Introduction aux traitements statistiques d'enquêtes sociologiques

# Cours 4 : Inférence sur les tableaux croisés

Damien Cartron et Martin Chevalier

Année universitaire 2020-2021

# 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 de l'intervalle de confiance d'une moyenne.

Analyse et commentaires de tableaux croisés

- vocabulaire : pourcentage en ligne, en colonne, sur- et sous-représentation ;
- indicateurs : $\chi^2$ de cellule, $\chi^2$ total, V de Cramer

# Le problème du jour

Les indicateurs calculés à partir d'un tableau croisé ne permettent pas seuls de **déterminer si la relation entre les deux variables est statistiquement significative ou pas**.

Autrement dit : à partir de quand les écarts à la situation d'indépendance sont-ils suffisamment important pour estimer que **les deux variables sont bel et bien liées** ?

Pour répondre à cette question, il va falloir prendre en compte une forme d'aléa différente de l'aléa de sondage : l'**aléa de modèle**.

Ce sera aussi l'occasion d'introduire un **aspect essentiel de l'inférence statistique**, la **théorie des tests**.

# Objectifs de la séance

1. Comprendre la nature de l'aléa en jeu dans un tableau croisé
2. Connaître les fondamentaux de la théorie des tests statistiques
3. Savoir mettre en œuvre le test du $\chi^2$ sur un tableau croisé
4. Savoir comment tenir compte des pondérations d'une enquête par sondage dans un test statistique

# Tableaux croisés et aléa de modèle

## 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 du cours, l'objectif est de déterminer si **certaines des sur- ou sous-représentations** dans le tri croisé des mentions et des prénoms correspondent à des **relations statistiquement significatives**.

# Données d'exemple : Mention au bac et prénoms

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

| Fréquence<br>Pourcentage<br>Pct de ligne |
|---|

**Table de Prenom par Mention**

| Prenom | Mention: Avec mention | Mention: Sans mention | Total |
|---|---|---|---|
| **Damien** | 462<br>27.34<br>50.77 | 448<br>26.51<br>49.23 | 910<br>53.85 |
| **Martin** | 354<br>20.95<br>45.38 | 426<br>25.21<br>54.62 | 780<br>46.15 |
| **Total** | 816<br>48.28 | 874<br>51.72 | 1690<br>100.00 |

## Données d'exemple : Prénom et mention au bac

En comparant la **valeur observée** à la **valeur théorique** si les deux variables étaient parfaitement indépendantes, on peut calculer la statistique du $\chi^2$ dans chaque cellule.

| |
|---|
| **Fréquence**<br>**Attendu**<br>**Khi-2 de cellule** |

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

# Les limites de la statistique du $\chi^2$

La statistique du $\chi^2$ pour l'ensemble du tableau vaut **4,8767**.

Son interprétation est difficile :

- une statistique du $\chi^2$ parfaitement **nulle** signifierait une **indépendance totale** entre les variables (valeur observée = valeur théorique pour toutes les cases) ;
- néanmoins en pratique cela n'arrive jamais : à partir de quelle valeur de la statistique du $\chi^2$ peut-on considérer qu'elle est **« trop grande » pour qu'elle corresponde à deux variables indépendantes** ?

En pratique ici, **4,8767 est-elle une valeur suffisamment élevée pour considérer que prénom et mention ne sont pas indépendants** ?

# La nature de l'aléa dans un tableau croisé

Cette interrogation traduit la **présence d'un aléa dans le tableau croisé**, de nature radicalement **différente de l'aléa de sondage**.

Il n'y a en effet 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 croisé

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 croisé, qui peut **différer de la probabilité théorique en raison de l'aléa**.

**Pour fixer les idées** 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**.

# La nature de l'aléa dans un tableau croisé

En effet, l'approche de l'indépendance adoptée ici est **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.

## Le retour des simulations

On ne sait **toujours pas si 4,8767** est une valeur suffisamment élevée pour considérer que prénom et mention ne sont pas indépendants.

Néanmoins, dès lors que le mécanisme à l'œuvre fait intervenir un **aléa**, une manière de l'aborder est de **procéder par simulation**.

En l'espèce ici, des simulations peuvent nous aider à **avoir une idée de la plage de variation** de la statistique du $\chi^2$ **quand les deux variables sont indépendantes**.

Il sera ainsi possible de **juger si 4,8767 est une valeur *a priori* compatible** avec l'idée que les deux variables sont **indépendantes** ou pas.

# Le retour des simulations

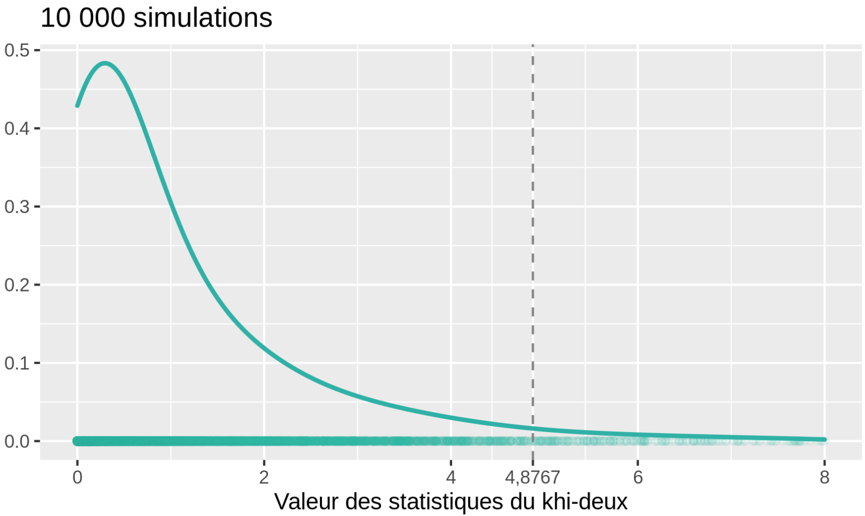
En pratique pour chaque simulation :

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 %** (fréquence marginale → **indépendance** entre prénom et mention) ;
2. On construit le tableau croisé correspondant et **on calcule la statistique du $\chi^2$ globale**.

## Remarques

- pour les simulations, on est bien obligé de procéder par **tirage au sort** (on ne peut pas reproduire le mécanisme d'attribution des mentions : examen, notation, etc.) ;
- si on dispose des données idoines, rien n'interdit d'utiliser pour ces simulations un **modèle dans lequel la probabilité varie selon les caractéristiques des candidats** (filière d'inscription, retard scolaire, etc.), à l'exception du prénom bien sûr.

# Distribution des statistiques du $\chi^2$



## Tableaux croisés et aléa de modèle

### 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 $\chi^2$ calculées sont inférieures à 4,8767 ;
- 2,8 % des statistiques du $\chi^2$ 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 $\chi^2$ est bien inférieure.

En même temps, **même pour deux variables totalement indépendantes**, il arrive que la statistique du $\chi^2$ 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 $\chi^2$ 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 $\chi^2$ 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, on cherchera à **éprouver la robustesse de ce résultat en examinant des variantes** (recodages, etc.).

# Théorie des tests et test du $\chi^2$

# Synthèse sur la démarche de la partie précédente

La démarche mise en œuvre dans la partie précédente est la suivante :

1. L'objectif est de répondre à une question binaire : les deux variables sont-elles liées ?
2. Le phénomène analysé peut être vu comme en partie aléatoire : l'existence d'un lien se traduit par un différentiel de probabilités.
3. La comparaison entre la valeur de la statistique du $\chi^2$ et une distribution sous l'hypothèse d'indépendance permet d'évaluer le risque de se tromper en affirmant que les deux variables sont liées.
4. La comparaison à des seuils de risque usuels permet de répondre à la question.

Cette démarche correspond exactement à celle proposée par la **théorie des tests statistiques**.

# Définition d'un test statistique

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 d'indépendance entre deux variables $X$ et $Y$ :

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

**Remarque** $H_0$ et $H_1$ ne sont pas interchangeables (*cf.* la fin de la partie)

# Notion de statistique de test

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.

**Exemple** Test d'indépendance entre deux variables qualitatives $X$ et $Y$ à $p$ et $q$ modalités respectivement $=$ **test du $\chi^2$**

Sous l'hypothèse $H_0$ d'indépendance entre $X$ et $Y$, la statistique du $\chi^2$

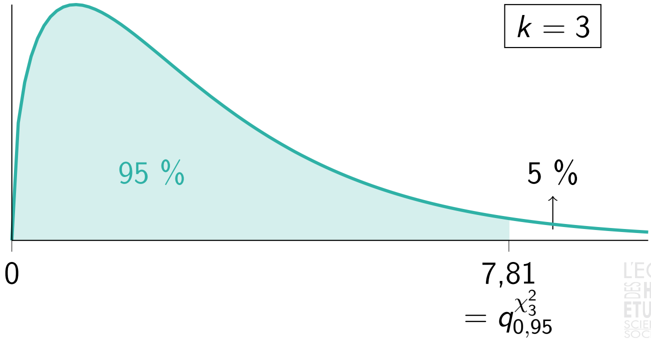
$$T = \sum_{ij} \frac{(n_{ij}^{obs} - n_{ij}^{exp})^2}{n_{ij}^{exp}}$$

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

# Parenthèse : la loi du $\chi^2$, définition

La loi du $\chi^2$ à $k$ degrés de liberté est notée $\chi^2_k$. 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 $\chi^2$, quantiles

| k \ $\gamma$ | **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 $\chi^2$ à 3 degrés de liberté a 95 % de chances de prendre une valeur inférieure à 7,81.
- Une loi du $\chi^2$ 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 vérifiée 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 vérifiée, 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 vérifiée 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 $\chi^2$ notée $T$ suit une loi du $\chi^2$ 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 $\chi^2$ 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 $\chi^2$ à 1 degré de liberté est 0,0272 ;
   - autrement dit, sous l'hypothèse d'indépendance 2,72 % des statistiques du $\chi^2$ 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

Sorties SAS sur deux sujets :

- 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 croisé en termes de sur- ou sous-représentations.
- Calculez à la main une statistique du $\chi^2$ de cellule dont on peut penser *a priori* qu'elle est élevée (justifiez votre choix).
- Interprétez la valeur de la statistique du $\chi^2$ globale en comparant aux quantiles de la loi du $\chi^2$ 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 % ?

# Types d'erreur dans un test statistique

Tout test statistique peut produire **deux types d'erreur**.

| Réalité \ Test | **$H_0$ acceptée** | **$H_0$ rejetée** |
|---|---|---|
| **$H_0$ vraie** | *Niveau*<br>1 - $\alpha$ | Faux négatifs<br>Erreur de 1$^{\text{ère}}$ espèce<br>$\alpha$ |
| **$H_0$ fausse** | Faux positifs<br>Erreur de 2$^{\text{nd}}$ espèce<br>$\beta$ | *Puissance*<br>1 - $\beta$ |

**Exemple** Test de dépistage

- faux négatifs : personnes malades non-détectées par le test ;
- faux positifs : personnes non-malades détectées par le test.

# Stratégie de Neyman-Pearson

En pratique il est **impossible de maximiser à la fois** le niveau d'un test (sa **sensibilité**) et sa puissance (sa **spécificité**).

**Exemple** Test de dépistage
Le plus sûr moyen de n'avoir aucun faux négatif est de considérer que tout le monde est malade : beaucoup de faux positifs, spécificité nulle.

Dans ce contexte, la stratégie dite de Neyman-Pearson conduit à **hiérarchiser les deux erreurs selon leur gravité** :

- l'erreur de première espèce est contrôlée de façon à pouvoir rendre $\alpha$ aussi petit qu'on le souhaite ;
- puis l'erreur de seconde espèce est minimisée autant que possible.

**Exemple** Test de dépistage
Les faux négatifs constituent clairement le risque de première espèce.

# Choix de l'hypothèse nulle

Le choix de l'hypothèse nulle d'un test est en fait le **résultat de la stratégie de Neyman-Pearson** :

1. Face à une issue binaire, $H_0$ est définie **en fonction de l'erreur qu'on considère la plus grave** (qu'on appellera l'erreur de première espèce).
2. Il ne reste alors plus qu'à **trouver une statistique dont le comportement sous $H_0$ est connu** : de la sorte, on pourra bien contrôler l'erreur de première espèce.
3. S'il en existe plusieurs, le choix entre statistiques de test s'effectue sur la base du **pouvoir discriminant** qu'elles confèrent au test.

Autrement dit : si on dispose de plusieurs statistiques de test, on préférera celle qui conduit en pratique à **la plus faible erreur de seconde espèce** pour un niveau donné.

# Choix de l'hypothèse nulle en sciences sociales

Dans le test du $\chi^2$, l'hypothèse nulle est l'**indépendance** :

- l'erreur de première espèce est d'affirmer que deux variables sont **liées alors qu'en réalité elle ne le sont pas** ;
- l'erreur de seconde espèce est d'affirmer qu'elles ne sont **pas liées alors qu'en fait elles le sont**.

Il s'agit en fait d'un **principe de prudence** : on préfère ne pas prendre le risque d'affirmer à tort que des relations entre variables existent, quitte à passer à côté de relations qui existent bel et bien.

**Mais attention** Quand on ne rejette pas l'hypothèse nulle à un certain seuil, on ne connaît pas l'erreur (de seconde espèce) à laquelle on s'expose.

# Choix de l'hypothèse nulle en sciences sociales

Ne pas pouvoir rejeter l'hypothèse d'indépendance au seuil de 5 % ne signifie pas qu'on a 5 % de se tromper en affirmant que les deux variables sont indépendantes. → On n'« **accepte** » jamais $H_0$ au seuil de 5 %.

Cela signifie seulement qu'on a plus de 5 % de chances de se tromper en affirmant que les deux variables ne sont pas indépendantes.

**Exemple** Discrimination

- Ne pas montrer qu'il y a une discrimination sur la base de données statistiques ne signifie pas montrer qu'il n'y a pas de discrimination.
- En pratique, on contrôle bien le risque d'affirmer à tort qu'il y a une discrimination mais pas du tout celui d'affirmer à tort qu'il n'y en a pas.
- Il faudrait un autre test, avec les hypothèses $H_0$ et $H_1$ inversées : mais sauf à postuler un comportement de discrimination précis, on ne connaît pas la loi que suit la statistique de test sous $H_0$ dans ce cas.

# Tenir compte des pondérations dans un test statistique

## 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.