Institut de Recherche sur l'Enseignement des Sciences

UNIVERSITÉ DE MONTPELLIER

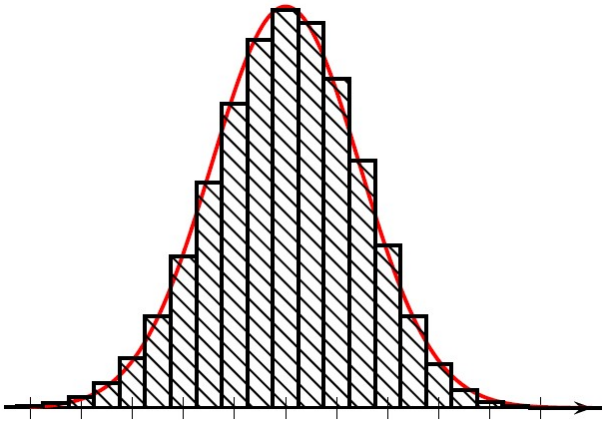
IRES MONTPELLIER

FACULTÉ DES SCIENCES DE MONTPELLIER

fde Faculté d'Éducation Université Montpellier 2

ACADÉMIE DE MONTPELLIER
Liberté
Égalité
Fraternité

Exercices de Probabilités et Statistique au Lycée



Auteurs : BRESSON Daniel, BRY Xavier, KINACH Hélène, SCHADECK Jean-Marie.

2020

Adresse : cc40, Place Eugène Bataillon
34095 MONTPELLIER CEDEX 05
Courriel : fds.ires@umontpellier.fr

☎ : (33) 04 67 14 33 83
Fax : (33) 04 67 14 39 09
Site : https://ires-fds.edu.umontpellier.fr/

Probabilités et Statistique au Lycée

Ce document a été réalisé par le groupe IRES *Probabilités et Statistique* de Montpellier.

Il est destiné aux enseignants de lycée général, technologique et professionnel.

Un travail de réflexion sur les exercices proposés dans différents manuels, notamment indiqués par des stagiaires, a été réalisé, ainsi que sur les exercices posés aux différentes épreuves du baccalauréat, lesquels pouvant parfois poser quelques difficultés.

Nous proposons une série d'exercices, originaux ou existants, avec corrections détaillées, voire très détaillées, en limitant les « implicites », et si nécessaire, accompagnés de commentaires sur la formulation, les erreurs à éviter et des prolongements et compléments de cours.
Les encadre=és peuvent être sautés en première lecture

Le document est composé de quatre sections :

- la première illustre les indices de la statistique, à savoir les quantiles, et les diverses notions de dispersion et de résumés statistiques associés ;
- la deuxième illustre les probabilités discrètes et les lois du programme ;
- la troisième section traite la statistique inférentielle ;
- enfin des annexes plus théoriques complètent le document.

Le groupe Probabilités et Statistique de l'IRES de Montpellier :

Daniel BRESSON

Xavier BRY

Hélène KINACH

Jean-Marie SCHADECK

Nous remercions tous les collègues rencontrés dans nos établissements et lors de différents stages pour leurs remarques et questions qui ont permis de nourrir notre réflexion.

Table des matières

Loi uniforme

Partie C : STATISTIQUE INFÉRENTIELLE

Intervalle de fluctuation et prise de décision

ANNEXES

Partie A

STATISTIQUE DESCRIPTIVE

Exercice 1 « Paradoxe » de Simpson

Thème abordé

- **Moyennes**

Énoncé

1. Dans une classe A d'un lycée, les notes des garçons (G) et des filles (F) en Histoire au baccalauréat blanc sont les suivantes :

Classe A

notes G	effectifs G	notes F	effectifs F
7	1	7	1
8	3	8	3
		9	2
		10	3
		11	2
		12	3

a. Calculer la note moyenne des garçons et celle des filles. Comparer ces deux moyennes.
b. En déduire la moyenne globale de la classe.

2. Mêmes questions qu'en **1.** pour la classe B suivante :

Classe B

notes G	effectifs G	notes F	effectifs F
14	3	19	3
15	2	20	1
16	3		
17	2		
18	3		
19	2		
20	1		

3. **a.** Calculer la note moyenne des garçons et celle des filles sur l'*ensemble* des deux classes, et comparer ces deux moyennes.
b. Construire un diagramme en bâtons qui distingue les filles des garçons et la classe A de la classe B.

4. Expliquer le paradoxe (dit de Simpson) constaté.

Corrigé

1. **a.** Moyenne : $\bar{x} = \frac{1}{N}\sum_i n_i x_i$,

ce qui donne pour les garçons : $\bar{x}_G^A = 7,75$ et pour les filles : $\bar{x}_F^A = 10,1875$.
Donc en moyenne, les filles ont mieux réussi que les garçons.

b. La moyenne dans la classe A est 9,7.

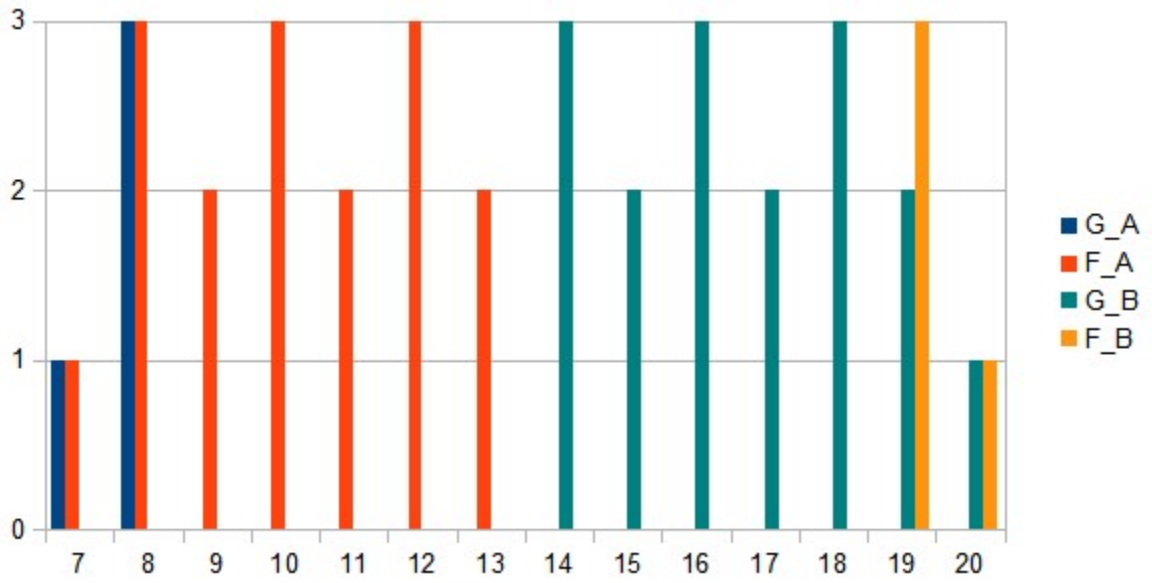
2. **a.** $\bar{x}_G^B = 16,625$ et $\bar{x}_F^B = 19,25$.
Ici aussi, les filles ont mieux réussi que les garçons.

b. La moyenne dans la classe B est 17,15.

3. a. $\overline{x}_G = \dfrac{n_G^A \overline{x}_G^A + n_G^B \overline{x}_G^B}{n_G^A + n_G^B} = 14{,}85$ et $\overline{x}_F = \dfrac{n_F^A \overline{x}_F^A + n_F^B \overline{x}_F^B}{n_F^A + n_F^B} = 12$.

Ça alors, sur l'ensemble des deux classes, les filles ont moins bien réussi que les garçons !
Pourtant, il n'y a pas d'erreur de calcul.

b. Diagramme en bâtons :



4. L'explication est la suivante : bien que les filles aient de meilleures notes que les garçons dans chaque classe, ici on a $\overline{x}_F^B \geqslant \overline{x}_G^B \geqslant \overline{x}_F^A \geqslant \overline{x}_G^A$, elles sont beaucoup plus nombreuses dans la classe qui a les notes plus basses.

Si on avait $\overline{x}_F^B \geqslant \overline{x}_F^A \geqslant \overline{x}_G^B \geqslant \overline{x}_G^A$, on aurait encore $\overline{x}_F \geqslant \overline{x}_G$ quelque soient les effectifs.

Exercice 2 Pas d'amalgame

Thème abordé

- **Histogrammes**

Énoncé

On a relevé les pointures de pied de 56 personnes d'un pays *A*. Par ailleurs, on note le sexe de chaque personne. Les résultats sont rassemblés dans le tableau suivant.

Femmes	Hommes
30, 31, 32, 33, 33, 34, 34, 35, 35, 36, 36, 37, 37, 38, 38, 38, 39, 39, 39, 39, 40, 40, 40, 41, 41, 42, 43, 44	36, 37, 38, 39, 39, 40, 40, 40, 41, 41, 41, 41, 42, 42, 42, 43, 43, 44, 44, 45, 45, 46, 46, 47, 47, 48, 49, 50

1. Tracer l'histogramme des fréquences des pointures tous sexes confondus.
À la vue du diagramme obtenu, la population vous semble-t-elle homogène ?

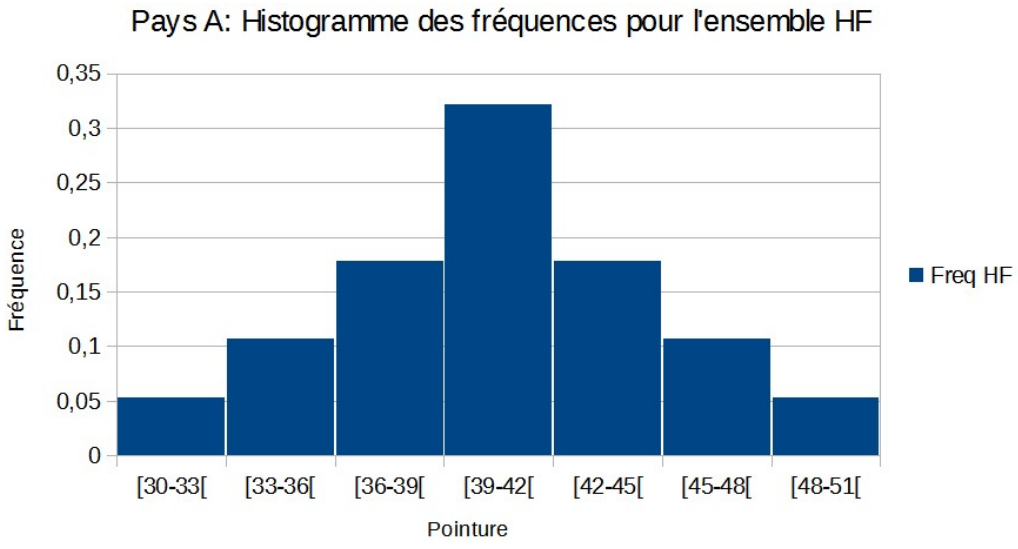
2. Tracer l'histogramme des fréquences des pointures des femmes, ainsi que celui des hommes dans un même repère. Ce nouveau graphique corrobore-t-il votre réponse à la question précédente ?

3. Reprendre la question **2.** puis la question **1.** avec les données suivantes, issues de l'étude d'habitants d'un pays *B* :

Femmes	Hommes
30, 31, 31, 32, 32, 32, 32, 33, 33, 33, 33, 33, 34, 34, 34, 34, 35, 35, 35, 36, 36, 36, 37, 37, 38, 39, 40, 41	39, 40, 41, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 45, 45, 45, 46, 46, 46, 46, 46, 46, 47, 47, 47, 48, 49, 50

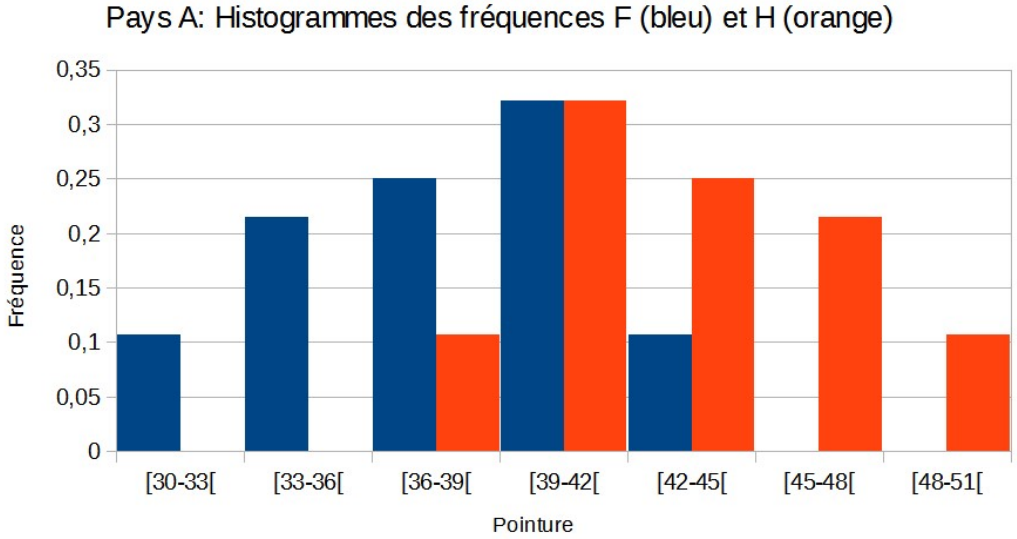
Corrigé

1.



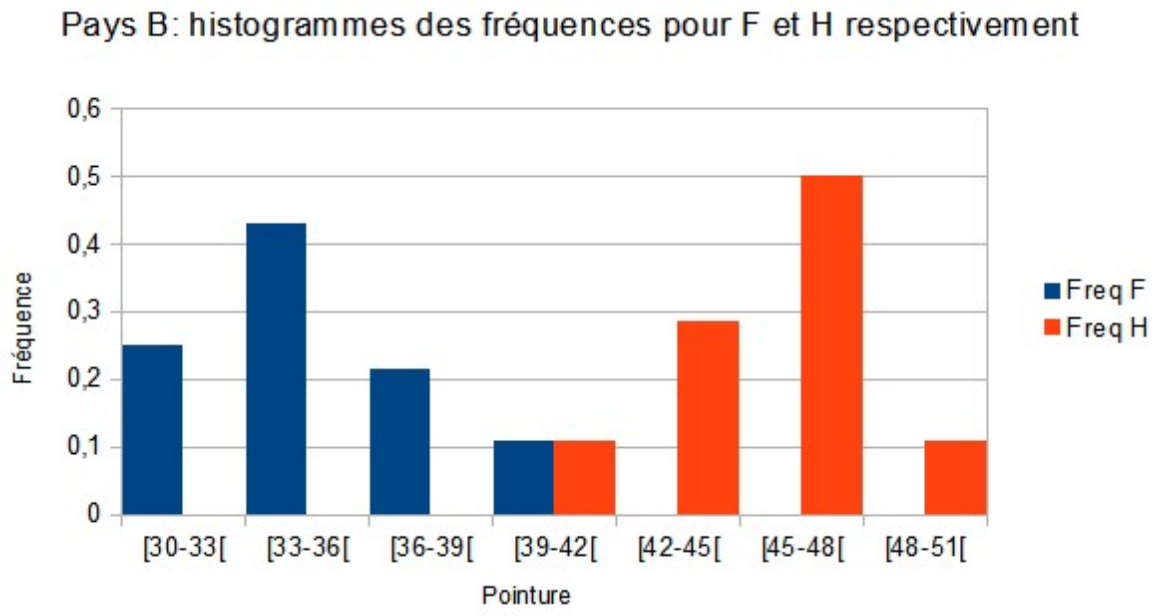
Cet histogramme ne fait pas apparaître de sous-populations bien distinctes. Rien n'y indique que la population soit hétérogène.

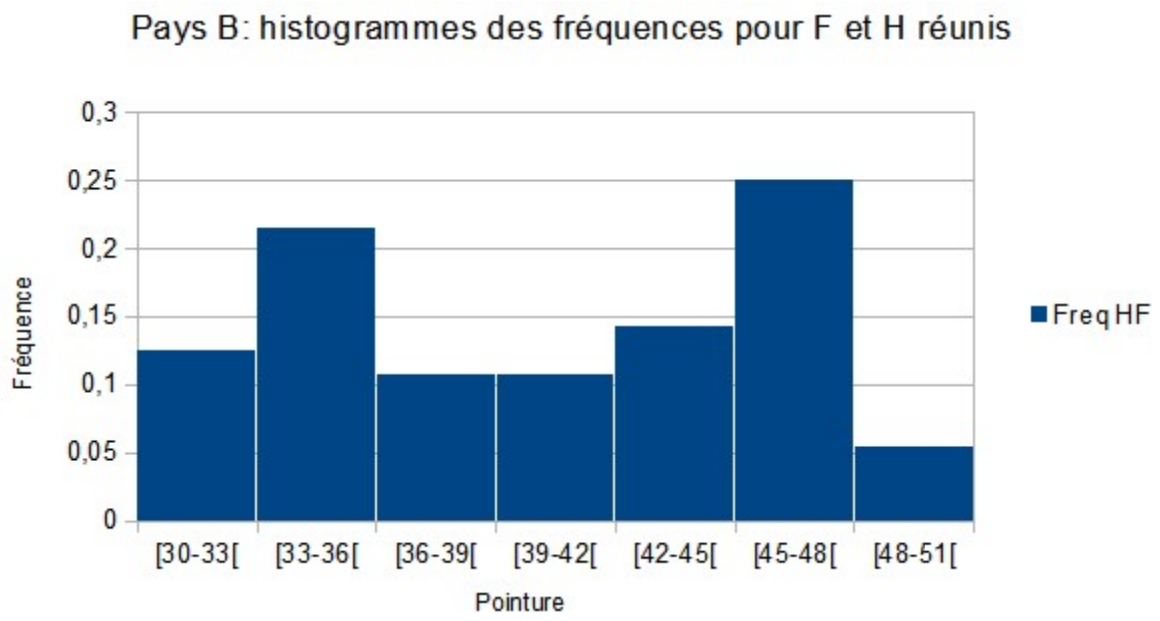
2.



Ce graphique fait apparaître que les femmes et les hommes du pays *A* sont, du point de vue de la pointure, deux populations très distinctes. L'histogramme qui les mélange ne permet pas de deviner leur existence, et laisserait même croire aux naïfs que la population est homogène et que la distribution des pointures est à peu près normale.

3.





Dans le pays *B*, les sous-populations H et F sont suffisamment éloignées l'une de l'autre pour que l'histogramme de leur mélange fasse apparaître deux modes, ce qui laisse penser que la population est constituée de deux sous-populations hétérogènes. Ce qui n'est pas le cas pour le pays *A*.

Exercice 3 Tendances centrales et dispersions

Thèmes abordés

- **Médiane et moyenne**
- **Écart interquartiles et écart type**

Énoncé

Le nombre de conversations téléphoniques par jour a été enregistré dans 15 centres d'appel. Les données sont :

35, 49, 225, 50, 30, 65, 40, 55, 52, 76, 48, 325, 47, 32, 60

Trouver :

a. la médiane du nombre de conversations téléphoniques ;
b. le nombre moyen de conversations ;
c. l'écart interquartiles ;
d. l'écart-type.

Commenter.

Corrigé

a. On ordonne la série par valeurs croissantes :
30, 32, 35, 40, 47, 48, 49, 50, 52, 55, 60, 65, 76, 225, 325
Il y a 15 valeurs, c'est un nombre impair ; $n = 2p + 1$ et la médiane est la p+1-ième valeur de la série ordonnée ; donc ici, la 8ème.
30, 32, 35, 40, 47, 48, 49, 50, 52, 55, 60, 65, 76, 225, 325
$m_e = x_{7+1} = 50$
(On rappellera que si les valeurs sont en nombre pair $n = 2p$, on prendra la demi somme des p-ième et $(p+1)$-ième valeurs classées)

b. La moyenne $\bar{x}$ est 79,26666... On notera deux choses :
- la moyenne n'est pas nécessairement entière, donc ne correspond pas à un nombre de conversations observables, alors que la médiane SI !, au moins lorsque le nombre des observations est impair.
- la moyenne est ici très supérieure à la médiane... Pourquoi ? Parce qu'il y a des observations atypiques dans les grandes valeurs. Celles-ci tirent la moyenne vers elles (donc vers le haut), tandis qu'elles n'influencent pas la médiane (remplacez 225 et 325 par 77 et 78, ou par n'importe quelle valeur supérieure ou égale à 50, et vous verrez que la médiane ne change pas)

c. Les quartiles sont obtenus en calculant leurs rangs :
$15 = 4 \times 3 + 3$ donc ici ce sont les $4^{\text{ème}}$ et $12^{\text{ème}}$ valeurs ordonnées, ce qui correspond aux valeurs 40 et 65.
$q_1 = x_{3+1} = 40$ et $q_3 = x_{3\times3+3} = 65$
L'écart inter-quartiles est donc : $q_3 - q_1 = 65 - 40 = 25$.

Taille de la série	q_1	q_3
$n = 4p$	x_p	x_{3p}
$n = 4p + 1$	x_{p+1}	x_{3p+1}
$n = 4p + 2$	x_{p+1}	x_{3p+2}
$n = 4p + 3$	x_{p+1}	x_{3p+3}

Remarque :

De façon générale, on dispose de la formule suivante pour la série ordonnée :

$q_\alpha = x_{E[(n+1)\alpha)} + \left((n+1)\alpha - E\left((n+1)\alpha\right)\right)\left(x_{E[(n+1)\alpha)+1} - x_{E[(n+1)\alpha)}\right)$ où E désigne la fonction partie entière et q_α est le quantile d'ordre α. Ce qui est bien sûr hors programme.

d. La moyenne des carrés des valeurs est de 12649,5.

On rappelle que $V(x) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2$ et $s(x) = \sqrt{V(x)}$.

Ici, l'écart-type est : $s(x) = 79,79$.

On remarque que l'écart-type est plus de trois fois supérieur à l'écart interquartile. Encore une fois, ceci est dû à l'influence des valeurs extrêmes, qui augmentent l'écart-type, tandis que les valeurs inférieures au 1er quartile ou supérieures au 3ème ne jouent pas sur l'écart interquartile.

Exercice 4 Observations incomplètes

Thèmes abordés

- **Médiane et moyenne**
- **Écart interquartiles et écart type**

Énoncé

On étudie la durée de vie d'un organisme contaminé par un virus à une date donnée. L'expérience, menée sur un échantillon de 20 organismes pendant 10 jours, à compter de la date d'inoculation prise pour origine des temps (date $t = 0$), a conduit aux observations suivantes :

Date t (jours)	[0,1[	[1,2[	[2,3[	[3,4[	[4,5[	[5,6[	[6,7[	[7,8[	[8,9[	[9,10[
Nombre de décès	0	0	1	4	5	5	2	1	0	0

1. Peut-on calculer la durée de vie moyenne ? L'écart-type ?
Que se passe-t-il si on les calcule sur les données du tableau ?

2. Peut-on calculer la durée de vie médiane ? L'écart interquartiles ? Si oui, les calculer.

Corrigé

1. Il n'y a que 18 organismes qui sont morts ! Il manque les durées de vie des organismes qui ont vécu le plus longtemps.
Les données sont incomplètes, elles ne permettent pas de calculer la moyenne, ni bien sûr l'écart-type. Calculer la moyenne sur les observations disponibles va en effet forcément conduire à sous-estimer la durée de vie moyenne (que l'on pourra calculer seulement quand tous les organismes seront morts !).

2. En revanche, comme plus de la moitié des organismes sont morts, on peut parfaitement déterminer la durée de vie médiane.

Date t (jours)	[0,1[	[1,2[	[2,3[	[3,4[	[4,5[	[5,6[	[6,7[	[7,8[	[8,9[	[9,10[
Nombre de décès	0	0	1	4	5	5	2	1	0	0
Nombres cumulés	0	0	1	5	10	15	17	18	18	18

Par simple lecture, la médiane est 4, c'est-à-dire qu'au moins la moitié de la population meurt avant le cinquième jour et au moins la moitié à partir du cinquième jour.

Et comme plus des 3/4 des organismes sont morts, on peut déterminer aussi les quartiles. (Mais attention : avec $n = 20$).
On lit, $q_1 = 3$ et $q_3 = 5$.

Exercice 5 Liaison et agrégation

Thème abordé

- **Ajustement affine**

Énoncé

On trouvera ci-dessous les tailles (t) de longueur de manche et les prix (p) de chemises pour hommes et de chemises pour femmes dans une boutique de vêtements parisienne.

Chemises pour hommes :

Taille : t_i	57	59	61	63	65	67	69	71	73	75
Prix : p_i	10	10	11	11	12	12,5	13	13,5	14	15

Chemises pour femmes :

Taille : t_i	35	37	39	41	43	45	47	49	51	53
Prix : p_i	14	14	15	15	16	17	18	19	21	22

1. **a.** Représenter le nuage de points $M(t_i ; p_i)$ correspondant aux chemises pour femmes dans le plan muni d'un repère adapté, ainsi que le point moyen.

b. Donner, à l'aide de la calculatrice ou d'un tableur, une équation de la droite D d'ajustement de p en t obtenue par la méthode des moindres carrés, les coefficients étant arrondis à 10^{-4} près, et la tracer sur le graphique précédent.

c. Commenter.

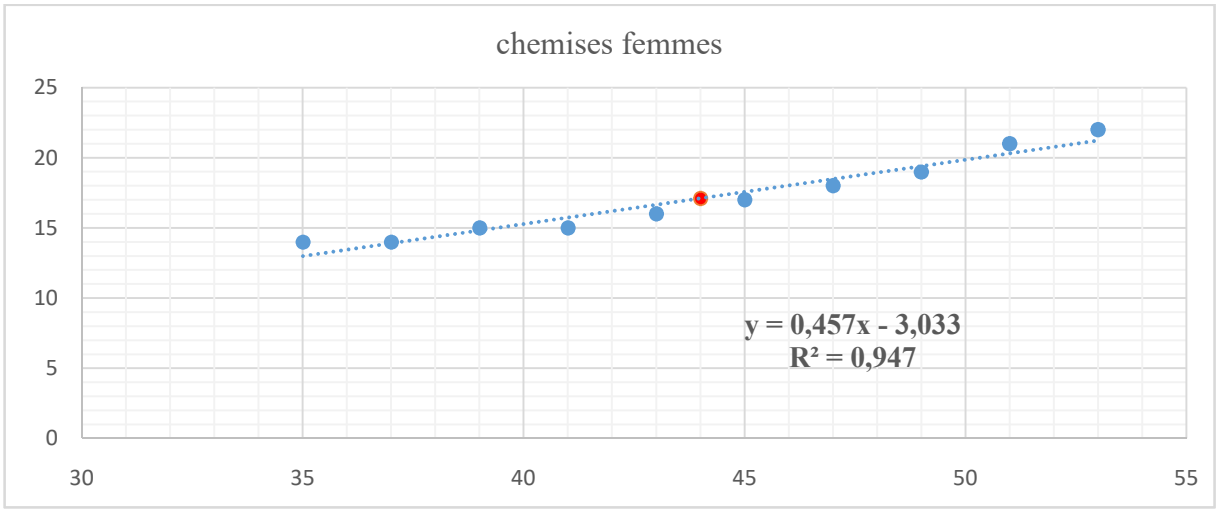
2. Reprendre ces trois questions pour les chemises pour hommes.

3. Faire de même enfin quand on regroupe toutes les chemises.

4. Expliquer le paradoxe, et tirer la morale de l'histoire.

Corrigé

1.

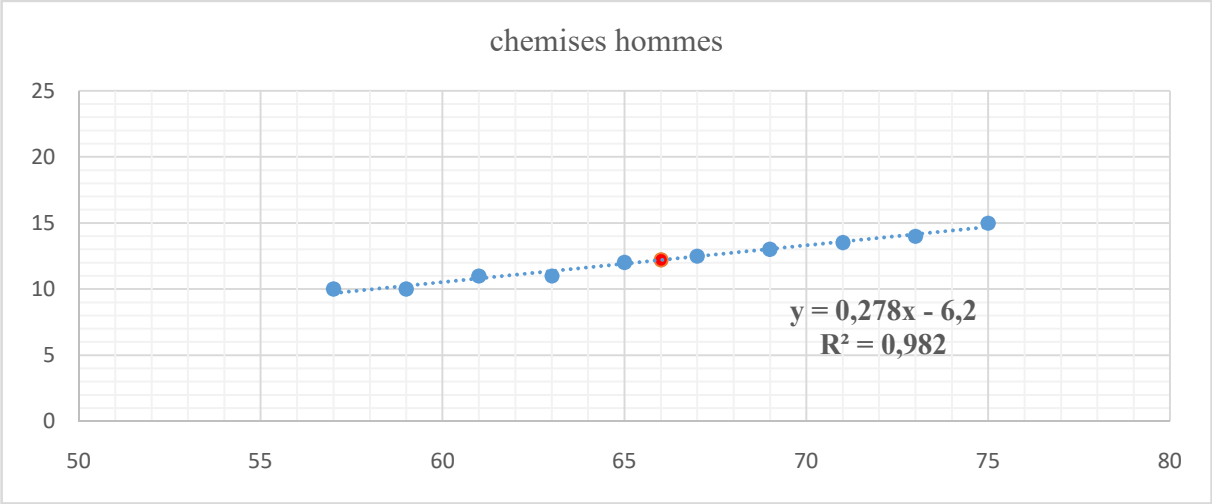


Le point moyen est $G_f(44 ; 17{,}1)$.

Le graphique montre des points tous pratiquement alignés. La droite d'ajustement est donc un bon modèle de la liaison entre t et p chez les femmes. La variable p y apparaît linéaire et croissante en fonction de t (le coefficient directeur de la droite est positif).

Remarque **hors programme** : Le carré de la corrélation linéaire est $R^2 = 0{,}9478$, très proche de 1. Par conséquent, la relation statistique est pratiquement linéaire (affine).

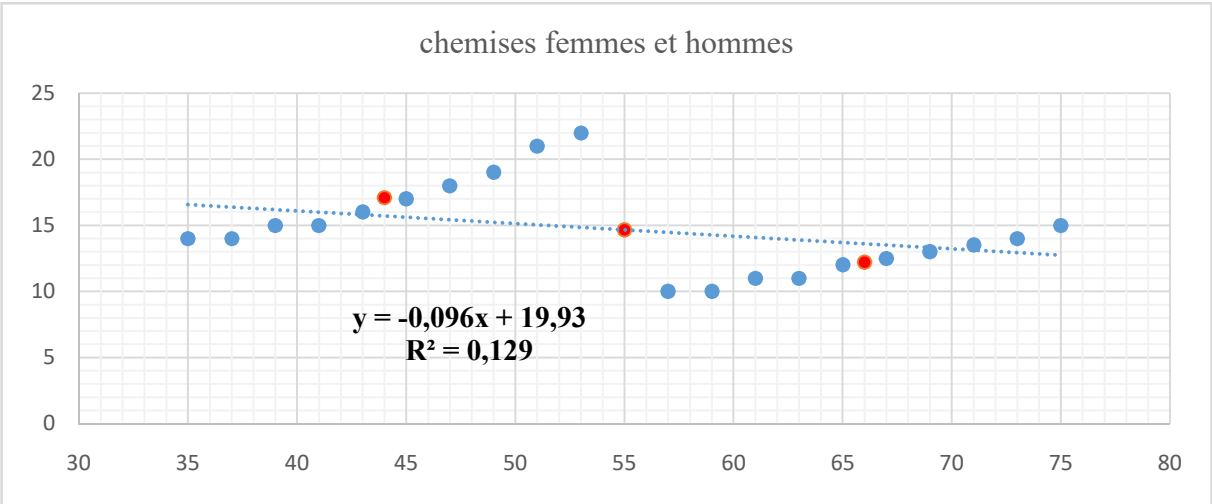
2.



Le point moyen est $G_h(66\,;12,2)$

Le graphique montre ici encore des points tous pratiquement alignés. La droite d'ajustement est donc un bon modèle de la liaison entre t et p chez les hommes. La variable p y apparaît linéaire et croissante en fonction de t. Chez les hommes, le carré de cette corrélation linéaire est $R^2 = 0{,}9827$, encore plus proche de 1.

3.



Le point moyen est $G(55\,;14,7)$

Les points du nuage total n'apparaissent ***pas*** alignés. (R^2 est beaucoup plus proche de 0 que de 1). La relation statistique globale entre p et t est très loin d'être linéaire. En vérité, elle est affine par morceaux, avec une nette rupture entre les sous-graphes des hommes et des femmes.

Par conséquent, la droite d'ajustement des moindres carrés est un très mauvais modèle de cette liaison. Néanmoins, cette droite a un coefficient directeur négatif, ce qui indique une liaison

statistique plutôt décroissante ***globalement***, alors qu'elle est croissante chez les hommes et chez les femmes !

4.
La situation de G_f et G_h nous éclaire sur ce point : en moyenne, les vêtements pour femmes sont plus petits mais plus chers que ceux des hommes, et leur différence de taille est grande ! Donc, même si le prix augmente avec la taille chez les femmes et fait de même chez les hommes, dans la population totale, le prix d'une chemise est "statistiquement plutôt décroissant" avec la taille. On comprend qu'il est indispensable de ne pas se contenter du signe du coefficient de la droite des moindres carrés pour "conclure à un sens de liaison" : en réalité, le graphe montre que la liaison n'est ***ni croissante ni décroissante***, globalement. Pour pouvoir statuer sur le sens d'une liaison à partir de la droite des moindres carrés, il est essentiel de s'assurer d'abord que le nuage en est très proche.

Compléments sur les ajustements affines en annexe

Exercice 6 Moyenne, médiane et quartiles

Thèmes abordés

- **Moyenne, médiane, quartiles**

Énoncé

1. a. On sait que la médiane des notes des 31 élèves d'une classe à un devoir est égale à 10. Que peut-on en déduire pour la moyenne ?

b. On sait que la moyenne des notes d'un partiel pour 31 étudiants est égale à 10. Que peut-on en déduire pour la médiane ?

2. On sait que la médiane des notes d'un partiel est égale à 10, et l'écart interquartiles est égal à 5, Que peut-on en déduire pour les quartiles q_1 et q_3 ?

3. a. On sait que, pour 43 étudiants, la médiane des notes d'un partiel est égale à 10, $q_1 = 5$ et $q_3 = 12$.
Que peut-on en déduire pour la moyenne ?

b. Et si on sait de plus que la médiane des notes est égale à 10 ?

4. On sait que la moyenne des notes d'un partiel est égale à $\overline{x} = 10$ et l'écart-type à $s = 3$.

a. On sait qu'il y a strictement plus de 75% des notes dans l'intervalle $[\overline{x} - 2s ; \overline{x} + 2s]$.
Que peut-on en déduire pour la médiane ? les quartiles ?

b. Et si on suppose la distribution normale ?

Corrigé

1.a. Si la médiane est 10, l'effectif étant impair, il y a une note égale à 10 (la seizième dans l'ordre croissant), 15 notes inférieures ou égales à 10, et 15 notes supérieures ou égales à 10.
Dans le cas où les notes sont les plus basses possibles, on a donc :
15 notes égales à 0, et 16 notes égales à 10, ce qui donne la moyenne :

$$\frac{15\times 0 + 10 + 15\times 10}{31} = \frac{360}{31} \simeq 5{,}16$$

Dans le cas où les notes sont les plus hauts possibles, on a donc :
16 notes égales à 10, et 15 notes égales à 20, ce qui donne la moyenne :

$$\frac{15\times 10 + 10 + 15\times 20}{31} = \frac{360}{31} \simeq 14{,}84$$

Finalement, on ne peut donc pas avoir une moyenne inférieure à 5,16 ou supérieure à 14,84.

1.b. La médiane est la seizième des notes ordonnées. On regarde les cas extrêmes :

Supposons que 16 notes aient la valeur la plus basse possible (notée y) sachant que les 15 autres sont égales à 20. On a $\frac{16y + 15\times 20}{31} = 10$, donc $y = \frac{10}{16} = \frac{5}{8} = 0{,}625$

À l'opposé, supposons que 16 notes aient la valeur la plus haute possible (notée z) sachant que les 15 autres sont égales à 0.

On a $\frac{16z+15\times 0}{31}=10$, donc $z=\frac{310}{16}=\frac{155}{8}=19{,}375$

La médiane est comprise entre 0,625 et 19,375.

On constate donc que la médiane conditionne plus fortement la moyenne que la réciproque.

2. Le premier quartile q_1 est inférieur ou égal à la médiane m_e et le troisième quartile q_3 est supérieur ou égal à la médiane m_e.

Au plus bas, $q_3=m_e=10$, et $q_3-q_1=5$ implique alors $q_1=5$.

De même, au plus haut, $q_1=m_e=10$, et $q_3-q_1=5$ implique alors $q_3=15$,

En somme, $q_1\in[5;10]$ et $q_3=\in[10;15]$.

3. a. Examinons les situations extrêmes :

Les notes étant ordonnées dans l'ordre croissant, on a :

– les notes les plus basses possibles :

Comme $q_1=x_{11}=5$, on pourrait avoir $x_1=...=x_{10}=0$.

Et comme $q_3=x_{33}=12$, on pourrait avoir $x_{11}=...=x_{32}=5$ et $x_{33}=...=x_{43}=12$.

Donc $\overline{x}=\frac{10\times 0+22\times 5+11\times 12}{43}\simeq 5{,}63$

– les notes les plus hautes possibles :

Comme $q_1=x_{11}=5$, on pourrait avoir $x_1=...=x_{11}=5$.

Et comme $q_3=x_{33}=12$, on pourrait avoir $x_{12}=...=x_{33}=12$ et $x_{34}=...=x_{43}=20$.

Donc $\overline{x}=\frac{11\times 5+22\times 12+10\times 20}{43}\simeq 12{,}07$

b. Les notes les plus basses possibles :

Comme $q_1=x_{11}=5$, on pourrait avoir $x_1=...=x_{10}=0$.

Comme $m_e=x_{22}=10$, on pourrait avoir $x_{11}=...=x_{21}=5$

Et $q_3=x_{33}=12$, on pourrait avoir $x_{22}=...=x_{32}=10$ et $x_{33}=...=x_{43}=12$.

Donc $\overline{x}=\frac{10\times 0+11\times 5+11\times 10+11\times 12}{43}\simeq 6{,}9$

Les notes les plus hautes possibles :

Comme $q_1=x_{11}=5$, on pourrait avoir $x_1=...=x_{11}=5$.

Comme $m_e=x_{22}=10$, on pourrait avoir $x_{12}=...=x_{22}=10$

Et $q_3=x_{33}=12$, on pourrait avoir $x_{23}=...=x_{33}=12$ et $x_{34}=...=x_{43}=20$.

Donc $\overline{x}=\frac{11\times 5+11\times 10+11\times 12+10\times 20}{43}\simeq 11{,}56$

4. a. On sait seulement qu'il y a strictement plus de 75% des données dans $[\overline{x}-2s;\overline{x}+2s]$.

Il y a donc strictement moins de 25% des données strictement inférieures à 4, donc $q_1\geqslant 4$,

et de même strictement moins de 25% des données strictement supérieures à 16 donc $q_3\leqslant 16$.

Ainsi, $4\leqslant q_1\leqslant m_e\leqslant q_3\leqslant 16$

b. On utilise la loi $N\left(\mu = 10\,,\sigma^2 = 3^2\right)$

$m_e = \overline{x} = 10$	$P\left(X \leqslant 10\right) = 0,5$
$q_1 = 8 \quad (7,9765...)$	$P\left(X \leqslant 18\right) = 0,25$
$q_3 = 12 \quad (12,02...)$	$P\left(X \leqslant 12\right) = 0,75$

De façon générale : si on pose $Z = \dfrac{X - \overline{x}}{\sigma}$, Z suit la loi $N\left(0\,,1\right)$ et

$m_e = \overline{x}$	$P\left(Z \leqslant 0\right) = 0,5$
$q_1 = \overline{x} - 0,6745\sigma$	$P\left(Z \leqslant -0,6745\right) = 0,25$
$q_3 = \overline{x} + 0,6745\sigma$	$P\left(X \leqslant -0,6745\right) = 0,75$

Exercice 7 Encadrements de quantiles

Thèmes abordés

- **Moyenne, médiane, quartiles**

Énoncé

Un échantillon statistique $\{x_1,...,x_n\}$ a une moyenne empirique $\bar{x}$ et un écart-type empirique $s\neq 0$. On cherche à en déduire certains encadrements sur les quantiles de cet échantillon.

Pour $\alpha>0$, on note $I_\alpha=[\bar{x}-\alpha s\,,\bar{x}+\alpha s]$, $E_\alpha=\{x_i\,/\,|x_i-\bar{x}|\leqslant \alpha s\}$,

$$E_\alpha^c=\{x_i\,/\,|x_i-\bar{x}|> \alpha s\} \text{ et } p_\alpha=card\left(E_\alpha^c\right)$$

1. Établir que $n-p_\alpha>n\left(1-\frac{1}{\alpha^2}\right)$.

2. Quels encadrements peut-on en déduire pour les quartiles avec $\alpha=2$ et avec $\alpha=\sqrt{2}$?

3. Peut-on faire mieux ?

Corrigé

1. La variance empirique est $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\left(\sum_{x_i\in E_\alpha}(x_i-\bar{x})^2+\sum_{x_i\in E_\alpha^c}(x_i-\bar{x})^2\right)$

d'où $ns^2=\sum_{x_i\in E_\alpha}(x_i-\bar{x})^2+\sum_{x_i\in E_\alpha^c}(x_i-\bar{x})^2$

Si E_α^c est non vide,

$$\forall x_i\in E_\alpha^c\ ,\ (x_i-\bar{x})>\alpha^2 s^2>0\Rightarrow \sum_{x_i\in E_\alpha^c}(x_i-\bar{x})^2>p_\alpha\alpha^2 s^2$$

$$\Rightarrow ns^2>\sum_{x_i\in E_\alpha}(x_i-\bar{x})^2+p_\alpha\alpha^2 s^2$$

$$\Rightarrow ns^2-p_\alpha\alpha^2 s^2>\sum_{x_i\in E_\alpha}(x_i-\bar{x})^2\geqslant 0$$

$$\Rightarrow s^2\left(n-p_\alpha\alpha^2\right)>0\ \Rightarrow\ n-p_\alpha\alpha^2>0\Rightarrow\ -p_\alpha>\frac{-n}{\alpha^2}$$

$$\Rightarrow\ Card\left(E_\alpha\right)=n-p_\alpha>n\left(1-\frac{1}{\alpha^2}\right)$$

si E_α^c est vide, $Card\left(E_\alpha\right)=n>n\left(1-\frac{1}{\alpha^2}\right)$.

La proportion des observations qui se trouvent dans l'intervalle $I_\alpha = [\bar{x} - \alpha s\,, \bar{x} + \alpha s\,]$ est donc strictement supérieure à $1 - \frac{1}{\alpha^2}$.

2. Pour $\alpha = 2$, on obtient $[\bar{x} - 2s; \bar{x} + 2s]$,

$Card(E_2) > n\left(1 - \frac{1}{4}\right) = \frac{3}{4}n \Leftrightarrow Card\left(E_2^c\right) < \frac{1}{4}n$

D'où $q_1 \geqslant \bar{x} - 2s$ et $q_3 \leqslant \bar{x} + 2s$

Pour $\alpha = \sqrt{2}$, on obtient $\left[\bar{x} - \sqrt{2}s\,; \bar{x} + \sqrt{2}s\right]$,

$Card\left(E_{\sqrt{2}}\right) > n\left(1 - \frac{1}{2}\right) = \frac{1}{2}n \Leftrightarrow Card\left(E_{\sqrt{2}}^c\right) < \frac{1}{2}n$

D'où $\boxed{\bar{x} - \sqrt{2}s \leqslant q_2 \leqslant \bar{x} + \sqrt{2}s}$

Ayant encadré la médiane et sachant qu'elle se trouve entre les deux autres quartiles, on en déduit :

$\bar{x} - 2s \leqslant q_1 \leqslant q_2 \leqslant \bar{x} + \sqrt{2}s$ donc $\boxed{\bar{x} - 2s \leqslant q_1 \leqslant \bar{x} + \sqrt{2}s}$

$\bar{x} - \sqrt{2}s \leqslant q_2 \leqslant q_3 \leqslant \bar{x} + 2s$ donc $\boxed{\bar{x} - \sqrt{2}s \leqslant q_3 \leqslant \bar{x} + 2s}$

Mais il s'agit de minorants et de majorants, pas d'extrema, comme on le verra ci-dessous.

3. Il semble que oui : (on donnera des exemples pour $\bar{x} = 10$ et $s = 3$)

On utilisera le fait que :

> $\sum_n (x_i - x)^2$, pour x fixé est minimum si $x_i = \bar{x} \quad \forall i \in [\![1\,;n]\!]$
>
> c'est-à-dire : $\sum_n (x_i - x)^2 \geqslant n(\bar{x} - x)^2 \qquad \forall (x_i)_{i=1,\ldots,n}$

En effet :

$$\sum_n (x_i - x)^2 - \sum_n (\bar{x} - x)^2 = \sum_n x_i^2 - 2x\sum_n x_i + nx^2 - n\bar{x}^2 + 2nx\bar{x} - nx^2 = n\left(\frac{1}{n}\sum_n x_i^2 - \bar{x}^2\right) = ns^2 \geqslant 0$$

On pourra regarder les cas particuliers ou directement le cas général donné en conclusion.

a. Minoration des quartiles

- Pour le premier quartile
  On se place dans le cas où n=4p.
  Les données étant ordonnées dans l'ordre croissant, $q_1 = x_p$.
  On a les deux contraintes : $\bar{x} = \frac{1}{n}\sum_{n\,\text{données}} x_i$ et $s^2 = \frac{1}{n}\sum_{n\,\text{données}} (x_i - \bar{x})^2$.

Si a est la moyenne des p premières données, alors

$\sum_p x_i = pa$ et $\sum_p (x_i - \bar{x})^2 \geqslant p(a - \bar{x})^2$. Il y a égalité si les p premières valeurs sont égales à a.

Si b est la moyenne des $3p$ x_i restants,

$\sum_{3p} x_i = 3\,pb$ et $\sum_{3p} (x_i - \bar{x})^2 \geqslant 3p\,(b-\bar{x})^2$. Il y a égalité si les *n-p* premières valeurs sont égales à *a*.

Donc $\bar{x} = \frac{pa+3pb}{4p} = \frac{1}{4}a + \frac{3}{4}b$. Ainsi : $\frac{1}{4}a + \frac{3}{4}b = \bar{x}$

et $ns^2 = \sum_{p} (x_i - \bar{x})^2 + \sum_{3p} (x_i - \bar{x})^2 \geqslant p\,(a-\bar{x})^2 + 3p\,(b-\bar{x})^2$

ce qui impose : $\frac{1}{4}(a-\bar{x})^2 + \frac{3}{4}(b-\bar{x})^2 \leqslant s^2$

$b = \frac{4\bar{x}-a}{3}$ donc $\frac{1}{4}(a-\bar{x})^2 + \frac{1}{12}(\bar{x}-a)^2 \leqslant s^2$

Donc $4(\bar{x}-a)^2 \leqslant 12s^2$ soit $(\bar{x}-a)^2 - 3s^2 \leqslant 0$

D'où $\bar{x} - \sqrt{3}s \leqslant a \leqslant \bar{x} + \sqrt{3}s$

Comme $q_1 \geqslant a$, la plus petite valeur de q_1 est donc $\bar{x} - \sqrt{3}s$, atteinte si les *p* premières valeurs sont égales à *a* et les *n-p* dernières valeurs égales à *b*.

$\boxed{q_1 \geqslant \bar{x} - \sqrt{3}s}$

et si $q_1 = a = \bar{x} - \sqrt{3}s$, $b = \bar{x} + \frac{1}{\sqrt{3}}s = m_e = q_3$, l'étendue est $\left(\sqrt{3} + \frac{1}{\sqrt{3}}\right)s$

$x_i = \bar{x} - \sqrt{3}s$ pour *i* allant de 1 à *p* et $x_i = \bar{x} + \frac{1}{\sqrt{3}}s$ pour *i* allant de *p* + 1 à *n*.

Exemple
La série (4,8 ; 4,8 ; 11,73 ; 11,73 ; 11,73 ; 11,73) vérifie les conditions. Ce minimum est donc atteint (sur l'ensemble des séries)

- Pour le second quartile
  on suppose ici que $n=2k$, et alors $q_2 = x_k$.

Si a est la moyenne des *k* premières données

$\sum_{k} x_i = ka$ et $\sum_{k} (x_i - \bar{x})^2 \geqslant k\,(a-\bar{x})^2$

Si *b* est la moyenne des *k* x_i restants,

$\sum_{k} x_i = kb$ et $\sum_{k} (x_i - \bar{x})^2 \geqslant k\,(b-\bar{x})^2$

Donc $\bar{x} = \frac{ka+kb}{2k} = \frac{1}{2}a + \frac{1}{2}b$

et $ns^2 = \sum_{k} (x_i - \bar{x})^2 + \sum_{k} (x_i - \bar{x})^2 \geqslant k\,(a-\bar{x})^2 + k\,(b-\bar{x})^2$

ce qui impose : $\frac{1}{2}(a-\bar{x})^2 + \frac{1}{2}(b-\bar{x})^2 \leqslant s^2$

Comme $b = \frac{2\bar{x}-a}{2}$, on a $(a-\bar{x})^2 \leqslant s^2$.

D'où $\bar{x} - s \leqslant a \leqslant \bar{x} + s$.

Comme $q_2 \geqslant a$, la plus petite valeur de q_2 est égale à $\bar{x} - s$.

$\boxed{q_2 \geqslant \bar{x} - s}$

Exemple
La série (7,7,7,7,13,13,13,13).

- Pour le troisième quartile

Si a est la moyenne des $3p$ premières données

$\sum_{3p} x_i = 3pa$ et $\sum_{3p} (x_i - \overline{x})^2 \geqslant 3p(a - \overline{x})^2$

Si b est la moyenne des p x_i restants,

$\sum_{p} x_i = pb$ et $\sum_{p} (x_i - \overline{x})^2 \geqslant p(b - \overline{x})^2$

Donc $\overline{x} = \frac{3}{4}a + \frac{1}{4}b$

et $ns^2 = \sum_{p} (x_i - \overline{x})^2 + \sum_{3p} (x_i - \overline{x})^2 \geqslant 3p(a - \overline{x})^2 + p(b - \overline{x})^2$

ce qui impose : $\frac{3}{4}(a - \overline{x})^2 + \frac{1}{4}(b - \overline{x})^2 \leqslant s^2$

Comme $\overline{x} = \frac{3}{4}a + \frac{1}{4}b$, $b = 4\overline{x} - 3a$ donc $12(a - \overline{x})^2 \leqslant 4s^2$

D'où $\overline{x} - \frac{1}{\sqrt{3}}s \leqslant a \leqslant \overline{x} + \frac{1}{\sqrt{3}}s$ soit $8{,}26... \leqslant a \leqslant 10$

Comme $q_3 \geqslant a$, la plus petite valeur de q_3 est donc égale à $\overline{x} - \frac{1}{\sqrt{3}}s$.

$$\boxed{q_3 \geqslant \overline{x} - \frac{1}{\sqrt{3}}s}$$

Exemple
La série (8,26 ; 8,26 ; 8,26 ; 8,26 ; 8,26 ; 8,26 ; 15,19 ; 15,19)

b. Majoration des quartiles

- le premier quartile.
  On se place dans le cas où $n = 4p$, donc $q_1 = x_p$.

Si a est la moyenne des p-1 premières données

$\sum_{p-1} x_i = (p-1)a$ et $\sum_{p-1} (x_i - \overline{x})^2 \geqslant (p-1)(a - \overline{x})^2$

Si b est la moyenne des $3p+1$ x_i restants,

$\sum_{3p+1} x_i = (3p+1)b$ et $\sum_{3p+1} (x_i - \overline{x})^2 \geqslant (3p+1)(b - \overline{x})^2$

De sorte que $(p-1)a + (3p+1)b = n\overline{x}$

et $ns^2 = \sum_{p-1} (x_i - \overline{x})^2 + \sum_{3p+1} (x_i - \overline{x})^2 \geqslant (p-1)(a - \overline{x})^2 + (3p+1)(b - \overline{x})^2$

soit $(p-1)(a - \overline{x})^2 + (3p+1)(b - \overline{x})^2 \leqslant ns^2$

Puisque $a = \frac{n\overline{x} - (3p+1)b}{p-1}$, on obtient

$$\frac{(3p+1)^2}{(p-1)}(\bar{x}-b)^2+(3p+1)(b-\bar{x})^2 \leqslant ns^2$$

$$(3p+1)^2(b-\bar{x})^2+(p-1)(3p+1)(b-\bar{x})^2 \leqslant n(p-1)s^2$$

$$4p(3p+1)(b-\bar{x})^2 \leqslant 4p(p-1)s^2$$

$$(b-\bar{x})^2 \leqslant \frac{(p-1)s^2}{(3p+1)}$$

Et $\bar{x}-s\sqrt{\frac{p-1}{3p+1}} \leqslant b \leqslant \bar{x}+s\sqrt{\frac{p-1}{3p+1}}$

Comme $q_1 \leqslant b$, la plus grande valeur de q_1 est donc égale à $\bar{x}+s\sqrt{\frac{p-1}{3p+1}}$.

$$\boxed{q_1 \leqslant \bar{x}+s\sqrt{\frac{p-1}{3p+1}}}$$

Exemple
(2,062 ; 11,134 ; 11,134 ; 11,134 ; 11,134 ; 11,134 ; 11,134 ; 11,134)

En particulier on a aussi $\boxed{\bar{x}-\sqrt{3}\,s \leqslant q_1 < \bar{x}+\frac{1}{\sqrt{3}}s}$ puisque $\frac{p-1}{3p+1} < \frac{p+1/3}{3p+1}$.

- le second quartile.
  On se place dans le cas où $n=2k$, donc $q_2 = x_k$

Si a est la moyenne des k-1 premières données

$\sum_{k-1} x_i = (k-1)a$ et $\sum_{k}(x_i-\bar{x})^2 \geqslant (k-1)(a-\bar{x})^2$

Si b est la moyenne des $k+1$ x_i restants,

$\sum_{k+1} x_i = (k+1)b$ et $\sum_{k}(x_i-\bar{x})^2 \geqslant (k+1)(b-\bar{x})^2$

De sorte que $(k-1)a+(k+1)b = n\,\bar{x}$

et $ns^2 = \sum_{k-1}(x_i-\bar{x})^2 + \sum_{k+1}(x_i-\bar{x})^2 \geqslant (k-1)(a-\bar{x})^2 + (k+1)(b-\bar{x})^2$

soit $(k-1)(a-\bar{x})^2+(k+1)(b-\bar{x})^2 \leqslant ns^2$

Puisque $a = \frac{nx-(k+1)b}{k-1}$, on obtient $(b-\bar{x})^2 \leqslant s^2\,\frac{k-1}{k+1}$

Et $\bar{x}-s\sqrt{\frac{k-1}{k+1}} \leqslant b \leqslant \bar{x}+s\sqrt{\frac{k-1}{k+1}}$

Comme $q_2 \leqslant b$, la plus grande valeur de q_2 est donc égale à $\bar{x}+s\sqrt{\frac{k-1}{k+1}}$.

$$\boxed{q_2 \leqslant \bar{x}+s\sqrt{\frac{k-1}{k+1}}}$$

Exemple
(6,13 ; 6,13 ; 6,13 ; 12,32 ; 12,32 ; 12,32 ; 12,32 ; 12,32)

En particulier on a aussi : $\boxed{\overline{x}-s\leqslant q_2\leqslant\overline{x}+s}$

- le troisième quartile.
  On se place dans le cas où $n = 4p$, donc $q_3 = x_{3p}$

Si a est la moyenne des $3p - 1$ premières données

$\sum_{3p-1} x_i=(3p-1)a$ et $\sum_{3p-1}(x_i-\overline{x})^2\geqslant(3p-1)(a-\overline{x})^2$

Si b est la moyenne des $p + 1$ x_i restants,

$\sum_{p+1} x_i=(p+1)b$ et $\sum_{p+1}(x_i-\overline{x})^2\geqslant(p+1)(b-\overline{x})^2$

De sorte que $(3p-1)a+(p+1)b=n\,\overline{x}$

et $ns^2=\sum_{3p-1}(x_i-\overline{x})^2+\sum_{p+1}(x_i-\overline{x})^2\geqslant(3p-1)(a-\overline{x})^2+(p+1)(b-\overline{x})^2$

soit $(3p-1)(a-\overline{x})^2+(p+1)(b-\overline{x})^2\leqslant ns^2$

Puisque $a=\dfrac{n\overline{x}-(p+1)b}{3p-1}$, on obtient

$$\frac{(p+1)^2}{(3p-1)}(\overline{x}-b)^2+(p+1)(b-\overline{x})^2\leqslant ns^2$$

$$(p+1)^2(b-\overline{x})^2+(3p-1)(p+1)(b-\overline{x})^2\leqslant n(3p-1)s^2$$

$$4p(p+1)(b-\overline{x})^2\leqslant 4p(3p-1)s^2$$

$$(b-\overline{x})^2\leqslant\frac{(3p-1)s^2}{(p+1)}$$

et $\overline{x}-s\sqrt{\dfrac{3p-1}{p+1}}\leqslant b\leqslant\overline{x}+s\sqrt{\dfrac{3p-1}{p+1}}$

Comme $q_3\leqslant b$, la plus grande valeur de q_3 est égale à $\overline{x}+s\sqrt{\dfrac{3p-1}{p+1}}$.

$$\boxed{q_3\leqslant\overline{x}+s\sqrt{\frac{3p-1}{p+1}}}$$

Exemple
(7,678 ; 7,678 ; 7,678 ; 7,678 ; 7,678 ; 13,87 ; 13,87 ; 13,87)

En particulier $\boxed{\overline{x}-\dfrac{1}{\sqrt{3}}s\leqslant q_3<\overline{x}+\sqrt{3}\,s}$

Conclusion

On sait que :

Taille de la série	q_1	q_2	m_e	q_3
$n = 4p$	x_p	x_{2p}	$\frac{x_{2p}+x_{2p+1}}{2}$	x_{3p}
$n = 4p+1$	x_{p+1}	x_{2p+1}	x_{2p+1}	x_{3p+1}
$n = 4p+2$	x_{p+1}	x_{2p+1}	$\frac{x_{2p+1}+x_{2p+2}}{2}$	x_{3p+2}
$n = 4p+3$	x_{p+1}	x_{2p+2}	x_{2p+2}	x_{3p+3}

On pourrait répéter ce raisonnement pour chacun des autres cas. On notera que la taille de la série peut intervenir dans les résultats, mais que l'on obtient ainsi les extrema, et des encadrements plus précis.

De façon générale

Pour minorer x_k :

Si *a* est la moyenne des *k* premières données

$\sum_k x_i = ka$ et $\sum_k (x_i - \overline{x})^2 \geqslant k(a-\overline{x})^2$. Il y a égalité si les *k* premières valeurs sont égales à *a*.

Si *b* est la moyenne des $n-k$ x_i restants,

$\sum_{n-k} x_i = (n-k)b$ et $\sum_k (x_i - \overline{x})^2 \geqslant (n-k)(b-\overline{x})^2$. Il y a égalité si les *n-k* dernières valeurs sont égales à *b*.

De sorte que $n\overline{x} = ka + (n-k)b$

et $ns^2 = \sum_k (x_i - \overline{x})^2 + \sum_{n-k} (x_i - \overline{x})^2 \geqslant k(a-\overline{x})^2 + (n-k)(b-\overline{x})^2$

ce qui impose : $k(a-\overline{x})^2 + (n-k)(b-\overline{x})^2 \leqslant ns^2$.

$$b = \frac{n\overline{x} - ka}{n-k} \quad \Rightarrow \quad k(a-\overline{x})^2 + (n-k)\left(\frac{k}{(n-k)}(a-\overline{x})\right)^2 \leqslant ns^2 \Rightarrow$$

$$(a-\overline{x})^2\left(k + \frac{k^2}{(n-k)}\right) \leqslant ns^2 \Rightarrow (a-\overline{x})^2 \leqslant \frac{n(n-k)}{nk}s^2 \Rightarrow \overline{x} - s\sqrt{\frac{(n-k)}{k}} \leqslant a \leqslant \overline{x} + s\sqrt{\frac{(n-k)}{k}}$$

Comme $x_k \geqslant a$, la plus petite valeur de x_k est donc $\overline{x} - s\sqrt{\frac{(n-k)}{k}}$, atteinte si les *k* premières valeurs sont égales à *a* et les *n-k* dernières valeurs égales à *b*.

$$\boxed{x_k \geqslant \overline{x} - s\sqrt{\frac{(n-k)}{k}}}.$$

Pour majorer x_k :

Si a est la moyenne des $k-1$ premières données,

$\sum_{k-1} x_i = (k-1)a$ et $\sum_{k-1} (x_i - \bar{x})^2 \geqslant (k-1)(a-\bar{x})^2$. Il y a égalité si les k premières valeurs sont égales à a.

Si b est la moyenne des $n-k+1$ x_i restants,

$\sum_{n-k+1} x_i = (n-k+1)b$ et $\sum_{n-k+1} (x_i - \bar{x})^2 \geqslant (n-k+1)(b-\bar{x})^2$. Il y a égalité si les $n-k$ dernières valeurs sont égales à b.

De sorte que $n\bar{x} = (k-1)a + (n-k+1)b$

et $ns^2 = \sum_{k-1} (x_i - \bar{x})^2 + \sum_{n-k+1} (x_i - \bar{x})^2 \geqslant (k-1)(a-\bar{x})^2 + (n-k+1)(b-\bar{x})^2$

ce qui impose : $(k-1)(a-\bar{x})^2 + (n-k+1)(b-\bar{x})^2 \leqslant ns^2$.

$$a = \frac{n\bar{x} - (n-k+1)b}{k-1} \Rightarrow (k-1)\left(\frac{(n-k+1)}{(k-1)}(b-\bar{x})\right)^2 + (n-k+1)(b-\bar{x})^2 \leqslant ns^2 \Rightarrow$$

$$(b-\bar{x})^2\left(\frac{(n-k+1)^2}{(k-1)} + (n-k+1)\right) \leqslant ns^2 \Rightarrow (b-\bar{x})^2 \leqslant \frac{(k-1)}{(n-k+1)}s^2 \Rightarrow$$

$$\bar{x} - s\sqrt{\frac{k-1}{(n-k+1)}} \leqslant b \leqslant \bar{x} + s\sqrt{\frac{k-1}{(n-k+1)}}$$

Comme $x_k \leqslant b$, la plus grande valeur de x_k est donc $\bar{x} + s\sqrt{\frac{k-1}{(n-k+1)}}$ atteinte si les k premières valeurs sont égales à a et les n-k dernières valeurs égales à b.

$$\boxed{x_k \leqslant \bar{x} + s\sqrt{\frac{k-1}{(n-k+1)}}}$$

Alternative :

Par symétrie, à partir de la minoration de x_k :

Comme $b = \frac{n\bar{x} - ka}{n-k}$, $a \geqslant \bar{x} - s\sqrt{\frac{(n-k)}{k}} \Rightarrow b \leqslant \frac{n\bar{x} - k\left(\bar{x} - s\sqrt{\frac{(n-k)}{k}}\right)}{n-k} = \bar{x} + s\sqrt{\frac{k}{(n-k)}}$

D'où $x_{k+1} \leqslant b \leqslant \bar{x} - s\sqrt{\frac{k}{(n-k)}}$, et enfin $x_k \leqslant \bar{x} + s\sqrt{\frac{k-1}{n-k+1}}$

Remarque

Pour la médiane, dans le cas pair, on prendra pour a la moyenne des $k+1$ premières données pour minorer, et $k-1$ premières données pour majorer.

Autre méthode

On utilisera une formule de Huygens : l'équation d'analyse de la variance.

- **Les données ne sont pas nécessairement ordonnées**
- $k \in \{1,...,n-1\}$
- On note $\overline{x}_1$ et s_1 la moyenne et l'écart type des *k* premières valeurs
- On note $\overline{x}_2$ et s_2 la moyenne et l'écart type des *n-k* dernières valeurs

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{k}(x_i - \overline{x})^2 + \sum_{i=k+1}^{n}(x_i - \overline{x})^2\right)$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{k}(x_i - \overline{x}_1 + \overline{x}_1 - \overline{x})^2 + \sum_{i=k+1}^{n}(x_i - \overline{x}_2 + \overline{x}_2 - \overline{x})^2\right)$$

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{k}(x_i - \overline{x}_1)^2 + \sum_{i=1}^{k}(\overline{x}_1 - \overline{x})^2 + \sum_{i=k+1}^{n}(x_i - \overline{x}_2)^2 + \sum_{i=k+1}^{n}(\overline{x}_2 - \overline{x})^2\right)$$

$$+\frac{1}{n}\left(2\sum_{i=1}^{k}(x_i - \overline{x}_1)(\overline{x}_1 - \overline{x}) + 2\sum_{i=k+1}^{n}(x_i - \overline{x}_2)(\overline{x}_2 - \overline{x})\right)$$

$$s^2 = \frac{1}{n}\left(ks_1^2 + (n-k)s_2^2 + k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2\right)$$

Puisque

$$\sum_{i=1}^{k}(x_i - \overline{x}_1)(\overline{x}_1 - \overline{x}) + \sum_{i=k+1}^{n}(x_i - \overline{x}_2)(\overline{x}_2 - \overline{x}) = (\overline{x}_1 - \overline{x})\sum_{i=1}^{k}(x_i - \overline{x}_1) + (\overline{x}_2 - \overline{x})\sum_{i=k+1}^{n}(x_i - \overline{x}_2).$$

et $\sum_{i=1}^{k}(x_i - \overline{x}_1) = \sum_{i=k+1}^{n}(x_i - \overline{x}_2) = 0$

D'où enfin

$$s^2 = \frac{ks_1^2 + (n-k)s_2^2}{n} + \frac{k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2}{n}$$

$$\boxed{s^2 = \left(\frac{k}{n}s_1^2 + \frac{(n-k)}{n}s_2^2\right) + \left(\frac{k}{n}(\overline{x}_1 - \overline{x})^2 + \frac{(n-k)}{n}(\overline{x}_2 - \overline{x})^2\right)}$$

s^2 est la « somme de la moyenne des variances et de la variance des moyennes ».

Remarque

On écrit aussi $s^2 = s_{\text{intra}}^2 + s_{\text{inter}}^2$

s_{intra}^2 est la variance intra-classe, moyenne des variances des classes.

s_{inter}^2 est la variance inter-classes, la variance des moyennes des classes.

Ici, nous avons deux classes : la classe de *k* valeurs, et la classe des autres valeurs.
Cette décomposition s'étend naturellement à plus de deux classes.

Cherchons la valeur minimale possible pour x_k

- Les données sont ordonnées dans l'ordre croissant.
- $\overline{x}$ et s sont fixés.
- x_k est la plus grande des *k* premières valeurs

On a : $n\overline{x} = k\overline{x}_1 + (n-k)\overline{x}_2$ **(1)**

et $ns^2 = ks_1^2 + (n-k)s_2^2 + k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2$ **(2)**

Pour $\overline{x}_1$ donnée, la plus petite valeur possible de x_k est $\overline{x}_1$, $\overline{x}_1 \leqslant x_k$, puisque $\overline{x}_1 \leqslant x_k$, ce qui impose $s_1 = 0$.

(2) devient $ns^2 = (n-k)s_2^2 + k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2$ **(2')**

Il faut donc à présent chercher $\overline{x}_1$ minimale. À $\overline{x}$ fixée, diminuer $\overline{x}_1$ implique via **(1)** d'augmenter $\overline{x}_2$. La quantité $k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2$ croît, en effet, d'après **(1)** encore

$$k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2 = \frac{(n-k)^2}{k}(\overline{x}_2 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2 = \frac{n(n-k)}{k}(\overline{x}_2 - \overline{x})^2$$

et $\overline{x} \leqslant \overline{x}_2$.

Ainsi, et en conséquence de **(2')**, s_2 diminue. Jusqu'où peut-elle aller ?

$s_2^2 \geqslant 0$ (3) $\Rightarrow$ $k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2 \leqslant ns^2$

La limite possible est donc atteinte lorsque $s_2 = 0$. Les $n-k$ plus grandes valeurs sont alors égales, et elles sont égales à $\overline{x}_2$.

(2') devient alors : $k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2 = ns^2$ **(2'')**

Il s'agit finalement de résoudre le système :

$$\begin{cases} k\overline{x}_1 + (n-k)\overline{x}_2 = n\overline{x} & \textbf{(1)} \\ k(\overline{x}_1 - \overline{x})^2 + (n-k)(\overline{x}_2 - \overline{x})^2 = ns^2 & \textbf{(2'')} \end{cases}$$

(1) $\Rightarrow$ $k(\overline{x}_1 - \overline{x}) + (n-k)(\overline{x}_2 - \overline{x}) = 0$ $\Rightarrow$ $(\overline{x}_2 - \overline{x}) = -\frac{k}{(n-k)}(\overline{x}_1 - \overline{x})$ **(1')**

(2'') $\Rightarrow$ $k(\overline{x}_1 - \overline{x})^2 + (n-k)\left(-\frac{k}{(n-k)}(\overline{x}_1 - \overline{x})\right)^2 = ns^2$ $\Rightarrow$ $(\overline{x}_1 - \overline{x})^2\left(k + \frac{k^2}{(n-k)}\right) = ns^2$

$\Rightarrow$ $(\overline{x}_1 - \overline{x})^2 = s^2\frac{(n-k)}{k}$.

Comme on a nécessairement $\overline{x}_1 \leqslant \overline{x}$, on prendra la racine négative, ce qui donne :

$\overline{x}_1 = \overline{x} - s\sqrt{\frac{(n-k)}{k}}$, ce qui entraîne $x_k \geqslant \overline{x} - s\sqrt{\frac{(n-k)}{k}}$ **(3)**

Par symétrie, la valeur *maximale* possible pour x_{k+1} est alors $\overline{x}_2$, donné par l'équation **(1')** :

$$(\overline{x}_2 - \overline{x}) = -\frac{k}{(n-k)}(\overline{x}_1 - \overline{x}) \Rightarrow \overline{x}_2 = \overline{x} + \frac{k}{(n-k)}s\sqrt{\frac{(n-k)}{k}} \Rightarrow \overline{x}_2 = \overline{x} + s\sqrt{\frac{k}{(n-k)}}$$

$\Rightarrow$ $\overline{x}_{k+1} \leqslant \overline{x} + s\sqrt{\frac{k}{(n-k)}}$ et donc $\overline{x}_k \leqslant \overline{x} + s\sqrt{\frac{k-1}{(n-k+1)}}$ **(4)**

Application aux quartiles :

Supposons $n = 4k-1$, soit $k = \frac{n+1}{4}$ qui est entier. Le 1er quartile est x_k .

On obtient : $\frac{(n-k)}{k}=\frac{3n-1}{n+1}$, et **(3)** $\Rightarrow$ $x_{\frac{n+1}{4}} \geqslant \overline{x}-s\sqrt{\frac{3n-1}{n+1}} \simeq \overline{x}-\sqrt{3}\,s$ pour n grand.

Pour le troisième quartile : **(4)** $\Rightarrow$ $x_{\frac{3}{4}(n+1)} \geqslant \overline{x}-s\sqrt{\frac{n-1}{3n-1}} \simeq \overline{x}-\frac{1}{\sqrt{3}}\,s$ pour n grand.

Cherchons à présent la valeur *maximale* pour x_k

Nous avons vu que $\forall k \in \{1,...,n-1\}$, $\overline{x}_{k+1} \leqslant \overline{x}+s\sqrt{\frac{k}{(n-k)}}$.

En prenant $k'=k-1$, on obtient : $\forall k' \in \{1,...,n-1\}$, $\overline{x}_{k+1} \leqslant \overline{x}+s\sqrt{\frac{k'}{(n-k')}}$

$$\forall k \in \{2,...,n\} \text{ , } \overline{x}_k \leqslant \overline{x}+s\sqrt{\frac{k-1}{(n-k+1)}}$$

Application aux quartiles :
Toujours avec $n=4k-1$

$$x_{\frac{n+1}{4}} \leqslant \overline{x}+s\sqrt{\frac{\frac{n+1}{4}-1}{\frac{3n-1}{4}+1}}=\overline{x}+s\sqrt{\frac{n-3}{3n+3}} \simeq \overline{x}+\sqrt{\frac{1}{3}}\,s \text{ pour } n \text{ grand.}$$

$$x_{\frac{3}{4}(n+1)} \leqslant \overline{x}+s\sqrt{\frac{\frac{3}{4}(n+1)-1}{\frac{1}{4}(n+1)+1}}=\overline{x}+s\sqrt{\frac{3n+2}{n+1}} \simeq \overline{x}+\sqrt{3}\,s \text{ pour } n \text{ grand.}$$

Application numérique :

$$\forall k \in \{1,...,n-1\} \text{ , } \overline{x}-s\sqrt{\frac{(n-k)}{k}} \leqslant \overline{x}_k \leqslant \overline{x}+s\sqrt{\frac{k-1}{(n-k+1)}}$$

Si $n=15$, $\overline{x}=10$ et $s=3$.
Pour le premier quartile : $q_1=x_4$

$$\overline{x}-s\sqrt{\frac{11}{4}} \leqslant q_1 \leqslant \overline{x}+s\sqrt{\frac{3}{12}}$$

$$10-\frac{3}{2}\sqrt{11} \leqslant q_1 \leqslant 10+\frac{3}{2}$$

$$5{,}025 \leqslant q_1 \leqslant 11{,}5$$

Pour le deuxième quartile (médiane) : $q_2=x_8$

$$\overline{x}-s\sqrt{\frac{7}{8}} \leqslant q_2 \leqslant \overline{x}+s\sqrt{\frac{7}{8}}$$

$$7{,}194 \leqslant q_2 \leqslant 12{,}806$$

Pour le troisième quartile : $q_3=x_{12}$

$$\overline{x}-s\sqrt{\frac{1}{4}} \leqslant q_2 \leqslant \overline{x}+s\sqrt{\frac{11}{4}}$$

$$8{,}5 \leqslant q_2 \leqslant 14{,}975$$

Exercice 8 Péréquations

Thèmes abordés

- **Moyenne, écart-type**
- **Transformations affines**

Énoncé 1

1. Les candidats à un concours ont eu le choix entre deux options, et on observe après correction que : pour l'option A, la moyenne est 13, et l'écart-type est 4 ;
pour l'option B, la moyenne est 9, et l'écart-type est 7.
Ne désirant pas discriminer par le choix de l'option, le jury décide de modifier par transformation affine toutes les notes de façon à obtenir la même moyenne, 10, et le même écart-type, 6 pour chaque option.
a. Déterminer les deux transformations affines nécessaires.
b. Un candidat ayant choisi l'option A avait obtenu la note 12. Quelle sera sa note après péréquation ?
Un candidat ayant choisi l'option B avait obtenu la note 9. Quelle sera sa note après péréquation ?
Lequel des deux candidats sera le mieux classé ?

Corrigé

1. a. On résout le système $\begin{cases} a\overline{x}+b=\overline{x}' \\ |a|s \quad =s' \end{cases}$ soit

Pour l'option A : $\begin{cases} 13a+b=10 \\ 4|a| \quad =6 \end{cases}$ donc $a=\dfrac{3}{4}$ et $b=\dfrac{1}{4}$

Pour l'option B : $\begin{cases} 9a+b=10 \\ 7|a| \quad =6 \end{cases}$ donc $a=\dfrac{6}{7}$ et $b=\dfrac{16}{7}$

b. $\dfrac{3}{4}\times 12+\dfrac{1}{4}=9,25$ pour le candidat ayant choisi l'option A

et $\dfrac{6}{7}\times 9+\dfrac{16}{7}=10$ pour le candidat ayant choisi l'option B

(ce candidat avait la moyenne de l'option B, il a encore la moyenne : par péréquation, la moyenne est conservée).
Le deuxième candidat est mieux classé.

Énoncé 2

Pendant un an on regarde les productions journalières de deux chaines de production d'un même véhicule :
Pour la chaine A, on observe $\overline{x}_A = 345$ unités et $s_A = 35{,}3$.
Pour la chaine B, on observe $\overline{x}_B = 754$ unités et $s_B = 43{,}5$.
Quelle chaine présente la plus grande dispersion dans sa production journaliére ?

Corrigé

Le calcul de l'écart-type annule l'effet de taille, mais il faut tenir compte de la différence des moyennes, on ramène donc les deux séries à la même moyenne, par exemple 1.
Pour la chaine A, on divise toutes les données par $\overline{x}_A$:

$$\overline{x'_A} = 1 \quad \text{et} \quad s'_A = \frac{s_A}{\overline{x}_A} = \frac{35{,}3}{345} \simeq 0{,}1023...$$

Pour la chaine B, on divise toutes les données par $\overline{x}_B$:

$$\overline{x'_B} = 1 \quad \text{et} \quad s'_B = \frac{s_B}{\overline{x}_B} = \frac{43{,}5}{754} \simeq 0{,}0576...$$

C'est donc la chaine A qui présente la plus grande dispersion dans sa production journaliére .

Remarque

Le quotient $\frac{\text{écart-type}}{\text{moyenne}}$ est le **coefficient de variation.**

Partie B

PROBABILITES

Probabilités discrètes finies

Exercice 1 Électroménager

D'après Nathan Technique, Exos et méthodes, Terminale Bac Pro, groupements A et B, 2017

Thèmes abordés

- **Probabilités**
- **Opérations sur les événements**

Énoncé

Une étude statistique, menée auprès d'un échantillon représentatif de familles, concernant l'équipement de cuisine, a donné les résultats suivants :

a. 80% ont un four à micro-ondes
b. 30% ont un lave-vaisselle
c. 15% n'ont ni four à micro-ondes, ni lave-vaisselle.

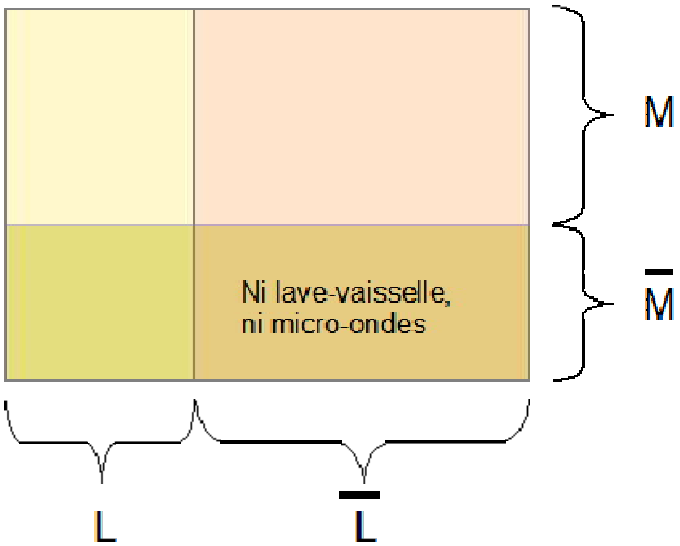
On choisit une famille de l'échantillon au hasard.

1. On note M l'événement "la famille possède un micro-ondes", L l'événement "la famille possède un lave-vaisselle".

a. Décrire par une phrase l'événement $\overline{M}$.

b. Décrire par une phrase l'événement $M \cup L$.

c. Compléter le tableau en explicitant dans chaque case l'événement.



2. Calculer $P(M \cup L)$.

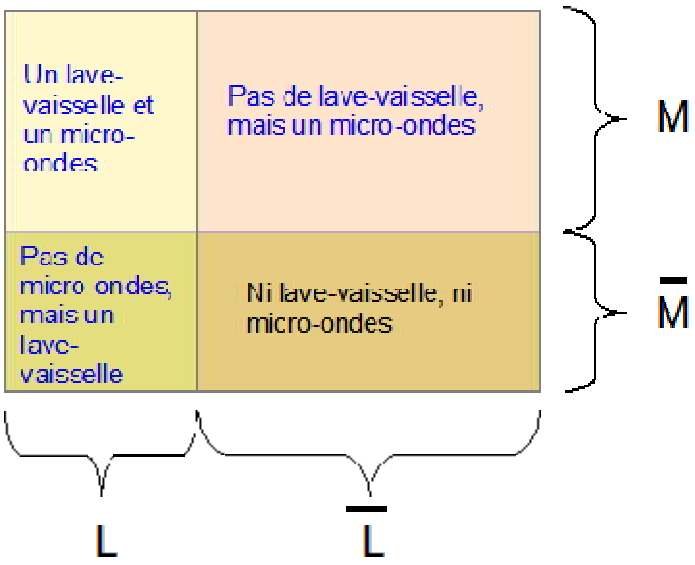
3. Calculer la probabilité que la famille ait un micro-onde et un lave-vaisselle. On pourra utiliser la formule : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Corrigé

1. a. $\overline{M}$: "la famille ne possède pas de micro-ondes"

b. $M \cup L$: "la famille possède un micro-ondes ou un lave-vaisselle".

c. Compléter le tableau en explicitant dans chaque case l'événement.



2. $P\left(\overline{M} \cap \overline{L}\right) = 0{,}15$ Or $M \cup L$ est l'événement complémentaire (ou contraire) de $\overline{M} \cap \overline{L}$ (voir graphique et c'est d'ailleurs une loi de Morgan),
par conséquent : $P(M \cup L) = 1 - P\left(\overline{M} \cap \overline{L}\right) = 1 - 0{,}15 = 0{,}85$

3. On sait donc que $P(M \cup L) = P(M) + P(L) - P(M \cap L)$.
Par conséquent : $P(M \cap L) = P(M) + P(L) - P(M \cup L)$,
ce qui donne $P(M \cap L) = 0{,}8 + 0{,}3 - 0{,}85 = 0{,}25$
La famille a une chance sur 4 d'être complètement équipée.

Exercice 2 La roue de la fortune

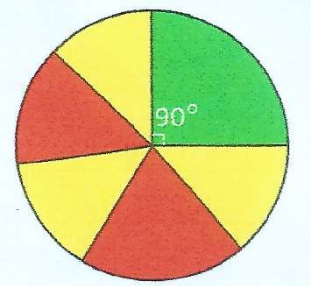
D'après Nathan Technique, Exos et méthodes, Terminale Bac Pro, groupements A et B, 2017

Thèmes abordés

- **Probabilités**
- **Espérance**

Énoncé

Dans un jeu de loterie, on utilise une roue découpée en six différents secteurs comme ci-dessous.



1. Calculer la probabilité que la roue s'arrête sur le vert.

2. La somme des mesures des angles des secteurs rouges (on la note *r*) est égale à la somme des mesures des angles des secteurs jaunes (on la note *j*).
Calculer la probabilité que la roue s'arrête sur le jaune.
En déduire la probabilité que la roue s'arrête sur le rouge.

3. Le gain est de 10 € pour le vert, de 2 € pour le rouge, et 0 € pour le jaune.
La mise est de 5 €.
« L'espérance » de gain (sans tenir compte de la mise) est obtenue par la formule

$$10 \times P(V) + 2 \times P(R) + 0 \times P(J).$$

a. Quelle est la probabilité de gagner plus que la mise ? Combien empoche-t-on alors ?

b. Quels événements conduisent à gagner moins que la mise ? Donner leur probabilité et la somme perdue à l'issue de chacun.

c. Au vu de ces résultats, trouvez-vous raisonnable de jouer à ce jeu ?

d. Calculer « l'espérance » de gain. La comparer à la mise
Cette comparaison corrobore-t-elle votre réponse à la question **c** ?

La **dernière question** est la suivante :
Si Matthieu décide de jouer 100 parties, est-il plus « probable » qu'il perde de l'argent ou qu'il en gagne ?

Corrigé

1. Les probabilités ne sont pas proportionnelles aux aires mais à la longueur des arcs, donc des angles.

On note V l'évènement «la roue s'arrête sur le vert »
On note R l'évènement «la roue s'arrête sur le rouge »
On note J l'évènement «la roue s'arrête sur le jaune »

La probabilité que la roue s'arrête sur un secteur est proportionnelle à la mesure de l'angle du secteur angulaire. Appelons b le coefficient de proportionnalité.

Le total des angles vaut 360° et la probabilité totale est 1, donc $b \times 360 = 1$, soit $b = \frac{1}{360}$

On a alors : $P(V) = 90 \times \frac{1}{360} = \frac{90}{360} = \frac{1}{4}$.

2. On a donc : (1) $r = j$
(2) $r + j + 90 = 360$

Alors $2r + 90 = 360 \Leftrightarrow 2r = 360 - 90 \Leftrightarrow r = \frac{270}{2}$

D'où $r = j = 135$

$P(J) = 135 \times \frac{1}{360} = \frac{135}{360} = \frac{3}{8}$ et $P(R) = \frac{3}{8}$.

3.a. Seul le secteur vert fait gagner plus que la mise. On empoche alors 10 € - 5 € = 5 €.

La probabilité demandée est $P(V) = \frac{1}{4}$

b. Tomber sur un secteur rouge : gain - mise = 2-5 = -3€, $P(R) = \frac{3}{8}$

Tomber sur un secteur jaune : gain - mise = 0-5 = -5€, $P(J) = \frac{3}{8}$

c. La probabilité de gagner de l'argent est $\frac{1}{4}$ et celle de perdre de l'argent est $\frac{3}{4}$

On répond par la négative.

d. L'espérance est $10 \times \frac{1}{4} + 2 \times \frac{3}{8} + 0 \times \frac{3}{8} = \frac{26}{8} = 3,25$ donc $3,25€ < 5€$

Par conséquent cela confirme qu'il n'est pas raisonnable de jouer à ce jeu.

N.B. L'espérance de gain est le gain qu'un joueur puisse espérer faire en moyenne lors d'une partie (c'est le gain moyen qu'il ferait sur un nombre infiniment grand de parties).

Dernière question

On peut dire qu'il perdra en moyenne 1,75€ par partie, s'il joue un grand nombre de parties, donc « probablement 175€ » pour 100 parties. Mais le calcul de la probabilité est difficile.

Pour aller plus loin

- S'il joue une partie, la probabilité de ne pas perdre de l'argent est $\frac{1}{4}=0,25$
- S'il joue 2 parties, la probabilité de ne pas perdre de l'argent est $P(X\geqslant 1)=0,375+0,0625=0,4375$ avec la loi $B(2;0,25)$
  (il suffit qu'il gagne au moins une partie) ou bien :

V	R	J	
2	0	0	0,0625
1	1	0	0,1875
1	0	1	0,1875
			0,4375

- S'il joue 3 parties, la probabilité de ne pas perdre de l'argent est $P(X\geqslant 2)=P(X=2)+P(X=3)=0,140625+0,015625=0,15625$
  avec la loi $B(3;0,25)$
  (il faut qu'il gagne au moins 2 parties) ou bien :

V	R	J	
3	0	0	0,015625
2	1	0	0,0703125
2	0	1	0,0703125
			0,15625

- S'il joue 4 parties, la probabilité de ne pas perdre de l'argent est $P(X\geqslant 2)=0,2109375+0,046875+0,00390625=0,26171875$
  avec la loi $B(4;0,25)$
  (il suffit qu'il gagne au moins 2 parties) ou bien :

V	R	J	
4	0	0	0,00390625
3	1	0	0,0234375
3	0	1	0,0234375
2	2	0	0,052734375
2	1	1	0,10546875
2	0	2	0,052734375
			0,26171875

- S'il joue 5 parties, la probabilité de ne pas perdre de l'argent est 0,136474609
  avec la loi multinomiale $M(n=5;p_1=0,25;p_2=0,375;p_3=0,375)$ ou bien :

V	R	J	
5	0	0	0,000976563
4	1	0	0,007324219
4	0	1	0,007324219
3	2	0	0,021972656

3	1	1	0,043945313
3	0	2	0,021972656
2	3	0	0,032958984
			0,136474609

- S'il joue 6 parties, la probabilité de ne pas perdre de l'argent est 0,169433594 avec la loi multinomiale $M\left(n=6\,;p_1=0,25;p_2=0,375;p_3=0,375\right)$ ou bien :

V	*R*	*J*	
6	0	0	0,000244141
5	1	0	0,002197266
5	0	1	0,002197266
4	2	0	0,008239746
4	1	1	0,016479492
4	0	2	0,008239746
3	3	0	0,016479492
3	2	1	0,049438477
3	1	2	0,049438477
3	0	3	0,016479492
			0,169433594

- S'il joue 7 parties, la probabilité de ne pas perdre de l'argent est 0,124629974 avec la loi multinomiale $M\left(n=7;p_1=0,25;p_2=0,375;p_3=0,375\right)$ ou bien :

V	*R*	*J*	
7	0	0	6,10352E-05
6	1	0	0,000640869
6	0	1	0,000640869
5	2	0	0,002883911
5	1	1	0,005767822
5	0	2	0,002883911
4	3	0	0,007209778
4	2	1	0,021629333
4	1	2	0,021629333
4	0	3	0,007209778
3	4	0	0,010814667
3	3	1	0,043258667
			0,124629974

Probabilités conditionnelles

Exercice 1 Liaison et agrégation

Thèmes abordés

- **Opérations sur les événements**
- **Indépendance**

Énoncé

On considère deux populations P_1 et P_2. On a ventilé les individus de chacune d'elles selon deux caractères : le genre (homme, femme) noté X, et le caractère fumeur (fumeur, non-fumeur) noté Y. On note $P = P_1 \cup P_2$.

Population P_1	Hommes	Femmes
Fumeur	10	20
Non-fumeur	30	5

Population P_2	Hommes	Femmes
Fumeur	50	10
Non-fumeur	10	15

On prélève au hasard un individu dans la population P_1, P_2 ou P.
On s'intéressera à la dépendance entre les variables X = "genre" et Y = "pratique du fumage".
On note

- F l'évènement : « l'individu est fumeur »
- H l'évènement : « l'individu est un homme »

a. Si on se restreint à la population P_1, F et H sont-ils indépendants ?
Que peut-on en déduire pour les événements $\overline{F}$ et H ? F et $\overline{H}$? $\overline{F}$ et $\overline{H}$?

b. Mêmes questions si on se restreint à la population P_2.

c. Mêmes questions pour la population totale P.

d. On aimerait comprendre le paradoxe. Pour ce faire :

a. Calculez, si l'expérience est réalisée dans P_1 les probabilités suivantes : $P_H(F)$, $P_H\left(\overline{F}\right)$ et $P(F)$. Comparez ces probabilités et interprétez.

b. Calculez, si l'expérience est réalisée dans P_2 les probabilités suivantes : $P_H(F)$, $P_H\left(\overline{F}\right)$ et $P(F)$. Comparez ces probabilités et interprétez.

c. Calculez, si l'expérience est réalisée dans P les probabilités suivantes : $P_H(F)$, $P_H\left(\overline{F}\right)$ et $P(F)$. Comparez ces probabilités et interprétez.

Corrigé

L'univers est l'ensemble des individus de la population considérée.
Une issue est un individu.
On considère que toutes les issues sont équiprobables.

1.

P_1	Homme	Femme	Totaux
Fumeur	10	20	30
Non-fumeur	30	5	35
Totaux	40	25	65

On a donc $P(F)=\frac{30}{65}=\frac{6}{13}$, $P(H)=\frac{40}{65}=\frac{8}{13}$ et $P(F\cap H)=\frac{10}{65}=\frac{2}{13}\neq P(F)\times P(H)$
Donc ces évènements sont dépendants.
Et comme $P(\overline{F})=\frac{35}{65}=\frac{7}{13}$ et $P(\overline{F}\cap H)=\frac{30}{65}=\frac{6}{13}\neq P(\overline{F})\times P(H)$
Donc $\overline{F}$ et H sont dépendants et il en est de même pour les autres couples d'événements.

Ainsi les variables X et Y ne sont pas indépendantes :
$P\big((X=a)\cap(X=b)\big)\neq P(X=a)\times P(Y=b)$

Rappel
On a le résultat suivant dont la démonstration est intéressante.
Si deux événements *A* et *B* sont indépendants, alors il en est de même pour les deux événements $\overline{A}$ et *B*.
(et il en est de même pour A et $\overline{B}$, ainsi que pour $\overline{A}$ et $\overline{B}$, puisque $\overline{\overline{A}}=A$… et de fait il y a équivalence) .
Démonstration
On sait que $P(A\cap B)=P(A)\times P(B)$
or $P(B)=P(\overline{A}\cap B)+P(A\cap B)$, d'après la formule des probabilités totales, puisque
$(\overline{A}\cap B)\cup(A\cap B)=(\overline{A}\cup A)\cap B=\Omega\cap B=B$ et
$(\overline{A}\cap B)\cap(A\cap B)=(\overline{A}\cap A)\cap B=\varnothing\cap B=\varnothing$).
ainsi $P(\overline{A}\cap B)=P(B)-P(A\cap B)=P(B)-P(A)\times P(B)$
soit $P(\overline{A}\cap B)=P(B)\big(1-P(A)\big)=P(B)\times P(\overline{A})$ d'où la conclusion.

2.

P_2	Homme	Femme	Totaux
Fumeur	50	10	60
Non-fumeur	10	15	25
Totaux	60	25	85

On a donc $P(F)=\frac{60}{85}=\frac{12}{17}$, $P(H)=\frac{60}{85}=\frac{12}{17}$ et $P(F\cap H)=\frac{50}{60}=\frac{5}{6}\neq P(F)\times P(H)$

Donc ces évènements sont dépendants et il en est de même pour les autres couples d'événements.

Ainsi les variables X et Y sont dépendantes : $P\big((X=a)\cap(X=b)\big)\neq P(X=a)\times P(Y=b)$

3.

P	Homme	Femme	Totaux
Fumeur	60	30	90
Non-fumeur	40	20	60
Totaux	100	50	150

On a $P(F)=\dfrac{90}{150}=\dfrac{3}{5}$, $P(H)=\dfrac{100}{150}=\dfrac{2}{3}$ et $P(F\cap H)=\dfrac{60}{150}=\dfrac{2}{5}=\dfrac{2}{3}\times\dfrac{3}{5}=P(F)\times P(H)$

Ces évènements sont cette fois indépendants, et il en est de même pour les autres couples d'événements.

Ainsi les variables X et Y sont indépendantes :
$P\big((X=a)\cap(X=b)\big)=P(X=a)\times P(Y=b)$

4. a. Pour P_1 :

	$P_H(F)$	$P_{\overline{H}}(F)$	
Y \ X	Homme	Femme	Prob. marginales
Fumeur	0,25	0,8	0,4615
Non-fumeur	0,75	0,2	0,5385
totaux	1	1	1

$P_H(F)=\dfrac{10}{40}=\dfrac{1}{4}$, $P_{\overline{H}}(F)=\dfrac{20}{25}=\dfrac{4}{5}$ et $P(F)=\dfrac{30}{65}=\dfrac{6}{13}$

Donc $P_H(F)\neq P_{\overline{H}}(F)$, $P_H(F)\neq P(F)$ et $P_{\overline{H}}(F)\neq P(F)$

et les événements sont dépendants.

b. Pour P_2 :

	$P_H(F)$	$P_{\overline{H}}(F)$	
Y \ X	Homme	Femme	Prob. marginales
Fumeur	0,833…	0,4	0,7059
Non-fumeur	0,166…	0,6	0,2941
totaux	1	1	1

$P_H(F)=\dfrac{50}{60}=\dfrac{5}{6}$, $P_{\overline{H}}(F)=\dfrac{10}{25}=\dfrac{2}{5}$ et $P(F)=\dfrac{60}{85}=\dfrac{12}{17}$

Donc $P_H(F)\neq P_{\overline{H}}(F)$, $P_H(F)\neq P(F)$ et $P_{\overline{H}}(F)\neq P(F)$

et les événements sont dépendants.

c. Pour P

	$P_H(F)$	$P_{\overline{H}}(F)$	
Y \ X	Homme	Femme	Prob. marginales
Fumeur	0,6	0,6	0,6
Non-fumeur	0.4	0,4	0,4
totaux	1	1	1

$P_H(F)=\frac{60}{100}=\frac{3}{5}$, $P_{\overline{H}}(F)=\frac{30}{50}=\frac{3}{5}$ et $P(F)=\frac{90}{150}=\frac{3}{5}$

Alors $P_H(F)=P_{\overline{H}}(F)=P(F)$

les événements sont cette fois indépendants.

Et on peut rappeler que :

- $P_H(F)=\frac{P(F\cap H)}{P(H)}=P(F)\Rightarrow P(F\cap H)=P(F)\times P(H)$
- $P_H(F)=P_{\overline{H}}(F)\Rightarrow \frac{P(F\cap H)}{P(H)}=\frac{P\left(F\cap\overline{H}\right)}{P\left(\overline{H}\right)}\Rightarrow \frac{P(F\cap H)}{P(H)}=\frac{P(F)-P(F\cap H)}{1-P(H)}$

$\Rightarrow P(F\cap H)\left(1-P(H)\right)=\left(P(F)-P(F\cap H)\right)P\left(\overline{H}\right)$

$\Rightarrow P(F\cap H)=P(F)\times P(H)$.

Bilan : Les probabilités conditionnelles montrent que dans P_1, un homme a moins de chances d'être fumeur que la moyenne (et une femme, plus), tandis que dans P_2, un homme a plus de chances d'être fumeur que la moyenne (et une femme, moins). Il se trouve que quand on réunit les deux populations, cela se compense et un homme, comme une femme, a exactement autant de chance d'être fumeur que la moyenne, ce qui signifie que le fait de fumer n'y dépend plus du genre.

Dans les trois cas il semble que l'expérience soit la même, ainsi que les évènements considérés. Mais les univers sont différents, et par conséquent les structures de probabilité des événements sont différentes. Ainsi les expériences aléatoires sont bien distinctes et les évènements aussi.

Exercice 2 Pâtisserie

Thèmes abordés

- **Probabilités conditionnelles, arbre pondéré.**
- **Intervalle de confiance**

Énoncé

Pour faire un gâteau, une pâtissière professionnelle a besoin d'un sachet de levure. Son nouveau fournisseur étant peu scrupuleux (ce qu'elle ignore), il y a une probabilité a inconnue de tomber sur un sachet de levure périmée (dont cet aigrefin a changé la date de péremption, le fourbe). Avec de la levure non périmée, il y a tout de même une probabilité de 0,08 que le gâteau ne lève pas, tandis qu'avec de la levure périmée, cette probabilité est de 0,75. Cela, la pâtissière (qui est un peu statisticienne) le sait.

1. Calculer, en fonction de a, la probabilité p que son gâteau, fabriqué avec un sachet de levure pris au hasard, lève.

2. Pour une grande réception, la pâtissière confectionne 100 gâteaux, et s'aperçoit que 27 n'ont pas levé. Attendant un moindre taux d'échec, elle commence à se douter que son fournisseur a triché.

a. Déterminer un intervalle de confiance à 95 % pour la probabilité p.

b. En déduire un intervalle de confiance à 95 % pour a. Conclure.

Corrigé

1. On choisit un sachet de levure au hasard et on considère les événements suivants :
A : « le sachet est un sachet de levure périmée »
B : « le gâteau confectionné avec le sachet de levure lève ».

On sait que : $P(A)=a$, $P_{\overline{A}}(\overline{B})=0,08$ et $P_A(\overline{B})=0,75$.

On cherche $p=P(B)$.

A et $\overline{A}$ forment une partition de l'univers donc, d'après la formule des probabilités totales :

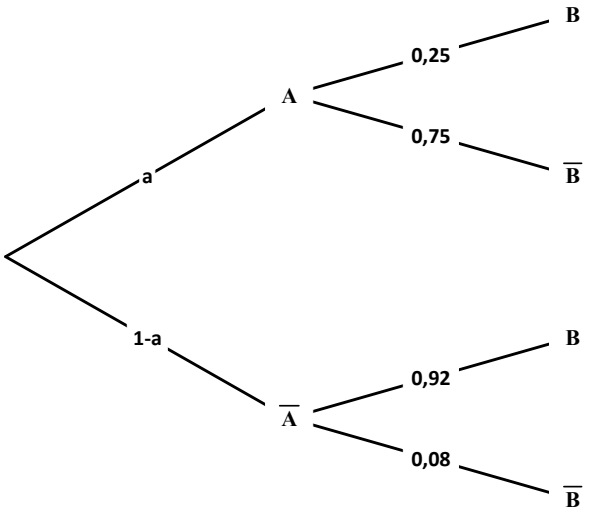
$P(B)=P(A\cap B)+P(\overline{A}\cap B)$

$P(B)=P(A)\times P_A(B)+P(\overline{A})\times P_{\overline{A}}(B)$

$P(B)=a(1-0,75)+(1-a)(1-0,08)$

$P(B)=0,92-0,67a$

Donc
$p=0,92-0,67a$



2. **a.** Dans cet échantillon de 100 gâteaux, 27 n'ont pas levé, donc 73 ont levé.

$n=100$ et $f=\dfrac{73}{100}$ (f est la fréquence observée des gâteaux qui ont levé).

Intervalle simplifié :

L'intervalle de confiance simplifié à 95 % pour la probabilité p est $\left[F_n-\frac{1}{\sqrt{n}},F_n+\frac{1}{\sqrt{n}}\right]$.

Ici la réalisation de cet intervalle est :

$I=\left[f-\frac{1}{\sqrt{n}}\,;\,f+\frac{1}{\sqrt{n}}\right]$ soit $I=\left[0,73-\frac{1}{\sqrt{100}}\,;\,0,73+\frac{1}{\sqrt{100}}\right]=[0,63\,;0,83]$

Intervalle non simplifié :

$\left[F_n-1,96\sqrt{\frac{F_n(1-F_n)}{n}},p+1,96\sqrt{\frac{F_n(1-F_n)}{n}}\right]$ est l'intervalle de confiance du paramètre p au niveau de confiance 0,95.

Ici une réalisation de cet intervalle est :

$I=\left[f-\frac{1,96}{\sqrt{n}}\sqrt{f(1-f)}\,;\,f+\frac{1,96}{\sqrt{n}}\sqrt{f(1-f)}\right]=\left[0,73-1,96\sqrt{\frac{0,73(1-0,73)}{100}}\,;\,0,73+1,96\sqrt{\frac{0,73(1-0,73)}{100}}\right]$

$I=[0,643\,;\,0,817]$.

Compléments sur les intervalles de confiance en annexe

b. On sait que $p=0,92-0,67a$ soit $a=\frac{0,92-p}{0,67}$.

Avec l'intervalle de fluctuation simplifié :

On notera F_n la variable aléatoire qui à chaque échantillon associe la fréquence des gâteaux qui ont levé.

$P\left(F_n-\frac{1}{\sqrt{n}}\leqslant p\leqslant F_n+\frac{1}{\sqrt{n}}\right)\simeq 0,95$

Puisque $x\longrightarrow\frac{0,92-x}{0,67}$ est strictement décroissante sur $\mathbb{R}$,

$F_n-\frac{1}{\sqrt{n}}\leqslant p\leqslant F_n+\frac{1}{\sqrt{n}}\;\Leftrightarrow\;\frac{0,92-\left(F_n+1/\sqrt{n}\right)}{0,67}\leqslant\frac{0,92-p}{0,67}\leqslant\frac{0,92-\left(F_n-1/\sqrt{n}\right)}{0,67}$

$\Leftrightarrow\;\frac{0,92-F_n}{0,67}-\frac{1}{0,67\sqrt{n}}\leqslant a\leqslant\frac{0,92-F_n}{0,67}+\frac{1}{0,67\sqrt{n}}$

Donc $P\left(a\in\left[\frac{0,92-F_n}{0,67}-\frac{1}{0,67\sqrt{n}}\,;\,\frac{0,92-F_n}{0,67}+\frac{1}{0,67\sqrt{n}}\right]\right)\simeq 0,95$.

L'intervalle de confiance de a au niveau 0,95 est donc :

$\left[\frac{0,92-F_n}{0,67}-\frac{1}{0,67\sqrt{n}};\frac{0,92-F_n}{0,67}+\frac{1}{0,67\sqrt{n}}\right]$

Et avec la réalisation de F_n :

$\left[\frac{0,92-0,73}{0,67}-\frac{1}{0,67\sqrt{100}};\frac{0,92-0,73}{0,67}-\frac{1}{0,67\sqrt{100}}\right]=[0,1343\,;\,0,4328]$.

Il y a donc au moins 13,43 % de sachets de levure périmée (ce qui paraît énorme !).

Exercice 3 Œufs d'or

Thèmes abordés

- **Probabilités conditionnelles**

Énoncé

Dans le jardin des Hespérides, outre les pommiers, il y a trois espèces de volailles capables de pondre des œufs d'or (mais pas à chaque œuf) : les canes (C), les poules (P) et les oies (O). Leurs proportions dans la population totale des volailles sont respectivement :
$\pi_C = 0,3$, $\pi_P = 0,5$, $\pi_O = 0,2$.
Ces volatiles n'ont pas le même pouvoir aurifère, c'est-à-dire ne pondent pas des œufs d'or avec la même fréquence : la probabilité d'avoir un œuf d'or est $p_C = 0,08$ chez les canes, $p_P = 0,04$ chez les poules et $p_O = 0,15$ chez les oies.
La nymphe Érythie trouve un œuf d'or dans son carré de laitues. Quelle est la probabilité qu'il provienne d'une cane ? D'une poule ? D'une oie ?

Corrigé

L'expérience aléatoire est : Érythie trouve un œuf dans le carré de laitues, l'absence d'œuf est un non-événement. (On ne dit rien sur Eglé et Hespérie)
L'univers est
$\Omega = \{(\text{cane, or});(\text{poule, or});(\text{oie, or});(\text{cane, non or});(\text{poule, non or});(\text{oie, non or})\}$
N.B. Il y a beaucoup d'implicites : chaque volatile, femelle en âge de pondre, anatidé ou gallinacé, peut pondre de façon équiprobable dans le carré de laitues. La fréquence des pontes est la même pour tous les volatiles. Chaque volatile pond de façon équiprobable dans le temps. Eglé et Hespérie ne perturbent pas l'expérience (ni Héraclès).
En d'autres termes : les événements « tel individu parmi ces volatiles a pondu cet œuf » sont équiprobables.

1. Avec un tableau (et/ou un tableur) :

Calculons la probabilité de chaque couple :
$P(\text{cane, or}) = P(\text{cane}) \times P_{\text{cane}}(\text{or}) = \pi_C \times p_C = 0,3 \times 0,08 = 0,024$
On calcule de même les autres probabilités.
On obtient le tableau en jaune ci-dessous.
Puis, on calcule les probabilités des différentes pondeuses sachant que l'œuf est en or.
Exemple : $P_{or}(\text{cane}) = \dfrac{P(\text{cane, or})}{P(\text{or})} = \dfrac{0,024}{0,074} = 0,324$.
Les probabilités cherchées sont dans la première ligne du tableau mauve.

Probas *or sachant pondeuse*	*cane*	*poule*	*oie*	
or	0,08	0,04	0,15	
non-or	0,92	0,96	0,85	
Probas *pondeuses*	0,3	0,5	0,2	1

Probas jointes	*cane*	*poule*	*oie*	Proba *or*
or	0,024	0,02	0,03	0,074
non-or	0,276	0,48	0,17	0,926
Probas *pondeuses*	0,3	0,5	0,2	1

Probas *pondeuse sachant or*	*cane*	*poule*	*oie*	
or	0,324	0,27	0,406	1
non-or	0,298	0,518	0,184	1

2. Avec des arbres :

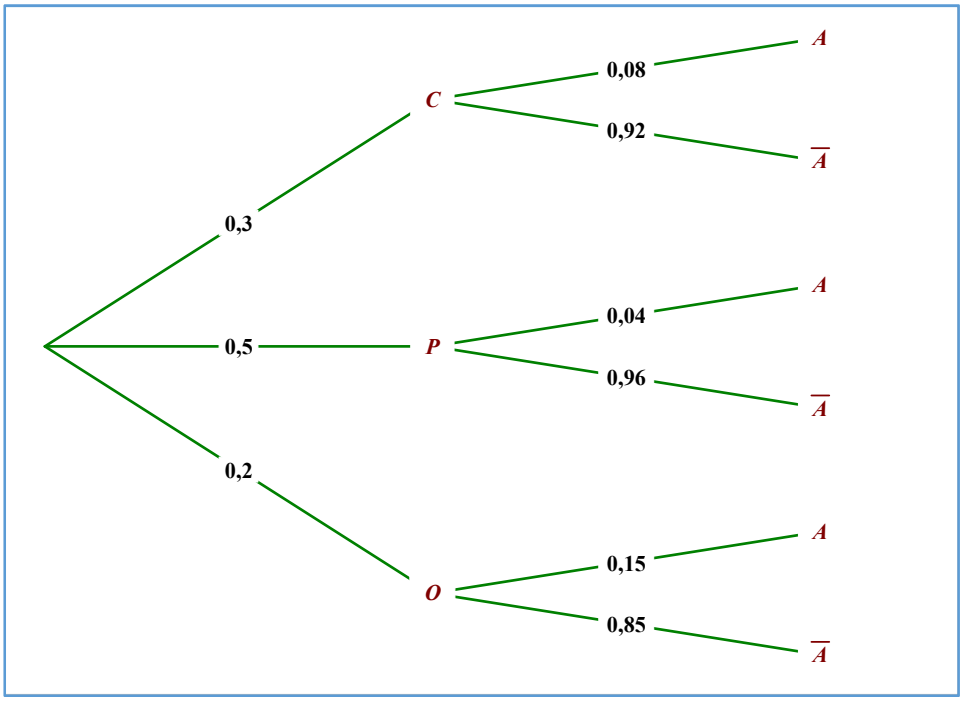
On note : C l'événement « l'œuf a été pondu par une cane »
P l'événement « l'œuf a été pondu par une poule »
O l'événement « l'œuf a été pondu par une oie »
A l'événement « l'œuf est en or »

On a donc :

$P(C)=\pi_C=0,3$, $P(P)=\pi_P=0,5$ et $P(O)=\pi_O=0,2$.

Et d'après l'énoncé : $P_C(A)=p_C=0,08$, $P_P(A)=p_P=0,04$ et $P_O(A)=p_O=0,15$.

Ce qui se traduit par l'arbre suivant :



On cherche $P_A(C)=\dfrac{P(A\cap C)}{P(A)}$.

Avec la formule des probabilités totales,

$P(A)=P(A\cap C)+P(A\cap P)+P(A\cap O)$

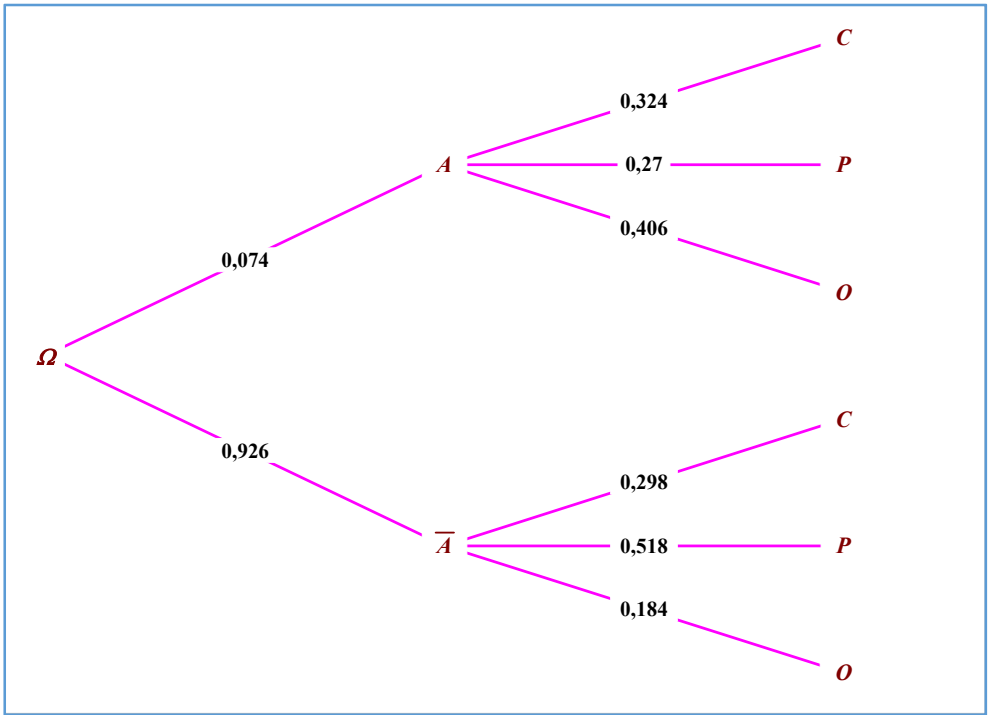
$P(A)=P_C(A)\times P(C)+P_P(A)\times P(P)+P_O(A)\times P(O)$

$P(A)=0,08\times 0,3+0,04\times 0,5+0,15\times 0,2=0,074$.

Donc $P_A(C) = \dfrac{0,3 \times 0,08}{0,074} \simeq 0,324$

Et de même $P_A(P) \simeq 0,27$, $P_A(O) \simeq 0,406$, $P_{\overline{A}}(C) \simeq \dfrac{0,3 \times 0,92}{0,926} = 0,298\ldots$

Ce qui permet de construire l'arbre suivant :



Les ramures sont différentes, mais on a les mêmes feuilles : $A \cap C$, $A \cap P$,…
Le premier arbre correspond au tableau vert, le second au tableau mauve.
Le tableau jaune correspond aux feuilles (probabilités des intersections).

Exercice 4 Un incontournable : le problème des pendus

Thème abordé

- **Probabilités conditionnelles**

Énoncé

Trois prisonniers sont condamnés à être pendus. Cependant le grand vizir a décidé de gracier l'un des prisonniers par tirage au sort. Les prisonniers n'en seront informés que le jour de l'exécution. Dans l'attente, l'un d'entre eux demande au geôlier « Je ne te demande pas si je serai pendu mais tu peux au moins me désigner l'un des deux autres prisonniers qui sera exécuté ». Le geôlier réfléchit et considère que cette information n'est d'aucune utilité pour le prisonnier et accède à sa demande. Le prisonnier lui répond alors : « Merci, tout à l'heure, j'avais une chance sur trois d'être gracié, et maintenant, j'ai une chance sur deux. ».

Qui a raison ?

N.B. Implicitement, il est interdit au geôlier de renseigner un prisonnier sur son sort.

Compte tenu des débats que suscite cet exercice, nous donnons deux formulations du corrigé.

Corrigé 1 (élémentaire et détaillé)

N.B. Comme il reste alors deux possibilités, avec un sentiment trompeur d'équiprobabilité, la tentation est grande de donner raison au prisonnier.

On identifie chaque prisonnier par une lettre : *a*, *b* et *c*, le prisonnier *a* est celui qui interroge le geôlier. Les prisonniers sont séparés, et ne peuvent communiquer.
Si on décrit ce qui se passe :
N.B. Une première expérience aléatoire, tirage au sort équiprobable, est réalisée mais les prisonniers en ignorent l'issue. Il y a trois issues, donc un univers très simple.
On note :
A est l'événement « le prisonnier *a* est gracié », et donc $\overline{A}$,« le prisonnier *a* est pendu ».
B est l'événement « le prisonnier *b* est gracié ».
C est l'événement « le prisonnier *c* est gracié ».
$\Omega = \{A; B; C\}$.

À ce moment, pour tout observateur, y compris *a* : $P(A) = P(B) = P(C) = \frac{1}{3}$.

On a, en particulier, $P_{(\overline{B} \cup \overline{C})}(A) = P_{\Omega}(A) = P(A) = \frac{1}{3}$

puisque $\overline{B}\cup\overline{C}=\{A;C\}\cup\{A;B\}=\{A;B;C\}=\Omega$.

Et, par exemple, l'information « le prisonnier *b* sera pendu » implique que la probabilité pour *a* d'être gracié, pour tout observateur, y compris *a,* est :

$P_{\overline{B}}(A)=\dfrac{P(A\cap\overline{B})}{P(\overline{B})}=\dfrac{P(A)}{P(\overline{B})}=\dfrac{1/3}{2/3}=\dfrac{1}{2}$, puisque $\overline{B}=\{A;C\}$. De même : $P_{\overline{C}}(A)=\dfrac{1}{2}$.

N.B. Du point de vue du prisonnier *a*, il y a une deuxième expérience aléatoire :
La réponse du geôlier, qui lui, sait qui est gracié. Il y a deux issues : le geôlier désigne *b*, le geôlier désigne *c.*

Avec l'enchainement de ces deux expériences, l'univers devient :

$$\Omega'=\begin{Bmatrix}(a\text{ est gracié};b\text{ est désigné});(a\text{ est gracié};c\text{ est désigné});(b\text{ est gracié};b\text{ est désigné})\\(b\text{ est gracié};c\text{ est désigné});(c\text{ est gracié};b\text{ est désigné});(c\text{ est gracié};c\text{ est désigné})\end{Bmatrix}$$

Si on note *Gb* l'événement « le geôlier désigne *b* » et *Gc* l'événement « le geôlier désigne *c* »
On a par exemple :

$$A=\{(a\text{ est gracié};b\text{ est désigné});(a\text{ est gracié};c\text{ est désigné})\}$$

$$G_b=\{(a\text{ est gracié};b\text{ est désigné});(b\text{ est gracié};b\text{ est désigné});(c\text{ est gracié};b\text{ est désigné})\}$$

$$G_c=\{(a\text{ est gracié};c\text{ est désigné});(b\text{ est gracié};c\text{ est désigné});(b\text{ est gracié};c\text{ est désigné})\}$$

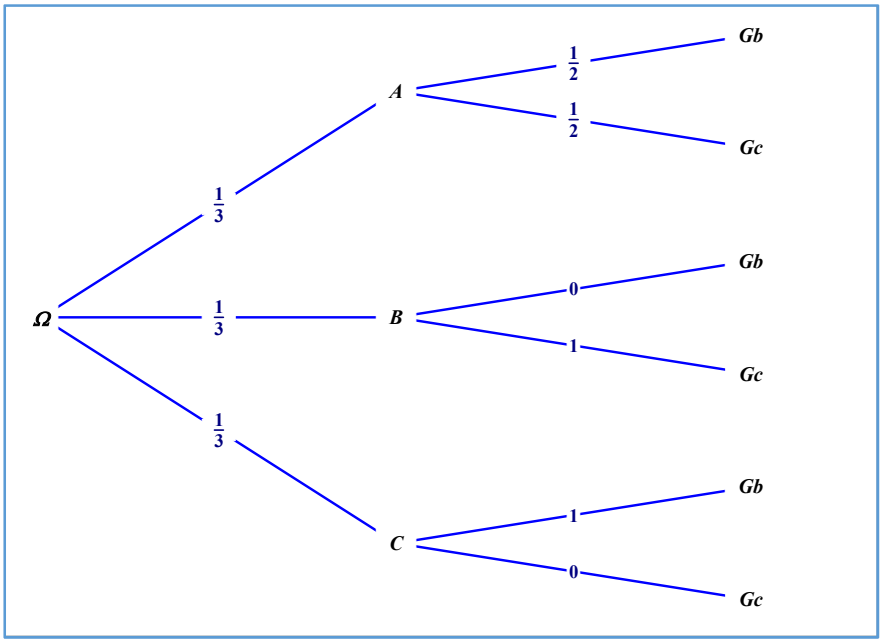
$$\overline{B}=\begin{Bmatrix}(a\text{ est gracié};b\text{ est désigné});(a\text{ est gracié};c\text{ est désigné});\\(c\text{ est gracié};b\text{ est désigné});(c\text{ est gracié};c\text{ est désigné})\end{Bmatrix}$$

Ainsi, $G_b\subset\overline{B}$ strictement, $A\cap G_b=\{(a\text{ est gracié};b\text{ est désigné})\}$,

N.B. Implicitement, on suppose comme une évidence que le geôlier désignera de façon équiprobable l'un des deux autres prisonniers quand il a le choix. Ce que ne dit pas l'énoncé !

Sous cette hypothèse, on peut alors résumer la situation avec l'arbre suivant pour connaitre la loi de probabilité :

Si on a encore $P(A)=P(B)=P(C)=\dfrac{1}{3}$, on obtient l'arbre :

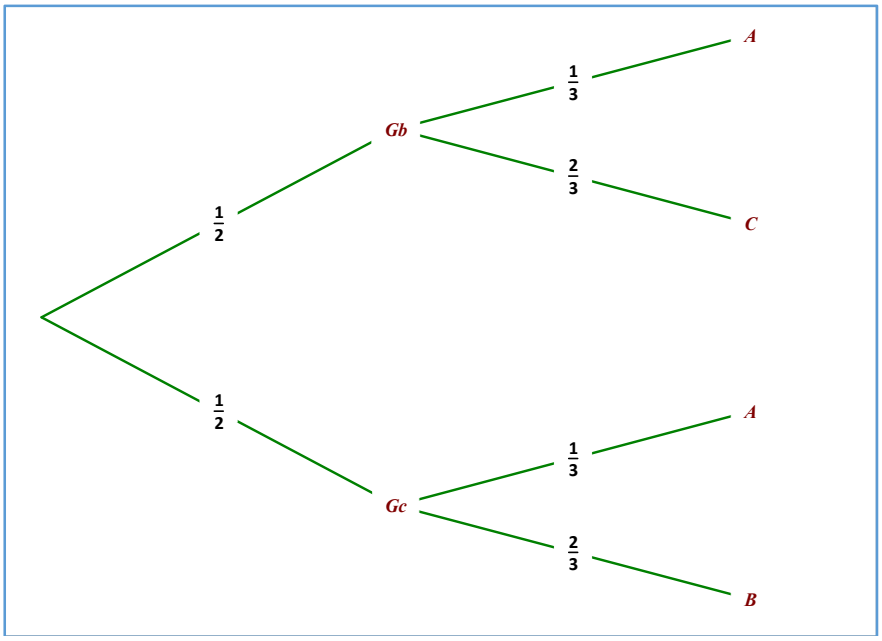


Avec la formule des probabilités totales :

$$P(G_b) = P(A \cap G_b) + P(B \cap G_b) + P(C \cap G_b) = \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 0 + \frac{1}{3} \times 1 = \frac{1}{2}$$

Et de même $P(G_c) = P(A \cap G_c) + P(B \cap G_c) + P(C \cap G_c) = \frac{1}{2}$

D'où $P_{G_b}(A) = \frac{P(A \cap G_b)}{P(G_b)} = \frac{1/3 \times 1/2}{1/2} = \frac{1}{3}$ et $P_{G_c}(A) = \frac{1}{3}$, et l'arbre :



Ainsi, quelle que soit la réponse du geôlier, la probabilité d'être pendu pour le prisonnier *a* est inchangée et ce dernier a donc tort. En revanche pour le prisonnier non désigné la probabilité d'être gracié est maintenant égale à 2/3, mais seul le prisonnier *a* le sait !

La réponse à la question posée par l'exercice est donc : Le geôlier a raison.

N.B. Cette situation est en apparence très simple mais la mise en forme de la solution n'est pas immédiate et le résultat contre-intuitif pour beaucoup.

Remarque

Bien sûr, il est tentant de faire directement les raisonnements suivants :

- On sait que l'un au moins des deux autres prisonniers sera pendu, donc que le geôlier désignera forcément l'un ou l'autre. On peut donc penser que la réponse du geôlier n'apportera à *a* aucune information supplémentaire. On sait alors que *b* ou *c* sera pendu (ou inclusif), c'est l'évènement $\overline{B} \cup \overline{C} = \Omega$.

  Ainsi, $P_{(\overline{B} \cup \overline{C})}(A) = P_{\Omega}(A) = P(A) = \frac{1}{3}$.

Mais :

- Ici on fait abstraction a priori de la seconde expérience, on fait abstraction du fait que le geôlier a désigné un condamné particulier, supposant que ça n'a pas d'incidence.
- En outre, si le geôlier a désigné un condamné et pas l'autre, c'est au terme d'un choix qui est une expérience aléatoire, ce qui n'est même pas considéré ici.
- On confond G_b et $\overline{B}$, G_c et $\overline{C}$.

- Et de plus, dans cette optique fausse, pour répondre à la question, on devrait déterminer $P_{\overline{B}}(A)$ et $P_{\overline{C}}(A)$

  On commet donc quatre erreurs, et ce raisonnement est inepte.

Ou encore de faire le raisonnement suivant :

- Le geôlier désigne nécessairement l'un des deux prisonniers, on sait alors que *b* ou *c* sera désigné (ou exclusif), c'est l'évènement $G_b \Delta G_c = \Omega'$ (encore).

Ainsi $P_{(G_b \Delta G_c)}(A) = P_{\Omega'}(A) = P(A) = \frac{1}{3}$.

Mais

- Le geôlier A RÉPONDU. Il a répondu *b*, ou il a répondu *c*. Il n'y a donc pas lieu de conditionner par $G_b \Delta G_c$ mais par G_b ou par G_c.

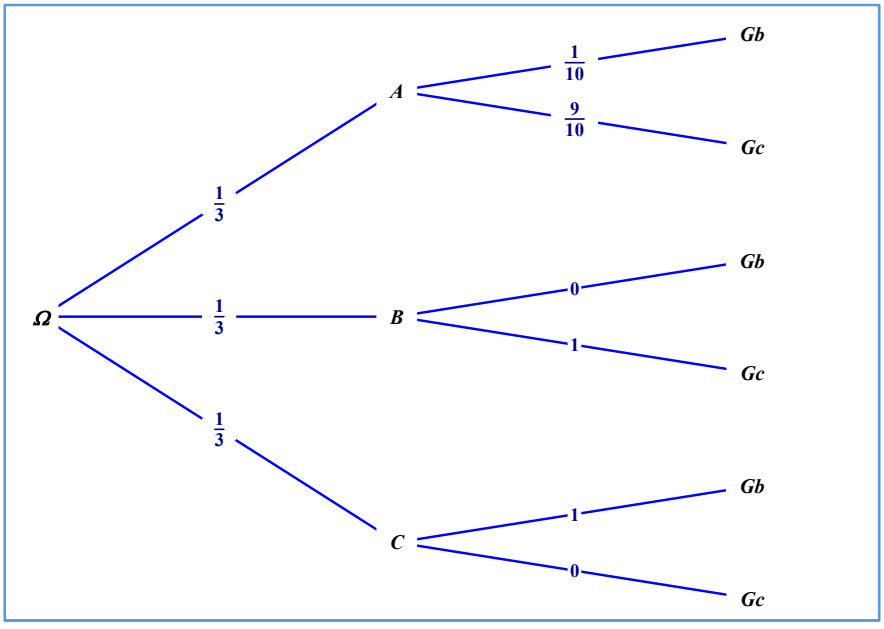
  (il faut déterminer $P_{G_b}(A)$ ou $P_{G_c}(A)$ pas $P_{(G_b \Delta G_c)}(A)$)

  Ce raisonnement est donc incorrect.

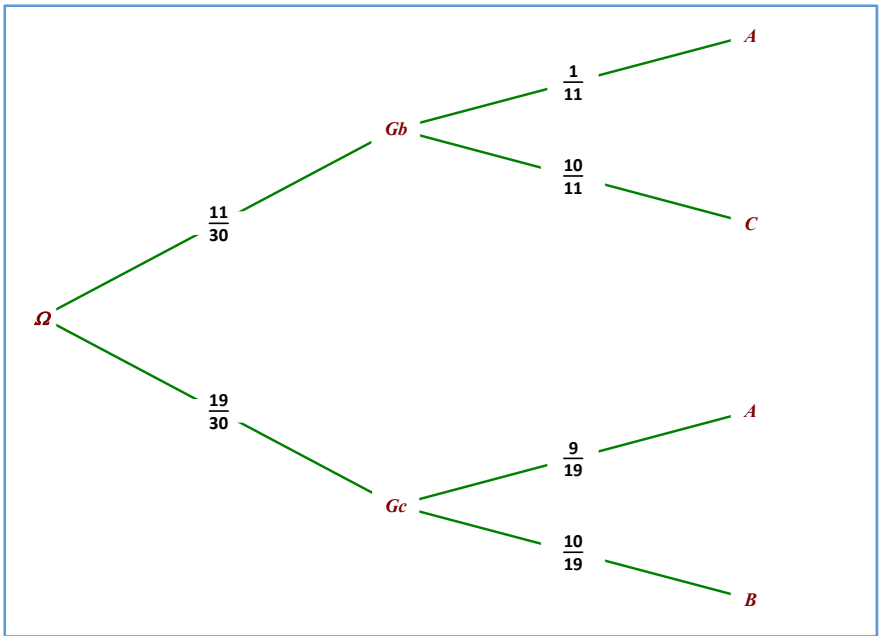
N.B. Le fait que le prisonnier *a* sache déjà que l'un au moins des deux autres sera pendu nous invite à penser (à tort) que le fait de savoir lequel ne change rien pour lui.

Ce qui peut surprendre ici est que la probabilité pour *a* d'être gracié sachant que *b* est désigné, qui est égale à la probabilité pour *a* d'être gracié sachant que *c* est désigné, est égale aussi à la probabilité pour *a* d'être gracié sachant que *b* ou *c* est désigné. Mais ceci n'est pas acquis a priori. Cela résulte du fait que l'on suppose que le geôlier n'a pas de raison de désigner *b* plutôt que *c* quand il a le choix. Si l'on modifie cela le résultat sera différent, voir ci-dessous. Par exemple, il se souvient mal du nom de *b* (on connaît des gens comme ça), donc donnera plus probablement *c,* mais *a* ne peut le savoir, ou du moins le quantifier.

Si on avait, par exemple, $P_A(G_b) = \frac{1}{10}$ et $P_A(G_c) = \frac{9}{10}$, on obtient les arbres :

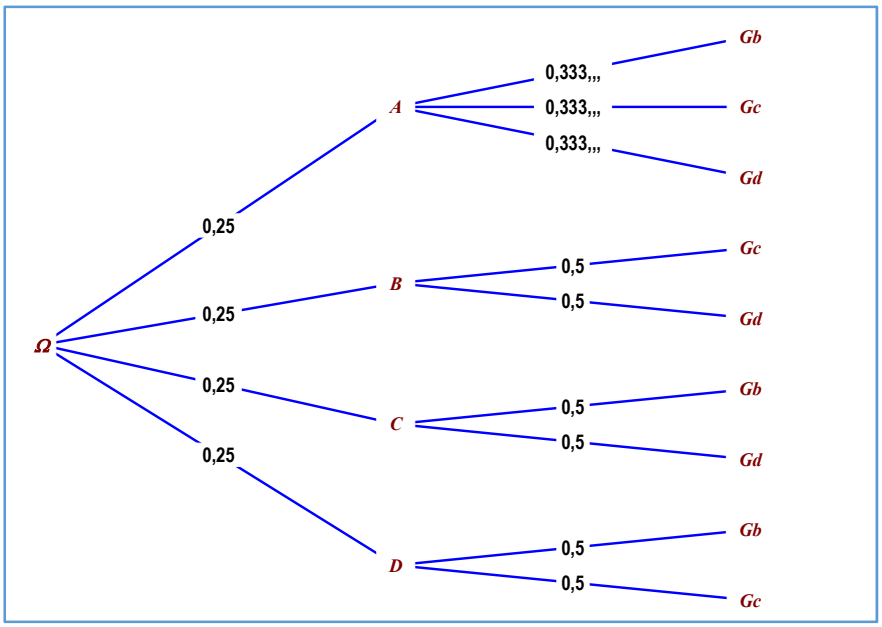


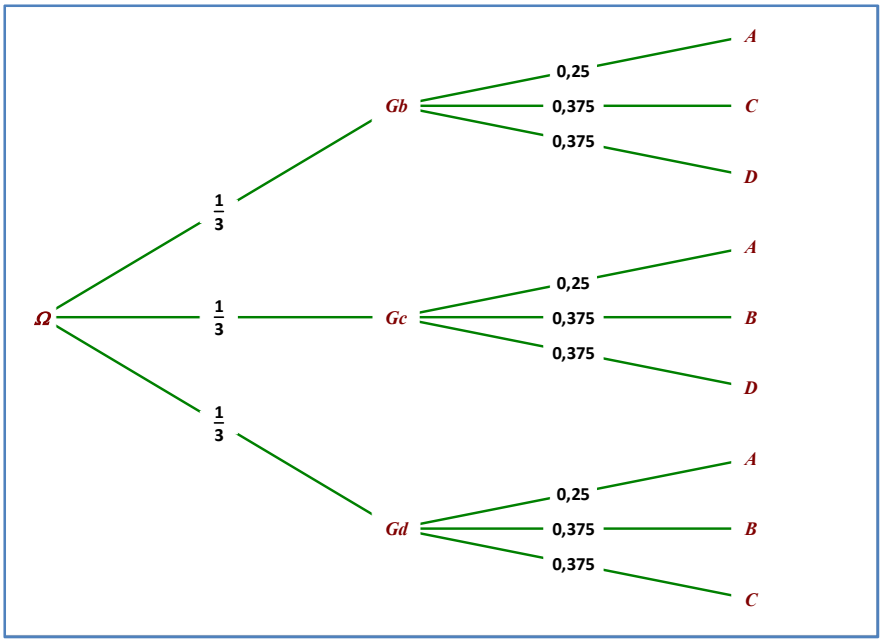
Alors :



Donc $P_{G_b}(A)=\frac{1}{11}$ et $P_{G_c}(A)=\frac{9}{19}$ et encore $P_{(G_b A G_c)}(A)=P_{\Omega'}(A)=P(A)=\frac{1}{3}$.

On peut aussi regarder ce qui se passe si on ajoute un prisonnier, en supposant encore une fois les issues, quant à la réponse du geôlier, équiprobables. On obtient :





Ainsi, on a encore : $P_{G_b}(A) = P_{G_c}(A) = P_{G_d}(A) = \frac{1}{4}$.

(Ceci devrait permettre de convaincre les indécis)

Corrigé 2 (plus subtil)

Un seul condamné est gracié (G), les autres sont exécutés (X).
L'univers est constitué de 3 événements élémentaires équiprobables (de probabilité = 1/3) :

1/3 : (a) = A gracié, B exécuté, C exécuté
1/3 : (b) = A exécuté, B gracié, C exécuté
1/3 : (c) = A exécuté, B exécuté, C gracié

"Solution apparente" A :

"G" = gracié. "X" = exécuté.

$$P_{B=X}(A=G) = \frac{P\big((A=G) et (B=X)\big)}{P(B=X)} = \frac{P(\{a\})}{P(\{a,c\})} = \frac{1/3}{2/3} = \frac{1}{2}$$

De la même façon, $P_{C=X}(A=G) = \frac{1}{2}$

Mais $P_{B=X\ ou\ C=X}(A=G) = \frac{P\big((A=G) et (B=X\ \ ou\ C=X)\big)}{P(B=X\ \ ou\ C=X)} = \frac{P(\{a\})}{P(\{a,b,c\})} = \frac{1/3}{1} = \frac{1}{3}$.

Donc, savoir que $B=X$ ou savoir que $C=X$ n'équivaut ***pas*** à savoir que ($B=X$ *ou* $C=X$) (qui est une certitude). Que le geôlier dise que B sera pendu ou qu'il dise que C sera pendu, il *élimine* une possibilité (celle que ce pendu soit gracié). La possibilité éliminée est une possibilité dans laquelle A serait pendu ! Il est donc "normal" que la probabilité a posteriori que A soit gracié augmente, *conditionnellement à cette information*. Si le geôlier disait "tu sais bien que l'un des deux sera exécuté, alors que t'importe lequel ?", il n'éliminerait aucune possibilité, donc ne changerait pas la probabilité a posteriori pour A d'être gracié.

Attention ! Ce ne sont que des probabilités a posteriori ***sachant une information donnée***. En réalité, ça ne change pas la situation : le geôlier connaît parfaitement la situation, c'est-à-dire qui sera exécuté. Pour lui (donc dans la réalité), il n'y a plus d'aléa : les probabilités valent 0 ou 1.
Il s'agit ici des probabilités que *A* soit gracié **sachant que *B* sera pendu**
ou que *A* soit gracié **sachant que *C* sera pendu.**
Mais sont-ce bien ces probabilités que *A* doit considérer lorsque le geôlier lui répond ?

Solution B :

En effet le paradoxe subsiste. Supposons que le geôlier tire au sort équiprobable l'exécuté qu'il désigne (d) parmi ceux qui seront exécutés, et que A le sache. Tout dépend de l'information que A possède et comment il s'en sert :
Dans (a), P(B désigné) = 1/2. Dans (b), P(B désigné) = 0. Dans (c), P(B désigné) = 1.
On en déduit :

$$P\left(B\,désigné\right)=P_{(a)}\left(B\,désigné\right)\times P(a)+P_{(b)}\left(B\,désigné\right)\times\ P(b)+P_{(c)}\left(B\,désigné\right)\times P(c)=\frac{1}{2}$$

$$P_{B=d}\left(A=G\right)=\frac{P\left(\left(A=G\right)et\left(B=d\right)\right)}{P\left(B=d\right)}=\frac{1/6}{1/2}=\frac{1}{3}$$

Du coup, la désignation de B n'a apporté aucune information à A sur sa probabilité d'être gracié.
Le pire est que si le geôlier, dans le cas où B et C seront exécutés, désigne B avec une probabilité *x* et C avec une probabilité $1-x$, et que A le sait, son calcul donne :

$$P_{B=d}\left(A=G\right)=\frac{P\left(\left(A=G\right)et\left(B=d\right)\right)}{P\left(B=d\right)}=\frac{x/3}{\left(1+x\right)/3}=\frac{x}{\left(1+x\right)}$$

et $P_{C=d}\left(A=G\right)=\frac{\left(1-x\right)/3}{\left(2-x\right)/3}=\frac{\left(1-x\right)}{\left(2-x\right)}$

Donc, si on utilise l'information dont A dispose sur les probabilités de désignation de l'exécuté par le geôlier, on obtient des probabilités spécifiques.
Mais il s'agit ici de la probabilité que A soit gracié **sachant que B est désigné**.
Alors, quelles probabilités A doit-il considérer ?
Ceci nous amène à réfléchir sur ce que sont les probabilités conditionnelles.
$P_{B=X}\left(A=G\right)=1/2$ signifie que dans l'ensemble de toutes les épreuves (tirage équiprobable d'un gracié parmi les 3) *telles que B est exécuté*, A est gracié une fois sur deux.
De même $P_{C=X}\left(A=G\right)=1/2$, signifie que dans l'ensemble de toutes les épreuves (tirage équiprobable d'un gracié parmi les 3) *telles que C est exécuté*, A est aussi gracié une fois sur deux.
Pourquoi A ***ne doit-il pas*** considérer la probabilité $P_{B=X}\left(A=G\right)=1/2$ à partir du moment où le geôlier lui a révélé B=X ? Tout simplement parce que si, logiquement, {le geôlier dit B=X} implique {B=X}, la ***réciproque n'est pas vraie*** : il existe une situation dans laquelle B=X et le geôlier peut dire C=X, c'est la situation (a) ci-dessus, où B=X et C=X. Cela fait que les deux événements suivants *ne sont pas équivalents* : {B=X} et {le geôlier dit que B=X}. A doit donc, au lieu de conditionner par {B=X}, conditionner par {le geôlier dit que B=X}, ce qui est la solution **B**. A doit absolument tenir compte du fait que ***ce que le geôlier dit*** est un événement ***aléatoire***.

Il existe un cas où cela revient au même, lequel ? Celui où la réciproque est vraie, autrement dit lorsque {le geôlier dit que B=X} et {B=X} sont des événements équivalents, en d'autres termes lorsque chaque fois que B=X, le geôlier dit B=X. C'est-à-dire encore que si B=X et C=X, le geôlier dira B=X avec une probabilité $x=1$. La formule donnée dans la solution B aboutit à : $P_{B=d}\left(A=G\right)=\frac{1}{\left(1+1\right)}=\frac{1}{2}$.

Variantes

- **Le jeu télévisé avec les 3 portes dit « Problème de Monty Hall »**

Qui est connu souvent des élèves, et dont la résolution est plus facilement acceptée.

N.B. Le résultat est dû au fait que le présentateur a deux possibilités quand le candidat a choisi la bonne porte, mais une seule sinon. Et en supposant que le présentateur désignera de façon équiprobable l'une des deux autres portes quand il a le choix.

On peut supposer aussi pour modifier l'exercice que le présentateur ouvrira de préférence (à quantifier) la porte la plus à gauche (ou à droite).

- **Les 3 jetons** : plus simple et moins ambigu, qui devrait susciter l'intérêt des élèves (« situation déclenchante », selon l'expression consacrée)

Dans un sac il y a 3 jetons : un qui a les deux faces blanches ;
un qui a les deux faces noires ;
un qui a une face noire et une face blanche.

On tire un jeton au hasard, on le pose sur la table, seule la face de dessus est visible.
Il faut alors parier sur la couleur de la face cachée.
À l'issue de l'expérience, la face visible est blanche ; sur quelle couleur doit-on parier ?

N.B. Le résultat est dû au fait que le jeton blanc/blanc peut montrer une face blanche de deux façons. Et il est légitime ici de supposer que les deux façons de poser le jeton sont équiprobables.

Exercice 5 Jeu des 4 cartes

Thèmes abordés

- **Probabilités conditionnelles**
- **Espérance**

Énoncé

Ce jeu est l'interprétation simplifiée du paradoxe imaginé par Robert Connelly.

Le joueur joue contre un banquier. Le jeu comporte quatre cartes dont les faces sont cachées. Deux de ces cartes sont noires et deux cartes sont rouges, indiscernables.

La règle du jeu est la suivante : le joueur doit miser une somme *m* au départ du jeu. Puis, le banquier s'apprête à retourner devant le joueur les cartes une par une. Le joueur doit annoncer « rouge ! » avant que le banquier retourne la première carte, ou la deuxième, ou la troisième ou encore la quatrième. Si la carte retournée juste après l'annonce du joueur est effectivement rouge, il double sa mise initiale ; si elle est noire, c'est le banquier qui récupère sa mise. Le but de cet exercice est de prévoir à quel moment le joueur a intérêt à faire son annonce.

1. Construire un arbre de dénombrement rassemblant toutes les issues possibles quand le banquier retourne successivement les quatre cartes.

2. Calcul de probabilités

a. Calculer la probabilité d'obtenir une carte rouge à la première carte.

b. Calculer la probabilité d'obtenir une carte rouge à la deuxième carte selon que la première carte retournée est noire ou rouge.

c. Calculer la probabilité d'obtenir une carte rouge à la troisième carte selon les issues concernant les deux premières cartes.

d. Calculer la probabilité d'obtenir une carte rouge à la quatrième carte selon les issues concernant les trois premières cartes.

3. Stratégies

a. Un joueur A décide de ne pas perdre de temps et de dire "rouge" avant même qu'on retourne la carte 1. Calculer l'espérance de son gain algébrique X (dont on a ôté la mise).

b. Un joueur B n'aimant pas l'incertitude remarque que lorsque trois cartes ont été retournées, il n'y a plus aucune incertitude. Il décide ainsi de dire "rouge" avant qu'on retourne la dernière carte. En s'aidant de l'arbre, calculer l'espérance de son gain algébrique Xet la comparer à celle de A.

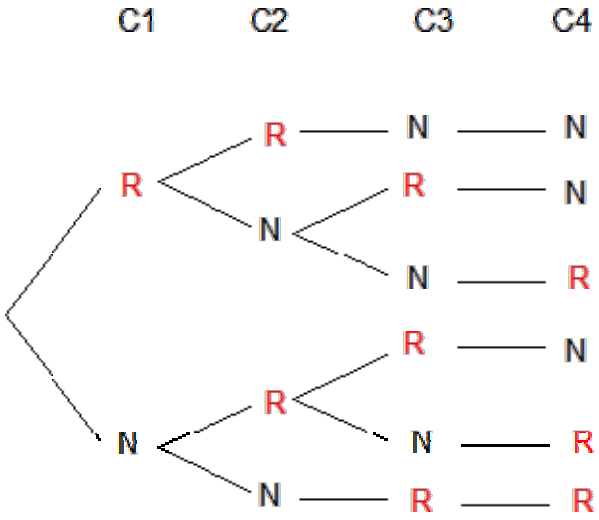
c. Un joueur C décide d'adapter à chaque instant son comportement (Attendre = A, ou Parier = P) selon les cartes retournées jusque-là. Voici sa stratégie : dès qu'une carte retournée est noire, il parie (observant que cela augmente la probabilité que la suivante soit rouge).

c1. Vérifier que la somme des probabilités associées à ces issues est égale à 1.

c2. Calculer l'espérance du gain algébrique correspondant à la stratégie de C, et la comparer à celles de A et B.

Corrigé

1.



2. a. La probabilité d'obtenir une carte rouge à la première carte est :

$$P(C1 = R) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}} = \frac{\text{nombre d'issues répondant à l'événement}}{\text{nombre d'issues de l'univers}}$$
$$= \frac{2}{4} = \frac{1}{2} = 0.5$$

b. La probabilité d'obtenir une carte rouge à la deuxième carte selon que la première carte retournée est noire ou rouge est :

$$P(C2 = R \mid C1 = R) = \frac{1}{3}; P(C2 = R \mid C1 = N) = \frac{2}{3}$$

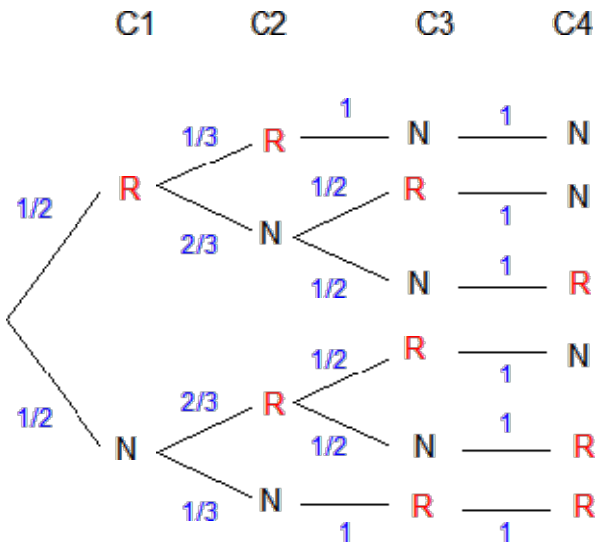
c. Les probabilités d'obtenir une carte rouge à la troisième carte selon les issues concernant les deux premières cartes sont :

$$P(C3 = R \mid C1 = R \; et \; C2 = R) = 0$$
$$P(C3 = R \mid C1 = R \; et \; C2 = N) = \frac{1}{2}$$
$$P(C3 = R \mid C1 = N \; et \; C2 = R) = \frac{1}{2}$$
$$P(C3 = R \mid C1 = N \; et \; C2 = N) = 1$$

d. Les probabilités d'obtenir une carte rouge à la quatrième carte selon les issues concernant les trois premières cartes sont :

$$P(C4 = R \mid C1 = R \; et \; C2 = R \; et \; C3 = N) = 0$$
$$P(C4 = R \mid C1 = R \; et \; C2 = N \; et \; C3 = R) = 0$$
$$P(C4 = R \mid C1 = R \; et \; C2 = N \; et \; C3 = N) = 1$$
$$P(C4 = R \mid C1 = N \; et \; C2 = R \; et \; C3 = R) = 0$$
$$P(C4 = R \mid C1 = N \; et \; C2 = R \; et \; C3 = N) = 1$$
$$P(C4 = R \mid C1 = N \; et \; C2 = N \; et \; C3 = R) = 1$$

d'où :



3. Stratégies

a. Un joueur A décide de ne pas perdre de temps et de dire "rouge" avant même qu'on retourne la carte 1. L'espérance de son gain algébrique X (dont on a ôté la mise) est :

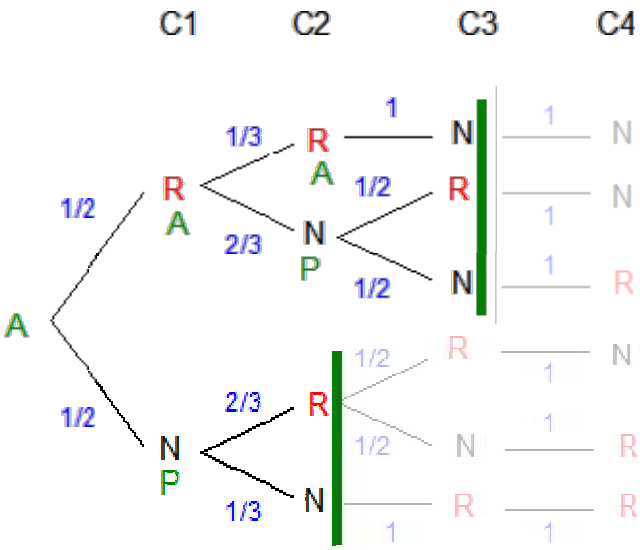
$$E(X) = \sum_i x_i p_i = (2m - m) \times \frac{1}{2} + (0 - m) \times \frac{1}{2} = 0$$

b. Un joueur B n'aimant pas l'incertitude remarque que lorsque trois cartes ont été retournées, il n'y a plus aucune incertitude. Il décide ainsi de dire "rouge" avant qu'on retourne la dernière carte. L'espérance de son gain algébrique X est :

$$E(X) = (0 - m) \times \frac{1}{6} + (0 - m) \times \frac{1}{6} + (2m - m) \times \frac{1}{6} + (0 - m) \times \frac{1}{6} + (2m - m) \times \frac{1}{6} + (2m - m) \times \frac{1}{6} = 0$$

C'est donc la même espérance.

c. Un joueur C décide d'adapter à chaque instant son comportement (Attendre = A, ou Parier = P) selon les cartes retournées jusque-là. Voici sa stratégie : dès qu'une carte retournée est noire, il parie (observant que cela augmente la probabilité que la suivante soit rouge). Cette stratégie est illustrée sur l'arbre ci-dessous, sa décision à chaque étape (A ou P) étant indiquée en vert.



c1. Le jeu s'arrête donc nécessairement aux traits verticaux verts.
$P(RRN) = \frac{1}{2} \times \frac{1}{3} \times 1 = \frac{1}{6}$
De même : $P(RNR) = \frac{1}{6}$; $P(RNN) = \frac{1}{6}$; $P(NR) = \frac{1}{3}$; $P(NN) = \frac{1}{6}$. D'où le résultat.

c2. Calculons l'espérance du gain algébrique correspondant à la stratégie de C.
RRN conduit à perdre : $X(RRN) = -m$. RNR conduit à gagner : $X(RNR) = m$.
De même : $X(RNN) = -m$, $X(NR) = m$ et $X(NN) = -m$.
D 'où :

$$E(X) = \frac{1}{6} \times -m + \frac{1}{6} \times m + \frac{1}{6} \times -m + \frac{1}{3} \times m + \frac{1}{6} \times -m = 0$$

Le joueur C a donc une stratégie qui ne l'avance à rien. C'était bien la peine !

Ces mauvaises perceptions des probabilités sont appelées des *biais cognitifs*.

Exercice 6 Jeu vidéo en ligne

D'après Baccalauréat S, Amérique du Sud, novembre 2017

Thèmes abordés

- **Graphe probabiliste**
- **Matrices, suites numériques**

Énoncé

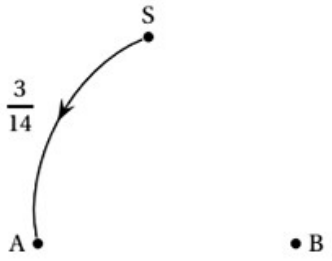
Dans un jeu vidéo en ligne, les joueurs peuvent décider de rejoindre l'équipe A (statut noté A) ou l'équipe B (statut noté B) ou bien de n'en rejoindre aucune et rester ainsi solitaire (statut noté S). Chaque jour, chaque joueur peut changer de statut mais ne peut pas se retirer du jeu.
Les données recueillies sur les premières semaines après le lancement du jeu ont permis de dégager les tendances suivantes :

- un joueur de l'équipe A y reste le jour suivant avec une probabilité de 0,6 ; il devient joueur solitaire avec une probabilité de 0,25. Sinon, il rejoint l'équipe B ;
- un joueur de l'équipe B y reste le jour suivant avec une probabilité de 0,6 ; sinon, il devient joueur solitaire avec une probabilité identique à celle de rejoindre l'équipe A ;
- un joueur solitaire garde ce statut le jour suivant avec une probabilité de $\frac{1}{7}$; il rejoint l'équipe B avec une probabilité 3 fois plus élevée que celle de rejoindre l'équipe A.

Au début du jeu, à la clôture des inscriptions, tous les joueurs sont solitaires.
On note $U_n = \begin{pmatrix} a_n & b_n & s_n \end{pmatrix}$ l'état probabiliste des statuts d'un joueur au bout de n jours. Ainsi a_n est la probabilité d'être dans l'équipe A, b_n celle d'être dans l'équipe B et s_n celle d'être un joueur solitaire, après n jours de jeu.
On a donc : $a_0 = 0$, $b_0 = 0$ et $s_0 = 1$.

1. On note p la probabilité qu'un joueur solitaire un jour donné passe dans l'équipe A le jour suivant. Justifier que $p = \frac{3}{14}$.

2. **a.** Recopier et compléter le graphe probabiliste ci-contre représentant la situation.



b. On admet que la matrice de transition est $T = \begin{pmatrix} \frac{3}{5} & \frac{3}{20} & \frac{1}{4} \\ \frac{1}{5} & \frac{3}{5} & \frac{1}{5} \\ \frac{3}{14} & \frac{9}{14} & \frac{1}{7} \end{pmatrix}$.

Pour tout entier naturel n, on a donc $U_{n+1} = U_n T$.
Montrer alors que, pour tout entier naturel n, on a $U_n = U_0 T^n$.

c. Déterminer l'état probabiliste au bout d'une semaine, en arrondissant au millième.

3. On pose $V = \begin{pmatrix} 300 & 405 & 182 \end{pmatrix}$.

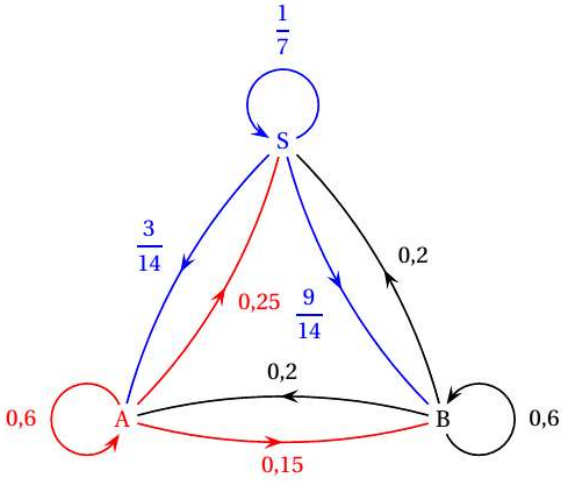
a. Donner, sans détailler les calculs, le produit matriciel VT. Que constate-t-on ?

b. En déduire un état probabiliste qui reste stable d'un jour sur l'autre.

Corrigé

1. $\frac{1}{7}+p+3p=1 \Leftrightarrow 4p=\frac{6}{7} \Leftrightarrow p=\frac{3}{14}$

2. **a.** Graphe probabiliste :



b. La matrice de transition est donnée : $T=\begin{pmatrix} \frac{3}{5} & \frac{3}{20} & \frac{1}{4} \\ \frac{1}{5} & \frac{3}{5} & \frac{1}{5} \\ \frac{3}{14} & \frac{9}{14} & \frac{1}{7} \end{pmatrix}$.

Remarque

Pour obtenir la matrice de transition à partir du graphe :

Si la relation est du type $U_{n+1}=U_nT$

↗	A	B	S
A	$\frac{3}{5}$	$\frac{3}{20}$	$\frac{1}{4}$
B	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$
S	$\frac{3}{14}$	$\frac{9}{14}$	$\frac{1}{7}$

L'ordre A B S est imposé par $U_n=\begin{pmatrix} a_n & b_n & s_n \end{pmatrix}$.

T est dite stochastique : la somme des termes de chaque ligne est égale à 1.

Si la relation est du type $U_{n+1}=MU_n$, comme dans d'autres sujets de bac :

↙	A	B	S
A	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{3}{14}$
B	$\frac{3}{20}$	$\frac{3}{5}$	$\frac{9}{14}$
S	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{7}$

avec $U_n=\begin{pmatrix} a_n \\ b_n \\ s_n \end{pmatrix}$

M est dite anti stochastique : la somme des termes de chaque colonne est égale à 1.

La matrice est la transposée de la précédente. $M=T^t$.

Si $T=\left(t_{ij}\right)_{\substack{1\leqslant i\leqslant 3\\1\leqslant j\leqslant 3}}$, le coefficient t_{ij} est la probabilité qu'un joueur rejoigne l'état *j* en partant de l'état *i*.

$U_n=\begin{pmatrix}a_n & b_n & s_n\end{pmatrix}$ est l'état probabiliste au bout de *n* jours ; et $U_0=\begin{pmatrix}0 & 0 & 1\end{pmatrix}$

Pout tout entier naturel *n*, on a $U_{n+1}=U_nT$.

On démontre par récurrence que, pour tout entier naturel *n*, $U_n=U_0T^n$.

c. Au bout d'une semaine, l'état probabiliste est :

$$U_7=U_0T^7=\begin{pmatrix}0 & 0 & 1\end{pmatrix}\begin{pmatrix}\frac{3}{5} & \frac{3}{20} & \frac{1}{4}\\ \frac{1}{5} & \frac{3}{5} & \frac{1}{5}\\ \frac{3}{14} & \frac{9}{14} & \frac{1}{7}\end{pmatrix}^7=(0{,}338 \quad 0{,}457 \quad 0{,}205)$$

(arrondis au millième)

3. a. Si $V=\begin{pmatrix}300 & 405 & 182\end{pmatrix}$ on constate que $VT=V$

b. Un état probabiliste stable est $S=\begin{pmatrix}a & b & s\end{pmatrix}$ tel que $a+b+c=1$ et $ST=S$.

Pour déterminer *S,* on peut demander aussi de résoudre le système : $\begin{cases}\frac{3}{5}x+\frac{1}{5}y+\frac{3}{14}z=x\\ \frac{3}{20}x+\frac{3}{5}y+\frac{9}{14}z=y\\ \frac{1}{4}x+\frac{1}{5}y+\frac{1}{7}z=y\\ x+y+z=1\end{cases}$

ou bien cet autre système : $\begin{cases}\frac{3}{5}x+\frac{1}{5}y+\frac{3}{14}z=x\\ \frac{3}{20}x+\frac{3}{5}y+\frac{9}{14}z=y\\ x+y+z=1\end{cases}$

On obtient ce système en utilisant, par exemple, les deux premières colonnes si les deux équations obtenues sont linéairement indépendantes. (La solution est un vecteur propre particulier).

L'ensemble des solutions du système $\begin{pmatrix}x & y & z\end{pmatrix}=\begin{pmatrix}x & y & z\end{pmatrix}T$ est le sous espace propre de la valeur propre 1. Ce sous espace est de dimension 1, donc une équation est combinaison linéaire des deux autres. On la remplace par $x+y+z=1$, qui elle ne peut être combinaison de ces deux autres puisque $x+y+z\neq 0$.

On considère donc $S=\frac{1}{300+405+182}V=\begin{pmatrix}\frac{300}{887} & \frac{405}{887} & \frac{182}{887}\end{pmatrix}$

$\frac{300}{887}+\frac{405}{887}+\frac{182}{887}=1$ et $ST=\frac{1}{887}VT=\frac{1}{887}V=S$

$S = \left(\frac{300}{887} \quad \frac{405}{887} \quad \frac{182}{887}\right) \approx (0{,}338 \;\; 0{,}457 \;\; 0{,}205)$ est un état stable.

Commentaire : L'état au bout de 7 jours est très proche de l'état stable.

En fait, $\lim\limits_{n\to\infty} U_n = S$, plus précisément, $\lim\limits_{n\to\infty} T^n = \begin{pmatrix} \frac{300}{887} & \frac{405}{887} & \frac{182}{887} \\ \frac{300}{887} & \frac{405}{887} & \frac{182}{887} \\ \frac{300}{887} & \frac{405}{887} & \frac{182}{887} \end{pmatrix}$

si bien que $\forall \; U_0 = \begin{pmatrix} a_0 & b_0 & s_0 \end{pmatrix}$ tel que $a_0 + b_0 + s_0 = 1$, $\lim\limits_{n\to\infty} U_0 T^n = \left(\frac{300}{887} \quad \frac{405}{887} \quad \frac{182}{887}\right)$

Compléments sur les matrices stochastiques en annexe

Loi géométrique

Exercice 1 Huîtres perlières

Thèmes abordés

- **Probabilités conditionnelles**
- **Loi géométrique**
- **Espérance, variance**
- **Intervalle de confiance**

Énoncé

Dans la (très grande) population d'huîtres perlières d'un parc à huîtres, il y a une proportion p d'huîtres qui contiennent une perle. On considèrera que cette population est tellement immense qu'en enlever quelques-unes ne change pas cette proportion.
Un pêcheur cherche une perle pour la vendre à un bijoutier. Il décide d'ouvrir successivement des huîtres prises au hasard jusqu'à trouver une perle. Ainsi, il est sûr de ne pas tuer trop d'huîtres.
Soit X le nombre total d'huîtres qu'il devra ouvrir jusqu'à trouver une perle.

0. Préliminaires.
a. Soient $A_1, A_2, ..., A_n$ n événements. Montrer par récurrence sur $n \geqslant 2$ que :

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) P_{A_1}(A_2) .... P_{A_1 \cap ... \cap A_{n-1}}(A_n)$$

b. Soit Y une variable aléatoire entière et A un événement. On définit l'espérance conditionnelle de Y comme : $E_A(Y) = \sum_{y \in \mathbb{N}} y P_A(Y = y)$.

Montrer que $E(Y) = P(A) E_A(Y) + P(\overline{A}) E_{\overline{A}}(Y)$.

1. Détermination de la loi de X.
a. Dessiner l'arbre de probabilités de l'expérience allant jusqu'à l'ouverture de 3 huîtres.
b. Quelle est la probabilité qu'il n'ait à ouvrir qu'une huître ?
c. Quelle est la probabilité qu'il ait à ouvrir deux huîtres ?
d. Supposons $k \geqslant 3$. Quelle est la probabilité qu'il ait à ouvrir k huîtres ?
e. Comparer $P_{X>0}(X=4)$ et $P_{X>1}(X=5)$.
f. Calculer $P_{X>k}(X=k+h)$, $P_{X>k}(X \geqslant k+h)$ et $P_{X>k}(X \leqslant k+h)$ pour $k > 0$ et $h > 0$.

2. Calcul de $\mu = E(X)$.
a. Montrer, en s'aidant de l'arbre de probabilité, que $E(X) = E_{X>1}(X-1)$.
b. Montrer que $E(X) = p \times 1 + (1-p) \times (1 + E_{X>1}(X-1))$
c. En déduire que $E(X) = \frac{1}{p}$.
d. Applications numériques :

- S'il y a une huître sur 20 qui contient une perle, combien le pêcheur devra-t-il ouvrir d'huîtres en moyenne pour trouver une perle ?
- Que devient μ lorsque $p \to 1$? Et quand $p \to 0$? Interprétation ?

3. Calcul de la variance de X.

a. Calculer $E\left(X^2\right)$. En déduire que : $V(X) = \dfrac{1-p}{p^2}$.

b. Que devient cette variance quand $p \to 1$? et quand $p \to 0$? Interpréter.

4. Intervalle de confiance pour p

Le pêcheur désire obtenir n perles. Il reproduit donc n fois indépendamment l'expérience précédente, et note X_i le résultat de la $i^{\text{ème}}$ expérience, c'est-à-dire le nombre d'huîtres supplémentaires qu'il lui a fallu ouvrir pour obtenir la $i^{\text{ème}}$ perle.

On considère la variable $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

a. Donner la loi approchée de la variable $\overline{X_n}$ (on admettra qu'une fois centrée et réduite, elle suit approximativement la loi normale $N(0;1)$).

b. En déduire un intervalle de confiance approché de niveau 0,95 pour p.

c. Application numérique : $n = 36$ et $\overline{X_n} = 11{,}8$.

Corrigé

0.a. (1) On sait que $P(A_1 \cap A_2) = P(A_1)P_{A_1}(A_2)$.

(2) Soit $n \in \mathbb{N}, n \geqslant 2$, on suppose que $P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P_{A_1}(A_2)....P_{A_1 \cap ... \cap A_{n-1}}(A_n)$.

Établissons l'égalité au rang $n+1$.

$$P(A_1 \cap A_2 \cap ... \cap A_{n+1}) = P\left((A_1 \cap A_2 \cap ... \cap A_n) \cap A_{n+1}\right) = P(A_1 \cap A_2 \cap ... \cap A_n)\, P_{(A_1 \cap A_2 \cap ... \cap A_n)}(A_{n+1})$$

Et d'après l'hypothèse de récurrence,

$$P(A_1 \cap A_2 \cap ... \cap A_{n+1}) = \left(P(A_1)P_{A_1}(A_2)....P_{A_1 \cap ... \cap A_{n-1}}(A_n)\right) P_{(A_1 \cap A_2 \cap ... \cap A_n)}(A_{n+1})$$

$$P(A_1 \cap A_2 \cap ... \cap A_{n+1}) = P(A_1)P_{A_1}(A_2)...P_{(A_1 \cap A_2 \cap ... \cap A_n)}(A_{n+1}).$$

0.b.

$$P(A)E_A(Y) + P\left(\overline{A}\right)E_{\overline{A}}(Y) = P(A)\sum_{y \in \mathbb{N}} y\, P_A(Y=y) + P\left(\overline{A}\right)\sum_{y \in \mathbb{N}} y\, P_{\overline{A}}(Y=y)$$

$$= \sum_{y \in \mathbb{N}} y\, P_A(Y=y)P(A) + \sum_{y \in \mathbb{N}} y\, P_{\overline{A}}(Y=y)P\left(\overline{A}\right)$$

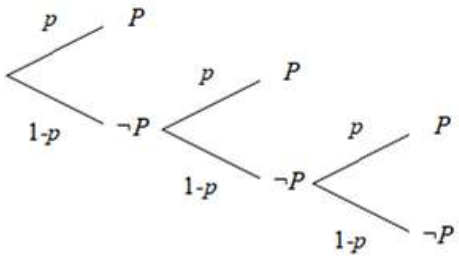
$$= \sum_{y \in \mathbb{N}} yP\left((Y=y) \cap A\right) + \sum_{y \in \mathbb{N}} yP\left((Y=y) \cap \overline{A}\right)$$

$$= \sum_{y \in \mathbb{N}} y\left[P\left((Y=y) \cap A\right) + P\left((Y=y) \cap \overline{A}\right)\right]$$

$$= \sum_{y \in \mathbb{N}} yP(Y=y) = E(y)$$

1.a.

- T_k est l'évènement « trouver une perle dans la *k*-ième huître ouverte ».
- $\neg T_k$ est l'évènement « ne pas trouver une perle dans la *k*-ième huître ouverte ».
- Les huîtres sont indépendantes (et même fermées, dit-on) et la taille de la population permet d'assimiler un tirage sans remise à un tirage avec remise. Ainsi on peut considérer que les tirages sont indépendants.
  Donc $P_{\neg T_1 \cap .. \cap \neg T_{k-1}}(T_k) = p$ et $P_{\neg T_1 \cap .. \cap \neg T_{k-1}}(\neg T_k) = 1-p \quad \forall k \in \mathbb{N}$
- L'expérience s'arrête quand une perle est trouvée.



1.b. $P(X=1) = P(\text{la première huître ouverte contient une perle}) = p$

$P(X=1) = P(T_1) = p$

1.c.

$P(X=2) = P(\text{la première huître ouverte ne contient pas de perle ET la seconde en contient une})$

$= P(\text{la première huître ouverte ne contient pas de perle}) \times P(\text{la seconde huître contient une perle sachant que la première n'en contenait pas})$

Soit :

$P(X=2) = P(\neg T_1 \cap T_2) = P(\neg T_1) \times P_{\neg T_1}(T_2) = (1-p)p.$

1.d.

$P(X=k) = P(\text{la } 1^{\text{ère}} \text{ huître ne contient pas de perle } \cap ... \cap \text{ la } (k-1)^{\text{ième}} \text{ ne contient pas de perle} \cap \text{la } k^{\text{ième}} \text{contient une perle})$

Et d'après la question **0.a.** :

$P(X=k) = P(\text{la} 1^{\text{ère}} \text{ne contient pas de perle})$

$\times P(\text{la } 2^{\text{ème}} \text{ ne contient pas de perle} \mid \text{la } 1^{\text{ère}} \text{ n'en contient pas})$

.....

$\times P(\text{la}(k-1)^{\text{ième}} \text{ ne contient pas de perle} \mid \text{les } (k-2) \text{ premières n'en contiennent pas})$

$\times P(\text{la } k^{\text{ième}} \text{ne contient pas de perle} \mid \text{les } (k-1) \text{ premières n'en contiennent pas})$

$= (1-p)^{k-1} \times p$

soit :

$P(X=k) = P(\neg T_1 \cap ... \cap \neg T_{k-1} \cap T_k)$

$= P(\neg T_1) \times P_{\neg T_1}(\neg T_2) \times ... \times P_{\neg T_1 \cap .. \cap \neg T_{k-2}}(\neg T_{k-1}) \times P_{\neg T_1 \cap .. \cap \neg T_{k-1}}(T_k)$

$P(X=k) = (1-p)^{k-1} \times p = p(1-p)^{k-1}$ pour $k \geqslant 2$.

Mais l'égalité est vraie aussi pour $k \geqslant 1$, puisque $P(X=1) = p$.

1.e. $P_{X>0}(X=4) = \dfrac{P((X>0) \cap (X=4))}{P(X>0)} = \dfrac{P(X=4)}{1} = p(1-p)^3$

car $P(X>0) = 1 - P(X=0) = 1$ car $P(X=0) = P(\varnothing) = 0$.

$$P_{X>1}(X=5)=\frac{P\big((X>1)\cap(X=5)\big)}{P(X>1)}=\frac{P(X=5)}{P(X>1)}=\frac{p(1-p)^4}{(1-p)}=p(1-p)^3$$

car $P(X>1)=1-P(X=0)-P(X=1)=1-p$

donc $P_{X>0}(X=4)=P_{X>1}(X=5)$.

1.f. $\boxed{P_{X>k}(X=k+h)}$

$$P_{X>k}(X=k+h)=\frac{P\big((X>k)\cap(X=k+h)\big)}{P(X>k)}=\frac{P(X=k+h)}{P(X>k)}=\frac{p(1-p)^{k+h-1}}{(1-p)^k}=p(1-p)^{h-1}$$

$$\begin{aligned}P(X>k)&=1-P(X=0)-P(X=1)\ldots-P(X=k)\\&=1-0-p-p(1-p)-\ldots-p(1-p)^{k-1}=1-p\left(1+\ldots+(1-p)^{k-1}\right)\\&=1-p\frac{1-(1-p)^k}{1-(1-p)}=1-1+(1-p)^k=(1-p)^k\end{aligned}$$

Donc $P_{X>k}(X=k+h)=P(X=h)$ **.**

$P_{X>k}(X=k+h)$ ne dépend pas de *k,* seulement de *h* .

$\boxed{P_{X>k}(X\geqslant k+h)}$

$$P_{X>k}(X\geqslant k+h)=\frac{P\big((X>k)\cap(X\geqslant k+h)\big)}{P(X>k)}=\frac{P(X\geqslant k+h)}{P(X>k)}=\frac{(1-p)^{k+h-1}}{(1-p)^k}=(1-p)^{h-1}$$

Le nombre d'ouvertures qu'il lui reste à faire ne dépend pas du nombre d'ouvertures infructueuses déjà faites.

Donc $P_{X>k}(X\geqslant k+h)=P(X\geqslant h)$

$P_{X>k}(X\geqslant k+h)$ ne dépend pas de *k,* seulement de *h*.

$\boxed{P_{X>k}(X\leqslant k+h)}$

Comme $\begin{aligned}P(X\leqslant k)&=P(X=0)+P(X=1)+\ldots+P(X=k)\\&=0+p+p(1-p)+\ldots+p(1-p)^{k-2}=p\left(1+\ldots+(1-p)^{k-1}\right)\\&=p\frac{1-(1-p)^k}{1-(1-p)}=1-(1-p)^k\end{aligned}$

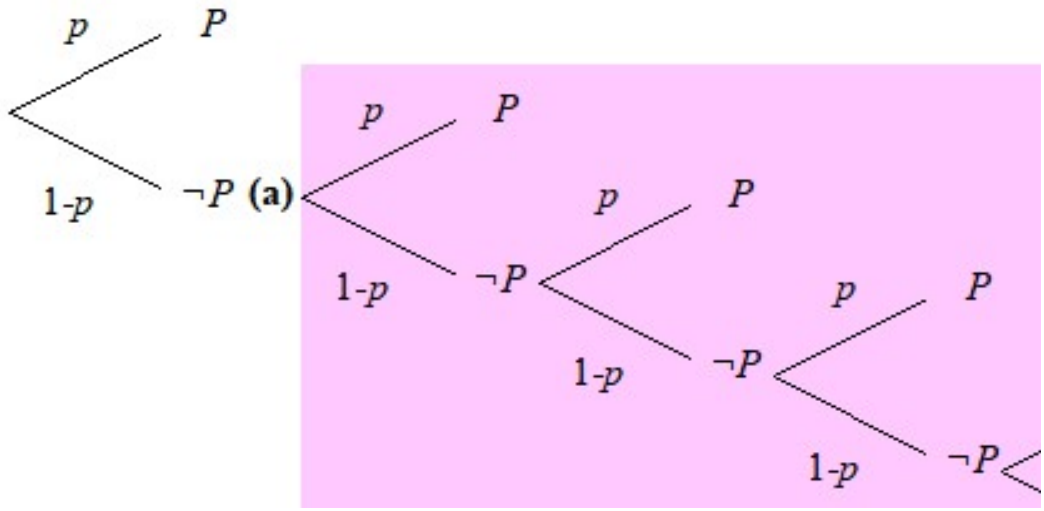
$$\begin{aligned}P_{X>k}(X\leqslant k+h)&=\frac{P\big((X>k)\cap(X\leqslant k+h)\big)}{P(X>k)}=\frac{P(k<X\leqslant k+h)}{P(X>k)}\\&=\frac{P(X\leqslant k+h)-P(X>k)}{P(X>k)}\\&=\frac{\left(1-(1-p)^{k+h}\right)-\left(1-(1-p)^k\right)}{(1-p)^k}\\&=\frac{(1-p)^k-(1-p)^{k+h}}{(1-p)^k}=1-\frac{(1-p)^{k+h}}{(1-p)^k}=1-(1-p)^h\end{aligned}$$

$P_{X>k}(X \leqslant k+h) = P(X \leqslant h)$

$P_{X>k}(X \leqslant k+h)$ ne dépend pas de *k,* seulement de *h*.

Discrète, la loi de *X* est aussi sans mémoire.

2.a. $\{X>1\}$ signifie exactement que la première huître ouverte ne contenait pas de perle.



Une fois la première huître ouverte, on considère donc l'arbre à partir du point **(a).**

Or, le sous-arbre qu'on obtient à partir de **(a)** est exactement le même que l'arbre total.

L'espérance du nombre de tirages d'huître *qui restent à partir du premier* est donc égale à μ .

Soit $E_{X>1}(X-1) = E(X)$.

(Formellement : $E_{X>1}(X) = \sum_{h\in\mathbb{N}} h P_{X>1}(X=h+1) = \sum_{h\in\mathbb{N}} h P(X=h) = E(X)$)

2.b. D'après la question **0.b.** :

Avec $A = \{X=1\}$ et donc $\overline{A} = \{X \neq 1\} = \{X>1\}$

$E(X) = P(X=1)E_{X=1}(X) + P(X>1)E_{X>1}(X)$

Or $E_{X=1}(X) = \sum_{x\in\mathbb{N}} x P_{X=1}(X=x) = 0\times0+1\times1+0\times0+\ldots = 1$

$E(X) = P(X=1)\times1 + P(X>1)E_{X>1}((X-1)+1)$

$E(X) = p\times1 + P(X>1)(1+E_{X>1}(X-1))$ (car $E(X+a) = a+E(X)$)

$E(X) = p\times1 + (1-p)(E_{X>1}(X-1)+1)$

2.c.

Comme $E(X) = E_{X>1}(X-1)$, on a : $E(X) = p\times1 + (1-p)(E(X)+1)$

Soit $\mu = p + (1-p)(\mu+1)$

$\mu = p + \mu + 1 - p\mu - p$

$p\mu = 1$

$\mu = \frac{1}{p}$.

On peut aussi calculer directement cette espérance :

Théorème :
La somme d'une série entière de rayon de convergence $R \neq 0$ *est indéfiniment dérivable sur l'intervalle* $]-R;R[$ *et la dérivée d'ordre p s'obtient en dérivant p fois terme à terme.*

Toutes les dérivées ont le même intervalle de convergence : $]-R;R[$.

Ce théorème permet de valider le calcul suivant :

ici, $R=1$

$$\sum_{k\in\mathbb{N}} k\,P(X=k) = \sum_{k\in\mathbb{N}} k\,p(1-p)^{k-1} = p\sum_{k\in\mathbb{N}} k(1-p)^{k-1}$$

$$= p\sum_{k\in\mathbb{N}} \frac{-d(1-p)^k}{dp} = p\frac{-d\left(\sum_{k\in\mathbb{N}}(1-p)^k\right)}{dp} = p\frac{-d\left(\frac{1}{1-(1-p)}\right)}{dp} = p\frac{-d\left(\frac{1}{p}\right)}{dp} = p\times\frac{1}{p^2}$$

d. Applications numériques

- S'il y a une huître sur 20 qui contient une perle, combien le pêcheur devra-t-il ouvrir d'huîtres en moyenne pour trouver une perle ?

$$p=\frac{1}{20} \Rightarrow \mu = 20$$

Le pêcheur devra ouvrir en moyenne 20 huîtres pour trouver une perle.

- Que devient μ lorsque $p\to 1$? Et quand $p\to 0$? Interprétation ?

$p\to 1 \Rightarrow \mu\to 1$: Lorsque toutes les huîtres contiennent une perle, le pêcheur n'a besoin d'en ouvrir qu'une pour trouver une perle.

$p\to 0 \Rightarrow \mu\to+\infty$: Lorsqu'il y a une proportion infime des huîtres qui contiennent une perle, le pêcheur doit en ouvrir un nombre immense pour trouver une perle.

3.a. $V(X)=E\left(X^2\right)-\left(E(X)\right)^2$ (a)

Or, d'après **0.b**, $E(Y)=P(A)E_A(Y)+P\left(\overline{A}\right)E_{\overline{A}}(Y)$.

En prenant $Y=X^2$ et $A=\{X=1\}$:

$$E\left(X^2\right)=P(X=1)E_{X=1}\left(X^2\right)+P(X>1)E_{X>1}\left(X^2\right)$$

$$E\left(X^2\right)=pE_{X=1}\left(X^2\right)+(1-p)E_{X>1}\left(X^2\right) \quad \text{(b)}$$

$$E_{X>1}\left(X^2\right)=E_{X>1}\left((1+X-1)^2\right)=E_{X>1}\left(1+2(X-1)+(X-1)^2\right)$$

$$E_{X>1}\left(X^2\right)=1+2E_{X>1}\left((X-1)\right)+E_{X>1}\left((X-1)^2\right)$$

Comme $E_{X>1}(X-1)=E(X)$ et pour la même raison, $E_{X>1}\left((X-1)^2\right)=E\left(X^2\right)$,

$$E\left(X^2\right)=1+2E(X)+E\left(X^2\right)$$

D'où, en notant : $\lambda = E\left(X^2\right)$

(b) $\Rightarrow \lambda = p+(1-p)(1+2\mu+\lambda)$

$$\Rightarrow \lambda = p+(1-p)\left(1+\frac{2}{p}+\lambda\right)$$

$\Rightarrow \lambda = \frac{2-p}{p^2}$

(a) $\Rightarrow V(X) = \lambda - \mu^2 = \frac{1-p}{p^2}$

On peut alternativement, avec le même théorème que pour l'espérance, calculer ainsi la variance :

$$E(X^2) = \sum_{k\in\mathbb{N}} k^2 P(X=k) = \sum_{k\in\mathbb{N}} k^2 p(1-p)^{k-1} = p\sum_{k\in\mathbb{N}} k^2 (1-p)^{k-1}$$

$$= p\sum_{k\in\mathbb{N}} k(k+1)(1-p)^{k-1} - p\sum_{k\in\mathbb{N}} k(1-p)^{k-1}$$

$$= p\sum_{k\in\mathbb{N}} \frac{d^2(1-p)^{k+1}}{dp^2} - \frac{1}{p} = p\frac{d^2}{dp^2}\left(\sum_{k\in\mathbb{N}}(1-p)^{k+1}\right) - \frac{1}{p} = p\times\frac{1}{p^2}$$

$$= p\frac{d^2}{dp^2}\left((1-p)\frac{1}{p}\right) - \frac{1}{p} = p\times\left(\frac{2}{p^3}\right) - \frac{1}{p} = \frac{2}{p^2} - \frac{1}{p} = \frac{2-p}{p^2}$$

Et $V(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$

3.b.

$V(X) \underset{p\to 1}{\to} 0$.

On retrouve bien le fait que lorsque toutes les huîtres contiennent une perle, le pêcheur n'a besoin d'en ouvrir qu'une pour trouver une perle, et donc que X est constante égale à 1, de variance nulle.

$V(X) \underset{p\to 0}{\to} +\infty$.

Donc, lorsque la proportion des huîtres qui contiennent une perle est infime, non seulement le pêcheur doit en ouvrir un nombre immense pour trouver une perle, mais en plus ce nombre varie immensément.

4.a.

$$E(\overline{X_n}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X) = \frac{n}{n}E(X) = E(X)$$

$E(\overline{X_n}) = \frac{1}{p}$

$V(\overline{X_n}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$, car les X_i sont indépendants.

$V(\overline{X_n}) = \frac{n}{n^2}V(X) = \frac{1}{n}V(X) = \frac{1-p}{np^2}$

La variable $Z = \frac{\overline{X_n} - E(\overline{X_n})}{\sqrt{V(\overline{X_n})}} = \frac{\overline{X_n} - \frac{1}{p}}{\sqrt{\frac{1-p}{np^2}}}$ suit approximativement la loi $N(0;1)$.

Donc $\overline{X_n}$ suit approximativement la loi $N\left(\mu = \frac{1}{p}; \sigma = \sqrt{\frac{1-p}{np^2}}\right)$.

b. On cherche tout d'abord un intervalle $[\alpha\,;\beta]$ tel que $P\left(Z \in [\alpha\,;\beta]\right) \simeq 0{,}95$:

$$Z = \frac{\overline{X_n} - \frac{1}{p}}{\sqrt{\frac{1-p}{np^2}}} = \sqrt{n} \times \frac{p\overline{X_n} - 1}{\sqrt{1-p}}.$$

On sait que si Z suit la loi $N(0;1)$, $P\left(-1{,}96 \leqslant Z \leqslant 1{,}96\right) \simeq 0{,}95$

$$P\left(-1{,}96 \leqslant \sqrt{n}\,\frac{p\overline{X_n} - 1}{\sqrt{1-p}} \leqslant 1{,}96\right) = P\left(\left(\sqrt{n}\,\frac{p\overline{X_n} - 1}{\sqrt{1-p}}\right)^2 \leqslant 3{,}84\right) \simeq 0{,}95$$

$$P\left(\frac{n}{1-p}\left(p^2\overline{X_n}^2 - 2p\overline{X_n} + 1\right) \leqslant 3{,}84\right) \simeq 0{,}95$$

$$P\left(n\left(p^2\overline{X_n}^2 - 2p\overline{X_n} + 1\right) \leqslant 3{,}84\left(1-p\right)\right) \simeq 0{,}95 \text{ car } \left(1-p\right) > 0$$

$$P\left(\left(n\overline{X_n}^2\right)p^2 + \left(3{,}84 - 2n\overline{X_n}\right)p + \left(n - 3{,}84\right) \leqslant 0\right) \simeq 0{,}95$$

Nous avons donc un trinôme en p, dont les deux premiers coefficients sont aléatoires et dont le signe est négatif lorsque p appartient à l'intervalle $[p_1\,;p_2]$ de ses racines (qui sont elles-mêmes aléatoires).

$$\Delta = \left(3{,}84 - 2n\overline{X_n}\right)^2 - 4n\overline{X_n}^2\left(n - 3{,}84\right) = 3{,}84^2 + 15{,}36\,n\,\overline{X_n}\left(\overline{X_n} - 1\right)$$

Il est aisé de vérifier que le discriminant est strictement positif :

$$X_i \geqslant 1 \quad \forall i \in \mathbb{N}^* \Rightarrow \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant 1 \;\; \forall n \in \mathbb{N}^* \Rightarrow \Delta > 0 \quad \forall n \in \mathbb{N}^*$$

Les deux racines sont : $p_1 = \frac{-\left(3{,}84 - 2n\overline{X_n}\right) - \sqrt{\Delta}}{2n\overline{X_n}^2}$ et $p_2 = \frac{-\left(3{,}84 - 2n\overline{X_n}\right) + \sqrt{\Delta}}{2n\overline{X_n}^2}$

Reste à établir les conditions pour que les deux racines soient positives.

Rappel : un trinôme $ax^2 + bx + c$ ayant deux racines x_1 et x_2 peut se factoriser :

$$ax^2 + bx + c = a(x - x_1)(x - x_2) = ax^2 - a(x_1 + x_2)x + ax_1x_2$$

Et par identification des coefficients, on a donc :

la somme des racines du trinôme : $x_1 + x_2 = -\dfrac{b}{a}$;

le produit des racines du trinôme : $x_1x_2 = \dfrac{c}{a}$.

ici $p_1p_2 = \dfrac{(n - 3{,}84)}{n\overline{X_n}^2}$ donc les racines sont de même signe si $n \geqslant 4$

$p_1 + p_2 = \dfrac{3{,}84 - 2n\overline{X_n}}{n\overline{X_n}^2}$, comme $\overline{X_n} \geqslant 1 \ \forall n \in \mathbb{N}^*$, la somme est positive si $n \geqslant 2$

En définitive, dès que $n \geqslant 4$, la somme des racines et leur produit étant positifs, les deux racines sont positives.

La plus grande des deux racines est-elle inférieure à 1 ?

$$p_2 = \frac{-\left(3{,}84 - 2n\overline{X_n}\right) + \sqrt{\Delta}}{2n\overline{X_n}^2} < \frac{2n\overline{X_n} + \sqrt{\Delta}}{2n\overline{X_n}^2}$$

$\Delta = 3{,}84^2 + 15{,}36\, n\overline{X_n}^2 - 15{,}36n\,\overline{X_n}$

$\overline{X_n} \geqslant 1 \ \forall n \in \mathbb{N}^* \Rightarrow 3{,}84^2 - 15{,}36n\,\overline{X_n} < 3{,}84^2 - 15{,}36n$

$n \geqslant 4$ donc $3{,}84^2 - 15{,}36\, n < 0$ donc $\Delta < 15{,}36\, n\,\overline{X_n}^2$

Dès lors : $p_2 = \dfrac{-\left(3{,}84 - 2n\overline{X_n}\right) + \sqrt{\Delta}}{2n\overline{X_n}^2} < \dfrac{2n\overline{X_n} + \sqrt{15{,}36\, n\overline{X_n}^2}}{2n\overline{X_n}^2} = \dfrac{1 + \dfrac{1{,}96}{\sqrt{n}}}{\overline{X_n}}$.

Cette quantité n'est pas toujours inférieure à 1, mais elle le sera si $\overline{X_n}$ n'est pas trop proche de 1 et que *n* est assez grand (ce qui doit de toutes façons être le cas pour que l'approximation normale soit licite). De plus la majoration est assez brutale…

On a donc $P\left(p \in \left[p_1\,;\,p_2\right]\right) \simeq 0{,}95$

$\left[p_1\,;\,p_2\right]$ est un intervalle de confiance de *p* au niveau approximatif 0,95.

c. Application numérique :

$n = 36$ et $\overline{X_{36}} = 11{,}8$ donne pour réalisation de l'intervalle de confiance pour *p* : [0.059 ; 0.112].

Si on se contentait de la fréquence calculée sur l'échantillon, on dirait qu'en moyenne, le pêcheur doit ouvrir environ 12 huîtres pour avoir une perle, et qu'il y a une huître sur 11,8 qui contient une perle, soit 8,47 % des huîtres. Mais on doit tenir compte de la fluctuation d'échantillonnage, ce qui est la raison d'être de l'intervalle de confiance. Ce dernier donne, avec un niveau de confiance de 95 %, une proportion d'huîtres perlières comprise entre 5,9 % et 11,2 %. Plus de confiance implique moins de précision !

$\bar{X}$	11,8
n	36
$-b$	853,4415
$2a$	10025,28
Δ	70510,75
$\sqrt{\Delta}$	265,54
p_1	0,0586
p_2	0,1116

Loi uniforme

Exercice 1 Temps d'attente (1)

Thèmes abordés

- **Loi uniforme**
- **Événements disjoints**
- **Densité**

Énoncé

Énoncé original du document « ressources pour la classe terminale générale et technologique »

1. A partir de 7 heures le matin, les bus passent toutes les quinze minutes à un arrê
se présente à cet arrêt entre 7h et 7h30. On fait l'hypothèse que l'heure exacte d
arrêt, représentée par le nombre de minutes après 7h, est la variable aléatoire unif
sur l'intervalle [0, 30].

1) Quelle est la probabilité que l'usager attende moins de cinq minutes le procl

N.B. Il faut commencer par « modéliser » la situation, c'est-à-dire définir la variable X qui suit la loi uniforme indiquée et la variable T correspondant au temps d'attente :
X suit la loi $U([0,30])$ et la variable T prend ses valeurs dans l'intervalle [0 ; 15 [

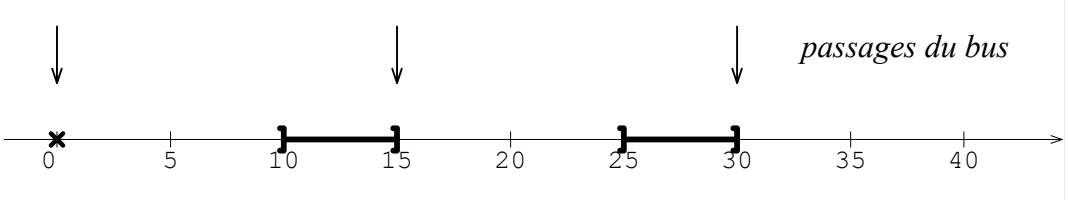
Reformulation possible de l'énoncé

N.B. Les bus passent toutes les 15 minutes à un arrêt précis dès 7h du matin : 7h00 ; 7h15 ; …
Un usager se présente à cet arrêt entre 7h et 7h30. On fait l'hypothèse que l'heure exacte de son arrivée à cet arrêt, représentée par le nombre de minutes après 7h00, est une variable aléatoire X uniformément répartie sur l'intervalle [0 ; 30]. Le temps d'attente par l'usager d'un bus est une variable aléatoire notée T.

1. Quelle est la probabilité que l'usager attende moins de cinq minutes le prochain bus ?
2. Quelle est la probabilité qu'il attende plus de dix minutes ?

Corrigé

1. Il faut déterminer $P(\{T < 5\})$



$T \in [0,5[\Leftrightarrow X \in \{0\}$ ou $X \in]10,15]$ ou $X \in]25,30]$

$T \in [0,5[\Leftrightarrow X \in \{0\} \cup [0,5] \cup]25,30]$

L'événement $\{T < 5\}$ est équivalent à $\{X = 0\} \cup \{X \in]10 ; 15]\} \cup \{X \in]25 ; 30]$

N.B. est une simplification de l'écriture $P\left(\{T<5\}\right)$

On identifie « $T<5$ » à l'événement $\{T<5\}$ qui est sous ensemble de l'univers.

Comme les événements sont disjoints deux à deux :
P ({T < 5 }) = P ({ X = 0 }) + P ({ $X \in$] 10 ; 15] }) + P ({ $X \in$] 25 ; 30] })
où X suit la loi $U\left([0,30]\right)$

$$P\left(\{T<5\}\right)=\int_0^0 \frac{1}{30}dt+\int_{10}^{15}\frac{1}{30}dt+\int_{25}^{30}\frac{1}{30}dt=0+\frac{5}{30}+\frac{5}{30}=\frac{1}{3}$$

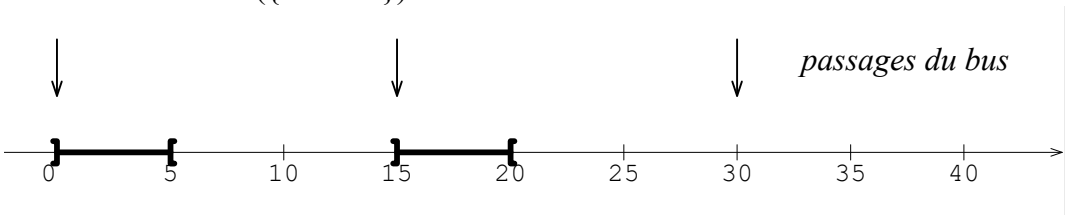
N.B. Il faut noter que $P\left(\{T<5\}\right)=P\left(\{T\leqslant 5\}\right)$

Pour les variables aléatoires réelles donc continues, changer une inégalité stricte à une inégalité large ne change pas la probabilité par définition de l'intégrale :

$$\int_a^{\lim\limits_{t\to b}} f(x)dx=\int_a^b f(x)dx$$

La probabilité P ({ X = 0 }) est nulle, mais l'événement { X = 0 } n'est pas impossible.
En revanche, si un événement est impossible alors sa probabilité est nulle.
En d'autres termes : $A=\varnothing \Rightarrow P(A)=0$ mais $P(A)=0 \not\Rightarrow A=\varnothing$

2. Il faut déterminer P({ T > 10 }).



L'événement {T > 10 } est équivalent à { $X \in$] 0 ; 5 [} ∪ { $X \in$] 15 ; 20 [}.
P ({ T > 10 }) = P ({ X ∈] 0 ; 5 [}) + P ({ X ∈] 15 ; 20 [}) où X suit la loi $U\left([0,30]\right)$

$$P\left(\{T>5\}\right)=\int_0^5 \frac{1}{30}dt+\int_{10}^{15}\frac{1}{30}dt=\frac{5}{30}+\frac{5}{30}=\frac{1}{3}$$

Complément

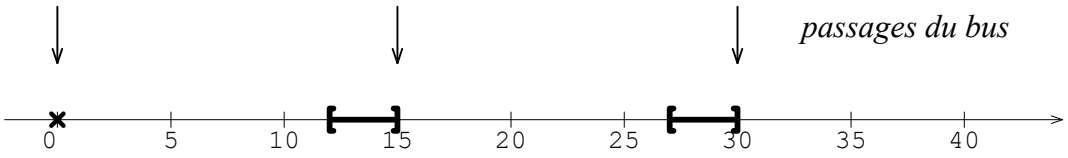
Quelle loi suit la variable aléatoire T ?

Notons F la fonction de répartition de la variable aléatoire T. Alors $F(t)=P(T\leqslant t)$

$t<0 \Rightarrow F(t)=0$

$t>15 \Rightarrow F(t)=1$, en effet $P(T>15)=1-P(T\leqslant 15)=0$.

$$t\in[0;15]\Rightarrow F(t)=P\left(X\in\{0\}\cup[15-t;15]\cup[30-t;30]\right)$$
$$=P\left(X\in\{0\}\right)+P\left(X\in[15-t;15]\right)+P\left(X\in[30-t;30]\right)$$
$$=\int_0^0\frac{1}{30}dx+\int_{15-t}^{15}\frac{1}{30}dx+\int_{30-t}^{30}\frac{1}{30}dx=0+\frac{t}{30}+\frac{t}{30}=\frac{t}{15}$$



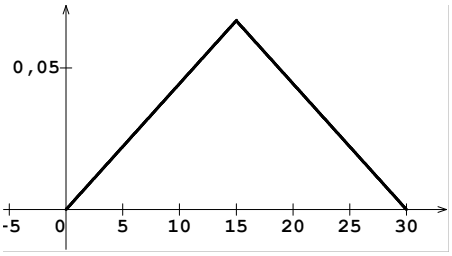
$$\begin{cases} F(t)=0 & \text{si} \quad t<0 \\ F(t)=\dfrac{t}{15} & \text{si} \quad t\in[\,0;15\,] \\ F(t)=1 & \text{si} \quad t>15 \end{cases}$$ donc la densité de T , $f=F'$ est $$\begin{cases} f(t)=0 & \text{si} \quad t<0 \\ f(t)=\dfrac{1}{15} & \text{si} \quad t\in[\,0;15\,] \\ f(t)=0 & \text{si} \quad t>15 \end{cases}$$

Donc la variable aléatoire T suit la loi uniforme sur $U\left([0,30]\right)$.

Prolongement – Reformulation possible de l'énoncé

On peut reprendre l'exercice avec une distribution moins simpliste et plus réaliste.

Les bus passent toutes les 15 minutes à un arrêt précis dès 7h du matin : 7h00 ; 7h15 ; …
Un usager a l'intention de prendre le bus de 7h15 mais son arrivée à l'arrêt est aléatoire. On fait l'hypothèse que l'heure exacte de son arrivée à cet arrêt, représentée par le nombre de minutes après 7h00, est une variable aléatoire X de densité f sur l'intervalle [0 ; 30] donnée par :



$f(t)=\dfrac{t}{225}$ sur [0 ; 15] et $f(t)=\dfrac{30-t}{225}$ sur [15 ; 30].

Le temps d'attente par l'usager d'un bus est une variable aléatoire notée T.

1. Quelle est la probabilité que l'usager attende moins de cinq minutes le prochain bus ?
2. Quelle est la probabilité qu'il attende plus de dix minutes ?

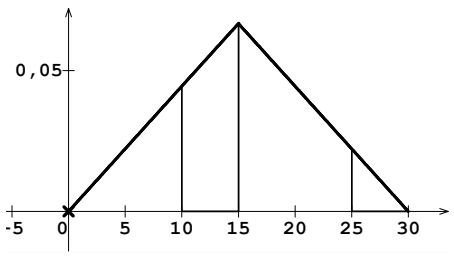
Corrigé

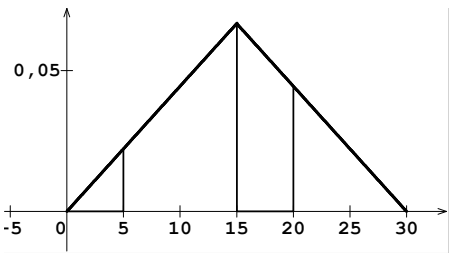
1. $\mathrm{P}(\{\,T<5\,\}) = \mathrm{P}(\{\,X=0\,\}) + \mathrm{P}(\{\,X\in\,]\,10\,;\,15\,]\,\}) + \mathrm{P}(\{\,X\in\,]\,25\,;\,30\,]\,\})$

$$\mathrm{P}(\{\,T<5\,\}) = \int_0^0 f(t)\,dt + \int_{10}^{15}\frac{t}{225}\,dt + \int_{25}^{30}\frac{(30-t)}{225}\,dt = 0 + \frac{5}{18} + \frac{1}{18} = \frac{1}{3}$$

2. $\mathrm{P}(\{\,T>10\,\}) = \mathrm{P}(\{\,X\in\,]\,0\,;\,5\,[\,\}) + \mathrm{P}(\{\,X\in\,]\,15\,;\,20\,[\,\})$

$$\mathrm{P}(\{\,T>10\,\}) = \int_0^5\frac{t}{225}\,dt + \int_{15}^{20}\frac{(30-t)}{225}\,dt = \frac{1}{18} + \frac{5}{18} = \frac{1}{3}$$





Exercice 2 Temps d'attente (2)

Thèmes abordés

- **Loi uniforme**

Énoncé

Énoncé original du document « ressources pour la classe terminale générale et technologique »

2. Partie A

Olivier vient tous les matins entre 7h et 7h 45 chez Karine prendre un café.

1) Sachant qu'Olivier ne vient jamais en dehors de la plage horaire indiquée et qu'il peut arriver à tout instant avec les mêmes chances, quelle densité peut-on attribuer à la variable aléatoire « heure d'arrivée d'Olivier » ?
2) Calculer la probabilité qu'Olivier sonne chez Karine :
   • Après 7h30 • Avant 7h10 • Entre 7h20 et 7h22 • A 7h30 exactement.

2. Partie B

Olivier et Karine décident de se retrouver au café de l'Hôtel de Ville entre 7h et 8h. Les instants d'arrivée d'Olivier et Karine sont assimilés à des variables aléatoires de loi uniforme sur $[0,1]$. Chacun attend un quart d'heure mais jamais au-delà de 8h. Quelle est la probabilité qu'ils se rencontrent ?

Reformulation possible de l'énoncé

Partie A

Olivier vient tous les matins entre 7h et 7h45 chez Karine prendre un café.

1. Sachant qu'Olivier ne vient jamais en dehors de la plage horaire indiquée et qu'il peut arriver à tout instant avec les mêmes chances, **quelle loi suit la variable aléatoire** « heure d'arrivée d'Olivier » ?

2. Calculer la probabilité qu'Olivier sonne chez Karine :
   • après 7h30 • avant 7h10 • entre 7h20 et 7h22 • à 7h30 exactement

Partie B

énoncé non modifié

Corrigé

Partie A

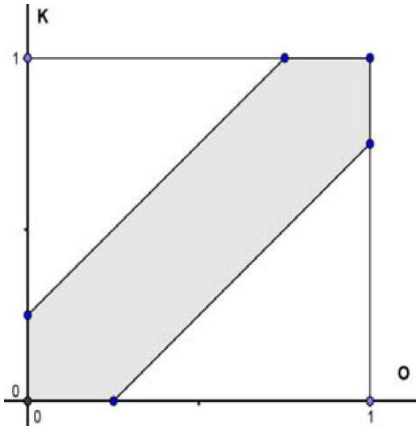
1. On note X la variable aléatoire « heure d'arrivée d'Olivier »
   D'après l'énoncé, elle suit la loi $U([0,45])$
   et la densité f de cette loi est définie par $f(t)=\frac{1}{45}$ si $t\in[0;45]$ et $f(t)=0$ sinon.
2. $P(X\geqslant 30)=\int_{30}^{45}\frac{1}{45}dx=\frac{45-30}{45}=\frac{1}{3}$

$$P(X \leqslant 10) = \int_0^{10} \frac{1}{45} dx = \frac{10}{45} = \frac{2}{15}$$
$$P(20 \leqslant X \leqslant 22) = \int_{20}^{22} \frac{1}{45} dx = \frac{2}{45}$$
$$P(X = 30) = \int_{30}^{30} \frac{1}{45} dx = 0$$

Partie B

Éléments de solution du document « ressources pour la classe terminale générale et technologique »

Pour la partie B, si on note *O* la variable aléatoire « instant d'arrivée d'Olivier » et *K* celle de Karine. La probabilité cherchée est $P(|O-K| \leq \frac{1}{4})$; en utilisant une représentation graphique, cette probabilité est l'aire de la zone grisée ci-dessous, ensemble des points de coordonnées (x, y) du carré tels que $|x-y| \leq 0{,}25$. (On trouve $\frac{7}{16}$).



Commentaires sur la question b de l'exercice 2 :

X et *Y* sont deux variables aléatoires indépendantes qui suivent la loi $U([0,1])$.

Si l'on *considère* que choisir deux nombres *x* et *y* (c'est l'expérience aléatoire), indépendamment l'un de l'autre, dans $[0;1]$ au hasard ,de façon uniforme, revient à choisir au hasard un couple (x, y) dans $[0;1]\times[0;1]$ de façon uniforme que l'on peut identifier à un point $M(x,y)$ du domaine plan

$C = \{M(x\ ; y) \in P /\ 0 \leqslant x \leqslant 1 \quad \text{et} \quad 0 \leqslant\ y \leqslant 1\}$, on peut « *admettre* » que $P((x,y) \in A)$ est proportionnelle à l'aire de *A* si $A \subset C$.

La constante de proportionnalité est égale à 1, puisque l'aire du carré (l'univers) est 1, ce qui revient à *accepter intuitivement* que la loi de (X, Y) est la loi uniforme sur $[0;1]\times[0;1]$.

$$P\left(|X-Y| \leqslant \frac{1}{4}\right) = Aire\left(\left\{M(x,y) \in P/\ 0 \leqslant x \leqslant 1 \quad \text{et} \quad 0 \leqslant\ y \leqslant 1 \text{ et } |x-y| \leqslant \frac{1}{4}\right\}\right)$$
$$P\left(|X-Y| \leqslant \frac{1}{4}\right) = Aire\left(\left\{M(x,y) \in C / -\frac{1}{4} \leqslant x-y \leqslant \frac{1}{4}\right\}\right)$$ (aire de la zone grisée)

Soit $1-2\times\frac{1}{2}\times\left(\frac{3}{4}\right)^2=\frac{7}{16}$.

De façon plus rigoureuse :

X et *Y* sont deux variables aléatoires **indépendantes** qui suivent la loi $U\left([0,1]\right)$.

$P\left((a<X\leqslant b)\cap(c<Y\leqslant d)\right)=P(a<X\leqslant b)\times P(c<Y\leqslant d)$

puisque les variables sont indépendantes, et donc :

$P\left((a<X\leqslant b)\cap(c<Y\leqslant d)\right)=\int_a^b 1_{[0;1]}(x)dx\times\int_c^d 1_{[0;1]}(y)dy=(b-a)(d-c)$

où 1_A est la fonction indicatrice de $A\subset E$ définie sur *E* par : $1_A(x)=1$ si $x\in A$

$1_A(x)=0$ sinon.

Ainsi $1_{[0;1]}(x)1_{[0;1]}(y)=1_{[0;1]^2}(x,y)\ \forall(x,y)\in\mathbb{R}^2$)

On a encore, avec le théorème de Fubini :

$P\left((a<X\leqslant b)\cap(c<Y\leqslant d)\right)=\int_c^d\int_a^b 1_{[0;1]}(x)1_{[0;1]}(y)dxdy$

$P\left((a<X\leqslant b)\cap(c<Y\leqslant d)\right)=\int_c^d\int_a^b 1_{[0;1]^2}(x,y)dxdy$

$P\left(|X-Y|\leqslant\frac{1}{4}\right)=P\left(-\frac{1}{4}\leqslant X-Y\leqslant\frac{1}{4}\right)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}1_D\,dxdy$ où

$D=\left\{(x,y)\in[0;1]^2\ /\ -\frac{1}{4}\leqslant x-y\leqslant\frac{1}{4}\right\}$

Ce qui justifie le calcul d'aire précédent.

Conclusion

La résolution de la question 2.b. fait appel à trop de notions implicitement admises et loin d'être évidentes ni même intuitives. Cet exercice est à déconseiller.

Partie C

STATISTIQUE INFÉRENTIELLE

Intervalle de fluctuation et prise de décision

Exercice 1 Discrimination ?

Thèmes abordés

- **Fréquences**
- **Loi normale**

Énoncé

Schématiquement, le sous-emploi comprend les personnes actives occupées (c'est-à-dire non chômeuses) qui travaillent à temps partiel et souhaitent travailler davantage, ou qui travaillent à temps complet mais ont connu un chômage partiel pendant une semaine de référence.
Le sous-emploi est mesuré par l'enquête Emploi périodique de l'INSEE. En 2016, celle-ci a donné les résultats suivants :

Hommes:	en sous-emploi	non en sous-emploi	total
< bac+3	439	9832	10271
>= bac+3	73	2908	2981
Femmes:			
< bac+3	1040	7508	8548
>= bac+3	165	2902	3067

Nous allons analyser les liens du sous-emploi avec le niveau de diplôme et avec le genre.

1. Statistique descriptive

a. Quelle est la proportion $f_{<3}$ des personnes en sous-emploi parmi celles ayant fait moins de trois ans d'études supérieures dans cet échantillon ? Quelle est la proportion $f_{\geq 3}$ des personnes en sous-emploi parmi celles ayant fait au moins trois ans d'études supérieures ?
Calculer la différence $\delta_{études} = f_{<3} - f_{\geq 3}$. Commenter.

b. Quelle est, parmi les femmes, la proportion f_F de celles qui sont en sous-emploi ? Quelle est, parmi les hommes, la proportion f_H de ceux qui sont en sous-emploi ?
Calculer la différence $\delta_{genre} = f_F - f_H$. D'après vous, y a-t-il discrimination entre hommes et femmes dans le sous-emploi ?

c. Quelle est, parmi les femmes, la proportion g_F de celles ayant fait au moins trois ans d'études supérieures ? Quelle est, parmi les hommes, la proportion g_H de ceux ayant fait au moins trois ans d'études supérieures ?
Calculer la différence $d_{genre} = g_F - g_H$.

d. Quelle est, parmi les femmes ayant fait moins de trois ans d'études supérieures, la proportion $f_{F<3}$ de celles ayant fait moins de trois ans d'études supérieures soit en sous-emploi ? Quelle cette proportion $f_{H<3}$ pour un homme ?
Calculer la différence $\delta_{genre,<3} = f_{F<3} - f_{H<3}$ et commenter.

e. Reprendre la question (**d**) pour les individus de l'échantillon ayant fait au moins trois ans

d'études supérieures (on notera respectivement $f_{F,\geq 3}$, $f_{H,\geq 3}$ et $\delta_{genre,\geq 3}$ les proportions correspondantes et leur différence).

2. Décision.

On cherche à savoir si les résultats de l'enquête sont significatifs.

a. Au vu de la différence $\delta_{études} = f_{<3} - f_{\geq 3}$, peut-on affirmer, au risque 5%, qu'il y a plus de personnes en sous-emploi parmi celles ayant fait moins de trois ans d'études supérieures que parmi celles ayant fait au moins trois ans d'études supérieures ?

b. Au vu de la différence $\delta_{genre} = f_F - f_H$, peut-on affirmer, au risque 5%, qu'il y a plus de sous-emploi parmi les femmes que parmi les hommes ?

Corrigé

1.a. (cf. tableau de calcul ci-dessous) :

$$f_{<3} = 0{,}079 \; ; \; f_{\geq 3} = 0{,}039 \; ; \; \delta_{études} = 0{,}039$$

Les personnes les moins diplômées sont donc deux fois plus touchées par le sous-emploi que les plus diplômées.

1.b. (cf. tableau de calcul ci-dessous) :

$$f_F = 0{,}104 \; ; \; f_H = 0{,}039 \; ; \; \delta_{genre} = 0{,}065$$

Le taux de sous-emploi des femmes est nettement supérieur à celui des hommes, de 6,5 points de pourcentage.

À la vue de δ_{genre}, on peut se demander s'il y a discrimination entre les genres. Cependant, ce calcul ne suffit pas pour répondre à cette question. En effet, on n'a pas tenu compte du niveau de formation des personnes : si les femmes étaient moins diplômées que les hommes, cela pourrait expliquer qu'elles soient plus touchées par le sous-emploi. Est-ce le cas ?

1.c. (cf. tableau de calcul ci-dessous) :

$$g_F = 0{,}264 \; ; \; g_H = 0{,}225 \; ; \; d_{genre} = 0{,}039$$

Les femmes ont plus fréquemment que les hommes fait au moins trois ans d'études supérieures. Ce n'est donc pas leur niveau d'étude qui peut expliquer le fait qu'elles soient plus touchées par le sous-emploi. Pour nous en convaincre, nous allons calculer les taux de sous-emploi des femmes et des hommes par niveau de formation.

1.c. (cf. tableau de calcul ci-dessous) :

$$f_{F<3} = 0{,}122 \; ; \; f_{H<3} = 0{,}043 \; ; \; \delta_{genre,<3} = 0{,}079$$

Chez les personnes les moins diplômées, les taux de sous-emploi sont plus grands que dans la population totale, et la différence entre celui des femmes et celui des hommes est plus grande elle aussi.

1.e. (cf. tableau de calcul ci-dessous) :

$$f_{F,\geq 3} = 0{,}054 \; ; \; f_{H,\geq 3} = 0{,}025 \; ; \; \delta_{genre,\geq 3} = 0{,}029$$

Chez les personnes les plus diplômées, les taux de sous-emploi sont nettement plus petits que dans la population totale, et la différence entre celui des femmes et celui des hommes est nettement plus petite elle aussi. Si l'on considère le rapport, $\frac{f_{F<3}}{f_{H<3}} = \frac{0{,}122}{0{,}043} \simeq 2{,}84$ tandis que $\frac{f_{F,\geq 3}}{f_{H,\geq 3}} = \frac{0{,}054}{0{,}025} = 2{,}16$. La discrimination à l'encontre des femmes est donc constatée chez les moins comme chez les plus diplômés, mais elle est plus sévère chez les premiers.

Tableau de calcul :

H:	sous-emploi	non sous-emploi	total	f	d(F-H)	
< bac+3	439	9832	10271	0.043	<b+3:	0.079
>=bac+3	73	2908	2981	0.025	>=b+3:	0.029
total H	**512**	**12740**	**13252**	**0.039**	ens:	**0.065**
F:						
< bac+3	1040	7508	8548	0.122	g_F:	0.264
>=bac+3	165	2902	3067	0.054	g_H:	0.225
total F	**1204**	**10410**	**11615**	**0.104**	delta:	0.039
ensemble				F	D	
< bac+3	1479	17341	18819	0.079	0.039	
>=bac+3	238	5809	6047	0.039		
total	**1717**	**23150**	**24867**	**0.069**		

2.a.

Soient $n_{<3}$ et $n_{\geq 3}$ les nombres respectifs de personnes ayant fait moins de trois ans et au moins trois ans d'études supérieures. On a $n_{<3} = 18819 > 30$ et $n_{\geq 3} = 6047 > 30$.
D'autre part, soient $n_{<3}^{se}$ et $n_{\geq 3}^{se}$ les nombres de personnes en sous-emploi dans ces sous-échantillons respectifs. On a $n_{<3}^{se} = 1479 > 5$ et $n_{\geq 3}^{se} = 238 > 5$.
Nous sommes donc dans les conditions où l'on peut admettre l'approximation normale. On a donc, parmi les $n_{<3}$ personnes ayant fait moins de trois ans d'études supérieures :

$$f_{<3} \sim N(p_{<3}; \frac{p_{<3}(1-p_{<3})}{n_{<3}})$$

De même, parmi les $n_{\geq 3}$ personnes ayant fait moins de trois ans d'études supérieures :

$$f_{\geq 3} \sim N(p_{\geq 3}; \frac{p_{\geq 3}(1-p_{\geq 3})}{n_{\geq 3}})$$

Rappelons que lorsque Xet Ysont deux variables aléatoires indépendantes, on a :

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

Les individus sont indépendants entre eux, donc les deux sous-échantillons constitués des personnes ayant fait respectivement moins de trois ans et au moins trois ans d'études supérieures sont indépendants, et l'on a:

$$\delta_{études} = f_{<3} - f_{\geq 3} \sim N(p_{<3} - p_{\geq 3}; \frac{p_{<3}(1-p_{<3})}{n_{<3}} + \frac{p_{\geq 3}(1-p_{\geq 3})}{n_{\geq 3}})$$

où l'on peut approximer la variance en remplaçant les probabilités par les fréquences correspondantes :

$$\delta_{études} = f_{<3} - f_{\geq 3} \sim N(p_{<3} - p_{\geq 3}; \frac{f_{<3}(1-f_{<3})}{n_{<3}} + \frac{f_{\geq 3}(1-f_{\geq 3})}{n_{\geq 3}})$$

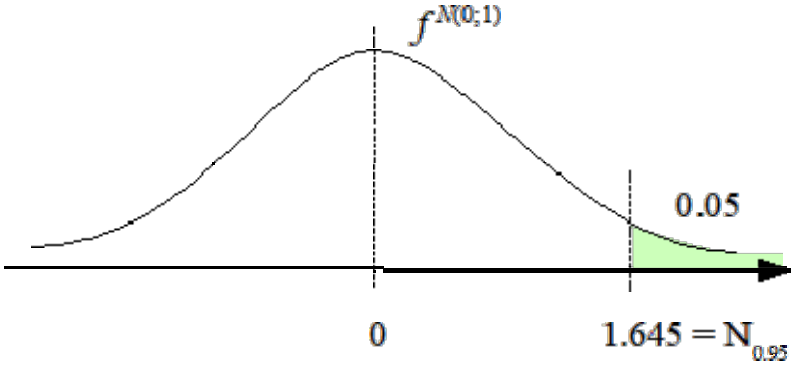
On voudrait tester $H_0 : p_{<3} - p_{\geq 3} = 0$ contre l'hypothèse alternative $H_1 : p_{<3} - p_{\geq 3} > 0$ avec un risque (de première espèce) de 0,05.
Sous H_0, $p_{<3} = p_{\geq 3} = p$, que l'on estime par la proportion de personnes en sous-emploi dans la population totale $f = \frac{1717}{24867}$. Sous H_0, on devrait donc avoir :

$$\delta_{études} = f_{<3} - f_{\geq 3} \sim N\left(0; f(1-f)(\frac{1}{n_{<3}} + \frac{1}{n_{\geq 3}})\right)$$

$$\Leftrightarrow T_{études} = \frac{\delta_{études}}{\sqrt{f(1-f)(\frac{1}{n_{<3}} + \frac{1}{n_{\geq 3}})}} \sim N(0;1)$$

On doit refuser H_0 en faveur de H_1 si la réalisation de T tombe dans une région de basse

densité et de faible probabilité (0,05) "orientée vers H_1, c'est-à-dire où $\delta_{études}$ est suffisamment grande (il s'agit en effet de décider $p_{<3} - p_{\geq 3} > 0$, et non $p_{<3} - p_{\geq 3} < 0$).



Calcul de $T_{études}$:

$$f = \frac{1717}{24867} = 0{,}069$$

$$\Leftrightarrow T_{études} = \frac{0{,}039}{\sqrt{0{,}069(1-0{,}069)(\frac{1}{18819}+\frac{1}{6047})}} = 10{,}41$$

Il faut donc refuser H_0 et conclure à $H_1 : p_{<3} > p_{\geq 3}$ au risque 5% de se tromper.
Si l'on s'était contenté de la constatation empirique $f_{<3} > f_{\geq 3}$ faite sur l'échantillon, et qu'on l'avait généralisée à la population entière en concluant $p_{<3} > p_{\geq 3}$, on n'aurait absolument pas géré le risque d'erreur. C'est le gros apport de la théorie de la décision: quantifier, et limiter les risques d'erreur.

2.b. La démarche est exactement la même.

$$\delta_{genre} = f_F - f_H \sim N(p_F - p_H; \frac{p_F(1-p_F)}{n_F} + \frac{p_H(1-p_H)}{n_H})$$

Il s'agit de tester $H_0 : p_F - p_H = 0$ contre l'hypothèse alternative $H_1 : p_F - p_H > 0$.
Sous H_0, $p_F = p_H = p$, que que nous avons déjà estimée par la proportion de personnes en sous-emploi dans la population totale $f = \frac{1717}{24867}$. Sous H_0, on devrait donc avoir :

$$\delta_{genre} = f_F - f_H \sim N\left(0; f(1-f)(\frac{1}{n_F}+\frac{1}{n_H})\right)$$

$$\Leftrightarrow T_{genre} = \frac{\delta_{études}}{\sqrt{f(1-f)(\frac{1}{n_F}+\frac{1}{n_H})}} \sim N(0;1)$$

La région de rejet de $H_0 : p_F - p_H = 0$ en faveur de $H_1 : p_F - p_H > 0$ est une fois de plus : $\{T_{genre} > 1{,}645\}$.
Calcul de T_{genre}:

$$\Leftrightarrow T_{genre} = \frac{0{,}039}{\sqrt{0{,}069(1-0{,}069)(\frac{1}{11615}+\frac{1}{13252})}} = 12{,}10$$

Il faut donc refuser H_0 et conclure à $H_1 : p_F > p_H$ au risque 5% de se tromper.

Exercice 2 Audiences

Thèmes abordés

- **Intervalles de fluctuation**

Énoncé

Lors de la coupe du monde de football de 2006 on a évalué le taux d'audience des retransmissions de matchs à 58%.

1. Donner l'intervalle de fluctuation à 95% de la fréquence de téléspectateurs regardant le match pour un échantillon de 1200 téléspectateurs en juin 2006.
2. Dans un village, il y a 1200 téléspectateurs potentiels, quelle indication peut-on donner sur le nombre de téléspectateurs ayant regardé la coupe du monde ?
3. Comparer l'intervalle déterminé à la question 1. avec celui obtenu avec la "formule" : $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$.
4. En juin 2010 sur un échantillon de taille 1200 (téléspectateurs), 636 affirment avoir regardé la coupe du monde. Que pensez-vous de ce résultat.

Corrigé

Lors de la coupe du monde de football de 2006 on a évalué le taux d'audience des retransmissions de matchs à 58%.

Population : l'ensemble des individus qui regardent la télévision pendant la retransmission du match.
Caractère étudié : un téléspectateur pris au hasard dans la population regarde-t-il ou pas le match retransmis ?
58% : c'est la probabilité p qu'un téléspectateur pris au hasard dans la population regarde le match retransmis

1. Donner l'intervalle de fluctuation pour un échantillon de 1200 téléspectateurs en juin 2006.
   Taille de l'échantillon : $n = 1200$
   - En seconde et en première bac pro on utilise l'approximation $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$;
     on obtient les résultats suivants :
     $$\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]=\left[0{,}58-\frac{1}{\sqrt{1200}}, 0{,}58+\frac{1}{\sqrt{1200}}\right]=[0{,}5511\ ;\ 0{,}6089]$$
   - En première, on utilise la loi binomiale (rappel du cours) :
     - a est le plus petit entier tel que $P(Y_n \leqslant a) > 0{,}025$
     - b est le plus petit entier tel que $P(Y_n \leqslant b) \geq 0{,}975$

   et on obtient les résultats suivants à l'aide d'un tableur : $a = 662$ et $b = 729$
   car $P(Y_n \leqslant 662) = 0{,}02527$ et $P(Y_n \leqslant 661) = 0{,}02203$

et $P(Y_n \leqslant 728) = 0{,}97163$ et $P(Y_n \leqslant 729) = 0{,}97525$

on pourra faire vérifier par les élèves les trois inégalités suivantes :

$P(a \leqslant Y_n \leqslant b) \geqslant 0{,}95$: $P(662 \leqslant Y_n \leqslant 729) = 0{,}97525 - 0{,}02203 = 0{,}95322$

$P(a+1 \leqslant Y_n \leqslant b) < 0{,}95$: $P(663 \leqslant Y_n \leqslant 729) = 0{,}97525 - 0{,}02527 = 0{,}94998$

$P(a \leqslant Y_n \leqslant b-1) < 0{,}95$: $P(662 \leqslant Y_n \leqslant 728) = 0{,}97163 - 0{,}02203 = 0{,}94960$

donc, l'intervalle de fluctuation à 95% est $\left[\frac{662}{1200}; \frac{729}{1200}\right] = [0{,}5517; 0{,}6075]$

- en terminale, on utilise l'intervalle de fluctuation asymptotique obtenu avec la loi normale (rappel du cours) :

$$\left[p - 1{,}96\sqrt{\frac{pq}{n}}; p + 1{,}96\sqrt{\frac{pq}{n}} \right] \text{ puisque } n \geqslant 30 \text{ , } np \geqslant 5 \text{ et } nq \geqslant 5$$

on obtient

$$\left[p - 1{,}96\sqrt{\frac{pq}{n}}; p + 1{,}96\sqrt{\frac{pq}{n}} \right] = \left[0{,}58 - 1{,}96\sqrt{\frac{0{,}58 \times 0{,}42}{1200}}; 0{,}58 + 1{,}96\sqrt{\frac{0{,}58 \times 0{,}42}{1200}} \right]$$

soit $[0{,}5521; 0{,}6079]$

2. Dans un village, il y a 1200 téléspectateurs potentiels, quelle indication peut-on donner sur le nombre de téléspectateurs ayant regardé la coupe du monde ?

Pour chacun des trois niveaux on multiplie par 1200 les bornes de chaque intervalle de fluctuation.

Le résultat est donné sous forme d'intervalle fermé ; la borne inférieure est arrondie à l'entier supérieur, la borne supérieure à l'entier inférieur.

- en seconde et en première bac pro on obtient :
  $[0{,}5511 \times 1200 ; 0{,}6089 \times 1200] = [662 ; 730]$
  donc, la probabilité que le nombre de téléspectateurs soit compris entre 662 et 730 est proche de 95%.
- en première, avec la loi binomiale, la question est sans objet : pour répondre à la question 1. on a identifié l'intervalle $[662; 729]$
- en terminale, on utilise l'intervalle de fluctuation asymptotique
  $[0{,}5521 \times 1200 ; 0{,}6079 \times 1200] = [663; 729]$

Remarque : pour n = 1200, les résultats sont quasiment identiques, et on constate bien que $[663; 729] \subset [662 ; 730]$ puisque l'approximation utilisée en classe de seconde est un majorant de celle utilisée en terminale.

3. Comparer l'intervalle déterminé à la question 1. avec celui obtenu avec la "formule" : $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$.

- en seconde et en première bac pro : sans objet
- en première, on utilise la loi binomiale:
  $\left[\frac{662}{1200}; \frac{729}{1200}\right] = [0{,}5517; 0{,}6075]$ qui est un intervalle à $0{,}95322$
  Il est à comparer avec l'approximation
  $$\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right] = \left[0{,}58 - \frac{1}{\sqrt{1200}}, 0{,}58 + \frac{1}{\sqrt{1200}} \right] = [0{,}5511 ; 0{,}6089]$$

- en terminale, on utilise l'intervalle de fluctuation asymptotique
  $[0,5521;0,6079]$ qui est un intervalle asymptotique à 95%, qui est à comparer avec l'approximation plus grossière $[0,5511\ ;\ 0,6089]$.

4. En juin 2010 sur un échantillon de taille 1200 (téléspectateurs), 412 affirment avoir regardé la coupe du monde. Que pensez-vous de ce résultat ? (Prise de décision)
$\frac{636}{1200}=0,53$ dans tous les cas, cette valeur est en dehors de l'intervalle de fluctuation. On peut donc considérer que sur l'échantillon observé en 2010 le taux d'audience n'est pas de 58% comme en 2006.
En d'autres termes l'échantillon remet en cause l'hypothèse selon laquelle le taux d'audience serait inchangé.

Prolongement

Dans ce cas on peut donc proposer de calculer un intervalle de confiance à 95% pour le taux d'audience en 2010.

$$\left[f-\frac{1}{\sqrt{n}},f+\frac{1}{\sqrt{n}}\right]=\left[0,53-\frac{1}{\sqrt{1200}},0,53+\frac{1}{\sqrt{1200}}\right]=[0,5011\ ;\ 0,5589]$$

et en terminale STL et STI2D

$$\left[f-1,96\sqrt{\frac{f(1-f)}{n}};f+1,96\sqrt{\frac{f(1-f)}{n}}\right]=\left[0,53-1,96\sqrt{\frac{0,53\times0,47}{1200}};0,58+1,96\sqrt{\frac{0,53\times0,47}{1200}}\right]$$
$$=[0,5018;0,5582]$$

Exercice 3 Préparation pour pancakes

D'après baccalauréat STL, Antilles Guyane, 2014

Thèmes abordés

- **Intervalle de fluctuation**
- **Prise de décision**

Énoncé

Une entreprise agroalimentaire dispose d'un parc de machines d'ensachage toutes identiques. Elle lance une production de paquets de préparation pour pancakes. Les paquets doivent contenir 650g de préparation. On considère qu'un paquet est commercialisable s'il contient entre 648g et 652g.
Les réglages de la machine d'ensachage ont été établis dans l'objectif d'obtenir une proportion $p = 97\%$ de paquets de préparation commercialisables. Afin d'évaluer l'efficacité de ces réglages, on effectue un contrôle qualité sur un premier échantillon de 400 paquets fabriqués.

1. En supposant que cet objectif a été atteint déterminer un intervalle de fluctuation asymptotique au seuil de 95% de la proportion de paquets de préparation commercialisables dans un échantillon de taille 400.

2. Parmi les 400 paquets de l'échantillon, 381 sont commercialisables.
En utilisant l'intervalle de fluctuation obtenu à la question 1, peut-on considérer au seuil de 95% que l'objectif a été atteint ?

Commentaires
Dans la question 2, l'énoncé original emploie le mot "estimer" au lieu du mot "considérer". "estimer" nous semble mal choisi car dans un exercice de statistique, le mot "estimer" renvoie à la procédure d'estimation alors qu'il s'agit ici d'une procédure de test (appelée prise de décision dans les programmes).

Corrigé

Population : les paquets de préparation pour pancakes produits par la machine.
Caractère étudié : "être commercialisable ".
97% : c'est la probabilité p qu'un paquet pris au hasard dans la population soit commercialisable.

1. On considère donc que $p = 0{,}97$.
Taille de l'échantillon : $n = 400$
L'expérience aléatoire est : prélever un échantillon de taille 400 dans la population.
Le tirage est assimilé à un tirage avec remise.
La variable aléatoire est X_n, le nombre de paquets commercialisables dans l'échantillon aléatoire.
La loi de probabilité de X_n est la loi binomiale $B(n,p)$.
Comme $n = 400 \geqslant 30$, $np = 388 \geqslant 5$ et $nq = 12 \geqslant 5$,

On considère que l'on peut approximer la loi binomiale par une loi normale (théorème de Moivre-Laplace), ce qui valide ce qui suit.
L'intervalle de fluctuation asymptotique de la proportion de paquets de préparation commercialisables au seuil de 95% du programme pour un échantillon de taille 400 est :

$$\left[p - 1{,}96\sqrt{\frac{p\times(1-p)}{n}} ; p + 1{,}96\sqrt{\frac{p\times(1-p)}{n}} \right]$$

soit $\left[0{,}97 - 1{,}96\sqrt{\frac{0{,}97\times 0{,}03}{400}} ; 0{,}97 + 1{,}96\sqrt{\frac{0{,}97\times 0{,}03}{400}} \right] = \left[0{,}9533 ; 0{,}9867 \right]$

Remarque 1
Parmi l'infinité d'intervalles de fluctuation de seuil 95% , c'est le seul qui soit centré sur *p*, et c'est celui de plus faible amplitude. En effet :

$\forall \alpha_1 \in [0\ ; 0{,}05]$: $P\left(p + q_{\alpha_1}\sqrt{\frac{pq}{n}} \leqslant F_n \leqslant p + q_{0{,}95+\alpha_1}\sqrt{\frac{pq}{n}} \right) = 0{,}95$, où q_α désigne le quantile d'ordre α

de *N(* 0 ;1*)*, c'est-à-dire la valeur q_α telle que $P(X \leqslant q_\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q_\alpha} e^{-\frac{x^2}{2}} dx = \alpha$ (donc pour *X* suivant la loi *N(* 0 ;1*)*).

L'amplitude de l'intervalle est $L = \sqrt{\frac{pq}{n}} \left(q_{0{,}95+\alpha_1} - q_{\alpha_1} \right)$ et on peut montrer que $\left(q_{0{,}95+\alpha_1} - q_{\alpha_1} \right)$ est minimum pour $\alpha_1 = 0{,}025$.

2. La fréquence observée sur l'échantillon est $\frac{381}{400} = 0{,}9525$, qui n'est pas dans l'intervalle de fluctuation précédemment calculé. Donc on doit rejeter l'hypothèse *p* = 0,97 au risque de 5% de se tromper (c'est à dire que pour 5% des échantillons, en moyenne, on rejettera l'hypothèse à tort, c'est à dire que l'on décidera que la machine n'est pas conforme alors qu'elle l'est)

Remarque 2
Pour cette procédure de test, on doit impérativement utiliser l'intervalle centré sur *p*. En effet il s'agit du test bilatéral de $H_0 : p = 0{,}97$ contre $H_1 : p \neq 0{,}97 \Leftrightarrow p < 0{,}97$ *ou* $p > 0{,}97$.
L'hypothèse *H1* est symétrique par rapport à la valeur $p = 0{,}97$.

L'énoncé demande de tester l'hypothèse $H_0 : p = 0{,}97$ contre l'hypothèse $H_1 : p \neq 0{,}97$.
Mais tester l'efficacité de la machine reviendrait à tester (test unilatéral) l'hypothèse $H_0 : p \leqslant 0{,}97$ contre l'hypothèse $H_1 : p > 0{,}97$ (en vue d'accepter significativement cette dernière, au vu d'une fréquence supérieure à un certain seuil, et non seulement suffisamment distante de 0.97).

Remarque 3

En première, avec la loi binomiale on obtient l'intervalle de fluctuation $\left[\frac{381}{400} ; \frac{394}{400} \right]$, soit $[0{,}9525\ ;\ 0{,}985\ ]$ au seuil de 96%, et la décision est la même.

Exercice 4 Détecteur de fraudes

Thèmes abordés

- **Intervalle de fluctuation**

Énoncé

Si on s'intéresse au premier chiffre significatif (*) de tous les nombres qui apparaissent dans un document chiffré, on s'aperçoit, contre toute attente, que dès lors que ces données chiffrées sont nombreuses et expriment des mesures (prix, documents comptables, mesures de grandeurs physiques ...) chaque chiffre significatif n'apparaît pas à la même fréquence. Franck Benford a su établir la distribution théorique d'apparition de chaque chiffre significatif que voici :

1er chiffre	1	2	3	4	5	6	7	8	9
probabilité	0,301	0,176	0,124	0,097	0,079	0,068	0,057	0,053	0,045

(*)exemples : le premier chiffre significatif de 452,25 est 4
le premier chiffre significatif de 0,25 est 2

Dans certains pays, cette loi est utilisée pour repérer d'éventuelles fraudes. Le service des impôts étudie les bilans financiers de certaines entreprises.

Le service des impôts a étudié les bilans financiers de deux entreprises A et B. Dans les documents comptables fournis par l'entreprise A, le chiffre significatif *1* apparait 4 479 fois sur les 15 500 données chiffrées étudiées. Pour l'entreprise B, ce chiffre significatif *1* apparait 707 fois sur les 2 500 données chiffrées étudiées.

Quelle entreprise peut être suspectée de fraudes ?

Corrigé

Population : l'ensemble des nombres apparaissant dans le bilan comptable d'une entreprise.
Le caractère étudié : le premier chiffre significatif du nombre, est-il ou pas le chiffre 1 ?
0,301 : c'est la probabilité p qu'un nombre pris au hasard ait 1 pour premier chiffre significatif, si on fait l'hypothèse que la proportion de 1 dans la population est celle de la distribution théorique de Benford.

$$\lim_{k\to 0}\frac{1}{k}\left(\lim_{h\to 0}\frac{F(a+h,b+k)-F(a+h,b)-F(a,b+k)+F(a,b)}{h}\right)=\lim_{k\to 0}\frac{1}{k}\left(F'_x(a,b+k)-F'_x(a,b)\right)$$
$$=F''_{xy}(a,b)=f(a,b)$$

Et de même pour le second membre,
Pour l'entreprise A on calcule l'intervalle de fluctuation I_A :
Taille de l'échantillon : $n = 15500$.

- en seconde et en première bac pro on utilise l'approximation $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$

  on obtient les résultats suivants :

$$\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]=\left[0{,}301-\frac{1}{\sqrt{15500}}, 0{,}301+\frac{1}{\sqrt{15500}}\right]=[0{,}2930\ ;\ 0{,}3090]$$

- en terminale, on utilise l'intervalle de fluctuation asymptotique obtenu avec la loi normale (rappel du cours) :

$$\left[p-1{,}96\sqrt{\frac{pq}{n}}; p+1{,}96\sqrt{\frac{pq}{n}}\right] \text{ puisque } n\geqslant 30 \ , \ np\geqslant 5 \text{ et } nq\geqslant 5$$

on obtient

$$\left[p-1{,}96\sqrt{\frac{pq}{n}}; p+1{,}96\sqrt{\frac{pq}{n}}\right]=\left[0{,}301-1{,}96\sqrt{\frac{0{,}301\times 0{,}699}{15500}}; 0{,}301+1{,}96\sqrt{\frac{0{,}301\times 0{,}699}{15500}}\right]$$
$$=[0{,}2938; 0{,}3082]$$

Pour l'entreprise B on calcule l'intervalle de fluctuation I_B :
Taille de l'échantillon : $n = 2500$;

- en seconde et en première bac pro on utilise l'approximation $\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]$

on obtient les résultats suivants :

$$\left[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}\right]=\left[0{,}301-\frac{1}{\sqrt{2500}}, 0{,}301+\frac{1}{\sqrt{2500}}\right]=[0{,}2810\ ;\ 0{,}3210]$$

- en terminale, on utilise l'intervalle de fluctuation asymptotique obtenu avec la loi normale (rappel du cours) :

$$\left[p-1{,}96\sqrt{\frac{pq}{n}}; p+1{,}96\sqrt{\frac{pq}{n}}\right] \text{ puisque } n\geqslant 30 \ , \ np\geqslant 5 \text{ et } nq\geqslant 5$$

on obtient

$$\left[p-1{,}96\sqrt{\frac{pq}{n}}; p+1{,}96\sqrt{\frac{pq}{n}}\right]=\left[0{,}301-1{,}96\sqrt{\frac{0{,}301\times 0{,}699}{2500}}; 0{,}301+1{,}96\sqrt{\frac{0{,}301\times 0{,}699}{2500}}\right]$$
$$=[0{,}2830; 0{,}3190]$$

Remarque 1 : On a pu déterminer les intervalles de fluctuation au seuil de 95%, car pour l'entreprise B les conditions d'application sont vérifiées.
En effet, $n_B = 2500$ donc $n_B \geq 30$, $n_B p = 2500\times 0{,}301 = 752{,}5$ donc $n_B p \geq 5$
et $n_B q = 2500\times 0{,}699 = 1747{,}5$ donc $n_B q \geq 5$.
Conditions évidemment vérifiées pour l'échantillon de l'entreprise A.
On calcule maintenant les fréquences : $f_A = \frac{4479}{15500} \simeq 0{,}2890$ et $f_B = \frac{707}{2500} \simeq 0{,}2828$
On constate que $f_A \notin I_A$, donc au risque $\alpha = 0{,}05$, on rejette l'hypothèse $p = 0{,}301$.
Il est légitime de suspecter une anomalie dans les comptes de la société et donc de proposer une investigation comptable plus poussée.
En revanche $f_B \in I_B$, on ne rejette pas l'hypothèse $p = 0{,}301$. Il n'y a pas de raison légitime de conseiller une investigation comptable plus poussée.

Remarque 2
Ces conclusions sont identiques quel que soit le niveau de la classe.

Remarque 3

f_A est plus proche de 0,301 que f_B, et pourtant il convient de rejeter l'hypothèse $p = 0{,}301$ pour la société A mais pas pour la société B : il y a une grande différence entre les tailles des deux échantillons.

Remarque 4

Pour aller plus loin : si l'on dispose de la répartition des 9 chiffres significatifs de tous les nombres qui apparaissent dans un document chiffré, on pourra alors construire un test de Khi2 d'adéquation à la loi de Benford.

On note p_k la probabilité d'apparition du chiffre *k*, et n_k le nombre de fois où le premier chiffre significatif est *k* dans un document chiffré contenant *N* nombres.

La statistique de test sera : $\sum_{k=1}^{9} \frac{(n_k - Np_k)^2}{Np_k}$ à comparer au quantile à 95% de la loi de Khi2 à 8 degrés de liberté.

Exercice 5 Problème de la surréservation (surbooking)

Thèmes abordés

- **Intervalle de fluctuation**
- **Centrer et réduire une variable aléatoire**
- **Théorème de Moivre Laplace**

Énoncé original du document « ressources pour la classe terminale générale et technologique »

Une compagnie aérienne possède des A340 (longs courriers) d'une capacité de 300 places.

Cette compagnie a vendu n billets pour le vol 2012.

La probabilité pour qu'un acheteur se présente à l'embarquement est p et les comportements des acheteurs sont indépendants les uns des autres.

On note X_n la variable aléatoire désignant le nombre d'acheteurs d'un billet se présentant à l'embarquement.

La compagnie cherche à optimiser le remplissage de l'avion en vendant éventuellement plus de places que la capacité totale de l'avion (surréservation ou surbooking) soit ici $n > 300$.

Comme il y a évidemment un risque que le nombre de passagers munis d'un billet se présentant à l'embarquement excède 300, la compagnie veut maîtriser ce risque.

1. Déterminer la loi de X_n.
2. On suppose que $0{,}5 \leq p \leq 0{,}95$. Écrire l'intervalle de fluctuation asymptotique I_n de $\frac{X_n}{n}$ au seuil de 0,95.
3. Montrer que si $I_n \subset \left[0, \frac{300}{n}\right]$ alors la probabilité que le nombre de passagers se présentant à l'embarquement excède 300 est proche de 0,05.
4. On cherche à déterminer la valeur de n maximale permettant de satisfaire la condition de l'inclusion $I_n \subset \left[0, \frac{300}{n}\right]$.

   a. Montrer que $I_n \subset \left[0, \frac{300}{n}\right] \Rightarrow pn + 1{,}96\sqrt{n}\sqrt{p(1-p)} - 300 \leq 0$.

   b. On pose $f(x) = px + 1{,}96\sqrt{x}\sqrt{p(1-p)} - 300$.

   Montrer qu'il existe un entier n_0 unique tel que si $n \leq n_0$ alors $f(n) \leq 0$ et si $n > n_0$ alors $f(n) > 0$.

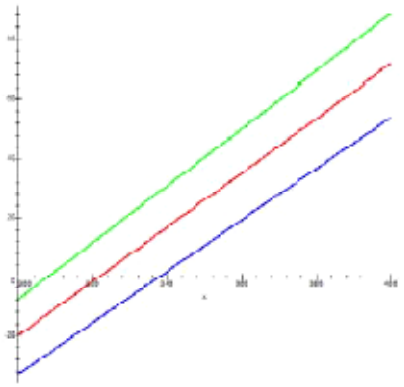
   c. Tracer la courbe représentative de f pour les valeurs $p = 0{,}85$; $p = 0{,}9$; $p = 0{,}95$.

   d. Déterminer à la calculatrice les valeurs de n_0 pour $p = 0{,}85$; $p = 0{,}9$; $p = 0{,}95$.

Corrigé du document « ressources pour la classe terminale générale et technologique »

1. X_n suit une loi binomiale de paramètres n et p.
2. Comme $n > 300$ et $0,5 \leq p \leq 0,95$ on a $np \geq 5$ et $n(1-p) \geq 5$ on peut utiliser l'intervalle de fluctuation asymptotique au seuil de 0,95 :
$$I_n = \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right].$$
3. Si $I_n \subset \left[0, \frac{300}{n}\right]$ alors $P(X_n > 300) \leq P\left(\frac{X_n}{n} \notin I_n\right)$.
Comme $P\left(\frac{X_n}{n} \notin I_n\right) \approx 0,05$ alors on peut dire que $P(X_n > 300)$ est proche également de 0,05 voire inférieur (l'événement $(X_n > 300)$ étant inclus dans la partie droite du complémentaire de I_n on pourrait vérifier avec le tableur que sa probabilité est en fait inférieure à 0,05 pour $n \geq 300$ et $0,5 \leq p \leq 0,95$).

4. a. $I_n \subset \left[0, \frac{300}{n}\right] \Rightarrow p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{300}{n} \Rightarrow np + 1,96\sqrt{n}\sqrt{p(1-p)} - 300 \leq 0$
b. En posant $y = \sqrt{x}$, on se ramène à une inéquation du second degré que l'on résout pour $x \geq 300$.
Les solutions de l'inéquation $f(x) \leq 0$ sont donc les réels de l'intervalle $[300, x_0]$ où
$$x_0 = \left(\frac{-1,96\sqrt{p(1-p)} + \sqrt{1200p + 1,96^2 p(1-p)}}{2p} \right)^2.$$
L'entier n_0 cherché est la partie entière de x_0.
c. $p = 0,85$ en bleu, $p = 0,9$ en rouge, $p = 0,95$ en vert.



d. Pour $p = 0,85$ on trouve $n_0 = 337$.
Pour $p = 0,9$ on trouve $n_0 = 321$,
Pour $p = 0,95$ on trouve $n_0 = 307$

Commentaires

Population : les acheteurs de billet
Caractère étudié : « se présenter à l'embarquement »

1. On suppose qu'il y a n acheteurs, $n > 300$, et que leur comportement est indépendant. Chaque acheteur a une probabilité p de se présenter à l'embarquement ; on appellera succès le fait qu'un acheteur se présente. X_n est donc le nombre de succès sur n épreuves de Bernoulli indépendantes. Par conséquent, X_n suit la loi binomiale de paramètres n et p.

2. $0{,}05 \leqslant p \leqslant 0{,}95$ et $n > 300$, donc $n > 300 \geqslant 30$, $np > 150 \geqslant 5$ et $n(1-p) > 12 \geqslant 5$ et donc on peut approximer la loi binomiale par la loi normale de paramètres $\mu = np$ et $\sigma^2 = np(1-p)$

Ou encore la loi de $\dfrac{X_n}{n}$ par la loi normale de paramètres $\mu = p$ et $\sigma^2 = \dfrac{p(1-p)}{n}$

Donc $I_n = \left[p - 1{,}96\sqrt{\dfrac{p(1-p)}{n}} ; p + 1{,}96\sqrt{\dfrac{p(1-p)}{n}} \right]$ car $P\left(\dfrac{X_n}{n} \in I_n\right) \simeq 0{,}95$

3. On suppose que $I_n \subset \left[0 ; \dfrac{300}{n}\right]$, alors $P\left(\dfrac{X_n}{n} \in \left[0 ; \dfrac{300}{n}\right]\right) \geqslant P\left(\dfrac{X_n}{n} \in I_n\right)$

$$\Rightarrow P\left(0 \leqslant \frac{X_n}{n} \leqslant \frac{300}{n}\right) \geqslant 0{,}95$$

$$\Rightarrow P\left(\frac{X_n}{n} > \frac{300}{n}\right) = 1 - P\left(0 \leqslant \frac{X_n}{n} \leqslant \frac{300}{n}\right) \leqslant 1 - 0{,}95$$

$$\Rightarrow P(X_n > 300) \leqslant 0{,}05$$

A ce stade, on sait que la probabilité est majorée par 0,05, mais on ne sait pas si elle est proche de 0,05.
Il faut donc remplacer, dans l'énoncé, « proche de 0,05 » par « majorée par 0,05 ».

4. On cherche n_0 tel que $n \geqslant n_0$ implique $I_n \subset \left[0 ; \dfrac{300}{n}\right]$.

4.a. Attention. On n'a qu'une implication simple mais pas une équivalence, la borne inférieure de I_n pouvant être négative.

$$I_n \subset \left[0 ; \frac{300}{n}\right] \Rightarrow p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \leqslant \frac{300}{n} \Rightarrow np + 1{,}96\sqrt{np(1-p)} \leqslant 300$$

4.b. En classe de Terminale, il semble plus « naturel » d'étudier la fonction proposée (variation de f puis corollaire du théorème des valeurs intermédiaires). Si x_0 est la racine réelle de f, alors $n_0 = E(x_0)$.

$$f(x) = px + 1{,}96\sqrt{p(1-p)}\sqrt{x} - 300 \quad , \quad f'(x) = p + 1{,}96\sqrt{p(1-p)}\frac{2}{\sqrt{x}}$$

4.c. Courbes : voir corrigé du document ressource

4.d. Cet énoncé ne propose pas de conclure (La problématique était : « la compagnie veut maitriser ce risque »).

De plus, avoir $n \geqslant n_0$ n'implique pas que $I_n \subset \left[0\,;\frac{300}{n}\right]$ (car il n'y a pas d'équivalence (cf. question 4a))

Pour avoir l'équivalence, il faut :

$$I_n \subset \left[0\,;\frac{300}{n}\right] \Leftrightarrow \begin{cases} p-1{,}96\sqrt{\dfrac{p(1-p)}{n}} \geqslant 0 \quad (1) \\ p+1{,}96\sqrt{\dfrac{p(1-p)}{n}} \leqslant \dfrac{300}{n} \quad (2) \end{cases}$$

- $(2) \Leftrightarrow n \geqslant n_0 \Rightarrow f(n) \geqslant 0$
- $(1) \Leftrightarrow p-1{,}96\sqrt{\dfrac{p(1-p)}{n}} \geqslant 0 \Leftrightarrow 1{,}96\sqrt{\dfrac{p(1-p)}{n}} \leqslant p$

$$\Leftrightarrow \quad 1{,}96\sqrt{\frac{(1-p)}{p}} \leqslant n \quad \Leftrightarrow \quad n \geqslant 1{,}96^2\,\frac{(1-p)}{p} = n_0'$$

Donc $n \geqslant \max\left\{n_0\,;n_0'\right\} \Leftrightarrow I_n \subset \left[0\,;\frac{300}{n}\right]$

Remarque

$p \longrightarrow \dfrac{(1-p)}{p} = \dfrac{1}{p} - 1$ est clairement strictement décroissante sur $[0{,}05\,;\,0{,}95]$,

Donc $\dfrac{(1-p)}{p} \leqslant \dfrac{(1-0{,}05)}{0{,}05} = 19 \qquad \forall p \in [0{,}05\,;\,0{,}95]$

ainsi $1 \leqslant n_0' \leqslant 75$, comme $n > 300$, implicitement, la seconde implication est vraie.

(Modification possible : on prend $\left]-\infty\,;\frac{300}{n}\right]$ au lieu de $\left[0\,;\frac{300}{n}\right]$:

$I_n \subset \left]-\infty\,;\frac{300}{n}\right]$ implique $P\left(\frac{X_n}{n} \in I_n\right) \leqslant P\left(\frac{X_n}{n} \leqslant \frac{300}{n}\right) \Leftrightarrow P(X_n > 300) \leqslant 0{,}05$

Mais en fait, pour cette situation, on ne doit pas utiliser l'intervalle I_n :

On veut *n* tel que $P(X_n > 300) \leqslant 0{,}05$.

Il suffit de centrer et réduire la variable X_n :

$$P(X_n > 300) \leqslant 0{,}05 \Leftrightarrow P(X_n \leqslant 300) \geqslant 0{,}95 \Leftrightarrow P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leqslant \frac{300-np}{\sqrt{np(1-p)}}\right) \geqslant 0{,}95$$

or $P(Z \leqslant a) \geqslant 0{,}95 \Leftrightarrow a = 1{,}645$

donc $\dfrac{300-np}{\sqrt{np(1-p)}} \geqslant 1{,}645 \Leftrightarrow P\left(\dfrac{X_n - np}{\sqrt{np(1-p)}} \leqslant \dfrac{300-np}{\sqrt{np(1-p)}}\right) \geqslant 0{,}95$

La résolution est plus simple, plus rapide, et fournit un exemple où il est nécessaire de centrer et réduire.

Reformulation possible de l'énoncé

Une compagnie aérienne possède des A340 (longs courriers) d'une capacité de 300 places.
Cette compagnie a vendu n billets pour le vol 2012.
La probabilité pour qu'un acheteur se présente à l'embarquement est p et les comportements des acheteurs sont indépendants les uns des autres.
On note X_n la variable aléatoire désignant le nombre d'acheteurs d'un billet se présentant à l'embarquement.
La compagnie cherche à optimiser le remplissage de l'avion en vendant éventuellement plus de places que la capacité totale de l'avion (surréservation ou surbooking) soit ici $n > 300$.
Comme il y a évidemment un risque que le nombre de passagers munis d'un billet se présentant à l'embarquement excède 300, la compagnie veut maîtriser ce risque.

1. Déterminer la loi de X_n.

2. On suppose que $0{,}05 \leqslant p \leqslant 0{,}95$. Justifier que l'on peut approximer la loi de X_n par une loi normale dont on précisera les paramètres.

3. **a.** Si Z suit la loi normale $\mathcal{N}(0, 1)$, déterminer le réel a tel que $P(Z \leqslant a) = 0{,}95$.

b. L'objectif de la compagnie est de majorer la probabilité de surréservation par 0,05.
Montrer que ceci équivaut à $\dfrac{300 - np}{\sqrt{np(1-p)}} \geqslant a$.

c. f est définie sur $\mathbb{R}_+$ par $f(x) = px + a\sqrt{p(1-p)}\sqrt{x} - 300$.
Résoudre, dans $\mathbb{R}_+$, l'équation $f(x) = 0$. On pourra poser $X = \sqrt{x}$.

d. Montrer qu'il existe un entier n_0 unique tel que si $n \leqslant n_0$ alors $f(n) \leqslant 0$ et si $n > n_0$ alors $f(n) > 0$.

e. Calculer pour $p = 0{,}85$, puis $p = 0{,}90$ et $p = 0{,}95$, et conclure dans chaque cas.

Remarque

On trouve $n_0 = 340$ pour $p = 0{,}85$
$n_0 = 330$ pour $p = 0{,}90$
$n_0 = 315$ pour $p = 0{,}95$

Exercice 6 Comment éviter l'insincérité dans un sondage délicat

Thèmes abordés

- **Probabilités conditionnelles (arbre pondéré)**
- **Intervalle de confiance**

Énoncé

L'une des difficultés pour un institut de sondage est le manque de sincérité dans les réponses des sondés. (Imaginez qu'un sondeur vous demande si vous avez déjà volé dans un magasin : que répondrez-vous ?)
Dans les années qui ont suivi l'engagement de l'armée américaine au Vietnam (1961-1975), les autorités américaines ont voulu évaluer par un sondage, parmi les soldats ayant combattu au Vietnam, la proportion de ceux qui avaient consommé de la drogue.
Voici une procédure *réelle* utilisée par les enquêteurs, et reproductible dans des situations analogues.
Le soldat tire au hasard une carte parmi trois :
Sur la première est écrit :
"Avez-vous, lors de votre déploiement, consommé une drogue illégale ?"
Sur la seconde est dessiné un triangle noir et écrit :
"Voyez-vous sur cette carte un triangle noir ?"
Sur la troisième est seulement écrit (il n'y a pas de triangle) :
"Voyez-vous sur cette carte un triangle noir ?"
Ayant tiré une carte, le soldat répond par oui ou par non à la question écrite.
Le sondeur ne sait pas quelle carte a été tirée et ignore donc à quelle question répond le soldat. Ainsi, on peut espérer que le soldat réponde franchement. On le supposera pour la suite.
L'enquêteur interroge 1200 soldats et 560 répondent "oui".

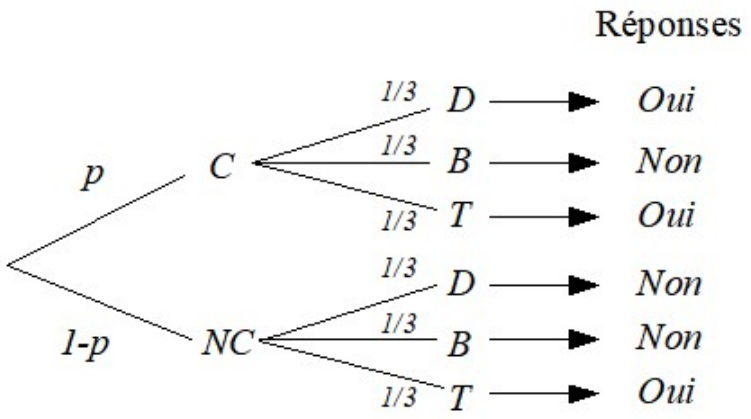
1. En déduire une estimation de la proportion p de soldats ayant consommé de la drogue.
Aide : On pourra construire un arbre pondéré utilisant les événements suivants :
C : "le soldat a consommé de la drogue",
D : "tirer la carte portant la question sur la drogue",
B : "tirer la carte sans triangle",
T : "tirer la carte portant le triangle noir",
O : "le soldat répond oui", et on notera $\pi = P(O)$, la probabilité d'obtenir « oui ».

2. Calculer un intervalle de confiance pour π, et en déduire un intervalle de confiance pour p.

3. Prolongements :
On peut aussi imaginer que le soldat tire une carte parmi 30 cartes dont la répartition est à définir, laissant croire au soldat qu'il y a 10 cartes de chaque sorte. On pourra envisager plusieurs répartitions et l'impact que la proportion des différentes cartes a sur l'estimation de p.
On note x la proportion de cartes "avez-vous consommé de la drogue ?" et y celle de triangles noirs, le reste étant constitué de cartes sans triangle :
a. Exprimer la probabilité π d'une réponse « Oui » en fonction de p, x et y.
b. Déterminer un intervalle de confiance pour p et calculer son l'amplitude.
c. Pour une valeur de p donnée, quelles sont les valeurs de x et y qui donnent l'intervalle de confiance de p le plus étroit ?

Corrigé

1. Soit p la proportion de soldats ayant combattu au Vietnam et qui ont consommé de la drogue. 1200 soldats sont tirés au hasard dans l'ensemble de ceux qui ont combattu au Vietnam. On assimile ce tirage à un tirage équiprobable avec remise.
Si on tire un soldat au hasard, il a une probabilité p d'avoir consommé de la drogue et $1-p$ de n'en avoir pas consommé. Ensuite, ce soldat tire une carte au hasard équiprobable et répond à la question qui figure dessus. Voici l'arbre correspondant :



$O=(C\cap D)\cup(C\cap T)\cup(\overline{C}\cap T)$, donc d'après la formule des probabilités totales :

$$\pi=P(O)=P(C\cap D)+P(C\cap T)+P(\overline{C}\cap T)=p\times\frac{1}{3}+p\times\frac{1}{3}+(1-p)\times\frac{1}{3}.$$

Soit : $\pi=\frac{1}{3}(1+p)$, ainsi $p=3\pi-1$.

Si on dépasse le programme, on peut écrire :
De la réalisation observée de l'échantillon aléatoire, on déduit :
une estimation ponctuelle de π : $\hat{\pi}=\frac{560}{1200}=\frac{7}{15}$,
d'où l'on tire une estimation ponctuelle de p : $\hat{p}=3\hat{\pi}-1=\frac{7}{5}-1=\frac{2}{5}$.

2. Mais une estimation rigoureuse implique la fourniture d'un intervalle de confiance, ce dernier donnant une information sur la précision de l'estimation. C'est l'objet de la question suivante.
L'intervalle de confiance à 0,95 pour π est donné par la formule :

$$\left[\hat{\pi}-1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\,;\hat{\pi}+1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

(après réalisation de l'expérience, on obtient l'intervalle $[0{,}4385;0{,}4949]$)

Or, $\pi\xrightarrow{g}3\pi-1$ est une fonction strictement croissante sur $\mathbb{R}$ On obtient un intervalle de confiance à 0,95 pour p en prenant l'image de celui pour π par la fonction g, soit:

$$\left[3\left(\hat{\pi}-1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)-1;3\left(\hat{\pi}+1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)-1\right]=\left[\hat{p}-3\times1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\,;\hat{p}+3\times1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Et après réalisation de l'expérience, on obtient l'intervalle $[0{,}315\,;0{,}485]$:

La justification rigoureuse de ce calcul est la suivante :

Rappelons que l'échantillon étant au départ aléatoire, $\hat{\pi}$**est une variable aléatoire**, et l'intervalle de confiance de niveau 0,95, souvent noté IC, est un intervalle aléatoire (puisque calculé à partir des observations qui sont aléatoires).

Il est défini par : $P\left(\hat{\pi}-1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\leqslant\pi\leqslant\hat{\pi}+1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)\cong0{,}95.$

Comme $x\longrightarrow3x-1$ est une fonction strictement croissante sur $\mathbb{R}$, elle conserve l'ordre, ainsi que sa réciproque.

Donc $\hat{\pi}-1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\leq\pi\leq\hat{\pi}+1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$ équivaut à :

$$3\left(\hat{\pi}-1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)-1\leq3\pi-1\leq3\left(\hat{\pi}+1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)-1$$

ainsi : $P\left(3\left(\hat{\pi}-1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)-1\leqslant3\pi-1\leqslant3\left(\hat{\pi}+1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)-1\right)\cong0{,}95.$

Soit : $P\left(\hat{p}-3\times1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\leqslant p\leqslant\hat{p}+3\times1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right)\cong0{,}95.$

L'intervalle de confiance de p au niveau 0,95 est alors :

$$\left[\hat{p}-3\times1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\,;\hat{p}+3\times1{,}96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

Revenons dans le cadre du programme :

Rappelons que dans les programmes du lycée, on appelle « intervalle de confiance » une réalisation de l'IC. Mais en toute rigueur, on ne devrait pas. En effet, seul un intervalle **aléatoire** peut contenir la vraie valeur du paramètre, qui est **fixe**, avec une probabilité de 0,95. Un intervalle fixe, lui, la contient ou ne la contient pas.

Pour utiliser une métaphore culinaire, un intervalle de confiance est comme une recette de soufflé : cette recette peut garantir que le soufflé lèvera dans 95% des cas, mais un soufflé particulier, lui, a levé ou non.

Pour faire comprendre cette idée, l'enseignant devrait appliquer la même formule d'IC à plusieurs échantillons simulés à partir d'une valeur connue de p, et montrer que contrairement à p, qui est fixe, l'intervalle varie, et que certains intervalles obtenus (réalisations particulières de l'IC) contiennent p tandis que d'autres, non.

Avec le programme de STI2D/STL :

L'intervalle de confiance de π au niveau 0,95, puisque $n\geqslant30$, $np\geqslant5$ et $nq\geqslant5$, est

$$\left[f-\frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\,;\,f+\frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)}\right]=\left[\frac{7}{15}-\frac{1{,}96}{\sqrt{1200}}\sqrt{\frac{7}{15}\times\frac{8}{15}}\,;\,\frac{7}{15}+\frac{1{,}96}{\sqrt{1200}}\sqrt{\frac{7}{15}\times\frac{8}{15}}\right]$$

$$=\left[\frac{7}{15}-0{,}0282;\,\frac{7}{15}+0{,}0282\right]=[0{,}4385;0{,}4949]$$

Et, **« intuitivement »**, on déduit l'intervalle de confiance de p par transformation affine croissante (difficile à justifier dans ces classes) :

$$[3\times0{,}4385-1\,;3\times0{,}4949-1]=[0{,}3155;\,0{,}4847].$$

Note :

Le fait que pour *n* assez grand, $P\left(p\in\left[F_n-1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\;;\;F_n+1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\right]\right)\cong0{,}95$

permette de définir l'intervalle $\left[F_n-1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\;;\;F_n+1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\right]$ comme l'IC de *p* au niveau de confiance 0,95, n'est pas au programme.

Avec le programme de S/ES

L'intervalle de confiance est $\left[f-\frac{1}{\sqrt{n}}\;;\;f+\frac{1}{\sqrt{n}}\right]$, donc les choses deviennent plus simples.

On notera encore F_n la variable aléatoire qui à chaque échantillon associe la fréquence des oui.

Comme : $F_n-\frac{1}{\sqrt{n}}\leqslant\pi\leqslant F_n+\frac{1}{\sqrt{n}}\Leftrightarrow\pi-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant\pi+\frac{1}{\sqrt{n}}$,

alors : $P\left(F_n-\frac{1}{\sqrt{n}}\leqslant\pi\leqslant F_n+\frac{1}{\sqrt{n}}\right)=P\left(\pi-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant\pi+\frac{1}{\sqrt{n}}\right)$,

$$=P\left(F_n-\frac{1}{\sqrt{n}}\leqslant\pi\leqslant F_n+\frac{1}{\sqrt{n}}\right)\simeq0{,}95.$$

On a : $P\left(\pi\in\left[F_n-\frac{1}{\sqrt{n}};F_n+\frac{1}{\sqrt{n}}\right]\right)\simeq0{,}95$, et on en déduit que l'intervalle de confiance de π au niveau 0,95 est :

$$\left[f-\frac{1}{\sqrt{n}}\,;\,f+\frac{1}{\sqrt{n}}\right]=\left[\frac{7}{15}-\frac{1}{\sqrt{1200}}\,;\,\frac{7}{15}+\frac{1}{\sqrt{1200}}\right]=\left[\frac{7}{15}-0{,}0289;\,\frac{7}{15}+0{,}0289\right]=[0{,}4378;\,0{,}4956].$$

Pour *p*, comme : $P\left(F_n-\frac{1}{\sqrt{n}}\leqslant\pi\leqslant F_n+\frac{1}{\sqrt{n}}\right)\simeq0{,}95$,

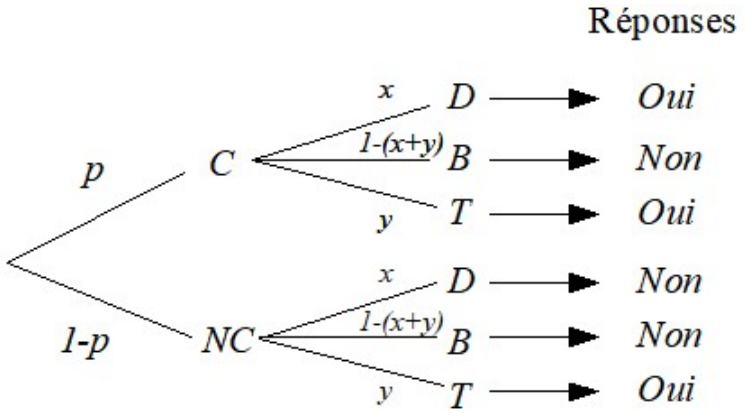
$$P\left(F_n-\frac{1}{\sqrt{n}}\leqslant\pi\leqslant F_n+\frac{1}{\sqrt{n}}\right)=P\left(3F_n-\frac{3}{\sqrt{n}}-1\leqslant3\pi-1\leqslant3F_n+\frac{3}{\sqrt{n}}-1\right)\simeq0{,}95.$$

Et $P\left((3F_n-1)-\frac{3}{\sqrt{n}}\leqslant p\leqslant(3F_n-1)+\frac{3}{\sqrt{n}}\right)\simeq0{,}95.$

L'intervalle de confiance de p est $\left[\frac{2}{5}-3\times0{,}0289\,;\,\frac{2}{5}+3\times0{,}0289\right]=[0{,}3133\,;\,0{,}4867]$.

Entre 31,33% et 48,67% des soldats se seraient drogués.

3. a. Voici l'arbre :



D'après la formule des probabilités totales

$\pi = p(x+y)+(1-p)y$ soit $\pi = px+y$ et par suite $p = \dfrac{\pi - y}{x}$ et $\hat{p} = \dfrac{\hat{\pi} - y}{x}$.

b. L'amplitude est : $2\times\dfrac{1{,}96}{x}\sqrt{\dfrac{f(1-f)}{1200}}$ pour STI/STL, ou $2\times\dfrac{1}{x}\times\sqrt{\dfrac{1}{1200}}$ pour S/ES.

Remarques générales sur l'intervalle de confiance

- $P\left(F_n - \dfrac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \dfrac{1}{\sqrt{n}}\right) = P\left(p \in \left[F_n - \dfrac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \dfrac{1}{\sqrt{n}}\right]\right) \simeq 0{,}95$,

signifie que si on pouvait faire une infinité de sondages, 0,95 est la proportion d'intervalles contenant la valeur p qu'on obtiendrait.

- On peut aussi interpréter « l'intervalle de confiance » comme suit :

$\left[f - \dfrac{1}{\sqrt{n}}\,;\, f + \dfrac{1}{\sqrt{n}}\right] = \left\{ p \in [0;1] \text{ tels que } f \in \left[p - \dfrac{1}{\sqrt{n}}\,;\, p + \dfrac{1}{\sqrt{n}}\right]\right\}$:

c'est l'ensemble des valeurs de p telles que la fréquence observée f appartienne à l'intervalle de fluctuation (simplifié) au seuil 0,95.

- En termes de « prise de décision » (c'est-à-dire de test) :

Comme $p \in \left[f - \dfrac{1}{\sqrt{n}}\,;\, f + \dfrac{1}{\sqrt{n}}\right] \Rightarrow f \in \left[p - \dfrac{1}{\sqrt{n}}\,;\, p + \dfrac{1}{\sqrt{n}}\right]$: l'intervalle de confiance est l'ensemble des valeurs p_0 telles que f n'infirmerait pas l'hypothèse $p = p_0$, avec un risque de 0,05. (On ne rejetterait pas l'hypothèse).

3.a. On rappelle que : $p = \dfrac{\pi - y}{x}$ donc $\hat{p} = \dfrac{\hat{\pi} - y}{x}$, et

$$P\left(\hat{p} - \frac{1{,}96}{x}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leqslant p \leqslant \hat{p} + \frac{1{,}96}{x}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) \cong 0{,}95.$$

b. L'amplitude de l'intervalle est proportionnelle à $\dfrac{1{,}96}{x}\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}$,

soit $\frac{1{,}96}{x}\sqrt{\frac{(\hat{p}x+y)(1-\hat{p}x-y)}{n}}$, donc dépend de $f(x,y)=\left(\hat{p}+\frac{y}{x}\right)\left(\frac{1-y}{x}-\hat{p}\right)$

$f'_x(x,y)=-\frac{1}{x^2}\left(y\left(\frac{1-y}{x}-\hat{p}\right)+(1-y)\left(\hat{p}+\frac{y}{x}\right)\right)$

$\hat{\pi}\geqslant 0 \Rightarrow \hat{p}+\frac{y}{x}\geqslant 0$, $1-\hat{\pi}\geqslant 0 \Rightarrow 1-\hat{p}x-y\geqslant 0 \Rightarrow \frac{1-\hat{p}x-y}{x}\geqslant 0 \Rightarrow \frac{1-y}{x}-\hat{p}\geqslant 0$

Donc $f'_x(x,y)\leqslant 0$

L'amplitude de l'IC est donc minimale pour $x=1$(donc $y=0$). Dans ce cas, $\pi = p$ et $\hat{\pi}=\hat{p}$.
On remarque qu'on n'a alors plus d'aléa supplémentaire car toutes les cartes sont équivalentes. On comprend donc que la précision soit la meilleure. Évidemment, il ne faut pas montrer au soldat que toutes les cartes sont des cartes demandant s'il s'est drogué.
Supposons donc qu'on fixe un *x* assez proche de 1 mais différent de 1. Ainsi, l'enquêteur ne **sait** pas à quelle question le soldat répond. On se demande alors quelle serait la meilleure répartition des cartes demandant si l'on voit un triangle noir, entre celles qui en contiennent un et celles qui n'en contiennent pas.
Autrement dit, quelle est la meilleure valeur de *y* entre 0 et $1-x$.

$f'_y(x,y)=\frac{1}{x^2}\left(1-2\left(y+x\hat{p}\right)\right)=\frac{1}{x^2}(1-2\hat{\pi})$

On suppose *x* proche de 1, donc *y* proche de 0,
si $x\hat{p}>0{,}5$, $f'_y(x,y)<0$, donc l'intervalle de confiance le plus précis sera obtenu pour *y* maximum, donc $y=1-x$.
si $x\hat{p}<0{,}5$, $f'_y(x,y)>0$, donc l'intervalle de confiance le plus précis sera obtenu pour *y* minimum, donc $y=0$.
Là non plus, en pratique, n'avoir aucune carte T n'est pas envisageable car une réponse Oui ne peut alors correspondre qu'à la question "vous êtes-vous drogué ?" mais *x* peut être maintenu très petit.

Exercice 7 Élections présidentielles

Exercice tel qu'il peut être posé et résolu dans le supérieur et une version adaptée au lycée

Thèmes abordés

- **Loi normale**
- **Intervalle de confiance**
- **Prise de décision**

Énoncé

La veille du 2ème tour d'une élection présidentielle opposants les candidats A et B, on effectue un sondage sur $n = 900$ personnes. Les résultats sont les suivants :

Candidat A : $X = 425$ voix

Candidat B : $Y = 475$ voix.

On note p_A et $p_B \in [0;1]$ les scores respectifs de A et B au second tour, et $\Delta = p_A - p_B$ l'écart entre les deux scores.

1 - a) On appellera estimateur ponctuel $\hat{p}_A$ de p_A la proportion empirique de votants pour A dans l'échantillon.
Calculer $\hat{p}_A$ en fonction de X, et donner sa valeur observée.
Donner la loi approchée de X et en déduire celle de $\hat{p}_A$.

b) En déduire l'estimateur $\hat{\Delta}$ de Δ, ainsi que sa loi approchée.

2 - a) Donner un intervalle de confiance de niveau 0.95 pour Δ.

b) Au vu des résultats, peut-on significativement (au risque 5%) anticiper une victoire de B ?

Corrigé

1 - a) $\hat{p}_A = \frac{X}{n}$. A.N. $\hat{p}_A = 0.4722$

Loi approchée de X: $\dfrac{X - np_A}{\sqrt{p_A(1-p_A)}} \sim N(0;1) \Rightarrow \dfrac{n\left(\dfrac{X}{n} - p_A\right)}{\sqrt{p_A(1-p_A)}} \sim N(0;1)$

$\sqrt{n}\dfrac{X - np_A}{\sqrt{p_A(1-p_A)}} \sim N(0;1)$, soit $\hat{P}_A \sim N\left(p_A; \dfrac{p_A(1-p_A)}{n}\right)$

1 - b) $\Delta = p_A - p_B = p_A - (1-p_A) = 2p_A - 1$ donc $\hat{\Delta} = 2\hat{p}_A - 1$

et par conséquent (transformation affine d'une variable normale:

$E(aX+b) = aE(X) + b$ et $V(aX+b) = a^2V(X)$

Donc $X \sim N(\mu, \sigma^2) \Rightarrow aX + b \sim N(a\mu + b, a^2\sigma^2)$

$$2\hat{p}_A - 1 \sim N\left(2p_A - 1\,;\, 4\frac{p_A(1-p_A)}{n}\right),\ \text{soit}\ \ \hat{\Delta} \sim N\left(\Delta;\, 4\frac{p_A(1-p_A)}{n}\right)$$

2 - a) $\hat{\Delta} \sim N\left(\Delta;\, 4\frac{p_A(1-p_A)}{n}\right) \Rightarrow \sqrt{n}\frac{\hat{\Delta}-\Delta}{2\sqrt{p_A(1-p_A)}} \sim N(0\,;\,1)$

Par conséquent : $P\left(-1{,}96 \leqslant \sqrt{n}\frac{\hat{\Delta}-\Delta}{2\sqrt{p_A(1-p_A)}} \leqslant 1{,}96\right) = 0{,}95$, ce qui donne :

$$P\left(-\hat{\Delta} - 1{,}96\frac{2\sqrt{p_A(1-p_A)}}{\sqrt{n}} \leqslant -\Delta \leqslant -\hat{\Delta} + 1{,}96\frac{2\sqrt{p_A(1-p_A)}}{\sqrt{n}}\right) = 0{,}95$$

Soit $P\left(\hat{\Delta} - 1{,}96\frac{2\sqrt{p_A(1-p_A)}}{\sqrt{n}} \leqslant \Delta \leqslant \hat{\Delta} + 1{,}96\frac{2\sqrt{p_A(1-p_A)}}{\sqrt{n}}\right) = 0{,}95$

Application numérique : IC observé est $[-0.1208; 0.0097]$.

2 - b) Pour savoir si ce que l'on observe (B est devant A) est statistiquement significatif, on regarde si l'on peut maintenir l'hypothèse contraire, ou si on doit la rejeter (i.e. anticiper la victoire de B) au risque 5%. DONC, il s'agit de tester : $H_0 : p_B \leq p_A$ contre $H_1 : p_B > p_A$.
Or $H_0 : p_B \leq p_A$ s'écrit aussi $H_0 : \Delta \geq 0$, et $H_1 : p_B > p_A$ s'écrit $H_1 : \Delta < 0$.
C'est donc le test de positivité de la moyenne d'une loi normale. On rejette H_0 en faveur de H_1 au risque 0.05 de le faire à tort, lorsque $\sqrt{n}\frac{\hat{\Delta}-0}{2\sqrt{p_A(1-p_A)}} < -1{,}645$

Application numérique : la statistique de test vaut $-1.67 < -1.645$

Commentaire : On anticipe donc la victoire de B, mais de justesse !

Refaisons le calcul avec $X = 426$ et $Y = 474$:

IC observé: $[-0.118; 0.012]$; la statistique de test vaut $-1.60 > -1.645$, donc on n'anticipe plus la victoire de B de façon significative !

Énoncé possible en lycée (passé sous les fourches caudines du programme)

La veille du 2ème tour d'une élection présidentielle opposant les candidats A et B, on effectue un sondage sur *n* personnes.
X est la variable aléatoire correspondant au nombre d'intentions de vote obtenu par le candidat A.
Y est la variable aléatoire correspondant au nombre d'intentions de vote obtenu par le candidat B.
F_A et F_B sont respectivement les variables aléatoires $\frac{X}{n}$ et $\frac{Y}{n}$.
p_A et p_B sont les pourcentages obtenus par les candidats à l'issue du scrutin et on pose $\delta = p_A - p_B$.

1. On suppose que le résultat des élections est conforme au sondage,
   a. Par quelle loi peut-on approcher celle de F_A ? et celle de X ?

b. Par quelle loi peut-on approcher celle de $\Delta = F_A - F_B$?

2. Le sondage a porté sur 900 personnes, et les résultats sont les suivants :
Candidat A : 425 voix
Candidat B : 475 voix

a. Donner un intervalle de confiance au niveau 0,95 pour p_A , donner sa signification ; et donner un intervalle de confiance au niveau 0,95 pour δ .

b. Peut-on anticiper une victoire de *B* au risque de 5% ?

Corrigé

1.

a. D'après le **théorème de Moivre-Laplace** :

Si $(X_n)_n$ est une suite de variables aléatoires indépendantes suivant la même loi de Bernoulli de paramètre p , donc d'espérance p et de variance $\sigma^2 = pq$,
alors si $F_n = \dfrac{X_1 + X_2 + ... + X_n}{n}$
$Z_n = \dfrac{F_n - E(F_n)}{\sigma(F_n)} = \dfrac{F_n - p}{\sqrt{pq/n}}$ ***converge en loi*** vers Z où Z suit la loi normale $N(0\,;1)$

Si $n \geqslant 30$, $np \geqslant 5$ et $nq \geqslant 5$, $\dfrac{F_A - p_A}{\sqrt{p_A q_A / n}}$ suit approximativement la loi $N(0\,;1)$

Alors, F_A suit approximativement la loi $N\left(\mu = p_A ; \sigma^2 = p_A q_A / n\right)$

Et $X = nF_A$ suit approximativement la loi $N\left(\mu = np_A ; \sigma^2 = np_A q_A\right)$

b. $\Delta = F_A - F_B = F_A - (1 - F_A) = 2F_A - 1 = 2p_A - 1$, donc

$V(\Delta) = V(2F_A - 1) = V(2F_A)$ (la variance est invariante par translation de la variable)

$$V(\Delta) = 2^2 V(2F_A) = 4\frac{p_A q_A}{n}$$

Et Δ suit approximativement la loi $N\left(\mu = 2p_A - 1 ; \sigma^2 = \dfrac{2p_A q_A}{n}\right)$

2. **a.** $\left[\dfrac{17}{36} - \dfrac{1{,}96}{36}\sqrt{\dfrac{17\times 19}{900}} \,;\, \dfrac{17}{36} + \dfrac{1{,}96}{36}\sqrt{\dfrac{17\times 19}{900}}\right] = [0{,}439\,;\,0{,}505]$

C'est une estimation par intervalle de p_A « au niveau » de confiance 0,95.
(même s'il n'est pas rare, mais cela est faux, de lire par exemple que $p_A \in [0{,}439\,;\,0{,}505]$ avec une probabilité de 95%)

$p_A \in [0{,}439\,;\,0{,}505] \Leftrightarrow \delta \in [2\times 0{,}439 - 1\,;\, 2\times 0{,}505 - 1] = [-0{,}122\,;\,0{,}01]$

b. Non, si on s'en tient au programme. Dans celui-ci, on peut remettre en cause une valeur de p, au risque de 5% , mais pas le fait que p appartienne à un intervalle :
Ici il faudrait pouvoir rejeter l'hypothèse $p_A \geqslant 0{,}5$ ou $\delta \geqslant 0$ **,** donc faire un test unilatéral.

Remarque et rappels

- Dans le supérieur, on appelle intervalle de confiance au niveau 0,95 l'intervalle $\left[F_n - 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\ ;\ F_n + 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\right]$.

  L'intervalle $\left[f - 1{,}96\sqrt{\frac{f(1-f)}{n}}\ ;\ f + 1{,}96\sqrt{\frac{f(1-f)}{n}}\right]$ est une réalisation de ce dernier.

- Une région de confiance du paramètre inconnu *p* au niveau $1-\alpha$ est un ensemble $R \subset [0;1]$ construit par rapport à une observation (X_n) tel que

  $\forall p \in [0;1]$, $P(p \in R) = 1-\alpha$. **(**Au lycée, $\alpha = 0,05$ **)**

  On utilise le théorème de Moivre-Laplace pour obtenir une région de confiance (asymptotique) qui ne vaudra donc que pour *n* assez grand (dans la pratique acceptable pour $n \geqslant 30$ **),** en fait un intervalle et de plus centré sur l'espérance, car pour une loi symétrique et pour une probabilité donnée, c'est l'intervalle le plus étroit.

  Donc F_n ***suit approximativement*** la loi normale $N\left(\mu = p; \sigma = \sqrt{pq/n}\right)$

  Et on sait que $P\left(p \in \left[F_n - 1{,}96\sqrt{pq/n}\ ;\ F_n + 1{,}96\sqrt{pq/n}\right]\right) \simeq 0{,}95$

  Mais on ne connait pas *p*, donc pas σ, on a deux façons de pallier cette difficulté :

  - On peut majorer $p(1-p)$ par $\frac{1}{4}$ sur $[0;1]$

    Ainsi $P\left(p \in \left[F_n - \frac{1}{\sqrt{n}}\ ;\ F_n + \frac{1}{\sqrt{n}}\right]\right) \geqslant P\left(p \in \left[F_n - \frac{1{,}96}{2}\frac{1}{\sqrt{n}}\ ;\ F_n + \frac{1{,}96}{2}\frac{1}{\sqrt{n}}\right]\right) \geqslant 0{,}95$.

    Autre approche : Il est évident que

    $P\left(F_n \in \left[p - \frac{1}{\sqrt{n}}\ ;\ p + \frac{1}{\sqrt{n}}\right]\right) \simeq 0{,}95 \Leftrightarrow P\left(p \in \left[F_n - \frac{1}{\sqrt{n}}\ ;\ F_n + \frac{1}{\sqrt{n}}\right]\right) \simeq 0{,}95$

  - Alternative plus précise : on utilise un estimateur de σ^2 : $\hat{\sigma}^2 = F_n(1-F_n)$.

    On montre que la loi de $\frac{F_n - p}{\sqrt{\hat{\sigma}^2/n}}$ converge en loi vers *X* où *X* suit la loi $N(0\ ;1)$

    Ainsi $P\left(\left[F_n - 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\ ;\ F_n + 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}\right]\right) \simeq 0.95$

N.B.

Il n'est pas évident que $P\left(F_n \in \left[p - 1{,}96\sqrt{pq/n}\ ;\ p + 1{,}96\sqrt{pq/n}\right]\right) \simeq 0{,}95 \Rightarrow$

$P\left(p \in \left[F_n - 1{,}96\sqrt{F_n(1-F_n)/n}\ ;\ F_n + 1{,}96\sqrt{F_n(1-F_n)/n}\right]\right) \simeq 0.95$.

Exercice 8 Les LED ne s'usent pas

Thèmes abordés

- **Loi exponentielle**
- **Fonction de répartition**
- **Loi conditionnelle**
- **Intervalle de confiance**

Énoncé

Une LED a une vie X suivant une loi exponentielle de paramètre λ .
1. À une date $t \geq 0$ donnée, on constate que la LED fonctionne toujours. On voudrait déterminer la loi de la durée de vie qu'il lui reste à "vivre". Déterminer pour cela la loi de $X-t$ conditionnellement à $X>t$.
2. À un échantillon aléatoire indépendant de n LED identiques on associe la variable aléatoire $\overline{X}_n$, la durée de vie moyenne des n LED.

On rappelle que $E(X)=\frac{1}{\lambda}$, $V(X)=\frac{1}{\lambda^2}$, et on admet que la variable $\frac{\overline{X}-E\left(\overline{X}\right)}{\sigma\left(\overline{X}\right)}$ suit approximativement la loi normale centrée-réduite (théorème central limite)

a. Montrer que la variable $Z=\sqrt{n}\left(\lambda\overline{X}-1\right)$ suit approximativement une loi normale centrée-réduite pour n assez grand.
b. Proposer un intervalle de fluctuation de probabilité 0,95 de Z.

En remplaçant Z par son expression en fonction de $\overline{X}$ et de λ dans l'intervalle précédent, déduire un intervalle de confiance approché de niveau 0,95 pour λ.
En déduire un intervalle de confiance de même niveau pour la durée de vie moyenne théorique de la LED.
3. Application numérique : Un échantillon indépendant de 49 LED identiques a fonctionné jusqu'à extinction. La moyenne empirique de leurs durées de vie observées a été de 34567 heures.
Une LED a déjà fonctionné 2000 heures. Calculez l'intervalle de confiance de niveau 0,95 pour la durée qu'il lui reste à fonctionner.

Corrigé

1. On pose $Y=X-t$

$$\forall y \geqslant 0, P_{(T\geqslant t)}(Y \geqslant y)=\frac{P\left((Y \geqslant y)\cap(X \geqslant t)\right)}{P(X \geqslant t)}$$

$$=\frac{P\left((X \geqslant y+t)\cap(X \geqslant t)\right)}{P(X \geqslant t)}=\frac{P(X \geqslant y+t)}{P(X \geqslant t)}=\frac{e^{-\lambda(t+y)}}{e^{-\lambda t}}=e^{-\lambda y}$$

Donc, conditionnellement au fait que la LED ait survécu jusqu'à t, la durée de la vie qui lui reste suit toujours la même loi exponentielle : cette LED ne s'use pas.

N.B. Mais Y ne suit pas une loi exponentielle, la densité de cette loi est $y \longrightarrow \lambda e^{\lambda(t+y)}$ sur $\mathbb{R}_+$ et 0 ailleurs.

2. a. $E\left(\overline{X}\right) = E\left(\frac{X_1 + ... + X_n}{n}\right) = \frac{1}{n}\left(E\left(X_1\right) + ... + E\left(X_n\right)\right) = \frac{1}{n} \times n \times \frac{1}{\lambda} = \frac{1}{\lambda}$

$V\left(\overline{X}\right) = V\left(\frac{X_1 + ... + X_n}{n}\right) = \frac{1}{n^2}\left(V\left(X_1\right) + ... + V\left(X_n\right)\right) = \frac{1}{n^2} \times n \times \frac{1}{\lambda^2} = \frac{1}{n\lambda^2}$

Car les variables X_i sont indépendantes.

$$\frac{\overline{X} - E\left(\overline{X}\right)}{\sigma\left(\overline{X}\right)} = \frac{\overline{X} - \frac{1}{\lambda}}{\frac{1}{\sqrt{n}\lambda}} = \sqrt{n}\lambda\left(\overline{X} - \frac{1}{\lambda}\right)$$

donc $Z = \sqrt{n}\left(\lambda\overline{X} - 1\right)$ suit approximativement la loi $N\left(0;1\right)$.

b. On en déduit un intervalle de fluctuation de $\overline{X}$ au seuil de 95% :
On sait que $P\left(-1,96 \leqslant Z \leqslant 1,96\right) \simeq 0,95$

$$P\left(-1,96 \leqslant \sqrt{n}\left(\lambda\overline{X} - 1\right) \leqslant 1,96\right) \simeq 0,95 \Leftrightarrow \quad P\left(1 - \frac{1,96}{\sqrt{n}} \leqslant \lambda\overline{X} \leqslant 1 + \frac{1,96}{\sqrt{n}}\right) \simeq 0,95$$

$$P\left(\frac{1}{\lambda}\left(1 - \frac{1,96}{\sqrt{n}}\right) \leqslant \overline{X} \leqslant \frac{1}{\lambda}\left(1 + \frac{1,96}{\sqrt{n}}\right)\right) \simeq 0,95$$

D'où l'intervalle de fluctuation de $\overline{X}$: $\left[\frac{1}{\lambda}\left(1 - \frac{1,96}{\sqrt{n}}\right) ; \frac{1}{\lambda}\left(1 + \frac{1,96}{\sqrt{n}}\right)\right]$

De plus $P\left(1 - \frac{1,96}{\sqrt{n}} \leqslant \lambda\overline{X} \leqslant 1 + \frac{1,96}{\sqrt{n}}\right) \simeq 0,95 \Leftrightarrow \quad P\left(\frac{1}{\overline{X}}\left(1 - \frac{1,96}{\sqrt{n}}\right) \leqslant \lambda \leqslant \frac{1}{\overline{X}}\left(1 + \frac{1,96}{\sqrt{n}}\right)\right) \simeq 0,95$

D'où l'intervalle de confiance de λ au niveau de confiance 0,95 pour un échantillon de moyenne observée $\overline{x}$: $\left[\frac{1}{\overline{x}}\left(1 - \frac{1,96}{\sqrt{n}}\right) \leqslant \lambda \leqslant \frac{1}{\overline{x}}\left(1 + \frac{1,96}{\sqrt{n}}\right)\right]$

La durée de vie moyenne est l'espérance, c'est-à-dire $\frac{1}{\lambda}$

Ainsi $\left[\frac{\overline{x}\sqrt{n}}{\sqrt{n} + 1,96} \leqslant \frac{1}{\lambda} \leqslant \frac{\overline{x}\sqrt{n}}{\sqrt{n} - 1,96}\right]$ est un intervalle de confiance de $\frac{1}{\lambda}$ au niveau de confiance 0,95.

3. Application numérique : $\left[\frac{\overline{x}\sqrt{n}}{\sqrt{n} + 1,96} \leqslant \frac{1}{\lambda} \leqslant \frac{\overline{x}\sqrt{n}}{\sqrt{n} - 1,96}\right] = \left[27005 ; 48010\right]$

Compléments sur les lois de durée de vie sans vieillissement en annexe

Exercice 9 Approximations

D'après Baccalauréat STL spécialité Biotechnologies, Métropole, juin 2017

Thèmes abordés

- **Lois binomiale et normale, approximation de l'une par l'autre.**
- **Intervalles de fluctuation.**
- **Prise de décision.**

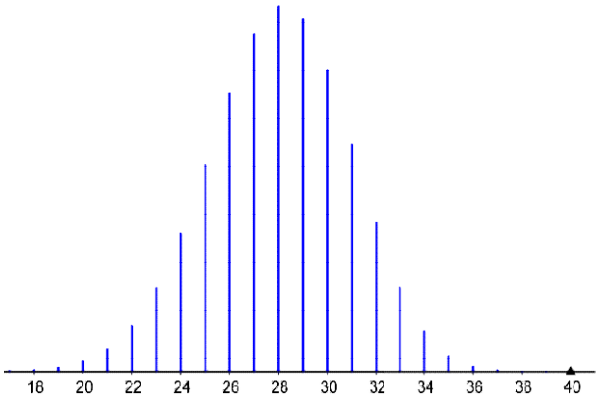
Énoncé

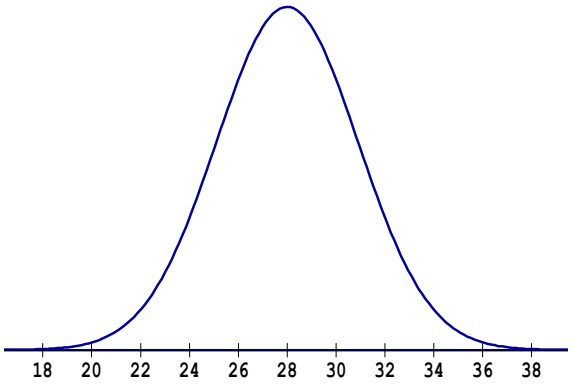
PARTIE A

En 2010, la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES) affirme qu'en France, 7 adultes sur 10 portent des lunettes.
On prélève au hasard un échantillon de 40 adultes parmi la population française. On assimile ce prélèvement à un tirage avec remise.
Soit X la variable aléatoire qui, à tout échantillon de ce type, associe le nombre de porteurs de lunettes dans l'échantillon.

1. Montrer que X suit une loi binomiale dont on précisera les paramètres.

2. Les figures ci-dessous présentent un diagramme en bâtons et une courbe.





a. À quoi correspond l'axe des abscisses commun à ces deux représentations ?
b. À laquelle des deux représentations est associée la loi binomiale ? Pourquoi ? Dire à quoi correspond son axe des ordonnées.

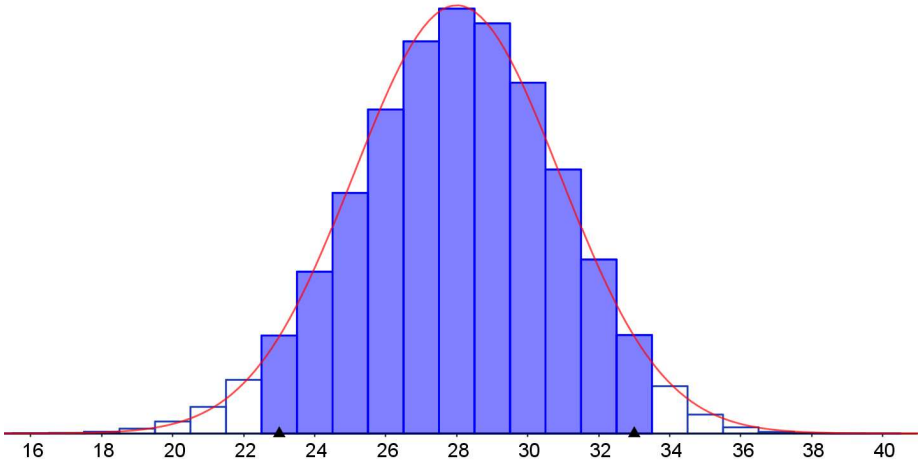
3. L'autre graphique représente la densité f de la loi normale de même espérance μ et de même écart type σ que la loi binomiale.
a. Donner, par lecture graphique, la valeur de μ. Justifier.
b. Dire à quoi correspond son axe des ordonnées.
c. Serait-il correct d'affirmer que l'écart type σ de la loi normale est égal à 8 ? Justifier votre réponse.

4. On considère la variable aléatoire W qui suit la loi de probabilité définie par la densité g, g étant la fonction en escalier définie sur $\mathbb{R}$ par :

$g(x) = P(X = k)$ si $x \in \left[k - \frac{1}{2}; k + \frac{1}{2}\right[$ pour $k \in [\![0; 40]\!]$

$g(x) = 0$ sinon.

On a représenté ci-dessous la fonction *g,* ainsi que la densité de la loi normale définie ci-avant.



a. Que représente l'aire de la partie colorée pour la variable W ? pour la variable X ?

b. (1) Calculer $p_1 = P(23 \leqslant X \leqslant 33)$.

(2) On voudrait vérifier que l'on peut approcher la loi binomiale de X par la loi normale de même espérance et de même écart-type. Soit Y une v.a. suivant cette loi normale.

- Calculer $p_2 = P(23 \leqslant Y \leqslant 33)$ et $p_3 = P(22{,}5 \leqslant Y \leqslant 33{,}5)$.
- Laquelle des 2 approxime le mieux p_1 ?
- Pouvait-on prévoir ce résultat à partir du graphique ?

c. Afin d'obtenir des approximations moins grossières, on décide d'augmenter la taille des échantillons, en prenant $n = 400$ et encore $p = 0{,}7$.

(1) Combien valent à présent μ et σ ?

(2) Calculer $P(263 \leqslant X \leqslant 297)$, $P(263 \leqslant Y \leqslant 297)$ et $P(263{,}5 \leqslant Y \leqslant 297{,}5)$.

d. Et on fait de même pour $n = 4000$:

Comparer $P(2745 \leqslant X \leqslant 2855)$, $P(2745 \leqslant Y \leqslant 2855)$ et $P(2744{,}5 \leqslant Y \leqslant 2855{,}5)$.

e. Commenter.

PARTIE B

Dans cette partie, on considère des échantillons de taille $n = 40$.

1. Déterminer un intervalle de fluctuation à 94.43% de la fréquence des porteurs de lunettes dans un échantillon de taille 40. Justifier.

2. **a.** Y suivant la loi normale de la question **4.**, si Z est la variable centrée réduite de $Y/40$, déterminer a tel que $P(-a \leqslant Z \leqslant a) = 0{,}9443$.

b. Si on considère que le nombre de porteurs de lunettes dans un échantillon de taille 40 suit la loi de Y, déduire un intervalle de fluctuation à 94,43% de $Y/40$.

c. Comparer cet intervalle à celui de la question précédente.

3. On rappelle que, en 2010, la DREES affirme que 7 adultes sur 10 portent des lunettes. Un opticien se demande si, en 2018, cette proportion est toujours la même. Dans un échantillon aléatoire de 40 adultes, il compte 24 porteurs de lunettes. Cette observation remet-elle en cause l'hypothèse que la proportion n'a pas changé ?

Corrigé

Partie A

1. Comme on assimile ce prélèvement à un tirage avec remise (la population étant très grande), on peut considérer que les tirages sont indépendants, on a donc un schéma de Bernoulli : on répète 40 épreuves identiques et indépendantes à deux issues :

- succès : « porter des lunettes » avec $p = P(succès) = 0{,}7$,
- et échec : « n'en point porter » avec $q = P(echec) = 0{,}3$.

Donc la v.a. X qui compte le nombre de succès suit la loi binomiale $B(40;0{,}7)$.

2. a. L'axe des abscisses correspond aux valeurs prises par la v.a. X, c'est-à-dire le nombre de porteurs de lunettes parmi les 40 sondés.

b. La variable aléatoire X est à valeurs entières, elle est donc discrète ;
sa loi binomiale est discrète : elle est associée au diagramme en bâtons.
L'axe des ordonnées correspond aux probabilités des valeurs de X, $P(X=k)$ pour $k \in \{0,\ldots,40\}$.

3. a. La courbe C est symétrique par rapport à la droite d'équation $x = \mu$ et par lecture graphique on constate que $\mu = 28$.

b. L'axe des ordonnées correspond aux valeurs prises par la fonction de densité f de la loi normale, et non pas à des probabilités : ces dernières correspondent à des aires délimitées par la courbe.

c. $\sigma = \sigma(X) = \sqrt{npq} = \sqrt{40 \times 0{,}7 \times 0{,}3} \simeq 2{,}898$ L'affirmation est donc erronée.
On peut aussi remarquer que si Y suit une loi normale de paramètres μ et σ,
$P(\mu - \sigma \leqslant Y \leqslant \mu + \sigma) \simeq 0{,}683$
Ainsi on aurait $P(20 \leqslant Y \leqslant 36) \simeq 0{,}683$, ce qui est manifestement faux sachant que l'aire totale sous la courbe est égale à 1.

4. Par définition de g, on constate que :
$\int_{k-\frac{1}{2}}^{k+\frac{1}{2}} g(x)\,dx = P\left(k-\frac{1}{2} \leqslant W \leqslant k+\frac{1}{2}\right) = P(X=k)$ et a et b étant deux entiers tels que $a \leqslant b$:

$$\int_{a-\frac{1}{2}}^{b+\frac{1}{2}} g(x)\,dx = \sum_{k=a}^{k=b} \int_{a-\frac{1}{2}}^{b+\frac{1}{2}} g(x)\,dx = \sum_{k=a}^{k=b} P(X=k) = P(a \leqslant X \leqslant b)$$

Ainsi l'aire colorée est $\int_{23-\frac{1}{2}}^{33+\frac{1}{2}} g(x)\,dx = P\left(23-\frac{1}{2} \leqslant W \leqslant 33+\frac{1}{2}\right)$

et aussi $\int_{23-\frac{1}{2}}^{33+\frac{1}{2}} g(x)\,dx = \sum_{k=23}^{33} P(X=k) = P(23 \leqslant X \leqslant 33)$.

b. (1) Avec la calculatrice :

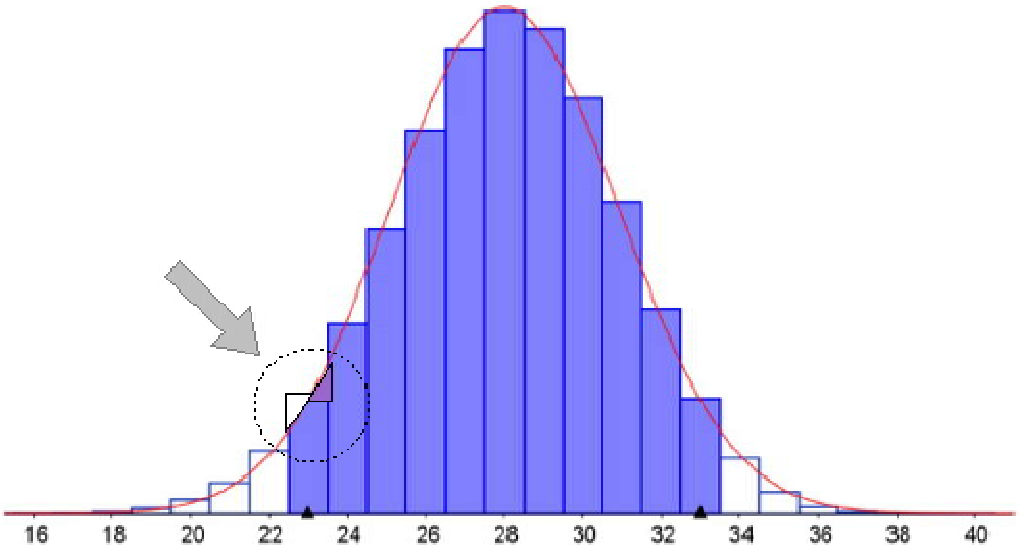
$p_1 = P(23 \leqslant X \leqslant 33) = P(X \leqslant 33) - P(X \leqslant 22) \simeq 0,9443$ (calcul effectué avec la loi binomiale $B(40;0,7)$, c'est-à-dire la vraie loi.)

(2) $p_2 = P(23 \leqslant Y \leqslant 33) \simeq 0,9155$ et $p_3 = P(22,5 \leqslant Y \leqslant 33,5) \simeq 0,9423$

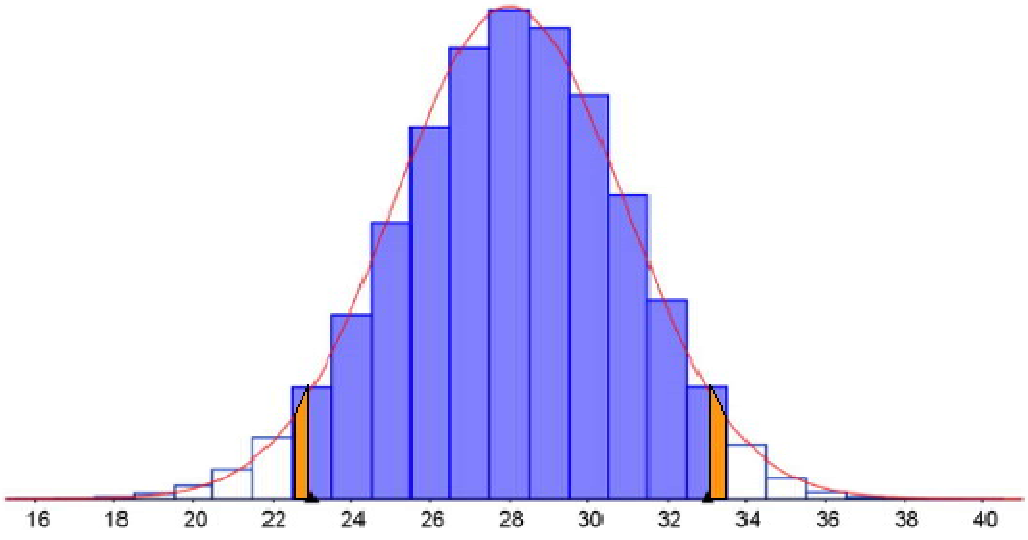
(calcul effectué avec la loi normale $N(\mu = 28;\sigma = 2,898)$)

p_3 est plus proche de p_1 que p_2 : $|p_1 - p_3| < |p_1 - p_2|$,

L'approximation par p_3 est meilleure. Graphiquement, cela peut se voir. En effet, regardons l'aire $_1p$: on constate qu'elle est très proche de l'aire sous la courbe de f qui est égale à p_3 (on observera la compensation des petites aires triangulaires).



D'autre part :



$$p_1 - p_2 = \int_{23-\frac{1}{2}}^{23} f(x)\,dx + \int_{33}^{33+\frac{1}{2}} f(x)\,dx$$

Cette différence représente l'aire des deux zones orange, qui ne sont pas négligeables.
Il est donc plus judicieux de considérer que :

$$\int_{23-\frac{1}{2}}^{33+\frac{1}{2}} f(x)dx = P(23 \leqslant X \leqslant 33)$$

C'est ce que l'on appelle « correction de continuité »

c. (1) $\mu = E(X) = np = 400 \times 0,7 = 280$.

$\sigma = \sigma(X) = \sqrt{npq} = \sqrt{400 \times 0,7 \times 0,3} \simeq 9,165$.

(2) et avec la calculatrice réglementaire :

$P(263 \leqslant X \leqslant 297) = P(X \leqslant 297) - P(X \leqslant 262) \simeq 0,9440$

$P(263 \leqslant Y \leqslant 297) \simeq 0,9364$

$P(262,5 \leqslant Y \leqslant 296,5) \simeq 0,9438$

d. $\mu = E(X) = np = 4000 \times 0,7 = 2800$.

$\sigma = \sigma(X) = \sqrt{npq} = \sqrt{4000 \times 0,7 \times 0,3} \simeq 28,983$

$P(2745 \leqslant X \leqslant 2855) = 0,9445$

$P(2745 \leqslant Y \leqslant 2855) = 0,9423$

$P(2744,5 \leqslant Y \leqslant 2855,5) = 0,9445$.

e. On observe que l'approximation est d'autant meilleure que *n* est grand, et que la pertinence de la correction de continuité demeure y compris pour des grandes valeurs de *n*.

Partie B

1. La v.a. X qui compte le nombre de porteur de lunettes suit la loi binomiale $B(40;0,7)$.

Et on a vu que $P(23 \leqslant X \leqslant 33) \simeq 0,9443$ donc $P\left(\frac{23}{40} \leqslant \frac{X}{40} \leqslant \frac{33}{40}\right) \simeq 0,9443$

Puisque $23 \leqslant X \leqslant 33 \Leftrightarrow \frac{23}{40} \leqslant \frac{X}{40} \leqslant \frac{33}{40} \Leftrightarrow 0,575 \leqslant \frac{X}{40} \leqslant 0,825$

Donc la probabilité (avant réalisation) que la fréquence de l'échantillon appartienne à l'intervalle $[0,575\,;0,825]$ est 94,43%.

C'est un intervalle de fluctuation à 94,43%.

2. a. On résout $P(-a \leqslant Z \leqslant a) = 0,9443$ où Z suit la loi normale de paramètres $\mu = 0$ et $\sigma = 1$.

$P(-a \leqslant Z \leqslant a) = P(Z \leqslant a) - P(Z \leqslant -a) = P(Z \leqslant a) - (1 - P(Z \leqslant a)) = 2P(Z \leqslant a) - 1$

Donc $P(Z \leqslant a) = \frac{1 + 0,9443}{2} = 0,97215$, avec la calculatrice, $a = 1,9134$.

b. Ainsi $P(-1,9134 \leqslant Z \leqslant 1,9134) = 0,9443$

d'où $P\left(-1,9134 \leqslant \frac{\frac{Y}{40} - 0,7}{\frac{2,898}{40}} \leqslant 1,9134\right) = 0,9443$

$P\left(0,7 - 1,9134 \times \frac{2,898}{40} \leqslant \frac{Y}{40} \leqslant 0,7 + 1,9134 \times \frac{2,898}{40}\right) = 0,9443$

et enfin $P\left(0,5614 \leqslant \frac{Y}{40} \leqslant 0,8386\right) = 0,9443$.

L'intervalle de fluctuation de probabilité 0,9443 est $\left[0,5614\,;\,0,8386\right]$.

On obtient donc un intervalle à partir de la loi normale dont l'amplitude est augmentée de 2 × 0,014, ce qui peut être considéré comme négligeable.
Donc la probabilité (avant réalisation) que la fréquence de l'échantillon appartienne à $[0,5614;0,8386]$ est 94,43%.

c. La fréquence observée dans l'échantillon est $\frac{24}{40} = 0,6$.

Donc la fréquence observée appartient à l'intervalle de fluctuation dans chacun des deux cas (avec la loi exacte qui est la loi binomiale ou son approximation normale), et on ne remet pas en cause l'affirmation.

En effet, on voudrait savoir si on peut considérer que la proportion *p* de porteur de lunettes en 2018 n'est pas différente de celle de 2010 : $p = p_0 = 0,7$. C'est l'hypothèse notée H_0.

Si tel était le cas, la fréquence d'un échantillon aléatoire de taille 40 : $F = \frac{Y}{40}$, appartiendrait à l'intervalle de fluctuation $I_{0,9443}$ avec une probabilité égale à 0,9443.

Si f, la fréquence observée, n'appartient pas à l'intervalle de fluctuation, on considère peu plausible que $p = p_0$: on remet en cause l'hypothèse H_0.

Parmi tous les échantillons possibles, seuls 5,57% donnerait une fréquence en dehors de l'intervalle. Dans le cas contraire, on ne remet pas en cause H_0.

Remarques

1. Il s'agit ici d'un « test bilatéral », la proportion en 2018 peut être supérieure ou inférieure à celle de 2010. Pour un test unilatéral, par exemple : la proportion a-t-elle augmenté ? la démarche est un peu différente et d'ailleurs hors programme.

2. Une fois le tirage réalisé, l'expérience aléatoire est réalisée, il n'y a donc plus de probabilité.
Si *f,* la fréquence observée, n'appartient pas à l'intervalle de fluctuation, on dit que l'on rejette l'hypothèse « au risque 0,0557 » de se tromper, pour tenir compte du fait que : avant réalisation, la probabilité pour la variable aléatoire *F* d'être en dehors de l'intervalle de fluctuation est égale à 0,0557 sous l'hypothèse $p = p_0$.
Si *f,* la fréquence observée, appartient à l'intervalle de fluctuation, on ne rejette pas l'hypothèse, mais sans préciser un risque car on ne connait pas la probabilité pour la variable aléatoire *F* d'être dans l'intervalle de fluctuation si $p \neq p_0$.(On ne connait pas alors la valeur de *p*).
Ainsi, un test ne sert pas à valider une hypothèse, mais à la rejeter.

Exercice 10 Centrer et réduire

Thèmes abordés

- **Centrer et réduire une loi normale**
- **Intervalles de fluctuation.**
- **Prise de décision.**

Énoncé

Une entreprise fabrique des tablettes de chocolat de 100 grammes. Le service de contrôle qualité effectue plusieurs types de contrôle. **On donnera les résultats à 10^{-3} près.**

Partie A

Une tablette de chocolat doit peser 100 grammes avec une tolérance de deux grammes en plus ou en moins. Elle est donc mise sur le marché si sa masse est comprise entre 98 et 102 grammes. Quand on prélève une plaque, la masse (exprimée en grammes) de cette tablette de chocolat peut être modélisée par une variable aléatoire X suivant la loi normale d'espérance $\mu = 100$ et d'écart-type $\sigma = 1{,}2$.

1. **a.** Calculer la probabilité de l'évènement M : « la tablette est mise sur le marché ».
   b. Calculer $P(X \leqslant 98) + P(X \geqslant 102)$, interpréter ce résultat.
   c. Exprimer l'intervalle de fluctuation [97,6 ; 102,4] en fonction de l'espérance et de l'écart-type.
   En déduire la valeur de $P(97{,}6 \leqslant X \leqslant 102{,}4)$.
2. Le réglage des machines de la chaîne de fabrication permet de modifier la valeur de σ. On souhaite modifier le réglage des machines de telle sorte que la probabilité de l'évènement M soit égale à 0,96.
   a. Quelle loi suit la variable aléatoire $Z = \dfrac{X - 100}{\sigma}$?
   b. Déterminer a tel que $P(-a \leqslant Z \leqslant a) = 0{,}96$.
   c. Déterminer la valeur de $\boldsymbol{\sigma}$ pour que la probabilité de l'évènement « la tablette est mise sur le marché » soit égale à 0,96.

Partie B

1. En fait, une tablette est commercialisable si sa masse est supérieure à 99 grammes.
   La masse d'une tablette de chocolat est encore modélisée par une variable aléatoire X suivant la loi normale d'espérance $\mu = 100$ et d'écart-type $\sigma = 1{,}2$.
   Quelle est la probabilité qu'une plaquette soit commercialisable ?
2. Le réglage des machines de la chaîne de fabrication permet de modifier la valeur de μ. Déterminer la valeur de μ pour que la probabilité de l'évènement « la tablette est non commercialisable » soit égale à 0,04.

Partie C

1. Une première étude a montré que 96% de la production est commercialisable.
   On veut vérifier si le réglage de la machine a changé, pour cela on prélève 300 plaques dont on vérifie la masse. Et on constate que 280 plaques sont commercialisables.
   a. Peut-on utiliser ici un intervalle de fluctuation asymptotique ?
   b. Quel est l'intervalle de fluctuation asymptotique au risque $\alpha = 0{,}05$?
   c. Peut-on penser que la machine est déréglée au seuil 95% ?

2. Le mois suivant, on teste 1500 plaques, il y a 1440 plaques commercialisables.

a. Donner un intervalle permettant d'estimer, au niveau de confiance de 95%, la proportion de plaques commercialisables.

b. Peut-on considérer que la machine est réglée comme initialement ?

Corrigé

Partie A

L'expérience aléatoire est : prélever une tablette dans la production.

La variable aléatoire X associe à chaque tablette sa masse en grammes, et cette variable suit une loi normale.

1. a. Une tablette est mise sur le marché si sa masse est comprise entre 98 gr et 102 gr, X suit la loi normale d'espérance $\mu = 100$ et d'écart-type $\sigma = 1{,}2$, donc

$P(M) = P(98 \leqslant X \leqslant 102) \simeq 0{,}904$ à 10^{-3} près.

b. On obtient par un calcul direct :

$$P(X \leqslant 98) + P(X \geqslant 102) \simeq 0{,}096 .$$

On pouvait aussi calculer en utilisant le résultat de la question précédente :

$$P(X \leqslant 98) + P(X \geqslant 102) = P\left(X \notin]98;102[\right) = 1 - P\left(X \in]98;102[\right),$$

$P(X \leqslant 98) + P(X \geqslant 102) = 1 - P(98 < X < 102) = 1 - P(98 \leqslant X \leqslant 102)$ puisque la loi est continue.

Il s'agit donc de la probabilité qu'une tablette soit non commercialisable, c'est-à-dire $P\left(\overline{M}\right)$.

c. On a : $P(97{,}6 \leqslant X \leqslant 102{,}4) = P(\mu - 2\sigma \leqslant X \leqslant \mu + 2\sigma)$.

Donc $P(97{,}6 \leqslant X \leqslant 102{,}4) \simeq 0{,}954$

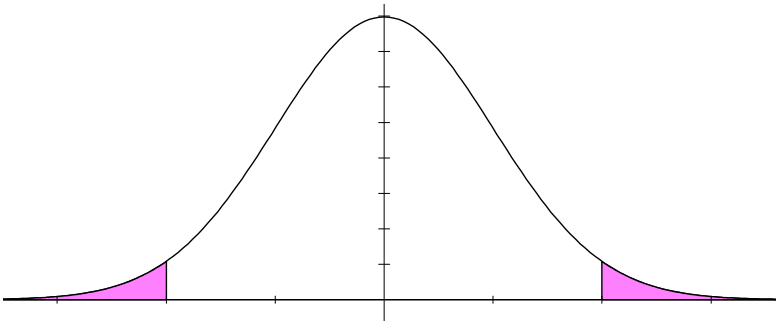
2. a. On sait que la variable aléatoire $Z = \dfrac{X - 100}{\sigma}$ suit la loi normale centrée réduite $N(0;1)$.

b. $P(-a \leqslant Z \leqslant a) = P(Z \leqslant a) - P(Z < -a) = P(Z \leqslant a) - P(Z \leqslant -a)$

$= P(Z \leqslant a) - P(Z \geqslant a)$ par symétrie

$= P(Z \leqslant a) - \left(1 - P(Z < a)\right)$ en passant à l'événement contraire

$= P(Z \leqslant a) - \left(1 - P(Z \leqslant a)\right) = 2P(Z \leqslant a) - 1$



Donc $P(-a \leqslant Z \leqslant a) = 0{,}96 \Rightarrow \quad 2P(Z \leqslant a) - 1 = 0{,}96 \Rightarrow \quad P(Z \leqslant a) = \dfrac{0{,}96 + 1}{2} = 0{,}98$.

Et avec la calculatrice, on obtient $a \simeq 2{,}054$.

c. On a $P(-2,054 \leqslant Z \leqslant 2,054) \simeq 0,96$ donc $P\left(-2,054 \leqslant \dfrac{X-100}{\sigma} \leqslant 2,054\right) \simeq 0,96$,

et $P(100-2,054\sigma \leqslant X \leqslant 100+2,054\sigma) = P(98 \leqslant X \leqslant 102) \simeq 0,96$

d'où $100+2,054\sigma=102$ soit $\sigma = \dfrac{2}{2,054} \simeq 0,974$.

Partie B

1. X suit la loi normale d'espérance $\mu = 100$ et d'écart-type $\sigma = 1,2$:

$P(X \geqslant 99) \simeq 0,798$

2. On veut déterminer μ tel que $P(X \geqslant 99) = 0,04$.

On doit se ramener, comme à la question **A.2.**, à la loi centrée réduite, c'est-à-dire considérer la variable aléatoire $Z = \dfrac{X-\mu}{1,2}$ qui suit la loi $N(0;1)$, **c'est-à-dire une loi normale dont on connait les paramètres.**

$P(X \geqslant 99) = 0,04 \Rightarrow \ P\left(\dfrac{X-\mu}{1,2} \geqslant \dfrac{99-\mu}{1,2}\right) = 0,04 \Rightarrow \ P\left(Z \geqslant \dfrac{99-\mu}{1,2}\right) = 0,04$.

Et on résout l'équation $P(Z \geqslant b) = 0,04$, ce qui revient à résoudre $P(Z \leqslant b) = 0,96$:

La calculatrice nous donne $b = 1,75068... \simeq 1,751$.

D'où $\dfrac{99-\mu}{1,2} = 1,751$, soit $\mu = 1,751 \times 1.2 + 99 = 101,101$.

Partie C

1. **a.** Ici, $n = 300$ et $p = 0,96$, donc $n \geqslant 30$, $np = 288 \geqslant 5$ et $nq = 12 \geqslant 5$.

On peut alors utiliser les intervalles de fluctuation asymptotique.

b. L'intervalle de fluctuation asymptotique au risque $\alpha = 0,05$ est

$$\left[p - 1,96\sqrt{pq/n}\,;\, p + 1,96\sqrt{pq/n}\right] = [0,938\,;0,982]$$

c. $f = \dfrac{280}{300} = 0,933...$ n'appartient pas à l'intervalle de fluctuation asymptotique.

On doit donc rejeter l'hypothèse que $p = 0,96$, considérant ainsi (au risque 5%) que la machine s'est déréglée.

2. **a.** Cette fois, il s'agit de l'intervalle de confiance $\left[f - \dfrac{1}{\sqrt{n}}\,;\, f + \dfrac{1}{\sqrt{n}}\right]$.

$f = \dfrac{1450}{1500} = 0,967...$ d'où l'intervalle $[0,941\,;\,0,993]$.

Si on utilise l'intervalle $\left[f - \dfrac{1,96}{\sqrt{n}}\sqrt{f(1-f)}\,;\, f - \dfrac{1,96}{\sqrt{n}}\sqrt{f(1-f)}\right]$,

on obtient $[0,958\,;0,976]$.

b. 0,96 est dans l'intervalle de confiance. On peut donc considérer au risque 5% que le réglage de la machine est conforme au réglage initial.

Compléments sur l'invariance de certaines familles de distributions par transformation affine en annexe

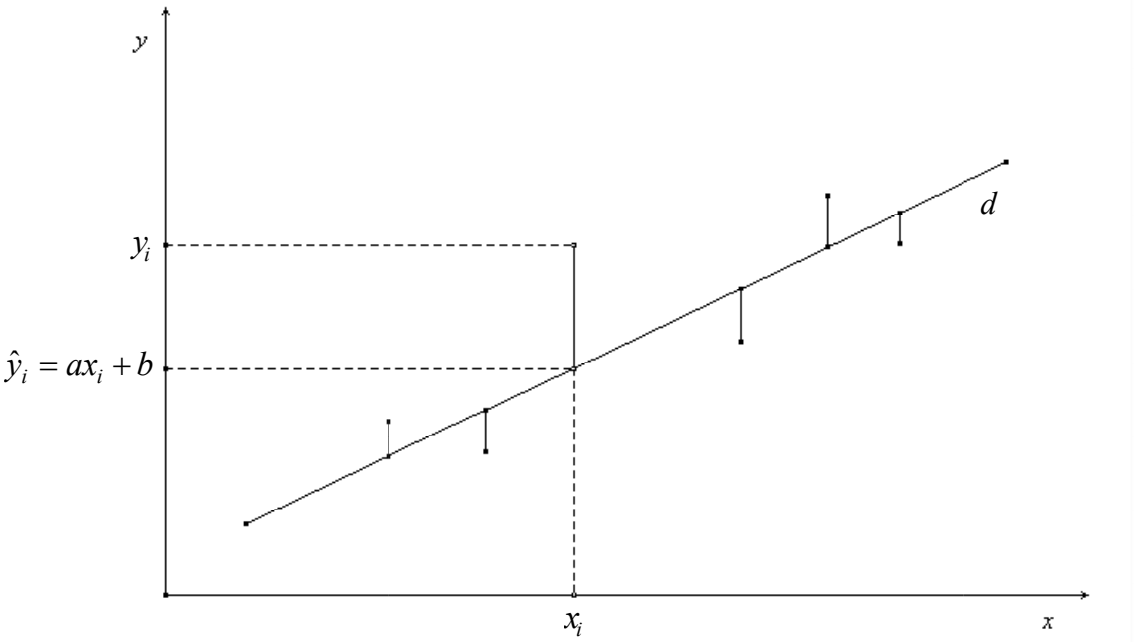
ANNEXES

Ajustements affines

Dans les programmes, il est demandé seulement aux élèves d'utiliser correctement l'outil numérique, de connaitre le nom de la méthode et une idée de celle-ci, puisque les commentaires précise : « on observe à l'aide d'un logiciel le caractère minimal de la somme des carrés des écarts ». En revanche, le coefficient de corrélation ne semble pas au programme.

Nous décrivons ici cette méthode, due à Gauss et Legendre.

Pour un ajustement de *y* en *x* :
On considère une série double $(x_i, y_i)_{i=1,\ldots,n}$, on suppose qu'il existe une liaison entre les grandeurs mesurées par les x_i et les y_i , représentée par une relation fonctionnelle simple *f,* affine dans le cas présent, définie par $f(x) = ax + b$.
Graphiquement :



d est la droite d'équation *y* = *ax* + *b*,
on choisit une « distance » entre le nuage de points et la droite *d* (ou la courbe de *f*)
en l'occurrence la somme des carrés des écarts : $S = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2$.

Soit $S = \sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2$, et on cherche les valeurs de *a* et *b* qui minimisent *S.*

- On regarde *S* comme un polynôme de la variable *b*

$$\sum_{i=1}^{n} \left(y_i - ax_i - b\right)^2 = nb^2 - 2b\sum_{i=1}^{p} \left(y_i - ax_i\right) + \sum_{i=1}^{p} \left(y_i - ax_i\right)^2$$

Or, si $\alpha > 0$, $\alpha x^2 + \beta x + \gamma$ est minimum pour $x = -\frac{\beta}{2\alpha}$.

Donc S est minimum pour $b=\frac{1}{n}\sum_{i=1}^{n}(y_i-ax_i)=\frac{1}{n}\sum_{i=1}^{n}y_i-a\frac{1}{n}\sum_{i=1}^{n}x_i$

on obtient : $b=\overline{y}-a\overline{x}$.

Ainsi, pour tout a : $y=ax+b=ax+\overline{y}-a\overline{x}$ soit $y-\overline{y}=a(x-\overline{x})$

et donc la droite passe par le point moyen $G(\overline{x},\overline{y})$.

- On regarde maintenant la somme obtenue comme un polynôme en a

$$S=\sum_{i=1}^{n}\left(y_i-\overline{y}+a\overline{x}-ax_i\right)^2=\sum_{i=1}^{n}\left(y_i-\overline{y}-a(x_i-\overline{x})\right)^2$$

$$S=a^2\sum_{i=1}^{n}(x_i-\overline{x})^2-2a\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+\sum_{i=1}^{n}(y_i-\overline{y})^2$$

S est minimum pour $a=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\dfrac{nCov(x,y)}{nV(x)}=\dfrac{Cov(x,y)}{V(x)}$

- $a=\dfrac{Cov(x,y)}{V(x)}$ et $b=\overline{y}-a\overline{x}$ (et $\overline{y}=a\overline{x}+b$)
- la droite d'équation $y = ax + b$ passe par le point moyen $G(\overline{x},\overline{y})$
- le minimum est $S=nV(y)-2n\dfrac{\left[Cov(x,y)\right]^2}{V(x)}+n\dfrac{\left[Cov(x,y)\right]^2}{V(x)}=nV(y)\left(1-\dfrac{\left[Cov(x,y)\right]^2}{V(x)V(y)}\right)$

  qui est la **somme quadratique des résidus**.
- La moyenne des écarts est nulle

  $\frac{1}{n}\sum_{i=1}^{n}(y_i-ax_i-b)=\frac{1}{n}\sum_{i=1}^{n}y_i-a\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}b=\overline{y}-a\overline{x}-b=0$
- La variance des écarts est donc égale à la moyenne quadratique des écarts

  $v_r=V(y)\left(1-\dfrac{\left[Cov(x,y)\right]^2}{V(x)V(y)}\right)$ appelée **variance résiduelle.**

La qualité de l'ajustement peut être mesurée par **somme quadratique des résidus**, ou encore par la **variance résiduelle** pour annuler l'effet de taille**.**

Elle est d'autant meilleure que v_r est proche de 0, donc que $\dfrac{\left(Cov(x,y)\right)^2}{V(x)V(y)}$ est proche de 1, ou encore que $\left|\dfrac{Cov(x,y)}{s(x)s(y)}\right|$ est proche de 1.

On peut remarquer que : $0\leqslant\dfrac{\left(Cov(x,y)\right)^2}{V(x)V(y)}\leqslant 1$.

$r=\dfrac{\mathrm{cov}(x,y)}{s(x)s(y)}$ est le **coefficient de corrélation linéaire**

$r^2 = \dfrac{\left(Cov(x,y)\right)^2}{V(x)V(y)}$ est le **coefficient de détermination.**

Remarques

Si on note $\hat{y}_i = ax_i + b$

$\overline{\hat{y}} = a\overline{x} + b = \overline{y}$

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2$$

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(ax_i + b - a\overline{x} - b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(ax_i - a\overline{x})$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = a\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i - a\overline{x}\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = a\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - \overline{y} + a\overline{x})x_i = \frac{1}{n}\sum_{i=1}^{n}y_i x_i - a\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{y}\frac{1}{n}\sum_{i=1}^{n}x_i + a\overline{x}\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = \frac{1}{n}\sum_{i=1}^{n}y_i x_i - a\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}\,\overline{y} + a\overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}y_i x_i - \overline{x}\,\overline{y} - a\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2\right)$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = Cov(x,y) - aV(x) = Cov(x,y) - \frac{Cov(x,y)}{V(x)}V(x) = 0 \qquad \text{ouf …}$$

Donc $\quad V(y) = \dfrac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = v_e + v_r$

$v_e = \dfrac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = V(\hat{y})$ est **la variance expliquée** par la variance des x_i.

$v_r = V(y)\left(1 - r^2\right)$ est la **variance résiduelle.**

On considère aussi le rapport $\dfrac{V(\hat{y})}{V(y)}$ qui mesure la part de la variance expliquée par la variable *x* dans la variance totale de la variable *y*, c'est le **coefficient de détermination** R^2 :

$$R^2 = \frac{V(\hat{y})}{V(y)} = \frac{\text{variance expliquée par } x}{\text{variance de } y}$$

$$R^2 = \frac{V(y) - v_r}{V(y)} = 1 - \frac{v_r}{V(y)} = 1 - \frac{\text{variance résiduelle}}{\text{variance de } y}$$

Donc dans le cas d'une corrélation linéaire, $R^2 = \dfrac{V(y) - v_r}{V(y)} = 1 - \dfrac{V(y)\left(1 - r^2\right)}{V(y)} = r^2$.

Intervalles de confiance

On note F_n la variable aléatoire qui à chaque échantillon associe la fréquence observée du caractère.
Comme d'après le Théorème de Moivre Laplace,

$$\lim_{n\to+\infty} P\left(p-1{,}96\sqrt{\frac{p(1-p)}{n}}\leqslant F_n\leqslant p+1{,}96\sqrt{\frac{p(1-p)}{n}}\right)=0{,}95,$$

on considère que $P\left(p-1{,}96\sqrt{\frac{p(1-p)}{n}}\leqslant F_n\leqslant p+1{,}96\sqrt{\frac{p(1-p)}{n}}\right)\simeq 0{,}95$ pour *n* assez grand.

Dans la pratique : si $n\geqslant 30$, $np\geqslant 5$ et $nq\geqslant 5$

Intervalle simplifié :

Comme $p(1-p)=-\left(p^2-p\right)=-\left(\left(p-\frac{1}{2}\right)^2-\frac{1}{4}\right)=-\left(p-\frac{1}{2}\right)^2+\frac{1}{4}\leqslant\frac{1}{4}\quad\forall p\in\mathbb{R}$,

$1{,}96\sqrt{p(1-p)}\leqslant 1{,}96\sqrt{\frac{1}{4}}\leqslant 1$ pour $\forall\, p\in]0\,;1[$,

$F_n\in\left[p-1{,}96\sqrt{\frac{p(1-p)}{n}},p+1{,}96\sqrt{\frac{p(1-p)}{n}}\right]\Rightarrow F_n\in\left[p-\frac{1}{\sqrt{n}},p+\frac{1}{\sqrt{n}}\right]$

Et $P\left(F_n\in\left[p-\frac{1}{\sqrt{n}},p+\frac{1}{\sqrt{n}}\right]\right)\geqslant 0{,}95$.

Et puisque $F_n\in\left[p-\frac{1}{\sqrt{n}},p+\frac{1}{\sqrt{n}}\right]\Leftrightarrow p\in\left[F_n-\frac{1}{\sqrt{n}},F_n+\frac{1}{\sqrt{n}}\right]$,

On considère que $P\left(p\in\left[F_n-\frac{1}{\sqrt{n}},F_n+\frac{1}{\sqrt{n}}\right]\right)\simeq 0{,}95$ pour *n* assez grand.

Dans la pratique : si $n\geqslant 25$, $0{,}2\leqslant p\leqslant 0{,}8$

L'intervalle de confiance simplifié à 95 % pour la probabilité *p* est $\left[F_n-\frac{1}{\sqrt{n}},F_n+\frac{1}{\sqrt{n}}\right]$.

Une réalisation de cet intervalle est : $I=\left[f-\frac{1}{\sqrt{n}}\,;\,f+\frac{1}{\sqrt{n}}\right]$.

Intervalle non simplifié :

$$p-1{,}96\sqrt{\frac{p(1-p)}{n}}\leqslant F_n\leqslant p+1{,}96\sqrt{\frac{p(1-p)}{n}}\Leftrightarrow F_n\in\left[p-1{,}96\sqrt{\frac{p(1-p)}{n}},p+1{,}96\sqrt{\frac{p(1-p)}{n}}\right]$$

Cependant

$$F_n \in \left[p - 1{,}96\sqrt{\frac{p(1-p)}{n}}, p + 1{,}96\sqrt{\frac{p(1-p)}{n}} \right] \not\Rightarrow p \in \left[F_n - 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}, p + 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}} \right]$$

Mais on peut montrer que $\lim_{n\to\infty} P\left(-1{,}96 \leqslant \sqrt{n}\frac{F_n - p}{\sqrt{F_n(1-F_n)}} \leqslant 1{,}96 \right) = 0{,}95$,

soit $\lim_{n\to\infty}\left(p \in \left[F_n - 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}, p + 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}} \right] \right) \simeq 0{,}95$.

On considère que $P\left(p \in \left[F_n - 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}, p + 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}} \right] \right) \simeq 0{,}95$ pour n assez grand…

$\left[F_n - 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}}, p + 1{,}96\sqrt{\frac{F_n(1-F_n)}{n}} \right]$ est l'intervalle de confiance du paramètre p au niveau de confiance 0,95.

Une réalisation de cet intervalle est : $I = \left[f - \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)} \; ; \; f + \frac{1{,}96}{\sqrt{n}}\sqrt{f(1-f)} \right]$.

Matrices stochastiques

Une matrice est dite stochastique si la somme des termes de chaque ligne est égale à 1.
On s'intéresse à l'existence de $\lim_{n\to\infty} A^n$, A étant une matrice stochastique.

En dimension 2

On considère les matrices $A=\begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}$ avec $0\leqslant a\leqslant 1$ et $0\leqslant b\leqslant 1$.

Le polynôme caractéristique est $\lambda^2-(a+b)\lambda-1+a+b$.

Les racines sont $\lambda_1=1$ ce qui était prévisible, et $\lambda_2=a+b-1$.

Donc la matrice est diagonalisable, évidemment même si $\lambda_2=1$.

Et $D=\begin{pmatrix} 1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$.

Si on exclut les cas limites $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$,

a ou b différents de 1 et a ou b différents de 0, $-1<\lambda_2<1$ et $a+b\neq 2$

et $\lim_{n\to\infty} D^n=\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$.

On choisit deux vecteurs propres :

$\begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}=\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et $\begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}\begin{pmatrix} 1-a \\ b-1 \end{pmatrix}=(a+b-1)\begin{pmatrix} 1-a \\ b-1 \end{pmatrix}$

(ici on regarde la matrice comme celle d'une application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^2$, remarquons par ailleurs que si une matrice et sa transposée ont les mêmes valeurs propres, les sous espaces propres sont en général différents).

La matrice de passage est $P=\begin{pmatrix} 1 & 1-a \\ 1 & b-1 \end{pmatrix}$, $P^{-1}=\dfrac{1}{a+b-2}\begin{pmatrix} b-1 & a-1 \\ -1 & 1 \end{pmatrix}$

$\lim_{n\to\infty} A^n=\lim_{n\to\infty} PD^nP^{-1}=\dfrac{1}{a+b-2}\begin{pmatrix} b-1 & a-1 \\ b-1 & a-1 \end{pmatrix}$

Et $\forall (x,y)\in\mathbb{R}^2$ tel que $x+y=1$, $\begin{pmatrix} x & y \end{pmatrix}\left(\dfrac{1}{a+b-2}\begin{pmatrix} b-1 & a-1 \\ b-1 & a-1 \end{pmatrix}\right)=\begin{pmatrix} \dfrac{b-1}{a+b-2} & \dfrac{a-1}{a+b-2} \end{pmatrix}$

La limite est un état stable quelque soit l'état initial $\begin{pmatrix} x_0 & y_0 \end{pmatrix}$ avec $x_0+y_0=1$.

Remarque 1

On a aussi : $A^n=PD^nP^{-1}=\dfrac{1}{a+b-2}\begin{pmatrix} (b-1)-\lambda_2^n(1-a) & a-1+\lambda_2^n(1-a) \\ (b-1)-\lambda_2^n(b-1) & a-1+\lambda_2^n(b-1) \end{pmatrix}$

(Ce qui permet de demander aux élèves une récurrence)

Remarque 2

Si on écrit $A=\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$, $\lambda_2=1-a-b$, et si $a+b\neq 0$, $\lim_{n\to\infty} A^n=\begin{pmatrix} \dfrac{b}{a+b} & \dfrac{a}{a+b} \\ \dfrac{b}{a+b} & \dfrac{a}{a+b} \end{pmatrix}$,

l'état stable est $\begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}$,

et $A^n = PD^nP^{-1} = \frac{1}{a+b}\begin{pmatrix} b+\lambda_2^n a & a-\lambda_2^n a \\ b-\lambda_2^n b & a+\lambda_2^n b \end{pmatrix}$

Remarque 3

Si $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $A^{2n} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $A^{2n+1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, donc la suite $\left(A^n\right)$ n'a pas de limite, et c'est le seul cas où la limite n'existe pas.

Exemples

1. $\begin{pmatrix} 0,2 & 0,8 \\ 0,6 & 0,4 \end{pmatrix}^n = \begin{pmatrix} \frac{3}{7}+\frac{4}{7}\left(-\frac{2}{5}\right)^n & \frac{4}{7}-\frac{4}{7}\left(-\frac{2}{5}\right)^n \\ \frac{3}{7}-\frac{3}{7}\left(-\frac{2}{5}\right)^n & \frac{4}{7}+\frac{3}{7}\left(-\frac{2}{5}\right)^n \end{pmatrix}$

Et $\lim\limits_{n\to\infty}\begin{pmatrix} 0,2 & 0,8 \\ 0,6 & 0,4 \end{pmatrix}^n = \begin{pmatrix} \frac{3}{7} & \frac{4}{7} \\ \frac{3}{7} & \frac{4}{7} \end{pmatrix}$ et $\lim\limits_{n\to\infty}\begin{pmatrix} x_n & y_n \end{pmatrix} = \begin{pmatrix} \frac{3}{7} & \frac{4}{7} \end{pmatrix}$.

2. $\lim\limits_{n\to\infty}\begin{pmatrix} 0,4 & 0,6 \\ 1 & 0 \end{pmatrix}^n = \begin{pmatrix} \frac{5}{8} & \frac{3}{8} \\ \frac{5}{8} & \frac{3}{8} \end{pmatrix}$

3. $\lim\limits_{n\to\infty}\begin{pmatrix} 0,4 & 0,6 \\ 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$

En dimension supérieure ou égale à 3

- Une matrice stochastique A n'est pas nécessairement diagonalisable.
- 1 est une valeur propre.
- Les autres valeurs propres ne sont pas nécessairement réelles, mais de module inférieur à 1.

Le produit de deux matrices stochastiques est stochastique, donc A^n est stochastique $\forall n \in \mathbb{N}$, donc la suite $\left(A^n\right)$ est bornée.

Mais on peut avoir par exemple :

- Si $A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $A^{2n} = I$ et $A^{2n+1} = A \quad \forall n \in \mathbb{N}$

  Les valeurs propres sont -1 et 1 qui est double.
  Le sous espace propre associé à 1 est de dimension 2.

- Si $A = \begin{pmatrix} 0 & 1 & 0 \\ 0,5 & 0 & 0,5 \\ 0 & 1 & 0 \end{pmatrix}$, $A^{2n} = A^2 = \begin{pmatrix} 0,5 & 0 & 0,5 \\ 0 & 1 & 0 \\ 0,5 & 0 & 0,5 \end{pmatrix}$ et $A^{2n+1} = A \quad \forall n \in \mathbb{N}$

  Les valeurs propres sont -1 ; 0 et 1

Voir le modèle d'urnes de T. & P. Ehrenfest dans le document ressource.

Cependant si $A=\begin{pmatrix}0,5 & 0,5 & 0\\ 0 & 1 & 0\\ 0,5 & 0 & 0,5\end{pmatrix}$, on a $\lim\limits_{n\to\infty} A^n=\begin{pmatrix}0 & 1 & 0\\ 0 & 1 & 0\\ 0 & 1 & 0\end{pmatrix}$.

Les valeurs propres sont 1 et $\frac{1}{2}$ qui est double, mais A non diagonalisable car les deux sous espaces propres sont de dimension 1.
Donc dans ces deux premiers cas, la suite $\left(A^n\right)$ n'a pas de limite.

En revanche (cas particulier du théorème de Perron-Frobenius)
Si tous les coefficients de la matrice ou d'une de ses puissances sont strictement positifs, on peut montrer que :

- toutes les valeurs propres différentes de 1, qui est simple, sont en module strictement inférieures à 1.
- la limite de la suite $\left(A^n\right)$ est une matrice stochastique dont toutes les lignes sont égales (en utilisant au besoin, si la matrice n'est pas diagonalisable, la décomposition de Dunford)

Ceci étant une condition suffisante mais non nécessaire.

Exemples

1. Si $A=\begin{pmatrix}1/2 & 1/3 & 1/6\\ 1/6 & 1/3 & 1/2\\ 1/3 & 1/6 & 1/2\end{pmatrix}$, les valeurs propres sont : 1, $\frac{1}{6}-\frac{i}{6}$ et $\frac{1}{6}+\frac{i}{6}$.

Et on a $\lim\limits_{n\to\infty} A^n=\begin{pmatrix}\frac{9}{26} & \frac{7}{26} & \frac{10}{26}\\ \frac{9}{26} & \frac{7}{26} & \frac{10}{26}\\ \frac{9}{26} & \frac{7}{26} & \frac{10}{26}\end{pmatrix}$.

2. Si $A=\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0,8 & 0,2 & 0\end{pmatrix}$, les valeurs propres sont : 1, $-\frac{1}{2}-\frac{i\sqrt{55}}{10}$ et $-\frac{1}{2}+\frac{i\sqrt{55}}{10}$.

Et on a $A^5=\begin{pmatrix}\frac{4}{25} & \frac{1}{25} & \frac{4}{5}\\ \frac{16}{25} & \frac{8}{25} & \frac{1}{25}\\ \frac{4}{125} & \frac{81}{125} & \frac{8}{25}\end{pmatrix}$ et $\lim\limits_{n\to\infty} A^n=\begin{pmatrix}\frac{2}{7} & \frac{5}{14} & \frac{5}{14}\\ \frac{2}{7} & \frac{5}{14} & \frac{5}{14}\\ \frac{2}{7} & \frac{5}{14} & \frac{5}{14}\end{pmatrix}$.

3. Si $A=\begin{pmatrix}0,2 & 0,5 & 0,3\\ 0,4 & 0,3 & 0,3\\ 0,5 & 0,4 & 0,1\end{pmatrix}$, les valeurs propres sont : 1 et $-\frac{1}{5}$.

La matrice n'est pas diagonalisable.

$$A = D + N = \begin{pmatrix} -\frac{1}{5} & \frac{9}{10} & \frac{3}{10} \\ 0 & \frac{7}{10} & \frac{3}{10} \\ 0 & \frac{9}{10} & \frac{1}{10} \end{pmatrix} + \begin{pmatrix} \frac{2}{5} & -\frac{2}{5} & 0 \\ \frac{2}{5} & -\frac{2}{5} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 \end{pmatrix} \qquad \text{avec } N^2 = 0$$

$$\left(A^n = (D+N)^n = \sum_{k=1}^{n} \binom{n}{k} D^k N^{n-k} = \sum_{k=1}^{n} \binom{n}{k} N^k D^{n-k} = D^n + nND^{n-1} \right.$$

$$\text{Avec } D = \begin{pmatrix} -\frac{1}{5} & \frac{9}{10} & \frac{3}{10} \\ 0 & \frac{7}{10} & \frac{3}{10} \\ 0 & \frac{9}{10} & \frac{1}{10} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & -\frac{1}{3} \\ \frac{1}{2} & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{5} & 0 \\ 0 & 0 & -\frac{1}{5} \end{pmatrix} \times \begin{pmatrix} 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & -\frac{3}{4} & -\frac{1}{4} \\ 0 & -\frac{3}{4} & \frac{3}{4} \end{pmatrix}$$

$$\text{Et on a } \lim_{n\to\infty} A^n = \begin{pmatrix} \frac{51}{144} & \frac{57}{144} & \frac{1}{4} \\ \frac{51}{144} & \frac{57}{144} & \frac{1}{4} \\ \frac{51}{144} & \frac{57}{144} & \frac{1}{4} \end{pmatrix}$$

Mais la condition ci-dessus (à savoir si tous les coefficients de la matrice ou d'une de ses puissances sont strictement positifs) **sur les coefficients n'est pas nécessaire :**

3. Si $A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0,2 & 0,8 \end{pmatrix}$, les valeurs propres sont : 1, $-\frac{1}{5}$ et 0 .

$$\text{Et on a } \lim_{n\to\infty} A^n = \begin{pmatrix} 0 & \frac{1}{6} & \frac{5}{12} \\ 0 & \frac{1}{6} & \frac{5}{12} \\ 0 & \frac{1}{6} & \frac{5}{12} \end{pmatrix}.$$

4. Si $A = \begin{pmatrix} 1 & 0 & 0 \\ 0,2 & 0 & 0,8 \\ 0 & 0,3 & 0,7 \end{pmatrix}$ les valeurs propres sont : 1, $\frac{7-\sqrt{145}}{20}$ et $\frac{7+\sqrt{145}}{20}$.

$$\text{Et on a } \lim_{n\to\infty} A^n = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

La loi uniforme

A. La loi uniforme sur $[a,b]$ est une loi à densité *f* caractérisée par :

- la probabilité que la variable uniforme soit inférieure strictement à *a* est nulle,
- la probabilité que la variable uniforme soit supérieure strictement à b est nulle,
- la probabilité que la variable uniforme appartienne à un intervalle donné est indépendante de la position de cet intervalle, mais dépend seulement de son amplitude, dès lors que cet intervalle est inclus dans [*a* ; *b*], le support de la loi .
- *f* est supposée intégrable sur $\mathbb{R}$ de somme 1.

Ainsi, si une variable aléatoire continue *X* suit la loi uniforme $U\left([a,b]\right)$,

$\forall x<a$, $P(x\leqslant X<a)=\int_x^a f(t)dt=0$ **,** donc **si** *f* est positive sur $\mathbb{R}$, *f* est nulle sur $]-\infty;a[$

$\forall x>b$, $P(b<X\leqslant x)=\int_b^x f(t)dt=0$ **,** donc **si** *f* est positive sur $\mathbb{R}$, *f* est nulle sur $]b;+\infty[$

$\forall\alpha$ et $\forall\beta$ tels que $a\leqslant\alpha\leqslant\beta\leqslant b$, et $\forall x$ tels que $a\leqslant a+x\leqslant b+x\leqslant b$

$P(\alpha\leqslant X\leqslant\beta)=\int_\alpha^\beta f(t)dt=\int_{\alpha+x}^{\beta+x} f(t)dt=P(\alpha+x\leqslant X\leqslant\beta+x)$

avec la relation de Chasles :

$$\int_\alpha^\beta f(t)dt=\int_{\alpha+x}^{\alpha} f(t)dt+\int_\alpha^\beta f(t)dt+\int_\beta^{\beta+x} f(t)dt$$

$$-\int_{\alpha+x}^{\alpha} f(t)dt=\int_\beta^{\beta+x} f(t)dt$$

$$\text{soit } \int_\alpha^{\alpha+x} f(t)dt=\int_\beta^{\beta+x} f(t)dt$$

en particulier, $\forall x>0$ $\frac{1}{x}\int_\alpha^{\alpha+x} f(t)dt=\frac{1}{x}\int_\beta^{\beta+x} f(t)dt$ **,** et **si** *f* est **continue** en α et β,

à la limite, $f(\alpha)=f(\beta)$ **:**

$$\frac{1}{x}\int_\alpha^{\alpha+x} f(t)dt-f(\alpha)=\frac{1}{x}\int_\alpha^{\alpha+x} f(t)dt-\frac{1}{x}\int_\alpha^{\alpha+x} f(\alpha)dt=\frac{1}{x}\int_\alpha^{\alpha+x}\left(f(t)-f(\alpha)\right)dt$$

or, $\forall\varepsilon>0,\exists\eta>0$ tel que $|t-\alpha|<\eta\Rightarrow\quad|f(t)-f(\alpha)|<\varepsilon$

donc, $\forall\varepsilon>0,\exists\eta>0$ tel que $0<x<\eta\ \Rightarrow$

$$\left|\frac{1}{x}\int_\alpha^{\alpha+x} f(t)dt-f(\alpha)\right|=\left|\frac{1}{x}\int_\alpha^{\alpha+x}\left(f(t)-f(\alpha)\right)dt\right|\leqslant\frac{1}{x}\int_\alpha^{\alpha+x}\left|f(t)-f(\alpha)\right|dt\leqslant\frac{1}{x}\int_\alpha^{\alpha+x}\varepsilon\,dt=\varepsilon$$

c'est-à-dire que $\lim\limits_{x\to0}\frac{1}{x}\int_\alpha^{\alpha+x} f(t)dt=f(\alpha)$.

Et **si** *f* est **continue** sur $[a,b]$, *f* est constante sur $[a,b]$.

Ou bien (plus simplement) :

pour deux intervalles d'amplitude *h* inclus dans le support,

$\int_\alpha^{\alpha+h} f(t)dt=\int_\beta^{\beta+h} f(t)dt$, si *f* est continue sur $[a,b]$, elle admet une primitive *F*, d'où

$$F(\beta+h)-F(\beta)=F(\alpha+h)-F(\alpha)$$

Pour $h \neq 0$, $\dfrac{F(\beta+h)-F(\beta)}{h} = \dfrac{F(\alpha+h)-F(\alpha)}{h}$

$$\lim_{h\to\infty} \frac{F(\beta+h)-F(\beta)}{h} = \lim_{h\to\infty} \frac{F(\alpha+h)-F(\alpha)}{h}$$

$F'(\beta)=F'(\alpha)$ puisque f est dérivable

C'est-à-dire $f(\alpha) = f(\beta)$.

Ainsi, $P(\alpha \leqslant X \leqslant \beta)$ est proportionnelle à $\beta-\alpha$, en particulier, pour $x \in [a,b]$,

$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(t)dt = \int_{a}^{x} f(t)dt = \lambda x$ et donc $F'(x) = f(x) = \lambda$

Et puisque $\int_{-\infty}^{+\infty} f(t)dt = \int_{a}^{b} f(t)dt = \lambda(b-a) = 1$, f est définie sur $\mathbb{R}$ par

$$\begin{cases} f(x) = \dfrac{1}{b-a} & x \in [a,b] \\ f(x) = 0 & \text{sinon} \end{cases}$$

Conclusion

Si une variable aléatoire continue X suit la loi uniforme $U([a,b])$

$P(\alpha \leqslant X \leqslant \beta) = \dfrac{1}{b-a}\int_{\alpha}^{\beta} dt = \dfrac{\beta-\alpha}{b-a}$ si $a \leqslant \alpha \leqslant \beta \leqslant b$

On a : $E(X) = \dfrac{a+b}{2}$ et $\text{Var(X)} = \dfrac{(b-a)^2}{12}$ donc $\sigma(X) = \dfrac{b-a}{2\sqrt{3}}$.

Sa fonction de répartition est définie sur $\mathbb{R}$ par :

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(t)dt \quad \text{donc} \quad F(x) = \begin{cases} 0 & si\ x \leqslant a \\ \dfrac{x-a}{b-a} & si\ a < x \leqslant b \\ 1 & si\ x > \text{b} \end{cases}$$

Exemples

1. Si l'on a rendez-vous avec une personne entre 10h et 11h (sans autre information), alors la probabilité que ce rendez-vous ait lieu entre 10h05 min et 10h10 min est $\dfrac{5}{60} = \dfrac{1}{12}$;

on utilise, faute de mieux, la loi $U([0,60])$.

Cependant, l'absence d'information ne justifie pas le caractère uniforme de la loi.
En fait il est bien difficile, nous semble-t-il, de trouver une situation concrète qui relève réellement de la loi uniforme.

2. Il est cependant un cas où la loi uniforme apparait de façon naturelle (non artificielle) :
On considère une variable aléatoire continue X suivant une loi de densité f, f continue sur un intervalle d'extrémités a et b éventuellement infinie(s) et nulle ailleurs.

Si on définit la variable aléatoire Y par $Y = F(X) = \int_{-\infty}^{X} f(t)dt$, alors la loi de Y est $U([0,1])$.

En effet : Y prend ses valeurs dans $[0,1]$ et si F_Y est la fonction de répartition de la loi de Y,

$$F_Y(x) = P(Y \leqslant x) = P(F(X) \leqslant x)$$

Puisque *f* est continue et strictement positive sur $[a;b]$ et nulle ailleurs (si a et/ou b sont infinis, il faut ouvrir les crochets en conséquence) alors *F* est continue et strictement croissante de $[a;b]$ sur $[0;1]$, et définit une bijection de $[0;1]$ sur $[a;b]$.

On note alors F^{-1} la fonction réciproque de la restriction de *F à* $[a;b]$, définie sur $[0;1]$

- $\forall x \in [0;1]$: $F(X) \leqslant x \Leftrightarrow F^{-1} \circ F(X) \leqslant F^{-1}(x)$ car F^{-1} est strictement croissante sur $[0;1]$

  Donc $F_Y(x) = P(F(X) \leqslant x) = P(X \leqslant F^{-1}(x)) = F(F^{-1}(x)) = x.$

- $\forall x < 0 \quad F_Y(x) = 0$ $\qquad \left(P(Y<0) = P\left(\int_{-\infty}^{X} F_Y(x) < 0\right) = 0\right)$
- $\forall x > 1 \quad F_Y(x) = 1$ $\qquad \left(P(Y>1) = P\left(\int_{-\infty}^{X} F_Y(x) > 1\right) = 1\right)$

F_Y est la fonction de répartition de la loi $U([0,1])$

Ainsi *Y* suit la loi $U([0,1])$.

3. La touche Randon des calculatrices simule une variable suivant la loi $U([0,1])$

Une application : on peut simuler n'importe quelle loi à l'aide de la loi uniforme si on connait explicitement sa fonction de répartition réciproque F^{-1}.

Exemple : Pour *X* suivant la loi exponentielle de paramètre λ , la fonction de répartition est $F(x) = 1 - e^{-\lambda x}$ pour $x > 0$ et 0 sinon.

$$\begin{cases} F^{-1}(0) = 0 \\ F^{-1}(x) = -\dfrac{1}{\lambda}\ln(1-x) \qquad \text{si } x \in]0;1[\\ \lim\limits_{x \to 1} F^{-1}(x) = +\infty \end{cases}$$

D'où l'algorithme de simulation : *(Puisque (Ramdon) et (1 – Ramdon) suivent la même loi)*

$$X \longleftarrow -\frac{\ln(Ramdon)}{\lambda}$$

B. La loi uniforme sur $[0;1]^2$ est une loi à densité *f* caractérisée par :

- si $A \subset \mathbb{R}^2$ et $A \cap [0;1]^2 = \varnothing$ alors $P((X,Y) \in A) = 0$,
- si $A \subset \mathbb{R}^2$ et $A \subset [0;1]^2$ alors $P((X,Y) \in A)$ est indépendante de la position de cet ensemble, mais dépend seulement de son aire.
- *f* est supposée intégrable sur $\mathbb{R}^2$ de somme 1, positive et continue sur $[0;1]^2$.

Ainsi, si une variable aléatoire continue *X* suit la loi uniforme $U([0,1]^2)$,

pour deux pavés élémentaires d'aire *hk* inclus dans le support :

$$\int_b^{b+k}\int_a^{a+h} f(x,y)dxdy = \int_d^{d+k}\int_c^{c+h} f(x,y)dxdy$$

avec le théorème de Fubini :

$$\int_b^{b+k}\left(\int_a^{a+h} f(x,y)dx\right)dy = \int_d^{d+k}\left(\int_c^{c+h} f(x,y)dx\right)dy$$

si f est continue sur $[0;1]^2$, l'application $x \longrightarrow f(x,y)$ est continue sur $[0;1]$ donc elle admet une primitive G :

$$\int_b^{b+k}\big(G(a+h,y)-G(a,y)\big)dy = \int_d^{d+k}\big(G(c+h,y)-G(c,y)\big)dy$$

Et l'application $y \longrightarrow G(x,y)$ est continue sur $[0;1]$ et elle admet une primitive F :

$$F(a+h,b+k)-F(a+h,b)-F(a,b+k)+F(a,b)$$
$$= F(c+h,d+k)-F(c+h,d)-F(c,d+k)+F(c,d)$$

On a alors

$$\lim_{k\to 0}\frac{1}{k}\left(\lim_{h\to 0}\frac{F(a+h,b+k)-F(a+h,b)-F(a,b+k)+F(a,b)}{h}\right) = \lim_{k\to 0}\frac{1}{k}\big(F_x'(a,b+k)-F_x'(a,b)\big)$$
$$= F_{xy}''(a,b) = f(a,b)$$

Et de même pour le second membre,

Ainsi $f(a,b) = f(c,d) = \lambda$ pour tout $\forall(a,b)\in[0;1]^2$ et $\forall(c,d)\in[0;1]^2$

f est constante sur $[0;1]^2$ et il est facile de voir qu'elle est nulle sur son complémentaire (puisque positive).

Enfin $\int_{\mathbb{R}^2} f(x,y)dxdy = \int_{[0\,,\,1]^2} f(x,y)dxdy = \int_0^1\left(\int_0^1 \lambda dx\right)dy = \lambda = 1$

La densité de la loi uniforme sur $[0;1]^2$ est donc $1_{[0;1]^2}$.

Loi de durée de vie sans vieillissement

T est la variable aléatoire : « durée de vie d'un individu », à valeur dans $[0;+\infty[$

$T \geqslant t$ signifie que l'individu est vivant l'instant t.

Si, pour tout $t \geqslant 0$ la probabilité que l'individu soit vivant à l'instant $t+h$, pour tout $h \geqslant 0$, sachant que l'individu est vivant l'instant t ne dépend pas de t :

$P_{(T\geqslant t)}(T \geqslant t+h)$ ne dépend pas de t pour tout $t \geqslant 0$ et pour tout $h \geqslant 0$.

On dit que T suit une loi de durée de vie sans vieillissement.

a. Toute loi exponentielle est une loi de durée de vie sans vieillissement :

$$P_{(T\geqslant t)}(T \geqslant t+h) = \frac{P\left((T \geqslant t+h)\cap(T \geqslant t)\right)}{P(X \geqslant t)} = \frac{P(T \geqslant t+h)}{P(X \geqslant t)} = \frac{e^{-\lambda(h+t)}}{e^{-\lambda t}} = e^{-\lambda h}$$

b. Réciproquement, une loi à densité qui « loi de durée de vie sans vieillissement » est une loi exponentielle.

En effet si $P_{(T\geqslant t)}(T \geqslant t+h)$ ne dépend pas de t pour tout $t \geqslant 0$ et pour tout $h \geqslant 0$.

En particulier :

$$P_{(T\geqslant t)}(T \geqslant t+h) = P_{(T\geqslant 0)}(T \geqslant t+0) == \frac{P\left((T \geqslant h)\cap(T \geqslant 0)\right)}{P(T \geqslant 0)} = \frac{P(T \geqslant h)}{P(T \geqslant 0)}$$

Donc $P_{(T\geqslant t)}(T \geqslant t+h) = P(T \geqslant h)$ puisque $P(T \geqslant 0) = 1$.

D'autre part :

$$P_{(T\geqslant t)}(T \geqslant t+h) = \frac{P\left((T \geqslant t+h)\cap(T \geqslant t)\right)}{P(X \geqslant t)} = \frac{P(T \geqslant t+h)}{P(X \geqslant t)}$$

D'où $$\frac{P(T \geqslant t+h)}{P(T \geqslant t)} = P(T \geqslant h)$$

On pose $F(t)=P(T \leqslant t)$ pour $t \geqslant 0$, (la fonction de répartition de la loi)

et $G(t)=1-F(t)$

la relation précédente s'écrit $\frac{G(t+h)}{G(t)} = G(h)$ soit $G(t+h) = G(h)\times G(t)$

Comme G est dérivable sur $[0;+\infty[$, cette équation fonctionnelle conduit à l'équation

$G'(t+h) = G'(h)\times G(t)$ (on dérive par rapport à h)

Et en particulier : $G'(t) = G'(0)\times G'(t)$ avec $G(0) = 1$

Les solutions de cette équation différentielle comme G est non nulle, sont de la forme $G(t) = e^{at}$.

Comme $\lim\limits_{t\to+\infty} G(t) = \lim\limits_{t\to+\infty}(1-F(t)) = 0$, ceci impose que $a = -\lambda$ avec $\lambda > 0$

D'où

La fonction de répartition est $F(t) = P(T \leqslant t) = 1-G(t) = 1-e^{-\lambda t}$ *pour* $t \geqslant 0$

Et la fonction de densité est $F'(t) = f(t) = \lambda e^{-\lambda t}$ $t \geqslant 0$

Conclusion : la v.a. T suit une loi exponentielle de paramètre λ.

De façon plus transversale mais à la limite du programme :

« la loi exponentielle à partir de la désintégration radioactive des noyaux »

La loi de décroissance radioactive de Rutherford et Soddy stipule que la probabilité qu'un noyau se désintégrant durant l'intervalle de temps Δt est égale à $\lambda \Delta t$ où λ est une constante réelle strictement positive, laquelle ne dépend que de la nature du noyau.
Ainsi la probabilité qu'un noyau soit désintégré à l'instant $t + h$, ne l'étant pas bien sûr à l'instant t, ne dépend pas de t, mais seulement de λ ;
ce qui signifie que la désintégration d'un noyau -
- ne dépend pas de son « âge »,
- ni du nombre de noyaux
- ni de la désintégration ou non des noyaux environnants.

Si T est la variable aléatoire : « durée de vie d'un noyau »

La loi de décroissance radioactive de Rutherford et Soddy se traduit par :

$P(T<t+\Delta t/T\geqslant t)=\lambda\Delta t$ ou $P(T<t+h/T\geqslant t)=\lambda h$

or $P(T<t+h/T\geqslant t)=\dfrac{P\big((T<t+h)\cap(T\geqslant t)\big)}{P(T\geqslant t)}$ (probabilité conditionnelle)

on pose $F(t)=P(T<t)$ donc $P(T\geqslant t)=1-F(t)$.

On a $P((T<t+h)\cap(T\geqslant t))=P(t\leqslant T\leqslant t+h)=P(T\leqslant t+h)-P(T\leqslant t)=F(t+h)-F(t)$

D'où $P(T<t+h/T\geqslant t)=\dfrac{F(t+h)-F(t)}{1-F(t)}=\lambda h$

D'où $\dfrac{F(t+h)-F(t)}{h}\times\dfrac{1}{1-F(t)}=\lambda$ et si on fait tendre λ vers 0,

$F'(t)\times\dfrac{1}{1-F(t)}=\lambda$ soit $F'(t)+\lambda F(t)=\lambda$ $t\in[0;+\infty[$.

Donc F est solution du système $\begin{cases} y'+\lambda y=\lambda \\ y(0)=0 \end{cases}$

c'est à dire $F(t)=Ke^{-\lambda t}-\dfrac{\lambda}{-\lambda}=Ke^{-\lambda t}+1$ et $F(t)=0\Rightarrow K=-1$

d'où enfin $P(T<t)=F(t)=1-e^{-\lambda t}$ et $F'(t)=f(t)=\lambda e^{-\lambda t}$

T suit la loi exponentielle de paramètre λ

La fonction de densité est $f(t)=\lambda e^{-\lambda t}$ si $t\geqslant 0$ et $f(t)=0$ sinon

La fonction de répartition est $P(T<t)=\int_0^t \lambda e^{-\lambda t}dt=1-e^{-\lambda t}$ $t\geqslant 0$

et donc $P(T\geqslant t)=1-\int_0^t \lambda e^{-\lambda t}dt=e^{-\lambda t}$ $t\geqslant 0$

$E(X)=\dfrac{1}{\lambda}$ et $\mathrm{Var(X)}=\dfrac{1}{\lambda^2}$ donc $\sigma(\mathrm{X})=\dfrac{1}{\lambda}$ (en recherchant des primitives de la forme $t\longrightarrow p(t)e^{\lambda t}$)

Invariance de certaines familles de distributions par transformation affine.

- On sait que si X suit une loi de probabilité admettant une espérance et une variance, pour tous réels a et b :

  $E(aX+b)=aE(X)+b$ et $V(aX+b)=a^2V(X)$ donc $\sigma(aX+b)=|a|\sigma(X)$

  On obtient donc l'espérance et la variance de la variable $Y=aX+b$ obtenue par transformation affine de la variable X.
  Il est donc possible, par exemple, de centrer et réduire toute loi de variance non nulle (la variable aléatoire n'est pas constante) en considérant la variable $Z=\dfrac{X-\mu}{\sigma}$.

- Dans le cas particulier de la loi normale, une transformation affine de la variable ne change pas la nature de la loi, mais seulement ses paramètres.

Si X suit la loi de densité définie sur $\mathbb{R}$ par $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$, la variable aléatoire $Y=aX+b$, $a\neq 0$, suit la loi de densité définie sur $\mathbb{R}$ par $f(x)=\dfrac{1}{|a|}\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-(a\mu+b)}{a\sigma}\right)^2}$,

En effet, en regardant la fonction de densité, si $a>0$:

$$G(t)=P(Y\leqslant t)=P(aX+b\leqslant t)=P\left(X\leqslant\frac{t-b}{a}\right)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\frac{t-b}{a}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}dx,$$

et on peut justifier que : $G'(t)=\left(\dfrac{t-b}{a}\right)'\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\frac{t-b}{a}-m}{\sigma}\right)^2}=\dfrac{1}{a}\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\frac{t-b}{a}-m}{\sigma}\right)^2}$.

Si $a<0$:

$$G(t)=P(Y\leqslant t)=P(aX+b\leqslant t)=P\left(X\geqslant\frac{t-b}{a}\right)=1-\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\frac{t-b}{a}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}dx$$

$$G'(t)=-\left(\frac{t-b}{a}\right)'\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\frac{t-b}{a}-m}{\sigma}\right)^2}=\frac{1}{|a|}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\frac{t-b}{a}-m}{\sigma}\right)^2}$$

On reconnait donc la loi normale d'espérance $a\mu+b$ et d'écart-type $|a|\mu$.

(si $a=0$, la v.a. Y ne prend qu'une valeur et $P(Y\leqslant t)=1$ si $t\geqslant b$, $P(Y\leqslant t)=0$ sinon)

- Pour une loi exponentielle de paramètre λ, on peut voir que la fonction de densité de la loi de la variable centrée réduite est : $\begin{cases}f(x)=\dfrac{1}{e}e^{-x} \text{ si } x\geqslant -1\\ f(x)=0 \quad \text{sinon}\end{cases}$

En effet :

$$H(t) = P(Y \leqslant t) = P\left(\frac{X - 1/\lambda}{1/\lambda} \leqslant t\right) = P(\lambda X - 1 \leqslant t) = \int_0^{\lambda t - 1} \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$$

$$H(t) = P\left(X \leqslant \frac{t+1}{\lambda}\right) = \int_0^{\frac{t+1}{\lambda}} \lambda e^{-\lambda x} dx = 1 - e^{-t-1}$$

$H'(t) = e^{-t-1}$.

Ce n'est donc plus exactement une loi exponentielle.

- En revanche une loi uniforme reste par transformation affine, en particulier centrage et réduction, une loi uniforme.

- La loi binomiale, et en particulier la loi de Bernoulli, n'est plus à support entier positif quand on la centre et réduit.