Rappels

Notions de Bases en Biostatistiques

Pr Roch Giorgi

roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr/

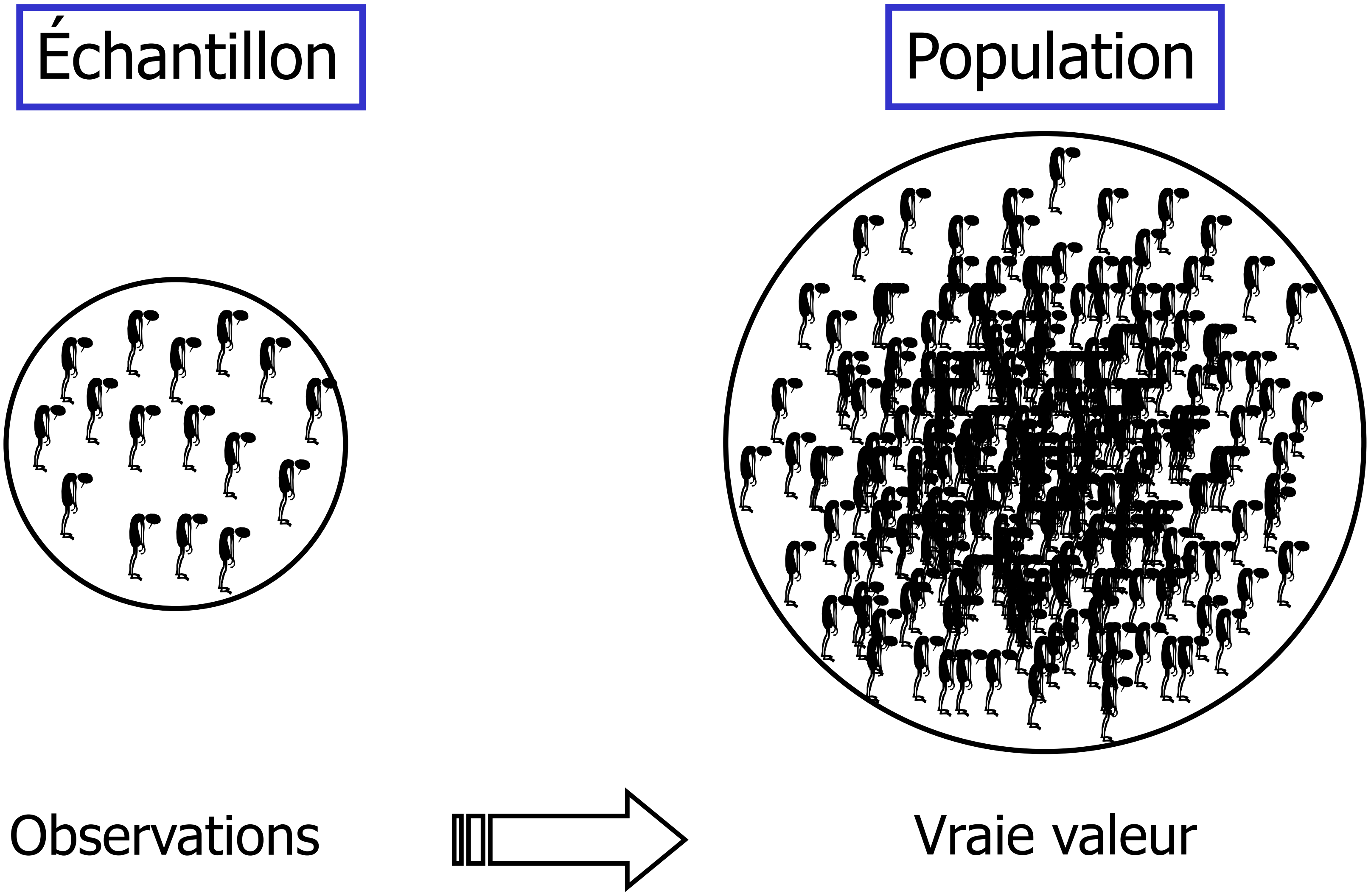
Population – Échantillon – Tirage au Sort

Population et Échantillon : Définitions

- Population
  - ✓ Ensemble d'individus ayant des caractéristiques qui leurs sont propres (français présentant des problèmes d'hypertension artérielle, ...)
  - ✓ Nombre d'individu souvent important

- Échantillon
  - ✓ Sous-ensemble d'une population
  - ✓ Sur chaque individu de l'échantillon on peut mesurer une caractéristique faisant l'objet de l'étude (impossible sur toute la population)

Population et Échantillon



Échantillon : Objectif

- Les observations faites sur l'échantillon servent à répondre aux questions que l'on se pose sur la population
- Les caractéristiques observées sont des variables aléatoires
- Leurs paramètres descriptifs permettent de connaître la distribution dans la population
    - Objectif : estimer les paramètres de la distribution de la population
    - Moyen : utiliser les observations faites sur l'échantillon

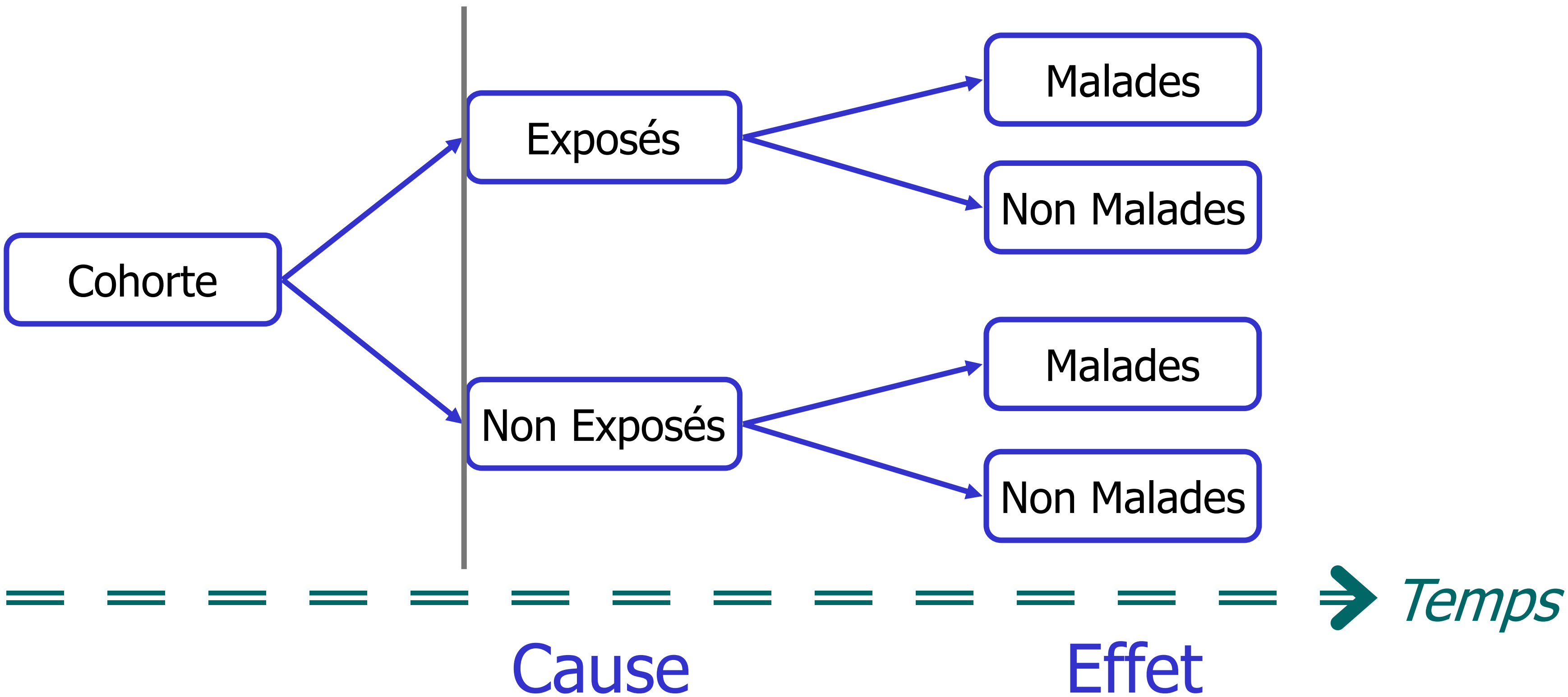
Population et Échantillon

Échantillon	Population
Critère d'intérêt	Critère d'intérêt
Estimation de la Caractéristique A	Caractéristique A ?
- *glycémie moyenne, écart-type* - *probabilité de décès à 5 ans* - *...*	- glycémie - décès à 5 ans - ...

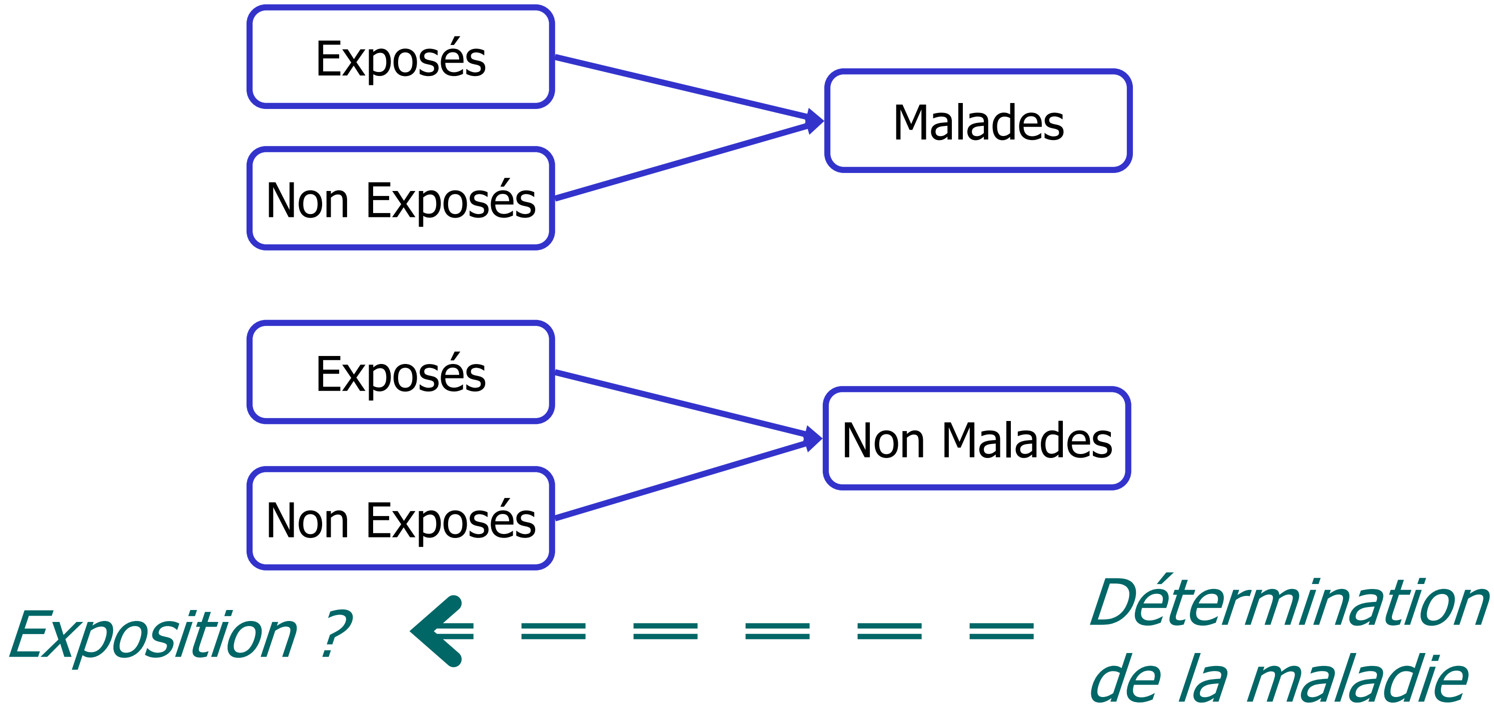
Constitution de l'Échantillon

- Un échantillon fourni des informations sur la population
- Un bon échantillon (« sans biais ») doit être représentatif de la population dont il est issu
- Nécessité de définir précisément la population
- L'échantillonnage aléatoire (tirage au sort) en est le meilleur moyen
- Le choix du processus peut dépendre de l'objectif de l'étude, donc du type d'étude

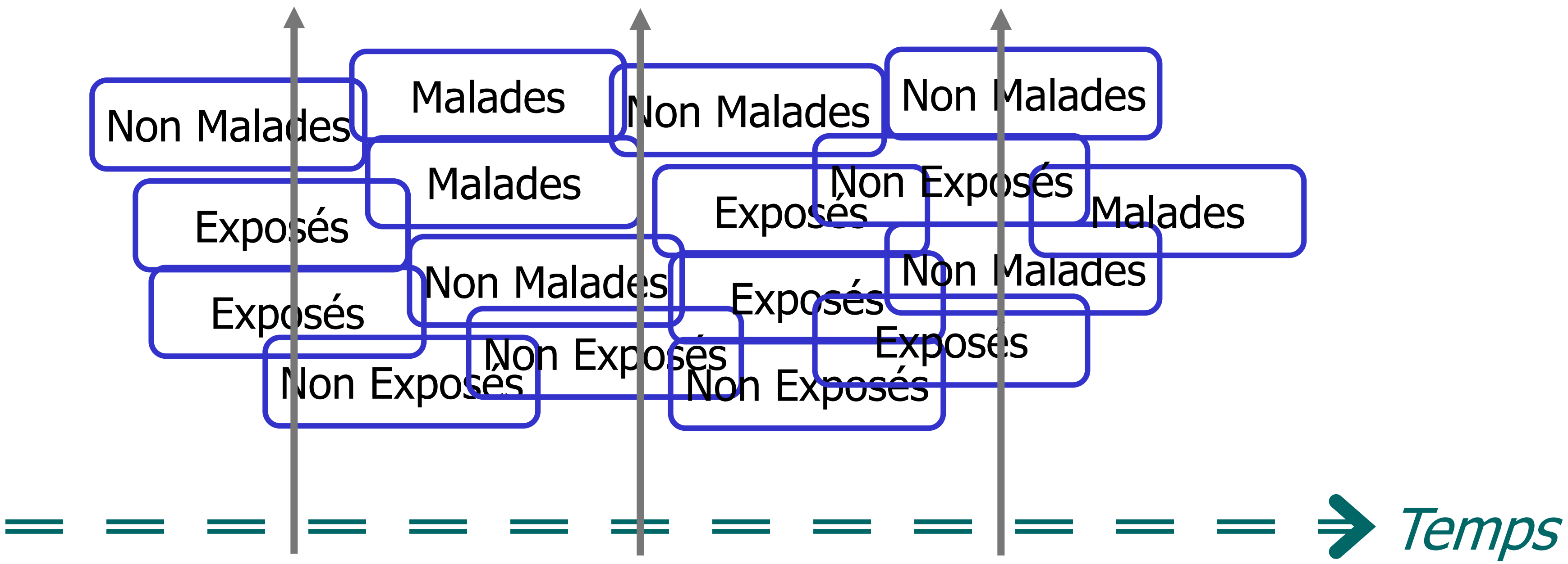
Types d'Études : Cohorte



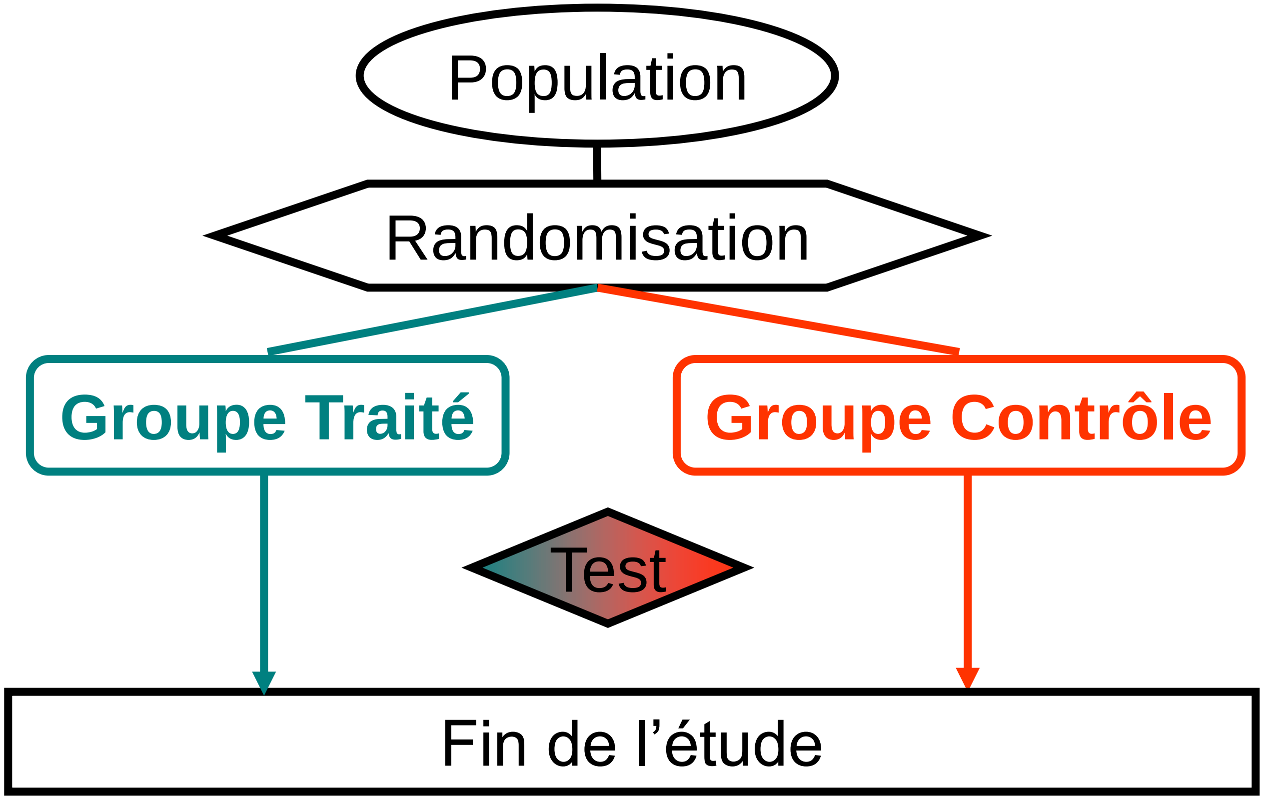
Types d'Études : Cas-Témoins



Types d'Études : Transversale



Randomisation



Randomisation
Répartition aléatoire des facteurs mesurables et non mesurables dans les deux groupes

Échantillon et Représentativité

- Méthode de sélection *a priori*
- Description des sujets de l'étude
- Critères d'inclusion et de non inclusion
- Écarts au protocole

Échantillonnage Aléatoire

- Chaque individus de la population a une chance égale de faire partie de l'échantillon (équiprobabilité)

- Échantillonnage simple
  - ✓ Tables de nombre au hasard
  - ✓ Générateur de nombres aléatoires

- Échantillonnage stratifié (ex : age, sexe, site, …)

Probabilités – Variables Aléatoires

Probabilités

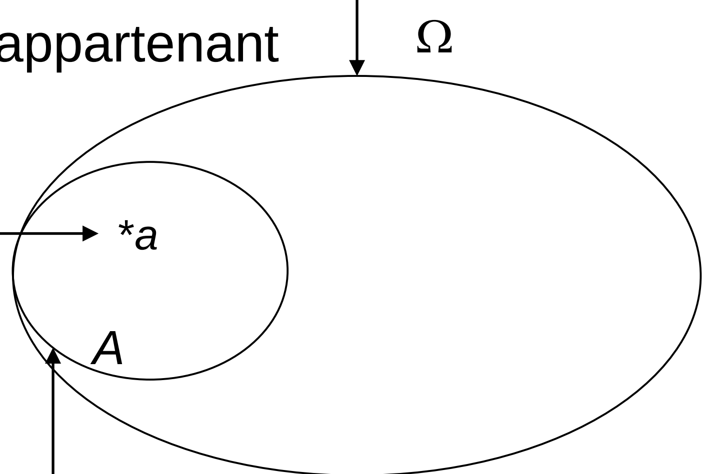
- Probabilité : modélise des phénomènes aléatoires dont les issues sont connues mais dont on ne peut en prédire la valeur car leur réalisation est incertaine

- Observation des issues d'un phénomène aléatoire sur des séries suffisamment grandes permet d'en déterminer leurs fréquences et par la suite la loi de distribution qui le dirige

Rappels sur les Ensembles

Ensemble : collection bien définie d'objets

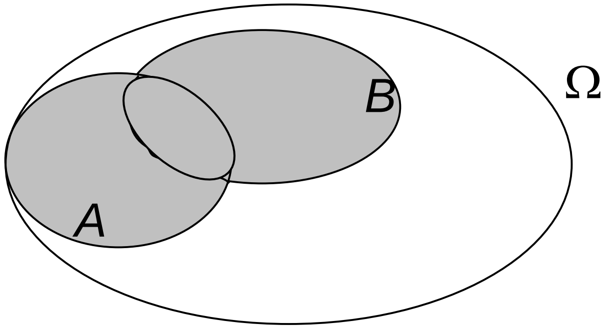
Elément : objet appartenant à l'ensemble



Sous-ensemble : *A* : ensemble d'éléments *a* appartenant tous à Ω

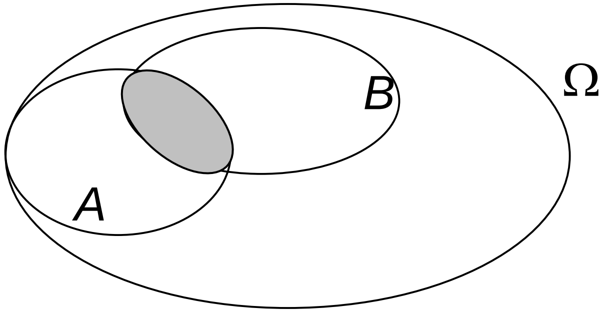
Rappels sur les Ensembles

Réunion : $A \cup B \Leftrightarrow A$ **ou** B

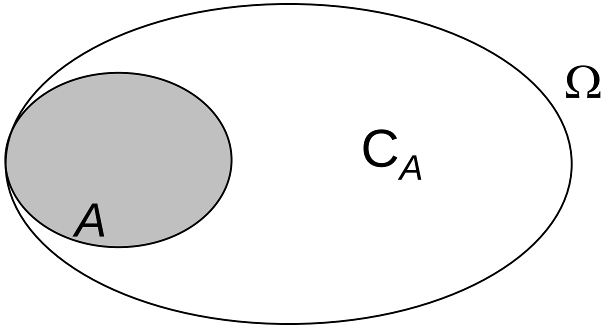


Intersection : $A \cap B \Leftrightarrow A$ **et** B

si $A \cap B = \varnothing$ alors A et B sont disjoints.



Complémentarité : C_A



Notion de Probabilité

- Probabilité : modélisation de phénomènes aléatoires

- Ensemble fondamental : Ω : ensemble des résultats possibles pour une expérience donnée (événements certains)

- Événement : c'est un sous-ensemble *A* de Ω, c'est-à-dire un ensemble de résultats. Un événement élémentaire est *a*

Exemple

Expérience aléatoire d'un jet de dé non pipé à 6 faces

Ensemble fondamental : $\Omega = \{f1, f2, f3, f4, f5, f6\}$

Événement A : face de nombre $\leq 2 = f1 \cup f2$

Événement B : face de nombre $\geq 5 = f5 \cup f6$

Événement C : face de nombre paire $\{2, 4, 6\} = f2 \cup f4 \cup f6$

$A \cup B = f1 \cup f2 \cup f5 \cup f6$, $A \cap B = \varnothing$

$A \cup C = f1 \cup f2 \cup f4 \cup f6$, $A \cap C \neq \varnothing$

Notion de Probabilité

Épreuve répétée n fois

	f1	f2	f3	f4	f5	f6	Total
Fréquences Absolues	n_1	n_2	n_3	n_4	n_5	n_6	n
Fréquences Relatives	n_1/n	n_2/n	n_3/n	n_4/n	n_5/n	n_6/n	1

$\text{Freq}(A)=(n_1+ n_2)/n$

$\text{Freq}(A\cup B)=(n_1+ n_2+ n_5+ n_6)/n=(n_1+ n_2)/n + (n_5+ n_6)/n = \text{Freq}(A) + \text{Freq}(B)$

$\text{Freq}(A\cup C)=(n_1+ n_2+ n_4+ n_6)/n \neq \text{Freq}(A) + \text{Freq}(C)$

Lorsque n $\rightarrow \infty$ la fréquence relative d'un événement tend vers la probabilité de cet événement

Probabilités Élémentaires

- Soit Ω un ensemble fondamental, P la fonction de probabilité qui à tout événement A associe un nombre réel positif ou nul. P(A) est appelée probabilité de l'événement A si :

  $P(A) \geq 0$

  $P(\Omega) = 1$

  si $A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B)$

  si $A_i \cap A_j = \varnothing \Rightarrow P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

  on en déduit que :

  - ✓ $P(\varnothing) = 0$
  - ✓ $P(A) \leq 1$
  - ✓ $P(C_A) = 1 - P(A)$
  - ✓ si $A \subset B$, alors $P(A) \leq P(B)$
  - ✓ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Probabilités Conditionnelles

- Exemple :
  - ✓ On s'intéresse au test de l'Hémocult dans le cadre du diagnostic du cancer colorectal
  - ✓ La probabilité d'avoir un cancer colorectal sachant le test Hémocult positif est une probabilité conditionnelle

$$P(CCR \setminus \text{Hémoc. Positif})$$

Probabilités Conditionnelles

- La probabilité de *A* sachant *B* est définie par

$$\boxed{\mathrm{P}(A \backslash B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}}$$

d'où $\mathrm{P}(A \cap B) = \mathrm{P}(A \backslash B)\mathrm{P}(B) = \mathrm{P}(B \backslash A)\mathrm{P}(A)$

et

$$\boxed{P(A \backslash B) = \frac{P(B \backslash A)P(A)}{P(B)}}$$

Théorème de Bayes

Indépendance en Probabilité

- A et B sont indépendants ssi

$$\boxed{\mathrm{P}(A \cap B) = \mathrm{P}(A).\mathrm{P}(B)}$$

- Si A et B sont indépendants et P(A) > 0, P(B) > 0, alors

$$\mathrm{P}(A \backslash B) = \mathrm{P}(A \cap B) \backslash \mathrm{P(B)} = \mathrm{P}(A).\mathrm{P}(B) \backslash \mathrm{P(B)}$$

- 2 événements disjoints de probabilités non nulles ne sont jamais indépendants
  - ✓ disjoints : $\mathrm{P}(\mathrm{A} \cap B) = 0$
  - ✓ indépendants : $\mathrm{P}(\mathrm{A} \cap B) = \mathrm{P(A).P(B)}$

Probabilités Conditionnelles (1)

- $A_1, \ldots, A_n$ événements formant une partition de Ω
- B un événement quelconque

Alors

$$P(B) = P(B \cap A_1) \cup P(B \cap A_2) \cup \ldots \cup P(B \cap A_n)$$

Et

$$\boxed{P(A_i \backslash B) = \frac{P(B \backslash A_i)P(A_i)}{P(B \backslash A_1)P(A_1) + \ldots + P(B \backslash A_n)P(A_n)}}$$

Formule développée de Bayes

Probabilités Conditionnelles (2)

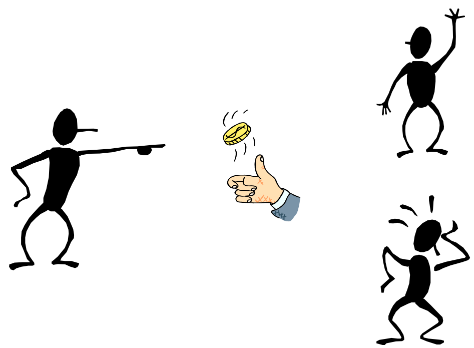
- Exemple :
  - ✓ La prévalence du SIDA dans une population est de 10 %
  - ✓ On sait qu'un test diagnostique est positif chez 95% des HIV^+ et qu'il est négatif chez 98% des HIV^-
  - ✓ Qu'elle est la probabilité d'être HIV^+ si le test est positif

P(HIV+) = 0,1
P(T+ \ HIV+) = 0,95
P(T+ \ HIV-) = 0,02

$$P\left(HIV+\backslash T+\right)=\frac{P\left(T+\backslash HIV+\right)P\left(HIV+\right)}{P\left(T+\backslash HIV+\right)P\left(HIV+\right)+P\left(T+\backslash HIV-\right)P\left(HIV-\right)}$$

P(HIV+\ T+) = (0,95*0,1)/(0,95*0,1 + 0,02*0,9) = 0,84

Variable Aléatoire



Si Pile, A gagne 1 F

Si Face, A perd 1 F

- Ω : {Pile, Face}
- P(Pile) = P(Face) = 0,5
- G : gain de A; G = +1, si Pile; G = -1, si Face
- P(G = +1) = P(G = -1) = 0,5
- Distribution de G : {(+1; 0,5), (-1; 0,5)}

G : variable aléatoire qui suit une certaine loi de probabilité

Variable Aléatoire : Définition

- Soit E un ensemble d'événements
- d'ensemble fondamental Ω fini, et
- *a* un événement élémentaire de E

Pour tout événement *a* appartenant à E on fait correspondre un nombre *x* (variable aléatoire) selon une loi bien définie

Variable Aléatoire : Exemple

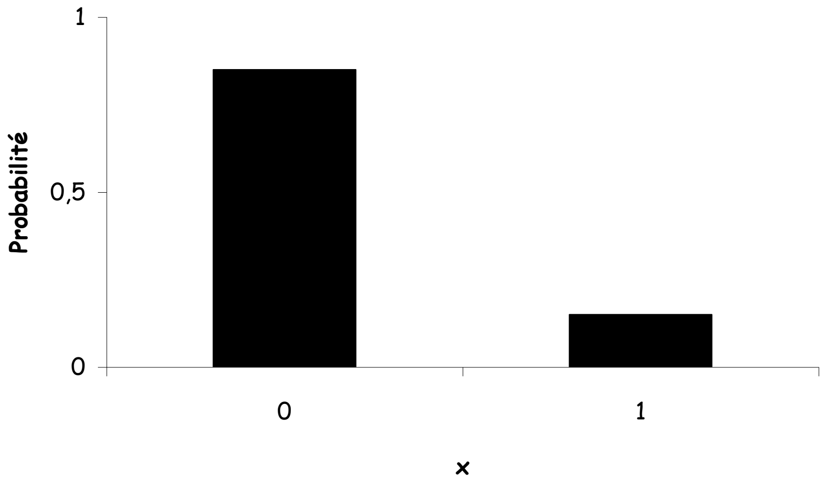
- Soit une maladie M pour laquelle il est nécessaire de débuter un TRT avant confirmation du diagnostic. Le médicament utilisé est cependant connu pour entraîner des effets indésirables.
- On sait que : $P(M^+)=5\%$; $P(EI^+ \backslash M+)=30\%$; $P(EI^- \backslash M^-)=85\%$

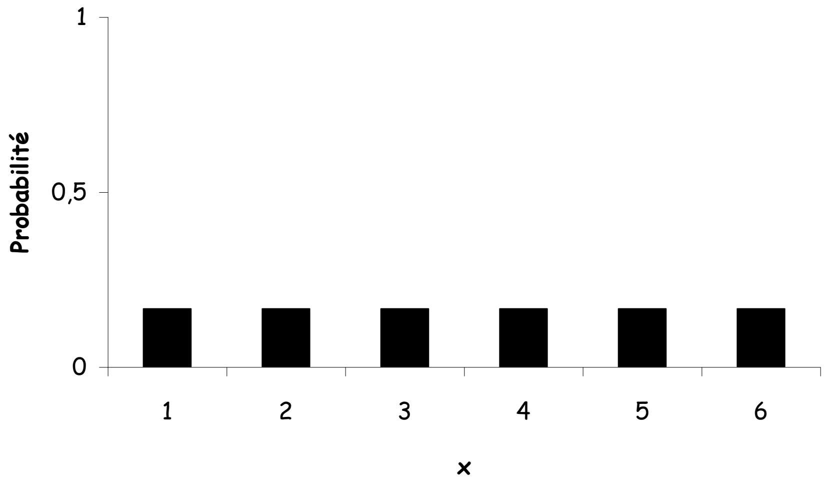
	M^+	M^-
EI^+	$P(EI^+ \cap M^+) = 0{,}3 x 0{,}05 = 1{,}5\%$ $X = 1$	$P(EI^+ \cap M^-) = (1-0{,}85) x (1-0{,}05) = 14{,}3\%$ $X = 1$
EI^-	$P(EI^- \cap M^+) = (1-0{,}3) x 0{,}05 = 3{,}5\%$ $X = 0$	$P(EI^- \cap M^-) = 0{,}85 x (1-0{,}05) = 80{,}8\%$ $X = 0$

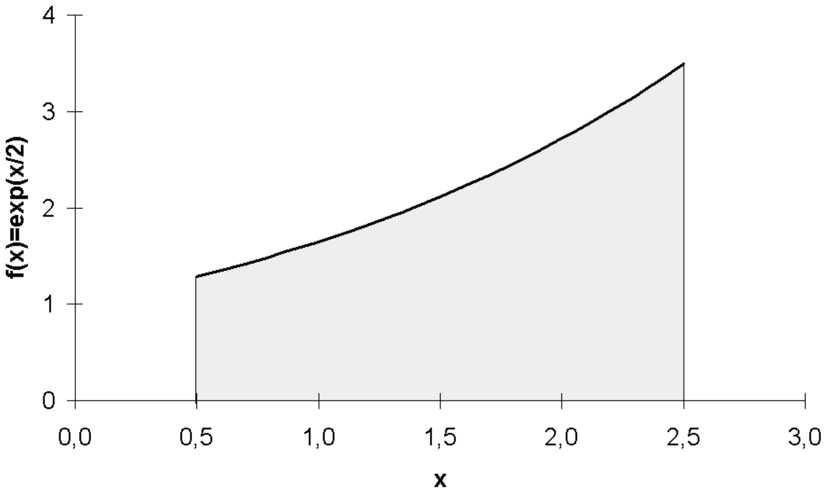
où X est une v.a. indicatrice des EI.

La distribution de X est : {(0; 0,84), (1; 0,16)}

Caractéristiques d'une Variable Aléatoire







Caractéristique de Position : Moyenne, Espérance

- Variable discrète X
  - ✓ soit X une va prenant les valeurs $x_1, x_2, \ldots, x_n$ avec les probabilités $p_1, p_2, \ldots, p_n$ et $\Sigma p_i = 1$, i = 1, …, n :

$$\boxed{\mu = E(X) = \sum p_i x_i}$$

- Cas d'une variable continue X
  - ✓ définie par une loi de densité $f(x)$

$$\boxed{\mu = E(X) = \int_a^b f(x)dx}$$

Caractéristique de Position :

Moyenne, Espérance

- Exemple 1 : μ = (p x 1) + (q x 0) = p

- Exemple 2 : μ = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 3,5

- Exemple 3 : $\mu = E(X) = \int_{0,5}^{2,5} f(x)dx = \int_{0,5}^{2,5} \exp\left(x/2\right)dx = \left[\exp\left(x/2\right)\right]_{0,5}^{2,5} = 2,21$

Caractéristique de Dispersion :

Variance, Écart-type

- Cas d'une variable discrète X :

$$\boxed{\begin{aligned}\sigma^2 &= \sum p_i \left[x_i - \mu\right]^2 \\ &= E\left(\left(X-\mu\right)^2\right) = E\left(X^2\right) - \left[E\left(X\right)\right]^2\end{aligned}}$$

- Cas d'une variable continue X :

$$\boxed{\sigma^2 = \int_a^b \left(x-\mu\right)^2 f\left(x\right)dx}$$

σ ² = variance

σ = écart-type

Caractéristique de Dispersion :

Variance, Écart-type

- Exemple 1 : $\sigma^2 = p \times (1 - p)^2 + q \times (0 - p)^2 = pq$

- Exemple 2 : $\sigma^2 = 1/6[(1 - 3{,}5)^2 + (2 - 3{,}5)^2 + ... + (6 - 3{,}5)^2]$

  $= 2{,}9$

- Exemple 3 : $\sigma 2 = \int_{0.5}^{2,5} (x - 2{,}21)^2 \exp\left(\frac{x}{2}\right) dx = \int_{0.5}^{2,5} x^2 \exp\left(\frac{x}{2}\right) dx - 2{,}21^2 = 0{,}68$

Estimation

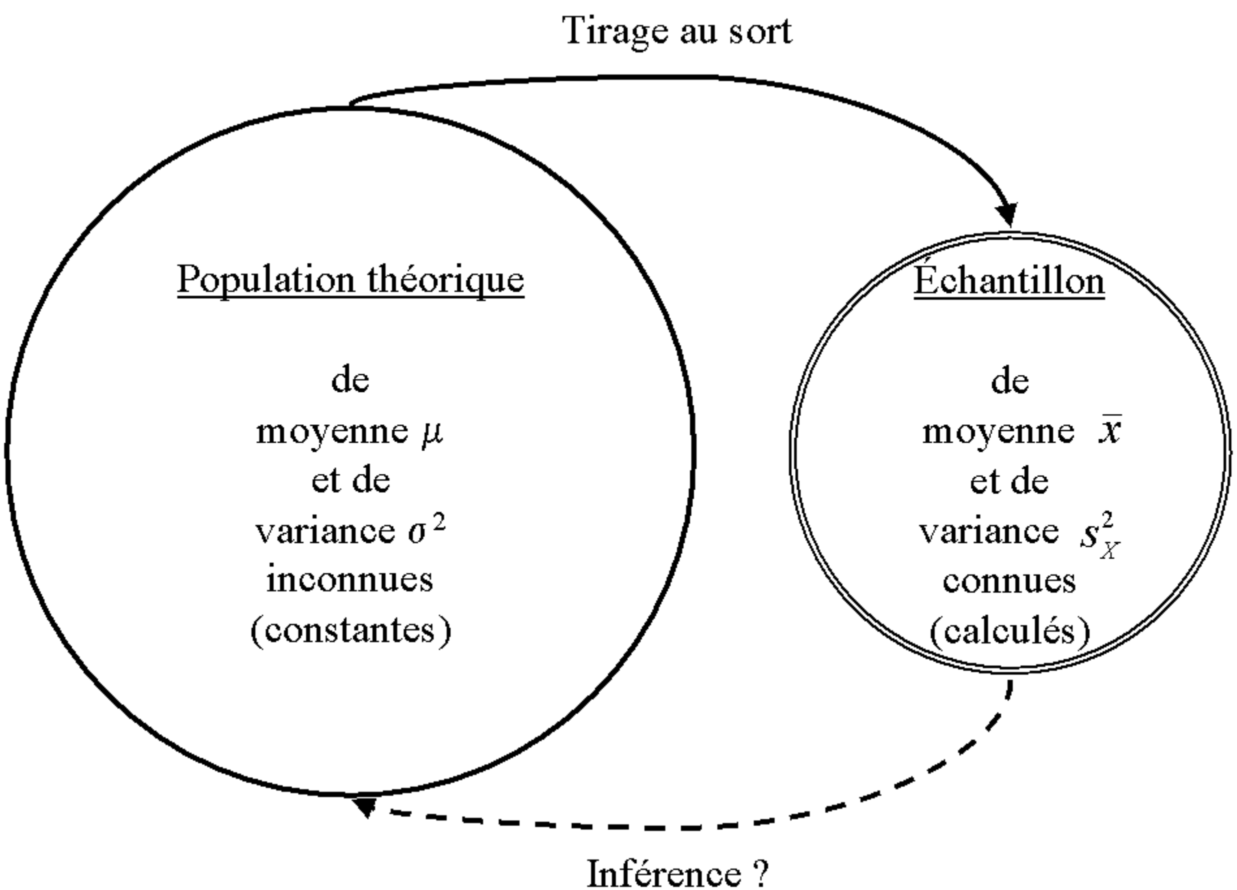
Introduction

- Connaître des valeurs de certaines grandeurs grâce à des observations réalisées sur un échantillon
  - ✓ Fréquence de la survenue du mélanome malin ?
  - ✓ Fréquence des infections nosocomiales ?
  - ✓ Valeur de la glycémie d'un patient ?
  - ✓ Variance de la glycémie mesurée chez ce patient ?

- Valeur la plus vraisemblable : estimation ponctuelle

- Intervalle de valeurs possibles, compatibles avec les observations : intervalle de confiance

Estimateur - Estimation



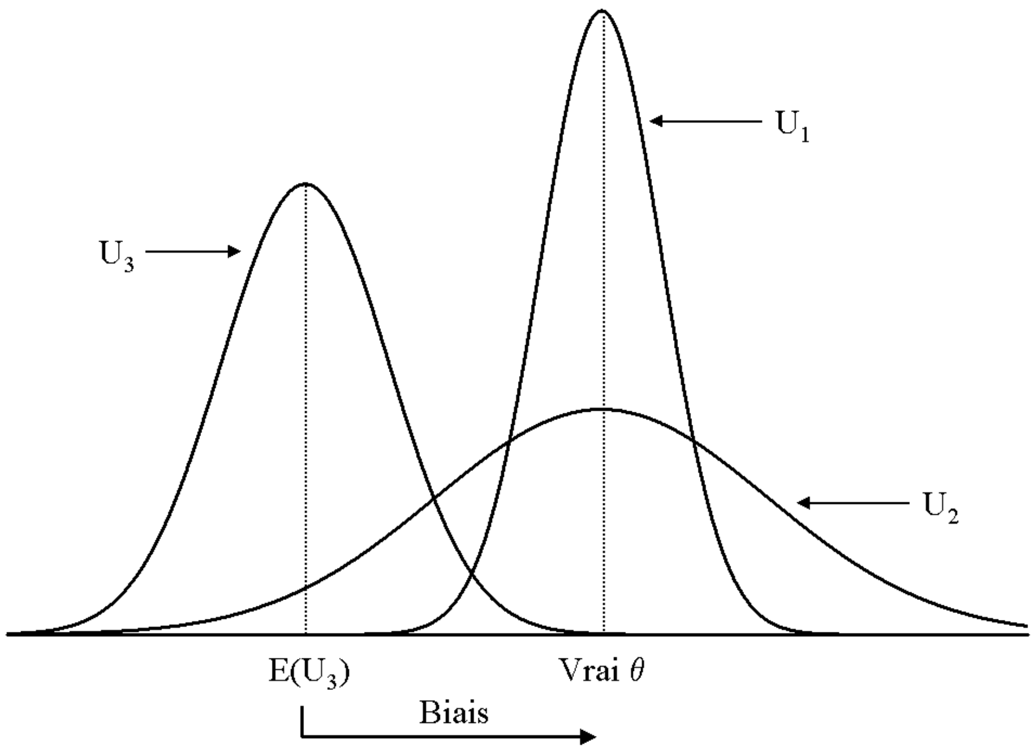
Estimateur

↓

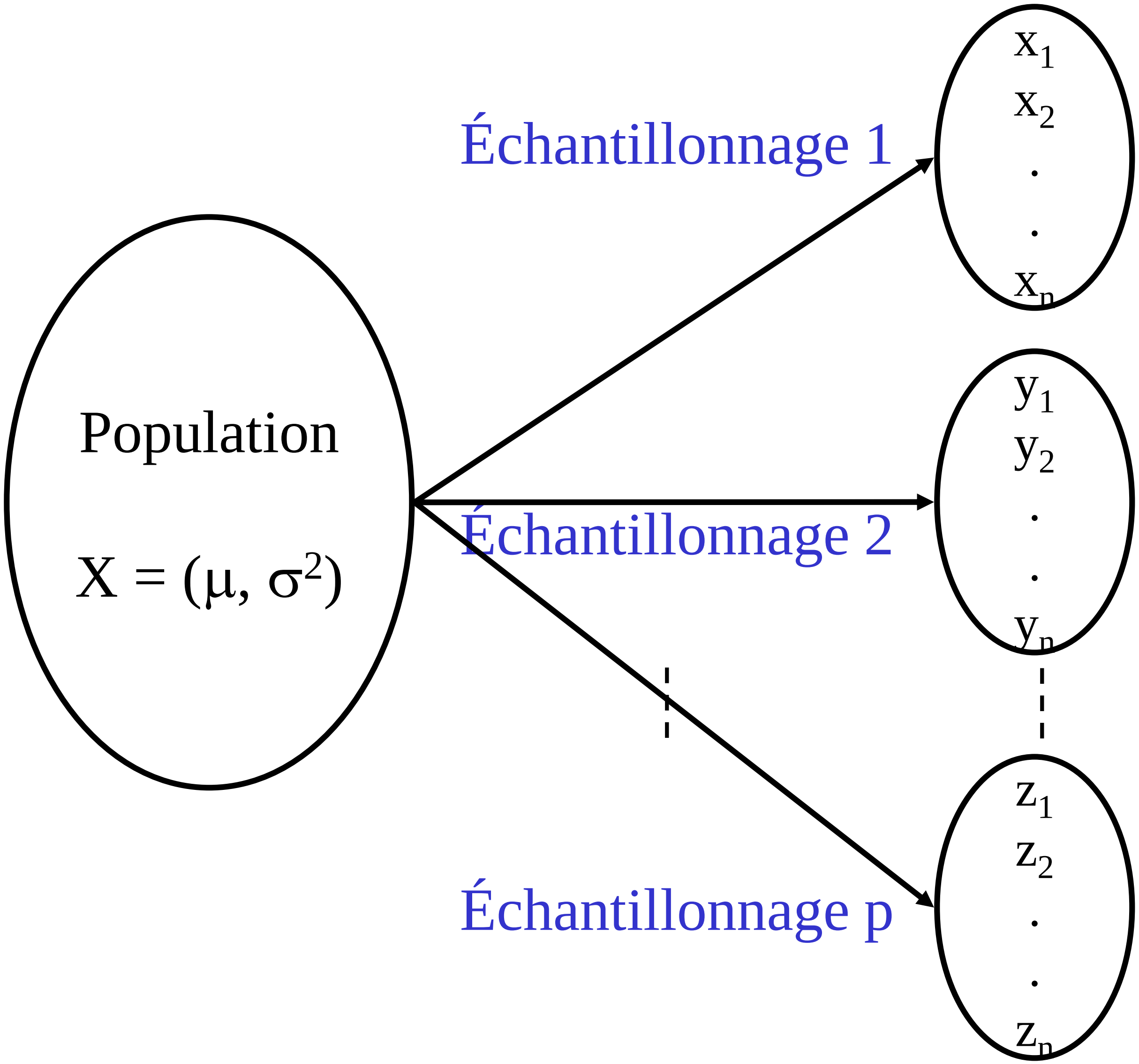
Estimation

Qualité d'un Estimateur

- U : estimateur sans biais de θ si E(U) = θ
- U : estimateur biaisé de θ si E(U) $\neq \theta$ et le biais = E(U) - θ



Estimation de la Moyenne d'une Population (1)



$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\bar{z} = \frac{z_1 + z_2 + \ldots + z_n}{n} = \frac{\sum_{i=1}^{n} z_i}{n}$$

Estimation de la Moyenne d'une Population (2)

$$\boxed{\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}}$$

$\bar{X}$ est l'estimateur de μ

C'est un estimateur sans biais car $E\left(\bar{X}\right) = \mu$

$$E\left(\bar{X}\right) = \frac{1}{n} E\left(X_1 + X_2 + \ldots + X_n\right) = \frac{1}{n} E\left(X_1\right) + E\left(X_2\right) + \ldots + E\left(X_n\right)$$

Comme l'espérance de la moyenne arithmétique des n variables est la moyenne arithmétique des espérances

$E\left(\bar{X}\right) = \mu$ et est **sans biais**

Estimation de la Variance d'une Population

- Soit $x_1, x_2, \ldots, x_n$ un échantillon tiré au hasard, d'effectif n et de moyenne $\bar{x} = \sum x_i / n$

- L'estimation de la variance de la population est

$$\boxed{s_x^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}}$$

- S_x^2 est un estimateur sans biais convergent de σ^2
- s_x est une bonne estimation de l'écart-type de la population

Estimation d'une Proportion d'une Population

- Soit *k* le nombre de fois où un caractère donné est présent dans un échantillon tiré au hasard d'effectif *n*
- Soit *p* la proportion inconnue du caractère étudié dans la population
- Fréquence *f* du caractère étudié dans l'échantillon

$$\boxed{f = \frac{k}{n}}$$

- On montre que E(F) = p

Estimation de la Variance d'une Proportion

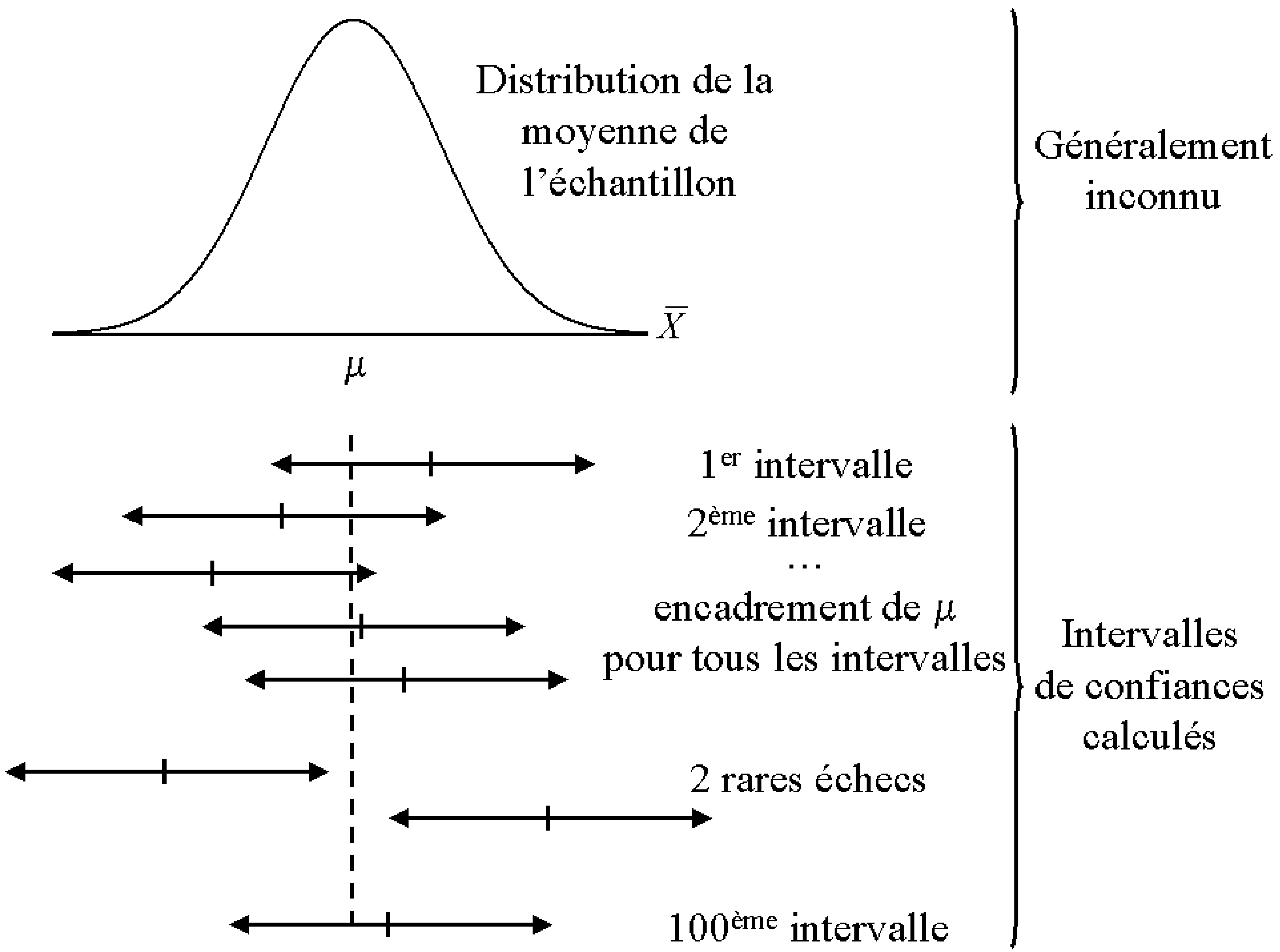
$$\boxed{Var\left(F\right)=\frac{p\cdot\left(1-p\right)}{n}}$$

- F est un estimateur convergent de p
- On estime la variance p.(1 – p)/n par *f* .(1-*f*)/n

Estimation par Intervalle (1)

- Soit Σ un paramètre inconnu (une moyenne ou une proportion) estimé à partir d'un échantillon au hasard par θ
- On souhaite avoir un degré de confiance acceptable comme quoi θ approche bien Σ
- Intervalle de confiance : déterminé à partir des données d'un échantillon dans lequel on peut parier, avec un risque de se tromper qui soit acceptable, que se situe réellement Σ dans la population

Estimation par Intervalle (2)



Estimation par Intervalle (3)

- Risque (α) généralement = 0,05 ; il correspond aux erreurs d'échantillonnages jugées acceptables
- Intervalle de confiance de Σ est de la forme

  θ - erreur d'échantillonnage ; θ + erreur d'échantillonnage
- Interprétation
  - ✓ On accepte qu'il y ait α.100 chances sur cent de se tromper en disant que Σ appartient à l'intervalle
  - ✓ On accepte qu'il y ait (1-α).100 chances sur cent de ne pas se tromper en disant que Σ appartient à l'intervalle

Intervalle de Confiance d'une Moyenne

$$\boxed{\bar{x} - L_{\alpha} \cdot \frac{s_x}{\sqrt{n}} \; ; \; \bar{x} + L_{\alpha} \cdot \frac{s_x}{\sqrt{n}}}$$

- Si $n \geq 30$ alors, $L_{\alpha} = N_{\alpha}$ ($N_{\alpha=5\%} = 1{,}96$)
- Si $n < 30$ et la loi de distribution de la variable dans la population est Normale alors, $L_{\alpha} = t_{\alpha,\nu}$

Intervalle de Confiance d'une Proportion

- Si f = k/n n'est pas voisin de 1 ou de 0
- Si $f*n \geq 5$ et $(1 - f)*n \geq 5$

$$\boxed{f - N_\alpha \cdot \sqrt{\frac{f \cdot (1 - f)}{n}} \;\; ; \;\; f + N_\alpha \cdot \sqrt{\frac{f \cdot (1 - f)}{n}}}$$

Estimation de la Prévalence

- Mesure du risque de maladie dans la population
- Proportion de malades présents (M^+) dans la population (N) à un instant donné

$$\boxed{P = \frac{M^+}{N}}$$

- ✓ C'est une probabilité
- ✓ Intègre la notion de durée de la maladie (↗ durée maladie $\Rightarrow$ ↗ nombre de M^+ $\Rightarrow$ ↗ prévalence)
- ✓ Intègre la notion de vitesse d'apparition des nouveaux M^+ (↗ vitesse d'apparition $\Rightarrow$ ↗ prévalence)

Estimation de l'Incidence

- Quantifie la production de nouveaux cas de maladie dans la population dans un certain intervalle de temps

$$\boxed{I = \frac{\text{nb nouveaux cas pendant } \Delta t}{N \cdot \Delta t}}$$

Estimation de l’Effet d’un Facteur Pronostique

- Risque = probabilité d’apparition d’un événement, d’une maladie
  - ✓ Risque d’infarctus du myocarde
  - ✓ Risque de récidive d’un cancer après rémission

- La probabilité d’apparition est-elle modifiée par la présence ou l’absence d’un facteur (pronostique) ?

Risque Relatif

- Considérons le cas où la maladie (M) est présente (M+) ou absente (M-) avec un seul facteur (F) qui peut être présent (F+) ou absent (F-)
  - ✓ Risque d'apparition de la maladie chez les exposés au facteur F : P(M+/F+)
  - ✓ Risque d'apparition de la maladie chez les non exposés au facteur F : P(M+/F-)
- Risque relatif : indicateur de l'influence du facteur

$$\boxed{RR = \frac{P(M+/F+)}{P(M+/F-)}}$$

Le RR varie de 0 à l'infini

Interprétation du Risque Relatif

- Si RR > 1
  - ✓ $\Leftrightarrow$ P(M+/F+) > P(M+/F-)
  - ✓ La présence du facteur F « favorise la maladie » = facteur de risque
- Si RR < 1
  - ✓ $\Leftrightarrow$ P(M+/F+) < P(M+/F-)
  - ✓ La présence du facteur F « favorise la non maladie » = facteur protecteur
- Si RR = 1
  - ✓ $\Leftrightarrow$ P(M+/F+) = P(M+/F-)
  - ✓ Le facteur F n'a pas d'effet sur la maladie

Estimation du Risque Relatif (1)

	M+	M-	Total
F+	a	b	m_1
F-	c	d	m_2
Total	n_1	n_2	N

$$P\left(M+/F+\right) = a/m_1$$

$$P\left(M+/F-\right) = c/m_2$$

$$\boxed{RR = \frac{a/m_1}{c/m_2}}$$

Estimation du Risque Relatif (2)

- Le RR peut être estimé
  - ✓ Dans les enquêtes sur un seul échantillon au hasard
  - ✓ Dans les enquêtes exposés / non exposés
- Le RR ne peut pas être estimé
  - ✓ Dans les enquêtes cas-témoins. Dans ce cas on peut estimer un Odds Ratio, (a.d)/(b.c), qui s'interprète qualitativement comme le RR par rapport à 1
- En plus de l'estimation ponctuelle du RR il est nécessaire de calculer son intervalle de confiance