

Intervalles de fluctuation

Raymond Moché

2/1S/TS

Résumé : Les logiciels de calcul actuels permettent de calculer rapidement la probabilité exacte d'un intervalle de fluctuation.

Cette facilité est utilisée intensivement pour ré-évaluer la validité de la recette (7). On s'aperçoit qu'elle est sans doute trop optimiste et contient une restriction sur p ($0.2 \leq p \leq 0.8$) qui est plutôt incompréhensible. C'est la partie utile et facile du papier.

On parle ensuite, dans la quatrième partie, de deux types de résultats qui pourraient conduire à une démonstration de la recette, obtenant ainsi des théorèmes inapplicables pratiquement mais qui, au moins, sont prouvés.

La liaison avec les intervalles de confiance occupe la dernière partie.

Les calculs ont été faits avec « scilab ». Nos commentaires et conclusions dépendent largement de la qualité des calculs, que nous sommes bien incapable d'évaluer. Ces calculs en soi ne sont pas probants. Ils permettent certainement de se faire une bonne idée de la situation ; ils peuvent aussi servir à démontrer, dans le cas considéré et moyennant une maîtrise suffisante des approximations, que certaines inégalités sont fausses.

Table des matières

1 Intervalles de fluctuation, sans théorie	1
1.1 Définitions	1
1.2 Comment on utilise l'intervalle de fluctuation : recette	2
2 Calcul exact de $P(E)$	3
3 Évaluation de $P(E)$ par expérimentation et simulation	15
4 Que peut-on démontrer sur la recette ?	16
4.1 Ce que dit le programme	16
4.2 Défaillance du contrôle de l'erreur	17
5 D'où vient l'intervalle de fluctuation ? Intervalles de confiance	20

1 Intervalles de fluctuation, sans théorie

1.1 Définitions

On répète une expérience aléatoire \mathcal{E} , les répétitions étant indépendantes les unes des autres (par exemple parce que l'on a pris soin de se remettre dans les mêmes conditions initiales à chaque répétition) et on s'intéresse aux réalisations d'un événement A de probabilité p ($0 < p < 1$).

Si on note X_i la variable aléatoire qui vaut 1 si A est réalisé au $i^{\text{ème}}$ coup, 0 sinon,

$$(1) \quad S_n = X_1 + \dots + X_n$$

est la variable aléatoire égale au nombre de réalisations de A au cours des n premières répétitions de l'expérience.

Simulation

Pour simuler ces n répétitions de \mathcal{E} , il suffit de simuler S_n . Pour cela, il est utile de savoir que l'on peut

démontrer que S_n suit la loi binomiale de taille n et de paramètre p , habituellement notée $\mathcal{B}(n, p)$, ce qui signifie que S_n est une variable aléatoire qui peut prendre les valeurs $0, 1, \dots, n$ avec les probabilités suivantes :

$$(2) \quad \text{pour } k \in \{0, 1, \dots, n\}, \quad P(S_n = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \quad \text{où } q = 1 - p$$

On s'intéressera plutôt à la fréquence de réalisation de l'événement A qui est

$$(3) \quad F_n = \frac{S_n}{n}$$

On sait que si l'on recommence la série des n répétitions de \mathcal{E} , F_n va fluctuer. F_n est une variable aléatoire.

Intervalle de fluctuation [6] ou [7], p. 16.

Le programme appelle *intervalle de fluctuation de F_n au seuil de 95%*¹ l'intervalle

$$(4) \quad \left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$$

C'est un intervalle centré en p d'amplitude $\frac{2}{\sqrt{n}}$. Il va se révéler très intéressant. Pour cela, considérons l'événement

$$(5) \quad E : \text{« } F_n \text{ appartient à l'intervalle de fluctuation » ou } E = \left\{ p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}} \right\}$$

L'événement E sera plus tard noté E_n ². Il a la propriété sensationnelle suivante : sous des conditions peu contraignantes, voir (7), $P(E) \geq 0.95$ ³.

On sera amené à écrire l'événement E sous d'autres formes, par exemple :

$$(6) \quad E = \left\{ F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}} \right\} = \left\{ p \in \left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}} \right] \right\}$$

Les bornes de l'intervalle $\left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}} \right]$ sont des variables aléatoires. C'est un intervalle aléatoire, contrairement à l'intervalle de fluctuation.

1.2 Comment on utilise l'intervalle de fluctuation : recette

Si p est inconnu, que peut-on déduire de la valeur (expérimentale) f_n prise par la variable aléatoire F_n à l'issue de n répétitions de \mathcal{E} ?

Recette⁴

On peut lire dans le document [7], p. 16, que

$$(7) \quad \boxed{\text{si } n \geq 25 \text{ et } 0.2 \leq p \leq 0.8, \text{ alors } P(E) \geq 0.95}$$

Ce que l'on utilise comme suit : *Si l'on pense qu'il est raisonnable de courir le risque de se tromper avec une probabilité de 0.05, on considèrera que l'événement E est réalisé* (à l'issue des n répétitions de \mathcal{E}), autrement dit, on considèrera que la double inégalité

$$(8) \quad f_n - \frac{1}{\sqrt{n}} \leq p \leq f_n + \frac{1}{\sqrt{n}}$$

1. « au seuil de 95% » est sous-entendu dans la suite.

2. La notation correcte serait $E_{n,p}$ car il dépend de n et de p .

3. On verra plus loin que l'on peut avoir des doutes.

4. En statistique, ce problème s'appelle *estimation par intervalle de confiance d'une proportion*, voir [2], p. 306-308.

est satisfaite, ce qui localise la valeur inconnue de p dans un intervalle de longueur $\frac{2}{\sqrt{n}}$, centré en f_n (la localisation sera d'autant plus précise que n sera plus grand).

Mise en garde

Supposons que $n = 100$ et $f_n = 0.42$. Si la vraie valeur de p est 0.45 , la double inégalité (8) est **vraie** ; si $p = 0.3$, elle est **fausse**. Une double inégalité entre 3 nombres est soit vraie, soit fausse. Il n'y a pas d'autre cas. Mais p étant inconnu, on fera comme si (8) était vérifié.

Il ne faut pas oublier que l'on fait ainsi un pari : on se dit que E étant un événement de probabilité au moins égale à 0.95 , ce qui est « grand », on parie qu'il est réalisé à l'issue de l'expérimentation et du calcul de f_n .

Quand on évalue une probabilité inconnue p à l'aide de l'intervalle de fluctuation, on doit dire que l'on parie que l'événement E est réalisé, parce que sa probabilité est au moins égale à 0.95 , et qu'il en résulte que la double inégalité (8) est vraie, ce qui nous donne un encadrement de p .

Cet encadrement n'est pas sûr. Il est vrai si l'événement E est réalisé.

2 Calcul exact de $P(E)$

Dans le but de tester la recette (7), calculons exactement $P(E)$ en écrivant E sous la forme :

$$(9) \quad E = \{np - \sqrt{n} \leq S_n \leq np + \sqrt{n}\}$$

Si on appelle a et b le plus petit et le plus grand entiers de l'intervalle $[np - \sqrt{n}, np + \sqrt{n}]$, autrement dit si l'on pose⁵

$$(10) \quad a = \max(-[\sqrt{n} - np], 0) \quad \text{et} \quad b = \min([\sqrt{n} + np], n)$$

il est clair, d'après l'additivité de la probabilité P , que

$$(11) \quad P(E) = \binom{n}{a} \cdot p^a \cdot q^{n-a} + \dots + \binom{n}{b} \cdot p^b \cdot q^{n-b}$$

$P(E)$ est donc la différence de 2 valeurs de la fonction de répartition de la loi binomiale $\mathcal{B}(n, p)$. Les calculs sont ultra-rapides : la fonction « BinInt(n, p, a, b) » décrite ci-dessous donne la probabilité pour qu'une variable aléatoire suivant la loi $\mathcal{B}(n, p)$ tombe dans l'intervalle $[a, b]$, autrement dit, prenne l'une des valeurs a, \dots, b .

Listing 1 – Algorithme BinInt.sci

```
function y=BinInt(n,p,a,b) // a,b,n entiers ; 1<=n, 0<=a<=b<=n ; 0<p<1.
c=ones(1,n+1);
FRBin=[0,cdfbin("PQ",0:n,n*c,p*c,(1-p)*c)]; // help cdfbin
y=FRBin(b+2)-FRBin(a+1);
endfunction ;
```

L'algorithme « GraphesPE » ci-dessous trace le graphe de la fonction $p \mapsto P(E)$ successivement quand $n = 15, 16, \dots, 24$, p prenant les valeurs $0.00001, 0.00002, \dots, 0.99999$:

Listing 2 – Algorithme GraphesPE.sce

```
// Graphes de P(En) lorsque n varie de 15 à 24.
clear ;
clf ;
```

5. $[x]$ désignant la partie entière de x

```

p=linspace(0.00001,0.99999,99999);
N=15:24;
for i=1:10
    n=N(i);
    PE=[];
    c=ones(1,n+1);
    for j=1:99999
        FRBin_n=[0,cdfbin("PQ",0:n,n*c,p(j)*c,(1-p(j))*c)];
        b=min(floor(sqrt(n)+n*p(j)),n);
        a=max(-floor(sqrt(n)-n*p(j)),0);
        PE=[PE,FRBin_n(b+2)-FRBin_n(a+1)];
    end;
    subplot(5,2,i)
    plot(p,PE);
    plot([0,1],[0.95,0.95])
end;

```

Les dix graphes obtenus se lisent ligne par ligne, de la gauche vers la droite sur la figure (1). Nous avons recommencé ce travail aux page suivantes, d'abord en faisant varier n de 25 à 34, puis de 35 à 44, *etc.*

Les graphes obtenus ont une discontinuité aux valeurs de p qui font changer le nombre de termes de la somme (11); de plus, entre deux points de discontinuité, il s'agit de graphes de polynôme. Mais en fait, seuls sont calculés les points correspondant aux valeurs de p variant de 0.00001 à 0.99999 avec un pas de 0.00001 et deux points consécutifs sont reliés par un segment de droite (ce qui se voit à l'aide de grossissements lors de l'affichage sur un écran). Ces graphes sont donc des lignes polynomiales. S'il y a une discontinuité entre deux points consécutifs calculés, elle n'apparaîtra pas. Remarquons aussi que *nous ignorons la qualité de la précision des calculs.*

On peut quand même penser que si un graphe passe visiblement en dessous de la droite $y = 0.95$, l'inégalité $P(E) \geq 0.95$ n'est pas satisfaite (pour la valeur de n considérée et pour certaines valeurs de p).

Notre conclusion en regardant les graphes qui suivent est double :

- Si $n \leq 114$ (c'est là que nous avons arrêté les calculs), l'inégalité $P(E) \geq 0.95$ ne sera pas satisfaite pour certaines valeurs de p .
- La condition $0 \leq p \leq 0.8$ semble étrangère au problème. Selon ces calculs, la recette (7) devrait être rectifiée. On pourrait énoncer, sans que cela soit démontré :

$$(12) \quad \boxed{\text{Si } n \geq 25, \quad P(E) \geq 0.93}$$

ce qui n'est pas si mal.

Figure 1 Graphes de $P(E)$ pour $n = 15, 16, \dots, 24$.

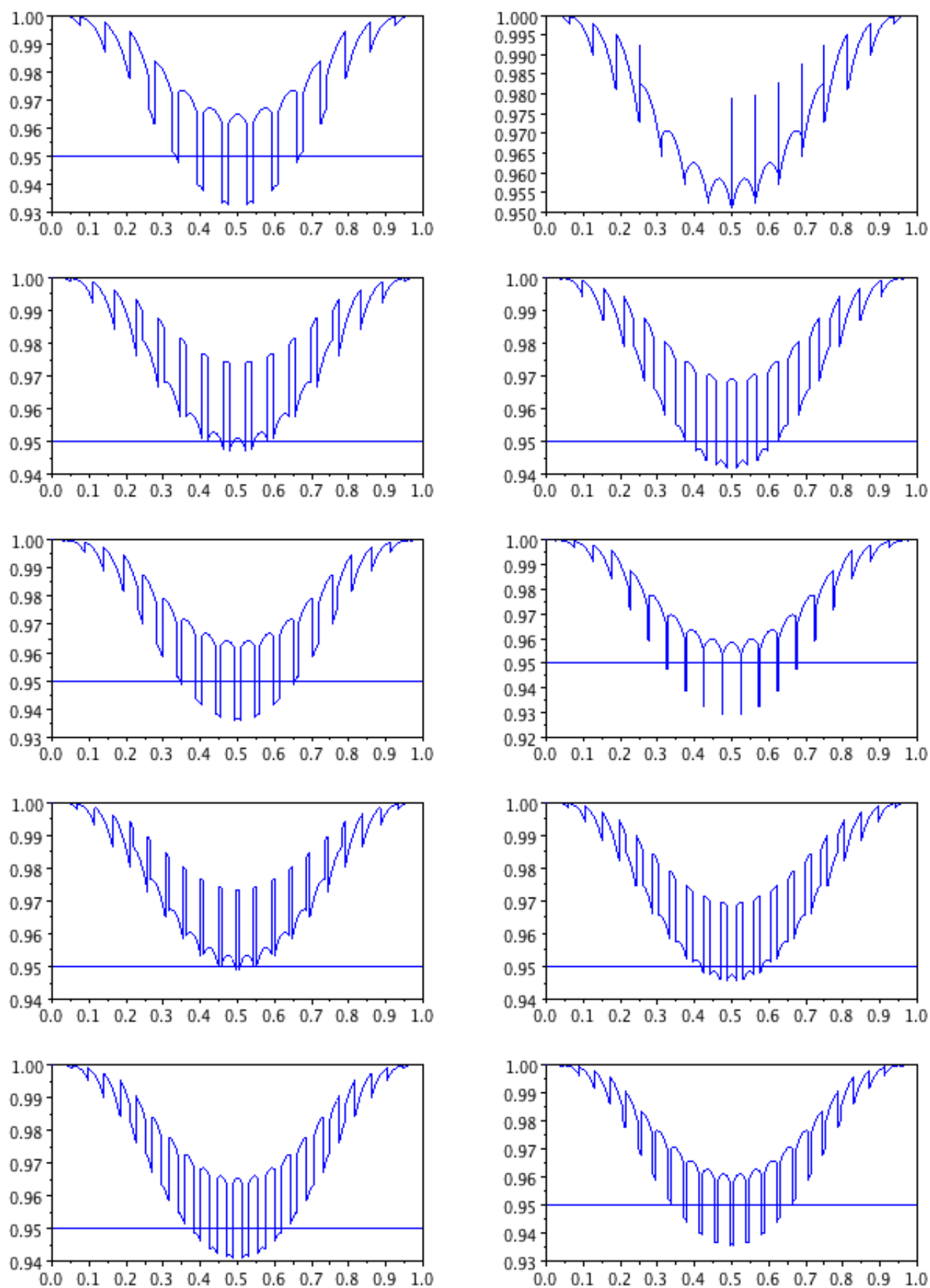


Figure 2 Graphes de $P(E)$ pour $n = 25, 26, \dots, 34$.

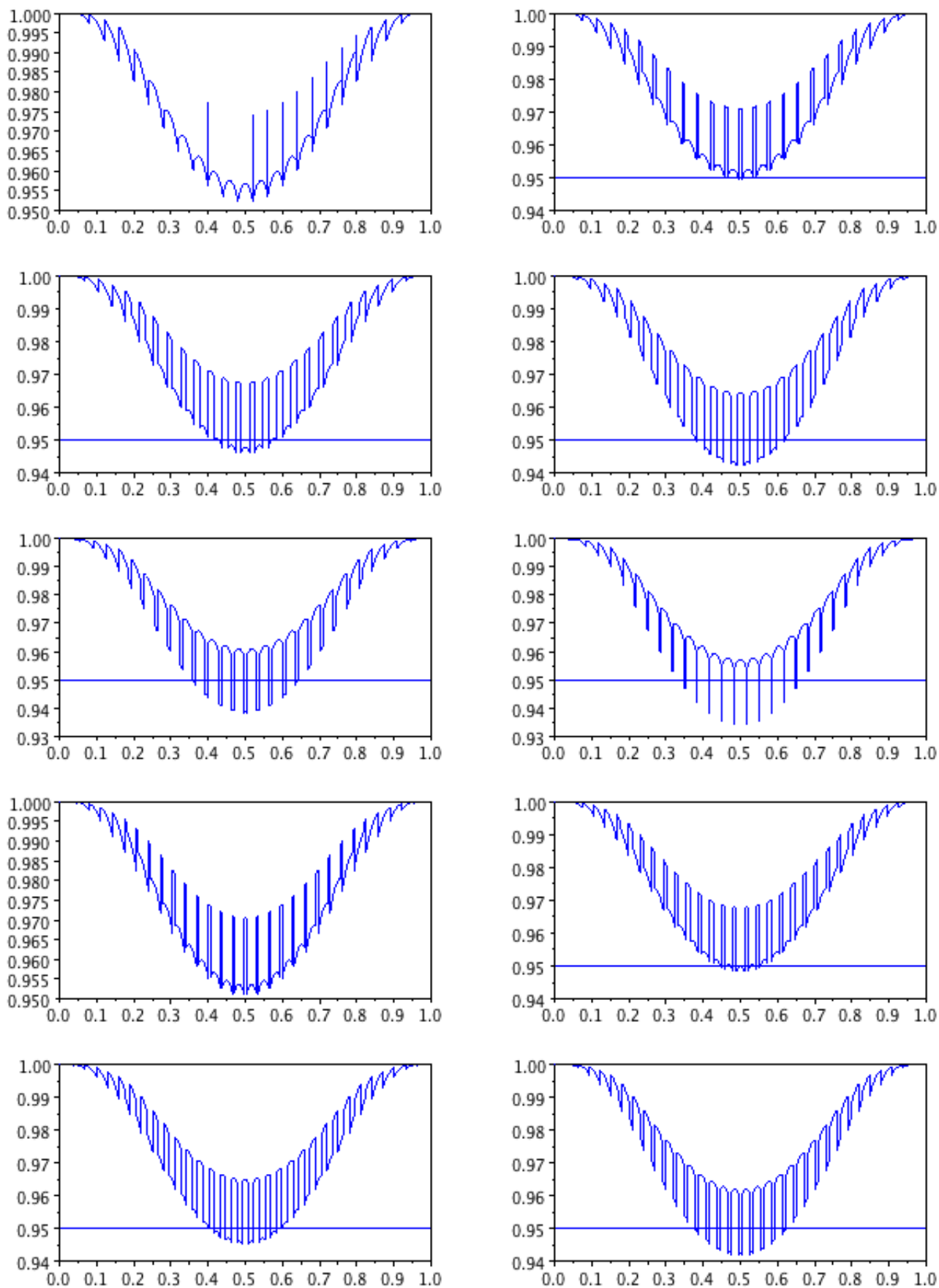


Figure 3 Graphes de $P(E)$ pour $n = 35, 36, \dots, 44$.

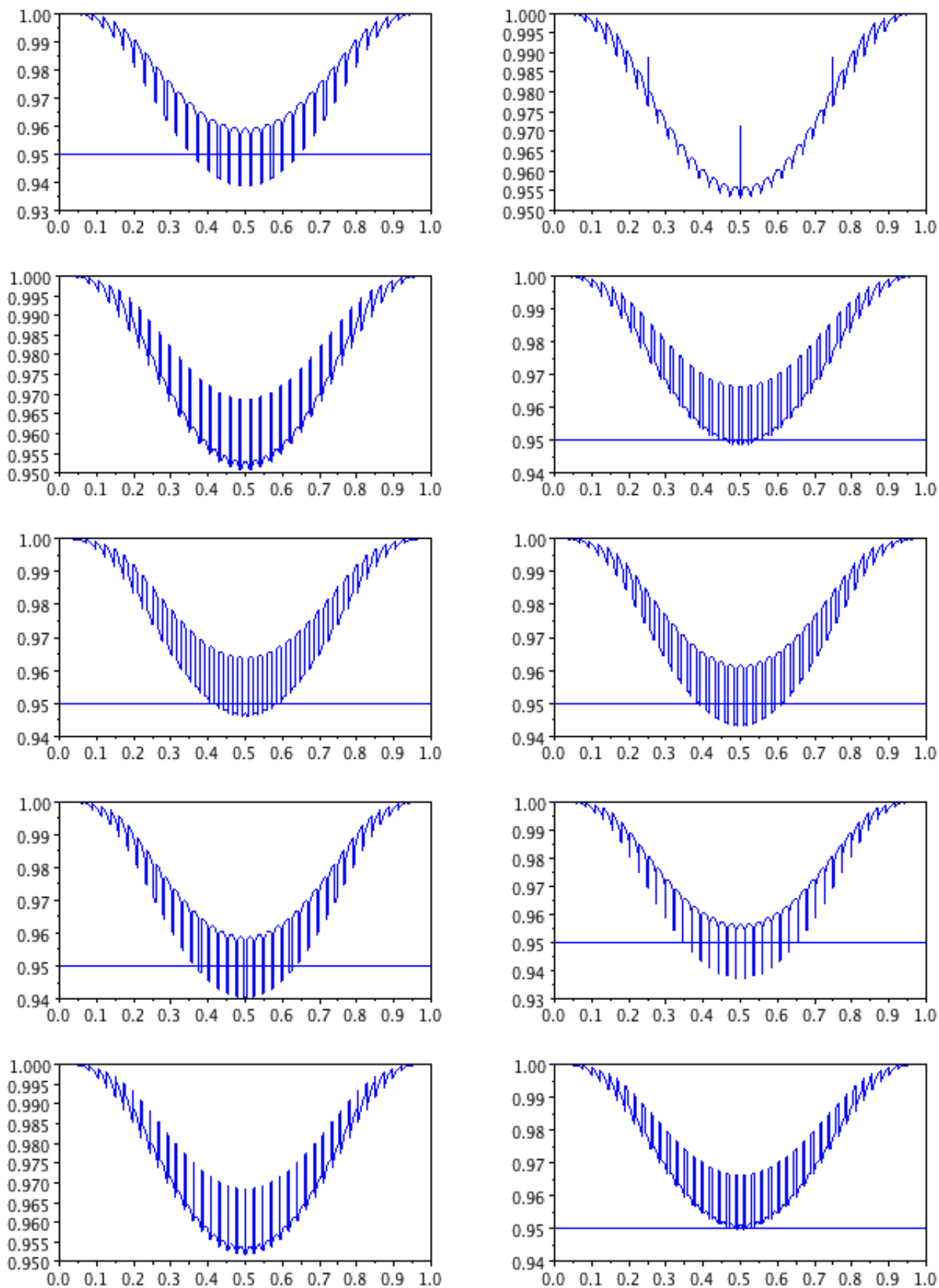


Figure 4 Graphes de $P(E)$ pour $n = 45, 46, \dots, 54$.

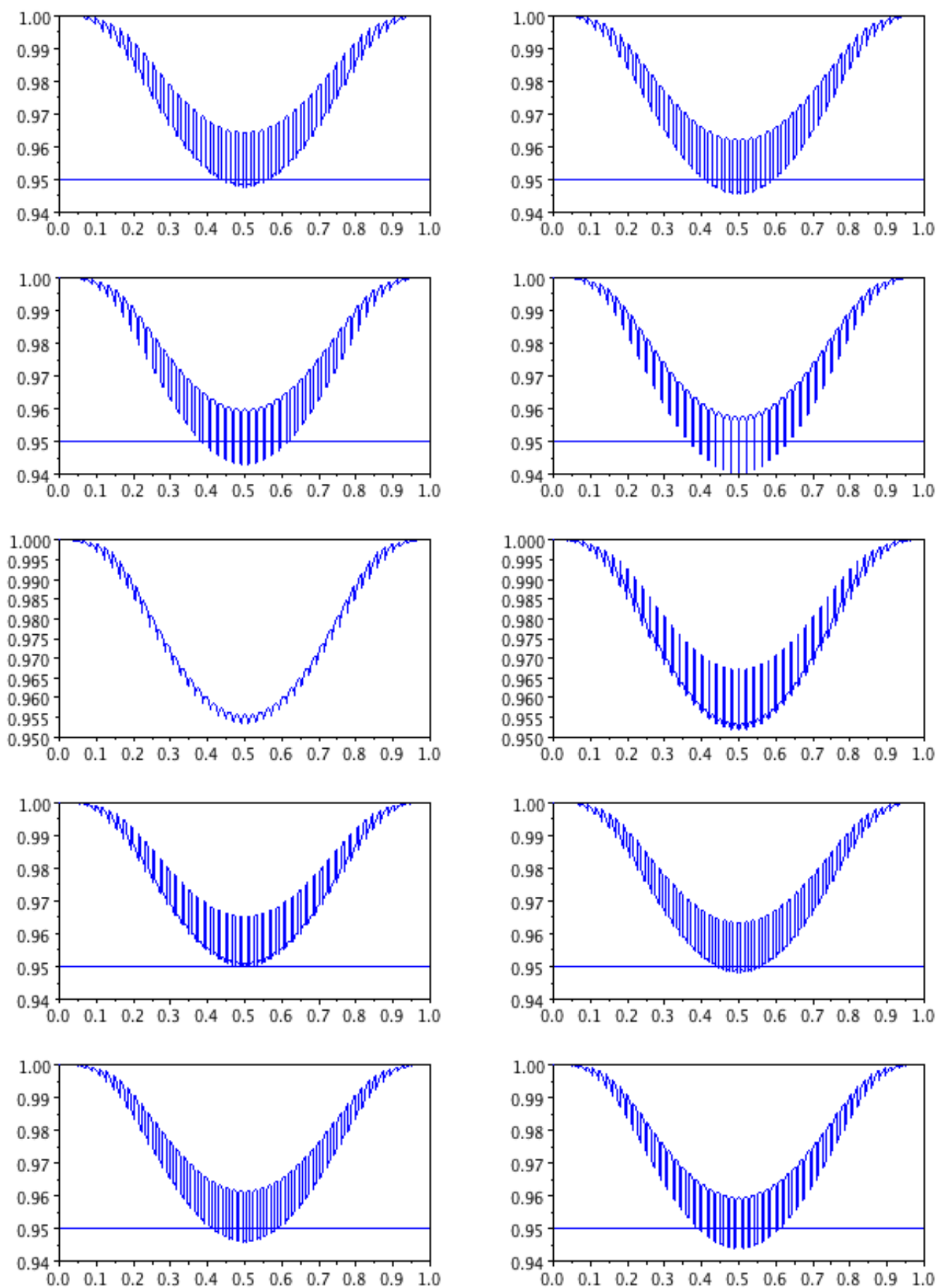


Figure 5 Graphes de $P(E)$ pour $n = 55, 56, \dots, 64$.

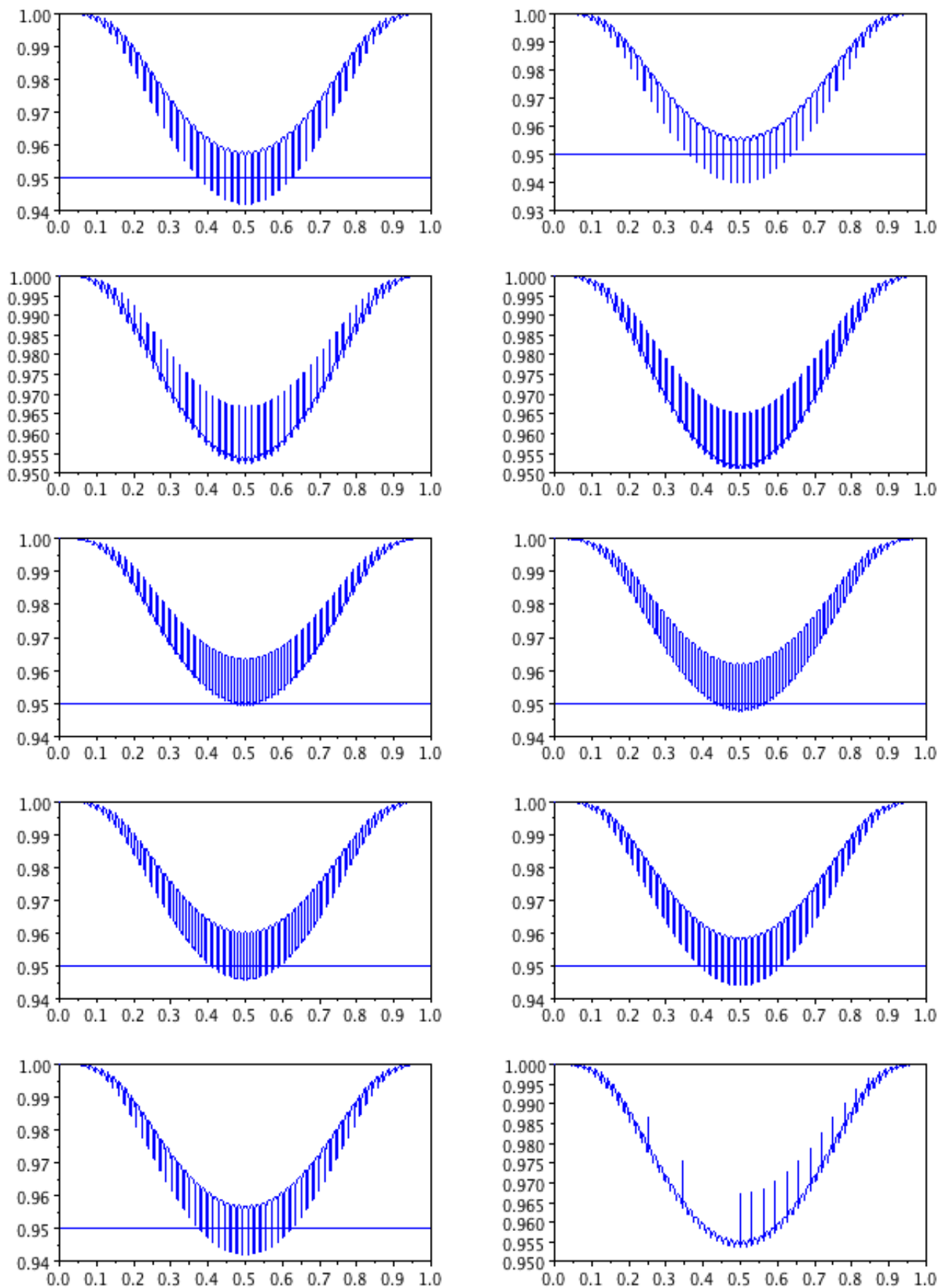


Figure 6 Graphes de $P(E)$ pour $n = 65, 66, \dots, 74$.

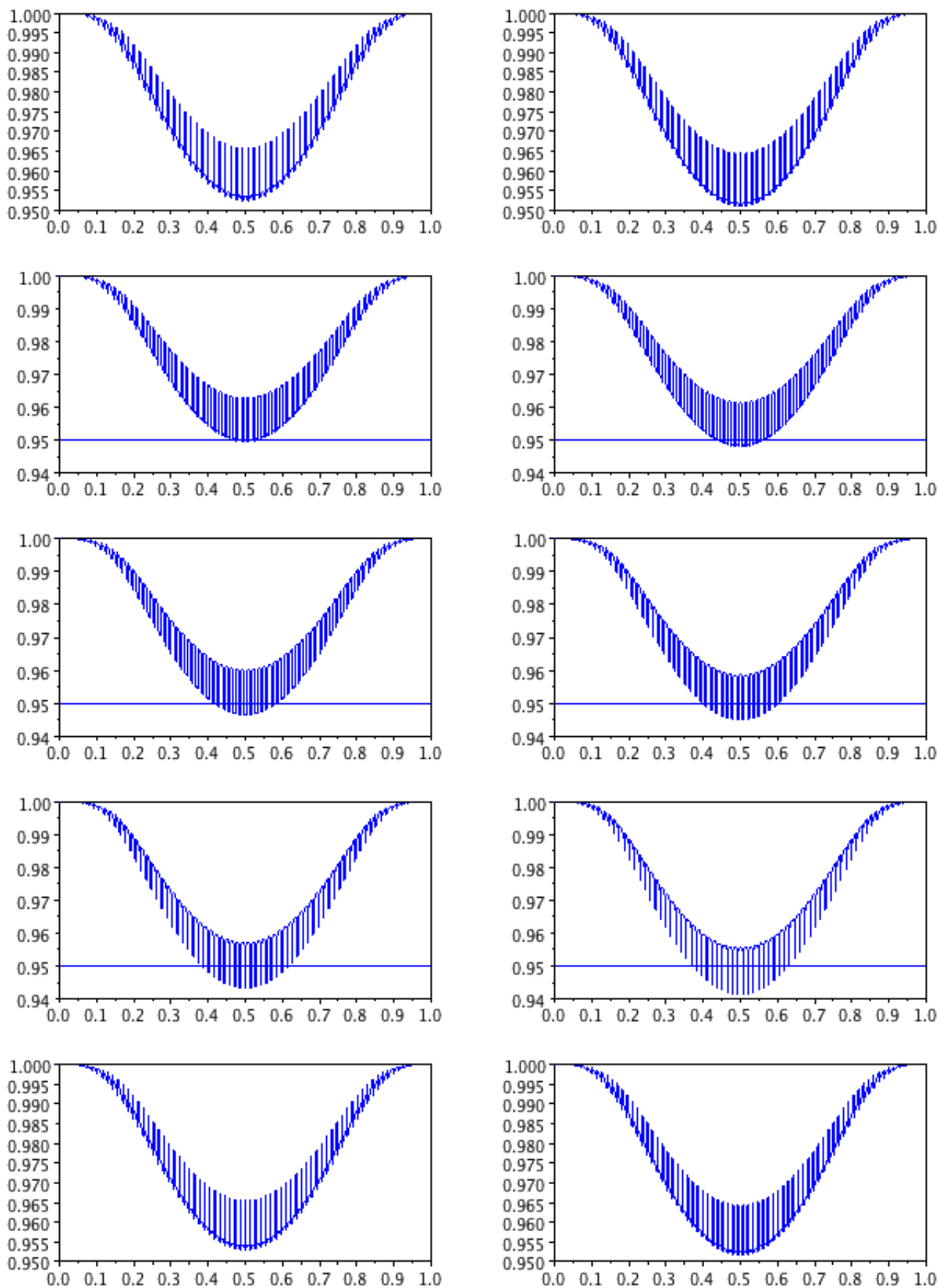


Figure 7 Graphes de $P(E)$ pour $n = 75, 76, \dots, 84$.

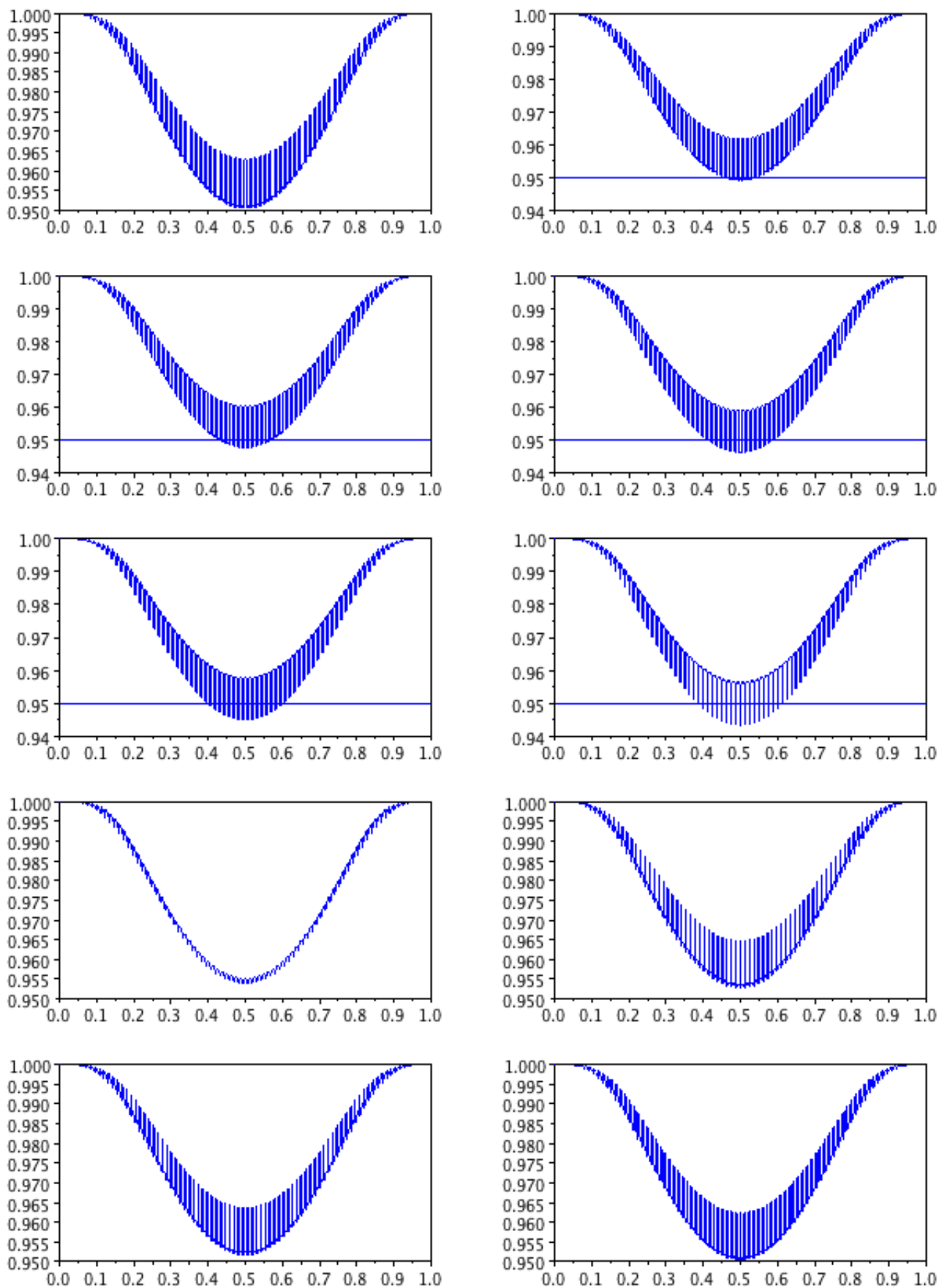


Figure 8 Graphes de $P(E)$ pour $n = 85, 86, \dots, 94$.

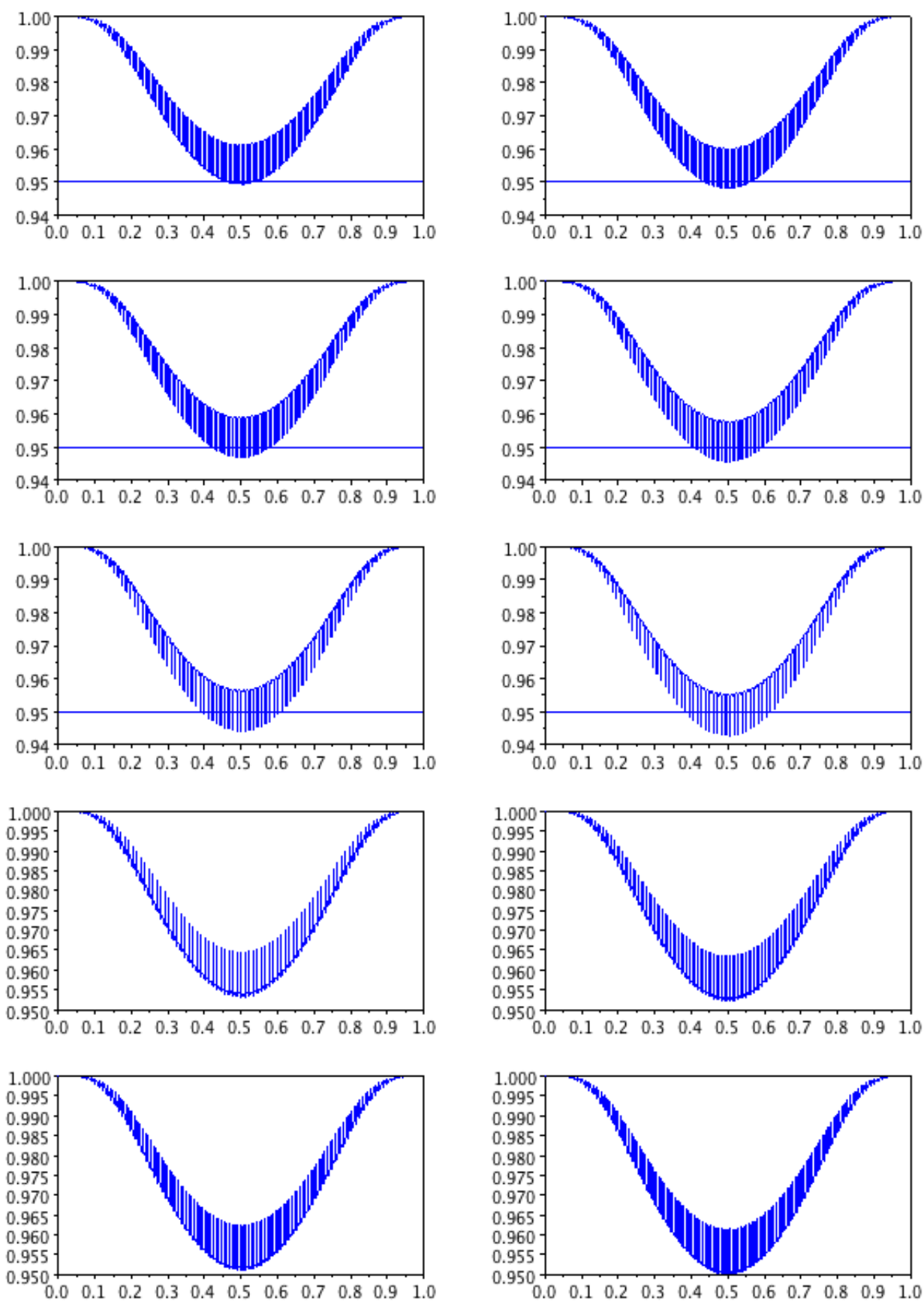


Figure 9 Graphes de $P(E)$ pour $n = 95, 96, \dots, 104$.

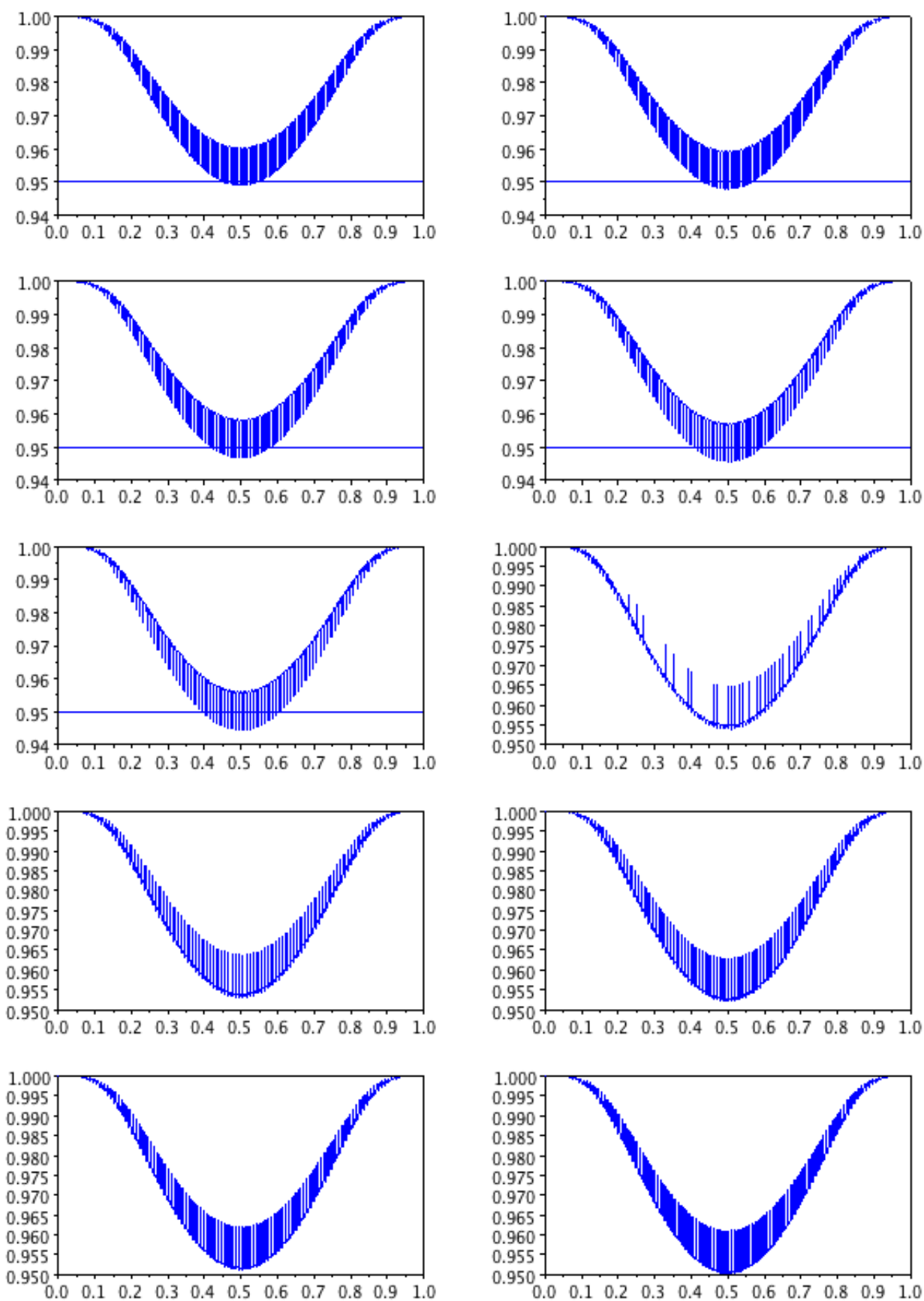
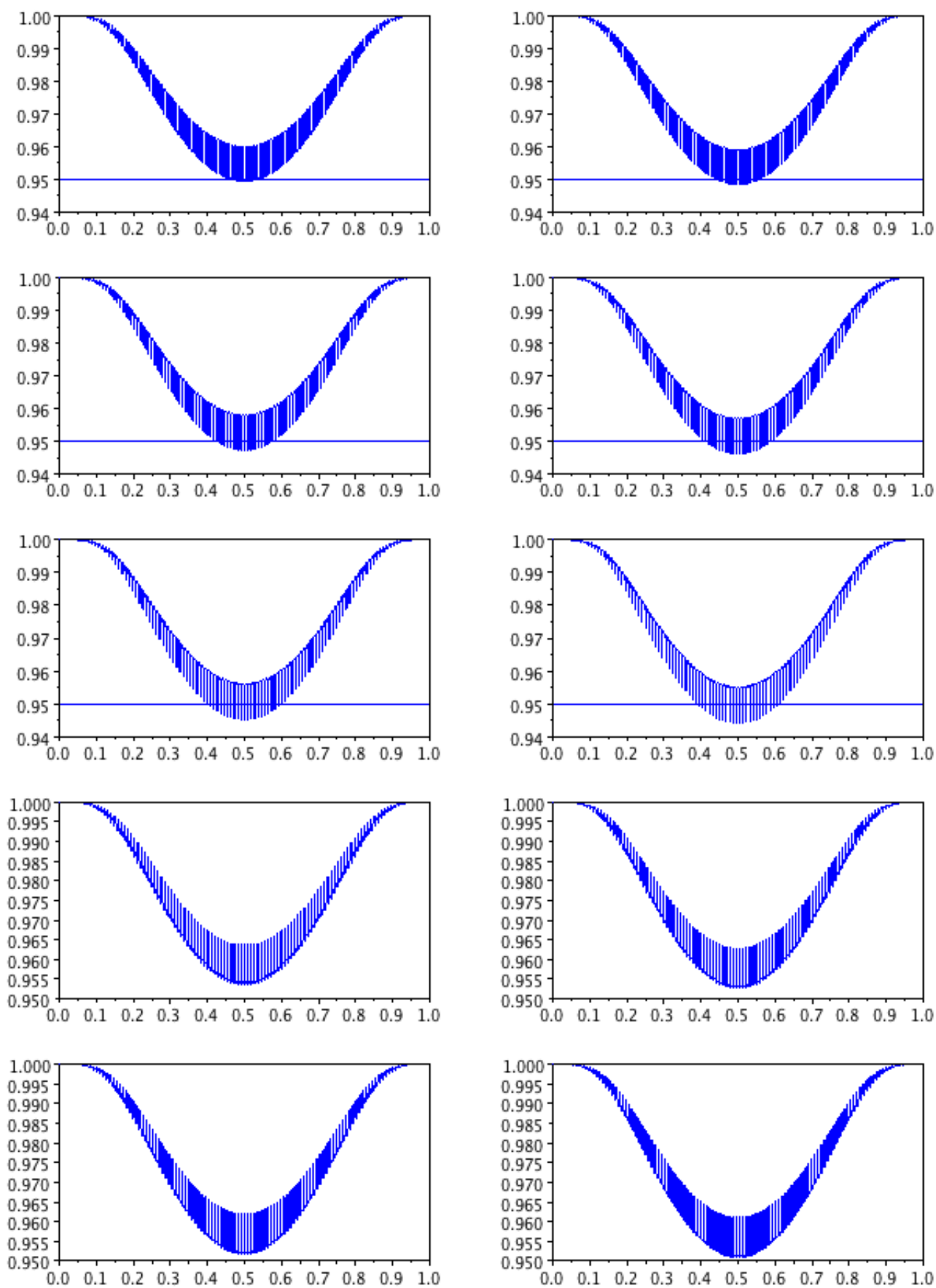


Figure 10 Graphes de $P(E)$ pour $n = 105, 106, \dots, 114$.



3 Évaluation de $P(E)$ par expérimentation et simulation

Il est suggéré dans [7], p. 15 d'évaluer la probabilité $P(E)$ par expérimentation et simulation. C'est un travail de routine. On procède comme suit :

Étape 1 : on répète n fois l'expérience aléatoire considérée⁶, on calcule la fréquence de réalisation de l'événement A et on regarde si cette fréquence appartient ou non à l'intervalle de fluctuation, c'est à dire si l'événement E est réalisé ou non.

Étapes suivantes : On répète $N - 1$ fois l'étape 1. On sait que si N est suffisamment grand, la fréquence de réalisation de E calculée au fur et à mesure se rapprochera de $P(E)$ et constituera une bonne approximation de cette probabilité si N est grand (loi des grands nombres).

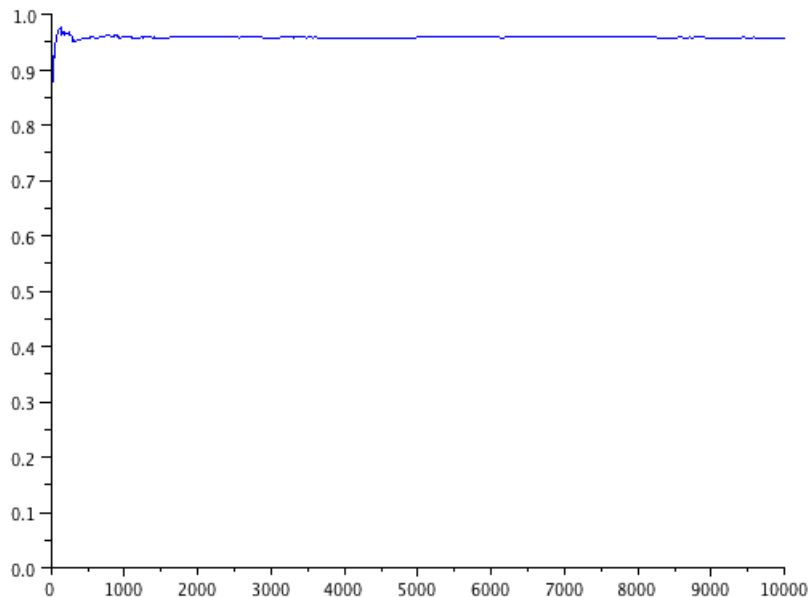
L'algorithme ci-dessous produit le graphe des fréquences successives de réalisation de l'événement E .

Listing 3 – Algorithme Simulation.sce

```
// Simulation
n=input('n=');
p=input('p=');
N=input('N=');// On réalise N fois l'expérience qui consiste à répéter n fois
// le jet d'une pièce qui a la probabilité p de tomber sur pile et à noter
// si l'événement E : "La fréquence de succès (pile) tombe dans
// l'intervalle de fluctuation".
NS=grand(1,N,"bin",n,p);//NS est la suite des nombres de succès obtenus.
Freq=[];
C=0;// Compteur des réalisations de E.
for i=1:N
    if (n*p-sqrt(n)<= NS(i))&(NS(i)<=n*p+sqrt(n)) then
        C=C+1;
    end;
    Freq=[Freq,C/i];
end;
scf
plot(Freq);
afficher('La valeur estimée de la probabilité de l''événement E est
//'+string(Freq(N)));
```

6. En fait, on simule une valeur d'une variable aléatoire suivant la loi binomiale $\mathcal{B}(n, p)$, comme on l'a expliqué ci-dessus.

Pour $n = 1000$, $p = 0.4$ et $N = 10000$, on obtient le graphe suivant :



La valeur estimée de $P(E)$ pour cette simulation (dernière fréquence de réalisation de E calculée) est 0.9574. La valeur exacte de $P(E)$ est 0.9580, en utilisant les instructions suivantes :

```
a=floor(10*sqrt(10)-400);
b=floor(10*sqrt(10)+400);
PE=cdfbin("PQ",b,1000,0.4,0.6)-cdfbin("PQ",a-1,1000,0.4,0.6)
```

4 Que peut-on démontrer sur la recette ?

Notons maintenant l'événement E , dorénavant appelé E_n , sous la forme :

$$(13) \quad E_n = \left\{ -\frac{1}{\sqrt{pq}} \leq S_n^* \leq \frac{1}{\sqrt{pq}} \right\} \quad \text{où} \quad S_n^* = \frac{S_n - np}{\sqrt{npq}}$$

Comme S_n suit la loi binomiale $\mathcal{B}(n, p)$, dont l'espérance est np et la variance est npq , S_n^* est la variable aléatoire centrée réduite associée à S_n . Son espérance est nulle et sa variance est égale à 1.

4.1 Ce que dit le programme

On sait que, quelle que soit la valeur du paramètre p ($0 < p < 1$), la loi de S_n^* converge vers la loi normale centrée réduite, quand $n \rightarrow +\infty$, d'après le théorème de De Moivre-Laplace⁷. Cela signifie, par définition de la convergence en loi, que si l'on appelle respectivement F_n^* et Φ les fonctions de répartition de S_n^* et de la loi normale centrée réduite, comme Φ est une fonction continue,

$$(14) \quad \forall x \in \mathbb{R}, \quad F_n^*(x) \xrightarrow[n \rightarrow +\infty]{} \Phi(x)$$

On en déduit que quels que soient les réels a et b tels que $a < b$,

$$(15) \quad P(a < S_n^* \leq b) = F_n^*(b) - F_n^*(a) \xrightarrow[n \rightarrow +\infty]{} \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$$

7. cf. [3], p. 147.

On démontre aussi que⁸

$$(16) \quad P(a \leq S_n^* \leq b) = F_n^*(b) - F_n^*(a - 0) \xrightarrow{n \rightarrow +\infty} \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$$

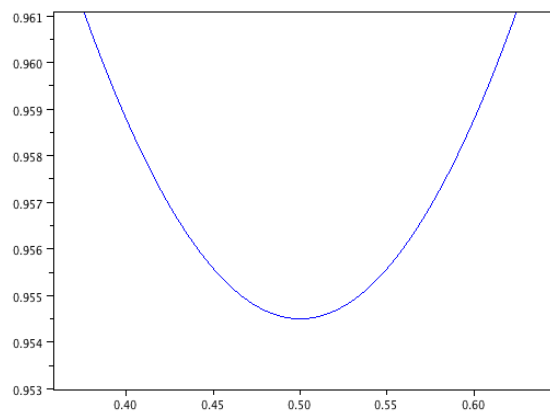
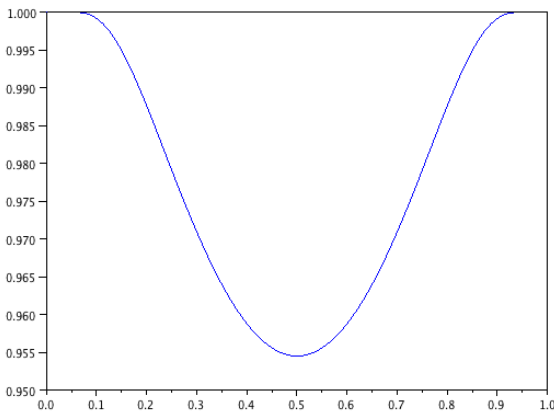
ce qui permet de conclure que

$$(17) \quad P(E_n) = F_n^*\left(\frac{1}{\sqrt{pq}}\right) - F_n^*\left(-\frac{1}{\sqrt{pq}} - 0\right) \xrightarrow{n \rightarrow +\infty} \Phi\left(\frac{1}{\sqrt{pq}}\right) - \Phi\left(-\frac{1}{\sqrt{pq}}\right) = 2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1$$

puisque l'on sait que la fonction Φ vérifie :

$$(18) \quad \forall x \in \mathbb{R}, \quad \Phi(x) + \Phi(-x) = 1$$

En considérant $p \mapsto 2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 = 2 \cdot \Phi\left(\frac{1}{\sqrt{p(1-p)}}\right) - 1$ comme une composée de fonctions monotones, on voit sans calcul que, quand p varie de 0 à 1 dans l'intervalle $]0, 1[$, $2 \cdot \Phi\left(\frac{1}{\sqrt{p(1-p)}}\right) - 1$ décroît de 1 à $2 \cdot \Phi(2) - 1$ entre 0 et 0.5 puis croît jusqu'à 1 (la droite $x = 0.5$ est axe de symétrie du graphe). Cela se voit mieux sur le graphe de droite dont le graphe de gauche est un simple grossissement.



Comme $\Phi(2) \approx 0.97724986$, $2 \cdot \Phi(2) - 1 \approx 0.95449973$ ⁹, on peut donc dire, d'après (17), que

$$(19) \quad \boxed{\text{Si } n \text{ est suffisamment grand, } P(E_n) \approx 2\Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 \geq 0.95}$$

C'est ici que commence le laisser-aller et que finissent les mathématiques. Dans la pratique, ceci devient :

$$(20) \quad \boxed{\text{Si } n \text{ est suffisamment grand, } P(E_n) = 2\Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 \geq 0.95}$$

c'est à dire que l'on oublie qu'il y a une erreur d'approximation. Or, il faut bien se pencher sur cette erreur et la contrôler si l'on veut préciser ce que signifie l'expression « Si n est suffisamment grand ».

4.2 Défaillance du contrôle de l'erreur

En fait, en choisissant $2 \cdot \Phi\left(\frac{1}{\sqrt{p(1-p)}}\right) - 1$ comme valeur approchée de $P(E_n)$, cf. (20), on fait une erreur qui dépend de n et de p , que l'on peut noter¹⁰

$$(21) \quad e(n, p) = \left| P(E_n) - \left(2\Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 \right) \right|$$

8. $F_n^*(a - 0)$ désigne la limite à gauche de la fonction F_n^* au point a .

9. Valeur approchée par défaut.

10. Il s'agit en fait de la valeur absolue de l'erreur.

Contrôler $e_{n,p}$ est difficile et n'a pas de solution qui conduise à la recette, sauf erreur.

Voie directe

On peut majorer $e(n,p)$ en étudiant les probabilités élémentaires de la loi $\mathcal{B}(n,p)$ qui figurent dans (11). C'est ainsi que De Moivre a découvert et démontré la première version du théorème-limite central en 1733. D'après un raffinement d'Uspensky (cf. [3], p. 177-178) ^{11, 12}

$$(22) \quad \text{si } npq \geq 25, \quad |P(E_n) - \left(2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1\right)| \leq \frac{0.588}{\sqrt{npq}}$$

ou, ce qui est équivalent,

$$(23) \quad \text{si } npq \geq 25, \quad \left(2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1\right) - \frac{0.588}{\sqrt{npq}} \leq P(E_n) \leq \left(2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1\right) + \frac{0.588}{\sqrt{npq}}$$

On en déduit que

Théorème 1 Si $npq \geq 25$ et $\left(2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1\right) - \frac{0.588}{\sqrt{npq}} \geq 0.95$, $P(E_n) \geq 0.95$. \square

On remarquera que la condition $npq \geq 25$ est différente de la condition de la recette. La minoration de $P(E_n)$ ci-dessus est sans intérêt si p est suffisamment petit (proche de 0) ou suffisamment grand (proche de 1) puisque

$$\left(2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1\right) - \frac{0.588}{\sqrt{npq}} \xrightarrow{p \rightarrow 0+} -\infty \quad \text{et} \quad \left(2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1\right) - \frac{0.588}{\sqrt{npq}} \xrightarrow{p \rightarrow 1-} -\infty$$

Par conséquent, imposons à p la condition $0.2 \leq p \leq 0.8$, ce qui implique $pq \geq 0.16$. On déduit alors que

$$2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 - \frac{0.588}{\sqrt{npq}} \geq 2 \cdot \Phi(2) - 1 - \frac{1.47}{\sqrt{n}} \geq 0.95449973 - \frac{1.47}{\sqrt{n}}$$

puis du théorème 1 que pour que $P(E_n) \geq 0.95$, il suffit que $0.95449973 - \frac{1.47}{\sqrt{n}} \geq 0.95$, ce qui donne :

Théorème 2 Si $0.2 \leq p \leq 0.8$ et si $n \geq 106723$, alors $P(E_n) \geq 0.95$. \square

Ce résultat, prouvé, est beaucoup moins bon que la recette ! Cette mauvaise performance n'est pas due au théorème d'Uspensky, qui est très fin, mais qui n'a pas été fait pour ça. Le problème est clairement de trouver une meilleure minoration de $P(E_n)$.

Avant d'abandonner cette voie qui paraît sans espoir alors qu'elle est directe, traçons les graphes des fonctions $p \mapsto P(E_n)$, de $p \mapsto 0.95$ et de la minoration de $P(E_n)$ donnée par (23) dans le cas $n = 250$, p prenant les valeurs 0.0001, 0.0002, 0.0003, ..., jusqu'à 0.9999. On voit bien que ce n'est pas $P(E_n)$ qui pose problème, mais bien la minoration utilisée.

11. **Théorème** : Si $npq \geq 25$, pour tous réels x_1 et x_2 tels que $x_1 < x_2$, $|P(x_1 \leq S_n^* \leq x_2) - (\Phi(x_2) - \Phi(x_1))| \leq \frac{0.588}{\sqrt{npq}}$.

On remarque que le majorant ne dépend que de n et de p puisque $q = 1 - p$. Autrement dit, cette majoration est uniforme en x_1 et x_2 . On remarque aussi que si p est proche de 0 ou de 1, il faudra que n soit grand pour que l'hypothèse $npq \geq 25$ soit satisfaite.

12. Le chapitre 7 de [3] : « Approximation gaussienne de la loi binomiale » est une très bonne référence sur le théorème de De Moivre-Laplace et sur la vitesse de convergence dans ce théorème.

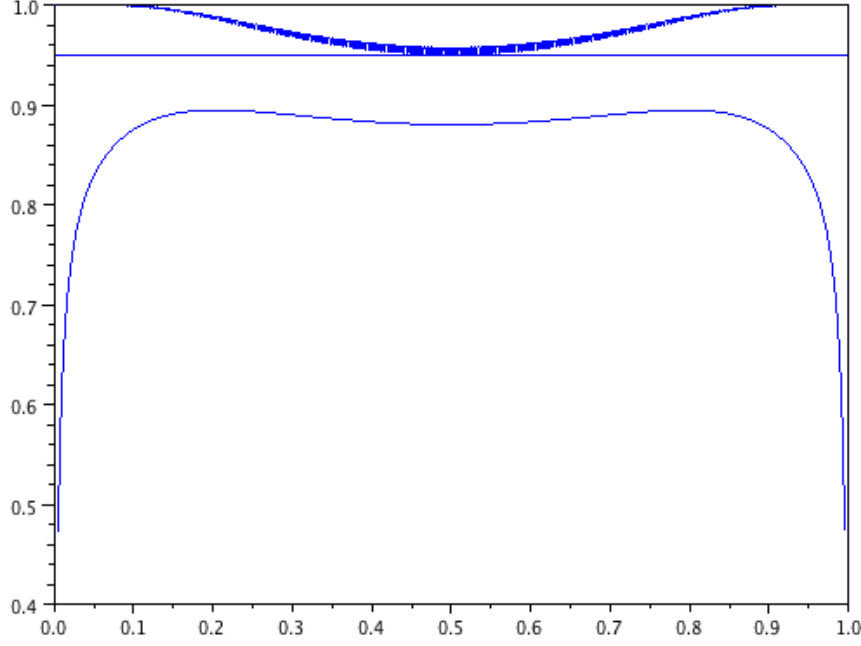


Figure 11

Voie du théorème de Berry-Esseen

Écrivons $e_{n,p}$ sous la forme :

$$(24) \quad e(n,p) = \left| \left(F_n^*\left(\frac{1}{\sqrt{pq}}\right) - \Phi\left(\frac{1}{\sqrt{pq}}\right) \right) - \left(F_n^*\left(-\frac{1}{\sqrt{pq}}\right) - \Phi\left(-\frac{1}{\sqrt{pq}}\right) \right) \right|$$

En introduisant la norme de la convergence uniforme $\| \cdot \|_\infty$ ¹³, on obtient à l'aide de l'inégalité triangulaire :

$$(25) \quad e(n,p) \leq 2 \| F_n^* - \Phi \|_\infty$$

ce qui donne, quand on majore $\| F_n^* - \Phi \|_\infty$ à l'aide du théorème de Berry-Esseen¹⁴ :

$$(27) \quad \forall n \in \mathbb{N}^* \text{ et } \forall p \in]0, 1[, \quad e(n,p) \leq 2 \cdot \frac{0.4784}{\sqrt{n}} \cdot \frac{p^2 + q^2}{\sqrt{pq}} \leq \frac{1.9136}{\sqrt{n}}$$

parce que $p^2 + q^2 \leq 1$ et $pq \geq \frac{1}{4}$. Pour que $P(E_n) \geq 0.95$, il suffit que $2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 - e_{n,p} \geq 0.95$, et pour cela, il suffit que $2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 - \frac{1.9136}{\sqrt{n}} \geq 0.95$, d'après (27), ce qui peut s'énoncer comme un théorème :

Théorème 3 $\forall n \in \mathbb{N}^* \text{ et } \forall p \in]0, 1[, \quad 2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 - \frac{1.9136}{\sqrt{n}} \geq 0.95 \implies P(E_n) \geq 0.95. \quad \square$

13. Pour toute fonction bornée f définie sur \mathbb{R} , $\| f \|_\infty = \sup_{x \in \mathbb{R}} |f(x)|$.

14. Voir [1], p. 542-546. Si $(X_n, n \geq 1)$ désigne une suite de variables aléatoires réelles indépendantes et équidistribuées dont le moment d'ordre 3 existe et si l'on pose $m = E(X_1)$, $\sigma = \sqrt{\text{var}(X_1)}$ et $\rho = \sqrt[3]{E(|X_1 - m|^3)}$, alors

$$(26) \quad \forall n \geq 1, \quad \sup_{x \in \mathbb{R}} \left| P\left(\frac{S_n - nm}{\sigma\sqrt{n}} \leq x\right) - \Phi(x) \right| \leq \frac{0.4784}{\sqrt{n}} \left(\frac{\rho}{\sigma}\right)^3$$

Cette majoration tient compte de la majoration de C_1 par 0.4784, voir [4].

Comme on sait que quel que soit p , $2 \cdot \Phi\left(\frac{1}{\sqrt{pq}}\right) - 1 \geq 0.95449973$, on en déduit que

Théorème 4 Si $n \geq 180854$, $P(E_n) \geq 0.95$ quelle que soit la valeur de p . \square

Ce résultat est plus conforme aux observations précédentes concernant le comportement de $P(E_n)$ quand p est voisin de 0 ou de 1. Il n'est ni meilleur ni moins bon que le théorème 2 (c'est à dire qu'aucun d'eux ne peut se démontrer à partir de l'autre). Bien sûr, il n'a pas d'intérêt pratique : qui s'astreindrait à n'utiliser que des échantillons de taille ≥ 180854 pour faire une estimation par intervalle de fluctuation ? Le théorème de Berry-Esseen est très général. Les théorèmes généraux sont souvent décevants quand on les applique à des cas particuliers. Mais l'approche très propre et débarrassée des conditions parasites sur p .

5 D'où vient l'intervalle de fluctuation ? Intervalles de confiance

Accepter un risque d'erreur de probabilité 5% (au plus) peut paraître déraisonnable. Peut-on réduire cette probabilité d'erreur à 0.01 ou même 0.001 ? Comment ?

α étant donné, vérifiant $0 < \alpha < 1$ (α sera un majorant de la probabilité d'erreur), on calcule le nombre β tel que

$$(28) \quad \Phi(\beta) - \Phi(-\beta) = 1 - \alpha \quad \text{ou} \quad \Phi(\beta) = 1 - \frac{\alpha}{2}$$

Ce β existe et est unique car Φ est une fonction continue strictement croissante de \mathbb{R} dans $]0, 1[$ qui a pour limites 0 et 1 au voisinage de $-\infty$ et de $+\infty$. C'est donc une bijection de \mathbb{R} dans $]0, 1[$. Si Φ^{-1} désigne sa fonction réciproque, $\beta = \Phi^{-1}(1 - \frac{\alpha}{2})$.

Considérons l'événement

$$(29) \quad (-\beta \leq S_n^* \leq \beta) = \left(p - \beta \sqrt{\frac{pq}{n}} \leq F_n \leq p + \beta \sqrt{\frac{pq}{n}} \right) = \left(p \in \left[F_n - \beta \sqrt{\frac{pq}{n}}, F_n + \beta \sqrt{\frac{pq}{n}} \right] \right)$$

L'intervalle de confiance dit *de niveau de confiance* α ¹⁵ est par définition

$$(30) \quad \left[F_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{pq}{n}}, F_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{pq}{n}} \right]$$

Il est utilisé comme dans le cas particulier ci-dessus : d'après le théorème de De Moivre-Laplace, quand n est grand (sic), on peut (sic) considérer que S_n^* suit la loi normale centrée réduite. Il en résulte que la probabilité que p appartienne à l'intervalle de confiance est approximativement

$$(31) \quad P(-\beta \leq S_n^* \leq \beta) \approx 2 \cdot \Phi(\beta) - 1 = 2 \cdot \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - 1 = 2 \cdot \left(1 - \frac{\alpha}{2}\right) - 1 = 1 - \alpha$$

Si $\alpha = 0.05$, $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \Phi^{-1}(0.975) \approx 1.96$. Il en résulte que

$$\left[F_n - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{pq}{n}}, F_n + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \cdot \sqrt{\frac{pq}{n}} \right] \subseteq \left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}} \right] = E_n$$

ce qui ramène à l'intervalle de fluctuation. Bien sûr, les problèmes évoqués pour l'intervalle de fluctuation sont analogues pour les intervalles de confiance.

15. Quand $\beta = \frac{1}{\sqrt{pq}}$, on obtient l'intervalle de fluctuation du programme.

Références

- [1] W. FELLER. *An Introduction to Probability Theory and its Applications*, Vol. 2, Wiley (Second Edition, 1970).
- [2] G. SAPORTA. *Probabilités, Analyse des données et statistique*, Éditions Technip, Paris (1990).
- [3] C. SUQUET. *Introduction au Calcul des Probabilités (Probabilités à Bac+2 et plus si affinités ...)*
<http://math.univ-lille1.fr/~suquet/ens/ICP/Cmd060902.pdf>
- [4] WIKIPEDIA. [en] *Berry-Esseen Theorem*.
http://en.wikipedia.org/wiki/Berry-Esseen_theorem
- [5] *Modifications des programmes de mathématiques de Première S et de Terminale S*, BO n° 18 du 6 mai 2010
http://media.education.gouv.fr/file/18/18/7/modifications_programmes_mathematiques_144187.pdf
- [6] *Programme de mathématiques pour la classe de Seconde*, Juin 2009
http://media.education.gouv.fr/file/30/52/3/programme_mathematiques_seconde_65523.pdf
- [7] *Ressources pour la classe de Seconde - Probabilités et Statistiques* - Juin 2009
http://media.eduscol.education.fr/file/Programmes/17/9/Doc_ressource_proba-stats_109179.pdf

