Marcel Délèze
Edition 2017

```
Needs["Statistique`",
nécessite
 "https://www.deleze.name/marcel/sec2/applmaths/packages/Statistique.m"]
```

§ 2 Variables aléatoires indépendantes

§ 2.1 Variables indépendantes

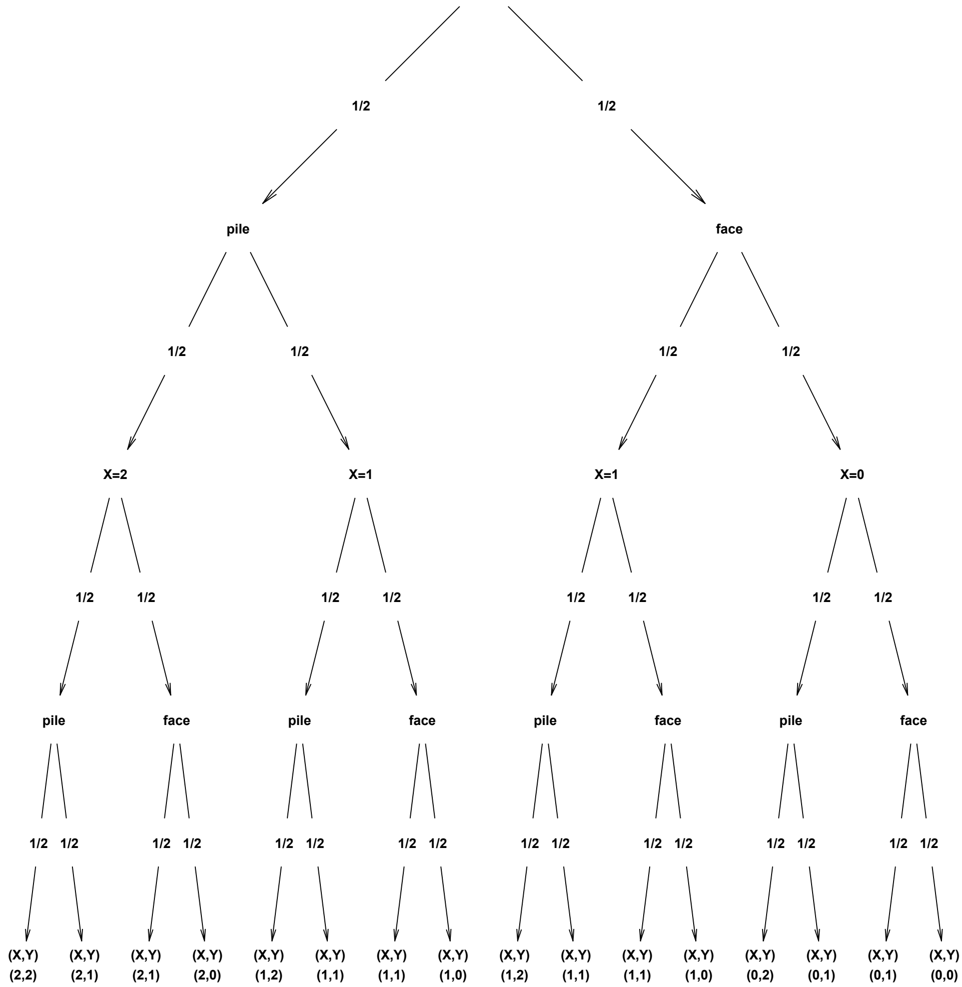
Exemple de deux variables indépendantes

On lance **quatre** fois une pièce de monnaie. On note

X = nombre de piles obtenus lors des deux premiers lancers,
Y = nombre de piles obtenus lors des deux derniers lancers.

Intuitivement, les valeurs de Y ne sont pas "influencées" par les résultats de X.
Les variables X et Y sont indépendantes.

Nous aimerions caractériser la situation par une relation mathématique. Représentons les différentes issues possibles pendant le déroulement des quatre lancers de la pièce de monnaie. A chaque lancer, il y a deux possibilités qu'on peut représenter par un embranchement.

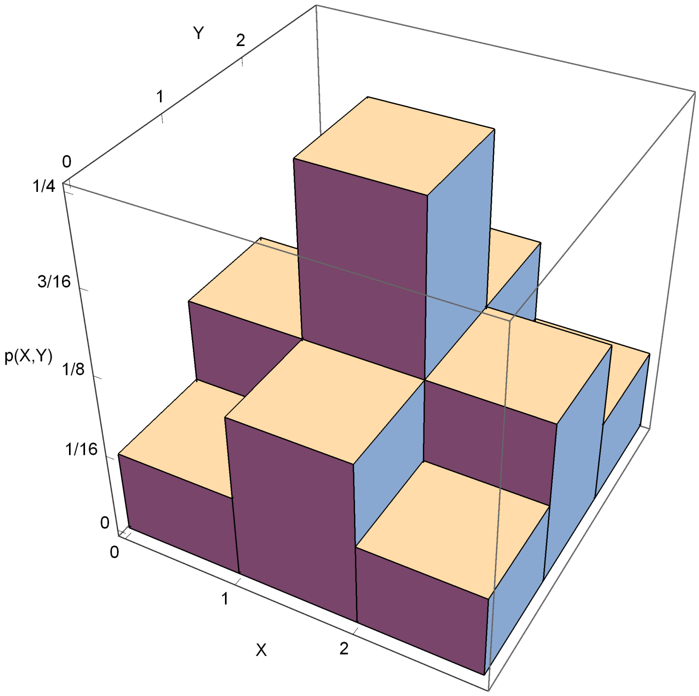


Chaque événement situé à l'extrémité d'une branche (tout en bas) a une 1 chance sur 16 de se produire. Donc

$$p((X=2, Y=2)) = \frac{1}{16}$$
$$p((X=2, Y=1)) = \frac{2}{16}$$
$$p((X=2, Y=0)) = \frac{1}{16}$$
$$p((X=1, Y=2)) = \frac{2}{16}$$
$$p((X=1, Y=1)) = \frac{4}{16}$$
$$p((X=1, Y=0)) = \frac{2}{16}$$
$$p((X=0, Y=2)) = \frac{1}{16}$$
$$p((X=0, Y=1)) = \frac{2}{16}$$
$$p((X=0, Y=0)) = \frac{1}{16}$$

Avec les résultats précédents, dressons un tableau des probabilités

	Y=0	Y=1	Y=2
X=0	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$
X=1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
X=2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$



Nous complétons le tableau précédent par les sommes de chaque ligne et de chaque colonne

	Y=0	Y=1	Y=2	
X=0	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$p_X(0)=\frac{1}{4}$
X=1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$p_X(1)=\frac{1}{2}$
X=2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$p_X(2)=\frac{1}{4}$
	$p_Y(0)=\frac{1}{4}$	$p_Y(1)=\frac{1}{2}$	$p_Y(2)=\frac{1}{4}$	

$p_X(0)=p((X=0,\,-))$ désigne la probabilité que $X=0$, la variable Y prenant toutes les valeurs possibles.

$p_X(0)=p((X=0,\,-))$ est égale à la somme de la première ligne du tableau :

$$p_X(0)=p((X=0,\,-))=p((X=0,\,Y=0))+p((X=0,\,Y=1))+p((X=0,\,Y=2))$$
$$=\frac{1}{16}+\frac{1}{8}+\frac{1}{16}=\frac{1}{4}$$

$p_Y(1)=p((-,\,Y=1))$ désigne la probabilité que $Y=1$, la variable X prenant toutes les valeurs possibles.

$p_Y(1)=p((-,\,Y=1))$ est égale à la somme de la deuxième colonne du tableau :

$$p_Y(1)=p((-,\,Y=1))=p((X=0,\,Y=1))+p((X=1,\,Y=1))+p((X=2,\,Y=1))$$
$$=\frac{1}{8}+\frac{1}{4}+\frac{1}{8}=\frac{1}{2}$$

Les nombres $p_X(i)$ sont les probabilités de la **distribution marginale** X, c'est-à-dire qui se rapportent à la première composante de la variable aléatoire à deux dimensions $Z = (X = i, -)$, la variable Y prenant toutes les valeurs possibles.

Les nombres $p_Y(j)$ sont les probabilités de la distribution marginale Y, c'est-à-dire qui se rapportent à la deuxième composante de la variable aléatoire à deux dimensions $Z = (-, Y = j)$, la variable X prenant toutes les valeurs possibles.

A quoi correspond le produit des probabilités marginales ?

$$p_X(0)\, p_Y(0) = \frac{1}{4}\,\frac{1}{4} = \frac{1}{16} = p((X = 0, Y = 0))$$

$$p_X(0)\, p_Y(1) = \frac{1}{4}\,\frac{1}{2} = \frac{1}{8} = p((X = 0, Y = 1))$$

etc.

Dans le tableau des probabilités, la propriété s'exprime comme suit:

pour des variables indépendantes,

la probabilité d'un événement est égale au produit des probabilités marginales:

	$Y=0$	$Y=1$	$Y=2$
$X=0$	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}$
$X=1$	$\frac{1}{2}\frac{1}{4}$	$\frac{1}{2}\frac{1}{2}$	$\frac{1}{2}\frac{1}{4}$
$X=2$	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}$

On a la propriété caractéristique suivante:

$$X, Y \text{ indépendantes} \Longleftrightarrow \text{ pour tous les couples } (i, j)$$
$$p((X = i, Y = j)) = p_X(i)\, p_Y(j)$$

Exemple de deux variables non indépendantes

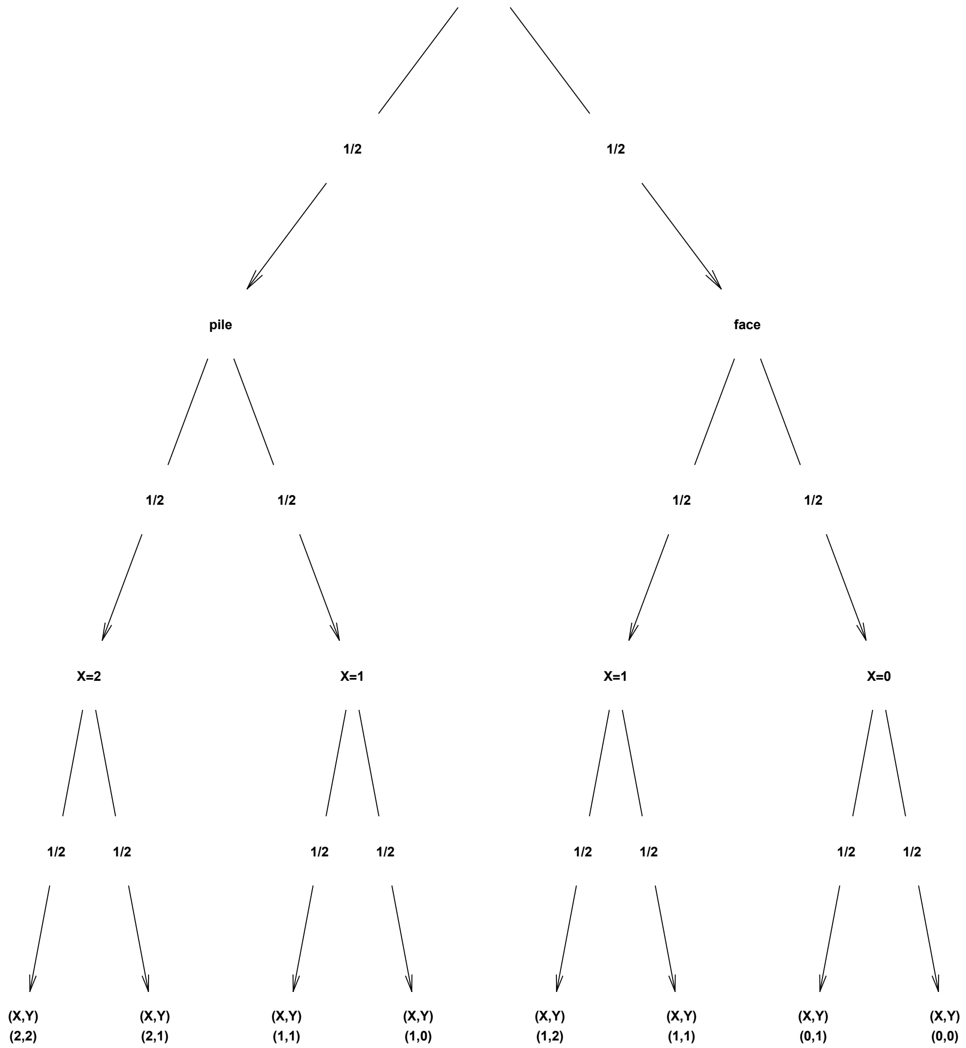
On lance **trois** fois une pièce de monnaie. On note

X = nombre de piles obtenus lors des deux premiers lancers,

Y = nombre de piles obtenus lors des deux derniers lancers.

Intuitivement, les valeurs de Y sont "influencées" par les résultats de X.
Par exemple, si $X = 0$, il est impossible que $Y = 2$;
ou encore, si $X = 2$, il est impossible que $Y = 0$.
Les variables X et Y ne sont pas indépendantes.

Nous aimerions caractériser la situation par une relation mathématique. Représentons les différentes issues possibles pendant le déroulement des trois lancers de la pièce de monnaie. A chaque lancer, il y a deux possibilités qu'on peut représenter par un embranchement.



Chaque événement situé à l'extrémité d'une branche (tout en bas) a une 1 chance sur 8 de se produire. Donc

$$p((X=2, Y=2)) = \frac{1}{8}$$
$$p((X=2, Y=1)) = \frac{1}{8}$$
$$p((X=2, Y=0)) = 0$$
$$p((X=1, Y=2)) = \frac{1}{8}$$
$$p((X=1, Y=1)) = \frac{2}{8}$$
$$p((X=1, Y=0)) = \frac{1}{8}$$
$$p((X=0, Y=2)) = 0$$
$$p((X=0, Y=1)) = \frac{1}{8}$$
$$p((X=0, Y=0)) = \frac{1}{8}$$

Avec les résultats précédents, dressons un tableau des probabilités

	Y=0	Y=1	Y=2
X=0	$\frac{1}{8}$	$\frac{1}{8}$	0
X=1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
X=2	0	$\frac{1}{8}$	$\frac{1}{8}$

Complétons le tableau précédent par les sommes de chaque ligne et de chaque colonne

	Y=0	Y=1	Y=2	
X=0	$\frac{1}{8}$	$\frac{1}{8}$	0	$p_X(0)=\frac{1}{4}$
X=1	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$p_X(1)=\frac{1}{2}$
X=2	0	$\frac{1}{8}$	$\frac{1}{8}$	$p_X(2)=\frac{1}{4}$
	$p_Y(0)=\frac{1}{4}$	$p_Y(1)=\frac{1}{2}$	$p_Y(2)=\frac{1}{4}$	

$p_X(0) = p((X = 0, -))$ désigne la probabilité que $X = 0$, la variable Y prenant toutes les valeurs possibles.

$p_X(0) = p((X = 0, -))$ est égale à la somme de la première ligne du tableau :

$$\begin{aligned} p_X(0) &= p((X=0,-)) = p((X=0, Y=0)) + p((X=0, Y=1)) + p((X=0, Y=2)) \\ &= \frac{1}{8} + \frac{1}{8} + 0 = \frac{1}{4} \end{aligned}$$

$p_y(1) = p((-, Y = 1))$ désigne la probabilité que $Y = 1$, la variable X prenant toutes les valeurs possibles.

$p_Y(1) = p((-, Y = 1))$ est égale à la somme de la deuxième colonne du tableau :

$$\begin{aligned} p_Y(1) &= p((-, Y=1)) = p((X=0, Y=1)) + p((X=1, Y=1)) + p((X=2, Y=1)) \\ &= \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2} \end{aligned}$$

Les nombres $p_X(i)$ sont les probabilités de la **distribution marginale** X, c'est-à-dire qui se rapportent à la première composante de la variable aléatoire à deux dimensions $Z = (X = i, -)$, la variable Y prenant toutes les valeurs possibles.

Les nombres $p_Y(j)$ sont les probabilités de la distribution marginale Y, c'est-à-dire qui se rapportent à la deuxième composante de la variable aléatoire à deux dimensions $Z = (-, Y = j)$, la variable X prenant toutes les valeurs possibles.

Le tableau des probabilités <u>ne</u> possède <u>pas</u> la propriété

la probabilité d'un événement est égale au produit des probabilités marginales

$$p_X(0)\, p_Y(0) = \frac{1}{4}\,\frac{1}{4} = \frac{1}{16} \neq \frac{1}{8} = p((X=0, Y=0))$$

$$p_X(0)\, p_Y(2) = \frac{1}{4}\,\frac{1}{4} = \frac{1}{16} \neq 0 = p((X=0, Y=2))$$

$$\ldots$$

Le tableau des produits des probabilités marginales <u>diffère</u> du tableau des probabilités:

	Y=0	Y=1	Y=2
X=0	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}$
X=1	$\frac{1}{2}\frac{1}{4}$	$\frac{1}{2}\frac{1}{2}$	$\frac{1}{2}\frac{1}{4}$
X=2	$\frac{1}{4}\frac{1}{4}$	$\frac{1}{4}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}$

Le fait que les deux variables ne soient pas indépendantes se traduit par la propriété caractéristique suivante:

$$p((X = i, Y = j)) \neq p_X(i)\, p_Y(j) \quad \text{pour au moins un couple } (i, j)$$

Définition de deux variables indépendantes

Dans le cas de deux variables discrètes *X*, *Y*, on définit les **probabilités marginales** comme suit:

$$\boxed{\begin{array}{l} p_X(x_i) = p((X = x_i, -)) = \sum_k p((X = x_i, Y = y_k)) \\ p_Y(y_j) = p((-, Y = y_j)) = \sum_r p((X = x_r, Y = y_j)) \end{array}}$$

Le fait que les deux variables *X* et *Y* soient ***indépendantes*** se traduit par la propriété caractéristique suivante:

la probabilité d'un événement est égale au produit des probabilités marginales

$$\boxed{\begin{array}{c} X, Y \text{ indépendantes} \iff \text{pour tous les couples } (i, j) \\ p((X = x_i, Y = y_j)) = p_X(x_i)\, p_Y(y_j) \end{array}}$$

En d'autres termes,

$$p((X = x_i, Y = y_j)) = \left(\sum_k p((X = x_i, Y = y_k))\right)\left(\sum_r p((X = x_r, Y = y_j))\right)$$

§ 2.2 *E*(*X Y*) = *E*(*X*) *E*(*Y*)

Exemple

On lance **quatre** fois une pièce de monnaie. On note

X = nombre de piles obtenus lors des deux premiers lancers,

Y = nombre de piles obtenus lors des deux derniers lancers.

Avec les valeurs calculées dans le § 2.1, on a

$$E(Y) = E(X) = 0 \cdot p_X(0) + 1 \cdot p_X(1) + 2 \cdot p_X(2)$$
$$= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

$$E(XY) = 0 \cdot 0 \cdot p((X = 0, Y = 0)) + 0 \cdot 1 \cdot p((X = 0, Y = 1)) + 1 \cdot 0 \cdot p((X = 1, Y = 0))$$
$$+ 1 \cdot 1 \cdot p((X = 1, Y = 1)) + 0 \cdot 2 \cdot p((X = 0, Y = 2)) + 2 \cdot 0 \cdot p((X = 2, Y = 0))$$
$$+ 1 \cdot 2 \cdot p((X = 1, Y = 2)) + 2 \cdot 1 \cdot p((X = 2, Y = 1)) + 2 \cdot 2 \cdot p((X = 2, Y = 2))$$
$$= 0 + 0 + 0 + 1 \cdot 1 \cdot \frac{1}{4} + 0 + 0 + 1 \cdot 2 \cdot \frac{1}{8} + 2 \cdot 1 \cdot \frac{1}{8} + 2 \cdot 2 \cdot \frac{1}{16} = 1$$

On a

$$E(XY) = 1$$
$$E(X)\,E(Y) = 1 \cdot 1 = 1$$

Cas général

Proposition

$$\boxed{X,\ Y \text{ indépendantes} \quad \Longrightarrow \quad E(XY) = E(X)\,E(Y)}$$

En mots :

> Pour des variables aléatoires indépendantes,
> l' espérance du produit est égale au produit des espérances

Démonstration pour des variables aléatoires discrètes

$$E(XY) = \sum_i \sum_j x_i\, y_j\, p((X = x_i, Y = y_j)) = \sum_i \sum_j x_i\, y_j\, p_X(x_i)\, p_Y(y_j) =$$
$$\sum_i x_i\, p_X(x_i) \left(\sum_j y_j\, p_Y(y_j)\right) = \left(\sum_i x_i\, p_X(x_i)\right) \left(\sum_j y_j\, p_Y(y_j)\right) = E(X)\,E(Y)$$

La proposition est valide pour des variables aléatoires quelconques, en particulier continues.

§ 2.3 $\quad V(X + Y) = V(X) + V(Y)$

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

Exemple

On lance **quatre** fois une pièce de monnaie. On note

X = nombre de piles obtenus lors des deux premiers lancers,

Y = nombre de piles obtenus lors des deux derniers lancers.

Avec les valeurs calculées dans les paragraphes 2.1 et 2.2, on a

$$E(Y) = E(X) = 1$$
$$E(X + Y) = E(X) + E(Y) = 2$$
$$V(Y) = V(X) = (0-1)^2 \frac{1}{4} + (1-1)^2 \frac{1}{8} + (2-1)^2 \frac{1}{4} = \frac{1}{2}$$
$$V(X+Y) = (0+0-2)^2\, p((X = 0, Y = 0)) +$$
$$(0+1-2)^2\, p((X = 0, Y = 1)) + (1+0-2)^2\, p((X = 1, Y = 0)) +$$
$$(1+1-2)^2\, p((X = 1, Y = 1)) + (0+2-2)^2\, p((X = 0, Y = 2)) +$$
$$(2+0-2)^2\, p((X = 2, Y = 0)) + (1+2-2)^2\, p((X = 1, Y = 2)) +$$
$$(2+1-2)^2\, p((X = 2, Y = 1)) + (2+2-2)^2\, p((X = 2, Y = 2))$$
$$= 4 \times \frac{1}{16} + 1\,\frac{1}{8} + 1\,\frac{1}{8} + 0\,\frac{1}{4} + 0\,\frac{1}{16} + 0 \times \frac{1}{16} + 1 \times \frac{1}{8} + 1\,\frac{1}{8} + 4 \times \frac{1}{16} = 1$$

Pour deux variables indépendantes, on a

$$V(X + Y) = V(X) + V(Y)$$

Cas général

Démonstration : faisons appel aux règles connues

$$V(X+Y) = E\left((X+Y)^2\right) - (E(X+Y))^2$$
$$= E\left(X^2 + 2XY + Y^2\right) - (E(X) + E(Y))^2$$

$$= E\left(X^2\right) + 2E(XY) + E\left(Y^2\right) - E^2(X) - 2E(X)E(Y) - E^2(Y)$$
$$= \left(E\left(X^2\right) - E^2(X)\right) + \left(E\left(Y^2\right) - E^2(Y)\right) + 2\left(E(XY) - E(X)E(Y)\right)$$
$$= V(X) + V(Y) + 2\cdot 0 = V(X) + V(Y)$$

On a démontré la proposition

$$\boxed{X,\ Y \text{ indépendantes} \implies V(X+Y) = V(X) + V(Y)}$$

En mots :

Pour des variables aléatoires indépendantes, la variance de la somme est égale à la somme des variances

En d'autres termes,

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

donc, pour l'écart-type de la somme,

$$\boxed{X,\ Y \text{ indépendantes} \implies \sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}}$$

Comparaison des variances empirique $s^2(x+y)$ et $s^2(x) + s^2(y)$

Partons de variables aléatoires indépendantes. Par exemple, pour les variables aléatoires

X =lancer d'une pièce de monnaie (convention 0 = pile, 1 = face);

Y =lancer d'un dé;

$Z = (X, Y)$ = lancer une pièce de monnaie et un dé (variable de dimension 2) ;

on a obtenu, en répétant 10 épreuves indépendantes de la variable Z, les échantillons suivants

```
piece = BernoulliDistribution[1/2];
        distribution de Bernoulli
```

```
de = DiscreteUniformDistribution[{1, 6}];
     distribution uniforme discrète
```

```
n = 10;
```

```
x = RandomInteger[piece, n]
    entier aléatoire
```

```
{0, 0, 0, 0, 0, 1, 0, 0, 0, 0}
```

```
y = RandomInteger[de, n]
    entier aléatoire
```

```
{3, 4, 2, 6, 3, 2, 1, 3, 2, 1}
```

```
N[VarianceMLE[x + y]]
valeur numérique
```

```
1.96
```

```
N[VarianceMLE[x] + VarianceMLE[y]]
valeur numérique
```

```
2.1
```

Ce que les variables aléatoires réalisent exactement dans un modèle théorique idéal

$$V(X+Y) = V(X) + V(Y)$$

les échantillons empiriques de variables indépendantes ne le réalisent qu'approximativement

$$s^2(x+y) \approx s^2(x) + s^2(y)$$

Les échantillons de grande taille ont un comportement plus proche du modèle théorique

```
n = 10 000;

x = RandomInteger[piece, n];
    entier aléatoire

y = RandomInteger[de, n];
    entier aléatoire

N[VarianceMLE[x + y]]
valeur numérique

3.16709

N[VarianceMLE[x] + VarianceMLE[y]]
valeur numérique

3.19019
```

Exercice 2.3 - 1

a) Soient X et Y deux variables aléatoires continues uniformément distribuées sur l'intervalle $[0,\ 1[$.
Par simulation, formez deux échantillons de même taille n :
x pour X, y pour Y. Calculez et comparez

$$m(x\,y),\quad m(x)\,m(y)$$

ainsi que

$$v(x+y),\ v(x)+v(y)$$

b) Soit (X, Y) une paire de variables aléatoires définie comme suit :
X est continue uniformément distribuée sur l'intervalle $[0,\ 1[$

$$\text{et } Y = \begin{cases} \text{unif. distr. sur } [0;\ 0.5[& \text{si } X < \frac{1}{2} \\ \text{unif. distr. sur } \left[\frac{1}{2},\ 1\right[& \text{si } X \geq \frac{1}{2} \end{cases}.$$

Par simulation, formez un échantillon $\{x, y\} = \{(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots\}$ de taille n.
Calculez et comparez

$$m(x\,y),\ m(x)\,m(y)$$

ainsi que

$$v(x+y),\ v(x)+v(y)$$

c) Expliquez pourquoi on obtient des quantités approximativement égales dans la situation a) et des quantités différentes dans la situation b).

Exercice 2.3 - 2

Soient X_1 et X_2 deux variables aléatoires indépendantes.
La distribution de X_1 est normale, d'espérance $\mu_1 = 0$, d'écart-type $\sigma_1 = 3$.
La distribution de X_2 est normale, d'espérance $\mu_2 = 0$, d'écart-type $\sigma_2 = 4$.
Au moyen d'une simulation, estimez la moyenne m et l'écart-type s de la somme $Y = X_1 + X_2$.
Comparez les résultats de la simulation avec les paramètres théoriques μ et σ.

Exercice 2.3 - 3

Soient X_1, X_2, ..., X_n des variables de Bernoulli indépendantes de même probabilité p
et
$S = X_1 + X_2 + ... + X_n$ la variable binomiale *(n, p)*.
Calculez l'écart-type théorique σ_S.

§ 2.4 Espérance et écart-type de la moyenne

Exemple introductif

On lance une pièce de monnaie et on compte le nombre de "face".
Dans le modèle théorique, notons

X le nombre de "face" en un lancer, à savoir : $X = \begin{cases} 0 & \text{avec prob. } \frac{1}{2} \\ 1 & \text{avec prob. } \frac{1}{2} \end{cases}$,

M_{10} le nombre moyen de "face" pour 10 lancers;

M_{1000} le nombre moyen de "face" pour 1000 lancers.

Nous allons comparer les valeurs théoriques à celles obtenues par des simulations.
Pour générer une moyenne de n lancers, on divise par n la variable binomiale $\left(n, \frac{1}{2}\right)$.
Une expérience consistant à lancer n fois une pièce, générons un échantillon de k expériences:

```
som[n_] := BinomialDistribution[n, 1/2];
           distribution binomiale
```

```
echant[n_, k_] := N[1/n RandomInteger[som[n], k]];
                  va  entier aléatoire
```

Le nombre moyen de "face" en 10 lancers est une **variable aléatoire** M_{10} dont on peut considérer un échantillon de taille 50

```
e10 = echant[10, 50]
```

```
{0.5, 0.5, 0.5, 0.4, 0.4, 0.6, 0.6, 0.3, 0.3, 0.4, 0.4, 0.5, 0.7, 0.6, 0.4, 0.4, 0.6,
 0.4, 0.4, 0.5, 0.7, 0.4, 0.7, 0.3, 0.6, 0.5, 0.5, 0.1, 0.5, 0.5, 0.4, 0.7, 0.4, 0.5,
 0.7, 0.3, 0.5, 0.4, 0.6, 0.3, 0.4, 0.6, 0.5, 0.6, 0.3, 0.3, 0.6, 0.2, 0.8, 0.4}
```

```
m10 = Mean[e10]
      valeur moyen
```

```
0.474
```

```
s10 = StandardDeviationMLE[e10]
```

```
0.142562
```

Le nombre moyen de "face" en 1000 lancers est une variable aléatoire M_{1000} dont on peut considérer un échantillon de taille 50

```
e1000 = echant[1000, 50]
```

```
{0.497, 0.507, 0.503, 0.509, 0.511, 0.503, 0.5, 0.493, 0.49, 0.479, 0.519, 0.485, 0.531,
 0.533, 0.516, 0.493, 0.49, 0.55, 0.499, 0.489, 0.51, 0.524, 0.484, 0.466, 0.512, 0.498,
 0.514, 0.501, 0.513, 0.497, 0.492, 0.478, 0.481, 0.472, 0.509, 0.506, 0.501, 0.515,
 0.495, 0.508, 0.508, 0.53, 0.498, 0.514, 0.469, 0.495, 0.502, 0.505, 0.489, 0.494}
```

On remarquera que les moyennes sont moins fluctuantes avec 1000 lancers qu'avec 10. En d'autres termes, **l'écart-type de la moyenne** est plus petit pour 1000 lancers que pour 10 lancers.

```
m1000 = Mean[e1000]
        valeur moyenne
```

0.50154

```
s1000 = StandardDeviationMLE[e1000]
```

0.0164222

On pourra observer que

* les valeurs moyennes de X, M_{10} et M_{1000} sont comparables (environ $\frac{1}{2}$);
* l'écart-type avec 1000 lancers vaut environ $\frac{1}{10}$de l'écart-type avec 10 lancers.

Ce paragraphe se propose d'expliquer pourquoi. Nous montrerons que, pour une moyenne comptant 100 fois plus de termes,

$$\sigma_{1000} = \frac{1}{\sqrt{100}} \sigma_{10} \qquad \text{où} \qquad \sigma_{1000} = \sqrt{V(M_{1000})} \quad \text{et} \quad \sigma_{10} = \sqrt{V(M_{10})}$$

Espérance et écart-type de la moyenne

Considérons la moyenne M de n variables aléatoires indépendantes X_1, X_2, ..., X_n de même espérance $\mu = E(X_1) =. .. = E(X_n)$ et de même écart-type $\sigma = \sqrt{V(X_1)} =. .. = \sqrt{V(X_n)}$

$$\boxed{M = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)}$$

M est une variable aléatoire. L'espérance mathématique de M est calculée avec les propriétés connues

$$E(M) =$$
$$E\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) = \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n)) = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \mu$$

La variance de M est calculée avec les propriétés précédemment établies

$$V(M) = V\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right) =$$
$$\left(\frac{1}{n}\right)^2 (V(X_1) + V(X_2) + \ldots + V(X_n)) = \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) = \frac{1}{n^2} n\, \sigma^2 = \frac{\sigma^2}{n}$$

L'écart-type théorique de M est

$$\sigma_M = S(M) = \sqrt{V(M)} = \frac{\sigma}{\sqrt{n}}$$

Retenons les formules suivantes qui sont valables, en particulier, pour des **épreuves répétées indépendantes** (μ et σ désignent l'espérance et l'écart-type d'une épreuve, n le nombre d'épreuves et M la variable aléatoire "moyenne des n épreuves"):

$$\boxed{\begin{array}{l} E(M) = \mu \\ V(M) = \dfrac{\sigma^2}{n} \end{array}}$$

En d'autres termes, pour l'écart-type de la moyenne $\sigma_M = \sqrt{V(M)}$, on a

$$\boxed{\sigma_M = \frac{\sigma}{\sqrt{n}}}$$

Retour à l'exemple introductif

Nous pouvons maintenant expliquer pourquoi l'écart-type avec 1000 lancers vaut environ $\frac{1}{10}$de l'écart-type avec 10 lancers

$$\sigma_{1000} = \frac{\sigma}{\sqrt{1000}} = \frac{\sigma}{10\sqrt{10}} = \frac{1}{10}\,\sigma_{10}$$

$$\text{où} \quad \sigma_{1000} = \sqrt{V(M_{1000})} \quad \text{et} \quad \sigma_{10} = \sqrt{V(M_{10})}$$

Plus généralement, pour gagner un chiffre caractéristique sur la précision de la moyenne (c'est-à-dire pour réduire l'écart-type d'un facteur $\frac{1}{10}$), il faut prendre un échantillon de taille non pas 10 fois plus grande mais **100 fois** plus grande !

Exemple

Lançons n fois une pièce de monnaie. Le jet numéro j est modélisé par la variable aléatoire X_j

$X_j = 0$ (pile) avec probabilité $\frac{1}{2}$

$X_j = 1$ (face) avec probabilité $\frac{1}{2}$

dont l'espérance mathématique est

$$\mu = E(X_k) = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2}$$

et dont l'écart-type théorique est

$$\sigma = \sqrt{\left(0 - \left(\frac{1}{2}\right)\right)^2 \frac{1}{2} + \left(1 - \left(\frac{1}{2}\right)\right)^2 \frac{1}{2}} = \frac{1}{2}$$

Ces n variables sont indépendantes.

La moyenne M des n variables représente le nombre moyen de "face" obtenu avec les n jets.
M est une variable aléatoire qui vérifie

$$E(M) = \mu = \frac{1}{2}$$

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{1}{2\sqrt{n}}$$

§ 2.5 Estimateurs non biaisés

Exemple introductif

On lance trois fois une pièce de monnaie et on compte le nombre moyen de "face".
A partir d'expériences, on veut estimer la moyenne et l'écart-type d'un lancer.

```
piece = BernoulliDistribution[1/2];
        distribution de Bernoulli
```

Une observation donne, par exemple,

```
obs = RandomInteger[piece, 3]
      entier aléatoire
```

```
{0, 1, 1}
```

```
m3 = Mean[obs]
     valeur moyen
```

$\frac{2}{3}$

```
v3 = VarianceMLE[obs]
```

$\frac{2}{9}$

Etant donné que les résultats précédents sont très fluctuants, organisons une série d'observations et faisons leur moyenne:

```
n = 1000;
```

```
obs = Table[RandomInteger[piece, 3], {n}];
      table entier aléatoire
```

```
mObs = Map[Mean, obs];
       app· valeur moyenne
m3m = N[Mean[mObs]]
      ·· valeur moyenne
```

```
0.494667
```

```
vObs = Map[VarianceMLE, obs];
       applique
v3m = N[Mean[vObs]]
      ·· valeur moyenne
```

```
0.163778
```

Comparons avec les valeurs théoriques. Nous partons de trois variables aléatoires indépendantes :

$X_j = \begin{cases} 0 & \text{avec prob. } \frac{1}{2} \\ 1 & \text{avec prob. } \frac{1}{2} \end{cases}$, $j = 1, 2, 3$, dont nous formons la moyenne $M_3 = \frac{1}{3}(X_1 + X_2 + X_3)$.

Nous nous demandons si les paramètres empiriques de M_3 constituent une estimation raisonnable des paramètres théoriques de X_j. Plus précisément, nous allons comparer

* la moyenne empirique de trois observations $m_3 = m(M_3)$ et l'espérance d'un jet $\mu = E(X_j)$;
* l'écart-type empirique de trois observations $v_3 = v(M_3)$ et l'écart-type théorique d'un jet $\sigma^2 = V(X_j)$.

$$\mu = E(X_j) = \frac{1}{2}$$

$$\sigma^2 = V(X_j) = \mu(1-\mu) = \frac{1}{4} = 0.25 \qquad \text{(voir l' exercice 3.1 .1 de Statistique I} - \S 3\text{)}$$

La comparaison de m_3 et μ donne bien le résultat attendu :

$$m_3 \approx \mu$$

Par contre, v_3 et σ^2 sont très différents. La raison est qu'il faudrait calculer non pas

$$v = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

où *m* est la moyenne empirique, mais plutôt

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

où *μ* est l'espérance mathématique. Vérifions-le sur l'exemple précédent:

```
vCor[{x1_, x2_, x3_}] := 1/3 ((x1 - 1/2)^2 + (x2 - 1/2)^2 + (x3 - 1/2)^2);
vObsCor = Map[vCor, obs];
          applique
v3Cor = N[Mean[vObsCor]]
        ·· valeur moyenne
0.25
```

Nous avons démontré, dans l'exercice 4.2-1 de Statistique I - § 4, que la valeur de *t* pour laquelle la fonction suivante est minimale

$$t \mapsto \frac{1}{n} \sum_{i=1}^{n} (x_i - t)^2$$

est la moyenne arithmétique $t = m$. C'est pourquoi si, dans l'expression $\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$, on remplace *μ* par *m*, on obtient une valeur trop petite. On dit que s^2 est un estimateur biaisé de σ^2.

Dans ce paragraphe, nous montrerons qu'une estimation sans biais de σ^2 est $\frac{3}{2} s^2$; numériquement,

```
3/2 v3m
0.245667
```

Hypothèses

On considère maintenant un échantillon donné de taille *n*

$$x_1, x_2, \ldots, x_n$$

auquel correspondent, dans le modèle théorique, *n* variables aléatoires **indépendantes identiquement distribuées**

$$X_1, X_2, \ldots, X_n$$

de même espérance *μ* et de même écart-type *σ*.

Remarquez que, dans le cas où l'on extrait d'une population finie un échantillon non exhaustif par un tirage sans remise, notre hypothèse n'est pas vérifiée.

Estimateur de l'espérance mathématique

A la moyenne empirique

$$m = m(x) = \frac{x_1 + x_2 \ldots + x_n}{n}$$

correspond, dans le modèle théorique, la variable aléatoire

$$M = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

D'une part, puisque les *n* variables sont indépendantes et identiquement distribuées,

$$E(M) = E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \ldots + E(X_n)}{n} = \frac{n\,\mu}{n} = \mu$$

D'autre part, m étant une réalisation de la variable aléatoire M, on a

$$E(m) = E(M)$$

ce qui montre que m est un estimateur sans biais de l'espérance mathématique de M. Nous notons

$$E\hat{(M)} = m$$

Calculons $E(M)$. D'après les règles de calcul

$$E(M) = \mu = E(X_k) \qquad \text{pour } k = 1, 2, \ldots, n$$

Finalement, **m est un estimateur non biaisé de l'espérance mathématique de chaque épreuve X_k**:

$$\boxed{\hat{\mu} = m}$$

Estimateur sans biais de l'écart-type théorique

A la variance empirique

$$s^2 = s^2(x) = m\left(x^2\right) - (m(x))^2$$

correspond la variable aléatoire

$$Z = \frac{X_1{}^2 + X_2{}^2 + \ldots + X_n{}^2}{n} - M^2$$

s^2 est une réalisation de la variable aléatoire Z. Puisque $E\left(s^2\right) = E(Z)$, s^2 est un estimateur sans biais de l'espérance mathématique de Z, ce qui se note

$$E\hat{(Z)} = s^2$$

Calculons $E(Z)$. D'après les propriétés du § 3.4 de Statistique I - § 3, on a

$$E(Z) = E\left(\frac{X_1{}^2 + X_2{}^2 + \ldots + X_n{}^2}{n} - M^2\right)$$
$$= \frac{1}{n}\left(E\left(X_1{}^2\right) + E\left(X_2{}^2\right) + \ldots + E\left(X_1{}^2\right)\right) - E\left(M^2\right)$$

Dans l'expression précédente, déterminons $E\left(X_i{}^2\right)$ et $E\left(M^2\right)$

$$V(X_i) = E\left(X_i{}^2\right) - (E(X_i))^2 \Longrightarrow E\left(X_i{}^2\right) = V(X_i) + (E(X_i))^2 = \sigma^2 + \mu^2$$

$$V(M) = E\left(M^2\right) - (E(M))^2 \Longrightarrow E\left(M^2\right) = V(M) + (E(M))^2 = \frac{\sigma^2}{n} + \mu^2$$

puis effectuons la substitution

$$E(Z) = \frac{1}{n}\left(E\left(X_1{}^2\right) + E\left(X_2{}^2\right) + \ldots + E\left(X_1{}^2\right)\right) - E\left(M^2\right)$$
$$= \frac{1}{n}\, n\left(\sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\,\sigma^2$$

On dit que s^2 est un estimateur biaisé de σ^2 parce que

à s^2 correspond la variable aléatoire Z dont l'espérance mathématique diffère de σ^2.

Pour obtenir un estimateur sans biais, on utilise la relation

$$\sigma^2 = \frac{n}{n-1}\,E(Z)$$

Finalement, $\frac{n}{n-1}$ **s^2 est un estimateur non biaisé de la variance théorique de chaque épreuve X_i**:

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

Par suite, $\sqrt{\frac{n}{n-1}}$ **s est un estimateur non biaisé de l'écart-type théorique de chaque épreuve X_i**:

$$\boxed{\hat{\sigma} = \sqrt{\frac{n}{n-1}}\, s}$$

C'est ainsi que se justifie la formule de l'écart-type corrigé que l'on utilise pour les échantillons de petite taille

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}}\, s = \sqrt{\frac{n}{n-1}} \sqrt{\frac{(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2}{n}} =$$

$$\sqrt{\frac{(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2}{n-1}}$$

La correction consiste à remplacer n par $(n - 1)$:

$$\boxed{\hat{\sigma} = \sqrt{\frac{(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2}{n-1}}}$$

Avec *Mathematica*, la fonction à utiliser n'est pas **StandardDeviationMLE** mais

```
? StandardDeviation
```

StandardDeviation[*list*] gives the sample standard deviation of the elements in *list*.
StandardDeviation[*dist*] gives the standard deviation of the symbolic distribution *dist*. >>

On remarquera que, pour des échantillons de grande taille, la correction est négligeable. Ainsi, pour $n = 100$, on a

$$\sqrt{\frac{n}{n-1}} \simeq 1.005$$

Reprenons l'exemple introductif:

```
vCorObs = Map[Variance, obs];
          app· variance
vCor3m = N[Mean[vCorObs]]
         ·· valeur moyenne
```

```
0.245667
```

Exercice 2.5 - 2

Pour $n = 5$, par une simulation, estimez $\hat{\sigma}_n$ = écart-type de la distribution binomiale $\left(5, \frac{1}{2}\right)$.

Comparez avec l'écart-type théorique $\sigma_n = \sqrt{n\,p\,(1 - p)}$ (selon l'exercice 2.3 - 3).

§ 2.6 Central limit theorem

Exercice 2-6-1 Exercice de révision

Dans Statistique I, § 4, voir l'exercice 4.2-3
https://www.deleze.name/marcel/sec2/applmaths/csud/statistique_1/4-stat_I.pdf

a) Considérons maintenant le nombre moyen de six obtenus en lançant 100 fois le dé. Cette variable aléatoire, désignée par M_{100}, est une distribution binomiale. *[Avec Mathematica]* Tracez son diagramme à bâtons. Comparez avec la distribution normale.

Remarque : En général, on ne peut pas superposer un diagramme de fréquences et un diagramme de densité sauf dans le cas particulier où tous les bâtons peuvent être interprétés comme des barres de largeur 1.

b) Effectuons maintenant une simulation de M_{100}. *[Avec Mathematica]* On demande de
- construire un échantillon de taille n;
- grouper les données en les traitant comme une variable continue;
- superposer l'histogramme et la distribution normale.

Problématique

Nous savons déjà que, si les variables que l'on somme sont de Bernoulli, indépendantes et identiquement distribuées, alors leur somme est une variable binomiale S, dont la moyenne est $M = \frac{1}{n} S$, qui est donc voisine d'une distribution normale. Qu'en est-il si la distribution des variables de départ n'est pas de Bernoulli ?

Considérons la moyenne M de n variables aléatoires identiquement distribuées, d'espérance mathématique μ et d'écart-type σ. Nous supposons que les variables obéissent à une distribution quelconque et que n est une constante entière pas trop petite.

Nous savons déjà que l'espérance mathématique de M est μ et que l'écart-type de M est $\sigma_M = \frac{\sigma}{\sqrt{n}}$.

Mais que peut-on dire de la **distribution** de M ? Il pourrait d'abord nous sembler que la distribution de M dépende de la distribution des variables participant à la moyenne. Fort étonnamment, pour de grandes valeurs de n, la distribution de M est voisine d'une distribution **normale** quelle que soit la distribution des variables dont on fait la moyenne.

Comme exemple, considérons la moyenne de n variables continues identiquement distribuées.

Distribution continue uniforme

```
distr = UniformDistribution[{0, 1}];
        distribution uniforme
```

La moyenne de la distribution uniforme est la valeur centrale:

```
μ = Mean[distr]
    valeur moyenne
```

$\frac{1}{2}$

Comme nous l'avons étudié dans le § 4.1 de Statistique I - § 4,

```
σ = StandardDeviation[distr]
    écart-type
```

$$\frac{1}{2\sqrt{3}}$$

```
N[σ]
valeur numérique
```

0.288675

Moyenne de 7 variables identiquement distribuées

```
M[] := Mean[RandomReal[distr, 7]]
       vale··· nombre réel aléatoire
```

```
Table[M[], {20}]
table
```

{0.551994, 0.42331, 0.466532, 0.615684, 0.537181, 0.483763, 0.369312, 0.603429, 0.560432, 0.425295, 0.612552, 0.446016, 0.599655, 0.578679, 0.347855, 0.629883, 0.635177, 0.59177, 0.492976, 0.329136}

La moyenne est une variable aléatoire dont on peut calculer la moyenne et l'écart-type (moyenne des moyennes, écart-type des moyennes):

```
n = 4000;
```

```
x = Table[M[], {n}];
    table
```

```
mm = Mean[x]
     valeur moy
```

0.496072

```
sm = StandardDeviationMLE[x]
```

0.111161

Comparaison avec la valeur théorique

$$\sigma m = N\left[\frac{\sigma}{\sqrt{7}}\right]$$

valeur num

0.109109

Pour effectuer une description statistique, choisissons un intervalle proche de $[m - 3\,s,\ m + 3\,s]$ et partageons-le en 9 classes comme suit

$$u = \text{Range}\left[\frac{1}{2} - \frac{3}{10}, \frac{1}{2} + \frac{3}{10}, \frac{6}{10 * 9}\right]$$

plage

$$\left\{\frac{1}{5}, \frac{4}{15}, \frac{1}{3}, \frac{2}{5}, \frac{7}{15}, \frac{8}{15}, \frac{3}{5}, \frac{2}{3}, \frac{11}{15}, \frac{4}{5}\right\}$$

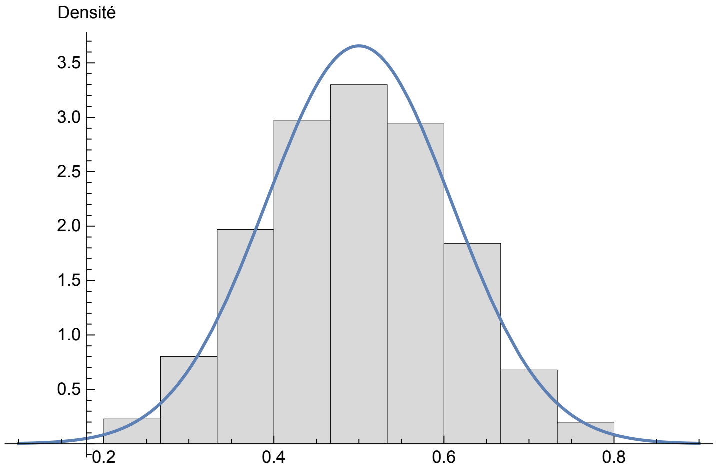
```
effectifs = BinCounts[x, {Join[u, {-Infinity, Infinity}]}];
            compte des huc··· joins    infini    infini
```

```
freq = Take[effectifs, {2, -2}] / n
```

$$\left\{\frac{61}{4000}, \frac{107}{2000}, \frac{21}{160}, \frac{793}{4000}, \frac{11}{50}, \frac{49}{250}, \frac{491}{4000}, \frac{181}{4000}, \frac{53}{4000}\right\}$$

```
norm = NormalDistribution[μ, σm];
       distribution normale
```

```
Show[histogramme[u, freq], Plot[PDF[norm, t], {t, 0.1, 0.9}],
montre                     trac··· fonction de densité de probabilité
 Ticks → Automatic, AxesLabel → {None, "Densité"}]
 graduat··· automatique titre d'axe aucun
```



Conclusion: On peut observer que la moyenne arithmétique de 7 variables est approximativement normale. Pourtant, la distribution des 7 variables de départ est très loin d'être normale puisqu'elle est uniformément distribuée sur [0, 1].

Central limit theorem

Le résultat précédent est généralisable : la distribution normale est une distribution particulière vers laquelle tendent toutes les moyennes. Le *central limit theorem* nous l'énonce plus précisément.

Hypothèses:

A partir de n variables aléatoires **indépendantes** identiquement distribuées $X_1, X_2, \ldots, X_n$, d'espérance $\mu = E(X_1) = .. = E(X_n)$ et d'écart-type $\sigma = \sqrt{V(X_1)} = .. = \sqrt{V(X_n)}$, on forme une nouvelle variable aléatoire $M = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$.

Conclusion:

Pour n tendant vers l'infini, la distribution de M tend, en probabilité, vers la distribution **normale** de moyenne μ et d'écart-type $\sigma_M = \frac{\sigma}{\sqrt{n}}$.

La démonstration du *central limit theorem* relève du niveau universitaire.

Application à la simulation : estimation de l'erreur

Rappel

En accord avec les résultats de Statistique I - § 4, la distribution normale de moyenne μ et d'écart-type σ_M possède les propriétés suivantes

l'intervalle $[\mu - \sigma_M, \mu + \sigma_M]$ contient environ 68.3 % des observations;

l'intervalle $[\mu - 2\,\sigma_M,\ \mu + 2\,\sigma_M]$ contient environ 95.5 % des observations;
l'intervalle $[\mu - 3\,\sigma_M,\ \mu + 3\,\sigma_M]$ contient environ 99.7 % des observations;

Conséquence:
Pour de grandes valeurs de n, la moyenne empirique de n tirages indépendants a une distribution approximativement normale. En notant m la moyenne empirique, s l'écart-type empirique et $\hat{\sigma}_M = \frac{s}{\sqrt{n-1}}$ l'estimation de σ_M, nous aurons approximativement

la probabilité que l'intervalle $\left[m - \hat{\sigma}_M,\ m + \hat{\sigma}_M\right]$ contienne μ est de 68.3 % environ;
la probabilité que l'intervalle $\left[m - 2\,\hat{\sigma}_M,\ m + 2\,\hat{\sigma}_M\right]$ contienne μ est de 95.5 % environ;
la probabilité que l'intervalle $\left[m - 3\,\hat{\sigma}_M,\ m + 3\,\hat{\sigma}_M\right]$ contienne μ est de 99.7 % environ.

Justification:
Dans la relation suivante, on peut échanger les rôles de m et μ.

$$\begin{aligned} m \in [\mu - \sigma_M,\ \mu + \sigma_M] &\Longleftrightarrow \mu - \sigma_M < m < \mu + \sigma_M \\ &\Longleftrightarrow -\sigma_M < m - \mu < \sigma_M \\ &\Longleftrightarrow \sigma_M > -m + \mu > -\sigma_M \\ &\Longleftrightarrow -\sigma_M < \mu - m < \sigma_M \\ &\Longleftrightarrow m - \sigma_M < \mu < m + \sigma_M \Longleftrightarrow \mu \in [m - \sigma_M,\ m + \sigma_M] \end{aligned}$$

σ_M est estimé par

$$\hat{\sigma}_M = s_M \sqrt{\frac{n}{n-1}} = \frac{s}{\sqrt{n}} \sqrt{\frac{n}{n-1}} = \frac{s}{\sqrt{n-1}}$$

Pour des valeurs de n pas trop petites, l'approximation $\sigma_M \simeq \hat{\sigma}_M$ est légitime. Par suite

$$m \in [\mu - \sigma_M,\ \mu + \sigma_M] \overset{\text{approx}}{\Longleftrightarrow} \mu \in \left[m - \hat{\sigma}_M,\ m + \hat{\sigma}_M\right]$$

Le *central limit theorem* nous fournit ainsi une estimation de l'erreur dont nous tirerons profit dans le chapitre *Simulation*.

Liens

Vers les corrigés des exercices :
https://www.deleze.name/marcel/sec2/applmaths/csud/corriges/statistique_2/2-stat_II-cor.pdf

Vers la page mère Applications des mathématiques
https://www.deleze.name/marcel/sec2/applmaths/csud/index.html