

## Échantillonnage et estimation

Dans ce chapitre, on s'intéresse à un caractère dans une population donnée dont la proportion est notée  $p$ . Cette proportion sera dans quelques cas connue (échantillonnage), dans certains cas supposée connue (prise de décision) et dans d'autres cas inconnue (estimation).

Pour des raisons généralement économiques, on étudie le caractère, non pas sur la population entière, mais sur des échantillons de taille  $n$  extraits de cette population. Pour ce faire, on prélève au hasard des individus de cette population un par un **avec remise**. On constitue ainsi un échantillon aléatoire de taille  $n$ .

Dans une situation tel qu'un sondage on pourrait donc être amené à interroger deux fois la même personne. Pour éviter cela on procède successivement et **sans remise**. Si la taille de l'échantillon n'excède pas 10 % de la taille de la population entière, ce prélèvement ne modifie pas sensiblement la proportion du caractère dans la population.

### 1. Échantillonnage et prise de décisions

---

#### ❖ Intervalle de fluctuation asymptotique d'une fréquence

Considérons une population dans laquelle un caractère qualitatif est répandu avec une proportion connue  $p$ .

Prélevons un échantillon de taille  $n$  de cette population et soit  $X$  la variable aléatoire qui compte le nombre d'individu possédant le caractère ; cette variable suit la loi binomiale de paramètres  $n$  et  $p$ . La variable aléatoire  $F = \frac{X}{n}$  est la fréquence aléatoire du caractère dans l'échantillon.

#### Exemple A

Dans une urne contenant 4 boules rouges et 6 boules blanches, on effectue 100 tirages au hasard avec remise.

Le nombre  $X$  de boules rouges suit la loi binomiale de paramètres  $n = 100$  et  $p = 0,4$ . La fréquence de boules rouges est donnée par la variable aléatoire  $F = \frac{X}{n} = \frac{X}{100}$ .

À l'aide de la calculatrice, on peut effectuer divers calculs. Par exemple :

- La probabilité qu'on observe 42 % de boules rouges est
$$P(F = 0,42) = P(X = 42) \approx 0,0742.$$
- La probabilité qu'on observe au moins 40 % de boules rouges est
$$P(F \geq 0,4) = P(X \geq 40) = 1 - P(X \leq 39) \approx 0,538.$$
- La probabilité qu'on observe entre 31 % et 50 % de boules rouges est
$$P(0,31 \leq F \leq 0,50) = P(X \leq 50) - P(X \leq 30) \approx 0,958.$$

**Définition.** On appelle intervalle de fluctuation au seuil de 95 % un intervalle  $I$  tel que

$$P(F \in I) \geq 0,95.$$

En seconde on a admis que l'intervalle  $I = \left[ p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}} \right]$  convient. Le résultat était énoncé sous cette forme.

**Théorème.** Si  $p$  est la proportion d'un caractère dans une population (avec  $0,2 \leq p \leq 0,8$ ), alors pour un échantillon de taille  $n$  avec  $n \geq 25$ , la fréquence  $f$  du caractère dans l'échantillon appartient à l'intervalle  $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$  avec une probabilité 95 %.

### Exemple A

L'intervalle de fluctuation au seuil de 95 % de la fréquence de boules rouges observées dans des échantillons de taille 100 est

$$\left[0,4 - \frac{1}{\sqrt{100}}; 0,4 + \frac{1}{\sqrt{100}}\right] = [0,3; 0,5].$$

Cela signifie qu'avec une probabilité de 95 %, la fréquence de boules rouges dans un échantillon de taille 100 sera comprise entre 30 % et 50 %

### ❖ Intervalle de fluctuation et la loi binomiale (rappel de Première ES)

La loi binomiale permet de calculer très exactement des intervalles de fluctuations au seuil de 95 %. On cherche deux entiers  $a$  et  $b$  tels que  $P(a \leq X \leq b) \geq 0,95$ . Pour cela, on prendra les plus petits entiers  $a$  et  $b$  tels

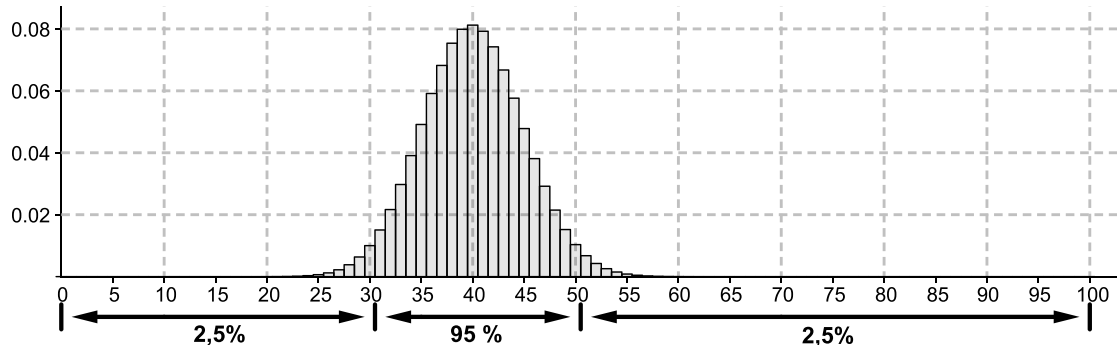
$$P(X \leq a) > 0,025 \text{ et } P(X \leq b) \geq 0,975.$$

Puisque  $a$  est le plus petit entier vérifiant  $P(X \leq a) > 0,025$ , on a  $P(X \leq a - 1) \leq 0,025$  et donc

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a - 1) \geq 0,975 - 0,025 = 0,95.$$

Il en résulte en divisant par  $n$  que  $P\left(\frac{a}{n} \leq F \leq \frac{b}{n}\right) \geq 0,95$ , et donc l'intervalle  $\left[\frac{a}{n}; \frac{b}{n}\right]$  est un intervalle de fluctuation au seuil de 95 %.

### Exemple A



À l'aide d'un tableur ou d'une calculatrice, on constate que  $a = 31$  est le plus petit entier tel que  $P(X \leq a) > 0,025$  et  $b = 50$  est le plus petit entier tel que  $P(X \leq b) \geq 0,975$ . Par conséquent on a  $P(31 \leq X \leq 50) \geq 0,95$  ou encore  $P(0,31 \leq F \leq 0,50) \geq 0,95$ .

```
Graph1 Graph2 Graph3
Y1=binomFRÉP(100,
0,4,X)
V2=
V3=
V4=
V5=
V6=
```

X	Y1
27	.0046
28	.00843
29	.01478
30	.02398
31	.03398
32	.0415
33	.04725

X=31

X	Y1
46	.90702
47	.93621
48	.9577
49	.9729
50	.98324
51	.98899
52	.99424

X=50

Donc  $[0,31; 0,50]$  est un intervalle de fluctuation au seuil de 95 % de la fréquence de boules rouges observées dans les échantillons de taille 100.

Avec la méthode de seconde on avait trouvé  $[0,3; 0,5]$ , ce qui à peine moins précis, mais avec des calculs bien plus simples.

## ❖ Intervalle de fluctuation asymptotique et loi normale

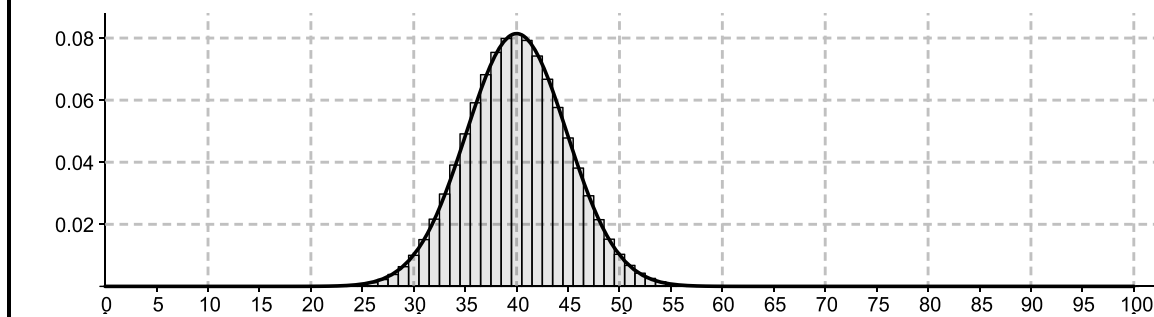
La variable aléatoire  $X$  suit une loi binomiale de paramètres  $n$  et  $p$ . On admet que sa loi peut être approximée par la loi normale d'espérance  $\mu = np$  et d'écart-type  $\sigma = \sqrt{np(1-p)}$ . Graphiquement, cela se traduit par le fait que l'histogramme de la loi binomiale est très bien approché par la courbe représentative de la densité de la loi normale  $\mathcal{N}(np; np(1-p))$ . En pratique dès que  $n \geq 30$ ,  $np \geq 5$  et  $n(1-p) \geq 5$  on utilise cette approximation.

### Exemple A

$X$  suit la loi binomiale de paramètre  $n = 100$  et  $p = 0,4$ .

Comme  $n \geq 30$ ,  $np = 40 \geq 5$  et  $n(1-p) = 60 \geq 5$ , le résultat admis précédemment affirme que la loi de  $X$  peut être approximée par une loi normale d'espérance  $\mu = np = 40$  et d'écart-type  $\sigma = \sqrt{np(1-p)} = \sqrt{24}$ , c'est-à-dire par la loi  $\mathcal{N}(40; 24)$ .

On a représenté sur un même graphique la distribution de  $X$  ainsi que la densité de la loi  $\mathcal{N}(40; 24)$ .



On a vu dans le chapitre précédent que  $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0,95$ .

Plus précisément, en utilisant `FracNormale(0.025)`, on a

$$P(\mu - 1,96\sigma \leq X \leq \mu + 1,96\sigma) \approx 0,95.$$

On en déduit

$$P\left(np - 1,96\sqrt{np(1-p)} \leq X \leq np + 1,96\sqrt{np(1-p)}\right) \approx 0,95$$

ou encore en divisant par  $n$ , puisque  $F = \frac{X}{n}$ ,

$$P\left(p - 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F \leq p + 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \approx 0,95.$$

Ce résultat conduit au théorème suivant.

**Théorème.** Si  $n \geq 30$ ,  $np \geq 5$  et  $n(1-p) \geq 5$ , l'intervalle

$$\left[ p - 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

est un intervalle de fluctuation au seuil de 95 % de la variable aléatoire fréquence  $F$ .

Plus précisément, l'intervalle de fluctuation IF vérifie  $P(F \in \text{IF}) \geq 0,95$ .

On prendra soin d'arrondir la borne inférieure par défaut et la borne supérieure par excès pour ne pas réduire l'intervalle et prendre le risque qu'il ne vérifie plus  $P(F \in \text{IF}) \geq 0,95$ .

Cet intervalle de fluctuation est dit asymptotique car il n'est valide que si les conditions sur  $n$  et  $p$  sont satisfaites (il faut que  $n$  soit suffisamment grand).

### Exemple A

Ici on a  $1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \approx 0,096$ , donc l'intervalle de fluctuation donné par cette méthode est  $[0,4 - 0,096; 0,4 + 0,096] \approx [0,304; 0,496]$ .

### ❖ **Prise de décision**

Dans ce paragraphe, la proportion du caractère étudié est supposée être égale à  $p$ .

La prise de décision consiste, à partir d'un échantillon de taille  $n$ , à valider ou non cette hypothèse faite sur la proportion  $p$ . Pour ce faire,

- on calcule la fréquence observée  $f$  du caractère étudié dans cet échantillon ;
- on détermine un intervalle de fluctuation seuil de 95 % de la fréquence ;
- enfin on applique la règle de décision suivante
  - si la fréquence observée  $f$  appartient à l'intervalle de fluctuation au seuil de 95 %, on ne rejette pas l'hypothèse faite sur  $p$  ;
  - si la fréquence observée  $f$  n'appartient pas à l'intervalle de fluctuation au seuil de 95 %, on rejette l'hypothèse faite sur la proportion  $p$  avec un risque de 5 % de se tromper.

### **Remarque.**

Le risque de 5 % signifie que la probabilité que l'on rejette à tort l'hypothèse faite sur la proportion  $p$  alors qu'elle est vraie est approximativement égale à 0,05. C'est une probabilité conditionnelle.

Dans le cas où l'on ne rejette pas l'hypothèse faite sur la proportion  $p$ , le risque d'erreur n'est pas quantifié.

### Exemple A

Une machine fabrique des composants électroniques qui présentent un défaut de fabrication avec une probabilité 0,4.

On constitue un échantillon de taille 100 de composants issus de la fabrication. On en détecte 49 présentant un défaut de fabrication. La machine nécessite-t-elle d'être réparée ?

On fait l'hypothèse que la machine est bien réglée, donc que  $p = 0,4$ .

Un intervalle de fluctuation au seuil de 95 % de la fréquence de pièces présentant un défaut sur les échantillons de taille 100 est  $[0,304; 0,496]$ .

Ici  $f = \frac{49}{100} = 0,49$  appartient à l'intervalle de fluctuation, donc on ne rejette pas l'hypothèse  $p = 0,4$ .

On peut donc dire que la machine ne nécessite pas d'être réparée, au seuil de confiance de 95 %.

## 2. Estimation et intervalle de confiance

---

Dans ce paragraphe, on suppose que la proportion  $p$  du caractère étudié est inconnue et on cherche à estimer  $p$  à partir de la fréquence observée  $f$  sur un échantillon.

**Définition.** Soit  $f$  la fréquence observée d'un caractère dans un échantillon de taille  $n$  extrait d'une population dans laquelle la proportion de ce caractère est  $p$ .

L'intervalle  $\left[ f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right]$  est un intervalle de confiance de la proportion  $p$  au niveau de confiance 95 %.

À chaque échantillon de taille  $n$  correspond un intervalle de confiance de  $p$ .

### Exemple

On cherche à estimer la proportion  $p$  d'élèves du lycée pratiquant une activité physique régulière.

Sur 100 élèves interrogés, on en a observé 63 pratiquant du sport. Ainsi l'intervalle de confiance au seuil de 95 % de la proportion d'élèves pratiquant du sport est

$$IC = \left[ \frac{63}{100} - \frac{1}{\sqrt{100}}; \frac{63}{100} + \frac{1}{\sqrt{100}} \right] = [0,53; 0,73].$$

Un deuxième sondage a révélé 68 élèves sportifs. Cela conduit à un nouvel intervalle de confiance :

$$IC' = [0,58; 0,78].$$

Le niveau de confiance 95 % que si l'on réalise 100 sondages,  $p$  se situera dans au moins 95 des intervalles de confiance ainsi construits.

### Remarques.

- Il est incorrect de dire « la proportion se situe avec une probabilité égale à 95 % dans l'intervalle de confiance » car la proportion n'est pas aléatoire.
- Dans d'autres disciplines on utilise l'intervalle de confiance suivant :

$$\left[ f - 1,96 \frac{\sqrt{f(1-f)}}{\sqrt{n}}; f + 1,96 \frac{\sqrt{f(1-f)}}{\sqrt{n}} \right].$$

- La proportion  $p$  étant inconnue, on ne peut pas vérifier si les paramètres  $n$  et  $p$  satisfont les conditions  $n \geq 30$ ,  $np \geq 5$  et  $n(1-p) \geq 5$ . Pour remédier à ce problème, on approche la proportion inconnue  $p$  par la fréquence observée  $f$  sur l'échantillon, puis on vérifie si les conditions  $n \geq 30$ ,  $nf \geq 5$  et  $n(1-f) \geq 5$  sont satisfaites.

### Exemple

Un sondage réalisé sur 989 personnes par l'institut IPSOS le 19 avril 2002 avant le premier tour des élections présidentielles crédite Chirac de 20 % des votes, Jospin de 18 % et Le Pen de 14 %.

Les intervalles de confiance des proportions de votes pour ces trois candidats étaient donc respectivement

$$IC_C = \left[ 0,2 - \frac{1}{\sqrt{989}}; 0,2 + \frac{1}{\sqrt{989}} \right] \approx [0,168; 0,232],$$

$$IC_J = [0,148; 0,212] \text{ et } IC_{LP} = [0,108; 0,172]$$

Au seuil de confiance 95 %, on constate qu'il était impossible de prédire le classement de ces trois candidats à l'issue du premier tour.

Leurs scores ont été respectivement 19,88 %, 16,18 % et 16,86 %, propulsant Le Pen au second tour contre toute attente (de la part des journalistes, mais pas des statisticiens).