Onze fiches de statistique

L'objectif du chapitre de statistique, en classe de seconde, est de sensibiliser les élèves à l'aléatoire et plus particulièrement de leur faire prendre conscience de l'existence de la fluctuation d'échantillonnage dans des cas simples : cette fluctuation est ici celle de distributions des fréquences entre des séries obtenues par répétition d'expériences identiques. La moyenne, la médiane, le maximum, le minimum se calculant à partir de cette distribution des fréquences sont aussi soumis à la fluctuation d'échantillonnage.

Les élèves devront se constituer un bagage d'expériences de référence : lancers de pièces, de dés, tirages de boules dans une urne, obtention de listes de nombres au hasard à l'aide d'une calculatrice ; ces expériences seront des éléments fondateurs de leur perception de l'aléatoire.

On simulera les expériences de référence et on pourra ainsi étudier de longues séries ; on simulera aussi de nouvelles situations à l'aide de listes de chiffres au hasard. Les outils de simulation sont puissants, mais pour les élèves une première étape importante est d'établir un lien entre l'expérience et une simulation de cette expérience : quand les thèmes s'y prêtent (et c'est le plus souvent le cas), la classe pourra être partagée en deux groupes : dans un groupe, les élèves feront effectivement des expériences, dans l'autre les élèves simuleront l'expérience.

Le langage utilisé est celui de la vie courante : égalité des chances, pièce ou dé équilibré, choix au hasard dans un ensemble fini, etc. Une série statistique dont les termes sont les résultats issus de la réplication d'une même expérience sera appelée un échantillon de cette expérience (ou plus simplement échantillon si cela ne prête pas à confusion) ; par exemple : les résultats de n lancers d'un dé, les couleurs de n boules tirées au hasard dans une urne, l'opinion de n personnes choisies au hasard dans une population bien définie (l'expérience que l'on répète est ici le recueil de l'opinion d'une personne choisie au hasard), la taille de n personnes choisies au hasard dans une population bien définie (l'expérience que l'on répète est ici la mesure de la taille d'une personne choisie au hasard), etc.

Les situations traitées mènent à l'observation de résultats qui appellent une explication ; par exemple, si on lance deux pièces équilibrées et qu'on regarde le nombre de *pile* obtenus, la distribution des fréquences de 0,1,2 semble voisine de (1/4,1/2,1/4) ; « pourquoi en est-il ainsi ? » est une question à adresser aux mathématiques, qui justifie de formaliser le langage utilisé et d'introduire à la théorie des probabilités. Le langage probabiliste permettra de se dégager de la polysémie des termes employés dans la vie courante (tels ceux de *chance* ou de *hasard*) ; des éléments simples de théorie des probabilités (introduits en classe de première) rendront intelligibles de nombreuses observations liées aux phénomènes aléatoires.

Le choix pédagogique de ce programme est d'aller de l'observation vers la conceptualisation et non d'introduire d'abord le langage probabiliste pour constater ensuite que tout se passe comme le prévoit cette théorie ; pour des enfants ou des adolescents, le langage des probabilités ne préexiste pas à la perception de l'aléatoire.

En seconde, les listes de chiffres au hasard produites par les ordinateurs sont présentées comme toute autre liste relevant de la répétition d'expériences identiques dont les issues ont des chances égales d'apparaître. Pour les élèves, lancer des pièces, des dés ou obtenir des listes de nombres au hasard avec une calculatrice sont des expériences aléatoires de référence. On simulera aussi quelques expériences simples : il est important que l'élève

conçoive lui-même de telles simulations (sans recherche systématique de subtilité algorithmique). En première, on précisera que pour simuler une expérience, il faut d'abord en établir un modèle puis simuler ce modèle.

Les fiches qui suivent couvrent la partie commune du programme et l'ensemble des thèmes proposés : il n'est donc pas question de toutes les traiter ! Elles indiquent, au-delà du contenu du programme, l'esprit dans lequel on peut le travailler. Ce ne sont pas des « fiches élèves », néanmoins quelques exemples de questions les concernant sont proposées (elles apparaissent en bleu). En en-tête de chaque fiche, on indique quelques-uns des éléments du programme abordés à l'occasion du thème considéré. Des aperçus théoriques à l'usage des enseignants sont donnés pour certains thèmes. Des résultats de simulation sont systématiquement fournis, que les enseignants pourront étudier avec leurs élèves, après que ceux-ci aient réalisé ou simulé quelques expériences. Ce sera l'occasion de manipuler des distributions de fréquences (c'est à dire des vecteurs ayant un nombre éventuellement assez grand de composantes) et de comprendre que pour regrouper des distributions de fréquences, il convient de faire des moyennes pondérées.

Le programme de statistique comporte dans son libellé deux parties : une première partie purement descriptive et la seconde partie qui parle de fluctuation d'échantillonnage ; comme on pourra le voir avec ces fiches, ces deux parties n'ont pas à être dissociées, les outils de description étant naturellement utilisés pour les séries statistiques, qu'elles soient réelles ou simulées.

Quelques adresses utiles :

Un logiciel pour la formation (articles, cours, simulations dynamique):
http://www.inrialpes.fr/sel/

Des simulations dynamiques des thèmes statistiques de seconde :
http://perso.wanadoo.fr/jpq

Des simulations dynamiques de quelques concepts de base en statistique :
http://www.kuleuven.ac.be/ucs/java/index.htm

Des simulations des thèmes statistiques de seconde avec le logiciel GEOPLANW2 et GEOSPACW:
http://www.irem.univ-mrs.fr

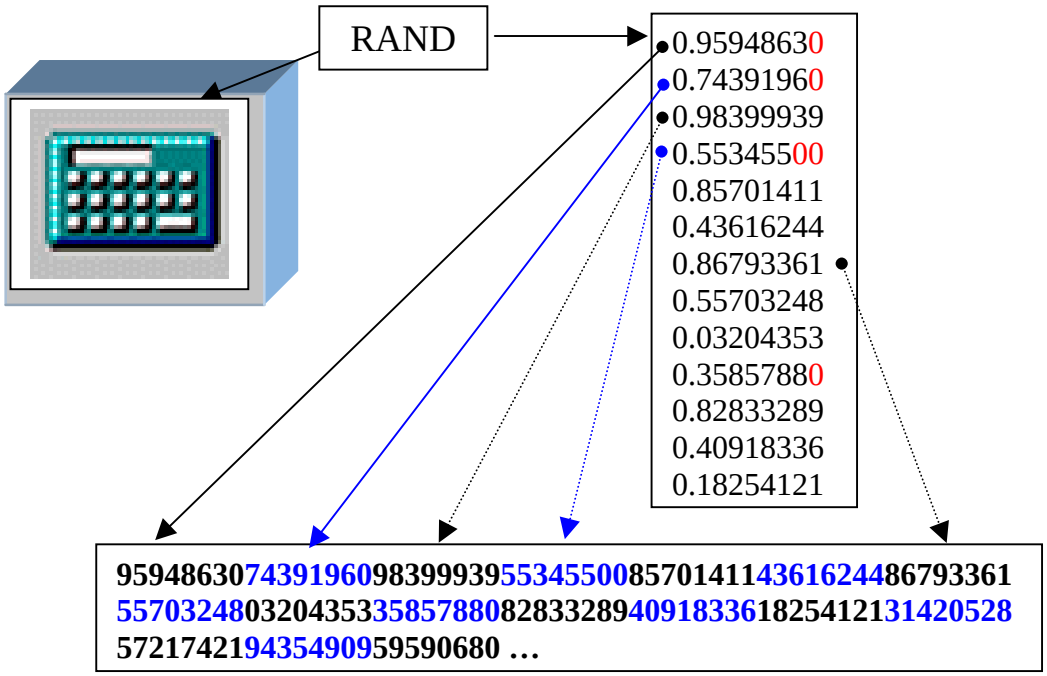
Liste de chiffres au hasard

Liste de chiffres au hasard
- Utilisation pour des simulations

1) Chiffres au hasard

Les calculatrices de poche et les ordinateurs possèdent une instruction dont le nom est *random*, *RDM*, *rand* ou *alea*. L'appel de cette touche, disons la touche *random*, fournit un nombre décimal compris entre 0 et 1 ; la partie décimale est une liste de k chiffres au hasard ; en faisant n appels de la touche *random*, on obtient une liste de kn chiffres au hasard. Une telle liste pourrait tout aussi bien résulter de kn tirages au hasard et avec remise dans une urne contenant en quantités égales des boules marquées 0 ou 1 ou 2,…, ou 9 ; au niveau de chaque choix, les chiffres ont des chances égales d'apparaître et les choix sont indépendants. Pour les élèves de seconde, la touche *random* fournit des chiffres au hasard de même que des lancers d'un dé équilibré fournissent des chiffres au hasard parmi les chiffres 1,2,3,4,5,6, ou que des lancers d'une pièce équilibrée produisent des listes de chiffres 0 ou 1 au hasard.

Certaines calculatrices affichant k chiffres après la virgule n'affichent cependant pas le dernier ou les derniers chiffres si ceux-ci valent 0 : il convient en ce cas de toujours compléter l'écriture par des 0 jusqu'à avoir k chiffres après la virgule.



2) Simulation de lancers de pièces ou de dés.

Pour simuler des lancers d'une pièce équilibrée avec une touche *random*, on codera les chiffres ; voici deux exemples de codages permettant de transformer une liste de chiffres au hasard en liste de lettres P, F choisies au hasard.

0	1	2	3	4	5	6	7	8	9
F	P	F	P	F	P	F	P	F	P

0	1	2	3	4	5	6	7	8	9
P	P	P	P	P	F	F	F	F	F

Si on voulait utiliser une *urne à chiffres*, (c'est à dire une urne contenant en quantités égales des boules marquées 0 ou 1 ou 2….ou 9) pour simuler des lancers d'un dé équilibré, on pourrait retirer de l'urne les boules marquées 0,7,8,9. S'il est impossible de retirer ces boules, on fera des tirages avec remise sans tenir compte des boules 0,7,8,9. Et on fait de même avec la touche *random* : on ne tient pas compte des chiffres qui n'existent pas dans le lancer d'un dé. Ainsi, a partir d'une liste de chiffres au hasard, on simule une série plus petite de lancers de dés :

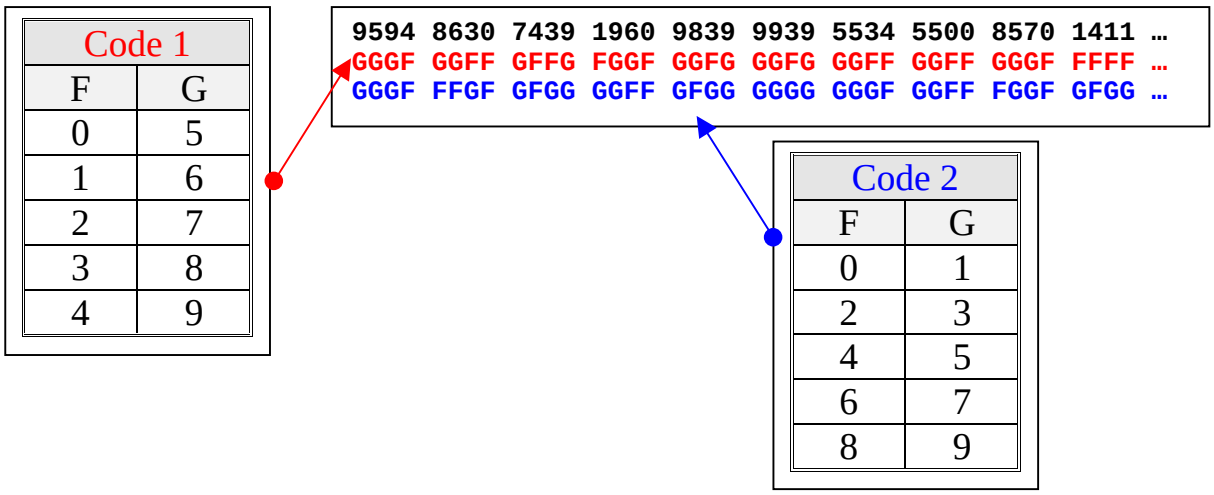
9	5	6	4	8	6	3	7	4	3	9	1	9	2	0	8	2	5	4	1
	5	6	4		6	3		4	3		1		2			2	5	4	1

Comparer le temps mis pour obtenir 100 résultats de lancer de dés et pour obtenir une simulation de même taille.

3) Simulation de familles de 4 enfants

Faisons les hypothèses suivantes :

- chaque naissance a autant de chances d'être celle d'un garçon ou celle d'une fille ,
- le sexe d'un enfant d'une famille ne dépend pas du sexe des enfants précédents.



4) Listes de nombres au hasard

On veut obtenir des listes de nombres au hasard entre 0 et 99 (resp. 0-999 et 0-9999) ; pour choisir au hasard dans 1-100, on ajoutera 1.

95948630743919609839993955345500857014114361624486793361
55703248032043533585788082833289409183361825412131420528
57217421943549095959068 0 …

95 94 86 30 74 39 19 60 98 39 99 39 55 34 55 00 85 70 14 11 43 61 62 44 86 79 33 61
55 70 32 48 03 20 43 53 35 85 78 80 82 83 32 89 40 91 83 36 18 25 41 21 31 42 05 28
57 21 74 21 94 35 49 09 59 59 06 80 … — 2

959 486 307 439 196 098 399 939 553 455 008 570 141 143 616 244 867 933 615 570
324 803 204 353 358 578 808 283 328 940 918 336 825 412 131 420 528 572 174 219
435 490 959 590 680 … — 3

9594 8630 7439 1960 9839 9939 5534 5500 8570 1411 4361 6244 8679 3361 5570 3248
0320 4353 3585 7880 8283 3289 4091 8336 1825 4121 3142 0528 5721 7421 9435 4909
5959 0680 … — 4

5) Utiliser les nombres au hasard pour faire des phrases aléatoires

0 ou 1 — **belle marquise**

2 ou 3 — **vos beaux yeux**

4 ou 5 — **me font**

6 ou 7 — **mourir**

8 ou 9 — **d'amour**

95948630743919609839993955345500857014114361624486793361
55703248032043533585788082833289409183361825412131420528
57217421943549095959068 0 …

95948630 | 74391 | 96098399939 5 | 53455008 57 | 0141143616244 8 | 67933615 |
5703248 | 03204353358 57 | 8808283328940918336 | 1825412131420528 57 | 217
4219 | 43549095959 06 | 80 …

95630 74391 96035 53087 04368 67315 57038 03487 80246 18257
21749 43906 80 …

95630

D'amour me font mourir vos beaux yeux belle marquise
ou
Mourir me font vos beaux yeux d'amour belle marquise ?

74391

exercices

Tous les tirages sont «avec remise».
Toutes les simulations seront faites en utilisant des tables de chiffres au hasard produits par une calculatrice ou un tableur. L'exercice consiste à écrire l'algorithme de simulation.

1) Simuler 10 lancers de 5 pièces équilibrées de couleurs différentes.
Simuler 10 lancers de 5 pièces équilibrées indiscernables.

Ces deux simulations consomment 50 chiffres au hasard. Pour la première, un résultat est une liste de 5 lettres P ou F choisies au hasard ; pour la seconde, un résultat est un nombre entre 0 et 5 qui est par exemple le nombre de piles.

2) Dans une urne, il y a 10 boules numérotées; simuler 20 tirages au hasard dans cette urne.
Dans une urne, il y a 94 boules numérotées; simuler 20 tirages au hasard dans cette urne.
Dans une urne, il y a 874 boules numérotées; simuler 20 tirages au hasard dans cette urne.

Pour la première simulation, on code les boules par les dix chiffres. Pour la deuxième, on prend les chiffres deux par deux et on ne tient pas compte de 00 et des entiers supérieurs ou égaux à 95. Pour la troisième, on prend les chiffres trois par trois et on ne tient pas compte de 0 et des entiers supérieurs ou égaux à 875.

3) Dans une urne, il y a 5 boules rouges et 5 boules noires.
Simuler les couleurs obtenues lors de 10 tirages au hasard dans cette urne.

Dans une urne, il y a 50 boules rouges et 50 boules noires.
Simuler les couleurs obtenues lors de 10 tirages au hasard dans cette urne.

Dans une urne, il y a 3645 boules rouges et 3645 boules noires.
Simuler les couleurs obtenues lors de 10 tirages au hasard dans cette urne.

Les trois simulations sont identiques puisqu'il y a autant de chances de tirer une boule rouge ou une boule blanche.

4) Dans une urne, il y a 73 boules rouges et 24 boules noires.
Simuler les couleurs obtenues lors de 10 tirages au hasard dans cette urne.

Dans une urne, il y a 7300 boules rouges et 2400 boules noires.
Simuler les couleurs obtenues lors de 10 tirages au hasard dans cette urne.

Les deux simulations sont identiques : on prendra les chiffres d'une liste de chiffres au hasard deux par deux, on ne tiendra pas compte du 00, 98,99. Les nombres de 1 à 73 seront codés rouge, les nombres de 74 à 97 seront codés noir.

5) Simuler 10 lancers d'une pièce déséquilibrée.

On pourrait simuler ce lancer si le fabriquant disait par exemple : à chaque lancer, on a trois fois plus de chances de tomber sur pile que sur face ; dans ce cas, sous réserve que ce que dit le fabriquant soit valide, on pourrait prendre les chiffres d'une liste au hasard deux par deux et coder par face les nombres de 00 à 74, par pile les autres.

6) A partir d'une liste de nombres choisis au hasard entre 0 et 99, comment construire une liste de nombres choisis au hasard parmi les nombres pairs compris entre 0 et 99 ?

Les élèves pour qui le résultat n'est pas intuitif pourront se reporter mentalement au tirage dans une urne à chiffres.

7) Si dans une liste de chiffres choisis au hasard, on élimine un chiffre sur deux, obtient-on encore une liste de chiffres au hasard ?

Les élèves pour qui le résultat n'est pas intuitif pourront se reporter mentalement au tirage dans une urne à chiffres.

8) Dans une liste de chiffres au hasard, on supprime un chiffre s'il est strictement inférieur au précédent ; obtient-on encore une liste de chiffres au hasard ?

Réfléchir sur la monotonie de la série obtenue.

De plus en plus de lancers d'un dé

- Echantillon de taille n d'une expérience.
- Notion de distribution de fréquences et de fluctuation d'échantillonnage.
- La distribution des fréquences de lancers d'un dé ***équilibré***, calculée sur une série de taille n varie d'une série à l'autre ; on observe que cette fluctuation est d'autant plus faible que n est grand ; quand n augmente, la distribution des fréquences se rapproche de (1/6, 1/6, 1/6, 1/6, 1/6, 1/6).
- La distribution des fréquences de lancers d'un dé ***non équilibré***, calculée sur une série de taille n, varie d'une série à l'autre ; on observe que cette fluctuation est d'autant plus faible que n est grand.

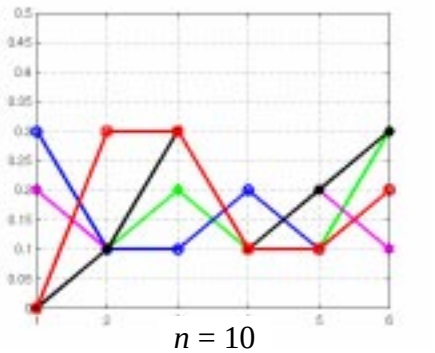
Dé équilibré

Une expérience consiste à lancer *n* fois un dé équilibré **:** les résultats constituent un échantillon de taille n de lancers d'un dé équilibré**.** La distribution de fréquences associée à cet échantillon est la liste des fréquences d'apparition de chacune des six faces. **A l'aide de listes de chiffres au hasard produites par un ordinateur, on simule 5 expériences pour chacune des valeurs de *n* suivantes : 10, 100, 1000 et 10000.**

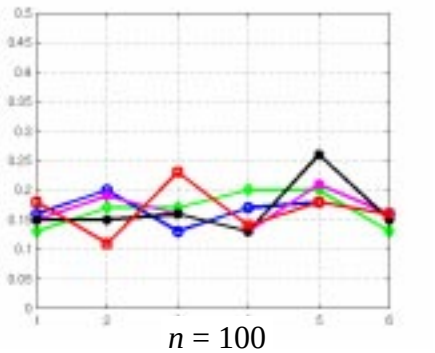
Les graphes ci-dessous représentent, pour chaque valeur de *n*, les 5 distributions de fréquences, avec en abscisse le numéro de face du dé et en ordonnée la fréquence associée.

On observe que :

- pour toute valeur de *n*, les distributions de fréquences des échantillons de taille n varient d'une expérience à l'autre : nous dirons qu'il y a **fluctuation d'échantillonnage**.
- lorsque *n* augmente, les distributions de fréquences « s'aplatissent » ; chaque fréquence semble se rapprocher de 1/6 (le 1/6 traduisant ici l'hypothèse que le dé est équilibré).

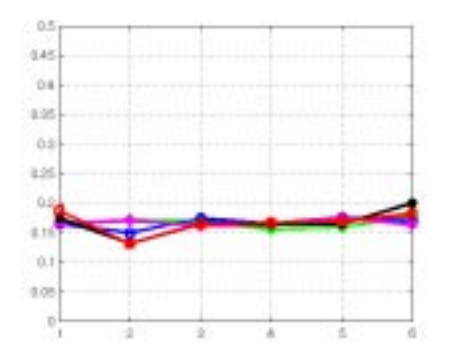


n = 10

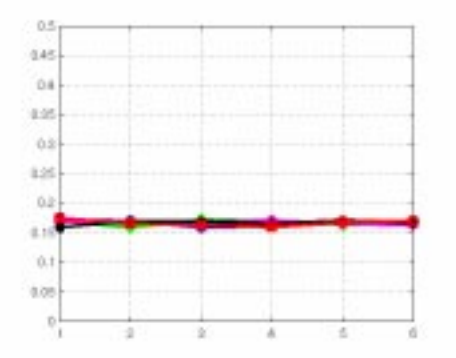


n = 100

Pour n=10, des représentations graphiques se chevauchent ; reconstituer les cinq distributions des fréquences.



n = 1000



n = 10000

Dé truqué

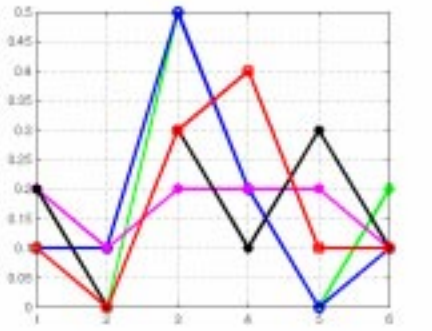
Une expérience consiste à lancer *n* fois un dé vendu comme dé truqué (i.e. à chaque lancer, toutes les faces n'ont pas les mêmes chances d'être observées) ; les résultats constituent un échantillon de taille n du lancer de ce dé**.** La distribution de fréquences associée à cet échantillon est la liste des fréquences d'apparition de chacune des six faces.

On fait 5 expériences dans lesquelles *n* prend successivement les valeurs 10, 100.

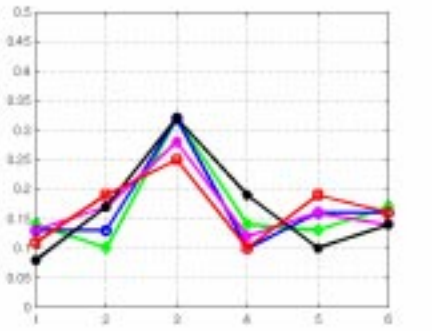
Les graphes ci-dessous représentent, pour différentes valeurs de *n*, les 5 distributions des fréquences, avec en abscisse le numéro de face du dé et en ordonnée la fréquence associée.

On observe que :

- pour toute valeur de *n*, les distributions de fréquences varient d'une expérience à l'autre : nous dirons qu'il y a **fluctuation d'échantillonnage**.
- on observe que pour n=10, les distributions des fréquences pour ce dé et pour un dé équilibré (cf. figure correspondante pour n=10) ne sont pas sensiblement différentes.
- Pour n=100, le déséquilibre commence à se voir.



n = 10

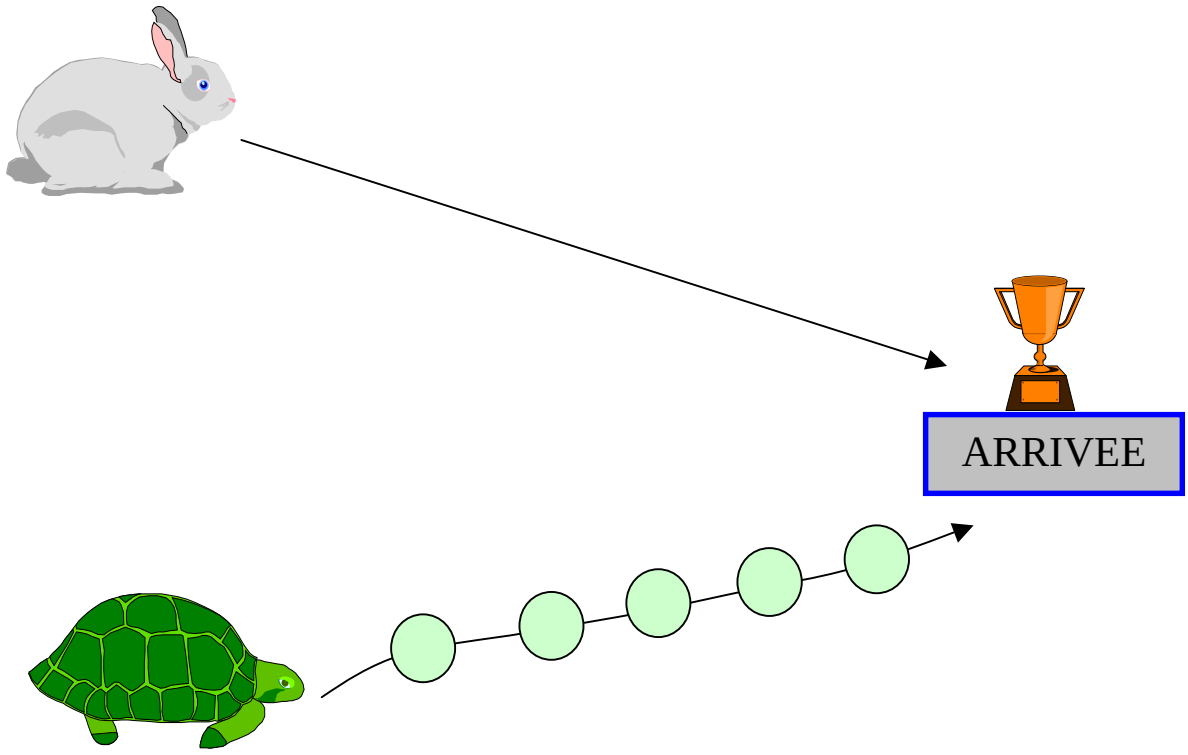


n = 100

Pour n=10, des représentations graphiques se chevauchent ; reconstituer les cinq distributions de fréquences.

Le lièvre et la tortue

- Expérimentation et simulation.
- Opérations sur les effectifs et sur les fréquences.
- Un exemple de jeu non équitable.
- La fluctuation d'échantillonnage diminue avec la taille des séries observées.



Une partie du jeu du lièvre et de la tortue se déroule ainsi :

1. **On lance un dé.**

   - ***Si le dé tombe sur 1, 2, 3, 4 ou 5 la tortue avance d'une case.***
     Elle a 5 cases à franchir avant d'atteindre l'arrivée.
     La partie est alors terminée, la tortue a gagné.

   - ***Si le dé tombe sur 6 le lièvre atteint directement l'arrivée.***
     La partie est alors terminée, le lièvre a gagné.

2. **La partie continue jusqu'à ce qu'il y ait un gagnant.**

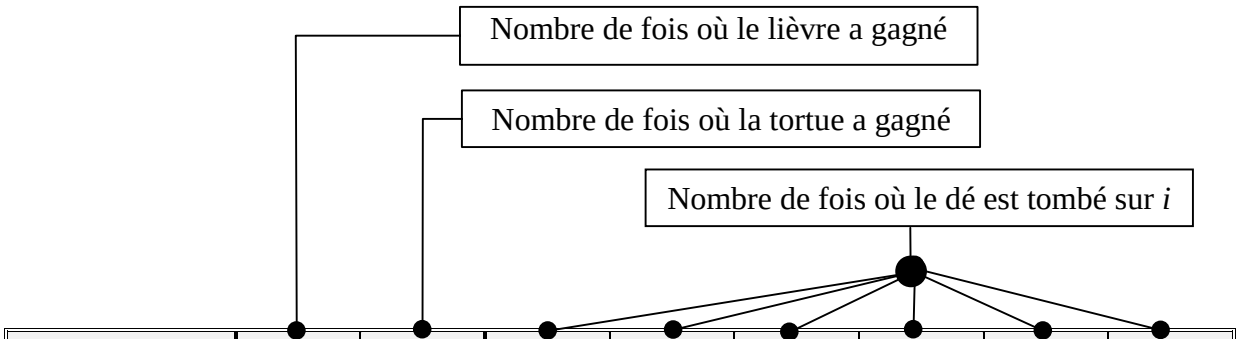
Quelle est la situation la plus enviable : celle du lièvre ou celle de la tortue ?

1) Un jeu est composé de 10 parties

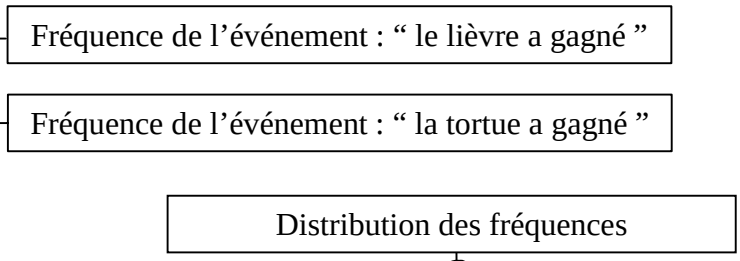
Pour un jeu, on recueille des données sous forme d'effectifs et on calcule des distributions de fréquences.

	(le lièvre, la tortue)	Les lancers de dés
Effectifs	$(n, 10 - n)$	$(n_1, n_2, n_3, n_4, n_5, n_6)$
Distribution des fréquences	$(f, 1 - f)$	$(f_1, f_2, f_3, f_4, f_5, f_6)$

Voici les résultats obtenus par dix élèves qui ont simulé un jeu.



	n	$10 - n$	n_1	n_2	n_3	n_4	n_5	n_6
Elève 1	7	3	5	6	7	3	6	7
Elève 2	6	4	2	7	10	9	11	6
Elève 3	6	4	4	8	6	9	6	6
Elève 4	6	4	8	7	8	8	6	6
Elève 5	5	5	8	10	10	5	7	5
Elève 6	8	2	7	3	9	2	4	8
Elève 7	7	3	3	5	11	13	9	7
Elève 8	7	3	5	2	5	8	9	7
Elève 9	5	5	9	5	9	9	9	5
Elève 10	4	6	4	10	13	10	4	4
Total	**61**	**39**	**55**	**63**	**88**	**76**	**71**	**61**

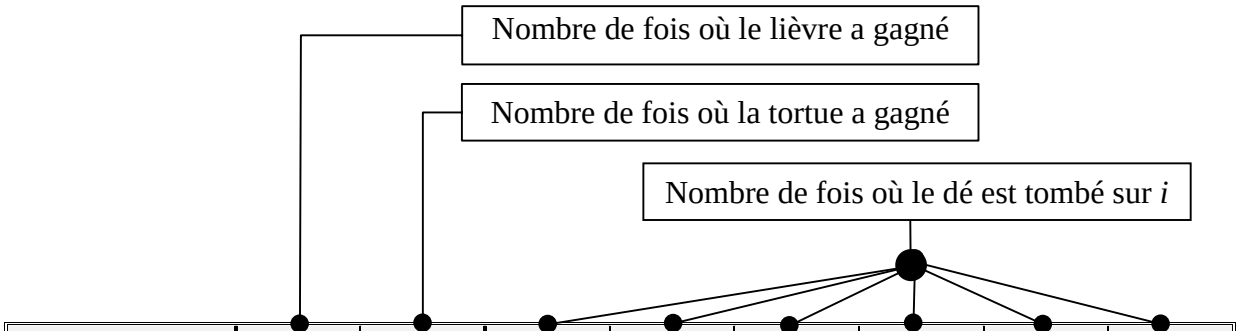


	f	1 - f	f_1	f_2	f_3	f_4	f_5	f_6
Elève 1	0.7	0.3	0.1471	0.1765	0.2059	0.0882	0.1765	0.2059
Elève 2	0.6	0.4	0.0444	0.1556	0.2222	0.2000	0.2444	0.1333
Elève 3	0.6	0.4	0.1026	0.2051	0.1538	0.2308	0.1538	0.1538
Elève 4	0.6	0.4	0.1860	0.1628	0.1860	0.1860	0.1395	0.1395
Elève 5	0.5	0.5	0.1778	0.2222	0.2222	0.1111	0.1556	0.1111
Elève 6	0.8	0.2	0.2121	0.0909	0.2727	0.0606	0.1212	0.2424
Elève 7	0.7	0.3	0.0625	0.1042	0.2292	0.2708	0.1875	0.1458
Elève 8	0.7	0.3	0.1389	0.0556	0.1389	0.2222	0.2500	0.1944
Elève 9	0.5	0.5	0.1957	0.1087	0.1957	0.1957	0.1957	0.1087
Elève 10	0.4	0.6	0.0889	0.2222	0.2889	0.2222	0.0889	0.0889
100 parties	**0.61**	**0.39**	**0.1356**	**0.1504**	**0.2116**	**0.1788**	**0.1713**	**0.1524**

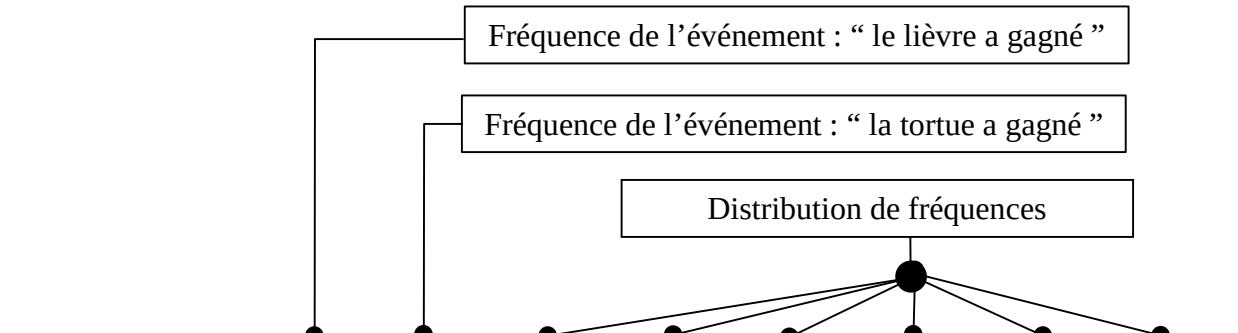
Représenter graphiquement les distributions de fréquences des élèves 1 et 2 et celle qui correspond aux 100 parties. On explicitera les calculs relatifs à la dernière ligne du tableau.

Mais le tableau ci-dessus concerne 100 parties et on peut se demander si avec cent autres parties, les résultats ne seraient pas nettement différents. Nous allons donc faire plus de simulations.

2) Un jeu est une suite de 1000 parties et sera donc simulée.



	n	$1000-n$	n_1	n_2	n_3	n_4	n_5	n_6
Jeu 1	672	328	674	645	646	667	682	672
Jeu 2	659	341	661	646	633	682	705	659
Jeu 3	666	334	671	643	631	657	672	666
Jeu 4	674	326	627	668	656	689	648	674
Jeu 5	671	329	670	671	660	634	661	671
Jeu 6	682	318	658	646	639	615	647	682
Jeu 7	667	333	688	657	643	637	729	667
Jeu 8	683	317	628	732	630	671	651	683
Jeu 9	656	344	678	639	703	651	650	656
Jeu 10	671	329	637	649	647	670	685	671
Total	**6701**	**3299**	**6592**	**6596**	**6488**	**6573**	**6730**	**6701**



	f	$1 - f$	f_1	f_2	f_3	f_4	f_5	f_6
Jeu 1	0.672	0.328	0.1691	0.1618	0.1621	0.1673	0.1711	0.1686
Jeu 2	0.659	0.341	0.1658	0.1621	0.1588	0.1711	0.1769	0.1653
Jeu 3	0.666	0.334	0.1703	0.1632	0.1602	0.1668	0.1706	0.1690
Jeu 4	0.674	0.326	0.1583	0.1686	0.1656	0.1739	0.1636	0.1701
Jeu 5	0.671	0.329	0.1689	0.1691	0.1664	0.1598	0.1666	0.1691
Jeu 6	0.682	0.318	0.1693	0.1662	0.1644	0.1582	0.1665	0.1755
Jeu 7	0.667	0.333	0.1711	0.1634	0.1599	0.1584	0.1813	0.1659
Jeu 8	0.683	0.317	0.1572	0.1832	0.1577	0.1680	0.1630	0.1710
Jeu 9	0.656	0.344	0.1705	0.1607	0.1768	0.1637	0.1634	0.1649
Jeu 10	0.671	0.329	0.1609	0.1639	0.1634	0.1692	0.1730	0.1695
Moyenne	**0.6701**	**0.3299**	**0.1661**	**0.1662**	**0.1635**	**0.1656**	**0.1696**	**0.1689**

- **On observe que les distributions varient moins d'une ligne à l'autre pour ce tableau que dans le cas où un jeu comporte 10 parties.**
- **Le tableau ci-dessus totalise 10 000 parties : c'est déjà beaucoup. On constate que le lièvre gagne plus souvent, et que sur les 10 000 parties simulées, il gagne dans 67% des cas.**
- **D'où sort ce 67% ? On conçoit que ce pourcentage ne dépend que du fait que le dé est équilibré, donc calculable à partir de cette hypothèse, mais comment faire ce calcul ?**

Expliquer le phénomène observé, c'est démontrer dans un cadre théorique que le lièvre gagne et calculer ses chances théoriques de gagner : mais pour cela, il faut donner une définition mathématique des hypothèses faites, à savoir que le dé est équilibré et les lancers indépendants les uns des autres.
La théorie qui permet de démontrer tout cela est la théorie des probabilités. On montre que la « chance théorique » du lièvre est 1 - (5/6)^6, soit à peu près 0,665.
Voici de nouvelles simulations ; on pourra observer que quand n augmente, les fréquences se rapprochent des valeurs théoriques.

NOMBRE DE PARTIES

VALEURS THEORIQUES

Pour f : $1 - (5/6)^6 \approx$ **0.6651**
Pour 1- f : $(5/6)^6 \approx$ **0.3349**
Pour les f_i : $1/6 \approx$ **0.1667**

	f	$1 - f$	f_1	f_2	f_3	f_4	f_5	f_6
10	**0.6100**	**0.3900**	**0.1356**	**0.1504**	**0.2116**	**0.1788**	**0.1713**	**0.1524**
1000	**0.6701**	**0.3299**	**0.1661**	**0.1662**	**0.1635**	**0.1656**	**0.1696**	**0.1689**
10000	**0.6667**	**0.3333**	**0.1667**	**0.1671**	**0.1660**	**0.1659**	**0.1668**	**0.1675**
100000	**0.6651**	**0.3349**	**0.1666**	**0.1667**	**0.1667**	**0.1669**	**0.1664**	**0.1671**

- On peut s'interroger sur la durée moyenne d'une partie (nombre de lancers de dés utilisés pour cette partie).

On pourra faire quelques remarques diverses ; par exemple :
- Si la tortue gagne, la partie dure 6 coups.
- Il y a une chance sur 6 que la partie dure 1 coup.
- Parmi les parties de 6 coups, le lièvre en gagne environ une sur six.

On n'a pas recueilli les durées des parties : on ne peut pas calculer la distribution des fréquences de cette durée ; en statistique, on a toujours intérêt à réfléchir à tout ce qui peut être intéressant avant de faire des expériences, pour ne pas risquer d'avoir à tout recommencer. On peut tout de même calculer des durées moyennes : pour les 100 parties du premier tableau, la durée moyenne est 4,14 et pour les 10 000 parties simulées, la durée moyenne est 3,958. Là aussi, la seule hypothèse faite concerne les lancers de dés et il devrait y avoir une formule qui permette de calculer une valeur théorique μ pour cette moyenne. Cette formule existe bien et se démontre par le calcul des probabilités :

$$\mu = \sum_{i=1}^{i=5} i \times \frac{5^{i-1}}{6^i} + 6 \times \frac{5^5}{6^5} \approx 3{,}99$$

- On peut modifier le jeu, c'est-à-dire changer le nombre de cases pour la tortue et faire franchir des cases au lièvre, en nombre inférieur, supérieur ou identique à celles de la tortue et étudier les distributions de fréquences (f, 1 - f) correspondantes.

Des chances inégales

- Fluctuation d'échantillonnage.
- Les distribution des fréquences calculées sur des échantillons de taille n fluctuent d'autant moins que n est grand.
- Une expérience dont les issues n'ont pas des chances égales d'apparaître.

Une expérience consiste à lancer deux pièces équilibrées et à compter le nombre de *pile*.

Un jeu est une suite de n=10 (resp.100) expériences.

Ci-dessous, on a cumulé les résultats de trois élèves ayant simulé des lancers de pièces ; on a donc un résultat sur 30 expériences pour n=10 (resp. sur 300 expériences pour n=100).

n = 10	Nombre de piles		
	0	1	2
Elève 1	2	4	4
Elève 2	4	5	1
Elève 3	1	7	2
30 expériences	***7***	***16***	***7***

n = 100	Nombre de piles		
	0	1	2
Elève 1	0.29	0.46	0.25
Elève 2	0.22	0.54	0.24
Elève 3	0.20	0.52	0.28
300 expériences	***0.24***	***0.51***	***0.25***

Calculer la distribution des fréquences du nombre de pile sur les 330 expériences totalisées dans les deux tableaux ci-dessus (attention, les données inscrites dans les deux tableaux ne sont pas de même nature : dans le premier cas, la taille de la série est petite et l'habitude est en ce cas de présenter les résultats en effectifs plutôt qu'en fréquences).

Certains s'étonnent de ces résultats et pensent qu'il faut aller plus loin ; voici donc des simulations de plus grandes tailles, pour n=1000, 10 000.

n = 1000		Nombre de piles		
		0	1	2
Simulation 1	Fréquence	0.252	0.494	0.254
Simulation 2	Fréquence	0.249	0.483	0.268
Simulation 3	Fréquence	0.255	0.492	0.253
3000 expériences		***0.252***	***0.48967***	***0.25833***

n = 10 000		Nombre de piles		
		0	1	2
Simulation 1	Fréquence	0.2426	0.5041	0.2533
Simulation 2	Fréquence	0.2457	0.4995	0.2548
Simulation 3	Fréquence	0.2601	0.4889	0.251
300 000 expériences		***0.24947***	***0.4975***	***0.25303***

Pour n grand, on observe que les distributions des fréquences sont proches de (0.25 , 0.50 , 0 .25). Pourquoi en est-il ainsi ? La seule hypothèse faite est que les deux pièces sont équilibrées : peut-on alors essayer d'expliquer l'observation faite ?

POUR ALLER PLUS LOIN

- Que peut-on dire de la distribution de fréquences du nombre de piles lorsque 3 pièces sont lancées un grand nombre de fois ?
- Etudier la distribution de fréquences du nombre de filles dans les familles de deux enfants (resp. trois ou quatre). On fait les hypothèses suivantes :

- chaque naissance a autant de chances d'être celle d'un garçon ou celle d'une fille,
- le sexe d'un enfant d'une famille ne dépend pas du sexe des enfants précédents.

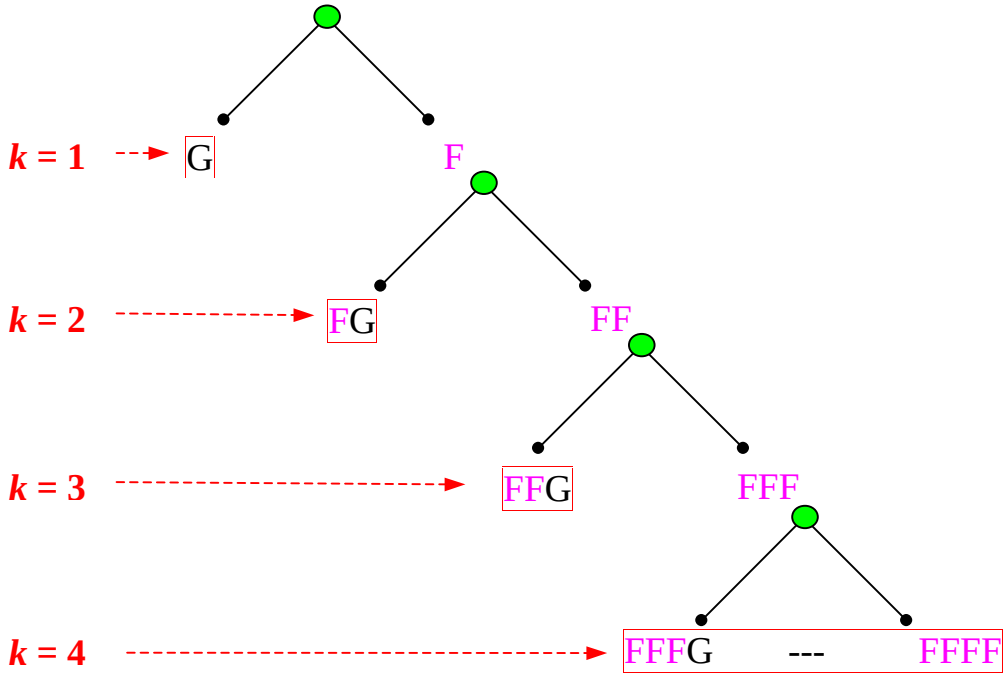
Politique nataliste

- Simuler une expérience.
- Fluctuation d'échantillonnage

Supposons qu'une politique nataliste soit mise en place à partir de la règle suivante. Les naissances au sein d'une famille s'arrêtent :

- soit à la naissance du premier garçon,
- soit lorsque la famille comporte quatre enfants.

On peut représenter ce processus par l'arbre suivant, où k est le nombre d'enfants de la famille :



Quelle sera l'influence de cette politique nataliste sur la répartition entre les sexes ?

Certains répondent qu'il y aura plus de garçons puisque environ 50% des familles auront un garçon et pas de filles ; d'autres répondent qu'il y aura plus de filles à cause des familles de plus de deux enfants ; d'autres n'ont pas d'opinion.

Nous allons simuler de telles familles, en faisant l'hypothèse que :

- chaque naissance a autant de chances d'être celle d'un garçon ou celle d'une fille,
- le sexe d'un enfant d'une famille ne dépend pas du sexe des enfants précédents.

Comme beaucoup se demandent si le fait de limiter à 4 enfants a une influence, nous allons faire plusieurs simulations avec diverses valeurs du nombre maximum n d'enfants et regarder ce qui se passe à la première génération qui suit l'application de cette politique nataliste.

n= 4

G|FG|FG|G|FFG|FG|G|G|G|FFG|G|FG|FFG|FFFF|G|FFFG|FG|FFG|G|FG|G|
FG|FG|G|G|FFFF|G|FG|FG|FG|G|G|FG|G|G|G|G|FFFG|G|FG|G|FG|G|FG|G|
G|G|G|G|FG|FFG|G|FFG|FFG|FFFG|G|FG|FFFG|G|G|G|FFG|G|FG|G|FG|FF
FG|FG|G|G|G|FFFG|FG|G|FG|FFG|G|FG|G|FFG|G|FG|FFFF|FFG|FFG|G|FF
FG|FFG|G|G|FFFF|FG|G|G|FG|FFG|G|FFG|FG|G|

n = 5

FFFFG|FG|FFG|FFFFF|G|FFFG|G|FFG|FG|G|G|G|FFG|FG|G|G|FFFFF|G|FG
|FFFFG|FFFG|G|G|G|FFG|G|FFFG|G|G|G|FG|G|FG|G|G|FG|G|FG|FG|G|FF
G|G|FFG|G|G|G|G|FFG|G|G|G|FFG|FFG|G|G|G|G|FFG|FG|FG|G|G|G|G|FG|
G|FG|G|G|FG|G|G|FG|FG|FG|FG|FFFFG|G|FG|G|G|G|G|FFG|G|FFFFF|G|G
|FFG|FG|G|G|FG|G|G|FFFFG|G|FG|FG|G|

n = 6

FG|G|G|FFFFFF|G|FFG|FG|FG|G|FG|G|FFFFFF|G|G|FFFFFG|G|FFG|G|G|F
G|FG|G|FFG|G|G|FFFFFF|G|G|G|G|FG|G|G|FG|G|G|G|FFFG|FFG|G|FFG|F
G|G|G|FFFG|G|FFG|FFFFG|FG|FG|FG|FG|G|FG|FFFG|FFG|FFFG|G|G|G|F
FFFG|G|FG|G|G|FG|FG|G|FG|G|G|FG|FG|G|G|G|G|FFG|FG|G|FFFFG|FG|G
|G|FFFFG|FG|G|G|G|G|FG|G|FFG|G|G|G|G|FFFFFG|

A partir de ces distributions, on peut recenser les effectifs des garçons, des filles ainsi que le nombre de familles ayant k enfants, pour 1;k;n, puis la distribution des fréquences de ces mêmes grandeurs.

RESULTATS POUR 100 FAMILLES OBSERVEES

	Effectifs						Total	Garçons	Filles	Total
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$				
$n = 4$	47	27	15	11			100	96	94	190
$n = 5$	55	23	12	3	7		100	97	87	184
$n = 6$	54	24	9	4	4	5	100	97	98	195

	Distributions de fréquences						Garçons	Filles
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$		
$n = 4$	0.47	0.27	0.15	0.11			0.505	0.495
$n = 5$	0.55	0.23	0.12	0.03	0.07		0.527	0.473
$n = 6$	0.54	0.24	0.09	0.04	0.04	0.05	0.497	0.503

Il semblerait que la répartition des sexes reste équilibrée. Mais on peut considérer que 100 familles, c'est peu et il se peut que la fluctuation d'échantillonnage ne permette pas, pour cette taille, de voir une éventuelle dissymétrie ; imaginons qu'on fasse la simulation relative à 10^6 familles : on *consommerait* pour cela une liste de N lettres F ou G choisies au hasard ($N>10^6$) ; pour compter le nombre de garçons dans l'ensemble de ces familles, on compterait le nombre de G dans la liste de N lettres. Or, comme c'est une liste résultant de choix au hasard, la fréquence des G sera proche de 0,5 : il est inutile de réaliser une simulation à grande échelle pour conclure que la répartition des sexes est inchangée.

Qu'observez-vous des distributions des fréquences du nombre d'enfants par famille ?
Calculer pour les trois lignes de ce tableau la moyenne, la médiane et le mode du nombre d'enfants par famille.

Les nombres moyens d'enfants observés dans cette simulation sont respectivement $m_4=1,90$ $m_5=1,84$, $m_6=1,95$. On pourrait s'attendre à ce que $m_6>m_5>m_4$. S'agit-il là d'un tour joué par la fluctuation d'échantillonnage ? On pourrait refaire d'autres simulations. Mais on peut démontrer, dans le cadre de la théorie des probabilités, une formule donnant la valeur théorique du nombre moyen d'enfants pour les trois cas envisagés:

$$\mu_4 = \frac{1}{2}+\frac{2}{4}+\frac{3}{8}+\frac{4}{8} \qquad \mu_5 = \frac{1}{2}+\frac{2}{4}+\frac{3}{8}+\frac{4}{16}+\frac{5}{16} \qquad \mu_6 = \frac{1}{2}+\frac{2}{4}+\frac{3}{8}+\frac{4}{16}+\frac{5}{32}+\frac{6}{32}$$

On pourra comparer ces valeurs théoriques entre elles.

Faites vos jeux

- Tenir compte de l'ordre dans des suites de chiffres 0 ou 1 au hasard.
- Formuler une question.
- Conception et écriture d'un algorithme.

En regardant 100 jeux consécutifs d'une roulette de casino, un joueur observe que dans la série des résultats, il y a 6 fois consécutivement la même couleur.

Une suite de résultats consécutifs aussi longue est-elle exceptionnelle ?

La locution « aussi longue » dans cette question mérite réflexion ; si on pense que 6 coups consécutifs égaux, cela fait beaucoup, on penserait a fortiori que 7 coups consécutifs égaux, ou plus, ce serait beaucoup ; aussi on ne se focalise pas sur 6, mais sur le fait d'observer au moins 6 coups consécutifs égaux.

Pour donner des éléments de réponse à cette question, considérons la situation suivante :

On effectue r tirages à la roulette, on note **N** pour noir et **R** pour rouge (le zéro n'est pas pris en compte).
Pour compter le nombre maximum m de coups consécutifs égaux, on fabrique un compteur : à partir d'une liste $x_1,.., x_r$ de lettre R ou N, on construit une liste $y_1,.., y_r$ par l'algorithme suivant :

Entrées : r ; $(x_1,.., x_r)$
$y_1 = 1$
Pour i=2 à r :
    Si $x_i=x_{i-1}$ alors $y_i=y_{i-1}+1$
    Sinon $y_i=1$
Résultat : $m=\max(y_1,.., y_r)$.

Par exemple :

R	R	R	N	R	N	N	R	N	R	R	N	N	R	N
1	2	3	1	1	1	2	1	1	1	2	1	2	1	1

Une expérience consiste à effectuer r tirages à la roulette et à calculer le nombre maximum de coups consécutifs égaux, noté m (a priori, m est un entier compris entre 1 et r).

Décrire les séries de résultats qui réalisent m=1 ou m=r .

n simulations pour r=10

m	1	2	3	4	5	6	7	8	9	10
n=10	0	5	1	4	0	0	0	0	0	0
n=100	0	11	37	20	21	9	2	0	0	0
n=1000	1	160	350	253	126	68	31	6	2	3
n=1000	1	186	347	258	121	54	24	7	1	1
n=1000	5	147	344	273	133	62	25	5	4	2
n=1000	4	163	357	239	141	62	23	7	3	1

Pour chaque ligne, déterminer la médiane, la moyenne, le mode, puis l'étendue.

Déterminer la médiane, la moyenne, le mode puis l'étendue pour les 4000 expériences obtenues avec les quatre dernières lignes.

n simulations pour r=100

m	≤ 4	5	6	7	8	9	10	11	12	13	≥14	max
n=1000	23	150	306	220	137	77	40	32	10	3	2	17
n=1000	21	164	240	255	146	85	47	26	9	6	1	15

Pour chaque ligne, peut-on déterminer la moyenne, la médiane, le mode, l'étendue ?

Déterminer la médiane et le mode pour les 2000 expériences obtenues avec les deux dernières lignes. Quel pourcentage d'expériences parmi ces 2000 réalisent l'événement $m<6$?

La question posée au début de cette fiche concerne l'événement $\{m\geq 6\}$. Or sur les 2000 simulations, environ 82% réalisent cet événement : on ne peut pas dire qu'un tel événement est exceptionnel !

La fréquence observée est bien sûr soumise à la fluctuation d'échantillonnage ; le calcul des probabilités permet de montrer que pour n expériences, lorsque n augmente, la fréquence observée se rapproche d'un nombre p, où $p\approx 0{,}8$.

APERÇU THEORIQUE

Certaines situations, pour lesquelles on ne sait faire les calculs théoriques, peuvent être simulées : les élèves de seconde, ne pouvant pas maîtriser les concepts et les calculs permettant d'apporter une solution à la question posée en début de cette fiche, peuvent donc utiliser des simulations. Connaître la théorie des chaînes de Markov n'est ainsi pas indispensable pour travailler cette fiche : c'est néanmoins un concept qui permet de répondre à cette question par la théorie et de dépasser ainsi le cas particulier d'une simulation, même si celle-ci est de grande taille.

On s'intéresse à la probabilité d'avoir observé au moins six coups consécutifs égaux lors d'une suite de lancers d'une pièce équilibrée. Pour cela considérons une suite de variables aléatoires indépendantes, $(X_r)_{r?1}$, de loi (1/2,1/2) sur {0,1}. On construit une nouvelle suite $(Y_r)_{r?1}$, où :

$Y_1 = 1$

Pour r>1 : Si $Y_{r-1} = 6$, alors $Y_r = 6$

Sinon : si $X_r = X_{r-1}$ alors $Y_r = Y_{r-1}+1$, sinon : $Y_r = 1$.

La suite $(Y_r)_{r?1}$ est une chaîne de Markov dont la matrice de transition est :

$$K = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Le terme k_{ij} de cette matrice, situé au croisement de la ligne i et de la colonne j, est la probabilité que Y_r soit égal à j sachant que Y_{r-1} vaut i, pour tout r, r>1, ce qui s'écrit aussi :

$$k_{ij} = P(Y_r = j / Y_{r-1} = i) \text{ , } 1;i;6,\ 1;j;6.$$

La théorie des chaînes de Markov montre que la loi P_r de Y_r est donnée par :

$$P_r = (K^*)^{r-1} P_1$$

où K* est la matrice transposée de K , P_r est la matrice unicolonne dont le terme de la ligne j est la probabilité que Y_r soit égal à j ; P_1 est la matrice unicolonne dont le premier terme vaut 1 (on a posé Y_1=1), les autres composantes étant nulles.

La probabilité d'avoir observé au moins 6 coups consécutifs égaux lors de r lancers d'une pièce équilibrée est la sixième composante de P_r, notée ici $p_r(6)$. On trouve les résultats suivants :

r	10	50	100	200	300
$P_r(6)$	0,094	0,544	0,807	0,965	0,994

Jeu de pile ou face

- Analyser une situation pour choisir une stratégie.
- Simuler à grande échelle
- Observer une distribution des fréquences et faire des conjectures.

On joue à pile ou face avec la règle du jeu suivante :

- Une somme s est misée.
- Si pile sort, le gain est égal à 2s.
- Si face sort, la mise s est perdue.

Un joueur dispose de S=1000 francs. Il hésite entre les deux stratégies suivantes :

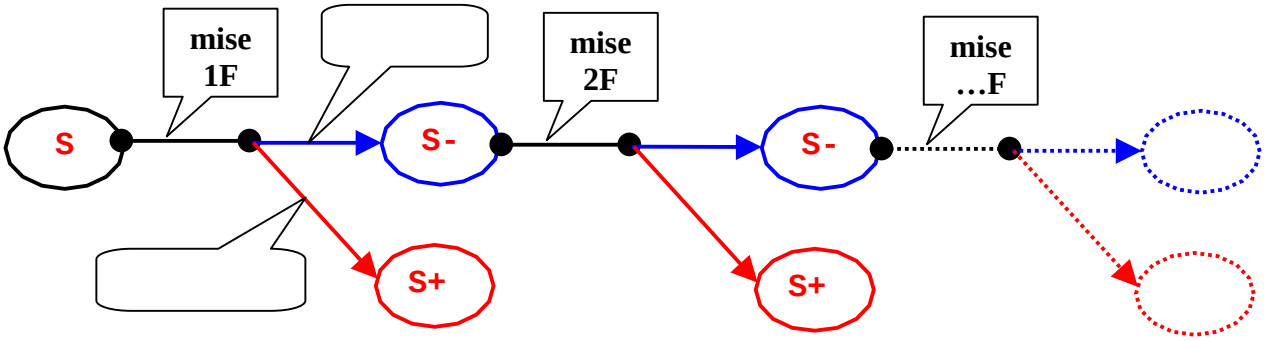
Stratégie1 : Miser 1 franc pour commencer.
Jouer tant que l'on perd et qu'il reste de l'argent,
en doublant la mise à chaque partie.
S'arrêter à la première partie gagnée ou lorsqu'on ne peut plus miser.

Stratégie 2 : Miser 1 franc pour commencer.
Jouer tant que l'on perd et qu'il reste de l'argent,
en triplant la mise à chaque partie.
S'arrêter à la première partie gagnée ou lorsqu'on ne peut plus miser.

La durée d'une partie est le nombre de fois où on a lancé une pièce.

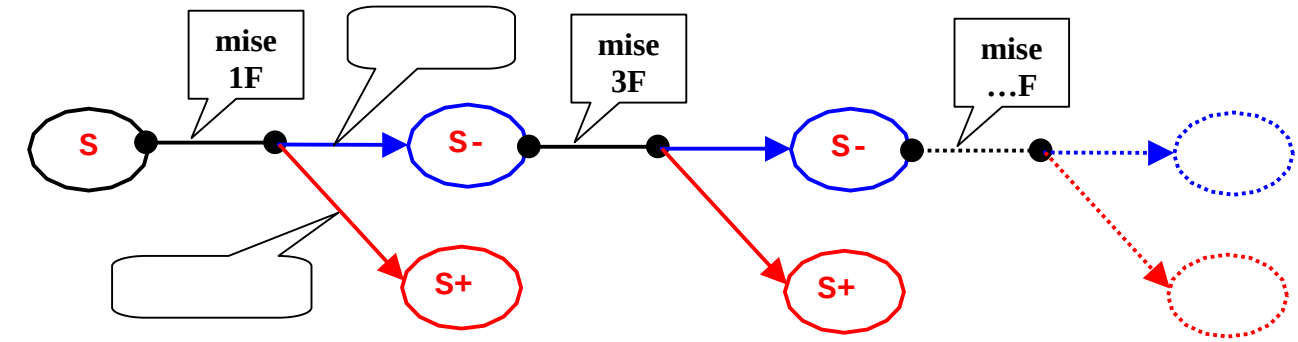
Quelle stratégie choisissez-vous ?

STRATEGIE 1
(S1)



Simulation avec 100 000 parties											
durée	1	2	3	4	5	6	7	8	9	10 gagné	10 perdu
Effectif	50223	24933	12400	6143	3075	1577	788	443	198	114	106
Fréquence	0.50223	0.24933	0.12400	0.6143	0.03075	0.01577	0.00788	0.00443	0.00198	0.00114	0.00106

STRATEGIE 2
(S2)



Simulation avec 100 000 parties								
durée	1	2	3	4	5	6	7 gagné	7 perdu
Effectif	50223	24933	12400	6143	3075	1577	788	861
Fréquence	0.50223	0.24933	0.12400	0.6143	0.03075	0.01577	0.00788	0.00861

• Parmi les critères de choix, il y a entre autres :

- la stratégie dont le gain moyen, calculé sur les 100 000 parties simulées est maximum ;
- la stratégie qui permet éventuellement de jouer le plus longtemps ;
- la stratégie où la fréquence de perdre, calculée sur 100 000 parties, est minimale ;
- la stratégie où le gain maximum possible est le plus élevé ;
- la stratégie où le bénéfice maximum possible est le plus élevé.

On pourra faire construire les tableaux suivants :

Stratégie 1										
durée du jeu	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
mise à la partie	1	2	4	8	16	32	64	128	256	512
gain à la partie	2	4	8	16	32	64	128	256	512	1024
total engagé	1	3	7	15	31	63	127	255	511	1023
bénéfice total	1	1	1	1	1	1	1	1	1	1

Stratégie 2										
durée du jeu	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
mise à la partie	1	3	9	27	81	243	729	-	-	-
gain à la partie	2	6	18	54	162	486	1458	-	-	-
total engagé	1	4	13	40	121	384	972	-	-	-
bénéfice total	1	2	5	14	41	122	586	-	-	-

Remarque : en première, à l'occasion de l'étude des suites géométriques, on pourra établir les formules permettant de calculer directement les valeurs des cases de ces tableaux.

Promenades aléatoires

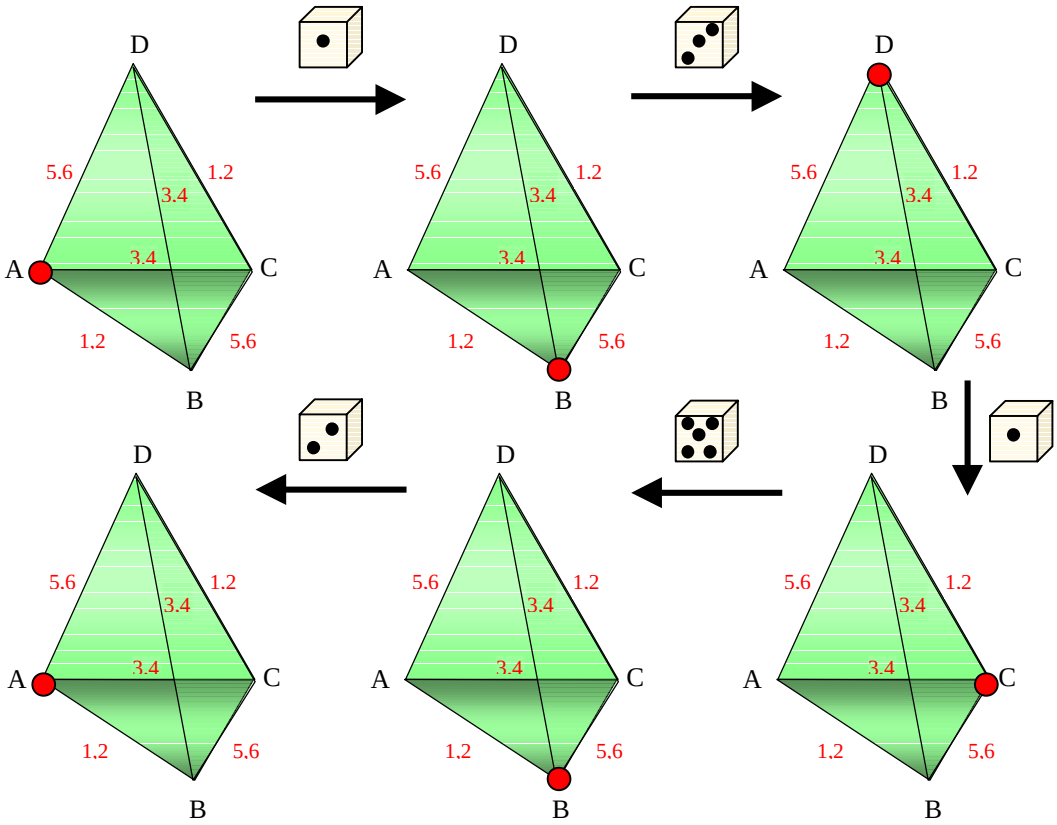
- Codage et simulation. Dénombrement.
- Etude d'une expérience par simulation.
- Calculs de moyennes, de médianes, de valeurs extrêmes. Observation de leur fluctuation entre séries d'expériences de même taille.
- Regroupement d'expériences.

PROMENADES ALEATOIRES SUR UN TETRAEDRE

On promène un pion sur les sommets d'un tétraèdre ; toutes les secondes, on déplace le pion d'un sommet à un autre, en choisissant au hasard parmi les trois sommets possibles. On s'intéresse au temps écoulé entre le début de la promenade du pion et le premier retour au point de départ. On limite la promenade à une minute. On utilise des lancers de dés, simulés ou non, pour les déplacements du pion.

Construire un codage des arêtes du tétraèdre pour une telle promenade.

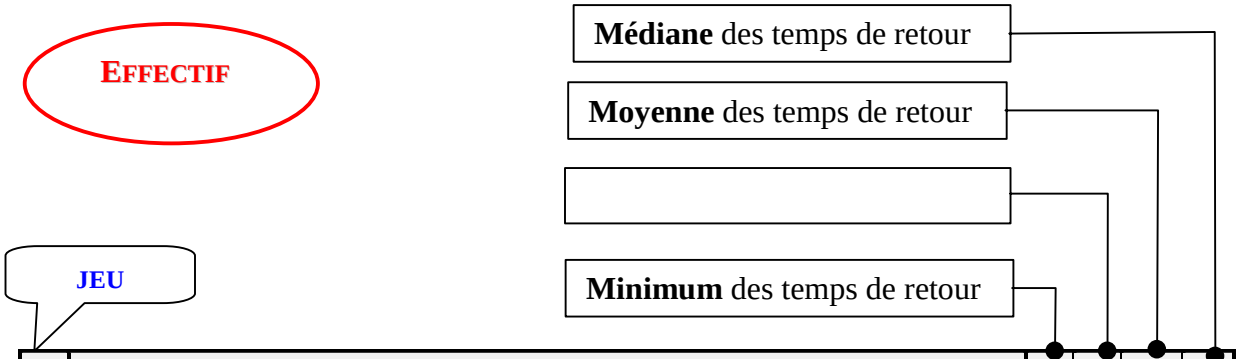
Voici un exemple d'une promenade en 5 coups, avec un codage particulier des arêtes :



Si on décide de donner le même code pour se déplacer d'un sommet à un autre, quel que soit le sens du trajet, combien y a-t-il de codages possibles ?

On appelle jeu un ensemble de 20 promenades. On simule 30 jeux. Pour chaque promenade, on étudie les temps de premier retour : ce sont des nombres entiers compris entre 1 et 60 (on a limité la promenade à 60 déplacements).

Il se trouve que pour les 20x30 promenades, le temps de premier retour n'excède jamais 19 secondes.

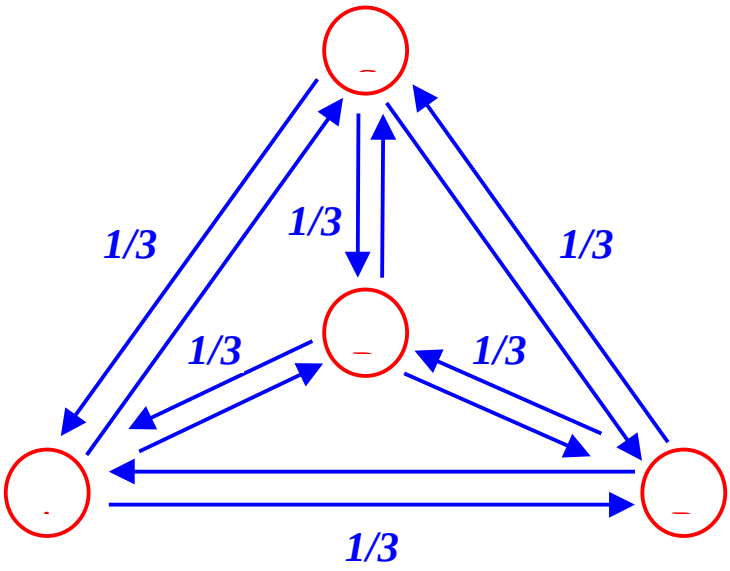


	Temps de premier retour																						
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19				
1	0	2	5	6	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	2	8	4.25	4
2	0	6	3	7	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	7	3.65	4
3	0	5	2	6	2	2	1	0	1	1	0	0	0	0	0	0	0	0	0	2	10	4.40	4
4	0	8	3	3	3	1	0	2	0	0	0	0	0	0	0	0	0	0	0	2	8	3.70	3
5	0	4	2	4	3	2	2	0	1	2	0	0	0	0	0	0	0	0	0	2	10	5.00	4.5
6	0	4	7	1	2	3	3	0	0	0	0	0	0	0	0	0	0	0	0	2	7	4.10	3
7	0	6	7	3	0	1	2	0	0	1	0	0	0	0	0	0	0	0	0	2	10	3.75	3
8	0	6	7	4	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	8	3.35	3
9	0	7	5	5	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2	9	3.45	3
10	0	10	4	1	1	0	2	0	0	0	1	1	0	0	0	0	0	0	0	2	12	3.90	2.5
11	0	6	2	4	5	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	3.85	4
12	0	8	3	2	3	1	1	2	0	0	0	0	0	0	0	0	0	0	0	2	8	3.85	3
13	0	4	5	6	2	0	2	1	0	0	0	0	0	0	0	0	0	0	0	2	8	3.95	4
14	0	6	5	3	3	2	0	0	0	0	0	0	0	0	0	1	0	0	0	2	16	4.10	3
15	0	7	6	2	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	2	9	3.55	3
16	0	8	6	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	3.00	3
17	0	6	6	4	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	3.40	3
18	0	6	2	2	3	3	2	1	0	0	0	0	0	1	0	0	0	0	0	2	14	4.75	4.5
19	0	6	7	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2	7	3.40	3
20	0	9	5	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	3.05	3
21	0	2	7	6	1	1	1	0	2	0	0	0	0	0	0	0	0	0	0	2	9	4.25	4
22	0	6	3	4	3	1	2	0	0	0	0	0	0	0	0	0	0	0	1	2	19	4.55	4
23	0	6	5	5	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	2	8	3.65	3
24	0	4	5	4	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	6	3.85	4
25	0	7	6	3	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	8	3.35	3
26	0	6	5	3	2	1	0	0	1	2	0	0	0	0	0	0	0	0	0	2	10	4.20	3
27	0	7	2	4	2	2	1	1	0	0	1	0	0	0	0	0	0	0	0	2	11	4.20	4
28	0	4	6	4	0	3	2	0	0	0	0	0	1	0	0	0	0	0	0	2	13	4.35	3.5
29	0	7	2	6	1	0	2	1	0	0	0	1	0	0	0	0	0	0	0				
30	0		3	4	3	1	0	1	0	0	0	0	0	0	0	0	0	0	0				
	0	**181**	**136**	**116**	**65**	**40**	**28**	**13**	**7**	**6**	**2**	**2**	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**2**	**19**	**3.88**	**3**

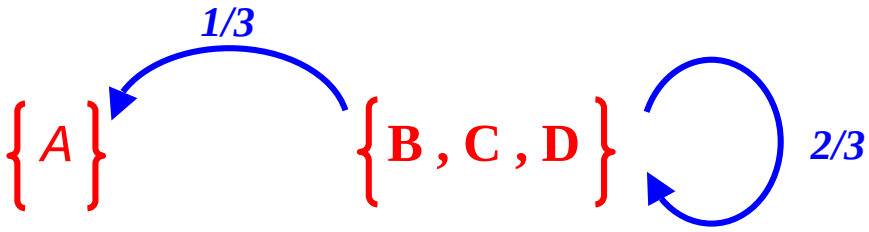
Que représente la dernière ligne ? Remplir les 9 cases vides des lignes 29 et 30, ainsi que le rectangle vide au dessus du tableau. Calculer les modes des séries présentées.

APERÇU THEORIQUE

Voici un schéma général pour la promenade considérée, où toutes les flèches sont codées 1/3 :



Si on s'intéresse plus particulièrement au premier retour en A, on peut illustrer la question avec le schéma suivant :



La probabilité p_k que la durée de la promenade du pion soit k, pour 1<k<60, est donnée par:

$$p_k = \left(\frac{2}{3}\right)^{k-2}\left(\frac{1}{3}\right).$$

La probabilité que la durée soit égale à 60 secondes (on arrête alors la promenade) est :

$$p_{60} = 1 - \sum_{k=2}^{59} p_k = \left(\frac{2}{3}\right)^{58}.$$

L'espérance ou moyenne théorique de la durée de la promenade est donnée par :

$$\mu = \sum_{k=2}^{60} k \times p_k = \frac{1}{3}\sum_{k=2}^{59} k \times \left(\frac{2}{3}\right)^{k-2} + 60 \times \left(\frac{2}{3}\right)^{58} \approx 4.$$

Plus généralement, si on limite la partie à n coups, la valeur moyenne théorique est :

$$\mu_n = \frac{1}{3}\sum_{k=2}^{n-1} k \times \left(\frac{2}{3}\right)^{k-2} + n \times \left(\frac{2}{3}\right)^{n-2},$$

d'où les valeurs suivantes :

n	4	5	10	20	60
μ_n	3,111	3,407	3,922	3,999	$4\text{-}10^{-10}$

Si on ne limitait pas la promenade, l'espérance de la durée est :

$$\mu_\infty = \sum_{k\geq 2} k \times p_k = \frac{1}{3}\sum_{k\geq 2} k \times \left(\frac{2}{3}\right)^{k-2} = 4$$

PROMENADES ALEATOIRES SUR UN CARRE

Un promeneur se déplace sur les sommets d'un terrain carré ; on s'intéresse au temps mis pour revenir au point de départ.

Les règles de déplacement sont les suivantes :

1. Départ du point SO (Sud Ouest),
2. On tire un chiffre au hasard : s'il est pair le promeneur va vers le sommet le plus proche en tournant dans le sens trigonométrique. Sinon, le promeneur va vers le sommet le plus proche en tournant dans le sens des aiguilles d'une montre. Des règles de déplacement similaires sont ensuite utilisées aux points NO, NE et SE, jusqu'au retour en SO où la promenade est terminée.
3. Il met une minute pour aller d'un sommet du carré à un sommet voisin. On limite la promenade à 50 minutes.

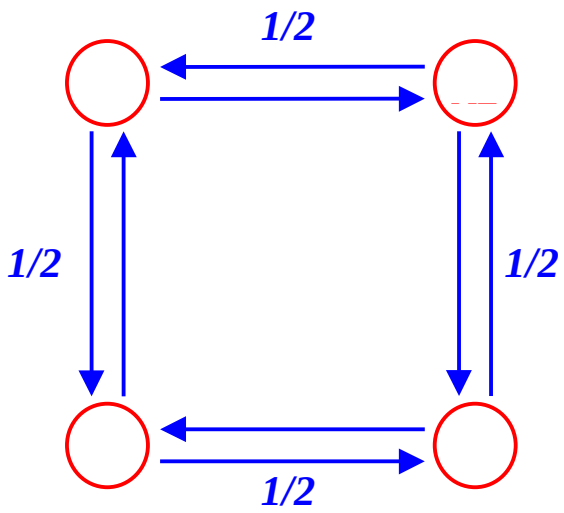
On appelle jeu un ensemble de 20 promenades. On simule 30 jeux. Pour chaque jeu, on étudie les temps de premier retour : ce sont des nombres entiers compris entre 1 et 50. Il se trouve que pour les 20x30 promenades, le temps de premier retour n'excède jamais 18 minutes.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18				
1	7	0	6	0	1	0	2	0	2	0	0	0	0	0	2	0	0	2	16	5.6	4
2	10	0	4	0	5	0	0	0	0	0	0	0	1	0	0	0	0	2	14	4.0	3
3	9	0	6	0	2	0	1	0	2	0	0	0	0	0	0	0	0	2	10	4.1	4
4	11	0	4	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2	8	3.5	2
5	14	0	4	0	1	0	0	0	0	0	1	0	0	0	0	0	0	2	12	3.1	2
6	11	0	4	0	1	0	2	0	1	0	0	0	1	0	0	0	0	2	14	4.2	2
7	10	0	8	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2	10	3.4	3
8	12	0	3	0	3	0	0	0	1	0	0	0	0	0	1	0	0	2	16	4.0	2
9	8	0	8	0	3	0	1	0	0	0	0	0	0	0	0	0	0	2	8	3.7	4
10	15	0	2	0	2	0	0	0	1	0	0	0	0	0	0	0	0	2	10	3.0	2
11	14	0	3	0	0	0	1	0	2	0	0	0	0	0	0	0	0	2	10	3.4	2
12	7	0	4	0	3	0	3	0	0	0	2	0	0	0	0	0	1	2	18	5.7	4
13	12	0	4	0	0	0	1	0	0	0	1	0	1	0	1	0	0	2	16	4.5	2
14	9	0	5	0	4	0	2	0	0	0	0	0	0	0	0	0	0	2	8	3.9	4
15	8	0	8	0	2	0	1	0	0	0	0	0	1	0	0	0	0	2	14	4.1	4
16	8	0	3	0	3	0	5	0	0	0	1	0	0	0	0	0	0	2	12	4.9	4
17	14	0	2	0	2	0	0	0	0	0	2	0	0	0	0	0	0	2	12	3.6	2
18	9	0	7	0	2	0	1	0	1	0	0	0	0	0	0	0	0	2	10	3.8	4
19	8	0	3	0	2	0	4	0	2	0	1	0	0	0	0	0	0	2	12	5.2	4
20	6	0	5	0	5	0	2	0	2	0	0	0	0	0	0	0	0	2	10	4.9	4
21	9	0	5	0	0	0	4	0	1	0	0	0	1	0	0	0	0	2	14	4.7	4
22	9	0	7	0	2	0	2	0	0	0	0	0	0	0	0	0	0	2	8	3.7	4
23	4	0	7	0	3	0	4	0	1	0	0	0	0	0	1	0	0	2	16	5.6	4
24	14	0	2	0	3	0	0	0	1	0	0	0	0	0	0	0	0	2	10	3.2	2
25	13	0	3	0	1	0	3	0	0	0	0	0	0	0	0	0	0	2	8	3.4	2
26	9	0	4	0	3	0	2	0	0	0	0	0	0	0	2	0	0	2	16	5.0	4
27	11	0	4	0	2	0	2	0	0	0	0	0	0	0	0	0	1	2	18	4.2	2
28	10	0	5	0	4	0	1	0	0	0	0	0	0	0	0	0	0	2	8	3.6	3
29	11	0	3	0	1	0	3	0	1	0	0	0	0	0	0	0	1	2	18	4.6	2
30	6	0	10	0	4	0	0	0	0	0	0	0	0	0	0	0	0	2	6	3.8	4
	298	**0**	**143**	**0**	**69**	**0**	**48**	**0**	**19**	**0**	**8**	**0**	**5**	**0**	**7**	**0**	**3**	**2**	**18**	**4.1**	**4**

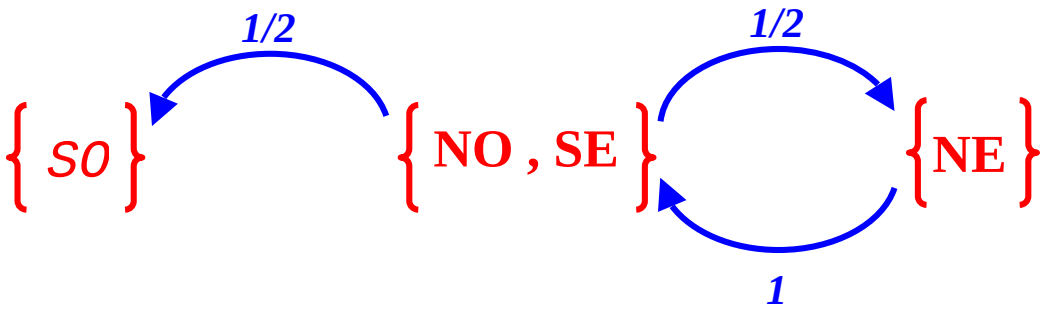
La légende de ce tableau s'est perdue ; que signifient les quatre dernières colonnes ? Et la dernière ligne ? Peut-on imaginer une simulation où les colonnes 3,5,7,9, etc. ne seraient pas remplies de 0 ?

APERÇU THEORIQUE

Voici un schéma général pour la promenade considérée, en admettant qu'elle ne se termine jamais (si le promeneur revient en A, il repart) :



Mais si on s'intéresse plus particulièrement au premier retour, on peut illustrer la question avec le schéma suivant :



La probabilité p_{2k} que la durée de la promenade du pion vaille 2k, pour 1;k;29, est donnée par : $p_k = \left(\frac{1}{2}\right)^k$.

La probabilité que la durée soit 60 (on arrête alors la promenade) est :

$$p_{60} = 1 - \sum_{k=1}^{29} p_k = \left(\frac{1}{2}\right)^{29}$$

L'espérance ou moyenne théorique de la durée de la promenade est :

$$\mu = \sum_{k=1}^{30} 2k \times p_{2k} = 2\sum_{k=1}^{29} k \times \left(\frac{1}{2}\right)^k + 60 \times \left(\frac{1}{2}\right)^{29} \approx 4$$

Si on limite la promenade à 2n coups, la moyenne théorique est :

$$\mu_{2n} = \sum_{k=1}^{n} 2k \times p_{2k} = 2\sum_{k=1}^{n-1} k \times \left(\frac{1}{2}\right)^k + 2n \times \left(\frac{1}{2}\right)^{n-1},$$

D’où les valeurs suivantes :

2n	4	6	8	10	20	60	100
μ_n	3	3,5	3,75	3,875	3,9961	$4 - 4\text{x}10^{-9}$	$4 - 3\text{x}10^{-15}$

Si on ne limite pas la promenade, la durée moyenne théorique est donnée par :

$$\mu_\infty = \sum_{k\geq 1} 2k \times p_{2k} = 2\sum_{k\geq 1} k \times \left(\frac{1}{2}\right)^k = 4\,.$$

Sondages

- Introduction à la notion de fourchette de sondage.
- Résultats de simulations.

UN PROBLEME DE SONDAGE SIMPLIFIE

Une urne contient une proportion p de boules numérotées 1, les autres étant numérotées 0. On ne peut pas compter les boules et tout ce que l'on peut faire pour connaître p est de tirer des boules (tirages avec remise). Si on fait n tirages et qu'on recueille la fréquence de 1 obtenue, quelle information cela apporte sur p ?
On appellera ici sondage de taille n dans une urne l'expérience consistant à faire n tirages avec remise dans cette urne.

Le résultat d'un sondage est un échantillon de l'expérience qui consiste à tirer avec remise une boule dans une urne et à regarder son numéro. Une même boule pouvant être tirée plusieurs fois, on ne peut pas dire que cet échantillon est celui des couleurs d'un sous-ensemble des boules de l'urne. Dans la pratique réelle des sondages, où un tirage au hasard d'individus dans une population importante est matériellement impossible, on remplace ce tirage par le choix d'une sous-population, appelée échantillon de cette population ou échantillon représentatif de cette population.

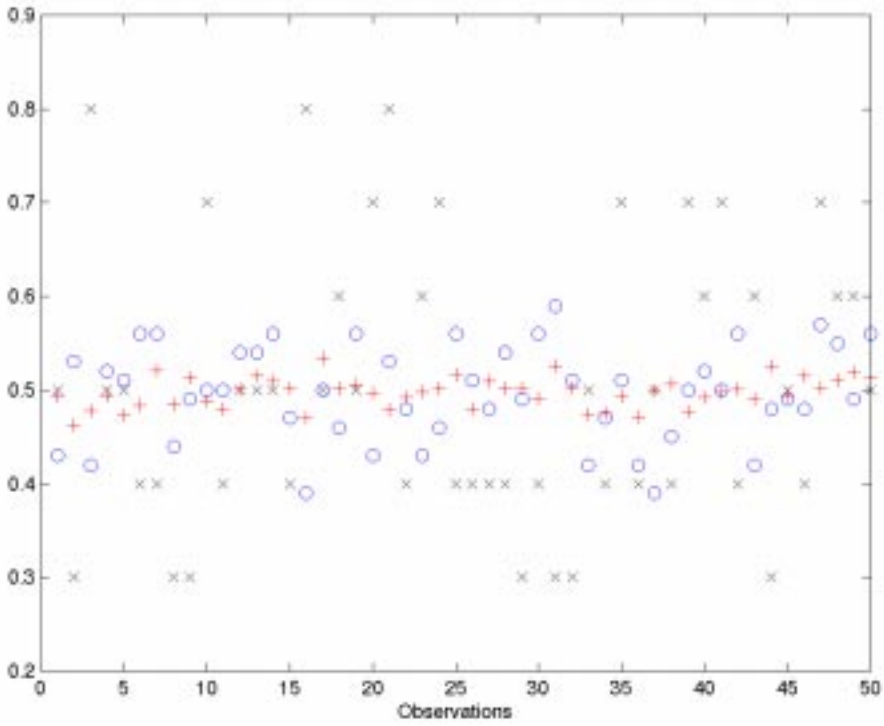
Cas où p est connu …

Dans un premier temps, nous allons, comme souvent en mathématiques, supposer le problème résolu : on fait un sondage dans une urne pour laquelle p est connu.
Un tel sondage peut aisément être simulé avec une liste de chiffres au hasard. Nous appellerons résultat d'un sondage la fréquence de 1 obtenue.
On simule 50 sondages de taille n pour diverses valeurs de n ; pour chaque valeur de n, on a donc une série de taille 50 de nombres compris entre 0 et 1, que nous résumons dans le tableau ci-dessous.

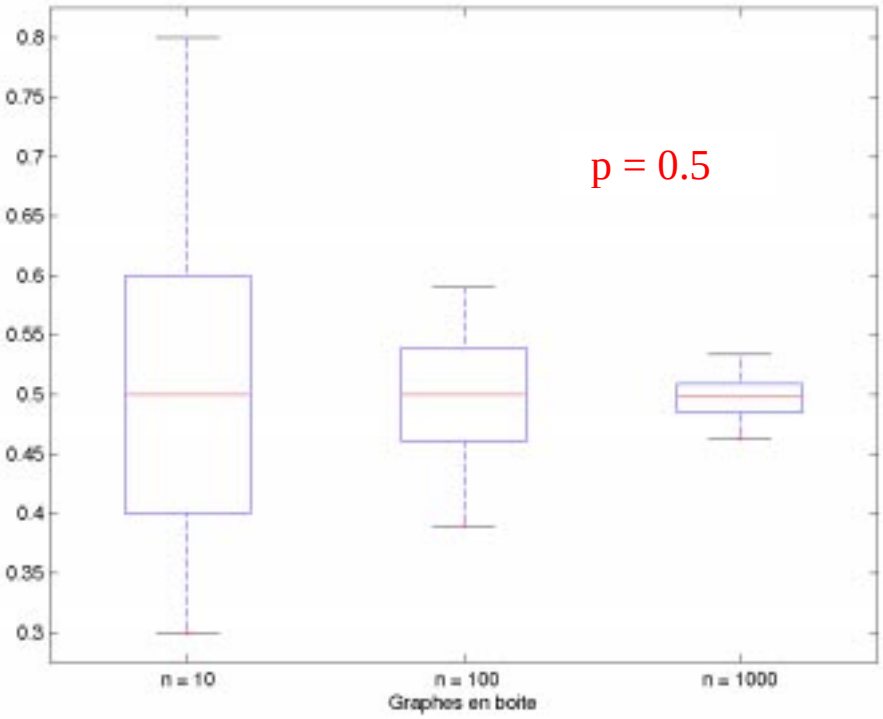
p = 0.5
50 sondages pour chaque valeur de n.

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
10	0.5	0.5	0.2	0.3	0.8	0.5	0.146
100	0.497	0.5	0.08	0.39	0.59	0.2	0.051
1000	0.497	0.498	0.024	0.463	0.534	0.071	0.016

L'intervalle interquartile et l'écart-type seront introduits en première. On pourra aussi utiliser les résultats de ces simulations en classe de première.



A l'aide du tableau ci-dessus, trouver à quelle taille de sondage correspondent les signes +, o, et x.



Les diagrammes en boîte ne sont pas au programme de seconde, mais sont au programme de première : on pourra aussi utiliser les résultats de ces simulations en classe de première. Dans ces diagrammes en boîtes, les ordonnées des segments extrêmes sont les valeurs extrêmes de la série.

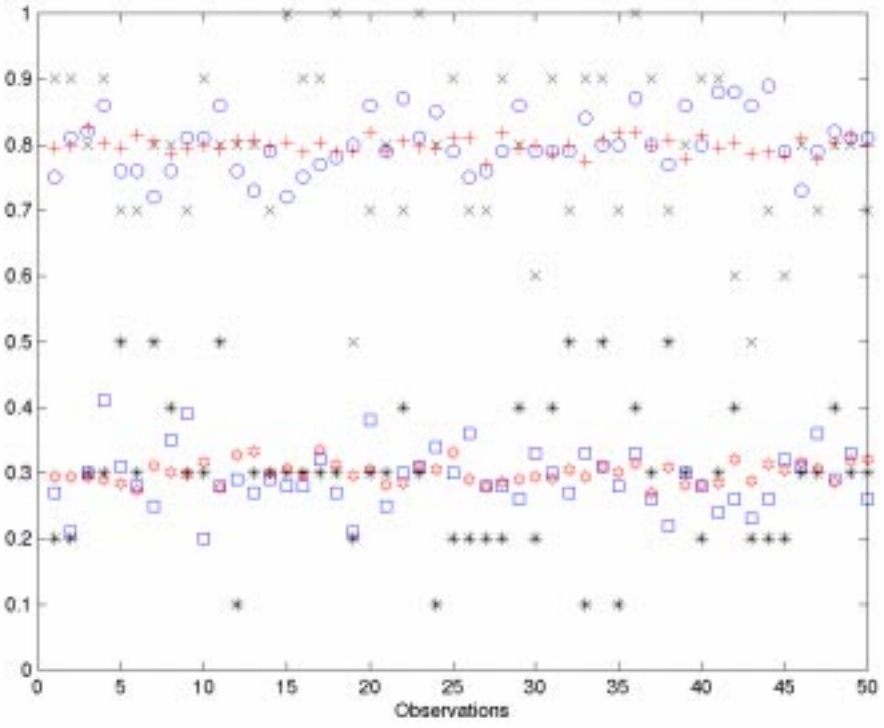
On recommence ensuite cette étude en fixant p successivement à 0.8 puis à 0.3 :
on obtient alors les résultats suivants :

p = 0.8
50 sondages pour chaque valeur de n.

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
10	0.792	0.8	0.2	0.5	1	0.5	0.123
100	0.802	0.8	0.07	0.72	0.89	0.17	0.045
1000	0.799	0.799	0.016	0.77	0.826	0.056	0.012

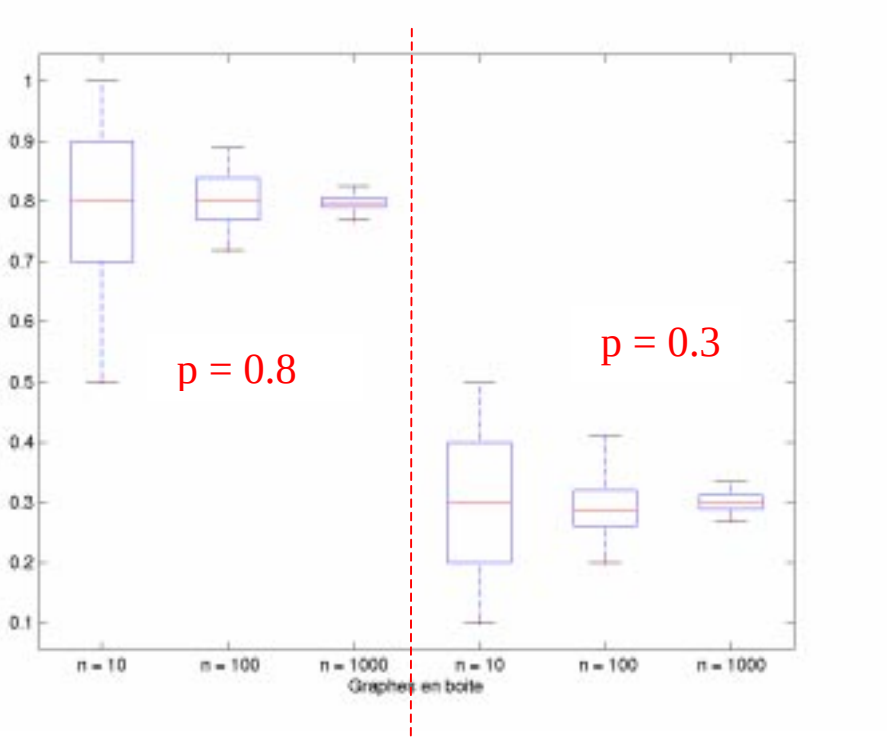
p = 0.3
50 sondages pour chaque valeur de n.

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
10	0.298	0.3	0.2	0.1	0.5	0.4	0.11
100	0.292	0.285	0.06	0.2	0.41	0.21	0.045
1000	0.301	0.299	0.023	0.27	0.336	0.066	0.016



p = 0.8 x ⟶ n = 10 o ⟶ n = 100 + ⟶ n = 1000

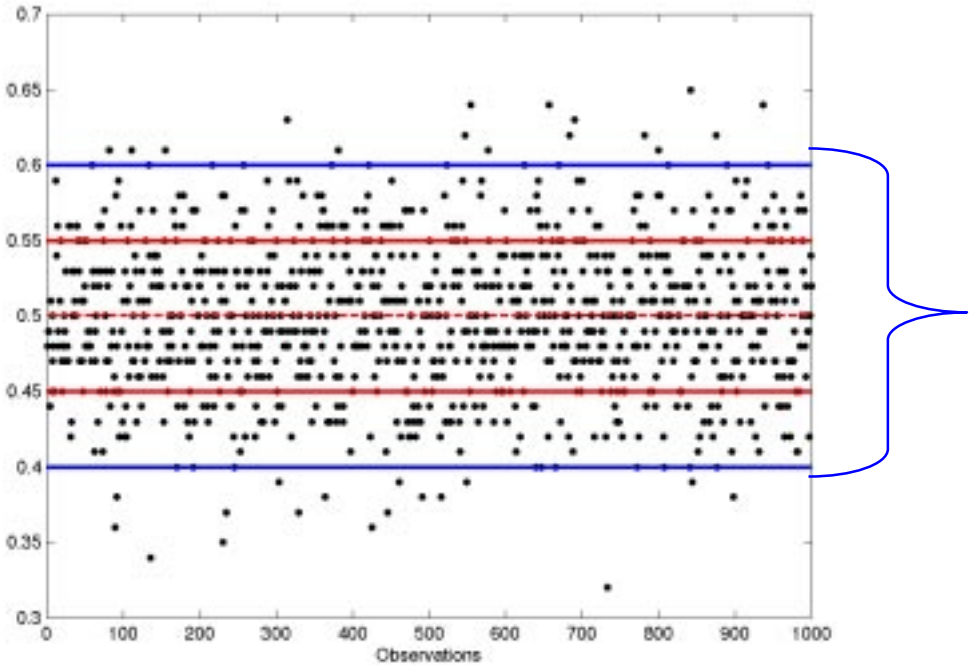
p = 0.3 * ⟶ n = 10 □ ⟶ n = 100 ☼ ⟶ n = 1000



On fait maintenant 1000 sondages de taille n = 100 pour p=0,5.

p = 0.5
1000 sondages de taille 100

n	Moyenne	Médiane	Interquartile	Minimum	Maximum	Etendue	Ecart-type
100	0.499	0.5	0.06	0.32	**0.65**	**0.33**	0.05



Un point représente la fréquence des 1 pour un sondage de taille 100 ; p=0,5.

Quelle est la proportion de sondages dans la bande délimitée par les traits bleus ? Dans la bande délimitée par les traits rouges ?

On peut démontrer, dans le cadre de la théorie de probabilités, la « formule » suivante :

Si on fait un grand nombre de sondages de taille n,
environ 95 % d'entre eux vérifient :

$$f \Sigma \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right] .$$

Est-ce que les résultats pour les 1000 sondages ci-dessus sont cohérents avec cette « formule » ?

Cas où p est inconnu

Les deux propositions suivantes sont équivalentes :

f est dans l'intervalle $[p - \delta ; p + \delta]$

p est dans l'intervalle $[f - \delta ; f + \delta]$

où f est la proportion observée de 1 et p la proportion de 1 dans l'urne.

Si on fait un grand nombre de sondages de taille n,
environ 95 % d'entre eux vérifient :

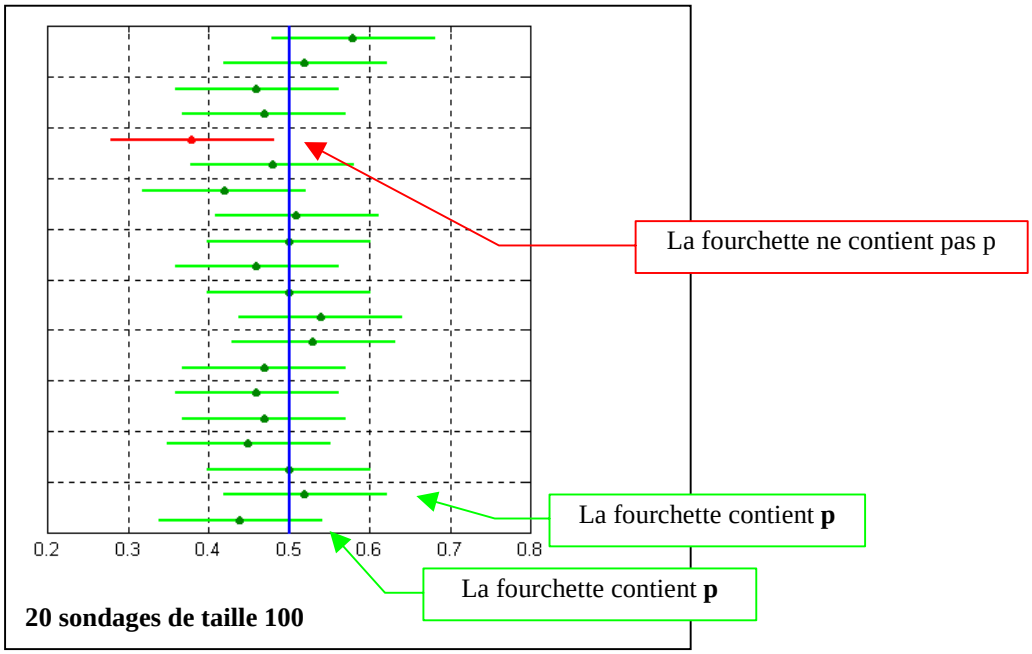
$$p \Sigma \left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right] .$$

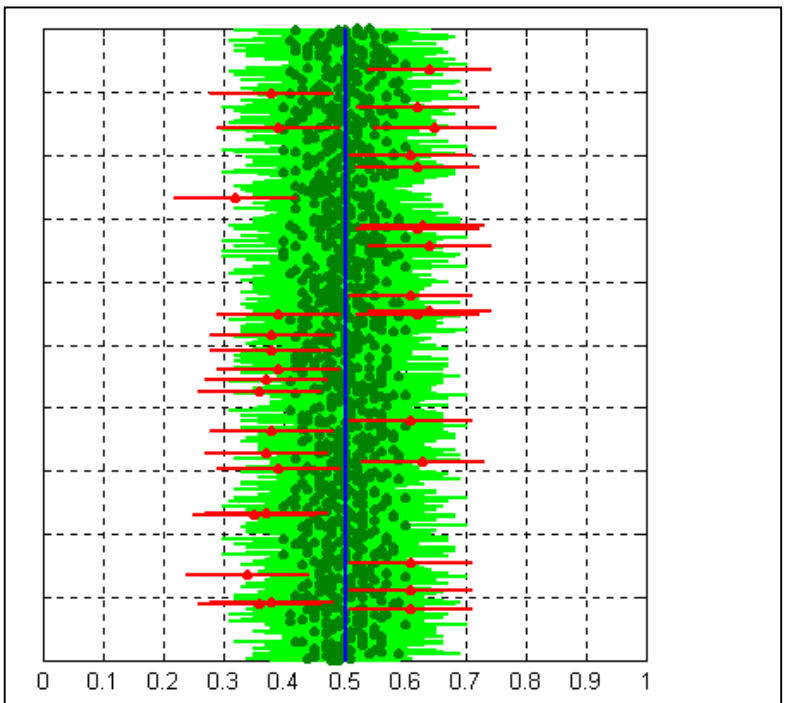
Pour un sondage de taille n dont le résultat est f, l'intervalle

$\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$ est appelé fourchette de sondage au niveau 0,95.

On dira aussi que f estime p avec une précision de $\frac{1}{\sqrt{n}}$, au niveau de confiance 0,95.

CONSTRUCTION DE FOURCHETTES AU NIVEAU 0,95





1000 sondages de taille 100

- 965 fourchettes (en vert) contiennent la valeur théorique p = 0.5.
- 35 fourchettes (en rouge) ne contiennent pas p.

APERÇU THEORIQUE

Soit S_n une variable aléatoire de loi binomiale B(n,p). Notons $R_n = \dfrac{S_n - np}{\sqrt{npq}}$.

Le théorème de Moivre énonce que la probabilité pour que R_n soit dans un intervalle [a,b] converge, lorsque n tend vers l'infini, vers l'intégrale :

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$$

Autrement dit, ce théorème énonce que la suite des variables R_n converge en loi vers la loi normale centrée réduite. La démonstration utilise essentiellement un développement asymptotique analogue à la formule de Stirling ainsi que la convergence d'une somme de Riemann vers une intégrale.
Notons Φ la fonction de répartition de la loi normale centrée réduite, soit :

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-u}^{u} e^{-\frac{x^2}{2}} dx$$

On a : $\Phi(1,96) \approx 0,95$.
Sous les hypothèses (H) suivantes :

(H) $n > 30$; $np > 5$; $nq > 5$ où $q = 1 - p$

on approxime avec une très bonne précision la probabilité pour que R_n soit dans un intervalle [a,b] par sa limite donnée dans le théorème de Moivre.

On a donc :

(1) $\text{Prob}(-1,96 \le R_n \le 1,96) \approx 0,95$

En notant $F_n = \dfrac{S_n}{n}$ la fréquence des 1 dans une suite d'expériences de Bernoulli de paramètre p, on peut écrire cette égalité sous la forme :

(2) $\text{Prob}\left(p - 1,96\sqrt{\dfrac{pq}{n}} \le F_n \le p + 1,96\sqrt{\dfrac{pq}{n}} \right) \approx 0,95$.

On dit que l'intervalle $\left[p - 1,96\sqrt{\dfrac{pq}{n}} ; p + 1,96\sqrt{\dfrac{pq}{n}} \right]$ est l'intervalle de dispersion de F_n au niveau 0,95.

Mais (2) peut s'écrire :

(3) $\text{Prob}\left(p \in \left[F_n - 1,96\sqrt{\dfrac{pq}{n}} ; F_n + 1,96\sqrt{\dfrac{pq}{n}} \right] \right) \approx 0,95$.

Remarquons que pq est toujours inférieur à $\dfrac{1}{4}$. Donc de (3), on déduit :

(4) $\text{Prob}\left(p \in \left[F_n - \dfrac{1}{\sqrt{n}} ; F_n + \dfrac{1}{\sqrt{n}} \right] \right) \ge 0,95$.

En pratique, si la valeur observée f_n (fréquence de 1 dans l'échantillon) est comprise entre 0,3 et 0,5, on admet que (H) est vérifiée et on en déduit que $\sqrt{\frac{pq}{n}}$ est peu différent de $\frac{1}{2\sqrt{n}}$.

On dit que l'intervalle $\left[f_n - \frac{1}{\sqrt{n}} ; f_n + \frac{1}{\sqrt{n}}\right]$ est la fourchette de sondage de *p* au niveau 0,95.

Si on estime *p* par f_n, on dira que la précision est $\frac{1}{\sqrt{n}}$ avec un niveau de confiance 0,95. Si *n* est de l'ordre de 100 (respectivement 1000), la précision de l'estimation d'un pourcentage au niveau 0,95 est de l'ordre de 10 % (respectivement 3 %).

Si on remplace le niveau 0,95 par 0,90, il faut remplacer 1,96 par 1,65.

Si on remplace le niveau 0,95 par 0,99, il faut remplacer 1,96 par 3.

D'où :

- L'intervalle $\left[f_n - \frac{1{,}65}{2\sqrt{n}} ; f_n + \frac{1{,}65}{2\sqrt{n}}\right]$ est la fourchette de sondage de *p* au niveau 0,90.

  Au niveau de confiance 0,90, la précision sur *p*, lorsqu'on estime *p* par F_n, est $\frac{1{,}65}{2\sqrt{n}}$.
- L'intervalle $\left[f_n - \frac{3}{2\sqrt{n}} ; f_n + \frac{3}{2\sqrt{n}}\right]$ est la fourchette de sondage de *p* au niveau 0,99.

  Au niveau de confiance 0,90, la précision sur *p*, lorsqu'on estime *p* par F_n, est $\frac{3}{2\sqrt{n}}$.

Linéarité de la moyenne

- Observer et réfléchir avant d'agir.
- Si une transformation linéaire ou affine est appliquée à des données, alors la même transformation s'applique à leur moyenne.

Comment calculer la moyenne arithmétique de la série ci-dessous ?

$$x_1 = 0{,}00432567189$$

$$x_2 = 0{,}00432567370$$

$$x_3 = 0{,}00432567127$$

$$x_4 = 0{,}00432567433$$

$$x_5 = 0{,}00432567156$$

$$x_6 = 0{,}00432567238$$

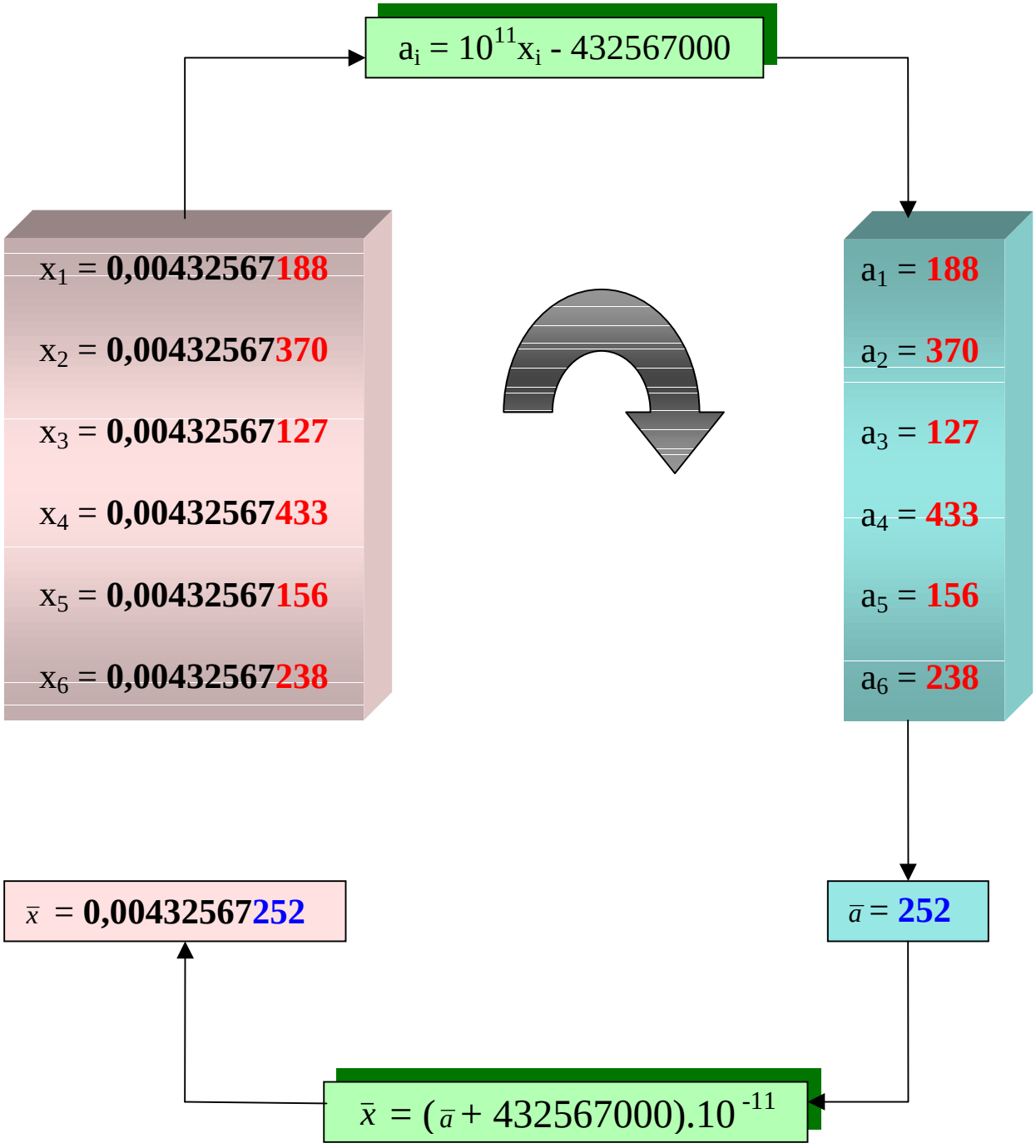
Observons cette série :

- *chaque donnée comporte beaucoup de chiffres, ce qui multiplie les risques d'erreurs de saisie ; de plus certaines calculatrices n'acceptent pas autant de chiffres ;*

- *seuls les trois derniers chiffres sont variables.*

Une transformation des données s'impose, à savoir :

$$a_i = 10^{11} x_i - 432567000$$



Un cube moyen ?

- Si une transformation non linéaire est appliquée à des données, alors la même transformation ne s'applique pas nécessairement à leur moyenne.
- Une série où médiane et moyenne diffèrent sensiblement.

Une entreprise fabrique sur commande des cubes creux de tailles différentes. En début du processus de fabrication, on détermine la longueur du côté des cubes à fabriquer et en fin de processus, on détermine leur capacité. Pour avoir une idée de la production, la direction a demandé de calculer deux moyennes ; ainsi, pour une série de 500 cubes,S la moyenne des côtés est de 10,7 cm ; pour ces mêmes cubes, le volume moyen est 2193 cm^3. Voulant tenir compte de toutes les mesures, la direction a conclu que le cube moyen fabriqué par cette entreprise était un cube de côté 10,7 cm et de volume 2193 cm^3 !

	n	Moyenne	Médiane	Minimum	Maximum	Etendue
côtés	500	10.7	10.5	1	20	19

	n	Moyenne	Médiane	Minimum	Maximum	Etendue
volumes	500	2193	1165.5	1	8000	7999

Pour mieux comprendre ce qui se passe pour les cubes, il suffit de prendre trois cubes de côtés respectifs 1,10 et 20 cm : le cube de la moyenne de leurs côtés n'est pas du tout égal à la moyenne des cubes des côtés.

Qu'en est-il des liens entre la médiane (resp. le maximum, le minimum, l'étendue) des côtés et des volumes ?