2de3 NOM : ….. **DS n°4** : Statistiques & Échantillonnage 13/12/16

Compléter la feuille directement, le reste sur la copie. N'oubliez pas de rendre la feuille ! durée 2*h* - barème : 5*pts par question*

Utilisateurs de FB

Voici la répartition des 26 millions d'utilisateurs actifs de Facebook en France en 2016, selon leur âge.

a) Compléter ce tableau : amplitude des intervalles, effectif et densité* de chaque classe d'âges (prendre 80 ans pour le maximum de la dernière classe).

* on rappelle que la densité est le rapport de l'effectif sur l'amplitude de l'intervalle

âge	fréquence	Effectif	Amplitude	Densité	Hauteur
[13 ; 18[	16%	4 160 000	5	832000	13,4
[18 ; 25[	25%	6 500 000	7	928571	15
[25 ; 35[	26%	6 760 000	10	676000	10,9
[35 ; 45[	16%	4 160 000	10	416000	6,7
[45 ; 55[	9%	2 340 000	10	234000	3,8
≥55	8%	2 080 000	25	83200	1,3

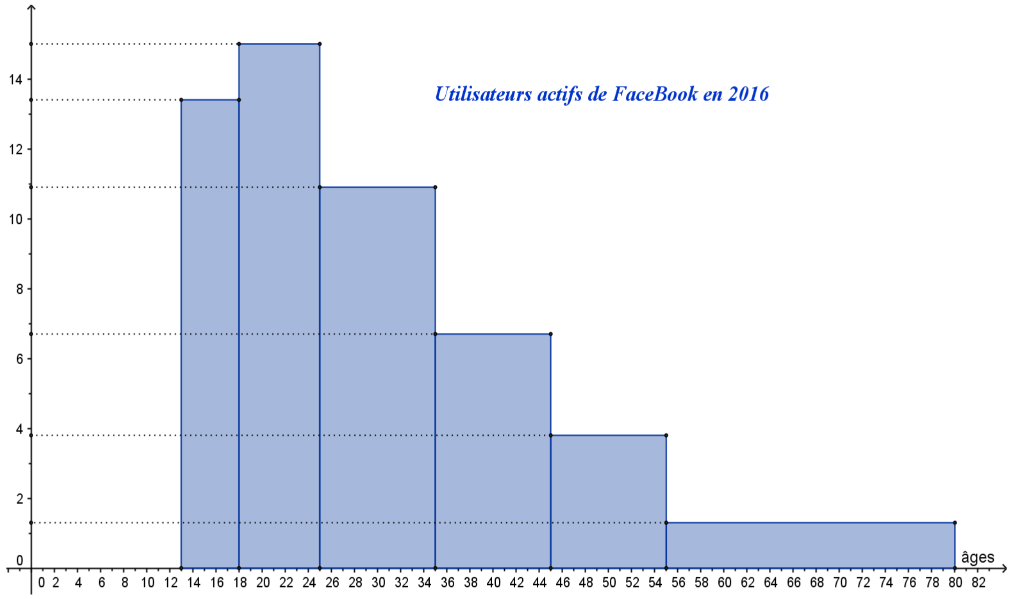
b) Tracer un histogramme représentant la répartition des utilisateurs actifs de Facebook en prenant 15 *cm* pour hauteur maximum.

On utilise les densités pour tracer des rectangles d'aire proportionnelle à l'effectif.

Quelle est la valeur modale de cette série ?

La série est unimodale, la classe modale est la classe [18;25[(c'est la classe de plus grand densité).

On peut estimer le mode au centre de cette classe, soit à 21 ans.

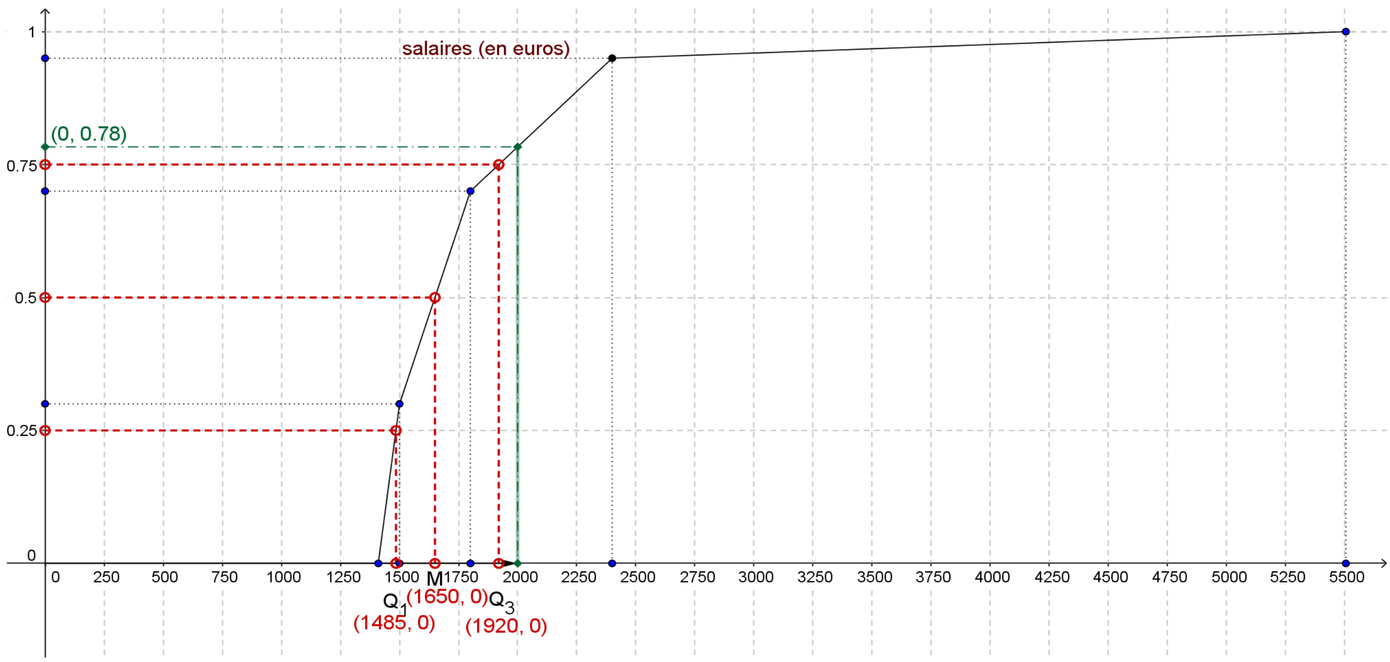


2) Salaires dans une entreprise

Une entreprise de 1250 personnes, soucieuse de transparence, publie la répartition des salaires :

Tranches de salaire (en euros)	Moins de 1500	De 1500 à 1800	De 1800 à 2400	Plus de 2400
Fréquences	0,3	0,4	0,25	0,05
Fréquences cumulées	0,3	0,7	0,95	1

a) Sachant que le salaire le plus faible dans l'entreprise est 1410 € et le salaire le plus élevé est 5505 €, représenter la série des fréquences cumulées par une ligne polygonale permettant de répondre à la question « quel pourcentage des salariés a un salaire inférieur ou égal à *x* euros ? ».



Déterminer, à l'aide de ce graphique, le nombre de salariés ayant un salaire inférieur ou égal à 2000 €.
On lit, sur ce graphique, que la fréquence des salariés ayant un salaire inférieur ou égal à 2000 € est 0,78. Comme il y a 1250 personnes dans cette entreprise, cela représente 975 salariés (0,78×1250=975).

b) Mettre en évidence sur ce graphique la médiane M et les quartiles Q_1 et Q_3. Donner alors ces paramètres.
En rouge, on voit que Q_1=1485 €, M=1650 € et Q_3=1920 €.
Montrer comment on obtient, par le calcul, la valeur de M.
Pour calculer M, on repère la classe dans lequel il se situe : la classe allant de 1500 à 1800.
On écrit l'égalité $\frac{0,5-0,3}{0,7-0,3}=\frac{M-1500}{1800-1500}$ qui vient du théorème de Thalès (ou bien de la condition de colinéarité des vecteurs), ce qui conduit à $M=1500+(1800-1500)\times\frac{0,5-0,3}{0,7-0,3}=1500+300\times\frac{0,2}{0,4}=1500+150=1650$.

c) Calculer l'écart interquartile, puis un autre paramètre de la dispersion.
Comparer, sur cet exemple, ces deux mesures de la dispersion.
L'écart interquartile est $Q_3 - Q_1$= 1920−1485=435 €.
L'étendue est un autre paramètre de la dispersion. Ici, l'étendue vaut 5505−1410=4095 €.
L'étendue porte sur les deux valeurs extrêmes (les moins représentatives) de la répartition alors que l'écart interquartile porte sur sa moitié la plus représentative. Si celle-ci est dix fois plus élevée que celui-là, cela signifie qu'il y a de grands écarts pour les valeurs extrêmes alors que la série est plutôt concentrée pour les valeurs intermédiaires. En fait ce sont les salaires élevés qui, s'écartant beaucoup de la moyenne, créent cette dispersion apparente.

3) Pains

Une association de consommateurs effectue des mesures du poids de la baguette « tradition » dans plusieurs boulangeries parisiennes. Dans les boulangeries K et C, les poids se répartissent comme suit :

Poids en grammes	[235;240[	[240;245[	[245;250[	[250;255[	[255;260[	[260;265[
Effectifs boulangerie K	12	25	30	28	15	7
Effectifs boulangerie C	23	37	31	7	0	0

a) Pour les baguettes de la boulangerie K, dresser un tableau donnant :
les centres x_i de classe, les effectifs n_i, les produits n_ix_i et $n_ix_i^2$, ainsi que les sommes Σn_i , $\Sigma n_i x_i$ et $\Sigma n_i x_i^2$.
En déduire la moyenne $\bar{x}$ et l'écart-type $\sigma=\sqrt{\frac{\Sigma n_i x_i^2}{\Sigma n_i}-\bar{x}^2}$ pour la boulangerie K.

Poids (g)	[235;240[	[240;245[	[245;250[	[250;255[	[255;260[	[260;265[	total	moyenne
x_i	237,5	242,5	247,5	252,5	257,5	262,5		
n_i	12	25	30	28	15	7	117	
$n_i x_i$	2850	6062,5	7425	7070	3862,5	1837,5	29107,5	248,78
$n_i x_i^2$	676875	1470156,25	1837687,5	1785175	994593,75	482343,75	7246831,25	61938,73

Pour la boulangerie K, on lit sur le tableau : $\Sigma n_i=98$, $\Sigma n_i x_i=29107,5$, $\Sigma n_i x_i^2=7246831,25$.
La moyenne est $\bar{x}=\frac{\Sigma n_i x_i}{\Sigma n_i}\approx\frac{29107,5}{117}=248,78\ g$,
l'écart-type $\sigma=\sqrt{\frac{\Sigma n_i x_i^2}{\Sigma n_i}-\bar{x}^2}=\sqrt{\frac{7246831,25}{117}-248,78^2}=\sqrt{61938,73-61891,49}=\sqrt{46,22}=6,8\ g$.

b) Donner les trois sommes pour la boulangerie C (utiliser le mode statistique pour les obtenir) ainsi que la moyenne $\bar{x}$ et l'écart-type σ pour la boulangerie C.
Pour la boulangerie C, on a : $\Sigma n_i=117$, $\Sigma n_i x_i=23875$, $\Sigma n_i x_i^2\approx5818000$.
Le mode statistique nous donne aussi la moyenne $\bar{x}=243,62\ g$ et l'écart-type $\sigma=4,4\ g$.
Comparer les baguettes des deux boulangeries.

	moyenne	écart-type	La boulangerie K fait des baguettes plus lourdes (5,16 g en moyenne en plus, soit 2% en plus) que la boulangerie C mais les poids sont plus variables, plus dispersés. On peut penser que la boulangerie K est plus artisanale alors que la C est plus industrielle (poids calibrés, recherche d'une économie de masse). On peut mentionner que l'étude porte sur davantage de baguettes K que C, mais cela ne change pas les résultats.
boulangerie C	243,62 g	4,4 g	
boulangerie K	248,78 g	6,8 g	

Pour compléter cette étude, il faudrait comparer les prix et la qualité du pain. Au passage, rappelons que qualité et quantité ne sont pas synonymes : ce n'est pas parce que les baguettes de K sont plus lourdes qu'elles sont meilleures...

4) <u>QCM</u>

*Un Questionnaire à Choix Multiples comporte cinq questions. Pour chacune d'entre elles, quatre réponses sont proposées mais une seule est exacte. Une réponse juste rapporte 2 points, une répon*se fausse enlève 1 *point. Pour n'avoir que des notes positives, on remplace une note finale négative par 0.*

a) Écrire un algorithme QCM1 qui simule un candidat qui répond au hasard à chaque question.
Cet algorithme doit afficher la note finale du candidat.
Traduire QCM1 en programme, le lancer cinq fois, noter les résultats obtenus.

Pour simuler une réponse au hasard : vu qu'il y a 4 questions, chacune a une probabilité 1/4 d'être choisie. Donc on examine si on obtient un nombre aléatoire R qui a une probabilité équivalente d'advenir : les nombres de [0;0,25[représentent 1/4 de ceux de l'intervalle [0;1[dans lequel on choisit ce nombre.

- S=0
- Pour I allant de 1 à 5 : Affecter un nombre aléatoire de [0;1[à R
  Si R≤0,25 Alors S=S+2 à S Sinon S=S-1
- Si S<0 alors S=0
- Afficher S

Traduisons cet algorithme en programme :

Casio	TI
0→S *For* 1→I To 5 *Ran#*→R *If* R<0,25 *Then* S+2→ S *Else* S-1→ S *IfEnd* *Next* *If* S<0 *Then* 0→ S *IfEnd* S ◢	0→S For (I ,1,5) *Rand*→R (ou *NbrAleat*, ou que sais-je encore) *If* R<0,25 *Then* S+2→ S *Else* S-1→ S *End* *End* *If* S<0 *Then* 0→ S *End* *Disp* S

Programmons maintenant cela sur la calculatrice et exécutons deux fois le programme 5 fois.

Essai n°	1	2	3	4	5	1	2	3	4	5
Note obtenue	0	1	0	4	0	0	1	7	1	0

Une petite remarque pour ceux qui seraient tentés de répondre au hasard : à ce QCM, toutes les notes ne sont pas possibles ! On ne peut avoir que 10, 7, 4, 1 ou 0 :

- 10/10 (5 bonnes réponses) ;
- 7/10 (4 bonnes réponses et 1 mauvaise : 4×2-1=7) ;
- 4/10 (3 bonnes réponses et 2 mauvaise1 : 3×2-2=4) ;
- 1/10 (2 bonnes réponses et 3 mauvaises : 2×2-3=1) ;
- 0 dans tous les autres cas (1 bonne réponse et 4 mauvaises : 1×2-4=-2 arrondi à 0 ; 0 bonne réponse et 5 mauvaises : 0×2-5=-5 arrondi à 0)

Mais toutes ces réponses ne sont pas équiprobables ! Pour anticiper sur le cours de probabilités :

- la probabilité d'avoir 10/10 est $\frac{1}{4^5}=\frac{1}{1024}=0{,}0009765625$
- la probabilité d'avoir 7/10 est $\frac{3\times 5}{4^5}=\frac{15}{1024}=0{,}0146484375$ car on doit échouer à une (n'importe laquelle) des 5 questions et réussir aux autres. Il suffit de choisir laquelle est réussie : 5 possibilités.
- la probabilité d'avoir 4/10 est $\frac{3^2\times 10}{4^5}=\frac{90}{1024}=0{,}087890625$ car on doit échouer deux (on choisit la 1ère parmi 5, la 2de parmi 4 et on divise par 2 car on a tout compté deux fois, cela fait 10 choix possibles: MMBBB, MBMBB, MBBMB, MBBBM, BMMBB, BMBMB, BMBBM, BBMMB, BBMBM, BBBMM) des 5 questions et réussir aux autres. Il suffit de choisir laquelle est réussie : 5 possibilités.
- la probabilité d'avoir 1/10 est $\frac{3^3\times 10}{4^5}=\frac{270}{1024}=0{,}263671875$ car on doit réussir deux des 5 questions (on choisit la 1ère parmi 5, la 2de parmi 4 et on divise par 2 car on a tout compté deux fois, cela fait 10 choix possibles : BBMMM, BMBMM, BMMBM, BMMMB, MBBMM, MBMBM, MBMMB, MMBBM, MMBMB, MMMBB) et échouer aux autres. Il suffit de choisir laquelle est réussie : 5 possibilités.

- la probabilité d'avoir 0/10 en est $\frac{3^4\times 5+3^5}{4^5}=\frac{648}{1024}=0{,}6328125$ pour les deux dernières possibilités. On peut vérifier que ce nombre est bien obtenu en faisant $1-\left(\frac{270}{4^5}+\frac{90}{1024}+\frac{15}{1024}+\frac{1}{1024}\right)=1-\frac{376}{1024}=\frac{1024-376}{1024}=648$, ce qui est la probabilité de l'événement contraire de « obtenir 1, 4, 7 ou 10/10 ».

Les QCM remplis au hasard on donc une probabilité de 63% environ d'avoir la note 0/10 ce que l'on va évaluer grâce à l'intervalle de confiance.

b) On veut maintenant simuler M candidats qui répondent au hasard. Écrire un nouvel algorithme QCM2 qui réalise cela : au départ, il demande la valeur de M et, à la fin, il affiche la note moyenne des candidats, puis, la fréquence des candidats ayant obtenu 0 au test (utiliser un compteur Z pour comptabiliser les notes nulles). Traduire QCM2 en programme, le lancer avec M=100, noter le résultat obtenu.
On doit inclure la boucle For de QCM1 dans une autre boucle For qui réalise l'échantillon de M candidats.
Pour ce qui est du compteur Z, il faut l'initialiser, l'incrémenter à la fin de la boucle extérieure, en modifier le contenu pour qu'il donne une fréquence, et afficher celle-ci.
Voici donc l'algorithme QCM2 :

- Saisir M (le nombre de candidats)
- Z=0 (compteur les notes nulles)
- T=0 (compteur du total des notes)
- Pour I allant de 1 à M :

        S=0 (la note du candidat)
        Pour J allant de 1 à 5 :
            Affecter un nombre aléatoire de [0;1[à R
            Si R≤0,25 Alors S=S+2 à S Sinon S=S-1
        Si S<0 alors S=0
        Si S=0 alors Z=Z+1
        T=T+S

- T=T÷M (calcul de la moyenne des notes)
- Afficher T
- Z=Z÷M×100 (calcul de la fréquence en % des notes nulles)
- Afficher Z

Traduisons cet algorithme en programme :

Casio	TI
?→M	Input M
0→Z	0→Z
0→T	0→T
For 1→J To M	For (J ,1,M)
0→S	0→T
For 1→I To 5	For (I ,1,5)
Ran#→R	*Rand*→R
If R<0,25	*If* R<0,25
Then S+2→ S	*Then* S+2→ S
Else S-1→ S	*Else* S-1→ S
IfEnd	*End*
Next	*End*
If S<0	*If* S<0
Then 0→ S	*Then* 0→ S
IfEnd	*End*
If S=0	*If* S=0
Then Z+1→ Z	*Then* Z+1→ Z
IfEnd	*End*
T+S→ T	T+S→ T
Next	End
T÷M→ T	T÷M→ T
T ◢	*Disp* T
Z÷M×100→ Z	Z÷M×100→ Z
Z ◢	*Disp* Z

Voici des valeurs obtenues pour M=100 candidats. Je donne dix réponses différentes, mais une seule était attendue.

	1	2	3	4	5	6	7	8	9	10
Moyenne des notes	0,58	0,78	0,68	0,79	1	0,87	0,66	0,77	1,21	0,96
Fréquences des notes nulles	72%	64%	62%	54%	54%	64%	70%	65%	51%	61%

En répondant au hasard, plus de la moitié des candidats auraient eu 0 au test (61,7% en moyenne pour les 10 valeurs obtenues), soit un peu moins de 2 candidats sur 3. La réponse des probabilité est la vraie valeur : $\frac{648}{1024}=0{,}6328125$. On s'en approche ici car on a essayé un grand nombre de fois.
La moyenne de ces 10×100=1000 résultats hasardeux est assez lamentable : 0,83 alors que les notes peuvent aller de 0 à 10. On voit déjà l'efficacité de ce barème...

c) Déterminer alors l'intervalle de confiance dans lequel se situe la fréquence réelle d'obtenir 0 à ce test.
Le barème est-il suffisamment dissuasif, pour ceux qui seraient tentés de répondre au hasard ?
L'intervalle de confiance est [$f-\frac{1}{\sqrt{n}}$; $f+\frac{1}{\sqrt{n}}$] où f est la fréquence expérimentale obtenue et n est le nombre de candidats dans l'échantillon. Pour en rester à ce qui était demandé, il y a n=100 candidats, et la fréquence obtenue est, par exemple, la première mentionnée : 72% (c'est la plus grand valeur obtenue). J'obtiens donc l'intervalle de confiance [$0{,}72-\frac{1}{\sqrt{100}}$; $0{,}72+\frac{1}{\sqrt{100}}$], soit [0,62 ; 0,82].
Avec f=51% (la plus faible), j'obtiendrais l'intervalle de confiance [0,41 ; 0,61], un intervalle qui n'a même pas d'intersection avec le premier... La conclusion est la même pourtant dans les deux cas : en répondant au hasard, un candidat a un peu plus d'une chance sur 2 d'obtenir 0 au QCM. Il semblerait que cela puisse dissuader les candidats de répondre ainsi.
Si on poursuivait l'étude de ce barème, on verrait que les candidats répondant au hasard ont près de 90% de chances d'avoir une note inférieure ou égale à 2/10.

Marches aléatoires

Une puce se déplace sur un axe gradué, d'une unité à chaque saut, dans un sens ou dans l'autre, de façon aléatoire et équiprobable. On note x l'abscisse de la puce ; la puce partant de l'origine, au départ x=0.

a) Écrire un algorithme ALG1 qui affiche le nombre N de sauts nécessaires pour atteindre $|x|=5$.
Traduire ALG1 en programme, le lancer cinq fois, noter les résultats obtenus.

• Lire D (a distance à atteindre ; ici on peut directement écrire D=5)
• X=0 (l'abscisse de la puce)
• N=0 (le nombre de sauts)
• Tant que X<D : A=Nombre aléatoire de [0;1[; Si A<0,5 alors X=X+1sinon X=X-1 ; N=N+1
• Affichage de N

Traduisons cet algorithme en programme :

Casio	TI
?→D	Input D
0→X	0→X
0→N	0→N
While abs(X)<D	While abs(X)<D
Ran#→A	*EntAleat*→A (ou Rand, ça dépend des calculatrices)
If A<0,5	If A<0,5
Then X+1→ X	Then X+1→ X
Else X-1→ X	Else X-1→ X
IfEnd	End
N+1→ N	N+1→ N
WhileEnd	End
N ◢	Disp N

Programmons maintenant cela sur la calculatrice et exécutons 5 fois le programme.

Essai	1	2	3	4	5
Nombre de sauts	9	35	39	5	17

Le résultat 5 est le plus petite valeur que l'on puisse obtenir car il faut au moins 5 étapes pour atteindre cette distance.

b) Écrire un algorithme ALG2 qui simule M=100 marches aléatoires de N=50 sauts. Comptabiliser les marches qui ont atteint au moins une fois $|x|=5$. Programmer cet algorithme, le lancer une fois et déterminer la fréquence des marches qui ont atteint au moins une fois $|x|=5$. Dans quel intervalle de confiance se situe la fréquence réelle des marches de 50 sauts ayant atteint au moins une fois $|x|=5$?

Voici le nouvel algorithme « ALG2 » qui réalise ce nouvel objectif :

- Lire M (ici M=100), Lire N (ici N=50)
- D=5, Z=0 (compteur des marches ayant atteint la distance 5)
- Pour J allant de 1 à M
  X=0 (l'abscisse de la puce) , Y=0 (indicateur de distance atteinte)
  Pour I allant de 1 à N : A=Nombre aléatoire de [0;1[
  Si A<0,5 alors X=X+1 sinon X=X-1
  Si abs(X)=5 alors Y=1
  Si Y=0 alors Z=Z+1
- affichage de Z
- Z=Z×100÷N (calcul de la fréquence en %) ; affichage de Z

Programmons cet algorithme :

Casio	TI
?→M	**input** M
?→N	**input** N
0→Z	0→Z
5→D	5→D
For 1→J **To** M	**For** (J,1,M)
0→X	0→X

0→Y	0→Y
For 1→I **To** N	**For** (I,1,N)
Ran#→A	*EntAleat*→A
If A<0,5	**If** A<0,5
Then X+1→ X	**Then** X+1→ X
Else X-1→ X	**Else** X-1→ X
IfEnd	**End**
If abs(X)=5	**If** abs(X)=5
Then 1→ Y	**Then** 1→ Y
IfEnd	**End**
Next	**End**
If Y=1	**If** Y=1
Then Z+1→Z	**Then** Z+1→Z
IfEnd	**End**
Next	**End**
Z ◢	**Disp** Z
Z×100÷N→ Z	Z×100÷N→ Z
Z ◢	**Disp** Z

Lançons ce programme. Il tourne un certain moment, puis affiche le résultat : 89, puis 89%.
Je le relance une 2ème fois, pour donner un autre exemple d'exécution. Je trouve 87 et 87% cette fois.
NB : C'est normal que l'on trouve 2 fois la même chose puisqu'on divise par 100 (pour avoir une fréquence) et qu'on multiplie par 100 (pour l'avoir en pourcentage).
Presque tous les marches aléatoires de 50 sauts atteignent la distance 5 au minimum (certaines vont plus loin).