Principes de l'échantillonnage et application aux activités des CNR

Yann Le Strat

Santé publique France

9ème séminaire des CNR - 15/11/2019

Échantillon

1. Un **échantillon** est une partie d'une population.

Échantillon

1. Un **échantillon** est une partie d'une population.
2. Il n'y a que des inconvénients à travailler sur un échantillon plutôt que sur l'ensemble de la population.

Échantillon

1. Un **échantillon** est une partie d'une population.
2. Il n'y a que des inconvénients à travailler sur un échantillon plutôt que sur l'ensemble de la population.
3. Travailler avec un échantillon est dicté par des contraintes budgétaires

Échantillon

1. Un **échantillon** est une partie d'une population.
2. Il n'y a que des inconvénients à travailler sur un échantillon plutôt que sur l'ensemble de la population.
3. Travailler avec un échantillon est dicté par des contraintes budgétaires
4. Entraînant des inconvénients : on ne connait pas la réalité
   $\Rightarrow$ estimation + incertitude

Échantillon

1. Un **échantillon** est une partie d'une population.
2. Il n'y a que des inconvénients à travailler sur un échantillon plutôt que sur l'ensemble de la population.
3. Travailler avec un échantillon est dicté par des contraintes budgétaires
4. Entraînant des inconvénients : on ne connait pas la réalité $\Rightarrow$ estimation + incertitude
5. Cela peut entraîner 2 choses :
   - un biais (au niveau de l'estimation)
   - une trop grande incertitude

Échantillon

1. Un **échantillon** est une partie d'une population.
2. Il n'y a que des inconvénients à travailler sur un échantillon plutôt que sur l'ensemble de la population.
3. Travailler avec un échantillon est dicté par des contraintes budgétaires
4. Entraînant des inconvénients : on ne connait pas la réalité $\Rightarrow$ estimation + incertitude
5. Cela peut entraîner 2 choses :
   - un biais (au niveau de l'estimation)
   - une trop grande incertitude
6. Malgré cela, les recensements n'existent plus (raison 3), même la statistique publique échantillonne.

Échantillon

1. Un **échantillon** est une partie d'une population.
2. Il n'y a que des inconvénients à travailler sur un échantillon plutôt que sur l'ensemble de la population.
3. Travailler avec un échantillon est dicté par des contraintes budgétaires
4. Entraînant des inconvénients : on ne connait pas la réalité $\Rightarrow$ estimation + incertitude
5. Cela peut entraîner 2 choses :
   - un biais (au niveau de l'estimation)
   - une trop grande incertitude
6. Malgré cela, les recensements n'existent plus (raison 3), même la statistique publique échantillonne.
7. Tout le monde travaille avec des échantillons !

Deux situations

- Situation favorable : on construit un échantillon en maîtrisant les différents points à respecter
  - On utilise des bases de sondage (listes)
  - On réalise des tirages aléatoires
  - On sait calculer des poids de sondages
  - On mobilise la théorie des sondages
  - **Risques de biais**

Deux situations

- Situation favorable : on construit un échantillon en maîtrisant les différents points à respecter
  - On utilise des bases de sondage (listes)
  - On réalise des tirages aléatoires
  - On sait calculer des poids de sondages
  - On mobilise la théorie des sondages
  - **Risques de biais**

- Situation défavorable : on dispose d'un échantillon dont on n'a pas contrôlé la construction
  - On ne sait pas comment il a été construit
  - Il ne provient pas de tirages aléatoires
  - On ne peut pas faire une inférence l'esprit tranquille
  - Seul outil à mobiliser : les redressements
  - **Risques de biais +++**

Raisons pouvant entraîner un biais

Raisons maîtrisables :

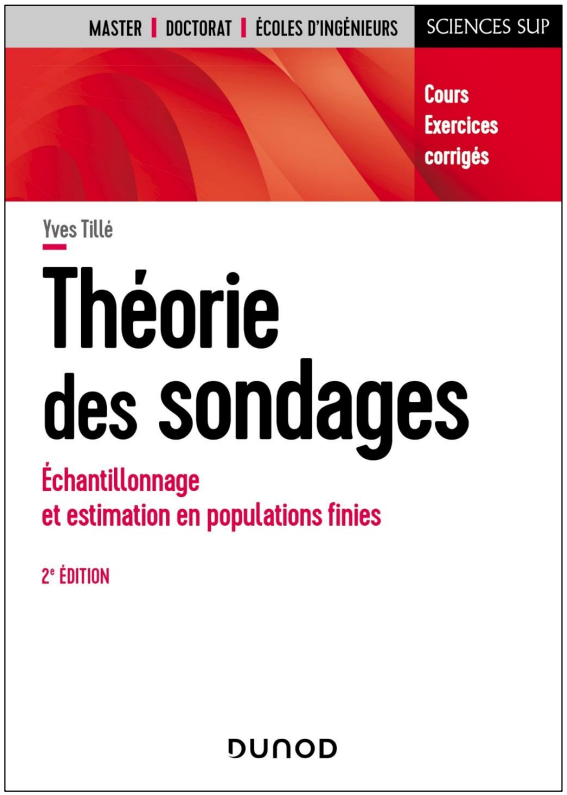
- La base de sondage a un défaut de couverture (la population n'est pas entièrement couverte par la base)
- Le tirage des unités de la base n'est pas aléatoire

Raisons non maîtrisables :

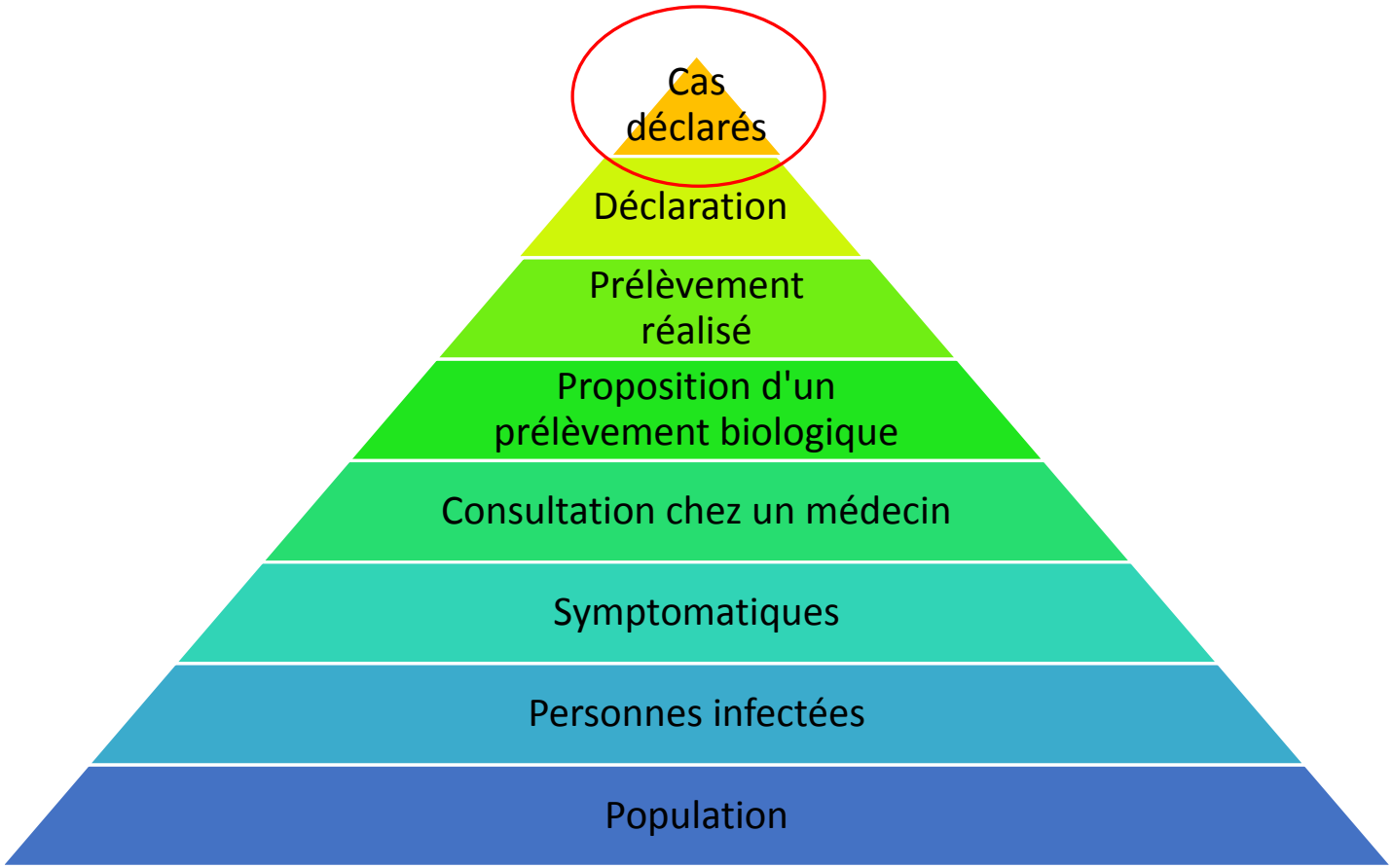
- Non-réponse totale : les raisons de la participation (ou non) à l'enquête sont corrélées à la variable d'intérêt (exemple : les personnes malades souhaitent moins répondre)
- Non-réponse partielle : une personne ne souhaite pas répondre à des questions et la raison n'est pas indépendante de la variable d'intérêt
- Il existe des erreurs de mesures lors de l'enquête, involontaires (incompréhension, ...) ou volontaires (mensonge, faux questionnaires remplis, etc.))

Situation favorable : on réalise une enquête

- On mobilise des méthodes de tirages aléatoires dans des listes plutôt de bonne qualité
- Des stratifications
- Des redressements
- Et toutes les méthodes d'estimations dans la population - Voir par exemple cet ouvrage de référence $\rightarrow$

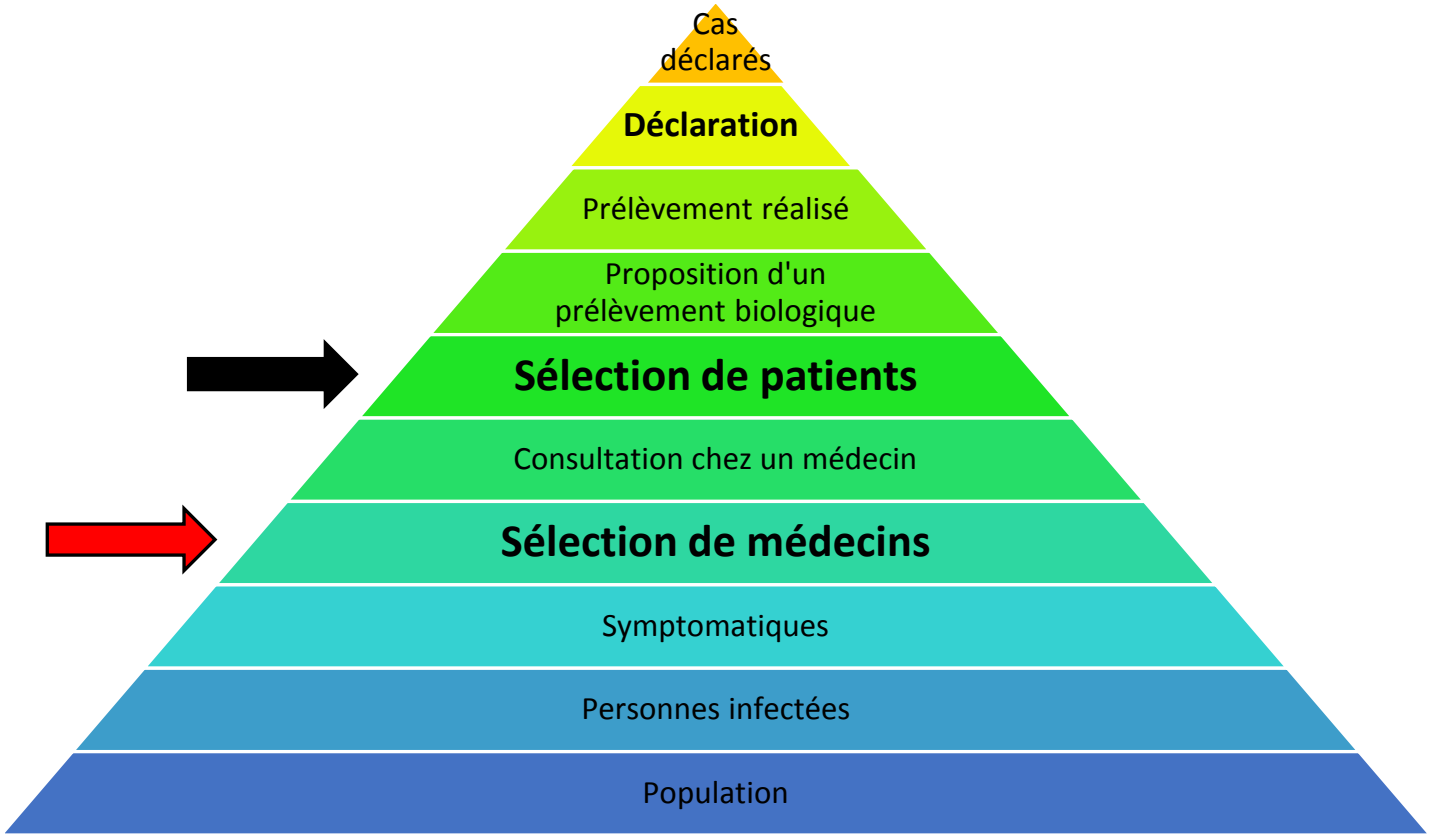


Situation défavorable : la surveillance épidémiologique



On ne contrôle pas l'ensemble du processus conduisant à l'échantillon des cas déclarés.

On va cependant essayer de contrôler certaines étapes

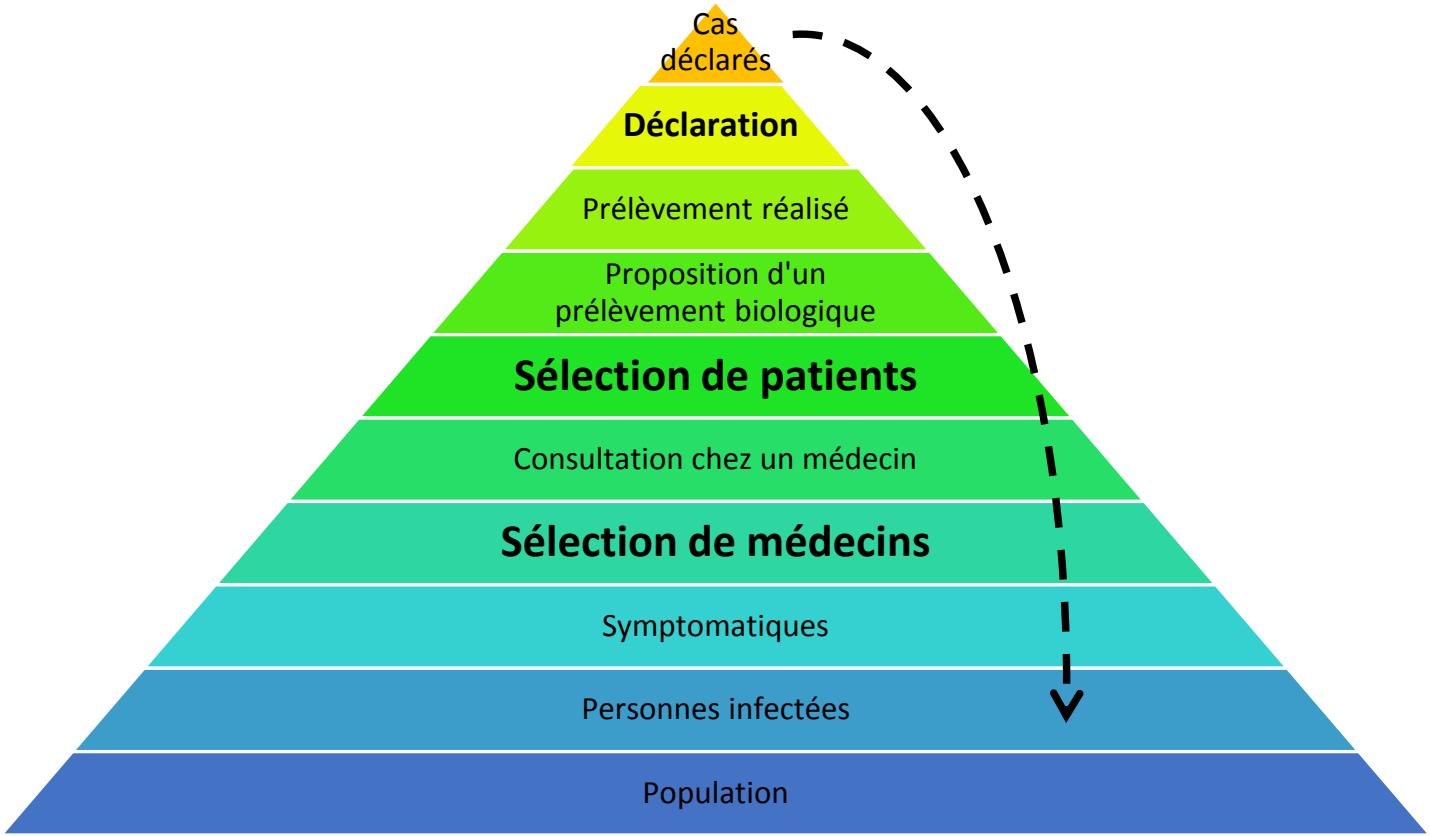


La sélection des déclarants est primordiale (ici des médecins). Cela permet de contrôler les étapes suivantes (sélection et nombre de patients, réalisation des prélèvements, déclaration, recueil de données sur les déclarants et les patients pour redresser).

Les sélections

- La sélection des médecins ou des biologistes ne repose pas sur l'aléatoire. La surveillance repose sur le volontariat (biais de participation éventuel).
- Il est à noter que même lorsqu'on fait des tirages aléatoires, on obtient au final que des volontaires pour participer.
- Les patients par contre peuvent être assez facilement tirés aléatoirement.
- Une fois qu'on obtient les prélèvements on doit tous les analyser mais on peut sans doute faire du tirage aléatoire pour l'utilisation de telles ou telle techniques (si plusieurs techniques sont utilisées - exemple séquençage/typage).
- Même lorsqu'il n'y a pas d'aléatoire, les échantillons de médecins, patients, biologistes, prélèvements méritent **toujours** d'être diversifiés en termes de région, type de médecins, caractéristiques patients (age, gravité), provenance des patients (ville/hôpital), etc.

L'inférence à la population



Mais les déclarants ne sont que des "intermédiaires". L'unité d'intérêt est le malade. On s'intéresse davantage à l'échantillon des malades qu'à l'échantillon des médecins (on redresse sur leurs activités pour estimer une incidence).

Comment gérer les nombreux objectifs d'un CNR ?

Comme dans toute étude ou enquête, la multiplicité des objectifs entraîne des choix cornéliens dans l'échantillonnage.

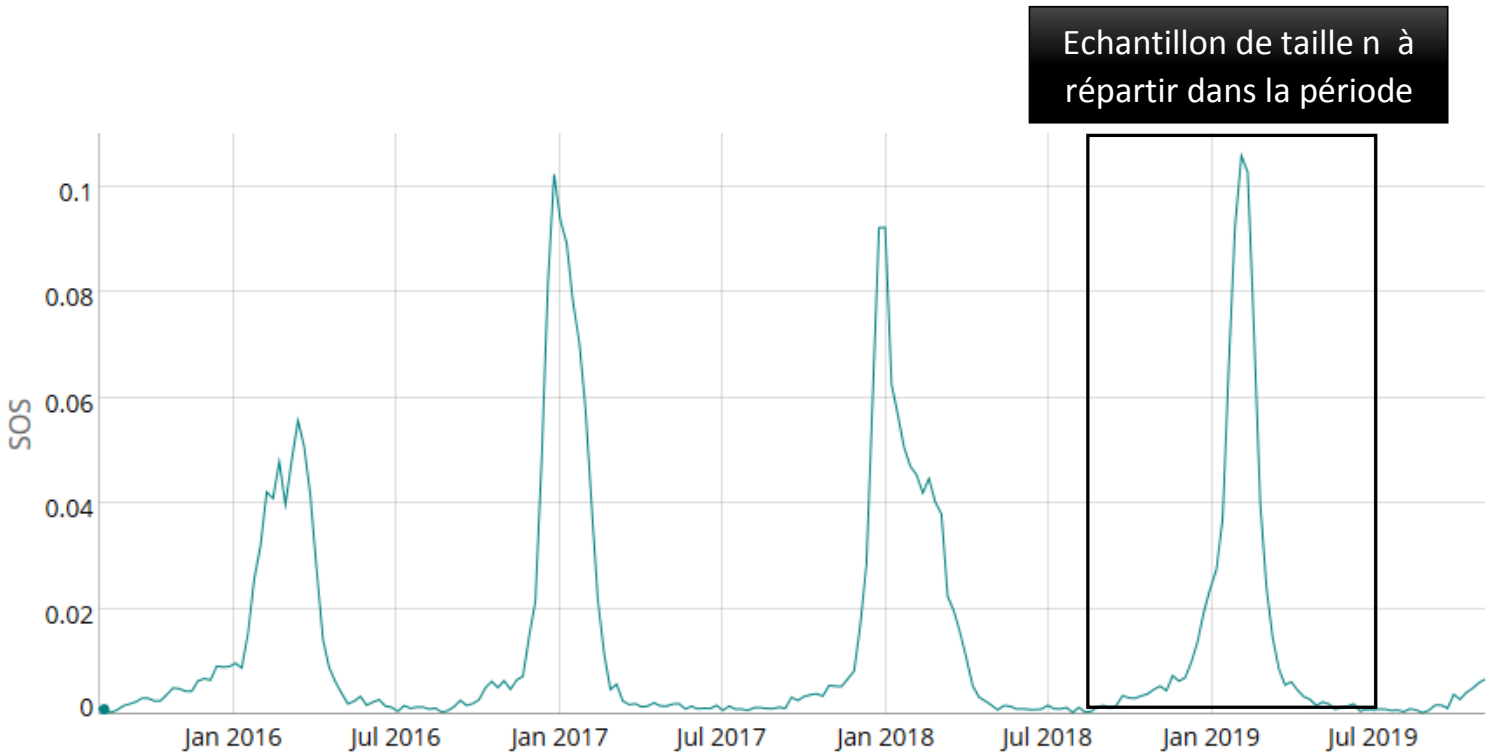
On peut alors :

- choisir de réduire le nombre d'objectifs (difficile de s'y résigner car on veut tout faire)
- faire un échantillonnage adaptatif au cours du temps (assez facile)

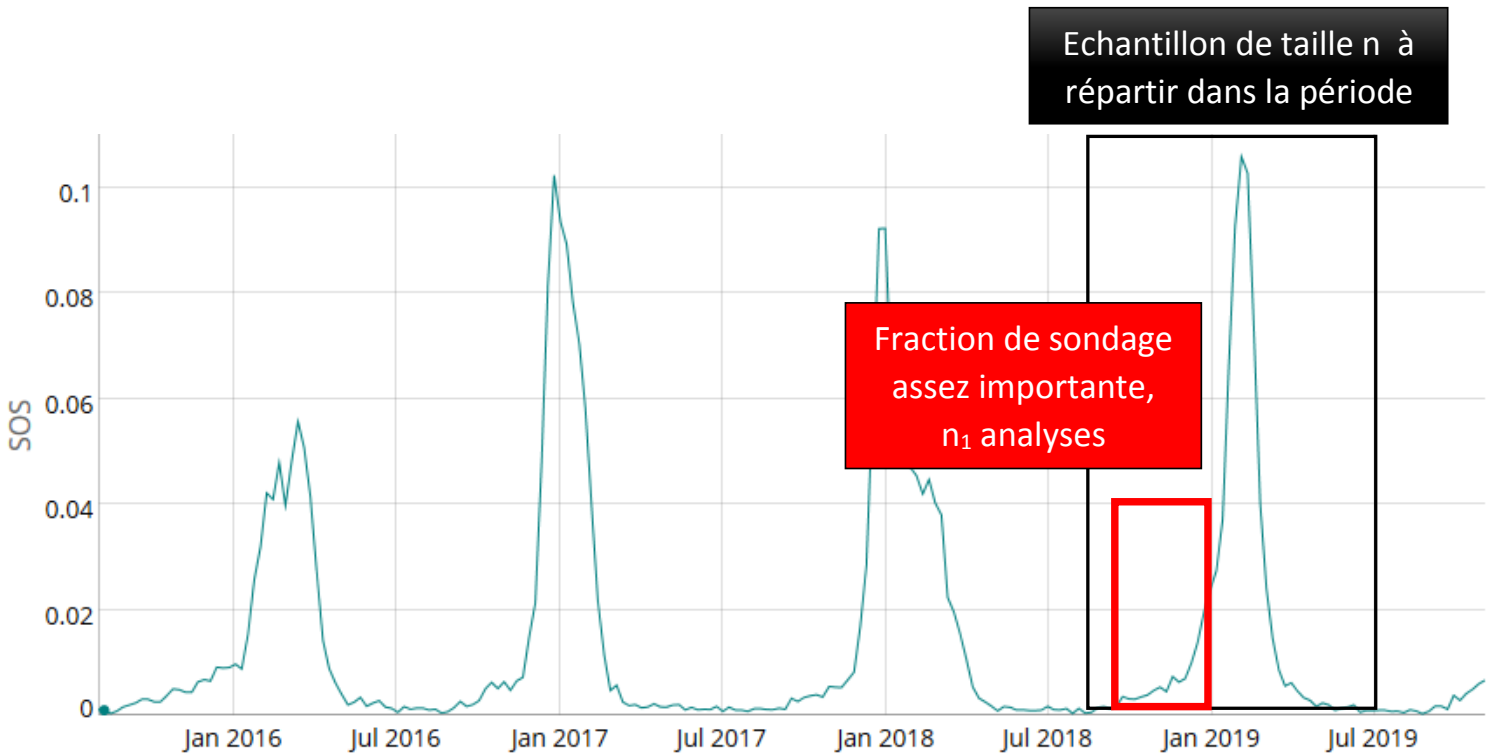
Exemple d'objectifs CNR grippe

- **Identifier** les virus circulants (typage, sous-typage)
- **Décrire** les caractéristiques antigéniques des virus circulants
- Réaliser des analyses phylogéniques
- Mesurer la résistance aux antibiotiques
- **Étudier l'adéquation** des souches vaccinales au regard des souches virales circulantes

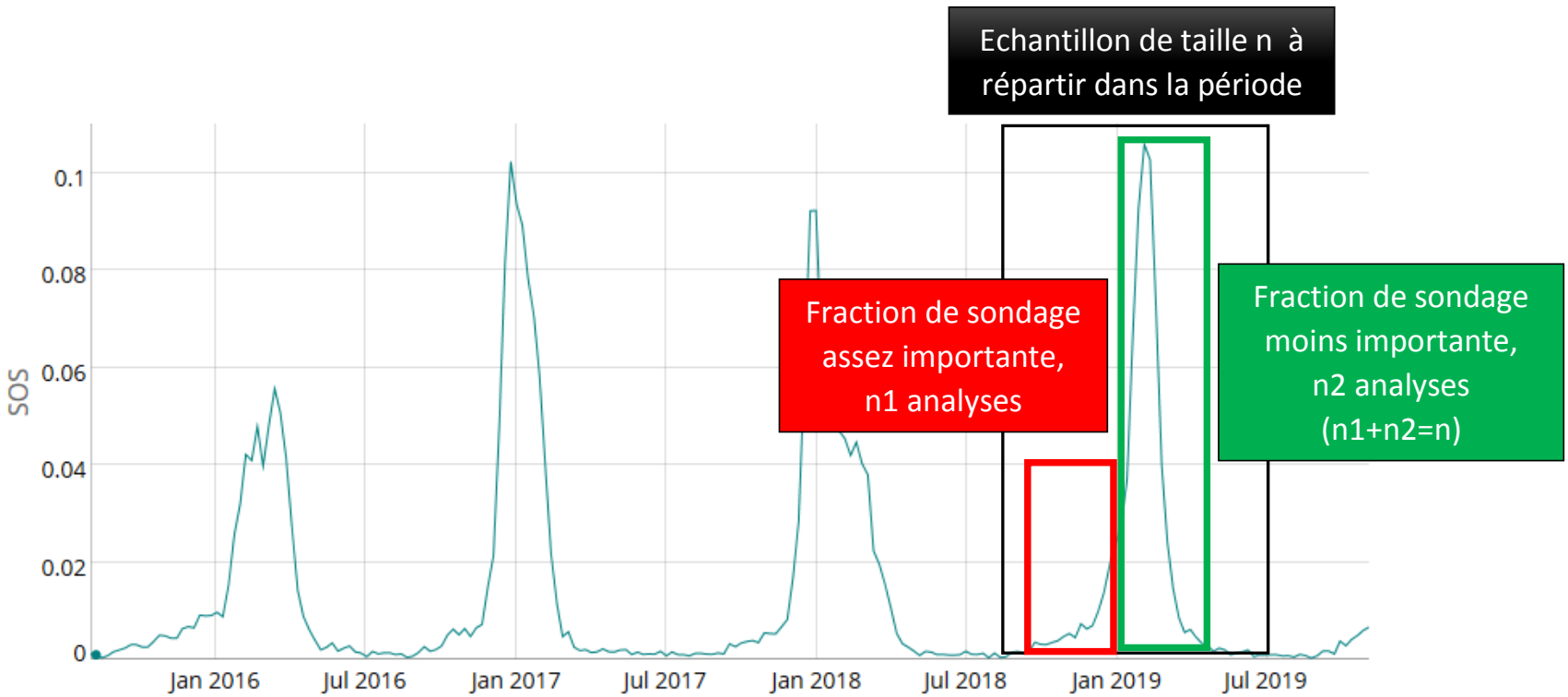
Sondage adaptatif - Illustration sur la grippe



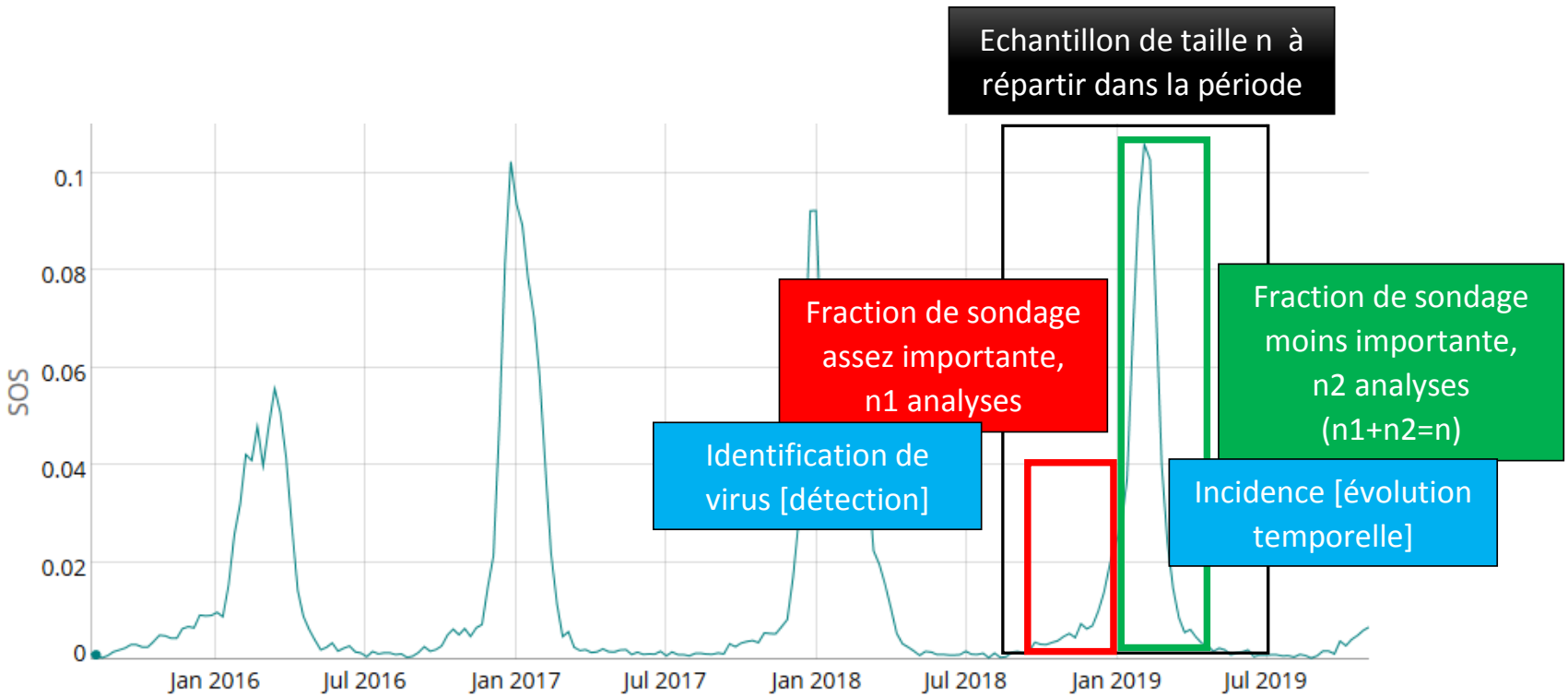
Sondage adaptatif - Illustration sur la grippe



Sondage adaptatif - Illustration sur la grippe



Sondage adaptatif - Illustration sur la grippe



En conclusion

- Lorsqu'on peut faire de l'aléatoire il faut le faire (pas de négociation possible)
- Quand on ne peut pas avoir un échantillon aléatoire, il faut au moins qu'il soit diversifié
- Se raccrocher à certaines techniques pour améliorer ce que l'on peut en dire (stratification, redressements)
- Réfléchir à stratifier les prélèvements pour appliquer différentes méthodes s'il y a un besoin (réduction de coût ou autres raisons)
- Essayer de ne pas avoir un échantillonnage gravé dans le marbre sur toute la saison de surveillance mais l'adapter s'il faut répondre à plusieurs objectifs qui dépendent du temps (détection versus description)