêmes (outsiders)	Si la distribution n'est pas normale, utiliser le test de corrélation de rang (selon Spearman).
Relation entre une variable quantitative et plusieurs variables quantitative	Régression linéaire multiple	La distribution de toutes les variables doit être normale.	Si la distribution n'est pas normale, ou pour des variables de distribution particulière, il existe des modèles de régression non linéaire, avec ou sans

			transformation des variables
Relation entre une variable dichotomique, et plusieurs variables dichotomiques, ordinales ou quantitatives	Régression logistique	Permet l'estimation d'interactions entre les variables, le contrôle des éléments confondants et modificateurs d'effet. Ne requiert pas la distribution normale pour les variables quantitatives.	
Construction d'une courbe de survie	Modèle de Kaplan-Meier	Etablit la probabilité conditionnelle de survie. Sa précision diminue au fur et à mesure du suivi et de la diminution du nombre de patients suivis	
Comparaison de deux courbes de survie	log-rank test (équivalent du test de Mantel-Haenszel)	Ne permet pas la prise en compte des co-facteurs intervenant dans la survie	
Comparaison de deux courbes de survie, avec prise en compte de co-facteurs, d'éléments confondants, ou de modificateurs d'effet	Modèle de Cox	Permet de déterminer le poids respectifs des différents facteurs intervenant dans la survie. Permet d'estimer le risque relatif pour les facteurs dichotomiques. Permet de repérer les facteurs confondants ou les facteurs modificateurs d'effet.	