ÉCHANTILLONNAGE ET SIMULATION
COMPLÉMENTS

I. Intervalle de fluctuation

PROPRIÉTÉ On suppose $n \geqslant 25$ et $0{,}2 \leqslant p \leqslant 0{,}8$.

Pour environ 95% des échantillons de taille n relevant du modèle de Bernoulli de probabilité p, la fréquence d'apparition du 1 appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.

Cet intervalle s'appelle ***intervalle de fluctuation au seuil de 95 %***.

Autrement dit, si les conditions sont respectées :

Si, pour échantillon de taille n, on obtient une fréquence f de l'issue choisie, alors :

$$\mathrm{p}\left(f \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]\right) \approx 0{,}95 .$$

Il faut insister sur le « environ 95 % », car dans le programme scolaire il est écrit « au moins 95 % », ce qui est faux !

En effet, en notant J_n l'intervalle de fluctuation de Seconde, voici quelques exemples où les conditions sont vérifiées, mais la proba est inférieure à 95 % :

n	p	$n \geqslant 25$ et $0{,}2 \leqslant p \leqslant 0{,}8$?	$\mathrm{p}\left(\mathrm{F}_n \in \mathrm{J}_n\right)$
30	0,484	oui	**≈ 0,9341**
528	0,5	oui	**≈ 0,9499**
28	0,59	oui	**≈ 0,9474**

En programmant des algorithmes (par exemple sur *Xcas*), on peut déterminer qu'en faisant varier p de 0 à 1, avec un pas de 0,0001, et en faisant varier n de 1 à 2000 :

- sur les 11 856 000 couples $(n ; p)$ qui vérifient les conditions $n \geqslant 25$ et $0{,}2 \leqslant p \leqslant 0{,}8$, 43 336 donnent $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{J}_n\right) < 0{,}95$. Cela fait seulement environ 0,37 % des couples !

- la plus grande valeur de n telle que $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{J}_n\right) < 0{,}95$ est $n = 528$ (et $p = 0{,}5$).
On trouve une probabilité d'environ 0,9499.

- la probabilité minimale des $\mathrm{p}\left(\mathrm{F}_n \in \mathrm{J}_n\right)$ (en vérifiant les conditions $n \geqslant 25$ et $0{,}2 \leqslant p \leqslant 0{,}8$) est atteinte lorsque $n = 30$ et $p = 0{,}484$. On trouve une probabilité d'environ 0,9341.

On est donc pas loin des 95 %, mais pas toujours au-dessus...

II. Prise de décision

II.1 Conjecturer une proportion et valider/invalider cette hypothèse

On considère un caractère dont la proportion dans la population 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 faire sur la proportion p :

1) On calcule la fréquence observée f du caractère dans cet échantillon.

2) Si les conditions d'approximation $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$ sont vérifiées, on détermine l'intervalle de fluctuation asymptotique au seuil 0,95.
Si les conditions d'approximation ne sont pas vérifiées, on peut déterminer l'intervalle de fluctuation étudié en Seconde ou en Première…

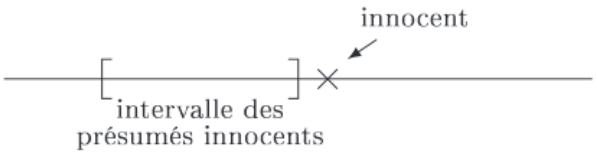
3) On applique la règle suivante :

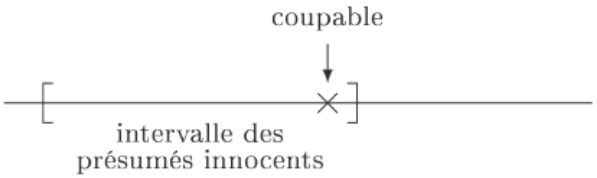
- Si $f \notin I$ alors on rejette l'hypothèse faite sur p.
  Dans ce cas, **il y a un risque[1] de se tromper de 5 %** :
  la probabilité qu'on rejette à tort l'hypothèse faite sur p alors qu'elle est vraie (proba. cond.) est environ égale à 5 %.

- Si $f \in I$ alors on accepte l'hypothèse faite sur p.
  Dans ce cas, **le risque d'erreur n'est pas quantifié[2] !**

Pourquoi ne pas abaisser le seuil de rejet d'une hypothèse ?

On pourrait penser qu'il suffit de réduire le risque d'erreur (de première espèce) de rejeter une hypothèse à tort, de façon à n'avancer que des hypothèses très fiables.
Mais en faisant cela, on augmente le risque de commettre une autre erreur (de seconde espèce) : accepter l'hypothèse alors qu'elle est fausse !
Une analogie simple suffit à faire comprendre la situation : une prise de décision est comme un jugement au tribunal. L'hypothèse est que le prévenu est présumé innocent.
Il y a deux risques au jugement : celui de condamner un innocent (rejet à tort de l'hypothèse, première espèce), ou d'innocenter un coupable (acceptation à tort de l'hypothèse, seconde espèce).
Plus on réduit un risque, plus on augmente l'autre :





Ainsi, la décision que l'on doit prendre est un compromis adapté à la situation.
Voilà pourquoi le seuil de 5 % est souvent utilisé.

1 On parle de « risque de première espèce ». Ce risque est défini à l'avance (le plus souvent 1 % ou 5 %).
2 On parle de « risque de seconde espèce ».
A taille d'échantillon égale, si l'on diminue le risque de première espèce, on augmente le risque de seconde espèce…

II.2 Tester la conformité d'un échantillon par rapport à la population

On considère un caractère dont la proportion dans la population est connue, égale à p.
La prise de décision consiste, à partir d'un échantillon de taille n, à valider ou non sa représentativité.

1) On calcule la fréquence observée f du caractère dans cet échantillon.

2) Si les conditions d'approximation $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$ sont vérifiées, on détermine l'intervalle de fluctuation asymptotique au seuil 0,95.
Si les conditions d'approximation ne sont pas vérifiées, on peut déterminer l'intervalle de fluctuation étudié en Seconde ou en Première…

3) On applique la règle suivante :

- Si $f \in I$ alors on considère que l'échantillon est représentatif de la population.

- Si $f \notin I$ alors on considère que l'échantillon n'est pas représentatif de la population.

Remarque : ici, on ne fait pas d'hypothèse sur la probabilité théorique p, puisqu'on la connaît.
On ne commet dont aucune erreur…

III. Intervalle de confiance

La notion d'intervalle de confiance est plus difficile à appréhender !

Elle est souvent « simplifiée » (comme sur les vidéos YouTube ou dans la plupart des livres) mais les conclusions sont souvent confuses.

Soit p une proportion inconnue dans une population.
Pour un échantillon de taille n de cette population, on calcule la fréquence f .
Alors la probabilité que l'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ contienne p est environ égale à 95 %.

Autrement dit, parmi tous les échantillons de taille n qu'on peut obtenir, au moins 95 % d'entre eux sont tels que l'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ contient la proportion p .

On dit donc que $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ *est* ***un intervalle de confiance*** *de la proportion inconnue* p *à un niveau de confiance de 95 %.*

On admet que l'on « peut » utiliser cet intervalle lorsque :

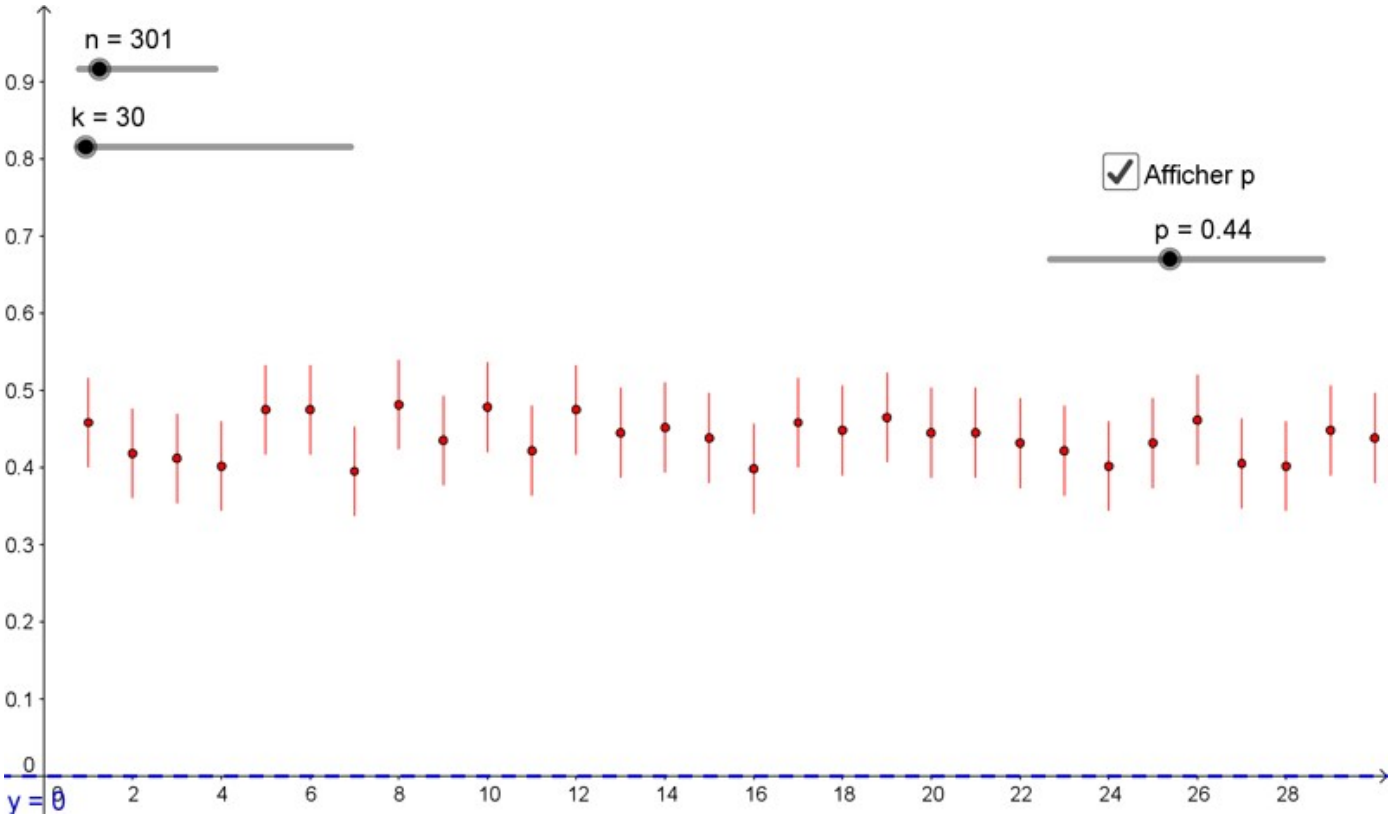
$$n\geqslant 30 \quad ; \quad nf\geqslant 5 \quad ; \quad n(1-f)\geqslant 5 .$$

1) Les conditions ci-dessus sont un peu du « n'importe quoi mathématique », ou plutôt issues de l'expérience des statisticiens…

2) Contrairement à ce qu'on peut lire dans certains manuels ou sur les vidéos YouTube, écrire $\mathrm{p}\left(p\in\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]\right)\approx 0{,}95$ n'a aucun sens ! Une probabilité n'a de sens que pour une variable aléatoire, pour quelque chose qui varie. Or ici, p ne varie pas du tout.
Voilà pourquoi on écrit plutôt $\mathrm{p}\left(\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]\ni p\right)\approx 0{,}95$.
Là, ça a du sens, car l'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ varie lui, c'est un intervalle aléatoire.
La nuance peut paraître d'abord un peu idiote, mais elle ne l'est pas.

D'ailleurs, l'intervalle de confiance permet juste de dire que « parmi tous les échantillons de taille n qu'on peut obtenir, au moins 95 % d'entre eux sont tels que l'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ contient la proportion p . »

Si on veut utiliser des intervalles de confiance intelligemment, on devrait plutôt faire plein de sondages (AVEC LE MEME NOMBRE DE PERSONNES), et représenter tous les intervalles obtenus sur un graphique « en peigne », pour en déduire des informations sur p :



Là, par exemple on a fait 30 sondages sur 301 personnes à chaque fois.
Avec tous les intervalles, on cherche le nombre qui fait qu'au moins 95 % des intervalles contiennent ce nombre… Et on tombe facilement sur la valeur réelle de p (ici choisie par le logiciel), qui est 0,44. Alors que si, par exemple, on prend le 4ème intervalle dessinée, on n'aurait pas envie de dire que $p=0{,}44$.

IV. L'exemple Royal contre Sarkozy

1) A partir d'un sondage, on ne peut jamais rien prévoir : on ne peut prévoir le futur, mais juste donner une probabilité.

2) Comme les intervalles ne se chevauchent pas, on peut dire, avec un « risque d'erreur d'environ 5 % », que Sarkozy gagnera l'élection face à Royal.

Mais cette notion de risque est sujette à de mauvaises interprétations… On devrait dire :

« Parmi tous les échantillons de taille 606 qu'on peut obtenir en faisant des sondages sur 606 personnes, au moins 95 % d'entre eux sont tels que l'intervalle [0,5092; 0,5907] contient la proportion réelle d'électeurs de Sarkozy. »

au lieu de

« La probabilité pour que l'intervalle [0,5092; 0,5907] contienne la proportion réelle d'électeurs de Sarkozy est d'environ 95 %. »

Parce que cette dernière phrase laisse entendre que quels que soient les sondages (de taille 606 ou 10000 ou 100) on peut parler de probabilité de 95 %…