Statistiques et probabilités

I Vocabulaire :

Les statistiques s'appliquent à un ensemble **d'individus** (êtres vivants, choses ou autres comme des entreprises) qualifié de **population**.

Elles s'intéressent à un caractère particulier de cette population qui peut être un nombre, on parle alors de **caractère quantitatif**, ou bien une qualité (on parle alors de **caractère qualitatif**)

Lorsque le caractère est quantitatif et qu'il ne peut prendre qu'un nombre fini de valeurs ou bien un nombre discret, on dit qu'il est **quantitatif discret.**

Exemples :

Si on s'intéresse aux notes d'un élève sur une année, alors la population est l'ensemble des devoirs notés de cet élève et le caractère est la note obtenue. Ce dernier est quantitatif discret.

Si on s'intéresse au nombre de parties de poker réalisés au cours d'une vie par un groupe d'individus alors la population est le groupe d'individus, le caractère est le nombre de parties de poker joués dans la vie de l'individu, ce caractère étant quantitatif discret avec des valeurs possibles considérées comme étant l'ensemble des entiers naturels, même si un individu ne peut évidemment pas jouer un milliard de parties.

Si on s'intéresse à la durée de vie des smartphones d'un groupe d'individus, la population est ce groupe, le caractère est la durée de vie et ce caractère est quantitatif continu, les durées de vie a priori possibles étant les nombres réels positifs ou nuls, même si la durée de vie encore une fois ne peut pas être de 1 milliard d'années.

Si on s'intéresse aux couleurs de cheveux des élèves d'une classe, la population est l'ensemble des élèves, le caractère, la couleur de cheveu, ce caractère étant qualitatif, les valeurs étant par exemple, blond, brun, roux, châtain.

II Représentation des données statistiques dans le cas d'un caractère quantitatif

Les statistiques sont des données recueillies par une enquête. La manière la plus simple d'enregistrer ces données est par individu. L'enquête se résume alors à un **tableau détaillé par individu** :

Numéro de l'individu : $1,\ \ 2\ , \ldots.\ \ , n$

Caractère : $x_1,\ x_2, \ldots.\ \ x_n$

Toutefois, quand il y a un trop grand nombre d'individus, il est préférable de présenter les données sous forme d'un **tableau d'effectif** , où les caractères sont présentés dans l'ordre croissant :

Caractère : $x_1,\ x_2, \ldots.\ \ x_p$

Effectif : $n_1,\ n_2, \ldots.\ \ n_p$

Lorsque le caractère est quantitatif continu et que dans l'enquête, il y a beaucoup de valeurs différentes de caractère, il est préférable de condenser l'information sous forme de classes d'intervalles :

Caractère : $[x_1,\ x_2[, [x_2,\ x_3[\ldots.\quad [x_p,\ x_{p+1}[$

Effectif : $n_1 \quad n_2 \quad \ldots. \quad n_p$

On perd alors de l'information sur la répartition des individus dans chacune des classes.

III Moyenne et variance d'une série statistique à caractère discret

Moyenne :

La moyenne est la valeur que devrait avoir chaque individu de la population pour que la somme des caractères soit la même.

Par exemple, le salaire moyen dans une entreprise est le salaire que devrait avoir chaque employé de cette entreprise pour avoir la même masse salariale.

Formules mathématiques :

Série détaillée :

$$m = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Série à variable discrète donnée par tableau d'effectif :

$$m = \frac{x_1\, n_1 + x_2\, n_2 + \cdots + x_p\, n_p}{n} = \frac{\sum_{i=1}^{p} x_i\, n_i}{n}$$

Où :

$$n = \sum_{i=1}^{p} n_i$$

Série à variable continue donnée par tableau d'effectif :

$$m = \frac{c_1\, n_1 + c_2\, n_2 + \cdots + c_p\, n_p}{n} = \frac{\sum_{i=1}^{p} c_i\, n_i}{n}$$

Où :

$$n = \sum_{i=1}^{p} n_i$$

Et pour tout $i \in \{1,2, \ldots, p\}$:

$$c_i = \frac{x_i + x_{i+1}}{2}$$

Variance :

La variance est la moyenne des carrés des écarts à la moyenne

Formules mathématiques :

Série détaillée :

$$\boldsymbol{V = \frac{(x_1 - m)^2 + (x_2 - m)^2 + \cdots + (x_n - m)^2}{n} = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n}}$$

Remarque :

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

La variance est donc aussi égale à la moyenne des carrés moins le carré de la moyenne

Série à variable discrète donnée par tableau d'effectif :

$$\boldsymbol{V = \frac{(x_1 - m)^2\,n_1 + (x_2 - m)^2\,n_2 + \cdots + (x_p - m)^2\,n_p}{n} = \frac{\sum_{i=1}^{p}(x_i - m)^2\,n_i}{n} = \sum_{i=1}^{p}(x_i - m)^2\,f_i}$$
$$\boldsymbol{= \sum_{i=1}^{p}{x_i}^2\,f_i - m^2}$$

Où :

$$n = \sum_{i=1}^{p} n_i$$

Et pour tout $i \in \{1,2,\dots,p\}$:

$$f_i = \frac{n_i}{n}$$

Série à variable continue donnée par tableau d'effectif :

$$\boldsymbol{V = \frac{(c_1 - m)^2\,n_1 + (c_2 - m)^2\,n_2 + \cdots + (c_p - m)^2\,n_p}{n} = \frac{\sum_{i=1}^{p}(c_i - m)^2\,n_i}{n} = \sum_{i=1}^{p}(c_i - m)^2\,f_i}$$
$$\boldsymbol{= \sum_{i=1}^{p}{c_i}^2\,f_i - m^2}$$

Où :

$$n = \sum_{i=1}^{p} n_i$$

Et pour tout $i \in \{1,2,\dots,p\}$:

$$c_i = \frac{x_i + x_{i+1}}{2}$$

Ecart-type :

L'écart-type est la racine carrée de la variance :

$$\boldsymbol{\sigma = \sqrt{V}}$$

L'écart-type est un indicateur de dispersion. Il est distinct de l'écart moyen qui est, dans le cas d'une série détaillée :

$$\sigma_{moy} = \frac{|x_1 - m| + |x_2 - m| + \cdots + |x_n - m|}{n}$$

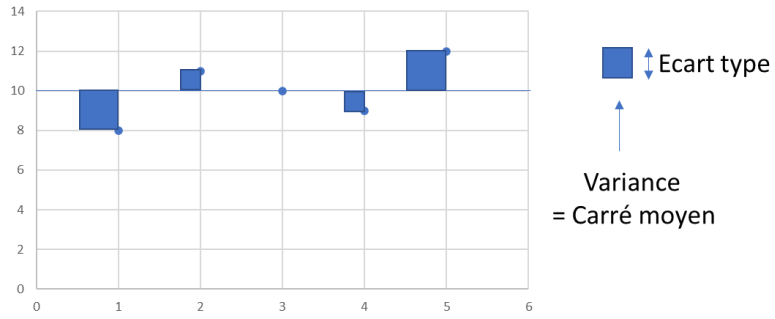
Illustrons les concepts de variance et d'écart-type par un exemple :

On donne les notes de deux élèves x et y à 5 devoirs de maths :

x : 8 11 10 9 12

y : 2 10 18 19 1

Les deux élèves ont 10 de moyenne. Pourtant, ils ont des profils différents. Le premier a des notes proches de la moyenne donc peu dispersées et le second a des notes très dispersées. Représentons les notes du premier élève sur un graphique :



Sur ce graphique, on a fait apparaitre des carrés formés sur les écarts à la moyenne. La moyenne des aires de ces carrés est la variance. Elle représente l'aire d'un « carré moyen » dont le côté est l'écart-type.

IV Propriétés de la moyenne et de la variance de séries statistiques :

Série somme :

Soit une série statistique double, c'est-à-dire pour laquelle on considère deux caractères quantitatifs x et y, soit sous forme détaillée :

Numéro de l'individu : $1,\quad 2\ , \ldots.\quad , n$

Caractère x : $x_1,\ x_2, \ldots.\quad x_n$

Caractère y : $y_1,\ y_2, \ldots.\quad y_n$

On peut alors s'intéresser au caractère somme $x + y$:

Numéro de l'individu : $1,\quad 2\ ,\ldots.\quad ,n$

Caractère x : $x_1+y_1,\ x_2+y_2,\ldots.\quad x_n+y_n$

Alors la moyenne de la série somme $x+y$ est la somme des moyenne des séries x et y. Cela s'écrit :

$$\overline{x+y}=\bar{x}+\bar{y}$$

Preuve :

$$\overline{x+y}=\frac{(x_1+y_1)+\ (x_2+y_2)+\cdots.\ +(x_n+y_n)}{n}=\frac{(x_1+x_2+\cdots+x_n)+(y_1+y_2+\cdots+y_n)}{n}$$
$$=\frac{x_1+x_2+\cdots+x_n}{n}+\frac{y_1+y_2+\cdots+y_n}{n}=\bar{x}+\bar{y}$$

Remarque : De façon générale, il est faux de dire que la variance de la série somme est la somme des variances des séries x et y.

Série fonction affine d'une autre série :

Soit une série statistique x et deux constantes réelles a et b. Formons la série $y=a\,x+b$:

Numéro de l'individu : $1,\quad 2\ ,\ldots.\quad ,n$

Caractère x : $x_1,\quad x_2,\ldots.\quad x_n$
Caractère y : $a\,x_1+b,\ a\,x_2+b,\ldots.\quad a\,x_n+b$

Alors :

$$\bar{y}=a\,\bar{x}+b$$

Cette propriété est qualifiée de **linéarité de la moyenne**

Preuve :

$$\bar{y}=\frac{(a\,x_1+b)+\ (a\,x_2+b)+\cdots.\ +(a\,x_n+b)}{n}=\frac{a\,(x_1+x_2+\cdots+x_n)+(b+b+\cdots+b)}{n}$$
$$=a\,\frac{x_1+x_2+\cdots+x_n}{n}+\frac{n\,b}{n}=a\,\bar{x}+b$$

De plus :

$$V(y)=a^2\,V(x)$$
$$\sigma(y)=|a|\,\sigma(x)$$

Preuve :

$$V(a\,x+b)=\overline{(a\,x+b)^2}-\left(\overline{a\,x+b}\right)^2=\overline{a^2\,x^2+2\,a\,b\,x+b^2}-(a\,\bar{x}+b)^2$$
$$=a^2\,\overline{x^2}+2\,a\,b\,\bar{x}+b^2-(a^2\,\bar{x}^2+2\,a\,b\,\bar{x}+b^2)=a^2\,\overline{x^2}-a^2\,\bar{x}^2=a^2\left(\overline{x^2}-\ \bar{x}^2\right)$$
$$=a^2\,V(x)$$

$$\sigma(a\,x+b)=\sqrt{V(a\,x+b)}=\sqrt{a^2\,V(x)}=|a|\,\sqrt{V(x)}=|a|\,\sigma(x)$$

V Probabilités

Introduction au concept :

Les probabilités sont un concept très proche de celui des statistiques que nous allons illustrer avec des expériences faites avec des dés :

Expérience 1 : lancer d'un dé équilibré

On lance un dé équilibré un certain nombre de fois n.

On fait alors des statistiques sur les faces obtenues. La population est alors formée par l'ensemble des lancés de ce dé, le caractère est la face obtenue, c'est un caractère quantitatif discret dont les valeurs possibles sont 1,2,3,4,5,6.

On peut simuler un lancé de dés sur un tableur avec la commande : =ENT(6*ALEA())+1

ALEA() est une fonction renvoyant un nombre aléatoire x de l'intervalle $[0,1[$. Donc 6*ALEA() renvoie un nombre aléatoire de l'intervalle $[0,6[$ et ENT(6*ALEA()) un nombre aléatoire parmi 0,1,2,3,4,5. Donc en rajoutant 1 on simule un lancer aléatoire de dé.

Voici les résultats obtenus pour différentes valeurs de n :

Pour $n=10$ lancers :

résultats possibles	1	2	3	4	5	6
fréquence des résultats	10,0%	0,0%	10,0%	0,0%	40,0%	40,0%
fréquence théorique	16,7%	16,7%	16,7%	16,7%	16,7%	16,7%
écart	6,7%	16,7%	6,7%	16,7%	-23,3%	-23,3%

Avec la touche F9 du clavier , on constate que les fréquences d'apparition des différentes faces fluctuent autour d'une même valeur 16,7% (1/6), les écarts par rapport à cette valeur pouvant aller jusqu'à plus de 20%.

Pour $n=100$ lancers :

résultats possibles	1	2	3	4	5	6
fréquence des résultats	19,0%	15,0%	15,0%	11,0%	21,0%	19,0%
fréquence théorique	16,7%	16,7%	16,7%	16,7%	16,7%	16,7%
écart	-2,3%	1,7%	1,7%	5,7%	-4,3%	-2,3%

On observe le même phénomène mais les écarts sont moindres allant jusqu'à environ 5%.

Pour n =10 000 lancers :

résultats possibles	1	2	3	4	5	6
fréquence des résultats	16,5%	16,6%	16,4%	17,2%	16,2%	17,1%
fréquence théorique	16,7%	16,7%	16,7%	16,7%	16,7%	16,7%
écart	0,2%	0,1%	0,3%	-0,6%	0,4%	-0,4%

Ici, les écarts sont encore réduits, fluctuant jusqu'à des valeurs autour de 0,5 %.

On formule ainsi l'idée que plus le nombre n d'expériences de lancers va être grand, plus les fréquences d'apparition des faces observées , appelées fréquences empiriques (empirique = issu de l'expérience) vont être proches d'une même valeur. Comme la somme de ces 6 fréquences doit faire 100% ou encore 1, cette valeur est donc la fraction 1/6. On la qualifie de probabilité d'apparition de chacune des faces du dé et on dit chaque face est **équiprobable**.

On dit que la face obtenue, est une variable aléatoire qu'on note généralement avec une lettre majuscule, X par exemple. L'ensemble des faces physiques possibles est appelé **univers** et noté **Ω**, soit en Mathématiques :

$$\Omega = \{face\ 1, face\ 2, face\ 3, face\ 4, face\ 5, face\ 6,\}$$

Et l'ensemble des numéros associés à ces faces est qualifié d'univers image :

$$X\ (\Omega) = \{1,2,3,4,5,6,\}$$

La variable aléatoire X n'est donc qu'une fonction de Ω dans $X\ (\Omega)$ qui à une face physique associe le numéro qui est marqué dessus. Le fait que cette fonction soit qualifiée d'aléatoire vient de ce que les différentes faces physiques sont affectées d'une même probabilité d'apparition, 1/6 pour un dé équilibré, ce qui ce résume par un tableau de probabilité :

Face	Face 1	Face 2	Face 3	Face 4	Face 5	Face 6
Probabilité	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Ce tableau permet alors de probabiliser les éléments de l'univers image :

numéro	1	2	3	4	5	6
Probabilité	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Pour dire que le 2 sera obtenu dans un sixième des cas d'un très grand nombre de lancers, on écrira :

$$P(X = 2) = \frac{1}{6}$$

Ce qui se lira : la probabilité d'obtenir 2 est égale à 1/6 (soit une chance sur 6).

Imaginons alors un dé à 6 faces équilibré dans lequel il y aurait les trois premières faces portant le même numéro 1, puis les autres, 4,5,6. Dans ce cas on aurait :

$$X\ (\Omega) = \{1,4,5,6,\}$$

Et le tableau de probabilité dit loi de probabilité de X serait :

numéro	1	4	5	6
Probabilité	$\frac{3}{6} = \frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Dans ce cas, si les faces physiques restent équiprobables, les numéros obtenus ne le seraient plus.

<u>Expérience 2</u> : **lancer de deux dés équilibrés**

On lance deux dés à 6 faces équilibrés, un rouge, un jaune et on s'intéresse à la somme des numéros obtenus qui est une variable aléatoire qu'on notera X.

On constate alors par l'expérience comme dans le cas précédent que les fréquences d'apparition des différents couples de faces physiques possibles (il y en a 36) se rapprochent d'une même valeur lorsque le nombre de répétition du lancer des deux dés devient de plus en plus grand. Cette valeur est donc 1/36. On peut résumer cette situation par un tableau à double entrée :

Dé rouge\Dé jaune	Face 1	Face 2	Face 3	Face 4	Face 5	Face 6
Face 1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Face 2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Face 3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Face 4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Face 5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Face 6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

L'univers Ω formé des 36 couples de faces possibles est donc formé de couples équiprobables. L'univers image des sommes possibles est alors :

$$X\,(\Omega) = \{2,3,4,5,6,7,8,9,10,11,12\}$$

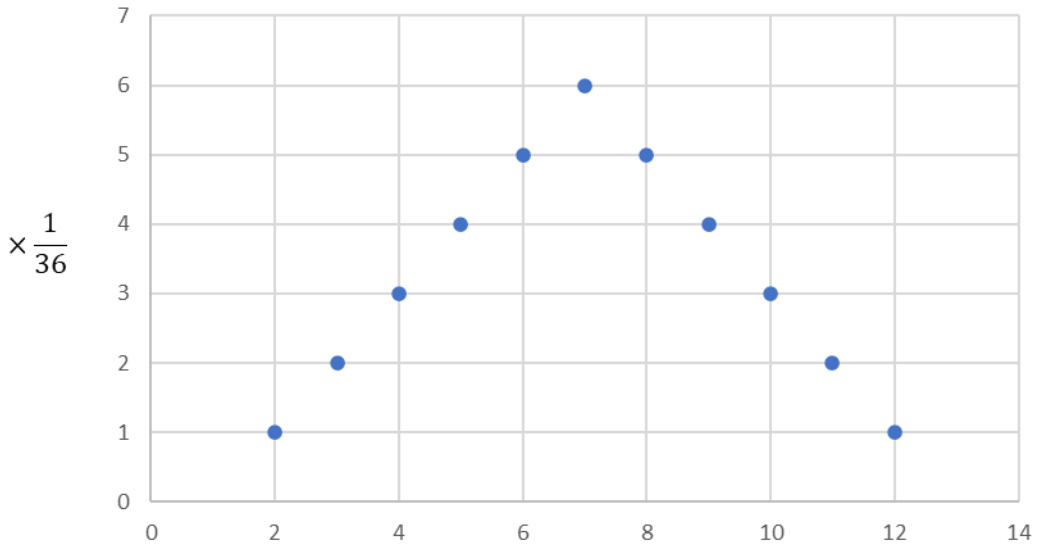
Pour probabiliser ses issues, nous allons à nouveau utiliser un tableau à double entrée faisant apparaitre ces sommes selon les couples de faces obtenus :

Dé rouge\Dé jaune	Face 1	Face 2	Face 3	Face 4	Face 5	Face 6
Face 1	2	3	4	5	6	7
Face 2	3	4	5	6	7	8
Face 3	4	5	6	7	8	9
Face 4	5	6	7	8	9	10
Face 5	6	7	8	9	10	11
Face 6	7	8	9	10	11	12

On peut donc prévoir que sur un grand nombre de répétitions de l'expérience des deux lancers, la somme 2 sera obtenue dans une fréquence proche de 1/36 des cas, la somme 3 dans 2/36 des cas, la somme 7 dans 6/36 des cas, etc…, ce qui permet d'établir la loi de probabilité de la somme :

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

Ce tableau peut être mis en graphique plus parlant et traduisant la **distribution de probabilités** des sommes possibles :



Une simulation sur tableur de 10 000 lancers de 2 dés a donné le tableau suivant :

Résultats possibles	2	3	4	5	6	7	8	9	10	11	12
Fréquence des résultats	2,6%	5,7%	8,7%	11,2%	13,2%	16,9%	13,6%	10,9%	8,8%	5,8%	2,6%
Fréquence théorique	2,8%	5,6%	8,3%	11,1%	13,9%	16,7%	13,9%	11,1%	8,3%	5,6%	2,8%
Ecarts	0,2%	-0,2%	-0,4%	-0,1%	0,7%	-0,2%	0,3%	0,2%	-0,4%	-0,3%	0,2%

On constate là encore que les fréquences observés des différentes sommes obtenues sont très proches des fréquences théoriques que sont les probabilités, les écarts allant jusqu'à un peu plus de 0,7%. Nous sommes donc en mesure, grâce à l'outil mathématique, de prédire (à une précision acceptable) les fréquences d'apparition de caractères associés à la répétition d'une même expérience aléatoire. C'est là tout l'intérêt des probabilités.

Les probabilités étant comme des statistiques, faites sur un grand nombre de répétitions d'une même expérience aléatoire, lorsque les issues de l'univers considéré sont des nombres, ce qui est le cas pour une variable aléatoire réelle, on peut caractériser cette variable aléatoire par la valeur vers laquelle tendrait la moyenne des résultats observés lorsque le nombre de répétitions de l'expérience tend vers l'infini, ce qu'on appelle **espérance de la variable aléatoire**. On peut également définir la variance et l'écart-type de cette variable de la même façon.

Dans le cas du lancer des deux dés, cela donne pour l'espérance :

$$E(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + 5 \times \frac{4}{36} + 6 \times \frac{5}{36} + 7 \times \frac{6}{36} + 8 \times \frac{5}{36} + 9 \times \frac{4}{36} + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = 7$$

Pour la variance :

$$\boldsymbol{V(X) = E\left(\left(X - E(X)\right)^2\right) = E\left(X^2\right) - \left(E(X)\right)^2}$$

$$== 4 \times \frac{1}{36} + 9 \times \frac{2}{36} + 16 \times \frac{3}{36} + 25 \times \frac{4}{36} + 36 \times \frac{5}{36} + 49 \times \frac{6}{36} + 64 \times \frac{5}{36} + 81 \times \frac{4}{36} + 100 \times \frac{3}{36} + 121 \times \frac{2}{36} + 144 \times \frac{1}{36} - 49 \approx 55$$

Pour l'écart-type :

$$\boldsymbol{\sigma(X) = \sqrt{V(X)}} = 1{,}96$$

Pour un organisateur de jeu aléatoire, connaitre par calcul mathématique son espérance de gain, lui permet de prévoir ce qu'il pourra gagner de façon quasi certaine s'il fait jouer un grand nombre de joueurs (10 000 par exemple). C'est le principe des casinos.

Propriétés de l'espérance et de la variance :

Soit une expérience aléatoire pour laquelle on considère deux variables aléatoires X et Y. On peut définir alors une variable somme $Z = X + Y$. Alors l'espérance de Z est la somme des espérance de X et de Y ce qui découle de la propriété vue en statistiques. On écrira ainsi :

$$\boldsymbol{E(X + Y)) = E(X) + E(Y)}$$

En particulier, si a et b sont deux constantes réelles alors, on a la propriété dite de linéarité de l'espérance :

$$\boldsymbol{E(a\,X + b)) = a\,E(X) + b}$$

En revanche il est faux d'écrire que la variance d'une somme est la somme des variances. Cependant :

$$\boldsymbol{V(a\,X + b)) = a^2\,V(X)}$$

$$\boldsymbol{\sigma(a\,X + b)) = |a|\,\sigma(X)}$$