Dénominateur des proportions  
*by André Gillibert*

# Dénominateur d’une proportion : Introduction

Souvent lorsqu’on analyse la relation entre deux maladies/événements/expositions, on analyse l’intersection entre les deux.

Exemple :

Supposons qu’on s’intéresse aux différences inter-communales de prévalence de diabète

Variables : diabète, rural vs urbain

La prévalence du diabète chez les habitants des communes rurales a du sens :

P(Diabète sachant Rural)

La prévalence du diabète chez les habitants des communes urbaines a du sens

P(Diabète sachant non Rural)

La proportion des diabétiques qui habitent dans des communes rurales (P(Rural sachant Diabète)) n’est a priori pas une donnée intéressante, car elle dépend énormément du ratio d’habitants rural/urbain, qui dépend fortement de la définition d’urbain et de rural.

# Quel dénominateur ? Cas général

Le plus souvent, on ne se trompe pas sur le numérateur d’un pourcentage. Il s’agit généralement du décompte d’observations vérifiant une propriété.

Exemple : Décompte(diabétiques qui ont une insuffisance rénale)

On peut clairement identifier deux variables distinctes « diabète » et « insuffisance rénale ».

Le « taux de diabétiques insuffisant rénaux » est une information ambiguë, car il ne décrit pas le dénominateur.

|  |  |  |
| --- | --- | --- |
|  | Non insuffisant rénal | Insuffisant rénal |
| Non Diabétique | a | b |
| Diabétique | c | d |

Ici, selon la population d’analyse et la maladie d’intérêt, les pourcentages en lignes et en colonnes peuvent être intéressants.

Pourcentages en lignes :

* Taux d’insuffisants rénaux chez les non diabétiques = b/(a+b)
* Taux d’insuffisants rénaux chez les diabétiques = d/(c+d)

Les pourcentages en lignes permettent d’analyser le problème de l’insuffisance rénale en tant que complication du diabète. Notamment, on peut analyser le risque relatif lié au diabète.

**Attention**: Dans une étude cas-témoin sur l’insuffisance rénale dans laquelle on fixera artificiellement le nombre d’insuffisants rénaux et de non insuffisants rénaux, ces taux ne sont pas calculables car on va fixer la prévalence de l’insuffisance rénale ! Par contre, un cas-témoin fixant le nombre de diabétiques ne pose pas de problème.

Pourcentages en colonnes :

* Taux de diabétiques chez les non insuffisants rénaux = c/(a+c)
* Taux de diabétiques chez les insuffisants rénaux = d/(b+d)

Ces pourcentages sont intéressants, surtout le second, si on s’intéresse au diabète en tant que comorbidité alourdissant le suivi de patients de néphrologie. On recherche la prévalence du diabète dans la population des insuffisants rénaux.

**Attention** : Là aussi, une étude cas-témoin sur le diabète fixera artificiellement le taux de diabétiques. Par contre, un cas-témoin fixant le nombre d’insuffisants rénaux ne pose pas de problème.

Pourcentages sur total général :

Le dénominateur est alors le total général (a+b+c+d)

Supposons qu’on veuille lancer un programme national d’éducation thérapeutique visant sélectivement les insuffisants rénaux diabétiques sur la gestion concomitante de leur traitement de l’insuffisance rénale et leur traitement antidiabétique associé au problème des règles hygiéno-diététiques. Pour ça, les insuffisants rénaux non diabétiques et les diabétiques sans insuffisance rénale ne nous intéressent pas. Nous avons juste à calculer le taux d’insuffisants rénaux diabétiques parmi l’ensemble de la population générale.

* Taux d’insuffisants rénaux diabétiques dans la population totale = d/(a+b+c+d)

En sondant sur un échantillon représentatif de la population française, puis en multipliant ce taux par le nombre de français (67 millions), nous obtenons le nombre de sujets concernés par le programme et nous pouvons en calculer le budget prévisionnel.

**Attention**: Les pourcentages sur total général n’ont pas de sens sur des cas-témoins dans lequel le nombre de diabétiques ou le nombre d’insuffisants rénaux sont fixés. Même si ça peut être dans une population limitée (p.e. population hospitalière), il faut que la prévalence de l’une et l’autre maladie soient naturelles dans cette population.

Avantages du total général : Si on présente les 4 taux, la présentation est très neutre puisqu’elle ne donne pas un rôle dominant ni à une variable ni à l’autre. Il n’y a pas de perte d’information. On peut facilement recalculer les pourcentages en ligne et en colonnes.  
L’inverse n’est pas vrai. Si on a les pourcentages en lignes mais qu’on ne connaît pas la prévalence du diabète, on n’a aucun moyen de calculer cette prévalence et de retrouver les 4 taux dont le dénominateur est le total général. De même, les pourcentages en colonnes, en l’absence de la connaissance de la prévalence de l’insuffisance rénale ne permettent pas de recalculer les 4 taux dont le dénominateur est le total général.

En bref, avant de présenter un pourcentage penser aux questions suivantes :

* Quels sont les dénominateurs interdits du fait d’un taux fixé de cas et témoins ?
* Quels sont les dénominateurs autorisés qui me fournissent le regard le plus intéressant ?
* Si aucun taux n’est fixé par un design cas-témoin, est-ce que je fournis assez de données pour qu’on puisse retrouver les 4 cases du tableau ?

# Quelques règles concernant une proportion

La majorité des pourcentages doit représenter une proportion se situant entre 0% et 100%.

Il y a quelques exceptions comme des promotions +200% de produit en plus pour -30% du prix, mais globalement, les pourcentages entre 0% et 100%, c’est plus commode.

Une proportion doit être sous la forme :

Le numérateur, c’est le nombre de personnes ayant un événement ou un état (maladie, guérison, décès)

Le dénominateur, c’est le nombre de sujets à risque de faire l’événement ou de se retrouver dans l’état.

Tout sujet situé dans le numérateur doit se retrouver au dénominateur. Seuls les événements concernant un sujet « à risque » peut être pris en compte. La violation de cette règle peut gonfler artificiellement la proportion qui peut dépasser 100%.

*Même si c’est dans le contexte d’un taux d’incidence (événement/temps) et pas d’une proportion (sujet/sujet = nombre sans dimension), j’ai déjà vu la violation de cette règle pour des effets indésirables médicamenteux (HTAP pour le Sofosbuvir) pour lesquels, au dénominateur, on mettait le temps total de traitement (personnes \* années), mais, au numérateur, on mettait des événements qui étaient survenus après l’arrêt de l’administration du traitement. C’est possible de comptabiliser les événements survenus après l’arrêt du traitement seulement si on comptabilise ce temps supplémentaire de suivi au dénominateur. Par exemple, si chaque sujet est traité 3 mois et qu’on souhaite comptabiliser les événements survenant jusqu’à 3 mois après l’arrêt du traitement, il faut multiplier par deux le dénominateur (6 mois de suivi par sujet).*

Tout sujet situé au dénominateur doit être susceptible de se retrouver au numérateur. C’est un sujet (ou du temps de sujet) **à risque**. Autrement, la proportion sera artificiellement réduite et pourra même changer alors qu’il ne se passe absolument rien dans la population à risque (pas de nouvel événement ni de nouveau sujet).

*J’ai vu la violation de cette règle dans une manière de comptabiliser le problème de l’emploi par le taux de « non employés » en population générale. Ce taux comptabilisait, au dénominateur, mais pas au numérateur, les retraités et les femmes au foyer. Ces sujets étaient donc, d’une certaine manière, considérés comme employés. Une augmentation du taux de retraité, sans changement de la population active, se traduirait alors par une baisse du taux de « non employés » parce que le dénominateur augmenterait sans que le numérateur change. De même, une augmentation du nombre de femmes au foyer serait une solution au problème de l’emploi !*

On peut alors réécrire la fraction :

Parfois certains rapports ne sont pas des proportions et ça reste quand même acceptable. Ce qui définit vraiment la proportion, c’est qu’elle est sans dimension (nombre de sujets / nombre de sujets). On peut parfois mettre en rapport des effets indésirables par des sujets et obtenir 3 effets indésirables par sujet en moyenne.

Tout sujet au dénominateur doit être susceptible de ne **pas se trouver** au numérateur et tout sujet au numérateur doit avoir eu **sa chance** de ne pas s’y retrouver. Il ne faut donc exclure les sujets qui, par principe, ont un risque de 100%.

*Exemple 1 : Dans une cohorte pour identifier l’apparition d’une maladie chronique dans l’année qui suit l’inclusion, il faut exclure les sujets qui étaient déjà malades à l’entrée de la cohorte. Leur risque est de 100% et ils sont sûrs de se trouver à la fois au numérateur et au dénominateur.*

*Exemple 2 : C’est la découverte fortuite d’un événement après la fin de suivi d’une étude (p.e. décès d’un patient). On est tenté d’aller rajouter 1 au dénominateur et au numérateur, mais c’est un sujet qui était sorti d’étude et qu’on fait rentrer de nouveau d’une manière qui garantit 100% de risque. On peut aussi dire qu’on viole le principe du « suivi » du sujet comme dans l’exemple du Sofosbuvir.*

# Différence ou rapport entre deux pourcentages

Il y a encore plus de manières de se planter !

Les règles concernant les pourcentages, devant alors s’appliquer à chacun des deux pourcentages, j’ai de nouvelles anecdotes :

*J’ai vu une comparaison du taux de connaissance de l’existence du don de lait maternel parmi les femmes en suites de couches qui ne l’ont jamais donné (X%) par rapport aux femmes qui l’ont déjà donné (100%). Ici, les femmes qui ont déjà donné leur lait (dénominateur), sont sûres à 100% de se retrouver au numérateur (sauf si elles n’ont pas compris la question ou ont une démence précoce). Par contre, il n’y a pas de problème pour le groupe qui n’a jamais donné son lait.*

Il faut en rajouter quelques nouvelles règles :

* L’événement dans les deux groupes doit être de même nature. Ce n’est pas particulièrement pertinent de comparer le taux de moustachus au taux de femmes parmi les médecins ayant un poste à responsabilité même s’ils l’ont fait dans le BMJ (Mackenzie R Wehner et al. Plenty of moustaches but not enough women: cross sectional study of medical leaders. *BMJ 2015;351:h6311*)

Globalement, trois situations se présentent :

1. Les deux groupes représentent des sujets distincts (p.e. diabétiques et non diabétiques) et on mesure un événement ou un état de la même manière dans les deux groupes. On fait alors un test de comparaison de pourcentages sur séries indépendantes (chi² de Pearson, Fisher exact)
2. Les deux groupes représentent les mêmes sujets à des moments différents (p.e. avant et après un traitement). L’événement ou l’état sont mesurés de la même manière dans les deux temps. On fait alors un test de comparaison de pourcentages sur séries appariées (chi² de McNemar, test binomial sur paires discordantes)
3. Les deux groupes sont les mêmes sujets au même moment mais l’état mesuré est opposé (p.e. on compare le taux d’hommes au taux de femmes dans une population). Dans ce cas, il ne s’agit pas vraiment d’une comparaison de pourcentages, mais d’une comparaison d’une répartition observée à une répartition théorique (50% d’hommes et 50% de femmes)