CORRECTION

1. Naissance

Au cours de l'année 1961, dans un village d'un pays d'Asie, il est né 52 enfants, parmi lesquels 37 garçons.

a) Calculer la fréquence f des naissances de garçons dans ce village puis déterminer l'intervalle de fluctuation dans lequel devrait se situer f dans 95% des cas dans l'hypothèse de 50% de garçons à la naissance. La fréquence f se situe t-elle dans cet intervalle ? Peut-on en tirer une conclusion ?

La fréquence f des naissances de garçons dans ce village est de $\frac{37}{52} \approx 0{,}7116$, soit 71,2% environ.

S'il y a 50% de garçons à la naissance, l'intervalle de fluctuation au seuil de 95% dans lequel devrait se situer f est l'intervalle $IF = [0{,}5 - \frac{1}{\sqrt{52}} ; 0{,}5 + \frac{1}{\sqrt{52}}]$ où 0,5 est la fréquence théorique des garçons dans la population des enfants. Cet intervalle est approximativement égal à $[0{,}361 ; 0{,}639]$, ce qui signifie que l'on peut s'attendre à avoir une fréquence expérimentale comprise entre ces deux valeurs (36,1% et 63,9%) dans 95% des cas. Ce résultat est *incompatible* avec la fréquence f observée dans ce village qui est de 71% environ.

On est donc en droit, pour ce qui concerne ce village d'Asie, de rejeter l'hypothèse d'une équirépartition des garçons à la naissance (hypothèse largement répandue mais souvent fausse : même en France, il ne naît pas autant de filles que de garçons).

b) Déterminer l'intervalle de confiance au seuil de 95% dans lequel devrait se situer la fréquence des garçons à la naissance dans ce pays d'Asie. Le professeur Sukimoto affirme qu'à cette époque, il naissait deux fois plus de garçons que de filles dans ce pays. Cette affirmation est-elle compatible avec l'énoncé ?

L'intervalle de confiance au seuil de 95% dans lequel devrait se situer la fréquence théorique est l'intervalle $IC = [0{,}712 - \frac{1}{\sqrt{52}} ; 0{,}712 + \frac{1}{\sqrt{52}}]$, soit $[0{,}5733 ; 0{,}8507]$. Cet intervalle contient la fréquence 0,67 qui traduit l'affirmation du professeur Sukimoto, ce qui signifie que celle-ci est compatible avec l'observation dans ce village.

Ce n'était pas demandé mais la fluctuation due au hasard, avec une fréquence théorique des garçons de 0,67 pour un village où s'observent 52 naissances est contenue, dans 95% des cas, dans l'intervalle $IF' = [0{,}67 - \frac{1}{\sqrt{52}} ; 0{,}67 + \frac{1}{\sqrt{52}}]$, soit $[0{,}5312 ; 0{,}8087]$. Il ne faut donc pas s'étonner d'avoir autant de garçons dans ce village.

c) Simuler 50 fois la naissance de 52 enfants quand il y a 67% de chance d'avoir un garçon à chaque naissance. Écrire l'algorithme ou le programme qui réalise cela et qui affiche le nombre de fois où les naissances de garçons dans l'échantillon sont au moins égales à 37. Donner un résultat d'exécution.

Voici l'algorithme demandé (il y a d'autres façons de l'écrire, nous nous sommes inspirés de l'algorithme donné à la question suivante, vous pouviez faire de même ou vous inspirer de celui donné en cours qui était présent, en principe, dans votre calculatrice) :

T = 0
Pour I allant de 1 à 50 : G = 0
  Pour J allant de 1 à 52 : R = nombre aléatoire de [0 ; 1[
    Si R ≤ 0,67 alors G=G+1
  Si G ≥ 37 Affecter T+1 à T
Afficher T

Voilà un programme Python qui traduit cet algorithme et trois résultats d'exécution. On voit que sur 50 villages où naissent 52 enfants avec une fréquence théorique de 67% de garçons à la naissance, il y en a entre 15 et 21 (cela pourrait être légèrement plus ou moins) qui ont au moins 37 garçons.

```
from random import *
total=0
taux=0.67
simulations=50
naissances=52
seuil=37
for i in range(simulations) :
  garçons=0
  for j in range(naissances) :
    if (random()<taux) : garçons=garçons+1
  if (garçons>=seuil) : total=total+1
print( "total=",total," - fréquence=",total/simulations*100,"%")
```

```
total= 21  - fréquence= 42.0 %
total= 20  - fréquence= 40.0 %
total= 15  - fréquence= 30.0 %
```

2. Nouveau-nés

a) Dans une maternité, l'étude du périmètre crânien de bébés de moins de 6 mois a donné les résultats :

x_i : périmètre crânien (cm)	40	41	42	43	44	45	46	totaux
n_i : effectifs	7	12	21	24	19	11	6	

Déterminer avec la calculatrice la moyenne $\bar{x}$, l'écart-type σ et le coefficient de variation $\frac{\sigma}{\bar{x}}$ de cette 1ère série. Sachant que $\sigma=\sqrt{\frac{\Sigma n_i x_i^2}{\Sigma n_i}-\bar{x}^2}$, justifier comment s'obtiennent $\bar{x}$ et σ (donner le calcul avec les valeurs de Σn_i, $\Sigma n_i x_i$ et $\Sigma n_i x_i^2$ que donne la calculatrice).

Série des périmètres crâniens des bébés :

Sommes des effectifs : $\Sigma n_i=100$, des valeurs : $\Sigma n_i x_i=4293$ et des carrés : $\Sigma n_i x_i^2=184547$.

- Moyenne $\bar{x}=\frac{4293}{100}=42{,}93\ cm$
- Écart-type $\sigma=\sqrt{\frac{184547}{100}-(\frac{4293}{100})^2}=\sqrt{1845{,}47-1842{,}9849}=\sqrt{2{,}4851}\approx 1{,}576\ cm$
- Coefficient de variation $\frac{\sigma}{\bar{x}}\approx\frac{1{,}576}{42{,}93}\approx 0{,}03671$, soit $\sigma\approx 3{,}7\%$ de $\bar{x}$

b) Dans la même maternité, l'étude des tailles de ces bébés a donné les résultats ci-dessous. Compléter le tableau. Déterminer la moyenne $\bar{x}'$, l'écart-type σ' et le coefficient de variation $\frac{\sigma'}{\bar{x}'}$ de cette 2ème série.

$taille$ (cm)	[45 ; 48[	[48 ; 51[	[51 ; 54[	[54 ; 57[
x_i : $taille\ moyenne$ (cm)	46,5	49,5	52,5	55,5
n_i : effectifs	20	70	6	4

Série des tailles des nouveau-nés :

Sommes des effectifs : $\Sigma n_i=100$, des valeurs : $\Sigma n_i x_i=4932$ et des carrés : $\Sigma n_i x_i^2=243621$.

- Moyenne $\bar{x}=\frac{4932}{100}=49{,}32\ cm$
- Écart-type σ $\sqrt{\frac{243621}{100}-(\frac{4932}{100})^2}=\sqrt{2436{,}21-2432{,}4624}=\sqrt{3{,}7476}\approx 1{,}936$
- Coefficient de variation $\frac{\sigma}{\bar{x}}\approx\frac{1{,}936}{49{,}32}\approx 0{,}039254$, soit $\sigma\approx 3{,}9\%$ de $\bar{x}$

c) Comparer les distributions de ces deux grandeurs (taille et périmètre crânien).

Les tailles des nouveau-nés et les périmètres crâniens des bébés de moins de 6 mois sont comparables car les moyennes sont proches 49,32 cm et 42,93 cm. La taille moyenne d'un nouveau-né est légèrement supérieure au périmètre crânien moyen d'un bébé mais la différence n'est que de 6,39 cm.

L'écart-type des tailles (1,936 cm) est supérieur à celui des périmètres crâniens (1,576 cm) mais pour pour mieux comparer les dispersions, il faut examiner les valeurs relatives : le coefficient de variation des périmètres crâniens (0,036711) est très légèrement plus faible que celui des tailles (0,039254), soit des valeurs de 3,7% et 3,9% de la moyenne environ, des valeurs assez proches pour pouvoir affirmer que la dispersion des tailles et des périmètres crâniens est sensiblement la même.

3. Enfants de 3 ans

a) Une étude récente montre la répartition des tailles chez des enfants de 3 ans :

$taille$ (cm)	[78 ; 82[	[82 ; 86[	[86 ; 90[	[90 ; 94[	[94 ; 98[	[98 ; 102[
Fréquences (en %)	10	17	20	23	19	11
cumul croissant des fréquences	10	27	47	70	89	100

Compléter ce tableau puis déterminer la médiane M par le calcul.

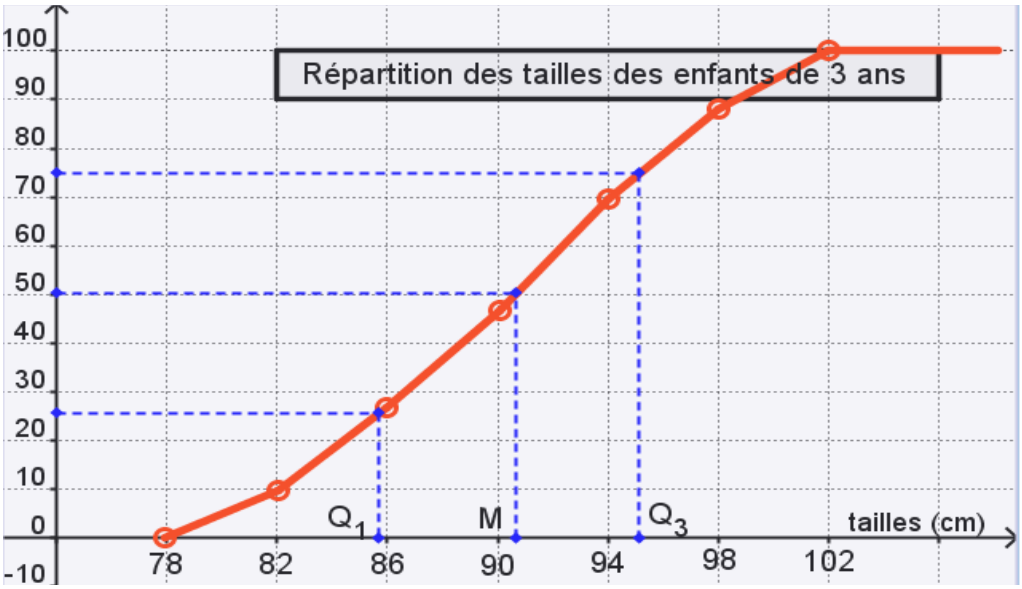
D'après les valeurs du tableau pour les tailles des enfants de 3 ans on a $M=90+4\times\frac{50-47}{23}\approx 90{,}52\ cm$.

b) Représenter la série des effectifs cumulés sur le graphique ci-dessous. Mettre en évidence sur ce graphique les tracés qui permettent de lire directement médiane et quartiles. Déterminer graphiquement les quartiles Q_1 et Q_3, puis calculer l'écart inter-quartile relatif

Désolé, j'avais écrit « effectifs » mais il fallait lire « fréquences », ou prendre les effectifs comme des fréquences.

Le polygone des fréquences cumulées est dessiné ci-dessous. On lit les quartiles : $Q_1\approx 85{,}5$ et $Q_3\approx 95$

(on peut accepter quelques différences ici puisque cela dépend de la construction et de la lecture du graphique, deux opérations approximatives). On en déduit l'écart inter-quartile relatif pour les tailles des enfants de 3 ans $\frac{Q_3-Q_1}{M}=\frac{95-85,5}{90,52}\approx 0,105$, l'écart inter quartile représente donc 10% environ de la médiane de cette série.



c) La série des tailles des nouveau-nés est caractérisée, quant à elle, par $M\approx 49,4$, $Q_1\approx 48,2$ et $Q_3\approx 50,5$ (valeurs déduites du tableau 2b). Comparer les deux séries de tailles (nouveau-nés et enfants de 3 ans) sur leurs valeurs centrales et leurs dispersions (absolues et relatives).

Les tailles des nouveau-nés et celles des enfants de 3 ans sont très différentes pour ce qui est de la valeur centrale : la médiane est, en effet, de 90,5 *cm* pour les enfants de 3 ans alors qu'elle vaut juste un peu plus de la moitié (49,5 *cm*) pour les nouveau-nés. Tout cela est, bien sûr, tout-à-fait normal et prévisible.

Pour ce qui est de la dispersion, il y a aussi une grande différence, du simple au double : l'écart inter-quartile relatif pour les tailles des nouveau-nés est $\frac{Q_3-Q_1}{M}=\frac{50,5-48,2}{49,4}\approx 0,0466$, soit moins de 5% de la médiane de cette série. Pour les enfants de 3 ans, il est de 10,5% soit plus du double. Les enfants de 3 ans ont donc des tailles 2 fois plus dispersées que celles des nouveau-nés. Ceux-ci sont, en gros tous à peu près aussi petit alors qu'à 3 ans, la croissance ne s'effectuant pas de la même façon et à la même vitesse pour tous les bébés, des écarts plus importants se relèvent.

4. Enfants de 12 ans

Une enquête a été effectuée pour étudier les durées du trajet domicile-collège pour deux collèges parisiens. Les résultats de cette enquête sont consignés (en minutes) dans le tableau suivant.

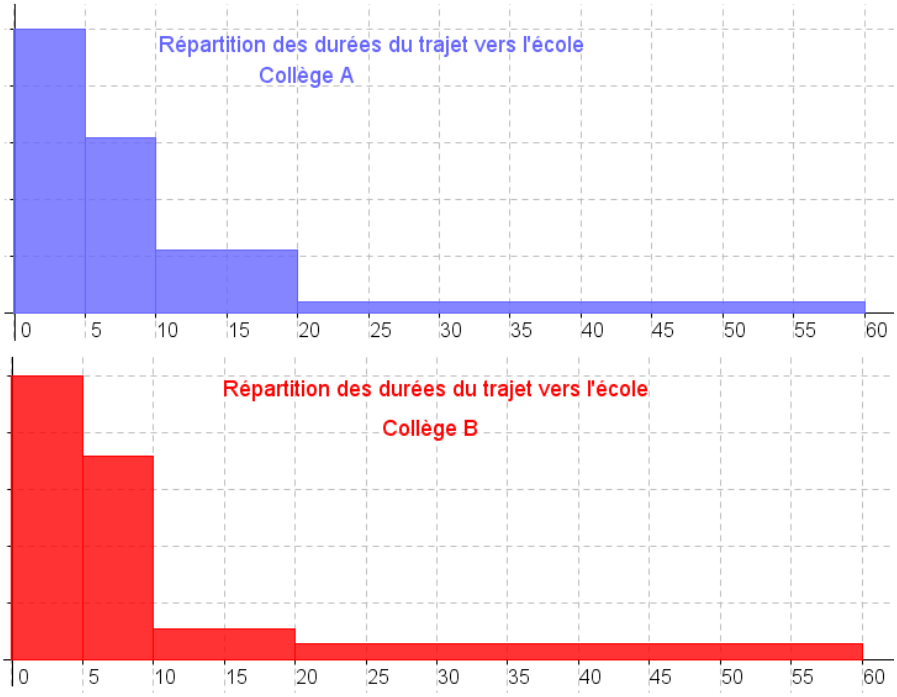
	Durées (*min*)	Moins de 5	De 5 à 10	De 10 à 20	Plus de 20	Totaux
	Amplitudes	5	5	10	40	-
Collège A	Effectifs	32	20	14	10	76
	Fréquences (en %)	42	26	18	13	100
	Densités	6,4	4	1,4	0,25	-
	Hauteurs	10	6,2	2,2	0,39	-
Collège B	Effectifs	65	47	14	29	154
	Fréquences (en %)	42	31	8	19	100
	Densités	13	9,4	1,4	0,73	-
	Hauteurs	10	7,2	1,1	0,56	-

a) En prenant 0 *min* comme minimum et 1 *h* comme maximum, déterminer les amplitudes, fréquences et densités de chaque classe de durées pour les deux collèges.

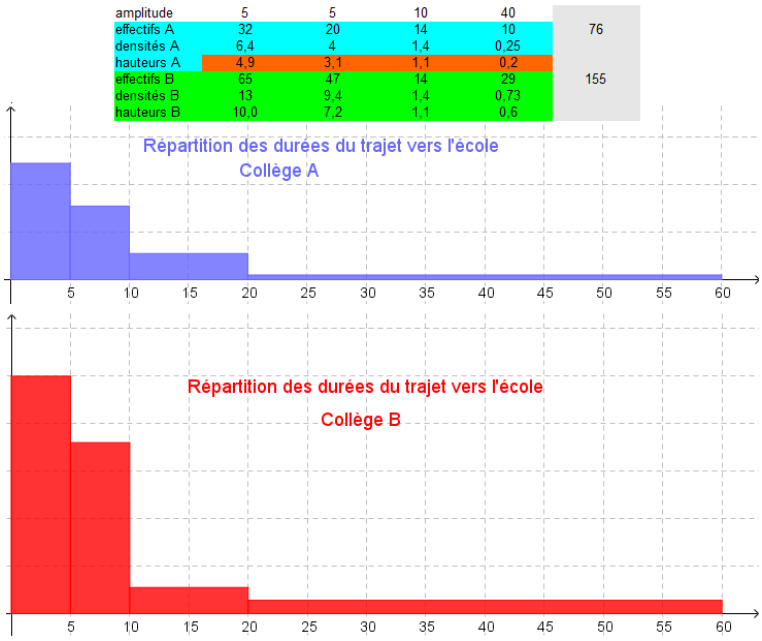
Nous avons effectué directement les calculs dans le tableau.

b) Pour la représentation ces deux séries sous forme d'histogrammes, on choisit de donner aux rectangles représentant la classe de plus grande densité, une hauteur de 10 *cm*. Déterminer les autres hauteurs.

Tracer côte-à-côte les deux histogrammes.
Pour le 1er rectangle qui a, à chaque fois la plus grand densité, on donne une hauteur de 10 *cm*. Cela conduit à appliquer un coefficient de proportionnalité de 10/6,4 pour le collège A et de 10/13 pour le collège B.
Les hauteurs se calculent alors sans difficulté et on peut tracer les histogrammes.



Certains d'entre vous m'ont posé la question « est-ce que c'est 10 *cm* pour la plus grande densité de chaque série ou bien pour la plus grande densité globalement ? » J'ai traité la question comme si c'était évident que c'était « 10 *cm* pour la plus grande densité de chaque série » (ce qui me semble indiqué par le pluriel dans « on choisit de donner **aux rectangles** représentant la classe de plus grande densité ») mais sinon, c'était une bonne idée de donner « 10 *cm* pour la plus grande densité globalement ». On obtient dans ce cas, quelque chose comme ce qui est à droite : les données du collège A seules sont modifiées. Les hauteurs recalculées avec le même coefficient que le collège B. Ainsi les deux histogrammes peuvent être comparés en grandeurs absolues (le A ayant deux fois moins d'élèves forme des rectangles deux fois moins haut) alors que l'autre méthode conduisait à comparer les deux collèges en valeurs relatives



c) Les affirmations suivantes sont-elles intéressantes, discutables ou fausses ? (Justifier)

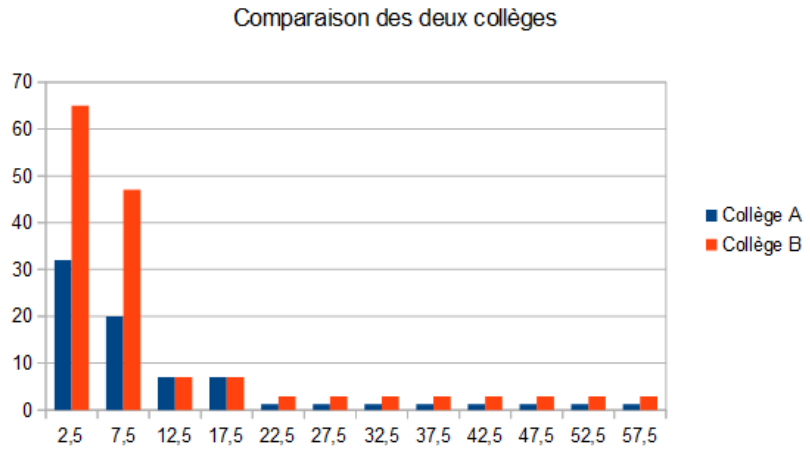
- Il y a autant d'élèves dans chaque collège dont le trajet dure entre 10 et 20 *mn*.

Cette affirmation est *discutable* : s'il est bien vrai qu'il y a 10 élèves dans cette classe pour les deux collèges, en proportion du total des élèves cela ne représente pas la même chose, relativement aux nombres d'élèves de chaque collège. Cela représente 18% pour le collège A et 8% seulement pour le collège B. Cette différence apparaît bien sur les histogrammes, les deux rectangles n'ayant pas la même hauteur.

- Il y a la même proportion d'élèves dans chaque collège dont le trajet dure moins de 15 *mn*.

Il faut faire le compte : pour le collège A cela représente 42+26+18/2 (on ne compte que la moitié de l'effectif de la dernière classe)=77% tandis que pour le collège B cela représente 42+31+8/2=77%. Ces deux valeurs sont bien exactes, au % près. L'affirmation est *intéressante* : comme 77% est proche de 75% on peut en déduire que les deux séries ont le même troisième quartile.

NB : Si on veut comparer les deux collèges sur ce point, la séparation des histogrammes n'est pas très parlante. Mieux vaut opter pour une représentation couplée des deux séries comme dans cette illustration (j'ai découpé les classes pour qu'elles aient la même amplitude afin d'utiliser le tableur pour cette représentation, ce qui n'est pas forcément une bonne idée).



5. Nombre d'enfants

Édouard est fier d'avoir écrit le programme ci-contre en Python.
Il dit que son algorithme permet de simuler *nF* familles de *nE* enfants lorsque la fréquence des filles à la naissance est de *tx*. Il se rappelle que ce programme compte certains types de familles mais il ne sait plus lesquels.

On rappelle que : random() *est une fonction du module* random *qui retourne un nombre aléatoire de l'intervalle [0 ; 1[. L'instruction* for I in range(n) : *est la traduction en Python de « pour I allant de 0 à n−1 ». Cette instruction conduit à une indentation (un retrait) de tout ce qui doit être exécuté dans cette boucle.*

a) Édouard a déterminé la fréquence *tx* en se basant sur l'information suivante : en France, il naît 105 garçons pour 100 filles. Vérifier le calcul d'Édouard et expliquer comment est simulé la naissance d'une fille ?

```
from random import *
y=0
z=0
tx=0.4878
nF=1000
nE=5
for I in range(nF) :
  x=0
  for J in range(nE) :
    bb= random()
    if (bb<tx) : x=x+1
  if (x==nE) : y=y+1
  z=z+x
print( y,100*z/nF/nE)
```

Il naît en France 105 garçons pour 100+105=205 enfants. Cela fait une fréquence des garçons de 51,22% car $\frac{105}{205}\approx 0{,}5122$ et une fréquence des filles de 48,78% car $\frac{100}{205}\approx 0{,}4878$ (ou $1-0{,}5122=0{,}4878$).
Ce dernier nombre a été choisi par Édouard pour simuler la naissance d'une fille : il tire un nombre aléatoire de [0 ; 1[et si ce nombre est inférieur à 0,4878 il en tire la conclusion qu'il s'agit d'une fille (les deux événements ont la même probabilité). Les instructions *bb=random()* et *if (bb<tx):x=x+1* réalisent cela (avec l'affectation *tx=0.4878* qui fixe la fréquence des filles).

b) Examiner cet algorithme et dire ce que représentent les variables *x*, *y* et *z*.

- *x* est le **nombre de filles** qui naissent dans cette simulation de famille à *nE*=5 enfants (à cause de ce qu'on a dit plus haut sur l'instruction *if (bb<tx):x=x+1*).
- *y* est le **nombre de familles à *nE*=5 enfants ayant exclusivement des filles** (à cause de l'instruction *if (x==nE) : y=y+1* qui signifie que si le nombre de filles dans la famille est égale au nombre d'enfants générés alors on incrémente ce compteur de 1).
- *z* est le **nombre total de filles** qui sont nées dans les *nF*=1000 familles à *nE*=5 enfants (à cause de l'instruction *z=z+x* qui ajoute le nombre de filles de chaque famille à *nE*=5 enfants générée).

c) Une exécution de ce programme affiche le résultat suivant : 35 49.92.
Que représentent ces nombres 35 et 49,92 ?

- Le nombre 35 signifie que sur les *nF*=1000 familles de *nE*=5 enfants générées, il y a eu 35 familles qui ont exclusivement eu des filles.
- Le nombre 49,92 signifie que sur les *nF*=1000 familles de *nE*=5 enfants générées, la fréquence des filles a été de 49,92%. Le calcul est, en effet, une fréquence $100\times z\div nF\div nE$ soit $z\div(nF\times nE)$ qui est multiplié par 100 pour avoir un taux de pourcentage. Le nombre total de filles est divisé par le nombre total d'enfants (*nF*=1000 familles de *nE*=5 enfant, cela fait 5000 enfants générés et z est le nombre de filles parmi ces 5000 enfants). Ce nombre est proche de la fréquence théorique des filles (48,78%), les fluctuations dues au hasard étant limitées car l'effectif est important.

d) Déterminer, à partir de cette exécution, l'intervalle de confiance contenant en théorie le nombre y dans 95% des cas. Anna dit qu'elle a obtenu le résultat suivant : 0 50.0. Que pensez-vous de ce zéro obtenu ?

Il y a eu 35 familles sur 1000 qui ont eu exclusivement des filles, soit une fréquence expérimentale de 0,035. L'intervalle de confiance que l'on peut déduire de ce résultat est $[0{,}035-\frac{1}{\sqrt{1000}};0{,}035+\frac{1}{\sqrt{1000}}]$, soit $[0{,}003377;0{,}066623]$. Multiplions par 1000 les bornes, on trouve $[3{,}377;66{,}623]$. Le nombre de familles sur 1000 qui ont eu exclusivement des filles est compris entre 3 et 67 dans 95% des cas. Si quelqu'un répond 0 ou 2, cela est peu crédible... Anna ayant répondu qu'il y a eu 0 familles sur 1000 qui ont eu exclusivement des filles a certainement inventé ce résultat, ou bien son programme n'est pas au point et ne fait pas ce qu'il devrait (par exemple, elle a oublié d'écrire l'instruction *if (x==nE) : y=y+1* ou bien elle a écrit l'instruction *if (x==nF) : y=y+1* et cela n'arrive jamais puisque *nF*=1000...)

Remarquons tout de même que nous ne sommes pas dans le domaine d'application de la formule : la fréquence de 3,5% est trop faible (elle doit être en principe, au moins de 20%). On devrait donc être moins catégorique sur nos conclusions.

Bonus (2 pts) : Déterminer un tableau contenant la valeur de y pour un entier nE compris entre 0 et 9. Donner la valeur de nF utilisée (cela peut être moins de 1000 si la simulation prend trop de temps)

nE	0	1	2	3	4	5	6	7	8	9
y	1000*	472	250	109	47	26	16	3	5	0

* le programme d'Édouard ne permet pas d'aboutir dans ce cas car il y a une division par 0 dans la dernière instruction, mais c'est évident que pour les familles sans enfants, le nombre de familles ayant eu exclusivement des filles n'a pas de sens. Pour suivre la décroissance observée, on peut mettre 1000 (elles ont eu exactement le même nombre de filles que d'enfants, c'est-à-dire 0).

Ces valeurs sont le produit d'une simulation. Elles ont une valeur indicative certaine (on pourrait donner pour chacune l'intervalle de confiance qui s'y rattache). Pour autant, les valeurs théoriques peuvent être obtenues par une approche probabiliste (voir le chapitre 8 qui traite de probabilité).