Fluctuations d'échantillonnage
Estimations ponctuelles et par intervalle de confiance

C. Bardel

PASS
Septembre 2024

Plan du cours

Statistiques descriptives

Exemple : étude du poids des nouveaux nés dans un hôpital donné

- Une **population**
  - ex : bébés nés à l'hôpital FME en 2022
- Une **variable d'étude**
  - le poids
- Analyse des **résultats**
  - calculs de la moyenne, de l'écart-type, etc...
    Description de cet ensemble d'individus

Conclusion de l'étude

Informations sur le poids des nouveaux nés **dans cet hôpital**, **en 2022**
Conclusions valables pour la population étudiée seulement

Statistiques inférentielles

Exemple : poids des nouveaux-nés en France

- Un **échantillon**
  - Choix aléatoire de n nouveaux nés en France
- Une **variable d'étude**
  - le poids
- Analyse des **résultats**
  - analyser l'échantillon pour inférer des résultats valables pour la population des nouveaux-nés en France $\rightarrow$ réaliser des estimations

Conclusion

Estimation du poids moyen des nouveaux nés **en France** et de sa variabilité

Analyse d'un échantillon mais conclusions valables pour la population

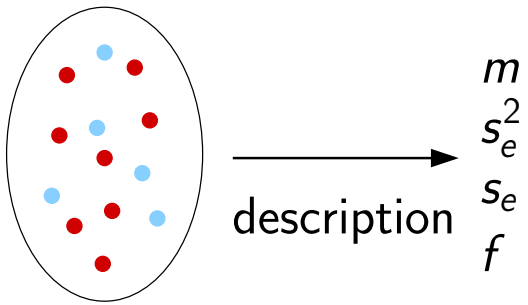
Autres exemples

Les sondages (marketing ou politique), estimation de la prévalence d'une maladie, estimation de la teneur en principe actif des comprimés produits par une machine, ...

Statistiques descriptives et statistiques inférentielles

Bilan

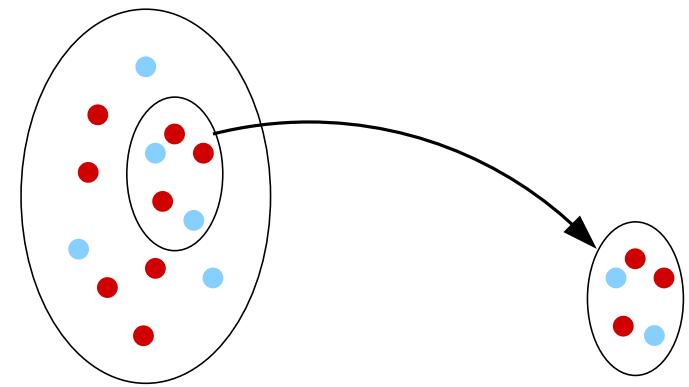
Statistiques descriptives



Population

"Echantillon" (description)

Statistiques inférentielles



Population — Échantillon

μ	estimation	m
σ^2		s^2
σ		s
π		f

- Estimations, intervalles de confiance
- Tests statistiques

Plan du cours

Problèmes liés à l'échantillonnage

Obtenir un échantillon représentatif

Échantillons non représentatifs

- les étudiants en biologie à Lyon / l'ensemble des étudiants français
- les 20 premières souris attrapées dans leur cage / l'ensemble des souris de laboratoire
- ...

Échantillon représentatif

- Réaliser un tirage aléatoire

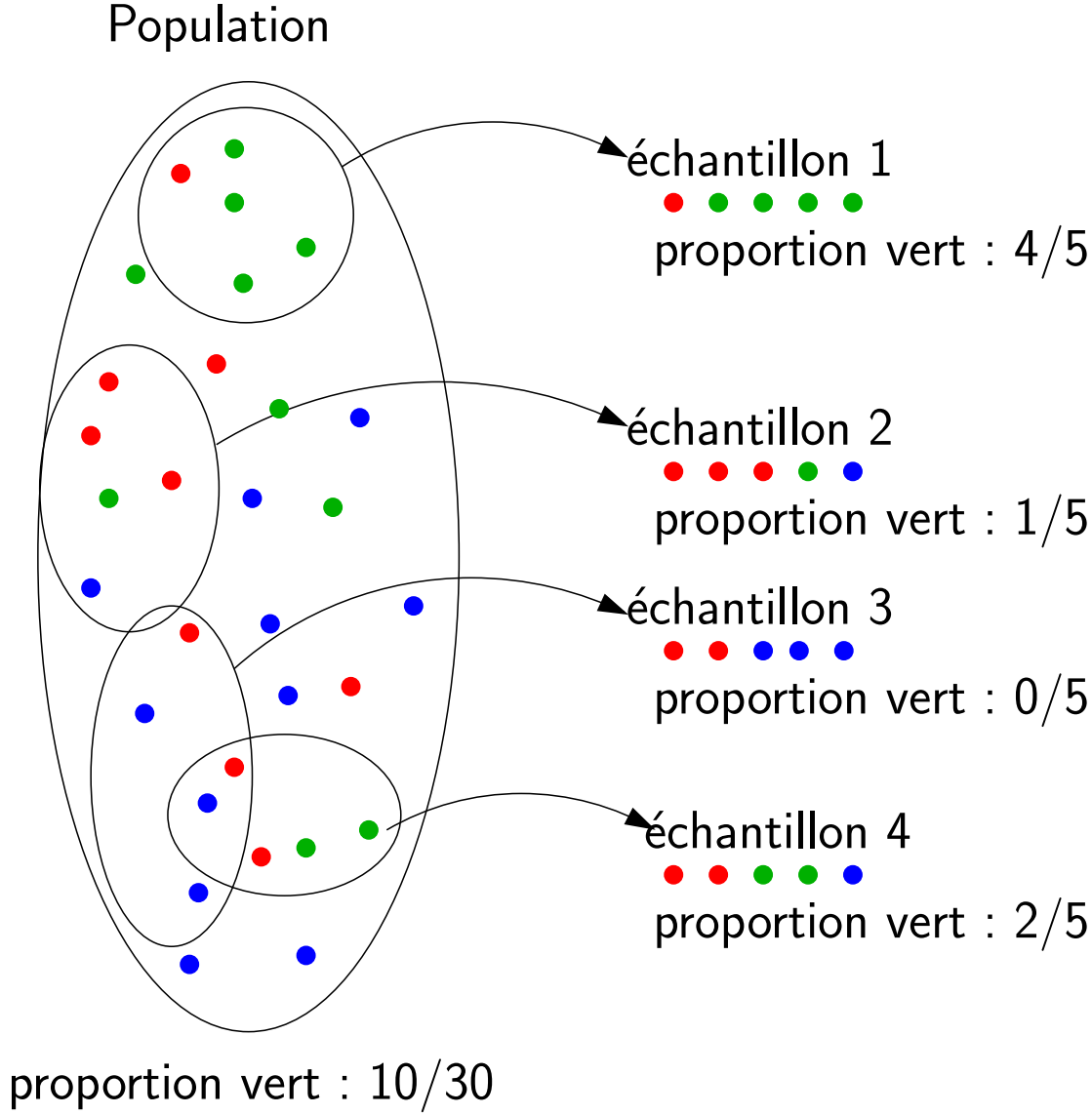
Définition : échantillon aléatoire simple

Tous les individus

- ont la **même probabilité** d'être choisis (tirage aléatoire)
- sont choisis de façon **indépendante** les uns des autres
  - tirage avec remise ou tirage dans une population de grande taille par rapport à celle de l'échantillon

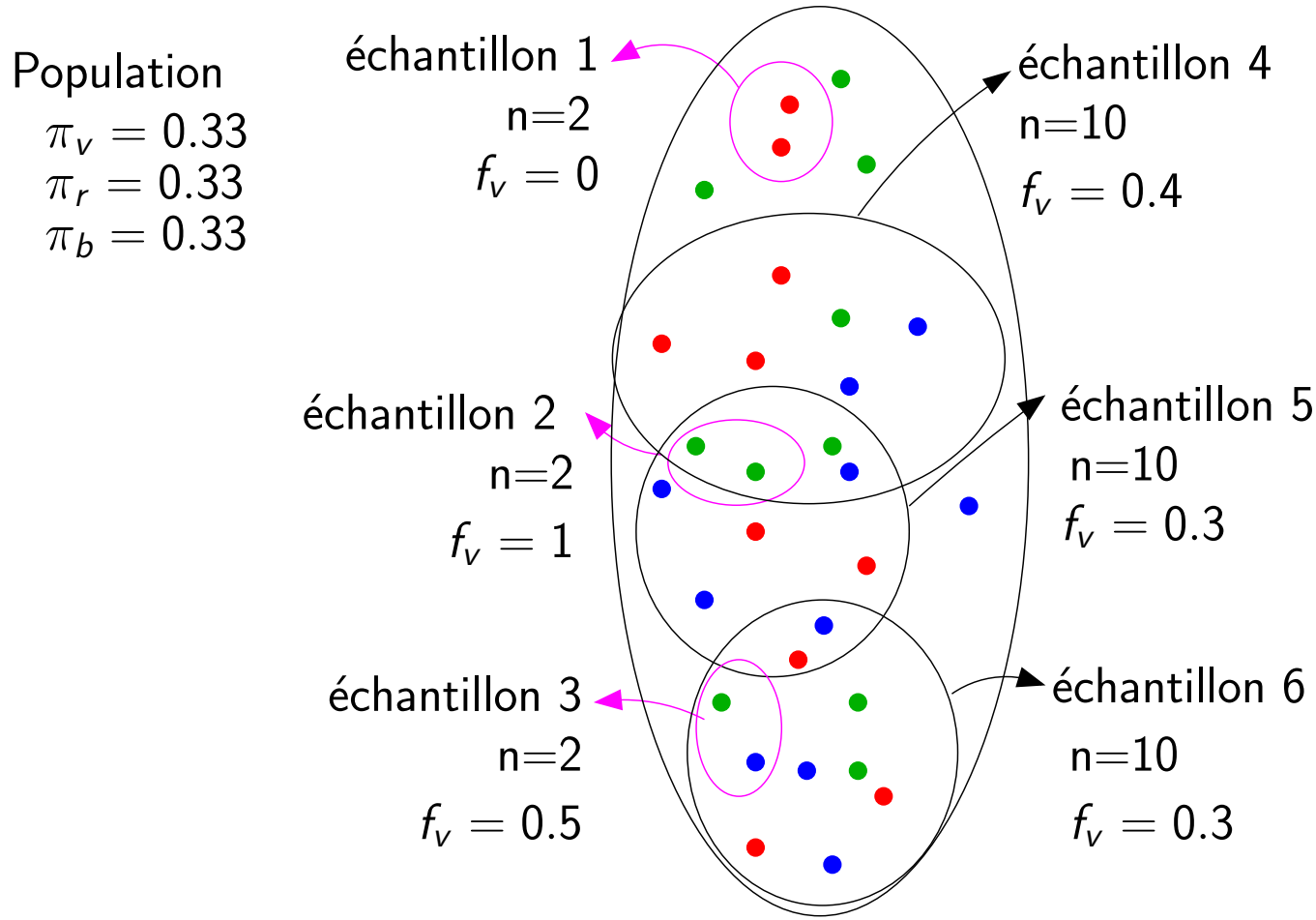
Dans le cadre du PASS, on se placera toujours dans ce cas

Fluctuations d'échantillonnage



Proportions observées dans les échantillons : estimations ponctuelles de la vraie proportion dans la population

Importance de la taille de l'échantillon



Plus l'échantillon est de **grande taille**, plus la **variabilité** des proportions observées est **faible** : l'estimation sera plus **précise**

Si l'échantillon est trop grand par rapport à la taille de la population : problème d'individus **non indépendants**

Formalisation du problème : notion d'échantillon statistique

Exemple

Soit X la va modélisant le poids (kg) des nouveaux-nés en France

$$X \to \mathcal{L}(\mu, \sigma) \quad \mu : \text{poids moyen} \quad \sigma : \text{ecart-type du poids}$$

Exemple d'échantillon : (2,8 ; 4,2 ; 3,8 ; 3,2 ; ...)

- $x_1 = 2.8$ est une réalisation de la v.a. X_1
  X_1 : « première valeur observée sur un échantillon », $X_1 \to \mathcal{L}(\mu, \sigma)$
- $x_2 = 4.2$ est une réalisation de la v.a. X_2
  X_2 « deuxième valeur observée sur un échantillon ». $X_2 \to \mathcal{L}(\mu, \sigma)$
- ...

Conclusion

- Chaque **valeur observée** sur un échantillon correspond à la **réalisation d'une v.a** X_i.
- Les X_i sont **indépendantes** et toutes de **même loi** que X

Notion d'échantillon statistique (2)

Échantillon statistique

On ne considère plus des individus mais des **variables aléatoires**

- échantillon statistique de taille n : ensemble de n va $(X_1; X_2; \ldots; X_n)$
- valeurs observées de l'échantillon : $(x_1; x_2; \ldots; x_n)$

Plan du cours

Estimateur et estimations ponctuelles

Notation

- $\theta =$ le paramètre à estimer
- θ : μ, σ, π, médiane, ...

Estimateur

Un estimateur de θ est une variable aléatoire exprimée en fonction des va d'échantillon X_i : $T = f(X_1; X_2; \ldots; X_n)$

Exemple : cas de M, l'estimateur de μ

$$M = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Estimation

Une estimation de θ est un nombre calculé en fonction des valeurs x_i observées dans l'echantillon : $t = f(x_1; x_2; \ldots; x_n)$

Exemple : cas de l'estimation de μ

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Ne pas confondre estimation, estimateur et le paramètre à estimer !

Lien entre estimateur et estimation

Le paramètre à estimer θ est **inconnu**

Soit T un estimateur de θ

T est une variable aléatoire

Un échantillon $\rightarrow$ calcul d'une valeur t, réalisation de T

t est une estimation de θ calculée à partir d'un échantillon donné

Notations

paramètre théorique	estimateur	estimation
μ	M	m
σ^2	S^2	s^2
σ	S	s
π	F	f

Attention : changement de notation. La proportion théorique est maintenant notée π et non plus p.

Qualité d'un estimateur (1)

Biais d'un estimateur

- Estimateur non biaisé : $\mathrm{E}(T) = \theta$
- Biais d'un estimateur $= \mathrm{E}(T) - \theta$

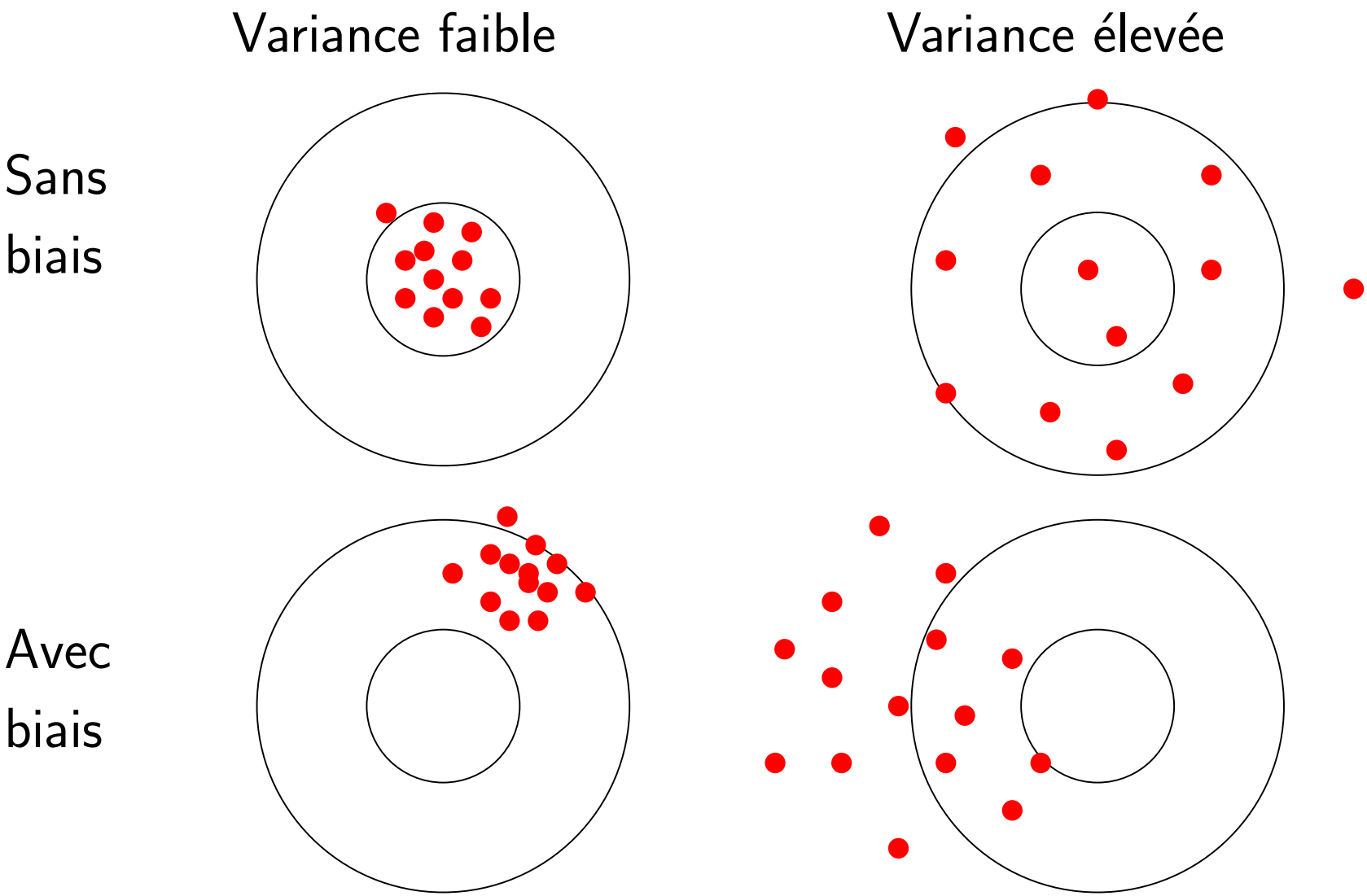
Variance d'un estimateur

Si $\mathrm{var}(T)$ est faible : les estimations sont peu dispersées

Erreur quadratique moyenne

- Erreur quadratique moyenne $= EQM$
- On peut montrer que $EQM = \mathrm{var}(T) + biais^2$
- Un bon estimateur a une EQM la plus faible possible
  Idéalement, $EQM \xrightarrow[n \to +\infty]{} 0$

Qualité d'un estimateur (2)



Estimateur de l'espérance μ

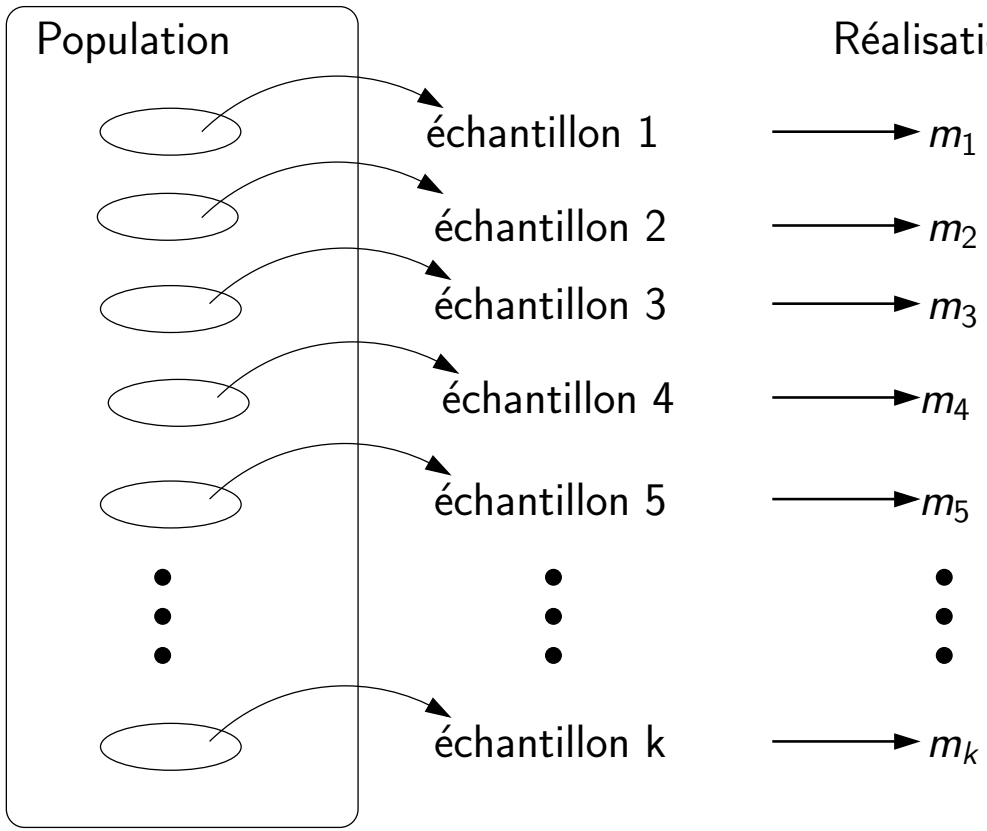
Moyenne d'échantillon ou moyenne empirique

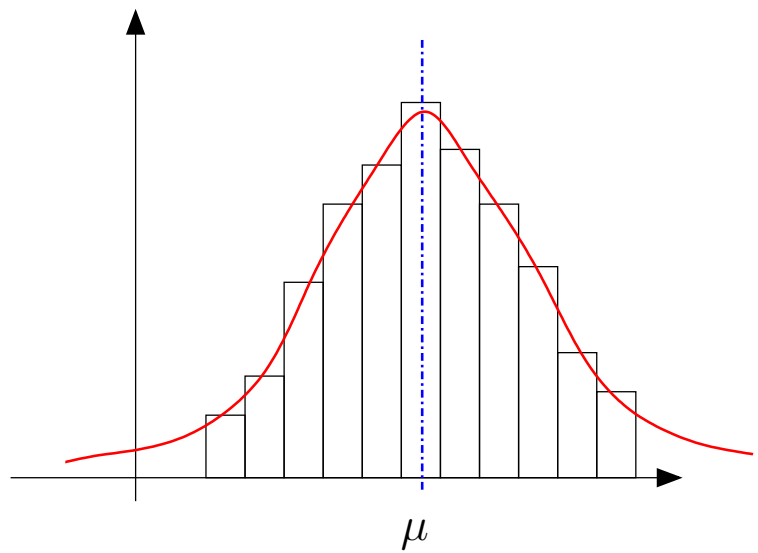
M : **moyenne d'échantillon** ou **moyenne empirique** (parfois notée $\bar{X}$)

$$M = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Rappel : M et les X_i sont des variables aléatoires

Loi de M, estimateur de l'espérance : approche intuitive





n grand (≥ 30)

Rappel : n est le taille des échantillons

Loi de $M = \frac{\sum_i X_i}{n}$

Les différents cas

Plusieurs cas :

- Si les X_i suivent une loi normale
- Si les X_i ne suivent pas une loi normale
  - Si $n \geq 30$
  - Si $n < 30$

Cas 1 : les X_i suivent une loi normale

M est une combinaison linéaire de va Gaussiennes

$$\boxed{M \to \mathcal{N}(\mu_M, \sigma_M)}$$

Cas 2 : les X_i ne suivent pas une loi normale

- Si $n \geq 30$: application du TCL (cf cours Proba/VA)

$$\boxed{M \rightsquigarrow \mathcal{N}(\mu_M, \sigma_M)}$$

- Si $n < 30$: on ne peut rien dire

Expression de μ_M et σ_M (1)

Rappel du cours de probabilités

$$\mathrm{E}(aX + bY) = a \times \mathrm{E}(X) + b \times \mathrm{E}(Y)$$

Espérance $\mathrm{E}(M) = \mu_M$ en fonction de $\mathrm{E}(X) = \mu$

$$\mathrm{E}(M) = \mathrm{E}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$$

$$\mathrm{E}(M) = \frac{1}{n} \times \mathrm{E}(X_1 + X_2 + \ldots + X_n)$$

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

$$\boxed{\mathrm{E}(M) = \frac{1}{n} \times n\mu = \mu}$$

Expression de μ_M et σ_M

Rappel du cours de probabilités

Si X et Y sont 2 variables aléatoires **indépendantes** :

$$\text{var}(aX + bY) = a^2 \times \text{var}(X) + b^2 \times \text{var}(Y)$$

Variance $\text{var}(M) = \sigma_M^2$ en fonction de $\text{var}(X) = \sigma^2$

$$\text{var}(M) = \text{var}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$$

$$\text{var}(M) = \frac{1}{n^2} \times \text{var}(X_1 + X_2 + \ldots + X_n)$$

$$\text{var}(M) = \frac{1}{n^2} \times (\text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n)) \quad \text{indépendance}$$

$$\boxed{\text{var}(M) = \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n}}$$

Qualité de l'estimateur M

Rappel

$$\mathrm{E}(M) = \mu \quad \text{et} \quad \mathrm{var}(M) = \frac{\sigma^2}{n} \quad \text{et} \quad \sigma_M = \frac{\sigma}{\sqrt{n}} \; (SEM)$$

SEM=Standard Error of the Mean

Absence de biais

$\text{Biais} = \mathrm{E}(M) - \mu = 0$

Erreur quadratique moyenne

$EQM = \mathrm{var}(M) + biais^2$

$EQM = \frac{\sigma^2}{n} + 0$

$\lim_{n \to +\infty} EQM = 0$

Conclusion

M est un « bon » estimateur de l'espérance

Intervalle de fluctuation (= intervalle de pari) de M

Principe

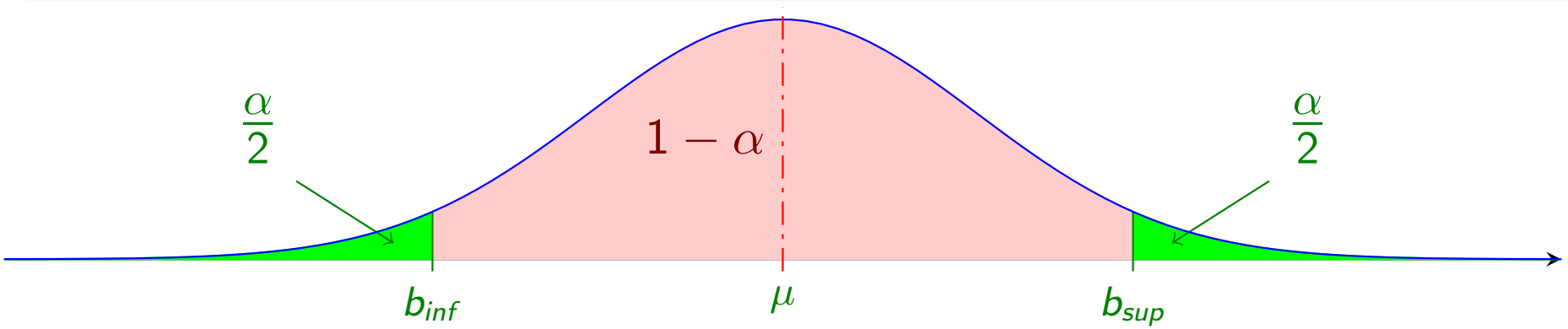
- On se place dans le cas ou $M \to (ou \rightsquigarrow)$ $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

  La loi de M est donc connue

- On cherche un intervalle centré sur μ qui a une probabilité $1-\alpha$ de contenir une réalisation de M

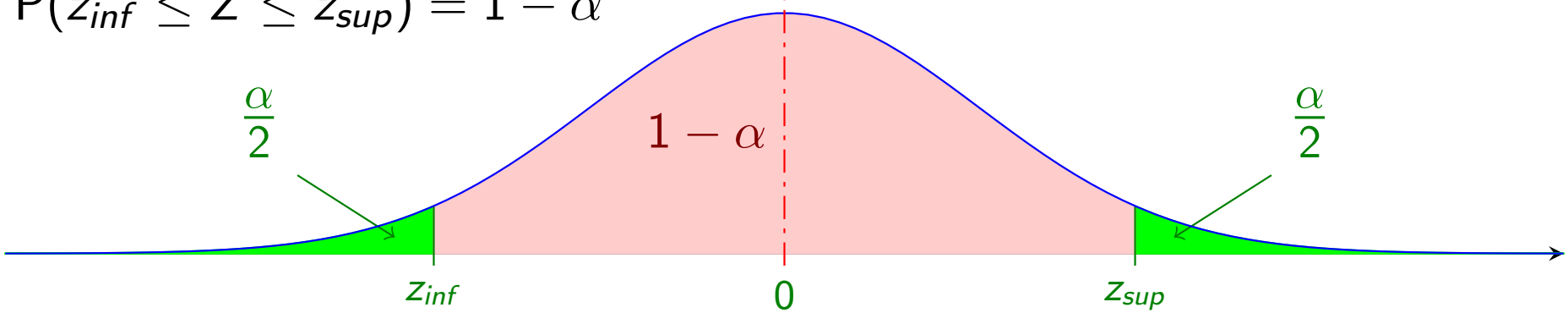
$$IF_{1-\alpha}(M) = [b_{inf}; b_{sup}] \quad \text{tel que} \quad P(b_{inf} \leq M \leq b_{sup}) = 1-\alpha$$

- $\alpha =$ risque (que la valeur ne soit pas dans l'intervalle)
  Classiquement $\alpha = 5 \cdot 10^{-2}$

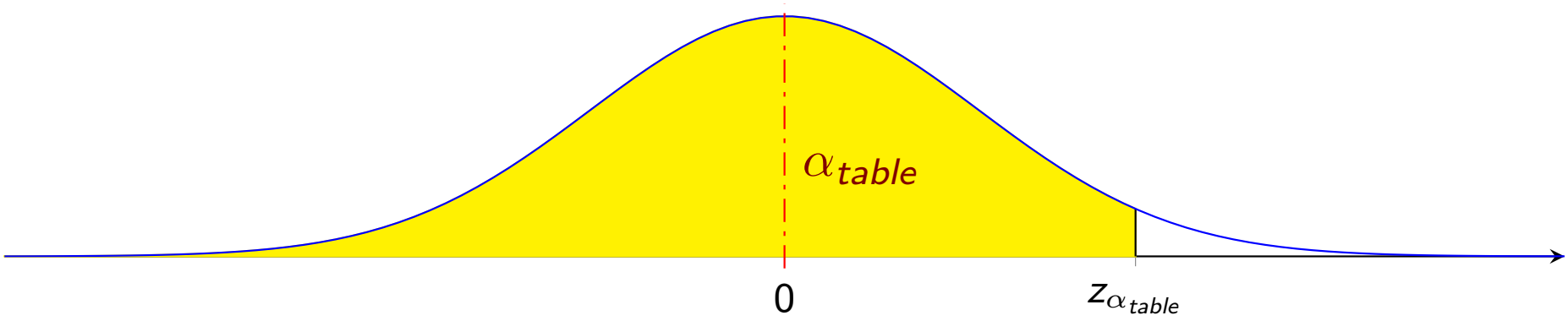


Rappel : lecture dans la table 2

On cherche les valeurs des bornes z_{inf} et z_{sup} d'un intervalle tel que $P(z_{inf} \leq Z \leq z_{sup}) = 1 - \alpha$



Dans la **table 2**, on lit directement la valeur $z_{\alpha_{table}}$ telle que $P(Z \leq z_{\alpha_{table}}) = \alpha_{table}$



Sur le graphe, on voit que $\alpha_{table} = \frac{\alpha}{2} + 1 - \alpha = 1 - \frac{\alpha}{2}$

Finalement, $\mathbf{z_{sup} = z_{1-\frac{\alpha}{2}}}$ et $\mathbf{z_{inf} = -z_{1-\frac{\alpha}{2}}}$

Attention, la notation utilisée pour les bornes dépend de la table

Intervalle de fluctuation (= intervalle de pari) de M (2)

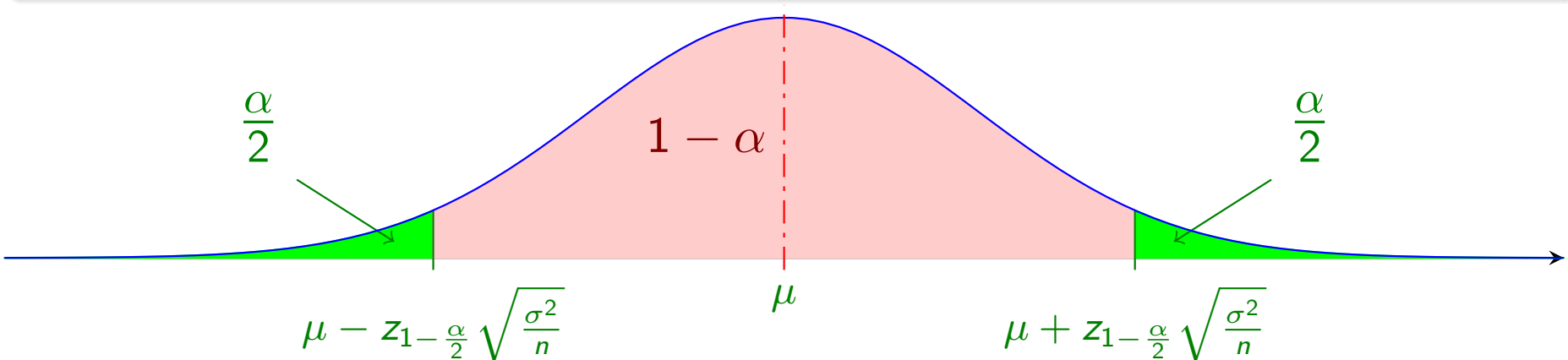
Bornes de l'IF

On cherche b_{inf} et b_{sup} telles que $\mathrm{P}(b_{inf} \leq M \leq b_{sup}) = 1 - \alpha$

- Centrage, réduction $\mathrm{P}\left(\frac{b_{inf} - \mu}{\sigma/\sqrt{n}} \leq \frac{M - \mu}{\sigma/\sqrt{n}} \leq \frac{b_{sup} - \mu}{\sigma/\sqrt{n}}\right) = 1 - \alpha$

  $\mathrm{P}(z_{inf} \leq Z \leq z_{sup}) = 1 - \alpha$
- Lecture de z_{sup} dans la table 2 ($\mathrm{P}(Z \leq z_{sup}) = 1 - \frac{\alpha}{2}$) : $z_{sup} = z_{1-\frac{\alpha}{2}}$
- Symétrie de la ddp de Z : $\frac{b_{inf}-\mu}{\sigma/\sqrt{n}} = -z_{1-\frac{\alpha}{2}}$ et $\frac{b_{sup}-\mu}{\sigma/\sqrt{n}} = z_{1-\frac{\alpha}{2}}$

$$IF_{1-\alpha}(M) = \left[\mu - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}; \mu + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right]$$



Exercice

QCM : Donnez la(les) propositions vraie(s)

On suppose que la glycémie des individus d'une population donnée est distribuée normalement, avec une moyenne de 1.00 g/L et un écart type de 0.03 g/L. On considère un échantillon de 100 personnes issues de cette population.

1. La probabilité que la glycémie d'une personne soit supérieure à 1.03 g/L est inférieure à $2 \cdot 10^{-5}$
2. La probabilité pour que la glycémie moyenne soit supérieure à 1.03 g/L est inférieure à $2 \cdot 10^{-5}$
3. L'intervalle de fluctuation au risque 5 % de la glycémie moyenne est : $[0.94; 1.06]$
4. L'intervalle de fluctuation au risque 5 % de la glycémie est : $[0.94; 1.06]$
5. Quand on donne un IF, on majore la borne supérieure et on minore la borne inférieure

Exercice

QCM : Donnez la(les) propositions vraie(s)

On suppose que la glycémie des individus d'une population donnée est distribuée normalement, avec une moyenne de 1.00 g/L et un écart type de 0.03 g/L. On considère un échantillon de 100 personnes issues de cette population.

1. La probabilité que la glycémie d'une personne soit supérieure à 1.03 g/L est inférieure à $2 \cdot 10^{-5}$
2. La probabilité pour que la glycémie moyenne soit supérieure à 1.03 g/L est inférieure à $2 \cdot 10^{-5}$
3. L'intervalle de fluctuation au risque 5 % de la glycémie moyenne est : $[0.94; 1.06]$
4. L'intervalle de fluctuation au risque 5 % de la glycémie est : $[0.94; 1.06]$
5. Quand on donne un IF, on majore la borne supérieure et on minore la borne inférieure

Correction

Réponses justes : 2 - 4 - 5

Généralités sur l'intervalle de fluctuation

IF d'une variable aléatoire suivant une loi normale

Soit Y une variable aléatoire suivant une loi normale d'espérance μ_Y et d'écart-type σ_Y

- Un intervalle de fluctuation de Y à la confiance $1-\alpha$ est :

$$IF_{1-\alpha}(Y) = \mu_Y \pm z_{1-\frac{\alpha}{2}} \times \sigma_Y$$

Application aux variables aléatoires M et X

- Soit $X \to \mathcal{N}(\mu, \sigma)$

$$IF_{1-\alpha}(X) = \mu \pm z_{1-\frac{\alpha}{2}} \times \sigma$$

- Soit $M = \frac{1}{n}\sum_{i=1}^{n} X_i$, $M \to \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

$$IF_{1-\alpha}(M) = \mu \pm z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

Estimateur de la variance : introduction

Variance d'un échantillon

Données de mortalité par hépatocarcinome (en mois de survie après le diagnostic)
Échantillon de 10 patients : 7 - 11 - 12 - 12 - 14 - 19 - 20 - 20 - 32 - 41

Variance de l'échantillon : $s_e^2 = \frac{\sum_{i=1}^{10}(x_i - m)^2}{10}$

Reformulation de la question

$S_e^2 = \frac{1}{n}\sum_i (X_i - M)^2$ est-il un bon estimateur de la variance σ^2 ?

On peut montrer que $\mathrm{E}(S_e^2) = \frac{(n-1)}{n}\sigma^2$

S_e^2 est un estimateur biaisé

Estimateur non biaisé de la variance : S^2

Définition d'un estimateur non biaisé de la variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n \times M^2\right)$$

Estimation ponctuelle de la variance

Estimation de la variance de la population à partir d'un échantillon

$$\boxed{s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - m)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n \times m^2\right)}$$

Attention à ne pas confondre

- la variance **descriptive** : $s_e^2 = \frac{SCE}{n} \rightarrow$ **statistiques descriptives**
- l'**estimation** de la variance de la population réalisée à partir d'un échantillon : $s^2 = \frac{SCE}{n-1} \rightarrow$ **statistiques inférentielles**

avec $SCE = \sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n} x_i^2 - n \times m^2$

Estimateur d'une proportion théorique π

Introduction

Soit X une va modélisant le statut maladie d'un patient
$X = 1$: le patient est malade $(\mathrm{P}(X = 1) = \pi)$

$$X \to Bern(\pi)$$

Soit S_n une va modélisant le nombre de patients malades parmi un ensemble de n patients

$$Sn = \sum_{i=1}^{n} X_i \quad S_n \to \mathcal{B}(n, \pi)$$

F est une va modélisant la **proportion** de malades

$$\boxed{\mathrm{F} = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}}$$

Remarque

F est un cas particulier de M lorsque X suit une **loi de Bernoulli**

Loi de F

Loi de F

Rappel : $\mathrm{F} = \frac{S_n}{n}$, avec $S_n \to \mathcal{B}(n, \pi)$

Si $n \geq 30$, $n\pi \geq 5$ et $n(1-\pi) \geq 5$, $S_n \rightsquigarrow \mathcal{N}(n\pi, \sqrt{n\pi(1-\pi)})$

Donc F suit **approximativement** une loi normale

Espérance et variance de F

$$\mathrm{E}(\mathrm{F}) = \mathrm{E}\left(\frac{S_n}{n}\right) = \frac{1}{n} \times \mathrm{E}(S_n) = \frac{1}{n} \times n\pi \times (1-\pi)$$

$$\mathrm{var}(\mathrm{F}) = \mathrm{var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2} \times \mathrm{var}(S_n) = \frac{1}{n^2} \times n\pi(1-\pi) = \frac{\pi \times (1-\pi)}{n}$$

Qualité de l'estimateur F

Rappel

$$\mathrm{E}(\mathrm{F}) = \pi \quad \text{et} \quad \mathrm{var}(\mathrm{F}) = \frac{\pi \times (1-\pi)}{n}$$

Absence de biais

$\mathrm{Biais} = \mathrm{E}(\mathrm{F}) - \pi = 0$

Erreur quadratique moyenne

$EQM = \mathrm{var}(\mathrm{F}) + biais^2$

$EQM = \frac{\pi \times (1-\pi)}{n} + 0 \qquad \lim_{n \to +\infty} EQM = 0$

Conclusion

F est un « bon » estimateur d'une proportion

Intervalle de fluctuation de F

Définition

- On se place dans le cas ou $F \rightsquigarrow \mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

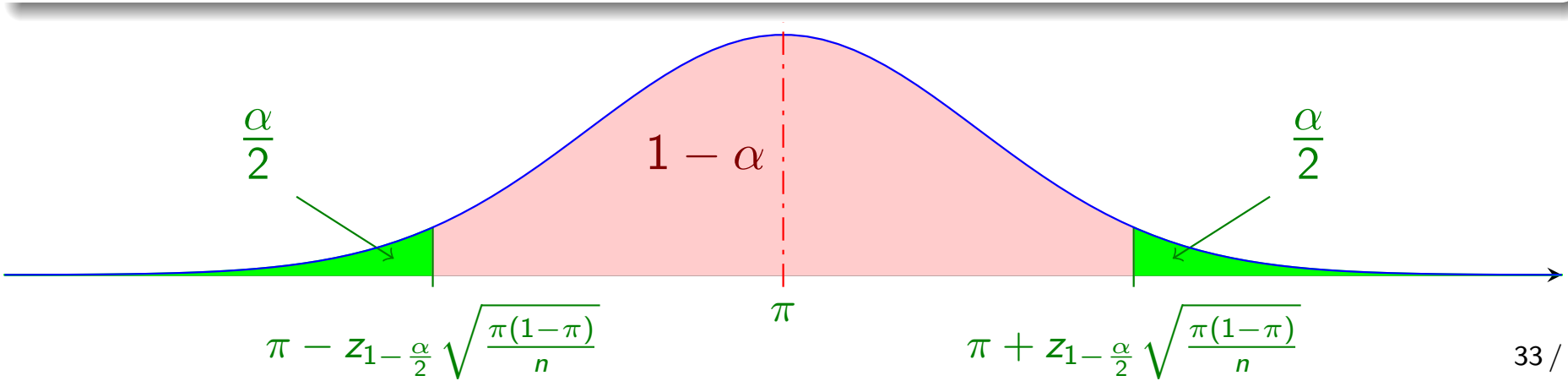
  On approxime la loi de F par une loi parfaitement connue

- On cherche un intervalle centré sur π qui a une probabilité $1-\alpha$ de contenir une réalisation de F

$$IF_{1-\alpha}(F) = [b_{inf}; b_{sup}] \quad \text{tel que} \quad P(b_{inf} \leq F \leq b_{sup}) = 1-\alpha$$

On détermine b_{inf} et b_{sup} comme dans le cas de l'$IF_{1-\alpha}(M)$

$$\boxed{IF_{1-\alpha}(F) = \left[\pi - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}; \pi + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}\right]}$$



Exercice

Énoncé

Lors d'une élection, un candidat a obtenu 40% des voix. On prélève aléatoirement un échantillon de 100 bulletins de vote.

Modélisation du problème

- Soit X la va modélisant le vote d'un électeur :

$X = 1$ si l'électeur vote pour le candidat donné.
$X \to Bern(\pi)$ avec $\pi = 0{,}40$

- Soit S_{100}, la va modélisant le nb d'électeurs ayant voté pour le candidat dans l'échantillon de 100 électeurs (n=100)

$S_{100} = \sum_{i=1}^{100} X_i$ $\qquad$ $S_{100} \to \mathcal{B}(100; 0{,}40)$
Or, $n \geq 30$, $n\pi \geq 5$ et $n(1-\pi) \geq 5$
$\to$ approximation de $\mathcal{B}(100; 0{,}40)$ par $\mathcal{N}(100 \times 0{,}4; \sqrt{100 \times 0{,}4 \times 0{,}6})$

- F : va modéliser la proportion d'électeurs ayant voté pour le candidat.

$F = \sum_{i=1}^{100} \frac{X_i}{100}$ $\quad$ $F \rightsquigarrow \mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$ $\quad$ $\pi = 0{,}4$ et $\sqrt{\frac{\pi(1-\pi)}{n}} \simeq 0{,}05$

Exercice

Énoncé

Lors d'une élection, un candidat a obtenu 40% des voix. On prélève aléatoirement un échantillon de 100 bulletins de vote.
1/ Quelle est la probabilité que, dans l'échantillon, le candidat ait obtenu entre 35% et 45% des suffrages ?

Correction de la question 1

On cherche $\mathrm{P}(0{,}35 < F < 0{,}45)$
Centrage et réduction : $\mathrm{P}\left(\frac{0{,}35-0{,}40}{0{,}05} < Z < \frac{0{,}45-0{,}40}{0{,}05}\right)$
$\mathrm{P}(-1 < Z < 1) = \phi(1) - \phi(-1) = \phi(1) - (1 - \phi(1)) = 2\phi(1) - 1$
On lit $\phi(1)$ dans la table de la fdr : $\phi(1) \simeq 0{,}84$
D'où $\mathrm{P}(0{,}35 < F < 0{,}45) \simeq 0{,}68$

Exercice

Énoncé

Lors d'une élection, un candidat a obtenu 40% des voix. On prélève aléatoirement un échantillon de 100 bulletins de vote.
2/ Donner un intervalle de fluctuation de niveau 0.95 de l'estimateur de la proportion des suffrages obtenus par le candidat

Correction de la question 2

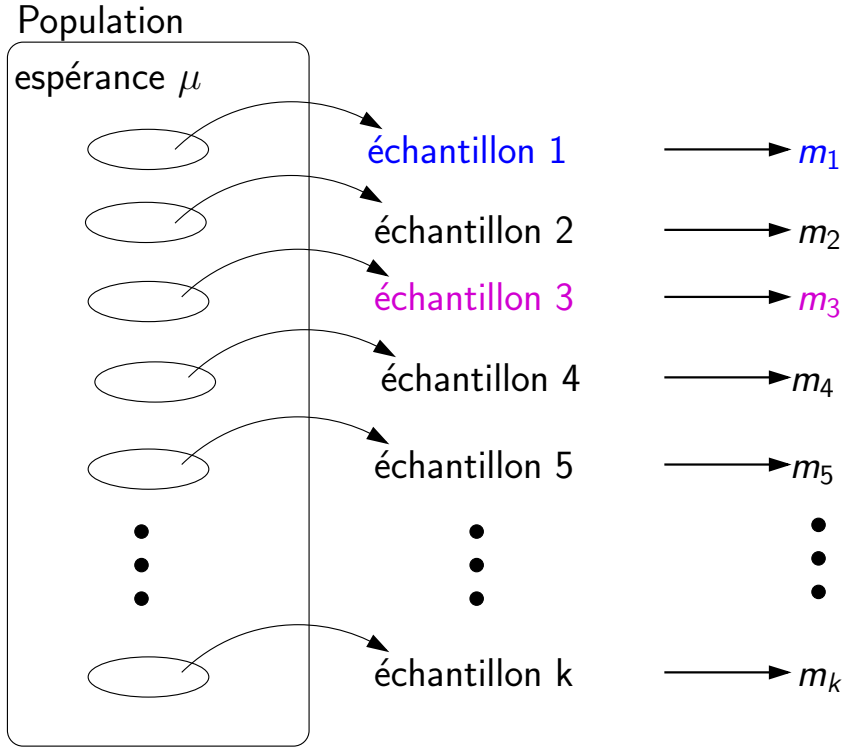
$$IF_{1-\alpha}(\mathrm{F}) = \left[\pi - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}; \pi + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

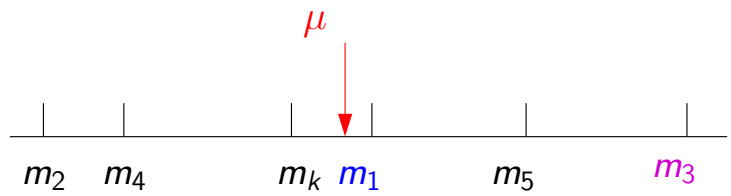
$\pi = 0{,}4$
$\sqrt{\frac{\pi(1-\pi)}{n}} \simeq 0{,}05$
$z_{0{,}975} = 1{,}96 \simeq 2$ (lecture dans la table 2)

$$\boxed{IF_{0{,}95}(\mathrm{F}) = [0{,}30; 0{,}50]}$$

Plan du cours

Introduction





Estimation ponctuelle

Exemple : X= « poids des nouveaux nés », $X \to \mathcal{L}(\mu, \sigma)$

Estimation ponctuelle de μ :

- définie par un **estimateur** (ici, la va M)
- calculée sur un échantillon $\to$ dépend de l'échantillon

Problème : précision de l'estimation inconnue

Estimation par intevalle de confiance (IC)

Principe

- Définir un **intervalle** dans lequel on a beaucoup de chances de trouver la valeur du paramètre à estimer
- Donner une **précision** sur la valeur de l'estimation ponctuelle

Définition d'un $IC_{1-\alpha}$

Un intervalle de confiance de niveau (de confiance) $1-\alpha$ est un intervalle calculé **pour un échantillon donné**, qui a (1-α)% de chances de contenir la vraie valeur de θ

$$IC_{1-\alpha}(\theta) = [B_{inf}; B_{sup}] \text{ tel que } P(B_{inf} \leq \theta \leq B_{sup}) = 1-\alpha$$

L'IC est **aléatoire** : ses bornes dépendent de l'échantillon considéré

- Calcul d'une infinité d'ic sur une infinité d'échantillons
- Alors on a une proportion $(1-\alpha)$ d'ic qui contiennent bien θ

Vocabulaire

Le risque que l'$IC_{1-\alpha}(\theta)$ ne contienne pas θ vaut α :

$$\mathrm{P}(\theta < B_{inf}) + \mathrm{P}(\theta > B_{sup}) = \alpha$$

IC bilatéral vs IC unilatéral

- Si $\mathrm{P}(\theta < B_{inf}) = 0$ ou $\mathrm{P}(\theta > B_{sup}) = 0$, l'IC est **unilatéral**
- Si $\mathrm{P}(\theta < B_{inf}) \neq 0$ et $\mathrm{P}(\theta > B_{sup}) \neq 0$, l'IC est **bilatéral**

IC symétrique vs IC asymétrique

- Si $\mathrm{P}(\theta < B_{inf}) = \mathrm{P}(\theta > B_{sup}) = \frac{\alpha}{2}$, l'IC est **symétrique**
- Sinon, l'IC est **asymétrique**

Pour le PASS

On ne calculera que des IC bilatéraux et symétriques

Conditions de validité

Pour tous les calculs d'intervalle de confiance, on se limitera au cas des grands échantillons ($n \to +\infty$)

$\to$ en pratique, $n \geq 30$

Le cas **petit échantillon** (n<30) n'est pas au programme du PASS cette année.

Intervalle de confiance d'une moyenne

Établissement de l'$IC_{1-\alpha}(\mu)$

On part de l'**intervalle de fluctuation**, mais on considère que μ est l'inconnue :

$$\mathrm{P}\left(\mu - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq M \leq \mu + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Or, $\left(\mu - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq M\right) \Leftrightarrow \left(\mu \leq M + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right)$

Et, $\left(M \leq \mu + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) \Leftrightarrow \left(\mu \geq M - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right)$

$$\Leftrightarrow \mathrm{P}\left(M - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq M + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Réalisation de l'IC pour un échantillon

Conclusion

Dans un échantillon

- La réalisation de M est m
- σ est estimé par s (sauf s'il est connu)

Une réalisation de l'$IC_{0.95}(\mu)$ est :

$$ic_{1-\alpha}(\mu) = \left[m - z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}; m + z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}\right]$$

Dans le cas « classique » où $\alpha = 0{,}05$, $z_{1-\frac{\alpha}{2}} = 1{,}96$. Un ic s'écrit alors :

$$ic_{0,95}(\mu) = \left[m - 1{,}96 \times \frac{s}{\sqrt{n}}; m + 1.96 \times \frac{s}{\sqrt{n}}\right]$$

Réalisation de l'IC pour un échantillon (2)

Règle d'arrondi

Pour garantir le niveau de confiance $1 - \alpha$:

- On **minore** la borne inférieure
- On **majore** la borne supérieure

Exemples

Donner les intervalles de confiance suivant avec un chiffre après la virgule :

- $ic_{0.95}(\mu) = [34{,}31; 39{,}09] \quad \rightarrow \quad ic_{0.95}(\mu) = [34{,}3; 39{,}1]$
- $ic_{0.95}(\mu) = [34{,}399; 39{,}001] \quad \rightarrow \quad ic_{0.95}(\mu) = [34{,}3; 39{,}1]$

Exercice

Énoncé

On veut quantifier la pollution de l'atmosphère par un gaz toxique. Sur un ensemble de 100 prélèvements, on observe $m = 50$ et $s^2 = 100$. Donner un intervalle de confiance de niveau (de confiance) 0,95 et un intervalle de confiance de niveau 0,99 de la quantité moyenne de gaz toxique (2 chiffres après la virgule)

Correction

Soit X : « quantité de gaz toxique », $X \to \mathcal{L}(\mu, \sigma)$
Estimateur de μ : $M = \frac{1}{n}\sum_i X_i$, $M \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

$$IC_{1-\alpha}(\mu) = M \pm z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

Estimation à partir de l'échantillon : $ic_{1-\alpha}(\mu) = m \pm z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

AN : $ic_{0,95}(\mu) = 50 \pm 1{,}96 \times \frac{10}{\sqrt{100}}$ $\boxed{ic_{0,95}(\mu) = [48{,}04; 51{,}96]}$

Exercice

Énoncé

On veut quantifier la pollution de l'atmosphère par un gaz toxique. Sur un ensemble de 100 prélèvements, on observe $m = 50$ et $s^2 = 100$. Donner un intervalle de confiance de niveau (de confiance) 0,95 et un intervalle de confiance de niveau 0,99 de la quantité moyenne de gaz toxique (2 chiffres après la virgule)

Correction (suite)

Intervalle de confiance de niveau 0,99 ($\alpha = 0{,}01$)
On lit dans la table la valeur de $z_{1-\frac{\alpha}{2}}$ pour $1 - \frac{\alpha}{2} = 0{,}995$:
$z_{1-\frac{\alpha}{2}} = 2{,}5758$
$ic_{0,99}(\mu) = m \pm 2{,}5758 \times \frac{s}{\sqrt{n}}$
AN : $ic_{0,99}(\mu) = 50 \pm 2{,}5758 \times \frac{10}{\sqrt{100}}$ $\quad ic_{0,99}(\mu) = [47{,}4242; 52{,}5758]$

$$\boxed{ic_{0,99}(\mu) = [47{,}42; 52{,}58]}$$

Rappel : $ic_{0,95}(\mu) = [48{,}04; 51{,}96]$

Conclusion : plus le niveau de confiance est élevé, plus l'ic est large

Exercice (suite : influence de la taille de l'échantillon)

Énoncé

On veut quantifier la pollution de l'atmosphère par un gaz toxique. Donner un intervalle de confiance à 95% de la quantité moyenne de gaz toxique dans les cas suivants (1 chiffre après la virgule) :

1. sur un ensemble de 81 prélèvements, on observe $m = 50$ et $s^2 = 100$
2. sur un ensemble de 900 prélèvements, on observe $m = 50$ et $s^2 = 100$

Exercice (suite : influence de la taille de l'échantillon)

Énoncé

On veut quantifier la pollution de l'atmosphère par un gaz toxique. Donner un intervalle de confiance à 95% de la quantité moyenne de gaz toxique dans les cas suivants (1 chiffre après la virgule) :

1. sur un ensemble de 81 prélèvements, on observe $m = 50$ et $s^2 = 100$
2. sur un ensemble de 900 prélèvements, on observe $m = 50$ et $s^2 = 100$

Correction

1. Échantillon de taille $n = 81$: $ic_{0,95}(\mu) = [47{,}8; 52{,}2]$
2. Échantillon de taille $n = 900$: $ic_{0,95}(\mu) = [49{,}3; 50{,}7]$

Conclusion : plus l'échantillon est grand, plus l'ic est étroit

Intervalle de confiance d'une proportion (1) : introduction

Le problème

On cherche à estimer la **prévalence** (π) d'une maladie. Pour cela, on étudie un échantillon de 10000 personnes. On observe 100 cas de maladie.
Estimation ponctuelle de la prévalence : f $= \frac{100}{10000} = 0.01$
Estimation par IC ?

Modélisation

X : « statut maladie d'un individu », $X \to Bern(\pi)$
Estimateur de la prévalence : F $= \frac{1}{n}\sum_{i=1}^{10000} X_i = \frac{1}{n} S_n$ avec $S_n \to \mathcal{B}(n, \pi)$
Si $n \geq 30$, $n\pi \geq 5$ et $n(1-\pi) \geq 5$, on peut approximer la loi de S_n par $\mathcal{N}(n\pi, \sqrt{n\pi(1-\pi)})$ et donc celle de F par $\mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$

Intervalle de fluctuation de F : rappels

IF : $[b_{inf}; b_{sup}]$ tel que $P(b_{inf} \leq F \leq b_{sup}) = 1 - \alpha$

$$IF_{1-\alpha}(F) = \left[\pi - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}; \pi + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}\right]$$

Intervalle de confiance pour une proportion (2)

Établissement de l'$IC_{1-\alpha}(p)$

On part de l'$IF_{1-\alpha}(\mathrm{F})$ mais on considère que π est l'inconnue

$$\mathrm{P}\left(\pi - z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\pi(1-\pi)}{n}} \leq \mathrm{F} \leq \pi + z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1-\alpha$$

Même type de raisonnement que pour l'$IF_{1-\alpha}(M)$
On obtient :

$$\boxed{\mathrm{P}\left(\mathrm{F} - z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\pi(1-\pi)}{n}} \leq \pi \leq \mathrm{F} + z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1-\alpha}$$

Intervalle de confiance pour une proportion (3)

Réalisation de l'IC pour un échantillon

$ic_{1-\alpha}(\pi) = \mathrm{f} \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\pi(1-\pi)}{n}}$

Or, π est inconnu...

Quand n grand, on peut **approximer** π par f

$$ic_{1-\alpha}(\pi) = \mathrm{f} \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\mathrm{f}(1-\mathrm{f})}{n}}$$

Vérification des conditions de validité

On a supposé les conditions d'approximation d'une loi $\mathcal{B}(n, \pi)$ par $\mathcal{N}(n\pi, \sqrt{n\pi(1-\pi)})$ vérifiées ($n \geq 30, n\pi \geq 5$ et $n(1-\pi) \geq 5$)

On vérifie que ces conditions sont réalisées aux bornes f_1 et f_2 de l'ic calculé

En pratique, on vérifie que :

$$n \geq 30, \quad n\mathrm{f}_1 \geq 5, \quad n(1-\mathrm{f}_1) \geq 5, \quad n\mathrm{f}_2 \geq 5, \quad n(1-\mathrm{f}_2) \geq 5$$

Exemple de la prévalence

Estimation par ic de la prévalence d'une maladie

- $n = 10000$
- $\mathrm{f} = 0{,}01$ (1%)
- $ic_{0,95}(\pi) = \mathrm{f} \pm 1{,}96 \times \sqrt{\frac{\mathrm{f}(1-\mathrm{f})}{n}}$

$ic_{0,95}(\pi) = 0{,}01 \pm 1{,}96 \times \sqrt{\frac{0{,}01 \times 0{,}99}{10000}}$

$ic_{0,95}(\pi) = 0{,}01 \pm 0{,}002$

$ic_{0,95}(\pi) = [0{,}008; 0{,}012]$

Conditions de validité vérifiées :

$n \geq 30$, $n\mathrm{f}_1 = 80$, $n\mathrm{f}_2 = 120$, $n(1-\mathrm{f}_1) = 9920$, $n(1-\mathrm{f}_2) = 9880$

Exemple des sondages

Le problème

Un institut de sondage étudie les intentions de vote pour le 2è tour des élections municipales. Sur un échantillon de 600 individus, 51% ont répondu avoir l'intention de voter pour le candidat A.

Intervalle de confiance de niveau 0.95 ?

- $n = 600$
- Estimation ponctuelle $\mathrm{f} = 0{,}51$
- $z_{1-\frac{\alpha}{2}} = 1{,}96$
- $ic_{0,95}(\pi) = \mathrm{f} \pm 1{,}96 \times \sqrt{\frac{\mathrm{f}(1-\mathrm{f})}{n}}$

$ic_{0,95}(\pi) = 0{,}51 \pm 1{,}96 \times \sqrt{\frac{0{,}51 \times 0{,}49}{600}}$

$ic_{0,95}(\pi) = 0{,}51 \pm 0{,}04 \quad ic_{0,95}(\pi) = [0{,}47; 0{,}55]$

Conditions de validité vérifiées :

$n \geq 30,\ n\mathrm{f}_1 = 282,\ n\mathrm{f}_2 = 330,\ n(1-\mathrm{f}_1) = 318,\ n(1-\mathrm{f}_2) = 270$

Précision d'un ic

Définition sur l'exemple des sondages

Rappel : $ic_{0.95}(\pi) = 0{,}51 \pm 0{,}04 \quad ic_{0,95}(\pi) = [0{,}47; 0{,}55]$

- Largeur l d'un ic : borne sup - borne inf. Ex : $l = 0{,}08$

  Dans le cas d'une proportion : $l = 2 \times z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{f(1-f)}{n}}$
- Précision i : précision $= 1/2$ largeur d'un ic.
  Ex : $i = 0{,}04$

i et l dépendent de n

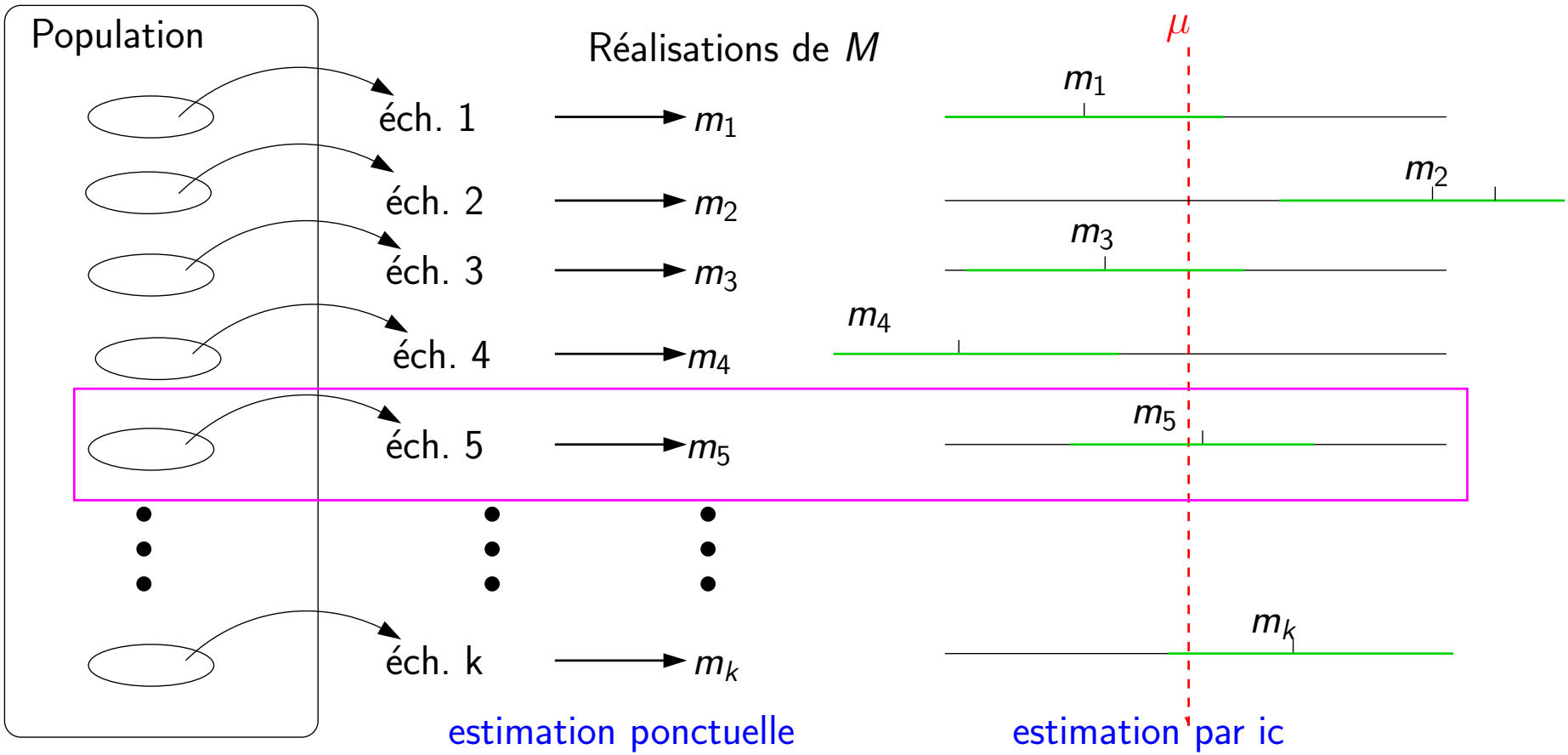
Nombre de sujets nécessaires pour avoir une 1/2 largeur inférieure ou égale à i_1 =1%

$$z_{1-\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} \leq i_1 \quad \Leftrightarrow \quad \frac{f(1-f)}{n} \leq \frac{i_1^2}{z_{1-\frac{\alpha}{2}}^2}$$

$$n \geq \frac{f(1-f) \times z_{1-\frac{\alpha}{2}}^2}{i_1^2}$$

AN (ex sur le sondage) : $n \geq 9600{,}1 \rightarrow n > 9601$

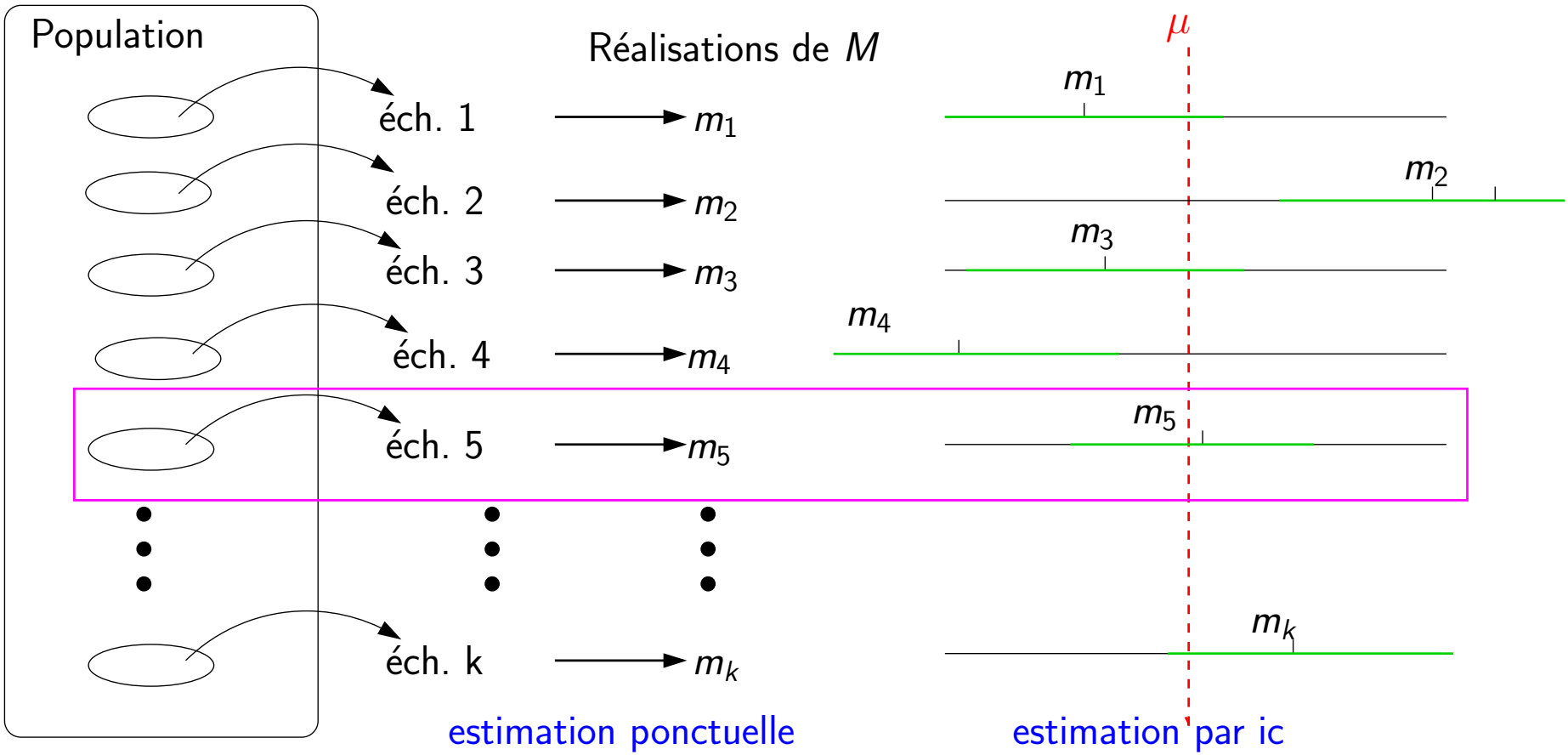
Bilan sur les ic



Un ic est centré sur la moyenne observée dans un échantillon

Il a $1-\alpha\%$ de chances de contenir la moyenne théorique μ

Bilan sur les ic

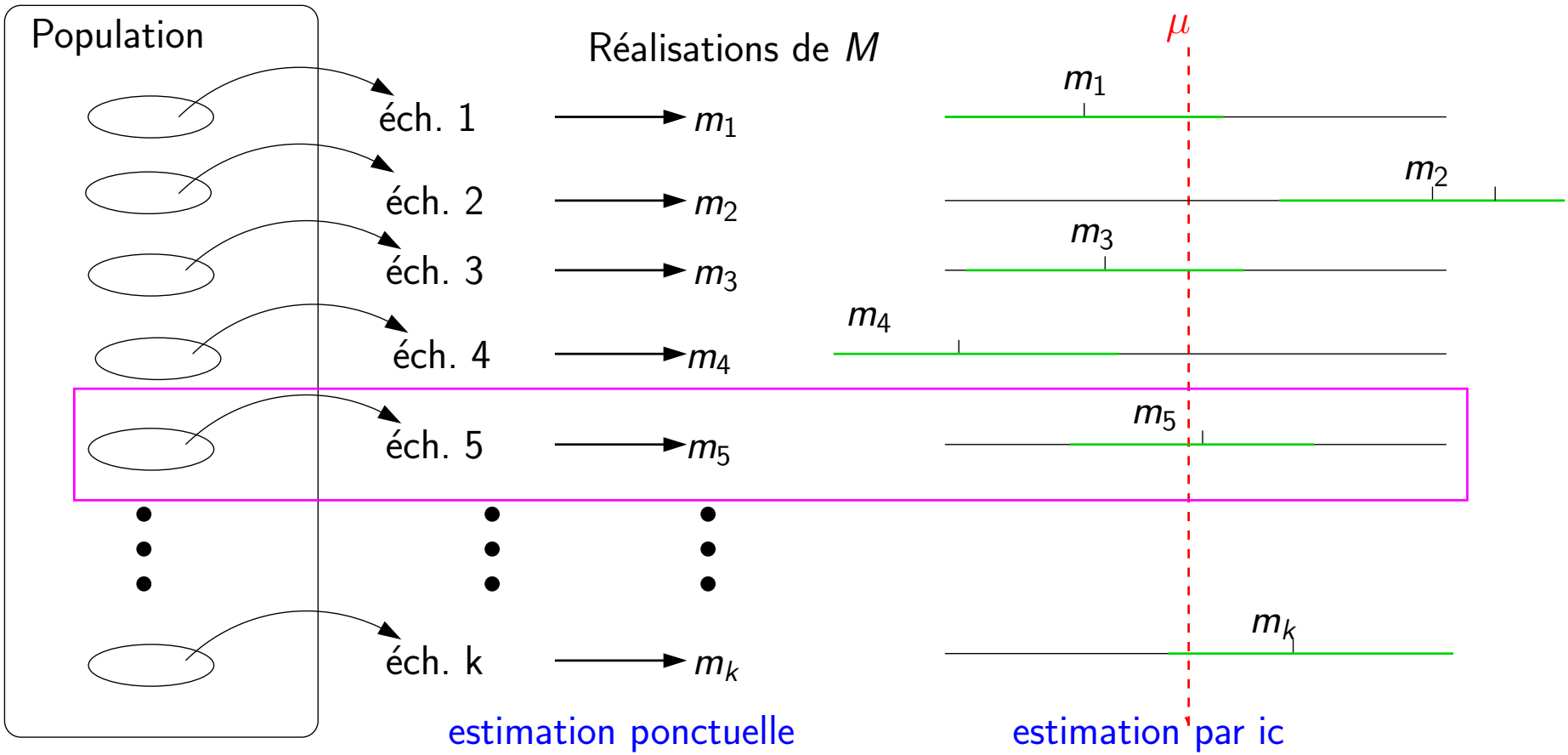


$$ic_{1-\alpha}(\mu) = m \pm z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

$$ic_{1-\alpha}(\pi) = \mathrm{f} \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\mathrm{f}(1-\mathrm{f})}{n}}$$

Toujours bien vérifier les conditions d'applications!

Bilan sur les ic



Plus la **taille d'échantillon** est grande, plus l'ic est étroit

Plus le **niveau de confiance** est grand, plus l'ic est large

Différences entre l'intervalle de confiance et l'intervalle de fluctuation (ou de pari)

Intervalle de confiance

- Calculé à partir d'un échantillon
- Centré sur m ou f
- Aléatoire (dépend de l'échantillon)
- A $1-\alpha\%$ de chances de contenir les valeurs théoriques μ ou π

Intervalle de fluctuation (ou de pari)

- Calculé à partir de la loi de probabilité de M ou F
- Centré sur μ ou π
- Fixe pour une va donnée
- Contient $(1-\alpha)\%$ des réalisations (m et f) de M et F

Conclusion

Éléments à retenir

- **Principe de l'estimation**
  - valeur théorique de la population : μ, π, σ
  - estimateur = va : M, F, S
  - estimation = réalisation d'une va : m, f, s
- **Intervalle de confiance**
  - pour une moyenne
  - pour une proportion
  - Conditions de validités des ic
  - différence avec l'intervalle de fluctuation
- **Calcul du nombre de sujets nécessaires pour une précision donnée** (cas d'une moyenne ou d'une proportion)