Titre : Le paradoxe de Simpson

*Les statistiques, c'est comme le bikini. Ce qu'elles révèlent est suggestif. Ce qu'elles dissimulent est essentiel.*

Aaron Levenstein

Degrés : 1e - 4e du collège /
PR, 1e - 2e de l'ECG

Durée : 45 minutes

Résumé :

Une manifestation est organisée devant le bureau du président de l'Université. Un groupe d'étudiantes dépose une pétition signée par de nombreuses étudiantes et …quelques étudiants : « Les résultats aux examens de cette année montrent que les filles ont obtenu un meilleur pourcentage de réussite aux examens de la licence de Physique que les garçons. Les filles ont aussi obtenu un meilleur pourcentage de réussite que les garçons aux examens de la licence de Biologie. Pourtant, en globalisant les résultats des deux licences, les filles sont moins nombreuses en proportion que les garçons à obtenir leur licence et cela bien qu'il y ait au total exactement le même nombre de candidates filles que de candidats garçons. Une manipulation sexiste a donc été faite. »

Le président de l'Université est ennuyé. Il vérifie les résultats et les calculs et ne trouve aucune erreur : il est parfaitement vrai que les filles ont un meilleur pourcentage de réussite que les garçons aux examens de la licence de Biologie (55,5% contre 40%), ainsi qu'aux examens de Physique (100% contre 89%). Pourtant en regroupant les résultats des deux licences, le pourcentage de réussite des filles est nettement moins bon que celui des garçons (60% contre 84%). Il n'y a absolument aucune erreur.
Que se passe-t-il ? Cela est-il possible ?

# Le paradoxe de Simpson

*Les statistiques, c'est comme le bikini. Ce qu'elles révèlent est suggestif. Ce qu'elles dissimulent est essentiel.*

Aaron Levenstein

## Problème de départ

Une manifestation est organisée devant le bureau du président de l'Université. Un groupe d'étudiantes dépose une pétition signée par de nombreuses étudiantes et …quelques étudiants : « Les résultats aux examens de cette année montrent que les filles ont obtenu un meilleur pourcentage de réussite aux examens de la licence de Physique que les garçons. Les filles ont aussi obtenu un meilleur pourcentage de réussite que les garçons aux examens de la licence de Biologie. Pourtant, en globalisant les résultats des deux licences, les filles sont moins nombreuses en proportion que les garçons à obtenir leur licence et cela bien qu'il y ait au total exactement le même nombre de candidates filles que de candidats garçons. Une manipulation sexiste a donc été faite. »

Le président de l'Université est ennuyé. Il vérifie les résultats et les calculs et ne trouve aucune erreur : il est parfaitement vrai que les filles ont un meilleur pourcentage de réussite que les garçons aux examens de la licence de Biologie (55,5% contre 40%), ainsi qu'aux examens de Physique (100% contre 89%). Pourtant en regroupant les résultats des deux licences, le pourcentage de réussite des filles est nettement moins bon que celui des garçons (60% contre 84%). Il n'y a absolument aucune erreur.
Que se passe-t-il ? Cela est-il possible ?

Avant de répondre à ce problème, nous allons nous intéresser au suivant.

## 1ère partie : Problème du scientifique

Un scientifique a effectué des expériences cliniques afin de déterminer les efficacités relatives de deux traitements. Il a obtenu les résultats suivants :

| | Traitement A | Traitement B |
|---|---|---|
| Succès | 219 | 1010 |
| Échec | 1801 | 1190 |

1) Quel traitement est le plus efficace ? (*Justifier par un ou plusieurs calculs.)*

Après avoir annoncé ce résultat (dites lequel…), un de ses assistants vient vers lui. Il est en désaccord avec l'interprétation des résultats. Il lui présente le tableau suivant, dans lequel les résultats précédents sont donnés en tenant compte du sexe des patients :

| | Femmes | | Hommes | |
|---|---|---|---|---|
| | Traitement A | Traitement B | Traitement A | Traitement B |
| Succès | 200 | 10 | 19 | 1000 |
| Échec | 1800 | 190 | 1 | 1000 |

2) Quel traitement est alors plus efficace suite à la remarque de l'assistant ? (*Justifier par un ou plusieurs calculs.)*
3) Qui a donc raison ? Trouvez une explication à ce problème. (*Justifier par des phrases.)*

**2ème partie : Retour au problème de départ *(A faire quand la première partie a été réussie et comprise)***

Voici les données de notre exemple dans le cas où il y a 100 filles et 100 garçons.

| | Physique | | Biologie | | Cumul | |
|---|---|---|---|---|---|---|
| | Garçons | Filles | Garçons | Filles | Garçons | Filles |
| Réussites | 80 | 10 | 4 | 50 | …. | …. |
| Échec | 10 | 0 | 6 | 40 | …. | …. |
| Total | …. | …. | …. | …. | …. | …. |

1) Compléter les trous dans le tableau.
2) Quel est le taux de réussite en Biologie et en Physique ?
3) Que s'est-il passé dans ce problème. Donnez une explication.
4) Doit-on conclure qu'on peut faire dire une chose et son contraire aux statistiques ? Risque-t-on toujours de rencontrer de telles situations ?

**3ème partie :**
Êtes vous capable de trouver une autre situation (même inventée) où le paradoxe de Simpson pourrait s'appliquer ? *(Une rédaction du problème bien rédigée est exigée.)*

Titre : Le paradoxe de Simpson

Degrés : 1ère-2ème-3ème année du collège / 2ème ECG / 1ère préparatoire

Prérequis :
Uniquement savoir calculer une moyenne.

Objectifs :
L'objectif principal est de travailler avec les statistiques et de savoir interpréter des données.

Matériel :
De quoi écrire.

Durée estimée :
45 minutes

Proposition de déroulement :

Annoncer un cours à l'avance qu'une activité de 45 minutes à réaliser par groupes de deux élèves aura lieu le cours suivant, avec un rapport écrit à rendre à la fin. Il est conseillé d'annoncer qu'il sera noté. Cela stresse un peu les élèves, mais cela les stimule et comme la note est souvent bonne, ils terminent avec une belle satisfaction.
Laisser les élèves avancer l'activité.
Il est recommandé de travailler l'activité en deux parties. S'assurer que les élèves aient compris la première partie pour pouvoir faire la seconde.

Analyse a priori de l'activité :
Ce qui sera difficile dans cette activité est de trouver le paradoxe. Hors de ceci, l'activité ne présente aucune difficulté puisque le seul objet mathématique qui intervient est l'utilisation de la moyenne.
Si l'élève ne comprend pas le paradoxe dans la première partie, il est conseillé de le lui expliquer afin qu'il puisse le trouver dans la deuxième partie.
La troisième partie est plutôt destinée aux meilleurs élèves, ceux qui auraient terminé à l'avance.

# Résolution :

## 1ère partie : Problème du scientifique

1) Le traitement A ayant été administré à 2020 personnes, et 219 d'entre elles ayant été guéries, son taux de succès est donc de 219/2020, ce qui est très inférieur aux taux correspondant pour le traitement B qui est de 1010/2200. Par conséquent, le traitement B est plus efficace que le traitement A.
2) Chez les femmes, le taux de succès des traitements sont de 1/10 et 1/20 respectivement, et chez les hommes de 19/20 et ½. Le traitement A est donc plus efficace dans les deux cas. Par conséquent, le traitement A est plus efficace que le traitement B.
3) C'est l'assistant qui a raison : quel que soit le sexe du patient, ses chances de guérir sont supérieures avec le traitement A. Ce paradoxe apparaît régulièrement dans des études statistiques. Observez aussi la difficulté suivante : si l'on n'avait pas relevé le sexe des patients, on aurait été obligé de baser notre analyse sur le premier raisonnement, et on serait arrivé à une conclusion erronée.

   En particulier, comment être certain qu'il n'existe pas d'autres paramètres que le sexe (l'âge, le poids, . . . ) dont on n'aurait pas tenu compte et qui modifieraient une fois de plus la conclusion ?

## 2ème partie : Retour au problème de départ

1)

| | Physique | | Biologie | | Cumul | |
|---|---|---|---|---|---|---|
| | Garçons | Filles | Garçons | Filles | Garçons | Filles |
| Réussites | 80 | 10 | 4 | 50 | 84 | 60 |
| Échec | 10 | 0 | 6 | 40 | 16 | 40 |
| Total | 90 | 10 | 10 | 90 | 100 | 100 |

2) L'examen de Physique a un taux de réussite de 90% et la licence de Biologie a un taux de réussite de 55%.
3) Un peu d'attention donne la clef du mystère. Les filles sont plus nombreuses en licence de Biologie et les garçons plus nombreux en licence de Physique. Or le taux de réussite est meilleur en Physique qu'en Biologie. Les filles tentent donc, en moyenne, un examen plus difficile que les garçons qui ne gagnent en réalité que parce qu'ils optent pour la facilité. Les filles sont 90 sur les 100 à tenter la licence de Biologie qui a un taux de réussite de 55% ; les garçons sont 90 sur 100 à tenter l'examen de Physique qui a un taux de réussite de 90%

4) La réponse est non et le paradoxe de Simpson provient de ce que dans l'agrégation des données qui le provoque, on ne mélange pas des données correspondant à des effectifs égaux pour les sous cas. Il est facile de démontrer qu'en regroupant les résultats des examens de
   - 100 filles passant la Biologie
   - 100 filles passant la Physique

- 100 garçons passant la Biologie
- 100 garçons passant la Physique

Alors le paradoxe de Simpson disparaît.