Année 2 – Semestre 3
2023/2024

UC 0213
Communication et réglementation dans la profession vétérinaire (CoRVet)

Biostatistique en Médecine Vétérinaire

« Ce n'est pas parce qu'une différence est non significative (p> 0,05) que les groupes comparés sont similaires ou comparables »

Auteur : Pr Loïc Desquilbet
Version : Juin 2023

TABLE DES MATIERES

INDEX DES FIGURES

I. INTRODUCTION

A. Comment lire ce polycopié

Les parties de ce polycopié doivent se lire dans l'ordre, elles sont tout sauf indépendantes. Je fais en effet souvent référence dans une partie à ce que j'ai écrit dans une partie précédente.

Par ailleurs, tous les articles que je cite dans ce polycopié sont présents sur la page EVE de l'UC-0213, tout en bas de la section du module de Biostatistique en Médecine Vétérinaire

B. Quel intérêt d'enseigner les stat' dans un cursus vétérinaire ?

1. *Compétences générales visées*

Quel intérêt d'enseigner les stat' dans un cursus vétérinaire ? Réponse (selon moi) : pour vous préparer à acquérir les compétences listées dans le référentiel de diplôme sous la macro-compétence 'Agir en Scientifique' (cf. Annexe 1).

La première compétence que vous devez acquérir au cours de votre cursus est celle d'être capable de porter une analyse critique des communications scientifiques auxquelles vous serez confrontés, d'abord au cours de votre cursus (articles scientifiques que vous devrez lire pour les exercices d'enseignements, lors de vos rotations cliniques, et pour votre thèse vétérinaire), mais aussi et surtout au cours de votre vie professionnelle (communications écrites, conférences, discussions entre collègues, …). Pour cela, vous devrez acquérir entre autres les bases en biostatistique.

La seconde compétence que vous devez acquérir est celle d'être capable d'appliquer l' « Evidence-based veterinary medicine » (EBVM ; médecine fondée sur les preuves (MFP), en français). Rapidement, l'EBVM est définie comme l'**utilisation** consciencieuse, explicite, et judicieuse de la **meilleure preuve** disponible pour **la prise de décision** concernant le **soin** du patient. Ce que l'on entend par « meilleure preuve », c'est une « ressource scientifique faisant état d'une démarche scientifique la plus rigoureuse possible ». Ce que l'on entend par « ressource scientifique », ce sont par exemple le chapitre d'un livre, un article publié dans une revue française ou internationale, un compte-rendu d'une conférence, ou l'opinion d'un expert entendue au cours d'une discussion. Ainsi, être capable d'appliquer l'EBVM, c'est être capable de repérer la meilleure preuve parmi celles que vous entendez ou lisez pour l'appliquer ensuite dans le soin de l'animal que vous, en tant que vétérinaire, avez en face de vous en consultation.

La troisième compétence est celle d'être capable de contribuer à l'accroissement des connaissances en médecine vétérinaire et plus largement dans le domaine des sciences du vivant. Là, encore, cette compétence nécessite des compétences en biostatistique.

2. *Développer l'esprit critique chez les étudiants au cours du cursus à l'EnvA*

Être capable de porter un regard critique sur une étude clinique avant d'appliquer un traitement à un animal malade ou avant de donner des conseils de prévention à un propriétaire, cela s'apprend. *Rigoureusement* douter (et non pas « douter pour douter »), cela s'apprend. Par exemple, si en tant que vétérinaire, vous assistez à une conférence au cours de laquelle une personne présente les résultats d'une étude clinique dont la conclusion est de

traiter des chats qui souffrent d'une insuffisance cardiaque avec un traitement A, vous devez prendre du recul et avoir un esprit critique : « est-ce que ce message clinique est 'evidence-based' ? », ou autrement dit, « est-ce que ce message clinique est soutenu par les résultats d'une étude clinique dont la méthodologie clinique utilisée est rigoureuse ? »

Pour vous préparer à l'acquisition de cet esprit critique de la méthodologie d'une étude clinique, nous allons nous voir cette année, en 3ème année, puis en 5ème année lors de rotations cliniques.

3. Biostatistique & épidémiologie : des pré-requis indispensables à l'analyse critique d'articles

Cette capacité à critiquer un document dans sa méthodologie statistique demande de bonnes connaissances de base en biostatistique, mais aussi et surtout, de bonnes connaissances en épidémiologie. Qu'est-ce que l'épidémiologie ? Pour faire simple, disons que c'est la science médicale permettant de comprendre les mécanismes de survenue d'un mauvais état de santé d'un être vivant. On peut distinguer l'épidémiologie **descriptive** et l'épidémiologie **analytique**. La première a principalement pour objectif de décrire et d'anticiper l'apparition d'un mauvais état de santé ; la seconde a principalement pour objectif de rechercher les facteurs de risque d'un mauvais état de santé afin, entre autres, de faire de la prévention efficace. Les articles scientifiques vétérinaires font souvent appel aux outils issus de l'épidémiologie. Soit parce qu'ils veulent décrire la maladie étudiée dans une population donnée, soit parce qu'ils veulent identifier les facteurs de risque d'une maladie. Or, pour maîtriser les outils issus de l'épidémiologie, il faut de façon indispensable maîtriser les bases de la biostatistique. Donc, si vous n'avez pas acquis ces bases en biostatistique, vous ne pourrez pas acquérir des connaissances / compétences solides en épidémiologie qui sont indispensables à la réalisation d'une analyse critique d'un article.

Par conséquent, des lacunes en biostatistique dès la 2ème année vous porteront préjudice en 3ème pour l'épidémiologie clinique, et des lacunes en épidémiologie clinique vous porteront préjudice ensuite quand nous travaillerons sur l'EBVM en 5ème année.

II. DEFINITIONS ET PRESENTATION DES CONCEPTS

A. La notion d' « étude »

Dans tout ce polycopié, je vais utiliser le terme « étude ». Ce terme, générique, peut faire référence à un essai clinique[1] ou une étude (ou « enquête ») épidémiologique, dont les objectifs peuvent être très variés. Le point commun parmi ces études ou enquêtes est le fait qu'elles soient constituées d'un *échantillon*, et qu'elles aient pour objectif de faire porter leurs résultats issus de l'échantillon vers une *population* d'individus.

B. Échantillon

L'échantillon est le groupe d' « individus » sur lesquels les analyses statistiques sont effectuées. Dans le domaine des animaux de production, il faut bien faire attention si l'échantillon est constitué d'élevages (auquel cas, l' « individu » est l'élevage, et les données recueillies le sont à l'échelle de l'élevage en entier ; citons par exemple la taille de l'élevage, l'hygiène de l'élevage, le type de stabulation, le type de l'élevage allaitant/laitier/mixte, …) ou bien constitué d'animaux (auquel cas, l' « individu » est l'animal au sein d'un élevage, et les données recueillies le sont à l'échelle de l'animal ; citons par exemple la note d'état corporel, la parité, les antécédents de mammites, …). La « taille de l'échantillon » est le nombre d'individus que compte l'échantillon.

C. Population cible

La population cible est la population que l'on vise quand on met en place l'étude ; c'est la population à laquelle on voudrait pouvoir étendre les résultats. Il est fondamental de correctement définir la population cible quand on met en place une étude, car elle va permettre de choisir la population source (cf. ci-dessous) de telle façon à ce que cette dernière soit la plus proche possible de la population cible. Il est par ailleurs tout aussi fondamental d'identifier la population cible quand vous lisez un article scientifique car vous, en tant que vétérinaire, saurez ainsi les individus sur lesquels vous pourrez *a priori* appliquer les résultats de l'étude, et ceux sur lesquels vous ne le pourrez *a priori* pas. Ainsi, quand, en tant que vétérinaire praticien, vous aurez un animal devant vous qui pourrait avoir besoin d'être traité par un traitement A, en lisant l'étude qui étudie ce traitement A et en remarquant la population cible de l'étude, vous saurez si les résultats de cette étude peuvent s'appliquer, ou non, à l'animal que vous avez en face de vous !

Dans la très grande majorité des communications scientifiques, la population cible est mentionnée dans l'objectif principal de l'étude. Ce sera naturellement systématiquement le cas dans tout le module de Biostatistique en Médecine Vétérinaire, que ce soit en TD ou en examen.

[1] Un essai clinique est une étude médicale ayant souvent pour objectif de montrer l'efficacité ou la tolérance d'une molécule, d'un traitement, ou plus généralement d'une intervention thérapeutique.

D. Population source

La population source est constituée des individus d'où sont extraits ceux qui ont fait partie de l'échantillon. Dit autrement, et plus pragmatiquement, la population source est l'ensemble des individus *susceptibles* de faire partie de l'échantillon. Ce mot « susceptible » est fondamental. Pour définir la population source, il faut *imaginer* tous les individus qui auraient pu faire partie de l'échantillon si le processus d'échantillonnage avait été réalisé une infinité de fois ! C'est uniquement la lecture du protocole d'une étude qui vous permet de définir la population source d'une étude.

E. « Echantillonner » et « fluctuation d'échantillonnage »

Le verbe « échantillonner » signifie « créer un échantillon à partir de la population source ». Dès que vous échantillonnez, de la fluctuation d'échantillonnage se produit.

La fluctuation d'échantillonnage est la manifestation du hasard dans la constitution de l'échantillon à partir de la population source. La fluctuation d'échantillonnage est i-né-luc-ta-ble, et elle doit toujours être dans votre tête lorsque vous lisez les résultats d'une étude. Quelle est la conséquence *majeure*[2] de cette « fluctuation d'échantillonnage » ? C'est le fait que si l'on tirait au sort deux échantillons issus de la *même* population source, les résultats dans chacun de ces deux échantillons (par exemple, les deux moyennes ou les deux pourcentages calculés) ne seraient jamais identiques (sauf coup de chance magistral). En pratique, il n'y a qu'un seul échantillon issu de la population source. Donc, vous ne verrez pas la manifestation de cette fluctuation d'échantillonnage, mais vous devrez toujours avoir en tête que la fluctuation d'échantillonnage s'*est* manifestée au moment d'échantillonner.

F. L'inférence

De façon générale, faire de l'inférence, ou « inférer », c'est étendre les résultats observés dans l'échantillon à la population *cible*. Toute étude a pour objectif de faire de l'inférence. En effet, quel serait l'intérêt d'une étude qui cantonne ses résultats à son propre échantillon ? Entre le moment où les données sont collectées pour les analyses statistiques et le moment où les résultats sont communiqués (oralement, par écrit dans un article, …), probablement qu'une partie non négligeable des individus est déjà morte[3]. Et quel est alors l'intérêt de parler d'animaux qui sont morts ?! Aucun ! ☺

G. Le « caractère » et la « variable »

Dans ce polycopié, un « caractère » est une caractéristique intrinsèque ou extrinsèque d'un individu de l'échantillon. Un caractère peut être de trois types : binaire, qualitatif, ou quantitatif. Un exemple de caractère binaire est le sexe d'un animal (mâle / femelle). Un exemple de caractère qualitatif est la race d'un chien (Bouledogue français, Golden Retriever,

[2] Certains vont probablement penser que j'abuse de l'écriture en italique ! Oui, c'est vrai, je ne vais pas la réserver aux seuls termes latins. Il ne s'agit pas d' « abus » ! Simplement, dans le domaine de la biostatistique et l'épidémiologie (comme dans beaucoup d'autres domaines – tous ?...), les mots sont très importants, et je les souligne en italique (plutôt qu'en les soulignant proprement dit).

[3] Cette année, je tâcherai de ne pas être trop cynique. Je ne peux rien vous promettre, mais j'essaierai…

Dalmatien, …). Notez qu'un caractère qualitatif comprend au moins trois classes. Un exemple de caractère quantitatif est l'âge d'un chat. Un caractère quantitatif, comme son nom l'indique, doit représenter une quantité. Lorsque le caractère possède une unité de mesure, il n'y a pas de question à se poser, le caractère est de fait quantitatif. Quand il n'y a pas d'unité de mesure, il est parfois possible d'hésiter entre le type quantitatif et le type qualitatif. Cela dit, en cas d'hésitation, si le caractère est sans unité de mesure et ne représente pas un nombre d'une chose, alors il y a de bonnes chances pour que le caractère soit de type qualitatif (on dira alors qu'il est de type qualitatif ordinal, par opposition à un caractère de type qualitatif nominal, où les classes ne sont pas ordonnées, comme par exemple la race d'un chien). Lorsqu'un caractère est collecté dans une étude, il est appelé « variable » dans le fichier de données de l'étude. Ainsi, les deux termes « caractère » et « variable » seront interchangeables dans ce polycopié.

III. STATISTIQUE DESCRIPTIVE

A. Introduction

1. *Définitions d' « indicateur » et d' « estimation »*

Un « indicateur » est un « dispositif fournissant des repères et servant à mesurer ». Par extension, un « indicateur » est un « élément permettant d'évaluer certains phénomènes »[4]. Les « indicateurs » que nous verrons dans le module de Biostatistique en Médecine Vétérinaire sont les suivants : moyenne, médiane, pourcentage, premier quartile (Q1), troisième quartile (Q3), minimum, maximum, variance, Standard Deviation, et Standard Error. A part le dernier indicateur cité (la Standard Error), tous les autres indicateurs que je viens de citer sont des indicateurs statistiques descriptifs qui permettent de *décrire* un échantillon. Il existe d'autres indicateurs statistiques, mais dont je ne parlerai pas.

Une « estimation » est le résultat d'un calcul d'un « indicateur » à partir des données d'un échantillon. Par exemple, on parlera de l'estimation d'une moyenne ou d'un pourcentage à partir des données d'un échantillon.

2. *Signification du mot « réel » dans ce polycopié*

Dès que j'emploierai le mot « réel » dans ce polycopié, mais aussi dans tous les exercices d'enseignement que vous aurez avec moi en 2ème, 3ème, et 5ème années, cela signifie que je parle de « la population cible ». Ainsi, « la valeur réelle du pourcentage de prévalence de pancréatite chronique » signifie « la valeur du pourcentage de pancréatite chronique dans la population cible ». Parfois, j'accole « réel » à « population cible » (par exemple, « la valeur réelle du pourcentage de pancréatite chronique dans la population cible ») : c'est une information redondante, mais c'est pour bien insister que je parle de la valeur dans la population cible.

3. *Objectif de la statistique descriptive*

L'un des objectifs de la statistique descriptive est de fournir une estimation d'un indicateur (calculée dans un échantillon) qui soit la plus proche possible de la valeur réelle de cet indicateur dans la population *cible*. Notez que la valeur de cet indicateur dans la population cible est forcément inconnue ! En effet, comme la population cible est composée de beaucoup trop d'individus, il ne sera jamais possible de collecter l'information sur le caractère étudié parmi *tous* les individus de la population cible pour ensuite calculer l'indicateur d'intérêt.

4. *Notations*

De façon générale, dans ce document, les lettres grecques vont toujours faire référence à des indicateurs dans la population cible. Je vais noter « θ » la valeur d'un indicateur quelconque (que ce soit une moyenne, une médiane, un pourcentage, …). Plus spécifiquement, « μ » sera la moyenne d'un caractère quantitatif, et « π » le pourcentage d'un caractère binaire. Les lettres grecques avec un chapeau au-dessus vont faire référence à la valeur *estimée* de

[4] https://www.cnrtl.fr/lexicographie/indicateur

l'indicateur dans l'échantillon (par exemple, $\hat{\mu}$ pour l'estimation d'une moyenne et $\hat{\pi}$ pour l'estimation d'un pourcentage).

Dans la mesure où θ est la valeur d'un indicateur dans la population cible, et dans la mesure où la population cible est inatteignable, θ est forcément inconnue.

B. Normalité d'une distribution d'une variable quantitative

Pour vérifier qu'une variable quantitative suit une loi normale, une des nombreuses méthodes est de dresser un histogramme. Un site Internet très simple d'utilisation permet de dresser un histogramme : http://www.socscistatistics.com/descriptive/histograms/ (attention, si vous avez des valeurs avec un chiffre après la virgule, le symbole décimal doit être le point, et non la virgule, au moment où vous copiez-collez vos valeurs sur le site).

Comment sait-on si une distribution suit une loi normale ou pas ? Déjà, la normalité mathématique parfaite n'existe (quasiment) pas dans la nature. Par conséquent, cette appréciation est subjective. La distribution peut être considérée comme normale si elle suit une forme de cloche, c'est-à-dire répondre aux trois critères ci-dessous :

1) Etre relativement symétrique,

2) Avoir peu de valeurs extrêmes et la majorité des valeurs autour de la moyenne,

3) N'avoir qu'une seule « grosse bosse ».

Ces critères sont bien évidemment subjectifs, et il est tout à fait accepté dans la littérature scientifique qu'ils le soient. Des critères dits « objectifs », notamment à partir de résultats d'un test statistique, ne doivent pas être utilisés pour montrer qu'une distribution peut être considérée comme normale. La justification de cela est présentée en Annexe 2.

C. Indicateurs usuels de statistique descriptive

1. *Le pourcentage et le pourcentage de prévalence d'une maladie*

Même si évidemment tout le monde (ou quasiment) sait ce qu'est un pourcentage, et même si la compétence requise pour en *calculer* un est *a priori* acquise depuis un certain nombre d'années, nous reviendrons dessus au moment où l'on comparera deux pourcentages pour savoir s'il existe une association entre deux caractères. Mais la règle magistralement importante est la suivante : quand vous citerez en français un pourcentage, vous devrez o-bli-ga-toi-re-ment utiliser le mot « parmi ». Par exemple, « le pourcentage de chats femelles *parmi* les chats atteints d'une pancréatite », ou bien « le pourcentage de chats atteints d'un lymphome *parmi* les chats sortant moins d'une heure par jour dehors ».

Le pourcentage de prévalence d'une maladie dans un échantillon est le pourcentage d'individus atteints de cette maladie parmi les individus de l'échantillon. Par exemple, « le pourcentage de chats atteints d'un lymphome *parmi* les chats sortant moins d'une heure par jour dehors » est aussi « le pourcentage de prévalence d'un lymphome *parmi* les chats sortant moins d'une heure par jour dehors ».

2. *La moyenne et hypothèse de calcul d'une moyenne*

Là encore, tout le monde (ou quasiment) sait ce qu'est une moyenne !... En revanche, il y a deux choses importantes à savoir sur la moyenne. Tout d'abord, une moyenne ne se calcule que pour un caractère quantitatif (il est en effet hors de question de calculer une moyenne sur un caractère binaire ou qualitatif, même si ce dernier est qualitatif *ordinal*). Deuxième chose, et ce que l'on sait moins, c'est que pour correctement interpréter une moyenne, la distribution du caractère quantitatif (dont on calcule la moyenne) doit être considérée comme normale.

3. *La variance et la Standard Deviation (SD) d'un caractère quantitatif*

Pour décrire un caractère quantitatif dont la distribution peut être considérée comme normale, on utilise la moyenne ainsi que la Standard Deviation (SD) ; la moyenne et la SD ne doivent pas être fournies si la distribution du caractère ne peut pas être considérée comme normale. Dans ce cas de figure-là, on fournit la médiane, le premier quartile et le troisième quartile.

Ne comptez pas sur moi pour vous fournir des formules mathématiques pour calculer une variance ou une SD[5] (vous savez probablement même mieux que moi où les retrouver sur Internet ; cela dit, pour information, la valeur de la variance est tout simplement la valeur de la SD élevée au carré).

En revanche, là encore, il y a plusieurs choses importantes à savoir. La première, c'est que la variance et la SD sont uniquement calculées pour un caractère quantitatif, et elles quantifient toutes les deux la *variabilité* de ce caractère. Cette variabilité est en quelque sorte fixée par la « nature », parfois avec l'aide de l'Homme… Plus la valeur de la variance ou celle de la SD augmente, plus le caractère est variable d'un individu à un autre. La valeur de la variance n'est pas interprétable en tant que telle. En revanche, la valeur de la SD est tout à fait interprétable. Tout d'abord, la SD s'exprime dans la même unité que celle du caractère quantitatif. Ensuite, sa valeur s'interprète de la façon suivante (cf. **Figure 1** ci-dessous) : si la distribution du caractère quantitatif peut être considérée comme normale, et si m et SD sont respectivement les estimations de la moyenne et de la SD de ce caractère dans l'échantillon, 68% (soit environ 2/3) des individus de l'*échantillon* ont une valeur de ce caractère comprise entre m-SD et m+SD, et 95% des individus de l'*échantillon* ont une valeur de ce caractère comprise entre m-1,96xSD et m+1,96xSD. La démonstration est faite en Annexe 3. Attention à ne surtout pas confondre cette interprétation (qui utilise la SD) avec celle de l'intervalle de confiance à 95% (qui utilise la Standard Error), que l'on verra plus loin dans ce polycopié.

[5] En français « Standard Deviation » = « écart-type dans l'échantillon ». En raison d'une confusion trop importante entre « écart-type dans l'échantillon » et « écart-type d'une estimation » (qui est la « Standard Error », cf. plus loin dans le polycopié), j'ai choisi d'utiliser la terminologie anglaise qui est moins source de confusion (et c'est celle que je vais vous demander d'utiliser).

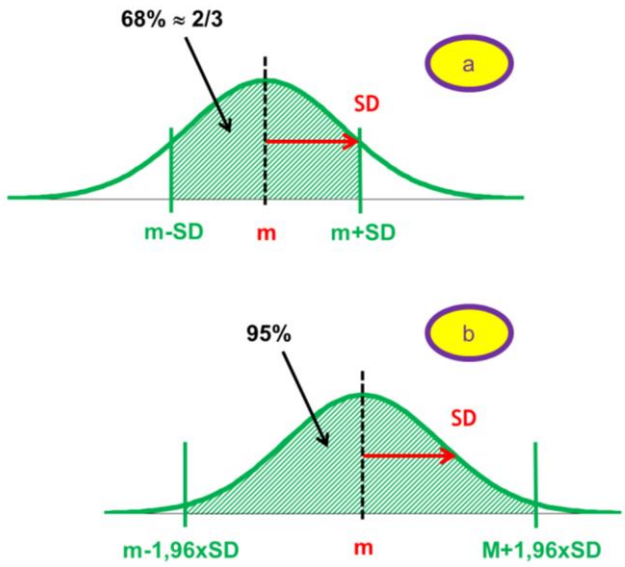


Figure 1. Interprétation graphique de la valeur de la Standard Deviation (SD). La distribution en vert représente la distribution du caractère quantitatif dans un échantillon dont on calcule la moyenne parmi les individus de l'échantillon. (a) Environ 2/3 des individus de l'échantillon ont une valeur du caractère quantitatif comprise entre m-SD et m+SD, et (b) 95% des individus de l'échantillon ont une valeur du caractère quantitatif comprise entre m-1,96xSD et m+1,96xSD.

La seconde chose importante, c'est que la variabilité d'un caractère quantitatif, fixée par la nature, ne dépend pas de la taille de l'échantillon ! En effet, le nombre de doigts sur une main (humaine) est un caractère quantitatif très peu variable. Certes, sur 10.000 individus, il y a plus de chances d'en observer au moins un avec 4 doigts, que sur 100 individus. Mais si, dans la population, la proportion d'individus avec 4 doigts sur une main est de 0,01%, alors en espérance, la proportion d'individus avec 4 doigts dans un échantillon tiré au sort de la population sera elle aussi de 0,01%, quelle que soit la taille de l'échantillon !

4. *La médiane, ses conditions d'utilisation, et les quartiles*

L'interprétation de la médiane et des quartiles que je vais vous présenter ci-dessous est une interprétation pragmatique. L'interprétation rigoureuse et mathématique s'écarte de cette interprétation pragmatique, et elle est présentée dans l'Annexe 4 (hors programme).

La médiane se calcule pour un caractère quantitatif. Soit V_{med} la valeur de la médiane calculée dans un échantillon. Cette valeur de la médiane V_{med} est telle que 50% des individus de l'échantillon ont une valeur inférieure ou égale à V_{med} (et donc 50% des individus de l'échantillon ont une valeur supérieure à V_{med}). Quand une médiane est fournie, on fournit aussi généralement les 1er et 3ème quartiles (cf. ci-dessous) pour fournir l'information sur la variabilité du caractère quantitatif. Au passage, si la distribution du caractère quantitatif est parfaitement normale, la moyenne et la médiane sont égales. Cela dit, ce n'est pas parce que la médiane et la moyenne sont égales que la distribution est normale[6]…

Contrairement à la moyenne, V_{med} peut s'interpréter quelle que soit la forme de la distribution du caractère quantitatif (distribution normale ou non normale). Ainsi, lorsque la distribution du caractère quantitatif peut être considérée comme normale, on peut présenter soit la

[6] Je vous laisse éventuellement méditer sur ce point !... (Si vous n'avez pas d'autres sources de méditation, bien entendu.)

moyenne, soit la médiane (au choix des chercheurs) ; mais lorsque la distribution du caractère quantitatif ne peut pas être considérée comme normale (ou lorsqu'il n'est pas possible de vérifier la normalité d'une distribution en raison d'une taille d'échantillon trop faible pour dresser un histogramme), la médiane *doit* être présentée.

L'interprétation pragmatique de la valeur du 1er quartile (Q1, qui est aussi le 25ème percentile de la distribution du caractère quantitatif) est la suivante : 25% des individus de l'échantillon présentent une valeur du caractère quantitatif inférieure ou égale à la valeur de ce 1er quartile. L'interprétation pragmatique du 3ème quartile (Q3, qui est aussi le 75ème percentile de la distribution du caractère quantitatif) est la suivante : 75% des individus de l'échantillon présentent une valeur du caractère quantitatif inférieure ou égale à la valeur de ce 3ème quartile.

La distance interquartile (« interquartile range », ou « IQR » en anglais) est l'intervalle [Q1 ; Q3] ; il fournit une indication de la variabilité du caractère mesuré (tout comme le fait la SD) car on peut dire que, dans l'échantillon, 50% des individus ont une valeur du caractère quantitatif comprise entre Q1 et Q3.

5. *L'étendue*

L'étendue (« Range » en anglais) ne peut être fournie que si le caractère est quantitatif. Elle composée de deux valeurs, à savoir la valeur minimale et la valeur maximale de ce caractère observées dans l'échantillon. L'étendue peut être fournie que le caractère suive, ou non, une distribution normale.

6. *En résumé*

La **Figure 2** résume les informations fournies ci-dessus.

Type de variable	Indicateur(s)	Condition d'utilisation ou commentaires
Variable binaire ou qualitative	Pourcentage (%)	Le mot « parmi » doit être obligatoirement utilisé en citant un %
Variable quantitative	Moyenne	La distribution de la variable quantitative doit être considérée comme normale
	SD	
	Médiane	Aucune condition de distribution de la variable quantitative
	Distance interquartile (Q1-Q3)	
	Etendue (min-max)	

SD = standard deviation
Q1 = 1er quartile
Q3 = 3ème quartile

Figure 2. Indicateurs utilisés et conditions éventuelles d'utilisation selon le type de variable (ou de caractère).

D. Qualité d'une estimation

1. *Problématique*

Supposons que vous souhaitiez connaître le pourcentage de chiens présentant des problèmes locomoteurs parmi les (millions de) chiens adultes en France. Vous allez mettre en place votre étude, c'est-à-dire demander à vos amis qui ont des chiens s'ils ont observé récemment des problèmes neurolocomoteurs. Supposons que parmi 18 chiens de votre entourage, 4 ont semble-t-il présenté récemment des problèmes neurolocomoteurs, ce qui conduit à un pourcentage estimé de 4/18=22%. La question est désormais la suivante : votre estimation est-elle suffisamment de *qualité* pour permettre d'inférer ce résultat de 22% à l'ensemble de la population des millions chiens adultes de France, en disant notamment que « en France, il y a des chances pour que le pourcentage de prévalence de problèmes locomoteurs parmi les chiens adultes soit proche de 22% » ? Pour y répondre, il faut s'assurer que cette estimation soit *précise* et *exacte* ! Nous allons voir ce que ces deux termes signifient.

2. *Précision et Standard Error (SE)*

Une estimation $\hat{\theta}$ est dite *précise* si (attention, ça va être théorique), en *imaginant* que l'on échantillonne *n* fois et que l'on calcule *n* fois $\hat{\theta}$ (cf. **Figure 3**), ces *n* valeurs de $\hat{\theta}$ sont très proches les unes des autres.

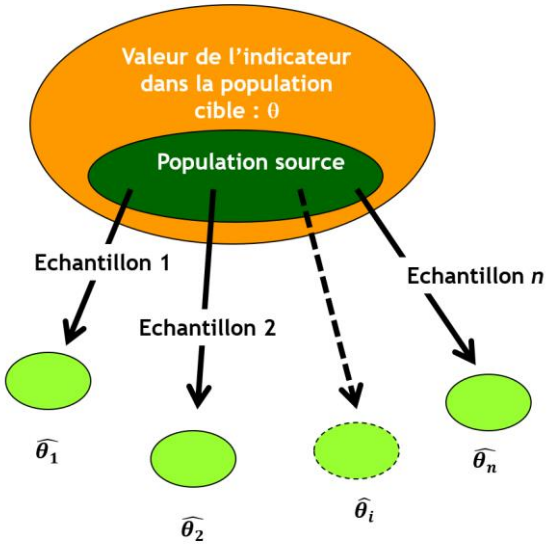


Figure 3. Processus théorique d'échantillonnage et d'estimations multiples.

Dire que ces valeurs $\widehat{\theta_i}$ sont très proches les unes des autres, c'est aussi dire qu'elles sont très peu « dispersées », ou sont très peu « variables » les unes par rapport aux autres. *A contrario,* plus les valeurs $\widehat{\theta_i}$ sont éloignées les unes par rapport aux autres, plus on peut se dire intuitivement que *chacune* de ces estimations est imprécise (car d'une estimation à l'autre, on obtient une valeur très différente).

Tout cela est très théorique puisqu'en pratique, on ne prélève qu'un seul échantillon de la population (source). En pratique, pour quantifier la précision d'une estimation (cette

variabilité théorique de toutes ces $\widehat{\theta_i}$), on calcule la Standard Error (SE)[7] de cette estimation. Plus la SE diminue, plus l'estimation est précise. Vous aurez à utiliser deux SE au cours des TD : la SE d'une moyenne (SE_m) et la SE d'un pourcentage (SE_p). Les formules ci-dessous permettant d'obtenir ces formules de calcul de la SE sont à connaître par cœur.

$SE_m = \frac{SD}{\sqrt{n}}$, où *SD* est la Standard Deviation, et *n* la taille de l'échantillon.

$SE_p = \sqrt{\frac{p.(1-p)}{n}}$, où *p* est le pourcentage estimé et *n* la taille de l'échantillon.

A partir de ces deux formules des SE_m et SE_p, nous pouvons remarquer une chose fondamentale : lorsque n augmente, la SE diminue. En français, cela signifie que plus la taille de l'échantillon augmente, plus l'estimation est précise.

3. *Exactitude et notion de « biais d'estimation »*

Une estimation $\hat{\theta}$ est dite exacte si elle n'est pas biaisée. Qu'est-ce qu'un « biais d'estimation » ? Le biais d'estimation est l'écart entre la moyenne de toutes les estimations $\widehat{\theta_i}$ que l'on aurait calculées à partir d'une infinité d'échantillons tirés de la population source et la vraie valeur inconnue θ dans la population cible (cf. **Figure 4**). Vous trouverez ci-dessous deux situations bien différentes. La première, celle où la façon d'estimer θ est précise, mais biaisée car les estimations $\widehat{\theta_i}$, bien que proches les unes des autres, sont systématiquement inférieures à la vraie valeur θ (**Figure 4**). La seconde situation présente une façon imprécise d'estimer θ car les $\widehat{\theta_i}$ sont éloignées les unes des autres, mais de façon cependant exacte (sans biais d'estimation) car les $\widehat{\theta_i}$ sont *autour* de la vraie valeur de θ, ce qui fait que la moyenne de toutes ces $\widehat{\theta_i}$ est égale à la vraie valeur θ (**Figure 5**).

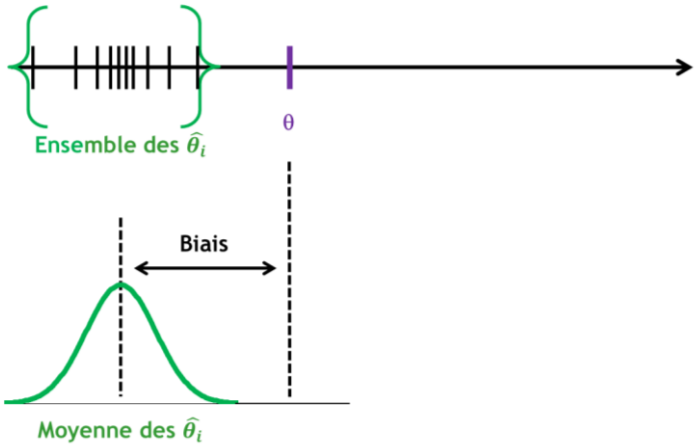


Figure 4. Estimation précise mais biaisée.

[7] En français : « écart-type de l'estimation »

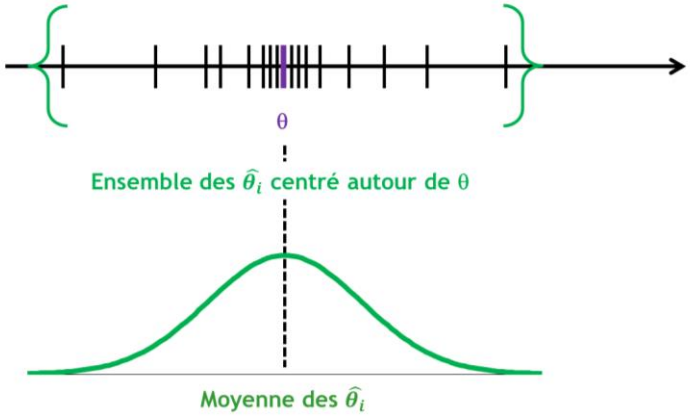


Figure 5. Estimation imprécise mais non biaisée.

On peut comprendre cette définition du « biais » d'une autre façon : le biais d'estimation est l'écart entre l'estimation d'un indicateur dans l'échantillon et la valeur réelle de cet indicateur dans la population cible *en retirant la part du hasard dans cet écart* (la part de hasard dans l'écart entre $\widehat{\theta_i}$ et θ, due à la fluctuation d'échantillonnage, c'est l'écart entre $\widehat{\theta_i}$ et la moyenne de ces $\widehat{\theta_i}$).

Une dernière façon de définir le biais d'estimation, en étant (un peu) plus proche de la réalité de terrain où il n'y a qu'une seule estimation $\hat{\theta}$ (et non une multitude de $\widehat{\theta_i}$) est la suivante : le biais d'estimation est l'écart *systématique* entre la valeur estimée $\hat{\theta}$ et la valeur réelle θ. « Systématique » dans le sens où si l'on refaisait l'échantillonnage une infinité de fois, on aurait un écart entre la valeur estimée $\hat{\theta}$ et la valeur réelle θ systématiquement du même ordre (non nul) de grandeur (et cet écart systématique est la valeur du « biais »).

Comme la vraie valeur de θ est inconnue, le biais ne peut donc jamais se quantifier. Il peut en revanche s'apprécier et se discuter[8].

L'origine des biais d'estimation n'est pas décrite dans ce polycopié, car n'étant pas au programme du module de Biostatistique en Médecine Vétérinaire de 2ème année. Cela dit, sachez qu'il existe deux biais d'estimation en épidémiologie descriptive : le biais d'échantillonnage et le biais de mesure.

4. *En résumé*

La **Figure 6** représente graphiquement l'ensemble des situations concernant la qualité d'une estimation. Une estimation est de bonne qualité si elle est précise (c'est-à-dire, avec une SE faible) *et* si elle est exacte (c'est-à-dire exempt de biais d'estimation). Dans la mesure où l'on arrive davantage à quantifier l'imprécision que l'inexactitude (car les biais ne sont pas quantifiables), il vaut à la limite mieux estimer un indicateur de façon imprécise mais exacte (**Figure 6**.c) plutôt que de façon précise mais inexacte (**Figure 6**.a) !...

[8] D'ailleurs, c'est l'une des immenses et non moins intéressantes tâches de l'épidémiologiste : discuter (et prendre en compte quand c'est possible) la présence de biais et l'impact qu'ont ces biais sur la capacité à faire de l'inférence statistique (et causale). Le module d'épidémiologie clinique que vous aurez avec moi dans l'UC-0324 en 3ème année est dédiée entièrement à la discussion des biais – super :'(.

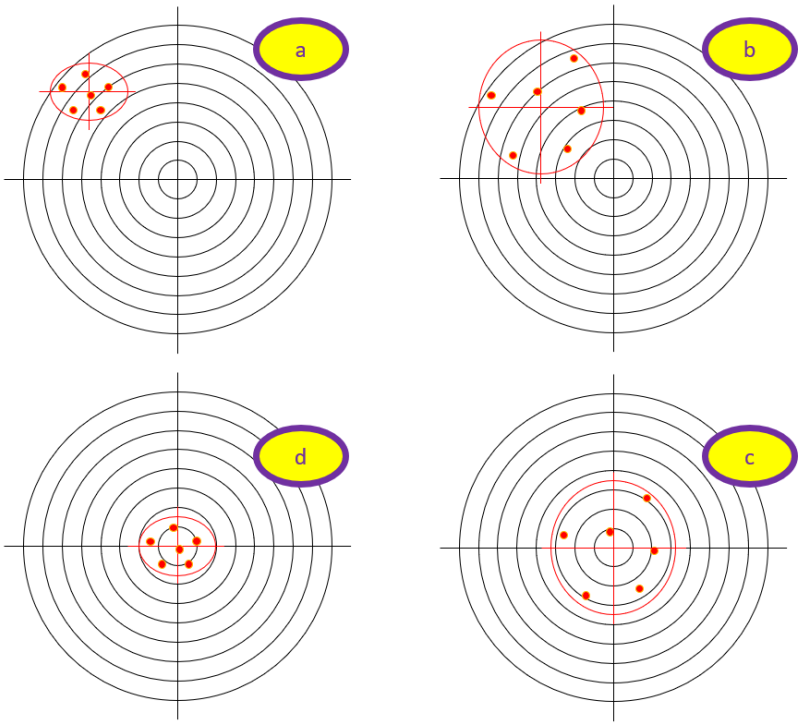


Figure 6. Illustration de la qualité d'une estimation. Chaque point rouge représente une estimation $\widehat{\theta}_i$. La figure (a) représente la situation d'une façon d'estimer $\hat{\theta}$ précise (car les estimations sont rapprochées les unes des autres) mais biaisée (car il existe un écart entre le centre de la cible et le centre du cercle rouge ⇔ existence d'un écart *systématique* entre chaque estimation et la valeur de θ dans la population cible). La figure (b) représente la situation d'une façon d'estimer $\hat{\theta}$ imprécise (car les estimations sont éloignées les unes des autres) et biaisée. La figure (c) représente la situation d'une façon d'estimer $\hat{\theta}$ imprécise mais non biaisée (car il n'existe aucun écart entre le centre de la cible et le centre du cercle rouge ⇔ absence d'écart *systématique* entre chaque estimation et la valeur de θ dans la population cible). La figure (d) représente la situation d'une façon d'estimer $\hat{\theta}$ précise et non biaisée, l'idéal à atteindre.

E. Inférence d'un indicateur à partir d'une estimation

Faire de l'inférence statistique à partir d'une estimation d'un indicateur dans un échantillon, c'est la mettre en rapport avec la valeur réelle inconnue de l'indicateur dans la population cible (**Figure 7**).

Soit $\hat{\theta}$ la valeur de l'estimation d'un indicateur dans un échantillon. On fait de l'inférence statistique en disant : « sous l'hypothèse d'absence de biais d'estimation, il y a des chances pour que la valeur réelle θ de l'indicateur dans la population *cible* soit proche de $\hat{\theta}$ ».

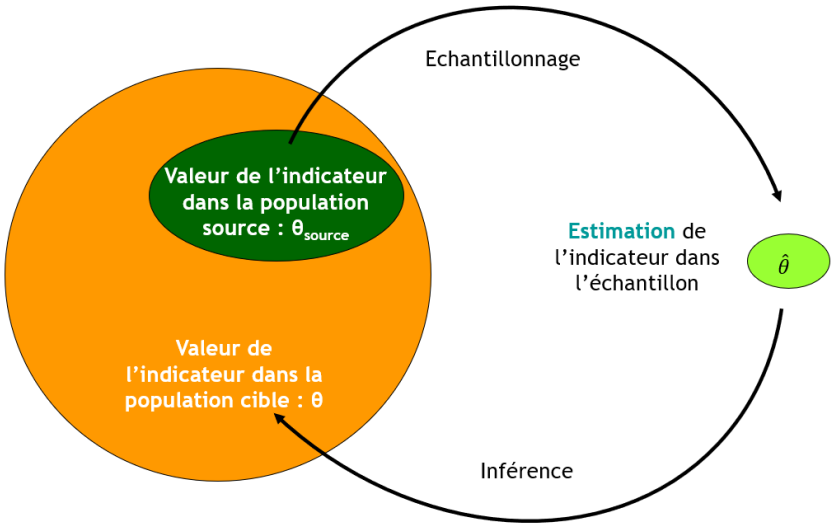


Figure 7. Représentation graphique de l'Inférence.

Attention, même s'il n'y a aucun biais d'estimation, il ne faut pas oublier que la fluctuation d'échantillonnage (la manifestation du hasard) peut conduire à une estimation $\hat{\theta}$ très éloignée de la valeur réelle θ dans la population cible, sans bien entendu que l'on s'en rende compte, puisque θ est inconnue. C'est normal, acceptable, et il faut vivre avec[9].

Je peux cependant vous rassurer en écrivant ceci : sous réserve d'absence de biais d'estimation, plus l'estimation est précise, plus θ a de chances d'être proche de la valeur estimée $\hat{\theta}$. (Alors, rassuré(e) ?! ☺)

F. Intervalle de confiance à 95% d'une estimation

1. *Théorie et interprétation*

Cette théorie ne va pas aller très loin. L'idée ici n'est pas de vous apprendre les statistiques pour que vous deveniez des biostatisticien(ne)s. J'ai donc pris le parti de bien davantage vous apprendre à interpréter les choses qu'à vous apprendre les démonstrations mathématiques / statistiques pour obtenir différentes formules.

Un intervalle de confiance d'une estimation $\hat{\theta}$ est un intervalle dans lequel on peut être confiant dans le fait d'affirmer que la valeur réelle θ dans la population cible se trouve dans cet intervalle. Cette « confiance » doit être quantifiée. Dans la très grande majorité des cas, on fixe ce degré de confiance à 95%. Ainsi, un intervalle de confiance à 95% de $\hat{\theta}$ est l'intervalle dans lequel il y a 95% de chances que la valeur réelle θ dans la population cible s'y trouve[10]. Pour que l'interprétation de l'intervalle de confiance soit complète, il faut ajouter « sous l'hypothèse que l'estimation ne soit pas biaisée par du biais d'estimation ».

Ainsi, l'interprétation de l'intervalle de confiance à 95% de $\hat{\theta}$ $[IC_{inf}\ ;\ IC_{sup}]_{95\%}$ doit être la suivante (en remplaçant ce qui est entre crochets par ce qu'il faut en fonction du contexte de

[9] De la même façon que dans la vie, on a appris à vivre avec cette incertitude, heureusement présente – sinon la vie serait atrocement prévisible et par conséquent tellement ennuyeuse, non ?!

[10] L'interprétation rigoureuse (mais hors programme) d'un intervalle de confiance est un peu plus compliquée que cela : il y a 95% de chances pour que l'intervalle de confiance à 95% comprenne la valeur réelle θ.

l'étude) : « si [$\hat{\theta}$] n'est pas biaisée par du biais d'estimation, il y a 95% de chances pour que la valeur réelle de [θ] dans la population [cible] soit comprise entre [IC_{inf}] et [IC_{sup}] ».

De façon très générale, les bornes de l'intervalle de confiance à 95% de $\hat{\theta}$ se calculent de la façon suivante :

IC_{inf} = $\hat{\theta}$ - 1,96 x $SE_{\hat{\theta}}$

IC_{sup} = $\hat{\theta}$ + 1,96 x $SE_{\hat{\theta}}$

2. *Intervalle de confiance à 95% d'un pourcentage*

Soit $\hat{\pi}$ l'estimation d'un pourcentage dans un échantillon de taille *n*. Je rappelle que la SE de l'estimation $\hat{\pi}$ vaut : $SE_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$.

Si *n* x $\hat{\pi}$ > 5 et si *n* x (1 - $\hat{\pi}$) > 5, alors la formule de l'intervalle de confiance à 95% d'un pourcentage estimé $\hat{\pi}$ est la suivante : $\hat{\pi}$ +/- 1,96 x $SE_{\hat{\pi}}$.

Si *n* x $\hat{\pi} \leq 5$ ou si *n* x (1 - $\hat{\pi}$) ≤ 5 (situation hors programme), alors je vous invite à aller sur un site Internet[11] pour calculer l'intervalle de confiance d'un pourcentage, en lisant les bornes calculées à l'aide de la méthode exacte de Clopper-Pearson.

3. *Intervalle de confiance à 95% d'une moyenne*

Soit $\hat{\mu}$ l'estimation d'une moyenne dans un échantillon de taille *n*. Je rappelle que la SE de la moyenne $\hat{\mu}$ vaut : $SE_{\hat{\mu}} = \frac{SD}{\sqrt{n}}$.

Si *n* > 30 et si la distribution de la variable quantitative peut être considérée comme normale, alors la formule de l'intervalle de confiance à 95% d'une moyenne estimée $\hat{\mu}$ est la suivante : $\hat{\mu}$ +/- 1,96 x $SE_{\hat{\mu}}$.

Si n < 30 (situation hors programme), je vous suggère d'aller sur le même site Internet, mais sur une autre page[12].

4. *Intervalle de confiance à 95% d'une médiane (hors programme)*

Le calcul de l'intervalle de confiance à 95% d'une médiane est hors programme. Il figure néanmoins dans l'Annexe 5 de ce polycopié, au cas où vous en auriez besoin pour votre thèse vétérinaire. Une fois calculé, l'interprétation de cet intervalle de confiance à 95% est la même que celle pour un pourcentage ou une moyenne.

5. *Intervalle de confiance et précision d'une estimation*

Comme on l'a vu ci-dessus, plus la $SE_{\hat{\theta}}$ diminue, plus l'estimation $\hat{\theta}$ est précise. Or, l'intervalle de confiance à 95% d'une estimation $\hat{\theta}$ est : $\hat{\theta}$ +/- 1,96 x $SE_{\hat{\theta}}$. Par conséquent, plus l'intervalle de confiance à 95% d'une estimation est resserré autour de l'estimation, plus cette estimation est précise (cette appréciation est forcément subjective, et en 2ème année, je ne vous demanderai pas de me dire si une estimation est précise ou non).

[11] https://epitools.ausvet.com.au/ciproportion

[12] https://epitools.ausvet.com.au/cimean

IV. INTRODUCTION AUX BASES THEORIQUES DES TESTS STATISTIQUES

Pour comprendre les bases théoriques sur lesquelles reposent tous les tests statistiques du monde[13], je vais commencer par (essayer de) vous montrer que ces bases théoriques reposent sur un raisonnement que nous appliquons tous les jours.

A. Rejet d'une « hypothèse » dans la vie de tous les jours

Voici un exemple de la vie de (plus ou moins) tous les jours : vous souhaitez partir travailler au Japon après vos études vétérinaires, et pour cela, vous devez réussir le JLPT (Japanes-Language Proficiency Test[14]) dont les épreuves sont des QCM. Le jour de l'examen, avant d'entrer dans la salle d'examen, vous croisez Pierre, un autre candidat au JLPT, qui vous dit « je n'ai jamais travaillé mon japonais, je n'ai jamais regardé un film en japonais, je ne sais pas lire le japonais ». Autrement dit, ce que vous dit Pierre, c'est que la durée pendant laquelle Pierre a travaillé son japonais au cours de sa vie est égale à 0 seconde (« durée d'apprentissage du japonais = 0 seconde »). Vous ne connaissez pas Pierre, vous ne savez pas du tout si ce qu'il dit est vrai ou faux, et vous ne le saurez jamais véritablement. Nous allons dire que « Pierre a travaillé son japonais pendant une durée = 0 seconde » est une *hypothèse*. Le jour des résultats, vous recevez un SMS de Pierre[15]. Imaginons deux situations exclusives.

1ère situation : Pierre a réussi son JLPT. Sachant cela, est-ce que vous croyez Pierre quand il vous a dit qu'il n'avait jamais travaillé son japonais ? Vous ne le croyez pas, bien sûr. Pourquoi, parce que s'il n'avait *effectivement* jamais travaillé son japonais de quelque façon que ce soit (autrement dit, si sa durée d'apprentissage du japonais au cours de sa vie était égale à 0 seconde), il n'aurait quasiment eu aucune chance de réussir le JLPT. Bien sûr que c'est *possible* de réussir un QCM sans jamais n'avoir travaillé : grâce à la chance. Mais même si, théoriquement, la réussite au JLPT est possible sans jamais n'avoir travaillé son japonais, vous[16] ne croyez pas Pierre. Là, ce que vous venez de faire, c'est de rejeter l'hypothèse « durée d'apprentissage du japonais = 0 seconde » car si cette hypothèse avait été vraie, Pierre n'aurait *quasiment* eu aucune chance de réussir cet examen JLPT (sauf coup de chance magistral). Autrement dit, l'*observation* « Pierre a réussi son examen JPLT » n'est pas *en accord* avec l'*hypothèse* selon laquelle sa durée d'apprentissage du japonais est égale à 0 seconde. De plus, cette hypothèse (« durée d'apprentissage du japonais = 0 seconde »), vous venez de la rejeter *avec force*, car vous ne le croyez pas *du tout* lorsqu'il vous a dit qu'il n'avait jamais travaillé son japonais.

2ème situation : Pierre n'a pas réussi son JLPT. Sachant cela, est-ce que vous croyez Pierre quand il vous a dit qu'il n'avait jamais travaillé son japonais ? Et bien là, vous êtes dans une situation d'incertitude. Tout d'abord, contrairement à la 1ère situation, vous ne pouvez pas ne pas croire Pierre car vous n'avez pas d'élément fort pour penser qu'il a menti. En effet, dans cette 2ème situation, l'*observation* « Pierre n'a pas réussi son examen JPLT » est *en accord* avec l'*hypothèse* selon laquelle durée d'apprentissage du japonais est égale à 0 seconde. Donc,

[13] Rien qu'ça !

[14] http://www.jlpt.jp/e/index.html

[15] L'histoire ne dit pas comment Pierre a votre numéro de portable

[16] Vos tripes, votre intestin, …

vous allez le croire, mais franchement, sans conviction du tout. En effet, croire Pierre avec conviction dans cette 2ème situation, c'est croire avec conviction que sa durée d'apprentissage du japonais est égale à 0 seconde, pas 1 seconde, pas 2 secondes. Vous imaginez bien que si sa durée d'apprentissage du japonais au cours de sa vie avait été de 20 secondes, le résultat au JLPT aurait été identique ! Donc, vous acceptez l'hypothèse « durée d'apprentissage du japonais = 0 seconde », sans aucune conviction.

La 1ère situation que je viens de décrire est une situation très classique, une situation qui fait que l'on ne croit pas en quelque chose (que l'on va appeler « hypothèse ») que l'on nous dit être vrai, si ce que l'on *observe* n'est pas *en accord* avec ce quelque chose (⇔ cette hypothèse). Et bien, croyez-moi, tous les tests statistiques du monde reposent sur ce comportement que l'on a de façon instinctive. De là à vous dire que la stat' est quelque chose de tout à fait instinctif, il n'y a qu'un pas[17].

Je vais revenir très souvent, dans la partie V sur cet exemple ci-dessus. Vous devez l'avoir compris pour comprendre le reste…

B. « Différence d'indicateurs » synonyme d' « association statistique »

Ce point est absolument fondamental, et nous ne cesserons de l'aborder en TD : « lorsque l'on compare les valeurs d'un indicateur[18] entre deux (ou plusieurs) groupes d'individus, l'égalité de ces indicateurs traduit l'*absence* d'association statistique entre les deux caractères étudiés. »

Je vais prendre plusieurs exemples pour illustrer ce propos.

Dans la population humaine adulte, il n'y a aucune association statistique (ou bien, « il n'y a aucun lien », ou bien encore « il y a parfaite indépendance ») entre la couleur préférée (« rouge » *versus* « pas rouge ») et la taille des personnes. Cela se traduit par le fait que, dans la population humaine adulte, la moyenne de la taille des millions de personnes dont la couleur préférée est le rouge est *égale* à la moyenne de la taille des millions de personnes dont la couleur préférée n'est pas le rouge.

Dans la population humaine adulte, il y a une association statistique (ou bien, « il y a un lien », ou bien encore « il y a non indépendance ») entre le poids des personnes et le fait que les personnes soient des hommes ou des femmes. Cela se traduit par le fait que, dans la population humaine adulte, la moyenne du poids des millions d'hommes est *différente* de la moyenne du poids des millions de femmes.

Dans la population humaine adulte, il y a une association statistique entre le fait d'avoir fumé pendant plus de dix ans des cigarettes et la présence de cancer du poumon. Cela se traduit par le fait que, dans la population humaine adulte, le pourcentage de personnes avec un cancer du poumon parmi les millions de personnes qui ont fumé des cigarettes pendant plus de dix ans est *différent* du pourcentage de personnes avec un cancer du poumon parmi les millions de personnes qui n'ont pas fumé de cigarettes pendant plus de 10 ans.

[17] Que je me garderais bien de franchir ☺

[18] A toujours garder en tête : indicateur = moyenne, médiane, ou pourcentage (entre autres ; ceux-là sont les trois principaux)

Dans un échantillon d'une étude de 46 personnes, si l'on observe que le pourcentage de personnes avec un cancer du poumon parmi les 16 personnes qui ont fumé des cigarettes pendant plus de dix ans (par exemple 18%) est égal au pourcentage de personnes avec un cancer du poumon parmi les 30 personnes qui n'ont pas fumé de cigarettes pendant plus de 10 ans (18% aussi[19]), nous dirons que, *dans l'échantillon*, il n'y avait pas d'association statistique entre le fait d'avoir fumé pendant plus de dix ans des cigarettes et la présence de cancer du poumon.

C. Remarque préliminaire avant d'entrer dans le vif du sujet

Dans toute la suite de cette partie IV et dans la partie V « Bases théoriques des tests statistiques » suivante, je vais me restreindre à un contexte particulier d'utilisation des tests statistiques : celui visant à montrer qu'il existe une association statistique entre deux caractères dans la population cible, dont l'un des deux est binaire, à partir des données d'un échantillon. Ce contexte particulier, mais pour autant très fréquent, est donc celui où l'on a deux groupes d'individus dont on cherche à montrer, au niveau de la population cible, qu'ils sont différents sur la valeur d'un indicateur (moyenne, médiane, ou pourcentage).

D. Question à laquelle répond un test statistique

Prenons l'exemple suivant. Supposons que l'on veuille montrer qu'il existe une différence réelle de production laitière quotidienne moyenne entre le groupe des (millions de) vaches Charolaises primipares et celui des (millions de) vaches Charolaises multipares. Pour répondre à cette question, on met en place une étude constituée d'un échantillon de 110 vaches, comprenant 63 vaches Charolaises primipares et 47 vaches Charolaises multipares, et on calcule (estime) la moyenne de la production laitière quotidienne *observée* dans le groupe des 63 vaches primipares de l'échantillon et celle *observée* dans le groupe des 47 vaches multipares.

Supposons que la différence entre la moyenne de production laitière quotidienne estimée dans le groupe des vaches primipares et celle estimée dans le groupe des vaches multipares soit égale à 2,5 kg/j. Ce résultat obtenu à partir d'une étude comprenant 110 vaches vous permet-il de montrer ce que l'on souhaitait montrer, à savoir « qu'il existe une différence réelle de production laitière quotidienne moyenne entre le groupe des (millions de) vaches Charolaises primipares et celui des (millions de) vaches Charolaises multipares. » ? Bien sûr que non ! En effet, car si en vrai, dans la population cible constituée de millions de vaches Charolaises, la moyenne de la production laitière quotidienne des vaches primipares était *égale* à celle des vaches multipares, il aurait été tout à fait possible, par le simple fait du hasard (de la fluctuation d'échantillonnage) d'observer, dans l'étude constituée de 110 vaches, une telle différence de 2,5 kg/j entre les deux groupes de vaches !

Vous devez ainsi prendre conscience de la problématique suivante : ce n'est pas parce que l'on observe, dans *un échantillon*, une différence numérique d'indicateurs que cette différence numérique d'indicateurs existe réellement dans la *population cible*. Car le hasard

[19] Observation qui est tout à fait possible – la fluctuation d'échantillonnage (le hasard) peut nous faire observer des indicateurs très éloignés de ceux de la population.

seul (via la fluctuation d'échantillonnage) peut conduire à une différence de valeurs d'un indicateur, parfois importante, entre deux groupes d'individus dans un échantillon.

Il n'y a qu'une seule façon d'utiliser les données d'un échantillon pour ensuite faire de l'inférence (étendre les résultats de l'échantillon à la population cible), c'est de réaliser un test statistique. En effet, le test statistique prend en compte le fait que la différence numérique d'indicateurs qui a été observée dans un échantillon peut provenir en totalité ou de façon partielle, du hasard via la fluctuation d'échantillonnage.

V. BASES THEORIQUES DES TESTS STATISTIQUES

A. Base de la démarche d'un test statistique

Revenons sur Pierre et son examen de Japonais. Qu'est-ce qui vous a fait fortement penser que l'hypothèse « sa durée d'apprentissage du japonais = 0 seconde » était fausse (⇔ rejet de l'hypothèse) dans la 1ère situation où Pierre a réussi son examen ? C'est parce que si Pierre n'avait *effectivement* jamais travaillé son japonais, il n'y aurait eu *quasiment* aucune chance pour qu'il réussisse son examen de japonais. Autrement dit, la probabilité d'observer l'événement qui a été observé dans la 1ère situation (« Pierre a réussi son examen ») sous l'hypothèse que « sa durée d'apprentissage du japonais = 0 seconde » est une probabilité jugée comme très faible.

Dans la 2ème situation, on ne pouvait pas fortement penser que l'hypothèse « sa durée d'apprentissage du japonais = 0 seconde » était fausse. Pourquoi ? Parce que la probabilité d'observer l'événement qui a été observé dans la 2ème situation (« Pierre n'a pas réussi son examen ») sous l'hypothèse que « sa durée d'apprentissage du japonais = 0 seconde » est une probabilité qui n'était pas jugée comme très faible (l'observation est en accord avec l'hypothèse). Mais pour autant, on ne pouvait pas non plus fortement penser que l'hypothèse « sa durée d'apprentissage du japonais = 0 seconde » était vraie ! Pourquoi ? Parce s'il avait travaillé 1, 2, …, 10, voire 50 secondes de japonais dans toute sa vie, le résultat à son examen de JLPT aurait été identique !

La démarche globale d'un test statistique est la suivante : pour penser avec force ou conviction que « quelque chose » existe au niveau de la population, il faut que l'observation d'un résultat dans l'échantillon ne soit pas en accord avec l'inverse de ce « quelque chose ». Autrement dit, il faut que l'observation d'un résultat dans une étude ait eu très peu de chances de se produire si l'inverse de ce « quelque chose » avait été vrai au niveau de la population. C'est du raisonnement par l'absurde !

B. Notations

Dans tout le reste de ce polycopié, je vais noter « θ_A » la valeur de l'indicateur, inconnue, parmi les individus du groupe A dans la population cible, « θ_B » la valeur de l'indicateur, inconnue, parmi les individus du groupe B dans la population cible, et « Δ » la valeur, inconnue, de la différence entre θ_A et θ_B ($\Leftrightarrow \Delta = \theta_A - \theta_B$). Soit $\Delta = 0$ (et auquel cas, il n'y a aucune différence réelle sur l'indicateur θ entre les individus du groupe A et ceux du groupe B dans la population cible), soit $\Delta \neq 0$. Mais une chose est certaine, Δ est inconnue 😊.

C. L'hypothèse nulle

L'hypothèse nulle (H_0) est l'hypothèse que l'on souhaite rejeter avec force, pour penser avec force que son alternative est vraie. Par exemple, si l'on souhaite penser avec force que, dans la population des chiens souffrant de PU-PD[20], les chiens traités avec un traitement A ont un taux de survie *différent* des chiens traités avec un traitement B, l'hypothèse nulle H_0 sera

[20] PU-PD = polyurie-polydipsie

« dans la population des chiens souffrant de PU-PD, le taux de survie parmi les chiens traités avec le traitement A est *identique* à celui parmi les chiens traités avec le traitement B ».

Un point très important : l'hypothèse nulle H_0 porte sur la population *cible*. En effet, un test statistique a pour objectif de donner une information sur ce qu'il se passe au niveau de la *population* (par exemple, « est-ce que, au niveau de la population, les deux groupes diffèrent sur la valeur d'un indicateur ? »). **Il n'y a pas besoin d'un test statistique pour uniquement savoir si*, dans un échantillon*, les groupes diffèrent sur la valeur d'un indicateur (vos yeux suffisent) !**

Ainsi, l'hypothèse nulle H_0 d'un test statistique est, dans la très grande majorité des cas : « dans la population [cible], la valeur réelle [de l'indicateur parmi les individus du groupe A (θ_A)] est *égale* à celle [parmi les individus du groupe B (θ_B)] ». On peut aussi écrire H_0 ainsi : « dans la population [cible], la *différence* réelle sur [la valeur de l'indicateur] entre [les individus du groupe A (θ_A) et les individus du groupe B (θ_B)] est égale à 0 ». H_0 peut aussi s'écrire très (trop) succinctement ainsi : « $\Delta = 0$ »[21]. On peut enfin écrire H_0 ainsi : « dans la population [cible], il n'existe pas d'association réelle entre [le fait d'appartenir au groupe A ou B] et [le caractère étudié] ».

Ce que je viens d'écrire peut tout à fait être étendu à la situation où il y a plus que deux groupes comparés (nous verrons cela en TD). Enfin, il existe des tests statistiques qui ne comparent pas deux groupes (ni trois, ni quatre, …) et que nous utiliserons aussi. Je vous fournirai alors spécifiquement l'hypothèse nulle H_0 pour ces tests statistiques.

Enfin, dans certaines situations (situations qui sont hors programme du module de Biostatistique en Médecine Vétérinaire), il arrive que l'on veuille penser avec force que, *dans la population*, la valeur d'un indicateur dans un groupe A est *égale* à celle dans un groupe B. A ce moment-là, l'hypothèse nulle H_0 devra être : « dans la population cible, la valeur réelle de l'indicateur parmi les individus du groupe A (θ_A) est *différente* de celle parmi les individus du groupe B (θ_B) ». Cette situation est celle des études cliniques d'équivalence[22].

D. Rejeter ou ne pas rejeter l'hypothèse nulle ?

Que signifie « rejeter H_0 » ? Cela signifie « rejeter l'hypothèse que, dans la population, la valeur réelle de l'indicateur dans le groupe A (θ_A) est égale à celle dans le groupe B (θ_B) ».

Comment faire pour rejeter H_0 ? Il faudra que la différence observée dans l'étude, notée d_{obs}, entre les deux indicateurs $\hat{\theta}_A$ et $\hat{\theta}_B$ estimés dans l'échantillon de l'étude soit un événement rare en faisant l'hypothèse que H_0 est vraie ($\Leftrightarrow$ « sous H_0 »). Autrement dit, il faudra que d_{obs} soit un événement rare sous l'hypothèse qu'en vrai, dans la population cible, les deux groupes A et B ne diffèrent strictement pas sur l'indicateur θ.

Alors, c'est un chtouille plus compliqué que cela. En fait, il faudra que l'événement « observer la valeur $|d_{obs}|$ ou une valeur plus élevée que $|d_{obs}|$ » soit un événement rare sous H_0. Si oui,

[21] Et là, j'espère que vous voyez l'analogie avec l'examen de japonais de Pierre (l'hypothèse était : « durée d'apprentissage du japonais = 0 seconde »)

[22] Référence à lire sur le sujet s'il vous intéresse : Jones, B., Jarvis, P., Lewis, J.A. and Ebbutt, A.F., 1996. Trials to assess equivalence: the importance of rigorous methods. BMJ. 313, 36-9

alors H_0 pourra être rejetée. Si ce n'est pas le cas, H_0 ne pourra pas être rejetée. Dans ce cas, on dira que H_0 est acceptée (acceptation par défaut) parce qu'elle n'a pas réussi à être rejetée.

Comment savoir si l'événement « observer la valeur $|d_{obs}|$ ou une valeur plus élevée que $|d_{obs}|$ » est un événement rare ou pas sous H_0 ? C'est en calculant la probabilité que cet événement soit observé sous H_0. Et cette probabilité s'appelle … « le degré de signification ».

Nous allons voir par la suite avec les risques d'erreur de 1ère et 2ème espèces que l'on ne peut que *rejeter* avec force l'hypothèse nulle H_0. On ne peut surtout pas *accepter* H_0 avec force (nous avions cependant déjà abordé ce point avec la 2ème situation de l'exemple de l'examen de japonais pour Pierre).

E. Le degré de signification

1. *Définition du degré de signification*

En mathématique / statistique, le degré de signification (noté *p* en italique) = Pr(observer une |différence| ≥ $|d_{obs}|$, sous H_0). En français, c'est plus compliqué, mais plus important à comprendre (car dans la vie de tous les jours, on communique en français, pas en math' !) : « le degré de signification *p* est la probabilité d'observer une différence en valeur absolue au moins égale à celle qui a été observée dans l'échantillon (d_{obs}) sous l'hypothèse qu'en vrai, il n'existe aucune différence réelle sur la valeur de l'indicateur étudié entre les deux groupes étudiés A et B dans la population cible ».

Autrement dit, si en vrai il n'y avait *aucune* différence réelle au niveau de la population entre les deux groupes A et B sur l'indicateur étudié ($\Leftrightarrow \Delta = 0$), il y aurait eu *p* % de chances d'observer une différence en valeur absolue au moins égale à celle que l'on a observée dans l'échantillon (d_{obs}). J'ai exceptionnellement souligné la première partie de cette phrase, car c'est (entre autres) son omission dans la tête de nombreux chercheurs qui génère les erreurs d'interprétation du degré de signification.

Ainsi, plus le degré de signification est faible, plus la différence qui a été observée fait partie des événements rares sous H_0.

2. *Illustration de l'interprétation du degré de signification sur un exemple*

Illustrons ces concepts assez théoriques à la réalité de terrain, ce à quoi vous serez confrontés (en TD, en examen, dans la suite de votre cursus, à la sortie de l'école, en formation continue, en conférence, …). La **Figure 8** ci-dessous est un extrait d'un tableau d'un article[23] présentant les résultats d'une étude dont l'objectif était d'identifier les facteurs de risque de développement d'une maladie respiratoire chez les chiens et les chats adultes (les auteurs avaient deux populations cibles, les chiens et les chats adultes, mais pour simplifier les choses, j'ai supprimé dans la **Figure 8** la colonne des chats). Les colonnes C1 et C2 présentent, sur la première ligne, la moyenne de l'âge (± SD) parmi les 83 chiens avec maladie respiratoire (9,8 ans) et parmi les 38 chiens sans maladie respiratoire (7,7 ans). La différence observée dans l'échantillon entre ces deux moyennes d_{obs} est donc égale à 9,8 – 7,7 = 2,1 ans.

[23] Lin, C.H., Lo, P.Y., Wu, H.D., Chang, C. and Wang, L.C., 2018. Association between indoor air pollution and respiratory disease in companion dogs and cats. J Vet Intern Med. 32, 1259-1267.

TABLE 1 Baseline characteristics of dogs and cats with and without respiratory disease, proportion of existence of selected household air pollutants, and household PM2.5 measurements

	Dogs			Cats		
Variable	Respiratory group (n = 83)	Control group (n = 38)	P	Respiratory group (n = 64)	Control group (n = 17)	P
Age (years)	9.8 ± 3.3	7.7 ± 4.3	.0092	8.0 (1–19)	5.0 (1–14)	.063
Sex (males)	51.8 (43/83)	44.7 (17/38)	.47	57.8 (37/64)	52.9 (9/17)	.72

C1 C2 C3

Figure 8. Extrait du tableau 1 de l'article de Lin, JVIM, 2018. Les « Ci » représentent les colonnes n°i (i ∈ {1, 2, 3}).

L'hypothèse nulle H_0 du test statistique dont le degré de signification vaut 0,0092 (cf. colonne C3) est la suivante : « dans la population des (millions de) chiens adultes, la moyenne de l'âge des chiens avec maladie respiratoire est égale à celle des chiens sans maladie respiratoire ». La valeur du degré de signification testant ces deux moyennes est 0,0092 (cf. colonne C3). Comment interprète-t-on cette valeur[24] ? De la façon suivante : « la probabilité d'observer une différence de moyennes d'âge des chiens en valeur absolue au moins égale à celle qui a été observée dans l'échantillon (d_{obs}) et qui vaut 2,1 ans, sous l'hypothèse que, dans la population des (millions de) chiens adultes, la moyenne de l'âge des chiens avec maladie respiratoire est égale à celle des chiens sans maladie respiratoire, est de 0,0092 ».

Au cas où vous vous poseriez la question, il n'est pas possible de raccourcir cette phrase. Et si vous le faites, elle deviendra fausse ☺.

3. *Le sacro-saint seuil de 0,05 et notion de « différence significative »*

Revenons sur Pierre et son examen de japonais. Dans la 1ère situation (Pierre avait réussi son examen), vous aviez rejeté l'hypothèse selon laquelle Pierre n'avait jamais travaillé son japonais parce que, si cela avait été le cas, il n'aurait eu quasiment aucune chance de réussir son examen (l'événement « réussir son examen » qui a été *observé* dans la 1ère situation est un événement *jugé* comme rare sous l'hypothèse de « n'avoir jamais travaillé son japonais »). Autrement dit, la probabilité de réussir un examen (ce qui a été observé dans la 1ère situation) sous l'hypothèse de ne jamais n'avoir travaillé de japonais a été *jugée* comme faible (mais cependant non nulle, car il est quand même possible de répondre par hasard correctement à des questions de QCM !). Cette probabilité est en quelque sorte un « degré de signification ».

Ce qui a été *jugé* dans l'exemple de Pierre est *calculé* dans les tests statistiques (cette probabilité d'observer ce qui a été observé sous une certaine hypothèse) : c'est le degré de signification d'un test statistique.

A partir de quelle valeur de probabilité peut-on dire qu'un événement est *rare* sous une hypothèse ?

Le seuil consenti par toute la communauté scientifique (ou presque[25]) est 0,05 (soit 5%). Par conséquent, il est admis par la communauté scientifique que si[26] la probabilité d'observer une différence en valeur absolue au moins égale à celle que l'on vient d'observer sous l'hypothèse

[24] Question qui peut tomber à l'examen de janvier ;-)

[25] Référence à lire sur le sujet s'il vous intéresse : Benjamin, D.J., Berger, J.O., Johannesson, M., et al., 2018. Redefine statistical significance. Nat Hum Behav. 2, 6-10

[26] Attention, prenez votre respiration. Mais sachez que vous devez parfaitement *comprendre* cette phrase car vous devrez être capable de la ressortir en examen. Si vous ne comptez que l'*apprendre* par cœur, ça risque de coincer.

qu'en vrai, il n'y a aucune différence réelle sur la valeur de l'indicateur étudié entre les deux groupes étudiés A et B dans la population cible, est inférieure ou égale à ce seuil de 0,05, alors on peut rejeter cette hypothèse (l'hypothèse qu'en vrai, il n'y a aucune différence réelle sur la valeur de l'indicateur étudié entre les deux groupes étudiés A et B dans la population cible). En rejetant cette hypothèse (l'hypothèse nulle H_0), on arrive à ce à quoi on voulait arriver : penser avec force que les deux groupes A et B, dans la population, *diffèrent* sur l'indicateur θ.

Ainsi, lorsque la valeur du degré de signification est inférieure ou égale à ce sacro-saint seuil de 0,05, on dit que la différence observée dans l'échantillon d_{obs} est « significative » (et l'on rejettera ensuite H_0 lorsque l'on souhaitera faire de l'inférence). En reprenant l'exemple de la **Figure 8**, la différence observée de 2,1 ans (9,8 – 7,7) était « significative » car le degré de signification du test statistique qui teste cette valeur de 2,1 ans est inférieur à 0,05 (0,0092 < 0,05).

F. Le risque d'erreur de 1ère espèce (α)

1. *Introduction*

Revenons (encore) sur Pierre et son examen de japonais. Dans la 1ère situation (Pierre avait réussi son examen), ce qui nous avait fait rejeter l'hypothèse selon laquelle Pierre n'avait jamais travaillé son japonais était que, s'il n'avait effectivement jamais travaillé son japonais, il n'aurait eu quasiment aucune chance de réussir son examen.

Mais il n'est pas *impossible* que Pierre ait eu raison lorsqu'il vous avait dit « je n'ai jamais travaillé mon japonais » ! Il est en effet possible qu'il ait coché les cases du QCM correctement par le simple fait du hasard. Par conséquent, dans la 1ère situation, vous vous rendez compte que vous avez peut-être commis une erreur en rejetant l'hypothèse selon laquelle « Pierre n'a jamais travaillé son japonais ».

Si l'on revient sur l'exemple de la **Figure 8**, l'hypothèse nulle H_0 selon laquelle « dans la population des chiens adultes, la moyenne de l'âge des chiens avec maladie respiratoire est égale à celle des chiens sans maladie respiratoire » avait été rejetée, car le degré de signification de 0,0092 était inférieur à 0,05. Mais là encore, il n'est pas *impossible* que dans la population des chiens adultes, la moyenne de l'âge des chiens avec maladie respiratoire soit *égale* à celle des chiens sans maladie respiratoire, et que ce soit donc uniquement la fluctuation d'échantillonnage (le hasard), qui nous fasse penser fortement (à tort) au fait qu'il existe une différence réelle de moyennes d'âge dans la population des chiens adultes, entre ceux souffrant d'une maladie respiratoire et les autres.

Par conséquent, vous prenez conscience (si ça n'avait pas déjà été le cas auparavant) que le rejet de H_0 peut avoir été réalisé *à tort*.

2. *Définition du risque d'erreur de 1ère espèce α*

La définition du risque d'erreur de 1ère espèce (α) est la suivante : c'est la probabilité de rejeter H_0 lorsque H_0 est vraie. Je ne vais pas vous faire la démonstration de ce qui suit, mais sachez qu'en fixant un seuil de 0,05 pour le degré de signification pour savoir si la différence observée d_{obs} est significative ou non, le risque d'erreur de 1ère espèce α est alors de fait fixé à cette valeur de 0,05. Si on avait choisi 0,01 comme seuil de significativité (c'est-à-dire que l'on rejette H_0 si $p \leq 0,01$), le risque d'erreur de 1ère espèce α aurait alors valu 0,01. Au passage, au

cas où vous vous poseriez la question, il est interdit de fixer la valeur de α *après* avoir vu la valeur du degré de signification dans l'échantillon…

Dans tout le module de Biostatistique en Médecine Vétérinaire, ainsi que dans probablement plus de 99% des articles scientifiques utilisant des tests statistiques, α est fixé à la valeur de 0,05.

3. *Mauvaise interprétation du risque d'erreur de 1ère espèce α*

Beaucoup de scientifiques pensent que α est la probabilité de se tromper en rejetant H_0, ou, autrement dit, que α est l'erreur que l'on commet en rejetant H_0. C'est faux. Le risque d'erreur de 1ère espèce α n'est que « la probabilité de rejeter H_0 *lorsque H_0 est vraie* ». La probabilité de se tromper en rejetant H_0, c'est : « la probabilité que H_0 soit vraie *lorsque l'on rejette H_0* ». Et cette phrase que je viens de mettre entre guillemet n'est pas du tout la même[27] que la définition du risque d'erreur de 1ère espèce α !

Et cette probabilité de se tromper en rejetant H_0, c'est une probabilité que l'on peut calculer, mais le raisonnement est trop compliqué pour être au programme du module de Biostatistique en Médecine Vétérinaire[28]. Cette probabilité peut être bien supérieure à 0,05. Elle dépend entre autres de la puissance statistique de l'étude (cf. plus loin) et du caractère exploratoire ou « confirmatoire » de l'étude :

- Lorsque la puissance statistique d'une étude est faible, la probabilité que les auteurs se trompent lorsqu'ils rejettent H_0 (c'est-à-dire, lorsque $p \leq 0,05$) est grande.
- Lorsqu'une étude qui obtient un résultat significatif ($p \leq 0,05$) est la première à l'obtenir, leurs auteurs ont une plus grande probabilité de se tromper en rejetant H_0 que si leur étude confirme un résultat que des études antérieures avaient déjà montré.

C'est la raison pour laquelle il ne faut pas être trop convaincu lorsque l'on rejette H_0, et ce d'autant plus que l'on a été les premiers au monde à l'avoir fait !

G. Le risque d'erreur de 2ème espèce (β)

1. *Introduction*

Revenons (encore et toujours) sur Pierre et son examen de japonais. Dans la 2ème situation (Pierre n'avait pas réussi son examen), ce qui nous avait fait accepter l'hypothèse selon laquelle Pierre n'avait jamais travaillé son japonais était que, s'il n'avait effectivement jamais travaillé son japonais, il aurait eu effectivement de bonnes chances de ne pas réussir son examen.

Mais il est cependant tout à fait possible que Pierre vous ait menti lorsqu'il vous avait dit « je n'ai jamais travaillé mon japonais » ! Ce n'est en effet pas parce que l'observation « Pierre n'a pas réussi son examen » est en accord avec l'hypothèse selon laquelle « il n'a jamais travaillé son japonais » que cette observation de non réussite à l'examen apporte une quelconque preuve que Pierre n'a effectivement *jamais* travaillé son japonais. En effet, s'il avait travaillé

[27] Regardez ce qui précède et ce qui suit le mot « lorsque » dans les deux expressions 😉.

[28] Référence à lire sur le sujet s'il vous intéresse : Desquilbet, L., 2020. Enhancing Clinical Decision-Making: Challenges of making decisions on the basis of significant statistical associations. J Am Vet Med Assoc. 256, 187-193

son japonais 50 secondes (et aurait abandonné tout de suite après), il vous aurait menti, et vous auriez eu tort de le croire en disant que sa durée d'apprentissage du japonais était égale à 0 seconde. Attention, je vous vois venir ! Vous allez penser « 0 seconde, ou 50 secondes, c'est pareil ! ». Surtout pas en recherche clinique : « pas d'effet d'un traitement » n'est pas synonyme en médecine de « ce traitement a un tout petit effet ».

Par conséquent, vous vous rendez compte que vous avez peut-être commis une erreur en ayant cru Pierre, c'est-à-dire en acceptant l'hypothèse selon laquelle « Pierre n'a jamais travaillé son japonais ».

Revenons sur la **Figure 8**. Regardez maintenant la ligne correspondant au sexe des chiens. On peut lire que le pourcentage de chiens mâles *parmi* les 83 chiens avec maladie respiratoire est de 51,8% et celui de chiens mâles *parmi* les 38 chiens sans maladie respiratoire est de 44,7%. Le test statistique testant ces deux pourcentages fournit un degré de signification *p* égal à 0,47. L'hypothèse nulle H_0 de ce test statistique était : « dans la population des (millions de) chiens adultes, le pourcentage de chiens mâles *parmi* les chiens avec maladie respiratoire est *égal* au pourcentage de chiens mâles *parmi* les chiens sans maladie respiratoire ». Comme *p* est supérieur à 0,05, la différence entre ces deux pourcentages (51,8% *versus* 44,7%) n'est pas significative. Les auteurs vont donc accepter H_0 (parce qu'ils ne peuvent pas la rejeter). Mais en acceptant H_0, ils peuvent tout à fait se tromper en disant qu'il n'existe donc pas de différence *réelle* de pourcentages de chiens mâles entre celui parmi les chiens avec maladie respiratoire et celui parmi les chiens sans maladie respiratoire, dans la population des chiens adultes. Il peut en effet réellement exister une différence entre ces deux pourcentages dans la population des chiens adultes, mais dans leur étude, leur différence de pourcentages était malheureusement non significative. Ils ont donc peut-être accepté (parce qu'ils n'avaient aucun moyen de la rejeter) cette hypothèse nulle H_0 à tort.

2. *Définition du risque d'erreur de 2ème espèce β*

La définition du risque d'erreur de 2ème espèce (β) est la suivante : c'est la probabilité d'accepter H_0 quand H_0 est fausse[29] et que la différence réelle vaut Δ ($\Delta \neq 0$).

3. *Lien entre β et Δ*

Petit commentaire, comme ça, en passant. Autant la situation correspondant à « H_0 est vraie » est unique ($\Delta = 0$), autant celle correspondant à « H_0 est fausse et que la différence réelle vaut Δ ($\Delta \neq 0$) » est multiple !

En effet, si $\Delta = 1$, H_0 est fausse, si $\Delta = 2$, elle fausse aussi, etc. Et vous pouvez intuiter[30] que la probabilité d'accepter H_0 quand H_0 est fausse (la valeur de β) ne vaudra pas la même valeur selon que Δ vaut 1, 2, ou 3, etc…

Par conséquent, cette probabilité β dépend de la valeur de Δ (inconnue), la différence réelle entre les deux indicateurs θ_A et θ_B, dans la population cible.

[29] Ecrire « H_0 est fausse » est strictement équivalent à écrire « il existe une réelle différence Δ ($\Delta \neq 0$) entre la valeur de l'indicateur dans le groupe A et celle dans le groupe B, au niveau de la population ».

[30] Et si ce n'est pas le cas, ce n'est pas grave.

4. *Mises en garde et conséquences*

Le risque d'erreur de 2ème espèce (β) est plus difficile à appréhender que le risque d'erreur de 1ère espèce α, mais son implication dans la vie (scientifique) de tous les jours est *beaucoup* plus importante, et sa méconnaissance conduit à de vraies et belles erreurs de communication scientifique. C'est d'ailleurs l'origine de la citation sur la page de couverture de ce polycopié !

J'ai écrit ci-dessus que β dépend de la valeur de Δ. Or, Δ est inconnue car Δ est la valeur de la différence réelle dans la population cible entre θ_A et θ_B, les valeurs des deux indicateurs dans les groupes A et B de la population cible qui sont inconnues (cf. partie III.A.4, page 14). Donc la valeur du risque d'erreur de 2ème espèce β est inconnue.

Alors, tout comme le risque d'erreur de 1ère espèce α ne quantifie pas la probabilité de se tromper lorsque l'on rejette H_0, le risque d'erreur de 2ème espèce β ne quantifie pas non plus l'erreur que l'on commet lorsque l'on accepte H_0. En revanche, ce que je peux vous dire, c'est que la probabilité de se tromper lorsque l'on accepte H_0 *dépend* de β. Or, puisque β est inconnue, la probabilité de se tromper quand on accepte H_0 est inconnue. Donc, elle est potentiellement très grande !! Par conséquent, le corollaire, excessivement important pour un vétérinaire (tout comme pour un médecin en médecine humaine) qui souhaite ne pas se faire avoir en lisant une communication scientifique, est le suivant : **lorsque H_0 n'a pas pu être rejetée ($\Leftrightarrow p > 0,05$), il n'est pas possible de dire ou de penser que H_0 est probablement vraie. Autrement dit, il est interdit de dire ou de penser que les deux indicateurs θ_A et θ_B sont probablement égaux (ou voisins) en vrai.** Pour paraphraser un auteur : « absence de preuve n'est pas preuve d'absence »[31]. Ainsi, penser, écrire ou dire « $p > 0,05$, donc les groupes A et B dans la population cible sont similaires, ou comparables, sur l'indicateur étudié » est **FAUX**. C'est l'erreur qu'ont commise les auteurs de l'étude[32] de la **Figure 9** en concluant à une absence de différence réelle (« confers no clinical benefit ») à partir d'une différence non significative observée dans l'échantillon de l'étude.

[31] Altman, D.G. and Bland, J.M., 1996. Absence of evidence is not evidence of absence. Aust Vet J. 74, 311

[32] Werner, M., Suchodolski, J.S., Straubinger, R.K., et al., 2020. Effect of amoxicillin-clavulanic acid on clinical scores, intestinal microbiome, and amoxicillin-resistant Escherichia coli in dogs with uncomplicated acute diarrhea. J Vet Intern Med. 34, 1166-1176

Received: 29 July 2019 | Accepted: 23 March 2020
DOI: 10.1111/jvim.15775

STANDARD ARTICLE

Journal of Veterinary Internal Medicine Open Access ACVIM American College of Veterinary Internal Medicine

Check for updates

Effect of amoxicillin-clavulanic acid on clinical scores, intestinal microbiome, and amoxicillin-resistant *Escherichia coli* in dogs with uncomplicated acute diarrhea

Conclusions and Clinical Importance: Our study suggests that treatment with amoxicillin-clavulanic acid confers no clinical benefit to dogs with AD, but predisposes the development of amoxicillin-resistant *E. coli*, which persist for as long as 3 weeks after treatment. These findings support international guideline recommendations that dogs with diarrhea should not be treated with antimicrobials unless there are signs of sepsis.

Figure 9. Exemple d'une belle erreur de communication scientifique dans l'un des « meilleurs » journaux de recherche clinique vétérinaire issu d'un article publié en mars 2020 (« Our study suggests that treatment with amoxicillin-clavulanic acid **confers no clinical benefit** to dogs with [acute diarrhea] » à partir d'une différence non significative).

Par conséquent, dans la suite de ce polycopié, dès qu'H_0 a été acceptée (parce qu'elle n'a pas été rejetée), aucune inférence ne pourra être faite, et nous nous limiterons à dire que « H_0 n'a pas été rejetée ».

La **Figure 10** ci-dessous provient de l'article de Amrhein[33], qui illustre le fait que dans une bonne partie (51%) des articles scientifiques, une différence non significative est interprétée à tort comme une absence de différence réelle.

[33] Amrhein, V., Greenland, S. and McShane, B., 2019. Scientists rise up against statistical significance. Nature. 567, 305-307

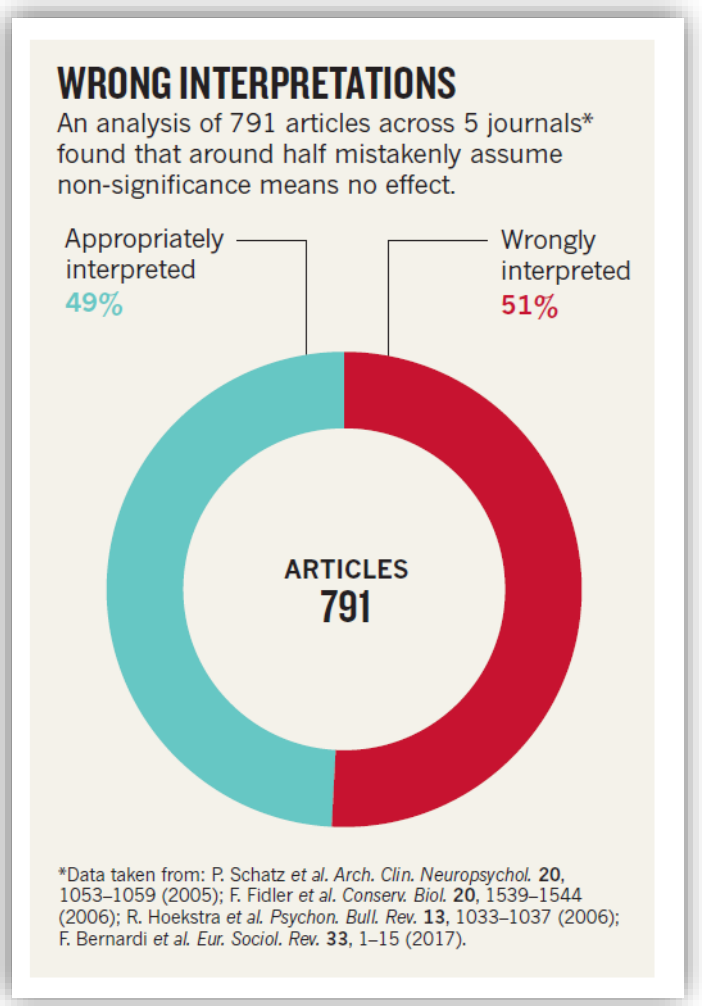


Figure 10. Figure issue de l'article de Amrhein, Nature, 2019, montrant que 51% des articles de recherche concluent de façon erronée à partir d'une différence non significative en prétendant qu'il n'existe donc pas de différence réelle.

H. Différence entre les termes « statistique » et « significatif »

Nous avons déjà vu la définition d' « association statistique » (cf. page 26). Ne faites surtout pas l'amalgame entre « association statistique » et « association significative » !

Par exemple, si l'on reprend la **Figure 8** (page 32), à la ligne de l'âge des chiens, il existait une différence *numérique* de moyennes d'âge entre les chiens avec (9,8 ans) et sans (7,7 ans) maladie respiratoire. Donc, dans l'échantillon, il existait une association statistique entre l'âge des chiens et la présence d'une maladie respiratoire. Et comme la différence numérique entre ces deux moyennes d'âge était significative (car $p \leq 0{,}05$), cette association statistique était donc significative. Ainsi, dans cette étude, il existait une association statistique significative entre l'âge des chiens et la présence d'une maladie respiratoire.

Restons sur cette **Figure 8**, mais regardons la ligne concernant le sexe des chiens. Il existait une différence *numérique* de pourcentages de mâles entre celui parmi les chiens avec maladie respiratoire (51,8%) et celui parmi les chiens sans maladie respiratoire (44,7%). Donc, dans l'échantillon, il existait aussi une association statistique entre le sexe des chiens et la présence d'une maladie respiratoire. MAIS cette différence numérique entre ces deux pourcentages

n'était *pas* significative (car $p > 0,05$), donc cette association statistique n'était *pas* significative. Ainsi, dans cette étude, il n'existait pas d'association statistique significative entre le sexe des chiens et la présence d'une maladie respiratoire. On peut aussi écrire « dans cette étude, l'association statistique entre le sexe des chiens et la présence d'une maladie respiratoire n'était pas significative ».

Nous venons de voir quelque chose de très important ! Dire « il existait une association statistique dans l'échantillon » n'apporte aucune information ! En effet, il existera quasiment toujours une association statistique dans l'échantillon, car les indicateurs comparés seront, quasiment toujours, numériquement différents. Ainsi, ce qui est pertinent, ce n'est pas de savoir s'il existe ou non une association statistique dans l'échantillon, mais de savoir si l'association statistique est significative ou non significative, dans l'échantillon !

I. Significativité et importance clinique d'une différence observée dans un échantillon : deux choses très différentes

1. *En théorie*

Je ne vais pas vous démontrer ce que je vais vous raconter ci-dessous, mais j'espère que cela va vous paraître suffisamment intuitif pour que ce manque de démonstration ne soit pas (trop) frustrant.

La valeur du degré de signification *p* dépend certes de la valeur de la différence observée dans l'échantillon d_{obs}, mais aussi de la taille des groupes A et B de l'échantillon.

Tout d'abord, plus la différence observée d_{obs} augmente en valeur absolue, plus la probabilité d'observer une valeur au moins égale à $|d_{obs}|$ sous l'hypothèse qu'en vrai, H_0 est vraie, diminue, donc plus *p* diminue.

Ensuite, plus la taille des groupes A et B de l'échantillon augmente, plus *p* diminue.

Ce que je viens d'écrire a deux gigantesques conséquences. (1) Une différence d_{obs} observée dans l'échantillon peut être petite, ou faible, cliniquement parlant, mais malgré tout significative ($p \leq 0,05$) si cette différence observée d_{obs} a été calculée dans un échantillon de très grande taille. (2) Une différence d_{obs} peut être grande, ou importante, cliniquement parlant, mais malgré tout *non* significative ($p > 0,05$) si cette différence observée d_{obs} a été calculée dans un échantillon de petite taille.

2. *Illustrations*

J'illustre ce que je viens d'écrire à partir de deux études fictives ci-dessous, testant la même association entre la race des vaches (Prim'Holstein *versus* autre race) et la présence de mammites, d'abord dans un échantillon de 36 vaches, et ensuite dans un échantillon de 360 vaches (où tous les effectifs ont été multipliés par 10). La 1ère étude obtient les résultats suivants :

		Présence de mammites		
		Oui	Non	Total
Race	Prim'Holstein	8	10	18
	Autre race	3	15	18
	Total	11	25	36

A partir des données ci-dessus, les deux pourcentages comparés pour montrer qu'il existe une association significative entre la race et la présence de mammites sont les suivants : le pourcentage de vaches avec mammites parmi les Prim'Hostein (8/18=44%) et le pourcentage de vaches avec mammites parmi les vaches d'une autre race (3/18=17%). Ces deux pourcentages sont cliniquement très différents (l'un est plus du double de l'autre). Si vous réalisez le test statistique pour tester la différence entre ces deux pourcentage (le test du Chi^2 que l'on verra plus loin dans ce polycopié), vous obtenez la valeur du degré de signification égale à 0,07. Donc, les deux pourcentages 44% et 17%, cliniquement parlant très différents, ne sont cependant pas *significativement* différents.

La 2ème étude fictive obtient les résultats suivants (tous les effectifs ont été multipliés par 10 par rapport à la 1ère étude) :

		Présence de mammites		
		Oui	Non	Total
Race	Prim'Holstein	80	100	180
	Autre race	30	150	180
	Total	110	250	360

Dans cette seconde étude, bien entendu, les pourcentages comparés sont strictement identiques à ceux de la première étude, puisque tous les effectifs ont été multipliés par 10 : 80/180=44% et 30/180=17%. La valeur du degré de signification du tests statistique testant la différence de ces pourcentages est cette fois-ci égale à 10^{-8} (donc < 0,05). Par conséquent, dans cette 2ème étude, la différence de pourcentages, identique à celle dans la 1ère étude (44% *versus* 17%), est cette fois-ci significative.

Dans la première étude, la faible taille d'échantillon n'a pas permis de rendre significative la différence de pourcentages pourtant cliniquement importante.

Par conséquent, « significativité » et « importance clinique » ne sont pas du tout synonymes ! De nombreux chercheurs pensent malheureusement le contraire, et notamment pensent à tort qu'une différence *non* significative est synonyme de « faible différence ».

VI. LA NOTION D'INDEPENDANCE DES INDIVIDUS

A. Introduction et précision sur le terme « individu »

Pour être valides, la majorité des tests statistiques nécessitent que les individus soient *indépendants* les uns vis-à-vis des autres. D'autres tests statistiques prennent justement en compte cette non indépendance. Il est donc important de déceler les situations où les individus sont indépendants, et celles où ils ne le sont pas.

Avant de définir ce terme d' « indépendance », je dois définir désormais précisément celui d' « individu ». Un « individu », dans une étude, est l'unité statistique sur laquelle sont calculés les indicateurs statistiques (moyenne, médiane, pourcentage, …). L'individu peut être, par exemple, le prélèvement sanguin, l'animal, l'élevage, ou le propriétaire d'un animal. En règle générale, dans un fichier de données, les « individus » sont placés en ligne, et il y a donc autant de lignes que d' « individus ». Et en colonne figurent les caractéristiques qui sont mesurées chez les « individus ».

B. Définition d' « indépendance » des données

On considère qu'il y a « indépendance » entre les individus sur le caractère (par exemple, la production laitière mensuelle d'une vache) dont on calcule un indicateur (par exemple, la moyenne), si la valeur du caractère d'un individu de l'échantillon est indépendante de la valeur de ce caractère d'un *autre* individu de l'échantillon. Dans l'exemple de la production laitière moyenne, si les individus constituant l'échantillon sont plusieurs vaches laitières d'un même élevage, et que l'on utilise plusieurs élevages pour constituer l'échantillon, les vaches de l'échantillon provenant d'un même élevage ne sont probablement pas indépendantes sur le caractère « production laitière mensuelle », car dans l'échantillon, la valeur de la production laitière d'une vache d'un élevage va probablement davantage ressembler à la production laitière d'une autre vache du *même* élevage qu'à celle d'une autre vache d'un élevage *différent* (car au sein d'un même élevage, il y a la même alimentation, les mêmes conditions d'élevage, etc.). Vous pouvez noter que pour un même échantillon, deux individus peuvent être considérés comme indépendants sur un caractère, et comme non indépendants sur un autre.

Ainsi, nous dirons que les *données* d'un échantillon concernant un caractère sont indépendantes si les individus sont indépendants sur ce caractère ; elles seront dites « non indépendantes » dans le cas contraire.

C. Situations classiques de données non indépendantes

Lorsque le phénomène de non indépendance est pressenti (plusieurs animaux d'un même élevage, plusieurs animaux d'une même portée, plusieurs prélèvements sanguins d'un même animal, …), il faut réfléchir au cas par cas s'il y a effectivement non indépendance sur *le* caractère dont on calcule l'indicateur. Classiquement, lorsqu'il y a au moins deux prélèvements sanguins d'un même animal, et si les analyses statistiques portent sur les marqueurs biologiques quantifiés à partir de ces prélèvements, il n'y a clairement pas indépendance des individus (ici, l'individu est le prélèvement sanguin). Une autre situation classique est celle qui va être décrite plus précisément dans la partie IX : un animal est vu deux

fois, une fois avant intervention (traitement, opération, …) et une fois après intervention, et l'on veut savoir s'il existe une *évolution* entre ces deux moments. On mesure un caractère deux fois sur un même animal, donc ces mesures (⇔ individus) ne sont pas indépendantes.

D. Que faire en cas de non indépendance ? (Hors programme)

Il y a principalement deux choses à faire (l'une ou l'autre, en fonction des compétences en statistique de la personne qui analyse les données) en situation de non indépendance.

Première chose, si vous utilisez les outils statistiques qui ne sont valides que pour des données indépendantes, vous devez dire que vous êtes dans une situation de non indépendance, que les méthodes statistiques que vous avez utilisées ne sont pas adaptées, que les résultats issus des tests statistiques sont à prendre avec précaution, et donc qu'il faudra confirmer vos résultats dans une autre étude (utilisant d'autres individus) en prenant en compte cette non indépendance.

Deuxième chose, vous utilisez les méthodes statistiques prenant en compte la non indépendance des données[34].

[34] Evident, non ? Oui, mais dans certaines situations, les méthodes statistiques sont trop compliquées ou demandent trop de compétences en statistiqus, et alors, on se rabat sur la première chose que je vous conseille de faire.

VII. LES TESTS STATISTIQUES SUR DONNEES INDEPENDANTES EN PRATIQUE

A. Introduction

1. *Préambule*

Mon objectif n'est pas de vous apprendre à réaliser un test statistique à la main. Je préfère passer du temps à vous apprendre à *interpréter* un résultat d'un test statistique que vous verrez dans les articles scientifiques (et cela inclut bien entendu de savoir ce que ces résultats ne veulent *pas* dire). Cette compétence que vous devez acquérir est beaucoup plus proche des compétences que l'on attend d'un vétérinaire à la sortie de l'école que celle de savoir réaliser un test statistique « à la main » (car pour avoir un esprit critique et pratiquer l' « evidence-based veterinary medicine », il *faut* savoir interpréter les résultats d'un test statistique)[35]. En revanche, vous devez connaître leurs conditions de validité, interpréter leurs résultats, et les réaliser à l'aide du site Internet BiostaTGV[36] que nous utiliserons en TD.

Je présenterai, pour information, la démarche de calcul pour réaliser certains de ces tests statistiques. Cette démarche n'est pas à apprendre pour l'examen.

Enfin, dans tout ce qui suit dans cette partie VII, tout ce qui est écrit entre crochets doit être remplacé par ce qu'il faut en fonction du contexte de l'étude et du ou des indicateurs estimés

2. *Vue d'ensemble des tests statistiques usuels*

La **Figure 11** ci-dessous présente les tests statistiques usuels lorsque les individus sont indépendants. Le tableau de cette figure doit être appris par cœur. Nous allons passer quelques heures en TD pour que vous « pratiquiez » ce tableau !

Dans la suite de cette partie VII, nous allons décrire un à un ces tests statistiques, et notamment leurs conditions de validité, l'hypothèse nulle H_0 qu'ils testent, et la conclusion que l'on peut écrire à partir du résultat du test statistique.

[35] Cet avis n'est que personnel, et peut tout à fait ne pas être partagé par d'autres.

[36] https://biostatgv.sentiweb.fr/?module=tests

Etudier l'association entre…	Indicateurs à comparer	Test statistique à utiliser
2 variables binaires	Compare 2 %	Chi-2, Fisher*
1 variable binaire x 1 variable qualitative	Compare ≥ 3 %	
2 variables qualitatives	Résultats ininterprétables ⇒ il faut transformer une des deux variables en une variable binaire	
1 variable binaire x 1 variable quantitative	Compare 2 moyennes	Student pour SNA
	Compare 2 médianes*	Mann-Whitney**
1 variable qualitative x 1 variable quantitative	Compare ≥ 3 moyennes	ANOVA
	Compare ≥ 3 médianes*	Kruskal-Wallis**
2 variables quantitatives	Calcule un coefficient de corrélation (pas d'indicateurs comparés)	Coefficient de corrélation de Pearson
		Coefficient de corrélation de Spearman***

SNA = séries non appariées
* Test de Fisher à utiliser à la place du test du Chi-2 si au moins un des effectifs attendus est < 5
** Test « non paramétrique », à utiliser si la distribution de la variable quantitative ne suit pas une loi normale
*** Coefficient de corrélation « non paramétrique », à utiliser si la distribution d'au moins l'une des deux variables quantitatives ne suit pas une loi normale

Figure 11. Récapitulatif des indicateurs à utiliser pour étudier l'association entre deux variables, et des tests statistiques à utiliser pour tester ces associations. Les individus doivent être indépendants pour utiliser ces tests statistiques.

3. *Règles générales de communication scientifique dans les conclusions à l'issue du résultat d'un test statistique*

a) Conclusion au niveau de l'échantillon

Quelle que soit la valeur du degré de signification d'un test statistique, il faut d'abord conclure au niveau de l'échantillon. Dans cette conclusion, le temps utilisé doit être au passé, car les résultats cités sont ceux provenant de données qui ont été collectées avant la date de rédaction de la conclusion.

Dans la conclusion au niveau de l'échantillon, il est *indispensable* de fournir les valeurs des indicateurs estimés dans chacun des groupes, quelle que soit la valeur du degré de signification, pour deux raisons cruciales : (1) une différence d'indicateurs peut être significative ($p \leq 0,05$) mais cliniquement faible, et (2) une différence d'indicateurs peut être non significative ($p > 0,05$) mais cliniquement importante (cf. partie V.I, page 39).

La valeur du degré de signification (qui doit figurer dans la conclusion au niveau de l'échantillon) doit toujours être arrondie à 2 chiffres après la virgule (écrire « $p = 1,00$ » si le degré de signification est égal à 1 ; écrire « $p < 0,01$ » si le degré de signification est inférieur à 0,01).

b) Conclusion éventuelle au niveau de la population cible (inférence)

Ensuite, si le degré de signification p du test statistique est inférieur ou égal à 0,05, il faut conclure au niveau de la population cible, c'est-à-dire faire de l'inférence. Dans cette conclusion (inférence), le temps utilisé doit être le présent car en science, le présent est réservé à la généralisation d'un résultat à l'ensemble d'une population.

De plus, comme nous l'avons dans la partie V.F.3 (page 34), dans de nombreuses situations, on ne peut pas être fortement convaincu de l'existence d'une association réelle dans la population cible, à partir d'une association significative dans l'échantillon ($p \leq 0,05$). Ainsi, au moment de faire de l'inférence, des « gants » doivent être mis, en écrivant qu'il y a « des chances » (pas « de grandes chances », et encore moins « 95% de chances ») pour que l'association existe réellement dans la population cible.

Si le degré de signification est supérieur à 0,05, il est interdit de faire une quelconque inférence. En effet, la probabilité de se tromper en acceptant H_0, c'est-à-dire en écrivant que H_0 est vraie, *dépend*[37] de β, de valeur inconnue donc potentiellement très grande (cf. partie V.G.3, page 36).

B. Le test de Student pour séries non appariées (comparaison de deux moyennes)

1. *Contexte du test de Student pour séries non appariées*

Le test de Student pour séries non appariées s'utilise lorsque l'on souhaite montrer que deux groupes A et B de la population cible diffèrent sur la *moyenne* d'un caractère quantitatif. Le test de Student utilise les lois de Student[38]. L'étude de l'association entre l'appartenance au groupe (A ou B) et la valeur de ce caractère consistera à comparer la valeur de la moyenne estimée dans le groupe A de l'échantillon (« $\widehat{\mu_A}$ » ou plus simplement, « m_A ») à celle estimée dans le groupe B de l'échantillon (« $\widehat{\mu_B}$ » ou plus simplement, « m_B »).

2. *Que veut dire « séries non appariées » ?*

Nous le reverrons en détails quand nous parlerons du test de Student pour séries appariées, plus loin dans ce polycopié. Les « séries » sont « non appariées » si les individus du groupe A sont indépendants des individus du groupe B. Les « séries » seront « appariées » quand, par exemple, les animaux sont mesurés deux fois, un fois à t_1, l'autre à t_2, et que l'on veut comparer la moyenne des valeurs à t_1 (les valeurs à t_1 constitueront la « série de données à t_1 ») à la moyenne des valeurs à t_2 (les valeurs à t_2 constitueront la « série de données à t_2 »).

3. *Hypothèse nulle dans le test de Student pour séries non appariées*

L'hypothèse nulle H_0 d'un test de Student pour séries non appariées est l'une des deux suivantes (au choix, car les deux sont équivalentes) :

- « Dans la population [cible], la moyenne [du caractère quantitatif] parmi [les individus du groupe A] est égale à la moyenne [du caractère quantitatif] parmi [les individus du groupe B] ».
- « Dans la population [cible], il n'existe pas d'association réelle entre [le caractère quantitatif] et [l'appartenance au groupe A ou B] ».

[37] Cette probabilité de se tromper en acceptant H_0 *n'est pas* égale à β.

[38] Cf. ici : https://fr.wikipedia.org/wiki/Loi_de_Student

4. *Conditions de validité du test de Student pour séries non appariées*

Trois conditions doivent être vérifiées pour que le test de Student pour séries non appariées soit valide.

1) Les individus de l'échantillon doivent être considérés comme indépendants.

2) Les variances dans les deux échantillons (SD_A^2 et SD_B^2) ne doivent pas être trop différentes. En pratique, dans le module de Biostatistique en Médecine Vétérinaire, on les considèrera comme « pas trop différentes » si $\frac{1}{3} < \frac{SD_A^2}{SD_B^2} < 3$.

3) La distribution de la variable quantitative dont on calcule les deux moyennes doit être considérée comme normale dans la population dont est issu l'échantillon. En pratique, on vérifiera la normalité de cette distribution dans l'échantillon à partir d'un histogramme dressé grâce à un site Internet (cf. partie III.B).

5. *Conclusion à l'issue du test de Student pour séries non appariées*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la moyenne [du caractère quantitatif] parmi [les individus du groupe A] ([valeur de m_A]) était significativement [supérieure, ou inférieure] à la moyenne [du caractère quantitatif] parmi [les individus du groupe B] ([valeur de m_B] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour que la moyenne [du caractère quantitatif] parmi [les individus du groupe A] soit réellement [supérieure, ou inférieure] à la moyenne [du caractère quantitatif] parmi [les individus du groupe B]. »

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la moyenne [du caractère quantitatif] parmi [les individus du groupe A] ([valeur de m_A]) n'était pas significativement différente de la moyenne [du caractère quantitatif] parmi [les individus du groupe B] ([valeur de m_B] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

C. Le test du Chi² avec comparaison de deux pourcentages

1. *Contexte du test du Chi² avec comparaison de deux pourcentages*

Le test du Chi² avec comparaison de deux pourcentages s'utilise lorsque l'on souhaite montrer que deux groupes de la population cible diffèrent sur le pourcentage de présence d'un caractère binaire. L'étude de l'association entre l'appartenance au groupe (A ou B) et la présence de ce caractère consistera à comparer la valeur du pourcentage de présence de ce caractère parmi les individus du groupe A de l'échantillon (« $\widehat{\pi_A}$ » ou plus simplement, « p_A ») au pourcentage de présence de ce caractère parmi les individus du groupe B de l'échantillon (« $\widehat{\pi_B}$ » ou plus simplement, « p_B »).

2. *Notations*

Je vais supposer que le caractère binaire est la présence d'une maladie M. Voici comment se répartissent les effectifs *observés* dans une étude, dans chacune des quatre cases du tableau (cf. **Tableau 1**).

Tableau 1. Tableau des effectifs observés dans la situation d'une comparaison de pourcentages avec deux variables binaires.

		Présence de la maladie		
		Malade (M)	Non malade (NM)	Total
Groupe	A	O_{AM} (p_{AM} %)	O_{ANM}	n_A
	B	O_{BM} (p_{BM} %)	O_{BNM}	n_B
	Total	n_M (p_M %)	n_{NM}	n_T

Par exemple, O_{AM} est le *nombre* d'individus malades observé au sein du groupe A, O_{BNM} le nombre d'individus non malades observé au sein du groupe B. Il y a en tout n_A individus dans le groupe A et en tout n_M individus malades dans l'échantillon. Les pourcentages observés d'individus malades *parmi* les individus des groupes A et B sont respectivement p_{AM} % et p_{BM} %, et le pourcentage d'individus malades parmi l'ensemble des individus de l'échantillon est p_M %.

3. *Citations correctes et incorrectes de pourcentages à comparer*

a) Préambule

Cette partie « Citations correctes et incorrectes de pourcentages à comparer » est absolument fondamentale à avoir comprise. Vous *serez* interrogés sur cette question aux examens (aucune incertitude, pour le coup, là).

b) Problématique

Avant de vous ruer sur le degré de signification du test du Chi^2 lorsque vous lisez un article, vous *devez* savoir quels sont les pourcentages qui sont comparés. En effet, écrire « l'effet d'un traitement (*versus* placebo) est significatif sur le pourcentage de guérison (p=0,02) », semble beaucoup apporter à la science (vétérinaire), sauf que ce n'est pas le cas. En effet, si l'on nous dit que le pourcentage de guérison chez les animaux traités est de 35% et chez les animaux non traités de 31% (guérison spontanée), tous deux significativement différents l'un de l'autre, alors là vous vous demandez, à juste titre, si la différence de pourcentages de guérison vaut le coup de traiter l'animal, sachant que le traitement a probablement des effets indésirables, et qu'il a de toute façon un coût pour le propriétaire !...

c) Comment bien citer deux pourcentages à comparer ?

Pour savoir si deux variables binaires sont associées, il faut comparer (puis tester) deux pourcentages. Les pourcentages que vous citez doivent être tels que s'ils sont *égaux*, ils traduisent une *absence* d'association statistique (cf. partie IV.B), et s'ils sont *différents*, ils traduisent la *présence* d'une association statistique (qu'elle soit ou non significative).

Par exemple, vous êtes d'accord avec le fait que la couleur bleue des yeux (*versus* une autre couleur) n'est pas du tout associée à la présence de cancer de l'œsophage. Par conséquent, les deux pourcentages que l'on va citer doivent être égaux. Il y a de nombreuses façons de *mal* citer ces pourcentages.

Si vous dites « je vais comparer le pourcentage de personnes avec les yeux bleus parmi les personnes qui ont le cancer de l'œsophage au pourcentage de personnes qui n'ont pas les yeux bleus parmi les personnes qui ont le cancer de l'œsophage », vous citez les *mauvais* pourcentages. En effet, le pourcentage de personnes avec les yeux bleus parmi les personnes qui ont le cancer de l'œsophage est probablement proche de 30% (il y a en effet, en France, environ 30% de personnes qui ont les yeux bleus). Par conséquent, le pourcentage de personnes qui n'ont *pas* les yeux bleus parmi les personnes qui ont le cancer de l'œsophage est de fait égal à environ 70% (la somme fait 100%). Or, 70% est très différent de 30%. Et comme ces pourcentages sont différents, vous diriez « il existe donc une association entre la couleur des yeux et la présence d'un cancer de l'œsophage », ce qui est faux puisque justement, il n'existe pas d'association entre la couleur des yeux et la présence d'un cancer de l'œsophage. Par conséquent, les deux pourcentages 70% et 30% (et leur expression en français) ne doivent pas être cités pour savoir s'il existe une association entre la couleur des yeux et la présence de cancer de l'œsophage.

Pour bien citer deux pourcentages, la règle que vous devez absolument appliquer est la suivante (si vous ne le faites pas, c'est à vos risques et périls, notamment en examen) : « le pourcentage de XX *parmi* les *uns* à comparer au pourcentage de XX *parmi* les *autres* », en mettant ce que vous voulez dans le XX, mais qui doit rester la même chose les deux fois !

Pour reprendre l'exemple des yeux bleus et du cancer de l'œsophage, cela donne quatre couples de pourcentages corrects que l'on peut citer :

- le pourcentage de personnes avec les yeux bleus *parmi* celles avec un cancer de l'œsophage à comparer au pourcentage de personnes avec les yeux bleus *parmi* celles qui n'ont pas de cancer de l'œsophage (XX ici est « avec les yeux bleus ») ;
- le pourcentage de personnes qui n'ont pas les yeux bleus *parmi* celles avec un cancer de l'œsophage à comparer au pourcentage de personnes qui n'ont pas les yeux bleus *parmi* celles qui n'ont pas de cancer de l'œsophage (XX ici est « qui n'ont pas les yeux bleus ») ;
- le pourcentage de personnes avec cancer de l'œsophage *parmi* celles avec les yeux bleus à comparer au pourcentage de personnes avec cancer de l'œsophage *parmi* celles qui n'ont pas les yeux bleus (XX ici est « avec cancer de l'œsophage ») ;
- le pourcentage de personnes sans cancer de l'œsophage *parmi* celles avec les yeux bleus à comparer au pourcentage de personnes sans cancer de l'œsophage *parmi* celles qui n'ont pas les yeux bleus (XX ici est « sans cancer de l'œsophage »).

4. *Hypothèse nulle dans le test du Chi² avec comparaison de deux pourcentages*

Comme dans le test de Student pour séries non appariées, il existe deux façons, au choix, de citer l'hypothèse nulle H_0 d'un test du Chi^2 avec comparaison de deux pourcentages :

- « Dans la population [cible], le pourcentage [du caractère binaire] parmi [les individus du groupe A] est égal au pourcentage [du caractère binaire] parmi [les individus du groupe B] ».
- « Dans la population [cible], il n'existe pas d'association réelle entre [la présence du caractère binaire] et [l'appartenance au groupe A ou B] ».

5. *Démarche de calcul du test du Chi² avec comparaison de deux pourcentages*

Le test du Chi² ne compare pas, *numériquement*, les pourcentages de malades dans les deux échantillons (p_{AM} et p_{BM}), contrairement au test de Student pour séries non appariées qui compare numériquement les moyennes m_A et m_B[39]. Le test du Chi² compare des *effectifs*. Il compare notamment des effectifs observés (cf. **Tableau 1**) à des effectifs attendus sous H_0.

Les effectifs attendus sous H_0 sont les effectifs que l'on aurait dû observer dans l'échantillon si H_0 avait été vraie, et s'il n'y avait eu aucune fluctuation d'échantillonnage dans la création de l'échantillon. Pour illustrer le calcul des effectifs attendus sous H_0 (que vous devez savoir réaliser pour l'examen), je vais utiliser un exemple.

Supposons une étude dont l'objectif était de montrer que parmi les vaches laitières en France, il existe une association entre la présence de mammites et la race Prim'Hostein (*versus* autres races). Pour cela, 120 vaches ont été recrutées dans l'étude, dont la répartition est présentée dans le **Tableau 2**. Les quatre effectifs observés (dans chacune des quatre cases du tableau) sont : 20, 20, 10, et 70 vaches. Les deux pourcentages observés qui vont être testés par le test du Chi² sont par exemple[40] : le pourcentage de présence de mammites parmi les vaches Prim'Hostein de l'échantillon (20/40=50%) et le pourcentage de présence de mammites parmi les vaches d'une autre race de l'échantillon (10/80=12,5%). Dans l'échantillon, nous pouvons remarquer que le pourcentage de vaches présentant une mammite parmi l'ensemble des 120 vaches de l'échantillon est de 30/120=25%.

Tableau 2. Répartition des 120 vaches incluses dans une étude souhaitant montrer l'association entre la présence de mammites et la race Prim'Holstein (*versus* autres races).

		Présence de mammites		
		Oui	Non	Total
Race	Prim'Holstein	20 (50%)	20	40
	Autre race	10 (12,5%)	70	80
	Total	30 (25%)	90	120

Je vais maintenant vous donner la règle (que vous devez connaître) pour calculer les quatre effectifs *attendus* sous H_0, puis je vous montrerai qu'effectivement ce sont bien des effectifs *attendus* sous H_0.

Pour obtenir l'effectif attendu sous H_0 de la case à la ligne i et à la colonne j du tableau (*i* et *j* variant de 1 à 2), vous devez faire le produit des effectifs totaux de la ligne i par les effectifs totaux de la colonne j, que vous divisez ensuite par l'effectif total dans l'échantillon. Attention, dans ce calcul, vous ne devez pas arrondir les chiffres à l'unité, vous devez les arrondir à un chiffre après la virgule (ce sont des effectifs *théoriques*, donc on accepte ce chiffre après la

[39] Une différence numérique entre deux pourcentages qui vaut par exemple 10% n'a pas du tout la même signification en fonction des valeurs de ces pourcentages. En effet, si p_{AM} = 5% et p_{BM} = 15%, il y a un écart de 10%, mais p_{BM} est trois fois plus élevé que p_{AM}. Tandis que si si p_{AM} = 40% et p_{BM} = 50%, il y a un même écart de 10%, mais p_{BM} n'est plus très différent de p_{AM} !

[40] Vous vous souvenez, il y a de nombreuses façons de citer correctement des pourcentages. Ici, je vous en présente une. Vous pouvez réfléchir aux les trois autres couples de pourcentages pour vous entrainer.

virgule). En suivant cette règle, vous obtenez les effectifs attendus sous H_0 présentés dans le **Tableau 3**.

Tableau 3. Effectifs attendus sous H_0 à partir des données du **Tableau 2**.

		Présence de mammites		
		Oui	Non	Total
Race	Prim'Holstein	10,0	30,0	40
	Autre race	20,0	60,0	80
	Total	30	90	120

10,0 = 40 x 30 / 120

20,0 = 80 x 30 / 120

30,0 = 40 x 90 / 120

60,0 = 80 x 90 /120

La raison de ce produit en croix (cf. ci-dessous) n'est pas à savoir pour l'examen. Je vous la donne quand même pour information. Ces effectifs sont effectivement ceux attendus sous H_0. En effet, si H_0 avait été vraie, et s'il n'y avait eu *aucune* fluctuation d'échantillonnage en créant cet échantillon de 120 vaches, nous aurions dû observer *exactement* le même pourcentage de vaches avec mammites entre celui parmi les vaches Prim'Holstein et celui parmi les vaches d'autre race (absence d'association statistique ⇔ égalité des indicateurs). Et si, dans l'échantillon, le pourcentage de vaches avec mammites parmi les vaches Prim'Holstein avait été égal au pourcentage de vaches avec mammites parmi les vaches d'autre race, ils auraient été forcément égaux au pourcentage de vaches avec mammites parmi les 120 vaches de l'échantillon : 30/120 (25%). Ensuite, en appliquant ce 30/120 (25%) aux 40 vaches Prim'Holstein, cela donne 40 x 30 / 120 (on retrouve la formule ci-dessus, 1ère ligne), et en appliquant ce 30/120 (25%) aux 80 vaches d'autre race, cela donne 80 x 30 / 120 (on retrouve la formule ci-dessus, 2ème ligne). (Et le raisonnement est bien entendu similaire pour obtenir les deux autres produits en croix : « 40 x 90 / 120 » et « 80 x 90 / 120 ».)

Une fois que ces quatre effectifs attendus sous H_0 sont calculés (et arrondis à un chiffre après la virgule), la démarche (hors programme) consiste à calculer la différence entre les effectifs observés et les effectifs attendus sous H_0. Si l'on note E_{AM}, E_{BM}, E_{ANM}, et E_{BNM} respectivement les effectifs attendus malades du groupe A, malades du groupe B, non malades du groupe A et non malades du groupe B (les nombres de 10, 20, 30, et 60 du **Tableau 3**), la formule (que vous n'avez pas à connaître) est la suivante :

$$d_{obs} = \frac{(O_{AM} - E_{AM})^2}{E_{AM}} + \frac{(O_{BM} - E_{BM})^2}{E_{BM}} + \frac{(O_{ANM} - E_{ANM})^2}{E_{ANM}} + \frac{(O_{BNM} - E_{BNM})^2}{E_{BNM}}$$

Plus cette différence entre effectifs observés et effectifs attendus sous H_0 est importante, plus ce que l'on a observé est éloigné de H_0, et plus on va donc avoir tendance à rejeter H_0.

Si en vrai, H_0 est vraie, alors l'ensemble des différences observables entre effectifs observés et effectifs attendus sous H_0 suit une loi du Chi^2 à 1 degré de liberté[41] (d'où la raison pour

[41] Cf. ici : https://fr.wikipedia.org/wiki/Loi_du_%CF%87%C2%B2

laquelle on parle de test statistique « du Chi^2 »). Le degré de signification se calcule à partir de la valeur de cette différence observée d_{obs}.

6. *Conditions de validité du test du Chi^2*

Deux conditions doivent être vérifiées avant de regarder le résultat d'un test du Chi^2.

1) Les individus de l'échantillon doivent être considérés comme indépendants.

2) Les quatre effectifs *attendus* sous H_0 (arrondis à un chiffre après la virgule) doivent être tous les quatre supérieurs ou égaux à 5.

Si au moins un des effectifs attendus sous H_0 est inférieur à 5, le test exact de Fisher doit alors être utilisé. Ainsi, puisque savoir si un test statistique est valide fait partie du programme du module, savoir calculer les effectifs attendus sous H_0 en fait partie aussi !

Vous vous rendez compte aussi de quelque chose d'un peu perturbant, c'est que pour savoir si l'on peut utiliser le test du Chi^2, il faut commencer la démarche de ce test en calculant les effectifs attendus sous H_0. Si au moins l'un d'entre eux est inférieur à 5, on s'arrête là et on passe au test exact de Fisher.

7. *Conclusion à l'issue du test du Chi^2*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, le pourcentage [du caractère binaire] parmi [les individus du groupe A] ([valeur de p_A]) était significativement [supérieur, ou inférieur] au pourcentage [du caractère binaire] parmi [les individus du groupe B] ([valeur de p_B] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour que le pourcentage [du caractère binaire] parmi [les individus du groupe A] soit réellement [supérieur, ou inférieur] au pourcentage [du caractère binaire] parmi [les individus du groupe B]. »

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, le pourcentage [du caractère binaire] parmi [les individus du groupe A] ([valeur de p_A]) n'était pas significativement différent du pourcentage [du caractère binaire] parmi [les individus du groupe B] ([valeur de p_B ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

D. Le test du Chi^2 avec comparaison de trois pourcentages ou plus

1. *Contexte du test du Chi2 avec comparaison de trois pourcentages ou plus*

Le test du Chi^2 avec comparaison de trois pourcentages ou plus s'utilise lorsque l'on souhaite montrer qu'il existe une association entre une variable binaire (le groupe ou le caractère étudié) et une variable qualitative à trois catégories ou plus (le caractère étudié ou le groupe). Pour faciliter le discours, je vais considérer que la variable binaire correspond au « groupe » (A ou B) et que la variable qualitative correspond au « caractère étudié ». Dans certains cas, ce sera un peu tiré par les cheveux (par exemple si la variable binaire est la présence de

mammites et si la variable qualitative est la race en plusieurs catégories), mais ce n'est pas grave, dans les phrases qui suivent, de toute façon vous n'utiliserez pas les mots « groupe » ni « caractère étudié », mais plutôt ce à quoi ces variables font référence.

Je vous rappelle (cf. partie VII.A.2, page 43) que vous ne pourrez pas mettre en évidence d'association entre deux variables qualitatives à trois catégories ou plus[42]. Pour cela, vous devrez rendre binaire au moins l'une de vos deux variables.

2. *Pourcentages comparés*

Le plus dur dans la situation d'une variable binaire croisée avec une variable qualitative est de savoir quels sont les pourcentages que vous allez comparer *avant* de les tester. Supposons que l'on veuille savoir si la parité de la vache est associée au type d'affection à l'origine du syndrome de la vache couchée (**Tableau 4**). Dans cet exemple, on peut considérer que nous avons deux groupes (les primipares et les multipares) et le caractère étudié (le type l'affection) est qualitatif en quatre catégories.

Tableau 4. Répartition du nombre de vaches selon leur race et la présence d'une affection à l'origine d'un syndrome de la vache couchée.

	Affection à l'origine du syndrome de la vache couchée				
	Métrite	Métabolique	Myoarthrosquelettique	Autre	Total
Primipares	5	8	11	10	34
Multipares	20	12	15	29	76
Total	25	20	26	39	110

Il y a de nombreuses façons de mal citer les pourcentages qui vont être comparés puis testés (cf. discussion partie VII.C.3). Je vous propose une bonne façon de citer les pourcentages qui vont ensuite être testés dans le cas du **Tableau 4** : le pourcentage de vaches multipares parmi les vaches ayant une métrite (20/25=80%), le pourcentage de vaches multipares parmi les vaches ayant une affection métabolique (12/20=60%), le pourcentage de vaches multipares parmi les vaches ayant une affection myoarthrosquelettique (15/26=58%), et le pourcentage de vaches multipares parmi les vaches ayant une autre affection (29/39=74%). S'il y avait une absence totale d'association entre la parité de la vache et le type d'affection à l'origine d'un syndrome de la vache couchée, ces quatre pourcentages auraient dû être égaux (en effet, on aurait dû retrouver le même pourcentage de vaches multipares parmi les vaches de chacune des quatre catégories). Pour savoir s'il y existe une association *significative* entre la parité de la vache et le type d'affection à l'origine d'un syndrome de la vache couchée, il faudra savoir si ces quatre pourcentages sont *significativement* différents les uns des autres.

Parenthèse, vous auriez tout à fait pu citer les quatre pourcentages suivants : le pourcentage de vaches primipares parmi les vaches ayant une métrite (5/20=20%), le pourcentage de vaches primipares parmi les vaches ayant une affection métabolique (8/20=40%), le pourcentage de vaches primipares parmi les vaches ayant une affection myoarthrosquelettique (11/26=42%), et le pourcentage de vaches primipares parmi les vaches ayant une autre affection (10/39=26%).

[42] Mathématiquement/statistiquement, c'est possible, mais les résultats sont *ininterprétables*. Donc, je vous empêche tout simplement de le faire !

La règle générale pour bien citer les pourcentages à comparer dans le cas d'un croisement entre une variable binaire et une variable qualitative est la suivante : vous fournissez le pourcentage d'individus appartenant au groupe A[43] parmi les individus de la catégorie n°1, le pourcentage d'individus appartenant au groupe A parmi les individus de la catégorie n°2, etc. jusqu'à la dernière catégorie.

3. *Hypothèse nulle dans le test du Chi² avec comparaison de trois pourcentages ou plus*

Il existe là encore deux façons, au choix, de citer l'hypothèse nulle H_0 d'un test du Chi² avec comparaison de trois pourcentages ou plus.

- « Dans la population [cible], le pourcentage [d'individus appartenant au groupe A] parmi [les individus de la catégorie n°1 du caractère étudié], celui parmi [les individus de la catégorie n°2 du caractère étudié], etc. sont tous égaux ».
- « Dans la population [cible], il n'existe pas d'association réelle entre [la présence du caractère qualitatif] et [l'appartenance au groupe A ou B] ».

4. *Démarche de calcul du test du Chi² avec comparaison de trois pourcentages ou plus*

La démarche est identique à celle du test du Chi² avec comparaison de deux pourcentages : calculs des effectifs attendus sous H_0, calcul de la différence d_{obs} entre les effectifs attendus sous H_0 et les effectifs observés, et calcul du degré de signification à partir de la valeur de cette différence observée d'effectifs. Le calcul des effectifs attendus sous H_0 lors de la comparaison de trois pourcentages ou plus n'est pas au programme du module de Biostatistique en Médecine Vétérinaire.

5. *Conditions de validité du test du Chi²*

Deux conditions doivent être vérifiées avant de regarder le résultat d'un test du Chi².

1) Les individus de l'échantillon doivent être considérés comme indépendants.

2) Tous les effectifs *attendus* sous H_0 (arrondis à un chiffre après la virgule) doivent être tous les quatre supérieurs ou égaux à 5.

Si au moins un des effectifs attendus sous H_0 est inférieur à 5, le test exact de Fisher doit alors être utilisé.

6. *Conclusion à l'issue du test du Chi²*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, le pourcentage [d'individus appartenant au groupe A] parmi [les individus de la catégorie n°1 du caractère étudié] ([p_1]), celui parmi [les individus de la catégorie n°2 du caractère étudié] ([p_2]), etc. étaient significativement différents (p = [valeur du degré de signification]). »

[43] Vous choisissez bien entendu ce que le groupe « A » représente, par rapport au groupe « B » !

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour qu'il existe une association réelle entre [le caractère qualitatif] et [le fait d'appartenir au groupe A ou B]. »

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, le pourcentage [d'individus appartenant au groupe A] parmi [les individus de la catégorie n°1 du caractère étudié] ([p_1]), celui parmi [les individus de la catégorie n°2 du caractère étudié] ([p_2]), etc. n'étaient pas significativement différents (p = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

7. Commentaires

Attention, lorsque l'on teste l'association entre une variable binaire et une variable qualitative (variable à trois classes ou plus), il est interdit de dire que le pourcentage observé le plus élevé parmi tous ceux cités est « significativement supérieur » à tous les autres qui sont moins élevés. L'hypothèse nulle étant l'égalité des pourcentages, le rejet de l'hypothèse nulle H_0 ne permet de dire qu'une seule chose, c'est qu'ils sont *globalement* significativement différents (mais pas *un* en particulier).

E. Le test exact de Fisher

Ce test statistique doit être utilisé quand le test du Chi^2 n'est pas applicable parce qu'au moins un des effectifs attendus sous H_0 est inférieur à 5. L'utilisation et l'interprétation restent identiques à celles d'un test du Chi^2 : comparaison de deux pourcentages ou comparaison de trois pourcentages ou plus, H_0, pourcentages à comparer, conclusion à l'issue du test. Seul le calcul du degré de signification est différent, car il utilise la loi binomiale, plutôt que la loi du Chi^2. Ainsi, pour tester l'association entre deux variables binaires, ou entre une variable binaire et une variable qualitative, je vous suggère les deux étapes suivantes :

1) Calculer ou regarder (quand ils ont déjà été calculés) les effectifs attendus sous H_0 ;

2) Si tous les effectifs attendus sous H_0 sont $\geq 5 \Rightarrow$ effectuez le test du Chi^2. Si au moins un des effectifs attendus est < 5, effectuez le test de Fisher.

F. L'analyse de variance (ANOVA, pour ANalysis Of VAriance)

1. Contexte du test de l'ANOVA

L'ANOVA, comme son nom (en anglais) ne l'indique par vraiment, permet de tester trois moyennes ou plus. L'ANOVA permet donc de tester l'association entre une variable qualitative et une variable quantitative lorsque la distribution de cette dernière peut être considérée comme normale. C'est en quelque sorte une généralisation du test de Student pour séries non appariées qui, lui, ne permet de tester que *deux* moyennes. L'ANOVA doit être l'une des méthodes les plus décrites dans les « choses » (cours, livres, pdf en ligne, forum de stat', …)

de statistique, donc je ne vais que très peu en parler ici[44]. Je vais en revanche vous parler de choses à ne pas oublier, quand vous faites une ANOVA ou quand vous lisez les résultats d'une ANOVA dans un article.

Le principe de l'ANOVA est de comparer la variance inter-groupe (le groupe étant l'une des catégories de la variable qualitative) à la variance intra-groupe (cf. **Figure 12**).

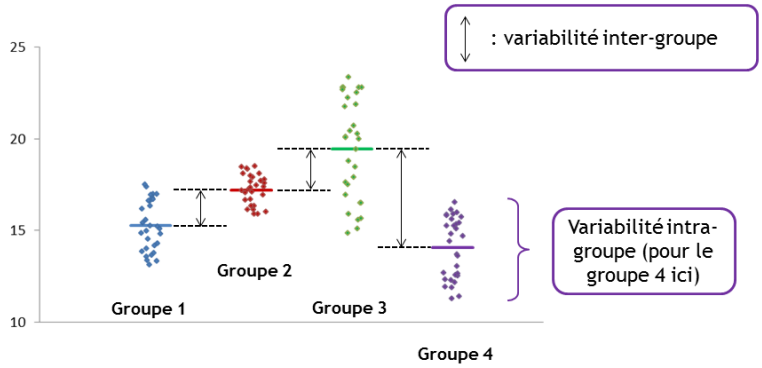


Figure 12. Représentation graphique de l'ANOVA. Le « groupe » représente la variable qualitative (ici, une variable qualitative en quatre catégories).

2. *Hypothèse nulle dans le test de l'ANOVA*

Il existe là encore deux façons, au choix, de citer l'hypothèse nulle H_0 d'un test de l'ANOVA.

- « Dans la population [cible], la moyenne [du caractère quantitatif] parmi [les individus du groupe n°1], celle parmi [les individus du groupe n°2], etc. sont toutes égales ».
- « Dans la population [cible], il n'existe pas d'association réelle entre [le caractère quantitatif] et [l'appartenance au groupe 1, 2, etc.] ».

3. *Conditions de validité du test de l'ANOVA*

Deux conditions doivent être vérifiées pour que le test de l'ANOVA soit valide.

1) Les individus de l'échantillon doivent être considérés comme indépendants.

2) La distribution de la variable quantitative dont on calcule les différentes moyennes doit être considérée comme normale dans la population dont est issu l'échantillon.

4. *Conclusion à l'issue du test de l'ANOVA*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau l'échantillon : « Dans l'échantillon, la moyenne [du caractère quantitatif] parmi [les individus du groupe n°1] ([m_1]), celle parmi [les individus du groupe n°2] ([m_2]), etc. étaient significativement différentes (p = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour qu'il existe une association réelle entre [le caractère quantitatif] et [le fait d'appartenir au groupe 1, 2, etc.]. »

[44] Car (1) si vous souhaitez en savoir plus sur l'ANOVA, vous trouverez les choses très facilement sur Internet, et (2) ce que vous trouverez sur Internet et qui n'est pas dans ce polycopié n'est pas au programme.

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la moyenne [du caractère quantitatif] parmi [les individus du groupe n°1] ([m_1]), celle parmi [les individus du groupe n°2] ([m_2]), etc. n'étaient pas significativement différentes (*p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

5. *Commentaires*

Attention, tout comme le test statistique du Chi^2 (ou de Fisher) testant l'association entre une variable binaire et une variable qualitative (variable à trois classes ou plus), il est interdit de dire que la moyenne observée la plus élevée parmi toutes celles citées est « significativement supérieure » à toutes les autres qui sont moins élevées. L'hypothèse nulle H_0 étant l'égalité des moyennes, le rejet de l'hypothèse nulle ne permet de dire qu'une seule chose, c'est qu'elles sont *globalement* significativement différentes (mais pas *une* en particulier).

G. Le test de Mann-Whitney (comparaison de deux médianes)

1. *Contexte du test de Mann-Whitney*

Le test statistique de Mann-Whitney fait partie des tests statistiques dits « non paramétriques », c'est-à-dire qu'ils ne sont pas basés sur des hypothèses de distribution de probabilités[45]. Il est notamment utilisé lorsque le test de Student pour séries non appariées ne peut pas être utilisé en raison d'une distribution du caractère quantitatif qui ne peut pas être considérée comme normale. Le test de Mann-Whitney teste si deux médianes sont ou non significativement différentes. Bien entendu, si l'on souhaite d'emblée comparer des médianes et plutôt que des moyennes, alors on réalise d'emblée un test de Mann-Whitney.

2. *Hypothèse nulle dans le test de Mann-Whitney*

Il existe là encore deux façons, au choix, de citer l'hypothèse nulle H_0 d'un test de Mann-Whitney.

- « Dans la population [cible], la médiane [du caractère quantitatif] parmi [les individus du groupe A] est égale à la médiane [du caractère quantitatif] parmi [les individus du groupe B] ».
- « Dans la population [cible], il n'existe pas d'association réelle entre [le caractère quantitatif] et [l'appartenance au groupe A ou B] ».

3. *Démarche de calcul du test de Mann-Whitney*

Le test de Mann-Whitney est un test de somme de rangs. Le principe est décrit dans la **Figure 13**, en prenant comme exemple la comparaison de la médiane de croissance pondérale de chatons entre deux groupes (les groupes 1 et 2 sur la figure). La première étape consiste à classer tous les chatons par ordre de croissance pondérale croissante, indépendamment du groupe d'appartenance. La deuxième étape consiste à regrouper les chatons par groupe, puis à faire la somme des rangs dans chacun des deux groupes. Le test de Mann-Whitney teste si la somme des rangs dans le premier groupe est significativement différente de la somme des

[45] Cf. https://en.wikipedia.org/wiki/Nonparametric_statistics

rangs dans le second groupe. Cela revient à tester si la médiane dans le premier groupe est significativement différente de la médiane dans le second groupe.

Groupe	1	2	2	1	1	2	1	1	2	2	1	1	2	2
Croissance ponderale	16.5	17.8	17.9	18.3	18.4	18.7	19.2	19.2	19.8	20.3	21.5	22.3	22.4	25.7
Rang	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Tri par croissance pondérale croissante →

Regroupement ↓

Groupe	1	1	1	1	1	1	1	2	2	2	2	2	2	2
Croissance ponderale	16.5	18.3	18.4	19.2	19.2	21.5	22.3	17.8	17.9	18.7	19.8	20.3	22.4	25.7
Rang	1	4	5	7	8	11	12	2	3	6	9	10	13	14
Somme des rangs	48							57						
Médiane	19.2							19.8						
Moyenne	19.4							20.4						

Figure 13. Principe du test de Mann-Whitney.

4. *Condition de validité du test de Mann-Whitney et commentaire*

Il n'y en a qu'une seule : l'indépendance des individus.

Notamment, le test statistique de Mann-Whitney fonctionne très bien lorsque le caractère quantitatif suit une loi normale. Alors pourquoi encore utiliser le test de Student quand celui de Mann-Whitney fonctionne très bien quelle que soit la distribution du caractère quantitatif ? Parce qu'une moyenne étant plus facile à interpréter qu'une médiane pour certains chercheurs (je trouve au contraire qu'une médiane est plus facile à interpréter qu'une moyenne, mais ce n'est qu'un jugement personnel), ces chercheurs préfèrent tester des moyennes – lorsqu'ils le peuvent[46].

5. *Conclusion à l'issue du test de Mann-Whitney*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la médiane [du caractère quantitatif] parmi [les individus du groupe A] ([valeur de med_A]) était significativement [supérieure, ou inférieure] à la médiane [du caractère quantitatif] parmi [les individus du groupe B] ([valeur de med_B] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour que la médiane [du caractère quantitatif] parmi [les individus du groupe A] ([valeur de med_A]) soit réellement [supérieure, ou inférieure] à la médiane [du caractère quantitatif] parmi [les individus du groupe B]. »

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la médiane [du caractère quantitatif] parmi [les individus du groupe A] ([valeur de med_A]) n'était pas significativement différente de la médiane [du caractère quantitatif] parmi [les individus du groupe B] ([valeur de med_B] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

[46] Lorsque la distribution du caractère quantitatif peut être considérée comme normale.

H. Le test de Kruskal-Wallis

1. *Contexte du test de Kruskal-Wallis*

Le test de Kruskal-Wallis peut être vu comme une généralisation du test de Mann-Whitney, de la même façon que l'ANOVA peut être vue comme une généralisation du test de Student pour séries non appariées. Le test de Kruskal-Wallis est un test non paramétrique testant trois médianes ou plus. Les détails de ce test peuvent se retrouver ici[47].

2. *Hypothèse nulle dans le test de Kruskal-Wallis*

Il existe là encore deux façons, au choix, de citer l'hypothèse nulle H_0 d'un test Kruskal-Wallis.

- « Dans la population [cible], la médiane [du caractère quantitatif] parmi [les individus du groupe n°1], celle parmi [les individus du groupe n°2], etc. sont toutes égales ».
- « Dans la population [cible], il n'existe pas d'association réelle entre [le caractère quantitatif] et [l'appartenance au groupe 1, 2, etc.] ».

3. *Condition de validité du test de Kruskal-Wallis*

Il n'y en a qu'une seule : l'indépendance des individus.

4. *Conclusion à l'issue du test de Kruskal-Wallis*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la médiane [du caractère quantitatif] parmi [les individus du groupe n°1] ([med_1]), celle parmi [les individus du groupe n°2] ([med_2]), etc. étaient significativement différentes (p = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour qu'il existe une association réelle entre [le caractère quantitatif] et [le fait d'appartenir au groupe 1, 2, etc.]. »

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la médiane [du caractère quantitatif] parmi [les individus du groupe n°1] ([med_1]), celle parmi [les individus du groupe n°2] ([med_2]), etc. n'étaient pas significativement différentes (p = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

5. *Commentaires*

Attention, tout comme le test statistique de l'ANOVA, il est interdit de dire que la médiane observée la plus élevée parmi toutes celles citées est « significativement supérieure » à toutes les autres qui sont moins élevées. L'hypothèse nulle H_0 étant l'égalité des médianes, le rejet de l'hypothèse nulle ne permet de dire qu'une seule chose, c'est qu'elles sont globalement significativement différentes (mais pas *une* en particulier).

[47] Cf. https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance

I. Les coefficients de corrélation

1. *Contexte des coefficients de corrélation*

Un coefficient de corrélation est un indicateur qui quantifie l'association entre deux variables quantitatives V_1 et V_2. Sa valeur est comprise entre -1 et +1. Un coefficient de corrélation égal à -1 signifie que $V_1 = k.V_2$ avec $k < 0$ (V_1 est alors parfaitement anti-corrélée à V_2). Un coefficient de corrélation égal à +1 signifie que $V_1 = k.V_2$ avec $k > 0$ (V_1 est alors parfaitement corrélée à V_2). Un coefficient de corrélation égal à 0 indique que V_1 et V_2 sont indépendantes (c'est-à-dire qu'elles ne sont absolument pas associées). Si la distribution de V_1 et celle de V_2 peuvent toutes les deux être considérées comme normales, il faut calculer un coefficient de corrélation paramétrique : le coefficient de corrélation de Pearson. Si ce n'est pas le cas, il faut calculer un coefficient de corrélation non paramétrique : le coefficient de corrélation de Spearman. Un test statistique est associé à la valeur du coefficient de corrélation. Il teste si la valeur du coefficient de corrélation est significativement différente de 0.

2. *Hypothèse nulle dans le test du coefficient de corrélation*

L'hypothèse nulle H_0 du test statistique testant le coefficient de corrélation est la suivante : « Dans la population [cible], il n'existe pas de corrélation réelle entre [le caractère quantitatif n°1] et [le caractère quantitatif n°2]. »

3. *Condition de validité du test statistique des coefficients de corrélation*

Que le coefficient de corrélation utilisé soit celui de Pearson ou bien de Spearman, les individus de l'échantillon doivent être considérés comme indépendants.

Ensuite, pour que le coefficient de corrélation de Pearson puisse être utilisé, les deux variables quantitatives doivent suive une distribution considérée comme normale dans la population dont est issu l'échantillon. Il n'y a pas de condition sur la distribution des deux variables quantitatives pour utiliser le coefficient de corrélation de Spearman.

4. *Conclusion à l'issue du test du coefficient de corrélation*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, [le caractère quantitatif n°1] était significativement corrélé [au caractère quantitatif n°2] ([valeur du coefficient de corrélation] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour qu'il existe une corrélation réelle entre [le caractère quantitatif n°1] et [le caractère quantitatif n°2]. »

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, [le caractère quantitatif n°1] n'était pas significativement corrélé [au caractère quantitatif n°2] ([valeur du coefficient de corrélation] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

VIII. LES TESTS STATISTIQUES SUR DONNEES NON INDEPENDANTES EN PRATIQUE

A. Préambule

Lorsque les individus ne sont pas indépendants, les méthodes statistiques sont beaucoup plus compliquées.

La seule situation d'utilisation de tests statistiques lorsque les individus ne sont pas indépendants qui est au programme est la suivante : celle où l'on veut montrer que, dans une population d'animaux, la valeur d'un indicateur évolue réellement au cours du temps entre un instant t_0 et un instant t_1. Dans cette situation, les animaux d'un échantillon sont évalués deux fois (une fois à t_0 et une fois à t_1). On parle dans cette situation de « séries appariées ». En effet, la série de mesures réalisées à t_0 parmi les animaux de l'échantillon est « appariée » à la série de mesures réalisées à t_1 sur ces *mêmes* animaux. Chaque animal possède une « paire de (deux) mesures » (une à chaque temps). Attention, l' « individu » n'est plus l'animal mais la « mesure ». Et la situation de « séries appariées » est donc une situation de non indépendance des « individus » car les valeurs des mesures réalisées à t_0 ne sont pas indépendantes de celles réalisées à t_1 : la mesure à t_1 a été réalisée sur le même animal que celle réalisée à t_0.

Dans la grande majorité des cas, la question de recherche est la suivante : est-ce que la valeur d'un caractère est réellement différente entre juste avant une intervention (t_0 ; la mise sous un traitement, par exemple) et « quelque temps » après cette intervention (t_1). La **Figure 14** illustre cette situation de « séries appariées ». Le caractère peut être binaire, qualitatif ou quantitatif. Il est évalué à t_0 (CAR_0), moment juste avant l'intervention, et à un instant t_1 après intervention, l'animal est ré-évalué (CAR_1).

Dans le cadre du module de Biostatistique en Médecine Vétérinaire, seul le cas de figure où le caractère est quantitatif sera traité. De plus, nous ferons l'hypothèse que la distribution de ce caractère quantitatif pourra être considérée comme normale. Ainsi, l'indicateur utilisé sera la moyenne du caractère quantitatif, qui sera estimée à t_0 et à t_1. Nous utiliserons le test de Student pour séries appariées pour savoir si la moyenne du caractère estimée à t_0 est significativement différente de celle estimée à t_1. Les autres tests statistiques sur séries appariées ne sont pas au programme du module. Ils sont néanmoins présentés dans ce polycopié en Annexe 6, au cas où vous en auriez besoin pour des analyses statistiques futures (notamment pour votre thèse vétérinaire).

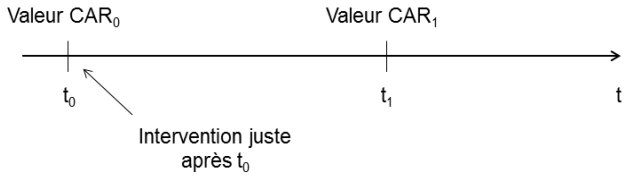


Figure 14. Illustration d'un cas fréquent de « séries appariées ». « CAR_0 » représente la valeur du caractère mesuré à t_0, et « CAR_1 » celle à t_1.

 – Biostatistique en Médecine Vétérinaire – version v3.7

B. Le test de Student pour séries appariées (comparaison de deux moyennes)

1. *Contexte du test de Student pour séries appariées*

Un caractère quantitatif est mesuré à deux moments différents (t_0 et t_1) sur les animaux d'un échantillon, et l'on souhaite montrer qu'il existe une réelle évolution entre t_0 et t_1 de ce caractère quantitatif au niveau de la population cible.

2. *Démarche de calcul du test de Student pour séries appariées*

La démarche de calcul de ce test est la suivante : pour chacun des N animaux de l'échantillon, le logiciel calcule la différence de la valeur du caractère entre t_1 et t_0. Ensuite, la moyenne de toutes ces N différences (une par animal) doit être testée à 0 (c'est-à-dire que l'on va vouloir savoir si la moyenne des différences observées entre t_1 et t_0 est significativement différente de 0). Nous verrons une application de ce test statistique en TD.

3. *Hypothèse nulle dans le test de Student pour séries appariées*

Il existe là encore deux façons, au choix, de citer l'hypothèse nulle H_0 d'un test de Student pour séries appariées :

- « Dans la population [cible], la moyenne [du caractère quantitatif] [à t_0] est égale à la moyenne [du caractère quantitatif] [à t_1] ».
- « Dans la population [cible], il n'existe pas d'évolution réelle [du caractère quantitatif] [entre t0 et t1] ».

4. *Conditions de validité du test de Student pour séries non appariées*

Deux conditions doivent être vérifiées pour que le test de Student pour séries appariées soit valide.

1) Les animaux de l'échantillon doivent être considérés comme indépendants.

2) La distribution de la variable quantitative dont on calcule les deux moyennes (à t_0 et à t_1) doit être considérée comme normale dans la population dont est issu l'échantillon.

5. *Conclusion à l'issue du test de Student pour séries non appariées*

a) Lors du rejet de H_0 ($p \leq 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la moyenne [du caractère quantitatif] [à t_0] ([valeur de m_{t0}]) était significativement [supérieure, ou inférieure] à la moyenne [du caractère quantitatif] [à t_1] ([valeur de m_{t1}] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : « Sous l'hypothèse d'absence de biais d'association, dans [la population cible], il y a des chances pour que [le caractère quantitatif] évolue[48] réellement [entre t_0 et t_1]. »

[48] On peut remplacer ici « évolue » par « augmente » ou « diminue » en fonction de la valeur des deux moyennes m_{t0} et m_{t1}.

b) Lors de l'acceptation de H_0 ($p > 0,05$)

Conclusion au niveau de l'échantillon : « Dans l'échantillon, la moyenne [du caractère quantitatif] [à t_0] ([valeur de m_{t0}]) n'était pas significativement différente de la moyenne [du caractère quantitatif] [à t_1] ([valeur de m_{t1}] ; *p* = [valeur du degré de signification]). »

Conclusion au niveau de la population cible (inférence) : impossible.

IX. LA PUISSANCE STATISTIQUE D'UNE ETUDE

A. Remarque préliminaire

Toutes les illustrations de cette partie font référence à la comparaison de deux moyennes, avec le test de Student pour séries non appariées. Mais bien entendu, tout le raisonnement que je vais appliquer pourrait s'appliquer à n'importe quel test statistique.

De plus, je vous rappelle que Δ est la différence réelle entre les indicateurs θ_A et θ_B dans la population cible (cf. page 29), et cette différence est inconnue.

B. Définition & commentaires

La puissance statistique d'une étude est « la probabilité d'une étude à obtenir une différence d_{obs} significative entre deux indicateurs estimés dans l'échantillon quand il existe une différence réelle de valeur Δ ($\Delta \neq 0$) entre les deux indicateurs correspondants (inconnus) dans la population cible ».

Énoncée de façon plus statistique, la puissance statistique d'une étude est « la probabilité qu'a cette étude de rejeter H_0 en supposant qu'en vrai, la différence réelle vaut Δ ($\Delta \neq 0$) ». Il existe par conséquent un lien très fort entre la puissance et le risque d'erreur de 2ème espèce β. En effet, je vous rappelle que β est la probabilité d'*accepter* H_0 quand H_0 est fausse, c'est-à-dire quand la différence réelle vaut Δ avec $\Delta \neq 0$. Par conséquent, la puissance statistique vaut $1 - \beta$.

C. De quoi dépend la puissance statistique d'une étude ?

1. *En théorie*

Premièrement, puisque la puissance statistique est égale à $1 - \beta$, et puisque β dépend de Δ (cf. page 35), la puissance statistique dépend de Δ : plus Δ est grande, plus la puissance statistique de l'étude est élevée. Ce premier point est intuitif : plus la différence réelle Δ est importante, plus ce sera facile d'obtenir dans une étude une différence d_{obs} significative ($p \leq 0,05$). Cela dit, comme Δ est en quelque sorte fixée par la nature, les investigateurs d'une étude ne pourront pas « jouer » sur Δ pour augmenter la puissance statistique de leur étude.

Ensuite, l'autre paramètre qui a un impact sur la puissance statistique est le nombre d'individus dans les groupes A et B : lorsque Δ est différente de 0, plus la taille des groupes est élevée, plus la puissance statistique de l'étude est élevée. La taille des groupes est LE paramètre sur lequel les investigateurs pourront « jouer » pour augmenter la puissance statistique de leur étude. Là encore, ce point est intuitif : plus la taille des groupes A et B dans l'échantillon de l'étude est importante, plus ce sera facile d'obtenir, dans cette étude, une différence d_{obs} significative ($p \leq 0,05$).

Enfin, lorsque le caractère est quantitatif, le troisième paramètre qui a un impact sur la puissance statistique d'une étude est la variabilité de ce caractère quantitatif : *moins* le caractère quantitatif est variable (c'est-à-dire, plus la SD est faible), et plus la puissance statistique est élevée. Comme pour Δ, la variabilité du caractère quantitatif est en quelque

sorte fixée par la nature. Les investigateurs ne pourront donc pas non plus « jouer » sur cette SD pour augmenter la puissance statistique de leur étude.

2. *Illustration*

Pour illustrer ce que je viens d'écrire ci-dessus, supposons trois études cliniques testant trois traitements différents (un traitement par étude clinique) contre placebo, avec comme caractère étudié le taux de rémission (en %) d'une maladie (la même dans les trois études). Chacune de ces études comporte deux groupes : le groupe traité avec le traitement d'intérêt et le groupe recevant du placebo (⇔ absence de traitement). Supposons de plus que les vrais taux de rémission dans la population cible soient les suivants :

- 10% en l'absence de tout traitement
- 20% si traité avec le traitement #1
- 40% si traité avec le traitement #2
- 60% si traité avec le traitement #3

On va considérer que, au vu des taux de rémission ci-dessus, la différence réelle de taux de rémission entre le traitement #1 et l'absence de traitement (Δ_1) est « faible » (⇔ effet faible du traitement #1), que la différence réelle de taux de rémission entre le traitement #2 et l'absence de traitement (Δ_2) est « modérée » (⇔ effet modéré du traitement #2), et que la différence réelle de taux de rémission entre le traitement #3 et l'absence de traitement (Δ_3) est « élevée » (⇔ effet élevé du traitement #3).

La **Figure 15** ci-dessous représente l'évolution de la puissance statistique pour chacune des trois études en fonction de la taille (que l'on suppose identique) de chacun des deux groupes de chaque étude.

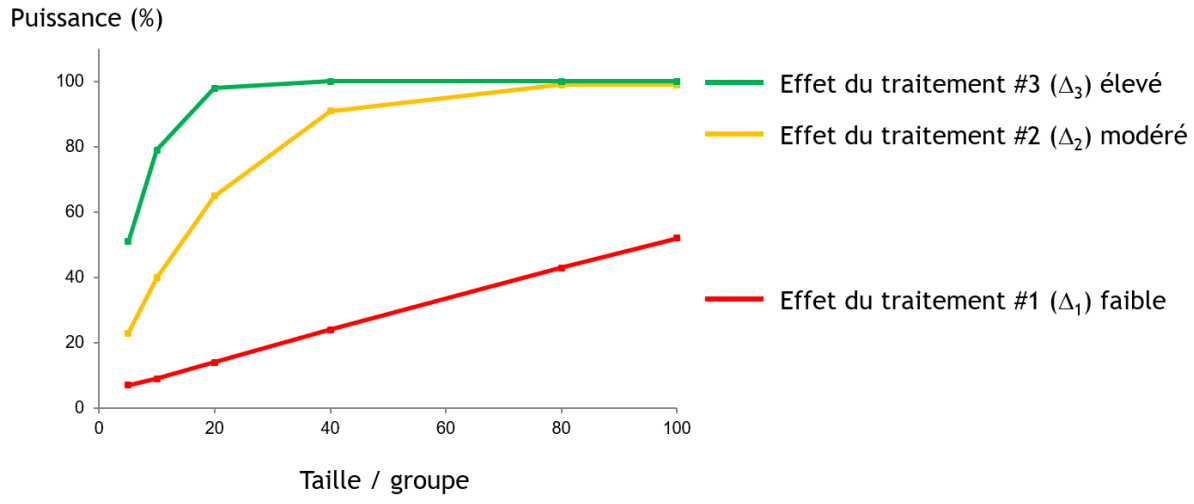


Figure 15. Lien entre puissance statistique d'une étude et taille d'échantillon dans chaque groupe, pour trois différentes études (chaque étude étudiant l'effet d'un traitement contre placebo).

Comme je l'ai écrit dans la partie théorique juste au-dessus, vous pouvez voir que, pour un effet réel fixé, la puissance statistique augmente avec l'augmentation de chacun des deux groupes, et qu'ensuite, à taille de groupe égale, la puissance statistique augmente avec l'augmentation de l'effet réel.

D. Invocation du manque de puissance statistique (hors programme)

Il arrive parfois (malheureusement pour les investigateurs d'une étude) qu'une étude ait manqué de puissance statistique. Qu'est-ce que cela signifie ? Cela signifie qu'une étude n'a pas réussi à obtenir une différence significative dans l'échantillon ($p > 0,05$) alors que cette différence existe pourtant réellement dans la population cible.

En pratique, on invoque le manque de puissance statistique quand on a *failli* obtenir une différence significative et que par ailleurs on pense qu'elle existe réellement. La plage de valeurs du degré de signification *p* communément admise pour avoir « *failli* obtenir une différence significative » est]0,05 ; 0,10]. Et ce qui fait penser qu'il existe une réelle différence, c'est le fait d'observer une différence (ou un effet) qui soit *cliniquement importante* dans l'échantillon. Ainsi, on invoque le manque de puissance statistique si les deux critères ci-dessous sont *tous les deux* vérifiés :

- Le degré de signification $p \in]0,05 ; 0,10]$;
- La différence observée entre les deux groupes comparés est jugée comme cliniquement importante (appréciation subjective à partir du ou des indicateurs estimés dans l'échantillon).

Dans l'exemple des deux études fictives sur l'association entre la race des vaches et la présence de mammites de la page 39, nous dirions que la 1ère étude a manqué de puissance statistique car (1) le degré de signification *p* (de valeur 0,07) était compris entre 0,05 et 0,10, et (2) la différence de pourcentages (44% *versus* 17%) pouvait être subjectivement jugée comme suffisamment importante pur qu'elle nous laisse penser à une différence réelle de pourcentages (l'appréciation du manque de puissance statistique *est* subjective).

X. NOMBRE D'INDIVIDUS A INCLURE DANS UNE ETUDE CLINIQUE

A. Introduction

Nous avons vu que lorsque H_0 est acceptée ($p > 0,05$), il est impossible de faire quelque inférence que ce soit (la seule conclusion possible est celle au niveau de l'échantillon, qui dit « il n'existait pas d'association significative entre les deux variables étudiées »). En revanche, sous certaines conditions (puissance statistique élevée et caractère « confirmatoire » de l'étude), on peut être convaincu que H_0 est fausse lorsqu'elle est rejetée (lorsque $p \leq 0,05$).

Par conséquent, il est fondamental de concevoir une étude qui maximise les chances de rejeter H_0. En effet, si, après des mois ou des années de mise en place de l'étude et de collecte des données, l'association étudiée est finalement non significative, tout ce temps (et tout l'argent dépensé) n'aura servi à rien[49]. Or, la définition de la puissance statistique d'une étude est : « la probabilité qu'a cette étude de rejeter H_0 en supposant qu'en vrai, la différence réelle vaut Δ ($\Delta \neq 0$) ». Ainsi, l'étude doit être mise en place avec une puissance statistique suffisante (suffisamment élevée pour avoir de bonnes chances de ne pas avoir travaillé toutes ces années pour rien !).

B. Taille d'échantillon et capacité d'une étude à rejeter H_0 : attention au piège !

Si je vous dis « dans certains cas, vous pourrez augmenter votre taille d'échantillon à l'infini, cela n'augmentera pas vos chances de rejeter H_0 », est-ce que vous êtes surpris ? Vous ne devriez pas ! Car c'est vrai !

Quelle est la définition du risque d'erreur de 1ère espèce α, déjà ? (Essayez de vous en souvenir, là, et non pas de vous ruer sur la réponse !) (…) C'est « la probabilité de rejeter H_0 lorsque H_0 est vraie ». Et on a vu que dès que le seuil du degré de signification pour dire qu'une association significative est fixé (à 0,05), la valeur de α est fixée à cette valeur (0,05). Par conséquent, lorsque H_0 est vrai, il y a *toujours* α% de risques de rejeter H_0, par conséquent, *indépendamment* de la taille de l'échantillon !!

Lorsque H_0 est vraie, la taille de l'échantillon n'a donc aucun impact, et *heureusement*, sur la capacité d'une étude à rejeter H_0. Pourquoi « *heureusement* » ? Nous allons voir cela tout de suite.

Est-ce que vous pensez que le produit Oscillococcinum®, produit homéopathique traditionnellement utilisé dans le traitement des états grippaux, est efficace pour prolonger la vie de patients atteints d'un cancer du poumon ? Evidemment que non. Autrement dit, H_0 est vraie (c'est-à-dire, « dans la population des millions de patients atteints d'un cancer du poumon, il y a égalité du taux de survie entre ceux qui reçoivent de l'Oscillococcinum® et ceux qui reçoivent un placebo »). Par conséquent, si le laboratoire Boiron, qui commercialise l'Oscillococcinum®, mettait en place un essai clinique chez des patients atteints d'un cancer du poumon comparant un groupe de patients mis sous placebo et un groupe de patients mis sous Oscillococcinum®, Boiron aurait 5% de risques[50] d'observer une différence significative

[49] Si, elle aura servi à écrire dans l'article « une nouvelle étude avec davantage de sujets est nécessaire pour montrer ce que nous, nous n'avons pas réussi à montrer » !

[50] Mais de « chances » en ce mettant du côté de Boiron 😊

sur le taux de survie entre les deux groupes : c'est la définition du risque d'erreur de 1ère espèce α. Et *heureusement* que cette probabilité de 5% est fixe, et ne dépend pas de la taille d'échantillon, car si cela n'avait pas été le cas, il suffirait à Boiron de mettre beaucoup d'argent sur la table pour recruter des millions de patients atteints de cancer du poumon pour augmenter ses chances (risques) de montrer quelque chose … de faux ! ☺

En revanche, lorsque H_0 est fausse (c'est-à-dire, lorsqu'il existe une réelle différence Δ entre les deux groupes A et B dans la population, aussi petite soit elle), là, oui, augmenter la taille de l'échantillon va augmenter les chances de rejeter H_0 ! (Donc, là, oui, tout est une question de temps et d'argent…)

C. Calcul du nombre d'individus à inclure dans une étude

1. *Remarque préliminaire*

Sur la **Figure 15**, vous pouvez vous rendre compte qu'il n'y a pas de nombre « magique » pour la taille d'échantillon. En effet, lorsque l'effet est élevé (courbe verte), avec 10 individus par groupe (ce qui pourrait être considéré par certains comme une « faible » taille d'échantillon), la puissance statistique de l'étude #3 est quand même de 80%. Et lorsque l'effet est faible (courbe rouge), avec 90 individus par groupe (ce qui pourrait être considéré par certains comme une taille d'échantillon « importante »), la puissance statistique de l'étude #1 est inférieure à 50%.

La bonne taille d'échantillon n'est donc pas un nombre unique (par exemple et par hasard, « 30 », comme on peut souvent l'entendre), mais il doit se calculer au cas par cas.

2. *Introduction à la démarche de calcul*

Pour calculer le nombre d'individus à inclure dans une étude, il faut se souvenir que la puissance statistique dépend de la différence réelle Δ entre les deux indicateurs θ_A et θ_B de chacun des deux groupes A et B dans la population cible, de la taille d'échantillon (autrement dit, du nombre d'individus dans chacun des deux groupes A et B de l'échantillon), et de la variabilité du caractère lorsque le caractère étudié est quantitatif (cf. partie IX.C, page 63).

On pourrait ainsi écrire la puissance comme une fonction F de plusieurs variables, en fonction du type de caractère (binaire ou quantitatif) :

Lorsque le caractère est binaire : Puissance = $F_{bin}(\Delta, N_A, N_B)$, avec Δ représentant la différence réelle dans la population entre le pourcentage du caractère parmi les individus du groupe A (π_A) et le pourcentage du caractère parmi les individus du groupe B (π_B), N_A et N_B respectivement le nombre d'individus dans les groupes A et B de l'échantillon.

Lorsque le caractère est quantitatif : Puissance = $F_{quant}(\Delta, N_A, N_B, \sigma)$, avec Δ la valeur de la différence réelle dans la population entre la moyenne du caractère parmi les individus du groupe A (μ_A) et la moyenne du caractère parmi les individus du groupe B (μ_B), N_A et N_B respectivement le nombre d'individus dans les groupes A et B de l'échantillon, et σ la valeur de la Standard Deviation (SD) dans la population.

Ce que l'on souhaite, c'est déterminer les tailles des groupes A (N_A) et B (N_B) dans l'échantillon. Ainsi, on peut ré-écrire les équations ci-dessus à l'aide d'une autre fonction G de la façon suivante :

Lorsque le caractère est binaire : $(N_A, N_B) = G_{bin}(\text{Puissance}, \pi_A, \pi_B)$

Lorsque le caractère est quantitatif : $(N_A, N_B) = G_{quant}(\text{Puissance}, \mu_A - \mu_B, \sigma)$

(J'ai écrit « π_A, π_B » et non pas « π_A - π_B », ce n'est pas une erreur. L'explication, hors programme, est en rapport avec ce que j'avais écrit dans la note de bas de page n°39, page 49, concernant la démarche de calcul du test du Chi^2 avec une différence entre deux pourcentages qui n'a pas vraiment de sens.)

3. *Difficultés psychologiques qu'il faut lever avant le calcul*

Ainsi, à partir des deux fonctions G_{bin} et G_{quant} écrites plus haut (dont il n'est pas question que je vous donne l'expression), vous vous rendez compte de quelque chose qui *va* vous dérouter, c'est que pour connaître le nombre d'individus à inclure dans les groupes A et B dont l'étude va avoir besoin, il faut (entre autres) fournir les valeurs de π_A et π_B ou la valeur de $\mu_A - \mu_B$.

Pourquoi est-ce déroutant ? Parce que l'on ne connait pas les valeurs de π_A et π_B ou celle de $\mu_A - \mu_B$ puisque ce sont les valeurs des indicateurs dans la population cible ! Donc, pour calculer N_A et N_B, il faut faire des hypothèses (encore !) sur les valeurs supposées de π_A et π_B ou sur la valeur supposée de $\mu_A - \mu_B$ (et aussi sur la valeur de σ si le caractère est quantitatif). Ces hypothèses doivent être faites par les investigateurs de l'étude clinique. Et ces investigateurs *peuvent* faire de telles hypothèses. Ils en ont les capacités.

En effet, quand les investigateurs d'une future étude clinique commencent à rédiger le protocole de cette future étude, ils ont une idée derrière la tête[51] ! Et ce que les calculs demandent, c'est justement d'avoir une « idée » de π_A et π_B ou celle de $\mu_A - \mu_B$, et non pas de fournir dans les calculs les vraies valeurs (inconnues) de π_A et π_B ou celle de $\mu_A - \mu_B$.

Par ailleurs, et pour vous rassurer, je ne vous demanderai jamais en examen de faire vous-même des hypothèses sur les valeurs de π_A et π_B ou sur la valeur de $\mu_A - \mu_B$. C'est moi qui fournirai ces valeurs auxquelles on peut s'attendre avant d'avoir réalisé l'étude.

4. *Utilisation d'un site Internet pour calculer le nombre d'individus à inclure dans une étude et interprétation des résultats fournis*

Nous utiliserons en TD le site Internet BiostaTGV dans sa partie dédiée au nombre d'individus à inclure dans une étude[52], à partir de la puissance statistique souhaitée, et des valeurs supposées de π_A et π_B (si le caractère est binaire) ou des valeurs supposées de μ_A et μ_B et de σ[53] (si le caractère est quantitatif). La **Figure 16** vous présente un exemple de remplissage du site Internet dans le cas d'un caractère quantitatif. Dans cet exemple, on voulait obtenir le nombre d'individus à inclure dans une étude dont on souhaitait qu'elle ait une puissance statistique de 80% pour mettre en évidence une différence significative (au seuil $\alpha = 0,05$) entre deux groupes 1 et 2 sur la moyenne d'un caractère quantitatif, en faisant l'hypothèse que les vraies moyennes de ce caractère au niveau de la population dans les groupes 1 et 2

[51] Basée sur de nombreuses choses : les articles déjà publiés, les conférences, l'expérience professionnelle, l'intuition médicale, etc.

[52] https://biostatgv.sentiweb.fr/?module=etudes/sujets

[53] la valeur de la SD dans la population

sont respectivement de 12 et 15 (soit une différence Δ = 3) et que la vraie valeur de la SD vaut 4 (σ = 4).

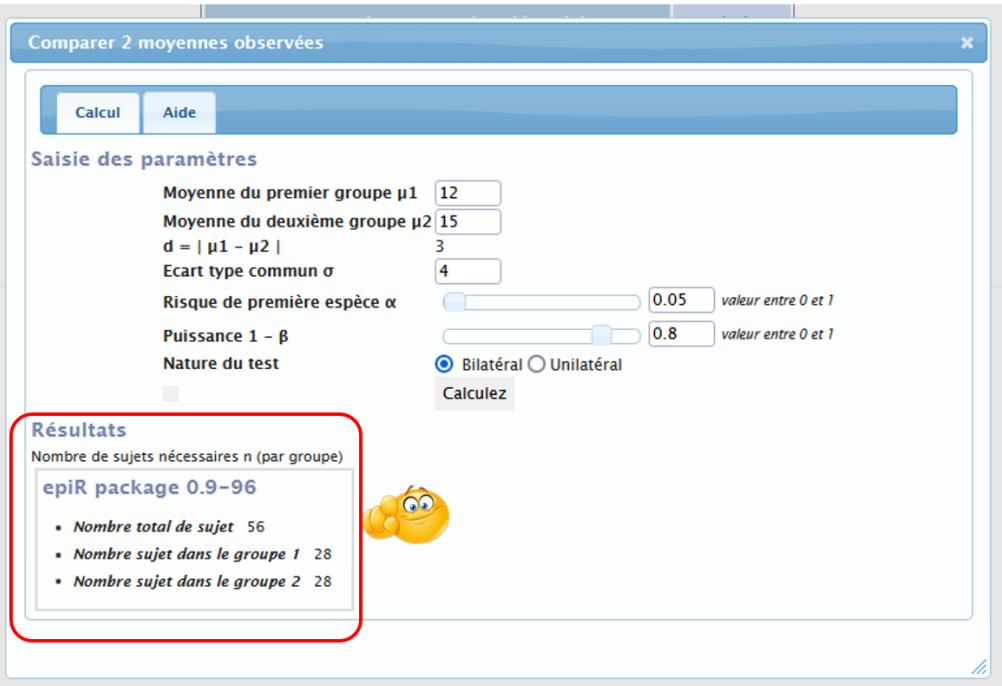


Figure 16. Copie d'écran du site Internet https://biostatgv.sentiweb.fr/?module=etudes/sujets# pour calculer la taille des groupes 1 et 2 dans le cas d'un caractère quantitatif (cf. texte pour les hypothèses formulées).

Après avoir cliqué sur « Calculez », le site fournit le nombre total d'individus à inclure de 56, soit 28 dans chaque groupe.

L'interprétation de ce « 28 » est fondamentale, et la voici : « si l'on veut mettre en place une étude clinique qui a 80% de chances d'obtenir dans l'échantillon une différence significative de moyennes du caractère quantitatif entre les groupes 1 et 2 au risque d'erreur de 1ère espèce α fixé à 0,05, en supposant que les moyennes de ce caractère au niveau de la population cible soient de 12 et 15 respectivement pour les groupes 1 et 2 (soit une différence réelle Δ = 3) et en supposant que la SD, dans la population cible, soit égale à 4, alors il faudra recruter 56 individus dans l'étude clinique au total, avec 28 individus dans le groupe 1 et 28 individus dans le groupe 2. »

XI. ANNEXES

A. Annexe 1 – Compétence « Agir en scientifique » du référentiel national du diplôme vétérinaire

AGIR EN SCIENTIFIQUE

Sc1. Porter une analyse critique et évaluer la bibliographie et des communications

Connaissances sous-jacentes

Anglais
Biostatistiques
Epidémiologie
Médecine
Sciences du vivant et vétérinaires
Techniques de documentation et bases de bibliométrie

Indicateurs (la compétence de l'étudiant sera évaluée sur…)

- Réalisation à l'écrit ou à l'oral d'une synthèse bibliographique sur un sujet de médecine vétérinaire ou de sciences du vivant
- Présentation à l'écrit ou à l'oral d'un article sur un sujet de médecine vétérinaire ou de sciences du vivant en en faisant une analyse critique

Capacités

Sc.1.1 Objectiver son besoin d'information en définissant une question scientifique et élaborant une stratégie de recherche de l'information

4A : a fait
5A : sait faire

Sc.1.2 Rechercher l'information (scientifique, réglementaire, recommandations et bonnes pratiques)

4A : a fait
5A : sait faire

Sc.1.3 Exploiter l'information en en faisant une analyse critique

	4A	5A
Critiquer l'information et analyser la pertinence des résultats d'une recherche par rapport à la question posée	a fait	sait faire
Interpréter les matériels et méthodes et les confronter à la conclusion des auteurs		
Reconnaître les biais méthodologiques		
Identifier un conflit d'intérêt		

Sc2. Appliquer l'analyse critique de l'organisation des soins et la médecine fondée sur les preuves (Evidence-based veterinary medicine, EBVM)

Connaissances sous-jacentes

Biostatistiques
Médecine vétérinaire fondée sur les preuves
Sciences du vivant et vétérinaires
Techniques de documentation

Indicateurs (la compétence de l'étudiant sera évaluée sur…)

- Présentation de travaux expérimentaux et de cas cliniques d'intérêt selon les règles de présentation scientifique (thèse vétérinaire, article, présentation orale en congrès, posters)
- Présentation de synthèses méthodologiques de manière scientifique
- Dans l'exercice clinique, reconnaître des situations de pratique de la médecine fondée sur les preuves et d'autres types de situations

Capacités

Sc.2.1 Appliquer la médecine fondée sur les preuves sur des ressources déjà existantes

	4A	5A
Justifier scientifiquement une démarche clinique existante	a fait	sait faire
Analyser de manière critique les approches sémiologiques, cliniques et thérapeutiques		
Reconnaître les biais méthodologiques		
Savoir reconnaître le besoin d'une information complémentaire, dans un cadre clinique et thérapeutique		

Sc.2.2 Appliquer la médecine fondée sur les preuves dans la prise de décision clinique et thérapeutique

	4A	5A
Savoir assurer la traçabilité et la démarche scientifique pour prendre une décision	a fait	sait faire
Justifier scientifiquement une nouvelle démarche clinique, différente de la pratique habituelle		

Sc.2.3 Présenter de manière critique et pondérée selon leur pertinence les différents résultats publiés concernant une approche sémiologique, clinique ou thérapeutique, en utilisant pour cela une grille issue de la médecine fondée sur les preuves

4A : a vu
5A : sait faire

Sc3. Contribuer à l'accroissement des connaissances en médecine vétérinaire et plus largement dans le domaine des sciences du vivant

Connaissances sous-jacentes

Anglais
Biostatistiques
Epidémiologie
Ethique
Organisation de la recherche
Réglementation
Sciences du vivant et vétérinaires

Indicateurs (la compétence de l'étudiant sera évaluée sur…)

- Présentation de travaux expérimentaux et de cas cliniques d'intérêt selon les règles de présentation scientifique (thèse vétérinaire, article, présentation orale en congrès, posters)
- Le cas échéant, présentation d'un dossier complet pour intégrer un Master 2 dans un domaine des sciences du vivant, pour obtenir une bourse de master ou de doctorat ou pour s'inscrire de manière dérogatoire en doctorat

Capacités

Sc.3.1 Identifier une question médicale ou scientifique à résoudre

4A : a vu
5A : sait faire selon dominante

Sc.3.2 Elaborer et appliquer la méthodologie scientifique requise (protocole expérimental et démarche qualité)

4A : a vu
5A : sait faire si dominante recherche

Sc.3.3 Etre acteur de la recherche : participer à un essai clinique ou un projet de recherche, incluant l'acquisition des données, leur analyse, leur présentation selon les règles scientifiques, leur interprétation et la formulation d'une discussion critique

4A : a vu
5A : a fait

B. Annexe 2 – Ne pas utiliser de test statistique pour vérifier la normalité d'une distribution d'une variable quantitative

Cette Annexe 2 ne peut pas être lue avant que la partie V « Bases théoriques des tests statistiques » ne soit lue.

Dans certains articles, on peut lire la chose suivante : « La normalité de la distribution des variables quantitatives a été testée à l'aide du test de Shapiro-Wilk[54]. »

L'hypothèse nulle (H_0) du test de Shapiro-Wilk (comme celle de tous les autres tests statistiques testant la normalité d'une distribution) est : « la variable quantitative suit une distribution normale dans la population cible ». Et les auteurs de ces articles considèrent que la variable quantitative suit approximativement une distribution normale si la valeur du degré de signification est > 0,05. Pourquoi ce raisonnement est-il faux et, par conséquent, pourquoi une valeur du degré de signification > 0,05 du test de Shapiro- Wilk ne signifie pas que la variable quantitative suit approximativement une distribution normale ?

Parce que, lorsque l'hypothèse nulle H_0 ne peut pas être rejetée ($\Leftrightarrow$ lorsque la valeur du degré de signification est > 0,05), on ne peut *rien* dire, et surtout pas dire que H_0 est vraie (dire « la variable quantitative suit une distribution normale dans la population cible » ; cf. page 36). Le test de Shapiro-Wilk (ou tous ceux testant la normalité) doit être utilisé uniquement pour suggérer qu'un caractère quantitatif *ne* suit *pas* une distribution normale dans la population. Il ne peut pas être utilisé pour suggérer qu'un caractère quantitatif *suit* une distribution normale dans la population.

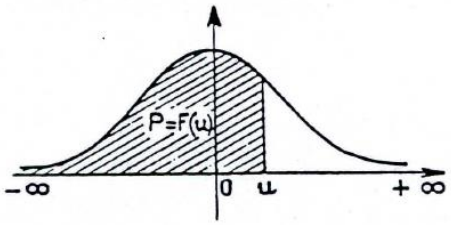
Par conséquent, le test de Shapiro-Wilk (ou tous les autres tests statistiques testant la normalité) ne doit pas être utilisé pour *vérifier* la normalité. La normalité doit être *vérifiée* visuellement (à l'aide d'un histogramme, par exemple).

[54] Il existe de nombreux autres tests statistiques testant la normalité d'une distribution. Celui-ci en est un parmi d'autres.

C. Annexe 3 – Démonstration de l'interprétation de la valeur de la SD

Soit une variable aléatoire X' qui suit une loi normale centrée réduite. A partir de la lecture de la table de cette loi, on peut lire la probabilité Pr(X' ≤ 1) et Pr(X' ≤ 1,96) :

FONCTION DE REPARTITION DE LA LOI NORMALE REDUITE
(Probabilité de trouver une valeur inférieure à u)



u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767

Pr(X' ≤ 1) = 0,84. Donc, Pr(X' ≥ 1) = 1 - 0,84 = 0,16.

On en déduit Pr(-1 ≤ X' ≤ 1) = 1 – (0,16 x 2) = 0,68.

De même, Pr(X' ≤ 1,96) = 0,975. Donc, Pr(X' ≥ 1,96) = 1 - 0,975 = 0,025.

On en déduit Pr(-1,96 ≤ X' ≤ 1,96) = 1 – (0,025 x 2) = 0,95.

Par conséquent, si X' est une variable aléatoire qui suit une loi normale centrée réduite, il y a 68% de chances pour que sa valeur soit comprise entre -1 et 1, et il y a 95% de chances pour que sa valeur soit comprise entre -1,96 et +1,96.

Soit maintenant X une variable aléatoire qui suit une loi normale de moyenne m et de Standard Deviation SD. Vous devez vous souvenir que pour passer d'une loi normale N(m,SD) à une loi normale centrée réduite N(0,1), on retire m et on divise par SD :

$$X' = \frac{X - m}{SD}$$

Par conséquent, si Pr(-1 ≤ X' ≤ 1) = 0,68, alors $Pr\left(-1 \leq \frac{X-m}{SD} \leq 1\right) = 0{,}68.$

Donc, Pr(m - SD ≤ X ≤ m + SD) = 0,68 = 68%.

De même, si Pr(-1,96 ≤ X' ≤ 1,96) = 0,95, alors $Pr\left(-1{,}96 \leq \frac{X-m}{SD} \leq 1{,}96\right) = 0{,}95.$

Donc, Pr(m - 1,96xSD ≤ X ≤ m + 1,96xSD) = 0,95 = 95%.

D. Annexe 4 – Interprétation rigoureuse des 1er et 3ème quartiles

La médiane est la valeur du caractère quantitatif telle qu'au moins 50% des individus de l'échantillon ont une valeur inférieure ou égale à la médiane, et au moins 50% des individus de l'échantillon ont une valeur supérieure ou égale à la médiane.

Le 1er quartile, ou 25ème percentile, noté « Q1 », est la valeur du caractère quantitatif telle qu'au moins 25% des individus de l'échantillon ont une valeur inférieure ou égale à Q1, et au moins 75% des individus de l'échantillon ont une valeur supérieure ou égale à Q1.

Le 3ème quartile, ou 75ème percentile, noté Q3, est la valeur du caractère quantitatif telle qu'au moins 75% des individus de l'échantillon ont une valeur inférieure ou égale à Q3, et au moins 25% des individus de l'échantillon ont une valeur supérieure ou égale à Q3.

Le tableau ci-dessous illustre ces notions qui peuvent paraître contre-intuitives.

Individu	Valeur
1	**0**
2	**1**
3	**1**
4	**1**
5	**1**
6	**1**
7	**1**
8	**1**
9	**2**
10	**2**
11	**2**
12	**2**

A partir de ce tableau, un logiciel de statistique fournira les indicateurs suivants : médiane = 1, Q1 = 1, Q3 = 2.

En effet, vous pouvez voir dans le tableau ci-dessus qu'au moins 50% des individus de l'échantillon ont une valeur inférieure ou égale à 1 (c'est vrai : 8/12=67%), et au moins 50% des individus de l'échantillon ont une valeur supérieure ou égale à 1 (c'est vrai : 11/12=92%).

De même, au moins 25% des individus de l'échantillon ont une valeur inférieure ou égale à Q1 = 1 (c'est vrai : 8/12=67%), et au moins 75% des individus de l'échantillon ont une valeur supérieure ou égale à Q1 = 1 (c'est vrai : 11/12=92%).

Enfin, au moins 75% des individus de l'échantillon ont une valeur inférieure ou égale à Q3 = 2 (c'est vrai : 12/12=100%), et au moins 25% des individus de l'échantillon ont une valeur supérieure ou égale à Q3 = 2 (c'est vrai : 4/12=33%).

E. Annexe 5 – calcul d'un intervalle de confiance à 95% d'une médiane

1. *En théorie*

Soit IC_{inf} et IC_{sup} respectivement les bornes inférieures et supérieures de l'intervalle de confiance à 95% d'une médiane MED d'une variable VAR_{quant} calculée dans un échantillon de taille N. Voici la procédure pour calculer IC_{inf} et IC_{sup}.

1ère étape : calculer les valeurs V_{inf} et V_{sup} à l'aide des formules ci-dessous (Campbell, Br Med J (Clin Res Ed), 1988).

$$V_{inf} = \frac{N}{2} - \frac{1{,}96 \times \sqrt{N}}{2}$$

$$V_{sup} = 1 + \frac{N}{2} + \frac{1{,}96 \times \sqrt{N}}{2}$$

2ème étape : arrondir V_{inf} à l'unité inférieure (soit V'_{inf} cette valeur) et V_{sup} à l'unité supérieure (soit V'_{sup} cette valeur).

3ème étape : trier le fichier de données par valeurs de VAR_{quant} croissantes.

4ème étape : identifier dans le fichier de données trié la V'_{inf}ème valeur de VAR_{quant} en partant de la plus petite ($\rightarrow IC_{inf}$), et la V'_{sup}ème valeur de VAR_{quant} en partant de la plus petite ($\rightarrow IC_{sup}$).

5ème étape : l'intervalle de confiance à 95% de MED est formé par ces deux valeurs IC_{inf} et IC_{sup} identifiées à la 4ème étape.

2. *En pratique avec Excel*

Soit un échantillon de 28 chats dont on souhaite calculer la médiane de l'âge ainsi que son intervalle de confiance à 95%.

	A	B
1	N° du chat	AGE
2	1	3
3	2	5
4	3	10
5	4	3
6	5	3
7	6	12
8	7	4
9	8	14
10	9	2
11	10	12
12	11	12
13	12	11
14	13	11
15	14	3
16	15	9
17	16	6
18	17	9
19	18	16
20	19	8
21	20	11
22	21	9
23	22	4
24	23	12
25	24	13
26	25	15
27	26	11
28	27	5
29	28	13
30		9.5

La médiane vaut 9,5 ans dans l'échantillon (formule tapée dans la cellule B30).

La 1ère étape consiste à calculer V_{inf} et V_{sup} :

$$V_{inf} = \frac{28}{2} - \frac{1{,}96 \times \sqrt{28}}{2} = 8{,}8$$

$$V_{sup} = 1 + \frac{28}{2} + \frac{1{,}96 \times \sqrt{28}}{2} = 20{,}2$$

La 2ème étape consiste à arrondir à l'unité inférieure V_{inf} et à l'unité supérieure V_{sup}. On obtient ainsi $V'_{inf} = 8$ et $V'_{sup} = 21$.

La 3ème étape consiste à trier le fichier de données selon l'âge croissant (une colonne « rang des valeurs » a été créée pour faciliter la 4ème étape) :

	A	B	C
1	N° du chat	AGE	Rang des valeurs
2	9	2	1
3	1	3	2
4	4	3	3
5	5	3	4
6	14	3	5
7	7	4	6
8	22	4	7
9	2	5	8
10	27	5	9
11	16	6	10
12	19	8	11
13	15	9	12
14	17	9	13
15	21	9	14
16	3	10	15
17	12	11	16
18	13	11	17
19	20	11	18
20	26	11	19
21	6	12	20
22	10	12	21
23	11	12	22
24	23	12	23
25	24	13	24
26	28	13	25
27	8	14	26
28	25	15	27
29	18	16	28

La 4ème étape consiste à identifier les 8ème et 21ème valeurs de AGE dans le fichier de données en partant de la plus petite valeur (2 ans ici), grâce à la colonne « rang des valeurs ». On peut lire ces valeurs là dans la figure ci-dessus : 5 ans et 12 ans.

5ème étape : l'intervalle de confiance à 95% de la médiane de 9,5 ans est donc : [5 ans ; 12 ans].

F. Annexe 6 – Autres tests statistiques pour séries appariées

1. *Le test de Wilcoxon pour séries appariées (comparaison de médianes)*

Ce test doit être utilisé lorsque le test de Student pour séries appariées ne peut pas s'utiliser si la distribution du caractère quantitatif ne peut pas être considérée comme normale. Ce test statistique permet de savoir si la médiane du caractère quantitatif estimée à t_0 est significativement différente de celle estimée à t_1.

2. *Le test de McNemar pour séries appariées (comparaison de deux pourcentages)*

Ce test doit être utilisé lorsque le caractère évalué à t_0 et à t_1 est binaire. L'exemple présenté dans le tableau ci-dessous est celui où le caractère évalué est la présence de symptômes. Les données doivent être présentées telles qu'indiqué dans le **Tableau 5**.

Tableau 5. Données théoriques pour illustrer le calcul du test statistique de McNemar.

		Symptômes à t_0		
		Oui	Non	Total
Symptômes à t_1	Oui	a	b	a+b
	Non	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Par exemple, c = « le nombre d'animaux qui présentaient des symptômes à t_0 *et* qui n'en présentaient pas à t_1 ». Si N animaux sont évalués deux fois (à t_0 et à t_1), alors a+b+c+d = N. Pour savoir si le pourcentage d'animaux symptomatiques à t_0 $\left(\frac{a+c}{a+b+c+d}\right)$ est significativement différent du pourcentage d'animaux symptomatiques à t_1 $\left(\frac{a+b}{a+b+c+d}\right)$, il faut utiliser le test de McNemar. Sans entrer dans les détails, ce test teste l'hypothèse nulle selon laquelle il n'y a pas de différence du nombre de paires discordantes entre t_0 (c) et t_1 (b)[55]. (D'ailleurs, vous voyez bien que si b=c, alors les deux pourcentages que j'ai cités plus haut seront égaux.) La **Figure 17** ci-dessous vous présente un exemple avec des données chiffrées telles que vous pourriez les recueillir pour une étude, et les placer dans le tableau de la bonne façon pour effectuer le test statistique[56].

[55] Cf. https://en.wikipedia.org/wiki/McNemar's_test

[56] Veuillez noter cependant que pour faire le test de McNemar sur de si faibles effectifs, une correction dite de « continuité » est nécessaire.

Num_chien	Symptomes à t0	Symptomes à t1
1	1	1
2	1	0
3	0	0
4	0	0
5	1	0
6	0	1
7	1	1
8	1	0
9	1	0

		Symptômes à t_0		
		Oui	Non	Total
Symptômes à t_1	Oui	2	1	3
	Non	4	2	6
	Total	6	3	9

Figure 17. Illustration pour le test de McNemar.

Dans l'exemple ci-dessus, on veut savoir si le pourcentage d'animaux symptomatiques à t_0 (6/9=66%) est significativement différent du pourcentage d'animaux symptomatiques à t_1 (3/9=33%)[57].

[57] Après avoir effectué le test statistique sur le site de BiotstatGV (http://marne.u707.jussieu.fr/biostatgv/?module=tests), p = 0,37 > 0,05 donc les deux pourcentages cités ne sont pas significativement différents.