TEST DU χ^2 DE PEARSON

I INTRODUCTION

Le test du χ^2 proposé par Pearson[1] (appelé aussi test de khi-carré ou de khi-deux, test d'indépendance ou test de χ^2 d'homogénéité) est un outil d'aide à la décision, notamment lorsque l'on s'interroge, en épidémiologie, sur l'existence ou non d'une différence dans les valeurs des indicateurs de fréquence d'une maladie, obtenues sur des échantillons.

La problématique est alors du type : ***la fréquence de la maladie étudiée dans deux populations (A et B) est-elle identique ou différente ?***

En général, on ne connait pas la réponse quant à la situation réelle pour les populations entières. On ne dispose que des résultats obtenus sur des échantillons.

La **situation réelle** peut être :

Soit 1 : il n'y a aucune différence entre les deux populations comparées ;

Soit 2 : il existe une différence, plus ou moins importante, entre les deux populations comparées.

- Dans le premier cas (***absence de différence réelle***), l'erreur de conclusion possible[2], à partir des résultats obtenus sur les échantillons étudiés, est de conclure, à tort, à l'existence d'une différence, due simplement aux fluctuations d'échantillonnage dépendant du hasard[3]. C'est l'***erreur alpha*** ou ***erreur par excès***.

  Afin de limiter les erreurs de ce type, il est nécessaire de déterminer *a priori* la probabilité de risque d'erreur alpha que l'on accepte, c'est-à-dire ***le risque accepté que la différence apparente observée sur les échantillons ne soit due qu'au hasard.***

  C'est ce risque d'erreur alpha qui est pris en compte pour l'interprétation des résultats d'un χ^2.

- Dans le second cas (***existence d'une différence réelle***), l'erreur de conclusion possible, à partir des résultats obtenus sur les échantillons étudiés, est de conclure, à tort, à l'absence de différence. C'est l'***erreur bêta*** ou ***erreur par défaut*** : une différence existe, mais elle n'a pas pu être détectée. Le risque de cette erreur est d'autant plus élevé :

  - Que le nombre d'unités dans les échantillons étudiés est faible ;
  - Que la différence réelle entre les populations est elle-même faible.

  On peut s'interroger sur la signification à donner à un niveau de différence existant entre deux populations : par exemple, pour deux populations (de 1 000 sujets chacune), étudiées en totalité avec un test de dépistage parfaitement sensible et spécifique, l'obtention de 200 réponses positives dans l'une, et de 201 dans l'autre, correspond certes à une ***différence***, mais ***minime***, sans ***aucune importance pratique***.

  À partir de quelle valeur d'écart entre les deux nombres de réponses positives (10, 20, 30 ou davantage), obtenues sur les populations entières, peut-on considérer que la différence est suffisamment grande pour être « ***utile*** » ***à connaitre*** et, donc, à détecter ?

  La réponse à cette question ***conditionne le nombre de sujets à introduire dans les échantillons,*** lorsque l'on procède par sondage : plus la valeur de l'écart que l'on veut être capable de détecter est faible, plus ***ce nombre*** doit être élevé. A la limite, en augmentant considérablement (jusqu'aux populations entières) la taille des échantillons, on peut, très souvent, détecter une différence, mais qui n'a aucune « ***signification sur le plan biologique*** », aucune utilité pratique (Encadré χ^2 1). C'est pourquoi effectuer un test de χ^2 sur des populations entières n'a aucun sens statistique, puisqu'en l'absence d'échantillonnage, on ne peut pas se poser la question de l'effet des fluctuations d'échantillonnage sur le résultat (ce à quoi vise ce test).

[1] Il existe d'autres types de test du χ^2: voir en fin de cette annexe.
[2] *cf.* « *Les erreurs* » dans l'*Epidémiologie pour tous* sur le site de l'AEEMA
[3] *cf.* l'annexe « *Les fluctuations d'échantillonnage et l'estimation d'un pourcentage par sondage* »

Encadré χ^2 1

Différence significative

L'expression *« Différence significative »* peut être rencontrée et utilisée avec deux significations différentes :

- ***« Différence significative au plan statistique »***

Cette expression est une conclusion possible à l'issue d'un test statistique utilisé pour comparer deux échantillons (ou plus) et consistant à estimer la probabilité que le hasard, source de fluctuations d'échantillonnage, puisse suffire à lui seul pour entrainer un écart aussi important que celui constaté à partir d'échantillons en provenance d'une même population.

Si cette probabilité estimée est plus faible que la probabilité d'erreur par excès choisie, on peut conclure à l'existence d'une ***« différence significative au plan statistique »***.

- ***« Différence significative au plan biologique »***

Cette expression est utilisée pour qualifier une différence entre deux situations comparées, suffisamment grande pour que l'on considère qu'elle soit importante à connaitre et à prendre en considération en pratique.

Plus une différence entre deux situations est grande, plus il est facile de la détecter au plan statistique à partir d'échantillons de petite taille.

Mais la détection d'une différence significative au plan statistique entre deux situations n'implique pas automatiquement qu'elle corresponde à une différence significative au plan biologique.

Et, à l'inverse, il peut exister une différence au plan biologique entre deux situations, en l'absence de différence significative au plan statistique, en raison de la taille insuffisante des échantillons utilisés (on parle alors de *« manque de puissance statistique »*).

II - COMPARAISON DE DEUX POURCENTAGES OBSERVÉS

La comparaison de deux pourcentages observés, en vue de déterminer de façon objective s'ils sont ou non « différents », nécessite de recourir à une méthode indirecte, un test statistique, dit aussi test d'hypothèse. Il n'est pas possible d'effectuer cette comparaison directement, par quelle méthode de calcul que ce soit, sans ce recours au test d'hypothèse : à moins d'avoir exactement les mêmes résultats, on est obligé de constater qu'ils ne sont pas identiques et diffèrent ne serait-ce que de quelques unités. Le test d'hypothèse permet de fixer un seuil à partir duquel on conviendra de conclure que les pourcentages sont différents, cela d'une façon objective et admise par tous.

En fait, on se demande si la différence est porteuse de sens, au plan épidémiologique.

Dans le cas de données exhaustives, les écarts éventuels de pourcentage entre deux zones peuvent être dus soit à une évolution différente de la maladie dans chacune des zones considérées, ce qui conduit à un pourcentage « notablement » plus élevé ici que là, soit au fait que la maladie n'évolue pas de façon différente mais les écarts résultent de différences, par exemple, dans la structure démographique de la population : la prise en compte de ces différences peut conduire à réduire l'importance des écarts observés et à tempérer la première impression (*cf.* Annexe *« Standardisation des taux démographiques»*).

Dans le cas de données obtenues par sondage, tout écart éventuel peut être rapporté soit aux fluctuations d'échantillonnage, soit au fait que la variable de santé dans les deux populations étudiées s'exprime par des pourcentages réellement différents. C'est ce double questionnement qui fournit le moyen indirect, fondamental en statistique, de comparer des pourcentages obtenus par sondage, selon une démarche appelée « test d'hypothèse ».

☐ *Principe du test d'hypothèse*

Le principe de la démarche consiste à formuler deux hypothèses :

- La première : « *on ne peut pas dire que les pourcentages sont réellement différents, car l'écart observé peut être simplement dû aux fluctuations d'échantillonnage* »,
- La deuxième : *« l'écart observé n'est pas compatible avec les fluctuations d'échantillonnage, et on peut en déduire que les pourcentages réels sont probablement différents ».*

La première hypothèse, en raison de sa formulation particulière, niant l'existence d'une différence réelle, appelée *« hypothèse nulle »* est notée H_0, l'autre, *« hypothèse alternative »*, H_1.

Les données d'observation, organisées en un tableau, sont utilisées pour aider à choisir entre ces deux hypothèses, en calculant, pour chacune des valeurs du tableau, les écarts entre les effectifs observés et les effectifs théoriques qui pourraient être observés en prenant pour base de calcul l'hypothèse nulle. Les écarts calculés pour chacune des cases du tableau sont agrégés, en en faisant tout simplement la somme, en un écart global pour l'ensemble des effectifs du tableau. Par la suite, nous emploierons, de façon raccourcie, le terme d'*effectifs calculés* à la place de l'expression plus complète d'*effectifs théoriques calculés sous l'hypothèse nulle.*

On juge l'importance de cet écart global à l'aide ***d'une table appropriée qui comporte les écarts les plus grands qui peuvent être observés par le simple fait du hasard***. Cette table (*cf.* Table du χ^2) présente la correspondance entre la valeur d'un écart et la probabilité de dépasser cette valeur ; généralement, on retient la probabilité dite « critique » de 0,05 car on considère que prendre une décision valide dans 95 p. cent des cas est supportable ; mais la probabilité peut être plus faible, pour augmenter la marge de validité.

L'interprétation dépend de l'importance des écarts observés :

- Si l'écart est jugé « important » par rapport à la valeur de la table pour un risque d'erreur consenti fixé à l'avance, on en déduit qu'il n'est pas compatible avec l'hypothèse nulle ; on décide de la rejeter et d'admettre que l'hypothèse alternative doit être préférablement retenue ;
- Dans le cas contraire, on dit qu'on n'a pas vu d'écart « significatif », on ne rejette pas l'hypothèse nulle… et on ne peut pas en dire davantage.

Il faut comprendre que le résultat du test ne s'impose pas à notre décision en nous révélant la « *réalité vraie* » : il s'agit seulement d'une information qui nous permet de faire le pari sur l'hypothèse la plus probable.

Table du χ^2

La table du χ^2 (extrait) donne la probabilité *p* du risque d'erreur α pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.) (*cf.* plus loin)

d.d.l. \ $p\alpha$	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588

☐ *Etude d'un exemple*

Prenons l'exemple, très courant en épidémiologie, de la comparaison de deux pourcentages obtenus par sondage dans deux régions A et B.

Région A : 200 exploitations dont 40 infectées, soit un pourcentage de prévalence de 20 %,
Région B : 300 exploitations dont 30 infectées, soit un pourcentage de prévalence de 10 %.

♦ *Raisonnement : l'hypothèse nulle et l'hypothèse alternative*

L'hypothèse nulle consiste à dire que les pourcentages d'exploitations infectées ne sont pas différents dans la région A et la région B, parce que la différence apparente des pourcentages obtenus (20 % et 10 %) n'est due qu'à

des fluctuations d'échantillonnage à partir d'une seule et même population. L'hypothèse alternative postule que cette différence apparente n'est pas compatible avec les fluctuations d'échantillonnage, et par conséquent on peut admettre, avec un risque d'erreur consenti au préalable, que les pourcentages correspondent à des populations différentes quant à la fréquence de la maladie.

Le raisonnement s'appuie sur la comparaison des effectifs observés aux effectifs théoriques calculés sous cette hypothèse nulle, c'est-à-dire selon laquelle les deux échantillons proviennent d'une seule population.

♦ ***Calcul des effectifs théoriques***

Mais, comment calculer ces effectifs théoriques, puisqu'on ne connaît pas le pourcentage d'élevages infectés dans la population correspondant à l'hypothèse nulle, et que justement on veut savoir si cette hypothèse nulle est vraie ou non ! Voici l'aspect le plus déroutant, mais qui justement va fournir la solution pour la résolution de ce problème.

Comme on ne connaît pas la composition de la population correspondant à l'hypothèse nulle, on doit en faire une estimation : la moins mauvaise approximation de cette population inconnue est obtenue en agrégeant les effectifs des deux échantillons. En effet, s'ils viennent bien d'une seule population, cette approximation sera de bonne qualité et permettra de faire les calculs.

Dans le cas contraire, si l'hypothèse nulle est fausse, et si, malgré cela, l'écart entre données observées et effectifs calculés permet de rejeter l'hypothèse nulle, alors c'est que cette différence est véritablement « notable », digne d'être prise en considération. C'est pourquoi ce test est considéré comme « ***robuste*** ».

D'un autre côté, il peut exister une différence réelle, au plan épidémiologique s'entend, la maladie ne sévissant pas de la même façon dans les deux régions, mais les conditions de la comparaison ne permettent pas de mettre cette différence en évidence ; c'est pourquoi on dit aussi que ce test est « ***conservateur*** », car il ne permet pas de révéler des différences si elles ne sont pas suffisamment importantes.

Le tableau χ^2.1 fournit les effectifs observés et les effectifs calculés (sous l'hypothèse nulle).

Tableau χ^2. 1

Effectifs observés dans deux régions et effectifs calculés correspondants (entre parenthèses) calculés sous l'hypothèse nulle

	Exploitations infectées		**Exploitations indemnes**		**Total**
Région A	40	(28)	160	(172)	200
Région B	30	(42)	270	(258)	300
Total	70		430		500

Le cumul des effectifs de chacun des échantillons fournit le nombre de 70 exploitations infectées sur un total de 500 exploitations, soit un pourcentage d'infection de 70/500 * 100 = 14 %.

Dans la région A, le nombre « calculé » d'exploitations infectées attendu sous l'hypothèse nulle est de :

200 × 14 p. cent = 28 (*placé entre parenthèses dans la case supérieure gauche*).

Dans la région B, par analogie, le nombre « calculé » d'exploitations infectées est de :

300 × 14 p. cent = 42 (*placé entre parenthèses dans la case inférieure gauche*).

Les effectifs calculés d'exploitations indemnes s'obtiennent par soustraction et sont placés entre parenthèses dans les cases de droite : on aurait pu procéder de cette façon pour le nombre correspond à la région B, puisque 70-28 = 42, ce qui illustre le fait qu'il est possible de reconstituer l'ensemble d'un tableau comportant quatre cases à partir d'une première donnée et des totaux de marges.

D'une façon générale, l'effectif calculé pour une case donnée peut être obtenu en multipliant le total de la ligne correspondante par le total de la colonne correspondante et en divisant par le total général du tableau.

♦ ***Calcul des écarts entre effectifs observés et effectifs calculés avec le χ^2***

On considère ensuite les écarts entre effectifs observés et effectifs calculés. S'agissant d'un tableau, c'est l'ensemble des écarts que l'on veut prendre en compte, d'un coup : il faut pour cela les « *réduire* » en une seule donnée.

On ne peut pas simplement faire la somme des écarts, car les valeurs algébriques s'annuleraient. Plutôt que d'utiliser les valeurs absolues, il est préférable d'utiliser le carré des écarts, car non seulement ils seront toujours positifs, mais en plus, ils amplifient l'importance des écarts les plus grands. L'inconvénient est alors de ne plus considérer des effectifs, mais des carrés d'effectifs. Le plus simple est de ramener à la dimension initiale d'un effectif en divisant par un autre effectif : l'effectif observé aurait l'inconvénient d'être dépendant des observations, tandis que la valeur calculée correspond à l'hypothèse nulle, beaucoup plus stable, et ce choix revient à privilégier l'hypothèse nulle.

Par conséquent, à partir des effectifs observés, on obtient par calcul une valeur qui reflète, après un calcul particulier, *globalement* les écarts de l'ensemble des effectifs observés avec les effectifs théoriques sous l'hypothèse nulle. Mais, si cette valeur a bien conservé la dimension d'un effectif, on ne peut plus réellement la considérer comme un effectif, en raison des transformations subies par le mode de calcul : on doit considérer le χ^2 comme un *indicateur*, ou un *indice*, qui nous aidera dans notre démarche.

Nous venons de définir le χ^2, qui correspond à la somme des quotients : carrés des différences entre effectifs observés (O) et effectifs calculés (C) divisés par les effectifs calculés, soit :

$$\chi^2 = \sum \frac{(O-C)^2}{C}$$

Dans le cas présent, le calcul du χ^2 est donc le suivant :

$$\chi^2 = \frac{(40-28)^2}{28} + \frac{(160-172)^2}{172} + \frac{(30-42)^2}{42} + \frac{(270-258)^2}{258} = 9{,}97$$

Ce chiffre est à comparer avec ceux de la table de χ^2. Cette table comporte en colonnes les probabilités pour que le χ^2 atteigne ou dépasse une valeur donnée du simple fait du hasard, et en lignes le « nombre de d.d.l. », qui est en fait un reflet de la taille du tableau (nombre de cases) : par conséquent, pour pouvoir utiliser la table, il faut déterminer à quel degré de liberté (d.d.l.) correspond le tableau étudié pour pouvoir choisir la valeur seuil appropriée.

Remarque

Avec les tableurs, il est facile d'effectuer ces calculs. On peut aussi utiliser la feuille de calcul fournie en annexe (Annexe *« Tableur pour le test du χ^2»*), un logiciel, comme Epi Info, ou encore une des nombreuses aides en ligne sur internet.

♦ *Degrés de liberté*

À ce stade, il est nécessaire de préciser ce concept de « nombre de degrés de liberté » (d.d.l.) (Encadré χ^2.2).

Encadré χ^2. 2

Les degrés de liberté

On appelle « *degré de liberté* » le nombre de cases du tableau des effectifs observés (case(s) « total » mise(s) à part) que l'on peut **remplir librement** pour pouvoir reconstituer l'ensemble du tableau à l'aide des marges : il constitue un reflet de la dimension du tableau.

❒ Lorsque l'on examine un échantillon selon **un seul caractère**, le degré de liberté est égal au **nombre d'éventualités** de ce caractère, **moins une**.

1ER EXEMPLE : un caractère (le sexe), avec deux éventualités (mâle ou femelle)

	Mâles	**Femelles**	**Total**
Effectifs			100

Ainsi, dans ce premier exemple, d'un caractère (le sexe), avec deux éventualités (mâle ou femelle), sur un total (quel qu'il soit) de 100 animaux, le degré de liberté est **UN**, c'est-à-dire que l'on peut **fixer librement** l'effectif **d'une seule case**, car lorsque l'on a fixé le nombre de mâles ou de femelles, l'effectif de l'autre case est automatiquement déterminé.

Dans ce cas, le degré de liberté (d.d.l.) est :

d.d.l. = nombre d'éventualités moins une
d.d.l. = 2 - 1 = 1

2EME EXEMPLE : un caractère (le type d'élevage), avec trois éventualités (élevage laitier, allaitant ou mixte)

3	**Elevages laitiers**	**Elevages allaitants**	**Elevages mixtes**	**Total**
Effectifs				100

Dans ce deuxième exemple, il est possible de **remplir librement deux cases** du tableau, mais pas une troisième.

Le degré de liberté est donc de deux, ce qui correspond bien à la règle énoncée ci-dessus :

d.d.l. = nombre d'éventualités moins une
d.d.l. = 3 - 1 = 2

❒ Lorsque l'on examine un échantillon selon **deux caractères**, le nombre de degrés de liberté est égal :

- Au **nombre d'éventualités** du premier caractère, **moins une**,
- **Multiplié** par le **nombre d'éventualités** du deuxième caractère, **moins une**.

3EME EXEMPLE : un caractère (le type d'élevage) avec trois éventualités (*cf.* ci-dessus) et un deuxième caractère (le département) avec deux éventualités (département A et département B).

Eventualités élevage / **Eventualités département**	**Elevages laitiers**	**Elevages allaitants**	**Elevages mixtes**	**Total**
Département A				50
Département B				150
Total	30	70	100	200

Dans un tel tableau, on peut **remplir librement deux cases**, **n'importe lesquelles**. Mais dès que les effectifs de deux cases sont fixés, il n'existe plus aucune liberté car les effectifs de toutes les autres cases sont automatiquement déterminés.

Le nombre de degrés de liberté est donc de deux, ce qui correspond bien à la règle citée ci-dessus :

d.d.l. = (nombre d'éventualités, moins une) (nombre d'éventualités, moins une)
d.d.l. = (3 − 1) (2 − 1) = 2

Le nombre de d.d.l. pour le tableau χ^2.1 est le produit :

(nombre de colonnes − 1) (nombre de lignes − 1)

Soit, ici : (2 − 1) (2 − 1) = 1

♦ ***Choix d'une probabilité critique***

Le choix du d.d.l. simplifie le tableau, en conduisant à ne retenir qu'une seule ligne.

Plutôt que de chercher dans quelle colonne peut se situer la valeur calculée, on doit avoir au préalable déterminé le degré de risque d'erreur consenti, *« **seuil de signification** »* au-delà duquel on considérera que le hasard a une probabilité faible de donner un écart de cette grandeur, suffisamment faible pour qu'on accepte de courir le risque de se tromper en choisissant de rejeter H_0.

Ce type d'erreur consentie est celui de conclure à tort à l'existence d'un écart significatif alors qu'en réalité c'est le hasard qui en est l'origine. Cette erreur est donc *par excès* ; elle aussi dénommée erreur alpha (α) (voir Annexe « *Les erreurs de jugement* »).

Finalement, cette ligne se résume alors à une seule case. Le plus souvent, on choisit la probabilité de 0,05, de façon totalement arbitraire et couramment admise dans la communauté scientifique.

Mais, on peut choisir, toujours au préalable, d'autres valeurs de probabilité :

- une probabilité plus faible (0,01, voire 0,001) pour accroître la robustesse des résultats, leur répétabilité ;
- ou, au contraire, plus élevée (de l'ordre de 0,10 à 0,20) :
    - soit dans les investigations sur le terrain lorsqu'on est à la recherche de pistes permettant d'orienter les recherches, quelles nouvelles données collecter,
    - soit dans des démarches d'exploration de données déjà collectées, pour éliminer les variables sans véritable intérêt, tout en en ayant conservé suffisamment, grâce à ce seuil plus complaisant, pour que l'exploration conserve sa puissance, avant le recours à des méthodes plus sophistiquées.

♦ ***Utilisation de la table de χ^2***

Dans la table de χ^2, le seuil de signification à 0,05 pour un d.d.l. est $\chi^2 = 3{,}84$.

La valeur trouvée, 9,97, dépasse cette limite et permet donc de considérer que **l'écart observé** entre les pourcentages d'exploitations infectées des régions A et B est **significatif** et probablement non dû simplement aux fluctuations d'échantillonnage. Cette formulation est le résultat qualitatif du test.

Cette valeur (9,97) est supérieure à celle correspondant à un seuil de risque d'erreur α de 1 p. cent (*cf.* table de χ^2, ligne 1 : 6,635) mais inférieure à celle correspondant à un seuil de risque d'erreur α de 1 p. mille (10,827). On peut ainsi considérer que la différence entre les deux régions A et B est donc significative, mais en précisant « avec un ***degré de signification*** compris entre 1 p. cent et 1 p. mille », c'est-à-dire que la probabilité qu'un tel écart résulte du seul fait du hasard est comprise entre ces valeurs ou encore, c'est le plus petit risque d'erreur pour lequel la différence est encore significative ; les logiciels statistiques peuvent même donner la valeur précise de cette probabilité, ici 0,002 (Annexe « *Tableur pour le test du χ^2*»).

Remarque

La table comporte en colonnes des probabilités et non des pourcentages. Dans la présentation des résultats, selon les auteurs, les indications peuvent être fournies soit sous forme de probabilités, soit sous forme de pourcentages.

♦ ***Conditions d'application du χ^2***

Dans tous les cas, le test du χ^2 ***ne peut être utilisé que si deux conditions sont remplies***. Elles sont rappelées dans l'encadré χ^2.3, commentées et illustrées ci-dessous.

Encadré χ^2. 3

Les conditions d'application du χ^2

1. ***Le test du χ^2 ne peut être utilisé que si les effectifs calculés de toutes les cases sont d'au moins 5.***
2. ***Les « individus » dénombrés doivent être indépendants sur le plan statistique.***

On peut noter qu'aucune condition sur la distribution des données n'a été formulée : pour d'autres tests, il faut s'assurer que la distribution des données est *« normale »* (à défaut, il faut transformer les données de façon appropriée pour rendre la distribution normale). Ce n'est pas le cas avec le χ^2 : c'est pourquoi on dit que ce test est *« non paramétrique »*, car il ne dépend pas des paramètres d'une loi normale.

Première condition

Ici la condition était respectée puisque les effectifs calculés les plus petits étaient de 28 et de 42.

Il arrive qu'une ou plusieurs cases d'un tableau de contingence corresponde(nt) à des effectifs calculés inférieurs à 5.

- Le test de χ^2 demeure applicable lorsque plus de 80 p. cent des cases ont des effectifs calculés supérieurs ou égaux à 5. Ainsi, pour un tableau de contingence avec **plusieurs degrés de liberté** (par exemple, 3x2, soit deux degrés de liberté), comportant seulement une case dont l'effectif calculé est inférieur à 5 (et donc 5 cases sur 6 avec des effectifs calculés supérieurs à 5, soit 5/6 = 83 p. cent), il est possible d'utiliser le test de χ^2. Si plus de

20 p. cent des cases comprennent des effectifs calculés inférieurs à 5, il faut utiliser le **test exact de Fisher**[4], qui nécessite l'utilisation d'un logiciel de statistique pour son calcul.

- Pour un tableau de contingence à **un seul degré de liberté** (tableau 2x2), il est possible d'utiliser la **correction de Yates** lorsqu'au moins un des effectifs calculés est ***inférieur à 5*** (mais ***supérieur ou égal à 3***). Elle consiste à retrancher 0,5 à chaque différence absolue « effectif observé-effectif calculé », avant de l'élever au carré :

$$\chi^2 = \sum \frac{(|O - C| - 0{,}5)^2}{C}$$

Lorsque l'effectif calculé est plus faible (< 3), il faut utiliser le **test exact de Fisher**.

Remarque : Les conditions d'indication de la correction de Yates sont très discutées par les statisticiens du point de vue de l'ajustement de l'approximation que constitue l'indicateur calculé avec la loi du χ^2 : cette approximation est surtout valide lorsqu'elle porte sur des nombres « suffisamment grands » (l'appréciation de la grandeur étant loin de faire l'unanimité). On remarquera toutefois que la correction diminue le numérateur et donc la valeur du χ^2 ; de ce fait, ce test est « conservateur » par rapport au χ^2 non corrigé car il tempère le risque de conclure à tort en considérant l'écart comme notable. D'un autre côté, il ne faut pas oublier que l'interprétation d'un tableau portant sur un nombre d'observations limité, cela quel que soit le mode de calcul, devra demeurer très prudente, du fait de la grande sensibilité du calcul à de faibles écarts d'effectifs ; cette remarque est d'autant plus valide pour des résultats proches du seuil de signification.

Deuxième condition

L'indépendance des « individus » sur le plan statistique signifie qu'il ne doit pas y avoir de lien entre eux, notamment au travers d'une appartenance à des groupes d'individus (par exemple, en médecine vétérinaire, tous les animaux d'un même élevage ont en commun le fait d'être soumis aux mêmes conditions d'élevage, au même environnement ; ils ne sont pas indépendants entre eux, même s'ils le sont par rapport aux animaux d'autres élevages).

Exemple :

Soit une étude faite sur l'administration d'antibiotiques lors du tarissement des vaches, avec 20 élevages dans lesquels on applique le tarissement systématique et 20 élevages dans lesquels on ne l'applique pas.

L'utilisation du test du χ^2 sur un tableau de contingence fondé sur l'unité « élevage » (avec, par exemple, un taux de mammites supérieur ou inférieur à 5 p. cent) est possible, car les élevages sont indépendants les uns des autres.

En revanche, elle ne le serait pas pour un tableau de contingence fondé sur l'unité « animal » car les vaches de chaque élevage ne sont pas indépendantes les unes des autres. L'inconvénient de négliger cette règle serait d'aboutir de façon non légitime (et donc pouvant induire en erreur) à un résultat montrant une différence significative du fait des plus grands effectifs mis en jeu avec l'unité « animal » qu'avec l'unité « élevage ».

Remarque

• Le test du χ^2 fournit les mêmes résultats que le test de l'écart-réduit. Ces deux tests sont identiques, le χ^2 étant le carré de ε pour un degré de liberté. La différence réside dans le fait que l'on peut utiliser le test du χ^2 lorsqu'il existe plusieurs degrés de liberté (**comparaison de plusieurs pourcentages**, *cf.* plus loin), alors qu'on ne le peut pas avec le test de l'écart-réduit.

[4] Pour effectuer un test exact de Fisher, utiliser un logiciel approprié.

III - COMPARAISON DE PLUSIEURS POURCENTAGES OBSERVÉS

Dans une zone, on a effectué une enquête descriptive sur une maladie et on a estimé par sondage aléatoire le pourcentage d'exploitations infectées dans trois catégories d'exploitations définies selon leur taille (strates) :

- Nombre d'animaux de 1 à 20 ;
- Nombre d'animaux de 21 à 50 ;
- Nombre d'animaux supérieur à 50.

Les résultats obtenus figurent dans le tableau χ^2.2.

Tableau χ^2. 2

Effectifs observés en fonction de la taille des exploitations et effectifs calculés

	Exploitations				
	infectées		**indemnes**		**Total**
Exploitations de 1 à 20 animaux	10	(20)	90	(80)	100
Exploitations de 21 à 50 animaux	12	(12)	48	(48)	60
Exploitations de plus de 50 animaux	18	(8)	22	(32)	40
Total	40		160		200

Comme dans le cas précédent, le calcul des effectifs théoriques fait appel à l'hypothèse nulle et conduit aux chiffres placés entre parenthèses dans le tableau χ^2. 2. Dans ce cas, on constate qu'il est nécessaire de calculer au moins deux effectifs avant d'en déduire les autres par soustraction car le nombre de degrés de liberté est : $(3 - 1)(2 - 1) = 2$.

Le calcul du χ^2 se présente comme suit :

$$\chi^2 = \frac{(10 - 20)^2}{20} + \frac{(90 - 80)^2}{80} + \frac{(12 - 12)^2}{12} + \frac{(48 - 48)^2}{48} + \frac{(18 - 8)^2}{8} + \frac{(22 - 32)^2}{32} = 21{,}87$$

La consultation de la table du χ^2 pour 2 d.d.l. ($2^{ème}$ ligne) montre que la valeur obtenue au seuil de 5 p. cent est de 5,99. Les pourcentages d'exploitations infectées dans les trois catégories d'exploitations classées en fonction de la taille diffèrent donc de façon significative puisque la valeur observée (21,87) est supérieure à la valeur seuil de 5,99.

Avant de procéder à l'interprétation de ce résultat, il faut s'assurer du respect des conditions de validité (effectifs calculés supérieurs ou égaux à 5 et indépendance des données) : c'est le cas.

On peut donc dire que les pourcentages d'infection des exploitations sont différents selon la taille des exploitations, au risque d'erreur α de 0,05.

Le degré de signification est inférieur à 1 p. mille (*cf.* table de χ^2) : χ^2 pour un risque d'erreur α de 0,001 pour 2 d.d.l. = 13,815.

Afin de préciser l'interprétation fournie, l'investigateur peut être amené à vouloir comparer les trois groupes deux à deux. Il convient alors de prendre en compte ***l'inflation du risque d'erreur de première espèce*** (erreur α) qui intervient dans le cas de ***comparaisons multiples***. En effet, le risque de conclure à tort quant à l'existence d'une différence significative augmente avec le nombre de tests statistiques réalisés. Si l'on réalise *k* tests statistiques indépendants, la probabilité de rejeter au moins une des *k* hypothèses nulles (donc de conclure à une différence significative) à tort sera de $1-(1-\alpha)^k$. Afin de prendre en compte ce phénomène d'inflation du risque d'erreur de première espèce lors de comparaisons multiples, il possible d'avoir recours à des méthodes d'ajustement telles que la ***correction de Bonferroni*** : elle consiste à diviser la valeur du risque d'erreur α initialement choisie (0,05 en général) par le nombre de comparaisons effectuées. Cette méthode a l'avantage d'être très conservatoire et simple à mettre en œuvre mais se révèle très peu puissante lorsque le nombre de tests réalisés est important.

Dans notre exemple, la variable correspondant à la taille d'exploitation comprend trois catégories (exploitations de 1 à 20 animaux, exploitations de 21 à 50 animaux et exploitations de plus de 50 animaux). Si l'on souhaite

effectuer des comparaisons deux à deux, qui seraient alors au nombre de trois, en appliquant la correction de Bonferroni, on considérerait une différence comme significative pour un degré de signification inférieur à α/3 soit 0,05/3 = 0,017 ou 1,7 p. cent.

IV – DISCUSSION

Nous avons vu que le test du χ^2 permet de déterminer si un écart observé entre deux pourcentages, ou plus, mérite d'être pris en considération. Il utilise les effectifs observés. C'est un « test d'hypothèses », c'est-à-dire que son résultat permet de choisir entre l'hypothèse nulle (l'écart est lié aux fluctuations d'échantillonnage) et l'hypothèse alternative (l'écart découle du fait que les échantillons proviennent de deux populations différentes). Ce choix est un *pari*, consistant à choisir l'hypothèse paraissant la plus probable.

Pour une utilisation optimale, il faut toutefois prendre en considération différents éléments.

1. Au plan de la méthode

Jamais de χ^2 sur des données exhaustives

Il ne faut jamais oublier qu'un test statistique comme le χ^2 ne peut être réalisée QUE sur des données obtenues par échantillonnage, JAMAIS sur des données exhaustives : l'objectif du test est d'aider à décider si l'écart observé peut être dû aux fluctuations d'échantillonnage ou non ; avec des données exhaustives, il n'y a aucune fluctuation d'échantillonnage, donc, il n'est pas nécessaire de faire le test pour obtenir cette réponse.

Toujours vérifier les conditions de validité

Avant de passer à l'interprétation des résultats, on doit toujours vérifier que les conditions de validité sont bien respectées. Celle des effectifs calculés est très facile à vérifier. Mais celle de l'indépendance des données est plus difficile à valider. En effet, comme évoqué précédemment, l'unité statistique en épidémiologie animale peut être tantôt l'animal, tantôt des groupes d'animaux (élevages, lots, ateliers…) en fonction de la conception du protocole, et souvent les deux unités sont présentes dans une étude : le risque est donc grand de tomber dans ce piège de réaliser le χ^2 sur les animaux (ce qui augmente les effectifs et peut conduire à la mise en évidence d'un écart significatif) alors que l'unité réelle de collecte est constituée par les élevages. Il convient donc d'être vigilant à ce sujet.

Utiliser des procédures adaptées si les données ne sont pas indépendantes

· Données appariées

Dans les études portant sur une comparaison entre deux échantillons, l'échantillonnage peut être conçu de façon à neutraliser les effets de différences non pertinentes, par exemple d'ordre démographique. Pour les études cas / témoins, il est fréquent de recourir à l'appariement des sujets, c'est-à-dire qu'après avoir sélectionné un cas, on choisit un ou plusieurs témoins dotés d'un certain nombre de caractéristiques similaires, par exemple sexe, âge, race, *etc.* De ce fait, les sujets ne sont plus indépendants, ils sont liés par cette contrainte de sélection. Le χ^2 de Pearson n'est donc plus valide et il faut lui préférer celui de Mac Nemar (voir plus loin) qui, non seulement convient dans ce type de situation, mais surtout a l'avantage de correspondre au gain de puissance important apporté par ce mode d'échantillonnage du fait de la diminution des causes parasites de variabilité. Autrement dit, l'utilisation inappropriée d'un χ^2 de Pearson sur des données appariées n'est pas une faute pouvant conduire à une erreur d'interprétation par excès, mais au contraire à une erreur par défaut, diminuant la chance de pouvoir conclure.

· Données répétées

Plusieurs observations peuvent être réalisées à des moments différents sur les mêmes individus, ou bien au même moment sur différentes unités statistiques (animaux) d'une même unité épidémiologique (élevage). Ces différentes observations ne sont pas indépendantes et ne peuvent pas être traitées par un test du χ^2. On dit qu'il s'agit de données répétées dont le traitement statistique particulier sort du cadre de cette annexe.

Présenter les résultats de façon appropriée

La présentation des résultats dans un tableau doit faire apparaître une information suffisante et pertinente, en règle générale la valeur de la probabilité d'observer nos échantillons sous l'hypothèse nulle (valeur p) fournie par le

logiciel, ou ***degré de signification***[5]. Les usages ne sont pas codifiés de façon universelle, et on peut voir des tableaux comportant des indications imagées sous forme d'étoiles dont le nombre correspond à différentes valeurs p (* : 0,05 ; ** : 0,01 ; *** : 0,001, *etc.*) : cette modalité a l'inconvénient de suggérer une possibilité d'interprétation quantitative du test ; il vaut mieux l'éviter et donner simplement la valeur objective du résultat obtenu, en indiquant d'une façon ou d'une autre (*italique* ou **gras**) le fait que la probabilité obtenue est inférieure à la valeur seuil définie au préalable (risque d'erreur α accepté *a priori*). L'indication de la valeur du χ^2 est non pertinente et à éviter, car la valeur seuil dépend du degré de liberté et du seuil de risque consenti : la valeur de probabilité (p) intègre ces différentes informations.

Dans le commentaire d'un tableau de résultats, la formulation correcte consiste à simplement constater si le ***degré de signification*** (valeur p) est ou non inférieur au ***seuil de signification*** (niveau de risque consenti d'erreur α), ce niveau étant fixé *avant* de récolter et de traiter les données (*cf.* Encadré χ^2. 4).

Encadré χ^2. 4

Seuil de signification et degré de signification

Il faut bien distinguer :

* ***le seuil de signification***, valeur du risque d'erreur α (erreur par excès) accepté,

*et ***le degré de signification*** (valeur p), probabilité d'obtenir un tel niveau d'écart sous l'hypothèse nulle, c'est-à-dire par le seul effet du hasard.

2. Au plan de l'interprétation

N'utiliser la table du χ^2 qu'après avoir défini le seuil de signification

Avant même de collecter les données, il faut avoir décidé du seuil à partir duquel on considérera l'écart comme significatif, autrement dit le ***seuil de signification***. Le plus souvent, dans l'esprit de l'épidémiologiste, ce choix est implicite (par exemple, 0,05 en conditions standards et 0,20 en situation d'investigation sur le terrain à la recherche d'hypothèses), mais il devient explicite dans la rédaction du compte-rendu.

Le seuil de 0,05, le plus souvent utilisé, est une convention arbitraire ne reposant sur aucun argument scientifique véritable qui permettrait de légitimer ce seuil.

Bien qu'arbitraire, ce seuil de 0,05, a le mérite d'être objectif, c'est-à-dire indépendant de la subjectivité de l'observateur, standard (répétable, reproductible), et par conséquent, largement admis dans la communauté scientifique. Les mêmes résultats, traités dans les mêmes conditions, donneront toujours la même interprétation statistique.

Mais il ne faut pas oublier que le seul fait du hasard peut aboutir au résultat observé dans 5 pour cent des cas. C'est pourquoi les travaux doivent être repris par d'autres chercheurs, pour établir sur d'autres jeux de données si les faits observés sont bien constants. Or, ce seuil est de plus en plus critiqué, car il laisse trop de chance au hasard et conduit à augmenter le risque de résultats dont la répétabilité est mauvaise.

Le χ^2 est un test qualitatif

Tout le monde s'accorde sur l'interprétation qualitative du résultat d'un χ^2: le degré de signification obtenu (valeur "p") est comparé au seuil de signification fixé avant l'étude ; ***si la valeur "p" est inférieure au seuil de signification***, on conclut qu'il y a de grandes chances pour que la différence observée entre les échantillons ne soit pas le fait du hasard (la ***différence*** est considérée comme « ***significative au plan statistique*** »). ***Si elle est supérieure*** au seuil de signification, il n'est pas possible de formuler ce type de conclusion et la différence est

[5] Pour les puristes, il n'est pas convenable d'utiliser des pourcentages pour évoquer des probabilités. En pratique, il faut bien reconnaître qu'il est plus facile d'utiliser la commodité de la formulation en pourcentage dès qu'on évoque des valeurs faibles : il est plus facile de dire 1 pour cent ou 1 pour mille, plutôt que 0,01 ou 0,001. On peut considérer toutefois ce mode d'expression comme « relâché », à éviter pour tout ce qui concerne la présentation des aspects méthodologiques d'un travail épidémiologique ; on retiendra l'habitude de formuler le seuil de signification sous forme de probabilité (par exemple, 0,05), de même, pour la présentation des résultats faisant appel à des degrés de signification différents (par exemple, 0,01 ou 0,001). Ce n'est que dans le feu d'une discussion des résultats que l'on pourrait admettre l'évocation de pourcentages, mais avec parcimonie.

considérée comme « ***non significative au plan statistique*** » (car le hasard pourrait expliquer la différence observée entre les échantillons). Il s'agit donc bien d'une ***interprétation qualitative***.

Toutefois, une divergence existe entre scientifiques pour aller ou non au-delà de cette interprétation qualitative unanimement acceptée. En effet, pour certains, en plus de la réponse qualitative, il est possible de prendre en considération la valeur de "p" et de considérer qu'en l'absence de biais et/ou d'erreur dans l'étude, un "p" de 10^{-6}, par exemple, correspond à une probabilité d'obtenir l'écart constaté, par le seul fait du hasard, plus faible que celle d'un "p" de 10^{-2} et, donc, conforte le rejet de l'hypothèse nulle. Cette proposition est critiquée par les tenants de l'autre conception, en raison de risques d'interprétation erronée qui pourraient en découler, surtout pour des usagers de la statistique insuffisamment expérimentés.

Un χ^2 significatif ne présage en rien de l'importance d'une différence ou de la force d'un effet.

Prenons un exemple : des visiteurs médicaux d'un laboratoire pharmaceutique présentaient les résultats d'un essai thérapeutique portant sur l'efficacité d'une nouvelle molécule par rapport au traitement conventionnel, en insistant tout d'abord sur le grand nombre des patients impliqués, respectivement 1 916 et 1 921, ce qui, d'après eux, garantissait la qualité des résultats : ils soulignaient aussi *« l'importance de l'écart : le degré de signification était de 0,005, soit dix fois plus faible que le seuil conventionnel de 0,05 ».*

Le praticien non averti pouvait se laisser convaincre par un tel argument, et finissait par choisir cette nouvelle molécule apparemment si efficace, mais, bien sûr, beaucoup plus chère.

Le tableau χ^2.3 montre les résultats de l'essai en question.

Tableau χ^2. 3

Résultats d'un essai comparant une ancienne molécule à une nouvelle

	Morts (%)	Vivants (%)	Total
Nouvelle	138 (7,2)	1 778 (92,8)	1 916
Ancienne	188 (9,8)	1 733 (90,2)	1 921

La nouvelle molécule a eu un effet favorable sur 92,8 % des sujets, l'ancienne sur 90,2 %. Ne serait-ce qu'au seul regard, cet écart d'un peu plus de 2 % ne semble pas aussi important que les visiteurs médicaux voulaient le faire paraître : l'écart absolu entre les pourcentages de résultats favorables obtenus avec les deux molécules est faible : 2,6 % (92,8-90,2). L'écart relatif l'est aussi : 2,9 % (2,6/90,2). On est loin du facteur 10 suggéré par les propos trompeurs des visiteurs. Le fait d'utiliser un grand nombre de sujets a permis de rendre statistiquement significatif (augmentation de la puissance statistique) un écart qui, au plan de l'efficacité réelle, mesurée par le rapport des performances respectives des molécules, ne l'est pas.

Le degré de signification obtenu en réalisant un test du χ^2 n'est pas lié qu'à l'importance de la différence ou de l'effet entre les deux groupes étudiés (ancien et nouveau traitement dans ce cas). Il dépend aussi beaucoup des nombres de sujets utilisés dans les échantillons : en augmentant suffisamment les effectifs, n'importe quel écart peut devenir significatif au plan statistique ! Le degré de signification permet seulement de savoir si cette différence (ou cet effet) est ***significative au plan statistique*** (et donc, probablement non liée au hasard), ou non.

C'est pourquoi il est essentiel, en plus de cette valeur, de fournir les pourcentages obtenus dans les deux groupes, constituant un ***indicateur quantitatif adapté*** (ou bien un risque relatif ou un odds-ratio, dans le cas d'enquêtes analytiques) afin de juger de ***l'importance de la différence au plan biologique*** (ou de l'effet) et donc de la « pertinence biologique » de la découverte. Ainsi, dans l'exemple évoqué il ne serait donc pas judicieux, malgré la valeur du degré de signification, d'utiliser la nouvelle molécule, compte tenu de son bénéfice très limité par rapport à son surcoût.

Ne pas oublier que le χ^2 n'est qu'une aide à la décision

Ce n'est pas le χ^2 qui donne l'interprétation, car il ne fait que fournir une indication, qui doit être intégrée dans le raisonnement portant sur l'étude dans sa globalité et notamment sur son protocole. Il faut tenir compte en effet de la conception du protocole, de la qualité de ses conditions de réalisation qui peuvent exposer à certains biais et invalider l'étude. C'est cet ensemble qui peut conduire à une interprétation suffisamment solide, et ce serait un très mauvais usage que de ne considérer qu'une faible valeur p obtenue pour conclure.

3. D'autres tests de χ^2

Il existe d'autres χ^2 que celui de Pearson.

Celui proposé par **Mac Nemar** est utilisé dans le cas des séries appariées, mais n'est utilisable que pour un tableau à un seul d.d.l. Il a l'avantage d'augmenter la puissance du résultat par rapport au χ^2 de Pearson, du fait qu'il tient compte de l'appariement.

Le test de **χ^2 d'Armitage** est aussi dénommé test de **χ^2 de tendance** ; il est utilisé lorsqu'on désire mettre en évidence une tendance évolutive d'un pourcentage selon une variable ordinale (ou quantitative discrète).

Le test de **χ^2 de Mantel-Haenszel** est utilisé dans l'analyse de variables catégorielles appariées ou stratifiées, en vue de tester l'association entre un facteur d'exposition et un état de santé. Contrairement au test de Mac Nemar, il peut être utilisé quel que soit le nombre de catégories.

Les formules de ces différents tests n'ont pas à être développées ici. Les noms de ces tests sont donnés à l'intention des utilisateurs des logiciels qui peuvent y faire référence de façon à savoir quelle information est pertinente par rapport au besoin.