Chapitre 2

Introduction à l'estimation

Université de Paris Ouest

2012–2013

Sommaire

Exemple 1 : Taille des Français

- Population $\mathcal{P} = \{$ Adultes français $\}$
- Taille $N = 45000000$
- Variable : $X =$ "Taille en cm", **quantitative**

Exemple 1 : Taille des Français

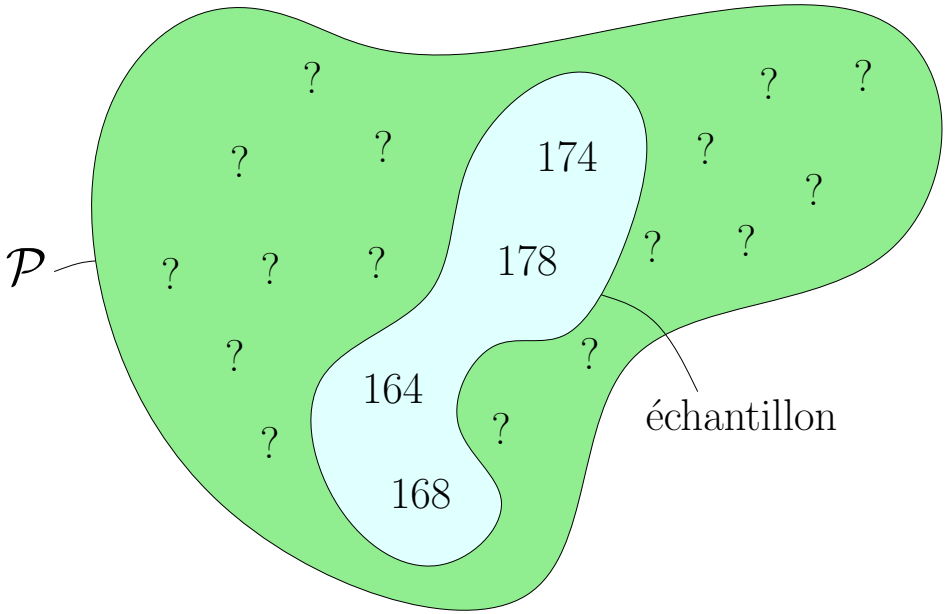
- Population $\mathcal{P} = \{$ Adultes français $\}$
- Taille $N = 45000000$
- Variable : $X =$ "Taille en cm", **quantitative**
- Modalités : intervalle $[0\mathrm{cm}; 300\mathrm{cm}]$
- 2 paramètres : $\mu =$ moyenne, $\sigma^2 =$ variance.

On cherche à connaître μ et σ^2.

Problème : N est trop grand !

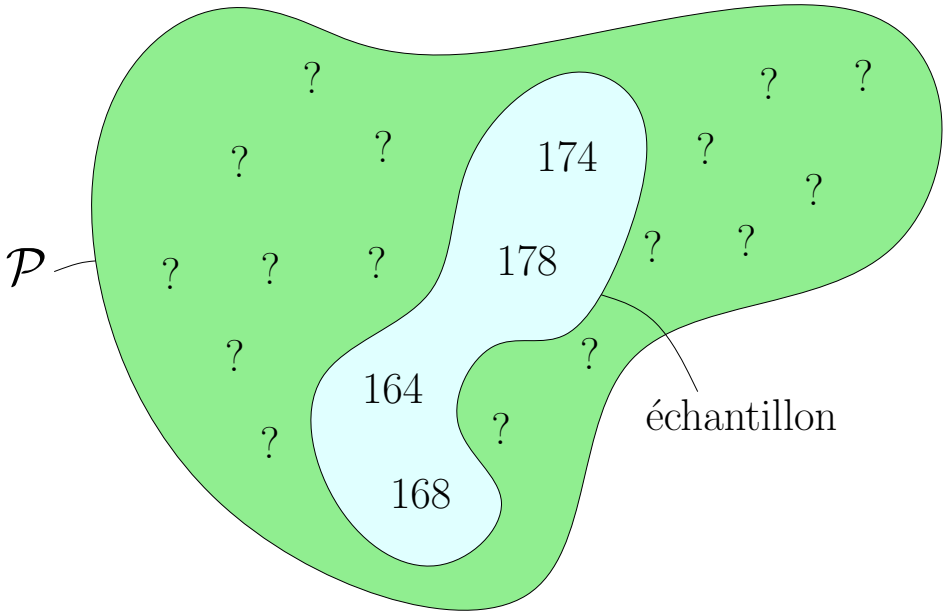
Exemple 1 : Taille des Français

Accès uniquement à un **échantillon** de taille 4 :



Exemple 1 : Taille des Français

Accès uniquement à un **échantillon** de taille 4 :



Dans cet échantillon, $\text{moyenne} = \dfrac{174+164+178+168}{4} = 171.$

On **extrapole** ces données à la population entière :

On ne connaît pas μ, mais on peut penser que μ est proche de 171.

Exemple 2 : Sondage pour un référendum

- Population $\mathcal{P} = \{$ Adultes français $\}$
- Taille $N = 45000000$
- Variable : $X =$ "réponse au référendum", **qualitative**

Exemple 2 : Sondage pour un référendum

- Population $\mathcal{P} = \{$ Adultes français $\}$
- Taille $N = 45000000$
- Variable : $X =$ "réponse au référendum", **qualitative**
- Modalités : oui/non
- 1 paramètre $p =$ proportion de "oui".

On cherche à connaître p.

Problème : N est trop grand !

Exemple 2 : Sondage pour un référendum

Accès uniquement à un **échantillon** de taille 1000 :

- On appelle 1000 adultes au téléphone, 540 disent voter "oui".

On **extrapole** ces données à la population entière :

On ne connaît pas p, mais on peut penser que p est proche de $0,54$.

Sommaire

Statistiques descriptives vs Statistiques inférentielles

Définition (Larousse)

L'**inférence statistique** consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon.

Statistiques descriptives vs Statistiques inférentielles

Définition (Larousse)

L'**inférence statistique** consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon.

Stat. descriptives (L1)	Stat. inférentielles (L2)
▶ petite population	▶ très grande population
▶ toutes les données	▶ données d'un échantillon
▶ on calcule les paramètres	▶ on **extrapole** à partir de l'échantillon

Estimation de μ pour une variable quantitative

Variable quantitative X, 2 paramètres μ, σ^2.
Échantillon de taille n, observations $x_1, x_2, \ldots, x_n$.

Définition

L'**estimation ponctuelle** de la moyenne μ est donnée par la **moyenne observée** dans l'échantillon

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Attention : μ est inconnue, seule $\bar{x}$ est observée !

Retour sur l'Exemple 1 : estimation de la moyenne

- Population $\mathcal{P} = \{$ Français $\}$
- Taille $N = 45000000$
- $\mu =$ moyenne

- Échantillon tiré au sort
- Taille $n = 4$
- $\bar{x} =$ **moyenne observée**$= 171.$

μ est inconnue, mais on **estime** *μ par la* **moyenne observée** $\bar{x} = 171.$

Estimation de σ^2 pour une variable quantitative

Variable quantitative X, 2 paramètres μ, σ^2, observations $x_1, x_2, \ldots, x_n$.
On note s^2 la **variance observée** :

$$s^2 = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2.$$

Estimation de σ^2 pour une variable quantitative

Variable quantitative X, 2 paramètres μ, σ^2, observations $x_1, x_2, \ldots, x_n$.
On note s^2 la **variance observée** :

$$s^2 = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2.$$

Définition

L'estimation ponctuelle de la variance σ^2 est donnée par la variance corrigée dans l'échantillon

$$s^{\star 2} = \frac{n}{n-1} s^2.$$

Attention : σ^2 est inconnue, seule s^2 et $s^{\star 2}$ sont observées !

Estimation de σ^2 pour une variable quantitative

Variable quantitative X, 2 paramètres μ, σ^2, observations $x_1, x_2, \ldots, x_n$.
On note s^2 la **variance observée** :

$$s^2 = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2.$$

Définition

L'estimation ponctuelle de la variance σ^2 est donnée par la variance corrigée dans l'échantillon

$$s^{\star 2} = \frac{n}{n-1} s^2.$$

Attention : σ^2 est inconnue, seule s^2 et $s^{\star 2}$ sont observées !

Définition bis

L'estimation ponctuelle de l'écart-type σ est donnée par l'écart-type corrigé $s^\star = \sqrt{s^{\star 2}}$.

Retour sur l'Exemple 1 : estimation de la variance

- Population $\mathcal{P} = \{$ Français $\}$
- Taille $N = 45000000$
- $\sigma^2 =$ variance

- Échantillon tiré au sort
- Taille $n = 4$
- $s^2 =$ **variance observée**

Dans l'échantillon,

$$\text{variance observée } s^2 = \frac{174^2 + 164^2 + 178^2 + 168^2}{4} - 171^2 = 29,$$

Retour sur l'Exemple 1 : estimation de la variance

- Population $\mathcal{P} = \{$ Français $\}$
- Taille $N = 45000000$
- $\sigma^2 =$ variance

- Échantillon tiré au sort
- Taille $n = 4$
- $s^2 =$ **variance observée**
- $s^{\star 2} =$ **variance corrigée**

Dans l'échantillon,

$$\text{variance observée } s^2 = \frac{174^2 + 164^2 + 178^2 + 168^2}{4} - 171^2 = 29,$$

$$\text{variance corrigée } s^{\star 2} = \frac{n}{n-1} s^2 = \frac{4}{3} \times 29.$$

σ^2 est inconnue, mais on **estime** *σ^2 par la* **variance corrigée** $s^{\star 2} = 38,67$.

Estimation de p pour une variable qualitative

Variable qualitative X à 2 modalités,
Un paramètre p = effectif de la 1ère modalité.
Échantillon de taille n,

n_1 = effectif de la 1ère modalité **dans l'échantillon**.

Estimation de p pour une variable qualitative

Variable qualitative X à 2 modalités,
Un paramètre p = effectif de la 1ère modalité.
Échantillon de taille n,

$$n_1 = \text{effectif de la 1ère modalité } \textbf{dans l'échantillon}.$$

Définition

L'**estimation ponctuelle** de la proportion p est donnée par la **fréquence observée** f de la première modalité dans l'échantillon :

$$f = \frac{n_1}{n}.$$

Attention : p est inconnue, seule f est observée !

Retour sur l'Exemple 2

- Population $\mathcal{P} = \{$ Français $\}$
- Taille $N = 45000000$
- $p =$ proportion de "oui".

- Échantillon tiré au sort
- Taille $n = 1000$
- $f =$ **fréquence observée** de "oui".

p est inconnue, mais on **estime** *p par la* **fréquence observée** $f = 0,54$.

Une notation pratique : $\sum$

x_i = la variable du i-ème individu :

$$x_1 = 174, \quad x_2 = 164, \quad x_3 = 178, \quad x_4 = 168$$

On note alors

$$\sum x_i = x_1 + x_2 + x_3 + x_4 = 174 + 164 + 178 + 168.$$

(se lit "somme des x_i")

Une notation pratique : $\sum$

x_i = la variable du i-ème individu :

$$x_1 = 174, \quad x_2 = 164, \quad x_3 = 178, \quad x_4 = 168$$

On note alors

$$\sum x_i = x_1 + x_2 + x_3 + x_4 = 174 + 164 + 178 + 168.$$

(se lit "somme des x_i")

On peut aussi noter

$$\sum x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = 174^2 + 164^2 + 178^2 + 168^2.$$

(se lit "somme des x_i au carré")

Un exemple d'exercice avec $\sum$

La santé des enfants prématurés est mesurée 5 minutes après la naissance par le **score d'Apgar** (une note entre 0 et 10).
Sur 60 nourrissons on recueille des scores $x_1, \ldots, x_{60}$ tels que

$$\sum x_i = 472, \qquad \sum x_i^2 = 3820.$$

Question : Donner une estimation du score moyen et de la variance du score parmi tous les prématurés.

Un exemple d'exercice avec $\sum$

La santé des enfants prématurés est mesurée 5 minutes après la naissance par le **score d'Apgar** (une note entre 0 et 10).
Sur 60 nourrissons on recueille des scores $x_1, \ldots, x_{60}$ tels que

$$\sum x_i = 472, \qquad \sum x_i^2 = 3820.$$

Question : Donner une estimation du score moyen et de la variance du score parmi tous les prématurés.

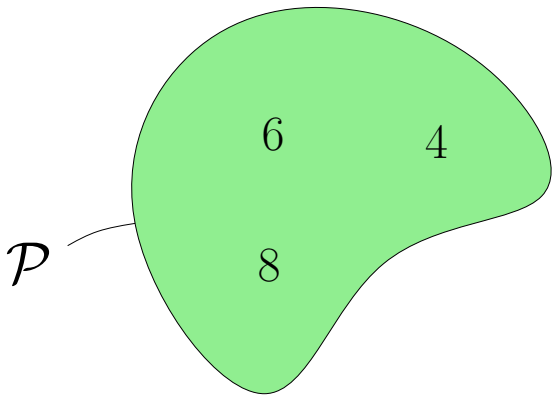
- Variable S =”score”, quantitative discrète. 2 paramètres μ, σ^2.
- Population $\mathcal{P} = \{$ prématurés $\}$, échantillon de taille $n = 60$.

On estime μ par $\bar{x} = \frac{\sum x_i}{60} \approx 7,87$.
On calcule la **variance observée** $s^2 = \left(\frac{\sum x_i^2}{60} - 7,87^2\right) \approx 1,73$.
On estime σ^2 par la **variance corrigée** $s^{\star 2} = \frac{60}{59}\left(\frac{\sum x_i^2}{60} - 7,87^2\right) \approx 1,76$.

Sommaire

Pourquoi $n-1$: un exemple

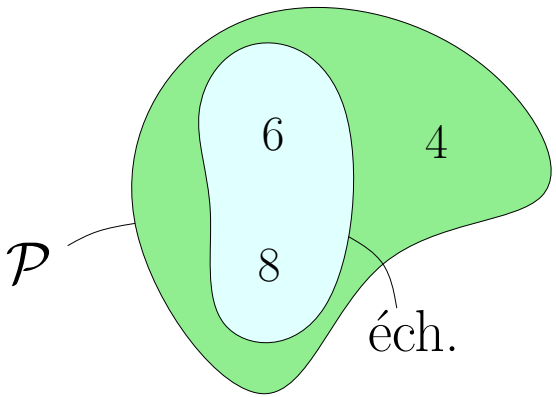
Imaginons une population de taille 3 :



$$\mu = 6$$
$$\sigma^2 = \frac{4^2+6^2+8^2}{3} - 6^2 = 2,66$$

Pourquoi $n-1$: un exemple

Imaginons une population de taille 3 :



$$\mu = 6$$
$$\sigma^2 = \frac{4^2+6^2+8^2}{3} - 6^2 = 2,66$$

Que donnerait une **estimation** à partir d'échantillons de taille 2 ?

Pourquoi $n - 1$: un exemple

Vraies valeurs des paramètres : $\mu = 6,\ \sigma^2 = 2,66$

Considérons toutes les estimations possibles :

Éch.	4; 4	4; 6	4; 8	6; 4	6; 6	6; 8	8; 4	8; 6	8; 8	moyenne
$\bar{x}$	4	5	6	5	6	7	6	7	8	6
s^2	0	1	4	1	0	1	4	1	0	1,33
$s^{\star 2}$	0	2	8	2	0	2	8	2	0	2,66

Pourquoi $n-1$: un exemple

Vraies valeurs des paramètres : $\mu = 6,\ \sigma^2 = 2,66$

Considérons toutes les estimations possibles :

Éch.	4; 4	4; 6	4; 8	6; 4	6; 6	6; 8	8; 4	8; 6	8; 8	moyenne
$\bar{x}$	4	5	6	5	6	7	6	7	8	6
s^2	0	1	4	1	0	1	4	1	0	1,33
$s^{\star 2}$	0	2	8	2	0	2	8	2	0	2,66

On voit que s^2 **sous-estime** la variance.

- Il faut corriger s^2 en $s^{\star 2}$ en multipliant par $\frac{n}{n-1} = \frac{2}{1}$.

Sommaire

Notations à retenir

Population	Échantillon
▶ Taille N (parfois inconnue)	▶ Taille n
▶ moyenne μ	▶ moyenne observée $\bar{x}$
▶ variance σ^2	▶ variance observée s^2
	▶ variance corrigée $s^{\star 2}$
▶ écart-type σ	▶ écart-type corrigé $s^\star$
▶ proportion p	▶ fréquence observée f

Prochains chapitres

- Échantillon suffisamment grand ?
- Choix de l'échantillon ?
- Qualité de l'estimation ?

$\Rightarrow$ Besoin d'un **modèle statistique**