Introduction aux traitements statistiques
d'enquêtes sociologiques

Cours 4 : Inférence sur les tableaux de contingence

Damien Cartron et Martin Chevalier

Année universitaire 2024-2025

Rappel des épisodes précédents

Calcul de statistiques descriptives univariées

Prise en compte de l'aléa de sondage des enquêtes statistiques :

- mise en évidence de l'impact du sondage ;
- principe de l'inférence statistique ;
- construction d'intervalles de confiance.

Interprétation de tableaux de contingence

- vocabulaire : pourcentage en ligne, en colonne, marginaux ;
- identification de sur- et sous-représentations.

Le problème du jour

L'interprétation des sur- ou sous-représentations permet d'identifier des relations qui semblent exister entre variables.

Mais **à partir de quand** peut-on estimer que la relation entre deux variables est **suffisamment forte** pour considérer qu'elles sont ***significativement*** liées ?

En d'autres termes : les sur- ou sous-représentations mises en évidence dans la lecture du tableau de contingence sont-elles **dues au hasard** ou peuvent-elles bien être **interprétées d'un point de vue sociologique** ?

Plan de la séance

1. Écarts à l'indépendance et statistique du χ^2
2. Nature de l'aléa en jeu dans un tableau de contingence
3. Théorie des tests et test du χ^2
4. Tests statistiques et enquêtes par sondage

Écarts à l'indépendance et statistique du χ^2

Données d'exemple : Prénom et mention au bac

Chaque année au moment des résultats du bac, le sociologue Baptiste Coulmont propose une **analyse du taux de mention par prénom**.

D'un point de vue sociologique, l'intérêt de ces travaux est d'illustrer que le prénom constitue un **indicateur pertinent de déterminants sociaux plus profonds**, la **classe sociale** notamment.

Dans le cadre de cette séance, l'objectif va être de déterminer sous quelles conditions un **différentiel entre des taux de mention** selon les prénoms peut bien être **interprété d'un point de vue sociologique**.

Écarts à l'indépendance et statistique du χ^2
Lien entre prénom et mention au bac

Afin de simplifier les représentations, l'analyse ne porte que sur **deux prénoms** : **les Damien et les Martin**.

Fréquence Pourcentage Pct de ligne	Table de Prenom par Mention		
	Mention		
Prenom	Avec mention	Sans mention	Total
Damien	462 27.34 50.77	448 26.51 49.23	910 53.85
Martin	354 20.95 45.38	426 25.21 54.62	780 46.15
Total	816 48.28	874 51.72	1690 100.00

Lien entre prénom et mention au bac

En comparant le pourcentage en ligne à la marge en colonne, on constate :

- une **légère sur-représentation des mentions chez les Damien** : 50,77 % contre 48,28 % dans l'ensemble de la population (des Damien et des Martin) ;
- symétriquement, une **légère sous-représentation des mentions chez les Martin** : 45,38 % contre 48,28 % dans l'ensemble de la population.

Doit-on en conclure que le prénom Damien est **significativement associé à une plus grande probabilité d'avoir une mention au bac** ?

Ou bien l'écart entre les taux de mention selon le prénom est-il finalement **trop faible pour que l'on puisse conclure** ?

Situation d'indépendance

Un tout premier élément pour répondre à ces questions est de déterminer quel tableau de contingence on obtiendrait si les **deux variables étaient parfaitement indépendantes**, *i.e.* si prénom et mention au bac n'étaient absolument pas liés.

En pratique, cela correspond à une situation où il n'y a absolument **aucune sur- ou sous-représentation** :

- **les pourcentages en ligne correspondent exactement aux marges en colonne** : il y a exactement autant de mentions chez les Damien et chez les Martin que dans l'ensemble de la population (des Damien et des Martin) ;
- symétriquement, **les pourcentages en colonne correspondent exactement aux marges en ligne** : il y a la même distribution de prénoms que le bac ait été obtenu avec ou sans mention.

Écarts à l'indépendance et statistique du χ^2
Situation d'indépendance

Fréquence Pourcentage Pct de ligne Pct de col.	Table de Prenom par Mention		
	Mention		
Prenom	Avec mention	Sans mention	Total
Damien	439.385 26.00 48.28 53.85	470.615 27.85 51.72 53.85	910 53.85
Martin	376.615 22.28 48.28 46.15	403.385 23.87 51.72 46.15	780 46.15
Total	816 48.28	874 51.72	1690 100.00

Écarts à l'indépendance et statistique du χ^2
Situation d'indépendance

Pour déterminer les effectifs attendus dans la situation d'indépendance, il suffit de **ventiler le marges en ligne selon les marges en colonne** :

- on répartit les 910 Damien afin qu'ils comportent le même pourcentage de mentions que l'ensemble de la population :

$$n^{\text{attendu}}_{\text{Damien, Avec mention}} = 910 \times 816/1690 = 439,385$$

- on procède de même pour les 780 Martin :

$$n^{\text{attendu}}_{\text{Martin, Avec mention}} = 780 \times 816/1690 = 376,615$$

Plus généralement, dans un tableau de contingence on définit l'**effectif attendu** (*expected* en anglais) pour les modalités i et j par :

$$n_{ij}^{exp} = \frac{n_i \times n_j}{n}$$

Écarts à l'indépendance et χ^2 de cellule

Une sur- ou sous-représentation est d'autant plus importante que l'écart entre l'effectif observé (n_{ij}^{obs}) et l'effectif attendu (n_{ij}^{exp}) est **grand devant l'effectif attendu**.

On définit ainsi le χ^2 **de cellule** :

$$\chi^2_{ij} = \frac{(n_{ij}^{obs} - n_{ij}^{exp})^2}{n_{ij}^{exp}}$$

Interprétation Pour une cellule donnée, le χ^2 de cellule est d'autant plus élevé que l'effectif observé est **différent de l'effectif attendu sous l'hypothèse d'indépendance**.

Remarque En raison de la mise au carré, le χ^2 de cellule ne peut jamais être négatif, **il est toujours positif**.

Écarts à l'indépendance et χ^2 de cellule

Fréquence Attendu Khi-2 de cellule

Table de Prenom par Mention			
	Mention		
Prenom	Avec mention	Sans mention	Total
Damien	462 439.38 1.164	448 470.62 1.0868	910
Martin	354 376.62 1.358	426 403.38 1.2679	780
Total	816	874	1690

Écarts à l'indépendance et statistique du χ^2

La statistique du χ^2 de l'ensemble du tableau de contingence est souvent notée T et définie comme la **somme des statistiques du χ^2 de toutes les cellules du tableau** :

$$T = \sum_{ij} \chi^2_{ij} = \sum_{ij} \frac{(n_{ij}^{obs} - n_{ij}^{exp})^2}{n_{ij}^{exp}}$$

Ici la statistique du χ^2 vaut **4,8767** (1,164 + 1,0868 + 1,358 + 1,2679).

Interprétation Plus la statistique du χ^2 d'un tableau est élevée, plus la distribution des variables est **éloignée de l'indépendance**.

Autrement dit, plus la statistique du χ^2 est élevée, plus la **liaison entre les deux variables est statistiquement significative**.

Écarts à l'indépendance et statistique du χ^2
Lien entre prénom et mention au bac

Avec la statistique du χ^2, on peut désormais **formuler de façon plus précise** la question de l'existence d'un **lien entre prénom et mention au bac** :

- Si les deux variables étaient **parfaitement indépendantes**, la statistique du χ^2 vaudrait exactement 0 ;
- Peut-on pour autant conclure que les deux variables sont liées de façon **statistiquement significative** dès lors que la statistique du χ^2 est **strictement positive** ?

$\rightarrow$ Non, car **il arrive qu'il y ait de « petits écarts » qui sont avant tout *le fait du hasard*** et **ne traduisent pas une relation interprétable**.

Ce phénomène est le signe de l'existence d'un **aléa particulier dans les tableaux de contingence**, qu'il convient de **prendre en compte** pour déterminer si deux variables sont significativement liées ou pas.

Nature de l'aléa dans un tableau de contingence

La nature de l'aléa dans un tableau de contingence

À noter tout d'abord qu'il n'y a ici **aucun sondage** :

- on s'intéresse à toute la **population des Damien et des Martin** d'une session donnée, pas à un échantillon ;
- les mentions au bac ne sont **pas attribuées par tirage au sort**.

C'est le **phénomène analysé en lui-même** (la relation entre prénom et mention au bac) **qui est porteur de l'aléa ici**.

- Dire que prénom et mention au bac sont liés ne renvoie pas à une **relation déterministe** du type : « Tous les Damien ont une mention, aucun Martin n'a de mention ».
- Elle renvoie plutôt à une **relation probabiliste** : « La probabilité d'avoir une mention des Damien est supérieure à celle des Martin ».

La nature de l'aléa dans un tableau de contingence

En pratique, cette approche suppose une distinction entre :

- la **probabilité théorique** : quantité en général inconnue, qui dépend ici peut-être du prénom ;
- le **fréquence empirique** : pourcentage en ligne (ou colonne) dans le tableau de contingence, qui peut **différer de la probabilité théorique en raison de l'aléa**.

Pour fixer les idées On pourrait tout à fait avoir la situation suivante : en général les Martin ont plus souvent des mentions au bac que les Damien (**probabilité théorique**), mais cela ne se voit pas pour cette session (**fréquence empirique**).

On parle d'un **aléa de modèle** : l'aléatoire provient du **caractère probabiliste des hypothèses (= du modèle) qu'on formule sur les données**.

Aléa et indépendance entre variables

L'approche de l'indépendance adoptée ici est ainsi **probabiliste** :

- prénom et mention au bac sont indépendantes dès lors que la **probabilité théorique** d'avoir une mention **est la même** pour les deux prénoms... ;
- ...et non si les deux prénoms ont **toujours exactement la même fréquence empirique** de mentions au bac.

Cette **approche aléatoire de l'indépendance** fait souvent sens en sciences sociales : un grand nombre de phénomènes sociaux peuvent être décrits sous la forme de **probabilités différentiées**.

Exemple Diplôme et emploi

- s'il est vrai que certaines personnes diplômées sont au chômage et certaines personnes non-diplômées sont en emploi...
- ...il n'en reste pas moins que **la probabilité d'être au chômage diminue significativement** avec le niveau de diplôme.

Aléa et statistique du χ^2

C'est la présence de cet aléa qui fait qu'**on ne peut pas conclure que deux variables sont significativement liées dès lors que la statistique du χ^2 est strictement supérieure à 0**.

En effet, du fait de la présence de cet aléa :

- deux variables **parfaitement indépendantes** (même probabilité théorique) peuvent présenter une distribution **ne coïncidant pas parfaitement** avec la distribution attendue sous l'hypothèse d'indépendance (fréquences empiriques différentes) $\rightarrow \chi^2 > 0$;
- deux variables **significativement liées** (probabilités théoriques différentes) peuvent présenter une distribution très proche voire qui **coïncide avec la distribution attendue sous l'hypothèse d'indépendance** (même fréquence empirique) $\rightarrow \chi^2 = 0$.

Aléa et statistique du χ^2

Néanmoins, on peut s'attendre à ce que **si les deux variables sont indépendantes**, la distribution empirique ne soit **pas trop éloignée de la distribution attendue sous l'hypothèse d'indépendance**.

Moralité Si les deux variables sont indépendantes, on s'attend à ce que la statistique du χ^2 prenne des **valeurs plutôt faibles**, mais **pas nécessairement une valeur nulle** en raison de l'aléa.

Aléa et statistique du χ^2

Mais qu'est-ce qu'une valeur « plutôt faible » ? En particulier, dans notre exemple la valeur 4,8767 doit-elle être considérée comme une **valeur plutôt faible ou plutôt élevée** ?

Dis autrement :

- 4,8767 est-elle une valeur de la statistique du χ^2 **compatible avec l'hypothèse d'indépendance entre prénom et mention au bac**, i.e. **suffisamment petite** pour n'avoir été produite que par l'aléa...
- ...ou bien il s'agit d'une **valeur déjà relativement grande** qu'il est peu probable d'obtenir avec deux variables vraiment indépendantes.

Pour apporter un premier élément de réponse à cette question, on va procéder **par simulation**.

Simulations : statistique du χ^2 sous l'hypothèse d'indépendance

L'objectif des simulations est d'**observer le comportement de la statistique du χ^2 sous l'hypothèse d'indépendance** entre les variables.

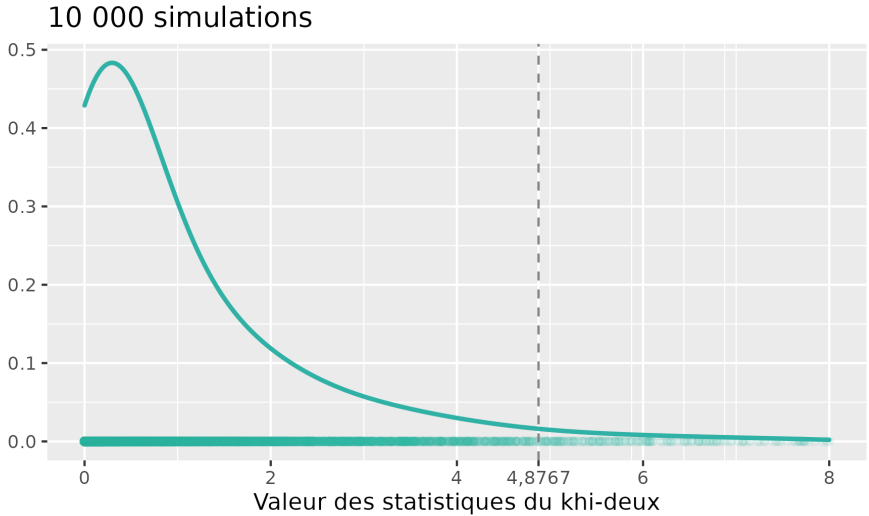
En pratique **pour chaque simulation** :

1. On détermine pour chacun des 1 690 candidats s'il a une mention ou pas en tirant au sort avec la **même probabilité 48,28 %** quel que soit son prénom (**indépendance** entre prénom et mention) ;
2. On construit le tableau de contingence correspondant et **on calcule sa statistique du χ^2**.

On réalise **10 000 simulations** et on **représente la valeur de la statistique du χ^2 obtenue** pour chacune d'entre elle.

Nature de l'aléa dans un tableau de contingence

Simulations : statistique du χ^2 sous l'hypothèse d'indépendance



Nature de l'aléa dans un tableau de contingence
Interprétation de la position de 4,8767 dans la distribution

Sur les 10 000 simulations réalisées **sous l'hypothèse d'indépendance** :

- 97,2 % des statistiques du χ^2 calculées sont inférieures à 4,8767 ;
- 2,8 % des statistiques du χ^2 calculées sont supérieures à 4,8767.

4,8767 est donc indéniablement une **valeur élevée**, qui semble **peu compatible avec l'hypothèse d'indépendance entre les deux variables** : dans la plupart des cas, quand les variables sont indépendantes la statistique du χ^2 est bien inférieure.

En même temps, **même pour deux variables totalement indépendantes**, il arrive que la statistique du χ^2 soit supérieure à 4,8767 dans 2,8 % des cas.

Moralité Sur la base des simulations menées, il semblerait qu'il y ait **2,8 % de chances de se tromper** en affirmant que prénom et mention au bac sont statistiquement liés **quand la statistique du χ^2 vaut 4,8767**.

Du risque de se tromper à la décision

Pour passer du risque de se tromper à la décision, on a recours à des **seuils usuels** : en sociologie, **on n'accepte pas un risque supérieur à 5 %** et on est plus à l'aise quand le risque est inférieur à 1 %.

Conclusions

- Sur la base de la valeur de statistique du χ^2 et de la distribution obtenue par simulation sous l'hypothèse d'indépendance, **on rejette l'hypothèse d'indépendance au seuil de 5 %**.
- En pratique, cela signifie que la relation entre prénom et diplôme peut être considérée comme **statistiquement significative** : les Damien ont **significativement plus souvent une mention que les Martin**.
- Le risque de se tromper étant relativement élevé (supérieur à 1 %), on cherchera à **éprouver la robustesse de ce résultat en examinant des variantes** (recodages, etc.).

Inférence dans un tableau de contingence

En situation réelle, on n'est **pas nécessairement en mesure** de simuler un grand nombre de tableaux de contingence sous l'hypothèse d'indépendance.

Pour déterminer si deux variables sons statistiquement liées, on va s'appuyer sur une **propriété notable** de la distribution des statistiques du χ^2 sous l'hypothèse d'indépendance : **cette distribution est régulière**.

Autrement dit : **sous l'hypothèse d'indépendance entre les variables, la statistique du χ^2 suit une loi connue**.

Un détour par la **théorie des tests statistiques** va nous permettre de comprendre comment déterminer en pratique si deux variables d'un tableau de contingence sont statistiquement liées.

Théorie des tests et test du χ^2

Test statistique : définition et statistique de test

Un test statistique est défini par une **alternative entre deux hypothèses**, appelées par convention l'hypothèse nulle (H_0) et l'hypothèse alternative (H_1).

Exemple Test qu'une variable aléatoire X vaut 0 :

$$H_0 : X = 0 \qquad \textbf{contre} \qquad H_1 : X \neq 0$$

Pour mener un test statistique, on s'appuie sur une **statistique de test** :

- « statistique » : il s'agit d'une quantité **calculable à partir des données** ;
- « de test » : son **comportement sous l'hypothèse H_0 est connu et décrit par une loi statistique**. Son comportement sous l'hypothèse H_1 diffère sensiblement.

Test d'indépendance entre deux variables qualitatives : test du χ^2

Définition Pour deux variables qualitatives X et Y comptant respectivement p et q modalités respectivement, on définit le test d'indépendance (ou **test du χ^2**) :

H_0 : (X et Y sont indépendantes) **contre** H_1 : (X et Y sont liées)

Statistique de test Sous l'hypothèse H_0 d'indépendance entre X et Y, la statistique du χ^2

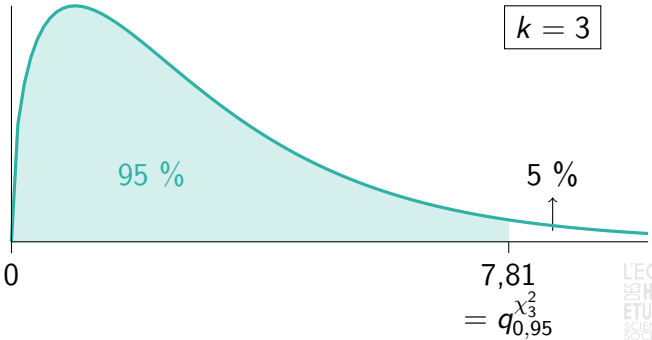
$$T = \sum_{ij} \frac{(n_{ij}^{obs} - n_{ij}^{exp})^2}{n_{ij}^{exp}}$$

suit une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté.

Parenthèse : la loi du χ^2, définition

La loi du χ^2 à k degrés de liberté est notée χ^2_k. Si $X_1, \ldots, X_k$ sont indépendantes et suivent une loi normale centrée réduite $\mathcal{N}(0,1)$, alors

$$\sum_{i=1}^{k} X_i^2 \hookrightarrow \chi^2_k$$



Parenthèse : la loi du χ^2, quantiles

k \ γ	0,900	0,950	0,990	0,999
1	2,71	3,84	6,63	10,83
2	4,61	5,99	9,21	13,82
3	6,25	7,81	11,34	16,27
4	7,78	9,49	13,28	18,47
5	9,24	11,07	15,09	20,52

Lecture

- Une loi du χ^2 à 3 degrés de liberté a 95 % de chances de prendre une valeur inférieure à 7,81.
- Une loi du χ^2 à 1 degré de liberté a 95 % de chances de prendre une valeur inférieure à 3,84 et 99 % de chances de prendre une valeur inférieure à 6,63.

Intuition sur la mécanique d'un test statistique

La statistique de test est une quantité dont le comportement est connu sous l'hypothèse nulle : quand l'hypothèse nulle est vraie la statistique de test doit avoir un **comportement régulier**, **conforme à la loi qu'elle suit sous cette hypothèse**.

Si la statistique de test prend une valeur **manifestement peu compatible avec la loi** qu'elle devrait suivre si l'hypothèse nulle était vraie, c'est que **c'est qu'on se trouve en fait sous l'hypothèse alternative**.

$\rightarrow$ **On rejette alors l'hypothèse nulle.**

En pratique cependant, même quand l'hypothèse nulle est vraie la statistique de test peut prendre des valeurs extrêmes : de ce fait, il y aura **toujours une probabilité non-nulle de se tromper en rejetant l'hypothèse nulle**, d'autant plus faible que la statistique de test prend une valeur élevée.

Mécanique d'un test statistique

1. Définition du test statistique : H_0 et H_1
2. Calcul de la statistique de test
3. Deux options (parfaitement équivalentes) :
   - 3.1 Option 1 : Comparaison de la statistique de test aux **quantiles de la loi qu'elle suit sous l'hypothèse nulle**
     $\rightarrow$ rejet de l'hypothèse nulle si la statistique est supérieure aux quantiles correspondants aux seuils usuels (5 %, 1 %)
   - 3.2 Option 2 : Calcul du **risque de se tromper** en rejetant l'hypothèse nulle étant donnée la valeur de la statistique de test, c'est-à-dire de la **p-valeur**
     $\rightarrow$ rejet de l'hypothèse nulle si la p-valeur est inférieure aux seuils usuels (5 %, 1 %)

Exemple : Prénom et mention au bac

On mène le test suivant :

$$H_0 : \text{(prénom et mention sont indépendants)}$$

contre

$$H_1 : \text{(prénom et mention sont liés)}$$

Les deux variables qualitatives ont deux modalités, aussi la statistique du χ^2 notée T suit sous l'hypothèse H_0 d'indépendance une loi du χ^2 à $(p-1)(q-1) = (2-1)(2-1) = 1$ degré de liberté :

$$T \hookrightarrow \chi^2_1$$

La valeur de la statistique de test est $T = 4,8767$.

Exemple : Prénom et mention au bac

1. Comparaison aux quantiles de la loi du χ^2 à 1 degré de liberté :
   - T est plus grande que le quantile à 95 % (3,84) : on rejette H_0 au seuil de 5 % ;
   - T est plus petite que le quantile à 99 % (6,63) : on ne peut pas rejeter H_0 au seuil de 1 %.

2. Calcul de la p-valeur :
   - la p-valeur correspondant à la valeur de T et à une loi du χ^2 à 1 degré de liberté est 0,0272 ;
   - autrement dit, sous l'hypothèse d'indépendance 2,72 % des statistiques du χ^2 sont plus grandes que T : il y a 2,72 % de risque de se tromper en affirmant sur la base de T que les deux variables sont liées ;
   - on peut donc rejeter H_0 au seuil de 5 % (2,72 % $<$ 5,00 %) mais pas au seuil de 1 % (2,72 % $>$ 1,00 %).

Exercices

- Genre et catégorie d'emploi à l'Insee : données extraites du rapport d'activité de l'Insee pour l'année 2019 ;
- Diplôme du supérieur et position sur le marché du travail : données extraites de l'enquête Emploi en continu.

Questions

- Interprétez le tableau de contingence en termes de sur- ou sous-représentations.
- Calculez à la main une statistique du χ^2 de cellule dont on peut penser *a priori* qu'elle est élevée (justifiez votre choix).
- Interprétez la valeur de la statistique du χ^2 globale en comparant aux quantiles de la loi du χ^2 avec le bon nombre de degrés de liberté.
- Interprétez la p-valeur du test.
- Rejetez-vous l'hypothèse d'indépendance entre les variables au seuil de 5 %, de 1 % ?

Tests statistiques et enquêtes par sondage

L'articulation de deux aléas

L'ensemble des exemples pris jusqu'à présent portent sur des données exhaustives :

- ensemble des Martin et des Damien de la session 2019 du bac ;
- ensemble des salariés de l'Insee ;
- échantillon d'individus de l'enquête Emploi en continu **non pondéré**.

En pratique cependant, ce sont sur des **données d'enquête** que sont menés la plupart des tests statistiques d'indépendance entre variables.

Ce faisant on est amené à articuler deux formes bien différentes d'inférence :

- l'**inférence sous le plan de sondage**, liée au fait que les données sont obtenues par sondage ;
- l'**inférence sous le modèle**, liée au fait qu'on formule des hypothèses sur le comportement des variables sous l'hypothèse d'indépendance.

Sensibilité des tests au nombre d'observations

L'articulation exacte de ces deux formes d'aléa est l'affaire de spécialiste.

En pratique, il convient de retenir une seule règle : **ne jamais pondérer un test statistique avec une pondération qui ramène l'échantillon à la taille de la population**.

Les tests statistiques sont en effet extrêmement sensibles au nombre d'observations :

- un écart de 1 pt entre deux proportions n'est pas significatif quand l'échantillon comporte 100 individus... ;
- ... mais il l'est très certainement quand il en comporte 10 000 000.

En effet, plus le nombre d'observations sur lequel porte un test statistique est élevé, moins il est probable qu'un écart à la situation d'indépendance soit **dû au hasard**, même s'il est faible.

Sensibilité des tests au nombre d'observations

Or les logiciels statistiques, quand on intègre les pondérations aux procédures usuelles de test statistique, ont tendance en général à considérer que **le nombre total d'observations correspond à la somme des poids**.

Si le poids utilisé se somme à la taille de la population, alors **tous les tests statistiques conduiront systématiquement à rejeter l'hypothèse d'indépendance entre variables** :

- des écarts faibles entre proportions, absolument non-significatifs quand ils sont calculés sur quelques milliers d'individus...
- ... seront jugés significatifs par le logiciel du fait qu'il pense qu'ils résultent de plusieurs millions d'observations.

Utilisation des pondérations

Une première option dans ce cas serait de **ne pas utiliser les pondérations** : de la sorte, on ne « trompe » pas le logiciel en lui faisant croire que l'échantillon est beaucoup plus grand qu'il n'est en réalité.

Cette option n'est en pratique **pas très satisfaisante** : en particulier dans des enquêtes dont le plan de sondage est complexe, elle peut conduire à ne **pas du tout tenir compte de fortes sur- ou sous-représentation dans l'échantillon**.

Utilisation des pondérations

L'option recommandée est la suivante :

- à partir de la pondération qui se somme à la taille de la population, construire une nouvelle variable de pondération **qui se somme à la taille de l'échantillon** :

$$\forall i, \quad w_i^{test} = w_i \times \frac{n}{\sum_i w_i}$$

- utiliser cette pondération pour mener à bien les tests statistiques, en vérifiant que le nombre d'observations pris en compte par le test correspond bien à la taille de l'échantillon.

De la sorte, le **poids relatif** des individus de l'échantillon dans la population est pris en compte sans perturber le calcul des tests statistiques.