Intervalles de confiance

Pierre Duchesne

August 1, 2017

- Habituellement il n'est pas suffisant de donner une estimation ponctuelle. Il faut aussi donner une appréciation de l'exactitude des estimations produites.
- C'est l'objectif principal entourant la construction des intervalles de confiance.

Intervalle de confiance de niveau 95%

- L'analyste a un certain paramètre d'intérêt à estimer.
- S'il est possible de prendre encore et encore des échantillons dans la population et si nous construisons un intervalle de confiance (IC) utilisant notre procédure pour chaque échantillon, il est attendu que 95% des intervalles résultants contiennent la vrai valeur du paramètre.

Nombre fini d'échantillons

- En échantillonnage d'une population finie, on dispose d'un maximum de 2^N échantillons différents.
- Il y a ainsi un nombre fini d'échantillons.
- Chaque échantillon a sa probabilité: en fait, chaque échantillon s a une probabilité $p(s)$ de survenir.

Calcul du niveau de confiance

- Supposons que l'on désire estimer le paramètre inconnu t_y.
- On s'intéresse à un intervalle de confiance de la forme:

$$IC(s) = [m_s, M_s]$$

- Posons $u(s) = 1$ si t_y est dans $IC(s)$, et $u(s) = 0$ sinon.
- Le niveau de confiance exact est donc:

$$\text{Niveau exact: } \sum_{s \in \mathcal{S}} u(s)p(s)$$

Exemple 2.7 (Lohr, pp. 35-36)

On considère pour fins d'illustration la population ayant pour valeurs:

k	y
1	1
2	2
3	4
4	4
5	7
6	7
7	7
8	8
total	40

Exemple (suite)

- Pour estimer le total inconnu $t_y = 40$, on procède à un tirage aléatoire simple (plan SI) de taille $n = 4$.
- Une procédure possible (sans forcément un fondement théorique), pourrait être:

$$[N\bar{y}_s - 4S_{ys}, N\bar{y}_s + 4S_{ys}]$$

- Si vous vous demandez pourquoi cet intervalle, la réponse est simplement: pourquoi pas!
- Quel est le niveau de confiance de cette procédure?

Échantillons				Estimateurs	Variance	Ecart-type	Borne inf	Borne sup	Inclus?
1	2	4	4	22	2.25	1.5	16	28	0
1	2	4	7	28	7	2.645751311	17.41699476	38.58300524	0
1	2	4	7	28	7	2.645751311	17.41699476	38.58300524	0
1	2	4	7	28	7	2.645751311	17.41699476	38.58300524	0
1	2	4	8	30	9.583333333	3.095695937	17.61721625	42.38278375	1
1	2	4	7	28	7	2.645751311	17.41699476	38.58300524	0
1	2	4	7	28	7	2.645751311	17.41699476	38.58300524	0
1	2	4	7	28	7	2.645751311	17.41699476	38.58300524	0
1	2	4	8	30	9.583333333	3.095695937	17.61721625	42.38278375	1
1	2	7	7	34	10.25	3.201562119	21.19375153	46.80624847	1
...	...	...	...	...	...	...	...	...	
2	4	4	8	36	6.333333333	2.516611478	25.93355409	46.06644591	1
2	4	7	7	40	6	2.449489743	30.20204103	49.79795897	1
2	4	7	7	40	6	2.449489743	30.20204103	49.79795897	1
2	4	7	8	42	7.583333333	2.753785274	30.98485891	53.01514109	1
2	4	7	7	40	6	2.449489743	30.20204103	49.79795897	1
2	4	7	8	42	7.583333333	2.753785274	30.98485891	53.01514109	1
...	...	...	...	...	...	...	...	...	
4	7	7	7	50	2.25	1.5	44	56	0
4	7	7	8	52	3	1.732050808	45.07179677	58.92820323	0
4	7	7	8	52	3	1.732050808	45.07179677	58.92820323	0
4	7	7	8	52	3	1.732050808	45.07179677	58.92820323	0
7	7	7	8	58	0.25	0.5	56	60	0

Niveau exact dans l'exemple précédent

- Utilisant les propriétés d'un plan SI, chaque échantillon a pour probabilité $1/70$ de survenir.
- On dénombre 54 échantillons qui contiennent la vraie valeur du paramètre valant $t_y = 40$.
- Ainsi le niveau exact de la procédure est $54/70 = 0.77$.

Que se passe-t-il en pratique?

- En pratique, il est impensable d'énumérer tous les 2^N échantillons et de calculer toutes les statistiques.
- Ainsi, en pratique, on ne peut déterminer le niveau de confiance exact.

Rappel: échantillonnage dans une population avec remise: Théorème limite central (TLC)

- Si on dispose d'un échantillon aléatoire $X_1, \ldots, X_n$, iid, $E(X_i) = \mu$, var$(X_i) = \sigma^2$, alors on sait par le TLC que:

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- Connaître (approximativement) la loi de la moyenne est l'ingrédient fondamental pour construire des intervalles de confiance.

Échantillonnage sans remise dans une population finie

- Le TLC ne s'applique pas directement car nous ne sommes pas dans les conditions du théorème.
- Il existe cependant d'autres types de TLC et l'intervalle de confiance suivant est valide:

$$\left[\bar{y}_s - z_{1-\alpha/2}\sqrt{1-\frac{n}{N}}\frac{S_{yU}}{\sqrt{n}}, \bar{y}_s + z_{1-\alpha/2}\sqrt{1-\frac{n}{N}}\frac{S_{yU}}{\sqrt{n}}\right]$$

  avec $z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ d'une $\mathcal{N}(0,1)$.
- En pratique on peut remplacer S_{yU} par S_{ys}.