I] Échantillonnage avec une calculatrice

Un sachet de *m&m's* est composé de bonbons de six couleurs également réparties entre : bleu, vert, jaune, rouge, orange, marron. Pablo ouvre un sachet de *m&m's*, et constate avec horreur que sur les 16 bonbons qu'il contient, aucun bonbon n'est rouge (sa couleur préférée) !!

a) Compléter le tableau qui donne les fréquences expérimentales (dans le sachet de Pablo) et théorique (en théorie, les couleurs sont également réparties).

Pour compléter le tableau, on calcule les rapports observés entre l'effectif d'une couleur et l'effectif total (fréquences exprimées en % en multipliant par 100 et en arrondissant à 0,01 près), par exemple 25,00% correspond au rapport 4÷16.

Couleur (code)	Effectifs	Fréquences expérimentales	Fréquences théoriques
Bleu (1)	4	25,00%	16,67%
Vert (2)	2	12,50%	16,67%
Jaune (3)	3	18,75%	16,67%
Rouge (4)	0	0,00%	16,67%
Orange (5)	5	31,25%	16,67%
Marron (6)	2	12,50%	16,67%

b) On décide de **simuler la production d'un sachet** de *m&m's* à l'aide de l'algorithme « SIMUL1 » :

- Lire N (ici on peut directement écrire N=16, cela évite d'avoir à le saisir à chaque fois)
- X=0 (ce sera notre compteur de m&m's rouges)
- Pour I allant de 1 à N : A=Nombre aléatoire de [0;1[; Si A<0,16666667 alors X=X+1
- affichage de X (nombre de m&m's rouges) ;
- Y=X×100÷N (calcul de la fréquence en %) ; affichage de Y

NB : Pour générer un nombre pseudo-aléatoire de l'intervalle [0 ; 1[: *Ran#* (Casio) ou *Rand* (TI)

Il peut y avoir des variantes selon les modèles de calculatrices. On peut écrire un programme très simple pour tester une fonction : avec TI par exemple, écrire le programme TEST qui contient une seule ligne : Disp *EntAleat*. Cela devrait à chaque fois afficher un nombre différent de [0 ; 1[. Si ce n'est pas le cas, s'il y a une erreur de syntaxe, revenir au catalogue ou, mieux, au manuel de la calculatrice.

pour Casio	Pour TI
?→N 0→X For 1→I TO N *Ran#*→A IF A<0,16666667 THEN X+1→ X IfEnd Next X ◢ X×100÷N→ Y Y ◢	input N 0→X For (*I,1,N*) *EntAleat*→A IF A<0,16666667 THEN X+1→ X End End Disp X X×100÷N→ Y Disp Y

>>Simuler la production de 10 sachets contenant 16 *m&m's* en utilisant 10 fois de suite ce programme.

N° du sachet	1	2	3	4	5	6	7	8	9	10
X=Nombre de *m&m's* rouges	3	2	3	2	5	1	3	2	2	4
Y=Fréquence de *m&m's* rouges	18,75	12,5	18,75	12,5	31,25	6,25	18,75	12,5	12,5	25

Combien de sachets contiennent 0 *m&m's* rouge?

Aucun, mais nous avons mis nos 10 premiers résultats, ils pourraient être différents...

Quelle est la fréquence moyenne des *m&m's* rouges? Nous avons obtenu 3+2+3+2+5+1+3+2+2+4=27 *m&m's* rouges sur 10×16=160 *m&m's*. Cela fait une fréquence moyenne de 27÷160=0,16875, soit 16,875%. Pas très loin de la fréquence théorique réelle : 16,6667%, mais nos résultats pourraient être différents...

Quel est leur nombre moyen par sachet? Nous avons obtenu 27 *m&m's* rouges sur 10 sachets. Cela fait un nombre moyen de 27÷10=2,7 par sachet. Pas très loin du nombre théorique réel : 16÷6=2,667 par sachet, mais nos résultats pourraient être différents...

c) On décide de **simuler la production d'un carton de M sachets** de *m&m's*, sans avoir à relancer M fois le programme. Pour cela on doit mettre au point une version améliorée de l'algorithme : « SIMUL2 ».

- On ajoute une boucle « Pour J allant de 1 à M » qui contient l'autre boucle (Pour J allant de 1 à M)
- On ajoute un compteur Y, pour comptabiliser les sachets contenant 0 *m&m's* rouge.
- On ajoute un compteur Z qui comptabilise le nombre total de *m&m's* rouges pour les M sachets.

Voici le nouvel algorithme « SIMUL2 » qui réalise ce nouvel objectif :

- Lire N (ici N=16), Lire M (ici M=100)
- Y=0 (compteur des sachets contenant 0 *m&m's* rouge), Z=0 (compteur de *m&m's* rouges au total)
- Pour J allant de 1 à M
  X=0 (compteur de *m&m's* rouges par sachet)
  Pour I allant de 1 à N : A=Nombre aléatoire de [0;1[;
  Si A<0,16666667 alors X=X+1
  Z=Z+X
  Si X=0 alors Y=Y+1
- affichage de Y (nombre de sachets contenant zéro *m&m's* rouge) ;
- affichage de Z (nombre de *m&m's* rouges au total)
- Z=Z×100÷N (calcul de la fréquence en %) ; affichage de Z

Programmons cet algorithme :

Casio	TI
?→N	input N
?→M	input M
0→Y	0→Y
0→Z	0→Z
For 1→J TO M	For (J,1,N)
0→X	0→X
For 1→I TO N	For (I,1,N)
Ran#→A	*EntAleat*→A
IF A<0,16666667	IF A<0,16666667
THEN X+1→ X	THEN X+1→ X
IfEnd	End
Next	End
Z+X→Z	Z+X→Z
IF X=0	IF X=0
THEN Y+1→ Y	THEN Y+1→ Y
IfEnd	End
Next	End
Y ◢	Disp Y
Z×100÷N→ Z	Z×100÷N→ Z
Z ◢	Disp Z

>>Simuler 10 fois la production d'un carton de 100 sachets contenant 16 *m&m's* en prenant M=100.

N° du carton de 100	1	2	3	4	5	6	7	8	9	10
Nombre de sachets contenant 0 *m&m's* rouge	1	7	10	3	4	7	6	6	4	6
Fréquence de *m&m's* rouges	18,13%	15,69%	16,75%	18,25%	17,68%	16,87%	15,93%	16,18%	18,31%	17,31%

Au vu de ces résultats que pouvez-vous conclure sur les sachets contenant 0 *m&m's* rouge? Sont-ils exceptionnels ou fréquents? S'agit-il d'une tricherie, d'un vol ou d'un phénomène dominé par le hasard?

Sur 100 sachets contenant 16 *m&m's*, il y a toujours quelques sachets ne contenant aucun *m&m's* rouge. Sur nos 10 cartons de 100, il y en a entre 3 et 10, avec une moyenne de 5,9 pour 100 sachets, car (6+7+10+3+4+7+6+6+4+6)÷10=59÷10=5,9. Pablo est donc tombé sur un de ces sachets (il y en a environ 6%). Ils ne sont ni exceptionnels ni fréquents (1 pour 17 environ), on pourrait dire qu'ils sont assez rares. On peut penser qu'il s'agit d'un phénomène dominé entièrement par le hasard. On ne pourrait pas comprendre l'intérêt de fabriquer des sachets ne contenant aucun *m&m's* rouge (il n'y a pas de mobile). La chaîne de fabrication semble fabriquer autant de bonbons de chaque couleur, ceux-ci sont mélangés et mis en sachet par des machines qui ne tiennent pas compte des couleurs. Il s'agit, en principe donc, de hasard.

d) Déterminer l'intervalle de fluctuation au seuil de 95% dans lequel doit se trouver la fréquence des *m&m's* rouge dans un sachet. Est-on dans le domaine d'application de la formule ?

La fréquence théorique étant de 1/6, soit 0,16666667 environ, et l'échantillon que l'on vient de générer contenant 10×100×16=16 000 bonbons, l'intervalle de fluctuation au seuil de 95% est

$$\left[\frac{1}{6}-\frac{1}{\sqrt{16000}}\approx 0{,}15876\ ;\ \frac{1}{6}+\frac{1}{\sqrt{16000}}\approx 0{,}17457\right]$$

La fréquence expérimentale d'obtention des bonbons rouges est, en moyenne pour nos 10 fois 100 sachets de 16 bonbons, de 17,11% (j'ai additionné mes 10 fréquences puis divisé par 10 pour en obtenir la moyenne), soit 0,1711. Est-on dans l'IF ? Certes, oui, et heureusement, puisque nous avons vraiment procédé par le seul effet du hasard. Cela aurait été étonnant d'être en dehors de l'IF (même si ça arrive, en principe dans 5% des cas, mais alors, on n'aurait pas été très loin...).

=SOMME(A1:J1)

A	B	C	D	E	F	G	H	I	J	K
18,13%	15,69%	16,75%	18,25%	17,68%	16,87%	15,93%	16,18%	18,31%	17,31%	171,10%

On peut cependant comprendre la question autrement : lorsque Pablo ouvre son paquet, il n'y a pas de bonbon rouge : la fréquence expérimentale d'obtention des bonbons rouges est de 0/16 dans son paquet. L'IF pour cet échantillon n'est pas celui que l'on a calculé plus haut sur 16000 bonbons. C'est un intervalle beaucoup plus large :

$$\left[\frac{1}{6}-\frac{1}{\sqrt{16}}\approx -0{,}0833\ ;\ \frac{1}{6}+\frac{1}{\sqrt{16}}\approx 0{,}416\right]$$

On peut prendre 0 comme valeur inférieure car une fréquence ne peut pas être négative.

Donc IF=$[0\ ;\ 0{,}416]$, et $0\in IF$. La fréquence obtenue est compatible avec le tirage par le seul fait du hasard. Est-on dans le domaine d'application de la formule ? Pas tout-à-fait ici car la taille de l'échantillon est trop petit (16 au lieu de 20 minimum) et la fréquence aussi est trop faible (0,16 au lieu de 0,2 minimum). Néanmoins, comme on n'est pas très loin du domaine d'application, on peut dire que le tirage est compatible avec l'hypothèse que seul le hasard prévaut dans sa constitution.

NB : on peut s'amuser à augmenter le nombre de bonbons par sachet pour voir à partir de quelle valeur l'IF ne contient plus 0. C'est à partir de n=37 bonbons dans un sachet que l'on doit commencer à s'étonner qu'il n'y ait pas de bonbons rouges dans le sachet. Car alors la fréquence 0 n'appartient pas à l'IF. Mais cela reste que dans 5% des cas, ce sera tout de même le cas...

n	16	20	25	30	35	36	37	38	39	40
$\frac{1}{6}-\frac{1}{\sqrt{n}}$	-0,0833	-0,0569	-0,0333	-0,0159	-0,0024	0,0000	0,0023	0,0044	0,0065	0,0086

e) Prolongement (la question n'était pas sur la feuille, elle n'est donc pas au programme du contrôle)

On décide de simuler la production de 100 sachets de *m&m's* à l'aide du tableur : ouvrir une page « classeur » du tableur « calc » d'*OpenOffice*. Pour simuler le tirage aléatoire d'une couleur codée par un chiffre allant de 1 à 6, on utilise les fonctions :

- ALEA() qui génère un nombre pseudo-aléatoire de l'intervalle [0 ; 1[
- ENT(*xy,zt*) qui donne la partie entière *xy* d'un nombre décimal *xy,zt*
- NB.SI(*plage;x*) qui calcule le nombre d'élément de la *plage* donnée contenant la valeur *x*

Par exemple, l'instruction =ENT(ALEA()*6)+1 permet de générer un nombre entier aléatoire allant de 1 à 6 (on peut aussi utiliser l'instruction =ALEA.ENTRE.BORNES(1;6)). L'instruction =NB.SI(B1:B20;4) permet de compter toutes les occurrences du nombre 4 (codant le *m&m's* rouge) dans la plage qui va de la cellule B1 à la cellule B20.

>>Simuler la production de 100 sachets contenant 16 *m&m's*.

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total PILE	f_exp
1																	0	0,000
2		ROUGE						ROUGE									2	0,125
3										ROUGE							1	0,063
4						ROUGE			ROUGE					ROUGE			3	0,188
5		ROUGE															1	0,063
6			ROUGE			ROUGE	ROUGE				ROUGE			ROUGE		ROUGE	6	0,375
7				ROUGE						ROUGE						ROUGE	3	0,188
8			ROUGE									ROUGE			ROUGE		3	0,188
9	ROUGE												ROUGE				2	0,125
10							ROUGE	ROUGE					ROUGE				3	0,188

Voici le début de notre tableau qui donne une simulation de 10 sachets de 16 bonbons. Pour afficher la couleur on a tapé : =SI(ALEA.ENTRE.BORNES(1;6)=4;"ROUGE";"")

Pour faire la somme des bonbons rouges à droite, on a tapé : =NB.SI(B2:Q2;"ROUGE")

En bas du tableau (après avoir étendu les formules des 19 colonnes jusqu'à N°=100 lignes), on a obtenu la somme des effectifs en tapant : =SOMME(R2:R101)
On a obtenu la fréquence moyenne en divisant par 16×100. La moyenne peut être obtenue en divisant par 100, sinon, le tableur donne ce résultat en faisant : =MOYENNE(R2:R101)
Pour l'écart-type, il suffit de faire :=ECARTYPEP(R2:R101)
Pour les valeurs extrêmes de la fréquence des bonbons rouges on tape : =MIN(S2:S101) et =MAX(S2:S101)

98									ROUGE	ROUGE							2	0,125
99	ROUGE										ROUGE						2	0,125
100				ROUGE	ROUGE		ROUGE		ROUGE	ROUGE							5	0,313

Total 100 échantillons	271
fréquence moyenne	0,16938
moyenne effectif	2,71
Ecart-type	1,52
f min	0,000
f max	0,375

Il faut noter que ces valeurs calculées, sur nos 100 échantillons aléatoires, vont varier d'une personne à une autre. On peut retirer (dans le sens de recommencer un tirage) les couleurs avec Ctrl+Maj+F9, on obtient alors d'autres résultats :

II] Une autre situation d'échantillonnage

Avec le même algorithme, en modifiant N, M et la fréquence théorique, étudier cette autre situation :
Dans la classe de 2de3 il y a 18 garçons et 21 filles. Est-ce le résultat du hasard ou bien est-ce qu'il y a plus de filles en seconde au Lycée Henri IV? Pour tester cette hypothèse, on part d'une *fréquence théorique* de 50% de filles dans une classe d'âge et on génère 5 classes de 39 élèves.
Faire les adaptations nécessaire et comptabiliser dans Y les classes d'au moins 21 filles.
Approche expérimentale : Combien y a t-il, parmi les 5 classes générées, de classes contenant au moins 21 filles ? ………..... Que pensez-vous de l'hypothèse testée?

N° de la classe de 39 élèves	1	2	3	4	5	6	7	8	9	10
Nombre de classes contenant 21 filles ou plus sur les 5 classes	1	3	1	1	2	1	2	2	1	3
Fréquence des filles	56,4%	49,2%	48,7%	48,2%	47,7%	47,0%	46,2%	46,2%	43,6%	53,3%

On génère 5 classes de 39 élèves (M=5, N=39).
La mise au point du programme pour faire ce décompte exige, en plus, deux petites modifications des tests

- à la place de l'instruction « Si X=0 alors Y=Y+1 » on écrit « Si X≥21 alors Y=Y+1 »
- à la place de l'instruction « Si A<0,16666667 alors X=X+1 » on écrit « Si A<0,5 alors X=X+1 »
- à la place de l'instruction « Z=Z×100÷N » on écrit « Z=Z×100÷N÷M »

Combien y a t-il, parmi les 5 classes générées, de classes contenant au moins 21 filles ?
Ici, nous avons généré 10 fois 5 classes. Sur 5 classes contenant 39 élèves, il y a entre 1 et 3 classes contenant au moins 21 filles (cinq groupes de 5 classes contiennent 1 classe ayant au moins de 21 filles). Sur nos 10 groupes de 5 classes, la moyenne est de 1,5 pour 10 classes, car (1×5+2×2+2×3)÷10=15÷10=1,5. Notre classe est donc une de ces classes (il y en a 15÷50=0,3 soit 30%). Elles sont assez fréquentes (un peu moins de 1 pour 3).
Que pensez-vous de l'hypothèse testée?
On peut, encore ici, penser qu'il s'agit d'un phénomène dominé entièrement par le hasard avec une équiprobabilité des filles et des garçons.

Approche théorique : Déterminer l'intervalle de fluctuation au seuil de 95% dans lequel doit se trouver la fréquence des filles dans une classe. Que pensez-vous de l'hypothèse testée?
Dans un échantillon de 39 éléments, avec une fréquence théorique de 0,5, l'intervalle de fluctuation pour la fréquence des filles dans une classe est [0,3399 ; 0,6601] (nous utilisons la formule du cours : $f\pm\frac{1}{\sqrt{n}}$).
Dans notre classe, la fréquence des filles est de $\frac{21}{39}\approx 0{,}53846$, soit une fréquence compatible avec l'hypothèse d'équiprobabilité des filles et des garçons, au seuil de 5% (en prenant le risque de se tromper 5 fois sur 100). Nous ne devons donc pas chercher à expliquer ce nombre supérieur de filles par un autre facteur que l'effet du hasard qui suffit, à lui seul, à expliquer ce léger surnombre.