Sondages

- Introduction à la notion de fourchette de sondage.
- Résultats de simulations.

Un problème de sondage simplifié

Une urne contient une proportion p de boules numérotées 1, les autres étant numérotées 0. On ne peut pas compter les boules et tout ce que l'on peut faire pour connaître p est de tirer des boules (tirages avec remise). Si on fait n tirages et qu'on recueille la fréquence de 1 obtenue, quelle information cela apporte sur p ?
On appellera ici sondage de taille n dans une urne l'expérience consistant à faire n tirages avec remise dans cette urne.

Le résultat d'un sondage est un échantillon de l'expérience qui consiste à tirer avec remise une boule dans une urne et à regarder son numéro. Une même boule pouvant être tirée plusieurs fois, on ne peut pas dire que cet échantillon est celui des couleurs d'un sous-ensemble des boules de l'urne. Dans la pratique réelle des sondages, où un tirage au hasard d'individus dans une population importante est matériellement impossible, on remplace ce tirage par le choix d'une sous-population, appelée échantillon de cette population ou échantillon représentatif de cette population.

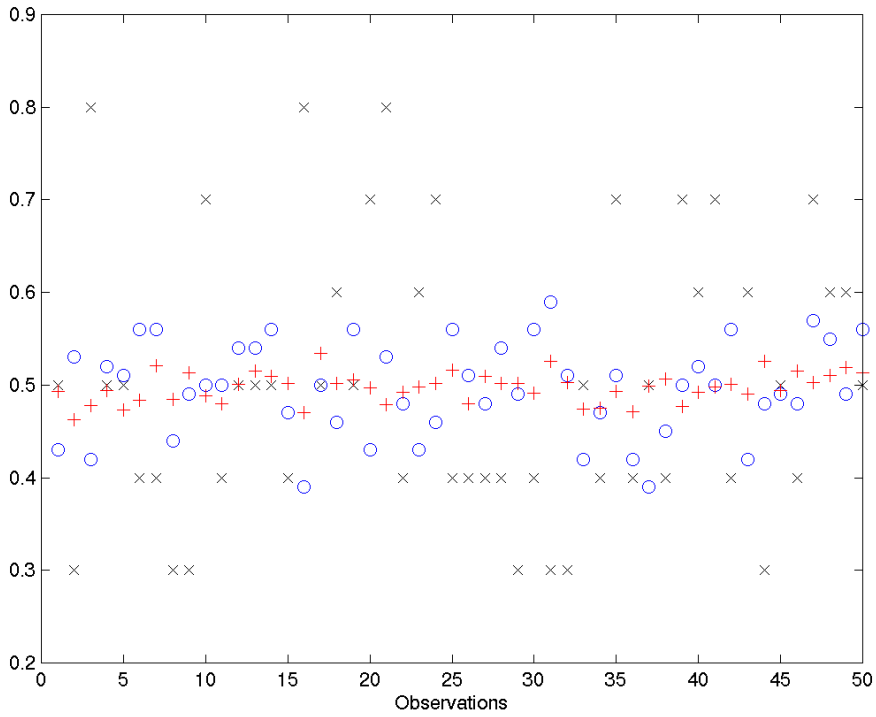
Cas où p est connu …

Dans un premier temps, nous allons, comme souvent en mathématiques, supposer le problème résolu : on fait un sondage dans une urne pour laquelle p est connu.
Un tel sondage peut aisément être simulé avec une liste de chiffres au hasard. Nous appellerons résultat d'un sondage la fréquence de 1 obtenue.
On simule 50 sondages de taille n pour diverses valeurs de n ; pour chaque valeur de n, on a donc une série de taille 50 de nombres compris entre 0 et 1, que nous résumons dans le tableau ci-dessous.

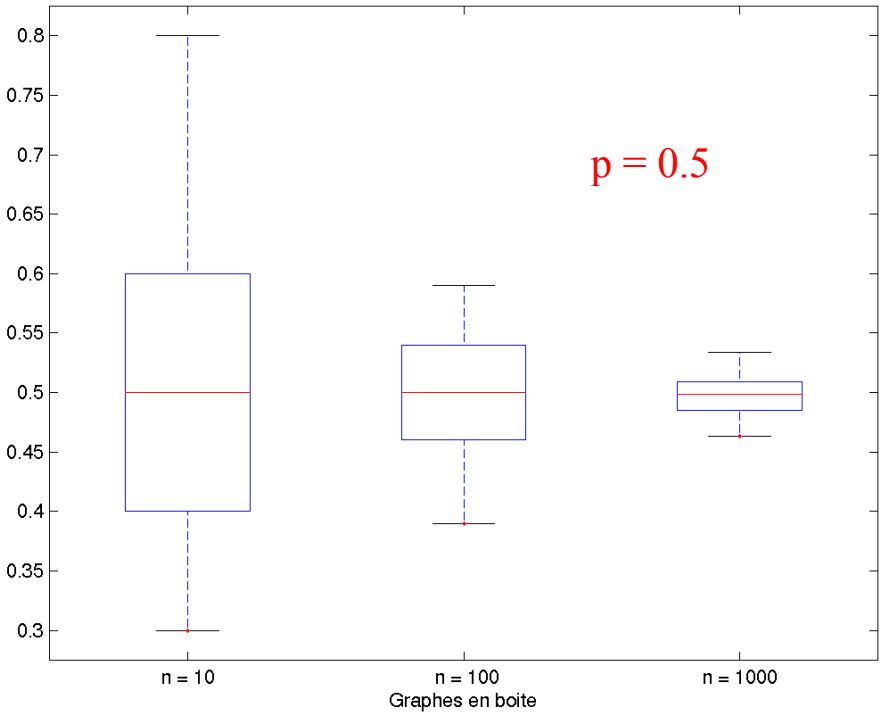
p = 0.5
50 sondages pour chaque valeur de n.

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
10	0.5	0.5	0.2	0.3	0.8	0.5	0.146
100	0.497	0.5	0.08	0.39	0.59	0.2	0.051
1000	0.497	0.498	0.024	0.463	0.534	0.071	0.016

L'intervalle interquartile et l'écart-type seront introduits en première. On pourra aussi utiliser les résultats de ces simulations en classe de première.



A l'aide du tableau ci-dessus, trouver à quelle taille de sondage correspondent les signes +, o, et x.



Les diagrammes en boîte ne sont pas au programme de seconde, mais sont au programme de première : on pourra aussi utiliser les résultats de ces simulations en classe de première. Dans ces diagrammes en boîtes, les ordonnées des segments extrêmes sont les valeurs extrêmes de la série.

On recommence ensuite cette étude en fixant p successivement à 0.8 puis à 0.3 :

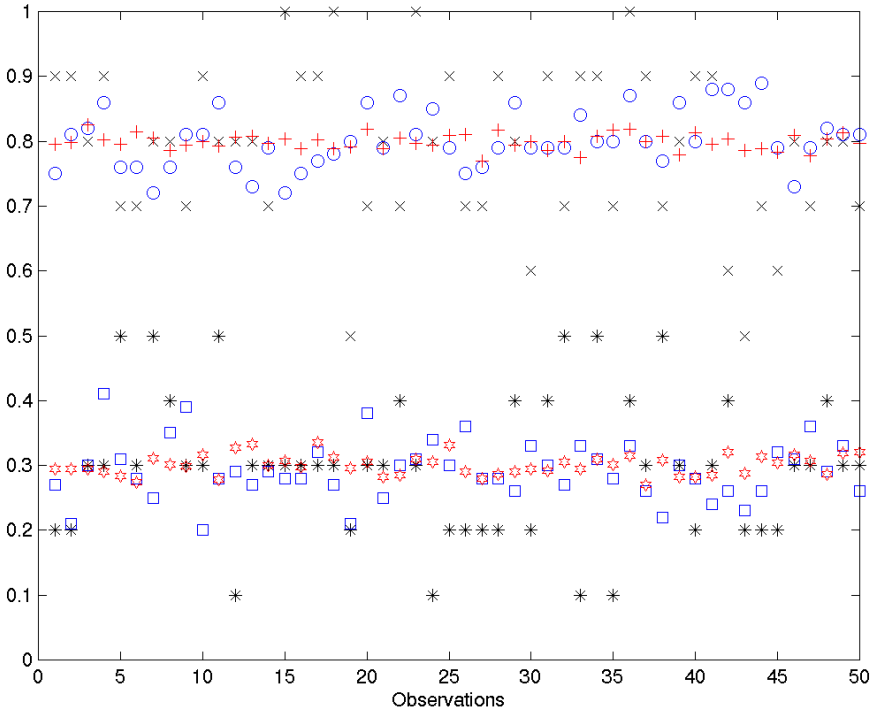
On obtient alors les résultats suivants :

p = 0.8
50 sondages pour chaque valeur de n.

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
10	0.792	0.8	0.2	0.5	1	0.5	0.123
100	0.802	0.8	0.07	0.72	0.89	0.17	0.045
1000	0.799	0.799	0.016	0.77	0.826	0.056	0.012

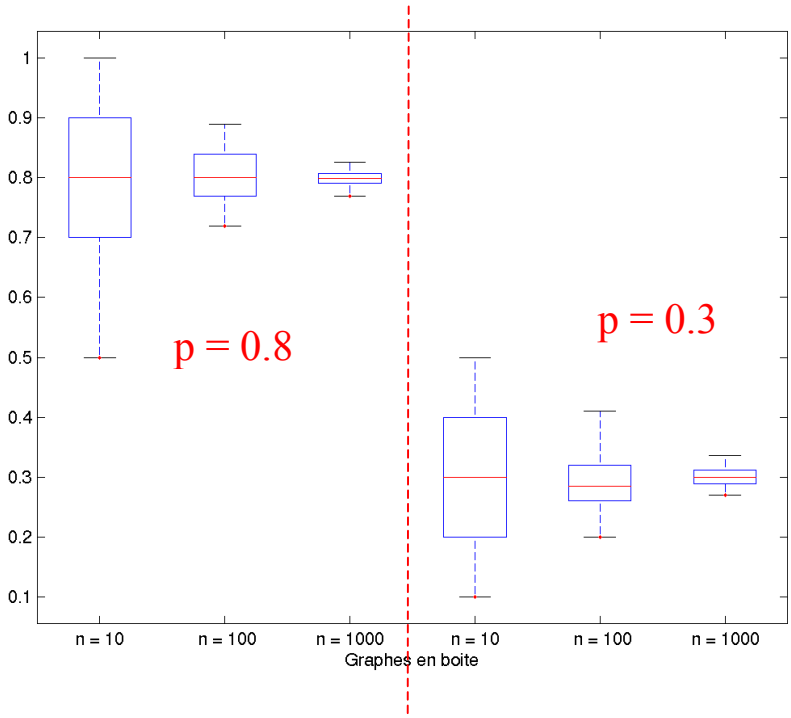
p = 0.3
50 sondages pour chaque valeur de n.

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
10	0.298	0.3	0.2	0.1	0.5	0.4	0.11
100	0.292	0.285	0.06	0.2	0.41	0.21	0.045
1000	0.301	0.299	0.023	0.27	0.336	0.066	0.016



p = 0.8 x ⟶ n = 10 o ⟶ n = 100 + ⟶ n = 1000

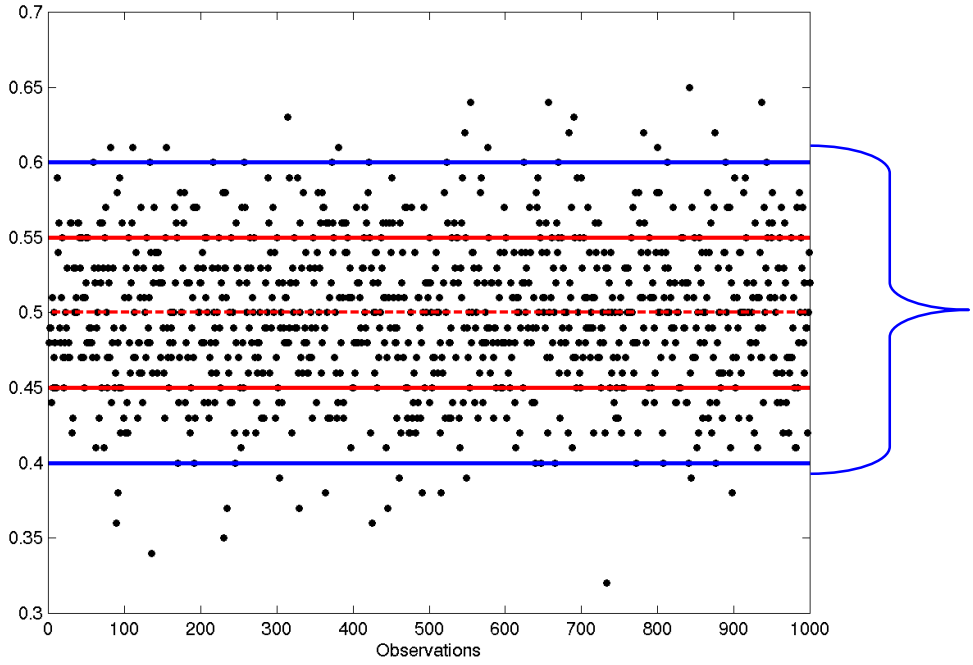
p = 0.3 * ⟶ n = 10 □ ⟶ n = 100 ✡ ⟶ n = 1000



On fait maintenant 1000 sondages de taille n = 100 pour p=0,5.

p = 0.5
1000 sondages de taille 100

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
100	0.499	0.5	0.06	0.32	0.65	0.33	0.05



Un point représente la fréquence des 1 pour un sondage de taille 100 ; p=0,5.

Quelle est la proportion de sondages dans la bande délimitée par les traits bleus ? Dans la bande délimitée par les traits rouges ?

On peut démontrer, dans le cadre de la théorie de probabilités, la " formule " suivante :

Si on fait un grand nombre de sondages de taille n,
environ 95 % d'entre eux vérifient :

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

Est-ce que les résultats pour les 1000 sondages ci-dessus sont cohérents avec cette " formule " ?

Cas où p est inconnu

Les deux propositions suivantes sont équivalentes :

f est dans l'intervalle $[p - \delta ; p + \delta]$

p est dans l'intervalle $[f - \delta ; f + \delta]$

où f est la proportion observée de 1 et p la proportion de 1 dans l'urne.

Si on fait un grand nombre de sondages de taille n,
environ 95 % d'entre eux vérifient :

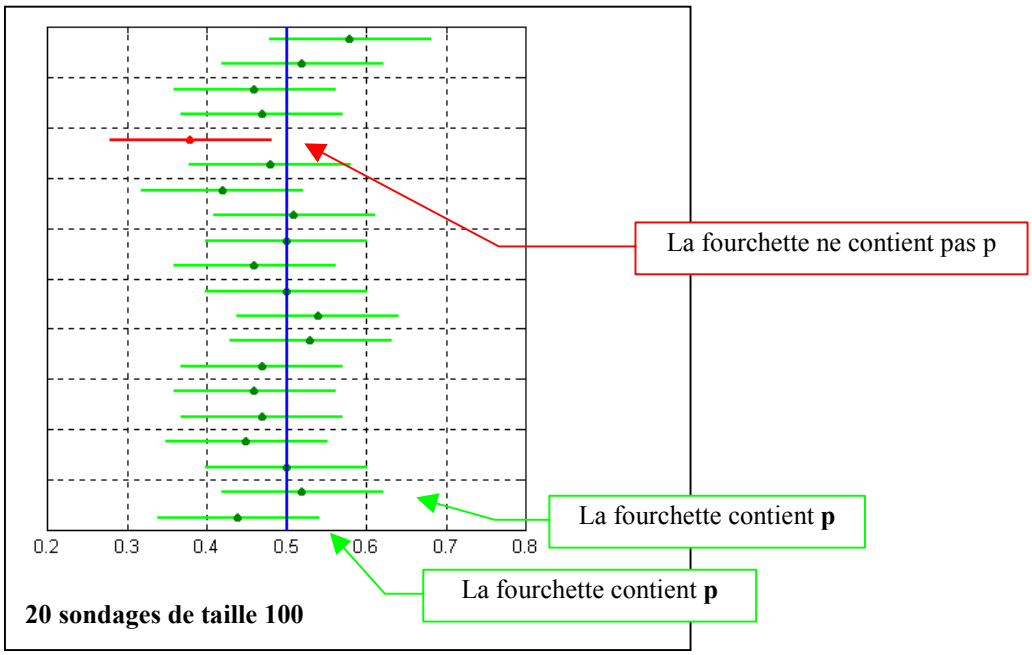
$$p \in \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right] .$$

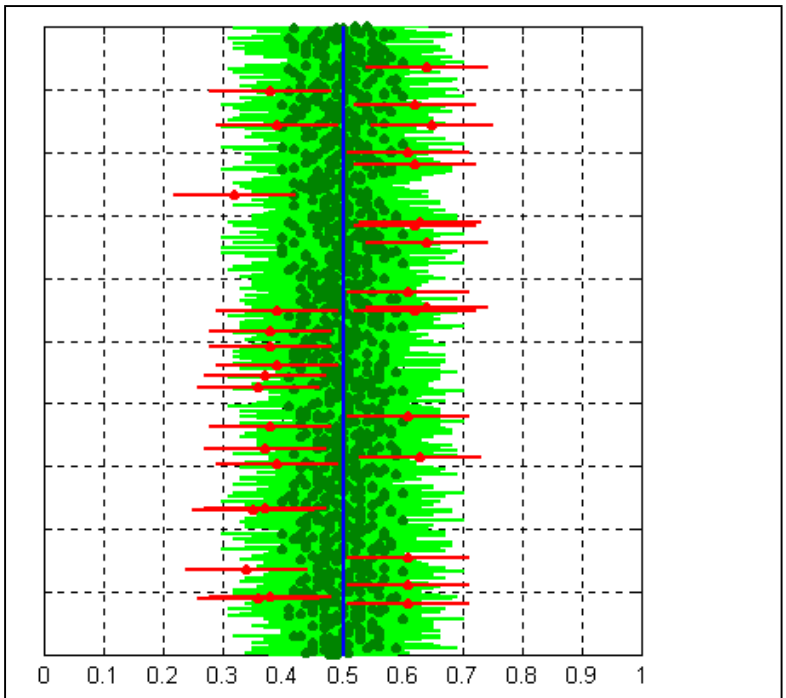
Pour un sondage de taille n dont le résultat est f, l'intervalle

$\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$ est appelé fourchette de sondage au niveau 0,95.

On dira aussi que f estime p avec une précision de $\frac{1}{\sqrt{n}}$, au niveau de confiance 0,95.

CONSTRUCTION DE FOURCHETTES AU NIVEAU 0,95





1000 sondages de taille 100

- 965 fourchettes (en vert) contiennent la valeur théorique p = 0.5.
- 35 fourchettes (en rouge) ne contiennent pas p.

APERÇU THEORIQUE

Soit S_n une variable aléatoire de loi binomiale B(n,p). Notons $R_n = \dfrac{S_n - np}{\sqrt{npq}}$.

Le théorème de Moivre énonce que la probabilité pour que R_n soit dans un intervalle [a,b] converge, lorsque n tend vers l'infini, vers l'intégrale :

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$$

Autrement dit, ce théorème énonce que la suite des variables R_n converge en loi vers la loi normale centrée réduite. La démonstration utilise essentiellement un développement asymptotique analogue à la formule de Stirling ainsi que la convergence d'une somme de Riemann vers une intégrale.
Notons Φ la fonction de répartition de la loi normale centrée réduite, soit :

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

On a : $\Phi(1,96) \approx 0,95$.
Sous les hypothèses (H) suivantes :

(H) $n > 30$; $np > 5$; $nq > 5$ où $q = 1 - p$

on approxime avec une très bonne précision la probabilité pour que R_n soit dans un intervalle [a,b] par sa limite donnée dans le théorème de Moivre.

On a donc :

(1) $\text{Prob}(-1,96 \le R_n \le 1,96) \approx 0,95$

En notant $F_n = \dfrac{S_n}{n}$ la fréquence des 1 dans une suite d'expériences de Bernoulli de paramètre p, on peut écrire cette égalité sous la forme :

(2) $\text{Prob}\left(p - 1,96\sqrt{\dfrac{pq}{n}} \le F_n \le p + 1,96\sqrt{\dfrac{pq}{n}} \right) \approx 0,95$.

On dit que l'intervalle $\left[p - 1,96\sqrt{\dfrac{pq}{n}} ; p + 1,96\sqrt{\dfrac{pq}{n}} \right]$ est l'intervalle de dispersion de F_n au niveau 0,95.

Mais (2) peut s'écrire :

(3) $\text{Prob}\left(p \in \left[F_n - 1,96\sqrt{\dfrac{pq}{n}} ; F_n + 1,96\sqrt{\dfrac{pq}{n}} \right] \right) \approx 0,95$.

Remarquons que pq est toujours inférieur à $\dfrac{1}{4}$. Donc de (3), on déduit :

(4) $\text{Prob}\left(p \in \left[F_n - \dfrac{1}{\sqrt{n}} ; F_n + \dfrac{1}{\sqrt{n}} \right] \right) \ge 0,95$.

En pratique, si la valeur observée f_n (fréquence de 1 dans l'échantillon) est comprise entre 0,3 et 0,5, on admet que (H) est vérifiée et on en déduit que $\sqrt{\frac{pq}{n}}$ est peu différent de $\frac{1}{2\sqrt{n}}$.

On dit que l'intervalle $\left[f_n - \frac{1}{\sqrt{n}} ; f_n + \frac{1}{\sqrt{n}}\right]$ est la fourchette de sondage de *p* au niveau 0,95.

Si on estime p par f_n, on dira que la précision est $\frac{1}{\sqrt{n}}$ avec un niveau de confiance 0,95. Si n est de l'ordre de 100 (respectivement 1000), la précision de l'estimation d'un pourcentage au niveau 0,95 est de l'ordre de 10 % (respectivement 3 %).

Si on remplace le niveau 0,95 par 0,90, il faut remplacer 1,96 par 1,65.

Si on remplace le niveau 0,95 par 0,99, il faut remplacer 1,96 par 3.

D'où :

- L'intervalle $\left[f_n - \frac{1,65}{2\sqrt{n}} ; f_n + \frac{1,65}{2\sqrt{n}}\right]$ est la fourchette de sondage de p au niveau 0,90.

  Au niveau de confiance 0,90, la précision sur p, lorsqu'on estime p par F_n, est $\frac{1,65}{2\sqrt{n}}$.
- L'intervalle $\left[f_n - \frac{3}{2\sqrt{n}} ; f_n + \frac{3}{2\sqrt{n}}\right]$ est la fourchette de sondage de p au niveau 0,99.

  Au niveau de confiance 0,90, la précision sur p, lorsqu'on estime p par F_n, est $\frac{3}{2\sqrt{n}}$.