Rappels et précisions préalables

Fiche résumé - Vocabulaire utilisé en statistique -

Désignation	Courte définition	Exemple
Population	Ensemble des individus étudiés, répondant à une propriété donnée (modalité d'une variable qualitative). La population étudiée est plus ou moins précise ; elle peut-être caractérisée par les paramètres exacts d'une variable aléatoire continue considérée dans l'étude. (*la taille de la population n'est en général pas connue avec précision, c'est pourquoi on ne la considère pas dans l'inférence statistique*)	Population des personnes atteintes par une maladie M (caractérisée, par exemple, par la moyenne exacte et l'écart-type exact de l'âge de ces personnes)
Individu	Un élément de la population étudiée, (vient du latin *individuum*, signifiant ce qui est indivisible) ; un des objets étudiés. (*en biologie, il n'existe pas deux individus totalement identiques*)	un patient, un écologue, une micropipette ; un yaourt, un ours des Pyrénées ; un grain de maïs ; une parcelle cultivée ; un tube à essai ; une personne atteinte par une maladie M ; un logiciel ; un profil Facebook
Unité statistique	Autre façon de désigner un individu	un patient ; un écologue ; une micropipette ; un yaourt ; un ours des Pyrénées ; un grain de maïs ; une parcelle cultivée ; un tube à essai ; une personne atteinte par une maladie M
Echantillon	Ensemble d'individus extraits d'une population. Il s'agit donc d'un sous ensemble de la population (de bien plus petite talle que la population, ce qui en rend l'étude réalisable) Un échantillon est dit représentatif de la population dont il est issu s'il a été tiré au hasard (indispensable en statistique inférentielle) (*l'analyse d'un Big Data peut conduire à étudier l'ensemble complet des données d'une population*)	28 poissons pêchés dans un étang normand ; 65 parcelles de culture d'une nouvelle variété céréalière ; N yaourts issus de la production d'une journée de la chaîne agroalimentaire de l'usine A; 42 micropipettes utilisées dans un laboratoire de recherche, 15 échantillons de salive de personnes touchées par un virus V
Variable	C'est le caractère étudié dans la population et qui varie d'un individu à l'autre. Variable aléatoire [VA] : application définie sur l'ensemble de définition. Les variables manipulées en statistiques sont qualitatives (nominales ou ordinales) ou quantitatives (discrètes ou bien continues). *(C'est le support de l'information, par exemple de l'information biologique étudiée. Dans le logiciel R, vous manipulez un vecteur)*	Concentration du micropolluant dans l'eau du robinet (en ng/ml); nombre de personnes ayant résisté au virus V dans un échantillon de N personnes atteintes par le virus ; Etat du patient (guéri/malade) ; stades de développement d'une maladie grave (non développée, premier stade, stade avancé, stade sévère).
Observation	Valeurs prises par la variable pour un individu donné (exemple de notation : $X=x_i$) On s'intéresse par exemple à la probabilité $P(X>x_i)$ On s'intéresse ainsi aux couples (i, x_i)	La mesure de la taille d'une des abeilles suivie à l'aide d'un nanoémetteur, a donné x_i=1,97 cm
Modalité	Les modalités sont les valeurs prises par la variable associée au caractère étudié. *(La variable manipulée dans un logiciel regroupe l'ensemble des valeurs prises par le caractère étudié. Il s'agit d'un vecteur)*	X : Nombre d'abeilles qui ne sont pas rentrées à leur ruche sur un échantillon de 8 abeilles suivies à l'aide d'un nanoémetteur; X peut prendre les modalités : 0,1,2,3,4,6,7,8 Etat du patient : «guéri(e)», «stationnaire», «moins malade», «encore plus malade »

Calcul des paramètres statistiques et de leurs estimations

▶ paramètres de position, paramètres de dispersion, facteurs de forme

Statistique descriptive **Décrire** *ab minima* **/ Résumer** **la distribution**

Paramètres

- **Position :** moyenne, mode, médiane, min, max…
  (*ordre de grandeur*); tendance centrale
- **Dispersion :** écart-type, variance, quantiles, étendue, …
  (*répartition autour de l'ordre de grandeur*)
- **Forme :** symétrie, aplatissement,…
  (tendance)

Tableaux synthétiques rassemblant **paramètres** caractéristiques et leurs **Intervalles de Confiance**

Graphes Histogrammes, boîtes à moustache, lissage normal, …

Test normalité de la distribution (Kolmogorov-Smirnov, Lilliefors, Shapiro-Wilk, Chi-deux,...)

⇨ **Comparer** **échantillon (s) - population (s)**

Informations (fichier) à décrire du mieux possible pour permettre leur analyse

Rappel : somme de n valeurs

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

i : indice de l'individu, de l'observation ou de la mesure

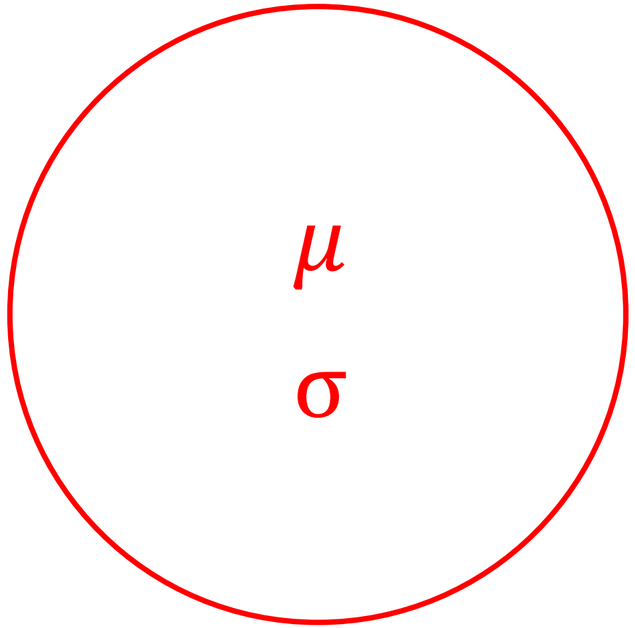
x_i : valeur observée ou mesurée (observation/mesure) pour l'individu i

n : nombre total d'individus, d'observations ou de mesures

Calcul des paramètres statistiques et de leurs estimations

▶ deux mots sur les codes couleurs adoptés

X : **variable aléatoire** d'étude, suit généralement une **loi de probabilité**



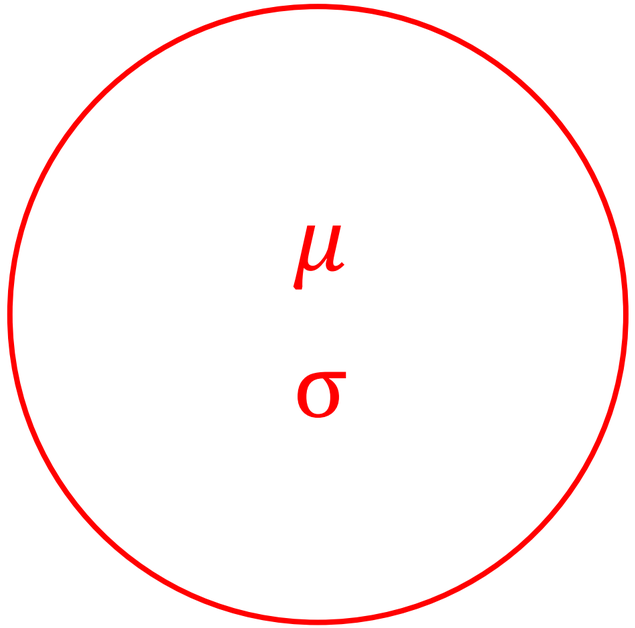
Population : caractérisée par

2 paramètres exacts : μ et σ

Calcul des paramètres statistiques et de leurs estimations

▶ deux mots sur les codes couleurs adoptés

X : variable aléatoire d'étude



Population : caractérisée par

2 paramètres exacts : μ et σ

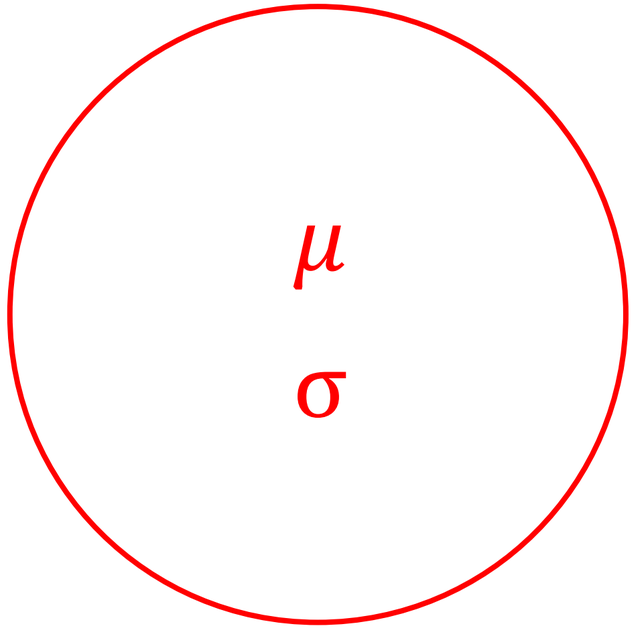
- **rouge** pour ce qui se rapporte **à la population**

- Lettres **Majuscules** ▶ **Variables Aléatoires**
- lettres **grecques** ▶ **paramètres exacts** d'une **population**

Calcul des paramètres statistiques et de leurs estimations

▶ deux mots sur les codes couleurs adoptés

X : variable aléatoire d'étude



Population : caractérisée par

2 paramètres exacts : μ et σ

Espérance : μ

Ecart-type : σ ; **Variance** : σ^2

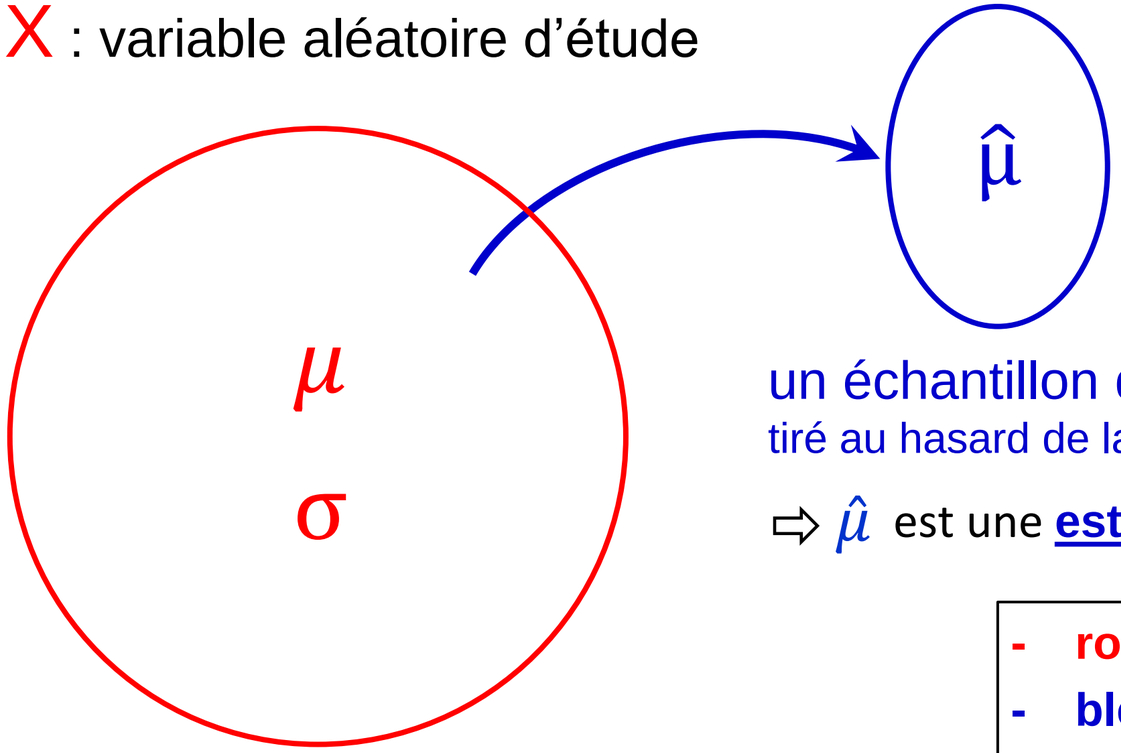
- **rouge** pour ce qui se rapporte **à la population**

- Lettres **Majuscules** ▶ **Variables Aléatoires**
- lettres **grecques** ▶ **paramètres exacts** d'une **population**

Calcul des paramètres statistiques et de leurs estimations

▶ deux mots sur les codes couleurs adoptés

X : variable aléatoire d'étude



⇨ $\hat{\mu}$ est également désignée par $\bar{x}$

$(x_1, x_2 , \dots x_i , \dots x_n)$

un échantillon de taille n
tiré au hasard de la population

⇨ $\hat{\mu}$ est une **estimation** de μ

Population : caractérisée par

2 paramètres exacts : μ et σ

Espérance : μ

Ecart-type : σ ; **Variance** : σ^2

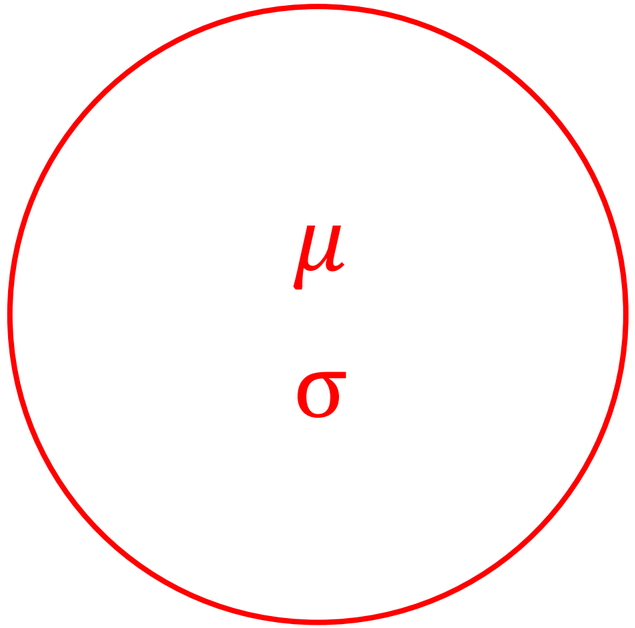
- **rouge** pour ce qui se rapporte **à la population**
- **bleu** pour ce qui se rapporte **à l'échantillon**
- lettres **minuscule** ▶ **observations**, **mesures**, **estimations**
- Lettres **Majuscules** ▶ **Variables Aléatoires**
- lettres **grecques** ▶ **paramètres exacts** d'une **population**
- **chapeaux** ▶ **estimations**

Calcul des paramètres exacts dans une population d’étude

Calcul des paramètres statistiques et de leurs estimations

▶ paramètres de position, espérance, moyenne exacte d'une population

X : variable aléatoire d'étude



Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

Espérance : **μ**

Ecart-type : **σ** ; **Variance** : σ^2

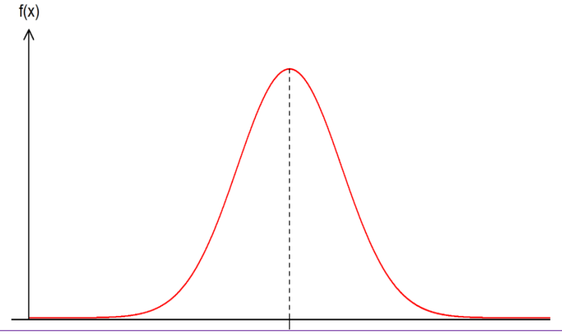
$$\mu = \mathrm{E}(X) = \frac{1}{N}\sum_{i=1}^{N} x_i \quad ; X \in \{x_1, x_2, \dots, x_N\}$$

si X suit une loi de probabilité discrète :

$$\mu = \mathrm{E}(X) = \sum_{i=1}^{N} x_i \,.\, P(X = x_i)$$

si X suit une loi de probabilité continue; $f(x)$ densité de probabilité :

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{+\infty} x.\, f(x)\, dx$$

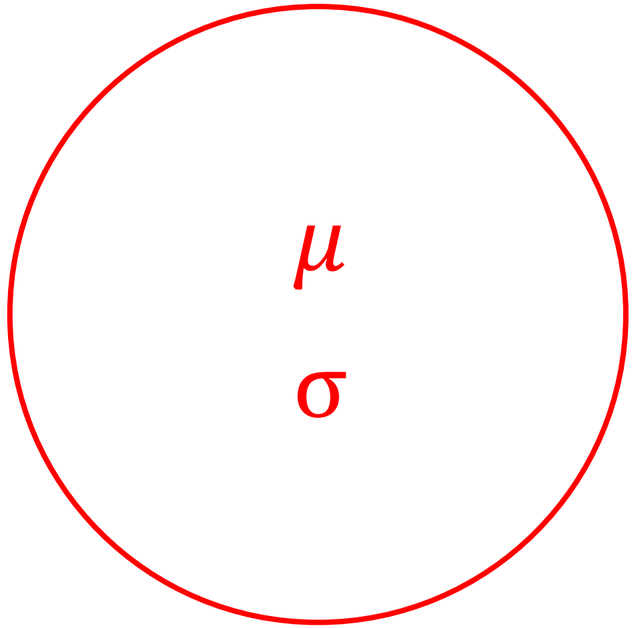


Une définition et un mode de calcul dépendant du contexte

Calcul des paramètres statistiques et de leurs estimations

▶ paramètres de dispersion, variance exacte d'une population

X : variable aléatoire d'étude



Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

Espérance : **μ**

Ecart-type : **σ** ; **Variance** : σ^2

$$X \in \{x_1, x_2, \dots, x_N\}$$

$$\sigma^2 = \mathrm{VAR}(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

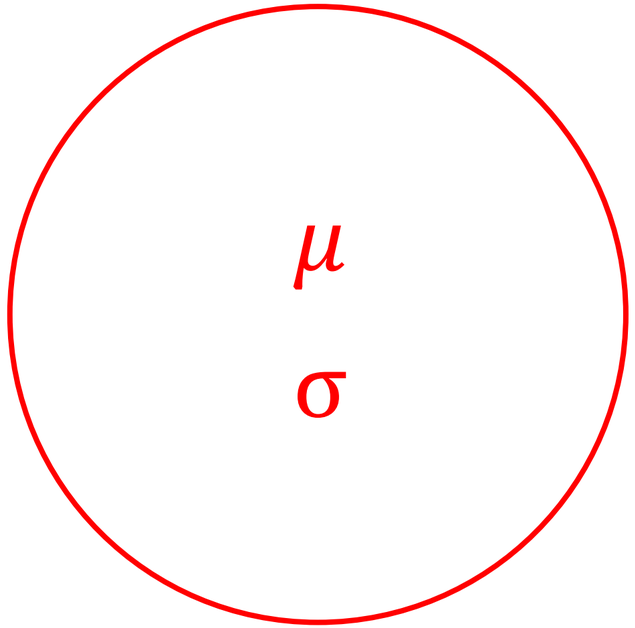
$$\sigma^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2$$ *[théorème de König-Huygens]*

$$\sigma^2 = \mathrm{VAR}(X) = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \mu^2$$

Calcul des paramètres statistiques et de leurs estimations

▶ paramètres de dispersion, variance exacte d'une population

X : variable aléatoire d'étude



Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

Espérance : **μ**

Ecart-type : **σ** ; **Variance** : σ^2

$$X \in \{x_1, x_2, \dots, x_N\}$$

$$\sigma^2 = \mathrm{VAR}(X) = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \mu^2$$ *(Könnig- Huyguens)*

si X suit une loi de probabilité discrète :

$$\sigma^2 = \mathrm{VAR}(X) = \sum_{i=1}^{N} {x_i}^2 . P(X = x_i) - \mu^2$$

si X suit une loi de probabilité continue; *f(x)* densité de probabilité :

$$\sigma^2 = \mathrm{VAR}(X) = \int_{-\infty}^{+\infty} {x_i}^2 . f(x) dx - \mu^2$$

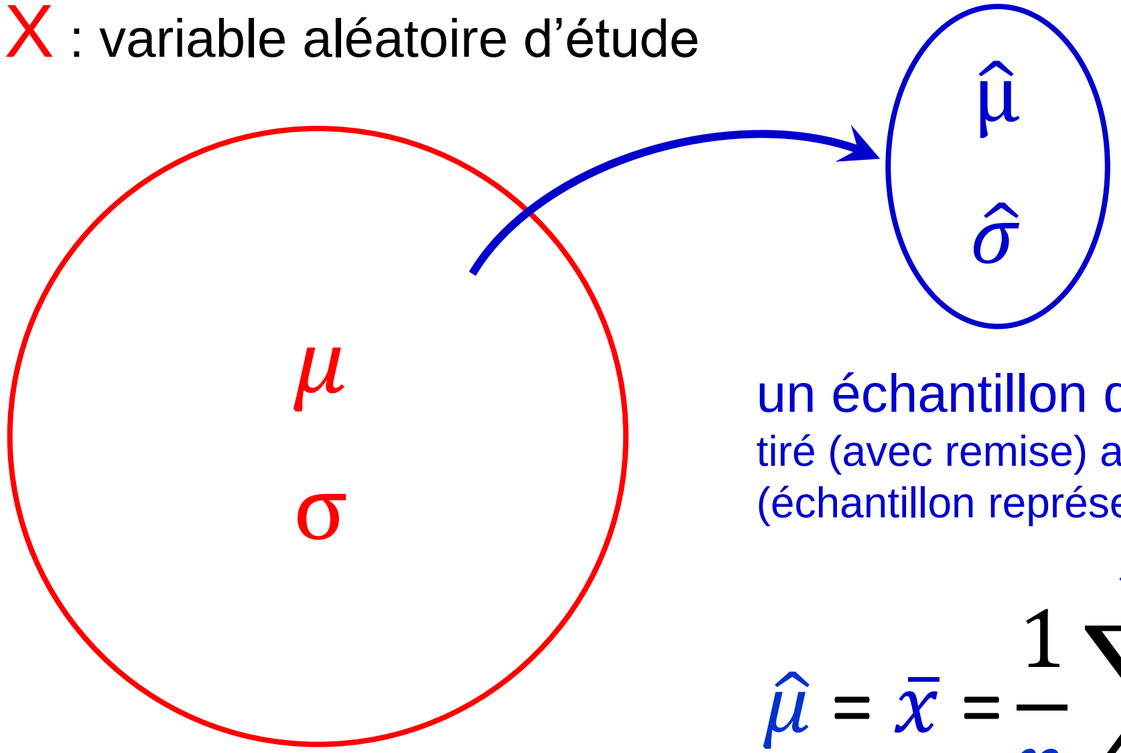
Une définition et un mode de calcul dépendant du contexte

Calcul des estimations des paramètres exacts dans un échantillon de taille n

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la moyenne exacte : moyenne calculée dans un échantillon

X : variable aléatoire d'étude



⇨ $\hat{\mu}$ est également désignée par $\bar{x}$

$(x_1, x_2 , \dots x_i , \dots x_n)$

un échantillon de taille n
tiré (avec remise) au hasard de la population
(échantillon représentatif de la population)

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

estimation ponctuelle de **μ**
(non biaisée)

Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

Espérance : **μ**

Ecart-type : **σ** ; **Variance** : σ^2

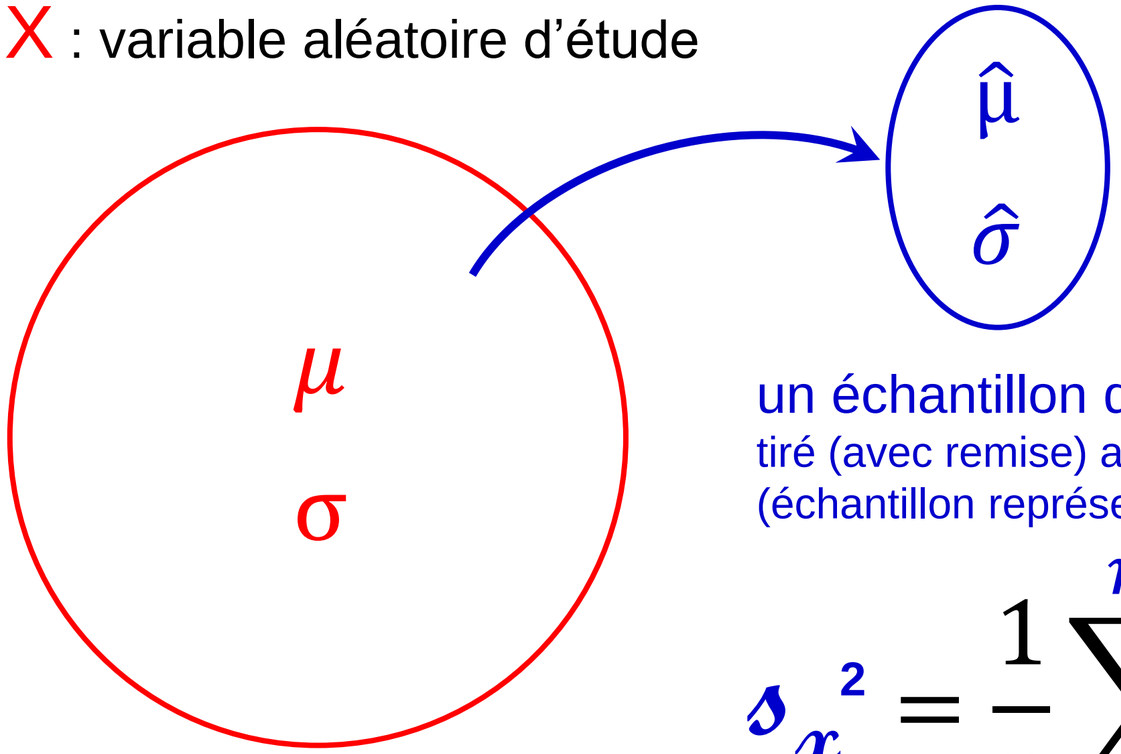
Si l'échantillon est tiré **au hasard** et de **taille suffisante**,
$\hat{\mu}$ est proche de la valeur de la moyenne exacte **μ** dans la population

La moyenne calculée dans un échantillon n'est qu'une estimation de la moyenne exacte dans la population

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la variance exacte : variance calculée dans un échantillon

X : variable aléatoire d'étude



$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

estimation ponctuelle de **μ**

(non biaisée)

un échantillon de taille n

tiré (avec remise) au hasard de la population

(échantillon représentatif de la population)

$$s_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

estimation ponctuelle de **σ**

(biaisée)

Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

Espérance : **μ**

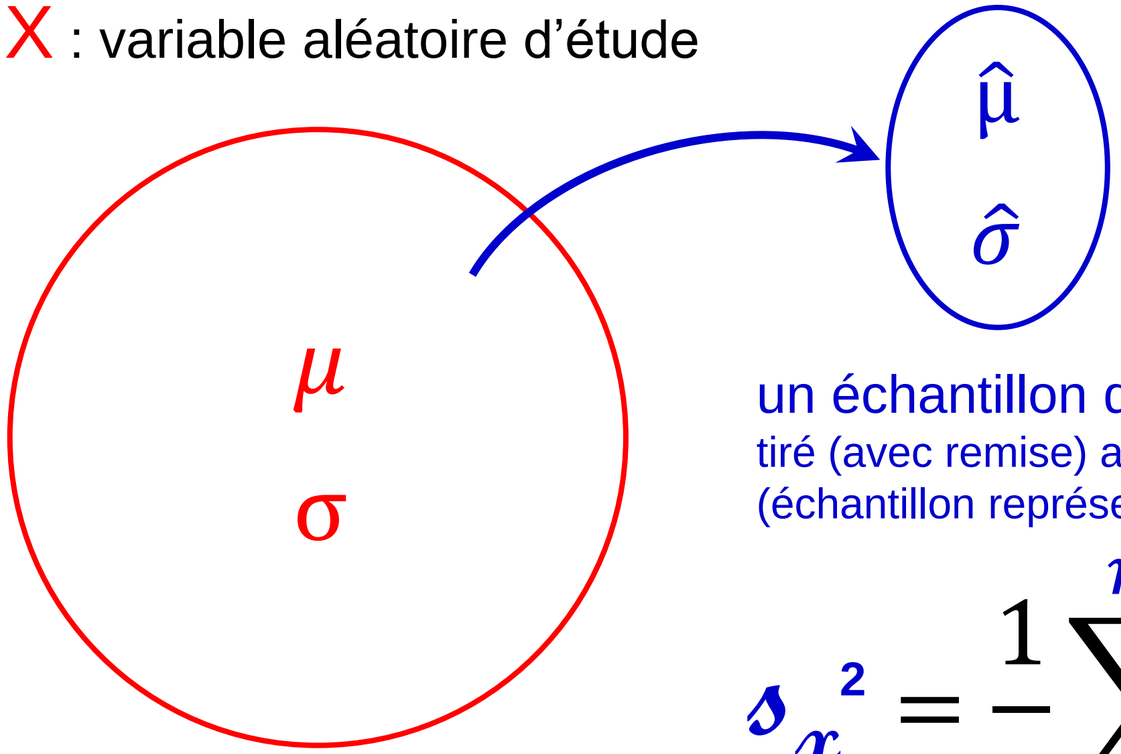
Ecart-type : **σ** ; **Variance** : σ^2

La variance calculée dans un échantillon est une estimation biaisée de la variance exacte de la population

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la variance exacte : variance calculée dans un échantillon

X : variable aléatoire d'étude



$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

estimation ponctuelle de **μ**

(non biaisée)

un échantillon de taille n

tiré (avec remise) au hasard de la population

(échantillon représentatif de la population)

$$s_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \hat{\mu}^2$$

formule de Könnig- Huyguens

Population : caractérisée par

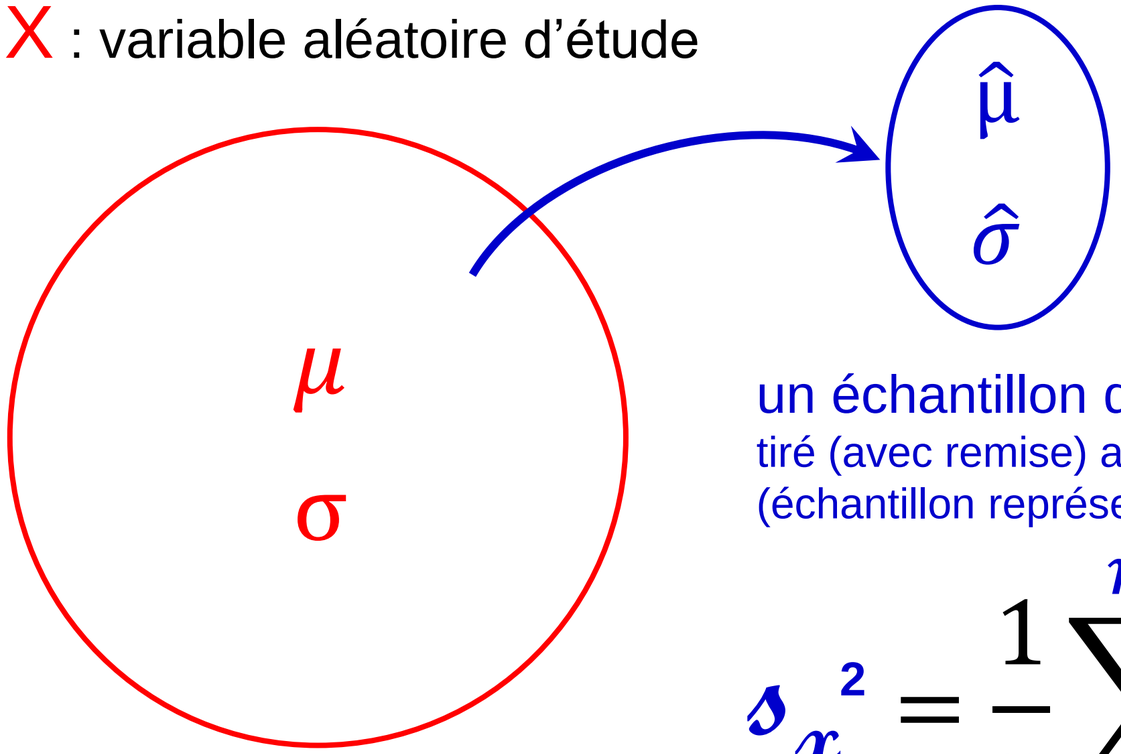
2 paramètres exacts : **μ** et **σ**

La variance calculée dans un échantillon est une estimation biaisée de la variance exacte de la population

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la variance exacte : estimation non biaisée de la variance

X : variable aléatoire d'étude



$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

estimation ponctuelle de **μ** (non biaisée)

un échantillon de taille n
tiré (avec remise) au hasard de la population
(échantillon représentatif de la population)

$$s_x^{\,2} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \hat{\mu}^2$$

formule de Könnig- Huyguens

Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

Espérance : **μ**

Ecart-type : **σ** ; **Variance** : σ^2

$$s^{*}_{x}{}^{2} = \frac{n}{n-1}\, s_x^{\,2}$$

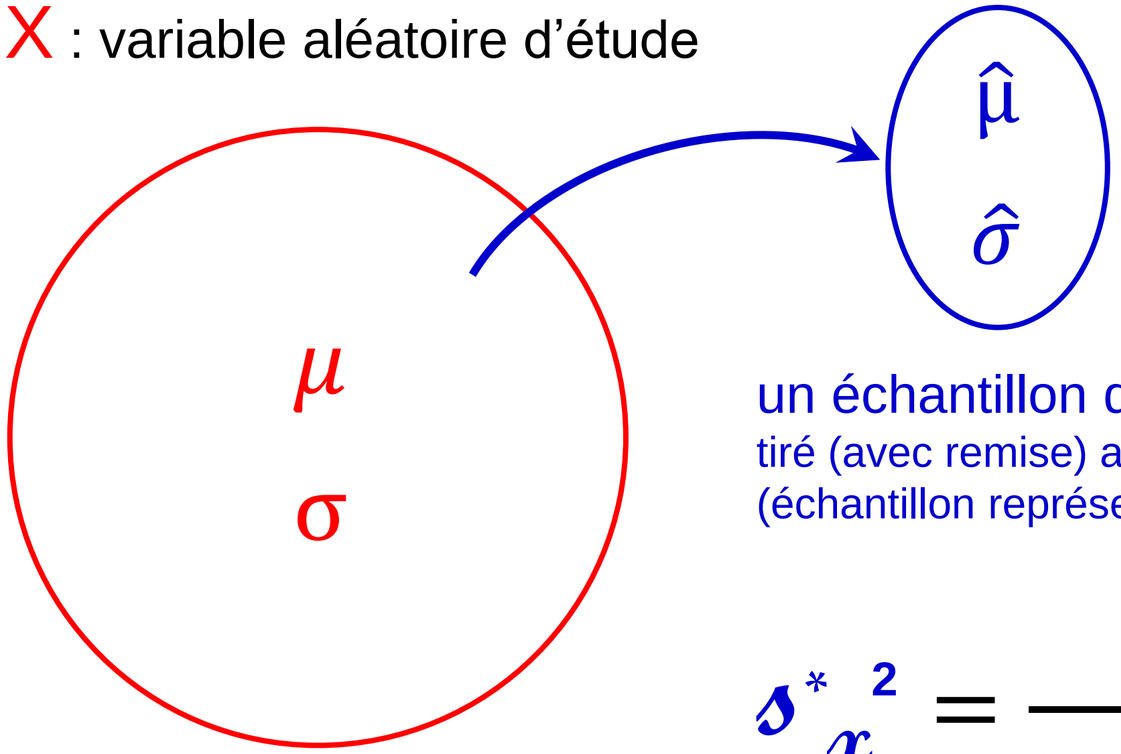
estimation ponctuelle de **σ** corrigée du biais

La variance calculée dans un échantillon est une estimation biaisée de la variance exacte de la population

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la variance exacte : estimation non biaisée de la variance

X : variable aléatoire d'étude



$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

estimation ponctuelle de **μ** (non biaisée)

un échantillon de taille n
tiré (avec remise) au hasard de la population
(échantillon représentatif de la population)

$$s^{*}{}_{x}^{2} = \frac{1}{n-1}\sum_{i=1}^{n} x_i^{2} - \frac{n}{n-1}\hat{\mu}^{2}$$

estimation ponctuelle de **σ** (corrigée du biais)

Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

Espérance : **μ**

Ecart-type : **σ** ; **Variance** : σ^2

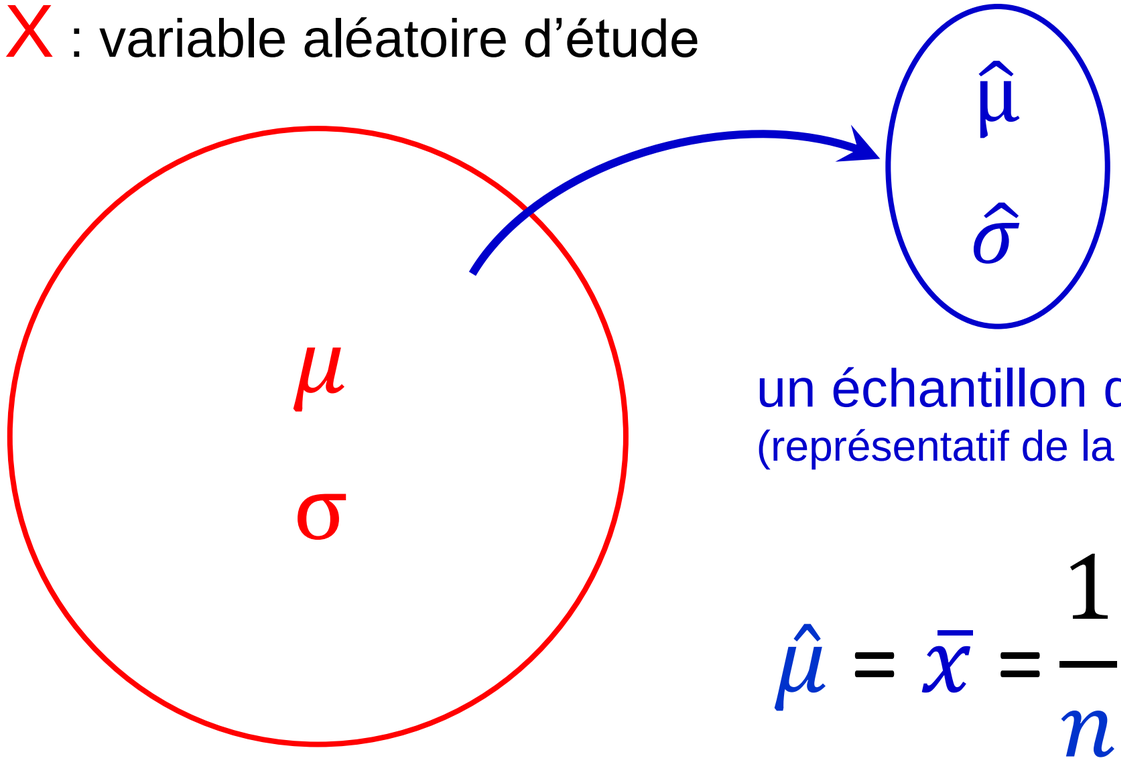
La variance calculée dans un échantillon est une estimation biaisée de la variance exacte de la population

Calcul des estimations des paramètres exacts lorsque les données sont ordonnées-classées (telles que représentées sur un histogramme)

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la moyenne exacte : moyenne calculée à partir de données ordonnées classées

X : variable aléatoire d'étude



un échantillon de taille n
(représentatif de la population)

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j x_{jc}$$

classe (qt/ha)	ni	x_{ic} (qt/ha)	$n_i x_{ic}$ (qt/ha)
]80,85]	0	82,5	0,0
]85,90]	1	87,5	87,5
]90,95]	3	92,5	277,5
]95,100]	5	97,5	487,5
]100,105]	8	102,5	820,0
]105,110]	11	107,5	1182,5
]110,115]	13	112,5	1462,5
]115,120]	13	117,5	1527,5
]120,125]	8	122,5	980,0
]125,130]	2	127,5	255,0
]130,135]	1	132,5	132,5
Sommes	65		7212,5

$\hat{\mu}$ = 7212,5/65 = 111,0 qt/ha

estimation ponctuelle de μ
(moins précise)

Population : caractérisée par
2 paramètres exacts : μ et σ

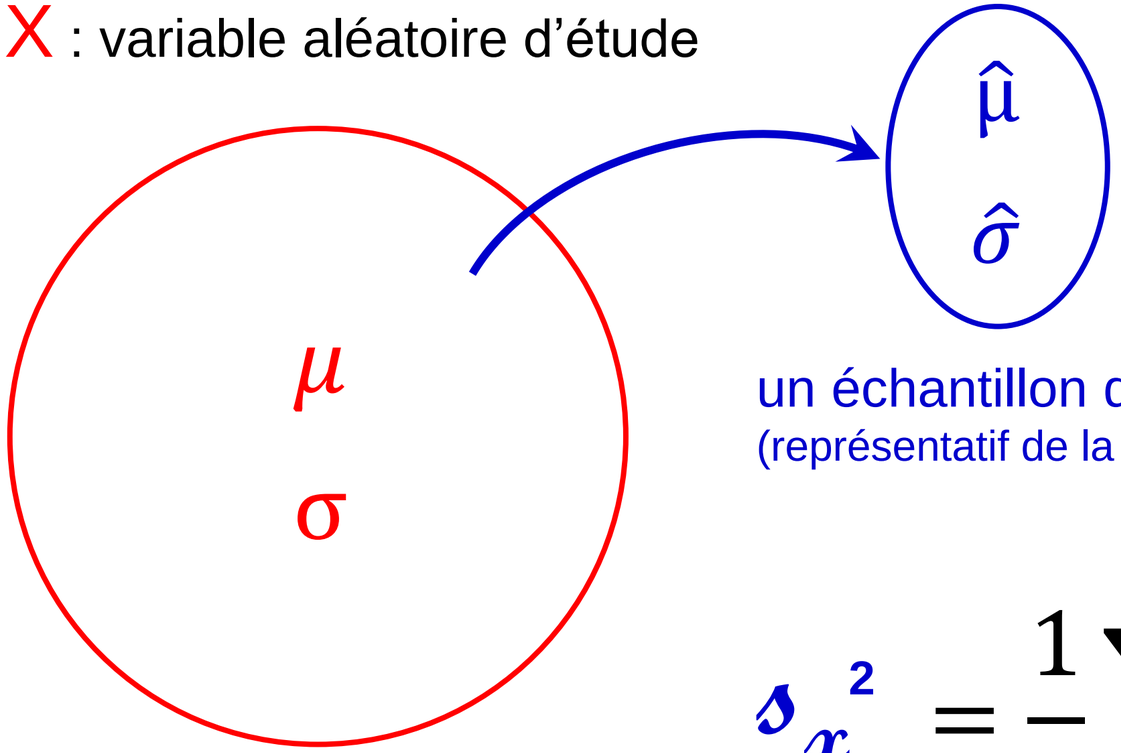
On utilise les **effectifs de classes** n_j et les valeurs centrées x_{jc} de chacune des **k classes** pour calculer une **moyenne approchée**; j est **l'indice de classe**.
On **perd de la précision** sur la valeur ciblée. L'estimation est moins bonne.

La moyenne calculée dans un échantillon n'est qu'une estimation de la moyenne exacte dans la population

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la moyenne exacte : variance calculée à partir de données ordonnées classées

X : variable aléatoire d'étude



un échantillon de taille n
(représentatif de la population)

classe (qt/ha)	ni	x_{ic} (qt/ha)	$n_i x_{ic}$ (qt/ha)
]80,85]	0	82,5	0,0
]85,90]	1	87,5	87,5
]90,95]	3	92,5	277,5
]95,100]	5	97,5	487,5
]100,105]	8	102,5	820,0
]105,110]	11	107,5	1182,5
]110,115]	13	112,5	1462,5
]115,120]	13	117,5	1527,5
]120,125]	8	122,5	980,0
]125,130]	2	127,5	255,0
]130,135]	1	132,5	132,5
Sommes	65		7212,5

$\hat{\mu}$ = 7212,5/65 = 111,0 qt/ha

$$s_x^2 = \frac{1}{n}\sum_{j=1}^{k} n_j (x_{jc} - \hat{\mu})^2$$

Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

estimation ponctuelle (moins précise) de **σ** (biaisée)

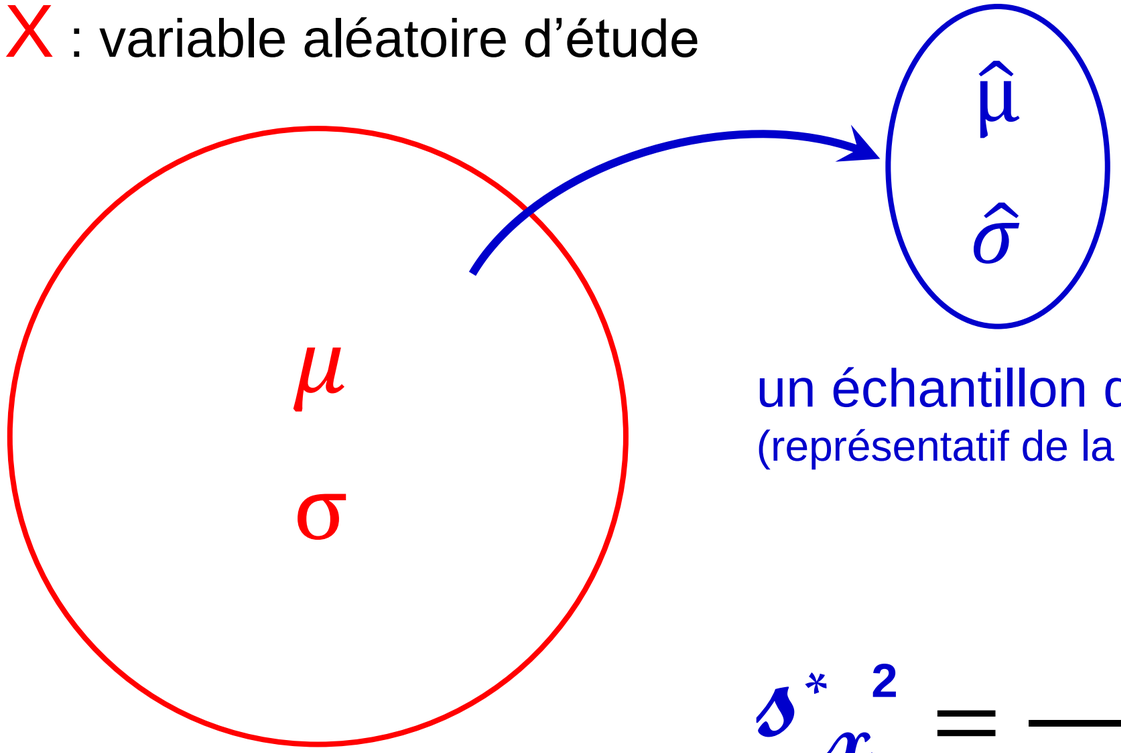
On utilise les **effectifs de classes** n_j et les valeurs centrées x_{jc} de chacune des k classes ; j est l'indice de classe.

La variance calculée dans un échantillon n'est qu'une estimation de la variance exacte dans la population

Calcul des paramètres statistiques et de leurs estimations

▶ estimation de la moyenne exacte : variance calculée à partir de données ordonnées classées

X : variable aléatoire d'étude



un échantillon de taille n
(représentatif de la population)

classe (qt/ha)	ni	x_{ic} (qt/ha)	$n_i x_{ic}$ (qt/ha)
]80,85]	0	82,5	0,0
]85,90]	1	87,5	87,5
]90,95]	3	92,5	277,5
]95,100]	5	97,5	487,5
]100,105]	8	102,5	820,0
]105,110]	11	107,5	1182,5
]110,115]	13	112,5	1462,5
]115,120]	13	117,5	1527,5
]120,125]	8	122,5	980,0
]125,130]	2	127,5	255,0
]130,135]	1	132,5	132,5
Sommes	65		7212,5

$\hat{\mu}$ = 7212,5/65 = 111,0 qt/ha

$$s^{*2}_{x} = \frac{1}{n-1}\sum_{j=1}^{k} n_j x_{jc}^{\,2} - \frac{n}{n-1}\hat{\mu}^2$$

Population : caractérisée par

2 paramètres exacts : **μ** et **σ**

estimation ponctuelle (moins précise) de **σ** (corrigée du biais)

On utilise les **effectifs de classes** n_j et les valeurs centrées x_{jc} de chacune des k classes ; j est l'indice de classe.

La variance calculée dans un échantillon n'est qu'une estimation de la variance exacte dans la population

Calcul des paramètres statistiques et de leurs estimations

Les calculs son réalisés avec un logiciel : EXCEL, R, RCMDR, SPSS, Jamovi, …

De nos jours, on ne calcule pratiquement plus rien à la main!