# Chapitre 3

# PRINCIPES DES STATISTIQUES INFERENTIELLES

# Chapitre 3

1. Problématique

2. Objectifs des statistiques inférentielles

**2.1 Estimation ponctuelle**

**2.2 Estimation par intervalles**

**2.3 Tests d'hypothèses statistiques**

3. Echantillonnage

**3.1 Echantillon représentatif**

**3.2 Tirage aléatoire**

Tirages aléatoires simples avec ou sans remise

En théorie

En pratique

**3.3 Echantillon statistique**

# Chapitre 3 *(suite)*

## 4. Estimation ponctuelle des paramètres

### 4.1 Variable qualitative dichotomique

### 4.2 Variable quantitative

Calcul des estimations

**a. données individuelles**

**b. données regroupées**

Estimation sans biais ou corrigée de la variance

Calcul de l'estimation sans biais de la variance

**a. données individuelles**

**b. données regroupées**

Calcul de l'estimation sans biais de l'écart-type

## 5. Justification des statistiques inférentielles

### 5.1 Loi des grands nombres

### 5.2 Interprétation des résultats

**Efficacité d'un traitement des troubles de l'anxiété de l'enfant**

**Intensité de la dépression**

**Durée de chômage**

# 1. Problématique (1)

La population $\mathcal{P}$ ne peut pas être étudiée dans son entier

- soit la population est de très grande taille (onéreux, long de faire une étude sur tous les sujets)
- soit la population ne peut pas être énumérée dans son entier
  - **Exemples :**

    *population française en dehors d'un recensement*

    *population des malades du SIDA*

    *population des SDF*
- soit la population est **virtuelle** (ou hypothétique) : elle est de taille infinie
  - **Exemples : études expérimentales**

    *population des malades qui **seront** traités avec un nouveau traitement dont l'efficacité est étudiée*

# 1. Problématique (2)

- **Exemple : durée de chômage**

  $\mathcal{P}$ = *{ chômeurs français }* ***N*** = *?*

  ***X*** = *"durée de chômage" (en mois)*

  *variable quantitative continue*

- **Exemple : intensité de la dépression de sujets dépressifs**

  $\mathcal{P}$ = *{ sujets dépressifs }* ***N*** = *?*

  ***X*** = *"score de dépression (CES-D)" (en points)*

  *variable quantitative discrète*

- **Exemple : efficacité d'un traitement des troubles de l'anxiété de l'enfant**

  $\mathcal{P}$ = *{ enfants atteints de troubles de l'anxiété, sous traitement }* ***N*** = $\infty$

  ***X*** = *"amélioration clinique" : oui, non*

  *variable qualitative dichotomique*

# 1. Problématique (3)

$\Rightarrow$ on ne peut pas recenser toutes les valeurs de la variable étudiée dans la population

$\Rightarrow$ on ne peut pas calculer la valeur numérique du paramètre d'intérêt $p$, ou $\mu$, ou $\sigma$ dans la population $\mathcal{P}$

$\Rightarrow$ la valeur numérique du paramètre d'intérêt $p$, ou $\mu$, ou $\sigma$ est ***inconnue*** dans la population $\mathcal{P}$

# 2. Objectifs des statistiques inférentielles

**$\mathcal{P}$ population de taille $N$**

**$N$ très grand ou infini, en général *inconnu***

A partir de l'analyse statistique des données d'un *sous-ensemble d'individus* ou **échantillon** de la population, de taille ***n*** avec ***n*** << $N$

⇒**tirer des conclusions concernant la population entière $\mathcal{P}$**

## 2.1 Estimation ponctuelle

- trouver une "bonne" *approximation* (ayant de bonnes propriétés statistiques) par une *valeur numérique unique* de la valeur inconnue du paramètre d'intérêt ***p***, ou $\mu$, ou $\sigma$ dans la population $\mathcal{P}$

  ➔**on "approche" la valeur du paramètre qui reste *inconnue***

# 2.2 Estimation par intervalles

- trouver un intervalle ("fourchette") de valeurs numériques qui contient vraisemblablement (a de grandes chances de contenir)
  la valeur inconnue du paramètre d'intérêt $p$, ou $\mu$, ou $\sigma$ dans $\mathcal{P}$

  → **elle fait intervenir le risque d'erreur que la "fourchette" trouvée ne contienne pas la valeur du paramètre**

  → **par rapport à l'estimation ponctuelle, elle intègre la notion de précision ou de marge d'erreur de l'estimation**

# 2.3 Tests d'hypothèses statistiques

- valider ou rejeter une hypothèse concernant la valeur inconnue du paramètre d'intérêt $p$, ou $\mu$, ou $\sigma$ dans $\mathcal{P}$

  → **il est impossible de répondre avec certitude, mais on peut établir les risques d'erreurs associés aux décisions envisagées (à l'aide du calcul probabiliste)**

# 3. Echantillonnage

## 3.1 Echantillon représentatif

**Dans quelles conditions est-il permis d'extrapoler à la population entière les résultats obtenus (observés) sur l'échantillon ?**

- si l'échantillon est convenablement choisi, il doit refléter assez fidèlement les caractéristiques (paramètres) de la population entière
- sinon, il y a un risque de **biais** (de sélection) c-a-d d'erreurs systématiques

$\Rightarrow$l'échantillon doit être **représentatif** de la population étudiée :

- **de taille $n$ suffisamment "grande"** ($n \geq 30$)
- **obtenu par tirage au sort des individus de la population**

- **Exemples :**

  ***échantillon*** *: groupe de TD d'étudiants en 2$^{d}$ année de licence de psychologie à Nanterre*

  $\boldsymbol{X}$ *= "sexe"*

  $\boldsymbol{X}$ *= "âge"*

  $\boldsymbol{X}$ *= "durée des études"*

  $\mathcal{P}$ *= { français }*

  $\mathcal{P}$ *= { français de 15 à 35 ans }*

  $\mathcal{P}$ *= { étudiants de Nanterre }*

  $\mathcal{P}$ *= { étudiants en psychologie de Nanterre }*

# 3.2 Tirage aléatoire

**$\mathcal{P}$ population de taille $N$**

**Pour assurer de bonnes propriétés aux résultats obtenus sur l'échantillon, deux conditions sont nécessaires :**

- les tirages doivent être équiprobables
  - **la probabilité de tirer au sort chaque individu de $\mathcal{P}$ doit être la même, c'est à dire égale à $1/N$**
- les tirages successifs doivent être indépendants

# Tirages aléatoires simples avec ou sans remise

- **tirage aléatoire simple avec remise (remplacement) d'un individu dans $\mathcal{P}$ :**

  **après avoir été tiré au sort, chaque individu est remis dans la population avant d'effectuer un nouveau tirage**

  $\Rightarrow$un même individu peut-être représenté plusieurs fois dans l'échantillon

  $\Rightarrow$assure l'indépendance entre les ***n*** tirages successifs

- **tirage aléatoire simple sans remise d'un individu dans $\mathcal{P}$ :**

  **après avoir été tiré au sort, l'individu tiré n'est pas remis dans la population**

  $\Rightarrow$un même individu n'est représenté qu'une seule fois dans l'échantillon

  $\Rightarrow$il n'y a pas indépendance des tirages successifs

# En théorie

Les propriétés théoriques seront données pour

- une population de taille infinie
- un échantillon obtenu par tirage aléatoire simple avec remise

# En pratique

- dans une population de grande taille (de l'ordre de milliers) on fait souvent un tirage aléatoire sans remise en appliquant les propriétés obtenues pour les tirages avec remise
  - **le tirage aléatoire sans remise ne modifie pas beaucoup la population initiale et les tirages sont donc quasiment indépendants**
- dans une population de taille plus faible (de l'ordre de centaines), il faut faire un tirage aléatoire avec remise pour assurer l'indépendance des tirages et appliquer les propriétés qui en découlent
- le plus souvent, il n'y a pas de tirage au sort d'où des difficultés d'extrapolation des résultats obtenus sur l'échantillon à la population étudiée

## 3.3 Echantillon statistique

**échantillon statistique de la variable X issu de la population $\mathcal{P}$ de taille $n$**

$n$ individus tirés au sort 1, 2, ... $n$

**échantillon (observé) ou observations** $(x_1, x_2, \ldots x_n)$ : $n$ valeurs de **X** recueillies (**observées**) sur les $n$ individus de l'échantillon

**En pratique :**

échantillon de **X** issu de $\mathcal{P}$ de taille $n$ :

$(x_1, x_2, \ldots x_n)$ sont des valeurs de la variable étudiée **X** recueillies indépendamment sur $n$ individus d'une même population dans des conditions identiques

- **s'il n'a pas été obtenu par tirage au sort, *l'échantillon n'est a priori pas représentatif* de la population $\mathcal{P}$ pour la variable X**

# 3.3 Echantillon statistique (2)

- **Exemple : durée de chômage**

  $\mathcal{P}$ = *{ chômeurs français }* ***N*** = *?*

  ***X*** = *"durée de chômage" (en mois)*
  *variable quantitative continue*

  ➔ *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
  ***n*** = *30*
  *observations* $(x_1, x_2, \ldots x_{30})$
  ( 3, 8, … 6 )

  30 durées observées sur 30 chômeurs

- **Exemple : efficacité d'un traitement des troubles de l'anxiété de l'enfant**

  $\mathcal{P}$ = *{ enfants atteints de troubles de l'anxiété, sous traitement }* ***N*** = $\infty$

  ***X*** = *"amélioration clinique" : oui, non*
  *variable qualitative dichotomique*

  ➔ *échantillon de* ***X*** *issu de* $\mathcal{P}$ *de taille*
  ***n*** = *63*
  *observations* ( $x_1$ , $x_2$ , $x_3$ , … $x_{63}$ )
  (oui, oui, non, … oui)

  63 valeurs "*oui*" ou "*non*" observées sur 63 enfants atteints de troubles de l'anxiété, sous traitement

# 4. Estimation ponctuelle des paramètres

On approche la valeur du paramètre étudié par une *valeur numérique unique*

➔on **estime** le paramètre d'intérêt

ATTENTION

➔la valeur de l'estimation n'est pas égale à celle du paramètre estimé

# 4.1 Variable qualitative dichotomique

**X** variable qualitative dichotomique définie sur **E = {oui, non}**

$p$ = proportion de "oui", inconnue dans $\mathcal{P}$

échantillon $(x_1, x_2, \ldots x_n)$ de **X** issu de $\mathcal{P}$ de taille $n$

$\Rightarrow$l'**estimation ponctuelle** de la proportion $p$ est donnée par la **fréquence observée** sur l'échantillon

- fréquence (proportion) observée de "oui" notée $f$

$$f = \frac{\text{effectif observé de "oui"}}{n} = \frac{n_1}{n}$$

- l'estimation ponctuelle de la proportion de "non" $1-p$ est donnée par la **fréquence observée** de "non" sur l'échantillon notée $1-f$

$$1-f = \frac{\text{effectif observé de "non"}}{n} = \frac{n_2}{n}$$

➔ **il y a très peu de chance pour que $f = p$ en général $f \neq p$**

# 4.2 Variable quantitative

**X** variable quantitative définie sur **E**

$\mu$ = **moyenne de X,** inconnue dans $\mathcal{P}$

$\sigma^2$ = **variance de X**, inconnue dans $\mathcal{P}$

$\sigma$ = **écart-type de X**, inconnu dans $\mathcal{P}$

**échantillon** $(x_1, x_2, \ldots x_n)$ **de X issu de** $\mathcal{P}$ **de taille** $n$

$\Rightarrow$l'**estimation ponctuelle** de la moyenne $\mu$ de **X** dans $\mathcal{P}$ est donnée par la **moyenne observée** sur l'échantillon notée $\bar{x}$

$\Rightarrow$une **estimation ponctuelle** de la variance $\sigma^2$ de **X** dans $\mathcal{P}$ donnée par la **variance observée** sur l'échantillon notée $s^2$

$\Rightarrow$une **estimation ponctuelle** de l'écart-type $\sigma$ de **X** dans $\mathcal{P}$ donnée par l'**écart-type observé** sur l'échantillon noté $s$

➔ **il y a très peu de chance pour que** $\bar{x} = \mu$ **ou** $s^2 = \sigma^2$ **(ou** $s = \sigma$**)**
**en général** $\bar{x} \neq \mu$ **et** $s^2 \neq \sigma^2$ **(et** $s \neq \sigma$**)**

# Calculs des estimations (1)

## a. données individuelles (*n* petit)

➔ observation (valeur observée) $x_i$ pour chaque individu ***i*** de l'échantillon, pour $i = 1,\ldots n$

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \dfrac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$

  avec $s^2 \geq 0$
- $s = \sqrt{s^2}$ avec $s \geq 0$

# Calculs des estimations (2)

## b. données regroupées (*n* grand)

➔ observation (valeur observée) $x_i$ pour $n_i$ individus de l'échantillon

➔ effectifs observés ($n_i$ , $i = 1,\ldots k$)

- $\overline{x} = \dfrac{\sum_{i=1}^{k} n_i x_i}{n}$

- $s^2 = \dfrac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^2}{n} = \dfrac{\sum_{i=1}^{k} n_i x_i^{\,2}}{n} - \overline{x}^2$

  avec $s^2 \geq 0$

- $s = \sqrt{s^2}$ avec $s \geq 0$

# Estimation sans biais ou corrigée de la variance

- la variance observée $s^2$ est toujours trop faible (**biaisée**) : elle **sous-estime** systématiquement la variance $\sigma^2$ de $X$ dans $\mathcal{P}$
- l'écart-type observé $s$ est toujours trop faible : sous-estimation systématique de l'écart-type $\sigma$ de $X$ dans $\mathcal{P}$

$\Rightarrow$on donne une *meilleure estimation* de la variance (et de l'écart-type) en calculant son **estimation sans biais**

$\Rightarrow$l'estimation ponctuelle de la variance $\sigma^2$ de $X$ dans $\mathcal{P}$ est donnée par la **variance observée sans biais** (ou **corrigée**) sur l'échantillon notée $s^{2*}$

$\Rightarrow$l'estimation ponctuelle de l'écart-type $\sigma$ de $X$ dans $\mathcal{P}$ : **écart-type observé sans biais** (ou **corrigé**) sur l'échantillon noté $s^*$

# Calcul de l'estimation sans biais (corrigée) de la variance (1)

## a. données individuelles (*n* petit)

➔observation $x_i$ pour chaque individu $i$ de l'échantillon, pour $i = 1,\ldots n$

- $$s^{2*} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}^2}{n-1}$$

avec $s^{2*} \geq 0$

## b. données regroupées (*n* grand)

➔observation $x_i$ pour $n_i$ individus de l'échantillon, pour $i = 1,\ldots k$

- $$s^{2*} = \frac{\sum_{i=1}^{k} n_i(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{k} n_i x_i^{\,2} - n\bar{x}^2}{n-1}$$

avec $s^{2*} \geq 0$

# Calcul de l'estimation sans biais (corrigée) de la variance (2)

## à partir de l'estimation biaisée

➔ $s^2$ estimation biaisée (trop faible) de $\sigma^2$

- $s^{2*} = \frac{n}{n-1} s^2$ avec $s^{2*} \geq s^2 \geq 0$

# Calcul de l'estimation sans biais (corrigée) de l'écart-type

- **à partir de l'estimation *sans biais* de la variance**

  → $s^{2*}$ estimation sans biais de $\sigma^2$

  - $s^* = \sqrt{s^{2*}}$ avec $s^* \geq 0$

- **à partir de l'estimation *biaisée* de la variance**

  → $s$ estimation biaisée (trop faible) de $\sigma$

  - $s^* = \sqrt{\frac{n}{n-1}}\, s$ avec $s^* \geq s \geq 0$

# 5. Justification des statistiques inférentielles

Si l'échantillon est convenablement choisi (échantillon représentatif), il doit refléter assez fidèlement les caractéristiques de la population entière

## 5.1 Loi des grands nombres

➔phénomène de régularité statistique

❶ la fréquence observée sur un très grand nombre d'expériences tend vers la proportion $p$ dans la population $\mathcal{P}$

❷ la moyenne observée sur un très grand nombre d'expériences tend vers la moyenne $\mu$ dans la population $\mathcal{P}$

$\Rightarrow$les inférences (conclusions) que l'on peut faire à partir d'un échantillon (représentatif) ne dépendent pas de la taille de la population $N$, mais de la taille de l'échantillon $n$

# 5.2 Interprétation des résultats

**Il est possible de tirer des conclusions, c'est à dire d'extrapoler les résultats observés sur l'échantillon :**

- à la population étudiée si l'échantillon est représentatif de cette population
  - **Exemples :**

    *population française*

    *population des chômeurs français*

    *population des sujets dépressifs*
- à la population virtuelle (hypothétique) créée à l'image de l'échantillon, c'est à dire dont l'échantillon étudié est supposé être représentatif
  - **Exemples : études expérimentales**

    *population des malades ayant les mêmes caractéristiques que ceux de l'échantillon (âge, gravité de la maladie, ...) qui **seront** traités avec le nouveau traitement dont l'efficacité a été étudiée*

# Efficacité d'un traitement des troubles de l'anxiété

$\mathcal{P}$ = {enfants atteints de troubles de l'anxiété, sous traitement} $N = ?$

X = "amélioration clinique": oui, non

⇨ X variable qualitative dichotomique

⇨ un paramètre :

proportion d'amélioration clinique = $\boldsymbol{p}$ inconnue dans $\mathcal{P}$

échantillon de X issu de $\mathcal{P}$ de taille $\boldsymbol{n} = 63$

effectifs observés sur l'échantillon

amélioration : "oui" $n_1 = 48$

"non" $n_2 = 15$

**estimation ponctuelle du paramètre proportion**

la fréquence d'amélioration observée $\boldsymbol{f} = 48/63 = 0{,}762$

➔ la proportion $\boldsymbol{p}$ d'amélioration clinique sous traitement dans la population des enfants atteints de troubles de l'anxiété, est estimée à 76,2%

# Intensité de la dépression (1)

$\mathcal{P}$ = {sujets dépressifs} $N$ = ?

X = "score de dépression (échelle CES-D)" (en points)

⇨ X variable quantitative (discrète) sur E = {0, 1, ... 60}

⇨ deux paramètres :

- score moyen = $\mu$ inconnu dans $\mathcal{P}$
- écart-type du score = $\sigma$ inconnu dans $\mathcal{P}$

échantillon de X issu de $\mathcal{P}$ de taille $\boldsymbol{n}$ = 36

score $x_i$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 19 | 32 | 30 | 21 | 45 | 48 | 52 | 33 | 27 |
| 27 | 19 | 19 | 24 | 49 | 31 | 41 | 38 | 34 |
| 33 | 32 | 17 | 46 | 21 | 30 | 28 | 27 | 44 |
| 18 | 24 | 25 | 31 | 20 | 32 | 31 | 22 | 18 |

$\Sigma\, x_i = 1\,088$ $\Sigma\, x_i^2 = 36\,234$

# Intensité de la dépression (2)

**estimation ponctuelle des paramètres**

**moyenne, variance et écart-type**

**données individuelles**

le score moyen observé $\bar{x} = 30{,}2$ (points)

la variance observée du score $s^2 = 93{,}1$

l'écart-type observé du score

$$s \approx \sqrt{93{,}1} \approx 9{,}6 \text{ (points)}$$

➔ le score de dépression moyen $\mu$ dans la population des sujets dépressifs est estimé à 30,2 (points)

➔ la variance du score de dépression $\sigma^2$ dans la population des sujets dépressifs est estimée à 93,1

➔ l'écart-type du score de dépression $\sigma$ dans la population des sujets dépressifs est estimé à 9,6 (points)

# Intensité de la dépression (3)

**estimation ponctuelle sans biais des paramètres variance et écart-type**

**données individuelles**

la variance observée sans biais du score $s^{2*} = 95,8$ ($s^{2*} \geq s^2 = 93,1$)

l'écart-type observé sans biais du score $s^* = 9,8$ (points) ($s^* \geq s = 9,6$)

➔ la variance du score de dépression $\sigma^2$ dans la population des sujets dépressifs est estimée à 95,8

➔ l'écart-type du score de dépression $\sigma$ dans la population des sujets dépressifs est estimé à 9,8 (points)

# Durée de chômage (1)

$\mathcal{P}$ = {chômeurs français} $N$ = ?

X = "durée de chômage" (en mois)

⇨ X variable quantitative (continue) sur E = (0 ; 120)

⇨ deux paramètres :

- durée moyenne = $\mu$ inconnue dans $\mathcal{P}$
- écart-type de la durée = $\sigma$ inconnu dans $\mathcal{P}$

échantillon de X issu de $\mathcal{P}$ de taille $\boldsymbol{n}$ = 30

| durée $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| effectif observé $n_i$ | 2 | 3 | 2 | 4 | 4 | 3 | 6 | 3 | 3 |

$\Sigma\, n_i x_i = 161$ $\Sigma\, n_i x_i^2 = 1\,033$

# Durée de chômage (2)

**estimation ponctuelle des paramètres**

**moyenne, variance et écart-type**

**données regroupées**

la durée moyenne observée $\bar{x} = 5{,}37$ mois

la variance observée de la durée $s^2 = 5{,}63$

l'écart-type observé de la durée

$$s \approx \sqrt{5{,}63} = 2{,}37 \approx 2{,}4 \text{ mois}$$

- la durée moyenne de chômage $\mu$ dans la population des chômeurs est estimée à 5,37 mois
- la variance de la durée de chômage $\sigma^2$ dans la population des chômeurs est estimée à 5,63
- l'écart-type de la durée de chômage $\sigma$ dans la population des chômeurs est estimé à 2,37 mois

# Durée de chômage (3)

**estimation ponctuelle sans biais des paramètres variance et écart-type**

**données regroupées**

la variance observée sans biais de la durée

$s^{2*} = 5{,}83$ ($s^{2*} \geq s^2 = 5{,}63$)

l'écart-type observé sans biais de la durée

$s^* = 2{,}41$ ($s^* \geq s = 2{,}37$)

➔ la variance de la durée de chômage $\sigma^2$ dans la population des chômeurs est estimée à 5,83

➔ l'écart-type de la durée de chômage $\sigma$ dans la population des chômeurs est estimé à 2,41 mois