STATISTIQUES POUR MEDECINS

Estimation

Christophe Combescure

Unité d'Appui Méthodologique du CRC

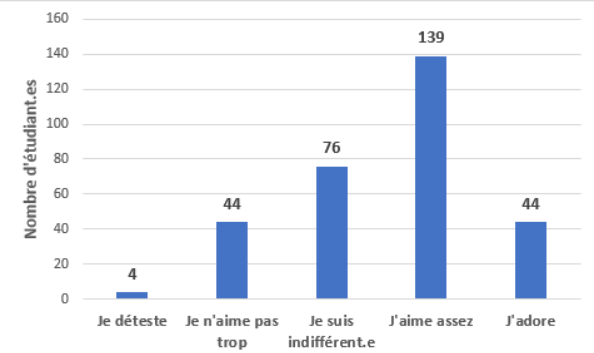
Rappel:
Analyses descriptives

- Description des données est une phase initiale importante de l'analyse des données
- Comprendre la composition de l'échantillon
- Décrire objectivement les résultats
- Statistiques descriptives
  - résumé simple et informatif des données
- Représentations graphiques:
  - message visuel clair
  - montrer toutes les données
  - rapport information/encre élevé

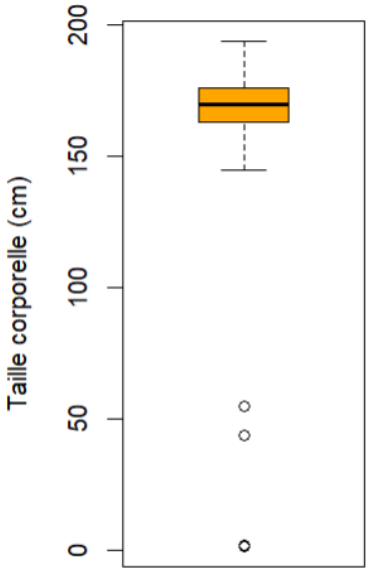
Rappel
Description d'une variable (1)

Relation aux chiffres
Variable qualitative ordinale
N=307 observations

	N (%)
Déteste	4 (1.3%)
N'aime pas trop	44 (14.3%)
Indifférent.e	76 (24.8%)
Aime assez	139 (45.3%)
Adore	44 (14.3%)
Total	**307 (100%)**



Votre taille corporelle (cm)
Variable quantitative continue
N=307 observations

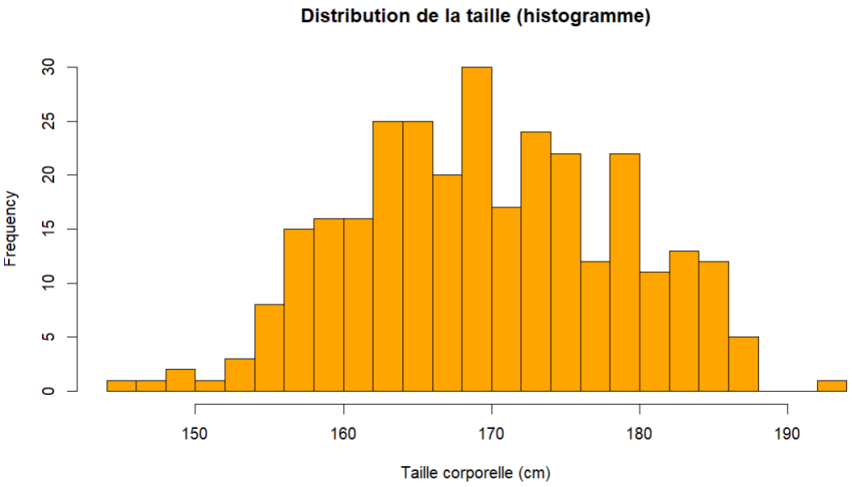


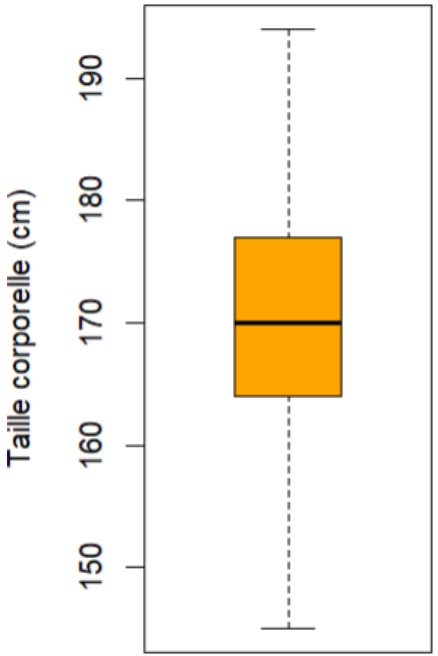
Moyenne : 167.7 cm
Ecart type: 21.1 cm
Médiane: 163 cm
Ecart inter-quartile: 163 à 176 cm

La représentation graphique de la distribution de la taille met en évidence des valeurs aberrantes

Rappel
Description d'une variable (2)

Votre taille corporelle (cm)
Variable quantitative continue
N=302 observations

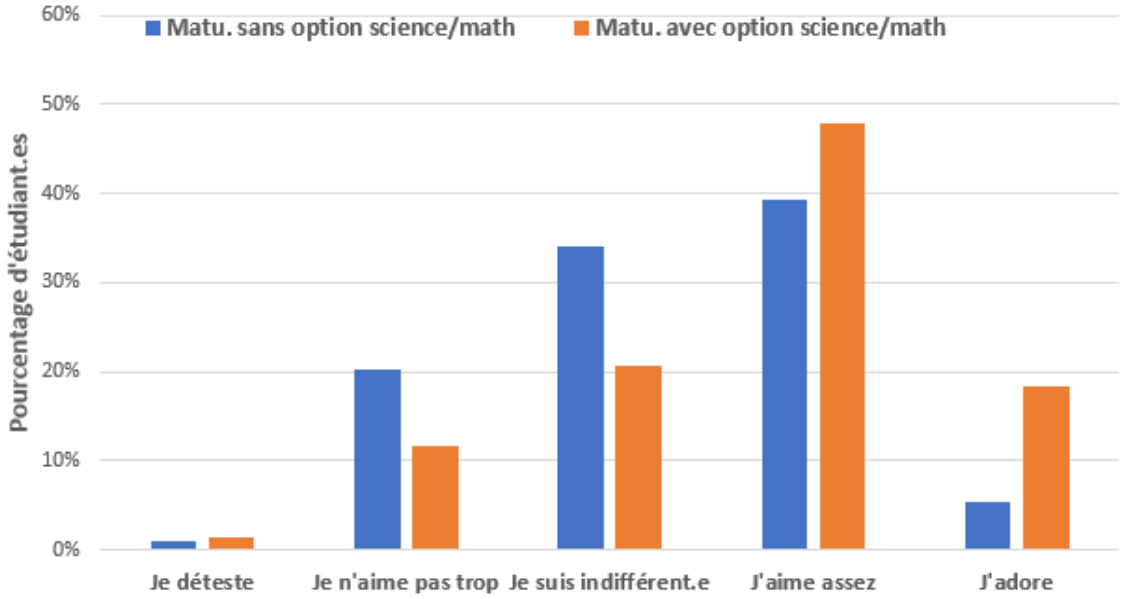




Moyenne : 170.1 cm
Ecart type: 8.9 cm
Médiane: 170 cm
Ecart inter-quartile: 164 à 177 cm

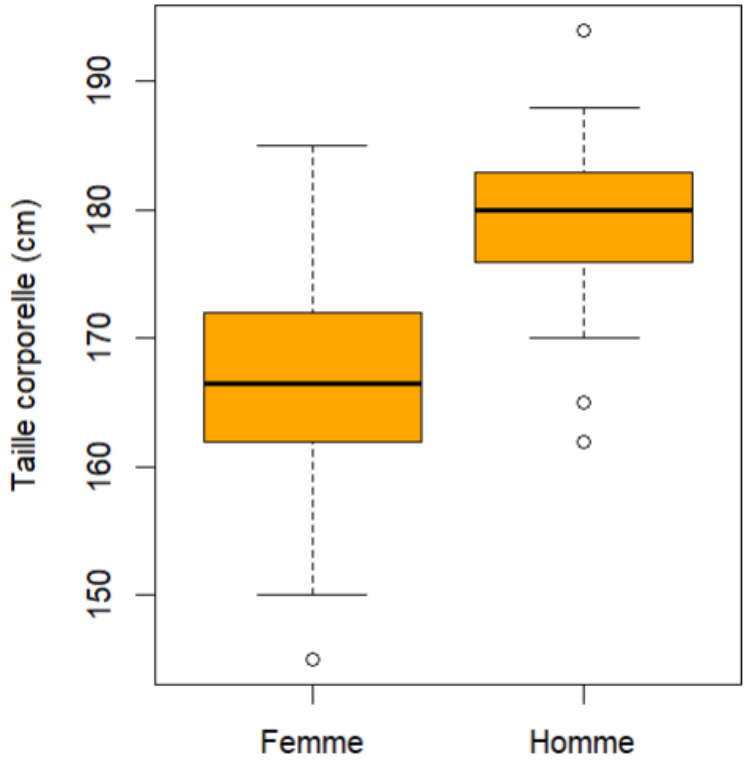
Rappel:
Description conjointe de deux variables (1)

	Maturité option science / math.	
	Non	**Oui**
Déteste	1 (1.1%)	3 (1.4%)
N'aime pas trop	19 (20.2%	25 (11.7%)
Indifférent.e	32 (34.0%)	44 (20.7%)
Aime assez	37 (39.4%)	102 (47.9%)
Adore	5 (5.3%)	39 (18.3%)
Total	**94 (100%)**	**213 (100%)**

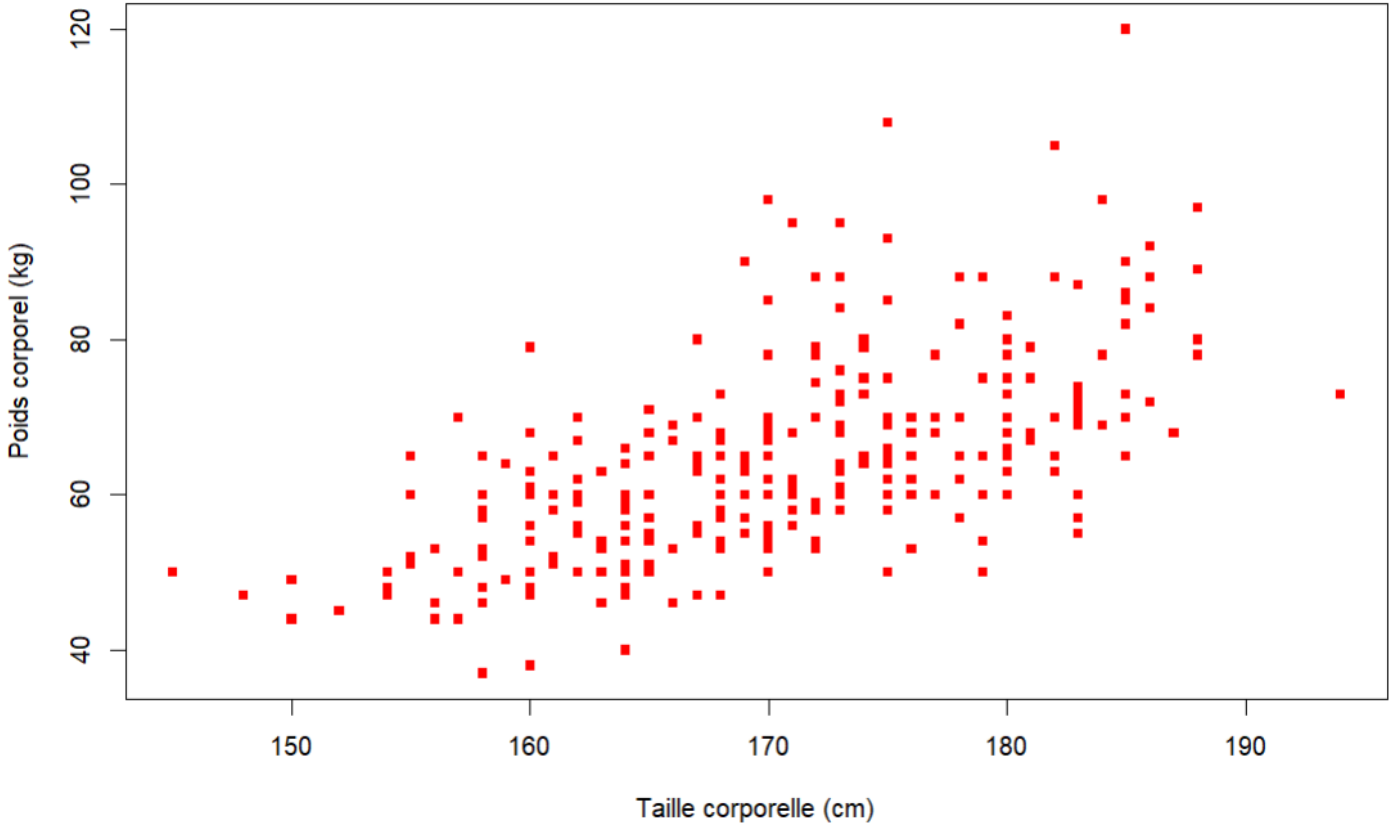


Rappel:
Description conjointe de deux variables (2)

Taille moyenne (écart type):
- étudiantes (n=216) : 166.5 cm (6.9 cm)
- étudiants (n=84) : 179.5 cm (5.9 cm)



Rappel:
Description conjointe de deux variables (3)



Rappel
Mesures d'association

- Pour évaluer:
  - effet d'une intervention expérimentale par rapport à un comparateur
  - ampleur de la relation entre une exposition et un problème de santé

- Le type de mesure d'association dépend de la nature statistique des variables:
  - différence de moyenne
  - différence ou ratio de proportion
  - …

Objectifs

- **Comprendre les notions suivantes**
  - Estimation de paramètres
  - Intervalle de confiance
  - Echantillonnage
  - Types d'erreur (aléatoire et systématique)

Chapitres Petrie/Sabin
7 – 10 – 11

Inférence
Principe

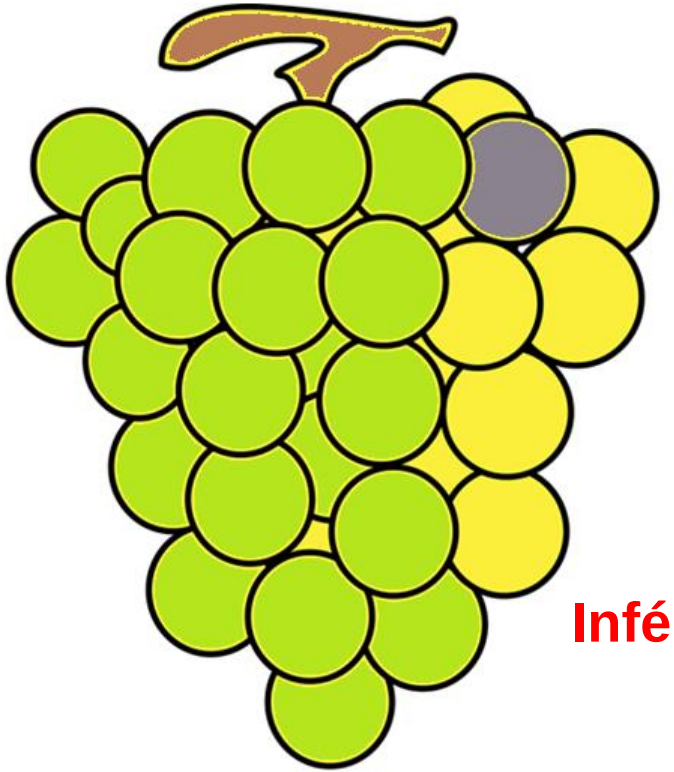
- Comment avoir une idée de la qualité de la grappe de raisin ?

- Goûter un des raisins :

  si le raisin est bon, vous concluez que la grappe est bonne

  si le raisin est mauvais, vous concluez que la grappe est mauvaise

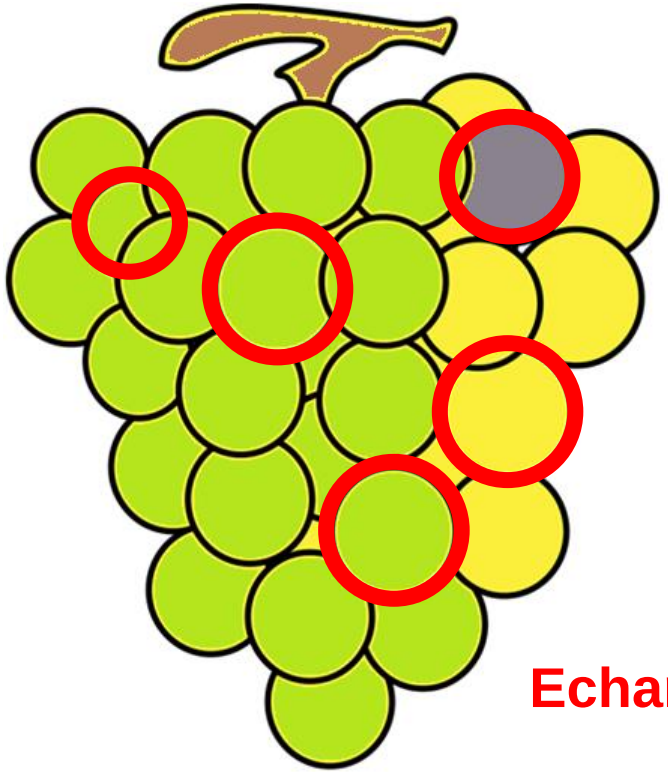
- Vous généralisez à l'ensemble de la grappe le goût du raisin testé



Inférence

Inférence
Echantillonnage

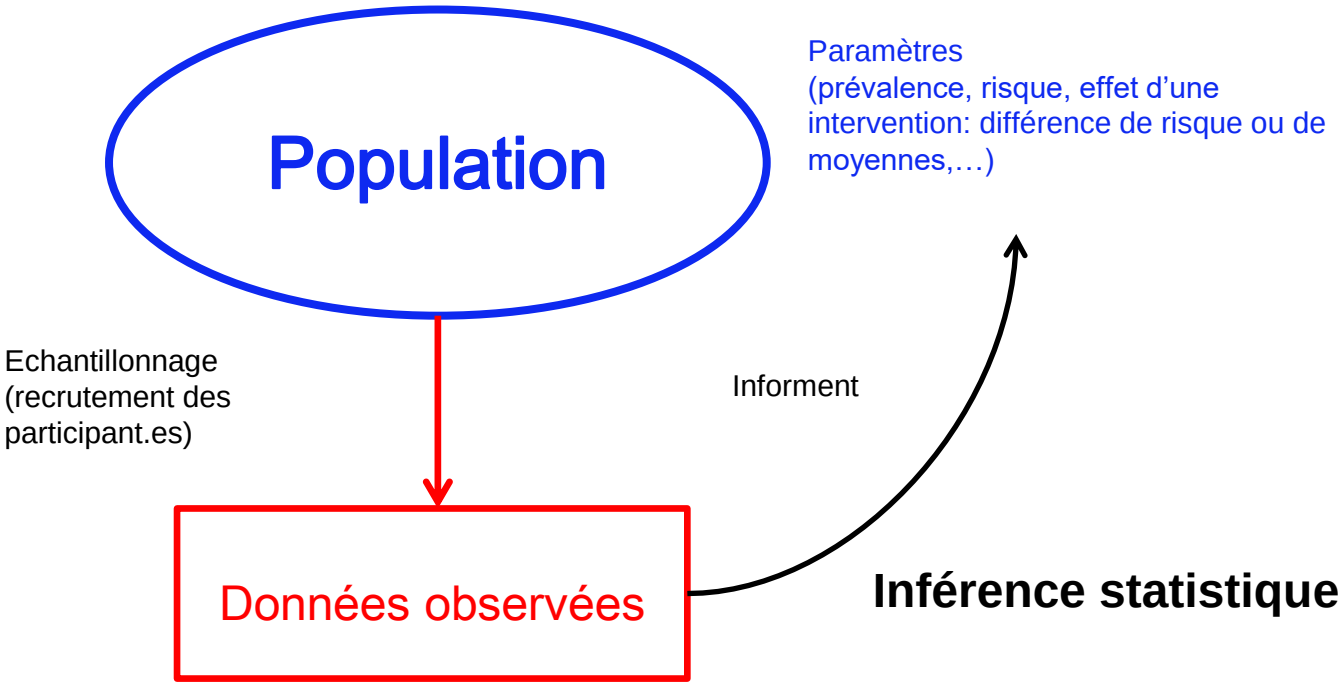
- Comment sélectionner le raisin à goûter ?
- Sélection subjective ?
  risque de goûter surtout les jolis raisins => inférence incorrecte
- Une sélection aléatoire protège d'une sélection subjective
- Plusieurs raisins: échantillon



Echantillon

Inférence

Inférence statistique (1)

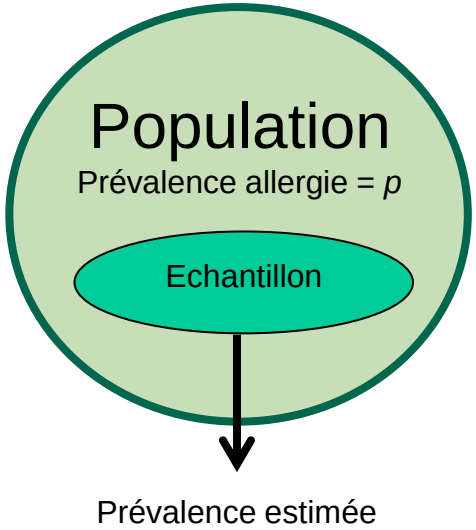


Inférence

Inférence statistique (2)

1. Question ou hypothèse de recherche posée à propos du monde en général (population cible), concernant un paramètre
2. Le paramètre n'est pas observable directement.
3. Sélection d'un échantillon approprié de la population:
   - représentatif
   - de taille suffisante
4. A partir des données de l'échantillon, on obtient une estimation du paramètre
5. Application des données de l'échantillon à l'hypothèse ou paramètre: inférence statistique
6. Réponse apportée avec un degré d'incertitude, une imprécision

Estimation d'une prévalence ou d'un risque



On veut connaître la prévalence d'allergie aux pollens dans la population (paramètre)

Allergie (non/oui): variable qualitative

Recueil de données dans un échantillon aléatoire de N personnes

Observation de x cas d'allergie

Estimation du paramètre «prévalence d'allergie»: $\hat{p} = x/N$

Estimation d'une prévalence ou d'un risque

Exemple

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Hyponatremia among Runners in the Boston Marathon

N Engl J Med 2005;352:1550-6

Excessive fluid intake is believed to be the primary risk factor for hyponatremia, on the basis of observations of marathon runners who have collapsed[2-5,7,11,12] and studies of elite athletes.[13-17] We undertook the present study to estimate prospectively the incidence of hyponatremia among marathon runners and to identify the principal risk factors involved.

Sur les 488 coureur.ses analysées, 62 cas d'hyponatrémie sont identifiés

Dans l'échantillon, le risque d'hyponatrémie est 12.7% (62/488)

Que peut-on dire du risque d'hyponatrémie dans la population des coureur.ses de marathon ?

Du paramètre à l'estimation

Simulations

- Objectif de la simulation:
  - observer la distribution des estimations d'une prévalence lorsqu'on échantillonne de manière aléatoire plusieurs fois la même population (avec la même prévalence dans la population)

- Etapes de la simulation:
  1) population avec une prévalence d'allergie de 30%
  2) échantillonnage aléatoire de 20 personnes de cette population
  3) distribution des estimations de prévalence sur un grand nombre d'échantillons
  4) mêmes procédures avec des échantillons de 50 personnes

http://www.distributome.org/V3/exp/BinomialExperiment.html

Du paramètre à l'estimation

Enseignement de la simulation

Echantillons aléatoires de n=20 personnes
Prévalence allergie dans la population: 30%

3 cas d'allergie observés dans cet échantillon
=> prévalence estimée = 3/20 = 15%

7 cas d'allergie observés dans cet échantillon
=> prévalence estimée = 7/20 = 35%

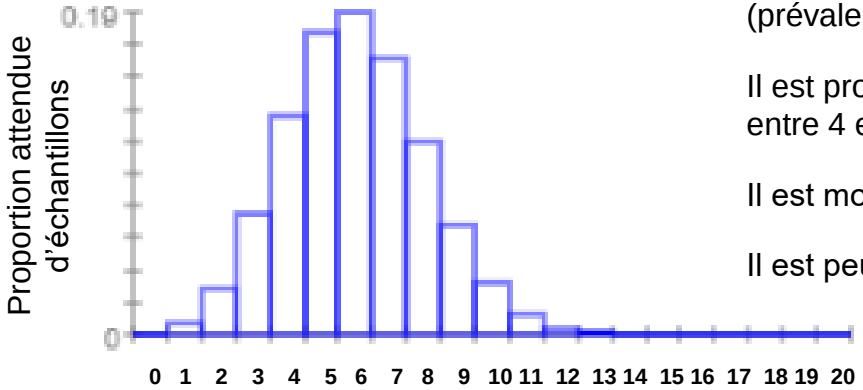
L'estimation du paramètre :

- peut varier d'un échantillon aléatoire à l'autre
- n'est pas forcément égale à la valeur du paramètre

Du paramètre à l'estimation
Enseignement de la simulation

Echantillons aléatoires de n=20 personnes
Prévalence allergie dans la population: 30%

Distribution du nombre de cas d'allergie par échantillon



Le nombre de cas le plus probable dans un échantillon est 6 (prévalence estimée = 30%)

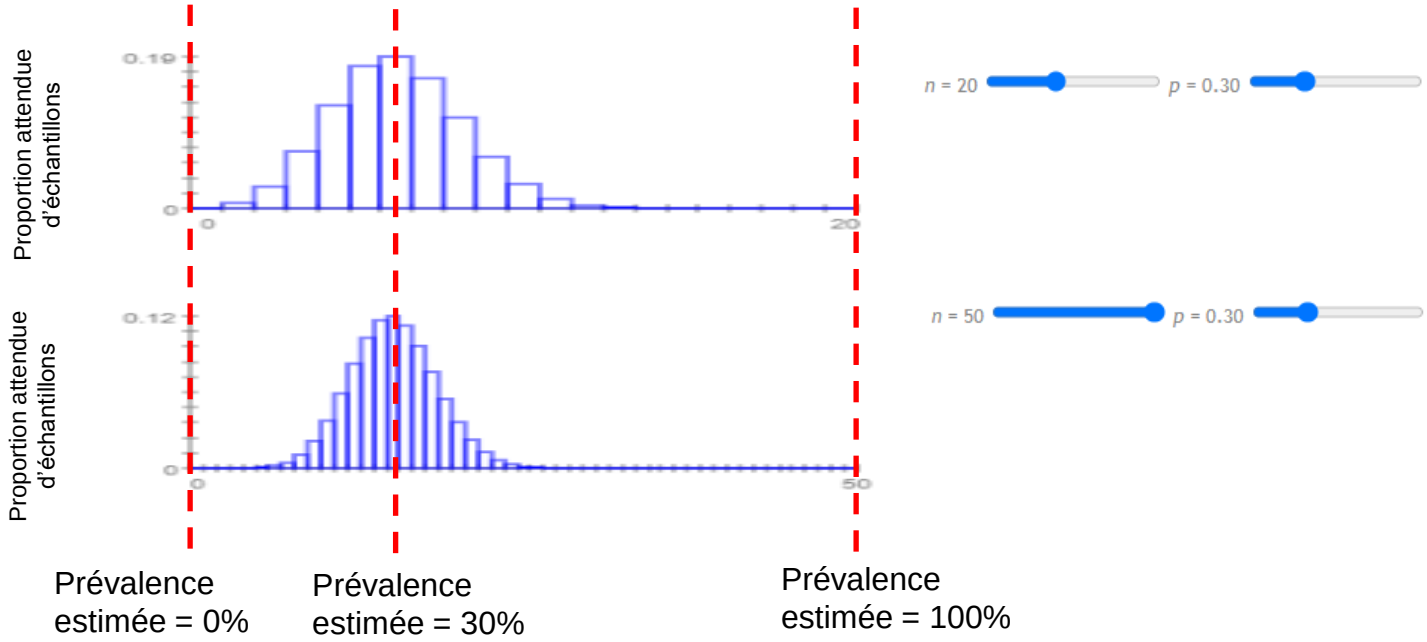
Il est probable que dans un échantillon aléatoire on observe entre 4 et 8 cas

Il est moins probable d'observer 2-3 ou 9-10 cas

Il est peu probable d'observer moins de 2 cas ou plus de 10 cas

Du paramètre à l'estimation

Enseignement de la simulation



Taille d'échantillon <u>augmente</u>

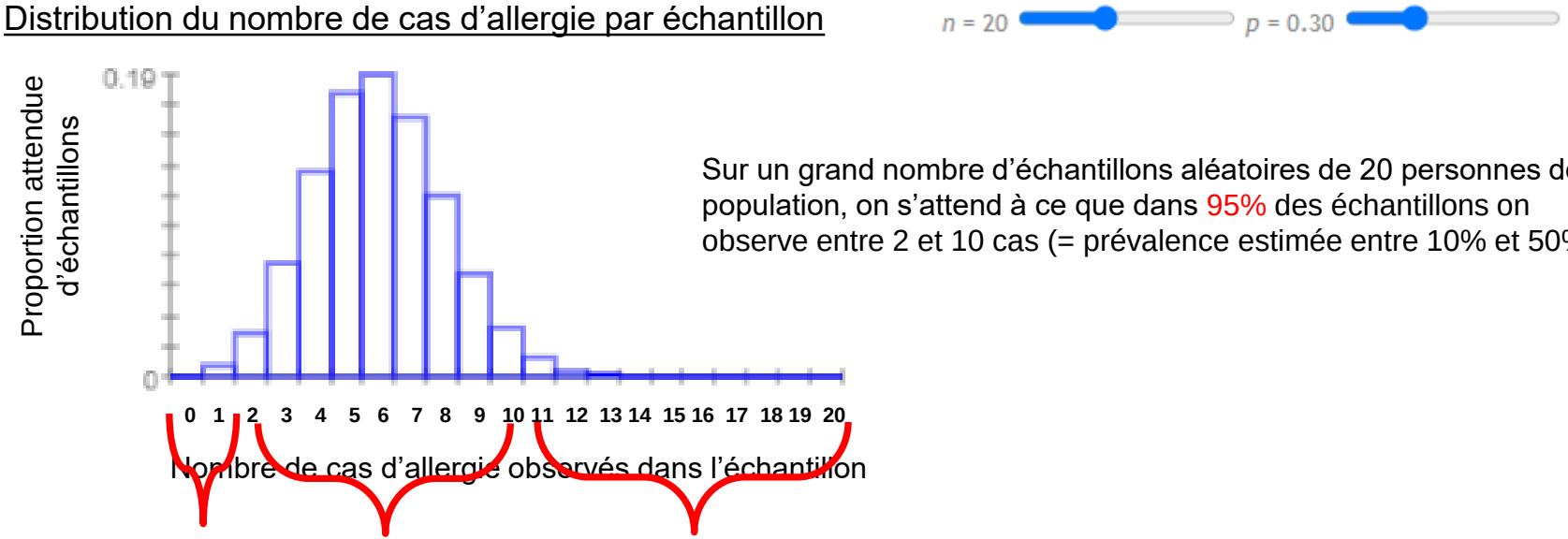
⇒ distribution des estimations plus resserrée autour de la prévalence 30%
(i.e. variance de cette distribution plus faible)

Du paramètre à l'estimation
Distribution des prévalences estimées

- L'estimation la plus fréquente est égale à la valeur du paramètre (pic de la distribution)
- Des estimations différentes de la valeur du paramètre sont possibles
  - plus l'estimation est éloignée de la valeur du paramètre, moins elle est fréquente (=> les estimations très éloignées sont très peu probables)
- Lorsque la taille d'échantillon augmente:
  - la variance de la distribution des estimations diminue (les estimations ont tendance à être plus proches de la valeur du paramètre)
- Cette distribution est connue grâce à la théorie des probabilités

Compatibilité entre estimation et paramètre

Distribution du nombre de cas d'allergie par échantillon



Sur un grand nombre d'échantillons aléatoires de 20 personnes de la population, on s'attend à ce que dans 95% des échantillons on observe entre 2 et 10 cas (= prévalence estimée entre 10% et 50%)

≈ 2.5% des échantillons

Prévalences estimées < 10%

≈ 95% des échantillons

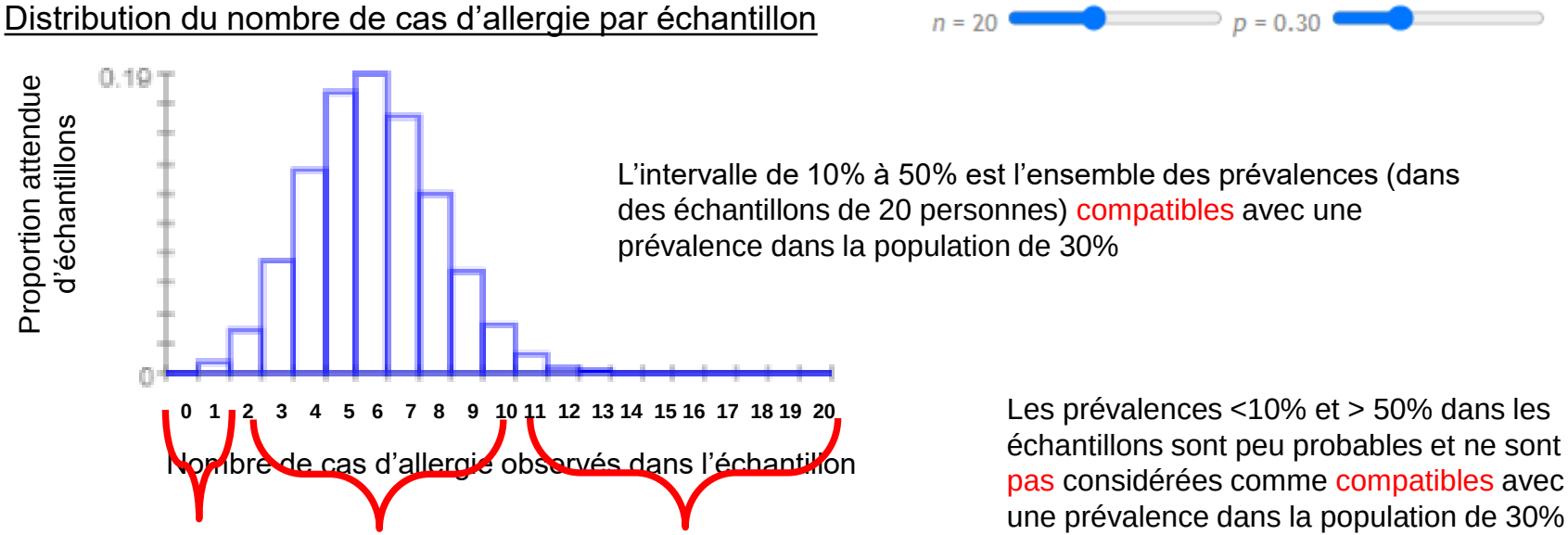
Prévalences estimées entre 10% et 50%

≈ 2.5% des échantillons

Prévalences estimées > 50%

Compatibilité entre estimation et paramètre

Distribution du nombre de cas d'allergie par échantillon



L'intervalle de 10% à 50% est l'ensemble des prévalences (dans des échantillons de 20 personnes) compatibles avec une prévalence dans la population de 30%

Les prévalences <10% et > 50% dans les échantillons sont peu probables et ne sont pas considérées comme compatibles avec une prévalence dans la population de 30%

≈ 2.5% des échantillons

Prévalences estimées < 10%

≈ 95% des échantillons

Prévalences estimées entre 10% et 50%

≈ 2.5% des échantillons

Prévalences estimées > 50%

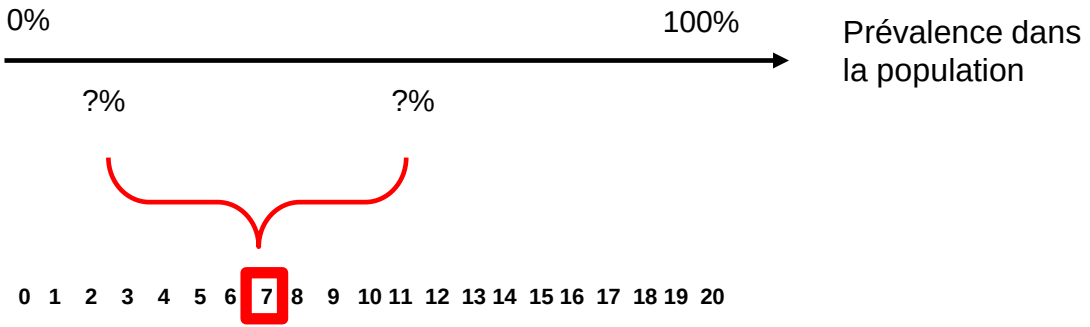
De l'estimation au paramètre
Incertitude de l'estimation

En pratique le.a chercheur.se :

1) n'a qu'un seul échantillon
2) ne connaît pas la prévalence dans la population

Si dans un échantillon de 20 personnes, 7 cas sont observés:

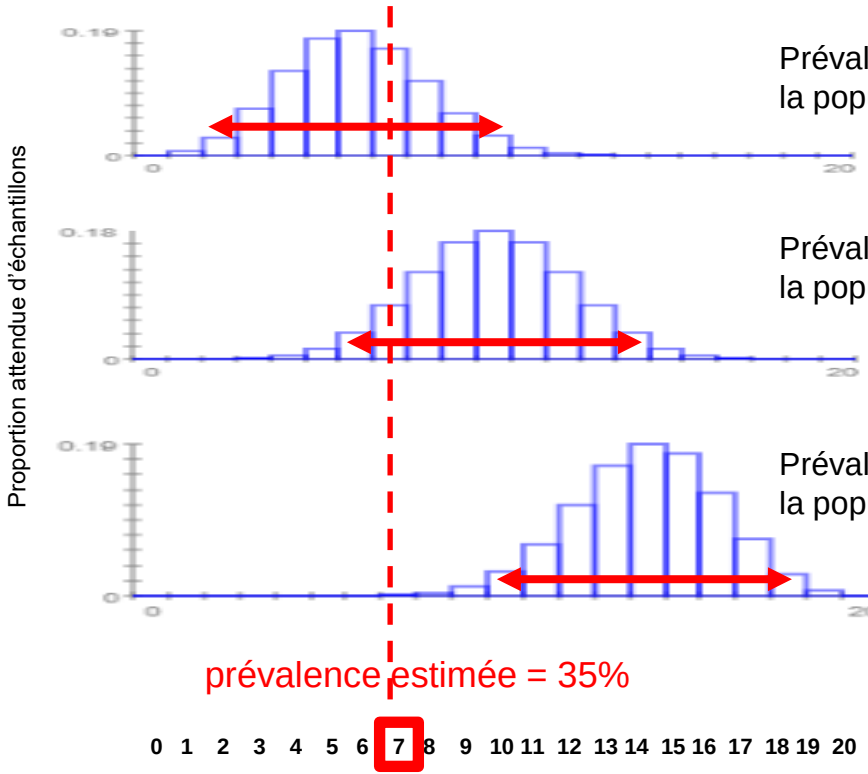
Ensemble des valeurs de prévalence dans la population compatibles avec les données de l'échantillon



7 cas : prévalence estimée = 35%

De l'estimation au paramètre

Incertitude de l'estimation



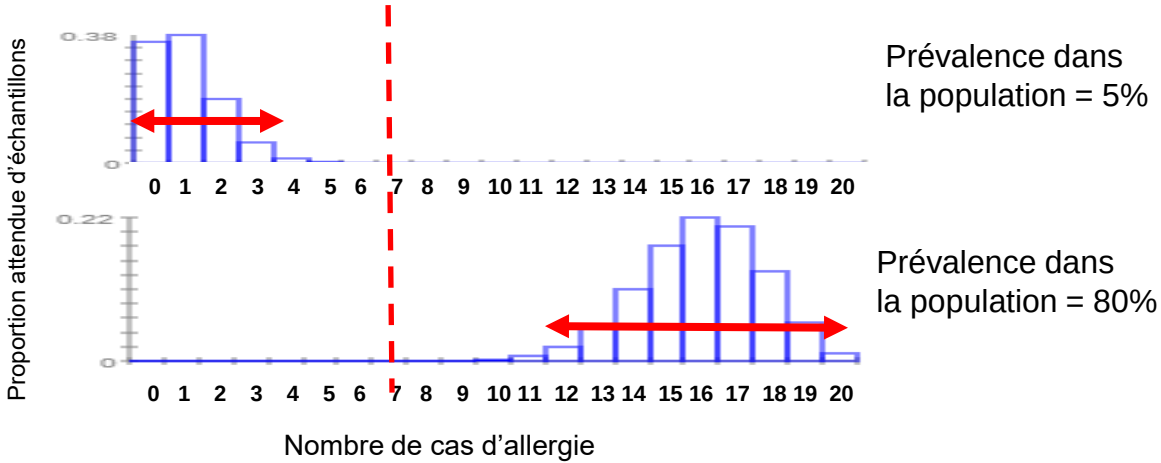
Compatible
avec les 7 cas observés

Compatible
avec les 7 cas observés

Pas compatible
avec les 7 cas observés

De l'estimation au paramètre

Incertitude de l'estimation

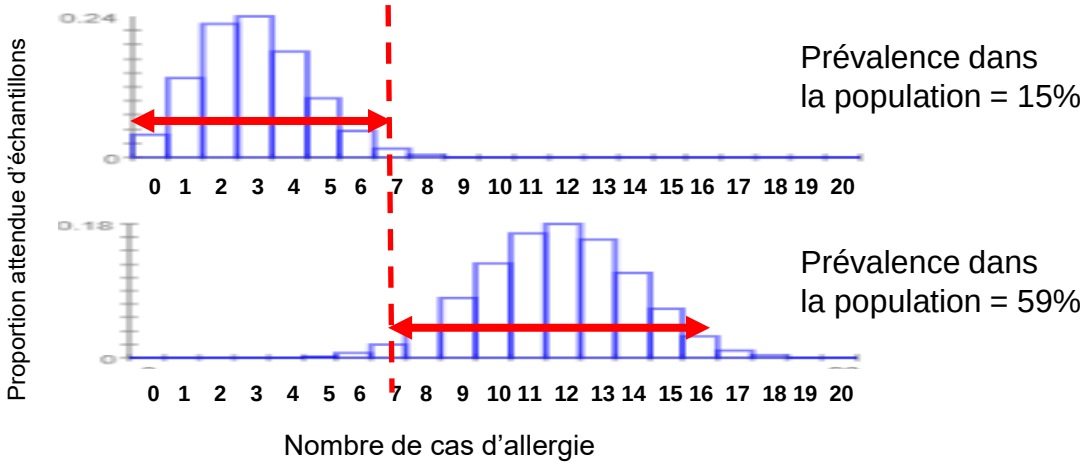


prévalence estimée = 35%

Des valeurs de paramètre de 5% ou 80% paraissent peu compatibles avec l'observation d'une estimation de 35%.
On rejette donc que la valeur du paramètre est ≤5% ou ≥80%.
Quid de 6%, 7%, 50% ou 78%?

De l'estimation au paramètre

Incertitude de l'estimation



prévalence estimée = 35%

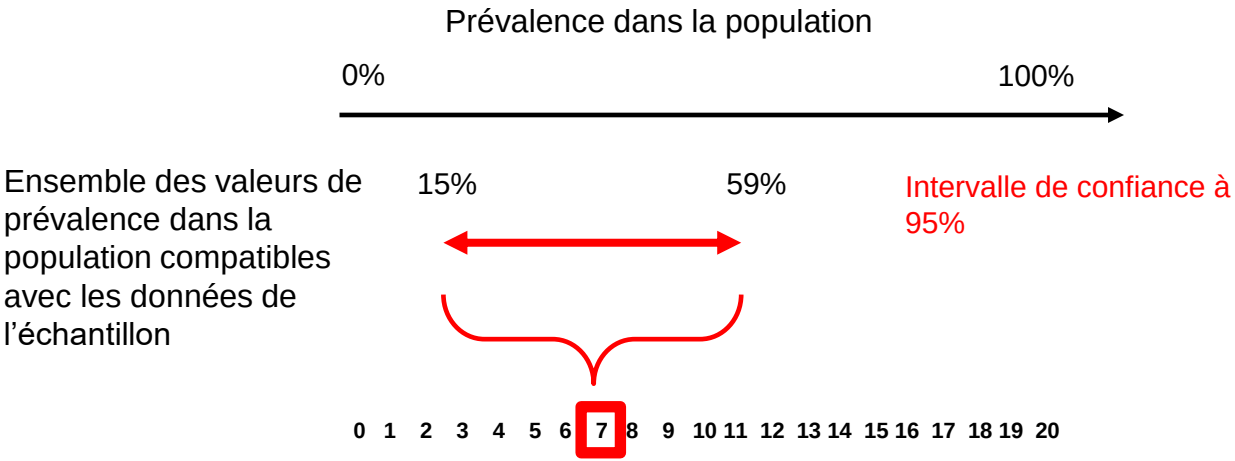
15% est la <u>plus petite valeur</u> de prévalence dans la population compatible avec les données
59% est la <u>plus grande valeur</u> de prévalence dans la population compatible avec les données

Toutes les valeurs de prévalence dans la population entre 15 et 59% sont compatibles
Aucune valeur de prévalence <15% et >59% n'est compatible avec les données.

De l'estimation au paramètre

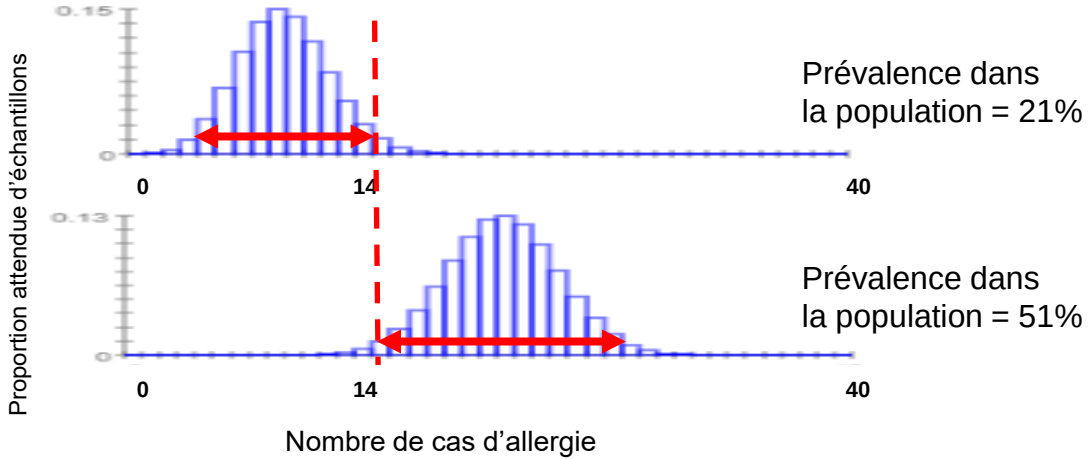
Intervalle de confiance (IC) à 95% d'une proportion

En pratique, le.a chercheur.se n'a qu'un seul échantillon
Si dans un échantillon de 20 personnes, 7 cas sont observés:



7 cas : prévalence estimée = 35%

IC 95% et taille d'échantillon



Si on observe 14 cas dans un échantillon de **40 personnes** (prévalence estimée = 35%):

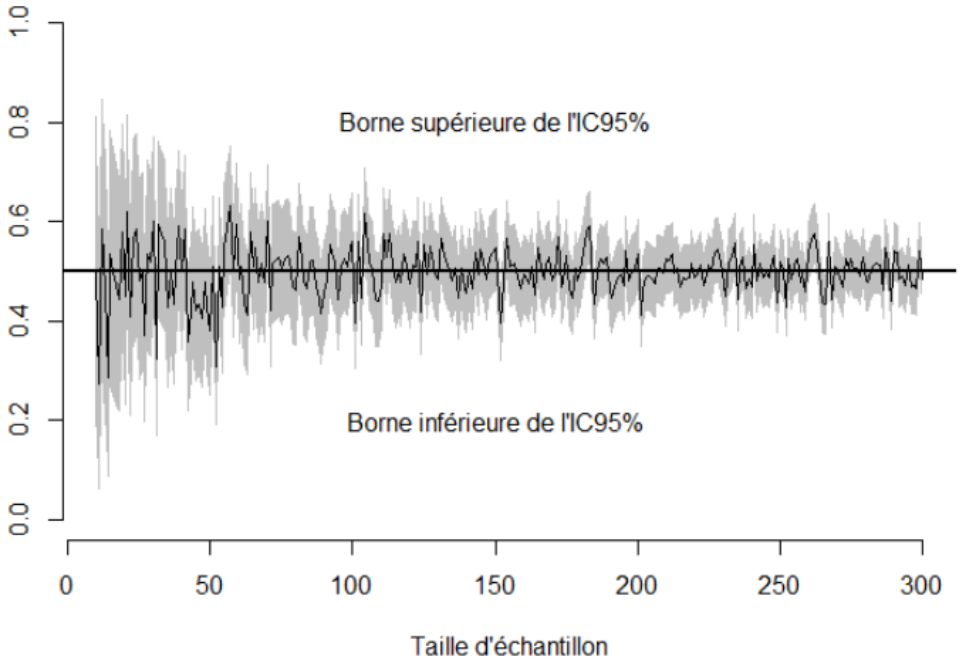
- 21% est la plus petite valeur de prévalence dans la population compatible
- 51% est la plus grande valeur de prévalence dans la population compatible

IC 95% = 21 à 51%

L'intervalle des valeurs du paramètre compatibles avec les données observées est plus étroit qu'avec un échantillon de 20 personnes

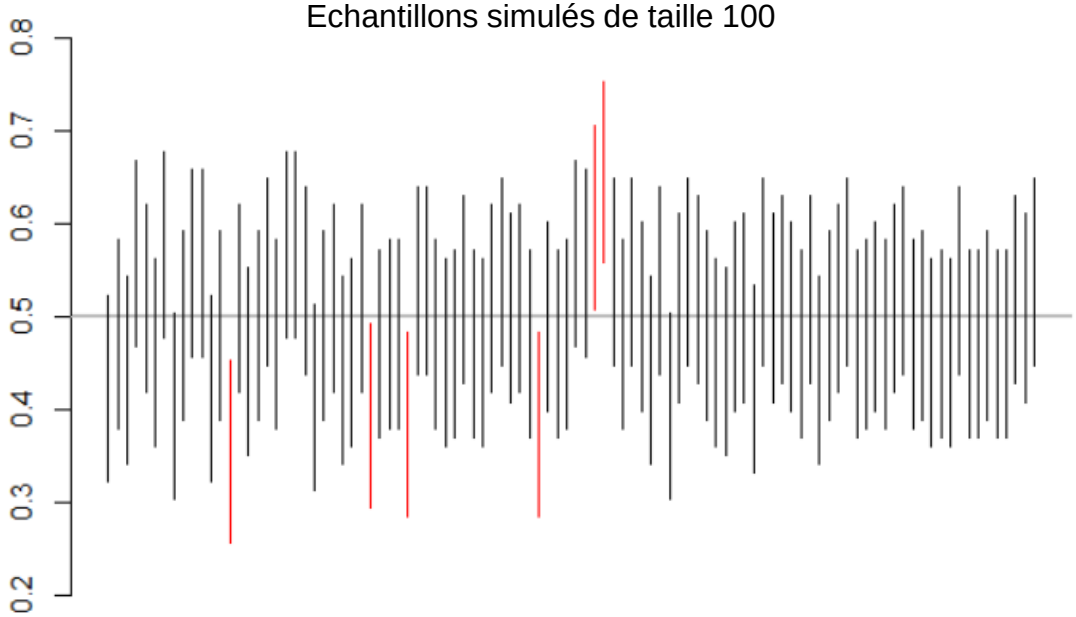
IC 95% et taille d'échantillon

Paramètre:
prévalence = 0.50



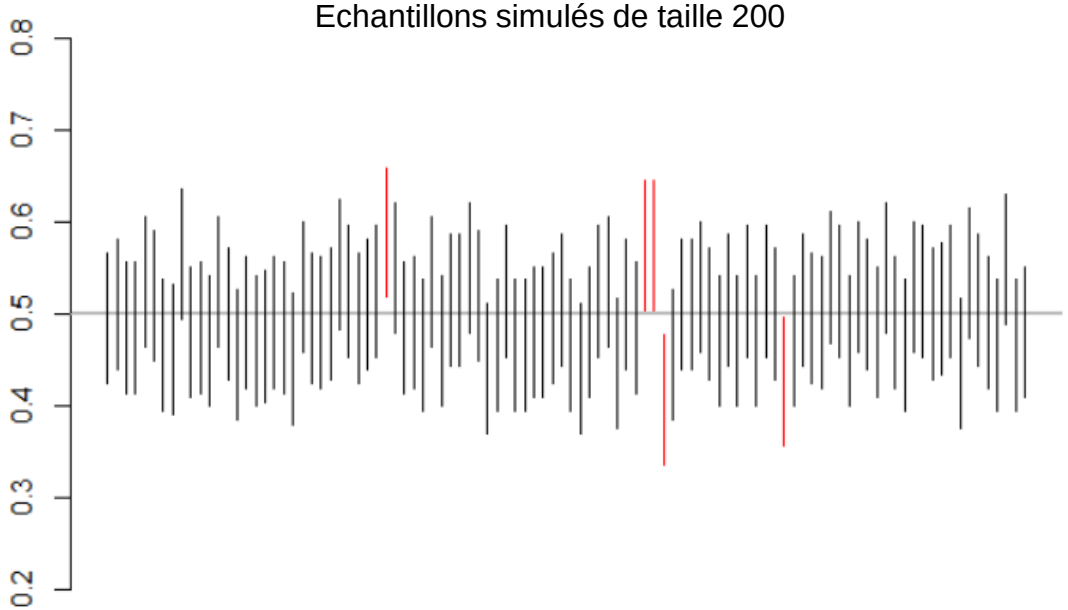
La largeur de l'IC95% tend à diminuer lorsque la taille d'échantillon augmente

Paramètre:
Prévalence = 0.50



- Si on obtient un intervalle de confiance à 95% de façon répétée à partir de la même population, 95% de ces ICs vont contenir le paramètre, et 5% ne le contiendront pas
- Quand on obtient un seul intervalle dans le cadre d'une étude, on ne sait pas dans quel cas de figure on se trouve. Si on n'a pas eu de chance, l'intervalle ne contient pas le paramètre.

Paramètre:
Prévalence = 0.50



- Si on obtient un intervalle de confiance à 95% de façon répétée à partir de la même population, 95% de ces ICs vont contenir le paramètre, et 5% ne le contiendront pas quelle que soit la taille d'échantillon
- Quand on obtient un seul intervalle dans le cadre d'une étude, on ne sait pas dans quel cas de figure on se trouve. Si on n'a pas eu de chance, l'intervalle ne contient pas le paramètre.

IC 95% d'une proportion
Résumé

- Définition:
  - Ensemble des valeurs du paramètre compatibles avec les données de l'échantillon
- Répond à la question:
  - Qu'est-ce que les données de l'échantillon permettent de conclure sur le paramètre?
- L'IC contient toujours la valeur estimée
- Plus l'IC est étroit, plus l'estimation est précise
- La largeur de l'IC diminue lorsque la taille d'échantillon augmente
- 95% des ICs contiennent la valeur du paramètre (5% ne la contiennent pas) quelle que soit la taille d'échantillon

IC 95% d'une proportion
Formule

- Formule permettant de calculer une approximation de l'IC95%
  - paramètre à estimer (dans la population): p
  - calculer la proportion dans l'échantillon: $\hat{p}$
  - calculer l'IC95%:

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Attention:
la largeur de l'intervalle de confiance est maximale si $\hat{p}$ =0.5, et diminue à mesure que $\hat{p}$ se rapproche de 0 ou 1

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Hyponatremia among Runners in the Boston Marathon

Sur les 488 coureur.ses analysées, 62 cas d'hyponatrémie sont identifiés

Dans l'échantillon, le risque d'hyponatrémie est 12.7% (62/488)

Que peut-on dire du risque d'hyponatrémie dans la population des coureur.ses de marathon ?

- Calculer la proportion dans l'échantillon: $\hat{p}$ =0.127
- Calculer l'IC95%: $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.127 \pm 2\sqrt{\frac{0.127(1-0.127)}{488}}$

$$= 0.127 \pm 2 \times 0.015$$

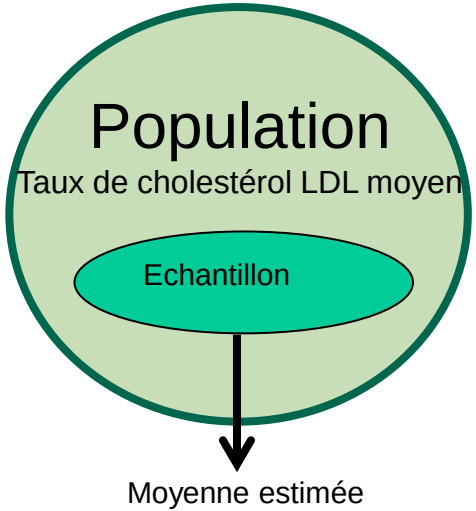
$$= 0.097 \text{ à } 0.157$$

Les données observées sont compatibles avec un risque d'hyponatrémie entre **9.7** et **15.7%** dans la population des coureur.ses

Autres exemples de paramètres à estimer

Paramètres
moyenne
écart type
coefficient de corrélation
coefficient de régression
taux d'incidence
odds ratio
différence de risque
risque relatif
…

Estimation de la moyenne d'une population



On veut connaître le taux moyen de cholestérol LDL dans une population de malades (paramètre)

Taux de cholestérol: variable quantitative continue

Recueil de données dans un échantillon aléatoire de N personnes

Observation de N valeurs de taux de cholestérol

Estimation du paramètre «moyenne dans la population»:

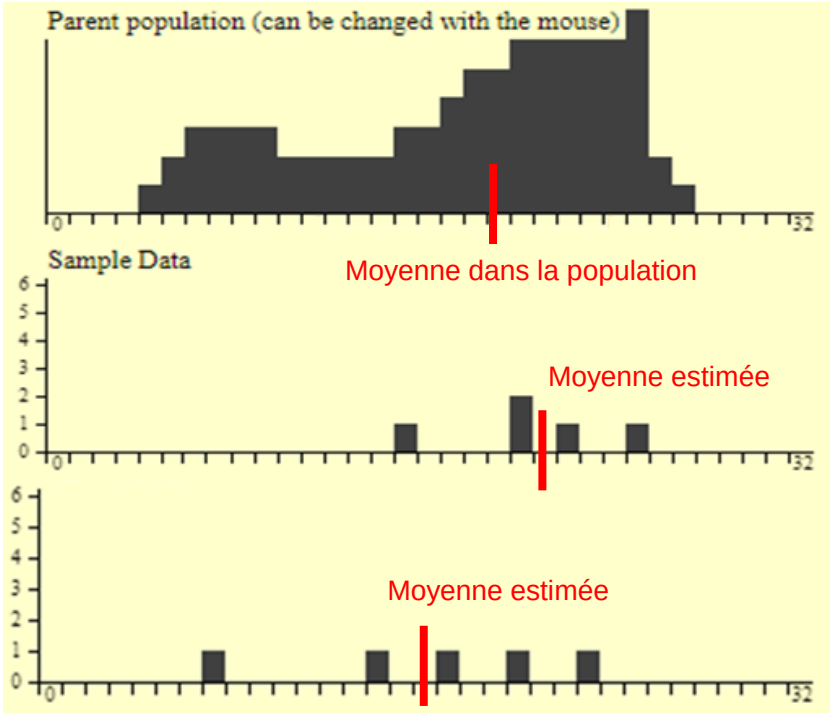
$$m = \frac{\sum_{i=1}^{n} x_i}{n}$$

x_i est la valeur de la variable ***X*** observée chez l'individu *i*

n est le nombre d'individus dans l'échantillon

Estimation et aléa de l'échantillonnage

Moyenne



Distribution des données dans la population

Distribution des observations dans un échantillon

Distribution des observations dans un autre échantillon

L'estimation du paramètre peut varier d'un échantillon à l'autre

Du paramètre « moyenne » à l'estimation
Simulation

- Objectif de la simulation:
  - observer la distribution de l'estimateur de la moyenne lorsqu'on échantillonne de manière aléatoire plusieurs fois la même population

- Etapes de la simulation:

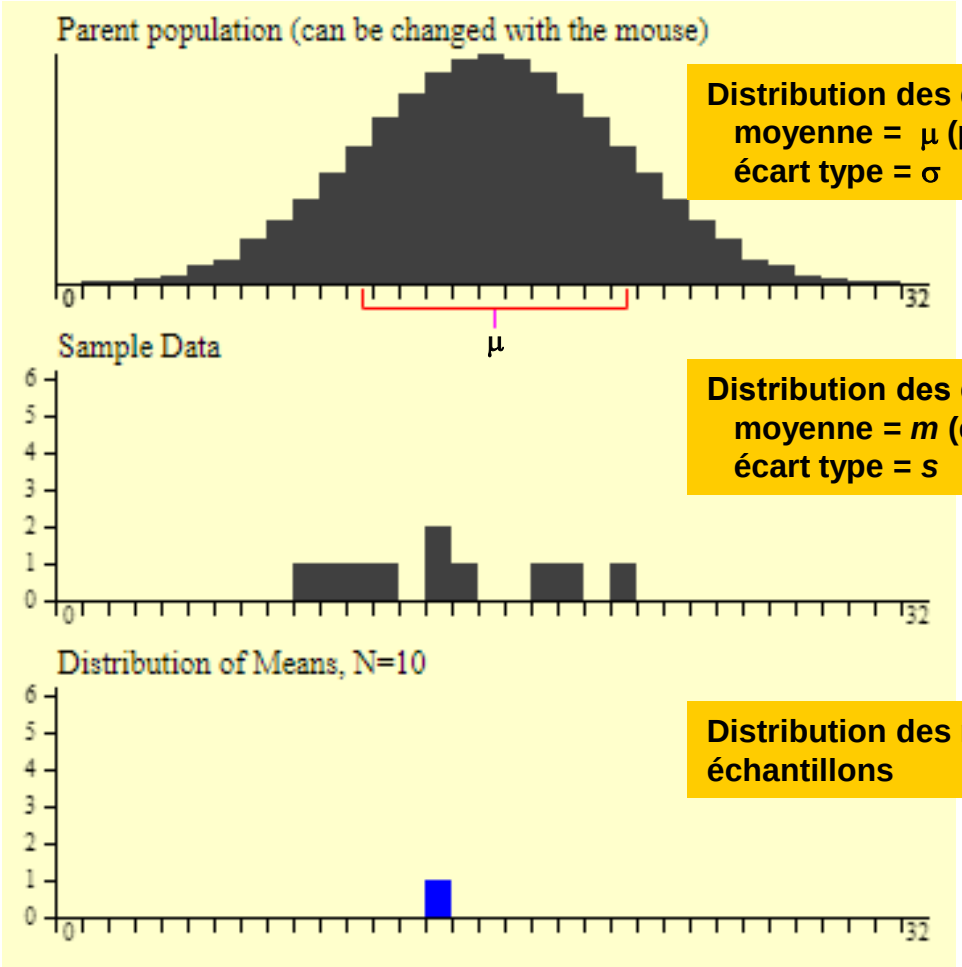
  1) distribution des données dans la population

  2) échantillonnage: sélection aléatoire de N individus dans cette population

  3) calcul de la moyenne estimée à partir des données de l'échantillon

  4) grand nombre d'échantillons => distribution des moyennes estimées

http://onlinestatbook.com/stat_sim/sampling_dist/index.html



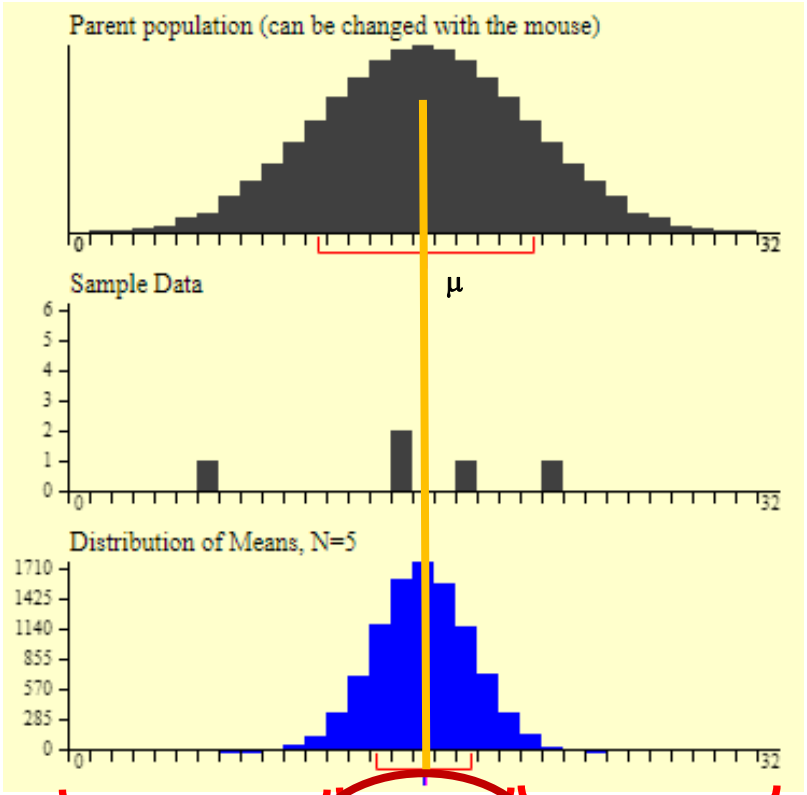
Distribution des données dans la population
moyenne = μ (paramètre à estimer)
écart type = σ

Distribution des données dans l'échantillon
moyenne = *m* (estimation du paramètre μ)
écart type = *s*

Distribution des moyennes des échantillons

Du paramètre « moyenne » à l'estimation

Enseignement de la simulation

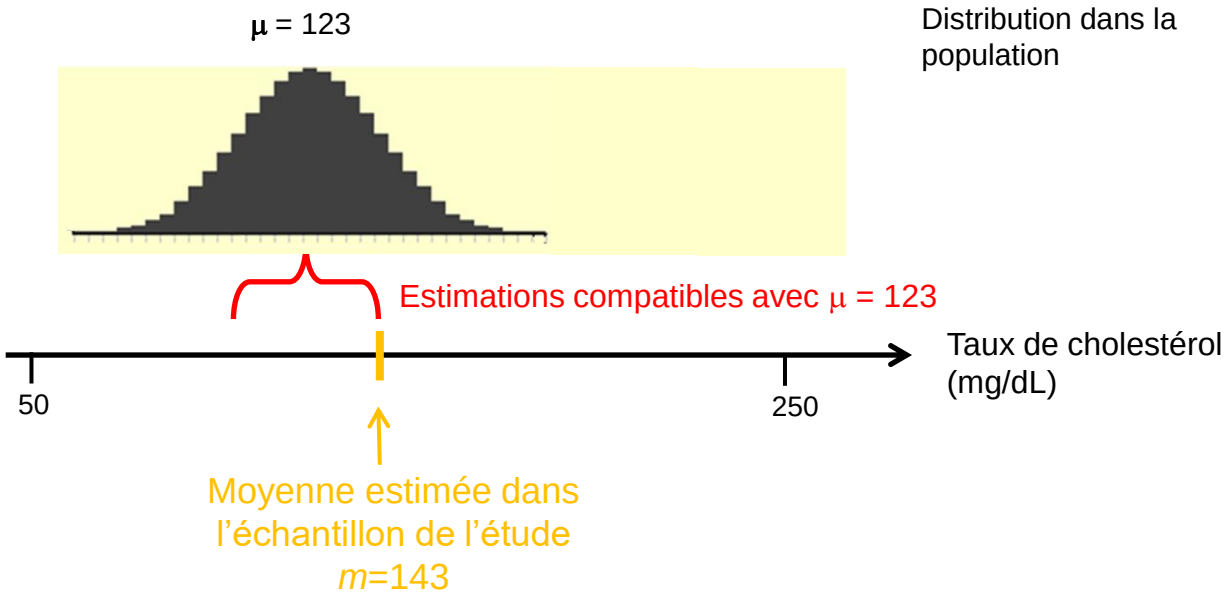


Distribution des moyennes des échantillons (échantillonnage répété) :

- forme en «cloche»
- centrée sur la moyenne μ dans la population

$\Rightarrow$ Intervalle des estimations de moyenne compatibles avec μ

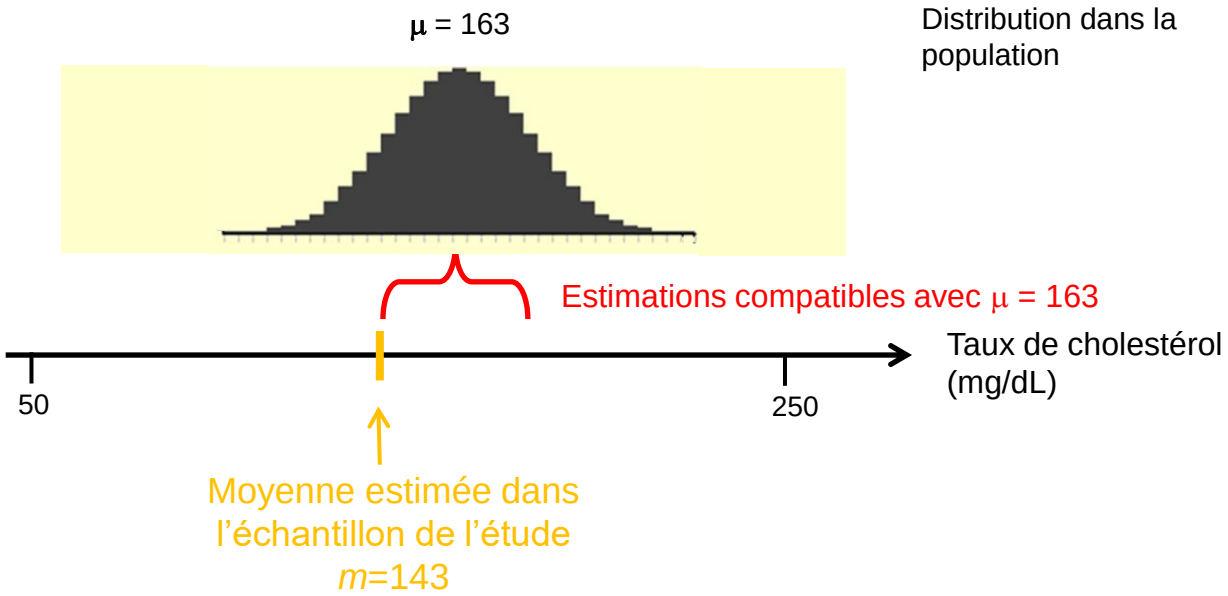
IC 95% d'une moyenne estimée



123 mg/dL est la plus petite valeur du paramètre μ compatible avec une moyenne estimée de 143 mg/dL avec les données de l'échantillon

=> La borne inférieure de l'IC 95% de l'estimation est 123 mg/dL

IC 95% d'une moyenne estimée



163 mg/dL est la plus grande valeur du paramètre μ compatible avec une moyenne estimée de 143 mg/dL avec les données de l'échantillon

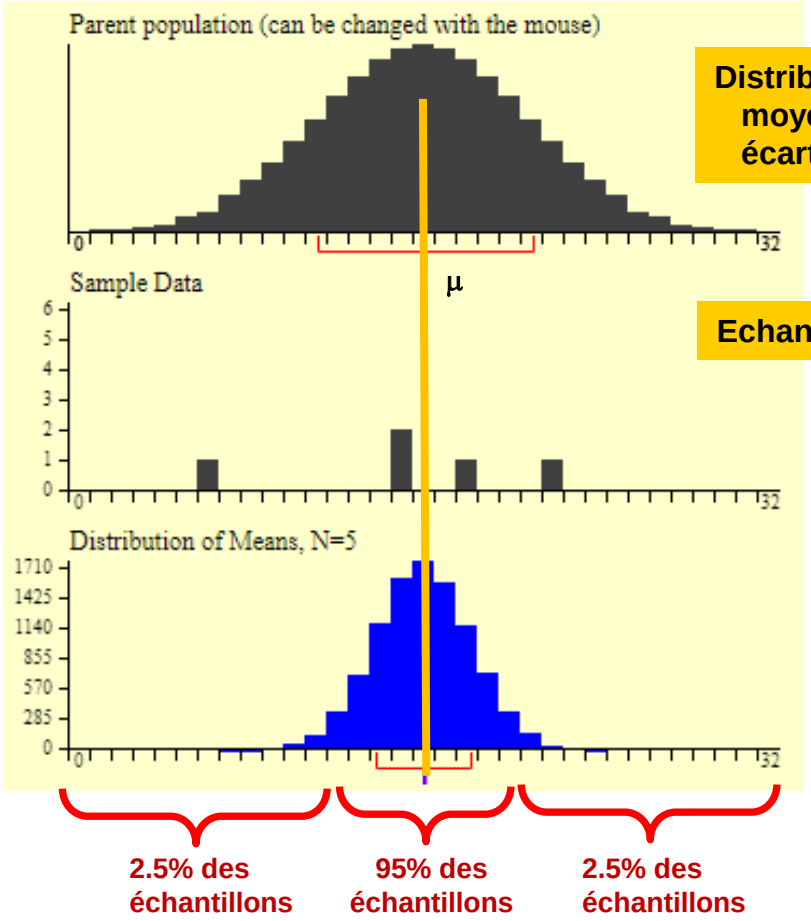
=> La borne supérieure de l'IC 95% de l'estimation est 163 mg/dL

Théorème de la limite centrale

- Soient
  - une variable dans la population de moyenne μ et de variance σ^2
  - un grand nombre d'échantillons de n observations indépendantes

- A mesure que la taille des échantillons n augmente (tend vers l'infini)
  - la distribution des moyennes des échantillons tend vers une distribution normale (i.e. gaussienne)
  - de moyenne μ
  - de variance σ^2/n

IC 95% d'une moyenne estimée



Distribution des données dans la population
moyenne = μ
écart type = σ

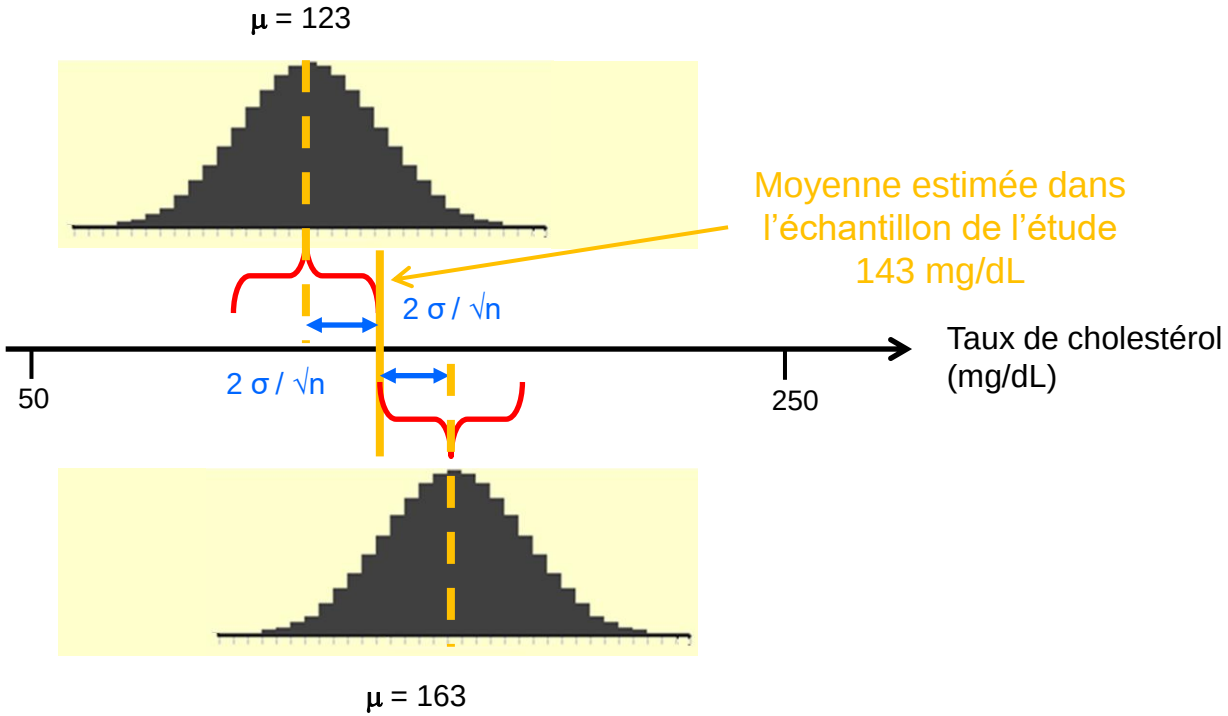
Echantillon de taille n

Distribution des moyennes des échantillons (échantillonnage répété) :

- distribution normale
- moyenne μ
- variance = σ^2/n
- écart type = $\sigma / \sqrt{n}$

⇒ Intervalle des estimations de moyenne compatibles avec μ
est: $\mu +/- 2\,\sigma / \sqrt{n}$

IC 95% d'une moyenne estimée



IC 95% d'une moyenne estimée
Formule

- IC95%: l'ensemble des valeurs du paramètre qui sont compatibles avec l'estimation obtenue avec les données de mon échantillon
  - Calculer la moyenne de l'échantillon: *m*
  - Calculer l'écart type des données dans l'échantillon: *s*
  - Calculer l'IC95%:

$$m \pm 2\,\frac{s}{\sqrt{n}}$$

L'intervalle de confiance à 95% est centré sur l'estimation
Sa largeur diminue lorsque :
la taille d'échantillon ***n*** augmente
et /ou l'écart type ***s*** diminue

Erreur type

- Erreur type d'une moyenne estimée = $s/\sqrt{n}$
- IC 95%: $m \pm 2\frac{s}{\sqrt{n}}$ ← erreur type

La largeur de IC95% diminue lorsque :

la taille d'échantillon ***n*** augmente
et /ou l'écart type ***s*** diminue

} erreur type diminue

- L'erreur type est une mesure de la précision de l'estimation

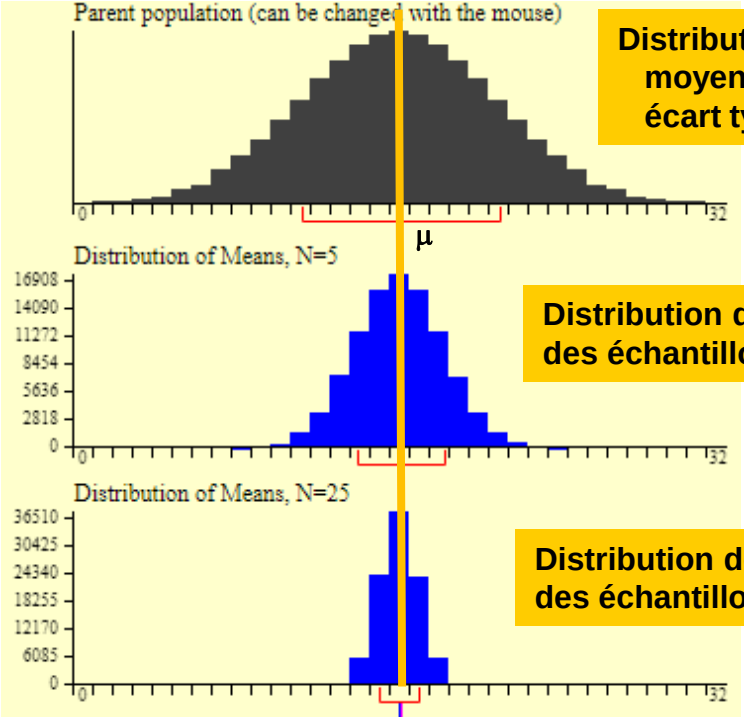
Exemple de l'enquête

- Taille chez les étudiantes (N=268):
  - Moyenne: 166.4 cm
  - Ecart type: 6.7 cm
  - Erreur type de la moyenne: 6.7/ √268 = 0.41
  - $IC_{95\%}$: 165.6 à 167.2 cm ↔ **1.6 cm** $IC_{95\%}$: 166.4 +/- 2*6.7/√268

La largeur de l'IC95% est plus petite chez les étudiantes que chez les étudiants car il y a plus d'étudiantes que d'étudiants

- Taille chez les étudiants (N=103):
  - Moyenne: 178.4 cm
  - Ecart type: 6.6 cm
  - Erreur type de la moyenne: 6.6/ √103 = 0.65
  - $IC_{95\%}$: 177.1 à 179.7 cm ↔ **2.6 cm** $IC_{95\%}$: 178.4 +/- 2*6.6/√103

IC 95% et taille d'échantillon



Distribution des données dans la population
moyenne = μ (paramètre à estimer)
écart type = σ

Distribution des moyennes des échantillons (n=5)

distribution normale
moyenne μ
variance = $\sigma^2 / 5$
écart type = $\sigma / \sqrt{5}$

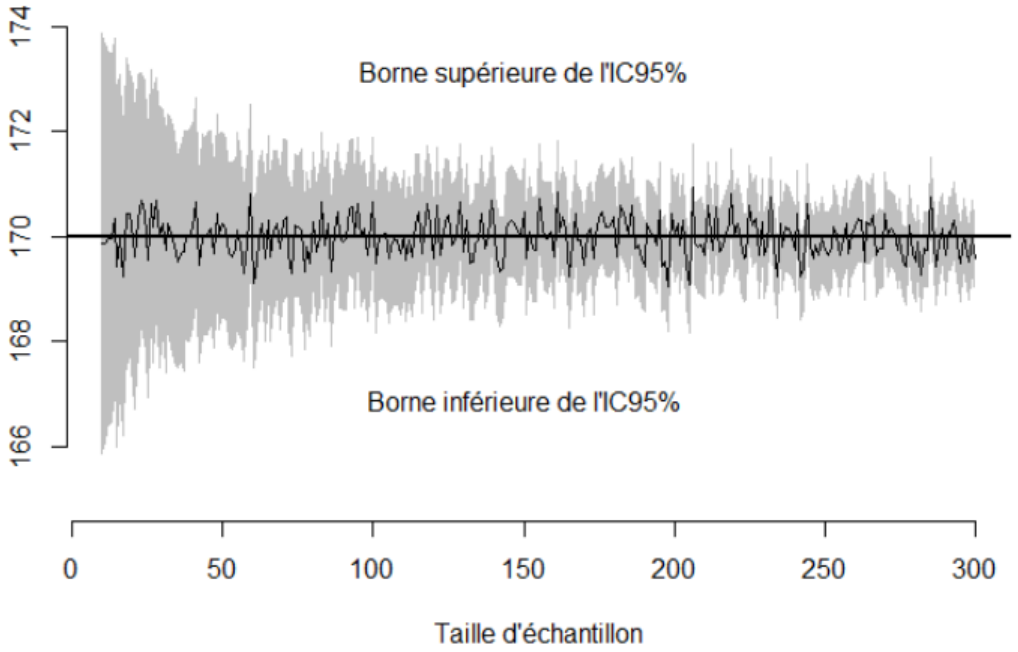
Distribution des moyennes des échantillons (n=25)

distribution normale
moyenne μ
variance = $\sigma^2 / 25$
écart type = $\sigma / \sqrt{25}$

⇒ quand n augmente, l'intervalle des estimations de moyenne compatibles avec μ rétrécit

IC 95% et taille d'échantillon

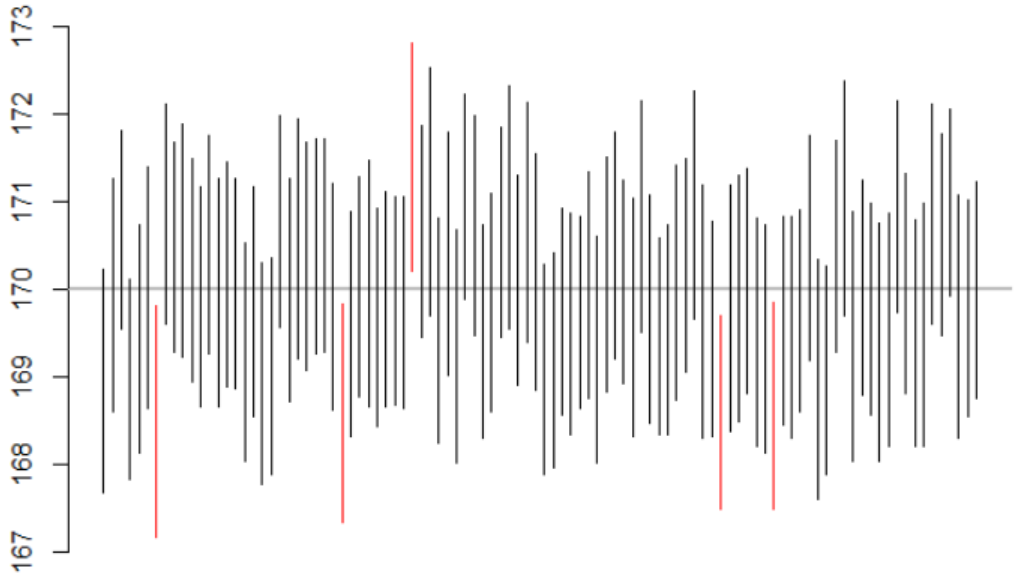
Population:
Taille moyenne = 170 cm (paramètre)
Ecart type = 6.5 cm



La largeur de l'IC95% tend à diminuer lorsque la taille d'échantillon augmente

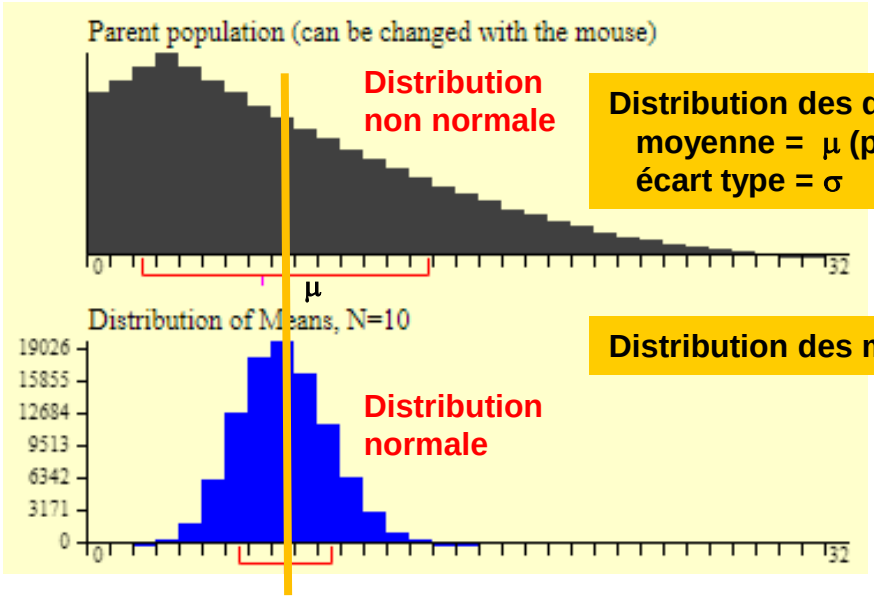
Echantillons simulés de 100 personnes

<u>Population</u>:
Taille moyenne = 170 cm (paramètre)
Ecart type = 6.5 cm



- Si on obtient un intervalle de confiance à 95% <u>de façon répétée</u> à partir de la même population, 95% de ces ICs vont contenir le paramètre, et 5% ne le contiendront pas quelle que soit la taille d'échantillon
- Quand on obtient <u>un seul intervalle</u> dans le cadre d'<u>une</u> étude, <u>on ne sait pas</u> dans quel cas de figure on se trouve. Si on n'a pas eu de chance, l'intervalle ne contient pas le paramètre.

IC 95% et distribution dans la population



Distribution des données dans la population
moyenne = μ (paramètre à estimer)
écart type = σ

Distribution des moyennes des échantillons (n=10)

Quelle que soit la distribution dans la population:

- la distribution des moyennes des échantillons est normale (à condition que la taille d'échantillon *n* est «suffisamment» grande)
- la formule de l'IC95% est applicable

IC 95% d'une moyenne
Résumé

- Définition:
  - Ensemble des valeurs du paramètre compatibles avec les données de l'échantillon
- Répond à la question:
  - Qu'est-ce que les données de l'échantillon permettent de conclure sur le paramètre?
- L'IC est centrée sur la moyenne estimée
- Plus l'IC est étroit, plus l'estimation est précise
- La largeur de l'IC diminue lorsque :
  - la taille d'échantillon augmente
  - l'écart dans la population diminue
- 95% des ICs contiennent la valeur du paramètre (5% ne la contiennent pas) quelle que soit la taille d'échantillon

IC 95%: généralités

- **Intervalle de confiance à 95%**
  - dépend des données observées
  - peut varier si on re-échantillonne la population
  - contient toujours l'estimation (en général IC centré sur l'estimation)
  - essentiel à l'interprétation des résultats d'une étude
  - permet de mesurer la précision de l'estimation
  - devrait toujours être rapporté avec l'estimation

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$m \pm 2\,\frac{s}{\sqrt{n}}$$

- Quel que soit le paramètre estimé, interprétation identique:
  - ensemble des valeurs du paramètre compatibles avec les données observées

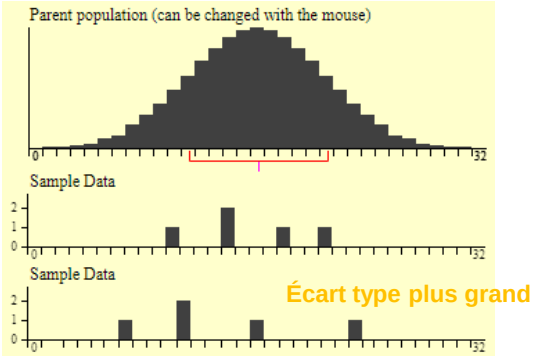
Echantillon

- Une inférence statistique juste nécessite un bon échantillon
- Un bon échantillon doit être
  - **représentatif** de la population cible
  - de **taille suffisante** (pour être assez précis)
- Toujours identifier la population cible!
- Echantillon représentatif:
  - **aléatoire** (idéal – participants sélectionnés par une méthode de tirage au sort)
  - **complet** (maladies rares)

Estimation: deux types d'erreur

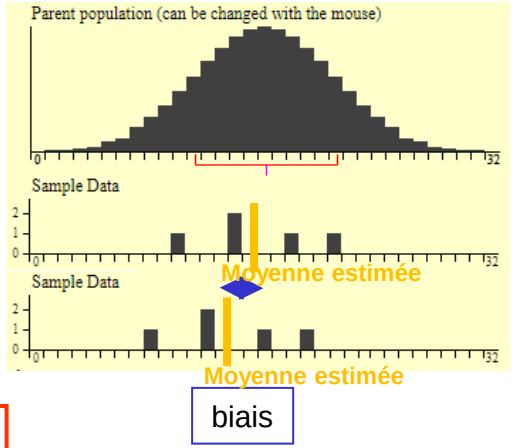
- Erreurs aléatoires:
  - cause potentielle: méthode de mesure imprécise
  - conséquence: IC95% tend à être plus large

précision



- Erreurs systématiques:
  - causes potentielles:
    - biais de sélection des participant.es
    - méthode de mesure mal calibrée
  - conséquence: estimation biaisée

validité



Attention: IC95% ne dit rien d'une éventuelle erreur systématique (biais)

Biais de sélection

- Erreur systématique dans la sélection des participants qui conduit à une estimation incorrecte du phénomène mesuré (ex, proportion, moyenne, etc).

- Echantillon **non-représentatif**:
  - volontaires
  - patients choisis par le médecin
  - nombreux patients perdus de vue
  - …

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Hyponatremia among Runners in the Boston Marathon

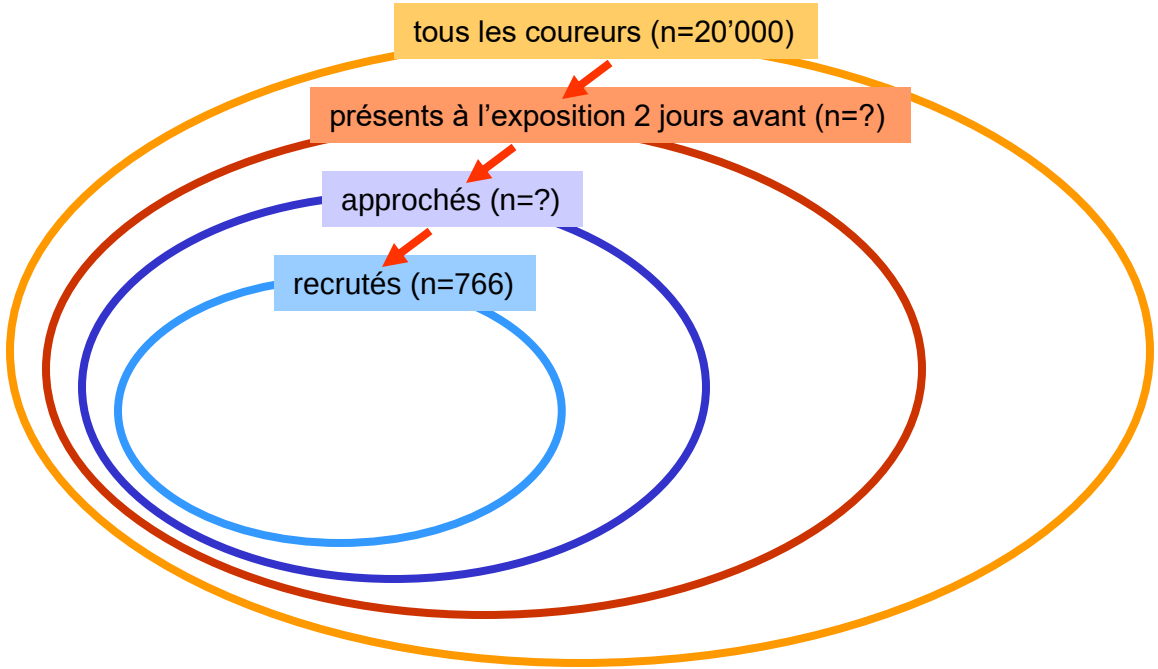
Excessive fluid intake is believed to be the primary risk factor for hyponatremia, on the basis of observations of marathon runners who have collapsed[2-5,7,11,12] and studies of elite athletes.[13-17]

We undertook the present study to estimate prospectively the incidence of hyponatremia among marathon runners and to identify the principal risk factors involved.

N Engl J Med 2005;352:1550-6

STUDY POPULATION

Marathon runners were recruited prospectively at an exposition one or two days before the Boston Marathon, in April 2002. All registered participants 18 years of age or older were eligible, regardless of whether they registered for the marathon on the basis of a competitive qualifying time or on behalf of a charitable organization — a mechanism for which no previous marathon experience was required. Subjects were approached at random in an area adjacent to race registration and invited to participate.



Processus de recrutement non-aléatoire

Table 1 summarizes the baseline demographic and training characteristics of the study population. Of 766 runners enrolled, 511 (67 percent) reported to the finish-line research station. Of these, 489 provided a blood sample (constraints such as plane flights precluded 22 runners from providing a sample). One sample was considered of insufficient quantity, leaving a total of 488 subjects for analysis.

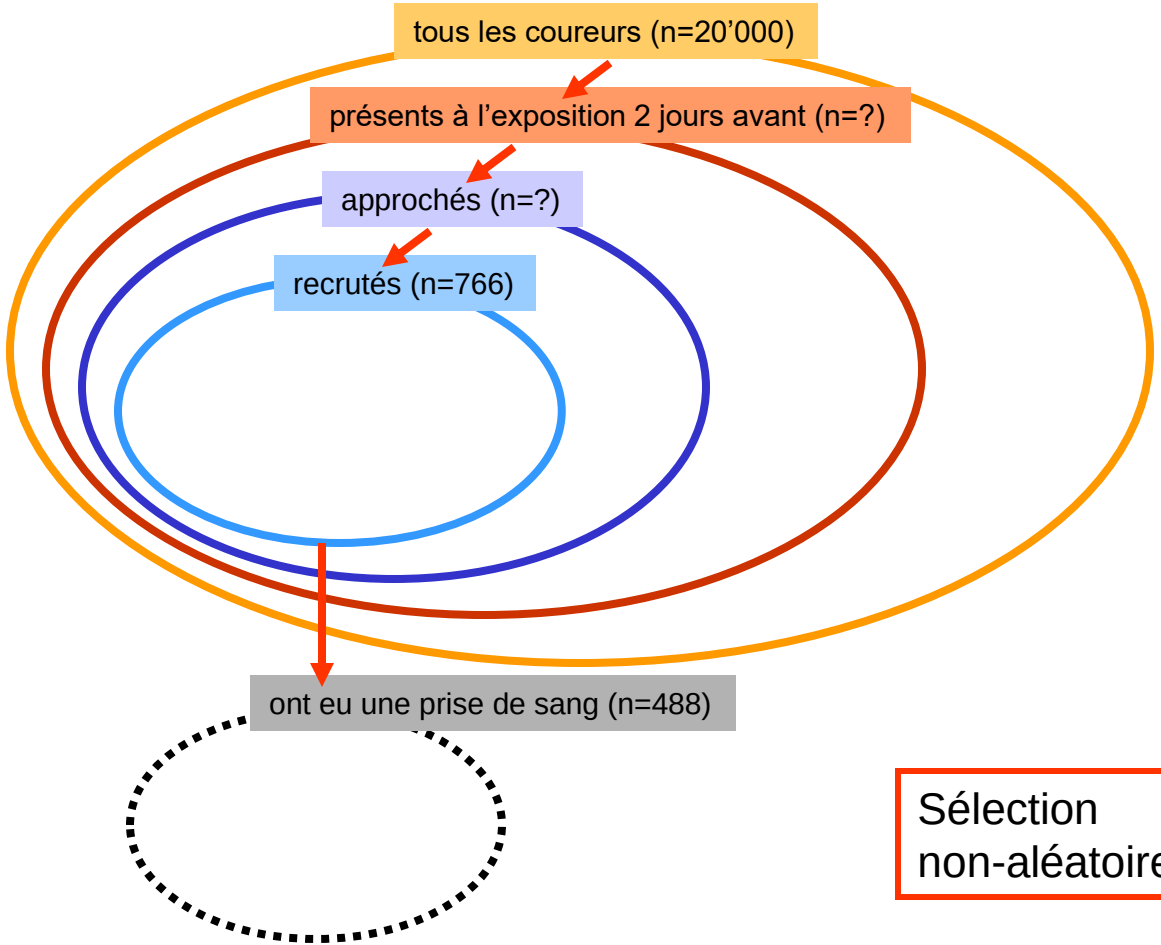
Analyse de 63.7% (488/766)
des personnes sélectionnées

N Engl J Med 2005;352:1550-6

Table 1. Baseline Characteristics of the 2002 Boston Marathon Study Population.*

Characteristic	Male Runners (N=473)	
	Reporting at Finish Line (N=336)	Not Reporting at Finish Line (N=137)
Age — yr	40.4±9.6	40.4±10.0
Nonwhite race — %	9	10
Prerace weight — kg	74.6±9.5	76.6±10.7
Body-mass index†	23.7±2.6	24.5±2.7
Training pace — min:sec/mi	7:53±1:02	8:04±1:09
Previous marathons — median no. (interquartile range)	5 (2–12)	4 (1–12)
Self-reported water loading — %‡	75	79
Self-reported use of NSAIDs — %§	51	54
Race duration — hr:min¶	3:37±0:42	3:46±0:40

N Engl J Med 2005;352:1550-6



Sélection non-aléatoire

At the finish line, the runners had a mean serum sodium concentration of 140±5 mmol per liter (range, 114 to 158). Thirteen percent (62 of 488) had hyponatremia, including 22 percent of women (37 of 166) and 8 percent of men (25 of 322). Three runners (0.6 percent) had critical hyponatremia (serum sodium concentrations, 119, 118, and 114 mmol per liter).

Risque de biais?

N Engl J Med 2005;352-1550-6

Problèmes potentiels

- Groupe sélectionné au départ est peut-être atypique
  - contactés lors d'une exposition
  - approchés « au petit bonheur »
  - refus de participer non décrits

- Un tiers des participants perdus de vue
  - ceux qui ont développé des symptômes d'hyponatrémie avaient-ils moins de chances de terminer la course?

- Suspicion de biais de sélection => risque de sous-estimation du risque d'hyponatrémie

Notions clefs

- Estimation d'un paramètre, inférence
- Echantillonnage, biais de sélection
- Erreurs aléatoires et systématiques
- L'intervalle de confiance à 95% d'une estimation décrit les valeurs du paramètre compatibles avec cette estimation: capte l'incertitude/imprécision de l'estimation mais pas un éventuel biais
- Savoir :
  - calculer l'IC95% d'une proportion et d'une moyenne estimée avec les formules de ce cours
  - interpréter un IC95%

Objectifs prochaine séance

- **Comprendre les notions suivantes**
  - Tests statistiques
  - Hypothèses nulles et alternes
  - Erreurs de type 1 et de type 2
  - Valeur p
  - Puissance

Chapitres Petrie/Sabin
17 – 18 – 36