# 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 l'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.

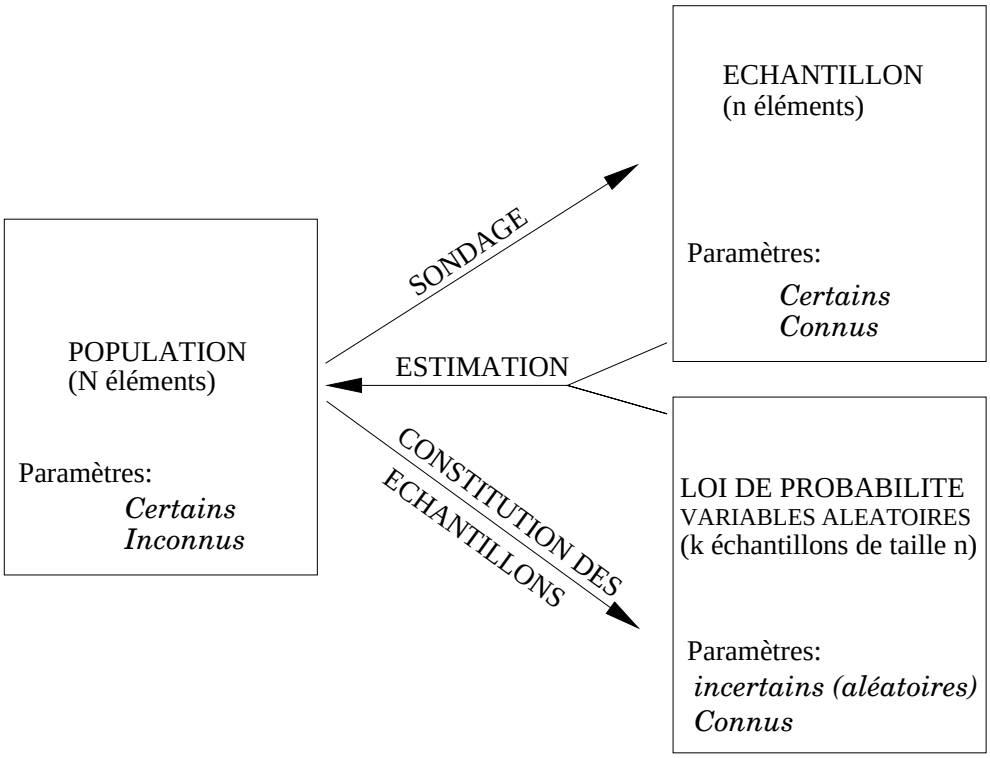


Figure 1: Principe général de l'estimation.

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? |
|---|

# Théorie des petits échantillons

On considère souvent que pour de grands échantillons ($N > 30$) les distributions d'échantillonnage des statistiques sont approximativement normales. L'approximation est d'autant meilleure que $N$ est grand. Pour des échantillons de petites tailles ($N < 30$), cette approximation n'est pas valable et est d'autant plus mauvaise que $N \to 0$. L'étude des distributions d'échantillonnage des statistiques de tels échantillons s'appelle la *théorie des petits échantillons* ou *théorie des tests exacts* car elle s'applique tout aussi bien quel que soit le nombre d'échantillons.

## La Distribution $t$ de Student

Si pour des échantillons de taille $N$ tirés d'une population normale de moyenne $\mu$, 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 $\mu$, 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.

La durée de vie de bactéries est en moyenne de 1120 $s$, avec un écart type de 125 $s$. Un échantillon de 8 bactéries provenant d'un autre laboratoire a une moyenne de 1070 $s$. Tester l'hypothèse que la durée de vie des bactéries est la même, au risque de 5% et 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.

Avec le carburant de type $A$, cinq voitures ont roulé dans des conditions identiques pendant 22.6 $km$ pour chaque litre de carburant, avec un écart type de 0.48 $km$. Avec la marque B, pendant 21.4 $km$ avec un écart type de 0.54 $km$. Au risque de 5% et 1% peut-on dire qu'une essence est meilleure que l'autre ?

## La Distribution du $\chi^2$

Si pour des échantillons de taille $N$ tirés d'une population normale de variance $\sigma^2$, 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 $\chi^2$

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

à $N - 1$ degré de liberté. Dès lors, il est possible d'estimer $\sigma$ 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 $\nu$ 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 $\chi^2$).

Dans le passé, l'écart type de paquets de 40 $kg$ était de 0.25 $kg$. Sur un échantillon de 20 paquets pris au hasard, on mesure un écart type de 0.32 $kg$. Est-ce que l'on a une fluctuation significative au risque 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 $\sigma_1^2$ et $\sigma_2^2$, 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 8 et 10 sont tirés de deux populations normales de variance 20 et 36. Déterminer la probabilité que la variance du premier échnatillon soit plus de deux fois celle du second.

# 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 $\beta$ |
| $H_0$ rejetée | Erreur $\alpha$ | Bonne décision |

Les longueurs d'ailes de mésanges mâles respectent une loi normale avec $\mu_M = 63.84\ mm$ et $\sigma_M = 1.20\ mm$. Les longueurs d'ailes de mésanges femelles respectent une loi normale avec $\mu_F = 61.16\ mm$ et $\sigma_F = 1.11\ mm$. Fixer des hypothèses et estimer leurs erreurs $\alpha$ et $\beta$.

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 $\alpha$, le test qui fournit l'erreur $\beta$ 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 $\alpha$ donné.

Soit un test pour vérifier qu'une population se compose exactement de 50% d'hommes et de 50% de femmes. L'hypothèse principale $H_0$ est donc qu'il y a la même proportion d'hommes que de femmes. A partir d'une sous-population de 22 individus, déterminer l'hypothèse principale pour avoir un test significatif. Quelles sont les hypothèse alternatives? Faire la courbe de puissance de test. Même question avec une sous-population de 100 personnes.