Les probabilités : Comment allier intuition et raisonnement ? Que peut apporter l'outil informatique ?

V. Henry et G. Haesbroeck

Département de Mathématique – Université de Liège

Février 2018

Outline

- Somme de variables aléatoires
  - En pratique
  - En théorie
- Chance de gagner à une tombola
- Vraisemblance d'une valeur observée par rapport à un modèle
  - En pratique
  - En théorie

LIÈGE université

Jet de dés

Exercice 1

On jette deux dés parfaitement équilibrés. Quelle est la probabilité que la somme soit égale à 6 ?

Jet de dés

Exercice 1

On jette deux dés parfaitement équilibrés. Quelle est la probabilité que la somme soit égale à 6 ?

Solution : $\frac{5}{36}$

Jet d'un dé

Résultat suit une loi uniforme discrète de paramètre $n = 6$,

Jet d'un dé

Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

Jet d'un dé

Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

x_k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Jet d'un dé

Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

x_k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

On a $E[X] = \sum_{k=1}^{n} p_k x_k = \frac{1}{n} \sum_{k=1}^{n} x_k \ (= 3,5)$.

Jet d'un dé

Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

x_k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

On a $E[X] = \sum_{k=1}^{n} p_k x_k = \frac{1}{n} \sum_{k=1}^{n} x_k \ (= 3,5)$.

On a $V[X] = \frac{1}{n} \sum_{k=1}^{n} x_k^2 - (E[X])^2$.

Jet d'un dé

Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

x_k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

On a $E[X] = \sum_{k=1}^{n} p_k x_k = \frac{1}{n} \sum_{k=1}^{n} x_k \ (= 3,5)$.

On a $V[X] = \frac{1}{n} \sum_{k=1}^{n} x_k^2 - (E[X])^2$.

Si $x_k = k$, avec $k \in \{1, \ldots, n\}$,

Jet d'un dé

Résultat suit une loi uniforme discrète de paramètre $n = 6$, soit X la variable « Résultat du dé », on écrit $X \sim U(6)$.

x_k	1	2	3	4	5	6
p_k	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

On a $E[X] = \sum_{k=1}^{n} p_k x_k = \frac{1}{n} \sum_{k=1}^{n} x_k \ (= 3,5)$.

On a $V[X] = \frac{1}{n} \sum_{k=1}^{n} x_k^2 - (E[X])^2$.

Si $x_k = k$, avec $k \in \{1, \ldots, n\}$, on a

$$V[X] = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n(n+1)}{2n}\right)^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12} \left(= \frac{35}{12}\right)$$

Jet de deux dés et somme obtenue

Jet de deux dés et somme obtenue

$Y =$ somme de deux variables de même loi (thm central limite)

Jet de deux dés et somme obtenue

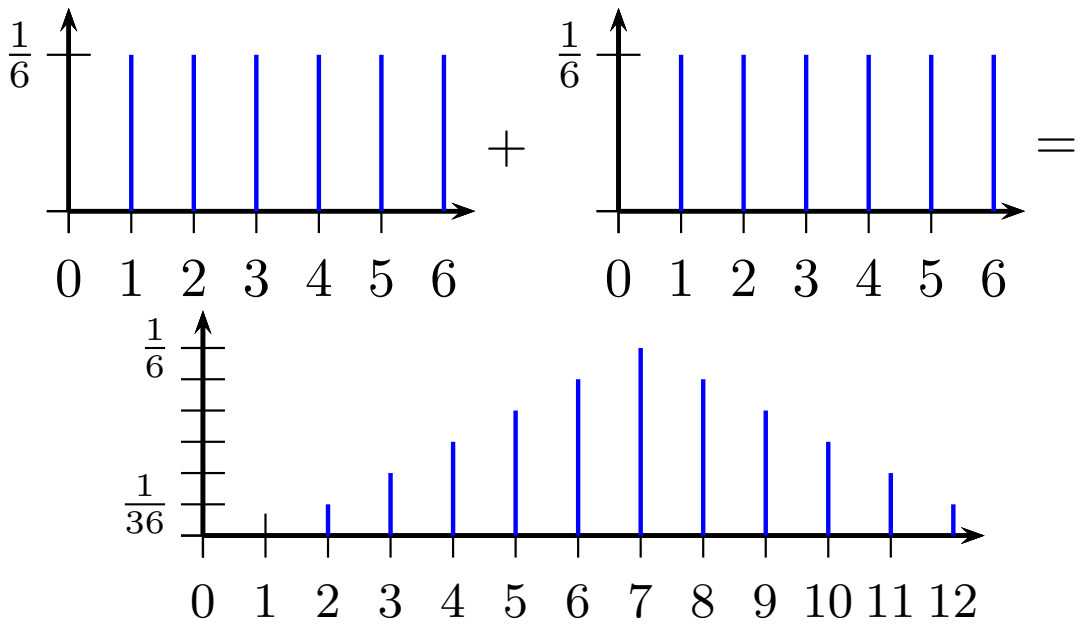
$Y =$ somme de deux variables de même loi (thm central limite)

k	2	3	4	5	6	7	8	9	10	11	12
p_k	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Jet de deux dés et somme obtenue

$Y =$ somme de deux variables de même loi (thm central limite)

k	2	3	4	5	6	7	8	9	10	11	12
p_k	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Somme de variables aléatoires iid et le “TCL”

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées (iid).

Somme de variables aléatoires iid et le "TCL"

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées (iid).

Que sait-on à propos de la distribution de $Y = X_1 + \ldots + X_n$?

Somme de variables aléatoires iid et le "TCL"

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées (iid).

Que sait-on à propos de la distribution de $Y = X_1 + \ldots + X_n$?

Si $E[X_i] = \mu$ et $V[X_i] = \sigma^2$ pour tout i, on a

$$E[Y] = E[X_1 + \ldots + X_n] = E[X_1] + \ldots + E[X_n] = n\mu$$

et

$$V[Y] = V[X_1 + \ldots + X_n] = n\sigma^2 \text{ (vu l'indépendance)}$$

Et la distribution ?

Considérons le cas uniforme discret sur $\{1, \ldots, k\}$.

Si X_1 et X_2 sont iid selon cette distribution, alors leur somme Y prend des valeurs entre 2 et $2k$. De plus, pour tout $2 \leq m \leq 2k$, on a

$$\boldsymbol{P}(Y = m) \quad = \quad \sum_{i=1}^{k} \boldsymbol{P}(Y = m, X_1 = i)$$

Et la distribution ?

Considérons le cas uniforme discret sur $\{1, \ldots, k\}$.

Si X_1 et X_2 sont iid selon cette distribution, alors leur somme Y prend des valeurs entre 2 et $2k$. De plus, pour tout $2 \leq m \leq 2k$, on a

$$\boldsymbol{P}(Y = m) \quad = \quad \sum_{i=1}^{k} \boldsymbol{P}(X_1 + X_2 = m, X_1 = i)$$

Et la distribution ?

Considérons le cas uniforme discret sur $\{1, \ldots, k\}$.

Si X_1 et X_2 sont iid selon cette distribution, alors leur somme Y prend des valeurs entre 2 et $2k$. De plus, pour tout $2 \leq m \leq 2k$, on a

$$\begin{aligned} \boldsymbol{P}(Y = m) &= \sum_{i=1}^{k} \boldsymbol{P}(X_1 + X_2 = m, X_1 = i) \\ &= \sum_{i=1}^{k} \boldsymbol{P}(X_2 = m - i, X_1 = i) \end{aligned}$$

Et la distribution ?

Considérons le cas uniforme discret sur $\{1, \ldots, k\}$.

Si X_1 et X_2 sont iid selon cette distribution, alors leur somme Y prend des valeurs entre 2 et $2k$. De plus, pour tout $2 \leq m \leq 2k$, on a

$$\begin{aligned} \boldsymbol{P}(Y = m) &= \sum_{i=1}^{k} \boldsymbol{P}(X_1 + X_2 = m, X_1 = i) \\ &= \sum_{i=1}^{k} \boldsymbol{P}(X_2 = m - i, X_1 = i) \\ &= \sum_{i=1}^{k} \boldsymbol{P}(X_2 = m - i)\boldsymbol{P}(X_1 = i) \end{aligned}$$

Et la distribution ?

Considérons le cas uniforme discret sur $\{1, \ldots, k\}$.

Si X_1 et X_2 sont iid selon cette distribution, alors leur somme Y prend des valeurs entre 2 et $2k$. De plus, pour tout $2 \leq m \leq 2k$, on a

$$\begin{aligned} \boldsymbol{P}(Y = m) &= \sum_{i=1}^{k} \boldsymbol{P}(X_1 + X_2 = m, X_1 = i) \\ &= \sum_{i=1}^{k} \boldsymbol{P}(X_2 = m - i, X_1 = i) \\ &= \sum_{i=1}^{k} \boldsymbol{P}(X_2 = m - i)\boldsymbol{P}(X_1 = i) \end{aligned}$$

Et la distribution ?

Considérons le cas uniforme discret sur $\{1, \ldots, k\}$.

Si X_1 et X_2 sont iid selon cette distribution, alors leur somme Y prend des valeurs entre 2 et $2k$. De plus, pour tout $2 \leq m \leq 2k$, on a

$$\begin{aligned} \boldsymbol{P}(Y = m) &= \sum_{i=1}^{k} \boldsymbol{P}(X_1 + X_2 = m, X_1 = i) \\ &= \sum_{i=1}^{k} \boldsymbol{P}(X_2 = m - i, X_1 = i) \\ &= \sum_{i=1}^{k} \boldsymbol{P}(X_2 = m - i)\boldsymbol{P}(X_1 = i) \end{aligned}$$

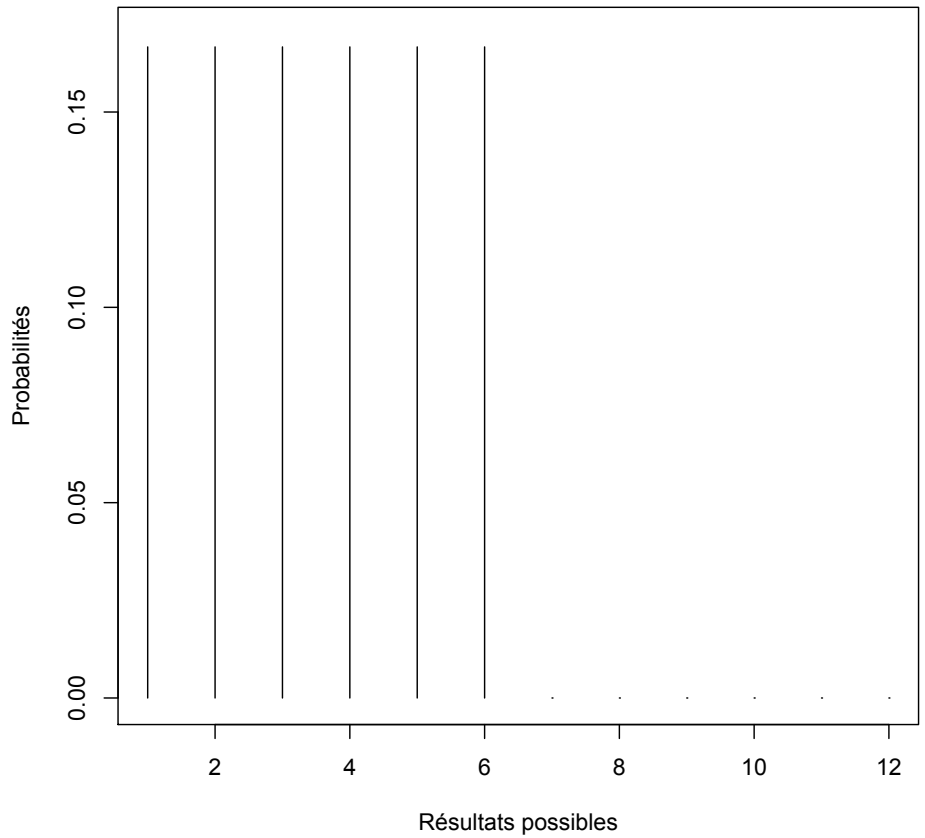
Il est possible de calculer facilement la distribution de Y, en passant en revue toutes les valeurs possibles pour m.

$\sum$ de 2 dés : $k = 6$; $P(X_j = i) = \frac{1}{6}$ $(1 \leq i \leq 6; j = 1, 2)$

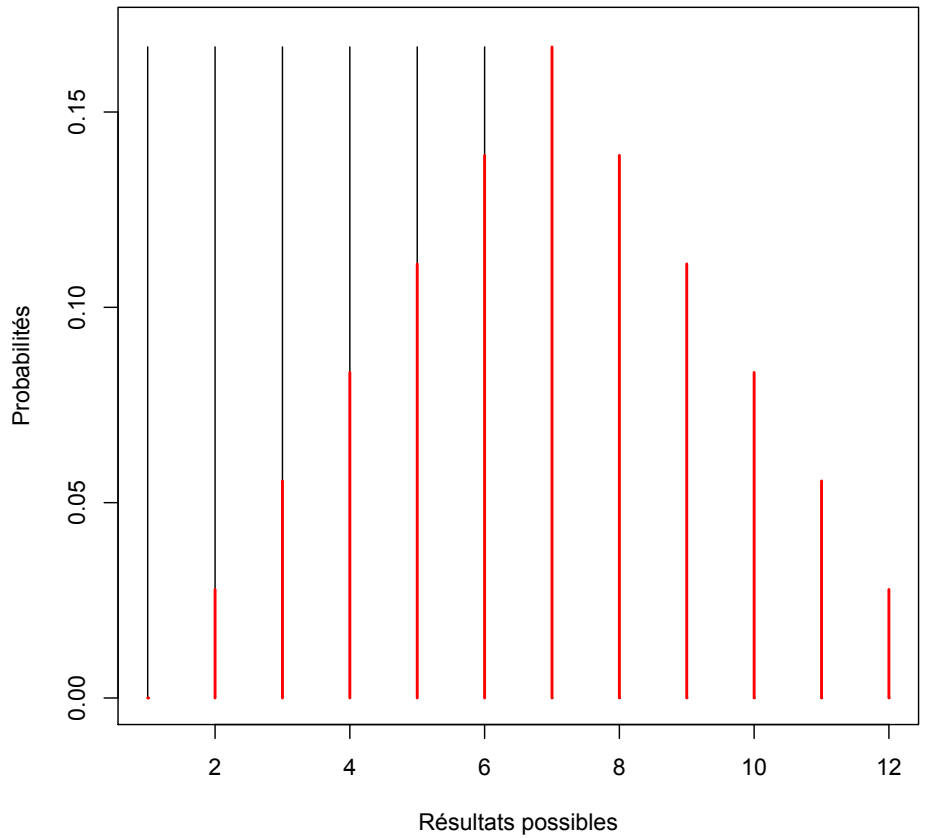
$$P(Y = m) = \sum_{i=1}^{k} P(X_2 = m - i) P(X_1 = i)$$

m	i	$m - i$	Probabilité
2	1	1	1/36
3	1	2	
	2	1	2/36
4	1	3	
	2	2	
	3	1	3/36
⋮			
10	4	6	
	5	5	
	6	4	3/36
⋮			

Graphiquement $n = 1$



Graphiquement $n = 2$



Et pour tout n ?

Soit $S_n = X_1 + \ldots + X_n$. Ses valeurs sont comprises entre n et $6n$.

$$P(S_n = m) = P(S_{n-1} + X_n = m)$$

où S_{n-1} et X_n sont indépendantes.

Et pour tout n ?

Soit $S_n = X_1 + \ldots + X_n$. Ses valeurs sont comprises entre n et $6n$.

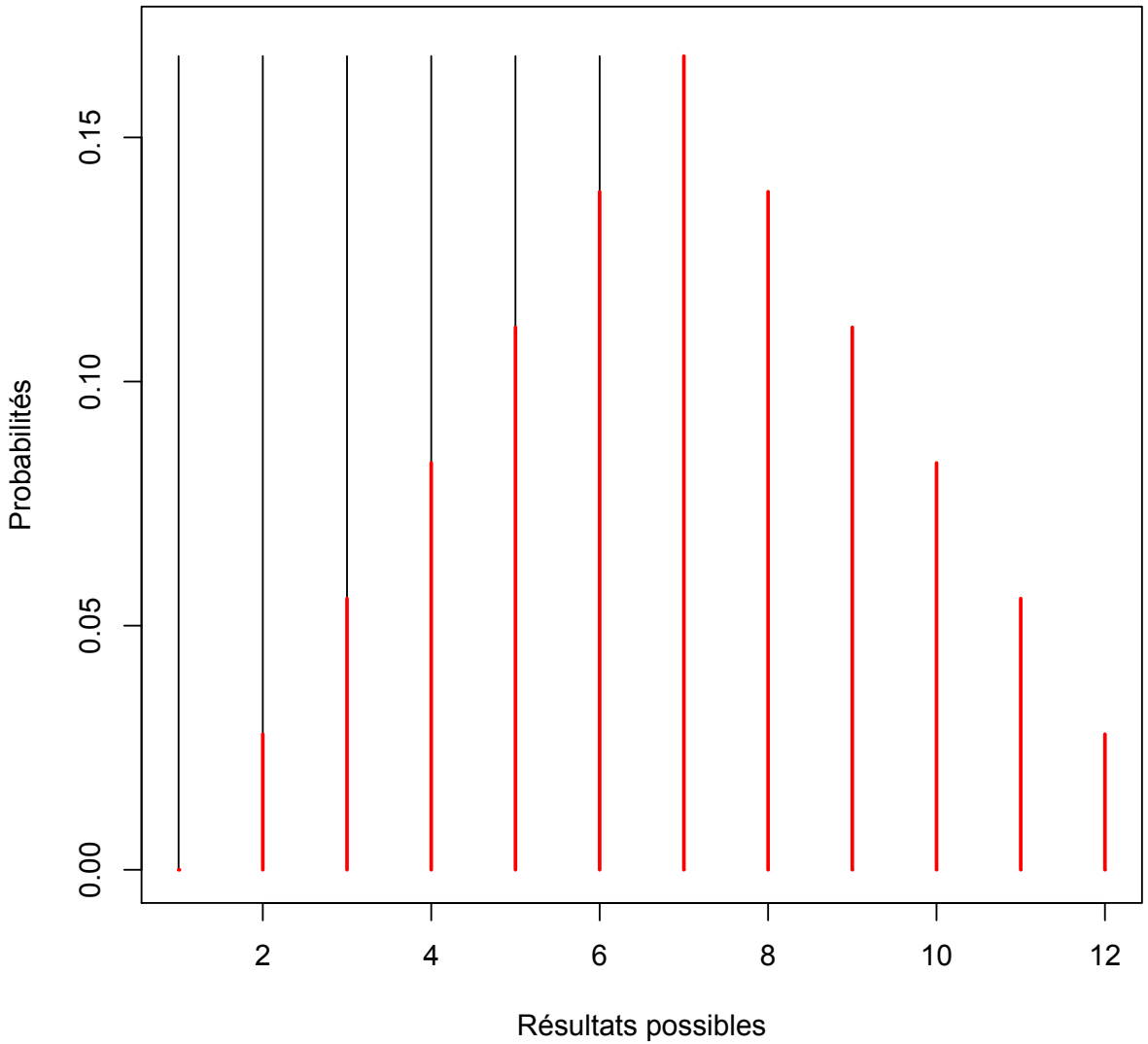
$$P(S_n = m) = P(S_{n-1} + X_n = m)$$

où S_{n-1} et X_n sont indépendantes.

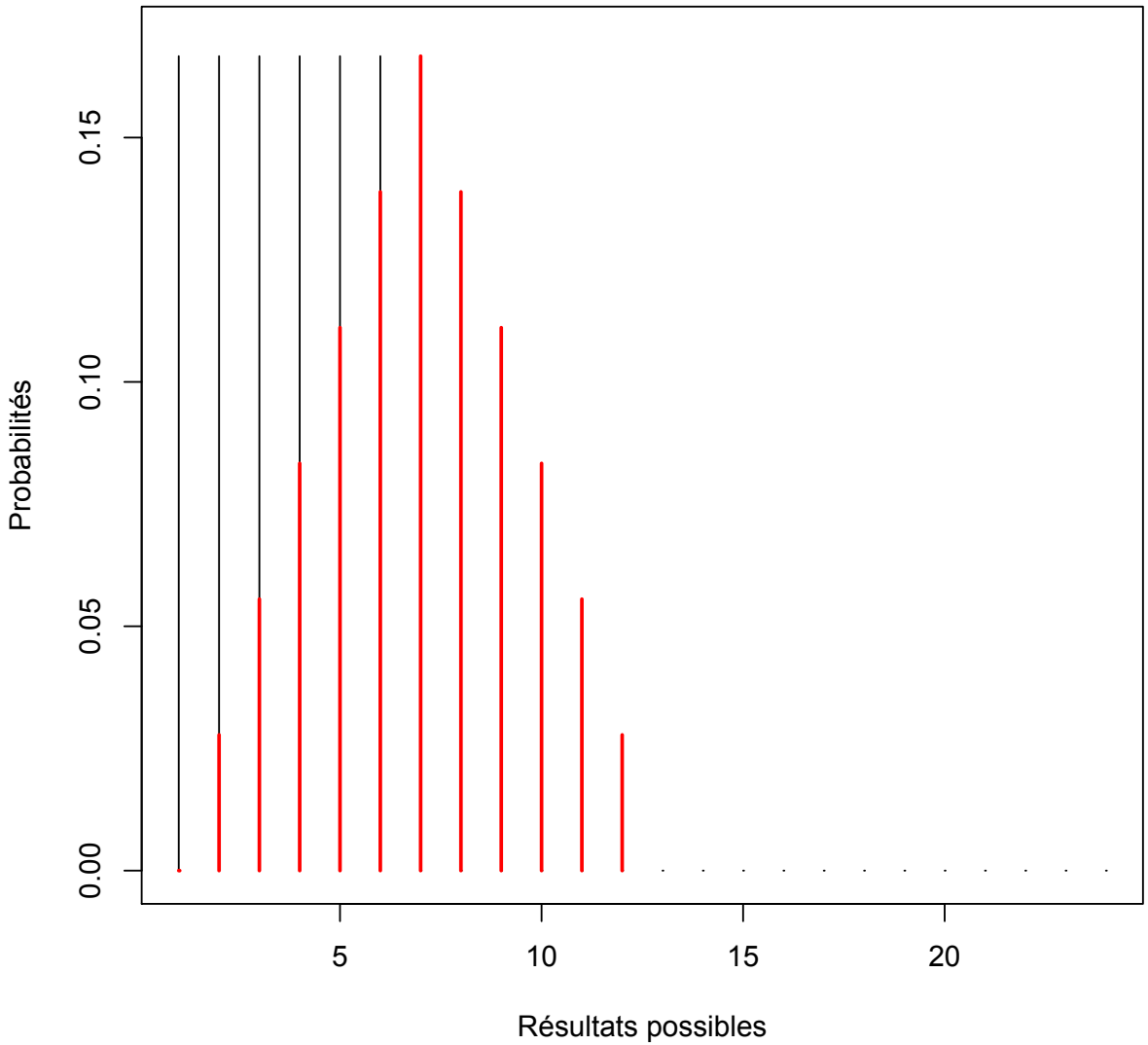
D'où

$$\begin{aligned} P(S_n = m) &= \sum_{i=1}^{k} P(S_n = m, X_n = i) \\ &= \sum_{i=1}^{k} P(S_{n-1} = m - i, X_n = i) \\ &= \sum_{i=1}^{k} P(S_{n-1} = m - i) P(X_n = i) \end{aligned}$$

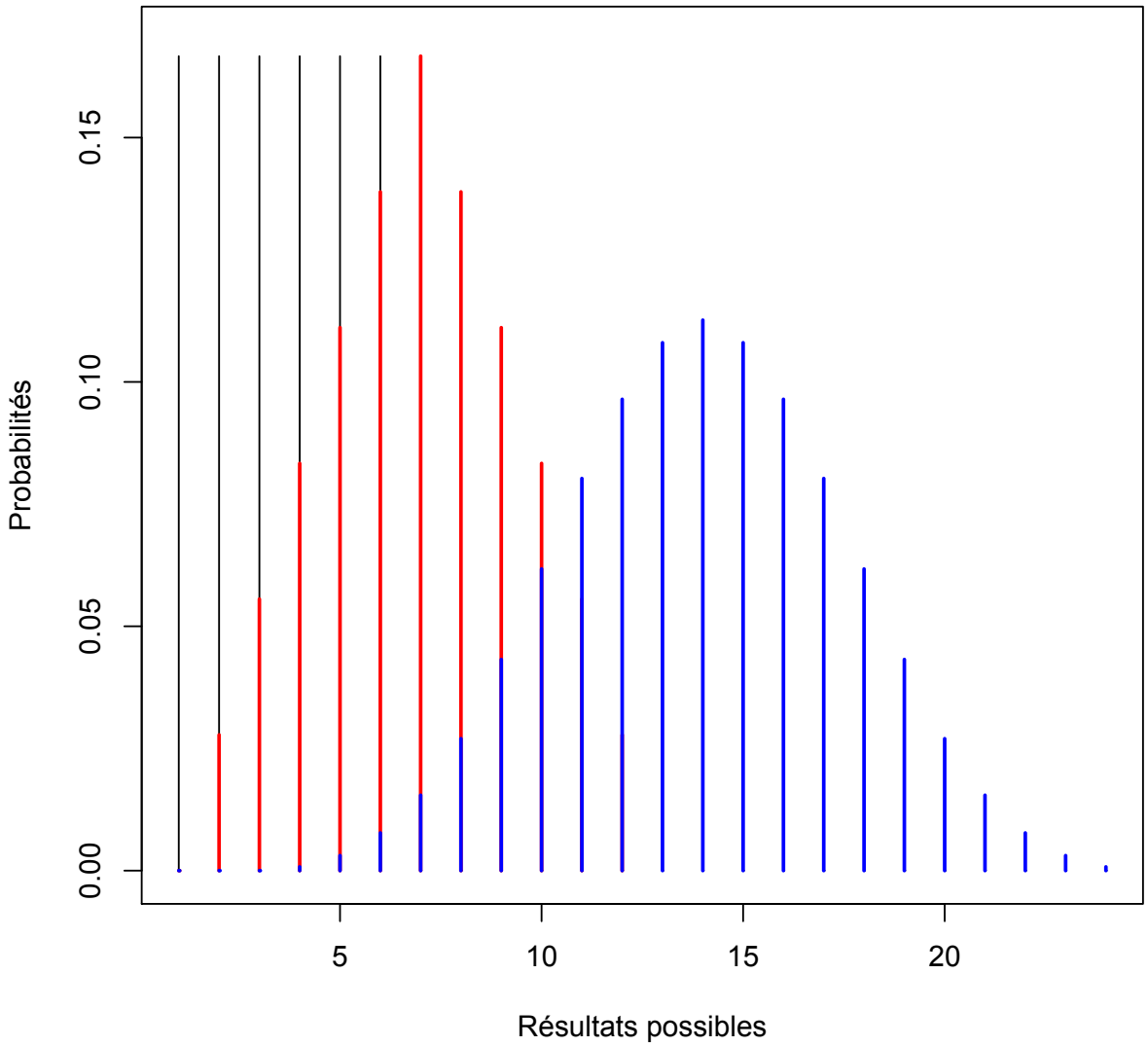
Graphiquement : $n = 1$ et $n = 2$



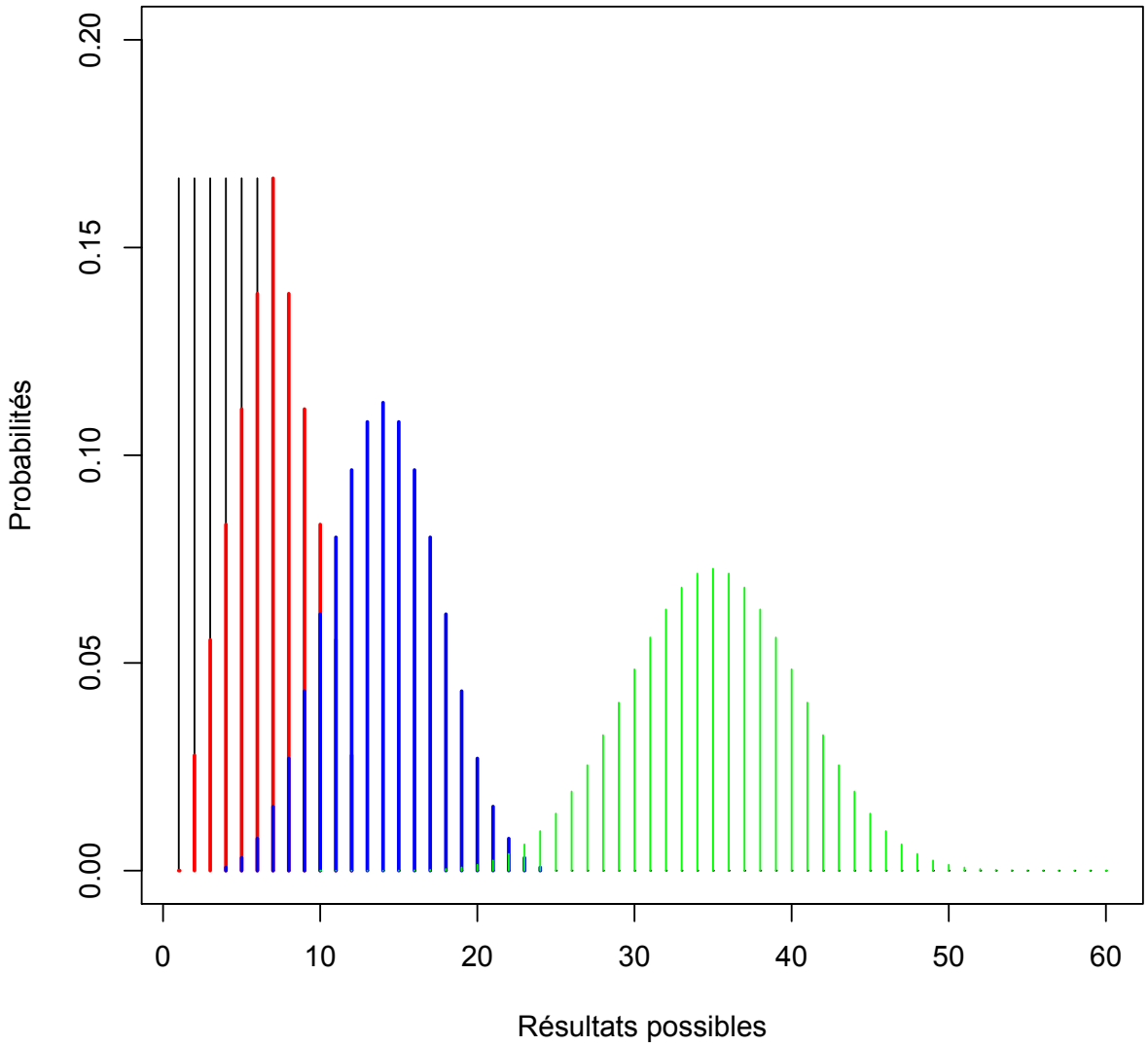
Graphiquement : $n = 1$ et $n = 2$



Graphiquement : $n = 1$ et $n = 2$; $n = 4$



Graphiquement : $n = 1$ et $n = 2$; $n = 4$ et $n = 10$



Dans le contexte continu

Soient X_1 et X_2 deux variables aléatoires continue iid de densité f_X. L'égalité

$$P(Y = m) = \sum_{i=1}^{k} P(X_2 = m - i)P(X_1 = i)$$

se transforme dans le contexte continu en

$$f_Y(m) = \int_0^m f_X(m - i)f_X(i)di$$

Dans le contexte continu

Soient X_1 et X_2 deux variables aléatoires continue iid de densité f_X. L'égalité

$$P(Y = m) = \sum_{i=1}^{k} P(X_2 = m - i)P(X_1 = i)$$

se transforme dans le contexte continu en

$$f_Y(m) = \int_0^m f_X(m - i)f_X(i)di = (f_X \star f_X)(m).$$

Dans le contexte continu

Soient X_1 et X_2 deux variables aléatoires continue iid de densité f_X. L'égalité

$$P(Y = m) = \sum_{i=1}^{k} P(X_2 = m - i)P(X_1 = i)$$

se transforme dans le contexte continu en

$$f_Y(m) = \int_0^m f_X(m - i)f_X(i)di = (f_X \star f_X)(m).$$

Et, de proche en proche,

$$f_{S_n}(m) = (f_{S_{n-1}} \star f_X)(m)$$

Illustration à partir de la loi uniforme $n = 1$

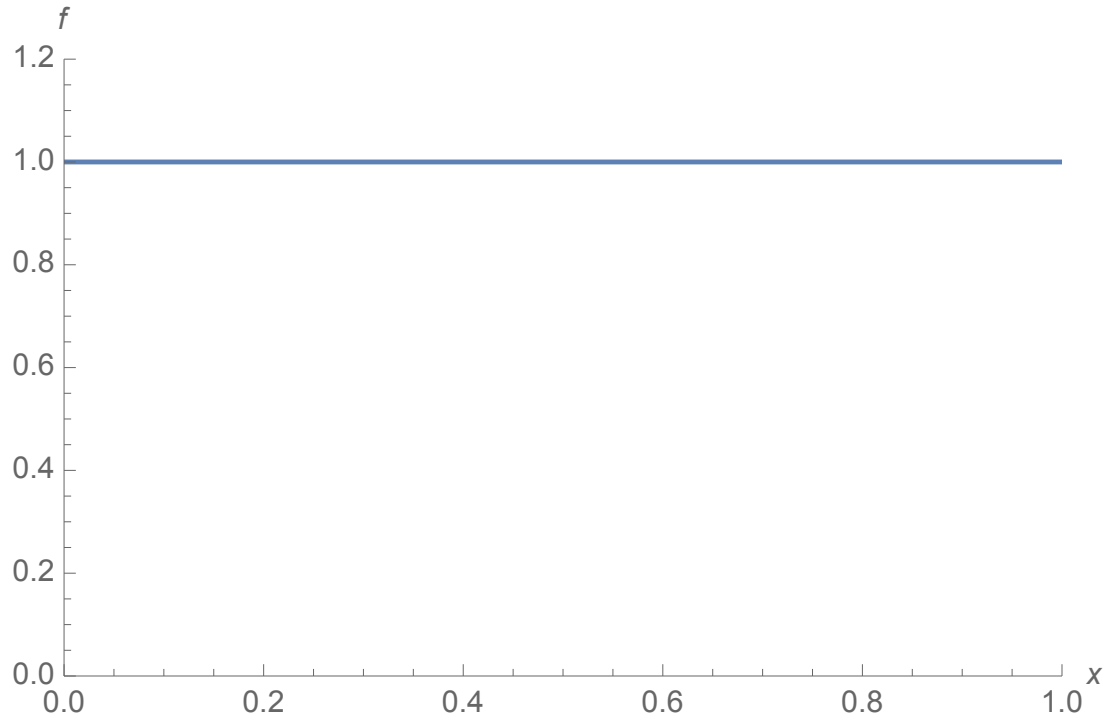


Illustration à partir de la loi uniforme $n = 2$

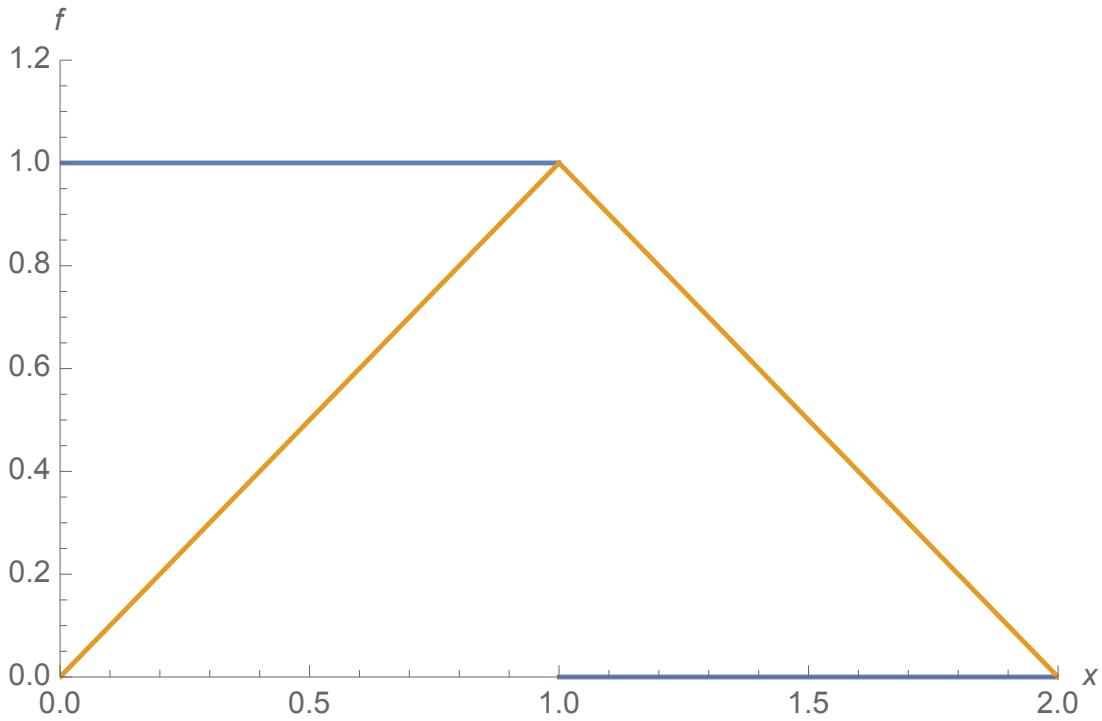


Illustration à partir de la loi uniforme $n = 4$

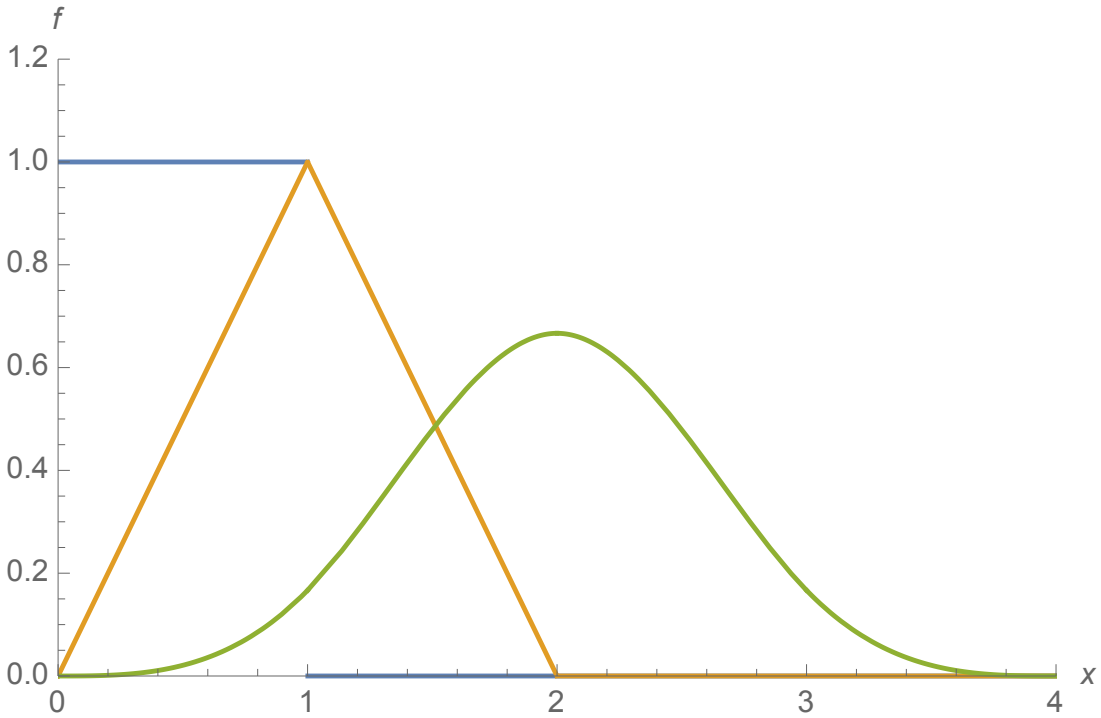
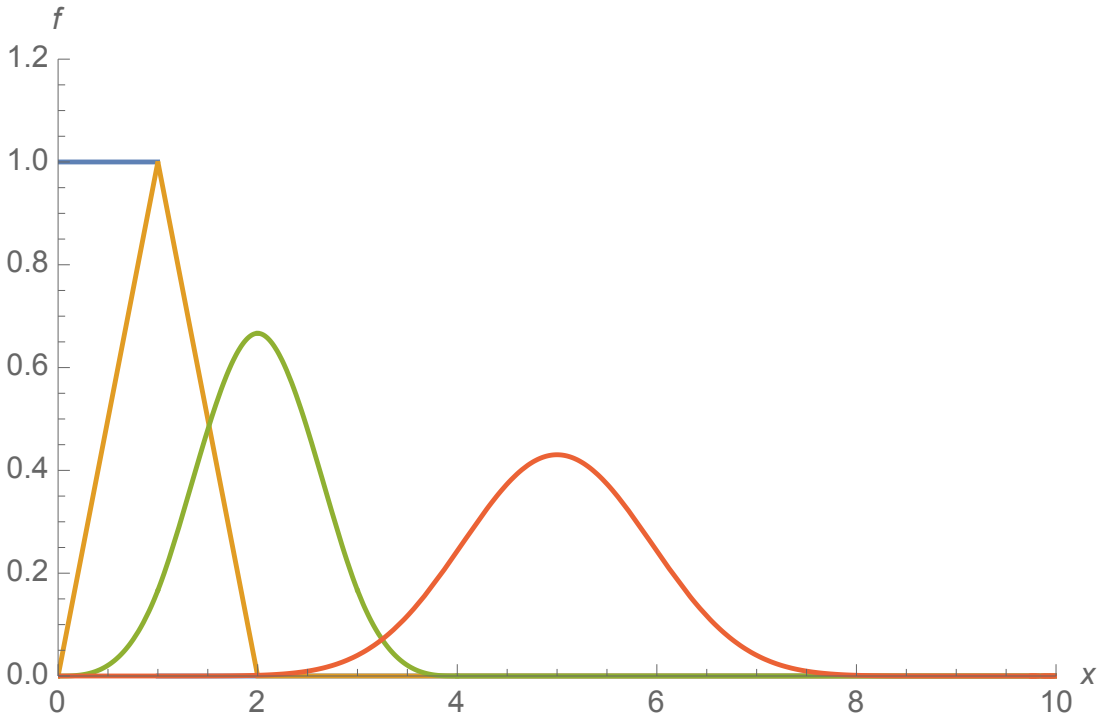


Illustration à partir de la loi uniforme $n = 10$



Pour n "grand", la normalité "saute aux yeux" !

C'est une conséquence du *Théorème central limit* (TCL, Pierre-Simon Laplace, 1809) :

Proposition

Soient $X_1, \ldots, X_n$ iid d'espérance μ et de variance σ^2. Soient $S_n = X_1 + \ldots + X_n$ et $Z_n = S_n/n$. On a

$$\frac{\sqrt{n}(Z_n - \mu)}{\sigma} \xrightarrow{\mathcal{L}} N(0,1).$$

Pour n "grand", la normalité "saute aux yeux" !

C'est une conséquence du *Théorème central limit* (TCL, Pierre-Simon Laplace, 1809) :

Proposition

Soient $X_1, \ldots, X_n$ iid d'espérance μ et de variance σ^2. Soient $S_n = X_1 + \ldots + X_n$ et $Z_n = S_n/n$. On a

$$\frac{\sqrt{n}(Z_n - \mu)}{\sigma} \xrightarrow{\mathcal{L}} N(0,1).$$

Pour n grand, on a donc $S_n \approx N(n\mu, n\sigma^2)$.

Pour n "grand", la normalité "saute aux yeux" !

C'est une conséquence du *Théorème central limit* (TCL, Pierre-Simon Laplace, 1809) :

Proposition

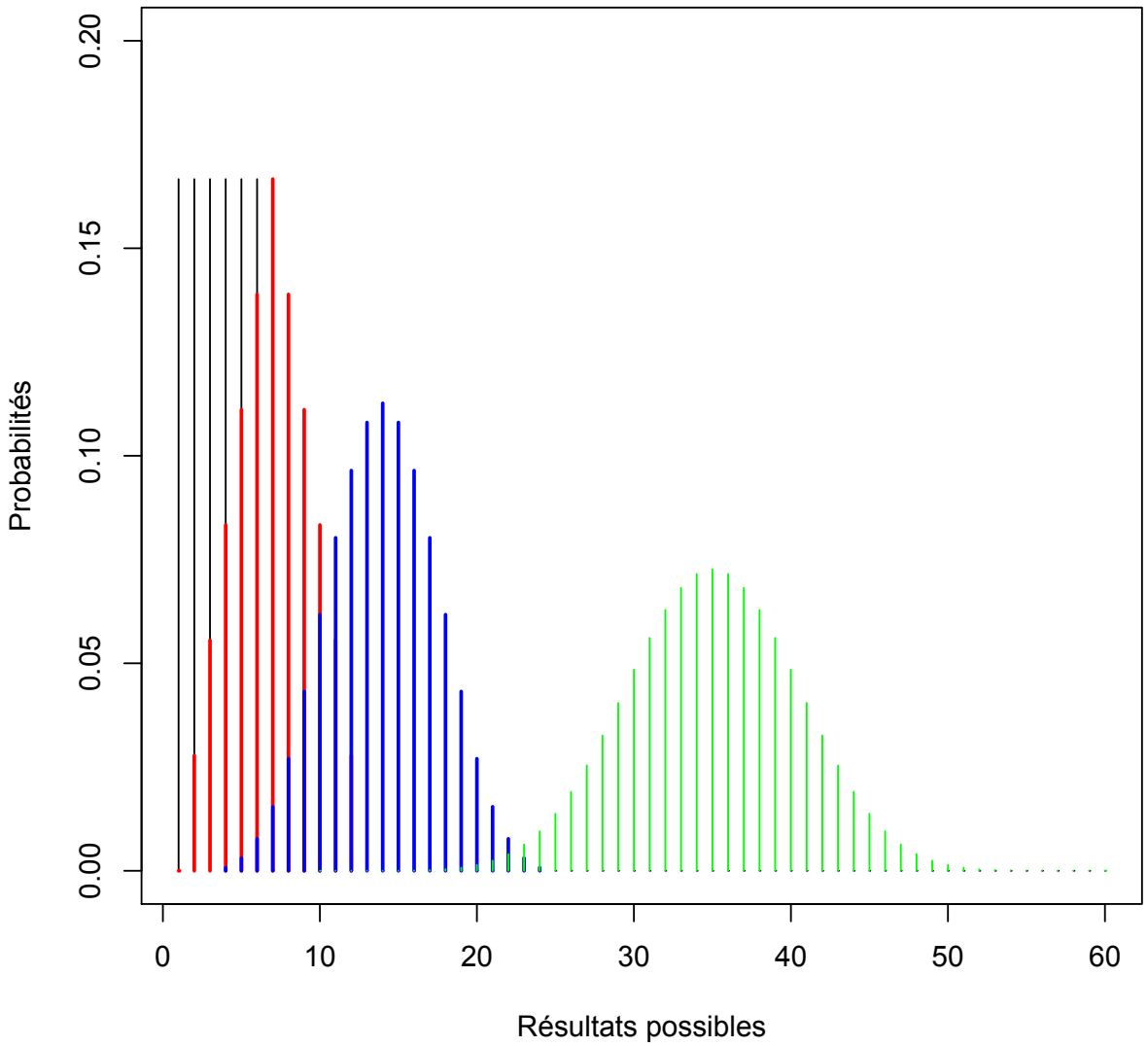
Soient $X_1, \ldots, X_n$ iid d'espérance μ et de variance σ^2. Soient $S_n = X_1 + \ldots + X_n$ et $Z_n = S_n/n$. On a

$$\frac{\sqrt{n}(Z_n - \mu)}{\sigma} \xrightarrow{\mathcal{L}} N(0,1).$$

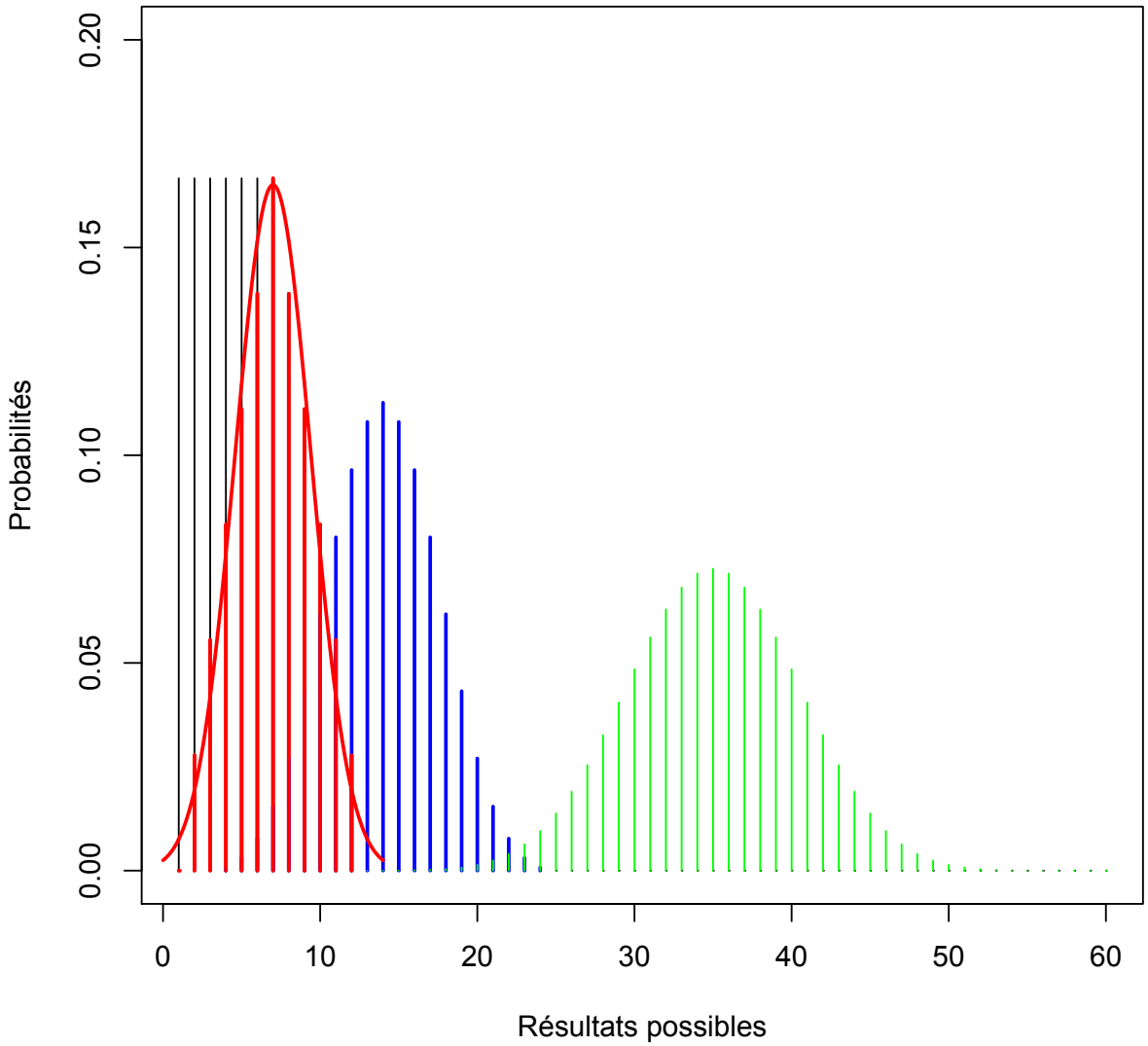
Pour n grand, on a donc $S_n \approx N(n\mu, n\sigma^2)$.

Avec $X_i \sim U(6)$, $\mu = 3.5$ et $\sigma^2 = 35/12$.

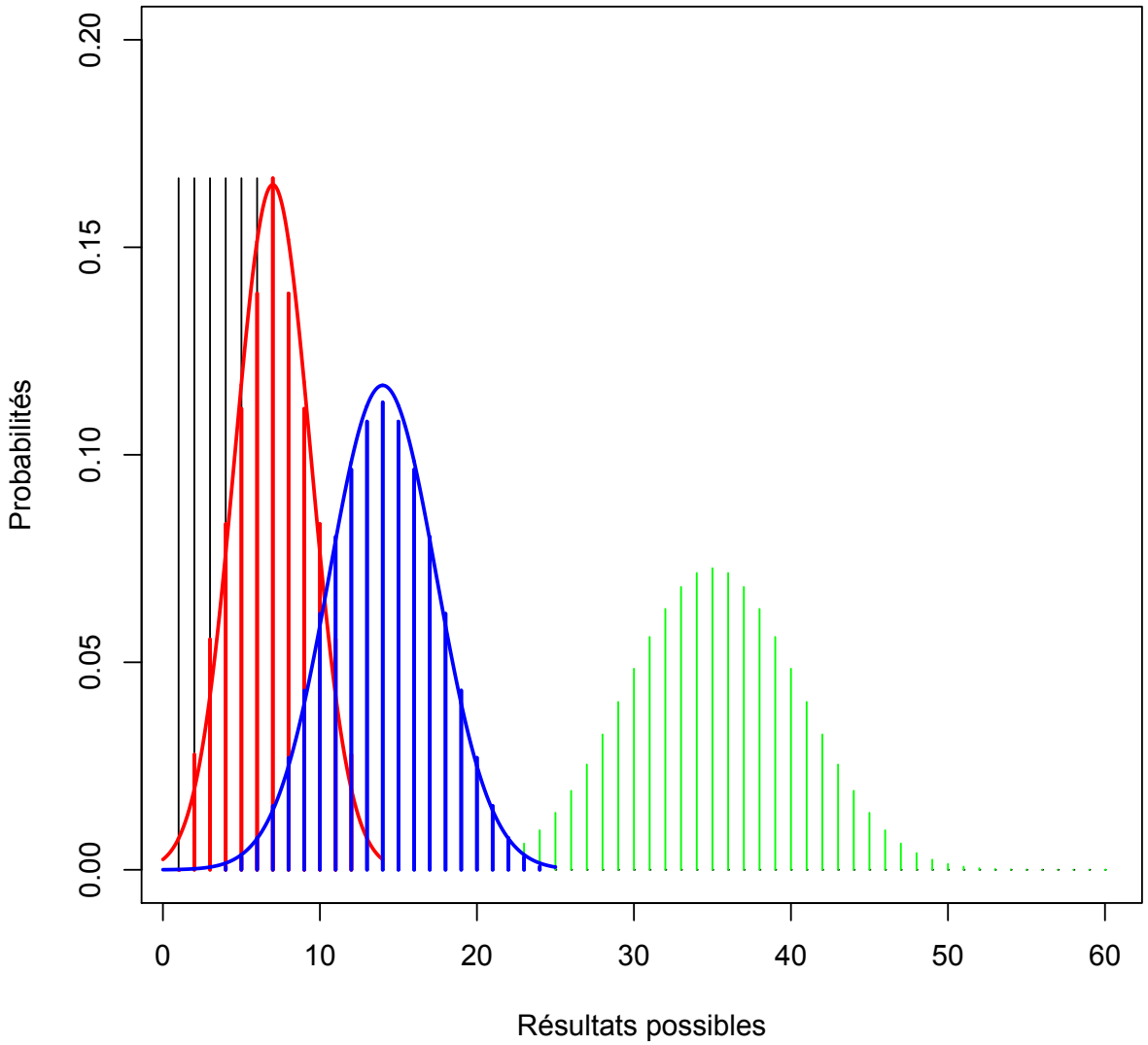
Approximation de la loi uniforme discrète par le TCL



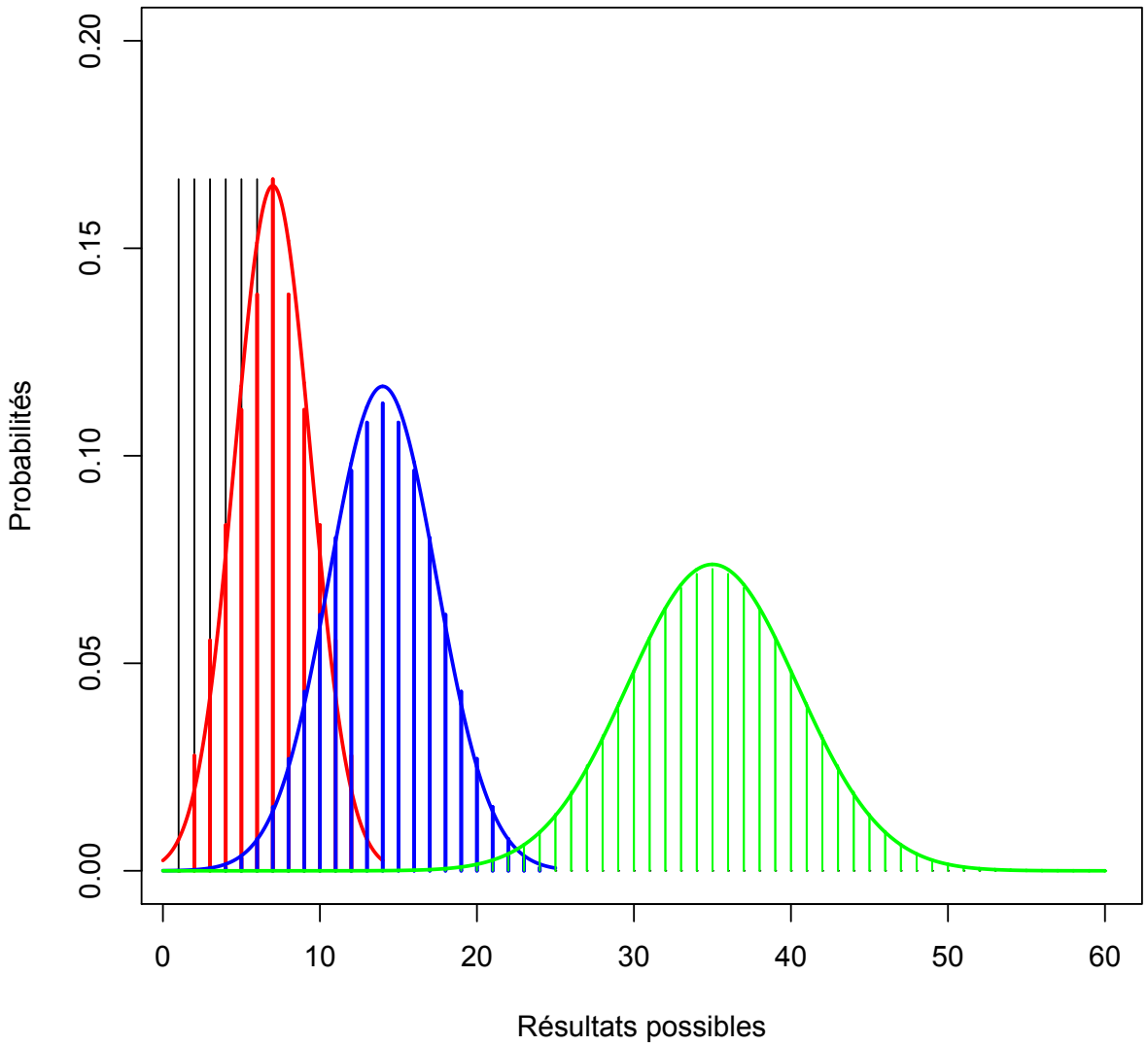
Approximation de la loi uniforme discrète par le TCL



Approximation de la loi uniforme discrète par le TCL



Approximation de la loi uniforme discrète par le TCL



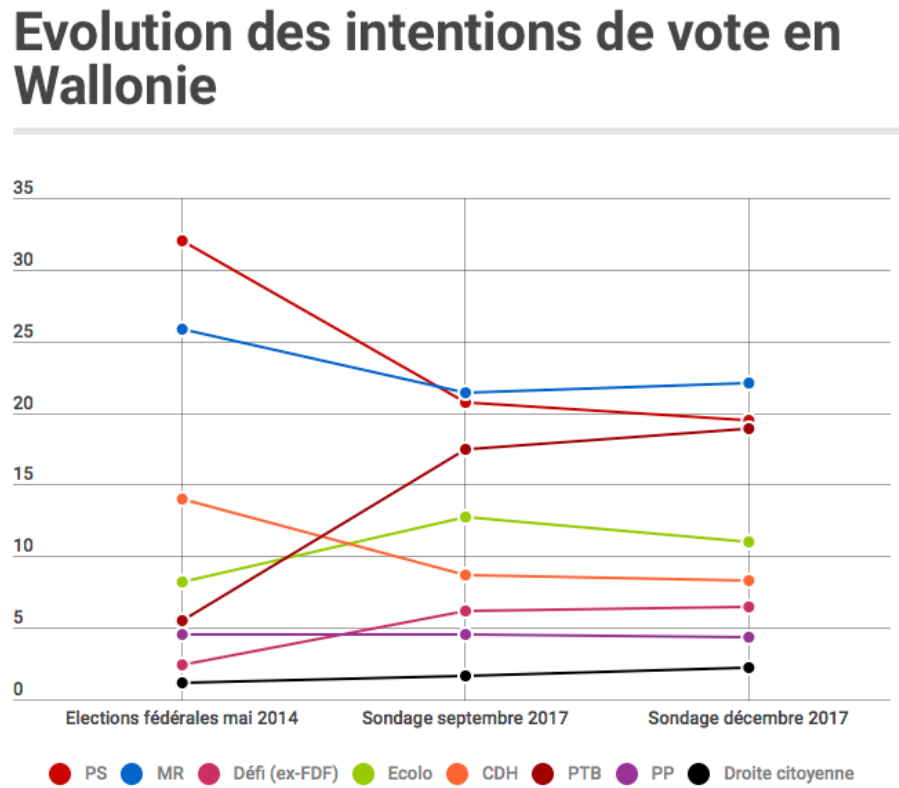
Utilisation dans les médias

LE SOIR

Grand Baromètre: PS et PTB au coude à coude en Wallonie

MIS EN LIGNE LE 8/12/2017 À 19:00 DAVID COPPI ET D.CI GRAND BAROMÈTRE POLITIQUE

Il n'y a plus que la marge d'erreur pour séparer les deux partis. Le MR reste le premier parti en Wallonie, selon le Grand Baromètre Ipsos – Le Soir – RTL – VTM et Het Laatste Nieuws.



Le TCL utilisé pour estimer la marge d'erreur

Méthodologie

Cette vague de 2.546 répondants, formant des échantillons représentatifs des Belges de 18 ans et plus à raison de 999 en Wallonie, 995 en Flandre et 552 dans les 19 communes de la Région Bruxelles-Capitale, a été réalisée du 27 novembre au 4 décembre 2017. Les interviews ont eu lieu en ligne. La marge d'erreur maximale, pour un pourcentage de 50 % et un taux de confiance de 95 % est de +- 3,1 en Wallonie, +- 3,1 en Flandre et de +— 4,2 à Bruxelles.

La marge d'erreur pour chaque résultat est basée sur l'approximation normale de la somme (moyenne) de n variables aléatoires iid selon la loi de Bernoulli (de probabilité de succès p) :

$$\overline{X}_n = Z_n \approx N\left(p, \frac{p(1-p)}{n}\right)$$

car $E[X_i] = p$ et $V[X_i] = p(1-p)$.

Marge d'erreur

A un niveau de confiance de 95%, la marge d'erreur est définie par k tq

$$P(|\bar{X}_n - p| \leq k) \geq 0.95.$$

A partir du TCL, on a

$$P\left(\frac{|\bar{X}_n - p|}{\sqrt{p(1-p)/n}} \leq \frac{k}{\sqrt{p(1-p)/n}}\right) = P\left(|Z| \leq \frac{k}{\sqrt{p(1-p)/n}}\right) \geq 0.95$$

où Z est distribuée selon la loi normale standard. D'où

$$k = 1.96\sqrt{\frac{p(1-p)}{n}}$$

expression dépendant de l'inconnue p.

La marge d'erreur **maximale** correspond à $p = 0.5$.

Calcul de la marge d'erreur

Méthodologie

Cette vague de 2.546 répondants, formant des échantillons représentatifs des Belges de 18 ans et plus à raison de 999 en Wallonie, 995 en Flandre et 552 dans les 19 communes de la Région Bruxelles-Capitale, a été réalisée du 27 novembre au 4 décembre 2017. Les interviews ont eu lieu en ligne. La marge d'erreur maximale, pour un pourcentage de 50 % et un taux de confiance de 95 % est de +- 3,1 en Wallonie, +- 3,1 en Flandre et de +— 4,2 à Bruxelles.

Si $n = 999$, on a

$$k = 1.96\sqrt{\frac{0.5(1-0.5)}{999}} = 0.03068$$

Probabilité continue

Exercice 2

Deux réels sont choisis au hasard entre 0 et 1. Quelle est la probabilité pour que la somme de ces deux nombres soit inférieure à $\frac{2}{3}$?

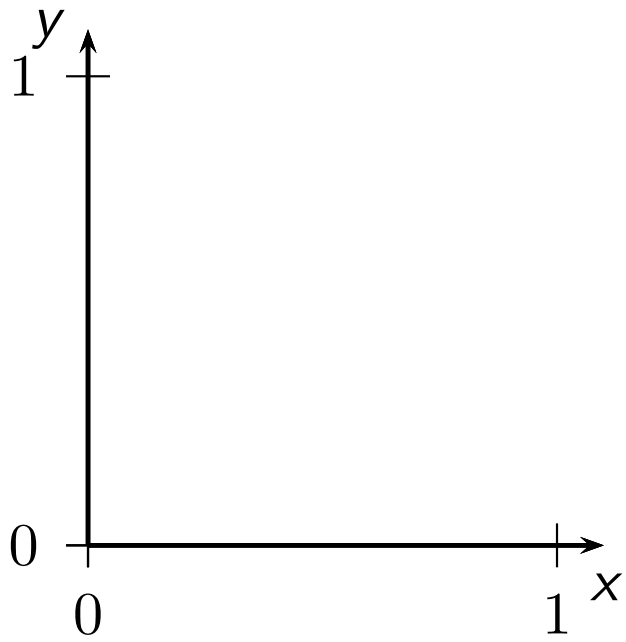
Probabilité continue

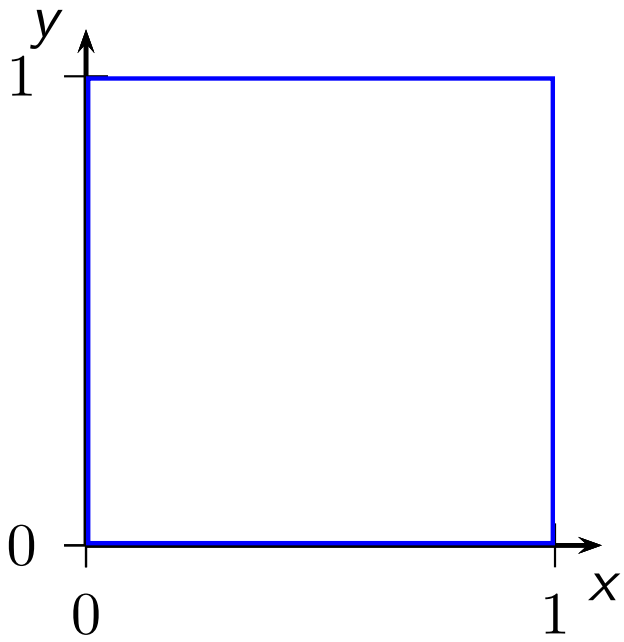
Exercice 2

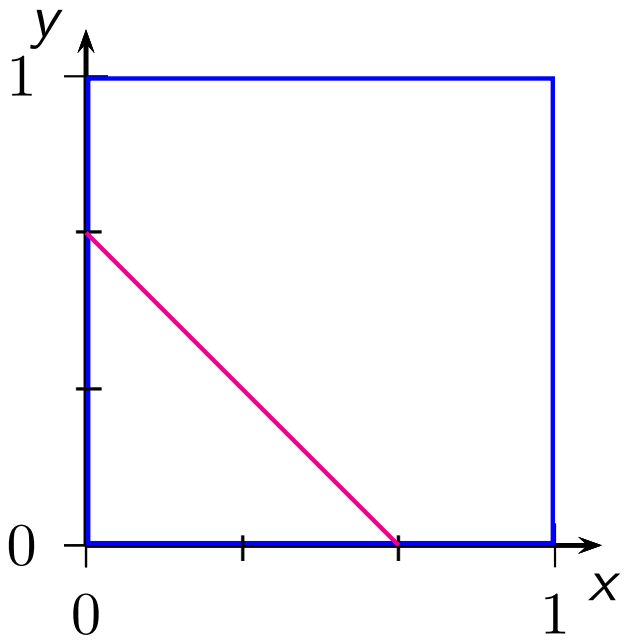
Deux réels sont choisis au hasard entre 0 et 1. Quelle est la probabilité pour que la somme de ces deux nombres soit inférieure à $\frac{2}{3}$?

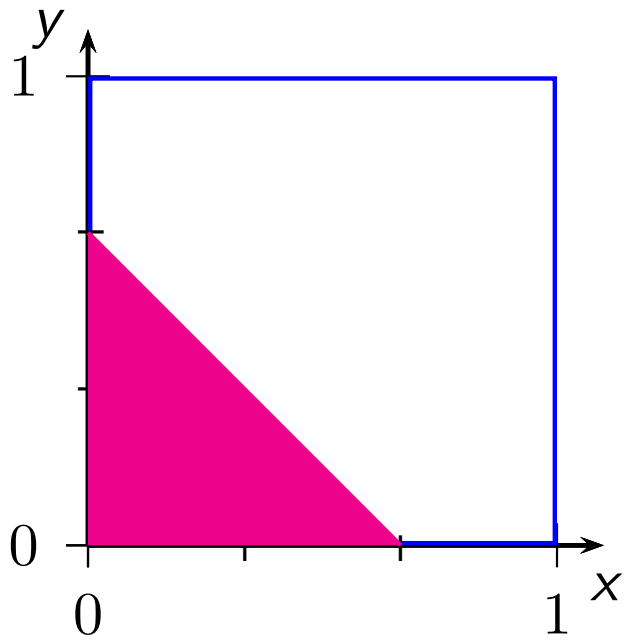
Le tableur et l'aléatoire : `=ALEA()` génère un nombre aléatoire entre 0 et 1 :

$$0 \leq \text{ ALEA() } < 1$$


y
1
0
0
1
x


y
1
0
0
1
x

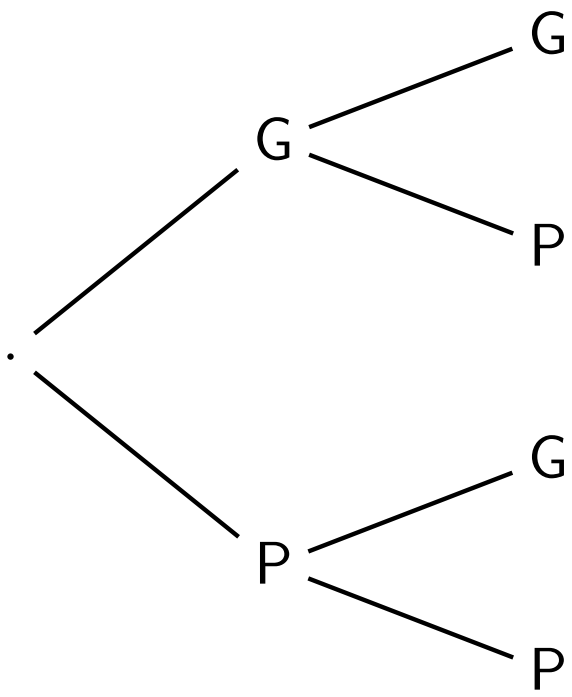

y
1
0
0
1
x

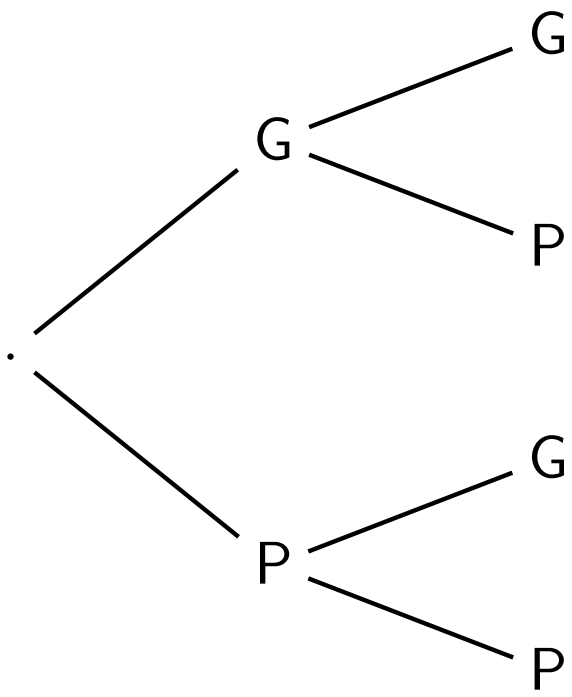


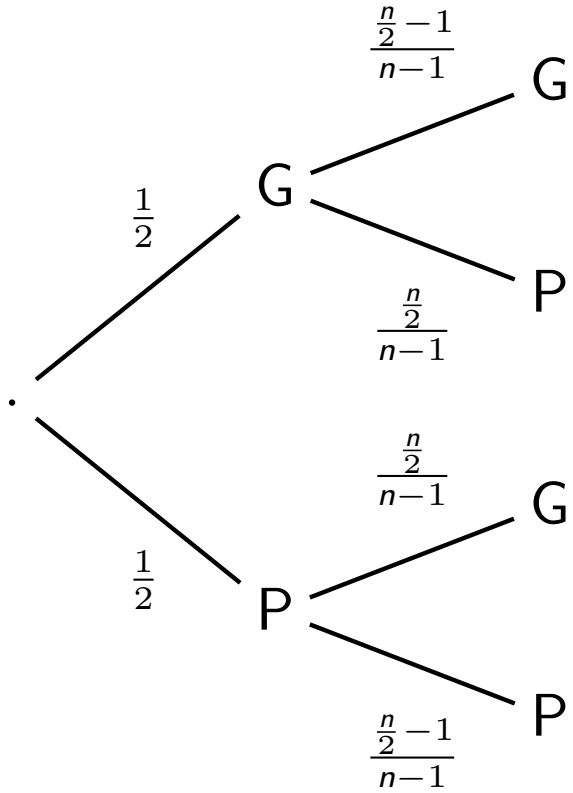
$$P = \frac{\text{Aire du triangle}}{\text{Aire du carré}} = \frac{\frac{2}{9}}{1} = \frac{2}{9}$$

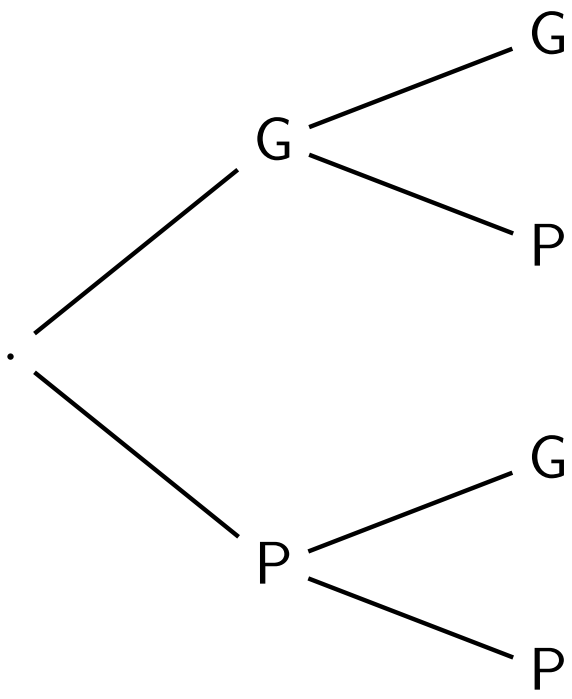
Tombola

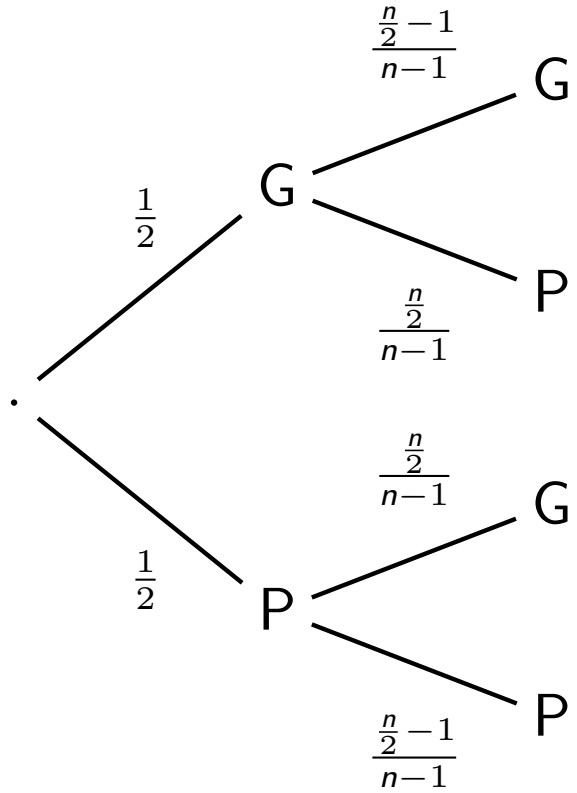
Exercice 3 : Petite histoire


G
G
P
G
P
P

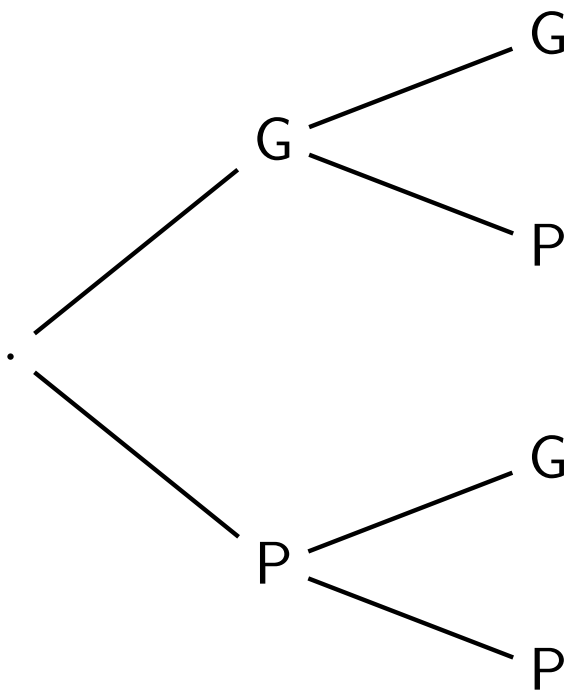

G
G
P
P
G
P

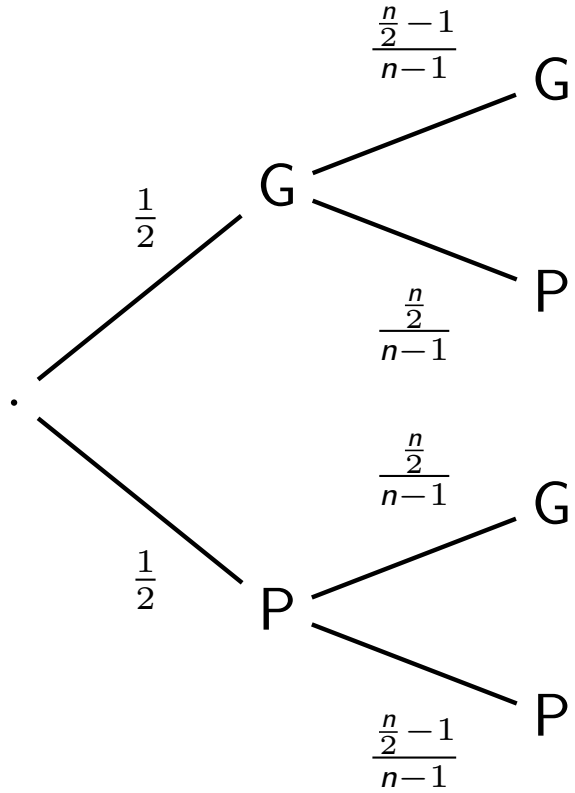

$\frac{\frac{n}{2}-1}{n-1}$
G
G
$\frac{1}{2}$
P
$\frac{\frac{n}{2}}{n-1}$
$\frac{\frac{n}{2}}{n-1}$
G
$\frac{1}{2}$
P
P
$\frac{\frac{n}{2}-1}{n-1}$



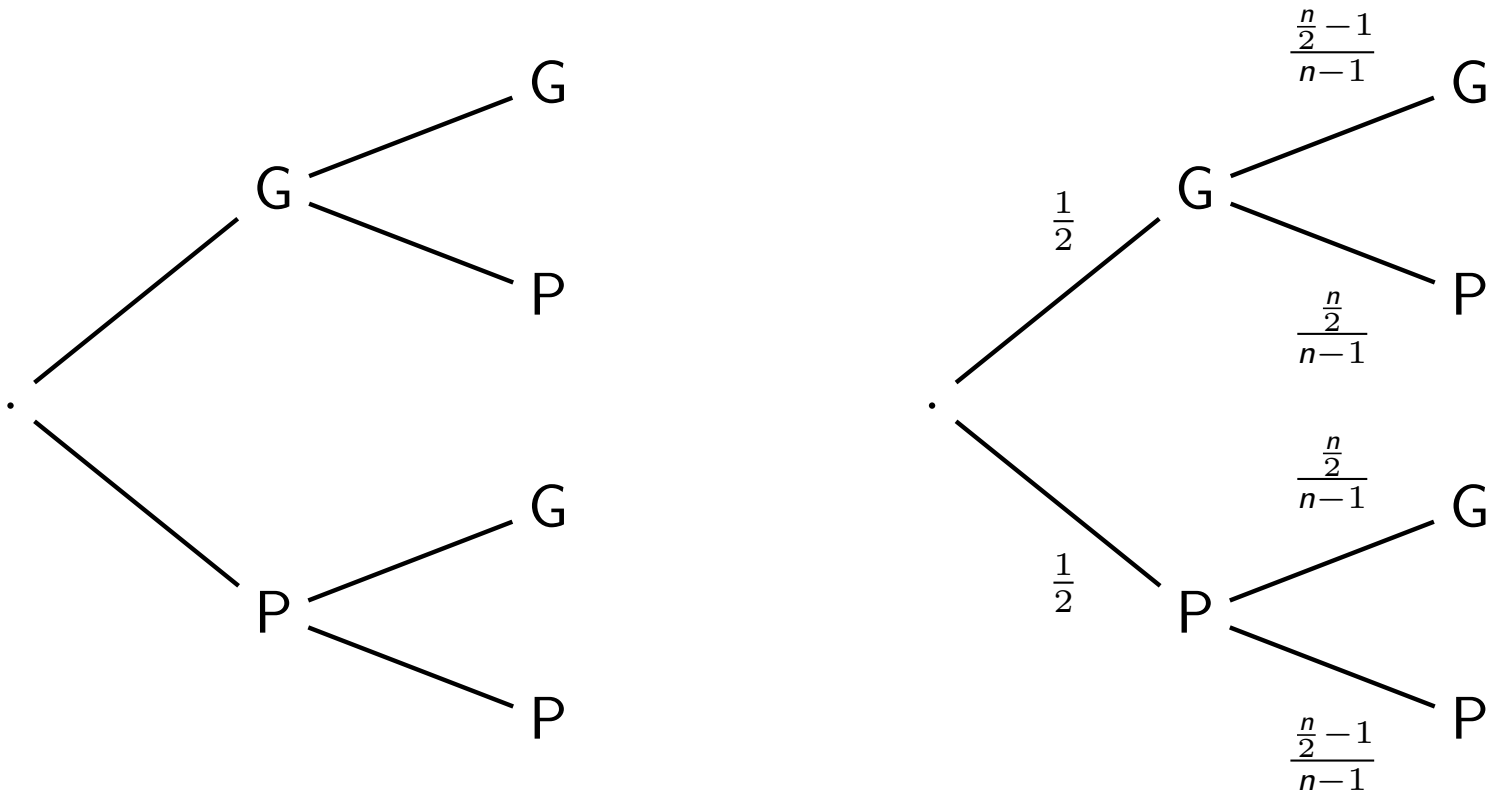


$$P(G) = \frac{1}{2}\left(\frac{\frac{n}{2}-1}{n-1} + 2\frac{\frac{n}{2}}{n-1}\right)$$





$$P(G) = \frac{1}{2}\left(\frac{\frac{n}{2}-1}{n-1} + 2\frac{\frac{n}{2}}{n-1}\right) = \frac{1}{2}\frac{3\frac{n}{2}-1}{n-1}$$



$$P(G) = \frac{1}{2}\left(\frac{\frac{n}{2}-1}{n-1} + 2\frac{\frac{n}{2}}{n-1}\right) = \frac{1}{2}\frac{3\frac{n}{2}-1}{n-1} = \frac{3n-2}{4n-4}$$

Naissances

Exercice 4 : Filles et garçons

Une étude publiée par des chercheurs de l'Université de Montréal en 2002 [a] à propos de l'influence des pesticides sur le rapport garçons/filles à la naissance a été menée dans la ville d'Ufa (fédération de Russie) auprès de 198 personnes (150 hommes et 48 femmes) ayant été exposés, dans une usine agrochimique de 1961 à 1988, à des pesticides contenant de la dioxine. Le rapport garçons/filles à la naissance pour cette ville est de 0,512 [b]. Sur la descendance des personnes étudiées, on observe 91 garçons et 136 filles, soit une fréquence observée de 0,4 garçons.

---

a. Sex Ratios of Children of Russian Pesticide Producers Exposed to Dioxin, *Environmental Health*, novembre 2002.

b. Ce rapport est reconnu comme valable dans le monde.

LIÈGE université

Générer un grand nombre d'échantillons

- Principe : simuler un grand nombre d'échantillons issus de la population et voir si la fréquence observée a des chances d'apparaître.
- Caractéristiques : $n = 227$ et $p = 0{,}512$
- Code : 1 = garçon, 0 = fille

Générer un grand nombre d'échantillons

- Principe : simuler un grand nombre d'échantillons issus de la population et voir si la fréquence observée a des chances d'apparaître.
- Caractéristiques : $n = 227$ et $p = 0{,}512$
- Code : $1 =$ garçon, $0 =$ fille

$$0 \leq \quad \text{ALEA()} \quad < 1$$

Générer un grand nombre d'échantillons

- Principe : simuler un grand nombre d'échantillons issus de la population et voir si la fréquence observée a des chances d'apparaître.
- Caractéristiques : $n = 227$ et $p = 0{,}512$
- Code : $1 =$ garçon, $0 =$ fille

$$\begin{array}{rcl} 0 \leq & \text{ALEA()} & < 1 \\ 0{,}512 \leq & \text{ALEA()} + 0{,}512 & < 1{,}512 \end{array}$$

Générer un grand nombre d'échantillons

- Principe : simuler un grand nombre d'échantillons issus de la population et voir si la fréquence observée a des chances d'apparaître.
- Caractéristiques : $n = 227$ et $p = 0{,}512$
- Code : $1 =$ garçon, $0 =$ fille

$$\begin{array}{rcl} 0 \leq & \text{ALEA()} & < 1 \\ 0{,}512 \leq & \text{ALEA()} + 0{,}512 & < 1{,}512 \end{array}$$

Dès lors, ENT(ALEA()) fournit 0 avec une probabilité égale à

Générer un grand nombre d'échantillons

- Principe : simuler un grand nombre d'échantillons issus de la population et voir si la fréquence observée a des chances d'apparaître.
- Caractéristiques : $n = 227$ et $p = 0{,}512$
- Code : $1 =$ garçon, $0 =$ fille

$$\begin{array}{rcl} 0 \leq & \text{ALEA()} & < 1 \\ 0{,}512 \leq & \text{ALEA()} + 0{,}512 & < 1{,}512 \end{array}$$

Dès lors, ENT(ALEA()) fournit 0 avec une probabilité égale à 0,488 et 1 avec une probabilité égale à 0,512.

Valeurs plausibles sous un modèle probabiliste

Supposons que le modèle théorique soit donné par

$$X \sim B(227, 0.512).$$

On peut se demander s'il est plausible d'observer un nombre de garçons égal à 91 sous ce modèle.

Valeurs plausibles sous un modèle probabiliste

Supposons que le modèle théorique soit donné par

$$X \sim B(227, 0.512).$$

On peut se demander s'il est plausible d'observer un nombre de garçons égal à 91 sous ce modèle.

Il est naturel de calculer la probabilité $P(X = 91)$. Celle-ci vaut

$$C_{227}^{91} 0.512^{91} (1 - 0.512)^{136} = 0.00019$$

Valeurs plausibles sous un modèle probabiliste

Supposons que le modèle théorique soit donné par

$$X \sim B(227, 0.512).$$

On peut se demander s'il est plausible d'observer un nombre de garçons égal à 91 sous ce modèle.

Il est naturel de calculer la probabilité $P(X = 91)$. Celle-ci vaut

$$C_{227}^{91} 0.512^{91} (1 - 0.512)^{136} = 0.00019$$

Vu qu'il y a 228 valeurs possibles pour une telle loi binomiale, il est difficile de juger.

Ecart plausible ?

Sous le modèle binomial $B(n, p)$, on s'attend, en moyenne, à observer np succès. Il devrait donc y avoir, en moyenne, 116.2 garçons.

Ecart plausible ?

Sous le modèle binomial $B(n, p)$, on s'attend, en moyenne, à observer np succès. Il devrait donc y avoir, en moyenne, 116.2 garçons.

Au lieu de se focaliser sur la valeur observée 91, il semble plus approprié de plutôt déterminer s'il est plausible d'observer un écart d'au moins $116.2 - 91 = 25.2$ garçons entre le nombre "attendu" de garçons sous le modèle et le nombre observé.

$$\boldsymbol{P}(|X - \mu| \geq 25.2)$$

Ecart plausible ?

Sous le modèle binomial $B(n, p)$, on s'attend, en moyenne, à observer np succès. Il devrait donc y avoir, en moyenne, 116.2 garçons.

Au lieu de se focaliser sur la valeur observée 91, il semble plus approprié de plutôt déterminer s'il est plausible d'observer un écart d'au moins $116.2 - 91 = 25.2$ garçons entre le nombre "attendu" de garçons sous le modèle et le nombre observé.

$$\boldsymbol{P}(|X - \mu| \geq 25.2) \quad = \quad \boldsymbol{P}(|X - \mu| > 25)$$

Ecart plausible ?

Sous le modèle binomial $B(n, p)$, on s'attend, en moyenne, à observer np succès. Il devrait donc y avoir, en moyenne, 116.2 garçons.

Au lieu de se focaliser sur la valeur observée 91, il semble plus approprié de plutôt déterminer s'il est plausible d'observer un écart d'au moins $116.2 - 91 = 25.2$ garçons entre le nombre "attendu" de garçons sous le modèle et le nombre observé.

$$\begin{aligned} \boldsymbol{P}(|X - \mu| \geq 25.2) &= \boldsymbol{P}(|X - \mu| > 25) \\ &= 1 - \boldsymbol{P}(|X - \mu| \leq 25) \end{aligned}$$

Ecart plausible ?

Sous le modèle binomial $B(n, p)$, on s'attend, en moyenne, à observer np succès. Il devrait donc y avoir, en moyenne, 116.2 garçons.

Au lieu de se focaliser sur la valeur observée 91, il semble plus approprié de plutôt déterminer s'il est plausible d'observer un écart d'au moins $116.2 - 91 = 25.2$ garçons entre le nombre "attendu" de garçons sous le modèle et le nombre observé.

$$\begin{aligned} \boldsymbol{P}(|X - \mu| \geq 25.2) &= \boldsymbol{P}(|X - \mu| > 25) \\ &= 1 - \boldsymbol{P}(|X - \mu| \leq 25) \\ &= 1 - \boldsymbol{P}(-25 \leq X - 116.2 \leq 25) \end{aligned}$$

Ecart plausible ?

Sous le modèle binomial $B(n, p)$, on s'attend, en moyenne, à observer np succès. Il devrait donc y avoir, en moyenne, 116.2 garçons.

Au lieu de se focaliser sur la valeur observée 91, il semble plus approprié de plutôt déterminer s'il est plausible d'observer un écart d'au moins $116.2 - 91 = 25.2$ garçons entre le nombre "attendu" de garçons sous le modèle et le nombre observé.

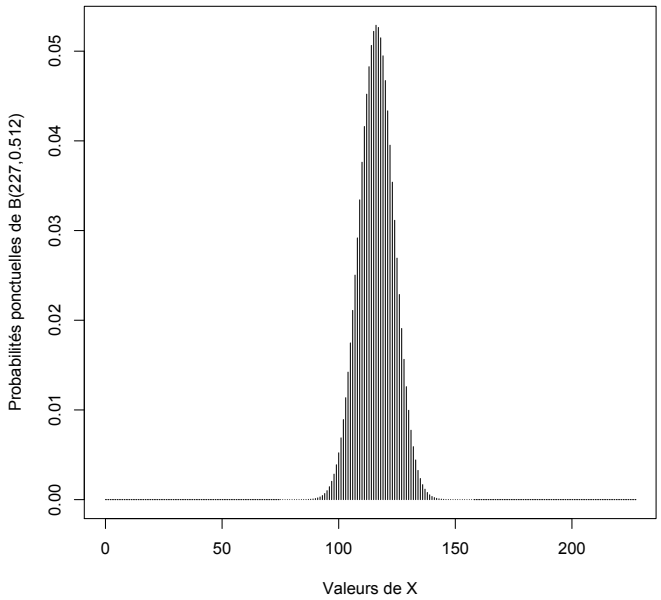
$$\begin{aligned} \boldsymbol{P}(|X - \mu| \geq 25.2) &= \boldsymbol{P}(|X - \mu| > 25) \\ &= 1 - \boldsymbol{P}(|X - \mu| \leq 25) \\ &= 1 - \boldsymbol{P}(-25 \leq X - 116.2 \leq 25) \\ &= 1 - \boldsymbol{P}(91.2 \leq X \leq 141.2) = 0.00087. \end{aligned}$$

Approximation de la loi Binomiale par une loi normale

Le calcul de la probabilité $\boldsymbol{P}(|X - \mu| \geq 25.2)$ sous la loi binomiale n'est pas compliqué mais il est "pénible" (sans ordinateur).

Approximation de la loi Binomiale par une loi normale

Le calcul de la probabilité $\boldsymbol{P}(|X-\mu| \geq 25.2)$ sous la loi binomiale n'est pas compliqué mais il est "pénible" (sans ordinateur).
Or, une loi binomiale peut être vue comme une somme de n variables aléatoires iid selon la loi de Bernoulli p : $X = X_1 + \ldots + X_n$.



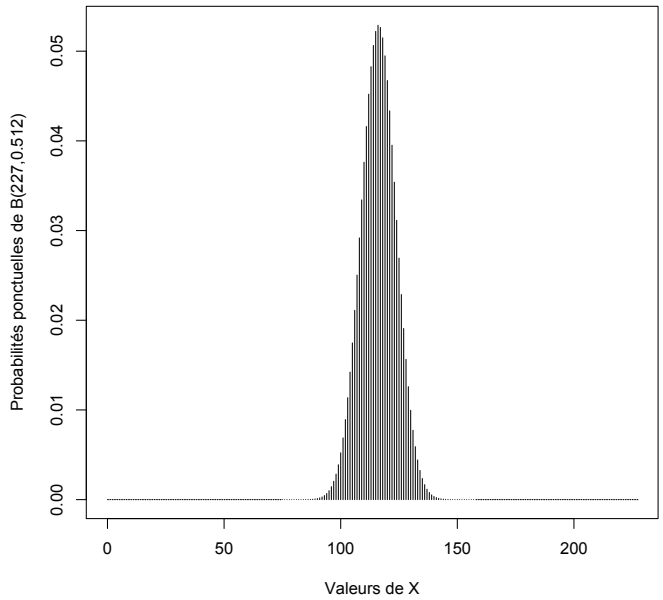
Par le théorème de De Moivre-Laplace (= le TCL "spécialisé" au cas binomial), on a, pour $n >>$

$$X = S_n \approx N(np, np(1-p))$$

car $E[X_i] = p$ et $V[X_i] = p(1-p)$.

Approximation de la loi Binomiale par une loi normale

Le calcul de la probabilité $\boldsymbol{P}(|X-\mu| \geq 25.2)$ sous la loi binomiale n'est pas compliqué mais il est "pénible" (sans ordinateur).
Or, une loi binomiale peut être vue comme une somme de n variables aléatoires iid selon la loi de Bernoulli p : $X = X_1 + \ldots + X_n$.



Par le théorème de De Moivre-Laplace (= le TCL "spécialisé" au cas binomial), on a, pour $n >>$

$$X = S_n \approx N(np, np(1-p))$$

car $E[X_i] = p$ et $V[X_i] = p(1-p)$.

On a donc

$$\boldsymbol{P}(|X-\mu| \geq 25.2) = 1 - \boldsymbol{P}(91.2 \leq S_n \leq 141.2) = 0.00090$$

Ecart ?

Quand on parle d'écart, il est souvent opportun d'exprimer celui-ci en termes d'écarts-types.

Si $X \sim B(n, p)$, alors $\sigma = \sqrt{np(1-p)}$ et on s'intéresse donc à la probabilité

$$\boldsymbol{P}(|X - \mu| \geq k\sigma)$$

pour un k donné.

Ecart ?

Quand on parle d'écart, il est souvent opportun d'exprimer celui-ci en termes d'écarts-types.

Si $X \sim B(n, p)$, alors $\sigma = \sqrt{np(1-p)}$ et on s'intéresse donc à la probabilité

$$\boldsymbol{P}(|X - \mu| \geq k\sigma)$$

pour un k donné.

Que sait-on sur cette probabilité ?

Ecart ?

Quand on parle d'écart, il est souvent opportun d'exprimer celui-ci en termes d'écarts-types.

Si $X \sim B(n,p)$, alors $\sigma = \sqrt{np(1-p)}$ et on s'intéresse donc à la probabilité

$$\boldsymbol{P}(|X-\mu| \geq k\sigma)$$

pour un k donné.

Que sait-on sur cette probabilité ? Tchebychev nous dit

$$\boldsymbol{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Ecart ?

Quand on parle d'écart, il est souvent opportun d'exprimer celui-ci en termes d'écarts-types.

Si $X \sim B(n,p)$, alors $\sigma = \sqrt{np(1-p)}$ et on s'intéresse donc à la probabilité

$$P(|X - \mu| \geq k\sigma)$$

pour un k donné.

Que sait-on sur cette probabilité ? Tchebychev nous dit

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Dans notre application, $\sigma = 7.53$, d'où, pour obtenir globalement un écart de 25,2 unités, il faut prendre $k = 25.2/\sigma \approx 3.35$ et l'inégalité de Tchebychev nous dit que la probabilité cherchée est inférieure ou égale à 0.089.

Une probabilité égale à 0.00090 ou 0.00087, est-ce trop faible pour être “honnête” ?

La probabilité ponctuelle étant non nulle, observer 91 garçons n’est pas impossible sous ce modèle... Observer un écart d’au moins 25.2 unités par rapport à la moyenne n’est pas impossible non plus.

Une probabilité égale à 0.00090 ou 0.00087, est-ce trop faible pour être “honnête” ?

La probabilité ponctuelle étant non nulle, observer 91 garçons n’est pas impossible sous ce modèle... Observer un écart d’au moins 25.2 unités par rapport à la moyenne n’est pas impossible non plus.

Néanmoins, ces événements sont tellement “rares” que l’on serait tenté de plutôt conclure que notre hypothèse de départ concernant le modèle n’était pas adéquate.

LIÈGE université

Une probabilité égale à 0.00090 ou 0.00087, est-ce trop faible pour être “honnête” ?

La probabilité ponctuelle étant non nulle, observer 91 garçons n’est pas impossible sous ce modèle... Observer un écart d’au moins 25.2 unités par rapport à la moyenne n’est pas impossible non plus.

Néanmoins, ces événements sont tellement “rares” que l’on serait tenté de plutôt conclure que notre hypothèse de départ concernant le modèle n’était pas adéquate.

Il faut en tout cas prendre une décision à propos de celle-ci !

LIÈGE université

Hypothèses, test statistique et risque d'erreur

Un test statistique confronte deux hypothèses :

$$H_0 : \begin{array}{c}\text{hypothèse vraie} \\ \text{par défaut}\end{array} \longleftrightarrow H_1 : \begin{array}{c}\text{hypothèse} \\ \text{alternative}\end{array}$$

Et, au final, on a les possibilités suivantes :

	Décision	
Réalité	Rejeter H_0	Ne pas rejeter H_0
H_0 est vraie		
H_0 est fausse		

Hypothèses, test statistique et risque d'erreur

Un test statistique confronte deux hypothèses :

$$H_0 : \begin{array}{c}\text{hypothèse vraie} \\ \text{par défaut}\end{array} \longleftrightarrow H_1 : \begin{array}{c}\text{hypothèse} \\ \text{alternative}\end{array}$$

Et, au final, on a les possibilités suivantes :

	Décision	
Réalité	Rejeter H_0	Ne pas rejeter H_0
H_0 est vraie		v
H_0 est fausse	v	

Hypothèses, test statistique et risque d'erreur

Un test statistique confronte deux hypothèses :

$$H_0 : \begin{array}{c}\text{hypothèse vraie} \\ \text{par défaut}\end{array} \longleftrightarrow H_1 : \begin{array}{c}\text{hypothèse} \\ \text{alternative}\end{array}$$

Et, au final, on a les possibilités suivantes :

	Décision	
Réalité	Rejeter H_0	Ne pas rejeter H_0
H_0 est vraie	x	v
H_0 est fausse	v	x

Hypothèses, test statistique et risque d'erreur

Un test statistique confronte deux hypothèses :

$$H_0 : \begin{array}{c}\text{hypothèse vraie} \\ \text{par défaut}\end{array} \longleftrightarrow H_1 : \begin{array}{c}\text{hypothèse} \\ \text{alternative}\end{array}$$

Et, au final, on a les possibilités suivantes :

	Décision	
Réalité	Rejeter H_0	Ne pas rejeter H_0
H_0 est vraie	x	v
H_0 est fausse	v	x

Le niveau du test est défini par α ($\in [0,1]$) tel que $\boldsymbol{P}(RH_0|H_0) \leq \alpha$.

Hypothèses, test statistique et risque d'erreur

Un test statistique confronte deux hypothèses :

$$H_0 : \begin{array}{c}\text{hypothèse vraie} \\ \text{par défaut}\end{array} \longleftrightarrow H_1 : \begin{array}{c}\text{hypothèse} \\ \text{alternative}\end{array}$$

Et, au final, on a les possibilités suivantes :

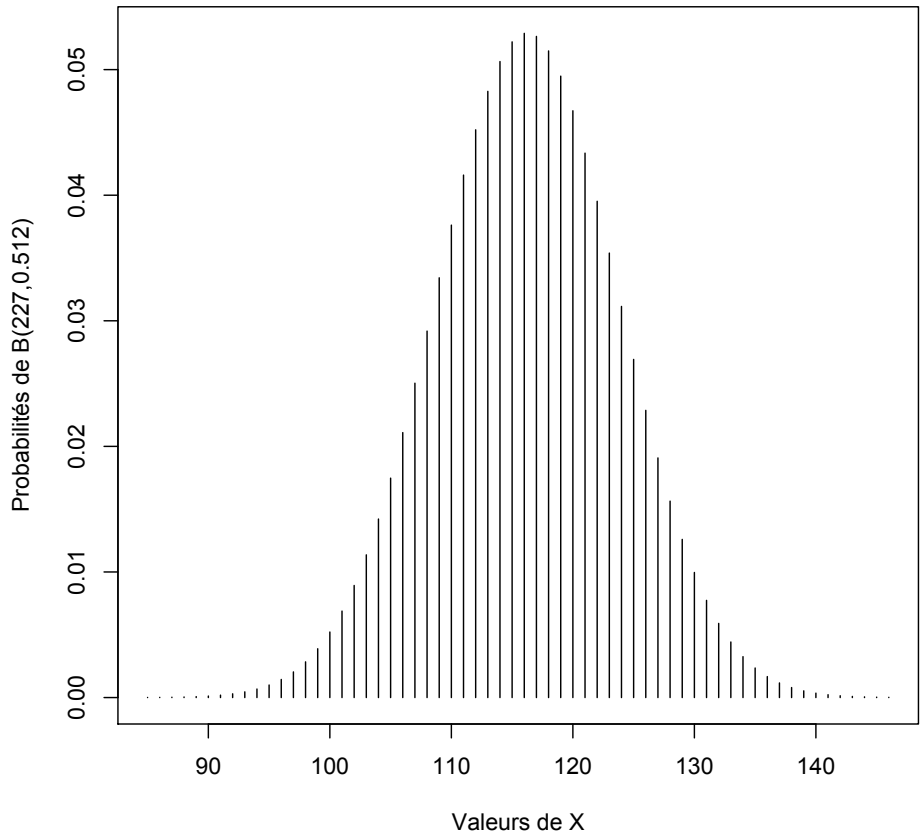
	Décision	
Réalité	Rejeter H_0	Ne pas rejeter H_0
H_0 est vraie	x	v
H_0 est fausse	v	x

Le niveau du test est défini par α ($\in [0,1]$) tel que $\boldsymbol{P}(RH_0|H_0) \leq \alpha$.

Procédure de test : rejeter H_0 si les données ne sont vraiment pas compatibles avec H_0, i.e. si la valeur observée de la statistique de test ne se trouve pas dans l'intervalle regroupant, avec une probabilité $1-\alpha$, les valeurs les plus plausibles sous H_0.

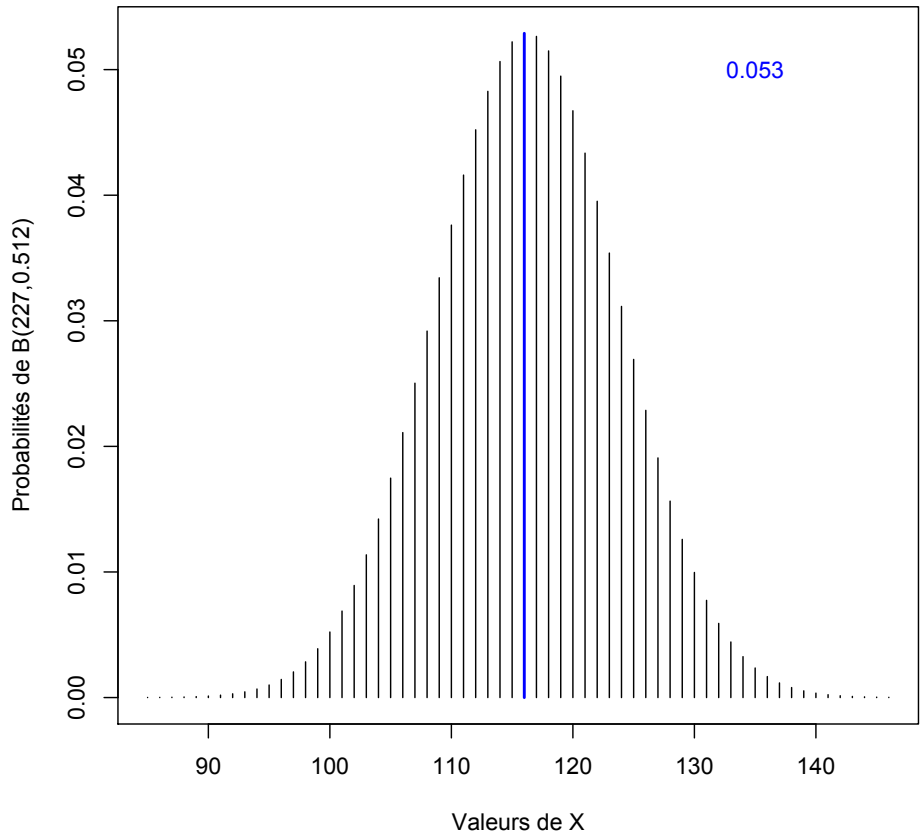
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



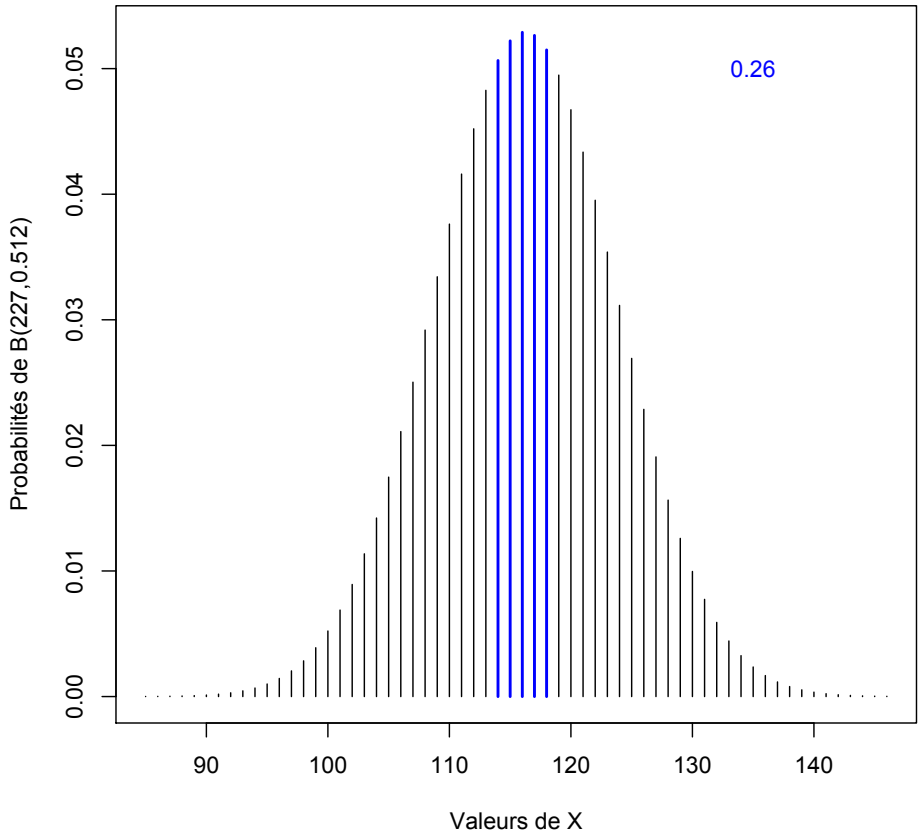
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



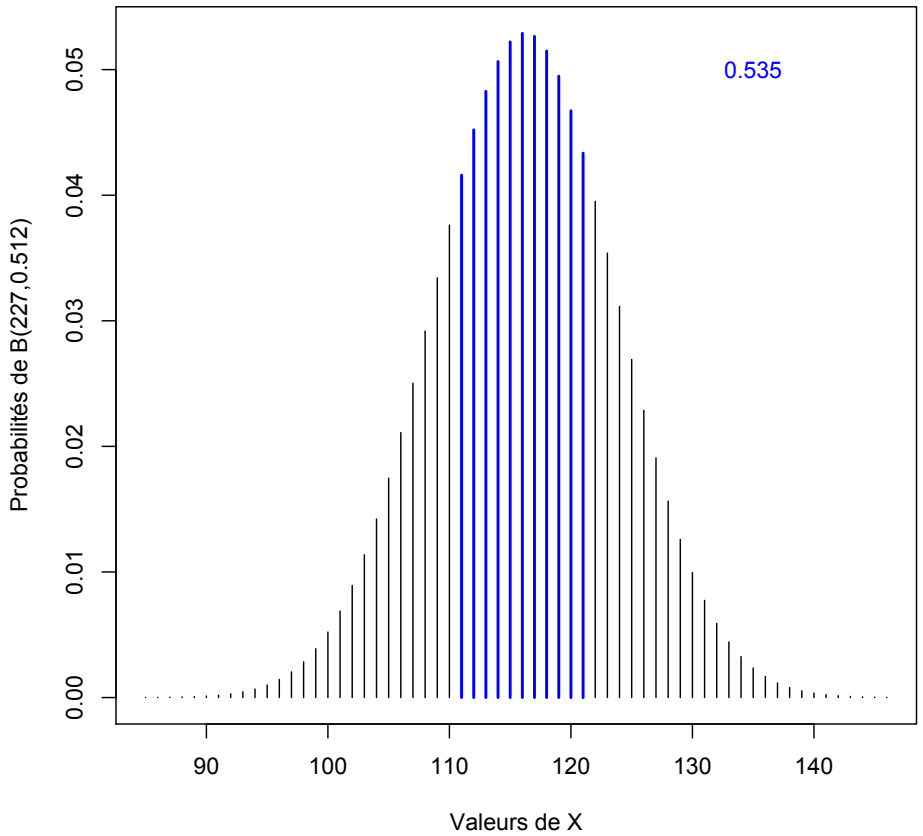
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



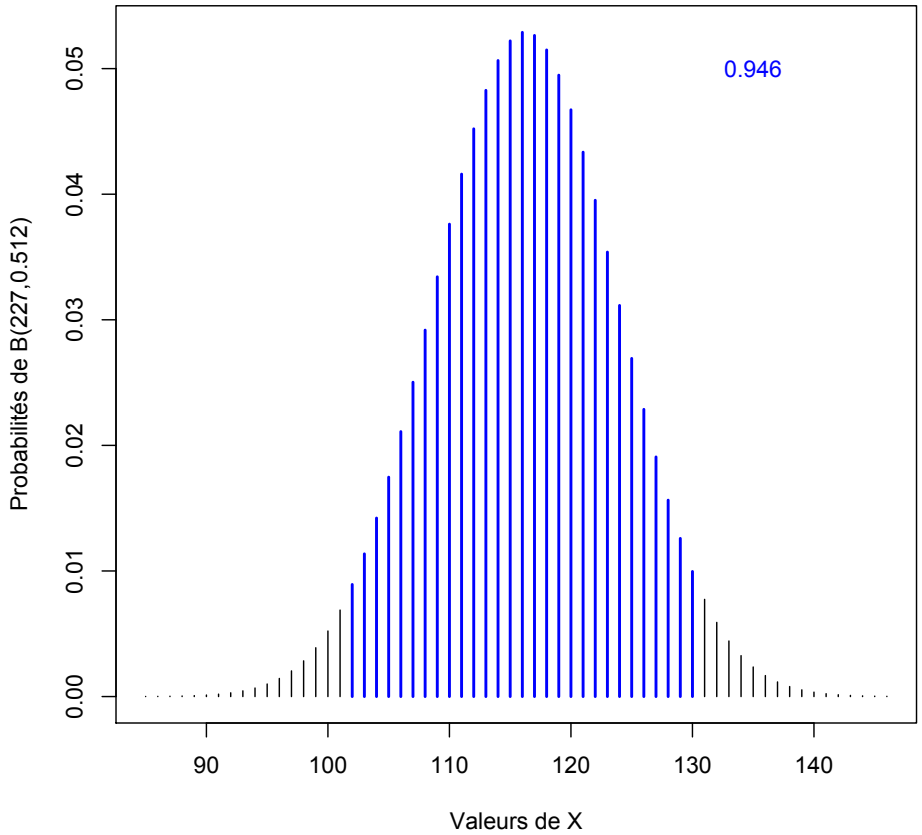
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



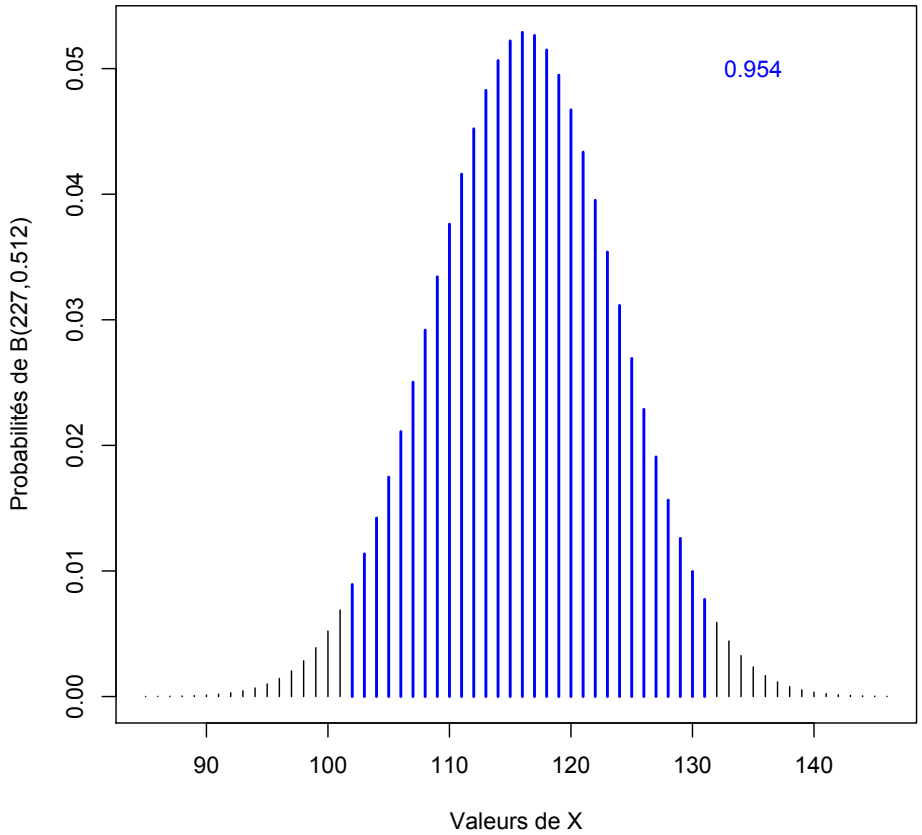
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



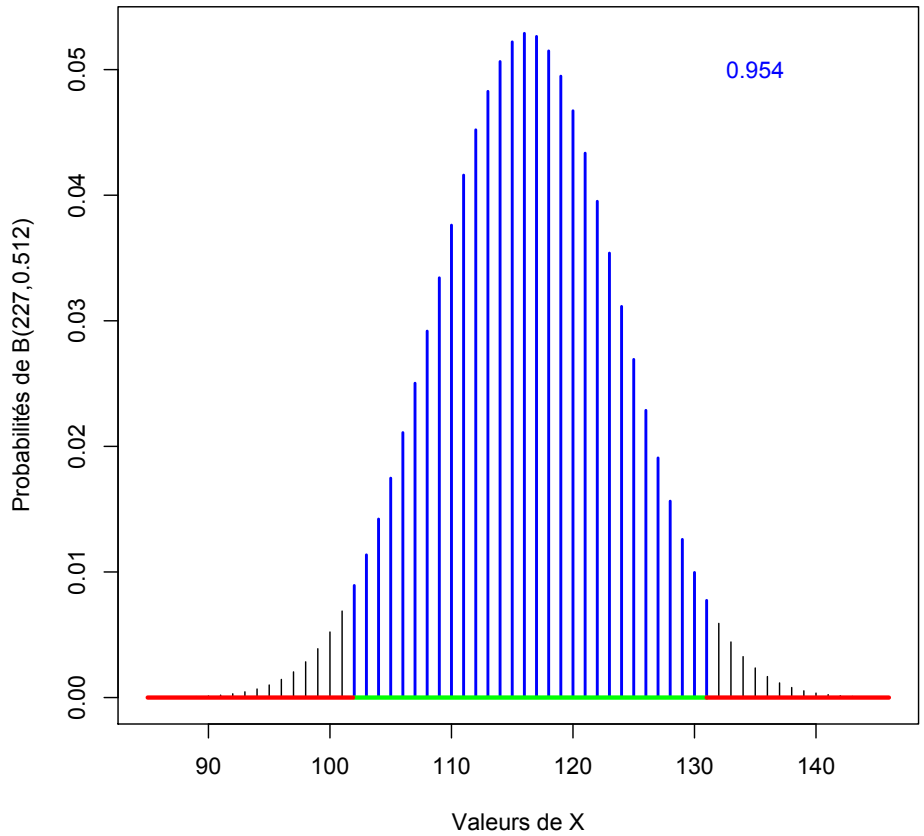
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



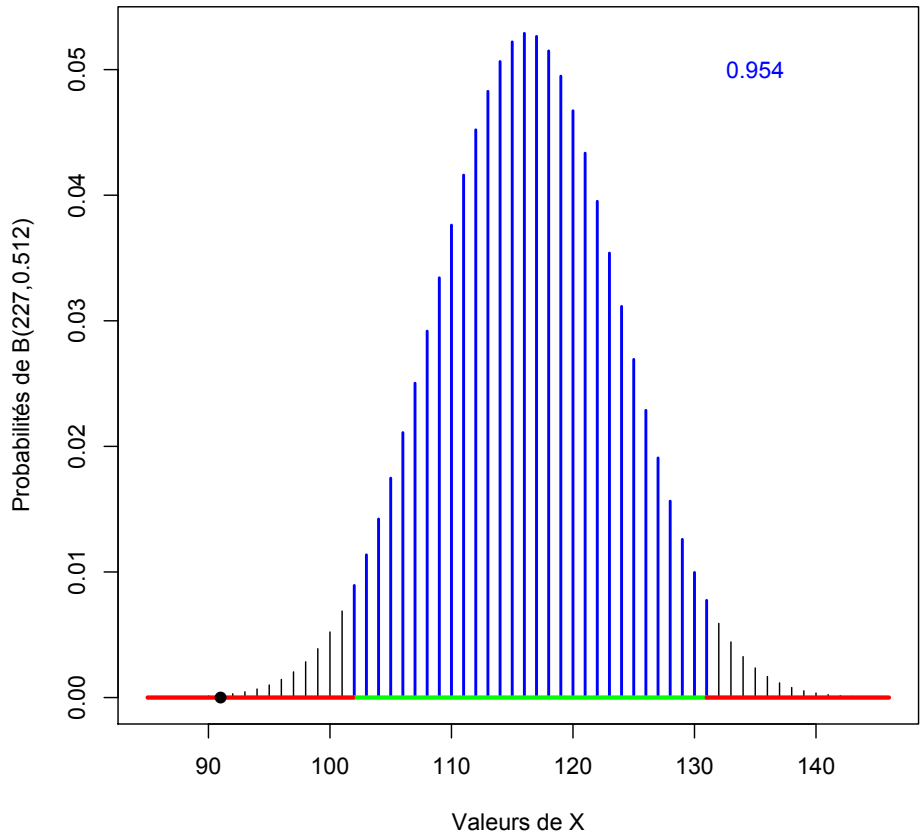
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



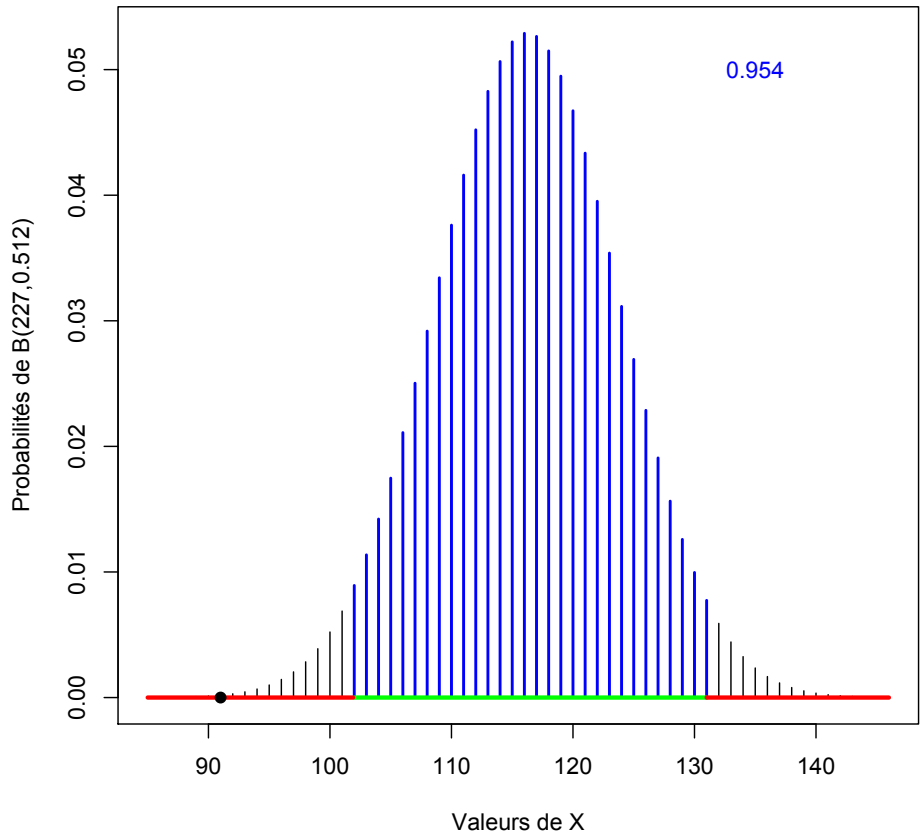
Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :



Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Prenons $\alpha = 0.05$. Sous H_0, la distribution du nombre de garçons est donnée par (version zoomée) :

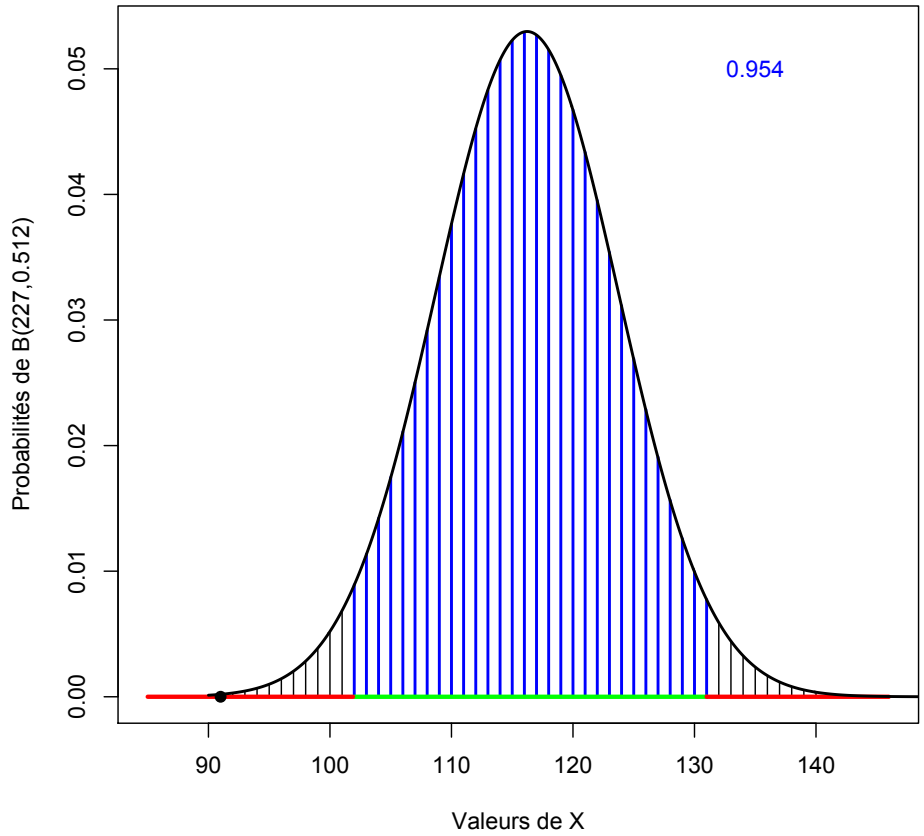


On obtient l'intervalle $[102, 131]$; on rejette H_0 au niveau de confiance 95%.

Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Sous l'approximation normale :

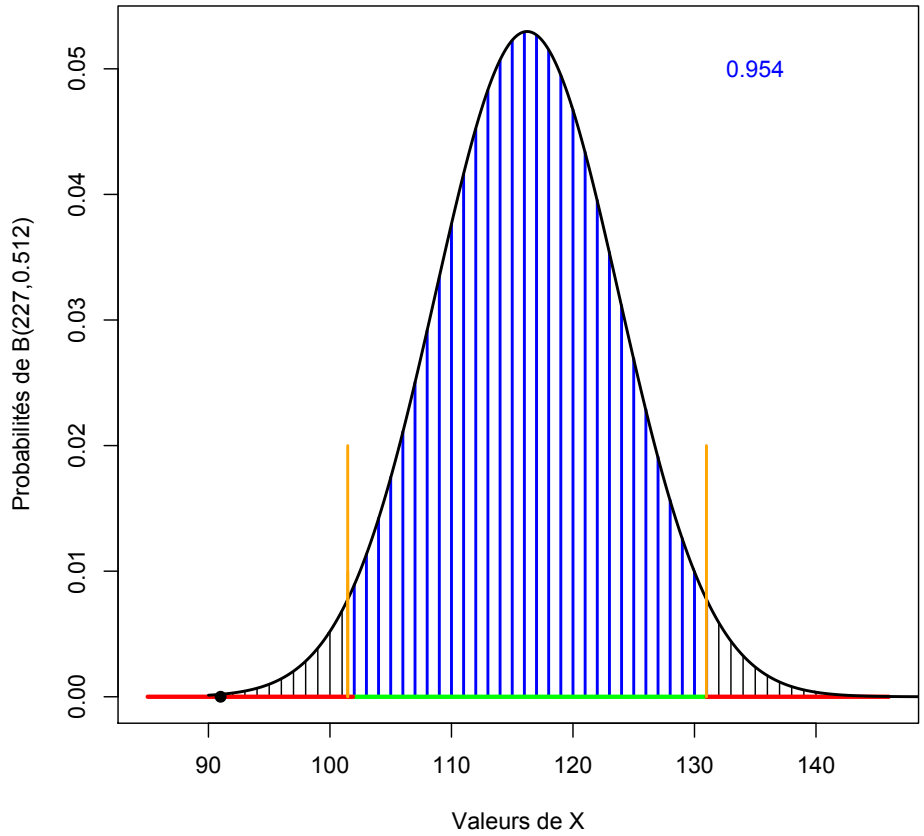
$$P(np - 1.96\sqrt{np(1-p)} \leq X \leq np + 1.96\sqrt{np(1-p)}) = 0.95$$



Dans notre contexte : $H_0 : p = 0.512 \longleftrightarrow H_1 : p \neq 0.512$

Sous l'approximation normale :

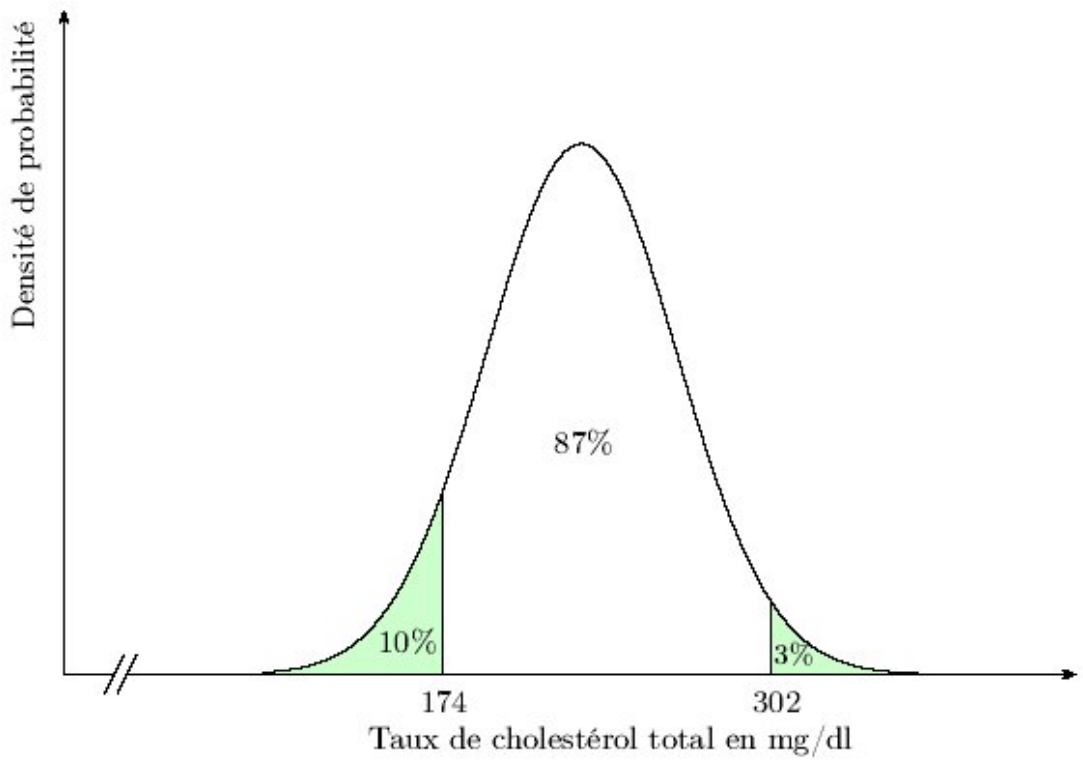
$$P(np - 1.96\sqrt{np(1-p)} \leq X \leq np + 1.96\sqrt{np(1-p)}) = 0.95$$



On obtient l'intervalle $[101.5; 131]$ menant à nouveau au rejet de H_0, avec un niveau de confiance de 95%.

Exercice 5

Un magazine médical présente le graphique ci-dessous donnant la répartition du taux de cholestérol total dans une population comportant un grand nombre d'individus.
Les personnes dont le taux de cholestérol total est supérieur à 190 mg/dl doivent subir un examen complémentaire. Pour un échantillon extrait de la population étudiée, on constate que 770 personnes sur 1000 sont dans le cas. Cet échantillon est-il représentatif de cette population ?



Loi normale

On note X le taux de cholestérol et on sait que $X \sim N(\mu, \sigma)$.

Loi normale

On note X le taux de cholestérol et on sait que $X \sim N(\mu, \sigma)$. Les informations du graphique fournissent le système suivant :

$$\begin{cases} P(X \leq 174) &= 0{,}1 \\ P(X \geq 302) &= 0{,}03 \end{cases}$$

Loi normale

On note X le taux de cholestérol et on sait que $X \sim N(\mu, \sigma)$. Les informations du graphique fournissent le système suivant :

$$\begin{cases} P(X \leq 174) &= 0{,}1 \\ P(X \geq 302) &= 0{,}03 \end{cases}$$

Soit $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$.

Loi normale

On note X le taux de cholestérol et on sait que $X \sim N(\mu, \sigma)$. Les informations du graphique fournissent le système suivant :

$$\begin{cases} P(X \leq 174) &= 0{,}1 \\ P(X \geq 302) &= 0{,}03 \end{cases}$$

Soit $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$. Le système devient

Loi normale

On note X le taux de cholestérol et on sait que $X \sim N(\mu, \sigma)$. Les informations du graphique fournissent le système suivant :

$$\begin{cases} P(X \leq 174) &= 0{,}1 \\ P(X \geq 302) &= 0{,}03 \end{cases}$$

Soit $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$. Le système devient

$$\begin{cases} P\left(Z \leq \dfrac{174 - \mu}{\sigma}\right) &= 0{,}1 \\ P\left(Z \geq \dfrac{302 - \mu}{\sigma}\right) &= 0{,}03 \end{cases}$$

Loi normale

On note X le taux de cholestérol et on sait que $X \sim N(\mu, \sigma)$. Les informations du graphique fournissent le système suivant :

$$\begin{cases} P(X \leq 174) &= 0{,}1 \\ P(X \geq 302) &= 0{,}03 \end{cases}$$

Soit $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$. Le système devient

$$\begin{cases} P\left(Z \leq \dfrac{174 - \mu}{\sigma}\right) &= 0{,}1 \\ P\left(Z \geq \dfrac{302 - \mu}{\sigma}\right) &= 0{,}03 \end{cases}$$

Excel ou Geogebra fournissent

Loi normale

On note X le taux de cholestérol et on sait que $X \sim N(\mu, \sigma)$. Les informations du graphique fournissent le système suivant :

$$\begin{cases} P(X \leq 174) &= 0{,}1 \\ P(X \geq 302) &= 0{,}03 \end{cases}$$

Soit $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$. Le système devient

$$\begin{cases} P\left(Z \leq \dfrac{174 - \mu}{\sigma}\right) &= 0{,}1 \\ P\left(Z \geq \dfrac{302 - \mu}{\sigma}\right) &= 0{,}03 \end{cases}$$

Excel ou Geogebra fournissent

$$\begin{cases} \dfrac{174 - \mu}{\sigma} &= -1{,}2816 \\ \dfrac{302 - \mu}{\sigma} &= 1{,}8808 \end{cases}$$

On trouve comme solution $\mu = 225{,}87$ et $\sigma = 40{,}48$.

On trouve comme solution $\mu = 225{,}87$ et $\sigma = 40{,}48$.
On cherche ensuite $\boldsymbol{P}(X \geq 190)$

On trouve comme solution $\mu = 225{,}87$ et $\sigma = 40{,}48$.
On cherche ensuite $\boldsymbol{P}(X \geq 190)$ et on trouve $0{,}8122$.