Théorie statistique de l'échantillonnage

L'estimation désigne le procédé par lequel on détermine les valeurs inconnues des paramètres de la population à partir des données de l'échantillon. Pour cela, il faut passer par des variables aléatoires dont on connait les lois de probabilité (Fig. 1).

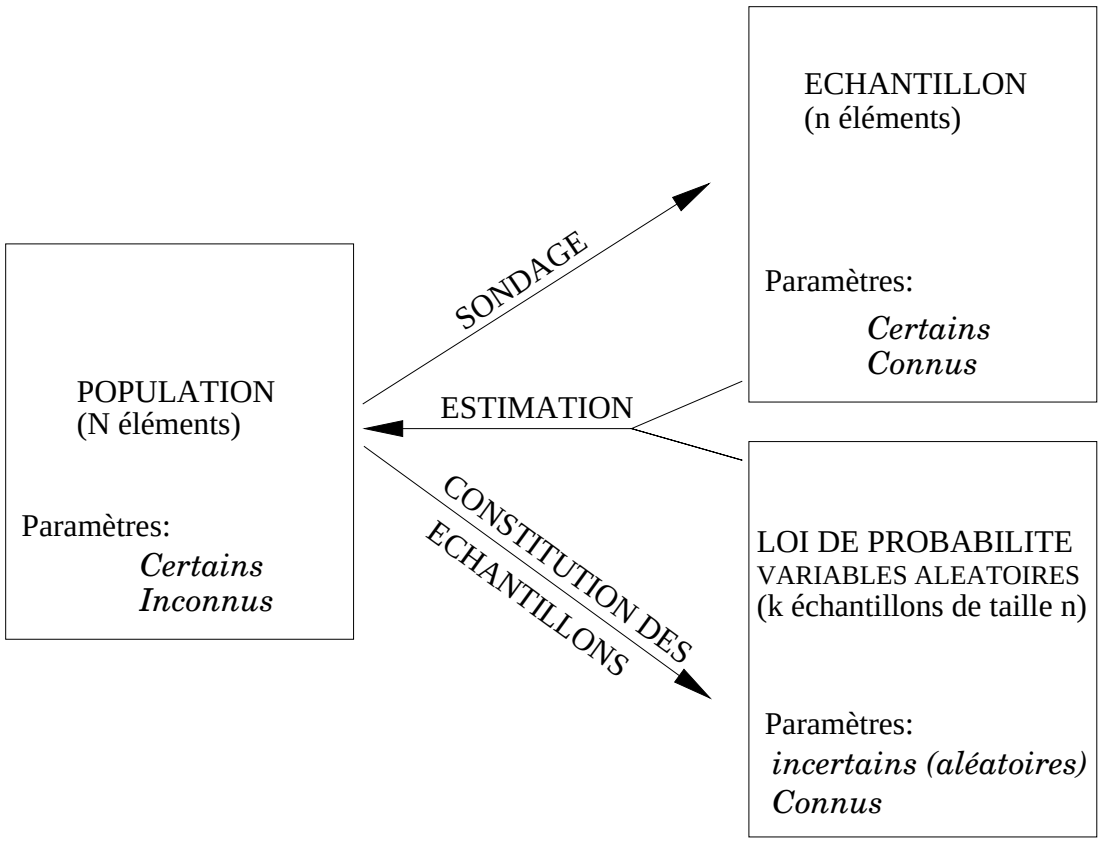

Figure 1: Principe général de l'estimation.

Prenons tous les échantillons possibles de taille N tirés d'une population donnée. Pour chaque échantillon, on peut calculer une statistique (moyenne, variance, etc...) qui variera avec l'échantillon. Pour tous les échantillons, on obtient alors une distribution de la statistique que l'on nomme *la distribution d'échantillonnage*. Pour la validité des résultat, il est important que les échantillons soient représentatif de la population concernée.

Combien d'échantillons de n éléments peuvent être isolés d'une population de N éléments?

L'échantillonnage peut se faire avec ou sans remise et une population peut être considérée comme finie ou infinie. Une population finie dans laquelle on procède à un échantillonnage avec remise peut être théoriquement considérée comme infinie. Dans la pratique, il en va de même pour des populations finies mais de grandes tailles.

Pour chaque distribution d'échantillonnage, on peut calculer une moyenne, un écart type, une variance etc ... On peut donc parler de la moyenne de la distribution d'échantillonnage des moyennes, de la moyenne de la distribution d'échantillonnage des variance, et de la variance de la distribution d'échantillonnage des moyennes. Il convient donc d'être prudent sur les termes utilisé car il est facile de se tromper. Par convention, on utilise souvent

- μ, σ et σ^2 comme les symboles de la moyenne, de l'écart type et de la variance d'une population.
- $\bar{X}$, s et s^2 comme les symboles de la moyenne, de l'écart type et de la variance d'un échantillon.
- $\mu_{\bar{X}}$, $\sigma_{\bar{X}}$ et $\sigma^2_{\bar{X}}$ comme les symboles de la moyenne, de l'écart type et de la variance de la distribution d'échantillonnage des moyennes.
- μ_{s^2}, σ_{s^2} et $\sigma^2_{s^2}$ comme les symboles de la moyenne, de l'écart type et de la variance de la distribution d'échantillonnage des variances.
- etc ...

Distribution d'échantillonnage des moyennes

Supposons tous les échantillons de taille N constitués sans remise à partir d'une population finie de taille $N_p > N$. On a alors

$$\mu_{\bar{X}} = \mu \qquad \text{et} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}}$$

Si la population est infinie ou que l'échantillonnage est avec remise,

$$\mu_{\bar{X}} = \mu \qquad \text{et} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

A partir du *thèorème central limite*, il est possible de montrer que la distribution d'échantillonnage est asymptotiquement normale. En d'autre termes, pour

N suffisamment grand ($N > 30$ ou de manière plus générale $N > 25\ \alpha_3^2$), la distribution d'échantillonnage se rapproche d'une distribution normale de moyenne $\mu_{\bar{X}}$ et d'écart type $\sigma_{\bar{X}}$. Lorsque N est trop petit, on utilisera alors la théorie des tests exacts plus communément appellée la théories des petits échantillons.

> Une population comporte 4 individus dont les masses sont respectivement de 3, 7, 11 et 15 kg. On considère tous les couples qu'il est possible d'extraire de cette population. Vérifier les relations entre la moyenne de la population et la moyenne de la distribution d'échantillonnage de la moyenne et entre l'écart type de la population et l'écart type de distribution d'échantillonnage de la moyenne.

Distribution d'échantillonnage des proportions

Un événement se réalise avec la probabilité et ne survient pas avec la probabilité $q = 1 - p$. La distribution d'échantillonnage des proportions de moyenne se caractérise alors par

$$\mu_P = p \qquad \text{et} \qquad \sigma_P = \sqrt{\frac{pq}{N}}.$$

Là encore cette distribution d'échantillonnage est pratiquement distribuée normalement. Elle suit une une loi binômiale de paramètres (N, p).

> Une agriculteur envoie 1000 lots contenant chacun 100 fruits. Si 5 % des fruits sont véreux, dans combien de lots devrait-il y avoir (a) moins de 90 fruits consommables et (b) plus de 98 fruits consommables ?

Distribution d'échantillonnage des sommes ou des différences

A partir de deux populations différentes, on extrait des échantillons indépendants de tailles N_1 et N_2. En procédant à toutes les combinaisons possibles des échantillons des deux populations, on peut obtenir des distributions d'échantillonnage de la variable V notée V_1 et V_2. On peut aussi obtenir une distribution d'échantillonnage de la différence des statistiques. La moyenne et l'écart type de cette distribution s'écrivent

$$\mu_{V_1 - V_2} = \mu_{V_1} - \mu_{V_2} \qquad \text{et} \qquad \sigma_{V_1 - V_2} = \sqrt{\sigma_{V_1}^2 + \sigma_{V_2}^2}.$$

Si V_1 et V_2 sont les moyennes des deux échantillons, nous avons

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} \qquad \text{et} \qquad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} - \sigma^2_{\bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

la moyenne et l'écart type de la distribution d'échantillonnage de la différence des moyennes.

De même, la moyenne et l'écart type de la distribution d'échantillonnage de la différence de deux populations binômiales de paramètres (N_1, p_1) et (N_2, p_2) s'écrit

$$\mu_{P_1-P_2} = \mu_{P_1} - \mu_{P_2} = p_1 - p_2 \qquad \text{et} \qquad \sigma_{P_1-P_2} = \sqrt{\sigma^2_{P_1} - \sigma^2_{P_2}} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}}.$$

Dans tous ces cas, la distribution d'échantillonnage de la différence des moyennes suit une loi normale pour $N_1, N_2 > 30$.

On peut aussi avoir à traiter la distribution d'échantillonnage de la somme des statistiques. La moyenne et la variance sont alors

$$\mu_{V_1+V_2} = \mu_{V_1} + \mu_{V_2} \qquad \text{et} \qquad \sigma_{V_1+V_2} = \sqrt{\sigma^2_{V_1} + \sigma^2_{V_2}}.$$

Le débit moyen d'un chenal artificiel est de 300 litres par minute avec un écart type de 50 litres. Quelle est la probabilité qu'en 25 minutes, le volume d'eau dépasse la contenance d'un bassin de rétention de 8.2 m^3 ?

La résistance à la rupture du hêtre et du bouleaux sont respectivement de 4500 kg et de 4000 kg avec des écarts type respectifs de 200 kg et 300 kg. Si l'on teste 100 bouleaux et 50 hêtres, quelle est la probabilité que la résistance moyenne des hêtres soit (a) supérieure de 600 kg à celle des bouleaux et (b) supérieure d'au moins 450 kg à celle des bouleaux ?