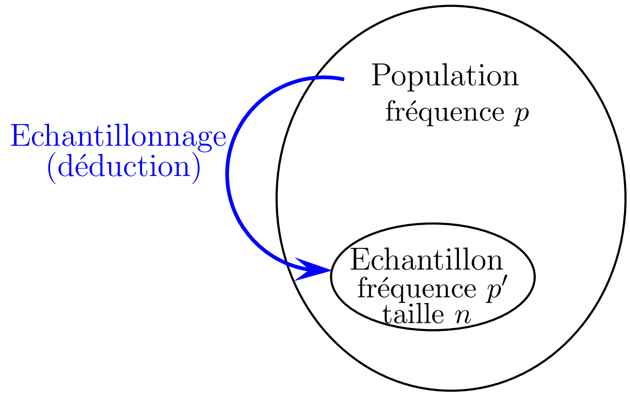
Echantillonnage - Estimation

I - Fluctuation d'échantillons

1) Position du problème

Dans une population donnée, on connaît la proportion p d'un caractère.

On constitue un échantillon de taille n à partir de la population : on prélève n fois, de façon identique et indépendante, un individu au hasard dans cette population.
On cherche alors à connaître, ou du moins à estimer, la proportion p' du caractère dans cet échantillon.



Exemple 1 : Il y a $p = 12\%$ de gauchers en France. Quelle est la proportion p' de gauchers dans un échantillon de $n = 30$ personnes (une classe de 30 élèves) ?

Exemple 2 : On lance une pièce bien équilibrée ; La probabilité d'obtenir Pile est $p = 0,5$. Sur $n = 100$ lancers, quelle est la proportion p' de Pile ?

2) Intervalle de fluctuation

La proportion p' dans l'échantillon n'est pas forcément égale à p ; la valeur de p' peut varier, ou **fluctuer**, d'un échantillon à un autre. On parle de **Fluctuation d'échantillonnage**.

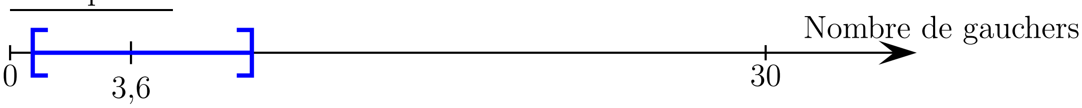
Exemple 1 : Dans une classe de 30 élèves on peut compter 3 gauchers, soit une proportion $p' = \dfrac{3}{30} = 10\%$, dans une autre "seulement" 1 gaucher soit $p' = \dfrac{1}{30} \simeq 3\%$, dans encore une autre 8 soit $p' = \dfrac{8}{30} \simeq 26\%$, ...

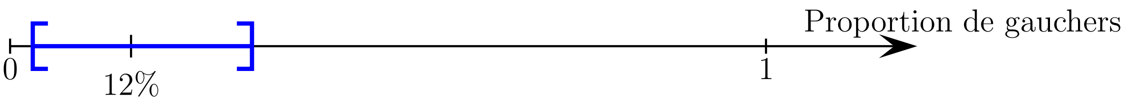
Exemple 2 : On lance une pièce bien équilibrée 100 fois, on obtient 54 fois "Pile", soit une proportion $p' = \dfrac{54}{100} = 0,54$, on recommence 100 lancers et on obtient cette fois 41 fois "Pile" soit une proportion $p' = \dfrac{41}{100} = 0,41$; on recommence une $3^{\text{ème}}$ fois ...

Bien que ces phénomènes soient aléatoires (le critère "gaucher" n'entre pas en compte dans la conception des classes...), on sait que, d'après la loi des grands nombres, plus la taille n des échantillons augmente, plus les proportions observées tendent vers une valeur limite qui est la probabilité $p = 12\%$ pour l'exemple 1 et $p = 0,5$ pour l'exemple 2.
Par contre, dans les exemples précédents, on sait aussi que même si le nombre de gauchers ou de Pile varie d'un échantillon à l'autre, il est quand même rare (c'est-à-dire la probabilité sera faible) d'avoir une classe que de gauchers, ou même avec plus de 15 gauchers, ou de même d'obtenir 100 "Pile" consécutifs,...

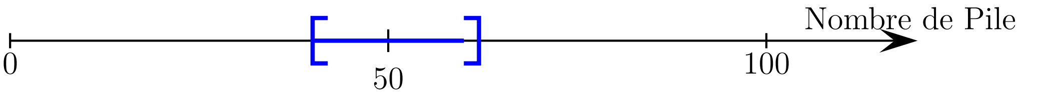
La notion d'intervalle de fluctuation permet de quantifier ce phénomène : c'est l'intervalle dans lequel la proportion observée varie "normalement" sous l'influence du hasard.

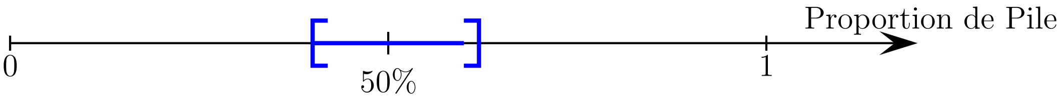
Exemple 1 :





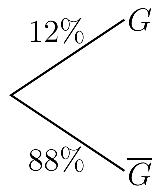
Exemple 2 :



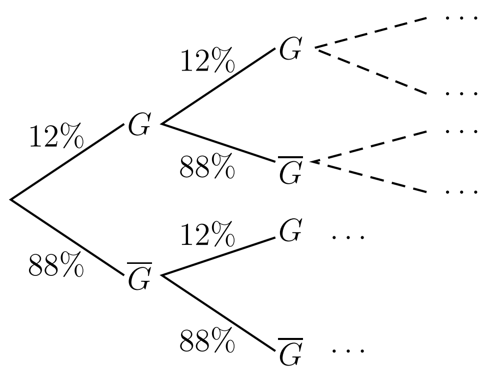


En pratique cet intervalle de variation, ou fluctuation, "normale" due à l'aléatoire est fixée comme celui dans lequel se trouve la proportion avec une probabilité d'environ 95%.

Pour constituer une classe de 30 élèves, on commence par en choisir un, la probabilité qu'il soit gaucher est de 12%



C'est une expérience de Bernoulli de paramètre $p = 0,12$ que l'on répète 30 fois.



La variable aléatoire X égale au nombre de gauchers suit donc la loi binomiale de parmètres $n = 30$ et $p = 0,12$.

On suppose alors qu'on peut approximer la loi binomiale par la loi normale $\mathcal{N}(\mu, \sigma)$ avec donc $\mu = np$ et $\sigma = \sqrt{np(1-p)}$.

Pour Y une variable aléatoire qui suit cette loi normale on sait que $P(\mu - 2\sigma \leqslant Y \leqslant \mu + 2\sigma) \simeq 0,95$, ou, un peu plus précisément, $P(\mu - 1,96\sigma \leqslant Y \leqslant \mu + 1,96\sigma) \simeq 0,95$.

La proportion $\dfrac{Y}{n}$ est alors dans l'intervalle $\left[\dfrac{\mu - 1,96\sigma}{n} ; \dfrac{\mu + 1,96\sigma}{n}\right]$ soit, avec $\mu = np$ et $\sigma = \sqrt{np(1-p)}$, l'intervalle $\left[\dfrac{np - 1,96\sqrt{np(1-p)}}{n} ; \dfrac{np + 1,96\sqrt{np(1-p)}}{n}\right] = \left[p - \dfrac{\sqrt{np(1-p)}}{n} ; p + \dfrac{\sqrt{np(1-p)}}{n}\right]$ ou encore

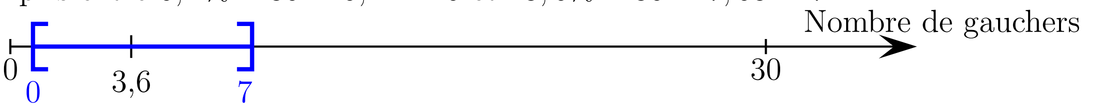
Propriété *L'intervalle de fluctuation asymptotique au seuil d'environ 95% d'une proportion obtenue sur un échantillon de taille n est :*

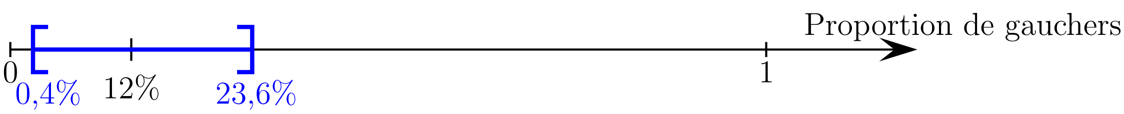
$$\left[p - 1,96\sqrt{\frac{p(1-p)}{n}} \; ; \; p + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$$

Exemple 1 : La proportion de gauchers parmi $n = 30$ personnes prises au hasard dans une population où il y a $p = 12\%$ de gauchers est, avec une probabilité d'environ 95%, dans l'intervalle

$$\left[p - 1,96\sqrt{\frac{p(1-p)}{n}}\ ;\ p + 1,96\sqrt{\frac{p(1-p)}{n}}\right] \simeq [0,12 - 0,116; 0,12 + 0,116] \simeq [0,004; 0,236]$$

La proportion est donc comprise entre 0,4% et 23,6%, soit, sur 30 personnes, un nombre de gauchers compris entre $0,4\% \times 30 = 0,12 \simeq 0$ et $23,6\% \times 30 = 7,08 \simeq 7$.

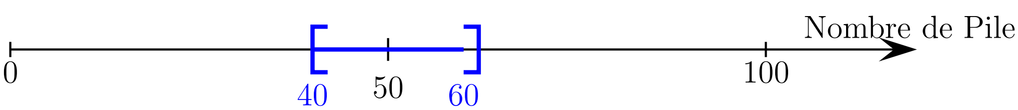


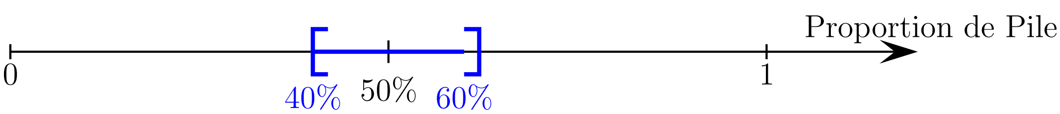


Exemple 2 : La proportion de Pile obtenus sur $n = 100$ lancers d'une pièce non truquée, donc $p = \frac{1}{2} = 50\%$ est, au seuil de 95%, dans l'intervalle

$$\left[p - 1,96\sqrt{\frac{p(1-p)}{n}}\ ;\ p + 1,96\sqrt{\frac{p(1-p)}{n}}\right] = [0,5 - 0,098; 0,5 + 0,098] \simeq [0,40; 0,60]$$

La proportion est donc comprise entre environ 40% et 60%, et le nombre de Pile entre 40 et 60.





Exemple : Avec les données de l'exemple précédent, l'intervalle de fluctuation approché au seuil de 95% est alors :

$$\left[p - \frac{1}{\sqrt{n}}\ ;\ p + \frac{1}{\sqrt{n}}\right] = \left[0,5 - \frac{1}{\sqrt{100}}\ ;\ 0,5 + \frac{1}{\sqrt{100}}\right] \simeq [0,4\ ;\ 0,6]$$

En comparant avec les résultats obtenus précédemment pour l'intervalle de fluctuation au seuil de 95%, on commet une erreur relative de seulement 0,2%=0,002 en utilisant cette formule approchée.

Exercice 1 Dimensionnement d'une cantine

Dans un établissement de 3000 personnes, chaque personne peut ou non, librement, manger à la cantine chaque jour. En moyenne 65% des personnes y mangent.

On souhaite estimer le nombre de places assises dans la cantine telle manière que toutes les personnes aient une place pour manger. Par ailleurs, par soucis d'économie, on souhaite aussi que le nombre de places prévues soit minimal.

Combien de places doit-on prévoir ?

3) Prise de décision

L'intervalle de fluctuation permet de juger, avec une probabilité d'erreur de 5%, de l'hypothèse de départ : la proportion dans la population complète est p.

Par exemple, on lance une pièce 100 et on obtient 35 Pile. Qu'en penser ?
On connaît l'intervalle de fluctuation pour une pièce non truquée, donc $p = 0,5$, qui est $[40; 60]$.

Le nombre observé est certes aléatoire, mais 35 est alors trop faible, trop éloigné de la valeur attendue de 50 si la pièce était équilibrée, car $35 < 40$.

L'intervalle de fluctuation permet ici de remettre en cause l'hypothèse : la pièce est ien équilibrée.

Propriété *Au seuil de 5% (d'erreur), si la proportion observée p' appartient à l'intervalle de fluctuation on accepte l'hypothèse selon laquelle la proportion est p dans la population, sinon on rejette cette hypothèse.*

Exercice 2 **Contrôle qualité**

Dans une usine, le responsable de la fabrication affirme que la proportion de produits défectueux fabriqués est de 20%.

Sur la chaîne de fabrication on a prélevé au hasard 72 produits, et on a constaté que 24 d'entre eux étaient défectueux.

1. Quelle est la proportion de produits défectueux dans l'échantillon prélevé ?
2. Que penser de l'affirmation du responsable de la fabrication ?

Exercice 3 **Influence du climat sur la couleur des yeux**

En France, la proportion de personnes ayant les yeux bleus est de 31%.

Dans une grande ville française, au micro-climat particulièrement ensoleillé, sur 50 personnes rencontrées au hasard, on a recensé 10 personnes ayant les yeux bleus.

1. Déterminer l'intervalle de fluctuation à 95% de la proportion de personnes ayant les yeux bleus dans un échantillon de 50 personnes.
2. Peut-on attribuer au micro-climat une influence spécifique sur la couleur des yeux ?

Exercice 4 **Influence d'une usine à proximité**

Selon l'Institut national des études démographiques (INED), il naît normalement 105 garçons pour 100 filles, soit une proportion de garçons $p = \dfrac{105}{205} \simeq 0,51$.

Aux abords d'une ville est venue s'implanter, il y a 20 ans, une usine chimique. La toxicité des substances manipulées et produites par cette usine est depuis grandement source de polémique.

Dans la maternité de cette ville, sont nés ces cinq dernières années 693 enfants, dont seulement 332 garçons. Les opposants à cette usine citent cette faible quantité de naissances de garçons comme une conséquence néfaste de l'exploitation de cette usine.

Ont-ils raison ?

Exercice 5 **Parité homme/femme**

Deux entreprises A et B recrutent leur personnel dans un bassin d'emploi où il y a autant d'hommes que de femmes. L'entreprise A emploie 60 personnes dont 26 femmes, tandis que l'entreprise B emploie 1050 personnes dont 480 femmes.

1. Calculer les proportions de femmes employées dans chaque entreprise.

   Laquelle de ces deux entreprises semble au mieux respecter la parité homme-femme ?
2. Déterminer pour chaque entreprise l'intervalle de fluctuation au seuil de 95 % de la proportion de femmes employées.

   Les deux entreprises respectent-elles la parité au seuil d'erreur de 5 % ?

II - Estimation

L'estimation, ou inférence, statistique consiste à essayer de déterminer les caractéristiques d'une population en ne connaissant des informations que sur un échantillon la composant.

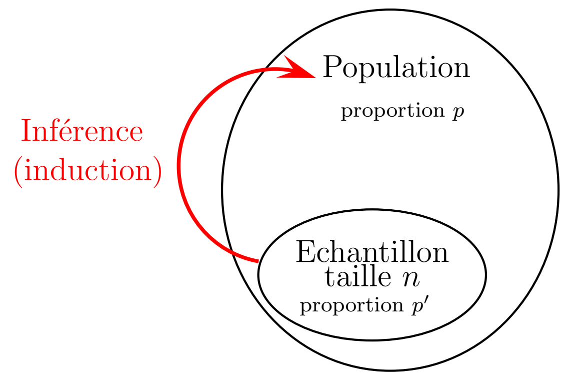
Un des exemples les plus médiatisés de nos jours est celui de sondage : en interrogeant un faible nombre de personnes sur leur intention de vote, on souhaite obtenir une information sur les intentions de vote de la population constituée par tous les électeurs.

Le journaliste et statisticien américain Georges Gallup a réussi à prédire en 1936 l'élection de Franklin Roosevelt contre Alfred Landon : les instituts de sondage étaient nés.

1) Position du problème

On connaît la proportion p' d'un caractère dans un échantillon aléatoire de la population complète.

On souhaite estimer la proportion p de ce caractère dans toute la population.



2) Intervalle de confiance

Propriété *L'intervalle* $\left[\ p'-1,96\sqrt{\dfrac{p'(1-p')}{n}}\ ;\ p'+1,96\sqrt{\dfrac{p'(1-p')}{n}}\ \right]$ *contient la proportion p avec une probabilité d'environ 95%.*

*Cet intervalle s'appelle l'***intervalle de confiance, au niveau de confiance de 95%***, de la proportion p dans la population.*

Exemple : Dans un village, lors d'un sondage effectué un mois avant le scrutin auprès de 200 personnes choisies de façon aléatoire, 109 personnes se déclarent favorables au candidat A.

La proportion d'électeurs favorables dans l'échantillon sondé est : $p'=\dfrac{109}{200}=54,5\,\%$.

L'intervalle de confiance au niveau de 95 % de la proportion p d'électeurs qui vont voter pour le candidat A est :

$$I=\left[\ p'-1,96\sqrt{\frac{p'(1-p')}{n}}\ ;\ p'+1,96\sqrt{\frac{p'(1-p')}{n}}\ \right]\simeq[0,545-0,069;0,545+0,069]\simeq[0,476;0,614]$$

On peut donc estimer, avec un niveau de confiance de 95 %, à partir du sondage effectué sur 200 personnes, que le score du candidat A aux prochaines élections sera dans la fourchette $\Big[47,6\%\,;\,61,4\%\Big]$.

En particulier, à partir de ce sondage, le candidat A ne peut pas en conclure qu'il sera élu car, au niveau de confiance de 95 %, il n'est pas exclu que la proportion de ses électeurs soit dans l'intervalle $[47,6\%\,;\,50\%[$, et donc inférieure à 50%.

Exercice 6 Avant le premier tour de l'élection présidentielle de 2002 un sondage IPSOS, réalisé auprès de 989 personnes constituant un échantillon national représentatif de la population française inscrite sur les listes électorales, annonçait les intentions de vote suivantes :

20 % pour J. Chirac, 18 % pour L. Jospin et 14 % pour J.M. Le Pen.

Les médias se préparaient donc pour un second tour entre J. Chirac et L. Jospin.

Le résultat réel des votes à ce premier tour a alors surpris bien des personnes ...

1. Déterminer, pour chaque candidat, l'intervalle de confiance au niveau de confiance de 0,95 de la proportion d'électeurs ayant eu l'intention de voter pour lui.
2. Les résultats à l'issue du premier tour ont été les suivants :
   19,88 % pour J. Chirac, 16,18 % pour L. Jospin et 16,86 % pour J.M. Le Pen.
   Ces pourcentages sont-ils en accord avec les calculs précédents ?
3. Pouvait-on au vu de ce sondage écarter avec un niveau de confiance de 0,95 l'un de ces trois candidats ?

Exercice 7 **Test de l'efficacité d'un médicament**

Un laboratoire pharmaceutique met en place un test pour estimer l'efficacité d'un nouveau médicament contre les migraines. Deux groupes de 125 patients souffrant de migraines, considérés comme des échantillons aléatoires, participent à ce test.

On administre aux patients du groupe A le nouveau médicament, tandis que les patients du groupe B reçoivent un placebo. Au bout de 4 jours de traitement, 73 patients du groupe A et 64 patients du groupe B déclarent ressentir une diminution de l'intensité de leurs migraines.

1. Déterminer les intervalles de confiance au niveau de confiance de 0,95 des proportions de patients déclarant ressentir une diminution de l'intensité de leurs migraines, dans chaque échantillon.
2. Les intervalles de confiance permettent-ils, au niveau de confiance 0,95, de considérer que le médicament est plus efficace que le placebo ?
3. Quelle devrait-être la taille minimale de chaque échantillon pour que, avec des proportions indentiques à celles observées précédemment, les résultats confirment l'efficacité du médicament, au niveau de confiance 0,95.

Exercice 8 **Commercialisation de deuc produits concurents**

Un magasin s'apprête à commercialiser deux modèles d'un même produit : le modèle A et le modèle B.

Une enquête préalable à la commande des produits par le magasin a montré que dans une ville 63 % des 400 personnes interrogées préfèrent le modèle A, et que dans une seconde ville, 69 % des 500 personnes interrogées préfèrent le modèle A.

Peut-on considérer, au niveau de confiance de 95 % qu'il y a une différence de préférence entre les personnes des deux villes ?

Quelle proportion de modèle A commanderiez-vous ?

Exercice 9 **Impact d'une campagne publicitaire**

Une entreprise possède 12% des parts de marché sur un de ses produits. Elle souhaite mieux se positionner sur ce produit et lance donc une vaste campagne publicitaire. Pour vérifier rapidement l'efficacité de cette campagne, elle procède à une étude auprès de 50 clients potentiels. 18 d'entre eux sont clients du produit de l'entreprise. Peut-on considérer que la campagne a eu un impact positif ?

Exercice 10 **Don de perception extra-sensorielle**

Dans une expérience de perception extra-sensorielle on demande à un sujet d'indiquer la couleur d'un jeton tiré aléatoirement dans un sac par un expérimentateur placé dans une autre pièce. Ni le sujet, ni l'expérimentateur ne connaissent la proportion de jetons de chaque couleur dans le sac.

Un sujet fait le test et identifie correctement la couleur de 32 jetons sur 50 essais.
A-t'il un don de perception extra-sensorielle ?