Démarche Statistique 1

Échantillonnage

Pierre Neuvial, http://stat.genopole.cnrs.fr/~pneuvial
Evry, M1 SGO, automne 2014

Introduction

Objectif

- statistique descriptive: sur l'échantillon
- statistique inférentielle: de l'échantillon à la population

→ comment choisir l'échantillon à partir de la population ?

Lois de probabilité

Moyenne et écart-type (population finie)

- X: variable d'intérêt
- population de taille N
- x_α: valeur de X mesurée sur l'unité d'indice α
- échantillon de taille n

Moyenne de la population

$$\mu = \frac{1}{N} \sum_{\alpha=1}^{N} x_\alpha$$

Moyenne de l'échantillon

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_{\alpha_i}$$

Ecart-type σ de la population

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

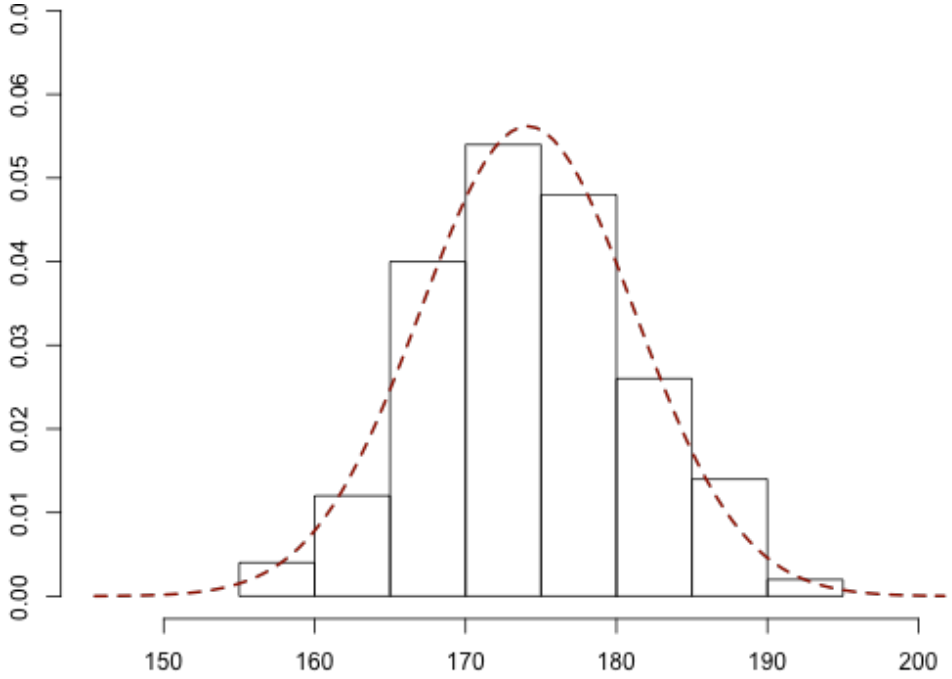
$$\sigma^{*2} = \frac{1}{N-1} \sum_{\alpha=1}^{N} (x_\alpha - \mu)^2$$

Ecart-type S de l'échantillon

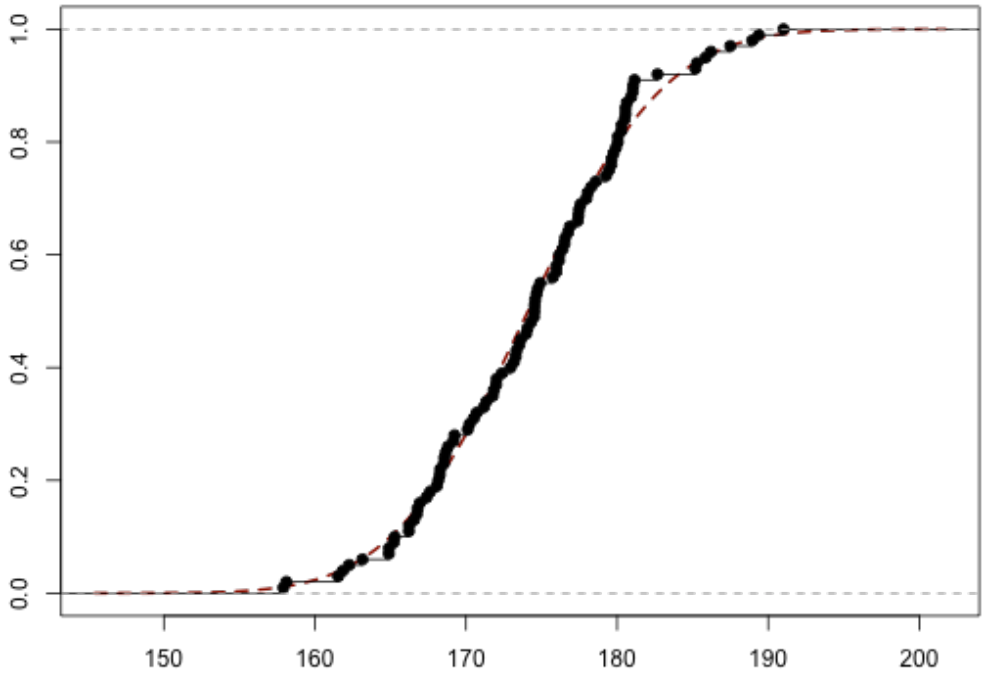
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{\alpha_i} - \bar{X})^2$$

Distributions empiriques: $n = 100$

Histogramme

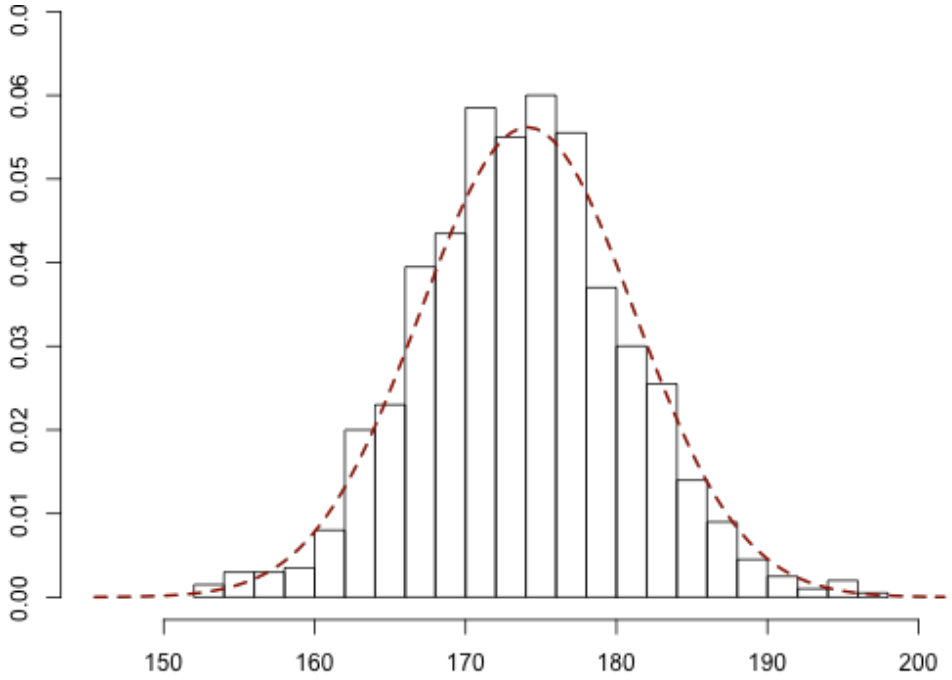


Fonction de répartition

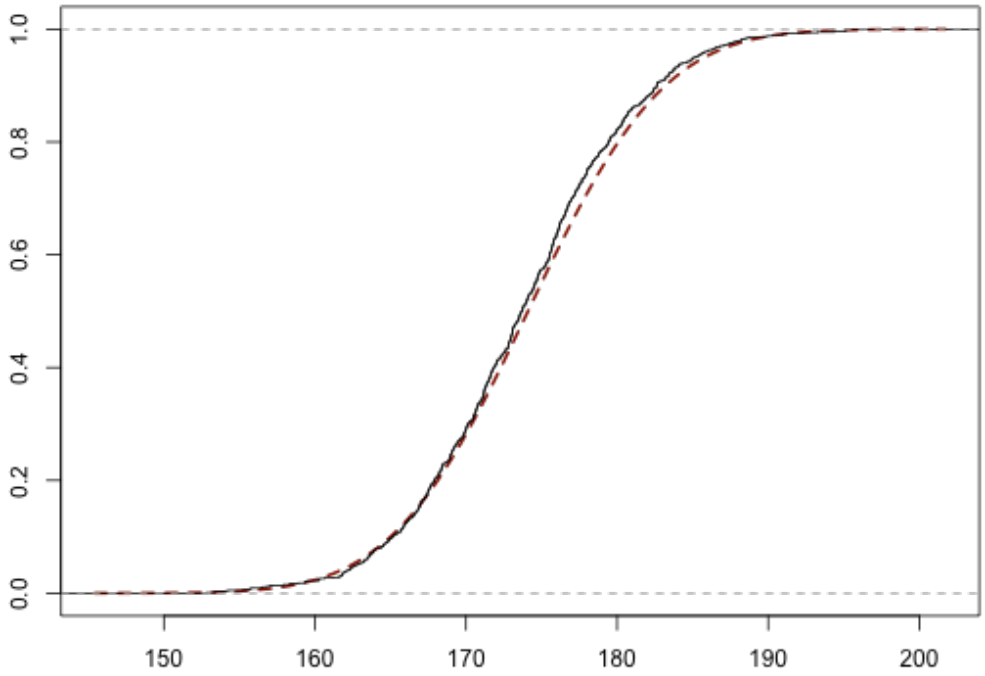


Distributions empiriques: $n = 1000$

Histogramme

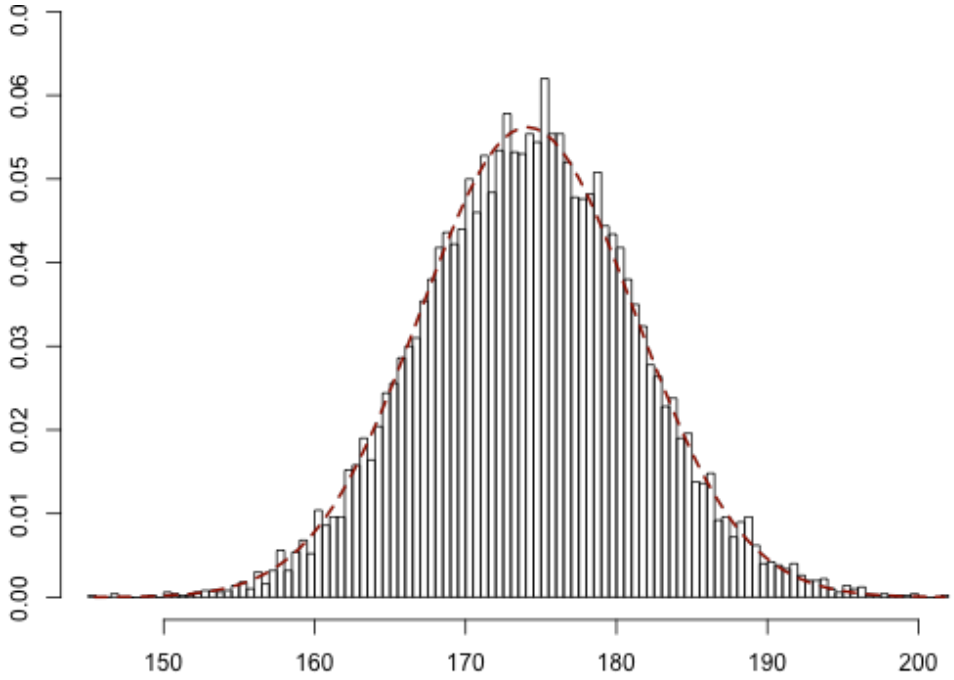


Fonction de répartition

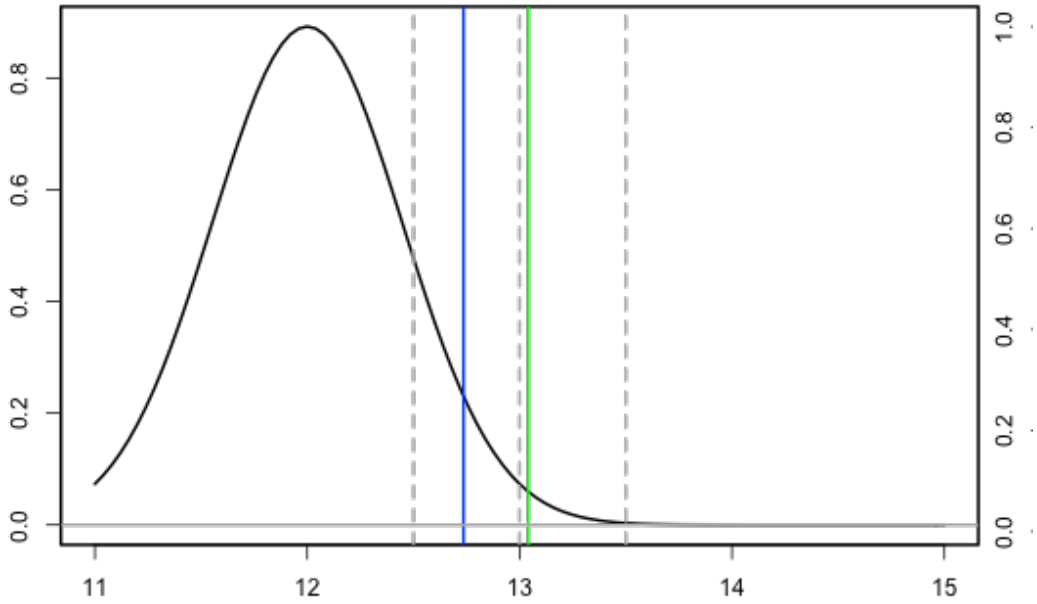


Distributions empiriques: $n = 10000$

Histogramme

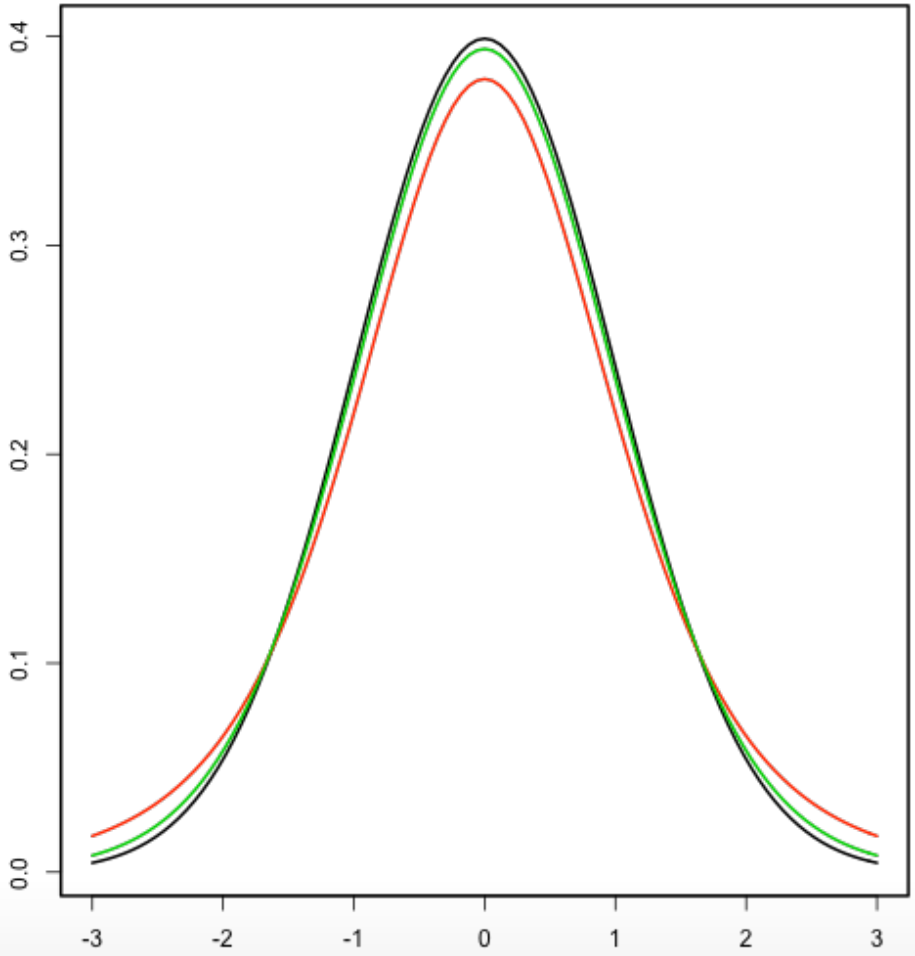


Fonction de répartition

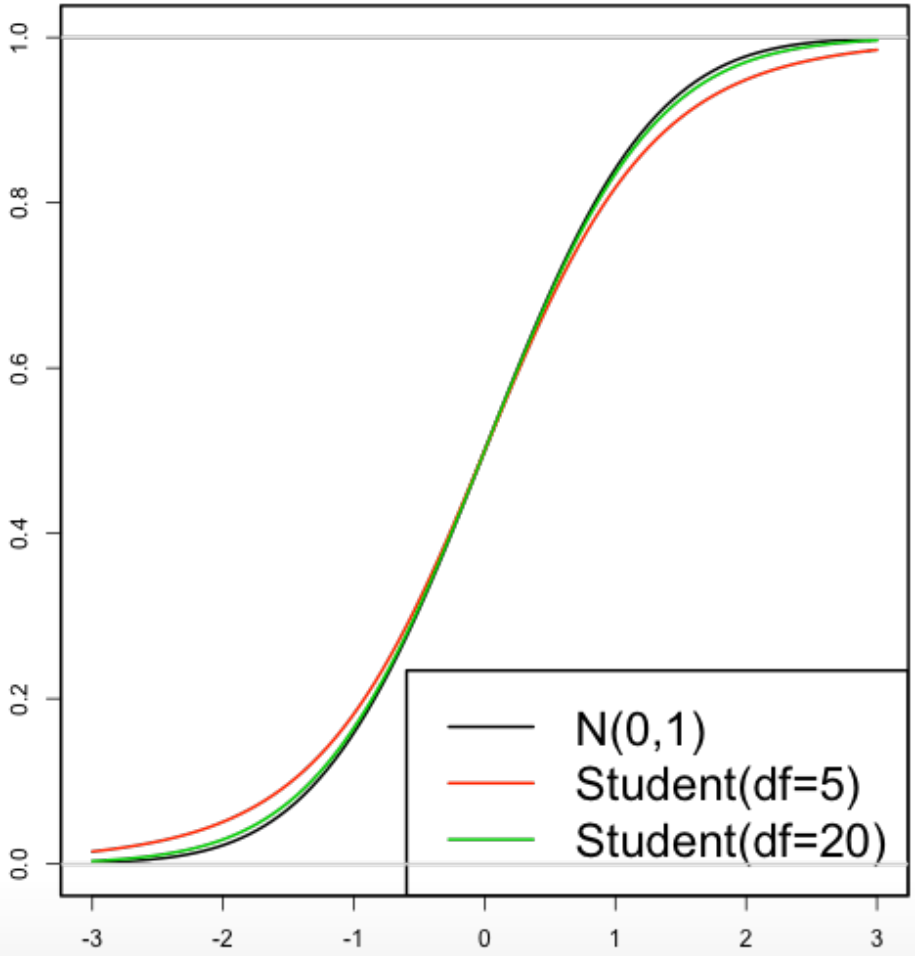


Distribution de la population ("loi théorique")

Densité ("dnorm")



Fonction de répartition (pnorm)

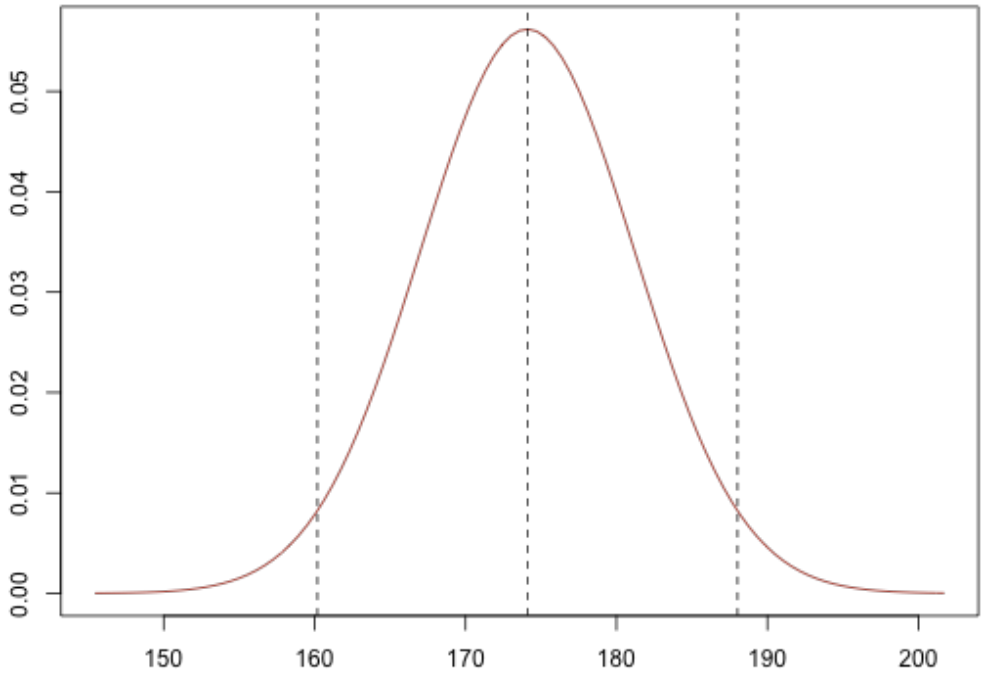


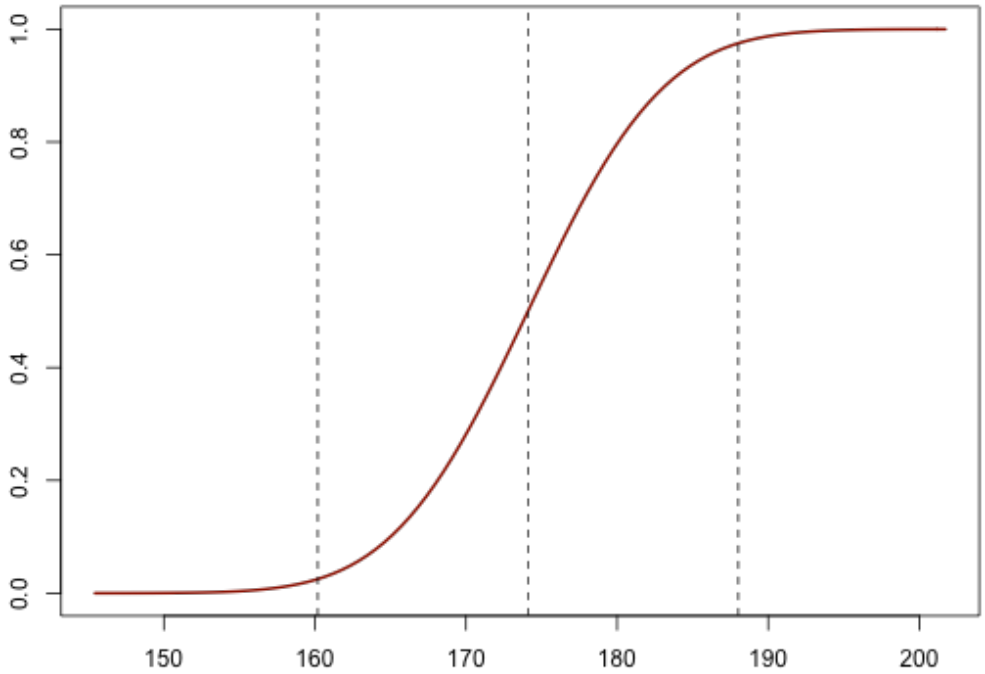
Dans notre exemple c'est la loi normale

Autres distributions

Situer un individu dans une distribution

Exemple: loi normale





Paramètres

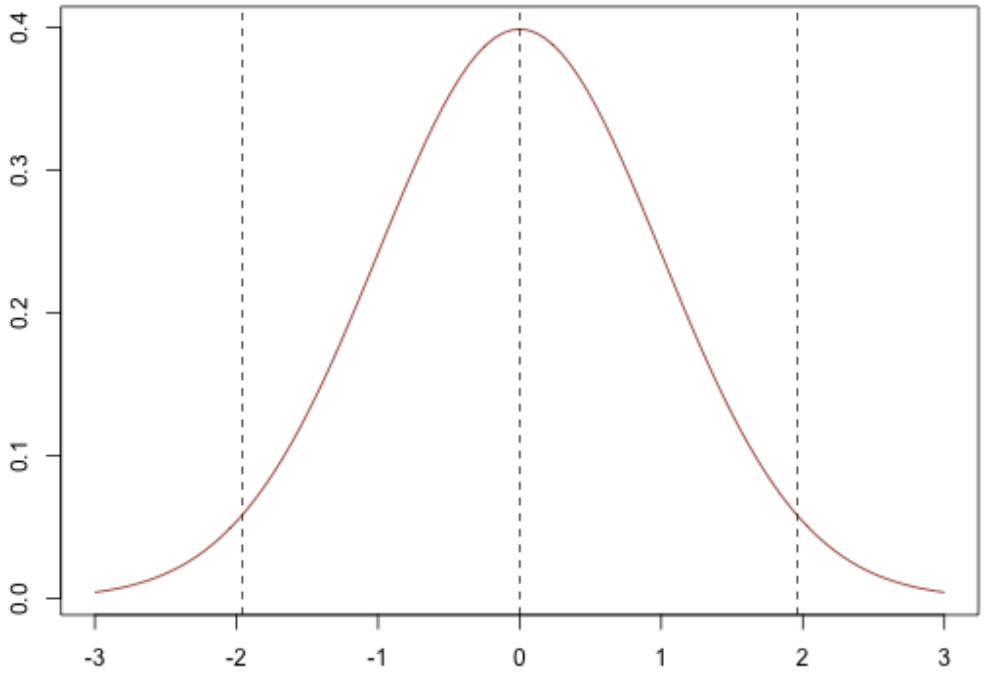
- moyenne: $\mu = 174.1$
- écart-type: $\sigma = 7.1$

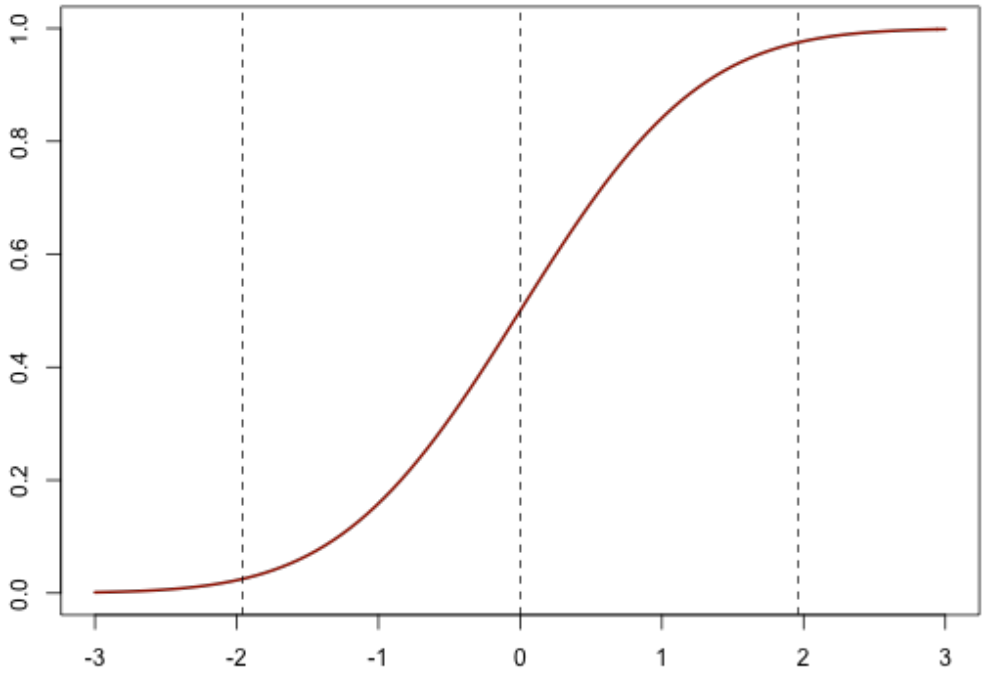
- Proportion d'hommes de plus de 1m90 ?
- Proportion d'hommes entre 1m75 et 1m80 ?

Score Z

$$Z = \frac{X - \mu}{\sigma}$$

Si X suit une loi $\mathcal{N}(\mu, \sigma^2)$, alors Z suit une loi $\mathcal{N}(0, 1)$





Échantillonnage

Introduction

Motivation

La mesure d'une caractéristique chez tous les individus de la population est impossible

On récolte donc seulement les données sur un échantillon

- Par nature un échantillon n'apporte qu'une **information partielle** sur la population
- **erreur d'échantillonnage**: écart entre le résultat obtenu sur l'échantillon et sur la population

Objectifs de la théorie de l'échantillonnage

- **déterminer** l'erreur d'échantillonnage
- la **minimiser** à coût fixé

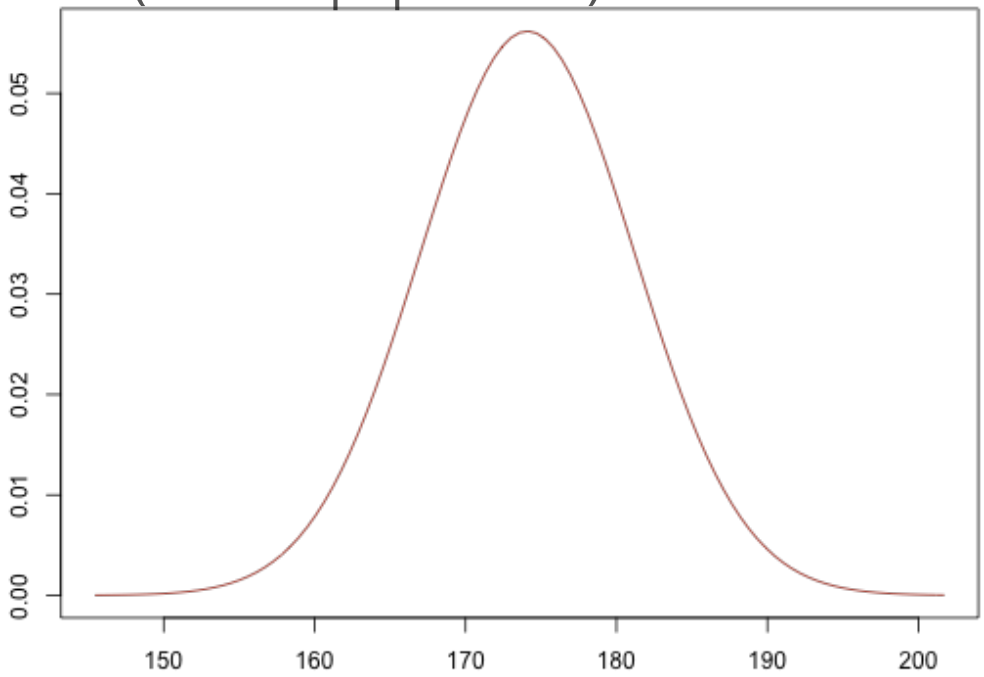
Erreur d'échantillonnage ≠ erreur de mesure !

Exemples de distribution d'échantillonnage

Distribution d'échantillonnage de la moyenne

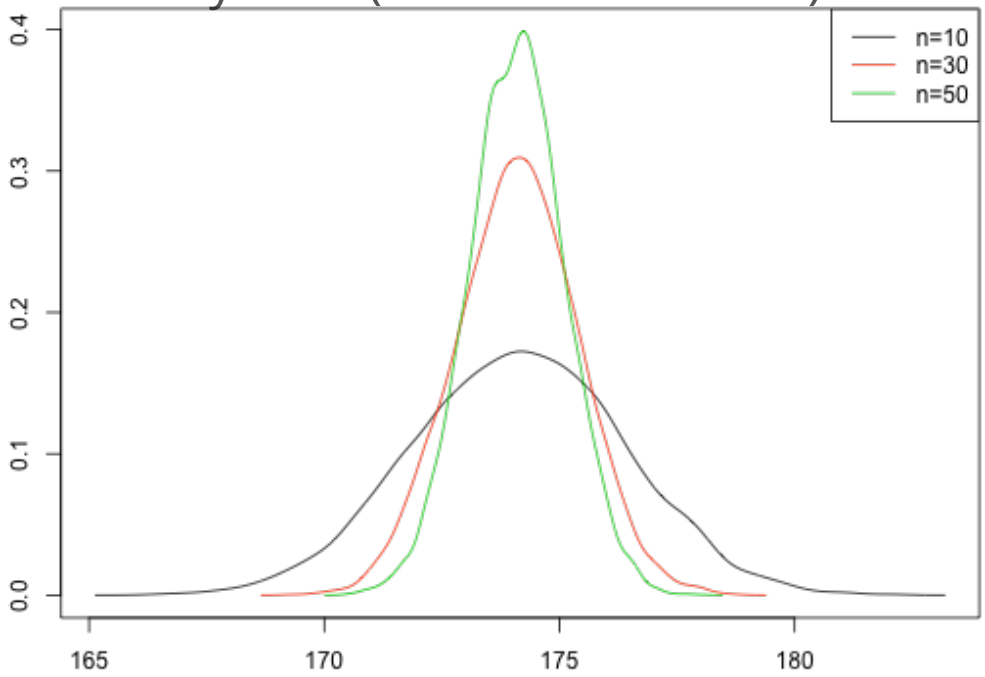
Exemple: taille des hommes français

Taille (dans la population)



un individu=un homme français

Taille moyenne (d'un n-échantillon)

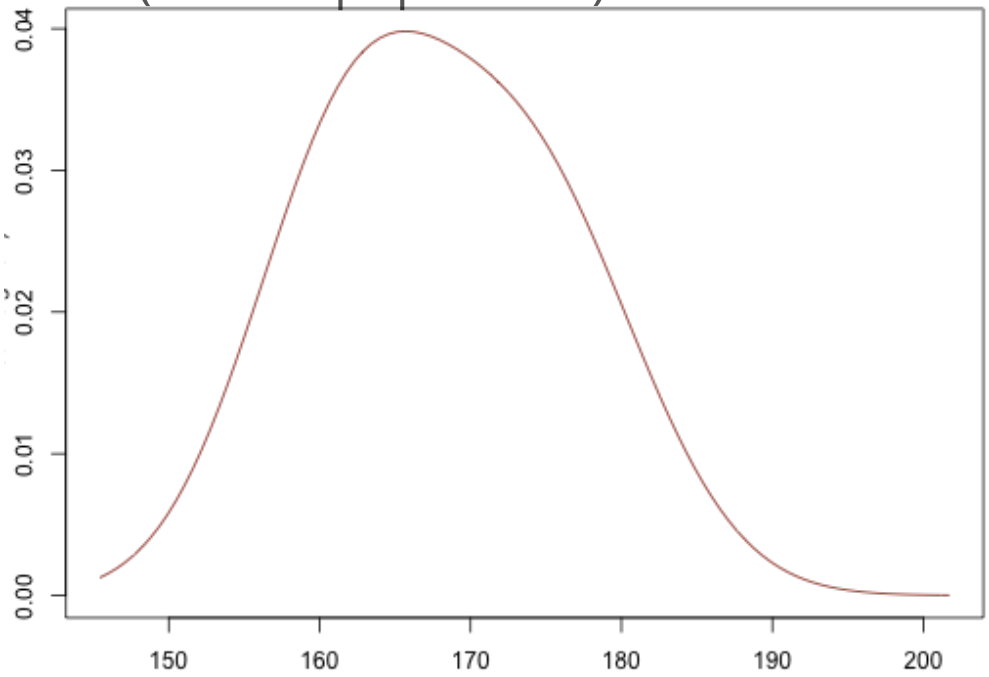


un individu=un échantillon de taille n

Distribution d'échantillonnage de la moyenne

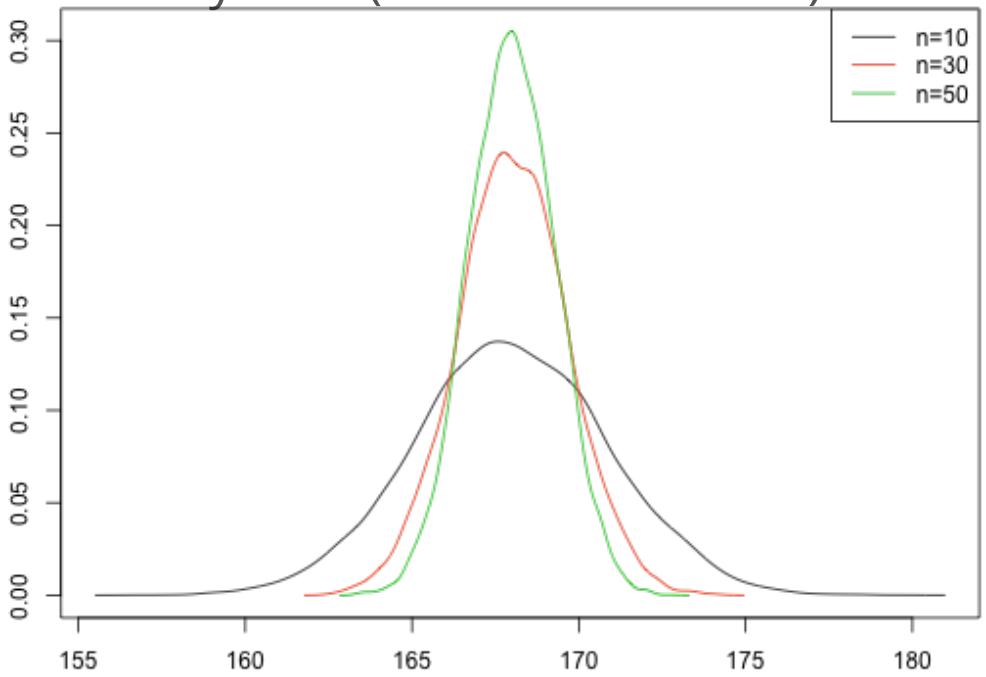
Exemple: taille des français (hommes+femmes)

Taille (dans la population)



un individu=un français

Taille moyenne (d'un n-échantillon)

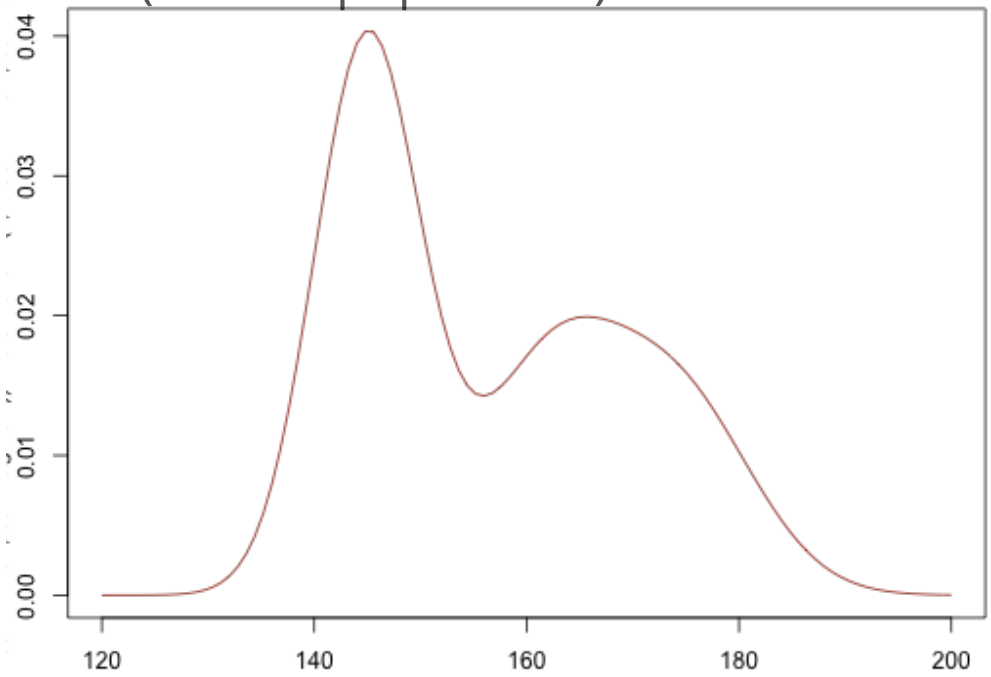


un individu=un échantillon de n français

Distribution d'échantillonnage de la moyenne

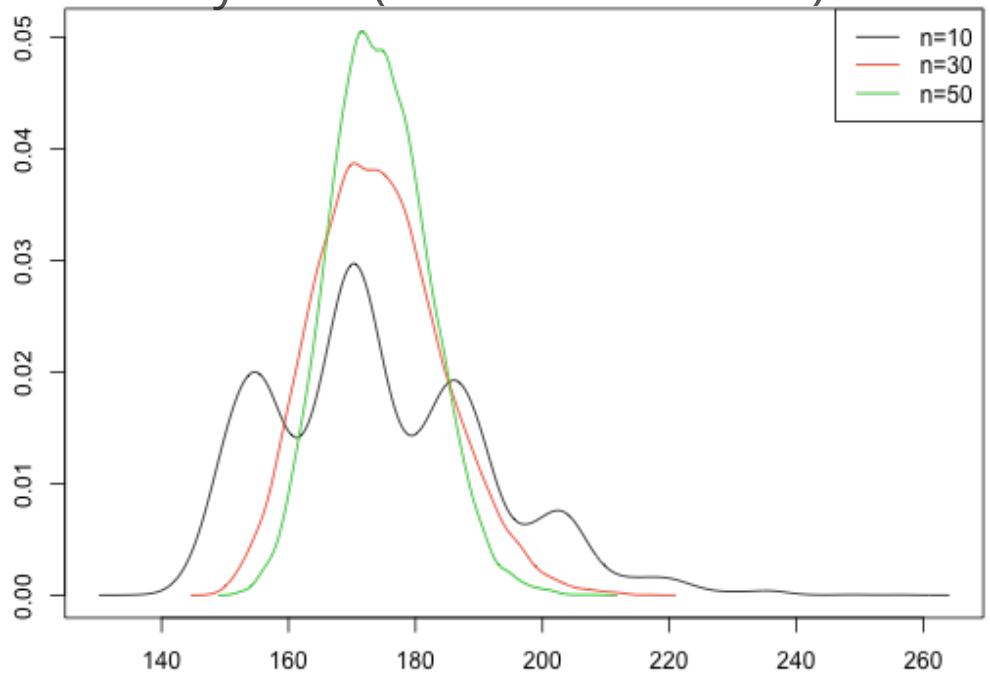
Exemple: taille des Evryens et des Pygmées

Taille (dans la population)



un individu=un Evryen ou un Pygmée

Taille moyenne (d'un n-échantillon)



un individu=un échantillon de n personnes, chacune soit Evryenne soit Pygmée

Echantillonnage aléatoire simple

Notations

On suppose l'absence d'erreur de mesure

- X: variable d'intérêt
- population de taille N
- x_α: valeur de X mesurée sur l'unité d'indice α
- échantillon de taille n
- $f = n/N$: taux de sondage

Paramètres (population)

- $\mu = \frac{1}{N} \sum_{\alpha=1}^{N} x_\alpha$
- $\sigma^{*2} = \frac{1}{N-1} \sum_{\alpha=1}^{N} (x_\alpha - \mu)^2$
- Erreur d'échantillonnage: $\bar{X} - \mu$

Estimateurs (échantillon)

- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_{\alpha_i}$
- $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{\alpha_i} - \bar{X})^2$

Echantillon aléatoire simple (EAS)

Objectif

- On s'intéresse au paramètre μ
- On utilise l'échantillon pour en donner une valeur approchée

$\Rightarrow$ comment construire l'échantillon ?

Définition

Un échantillon aléatoire simple est un échantillon obtenu par une méthode qui assure à chaque échantillon possible la même probabilité d'être sélectionné

Conséquence

Chaque unité a la même probabilité d'appartenir à l'échantillon: $\frac{n}{N}$

Propriétés de l'EAS

Ici la population est celle de *tous les échantillons possibles*

1. L'EAS est sans biais

Absence d'erreur systématique (en moyenne sur tous les échantillons possibles)

2. L'écart-type de l'erreur est $\frac{\sigma^*}{\sqrt{n}}\sqrt{1-f}$

- ne dépend des échantillons possibles qu'à travers σ^*, n, et f
- influence de f minime en pratique; influence de σ^* conforme à l'intuition
- erreur en $1/\sqrt{n}$: précision de plus en plus coûteuse

3. $\sqrt{n}(\bar{X}-\mu)$ tend vers une loi $\mathcal{N}(0, \sigma^{*2}(1-f))$

(quand la population devient infinie et f reste constant)

- permet de garantir que l'erreur n'est pas trop importante (avec grande probabilité)
- conséquence du "théorème de la limite centrale"

Exemple: taille des français

- $\mu = 174.1$
- $\sigma = 7.1$

On tire un échantillon de $n = 27$ individus, et on note $\bar{x}$ la moyenne empirique

- quelle est la moyenne théorique de $\bar{x}$?
- quel est l'écart-type théorique de $\bar{x}$?
- quel nombre d'échantillons n serait nécessaire pour diviser par 2 l'erreur d'échantillonnage ?
- avec ce n, quelle est la probabilité que la taille moyenne dans l'échantillon dépasse 1m80 ?

Illustration des propriétés 1 à 3

Echantillonnage stratifié

Uitle lorsqu'on sait *a priori* diviser la population en sous-groupes

- homogènes (de moyenne similaire)
- différents les uns des autres

Principe

- EAS dans chaque groupe
- Reconstitution de la moyenne des groupes

Bénéfice

diminution de l'erreur d'échantillonnage, d'autant plus importante que les groupes sont séparés

Exemple: revenu moyen des médecins (différentes spécialités médicales).