Chapitre 14

Échantillonnage et estimation

I. Introduction

On se situe dans deux domaines des statistiques : l'**échantillonnage** et l'**estimation**.
Ces deux domaines appartiennent au champ des **statistiques inférentielles** et ont des contextes d'application différents.

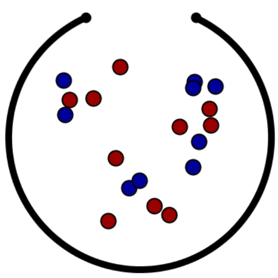
1) Identification de la situation

On s'intéresse à un caractère dans une population donnée dont la proportion est notée p.
Cette **proportion** sera, dans quelques cas, **connue** (échantillonnage) ou **supposée connue** (prise de décision) et, dans d'autres cas, **inconnue** (estimation).
Pour des raisons généralement économiques, on étudie le caractère, non pas sur la population entière, mais sur des échantillons de taille n extraits de cette population. Pour ce faire, on peut **prélever au hasard** des individus de cette population un par un avec remise. On parle d'**échantillons aléatoires non exhaustifs**.
Dans des situations telles qu'un sondage, un tel prélèvement est impensable : on pourrait interroger la même personne plusieurs fois. On prélève alors **successivement et sans remise** n individus de cette population. Si la taille de cet échantillon n'excède pas 10 % de la taille de la population entière, ce prélèvement ne modifie pas sensiblement la population. L'échantillon ainsi construit est assimilé à un échantillon aléatoire non exhaustif.

Modélisation

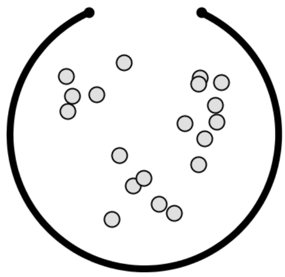
On considère deux urnes U_1 et U_2 contenant chacun un très grand nombre de boules, rouges ou bleues.

Dans l'urne U_1, **on connaît la proportion** ***p*** **de boules rouges**.



On procède à des tirages avec remise de n boules, et on observe la fréquence d'apparition d'une boule rouge.
Cette fréquence observée appartient « en général » à un « intervalle de fluctuation » de centre p, dont la longueur diminue avec n.
Cet intervalle est un « intervalle de fluctuation ».
On est alors dans le domaine de l'**échantillonnage** et de l'**intervalle de fluctuation**.

Dans l'urne U_2, **on ignore la proportion de boules rouges**.



En procédant à des tirages avec remise de n boules, on va essayer d'estimer la proportion p de boules rouges dans l'urne, proportion dont on n'a aucune idée *a priori*.
Cette estimation se fait au moyen d'un « intervalle de confiance ».
Cet intervalle dépend d'un coefficient, le « niveau de confiance », que l'on attribue à l'estimation.
On est alors dans le domaine de l'**estimation** et de l'**intervalle de confiance**.

2) Intervalle de confiance, intervalle de fluctuation

On s'intéresse à une population, dont on étudie un caractère particulier.

Échantillonnage	Estimation
On utilise un **intervalle de fluctuation** quand : • on **connaît la proportion *p*** de présence du caractère dans la population • on **fait une hypothèse sur la valeur de cette proportion** (on est dans le cadre de la **prise de décision**)	On utilise un **intervalle de confiance** quand : • on **ignore la valeur de la proportion *p*** de présence du caractère dans la population, et on ne formule pas d'hypothèse sur cette valeur.

Exemples :

- On dispose d'une pièce de monnaie.
  Comment décider qu'elle est « équilibrée » ou pas ?
  On va ici faire l'hypothèse que la fréquence d'apparition de « Pile », par exemple, est égale à 0,5, et on va tester cette hypothèse.
  On est dans une **situation d'échantillonnage**.
- Une usine fabrique des fusées de feux d'artifice. Sur 100 fusées choisies au hasard à l'issue du processus de fabrication et mises à feu, on trouve 12 fusées qui ne fonctionnent pas.
  Comment se faire une idée de la proportion des fusées défectueuses dans la production ?
  On est dans une **situation d'estimation** : on n'a, au départ, aucune idée de la valeur de la proportion étudiée dans la production.

3) Variable aléatoire fréquence

Propriété :
La variable aléatoire X qui à tout échantillon de taille n associe le nombre d'individus qui possèdent le caractère étudié, suit la loi binomiale de paramètre n et p.

Définition :
La variable aléatoire F qui à tout échantillon de taille n associe la fréquence f du caractère étudié dans cet échantillon est appelé **variable aléatoire fréquence** et elle est définie par :

$$\mathrm{F}=\frac{\mathrm{X}}{n}$$

Remarque :
Comme la variable aléatoire X prend les valeurs entières de 0 à n, la variable aléatoire fréquence F prend les valeurs 0, $\frac{1}{n}$, $\frac{2}{n}$, … , $\frac{n-1}{n}$, 1. Par conséquent, la variable aléatoire F est une variable aléatoire discrète (nombre fini de valeurs). Mais pour $n\geqslant 2$, cette variable aléatoire ne suit pas une loi binomiale (valeurs non toutes entières).

II. Échantillonnage

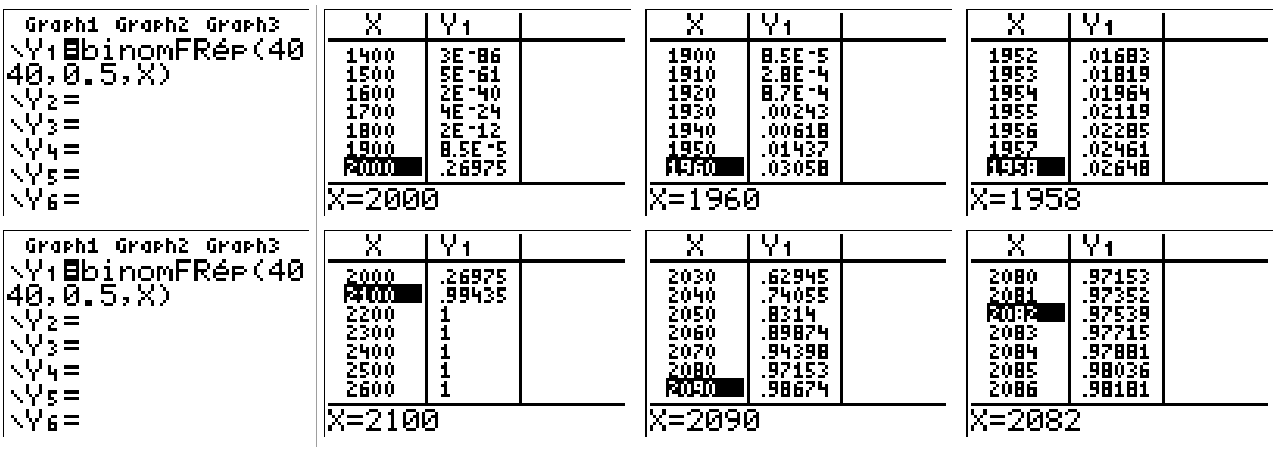
1) Cadre général

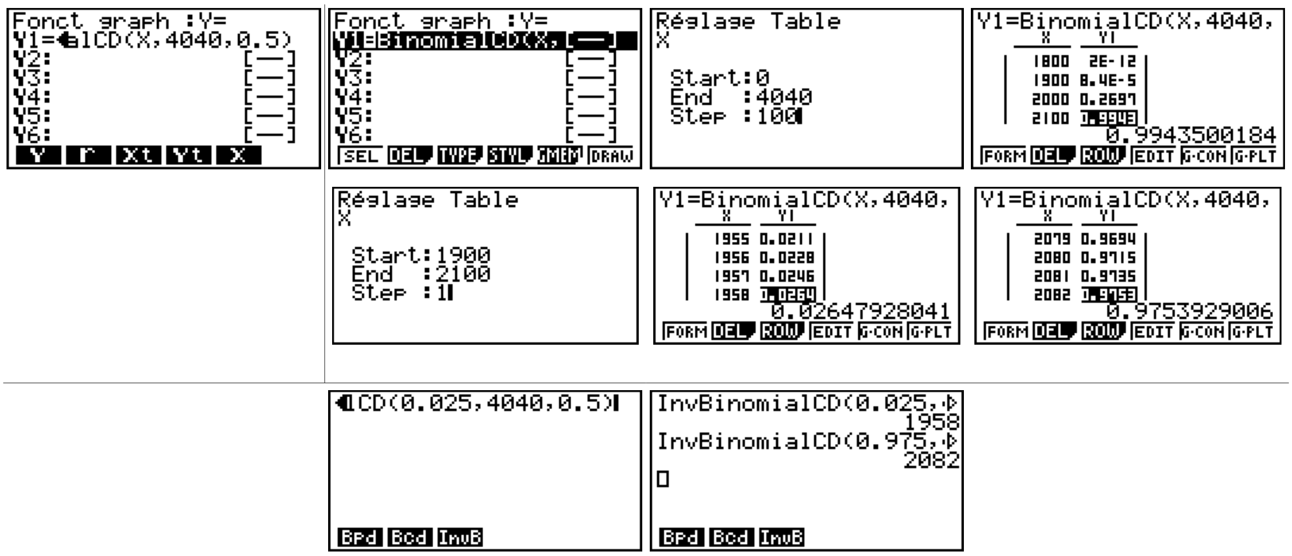
Définition :
X est une variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\,;p)$.
α est un nombre réel de l'intervalle $]0\,;1[$ et a, b des nombres réels.
Dire que $[a\,;b]$ est un **intervalle de fluctuation de X au seuil 1-α** signifie que :

$$\mathrm{P}(a \leqslant \mathrm{X} \leqslant b) \geqslant 1-\alpha$$

Exemple :
Un enfant a réalisé 4040 lancers d'une pièce de monnaie, et il a obtenu 2048 fois le résultat « Pile ». Peut-on considérer que la pièce est équilibrée ?

- En Seconde, on a vu que pour $n>25$ et $0{,}2<p<0{,}8$, l'intervalle $\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$ constitue un intervalle de fluctuation au seuil 95 %.
  p est la « proportion théorique » de résultats « Pile » obtenus sur un très grand nombre de lancers, on veut tester l'hypothèse selon laquelle la pièce est équilibrée : on fait l'hypothèse $p=0{,}5$.
  Pour l'échantillon considéré, on a $n=4040$, les conditions d'application sont donc réalisées, et on obtient l'intervalle $\left[0{,}5-\frac{1}{\sqrt{4040}}\,;0{,}5+\frac{1}{\sqrt{4040}}\right]$.
  En arrondissant à 10^{-4} près, on obtient l'intervalle de fluctuation : $\mathrm{I}_1=[0{,}4843\,;0{,}5157]$.
  La fréquence observée dans l'échantillon est donnée par :
  $$f=\frac{2048}{4040}\simeq 0{,}5096 \text{ à } 10^{-4} \text{ près.}$$
  Ainsi, on a $f \in \mathrm{I}_1$, donc on accepte l'hypothèse de pièce équilibrée.
- En Première, on suppose que la pièce est équilibrée, la variable aléatoire X qui dénombre les résultats « Pile » obtenus parmi les 4040 lancers réalisés suit une loi binomiale de paramètres $n=4040$ et $p=0{,}5$.
  L'intervalle de fluctuation à 95 % associé à la variable aléatoire X est l'intervalle $\left[\frac{a}{n}\,;\frac{b}{n}\right]$, où a est le plus petit entier tel que $p(\mathrm{X}\leqslant a)>0{,}025$ et b le plus petit entier vérifiant $p(\mathrm{X}\leqslant b)\geqslant 0{,}975$.
  On détermine les entiers a et b à l'aide de la calculatrice.





On obtient $a=1958$ et $b=2082$, qui donne un intervalle de fluctuation $\left[\frac{1958}{4040};\frac{2082}{4040}\right]$.

En arrondissant à 10^{-4} près, on obtient l'intervalle de fluctuation : $I_2=[0{,}4847;0{,}5153]$.

La fréquence observée $f\simeq 0{,}5069$, vérifie donc $f\in I_2$.

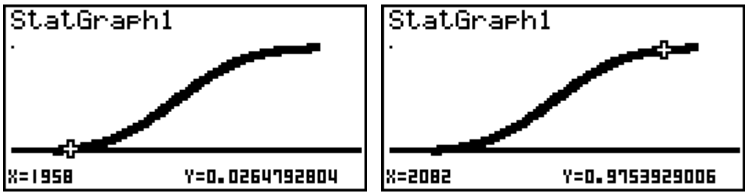
On est ici aussi conduit à accepter l'hypothèse de pièce équilibrée.

Remarque :

L'intervalle de fluctuation étudié ici est bilatéral.

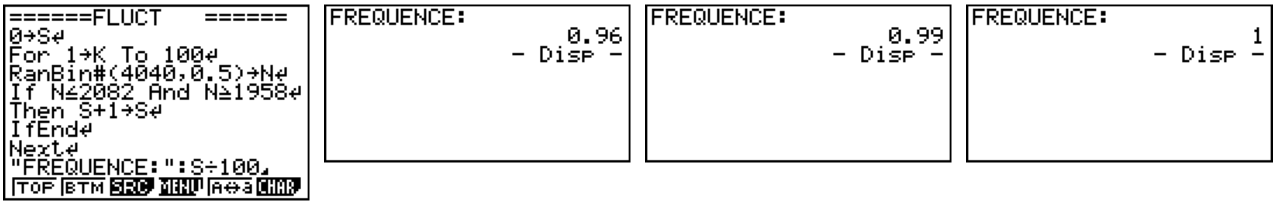
En effet, on est amené à rejeter l'hypothèse de pièce équilibrée dans deux situations symétriques : celle où on aurait observé « trop peu » de résultats « Pile » comme celle où on en aurait observé « trop ».

Dans le premier cas, la fréquence observée est du côté des réels inférieurs à ceux de l'intervalle de fluctuation, dans l'autre cas du côté des réels supérieurs à ceux de l'intervalle de fluctuation.



Intervalle de fluctuation au seuil de 95 % :

$$I=[\frac{1958}{4040};\frac{2082}{4040}]=[0.4847;0.5153]$$



2) *Intervalle de fluctuation asymptotique*

Définition :

Pour tout α dans $]0;1[$, un **intervalle de fluctuation asymptotique** de la variable aléatoire X_n au seuil $1-\alpha$ est un intervalle déterminé à partir de *p* et de *n* et qui contient X avec une probabilité d'autant plus proche de $1-\alpha$ que *n* est grand.

Remarques :

- Il n'existe donc pas un unique intervalle de fluctuation asymptotique à un seuil donné. L'expression « l'intervalle de fluctuation asymptotique » désigne l'intervalle considéré comme celui de référence en classe de Terminale.
- La probabilité considérée dans cette définition : $P(a \leqslant X \leqslant b)$ (*a* et *b* étant les bornes d'un éventuel intervalle de fluctuation asymptotique qui dépendent des paramètres *n* et *p*) n'est pas nécessairement égale à 1-α. Elle s'en approche quand la taille de l'échantillon devient de plus en plus grande.
- Quelles que soient les valeurs des paramètres *n* et *p*, on peut toujours construire l'intervalle de fluctuation déterminé à l'aide de la loi binomiale (étudié en classe de première). Cet intervalle dépend implicitement des paramètres *n* et *p*.

Propriété :

Si la variable aléatoire X_n suit la loi binomiale $\mathscr{B}(n\,;p)$ avec *p* dans l'intervalle $]0\,;1[$, alors pour tout nombre réel α de $]0\,;1[$,

$$\lim_{n \to +\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha$$

où $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$

et u_α désigne le nombre réel tel que $P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$ lorsque Z suit la loi normale $\mathscr{N}(0\,;1)$.

Démonstration :

On pose $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$.

D'après le théorème de Moivre-Laplace, $\lim_{n \to +\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = P(-u_\alpha \leqslant Z \leqslant u_\alpha)$.

Or $\quad P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = P\left(np - u_\alpha\sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha\sqrt{np(1-p)}\right)$

$$P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = P\left(p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

Donc $\lim_{n \to +\infty} P\left(\frac{X_n}{n} \in I_n\right) = P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$

Remarque :

On admet que pour $n \geqslant 30$, $np \geqslant 5$, $n(1-p) \geqslant 5$, on peut approcher $P\left(\frac{X_n}{n} \in I_n\right)$ par $1-\alpha$.

Définition :

X_n est une variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\,;p)$ avec $p \in\,]0\,;1[$ et α est un nombre réel de $]0\,;1[$.

L'intervalle I_n de la propriété ci-dessus est appelé **un intervalle de fluctuation asymptotique au seuil** $1-\alpha$ de la variable aléatoire fréquence $F_n = \frac{X_n}{n}$ qui, à tout échantillon de taille *n*, associe la fréquence obtenue f .

Exemples :

- Pour $\alpha = 0{,}05$ on a vu que $u_\alpha \simeq 1{,}96$.

  On prendra pour **intervalle de fluctuation au seuil 95 %** de $\frac{X_n}{n}$ l'intervalle :

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

  où p désigne la proportion dans la population.

- Pour $\alpha = 0{,}01$ on a vu que $u_\alpha \simeq 2{,}58$.

  On prendra pour **intervalle de fluctuation au seuil 99%** de $\frac{X_n}{n}$ l'intervalle :

  $$I_n = \left[p - 2{,}58 \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 2{,}58 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

  où p désigne la proportion dans la population.

Exemple :

En reprenant l'exemple traité précédemment.
Un enfant a réalisé 4040 lancers d'une pièce de monnaie, et il a obtenu 2048 fois le résultat « Pile ».
Peut-on considérer que la pièce est équilibrée ?

On peut prendre pour intervalle de fluctuation au seuil 95 % de $\frac{X_n}{n}$ l'intervalle J défini par :

$$J = \left[0{,}5 - 1{,}96 \frac{\sqrt{0{,}5(1-0{,}5)}}{\sqrt{4040}} ; 0{,}5 + 1{,}96 \frac{\sqrt{0{,}5(1-0{,}5)}}{\sqrt{4040}} \right].$$

Ainsi $J = [0{,}4846 ; 0{,}5154]$.
La fréquence observée $f \simeq 0{,}5069$, vérifie donc $f \in J$.
On est ici aussi conduit à accepter l'hypothèse de pièce équilibrée.

Calculatrice

```
PROGRAM:INTFLUC
:Prompt N,P
:If N≥30 et N*P≥
5 et N*(1-P)≥5
:Then
:Input "SEUIL=",
S
:FracNormale(0.5
+S/2)→U
:Disp "INT FLUCT
 ASYMPT",P-U*√(P
*(1-P)/N),P+U*√(
P*(1-P)/N)
:Else
:Disp "PAS DE CO
NDITION"
:End
```

```
prgmINTFLUC
N=?4040
P=?0.5
SEUIL=0.95
INT FLUCT ASYMPT
      .4845820223
      .5154179777
             Done
```

```
prgmINTFLUC
N=?4040
P=?0.5
SEUIL=0.99
INT FLUCT ASYMPT
      .4797373426
      .5202626574
             Done
```

```
prgmINTFLUC
N=?32
P=?0.1
PAS DE CONDITION
             Done
```

```
prgmINTFLUC
N=?28
P=?0.5
PAS DE CONDITION
             Done
```

```
======INTFLUCT======
"N="?→N
"P="?→P
If N≥30 And N×P≥5 And
 N×(1-P)≥5
Then
"SEUIL="?→S
InvNormCD(0.5+S÷2)→U
"INT FLUCT ASYMPTOT":
P-U×√(P×(1-P)÷N)◢
P+U×√(P×(1-P)÷N)◢
Else
"PAS DE CONDITIONS"
IfEnd
TOP BTM SRC MENU A↔a CHAR
```

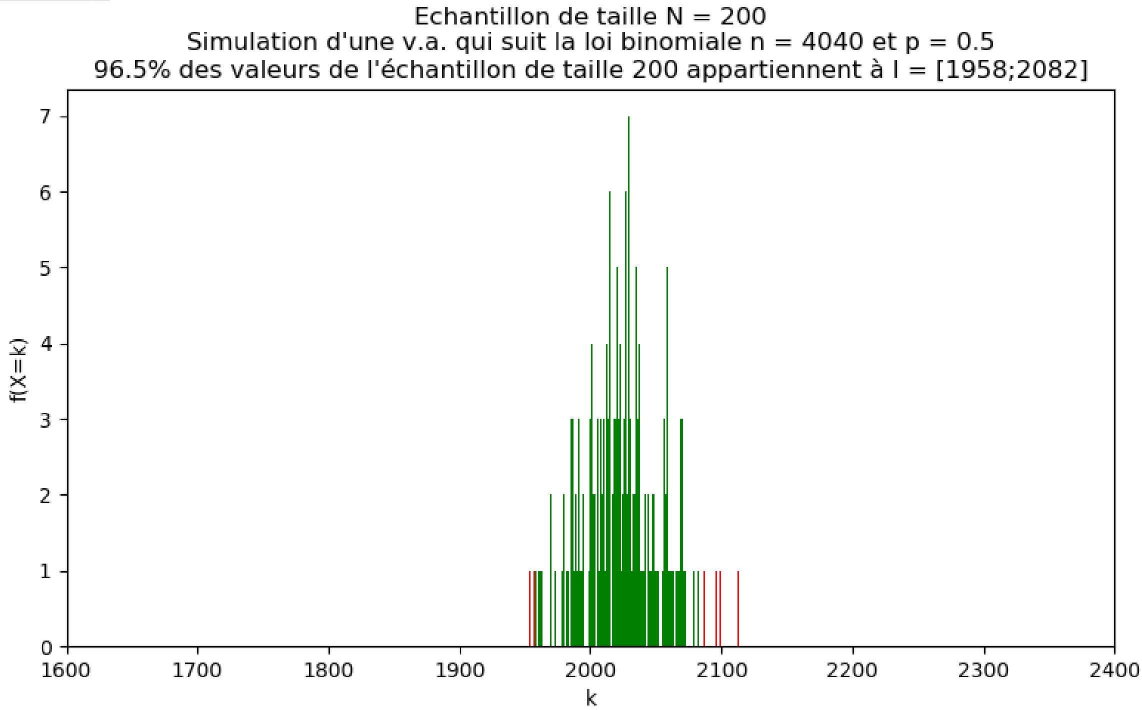
```
N=?
4040
P=?
0.5
SEUIL=?
0.95
INT FLUCT ASYMPTOT
          0.4845820223
          0.5154179777
                - Disp -
```

```
N=?
4040
P=?
0.5
SEUIL=?
0.99
INT FLUCT ASYMPTOT
          0.4797373426
          0.5202626574
                - Disp -
```

```
N=?
32
P=?
0.1
PAS DE CONDITIONS
```

```
N=?
28
P=?
0.5
PAS DE CONDITIONS
```

Simulation :



Remarque :

On montre facilement que pour $p \in]0;1[$: $\sqrt{p(1-p)} \leqslant \frac{1}{2}$ et ainsi que J_n est contenu dans l'intervalle de fluctuation au seuil de 95 % vu en seconde : $\left[p-\frac{1}{\sqrt{n}}; p+\frac{1}{\sqrt{n}}\right]$.

Propriété :

Soit X_n une variable aléatoire de loi $\mathscr{B}(n\,;p)$ et $F_n=\frac{X_n}{n}$.

Pour tout p de $]0;1[$, il existe $n_0 \in \mathbb{N}$ tel que si $n \geqslant n_0$,

$$P\left(p-\frac{1}{\sqrt{n}} \leqslant F_n \leqslant p+\frac{1}{\sqrt{n}}\right) \geqslant 0,95$$

3) Prise de décision

La **proportion** du caractère étudié 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 faite sur la proportion p.

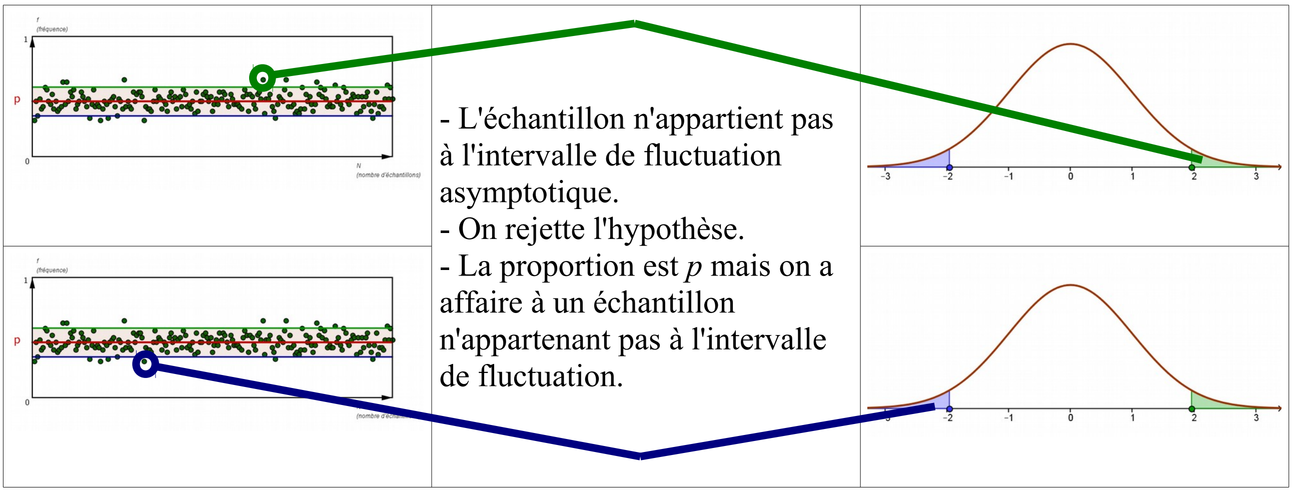
- on calcule la fréquence observée f du caractère étudié dans cet échantillon
- puis, si les conditions sur les paramètres n et p sont vérifiées ($n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$), on détermine l'intervalle de fluctuation asymptotique au seuil 0,95. Si ce n'est pas le cas, on peut déterminer l'intervalle de fluctuation étudié en classe de seconde ou de première
- On applique la règle de décision suivante :

  Règle de décision :
  - Si la fréquence observée f appartient à l'intervalle de fluctuation asymptotique au seuil 0,95, on accepte l'hypothèse faite sur la proportion p
  - Si la fréquence observée f n'appartient pas à l'intervalle de fluctuation asymptotique au seuil 0,95, on rejette l'hypothèse faite sur la proportion p avec un risque de 5 % de se tromper.

Remarques :

- Dans le cas où on accepte l'hypothèse faite sur la proportion p, le risque d'erreur n'est pas quantifié.
- Le risque de 5 % signifie que la probabilité (que l'on rejette, à tort, l'hypothèse faite sur la proportion p alors qu'elle est vraie) est **approximativement** égale à 0,05.
  C'est une probabilité conditionnelle.

Lorsque $f \notin$ I, on rejette l'hypothèse concernant p : on considère ainsi que la proportion au sein de la population n'est pas p.
Mais il existe la possibilité que notre échantillon fasse partie des 5 % qui n'appartiennent pas à I : il s'agit de la fluctuation d'échantillonnage.
C'est ainsi que le risque d'erreur lorsque l'on rejette l'hypothèse est de 5 %.



Exemple :

Dans un casino, il a été décidé que les « machines à sous » doivent être réglées sur une fréquence de gain du joueur de $g=0{,}06$.

Une fréquence inférieure est supposée faire « fuir le client », et une fréquence supérieure est susceptible de ruiner le casino.

Trois contrôleurs différents vérifient une même machine.

Le premier a joué 50 fois et gagné 2 fois, le second a joué 120 fois et gagné 14 fois, le troisième a joué 400 fois et gagné 30 fois.

En utilisant des intervalles de fluctuation au seuil de 95 %, examiner dans chaque cas la décision à prendre par le contrôleur, à savoir accepter ou rejeter l'hypothèse $g=0{,}06$.

<table>
<tr><td></td><td>1^{er} contrôleur
$n=50$, $p=0{,}06$,
$f=\frac{2}{50}=0{,}04$</td><td>2^e contrôleur
$n=120$, $p=0{,}06$,
$f=\frac{14}{120}\simeq 0{,}1167$</td><td>3^e contrôleur
$n=400$, $p=0{,}06$,
$f=\frac{30}{400}=0{,}075$</td></tr>
<tr><td rowspan="2">Seconde</td><td colspan="3">Si les conditions d'application sont vérifiées ($n\geqslant 25$ et $0{,}2<p<0{,}8$).
On observe si $f\in \mathrm{I}=\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$</td></tr>
<tr><td colspan="3">Conditions d'application :
$n\geqslant 25$ mais $p<0{,}2$.
On n'est pas dans les conditions d'application d'un intervalle de fluctuation.
Ce contrôle ne peut rien donner de probant en termes de prise de décision.</td></tr>
<tr><td rowspan="2">Première</td><td colspan="3">La variable aléatoire X qui compte le nombre de fois ou le contrôleur gagne suit une loi binomiale de paramètres n et p.
On cherche les plus petits entiers a et b tels que $\mathrm{P}(\mathrm{X}\leqslant a)>0{,}025$ et $\mathrm{P}(\mathrm{X}\leqslant b)\geqslant 0{,}975$.
On observe donc si $f\in \mathrm{I}=\left[\frac{a}{n};\frac{b}{n}\right]$</td></tr>
<tr><td>$a=0$, $b=7$
donc $\mathrm{I}=[0;0{,}1]$.
$f\in \mathrm{I}$.
On accepte l'hypothèse $g=0{,}06$.</td><td>$a=3$, $b=13$
donc $\mathrm{I}=[0{,}025;0{,}109]$.
$f\notin \mathrm{I}$.
On rejette l'hypothèse $g=0{,}06$.</td><td>$a=15$, $b=34$
donc $\mathrm{I}=[0{,}0375;0{,}085]$.
$f\in \mathrm{I}$.
On accepte l'hypothèse $g=0{,}06$.</td></tr>
<tr><td rowspan="2">Tale</td><td colspan="3">Si les conditions d'application sont vérifiées ($n\geqslant 30$ et $n\times p\geqslant 5$ et $n\times(1-p)\geqslant 5$).
On observe donc si $f\in \mathrm{I}=\left[p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$</td></tr>
<tr><td>Conditions d'application :
$n=50$ mais $n\times p=3$.
On n'est pas dans les conditions d'application d'un intervalle de fluctuation asymptotique.
Ce contrôle ne peut rien donner de probant en termes de prise de décision.</td><td>Conditions d'application :
$n=120$, $n\,p=7{,}2$ et $n(1-p)=112{,}8$.

$\mathrm{I}=[0{,}0175;0{,}1025]$.
$f\notin \mathrm{I}$.

On rejette l'hypothèse $g=0{,}06$.</td><td>Conditions d'application :
$n=400$, $np=24$ et $n(1-p)=376$.

$\mathrm{I}=[0{,}0367;0{,}0833]$.
$f\in \mathrm{I}$.

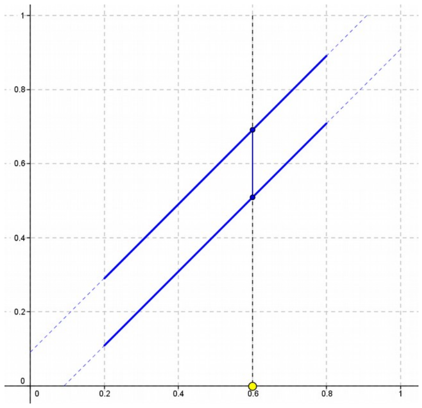
On accepte l'hypothèse $g=0{,}06$.</td></tr>
</table>

Remarque :

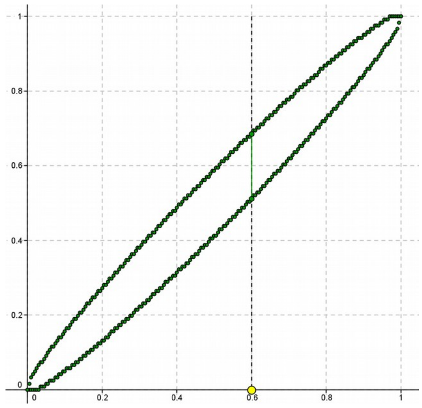
L'intervalle de fluctuation étudié ici est « bilatéral » : on est conduit à rejeter l'hypothèse dans le cas où la fréquence observée est « trop grande » ou « trop petite ».

Illustration :

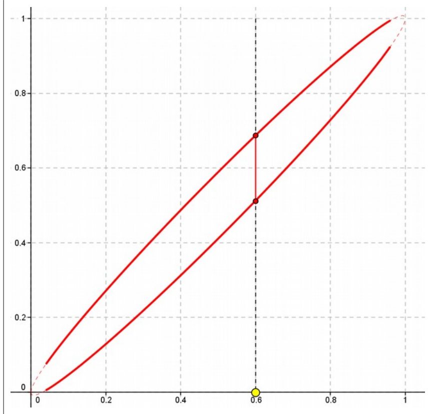
En seconde
Sensibilisation

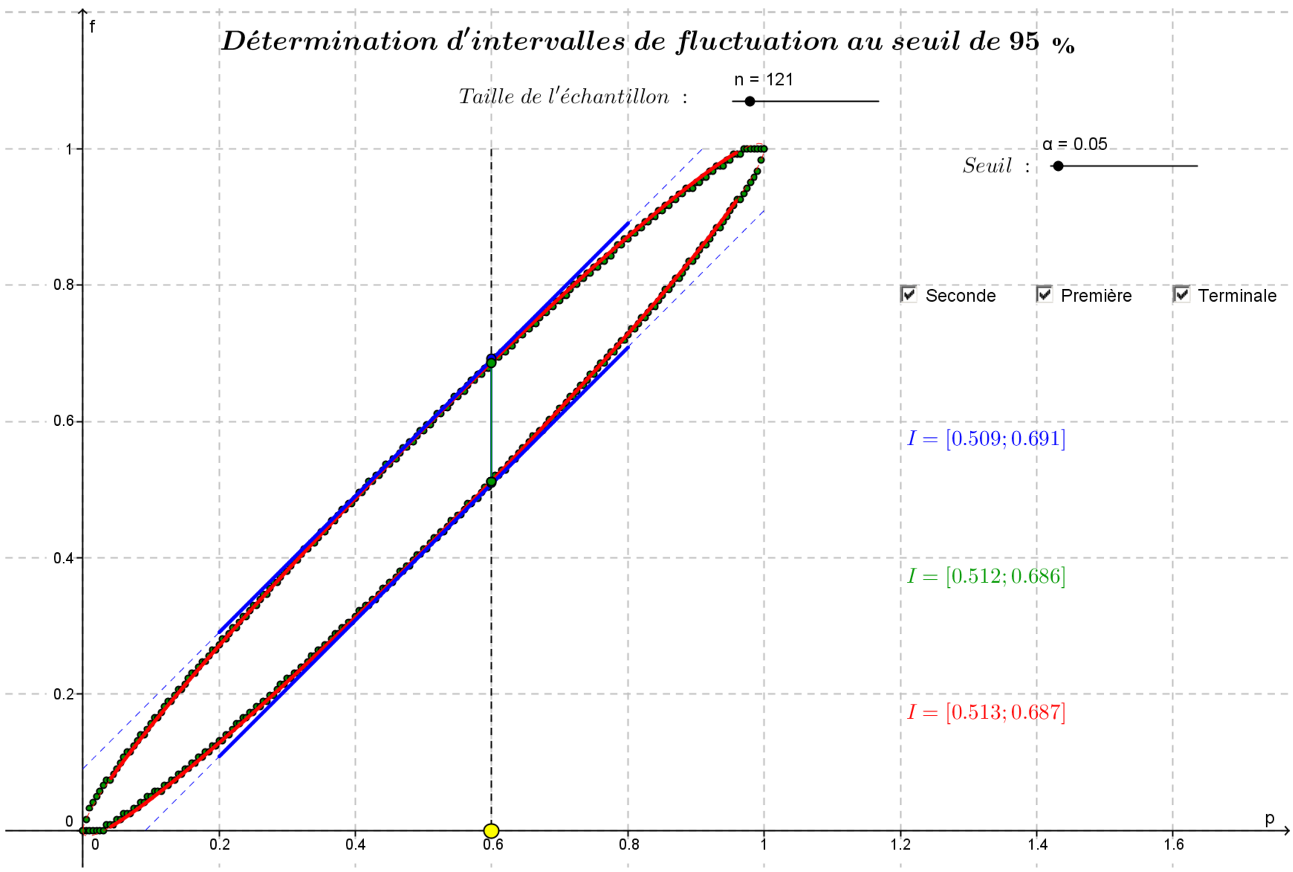


En première
Intervalle de fluctuation exact



En terminale
Intervalle de fluctuation asymptotique

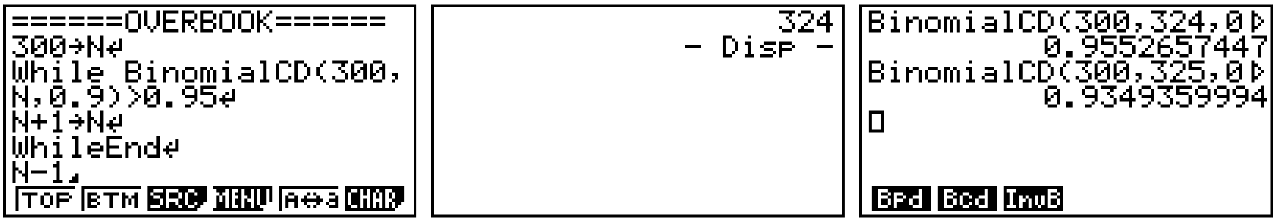




Exemple :

Une compagnie aérienne dispose d'un avion de 300 places et vend *n* réservations $(n>300)$.
La probabilité qu'un acheteur se présente à l'embarquement est $p=0,9$ et l'absence d'une personne à l'embarquement n'influe pas sur celle d'une autre.
L'objectif est d'évaluer la probabilité de surréservation de cette compagnie, autrement dit le risque que plus de 300 passagers se présentent à l'embarquement.
On cherche à maîtriser le risque de telle façon que la probabilité de surbooking ne dépasse pas 5 %.
Quelle est la valeur maximum de *n* ?

- Soit X_n la variable aléatoire dont la valeur est égale au nombre de passagers se présentent à l'embarquement.
  X_n Suit une loi binomiale de paramètres $(n;0,9)$.
  On cherche le plus grand *n* tel que $p(X_n>300)<0,05$ soit $p(X_n\leqslant 300)>0,95$.



  On obtient donc que la valeur maximale de *n* est 324
- On s'intéresse à la fréquence $\dfrac{X_n}{n}$.
  Comme $n>300$ et $p=0,9$ on a $np>5$ et $n(1-p)>5$, on peut utiliser l'intervalle de fluctuation asymptotique au seuil de 0,95 de $\dfrac{X_n}{n}$ et par conséquent $p\left(\dfrac{X_n}{n}<\dfrac{300}{n}\right)$ sera supérieur à 0,95 dès que $\left[0;\dfrac{300}{n}\right]$ contiendra l'intervalle de fluctuation :
  $$I_n=\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$
  Ce qui se traduit par $p+1,96\sqrt{\dfrac{p(1-p)}{n}}\leqslant\dfrac{300}{n}$ soit $0,9n+1,96\sqrt{0,9\times 0,1}\sqrt{n}-300\leqslant 0$
  En posant $x=\sqrt{n}$ on obtient $x<17,934$ soit $n\leqslant 323$.
- Soit X_n la variable aléatoire dont la valeur est égale au nombre de passagers se présentant à l'embarquement.
  X_n Suit une loi binomiale de paramètres $(n;0,9)$.
  On cherche *n* tel que $p(X_n\leqslant 300)>0,95$.
  Or $\lim\limits_{n\to+\infty} P(X_n\leqslant 300)=P\left(\dfrac{X_n-0,9n}{\sqrt{0,09n}}\leqslant\dfrac{300-0,9n}{\sqrt{0,09n}}\right)$.
  Ce qui est très proche de $P\left(Z\leqslant\dfrac{300-0,9n}{\sqrt{0,09n}}\right)$ où Z suit une loi normale centrée réduite.
  On a donc $P(X_n\leqslant 300)\simeq P\left(Z\leqslant\dfrac{300-0,9n}{\sqrt{0,09n}}\right)$ et on s'intéresse alors à $\dfrac{300-0,9n}{\sqrt{0,09n}}>1,64$ ce qui se traduit par $300-0,9n-1,64\sqrt{0,09}\sqrt{n}>0$.
  En posant $x=\sqrt{n}$ on obtient $x<17,987$ soit $n\leqslant 323$.

III. Estimation

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

1) Intervalle de confiance

Définition :
Pour tout α dans $]0;1[$, **un intervalle de confiance** pour une proportion *p* au niveau de confiance $1-\alpha$ est la réalisation, à partir d'un échantillon, d'un **intervalle aléatoire** contenant la proportion *p* avec une probabilité supérieure où égale à $1-\alpha$.

Remarques :

- Il est implicitement supposé que l'intervalle aléatoire considéré est déterminé à partir de la variable aléatoire qui, à tout échantillon de taille *n*, associe la fréquence observée du caractère étudié dans cet échantillon.
- De la même façon que pour les intervalles de fluctuation, l'emploi de « un » dans cette définition, souligne la non-unicité d'un intervalle de confiance à un niveau de confiance donné.

Propriété :
Soit X_n une variable aléatoire suivant une loi binomiale $\mathscr{B}(n\,;\,p)$ où *p* est la proportion inconnue d'apparition d'un caractère, et $F_n=\frac{X_n}{n}$ la fréquence associée à X_n . Alors, pour *n* suffisamment grand, *p* appartient à l'intervalle $\left[F_n-\frac{1}{\sqrt{n}};F_n+\frac{1}{\sqrt{n}}\right]$ **avec une probabilité supérieur ou égale à 0,95**.

Démonstration :
Nous avons vu que l'intervalle $\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$ est, pour *n* assez grand, un intervalle de fluctuation au seuil de 95 % pour $F_n=\frac{X_n}{n}$.

Donc, $\exists\, n_0 \in \mathbb{N}^*$, $\forall\, n \geqslant n_0$, $P\left(p-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant p+\frac{1}{\sqrt{n}}\right)\geqslant 0{,}95$.

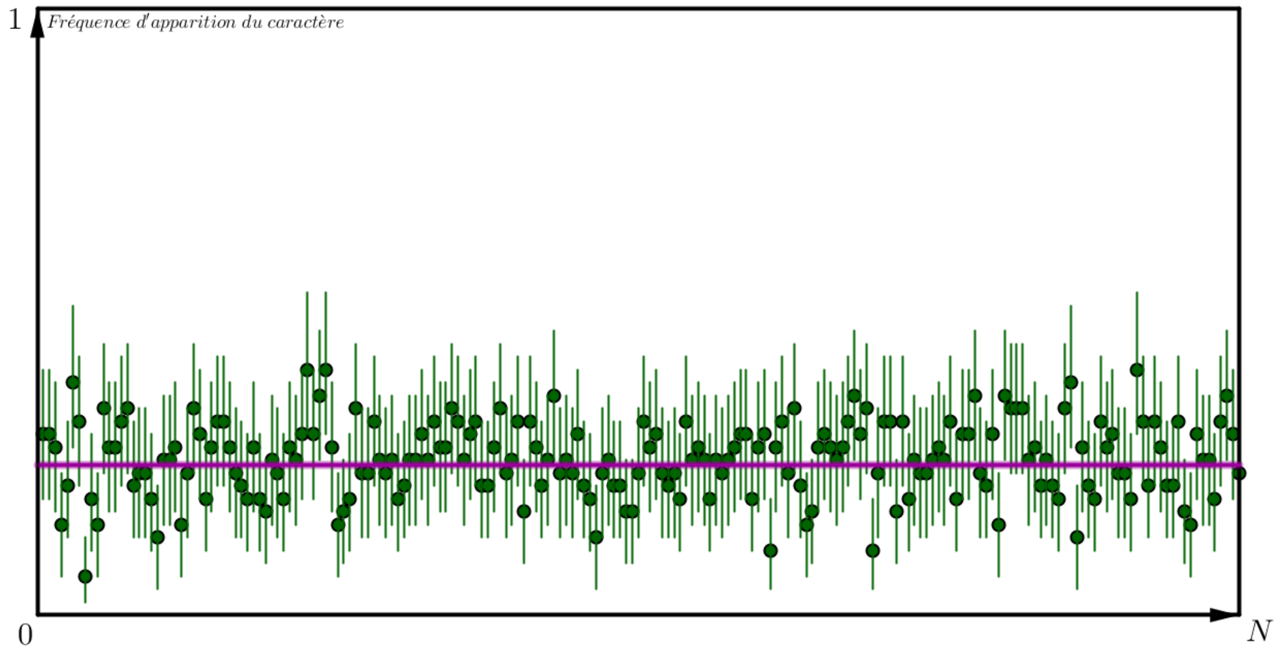
Or $\left(p-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant p+\frac{1}{\sqrt{n}}\right) \Leftrightarrow \left(F_n-\frac{1}{\sqrt{n}}\leqslant p\leqslant F_n+\frac{1}{\sqrt{n}}\right)$

Donc, $\exists\, n_0 \in \mathbb{N}^*$, $\forall\, n \geqslant n_0$, $P\left(F_n-\frac{1}{\sqrt{n}}\leqslant p\leqslant F_n+\frac{1}{\sqrt{n}}\right)\geqslant 0{,}95$.

Définition :
On **réalise** l'expérience aléatoire de *n* tirages au hasard, et on appelle f la fréquence observée d'apparition du caractère.
L'intervalle $\left[f-\frac{1}{\sqrt{n}};f+\frac{1}{\sqrt{n}}\right]$ est appelé **intervalle de confiance de *p* au niveau de confiance 0,95**, où *p* est la proportion (inconnue) d'apparition du caractère dans la population.

Remarques :

- Le « niveau de confiance 95 % » signifie que si l'on effectuait un très grand nombre de tirages de 100 boules, on devrait obtenir moins de 5 % d'intervalles de confiance ne contenant pas la proportion p de boules rouges.
- On ne peut faire aucun pronostic sur une localisation possible de cette proportion dans l'intervalle de confiance. En particulier, la proportion inconnue p n'est pas nécessairement le centre de l'intervalle de confiance. L'intervalle de confiance au niveau de confiance 0,95 est centré en la fréquence observée f . Cette condition n'est pas imposée par la définition générale, ce n'est donc pas le cas de tous les intervalles de confiance.
- Dans d'autres disciplines, on utilise l'intervalle de confiance (au niveau de confiance 0,95) suivant :
$$\mathrm{I}=\left[f-1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}}\,;\,f+1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}}\right]$$
On remarque que cet intervalle de confiance est également centré sur la fréquence observée f .
- La proportion p étant inconnue, on ne peut pas vérifier si les paramètres n et p satisfont les conditions exigées en ce début de chapitre afin d'utiliser l'intervalle de confiance au niveau de confiance 0,95. Pour remédier à ce problème, **on approche la proportion inconnue p par la fréquence observée f** sur l'échantillon considéré, puis on vérifie si les conditions suivantes sont satisfaites :
$$n\geqslant 30\ ,\ n\times f\geqslant 5\ \text{ et }\ n\times(1-f)\geqslant 5$$



Exemple :
Dans une urne contenant des boules rouges et des boules bleues, on obtient 59 rouges et 41 bleues.
La fréquence observée de sortie du rouge est donc 0,59.
L'intervalle $\left[0{,}59-\frac{1}{\sqrt{100}}\,;0{,}59+\frac{1}{\sqrt{100}}\right]=[0{,}49\,;0{,}69]$ est un intervalle de confiance de la proportion de boules rouges dans l'urne au niveau de confiance 95 %.

2) Précision d'une estimation et taille de l'échantillon

Un intervalle de confiance au niveau 95 % est d'amplitude $\frac{2}{\sqrt{n}}$.
Plus la taille de l'échantillon est grande, plus les intervalles de confiance obtenus sont précis.

Exemple :
Pour obtenir un intervalle de confiance d'amplitude inférieure à 0,01 de la proportion de boules rouges dans l'urne, il faut procéder à des tirages de n boules, avec $\frac{2}{\sqrt{n}} \leqslant 0{,}01$, soit $\frac{4}{n} \leqslant 10^{-4}$, ou encore $n \geqslant 4 \times 10^{4}$. Il faut procéder à au moins 40000 tirages.

Exemple :
Voici les résultats d'un sondage IPSOS réalisé avant l'élection présidentielle de 2002 pour le Figaro et Europe 1, les 17 et 18 avril 2002 auprès de 989 personnes, constituant un échantillon national représentatif de la population française âgée de 18 ans et plus inscrite sur les listes électorales.
On suppose que cet échantillon est constitué de manière aléatoire (même si, en pratique, ce n'est pas le cas).
Les intentions de vote au premier tour pour les principaux candidats sont les suivants :
- 20 % pour Jacques Chirac (198 personnes)
- 18 % pour Lionel Jospin (178 personnes)
- 14 % pour Jean Marie Le Pen (138 personnes)

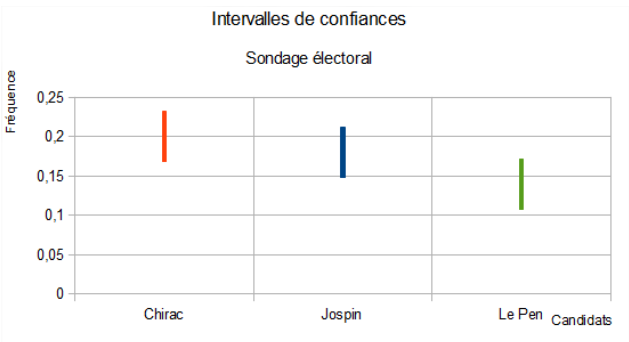
Les médias se préparent pour un second tour entre J. Chirac et L. Jospin.

- L'échantillon étudié comprend 989 personnes. Pour chaque candidat, l'intervalle de confiance au niveau de confiance de 95 % est de la forme $\left[f - \frac{1}{\sqrt{989}} ; f + \frac{1}{\sqrt{989}}\right]$.
  Cela donne :
  - $[0{,}168 ; 0{,}232]$ pour Jacques Chirac
  - $[0{,}148 ; 0{,}212]$ pour Lionel Jospin
  - $[0{,}108 ; 0{,}172]$ pour Jean Marie Le Pen
- Le 21 avril, les résultats du premier tour des élections sont les suivantes :
  - 19,88 % pour Jacques Chirac
  - 16,18 % pour Lionel Jospin
  - 16,86 % pour Jean Marie Le Pen

  Les résultats constatés sont bien dans les intervalles de confiance.
- Ces trois intervalles de confiance ont une intersection non vide $[0{,}168 ; 0{,}172]$. Il n'est donc pas possible, avec un niveau de confiance de 0,95, de désigner le classement final des trois candidats.

C1 =B1-1/RACINE(989)

	A	B	C	D
1	Chirac	0,2	0,16820185	0,23179815
2	Jospin	0,18	0,14820185	0,21179815
3	Le Pen	0,14	0,10820185	0,17179815



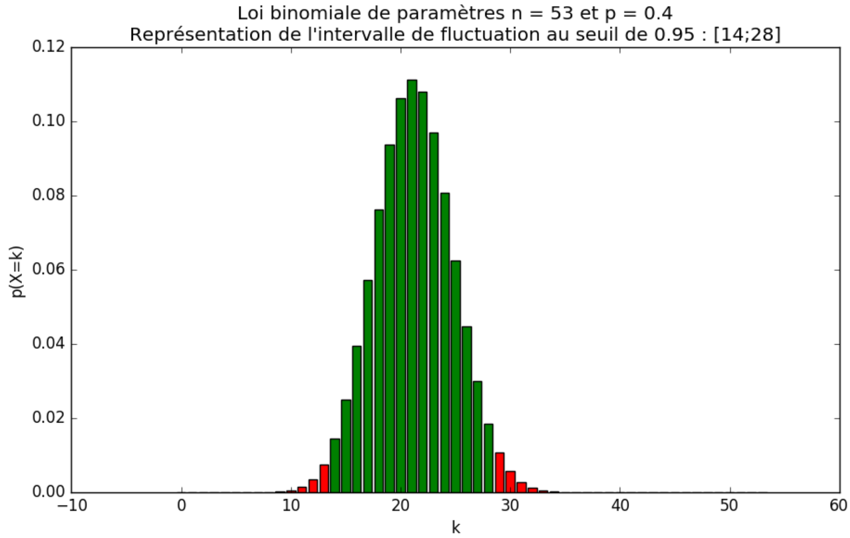
Annexe 1 : Fluctuation et programmation

- **Intervalle de fluctuation**

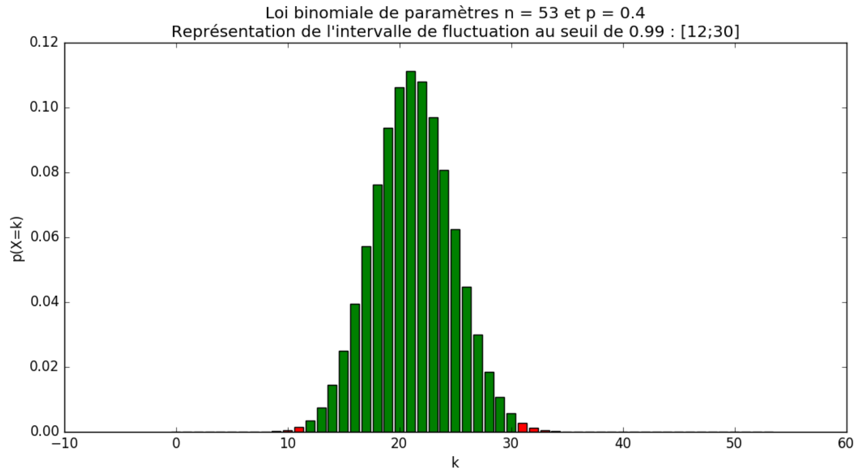
A partir de la loi binomiale (valable pour des valeurs de n et p limitées) : `binomiale.py`

```
#On importe la fonction factorielle et la librairie pyplot pour les tracés
from math import factorial
import matplotlib.pyplot as plt

# Combinaison
def combin(n, k):
    """Nombre de combinaisons de k objets parmi n
    nCk = n!/(k!*(n-k)!)
    Attention! Ne fonctionne pas pour n trop grand"""
    return factorial(n) / (factorial(k) * factorial(n - k))

# Loi binomiale
def binom(k, n, p):
    """p(X=k) ou X est une v.a. qui suit la loi binomiale B(n,p)"""
    return combin(n,k) * pow(p,k) * pow(1 - p,n - k)

# Inverse binomiale
def invBin(n, p, s):
    """Intervalle centré [a;b] tel que p(a<=X<=b) >= s
        ou X est une v.a. qui suit la loi binomiale B(n,p)"""
    # Calcul de a
    somme = 0
    k = 0
    while somme <= (1 - s) / 2:
        somme += binom(k,n,p)
        k += 1
    a = k - 1
    # Calcul de b
    somme = 0
    k = 0
    while somme < (1 + s) / 2:
        somme += binom(k,n,p)
        k += 1
    b = k - 1
    return a, b

# Représentation graphique
def graphBinomiale(n, p, s=0.95):
    """Représentation graphique de la loi binomiale B(n,p)
    s représente le seuil définissant un intervalle de fluctuation centré"""
    a, b = invBin(n,p,s)
    x_in = []
    datas_in = []
    x_out = []
    datas_out = []
    for i in range(n+1):
        valeur = binom(i,n,p)
        if i>=a and i<=b:
            x_in.append(i)
            datas_in.append(valeur)
        else:
            x_out.append(i)
            datas_out.append(valeur)
    # diagramme en barres des données
    plt.bar(x_in,datas_in, align='center', color='green')
    plt.bar(x_out,datas_out, align='center', color='red')
    plt.title("Loi binomiale de paramètres n = " + str(n) + " et p = " + str(p) +
              "\nReprésentation de l'intervalle de fluctuation au seuil de "+
              str(s) + " : [" + str(a) + ";" + str(b) + "]")
    plt.xlabel('k')
    plt.ylabel('p(X=k)')
    plt.show()
```

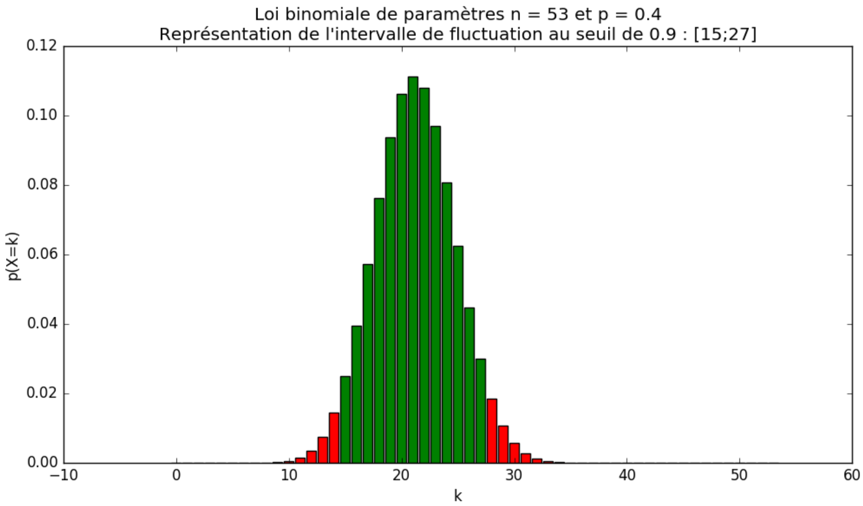
```
>>> graphBinomiale(53,0.4)
```




```
>>> graphBinomiale(53,0.4,0.99)
```



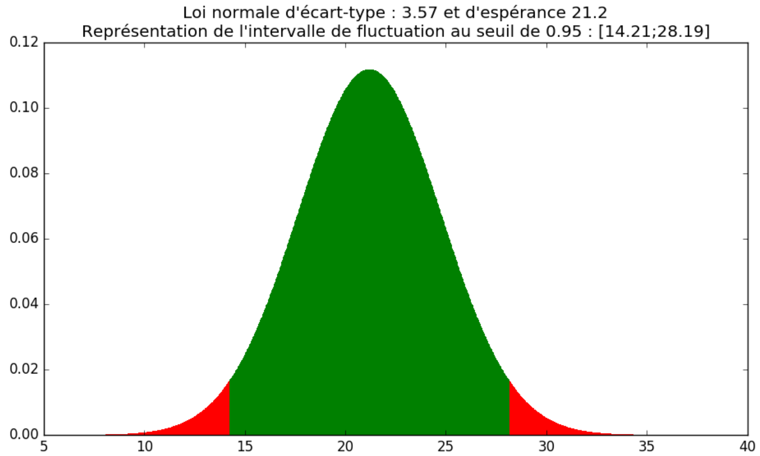

```
>>> graphBinomiale(53,0.4,0.9)
```



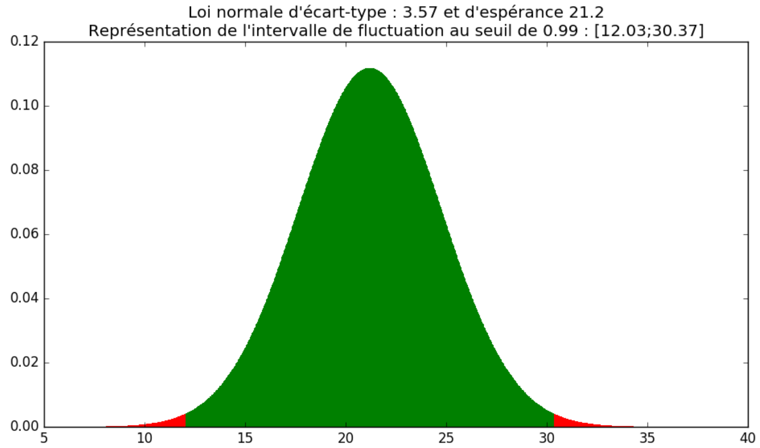
À partir de la loi normale (intervalle de fluctuation asymptotique) : `normale.py`

```
# On importe les items nécessaires pour définir la loi normale
# et la librairie pyplot pour les tracés
from math import sqrt, pi, exp
import matplotlib.pyplot as plt

# On définit la fonction de densité de la loi normale
def normale(x, σ=1, μ=0):
    """ Loi Normale par défaut espérance (μ) = 0 et écart-type (σ) = 1"""
    return 1/(σ*sqrt(2*pi))*exp(-0.5*((x-μ)/σ)**2)

# On implémente la méthode des rectangles pour l'intégration numérique
def integRect(f, a, b, n=1000):
    """ Approximation de l'intégrale de f(x) sur [a;b]
    n détermine le nombre de subdivisions sur l'intervalle (par défaut 1000)"""
    somme = 0
    dx = (b-a)/n
    for k in range (n):
        somme += f(a+(k*dx))*dx
    return somme

# Fonction de répartition de la loi normale
def normCum(a, b, σ=1, μ=0, n=1000):
    """p(a<=X<=k) ou X est une v.a. qui suit la loi normale N(σ,μ)
    n détermine le nombre de subdivisions sur l'intervalle (par défaut 1000)"""
    densite = lambda x: normale(x,σ,μ)
    return integRect(densite,a,b,n)

# Inverse normale
def invNorm(s, σ=1, μ=0, n=1000):
    """Intervalle centré [a;b] tel que p(a<=X<=b) >= s
    ou X est une v.a. qui suit la loi normale N(σ,μ)
    n détermine le nombre de subdivisions de l'unité"""
    densite = lambda x: normale(x,σ,μ)
    # Calcul de b
    x = μ
    dx = 1/n
    somme = 0
    k = 0
    while somme < s / 2:
        somme += densite(x+(k*dx))*dx
        k += 1
    b = μ + (k - 1)*dx
    # Calcul de a
    a = μ - (k - 1)*dx
    return a, b

# Représentation graphique
def graphNormale(σ=1, μ=0, s=0.95, n=1000):
    """Représentation graphique de la loi normale N(σ,μ)
    s représente le seuil définissant un intervalle de fluctuation centré"""
    a, b = invNorm(s,σ,μ)
    x_moins, x_in, x_plus = [], [], []
    datas_moins, datas_in, datas_plus = [], [], []
    start = μ - 4 * σ
    end = μ + 4 * σ
    dx = (end - start) / n
    for i in range(n):
        x = start + i * dx
        valeur = normale(x, σ, μ)
        if x<=a:
            x_moins.append(x)
            datas_moins.append(valeur)
        elif x <= b:
            x_in.append(x)
            datas_in.append(valeur)
        else:
            x_plus.append(x)
            datas_plus.append(valeur)
    # diagramme en barre des données
    largeur = 1 / n**0.5
    plt.bar(x_in, datas_in, align='center', color='green', width = largeur, linewidth=0)
    plt.bar(x_moins, datas_moins, align='center', color='red', width = largeur, linewidth=0)
    plt.bar(x_plus, datas_plus, align='center', color='red', width = largeur, linewidth=0)
    plt.title("Loi normale d'écart-type : " + str(round(σ,2)) + " et d'espérance " +
              str(round(μ,2)) + "\nReprésentation de l'intervalle de fluctuation au seuil de "+
              str(s) + " : [" + str(round(a,2)) + ";" + str(round(b,2)) + "]")
    plt.show()
```

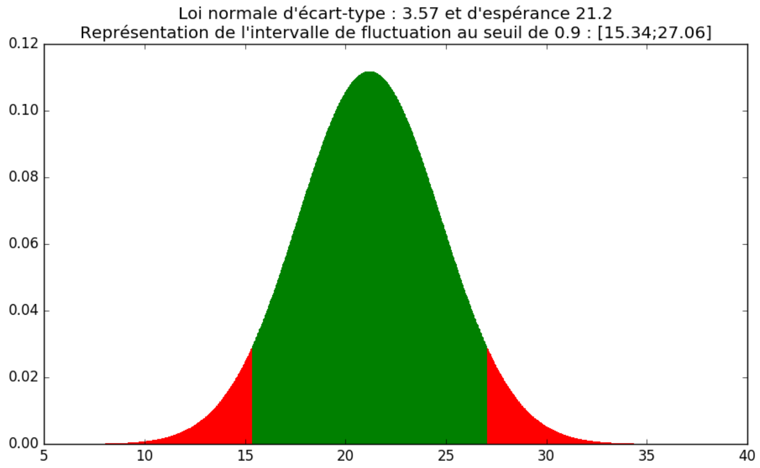
```
>>> mu=53*0.4
>>> sig=(53*0.4*0.6)**0.5
>>> graphNormale(sig,mu)
```




```
>>> graphNormale(sig,mu,0.99)
```



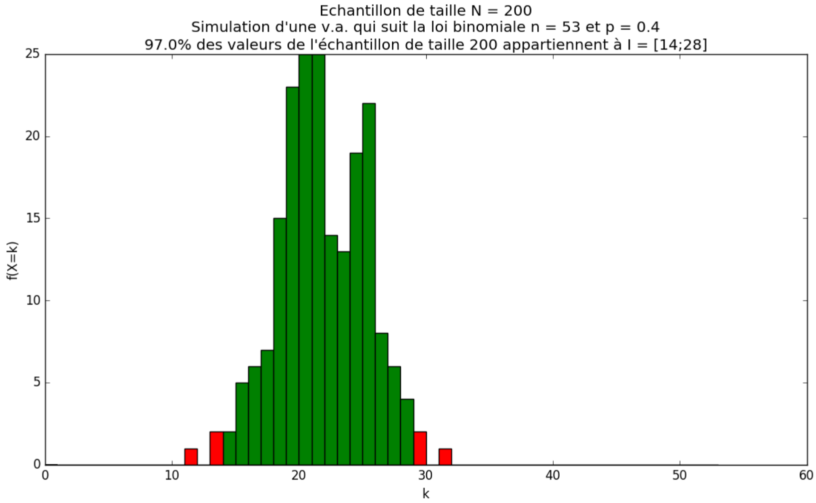

```
>>> graphNormale(sig,mu,0.9)
```



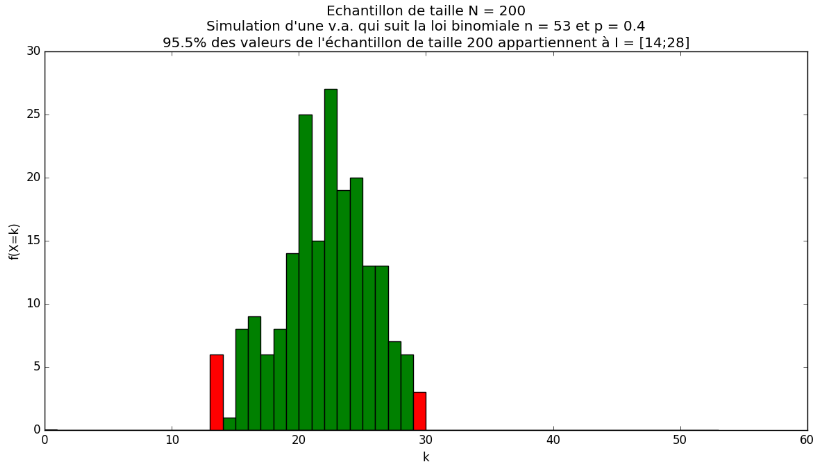
- **Simulation**

```
from random import random           # pour la simulation
from normale import invNorm         # pour l'intervalle de fluctuation
import matplotlib.pyplot as plt     # pour la représentation

# Loi de Bernoulli
def randBernoulli(p):
    """génère un nombre aléatoire compris entre 0 et 1
    en conformité avec la loi de Bernoulli de paramètres p"""
    if random() < p:
        return 1
    else:
        return 0

# Loi binomiale
def randBinomiale(n, p):
    """génère un nombre aléatoire compris entre 0 et n
    en conformité avec la loi binomiale de paramètres n et p"""
    somme = 0
    for k in range(n):
        somme += randBernoulli(p)
    return somme

# Simulation d'un échantillon de taille N
def echantillon(n, p, N, s=0.95):
    """Observation empirique de la fluctuation d'échantillonnage
    X est une v.a. qui suit B(n,p)
    s est le seuil définissant l'intervalle centré [a;b] tel que p(a<=B<=b)>=s
    N est la taille de l'échantillon"""
    # On récupère l'intervalle de fluctuation asymptotique au seuil s
    sigma = (n*p*(1-p))**0.5
    mu = n*p
    a,b = invNorm(s, sigma, mu)
    a,b = round(a), round(b)
    # On effectue la simulation
    datas_in = []
    datas_out = []
    compteur = 0
    for i in range(N):
        valeur = randBinomiale(n,p)
        if valeur>=a and valeur<=b:
            datas_in.append(valeur)
            compteur += 1
        else:
            datas_out.append(valeur)
    # Histogramme des données
    plt.hist(datas_in, range = (0, n), bins = n, color='green')
    plt.hist(datas_out, range = (0, n), bins = n, color='red')
    plt.title("Echantillon de taille N = " + str(N) +
              "\nSimulation d'une v.a. qui suit la loi binomiale n = " + str(n) +
              " et p = " + str(p) + "\n" + str(round(compteur*100/N,2)) +
              "% des valeurs de l'échantillon de taille "+str(N)+
              " appartiennent à I = [" + str(a) + ";" + str(b) + "]")
    plt.xlabel('k')
    plt.ylabel('X=k')
    plt.show()
```

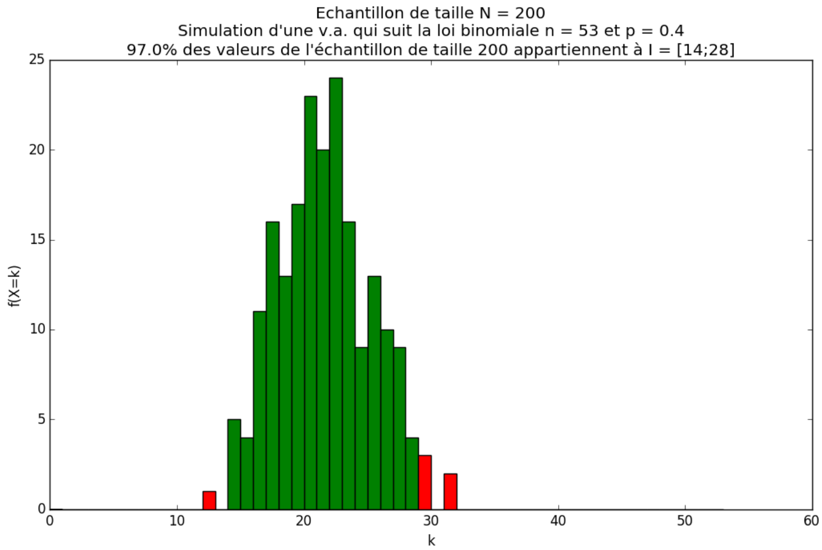
```
>>> echantillon(53,0.4,200)
```




```
>>> echantillon(53,0.4,200)
```

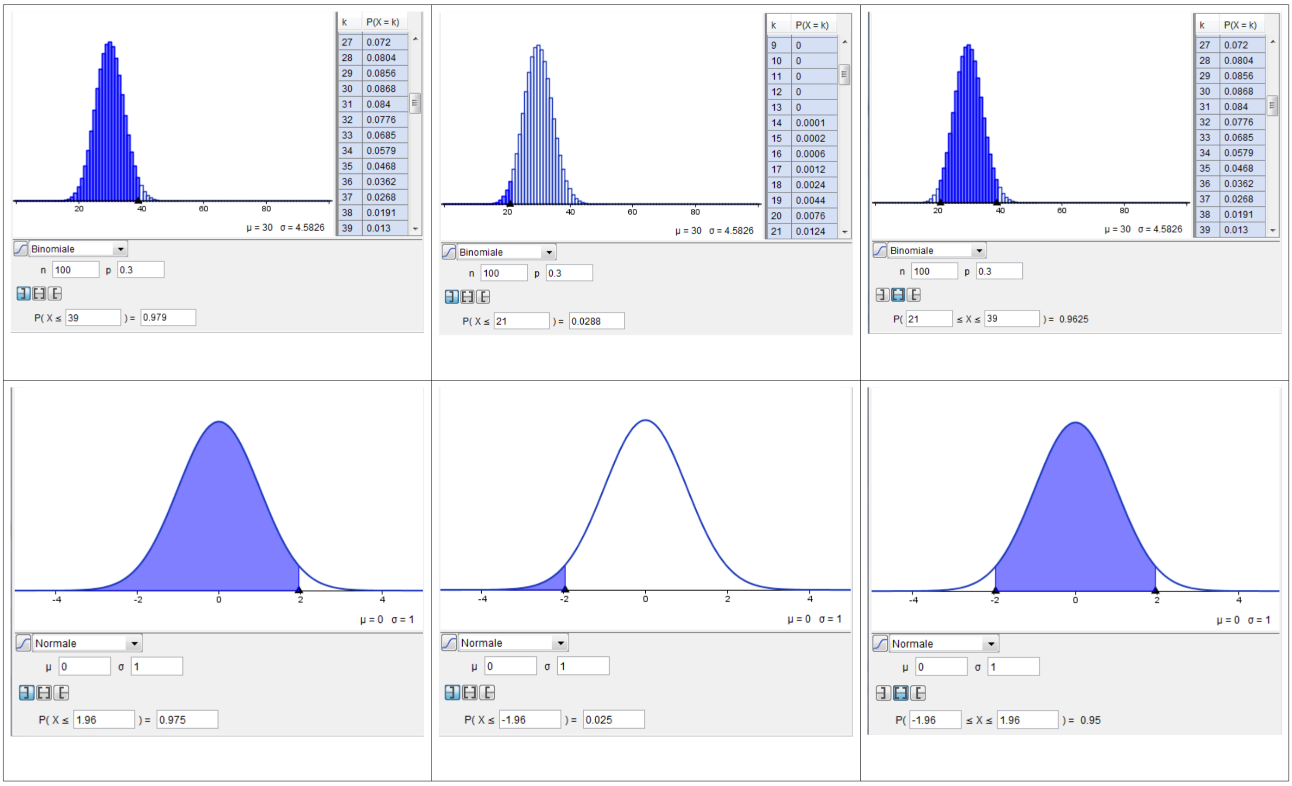



```
>>> echantillon(53,0.4,200)
```



Annexe 2 : Tests unilatéraux et bilatéraux

Dans un problème de prise de décisions, avant d'appliquer tout test statistique, il s'agit de bien définir le problème posé.
En effet, selon les hypothèses formulées, on applique soit un test bilatéral, soit un test unilatéral.



Test bilatéral

- **Intervalle de fluctuation au seuil de 95 %**

Pour les tests bilatéraux, le risque de 5% (1-0,95=0,05) est également réparti entre les valeurs "trop petites" (Z<-1,96) et les valeurs "trop grandes" (Z>1,96).
Si les conditions d'application sont vérifiées ($n \geqslant 30$ et $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$).
On observe donc si $f \in \mathrm{I} = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ avec $u_\alpha \simeq 1{,}96$.

<u>Exemple :</u>
Dans un casino, il a été décidé que les « machines à sous » doivent être réglées sur une fréquence de gain du joueur de $g = 0{,}06$.
Une fréquence inférieure est supposée faire « fuir le client », et une fréquence supérieure est susceptible de ruiner le casino.
Trois contrôleurs différents vérifient une même machine.
Le premier a joué 50 fois et gagné 2 fois, le second a joué 120 fois et gagné 14 fois, le troisième a joué 400 fois et gagné 30 fois.
En utilisant des intervalles de fluctuation au seuil 95 %, examiner dans chaque cas la décision à prendre par le contrôleur, à savoir accepter ou rejeter l'hypothèse $g = 0{,}06$.

<table>
<tr><td>1er contrôleur
$n=50$, $p=0{,}06$, $f=\dfrac{2}{50}=0{,}04$</td><td>2e contrôleur
$n=120$, $p=0{,}06$,
$f=\dfrac{14}{120}\simeq 0{,}1167$</td><td>3e contrôleur
$n=400$, $p=0{,}06$, $f=\dfrac{30}{400}=0{,}075$</td></tr>
<tr><td>Conditions d'application :
$n=50$ mais $n\times p=3$.
On n'est pas dans les conditions d'application d'un intervalle de fluctuation asymptotique.
Ce contrôle ne peut rien donner de probant en termes de prise de décision.</td><td>Conditions d'application :
$n=120$, $n\,p=7{,}2$ et $n(1-p)=112{,}8$.

$\mathrm{I}=[0{,}0175;0{,}1025]$.
$f\notin \mathrm{I}$.

On rejette l'hypothèse $g=0{,}06$.</td><td>Conditions d'application :
$n=400$, $np=24$ et $n(1-p)=376$.

$\mathrm{I}=[0{,}0367;0{,}0833]$.
$f\in \mathrm{I}$.

On accepte l'hypothèse $g=0{,}06$.</td></tr>
</table>

Test unilatéral

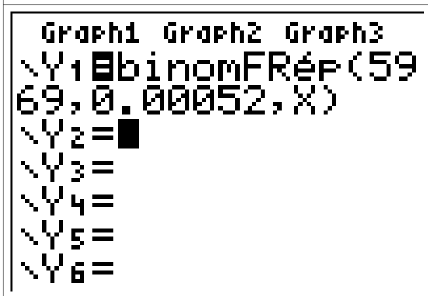
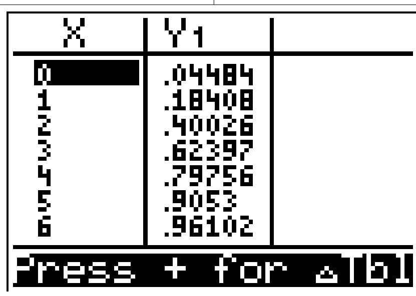
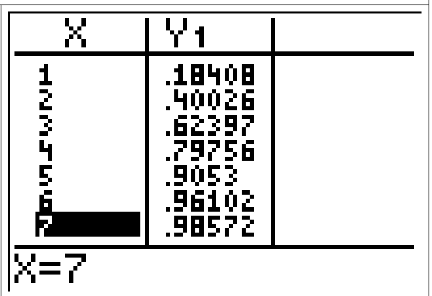
Pour les tests unilatéraux, on ne s'intéresse par exemple qu'aux valeurs « trop grandes » ou « trop petites ».

<u>Remarque :</u>
Ces tests concernent généralement des situations où les conditions d'application ne permettent pas d'utiliser l'intervalle de fluctuation asymptotique.

<u>Exemples :</u>

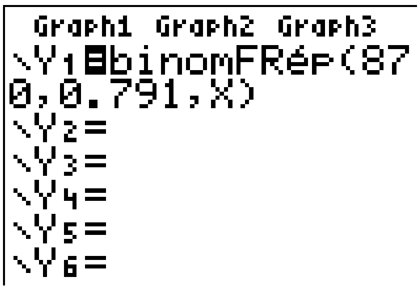
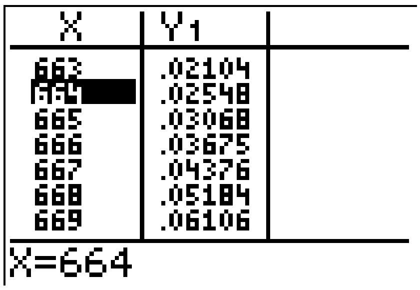
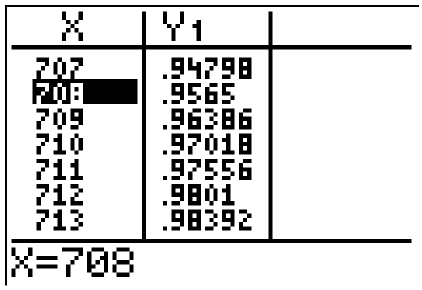
- Une petite ville des États-Unis, Woburn, a connu 9 cas de leucémie parmi les 5969 garçons de moins de 15 ans sur la période 1969-1979.
  La fréquence des leucémies pour cette tranche d'âge aux États-Unis est égale à 0,00052.

$$f=\frac{9}{5969}\simeq 0{,}0015$$

<table>
<tr><td colspan="2">Test bilatéral</td><td>Test unilatéral
S'agissant d'un problème de santé publique, un taux anormalement bas de leucémies ne permet pas de considérer que la situation n'est pas dangereuse alors qu'un taux anormalement élevé permet de considérer la situation comme étant dangereuse.</td></tr>
<tr><td colspan="2">a et b tels que $\mathrm{P}(\mathrm{X}\leqslant a)>0{,}025$ et $\mathrm{P}(\mathrm{X}\leqslant b)\geqslant 0{,}975$</td><td>$c$ tel que $\mathrm{P}(\mathrm{X}\leqslant c)\geqslant 0{,}95$</td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td>$a=0$</td><td>$b=7$</td><td>$c=6$</td></tr>
<tr><td colspan="2">Donc $\mathrm{I}=\left[\dfrac{a}{n};\dfrac{b}{n}\right]=\left[0;\dfrac{7}{5969}\right]=[0;0{,}0012]$
et $f\notin\mathrm{I}$ donc on considère que la situation est dangereuse.</td><td>Donc $\mathrm{I}=\left[0;\dfrac{c}{n}\right]=\left[0;\dfrac{6}{5969}\right]=[0;0{,}0010]$
et $f\notin\mathrm{I}$ donc on considère (de façon plus évidente encore) que la situation est dangereuse.</td></tr>
</table>

- En novembre 1976 dans un comté du sud du Texas, Rodrigo Partida était condamné à huit ans de prison.
  Il attaqua ce jugement au motif que la désignation des jurés de ce comté était discriminante à l'égard des Américains d'origine mexicaine.
  Alors que 79,1% de la population de ce comté était d'origine mexicaine, sur les 870 personnes convoqués pour être jurés lors d'une certaine période de référence, il n'y eut que 339 personnes d'origine mexicaine.

$$f=\frac{339}{870}\simeq 0{,}390$$

<table>
<tr><td colspan="2">Test bilatéral</td><td>Test unilatéral
S'agissant d'un problème d'interprétation, une proportion anormalement élevée de jurés d'origine mexicaine ne permet pas d'argumenter en faveur de M Partida alors qu'une proportion anormalement basse est à l'origine de l'argumentation.</td></tr>
<tr><td colspan="2">a et b tels que $\mathrm{P}(\mathrm{X}\leqslant a)>0{,}025$ et $\mathrm{P}(\mathrm{X}\leqslant b)\geqslant 0{,}975$</td><td>$c$ tel que $\mathrm{P}(\mathrm{X}\leqslant c)>0{,}05$</td></tr>
<tr><td>
</td><td>
</td><td>
</td></tr>
<tr><td>$a=664$</td><td>$b=711$</td><td>$c=668$</td></tr>
<tr><td colspan="2">Donc $\mathrm{I}=\left[\frac{a}{n};\frac{b}{n}\right]=\left[\frac{664}{870};\frac{711}{870}\right]=[0{,}763;0{,}817]$
et $f\notin\mathrm{I}$ donc on considère que la situation est « anormale ».</td><td>Donc $\mathrm{I}=\left[\frac{c}{n};1\right]=\left[\frac{668}{870};1\right]=[0{,}768;1]$
et $f\notin\mathrm{I}$ donc on considère (de façon plus flagrante encore) que la situation est « anormale ».</td></tr>
</table>

Annexe 3 : Risque d'erreurs

Test statistique et prise de décision

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Décision
(à partir d'un échantillon représentatif)

Hypothèse : la proportion du caractère étudié est p</td></tr>
<tr><td>On rejette l'hypothèse</td><td>On accepte l'hypothèse</td></tr>
<tr><td rowspan="2">Réalité</td><td>p est la proportion du caractère étudiée</td><td>Erreur ·
(avec un risque de 5%)</td><td>Bonne décision</td></tr>
<tr><td>p n'est pas la proportion du caractère étudiée</td><td>Bonne décision</td><td>Erreur √2
(non quantifiée)</td></tr>
</table>

<u>Exemple :</u>
Fabrication industrielle (d'ampoules par exemple).
Comment décider qu'un lot sur lequel on effectue des tests (sur des échantillons) est conforme ?
Sur un échantillon représentatif on obtient une proportion p de pièces défectueuses.
Hypothèse : la proportion de pièces défectueuses dans le lot est p.

Erreur de première espèce α

On rejette l'hypothèse et on se trompe (ceci avec une probabilité de ~ 5% dans le cours).
On obtient une proportion p « trop élevée » dans l'échantillon et on rejette donc le lot (à tort).

L'entreprise est perdante.

Le test est **significatif**.

Erreur de deuxième espèce β

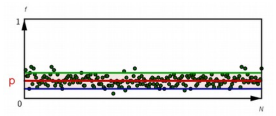
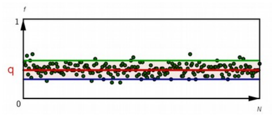
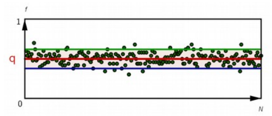
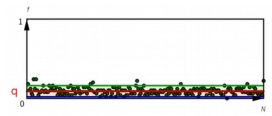
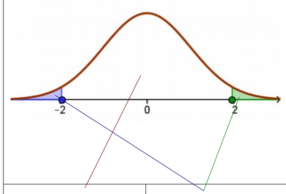
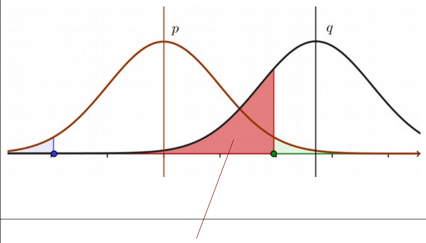
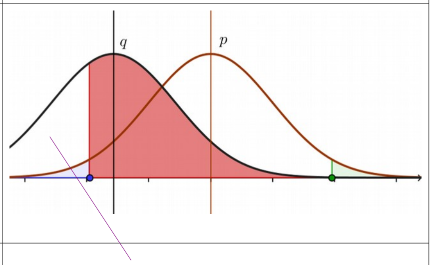
On accepte l'hypothèse et on se trompe.
On obtient une proportion p « correcte » dans l'échantillon et on accepte donc le lot (à tort).

Le client est perdant.

Le test est **non significatif**.
D'où l'habitude de conclure : « on ne rejette pas l'hypothèse » plutôt qu'« on accepte l'hypothèse ».

Il est donc nécessaire de réfléchir au choix des hypothèses.

Prise de décision

<table>
<tr><td></td><td colspan="2">Erreur α</td><td colspan="6">Erreur β</td></tr>
<tr><td>Réalité</td><td colspan="2">La proportion du caractère étudié dans la population est p</td><td colspan="6">La proportion du caractère étudié dans la population n'est pas p</td></tr>
<tr><td rowspan="2">Hypothèse</td><td colspan="8">La proportion du caractère étudié est p
Remarque : l'hypothèse se fait en général à partir de données statistiques</td></tr>
<tr><td colspan="2">L'hypothèse est vraie</td><td colspan="6">L'hypothèse est fausse</td></tr>
<tr><td>Échantillonnage</td><td colspan="8">On prélève un échantillon représentatif au sein de la population</td></tr>
<tr><td>Fluctuation</td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Étude</td><td>$f \in \mathrm{I}$
I centré sur p</td><td>$f \notin \mathrm{I}$
I centré sur p</td><td colspan="3">$f \in \mathrm{I}$
I centré sur p</td><td colspan="3">$f \notin \mathrm{I}$
I centré sur p</td></tr>
<tr><td rowspan="4">Décision</td><td>On accepte l'hypothèse.
La proportion du caractère étudiée dans la population est p</td><td>On rejette l'hypothèse.
La proportion du caractère étudiée dans la population n'est pas p</td><td colspan="3">On accepte l'hypothèse
La proportion du caractère étudiée dans la population est p</td><td colspan="3">On rejette l'hypothèse
La proportion du caractère étudiée dans la population n'est pas p</td></tr>
<tr><td colspan="2"></td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td>La probabilité d'accepter l'hypothèse alors qu'elle est vraie est de 0,95</td><td>La probabilité de rejeter l'hypothèse alors qu'elle est vraie est de 0,05</td><td colspan="3">Probabilité d'accepter l'hypothèse alors qu'elle est fausse : ?</td><td colspan="3">Probabilité de rejeter l'hypothèse alors qu'elle est fausse : ?</td></tr>
<tr><td>On ne conclut pas à une « anormalité » au sein de la population étudiée</td><td>On conclut à une « anormalité » au sein de la population étudiée (avec une marge d'erreur de 5%)</td><td colspan="6">Le test n'est pas significatif.</td></tr>
<tr><td>Conclusion</td><td colspan="8">On ne se « trompe » pas lorsque l'on rejette l'hypothèse :
• si l'hypothèse est vraie alors on rejette l'hypothèse « à tort » dans 5 % des cas.
• si l'hypothèse est fausse alors on rejette l'hypothèse qui est ... fausse.
Par contre, accepter l'hypothèse ne permet pas de conclure que la proportion est p (même avec une marge d'erreur de 5%).</td></tr>
</table>

Annexe 4 : Intervalle de confiance

Un **intervalle de confiance** permet de définir une marge d'erreur entre les résultats d'un sondage et un relevé exhaustif de la population totale. Plus généralement, l'intervalle de confiance permet d'évaluer la précision de l'estimation d'un paramètre statistique sur un échantillon.
Contrairement à l'intervalle de fluctuation, qui est déterminé par le paramètre et vise à encadrer l'estimateur (la proportion), l'intervalle de confiance est aléatoire car dépend de l'échantillon et vise à encadrer le paramètre réel.

Formulation de l'**intervalle de confiance** centré autour d'une moyenne observée $\bar{x}$ avec un écart type observé s sur un échantillon de taille n.

$$\mathrm{I_c} = \left[\bar{x} - t_\alpha \frac{s}{\sqrt{n}} ; \bar{x} + t_\alpha \frac{s}{\sqrt{n}} \right]$$

t_α est le quantile d'ordre $\frac{\alpha}{2}$ de la loi normale.

La notion d'intervalle de confiance apparaît lorsqu'on tente d'obtenir des informations synthétiques sur une population que l'on ne connaît pas entièrement. Il faut donc associer à la population une loi de probabilité dont la pertinence doit être justifiée. Ceci conduit à interpréter un élément de la population comme une variable aléatoire et un échantillon comme un ensemble de telles variables.

En particulier, la moyenne et la variance, dites empiriques, calculées à partir de l'échantillon selon les règles algébriques applicables en statistique descriptive, sont elles-mêmes des variables aléatoires dont il est possible de calculer la moyenne et la variance, sous réserve d'indépendance des éléments de l'échantillon.

Estimation d'une moyenne

L'usage le plus simple des intervalles de confiance concerne les populations à distribution normale dont on cherche à estimer la moyenne μ.
Si on mesure la moyenne $\bar{x}$ sur un échantillon de taille n pris au hasard alors on détermine l'intervalle de confiance de μ à un seuil donné en utilisant les valeurs de t_α obtenus à partir de la loi normale.
Par exemple, $\mathrm{I_c} = \left[\bar{x} - 2 \frac{\sigma}{\sqrt{n}} ; \bar{x} + 2 \frac{\sigma}{\sqrt{n}} \right]$ est un intervalle de confiance de μ à environ 95 %.
Encore faut-il connaître ou avoir une estimation de l'écart type σ .
En pratique, on prend comme estimation de σ la valeur s où s est l'écart-type de la série de mesures issues de l'échantillon.

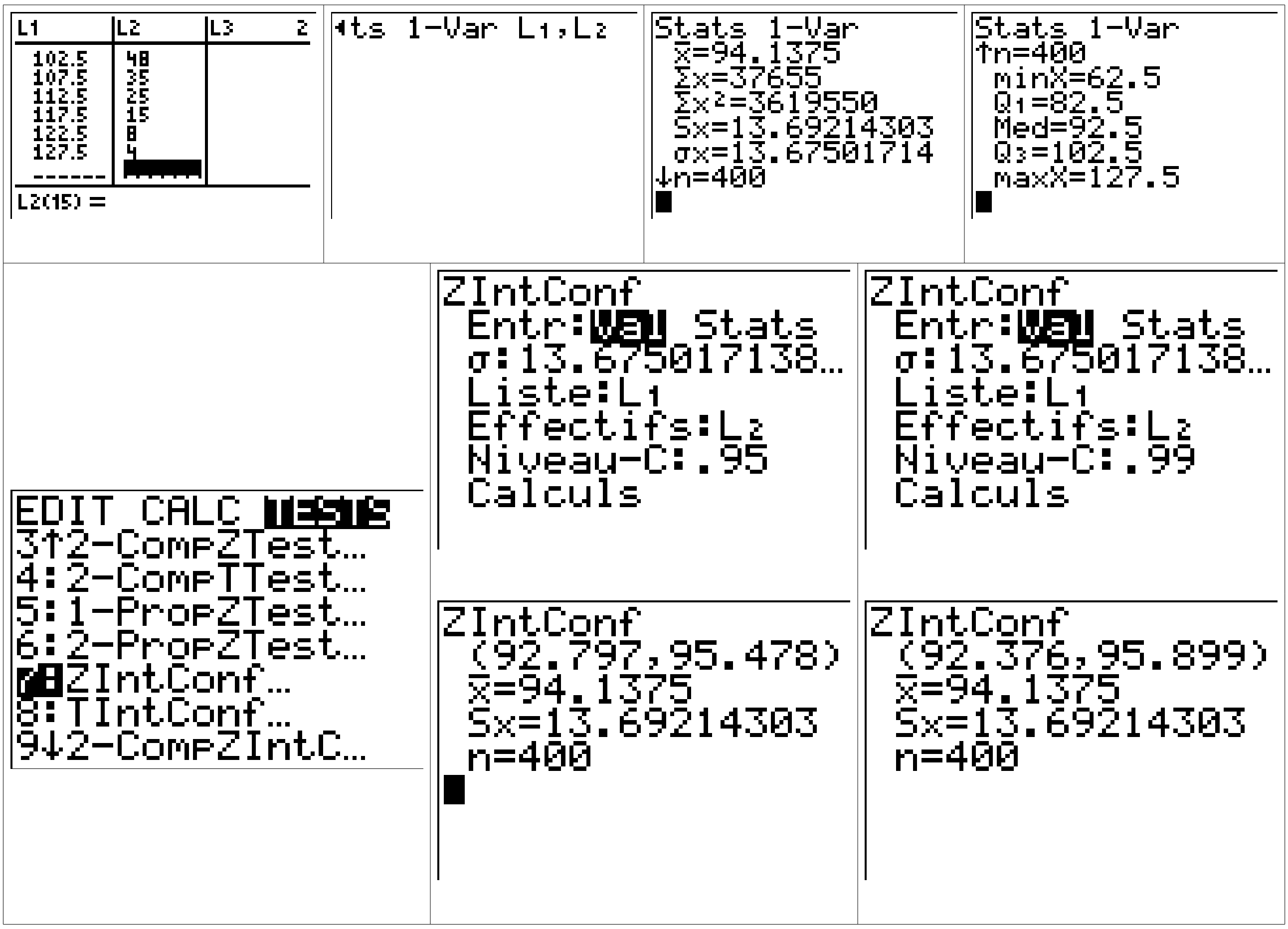
Remarque :
t_α est un quantile obtenu à partir des tables de la loi normale pour $n > 100$. Dans le cas d'échantillon plus petit, la consultation d'une table de distribution de la loi de Student est nécessaire.

Exemple :
Pour étudier l'érythroblastose, on injecte du fer radioactif par voie veineuse, on constate que sa concentration plasmatique décroît au cours du temps ; cette décroissance est caractérisée par une période T (temps en minutes au bout duquel la concentration a diminué de moitié).
Cet examen effectué sur un échantillon de 400 sujets sains a donné les résultats suivants (en remplaçant les intervalles par leur centre) :

période	62,5	67,5	72,5	77,5	82,5	87,5	92,5	97,5	102,5	107,5	112,5	117,5	122,5	127,5
nombre de sujets	5	11	18	29	40	51	57	54	48	35	25	15	8	4

On souhaite déterminer un intervalle de confiance pour la moyenne μ, au seuil de risque 5 %.

Calculatrice :



Remarque :
Pour la calculatrice, Sx est l'écart-type de l'échantillon et =x est l'écart-type « estimé » de la population :

$$\frac{\sigma^2}{n-1}=\frac{s^2}{n}$$

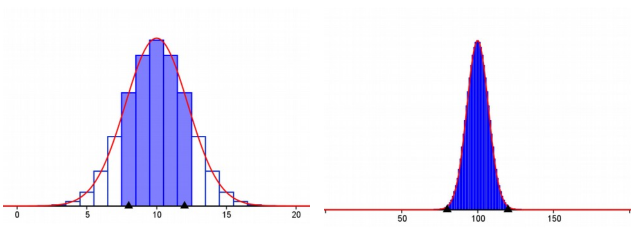
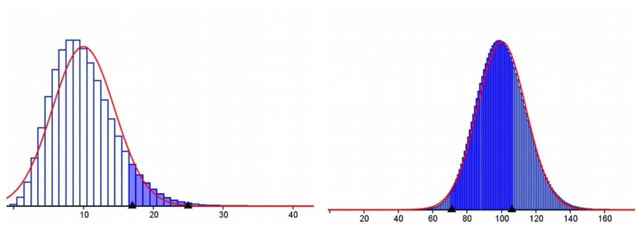
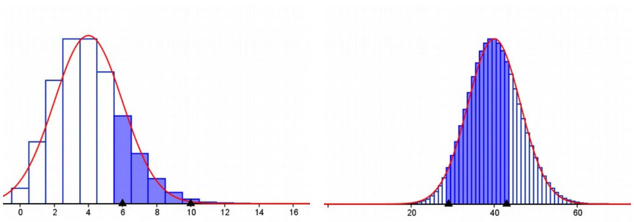
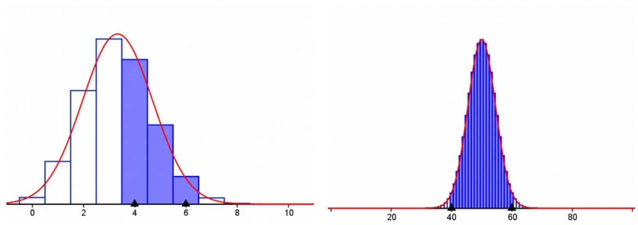
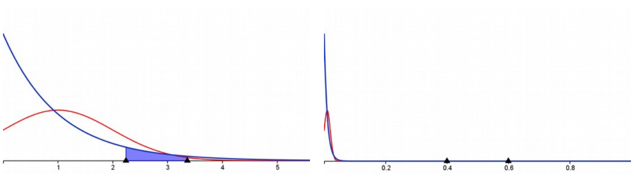
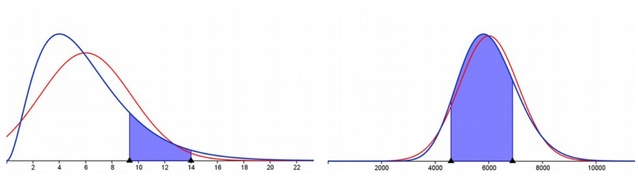
Le théorème central limite

Le **théorème central limite** (parfois appelé **théorème de la limite central**) établit la convergence en loi de la somme d'une suite de variables aléatoires vers la loi normale. Intuitivement, ce résultat affirme que toute somme de variables aléatoires indépendantes et identiquement distribuées tend vers une variable aléatoire gaussienne (qui suit une loi normale).

Soit X_1, X_2, ..., X_n une suite de variables aléatoires réelles indépendantes et suivant une même loi de probabilité ayant pour espérance μ et pour écart-type σ.
Considérons la somme $S_n = X_1 + \ldots + X_n$. L'espérance de S_n est $n\mu$ et son écart-type est $\sigma\sqrt{n}$.
On pose $F_n = \dfrac{S_n}{n}$.
Pour tout réel z :

$$\lim_{n \to +\infty} P\left(\frac{F_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leqslant z\right) = \phi(z)$$

ou ϕ est la fonction de répartition de $\mathcal{N}(0\,;1)$.

Loi binomiale 	Loi de Pascal 
Loi de Poisson 	Loi hypergéométrique 
Loi exponentielle 	Loi Gamma 

Sondage : estimation d'une proportion

On cherche une proportion p dans une population.

Pour des échantillons de taille n suffisamment grand, chaque élément de la population étant une variable aléatoire, en appliquent le théorème central limite, la variable a une moyenne p et un écart type $\sigma=\sqrt{p(1-p)}$.

t_α sera obtenu comme étant le quantile d'ordre $\frac{\alpha}{2}$ de la loi normale.

En estimant p par f (où f est la fréquence obtenue par $f=\frac{n_S}{n}$ où n_s est le nombre de succès) on pourra ainsi encadrer p.

$$P\left(f-t_\alpha\frac{\sqrt{f(1-f)}}{\sqrt{n}}\leqslant p\leqslant f+t_\alpha\frac{\sqrt{f(1-f)}}{\sqrt{n}}\right)\simeq 1-\alpha$$

<u>Exemple :</u>

Voici les résultats d'un sondage IPSOS réalisé avant l'élection présidentielle de 2002 pour le Figaro et Europe 1, les 17 et 18 avril 2002 auprès de 989 personnes, constituant un échantillon national représentatif de la population française âgée de 18 ans et plus inscrite sur les listes électorales.

On suppose cet échantillon constitué de manière aléatoire (même si, en pratique, ce n'est pas le cas).

Les intentions de vote au premier tour pour les principaux candidats sont les suivants :

- 198 personnes pour Jacques Chirac
- 178 personnes pour Lionel Jospin
- 138 personnes pour Jean Marie Le Pen

<u>Calculatrice :</u>

	Chirac	Jospin	Le Pen
	1-PropZInt x:198 n:989 Niveau-C:.95 Calculs	1-PropZInt x:178 n:989 Niveau-C:.95 Calculs	1-PropZInt x:138 n:989 Niveau-C:.95 Calculs
EDIT CALC TESTS 5↑1-PropZTest… 6:2-PropZTest… 7:ZIntConf… 8:TIntConf… 9:2-CompZIntC… 0:2-CompTIntC… A:1-PropZInt…	1-PropZInt (.17526,.22514) p̂=.2002022245 n=989	1-PropZInt (.15604,.20392) p̂=.1799797776 n=989	1-PropZInt (.11794,.16113) p̂=.1395348837 n=989