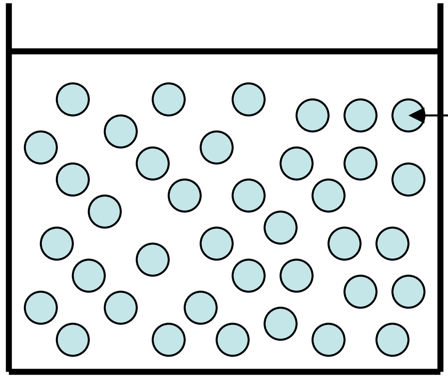
1. Introduction aux méthodes statistiques

Schéma de l'urne:



X_i **Variable aléatoire, réalisation de X (ou mesure X de l'individu i)**

Population cible, N individus
Variable aléatoire X, Loi P

On peut compliquer ce schéma pour un problème **multivarié**:
Revient à considérer une variable ayant plusieurs composantes.

$$\begin{cases} X = X^c, \ c = 1,\ldots k \\ X = (X^1, X^2, \ldots, X^k) \end{cases}$$

On peut alors calculer les moments de la population:
Moyenne théorique:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Variance:

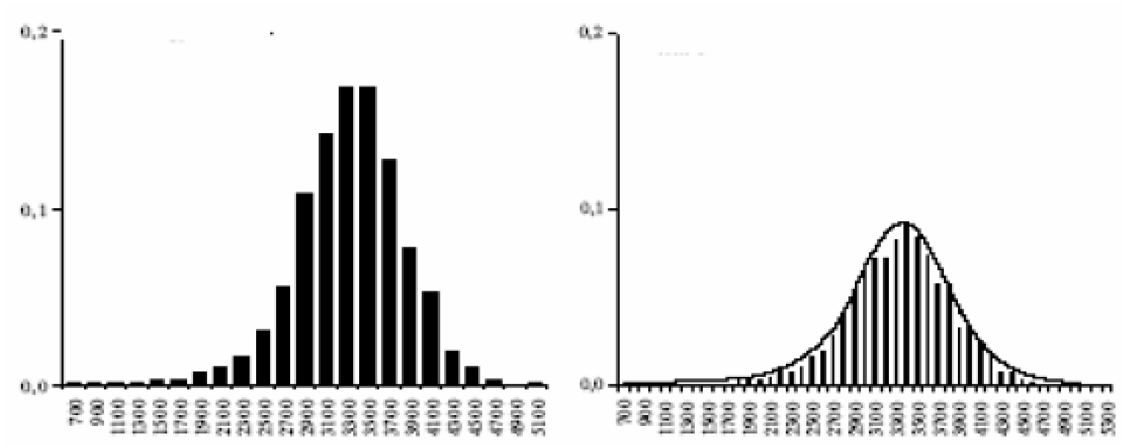
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

Statistique descriptive ou exploratoire ou analyse des données:
Description de la distribution de ces variables en population.
Suppose que l'on dispose de toutes les valeurs de la population.

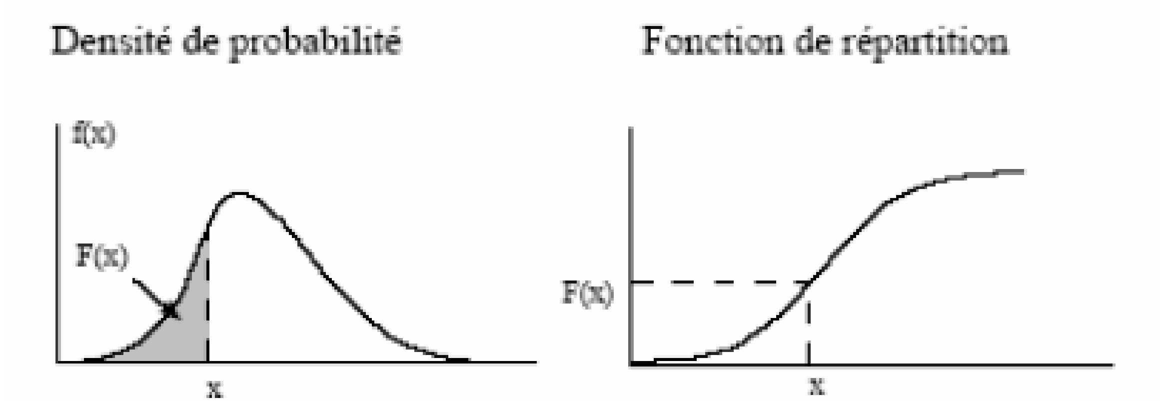
Principaux paramètres caractérisant les distributions (variables discrètes ou continues):

- Moyenne
- Variance → Permettent comparaisons entre les groupes
- Distribution elle-même

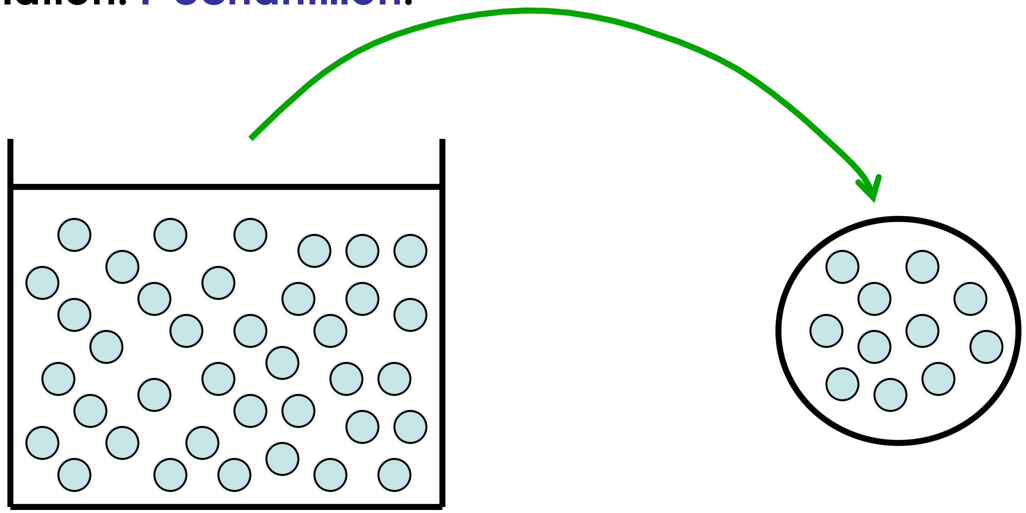
Cas discret: histogramme



Cas continu: fonction densité



Souvent difficile d'accéder à l'ensemble de la population. On doit alors se baser sur une partie de cette population: l'échantillon.



Population cible, N individus
Variable aléatoire X, Loi P

$$X = (X^1, X^2, \ldots, X^k)$$

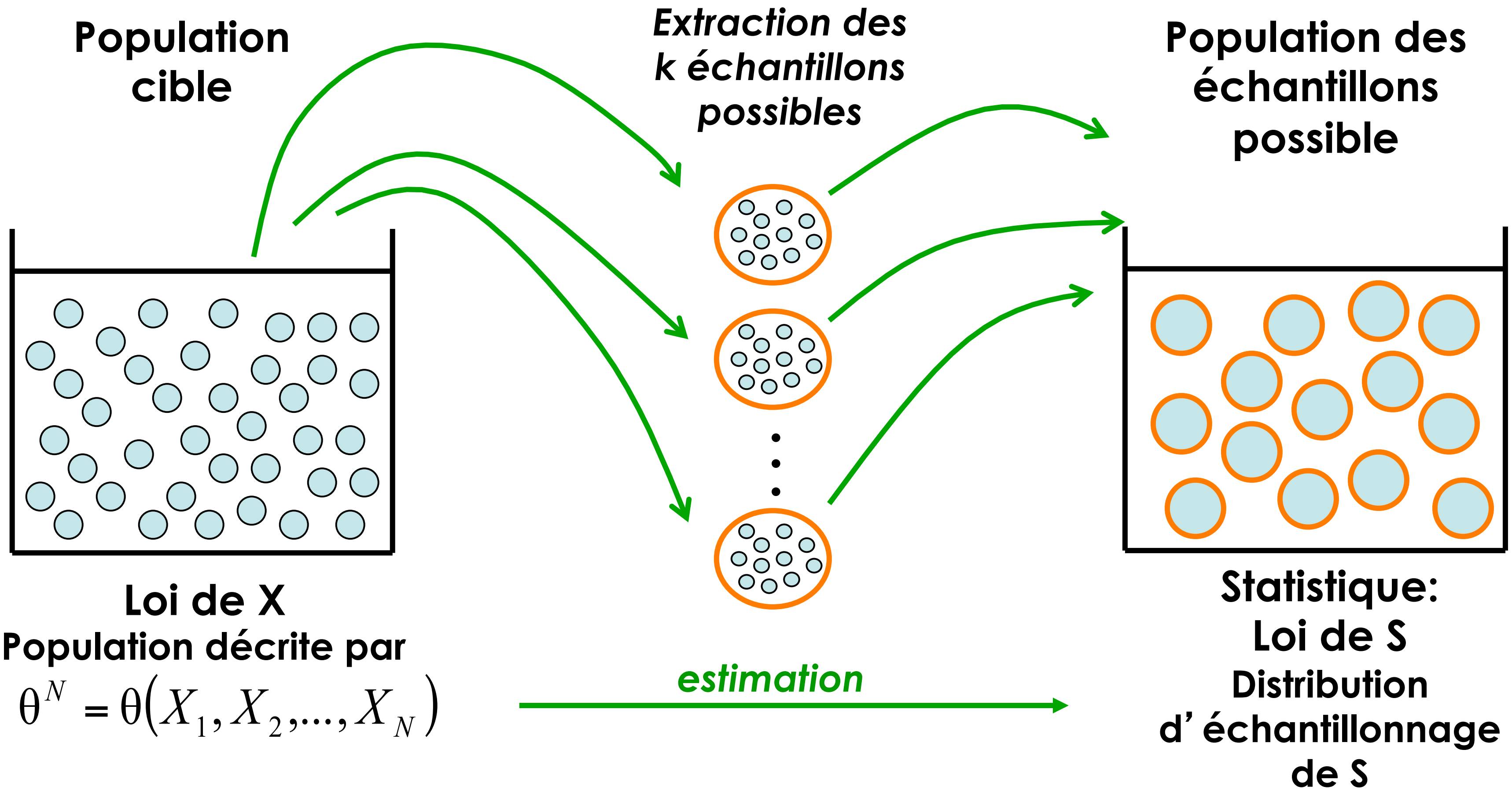
$$X_1, X_2, X_3, \ldots, X_N$$

$$X_1, X_2, \ldots, X_n$$

On considère ici que les variables de l'échantillon :

- **sont extraites de la population de manière indépendante;**
- **sont identiquement distribuées.**

Le même loi de probabilité s'applique: loi P, dépendant d'un paramètre inconnu θ, paramètre de population ou paramètre théorique.
Estimation des paramètres θ de P == S.



$$\begin{pmatrix} X_{1,1}, X_{1,2}, ..., X_{1,n} \\ X_{2,1}, X_{2,2}, ..., X_{2,n} \\ ... \\ X_{k,1}, X_{k,2}, ..., X_{k,n} \end{pmatrix} \Rightarrow \begin{pmatrix} S_1^n(X) \\ S_2^n(X) \\ ... \\ S_k^n(X) \end{pmatrix} = S^n(X)$$

Démarche probabiliste: démarche déductive partant de la population, connaissant sa distribution, vers les observations.
On parle de calcul de probabilités.
On parlera d'intervalle de pari, de calcul de risque, etc ...

$\neq$

Démarche Statistique: démarche inductive : partant d'une partie de la population appelé échantillon, on estimera des paramètres inconnus de cette population, et d'une manière plus générale, on essayera de répondre à des questions inférentielles sur cette population observée qu'en partie ...

2. Fluctuations d' échantillonnage

Considérons une variable aléatoire X suivant une loi f dans une population donnée.
n-échantillon: échantillon de taille n regroupant un ensemble de **variables aléatoires indépendantes et identiquement distribuées, de même loi f que X.**
En pratique: $\left(X_1, X_2, \ldots, X_n\right)$ **== observations réalisées**

avec $E(X_i) = \mu \quad ; \quad VAR(X_i) = \sigma^2$

Fluctuations d' une moyenne empirique

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Rightarrow \quad E(\overline{X}) = \mu \ ; \ VAR(\overline{X}) = \frac{\sigma^2}{n}$$

Fluctuations d' une variance empirique

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \Rightarrow \quad E\left(\frac{n}{n-1}S^2\right) = \sigma^2$$

Théorème Central-Limite (TCL)

Sous les hypothèses précédentes, on montre que

Moyenne échantillonnage standardisée

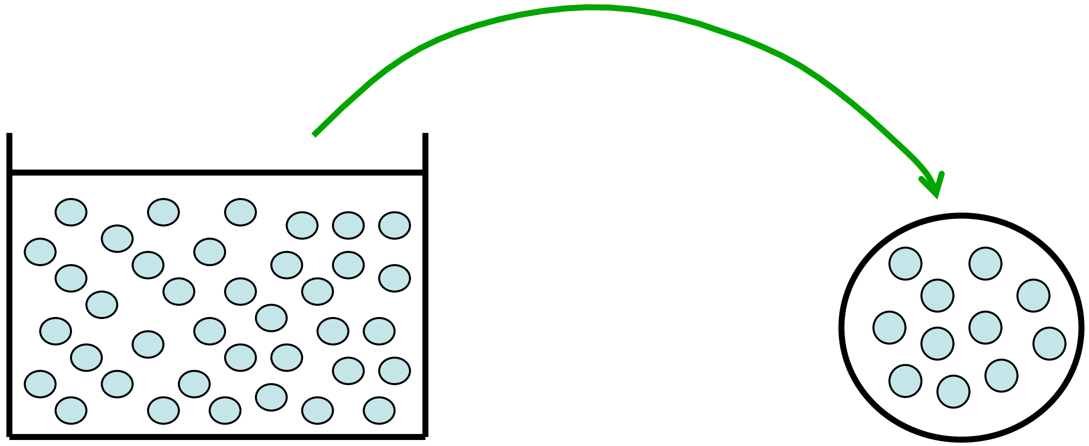
$$Z = \frac{\left(\overline{X} - \mu\right)}{\sigma / \sqrt{n}} \xrightarrow{n \to \infty} N(0,1)$$

∀ distribution sous-jacente, pour de ***<u>grands échantillons</u>***:
distribution d'échantillonnage connue = la loi normale.

En pratique, on considère que la moyenne échantillonnage suit une loi normale

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

3. Estimation



Population cible, N individus
Variable aléatoire X, Loi P
$X_1, X_2, X_3, \ldots, X_N$
P de paramètre $\theta = \theta^N(X)$

Observations:
Echantillon
n individus $X_1, X_2, \ldots, X_n$
Paramètre empirique $S^n(X)$

Démarche statistique:
estimer un paramètre théorique (ou paramètre en population ou paramètre de la distrib. théorique) à l' aide d' un échantillon de n observations issues de cette population.
On supposera vérifiée l' hypothèse d' échantillonnage aléatoire simple.

Exemple: $\overline{X}, s^2, F$ *estimateurs de* $\mu_X, \sigma_X{}^2, P$

3.1 Estimateur naturel ou estimateur fonctionnel

Estimateur naturel: estimateur qui se calcule dans l'échantillon de la même manière que dans la population.

Exemple:

Moyenne théorique en population: $$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Moyenne empirique, estimateur naturel de μ: $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

De même: variance empirique S^2 estimera de façon naturelle la variance théorique σ^2.

3.2 Biais d' un estimateur

Soit S(X) un estimateur d' un paramètre θ.

Biais de S:

$$\boxed{b(\theta) = E[S(X)] - \theta}$$

Calculé en utilisant la distribution d' échantillonnage de S.

Applications:

• **Biais d' une moyenne empirique = 0**, moyenne empirique est donc un estimateur sans biais de la moyenne théorique;
• **Biais d' une variance empirique**,
Variance empirique, estimateur naturel:

$$E\left(S^2\right) = \frac{n}{n-1}\sigma^2 \quad \Longrightarrow$$

Estimateur biaisé

• **Estimateur de la variance: préférable d' utiliser**

$$\overline{\overline{S}}^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{n}{n-1}S^2 \quad \text{et donc} \quad E\left(\overline{\overline{S}}^2\right) \approx \sigma^2 \quad \Longrightarrow$$

Considéré comme un estimateur non biaisé

Estimateur sans biais: S(X), fonction des observations Xi, est un estimateur sans biais du paramètre θ si et seulement si E(S(X))= θ.
On préférera tjs un estimateur de biais inférieur, si possible nul, et de variance minimale.

3.3 Variance d'un estimateur

Variance d'un estimateur S(X) d'un paramètre θ connu
= variance de sa distribution d'échantillonnage

$$Var(S(X)) = E\left[\left(S(X) - E(S(X))\right)^2\right]$$

Si S(X) est un estimateur sans biais de θ alors

$$Var(S(X)) = E\left[\left(S(X) - \theta\right)^2\right]$$

Entre deux estimateurs de θ, l'un non-biaisé et l'autre biaisé, on choisira celui qui dont la variance est la plus faible, c'est-à-dire celui qui, dans l'échantillon, s'écartera en moyenne le moins du paramètre.

Un estimateur est dit optimal s'il est celui parmi les estimateurs sans biais qui a la variance minimum.

3.3 Estimateur consistent

Considérons un échantillon constitué par toute la population: n=N;
Estimateur **consistent** si il est alors égal au paramètre en population:

$$S^N(X) = \theta^N$$

Note: toujours le cas des estimateurs naturels.

Estimateur est dit **convergent** si lorsque n → ∞
(échantillon de la taille de la population entière):

$$S^n(X) = \theta$$

Note: il suffit pour cela qu'il soit sans biais.

4. Intervalle de fluctuation d'une moyenne empirique

La variance permet de quantifier la dispersion d'une variable. On s'intéresse maintenant ici *aux fluctuations d'échantillonnage résultant de la dispersion de la variable*.
Donner l'intervalle de fluctuation d'une variable == donner dans quel intervalle les valeurs de l'échantillon (ou les valeurs de paramètres (moyenne, variance…) calculés à partir de cet échantillon) doivent se situer. Cet intervalle ne peut être qu'approché.
C'est donc un problème probabiliste!

L'intervalle de fluctuation (ou de variation) d'une moyenne empirique (observée) d'une variable X est caractérisée par:
- **Une distribution normale N(μ,σ^2) avec μ et σ connus.**
- **Une distribution qcq si n>30 (grand échantillon).**

= intervalle fixe vérifiant

$$P(\overline{X} \in I_f) = 1 - \alpha$$

$$\boxed{I_f = \left[\mu - \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}}, \mu + \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}}\right]}$$

(Z variable normale centrée réduite)

Où $e_{1-\alpha/2}$ = quantile défini par

$$\boxed{P\left(Z \in \left[-e_{1-\frac{\alpha}{2}}, +e_{1-\frac{\alpha}{2}}\right]\right) = 1 - \alpha \quad \text{avec} \quad Z \approx \mathrm{N}(\mu, \sigma^2)}$$

Intervalle de fluctuation (ou de variation) d' une proportion observée P0 d' une variable binaire X distribuée suivant une loi de Bernouilli de probabilité π

= intervalle fixe vérifiant

$$P(P_0 \in I_f) = 1 - \alpha$$

$$\boxed{I_f = \left[\pi_0 - \frac{\pi_0(1-\pi_0)}{\sqrt{n}} e_{1-\frac{\alpha}{2}}, \pi_0 + \frac{\pi_0(1-\pi_0)}{\sqrt{n}} e_{1-\frac{\alpha}{2}}\right]}$$

Valable si

$$\boxed{\begin{array}{l} n\pi_0 \geq 5 \\ n\pi_0(1-\pi_0) \geq 5 \end{array}}$$

5. Intervalle de confiance d’une moyenne théorique

Intervalle de confiance d’une moyenne théorique d’une variable X suivant une distribution normale N(μ,σ²) avec σ connue.

= intervalle fixe vérifiant

$$P(I_c \ni \mu) = 1 - \alpha$$

Construit à partir de l’intervalle de fluctuation:

$$P(\overline{X} \in I_f) = 1 - \alpha$$

$$\Leftrightarrow P\left(\mu - \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}} \leq \overline{X} \leq \mu + \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\boxed{P(I_c \ni \mu) = 1 - \alpha}$$

OU

$$\boxed{I_c = \left[\overline{X} - \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}} e_{1-\frac{\alpha}{2}}\right]}$$