L’emploi de méthodes de capture recapture pour l’estimation de la taille de populations humaines et animales

Louis-Paul Rivest
lpr@mat.ulaval.ca

Résumé

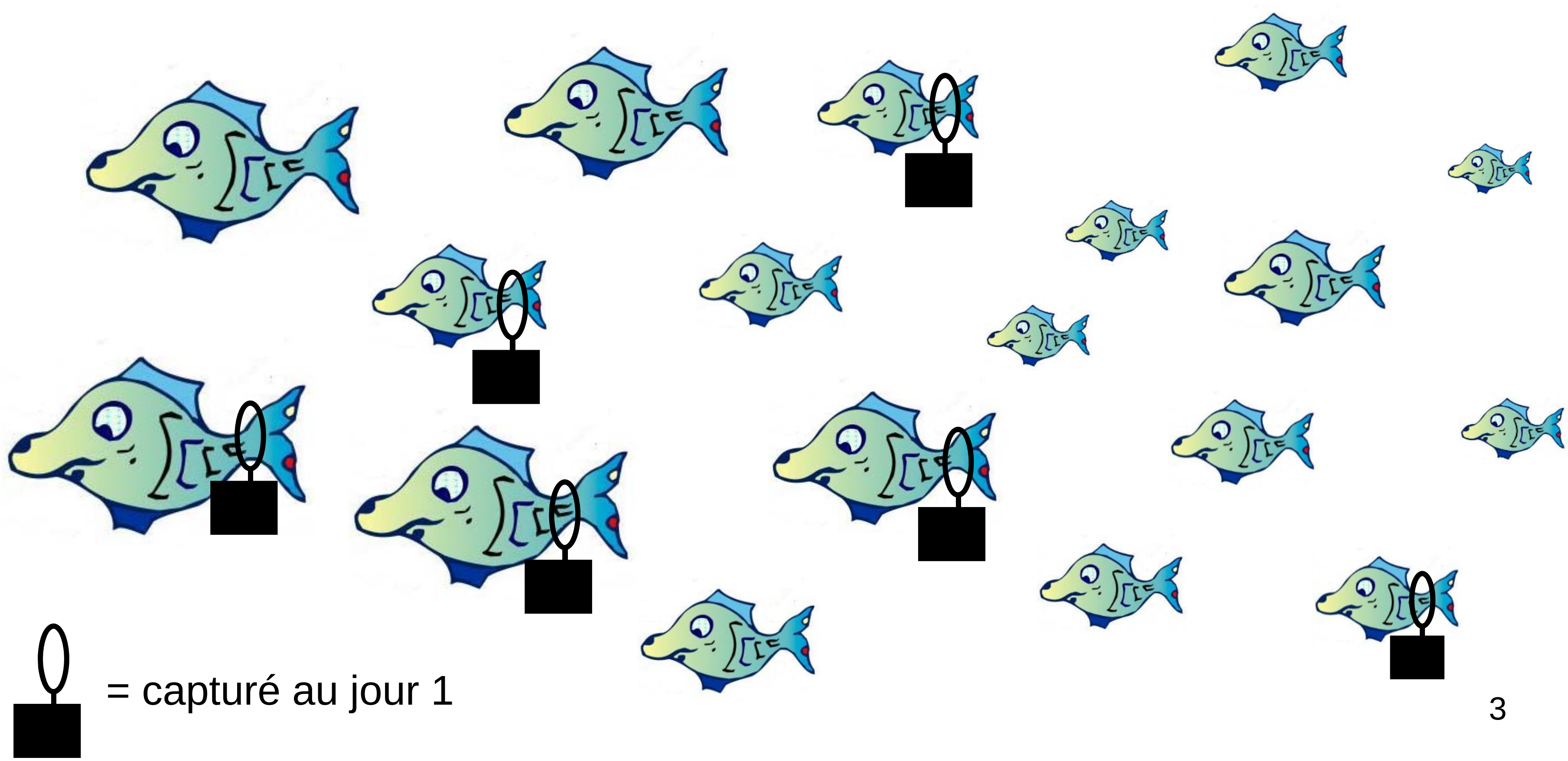
1- Populations fermées

- Les données
- Les modèles

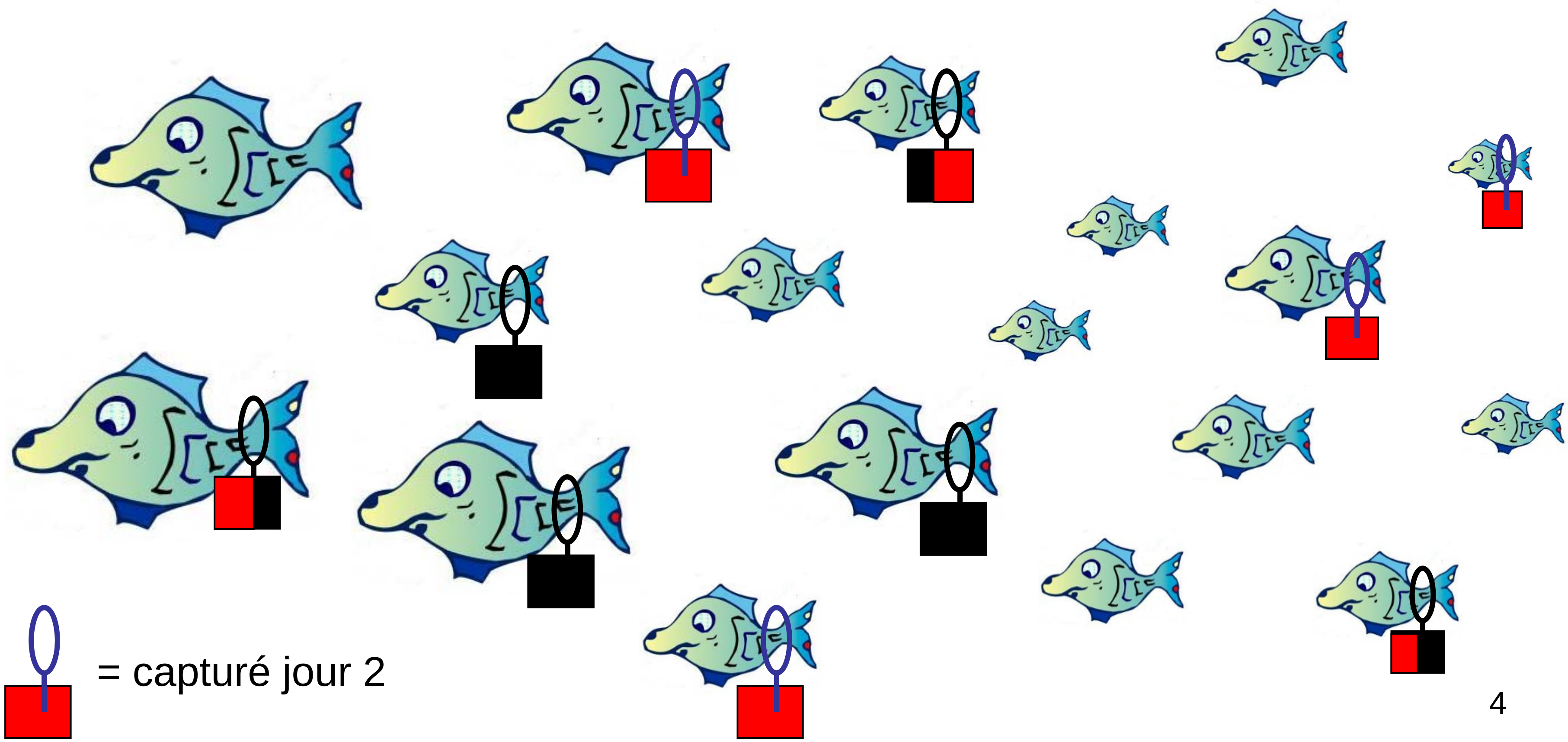
2- Populations ouvertes

- Données+modèles

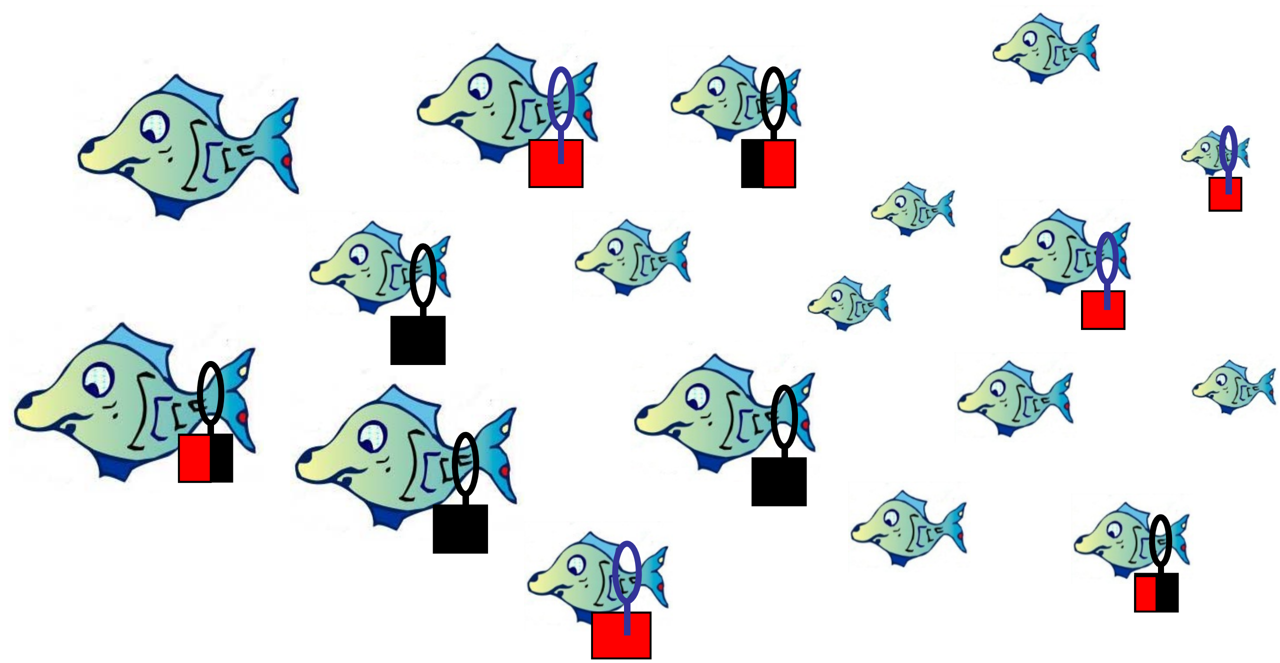
Exemple des poissons: au jour 1 on capture des poissons que l’on marque et que l’on relache



On fait de même au jour 2



Trois historiques de capture sont observables ω=(1,0), (0,1), (1,1)



n_{10}=# poissons capturés au jour 1 seulement =3
n_{01}=# poissons capturés au jour 2 seulement = 4
n_{11}=# poissons capturé les deux jours=3

Population de poissons

Les données sont

	manqué jour 2	capturé jour 2
manqué jour 1	n_{00}=??	4
capturé jour 1	3	3

Paramètre inconnu

N=Nombre total de poissons dans le lac=$n_{11}+n_{10}+n_{01}+n_{00}$

Enfants mendiants à Dakar

- **Recensement de tous les enfants mendiants à différents site de rassemblement, c'est la capture**
- **La recapture se fait à l'aide d'incitatifs qui amènent les enfants à contacter les autorités (**sondage déterminé selon les répondants)
- **Ref: MauriziaTovo_WorldBank**

Enfants mendiants à Dakar

	Capture I	Capture II	Recapturé
Dakar	517	578	92
Rufisque	223	304	100
Guédiawaye	238	363	53
Pikine	84	370	46
REGION DE DAKAR	1 062	1 615	291

Enfants mendiants à Dakar

	Capture I	Capture II	Recapturé
Dakar	517	578	92

	manqué SDR	capturé SDR
manqué recen	n_{00}=??	486
capturé recen	425	92

Combien de personnes présentes dans les tours ont survécu le 9 sept 2001?

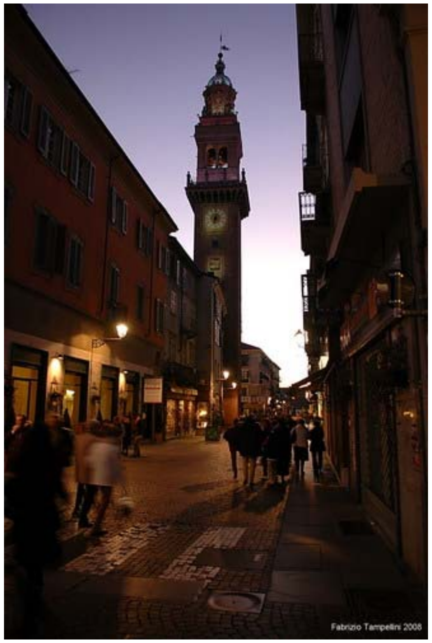
Murphy, J. (2009): "Estimating the world trade center tower population on september 11, 2001: A capture-recapture approach," *American Journal of Public Health*, 99, 65–67.

On dispose de trois listes partielles de personnes présentes dans les tours

- L1= Auto identification
- L2 = Liste d'affaires
- L3 = Liste du Port authority

Capture Hist.			n_ω
1	1	1	174
1	1	0	88
1	0	1	1658
1	0	0	1702
0	1	1	750
0	1	0	270
0	0	1	4323

n= 8 965 survivants identifés



Estimation du nombre de cas de diabètes à Casale Monferrato, Italy

Bruno, G., A. Biggeri, R. E. LaPorte, D. McCarty, F. Merletti, and G. Pagano (1994): "Appplication of capture-recapture to count diabetes?" *Diabetes Care*, 17, 548–556.

4 Listes:

- E1=Clinique pour le diabète
- E2=Rapport d'hopitaux
- E3 = Médicaments
- E4=Seringues à insuline.

		Total			
		E_1=1		E_1=0	
		E_2=1	E_2=0	E_2=1	E_2=0
E_3=1	E_4=1	58	46	14	8
E_3=1	E_4=0	157	650	20	182
E_3=0	E_4=1	18	12	7	10
E_3=0	E_4=0	104	709	74	NA

n= 2 069

cas connus

Mortalité au Kosovo au printemps 1999

Ball, P., W. Betts, F.J. Scheuren, J. Dudukovich, and J. Asher, *Killings and Refuge Flow in Kosovo, March – June 1999*, American Academy for the Advancement of Science, January 3, 2002.

Quatre listes:

- Association américaine du bareau
- Exhumation
- Human Rights Watch
- Organisation pour la sécurité et la coopération en Europe

	ABA	Yes	Yes	No	No	
	EXH	Yes	No	Yes	No	
HRW	OSCE					Total
Yes	Yes	27	32	42	123	
Yes	No	18	31	106	306	
No	Yes	181	217	228	936	
No	No	177	845	1,131	??	
Total						4,400

n= 4400 morts répertoriés

Modèle pour 2 captures

N=taille de la population

Indépendance des captures aux deux occasions:

p_1=Prob (capture à l'occasion 1)

p_2=Prob (capture à l'occasion 2)

Valeurs prédites	manqué jour 2	capturé jour 2
manqué jour 1	$\mu_{00}=N(1-p_1)(1-p_2)$	$\mu_{01}=N(1-p_1)p_2$
capturé jour 1	$\mu_{10}=Np_1(1-p_2)$	$\mu_{11}=Np_1p_2$

Note

$$\text{Rapport de cotes} = \theta = \frac{\mu_{00} \times \mu_{11}}{\mu_{10} \times \mu_{01}} = 1$$

Modèle pour 2 captures

N=taille de la population

Indépendance des captures aux deux occasions:

p_1=Prob (capture à l'occasion 1)

p_2=Prob (capture à l'occasion 2)

Modèle et estimateur de Petersen

Valeurs prédites	manqué jour 2	capturé jour 2
manqué jour 1	$\mu_{00}=N(1-p_1)(1-p_2)$	$\mu_{01}=N(1-p_1)p_2$
capturé jour 1	$\mu_{10}=Np_1(1-p_2)$	$\mu_{11}=Np_1p_2$

Note $\mu_{00}=\dfrac{\mu_{10}\times\mu_{01}}{\mu_{11}}$ et $\hat{n}_{00}=\dfrac{n_{10}\times n_{01}}{n_{11}}$

Modèle de Petersen à deux occasions de capture

- Estimateur de N

$$\hat{N} = \boxed{\overset{n\,=}{n_{11} + n_{01} + n_{10}}} + \boxed{\overset{\hat{n}_{00}\,=}{\frac{n_{10} \times n_{01}}{n_{11}}}} = (n_{11} + n_{10}) \div \frac{n_{11}}{n_{11} + n_{01}}$$

Nombre de poissons capturés à l'occasion 1

Proportion de poissons marqués parmi ceux capturés à l'occasion 2

Variance de $\hat{N}$?

Résultat général qui fonctionne pour la plupart des modèles à 2 et 3 occasions:

$$v(\hat{N}) = \hat{n}_{00} + \hat{n}_{00}^2 \left[\sum_{n_\omega \text{ dans } \hat{n}_{00}} \frac{1}{n_\omega} \right]$$

Pour le modèle de Petersen:

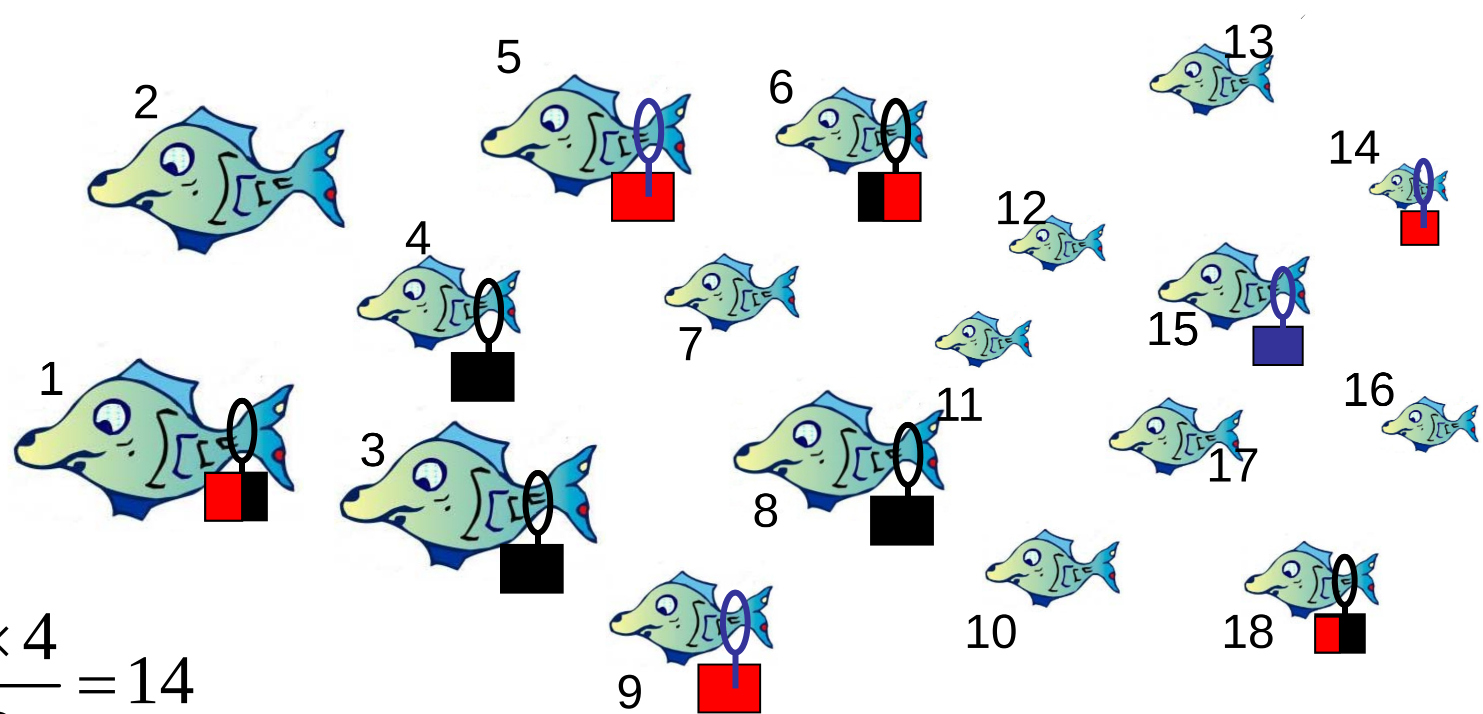
$$v(\hat{N}) = \frac{n_{10} \times n_{01}}{n_{11}} + \left[\frac{n_{10} \times n_{01}}{n_{11}} \right]^2 \left[\frac{1}{n_{10}} + \frac{1}{n_{11}} + \frac{1}{n_{01}} \right]$$

$$= \frac{(n_{11} + n_{10}) \times (n_{11} + n_{01}) \times n_{01} \times n_{10}}{n_{11}^3}$$

$$= \hat{N} \frac{n_{01} \times n_{10}}{n_{11}^2}$$

$$IC \text{ à } 95\% : \hat{N} \pm 2\sqrt{v(\hat{N})}$$

Premier exemple



Données

$n_{11}=3$

$n_{01}=4$

$n_{10}=3$

$$\hat{N} = 10 + \frac{3\times 4}{3} = 14$$

$$v(\hat{N}) = 14 \times \frac{3\times 4}{3^2} = 4.3^2$$

10

IC à 95%: $[\cancel{8},22]$

Enfants mendiants à Dakar

	Capture I	Capture II	Recapturé	Population estimée	Intervalle de confiance à 95%	
Dakar	517	578	92	**3 224**	2 683	3 765
Rufisque	223	304	100	**675**	596	755
Guédiawaye	238	363	53	**1 610**	1 264	1 956
Pikine	84	370	46	**670**	596	755
REGION DE DAKAR	1 062	1 615	291	**5 882**	5 363	6 401

Enfants mendiants à Dakar

District de Dakar	manqué SDR	capturé SDR
manqué recen	n_{00}=??	486
capturé recen	425	92

$$\hat{N} = 1003 + \frac{486 \times 425}{92} = 3248$$

$$v(\hat{N}) = 3248 \times \frac{486 \times 425}{92^2} = 278.5^2$$

IC à 95%: [2691,3805]

Validité: attention à l'hétérogénéité

Si les petits poissons sont plus difficiles à capturer que les gros il faut faire des analyses séparées pour obtenir un estimateur de N non biaisé

Petit	Raté J 2	Pris J 2	Gros	Raté J 2	Pris J 2
Raté J 1	n_{00}=??	11	Raté J 1	n_{00}=??	7
Pris J 1	14	6	Pris J 1	5	37

$$\hat{N}_{petit} = 31 + \frac{14 \times 11}{6} \approx 57 \qquad \hat{N}_{gros} = 49 + \frac{5 \times 7}{37} \approx 50$$

$$\hat{N}_{gros} + \hat{N}_{petit} = 107$$

Validité: attention à l'hétérogénéité

Si on combine les petits et les gros poissons avant de calculer N on trouve

	Raté J2	Pris J2
Raté J1	n_{00}=??	18
Pris J 1	19	43

$$\hat{N}_{combi} = 43 + 19 + 18 + \frac{19 \times 18}{43}$$

$$\approx 88$$

L'estimateur combiné est inférieur de 20% à la somme des estimateurs pour les petits et les gros poissons

L'estimateur combiné est biaisé! Il sous-estime la taille de la population

Modèle à trois occasions de capture

Pris J1	Raté J3	Pris J3	Raté J1	Raté J3	Pris J3
Raté J 2	n_{100}	n_{101}	Raté J 2	n_{000}=??	n_{001}
Pris J 2	n_{110}	n_{111}	Pris J 2	n_{010}	n_{011}

On estime le rapport de cotes

$$\hat{\theta}_1 = \frac{n_{100} \times n_{111}}{n_{101} \times n_{110}}$$

On l'utilise pour estimer n_{000}

$$\hat{n}_{000} = \hat{\theta}_1 \times \frac{n_{010} \times n_{001}}{n_{011}}$$

Modèle à trois occasions de capture

L’estimateur le plus général de N est

$$\hat{N} = n + \hat{n}_{000} = n + \frac{n_{100} \times n_{111} \times n_{010} \times n_{001}}{n_{110} \times n_{101} \times n_{011}}$$

avec

$$v(\hat{N}) = \hat{n}_{000} + \hat{n}_{000}^{2}\left[\sum \frac{1}{n_{\omega}}\right].$$

C’est l’estimateur à utiliser lorsque tous les $\hat{\theta}_j$ sont différents de 1

Modèle à trois occasions de capture

Si le $\hat{\theta}_j$ pour les unités capturées à l'occasion j vaut 1, on peut simplifier l'estimeur précédent. Si $\hat{\theta}_2 \approx 1$ on peut prendre

$$\hat{N} = n + \hat{n}_{000} = n + n_{100} \times \frac{n_{111} \times n_{010} \times n_{001}}{n_{110} \times n_{011} \times n_{101}}$$

vaut 1

$$\hat{N} = n + \hat{n}_{000} = n + \frac{n_{100} \times n_{001}}{n_{101}}$$

Modèle à trois occasions de capture

Si deux des trois $\hat{\theta}_j$ valent 1, on combine les deux variables correspondantes et on calcule l'estimateur de Petersen

Si $\hat{\theta}_1 \approx \hat{\theta}_2 \approx 1$ on calcule des fréquences *m* en combinant les deux premières occasions,

$m_{10}=n_{100}+n_{010}+ n_{110}$, $m_{01}=n_{001}$

$m_{11}=n_{101}+n_{011}+ n_{111}$

et on calcule l'estimateur de Petersen sur les m_{ω}.

Modèle à trois occasions de capture

Si $\hat{\theta}_1 \approx \hat{\theta}_2 \approx 1$ on prend

$$\hat{N} = n + \hat{m}_{00} = n + \frac{m_{10} \times m_{01}}{m_{11}}$$

$$\hat{N} = n + \hat{n}_{000} = n + \frac{\boxed{(n_{100} + n_{010} + n_{110})} \times n_{001}}{\boxed{n_{111} + n_{101} + n_{011}}}$$

Modèle à trois occasions de capture

Si les trois $\hat{\theta}_j$ valent 1, les trois occasions de capture sont indépendantes. L’estimateur de N n’a pas une forme explicite dans ce cas.

Inférence sur les rapports de cotes: θ_1=1?

On rejette l'hypothèse H_0: θ_1=1 au seuil 5% si

$$\left| \frac{\log(\hat{\theta}_1)}{\sqrt{\boxed{1/n_{111} + 1/n_{100} + 1/n_{101} + 1/n_{110}}}} \right| > 2$$

Estimation de la variance du log-rapport de cote: $= v\left\{\log(\hat{\theta}_1)\right\}$

Capture Hist.			n_ω
1	1	1	174
1	1	0	88
1	0	1	1658
1	0	0	1702
0	1	1	750
0	1	0	270
0	0	1	4323

Tot= 8 965

$$\hat{\theta}_1 = \frac{174 \times 1702}{88 \times 1658} = 2.03 = \exp(.71)$$

$$v(\hat{\theta}_1) = \frac{1}{174} + \frac{1}{1702} + \frac{1}{88} + \frac{1}{1658} = .14^2$$

$$\hat{\theta}_2 = 0.71 \; (\log \hat{\theta}_2 = -.34) \quad v(\hat{\theta}_1) = .15^2$$

$$\hat{\theta}_3 = 0.60 \; (\log \hat{\theta}_2 = -.50) \quad v(\hat{\theta}_1) = .09^2$$

Interprétation: Les listes 2 et 3 (Affaires et Port Authority) concernent des gens qui travaillent dans les tours et y vont sur une base régulière. Ces deux listes sont associées positivement.

La première liste (auto-identification par téléphone) est peut-être associée aux visiteurs occasionnels. Les gens qui travaillent régulièrement dans les tours ont eu moins tendance à s’identifier sur cette liste (dépendance négative).

Capture Hist.			n_ω
1	1	1	174
1	1	0	88
1	0	1	1658
1	0	0	1702
0	1	1	750
0	1	0	270
0	0	1	4323

Tot= 8 965

Hypothèse: Tous les θ_j sont différents de 1

3159

$$\hat{N} = n + \hat{n}_{000} = 8965 + \boxed{\frac{174 \times 1702 \times 270 \times 4323}{88 \times 1658 \times 750}} = 12124$$

$$v(\hat{N}) = 3159 + 3159^2 \left(\frac{1}{174} + \ldots + \frac{1}{4323} \right) = 488^2$$

$$IC = (11147,\ 13100)$$

Capture Hist.			n_ω
1	1	1	174
1	1	0	88
1	0	1	1658
1	0	0	1702
0	1	1	750
0	1	0	270
0	0	1	4323

Tot= 8 965

Hypothèse: $\theta_2 = 1$
(un modèle plus simple donne une estimation plus précise)

4438

$$\hat{N} = n + \hat{n}_{000} = 8965 + \boxed{\frac{1702 \times 4323}{1658}} = 13403$$

$$v(\hat{N}) = 4338 + 4338^2 \left(\frac{1}{1702} + \frac{1}{1658} + \frac{1}{4323} \right) = 180^2$$

$$IC = (13402,\ 13762)$$

Deux IC disjoints!

Cas général

- On écrit un modèle comme une somme d’effets simple et d’interactions entre les occasions de capture ci.
- Avec 3 occasions le modèle le plus général s’écrit

  c1+c2+c3+c1*c2+c1*c3+c2*c3
- Un rapport de cotes porte sur une interaction particulière, si θ_1=1 il n’y a pas de lien entre les occasions 2 et 3. On peut enlever c2*c3:

  c1+c2+c3+c1*c2+c1*c3
- Le modèle de Petersen s’écrit

  c1+c2

Résumé

1- Populations fermées

- Les données
- Les modèles

2- Populations ouvertes

- **Les données et les modèles**

Populations ouvertes

Une population ouverte est observée sur une plus longue période.

Les occasions de capture sont espacées dans les temps

La population change entre deux occasions: il y a des naissances et des morts

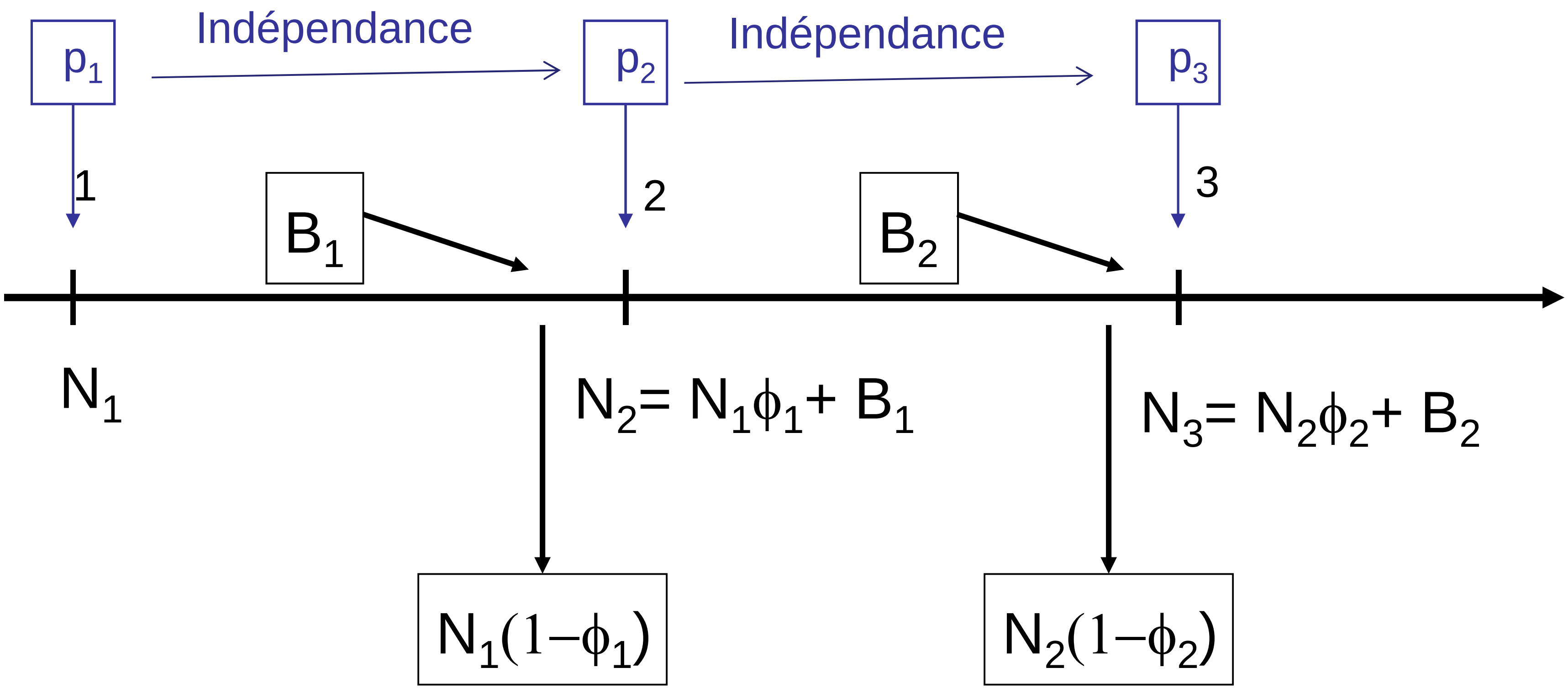
Populations ouvertes : Exemples

- Une population animale où les occasions de capture sont des années. Un animal peut-être observé à une anné donnée s’il est présent sur le site de reproduction de l’espèce
- Les clients d’un magasin (ou d’un site internet): naissance= arrivée d’un nouveau client, mort=perte d’un client. Un achat enregistré est une capture.

Populations ouvertes

Population — Echantillonnage

B=# Naissances — ϕ=taux de survie



Populations ouvertes : 3 occasions de capture

Les données sont les fréquences n_ω pour les 7 occasions de capture observables.

Les paramètres sont

- les tailles N_1, N_2, N_3
- les taux de survie entre 2 occasions ϕ_1, ϕ_2
- Les probabilités de capture p_1, p_2, p_3
- Les naissances entre 2 occasions B_1, B_2

Populations ouvertes : historique de capture

L’historique de capture (0,1,0) d’un animal vu au temps 2 est associé aux cas suivants:

- L’animal était dans la population aux temps 1,2 et 3. Il n’a été capturé qu’au temps 2
- L’animal est né entre les temps 1 et 2, a été capturé au temps 2 et est mort entre les temps 2 et 3
- …..

3 ans de données sur des alligators

Histor.	n_ω
1 1 1	4
1 1 0	5
1 0 1	1
1 0 0	7
0 1 1	18
0 1 0	22
0 0 1	14

Populations ouvertes

- Puisque la probabilité d'être capturé à l'occasion 3 sachant qu'on est capturé en 2 ne dépend pas de la capture en 1 on a:

$$\theta_2 = \frac{\mu_{111} \times \mu_{010}}{\mu_{110} \times \mu_{011}} = 1$$

- Les 6 paramètres estimables sont: $N_1 p_1$, $\boxed{\phi_1, N_2,}$ $p_{2,}$ $\phi_2 p_3$, et B_2/N_3

Estimation de N_2

On utilise l'estimateur de Petersen:

- liste 1= animaux capturés aux temps 1 et 3
- liste 2= population au temps 2

Estimation de N_2

Estimation de N_2	manqué liste 2	capturé liste 2
manqué liste 1	??	$n_{010} + n_{110} + n_{011}$
capturé liste 1	n_{101}	n_{111}

$$\hat{N}_2 = n_{101} + \sum n_{i1j} + \frac{n_{101} \times (n_{010} + n_{011} + n_{110})}{n_{111}}$$

Captures au temps 2

$$= \boxed{\sum n_{i1j}} \div \boxed{\frac{n_{111}}{n_{111} + n_{101}}}$$

Proportion des capturés au temps 2 parmi les animaux capturés aux temps 1 et 3.

Estimation de N_2

$$\hat{N}_2 = n_{101} + \sum n_{i1j} + \boxed{\frac{n_{101} \times (n_{010} + n_{011} + n_{110})}{n_{111}} = \hat{n}_{00}}$$

$$v(\hat{N}_2) = \hat{n}_{00} + \hat{n}_{00}^2 \left(\frac{1}{n_{101}} + \frac{1}{n_{111}} + \frac{1}{n_{010} + n_{110} + n_{011}} \right)$$

Exemple: alligators

$$\hat{N}_2 = 49 + \frac{1 \times 45}{4} = 61.2$$

$$v(\hat{N}_2) = 11.2 + 11.2^2 \left(1 + \frac{1}{4} + \frac{1}{45}\right) = 13.12^2$$

$$\hat{p}_2 = \frac{4}{5} = .8 \qquad v(\hat{p}_2) = .18^2$$

Estimation de ϕ_1

Parmi les $n_1 = \sum n_{1ij}$ personnes capturées au temps 1, combien ont survécu jusqu'au temps 2?

Liste 1= capturé au temps 2 et liste 2=capturé au temps 3

$$\hat{S}_1 = n_{111} + n_{110} + n_{101} + \frac{n_{110} \times n_{101}}{n_{111}}$$

$$\hat{\phi}_1 = \left(n_{111} + n_{110} + n_{101} + \frac{n_{110} \times n_{101}}{n_{111}} \right) / \, n_1$$

Populations ouvertes: ϕ_1

$$\hat{\phi}_1 = \left(n_{111} + n_{110} + n_{101} + \frac{n_{110} \times n_{101}}{n_{111}} \right) / n_1$$

Attention il est possible d'avoir $\hat{\phi}_1 > 1$

$$v(\hat{\phi}_1) = \frac{1}{n_1^2} \left\{ \frac{n_{110} \times n_{101}}{n_{111}} + \left(\frac{n_{110} \times n_{101}}{n_{111}} \right)^2 \left(\frac{1}{n_{110}} + \frac{1}{n_{101}} + \frac{1}{n_{111}} \right) \right\}$$

$$+ \frac{\hat{\phi}_1 (1 - \hat{\phi}_1)}{n_1}$$

Alligators: ϕ_1

$$\hat{\phi}_1 = \left(10 + \frac{5 \times 1}{4}\right) / 17 = 0.66$$

$$v(\hat{\phi}_1) = \frac{1}{17^2}\left\{1.25 + 1.25^2\left(\frac{1}{5} + 1 + \frac{1}{4}\right)\right\} + \frac{.66 \times .34}{17}$$

$$= 0.16^2$$

Discussion: Optimalité

- Les estimateurs de N_2 et ϕ_1 obtenus en appliquant le modèle de Petersen font-ils le meilleur usage de l'information disponible?

- Puisque θ_2=1 $\frac{\mu_{110}}{\mu_{111}} = \frac{\mu_{010}}{\mu_{011}} = \frac{\mu_{110}+\mu_{010}}{\mu_{111}+\mu_{011}}$ et

  $\frac{n_{110}+n_{010}}{n_{111}+n_{011}}$ est meilleur que $\frac{n_{110}}{n_{111}}$ pour estimer $\frac{\mu_{110}}{\mu_{111}}$

- Un meilleur estimateur de la probabilité de survie est

$$\hat{\hat{\phi}}_1 = \left(n_{111} + n_{110} + n_{101} + \frac{(n_{110}+n_{010}) \times n_{101}}{n_{111}+n_{011}} \right) / \, n_1$$

Discussion: Optimalité

- On a

$$\hat{N}_2 = n_2 \left[1 + \frac{n_{101}}{n_{111}} \right]$$

- Un estimateur plus précis est

$$\hat{\hat{N}}_2 = n_2 \left[1 + n_{101} \times \frac{n_2}{(n_{111} + n_{110}) \times (n_{111} + n_{011})} \right]$$

- avec $n_2 = (n_{010} + n_{011} + n_{110} + n_{111})$ égal au nombre d'animaux capturés à l'occasion 2

Discussion: Optimalité

- La statistique mathématique s’intéresse à la construction d’estimateurs optimaux qui utilisent, au mieux, l’information disponible;

- Cette théorie a vu le jour au XX^{ième} siècle et fait, encore aujourd’hui, l’objet de nombreux travaux.

Discussion: Probabilité

- Distribution multinomiale (la loi conjointe des n_ω)
- Distribution de Poisson (il est parfois utile de considérer des variables aléatoires multinomiales comme des variables Poisson indépendantes)
- Les estimations de variance s’appuient des expansions en séries de Taylor d’ordre 1

$$\hat{N} - N = N\left\{g(\hat{p}_\omega) - g(p_\omega)\right\}$$

$$\approx N \frac{\partial g(p_\omega)}{\partial p_\omega}^T \left\{\hat{p}_\omega - p_\omega\right\}$$

$$\text{Avec } \hat{p}_\omega = \frac{n_\omega}{N} \text{ et } p_\omega = \frac{\mu_\omega}{N}.$$

Discussion: Mise en oeuvre

Les biologistes ou les épidémiologistes qui veulent réaliser des analyses de capture-recapture n'ont pas à connaître les mathématiques sous-jacentes à ces dérivations.

Il suffit d'utiliser un logiciel qui ajuste ces modèles. Le gratuiciel R permet la diffusion de logiciels spécialisés permettant d'ajuster de tels modèles.

Est-ce tout à fait exact?

Discussion: Mise en oeuvre

Journal of Statistical Software

April 2007, Volume 19, Issue 5. *http://www.jstatsoft.org/*

Rcapture: Loglinear Models for Capture-Recapture in R

Sophie Baillargeon
Université Laval, Québec

Louis-Paul Rivest
Université Laval, Québec

Merci de votre attention!