LIMITES D'UTILISATION DES ESTIMATEURS PROPORTIONS EN HALIEUTIQUE

Olivier Schaan, Didier Gascuel[a]

I- INTRODUCTION

L'obtention de données en halieutique impose généralement d'avoir recours aux techniques d'échantillonnage et les paramètres à estimer résultent fréquemment de combinaisons de variables intermédiaires. Aussi les précisions finales sont couramment quantifiées par le calcul de variances d'estimation à partir de formules usuelles. En outre, les procédures d'échantillonnage mises en œuvre font souvent appel à des estimations de proportions. Ainsi, par exemple, un nombre de pêcheurs peut être estimé en utilisant un estimateur proportion du nombre de pêcheurs par engin ou par secteur de pêche. Un autre exemple courant d'application de cet estimateur réside dans l'élaboration de clés taille/âge, qui indiquent la proportion de chacun des groupes d'âge au sein de chaque classe de taille.

Lorsqu'on se réfère aux équations usuellement mises en œuvre, l'estimation de la variance tend vers zéro quand la proportion p estimée par échantillonnage tend vers zéro ou un. En outre, les conditions d'approximation normale de la distibution d'échantillonnage d'une proportion ne sont pas clairement définies et diffèrent selon les auteurs : les proportions ne doivent pas être "trop proches de zéro ou un" et la taille de l'échantillon n doit être supérieur à 30 (Cochran 1977, Scherrer 1983) ou à 100 (Wonnacott et Wonnacott 1991).

Cette note vise à fixer, pour un risque d'erreur donné, des limites claires d'utilisation des formules usuelles de variances, à proposer hors de ces limites des variances corrigées et non nulles.

II- RECHERCHE D'UNE MESURE D'INCERTITUDE NON NULLE POUR P=0 OU P=1

Pour une population statistique finie de taille N, l'espérance d'une proportion p observée dans un échantillon de taille n est égale à la proportion P issue de la

[a] - ENSAR halieutique - 65 route de St-Brieuc- 35042 RENNES Cédex

population (Kendall et Stuart, 1958). À l'estimateur de P peut être associé une variance usuellement estimée pour un tirage sans remise (loi hypergéométrique) par (Cochran, 1977 ; Grosbras, 1987) :

$$\hat{Var}_e(\hat{p}) = \frac{(N-n)}{(N-1)} \times \frac{p(1-p)}{n-1}, \qquad (1)$$

Ainsi, la formule usuelle conduit à une estimation nulle lorsque p = 0 ou 1. De toute évidence, une telle estimation n'est pas valide. Elle impliquerait que l'estimation soit égale à la valeur vraie, sans aucune incertitude, dès lors que la proportion observée au sein de l'échantillon est zéro ou un.

Le paramètre étudié peut aussi être estimé par intervalle (Scherrer, 1984). Cette démarche permet de fixer, pour un risque donné α, un intervalle de confiance autour de l'estimation. Pour des petits échantillons, la distribution d'échantillonnage suit une loi binomiale (avec remise) ou hypergéométrique (sans remise). Dans le cas d'une loi binomiale, les bornes supérieures (ps) et inférieures (pi) de l'intervalle d'estimation, à divers seuils de certitude α, peuvent être obtenues à l'aide du test binomial exact (logiciel S-Plus). Connaissant l'effectif de l'échantillon n et le nombre a d'observations présentant la caractéristique étudiée, le logiciel calcule les bornes pi et ps. La largeur de l'intervalle de confiance correspondant, pour un seuil α donné, est :

$$It_\alpha = p_s - p_i \qquad (2)$$

Pour passer d'une variable binomiale à une variable hypergéométrique, il faut corriger l'équation 2 par un coefficient d'exhaustivité (Abboud et Audroing, 1989), soit :

$$It_\alpha = p_{sc} - p_{ic} \text{ avec } \begin{cases} p_{ic} = p - \left((p - p_i) \times \sqrt{\frac{N-n}{N-1}} \right) \\ p_{sc} = p + \left((p_s - p) \times \sqrt{\frac{N-n}{N-1}} \right) \end{cases} \qquad (3)$$

Cette quantification de l'incertitude par un intervalle de confiance ne peut pas être combiner à d'autres incertitudes pour aboutir à une incertitude finale. Autrement dit, la prise en compte conjointe de plusieurs sources d'incertitudes (i.e. combinaison de plusieurs variables) impose *a priori* de revenir à une

expression analytique des variances d'estimation. C'est pourquoi, une variance conventionnelle, Var_c, déduite de la largeur d'intervalle est construite comme suit :

$$It_\alpha = 2 \cdot t_{\alpha/2} \cdot \sqrt{Var_{c(\alpha)}(\hat{p})} \qquad \text{où t est le t de Student soit :}$$

$$Var_{c(\alpha)}(\hat{p}) = \left(\frac{p_{sc} - p_{ic}}{2 \cdot t_{\alpha/2}} \right)^2 \qquad (4)$$

Cette équation permet notamment d'obtenir une mesure non nulle de l'incertitude associée à l'estimation d'une proportion nulle ou égale à un.

III- LIMITES DE VALIDATION DE L'ESTIMATEUR VARIANCE USUEL

Posons le rapport R_α des variances estimée et conventionnelle :

$$R_\alpha = \frac{\sqrt{Var_e(\hat{p})}}{\sqrt{Var_{c(\alpha)}(\hat{p})}} \qquad (5)$$

Déterminer si le rapport R_α diffère ou non de 1, revient à étudier si l'estimateur usuel fournit une mesure acceptable de l'incertitude. Aussi, les coefficients R_α sont calculés pour diverses tailles d'échantillon n et proportions p. On en déduit par des méthodes graphiques des isoplètes du rapport R_α dans le plan n x p (figure 1). Celles-ci permettent de déterminer pour chaque valeur de n, quels sont les valeurs de p qui conduisent à accepter les variances Var_e comme mesure de l'incertitude sur l'estimation. Cette démarche peut être conduite en acceptant des écarts entre variance estimée et conventionnelle plus ou moins importants. La figure 1 présente les résultats obtenus pour un risque $\alpha = 0,05$ avec $0,95 < R_\alpha < 1,05$ et pour $\alpha = 0,20$ avec $0,8 < R_\alpha < 1,25$.

Un autre problème mérite attention. Pour une proportion p s'éloignant de 1/2, l'asymétrie de l'intervalle de confiance s'accroit. Bien que n'affectant pas le poids de l'incertitude globale, elle rentre en ligne de compte dans l'estimation des bornes de l'intervalle. Aussi, un indice de symétrie S correspondant au rapport des deux demi-intervalles est défini par : $S_\alpha = \frac{p_{sc} - p}{p - p_{ic}}$ (6)

Comme pour le rapport R_α, S_α est calculé pour différentes valeurs de n et p. On en déduit par des abaques dans quel domaine (n,p), l'intervalle de confiance autour de la valeur estimée n'est pas "trop" dissymétrique (figure 1).

Le respect des conditions : R et S proches de 1, impose des restrictions conséquentes sur n et p (Schaan, 1993). Même si les contraintes diminuent avec un risque d'erreur croissant, des proportions proches de 0 ou 1 conduisent toujours à des sous-estimations de l'incertitude qui restent importantes.

En dehors des limites définies, les formules usuelles de variance ne fournissent pas une mesure correcte de l'incertitude. Toutefois, les rapports R_α peuvent être assimilés à des facteurs de correction (tableau 1) et donc permettre de calculer une variance corrigée :

$$\hat{Var}_{cor\,\alpha}(\hat{P}) = \frac{Var_e(\hat{P})}{R_\alpha^2} \qquad \text{, où "cor" signifie corrigée} \qquad (7)$$

Bien que ne résolvant pas le problème de l'asymétrie, cette variance corrigée fournit une mesure plus acceptable que celle obtenue d'après les formules usuelles.

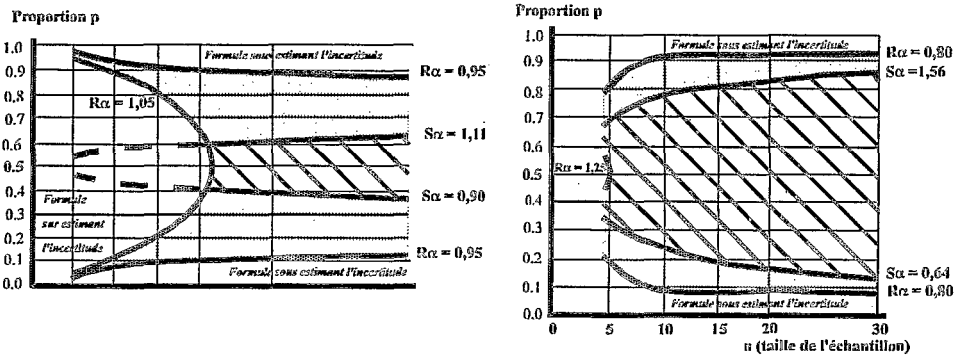

Figure 1. Limites de validité pour un risque d'erreur de 5 et 20%. Les zones en grisée et en hachurée indiquent l'aire de validité des formules usuelles de variance comme mesure de l'incertitude et la zone hachurée l'aire où en plus les conditions de symétrie sont respectées.

Tableau 1 : Facteurs de correction (ratios Rα des incertitudes) pour des seuils de confiance de 95 et 80 %. Pour les valeurs de a>n/2, R correspond au complément à n de la valeur recherchée, donc pour n-a. (Pour n = 22 et a= 18, lire R pour a = 4). (Pour a = 0, R = 0)

a= np	Taille n de l'échantillon									α = 95%
	2	3	4	5	6	7	8	9	10	11
1	12.36	3.19	1.99	1.56	1.35	1.22	1.13	1.07	1.02	0.99
2	0.00	3.19	2.13	1.70	1.48	1.34	1.25	1.19	1.14	1.10
3		0.00	1.99	1.70	1.51	1.38	1.29	1.23	1.18	1.14
4			0.00	1.56	1.48	1.38	1.30	1.25	1.20	1.16
5				0.00	1.35	1.34	1.29	1.25	1.21	1.17
	n*									
	12	14	16	18	20	22	24	26	28	30
1	0.96	0.92	0.89	0.86	0.85	0.83	0.82	0.81	0.80	0.80
2	1.07	1.02	0.99	0.97	0.95	0.93	0.92	0.91	0.90	0.89
3	1.11	1.07	1.03	1.01	0.99	0.97	0.96	0.95	0.94	0.93
4	1.13	1.09	1.06	1.03	1.01	1.00	0.99	0.97	0.97	0.96
5	1.15	1.10	1.07	1.05	1.03	1.01	1.00	0.99	0.98	0.97
6	1.15	1.11	1.08	1.06	1.04	1.02	1.01	1.00	0.99	0.98
7		1.11	1.08	1.06	1.05	1.03	1.02	1.01	1.00	0.99
8			1.09	1.07	1.05	1.04	1.02	1.02	1.01	1.00
9				1.07	1.05	1.04	1.03	1.02	1.01	1.00
10					1.05	1.04	1.03	1.02	1.01	1.01
11						1.04	1.03	1.02	1.02	1.01
12							1.03	1.03	1.02	1.01
13								1.03	1.02	1.01
14									1.02	1.02
15										1.02

α = 80%	a= np	Taille n de l'échantillon					
		2	5	10	15	20	30
	1	3.43	0.81	0.85	0.78	0.76	0.73
	2	0.00	1.20	0.93	0.87	0.84	0.80
	3		1.20	0.97	0.91	0.88	0.85
	4		0.81	0.99	0.93	0.90	0.87
	5		0.00	0.99	0.94	0.92	0.89
	6			0.99	0.95	0.93	0.90
	7			0.97	0.95	0.93	0.91
	8			0.93		0.94	0.92
	9			0.85		0.94	0.92
	10			0.00		0.95	0.93
	11						0.93
	12						0.93
	15						0.94

IV- CONCLUSION

Le prélèvement d'échantillons de grande taille demeure bien évidemment souhaitable pour aboutir à des précisions d'estimations élevées et fiables. Les méthodes proposées ici doivent être considérées comme outils palliatifs, lorsque les contraintes sont telles que seuls des échantillons de petites tailles sont pratiquement réalisables ou quand les proportions observées tendent vers zéro ou un.

Une autre possibilité consisterait à travailler non plus sur les variances d'estimation mais sur les intervalles de confiance. Ceci nécessiterait cependant de construire les lois successives des variables et d'aboutir à la fonction de densité finale. Une telle démarche ne pourrait être envisagée que sous hypothèses d'indépendance des variables et de symétrie des distributions.

BIBLIOGRAPHIE

ABBOUD (N.), AUDROING (J.F.), 1989 - *Probabilités et Inférences Statistiques*. Nathan éditeur, Collection Supérieur/Économie, 351 p.

COCHRAN (W.J), 1977 - *Sampling technics*. Third edition, Wiley & Son, New York, 413 p.

GROSBRAS (J.M.), 1987 -. *Méthodes statistiques des sondages*. Economica, Paris, 331p.

KENDALL (M.G.), STUART (A.), 1958 - *The Advanced Theory of Statistics*. Griffin and Co. LTD, London, 3 vol., 690 p.

SCHAAN (O.), 1993. L'exploitation des anguilles sub-adultes (*Anguilla anguilla*, L.) dans les estuaires de la Loire et de la Vilaine : Méthodes d'estimation des captures par âge. Thèse de Doctorat, ENSAR, Rennes, 156 p.

SCHERRER (B.), 1983 - Techniques de sondage en écologie. *In : Stratégies d'échantillonnage en écologie*. Frontier S.(éd.), Masson, Paris, 63-162.

SCHERRER (B.), 1984 - *Biostatistique*. Gaëtan Morin, Chicoutimi, Québec, 850 p.

WONNACOTT (T.H.), WONNACOTT (R.J.), 1991 - *Statistique : Économie, Gestion, Sciences, Médecine*. Économica, 4ème édition, Paris, 919 p.