Séminaire méthodologique

Capture-Recapture

Département d 'Épidémiologie et de Promotion de la Santé

Septembre 2003

Plan de la présentation

1. Introduction et genèse de la méthode
2. Application à l'épidémiologie
3. Différents domaines d'application
4. Exemples

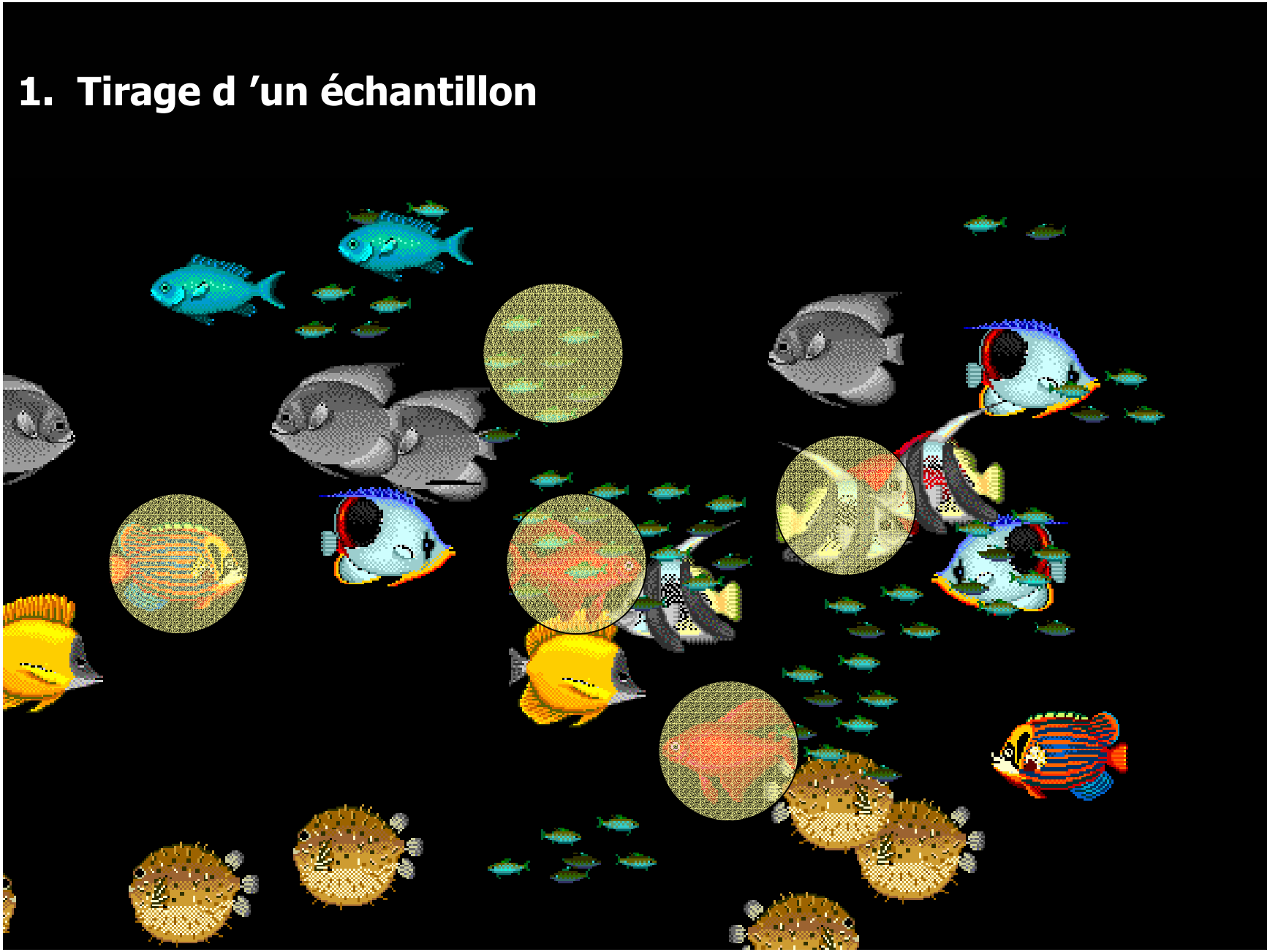
I. Introduction et genèse de la méthode Capture-recapture

Appliquer d'abord en **Zoologie** pour estimer la taille de populations animales (notamment oiseaux, poissons).

Principe utilisé:

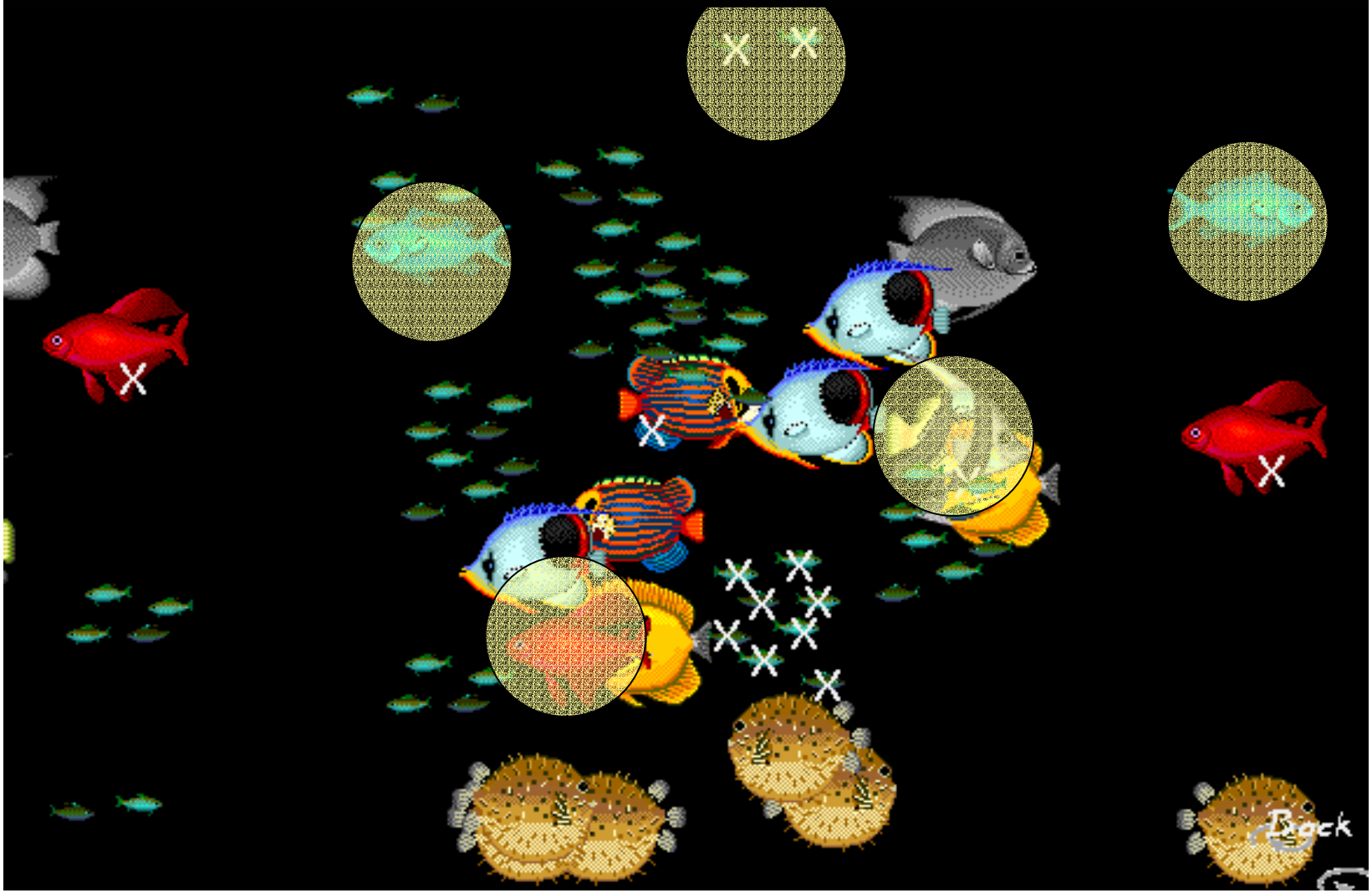
- **Tirer un échantillon aléatoire d'une espèce animale**
- **"Marquer" les animaux tirés au sort, puis les relâcher**
- **Tirer un second échantillon et compter le nombre d'animaux marqués**
- **Estimer la population totale en appliquant une règle de 3**

1. Tirage d ’un échantillon



2. Marquage des tirés au sort

3. Remise de l ’échantillon



4. Retirage d ’un échantillon

$$\frac{\#\text{ recapture échantillon 2}}{\#\text{ échantillon 2}} \approx \frac{\#\text{ capture dans échantillon 1}}{\#\text{ population totale}}$$

N= Ech1 x Ech2/ #recapture

I. Introduction et genèse de la méthode Capture-recapture

Démographie:

- Estimer la taille des populations
- Compenser les sous-dénombrements
- Estimer les taux de natalité et de mortalité

Epidemiologie:

Première utilisation : Travaux de Wittes (1968)

(évaluation de la fréquence des malformations à la naissance).

Dévelloppement réel: Années 80

II. Application de la méthode à l'épidémiologie

- **Objectif:**

- **Evaluation de l'exhaustivité (et qualité) des systèmes de surveillance**
  - » Rarement exhaustif
  - » Notification dépend de nombreux facteurs:
    - Facteurs personnels (statut socio-éco, géographique,hisorique médical)
    - Sévérité de la maladie
    - caractéristique des tests de diagnostiques (Sens, spec,VPP,VPN)

→ **Estimer le nombre de sujets non répertoriés**

II. Application de la méthode à l'épidémiologie

- **Principe:**

Croisement/comparaison de plusieurs sources

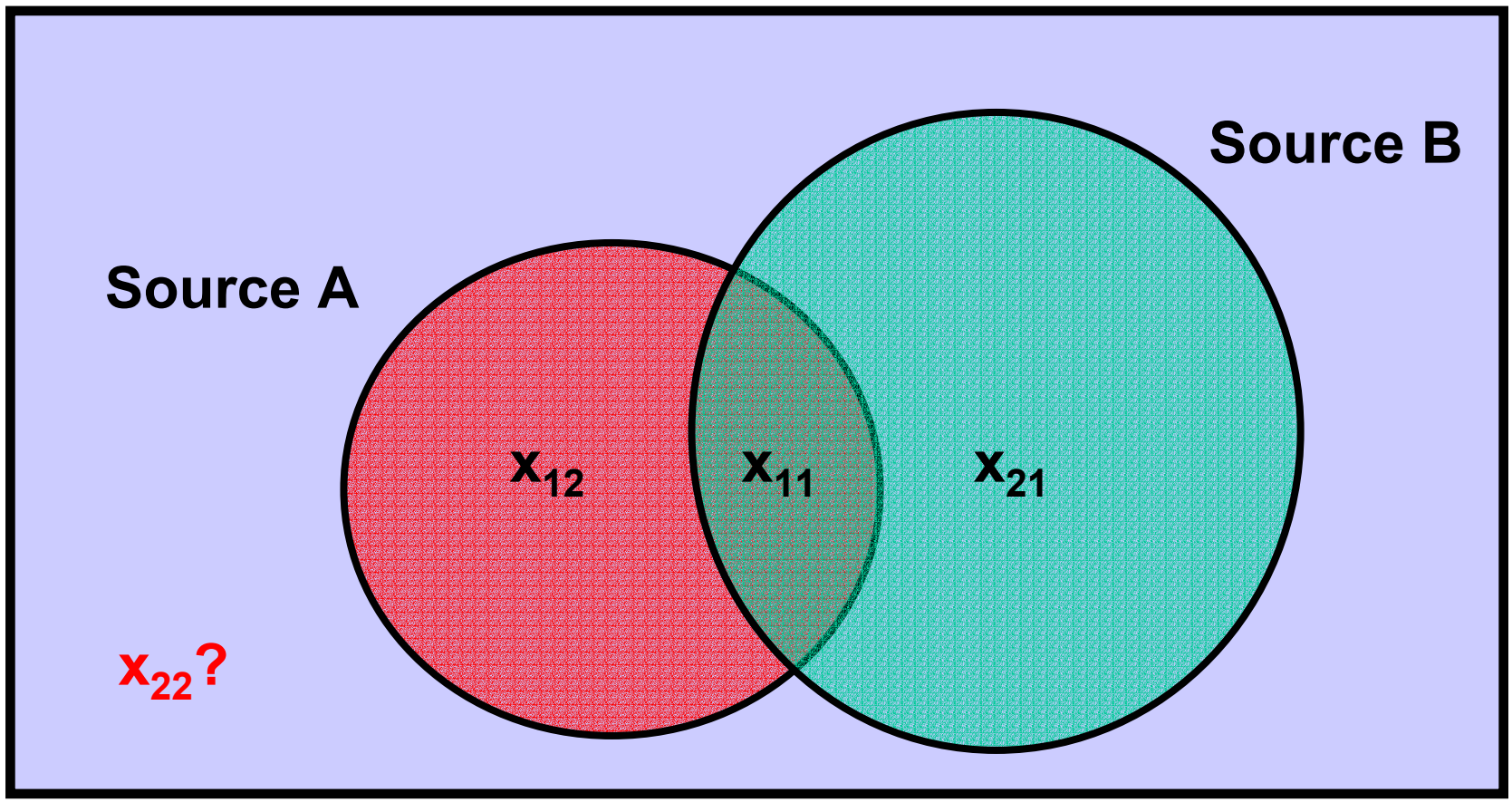
Estimer le nombre cas non répertoriés

Estimer le nombre total de cas

Evaluer l'exhaustivité de la source

Sources : Registres, déclarations obligatoires, les dossiers médicaux, certificats de décès,

II. Application de la méthode à l'épidémiologie



1 included in source
2 not included in source

II. Application de la méthode à l'épidémiologie

Deux méthodes:

- Croisement de 2 sources
- Croisement de plus de 2 sources

Chaque cas se caractérise par sa présence/ absence sur une liste
Table de contigence

Table 2x2: croisement de deux listes
Table 2x k: croisement de plus de deux listes

II. 1. Croisement de deux sources

		Source A		
		+	-	
Source B	+	X_{11}	X_{21}	N_2
	-	X_{12}	X_{22}	
		N_1		N

II. 1. Croisement de deux sources

1. Estimation

a) Estimation du nombre de cas non répertoriés

$$x_{22} = \frac{x_{12}\, x_{21}}{x_{11}}$$

		Source A		
		+	-	
Source B	+	X_{11}	X_{21}	N_2
	-	X_{12}	X_{22}	
		N_1		N

$$N = \frac{N_1\, N_2}{x_{11}}$$

$$Var_N = \frac{N_1\, N_2\, x_{12}\, x_{21}}{x_{11}^3}$$

$$95\%CI = N \pm 1.96 \sqrt{Var_N}$$

II. 1. Croisement de deux sources

b)Exhaustivité des sources:

$$\text{Sens A} = \frac{N_1}{N}$$

$$\text{Sens B} = \frac{N_2}{N}$$

3. Estimateurs de Chapman et Seber

Lorsque les effectifs sont faibles, la probabilité que $x_{11} = 0$ n 'est pas nulle:

$$N = \frac{(N_1 + 1)\,(N_2 + 1)}{x_{11} + 1}$$

$$\text{Var}_N = \frac{(N_1 + 1)(N_2 + 1)\; x_{12}\, x_{21}}{(x_{11}^2)(x_{11} + 2)}$$

II. 1. Croisement de deux sources

1. Estimations

2. Conditions d'application:

- Indépendance des sources
- Homogénéité de capture
- Appariement des cas
- Vrais cas
- Population close
- Cas identifiés survenu sur une période et zone identique

II. 1. Croisement de deux sources

2. Conditions d'application:

- **Indépendance des sources :**

Probabilité d'être recensé sur une liste ne dépend pas ou n'a pas d'incidence sur la probabilité de figurer sur une autre liste.

Dépendance positive : l'identification des cas par une source augmente la probabilité pour ces cas d'être identifiés par l'autre . Cela induit une sous estimation de N.

Dépendance négative: l'identification des cas par l'une diminue la probabilité pour ces cas d'être identifiés par l'autre. Il y a alors sur estimation de N.

II. 1. Croisement de deux sources

		Source A +	Source A -	
Source B	+	a	b	N_2
	-	c	d	
		N_1		N

$$OR = \frac{ad}{bc}$$

OR= 1 Indépendance entre les sources

OR > 1 (dépendance positive): → sous estimation de N

OR < 1 (dépendance négative): → sur estimation de N

II. 1. Croisement de deux sources

2. Conditions d'application:

- **Homogénéité de capture :**

Même probabilité pour tous les cas d'être identifiés par une même source.

L'identification des cas sur une liste ne peut être liée à certaines caractéristiques personnelles (âges, sexe, le niveau socio-économique, gravité de la maladie, etc.).

Quand hétérogénéité: Estimation dans chaque strate de la variable d 'hétérogénéité

II. 1. Croisement de deux sources

- Stratification pour la variable qui introduit l 'hétérogénéité
- Calcul des estimations dans chaque strate

Stratum 1

		Source A +	Source A -	
Source B	+	X_{11}	X_{21}	B_1
	-	X_{12}	X_{22}	
		A_1		N_1

$$N_1 = \frac{A_1 B_1}{x_{11}}$$

Stratum 2

		Source A +	Source A -	
Source B	+	X_{11}	X_{21}	B_2
	-	X_{12}	X_{22}	
		A_2		N_2

$$N_2 = \frac{A_2 B_2}{x_{11}}$$

$$N = \Sigma N_i = N_1 + N_2$$

II. 1. Croisement de deux sources

2. Conditions d'application:

- **appariement:**

Cas identifiés doivent êtres de vrais doublons ⇨ Identifiant commun entre les listes

- **Vrais cas:**

Cas identifiés sont des vrais cas ⇨ Même définition de ce qu 'est un cas entre les listes

II. 1. Croisement de deux sources

2. Conditions d'application:

- **Population close**

Pas de mouvement de population (pas de migration, pas de décès et pas de naissance) dans la zone géographique et la période étudiées

Si population ouverte: Probabilité de capture $\downarrow \rightarrow x_{11} \downarrow \rightarrow$

Surestimationde N

$$N = \frac{N_1 N_2}{x_{11}}$$

- **Cas identifiés sur période et zone identique**

Tous les cas identifiés sont survenus pendant la période et dans la zone géographique étudiées

II. 1. Croisement de deux sources

1. Estimations
2. Conditions d'application

3. Problèmes de l'approche à deux sources:

- L **'indépendance** entre sources est **rarement rencontrée**. La dépendance entre deux sources ne peut être évaluée statistiquement. Elle ne peut se faire que qualitativement en analysant le circuit de l'information et les modes de recrutement des deux sources.

- L '**homogénéité** de capture est également **rarement rencontrée**. la probabilité d 'être sur une liste dépend souvent de plusieurs facteurs (âge, sexe, lieu de résidence, gravité de la maladie, etc.)

II. 2. Croisement de plus de 2 sources

		Source A			
		included		not included	
		Source B		Source B	
		included	not included	included	not included
Source C	included	m111	m121	m211	m221
	not included	m112	m122	m212	m222

Comment estimer m222?

II. 2. Croisement de plus de 2 sources

Le recoupement de plusieurs sources permet :

- **calculer la dépendance** entre les différentes sources:

  Table 2 x 2, Chi-carré de dépendance et un OR (IC)

- **prendre en compte la dépendance** dans l'estimation du nombre total de cas. Deux méthodes:

  1. grouper les sources dépendantes en une source unique en additionnant le nombre total de cas identifié par chaque source moins les cas communs aux deux sources.

  2. L'utilisation des modèles log-linéaires

- **prendre en compte l'hétérogénéité** de capture .

II. 2. Croisement de plus de 2 sources

Log-linear modelling

- Modèle qui permet d 'analyser les tables de contingence multiple
- Logarithme néperien de la fréquence attendue d' une cellule du tableau comme une combinaison linéaire d'effets principaux (sources) et d'interaction entre sources.

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ij}^{AC} + \lambda_{ij}^{BC} + \lambda_{ijk}^{ABC}$$

λ_I^A, λ_j^B, λ_K^C= effets principaux (les différentes sources)

λ_{IJ}^{AB} ,λ_{IK}^{AC}, λ_{jk}^{BC}= effets d'interaction (prise en compte de la dépendance entre les sources.)

Possiblilité d'introduire dans le modèle les facteurs d'hétérogénéité (âge, sexe, la sévérité de la maladie, le statut socio-économique, etc.).

II. 2. Croisement de plus de 2 sources

- **1 Modèle sans interactions: sources sont independantes**

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

- **3 modèles avec une interaction**

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$$

- **3 modèles avec 2 interactions**

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ij}^{AC}$$

- **1 modèle avec toutes les interactions = modèle saturé**

$$LnFijk = \theta + \lambda_I^A + \lambda_j^B + \lambda_K^C + \lambda_{IJ}^{AB} + \lambda_{IK}^{AC} + \lambda_{jk}^{BC} + \lambda_{jk}^{BAC}$$

- **Le choix est basé sur la statistique de vraissemblance G^2**
- **Le modèle choisi sera celui qui a le moins d'interactions tout en ayant une bonne adéquation avec les données observées (G^2NS)**

Problèmes de l'application de la méthode en épidémiologie

- **Plusieurs sources distinctes sur un même pathologie**
- **Identifiant commun aux sources pour identifier les cas communs → Problème éthique**
- **Problèmes de dépendance et d 'hétérogénéité entre source qui nécéssite de disposer de plus de 2 sources**
- **Définition uniforme d'un cas entre les sources.**
- **Condition d 'une population close et stable jamais satisfaisante (Impact est variable selon le sujet d 'étude: usager de drogue versus méningocoques)**

III. Domaines d 'application de la méthode en épidémiologie

Domaines	Buts
• **Cancer**	l'exhaustivité de registre de cancer notamment aux Etats-Unis, Pays-Bas, etc.
• **Drogues**	Estimer et rectifier la prévalence de l'utilisation de drogues injectables et à estimer la taille de population de certains groupes d'utilisateurs.
• **Malformations Congénitales**	Améliorer le taux de déclaration
• **Accidents**	Réajuster des estimations
• **Maladies infectieuses**	Estimer soit la prévalence de cas soit l'efficacité des systèmes de déclaration, soit les deux.
• **Diabètes**	Vérifier le niveau de dénombrement et obtenir des estimations corrigées.

IV. Exemples:

- **Application de la méthode à deux sources:**

Surveillance des infections à méningocoque en France en 1989 et 1990

- **Application de la méthode à sources multiples**

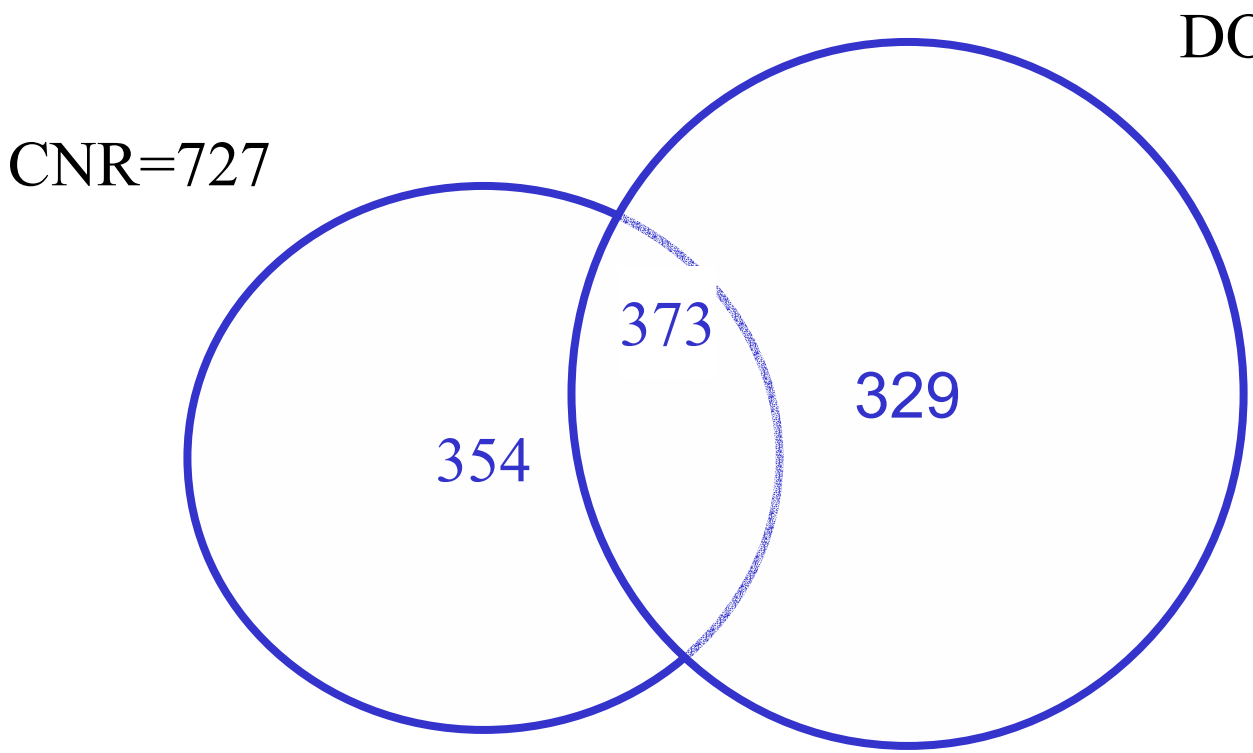
Evaluation du système de surveillance de la Légionellose en France

Surveillance des infections à méningocoque en France, 1989 - 1990

Sources:

- Fiches détaillées de déclaration obligatoire (DO) : 702 cas recensés
- Centre National de Référence(CNR) : 727
- La comparaison des deux sources: 373 cas communs

Surveillance des infections à méningocoque en France, 1989 - 1990



??????

Surveillance des infections à méningocoque en France, 1989 - 1990

Estimations:

1.Estimation du nombre de cas non répertoriés:

$$X_{22} = \frac{X_{12} X_{21}}{X_{11}}$$

(354 x 329)/373 = 312 cas qui ne sont sur aucunes des listes

2. Estimation du nombre total de cas = N

N= N1. N2/X11

(727 x 702)/ 373 = 1368 cas de méningocoques = nombre total

Surveillance des infections à méningocoque en France, 1989 - 1990

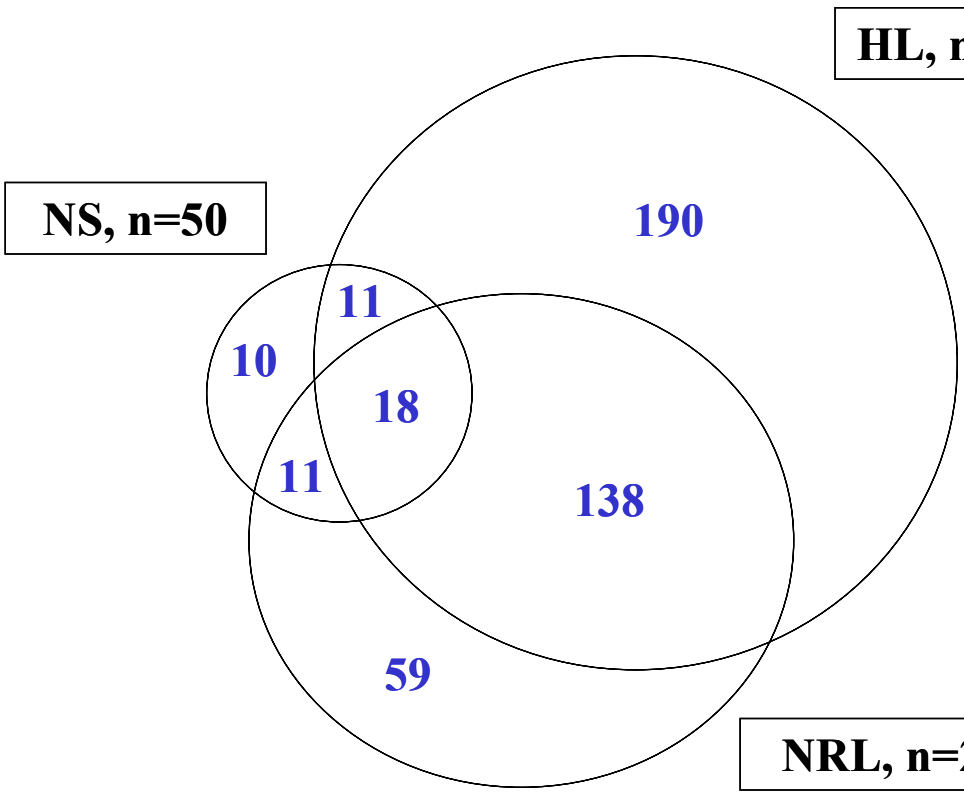
3. Taux d ’exhaustivité

DO: 702/1368 = 51% .

CNR: 727/1368 =53%

Des deux listes combinées:[(702+727)-373] / 1368 = 77%

Evaluation du système de surveillance de la Légionellose en France



NS: Système de déclaration obligatoire
NRL: Centre National de référence
HL: Enquête auprès des laboratoires hospitaliers

Evaluation du système de surveillance de la Légionellose en France

1. Calcul de l ’indépendance entre les sources:

- **Reconstruire les tables 2x2**
- **Calcul du OR de dépendance**
- **Combiner deux sources dépendantes en une seule**

Evaluation du système de surveillance de la Légionellose en France

Indépendance entre NS et NRL

	NS	
NRL	Présent	Absent
Présent	18	138
Absent	11	190

OR : 18 x 190/11 x 138 = 2.3 [1- 5.7] → Dépendance positive entre les deux sources

Evaluation du système de surveillance de la Légionellose en France

Indépendance entre NS et HL

	NS	
HL	Présent	Absent
Présent	18	138
Absent	11	59

OR : 18 x59/11 x 138 = 0.7 [0.3-1.8] → Pas de dépendance

Indépendance entre HL et NRL

OR : 18 x10/11 x 11 = 1.5[0.4-5.4] → Pas de dépendance

Evaluation du système de surveillance de la Légionellose en France

2. Estimation du nombre total de cas :

Deux possibilités :

- Regrouper les sources dépendantes en 1 source: approche à deux sources indépendantes
- Utiliser le modèle log –linéaire

NS + NRL --> 2 sources indépendantes : HL et la combinaison NS +NRL.

Le nombre total de cas répertorié par 2 sources

NS∪NRL / HL 528 [495–561] cas

Evaluation du système de surveillance de la Légionellose en France