Échantillonnage

académie Bordeaux
MINISTÈRE DE L'ÉDUCATION NATIONALE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Équipe Académique Mathématiques - 2011

Fluctuation des échantillons

Considérons une urne « de Bernoulli »
(la population) contenant une proportion *p* de boules blanches, dont on extrait *n* boules ;
la proportion de boules blanches dans le tirage (ou échantillon) est notée *f*.

p est connu

On note *X* la variable aléatoire correspondant au nombre de boules blanches dans un échantillon de taille *n*.

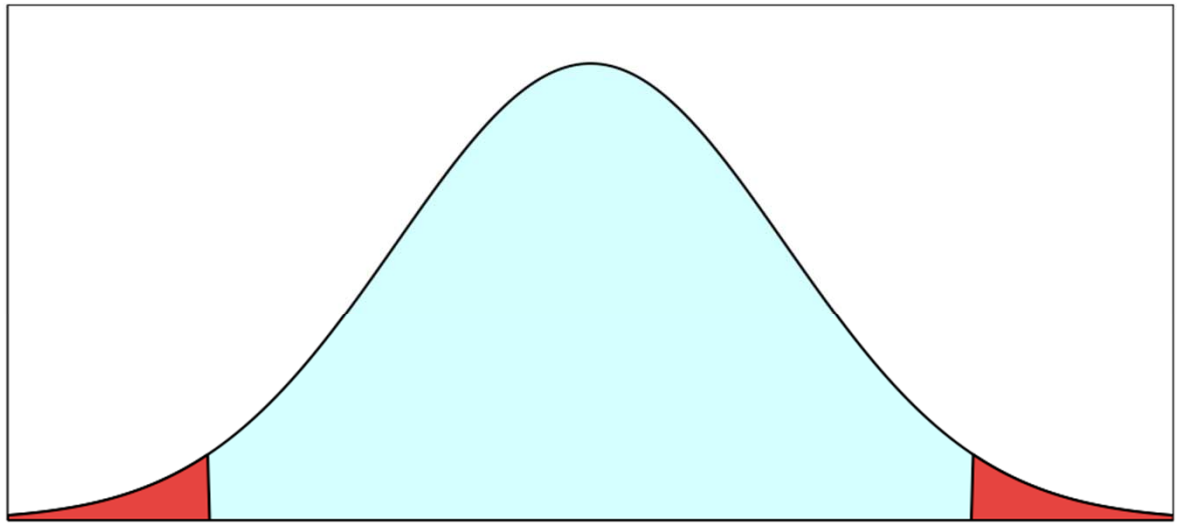
***X* suit la loi binomiale de paramètres *n* et *p*.**

Un raisonnement probabiliste, permet d'approcher la variable aléatoire $f = X/n$

par la loi normale $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Sous certaines conditions sur *n* et *p*:

$n \geq 25$ et *p* compris entre 0,2 et 0,8.



Si *T* suit la loi normale d'espérance p et d'écart type σ , on a :

$$P(p - 1{,}96\,\sigma \leq T \leq p + 1{,}96\,\sigma) \approx 0{,}95.$$

La probabilité que *f* appartienne à l'intervalle

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

est environ égale à 0,95.

C'est l'intervalle de fluctuation au seuil de 95 %

Remarquons que le produit $p(1-p)$ est toujours inférieur à 1/4. On élargit donc l'intervalle de fluctuation au seuil de 95% à :

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

Les situations étudiées correspondent aux cas où les nombres n et p vérifient $n \geq 25$ et p compris entre 0,2 et 0,8.

La connaissance de ces conditions n'est pas exigible. La formule de l'intervalle est donnée.

Cf. Algorithme et fichier tableur

Prise de décision

Un médecin de la santé publique s'interroge sur la proportion de patients souffrant d'hypertension dans sa commune.
Une récente étude dans des populations semblables indique une proportion égale à 23%.

Il étudie un échantillon de taille 1000 et calcule la fréquence *f* obtenue.

f appartient-elle à l'intervalle de fluctuation au seuil de 95% ?

$$[0,20\,;0,26]$$

Si la réponse est non

il rejette l'hypothèse $p = 0,23$ avec un risque d'erreur de 5%.

Si la réponse est oui

l'hypothèse est acceptable sans connaître le risque d'erreur.

Contrôle de qualité industrielle

Dans une usine automobile, on contrôle les défauts de peinture de type « grains ponctuels sur le capot ».
Lorsque le processus est sous contrôle, on a 20 % de ce type de défauts.
Lors du contrôle aléatoire de 50 véhicules, on observe 26 % de défauts.
Que faut-il en penser ?

On est ici dans une situation de test unilatéral ; l'intervalle de fluctuation ne fournit une réponse que pour un test bilatéral.

Estimation ou « fourchette de sondage »

Si *p* est inconnu mais que l'on procède à un tirage donnant une valeur de *f* ; on peut dire que dans environ 95% des tirages on a :

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

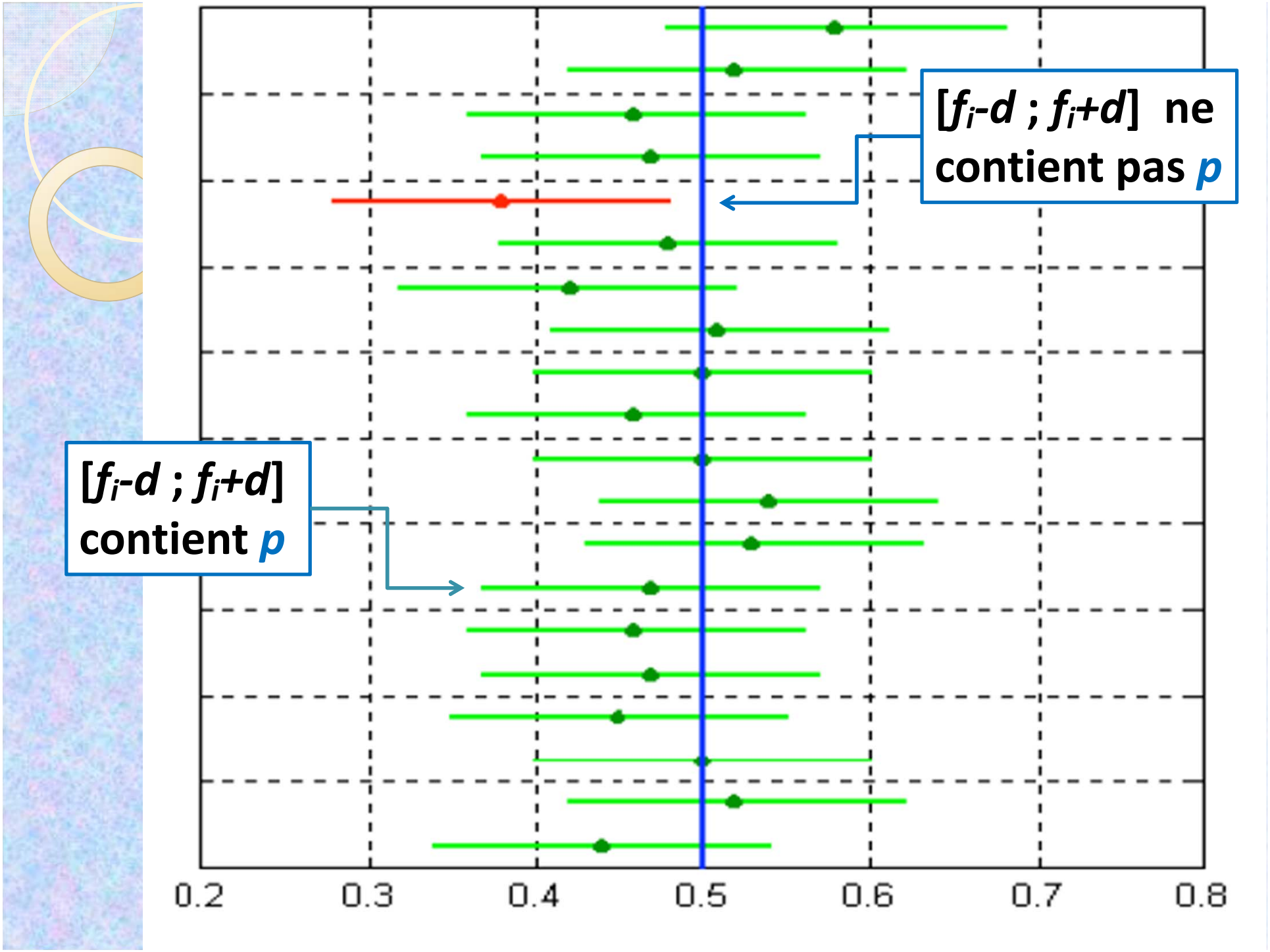
Cet intervalle est appelé intervalle de confiance ou fourchette de sondage.

Si on calcule l'intervalle de confiance précédent pour 100 sondages indépendants de taille n, on peut s'attendre à ce qu'environ :

95 contiennent la vraie valeur de p

5 ne la contiennent pas.

Ceci est illustré dans la figure suivante.

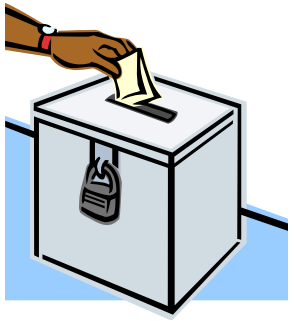

$[f_i-d\ ;\ f_i+d]$ ne contient pas p
$[f_i-d\ ;\ f_i+d]$ contient p
0.2
0.3
0.4
0.5
0.6
0.7
0.8

Lors du premier tour des élections présidentielles en 2002, le dernier sondage publié par l'institut B.V.A., effectué sur 1000 électeurs le vendredi 19/04/02 prévoyait :

Jacques Chirac	**19%**
Lionel Jospin	**18%**
Jean-Marie Le Pen	**14%**

La surprise a été grande le dimanche 21/04/02 au vu des résultats, puisque Jean-Marie Le Pen figurait au second tour :

Jacques Chirac	**19,88%**
Lionel Jospin	**16,18%**
Jean-Marie Le Pen	**16,86%**



(extrait de la brochure Statistique et citoyenneté : IREM Paris-Nord)

1) Calculer les trois fourchettes de sondages à partir du sondage B.V.A. et les représenter sur un graphique. Peut-on « prévoir » l'ordre des candidats au premier tour de l'élection ?

Doit-on considérer que ce sondage était « faux » ?

2) Un autre sondage aurait-il pu prévoir le résultat de ces élections ?

On procède à une simulation d'un sondage de taille 1000 en utilisant les pourcentages obtenus à l'issue des résultats.

Pour cela on simule le tirage au hasard d'un nombre entre 1 et 100 :

Si le nombre tiré est compris entre 1 et 20 alors on convient qu'il s'agit d'un électeur de J. Chirac.

Si le nombre tiré est compris entre 20+1 =21 et 20+16=36 alors il s'agit d'un électeur de L. Jospin.

Si le nombre tiré est compris entre 36+1 =37 et 36+17=53 alors il s'agit d'un électeur de J-M Le Pen.

Observer plusieurs sondages de taille 1000.

Observez-vous des sondages analogues au dernier sondage B.V.A. qui donnait l'ordre : Chirac - Jospin - Le Pen ?

En conclusion :

- ***Intervalle de fluctuation :***

  On connaît p et on prend une décision concernant une population à partir de la fréquence f observée sur un échantillon.

  Si l'on admet qu'un quart de la population des girafes adultes a une taille supérieure à 5,5 m, peut-on dire qu'un échantillon de 64 girafes dont le tiers dépasse 5,5 m est « anormal » ?

  (Math'X seconde)

- ***Intervalle de confiance :***

  On ne connaît pas p mais on cherche à l'estimer à partir de la fréquence f observée sur un échantillon.

  Dans une grande ville la municipalité a recensé sur 400 foyers choisis aléatoirement 78 foyers propriétaires de chiens. Que peut-on en déduire pour la proportion de propriétaires de chiens dans cette ville ?

  (Math'X seconde)

En seconde seul l'intervalle de fluctuation est dans les contenus exigibles du programme.