Notions sur la théorie statistique de l'estimation

L'estimation désigne le procédé par lequel on détermine les valeurs inconnues des paramètres d'une population à partir des données d'un échantillon. Pour cela, il faut passer par des variables aléatoires dont on connait les lois de probabilité (Fig. 1). Les informations fournies par un échantillon ne sont inteprétable que si elles sont accompagnées d'informations quantitatives fixant le degré de confiance qu'on peut leur accorder.

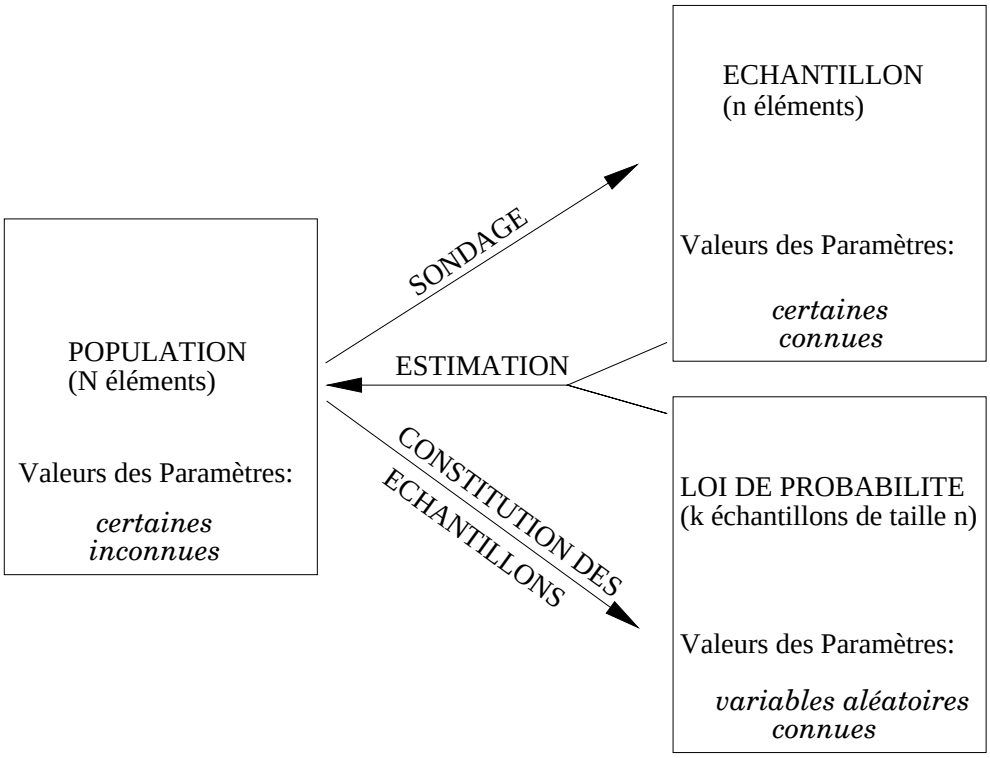

Figure 1: Principe général de l'estimation.

La **distribution d'échantillonnage** d'un paramètre (proportion, moyenne, variance, quantité totale etc...) est la distribution de ce paramètre obtenue à partir de l'ensemble des échantillons.

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

A	7	Nombre d'heures consacrées au travail universi-
B	3	taire pour cinq étudiants $\{A, B, C, D, E\}$. Pour
C	6	des sous-échantillons de 3 étudiants, retrouvez
D	10	la moyenne et l'écart type de la distribution
E	4	d'échantillonnage des moyennes.

Théorie des petits échantillons

On considère souvent que pour de grands échantillons ($N > 30$) les distributions d'échantillonnage des statistiques suivent des lois normales. Cette approximation est d'autant meilleure que N est grand. Pour des échantillons de petites tailles ($N < 30$), cette approximation n'est plus valable et se détériore lorsque $N \to 0$. Il faut donc faire appel à la *théorie des tests exacts* pour l'étude des distributions d'échantillonnage des statistiques des petits échantillons. Cette théorie utilise une distribution dite de Student qui présente l'énorme avantage de pouvoir s'appliquer quel que soit le nombre d'échantillons. Par exemple, la loi normale est une loi de Student si $N \to \infty$

La Distribution t de Student

Si pour des échantillons de taille N tirés d'une population normale de moyenne μ, on calcule t

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1}$$

en utilisant les moyennes $\bar{X}$ et la variance s^2 de chaque échantillon, on obtient une distribution d'échantillonnage qui respecte une loi de Student. On peut alors définir des intervalles de confiance à différents niveaux de risque en utilisant une table de Student.

Par exemple pour un intervalle de confiance de 95%, on utilise $-t_{0.975}$ et $t_{0.975}$ pour limiter 2.5% de l'aire dans chaque queue de distribution ($\bar{X} \to \pm\infty$). Dès lors,

$$-t_{0.975} \ < \ \frac{\bar{X} - \mu}{s}\sqrt{N-1} \ < \ t_{0.975},$$

et l'intervalle de confiance pour la moyenne de la population globale s'écrit

$$\bar{X} - t_{0.975}\frac{s}{\sqrt{N-1}} \ < \ \mu \ < \ \bar{X} + t_{0.975}\frac{s}{\sqrt{N-1}}.$$

En général, les limites de confiance pour la moyenne de la population s'écrivent

$$\bar{X} \pm t_c \frac{s}{\sqrt{N-1}}$$

où les valeurs $\pm t_c$ dites valeurs critiques ou coefficients de l'intervalle de confiance sont fonction du niveau de confiance recherché et de la taille de l'échantillon.

Test d'hypothèse sur les moyennes

Pour tester l'hypothèse H_0 que la population normale a pour moyenne μ, on utilise le score

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1}.$$

La distribution de t est une loi de Student à $N-1$ degrés de liberté.

Sur 7 jours, on observe un débit moyen journalier de 32 m^3 et un écart type de 12 m^3. Estimer l'interval moyen des débits journaliers avec un risque de 5% et de 1%.

Test d'hypothèse sur les différences de moyennes

Pour tester l'hypothèse H_0 que deux sous-populations de N_1 et N_2 éléments, de moyenne $\bar{X}_1$ et $\bar{X}_2$, et de variance s_1^2 et s_2^2 sont issues de la même population, on utilise le score

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \qquad \text{avec} \qquad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}.$$

La distribution de t est une loi de Student a $N_1 + N_2 - 2$ degrés de liberté.

Deux sous-populations d'agneaux ont été soumises à différents régimes alimentaires. Après huit mois, voici les masses mesurées:

Régime I	91	97	102	93	95	90	99	119	86	97	87	97
Régime II	90	99	106	95	97	93	101	116	89	100	89	95

Les régimes alimentaires ont-ils un effet sur la masse des agneaux?

La Distribution du χ^2

Si pour des échantillons de taille N tirés d'une population normale de variance σ^2, on calcule

$$\chi^2 \;=\; \frac{Ns^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \;...\; + (X_n - \bar{X})^2}{\sigma^2}$$

en utilisant les moyennes $\bar{X}$ et la variance s^2 de chaque échantillon, on obtient une distribution d'échantillonage qui respecte une loi du *chi-carré*. On peut alors définir des intervalles de confiance à différents niveaux de confiance en utilisant une table du χ^2

$$\chi^2_{0.025} \;<\; \frac{Ns^2}{\sigma^2} \;<\; \chi^2_{0.975}$$

à $N - 1$ degré de liberté. Dès lors, il est possible d'estimer σ dans l'intervalle

$$\frac{s\sqrt{N}}{\sqrt{\chi^2_{0.975}}} \;<\; \sigma \;<\; \frac{s\sqrt{N}}{\sqrt{\chi^2_{0.025}}}.$$

Noter que le nombre de degré de liberté ν est systématiquement le nombre N d'observations au sein de l'échantillon moins le nombre k de paramètres de populations que l'on doit estimer à partir de l'échantillon. Dans les exemples ci-dessus, il s'agit de la ou des moyennes (test de Student), et de la variance (test du χ^2).

L'écart type de la durée de vie de 10 batteries est de 120 heures. Quel est l'intervalle de confiance pour les batteries de la même marque aux risques de 5% et 1%?

La Distribution F de Fisher

Comme pour la moyenne, il est parfois important de déterminer la distribution d'échantillonnage de la différence de deux variances. Très diffcile à mettre en oeuvre, il est plus facile d'étudier le rapport s_1^2/s_2^2 de deux variances de sous-populations donnees. Cette statistique suit la loi de Fisher. Plus exactement, pour deux échantillons de tailles N_1 et N_2 et de variance s_1^2 et s_2^2 provenant de populations normales de variances σ_1^2 et σ_2^2, la statistique

$$F = \frac{\frac{N_1 s_1^2}{(N_1-1)\sigma_1^2}}{\frac{N_2 s_2^2}{(N_2-1)\sigma_2^2}}$$

suit une distribution de Fisher avec $\nu_1 = N_1 - 1$ et $\nu_2 = N_2 - 1$ degrés de liberté.

Deux échantillons de taille 11 et 15 sont tirés de deux populations normales de variance 40 et 60. Si les variances des échantillons sont respectivement 90 et 50, déterminer si la variance du premier échantillon est significativement plus grande que la variance du second au risques de 5% et de 1%.

Notions sur la théorie statistique de la décision

Hypothèses et risques d'erreur statistique

H_0 est une hypothèse statistique. H_1 est une hypothèse alternative qui suppose généralement le fait contraire de H_0.

Hypothèses	H_0 est vraie	H_1 est vraie
H_0 acceptée	Bonne décision	Erreur β
H_0 rejetée	Erreur α	Bonne décision

Les masses des tortues de mer mâles respectent une loi normale de moyenne $\mu_M = 112.3\ mm$ et de variance $\sigma_M = 10.5\ mm$. Les masses des tortues de mer femelles respectent une loi normale de moyenne $\mu_F = 98.6\ mm$ et de variance $\sigma_F = 8.7\ mm$. On veut établir un test qui permet de discriminer entre les individus mâles et les individus femelles. Fixer les hypothèses H_0 et H_1, et estimer leurs erreurs α et β.

Par définition,

le seuil de probabilité de 5% est "*significatif*",
le seuil de probabilité de 1% est "*hautement significatif*",
le seuil de probabilité de 0.1% est "*très hautement significatif*".

Puissance et robustesse d'un test

Pour une même erreur α, le test qui fournit l'erreur β la plus petite est, par définition, le plus puissant. En pratique, il s'agit de tracer la courbe de puissance du test ou courbe caractéristique d'efficacité. Elle indique la probabilité de prendre une bonne décision si H_1 est vraie. La puissance est donc mesurée par la probabilité $1 - \beta$ pour un α donné.

Une machine outil produit des boulons de $0.574\ g$ et de variance $0.008\ g^2$ à la fréquence de 1 boulons par seconde. Toutes les minutes, 10% des boulons sont prélevés pour vérifier le bon fonctionnement de la machine. En prenant 1% de risque, établir l'hypothèse principale H_0. Quelles sont les hypothèse alternatives? En considérant que seule la masse moyenne des boulons peut varier si la machine outil commence à mal fonctionner, faire la courbe de puissance de test. Même question avec un prélévement de 50% de la production.