Statistiques pour audioprothésistes
ou
La bonne utilisation des statistiques[1]

Julie Bestel, audioprothésiste & PhD

[1] *Titre original modifié avec autorisation de J.Bestel. Ce document est paru dans les cahiers de l'audition en Novembre 2019. Le document a été édité et modifié par Nicolas Vannson, PhD pour Memau.*

TABLE DES MATIERES

I. Introduction

Cet article est issu du cours que je donne aux étudiant de 3e année dans l'école d'audioprothèse de Paris. Il rassemble plusieurs aspects de la statistique inférentielle fréquentiste. Il pourrait se sous-titrer ainsi : « **Quelle information tirer sur la population totale à partir de l'observation faite sur un échantillon restreint de sujets** ? ». Un autre titre, plus provocateur, serait : « De la dictature du petit p », qui prendra tout son sens au fur et à mesure de votre lecture. Ce cours ne contient presque pas de formules mathématiques, à dessein. Nous allons partir d'un exemple simple, que nous allons complexifier peu à peu. Nous parlerons d'acouphènes, mais il n'est pas nécessaire de connaître ce domaine pour comprendre cet article. J'ai simplement trouvé que les données collectées autour des acouphènes me donnaient toute la matière dont j'avais besoin.

Chose importante : nous ne ferons pas de recherche clinique, donc pas d'interprétation clinique des effets observés, d'où le choix d'exemples un peu farfelus. C'est la méthodologie de l'essai clinique, si elle est bien conçue, qui doit permettre d'attribuer l'effet observé à une cause, et plus précisément au traitement appliqué (port d'une aide auditive, écoute d'un bruit masquant, etc.). Cela donnerait lieu à un autre cours, notamment sur la méthodologie des essais cliniques.

Je me place dans un cadre fictif : imaginons que je me rende régulièrement à la consultation « Acouphènes » d'un service hospitalier spécialisé. J'ai eu l'autorisation de rester en salle d'attente et d'interroger les patients avant leur rendez-vous. Pas de considération éthique, ce n'est pas notre propos. Je demande à chacun d'évaluer le désagrément causé par son acouphène, sur une échelle de 0 à 10, à l'aide d'une échelle visuelle analogique (EVA), présentée comme une règle graduée. On parle d'échelle continue et linéaire, car l'on suppose qu'il y a la même « distance » entre la note 1,5 et 2 qu'entre 4 et 4,5, ou qu'entre 0 et 0,5, etc. Cette hypothèse est assez forte, et bon nombre d'échelles utilisées dans notre domaine ne la remplissent pas. Elle est importante au regard des traitements mathématiques que nous ferons subir aux notes obtenues. Par exemple, nous avons le « droit » de calculer la moyenne de plusieurs notes (collectées chez plusieurs individus), car la note est une variable linéaire continue (variable quantitative).

Notons au passage que je parle maintenant de « variable », terme emprunté à la statistique, qui signifie que je considère implicitement une variable aléatoire « note entre 0 et 10 » dont j'observe un ensemble de réalisations. Une réalisation de cette variable est la note donnée par un individu quand je l'interroge. Nous avons autant de réalisations ou observations de notes que d'individus interrogés.

II. Observations... Statistiques descriptives

Construction du fichier de données

Je range les données collectées dans un fichier Excel très simple, dans lequel j'ai défini plusieurs colonnes : la première contient des « codes ou identifiants » des sujets, avec un code par ligne. Une deuxième colonne contient l'EVA de gêne de l'acouphène pour chacun des sujets. J'ai donc autant de lignes que d'individus. La première ligne de ce fichier est différente des autres, car elle contient les libellés des variables : « Code sujet », « EVA », etc. Si je mesure d'autres choses, j'aurai autant de colonnes que de variables mesurées. Par exemple, je peux avoir une colonne « Âge » (en années), une colonne « EVA intensité », une colonne « Score THI » (Tinnitus Handicap Inventory, codée de 0 à 100), etc. Le tableau ci-dessous pourrait être extrait de cette collecte :

Code sujet	Age	EVA intensité	Score THI
S1	57	6	78
S2	75	3	34
etc	…	…	…

Table 1 : Donnes patients

Nuages de points

Pour servir notre propos, supposons que j'aie accumulé beaucoup de données, et que je dispose d'un fichier de 600 sujets. Concentrons-nous sur l'EVA de gêne de l'acouphène. Une première représentation graphique est un nuage de points, avec en abscisse le numéro du sujet et en ordonnée son EVA. Elle peut donner lieu à la Figure 1.

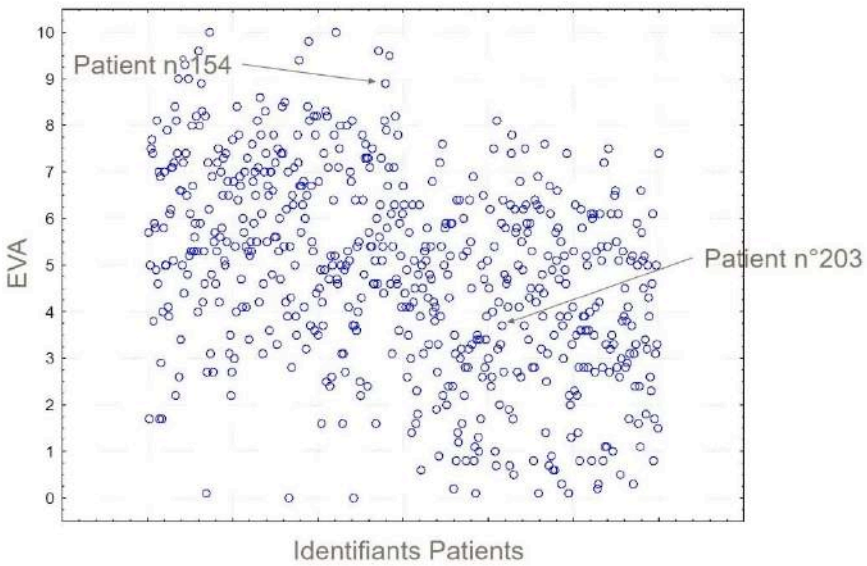


Figure 1 : Nuage de points simple, où les codes patients sont en abscisses et les EVA en ordonnées.

L'observation de ce nuage nous inspire un premier commentaire : toutes les valeurs d'EVA sont possibles. En y regardant de plus près, on peut noter néanmoins qu'il y a « peu » de sujets qui ont une EVA inférieure à 1, et peu de sujets qui ont une EVA supérieure à 9.

Si j'ai une deuxième variable dans mon tableau, comme le score THI, je peux choisir de faire un nuage de points à deux dimensions, où cette fois-ci chaque individu est caractérisé par deux notes : son EVA et son score THI. Nous verrons cela plus en détails dans la partie7. Si on collecte n données par sujet, chacun sera caractérisé par un ensemble de n valeurs (un « n-uplet »), que l'on pourrait représenter dans un espace à n dimensions. Evidemment dès que n> 3, on a du mal à se représenter les choses, et on fait donc des projections en 2D ou 3D par exemple.

Distributions et histogrammes

Revenons à notre premier nuage de points, très simple, celui de la Figure 1. Chaque sujet n'est caractérisé que par son EVA. On souhaite maintenant connaître la façon dont les EVA se « distribuent » ; on construit pour cela un histogramme. Dans cette représentation graphique, on met en abscisses les intervalles possibles de la variable, et en ordonnées le nombre d'individus dont l'EVA tombe dans chaque intervalle. Si on n'utilise pas de logiciel adapté, on doit passer par une étape intermédiaire, qui consiste à mettre en classes la variable continue EVA, c'est-à-dire construire une nouvelle variable catégorielle ordinale, qui prend des valeurs discrètes ordonnées. Par exemple, on peut choisir de découper en 11 classes : [-0,5 ; 0,5[(= classe 1), [0,5 ; 1,5[(= classe 2),..., [9,5 ;10,5] (= classe 11).

Dans mon exemple, la distribution des EVA est représentée par l'histogramme de la Figure 2.

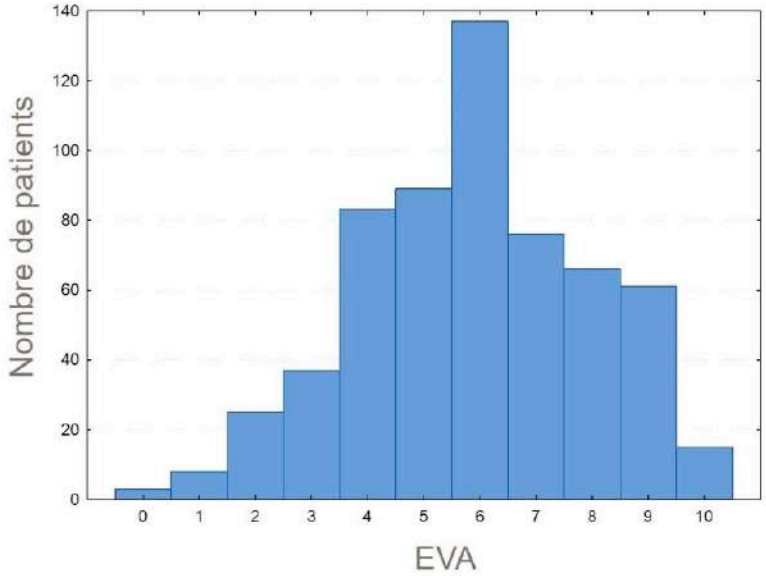


Figure 2 : Histogramme de la variable EVA construit sur un échantillon de 600 sujets, avec un découpage des valeurs d'EVA en 11 classes égales.

Sur ce graphique, on peut faire de nouvelles observations : il y a une « bosse », ou un pic de fréquence, qui signifie que l'intervalle [5,5 ; 6,5[est le plus représenté dans les valeurs d'EVA de mon échantillon. Il y a précisément 138 individus sur 600 qui donnent une EVA entre 5,5 et 6,5 à leur acouphène. On parle de **mode**. Comme notre distribution a un mode, on dit qu'elle est **unimodale**. NB : une distribution **bimodale** a « deux pics », c'est-à-dire deux valeurs (intervalles de valeurs) qui sont plus souvent données que les autres.

Toujours sur cet exemple, on peut également observer que la distribution est assez symétrique autour du mode ; il y a en effet à peu près autant de personnes qui ont des EVA entre 6,5 et 7,5

qu'entre 4,5 et 5,5. On retrouve également que très peu de gens ont une EVA autour de 0 ou une EVA proche de 10. Notons au passage que nous ne faisons que des approximations, nous ne sommes pas intéressés par le fait de savoir si notre distribution est parfaitement symétrique (modèle)

Loi normale, moyenne, écart-type, médiane, mode

La distribution de la Figure 2 est très bien approchée par une **distribution « normale »**, ou « **gaussienne** », modèle qui nous intéresse beaucoup en statistique. Elle a la forme d'une courbe en cloche et symétrique. Une gaussienne est entièrement caractérisée par deux paramètres : la moyenne et l'écart-type, souvent notés respectivement **μ** et **σ**. Quand on connaît ces deux valeurs, on sait représenter totalement la courbe. Des exemples sont rassemblés sur la Figure 3 (extraite de Wikipedia). En ordonnée figure la « densité de probabilité », qui peut être vue comme une proportion (ou fréquence) de sujets. L'aire sous chacune des courbes vaut 1.

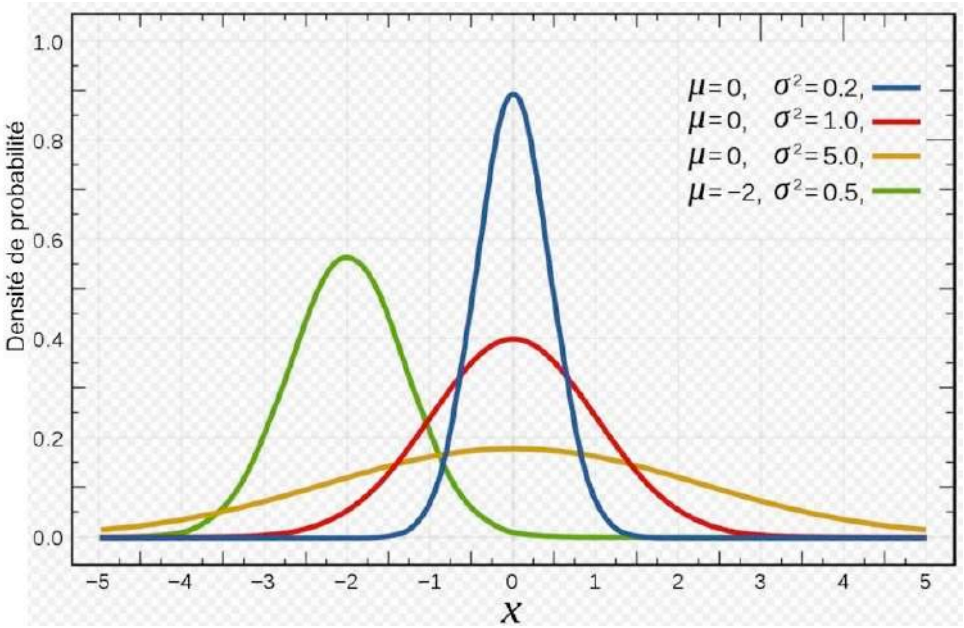


Figure 3 : Distributions gaussiennes ou normales, avec différentes valeurs de moyenne et d'écart-type. La courbe rouge correspond à une gaussienne centrée-réduite : moyenne nulle et écart-type de 1. Plus l'écart-type est grand, plus la courbe est étalée autour de la moyenne.

Pour l'instant nous n'avons fait qu'observer ce qu'il y a dans nos données, sans tenter d'expliquer quoi que ce soit. Dans cette étape indispensable, on calcule des « indicateurs » à partir des données. Puisque l'EVA est une variable quantitative, je peux calculer, entre autres : la moyenne, la médiane, l'écart-type, etc. Je peux aussi le faire sur d'autres variables du fichier, mais à ce stade je n'exploite pas les relations qui peuvent exister entre elles. **On parle de statistiques descriptives, étape essentielle pour connaître ses données.**

Moyenne	**Somme de toutes les valeurs divisées par le nombre de valeurs**
Médiane	Valeur en-dessous de laquelle se trouvent 50% des données (donc les autres 50% sont au-dessus).
Mode	Valeur la plus fréquente dans l'échantillon de données.
Variance	Moyenne des carrés des écarts entre chaque valeur de la série et la moyenne
Ecart-type	Racine carrée de la variance. Plus l'écart-type est grand, plus les valeurs sont « dispersées », ou étalées, autour de la moyenne.
1er quartile	Percentile à 25%, soit la valeur en dessous de laquelle se trouvent 25% des données
3e quartile	Percentile à 75%, soit la valeur en dessous de laquelle se trouvent 75% des données

Tableau 2 : Rappel des principaux indicateurs calculés sur des variables quantitatives (statistiques descriptives)

III. Echantillon et inférence

Echantillon et inférence Fluctuations d'échantillonnage

Les 600 sujets interrogés constituent un « **échantillon** », qui est une sous-population de la population totale. Quelle est donc cette population totale ? on pourrait dire que c'est celle de tous les acouphéniques du monde... ou de région parisienne... C'est à moi de la définir, mais c'est important d'en avoir conscience car c'est sur elle que portera le résultat « inféré » : qu'est-ce que j'apprends sur la population totale à partir de mon observation sur l'échantillon, tiré parmi cette population ?

Plus précisément, si je refaisais mon expérience en tirant un autre échantillon de 600 personnes, j'observerais des EVA et des THI peut-être légèrement différents. Par exemple, peut-être que la moyenne calculée sur toutes les EVA de mon premier échantillon serait « différente » de celle calculée sur le deuxième échantillon. Le fait que ces deux moyennes diffèrent est dû à ce qu'on appelle les « **fluctuations d'échantillonnage** ». Si notre échantillon était la population tout entière, on ne se poserait pas de question, la moyenne des EVA serait par construction la vraie moyenne des EVA de la population totale. Evidemment, cette population n'est pas observable, c'est précisément pour cela qu'on travaille sur un échantillon. Je n'ai accès qu'à une sous-partie de la population, sur laquelle je peux calculer une moyenne, et je vais donc chercher à savoir dans quelle mesure cette moyenne est « proche » de la vraie moyenne des EVA. **C'est toute la question de la statistique inférentielle.**

Intervalle de confiance autour de la moyenne

Comme on ne peut pas connaître la vraie valeur de la moyenne, on en donne un encadrement, avec un risque de se tromper en disant qu'il contient la vraie moyenne. Ce risque est en général fixé à 5%, mais il pourrait l'être à 1% ou à une autre valeur. Cet encadrement est calculé à partir de la moyenne de l'échantillon, et s'appelle **l'intervalle de confiance à 95%** si l'on fixe le taux d'erreur à 5%. L'interprétation est la suivante : si nous faisions 100 tirages aléatoires

d'échantillons (de 600 sujets pour notre exemple), pour 95 d'entre eux l'intervalle de confiance calculé contiendrait la vraie moyenne.

Les intervalles de confiance (IC) sont calculés par des logiciels, et on trouve dans les livres les explications pour calculer un intervalle de confiance : autour d'une moyenne, d'une proportion, dans le cas de grands ou de petits échantillons, etc (voir par exemple le livre « Statistics with Confidence »). Ce n'est pas le propos de cet article. Retenons ici une formule générale :

IC95%= [moyenne – coeff*SE ; moyenne + coeff*SE]

Où SE est « **l'erreur type** » ou « **erreur standard** », autour de la moyenne. Elle vaut $\sigma / \sqrt{n}$ où σ est l'écart-type des données (estimé sur mon échantillon), n est le nombre de données (600 dans mon exemple). Coeff = 1,96 si n > 30, un peu plus grand si n < 30 (on trouve cette valeur dans des tables ou on utilise un logiciel qui fait cela tout seul).

On constate donc que l'IC est d'autant plus étroit, donc l'estimation de la moyenne est d'autant plus précise que :

- il y a peu de dispersion autour de la moyenne : σ faible
- Le nombre de données est grand.

On retrouve ce que nous intuitons facilement : moins les données fluctuent autour de la moyenne, et plus il y a de sujets dans notre base, plus nous pouvons donner une estimation précise de la « vraie » moyenne. On représente l'IC par une barre encadrant la moyenne. On parle de façon générique de « **barre d'erreur** », et il faut donc toujours préciser quelle est cette barre d'erreur. En effet, on peut choisir plusieurs barres d'erreur dans les logiciels : l'IC — et c'est la barre d'erreur la plus informative quand on représente une moyenne —, l'écart type, et l'erreur type. Deux remarques :

- afficher la moyenne seule, sans barre d'erreur, n'est pas suffisant : il peut se passer beaucoup de choses autour de cette moyenne, comme nous le verrons plus loin.
- Certains auteurs représentent l'erreur type, car elle est plus petite que l'écart-type (forcément car elle est en $1/\sqrt{n}$) ! Cette justification n'est évidemment pas acceptable.

Nous retiendrons qu'il faut mettre comme barre d'erreur autour de la moyenne l'intervalle de confiance (à 95%). Il donne une idée de la précision de la valeur, et il nous renseigne sur la **taille de l'effet**, comme nous le verrons plus loin. Sur notre exemple, l'estimation de la moyenne est très précise car nous avons un grand nombre de patients (600), comme le montre la Figure 4. A écart-type égal, l'estimation serait moins précise si nous avions moins de sujets (IC plus large).

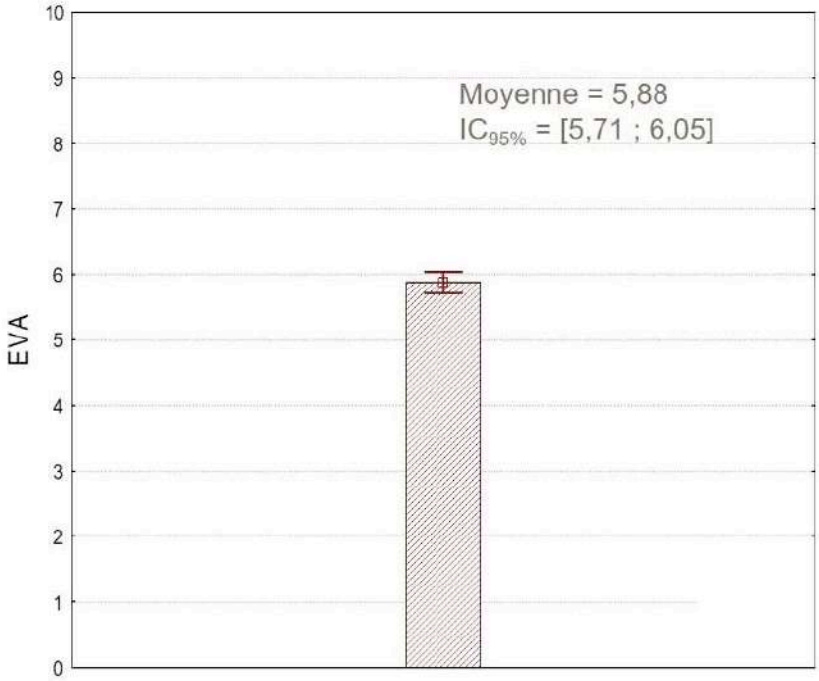


Figure 4 : Représentation correcte de la moyenne des EVA avec l'intervalle de confiance à 95% comme barre d'erreur

Comparaison de deux moyennes et inférence statistique

Supposons maintenant que les patients que j'ai interrogés soient séparables en deux groupes, selon leur régime alimentaire : soit omnivore, soit végétarien. Cette information est entrée dans le fichier Excel comme nouvelle variable « Régime », qui peut prendre deux **modalités** : « Omnivore » et « Végétarien ». Mon fichier (fictif) est bien fait : il y a à peu près autant de représentants des deux catégories, soit environ n = 300 sujets dans chacun des deux groupes. Dans le logiciel, je peux choisir, dans le nuage de points, d'étiqueter différemment les sujets en fonction de leur régime, comme le montre la Figure 5.

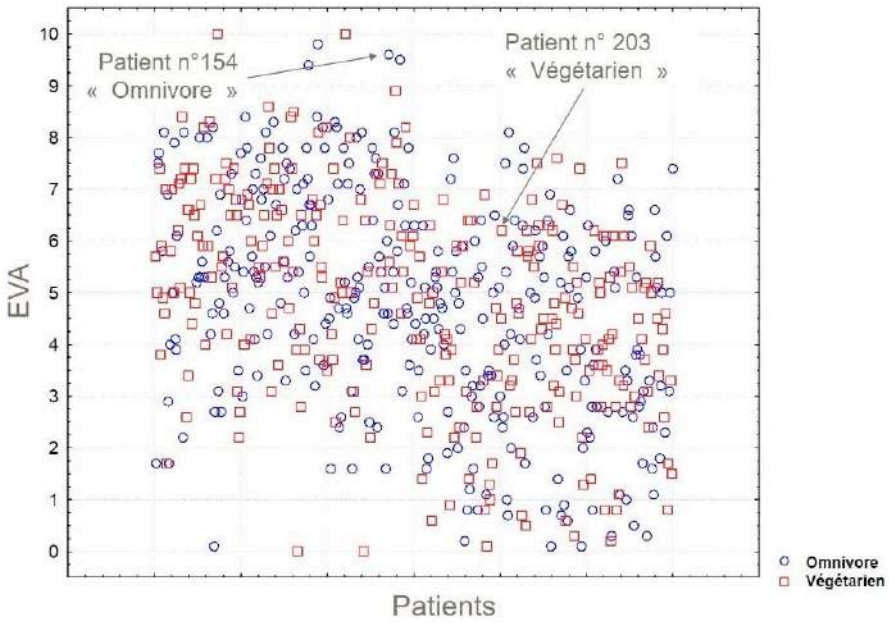


Figure 5 : Nuage de points étiquetés selon une variable de catégorisation à deux modalités

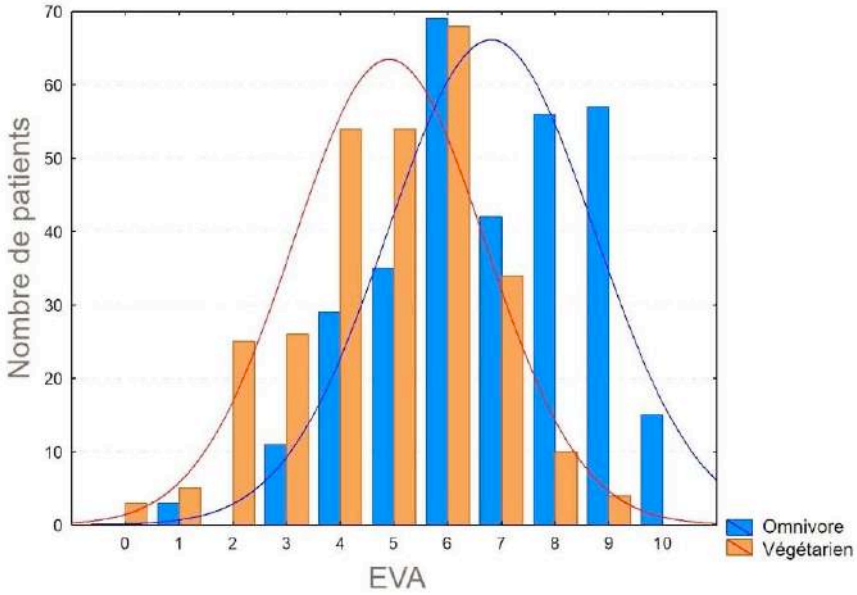


Figure 6 : Histogrammes (et courbes gaussiennes) catégorisés en fonction de la variable « Régime » à deux modalités : « Omnivore » et « Végétarien ».

Les histogrammes correspondant à chacun des deux sous-échantillons sont représentés sur la Figure 6. On voit sur cette figure que les distributions diffèrent légèrement. Si on demande en plus au logiciel de tracer les gaussiennes qui s'ajustent au mieux aux distributions, on constate qu'elles ont à peu près le même étalement (même écart-type), mais qu'elles ne sont pas centrées sur les mêmes valeurs. On peut calculer les deux moyennes : elles valent 4,88 et 6,76 respectivement pour les patients végétariens et pour les patients omnivores. Il y a donc une différence de moyennes d'EVA de 1,88 entre les deux échantillons.

La question que l'on se pose en statistique (inférentielle) est : « Qu'en est-il de la différence de moyennes dans la population tout entière (la population de tous les omnivores et végétariens acouphéniques) ? ». Rappelons que l'on souhaite répondre à cette question à partir de l'observation d'un échantillon (celui de ma base). Je vous propose ci-dessous plusieurs reformulations de cette question, qui disent toutes la même chose, et sur lesquelles il est important de s'arrêter pour comprendre ce que l'on fait :

1) « Est-ce que les deux moyennes observées proviennent de populations dont les moyennes sont différentes ? ».

2) « Est-ce que la différence observée sur mon échantillon est due aux fluctuations d'échantillonnage, ou existe-t-elle réellement ? »

3) « Si je refaisais plusieurs fois cette expérience, à savoir interroger 600 patients sur leur EVA, retrouverais-je souvent une différence dans les moyennes d'EVA entre les omnivores et les végétariens, ou au contraire est-ce que cette différence serait nulle la plupart du temps » ?

Pour répondre à la question 3), on commence par calculer les intervalles de confiance pour chaque échantillon, représentés sur la Figure 7. On peut déjà constater qu'ils ne se recouvrent pas.

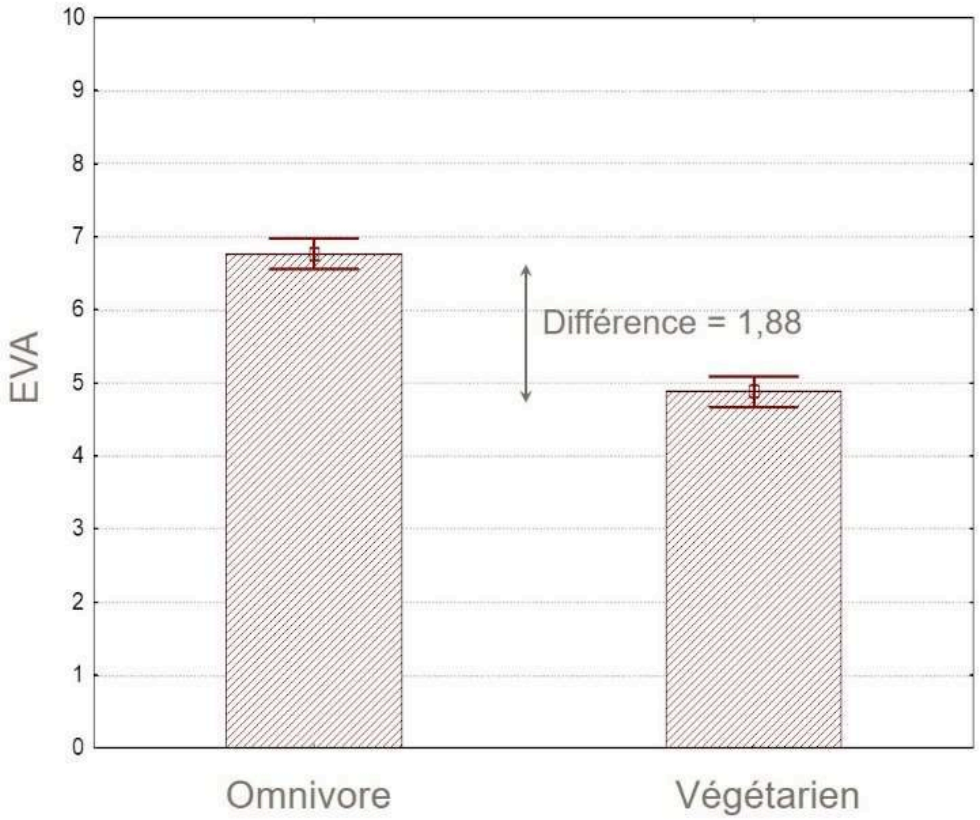


Figure 7 : Moyennes des EVA et Intervalles de confiance à 95% pour les deux groupes de sujets « omnivores » et « végétariens »

Comparaison de deux moyennes : le « t-test » ou test de Student

Pour aller plus loin, on va attacher à la différence observée entre les moyennes un indice de «crédibilité» ou «fiabilité». On le nomme «**petit p**», ou «**p-value**» en anglais. C'est une **probabilité**, donc comme toute bonne probabilité qui se respecte, elle est comprise entre 0 et 1. On l'interprète comme suit : plus le petit p est petit (proche de 0), plus la différence observée entre les moyennes a peu de chances d'être due aux fluctuations d'échantillonnage (c'est-à-dire qu'elle existe réellement, entre les deux vraies moyennes). Notons que nous n'avons encore rien dit de l'amplitude de la différence.

Le petit p est le résultat d'un « test statistique », ou test d'hypothèses (voir plus loin). Dans le cas des deux moyennes, ce test s'appelle t-test ou test de Student.

Le petit p est le résultat d'un calcul qui est valide à condition que les données respectent certaines conditions. Comme il s'agit de moyennes, il est logique d'attendre que les distributions soient bien résumées par leurs moyennes. C'est bien sûr le cas des gaussiennes, qui sont entièrement connues grâce à deux **paramètres** : la moyenne et l'écart-type, comme nous l'avons vu dans la partie 2 (*Loi normale, moyenne, écart-type, médiane, mode)*

Quand le nombre de sujets devient grand, peu importe la forme de la distribution des données, car

ce qui compte pour que le calcul du t-test soit valide, c'est que la distribution des moyennes suive une loi de probabilité définie par des paramètres.

Ceci découle de la « **loi des grands nombres** » (ou « **théorème central limite** »). Appliquée à notre problématique, cette loi dit que si l'on tire au sort plein d'échantillons de n individus, que l'on calcule la moyenne des EVA pour chacun des échantillons de taille n alors la distribution des **moyennes d'échantillons suit une loi normale dès que n est supérieur ou égal à 30**. Cette propriété est vraie quelle que soit la forme de la distribution des données pour n > 30 (environ). En dessous de 30 (environ), il faut une condition sur la distribution des données elles-mêmes : si elle est normale, alors la distribution des moyennes est encore régie par des paramètres ; si elle ne l'est pas, nous ne pouvons plus rien dire sur la distribution des moyennes d'échantillons.

Ainsi, le calcul du t-test requiert, selon le nombre de sujets, que les distributions soient régies par des paramètres. C'est pour cela qu'on parle de test paramétrique.

En pratique, on n'a pas d'hypothèse à faire pour réaliser un t-test dès qu'on a plus de 30 individus, et on doit évaluer la normalité si on a moins de 30 individus (30 valeurs d'EVA). Notons qu'il faudrait également vérifier l'égalité des variances dans chacun des deux groupes, même si le t-test reste robuste à une violation de cette hypothèse (il faut juste que les distributions ne soient « pas trop » asymétriques). Dans le cas où n < 30 et où la distribution des données n'est pas normale, il faudra utiliser un autre type de test (**non paramétrique**).

Pour information, la distribution des moyennes des échantillons s'appelle **la distribution d'échantillonnage**.

Mais comment calcule-t-on un petit p ?

Plaçons-nous dans le cas du t-test. Le raisonnement est assez tordu, et se rapproche de ce que vous avez vu dans vos cours de maths : le **raisonnement par l'absurde**. On veut savoir s'il y a une différence, donc on va commencer par dire qu'il n'y en a pas. Logique, non ? On suppose donc qu'il n'y a pas de différence EN VRAI, c'est-à-dire entre les deux moyennes des deux sous-populations totales (tous les omnivores et tous les végétariens acouphéniques du monde). Cette hypothèse est notée H_0, également appelée « hypothèse nulle ».

Si cette hypothèse est vraie, alors, quand je tire au hasard un échantillon de n sujets (n = 30, 100, 300, 1000...) et que je calcule la moyenne des EVA pour les deux sous-groupes, il se peut, à cause des fluctuations d'échantillonnage, que la différence des moyennes soit quand même non nulle. Toutes les valeurs de différences ont une certaine probabilité d'apparition, que l'on sait (les matheux savent) calculer. Pour n ≥ 30, d'après la loi des grands nombres, c'est une gaussienne. Si n < 30 et si la distribution des données suit une loi normale, la distribution des moyennes des échantillons suit une autre loi, celle de Student, qui est un autre type de courbe en cloche et symétrique. On sait aussi calculer la fréquence d'apparition de toutes les différences de moyennes (la distribution des moyennes).

On note Δm la variable « différence des deux moyennes » (c'est une variable continue sur l'ensemble des nombres réels). Pour simplifier les calculs, on suppose que les deux échantillons sont de même taille n ≥ 30, et de même écart-type s (l'écart-type estimé sur l'échantillon n'est plus noté σ mais s). Alors la loi des grands nombres nous dit que la distribution des Δm suit une loi normale centrée sur 0 et d'écart-type $\sqrt{(s^2/n)}$. On normalise cela en disant que la variable z = Δm $/\sqrt{(s^2/n)}$ suit une loi normale centrée et réduite (écart-type de 1), ce qui nous permet de représenter facilement sa distribution. Quand on représente la distribution en mettant en ordonnées les fréquences d'apparition des valeurs, la distribution est alors égale à la « densité de probabilités de la variable ». Cela signifie que la probabilité pour qu'une valeur soit supérieure à

une valeur seuil est égale à la portion d'aire sous la courbe au-delà de la valeur seuil (en valeur absolue), comme on l'a représenté sur la Figure 8.

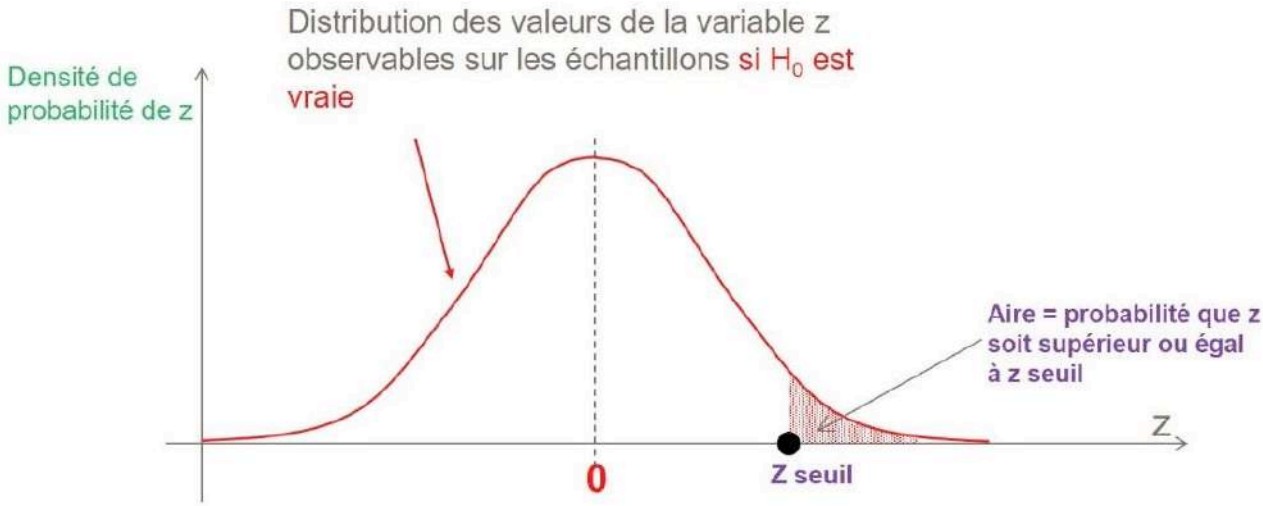


Figure 8 : Densité (distribution) de probabilités de z si H0 est vraie. L'aire totale sous la courbe vaut 1, par définition de la densité de probabilités. La probabilité d'observer une valeur de z plus grande qu'une valeur donnée seuil z-seuil est égale à la portion d'aire sous la courbe entre z-seuil et l'infini.

Comme on le voit sur la distribution, plus z est grand plus sa probabilité d'être observé est faible : on est sur la « queue » de la gaussienne. Au contraire, les petites valeurs de z sont très probables (vers le centre de la gaussienne). Dans notre échantillon, on observe une valeur particulière de Δm que l'on note $m_1 - m_2$ (c'est-à-dire : moyenne des EVA dans l'échantillon 1 – moyenne des EVA dans l'échantillon 2. On note : z_0 = m1-m2 //√(s^2/n) . Et on se demande alors si z_0 a de fortes chances de se produire ou pas, c'est-à- dire sa probabilité d'apparition si H_0 est vraie. **C'est cette probabilité que l'on appelle le « petit p »**.

Lien entre seuil de significativité alpha (α) = 0,05 et petit p

Pour calculer le petit p associé à la valeur z_0 observée, on reporte z_0 sur l'axe des abscisses de la distribution d'échantillonnage. La valeur seuil à laquelle on compare z_0 est la valeur notée z/2 sur la Figure 9. est appelé « **risque de première espèce »**, ou « taux d'erreur » que l'on juge acceptable. Dans bon nombre de domaines, on a pris l'habitude de fixer ce seuil à 0,05 (c'est une probabilité seuil). On a aussi l'habitude de répartir ce risque de manière « **bilatérale** » : 0,025 d'un côté — pour les valeurs de z grandes et positives — et 0,025 de l'autre côté — pour les valeurs de z grandes et négatives. Si la probabilité d'observer une différence donnée est inférieure à 0,025, alors on décide que c'est peu probable. Cela est tellement peu probable que l'on se dit que l'hypothèse de départ n'était pas réaliste : ce qu'on observe n'est pas compatible avec l'hypothèse de départ. Donc on « rejette H_0 » selon laquelle il n'y a pas de différence entre les deux moyennes.

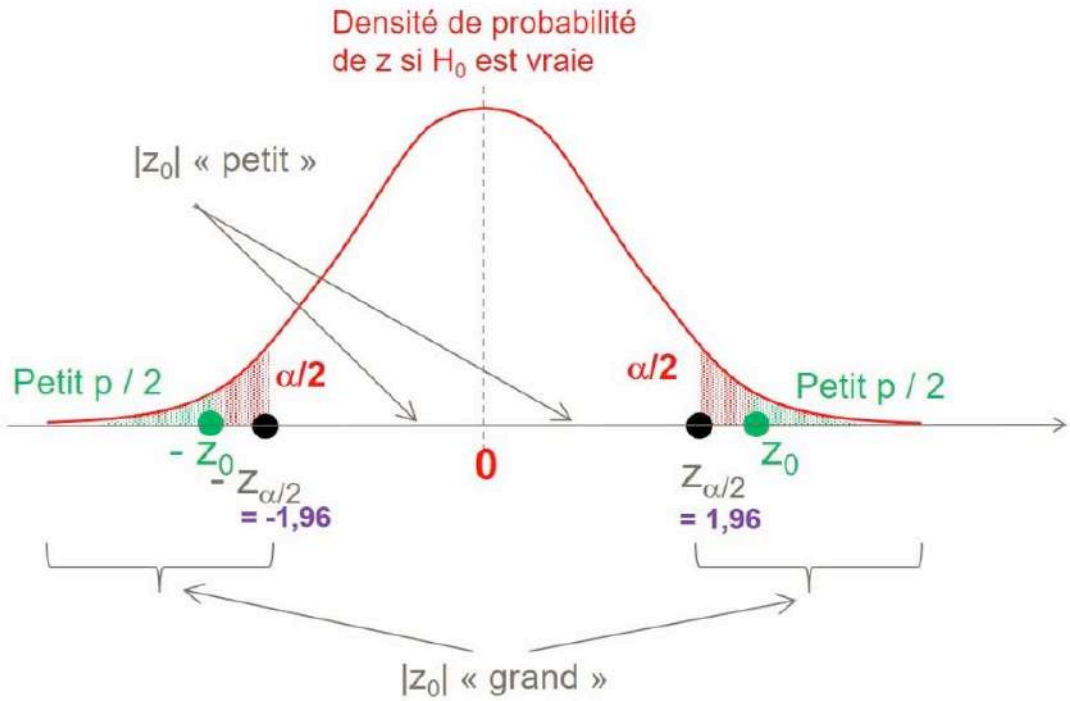


Figure 9 : Distribution de probabilités de la variable z = Δm /√(s^2/n). quand H0 est vraie. La valeur z0 = m1-m2 /√(s^2/n) est calculée sur un échantillon particulier. Sous H0 : les grandes valeurs de z0 (supérieures au seuil) ont une probabilité faible d'être observées, alors que les petites valeurs de z0 ont une grande probabilité d'être observées. On compare z0 en valeur absolue à deux seuils fixés arbitrairement. Le petit p se divise en deux parts égales, correspondant aux deux aires hachurées en vert. Ici chaque aire verte est plus petite que l'aire rouge, qui correspond à la valeur critique ⍰/2, où ⍰ est appelé risque de première espèce.

Les valeurs seuils auxquelles il faut alors comparer z0 sont respectivement -1,96 et 1,96. On les trouve dans les tables pour α = 0,05, test bilatéral, et n > 30.

Fixer α à 0,05 signifie que l'on n'interprète pas le résultat quand petit p est supérieur à 0,05. Dans ce cas on ne peut pas rejeter H0. On interprète un résultat dès que petit p < 0,05. Si par exemple petit p = 0,03, on dit que l'on rejette H0 avec 3% de chances de le faire à tort.

Quand nous disons « on rejette H0 », il faut bien faire attention qu'à ce stade, on a seulement établi que z0 = m1-m2 /√(s^2/n) est statistiquement différent de 0. Exemple : petit p = 0,03 signifie que si je recommence plein de fois mon expérience, à savoir « tirer deux échantillons de taille n et calculer la moyenne des EVA dans chacun », alors 97 fois sur 100 je vais trouver que la quantité z0 n'est pas nulle. Or, l'expression de z0 implique : z0 grand en valeur absolue (on note |z0|) ⇔ $|\boldsymbol{m1} - \boldsymbol{m2}|$ grand ET/OU s petit ET/OU n grand

On retrouve donc le fait que plus la taille de l'échantillon est grande, et/ou plus les données sont rassemblées autour de la moyenne (écart-type faible), plus la différence devient statistiquement significative. A ce stade, ce n'est donc pas parce qu'une différence est statistiquement significative que l'on connaît « l'amplitude » de cette différence. Malheureusement (pour ceux en quête d'une certaine « vérité »), ou heureusement (pour ceux dont la vie semble dépendre du petit p) : dans les grands échantillons, tout devient statistiquement significatif...

Différence statistiquement significative et cliniquement significative

Nous venons de voir que nous ne faisons pas de magie : le petit p est calculé en utilisant les données de l'échantillon : moyenne, nombre de sujets, écart-type. En particulier : le petit p diminue quand le nombre de sujets augmente, et donc dans les grands échantillons tout devient « statistiquement » significatif, même une toute petite différence qui n'a pas d'intérêt en clinique. Seul l'expert du domaine, donc vous, peut statuer sur l'intérêt clinique d'une différence. Il faut donc toujours regarder les valeurs, et ne jamais interpréter un petit p sans commenter la différence à laquelle il correspond.

Illustration 1 (Figure 10 gauche) : j'ai trafiqué mes valeurs d'EVA, et justement parce que j'ai beaucoup de patients (environ 300 dans chaque groupe), j'ai pu produire un jeu de données qui mène à une différence d'EVA de 5,76 – 5,41 = 0,35. Le t-test donne un petit p associé de 0,023. La présentation du résultat, telle qu'on pourrait la trouver dans un article scientifique, se trouve sur la partie gauche de la Figure 10. On peut très vite conclure : « Houra, il y a une différence d'EVA statistiquement significative ! ». Mais regardons de plus près : quelle est cette différence ? Elle vaut 0,35... Tiens...Hum... Quel intérêt clinique cela peut-il bien représenter ? Sur une échelle de 0 à 10 ? Passer de 5,76 à 5,41 ? Est-ce que cela est le reflet d'une gêne moindre pour les végétariens ? Cette différence a beau être statistiquement significative, elle n'est PAS cliniquement significative. Je vous laisse méditer...

Illustration 2 (Figure 10 droite) : j'ai de nouveau manipulé mes données, et réduit considérablement la taille de mon échantillon, de telle sorte que j'ai 30 sujets dans chaque groupe. J'ai calculé les moyennes des EVA dans chaque échantillon, et je trouve respectivement 5,5 et 5,03 soit une différence de 0,47. Cette différence est donc plus grande que dans l'exemple qui précède. Cependant, le nombre de sujets est bien plus petit, et le logiciel me donne comme petit p associé à cette différence : 0,27. La différence observée n'est donc absolument pas statistiquement significative. Au clinicien de juger si 0,47 de différence sur une EVA est cependant cliniquement intéressant. Si c'est le cas, il faudrait plus de patients pour montrer qu'elle est statistiquement significative (qu'elle a peu de chances d'être due au hasard de l'échantillonnage). On dit dans ce cas que mon expérience manque de « puissance » (*voir la partie 8 : DIVERS*).

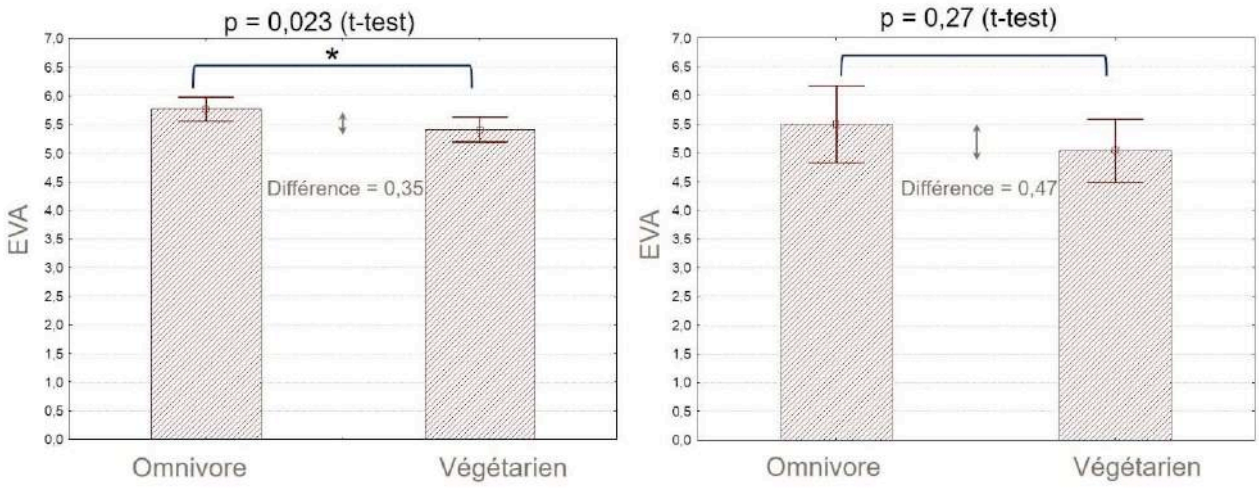


Figure 10 : A gauche, jeu de données menant à une différence de moyennes d'EVA de 0,35 et un petit p < 0,05 (résultat statistiquement significatif). A droite : autre jeu de données menant à une différence de moyennes d'EVA de 0,47 (donc plus grande) et un petit p > 0,05 (résultat non statistiquement significatif).

Les tests d'hypothèses

Revenons sur notre fameuse hypothèse H_0. Dans le cas de la comparaison de moyennes, H_0 se formule ainsi : il n'y pas de différence de moyennes dans la population totale, c'est-à-dire $m_1 = m_2$. Le raisonnement que nous avons conduit dans le cas du t-test consiste à confronter notre observation expérimentale de $m_1 - m_2$, et d'en conclure (ou pas) à l'incompatibilité avec H_0. On rejette H_0 quand on trouve que la différence observée a tellement peu de chances de se produire si H_0 est vraie que l'on considère que H_0 devait être fausse (raisonnement par l'absurde).

En statistique, on fait toujours des hypothèses... On va toujours poser une hypothèse H_0 (hypothèse nulle), chaque fois que l'on veut prendre une décision, donc toujours avec un risque de la prendre à tort (risque acceptable fixé arbitrairement à 5%). On ne peut jamais dire par exemple : mes données sont distribuées selon une loi normale. On peut seulement dire : « je ne fais pas trop d'erreur (je ne prends pas trop de risques) en disant que les EVA observées sont issues d'une population dans laquelle les EVA se distribuent selon une loi normale ». La précision sémantique peut paraître tenue, mais elle est vraiment à la base du raisonnement en statistique inférentielle.

Ainsi, on va aussi faire un test d'hypothèses pour justifier des conditions d'application du t-test (qui est lui-même un test d'hypothèses). Dans un petit échantillon ($n < 30$), on doit savoir si la distribution des données, dans chaque groupe, suit une loi gaussienne (normale). On pose alors comme hypothèse H_0 : « ma distribution et la distribution normale ne sont pas différentes ». Cette hypothèse nulle n'a rien à voir avec celle posée pour faire le t-test, elle correspond à celle qu'on doit formuler pour explorer la condition principale d'application du t-test.

Tester la normalité, tester l'égalité des variances...

Les tests de Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors,... et j'en passe, sont autant de tests statistiques utilisés pour évaluer si l'on ne fait pas trop d'erreur en disant qu'un jeu de données est issu d'une distribution normale. Certains sont plus conservatifs que d'autres, ont des comportements asymptotiques différents, sont plutôt recommandés pour les petits ou les grands échantillons, etc. Bref, restons modestes, dans ce que nous faisons il s'agit juste de ne pas raconter trop de bêtises. Le test qui fait consensus et nous suffira amplement est celui de Shapiro-Wilk (S-W).

Le test de S-W calcule une quantité, non pas un z comme dans le cas du t-test, mais une autre « statistique » (c'est le terme consacré), que l'on peut appeler SW (comme Shapiro-Wilk). Elle va être reportée sur une courbe, qui n'est pas une gaussienne ou une courbe de Student, mais une autre courbe « tabulée », et l'on regarde où se situe la grandeur SW calculée sur notre échantillon sur l'axe des abscisses. En réalité c'est toujours le logiciel de statistique qui fait cela, et produit aussi le petit p associé. Si $|SW|$ est « grand », le petit p va être petit (alpha = 0,05), et si $|SW|$ est « petit », le petit p associé va être grand (> 0,05). La règle de rejet est toujours la même : on rejette H_0 si $p < 0{,}05$ et on accepte H_0 si $p > 0{,}05$.

Il y a souvent une confusion chez les étudiants à propos de ces tests : pour tous les tests de normalité, on est « content » quand on ne PEUT PAS rejeter H_0, car, souvenons-nous que H_0 est l'hypothèse selon laquelle « ma distribution et la distribution normale ne sont pas différentes ».

Autre remarque : la plupart des tests utilisés pour évaluer la normalité sont très mauvais pour rejeter H_0 pour les « petits effectifs », autrement dit, ils sont très « sympas ». Raison de plus pour vraiment suivre leur conseil quand ils rejettent l'hypothèse de normalité. Et surtout, ne pas se contenter d'appliquer bêtement un test, mais toujours regarder la distribution des données.

Une autre condition à vérifier pour que le calcul du t-test soit valide est l'égalité des variances de chacun des échantillons. Cependant, le t-test est assez robuste à la violation de cette hypothèse, notamment pour n grand. Néanmoins, si l'on veut vérifier cette hypothèse, on utilisera un test de Levene (ou Bartlett).

IV. Généralisation du t-test et AnOVa

Revenons à notre jeu de données. En réalité, on pense que la distinction « omnivore/végétarien » est trop grossière : parmi les végétariens il y a des végans purs et des végétariens qui mangent des protéines animales. On souhaite donc être plus précis en séparant donc le régime en 3 catégories : omnivore, végétarien, végan. La question que l'on se pose est : « Y a-t-il une différence de moyenne des EVA selon le régime ? ». Pour cela, on réalise un test appelé « **AnOVa** », qui est l'abréviation de « **Analysis of Variance** ». NB : on note souvent **ANOVA** plutôt qu'AnOVa.

ANOVA à un facteur

Contrairement à ce que pourrait laisser imaginer ce terme, ce test sert à comparer des moyennes. Il répond à la question : « Y a-t-il au moins une moyenne qui diffère d'une autre ? ». Pour répondre à cette question, on (le logiciel) décompose l'ampleur de la dispersion totale des données (la variance totale) en deux termes : celle qui est due aux dispersions au sein de chaque sous-groupe (variance intra population) et celle qui est due à la dispersion entre les sous-populations (variance inter-population).

En gros, si les moyennes dans chaque sous-groupe sont toutes égales, la variance totale est égale à la variance dans chaque population ; si une paire de moyennes diffère, la variance totale est plus grande que la variance de chaque population : il y a un terme de « variance résiduelle » non nul. C'est donc l'importance du terme de variance résiduelle qui va permettre de statuer sur l'existence ou non d'une différence de moyennes. Nous n'irons pas plus loin dans les calculs, mais ce début d'explication permet de comprendre pourquoi on parle d'analyse de la variance.

Le résultat d'un test d'ANOVA est de nouveau un petit p, donc il faut savoir à quelle hypothèse H_0 il correspond. L'hypothèse nulle sous-jacente est : « il n'y a pas de différence entre toutes les moyennes prises 2 à 2 », c'est-à-dire, dans notre cas : la moyenne des EVA chez les omnivores = moyenne des EVA chez les végétariens = moyenne des EVA chez les végans. L'hypothèse alternative H_1 est : « il existe au moins une paire de moyennes dont les moyennes sont différentes ». Attention donc, petit $p < 0,05$ dans une AnOVA signifie seulement : l'une des moyennes diffère d'une autre moyenne, mais nous ne savons pas de quelles moyennes il s'agit. Cela peut être : $EVA_{omnivore} \neq EVA_{végétarien}$ ou (/et) $EVA_{omnivore} \neq EVA_{végan}$ ... C'est pourquoi, quand le petit p issu d'une ANOVA est < 0,05, on réalise ensuite des **tests « post hoc ».** Ils consistent à rechercher quelles paires sont concernées par une différence statistiquement significative. Cela revient à faire des comparaisons 2 à 2 avec un seuil ? qui ne vaut pas 0,05 mais qui est « corrigé » par le nombre de comparaisons souhaitées. Si l'on fait 3 comparaisons : « omnivore vs végétarien », « omnivore vs végan » et « végétarien vs végan », on utilisera un seuil α^* (on note parfois le α corrigé de cette manière) = 0,05/3 = 0,0167. Cela revient presque au même de ne pas faire une ANOVA mais de réaliser 3 t-tests en prenant α * comme seuil de significativité. J'ai écrit « **presque** » car ce n'est pas tout à fait exact ; les calculs ne sont pas les mêmes selon que l'on fait une ANOVA puis des tests post-hoc ou si l'on fait 3 t-tests, mais cette nuance dépasse largement le cadre de ce cours.

Le test post-hoc le plus utilisé est celui de Bonferroni, c'est aussi le plus conservatif : il suppose que l'on fait toujours toutes les comparaisons 2 à 2 possibles. Dans la pratique, certaines comparaisons peuvent ne pas nous intéresser, et on a alors le « droit » de ne diviser α que par le nombre de comparaisons effectivement réalisées.

Sur la Figure 11 on a représenté les moyennes des EVA et les intervalles de confiance correspondant à un exemple fictif avec séparation des individus en 3 groupes. On observe que le modèle total conduit à un petit p de 0,01 donc largement statistiquement significatif. Le test de Bonferroni permet de conclure sur les paires de moyennes statistiquement différentes l'une de l'autre. Ainsi : l'EVA moyenne des végétariens est statistiquement différente de la moyenne des EVA chez les omnivores, et l'EVA moyenne des végans est statistiquement différente de la moyenne des EVA chez les omnivores. En revanche les EVA moyennes des végans et des végétariens ne sont pas statistiquement différentes l'une de l'autre.

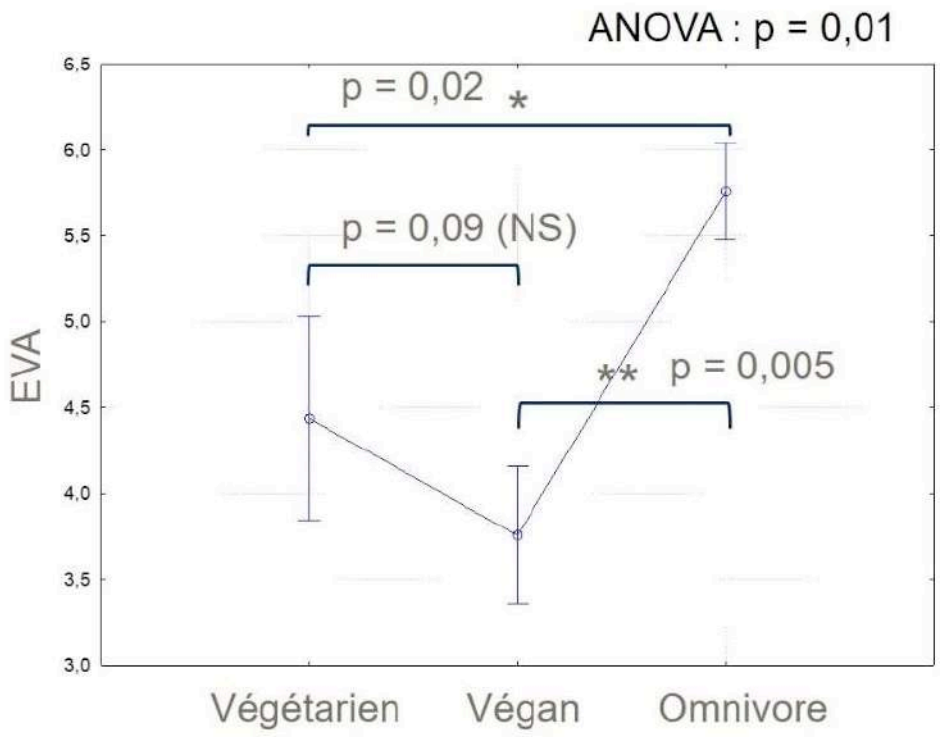


Figure 11 : Moyennes et intervalles de confiance à 95% des EVA catégorisées par le régime à 3 modalités. Le test global est statistiquement significatif avec un petit p de 0,01. L'analyse post-hoc (Bonferroni) montre que la moyenne des EVA des végétariens est statistiquement différente de la moyenne des EVA chez les omnivores, et l'EVA moyenne des végans est statistiquement différente de la moyenne des EVA chez les omnivores. En revanche les EVA moyennes des végans et des végétariens ne sont pas statistiquement différentes l'une de l'autre.

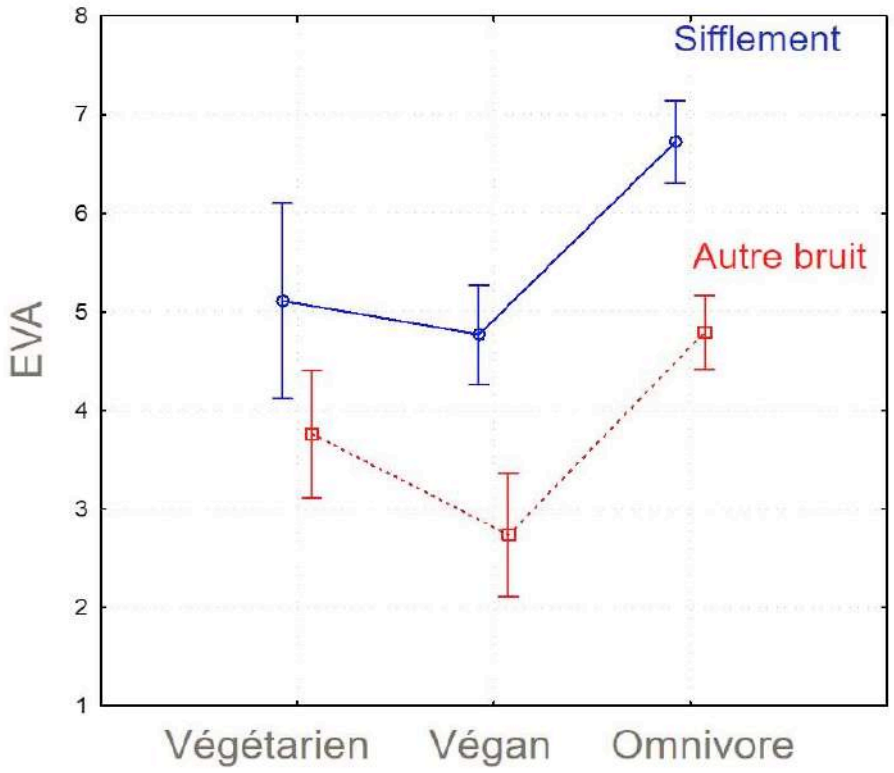
Dans la terminologie utilisée en statistique, la variable « régime alimentaire » est appelée un **facteur** et ce facteur peut prendre 3 **niveaux**, qui sont les 3 catégories « omnivore/végétarien/végan ». Le facteur est la variable qui sert à catégoriser les individus. Un facteur peut bien sûr avoir plus de 3 niveaux (quid des flexitariens ?...). Nous avons traité ici le cas d'une ANOVA à 1 facteur. Comme ce test permet de comparer 3 moyennes ou plus, on dit qu'il généralise le t-test à 3 moyennes et plus.

ANOVA à deux facteurs

Il peut y avoir plusieurs variables de catégorisation, et on réalise alors une ANOVA à plusieurs facteurs. Par exemple, dans notre exemple, imaginons que nous collections une deuxième variable de classement : le type de son de l'acouphène dominant. Selon cette deuxième variable,

les personnes sont par exemple classées en 2 catégories, selon que le son perçu est un sifflement ou un autre bruit. Cette variable a donc 2 niveaux, et l'on peut étudier dans quelle mesure la moyenne des EVA des personnes qui entendent un sifflement diffère de celle des individus qui entendent un autre type de bruit. Si on mixe avec le premier facteur de catégorisation (le régime alimentaire), on peut étudier 3 choses : l'impact du régime, l'impact du type de son, et l'impact d'un facteur croisé, ou **interaction**, entre le régime et le type de son.

Par curiosité, j'ai construit un jeu de données menant à la Figure 12



Sur cette figure, on constate que l'effet du régime sur la moyenne de l'EVA perdure après avoir « ajusté » sur le type de bruit : que l'on soit sur la courbe bleue (« sifflement ») ou sur la courbe rouge (« autre bruit »), on retrouve les recouvrements et non-recouvrements de la Figure 11 des intervalles de confiance (il faudrait refaire les tests post-hoc). A régime donné, il semble qu'il n'y ait pas d'effet significatif du type de bruit pour les végétariens (recouvrement des IC). Pour les deux autres régimes, les IC étant disjoints, on a certainement un effet statistiquement significatif du type de bruit : les moyennes d'EVA sont plus élevées pour les sujets qui entendent un « sifflement » que pour ceux qui entendent un « autre bruit », que ce soit pour les végans ou les omnivores. Enfin, sur ce genre de graphe, on peut commenter l'absence d'interaction entre les deux facteurs « régime » et « type de bruit ». En effet, les deux courbes (bleue et rouge) ne se croisent pas, elles restent d'une certaine manière « parallèles », ce qui signifie que le signe de l'effet du type de bruit sur le régime est le même entre « végétarien et végan » et entre « végan et omnivore ». Les logiciels de statistique attribuent un petit p à l'effet d'interaction noté « type de bruit * régime », à condition qu'il y ait assez de sujets pour l'estimer.

V. Tests sur mesures appariées, répétées, sujets comparés à eux-mêmes

Revenons à notre exemple de départ : nous collectons des EVA chez un ensemble de sujets, sans se préoccuper de leur provenance. Supposons que nous les interrogions sur leur EVA de gêne à un instant t, que nous appelons T_0, et supposons que nous les revoyions tous après 3 mois, et

collectons de nouveau leurs EVA, à un instant cette fois-ci désigné par M_3.

La question de statistique inférentielle posée est alors : la moyenne des EVA à T_0 est-elle différente de celle des EVA à M_3 ? Dans ce cas, nous réalisons un t-test pour mesures **appariées**. Et l'on dit que les sujets sont comparés à eux-mêmes, car les EVA sont mesurées en deux instants, chez les mêmes sujets (il y a deux mesures par sujet). D'un point de vue calcul, la distinction entre mesures appariées et mesures qui ne le sont pas (on parle de mesures parallèles, comme dans l'exemple précédent) est importante. Sans entrer dans le détail, le fait de préciser que les mesures sont appariées fait intervenir la variance des différences des EVA, alors que lors d'un calcul de t-test pour mesures parallèles ce sont les variances respectives des deux groupes qui interviennent. Or la variance des différences est inférieure à la somme des variances des deux groupes, ce qui conduit au fait que pour des valeurs identiques de moyennes et d'écarts-types, un t-test pour mesures appariées sera « plus facilement » statistiquement significatif qu'un t-test pour mesures parallèles. Il serait donc dommage d'oublier de le spécifier dans le logiciel quand on mène le calcul. Inversement, quelqu'un qui aurait par mégarde coché la case « mesures appariées » pour un t-test sur groupes parallèles pourrait trouver un résultat statistiquement significatif qui ne l'est pas en réalité quand on corrige par « mesures parallèles ».

Remarque sur la condition de normalité : la condition de validité des calculs repose cette fois-ci sur la normalité de la distribution des différences des données, et non sur la distribution de chacune des deux distributions.

Mesures répétées

Si l'on étend le nombre de visites, ici T_0 et M_3, à plus de deux, par exemple avec des mesures à 6 mois (M_6), on parle alors de **mesures répétées**. NB: «Répétées» est donc une généralisation de « appariées », dans notre contexte. Et comme nous l'avons vu précédemment dans le cas de groupes parallèles, la généralisation du t-test à plus de 2 groupes s'appelle une ANOVA donc s'il s'agit de mesures répétées on parlera d'**ANOVA pour mesures répétées**.

Il n'y a pas que le temps qui peut être répété...

Dans ce qui précède, nous avons pris le « temps » comme variable répétée. Les EVA des sujets sont comparées chez les mêmes sujets, au cours du temps. On dit que les sujets sont comparés à eux-mêmes, mais il faut bien comprendre que ce sont les EVA (ou une autre mesure) mesurées à différents instants, chez le même groupe de sujets, qui sont comparées.

Les EVA pourraient bien sûr être comparées à elles-mêmes en fonction d'une autre grandeur que le temps. Par exemple, dans notre domaine, il peut s'agir de la comparaison de plusieurs réglages proposés aux mêmes sujets. S'il ne s'agit que de deux programmes — par exemple avec et sans générateur de bruit (GB) — le test sera un t-test pour mesures appariées. Si nous comparons 3 programmes — sans GB, avec GB « blanc », avec GB « rose » — nous réaliserons une ANOVA pour mesures répétées.

Remarque importante : on ne peut évidemment pas « séparer les effets » (formule consacrée en statistique) si l'on fait varier deux choses en même temps : par exemple mettre un programme P_1 à T_0 et mettre un autre programme P_2 à M_3. En proposant un tel changement, on confond les effets « programme » et « temps », et ce n'est pas un simple t-test qui va permettre de séparer les effets. Il faut que le protocole de l'étude soit pensé différemment, ce qui sort du cadre de cet article.

Mélange de facteurs répétés et facteurs non répétés

Evidement on peut envisager de mixer les deux types de facteurs (ou variables) répétés et non répétés. Par exemple, on peut étudier les EVA en fonction du type de bruit de l'acouphène dominant, et regarder quel est l'effet au cours du temps. Cela implique que pendant la durée de l'étude le son de l'acouphène dominant ne change pas ; c'est soit un sifflement, soit un autre bruit. Les sujets peuvent être vus à 3 instants sur 6 mois : T_0, M_3 et M_6. Lorsque l'on paramètre l'ANOVA dans le logiciel de statistique on doit spécifier que le facteur « temps » a 3 modalités et qu'il est répété, et que le facteur « type de bruit » a 2 modalités et qu'il n'est pas répété. Les logiciels s'attendent à ce que les données soient « rangées » d'une certaine manière, non détaillée ici.

Dans certains cas, le caractère répété ou non répété n'est pas évident, il dépend de la manière dont est conçue l'expérience et il faut le préciser. Le temps est forcément répété, mais dans le cas de deux réglages par exemple, on pourrait soit attribuer un réglage A à un groupe de sujets et un réglage B à un deuxième groupe de sujets, soit faire essayer les deux programmes aux mêmes sujets (ce que l'on fait bien sûr chaque fois que c'est possible). Certains facteurs sont structurellement non répétés, ce sont souvent des caractéristiques propres aux patients, comme le sexe, l'âge, le degré de la perte auditive, etc. En toute rigueur on devrait distinguer les effets des traitements appliqués des autres effets. Un traitement est quelque chose que l'on contrôle et que l'on peut modifier par l'expérience, les données démographiques comme l'âge ou le sexe ne sont pas des données imposées lors de l'expérience.

VI. Tests paramétriques et tests non paramétriques

Jusqu'à maintenant, nous avons réalisé des tests statistiques sur des moyennes de variables quantitatives : qu'il s'agisse du t-test ou de l'ANOVA, nous avons comparé des moyennes d'EVA, en fonction de deux (ou plus) conditions, qu'elles soient répétées ou pas.

Faire des calculs de moyennes est licite quand les données respectent certaines conditions : soit elles sont en grand nombre, soit elles ne le sont pas mais leur distribution peut être assimilée à une loi normale. Quand ce n'est pas le cas, et qu'en réalité la moyenne n'a pas de sens (ce n'est pas un bon résumé de nos données), on ne fait pas de calcul à partir de la moyenne, mais à partir d'une transformation des données. Je ne parlerai pas ici des transformations comme le logarithme ou l'arc sinus, en réalité très peu utilisées par les statisticiens.

Quand on ne peut pas travailler sur la moyenne des données, on travaille sur la moyenne des **rangs** des données. Transformer les données en leurs rangs consiste à attribuer un ordre aux valeurs contenues dans la série. Donc cela s'applique bien évidemment à des variables qui sont au moins ordinales voire quantitatives. Par exemple: une EVA (bien sûr), mais également des réponses ordonnées à un questionnaire, un temps de port découpé en classes, etc.

Reprenons notre tableau du début, et intéressons-nous aux EVA de gêne, par exemple sur 15 sujets. Pour attribuer les rangs, il faut trier les sujets en fonction de la variable qui nous intéresse (ici l'EVA). La première opération est un tri des sujets par ordre croissant d'EVA. La nouvelle colonne « Rang » contient les numéros d'ordre correspondants. Il faut noter la gestion des ex-aequo ou « ties » en anglais : quand deux sujets ont la même EVA, on leur attribue le même rang, qui est la moyenne des deux rangs qu'on leur attribuerait « sans réfléchir ». Par exemple, les sujets n°2 et n°10 arrivent en 2e et 3e positions avec la même valeur, donc au lieu d'avoir les rangs 2 et 3, on leur attribue le même rang moyen, qui vaut donc (2+3)/2 = 2,5. Idem, pour les 3 sujets n°4, n°7 et n°12 : ils ont tous les trois la même EVA, qui vaut 4. Ils arrivent avec les rangs 5, 6, et 7, mais comme il n'y a aucune raison pour qu'ils aient un rang différent, on leur attribue à tous la moyenne de 5, 6, et 7, qui vaut 6. La succession de ces opérations est reprise sur la Figure 13 :

Code Sujet	EVA Gêne
S1	9,5
S2	1,5
S3	2,5
S4	4
S5	9,5
S6	8,5
S7	4
S8	6,25
S9	9,25
S10	1,5
S11	7,5
S12	4
S13	4,5
S14	7
S15	1

Code Sujet	EVA Gêne	Tri	Rang
S15	1	1	1
S2	1,5	2	2,5
S10	1,5	3	2,5
S3	2,5	4	4
S4	4	5	6
S7	4	6	6
S12	4	7	6
S13	4,5	8	8
S8	6,25	9	9
S14	7	10	10
S11	7,5	11	11
S6	8,5	12	12
S9	9,25	13	13
S1	9,5	14	14,5
S5	9,5	15	14,5

Code Sujet	Rang
S1	14,5
S2	2,5
S3	4
S4	6
S5	14,5
S6	12
S7	6
S8	9
S9	13
S10	2,5
S11	11
S12	6
S13	8
S14	10
S15	1

Figure 13 : Etapes de la transformation en rangs sur un exemple de 15 observations

Après avoir transformé les valeurs d'EVA en rangs, c'est sur eux qu'on va travailler, à la place des valeurs de départ. On dit alors qu'on réalise des **tests non paramétriques**, car on ne fait plus d'hypothèse sur la distribution des données. En particulier, on n'a plus besoin d'étudier la normalité de la distribution. Les tests non paramétriques sont pour cette raison également dénommés « **distribution free** » en anglais. Comme on a transformé les données en rangs, on a perdu de l'information : on sait seulement que tel sujet a une EVA plus grande ou plus petite que tel autre sujet, mais on ne sait pas de combien elle plus grande ou plus petite.

Cette manipulation est transparente pour l'utilisateur qui sélectionne un test non paramétrique, c'est le logiciel de statistique qui fait ce recodage. Les tests non paramétriques font exactement tout ce que nous avons vu précédemment sur le t-test et l'ANOVA, mais à partir des rangs et non à partir des données brutes. Voici les noms des équivalents non paramétriques des tests principaux abordés dans cet article :

- t-test pour groupes parallèles → test de **Mann-Whitney** (ou test des rangs signés)
- t-test pour groupes appariés → test de **Wilcoxon**
- ANOVA à un facteur non répété → **ANOVA de Kruskal-Wallis**
- ANOVA à un facteur répété → **ANOVA de Friedman**

A quelle question répond-on avec un test non paramétrique ?

Même si on ne fait jamais le calcul des rangs « à la main », il est intéressant de comprendre ce que fait le logiciel quand on classe les individus selon les valeurs d'une variable. Sur notre exemple, il s'agit de l'EVA. Si l'on compare deux groupes de sujets indépendants, par exemple les végétariens et les omnivores, il faut d'abord affecter les rangs, indépendamment de cette variable de catégorisation, autrement dit tous régimes confondus. Imaginons un effectif total de 30 sujets : 15 végétariens et 15 omnivores. Il faut d'abord classer les 30 sujets, donc attribuer les rangs de 1 à 30. Ensuite, le test (de Mann Whitney) va comparer les moyennes des rangs dans chacun des groupes « végétarien » et « omnivore ». Ainsi, il permettra de dire si la moyenne des rangs des EVA des « végétariens » et plus élevée ou plus faible que celle des rangs des EVA « omnivores ». Comme, en passant aux rangs, on n'a gardé que la notion d'ordre, on saura alors répondre à la question suivante sur les EVA elles-mêmes : « Y a-t-il plus de valeurs élevées ou plus de valeurs faibles chez les omnivores que chez les végans ? ».

On note donc que cette question est moins « ambitieuse » que celle posée lorsqu'on peut faire un test paramétrique. C'est normal, ou plutôt logique, voire réconfortant : on est moins exigeant sur les données, en échange on peut en dire quelque chose de plus vague...

Tests non paramétriques, inférence statistique et représentations graphiques associées

Comme pour les tests paramétriques, on peut associer un indice de crédibilité (un petit p) à un test non paramétrique. Le raisonnement est identique : nous formulons implicitement une hypothèse nulle H_0 du genre « Il n'y a pas de différence entre ...et ... ». Par exemple, dans le cas d'un test de Mann-Whitney, H_0 correspond à : « Il n'y a pas de différence de moyenne des rangs de la variable entre les deux groupes comparés ». Comme on n'énonce en général pas de résultat sur les rangs, H_0 peut se formuler par : « les valeurs ne sont pas plus grandes ou plus petites dans l'un des deux groupes ».

On fera particulièrement attention aux représentations graphiques associées aux tests non paramétriques. Trop souvent, on voit des moyennes et un petit p issu d'un test non paramétrique comme Wilcoxon ou Mann-Whitney sur le même graphique. Or, comme ces tests ne comparent pas les moyennes des valeurs, illustrer par des moyennes n'a pas de sens, et, pire, cela peut signifier que l'auteur ne comprend pas le test statistique qu'il a utilisé. Pour un test de Wilcoxon ou Mann-Whitney, il faut utiliser une représentation graphique avec des indicateurs de position. Il s'agit de la médiane, des quartiles, du min, du max... Ces indicateurs donnent des informations sur les valeurs qui séparent les sujets : en deux effectifs égaux pour la médiane, en 4 pour les quartiles. Ces indicateurs peuvent être repris dans une boîte à moustaches, comme celle de Figure 14.

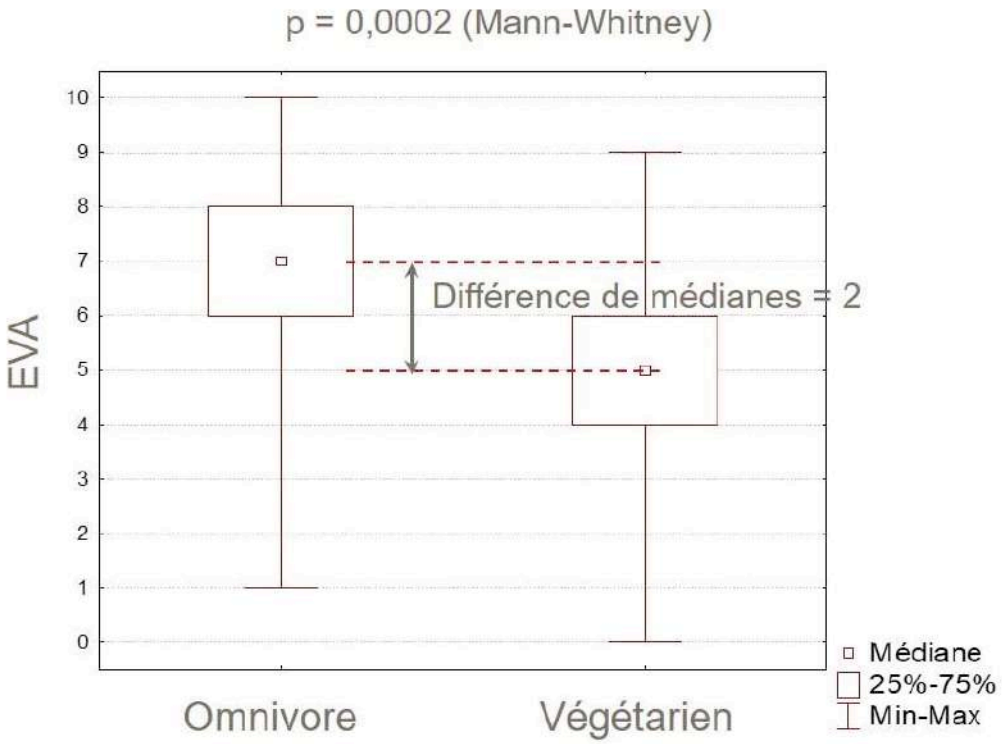


Figure 14 : Illustration d'un test de Mann-Whitney avec des boîtes à moustaches représentant la médiane, les quartiles à 25% et 75%, le min et le max des deux séries.

VII. Etude du lien entre deux variables

Jusqu'à maintenant, nous avons parlé des différences entre des moyennes (ou des moyennes de rangs), calculées sur une même variable. Par exemple, nous avons parlé de la différence entre la moyenne des EVA dans un groupe de sujets et la moyenne des EVA dans un autre groupe de sujets. Ce qui définit le groupe est une variable de classement ou de catégorisation, par exemple le régime alimentaire, à 2 modalités.

Nous allons maintenant nous intéresser à la ressemblance qui peut exister entre deux variables distinctes. Par exemple : entre un score THI et une EVA, entre une EVA de gêne et une EVA d'intensité, entre un score d'intelligibilité avec réducteur de bruit et un score d'intelligibilité sans réducteur de bruit, etc. Il s'agit alors d'étudier si, quand l'une des variables croît, l'autre croît aussi, ou au contraire décroît ; est- ce que si cette tendance est avérée, on peut être plus précis ? S'il s'agit de deux variables quantitatives, quelle est la forme de la courbe qui relie les deux ? Est-ce qu'elle ressemble à une droite ? Que faire si les deux variables ne sont pas des variables quantitatives ?

Corrélation entre deux variables quantitatives et droite des moindres carrés

Revenons aux sujets acouphéniques, et considérons qu'ils sont évalués sur deux échelles : leur « score THI » (note entre 0 et 100) et l'EVA de gêne (note entre 0 et 10). On peut voir cela comme deux « dimensions », et dire que nos sujets sont caractérisés par deux coordonnées : leur THI et leur EVA. Evidemment beaucoup plus de dimensions les caractérisent, mais nous choisissons de projeter les individus dans cet espace en 2D.

Naturellement, on peut donc représenter le « nuage de points » correspondant, où chaque sujet est représenté par son abscisse (EVA ou score THI) et son ordonnée (la deuxième variable). On a représenté sur la Figure 15 un exemple (toujours fictif) d'un tel nuage de points. Chaque cercle représente un individu, qui est caractérisé par un couple de coordonnées (EVA, score THI). Par exemple, on retrouve l'individu n°154, déjà étiqueté sur le nuage simple de la Figure 1, qui a une EVA de 9,1 et un score THI de 72.

Quand on regarde ce nuage de points, on peut noter plusieurs choses :

- Il y a une tendance : les gens qui ont des EVA élevées (respectivement faibles) ont plutôt des scores THI également élevés (respectivement faibles).

- À score THI donné (ou à EVA donnée), il peut y avoir plusieurs valeurs possibles d'EVA (respectivement de score THI). Par exemple, imaginez une ligne horizontale qui passerait par la valeur de THI = 40 : il y a de nombreux sujets sur cette ligne, qui correspondent à autant de valeurs d'EVA différentes. Idem, si l'on se place sur une droite verticale passant par une valeur d'EVA donnée, par exemple EVA = 4, il existe une grande étendue de valeurs possibles pour le THI : de 10 à 60 environ.

- Il existe plusieurs personnes qui ne suivent pas la tendance générale : elles ont par exemple un THI bas et une EVA élevée, ou inversement.

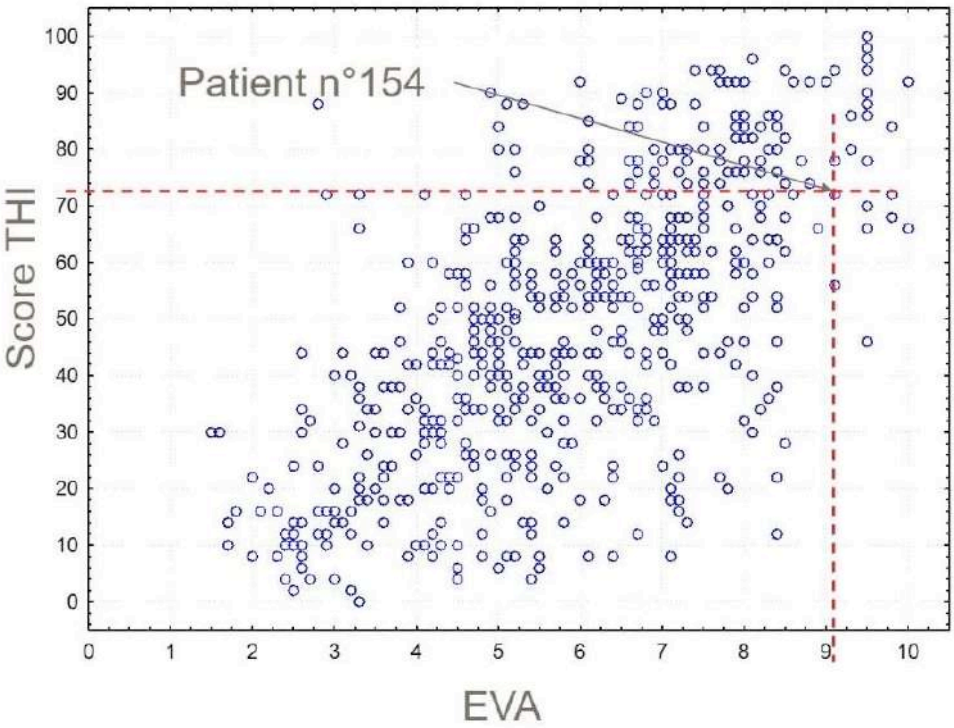


Figure 15 : Nuage de points en deux dimensions, où chaque individu est représenté par deux coordonnées : son EVA en abscisse et son score THI en ordonnée. Pour exemple, on a repéré l'individu 154, déjà étiqueté sur le nuage de points simple de la Figure 1.

L'observation n°1 va conduire beaucoup de personnes à se poser la question très saugrenue : « Est-ce que le score THI et l'EVA sont reliés par une droite ? ». Bien évidemment non, puisque le nuage est un vrai nuage, et non une droite parfaite. Mais on peut se demander dans quelle mesure une droite approche bien ce nuage.

Quelle est cette droite ? Elle s'appelle la droite des moindres carrés. Pour établir son équation, on (le logiciel) cherche le meilleur couple (a, b) tel que : « THI = a + b*EVA ». On montre en mathématiques que ce couple est unique ; c'est celui qui minimise l'erreur quadratique totale (somme de toutes les différences au carré entre les valeurs données par la droite et les valeurs réellement observées dans notre échantillon). Le logiciel va donc trouver la droite optimale au sens de ce critère, celle qui passe « le plus près possible de tous les points du nuage », et donner les valeurs de a et b correspondantes. Sur l'exemple de la Figure 15, l'équation donnée par le logiciel est THI = – 1,4 + 8,46 *EVA. Elle apparaît en rouge sur le graphique suivant (Figure 16).

Remarque importante : à ce stade, je n'ai fait aucune inférence statistique. Nous faisons des statistiques descriptives. Quelles que soient mes données, j'ai le « droit » de calculer l'équation de la droite des moindres carrés. En particulier, mes données n'ont pas à suivre une distribution normale ou autre, je n'ai aucune hypothèse à faire.

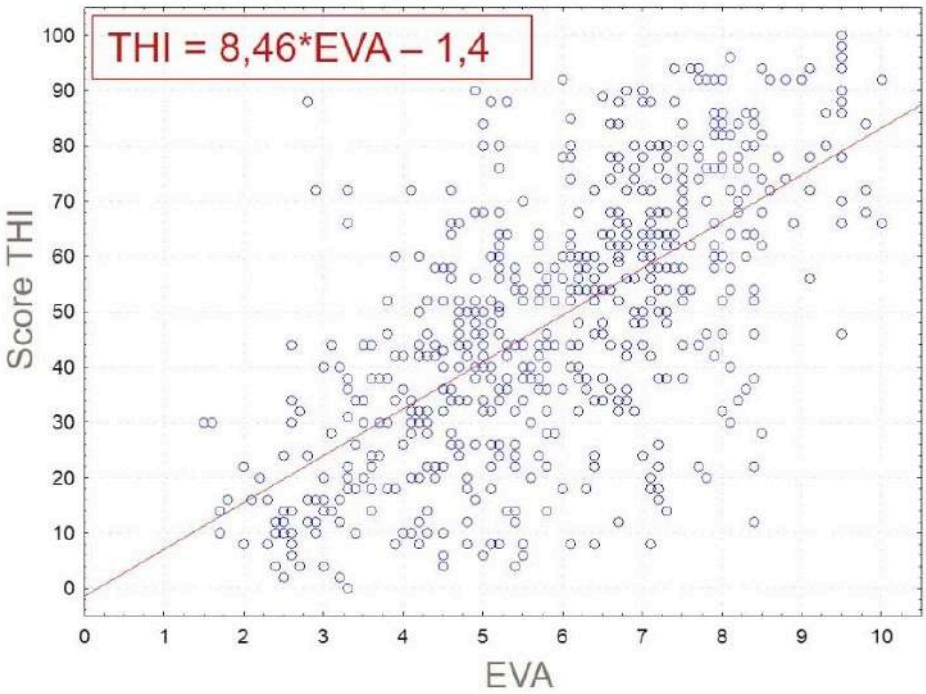


Figure 16 : Nuage de points avec interpolation par la droite des moindres carrés (en rouge), dont l'équation est « Score THI = – 1,4 + 8,46 *EVA »

On parle aussi « d'ajustement linéaire » ou de régression linéaire simple. Ce terme, par opposition à « multiple », précise que le score THI n'est relié qu'à une seule autre variable : l'EVA. Dans cette terminologie, le score THI est appelé « variable à expliquer » (ou « dépendante »), et l'EVA est appelée « variable explicative » ou « indépendante ». Attention à ce terme qui peut prêter à confusion, tout comme « à expliquer » et « explicative », car nous n'avons aucun moyen de dire qui explique quoi (relation causale) ; nous statuons seulement sur une association. On pourrait d'ailleurs tout à fait tracer l'EVA en fonction du score THI (je vous laisse inverser l'équation de la droite...). C'est le protocole, si c'est possible, qui peut aider à répondre à la question de la causalité, et non la mise en évidence mathématique de la liaison.

On peut s'interroger sur la pertinence de cette droite : pourquoi est-ce que l'EVA et le THI seraient liés par une relation aussi forte ? Evidemment, il s'agit d'un « modèle », et faire coller un modèle va permettre de tirer de l'information. On a l'habitude de quantifier la force de la relation linéaire par le coefficient de détermination R^2. Pour comprendre comment il est construit, et son lien avec la qualité d'ajustement du modèle linéaire, il faut introduire quelques notations sur notre exemple. Soient :

- la moyenne de tous les THI de notre échantillon
- le score THI observé dans notre échantillon pour l'individu n°i
- e score THI de l'individu i calculé à partir de la droite des moindres carrés

Alors on peut écrire la formule de décomposition suivante : Petit rappel de terminale : le signe S désigne la somme sur tous les individus, de i = 1 à n (si notre échantillon comporte 600 sujets, n = 600). A partir de cette formule on peut introduire le coefficient de détermination, qui est le

rapport entre la somme des carrés totale et la somme des carrés expliqués (par la droite) : R^2 quantifie la part des variations de y expliquées par les variations de x. Il est compris entre 0 et 1, et s'exprime souvent en %. Par exemple : R^2 = 0,64 signifie que 64% des variations de y sont expliquées par les variations de x. En particulier (cf. Figure 17) :

. - R^2 = 1 signifie que tous les points sont alignés, le modèle de droite est parfait, il explique totalement les variations de y (score THI) par les variations de x (EVA).

. - R^2 = 0 signifie que la droite n'a aucun intérêt, elle n'explique aucune des variations de y par celles de x (ou inversement).

$$\underbrace{\sum(y_i-\bar{y})^2}_{\text{Somme des carrés totale}} = \underbrace{\sum(\hat{y}_i-\bar{y})^2}_{\text{Somme des carrés expliquée}} + \underbrace{\sum e_i^2}_{\text{Somme des carrés résiduelle}}$$

$$R^2 = \frac{\textit{Somme des carrés expliquée}}{\textit{Somme des carrés totale}} = \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}$$

$$\boxed{0 \leq R^2 \leq 1}$$

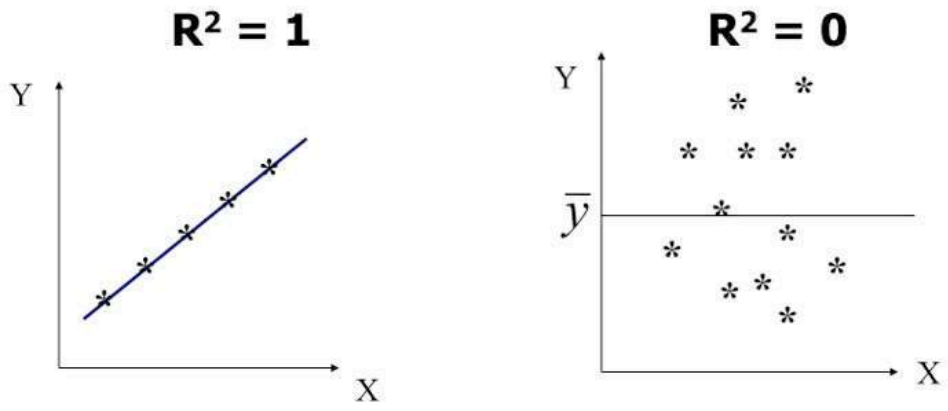


Figure 17 : Deux nuages de points extrêmes, montrant une relation linéaire parfaite (gauche) et une absence totale de relation linéaire (droite)

R est quant à lui le coefficient de corrélation linéaire de Pearson. On peut l'obtenir par une autre formule, qui fait intervenir la covariance entre x et y et les variances respectives de x et de y (pas détaillé ici).

R a le même signe que la pente de la droite des moindres carrés : si cette pente est positive (relation monotone croissante), les EVA et les THI sont corrélés positivement ; si la pente est négative, cela signifie que les EVA et les THI sont corrélés négativement. Evidemment, dans notre exemple, EVA et THI évoluent dans le même sens : la corrélation est positive. Il est courant (mais empirique) de considérer les valeurs suivantes pour quantifier l'importance de la liaison linéaire entre x et y :

Valeur de R	Valeur de R2	Importance de la liaison linéaire	Pourcentage de variance expliquée
0 – 0,2	Très faible	0 – 0,04	0 – 4%
0,2 – 0,4	Faible	0,04 – 0,16	4% – 16%
0,4 – 0,7	Moyenne	0,16 – 0,49	16% – 49%
0,7 – 0,9	Forte	0,49 – 0,81	49% – 81%
0,9 – 1	Très forte	0,81 – 1	81% – 100%

Tableau 3 : Valeurs empiriques communément admises pour quantifier l'importance de la liaison linéaire entre deux variables quantitatives, à partir du coefficient de corrélation linéaire R (Pearson)

Il faudra donc toujours être prudent, avec ce nombre R qui, parce qu'il est plus petit que 1 en valeur absolue, va être encore plus petit que 1 quand on va l'élever au carré...

Remarques : ces valeurs ne sont pas gravées dans le marbre, il s'agit simplement d'un petit guide pour se situer. Il n'est cependant pas raisonnable, comme on l'entend parfois, de « s'arranger avec ses résultats » en disant que dans un domaine on peut considérer telle valeur de R comme une corrélation forte, alors que dans un autre domaine elle serait plutôt faible...

Revenons à notre exemple des THI et des EVA et la Figure 16 : le logiciel fournit, en plus de l'équation de la droite des moindres carrés, la valeur du R (et donc du R^2) : R = 0,65 et R^2 = 0,43. On peut dire que « ce n'est pas mal » : la droite explique une petite moitié des variations de THI par les variations des EVA. Mais encore ? Je dirais : HEUREUSEMENT ! Cela signifie que toute l'information n'est pas contenue dans une seule variable. Si la droite était parfaite, à quoi bon faire une mesure de THI en plus de celle de l'EVA ? Sans compter que ce serait très suspect, non ? Au contraire, ce résultat nous dit que nous avons raison d'évaluer les gens sur deux dimensions, car elles apportent chacune de l'information qui n'est pas apportée par l'autre. La part de variance du THI non expliquée par l'EVA vient d'autre chose, mais de quoi ? Ce n'est pas le propos de cet article, mais il faudrait sûrement introduire d'autres variables. C'est là qu'arrivent les modèles « multivariés » mentionnés plus haut sous le terme de régression multiple. Au lieu de regarder y en fonction de x, on regarde y en fonction de la combinaison linéaire d'une variable x_1, d'une autre variable x_2....

Régression linéaire et petit p

Jusque-là, nous avons tenu sans faire de statistique inférentielle, c'est-à-dire sans parler de petit p... Ne soyez pas inquiet, j'y viens. Pour calculer un petit p, il va falloir faire des hypothèses sur le terme d'erreur entre l'estimation du THI (par le modèle de la droite) et les valeurs réellement observées de THI dans l'échantillon. Cette erreur, qui s'appelle aussi « résidu », est souvent notée epsilon (e) ; le modèle s'écrit alors, pour un individu i : $THI_i = a + b*EVA_i + e_i$

Nous n'allons pas passer trop de temps là-dessus, mais l'hypothèse qui rend les calculs de petit p licites est que les (e_i) suivent une loi normale (centrée sur 0). Normalement, il faudrait toujours vérifier cette hypothèse après avoir réalisé une régression linéaire (les logiciels de statistiques proposent systématiquement un menu de vérification d'hypothèses sur les résidus).

Cependant, dans la pratique, on se contente souvent de regarder si la distribution des EVA et celle des scores THI sont normales. Sous cette condition, nous pouvons faire de l'inférence et calculer un petit p. En fait, on peut calculer plusieurs petit p : celui qui correspond à l'hypothèse H_0 formulée sur le coefficient a de la droite de régression, celui qui correspond à H_0 formulée sur le coefficient b, et celui sur R. Nous nous concentrons sur le petit p associé à R, puisque c'est lui que l'on voit partout...

Le petit p associé au coefficient de corrélation R vient du test d'hypothèses suivant :

- $H_0 : R = 0$
- $H_1 : R \neq 0$

Il est très important de réaliser que le petit p associé à un coefficient de corrélation est basé sur CE test d'hypothèse. AInsi, un petit $p < 0,05$ permet UNIQUEMENT de conclure que R est statistiquement différent de 0.

On peut montrer que :

$$R \text{ significatif au seuil de } 0,05 \Leftrightarrow |R| \geq 2 / \sqrt{(n+2)}$$

Conséquence très importante: dès que n est «grand», une petite valeur de R peut devenir statistiquement significative... Exemple : pour n = 34 la quantité **$2/\sqrt{(n+/2)}$** soit **$2/\sqrt{36}$** où 2/6 = 0,33.Toute valeur de R supérieure ou égale à 0,33 va donc être statistiquement significative. Or R 0,33 donne un R2 d'environ 0,1 soit une explication d'environ 10% des variations d'une variable par l'autre. Autant dire que le modèle linéaire, dans ce cas, n'a pas grand intérêt... Je vous laisse faire les calculs pour des n plus petits ou plus grands. Décidément, dans les grands échantillons, tout devient statistiquement significatif... Nouvelle pause pour méditer... Nous arrivons à une autre remarque intéressante, qui doit encore vous convaincre de bien regarder vos nuages de points, d'être critiques devant certaines valeurs de R et certaines droites de régression, ou encore quand l'auteur ne montre pas le nuage de points associé. La Figure 18 représente une variante du « quartet d'Anscombe », statisticien ayant publié pour la première fois en 1973 quatre nuage de points ayant les « mêmes statistiques » mais n'ayant pas grand-chose à voir visuellement. Cette version-là est postée par la communauté de statisticiens JMP (www.jmp.com/en_us/about.html)

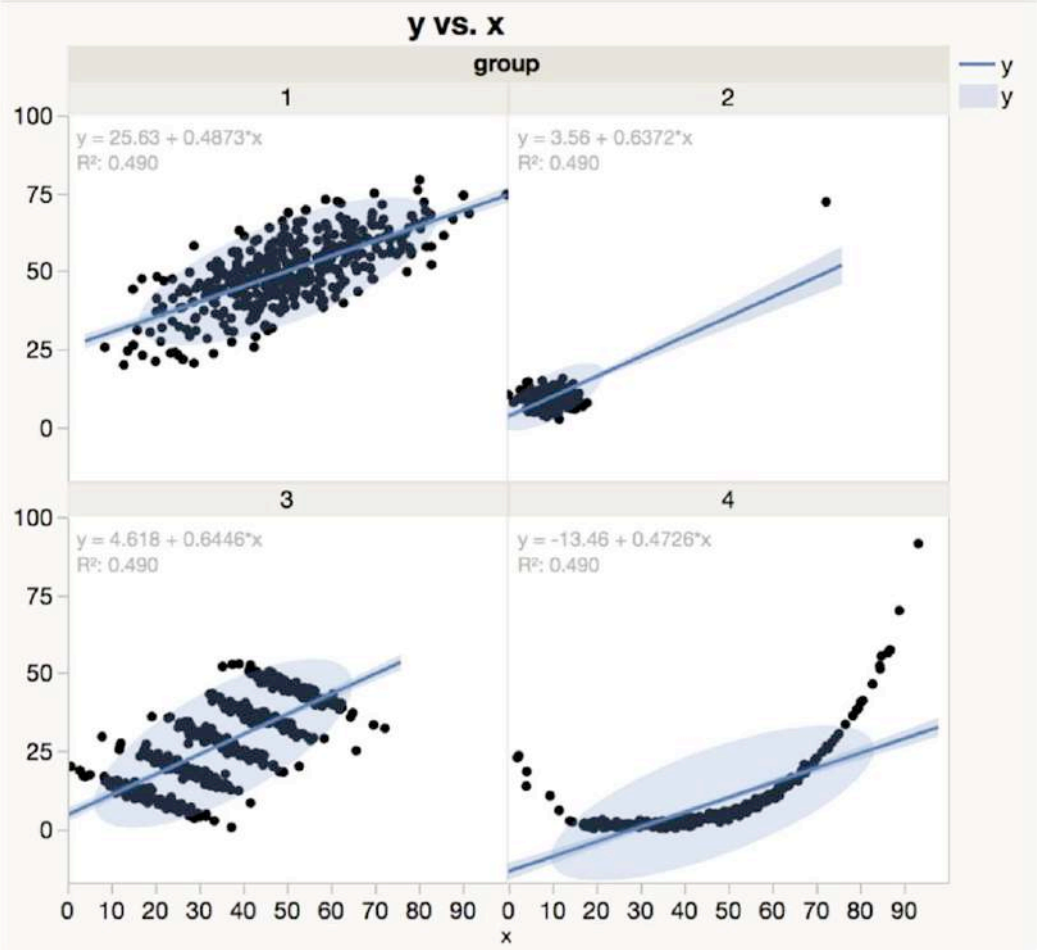


Figure 18 : Variante du quartet d'Anscombe, montrant 4 nuages de points très différents, cependant modélisés par la même droite des moindres carrés et le même coefficient de détermination R2 (corrélation linéaire « moyenne ») Le nuage 1 en haut à gauche est le seul qui mérite qu'on y fasse passer une droite. Le nuage 2 montre un point « levier », qui tire à lui seul toute la droite. Il serait intéressant de refaire le calcul de la droite sans ce point... Le nuage 3 montre une sous-structure évidente qui va dans le sens opposé de la pente de la droite (il faut certainement regarder de plus près la sous-structure). Malheureusement, un calcul « bête » sur l'ensemble des points du nuage mène inéluctablement à cette belle droite. Enfin, le nuage 4 montre une relation entre x et y qui n'est certainement pas linéaire ; la droite est très peu informative de la véritable association entre les deux variables.

Corrélation sur les rangs : le R de Spearman

Lorsque les résidus du modèle linéaire ne suivent pas une loi normale, on ne peut pas faire d'inférence sur le coefficient de corrélation linéaire. On a alors deux choix : soit les résidus suivent une autre loi que la loi normale et présentent une structure évidente, auquel cas on peut transformer les données pour faire apparaître cette structure et faire une régression linéaire sur les données ainsi transformées. L'autre possibilité est de ne faire aucune hypothèse, et de travailler sur les rangs des données. Le coefficient de corrélation sur les rangs s'appelle le R de Spearman.

Première remarque importante : accompagner un R de Spearman avec une belle droite de régression passant par le nuage de points n'est pas cohérent. Si on a calculé un R de Spearman, c'est justement parce que la droite n'est pas un bon modèle. C'est exactement la même erreur que

lorsqu'on illustre le résultat d'un test de Wilcoxon avec des moyennes...

Deuxième remarque : transformer les données en rangs permet d'explorer une relation non linéaire, mais il faut qu'elle reste « monotone ». Par exemple dans le cas d'une relation quadratique entre x et y ($y = x^2$), le R de Pearson vaut 0, et le R de Spearman vaut tout autant 0. En effet, à x et (-x) donnés correspondent la même valeur de y. Cette relation n'est pas monotone, il y a pour chaque couple (x ; - x) le même rang moyen. Donc il n'y aura pas plus de relation linéaire entre les rangs des données qu'entre les données elles-mêmes.

On note parfois le R de Spearman par la lettre grecque « Rho » ρ: r, pour le différencier du R de Pearson. De manière générale, il faudra se méfier des cas où le r de Spearman et le R de Pearson sont très différents. En général, leurs valeurs sont proches, mais ce sont les conditions de calcul du petit p qui sont différentes. En effet, nous pouvons calculer un petit p associé à un r de Spearman. Il est de nouveau issu d'un test d'hypothèses :

- $H_0 : r = 0$
- $H_1 : r \neq 0$

Et comme nous avons perdu la « puissance » donnée par les tests paramétriques, r est souvent moins facilement statistiquement significatif qu'un R (de Pearson) calculé sur les mêmes données (nous retrouvons des résultats similaires au test non paramétrique de Mann Whitney). Autrement dit : il faut souvent un peu plus d'individus pour atteindre la même significativité statistique sur r que sur R. N'attendons donc pas de miracle de ce côté-là, mais gardons en tête que c'est le bon test à faire si nous voulons associer un petit p et que le modèle linéaire n'est pas adapté à notre nuage de points.

Données qualitatives et autres mesures d'association non paramétriques : le test du Khi-2

La grande star de ce paragraphe s'appelle Khi-deux, « Chi Square » en anglais, noté avec la lettre grecque : $\chi 2$. (NDE : On retrouve parfois dans la littérature chi-2). Dès que l'on cherche à savoir s'il existe une liaison entre deux variables catégorielles (indépendantes), on réalise un test du $\chi 2$. Notons que ce test ne s'applique pas quand les variables sont dépendantes, ou appariées, par exemple quand on collecte une variable catégorielle chez les mêmes sujets dans deux conditions différentes (par exemple avant/après traitement).

Donc, lorsque deux variables sont catégorielles et que l'on cherche à savoir si elles sont liées, on réalise un test du Khi-deux. Il faut avant cela introduire ce qu'on appelle un tableau de contingence, ou tableau croisé (cf. tableau croisé dynamique sous Excel), « cross-tabulation » en anglais. Nos individus étant caractérisés, entre autres, par deux variables catégorielles, on compte le nombre d'individus qui ont telle catégorie de la première variable et telle catégorie de la deuxième variable.

Par exemple, supposons que l'on connaisse, pour nos sujets acouphéniques : le régime alimentaire codé en 3 modalités (végan, végétarien, omnivore) et le signe du zodiaque, donc codé sur 12 modalités. On construit le tableau à 3 x 12 = 36 cases (3 lignes et 12 colonnes ou le contraire), issu du croisement des modalités de chaque variable prises deux à deux : végan et Bélier, végan et Taureau, ..., omnivore et Poissons.

Sur un échantillon de n personnes, on peut se poser la question de savoir s'il y a « indépendance » ou « équi-répartition » entre les deux variables, dans la population TOTALE, c'est-à-dire s'il y a autant de gens de chaque signe qui sont soit végans, soit végétariens, soit omnivores, ou si, au contraire, il y a une tendance à ce que les gens de certains signes du zodiaque mangent plutôt végan qu'autre chose... Evidemment nous ne saurions pas pourquoi, mais là n'est pas notre propos.

S'il n'y a pas « équi- répartition », on dit que les variables sont « liées », qu'elles ne sont pas indépendantes.

Pour statuer sur la « force de la liaison », on regarde la valeur du Khi-deux qui est calculé à partir du tableau de contingence. Ce calcul fait intervenir les effectifs observés sur l'échantillon et les effectifs théoriques, c'est-à-dire les effectifs attendus si jamais il y a équi-répartition en VRAI. Le Khi-deux donné par le logiciel est assorti d'un nombre de degrés de liberté (« ddl » ou « df » en anglais, pour « degrees of freedom »), qui est égal au (nombre de lignes-1)*(nombre de colonnes-1).

On peut faire de l'inférence sur le khi-deux en l'accompagnant d'un petit p, associé au test

- $H_0 : \chi 2 = 0$

- $H_1 : \chi 2 \neq 0$

On ne va pas donner les détails du calcul ici, mais, en gros : plus le χ2 calculé sur notre échantillon est grand, plus la probabilité d'observer une telle valeur, si en réalité le χ2 était nul sur la population totale, est faible. Encore une fois, les valeurs de petit p petites sont associées à des valeurs de c2 grandes. En pratique, c'est le logiciel qui calcule le nombre de ddl, le χ2 associé et son petit p. Attention à la conclusion d'un tel test : si le petit p est faible (inférieur à 0,05), alors on rejette l'hypothèse nulle selon laquelle χ2 = 0 (hypothèse d'indépendance des deux variables). Dans ce cas, on conclut donc que les deux variables qualitatives sont liées. Normalement, si p > 0,05, on ne sait pas conclure. On ne peut pas rejeter H_0. En pratique, on conclut quand même, simplement parce qu'on ne sait pas faire autre chose, et l'on dit que les deux variables sont indépendantes l'une de l'autre. Dans le cas des sujets acouphéniques dont on connaît le signe du zodiaque et le régime alimentaire : un petit p < 0,05 signifie qu'il n'y a pas indépendance entre le régime et le signe du zodiaque (quel exemple tordu). Ensuite, pour savoir quel signe mange plutôt quoi, il faut regarder les cellules du tableau, et parfois ce n'est pas évident. Un conseil, quand on regarde le croisement de deux variables qualitatives, est de ne pas avoir « trop de catégories » pour chaque variable, d'autant plus si on a peu de sujets (c'est logique). Par exemple, si on était vraiment intéressé par cette question du signe du zodiaque, peut-être que l'on commencerait à dégrossir en découpant en 4 classes et non 12 : signes d'eau, de terre, de feu ou d'air (j'ai demandé à une spécialiste). Cette remarque est valable pour n'importe quel découpage en classes d'une variable quantitative. Par exemple, on transforme souvent l'âge en classes d'âges. Le contexte et les données de l'échantillon doivent nous guider pour définir les bornes des classes et leur nombre. Remarque sur le khi-deux et les facteurs de confusion : on fait souvent un test du khi-deux pour explorer des biais possibles qui pourraient compromettre l'interprétation d'autres analyses. Dans notre exemple sur les acouphéniques dont on étudie le lien entre l'EVA et le régime alimentaire, nous avons montré une différence entre les omnivores et les végans. Or, dans notre échantillon, il y a des femmes et des hommes. Et il se peut que la différence entre omnivores et végans soit en fait une différence entre hommes et femmes. Il faut alors vérifier qu'il n'y a pas de lien entre le régime alimentaire et le sexe, sinon nous ne pouvons simplement pas conclure sur l'impact du régime. Typiquement, dans ce cas, on croisera les deux variables « régime » et « sexe », et la valeur du Khi-deux associé nous conduira à statuer sur la présence d'un biais (par exemple une sur-représentation des végans chez les femmes) ou pas.

VIII. Divers

Le drame du petit p supérieur à 0,05

Bien souvent, les étudiants, parce qu'ils ont aussi été « poussés » à cela, sont désespérés quand leurs résultats ne sont pas statistiquement significatifs. Bien souvent, ils essaient de « manipuler » ou « tordre » les hypothèses de départ pour pouvoir conclure quand même, comme si le petit $p < 0,05$ était le saint Graal. Au contraire ! Il faut savoir commenter, interpréter et utiliser un tel résultat. Il ne faut pas avoir peur de soutenir que c'est aussi un résultat. Rappelons notre démarche scientifique :

- 1) Formuler une hypothèse.
- 2) Construire une expérience afin de l'éprouver.
- 3) Conclure à partir des données récoltées.

C'est le mathématicien et biologiste R.A. Fisher (1890 – 1962) qui a, le premier, introduit la notion du petit p. Il travaillait dans un contexte d'expérimentation appliquée, et réalisait des séries d'expériences. Voici un extrait de l'excellent document, plein d'humour, qui se nomme «Statistiques pour Statophobes » (voir les références bibliographiques) : « Le seuil de a = 0,05 est un simple premier crible. Lorsque $p > 0,05$ (test « non significatif ») Fisher préfère passer à autre chose car il effectue un travail de défrichage, il y a beaucoup d'effets à découvrir et la vie est courte, mais si $p < 0,05$, Fisher ne crie pas sur les toits « Hourrah ! Les amis, j'ai découvert un effet ! ». Il se contente de conclure que ce traitement vaut la peine qu'on s'y intéresse, et lance une série d'expériences pour essayer de répliquer l'effet qu'il a peut-être découvert. C'est seulement lorsqu'on connaît un protocole tel qu'une probabilité de $p <0,05$ est presque systématiquement obtenue répétition après répétition de l'expérience, qu'il s'estime satisfait. Cette attitude est à des années-lumière de la manière moderne d'utiliser les tests. »... Je vous invite à lire la suite, très intéressante, et qui vous fera relativiser l'importance de ce petit p, et vous incitera à ne pas tirer de conclusions ayant un caractère définitif à partir d'une seule valeur de petit p, aussi petite fusse-t-elle...

Ces réflexions sont largement d'actualité, elles constituent notamment le sujet de l'article paru récemment dans la revue Nature, cité à la fin des références bibliographiques (l'extrait vulgarisé dans Sciences et Avenir, en français, en donne un bon aperçu).

Manque de puissance ?

On entend et on peut souvent lire, devant un petit $p > 0,05$: « le test n'était peut-être pas assez puissant ». Certes... La puissance d'un test est la probabilité qu'il mette en évidence un effet (par exemple une différence) s'il existe en vrai (c'est-à-dire dans la population totale). On la note $P = 1 - b$, où b est le « risque de deuxième espèce ».

On va ainsi plus loin que le simple rejet ou non-rejet de H_0 : en cas de rejet de H_0, on accepte H_1 (l'hypothèse alternative), avec une probabilité b de se tromper en le faisant. On fixe β à 10, 15 ou 20% en général, ce qui nous donne respectivement une puissance de 90%, 85% ou 80%. Dans le cas simple d'une différence de moyennes entre deux groupes, quand on a fixé α et β, que l'on connaît la plus petite différence que l'on veut pouvoir montrer statistiquement (on veut pouvoir dire qu'elle n'est pas due au hasard), et que l'on connaît l'allure des distributions de données qu'on va récolter (mais si on le sait pourquoi fait-on une expérience ?!!), alors le logiciel sait calculer le fameux nombre de sujets nécessaire.

Les calculs de nombre de sujets sont compliqués (il s'agit d'inversions de formules à plusieurs variables), et l'on a recours à des tables ou des logiciels. Rien ne sort du chapeau, il faut définir plusieurs hypothèses, qui sont des hypothèses de travail. Par exemple, dans le cas le plus simple d'une différence de moyennes (t-test), il faut définir :

- La plus petite différence de moyennes que l'on souhaite mettre en évidence

- L'écart-type dans chacun des deux groupes

- La valeur de a (en général 0,05, test bilatéral)

- La valeur de la puissance recherchée (au minimum 80%). Si l'on fait une expérience, il faut regarder ce que d'autres expériences « similaires » ont déjà donné. Cela suppose qu'elles aient été publiées et, malheureusement, beaucoup de résultats non statistiquement significatifs ne sont pas publiés (pour la seule raison qu'ils sont non statistiquement significatifs). On ira aussi consulter des sources réputées moins fiables, car non parues dans des revues à comité de lecture, cependant intéressantes à consulter.

Il est toujours informatif et donc recommandé de faire plusieurs scenarii (pessimistes et optimistes), quand on ne sait pas grand-chose (études exploratoires), et de comprendre à l'avance ce que l'on sera capable de montrer si on est contraint dès le départ par un nombre de sujets « faible » en pratique (ce qui est souvent le cas).

Traitement des réponses à des questionnaires

Le traitement des questionnaires (essentiellement de qualité de vie) dépend du type de réponses demandées. Il peut s'agir de :

1. **Réponses quantitatives**. Les questions appellent une réponse continue, par exemple de 0 à 10 (comme une EVA). On peut citer le SSQ, dont chaque réponse doit être reportée sur une règle graduée. L'ERSA présente aussi une échelle de 0 à 10.

2. **Réponses catégorielles ordinales**. C'est la majorité des questionnaires. Citons : le THI, l'APHAB, le HHIES, l'IOI-HA...

3. **Réponses catégorielles non ordinales**. Il en existe peu dans notre domaine ; il peut s'agir de questionnaires ad hoc, où les réponses ne contiennent pas de notion d'ordre. Par exemple, une question peut concerner une préférence pour un réglage d'aide auditive.

Les réponses à un questionnaire doivent être vues comme des variables statistiques, auxquelles on applique donc les méthodes exposées dans cet article. Comme un questionnaire contient plusieurs questions, il y a donc un ensemble de variables à traiter. On commence en général par faire des statistiques descriptives : s'il s'agit de réponses quantitatives, on pourra calculer des moyennes, médianes, etc. S'il s'agit de variables catégorielles, on fera des tris à plat (nombre de sujets dans chaque catégorie de réponses, pour une question donnée).

Il faudrait normalement faire attention aux réponses de type 2 : les catégorielles ordinales. Il faut se demander s'il y a une échelle continue linéaire derrière, ou pas. Par exemple, le traitement des réponses à l'PAHAB est souvent très « limite ». En effet, les 7 modalités de réponses de A à G correspondent aux réponses : «toujours, presque toujours, en général, la moitié du temps, parfois, rarement, jamais ». Ces réponses viennent des percentiles des distributions de réponses possibles. Donc traiter des réponses à l'APHAB comme des échelles linéaires n'est pas rigoureux.

De manière générale, il est très tentant de traiter les réponses comme des variables continues, et

d'en faire la moyenne par exemple. Il ne faut pas être trop rigide, mais se poser le sens que cela a... Dans ce même esprit, il faudrait en toute rigueur ne pas traiter les seuils d'audiométrie recueillis par pas de 5 dB comme des variables quantitatives, surtout quand on sait que la perception de l'intensité sonore n'est certainement pas linéaire, mais tout le monde le fait... Il ne s'agit pas d'être rigide ou de faire du zèle, il s'agit une fois de plus de se demander le sens que l'on va attribuer à un petit p calculé sur une succession de transformations plus ou moins licites sur les données brutes (elles-mêmes très soumises à la subjectivité des réponses...).

Parfois, les auteurs d'un questionnaire (validé) proposent un score composite ; on peut citer l'exemple du THI, où l'on attribue 0/2/4 points à chaque réponse pour construire un score sur 100. Il peut ensuite être intéressant de le découper en classes, comme le grade THI (5 grades). Vu les traitements successifs infligés aux réponses brutes, on fera donc attention à ne pas considérer le grade final comme une échelle linéaire ! Le traitement statistique qui s'impose est bien celui d'une variable catégorielle.

Après avoir décrit, on peut chercher à expliquer. On pourra relier les réponses à des questionnaires à des variables de classement. On fera alors des t-tests et des ANOVA si la variable « réponse » est quantitative. Si elle est catégorielle il s'agira de tableaux croisés (de contingence) et de tests du Khi-deux.

Nous n'avons pas parlé d'épidémiologie

Dans tout ce qui précède, nous avons fait de l'expérimentation. Quand l'expérimentation n'est pas possible, on observe... La réalité observée est celle qui se présente à l'observateur, sans qu'il intervienne, sans qu'il puisse influer sur certains facteurs. Il est alors nécessaire de collecter un grand nombre de données, car beaucoup de facteurs non connus ou non maîtrisés peuvent expliquer les observations. C'est le champ de l'épidémiologie, qui sort du cadre de ce cours, mais dont nous pouvons dire quelques mots.

En épidémiologie on parle d'enquête plutôt que d'étude, et l'on en distingue 3 grands types :

- Enquêtes de cohorte
- Enquêtes cas-témoins
- Enquêtes transversales

Quelques mots sur les deux premières :

Dans une enquête de cohorte, on recrute des sujets qui sont exposés à un « risque », et d'autres qui ne le sont pas, et on les suit au cours du temps. Il s'agit donc d'une étude longitudinale prospective. On va observer le développement d'une maladie donnée dans les deux groupes. Tous ces termes s'entendent au sens large, et sont en fait génériques. Par exemple, l'exposition peut être, dans notre domaine, l'exposition au bruit, et la maladie peut être la déficience auditive (ne polémiquons pas sur l'emploi du terme « maladie » pour parler de surdité, merci !).

L'enquête cas-témoins est au contraire rétrospective. On recrute dans la population des personnes « atteintes d'une maladie » : ce sont les « cas ». On recrute également des personnes qui n'ont pas cette maladie, ce sont les « témoins ». On cherche alors à savoir si, dans leur passé, les sujets ont été exposés ou pas à un certain facteur. Pour reprendre le même exemple que précédemment, les « cas » sont les gens « sourds », et les témoins sont les normo-entendants. L'exposition peut être l'exposition au bruit (ou la prise d'un certain médicament, par exemple). Les

deux types d'enquêtes répondent à des problématiques différentes (par exemple maladies rares pour les enquêtes « cas-témoins » et expositions rares pour les enquêtes de cohorte). Les indicateurs statistiques sont des risques et des « odds-ratio » (« rapport de cote », comme pour la cote d'un pari). Nous renvoyons par exemple à l'abrégé de médecine (Ed. Masson) sur l'épidémiologie pour ceux qui souhaitent aller plus loin.

IX. Récapitulatif des principaux tests statistiques

Voici un récapitulatif par type de test :

T-test (test de Student) pour groupes parallèles (ou indépendants)	
Objectifs	Comparer les moyennes d'une variable X continue mesurée dans deux groupes de sujets différents
Exemples	- Comparer les moyennes de SRT mesuré chez des sujets femmes (groupe 1) et des sujets hommes (groupe 2) - Comparer les moyennes des EVA de douleur chez des sujets végétaliens et des sujets omnivores quand on leur présente une assiette de charcuterie
Conditions d'application	- Si n1 ≥ 30 et n2 ≥ 30 : aucune condition sur X - Si n1 ou n2 < 30 : les distributions de X dans les deux populations sont normales et de même variance
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative des moyennes entre les deux groupes.
Représentations graphiques associées	Boîtes à moustaches avec la moyenne et l'intervalle de confiance pour chaque groupe.

Alternative non paramétrique au t-test : test de Mann-Whitney (test des rangs signés)	
Objectifs	Comparer les ordres des valeurs de X dans deux groupes de sujets différents. X peut être continue ou catégorielle ordinale
Exemples	- Comparer les valeurs de SRT mesuré chez des sujets femmes (groupe 1) et des sujets hommes (groupe 2) - Comparer les réponses à une question (variable X) dont les réponses sont ordonnées. Ex : grades THI (5 valeurs possibles ordonnées)

	entre un groupe de sujets acouphéniques végétaliens et un groupe de sujets acouphéniques omnivores.
Conditions d'application	Aucune condition, adapté aux échantillons de taille < 30 quand la distribution de X n'est pas normale
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative des moyennes des rangs entre les deux groupes. On conclut donc souvent sur la différence des médianes.
Représentations graphiques associées	Boîtes à moustaches avec la médiane, les quartiles, et les extrêmes pour chaque groupe.

T-test (test de Student) pour sujets appariés	
Objectifs	Comparer les moyennes d'une variable X continue mesurée deux fois dans des conditions différentes ou à deux instants chez un même groupe de sujets
Exemples	- Comparer les moyennes de SRT mesuré chez des sujets avant et après un mois d'appareillage - Comparer les moyennes des fréquences cardiaques chez des sujets avant et après avoir couru un marathon
Conditions d'application	Par définition : n1 = n2 (on travaille sur des paires). On note n le nombre de sujets. Si n ≥ 30 : aucune condition Si n < 30 : la distribution des différences de X entre les deux conditions est normale
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative de la moyenne de X entre les deux conditions
Représentations graphiques associées	Boîtes à moustaches avec la moyenne et l'intervalle de confiance pour chaque condition.

Alternative non paramétrique au t-test : test de Wilcoxon	
Objectifs	Comparer les ordres d'une variable X mesurée deux fois dans des conditions différentes ou à deux instants chez un même groupe de sujets. X peut être continue ou catégorielle ordinale
Exemples	- Comparer les valeurs de SRT mesuré chez des sujets avec deux réglages différents (ex : avec et sans réducteurs de bruit). - Comparer les réponses à une question (variable X) dont les réponses sont ordonnées. Ex : grades THI (5 valeurs possibles ordonnées) chez des sujets acouphéniques récoltées avant appareillage et après 3 mois d'appareillage.
Conditions d'application	Aucune condition, adapté aux échantillons de taille < 30 quand la distribution des différences de X entre les deux conditions n'est pas normale
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative des moyennes des rangs entre les deux conditions. On conclut donc souvent sur la différence des médianes.
Représentations graphiques associées	Boîtes à moustaches avec la médiane, les quartiles, et les extrêmes pour chaque condition.

ANOVA (paramétrique) à 1 facteur non répété	
Objectifs	Comparer les moyennes d'une variable X continue mesurée dans plus de deux groupes de sujets différents
Exemples	- Comparer les moyennes de SRT mesuré chez des sujets implantés cochléaires avec 4 marques d'implant : A, B, C, D. - Comparer les moyennes des EVA de douleur chez des sujets végétaliens (groupe 1), des sujets végétariens (groupe 2), et des sujets omnivores (groupe 3) quand on leur présente une assiette de charcuterie
Conditions d'application	Distribution normale de X dans chaque groupe et égalité des variances. Si chaque échantillon

	est de taille > 30, on peut se passer de ces hypothèses (et bonne robustesse par rapport à la violation de celle d'homoscédasticité).
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative entre au moins une paire de moyennes parmi toutes les paires possibles. Pour savoir quelles moyennes diffèrent, on réalise des tests post hoc (typiquement Bonferroni). Cela revient (presque) à réaliser plusieurs t-tests avec un risque de 5% divisé par le nombre de comparaisons 2 à 2.
Représentations graphiques associées	Boîtes à moustaches avec la moyenne et l'intervalle de confiance pour chaque groupe.

ANOVA non paramétrique à 1 facteur non répété : ANOVA de Kruskall-Wallis	
Objectifs	Comparer les ordres des valeurs de X dans plus de deux groupes de sujets différents. X peut être continue ou catégorielle ordinale.
Conditions d'application	Aucune condition, adaptée quand les résidus d'une ANOVA paramétrique ne suivent pas une loi normale. Dans la pratique : adapté quand l'une au moins des distributions de X par condition n'est pas normale ou que l'une au moins des variances diffère d'une autre variance, dans le cas de variables continues. Adapté également aux variables ordinales.
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative entre au moins une paire de moyennes des rangs parmi toutes les paires possibles. Comme on travaille sur les rangs, cela revient à statuer sur une différence statistiquement significative entre au moins une paire de médianes parmi toutes les paires possibles. Pour savoir quelles médianes diffèrent, on réalise des tests post hoc (typiquement Bonferroni). Cela revient (presque) à réaliser plusieurs tests de Mann-Whitney avec un risque de 5% divisé par le nombre de comparaisons 2 à 2.
Représentations graphiques associées	Boîtes à moustaches avec la médiane, les quartiles, et les extrêmes pour chaque condition.

ANOVA (paramétrique) à 1 facteur répété	
Objectifs	Comparer les moyennes d'une variable X continue mesurée plus de deux fois dans des conditions différentes ou à plus de deux instants chez un même groupe de sujets
Exemples	- Comparer les moyennes de SRT mesuré chez des sujets avant appareillage, à 1 mois, 3 mois puis 6 mois d'appareillage - Comparer les moyennes des fréquences cardiaques chez des sujets avant un marathon, 2 secondes après l'arrivée, 1 heure après, 2 jours après...
Conditions d'application	Distribution normale de toutes les différences Xi – Xj (groupes i et j) et égalité des variances des distributions des différences. Si chaque échantillon est de taille > 30, on peut se passer de ces hypothèses. La condition sur les variances des distributions de différences est appelée hypothèse de sphéricité et se teste avec le test de Mauchly.
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative entre au moins une paire de moyennes parmi toutes les paires possibles. Pour savoir quelles moyennes diffèrent, on réalise des tests post hoc (typiquement Bonferroni). Cela revient (presque) à réaliser plusieurs t-tests pour groupes appariés avec un risque de 5% divisé par le nombre de comparaisons 2 à 2.
Représentations graphiques associées	Boîtes à moustaches avec la moyenne et l'intervalle de confiance pour chaque groupe.

ANOVA non paramétrique à 1 facteur non répété : ANOVA de Friedman	
Objectifs	Comparer les ordres d'une variable X mesurée plus de deux fois dans des conditions différentes ou à plus de deux instants chez un même groupe de sujets. X peut être continue ou catégorielle ordinale.
Exemples	- Comparer les valeurs de SRT mesuré chez des sujets avec deux réglages différents (ex : avec et sans réducteurs de bruit). - Comparer les réponses à une question (variable X) dont les réponses sont ordonnées. Ex : catégories THI chez des sujets acouphéniques récoltées avant appareillage et après 3 mois d'appareillage.
Interprétation	Si petit p < 0,05 on conclut à une différence statistiquement significative entre au moins une paire de moyennes des rangs parmi toutes les paires possibles. Comme on travaille sur les rangs, cela revient à statuer sur une différence statistiquement significative entre au moins une paire de médianes parmi toutes les paires possibles. Pour savoir quelles médianes diffèrent, on réalise des tests post hoc (typiquement Bonferroni). Cela revient (presque) à réaliser plusieurs tests de Wilcoxon avec un risque de 5% divisé par le nombre de comparaisons 2 à 2.

Test du Khi-deux	
Objectifs	Etudier la liaison entre deux variables catégorielles (ordinales ou non ordinales) mesurées dans un groupe de sujets.
Exemples	- Etudier le lien entre la satisfaction d'appareillage (mesurée par exemple en 3 grades : satisfait/neutre/insatisfait) et le type d'appareillage : CROS/biCROS/bilatéral conventionnel. - Etudier le lien entre la classe d'âge (par exemple découpé en 5 classes) et l'environnement de résidence principale (urbain/périurbain/rural). - Etudier le lien entre le grade THI et le type de régime alimentaire

	(végétalien/végétarien/omnivore) chez des sujets acouphéniques.
Conditions d'application	Les deux variables doivent être indépendantes. Par exemple, il ne peut pas s'agir de la même variable mesurée à deux instants différents (dans ce cas il faut faire un test de Cochran, ou de Mc Nemar si les deux variables sont dichotomiques). Dans le cas de deux variables dichotomiques (variables à 2 modalités), les effectifs théoriques (c'est-à-dire sous l'hypothèse H0 selon laquelle il n'y a pas de lien entre les deux variables) doivent être > 5. Si ce n'est pas le cas il faut réaliser un test (exact) de Fisher.
Interprétation	Si petit p < 0,05 on conclut que les deux variables sont reliées, sans pour autant savoir quelle variable influe sur l'autre.

Corrélation linéaire de Pearson	
Objectifs	Etudier la liaison linéaire entre deux variables continues mesurées dans un groupe de sujets.
Exemples	- Etudier le lien entre l'âge et l'EVA de gêne chez des sujets acouphéniques. - Etudier le lien entre la pression artérielle systolique et la fréquence cardiaque chez des champions de curling à la fin d'un entraînement intensif.
Conditions d'application	On peut calculer un coefficient de corrélation linéaire R sans AUCUNE condition sur les données. Si l'on veut faire de l'inférence et donc associer un petit p au test d'hypothèse « R = 0 vs R ≠ 0 », il faut que les résidus (terme d'erreur ou part non expliquée par le modèle linéaire) suivent une loi normale. En pratique, si n > 30, aucune condition n'est requise. Pour n < 30, il faut que les distributions de X et de Y suivent une loi normale.
Interprétation	Si petit p < 0,05 on conclut que R est statistiquement différent de 0. C'est la valeur de R qui renseigne sur la force de la liaison linéaire entre les deux variables. Si les variables

	sont corrélées linéairement (typiquement pour |R| > 0,7), on ne sait pas pour autant quel est le lien de causalité entre les deux variables.
Représentations graphiques associées	Nuage de points (Y en fonction de X ou vice versa) avec la droite des moindres carrés (droite de régression linéaire)

Corrélation non linéaire de Spearman (corrélation non paramétrique, effectuée sur les rangs)	
Objectifs	Etudier la liaison non linéaire mais tout de même monotone (à une valeur de X correspond une seule valeur de Y) entre deux variables continues mesurées dans un groupe de sujets, ou entre deux variables ordinales.
Conditions d'application	Aucune condition.
Interprétation	Si petit p < 0,05 on conclut que le coefficient de corrélation sur les rangs est statistiquement différent de 0. C'est la valeur de ρ qui donne la force de la liaison non linéaire entre les deux variables. Si les variables sont corrélées (typiquement pour |ρ| > 0,7), cela signifie que les deux variables sont « associées », qu'elles tendent à évoluer « ensemble ». Par exemple : l'une croît quand l'autre décroît, mais pas selon une droite. On ne connaît toujours pas pour autant le lien de causalité entre les deux variables.
Représentations graphiques associées	Nuage de points (sans droite d'ajustement, puisque l'on n'étudie pas la présence d'une relation linéaire).

Tests utilisés pour vérifier des hypothèses	
Tester la normalité	Tests de Shapiro-Wilk (recommandé), Kolmogorov-Smirnov, Lilliefors, Jarque-Bera, Cramer-von Mises, ... **Interprétation** : conclure à la normalité quand petit p > 0,05
Tester l'égalité des variances (homoscédasticité)	Tests de Levene, Bartlett, ... **Interprétation** : conclure à l'égalité des variances quand petit p > 0,05

X. Conclusion

Nous avons fait un petit tour des principaux tests statistiques, à partir d'un exemple fictif. Je concède qu'il y a un certain parti pris dans le ton. Je souhaite pour autant ne pas me poser en donneuse de leçon, j'essaie juste de transmettre ce que j'ai compris à force de pratiquer dans notre domaine. Je précise aussi que ce cours n'est pas exhaustif.

Voici quelques points pour résumer :

- Regardez vos données, et quand ce ne sont pas les vôtres, restez critique.

- Restez critique aussi si ce sont les vôtres, et soyez raisonnable sur ce que vous allez leur faire dire.

- Un résultat non statistiquement significatif n'est pas la fin du monde, peut-être même le contraire... Ce cours contient des approximations, parfois au détriment d'une certaine rigueur mathématique. J'en suis désolée. Enfin, pour expliquer le succès du petit p : je pense que nous souhaitons tous que la réalité soit résumable par une seule chose, quelque chose de simple, Malheureusement la plupart du temps, et en l'occurrence dans notre domaine, cela n'est pas possible. Si nous avions abordé ce cours par la recherche clinique, nous aurions commencé par parler d'essai contrôlé randomisé (« gold standard »), dont le calcul du nombre de sujets nécessaire repose sur un seul objectif (quantifié). Ce « gold standard » est lui aussi de plus en plus critiqué, mais il s'est imposé car il est facile à comprendre. Nous n'avons pas parlé de statistique bayésienne, qui est une manière plus rigoureuse, et ancienne, de « remonter des effets aux causes ». On parle d'inférence bayésienne, par opposition à l'inférence « classique » ou « fréquentiste » (celle qui nous a occupés ici). Elle requiert de définir, a priori (avant l'expérimentation) la distribution de probabilité d'un paramètre que l'on veut estimer (par exemple la moyenne). Cette distribution a priori modélise la connaissance antérieure que nous avons du sujet (avis d'experts, résultats publiés antérieurement). Le théorème de Bayes permet ensuite de calculer la probabilité a posteriori, par exemple la probabilité d'observer telle moyenne étant données les valeurs observées dans l'expérience. Grâce à la puissance de calcul actuelle, ce champ est en plein développement, y compris dans le dispositif médical. Il est particulièrement employé dans des domaines de la santé où l'on a besoin de faire des analyses intermédiaires pour stopper ou poursuivre un essai clinique, quand le nombre de sujets est faible et «couteux» (pénible pour les patients et économiquement pour la société), typiquement en oncologie. Je terminerai avec ce dinosaure et ses compagnons, tirés d'un article intitulé « Same Stats, Different Graphs » (voir la biblio), que je vous laisse commenter et apprécier :

-