4.1 DISTRIBUTIONS ÉCHANTILLONALES

cours 22

Aujourd'hui, nous allons voir

- ✓ Distributions échantillonnales

Étant donné une population, on peut s'intéresser à toute sorte de chose la concernant.

Définition

Soit P une population, une **variable statistique** X sur P est une fonction qui attribue à chaque individu un élément d'un ensemble M.

$$X : P \longrightarrow M$$

On discernera deux cas en fonction de l'ensemble d'arrivé de la fonction.

Définition

On dira d'une **variable statistique** qu'elle est **quantitative** si l'ensemble des **valeurs** qu'elle peut prendre est un ensemble de nombres.

Exemple

L'âge, poids, taille, etc.

Définition

On dira d'une **variable statistique** qu'elle est **qualitative** si l'ensemble des **modalités** qu'elle peut prendre n'est pas un ensemble de nombres.

Exemple

Le sexe, la couleur des cheveux, fumeur, etc.

- Variables statistiques
  - Quantitatives
    - Discrètes
    - Continues
  - Qualitatives
    - Nominales
    - Ordinales

Si par exemple, on s'intéresse au poids des chevreuils du Québec, avoir disons 3 000 000 de poids est beaucoup d'informations!

On s'intéresse donc à certaines mesures plus simples pour d'écrire une variable statistique.

Mesures de tendance centrale

Mode: La donnée qui apparait le plus souvent

Médiane: Si les données sont placées en ordre, c'est la donnée qui divise les données en deux.

Moyenne:

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_i$$

Mesures de dispersion

Étendue: $x_{max} - x_{min}$

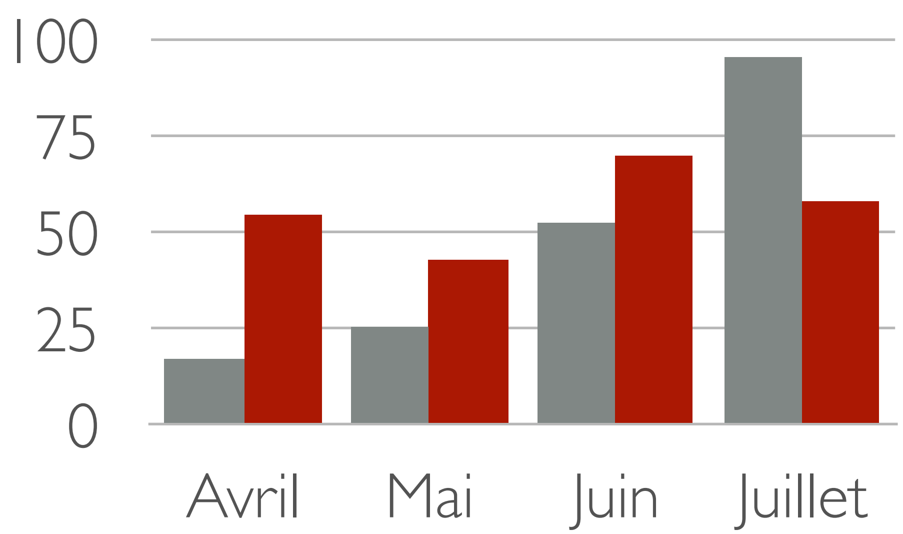
Écart moyen: $$EM = \frac{1}{n}\sum_{k=1}^{n} |x_i - \mu|$$

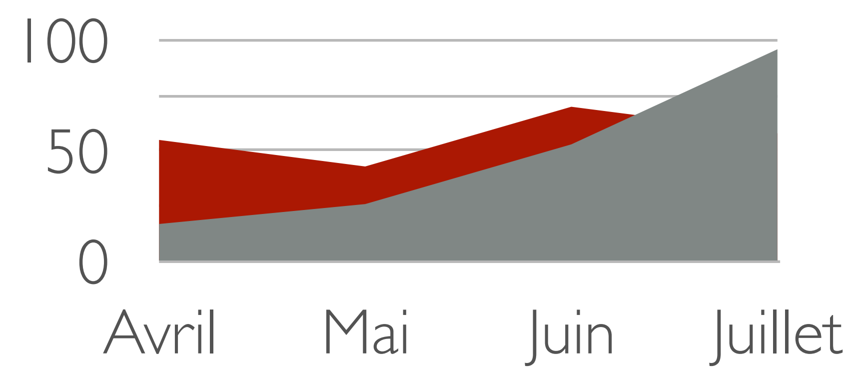
Variance: $$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n} (x_i - \mu)^2$$

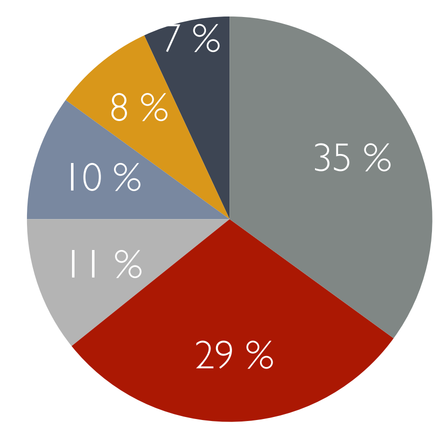
Statistique descriptive.

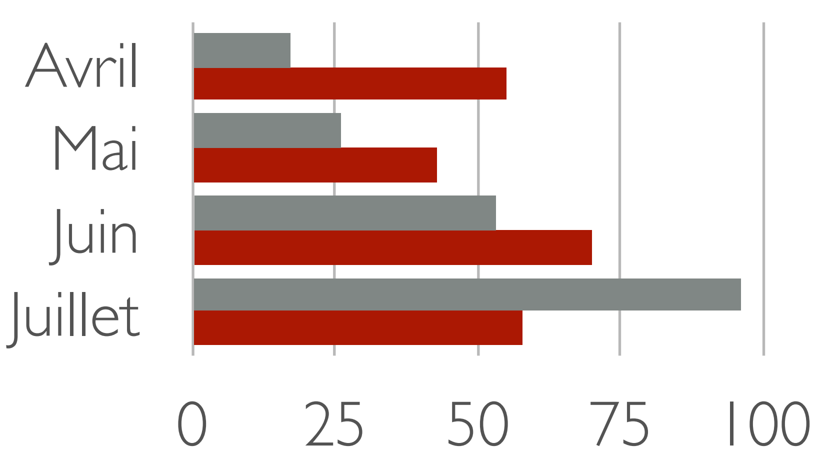
En statistique descriptive, on considère des variables statistiques sur des populations

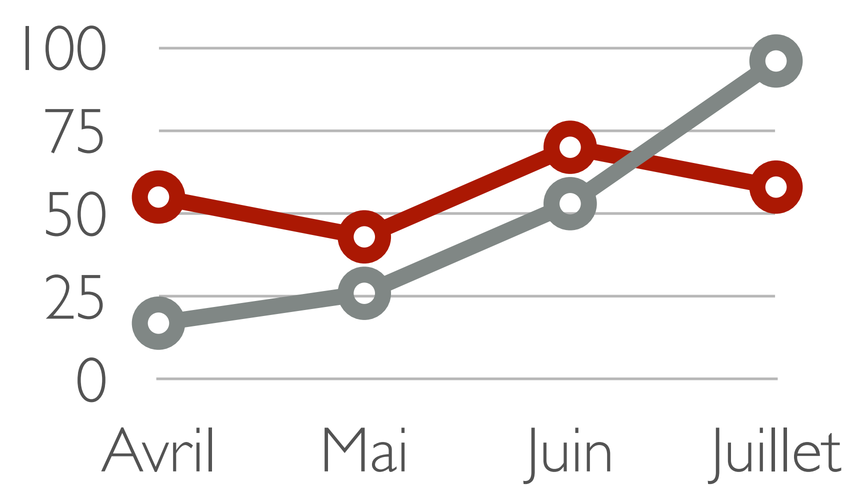
On recueille l'information et on la présente.

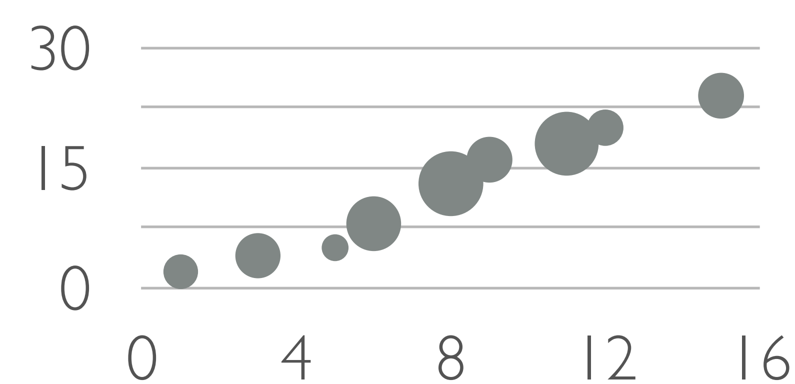












Inférence statistique

Dans les faits, allez chercher l'information pour toute une population est très difficile, voire impossible.

C'est pour cette raison qu'on observe plutôt un échantillon.

Mais comment fait-on pour déterminer si notre échantillon représente bien la population?

Avec les probabilités.

	Population	Échantillon
Taille	N	n
Moyenne	μ	$\bar{x}$
Variance	σ^2	s^2
Écart type	σ	s
Proportion	π	p

Exemple Regardons une population de 5 personnes et considérons la variable statistique: le nombre de cellulaires dans les 5 dernières années.

A	2
B	1
C	0
D	3
E	4

$$\mu = \frac{1}{5}(2+1+0+3+4) = 2$$

$$\sigma^2 = \frac{1}{5}((2-2)^2+(1-2)^2+(0-2)^2+(3-2)^2+(4-2)^2)$$

$$= 2$$

Exemple

Regardons une population de 5 personnes et considérons la variable statistique: le nombre de cellulaires dans les 5 dernières années.

$\mu = 2 \qquad \sigma^2 = 2$

A	2
B	1
C	0
D	3
E	4

Si on prend un échantillon avec remise de taille 2

	$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$
AA	2	BA	1,5	CA	1	DA	2,5	EA	3
AB	1,5	BB	1	CB	0,5	DB	2	EB	2,5
AC	1	BC	0,5	CC	0	DC	1,5	EC	2
AD	2,5	BD	2	CD	1,5	DD	3	ED	3,5
AE	3	BE	2,5	CE	2	DE	3,5	EE	4

Exemple

Regardons une population de 5 personnes et considérons la variable statistique: le nombre de cellulaires dans les 5 dernières années.

A	2
B	1
C	0
D	3
E	4

$$\mu = 2$$

$$\sigma^2 = 2$$

Considérons l'expérience aléatoire de piger un échantillon au hasard

$\bar{X}$: moyenne de l'échantillon

	$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$
AA	2	BA	1,5	CA	1	DA	2,5	EA	3
AB	1,5	BB	1	CB	0,5	DB	2	EB	2,5
AC	1	BC	0,5	CC	0	DC	1,5	EC	2
AD	2,5	BD	2	CD	1,5	DD	3	ED	3,5
AE	3	BE	2,5	CE	2	DE	3,5	EE	4

Exemple Regardons une population de 5 personnes et considérons la variable statistique: le nombre de cellulaires dans les 5 dernières années.

$\bar{X}$: moyenne de l'échantillon

L'ensemble de réalisation est

$$\begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 4 & 3 & 2 & 1 \end{array}$$

$$\{0, 0{,}5, 1, 1{,}5, 2, 2{,}5, 3, 3{,}5, 4\}$$

	$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$		$\bar{x}$
AA	2	BA	1,5	CA	1	DA	2,5	EA	3
AB	1,5	BB	1	CB	0,5	DB	2	EB	2,5
AC	1	BC	0,5	CC	0	DC	1,5	EC	2
AD	2,5	BD	2	CD	1,5	DD	3	ED	3,5
AE	3	BE	2,5	CE	2	DE	3,5	EE	4

Exemple

Regardons une population de 5 personnes et considérons la variable statistique: le nombre de cellulaires dans les 5 dernières années.

$\bar{x}_i$	$f(\bar{x}_i)$
0	1/25
0,5	2/25
1	3/25
1,5	4/25
2	5/25
2,5	4/25
3	3/25
3,5	2/25
4	1/25

1 2 3 4 5 4 3 2 1

{0, 0,5, 1, 1,5, 2, 2,5, 3, 3,5, 4}

$$E(\bar{X}) = \sum_{i=1}^{9} \bar{x}_i f(\bar{x}_i) \quad = 2$$

$$\text{Var}(\bar{X}) = \sum_{i=1}^{9} \bar{x}_i^2 f(\bar{x}_i) - \mu_{\bar{X}}^2 \quad = 1$$

L'exemple précédant illustre les définitions suivantes

Définition

Étant donnée une variable aléatoire quantitative d'une population de taille N. Considérons l'expérience aléatoire qui consiste à piger un échantillon de taille n.

La variable aléatoire

$\bar{X}$: la moyenne de l'échantillon

On nomme la distribution de cette variable aléatoire,
la **distribution échantillonnale de la moyenne** $\bar{X}$

Échantillon avec remise

Étant donnée une population de taille n ainsi qu'une variable statistique X. On aimerait avoir un lien entre la moyenne et l'écart type de X et l'espérance et l'écart type de la variable aléatoire $\bar{X}$.

Soit μ la moyenne de X et σ^2 la variance de X

On peut voir l'expérience aléatoire de piger un échantillon de taille n comme une suite d'expérience consistant à piger un individu.

X_i : la valeur de la variable statistique.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Échantillon avec remise

Soit μ la moyenne de X et σ^2 la variance de X

X_i : la valeur de la variable statistique.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Mais $E(X_i) = \mu$ $\mathrm{Var}(X_i) = \sigma^2$

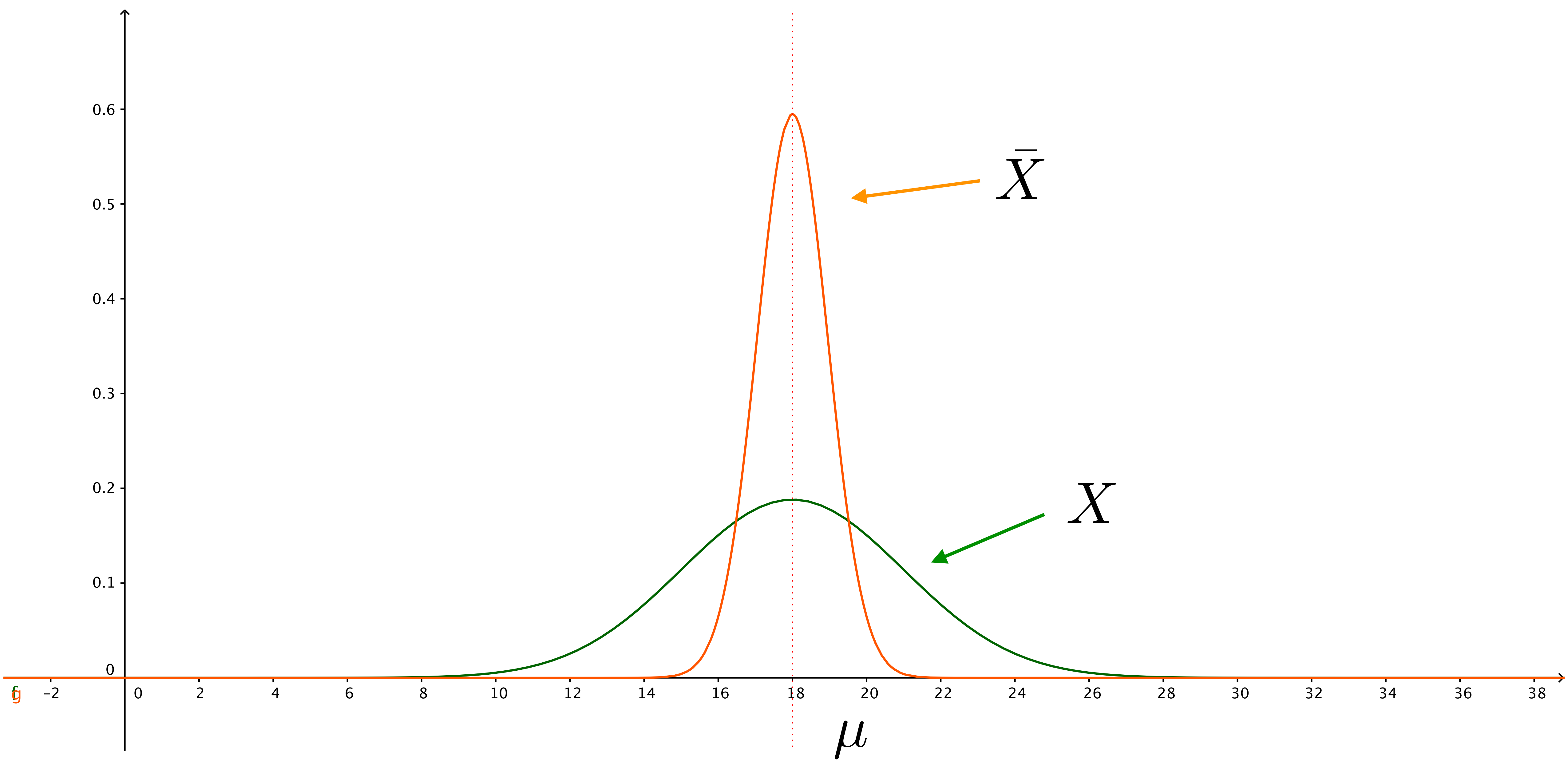
$$
\begin{aligned}
E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n} E(X_1 + X_2 + \cdots + X_n) \\
&= \frac{1}{n}\big(E(X_1) + E(X_2) + \cdots + E(X_n)\big) \\
&= \frac{1}{n}\big(\mu + \mu + \cdots + \mu\big) \\
&= \frac{n\mu}{n} \\
&= \mu
\end{aligned}
$$

$$E(\bar{X}) = \mu$$

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(X_1 + X_2 + \cdots + X_n\right) \\
&= \frac{1}{n^2}\big(\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\big) \\
&= \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right) \\
&= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
\end{aligned}$$

$$E(\bar{X}) = \mu \qquad\qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Ici on peut remarquer que plus la taille de l'échantillon est grande plus la variance de $\bar{X}$ diminue



À priori, on ne connait pas la distribution de la variable statistique X et donc on ne connait pas non plus la distribution de la variable aléatoire $\bar{X}$

Théorème (Théorème central limite)

Soit $X_1, X_2, \ldots, X_n$ des variables aléatoires indépendantes ayant toutes la même distribution de probabilité.

$$S = \sum_{k=1}^{n} X_k \qquad \text{alors quand} \quad n \to \infty$$

$$Y = \frac{S - E(S)}{\sqrt{\mathrm{Var}(S)}} \sim N(0,1)$$

À l'aide du théorème central limite, on peut déduire que si

X a comme moyenne μ et variance σ^2

$$E(\bar{X}) = \mu \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} = \sigma_{\bar{X}}^2 \qquad \sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}}$$

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0, 1)$$

et ce peut importe la distribution de X

Faites les exercices suivants

Faites les exercices suivants

#4.1 à 4.3

Échantillon sans remise

Soit μ la moyenne de X et σ^2 la variance de X

X_i : la valeur de la variable statistique.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Ici puisque les variables aléatoires X_i suivent des lois hypergéométriques

on aura plutôt

$$E(\bar{X}) = \mu \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

Échantillon sans remise

$$E(\bar{X}) = \mu \qquad\qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

On a donc un facteur de correction pour tenir en compte le fait que l'échantillon est fait sans remise

Or, très souvent, la taille de l'échantillon est très petite en comparaison à la taille de la population.

$$N = 10\ 000 \qquad\qquad n = 100$$

$$\frac{N-n}{N-1} = \frac{9\ 900}{9\ 999} = 0,9901 \approx 1$$

La distribution d'échantillonnage suit donc une loi normale

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0,1)$$

Si l'échantillon est avec remise $$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}}$$

Si l'échantillon est sans remise $$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$$

Faites les exercices suivants

#4.4

Aujourd'hui, nous avons vu

La distribution d'échantillonnage suit donc une loi normale

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0, 1)$$

Si l'échantillon est avec remise $\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}}$

Si l'échantillon est sans remise $\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$

Devoir: 4.1 à 4.4