INTRODUCTION À L'INFÉRENCE STATISTIQUE

1. Introduction
2. Notion de variable aléatoire
   - Présentation
   - Variables aléatoires discrètes
   - Variables aléatoires continues
3. Représentations d'une distribution
   - Représentations graphiques
   - Résumés numériques
   - Représentations semi-graphiques.
4. Estimation

INTRODUCTION
Populations - Échantillons

En statistique, on appelle *population* une collection d'éléments possédant au moins une caractéristique commune permettant de les regrouper.

Un élément est un individu ou une unité statistique.

Si le nombre d'éléments est limité, la population est dite *finie*. Si ce nombre est illimité ou difficilement calculable, la population est dite *infinie*.

On définit un *échantillon* comme un sous-ensemble de la population statistique.

INTRODUCTION
Les deux types de démarches statistiques

finie, petite et connue

POPULATION → RECENSEMENT

infinie et connue

SONDAGE

ÉCHANTILLON →

STATISTIQUE *EXPLORATOIRE*
on ne s'intéresse qu'aux individus
dont on a les mesures

STATISTIQUE *INFÉRENTIELLE*

VARIABLE ALÉATOIRE
Notion de phénomène aléatoire

Dans de nombreux cas, la répétition d'une expérience dans des conditions apparemment identiques ne conduit pas toujours au même résultat.

Exemples:

- mélange à parts égales d'un produit **A** et d'un produit **B** et examen du résultat du mélange: produit **C**;
- semis de graines dans une terrine et comptage du nombre de levées après 5 jours;
- lancement d'une pièce de monnaie;
- jet d'un dé et examen du nombre indiqué sur la face supérieure.

Si le résultat d'une expérience ne peut être déterminé par la connaissance des conditions initiales, nous dirons que le phénomène est *aléatoire*.

VARIABLE ALÉATOIRE
Définition d'une variable

Une *variable* X est une application d'un ensemble (Ω) d'événements dans un ensemble S de valeurs numériques ou non (appelées réalisations).

Ω est un ensemble discret d'objets, d'individus, d'occasions, ...:

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$$

alors que S peut être n'importe quoi.

En particulier, les valeurs de S peuvent être *numériques*, *ordinales* ou *nominales*.

VARIABLE ALÉATOIRE
Exemples de variables

Exemples:

Soit Ω une population d'individus et

- X_1 la variable `sexe` prenant ses valeurs dans S ={homme, femme};
- X_2 la variable `diplôme` prenant ses valeurs dans $S = \{$certificat d'étude, $\ldots$, thèse $\}$
- X_3 la variable `poids` (en kg) prenant ses valeurs dans l'intervalle $S = [0, 200]$;

VARIABLE ALÉATOIRE
Définition

Une *variable aléatoire* X est une variable associée à une expérience **aléatoire** et servant à caractériser le résultat de cette expérience. Autrement dit, à chaque réalisation (valeur de S) est associée

- une probabilité si la variable est discrète;
- une densité de probabilité si la variable est continue (Ces deux notions seront vues plus loin.)

Exemples

- On jette un dé bleu et un dé rouge et on considère la somme X du dé bleu et du dé rouge;
- On jette un dé bleu et un dé rouge et on considère la valeur Y correspondant à la valeur absolue de la différence entre les valeurs des 2 dés.
- On prend au hasard un ananas dans la récolte d'un champ et on considère le poids Z de l'ananas.

X, Y, Z sont des *variables aléatoires*.

VARIABLE ALÉATOIRE

Il existe plusieurs types de *variable aléatoire*.

Les types les plus fréquents qui seront définis sont:

- les *variables aléatoires discrètes*
  Ex: somme de 2 dés, ...
- les *variables aléatoires continues*
  Ex: taille des individus dans une population, ...

VARIABLE ALÉATOIRE DISCRÈTE
Définition

- l'ensemble des réalisations possibles (S) d'une telle variable aléatoire (notée X) a un *nombre fini (ou infini dénombrable) d'éléments*
- à chacune des valeurs $x \in S$ que peut prendre la variable aléatoire X, correspond une probabilité $P(x)$ ou P_x;

$$P(x) = P_x = P(X = x)$$

- l'ensemble des valeurs x et des probabilités correspondantes P_x définit une *distribution de probabilité*;
- l'ensemble des probabilités cumulées définit une *fonction de répartition*:
  $F(x) = P(X \leq x)$

VARIABLE ALÉATOIRE DISCRÈTE
Exemple

Exemple: Jet de 2 dés et calcul de la somme.

x	$\Pr\{X = x\}$	$F(x)$
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36

VARIABLE ALÉATOIRE CONTINUE
Définition

- l'ensemble des réalisations possibles d'une telle variable aléatoire (notée X) a un *nombre de valeurs non dénombrables*;
- il n'est plus possible d'associer à chacune des valeurs x que peut prendre la variable aléatoire X une probabilité $P(x)$ ou P_x;
- par contre, il est possible de définir une *fonction de répartition*:

$$F(x) = P(X \leq x)$$

- de même on peut définir la probabilité d'observer une valeur comprise dans un intervalle donné $[a; b]$

$$P(a \leq X \leq b) = F(b) - F(a)$$

- si F est dérivable on peut encore écrire

$$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta F}{\Delta x} = f(x)$$

-

$$F(x) = \int_{-\infty}^{x} f(x) d(x)$$

f est appelée densité

REPRÉSENTATION DES DONNÉES
Exemple: répartition par âge des agents INRA

Âge	Effectif	Effectif cumulé	Fréquence	Fréquence cumulée
19	2	2	0.0002	0.0002
20	0	2	0.0000	0.0002
21	7	9	0.0008	0.0010
22	41	50	0.0049	0.0059
23	66	116	0.0079	0.0138
24	121	237	0.0145	0.0283
25	128	365	0.0153	0.0436
26	191	556	0.0230	0.0666
⋮	⋮	⋮	⋮	⋮
59	152	7992	0.0183	0.9604
60	102	8094	0.0123	0.9727
61	66	8160	0.0079	0.9806
62	62	8222	0.0075	0.9881
63	52	8274	0.0063	0.9944
64	42	8316	0.0050	0.9994
65	5	8321	0.0006	1.0000
	8321		1.0000	

REPRÉSENTATION DES DONNÉES

Les représentations graphiques

- Diagramme en Bâtons
- Histogramme
- Densité
- Fonction de répartition

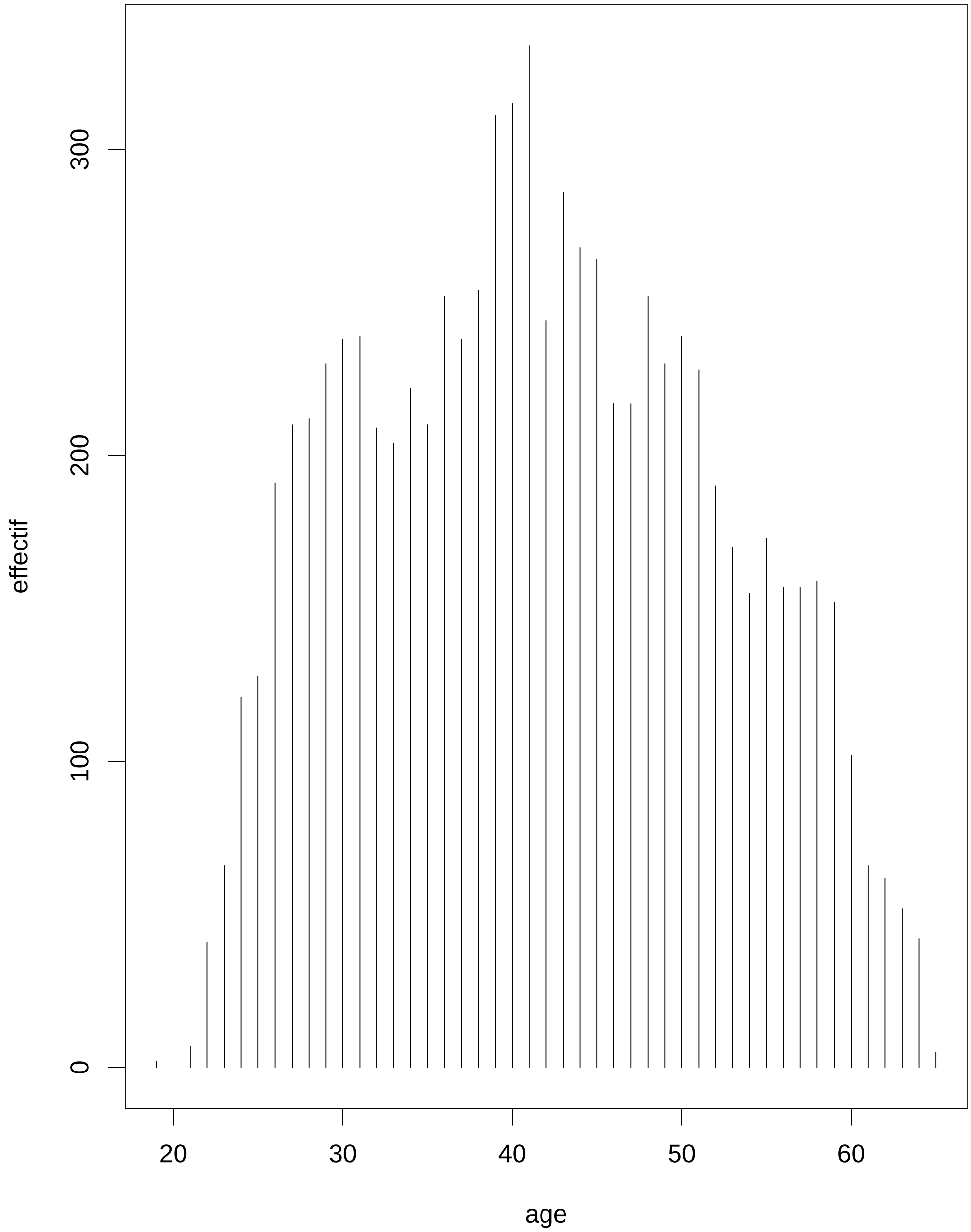
Les représentations numériques

- de tendance centrale: médiane, moyenne
- de dispersion: variance, écart-type, quantiles, étendue

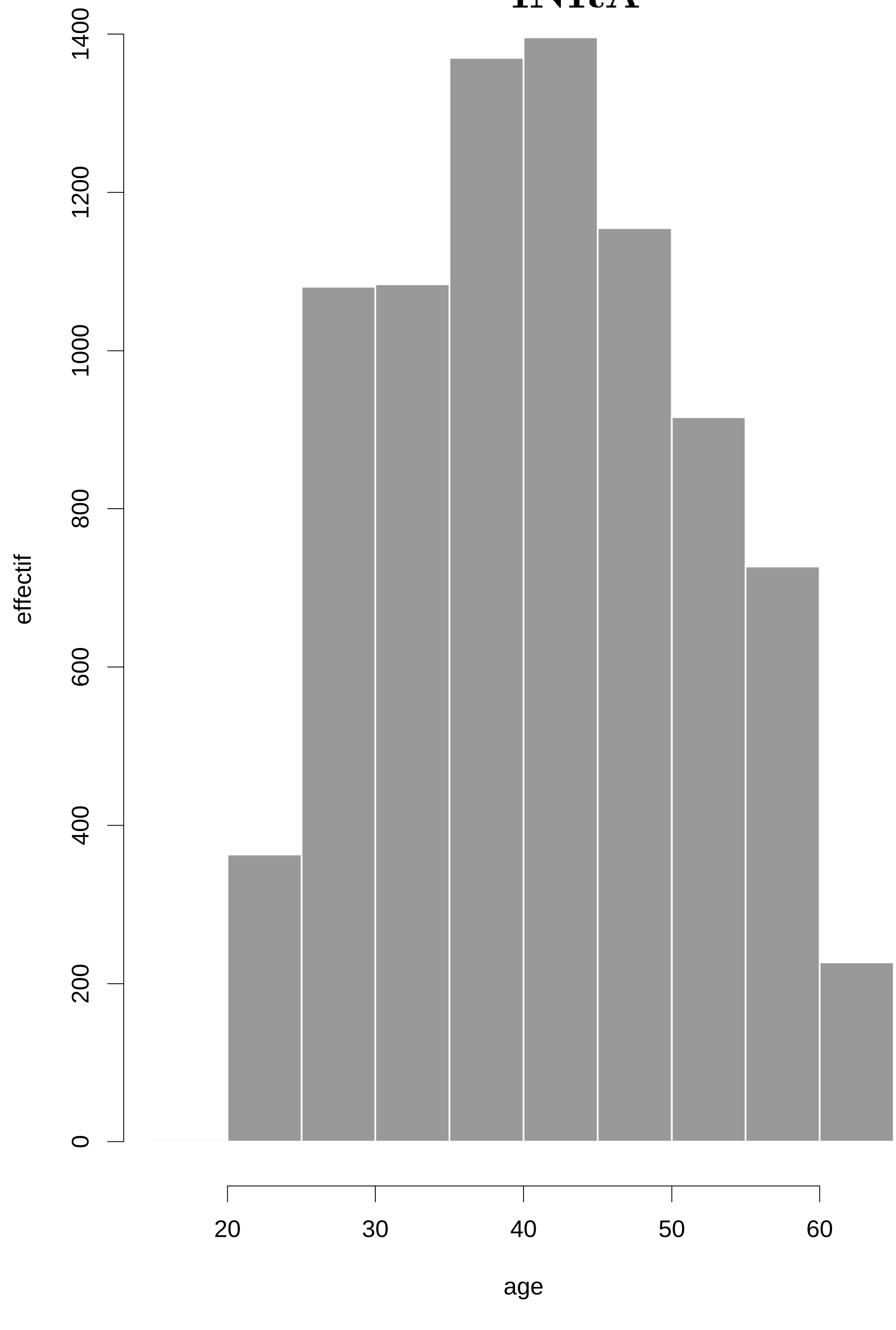
Les représentations semi-graphiques

- boîte à pattes (box-plot)
- branchage (stem and leaf)

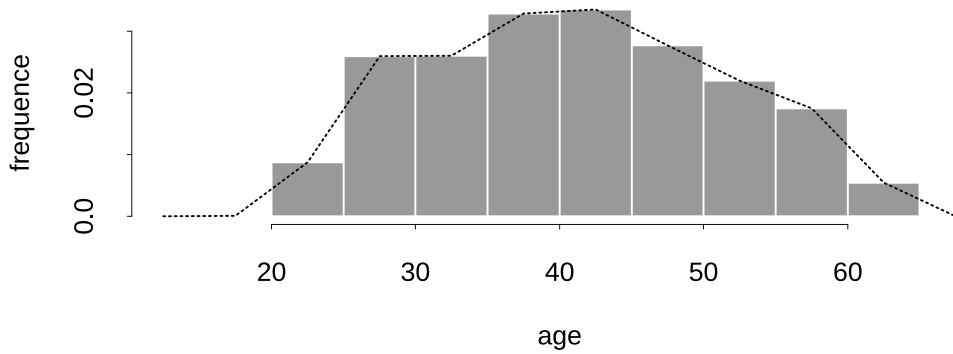
LES REPRÉSENTATIONS GRAPHIQUES
Diagramme en bâton de la population INRA



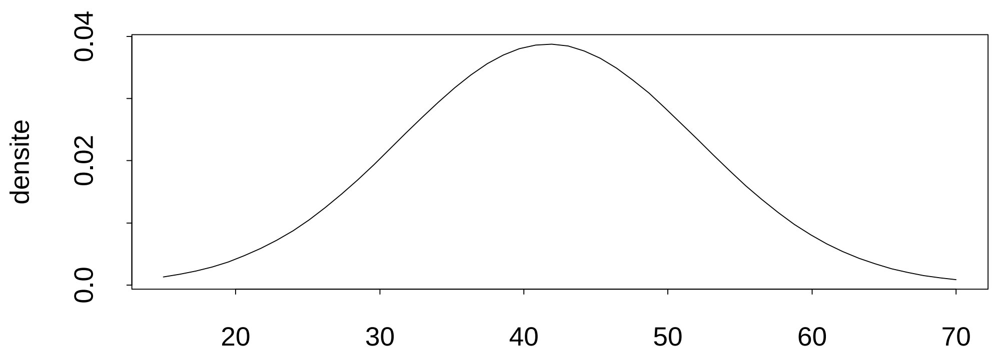
LES REPRÉSENTATIONS GRAPHIQUES
Histogramme (des effectifs) des âges des agents INRA



LES REPRÉSENTATIONS GRAPHIQUES
Polygone des fréquences - Courbe de densité

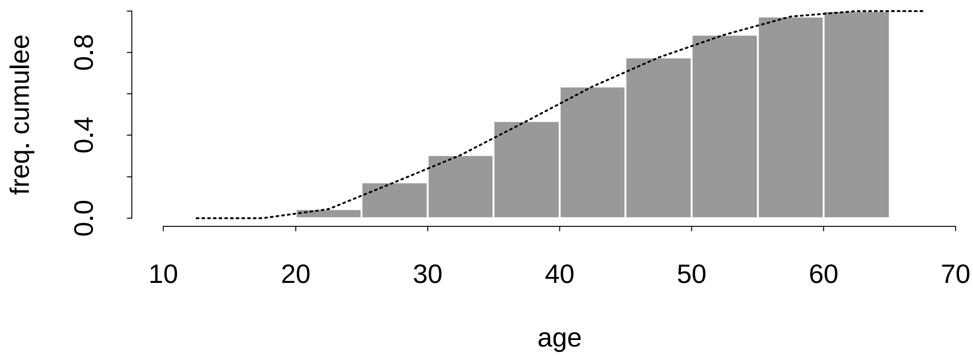


histogramme et polygone des fréquences d'âges

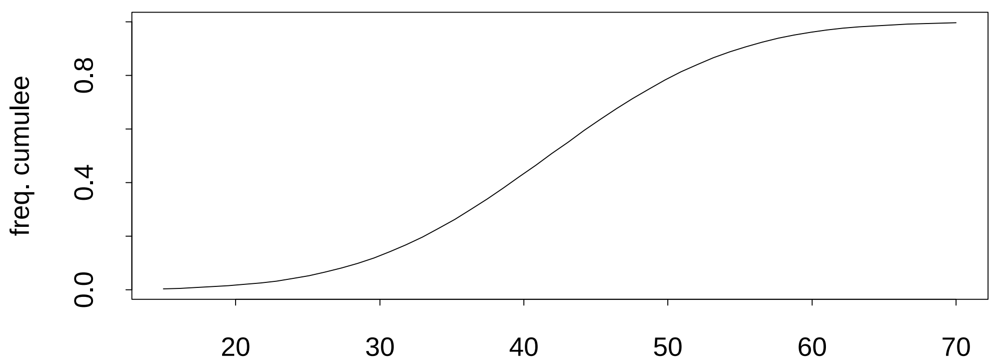


courbe de densité pour une variable continue

LES REPRÉSENTATIONS GRAPHIQUES
Polygone des fréquences cumulées
Courbe de fonction de répartition



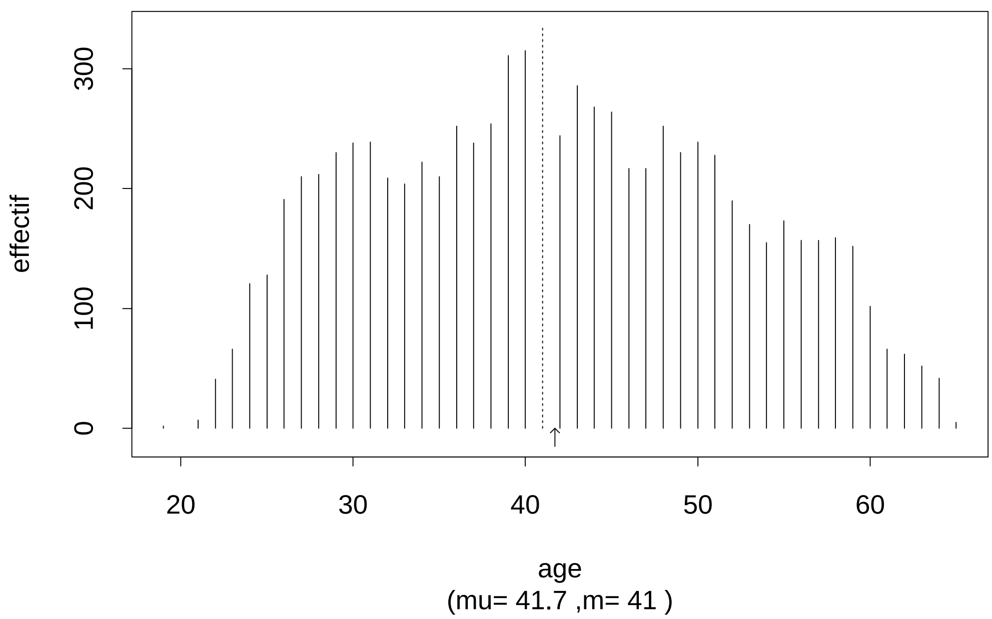
histogramme et polygone des fréquences cumulées d'âges



courbe de fonction de répartition pour une variable continue

RÉSUMÉS NUMÉRIQUES:

Tendance centrale : espérance et médiane



espérance : moyenne arithmétique des réalisations pondérées par leur probabilité.

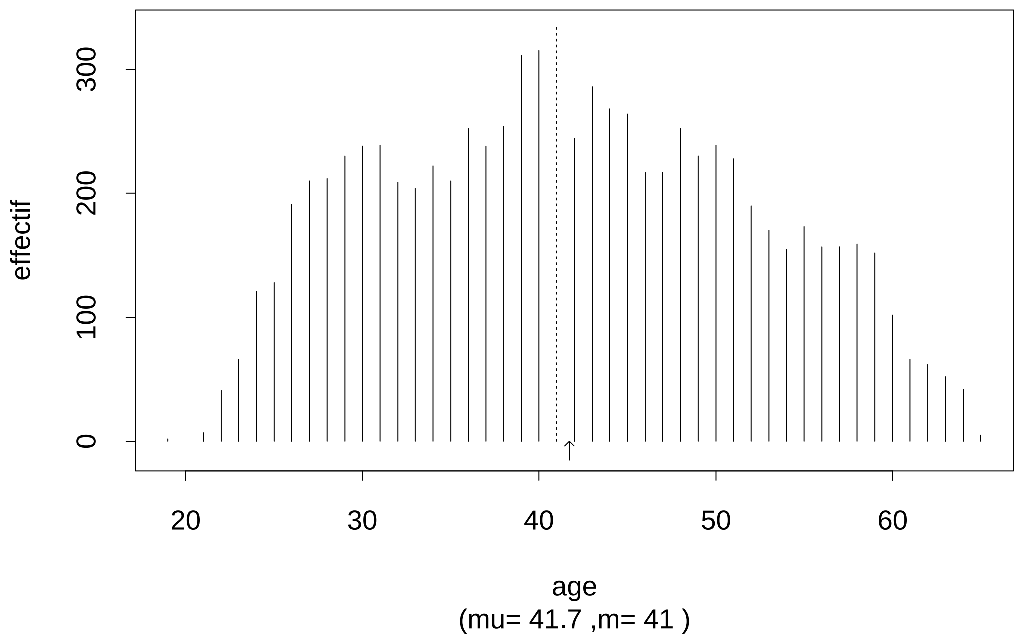
$$\mathrm{E}\,(X) = \mu = \sum_{x \in S} x P_x.$$

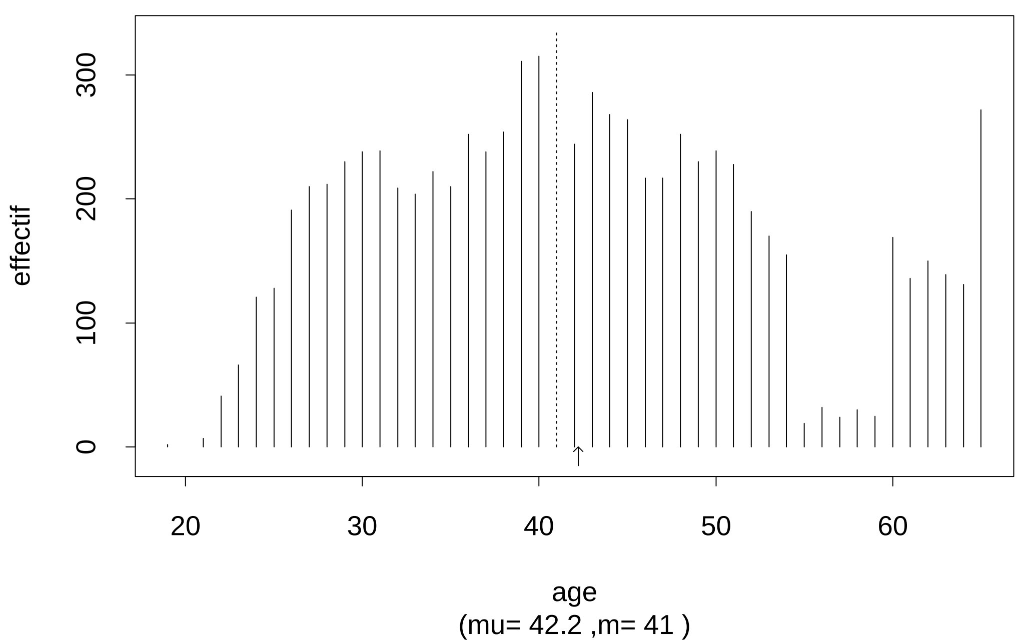
médiane : valeur m telle que

$$\mathrm{P}\,(X < m) \simeq \mathrm{P}\,(X > m) \simeq 1/2.$$

RÉSUMÉS NUMÉRIQUES :

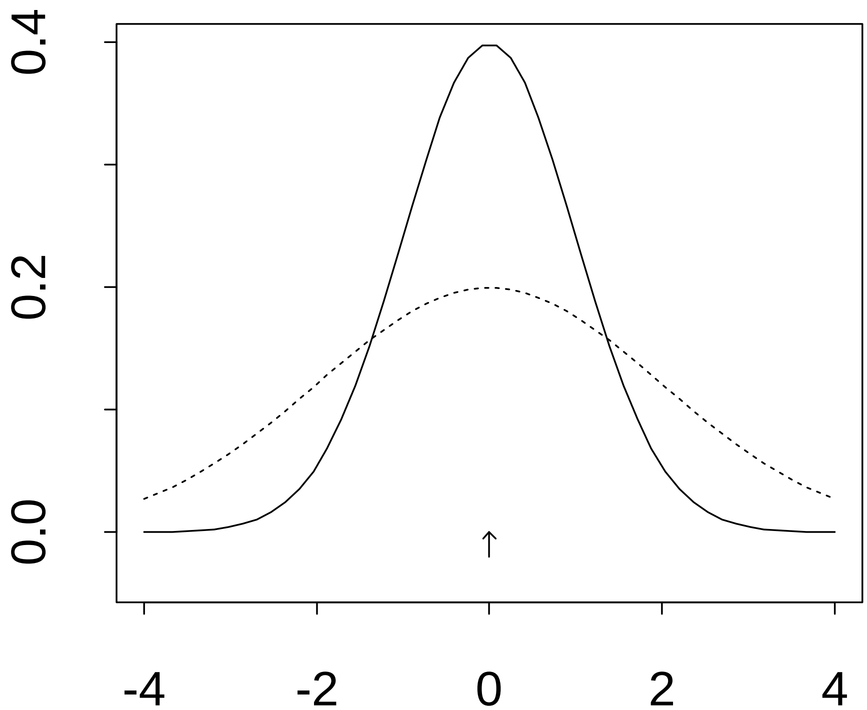
Stabilité de la médiane





RÉSUMÉS NUMÉRIQUES:

tendance centrale $\neq$ dispersion



RÉSUMÉS NUMÉRIQUES:

Dispersion : variance et écart-type

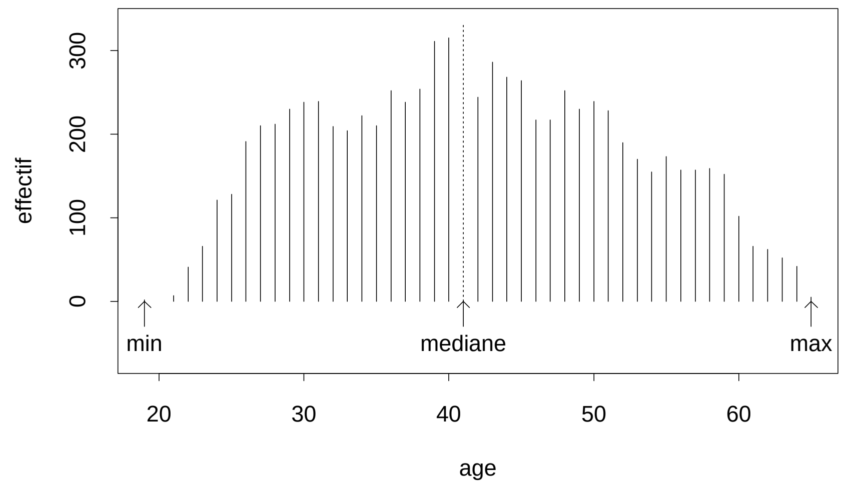
Écart à l'espérance : $X - \mu$.
Mesurer la dispersion par $\mathrm{E}\,(X - \mu)$? $(= 0)$

Carré de l'écart à l'espérance : $(X - \mu)^2$.
Variance : $\sigma^2 = \mathrm{E}\left((X - \mu)^2\right)$.
Formule de calcul :

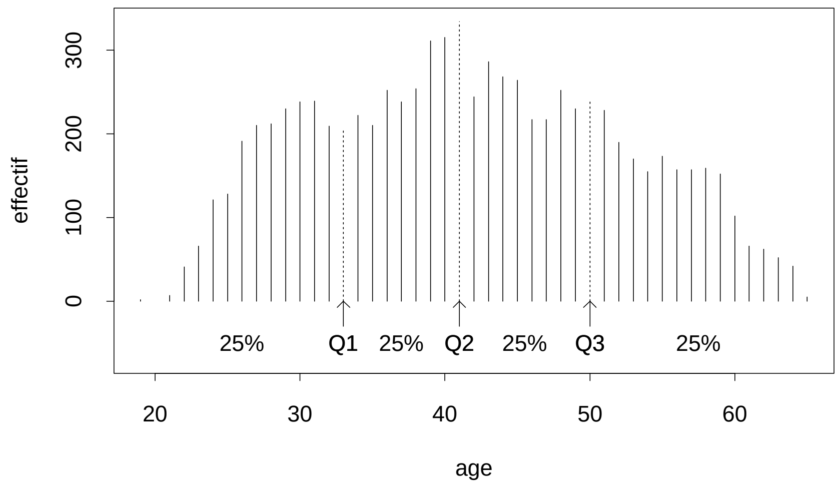
$$\sigma^2 = \sum_{x \in S} (x - \mu)^2 \, P_x .$$

Écart-type : σ.

RÉSUMÉS NUMÉRIQUES
Dispersion : étendue, quartiles et IQR



Étendue : valeur maximale − valeur minimale

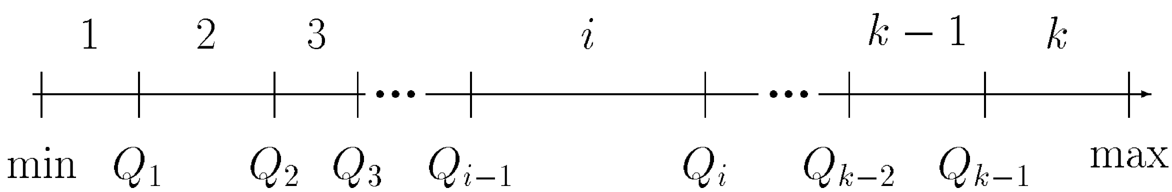


Quartiles : Q_1, Q_2, Q_3.

Intervalle interquartile : $Q_3 - Q_1$.

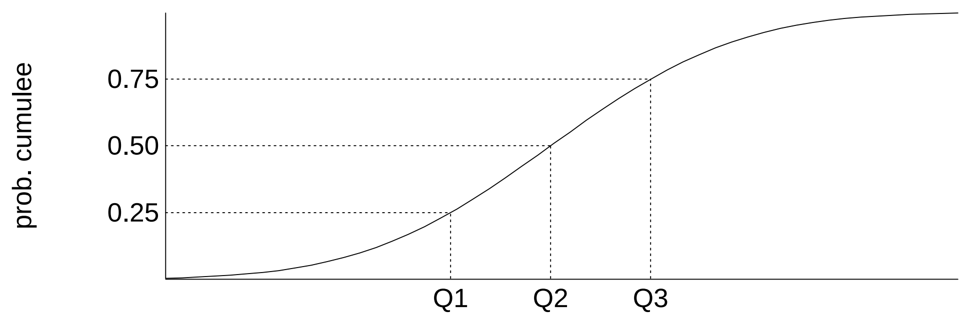
RÉSUMÉS NUMÉRIQUES
Dispersion : quantiles

k parties :

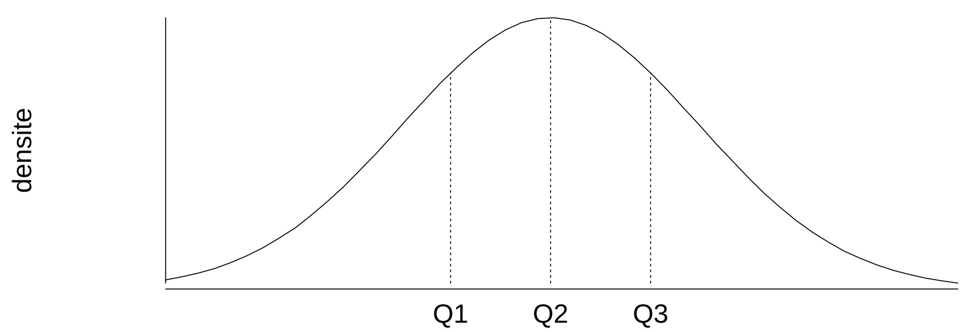


$100\ (\frac{i}{k})$ % des observations ont une valeur inférieure à Q_i

RÉSUMÉS NUMÉRIQUES
Dispersion : quantiles



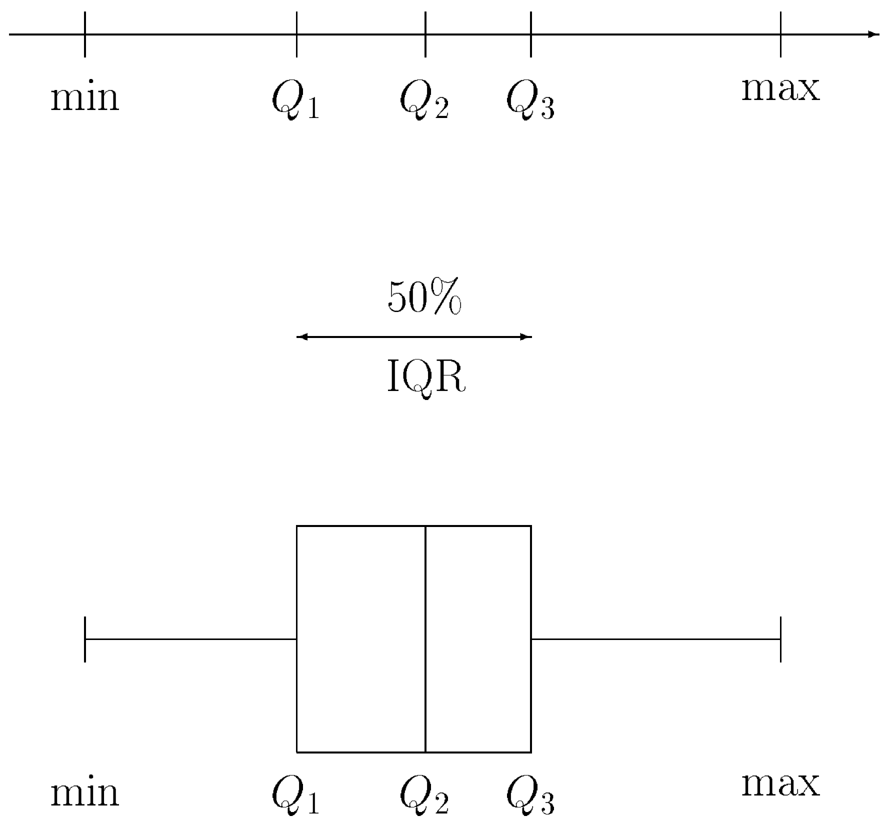
courbe de fonction de répartition



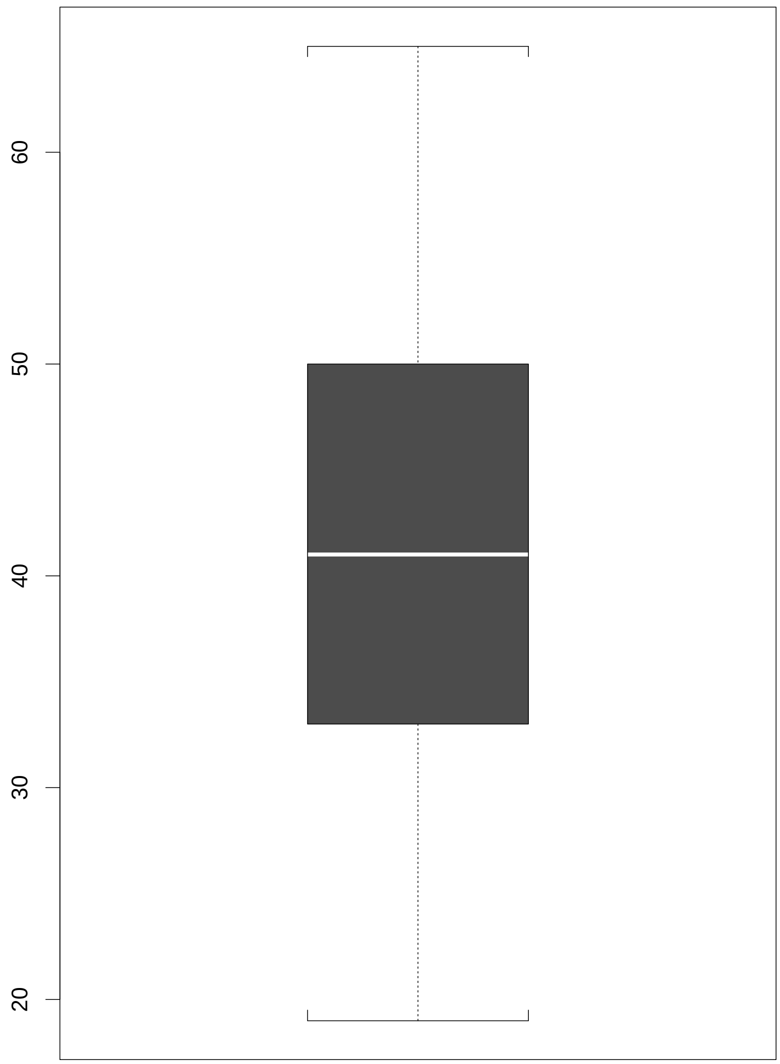
courbe de densité

LES REPRÉSENTATIONS SEMI-GRAPHIQUES

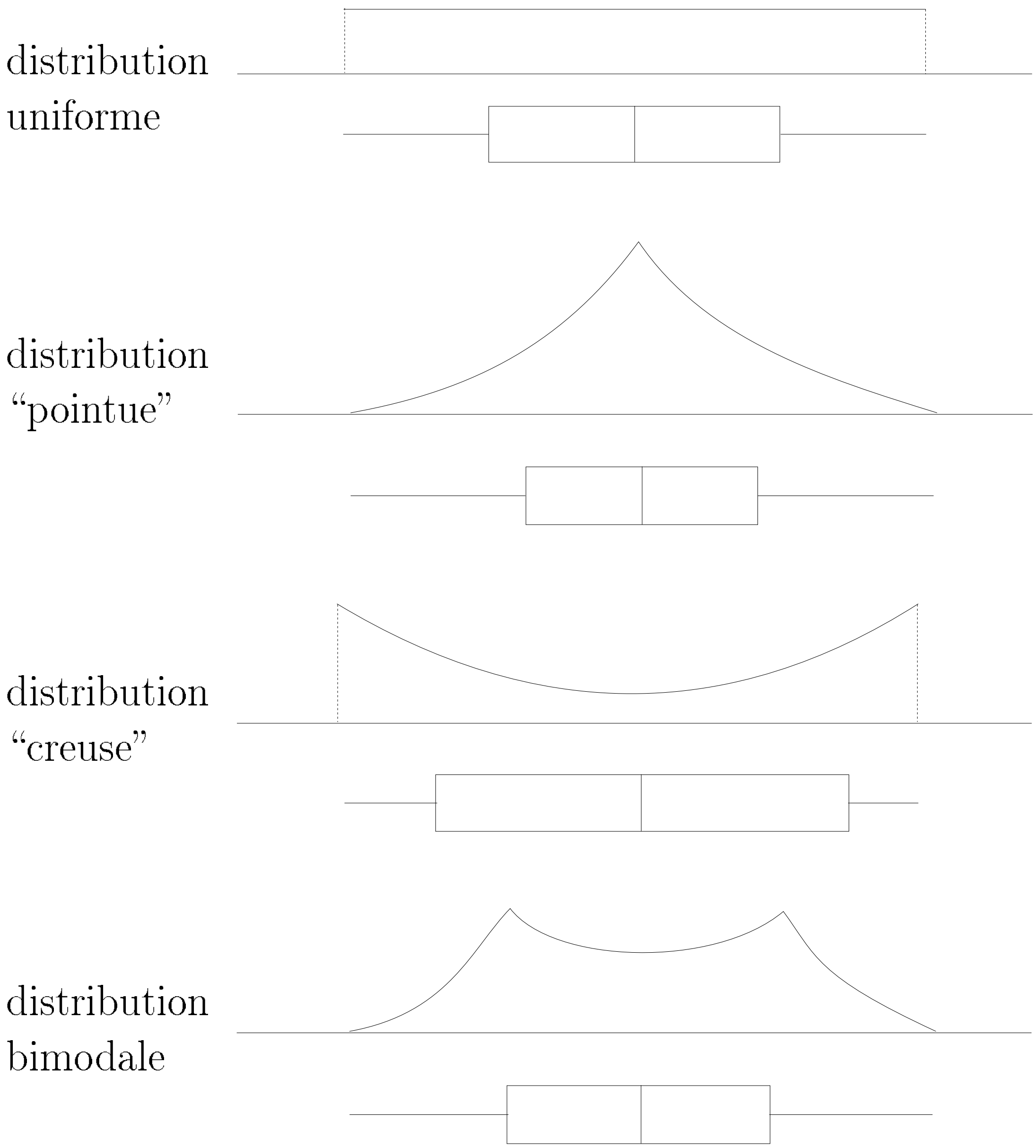
La boîte à pattes



LES REPRÉSENTATIONS SEMI-GRAPHIQUES
Exemple : la boîte à pattes de l'âge des agents INRA

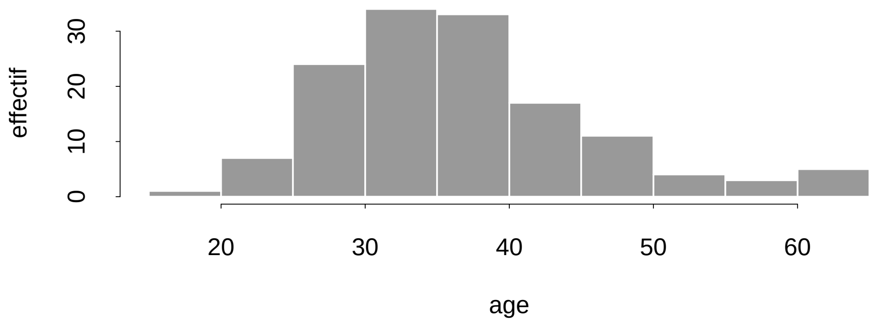


LES REPRÉSENTATIONS SEMI-GRAPHIQUES
Exemples de boîtes à pattes



LES REPRÉSENTATIONS SEMI-GRAPHIQUES
Exemple : branchage de l'âge des agents INRA

Histogramme des âges des 139 agents de Lille



Branchage des âges du Centre de Lille

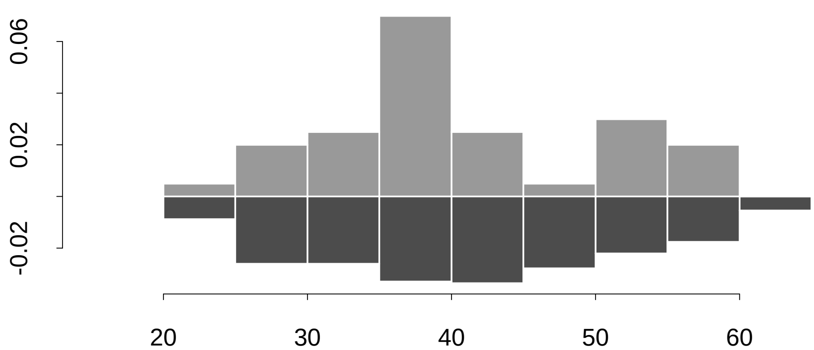
```
N = 139    Median = 36
Quartiles = 31, 42
Decimal point is 1 place to the right of the colon
     1 :  9
     2 :  13444
     2 :  556666666777888889999
     3 :  00000011111111111222222333333334444
     3 :  55555666666777777888888899999999
     4 :  000000112222222334444
     4 :  5566667889
     5 :  0004
     5 :  55566
     6 :  012233
```

ESTIMATION
estimation empirique (1)

Sondage : 40 agents (environ 0.5% de la population totale)

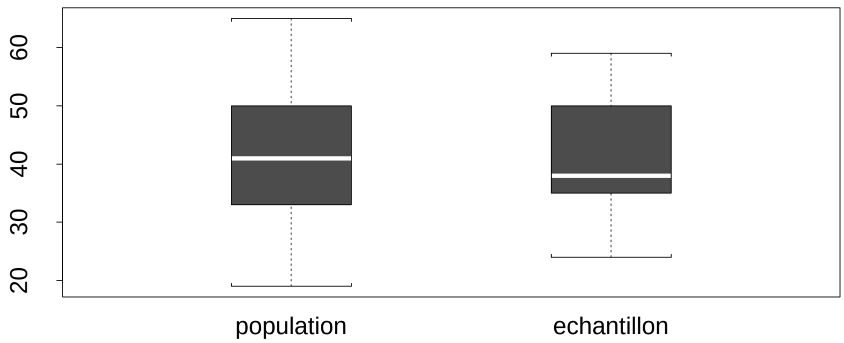
24	31	36	36	38	42	51	55
27	31	36	36	38	43	52	56
27	32	36	37	40	44	53	56
29	34	36	37	40	44	53	58
29	34	36	38	42	49	54	59

Distribution de la population et distribution de l'échantillon



ESTIMATION
estimation empirique (2)

Boîtes à pattes



Résumés numériques

paramètres	population	échantillon
espérance	41.7	40.7
variance	105.9	92.5
Q_1	33.0	35.5
Q_2	41.0	38.0
Q_3	50.0	49.5

PARAMÈTRES - ESTIMATEURS
Définitions

On appelle *paramètre* la caractéristique quantitative qui permet une représentation condensée de l'information contenue dans une ou plusieurs populations.

L'expression mathématique permettant de mesurer, à partir des données de l'échantillon, un paramètre de la population s'appelle un *estimateur* d'un paramètre.

C'est une variable aléatoire dont on espère que la valeur sera "souvent proche" du paramètre que l'on cherche à estimer.

EXEMPLE : Si on a observé $S = \{1, 1, 4, 5\}$, alors la moyenne arithmétique des observations

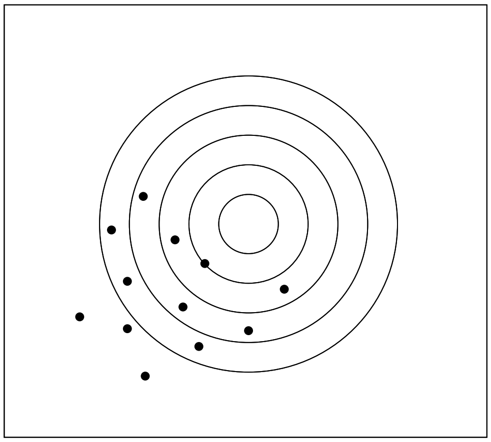
$$\overline{X} \;=\; \frac{1}{4}(1+1+4+5)$$

est un estimateur de l'espérance μ.

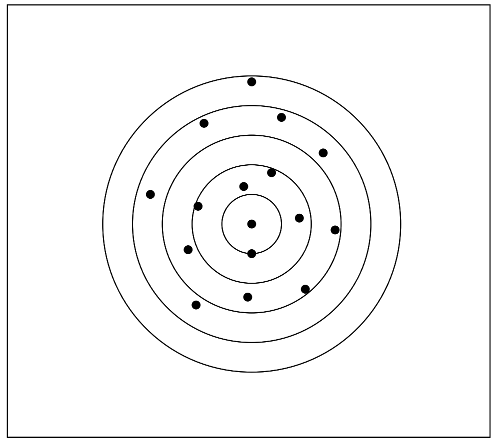
Formule générale :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_n.$$

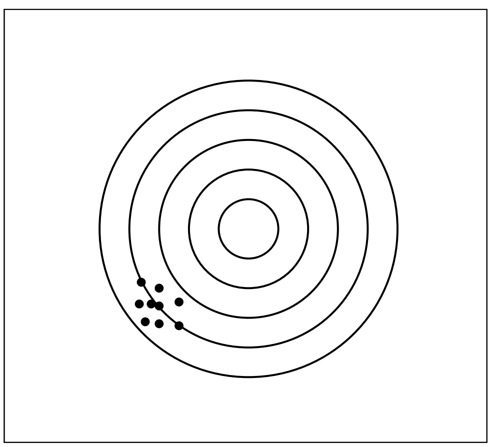
PARAMÈTRES - ESTIMATEURS
Propriétés des estimateurs



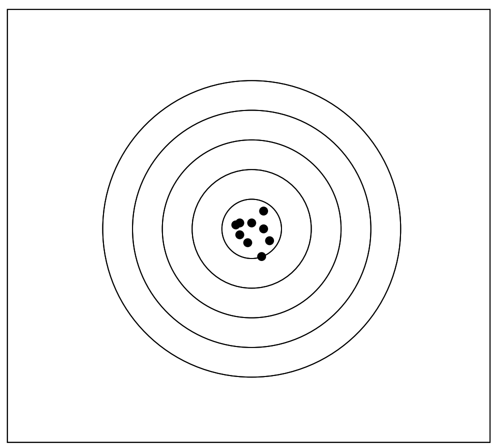
estimateurs biaisés avec grande variance



estimateurs non biaisés avec grande variance



estimateurs biaisés avec petite variance



estimateurs non biaisés avec petite variance

ESTIMATION
Définition

L'*estimation* est la valeur prise par un estimateur pour un échantillon particulier.

L'estimation d'un paramètre à partir d'un échantillon unique ne conduit généralement pas à la vraie valeur du paramètre. Cette estimation va varier d'un échantillon à l'autre.

La réalisation d'un très grand nombre d'échantillons de même taille permet de construire la *distribution (d'échantillonnage)* de l'estimateur.

L'estimation d'un paramètre peut être *ponctuelle* ou par *intervalle*.

PARAMÈTRES - ESTIMATEURS

POPULATION	ÉCHANTILLON
fixe	aléatoire
paramètres théoriques	versions empiriques = estimateurs
	$\downarrow$
"tout est connu"	INFÉRENCE
	$\downarrow$
	"information" sur les paramètres de la population inconnue

PARAMÈTRES - ESTIMATEURS

NOTATIONS

	paramètres théoriques (population)	version empiriques (échantillon)
médiane	m	$\widetilde{X}$ $(\widehat{m})$
moyenne	μ	$\overline{X}$ $(\widehat{\mu})$
variance	σ^2	$\widehat{\sigma}^2$ (S^2)
fonction de répartition	F	$\widehat{F}$