MIASS 241 EC1

Mathématiques (appliquées aux sciences sociales) 4

© El Hadj Touré, PhD Sociologie
Spécialiste des statistiques sociales et méthodes de sondage
Section de sociologie
Université Gaston Berger de St-Louis

7:41

Présentation générale du cours

Objectifs d'apprentissage

1. Comprendre la logique de l'inférence statistique
   - Marge d'erreur, estimation par intervalle de confiance
2. Analyser une relation entre deux variables qualitatives
   - Tableaux croisés et test du chi-carré
3. Comparer deux moyennes de groupes
   - Tableau des moyennes et test t de Student
4. Utiliser le tableur Excel pour procéder à des calculs et analyses statistiques

Présentation générale du cours

Moyens mobilisés

- Démarche pédagogique
  - 4 présentations théoriques à l'aide de PowerPoint (Auditorium)
  - 4 applications pratiques avec le tableur Excel (salle multimédia 6)
- Espace virtuel du cours (Moodle)
  - Moodle, une plateforme interactive de gestion de cours en ligne
  - Accès au site web du cours (Moodle)
    - Dans l'url de **votre navigateur, saisir: foad.ugb.sn**
    - Utilisez votre courriel ugb comme nom d'utilisateur
    - Mettez votre mot de passe
  - Planification des activités
  - Contenu (plan, notes de cours, exercices corrigés, TP, etc.)
  - Test d'autoévaluation formative et sommative hebdomadaire (quiz)

Présentation générale du cours

Modalités d'évaluation

Évaluations	Date de disponibilité	Date de remise	Pondéra-tion
4 Quiz	Après chaque leçon	Une semaine après	25%
TP en équipe	3 juin	17 juin	25%
Examen	29 juin		50%
Total	--	--	100%

- Les quiz sont individuels et sont disponibles sur Moodle
- Le TP se fait en équipe de 5 à 6 étudiants, disponible et remis sur Moodle
- L'examen se fait sur table

7:41

Leçon 1

Introduction à l'inférence statistique

Au programme

- Comment inférer à toute la population des résultats obtenus à partir d'un échantillon aléatoire?
  - Différencier un paramètre d'une statistique
  - Conditions d'application de l'inférence statistique
  - Savoir ce qu'est une distribution d'échantillonnage
  - Comprendre la logique de l'estimation d'un paramètre: estimation ponctuelle et par intervalle de confiance
  - Estimer une moyenne ou un pourcentage d'une population à partir de la moyenne ou du pourcentage d'un échantillon
  - Remarques concernant l'estimation de paramètres
- Remue-méninges et exercices « éclair »

Inférence statistique

Paramètre & statistique: définition

- Elle consiste à connaître les caractéristiques d'une population (paramètres) à partir de celles d'un échantillon (statistiques) en déterminant l'erreur d'échantillonnage
- Une statistique
  - Valeur décrivant une caractéristique d'un échantillon ***n*** : $\overline{\boldsymbol{X}}$ ou ***p***
  - Une statistique peut être vue comme l'estimé d'un paramètre
  - Elle est un nombre aléatoire, c'est-à-dire soumis au hasard
- Un paramètre
  - Valeur décrivant une caractéristique d'une population ***N*** : μ ou π
  - La plupart du temps, cette valeur est inconnue et il faut l'estimer
  - Un paramètre est un nombre déterministe, non soumis au hasard

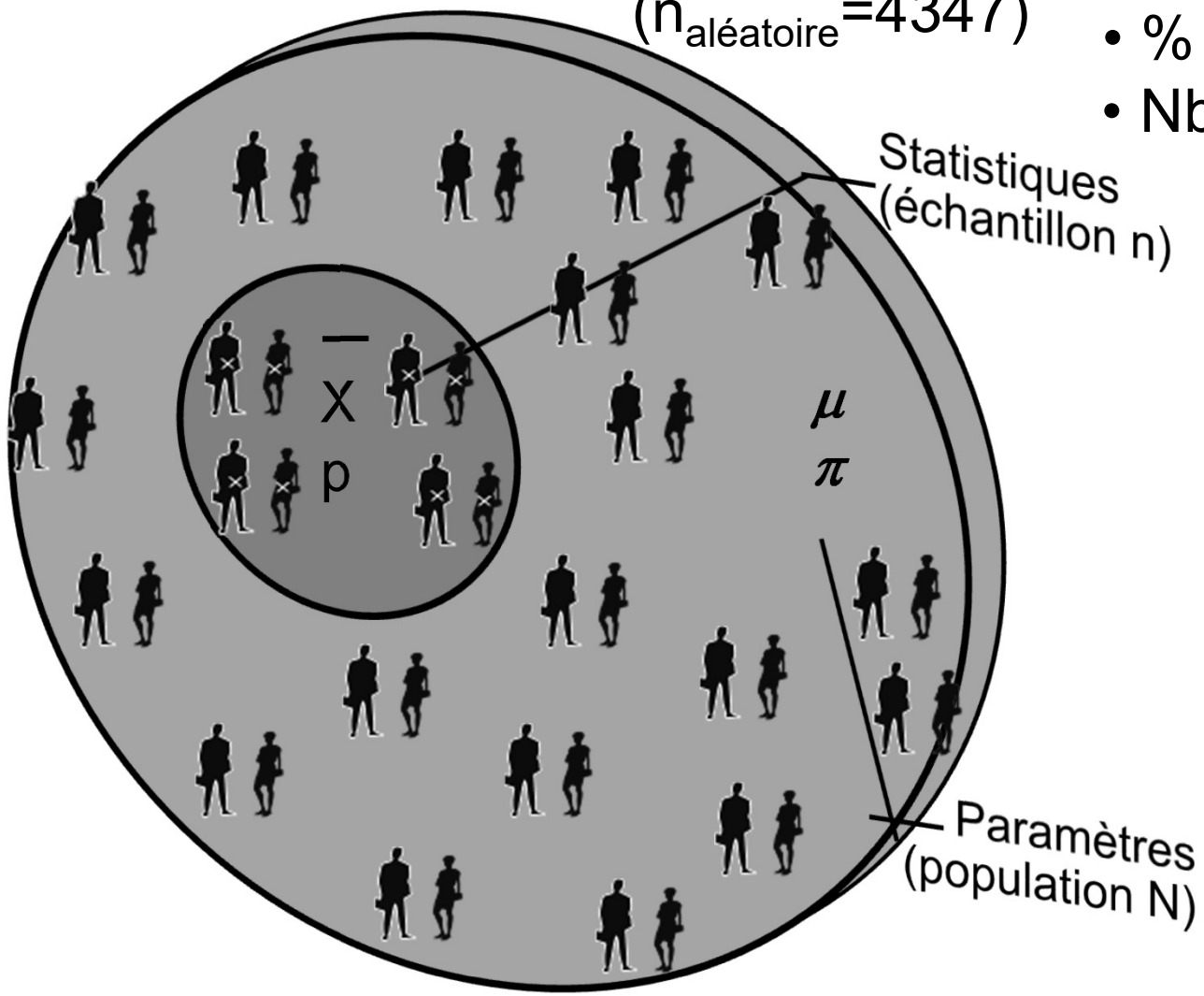
Inférence statistique

Statistique & paramètre: illustration

→Tabagisme chez les sénégalais âgés de 15 ans et plus (GATS 2015)

($n_{aléatoire}$=4347)

- % fumeurs p= 5,4% (235/4347)
- Nbre cig. fumées/j $\bar{X}$ = 9,4



Dans quelle mesure la proportion ou la moyenne de l'échantillon reflète-t-elle la valeur réelle, c'est-à-dire la proportion ou la moyenne qui serait obtenue si l'étude portait sur l'ensemble de la population sénégalaise?

Inférence statistique

Deux conditions d'application

1. L'échantillon doit être constitué de 30 cas au minimum
   - Les caractéristiques de ***n*** se rapprochent d'autant plus de celles de ***N*** que ***n*** est grand (loi des grands nombres)
2. L'échantillon doit être aléatoire et non empirique
   - Échantillon aléatoire ou probabiliste
     - Les individus de ***n*** sont choisis au hasard, par tirage au sort
     - La population parente ***N*** est connue (liste des individus)
     - Ex: aléatoire simple, systématique, stratifié, par grappes
   - Échantillon empirique ou non probabiliste
     - Les individus de ***n*** sont choisis délibérément (selon des critères)
     - La population parente ***N*** est inconnue (absence d'une liste)
     - Ex: par quotas, accidentel, volontaire, typique, boule de neige

Inférence statistique

Pourquoi opter pour l'échantillon aléatoire?

- Sémantiquement, un échantillon ***n*** est vraiment aléatoire si chacun des individus de la pop. ***N*** a une chance égale non nulle (équiprobabilité) et indépendante d'être sélectionné
- À la longue, l'échantillon aléatoire produit un échantillon "représentatif" de la population, permettant ainsi de pouvoir se servir en toute confiance de la statistique de ***n*** pour connaître, estimer le paramètre de ***N*** (estimation fiable)
- Avec un échantillon aléatoire, on peut calculer la probabilité générale qu'un ind. de ***N*** fasse partie de ***n*** (p=n/N), permettant ainsi de calculer l'erreur d'échantillonnage, d'établir une probabilité connue de la précision de l'estimation

Estimation d'un paramètre

Estimation ponctuelle & par intervalle de confiance

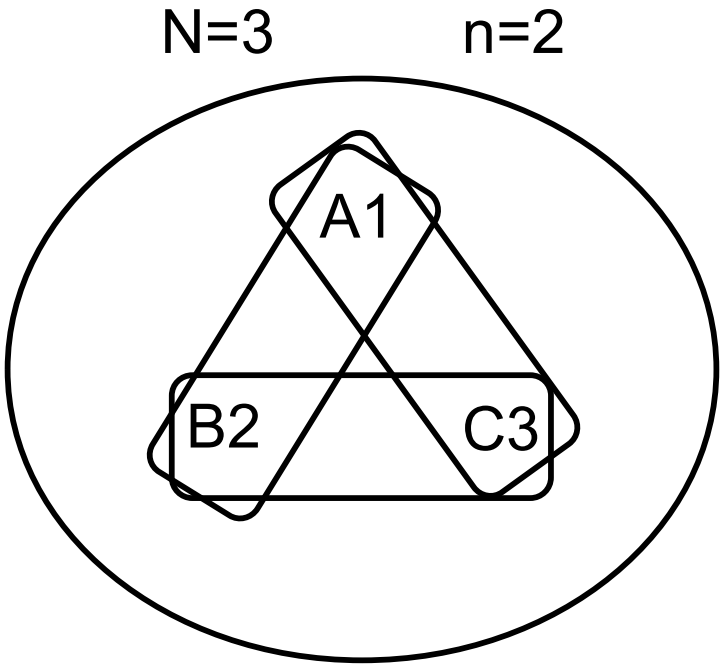
→L'estimation consiste, à partir d'une statistique de ***n*** ($\overline{X}$ ou p), à estimer le paramètre de ***N*** (μ ou π) (sondages surtout)

- Estimation ponctuelle
  - Elle consiste à estimer un paramètre d'une population par une valeur unique: une statistique de l'échantillon
  - Ex: Un sondage aléatoire mené en 2002 auprès de 1500 enfants (5^e-9^e) canadiens montre que 13% (195) fument du tabac. La proportion 13% s'applique à toute la population canadienne étudiée
- Estimation par intervalle de confiance
  - Elle consiste à estimer les deux valeurs qui encadrent un paramètre recherché: on parle d'intervalle de confiance
  - Ex: … 13% avec une marge d'erreur de ±2% (entre 11% et 15%), le niveau de confiance étant fixé à 95% (95 fois sur 100)

Estimation d'un paramètre

Trois sortes de distribution à distinguer

- Distribution d'une population
  - Distribution des scores dans ***N***. Elle donne un paramètre
- Distribution d'un échantillon
  - Distribution des scores dans ***n***. Elle donne une statistique
- Distribution d'échantillonnage
  - Distribution d'une statistique quelconque [moyenne ou %] de tous les échantillons possibles [***n***] d'une taille donnée [dans ***N***]. Elle est au fondement de l'inférence statistique

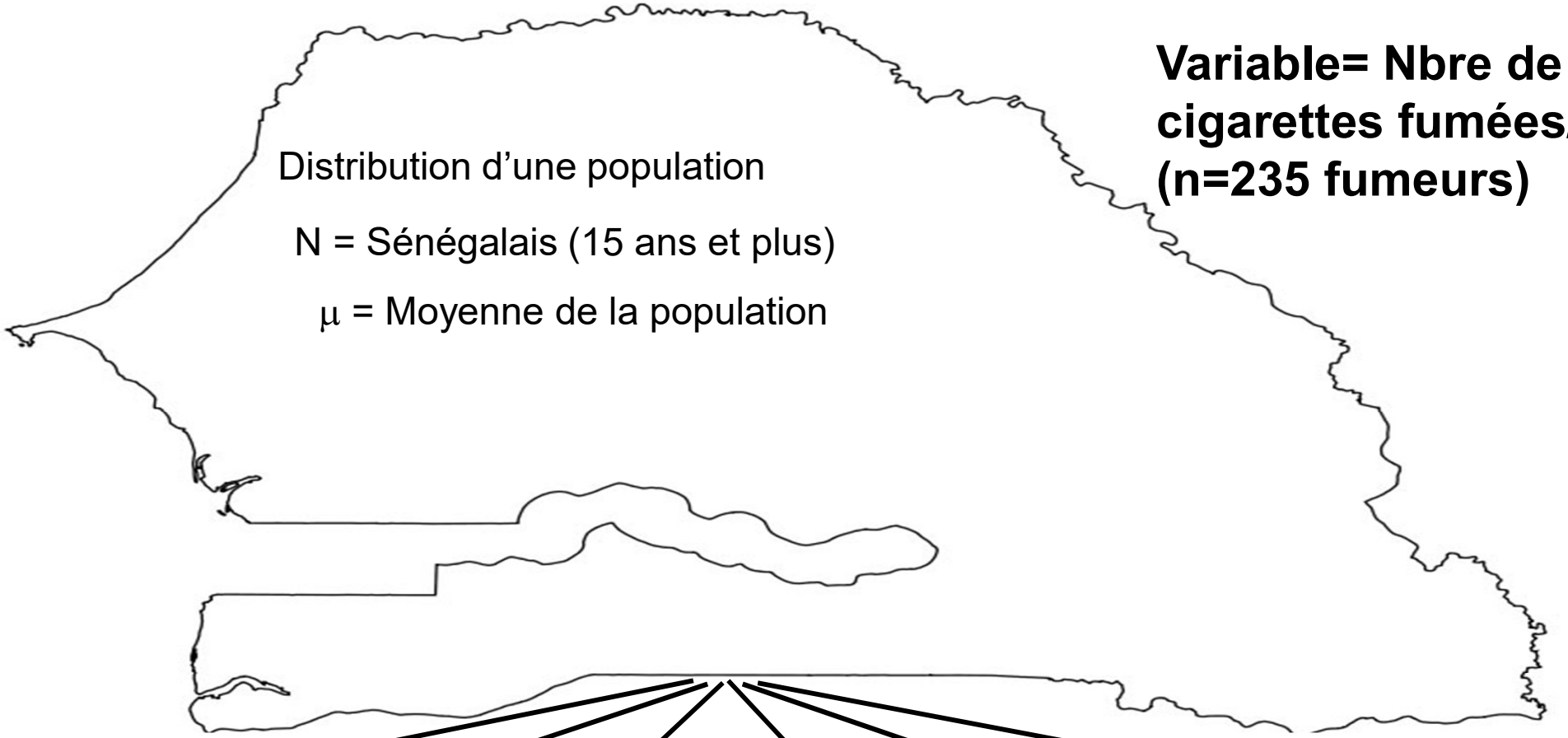


Estimation d'un paramètre

Distribution d'échantillonnage: Théorème central limite

- Au fur et à mesure que la taille de ***n*** augmente, la distribution des moyennes $\overline{\mathbf{X}}$ ou proportions **p** provenant d'échantillons aléatoires tend à se distribuer normalement autour de la moyenne μ ou de la proportion π
- La moyenne de la distribution d'échantillonnage des $\overline{\mathbf{X}}$ ($\mu_{\overline{x}}$) est semblable à celle de la population μ et son écart-type $\sigma_{\overline{x}}$, appelé encore erreur-type, est : $\sigma_{\overline{x}} = \sigma / \sqrt{n}$
- La moyenne de la distribution d'échantillonnage des **p** ($\mu_{\mathbf{p}}$) est semblable à la proportion de la population π et son écart-type $\sigma_{\mathbf{p}}$, appelé aussi erreur-type, est: $\sigma_p = \sqrt{pq / n}$

Distribution d'une population, distribution d'un échantillon et distribution d'échantillonnage d'une **moyenne**



Variable= Nbre de cigarettes fumées/j (n=235 fumeurs)

Distribution d'une population

N = Sénégalais (15 ans et plus)

μ = Moyenne de la population

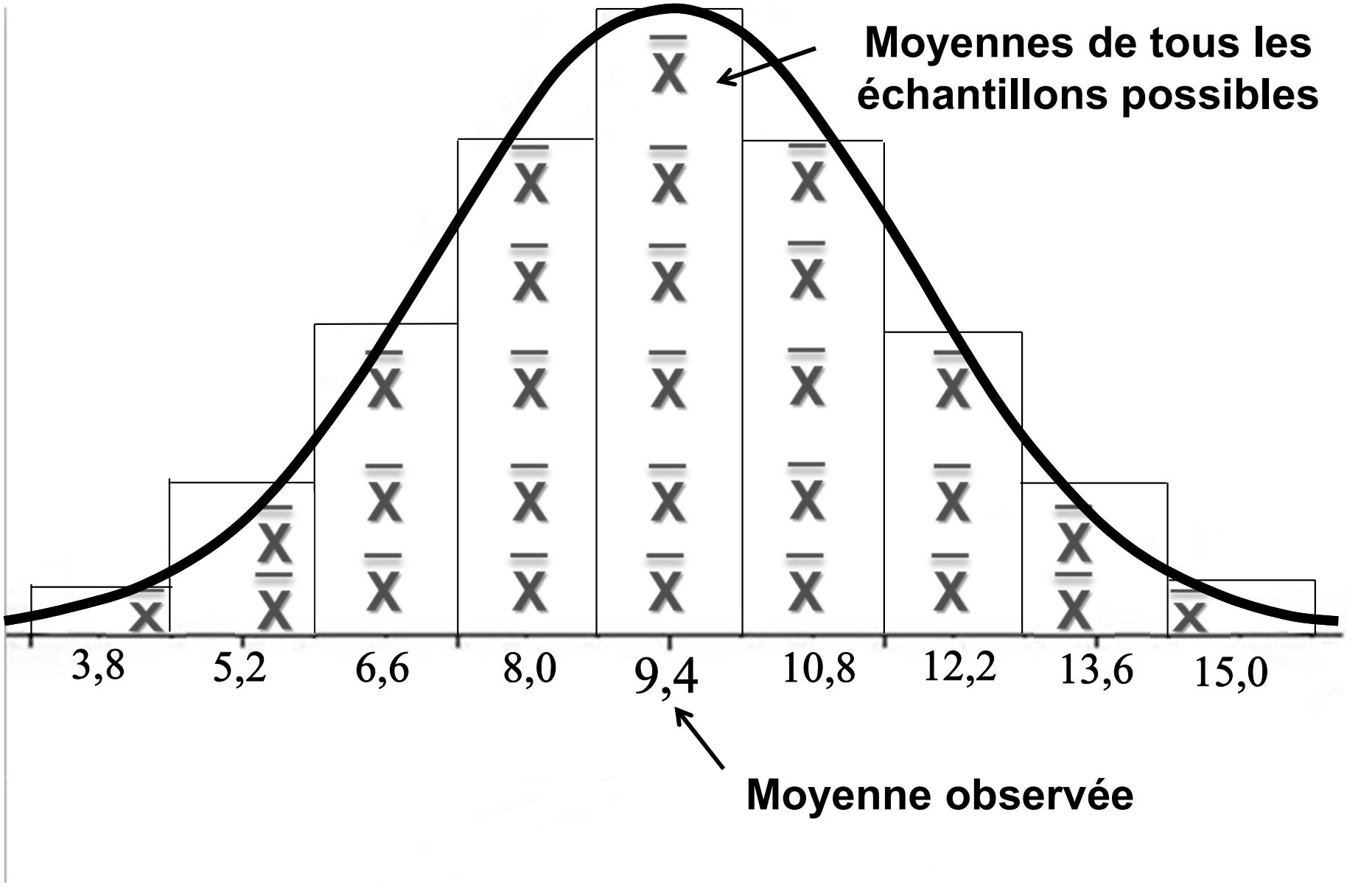
1er échant.	2e échant.	3e échant.	4e échant.	5e échant.	. . .	Der. échant.
n = 235	n = 235	n = 235	n = 235	n = 235		n = 235
$\overline{X}$ = **9,4**	$\overline{X}$ = 9,0	$\overline{X}$ = 10	$\overline{X}$ = 8,5	$\overline{X}$ = 8,0		$\overline{X}$ = ?

Distribution d'échantillonnage

$\mu_{\overline{X}} = \mu$

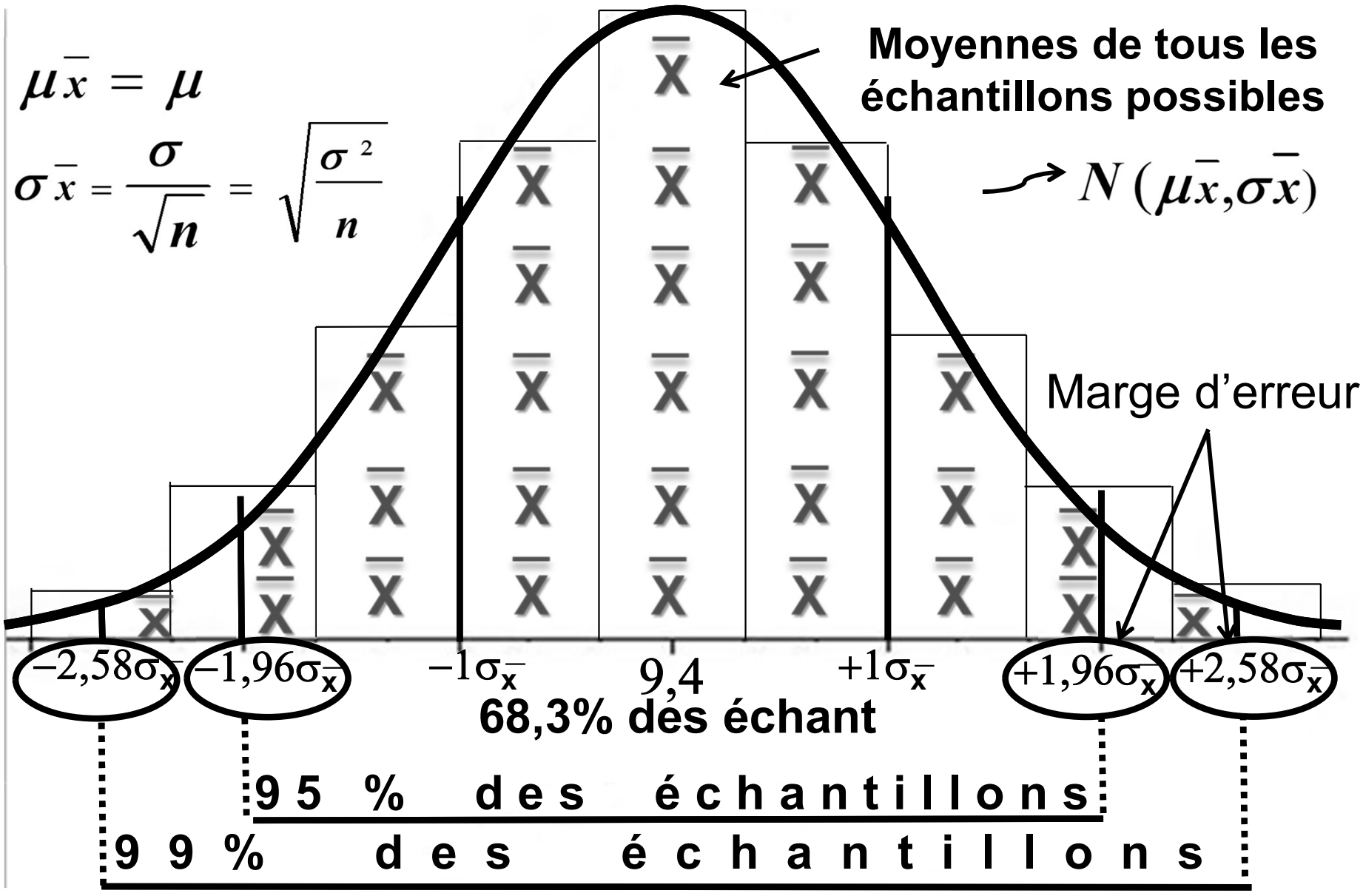
Intervalle de confiance d'une moyenne

Illustration



Intervalle de confiance d'une moyenne

Illustration



Intervalle de confiance d'une moyenne

Formule

- Pour le trouver, rappelons-nous de deux choses:
  - La distribution d'échantillonnage d'une moyenne suit **N** (μ, $\sigma_{\bar{x}}$)
  - Or, 95% des données d'une distribution normale **N** sont comprises à l'intérieur de 1,96 écart-type de part et d'autre de la moyenne
- Pour connaître les limites inf. et sup. de l'intervalle, il faut :
  - Soustraire 1,96 erreur-type de la moyenne de l'échantillon
  - Et additionner 1,96 erreur-type à la moyenne de l'échantillon
- Formule de l'intervalle de confiance (IC) d'une moyenne:

  IC à 95% = $\overline{X} \pm 1{,}96\sigma_{\bar{x}}$ $\Longrightarrow$ $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

  IC à 99% = $\overline{X} \pm 2{,}58\sigma_{\bar{x}}$

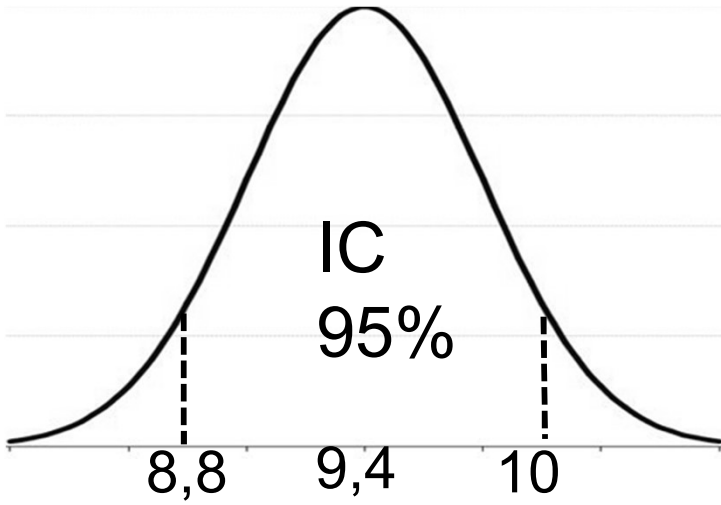
  - 95% | 99% = <u>niveau de confiance,</u> $1{,}96\sigma_{\bar{x}}$,| $2{,}58\sigma_{\bar{x}}$= <u>marge d'erreur</u>

Intervalle de confiance d'une moyenne

Exercice-éclair

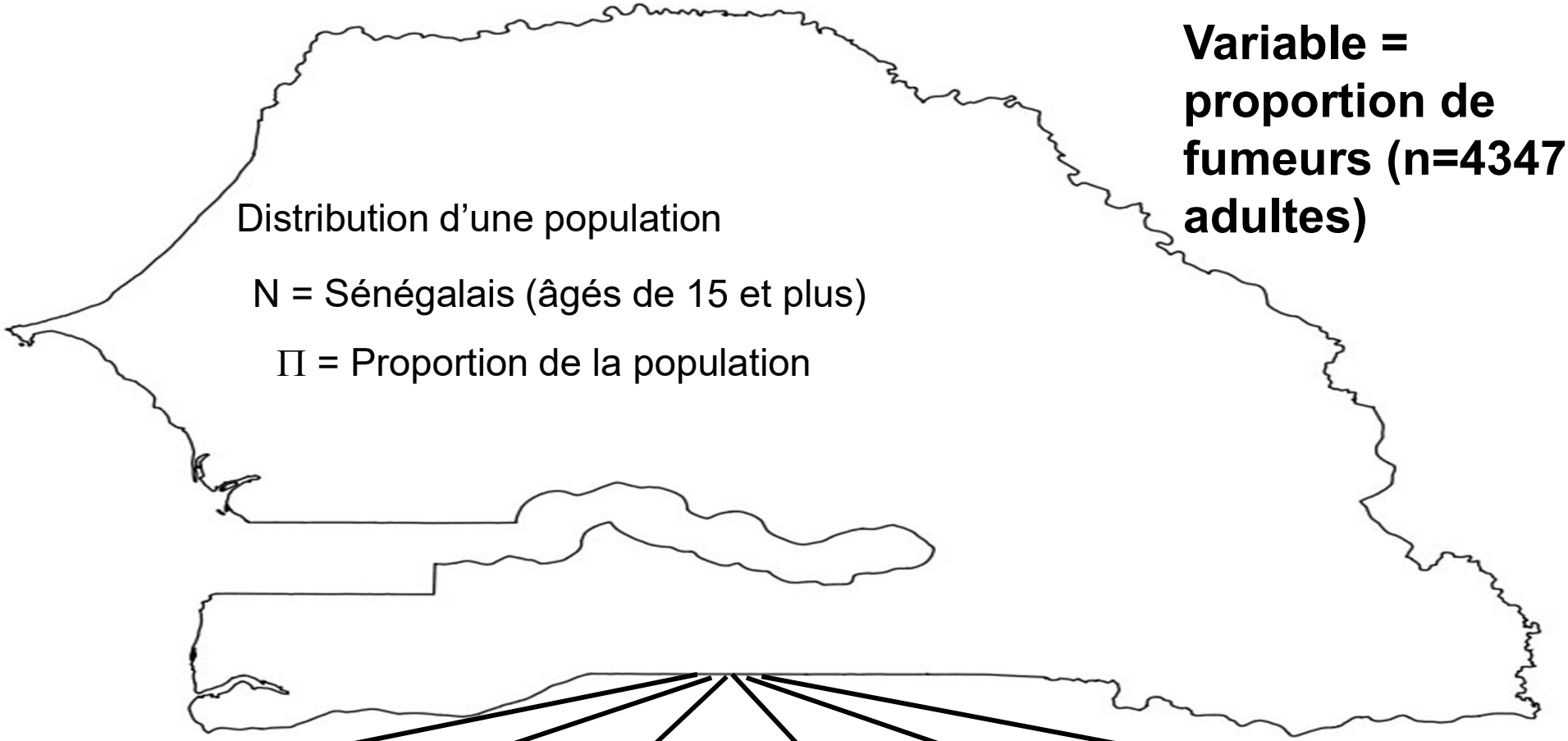
→ Un sondage aléatoire montre que chez les 235 (n) fumeurs sénégalais, le nbre moyen de cigarettes fumées/ jour est de 9,4 avec un écart-type de 4,7. Estimez par intervalle de confiance à 95%, la vraie moyenne dans la pop.

1. Calculez l'erreur type: $\sigma_{\bar{x}} = \sigma / \sqrt{n} = 4,7/\sqrt{235} = 0,31$
2. Calculez la marge d'erreur: $E = 1,96\sigma_{\bar{x}} = 1,96*0,31 = 0,6$
3. Déterminez l'intervalle: $IC = \overline{X} \pm E = 8,8$ et 10
4. Représentez graphiquement l'intervalle



→ On est sûr à 95% que le nbre moyen de cig. fumées par jour se situe entre 8,8 et 10 dans la population sénégalaise adulte. Ou le nbre moyen est de 9,4 cigarettes, avec une marge d'erreur de ±0,6, 95 fois sur 100

Distribution d'une population, distribution d'un échantillon et distribution d'échantillonnage d'une **proportion**



Variable = proportion de fumeurs (n=4347 adultes)

Distribution d'une population

N = Sénégalais (âgés de 15 et plus)

Π = Proportion de la population

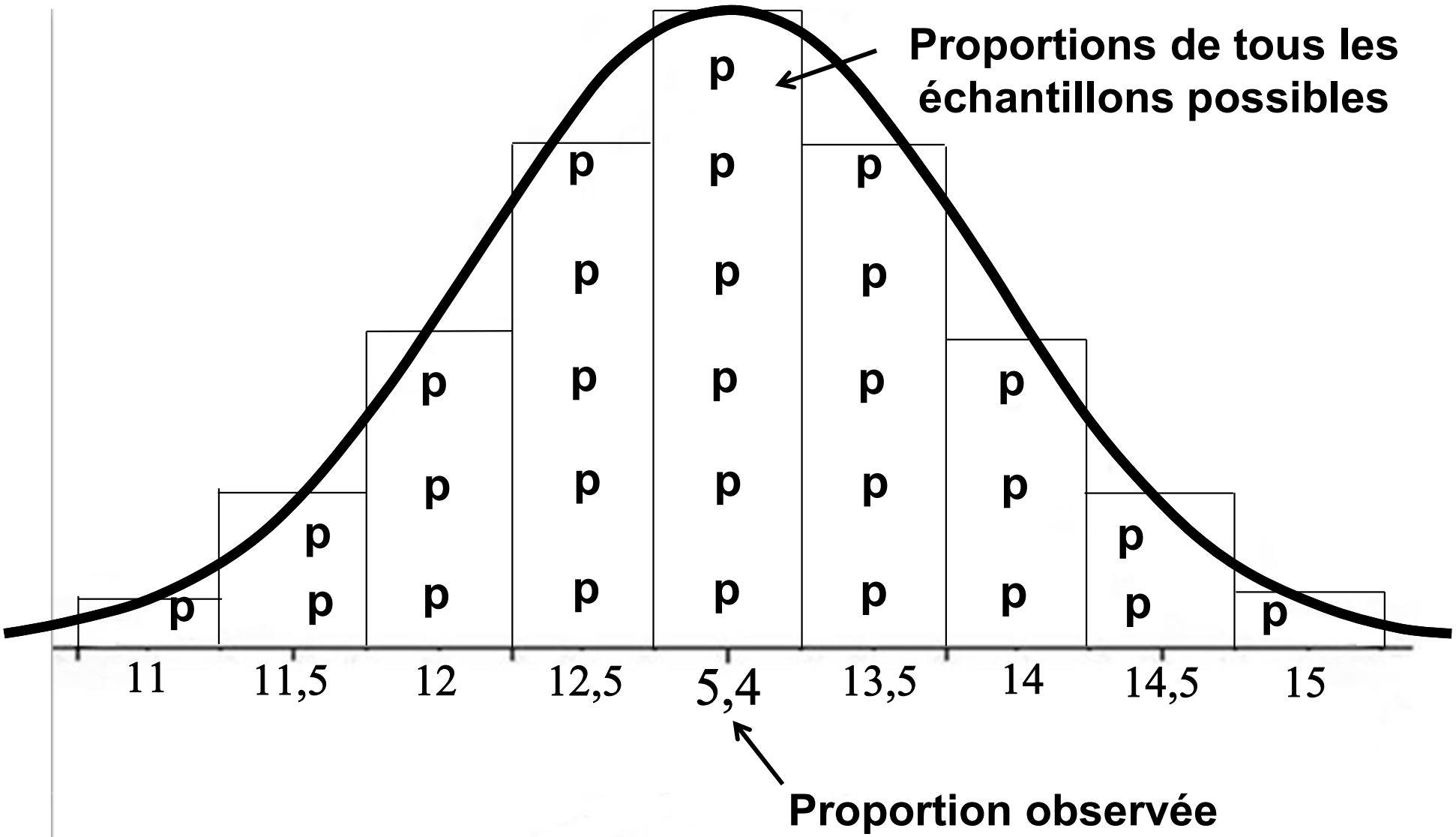
1er échant.	2e échant.	3e échant.	4e échant.	5e échant.	...	Der. échant.
n = 4347	n = 4347	n = 4347	n = 4347	n = 4347		n = 4347
p = 5,4	p = 5,0	p = 4,8	p = 6,0	p = 5,9		p = ?

Distribution d'échantillonnage

$\mu_p = \Pi$

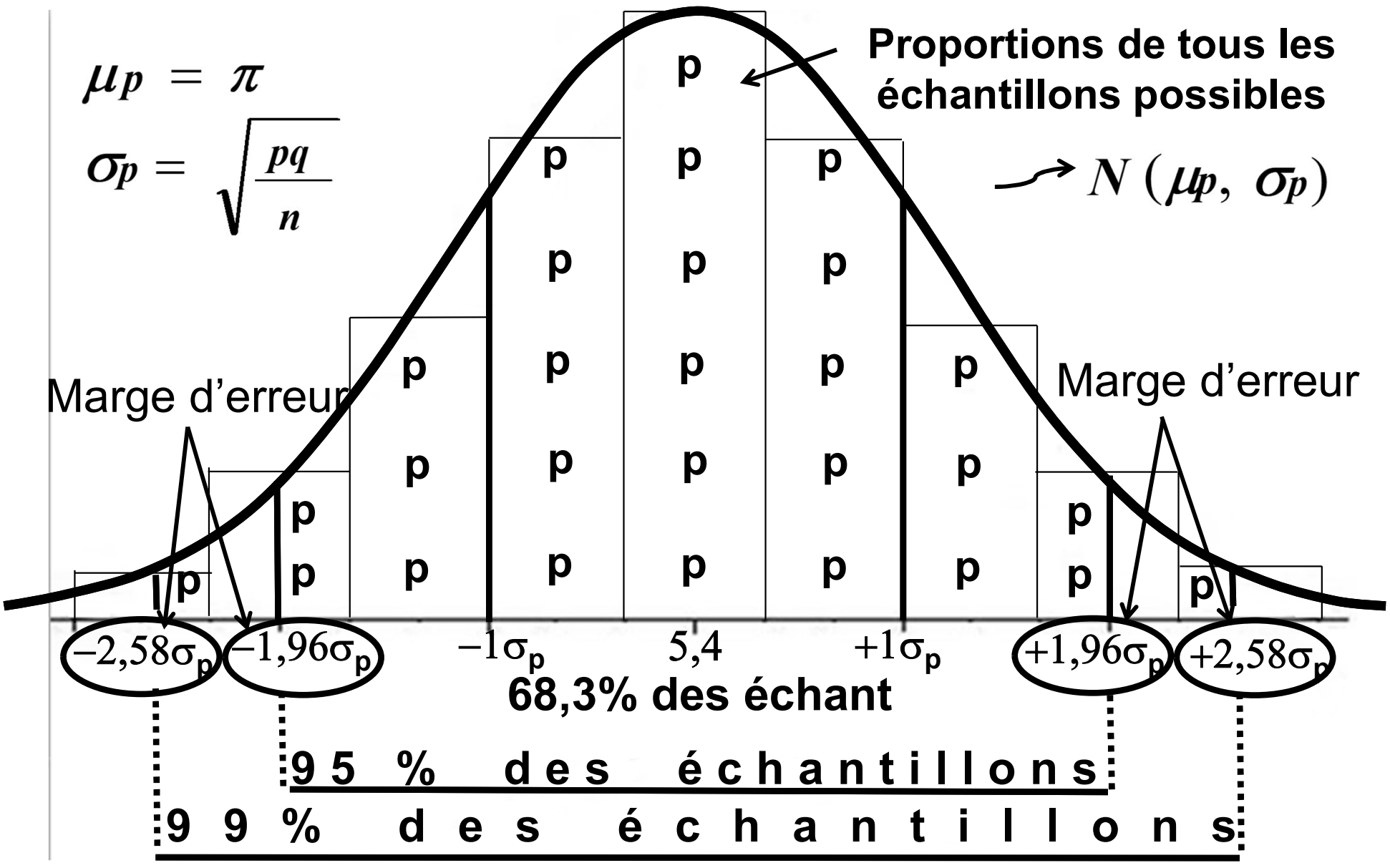
Intervalle de confiance d'une proportion

Illustration



Intervalle de confiance d'une proportion

Illustration



Intervalle de confiance d'une proportion

Formule

- Comme la moyenne, un pourcentage (ou proportion) peut se situer à l'intérieur d'un intervalle de confiance
- Exemple du sondage où 5,4% des 4347 sénégalais sondés en 2002 fument du tabac
  - p: probabilité du pourcentage à estimer, soit 5,4%
  - q: probabilité du pourcentage complémentaire, soit 100-p = 94,6%
  - n : nombre de cas enquêtés, soit 4347
- Formule de l'intervalle de confiance (IC) d'une proportion:

IC à 95% = $p \pm 1{,}96\sigma_p$

IC à 99% = $p \pm 2{,}58\sigma_p$

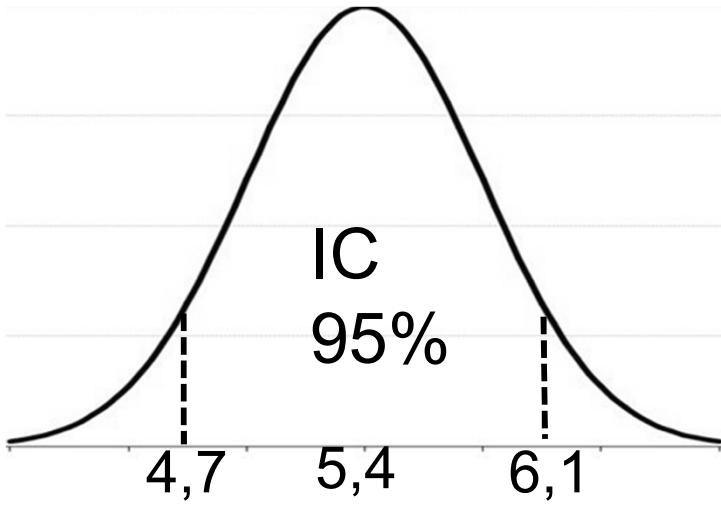
$\Longrightarrow$ $\sigma_p = \sqrt{pq / n}$

Intervalle de confiance d'une proportion

Exercice-éclair

→Un sondage aléatoire sur 4347 (n) sénégalais âgés de 15 ans et plus révèle que 5,4% fument du tabac. Estimez par intervalle de confiance à 95%, la vraie proportion de la pop.

1. Calculez l'erreur type: $\sigma_p = \sqrt{(pq)/n} = \sqrt{5,4*94,6/4347} = 0,35$
2. Calculez la marge d'erreur $E = 1,96\sigma_p = 1,96*0,35 = 0,7$
3. Déterminez l'intervalle: $IC = p \pm E = 4,7\%$ et $6,1\%$
4. Représentez graphiquement l'intervalle



→On est certain à 95% que la proportion de fumeurs se situe entre 4,7% et 6,1% dans la population sénégalaise adulte. Ou la proportion de fumeurs est de 5,4%, avec une marge d'erreur de ±0,7%, 95 fois sur 100

Estimation de paramètres

Effets de variation

- Variation de la taille de l'échantillon sur la marge d'erreur
  - Plus la taille de l'échantillon augmente, plus l'erreur-type (erreur standard moyenne) diminue
  - Ce qui, par ricochet, diminue la marge d'erreur ou l'IC
  - Donnant ainsi une estimation plus précise du paramètre
- Variation du niveau de confiance sur la marge d'erreur
  - Plus le niveau de confiance augmente, plus les chances de voir l'intervalle contenir la valeur de la population augmentent
  - Ce qui, toutefois, augmente la marge d'erreur ou l'IC
  - Donnant ainsi une estimation moins précise du paramètre

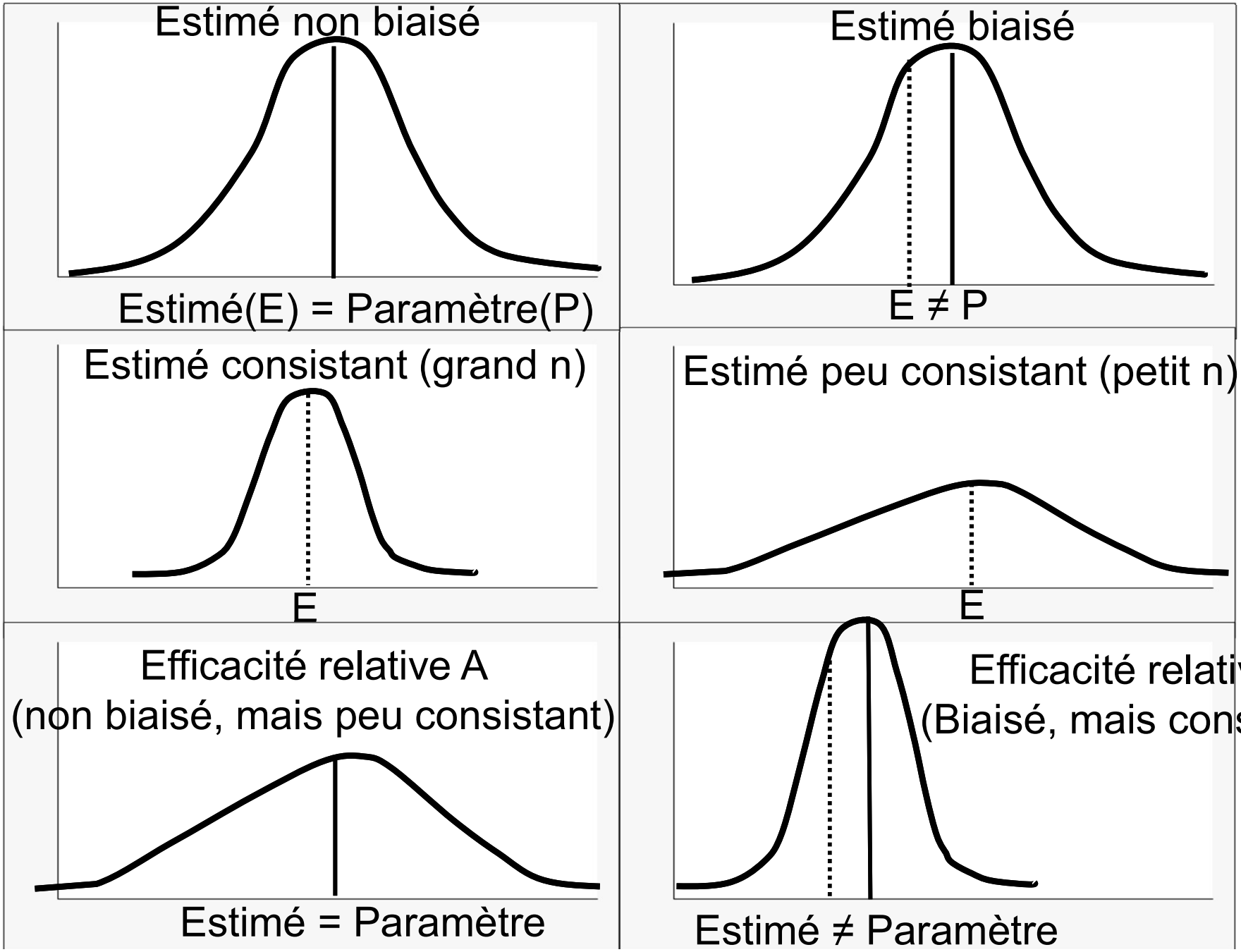
Estimation de paramètres

Propriétés des estimés | statistiques: définition

- Absence de biais
  - Un estimé est dit non biaisé s'il s'apparente (=) au paramètre, lorsque l'on constitue tous les échantillons possibles
    - Réduire le biais d'un estimé revient à constituer un ***n*** aléatoire
- Consistance
  - Un estimé est consistant s'il se rapproche du paramètre, sa distribution d'échantillonnage comportant une faible variabilité
    - Augmenter la consistance revient à augmenter la taille de ***n***
- Efficacité relative
  - L'efficacité relative rend compte de la précision relative avec laquelle un estimé estime un paramètre. Un estimé est d'autant plus efficace qu'il est non biaisé et consistant

Estimation de paramètres

Propriétés des estimés | statistiques: illustration



Tout prochainement

- Prochaine leçon
  - Analyse de tableaux croisés et test du chi-carré
- Au labo Excel de cette semaine
  - Calculer l'erreur-type, la marge d'erreur et l'intervalle de confiance d'une moyenne et les représenter graphiquement
  - Calculer l'erreur-type, la marge d'erreur et l'intervalle de confiance d'une proportion et les représenter graphiquement
  - Comparer deux groupes ou plus à l'aide de la barre d'erreur de façon à établir s'il y a une différence
  - Calculer la marge d'erreur maximale d'un sondage

7:41