Chapitre 1

# Notations et notions utilisées

## 1.1 Bases de la Statistique

### 1.1.1 Définitions

- **Population et individus :** Une population est un ensemble fini d'objets sur lesquels une étude se porte. Ces objets sont appelés individus.

- **Caractère :** Un caractère est une qualité que l'on étudie chez des individus.

- **Échantillon :** Un échantillon est un ensemble d'individus issus d'une population.

- **Données :** Les données en notre possession sont les observations d'un caractère sur les individus d'un échantillon.

- **Estimation paramétrique :** L'enjeu de l'estimation paramétrique est d'évaluer/estimer avec précision un paramètre inconnu émanant d'un caractère à partir des données.

- **Moyenne et écart-type corrigé :** La moyenne et l'écart-type corrigé des données sont les principales mesures statistiques intervenant en estimation paramétrique.

En notant $X$ un caractère numérique, $n$ le nombre d'individus d'un échantillon et $x_1, \ldots, x_n$ les données associées, on définit :

– La moyenne de $x_1, \ldots, x_n$ :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

C'est une estimation ponctuelle de la valeur moyenne de $X$.

– L'écart-type corrigé de $x_1, \ldots, x_n$ :

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

C'est une estimation ponctuelle de la variabilité de $X$ autour de sa moyenne. La valeur obtenue a la même unité que $X$.

### 1.1.2 Exemple

◦ **Contexte :**

| | |
|---|---|
| Population | Ensemble des kiwis d'une ferme |
| Individu | Kiwi |
| Caractère | Poids d'un kiwi (en grammes) |
| Paramètre inconnu | Poids moyen d'un kiwi |
| Échantillon | 6 kiwis choisis au hasard ($n = 6$) |
| Données | $x_1$ \| $x_2$ \| $x_3$ \| $x_4$ \| $x_5$ \| $x_6$<br>72 \| 75 \| 78 \| 71 \| 71 \| 75<br>(par exemple, $x_1$ est le poids du premier kiwi de l'échantillon, soit 72 grammes) |
| Objectif | Évaluer le poids moyen inconnu d'un kiwi à l'aide des données $x_1, \ldots, x_6$ |

◦ **Mesures statistiques :**

| | |
|---|---|
| Moyenne | $\overline{x} = \frac{1}{6} \sum_{i=1}^{6} x_i = 73.6666$ |
| Écart-type corrigé | $s = \sqrt{\frac{1}{6-1} \sum_{i=1}^{6} (x_i - \overline{x})^2} = 2.8047$ |

## 1.2 Éléments de probabilités

### 1.2.1 Notations

- $\mathbb{P}(A)$ = probabilité que $A$ se réalise.
- *var* = variable aléatoire réelle.
- $X \sim \ldots$ = la *var* $X$ suit la loi $\ldots$
- $\mathbb{E}(X)$ = espérance de $X$.
- $\mathbb{V}(X)$ = variance de $X$.
- $\sigma(X) = \sqrt{\mathbb{V}(X)}$ = écart-type de $X$.
- $\mathcal{N}(\mu, \sigma^2)$ = loi normale de moyenne $\mu$ et de variance $\sigma^2$ (écart-type $\sigma$).
- $\mathcal{B}(p)$ = loi de Bernoulli de paramètre $p$.
- $\mathcal{T}(\nu)$ = loi de Student à $\nu$ degrés de liberté.
  Si $\nu \geq 31$, on a l'approximation $T(\nu) \approx \mathcal{N}(0, 1)$.
- $\chi^2(\nu)$ = loi du Chi-deux à $\nu$ degrés de liberté.
- $\mathcal{F}(\nu_1, \nu_2)$ = loi de Fisher à $(\nu_1, \nu_2)$ degrés de liberté.

### 1.2.2 Modélisations

- **Loi normale :** On adopte la modélisation

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

  quand $X$ représente une grandeur sujette à une somme d'erreurs mineures indépendantes.
  Par exemple, $X$ peut être : poids, taille, temps, distance, masse, vitesse, température, indice, score, salaire, note, quantité ou teneur.
  Dans ce cas, $\mu$ est la moyenne de $X$ et $\sigma^2$ (ou $\sigma$) mesure la variabilité/dispersion de $X$ autour de $\mu$.

- **Loi de Bernoulli :** On adopte la modélisation

$$X \sim \mathcal{B}(p)$$

quand $X$ prend deux valeurs : 0 ou 1, correspondant souvent à un codage binaire.
Par exemple, $X = 1$ peut caractériser :
– le succès à une épreuve,
– la présence d'un élément caractéristique.
Le paramètre $p$ est la probabilité que $X = 1$ se réalise, laquelle peut aussi s'interpréter en terme de proportion d'individus dans la population vérifiant $X = 1$.

- **Exemple :**

| Population | Ensemble des plaquettes de beurre d'une production |
|---|---|
| Individu | Plaquette de beurre |
| Caractère 1 | $X$ = Poids d'une plaquette (en grammes) |
| Modélisation | $X \sim \mathcal{N}(\mu, \sigma^2)$ |
| Paramètres : $\mu$ et $\sigma^2$ | $\mu$ = Poids moyen d'une plaquette<br>$\sigma^2$ mesure la dispersion du poids d'une plaquette autour de $\mu$ |
| Caractère 2 | $Y = 1$ si la plaquette présente un défaut de conditionnement et $Y = 0$ sinon |
| Modélisation | $Y \sim \mathcal{B}(p)$ |
| Paramètre $p$ | $p$ = Proportion de plaquettes présentant un défaut de conditionnement |

Chapitre 2

# Intervalles de confiance

## 2.1 Méthodes

### 2.1.1 Bases des intervalles de confiance

- **Objectif :** On veut évaluer avec précision un paramètre inconnu émanant d'un caractère à partir des données. On note $\theta$ ce paramètre.

- **Intervalle de confiance ; première approche :** Un intervalle de confiance pour $\theta$ est une "fourchette de valeurs", déterminée à partir des données, qui a de fortes chances de contenir $\theta$.

- **Niveau :** Le niveau est le pourcentage de chances que l'intervalle de confiance contienne $\theta$.
  On veut que ce niveau soit aussi élevé que possible.
  Il s'écrit sous la forme : $100(1-\alpha)\%$, avec $\alpha \in ]0,1[$ (par exemple, 95%, soit $\alpha = 0.05$, ou 99%, soit $\alpha = 0.01$).

- **Intervalle de confiance ; bilan :** Un intervalle de confiance pour $\theta$ au niveau $100(1-\alpha)\%$, $\alpha \in ]0,1[$, est un intervalle :

  $$i_\theta = [a,\ b],$$

  où $a$ et $b$ sont des réels calculés à partir des données de sorte qu'il y ait $100(1-\alpha)\%$ de chances que $i_\theta$ contienne $\theta$.

### 2.1.2 Intervalles de confiance pour une moyenne : Z-IntConf, T-IntConf et Z-IntConf-Lim

#### Contexte

On étudie un caractère représenté par une *var* $X \sim \mathcal{N}(\mu, \sigma^2)$.

Les données sont constituées de la valeur de $X$ pour chacun des $n$ individus d'un échantillon. Ces valeurs sont notées $x_1, \ldots, x_n$.

Le paramètre $\mu$, représentant la moyenne de $X$, est inconnu.

#### Objectif

On veut déterminer un intervalle de confiance pour $\mu$ au niveau $100(1-\alpha)\%$, $\alpha \in ]0,1[$.

On distingue alors 2 cas :

- le cas où $\sigma$ est connu,
- le cas où $\sigma$ est inconnu.

La connaissance ou non de $\sigma$ va impacter sur la construction théorique de l'intervalle de confiance.

#### Cas où $\sigma$ est connu : Z-IntConf

Dans le cas où $\sigma$ est connu, on utilise le Z-IntConf.

Sa construction repose sur la loi normale centrée réduite $\mathcal{N}(0,1)$ (symbolisée par le "Z" dans "Z-IntConf").

Pour obtenir le Z-IntConf, il faut calculer :

- $\overline{x}$ : la moyenne de $x_1, \ldots, x_n$,
- le réel $z_\alpha$ vérifiant :

$$\mathbb{P}(|Z| \geq z_\alpha) = \alpha,$$

où $Z \sim \mathcal{N}(0,1)$.

Ce réel est évaluable dans la table 1 page 223.

Le Z-IntConf est alors donné par :

$$i_\mu = \left[\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right].$$

Il y a donc $100(1-\alpha)\%$ de chances que $i_\mu$ contienne $\mu$.

**Cas où $\sigma$ est inconnu : T-IntConf**

Dans la pratique, le $\sigma$ est souvent inconnu. On utilise alors le T-IntConf.

Sa construction repose sur la loi de Student $\mathcal{T}(\nu)$ (symbolisée par le "T" dans "T-IntConf").

Pour obtenir le T-IntConf, il faut calculer :
- $\overline{x}$ : la moyenne de $x_1, \ldots, x_n$,
- $s$ : l'écart-type corrigé de $x_1, \ldots, x_n$,
- le réel $t_\alpha(\nu)$ vérifiant :

$$\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha,$$

où $T \sim \mathcal{T}(\nu)$, $\nu = n - 1$.

Ce réel est évaluable dans la table 3 page 225.

Le T-IntConf est alors donné par :

$$i_\mu = \left[\overline{x} - t_\alpha(\nu)\frac{s}{\sqrt{n}},\ \overline{x} + t_\alpha(\nu)\frac{s}{\sqrt{n}}\right].$$

Il y a donc $100(1-\alpha)\%$ de chances que $i_\mu$ contienne $\mu$.

*Approximation :* Lorsque $\nu \geq 31$, on peut utiliser l'approximation $\mathcal{T}(\nu) \approx \mathcal{N}(0,1)$ qui entraîne

$$t_\alpha(\nu) \simeq z_\alpha,$$

où $z_\alpha$ est le réel vérifiant :

$$\mathbb{P}(|Z| \geq z_\alpha) = \alpha,$$

$Z \sim \mathcal{N}(0,1)$ (celui-ci est évaluable dans la table 1 page 223).

**Complément : Z-IntConf-Lim**

Dans le cas où $n \geq 1000$, l'hypothèse : "$X \sim \mathcal{N}(\mu, \sigma^2)$" peut être omise.

On représente alors toujours le caractère par une *var* $X$ et sa moyenne est donnée par $\mathbb{E}(X)$. Elle est inconnue.

On veut déterminer un intervalle de confiance pour $\mathbb{E}(X)$ au niveau $100(1-\alpha)\%$, $\alpha \in ]0,1[$.

Dans ce cas, on utilise le Z-IntConf-Lim.

Sa construction repose sur la loi normale centrée réduite $\mathcal{N}(0,1)$, laquelle est obtenue en tant que loi limite (d'où le "Z" et le "Lim").

Pour obtenir le Z-IntConf-Lim, il faut calculer :

- $\overline{x}$ : la moyenne de $x_1, \ldots, x_n$,
- $s$ : l'écart-type corrigé de $x_1, \ldots, x_n$,
- le réel $z_\alpha$ vérifiant :

$$\mathbb{P}(|Z| \geq z_\alpha) = \alpha,$$

où $Z \sim \mathcal{N}(0,1)$.

Ce réel est évaluable dans la table 1 page 223.

Le Z-IntConf-Lim est alors donné par :

$$i_{\mathbb{E}(X)} = \left[\overline{x} - z_\alpha \frac{s}{\sqrt{n}},\ \overline{x} + z_\alpha \frac{s}{\sqrt{n}}\right].$$

Il y a donc $100(1-\alpha)\%$ de chances que $i_{\mathbb{E}(X)}$ contienne $\mathbb{E}(X)$.

### 2.1.3 Intervalles de confiance pour la variance : Chi2-IntConf

#### Contexte

On étudie un caractère représenté par une *var* $X \sim \mathcal{N}(\mu, \sigma^2)$.

Les données sont constituées de la valeur de $X$ pour chacun des $n$ individus d'un échantillon. Ces valeurs sont notées $x_1, \ldots, x_n$.

Les paramètres $\mu$ et $\sigma$ sont inconnus. On rappelle que $\sigma^2$ (ou $\sigma$) représente la dispersion de $X$ autour de $\mu$.

#### Objectif

On veut déterminer un intervalle de confiance pour $\sigma^2$ au niveau $100(1-\alpha)\%$, $\alpha \in ]0,1[$.

#### Chi2-IntConf

On utilise un Chi2-IntConf.

Sa construction repose sur la loi du Chi-deux $\chi^2(\nu)$.

Pour obtenir le Chi2-IntConf, il faut calculer :

- $s$ : l'écart-type corrigé de $x_1, \ldots, x_n$,