Échantillonnage, estimation

La théorie comporte deux applications principales :

- 1/ La prise de décision à partir d'un échantillon lorsque la proportion dans la population entière est connue ou supposée connue
- 2/ L'estimation, à partir d'un échantillon, d'une proportion inconnue dans la population.

Échantillonnage en 2nde

Objectifs

(programme officiel - préambule de la partie « Statistiques et probabilités »)

- faire réfléchir les élèves à la conception et à la mise en œuvre d'une simulation ;
- sensibiliser les élèves à la fluctuation d'échantillonnage, aux notions d'intervalle de fluctuation et d'intervalle de confiance et à l'utilisation qui peut en être faite.

Échantillon de taille *n* :

(programme officiel)

constitué des résultats

de *n* répétitions indépendantes

de la même expérience

On accepte comme indépendants des « tirages sans remise » si la population est grande.

On ne se pose pas la question de la représentativité de l'échantillon.

Dans le sens commun des sondages, un échantillon est un sous-ensemble obtenu par prélèvement aléatoire dans une population.

Capacités attendues (programme officiel) :

- Concevoir, mettre en œuvre et exploiter des simulations de situations concrètes à l'aide du tableur ou d'une calculatrice

- Exploiter et faire une analyse critique d'un résultat d'échantillonnage

Après simulations, on se rend compte que les fréquences obtenues varient d'un échantillon à l'autre pour une même expérience : c'est ce qu'on appelle la **fluctuation d'échantillonnage**.

Lorsque la taille de l'échantillon augmente, les fluctuations diminuent …

Questionnement

(2nde – programme officiel) :

- estimation d'une proportion inconnue à partir d'un échantillon
- prise de décision à partir d'un échantillon

Intervalle de fluctuation au seuil de 95 % : centré autour de p (proportion du caractère dans la population), contient, avec une probabilité (environ) égale à 0,95, la fréquence d'apparition du caractère observée dans un échantillon de taille n.

Règle de décision

Si la fréquence observée sur l'échantillon n'appartient pas à l'intervalle de fluctuation à 0,95, on rejette, au risque d'erreur de 5 %, l'hypothèse que l'échantillon est compatible avec le modèle ; on considère que l'observation n'est pas compatible avec le modèle, en ce sens que, dans un tel modèle, elle ne s'observerait que dans 5% des échantillons de taille *n*.

La probabilité de rejeter l'hypothèse alors qu'elle est vraie est inférieure à 5%.

Si la fréquence observée appartient à l'intervalle de fluctuation, on ne peut pas rejeter l'hypothèse …

Intervalle de fluctuation ….

Intervalle de confiance ….

Ne pas confondre !

Intervalle de fluctuation

La proportion du caractère dans la population est **connue** ou **supposée connue**.

Un intervalle de fluctuation à 95% est un intervalle auquel appartient la fréquence observée sur un échantillon dans (environ) 95% des cas (ou : avec une probabilité approximativement égale à 0,95).

Intervalle de confiance

On est dans la théorie de l'estimation.

La proportion du caractère dans la population est **inconnue**.

On essaie d'estimer cette proportion à partir de la fréquence du caractère dans l'échantillon. On détermine un intervalle dans lequel la proportion du caractère dans la population se trouve avec une probabilité (au moins) égale à 0,95.

Comment déterminer un intervalle de fluctuation ?

- $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ (justifié en 2nde par des simulations)

(si *n* supérieur ou égal à 25 (30 ?) et

p compris entre 0,2 et 0,8 ; voir développements théoriques doc ressource Probabilités au clg, p23 et suivantes)

- À l'aide de la loi binomiale (première)
- À l'aide de la loi normale (intervalle de fluctuation asymptotique – prgm de terminale)

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

Cet intervalle ne dépend que de la taille de l'échantillon, pas de celle de la population.

On peut le « vérifier » par des simulations.

Il est le résultat de deux approximations : approximation de la loi binomiale par une loi normale, approximation de l'intervalle correspondant à la loi normale.

Voir doc ressource seconde page 15.

On simule 1000 échantillons de taille 100

d'un modèle de Bernoulli avec $p = 0,4$.

Chaque échantillon est représenté par un point dont l'ordonnée est la fréquence d'apparition du 1.

On observe que la plupart des échantillons ont des fréquences d'apparition du 1 dans l'intervalle [0,3 ; 0,5].

fichier

D'après Doc ressource Probabilités au collège, p24 …

On a vu qu'un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience.

On peut modéliser n répétitions indépendantes de la même expérience de probabilité de succès p en introduisant n variables aléatoires X_i prenant la valeur 1 en cas de succès, 0 en cas d'échec ; X_i suit une loi de Bernoulli. La somme S_n des X_i pour i variant de 1 à n donne le nombre de succès lors des n répétitions.

S_n suit la loi binomiale de paramètres n et p.

Son espérance est np, sa variance est $np(1 - p)$.

Soit F_n la fréquence du nombre de succès dans l'échantillon de taille n

(*supposée « voisine » de p*).

$$F_n = \frac{S_n}{n}$$

- Soit R_n la variable centrée réduite associée à S_n

$$R_n = \frac{S_n - np}{\sqrt{np(1-p)}} \left(= \frac{S_n - m}{\sigma} \right)$$

- La variable centrée réduite associée à F_n est également R_n car :

$$R_n = \frac{F_n - p}{\sqrt{\frac{np(1-p)}{n^2}}} = \frac{F_n - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Le théorème de Moivre-Laplace (cas particulier du théorème « central limite » ou théorème de la limite centrale) permet de dire que la variable R_n converge en loi vers la loi normale centrée réduite.

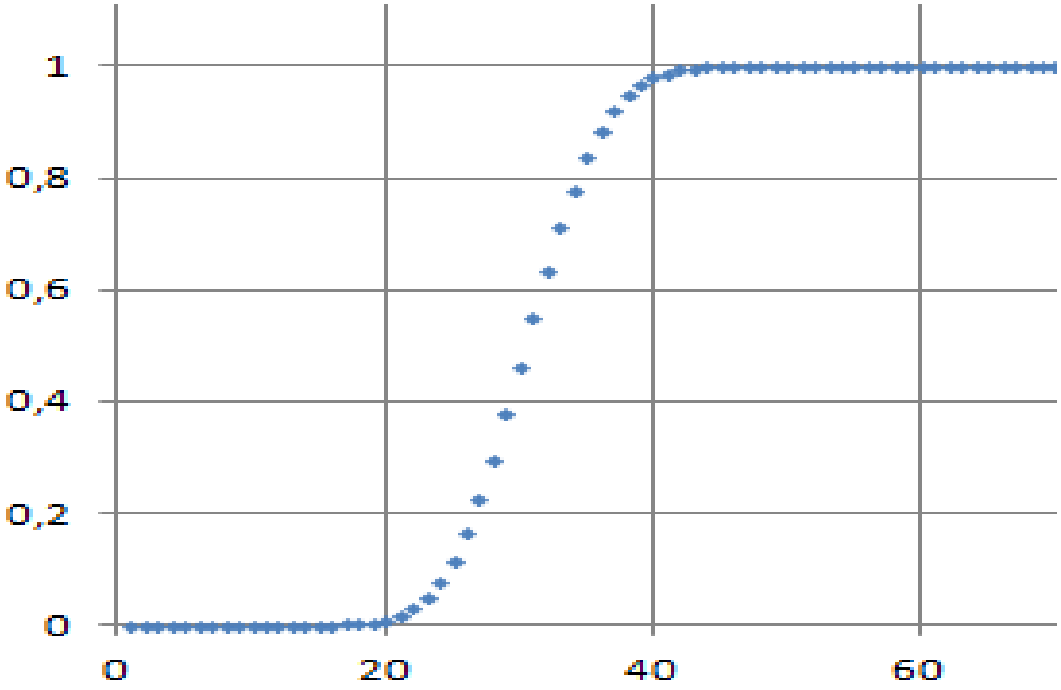
Cette convergence en loi signifie que, lorsque n tend vers l'infini, la probabilité que R_n prenne des valeurs entre a et b tend vers

$$\frac{1}{\sqrt{2\,\pi}}\int_a^b \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x$$

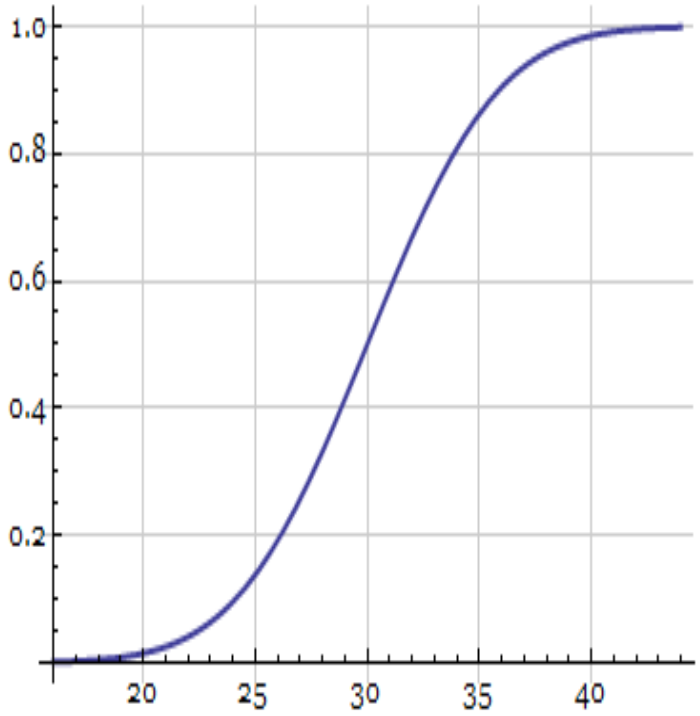
Avec n > 30 (ou n > 25),

on approxime avec une très bonne précision la probabilité que R_n prenne des valeurs entre a et b par sa limite donnée par le théorème de Moivre-Laplace.

Graphiquement
(fonctions de répartition avec *n*=100)

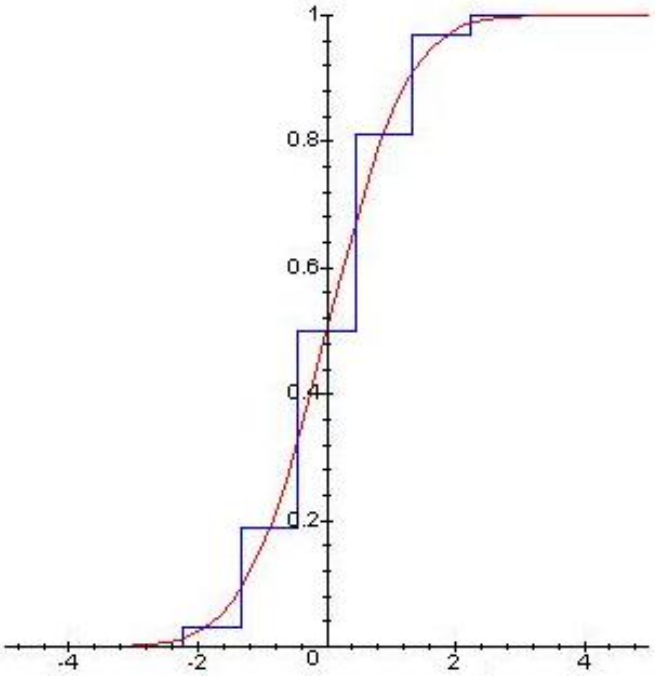


Loi binomiale

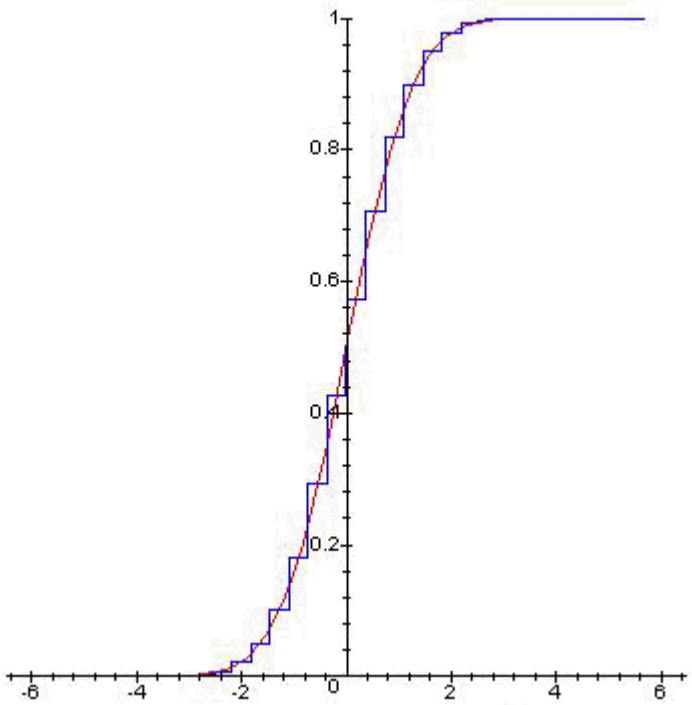


Loi normale (*euler*)

centré réduit (avec *p*=0,5) :



n = 5



n = 30

Une table de la loi normale (centrée réduite) donne la valeur de *u* telle que

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-u}^{u} \mathrm{e}^{-\frac{x^2}{2}} \, \mathrm{d}x$$

prenne une valeur donnée.

On retient que :

$$\Phi(1{,}96) \approx 0{,}95 \ ; \ \Phi(1.65) \approx 0{,}9 \ ; \ \Phi(3) \approx 0{,}99$$

- D’où :

$$P(-1{,}96 \leq R_n \leq 1{,}96) \approx 0{,}95$$

- Or $-1{,}96 \leq R_n \leq 1{,}96$ équivaut à

$$p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \leq F_n \leq p + 1{,}96\sqrt{\frac{p(1-p)}{n}}$$

inégalité qui est donc vraie avec une probabilité voisine de 0,95.

L'intervalle de bornes

$$p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \text{ et } \quad p + 1{,}96\sqrt{\frac{p(1-p)}{n}}$$

est appelé intervalle de fluctuation de niveau 95%

(ou intervalle de fluctuation asymptotique au seuil de 95% …).

Dans environ 95% des séries de n tirages que l'on peut faire, la fréquence empirique de succès f_n obtenue expérimentalement (modélisée par F_n) doit appartenir à cet intervalle.

En seconde, on simplifie …

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

1,96 ≤ 2 et, en étudiant sur

[0 ; 1] la fonction $x \mapsto \sqrt{x(1-x)}$,

on montre que cette fonction a un maximum égal à 0,5 atteint en $x = 0,5$.

Donc : $1,96\sqrt{p(1-p)} \leq 1$

(pour que la majoration ne soit pas trop "grossière", on considère que le résultat est valable si p pas trop éloigné de 0,5 : si $0,2 \leq p \leq 0,8$, alors :

$0,4 \leq \sqrt{p(1-p)} \leq 0,5$).

L’intervalle précédent est donc inclus dans l’intervalle $\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right]$

Environ 95 % des échantillons de taille *n* fournissent une fréquence appartenant à cet intervalle

(ce qu'on peut vérifier par des simulations).

UTILISATION DE LA LOI BINOMIALE

- Avec la notion de variable aléatoire et la loi binomiale, il n'est plus nécessaire d'approximer par la loi normale.
- L'intervalle de fluctuation déterminé est valable « en toutes circonstances ».
- Mais comme la loi binomiale est discrète, on ne pourra pas déterminer précisément un intervalle où la fréquence observée se situe avec une probabilité exactement égale à 0,95.

S_n suit la loi binomiale de paramètres n et p.

Elle prend donc ses valeurs (entières) dans

$[0 ; n]$.

On cherche à partager cet intervalle en trois intervalles $[0, a - 1]$, $[a, b]$ et $[b + 1, n]$ de sorte que S_n prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de 0,025, sans dépasser cette valeur

(donc S_n prend ses valeurs dans l'intervalle

$[a, b]$ avec une probabilité voisine de - ou légèrement supérieure à - 0,95).

a et b dépendent de n.

On fait une hypothèse de « symétrie ».

a est le plus grand entier tel que $P(S < a) \leq 0{,}025$

(S dans l'intervalle $[0, a - 1]$), donc

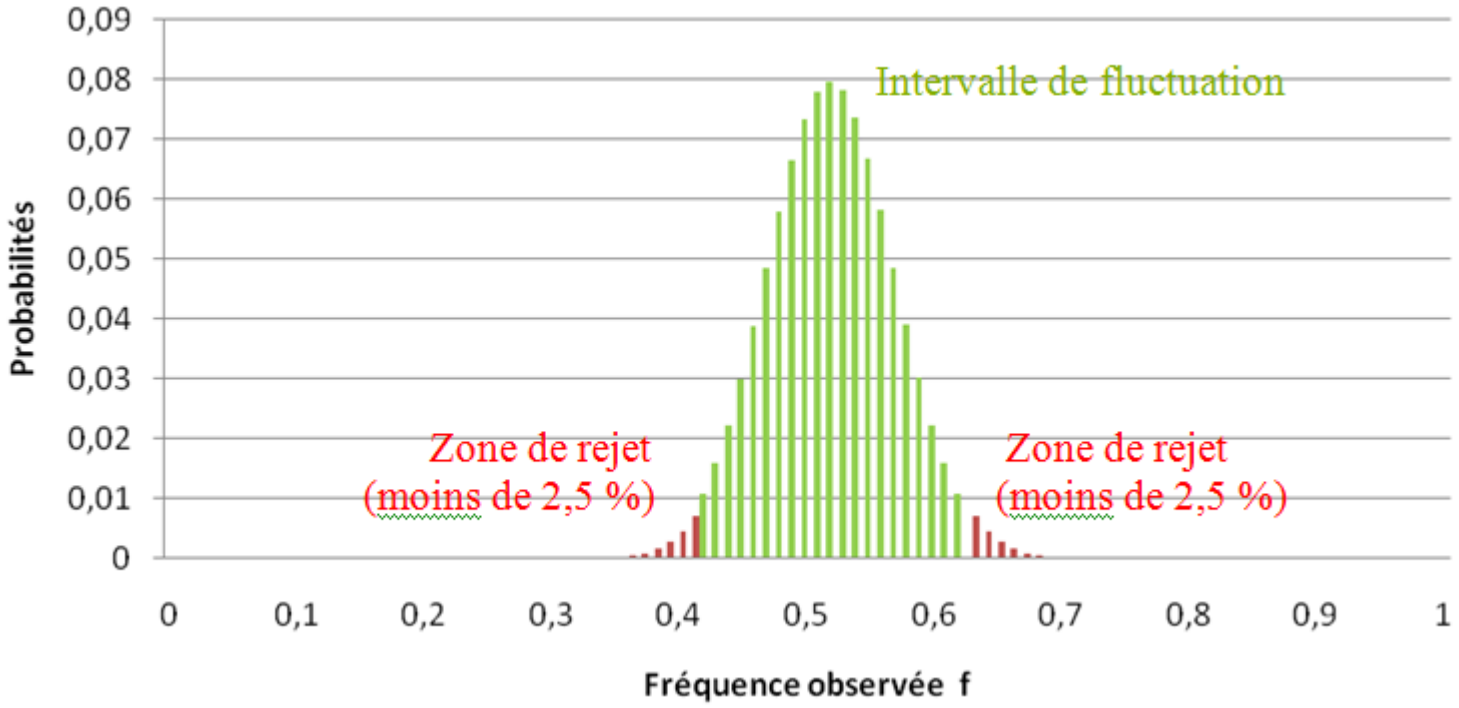
a est le plus petit entier tel que $P(S \leq a) > 0{,}025$

b est le plus petit entier tel que $P(S > b) \leq 0{,}025$

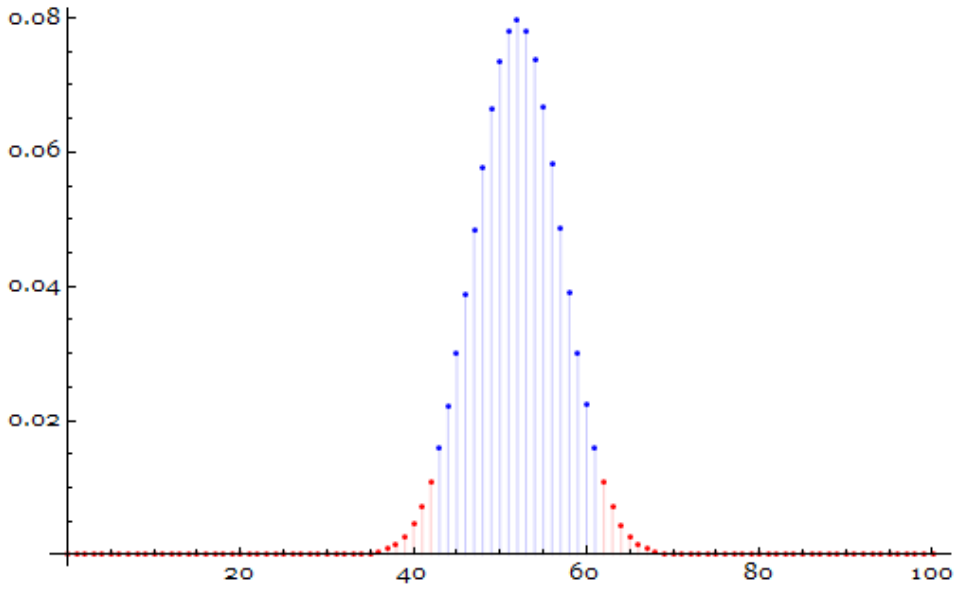
(S dans l'intervalle $[b + 1, n]$), donc

b est le plus petit entier tel que $P(S \leq b) \geq 0{,}975$

(on se ramène à des probabilités du type $P(S \leq k)$ que l'on peut obtenir à l'aide d'un tableur).


Intervalle de fluctuation
Zone de rejet
(moins de 2,5 %)
Zone de rejet
(moins de 2,5 %)
Probabilités
Fréquence observée f
0,09
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1

Outil 3944


0.08
0.06
0.04
0.02
20
40
60
80
100

- On divise par *n* pour passer aux fréquences;
- D'où (définition) :

L'intervalle de fluctuation à 95 % d'une fréquence correspondant à la réalisation, sur un échantillon aléatoire, d'une variable aléatoire X de loi binomiale de paramètres *n* et *p* est … l'intervalle $\left[\frac{a}{n} ; \frac{b}{n}\right]$ défini par :

a est le plus petit entier tel que $P(X \leq a) > 0{,}025$

b est le plus petit entier tel que $P(X \leq b) \geq 0{,}975$ …

- Voir autres précisions sur la détermination de a et b d.r. page 39.
- La « règle de décision » est la même …
- L'intervalle de fluctuation déterminé avec la loi binomiale est « quasiment » centré sur p dès que n est assez grand.
- Cet intervalle est « quasiment » le même que celui vu en seconde pour les « grandes binomiales »

  ($n > 25$; $0{,}2 < p < 0{,}8$) (voir page 48).

EULER

Déterminer l'intervalle de fluctuation à 95 % d'une variable aléatoire suivant une loi binomiale de paramètres donnés

ressource 3943

Soit Ω l'univers d'une expérience aléatoire.
Soit X une variable aléatoire définie sur Ω qui suit la loi binomiale $B\left(30, \frac{12}{19}\right)$.
On admet les résultats donnés par le tableau suivant (probabilités arrondies au millionième).

Déterminez l'intervalle I de fluctuation à 95 % de X.

$I =$ [;]

Valider

INTERVALLE DE CONFIANCE

- On ne connaît pas la proportion *p* du caractère dans la population
- On cherche à estimer *p* à partir de la fréquence empirique *f* (expérimentale) mesurée dans un échantillon

On est dans la théorie de l'**estimation**.

Comment estimer p?

- On peut faire une **estimation ponctuelle**

en posant $p = f$

- *Mieux :*

  On peut chercher un **intervalle de confiance**

  de la proportion p (c'est-à-dire un intervalle contenant « très vraisemblablement » p) à partir de la fréquence f mesurée dans un échantillon de taille n

- Parmi tous les échantillons de taille *n* possibles, 95% des intervalles associés de la forme $\left[f - \frac{1}{\sqrt{n}} \; ; \; f + \frac{1}{\sqrt{n}}\right]$ contiennent le nombre *p*.

- La fréquence empirique *f* étant connue, on dit que cet intervalle est un **intervalle de confiance à 95%** pour la proportion *p*

  (on parle aussi de « fourchette » au niveau de confiance de 95%).

  Dans plus de 95 % des cas, la « fourchette » recouvre effectivement la valeur *p*.

$$p \in \left[f - \frac{1}{\sqrt{n}} \; ; \; f + \frac{1}{\sqrt{n}} \right]$$

est équivalent à :

$$f \in \left[p - \frac{1}{\sqrt{n}} \; ; \; p + \frac{1}{\sqrt{n}} \right]$$

Quand on entend comme résultat d'un sondage : *Il y a actuellement 52 % de gens qui voteraient pour M. X au deuxième tour (sondage effectué auprès de 948 personnes),* il faut comprendre *:*

« Il y a 95 % de chance pour que l'intervalle [49% ; 55%] contienne le pourcentage de gens prêts à voter pour M. X au deuxième tour »,

ce qui n'est pas tout à fait la même chose que de dire que 52 % des électeurs sont prêts à voter pour lui !

Une réponse à la question : « combien faut-il tirer de boules dans une urne de Bernoulli pour pouvoir faire une estimation de sa composition avec une précision donnée a priori » est donc : en tirant n boules avec remise, on obtient une estimation de p par un intervalle d'amplitude $\frac{2}{\sqrt{n}}$, avec une confiance de plus de 95 %.

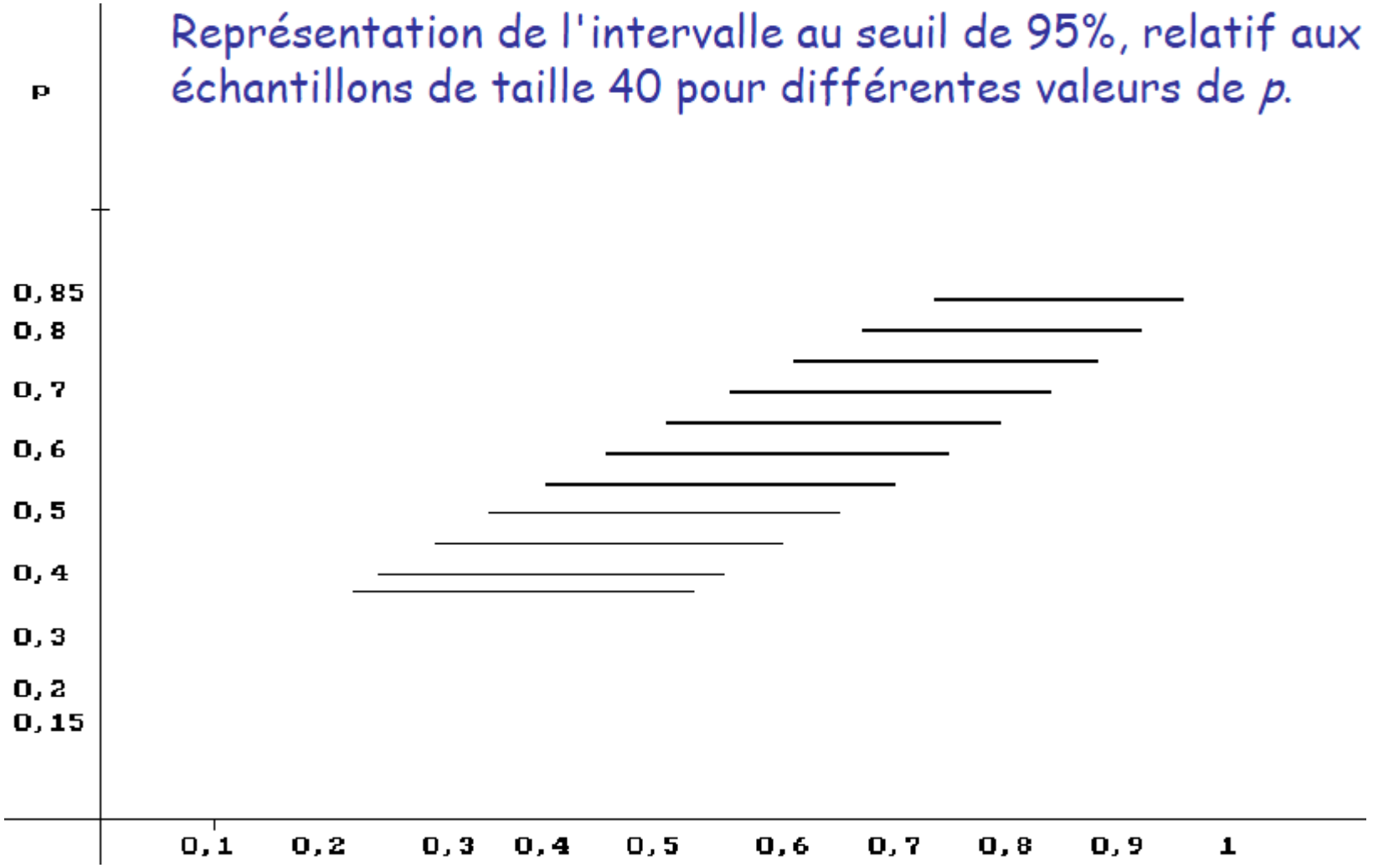
Si l'on tire n = 1 000 boules (avec remise) on a une estimation de p à plus de 95 % de confiance par un intervalle d'amplitude 6 %.

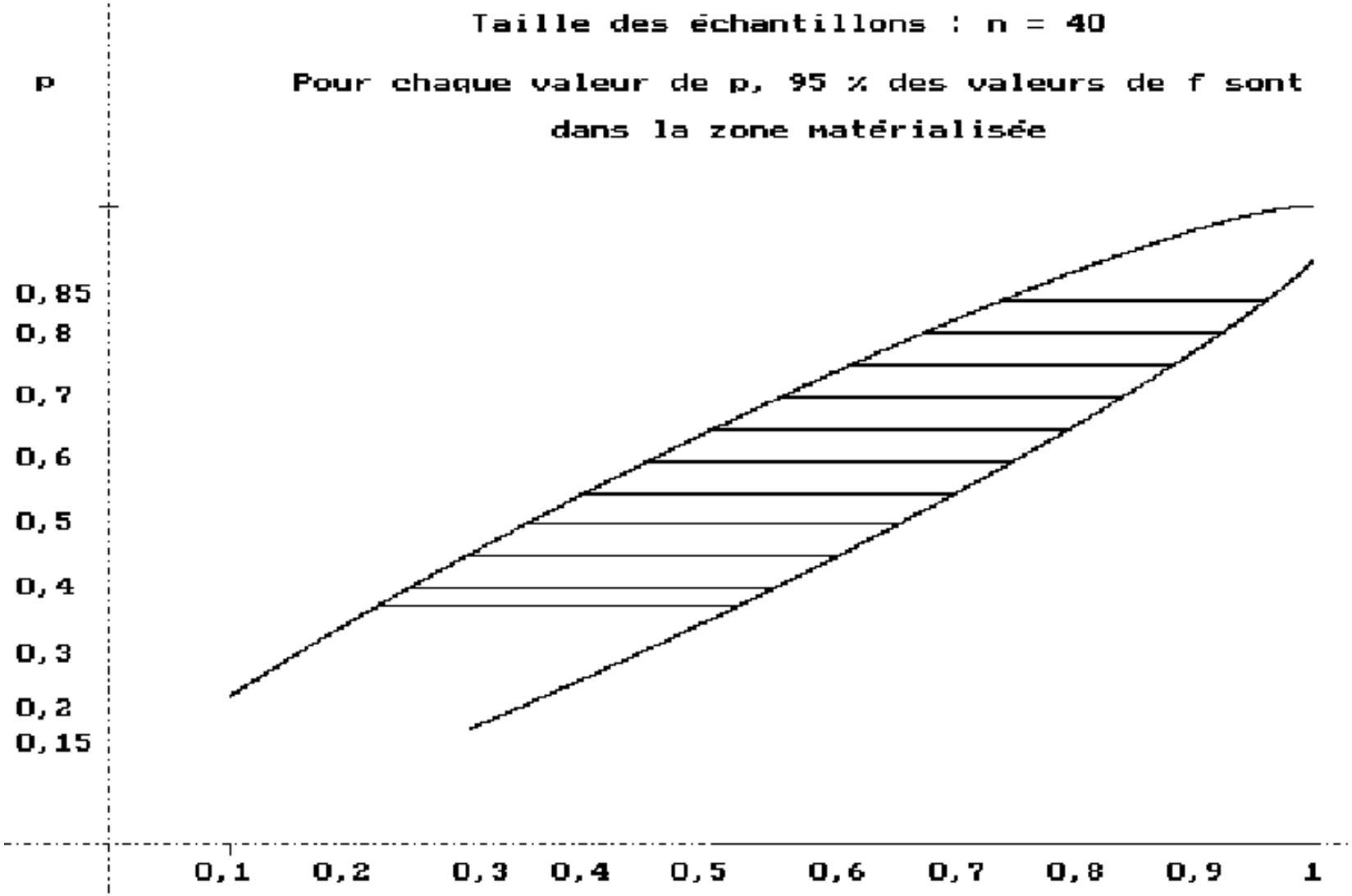
Si par exemple le tirage de 1000 boules avec remise fournit une fréquence de boules noires égale à 0,47, on peut estimer avec plus de 95 % de confiance que la proportion p de boules noires dans l'urne est comprise entre 0,44 et 0,50.

Les sondages, par exemple, sont souvent pratiqués sur des échantillons d'environ 1 000 personnes.

- Construction d'un abaque

(*f* en abscisse, *p* en ordonnée) :

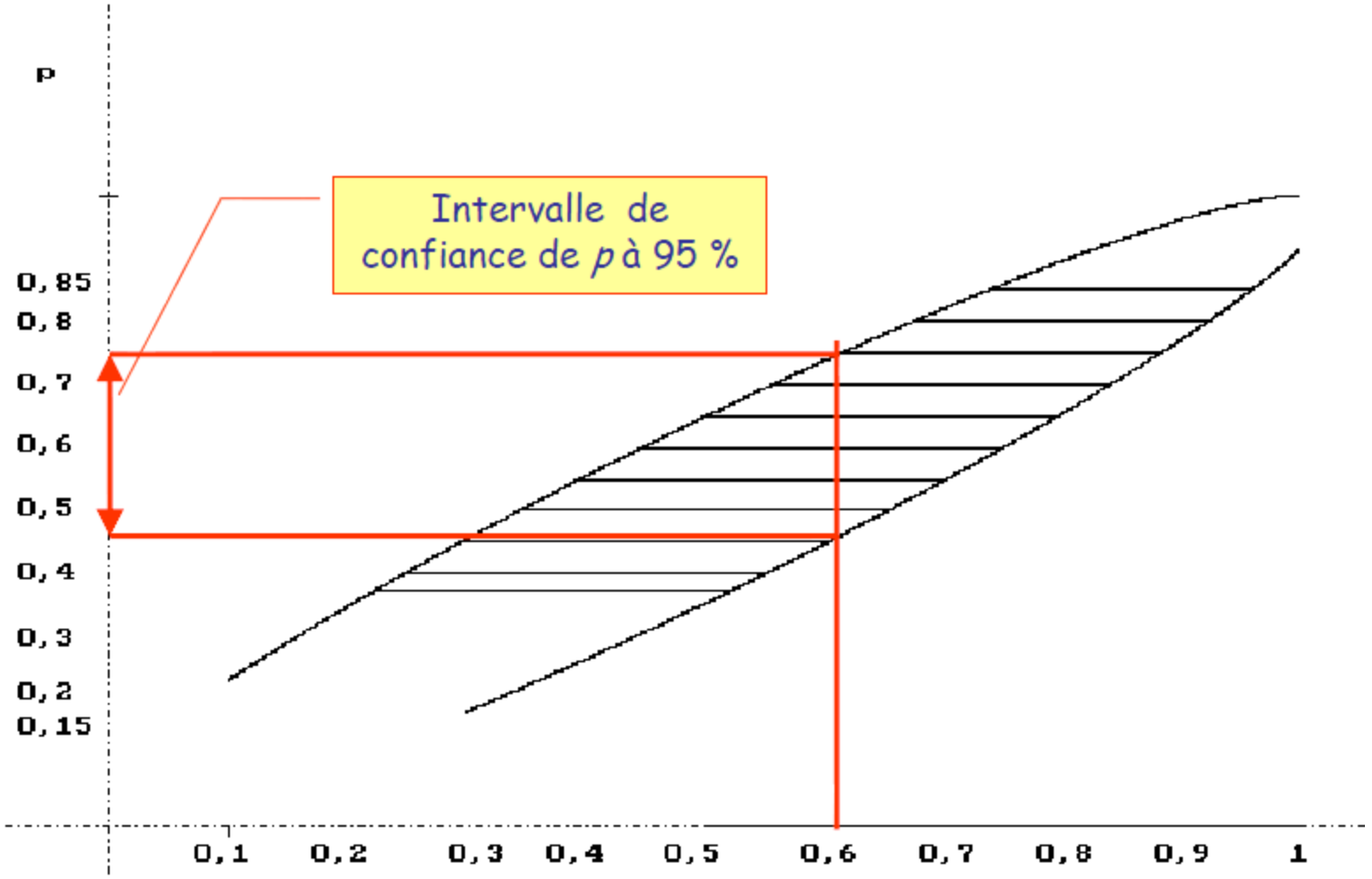



Taille des échantillons : n = 40
Pour chaque valeur de p, 95 % des valeurs de f sont dans la zone matérialisée
p
0,85
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,15
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1

On souhaite estimer la proportion p (inconnue) d'individus présentant une propriété donnée dans une population à partir d'un échantillon de taille 40.

Supposons que la propriété est observée dans l'échantillon avec une fréquence de 60 %.

On détermine les valeurs de p qui font en sorte que 0,6 appartienne à l'intervalle de fluctuation au seuil de 95 %, relatif aux échantillons de taille 40, associé à p.


p
Intervalle de confiance de p à 95 %
0,85
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,15
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1